Montpellier SupAgro
Centre International d'Études Supérieures en Sciences Agronomiques

THESE
présentée pour obtenir le diplôme de Doctorat

Discipline :
Évolution, Écologie, Ressources Génétiques, Paléontologie

Ecole doctorale :
SIBAGHE – Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement

Diversité et évolution des paysages nucléotidiques des plantes

par Laurana Serres-Giardi

Sous la direction de Jacques David et Sylvain Glémin

LABORATOIRES D'ACCUEIL :
ISEM – Institut des Sciences de l'Évolution de Montpellier – UM2 CNRS UMR 5554
AGAP – Amélioration Génétique et Adaptation des Plantes – INRA SupAgro UMR 1334

SOUTENUE LE 28 JUIN 2012 DEVANT LE JURY COMPOSE DE :

Vincent Castric, Chargé de Recherche, CNRS Lille	Rapporteur
Jérôme Salse, Directeur de Recherche, INRA Clermont-Ferrand	Rapporteur
Pierre Boursot, Directeur de Recherche, CNRS Montpellier	Examinateur
Sylvain Glémin, Chargé de Recherche, CNRS Montpellier	Directeur de thèse

```
# Script remerciements de thèse
# LSG, été 2012


#.........Libraries
library(thanks)
library(mail)

#.........Elements
jury_de_thèse <- c("Vincent Castric", "Jérôme Salse", "Pierre Boursot")
encadrants <- c("Sylvain", "Jacques")
collègues <- c("Nathalie", "Marie-Hélène", "Muriel", "Joëlle", "Marie-France",
"Martin", "Aline", "Laurène", "Concetta", "Pierre", "Jean-Marie", "agap",
"davem", "les gens de Mauguio", "Adrienne", "Khalid", "Aline", "Philippe",
"l'équipe phylmol", "les paléontos", "les gens de l'ISEM", "les lyonnais",
"Agnès et les collègues de l'enseignement")
cothésards <- c("Marie", "Laurence", "YungFen", "Greg", "Ingrid", "Constance",
"Jonathan", "João", "Fidel")
les_girapoix <- c("l'Impératrice suprême", "notre gourou à tous",
"trésorier_bien_shampouiné", "Anushka", "Mumu", "le baigneur", "l'amateur de
biches", "Romain", "Flore", "l'huître rieuse")
mes_proches <- c("famille", "amis", "Rodentia", "Lagomorpha")

pan <- c(jury_de_thèse, encadrants, collègues, cothésards, les_girapoix,
mes_proches)

#.........Using thanks functions
for j in (1:length(jury_de_thèse))
  { thanksalot(jury_de_thèse[j], for=c("d'avoir accepté d'évaluer mon travail",
"pour vos commentaires et vos questions constructives") }

dualthanks(encadrants, for="pour votre encadrement, votre passion de la
recherche (contagieuse) et votre humanité")
doublethanks(encadrants[1], for="pour ta disponibilité au quotidien, ta rigueur
de réflexion, ta patience")
doublethanks(encadrants[2], for="pour ton enthousiasme, tes idées
foisonnantes... et pour m'avoir entraînée dans cette histoire !")

individualthanks(collègues, for=c(discussions, aide, soutien, cafés,
moultes_et_moultes_choses), chaleureusement=TRUE)
thanksthanksthanks(cothésards, veryveryveryspecialmention=first.element,
pintadeoption=TRUE)
greenspottedthanks(les_girapoix, for=traductfromananas("$^ù*é$^&^ù*µé&&é~*^ù"),
option=M&M's, web=www.despoisetdesgirafes.com)
thanksthanksthanksandrethanks(mes_proches, carrotmention=TRUE)
thanksofcourse(R, perl, bash, worldpeople="tous les gens qui ont contribué à
l'alimentation en données des bibliothèques publiques de séquences")
absolutelynothanks(chaleur_montpelliéraine, ma_montre)

#.........Using mail function
for p in (1:length(pan))
  {sendmail(pan[p], subject="Merci !", message=remerciements_personnalisés[p])}




#                                                        O..O
#......................................................"..o.."........
```

une pensée pour Laurence

Mise en garde au lecteur

Tout aventureux lecteur doit être averti que, contrairement aux graminées, ce document est fortement biaisé en faveur des lettres A et T, au détriment des G et C (AT/(AT+GC) = 0,69).

Diversity and Evolution of Nucleotide Landscapes in Plants

The nucleotide landscape – the way base composition varies along a genome – is a striking feature of genome organization and is highly variable between species. The evolutionary causes of such heterogeneity in GC content have been much debated. Biased gene conversion towards G and C (BGC) and selection on codon usage (SCU) are thought to be main forces. BGC is a neutral process associated with recombination favouring G and C alleles over A and T ones. SCU is a selection process favouring the so-called "preferred" codons, i.e., those whose translation is the most efficient. Contrary to vertebrates, plant nucleotide landscapes are still poorly known. Most studies focused on the GC-poor and homogeneous *Arabidopsis thaliana* genome and on the GC-rich and heterogeneous rice genome. The contrast between these two genomes was often generalized as a dicot/monocot dichotomy but this vision is clearly phylogenetically biased.

The objectives of this study are to characterize angiosperm nucleotide landscapes on a wide phylogenetic scale and to better understand the evolutionary mechanisms acting upon the evolution of nucleotide landscapes. To what extent do nucleotide landscapes vary across angiosperm phylogeny? Are nucleotide landscapes shaped by BGC and SCU? Are taxa affected with the same intensity?

To tackle these issues, I used a comparative genomic approach relying on EST data analysis on over 230 angiosperm and gymnosperm species. Through the nucleotide landscape survey for such a wide range of species I found a continuum of GC-heterogeneity patterns across phylogeny, some taxa such as Poaceae being strikingly GC-rich and heterogeneous. My results suggest that nucleotide landscapes could have been shaped by BGC and, to a lesser extent, by SCU. GC-content enrichment and impoverishment are likely to have occurred several times independently during plant evolution and could be explained by intensity variations of BGC and SCU. Using a proxy for EST expression level, I also characterized the diversity of preferred codons between species. Codon usage preferences were shown to be evolutionarily more unstable for four- and six-fold degenerate codon families. Finally, I could link the evolution of codon usage preferences to the evolution of genome base composition. My results suggest that genome base composition, partially shaped by BGC, seems to drive the coevolution between codon usage preferences and tRNAs.

Keywords: base composition, biased gene conversion towards G and C, selection on codon usage, codon usage preferences, Gymnosperms, Angiosperms, Expressed Sequence Tags.

Diversité et évolution des paysages nucléotidiques des plantes

Le paysage nucléotidique – la manière dont la composition nucléotidique varie le long du génome – est une caractéristique marquante de l'organisation des génomes et varie fortement entre espèces. Plusieurs hypothèses émergent des nombreux débats autour des mécanismes évolutifs à l'origine de ces hétérogénéités du taux de GC, parmi lesquelles la conversion génique biaisée vers G et C (BGC) et la sélection sur l'usage du code (SUC). La BGC est un processus neutre associé à la recombinaison qui favorise les allèles en G ou C au détriment des allèles en A ou T. La SUC est une force de sélection qui favorise les codons dits « préférés », ceux dont la traduction serait la plus efficace. Contrairement à ceux des vertébrés, les paysages nucléotidiques des plantes sont peu connus. La plupart des études ont été consacrées au génome d'*Arabidopsis thaliana*, pauvre en GC et homogène, et à celui du riz, riche en GC et hétérogène. Le contraste entre ces deux génomes a souvent été généralisé comme une dichotomie entre dicotylédones et monocotylédones, mais cette vision est clairement phylogénétiquement biaisée.

Les objectifs de ce travail de thèse sont de caractériser les paysages nucléotidiques des angiospermes à une large échelle phylogénétique et de mieux comprendre quels sont les mécanismes évolutifs jouant sur l'évolution de ces paysages nucléotidiques. Comment varient les paysages nucléotidiques le long de la phylogénie des angiospermes ? SUC et BGC façonnent-elles ces paysages nucléotidiques ? Les différents taxons sont-ils affectés avec la même intensité ?

Pour répondre à ces questions, j'ai utilisé une approche de génomique comparative basée sur l'analyse de données EST chez plus de 230 espèces d'angiospermes et de gymnospermes. L'exploration des paysages nucléotidiques de ce large éventail de plantes a montré que les patrons d'hétérogénéité des paysages nucléotidiques suivent un continuum le long de la phylogénie, avec des groupes particulièrement riches et hétérogènes en GC, les graminées par exemple. Mes résultats suggèrent que les paysages nucléotidiques des plantes pourraient avoir été façonnés par la BGC et, dans une moindre mesure, par la SUC. Des épisodes indépendants d'enrichissement et d'appauvrissement en GC ont vraisemblablement eu lieu au cours de l'évolution des plantes, et pourraient être expliqués par des variations d'intensité de ces mécanismes. En utilisant une prédiction du degré d'expression des EST, j'ai également mis en évidence une diversité dans les codons préférés entre espèces. Les préférences d'usage des codons se sont révélées plus labiles au cours de l'évolution pour les codons des acides aminés au code quatre et six fois dégénéré. J'ai pu lier l'évolution des préférences d'usage des codons à l'évolution de la composition nucléotidique des génomes. Mes résultats suggèrent que la composition en base des génomes, affectée en partie par la BGC, orienterait la coévolution entre préférence d'usage du code et ARNt.

Mots-clés : composition nucléotidique, conversion génique biaisée vers G et C, sélection sur l'usage du code, préférences d'usage du code, gymnospermes, angiospermes, *Expressed Sequence Tags*.

Sommaire

Introduction générale : Géographie et orogenèse des paysages nucléotidiques 1
I. Hétérogénéité des paysages nucléotidiques des eucaryotes 1
II. Hypothèses sélectionnistes quant à l'émergence de paysages nucléotidiques accidentés 8
III. Hypothèses neutralistes quant à l'émergence de paysages nucléotidiques accidentés 14
IV. Evolution des paysages nucléotidiques des plantes 18
Références bibliographiques 22

Chapitre 1 : Patrons et évolution des paysages nucléotidiques des plantes 29
Introduction 29
Résultats 31
Discussion 38
Méthodes 42
Références bibliographiques 45

Chapitre 2 : Evolution des préférences d'usage du code chez les plantes 48
Introduction 48
Matériel et méthodes 59
Résultats 63
Discussion 87
Références bibliographiques 100

Conclusion générale et perspectives 107
I. Paysages nucléotidiques des plantes : un continuum de paysages hétérogènes vraisemblablement façonnés (en partie) par la BGC et la SUC 107
II. Vers une connaissance plus complète des paysages nucléotidiques 108
III. Vers une meilleure connaissance de l'intensité de la SUC et de la BGC 109
IV. Mieux comprendre les facteurs génomiques à l'origine de l'évolution de la BGC et de la SUC 111

Annexes 113
Annexe de l'introduction .....
Annexes du chapitre 1 .....
Annexes du chapitre 2 .....

Introduction générale

Géographie et orogenèse des paysages nucléotidiques

I. Hétérogénéité des paysages nucléotidiques des eucaryotes

A. Découverte de la structure en isochore des génomes

A la fin des années 1960, des travaux basés sur l'étude des gradients de densité de l'ADN après centrifugation en présence de ligands spécifiques ont permis de découvrir l'hétérogénéité de la composition nucléotidique des génomes eucaryotes. D'après ces premiers travaux, menés sur quelques espèces de vertébrés à sang chaud, les génomes nucléaires sont des mosaïques de longues régions d'ADN (> 200 kilobases) caractérisées par différentes proportions de bases G et C par rapport aux bases A et T (Bernardi et al. 1985) (Figure 1 A et B). Ces régions ont été baptisées *isochores*, d'après le grec ισος, égal, et χορος, région.

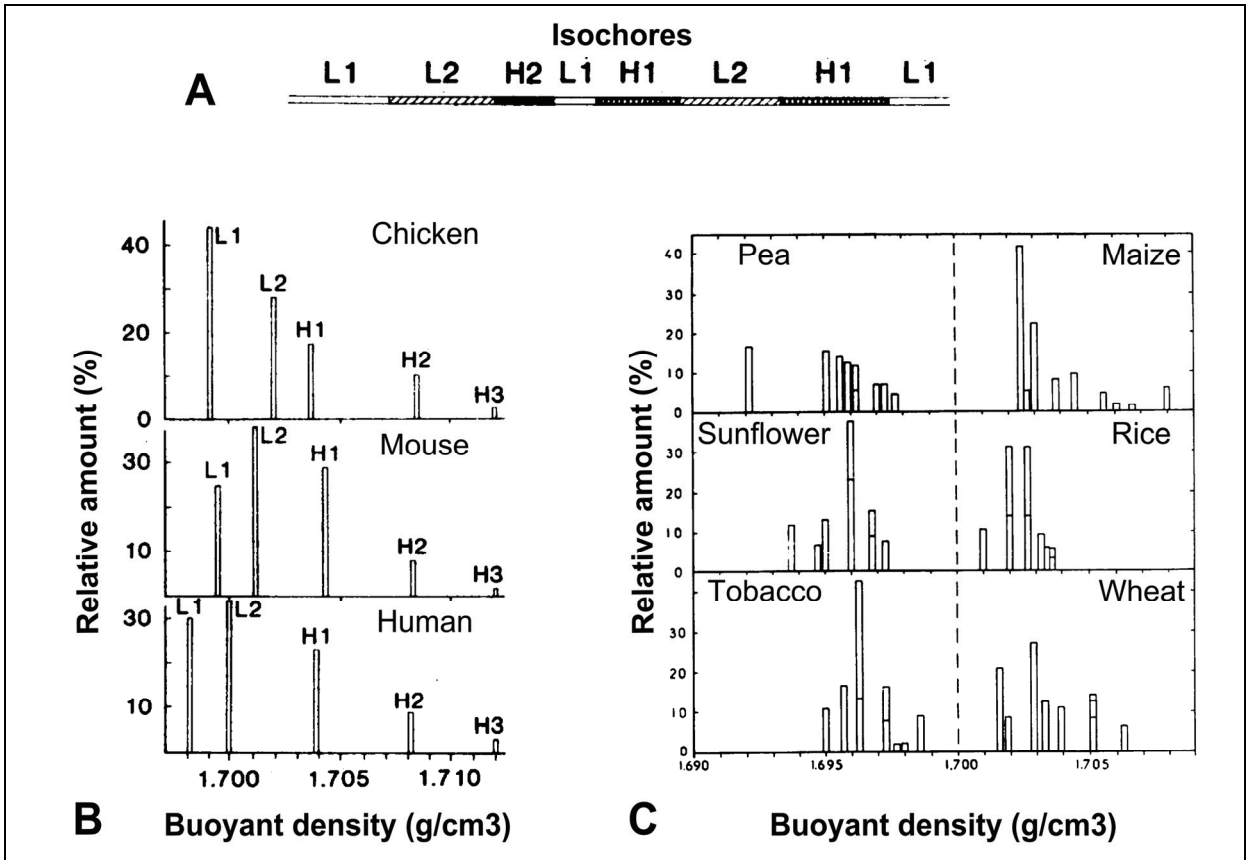


Figure 1 – Organisation des génomes en une mosaïque d'isochores

A). Schématisation de la succession d'isochores riches en GC (H1 et H2) et pauvres en GC (L1 et L2) le long d'un chromosome. **B)**. Histogrammes montrant les proportions des isochores mis en évidence par gradients de densité chez trois vertébrés. **C)**. Histogrammes montrant les proportions des isochores mis en évidence par gradients de densité chez six plantes.

(**A** et **B** : tirés de Bernardi et al. 1985 ; **C** : tiré de Salinas et al. 1988)

La présence d'isochores a également été mise en évidence par les mêmes méthodes chez des plantes, le pois, le tournesol, le tabac, le riz, le maïs et le blé notamment (Salinas et al. 1988) (Figure 1 C). Les isochores correspondraient globalement aux bandes chromosomiques obtenues par coloration au Giemsa (Cuny et al. 1981 ; Bernardi et al. 1985 ; Niimura & Gojobori 2002) (Figure 2).

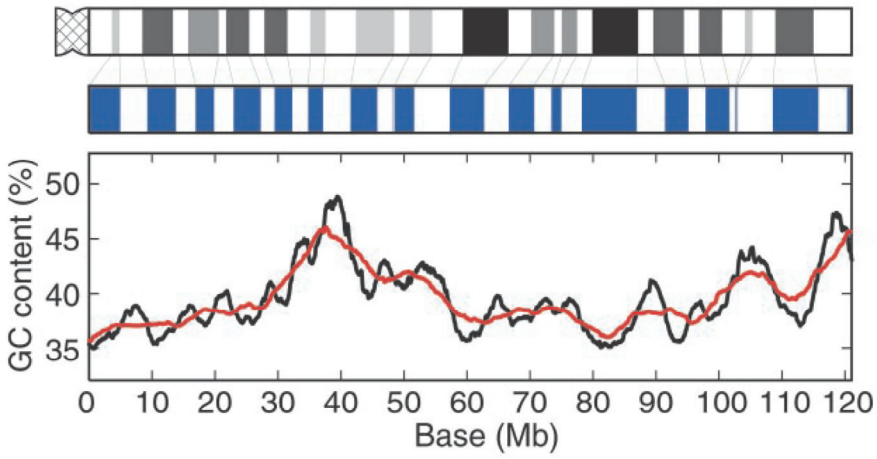


Figure 2 – Correspondance entre bandes chromosomiques obtenues par coloration Giemsa et bandes de composition nucléotidique homogène déterminées *in silico*

Variations du taux de GC le long d'un bras chromosomique humain (en rouge : fenêtre glissante de 9,3 Mb, pas de 10 kb ; en noir : fenêtre glissante de 2,5 Mb, pas de 10 kb). Diagramme en gris : bandes obtenues expérimentalement par coloration Giemsa. Diagramme en bleu : bandes de composition nucléotidique déterminées *in silico* à partir de la séquence génomique (fenêtre glissante de 2,5 Mb)

(Tiré de Niimura & Gojobori 2002)

B. Les paysages nucléotidiques à l'ère du séquençage

L'avènement de l'ère du séquençage a permis de confirmer l'hétérogénéité de la composition nucléotidique au sein des génomes. Du séquençage de quelques gènes au séquençage de génomes complets, la connaissance des paysages nucléotidiques – la manière dont la composition nucléotidique varie le long des génomes – s'est affinée. Le taux de GC s'est effectivement avéré bien variable au sein des génomes, mais l'analyse des séquences le long des chromosomes a montré que le taux de GC varie plutôt par une succession de petites oscillations

locales que par une juxtaposition nette de grandes régions homogènes, ainsi qu'avaient été définies les isochores (Figure 3). La vision initiale des isochores a donc été mise à mal et le terme de *structure en isochore* est aujourd'hui employé dans un sens plus large pour qualifier des paysages nucléotidiques marqués par de fortes variations locales de taux de GC. Le taux de GC est également variable entre espèces, avec des valeurs moyennes variant de 20 % à 60 % chez les eucaryotes (Lynch 2007).

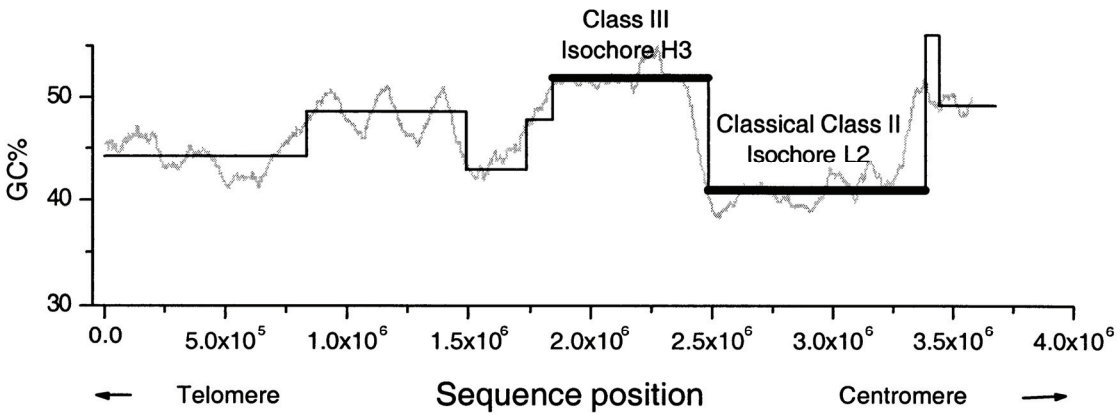


Figure 3 – Contraste entre la vision initiale des isochores et les variations réelles du taux de GC

Variations de la composition nucléotidique le long d'un bras chromosomique de l'homme (région MHC). En noir : cartographie des isochores de composition nucléotidique homogène, la vision initiale, avec en gras deux isochores déterminés expérimentalement. En gris : taux de GC réel (fenêtre glissante de 100 kb, pas de 1 kb).

(Tiré de Oliver et al. 2001)

La structure en isochore des génomes a été particulièrement bien décrite chez les mammifères et les oiseaux (pour revue Eyre-Walker & Hurst 2001 ; Duret & Galtier 2009). Les variations de taux de GC affectent à la fois les régions codantes et non codantes. A petite échelle les taux de GC des différents compartiments sont corrélés : corrélations au sein des régions codantes entre les taux de GC en première, deuxième et troisième position de codon (GC1, GC2, GC3) (Clay et al. 1996), et corrélations locales entre taux de GC des régions codantes et des régions flanquantes non codantes, que ce soient des UTRs, des introns ou d'autres types de régions non codantes (Mouchiroud et al. 1991 ; Costantini & Bernardi 2008a) (Figure 4). Le degré d'hétérogénéité du paysage nucléotidique est plus ou moins marqué selon l'espèce. Par exemple,

chez les mammifères, les musaraignes et les tenrecs ont des paysages nucléotidiques particulièrement riches et hétérogènes en GC alors que chez l'homme et d'autres grands singes l'hétérogénéité est moins prononcée, vraisemblablement à cause de l'érosion de leurs isochores riches (Lander et al. 2001 ; Duret et al. 2002 ; Romiguier et al. 2010).

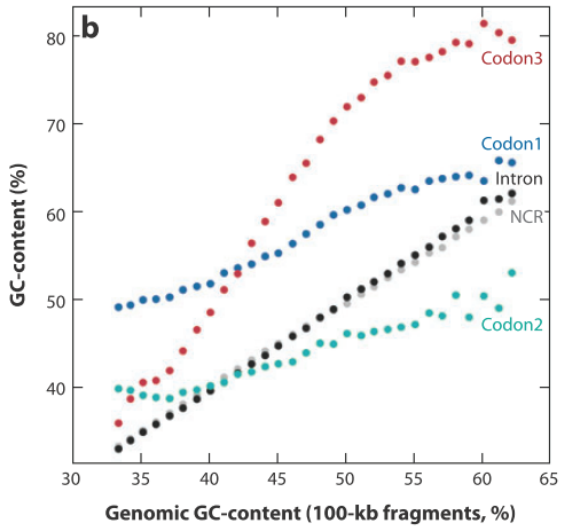


Figure 4 – Corrélations entre taux de GC de différents compartiments génomiques

Corrélations entre taux de GC calculés pour des fragments de 100 kb du génome humain : en abscisse, taux de GC global, en ordonnée, taux de GC pour différents compartiments génomiques – première position de codon (bleu), deuxième position de codon (vert), troisième position de codon (rouge), introns (noir) et autres régions non codantes (gris).

(Tiré de Duret & Galtier 2009)

Les paysages nucléotidiques ont été moins étudiés pour les autres groupes d'eucaryotes. Des variations de la composition nucléotidique ont été mises en évidence chez la levure (Sharp & Lloyd 1993), chez des invertébrés (drosophile : Carulli et al. 1993 ; abeille : Jørgensen et al. 2007, nématodes, échinoderme et chordés : Cammarano et al. 2009) et, chez les vertébrés, chez des poissons (Costantini et al. 2007), des reptiles (Hughes et al. 1999), et un amphibien (xénope : Costantini et al. 2009).

Chez les plantes, des travaux ont été consacrés à l'étude des paysages nucléotidiques des espèces modèles eudicotylédones (principalement *Arabidopsis thaliana*, et aussi des espèces du genre *Brassica*, le pois *Pisum sativum*, la pomme de terre *Solanum tuberosum*, la tomate *Solanum*

lycopersicum, le tabac *Nicotiana tabacum*, le soja *Glycine max* et le tournesol *Helianthus annuus* notamment) et monocotylédones (principalement le riz *Oryza sativa*, et d'autres graminées : l'orge *Hordeum vulgare*, le maïs *Zea mays*, l'avoine *Avena sativa*, le sorgho *Sorghum bicolor* et le blé tendre *Triticum aestivum* notamment) (par exemple Carels & Bernardi 2000 ; Wang et al. 2004 ; Guo et al. 2007 ; Zhang & Zhang 2004 ; Wang & Hickey 2007 ; Tatarinova et al. 2010). Les paysages nucléotidiques de ces plantes se sont révélés hétérogènes, tout particulièrement ceux des graminées. Une structure en isochore similaire à celle des mammifères est suggérée par la corrélation entre GC3 et GC des régions flanquantes chez *Arabidopsis* (Barakat et al. 1998). Le GC3 des graminées étudiées a une distribution bimodale, contrairement à celui des eudicotylédones (Figure 5). Chez le riz et le maïs, le taux de GC décroît fortement le long des gènes de 5' à 3', tandis que ce gradient est beaucoup plus faible chez le tabac et *Arabidopsis* (Figure 6) (Wong et al. 2002). Ces différences marquantes entre les paysages nucléotidiques des graminées et des eudicotylédones étudiées sont souvent généralisées en une dichotomie monocotylédone/eudicotylédone (par exemple Wong et al. 2002 ; Wang et al. 2004), mais cette vision est phylogénétiquement biaisée. Les paysages nucléotidiques des monocotylédones non graminées semblent différents de ceux des graminées : les paysages nucléotidiques des bananes (genre *Musa*) seraient intermédiaires entre ceux du riz et d'*Arabidopsis* (Lescot et al. 2008), et le paysage nucléotidique de l'oignon *Allium cepa* serait semblable à celui d'*Arabidopsis* (Kuhl et al. 2004). Globalement, la distribution phylogénétique des paysages nucléotidiques reste encore peu explorée chez les angiospermes.

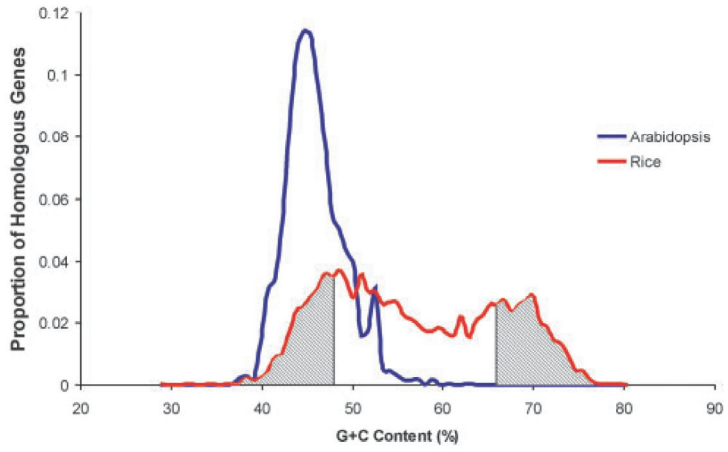


Figure 5 – Distribution du taux de GC génique chez le riz et *Arabidopsis thaliana*

Taux de GC calculés pour 4 447 paires de gènes homologues.

(Tiré de Wang et al. 2004)

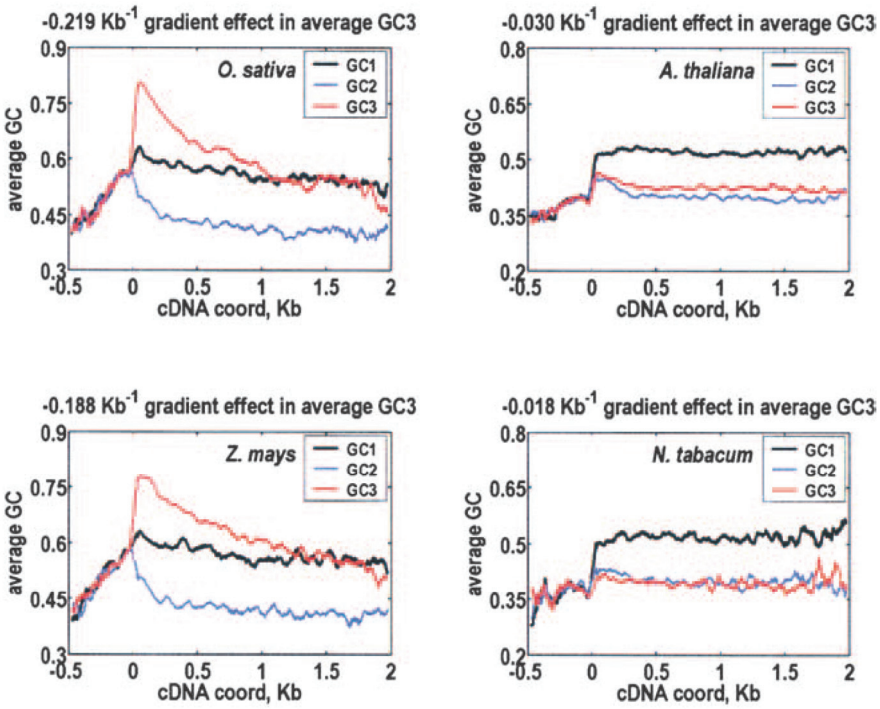


Figure 6 – Gradients de GC1, GC2 et GC3 le long des gènes du riz, du maïs, d'*Arabidopsis thaliana* et du tabac

(Tiré de Wong et al. 2002)

C. Caractéristiques génomiques associées à la composition nucléotidique

Les variations de taux de GC sont souvent associées à d'autres caractéristiques génomiques. Chez les mammifères, les régions riches en GC ont une plus forte densité génique et leurs gènes sont plus courts et plus compacts (Figure 7) (Mouchiroud et al. 1991 ; Duret et al. 1995 ; Duret & Galtier 2009 ; Freudenberg et al. 2009). Elles sont répliquées plus tôt que les régions pauvres en GC (Costantini & Bernardi 2008b). La composition nucléotidique est également corrélée à la distribution des éléments transposables (Smit 1999), aux niveaux d'expression (Konu & Li 2002 ; Arhondakis 2004), aux patrons de recombinaison (Fullerton et al. 2001 ; Duret & Arndt 2008) et de méthylation (Cacciò et al. 1997).

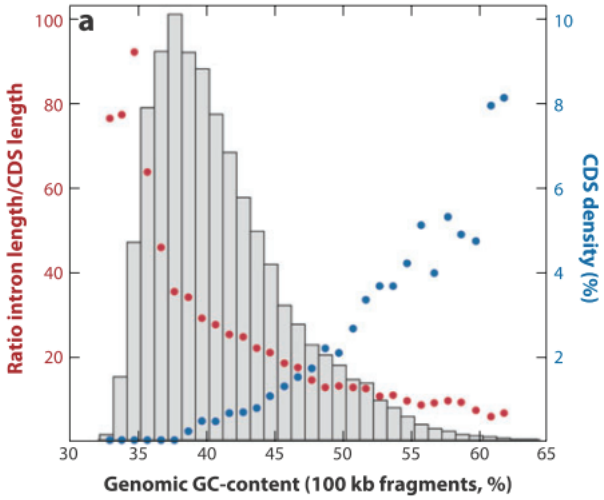


Figure 7 – Corrélation entre GC génomique, compaction des gènes et densité génique

Taux de GC, compaction et densité génique pour des fragments de 100 kb du génome humain. Histogramme : distribution des taux de GC. En rouge : médiane du ratio longueur des introns / longueur des séquences codantes. En bleu : médiane de la densité en séquence codante.

(Tiré de Duret & Galtier 2009)

Pour expliquer de telles hétérogénéités, plusieurs hypothèses sélectionnistes et neutralistes, non mutuellement exclusives, ont été proposées et ont fait l'objet de nombreux débats (pour revue : Eyre-Walker & Hurst 2001 ; Duret & Galtier 2009).

II. Hypothèses sélectionnistes quant à l'émergence de paysages nucléotidiques accidentés

A. Sélection pour une adaptation à l'homéothermie

Selon Giorgio Bernardi, c'est la sélection naturelle pour une meilleure stabilité de l'ADN qui serait à l'origine de la structure en isochore (Bernardi et al. 1985 ; Bernardi & Bernardi 1986 ; Bernardi 2000). Les régions d'ADN riches en GC ayant une meilleure thermostabilité que les régions riches en AT, l'augmentation du GC serait avantageuse et les isochores résulteraient d'une adaptation à l'homéothermie des génomes des mammifères et des oiseaux. Cette hypothèse est compatible avec les forts taux de GC à la fois des régions codantes et des régions non codantes, mais explique difficilement les variations intra génomiques. Plusieurs arguments ont contribué à exclure cette hypothèse. Une structure en isochore marquée a été découverte chez des reptiles, organismes non homéothermes, et l'évolution des isochores serait antérieure à l'évolution de l'homéothermie (Hughes et al. 1999). Chez les procaryotes, l'absence de corrélation entre GC génomique et température optimale de croissance montre qu'un fort taux de GC n'est pas nécessaire à la vie à haute température (Galtier & Lobry 1997 ; Hurst & Merchant 2001).

B. Sélection sur l'usage du code pour l'efficacité de la traduction

En lien avec les hétérogénéités de la composition nucléotidique, les codons synonymes – codons codant pour le même acide aminé (voir Encadré 1) – ne sont pas présents en fréquences égales entre gènes au sein d'un même génome. Une hypothèse sélectionniste spécifique a été avancée pour expliquer ce biais d'usage du code : l'hypothèse de sélection traductionnelle, ou sélection sur l'usage du code (SUC ; Akashi 2001 pour revue). Selon cette hypothèse, parmi les différents codons codant pour un même acide aminé, certains permettent une traduction plus efficace : leur traduction est plus rapide et/ou plus fidèle. Ces codons sont dits codons préférés ou codons optimaux.

Plus un gène est exprimé, plus la pression de sélection pour que sa traduction soit optimale est forte. Il est donc attendu que les gènes les plus exprimés soient plus fortement enrichis en codons préférés que ne le sont les gènes moins exprimés. Il a effectivement été démontré que le biais d'usage du code des gènes est corrélé à leur niveau d'expression (par exemple chez *E. coli* : Post & Nomura 1980 ; Gouy & Gautier 1982). Les gènes codant pour les protéines ribosomales notamment, connues pour leur forte abondance, ont un biais d'usage du code parmi les plus forts

(Ikemura 1985 ; Kanaya et al. 2001). Cette corrélation entre niveau d'expression des gènes et fréquences d'usage des codons synonymes peut être utilisée pour déterminer quels sont les codons préférés d'une espèce (par exemple Tanguy et al. 2008). L'amélioration de l'efficacité de la traduction conférée par la présence de codons préférés a été prouvée expérimentalement. Par exemple pour les codons synonymes codant pour l'acide glutamique chez *Escherichia coli*, le codon préféré GAA est traduit 3,4 fois plus rapidement que GAG (Sørensen & Pedersen 1991). Chez la drosophile, l'introduction de quelques codons synonymes non préférés dans la séquence du gène *Adh* entraîne une réduction significative de la synthèse de la protéine ADH (Carlini & Stephan 2003).

C'est la composition de la population d'ARNt cellulaire qui permettrait d'expliquer pourquoi la traduction des codons préférés est plus efficace que celle des autres codons synonymes. Les codons préférés sont en effet ceux qui s'apparient avec l'espèce d'ARNt la plus fréquente dans la cellule pour l'acide aminé correspondant par rapport aux autres ARNt isoaccepteurs ou qui s'apparient avec la plus grande force avec leur ARNt (voir Encadré 2) (Stoletzki & Eyre-Walker 2007). Cette corrélation entre l'abondance des ARNt et les préférences d'usage du code a été montrée de manière directe chez *Escherichia coli* et *Saccharomyces cerevisiae* (Ikemura 1981 ; Ikemura 1985). L'abondance des ARNt dans la cellule étant bien corrélée au nombre de copies des gènes codants pour ces ARNt (Dong et al. 1996), le nombre de copies de gènes d'ARNt peut être utilisé comme estimateur de la composition de la population d'ARNt cellulaire, et les préférences d'usage du code sont effectivement corrélées au nombre de copies de gènes d'ARNt (Duret 2000).

Des arguments en faveur de la SUC ont été mis en évidence chez de nombreux procaryotes (par exemple Kanaya et al. 1999) et eucaryotes (par exemple Bennetzen & Hall 1981 ; Shields et al. 1988 ; Stenico et al. 1994 ; Musto et al. 2001), y compris chez les plantes (par exemple chez *Arabidopsis thaliana* : Morton & Wright 2007 ; chez le maïs : Liu et al. 2010).

Parmi les espèces pour lesquelles les préférences d'usage du code ont été établies, il apparaît que certaines, comme le xénope (Musto et al. 2001), préfèrent majoritairement des codons finissant en A ou T, alors que d'autres, comme *Caenorhabditis elegans* (Stenico et al. 1994) préfèrent au contraire majoritairement des codons en G ou C. Si les préférences d'usage du code sont ainsi biaisées en faveur de A/T ou de G/C, alors la SUC, en enrichissant les gènes en codons préférés, est susceptible de modifier les paysages nucléotidiques. Une corrélation entre biais de composition nucléotidique et niveaux d'expression est alors attendue. Une telle corrélation a par exemple été mise en évidence chez la souris (Figure 8). Suivant les types de préférences et

l'intensité de la sélection, la SUC peut affecter la composition nucléotidique des régions codantes, les variations de GC entre gènes étant explicables par les patrons d'expression. Si la SUC peut ainsi expliquer les variations des taux de GC en troisième position des codons, et en partie en première position, elle ne peut par contre pas expliquer les variations de taux GC des régions non codantes et les corrélations locales observées.

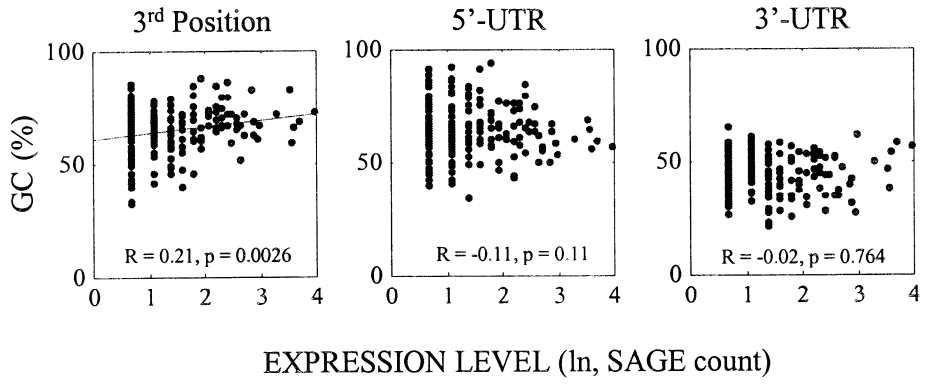


Figure 8 – Corrélation entre GC3 et niveaux d'expression et absence de corrélation entre taux de GC des UTR et niveaux d'expression chez la souris

(Tiré de Konu & Li 2002)

Encadré 1 – Dégénérescence du code génétique

L'information encodée par les séquences de nucléotides des gènes est traduite en séquences d'acides aminés, les protéines, en suivant un certain nombre de règles qui constituent le code génétique. Au cours de la traduction, aux triplets de nucléotides des ARN messagers (ARNm) s'apparient des ARN de transfert (ARNt) spécifiques qui permettent ainsi d'ajouter à la chaîne protéique en formation l'acide aminé correspondant. Avec seulement 20 acides aminés pour 64 (4^3) triplets de nucléotides, le code génétique est dégénéré : hormis la méthionine et le tryptophane pour lesquels le code génétique n'est pas redondant, les autres acides aminés et le codon stop sont codés par 2, 3, 4 ou 6 codons dits synonymes (Tableau 1). La troisième position de codon et la première position dans une moindre mesure sont souvent synonymes, leur changement n'entraînant pas de changement d'acide aminé, alors que la deuxième position de codon est toujours non synonyme.

Tableau 1 – Code génétique standard

Acide aminé		**Codons**	**Acide aminé**		**Codons**
Alanine	Ala	GCU, GCC, GCA, GCG	Lysine	Lys	AAA, AAG
Arginine	Arg	CGU, CGC, CGA, CGG, AGA, AGG	Méthionine, START	Met, START	AUG
Asparagine	Asn	AAU, AAC	Phénylalanine	Phe	UUU, UUC
Acide aspartique	Asp	GAU, GAC	Proline	Pro	CCU, CCC, CCA, CCG
Cystéine	Cys	UGU, UGC	Sérine	Ser	UCU, UCC, UCA, UCG, AGU, AGC
Glutamine	Gln	CAA, CAG	Thréonine	Thr	ACU, ACC, ACA, ACG
Acide glutamique	Glu	GAA, GAG	Tryptophane	Trp	UGG
Glycine	Gly	GGU, GGC, GGA, GGG	Tyrosine	Tyr	UAU, UAC
Histidine	His	CAU, CAC	Valine	Val	GUU, GUC, GUA, GUG
Isoleucine	Ile	AUU, AUC, AUA	STOP	STOP	UAA, UGA, UAG
Leucine	Leu	UUA, UUG, CUU, CUC, CUA, CUG			

Encadré 2 – Les ARN de transfert

Les ARNt, qui jouent le rôle d'adaptateurs entre les codons des ARNm et les acides aminés à assembler lors de la traduction, ont une structure secondaire avec 4 tiges appariées, en forme dite de feuille de trèfle (Figure 9 a). Côté « tige » se trouve le bras accepteur où s'attachera l'acide aminé correspondant à l'ARNt. Sur la boucle opposée se trouve le bras anticodon, avec l'anticodon qui s'appariera avec le codon de l'ARNm. L'anticodon correspond aux positions 34, 35 et 36 de la séquence des ARNt. Les ARNt présentent la particularité de posséder des nucléotides modifiés, c'est-à-dire autres que l'adénine, la cytosine, la guanine et l'uracile. Ces modifications de nucléotides sont dues à l'action d'enzymes spécifiques lors de la maturation post-transcriptionnelle des ARNt (Figure 9 b). Par exemple l'inosine (I) est obtenue par la déamination d'une adénosine. Les nucléotides modifiés sont notamment fréquents en position 34 de l'anticodon et en position 37 adjacente en 3' (Figure 10). Ils sont capables de réaliser des appariements bancals (*wobble pairing*), des appariements autres que les appariements canoniques de Watson-Crick A•U et G•C (*position 1 de l'anticodon • position 3 du codon*). Les principaux appariements bancals des interactions anticodon-codon sont I•U, I•A, I•C et G•U (Agris et al. 2007). Cette propriété d'appariement bancal permet à une même espèce d'ARNt de pouvoir s'apparier avec jusqu'à 3 codons différents. L'existence de cet effet *wobble* avait été prédite par Crick en 1966 (Crick 1966) puis démontrée expérimentalement (par exemple Söll et al. 1966). Même si une espèce d'ARNt peut s'apparier à plusieurs codons synonymes, elle a toutefois une préférence pour l'un de ces codons, déterminée par la stabilité de l'interaction codon-anticodon. La préférence est plus forte pour le codon avec lequel un appariement canonique est possible (Stoletzki & Eyre-Walker 2007). Pour un acide aminé donné, il n'y a donc pas autant d'espèces d'ARNt isoaccepteurs (ceux qui portent cet acide aminé) que de codons synonymes. Le répertoire d'ARNt d'une espèce comprend ainsi un nombre d'espèces d'ARNt intermédiaire entre 21 – nombre d'acides aminés – et 64 – nombre de codons possibles. Chez les plantes, il y a de l'ordre de 45 espèces d'ARNt (Michaud et al. 2011). Certains codons n'ont pas d'ARNt capables de les décoder : c'est le cas de TGA, TAA et TAG dans le code standard, qui stoppent la synthèse de la chaîne protéique.

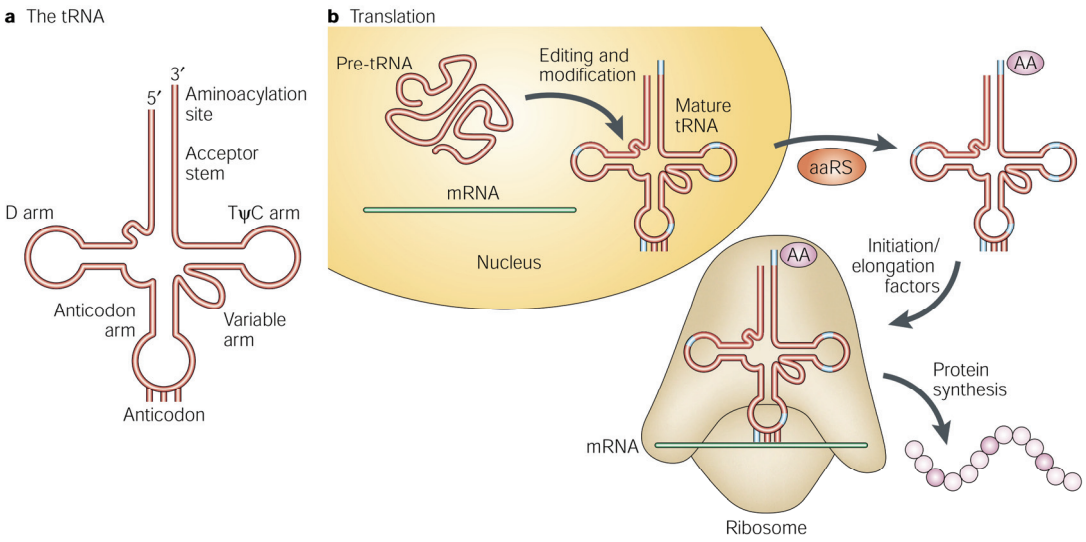


Figure 9 – Schémas d'un ARNt et des étapes et acteurs de la traduction

a | Structure secondaire d'un ARNt, montrant ses principales caractéristiques. L'acide aminé s'attachera sur la tige accepteuse. L'anticodon se trouve sur la boucle opposée. **b** | Acteurs et étapes de la traduction. ARNt et ARNm sont transcrits d'après les gènes dans le noyau de la cellule. ARNt et ARNm peuvent tous deux être édités avant la traduction, c'est-à-dire subir des modifications post-transcriptionnelles ; en particulier, des enzymes spécifiques modifient certains nucléotides de l'ARNt (représentés en bleu), ce qui peut changer sa capacité de décodage. Dans le cytoplasme, les aminoacyl-ARNt synthétases (aaRS) chargent les ARNt avec leurs acides aminés spécifiques (AA). Dans le ribosome, les trois bases de l'anticodon de l'ARNt s'apparient avec les trois bases complémentaires du codon de l'ARNm. L'acide aminé est alors ajouté à la chaîne polypeptidique en cours d'élongation.

(Tiré de Knight et al., 2001)

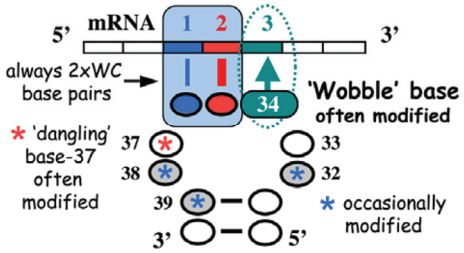


Figure 10 – Appariements de type bancal et de type Watson-Crick lors de l'interaction codon-anticodon

L'anticodon est constitué par les nucléotides en position 34, 35 et 36 de l'ARNt. Les nucléotides 34 et 37 sont souvent modifiés, et les nucléotides 32, 38 et 39 le sont occasionnellement. Les nucléotides en position 3 et 2 de l'anticodon s'apparient respectivement avec les nucléotides en position 1 (bleu) et 2 (rouge) du codon par un appariement de type Watson-Crick. Le nucléotide en position 1 de l'anticodon s'apparie avec le troisième nucléotide du codon (vert) soit par un appariement de type Watson-Crick, soit par un appariement bancal.

(Tiré de Grosjean, de Crécy-Lagard, & Marck, 2010)

III. Hypothèses neutralistes quant à l'émergence de paysages nucléotidiques accidentés

La SUC, force de sélection agissant sur les régions codantes, ne permet pas d'expliquer intégralement les variations de composition nucléotidique des génomes. Il existe de nombreux arguments plaidant en faveur de mécanismes évolutifs neutres. Par exemple, chez les primates, deux pseudogènes de globine ont vu leur GC évoluer après leur inactivation : le GC du pseudogène situé dans une région pauvre en GC a diminué, alors que le GC du pseudogène situé dans une région riche en GC a augmenté (Francino & Ochman 1999).

Deux types de mécanismes neutres ont été proposés pour expliquer l'émergence des paysages nucléotidiques accidentés : biais mutationnel et conversion génique biaisée.

A. Biais mutationnel

Différentes hypothèses de biais mutationnel ont été proposées. Selon ces hypothèses, des biais mutationnels variables le long des génomes modifieraient de manière hétérogène la composition nucléotidique, affectant à la fois le GC des régions codantes et non codantes. Les hypothèses proposées diffèrent sur les mécanismes moléculaires supposés être à l'origine de ces variations de biais mutationnel. Il a été proposé que les variations de concentrations des nucléotides libres au cours du cycle de réplication de l'ADN entraînent des variations de biais mutationnel selon le *timing* de réplication des régions : les régions se répliquant en premier, lorsque la population de nucléotides libres est riche en G et C, subiraient plus de mutations vers G et C (Wolfe et al. 1989 ; Eyre-Walker & Hurst 2001). Il a également été suggéré que le GC local pouvait avoir un effet sur le biais mutationnel. Le taux de déamination des cytosines – la principale cause de mutation de C vers T et de G vers A – diminuerait de moitié à chaque augmentation de 10% du GC. Ce mécanisme fonctionnerait ainsi en une boucle de rétroaction positive, les régions d'ADN s'enrichissant ou s'appauvrissant en GC en fonction de leur GC local (Fryxell & Zuckerkandl 2000). D'autres hypothèses attribuent les variations de biais mutationnel à des variations de l'efficacité de la machinerie cellulaire de réparation de l'ADN (Filipski 1987), aux variations d'état de la chromatine (Sueoka 1988) ou encore à un effet mutagène de la recombinaison (Lercher & Hurst 2002 ; Hellmann et al. 2003).

Ces hypothèses ont cependant été fortement remises en cause par des études basées sur des données de polymorphisme et des modèles de génétiques des populations, qui ont révélé

l'existence d'un biais mutationnel vers A/T et la présence d'un biais de fixation en faveur de G/C (par exemple Eyre-Walker 1999 ; Smith & Eyre-Walker 2001 ; Duret et al. 2002 ; Capra & Pollard 2011). Un biais de fixation peut être interprété comme une signature typique d'un processus de sélection ou comme le fruit d'un processus mimant la sélection.

B. Conversion génique biaisée vers G et C

La conversion génique est un processus associé à la recombinaison, qui résulte en un transfert non réciproque d'information génétique : c'est le « copier/coller » d'un allèle sur l'autre lors d'une recombinaison à un site hétérozygote (pour revue, Marais 2003).

Lors de la méiose, la recombinaison est indispensable à la ségrégation chromosomique et est responsable de la création de nouvelles combinaisons alléliques, lorsqu'elle donne lieu à un crossing-over. Permettant de réparer l'ADN endommagé, c'est également une voie de maintenance de l'ADN. La recombinaison est initiée par une cassure double-brin sur un chromosome. Les extrémités de la séquence ayant subi la coupure sont dégradées en 5'. Les extrémités 3' envahissent alors une séquence homologue – qui peut être allélique ou ectopique. Des hétéroduplexes, c'est-à-dire de l'ADN double-brin dans lequel les deux brins sont d'origines différentes, se forment alors. L'ADN manquant est ensuite synthétisé en prenant comme patron la séquence homologue, puis les brins se séparent, donnant lieu ou pas à un crossing-over. Si les hétéroduplexes formés concernent des sites hétérozygotes, il y a des mésappariements qui doivent être corrigés par la machinerie de réparation d'ADN. La conversion génique est ainsi la conséquence de la réparation des mésappariements des hétéroduplexes lors d'une recombinaison (Figure 11).

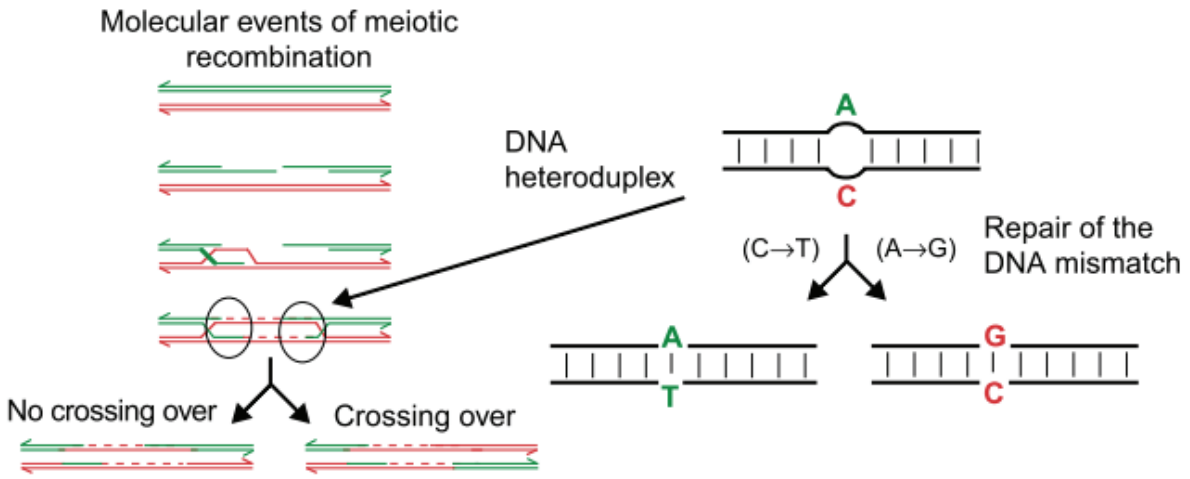


Figure 11 – Mécanisme de conversion génique lors d'une recombinaison méiotique à un site hétérozygote

Evènement de recombinaison entre deux bras chromosomiques (vert et rouge). La correction des mésappariements des hétéroduplexes aboutit à la conversion, c'est-à-dire la perte, d'un des deux allèles initiaux.

(Tiré de Duret 2002)

Chez certaines espèces, ce processus de réparation des mésappariements serait biaisé en faveur de G et C (Brown & Jiricny 1989 ; Holmquist 1992 ; Eyre-Walker 1993 ; Galtier et al. 2001 ; Marais 2003). Chez la levure, ce biais associé à la recombinaison a été mis en évidence expérimentalement (Mancera et al. 2008). Ce biais de réparation en faveur de G/C aurait pu évoluer pour compenser le biais mutationnel vers A/T (Birdsell 2002). La conversion génique biaisée vers G et C (BGC) conduirait alors, lors d'une méiose chez un individu hétérozygote portant un allèle en A ou T et un allèle en G ou C, à la production en excès des gamètes porteurs de l'allèle en G ou C au détriment des gamètes porteurs de l'allèle en A ou T. Conférant ainsi un avantage aux allèles en G ou C, la BGC mime la sélection (Nagylaki 1983) et peut même conduire à la fixation d'allèles en G ou C délétères (hypothèse du talon d'Achille génomique, voir Galtier & Duret 2007 ; Galtier et al. 2009). Cependant, on peut considérer la BGC comme un processus neutre puisque le choix de l'allèle à favoriser n'est pas fonction de son effet sur la valeur sélective.

Sous l'hypothèse de BGC, les séquences situées dans les régions les plus recombinantes du génome ont une probabilité plus forte de subir ce mécanisme et donc d'être enrichies en GC. Les variations du taux de recombinaison le long du génome, couplées à la BGC, seraient ainsi à l'origine des paysages nucléotidiques accidentés. L'hypothèse de BGC permet d'expliquer

pourquoi le génome entier est affecté par des variations de la composition nucléotidique – contrairement à l'hypothèse de SUC – et est compatible avec les signatures de sélection – contrairement aux hypothèses de biais mutationnels variables. Depuis une quinzaine d'années, l'hypothèse BGC est de plus en plus reconnue comme étant la force évolutive majeure à l'origine de l'hétérogénéité des paysages nucléotidiques des mammifères et d'autres groupes taxonomiques (pour revue Duret & Galtier 2009), notamment grâce à de nombreuses études ayant mis en évidence les corrélations attendues entre recombinaison et patrons de GC (Figure 12) ou de substitution nucléotidique (Meunier & Duret 2004 ; Duret & Arndt 2008).

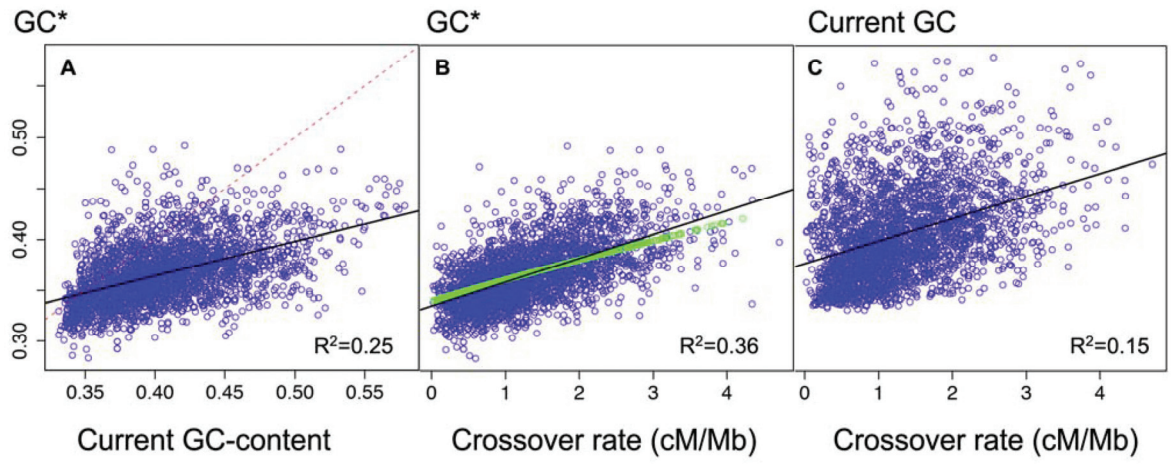


Figure 12 – Corrélation entre le GC d'équilibre (GC*), le GC actuel et le taux de crossing-over chez les autosomes de l'homme

Le taux de GC d'équilibre (GC*) correspond au taux de GC qu'atteindrait une séquence si la dynamique évolutive en cours se poursuivait suffisamment longtemps.

(Tiré de Duret & Arndt 2008)

IV. Evolution des paysages nucléotidiques des plantes

A. Processus évolutifs façonnant les paysages nucléotidiques des plantes : quelques arguments en faveur des hypothèses SUC et BGC

Les mécanismes évolutifs à l'origine des paysages nucléotidiques des plantes sont encore mal connus, mais quelques travaux apportent des arguments en faveur des hypothèses de la SUC et de la BGC, principalement chez les graminées.

Des codons préférés ont été détectés chez plusieurs espèces de plantes (par exemple Kawabe & Miyashita 2003). Ces codons préférés sont en majorité des codons finissant en G ou C, tout particulièrement chez le riz, le blé, l'orge, le sorgho et le maïs (Wang & Roossinck 2006). La SUC pourrait ainsi expliquer une partie des fortes variations de GC3 de ces graminées.

D'autres résultats semblent confirmer la présence du phénomène de BGC chez les plantes (Wright et al. 2007). Les graminées paraissent être affectées plus fortement par la BGC que d'autres angiospermes (Glémin et al. 2006 ; Haudry et al. 2008 ; Escobar et al. 2010 ; Escobar et al. 2011). Une étude que nous avons menée chez des espèces du genre *Oryza* a montré que leur composition nucléotidique est vraisemblablement affectée à la fois par la SUC et la BGC mais pas par des variations de biais mutationnel (voir article complet en annexe). En effet, nous avons montré que (i) le taux de GC d'équilibre (GC*) était positivement corrélé à la recombinaison et au taux d'allogamie aussi bien pour les positions 3 des codons que pour les introns (Figure 13), (ii) le GC3* était également positivement corrélé au niveau d'expression des gènes (Figure 14), et que (iii) les patrons de polymorphisme montraient un biais de fixation en faveur des allèles G et C (Figure 15).

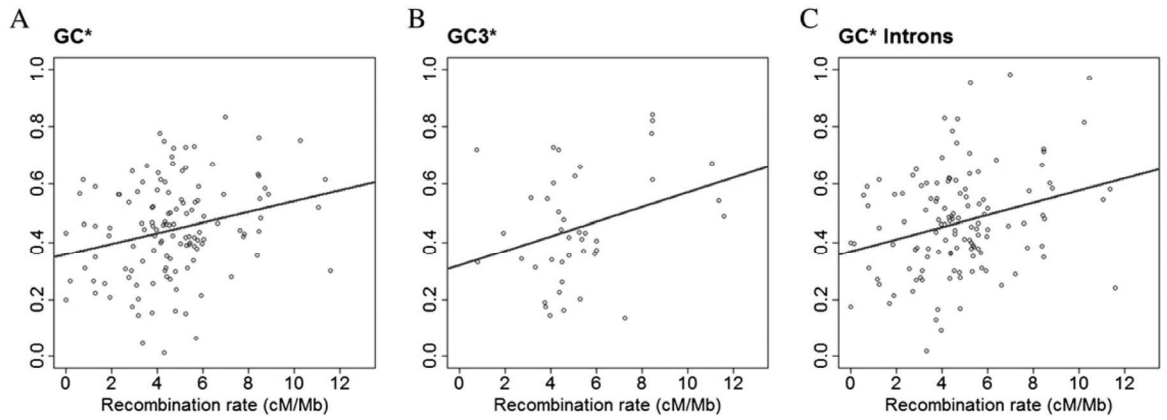


Figure 13 – Corrélations entre GC*, GC3* et GC* introns et le taux de recombinaison chez le riz cultivé *Oryza sativa*

(Tiré de Muyle et al. 2011)

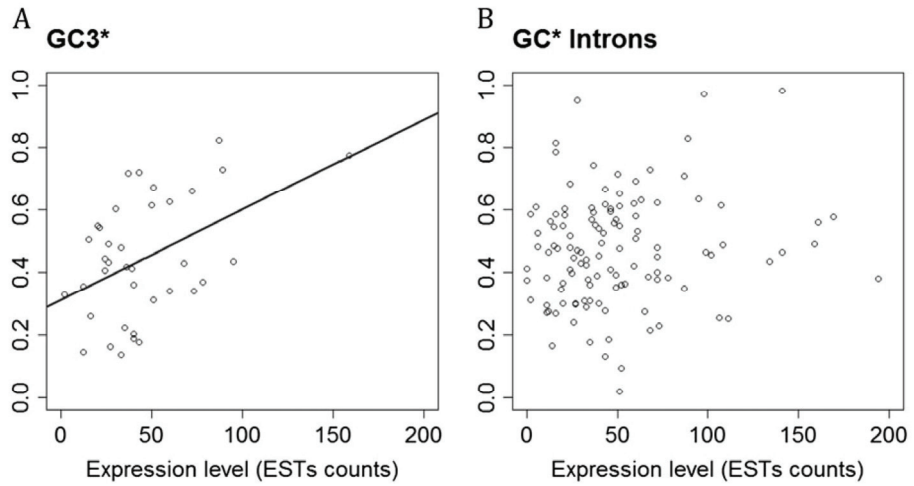


Figure 14 – Corrélation entre GC3* et niveaux d'expression et absence de corrélation entre GC* introns et niveaux d'expression chez le riz cultivé *Oryza sativa*

(Tiré de Muyle et al. 2011)

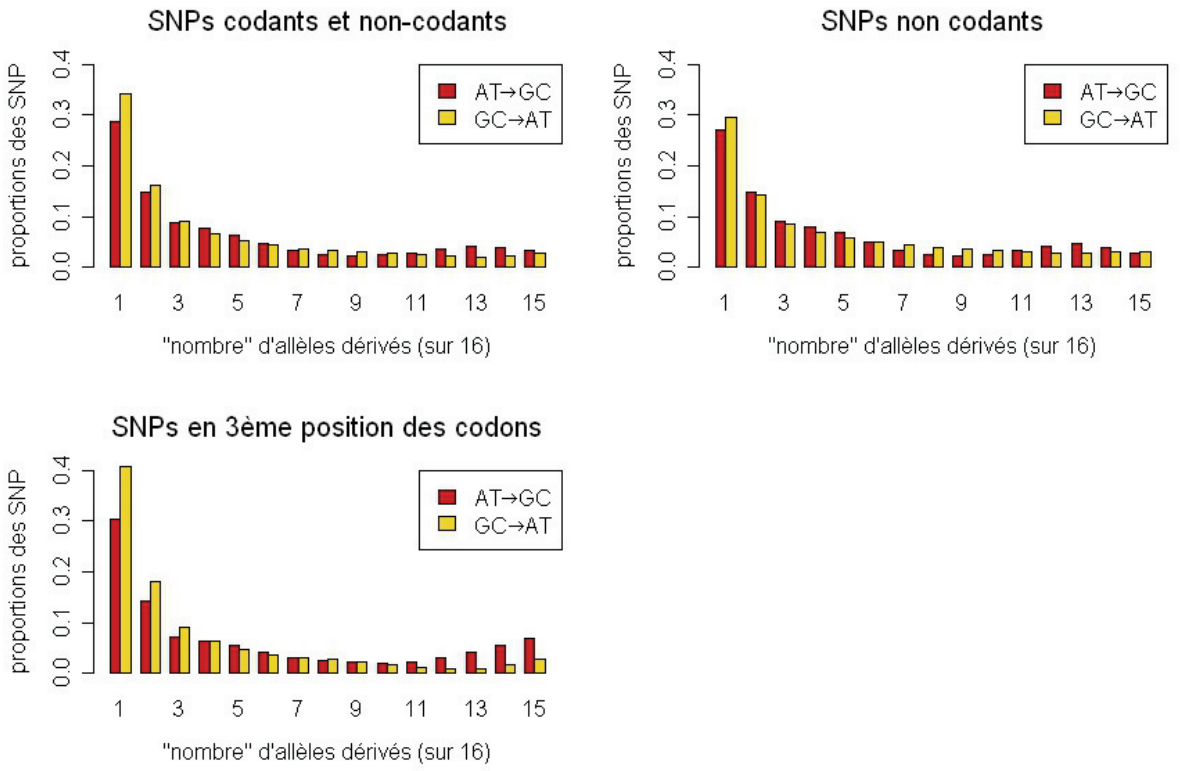


Figure 15 – Spectres des fréquences alléliques dérivées pour le riz sauvage *Oryza rufipogon*

Spectre des fréquences alléliques dérivées du riz sauvage asiatique *Oryza rufipogon* (orientation des SNPs avec le riz sauvage africain *Oryza barthii*). En rouge : SNPs de A ou T vers G ou C. En jaune : SNPs de G ou C vers A ou T.

(Tiré du travail de Muyle et al. 2011)

B. Diversité et évolution des paysages nucléotidiques des plantes : objectifs de la thèse

Chez les plantes, il y a actuellement un déficit de connaissances quant à la composition nucléotidique de nombreux groupes taxonomiques. Les paysages nucléotidiques des monocotylédones non graminées et des angiospermes basales par exemple sont mal connus. Les forces évolutives ayant façonné ces paysages font toujours l'objet de débats.

Les objectifs de ce travail de thèse sont de caractériser les paysages nucléotidiques des angiospermes à une large échelle phylogénétique et de mieux comprendre quels sont les mécanismes évolutifs jouant sur l'évolution de ces paysages nucléotidiques. Comment varient les paysages nucléotidiques le long de la phylogénie des angiospermes ? La dichotomie dicotylédone / monocotylédone est-elle confirmée ? SUC et BGC façonnent-elles ces paysages nucléotidiques ? Les différents taxons sont-ils affectés avec la même intensité ? Pour répondre à ces questions, j'ai utilisé une approche de génomique comparative basée sur l'analyse de données EST (E*xpressed Sequence Tag*) correspondant au séquençage d'ARN messagers. Cette approche m'a permis d'étudier la composition en base de la fraction codante du génome chez plus de 200 espèces d'angiospermes et de gymnospermes.

La première partie est consacrée à l'exploration des paysages nucléotidiques d'un large éventail de plantes couvrant la phylogénie des angiospermes et à la recherche d'arguments permettant de confirmer ou d'invalider la présence de SUC et de BGC à l'œuvre chez les plantes. Ce travail a fait l'objet d'un article publié dans *The Plant Cell* qui est présenté dans cette première partie.

La deuxième partie est plus particulièrement consacrée à l'évolution des préférences d'usage des codons chez les plantes en relation avec les taux de GC des génomes étudiés dans la partie précédente. Cette partie est rédigée en français en suivant la structure d'un article scientifique.

Agris, P.F., Vendeix, F.A.P. & Graham, W.D., 2007. tRNA's wobble decoding of the genome: 40 years of modification. *Journal of molecular biology*, 366(1), pp.1-13.

Akashi, H., 2001. Gene expression and molecular evolution. *Current Opinion in Genetics & Development*, 11, pp.660-666.

Arhondakis, S., 2004. Base composition and expression level of human genes. *Gene*, 325, pp.165-169.

Barakat, A., Matassi, G. & Bernardi, G., 1998. Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. *Proceedings of the National Academy of Sciences of the United States of America*, 95(17), pp.10044-9.

Bennetzen, J.L. & Hall, B.D., 1981. Codon Selection in Yeast. *The Journal of biological chemistry*, 257(6), pp.3026-3031.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1), pp.3-17.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., et al., 1985. The mosaic genome of warm-blooded vertebrates. *Science*, *228*(4702), pp.953-8.

Bernardi, G. & Bernardi, G., 1986. Compositional constraints and genome evolution. *Journal of Molecular Evolution*, 24(1-2), pp.1-11.

Birdsell, J.A., 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular biology and evolution*, 19(7), pp.1181-97.

Brown, T.C. & Jiricny, J., 1989. Repair of base-base mismatches in simian and human cells. *Genome*, 31(2), pp.578-83.

Cacciò, S., Jabbari, K., Matassi, G., Guermonprez, F., Desgrès, J., & Bernardi, G., 1997. Methylation patterns in the isochores of vertebrate genomes. *Gene*, 205(1-2), pp.119-24.

Cammarano, R., Costantini, M. & Bernardi, G., 2009. The isochore patterns of invertebrate genomes. *BMC genomics*, 10(538).

Capra, J.A. & Pollard, K.S., 2011. GBE Substitution patterns are GC-biased in divergent sequences across the metazoans. *Genome*, pp.1-16.

Carels, N. & Bernardi, G., 2000. Two classes of genes in plants. *Genetics*, 154(4), pp.1819-25.

Carlini, D.B. & Stephan, W., 2003. *In vivo* introduction of unpreferred synonymous codons into the Drosophila *Adh* gene results in reduced levels of ADH protein. *Genetics*, 163(1), pp.239-43.

Carulli, J. P., Krane, D. E., Hartl, D. L., & Ochman, H., 1993. Compositional Heterogeneity and Patterns of Molecular Evolution in the Drosophila Genome. *Genetics*, 134, pp.837-845.

Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D., & Bernardi, G., 1996. Human coding and noncoding DNA: compositional correlations. *Molecular phylogenetics and evolution*, 5(1), pp.2-12.

Costantini, M., Auletta, F. & Bernardi, G., 2007. Isochore patterns and gene distributions in fish genomes. *Genomics*, 90(3), pp.364-71.

Costantini, M. & Bernardi, G., 2008a. Correlations between coding and contiguous non-coding sequences in isochore families from vertebrate genomes. *Gene*, 410(2), pp.241-8.

Costantini, M. & Bernardi, G., 2008b. Replication timing, chromosomal bands, and isochores. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), pp.3433-7.

Costantini, M., Cammarano, R. & Bernardi, G., 2009. The evolution of isochore patterns in vertebrate genomes. *BMC genomics*, 10(146).

Crick, F.H., 1966. Codon–anticodon pairing: the wobble hypothesis. *Journal of molecular biology*, 19(2), pp.548-55.

Cuny, G., Soriano, P., Macaya, G., & Bernardi, G., 1981. The Major Components of the Mouse and Human Genomes 1. Preparation, Basic Properties and Compositional Heterogeneity. *Eur. J. Biochem.*, (115), pp.227-233.

Dong, H., Nilsson, L. & Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *Journal of molecular biology*, 260(5), pp.649-63.

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development*, 12(6), pp.640-9.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., & Galtier, N., 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics*, 162(4), pp.1837-47.

Duret, L., 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends in genetics: TIG*, 16(7), pp.287-9.

Duret, L. & Arndt, P.F., 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS genetics*, 4(5), p.e1000071.

Duret, L. & Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10, pp.285-311.

Duret, L., Mouchiroud, D. & Gautier, C., 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *Journal of Molecular Evolution*, 40(3), pp.308-317.

Escobar, J.S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J., & Glémin, S., 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). *Evolution*, pp.1-18.

Escobar, J.S., Glémin, S. & Galtier, N., 2011. GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms, and Other Eukaryotes. *Molecular biology and evolution*, 28(9), pp.2561-2575.

Eyre-Walker, A., 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*, 152(2), pp.675-83.

Eyre-Walker, A., 1993. Recombination and mammalian genome evolution. *Proc. R. Soc. Lond.*, 252(1335), pp.237-43.

Eyre-Walker, A. & Hurst, L.D., 2001. The evolution of isochores. *Nature reviews. Genetics*, 2(4), pp.549-555.

Filipski, J., 1987. Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS letters*, 217(2), pp.184-6.

Francino, M.P. & Ochman, H., 1999. Isochores result from mutation not selection. *Nature*, 400(July), pp.30-31.

Freudenberg, J., Wang, M., Yang, Y., & Li, W., 2009. Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. *BMC bioinformatics*, 10 (Suppl 1), S66.

Fryxell, K.J. & Zuckerkandl, E., 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Molecular biology and evolution*, 17(9), pp.1371-83.

Fullerton, S.M., Carvalho, A.B. & Clark, A.G., 2001. Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. *Molecular biology and evolution*, 18(6), pp.1139-1142.

Galtier, N., Piganeau, G., Mouchiroud, D., & Duret, L., 2001. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics*, 159, pp.907-911.

Galtier, N., Duret, L., Glémin, S., & Ranwez, V., 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in genetics: TIG*, 25(1), pp.1-5.

Galtier, N. & Duret, L., 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in genetics: TIG*, 23(6), pp.273-7.

Galtier, N. & Lobry, J.R., 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of molecular evolution*, 44(6), pp.632-6.

Glémin, S., Bazin, E. & Charlesworth, D., 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc. R. Soc. B.,* 273(1604), pp.3011-9.

Gouy, M. & Gautier, C, 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10(22), pp.7055-7074.

Grosjean, H., de Crécy-Lagard, V. & Marck, C., 2010. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS letters*, 584(2), pp.252-64.

Guo, X., Bao, J. & Fan, L., 2007. Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes. *FEBS letters*, 581(5), pp.1015-21.

Haudry, A., Cenci, A., Guilhaumon, C., Paux, E., Poirier, S., Santoni, S., David, J., et al., 2008. Mating system and recombination affect molecular evolution in four Triticeae species. *Genetics research*, 90(1), pp.97-109.

Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S., & Przeworski, M., 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *American journal of human genetics*, 72(6), pp.1527-35.

Holmquist, G.P., 1992. Chromosome Bands, Their Chromatin Flavors, and Their Functional Features. *American journal of human genetics*, 51, pp.17-37.

Hughes, S., Zelus, D. & Mouchiroud, D., 1999. Warm-blooded isochore structure in Nile crocodile and turtle. *Molecular biology and evolution*, 16(11), pp.1521-7.

Hurst, L.D. & Merchant, A.R., 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings. Biological sciences / The Royal Society*, 268(1466), pp.493-7.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), pp.13-34.

Ikemura, T., 1981. Correlation between the abundance of transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the translational system. *Journal of Molecular Biology*, 151(3), pp.389-409.

Jørgensen, F.G., Schierup, M.H. & Clark, A.G., 2007. Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of *Apis mellifera*. *Molecular biology and evolution*, 24(2), pp.611-9.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., & Ikemura, T., 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of molecular evolution*, 53(4-5), pp.290-8.

Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1), pp.143-55.

Kawabe, A. & Miyashita, N.T., 2003. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes & genetic systems*, 78(5), pp.343-52.

Knight, R.D., Freeland, S.J. & Landweber, L.F., 2001. Rewiring the keyboard: evolvability of the genetic code. *Nature reviews. Genetics*, 2(1), pp.49-58.

Konu, O. & Li, M.D., 2002. Correlations between mRNA expression levels and GC contents of coding and untranslated regions of genes in rodents. *Journal of molecular evolution*, 54(1), pp.35-41.

Kuhl, J.C., Cheung, F., Yuan, Q., Martin, W., Zewdie, Y., Mccallum, J., Catanach, A., et al., 2004. A Unique Set of 11,008 Onion Expressed Sequence Tags Reveals Expressed Sequence and Genomic Differences between the Monocot Orders Asparagales and Poales. *The Plant Cell*, *16*, pp.114-125.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860-921.

Lercher, M.J. & Hurst, L.D., 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in genetics: TIG*, 18(7), pp.337-40.

Lescot, M., Piffanelli, P., Ciampi, A.Y., Ruiz, M., Blanc, G., Leebens-Mack, J., da Silva, F.R., et al., 2008. Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. *BMC genomics*, 9(58).

Liu, H., He, R., Zhang, H., Huang, Y., Tian, M., & Zhang, J., 2010. Analysis of synonymous codon usage in *Zea mays*. *Molecular biology reports*, 37(2), pp.677-84.

Lynch, M., 2007. *The Origins of Genome Architecture*, Sunderland, USA: Sinauer Associates Inc.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., & Steinmetz, L.M., 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203), pp.479-85.

Marais, G., 2003. Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics*, 19(6), pp.330-338.

Meunier, J. & Duret, L., 2004. Recombination drives the evolution of GC-content in the human genome. *Molecular biology and evolution*, 21(6), pp.984-90.

Michaud, M., Cognat, V., Duchêne, A.-M., & Maréchal-Drouard, L., 2011. A global picture of tRNA genes in plant genomes. *The Plant Journal*, 66(1), pp.80-93.

Morton, B.R. & Wright, S.I., 2007. Selective constraints on codon usage of nuclear genes from *Arabidopsis thaliana*. *Molecular biology and evolution*, 24(1), pp.122-9.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., & Bernardi, G., 1991. The distribution of genes in the human genome. *Gene*, 100(3), pp.181-7.

Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H., & Bernardi, G., 2001. Translational selection on codon usage in *Xenopus laevis*. *Molecular biology and evolution*, 18(9), pp.1703-7.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., & Glémin, S., 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular biology and evolution*.

Nagylaki, T., 1983. Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 80(20), pp.6278-81.

Niimura, Y. & Gojobori, T., 2002. *In silico* chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2), pp.797-802.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., & Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. *Gene*, 276(1-2), pp.47-56.

Post, L.E. & Nomura, M., 1980. DNA Sequences from the str Operon of *Escherichia coli*. *The Journal of biological chemistry*, 255(10), pp.4660-4666.

Romiguier, J., Ranwez, V., Douzery, E. J. P., & Galtier, N., 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome research*, pp.1001-1009.

Salinas, J., Matassi, G., Montero, L.M., & Bernardi, G., 1988. Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. *Nucleic acids research*, 16(10), pp.4269-4285.

Sharp, P.M. & Lloyd, A.T., 1993. Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. *Nucleic acids research*, 21(2), pp.179-183.

Shields, D.C., Sharp, P.M., Higgins, D.G., & Wright, F., 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular biology and evolution*, 5(6), pp.704-16.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development*, 9, pp.657-663.

Smith, N.G. & Eyre-Walker, a, 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. *Molecular biology and evolution*, 18(6), pp.982-6.

Stenico, M., Lloyd, A.T. & Sharp, P.M., 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic acids research*, 22(13), pp.2437-46.

Stoletzki, N. & Eyre-Walker, A., 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Molecular biology and evolution*, 24(2), pp.374-81.

Söll, D., Jones, D.S., Ohtsuka, E., Faulkner, R.D., Lohrmann, R., Hayatsu, H., Khorana, H.G., et al., 1966. Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique. *Journal of Molecular Biology*, 19(2), pp.556-573.

Sørensen, M.A. & Pedersen, S., 1991. Absolute *in vivo* translation rates of individual codons in *Escherichia coli*: The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *Journal of molecular biology*, 222(2), pp.265-280.

Tanguy, A., Bierne, N., Saavedra, C., Pina, B., Bachère, E., Kube, M., Bazin, E., et al., 2008. Increasing genomic information in bivalves through new EST collections in four species: development of new genetic markers for environmental studies and genome evolution. *Gene*, 408(1-2), pp.27-36.

Tatarinova, T.V., Alexandrov, N.N., Bouck, J.B., & Feldmann, K.A., 2010. GC3 biology in corn, rice, sorghum and other grasses. *BMC genomics*, 11(308).

Wang, H.-C. & Hickey, D.A., 2007. Rapid divergence of codon usage patterns within the rice genome. *BMC evolutionary biology*, 7 Suppl 1, S6.

Wang, H.-C., Singer, G.A.C. & Hickey, D.A., 2004. Mutational bias affects protein evolution in flowering plants. *Molecular biology and evolution*, 21(1), pp.90-6.

Wang, L. & Roossinck, M.J., 2006. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. *Plant molecular biology*, 61(4-5), pp.699-710.

Wolfe, K.H., Sharp, P.M. & Li, W.H., 1989. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204), pp.283-5.

Wong, G.K.-S., Wang, J., Tao, L., Tan, J., Zhang, J., Passey, D.A., & Yu, J., 2002. Compositional gradients in Gramineae genes. *Genome research*, 12(6), pp.851-6.

Wright, S.I., Iorgovan, G., Misra, S., & Mokhtari, M., 2007. Neutral evolution of synonymous base composition in the Brassicaceae. *Journal of molecular evolution*, 64(1), pp.136-41.

Zhang, R. & Zhang, C.-T., 2004. Isochore structures in the genome of the plant *Arabidopsis thaliana*. *Journal of molecular evolution*, 59(2), pp.227-38.

Chapitre 1

Patrons et évolution des paysages nucléotidiques des plantes

This article is a *Plant Cell* Advance Online Publication. The date of its first appearance online is the official date of publication. The article has been edited and the authors have corrected proofs, but minor changes could be made before the final version is published. Posting this version online reduces the time to publication by several weeks.

Patterns and Evolution of Nucleotide Landscapes in Seed Plants [W]

Laurana Serres-Giardi,[a,b] Khalid Belkhir,[a] Jacques David,[b] and Sylvain Glémin[a,1]

[a] Institut des Sciences de l'Evolution de Montpellier, Unité Mixte de Recherche 5554, Centre National de la Recherche Scientifique, Université Montpellier 2, F-34095 Montpellier, France

[b] Montpellier SupAgro, Unité Mixte de Recherche 1334, Amélioration Génétique et Adaptation des Plantes Méditerranéennes et Tropicales, F-34398 Montpellier, France

Nucleotide landscapes, which are the way base composition is distributed along a genome, strongly vary among species. The underlying causes of these variations have been much debated. Though mutational bias and selection were initially invoked, GC-biased gene conversion (gBGC), a recombination-associated process favoring the G and C over A and T bases, is increasingly recognized as a major factor. As opposed to vertebrates, evolution of GC content is less well known in plants. Most studies have focused on the GC-poor and homogeneous *Arabidopsis thaliana* genome and the much more GC-rich and heterogeneous rice (*Oryza sativa*) genome and have often been generalized as a dicot/monocot dichotomy. This vision is clearly phylogenetically biased and does not allow understanding the mechanisms involved in GC content evolution in plants. To tackle these issues, we used EST data from more than 200 species and provided the most comprehensive description of gene GC content across the seed plant phylogeny so far available. As opposed to the classically assumed dicot/monocot dichotomy, we found continuous variations in GC content from the probably ancestral GC-poor and homogeneous genomes to the more derived GC-rich and highly heterogeneous ones, with several independent enrichment episodes. Our results suggest that gBGC could play a significant role in the evolution of GC content in plant genomes.

INTRODUCTION

The nucleotide landscape, which is the way base composition varies along a genome, is a striking characteristic of genome organization that strongly varies among species. In eukaryotes, the mean GC content varies from ~20 to 60% (Lynch, 2007). In some species, such as humans, the base composition is highly heterogeneous, and genomes appear as patchworks of GC-rich and GC-poor regions at the 100-kb scale (Lander et al., 2001). This peculiar feature of base composition variation was called isochore structure after its discovery by ultracentrifugation (Bernardi et al., 1985). It has been well described in other vertebrates, especially in mammals and birds (reviewed in Eyre-Walker and Hurst, 2001). In these groups, the GC content varies locally and correlations are found on a small scale among GC content in coding, introns, and flanking regions (Aïssani et al., 1991) as well as among GC content in the first, second, and third codon positions (GC1, GC2, and GC3 hereafter) (Clay et al., 1996). Strikingly, other genome characteristics are associated with GC content: For instance, in GC-rich regions, gene density is higher and genes are more compact (shorter introns) (Mouchiroud et al., 1991). Recombination (Fullerton et al., 2001; Duret and Arndt, 2008), expression level (Kudla et al., 2006), and replication timing (Costantini and Bernardi, 2008) were also shown to correlate with GC content.

The evolutionary causes of isochore structure, in particular the existence of GC-rich isochores, have been much debated. The initial hypothesis of adaptation to homeothermy (Bernardi et al., 1985) has been ruled out for several reasons, including the discovery of GC-rich isochores in some reptiles, which are not homeothermic (Hughes and Mouchiroud, 2001). Three main hypotheses, not mutually exclusive, are now commonly proposed to explain nucleotide landscape evolution: selection on codon usage (SCU) in coding regions, mutational bias (MB), and GC-biased gene conversion (gBGC) (reviewed in Eyre-Walker and Hurst, 2001; Duret and Galtier, 2009). Selection for the accuracy of translation can affect synonymous codon usage (reviewed in Akashi, 2001). The GC content of coding sequences (CDSs), especially in the third codon position, could thus be driven by translational selection.

Under the SCU hypothesis, highly expressed genes are expected to undergo stronger translational selection, hence higher GC contents, if preferred codons mainly end in G or C. Variation in translational selection intensity among genes could explain within-genome variations in GC content, and variations in codon usage preferences could explain differences in GC content between species. Nevertheless, the SCU hypothesis cannot explain the GC content variations in noncoding regions and in nonsynonymous positions.

The MB hypothesis stipulates that GC content is driven by variation in MB along the genome, which can explain GC content variations at all codon positions. Several kinds of MBs have been

[1] Address correspondence to glemin@univ-montp2.fr.
The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Laurana Serres-Giardi (laurana.serres-giardi@univ-montp2.fr) and Sylvain Glémin (glemin@univ-montp2.fr).
[W] Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.111.093674

 © 2012 American Society of Plant Biologists. All rights reserved.

suggested. They include variation in bias associated with replication timing due to variations in free nucleotide availabilities during the cell cycle (Wolfe et al., 1989) and the negative effect of the local GC content on the rate of cytosine deamination, which is the major cause of C→T and G→A mutations. The latter mechanism would drive either impoverishment or enrichment in GC content in a positive feedback (Fryxell and Zuckerkandl, 2000). However, MB hypotheses were seriously weakened by polymorphism data. Several studies revealed that the mutational pattern was biased toward AT mutations and that a bias of fixation occurred in favor of GC alleles, a typical selection-like signature (e.g., Galtier et al., 2001; Smith and Eyre-Walker, 2001; Katzman et al., 2011).

Since the last decade, the gBGC hypothesis has increasingly been recognized as the main force driving GC content evolution in mammalian genomes and in other taxonomic groups (reviewed in Duret and Galtier, 2009). gBGC is a recombination-associated process that favors GC over AT bases in alleles during mismatch repair following heteroduplex formation at meiosis (reviewed in Marais, 2003). Because this process mimics selection (Nagylaki, 1983), the gBGC hypothesis is compatible with both selection-like signatures (contrary to the MB hypothesis) and with genome-wide effects (contrary to the SCU hypothesis). Moreover, as gBGC is driven by recombination, the strong correlations between recombination and GC content and GC substitution patterns are well predicted by this hypothesis (Meunier and Duret, 2004; Duret and Arndt, 2008). Isochore structure can thus be explained by the heterogeneity of recombination throughout the genome (Duret and Galtier, 2009).

As opposed to vertebrates, the base composition has only been investigated in a few plant species (e.g., Carels and Bernardi, 2000; Tatarinova et al., 2010) mainly focused on the comparison between rice (*Oryza sativa*) and *Arabidopsis thaliana* (e.g., Wong et al., 2002; Wang et al., 2004; Guo et al., 2007; Wang and Hickey, 2007). Gene GC content appeared to be heterogeneous, especially in grasses that exhibit a bimodal GC3 distribution, as opposed to other species studied so far (Carels and Bernardi, 2000; Wang et al., 2004; Tatarinova et al., 2010). In *Arabidopsis*, the correlation between GC3 and GC content of flanking regions suggested a similar isochore structure as in vertebrates (Barakat et al., 1998). Based on only several genes, such local correlations were also suggested to occur in other plants, including grasses (Matassi et al., 1989). However, a recent study in rice showed that GC3 is correlated to GC1, GC2, and GC in introns, as initially suggested, but not to flanking (1 kb) sequences (Tatarinova et al., 2010). This suggests that GC content could be structured at a very local scale around genes and/or that transposable element dynamics might have disrupted local correlations. Note that in mammals, the GC content correlation between third positions and flanking sequences is weak when using all genes (Elhaik et al., 2009) but much more clear when using orthologous more highly conserved genes (Romiguier et al., 2010). In rice, GC content was also found to strongly decrease from 5′ to 3′ along genes, whereas this GC gradient is much weaker in *Arabidopsis* (Wong et al., 2002). These striking differences between rice and *Arabidopsis* have been extended to few other eudicots species, including *Brassica* sp, pea (*Pisum sativum*), potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*), tobacco (*Nicotiana tabacum*), soybean (*Glycine soja*), and sunflower (*Helianthus annuus*), and other grasses, including barley (*Hordeum vulgare*), maize (*Zea mays*), oat (*Avena sativa*), sorghum (*Sorghum bicolor*), and wheat (*Triticum aestivum*), and are often generalized as a dicots/monocots dichotomy (e.g., Wong et al., 2002; Wang et al., 2004). This vision is clearly phylogenetically biased, and the peculiar pattern observed in Poaceae has not been observed in the few other monocots studied so far: GC content in onion (*Allium cepa*) is more similar to the one in *Arabidopsis* than in rice (Kuhl et al., 2004), and in banana (*Musa* spp), GC content appears intermediate between *Arabidopsis* and rice.

In addition, the mechanisms involved in GC content variations in plants are still weakly established, and most studies have exclusively focused on explaining the peculiar structure of Poaceae genomes. In rice and other grasses, most preferred codons end in G or C (Wang and Roossinck, 2006), suggesting that SCU can partly explain GC3 pattern in these species. However, the SCU hypothesis alone is not sufficient because weakly expressed genes also show a strongly bimodal GC3 distribution (Mukhopadhyay et al., 2007) and other positions are also affected (Carels and Bernardi, 2000; Shi et al., 2006). Other authors have thus invoked a general MB toward GC (Wang et al., 2004; Wang and Hickey, 2007). However, a recent study ruled out this hypothesis in rice by showing that, as in mammals, mutation is AT biased and fixation GC biased (Muyle et al., 2011). Finally, as in mammals and birds, gBGC was suggested to affect GC content in grasses (Glémin et al., 2006; Haudry et al., 2008; Escobar et al., 2010; Muyle et al., 2011), but we know very little on the possible occurrence of gBGC in other plant species (but see Wright et al., 2007).

Our view of both the patterns and the process of GC content evolution in plant genomes is thus very scattered. We do not know whether GC-rich and heterogeneous gene content is a specific property of grasses or whether it is shared with other monocots. We ignore whether all eudicots have GC-poor and homogeneous genomes or not. The ancestral nucleotide landscape of angiosperms remains unknown. In addition, we do not know whether the underlying mechanisms differ between groups. We thus clearly need a broader phylogenetic view of GC-content variations in flowering plants. As a comparison, even in well-studied groups as mammals, a recent study extending GC content analyses from six to eight to 33 complete genomes substantially modified the view of isochore dynamics and pinpointed the role of life history traits (body mass and longevity) and genome size on isochore evolution (Romiguier et al., 2010). In plants, there are currently too few complete genomes available, and key phylogenetic groups are still lacking (e.g., basal angiosperms, nongrass monocots, and asterids). Recently, Escobar et al. (2011) used an alternative approach and analyzed only one gene (18S rDNA) in more than 1000 angiosperm species. They showed that 18S rDNA GC content had enriched within monocots (not only in Poaceae) and suggested that gBGC could explain this pattern. They argued that 18S rDNA GC content could be a good marker of the processes involved in GC content variations among species. Yet, the 18S rDNA is clearly not sufficient to characterize GC content distributions.

Here, we propose an intermediate approach using EST sequence data from 232 species covering most angiosperm clades, plus a few gymnosperms used as outgroups, to describe global base composition heterogeneities and explore local variations at the transcript scale within and across species. EST data allow characterizing GC content in many genes in a large phylogenetic sample. We provide the most comprehensive description of gene GC content in angiosperms and gymnosperms and specifically address the following questions. How are GC richness and base composition heterogeneity distributed across the phylogeny of seed plants? Are the genomic patterns observed in rice specific to Poaceae? What are the underlying mechanisms involved in the evolution of base composition? We find a continuous pattern of nucleotide landscapes from the possibly ancestral GC-poor and homogeneous genomes to the more derived GC-rich and highly heterogeneous genomes. The data strongly suggest that GC content enrichment occurred several times independently, especially among commelinid monocots. Finally, we discuss why the data appear more compatible with the gBGC hypothesis than with the alternative ones.

RESULTS

We characterized GC content distributions in 232 seed plant species (16 gymnosperms, six basal angiosperms, 56 monocots, and 154 eudicots) using a set of 3,435,183 EST unigenes. A unigene corresponds to the assembly of a set of transcript sequences that appear to come from the same transcription locus, representing the mature mRNA (or only a fragment) of an expressed gene. After assembly, annotation, and filtering (see Methods), each species data set contained at least 1000 sequences with more than 99 nucleotides. GC contents were analyzed separately for the different nucleotide positions: 5′ untranslated region (UTR; GC5UTR), 3′UTR (GC3UTR), first codon position (GC1), second codon position (GC2), and third codon position (GC3). We focused mainly on GC3. This is because most variations in the third coding position are synonymous and almost neutral or weakly selected. Therefore, GC3 should better reflect genome-wide variations in base composition. Results for all species are available in Supplemental Data Set 1 online.

Validation of EST Data by Comparison with Complete Transcriptome Data

Complete transcriptome sequences (i.e., the set of complete annotated CDS) of seven species (three dicots: *Arabidopsis*, poplar [*Populus trichocharpa*], and grape [*Vitis vinifera*]; and four grasses: *Brachypodium distachyon*, rice, sorghum, and maize) were used as controls to test whether EST data gave representative and unbiased results of all transcripts to describe the base composition patterns. Mean GC3 contents computed with complete transcriptomes were very similar to those computed with EST data (see Supplemental Data Set 2 online). Because many EST corresponded to gene fragments, sampling variance on GC content inflated the SDs computed with EST data compared with those computed with complete genomes (see Supplemental Data Set 2 and Supplemental Figure 1 online). To tackle this problem, we modeled the GC content distributions and took sampling variance into account to fit the model to the data. To capture the heterogeneity of GC content distribution, especially the cases with bimodality, we assumed that GC content followed a bi-Beta distribution that is the mix of two BetaAU: distributions in proportions p and $1 - p$ with density function:

$$f(x,\alpha_1,\beta_1,\alpha_2,\beta_2,p) = p\frac{x^{\alpha_1-1}(1-x)^{\beta_1-1}}{B(\alpha_1,\beta_1)} + (1-p)\frac{x^{\alpha_2-1}(1-x)^{\beta_2-1}}{B(\alpha_2,\beta_2)} \quad (1)$$

where x is the GC content and $B(\alpha, \beta)$ is the Beta function (Abramowitz and Stegun, 1970). The use of two distributions is sufficient to obtain good fits and to give results with the simple biological meaning of two classes of genes regarding GC content (Carels and Bernardi, 2000). We fitted the five parameters of the bi-Beta using maximum likelihood (see Methods). Fitted parameters were used to obtain estimates of the mean and the SD of GC content distributions that (partly) correct for sampling variance. The estimated means of the bi-Beta fits were very close to the raw averages and the complete transcriptome averages (see Supplemental Data Set 2 online). For the whole species set, raw averages and bi-Beta means were also mostly identical (see Supplemental Figure 2 online; Spearman's rho = 0.999, P value < 10^{-15}). The estimated SDs of the bi-Beta fits were lower than the raw SDs and much more similar to the SDs of the complete CDS sets than the raw SDs (see Supplemental Data Set 2 online). However, though overestimated, raw SDs were strongly correlated to the bi-Beta SDs (see Supplemental Figure 2 online; Spearman's nonparametric coefficient of correlation rho = 0.997, P value < 10^{-15}). More generally, the GC3 bi-Beta distribution estimated through EST data was also similar to the full GC3 distribution obtained from whole-genome transcripts (see Supplemental Figure 1 online). Because estimates obtained through the bi-Beta distribution were closer to the full distribution than raw estimates, we chose to further use these corrected estimates instead of raw values. This approach also allowed us to characterize the properties of the two underlying distributions possibly corresponding to two classes of genes (see below).

Since we found the results for GC1 and GC2 to be similar to those for GC3 (see Supplemental Data Set 2 and Supplemental Figure 2 online), we also used corrected estimates for GC1 and GC2. GCUTR were more problematic because there were fewer sequences and they were shorter, which resulted in irregular raw distributions. For many species, we were thus not able to fit a bi-Beta distribution because of convergence problems during maximum likelihood maximization. We thus used the raw averages and SDs. However, because raw and corrected averages and SDs were highly correlated, relative results should be very similar. Moreover, raw means and SDs were quite close to the transcriptome estimates (see Supplemental Data Set 2 online).

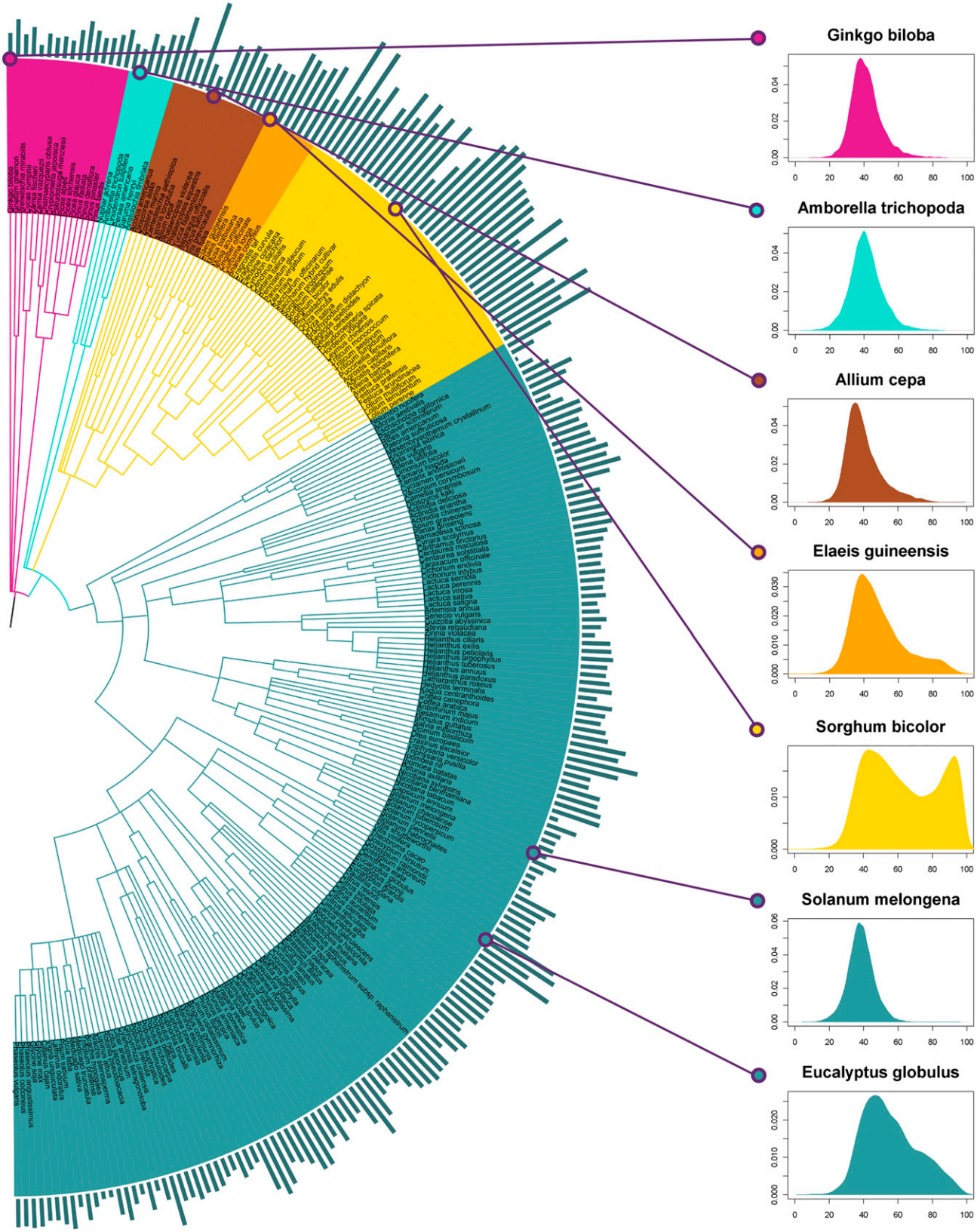


Figure 1. Taxonomic (NCBI) Trees with Unbiased Mean GC Content at Third Codon Position (GC3) per Species and GC3 Distributions for Seven

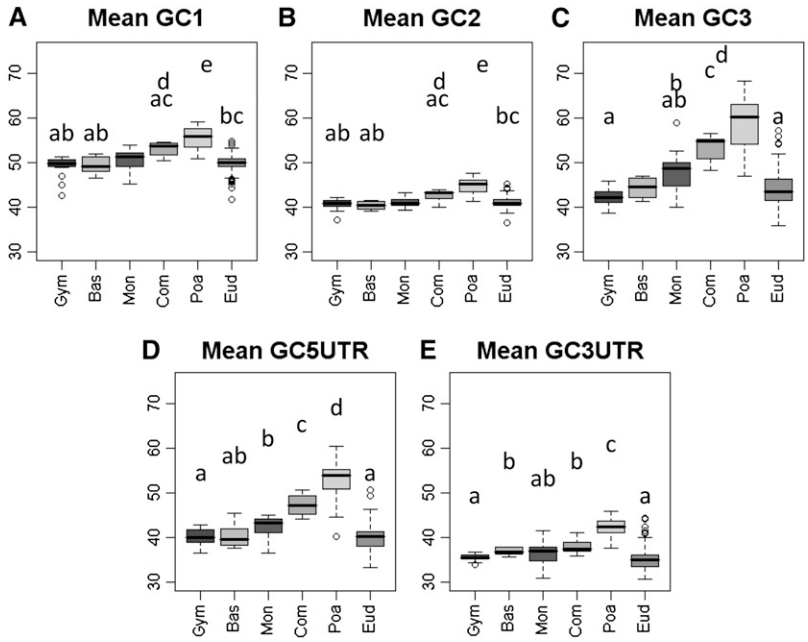


Figure 2. Box Plots of Mean GC per Species by Taxonomic Group.

(A) to **(E)** Box plots showing the among-species variations of GC content at different coding and UTR positions for six taxonomic groups. On each box plot, the dark midline is the median, the bottom and top of the box are the lower and upper quartiles, respectively, and the ends of the whiskers are the lowest and highest data still within 1.5 times the interquartile range of the lower and higher quartiles, respectively. Significant differences between groups according to Kruskal-Wallis tests are indicated with letters. Bas, basal angiosperms; Com, non-Poaceae commelinids; Eud, eudicots; Gym, gymnosperms; Mon, noncommelinid monocots; Poa, Poaceae.
(A) Mean GC1 (unbiased estimate).
(B) Mean GC2 (unbiased estimate).
(C) Mean GC3 (unbiased estimate).
(D) Mean GC5UTR (raw estimate).
(E) Mean GC3UTR (raw estimate).

Finally, we verified that the size of the EST bank or the number of unigenes did not affect the results (see Supplemental Figure 3 online).

Strong Variation in GC Richness across the Seed Plant Phylogeny

We found strong variations in the mean GC content among species for all coding positions, with GC3 being the most contrasted (Figure 1). Unbiased mean GC3 ranged from 35.9% for the Papaveraceae species *Eschscholzia californica* to 68.2% for the Poaceae species *Secale cereale* (rye; see Supplemental Data Set 1 online). Phylogeny strongly impacts GC3 variations (Figure 1). To detect how variations in GC3 are distributed on the phylogeny, we estimated the variance component among six nested taxonomic levels (Table 1). We found that most variations (38.0%) occurred at the highest taxonomic level that included gymnosperms (mean GC3 = 42.4%), basal angiosperms (44.4%), monocots (55.8%), and eudicots (43.9%). Most of the variation was due to the difference between monocots and the other three groups (P values < 0.0001), whereas no significant difference was found between gymnosperms, basal angiosperms, and eudicots (Figure 2; see Supplemental Table 1 online). The rest of the variance mainly arose at the lower taxonomic levels, species (21.2%), family (18.9%), and genus (11.3%) (Table 1).

Figure 1. (continued).

Representative Species.
The mean GC3 of each species is plotted on the phylogenetic tree of seed plants. The height of the blue bars is proportional to mean GC3. The GC3 distribution is also plotted for seven representative species (*x* axis, GC3 in %; *y* axis, proportion of unigenes). Species colors: violet red, gymnosperms; turquoise, basal angiosperms; brown, noncommelinid monocots; orange, non-Poaceae commelinids; gold, Poaceae; and cadet blue, eudicots. Clade colors: violet red, gymnosperms; turquoise, basal angiosperms; yellow, monocots; and cadet blue, eudicots. The tree was plotted with the phylogenetic display and manipulation online tool Interactive Tree Of Life version 2 (Letunic and Bork, 2011).

Table 1. Variance Component Estimates (in %) of the Effect of Taxonomy on GC3 (Mean and SD)

Data Set	Statistics	Group	Infra Group	Super Order	Order	Family	Genus	Species
Total	Mean	38.0	0.0	8.6	2.0	18.9	11.3	21.2
	SD	52.1	0.0	10.6	0.0	13.5	12.0	11.8
Monocots	Mean	–	–	38.6	0.0	14.7	0.0	46.7
	SD	–	–	48.9	0.0	16.2	0.0	34.9
Eudicots	Mean	–	0.0	0.0	19.0	45.3	21.8	14.0
	SD	–	0.0	11.6	7.6	35.5	32.6	12.6

The variations in GC3 mean and SD are decomposed into the proportion due to every nested taxonomic level. For instance, for the full data set, 38% of mean GC3 variation is due to average differences among the four groups (gymnosperms, basal angiosperms, monocots, and eudicots), and 21.2% of variations are due to differences among species that are not attributable to higher taxonomic levels. Values for Group and Infra Group are not defined for the monocot data set because these taxonomic levels are higher than the monocot level and, hence, not nested within it. Similarly, values for Group are missing for the eudicot data set because these taxonomic levels are higher than the eudicot level.

We also found strong variations in the mean GC3 within monocots, from 40.0% in onion to 68.2% in *S. cereale*, and 38.6% of the variance was due to the difference between commelinids (mean GC3 = 58.2%) and other monocots (mean GC3 = 48.1%) (P value < 0.0001; Table 1; see Supplemental Table 1 and Supplemental Figure 4 online). Because grasses were known to have peculiar GC content (e.g., Carels and Bernardi, 2000; Wang et al., 2004), we also distinguished them from other commelinid species. Surprisingly, we also found significant differences at this level (mean GC3: Poaceae = 59.2%, other commelinids = 53.1%; P value = 0.0163). However, the major part of the variance (46.7%) is due to differences between species that are not due to higher taxonomic levels. This is well illustrated by the wide variation observed within grasses, from bamboo (*Phyllostachis edulis*) (mean GC3 = 46.9%) to rye (mean GC3 = 68.5%). Important variations also occurred in eudicots (from 35.9% in *E. californica* to 57.1% in *Eucalyptus grandis*), but variation was more evenly distributed among the different taxonomic levels (Table 1) and no clear phylogenetic pattern emerged (Figure 1; see Supplemental Figure 1 online). Some orders appeared much more GC rich than the average, such as Myrtales, whereas others were somewhat lower, such as Ranunculales (see Supplemental Figure 1 online). We found much lower variations in gymnosperms (from 38.7% in *Zamia fischeri* to 45.8% in *Picea sitchensis*) and basal angiosperms (from 41.4% in *Amborella trichopoda* to 47.0% in *Aristolochia fimbriata*). Consequently, we compared the four main groups (gymnosperms, basal angiosperms, monocots, and eudicots), and among monocots, we distinguished Poaceae, other commelinids, and other monocots.

GC1, GC2, GC5UTR, and GC3UTR were similar to GC3 (Figure 2). No significant difference was found between gymnosperms, basal angiosperms, and eudicots (except for GC3UTR; significant with P value < 0.05). Monocots were significantly more GC rich than other species (P values < 0.01) and among monocots, commelinids, and Poaceae were the richest (P value < 0.001 for commelinid versus noncommelinid monocots; for Poaceae versus non-Poaceae commelinids, P value < 0.05 for GC1, P value < 0.01 for GC2 and GC5UTR, P value < 0.001 for GC3UTR) (Figure 2; see Supplemental Table 1 online). These GC contents were more variable in monocots than in eudicots (Figure 2). On average, GC1 was higher than GC2 and GC5UTR higher than GC3UTR (see Supplemental Table 1 online).

For all positions, the highest values were found among monocots, with punctually high values in eudicots (e.g., *Eucalyptus* spp and Myrtales in general; see Supplemental Figure 4 online), whereas gymnosperms, basal angiosperms, and the majority of eudicots showed low values (Figures 1 and 2; see Supplemental Figure 5 online). This suggests that GC richness is a derived state among flowering plants and that the processes or conditions favoring GC enrichment have emerged in different taxa. It is also worth noting that strong variation between species also occurred at low taxonomic levels. For instance, in Poaceae, some species have rather low GC content, lower than other monocots and even than some eudicots (such as Myrtales).

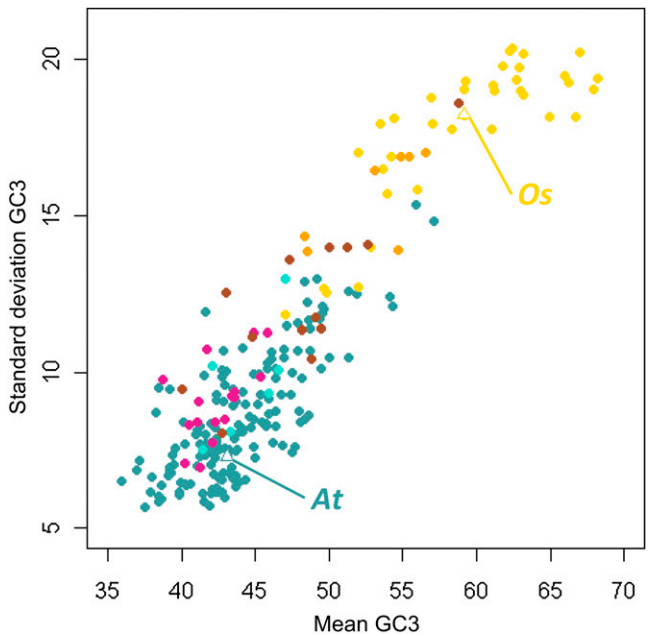


Figure 3. Unbiased SD GC3 as a Function of Unbiased Mean GC3.

The scatterplot shows the strong positive correlation between GC richness (mean GC3) and GC heterogeneity (GC3 SD). Violet red, gymnosperms; turquoise, basal angiosperms; brown, noncommelinid monocots; orange, non-Poaceae commelinids; gold, Poaceae; cadet blue: eudicots. At, *Arabidopsis*; Os, *O. sativa*.

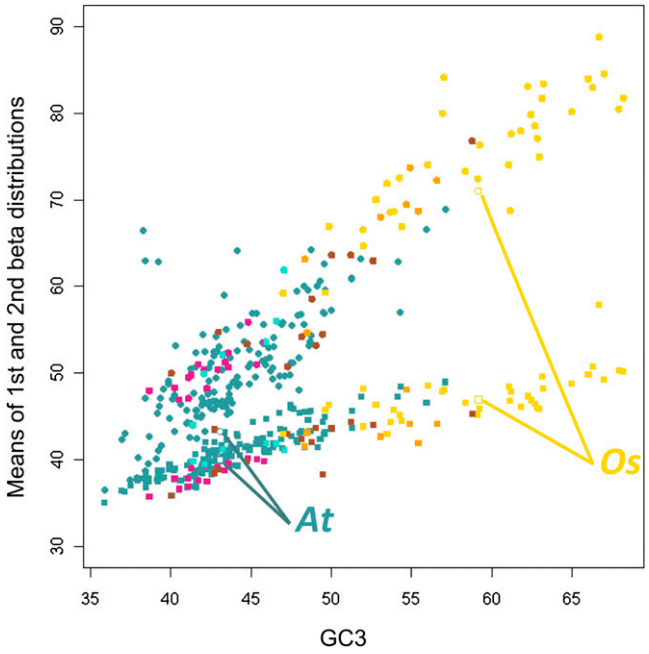


Figure 4. Means of the Two Beta Distributions as a Function of Unbiased Mean GC3.

The scatterplot shows that the means of the two Beta distributions correlate positively with the mean GC3. Squares, mean of the first Beta distribution; circles, mean of the second Beta distribution. Violet red, gymnosperms; turquoise, basal angiosperms; brown, noncommelinid monocots; orange, non-Poaceae commelinids; gold, Poaceae; cadet blue, eudicots. At, *Arabidopsis*; Os, *O. sativa*.

Covariation of GC Richness and GC Heterogeneity among Species

Differences in base composition between species do not necessarily imply differences in GC content heterogeneity. We thus tested whether the GC content variation among species was a homogeneous process at the genome scale or whether only some gene categories were affected. GC3 SDs were strongly variable among species (see Supplemental Figure 5 online). They ranged from 5.65% for the Solanaceae *Petunia axillaris* to 20.38% for the Poaceae *Panicum virgatum* (see Supplemental Data Set 1 online). The phylogenetic pattern is similar to the mean GC3, with most variations (52.1%) occurring at the highest taxonomic levels (Table 1). GC3 SD was not significantly different between gymnosperms, basal angiosperms, and eudicots but significantly higher in monocots compared with the other groups (P values < 0.001). As for the mean GC3, we also found most variations within monocots: Commelinids were significantly more variable than other monocots (P value = 1.10^{-5}) and Poaceae more variable than other commelinids (P value = 0.0077) (see Supplemental Table 2 and Supplemental Figure 5 online). GC1, GC2, GC5UTR, and GC3UTR SDs followed the same pattern: Monocots, commelinids, and Poaceae were the most heterogeneous groups (except GC3UTR for commelinids, nonsignificant) (see Supplemental Table 2 online). As above, the patterns of variation were similar for all positions and accentuated for GC3. Therefore, we chose to focus on GC3 for the next analyses.

This global analysis showed that monocots and especially Poaceae were both more GC rich and more heterogeneous in GC content than the other clades of seed plants. This relation held true at the whole species range. We found a strong positive correlation between mean GC3 and GC3 SD (Spearman's rho = 0.83, P value < 10^{-15}). The richest GC3 species were the most heterogeneous (Figure 3). It is worth noting that this pattern is not expected under a null neutral model under which variance should be binomially related to the mean with a maximum variance for a mean of 0.5. This strong correlation held true after several controls for taxonomy (see Methods). We found the same positive correlation at all taxonomic levels and within almost all taxonomic groups (see Methods). We also verified that this relationship was neither affected by the number of raw EST nor by the number of unigenes (see Supplemental Figure 3 online). We found the same kind of pattern for GC1, GC2, GC5UTR, and GC3UTR.

Our results showed a continuous relationship between GC richness and heterogeneity, which suggests that the same evolutionary mechanisms have likely been involved in all seed

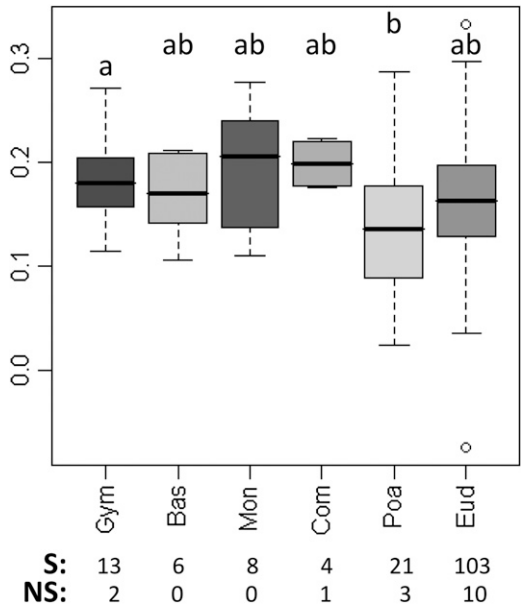


Figure 5. Spearman's Rho between GC3 and Expression Level.

The box plot shows that GC3 is positively correlated with expression level in most species, irrespective of their taxonomic group. Only species with significant correlation (Spearman's test, P value ≤ 0.05) were plotted. On each box plot, the dark midline is the median, the bottom and top of the box are the lower and upper quartiles, respectively, and the ends of the whiskers are the lowest and highest data still within 1.5 times the interquartile range of the lower and higher quartiles, respectively. The number of species with significant correlation (S) and nonsignificant correlation (NS) is indicated below the graphs. Significant differences between groups according to Kruskal-Wallis tests are indicated with letters. Bas, basal angiosperms; Com, non-Poaceae commelinids; Eud, eudicots; Gym, gymnosperms; Mon, noncommelinid monocots; Poa, Poaceae.

plants. An intensification of this process may have occurred independently in some groups, such as commelinids and Myrtales.

Evolution of the GC Content Distribution among Species

GC Content Classes of Genes within Species

The distribution of gene GC content, especially GC3, varied among species, from a symmetric unimodal distribution to strongly skewed distributions toward high GC content and even to bimodal distributions in some grasses (Figure 1). The sole mean and SD of these distributions were not sufficient to capture such heterogeneity. The characteristics of the bi-Beta distribution allowed us to better capture this heterogeneity, especially the means of the two underlying Beta distributions and their proportion.

Figure 4 shows that the two bi-Beta means increases when global GC3 increases. The increase in the second mean was much stronger than in the first one. Both means were strongly correlated to the global mean (Spearman's rho = 0.92 and 0.84, respectively; P value < 10^{-15}). This phenomenon was observed in a continuous way in gymnosperms, basal angiosperms, and eudicots up to monocots (Figure 4). Outliers for the second mean corresponded to species with a very low proportion of the second distribution. We also found that the proportion of the second Beta distribution increased with the mean GC3 (Spearman's rho = 0.36; P value = 1.80 10^{-8}), though the pattern was weaker than those for the two means. These results were robust to taxonomic control (see Methods). These results suggest that the GC3 enrichment is partly due to the increase in the proportion of the GC3-rich class of genes. However, they reject the hypothesis that the global GC3 increase is only due to the appearance of this GC3-rich class without enrichment of the GC3-poor class. On the contrary, it suggested that there was a global, but uneven, trend to enrichment leading to the evolution of both GC3-poor and GC3-rich classes of genes.

Association between GC Richness, Size, and Expression

Several studies in various organisms, including several plants, have reported that GC3-rich genes tend to be shorter than GC3-poor genes (e.g., Duret and Mouchiroud, 1999; Carels and Bernardi, 2000; Stoletzki, 2011). To test whether this association is also found in seed plants, we compared the length of GC3-rich and GC3-poor genes. We only kept species with at least 500 GC3-poor and 500 GC3-rich complete CDSs (see Methods). For 76 species with enough data over 78, GC3-rich genes were shorter than GC3-poor genes. The difference was significant in 69 species (i.e., three out of four gymnosperms, 56 out of 64 eudicots, and all 10 Poaceae monocots) (see Supplemental Data Set 1 online).

We also tested whether GC3-rich genes were more expressed than GC-poor genes, as expected under the SCU hypothesis. For 154 species of the 171 we tested, the expression levels were positively and significantly correlated with GC3 (Figure 5). Importantly, as opposed to other characteristics, we found no differences in the correlation coefficients between groups (Figure 5) and no positive relationship between mean GC3 and the correlation strength (Spearman's rho = 0.12; P value = 0.127). In particular, we did not find a stronger correlation between GC3 and expression in Poaceae. These results were probably not due to statistical noise induced by small EST databases because the same patterns were observed excluding the 20% or the 50% smaller data sets. However, because of other heterogeneities among EST libraries, this absence of differences between groups must be viewed with caution. Strong differences between groups should have been captured by this analysis, but we might have missed smaller differences.

Local Variations in GC Content within Species

Local Correlations of GC Content between Positions

We found that all coding positions globally exhibited the same patterns, the third position being more contrasted. This suggests that the same process could affect all positions and that intragenomic variations could result from local variations in GC content, as observed in species exhibiting isochore structure like mammals and birds (Eyre-Walker and Hurst, 2001). In our data set, we found significant positive correlations among GC3, GC1, and GC2 (Figure 6). This is in agreement with the hypothesis of local variation in GC content. The strength of these correlations differed between species groups (Figure 6). Local correlations were significantly stronger for monocots than eudicots, gymnosperms, and basal angiosperms (Kruskal-Wallis nonparametric identity test between samples: P value < 0.01). Correlations between GC1 and GC3, and between GC2 and GC3, strongly correlated with the mean GC3 (Figure 6; P value < 10^{-15}). This suggests that the genome base composition was more locally structured in GC-rich than in GC-poor species. For these correlations, significant quantitative differences were observed between the seven complete transcriptome data sets and their corresponding EST data sets, especially in eudicots. This is not unexpected because correlation coefficients were less robust than means to the noise introduced by sampling EST. However, despite these quantitative differences, the qualitative trends remained robust. The extension of local correlation to flanking regions has been questioned (Tatarinova et al., 2010). Accordingly, the correlations between GC3 and GC3UTR and GC5UTR were low but still significant in most species (see Supplemental Data Set 1 online). This was also true for the seven complete transcriptomes (see Supplemental Data Set 2 online). Variations between groups appeared noisy but the intensity of the correlation between GC3 and GCUTR is also positively correlated with the mean GC3 (Spearman's rho = 0.16, P value = 0.016 for GC3UTR; Spearman's rho = 0.28, P value = 1.38 10^{-5} for GC5UTR).

Gradients of GC Content along Transcripts

At a more local scale, gradients of GC content within genes have been previously documented in rice (Wong et al., 2002) and nonplant species, such as yeast (*Saccharomyces cerevisiae*) and drosophila (*Drosophila melanogaster*) (Qin et al., 2004; Stoletzki, 2011). In these species, GC3 decreases from the 5′ to the 3′

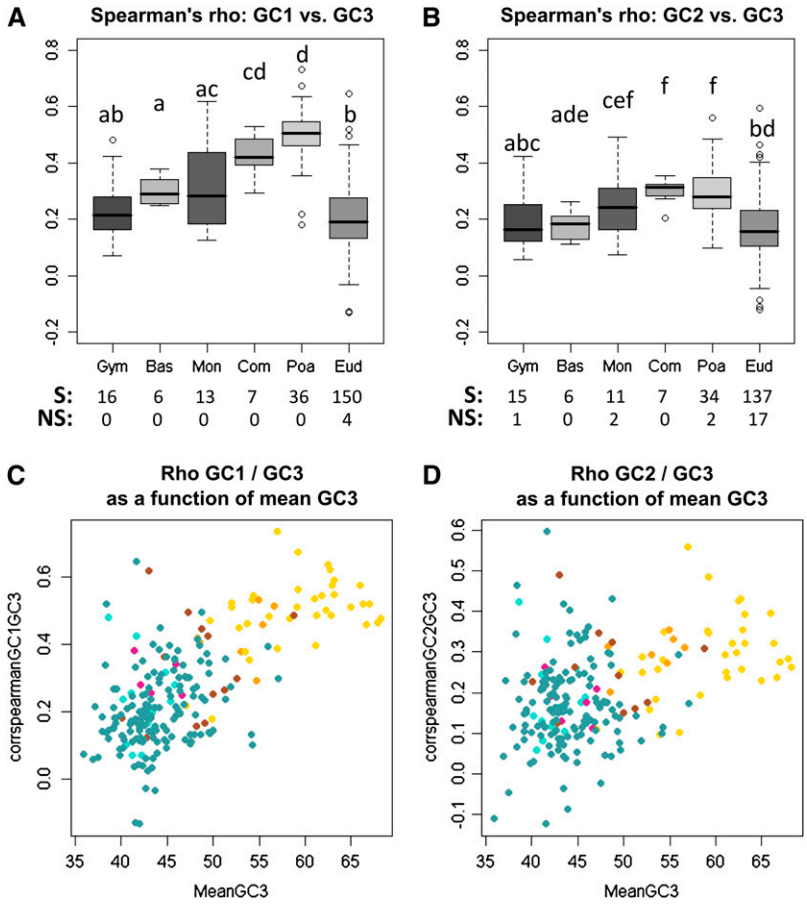


Figure 6. Spearman's Rho between Local GC3 and GC1 and between Local GC3 and GC2: Box Plots and Plots as Functions of Unbiased Mean GC3.

(A) and **(B)** Box plots of Spearman's correlation coefficients between GC1 and GC3 **(A)** and GC2 and GC3 **(B)** within the genome for six taxonomic groups

(C) and **(D)** Scatterplots of Spearman's correlation coefficients between GC1 and GC3 **(C)** and GC2 and GC3 **(D)** as a function of mean GC3.

Only species with significant correlation (Spearman's test, P value ≤ 0.05) were plotted. On each box plot, the dark midline is the median, the bottom and top of the box are the lower and upper quartiles, respectively, and the ends of the whiskers are the lowest and highest data, respectively, still within 1.5 times the interquartile range of the lower and higher quartiles, respectively. The number of species with significant correlation (S) and nonsignificant correlation (NS) are indicated below the graphs. Significant differences between groups according to Kruskal-Wallis tests are indicated with letters. Violet red, gymnosperms (Gym); turquoise, basal angiosperms (Bas); brown, noncommelinid monocots (Mon); orange, non-Poaceae commelinids (Com); gold, Poaceae (Poa); cadet blue, eudicots (Eud).

ends. As previously noted in rice (Wong et al., 2002), we also found that GC5UTR is higher and more variable than GC3UTR in most species (Figure 2). To go further, we measured GC3 decay along transcripts by the regression slope between GC3 and the distance from the start codon (see Methods). For all of the species with sufficient data to perform the test (four gymnosperms, 65 eudicots, and 12 Poaceae), we found a negative GC3 gradient along transcripts (Figure 7). The strength of this gradient increased with the mean GC3 content of species in a continuous way from eudicots to monocots. GC3 gradients were significantly stronger in monocots than in eudicots (P value = 2.10^{-7}). Similar gradients were also found for the first and second positions (see Supplemental Data Set 1 online). These results suggests that the negative GC gradient from the 5′ to the 3′ ends seemed to be a general feature of angiosperm and gymnosperm genomes and that the most GC-rich and GC-structured species also exhibited the strongest GC gradient.

Relationship between Recombination and GC Content

In mammals (Duret and Arndt, 2008), birds (Nabholz et al., 2011), and yeast (Birdsell, 2002) (among others), within-species variations in GC content are partly explained by variation in recombination rates: In these groups, GC content strongly and positively correlates with local recombination rate, which is seen as a strong argument in favor of the gBGC hypothesis. For most species of our data set, the effect of recombination on GC content cannot be tested because physical location of EST and recombination map are lacking. However, it could be tested in the few species for which good physical and genetic maps are

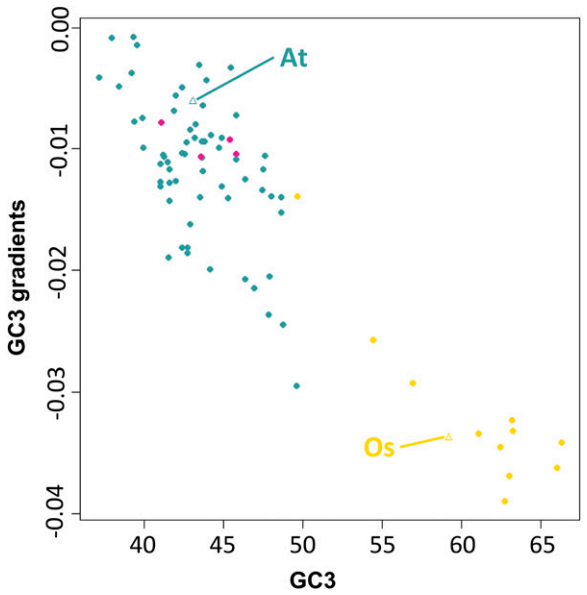


Figure 7. GC3 Gradients along Transcripts as a Function of Mean GC3.

The scatterplot shows the strong correlation between GC richness (mean GC3) and the steepness of the GC gradient along transcripts. Violet red, gymnosperms; cadet blue, eudicots; gold, monocots (Poaceae). At, *Arabidopsis*; Os, *O. sativa*.

available: *Arabidopsis* and the three grasses rice, maize, and *B. distachyon*. In *Arabidopsis*, GC3 does not correlate with the local recombination rate (Marais et al., 2004) of GC1 or GC2 (Giraut et al., 2011). In grasses, this relationship has been overlooked and has not been studied at the gene level. Using complete genome data and available genetic and physical map, we analyzed the relationship between local recombination rate and GC content in rice, maize, and *B. distachyon*. In the three species, we found a highly significant positive correlation between local recombination rate and GC3. At the gene level, the correlations are weak but highly significant (Spearman rho = 0.081, 0.068, and 0.052 for rice, maize, and *B. distachyon*, respectively; P value < 10^{-15} for the three species), but the relationship appears very clearly by grouping genes by recombination rate bins (Figure 8). We also found a positive correlation with GC1 and GC3UTR in the three species, with GC5UTR in maize and *B. distachyon*, and with GC2 in *B. distachyon* (see Supplemental Table 3 and Supplemental Figure 6 online). We found the same results when genes annotated as transposable elements were included in the rice and maize data set (see Supplemental Table 3 online). Surprisingly, in previous studies, only weak (but significant) correlation was found in maize (Gore et al., 2009), and no correlation was found in rice (Tian et al., 2009) and *B. distachyon* (Huo et al., 2011). However, in these studies, recombination rate has been correlated to GC content at the hundreds of kilobase or at the megabase scale without distinguishing genic and intergenic regions, which might have obscured the relationship between recombination and base composition.

DISCUSSION

Robustness and Validation of the EST Data Approach

Thanks to the increasing availability of EST data, we explored GC content distributions in 232 seed plant species covering the major clades of both gymnosperms and angiosperms. This is more than one order of magnitude larger than previous studies (e.g., Tatarinova et al., 2010). The downside of this approach is the potentially lower quality and partial representativeness of data sets compared with complete genome data. However, several controls strongly support the validity of our results. First, we applied a stringent filtering procedure to select confidently annotated unigenes within each species and species with at least 1000 good quality unigenes. More stringent filters were then applied for specific analyses. Despite stringent filtering, strong variations in the size of the data sets remained. Yet, the smallest data sets were evenly distributed across the seed plant phylogeny, and if such data sets were problematic, they would have obscured the observed patterns. On the contrary, we found strong and significant patterns, suggesting that the data quality was good and/or that the signal was strong enough to clearly emerge from our analyses.

We also verified that the size of the EST bank or the number of unigenes did not affect the results (see Supplemental Figure 3 online). Second, we checked the validity of our results by comparing them to complete genome data in three eudicots and three grasses. We found that GC content distribution among ESTs mirrored very well the complete transcriptome distribution. In addition, the bi-Beta fit we applied was efficient to reduce additional variance in GC content distribution due to the

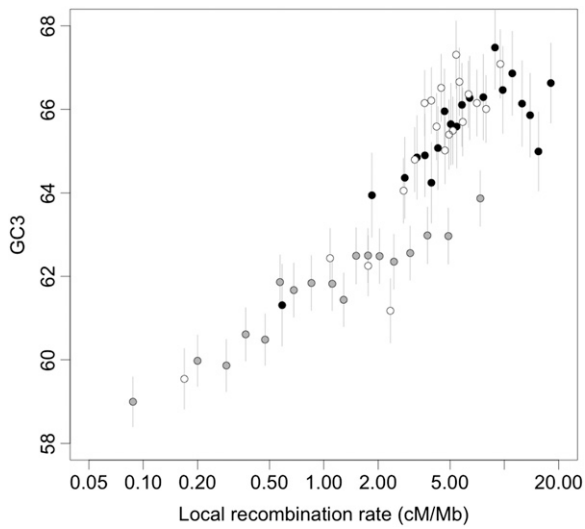


Figure 8. Relationship between Local Recombination Rate and GC3 in Three Grass Species.

The scatterplot shows the strong positive correlation between local recombination rate and GC3 in three grass species. Genes have been grouped into 20 bins according to their local recombination rate. Dots correspond to the mean GC3 of each bin and bars to the SEs. Black dots, *B. distachyon*; gray dots, maize; white dots, rice.

sampling of gene fragments inherent to EST data. The absolute values of the correlation between the three codon positions appeared less robust (this is not unexpected because of the composed nature of correlation coefficients), but the relative order among species was perfectly kept. Finally, the correlations between GC3 and GCUTR appeared more problematic, but they were weak using both EST and complete genome data sets. The lack of correlation between GC3 and flanking GC content has already been mentioned and debated in plants (Tatarinova et al., 2010), and it is not specific to our data set. All in all, our data set was appropriate to characterized GC content patterns at the gene level and gives a general view of nucleotide landscape variations in seed plants. In the near future, the increasing availability of complete genome data will then help to refine our results and to test further our conclusions.

A More Complete and Complex View of Seed Plant Nucleotide Landscapes

GC content of plant genomes is generally reported as a dichotomy opposing GC-rich and heterogeneous monocots (at least grass) to more GC-poor and homogeneous (eu)dicots (e.g., Wong et al., 2002; Wang et al., 2004). Although our results confirmed this pattern, that is, on average Poaceae genomes are GC richer and more heterogeneous than genomes of eudicots, our broad phylogenetic survey gave a more complete and rather different picture: Variation in nucleotide landscapes in plants is continuous. Moreover, this continuous variation holds for all the characteristics we analyzed: GC-rich genomes are also more heterogeneous (Figure 3) and display stronger local correlations among codon positions (Figure 6) and stronger GC content gradient along genes (Figure 7). As previously proposed (Carels and Bernardi, 2000), two classes of genes naturally emerged from our bi-Beta parameterization. It allowed us to test how GC content evolved within and between these two classes across the seed plant phylogeny. We found that GC enrichment is not only due to the appearance of a GC-rich class of genes but that the GC-poor gene category is also affected. However, GC-rich genomes are characterized by a higher proportion of GC-rich genes and a larger difference in mean GC content of the two classes (Figure 4). We thus found a continuous pattern of GC3 distributions from slightly to strongly skewed toward high GC, up to bimodal in some Poaceae, some commelinid monocots, and some eucalyptus species in eudicots (see example in Figure 1). Such bimodal distributions have already been described in grasses (Tatarinova et al., 2010) and seem to also occur in some mammals (J. Romiguier, personal communication) and birds (B. Nabholz, personal communication). Bimodal distributions seem to occur only in very GC-rich genomes, but it is still unclear which process drives the evolution of bimodality. Overall, the correlations between the various properties of GC content distributions among species and the continuous variations in these patterns suggest that the same evolutionary process might be at work for all plant groups, with varying intensities, both at the global and the local genomic scales.

Because we did not compare orthologous sequences, we were not able to reconstruct ancestral GC content distribution. However, using parsimony reasoning, it is possible to propose a global scenario for the evolution of nucleotide landscapes in seed plants. Most species, including gymnosperms, basal angiosperms, noncommelinids monocots, and early-diverged eudicots, exhibit GC-poor and homogeneous genomes, which was likely the ancestral state. Enrichment in GC content likely occurred independently several times, especially in commelinids and Poaceae, as recently suggested by Escobar et al. (2011), but also in some eudicot orders such as Myrtales (including *Eucalyptus*) (see Supplemental Figure 4 online). Moreover, some grasses are relatively GC poor and homogeneous, offering the possibility of independent enrichment even within grasses. GC content impoverishment would also be possible, such as in Ranunculales: It is the GC-poorest order of our data set, especially compared with gymnosperms and basal angiosperms, which suggest the ancestral angiosperm genome was richer than Ranunculales (see Supplemental Figure 4 online). These GC content dynamics parallel the one observed in vertebrates where a strong increase in GC content occurred independently in several groups, such as mammals, birds, and probably some fishes (Duret and Galtier, 2009; Escobar et al., 2011). Independent enrichment in GC content has also been documented within mammals (Romiguier et al., 2010). These parallels suggest that comparisons between plants and vertebrates could help understand the underlying mechanisms of these GC content variations, as discussed below.

Explanatory Hypotheses for GC Content Variations among and within Plant Genomes

From studies in mammals and yeast, three main nonexclusive hypotheses have been put forward to explain GC content variations (reviewed in Eyre-Walker and Hurst, 2001; Duret and Galtier, 2009): selection, MB, and gBGC. We now discuss whether the patterns we observed are or are not in agreement with each of these three hypotheses. We also consider the possibility that other (unrecognized) mechanisms could affect the evolution of GC content in plant.

SCU

In many species, selection for translation accuracy affects codon usage (Akashi, 2001; Duret, 2002). If the intensity and the direction (in favor of codons ending in GC or AT) vary from one species to another, SCU could explain GC content variation across species but also within the genome through the variation in selection pressure among genes. However, the SCU hypothesis can only explain GC3 (and partly GC1) patterns and is thus insufficient to explain other observations, especially local correlations among GC3, , GC2, and GCUTR that were also observed with introns in few species (e.g., Tatarinova et al., 2010). However, variations in GC3 are much stronger than at other positions. Could SCU explain most of the variation in GC3 anyway? In most species, GC3 is positively correlated with expression levels (Figure 5). This is consistent with SCU because in all plants studied so far, most preferred codons, which are more frequent in highly expressed genes, end in G or C (Wang and Roossinck, 2006). However, as opposed to other patterns, there is no clear difference between the species groups (Figure 5). Variation in

SCU intensities can thus hardly explain the strong variations in GC3 we observed. In addition, as previously observed (Ren et al., 2006), highly expressed genes tend to be longer in our data set. We would thus expect GC-rich genes to be longer. However, we observed the reverse pattern in most species, suggesting SCU is not strong enough to overwhelm other causes making GC-rich genes shorter. Finally, in rice, the distributions of both highly and weakly expressed genes are bimodal (Mukhopadhyay et al., 2007); hence, SCU is not sufficient to explain bimodality in this species and likely in other grass species. Further analyses will be necessary to quantify more precisely the role of SCU and whether it plays a marginal or significant role in the evolution of GC3.

MB

Variation in MB, both among species and within the genome, could alternatively drive the variations in GC content we observed. Contrary to the SCU hypothesis, the MB hypothesis could explain variations at all positions. However, the MB hypothesis must also explain the association between GC richness and heterogeneity, so that mutation bias evolution should be linked to an increase in mutation bias variation along genomes and along genes (5′-3′ gradient). Direct data on mutation rate and bias are still too scarce to give firm conclusions, in particular because we do not known whether bias varies with GC content, as proposed by Fryxell and Zuckerkandl (2000). Wong et al. (2002) suggested that the GC gradient could be due to transcription-coupled DNA repair (TCR) mechanism, whose rate of repair decreases from 5′ to 3′ along transcribed regions (Svejstrup, 2002). However, they did not explain why the differential repair rate should induce a GC MB. Overall, the data available so far mainly disagree with the MB hypothesis. First, mutation seems to be AT biased in most eukaryotes (Lynch, 2007), as reflected by the most early-diverged species with a mean GC3 lower than 50%. An inversion of this mutation bias would thus be required to explain the emergence of GC-rich genomes, as in grasses. But mutation has been showed to be also AT biased in rice (Muyle et al., 2011). Variation in methylation patterns would be a possible cause. In plants, genomic methylation patterns appear highly conserved (Feng et al., 2010; Zemach et al., 2010), but the methylation level is much higher in rice than in poplar and *Arabidopsis* (Feng et al., 2010). However, methylation tends to increase the mutation bias toward AT as methylated CpGs are usually highly mutable toward TpG (Nachman and Crowell, 2000). This would thus predict lower GC content in rice than in *Arabidopsis* and poplar. Finally, the analysis of polymorphism data in rice showed that AT→GC mutations experienced higher probability of fixation than GC→AT mutations, which is incompatible with the neutral MB hypothesis that predict equal probability of fixation for both kind of mutations (Muyle et al., 2011). Such analyses of polymorphism data in many groups would be necessary to definitively reject (or not) the MB hypothesis as a global explanation.

gBGC

The third hypothesis posits that variation in the occurrence and strength of gBGC can explain both GC content heterogeneity within the genome and the difference between species and taxonomic groups (Eyre-Walker and Hurst, 2001; Duret and Galtier, 2009). gBGC is a mechanism associated with recombination: Mismatches formed during recombination at heterozygote sites are preferentially repaired in favor of GC bases in alleles (Marais, 2003). The population dynamics of this process is similar to selection (Nagylaki, 1983) and contributes to GC enrichment in highly recombining regions. GC content heterogeneity would thus be the result of variations in recombination along genomes. Direct evidence of gBGC (biased segregation at meiosis) exists in yeast (Mancera et al., 2008), and strong indirect evidence has been found in mammals (Duret and Arndt, 2008) and birds (Webster et al., 2006; Nabholz et al., 2011), mainly through the correlation between recombination and GC content dynamics. In plants, correlations between recombination and GC content have also been found in grasses (Haudry et al., 2008; Escobar et al., 2010; Muyle et al., 2011), in agreement with the gBGC hypothesis. In addition, in grasses, selfing species showed lower GC content and GC enrichment than outcrossing. This is also expected under the gBGC hypothesis because gBGC is expected to be weak or absent in mainly homozygote selfing genomes (Marais et al., 2004; Glémin et al., 2006; Haudry et al., 2008; Muyle et al., 2011).

Although our results do not directly prove the role of gBGC, they are mainly compatible with this hypothesis: Differences in the occurrence, intensity, and patterns of gBGC could explain the variations in nucleotide landscapes across seed plants. gBGC acts locally disregarding the nucleotide position type and can thus explain variations at all positions (Figure 2) and the local correlations between positions (Figure 6). As predicted by the gBGC hypothesis, we also found a positive correlation between local recombination rate and GC content in three grass species with GC-rich and heterogeneous genomes (Figure 8; see Supplemental Figure 6 and Supplemental Table 3 online), whereas there is no such a correlation in the GC-poor and more homogeneous genome of *Arabidopsis* (Marais et al., 2004; Giraut et al., 2011). Moreover, in this three grass species and in all species studied so far, recombination is heterogeneous along genomes, including chromosomic gradients and/or local recombination hotspots (e.g., Akhunov et al., 2003; McVean et al., 2004; Drouaud et al., 2006; Mancera et al., 2008; Gore et al., 2009; Rockman and Kruglyak, 2009; Saintenac et al., 2009). Species with strong gBGC are thus expected to exhibit both GC richer and more heterogeneous genomes.

Beyond the case study of the three grasses, the strong positive correlation between mean and variance in GC content we observed perfectly matches with this prediction. Such a correlation was also observed in mammals for which gBGC is supposed to play a central role in isochore evolution (Romiguier et al., 2010). Other observations can also be interpreted under the gBGC hypothesis, though the underlying processes are more speculative. As in mammals (Duret et al., 1995), we observed that GC-rich genes are shorter than GC-poor ones. This correlation could emerge if recombination drives gene compaction, as it was proposed for mammalian isochores (Montoya-Burgos et al., 2003). Finally, we found a strong relationship between GC richness and the steepness of the 5′-3′ gradient in the GC content along genes (Figure 7). Such a gradient was already

observed in grasses and banana and much more weakly in *Arabidopsis* (Wong et al., 2002; Tatarinova et al., 2010). Our results suggest that such a gradient seems to be quite universal in plants and directly linked to GC richness. In rice, this gradient extends to introns (Wong et al., 2002), and in yeast, which has both AT- and GC-ending preferred codons, it was shown to be a GC gradient, not a preferred codon gradient (Stoletzki, 2011). In yeast, 5′-3′ recombination gradients have been documented (Detloff et al., 1992), and the recombination often initiates within gene promoters (Baudat and Nicolas, 1997; Mancera et al., 2008). Stoletzki (2011) proposed this can explain the 5′-3′ GC gradient found in yeast genes.

Very few data are currently available in plants, but a 5′-3′ gene conversion gradient at the highly recombining bronze locus in maize has been suggested and debated (Dooner and Martínez-Férez, 1997; Thijs and Heyting, 1998). If intragenic recombination gradients also occur in plants, with recombination initiation concentrated in 5′ regions, it would explain the decreasing GC gradients we found along transcripts. If so, for a given GC gradient, the distribution of transcript length would thus partly control the distribution of GC content. More detailed analyses of local recombination patterns in plants will be necessary in the future to tackle this issue. Alternatively, gBGC could also be coupled with the hypothesis of Wong et al. (2002) involving the TCR mechanism (see above). TCR can occur through the base excision repair pathway (Svejstrup, 2002), which is known to be GC biased and proposed to be involved in the gBGC mechanism (Marais, 2003). gBGC could thus also generate the GC gradient through the TCR mechanism.

Other Causes?

Beyond the three hypotheses discussed above, alternative, still unknown, mechanisms could also be involved. For instance, Jiang et al. (2011) recently proposed that Pack-Mule transposable elements could contribute to GC content heterogeneity and GC gradient along genes in grasses by preferentially integrating GC-rich genes or gene fragments and inserting preferentially into the 5′ end of genes, sometimes evolving into additional exons at the 5′ end of genes. The authors suggested this reinforcing mechanism might have operated on initially heterogeneous genomes created by other causes. Finally, some form of selection on GC content at the transcript level in relation to gene functional classes or to the regulation of gene expression has been proposed and remains to be explored (Tatarinova et al., 2010).

Why Do Patterns Differ across the Phylogenetic Groups?

We suggested that gBGC should play a significant role in shaping nucleotide landscapes in plant genomes. However, what can explain, for instance, the huge differences between commelinids, especially grasses, and early-diverged eudicots? So far, this remains a fully open question, and we can only propose some directions for future work. Under gBGC, the emergence of more contrasted recombination landscapes and/or the increase in gBGC intensity would lead to GC richer and more heterogeneous genomes. Recombination is generally stronger and more variable in plants than in animals (Gaut et al., 2007). For example, several grass species show strong recombination gradients along chromosomes, mainly from low recombining centromeric regions to highly recombining telomeric regions (Gore et al., 2009; Saintenac et al., 2009; Huo et al., 2011). In large maize and wheat genomes, recombination also seems to be concentrated in gene-rich regions, with large noncoding transposon-rich regions contributing little to genetic length (Fu et al., 2002; Saintenac et al., 2011). In agreement with previous studies (Escobar et al., 2010; Muyle et al., 2011), we showed that recombination correlates to GC content in three grass species (Figure 8). In these species, and likely in other grasses, recombination heterogeneities at different genomic scales could contribute to their specific GC content. Though GC content patterns are sharply different, recombination gradients and hot spots were also found in *Arabidopsis*, which does not seem to be very different from grasses in this respect (Drouaud et al., 2006). However, recombination patterns must be relatively stable on the long term so that sufficient time allows heterogeneity in GC content to build up. For instance, frequent chromosome rearrangements were invoked to explain the homogenization of GC content in rat and mouse genome (Romiguier et al., 2010). The reconstruction of ancestral karyotypes suggests that Poaceae experienced less chromosome rearrangements than eudicots (Salse et al., 2009), which could contribute to explain its specific patterns. However, detailed features on the evolution of recombination landscapes among plants are still poorly known and will be clearly needed to test the role of recombination in building up plant nucleotide landscapes.

Changes in the intensity of the mismatch correction bias in favor of GC could also directly impact gBGC strength and, hence, GC patterns. gBGC may have evolved as a response to mutation bias toward AT (Birdsell, 2002): Higher mutation bias could strengthen the intensity of gBGC and, hence, paradoxically increases GC content. Among vertebrates, the appearance of GC-rich isochores in amniotes has been related to an increase in the level of CpG methylation leading to an increase in the mutation rate of methylated cytosines toward thymine (Duret et al., 2006). As mentioned above, the methylation level is much higher in rice than in poplar and *Arabidopsis* (Feng et al., 2010). In particular, methylation is higher in genes and genic regions, which is a key point to the evolution of gBGC: A higher mutation rate toward AT in these regions would be likely deleterious and could select a stronger gBGC to compensate the induced load (for theoretical arguments, see Glémin, 2010). On the contrary, such an effect is not expected if the increase in CpG methylation occurs in noncoding repetitive regions, such as observed in poplar (Feng et al., 2010). Data on methylation patterns are still too scarce in plants, but we think this hypothesis is worth investigating.

Conclusion

Detailed analyses of genomic patterns in model species is of fundamental importance in comparative and evolutionary genomics. However, the number of model species is still restrictive. We showed that broadening analyses to a much larger phylogenetic scale would shed new light on the evolution of nucleotide landscapes in plants. Instead of the classical monocot/dicot dichotomy, we proposed a much more continuous view of the

evolution of nucleotide landscapes in seed plants and suggested that GC content enrichment occurred several times independently from ancestral GC-poor and homogeneous genomes. This continuous view also supports the view of common evolutionary causes, and we suggest that gBGC might have played a central role in shaping nucleotide landscapes in plants, as it likely does in vertebrates and other species. Strong support for gBGC has already been obtained in rice (Muyle et al., 2011), and it will be worth extending such analyses to other plant species, especially in the key groups found in our analyses (e.g., nonpoaceae commelinids and Myrtales). However, other, still unrecognized, causes also could be explored, for instance, in relation with transcription and transposition mechanisms. Emergence of GC3 bimodal distributions in the most GC-rich genomes also remains to be explained. Is it the direct result of the combination of recombination patterns and gene distribution along the genome, or is it due to other specific, possibly functional, causes? Finally, if the role of gBGC is confirmed, it will be necessary to take it into account in the study of plant genome evolution as it can have a deep impact on many analyses such as the detection of selection signatures (Galtier and Duret, 2007; Berglund et al., 2009; Galtier et al., 2009; Ratnakumar et al., 2010).

METHODS

Sequence Data Sets

We retrieved all EST unigene data sets from three public EST databases: the PlantGDB versions 157a to 171a (http://www.plantgdb.org/), The Gene Index Project (http://compbio.dfci.harvard.edu/tgi/) releases 1 to 19, and The Institute for Genomic Research Plant Transcript Assemblies releases 1 to 5 (Childs et al., 2007). To increase the phylogenetic coverage, we also used raw plant EST data sets available in GenBank (http://www.ncbi.nlm.nih.gov/genbank/). We did not use all the data sets available, but we specifically focused on underrepresented and key groups, such as gymnosperms, used here as outgroups, basal angiosperms, and monocotyledons, and we tried to cover most orders within eudicotyledons. We assembled the EST sequence data sets retrieved from GenBank with the EST analysis pipeline EST2uni (Forment et al., 2008). After filtering (detailed below), we obtained 16 gymnosperms, six basal angiosperms, 56 monocots, and 154 eudicots. A total of 115 species came from the PlantGDB, 24 from the Gene Index Project, and 46 from The Institute for Genomic Research Plant Transcript Assemblies, and 47 were directly assembled from GenBank.

As a control, we used the complete genome sequences of three eudicots, *Arabidopsis thaliana* (TAIR9), *Populus trichocarpa* (JGI2.0), and grape (*Vitis vinifera*) (IGGP_12x), and four monocots, *Brachypodium distachyon* (Brachy1.0), rice (*Oryza sativa*) (MSU6), sorghum (*Sorghum bicolor*) (Sbi1), and maize (*Zea mays*) (AGPc2). They were retrieved from EnsemblPlants release 5 via BioMart (http://plants.ensembl.org). CDS and UTR regions were retrieved according to the annotations given in EnsemblPlants.

Phylogenetic Representation

To plot the species into a general phylogenetic context, we used the National Center for Biotechnology Information (NCBI) taxonomic tree. Though it leaves some relationships unresolved at low taxonomic levels, it gives the phylogenetic relationships among major clades that are strongly supported, and it is in agreement with the ordinal phylogeny of the Angiosperm Phylogeny Group (Bremer et al., 2009; Chase and Reveal, 2009). The tree was plotted with the phylogenetic display and manipulation online tool Interactive Tree of Life version 2 (Letunic and Bork, 2011)

EST Annotation and Filtering

All unigenes sequences were then processed through the EST protein translation pipeline prot4EST version 2.2 (Wasmuth and Blaxter, 2004) to determine codon positions and UTRs. The BLASTX search step of prot4EST analysis was performed against the UniProtKB/Swiss-Prot plant database (http://www.uniprot.org/). CDSs with frame shifts, abnormal stop codons, or a length under 100 nucleotides were removed from the data sets. The 232 final data sets for CDSs were all composed of at least 1000 EST unigenes, for a global amount of $3.4 \cdot 10^6$ EST unigenes. After filtering to remove UTRs with a length under 30 nucleotides and to retain only species having at least 1000 UTR sequences, the final data sets were composed of 185 species for 5′UTR and 197 species for 3′UTR. The number of raw EST sequences matching with each EST unigene sequence was recorded as a proxy of expression. We applied the same filters to complete genome sequences: UTRs with a length under 30 nucleotides and CDSs with a length under 100 nucleotides were removed.

Data Analyses

Descriptive Statistics

We used homemade Perl scripts to compute GC content in the first codon position (GC1), the second codon position (GC2), the third codon position (GC3), in 3′UTR (GC3UTR), and in 5′UTR (GC5UTR). For each species, we computed the mean and the SD of each GC category and the Spearman's correlation coefficient (a nonparametric rank correlation coefficient measuring the degree of dependence between two variables) of GC content between all pairs of positions, using the R package (R Development Core Team, 2011).

Fit of the GC Content Distributions

To capture the heterogeneity of GC content distribution, we fitted a bi-Beta distribution that is the mix of two Beta distributions in proportions p and $1-p$, and with the density function given by Equation 1. We chose the Beta distribution because it allows a large flexibility of shape for distribution ranging from 0 to 1. To get a better fit of the data, we could extend this rationale to the mix of multiple Beta distributions. However, the use of two distributions is sufficient to obtain relatively good fits and to give results with the simple biological meaning of two classes of genes regarding GC content, as it has already been proposed (Carels and Bernardi, 2000).

We used a maximum likelihood procedure to estimate the five parameters of the bi-Beta distribution. Because EST correspond to gene fragments, the number of G and C nucleotides of the i^{th} unigene, k_i, follows a hypergeometric distribution with parameters x_i the GC content of the full CDS, N_i the total length of the CDS, and n_i the length of the unigene. If $n_i = N_i$, x_i is perfectly known and equals k_i/n_i, otherwise, the sampling variance should be taken into account, short ESTs being less informative than longer ones. Unfortunately, N_i is not known (except for full CDS) so that the hypergeometric sampling variance cannot be incorporated into the likelihood function. As an approximation, we assumed that k_i follows a binomial distribution with parameter n_i and x_i, x_i following a bi-Beta distribution as described above. The likelihood of observing k_i for the i^{th} EST is given by:

$$l_i = \int_0^1 C_{n_i}^{k_i} x_i^{k_i} (1 - x_i)^{n_i - k_i} f(x_i, \alpha_1, \beta_1, \alpha_2, \beta_2, p) dx_i \quad (2)$$

where C is the binomial coefficient. To facilitate further numerical computations, Equation 2 can be written analytically as:

$$l_i = C_{n_i}^{k_i}\left(p\frac{\Gamma(\alpha_1+k_i)\Gamma(\beta_1+n_i-k_i)}{B(\alpha_1,\beta_1)\Gamma(\alpha_1+\beta_1+n_i)}+(1-p)\frac{\Gamma(\alpha_2+k_i)\Gamma(\beta_2+n_i-k_i)}{B(\alpha_2,\beta_2)\Gamma(\alpha_2+\beta_2+n_i)}\right) \quad (3)$$

where Γ is the gamma function (Abramowitz and Stegun, 1970). Assuming independence between unigenes, the likelihood of the full data for a given species is thus:

$$L = \prod_{i=1}^{N} l_i \quad (4)$$

where N is the total number of unigenes. We maximized the log likelihood function using the optim function of the R package with the option BFGS (Broyden-Fletcher-Goldfarb-Shanno method) (R Development Core Team, 2011). We checked for convergence using different initial values and inspecting by eye the fitted distribution. For each species, we then computed the mean of each Beta distribution from the α and β parameters given by:

$$\mu_x = \frac{\alpha_x}{\alpha_x+\beta_x},\ x = 1,2 \quad (5)$$

We then used the fitted distribution to obtain estimates of the mean and the SD of GC content distribution that (partly) correct for sampling variance.

$$\mu_{GC} = \int_0^1 xf(x,\alpha_1,\beta_1,\alpha_2,\beta_2,p)dx = p\frac{\alpha_1}{\alpha_1+\beta_1}+(1-p)\frac{\alpha_2}{\alpha_2+\beta_2} \quad (6A)$$

$$\sigma_{GC} = \left(\int_0^1 x^2 f(x,\alpha_1,\beta_1,\alpha_2,\beta_2,p)dx - {\mu^2}_{GC}\right)^{\frac{1}{2}} \quad (6B)$$

$$= \left(p\frac{\alpha_1(1+\alpha_1)}{(\alpha_1+\beta_1)(1+\alpha_1+\beta_1)}+(1-p)\frac{\alpha_2(1+\alpha_2)}{(\alpha_2+\beta_2)(1+\alpha_2+\beta_2)} - \left(p\frac{\alpha_1}{\alpha_1+\beta_1}+(1-p)\frac{\alpha_2}{\alpha_2+\beta_2}\right)^2\right)^{1/2}$$

Expressions 6A and 6B were compared with the mean and variance directly computed from the raw distributions to check the robustness of data.

Difference in Transcript Length

We tested for the difference in mean transcript length between the GC3-rich and the GC3-poor classes of genes. We only retained transcripts assumed to be complete, that is, when they began with the start codon ATG and when their 5′UTR and 3′UTR were more than 19 nucleotides long. We used the GC3 median to define the GC3-rich and the GC3-poor gene classes. We then filtered the species data set and retained species having at least 500 complete transcripts in the GC3-rich class and 500 in the GC3-poor class. For each of the 78 remaining species (four gymnosperms, 64 eudicots, and 10 monocots, all Poaceae), we performed nonparametric Kruskal-Wallis tests to compare coding length of transcripts between the GC3-rich class and the GC3-poor class.

GC Gradients along Transcripts

Gradients of GC content along transcripts were measured by the linear regression slope of GC content against the distance to the starting codon. We only kept CDSs longer than 600 nucleotides with a confident start codon (defined as beginning with the starting codon ATG and having a 5′UTR >19 nucleotides long). We removed species data sets having <1000 transcripts meeting the criteria. There remained 81 species, including four gymnosperms, 65 eudicots, and 12 monocots that were all Poaceae. The coding regions of all transcripts were aligned from the start codon, and the mean GC content was computed for each nucleotide position. After visual inspection of the data, the GC content of the first nucleotides appeared messy and the number of sequences longer than 600 bp drastically decreased. We thus performed the linear regressions from positions 60 to 600. We computed the slope separately for positions 1, 2, and 3.

GC Correlations with Expression

EST singletons (i.e., EST with an expression level equal to 1) were first removed from data sets. To retain species with enough expression variance, the species data set was then filtered to remove species having a maximal expression value under 20 and an expression SD under 2. For each of the 171 remaining species (14 gymnosperms, six basal angiosperms, 28 monocots, and 94 eudicots), Spearman's correlation tests were performed to test the correlation between expression levels and GC3. Differences of Spearman's rho between species groups were tested with Kruskal-Wallis tests.

Correlation between Recombination and GC Content

To estimate local recombination rate, we built a genetic versus physical distances map (Marey's map) using available genetic maps for which markers have been physically mapped on genome. For rice, we used the 1202 markers used by Muyle et al. (2011), corresponding to a cleaned subset of those available at the Rice Genome Program website (http://rgp.dna.affrc.go.jp/E/publicdata/geneticmap2000/index.html). For maize, we used the "skeleton" markers used by Liu et al.(2009). We only kept the 1366 markers with consistent mapping that is when genetic and physical positions agree. For *B. distachyon*, we used the 558 markers developed by Huo et al. (2011). All markers data are available in Supplemental Data Set 3 online. Recombination rates were computed with the MareyMap program (Rezvoy et al., 2007). As in Muyle et al. (2011), we used the loess function that locally adjusts a polynomial second degree curve, using a weight attributed to each marker depending on how far it is from the center of the window (Rezvoy et al., 2007). We used windows containing 20% of the total number of markers to get rather smooth recombination rate curves. For the three species, we retrieved the longest transcript for each gene, and we estimated its local recombination rate using the center of the pretranscript as coordinate. For maize and rice, we used either all the annotated transcripts or the protein coding transcripts only. In maize, we simply selected the "protein_coding" gene biotype using Biomart, and we thus excluded transposable element, pseudogene, and miRNA. In rice, the biotype entry is noted as "protein_coding" for all genes. We thus used gene description to exclude transposable element, pseudogene, and miRNA. For *B. distachyon*, all genes are considered as "protein_coding," and no description is easily available via Biomart. We thus only used the complete data set.

Comparison between Taxonomic Groups and Control for Taxonomy

To explore phylogenetic variations in GC content distribution, we grouped species according to the NCBI taxonomy. We used six levels starting from the genus level. For the genus, the family, and the order levels, all groups used are monophyletic according to NCBI taxonomy. Beyond the order level, we used some paraphyletic groups (e.g., early-diverged core eudicots at the same level as Rosids and Asterids) to avoid multiplying the number of singleton groups. We estimated the variance components at each taxonomic level by fitting a linear model with the six nested taxonomic levels as random effects using the *lme* function of the R package and then extracting the variance components with the *varcomp* function of the R package (R Development Core Team, 2011). GC content characteristics between groups were also compared with nonparametric Kruskal-Wallis tests using the R package (R Development Core Team, 2011).

We also correlated the mean GC3 with the other GC content characteristics between species (e.g., mean GC3 versus SD) using nonparametric Spearman's correlations using the R package (R Development Core Team, 2011). To our knowledge, there is no phylogeny with branch lengths available for the set of species we studied, and some nodes are not well resolved. We thus did not use a tree-based method for phylogenetic control. Instead, we tested the robustness of the observed correlations through several controls using six nested taxonomic levels: level 1, group; level 2, infra-group; level 3, super-order; level 4, order; level 5, family; level 6, genus. Levels 1 to 3 are based on the phylogenetic relationships given by APG III for angiosperms (Bremer et al., 2009; Chase and Reveal, 2009) and by Burleigh and Mathews (2004) for gymnosperms, which are supposed to be monophyletic. At these levels, we used paraphyletic groups to avoid numerous singletons. Their names are given for practical reasons here and do not correspond to classical taxonomic levels. The composition of these three levels is given in Supplemental Table 4 online. Levels 4 to 6 are the classical ones.

First Control

In addition to nonparametric correlation, we performed mixed linear models including mean GC3 as a fixed effect and the six nested taxonomic levels as random effects using the *lme* function of the R package (R Development Core Team, 2011). We thus fitted the following model:

$$SD(GC3)_{ijklmno} = \mu + mean(GC3)_{ijklmno} + Group_i + Infra_Group_{ij} + Super_Order_{ijk} + Order_{ijkl} + Family_{ijklm} + Genus_{ijklmn} + \varepsilon_{ijklmno}$$

where mean(GC3) was assumed to be a fixed effect, and nested taxonomic levels were assumed to be random effects. The fixed effect was still highly significant after taxonomy control:

$$\text{Without control: t-value} = 35.72, \text{df} = 230, \text{p-value} < 10^{-15}$$

$$\text{With control: t-value} = 19.88, \text{df} = 74, \text{p-value} < 10^{-15}$$

Similarly, we tested the effect of taxonomy on the relationship between the global mean GC3 and the means of the two Beta distributions and the proportion of the two distributions. Once again, the fixed effects were still highly significant after taxonomy control:

Mean 1:

$$\text{Without control: t-value} = 30.81, \text{df} = 230, \text{p-value} < 10^{-15}$$

$$\text{With control: t-value} = 20.02, \text{df} = 74, \text{p-value} < 10^{-15}$$

Mean 2:

$$\text{Without control: t-value} = 35.37, \text{df} = 230, \text{p-value} < 10^{-15}$$

$$\text{With control: t-value} = 19.12, \text{df} = 74, \text{p-value} < 10^{-15}$$

Proportion:

$$\text{Without control: t-value} = 5.26, \text{df} = 230, \text{p-value} = 3.37\ 10^{-7}$$

$$\text{With control: t-value} = 7.17, \text{df} = 74, \text{p-value} = 9.62\ 10^{-7}$$

Second Control

We redid the nonparametric Spearman's correlation analyses recursively at every taxonomic level for groups containing at least four species. Results are summarized in the Supplemental Table 5 online. In most cases, the correlation is positive significant when there are enough points.

Third Control

We performed Spearman's correlations on averages computed recursively at every taxonomic level using the R environment (R Development Core Team, 2011). We also redid the nonparametric Spearman's correlation on averages computed recursively at every taxonomic level. We found a strong positive correlation at all levels, as shown in Supplemental Figure 7 online. Using the same averaging procedure, we also verified that the increased in the two means of the bi-Beta distribution with the total mean GC3 (shown in Figure 4) also held true at all taxonomic levels (see Supplemental Figure 8 online)

Accession Numbers

Sequence data from this article can be found in the GenBank/EMBL data libraries under the accession numbers given in Supplemental Data Set 4 online for EST data and Supplemental Data Set 5 online for complete genome data. These data are publicly available at http://datadryad.org/ under the DOI http://dx.doi.org/10.5061/dryad.p12305b2. Genome sequences used were *Arabidopsis* (TAIR9), *P. trichocarpa* (JGI2.0), grape (IGGP_12x), *B. distachyon* (Brachy1.0), rice (MSU6), sorghum (Sbi1), and maize (AGPc2), as described above.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Comparisons of GC3 Distribution Computed with Raw EST Data, Bi-Beta Fits, and Complete Genome Sequences for Six Species.

Supplemental Figure 2. Comparison of the Means and SDs Estimated from the Raw Data and from the Bi-Beta Distributions.

Supplemental Figure 3. Relationship between the Residuals of the Regression of GC3 SD against Mean GC3 (Bi-Beta Estimates) and the Number of EST Sequences and Unigenes.

Supplemental Figure 4. Box Plot of Mean GC3 by Super Order and Order.

Supplemental Figure 5. Box Plot of GC3 SD by Super Order and Order.

Supplemental Figure 6. Relationship between Local Recombination Rate and GC1, GC2, GC5UTR, and GC3UTR in Rice, Maize, and *Brachypodium distachyon*.

Supplemental Figure 7. Correlation between Mean GC3 and GC3 Standard Variation Averaged by Taxonomic Level

Supplemental Figure 8. Correlation between Mean GC3 and the Two Means of the Bi-Beta Distribution Averaged by Taxonomic Level.

Supplemental Table 1. Mean GC Content by Taxonomic Group and Kruskal-Wallis Test for Differences between Groups.

Supplemental Table 2. GC Content SDs by Taxonomic Group and Kruskal-Wallis Tests for Differences between Groups.

Supplemental Table 3. Nonparametric Spearman's Correlation between Local Recombination Rate and GC1, GC2, GC3, GC5UTR, and GC3UTR in Rice, Maize, and *Brachypodium distachyon*.

Supplemental Table 4. Taxonomy Levels Used in the Analyses.

Supplemental Table 5. Correlation between Mean GC3 and GC3 Standard Variation by Taxonomic Group

Supplemental Data Set 1. GC Statistics for All Species.

Supplemental Data Set 2. Comparison of GC Statistics between Raw EST Data, Bi-Beta Fits, and Complete Transcriptome Data.

Supplemental Data Set 3. List of Markers Used for Marey's Maps in Rice, Maize, and *Brachypodium distachyon*.

Supplemental Data Set 4. List of EMBL/GenBank Accession Numbers for the 47 Species for Which Unigenes Were Directly Assembled from EST Sequences.

Supplemental Data Set 5. List of EMBL/GenBank Accession Numbers for the Seven Complete Genome Data.

ACKNOWLEDGMENTS

We thank Adrienne Ressayre, Juan Escobar, and three anonymous reviewers for their helpful comments on the article and Nicolas Galtier, Benoit Nabholz, Gabriel Marais, and Sylvain Mousset for their insightful discussions and remarks on this work. This publication is contribution ISEM-2012-020 of the Institut des Sciences de l'Evolution de Montpellier (Unité Mixte de Recherche 5554, Centre National de la Recherche Scientifique). This work was supported by Agropolis Resource Centre for Crop Conservation, Adaptation and Diversity, a flagship project of Agropolis Foundation and by the Centre National de la Recherche Scientifique and Agence Nationale de la Recherche (ANR-08-GENM-036-01).

AUTHOR CONTRIBUTIONS

L.S.-G. performed research, analyzed data, and wrote the article. K.B. contributed to bioinformatics analyses. J.D. designed research and contributed to data analyses. S.G. designed research, analyzed data, and wrote the article.

Received November 13, 2011; revised February 5, 2012; accepted March 16, 2012; published April 6, 2012.

REFERENCES

Abramowitz, M., and Stegun, I.A. (1970). Handbook of Mathematical Functions. (New York: Dover).

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., and Bernardi, G. (1991). The compositional properties of human genes. J. Mol. Evol. **32:** 493–503.

Akashi, H. (2001). Gene expression and molecular evolution. Curr. Opin. Genet. Dev. **11:** 660–666.

Akhunov, E.D., et al. (2003). The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res. **13:** 753–763.

Barakat, A., Matassi, G., and Bernardi, G. (1998). Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. Proc. Natl. Acad. Sci. USA **95:** 10044–10049.

Baudat, F., and Nicolas, A. (1997). Clustering of meiotic double-strand breaks on yeast chromosome III. Proc. Natl. Acad. Sci. USA **94:** 5213–5218.

Berglund, J., Pollard, K.S., and Webster, M.T. (2009). Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. **7:** e26.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science **228:** 953–958.

Birdsell, J.A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. **19:** 1181–1197.

Bremer, B., et al. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. **161:** 105–121.

Burleigh, J.G., and Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. Am. J. Bot. **91:** 1599–1613.

Carels, N., and Bernardi, G. (2000). Two classes of genes in plants. Genetics **154:** 1819–1825.

Chase, M.W., and Reveal, J.L. (2009). A phylogenetic classification of the land plants to accompany APG III. Bot. J. Linn. Soc. **161:** 122–127.

Childs, K.L., Hamilton, J.P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P.D., Town, C.D., Buell, C.R., and Chan, A.P. (2007). The TIGR Plant Transcript Assemblies database. Nucleic Acids Res. **35**(Database issue): D846–D851.

Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D., and Bernardi, G. (1996). Human coding and noncoding DNA: Compositional correlations. Mol. Phylogenet. Evol. **5:** 2–12.

Costantini, M., and Bernardi, G. (2008). Replication timing, chromosomal bands, and isochores. Proc. Natl. Acad. Sci. USA **105:** 3433–3437.

Detloff, P., White, M.A., and Petes, T.D. (1992). Analysis of a gene conversion gradient at the HIS4 locus in *Saccharomyces cerevisiae*. Genetics **132:** 113–123.

Dooner, H.K., and Martínez-Férez, I.M. (1997). Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. Plant Cell **9:** 1633–1646.

Drouaud, J., Camilleri, C., Bourguignon, P.Y., Canaguier, A., Bérard, A., Vezon, D., Giancola, S., Brunel, D., Colot, V., Prum, B., Quesneville, H., and Mézard, C. (2006). Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". Genome Res. **16:** 106–114.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12:** 640–649.

Duret, L., and Arndt, P.F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. **4:** e1000071.

Duret, L., Eyre-Walker, A., and Galtier, N. (2006). A new perspective on isochore evolution. Gene **385:** 71–74.

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. **10:** 285–311.

Duret, L., and Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

Duret, L., Mouchiroud, D., and Gautier, C. (1995). Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40:** 308–317.

Elhaik, E., Landan, G., and Graur, D. (2009). Can GC content at third-codon positions be used as a proxy for isochore composition? Mol. Biol. Evol. **26:** 1829–1833.

Escobar, J.S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J., and Glémin, S. (2010). An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). Evolution **64:** 2855–2872.

Escobar, J.S., Glémin, S., and Galtier, N. (2011). GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. Mol. Biol. Evol. **28:** 2561–2575.

Eyre-Walker, A., and Hurst, L.D. (2001). The evolution of isochores. Nat. Rev. Genet. **2:** 549–555.

Feng, S., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. USA **107:** 8689–8694.

Forment, J., Gilabert, F., Robles, A., Conejero, V., Nuez, F., and Blanca, J.M. (2008). EST2uni: An open, parallel tool for automated

EST analysis and database creation, with a data mining web interface and microarray expression data integration. BMC Bioinformatics **9:** 5.

Fryxell, K.J., and Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol. **17:** 1371–1383.

Fu, H., Zheng, Z., and Dooner, H.K. (2002). Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc. Natl. Acad. Sci. USA **99:** 1082–1087.

Fullerton, S.M., Bernardo Carvalho, A., and Clark, A.G. (2001). Local rates of recombination are positively correlated with GC content in the human genome. Mol. Biol. Evol. **18:** 1139–1142.

Galtier, N., and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. **23:** 273–277.

Galtier, N., Duret, L., Glémin, S., and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. **25:** 1–5.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. Genetics **159:** 907–911.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J., and Anderson, L.K. (2007). Recombination: An underappreciated factor in the evolution of plant genomes. Nat. Rev. Genet. **8:** 77–84.

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O.C., and Mézard, C. (2011). Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. PLoS Genet. **7:** e1002354.

Glémin, S. (2010). Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics **185:** 939–959.

Glémin, S., Bazin, E., and Charlesworth, D. (2006). Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc. Biol. Sci. **273:** 3011–3019.

Gore, M.A., Chia, J.M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., Ware, D.H., and Buckler, E.S. (2009). A first-generation haplotype map of maize. Science **326:** 1115–1117.

Guo, X., Bao, J., and Fan, L. (2007). Evidence of selectively driven codon usage in rice: Implications for GC content evolution of Gramineae genes. FEBS Lett. **581:** 1015–1021.

Haudry, A., Cenci, A., Guilhaumon, C., Paux, E., Poirier, S., Santoni, S., David, J., and Glémin, S. (2008). Mating system and recombination affect molecular evolution in four Triticeae species. Genet Res (Camb) **90:** 97–109.

Hughes, S., and Mouchiroud, D. (2001). High evolutionary rates in nuclear genes of squamates. J. Mol. Evol. **53:** 70–76.

Huo, N., Garvin, D.F., You, F.M., McMahon, S., Luo, M.C., Gu, Y.Q., Lazo, G.R., and Vogel, J.P. (2011). Comparison of a high-density genetic linkage map to genome features in the model grass *Brachypodium distachyon*. Theor. Appl. Genet. **123:** 455–464.

Jiang, N., Ferguson, A.A., Slotkin, R.K., and Lisch, D. (2011). Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. Proc. Natl. Acad. Sci. USA **108:** 1537–1542.

Katzman, S., Capra, J.A., Haussler, D., and Pollard, K.S. (2011). Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. Genome Biol. Evol. **3:** 614–626.

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., and Zylicz, M. (2006). High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. **4:** e180.

Kuhl, J.C., et al. (2004). A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders Asparagales and Poales. Plant Cell **16:** 114–125.

Lander, E.S., et al; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature **409:** 860–921. Erratum. Nature **411:** 720.

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. **39**(Web Server issue): W475–W478.

Liu, S., Yeh, C.T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D., and Schnable, P.S. (2009). Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. **5:** e1000733.

Lynch, M. (2007). The Origin of Genome Architecture. (Sunderland, MA: Sinauer).

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L.M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature **454:** 479–485.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. Trends Genet. **19:** 330–338.

Marais, G., Charlesworth, B., and Wright, S.I. (2004). Recombination and base composition: The case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. **5:** R45.

Matassi, G., Montero, L.M., Salinas, J., and Bernardi, G. (1989). The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Res. **17:** 5273–5290.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Meunier, J., and Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. **21:** 984–990.

Montoya-Burgos, J.I., Boursot, P., and Galtier, N. (2003). Recombination explains isochores in mammalian genomes. Trends Genet. **19:** 128–130.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., and Bernardi, G. (1991). The distribution of genes in the human genome. Gene **100:** 181–187.

Mukhopadhyay, P., Basak, S., and Ghosh, T.C. (2007). Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes. Gene **400:** 71–81.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glémin, S. (2011). GC-biased gene conversion and selection affect GC content in the Oryza genus (rice). Mol. Biol. Evol. **28:** 2695–2706.

Nabholz, B., Künstner, A., Wang, R., Jarvis, E.D., and Ellegren, H. (2011). Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol. Biol. Evol. **28:** 2197–2210.

Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. USA **80:** 6278–6281.

Qin, H., Wu, W.B., Comeron, J.M., Kreitman, M., and Li, W.H. (2004). Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics **168:** 2245–2260.

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M.T. (2010). Detecting positive selection within genomes: The problem of biased gene conversion. Philos. Trans. R. Soc. Lond. B Biol. Sci. **365:** 2571–2580.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. (Vienna, Austria: R Foundation for Statistical Computing).

Ren, X.Y., Vorst, O., Fiers, M.W., Stiekema, W.J., and Nap, J.P.

(2006). In plants, highly expressed genes are the least compact. Trends Genet. **22:** 528–532.

Rezvoy, C., Charif, D., Guéguen, L., and Marais, G.A. (2007). MareyMap: An R-based tool with graphical interface for estimating recombination rates. Bioinformatics **23:** 2188–2189.

Rockman, M.V., and Kruglyak, L. (2009). Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLoS Genet. **5:** e1000419.

Romiguier, J., Ranwez, V., Douzery, E.J., and Galtier, N. (2010). Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. **20:** 1001–1009.

Saintenac, C., Falque, M., Martin, O.C., Paux, E., Feuillet, C., and Sourdille, P. (2009). Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). Genetics **181:** 393–403.

Saintenac, C., Faure, S., Remay, A., Choulet, F., Ravel, C., Paux, E., Balfourier, F., Feuillet, C., and Sourdille, P. (2011). Variation in crossover rates across a 3-Mb contig of bread wheat (*Triticum aestivum*) reveals the presence of a meiotic recombination hotspot. Chromosoma **120:** 185–198.

Salse, J., Abrouk, M., Bolot, S., Guilhot, N., Courcelle, E., Faraut, T., Waugh, R., Close, T.J., Messing, J., and Feuillet, C. (2009). Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc. Natl. Acad. Sci. USA **106:** 14908–14913.

Shi, X., Wang, X., Li, Z., Zhu, Q., Tang, W., Ge, S., and Luo, J. (2006). Nucleotide substitution pattern in rice paralogues: Implication for negative correlation between the synonymous substitution rate and codon usage bias. Gene **376:** 199–206.

Smith, N.G., and Eyre-Walker, A. (2001). Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. **18:** 982–986.

Stoletzki, N. (2011). The surprising negative correlation of gene length and optimal codon use—Disentangling translational selection from GC-biased gene conversion in yeast. BMC Evol. Biol. **11:** 93.

Svejstrup, J.Q. (2002). Mechanisms of transcription-coupled DNA repair. Nat. Rev. Mol. Cell Biol. **3:** 21–29.

Tatarinova, T.V., Alexandrov, N.N., Bouck, J.B., and Feldmann, K.A. (2010). GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics **11:** 308.

Thijs, H., and Heyting, C. (1998). Polarity of meiotic recombination in the bronze locus of maize. Plant Cell **10:** 645–648.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J.L., Jackson, S.A., Gaut, B.S., and Ma, J. (2009). Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. **19:** 2221–2230.

Wang, H.C., and Hickey, D.A. (2007). Rapid divergence of codon usage patterns within the rice genome. BMC Evol. Biol. **7** (suppl. 1)**:** S6.

Wang, H.C., Singer, G.A., and Hickey, D.A. (2004). Mutational bias affects protein evolution in flowering plants. Mol. Biol. Evol. **21:** 90–96.

Wang, L.J., and Roossinck, M.J. (2006). Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. Plant Mol. Biol. **61:** 699–710.

Wasmuth, J.D., and Blaxter, M.L. (2004). prot4EST: Translating expressed sequence tags from neglected genomes. BMC Bioinformatics **5:** 187.

Webster, M.T., Axelsson, E., and Ellegren, H. (2006). Strong regional biases in nucleotide substitution in the chicken genome. Mol. Biol. Evol. **23:** 1203–1216.

Wolfe, K.H., Sharp, P.M., and Li, W.H. (1989). Mutation rates differ among regions of the mammalian genome. Nature **337:** 283–285.

Wong, G.K., Wang, J., Tao, L., Tan, J., Zhang, J., Passey, D.A., and Yu, J. (2002). Compositional gradients in Gramineae genes. Genome Res. **12:** 851–856.

Wright, S.I., Iorgovan, G., Misra, S., and Mokhtari, M. (2007). Neutral evolution of synonymous base composition in the Brassicaceae. J. Mol. Evol. **64:** 136–141.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science **328:** 916–919.

Chapitre 2

Évolution des préférences d'usage du code chez les plantes

Usage du code : biais, sélection traductionnelle et préférences

Le code génétique étant dégénéré, 18 acides aminés parmi les 20 existants peuvent être encodés par différents codons synonymes. La troisième base du codon est toujours synonyme, partiellement ou totalement, pour les acides aminés au code dégénéré. La première base est partiellement synonyme pour deux acides aminés. Chez de nombreux organismes les codons synonymes ne sont pas utilisés à fréquences égales, contrairement à ce qu'on attendrait *a priori* (Grantham et al. 1980a ; Grantham et al. 1980b). L'usage du code est biaisé en faveur de certains codons et au détriment d'autres. L'intensité de ce biais d'usage varie entre les espèces (Ikemura 1985 ; Sharp et al. 2005) et, au sein d'un même génome, selon les acides aminés et selon les gènes (Sharp et al. 1988). Les patrons d'usage du code (fréquences d'usage des codons et leurs variations à l'échelle des régions codantes du génome) sont spécifiques à chaque espèce. Selon la métaphore proposée par Ikemura (Ikemura 1985), le langage universel du code génétique est interprété au niveau des espèces en différents dialectes d'usage du code.

Les fréquences d'utilisation des codons synonymes permettent de distinguer les codons minoritaires et les codons majoritaires, ceux dont la fréquence diffère de celle attendue sous équiprobabilité. Les différences de fréquences entre codons minoritaires et majoritaires s'expliquent principalement par la composition nucléotidique des génomes (Kanaya et al. 2001 ; Lynn et al. 2002 ; Chen et al. 2004). Ainsi, dans un génome riche en GC, les codons majoritaires sont ceux finissant en G ou C. La composition génomique n'explique toutefois pas intégralement les patrons d'usage du code, en particulier les différences de biais d'usage observées entre les gènes peu exprimés et les gènes les plus exprimés (Kanaya et al. 2001).

Ces différences d'usage du code selon le niveau d'expression des gènes ont mené à la formulation de l'hypothèse de sélection sur l'efficacité de la traduction, selon laquelle certains codons permettent une traduction plus efficace et sont sélectionnés pour réduire le coût de la traduction pour l'organisme (pour revue Kurland 1991 ; Akashi 2001 ; Duret 2002 ; Hershberg & Petrov 2008).

L'efficacité de la traduction des codons dépend des ARNt présents dans la cellule. Pour un acide aminé donné, il y a généralement plus de codons synonymes que d'ARNt isoaccepteurs – les ARNt associés à cet acide aminé. Trois codons, TAA, TAG et TGA, ne sont ainsi reconnus par aucun ARNt, ce qui conduit à l'arrêt de la synthèse protéique. Tous les autres codons peuvent

néanmoins s'apparier à l'anticodon d'un ARNt, soit par appariement canonique de Watson-Crick (A•U et G•C), soit par « appariement bancal » (*wobble pairing*, principalement I•U, I•C et G•U). Une même espèce d'ARNt peut ainsi s'apparier avec jusqu'à trois codons différents. L'efficacité de la traduction d'un codon est liée à la nature de son interaction avec l'anticodon. Cette interaction est plus stable lorsque l'appariement est canonique. Les isoaccepteurs ne sont de plus pas présents en proportions égales dans la cellule. Les gènes codant pour les ARNt sont présents en multiples copies dans le génome et l'abondance des ARNt dans la cellule est déterminée par le nombre de ces copies. La cinétique de la traduction d'un codon est affectée par l'abondance des différents isoaccepteurs capables de s'apparier à ce codon. Parmi un groupe de codons synonymes, les codons dont la traduction est la plus efficace sont donc ceux qui s'apparient avec la plus grande force avec leur ARNt, ou qui s'apparient avec l'espèce d'ARNt la plus fréquente dans la cellule par rapport aux autres ARNt isoaccepteurs (Stoletzki & Eyre-Walker 2007). Ces codons dits codons préférés ou optimaux sont plus fréquents chez les gènes les plus exprimés : la pression de sélection sur l'efficacité de la traduction augmente avec le niveau d'expression des gènes. A l'inverse, les codons dits sous-optimaux, que nous dénommerons ici codons détestés, sont plus fréquents chez les gènes faiblement exprimés.

La sélection sur l'efficacité de la traduction, ou sélection sur l'usage du code (SUC), a été mise en évidence chez de nombreux procaryotes (premiers résultats chez *Escherichia coli* : Ikemura, 1981 ; exemples d'autres travaux : Lobry & Necşulea, 2006 ; Sharp et al., 2005 ; Stoletzki & Eyre-Walker, 2007) et chez des eucaryotes tels que la levure *Saccharomyces cerevisiae* (Bennetzen & Hall 1981), la drosophile (Shields et al. 1988), le nématode *Caenorhabditis elegans* (Stenico et al. 1994) et un vertébré, le xénope (Musto et al. 2001) – mais aucune trace de SUC n'a été trouvée chez les mammifères. Des arguments en faveur de la SUC ont été trouvés chez *Arabidopsis thaliana* (Chiapello et al. 1998 ; Morton & Wright 2007) puis chez d'autres plantes (Murray et al. 1989 ; Campbell & Gowri 1990 ; Wang & Roossinck 2006 ; Kawabe & Miyashita 2003), dont notamment les graminées maïs (Liu et al. 2010 ; Fennoy & Bailey-Serres 1993), blé (Zhang et al. 2007) et riz (Guo et al. 2007 ; Wang & Hickey 2007), et les eudicotylédones silène (Qiu et al. 2011a), capselles (Qiu et al. 2011b) et peupliers (Zhou et al. 2007 ; Ingvarsson 2008).

Si les codons minoritaires et majoritaires reflètent donc surtout la composition en base du génome, les codons détestés et préférés sont quant à eux ceux que pénalise ou favorise la SUC. Nous grouperons sous le terme de préférences d'usage du code les statuts (préféré, détesté) de l'ensemble des codons. Les codons détestés et préférés étant respectivement définis comme étant les codons dont la fréquence diminue ou augmente avec le niveau d'expression, ils peuvent être

déterminés en utilisant des données de séquences de gènes et leur niveau d'expression. Concrètement, la corrélation entre les fréquences des codons et les niveaux d'expression peut être directement utilisée pour déterminer les préférences d'usage du code (Tanguy et al. 2008). Une autre méthode proposée par Duret et Mouchiroud (Duret & Mouchiroud 1999) est basée sur une mesure du biais d'usage des codons, le RSCU (*Relative Synonymous Codon Usage* : fréquence d'usage d'un codon calculée relativement aux fréquences des autres codons synonymes de l'acide aminé pour lequel il code et pondérée par le degré de dégénérescence de l'acide aminé – voir Matériel et Méthodes). Cette méthode consiste à définir des classes de gènes selon leur niveau d'expression, à calculer pour chaque codon le RSCU moyen pour la classe de gènes la moins exprimée et pour la classe de gènes la plus exprimée, puis à utiliser la différence entre ces deux valeurs (ΔRSCU = RSCU expression forte – RSCU expression faible) pour déterminer les préférences (méthode reprise par exemple par Cutter et al 2008 ; Duret 2000 ; Ingvarsson 2008 ; Marais et al 2001 ; Wang & Roossinck 2006). Il est également possible de déterminer les préférences sans disposer de données d'expression, en profitant du fait que l'intensité de la sélection sur l'usage du code augmente avec le niveau d'expression. Selon l'intensité de leur biais d'usage du code, les gènes sont attribués à des classes d'expression. Les variations de fréquences des codons entre ces classes sont alors utilisées pour déterminer les préférences (par exemple Liu et al. 2010).

Evolution des préférences d'usage du code

Entre quatre espèces d'eucaryotes, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster*, et *Xenopus laevis* , les préférences d'usage du code sont différentes pour la plupart des codons (Kanaya et al. 2001). Les préférences d'usage du code varient également à large échelle entre bactéries, archées et champignons (Hershberg & Petrov 2009). Chez les plantes, une certaine conservation des préférences a été trouvée entre quelques graminées et eudicotylédones pour certains acides aminés alors que pour d'autres acides aminés ces préférences ont évolué (Kawabe & Miyashita 2003 ; Wang & Roossinck 2006). Les préférences d'usage du code évoluent donc et ont divergé entre ces espèces phylogénétiquement éloignées.

Quelques études apportent des éléments quant à l'évolution des préférences d'usage du code à plus faible échelle phylogénétique. Chez cinq espèces de nématodes (*Ancylostoma ceylanicum, Meloidogyne incognita, Onchocerca volvulus, Strongyloides ratti* et *Trichinella spiralis*) appartenant à des clades différents, seuls trois codons (AGC, CAC et TAC) sont toujours significativement préférés (Mitreva et al. 2006). L'usage global du code chez 32 espèces de nématode reste très similaire au niveau intragénérique puis perd rapidement en similarité aux niveaux taxonomiques supérieurs (Mitreva et al. 2006). Chez trois espèces de bivalves (l'huître *Crassostrea gigas*, la moule commune *Mytilus edulis* et la moule abyssale *Bathymodiolus azoricus*), seuls six codons sont systématiquement préférés (CTC, GCC, GTC, ATC, AAC et TTC) (Tanguy et al. 2008). Chez ces bivalves marins riches en AT (GC3 respectifs : 43%, 31%, 30%), les codons préférés sont souvent les codons finissant en G ou C : la SUC y favorise des codons minoritaires. Quant aux plantes, des études menées sur sept espèces de peupliers ont montré que leurs préférences d'usage du code sont déjà quelque peu variables à ce niveau intragénérique (Zhou et al. 2007 ; Ingvarsson 2008).

Forces évolutives façonnant l'usage du code

Les patrons d'usage du code sont façonnés par la sélection sur l'usage du code mais aussi par d'autres processus évolutifs, non sélectifs. La dérive génétique a des effets plus ou importants selon la taille efficace de la population. Son action n'est pas biaisée par la composition nucléotidique. Deux forces évolutives neutres sont susceptibles d'affecter directement la composition nucléotidique si elles présentent un biais en faveur de certains nucléotides : la mutation et la conversion génique. Chez la plupart des eucaryotes, la mutation semble être biaisée en faveur des nucléotides A et T (Lynch 2007). Ce biais mutationnel vers A/T a récemment été montré chez le riz (Muyle et al. 2011). La conversion génique est un transfert non réciproque d'information de séquence lors d'un évènement de recombinaison à un site hétérozygote : l'un des deux allèles présents est remplacé par l'autre. Elle résulte de la correction des mésappariements de l'hétéroduplex qui se forme lors de la recombinaison. Si la réparation se fait en prenant comme patron l'un ou l'autre des deux brins de manière équiprobable, la conversion génique n'a pas d'impact sur la composition nucléotidique. Mais il a été montré, chez les mammifères et les oiseaux notamment, l'existence d'un biais de réparation : c'est le plus souvent l'allèle en G ou C qui est pris comme patron et convertit l'allèle en A ou T (Duret & Galtier 2009). Lorsque cette conversion génique biaisée a lieu lors de la méiose, elle mène à une ségrégation non-Mendélienne des gamètes. En favorisant les allèles en G ou C au détriment de ceux en A ou T, la conversion génique biaisée vers G et C (BGC) mime la sélection (Nagylaki 1983). Les résultats du chapitre précédent ainsi que d'autres travaux (Glémin et al. 2006 ; Haudry et al. 2008 ; Escobar et al. 2010 ; Muyle et al. 2011) suggèrent la présence de BGC chez les plantes.

Sélection sur l'usage du code, dérive génétique, biais mutationnel et conversion génique biaisée vers G et C le cas échéant façonnent donc les patrons d'usage du code. L'usage du code d'un gène est ainsi le fruit d'un équilibre entre la SUC, qui favorise les codons préférés, la mutation et la BGC, qui selon les codons et les préférences d'usage du code agissent dans le même sens que la SUC ou dans le sens contraire, et la dérive qui joue indifféremment sur les codons. La présence des codons non préférés peut ainsi être maintenue par la mutation, la BGC et la dérive, forces dont l'intensité ne dépend pas du niveau d'expression, alors que les codons préférés sont favorisés par la SUC, dont l'intensité dépend de l'expression. Les gènes peu exprimés reflètent donc plutôt l'équilibre mutation-dérive-BGC et les gènes les plus exprimés sont ceux pour lesquels l'usage du code est le plus biaisé en faveur des codons préférés.

Des modèles théoriques de génétique des populations ont été développés pour mieux comprendre la manière dont les différentes forces en jeu s'équilibrent (par exemple Kimura 1981 ; Bulmer 1991 ; Akashi 1997 ; Higgs & Ran 2008). Ces modèles prédisent que la proportion de codons préférés à l'équilibre dans une population de taille finie dépend des taux de mutation entre codons préférés et non préférés, de la taille efficace de la population, de la force de la BGC et de l'intensité de la sélection sur l'usage du code. Pour un codon donné dans un gène donné, l'intensité de la SUC est déterminée par le niveau d'expression du gène et par les différences de valeur sélective des codons synonymes. Ces valeurs sélectives des codons synonymes sont elles-mêmes déterminées par la composition de la population d'ARNt présents dans la cellule et la capacité des codons à s'apparier à ces ARNt. Les types d'ARNt isoaccepteurs disponibles pour chaque acide aminé et la proportion de ces différents isoaccepteurs présents sont donc des paramètres majeurs de l'intensité de la SUC.

Les modèles de génétique des populations prédisent que l'efficacité de la sélection est plus faible dans les régions peu recombinantes, à cause de l'effet Hill-Robertson qui réduit la taille efficace locale (Gordo & Charlesworth 2001). L'effet de la SUC serait donc plus fort pour les séquences soumises aux plus forts taux locaux de recombinaison, celles-là même qui subissent le plus fortement la BGC. Les effets de la SUC et de la BGC peuvent alors se confondre si les codons préférés sont majoritairement des codons riches en GC : SUC et BGC enrichissent toutes deux en GC les séquences codantes, et ceci de manière hétérogène, selon les niveaux d'expression et les taux locaux de recombinaison. L'amélioration de l'efficacité de la sélection par la recombinaison reste toutefois faible et les effets d'une SUC plus efficace chez les séquences les plus recombinantes semblent masqués par les modifications de la composition nucléotidique directement associées à la recombinaison (BGC, biais mutationnel associé à la recombinaison ?) (Marais et al. 2001 ; Marais & Piganeau 2002).

Coévolution de l'usage du code et des ARNt

La composition de la population d'ARNt de la cellule – nature et proportions des différents ARNt isoaccepteurs – est déterminée par les gènes codant pour les ARNt. Ce sont donc ces gènes d'ARNt qui gouvernent le sens et l'intensité de la SUC. Or ces gènes sont eux-mêmes soumis à l'évolution et le répertoire de gènes ARNt est ainsi susceptible d'évoluer, par gain ou perte de copies, conversion génique ou mutation. Il est donc possible que l'évolution de ces gènes soit soumise à une pression exercée par l'usage du code lui-même – alors même que le répertoire de gènes ARNt est un déterminant majeur de l'évolution de l'usage du code…

Des modèles théoriques ont été développés autour de cette hypothèse de coévolution entre usage du code (au sens biais de composition) et ARNt. En modélisant l'évolution de l'usage du code en fonction de proportions fixées d'ARNt isoaccepteurs, le modèle de Bulmer (Bulmer 1987) prédit que plus les proportions d'ARNt sont inégales, plus l'usage des codons reconnus par ces ARNt est biaisé. En considérant à l'inverse l'évolution des proportions d'ARNt en fonction de biais d'usage du code fixés, ce modèle prédit que les proportions d'ARNt sont d'autant plus inégales que l'usage du code est biaisé. En combinant ces deux analyses, Bulmer prédit qu'il existe un équilibre stable de coévolution entre usage du code et ARNt et que cet équilibre est fonction du niveau d'expression. Si l'expression est forte, l'équilibre attendu correspond à un fort biais d'usage du code et à une forte inégalité des proportions d'ARNt.

Afin de prédire les états d'équilibre attendus d'une coévolution entre ARNt et usage du code, Higgs et Ran (Higgs & Ran 2008) ont développé un modèle plus complexe où les proportions d'ARNt et les biais d'usage du code ne sont pas fixés et peuvent véritablement coévoluer. Leur modèle prédit l'existence de différents états stables possibles d'usage du code et de répertoire d'ARNt. Cet attendu théorique est corroboré par la variabilité connue des préférences d'usage du code et des gènes ARNt isoaccepteurs entre acides aminés et entre espèces. La stabilité des états stables alternatifs prédits par le modèle de Higgs et Ran dépend très fortement du taux de bases G et C (GC) génomique, au point qu'une évolution graduelle du GC génomique entraîne des changements d'états stables (Figure 1). Une évolution de l'intensité de la BGC ou du biais mutationnel peut mener à une telle situation. Cependant, d'après ce modèle, le GC actuel ne permet pas de prédire l'état stable correspondant. Les changements de stabilité des états répondent en effet avec retard aux changements de GC : les seuils de GC correspondant à la

succession de changements d'états stables lorsque le GC croît diffèrent des seuils de changements lorsque le GC décroît (phénomène d'hystérésis, Figure 1).

D'après ces modèles donc, l'usage du code évoluerait vers un équilibre adapté au répertoire d'ARNt, où seraient favorisés les codons les plus efficacement traduits par ce répertoire donné. Et les populations d'ARNt évolueraient, tant du point de vue du type d'ARNt que des proportions relatives des ARNt isoaccepteurs, vers un équilibre adapté à la composition nucléotidique à l'équilibre mutation-BGC. L'état actuel d'usage du code et de répertoire d'ARNt serait donc le fruit d'une coévolution menant à un état d'équilibre où usage du code et ARNt sont mutuellement adaptés. Cette coévolution serait fortement marquée par l'histoire de l'évolution du GC génomique. Les possibilités d'appariement bancal entre espèces d'ARNt et codons devraient également intervenir.

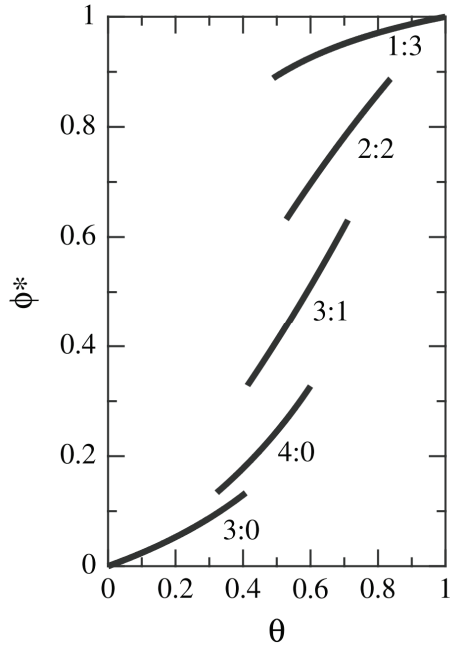


Figure 1 – Variation de l'usage du code (φ*) en fonction de la composition nucléotidique déterminée par les forces évolutives neutres telles que la mutation (θ), avec les combinaisons de nombre de copies des gènes ARNt correspondants aux différents états stables.

Modèle développé pour le cas d'un acide aminé au code deux fois dégénéré, avec des codons synonymes finissant en A et G. Les ARNt correspondants sont en U et C, avec un appariement de type *wobble* possible entre le codon en G et les ARNt en U. **φ*** est la fréquence d'usage du codon en G. A côté des états stables sont indiquées les combinaisons correspondantes de nombres de copies des gènes ARNt, sous la forme N_U:N_C.

(Tiré de Higgs & Ran 2008)

Diversité et évolution des préférences d'usage du code chez les plantes

Nous avons précédemment étudié l'histoire particulière de l'évolution du taux de GC chez les plantes. Nous avons montré que les patrons de GC des plantes sont très contrastés, avec une gamme continue de paysages nucléotidiques allant de génomes pauvres en GC et plutôt homogènes chez les gymnospermes, les angiospermes basales et la plupart des eudicotylédones, à des génomes riches en GC et très hétérogènes chez certains groupes d'eudicotylédones, chez les monocotylédones et tout particulièrement chez les Poaceae. Par parcimonie, il semble ainsi que le génome ancestral des angiospermes ait été homogène et pauvre en GC. Pour présenter aujourd'hui une telle palette de variation de leur composition nucléotidique, les génomes des plantes ont vraisemblablement subi au cours de leur évolution et de leur diversification taxonomique des épisodes indépendants d'enrichissement en GC (par exemple pour les Poaceae et les Myrtales), voire des épisodes d'appauvrissement du GC (Ranunculales par exemple).

La BGC pourrait être à l'origine de ces variations de GC. Cette hypothèse est étayée par nos précédents résultats. Nous avons notamment mis en évidence des variations importantes du taux de GC de régions codantes et non codantes, des corrélations locales entre GC1, GC2, GC3 et GCUTR (taux de bases G et C en première, deuxième et troisième position de codon et aux UTR) et une corrélation entre richesse et hétérogénéité en GC du génome : ce sont des effets attendus de la BGC au niveau génomique. A l'échelle des angiospermes, cette diversité des paysages nucléotidiques pourrait être expliquée par des variations d'intensité de la BGC et des patrons de recombinaison.

Si la BGC semble ainsi jouer un rôle primordial dans le façonnement des patrons de GC des plantes, la SUC pourrait également entrer en jeu. Nous avons en effet détecté la présence de SUC chez la quasi-totalité des espèces de gymnospermes et d'angiospermes étudiées, en mettant en évidence une corrélation entre niveau d'expression et GC3. La SUC n'a par définition pas d'effet sur le GC de l'ADN non codant mais elle est susceptible, selon son intensité et son sens et selon les patrons d'expression, de générer des variations de GC1 et surtout de GC3 entre gènes. Si la SUC jouait de manière significative sur les patrons de GC, elle pourrait contribuer à la diversité observée entre espèces à travers des variations de son sens (changement des préférences d'usage du code : codons en G/C préférés ou non) ou de son intensité (variations des différences de valeur sélective entre codons : changements des proportions entre ARNt).

Il apparaît effectivement que les préférences d'usage du code varient pour les quelques espèces de plantes pour lesquelles elles ont été établies : chez les graminées, riches en GC, les

codons préférés sont très majoritairement ceux finissant en G/C, alors que chez les eudicotylédones la proportion de codons préférés finissant en A/T peut atteindre un tiers (Kawabe & Miyashita 2003 ; Wang & Roossinck 2006). Il reste un déficit de connaissances quant aux préférences d'usage du code et à leur éventuelle coïncidence avec les patrons de GC chez les gymnospermes, les angiospermes basales et de nombreux groupes d'eudicotylédones et de monocotylédones, notamment chez les monocotylédones basales et les eudicotylédones riches en GC. Quant aux différences de valeur sélective entre codons, une variabilité a été montrée entre *Arabidopsis thaliana* et *Oryza sativa*, espèces modèles aux paysages nucléotidiques et aux préférences d'usage du code contrastés, puisque le nombre et la proportion relative de leurs ARNt isoaccepteurs diffèrent (Michaud et al. 2011).

Chez les plantes donc, les paysages nucléotidiques, très variables, seraient vraisemblablement façonnés par la BGC avec un rôle possible de la SUC, avec des répertoires d'ARNt et des préférences d'usage du code eux aussi variables. Or, selon les modèles théoriques, l'évolution du taux de GC génomique joue un rôle déterminant dans la coévolution entre l'usage du code et la SUC (Higgs & Ran 2008). Si ce taux de GC génomique est lui-même soumis à la BGC et à son évolution, alors il y a nécessairement une coévolution entre l'intensité de la BGC, l'usage du code et la SUC. A la lumière de ces prédictions théoriques, il apparaît donc important d'évaluer dans quelle mesure la SUC peut affecter les patrons de GC et leur évolution, et, à l'inverse, de chercher à déterminer si la coévolution entre SUC et usage du code a pu être affectée par les variations de GC, dues par exemple à la BGC. Nous pourrions par exemple nous attendre à ce que chez les groupes riches en GC, qui ont été soumis à l'action d'une BGC forte, la coévolution entre usage du code et SUC ait mené à ce que les codons préférés finissent majoritairement en G/C, alors que chez les espèces moins marquées par la BGC, la préférence de codons finissant en A/T n'ait pas été contre-sélectionnée.

Nous utiliserons des banques d'EST (*Expressed Sequence Tag*) pour 130 espèces de plantes largement réparties sur la phylogénie (neuf gymnospermes, six angiospermes basales, 26 monocotylédones et 89 eudicotylédones) afin d'étudier l'évolution des préférences d'usage du code chez les plantes et d'interpréter leur variabilité actuelle et leur évolution en les mettant en parallèle avec l'évolution des patrons de GC. Nous déterminerons dans un premier temps les codons préférés et détestés pour chaque espèce puis nous reconstruirons les préférences ancestrales. Nous discuterons de la stabilité des préférences selon les acides aminés, des types de changements de préférences (changement vers A/T ou G/C) selon le contexte en GC, et de la répartition des changements sur la phylogénie (groupes variables, niveaux taxonomiques affectés).

Nous chercherons à voir si l'évolution des préférences d'usage du code et l'évolution des taux de GC sont liées ou si elles peuvent être découplées, et dans quelle mesure l'évolution des taux de GC a pu orienter la coévolution entre répertoire d'ARNt et usage du code.

Jeux de données de séquences EST

Les jeux de données d'unigènes EST (*Expressed Sequence Tag*) ont été récupérés à partir de trois bases de données publiques : PlantGDB versions 157a à 181a (http://www.plantgdb.org/), The Gene Index Project (http://compbio.dfci.harvard.edu/tgi/), et TIGR Plant Transcript Assemblies versions 1 à 5 (Childs et al., 2007 ; http://plantta.jcvi.org/). Des données d'EST bruts ont également été téléchargées sur GenBank (http://www.ncbi.nlm.nih.gov/genbank/) afin d'accroître la couverture phylogénétique du jeu d'espèces, notamment en couvrant des groupes souvent peu représentés tels que les angiospermes basales et les monocotylédones hors graminées. Ces séquences d'EST bruts ont été assemblées grâce au pipeline d'analyses d'EST EST2uni (Forment et al. 2008 ; voir chapitre 1). Après avoir appliqué les différents filtres (voir détails plus bas), notre jeu de données est constitué de 130 espèces : neuf gymnospermes, six angiospermes basales, 26 monocotylédones et 89 eudicotylédones. Les données de 86 espèces proviennent de PlantGDB, 22 de The Gene Index Project, six de TIGR Plant Transcript Assemblies et 16 de GenBank.

Annotation des EST, expression et filtrage

Nous avons traité les séquences d'unigènes EST par le pipeline d'annotation et de traduction d'EST prot4EST version 2.2 (Wasmuth & Blaxter 2004) afin de déterminer les positions de codon et les régions traduites (voir chapitre 1). L'étape de recherche BLASTX de cette analyse prot4EST a été effectuée contre la base de données UniProtKB/Swiss-Prot plant version 2010_8 (http://www.uniprot.org/). Les séquences présentant des décalages de cadre de lecture, des codons stop anormaux, ou d'une longueur inférieure à 100 nucléotides ont été supprimées des données. Les espèces présentant moins de 1 000 unigènes après application de ces filtres ont été supprimées.

Nous avons utilisé comme approximation du niveau d'expression le nombre d'EST bruts s'appariant avec chaque unigène (nombre de « hits »). Afin de ne conserver que les espèces dont les données d'expression présentaient une gamme de variation suffisante, nous avons gardé les espèces dont les données remplissaient les trois critères suivant : plus de 300 séquences de valeurs d'expression supérieures à 1 (séquences non-singletons), valeur d'expression maximale supérieure à 20, et écart-type des valeurs d'expression différentes de 1 supérieur à 2.

Pour chaque espèce, nous avons défini deux groupes de séquences en fonction de ces niveaux d'expression approximés (Duret & Mouchiroud 1999). Le groupe à expression faible comprend les singletons, c'est-à-dire les séquences dont le niveau d'expression approximé est égal à 1. Le groupe à expression forte comprend le tiers le plus exprimé des séquences restantes, c'est-à-dire les séquences dont le niveau d'expression est au-delà du deuxième tercile parmi les séquences non-singletons. Grâce aux filtres précédents il y a au moins 100 séquences dans chacune des classes.

L'ensemble du jeu de données est constitué de 154 espèces après application de ces filtres. Un filtre supplémentaire sera appliqué après la détermination des préférences d'usage du code pour éliminer les espèces ayant trop d'acides aminés sans préférences significatives (voir plus bas la partie « Détermination des préférences d'usage du code »). Après application de ce dernier filtre, les 130 espèces finalement retenues constituent un jeu global de près de 2,9 millions d'unigènes EST, pour un total de plus de 408 millions d'acides aminés.

Statistiques descriptives de l'usage du code et de la composition nucléotidique

Les taux de G et C en troisième position de codon (GC3), les fréquences relatives des codons, les fréquences des acides aminés et les taux de GC3 particuliers par niveau de dégénérescence des acides aminés et par classe d'expression ont été calculés pour chaque espèce grâce à des scripts Perl. Les analyses statistiques ont été réalisées sous l'environnement d'analyses statistiques R (R Development Core Team 2008).

L'usage relatif d'un codon synonyme, le RSCU (*Relative Synonymous Codon Usage*), est défini comme étant la fréquence observée d'un codon divisée par la fréquence attendue si tous les codons synonymes de l'acide aminé correspondant étaient utilisés à fréquences égales (Sharp et al. 1986 ; Duret & Mouchiroud 1999). Cette mesure permet de s'affranchir des biais de composition en acide aminé entre séquences. Elle se calcule pour chaque séquence pour le codon c de l'acide aminé a comme $RSCU_{ca} = d_a * freq(c) / freq(a)$, où d_a est le degré de dégénérescence du code de l'acide aminé a, $freq(c)$ est la fréquence du codon c dans la séquence, et $freq(a)$ est la fréquence de l'acide aminé a dans la séquence. Ainsi pour un codon majoritaire (minoritaire) le RSCU est supérieur (inférieur) à 1, et le RSCU vaut 1 s'il n'y a pas de biais d'usage des codons synonymes. Les RSCU ont été calculés pour chaque séquence pour chaque codon – hormis ceux de la méthionine et du tryptophane, les deux acides aminés codés par un seul codon. La moyenne de

ces RSCU a été calculée sur l'ensemble des séquences et pour les groupes de séquences à expression faible et forte. Nous avons également calculé le taux de GC3 séparément pour les différentes catégories de codons (codons codant pour des acides aminés au code deux fois, trois fois, quatre fois et six fois dégénéré) ainsi que pour les groupes de séquences à expression faible et forte pour chacune de ces catégories de codons.

Analyse en composantes principales

Une analyse en composantes principales (ACP) est réalisée avec pour variables les RSCU moyens par espèce pour 62 codons (les 59 codons des acides aminés au code dégénéré et les trois codons stop). L'analyse est réalisée sous l'environnement d'analyses statistiques R (R Development Core Team 2008) avec le paquet ade4 (Dray et al. 2007). Des tests de corrélation de Spearman sont effectués pour tester la corrélation entre les GC3 moyens des espèces et les coordonnées des espèces sur les différents axes de l'ACP.

Détermination des préférences d'usage du code

Le ΔRSCU, différence de RSCU moyen entre les séquences à expression forte et faible, permet de déterminer les codons dont l'usage varie avec l'expression (Duret & Mouchiroud 1999). Si le ΔRSCU est nul, il n'y a pas de préférences. Si le ΔRSCU est significativement positif (négatif), alors le codon est préféré (détesté). Les ΔRSCU ont été calculés pour tous les codons et leur significativité au seuil de 10^{-3} a été testée par le test non paramétrique de Mann-Whitney-Wilcoxon. Ce seuil de significativité a été choisi pour sa forte stringence, nos tests étant multiples (correction de Bonferroni du seuil 0,05 pour 41 codons indépendants : seuil ≈ 0,00125). En fonction de ces résultats l'état de préférence de chaque codon a été déterminé parmi les quatre états possibles suivants : préféré significatif, préféré non significatif, détesté non significatif et détesté significatif.

Pour certaines espèces, le nombre de codons sans préférence significative était particulièrement élevé. Les préférences d'usage des codons des neuf acides aminés au code deux fois dégénéré étant très conservées chez les plantes, l'absence de préférence à ces codons pour ces espèces peut être attribuable à un manque de puissance statistique des jeux de données ou bien à une SUC de faible intensité. Or ce sont les préférences en elles-mêmes et leurs changements qui nous intéressent. Nous avons donc éliminé les espèces présentant une absence

de préférence significative pour au moins quatre des acides aminés au code deux fois dégénéré. Les 130 espèces conservées constituent le jeu de données final.

Représentation des préférences sur la phylogénie

L'arbre phylogénétique des 130 espèces a été reconstitué à l'échelle ordinale (Chase et al. 2009) puis a été complété à l'échelle des familles et des genres à partir de différents travaux (gymnospermes : Burleigh & Mathews 2004 ; Poaceae : Bouchenak-Khelladi et al. 2008 ; Triticeae : Escobar et al. 2010 ; Caryophyllales : Cuénoud et al. 2002 ; Brassicales : Hall et al. 2002 ; Brassicaceae et *Brassica* : Bailey et al. 2006 ; Fabaceae : Wojciechowski et al. 2004 ; *Actinidia* : Chat et al. 2004 ; *Citrus* : Froelicher et al. 2011 ; *Gossypium* : Cronn et al. 2002 ; *Populus* : Hamzeh & Dayanandan 2004 ; Ericales : Schonenberger et al. 2005 ; Lamiales : Schäferhoff et al. 2010 ; Solanaceae : Olmstead et al. 2008 ; *Solanum* : Goldberg et al. 2010 ; Asteraceae : Goertzen et al. 2003 ; *Lactuca* : Koopman et al. 2001 ; *Helianthus* : Timme et al. 2007). Les préférences d'usage des codons ont ensuite été placées en regard de cet arbre.

Reconstruction des préférences ancestrales

Les préférences ancestrales d'usage du code ont été reconstruites par parcimonie pour tous les nœuds de la phylogénie, avec le paquet Ancestral State Reconstruction du logiciel Mesquite (Maddison & Maddison 2005). Les préférences pour chaque codon ont été recodées en trois états : préféré significatif, détesté significatif, et neutre, qui regroupe les préférences non significatives. La méthode de parcimonie utilisée attribue le même coût à tous les changements d'état.

Biais d'usage du code

Biais d'usage pour tous les codons chez toutes les espèces

En moyenne sur l'ensemble des espèces étudiées, tous les codons ont un RSCU différent de 1, valeur attendue si les codons synonymes d'un acide aminé sont utilisés de manière équiprobable. Ces RSCU moyens sont variables entre les codons (écart-type = 0,26), avec des moyennes extrêmes de 0,51 pour GCG (alanine) et de 1,69 pour AGA (arginine). L'usage des codons est également variable entre espèces, avec un écart-type moyen des RSCU par codon de 0,17, et un écart-type maximal de 0,34 pour les RSCU moyens de AGA (arginine). Il y a donc bel et bien des biais d'usage pour l'ensemble des codons et chez toutes les espèces – y compris chez les espèces que nous avons supprimées du jeu de données pour cause de manque de préférences détectées sur les acides aminés au code deux fois dégénéré. L'intensité de ces biais d'usage varie selon les codons et les espèces.

Le premier axe de l'ACP réalisée sur les RSCU des codons explique 75,5 % de la variance totale, contre seulement 6,2 % et 5,4 % respectivement portés par les deuxième et troisième axes (Figure 2 C). Les coordonnées des espèces sur l'axe 1 sont extrêmement bien corrélées avec le taux de G et C en troisième position de codon (GC3) des espèces (corrélation de Spearman : rho = - 0,99, pvalue < $2{,}2*10^{-16}$; Figure 2 D). Aucune corrélation significative entre coordonnées et GC3 n'est trouvée pour les autres axes. Comme attendu, les biais de composition nucléotidique des génomes ont un impact majeur sur l'usage du code. La dispersion des espèces sur cet axe 1 permet de retrouver les groupes à patrons de GC similaires tels que déterminés dans le chapitre précédent, des plus riches en GC, les graminées, jusqu'aux eudicotylédones les plus pauvres (Figure 2 D). Les variables codons s'organisent globalement en deux groupes opposés sur l'axe 1, avec les codons finissant en G ou C d'une part et les codons finissant en A ou T d'autre part (Figure 2 A). Les codons des acides aminés deux fois dégénérés – qui ont systématiquement un codon en G ou C, et un codon en A ou T – expliquent en particulier une part importante de la variation sur l'axe 1, alors que les codons les moins bien représentés appartiennent à des acides aminés six fois dégénérés.

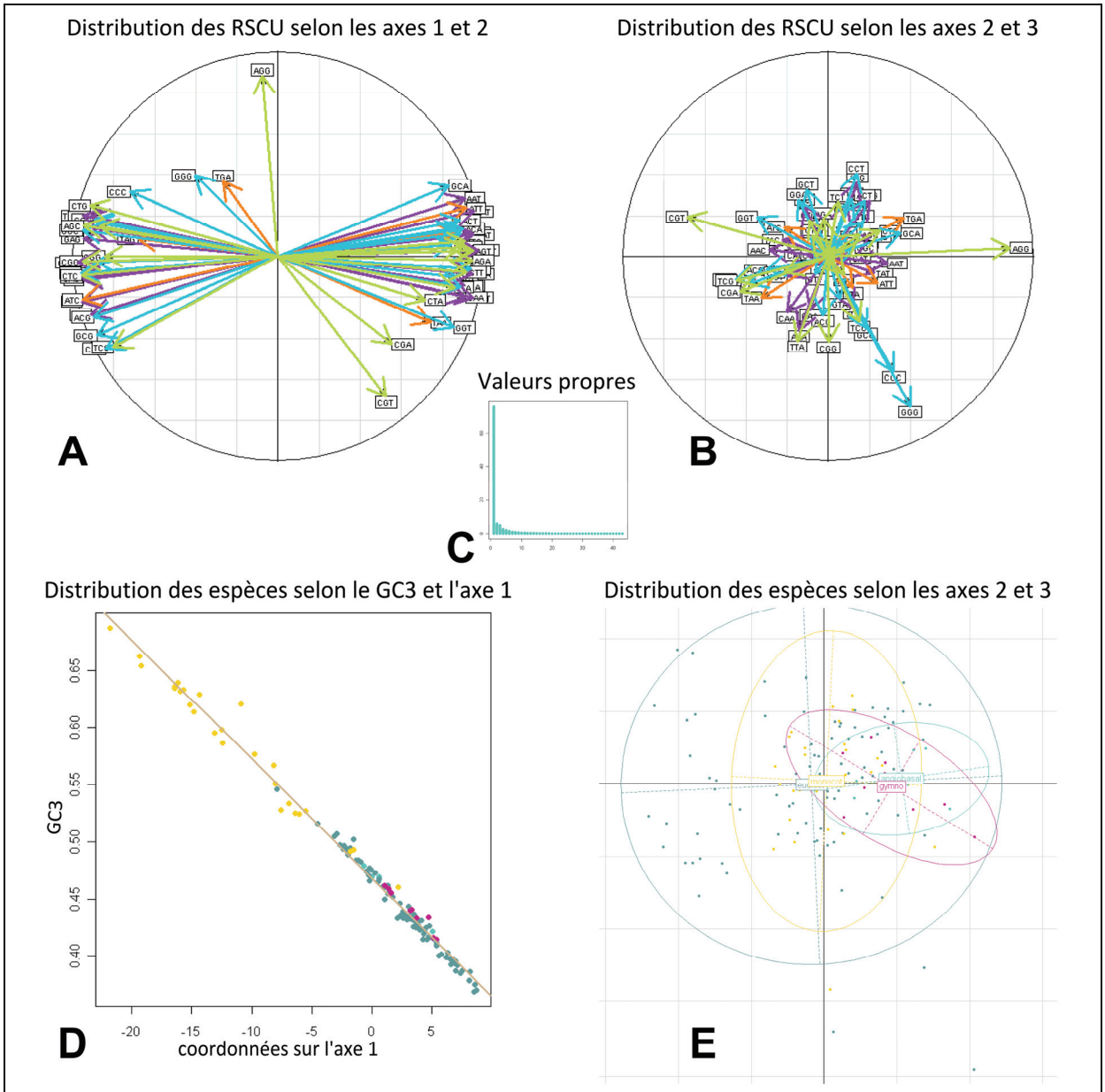


Figure 2 – ACP sur les RSCU moyens des codons pour les 130 espèces.

A). Cercle de corrélation des variables RSCU moyens des codons, sur les axes 1 et 2. **B)**. Cercle de corrélation des variables RSCU moyens des codons, sur les axes 2 et 3. **C)**. Histogramme des valeurs propres des axes de l'ACP. **D)**. Corrélation entre le GC3 moyen des espèces et leur coordonnée sur l'axe 1 de l'ACP. **E)**. Distribution des espèces sur le plan de l'ACP défini par les axes 2 et 3, avec les ellipses d'inertie par groupe.

La couleur des variables (**A**, **B**) est attribuée selon le niveau de dégénérescence des acides aminés : violet pour les acides aminés deux fois dégénérés, orange pour l'acide aminé trois fois dégénéré et les codons stop, bleu pour les quatre fois dégénérés et vert pour les six fois dégénérés.

Les groupes d'espèces sont figurés par les couleurs suivantes (**D**, **E**) : gymnospermes en fuchsia, angiospermes basales en turquoise, monocotylédones en jaune et eudicotylédones en bleu.

Sur le plan défini par les axes 2 et 3, il y a une variation certaine d'usage du code entre les espèces mais elle ne permet pas de retrouver les grands groupes taxonomiques aussi nettement que sur l'axe 1 (Figure 2 E). Les codons des acides aminés deux fois dégénérés expliquent une faible part de la variation sur ces axes et les sources majeures de variation sont dues à quelques uns des codons des acides aminés quatre et six fois dégénérés. (Figure 2 B). Ces résultats suggèrent plutôt des changements assez ponctuels d'usage du code, pour certains acides aminés et certaines espèces.

Sélection sur l'usage du code

L'existence de biais d'usage du code n'implique pas nécessairement la présence de sélection sur l'usage des codons synonymes, dont on détecte l'existence par la présence d'une corrélation entre biais d'usage et expression.

La troisième position de codon étant considérée comme position synonyme à cause de la redondance du code génétique, le GC3 reflète globalement l'usage des codons synonymes – ce que nous avons confirmé précédemment avec l'ACP. Sous l'hypothèse de sélection sur l'usage du code (SUC), les gènes les plus exprimés subissent les pressions de SUC les plus fortes et s'enrichissent en codons préférés. Si les codons préférés sont majoritairement les codons finissant en G/C, ce qui semble être le cas chez les plantes (Kawabe & Miyashita 2003 ; Wang & Roossinck 2006), alors une corrélation entre GC3 et expression au niveau des transcrits doit être trouvée. Comme nous l'avons montré dans le chapitre précédent, le GC3 est corrélé de manière significative avec le niveau d'expression, ce que nous retrouvons ici pour 121 espèces parmi les 130 espèces de notre jeu de données. Ce résultat suggère la présence de SUC chez la plupart des espèces d'angiospermes et de gymnospermes avec une majorité de codons préférés terminant en G ou C comme déjà montré pour quelques espèces (Kawabe & Miyashita 2003 ; Wang & Roossinck 2006). Nous pouvons maintenant déterminer plus précisément les tables de codons préférés pour l'ensemble des espèces du jeu de données.

Etudier la corrélation entre expression et fréquence pour chaque codon permet de caractériser précisément les préférences d'usage du code. Pour exemple, la Figure 3 montre les corrélations entre les fréquences des quatre codons synonymes de l'alanine pour deux espèces contrastées, le cacaoyer – eudicotylédone pauvre en GC – et le seigle – graminée riche et hétérogène en GC. Les codons préférés (détestés) sont ceux dont la fréquence augmente (diminue) significativement avec l'expression. Pour l'alanine, les préférences d'usage sont donc

variables entre le cacaoyer et le seigle : leurs codons préférés sont respectivement GCA et GCT, et GCC et GCG. La Figure 4 montre ces corrélations pour l'ensemble des codons pour ces deux mêmes espèces. Chez le seigle, les 25 codons préférés détectés sont des codons finissant en G ou C, alors que chez le cacaoyer, 14 codons préférés sont des codons en A ou T, et 19 sont des codons en G ou C. Indépendamment de la détection des préférences d'usage du code, les graphes de la Figure 4 permettent de bien faire la distinction entre codons préférés et détestés d'une part, et codons minoritaires et majoritaires d'autre part. Les codons majoritaires (minoritaires) sont ceux dont la fréquence relative observée est supérieure (inférieure) à la fréquence relative attendue, 0,25 par exemple pour les codons des acides aminés quatre fois dégénérés. Ces graphes reflètent bien les biais de composition nucléotidique connus pour ces deux espèces. Les préférences ne vont pas nécessairement dans le même sens que les biais de composition. Par exemple chez le cacaoyer, TTC est le codon préféré pour la phénylalanine alors qu'il est minoritaire (fréquence relative moyenne = 34,2 %).

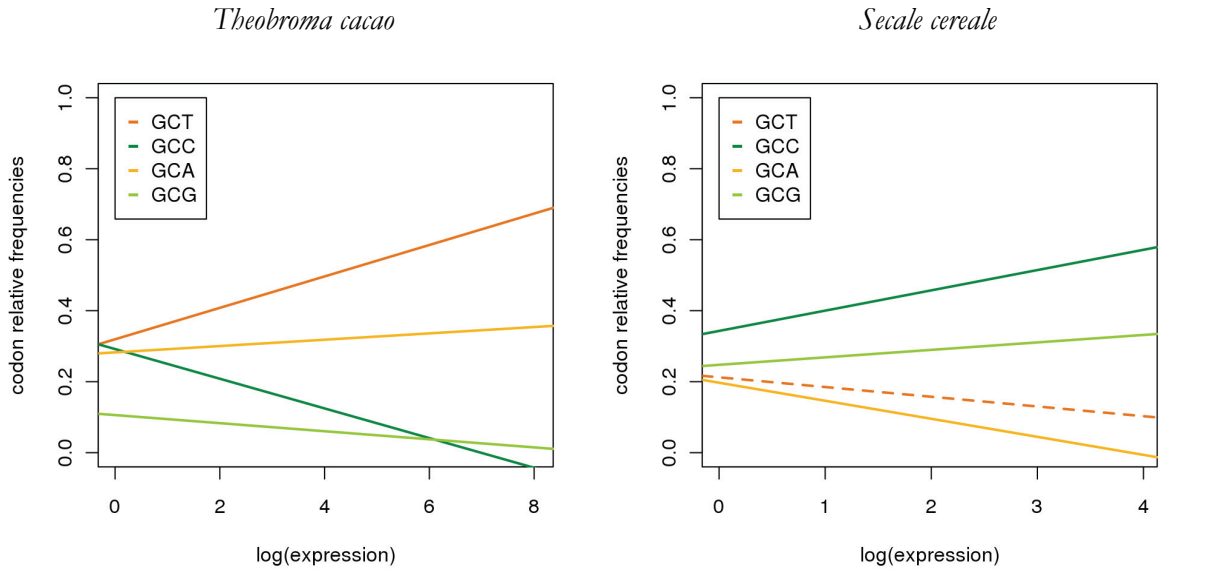
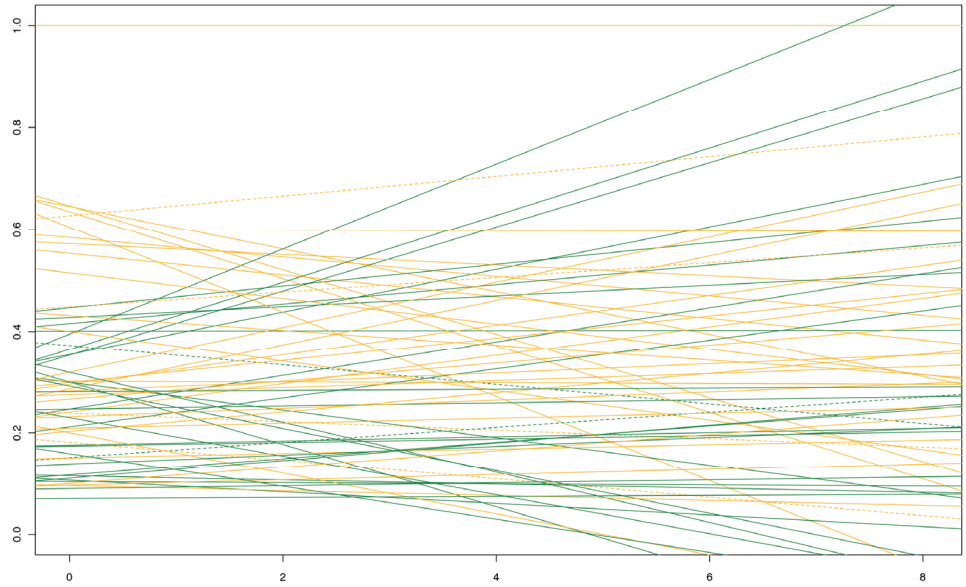


Figure 3 – Corrélation linéaire entre les fréquences relatives des codons synonymes de l'alanine et le logarithme de l'expression pour le cacaoyer *Theobroma cacao* (eudicotylédone) et le seigle *Secale cereale* (monocotylédone).

En vert : codons finissant en G ou C. En orange : codons finissant en A ou T. Trait plein : préférence significative. Trait pointillé : préférence non significative.

Theobroma cacao



Secale cereale

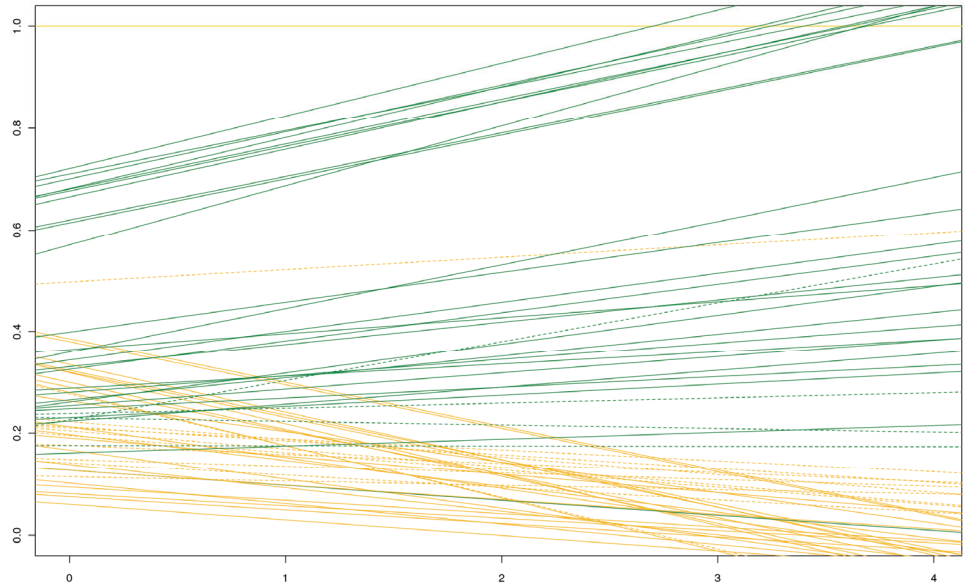


Figure 4 – Corrélation linéaire entre les fréquences relatives des 64 codons et le logarithme de l'expression pour le cacaoyer *Theobroma cacao* (eudicotylédone) et le seigle *Secale cereale* (monocotylédone).

En vert : codons finissant en G ou C. En orange : codons finissant en A ou T. En jaune : codons des acides aminés non dégénérés. Trait plein : préférence significative. Trait pointillé : préférence non significative.

Variabilité et évolution des préférences d'usage du code

Les préférences actuelles d'usage du code sont déterminées à partir du ΔRSCU, une mesure du biais d'usage selon l'expression. Pour toutes les espèces, des codons préférés et des codons détestés ont été détectés pour la quasi-totalité des acides aminés. Ces préférences d'usage du code sont souvent partagées par les espèces, mais présentent tout de même une certaine labilité.

Les codons stop, qui ne codent pourtant pas pour des acides aminés, présentent également des préférences d'usage, puisque vingt-trois espèces ont une préférence significative pour au moins l'un des trois codons stop (Figure 5). Ces préférences ne varient pas entre les espèces. TAG et TGA sont préférés, TAA est détesté.

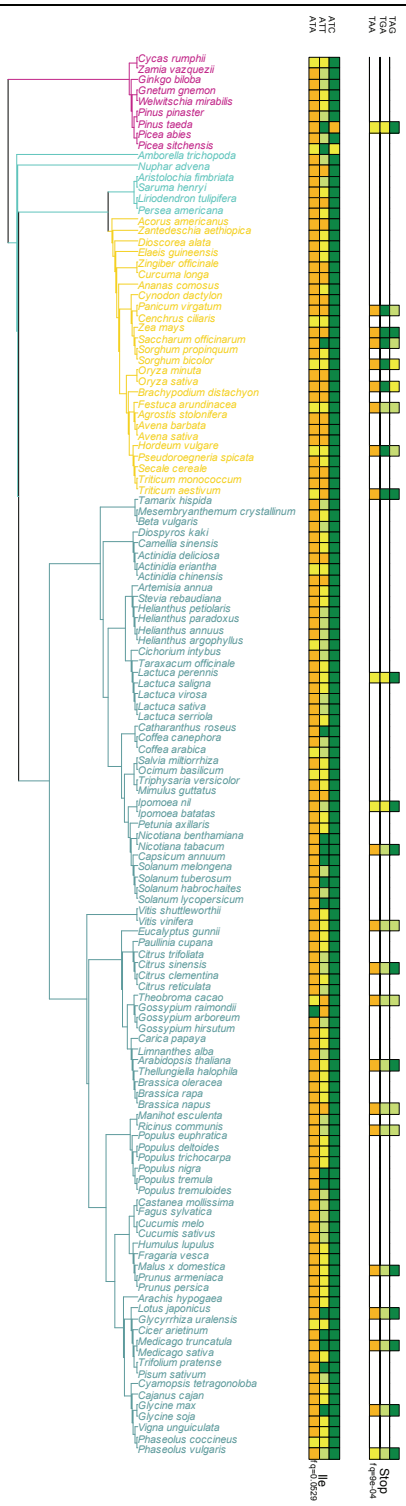


Figure 5 – Préférences d'usage du code pour les codons synonymes de l'acide aminé trois fois dégénéré et des codons stop, représentées sur l'arbre phylogénétique des espèces.

En vert foncé : codons significativement préférés. En vert clair : codons préférés de manière non significative. En jaune : codons détestés de manière non significative. En orange : codons significativement détestés. Les préférences n'ont pas été représentées lorsque aucun des codons synonymes n'avait de préférence significative.

Préférences en faveur de G ou C très conservées pour les acides aminés au code deux et trois fois dégénéré

Pour les codons codant pour l'asparagine, la cystéine, l'acide glutamique, la glutamine, la lysine, la phénylalanine et la tyrosine, les préférences d'usage ne varient pas parmi les 130 espèces (Figure 6). Les codons se terminant par G ou C sont toujours préférés, et les codons en A ou T toujours détestés.

Pour l'histidine et l'acide aspartique, les deux autres acides aminés au code deux fois dégénéré, les préférences d'usage des codons varient très peu (Figure 6). Les codons préférés sont chez la grande majorité des espèces les codons en G ou C. Pour l'histidine, deux espèces préfèrent tout de même le codon en T : le pin *Pinus taeda* (gymnosperme) et la luzerne *Medicago truncatula* (eudicotylédone). Les préférences pour les codons de l'acide aspartique sont plus variables, avec une préférence en faveur du codon en T chez 13 espèces réparties sur la phylogénie (une gymnosperme, une angiosperme basale, une monocotylédone, dix eudicotylédones).

Les préférences pour l'usage des codons synonymes de l'isoleucine, le seul acide aminé au code trois fois dégénéré, sont assez peu variables (Figure 5). Hormis pour deux gymnospermes (le pin *Pinus taeda* et l'épicea *Picea sitchensis*), le codon en C est toujours préféré. Le codon en T est préféré ou détesté pour quelques espèces, et souvent ne présente pas de préférence significative. Le codon en A est très majoritairement détesté – mais est préféré chez le coton *Gossypium raimondii*.

Les préférences d'usage des codons des acides aminés au code deux et trois fois dégénéré ont été très conservées chez les plantes au cours de leur évolution. Les quelques changements de préférences sont récents et ne se concentrent pas dans des groupes taxonomiques particuliers (pour exemple, le codon CAT de l'histidine en Figure 7 ; pour les autres codons, voir en annexe).

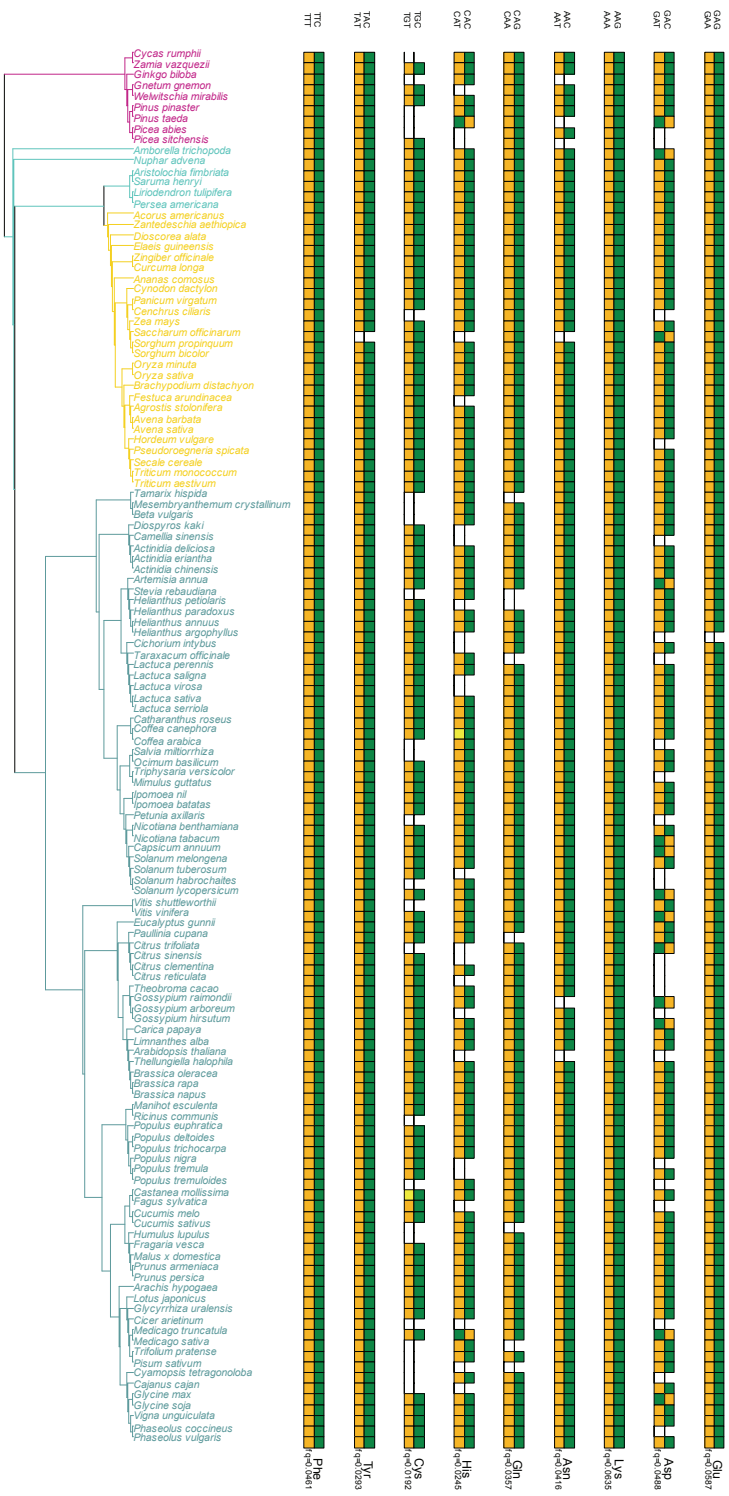


Figure 6 – Préférences d'usage du code pour les codons synonymes des neuf acides aminés deux fois dégénérés, représentées sur l'arbre phylogénétique des espèces.

En vert foncé : codons significativement préférés. En vert clair : codons préférés de manière non significative. En jaune : codons détestés de manière non significative. En orange : codons significativement détestés. Les préférences n'ont pas été représentées lorsque aucun des codons synonymes n'avait de préférence significative.

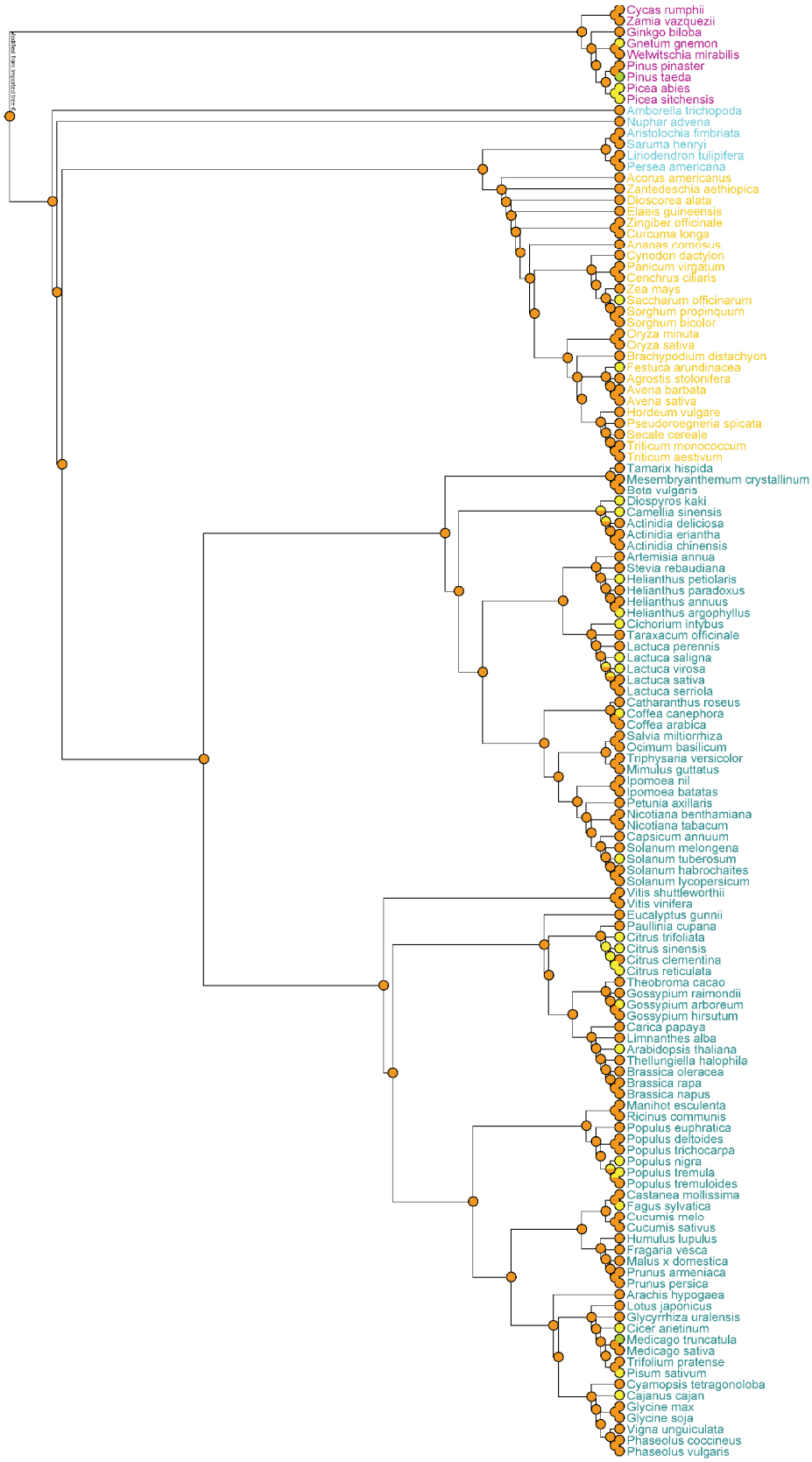


Figure 7 – Préférences ancestrales reconstruites pour le codon CAT de l'histidine

Pastille verte : codon préféré. Pastille jaune : codon neutre. Pastille orange : codon détesté.

Labilité des préférences pour les acides aminés au code quatre et six fois dégénéré

Les préférences d'usage pour les codons de l'alanine, de la glycine, de la proline, de la thréonine et de la valine, les cinq acides aminés au code quatre fois dégénéré, sont beaucoup plus variables que pour les acides aminés précédents (Figure 8). Résumons les patrons de variation des préférences d'usage de ces codons, de l'acide aminé le moins variable au plus variable :

- Valine : ses préférences sont les moins variables parmi les acides aminés au code quatre fois dégénéré. Les codons en G/C sont très majoritairement préférés et le codon en A détesté. Le codon en T est plutôt préféré chez les eudicotylédones mais détesté chez les monocotylédones.

- Glycine : le codon en C est nettement préféré, sauf chez quelques groupes d'eudicotylédones, alors que le codon en G est plutôt détesté. Les codons en A/T sont préférés chez la plupart des eudicotylédones.

- Thréonine : le codon en C est très majoritairement préféré. Le codon en G est préféré chez les gymnospermes, les angiospermes basales et les monocotylédones basales et est souvent détesté chez les autres espèces. Le codon en T est préféré chez les eudicotylédones, et le codon en A est détesté chez la plupart des espèces.

- Alanine : les codons en G/C sont le plus souvent préférés chez les gymnospermes, les angiospermes basales et les monocotylédones, alors que chez les eudicotylédones ils sont fréquemment détestés. Le codon en T a des préférences variables chez les monocotylédones et est toujours préféré chez les autres groupes. Le codon en A est majoritairement détesté.

- Proline : les préférences sont très variables. Les codons en C et G sont plus souvent préférés chez les non eudicotylédones que chez les eudicotylédones. Les codons en A/T sont souvent préférés chez les eudicotylédones.

Les préférences pour les codons codant pour l'arginine, la leucine et la sérine, les trois acides aminés au code six fois dégénéré, sont également assez variables (Figure 9) :

- Leucine : chez les non-eudicotylédones CTC et CTG sont préférés, TTG, CTA et TTA sont détestés et CTT est variable. Chez les eudicotylédones, les trois codons en G/C et CTT sont très majoritairement préférés et CTA et TTA détestés.

- Arginine : CGC et AGG sont très souvent préférés alors que CGG est plutôt détesté. CGT est largement préféré chez les non monocotylédones. CGA et AGA sont généralement détestés.

- Sérine : les trois codons en G/C sont très souvent préférés. Les trois codons en A/T sont le plus souvent détestés mais sont préférés chez quelques petits groupes d'eudicotylédones proches.

Les préférences d'usage des codons synonymes de ces acides aminés au code quatre et six fois dégénéré sont donc très variables, avec au minimum un codon préféré et un codon détesté, et jusqu'à respectivement trois et cinq codons préférés et trois et quatre codons détestés. La distribution de ces variations sur la phylogénie est souvent ponctuelle, avec parfois même des variations entre espèces du même genre. C'est le cas par exemple pour les quatre espèces de *Solanum* : l'aubergine *S. melongena* et la tomate sauvage *S. habrochaites* préfèrent les codons ACC (Thr), GTC (Val) et GCC (Ala), alors que la pomme de terre *S. tuberosum* et la tomate cultivée *S. lycopersicum* les détestent. Les variations concernent parfois de petits groupes taxonomiques entiers. Les Ebénacées (kaki, thé, kiwis) par exemple ont souvent des préférences différentes des autres groupes proches sur la phylogénie. Les variations sont aussi présentes à des niveaux taxonomiques supérieurs, variations communes aux eudicotylédones par exemple, avec toutefois souvent quelques variations ponctuelles qui atténuent ce patron.

Les changements de préférence d'usage des codons au cours de l'évolution des plantes ont été plus nombreux pour les codons des acides aminés au code quatre ou six fois dégénéré que pour ceux des acides aminés au code deux ou trois fois dégénéré (pour l'ensemble des codons, voir en annexe). La plupart des changements sont récents et bien distribués sur la phylogénie. Par exemple le codon TCC de la sérine a évolué du statut de codon préféré à codon détesté à 17 reprises de manière indépendante dans des groupes taxonomiques bien différents (Figure 10). Certains changements affectent des nœuds plus profonds. Par exemple le codon GCG de l'alanine est devenu détesté à partir de l'ancêtre commun aux Campanulidées et aux Lamiidées, ainsi qu'à plusieurs reprises plus récemment chez les Rosidées, les Poacées et les Pinacées (Figure 11). Les cas d'inversion de préférence ne sont pas rares. Ainsi GCG est redevenu préféré chez le basilic *Ocimum basilicum* (Figure 11). Les changements de préférences d'un codon ne sont généralement pas synchrones avec des changements de préférences pour les autres codons synonymes de l'acide aminé concerné.

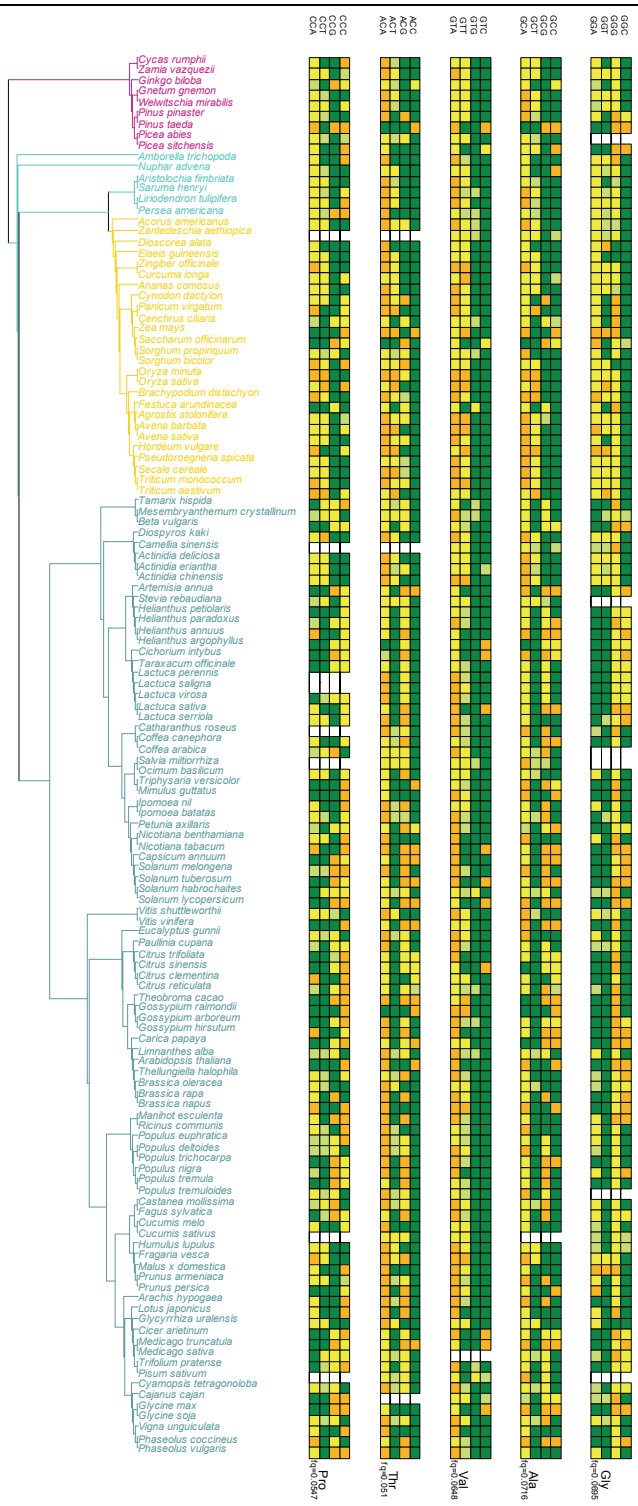


Figure 8 – Préférences d'usage du code pour les codons synonymes des cinq acides aminés quatre fois dégénérés, représentées sur l'arbre phylogénétique des espèces.

En vert foncé : codons significativement préférés. En vert clair : codons préférés de manière non significative. En jaune : codons détestés de manière non significative. En orange : codons significativement détestés. Les préférences n'ont pas été représentées lorsque aucun des codons synonymes n'avait de préférence significative.

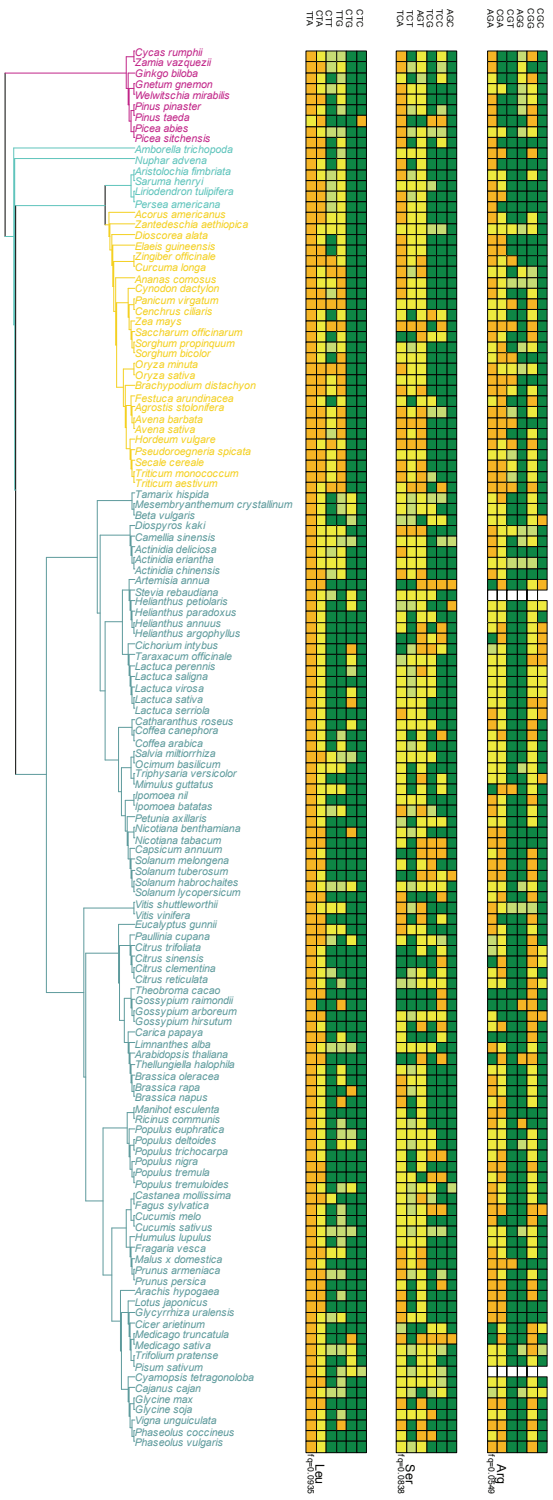


Figure 9 – Préférences d'usage du code pour les codons synonymes des trois acides aminés six fois dégénérés, représentées sur l'arbre phylogénétique des espèces.

En vert foncé : codons significativement préférés. En vert clair : codons préférés de manière non significative. En jaune : codons détestés de manière non significative. En orange : codons significativement détestés. Les préférences n'ont pas été représentées lorsque aucun des codons synonymes n'avait de préférence significative.

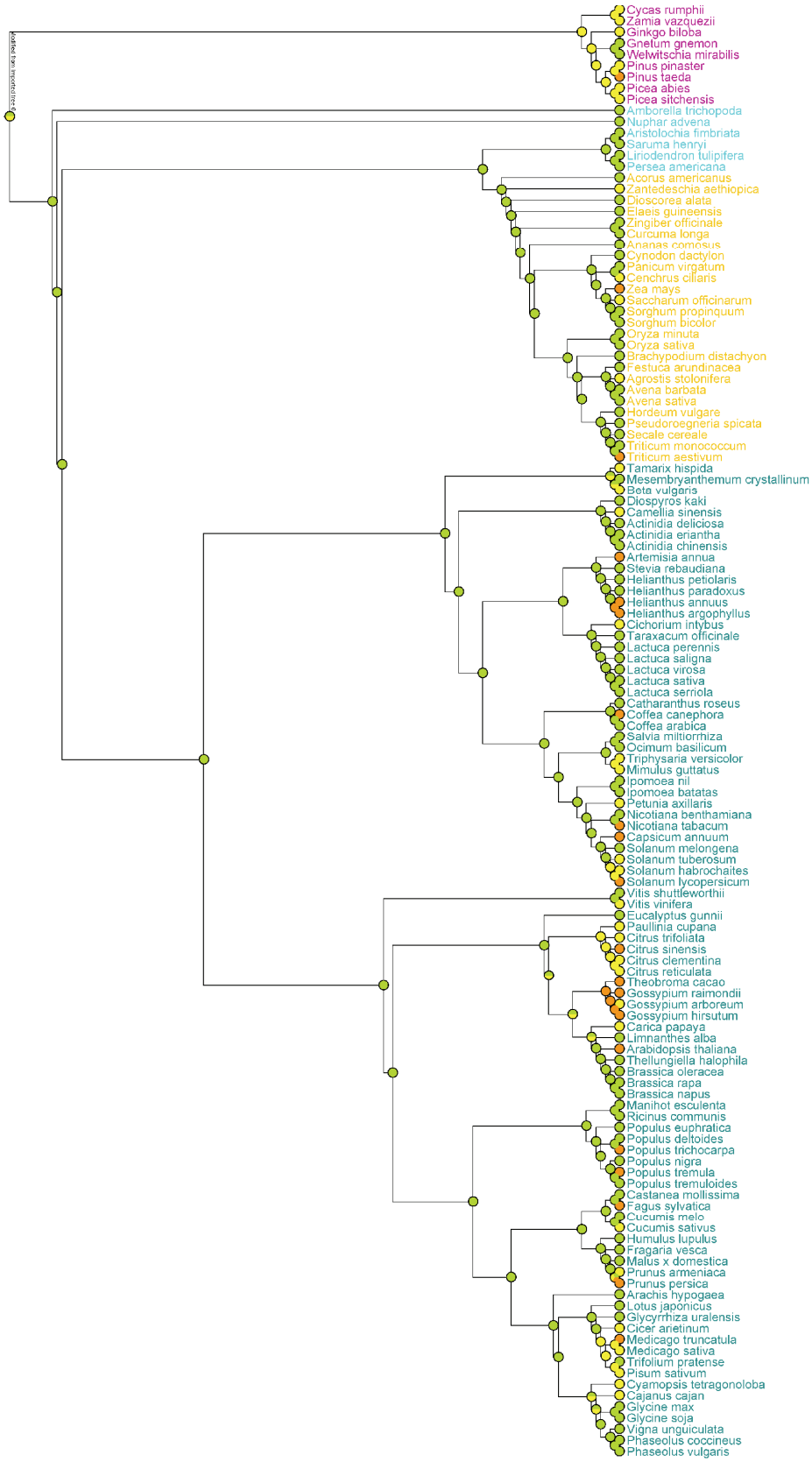


Figure 10 – Préférences ancestrales reconstruites pour le codon TCC de la sérine

Pastille verte : codon préféré. Pastille jaune : codon neutre. Pastille orange : codon détesté.

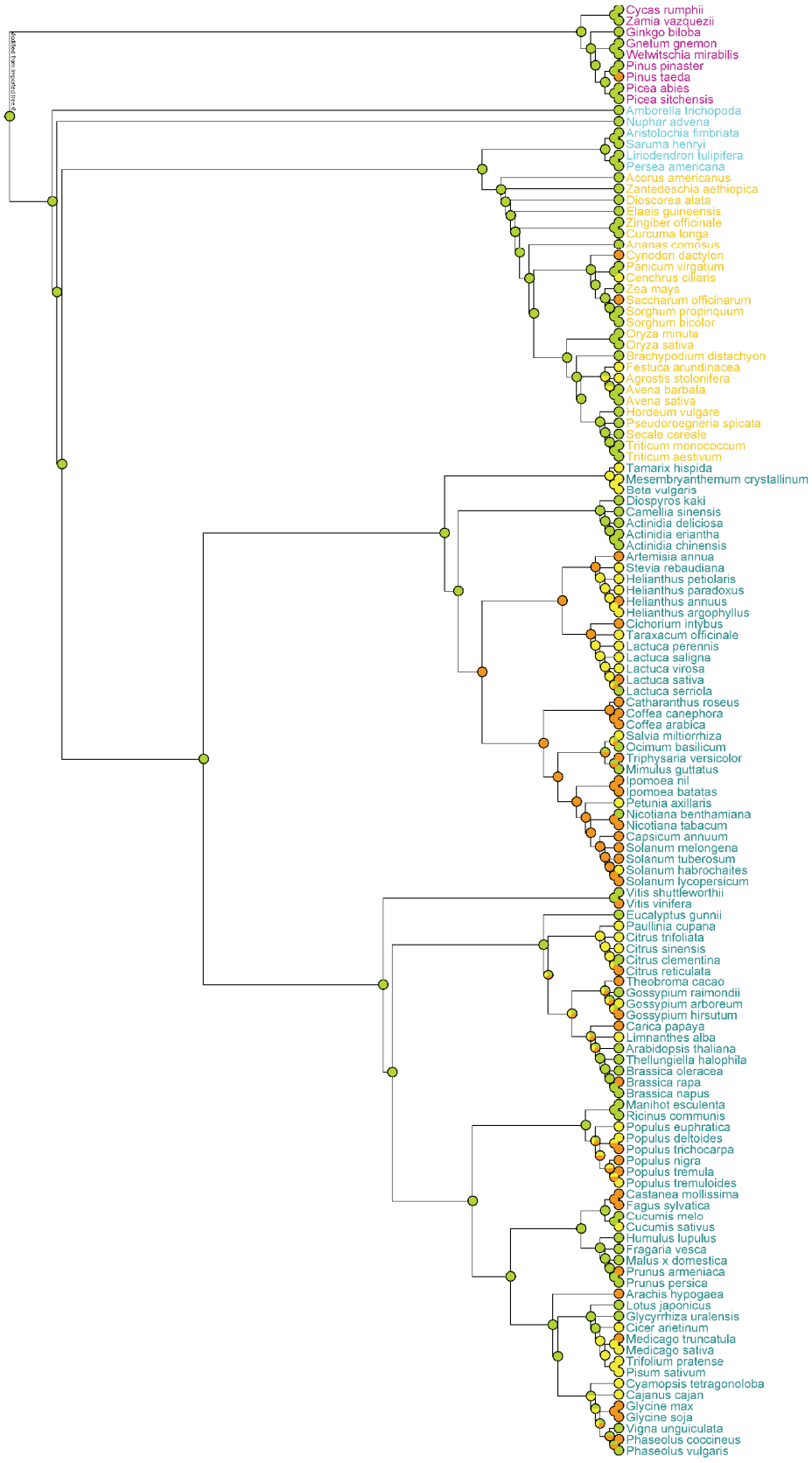


Figure 11 – Préférences ancestrales reconstruites pour le codon GCG de l'alanine

Pastille verte : codon préféré. Pastille jaune : codon neutre. Pastille orange : codon détesté.

Préférences d'usage du code et composition nucléotidique

Des points communs semblant apparaître entre les patrons de préférences des codons finissant par le même troisième nucléotide, nous allons chercher maintenant s'il émerge des patrons de préférence particuliers selon la composition nucléotidique des codons, s'ils sont en lien avec la composition nucléotidique des espèces, et quels peuvent être leurs effets l'un sur l'autre.

Adéquation entre préférences d'usage du code et composition nucléotidique

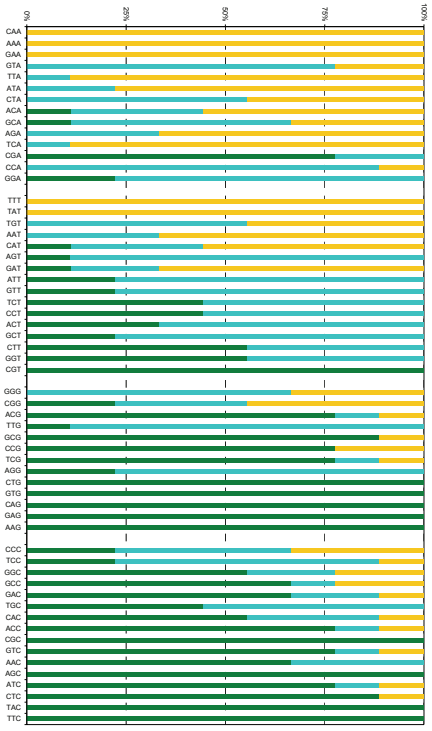
Si l'on considère les proportions de types de préférence (préféré, neutre, détesté) selon la composition nucléotidique des codons, il apparaît que globalement, sur l'ensemble des espèces, les codons finissant en G/C sont plus souvent préférés que ceux en A/T (en moyenne sur les 130 espèces, proportion de « préféré » pour les codons en G/C : 71,6 % ; pour les codons en A/T : 18,3 % ; test de Mann-Whitney-Wilcoxon : p-value = $4{,}0*10^{-8}$), alors que les codons en A/T sont plus souvent détestés (en moyenne sur les 130 espèces, proportion de « détesté » pour les codons en G/C : 10,2 % ; pour les codons en A/T : 47,3 % ; test de Mann-Whitney-Wilcoxon : p-value = $2{,}1*10^{-5}$). Des différences apparaissent également au sein de ces catégories de codon : les codons en C sont encore plus souvent préférés que ceux en G (en moyenne sur les 130 espèces, proportion de « préféré » pour les codons en C : 80,3 % ; pour les codons en G : 60,9 % ; test de Mann-Whitney-Wilcoxon : p-value = 0,1), et les codons en A sont plus souvent détestés que les codons en T (en moyenne sur les 130 espèces, proportion de « détesté » pour les codons en A : 57,3 % ; pour les codons en T : 38,4 % ; test de Mann-Whitney-Wilcoxon : p-value = 0,09).

Entre les quatre grands groupes phylogénétiques (gymnospermes, angiospermes basales, monocotylédones, eudicotylédones), ces patrons sont conservés mais accentués de manière différente (Figure 12). En particulier, chez les eudicotylédones la proportion d'espèces préférant les codons en A/T est plus forte (22,1 % chez les eudicotylédones contre 10 % chez les angiospermes basales et 7 % chez les monocotylédones) et la proportion d'espèces préférant les codons en G/C est plus basse (68,2 % chez les eudicotylédones contre 86,2 % chez les angiospermes basales et 80,6 % chez les monocotylédones). Chez les monocotylédones la proportion d'espèces détestant les codons en A/T est plus forte (60 % chez les monocotylédones contre 51,7 % chez les angiospermes basales et 43,9 % chez eudicotylédones), alors que les

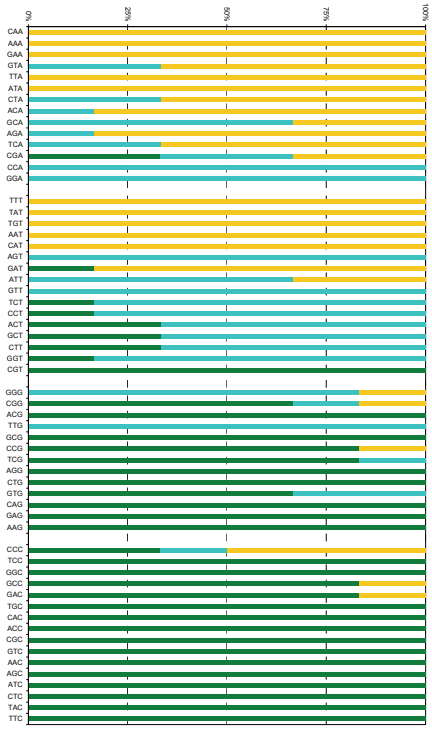
codons en G/C sont plus souvent détestés chez les eudicotylédones (11,9 % chez les eudicotylédones contre 4,6 % chez les angiospermes basales et 5,6 % chez les monocotylédones).

Cette modulation des patrons de préférences selon la composition nucléotidique des codons entre les grands groupes phylogénétiques semble coïncider avec les patrons de composition nucléotidique de ces groupes, les génomes des monocotylédones étant riches en GC alors que ceux des eudicotylédones sont riches en AT. Pour tenir compte de l'importante variabilité des patrons de composition nucléotidique au sein de ces groupes, il est nécessaire d'appréhender plus finement l'adéquation entre composition génomique et préférences d'usage du code en prenant en considération les GC propres des espèces.

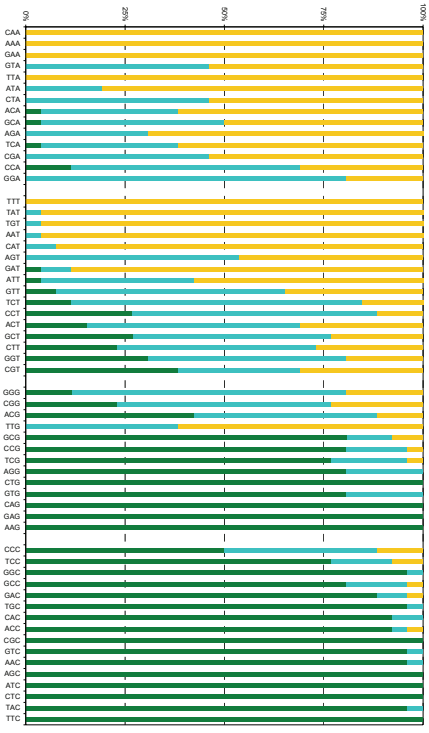
Qu'elles soient riches ou pauvres en GC, la quasi-totalité des espèces préfèrent en majorité des codons en G /C et détestent en majorité des codons en A/T (Figure 13). La proportion de codons préférés en G ou C est corrélée positivement au GC3 moyen des espèces (corrélation de Spearman : rho = 0,58, pvalue = $1,4*10^{-11}$), en particulier pour les gymnospermes, les angiospermes basales et les eudicotylédones, pauvres en GC (Figure 13 A). La proportion de codons détestés en G ou C est quant à elle corrélée négativement au GC3 moyen des espèces (corrélation de Spearman : rho = -0,81, pvalue = $1,8*10^{-7}$, Figure 13 B).



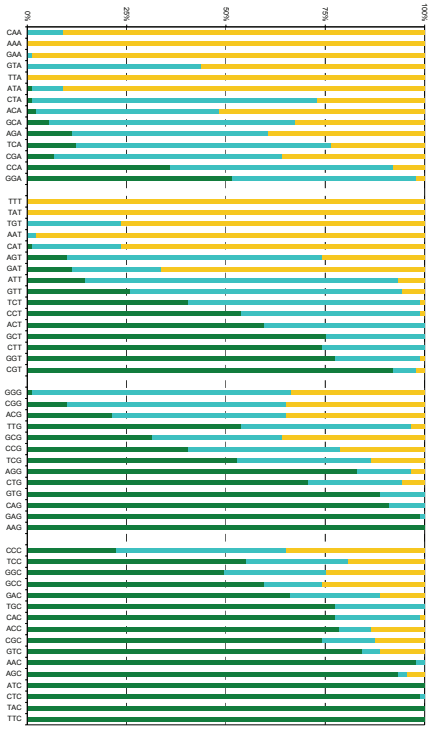
A. Gymnospermes



B. Angiospermes basales



C. Monocotylédones



D. Eudicotylédones

Figure 12 – Proportion des types de préférence par codon et par groupe d'espèces.

Les codons sont organisés selon leur troisième nucléotide : codons en A, puis en T, en G et en C. Les couleurs correspondent aux types de préférences. Vert : préféré. Bleu : neutre. Orange : détesté.

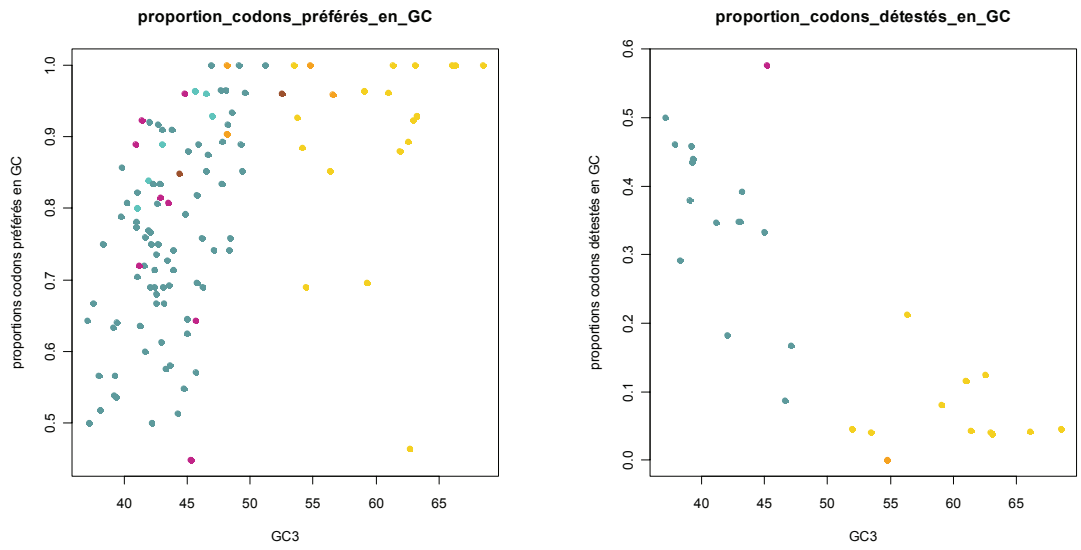


Figure 13 – Proportions de codons en G/C parmi les codons préférés et détestés, en fonction des GC3 moyens des espèces.

Seules les espèces possédant au moins 20 codons significativement préférés (**A**) ou détestés (**B**) sont représentées.

Les groupes d'espèces sont figurés par les couleurs suivantes : gymnospermes en fuchsia, angiospermes basales en turquoise, monocotylédones en jaune et eudicotylédones en bleu.

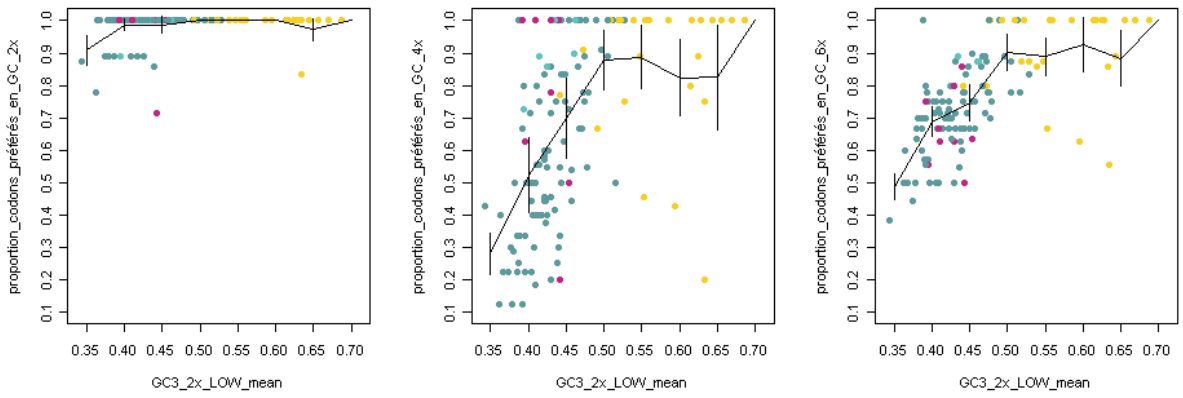
Effet de la composition nucléotidique sur les préférences d'usage du code ?

L'existence de cette corrélation entre préférences d'usage du code et composition nucléotidique soulève un problème de causalité : quels peuvent être les effets de la composition nucléotidique sur l'évolution des préférences d'usage du code, et réciproquement ?

Pour tenter de dénouer cette dépendance entre les préférences d'usage du code et la composition nucléotidique génomique, nous avons besoin d'un proxy du GC génomique non influencé par la SUC – un proxy du θ de Higgs et Ran (Higgs & Ran 2008). Avec nos données EST, nous ne disposons pas directement de GC génomique non codant, séquences d'introns ou séquences intergéniques. Nous disposons de séquences UTR, qui pourraient être pertinentes puisque non soumises à la SUC, mais leur qualité est insuffisante dans nos jeux de données : les UTR de nos unigènes EST sont souvent manquants ou très courts. Parmi les GC codants, le GC le plus proche de l'équilibre mutation/BGC est le GC3 des séquences les plus faiblement exprimées, les moins susceptibles de subir fortement la SUC. Les préférences étant très conservées pour les codons des acides aminés au code deux fois dégénéré, le GC3 de ces codons-là subit les effets de la SUC dans le même sens pour toutes les espèces. Parmi les GC dont nous disposons, le GC3 calculé sur les codons des acides aminés au code deux fois dégénéré des séquences faiblement exprimées (noté GC3_2x_LOW) est donc le moins affecté par la SUC et le moins différentiellement biaisé entre les espèces par les différences de préférences d'usage du code : le GC3_2x_LOW s'avère être le plus pertinent proxy du GC génomique attendu sans SUC.

La proportion de codons préférés en G/C parmi les codons des acides aminés au code deux, quatre et six fois dégénéré est positivement corrélée au GC3_2x_LOW pour l'ensemble des espèces, et de manière particulièrement forte pour les non-monocotylédones, dont le GC3_2x_LOW est inférieur à 50 % (corrélation de Spearman pour les codons 2x : rho = 0,2625, pvalue = 0,0025 ; corrélation de Spearman pour les codons 4x : rho = 0,5894, pvalue = $1,6*10^{-13}$; corrélation de Spearman pour les codons 6x : rho = 0,6427, pvalue < $2,2*10^{-16}$) (Figure 14). Cette corrélation présente moins d'ampleur pour les codons des acides aminés au code deux fois dégénéré, pour lesquels la variance de la proportion de codons préférés en G/C est très faible. Si la proportion de codons préférés en G/C parmi les codons des acides aminés au code six fois dégénéré est généralement supérieure à 50 %, la proportion de codons préférés en G/C parmi les codons des acides aminés au code quatre fois dégénéré est inférieure à 50 % pour les espèces pauvres en GC3_2x_LOW, et atteint même 12,5 % pour la luzerne tronquée, la pomme de terre et la tomate. Au-delà de 50 % de GC3_2x_LOW, les préférences sont très nettement

majoritairement en faveur des codons en G/C. Ces résultats tendent à montrer que les préférences d'usage du code sont adaptées au GC génomique, et que cette adaptation est particulièrement sensible pour les codons des acides aminés au code quatre fois dégénéré.



A **B** **C**

Figure 14 – Proportion de codons préférés en G/C pour les acides aminés au code deux (A), quatre (B) et six (C) fois dégénéré, en fonction du GC3 calculé pour les codons des acides aminés au code deux fois dégénéré des séquences faiblement exprimées.

Courbes de tendance en noir : moyennes (points) et écart-types (barres verticales) des proportions de codons préférés calculées pour des classes de GC3_2x_LOW de 0,05 de large (± 0,025 autour de chaque point représenté).

Les groupes d'espèces sont figurés par les couleurs suivantes : gymnospermes en fuchsia, angiospermes basales en turquoise, monocotylédones en jaune et eudicotylédones en bleu.

Effet des préférences d'usage du code sur la composition nucléotidique ?

Réciproquement, la composition nucléotidique doit être impactée par la SUC et les préférences d'usage du code. Pour évaluer les effets de la SUC et des préférences d'usage sur la composition nucléotidique et voir si l'intensité de ces effets diffère entre espèces et entre types de codons, nous avons comparé le GC3 des séquences les plus fortement exprimées (GC3_Xx_HIGH), celles qui subissent le plus fortement la SUC, au proxy du GC génomique (GC3_2x_LOW) (Figure 15).

Pour toutes les espèces, la SUC augmente de manière significative le GC3_2x_HIGH, GC3 des codons des acides aminés au code deux fois dégénéré (enrichissement moyen : 7 %) (Figure 15 A). Ces résultats sont en cohérence avec les proportions extrêmement fortes de préférence en faveur des codons en G/C trouvées pour les acides aminés au code deux fois dégénéré. L'effet positif de la SUC sur le GC3_2x est d'autant plus fort que les espèces sont riches en GC (corrélation de Spearman entre GC3_2x_LOW et « GC3_2x_HIGH - GC3_2x_LOW » : rho = 0,3746, pvalue = $1{,}3*10^{-5}$). Ce résultat révèle que la différence d'intensité de SUC entre séquences faiblement et fortement exprimées est plus forte chez les monocotylédones que chez les autres espèces, ce qui suggère une SUC plus intense chez les monocotylédones ou une différence des patrons d'expression.

Pour les codons des acides aminés au code quatre fois dégénéré, le GC3_4x_HIGH est inférieur au GC3_2x_LOW pour la majorité des espèces non-monocotylédones et pour une partie des monocotylédones (Figure 15 B). Cette différence est la plus marquée pour les espèces les plus pauvres en GC3 (corrélation de Spearman entre GC3_2x_LOW et « GC3_4x_HIGH - GC3_2x_LOW » : rho = 0,2826, pvalue = 0,0012) : la SUC diminue notablement le GC3 pour ces espèces, en lien avec leurs proportions de préférence en faveur de codons en G/C inférieures à 50 %.

Pour les codons des acides aminés au code six fois dégénéré, le GC3_6x_HIGH est le plus souvent supérieur au GC3_2x_LOW (Figure 15 C). Cette augmentation de GC3 due à la SUC semble plus forte chez les monocotylédones (corrélation de Spearman entre GC3_2x_LOW et « GC3_6x_HIGH - GC3_2x_LOW » : rho = 0,2700, pvalue = 0,0019), ce qui est toujours bien en rapport avec les proportions de préférence en faveur des codons en G/C pour ces codons.

Etant donné les très fortes proportions de préférence – jusqu'à 100 % – en faveur des codons en G/C trouvées pour les espèces les plus riches en GC pour les codons des acides

aminés au code quatre et six fois dégénéré, nous attendions que la SUC ait des effets aussi marqués pour ces codons que pour les codons des acides aminés au code deux fois dégénéré. Or le GC3_4x_HIGH est globalement plus bas que le GC3_2x_LOW, ce qui est surtout étonnant pour les monocotylédones, et les différences entre GC3_6x_HIGH et GC3_2x_LOW sont le plus souvent plus faibles qu'entre GC3_2x_HIGH et GC3_2x_LOW. Peut-être est-ce dû au fait que les gènes faiblement exprimés de nos données ne représentent pas parfaitement la classe de gène la moins exprimée du transcriptome, puisque leur expression a été suffisante pour qu'ils soient capturés lors du séquençage EST ? Le GC3_2x_LOW serait alors déjà marqué par la SUC et serait un mauvais proxy du GC génomique à l'équilibre BGC/mutation. Il est également possible que la pression de la SUC soit effectivement plus forte pour les codons des acides aminés au code deux fois dégénéré. Cette hypothèse est d'autant plus plausible si pour ces codons il n'y a qu'un seul ARNt pour les deux codons synonymes, ce qui accentuerait la différence de valeur sélective entre codons synonymes.

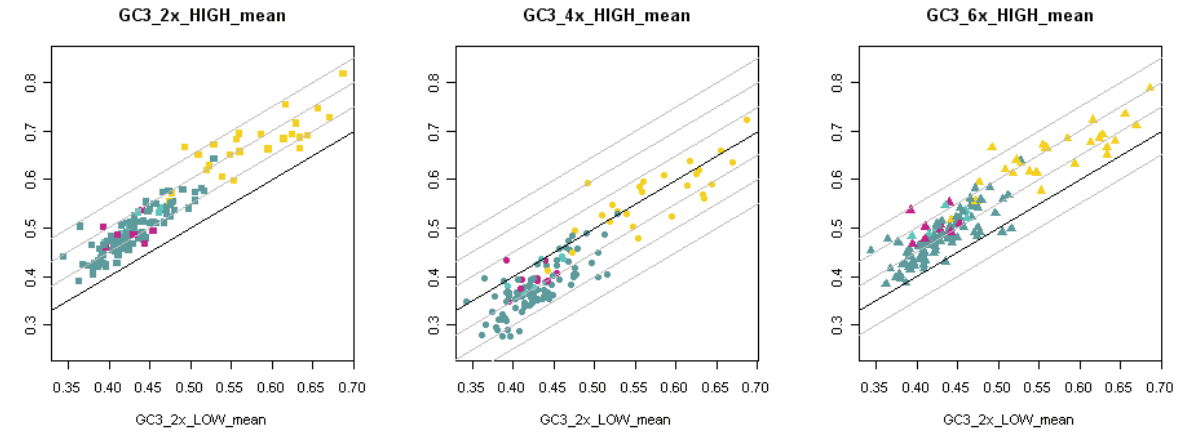


Figure 15 – GC3 calculés pour les codons des acides aminés au code deux (A), quatre (B) et six (C) fois dégénéré des séquences fortement exprimées (GC3_Xx_HIGH) en fonction des GC3 calculés pour les codons des acides aminés au code deux fois dégénéré des séquences faiblement exprimées (GC3_2x_LOW).

En noir : droites d'équation y = x. En gris : droites d'équation y = x ±0,05, y = x ±0,10 et y = x ±0,15.

Les groupes d'espèces sont figurés par les couleurs suivantes : gymnospermes en fuchsia, angiospermes basales en turquoise, monocotylédones en jaune et eudicotylédones en bleu.

Déterminer les préférences d'usage du code à partir de données EST

Les données de séquences EST (*Expressed Sequence Tag*) sont générées à partir de courts fragments d'ADNc obtenus par rétrotranscription d'ARNm, séquencés puis contigués pour reconstituer les séquences codantes. Plus un gène est exprimé, plus ses ARNm, présents dans la cellule en forte proportion par rapport aux ARNm des gènes moins exprimés, ont une forte probabilité d'être échantillonnés puis séquencés. Cette corrélation entre niveau d'expression des gènes et probabilité d'échantillonnage des EST permet d'utiliser le nombre d'EST s'appariant à chaque unigène (nombre de « hits ») comme un proxy du niveau d'expression (par exemple Duret & Mouchiroud 1999). Des bibliothèques de données de séquences EST sont aujourd'hui disponibles pour plus de 200 espèces de plantes mais ces bibliothèques ne sont pas de qualité homogène, avec notamment des différences quant aux types de tissus cellulaires utilisés, à la taille des bibliothèques, et à leur normalisation ou non. Dans quelle mesure ces caractéristiques peuvent-elles affecter la robustesse du proxy du niveau d'expression ? La représentativité des patrons d'expression est-elle suffisante pour déterminer de manière fiable les préférences d'usage du code à partir de données EST ?

Les tissus cellulaires à partir desquels ont été réalisés les séquençages EST diffèrent selon les bibliothèques. Les tissus peuvent provenir d'organes différents tels que les racines, les feuilles, les fleurs ou les fruits, isolés ou en mélange. Ils peuvent être prélevés à des stades de développement différents, stades de germination ou de fructification par exemple, ou dans des contextes particuliers, stress hydrique ou présence de pathogène par exemple. Ces bibliothèques sont généralement conçues de la sorte pour identifier les gènes particuliers impliqués dans les processus propres à ces contextes. Si pour ces gènes-là le patron d'expression est effectivement modifié par le contexte, la majeure part du transcriptome ne verrait cependant pas son expression impactée.

Pour les petites bibliothèques, la représentativité de l'échantillonnage est moins bonne : le biais d'échantillonnage en faveur des ARNm les plus nombreux est tel que les gènes peu exprimés sont sous-échantillonnés, voire non-échantillonnés. La classe de gènes la moins exprimée de la base de données EST ne correspond alors pas à la classe de gènes la moins exprimée du transcriptome, mais à une classe d'expression intermédiaire. Pour limiter ce risque,

nous avons filtré les bibliothèques disponibles pour éliminer les plus petites d'entre elles ainsi que celles avec peu de variance des niveaux d'expression.

Certaines bibliothèques de données EST ont subi une étape de normalisation qui vise à réduire l'écart entre les nombres d'occurrence des gènes faiblement et fortement exprimés. La normalisation, qui a lieu avant l'étape de séquençage, permet d'augmenter la probabilité de séquencer les gènes rarement exprimés et de limiter la surreprésentation des gènes les plus exprimés. Si la variance des niveaux d'expression est ainsi artificiellement réduite, l'ordre relatif des gènes ordonnés selon leur niveau d'expression est par contre peu affecté. Pour déterminer les préférences d'usage du code, les bibliothèques normalisées pourraient donc poser problème en terme de manque de puissance, mais ne devraient pas mener à faussement déterminer les préférences. Notre filtre des bibliothèques sur le critère de leur variance d'expression permet d'éliminer les bibliothèques pour lesquelles la normalisation paraît trop problématique.

Ces différentes caractéristiques de taille de base, de type et de contexte des tissus utilisés, et la normalisation ou non des bibliothèques, sont de plus bien distribuées dans les différents groupes phylogénétiques et ne devraient donc pas créer de signal phylogénétique particulier quant aux préférences d'usage du code.

L'estimation des niveaux d'expression à partir du nombre d'EST restant assez rudimentaire, il est attendu que les corrélations détectées à partir des EST entre usage du code et expression puissent être moins fortes que réellement. Nous avons donc préféré définir trois classes d'expression plutôt que de chercher des corrélations directement à partir des valeurs d'expression estimées, comme proposé par Duret & Mouchiroud (1999). Cette méthode est plus robuste aux imprécisions sur les niveaux d'expression estimés et au manque de puissance. Si des problèmes de manque de puissance persistent, ils peuvent mener à ne pas détecter de préférence significative d'usage pour un codon mais ne devraient pas mener à déterminer de fausses préférences. Dans les cas où nous n'avons trouvé aucune préférence d'usage, la cause peut donc être soit une absence réelle de préférence, soit un manque de puissance des données. Les intensités des corrélations entre usage du code et expression sont par contre trop dépendantes des caractéristiques des bibliothèques de données EST pour pouvoir être comparées entre espèces. Elles ne permettent donc pas d'étudier directement les variations d'intensité de la SUC entre espèces – une problématique autre que celle des variations des préférences d'usage du code.

Wang & Roossinck ont déterminé les préférences d'usage du code pour 11 espèces de plantes à partir de données EST avec une méthode similaire à celle que nous avons utilisée (Wang

& Roossinck 2006). En comparant leurs résultats aux nôtres, il s'avère que les préférences détectées sont très généralement les mêmes, en particulier pour les acides aminés au code deux fois dégénéré, mais que les préférences diffèrent pour certains acides aminés, par exemple pour la proline des monocotylédones – mais la proline montre une assez grande variabilité entre espèces dans les deux études. Ces points de différence semblent généralement attribuables à un problème de puissance. La fiabilité de nos résultats paraît suffisante pour étudier les patrons globaux de préférences d'usage du code mais ne permet probablement pas d'accorder assez de crédit aux variations les plus rares pour une interprétation fine et détaillée des préférences au niveau de chaque espèce.

Nous avons également comparé nos résultats aux préférences d'usage du code déterminées à partir de méthodes de classification des gènes d'après leurs biais d'usage du code (*Codon Adaptation Index* et *Effective Number of Codons*) (Kawabe & Miyashita 2003 ; Zhang et al. 2007 ; Liu et al. 2010). Pour la plupart des codons les préférences déterminées sont les mêmes, mais elles diffèrent pour certains. Ces méthodes n'utilisent pas de données d'expression pour déterminer les préférences d'usage. Elles se basent sur les différences de biais d'usage entre gènes pour distribuer les gènes en deux catégories : une catégorie de gènes à usage du code peu biaisé, censée correspondre à la catégorie des gènes peu exprimés et donc peu affectés par la SUC, et une catégorie de gènes à fort biais d'usage du code, censée correspondre à la catégorie des gènes fortement marqués par les effets de la SUC. Ces méthodes paraissent peu judicieuses en cas de génome riche et hétérogène en GC. En effet, les catégories de gènes à faible et à fort biais d'usage peuvent être dues à d'autres facteurs que l'expression, à la recombinaison par exemple. L'emploi de ces méthodes chez une espèce à forte BGC pourrait ainsi mener à détecter de faux codons préférés en G/C, et à ne pas détecter certains codons préférés en A/T. Pour le maïs, monocotylédone subissant vraisemblablement une forte BGC, les neuf différences de préférences entre les résultats obtenus par ces méthodes et nos résultats correspondent systématiquement à l'un ou l'autre de ces cas (Liu et al. 2010).

Variabilité et évolution des préférences d'usage du code

Nous avons montré que, chez les 130 espèces de plantes étudiées, l'usage de tous les codons est biaisé. Nos résultats indiquent que le principal facteur explicatif de ces biais d'usage du code est la composition nucléotidique globale des génomes. Leur composition nucléotidique est façonnée par des forces évolutives neutres, biais mutationnel et conversion génique biaisée vers G et C (BGC) vraisemblablement. La composition nucléotidique des régions intergéniques porte trace de leurs effets et permet déjà de prédire en partie les patrons globaux d'usage du code (Chen et al. 2004 ; Knight et al. 2001a). Nous avons montré que la SUC est un autre facteur explicatif de ces biais d'usage du code. Des arguments en faveur de la SUC avaient déjà été mis en évidence chez quelques espèces de plantes. Nous avons confirmé l'existence de SUC à grande échelle chez les plantes en révélant sa présence chez l'ensemble des espèces étudiées, couvrant un large éventail de gymnospermes et d'angiospermes.

Nous avons pu détecter des préférences d'usage des codons significatives pour la plupart des acides aminés chez les espèces étudiées. Ces préférences sont variables entre espèces. A partir des préférences actuelles nous avons reconstruit les préférences ancestrales d'usage du code sur la phylogénie et mis en évidence des patrons d'évolution différents selon les acides aminés : les préférences d'usage du code sont très conservées entre espèces pour les codons des acides aminés au code deux et trois fois dégénéré, alors que pour les codons des acides aminés au code quatre et six fois dégénéré les préférences d'usage sont plus labiles. Les changements de préférence au cours de l'évolution des plantes sont souvent récents, avec parfois même des changements de préférence au niveau intergénérique, mais peuvent aussi être localisés à des niveaux phylogénétiques plus profonds. Ils n'affectent pas préférentiellement certains groupes taxonomiques. Les changements de préférence des différents codons sont distribués de manière ponctuelle sur la phylogénie et ne semblent pas coïncider en des points de changement majeur des préférences. De précédentes études consacrées aux préférences d'usage du code au sein de groupes d'espèces proches, chez cinq nématodes (Mitreva et al. 2006) et chez douze espèces de drosophiles (Vicario et al. 2007) par exemple, avaient mis en évidence des variations de préférences plus ou moins récentes entre espèces, et la différence de variabilité des préférences d'usage des codons selon le degré de dégénérescence du code des acides aminés est également retrouvée pour leurs résultats. Les préférences d'usage du code évoluent donc fréquemment à l'échelle du vivant, tout particulièrement pour les codons des acides aminés au code quatre et six fois dégénéré.

D'après nos résultats, les préférences d'usage sont liées à la composition nucléotidique des codons, à leur troisième nucléotide en particulier : les codons préférés sont majoritairement des codons en G/C, alors que les codons en A/T sont plus souvent détestés. Et les codons en C sont plus préférés que ceux en G, tandis que les codons en A sont plus souvent détestés que ceux en T. Ce patron est globalement retrouvé pour les différents acides aminés et pour les grands groupes taxonomiques, mais les préférences sont accentuées en faveur des codons en G/C chez les monocotylédones : les eudicotylédones se démarquent par des préférences plus favorables aux codons en A/T. A l'échelle des espèces, les préférences en faveur des codons en G/C et au détriment des codons en A/T s'avèrent être positivement corrélées à la composition nucléotidique globale, estimée par le GC3. Cette corrélation entre composition génomique et préférences en faveur de codons en G/C a également été mise en évidence chez les bactéries, chez les archées et chez des eucaryotes, les champignons (Hershberg & Petrov 2009). Chez ces espèces tout comme chez les angiospermes et les gymnospermes que nous avons étudiées, il y a une même correspondance entre gamme de composition nucléotidique des espèces (GC intergénique ou GC3_2x_low, que nous avons utilisé comme proxy du GC sans SUC) et gamme de proportion de préférences en faveur de codons en G/C. Les espèces riches en GC (GC3_2x_low > 50 %) préfèrent ainsi majoritairement des codons en G/C et les espèces les plus riches en AT (GC3_2x_low < 40 %) préfèrent majoritairement des codons en A/T pour les acides aminés au code quatre et six fois dégénéré. Mais il existe une classe intermédiaire d'espèces pauvres en GC (gamme de GC3_2x_low : 40 à 50 %) pour lesquelles les préférences sont majoritairement en faveur de codons en G/C pour ces acides aminés, supérieures donc à ce qu'on attendrait si l'adéquation entre préférences et composition nucléotidique était optimisée. Ce décalage de préférences est également retrouvé chez la drosophile et chez *Caenorhabditis elegans* (Duret & Mouchiroud 1999) ainsi que chez quatre espèces de bivalves (Tanguy et al. 2008) : les préférences de ces espèces riches en AT sont nettement en faveur de codons en G/C.

Répertoire d'ARNt : variabilité et adéquation avec les préférences d'usage du code

Les préférences d'usage des codons sont directement liées à la population d'ARNt cellulaire : ce sont les espèces d'ARNt isoaccepteurs disponibles et leurs abondances relatives qui déterminent l'efficience de la traduction des codons synonymes, et donc le sens et l'intensité de la SUC. Les quantifications expérimentales des proportions d'ARNt cellulaires sont rares chez les eucaryotes. Pour les espèces dont le génome est entièrement séquencé il est cependant possible d'inférer *in silico* les gènes codants pour des ARNt et d'estimer à partir du nombre de copies de ces gènes les proportions relatives des ARNt cellulaires (Duret 2002). Chez les plantes, des données sur le nombre de gènes d'ARNt sont disponibles pour six espèces : deux monocotylédones, le riz et *Brachypodium distachyon*, et trois eudicotylédones, *Arabidopsis thaliana*, *Medicago truncatula* et *Populus trichocarpa*, ainsi qu'une chlorophyte, *Chlamydomonas reinhardtii* (Michaud et al. 2011). A partir de ces données, nous avons calculé une mesure du biais d'abondance des ARNt, le RGF (*Relative Gene Frequency of tRNA genes*, voir tableau 1 en annexe), et mettons en parallèle ces informations avec les préférences d'usage du code que nous avons détectées pour les cinq angiospermes (tableau 1 en annexe). Les répertoires d'ARNt de ces espèces sont-ils bien en lien avec les préférences que nous avons déterminées, et permettent-ils d'expliquer les patrons de variabilité que nous avons trouvés ?

Nous avons indiqué (tableau 1 en annexe) les codons reconnus par l'anticodon de chaque espèce d'ARNt présente, en distinguant les types d'appariements : appariement canonique de Watson-Crick, l'interaction la plus efficace, et appariement bancal (*wobble pairing*). Les types d'appariements bancals dépendent fortement des modifications post-transcriptionnelles des ARNt pré-matures, en particulier des modifications du premier nucléotide de l'anticodon. Ces modifications varient à l'échelle du vivant et sont difficilement prédictibles *in silico* (Knight et al 2001b). Elles ont fait l'objet d'études approfondies chez les procaryotes (par exemple Näsvall et al. 2007). Chez les plantes, les appariements bancals sont mal connus, faute notamment de données expérimentales précises sur les modifications post-transcriptionnelles des ARNt. Il semble cependant probable que les adénosines (A) en première position d'anticodon soient modifiées en inosines (I) (Percudani 2001) et que les uridines (U) en première position de l'anticodon soient modifiées en 5-méthoxycarbonylméthyl-uridines (mcm^5U), 5-carbamoylméthyl-uridines (ncm^5U) et leurs dérivés (Leihne et al. 2011). Nous avons donc utilisé ici les règles d'appariement bancal suivantes : G•U, U•G et I•C (*position bancale de l'anticodon* • *troisième position de codon*). Ces règles semblent communément admises chez les eucaryotes (Ikemura 1985 ; Guthrie & Abelson 1982 ; Percudani 2001 ; Johansson et al. 2008). Nous

indiquons également un autre type d'appariement bancal, I•A (Crick 1966), dont l'existence et l'importance chez les eucaryotes sont toujours discutées mais qui, s'il était bien confirmé, semblerait trop marginal pour être susceptible d'intervenir de manière significative dans le sens et l'intensité de la SUC (Percudani 2001).

Les répertoires d'ARNt sont variables entre espèces. L'ARNt de la sélénocystéine est propre à l'algue verte. L'ARNt portant l'anticodon TAG (leucine) semble avoir été perdu chez le peuplier et l'ARNt portant l'anticodon GAT (isoleucine) semble avoir été perdu indépendamment chez le riz et chez *Arabidopsis* et *Medicago*. Trois espèces d'ARNt (anticodons ATA, GGC, GGT) ne seraient présentes que chez une seule espèce d'après ces données. Ces variations des espèces d'ARNt sont possibles, mais pourraient également être attribuables à des erreurs de séquençage ou d'annotation. Les nombres mêmes des gènes d'ARNt ont également évolué, par des évènements de gain ou perte et de conversion génique. Chez les cinq angiospermes, près de la moitié des gènes sont actuellement présents en plus de dix copies, jusqu'à 70 copies pour l'ARNt de la tyrosine d'*Arabidopsis*, qui est pourtant un des acides aminés les moins fréquents. Les gènes d'ARNt sont les moins nombreux chez l'algue verte, en lien avec la petite taille de son génome. Pour une même espèce d'ARNt, le nombre de gènes est très variable entre espèces, y compris au sein des monocotylédones et des eudicotylédones – jusqu'à une différence du simple au sextuple pour l'ARNt portant l'anticodon TGG de la proline. Ces variations de nombre de gènes modifient les proportions relatives des ARNt isoaccepteurs, au point d'inverser parfois ARNt minoritaires et majoritaires. Ce sont ces évolutions du nombre des gènes d'ARNt qui entraînent des changements d'intensité de la SUC, voire des changements de préférence d'usage des codons.

Il est attendu que les proportions relatives des ARNt (RGF) et les préférences d'usage du code coïncident. Les préférences d'usage du code que nous avons détectées sont effectivement bien corrélées avec les RGF (tableau 1 en annexe). Généralement les codons préférés sont ceux qui ont un appariement de type Watson-Crick avec les espèces d'ARNt à fort RGF, et les codons détestés correspondent aux espèces d'ARNt à faible RGF. Pour certaines espèces d'ARNt à fort RGF, c'est le codon en appariement bancal qui est préféré, alors que le codon en appariement Watson-Crick est neutre ou détesté. Pour quelques cas les préférences détectées ne correspondent pas aux attendus. Peut-être est-ce dû à un manque de puissance de nos données, ou à des erreurs de séquençage ou d'annotation des ARNt. L'annotation des ARNt est compliquée par la présence de nombreux pseudogènes et d'insertions de séquences d'ARNt chloroplastiques et mitochondriales. De plus, les répertoires d'ARNt estimés à partir de nombres

de gènes d'ARNt inférés *in silico* ne correspondent pas avec exactitude aux ARNt cellulaires. D'une part, l'expression de ces gènes d'ARNt peut être modulée. Chez la bactérie *Escherichia coli* par exemple l'expression des ARNt est fonction de la vitesse de croissance (Dong 1996). D'autre part il y a une incertitude quant aux modifications post-transcriptionnelles des ARNt chez les plantes, susceptibles d'affecter les capacités d'appariement bancal des ARNt.

Parmi les neuf acides aminés au code deux fois dégénéré, les six acides aminés de type C/T ne disposent que d'une seule espèce d'ARNt, l'ARNt portant l'anticodon en G (tableau 1 en annexe). La forte conservation des préférences en faveur des codons en C pour ces acides aminés à l'échelle des angiospermes serait donc expliquée par la difficulté d'acquérir une nouvelle espèce d'ARNt. Pour une espèce, la perte complète des gènes d'un type d'ARNt a un effet cliquet. Pour les codons des acides aminés au code quatre et six fois dégénéré, la plus grande labilité des préférences d'usage du code est également éclairée par les répertoires de gènes ARNt. Il existe en effet toujours trois espèces d'ARNt pour les acides aminés au code quatre fois dégénéré, et quatre ou cinq espèces d'ARNt pour les acides aminés au code six fois dégénéré. Le sens et l'intensité de la SUC sont donc plus rapidement modifiables, par des mutations affectant simplement le nombre de copies de ces gènes ARNt déjà présents.

L'évolution des préférences affecte-t-elle l'évolution du GC ?

Nous avons montré que les préférences d'usage du code sont liées à la composition nucléotidique des codons et sont le plus souvent en faveur des codons en G/C : il est donc attendu que la SUC, en augmentant la fréquence de ces codons préférés, ait un impact sur le GC3. Pour les codons des acides aminés au code deux fois dégénéré, pour lesquels les codons préférés sont très majoritairement, sinon exclusivement, les codons en G/C, nos résultats montrent effectivement que la SUC a pour effet d'augmenter le GC3 des séquences les plus exprimées de l'ordre de 3 à 15 % pour la plupart des espèces. Les résultats sont moins clairs pour les codons des acides aminés au code quatre et six fois dégénéré. Pour les monocotylédones qui préfèrent 100 % de codons en G/C pour ces acides aminés, le GC3 devrait être augmenté par la SUC, de la même manière que pour les acides aminés au code deux fois dégénéré. Pour les autres espèces, notamment la plupart des eudicotylédones, pour lesquelles les préférences s'orientent à la fois en faveur de codons en G/C et de codons en A/T, la SUC pourrait avoir pour effet global d'augmenter ou de diminuer le GC3, selon le type de préférences et l'intensité de la SUC des différents acides aminés. La SUC agit certes au second ordre sur la détermination de l'usage du code, derrière les forces neutres qui façonnent les patrons de composition nucléotidique des génomes, mais elle a tout de même un effet certain sur le GC3 pour les espèces préférant très fortement les codons en G/C.

Les préférences d'usage du code varient entre eudicotylédones et monocotylédones, en particulier du point de vue de la composition nucléotidique des codons préférés, tandis que les paysages nucléotidiques des eudicotylédones et des monocotylédones diffèrent. Comme nous l'avons montré dans le chapitre précédent, les génomes des eudicotylédones sont généralement pauvres et plutôt homogènes en GC, alors que les monocotylédones, et particulièrement les Poacées, ont des génomes plus riches et hétérogènes en GC. Nous avons suggéré que ces patrons aient été façonnés par la BGC. La SUC contribue-t-elle aussi à ces différences de patron de GC entre eudicotylédones et monocotylédones ? Par essence, la SUC n'affecte pas les régions non codantes, contrairement à la mutation et à la BGC ; parmi les GC codants, la SUC ne peut affecter que le GC3 et, de manière plus marginale, le GC1. La SUC ne peut donc pas être une explication majeure des contrastes des patrons de GC observés, mais il est probable qu'elle puisse y contribuer dans une certaine mesure. En effet, chez les espèces qui préfèrent exclusivement des codons en G/C, certaines Poacées par exemple, la SUC agit toujours dans le sens d'un enrichissement en GC3 et accentue ainsi vraisemblablement les patrons de GC dus aux forces évolutives neutres : en augmentant le GC3 des gènes les plus exprimés, elle augmente à la fois la

richesse et l'hétérogénéité en GC de ces espèces, et irait donc dans le même sens que la BGC. Chez les autres espèces, la plupart des eudicotylédones par exemple, les préférences pour certains acides aminés sont en faveur des codons en A/T : pour ces codons-là la SUC a pour effet de diminuer le GC3, et s'oppose ainsi à la BGC. Cependant, même chez ces espèces, les codons préférés restent majoritairement des codons en G/C, notamment pour les acides aminés au code deux fois dégénéré. La manière dont se compensent les effets positifs et négatifs de la SUC sur le GC3 est donc incertaine. Quantifier l'intensité de la SUC pour les différents aminés permettrait de mieux estimer l'effet potentiel de la SUC sur le GC3 de ces espèces, et d'évaluer s'il reste possible, quoique cela semble peu probable, que la SUC y amenuise les effets de la BGC.

Les effets de la SUC sur la composition nucléotidique dépendent également de l'intensité de la SUC. Les pressions de sélection sur l'usage du code sont susceptibles de varier entre acides aminés – selon les proportions relatives des ARNt – mais aussi entre espèces, selon les patrons d'expression et les paramètres populationnels affectant l'intensité des forces de sélection. Des changements importants d'intensité de SUC entre acides aminés et entre espèces pourraient modifier fortement l'impact de la SUC sur la composition nucléotidique, sans nécessairement impliquer de changements de préférences. A l'échelle intragénérique déjà, au sein du genre *Populus,* des différences notables d'intensité de la SUC entre espèces ont été mises en évidence (Ingvarsson 2008).

L'évolution du GC affecte-t-elle l'évolution des préférences ?

Si les préférences d'usage du code étaient établies de manière aléatoire au cours de l'évolution, par un processus de type *frozen accidents* par exemple, on n'attendrait *a priori* pas de correspondance entre les préférences d'usage du code et la composition nucléotidique des génomes. Or nous avons mis en évidence une telle corrélation, ce qui accrédite l'hypothèse d'une coévolution entre les forces affectant la composition nucléotidique et l'usage du code. Des modèles ont été consacrés à cette théorie de coévolution entre ARNt et composition nucléotidique (Bulmer 1987 ; Higgs & Ran 2008). Le modèle de Higgs & Ran prédit qu'il existe différents états stables possibles de coadaptation entre le répertoire d'ARNt isoaccepteurs et la composition nucléotidique à l'équilibre mutationnel, et qu'ils sont de plus affectés par l'histoire de l'évolution de la composition génomique (augmentation ou diminution graduelle du GC : voir Figure 1). Ces prédictions permettent d'expliquer le décalage de la corrélation entre composition nucléotidique et préférences d'usage en faveur de G/C pour la gamme d'espèces pauvres en GC.

Nos précédents résultats ont suggéré que la BGC serait largement présente chez les plantes et que son intensité ait pu avoir augmenté au cours de l'évolution chez certains groupes d'espèces, notamment chez les Poacées. A l'instar de la mutation, la BGC est une force évolutive neutre agissant sur la composition nucléotidique – mais en sens inverse du biais mutationnel, orienté vers A/T chez les plantes. Il faudrait donc considérer un modèle de coévolution entre d'une part les ARNt et d'autre part la composition nucléotidique dictée par l'action conjointe de la mutation et de la BGC. Avec un tel modèle il est attendu, sous l'hypothèse d'une intensification de la BGC au cours de l'évolution des plantes, que la coévolution entre ARNt et usage du code soit affectée. D'après nos résultats, la proportion de préférences en faveur de codons en G/C augmente conjointement avec notre proxy du GC à l'équilibre BGC/mutation. Il semblerait donc bien que la possible évolution de l'intensité de la BGC ait orienté la coévolution entre ARNt et usage du code, aboutissant, à travers son effet sur la composition nucléotidique, à l'établissement de préférences fortement en faveur de G/C chez les espèces les plus fortement marquées par la BGC.

Nous pourrions alors esquisser un scénario hypothétique d'évolution de la composition nucléotidique, des préférences d'usage du code et de la BGC chez les plantes. D'après nos reconstructions, les préférences ancestrales à la base des angiospermes, des monocotylédones et des eudicotylédones étaient majoritairement en faveur de codons en G/C, proches des patrons de préférences des angiospermes basales actuelles. Les paysages nucléotidiques de ces génomes

ancestraux n'ont pas encore été inférés avec précision mais nous pouvons avancer par parcimonie qu'ils étaient probablement plutôt pauvres et homogènes en GC, de type *Amborella trichopoda* (voir chapitre précédent). A partir de ces états ancestraux, la composition nucléotidique, la BGC et les préférences d'usage du code auraient ensuite évolué chez les monocotylédones et les eudicotylédones en suivant deux voies différentes.

Au cours de l'évolution des monocotylédones, la BGC se serait intensifiée, avec pour effet d'enrichir en GC les paysages nucléotidiques et par là même d'accentuer la contrainte pesant sur les préférences d'usage du code pour que la SUC favorise plus fortement les codons en G/C. Les préférences auraient ainsi évolué : les codons en G/C ont été encore plus souvent préférés et moins souvent détestés, tandis que le nombre de codons détestés en A/T a augmenté. Une telle coévolution a pu mener aux patrons extrêmes trouvés chez certaines Poacées, avec des préférences exclusivement en faveur de codons en G/C, et des génomes extrêmement hétérogènes et riches en GC, au point de dépasser les 60 % de GC3 moyen.

A l'inverse, au cours de l'évolution des eudicotylédones, la BGC aurait peu évolué, et aurait même probablement diminué d'intensité pour certains groupes dont le GC s'est appauvri, comme les Ranunculales (néanmoins pour d'autres groupes d'eudicotylédones, comme les Myrtales, la BGC se serait plus vraisemblablement accentuée, et il convient alors plutôt d'invoquer pour ces espèces un scénario du type de celui proposé pour les monocotylédones). En l'absence d'intensification de la BGC, la composition nucléotidique serait restée proche de la composition ancestrale. Dans ce contexte génomique pauvre en GC, les pressions de sélection pour préférer les codons en G/C aurait été moindres par rapport à celles probablement subies par les monocotylédones. Des préférences en faveur de codons en A/T ont ainsi pu remplacer des préférences en faveur de G/C. Le fait de détester des codons en G/C aurait été moins contre-sélectionné que chez les monocotylédones, et le nombre de codons détestés en G/C a pu augmenter tandis le nombre de codons détestés en A/T a diminué. Sous ces contraintes moins fortes, des états stables différents peuvent être fixés selon les espèces et les acides aminés. Ce scénario permettrait ainsi d'expliquer la forte variabilité des préférences d'usage du code au sein des eudicotylédones pour les acides aminés au code quatre et six fois dégénéré. Les préférences pour les acides aminés au code deux fois dégénéré ont par contre été très conservées, vraisemblablement à cause d'une contrainte forte : l'existence d'une seule espèce d'ARNt pour la plupart de ces acides aminés chez les plantes.

Les états actuels de biais d'usage du code et de préférences d'usage du code chez les plantes seraient donc très vraisemblablement le fruit d'une coévolution entre ARNt et composition nucléotidique, coévolution probablement orientée par l'évolution de la BGC – il est également possible que l'intensité même de la SUC ait aussi évolué en lien avec cette coévolution. Ces caractéristiques génomiques, ces forces évolutives et leurs effets sur les génomes sont très dépendantes d'autres éléments tels que le biais mutationnel, les patrons de recombinaison et les patrons d'expression… dont on sait qu'ils évoluent, eux aussi. Un modèle de coévolution encore plus complexe est donc à envisager pour mieux comprendre l'orogenèse des paysages nucléotidiques si contrastés du monde des plantes.

Akashi, H., 1997. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene*, 205(1-2), pp.269-78.

Akashi, H., 2001. Gene expression and molecular evolution. *Current Opinion in Genetics & Development*, 11, pp.660-666.

Bailey, C.D., Koch, M.A., Mayer, M., Mummenhoff, K., O'Kane Jr, S.L., Warwick, S.I., Windham, M.D., et al., 2006. Toward a Global Phylogeny of the Brassicaceae. *Molecular Biology*, 23(11), pp.2142-2160.

Bennetzen, J.L. & Hall, B.D., 1981. Codon Selection in Yeast. *The Journal of biological chemistry*, 257(6), pp.3026-3031.

Bouchenak-Khelladi, Y., Salamin, N., Savolainen, V., Forest, F., Bank, M.V.D., Chase, M.W., & Hodkinson, T.R., 2008. Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. *Molecular phylogenetics and evolution*, 47(2), pp.488-505.

Bulmer, M., 1987. Coevolution of codon usage and transfer RNA abundance. *Nature*, 325, pp.728-730.

Bulmer, M., 1991. The Selection-Mutation-Drift Theory of Synonymous Codon Usage. *Genetics*, 129, pp.897-907.

Burleigh, J.G. & Mathews, S., 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American journal of botany*, 91(10), pp.1599-1613.

Campbell, W.H. & Gowri, G., 1990. Codon usage in higher plants, green algae, and cyanobacteria. *Plant physiology*, 92(1), pp.1-11.

Chase, M.W., Fay, M.F., Reveal, J.L., Soltis, D.E., Soltis, P.S., Peter, F., Anderberg, A.A., et al., 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*, 161, pp.105-121.

Chat, J., Jauregui, B., Petit, R., & Nadot, S., 2004. Reticulate evolution in kiwifruit (*Actinidia*, Actinidiaceae) identified by comparing their maternal and paternal phylogenies. *American journal of botany*, 91(5), pp.736-747.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., & McAdams, H.H., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10), pp.3480-5.

Chiapello, H., Lisacek, F., Caboche, M., & Hénaut, A., 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. *Gene*, 209(GC1-GC38).

Childs, K.L., Hamilton, J.P., Zhu, W., Ly, E., Cheung, F., Wu, H., Rabinowicz, P.D., et al., 2007. The TIGR Plant Transcript Assemblies database. *Nucleic Acids Research*, 35(Database issue), D846-D851.

Crick, F.H., 1966. Codon–anticodon pairing: the wobble hypothesis. *Journal of molecular biology*, 19(2), pp.548-55.

Cronn, R.C., Small, R.L., Haselkorn, T., & Wendel, J.F., 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *American journal of botany*, 89(4), pp.707-725.

Cutter, A.D., Wasmuth, J.D. & Washington, N.L., 2008. Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. *Genetics*, 178(4), pp.2093-104.

Cuénoud, P., Savolainen, V., Chatrou, L.W., Powell, M., Grayer, R.J., & Chase, M.W., 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. *American journal of botany*, 89(1), pp.132-144.

Dray, S., Dufour, A.B. & Chessel, D., 2007. The ade4 package-II: Two-table and K-table methods. *R News*, 7(2), pp.47-52.

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development*, 12(6), pp.640-9.

Duret, L., 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends in genetics: TIG*, 16(7), pp.287-9.

Duret, L. & Mouchiroud, D., 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8), pp.4482-7.

Duret, L. & Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual review of genomics and human genetics*, 10, pp.285-311.

Escobar, J.S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J., & Glémin, S., 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). *Evolution; international journal of organic evolution*, 64(10), pp.2855-72.

Fennoy, S.L. & Bailey-Serres, J., 1993. Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C and G-ending codons. *Nucleic acids research*, 21(23), pp.5294-5300.

Forment, J., Gilabert, F., Robles, A., Conejero, V., Nuez, F., & Blanca, J.M., 2008. EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. *BMC bioinformatics*, 9(5).

Froelicher, Y., Mouhaya, W., Bassene, J.-B., Costantino, G., Kamiri, M., Luro, F., Morillon, R., et al., 2011. New universal mitochondrial PCR markers reveal new information on maternal citrus phylogeny. *Tree Genetics & Genomes*, 7, pp.49-61.

Glémin, S., Bazin, E., & Charlesworth, D., 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc. R. Soc. B.*, 273(1604), pp.3011-9.

Goertzen, L.R., Cannone, J.J., Gutell, R.R., & Jansen, R.K., 2003. ITS secondary structure derived from comparative analysis: implications for sequence alignment and phylogeny of the Asteraceae. *Molecular Phylogenetics and Evolution*, 29, pp.216-234.

Goldberg, E.E., Kohn, J.R., Lande, R., Robertson, K.A., Smith, S.A., & Igic, B., 2010. Species selection maintains self-incompatibility. *Science*, 330, pp.493-495.

Gordo, I. & Charlesworth, B., 2001. Genetic linkage and molecular evolution. *Current biology: CB*, 11(17), pp.R684-6.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pavé, A., 1980a. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1), pp.49-62.

Grantham, R., Gautier, C., & Gouy, M., 1980b. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, 8(9), pp.1893-1912.

Guo, X., Bao, J. & Fan, L., 2007. Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes. *FEBS letters*, 581(5), pp.1015-21.

Guthrie, C. & Abelson, J., 1982. Organization and expression of tRNA genes in *Saccharomyces cerevisiae*. In J.N. Strathern, E.W. Jones, & J.R. Broach, eds. *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory, pp. 487-528.

Hall, J.C., Sytsma, K.J. & Iltis, H.H., 2002. Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. *American journal of botany*, 89(11), pp.1826-1842.

Hamzeh, M. & Dayanandan, S., 2004. Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast trnT-trnF region and nuclear rDNA. *American journal of botany*, 91(9), pp.1398-1408.

Haudry, A., Cenci, A., Guilhaumon, C., Paux, E., Poirier, S., Santoni, S., David, J., et al., 2008. Mating system and recombination affect molecular evolution in four Triticeae species. *Genetics research*, 90(1), pp.97-109.

Hershberg, R. & Petrov, D.A., 2009. General rules for optimal codon choice. *PLoS genetics*, 5(7), p.e1000556.

Hershberg, R. & Petrov, D.A., 2008. Selection on codon bias. *Annual review of genetics*, 42(IV), pp.287-99.

Higgs, P.G. & Ran, W., 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Molecular biology and evolution*, 25(11), pp.2279-91.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), pp.13-34.

Ikemura, T., 1981. Correlation between the abundance of transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the translational system. *Journal of Molecular Biology*, 151(3), pp.389-409.

Ingvarsson, P.K., 2008. Molecular evolution of synonymous codon usage in *Populus*. *BMC evolutionary biology*, 8(307).

Johansson, M.J.O., Esberg, A., Huang, B., Björk, G.R., & Byström, A.S., 2008. Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Molecular and cellular biology*, 28(10), pp.3301-12.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., & Ikemura, T., 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of molecular evolution*, 53(4-5), pp.290-8.

Kawabe, A. & Miyashita, N.T., 2003. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes & genetic systems*, 78(5), pp.343-52.

Kimura, M., 1981. Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. *Proceedings of the National Academy of Sciences of the United States of America*, 78(9), pp.5773-7.

Knight, R.D., Freeland, S.J. & Landweber, L.F., 2001a. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome biology*, 2(4).

Knight, R.D., Freeland, S.J. & Landweber, L.F., 2001b. Rewiring the keyboard: evolvability of the genetic code. *Nature reviews. Genetics*, 2(1), pp.49-58.

Koopman, W.J.M., Zevenbergen, M.J. & Van den Berg, R.G., 2001. Species relationships in *Lactuca* S.L. (Lactuceae, Asteraceae) inferred from AFLP fingerprints. *American journal of botany*, 88(10), pp.1881-1887.

Kurland, C.G., 1991. Codon bias and gene expression. *FEBS letters*, 285(2), pp.165-9.

Leihne, V., Kirpekar, F., Vågbø, C.B., van den Born, E., Krokan, H.E., Grini, P.E., Meza, T.J., et al., 2011. Roles of Trm9- and ALKBH8-like proteins in the formation of modified wobble uridines in *Arabidopsis* tRNA. *Nucleic acids research*, 39(17), pp.7688-701.

Liu, H., He, R., Zhang, H., Huang, Y., Tian, M., & Zhang, J., 2010. Analysis of synonymous codon usage in *Zea mays*. *Molecular biology reports*, 37(2), pp.677-84.

Lobry, J.R. & Necşulea, A., 2006. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene*, 385, pp.128-36.

Lynch, M., 2007. *The Origins of Genome Architecture*, Sunderland, USA: Sinauer Associates Inc.

Lynn, D.J., Singer, G.A.C. & Hickey, D.A., 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic acids research*, 30(19), pp.4272-7.

Maddison, W. & Maddison, D., 2005. Mesquite: A modular system for evolutionary analysis. *Evolution*, 62(5), pp.1103-1118.

Marais, G., Mouchiroud, D. & Duret, L., 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10), pp.5688-92.

Marais, G. & Piganeau, G., 2002. Hill-Robertson interference is a minor determinant of variations in codon bias across *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. *Molecular biology and evolution*, 19(9), pp.1399-406.

Michaud, M., Cognat, V., Duchêne, A.-M., & Maréchal-Drouard, L., 2011. A global picture of tRNA genes in plant genomes. *The Plant Journal*, 66(1), pp.80-93.

Mitreva, M., Wendl, M.C., Martin, J., Wylie, T., Yin, Y., Larson, A., Parkinson, J., et al., 2006. Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. *Genome Biology*, 7(R75).

Morton, B.R. & Wright, S.I., 2007. Selective constraints on codon usage of nuclear genes from *Arabidopsis thaliana*. *Molecular biology and evolution*, 24(1), pp.122-9.

Murray, E., Lotzer, J. & Eberle, M., 1989. Codon usage in plant genes. *Nucleic acids research*, 17(2), pp.477-498.

Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H., & Bernardi, G., 2001. Translational selection on codon usage in *Xenopus laevis*. *Molecular biology and evolution*, 18(9), pp.1703-7.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., & Glémin, S., 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Molecular biology and evolution*.

Nagylaki, T., 1983. Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 80(20), pp.6278-81.

Näsvall, S.J., Chen, P. & Björk, G.R., 2007. The wobble hypothesis revisited: uridine-5-oxyacetic acid is critical for reading of G-ending codons. *RNA*, 13(12), pp.2151-64.

Olmstead, R.G., Bohs, L., Migid, H. A., Santiago-Valentin, E., Garcia, V.F., & Collier, S.M., 2008. A molecular phylogeny of the Solanaceae. *Taxon*, 57(4), pp.1159-1181.

Percudani, R., 2001. Restricted wobble rules for eukaryotic genomes. *Trends in genetics*, 17(3), pp.133-135.

Qiu, S., Bergero, R., Zeng, K., & Charlesworth, D., 2011a. Patterns of codon usage bias in *Silene latifolia*. *Molecular biology and evolution*, 28(1), pp.771-780.

Qiu, S., Zeng, K., Slotte, T., Wright, S., & Charlesworth, D., 2011b. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. *Genome biology and evolution*, 3, pp.868-880.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing.

Schonenberger, J., Anderberg, A.A. & Sytsma, K.J., 2005. Molecular phylogenetics and patterns of floral evolution in the Ericales. *Int. J. Plant Sci.*, 166(2), pp.265-288.

Schäferhoff, B., Fleischmann, A., Fischer, E., Albach, D.C., Borsch, T., Heubl, G., & Müller, K.F., 2010. Towards resolving Lamiales relationships: insights from rapidly evolving chloroplast sequences. *BMC Evolutionary Biology*, 10(352).

Sharp, P.M, Tuohy, T.M. & Mosurski, K.R., 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic acids research*, 14(13), pp.5125-43.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., & Wright, F., 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic acids research*, 16(17), pp.8207-8211.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., & Sockett, R.E., 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic acids research*, 33(4), pp.1141-53.

Shields, D.C., Sharp, P.M., Higgins, D.G., & Wright, F., 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Molecular biology and evolution*, 5(6), pp.704-16.

Stenico, M., Lloyd, A.T. & Sharp, P.M., 1994. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic acids research*, 22(13), pp.2437-46.

Stoletzki, N. & Eyre-Walker, A., 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Molecular biology and evolution*, 24(2), pp.374-81.

Tanguy, A., Bierne, N., Saavedra, C., Pina, B., Bachère, E., Kube, M., Bazin, E., et al., 2008. Increasing genomic information in bivalves through new EST collections in four species: development of new genetic markers for environmental studies and genome evolution. *Gene*, 408(1-2), pp.27-36.

Timme, R.E., Simpson, B.B. & Randal, L.C., 2007. High-resolution phylogeny for *Helianthus* (Asteraceae) using the 18S-26S ribosomal DNA external transcribed spacer. *Taxon*, 94(11), pp.1837-1852.

Vicario, S., Moriyama, E.N. & Powell, J.R., 2007. Codon usage in twelve species of *Drosophila*. *BMC evolutionary biology*, 7(226).

Wang, H.-C. & Hickey, D.A., 2007. Rapid divergence of codon usage patterns within the rice genome. *BMC evolutionary biology*, 7 Suppl 1, S6.

Wang, L. & Roossinck, M.J., 2006. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. *Plant molecular biology*, 61(4-5), pp.699-710.

Wasmuth, J.D. & Blaxter, M.L., 2004. prot4EST: translating expressed sequence tags from neglected genomes. *BMC bioinformatics*, 5(187).

Wojciechowski, M.F., Lavin, M. & Sanderson, M.J., 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. *American journal of botany*, 91(11), pp.1846-1862.

Zhang, W.-J., Zhou, J., Li, Z.-F., Wang, L., Gu, X., & Zhong, Y., 2007. Comparative Analysis of Codon Usage Patterns Among Mitochondrion, Chloroplast and Nuclear Genes in *Triticum aestivum* L. *Journal of Integrative Plant Biology*, 49(2), pp.246-254.

Zhou, M., Tong, C. & Shi, J., 2007. Analysis of codon usage between different poplar species. *Journal of genetics and genomics = Yi chuan xue bao*, 34(6), pp.555-61.

Conclusion générale et perspectives

I. Paysages nucléotidiques des plantes : un continuum de paysages hétérogènes vraisemblablement façonnés (en partie) par la BGC et la SUC

Mon travail a permis d'apporter une meilleure connaissance de la diversité des paysages nucléotidiques et des préférences d'usage du code chez les plantes. En étudiant les variations de la composition nucléotidique de plus de 200 espèces de gymnospermes et d'angiospermes, j'ai montré que, plutôt qu'une dichotomie nette dicotylédones / monocotylédones, les patrons d'hétérogénéité des paysages nucléotidiques suivent un continuum le long de la phylogénie des plantes, avec des groupes particulièrement riches et hétérogènes en GC, comme les graminées chez les monocotylédones ou, dans une moindre mesure, comme les Myrtales chez les eudicotylédones. Les forces évolutives à l'œuvre chez les plantes auraient donc façonné leurs paysages nucléotidiques avec une intensité variable.

J'ai testé indirectement deux hypothèses de mécanismes évolutifs susceptibles d'affecter la composition nucléotidique : la conversion génique biaisée vers G et C (BGC) et la sélection sur l'usage du code (SUC). Mes résultats suggèrent que la BGC – et dans une moindre mesure la SUC – pourraient être à l'origine de l'hétérogénéité des paysages nucléotidiques des plantes. Des épisodes indépendants d'enrichissement et d'appauvrissement en GC ont vraisemblablement eu lieu au cours de l'évolution des plantes, et pourraient être expliqués par des variations d'intensité de ces mécanismes.

Les variations de composition en base se traduisent aussi par des variations d'usage et de préférence des codons. En établissant les préférences actuelles d'usage du code et en reconstituant les préférences ancestrales aux nœuds de la phylogénie, j'ai révélé des patrons d'évolution différents selon les types d'acides aminés : bonne conservation des préférences pour les codons des acides aminés au code deux et trois fois dégénéré, labilité des préférences pour les codons des acides aminés au code quatre et six fois dégénéré. Ces patrons d'évolution peuvent être mis partiellement en relation avec la population d'ARNt disponibles. Enfin j'ai pu lier l'évolution des préférences d'usage du code à l'évolution de la composition nucléotidique des génomes. Ainsi mes données suggèrent-elles, en accord avec certains modèles théoriques, que le taux de GC des génomes, probablement déterminé en partie par la BGC, oriente la coévolution entre ARNt et composition nucléotidique.

II. Vers une connaissance plus complète des paysages nucléotidiques

Afin de compléter cette nouvelle vision des paysages nucléotidiques des plantes, il serait intéressant d'élargir l'éventail d'espèces étudiées, notamment en densifiant l'échantillonnage au sein des gymnospermes et en l'étendant aux plantes non spermaphytes telles que les mousses et les fougères. Cela apporterait des éléments quant à l'émergence de la BGC et de la SUC au cours de l'évolution des Embryophytes.

Avec les séquences d'EST dont je disposais, j'ai exploré la fraction exprimée des paysages nucléotidiques. La description des paysages nucléotidiques devrait être étendue à la fraction non exprimée du génome. L'étude de la composition nucléotidique des introns par exemple permettrait de dissocier les effets de la BGC et de la SUC sur la composition nucléotidique des gènes. L'utilisation de séquences génomiques complètes pourrait permettre de mieux comprendre la bimodalité du GC3 observée pour certaines espèces, en caractérisant plus finement les gènes riches et pauvres en GC (distribution dans le génome, niveau de recombinaison, gradients de GC intragéniques, distribution et longueur des introns par exemple), et de comparer plus finement les paysages nucléotidiques des plantes à la structure en isochore bien décrite chez les mammifères.

L'étude de l'évolution des préférences d'usage du code pourrait être élargie à l'échelle du vivant, afin de chercher si les patrons de conservation ou d'évoluabilité des préférences selon les acides aminés sont universels et si des épisodes d'intensification de la BGC ou du biais mutationnel ont entraîné les préférences d'usage du code vers des voies d'évolution similaires.

Les préférences d'usage du code des génomes chloroplastiques et mitochondriaux pourraient également être étudiées. Diffèrent-elles des préférences d'usage des génomes nucléaires ? Ont-elles suivies la même évolution ? Les ARNt mitochondriaux et chloroplastiques peuvent faire l'objet d'insertions dans le génome nucléaire. Constituent-ils un réservoir d'évolution pour les ARNt nucléaires, par exemple à travers des évènements de conversion génique qui permettraient l'acquisition de nouvelles espèces d'ARNt nucléaires ?

III. Vers une meilleure connaissance de l'intensité de la SUC et de la BGC

D'après mes résultats, la BGC et la SUC évolueraient en intensité et affecteraient la composition nucléotidique des génomes de manière plus ou moins marquée selon les espèces. Il serait intéressant de quantifier plus précisément leur impact sur les taux de GC et de mieux comprendre quels sont les facteurs à l'origine de ces variations.

Les outils de la génétique des populations permettront de prédire les paramètres génomiques et populationnels influençant la force de la SUC et de la BGC, d'estimer leur intensité et de quantifier l'intensité des effets de la SUC et de la BGC sur la composition nucléotidique. Par exemple, des données de polymorphisme pour des espèces aux paysages nucléotidiques contrastés et bien réparties dans la phylogénie des angiospermes permettraient d'inférer l'intensité de la SUC et de la BGC pour chacune de ces espèces et de vérifier si, comme je le prédis, la SUC et la BGC varient en intensité entre les groupes. L'utilisation de jeux de données contrastés pour les paramètres susceptibles d'affecter l'intensité de la SUC et de la BGC (par exemple régions peu recombinantes versus régions très recombinantes, expression faible versus expression forte, espèces allogames versus espèces autogames) permettrait de déterminer quels sont les paramètres majeurs jouant sur l'intensité de ces forces. Ce type d'approche permettrait également d'estimer le biais mutationnel chez différentes espèces. Il serait ainsi possible de vérifier si la mutation est bien toujours biaisée en faveur de A et T et de mettre en évidence de possibles variations d'intensité du biais mutationnel entre groupes d'espèces.

J'ai montré l'existence de gradients de composition nucléotidique le long des séquences codantes. Il existe une hypothèse proposant que l'intensité de la SUC varie le long des gènes, en lien avec la cinétique de la traduction : la vitesse de recrutement des ribosomes à l'initiation de la traduction étant limitée par les risques d'engorgement, la traduction débuterait lentement en 5', puis s'accélérerait ensuite. Il serait donc attendu que les codons préférés soient moins fréquents en 5', là où la contrainte sur la vitesse de traduction pèse moins. Un gradient intragénique de la recombinaison serait également possible, avec des taux de recombinaison plus élevés aux extrémités du gène. Il serait donc intéressant de tester ces hypothèses en cherchant la présence de gradients intragéniques de BGC et d'intensité de SUC, et de voir si ces patrons diffèrent entre espèces. Il est possible de démêler les effets respectifs de la SUC et de la BGC sur le taux de GC en contrastant les gradients de fréquence de codons A/T et G/C préférés et détestés dans des gènes appartenant à des classes d'expression et de recombinaison faibles et fortes.

Au sein d'un même génome, les différences d'efficacité de la traduction des codons synonymes détermineraient les variations d'intensité de la SUC entre codons. Il serait donc intéressant d'essayer de quantifier expérimentalement les avantages sélectifs associés à chaque codon, en estimant les différences de vitesse de traduction entre codons synonymes. Une méthode envisageable consisterait à insérer par transgénèse des gènes dont l'usage du code aurait été au préalable modifié en faveur de différents codons synonymes, d'induire leur expression et de suivre la cinétique de leur traduction. Une connaissance fine des différences d'efficacité de la traduction entre codons pourrait permettre d'expliquer les patrons d'évolution des préférences d'usage du code que j'ai mis en évidence. Par exemple, les groupes de codons dont les préférences d'usage se sont révélées les plus variables au cours de l'évolution sont-ils ceux pour lesquels les différences d'efficacité de la traduction sont les plus faibles ?

Les différences de cinétique de traduction entre codons seraient déterminées par la composition de la population d'ARNt cellulaires. J'ai utilisé comme proxy des proportions d'ARNt cellulaires les nombres de gènes d'ARNt inférés *in silico*. Or il y a des incertitudes quant à la régulation de leur expression et à leurs modifications post-transcriptionnelles – qui peuvent avoir des conséquences sur leur capacité d'appariement avec les différents codons synonymes. Il serait envisageable de quantifier directement les proportions cellulaires des différents ARNt. La composition de la population d'ARNt cellulaires est-elle bien en adéquation avec les préférences d'usage du code ? Les concentrations cellulaires des ARNt et les différences de vitesse de traduction entre codons sont-elles finement corrélées aux nombres de gènes ARNt ? Dans ce cas, l'inférence des répertoires ancestraux de gènes ARNt à partir des répertoires actuels pourrait permettre de clarifier les hypothèses autour de l'évolution de l'intensité de la SUC.

IV. Mieux comprendre les facteurs génomiques à l'origine de l'évolution de la BGC et de la SUC

L'hétérogénéité de la composition nucléotidique s'est accentuée au cours de l'évolution de certains groupes, les graminées par exemple, vraisemblablement à cause d'une intensification de la BGC. Certaines caractéristiques génomiques – patrons de recombinaison, d'expression et de méthylation par exemple – pouvant être associées à la composition nucléotidique, il serait intéressant de reconstituer également leur histoire évolutive au cours de l'évolution des plantes. La mise en parallèle des patrons d'évolution de la composition nucléotidique, des préférences d'usage du code et de ces caractéristiques génomiques pourrait apporter des éléments de réponse quant aux forces évolutives qui en sont à l'origine et à la manière dont elles ont coévolué et se sont mutuellement influencées. En particulier cela pourrait permettre de mieux comprendre quels ont été les facteurs à l'origine de l'intensification de la BGC, et dans quelle mesure l'évolution de la BGC a pu influencer en retour l'évolution de ces caractères génomiques. Cela pourrait également permettre de comprendre pourquoi des paysages nucléotidiques plus hétérogènes ont émergé dans certains groupes taxonomiques.

Plusieurs hypothèses pourraient être plus précisément testées. Les éléments transposables pourraient avoir joué un rôle déterminant dans l'évolution des génomes. En les envahissant, ils peuvent affecter profondément les caryotypes, augmentant de manière conséquente la taille des génomes et créant des déserts de gènes par exemple. Ils affectent ainsi les patrons de recombinaison de manière directe. Les patrons d'état de la chromatine se trouvent également modifiés, les déserts de gènes correspondant généralement à des régions d'hétérochromatine, moins sensibles aux cassures double-brin, ce qui affecte de manière indirecte les patrons de recombinaison. Or la distribution des effets de la BGC sur la composition nucléotidique a pour déterminants majeurs l'hétérogénéité des patrons de recombinaison et leur stabilité au cours de l'évolution – les paysages nucléotidiques seront plus accidentés si la BGC martèle pendant une longue période les mêmes régions à forte recombinaison. Quant au biais de réparation vers G/C, qui détermine également l'intensité de la BGC, il aurait pu évoluer en réponse à l'évolution du biais de mutation vers A/T, qui peut être une conséquence de l'évolution des patrons de méthylation. L'hyperméthylation des génomes pourrait être une stratégie de lutte contre les invasions d'éléments transposables. Les génomes des graminées paraissent avoir été fortement marqués par les éléments transposables; peut-être l'accentuation de l'hétérogénéité de leurs paysages nucléotidiques serait-elle en lien ? Les évènements de polyploïdisation et les remaniements chromosomiques au cours de l'histoire évolutive des plantes pourraient également

avoir eu un impact fort sur l'évolution de leur composition nucléotidique, en augmentant à court terme le taux de conversion génique et en modifiant à plus long terme les caryotypes et donc les patrons de recombinaison.

A l'inverse, la connaissance fine de l'histoire évolutive de la composition nucléotidique pourrait contribuer à apporter un éclairage à l'histoire évolutive d'autres caractères génomiques. En particulier, l'estimation des taux de GC d'équilibre (GC*) des gènes ancestraux pourrait permettre d'inférer les taux de recombinaison ancestraux, grâce à la forte corrélation entre GC* et taux de recombinaison local. La reconstruction des patrons de recombinaison des paléogénomes permettrait d'apporter des éléments de réponse aux questions touchant à l'évolution de la recombinaison.

Annexes

Annexes de l'introduction

- Article "GC-Biased Gene Conversion and Selection Affect GC Content in the Oryza Genus (rice)", publié dans Molecular Biology and Evolution

- Supplementary figure S1 : Marey's maps and recombination maps for the twelve rice chromosomes

- Supplementary figure S2 : Derived allele frequency spectra

Molecular Biology and Evolution

GC-Biased Gene Conversion and Selection Affect GC Content in the *Oryza* Genus (rice)

Aline Muyle,[1] Laurana Serres-Giardi,[1] Adrienne Ressayre,[2] Juan Escobar,[1] and Sylvain Glémin*[,1]

[1]Institut des Sciences de l'Evolution, UMR 5554 CNRS, Université Montpellier II, France

[2]INRA, UMR de Génétique Végétale, INRA/CNRS/Univ Paris-Sud/AgroParistech, Ferme du Moulon, Gif sur Yvette, France

***Corresponding author:** E-mail: glemin@univ-montp2.fr.

Associate editor: Naoki Takebayashi

Abstract

Base composition varies among and within eukaryote genomes. Although mutational bias and selection have initially been invoked, more recently GC-biased gene conversion (gBGC) has been proposed to play a central role in shaping nucleotide landscapes, especially in yeast, mammals, and birds. gBGC is a kind of meiotic drive in favor of G and C alleles, associated with recombination. Previous studies have also suggested that gBGC could be at work in grass genomes. However, these studies were carried on third codon positions that can undergo selection on codon usage. As most preferred codons end in G or C in grasses, gBGC and selection can be confounded. Here we investigated further the forces that might drive GC content evolution in the rice genus using both coding and noncoding sequences. We found that recombination rates correlate positively with equilibrium GC content and that selfing species (*Oryza sativa* and *O. glaberrima*) have significantly lower equilibrium GC content compared with more outcrossing species. As recombination is less efficient in selfing species, these results suggest that recombination drives GC content. We also detected a positive relationship between expression levels and GC content in third codon positions, suggesting that selection favors codons ending with G or C bases. However, the correlation between GC content and recombination cannot be explained by selection on codon usage alone as it was also observed in noncoding positions. Finally, analyses of polymorphism data ruled out the hypothesis that genomic variation in GC content is due to mutational processes. Our results suggest that both gBGC and selection on codon usage affect GC content in the *Oryza* genus and likely in other grass species.

Key words: codon usage, GC-biased gene conversion, GC content, mating systems, *Oryza*, recombination..

Research article

Introduction

Genomic nucleotide landscapes strongly vary among organisms. In eukaryotes, mean GC content ranges from about 20% to 60% and from 30% to 50% in animals and land plants (Lynch 2007). Moreover, several groups show strong heterogeneity of base composition along chromosomes, the so-called isochore structure, initially discovered in vertebrates (for review, see Eyre-Walker and Hurst 2001). For example, the GC content of 100-kb regions in the human genome ranks from 35% to 65%. This structure affects both coding and noncoding sequences and the GC content of a gene is highly correlated to the GC content of its flanking regions. GC content is also associated with several genome features: GC-rich regions tend to have more genes, with shorter introns; GC-rich regions replicate earlier and some classes of transposable elements are only found in regions of particular base composition (Eyre-Walker and Hurst 2001). Resolving the origin and causes of isochores is thus of major importance to understand genomes organization.

Three major kinds of hypotheses have been proposed to explain isochores evolution: selection, mutational biases, and GC-biased gene conversion (hereafter gBGC) (reviewed in Eyre-Walker and Hurst 2001). Selectionist theories argue that high GC content can be selected for by their thermal stability (Bernardi 2000). However, this theory assumes significant differences in thermal stability, hence fitness, between two individuals differing only by a few point mutations even in noncoding sequences. Given the low effective population sizes N_e and hence the low efficacy of natural selection of some groups exhibiting isochores like mammals, such assumption seems unrealistic (reviewed in Duret and Galtier 2009). In addition, this theory fails to explain why selection for thermal stability should only affect some regions of the genome and the occurrence of isochores in cold-blooded vertebrates (Kuraku et al. 2006; Chojnowski et al. 2007). There are, however, other selective forces that can influence base composition. For instance, alternative codon usage in protein-coding genes can be driven by selection in order to increase the speed or accuracy of translation (Akashi 2001). If the preferred codons have a particular base composition, then selection on codon usage can drive GC content of coding sequences (CDS), especially for highly expressed genes.

Mutational theories suggest that heterogeneous mutational biases along genomes could drive GC content. Several mutational biases have been proposed to influence base composition. High GC content was suggested to favor GC mutations, thus driving isochores evolution thanks to a positive feedback loop on base composition (Fryxell and Zuckerkandl 2000). G and C nucleotides were also proposed to be more often misincorporated in early-replicating regions, when the pool of free nucleotides is still GC rich (Eyre-Walker and Hurst 2001). Finally, as recombination

© The Author 2011. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

http://mbe.oxfordjournals.org/

Table 1. Predictions of the Three Hypotheses to Explain GC Content Evolution. GC*: equilibrium GC content (Sueoka 1962)

	gBGC	Selection on Codon Usage*	Mutational Bias
Recombination	**GC* increased with recombination**	**GC3* increased with recombination**	**?**
Mating systems	**GC* higher in outcrossers**	**GC3* higher in outcrossers**	**—**
Expression level	**—**	**GC3* increased with expression**	**—**
DAF spectra	**Increased GC DAF**	**Increased GC3 DAF**	**Similar spectra for AT and GC derived alleles**

*Predictions valid if most preferred codons end in G or C.

was proposed to be mutagenic (Lercher and Hurst 2002; Hellmann et al. 2003), recombination could also affect base composition, provided that recombination affects both mutation rates and bias.

The third hypothesis involves gBGC, a process taking place during recombination (see review in Marais 2003; Duret and Galtier 2009). During meiosis, after double-strand break, single-stranded DNA invades the homologous sequence to form a heteroduplex. When parental alleles are different, the mismatch created in the heteroduplex is repaired by changing one of the nucleotides. Experimental results in yeast and human culture cell showed that this repair is biased toward GC alleles (Birdsell 2002). This was recently confirmed by a broad survey of segregation patterns in individual meioses in yeast (Mancera et al. 2008). Biased gene conversion can also arise from a bias in initiation of double-strand breaks at meiosis, but this process is less documented and not necessarily GC biased (Marais 2003). From the point of view of population genetics, gBGC is a kind of meiotic drive that mimics selection in favor of G and C alleles and increases their probability of fixation (Nagylaki 1983). However, gBGC can be viewed as a neutral process because it does not depend on the fitness effect of alleles on the individuals that carry them. As the intensity of gBGC directly depends on recombination in heterozygotes, it will be more efficient in large outcrossing populations and in highly recombining regions along genomes.

Selection and gBGC can be distinguished from mutational bias using polymorphism data (table 1). Both gBGC and selection increase the probability of fixation of W → S mutations (weak to strong: from A or T to G or C) and decrease it for S → W mutations (strong to weak: from G or C to A or T), whereas mutational bias does not affect probabilities of fixation. Under selection or gBGC, W → S mutations are thus expected to segregate at higher frequencies than S → W mutations. gBGC can be distinguished from selection because it affects both coding and noncoding regions, whereas selection on codon usage is mainly restricted to third codon positions (table 1).

There is evidence for genomic effect of gBGC, especially in yeast (Birdsell 2002), mammals (Duret and Arndt 2008; Duret and Galtier 2009), and birds (Webster et al. 2006), but also in turtle, nematode, Drosophila, paramecium, and green algae, supporting the gBGC hypothesis (reviewed in Duret and Galtier 2009). In these organisms, correlations between recombination rate (or proxies of recombination) and GC content were observed for both coding and noncoding regions, ruling out selection on codon usage as the sole cause. More precisely, in human, the better correlation with the equilibrium GC content than the actual GC content supports the fact that recombination drives GC content, not the reverse (Meunier and Duret 2004; Duret and Arndt 2008). Polymorphism analyses in humans also showed an excess of derived G and C alleles in high frequency compared with A and T alleles, ruling out the mutation bias hypothesis. Moreover, this excess is greater in highly recombining single-nucleotide polymorphisms (SNPs) compared with low recombining ones (Spencer et al. 2006).

Among plants, grasses present a peculiar genomic structure: they exhibit a remarkable GC content heterogeneity (Carels and Bernardi 2000; Wang et al. 2004) that parallels the mammalian genome organization (e.g., regions with high GC content having numerous genes with short introns). More strikingly, GC3 distribution is strongly bimodal, as observed in rice (Shi et al. 2006). However, the causes of this high and heterogeneous GC content are still unclear. Several studies suggested that selection on codon usage occurs in rice as codon bias is related to gene expression levels (Guo et al. 2007; Mukhopadhyay et al. 2007), and most preferred codons end in G or C (Wang and Roossinck 2006). However, codon usage alone can hardly explain the overall high GC content and the strong heterogeneity in base composition observed in grasses. This is because first and second codon positions, as well as introns, are also affected (Carels and Bernardi 2000; Shi et al. 2006) and lowly expressed genes also show a strongly bimodal GC3 distribution (Mukhopadhyay et al. 2007). Other authors have thus invoked a general mutational bias toward GC to explain the isochore structure in rice and more generally in grasses (Wang et al. 2004; Wang and Hickey 2007). However, these studies are mainly based on the comparison of *Arabidopsis* and rice genes, making difficult to infer the underlying processes. Alternatively, comparative and phylogenetic approaches within grasses showed that GC content and the equilibrium GC content correlated with recombination and mating system, suggesting that gBGC could be involved (Glémin et al. 2006; Haudry et al. 2008; Escobar et al. 2010). gBGC is also a strong alternative candidate to mutational bias to explain the global shift in GC content in grasses.

The aim of this study is to assess which evolutionary forces have a significant impact on GC content in grasses (Poaceae). We chose to focus on the rice genus (*Oryza*) because it is currently one of the few grasses for which genomic data, genetic and physical maps, and polymorphism data are available. Furthermore, the existence of selfers and outcrossers among the *Oryza* genus allows for additional tests: selfing species show high levels of homozygosity such that gBGC should be inefficient. Selection is also predicted to be less efficient in selfing species because the effective size and effective recombination are reduced (Charlesworth 1992; Glémin 2007). To test the three

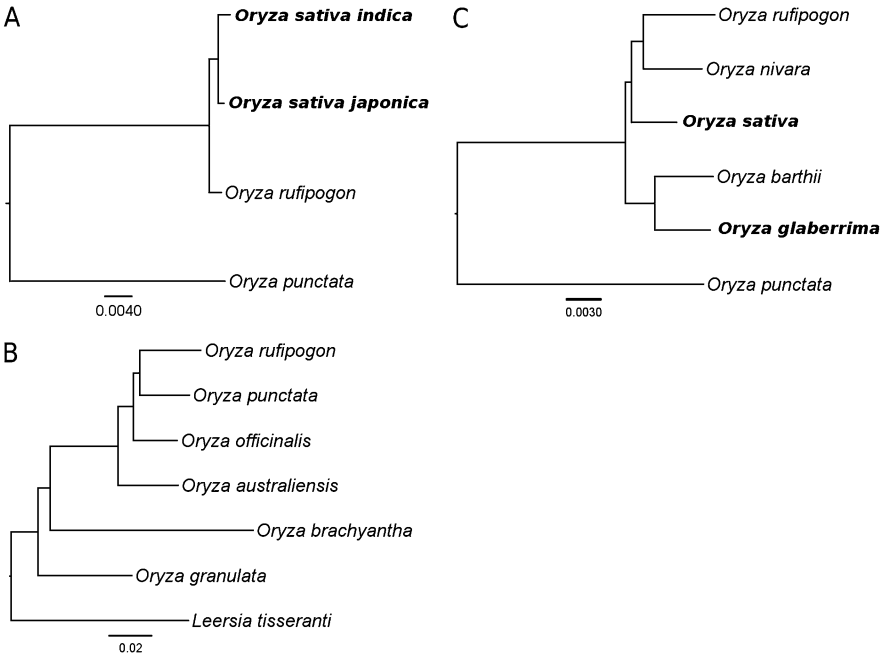


Fig. 1. Trees used in phylogenetic analyses. Bolded names indicate self-fertilizing species. (*A*) Data set 1 (132 genes, 166,648 bp). (*B*) Data set 2 (102 genes, 114,868 bp). (*C*) Data set 3 (307 genes, 136,686 bp). Branch lengths are given in units of per site substitution rate.

hypotheses mentioned above, we studied correlations between GC content, the equilibrium GC content, and recombination rates along the genomes of several *Oryza* species and compared the equilibrium GC content between selfing and outcrossing species. To contrast gBGC and selection hypotheses, we used intronic sequences as a neutral control. We also studied correlations between GC equilibrium at third codon position and expression level to assess whether selection was at work on codon usage. To discriminate between gBGC/selection and mutational biases, we studied fixation probabilities with polymorphism data.

Materials and Methods

Test of the Effect of Recombination and Expression on GC Content

Data Sets

We used two data sets published by Zou et al. (2008) and kindly provided by Fu-Min Zhang. They included exon and intron sequences. The first one contains four taxa, the two cultivated rices, *Oryza sativa* ssp. *japonica* and *O. sativa* ssp. *indica*, their wild ancestor, *O. rufipogon*, and an outgroup *O. punctata* (fig. 1*A*). It is composed of 132 genes scattered along the genome for a total length of 166,648 bp. This data set was used to study correlations between recombination rates, expression levels, and equilibrium GC content at small phylogenetic scale. The second data set gives an overview of the rice genus and consists of seven species covering major clades of the genus, including the outgroup *Leersia tisseranti* (fig. 1*B*). This data set contains 102 genes (most of which are in common with the first data set) and has a total length of 114,868 bp. This data set was used to study correlations between recombination rates, expression levels, and equilibrium GC content for each gene at a larger phylogenetic scale. For the two data sets, sequences were aligned with Muscle v3.8.31 (Edgar 2004) and the resulting alignments were manually checked. These data sets were already annotated by the authors (Zou et al. 2008), and we directly used these annotations (coding vs. noncoding, positions 1, 2, and 3).

Expression Level Data

We retrieved expression level data from the Rice Genome Annotation Project Web site (http://rice.plantbiology.msu.edu/expression_anatomy.shtml). To get the number of EST counts, we used locus names directly provided in the data sets. EST counts were used as the proxy for expression level.

Recombination Rate Estimates

To estimate recombination rates, we built a genetic versus physical distances map (Marey's map) using the 3,267 markers of the latest high density rice genetic map available on the Rice Genome Program Web site (http://rgp.dna.affrc.go.jp/E/publicdata/geneticmap2000/index.html). We then retrieved the physical position of these markers from the Rice Genome Annotation Project Web site (http://rice.plantbiology.msu.edu/annotation_insilico_genmarkers.shtml) and excluded those with multiple locations. We got 2114 markers. On the RGAP Web site, some markers are redundant because both the 5′ and the 3′ ends of the markers have been physically mapped. In such a case, we only kept one marker, which reduces the data set to 1219 markers. For each marker, we attributed the middle position of the marker (or the average of the middle position for redundant markers) as its physical location. After visual inspection of the Marey's maps, we excluded 17 markers with anomalous positions. We thus finally obtained 1202 markers along the rice genome

(see supplementary table S1, Supplementary Material online). Recombination rates were computed with the MareyMap program (Rezvoy et al. 2007). To fit the data, we first used the loess function that locally adjusts a polynomial second-degree curve, using a weight attributed to each marker depending on how far it is from the center of the window (Rezvoy et al. 2007). The size of the window is defined as a percentage of the total number of markers. We used 20% to get rather smooth recombination rate curves. We also used the cubic splines method with the cross-validation option (for details, see Rezvoy et al. 2007), which captures variation in recombination rates at a more local scale (but it is also more sensible to imprecision and errors in the genetic map). We then estimated the local recombination rate of each marker from the fitted curves using the physical location of each gene according to the release 6.1 of MSU Rice Genome Annotation. Marey's maps and recombination maps are available on supplementary figure S1 (Supplementary Material online).

GC Estimates and Analyses

GC content was computed for all genes. To investigate the forces affecting the evolution of GC content, we also computed the equilibrium GC content (hereafter GC*) reached by a sequence if substitution rates remain constant over time (Sueoka 1962):

$$GC^* = \frac{U_{W\to S}}{U_{W\to S} + U_{S\to W}}, \quad (1)$$

where $U_{W\to S}$ is the W to S substitution rate and $U_{S\to W}$ the reverse one. We used the BPPML program (Dutheil and Boussau 2008) which uses a maximum-likelihood approach to estimate GC* along a given phylogeny. It allows nonstationarity (ancestral and current GC content can differ) and nonhomogeneity (branches can have distinct GC*), which allows to test whether branches underwent similar evolution. We used the tree topologies given by Zou et al. (2008) (see fig. 1). To get a reasonable starting tree for ML search, we re-estimated branch lengths using the HKY model using PhyML v3.0 (Guindon and Gascuel 2003). In BPPML, we used the T92 model of sequence evolution (Tamura 1992) with a uniform + invariant distribution for rate evolution. This distribution was preferred to the more classical gamma distribution because data sets contain few variable sites and the gamma shape parameter was wrongly estimated for some genes. In all cases, the transition/transversion ratio was shared by all branches, whereas GC* was allowed to vary between branches or groups of branches. In nonstationary models, the root position is usually badly estimated and we preferred to fix it as the one obtained with PhyML. For each gene separately, we computed GC* for all sites (coding and noncoding), GC* at third codon position (GC3*), and at noncoding positions alone (GC* noncoding).

For the first data set, GC* were computed using *O. punctata* as outgroup and by grouping *O. rufipogon*, *O. sativa indica*, and *O. sativa japonica* together to get sufficient power as individual branches are short (fig. 1A). With the second data set, we investigated whether GC* evolved at a larger phylogenetic scale. We used *L. tisseranti* as outgroup and estimated either one GC* for the whole *Oryza* clade or one GC* for each species (terminal branch) separately: *O. rufipogon*, *O. punctata*, *O. officinalis*, *O. australiensis*, *O. brachyantha*, and *O. granulata* (fig. 1B). For these analyses, we used the species phylogeny and, as a control, the gene phylogenies provided in the original publication.

GC and GC* were arcsin squared root transformed and regressed against recombination rates and expression levels. Genes for which GC* = 0 or 1 were excluded of these analyses. Kendall nonparametric correlations were also performed. Note that recombination rates and expression levels were only available for *O. sativa* and were used for all other species.

Test of the Effect of Mating System on GC Content

Data Set

To test for the effect of mating system on the GC content dynamic, we used the data set published by Cranston et al. (2009). We directly used the alignment provided in their supplementary material after having annotated it (see below, Supplementary Material online). It consists of CDS and flanking noncoding sequences (introns, untranslated regions) corresponding to 307 genes and has a total length of 136,686 bp. It includes the two cultivated and mostly self-fertilizing species *O. sativa* (Asian rice) and *O. glaberrima* (African rice) and four wild species (*O. rufipogon*, *O. nivara*, *O. barthii* and *O. punctata*) (fig. 1C). *Oryza barthii*, *O. nivara*, and *O. rufipogon* have mixed-mating system with outcrossing rates ranging from 10 to 50% (Sweeney and McCouch 2007), and *O. punctata* could also exhibit significant level of outcrossing, at least the perennial forms (Sano 1980). For concision, all these four species were considered as outcrossers thereafter. As evolution is mostly unidirectional from outcrossing to selfing (Takebayashi and Morrell 2001; Igic et al. 2006), all internal branches were assumed to be outcrossing.

Gene Annotation

We annotated this data set by blasting sequences on the complete rice genome (version 6.1) provided by the Rice Genome Annotation Project Web site. For this, we used BlastALL (Altschul et al. 1990) (options: −p BlastN − e 1×10^{-5}) and PERL and BASH homemade scripts. We only kept hits with 100% identity. Sequence fragments that had no match on the whole CDS data were considered as noncoding. Reading frames of the CDS fragments were determined using the distance that separated their own start from the real start (ATG) of the whole-gene CDS. This distance was obtained with a blast of the CDS fragments on the whole-gene CDS. We confirmed the accuracy of this method by checking for stop codons along the fragments of CDS once they were framed.

Statistical Analyses

We used the ML framework described above to estimate GC* and compared various nested hierarchical models of sequence evolution using likelihood ratio tests (LRTs). Note that we were not interested in comparing GC* among genes but among branches. We thus concatenated all sequences (or all noncoding and all third codon positions) to get sufficient power. We could thus infer whether selfing versus outcrossing branches, and internal versus external

branches, underwent significantly different evolution (table 1). Under gBGC or selection, GC* is expected to be higher on internal than on external branches because polymorphic mutations can be counted as substitutions. As $W \rightarrow S$ mutations have a higher probability of fixation than $S \rightarrow W$ ones, $U_{W\rightarrow S}$is underestimated and $U_{S\rightarrow W}$ overestimated. In addition to the null model, M0, with homogeneous GC*, we run three nonhomogenous models. In model M1a, selfing and outcrossing branches have a different GC*. In model M1b, internal and external branches have a different GC*. In model M2, external selfing, external outcrossing, and internal outcrossing branches have different GC* (note that there are no internal selfing branches). To test the effect of internal *versus* external branches, model M2 was compared with model M1a (LRT with 1 degree of freedom). To test the effect of mating systems, model M2 was compared with model M1b (LRT with 1 degree of freedom).

Analyses of Polymorphism Data

Data Set

We used the data set published by Caicedo et al. (2007) who analyzed polymorphism in 111 gene fragments in the wild rice species *O. rufipogon* (21 individuals) and the two cultivated rice subspecies, *O. sativa* ssp. *japonica* (46 individuals) and *O. sativa* ssp. *indica* (26 individuals). *Oryza barthii* and *O. meridionalis*, their sister species, were used as outgroups to orientate polymorphism as in Caicedo et al. (2007). We retrieved the sequences corresponding to a total length of 55,562 bp from GeneBank and aligned them with Muscle v3.8.31 (Edgar 2004). All the resulting alignments were manually checked.

Determination of the Derived Allele Frequency Spectrum

We used the derived allele frequency (DAF) spectrum to quantify the strength of gBGC or selection: $B = 4N_e b$, where N_e is the effective population size and b the gBGC or selection coefficient. We analyzed separately the *O. rufipogon*, *O. s. indica*, and *O. s. japonica* data sets. To build DAF spectra, we inferred the ancestral states of SNPs using the sequences of *O. barthii* or *O. meridionalis* as outgroups, as in Caicedo et al. (2007). Because CpG hypermutability can lead to orientation problems and to spurious signatures of selection and/or gBGC (Hernandez et al. 2007), we used the method developed by Duret and Arndt (2008) that takes CpG into account to infer ancestral states, thanks to a program kindly provided by Peter Arndt. As this method requires rather large alignments, we concatenated all genes. Then we built a triple sequence alignment including an outgroup and two sequences corresponding to the two polymorphic alleles randomly assigned to one sequence or to the other. We then obtained a probability distribution for the identity of the ancestral state for each position in the alignment. SNP counts were thus weighted by these probabilities. For instance, consider a site with 12 A and four G with the ancestral state being A with a probability 0.9 and G with a probability 0.1. This site counts as 0.9 W $\rightarrow$ S SNP with DAF = 4/16 and as 0.1 S $\rightarrow$ W SNP with DAF = 12/16. We ignored missing data when it occurred in polymorphic sites, so that the number of sampled sequences, n, varied from site to site. We thus transformed our data by resampling polymorphic sites in order to have the same sample size, n', for all sites. As in Caicedo et al. (2007), we used $n' = 16$. The few SNPs with $n < 16$ were removed. The frequency of the derived allele in the reduced sample follows a hypergeometric distribution. Given k the frequency of the derived allele in the original sample of size n, the probability that i copies were observed in the reduced sample of size n' is (Hernandez et al. 2007):

$$P(i/n'|k/n) = \frac{C_k^i C_{n-k}^{n'-i}}{C_n^{n'}}. \quad (2)$$

Fit of a gBGC/Selection Population Genetics Model

We fitted a population genetic model to the DAF spectrum to estimate B. To take demography into account, we adapted the model developed by Eyre-Walker et al. (2006) to estimate selection against deleterious mutations. Although Eyre-Walker et al. (2006) distinguished the nonsynonymous and the synonymous spectra, we distinguished the W $\rightarrow$ S, the S $\rightarrow$ W, and the W $\leftrightarrow$ W + S $\leftrightarrow$ S spectra, the latter two being used as a neutral reference. Polymorphic sites were then divided into these three groups. The probability of observing k_i SNPs having i derived alleles out of n follows a Poisson distribution, $P(\mu,k_i)$, with mean:

$$\mu_{\text{neutral}} = \frac{4N_e \nu L r_i}{i}, \quad (3a)$$

for neutral W $\leftrightarrow$ W and S $\leftrightarrow$ S mutations,

$$\mu_{W\rightarrow S}(i) = 2N_e uL(1 - p_{GC})r_i \int_0^1 C_n^i x^i (1-x)^{n-i} H(b,x)dx, \quad (3b)$$

for W $\rightarrow$ S mutations, and

$$\mu_{S\rightarrow W}(i) = 2N_e \lambda uL p_{GC} r_i \int_0^1 C_n^i x^i (1-x)^{n-i} H(-b,x)dx, \quad (3c)$$

for S $\rightarrow$ W mutations, where

$$H(b,x) = 2\frac{1 - e^{-4N_e b(1-x)}}{(1 - e^{-4N_e b})(1-x)} \quad (3c)$$

is the time that a mutation with gBGC coefficient b spends between x and $x + dx$. The first term within the integral corresponds to the binomial sampling of i alleles over n given their frequency x. N_e is the effective population size, ν the mutation rate from W to W and from S to S mutations, u the mutation rate from W to S, $u\lambda$ the mutation rate from S to W, λ being the mutational bias toward AT, L the sequence length, and p_{GC} the GC content of the sequence. We assumed that p_{GC} is constant, that is the ongoing substitution process does not significantly affect base composition at the polymorphism time scale. The r_i have been introduced by Eyre-Walker et al. (2006) to take demography and/or population structure (and sampling) into account. There is one r_i for each SNP class,

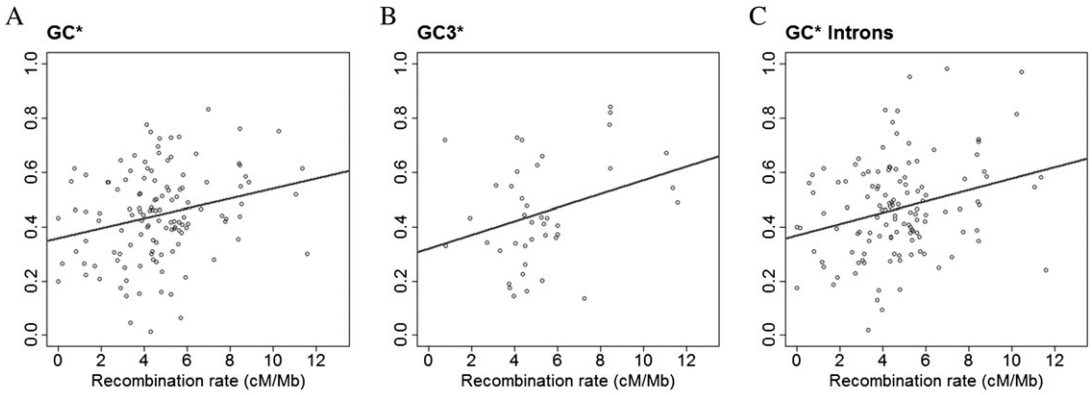


FIG. 2. Correlations between recombination rates, estimated by the loess method, and GC*. Each point corresponds to one gene in data set 1. Genes with GC* = 0 or 1 have been excluded. Regression lines are plotted when significant. (*A*) Total GC*. $n = 127$, $R^2 = 0.062$, *P* value = 0.005, $\tau_{Kendall} = 0.155$, *P* value = 0.010 for the loess method, and $n = 130$, $R^2 = 0.062$, *P* value = 0.004, $\tau_{Kendall} = 0.156$, *P* value = 0.009 for the cubic splines method. (*B*) GC3*. $n = 41$, $R^2 = 0.099$, *P* value = 0.044, $\tau_{Kendall} = 0.177$, *P* value = 0.103 for the loess method, and $n = 127$, $R^2 = 0.077$, *P* value = 0.079, $\tau_{Kendall} = 0.170$, *P* value = 0.118 for the cubic splines method. C: GC* in noncoding positions. $n = 124$, $R^2 = 0.080$, *P* value = 0.001, $\tau_{Kendall} = 0.146$, *P* value = 0.016 for the loess method, and $n = 41$, $R^2 = 0.054$, *P* value = 0.008, $\tau_{Kendall} = 0.142$, *P* value = 0.018 for the cubic splines method.

corresponding to the deviation from the standard equilibrium model relative to the singleton class for which r_1 is set to one. This parameterization makes the assumption that deviation is the same for neutral and selected sites. Although this is not necessarily true, Eyre-Walker et al. (2006) showed this approximation is robust. This is especially true in our case where low gBGC/selection coefficients are expected. To simplify the analysis and to get sufficient power, we assumed that the parameters are shared by all SNPs. Assuming independence between SNPs, the likelihood of the model can thus be written down as:

$$\Gamma = \prod_{i=1}^{n-1} P(\mu_{\text{neutral}}, k_i^{W\leftrightarrow W, S\leftrightarrow S}) P(\mu_{W\to S}, k_i^{W\to S}) P(\mu_{S\to W}, k_i^{S\to W}). \quad (4)$$

The log-likelihood was maximized by using the FindMaximum function of the Mathematica software v6 (Wolfram 1996) with the MaxIteration option set to 8. $4N_e u$, $4N_e v$, λ, and the r_i were constrained to be positive, whereas $B = 4N_e b$ was let free. *B* was estimated for all SNPs, for SNPs in third codon position, and for SNPs in noncoding positions. Then we used an LRT with one degree of freedom to compare the gBGC model for which *B* is estimated to the neutral model for which *B* is fixed to zero. Because they were sampled in several gene fragments in which linkage disequilibrium can occur (especially in the *O. sativa* datasets because of selfing), not all SNPs are fully independent. r_i coefficients potentially capture the effects of the sampling scheme; however, *P* values given by LRT must be viewed with caution (especially for *O. sativa*). We then determined the expected GC* given the estimates of *B* and λ, using the so-called Li-Bulmer equation, initially given for codon usage (Bulmer 1991):

$$GC^* = \frac{1}{1 + \lambda e^{-B}}. \quad (5)$$

Results

Analyses on a Small Phylogenetic Scale

We first carried out analyses at a small phylogenetic scale, that is, for the species closely related to *O. sativa*, the cultivated Asian rice species. We computed GC* for each of the 132 genes using the species tree provided in figure 1*A* (data available on supplementary table S2, Supplementary Material online). We first investigated the relationship between GC*, current GC content, and recombination rates by regression and its strength using the Kendall nonparametric correlation coefficient. We found a weak but significant correlation between current GC content and recombination rates estimated both with the loess (R^2 of the regression = 0.064, *P* value = 0.003, $\tau_{Kendall} = 0.145$, *P* value = 0.0213) and the cubic splines methods ($R^2 = 0.057$, *P* value = 0.005, $\tau_{Kendall} = 0.131$, *P* value = 0.022). We also found a correlation between GC* and recombination rates ($R^2 = 0.062$, *P* value = 0.005, $\tau_{Kendall} = 0.155$, *P* value = 0.010 and see figure 2*A* for the loess method, and $R^2 = 0.062$, *P* value = 0.004, $\tau_{Kendall} = 0.156$, *P* value = 0.009 for cubic splines method). The correlation between GC* and recombination was still significant when looking only at noncoding positions ($R^2 = 0.080$, *P* value = 0.001, $\tau_{Kendall} = 0.146$, *P* value = 0.016 and see figure 2*C* for the loess method, and $R^2 = 0.054$, *P* value = 0.008, $\tau_{Kendall} = 0.142$, *P* value = 0.018 for the cubic splines method) and only marginally significant at third codon positions ($R^2 = 0.099$, *P* value = 0.044, $\tau_{Kendall} = 0.177$, *P* value = 0.103 and see figure 2*B* for the loess method, and $R^2 = 0.077$, *P* value = 0.079, $\tau_{Kendall} = 0.170$, *P* value = 0.118 for the cubic splines method), probably because of lower statistical power ($n = 41$ for third positions vs. $n = 124$ for noncoding positions).

We also tested for the effect of recombination on GC* through the comparison of outcrossing and selfing lineages using the third data set (fig. 1*C*). As there are only terminal selfing branches while there are both internal and terminal

Table 2. Estimates of GC* on the Branches of the Tree Corresponding to Figure 2C and *P* values of the Nested Hierarchical Models of Sequence Evolution Likelihood-Ratio Tests.

		GCtotal*	GC3*	GCnoncoding*
	Number of sites	**136,685**	**41,730**	**12,772**
GC* on branches	**External selfing**	**0.35**	**0.34**	**0.27**
	External outcrossing	**0.39**	**0.37**	**0.32**
	Internal outcrossing	**0.49**	**0.47**	**0.45**
***P* values**	**Selfing vs. outcrossing**	***$<10^{-6}$***	***0.003***	***0.04***
	External selfing vs. external outcrossing	***$<10^{-9}$***	***$<10^{-4}$***	***0.03***
	External vs. internal	***$<10^{-4}$***	***0.04***	***0.06***
	External outcrossing vs. internal outcrossing	***$<10^{-7}$***	***$<10^{-3}$***	***0.04***

Significant values at the 5% level are italic.

outcrossing branches, we compared external selfing and external outcrossing branches to test for the effect of mating system. Similarly, we compared internal outcrossing and external outcrossing branches to test for the possible effect of branch positions. External selfing branches have a significantly lower GC* compared with external outcrossing branches, for all positions, third codon positions, and noncoding positions (table 2). Likewise, for all subsets, outcrossing external branches have a lower GC* compared with internal outcrossing branches, as expected if either gBGC or selection on codon usage is driving GC content (table 2).

To test for possible effects of selection on third codon position, we investigated the relationship between GC3* and noncoding GC* with the expression level using data set 1. A positive correlation between GC3* and the expression level was detected ($R^2 = 0.209$, *P* value = 0.0044, $\tau_{Kendall} = 0.233$, *P* value = 0.044, fig. 3*A*), whereas the expression level and noncoding GC* were not correlated (fig. 3*B*).

Analyses on a Larger Phylogenetic Scale

We then explored the relationship between recombination, expression, and GC content at the whole *Oryza* genus scale. We computed GC* for each gene of the second data set, using the species tree provided in fig. 1*B* and allowing one GC* for all the *Oryza* clade and one for the outgroup, *L. tisserandi* (data available on supplementary table S2, Supplementary Material online). We reasoned that if the same processes have been acting to modify the GC content at this phylogenetic scale, GC* would be better estimated because of the higher number of substitutions in this data set compared with the first data set. We observed no correlation between GC* and recombination rates nor between GC3* and expression levels (data not shown). We controlled that these results were not affected by the topology of the tree by using gene trees rather than the species tree.

In order to assess at which point of the phylogeny the correlation between GC* and recombination or expression is lost, we computed GC* and GC3* for each gene and for each species of the second data set and studied correlations with recombination and expression. GC* are correlated with recombination rates in *O. rufipogon* but not in other species, except in one case (*O. brachyanta*, table 3). We observed similar results for correlations between expression levels and GC3* (table 3).

Polymorphism Data Analyses

From the Caicedo et al. (2007) data set, we obtained W → S, S → W, and neutral DAF spectra for wild and cultivated rices. In *O. rufipogon*, we observed that W → S SNPs tend to segregate at higher frequencies than S

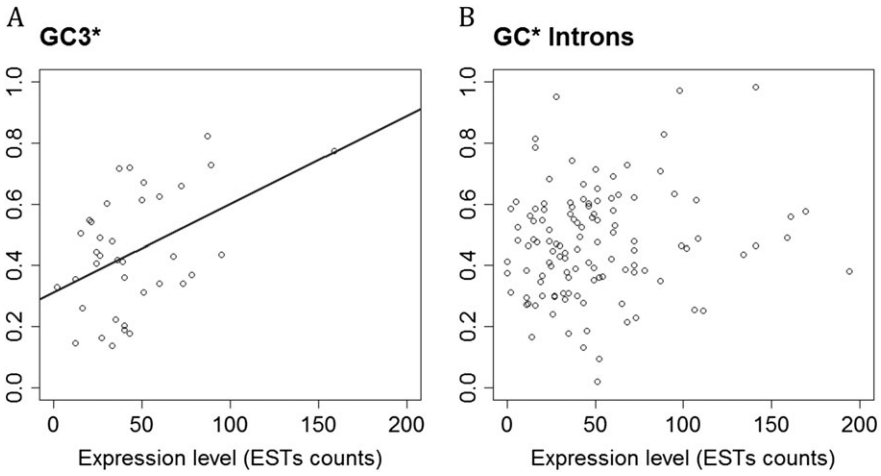


Fig. 3. Correlations between raw expression level (EST counts) and GC*. Each point corresponds to one gene in data set 1. Genes with GC* = 0 or 1 have been excluded. Regression lines are plotted when significant. (*A*) GC3*. $n = 41$, $R^2 = 0.209$, *P* value = 0.0044, $\tau_{Kendall} = 0.233$, *P* value = 0.044. (*B*) GC* in noncoding positions. $n = 127$, $R^2 = 0.0002$, *P* value = 0.8805, $\tau_{Kendall} = 0.070$, *P* value = 0.4443.

Table 3. Correlation between Recombination and GC* and Expression and GC3* for Each Branch of the Tree Corresponding to figure 2*B*.

	Correlation between GC* and Recombination				Correlation between GC3* and Expression			
	Linear Regression		Kendall Correlation		Linear Regression		Kendall Correlation	
	R^2	*P* Value	τ	*P* Value	R^2	*P* Value	τ	*P* Value
Oryza rufipogon	***0.049***	***0.028***	***0.138***	***0.043***	***0.101***	***0.009***	***0.227***	***0.007***
O. punctata	**0.017**	**0.201**	**−0.096**	**0.166**	**0.001**	**0.780**	**0.039**	**0.650**
O. officinalis	**0.006**	**0.460**	**−0.050**	**0.472**	**0.034**	**0.150**	**−0.074**	**0.390**
O. australiensis	**0.018**	**0.191**	**−0.119**	**0.082**	**0.002**	**0.750**	**−0.019**	**0.820**
O. brachyantha	**0.033**	**0.072**	***0.149***	***0.029***	**0.015**	**0.240**	**0.072**	**0.310**
O. granulata	**0.010**	**0.320**	**0.112**	**0.104**	**0.007**	**0.460**	**0.070**	**0.360**

Significant values at the 5% level are in italic type.

→ W SNPs, especially in third codon positions (fig. 4 and supplementary fig S2, Supplementary Material online). For all positions, the mean frequency of the derived allele is 0.28 for S → W SNPs versus 0.31 for W → S SNPs when SNPs were orientated using *O. meridionalis* as outgroup, and 0.29 and 0.34, respectively, when orientated using *O. barthii*. These differences are, however, not significant (sign-rank test *P* value = 0.474 and *P* value = 0.090, respectively). In *O. sativa*, patterns are less biased and less clear because of the strong impact of demography (see Caicedo et al. 2007) and because of the lower number of SNPs (see supplementary table S2, Supplementary Material online).

To better capture the potential differences between the two spectra (and not only differences between mean), we fitted a gBGC/selection model on the spectra and compared it with a neutral model. In *O. rufipogon*, the estimated *B* coefficient is positive in most cases. However, the gBGC/selection model was significantly better than the neutral model only for third codon positions (table 4). In contrast, in *O. sativa*, *B* is lower and never significant. In all cases, the mutational bias, λ, is higher than 1, showing a global mutational bias toward A and T bases. In *O. rufipogon*, λ is consistently similar for all sites, with values around 3. In *O. sativa*, estimates are lower and less stable. Using *B* and λ estimates, we computed the predicted GC* (table 5). Results suggest that GC is decreasing in introns in both *O. rufipogon* and *O. sativa*, whereas in third codon positions it is increasing in *O. rufipogon* and seems to be at equilibrium in *O. sativa*.

Discussion

Among plants, grasses show rich and highly heterogeneous GC content along their genomes (Carels and Bernardi 2000), which is reminiscent of the mammalian isochores. Although gBGC is thought to play a crucial role in shaping mammalian nucleotide landscape (reviewed in Duret and Galtier 2009), which forces act on grass genomes is still a controversial issue (Wang et al. 2004; Glémin et al. 2006; Guo et al. 2007; Mukhopadhyay et al. 2007; Wang and Hickey 2007; Haudry et al. 2008; Escobar et al. 2010). Here we tested in the *Oryza* genus the various

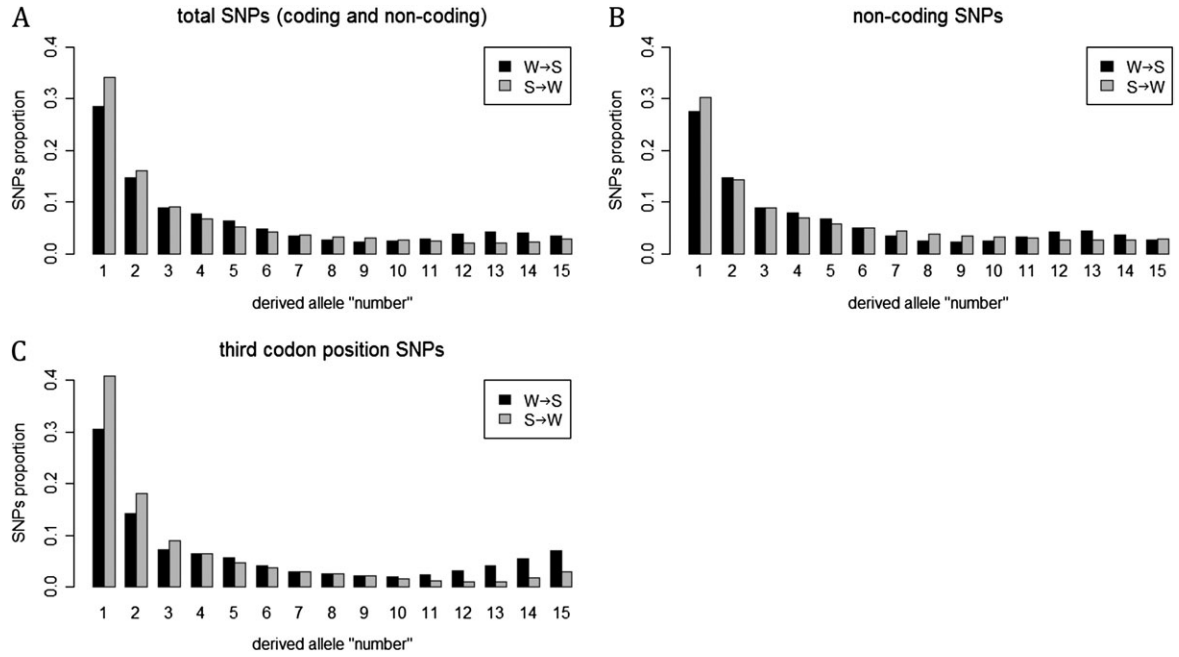


Fig. 4. DAF spectra of *Oryza rufipogon* SNPs orientated with *O. barthii*. (*A*) Total SNPs (coding and noncoding). (*B*) Noncoding SNPs. (*C*) Third codon position SNPs. To cope with SNPs with missing data, spectra have been computed for a subsample of size $n = 16$ as in Caicedo et al. (2007).

Table 4. Estimations of the Intensity of gBGC/selection ($B = 4N_e b$) and mutation bias, λ, from DAF spectra. *n*: number of W → S and S → W SNPs used, $n_{(WW + SS)}$: number of "neutral" SNPs used (see main text for details).

		GCtotal				GC3				GCnoncoding				
Species	Outgroup	*B*	λ	*P* value	*n*	*B*	λ	*P* value	*n*	*B*	λ	*P* value	*n*	$n_{(WW+SS)}$
0ryza rufipogon	*O. meridionalis*	0.463	2.934	0.225	517	*1.677*	2.932	*0.030*	112	−0.001	3.234	0.999	326	82
	O. barthii	0.672	3.014	0.062	536	*1.864*	2.774	*0.017*	114	0.273	3.512	0.535	345	94
0. sativa japonica	*O. meridionalis*	−0.018	2.263	0.967	224	0.341	1.206	0.721	47	0.219	3.388	0.682	143	30
	O. barthii	0.055	2.605	0.897	233	1.176	2.133	0.205	49	−0.107	3.290	0.833	148	36
0. sativa indica	*O. meridionalis*	−0.473	2.156	0.347	223	0.002	1.425	0.999	50	−0.777	2.509	0.216	141	41
	O. barthii	0.026	2.708	0.957	237	0.820	1.846	0.416	51	−0.420	3.158	0.494	144	54

Significant values at the 5% level are in italic type.

hypotheses previously proposed to explain GC content variations in other species: biased gene conversion, selection, and mutational biases (Eyre-Walker and Hurst 2001).

Recombination and Mating Systems Affect GC Content

We found a weak positive correlation between GC content and recombination rates in *Oryza*. However, such a correlation cannot distinguish between recombination driving GC content and the reverse. For this reason, we studied correlations between recombination and equilibrium GC content (GC* or "future" GC content). We also found a positive and significant correlation between GC* and recombination, suggesting that recombination drives GC content (Duret and Arndt 2008). In addition, we found that selfing species have lower GC* values compared with outcrossing species (table 2). Because of their high homozygosity, selfing species should be weakly affected by gBGC (Marais et al. 2004). This result suggests that recombination, through mating systems, drives GC content and not the other way around as GC content cannot drive mating systems.

These observations could be explained by a mutagenic effect of recombination. If recombination induced more W → S than S → W mutations, GC content would mechanically increase with recombination. However, several results are not consistent with this hypothesis. First, selfing species have lower GC* compared with outcrossing species (table 2). Recombination is still at work in selfing species, whereas it is not genetically efficient because of homozygosity. If recombination were mutagenic we would not expect an effect of mating system. Second, under the mutational bias hypothesis, we expect mutations to have similar fixation probabilities. The fact that external branches have lower GC* compared with internal branches (table 2) shows that polymorphism present in external branches is enriched in W alleles that are eventually lost more frequently than S alleles, that is, they are selected against. Polymorphism data also show that W → S mutations segregate at higher frequency than S → W mutations, at least for third codon positions (fig. 3 and table 4). As expected, this effect is higher in *O. rufipogon*, a species that exhibit higher outcrossing rate than *O. sativa*. Third, we also showed that mutations are biased towards A and T bases ($\lambda \approx 3$), which cannot explain the high GC content of rice genome and grass genomes in general. Only two hypotheses thus remain to explain the effect of recombination: gBGC and selection.

gBGC, Selection, or Both?

Previous studies have already shown that mating system and recombination affect GC content in grasses (Haudry et al. 2008; Escobar et al. 2010). However, as these studies only used CDS, they could not disentangle selective and gBGC effects. Indeed, recombination and outcrossing affect selection and gBGC in a similar way: gBGC is driven by recombination and selection is more efficient in outcrossing and recombining genomes (Charlesworth and Wright 2001; Gordo and Charlesworth 2001). We thus expect selection on codon usage and gBGC to be less efficient in genomic regions of low recombination and in selfing species. To distinguish gBGC from selection on codon usage, we used noncoding sequences, affected only by gBGC, and expression levels, which mainly affects third codon positions.

GC* is positively correlated with recombination both for third codon positions and noncoding positions (fig. 2), showing that recombination affects GC content at all positions, with no distinction whether they are coding or noncoding. Similarly, GC* values for both noncoding and third codon positions are significantly lower in selfing than in outcrossing branches (table 2). GC* is also significantly lower in external than internal branches for all site categories (table 2). These observations cannot be explained by selection on codon usage alone as only third codon positions would be affected. We thus have strong evidence for gBGC in the *Oryza* genus.

Besides, we observed a positive correlation between GC3* and the expression level but not with noncoding GC* (fig. 3). This clearly suggests selection on codon usage for translation efficiency, as almost all favored codons in grasses, and in rice in particular, end with a G or C base (Wang and Roossinck 2006). Another result suggesting selection on codon usage comes from the polymorphism analysis. The gBGC/selection model is significantly better than the neutral model only for third codon positions and the estimated coefficient is higher than for noncoding sites (table 4). This strongly suggests that selection on codon usage affects third codon positions in addition of any possible gBGC effect. These results show that selection is at work on codon usage and that it also drives the GC3-content in the *Oryza* genomes. The fact that these effects are weaker and not significant in *O. sativa* is also consistent

Table 5. Estimate of GC* from equation (5) and estimates given in Table 4.

Species		GCtotal	GC3	GCnoncoding
	Current GC	0.44	0.55	0.36
0ryza rufipogon	GC*/*O. meridionalis*	0.35	0.65	0.24
	GC*/*O. barthii*	0.39	0.70	0.27
0ryza sativa japonica	GC*/*O. meridionalis*	0.30	0.54	0.27
	GC*/*O. barthii*	0.29	0.60	0.21
0. sativa indica	GC*/*O. meridionalis*	0.22	0.41	0.15
	GC*/*O. barthii*	0.27	0.55	0.17

with a reduced efficacy of both selection and gBGC due to the high selfing rate.

Nonequilibrium Processes

We observed significant correlations between GC* and recombination only at a small phylogenetic scale, when we focused on *O. sativa* and *O. rufipogon* species. As recombination rates were obtained from *O. sativa* ssp. *japonica*, this suggests that recombination patterns vary quite quickly along the *Oryza* genus, corresponding to about 10 My (Zou et al. 2008; Kellog 2009). If the recombination pattern and the associated gBGC dynamics were mostly conserved, we would have expected a stronger relationship between recombination and GC* at the whole genus scale, simply because, if constant, GC* is better estimated in a larger phylogeny. In contrast, in Triticeae, recombination pattern is thought to be partly conserved. Accordingly, Escobar et al. (2010) found a significant positive correlation between recombination and GC* at the whole Triticeae scale, corresponding to about 12–15 Mya, similar to the divergence of the *Oryza* genus. We observed similar results for correlations between GC3* and expression levels (obtained from *O. sativa*), suggesting that expression levels may also vary quickly at this evolutionary scale.

The comparison between current GC content and GC* in the fourth data set (table 5) also shows that the *O. rufipogon* genome is not at equilibrium. Noncoding GC content seems to be decreasing. This is consistent with a mutational bias toward A and T bases. However, to allow this decrease, the GC content of noncoding positions must have previously increased. Only gBGC can cause an increase in the GC content in noncoding regions. gBGC strength could have gone down recently. A change in gBGC strength is quite plausible considering that recombination is known to vary quite quickly (Coop and Przeworski 2007). Moreover, if recombination occurs in short-lived hot spots as in primates (Spencer et al. 2006), only a small fraction of polymorphism should be driven by gBGC, making its detection difficult with as few as 800 SNPs (compared with the few millions of SNPs available in humans). In noncoding regions, selection may favor AT variants in some positions because of their involvement in regulatory functions, which would oppose to gBGC.

Comparison with Mammalian Isochores

Nucleotide landscapes of grass genomes show striking similarities with the isochore structure found in mammals and other vertebrates. Assuming our results in *Oryza* could be generalized to other grass species, we can draw parallels between the mechanisms involved in the two groups. As in mammals, recombination seems to drive GC content, likely through the effect of gBGC. Our results are less clear than in mammals, simply because the data sets we used are much smaller than those used in mammals (e.g. Duret and Arndt 2008). Although we found evidence of gBGC in noncoding regions, we were not able to estimate significant gBGC coefficients (table 4), which suggests gBGC is too weak to be detectable with the SNPs data set we used. In humans, the gBGC intensity is also low, $4N_e b \approx 1.3$ for the 20% of the genome with the highest recombination rates (Spencer et al. 2006). Grouping SNPs for all recombination categories as we did would lead to $4N_e b \approx 0.3$, which is indeed too low to be detectable with the data set we used. However, even such a low level is sufficient to explain GC content patterns in primates (Duret and Arndt 2008).

Contrary to mammals, we also found evidence of efficient selection on codon usage in the *Oryza* genus, whereas it is very low, if any, in mammals. As most preferred codons end in G or C in grasses (Wang and Roossinck 2006), selection on codon usage may also contribute to the high GC3 content found in grasses. Moreover, analyses of polymorphism data suggest that selection on codon usage is stronger than gBGC, as positive and significant selection/gBGC coefficients were only found at third codon positions (table 4).

Conclusions

Our results shed light on the likely causes of the peculiar nucleotide landscape in rice and very likely in other grass species. We showed that gBGC is an active process in this group and may have driven GC content evolution in addition to selection on codon usage. Because it is a selection-like process, gBGC can lead to spurious signatures of relaxed or even positive selection (Galtier and Duret 2007; Galtier et al. 2009; Ratnakumar et al. 2010). We thus strongly suggest taking gBGC into account in future genomic studies in grasses. Quantifying the genomic and the possible fitness impacts of gBGC will be now possible thanks to the increasing amount of genomic data available in several species. Finally, why gBGC seems to be more active in grasses than in other angiosperms studied so far remains an open and challenging question.

Supplementary Material

Supplementary tables S1and S2 and figures S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Fu-Min Zhang and Song Ge who kindly provided us with their data set and Peter Arndt who kindly provided his programs for CpG analyses. Julien Dutheil gave very helpful guidelines for the use of BPPML and Gabriel Marais

for the use of MareyMap. We also thank Nicolas Galtier and Laurent Duret for helpful discussions and two anonymous reviewers for their constructive comments. This publication is the contribution ISEM 2011-030 of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554—CNRS). This work was supported by the French *Centre National de la Recherche Scientifique* and *Agence Nationale de la Recherche* (ANR-08-GENM-036-01).

References

Akashi H. 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev*. 11:660–666.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241:3–17.

Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol*. 19:1181–1197.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.

Caicedo AL, Williamson SH, Hernandez RD, et al (12 co-authors) 2007. Genome-wide patterns of nucleotide polymorphism in domesticated. rice. *PLoS Genetics* 3:1745–1756.

Carels N, Bernardi G. 2000. Two classes of genes in plants. *Genetics* 154:1819–1825.

Charlesworth B. 1992. Evolutionary rates in partially self-fertilizing species. *Am Nat*. 140:126–148.

Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. *Curr Opin Genet Dev*. 11:685–690.

Chojnowski JL, Franklin J, Katsu Y, Iguchi T, Guillette LJ Jr., Kimball RT, Braun EL. 2007. Patterns of vertebrate isochore evolution revealed by comparison of expressed mammalian, avian, and crocodilian genes. *J Mol Evol*. 65:259–266.

Coop G, Przeworski M. 2007. An evolutionary view of human recombination. *Nat Rev Genet*. 8:23–34.

Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA. 2009. Species trees from highly incongruent gene trees in rice. *Syst Biol*. 58:489–500.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4:e1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*. 10:285–311.

Dutheil J, Boussau B. 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol*. 8:255.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 5:113.

Escobar JS, Cenci A, Bolognini JS, Haudry A, Laurent S, David J, Glemin S. 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (poaceae). *Evolution* 64:2855–2872.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet*. 2:549–555.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900.

Fryxell KJ, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol*. 17:1371–1383.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet*. 23:273–277.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*. 25:1–5.

Glémin S. 2007. Mating systems and the efficacy of selection at the molecular level. *Genetics* 177:905–916.

Glémin S, Bazin E, Charlesworth D. 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc Biol Sci*. 273:3011–3019.

Gordo I, Charlesworth B. 2001. Genetic linkage and molecular evolution. *Curr Biol*. 11:R684–R686.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Guo X, Bao J, L Fan . 2007. Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes. *FEBS Lett*. 581:1015–1021.

Haudry A, Cenci A, Guilhaumon C, Paux E, Poirier S, Santoni S, David J, Glémin S . 2008. Mating system and recombination affect molecular evolution in four Triticeae species. *Genet Res*. 90:97–109.

Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet*. 72:1527–1535.

Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. *Mol Biol Evol*. 24:2196–2202.

Igic B, Bohs L, Kohn JR. 2006. Ancient polymorphism reveals unidirectional breeding system shifts. *Proc Natl Acad Sci U S A*. 103:1359–1363.

Kellog EA. 2009. The evolutionary history of Ehrhartoideae, Oryzeae, and Oryza. *Rice* 2:1–14.

Kuraku S, Ishijima J, Nishida-Umehara C, Agata K, Kuratani S, Matsuda Y. 2006. cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a chromosomal size-dependent GC bias shared by sauropsids. *Chromosome Res*. 14:187–202.

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet*. 18:337–340.

Lynch M. 2007. The origin of genome architecture. Sunderland (MA): Sinauer.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454:479–485.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet*. 19:330–338.

Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. *Genome Biol*. 5:R45.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*. 21:984–990.

Mukhopadhyay P, Basak S, Ghosh TC. 2007. Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes. *Gene* 400:71–81.

Nagylaki T. 1983. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A*. 80:6278–6281.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci*. 365:2571–2580.

Rezvoy C, Charif D, Gueguen L, Marais GA. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*. 23:2188–2189.

Sano Y. 1980. Adaptive strategies compared between the diploid and tetraploid forms of Oryza punctata. *Bot Mag*. 93:171–180.

Shi X, Wang X, Li Z, Zhu Q, Tang W, Ge S, Luo J. 2006. Nucleotide substitution pattern in rice paralogues: implication for negative correlation between the synonymous substitution rate and codon usage bias. *Gene* 376:199–206.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet*. 2:e148.

Sueoka N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A*. 48:582–592.

Sweeney M, McCouch S. 2007. The complex history of the domestication of rice. *Ann Bot*. 100:951–957.

Takebayashi N, Morrell PL. 2001. Is self-fertilization an evolutionary deed end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. *Am J Bot*. 88:1143–1150.

Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*. 9:678–687.

Wang HC, Hickey DA. 2007. Rapid divergence of codon usage patterns within the rice genome. *BMC Evol Biol*. 7 (Suppl 1):S6.

Wang HC, Singer GA, Hickey DA. 2004. Mutational bias affects protein evolution in flowering plants. *Mol Biol Evol*. 21:90–96.

Wang LJ, Roossinck MJ. 2006. Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. *Plant Mol Biol*. 61:699–710.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol*. 23:1203–1216.

Wolfram S. 1996. The Mathematica book. Cambridge: Cambridge University Press.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang LL, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol*. 9:R49.

Supplementary figure S1

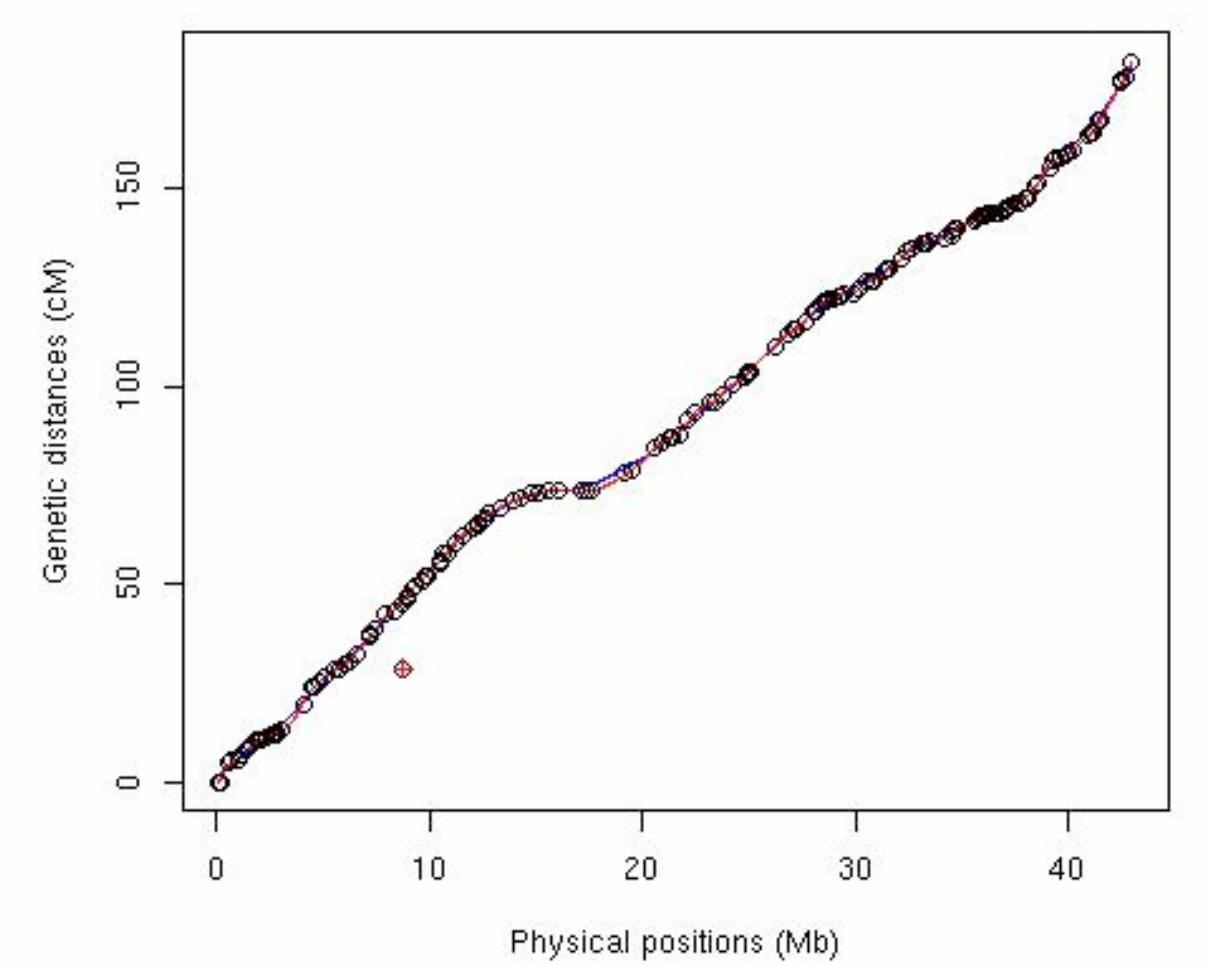
Marey's maps and recombination maps for the twelve rice chromosomes.

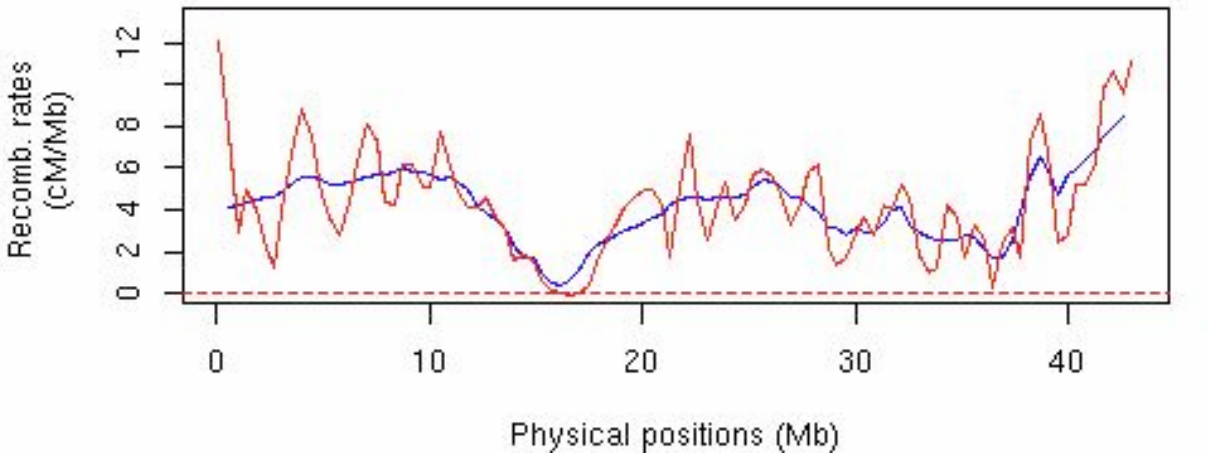
We started with the 3267 markers of the latest high density rice genetic map available on the Rice Genome Program website (http://rgp.dna.affrc.go.jp/E/publicdata/geneticmap2000/index.html). We then retrieved the physical position of these markers from the Rice Genome Annotation Project website (http://rice.plantbiology.msu.edu/annotation_insilico_genmarkers.shtml) and excluded those with multiple locations. We got 2114 markers. On the RGAP website, some markers are redundant because both the 5' and the 3' ends of the markers have been physically mapped. In such a case, we only kept one marker, which reduces the dataset to 1219 markers. For each marker we attributed the middle of the marker (or the average of the middle position for redundant markers) as its physical location. After visual inspection of the Marey's maps, we excluded 17 markers with aberrant positions (marked dots on figures). We thus finally obtained 1202 markers along the rice genome (See Supplementary Table S1). Recombination rates were computed with the MareyMap program (Rezvoy et al. 2007).
Recombination maps were obtained using the software MareyMap (Rezvoy et al. 2007). The functions used to fit the data were (i) in blue: the lœss function (span = 0.2 degree = 2), (ii) in red : the cubic spline function with cross-validation.

Reference

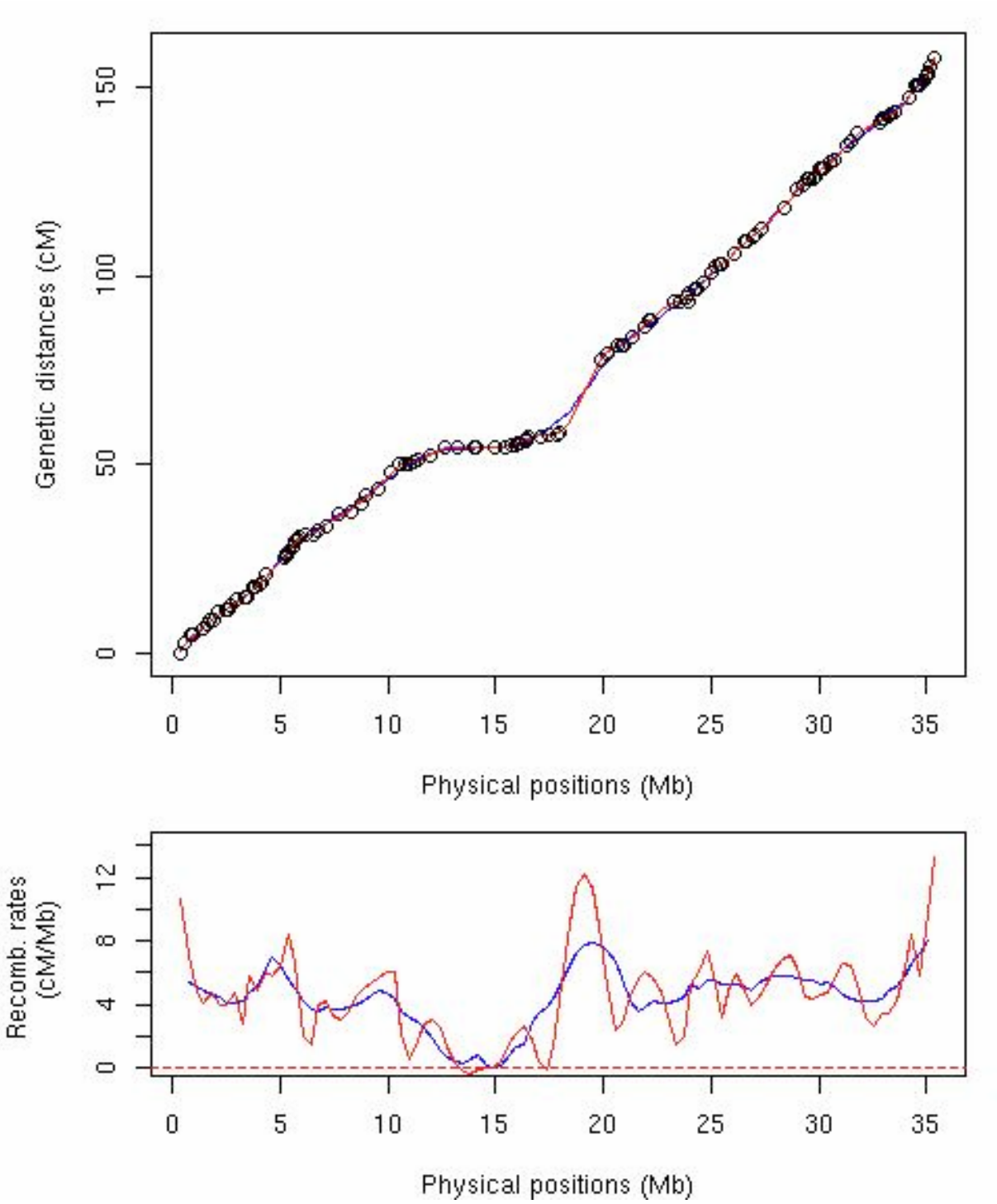
Rezvoy, C, D Charif, L Gueguen, GA Marais. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics 23:2188-2189.

Chromosome 1

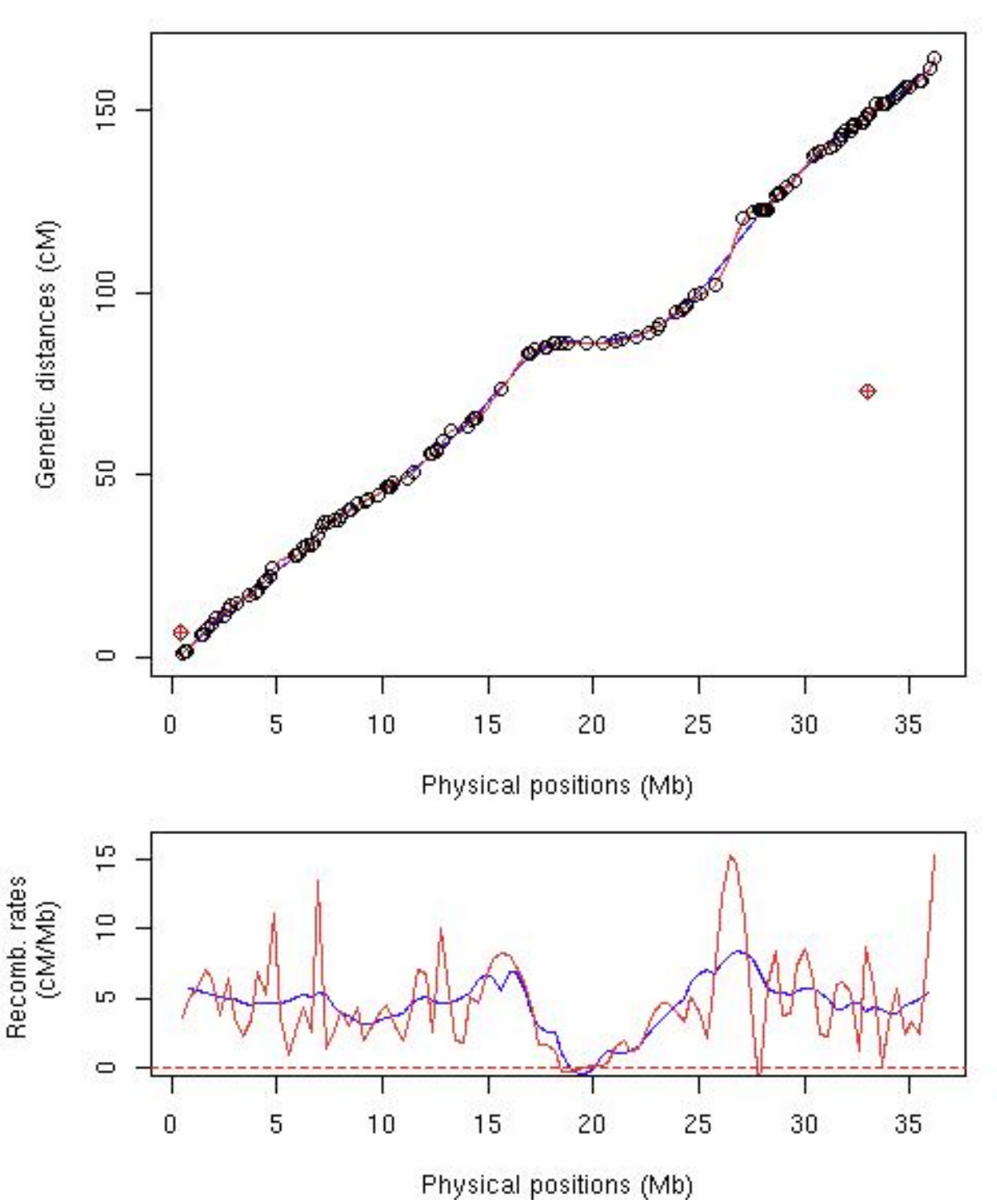




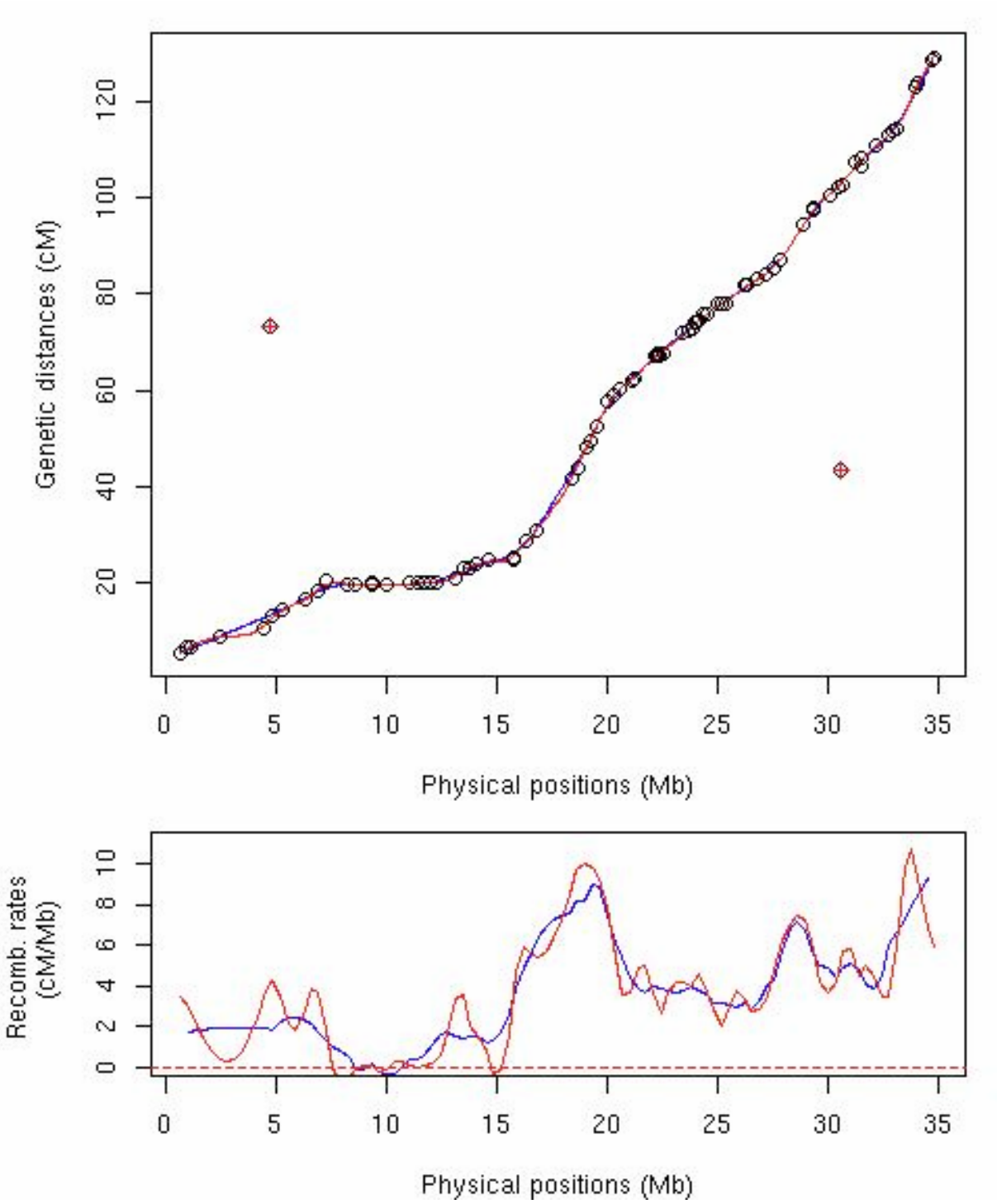
Chromosome 2



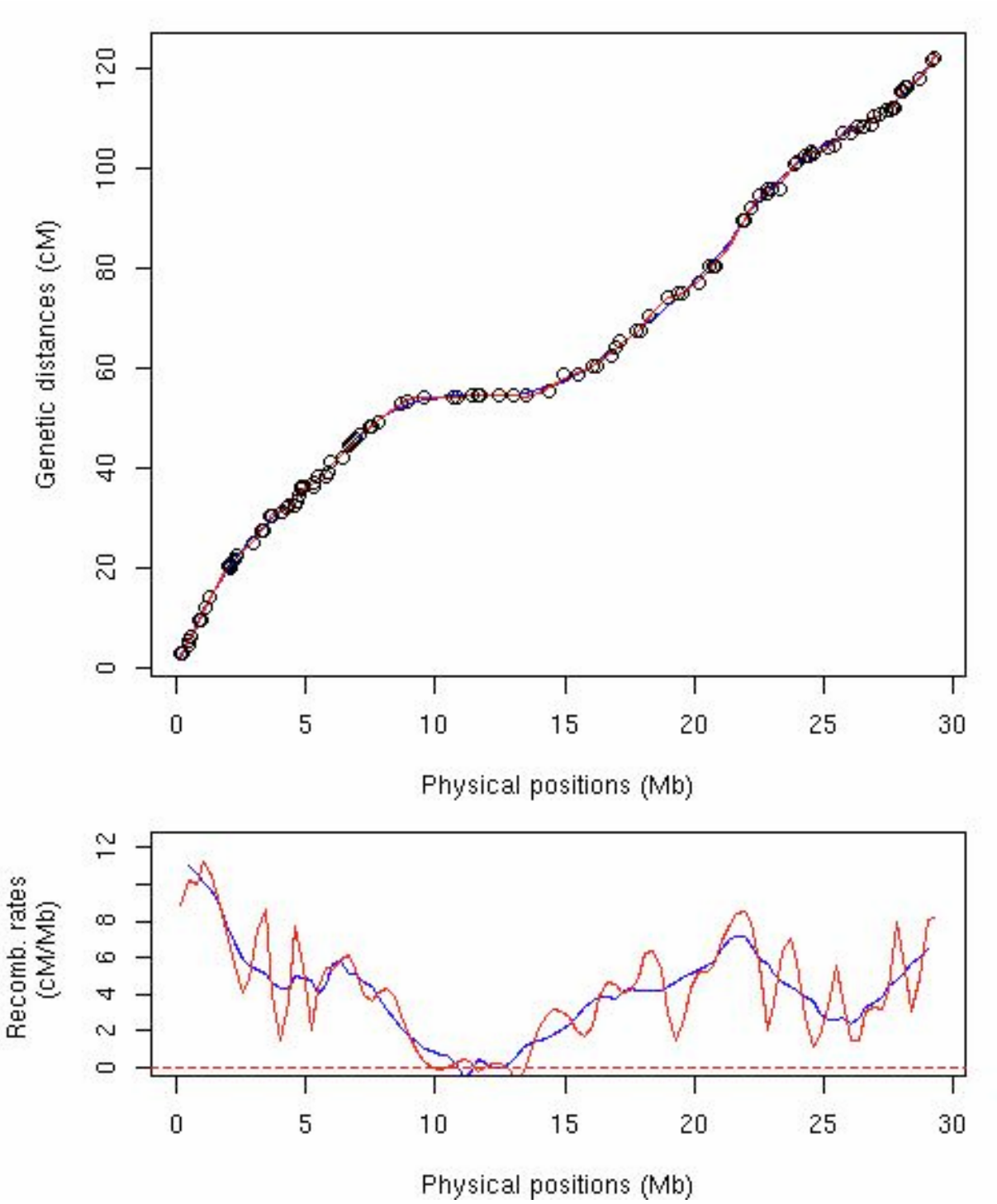
Chromosome 3



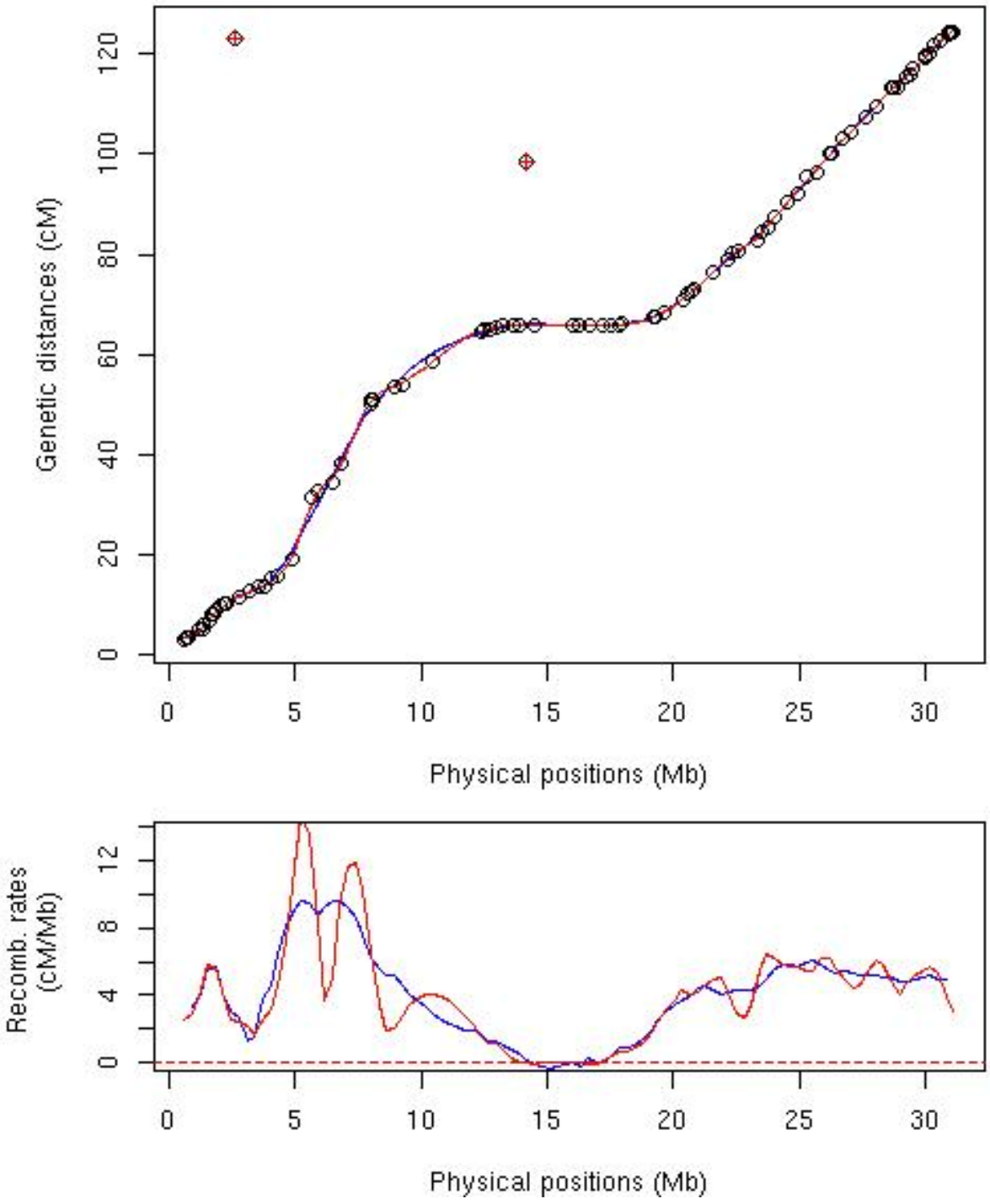
Chromosome 4



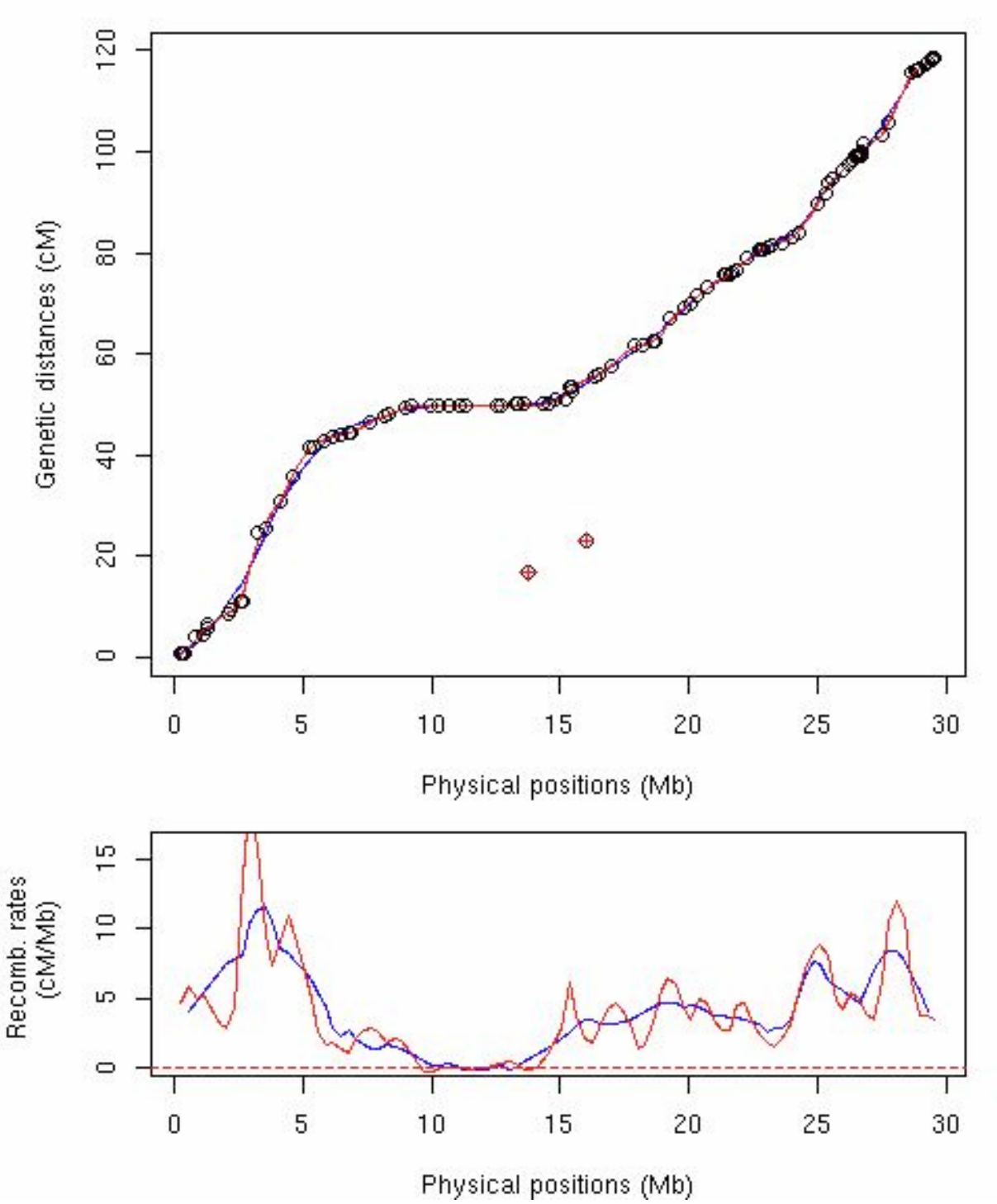
Chromosome 5



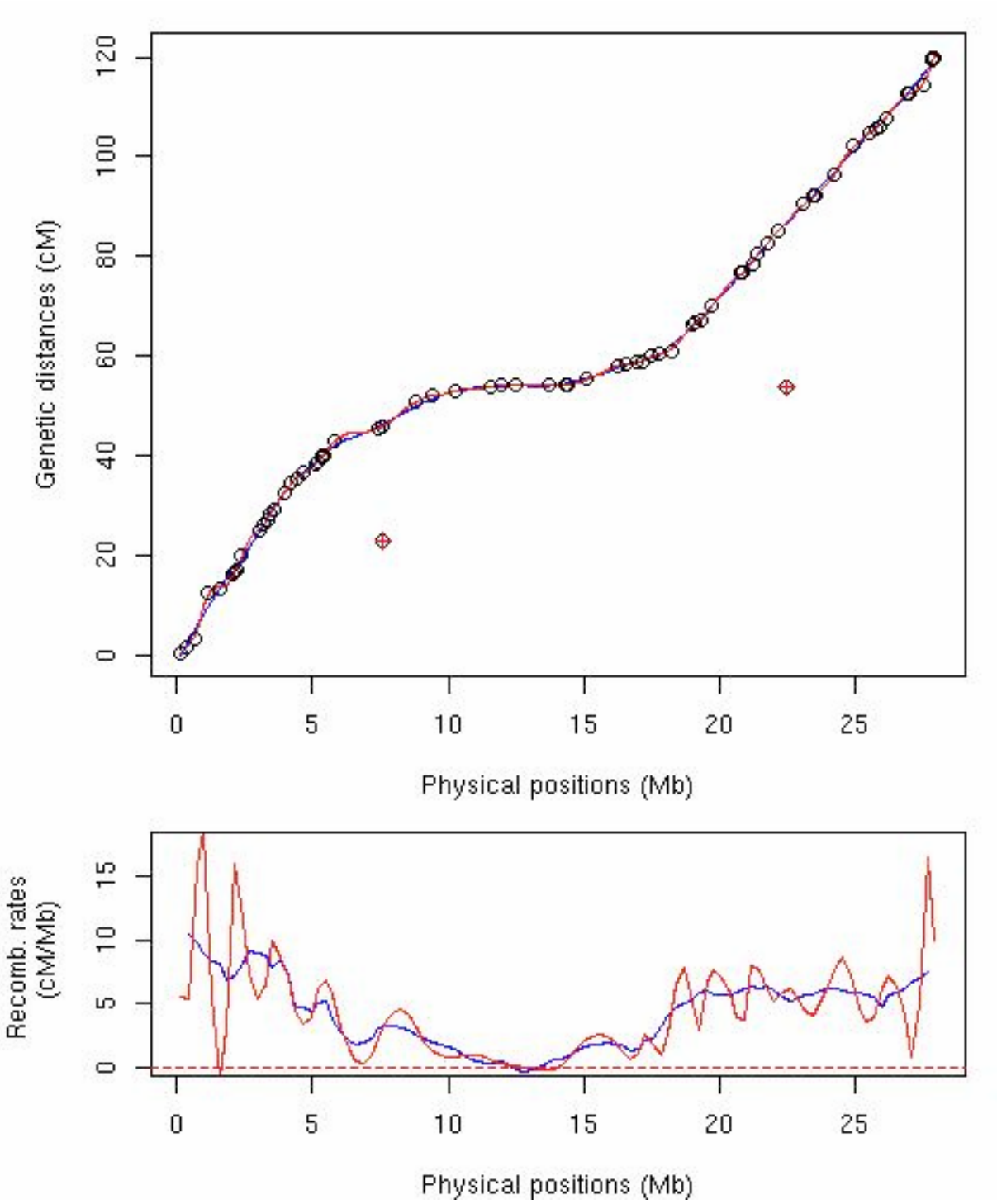
Chromosome 6



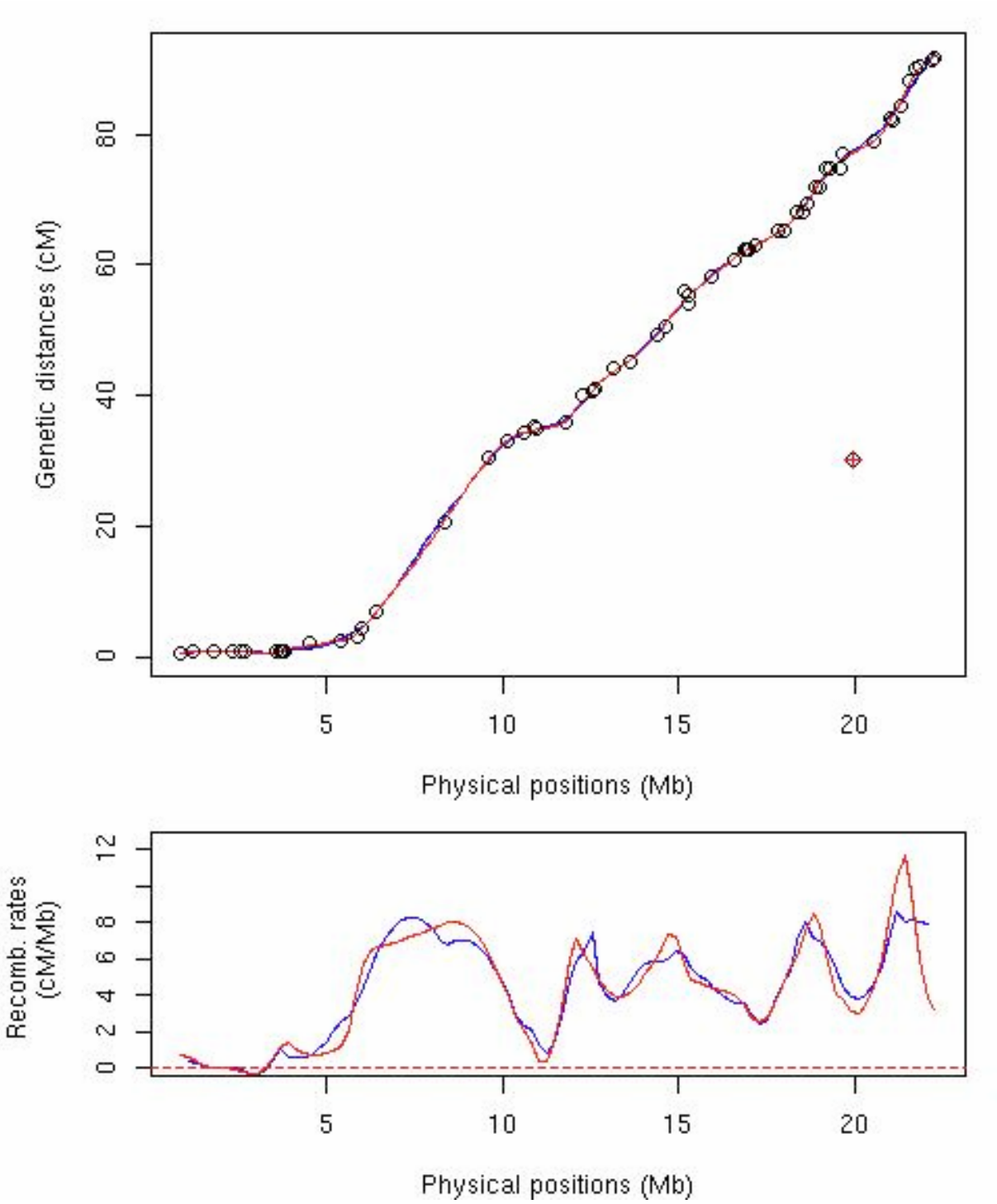
Chromosome 7



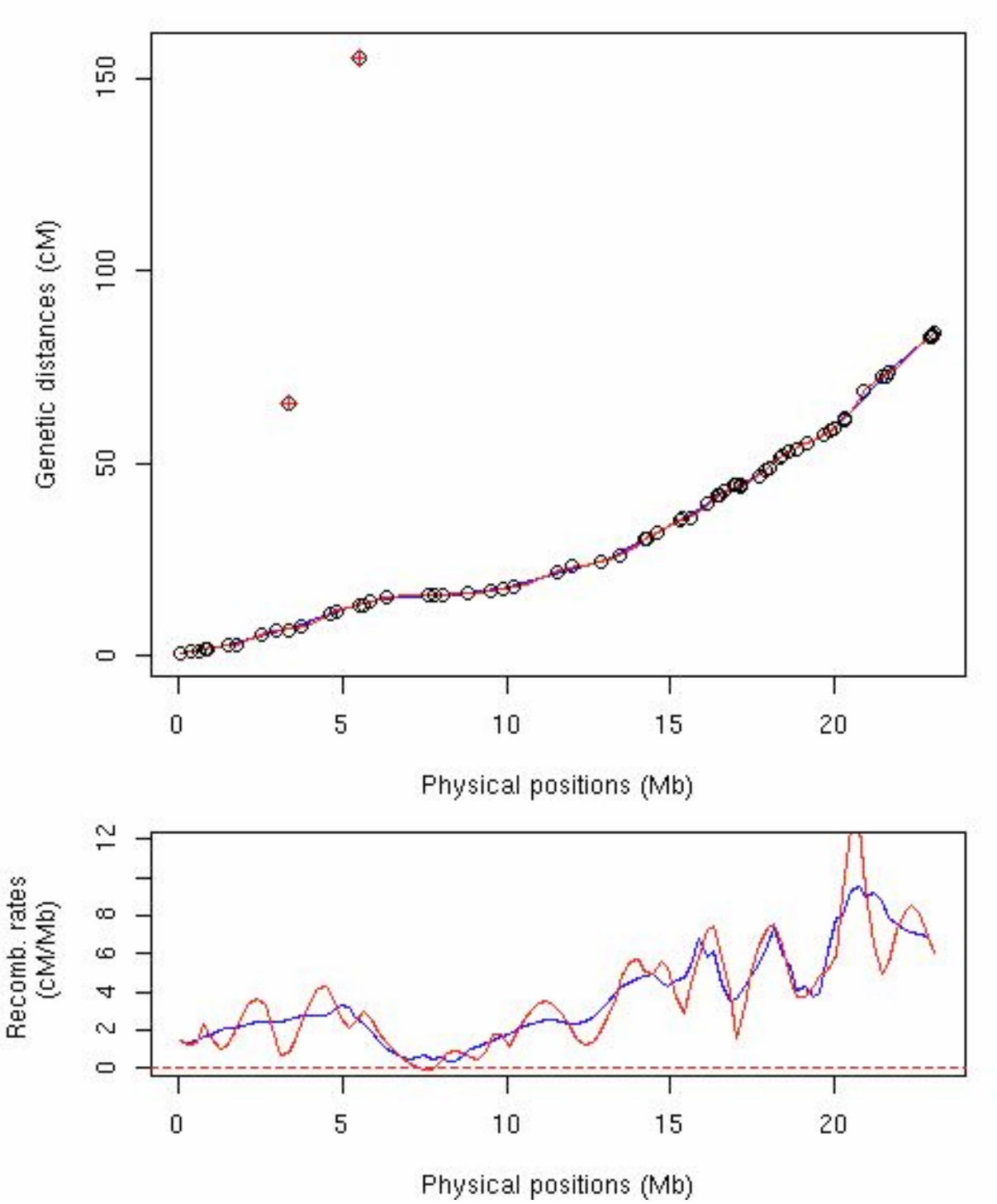
Chromosome 8



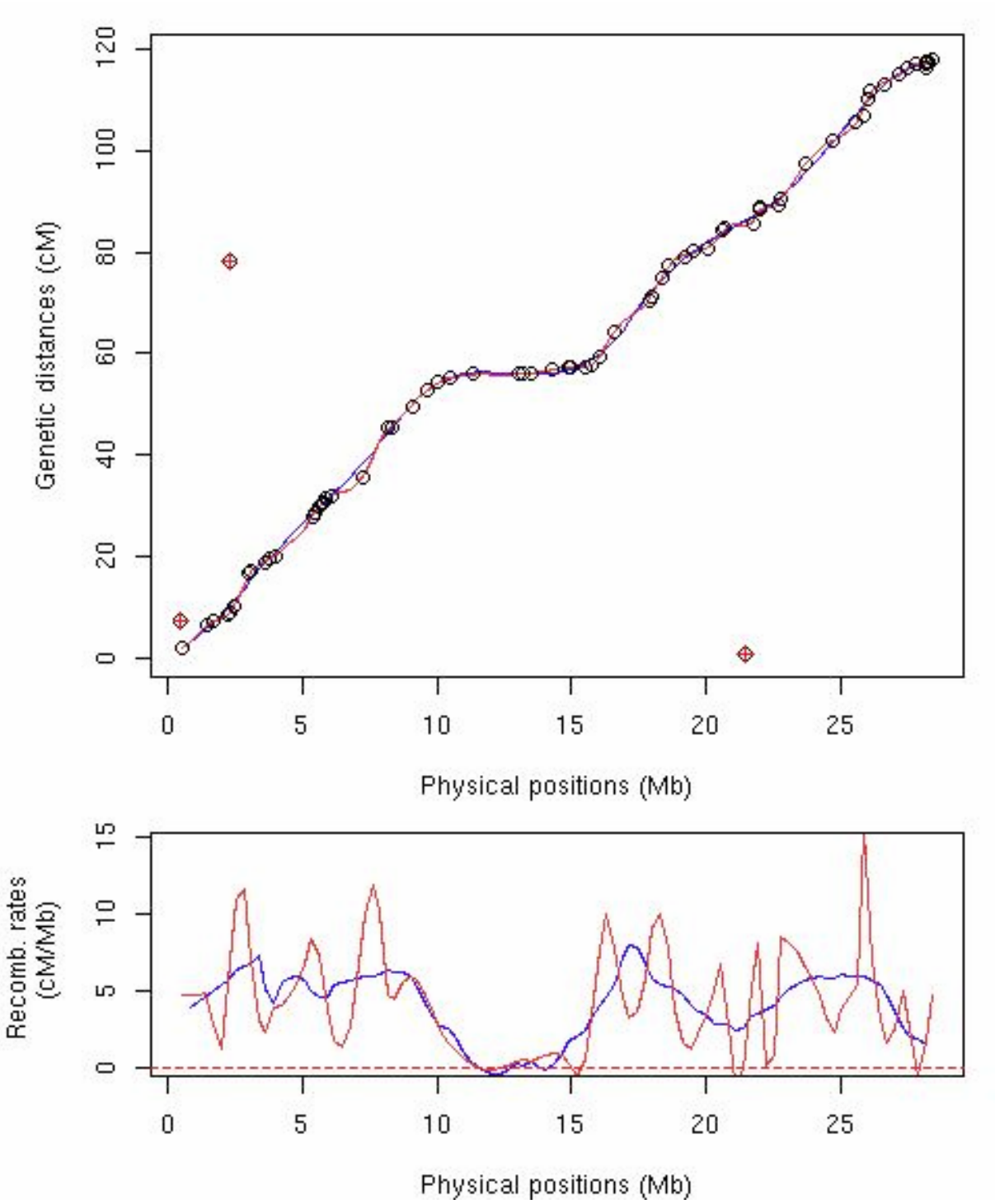
Chromosome 9



Chromosome 10



Chromosome 11



Chromosome 12

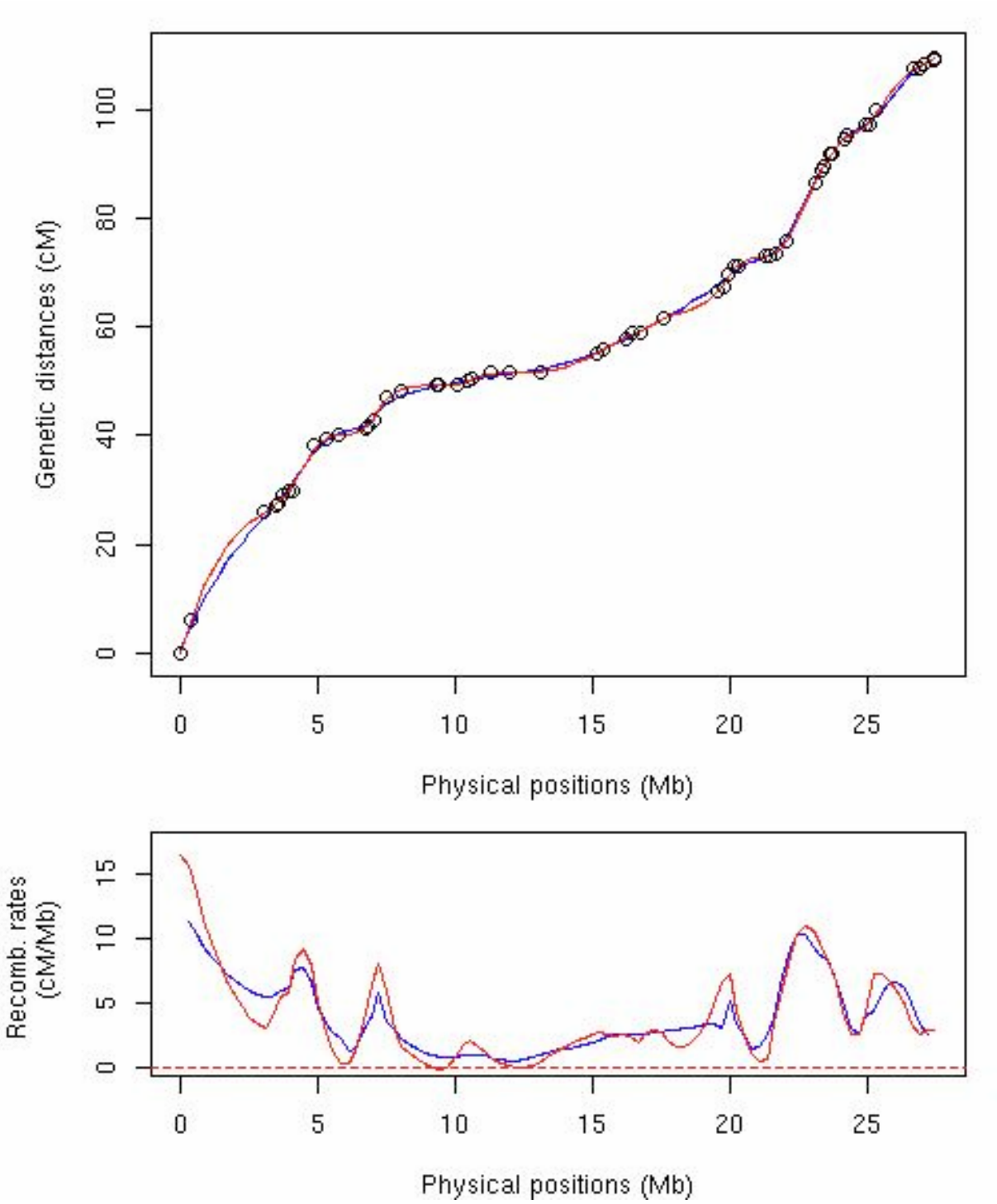


Figure S2

Derived allele frequency spectra. To cope with SNPs with missing data, spectra have been computed for a subsample of size $n = 16$ as in (Caicedo et al. 2007).

A: *O. rufipogon* orientation with *O. meridionalis*

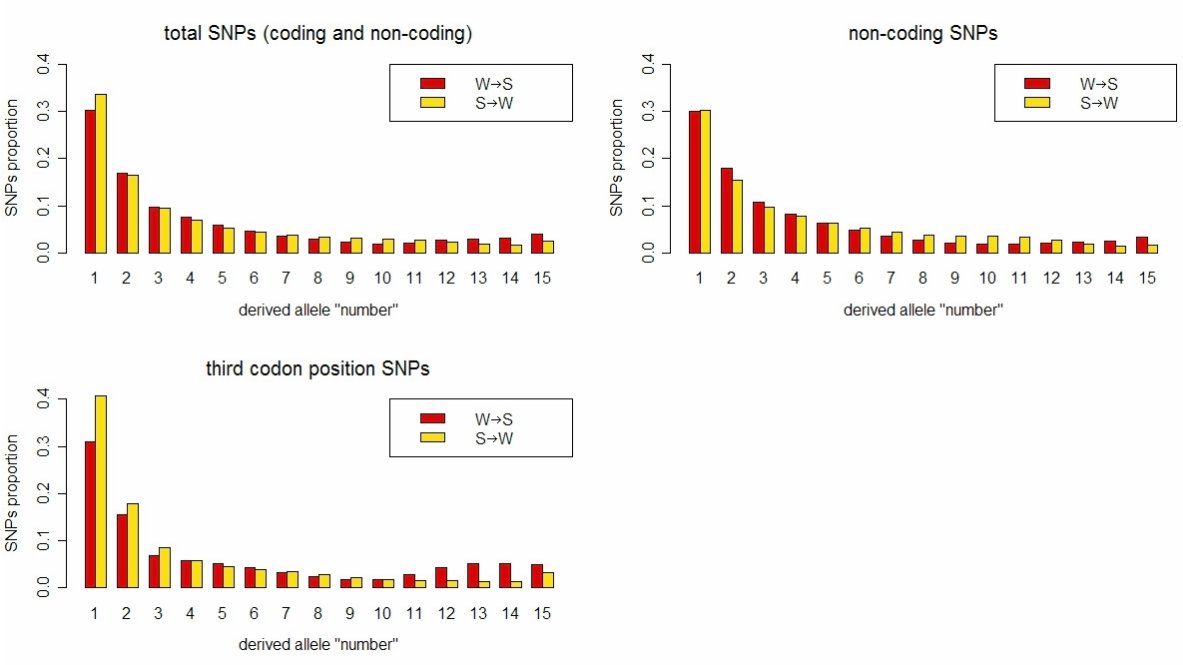
B: *O. s. japonica* orientation with *O. meridionalis*

C: *O. s. japonica* orientation with *O. barthii*

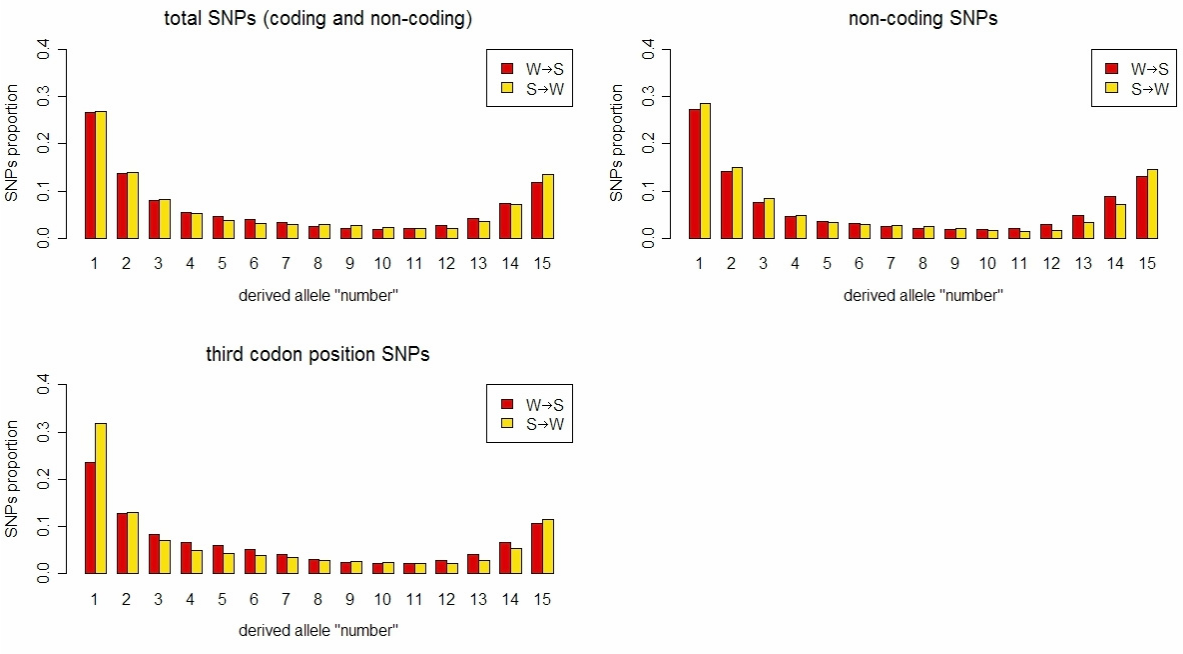
D: *O. s. indica* orientation with *O. meridionalis*

E: *O. s. indica* orientation with *O. barthii*

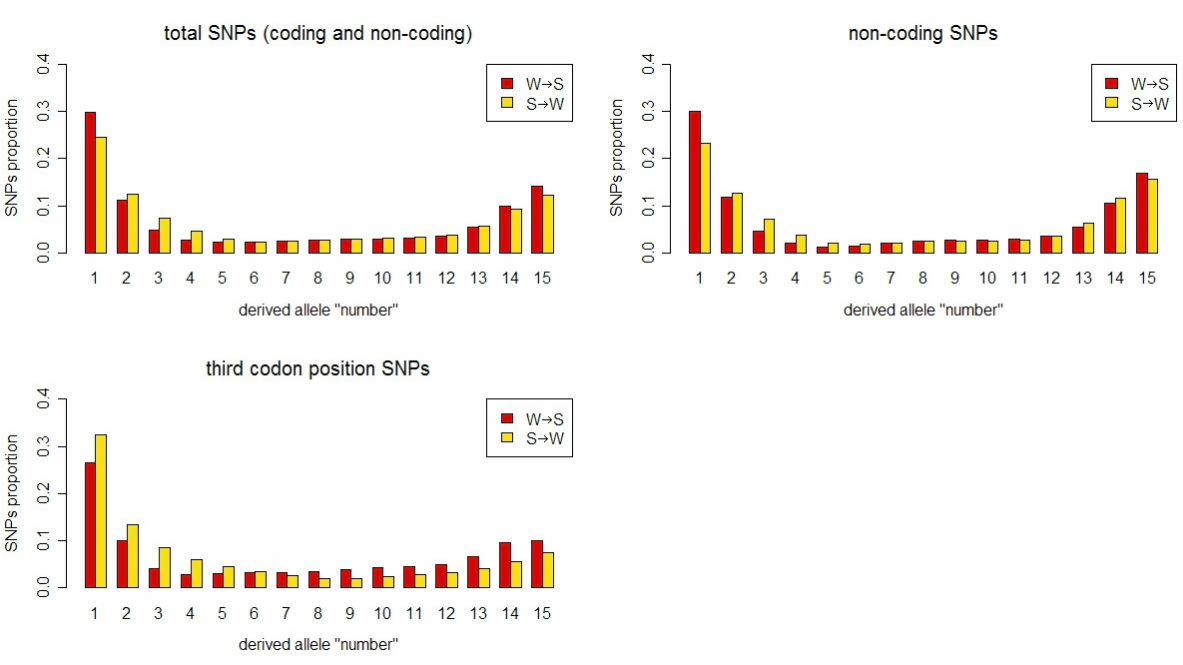
A


total SNPs (coding and non-coding)
non-coding SNPs
third codon position SNPs
W→S
S→W
SNPs proportion
derived allele "number"

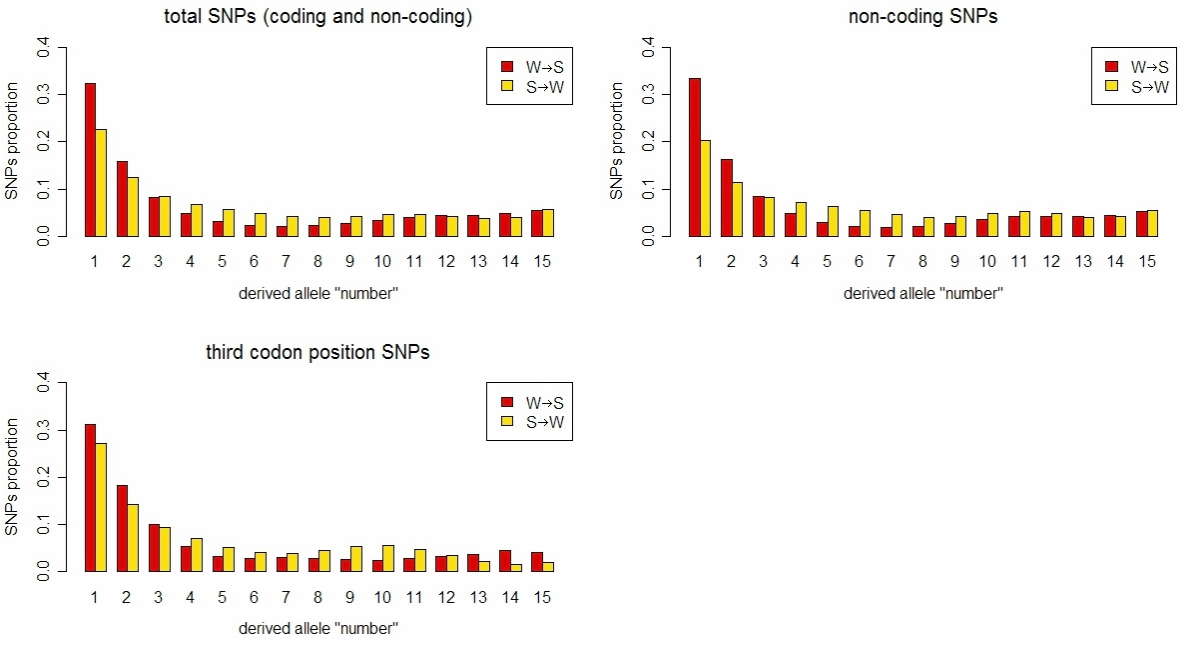
B


total SNPs (coding and non-coding)
non-coding SNPs
third codon position SNPs
W→S
S→W
SNPs proportion
derived allele "number"

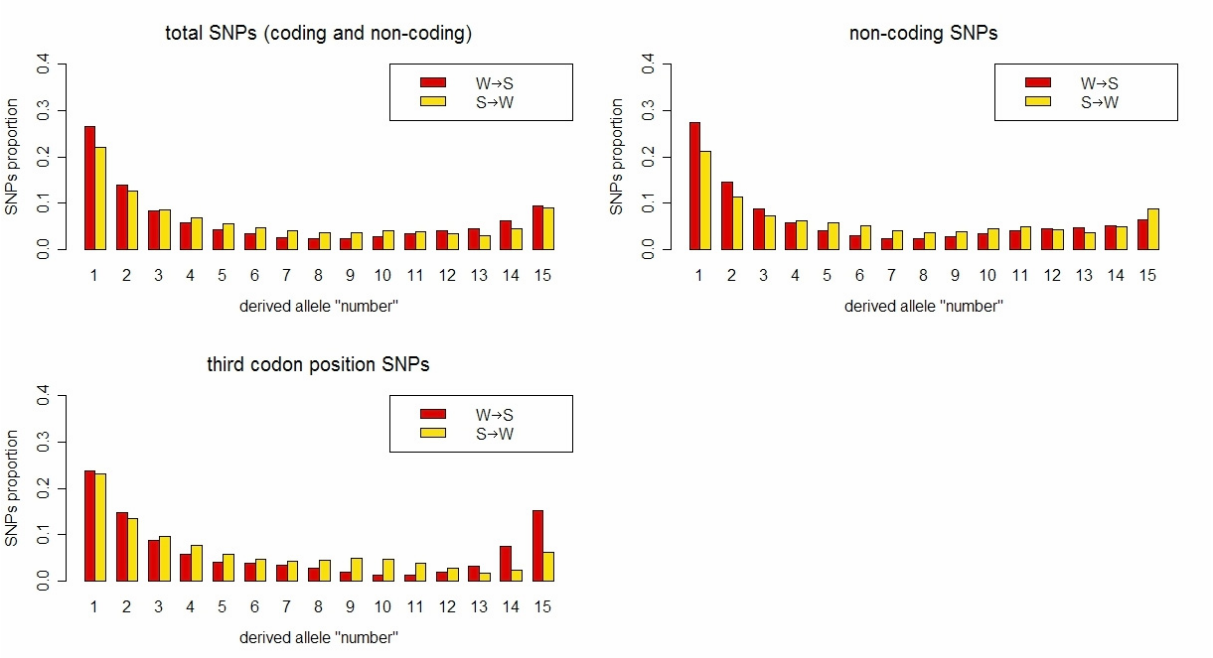
C


total SNPs (coding and non-coding)
non-coding SNPs
third codon position SNPs
W→S
S→W
SNPs proportion
derived allele "number"
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

D


total SNPs (coding and non-coding)
non-coding SNPs
third codon position SNPs
W→S
S→W
SNPs proportion
derived allele "number"
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

E


total SNPs (coding and non-coding)
non-coding SNPs
third codon position SNPs
W→S
S→W
SNPs proportion
derived allele "number"
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Annexes du chapitre 1

- Supplemental Figure 1. Comparisons of GC3 distribution computed with raw EST data, bi-Beta fits and complete genome sequences for 6 species

- Supplemental Figure 2: Comparison of the means and standard deviations estimated from the raw data and from the bi-Beta distributions

- Supplemental Figure 3: Relationship between the residuals of the regression of GC3 standard deviation against mean GC3 (bi-Beta estimates) and the number of EST sequences and unigenes

- Supplemental Figure 4: Boxplots of mean GC3 by Super Order and Order

- Supplemental Figure 5: Boxplots of GC3 standard deviation by Super Order and Order

- Supplemental Figure 6: Relationship between local recombination rate and GC1, GC2, and GCUTR in three grass species

- Supplemental Figure 7: Correlation between mean GC3 and GC3 standard variation averaged by taxonomic level

- Supplemental Figure 8: Correlation between mean GC3 and the two means of the bi-Beta distribution averaged by taxonomic level

- Supplemental Table 1: mean GC-content by taxonomic group and Kruskal-Wallis test for differences between groups

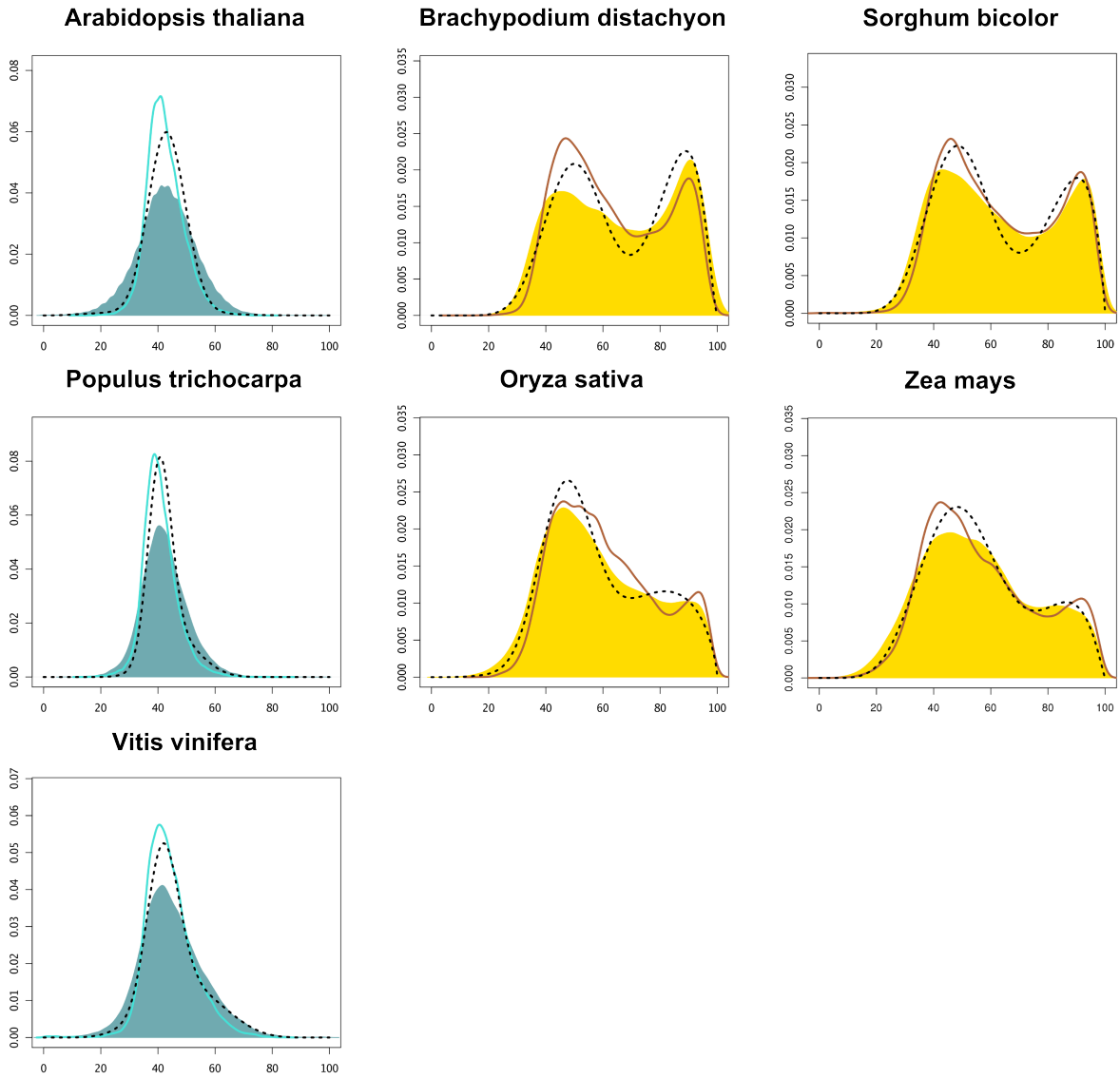
- Supplemental Table 2: GC-content standard deviations by taxonomic group and Kruskal-Wallis tests for differences between groups

- Supplemental Table 3: Non-parametric Spearman correlation coefficient (rho) and p-value of the correlation between local recombination rate and GC1, GC2, and GCUTR in three grass species

- Supplemental Table 4: Taxonomy levels used in the analyses. Supplemental Table 5: Correlation between mean GC3 and GC3 standard variation by taxonomic group

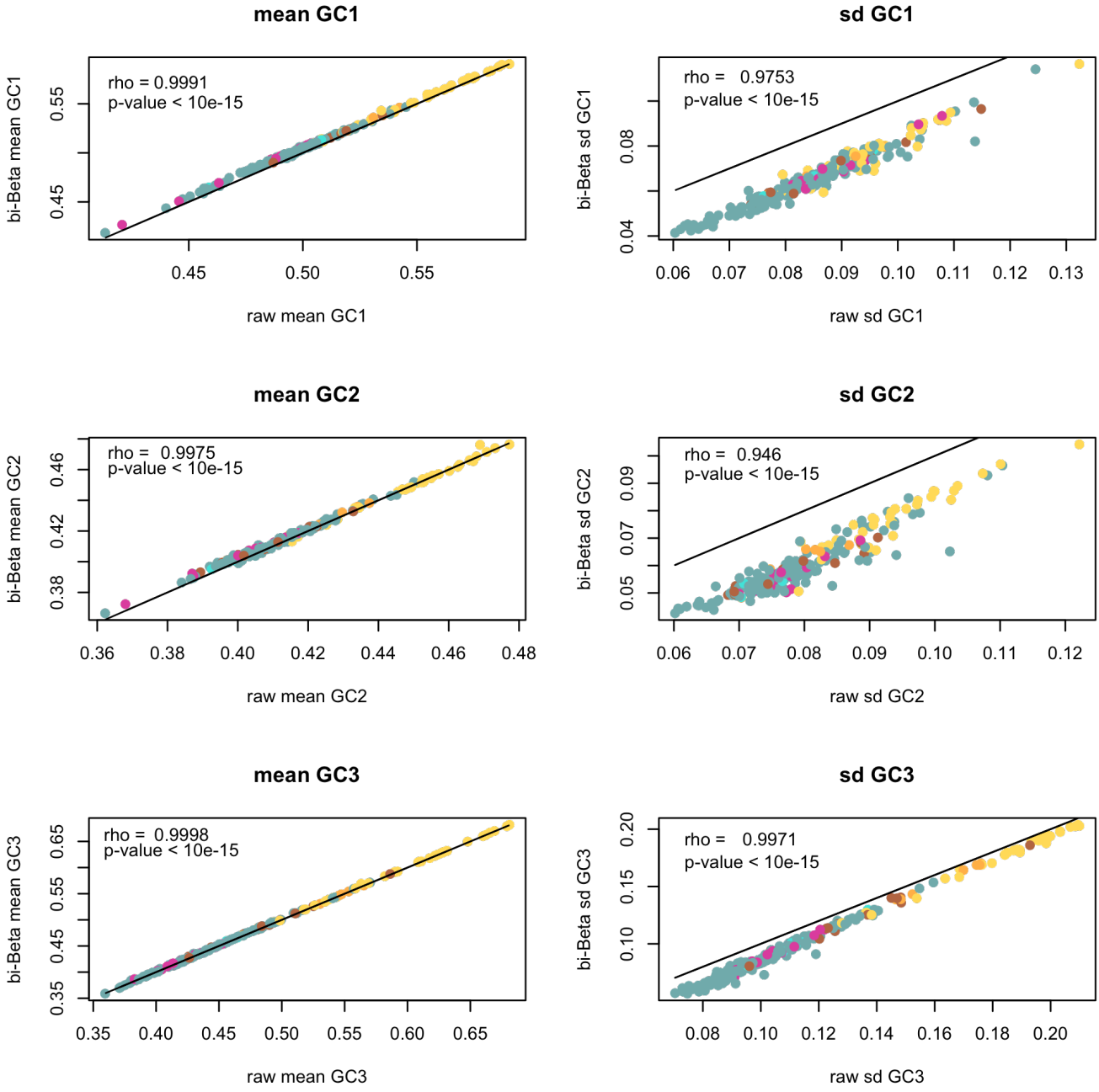
Supplemental Figure 1. Comparisons of GC3 distribution computed with raw EST data, bi-Beta fits and complete genome sequences for 6 species.

Filled distributions: GC3 computed with raw EST data. Dashed distribution: GC3 bi-Beta distribution fitted with raw EST data. Unfilled distributions: GC3 computed with complete genome sequences. Eudicots (in cadet blue): *Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera.* Monocots (in gold): *Brachypodium distachyon, Oryza sativa, Sorghum bicolor*, *Zea mays*.



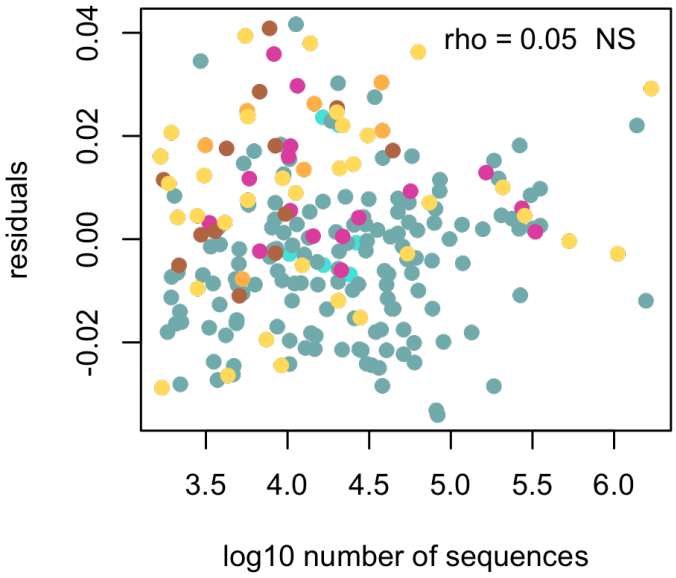
Supplemental Figure 2: Comparison of the means and standard deviations estimated from the raw data and from the bi-Beta distributions.

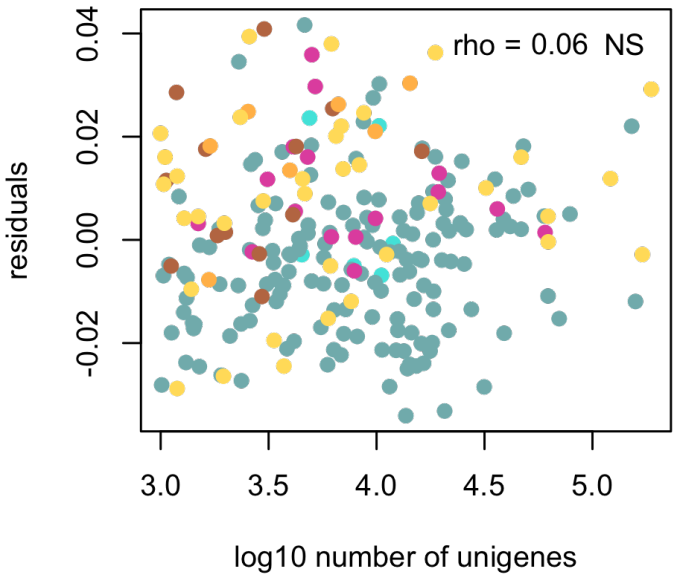
The colors correspond to the groups used in the main text: violet red: gymnosperms, turquoise: basal angiosperms, brown: non-commelinids monocots; orange: non-Poaceae commelinids; gold: Poaceae; cadet blue: eudicots. Black lines correspond to the bissectrices.



Supplemental Figure 3: Relationship between the residuals of the regression of GC3 standard deviation against mean GC3 (bi-Beta estimates) and the number of EST sequences and unigenes.

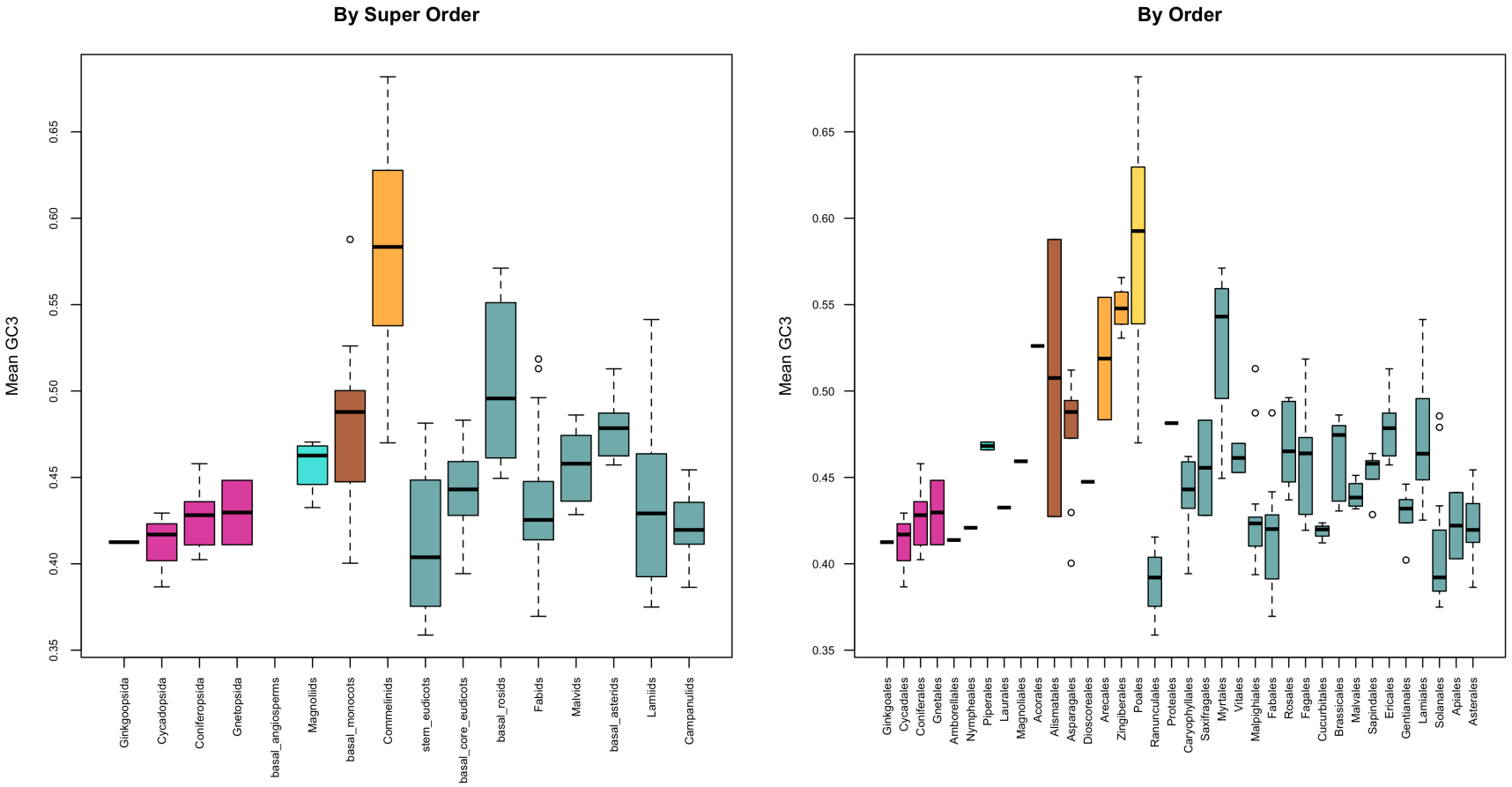
The colors correspond to the groups used in the main text: violet red: gymnosperms, turquoise: basal angiosperms, brown: non-commelinids monocots; orange: non-Poaceae commelinids; gold: Poaceae; cadet blue: eudicots.





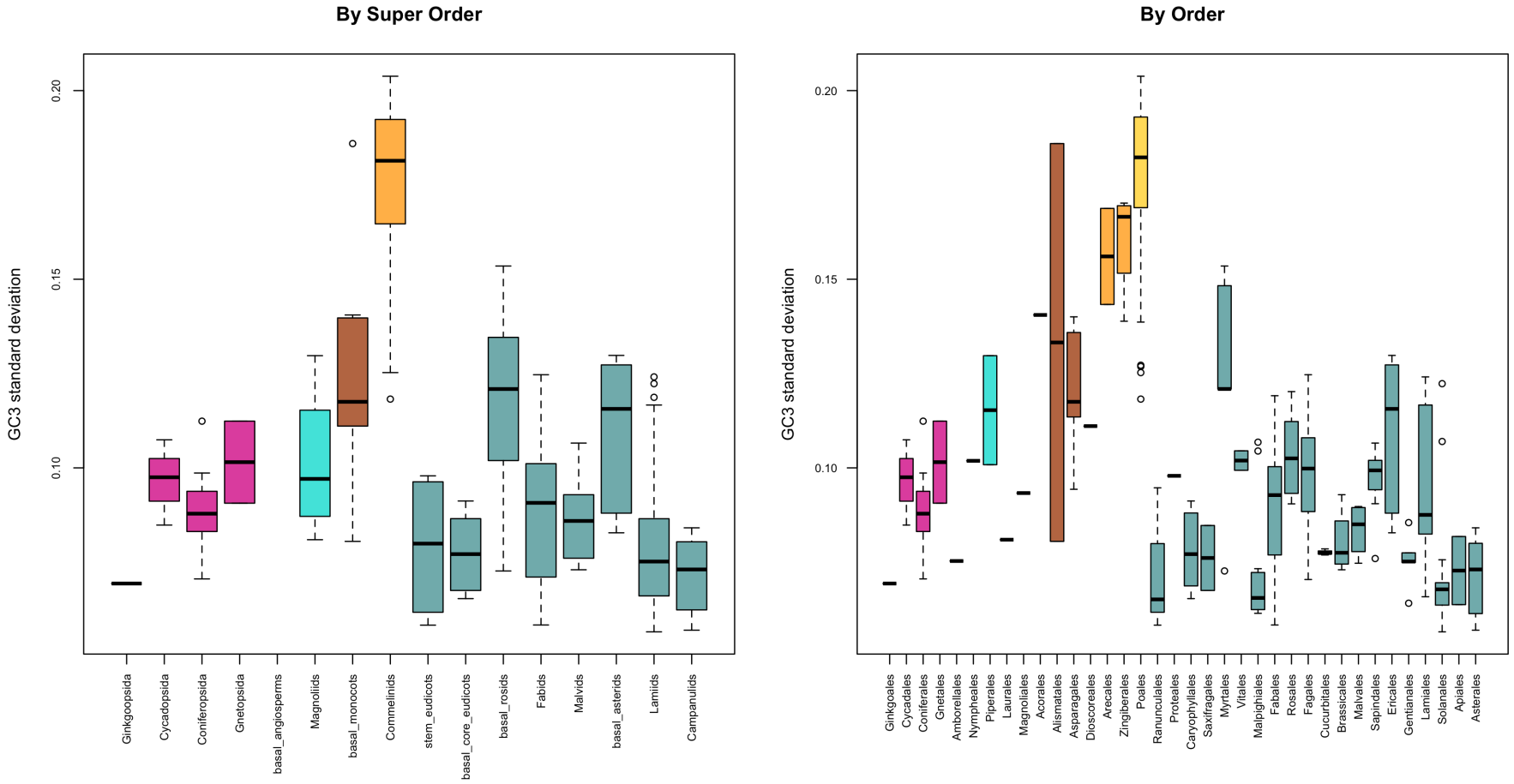
Supplemental Figure 4: Box plots of mean GC3 by Super Order and Order.

The colors correspond to the groups used in the main text: violet red: gymnosperms, turquoise: basal angiosperms, brown: non-commelinids monocots; orange: non-Poaceae commelinids; gold: Poaceae; cadet blue: eudicots. On each box plot the dark midline is the median, the bottom and top of the box are the lower and upper quartiles, respectively, and the ends of the whiskers are the lowest, respectively highest, data still within 1.5 times the inter-quartile range of the lower, respectively higher, quartile.



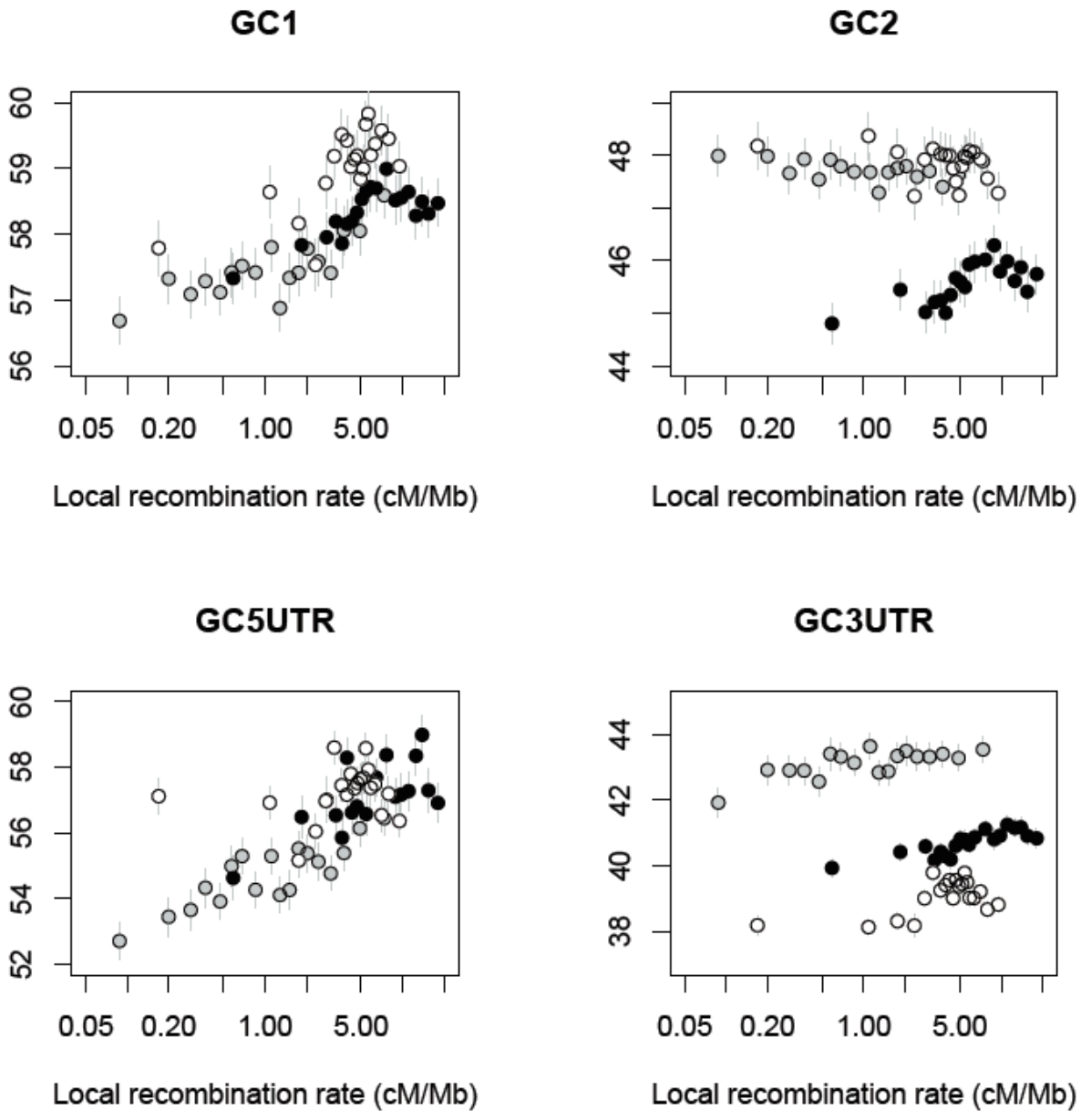
Supplemental Figure 5: Box plots of GC3 standard deviation by Super Order and Order.

The colors correspond to the groups used in the main text: violet red: gymnosperms, turquoise: basal angiosperms, brown: non commelinids monocots; orange: non Poaceae commelinids; gold: Poaceae; cadet blue: eudicots. On each box plot the dark midline is the median, the bottom and top of the box are the lower and upper quartiles, respectively, and the ends of the whiskers are the lowest, respectively highest, data still within 1.5 times the inter-quartile range of the lower, respectively higher, quartile.

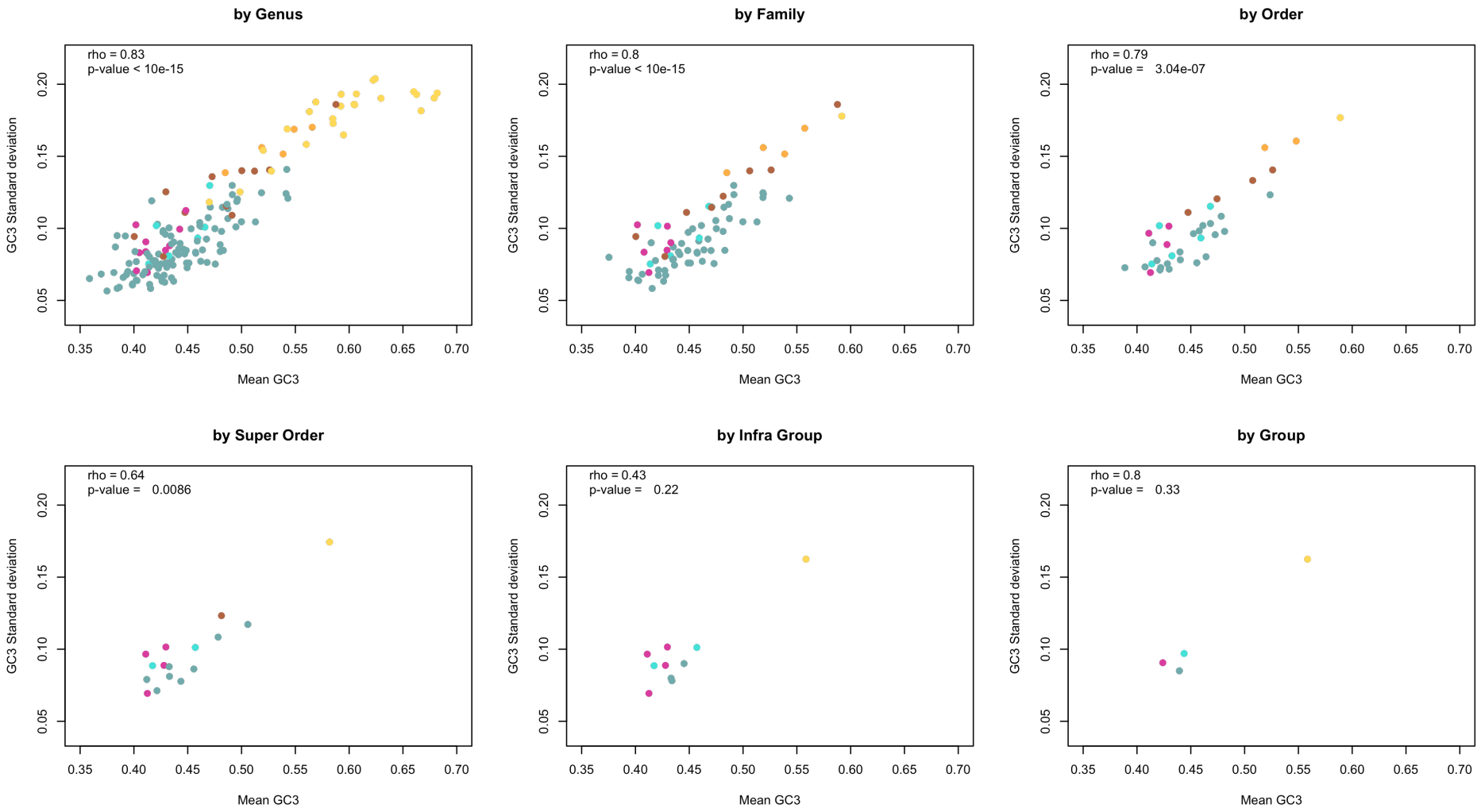


Supplemental Figure 6: Relationship between local recombination rate and GC1, GC2, and GCUTR in three grass species

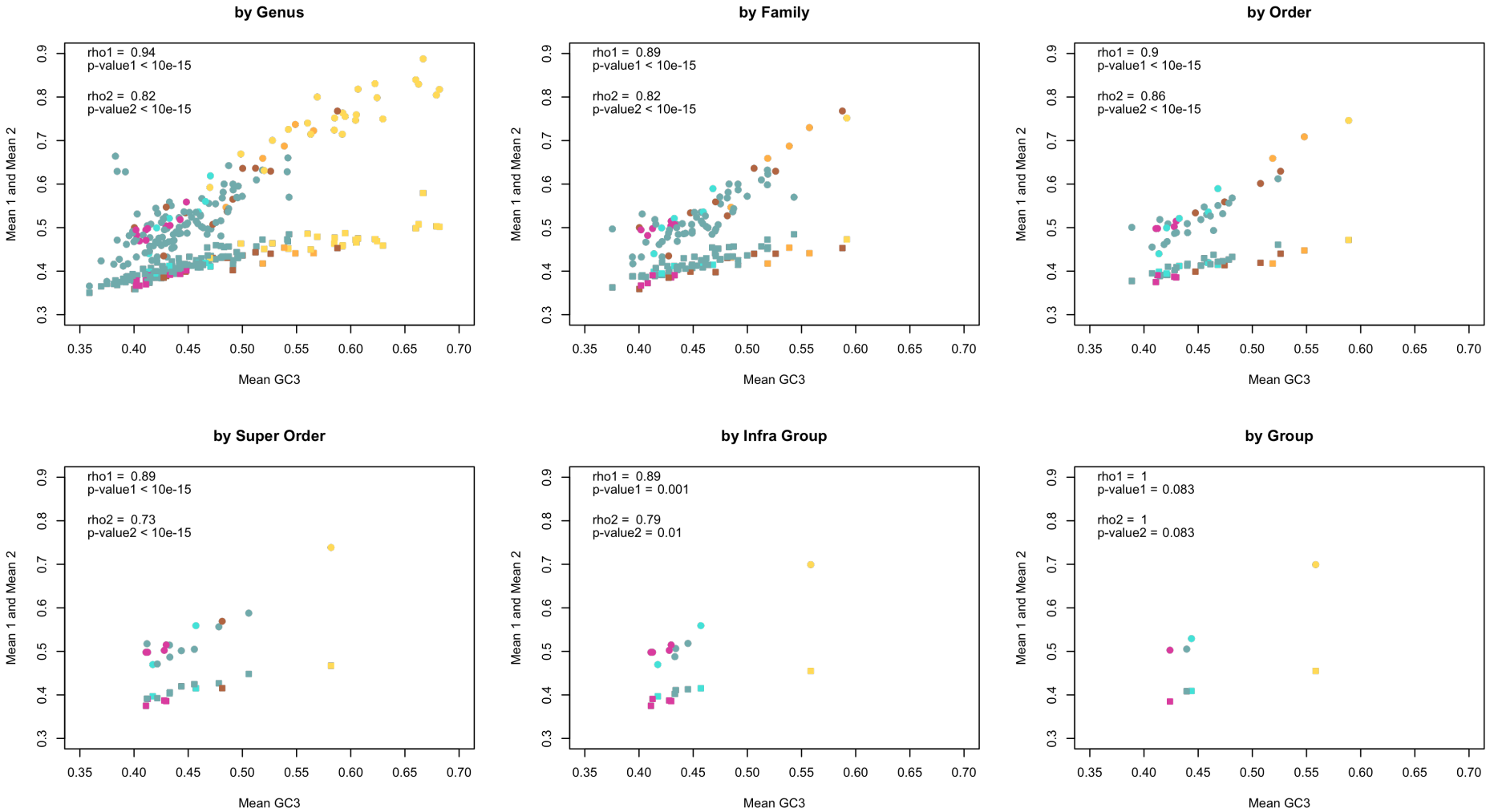
Genes have been grouped into 20 bins according to their local recombination rate. Dots correspond to the mean GC-content of each bin and bars to the standard error. Black dots: *Brachypodium distachyon*, grey dots: maize, white dots: rice.



Supplemental Figure 7: Correlation between mean GC3 and GC3 standard variation averaged by taxonomic level



Supplemental Figure 8: Correlation between mean GC3 and the two means of the bi-Beta distribution averaged by taxonomic level



Supplemental Table 1: mean GC-content by taxonomic group and Kruskal-Wallis test for differences between groups.

GC mean	GC1	GC2	GC3	GC5UTR	GC3UTR
Means per group					
gymnos	49.1%	40.6%	42.4%	40.1%	35.7%
basal angios	49.3%	40.4%	44.4%	40.4%	36.9%
monocots	53.9%	43.6%	55.8%	49.6%	40.4%
non commelinids monocots	50.6%	41.1%	48.1%	43.3%	36.5%
commelinids	55.0%	44.4%	58.2%	51.5%	41.6%
non Poaceae commelinids	53.0%	42.6%	53.1%	47.2%	38.3%
Poaceae	55.4%	44.8%	59.2%	52.3%	42.2%
eudicots	49.9%	41.2%	43.9%	40.3%	35.4%
Kruskal-Wallis tests					
Gym, Bas, Eud	NS	NS	NS	NS	*
Gym, Mon	***	***	***	***	***
Bas, Mon	**	**	***	**	**
Eud, Mon	***	***	***	***	***
noCom Mon, Com	***	***	***	***	***
noPoa Com, Poa	*	**	*	**	***

NS: p-value>0.05; *: p-value<0.05; **: p-value<0.01; ***: p-value<0.001.
Gym: Gymnosperms; Bas: basal angiosperms; Mon: all monocots; noCom Mon: non-commelinids monocots; Com: all commelinids; noPoa Com: non-Poaceae commelinids; Poa: Poaceae; Eud: eudicots.
Unbiased mean GC for GC1, GC2, GC3. Raw mean GC for GC5UTR and GC3UTR.

Supplemental Table 2: GC-content standard deviations by taxonomic group and Kruskal-Wallis tests for differences between groups.

GC standard-deviation	GC1	GC2	GC3	GC5UTR	GC3UTR
Means per group					
gymnos	6.93	5.56	9.06	7.58	5.69
basal angios	6.89	5.20	9.7	8.08	5.73
monocots	7.40	6.97	16.25	11.67	7.32
Non-commelinids monocots	6.74	5.75	12.33	9.94	7.01
commelinids	7.61	7.34	17.43	12.19	7.42
Non-Poaceae commelinids	6.73	6.21	15.61	10.63	6.67
Poaceae	7.78	7.56	17.79	12.49	7.57
eudicots	6.35	5.74	8.5	8.32	6.82
Kruskal-Wallis tests					
Gym, Bas, Eud	NS	NS	NS	**	***
Gym, Mon	NS	***	***	***	***
Bas, Mon	NS	**	***	***	***
Eud, Mon	***	***	***	***	***
noCom Mon, Com	*	***	***	***	NS
noPoa Com, Poa	*	**	**	***	*

NS: p-value>0.05; *: p-value<0.05; **: p-value<0.01; ***: p-value<0.001.
Gym: Gymnosperms; Bas: basal angiosperms; Mon: all monocots; noCom Mon: non-commelinids monocots; Com: all commelinids; noPoa Com: non-Poaceae commelinids; Poa: Poaceae; Eud: eudicots.
Unbiased standard deviation GC for GC1, GC2, GC3. Raw standard deviation GC for GC5UTR and GC3UTR.

Supplemental Table 3: Non-parametric Spearman correlation coefficient (rho) and p-value of the correlation between local recombination rate and GC1, GC2, and GCUTR in three grass species.

The correlations are computed using every gene as a data point. For *O. sativa* and *Z.mays* results are given for the total dataset including all annotated genes or only protein coding genes (excluding pseudogenes, transposable elements, and miRNA).

Species	dataset		GC1	GC2	GC3	GC5UTR	GC3UTR
O. sativa	protein-coding	rho	0.045	-0.001	0.081	0.002	0.027
		p-value	<10e-15	0.853	<10e-16	0.7628	7.49E-05
	total	rho	0.089	0.077	0.138	0.005	0.025
		p-value	<10e-15	<10e-15	<10e-15	0.4434	1.74E-04
Z. mays	protein-coding	rho	0.027	-0.014	0.068	0.069	0.047
		p-value	5.04E-11	0.0007625	<10e-15	<10e-15	<10e-15
	total	rho	0.054	0.038	0.081	0.075	0.075
		p-value	<10e-15	<10e-15	<10e-15	<10e-15	<10e-15
B. distachyon	unfiltered	rho	0.036	0.041	0.052	0.057	0.063
		p-value	7.34E-09	3.38E-11	<10e-16	2.47E-07	<10e-16

Supplemental Table 4: Taxonomy levels used in the analyses.

Group	Infra group	Super order
Gymnosperms	Coniferopsida	Coniferopsida
	Cycadopsida	Cycadopsida
	Ginkgoopsida	Ginkgoopsida
	Gnetopsida	Gnetopsida
basal angiosperms	basal angiosperms	basal angiosperms
Monocots+Magnoliids	Magnoliids	Magnoliids
	Monocots	basal monocots
		Commelinids
Eudicots	basal eudictos	stem eudicots
		basal core eudicots
	Rosids	basal rosids
		Fabids
		Malvids
	Asterids	basal asterids
		Lamiids
		Campanulids

Upper case (lower case) is used for monophyletic (paraphyletic) groups, respectively.

Supplemental Table 5: Correlation between mean GC3 and GC3 standard variation by taxonomic group

Group	Infra Group	Super Order	Order	Family	Genus	# species	Spearman's rho	p-value
Gymnosperms						16	**0.60**	**0.016**
		Coniferopsida	Coniferales			10	**0.95**	**0.000**
				Pinaceae		8	**0.98**	**0.000**
					Pinus	4	*1.00*	*0.083*
basal angiosperms						6	0.66	0.175
		Magnoliids	Magnoliales			4	*1.00*	*0.083*
Monocots						56	**0.91**	**0.000**
		basal monocots				13	**0.80**	**0.002**
			Asparagales			9	0.55	0.133
		Commelinids				43	**0.85**	**0.000**
			Zingiberales			4	0.80	0.333
			Poales			37	**0.81**	**0.000**
				Poaceae		36	**0.80**	**0.000**
Eudicots						154	**0.70**	**0.000**
	basal eudicots					13	0.45	0.130
		stem eudicots				4	0.40	0.750
		basal core eudicots			9	0.43	0.250	
			Caryophyllales			7	0.36	0.444
	Rosids					80	**0.65**	**0.000**
		basal rosids				7	**0.93**	**0.007**
			Myrtales			5	0.80	0.133
				Myrtaceae		4	0.80	0.333
		Fabids				52	**0.69**	**0.000**
			Malpighiales			13	0.32	0.289
				Euphorbiaceae	5	0.00	1.000	
				Salicacecae	*Populus*	6	0.54	0.297
			Fabales	Fabaceae		21	**0.68**	**0.001**
			Rosales			8	**0.83**	**0.015**
				Rosaceae		6	0.77	0.103
			Fagales			7	**0.89**	**0.012**
				Fagaceae		4	*1.00*	*0.083*
		Malvids				20	0.16	0.484
			Brassicales			10	0.54	0.113
				Brassicaceae		5	0.10	0.950
			Malvales	Malvaceae		4	0.80	0.333
			Sapindales			7	0.57	0.200
				Rutaceae	*Citrus*	6	0.37	0.497
	Asterids					61	**0.75**	**0.000**
		basal asterids	Ericales			7	**0.86**	**0.024**
		Lamiids				27	**0.92**	**0.000**
			Gentianales			5	0.60	0.350
				Rubiaceae		4	0.20	0.917
			Lamiales			9	**0.90**	**0.002**
			Solanales			13	**0.88**	**0.000**
				Solanaceae		11	**0.81**	**0.004**
					Solanum	6	0.71	0.136
		Campanulids				27	0.21	0.287
			Asterales	Asteraceae		25	0.11	0.589
					Helianthus	7	*0.75*	*0.066*
					Lactuca	5	0.30	0.683

Annexes du chapitre 2

- Tableau 1 – Gènes ARNt et préférences d'usage du code chez 5 espèces de plantes
- Synopsis des préférences d'usage du code de 130 espèces de plantes
- Arbres de préférences d'usage des codons pour 130 espèces de plantes
- Reconstruction des préférences ancestrales pour 62 codons

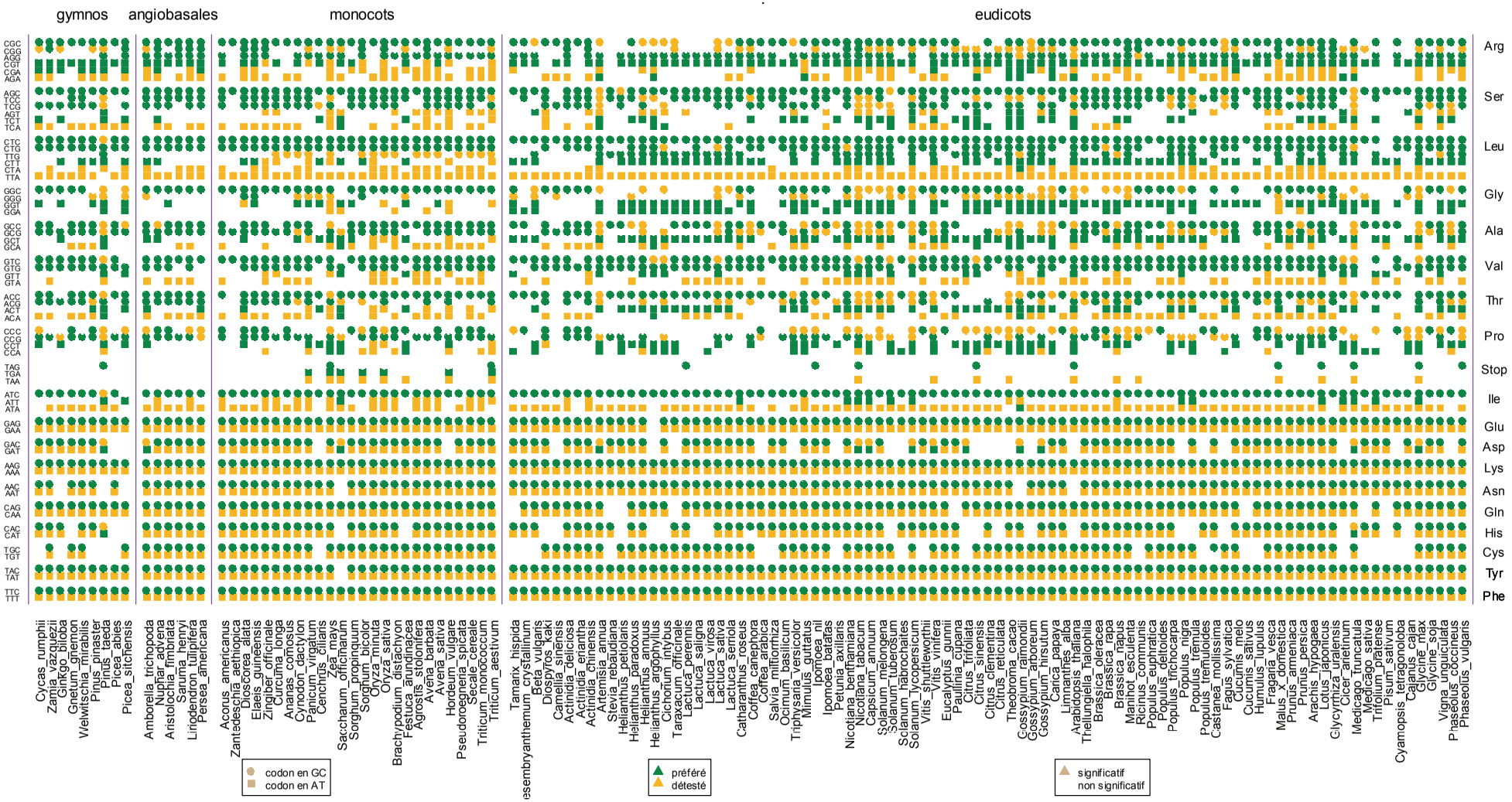
Tableau 1 – Gènes ARNt et préférences d'usage du code chez 5 espèces de plantes

Les gènes ARNt ont été déterminés à partir des séquences de génomes complets à l'aide du logiciel tRNAscan-SE et leur annotation a été vérifiée à la main, par Michaud et collaborateurs (données de Michaud et al. 2011 et données supplémentaires disponibles sur http://www-ibmp.u-strasbg.fr/Drouard/tRNA.html). Nous avons comptabilisé les gènes des différents ARNt après avoir éliminé les séquences annotées comme étant chloroplastiques, mitochondriales ou pseudogéniques.
Le RGF (*Relative gene frequency of tRNA genes*) est la fréquence observée d'un gène d'ARNt divisée par la fréquence attendue si tous les gènes d'ARNt isoaccepteurs possibles pour cet acide aminé étaient présents en proportions égales (Duret 2000). Un ARNt dont le RGF est supérieur à 1 est donc plus fréquent qu'attendu.
Cre : *Chlamydomonas reinhardtii*. Osa : *Oryza sativa*. Bdi : *Brachypodium distachyon*. Ath : *Arabidopsis thaliana*. Mtr : *Medicago truncatula*. Ptr : *Populus trichocarpa*.
Codons préférés : en vert. Codons neutres : en jaune. Codons détestés : en orange.

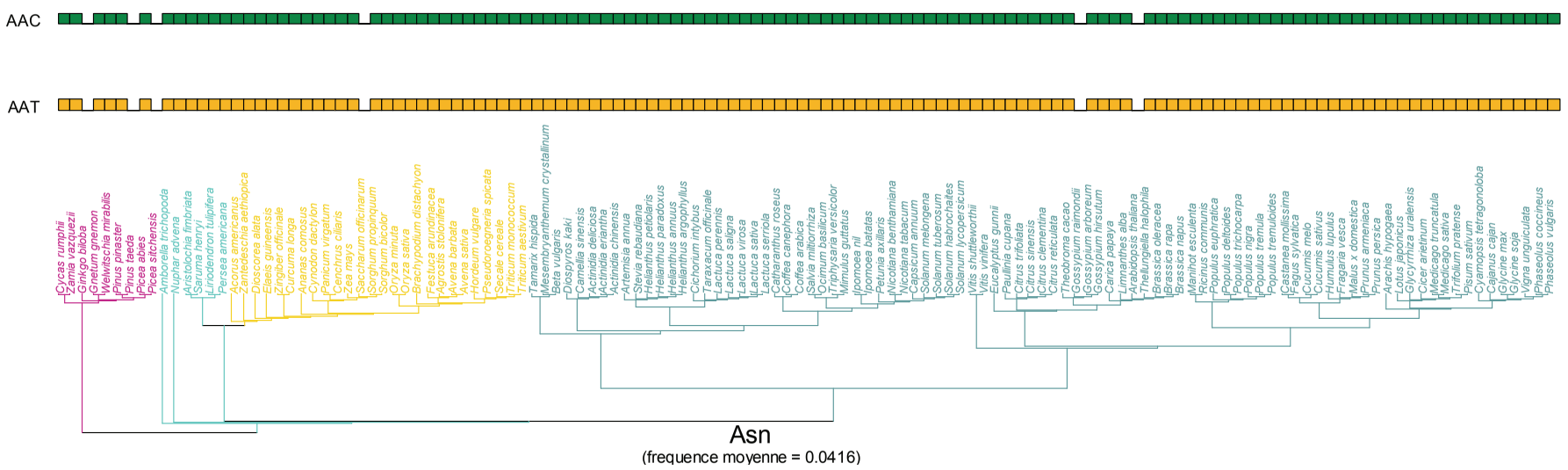
Acide aminé	Anticodon	Codon W&C	Codon wobble	Nombre de gènes						RGF					
				Cre	Osa	Bdi	Ath	Mtr	Ptr	*Cre*	Osa	Bdi	Ath	Mtr	Ptr
Met-e	CAT	ATG		*6*	13	17	11	12	13	*0.43*	0.52	0.65	0.52	0.46	0.52
Met-i	CAT	ATG		*8*	12	9	10	14	12	*0.57*	0.48	0.35	0.48	0.54	0.48
Trp	CCA	TGG		*5*	13	11	13	8	14	*1*	1	1	1	1	1
Asn	GTT	AAC	AAT	*7*	15	11	15	23	19	***2***	**2**	**2**	**2**	**2**	**2**
Asp	GTC	GAC	GAT	*11*	24	19	25	25	32	***2***	**2**	**2**	**2**	**2**	**2**
Cys	GCA	TGC	TGT	*7*	9	10	14	9	11	***2***	**2**	**2**	**2**	**2**	**2**
Gln	CTG	CAG		*6*	10	11	9	5	11	***1.71***	0.87	**1.29**	**1.13**	0.71	1
Gln	TTG	CAA	CAG	*1*	13	6	7	9	11	*0.29*	**1.13**	0.71	0.88	**1.29**	1
Glu	CTC	GAG		*13*	26	19	13	11	21	***1.86***	**1.41**	**1.31**	**1.13**	0.85	**1.20**
Glu	TTC	GAA	GAG	*1*	11	10	10	15	14	*0.14*	0.59	0.69	0.87	**1.15**	0.80
His	GTG	CAC	CAT	*5*	11	10	10	12	12	***2***	**2**	**2**	**2**	**2**	**2**
Lys	CTT	AAG		*11*	18	18	17	16	17	***1.83***	**1.38**	**1.29**	**1.13**	1	1
Lys	TTT	AAA	AAG	*1*	8	10	13	16	17	*0.17*	0.62	0.71	0.87	1	1
Phe	GAA	TTC	TTT	*9*	14	13	16	19	18	***2***	**2**	**2**	**2**	**2**	**2**
Tyr	ATA	TAT		*0*	0	0	0	1	0	*0*	0	0	0	0.20	0
Tyr	GTA	TAC	TAT	*8*	13	12	70	9	17	***2***	**2**	**2**	**2**	**1.80**	**2**

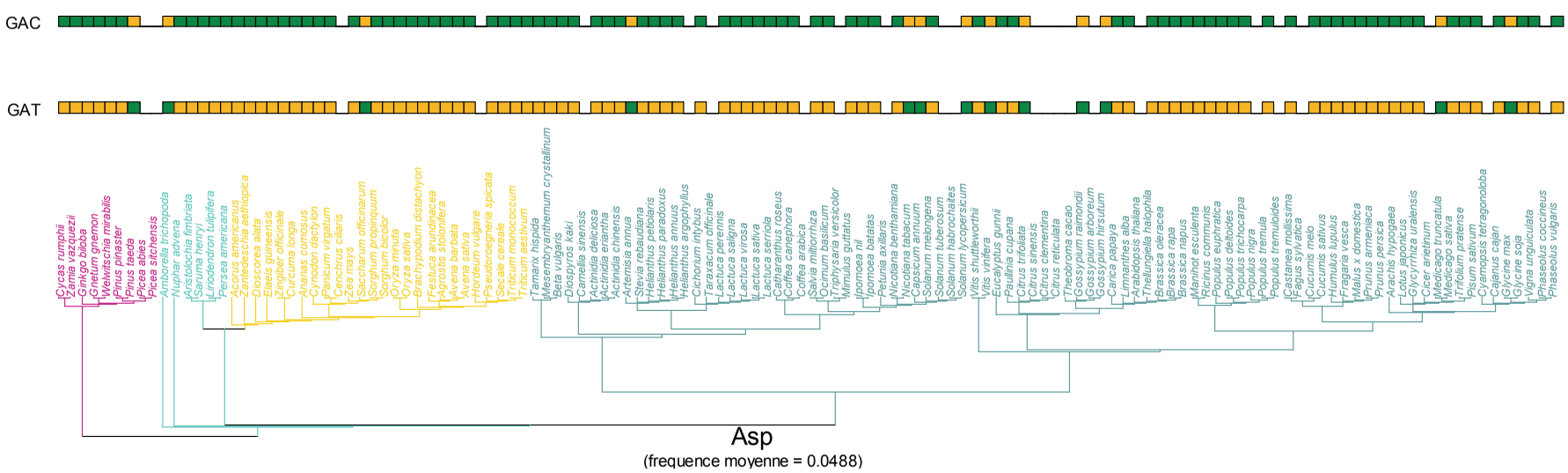
Acide aminé	Anticodon	Codon W&C	Codon wobble	Nombre de gènes						RGF						RGF, préférence du codon W&C, préférence du codon wobble					
				Cre	Osa	Bdi	Ath	Mtr	Ptr	*Cre*	Osa		Bdi		Ath		Mtr		Ptr		
Ile	AAT	ATT	ATC - (ATA)	*7*	18	19	17	18	19	***2.33***	**2.35**		**2.28**		**2.32**		**2**		**2.11**		
Ile	GAT	ATC		*1*	0	2	0	0	1	*0.33*	0		0.24		0		0		0.11		
Ile	TAT	ATA		*1*	5	4	5	9	7	*0.33*	0.65		0.48		0.68		1		0.78		
Ala	AGC	GCT	GCC - (GCA)	*13*	20	19	16	25	20	***1.86***	**1.86**		**1.69**		**1.94**		**2.33**		**1.95**		
Ala	GGC	GCC		*0*	1	0	0	0	0	*0*	0.09		0		0		0		0		
Ala	CGC	GCG		*10*	12	13	7	2	5	***1.43***	**1.12**		**1.16**		0.85		0.19		0.49		
Ala	TGC	GCA	GCG	*5*	10	13	10	16	16	*0.71*	0.93		**1.16**		**1.21**		**1.49**		**1.56**		
Gly	GCC	GGC	GGT	*17*	22	20	22	19	25	***3.58***	**2.15**		**2.11**		**2.26**		**2.05**		**2.04**		
Gly	CCC	GGG		*1*	9	9	5	3	7	*0.21*	0.88		0.95		0.51		0.32		0.57		
Gly	TCC	GGA	GGG	*1*	10	9	12	15	17	*0.21*	0.98		0.95		**1.23**		**1.62**		**1.39**		
Pro	AGG	CCT	CCC - (CCA)	*13*	14	13	16	7	14	***2.60***	**1.87**		**1.73**		0.98		**1.87**		**1.51**		
Pro	CGG	CCG		*6*	9	8	5	1	4	***1.20***	**1.20**		**1.07**		0.31		0.27		0.43		
Pro	TGG	CCA	CCG	*1*	7	9	44	7	19	*0.20*	0.93		**1.20**		**2.71**		**1.87**		**2.05**		
Thr	AGT	ACT	ACC - (ACA)	*6*	11	8	10	10	6	***2.18***	**2**		**1.60**		**1.74**		**2.11**		**1.20**		
Thr	CGT	ACG		*3*	5	6	5	1	4	***1.09***	0.87		**1.20**		0.87		0.21		0.80		
Thr	GGT	ACC		*0*	1	0	0	0	0	*0*	0.17		0		0		0		0		
Thr	TGT	ACA	ACG	*2*	6	6	8	8	10	*0.73*	**1.04**		**1.20**		**1.39**		**1.68**		**2**		
Val	AAC	GTT	GTC - (GTA)	*7*	17	13	15	17	17	***1.56***	**2.19**		**1.79**		**2**		**2.27**		**1.94**		
Val	CAC	GTG		*10*	10	11	8	7	10	***2.22***	**1.29**		**1.52**		**1.07**		0.93		**1.14**		
Val	TAC	GTA	GTG	*1*	4	5	7	6	8	*0.22*	0.52		0.69		0.93		0.80		0.91		
Arg	ACG	CGT	CGC - (CGA)	*11*	12	11	9	9	9	***3.67***	**1.71**		**1.69**		**1.50**		**1.86**		**1.26**		
Arg	CCG	CGG		*3*	8	6	4	2	5	***1.00***	**1.14**		0.92		0.67		0.41		0.70		
Arg	CCT	AGG		*2*	11	9	8	6	11	*0.67*	**1.57**		**1.38**		**1.33**		**1.24**		**1.53**		
Arg	TCG	CGA	CGG	*1*	4	5	6	5	7	*0.33*	0.57		0.77		1		**1.03**		0.98		
Arg	TCT	AGA	AGG	*1*	7	8	9	7	11	*0.33*	1		**1.23**		**1.50**		**1.45**		**1.53**		
Leu	AAG	CTT	CTC - (CTA)	*3*	15	14	11	8	14	***1.06***	**2.20**		**2.27**		**1.69**		**1.45**		**3**		
Leu	CAG	CTG		*10*	8	8	3	4	6	***3.53***	**1.17**		**1.30**		0.46		0.73		**1.29**		
Leu	CAA	TTG		*2*	8	6	10	9	1	*0.71*	**1.17**		0.97		**1.54**		**1.64**		0.21		
Leu	TAG	CTA	CTG	*1*	6	5	9	7	0	*0.35*	0.88		0.81		**1.38**		**1.27**		0		
Leu	TAA	TTA	TTG	*1*	4	4	6	5	7	*0.35*	0.59		0.65		0.92		0.91		**1.50**		
Ser	GCT	AGC	AGT	*8*	10	9	11	13	13	***2.53***	**1.88**		**1.86**		**1.12**		**2.29**		**1.95**		
Ser	AGA	TCT	TCC - (TCA)	*5*	10	10	37	12	13	***1.58***	**1.88**		**2.07**		**3.76**		**2.12**		**1.95**		
Ser	CGA	TCG		*5*	6	5	4	2	5	***1.58***	**1.13**		**1.03**		0.41		0.35		0.75		
Ser	TGA	TCA	TCG	*1*	6	5	7	7	9	*0.32*	**1.13**		**1.03**		0.71		**1.24**		**1.35**		
SeC	TCA			*1*	0	0	0	0	0												

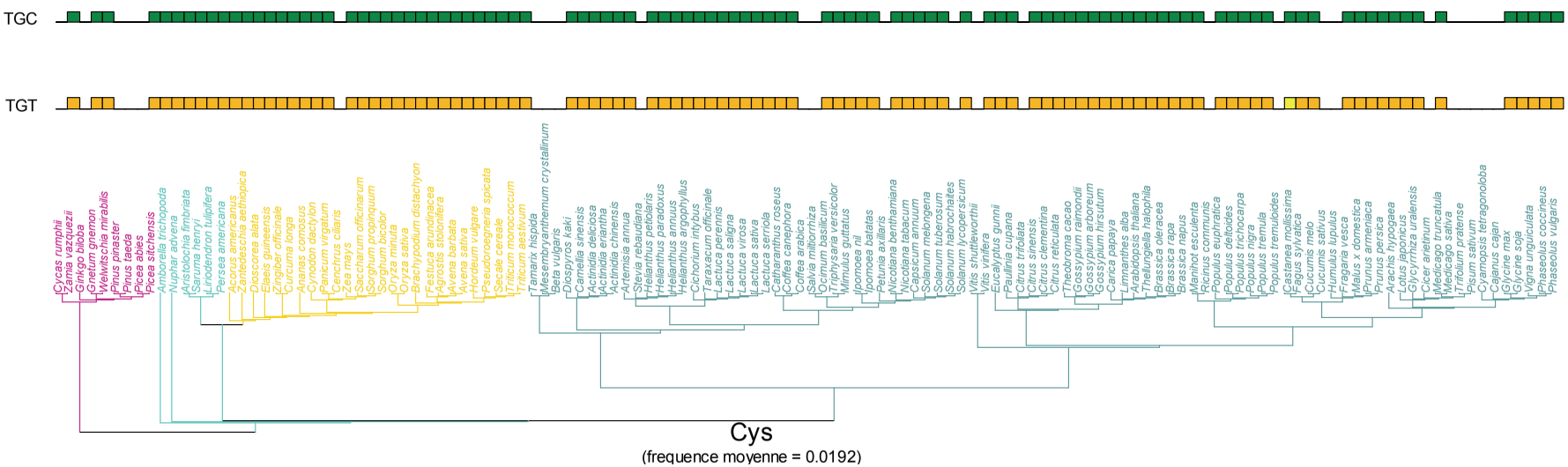
Synopsis des préférences d'usage du code de 130 espèces de plantes

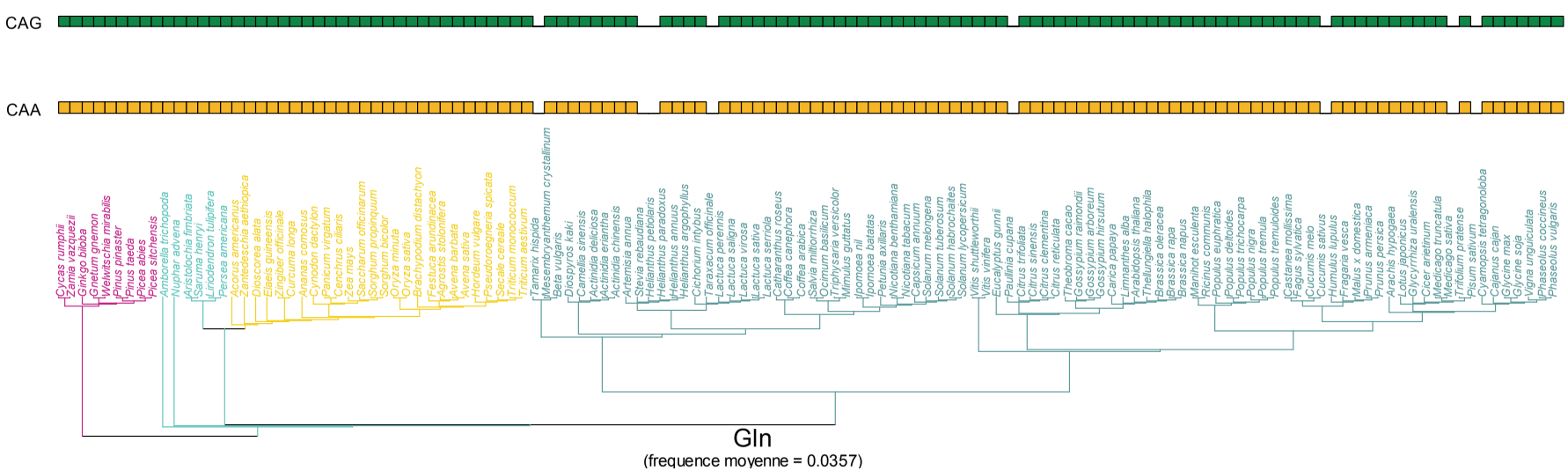


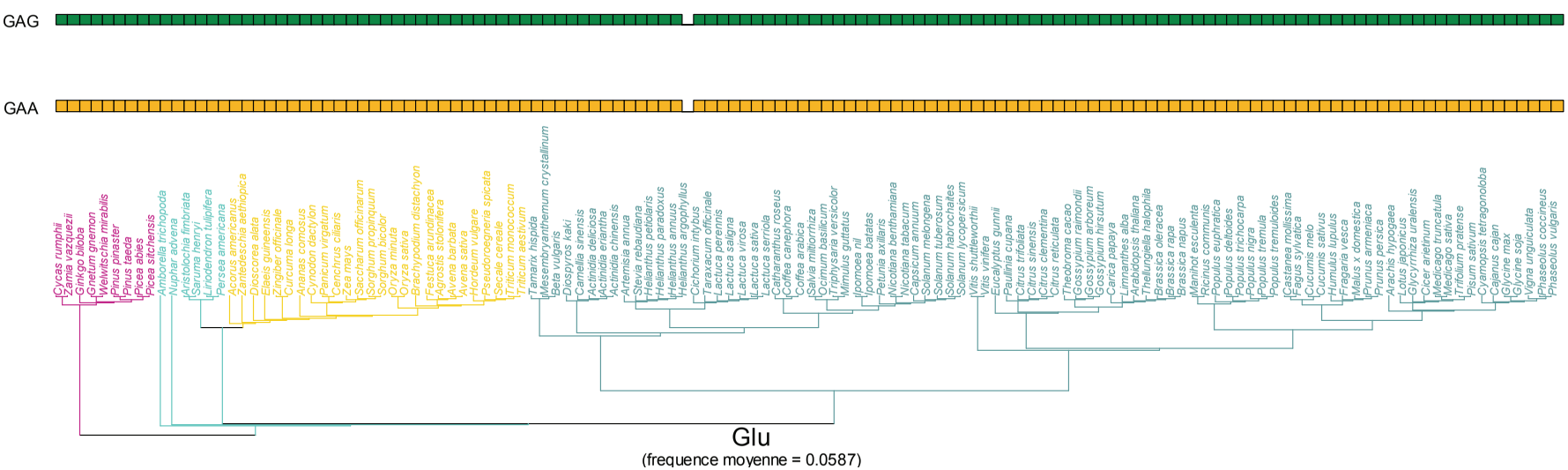
Arbres de préférences d'usage des codons pour 130 espèces de plantes

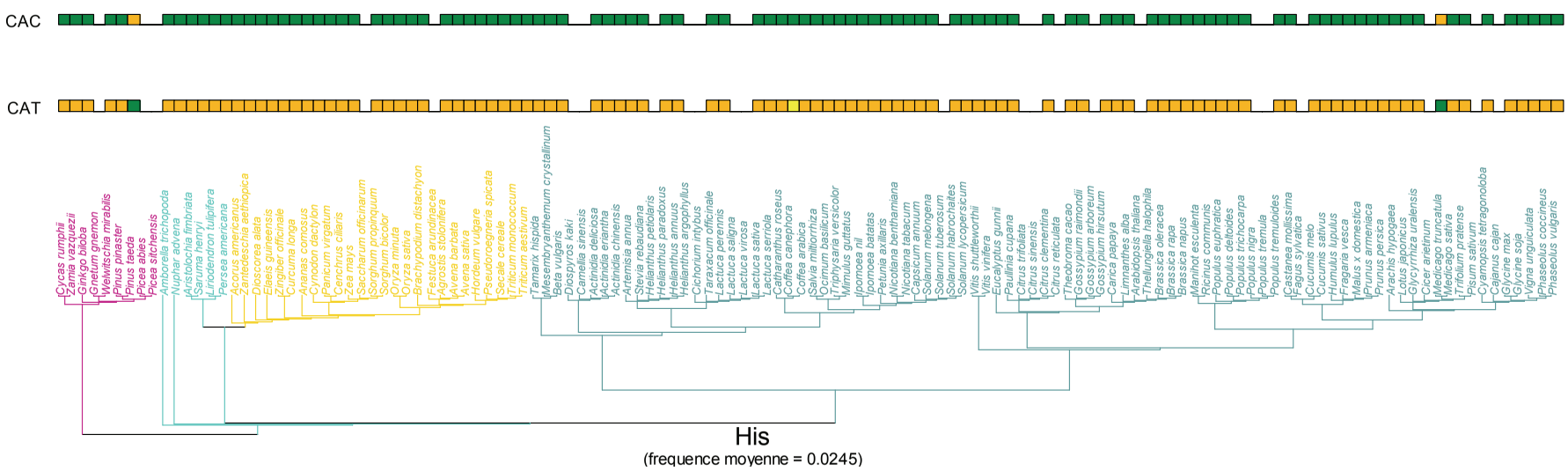


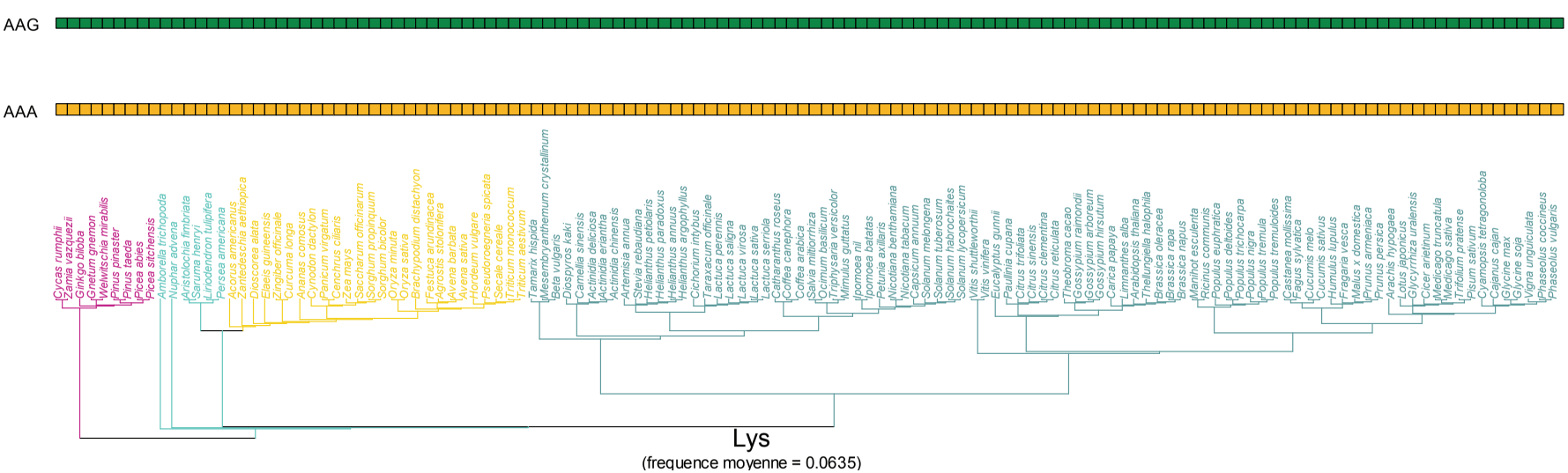

GAC
GAT
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Dendrocalamus latiflorus
Panicum virgatum
Lemchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Asp
(frequence moyenne = 0.0488)

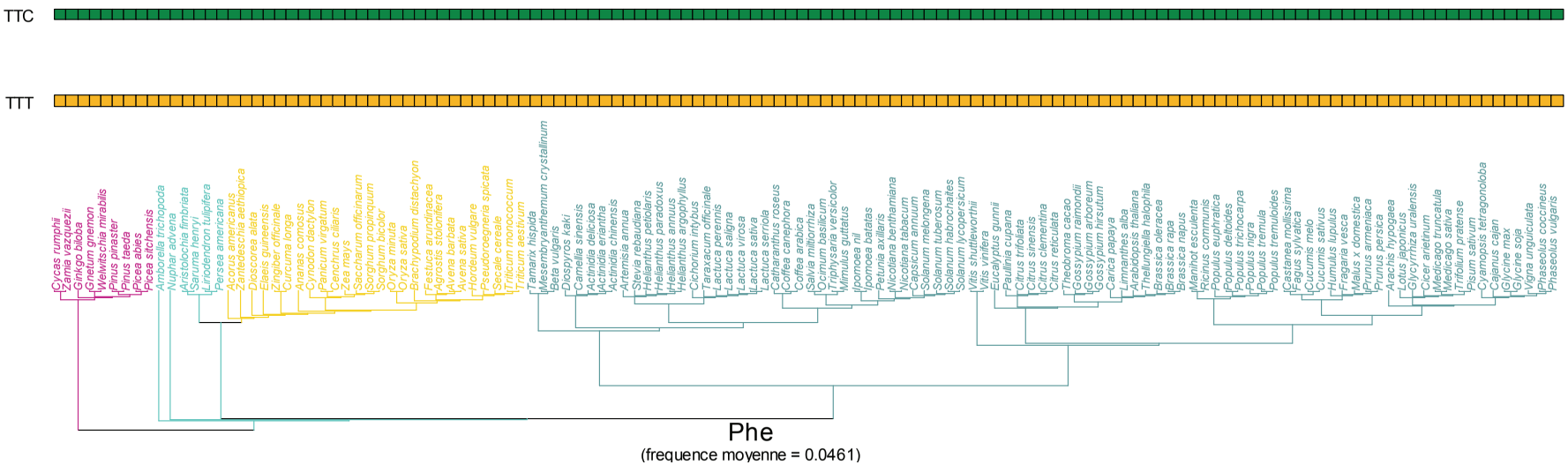

TGC
TGT
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Dendrocalamus latiflorus
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Cys
(frequence moyenne = 0.0192)

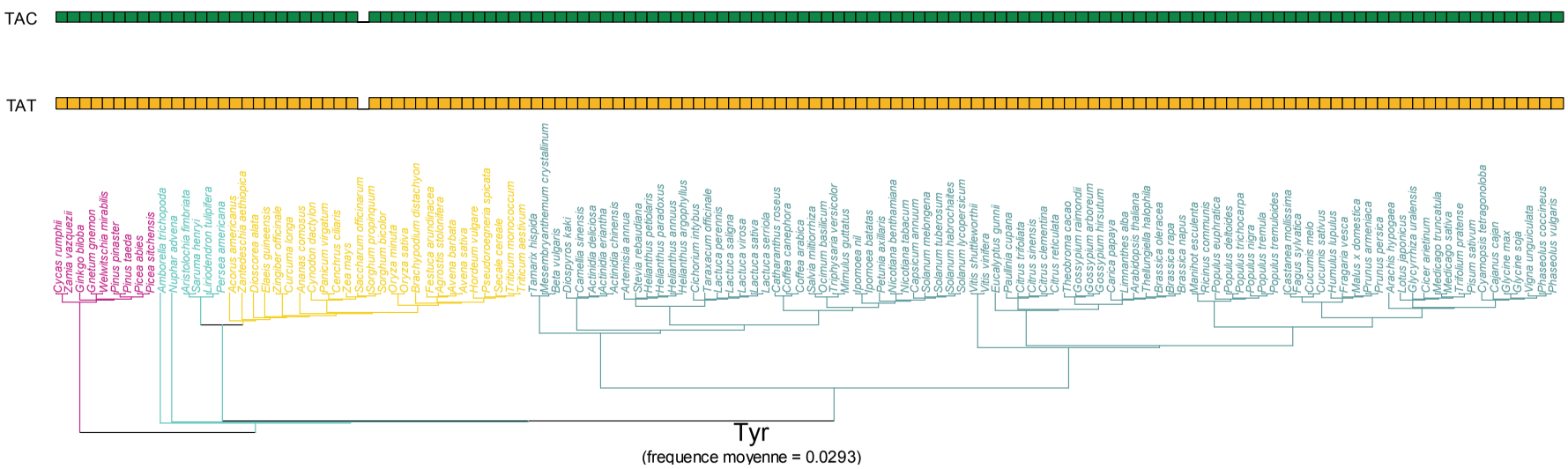

CAG
CAA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Gln
(frequence moyenne = 0.0357)

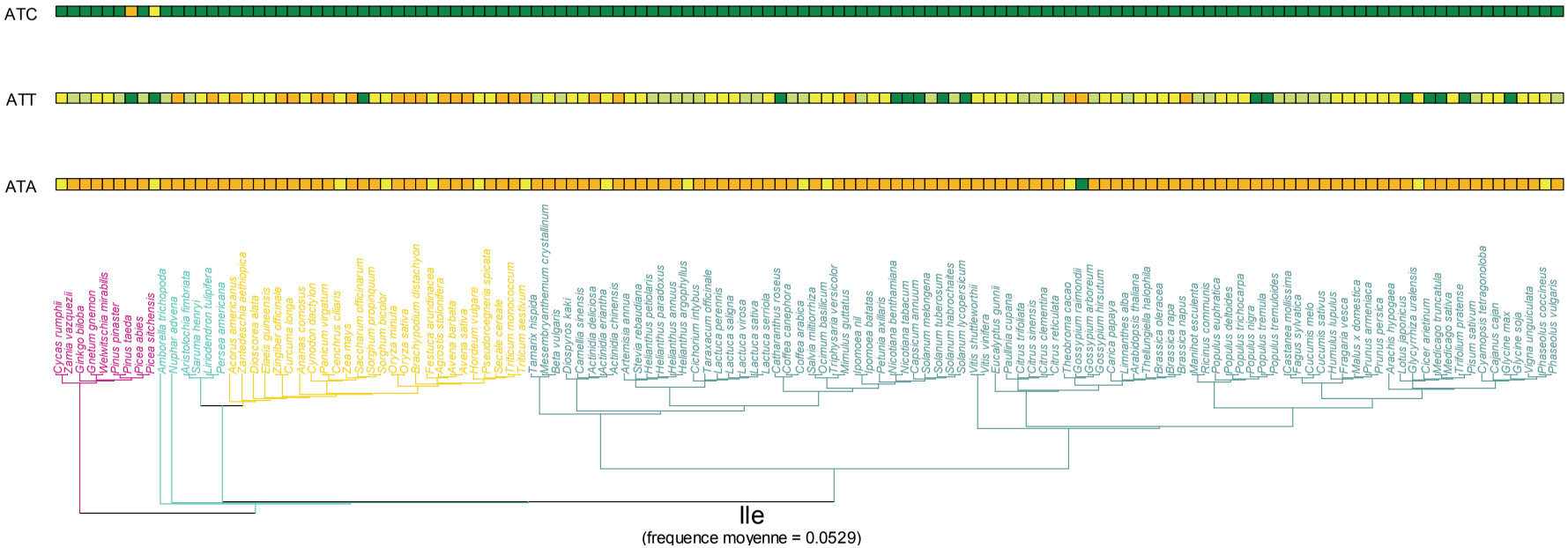

GAG
GAA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Glycyrrhiza uralensis
Lotus japonicus
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Glu
(frequence moyenne = 0.0587)

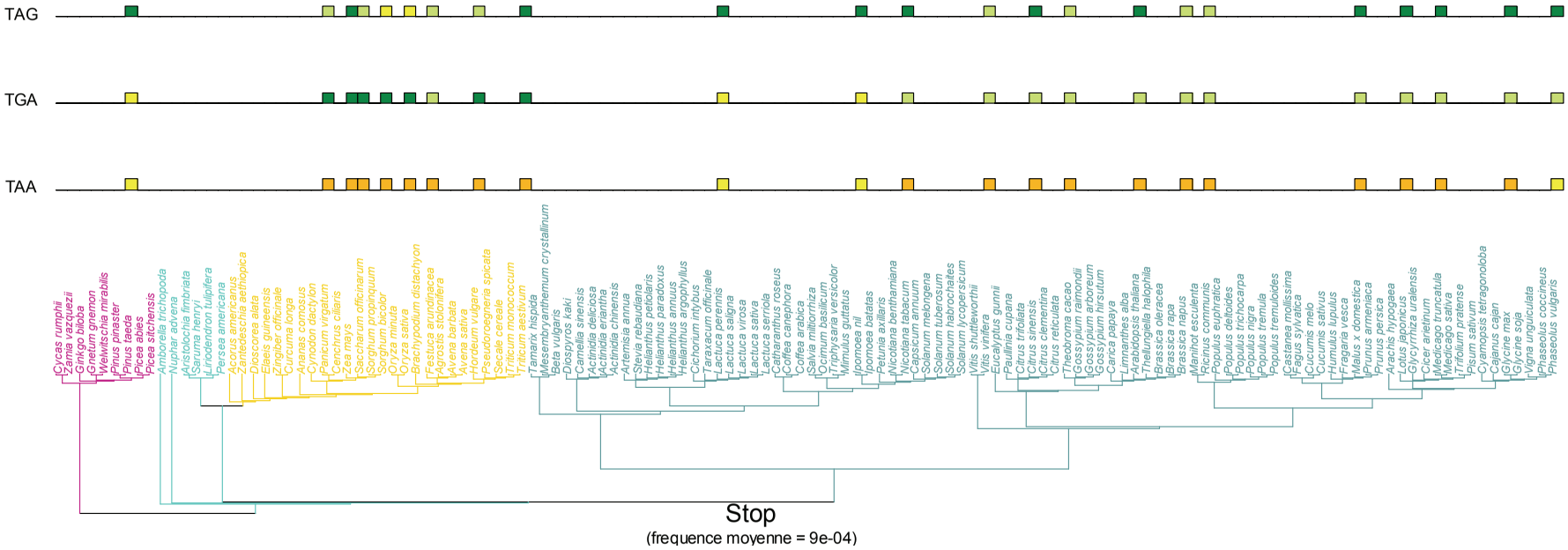

CAC
CAT
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
His
(frequence moyenne = 0.0245)

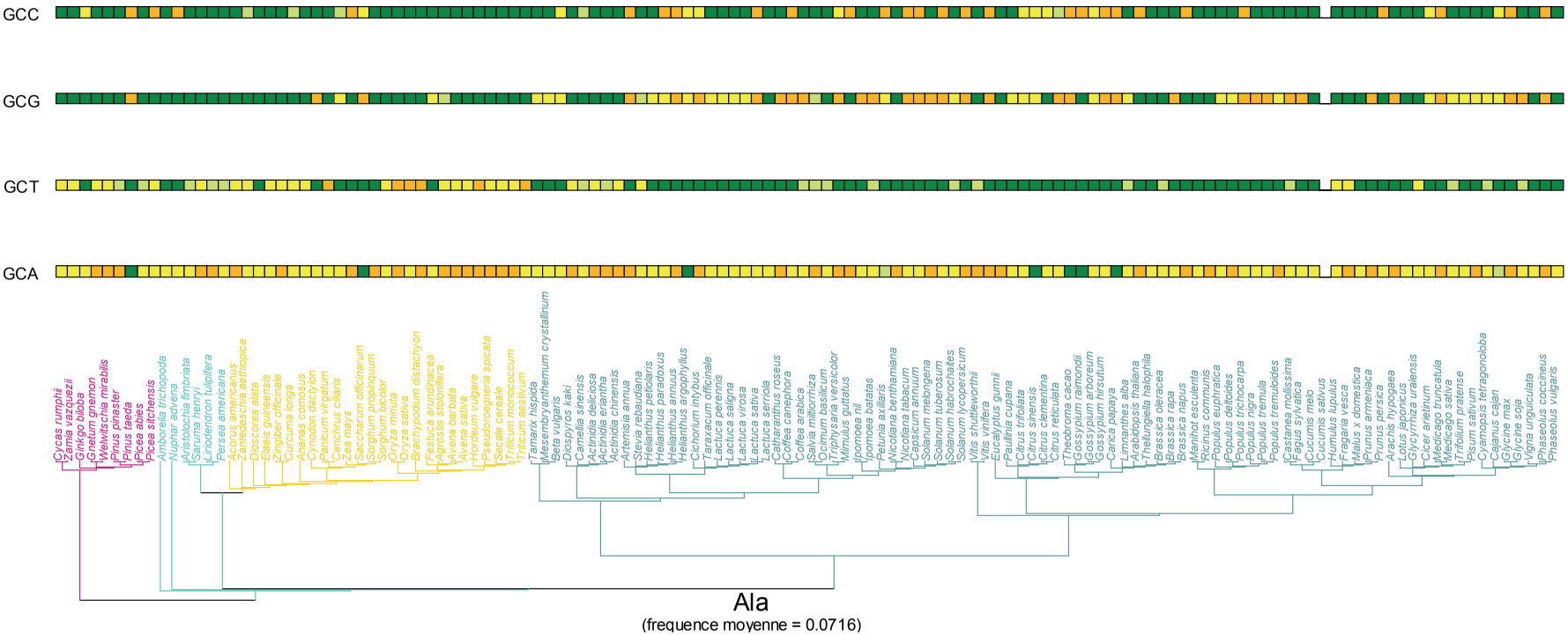

AAG
AAA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Lys
(frequence moyenne = 0.0635)

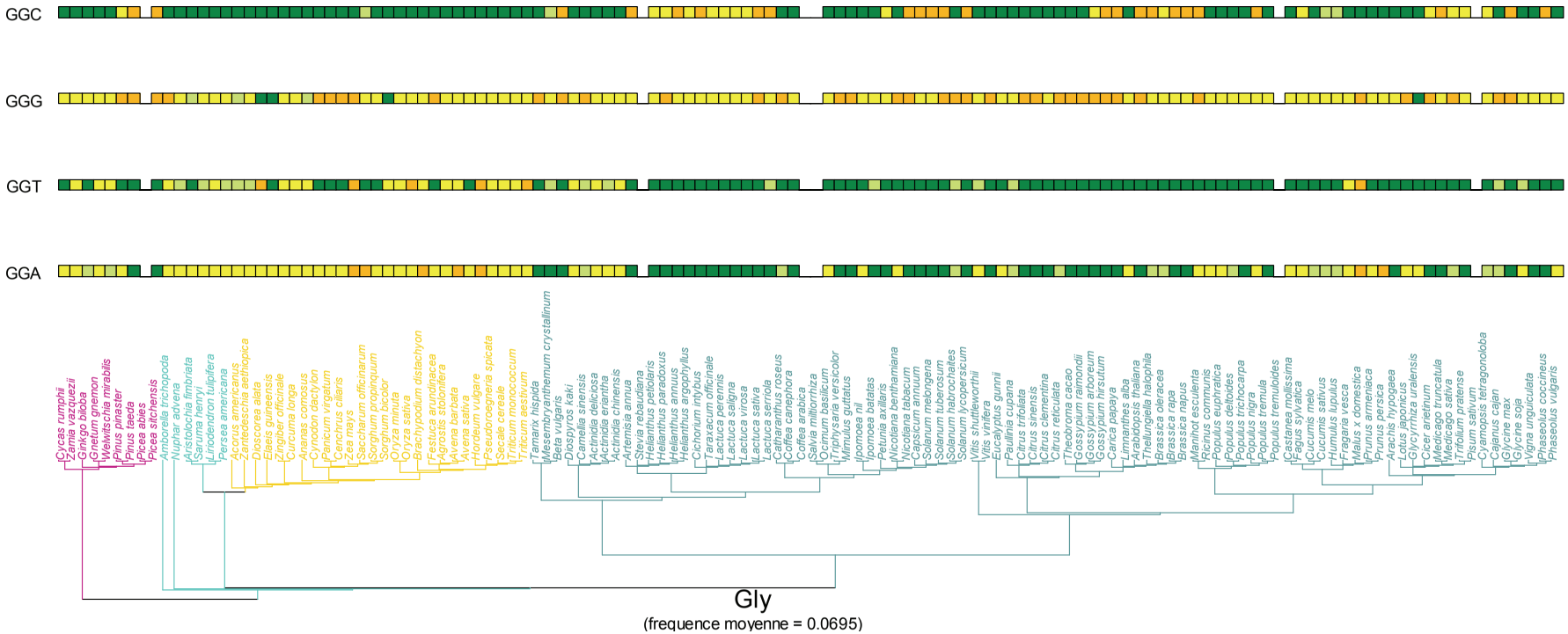

TTC
TTT
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Phe
(frequence moyenne = 0.0461)

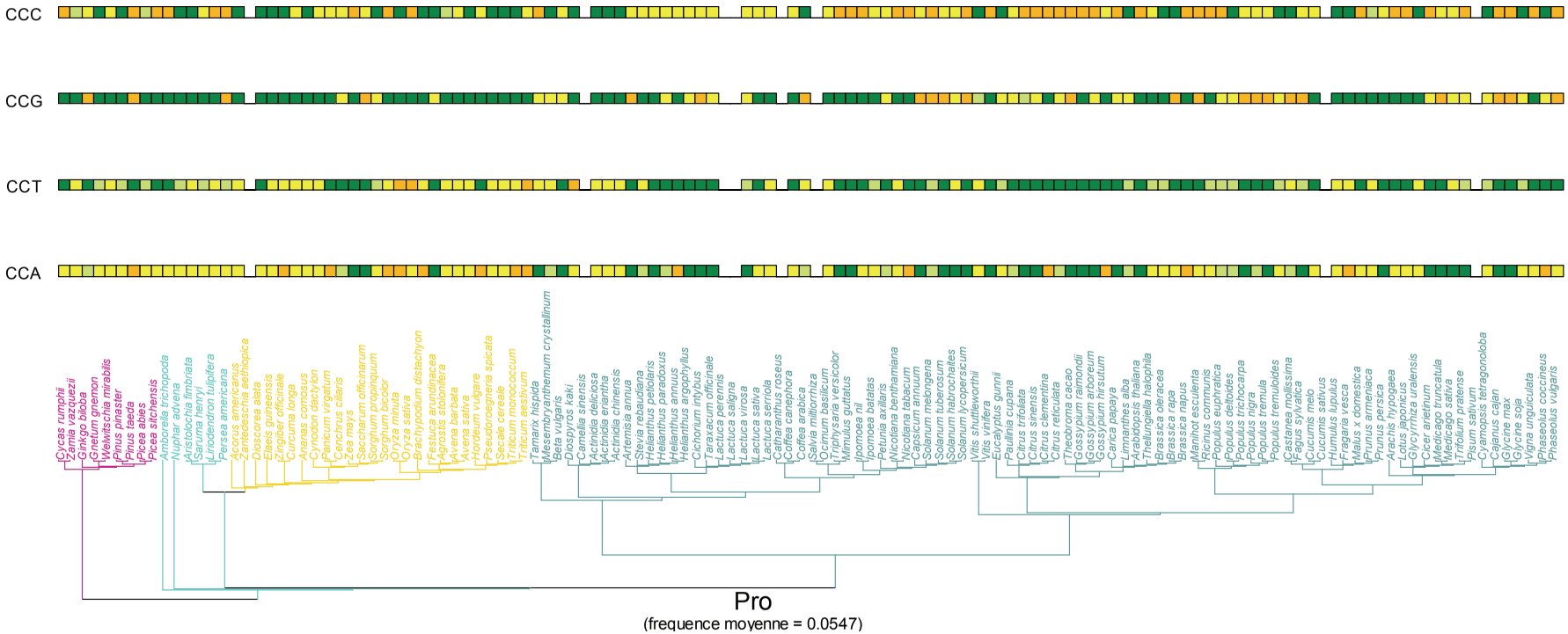

TAC
TAT
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Glycyrrhiza uralensis
Lotus japonicus
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Tyr
(frequence moyenne = 0.0293)

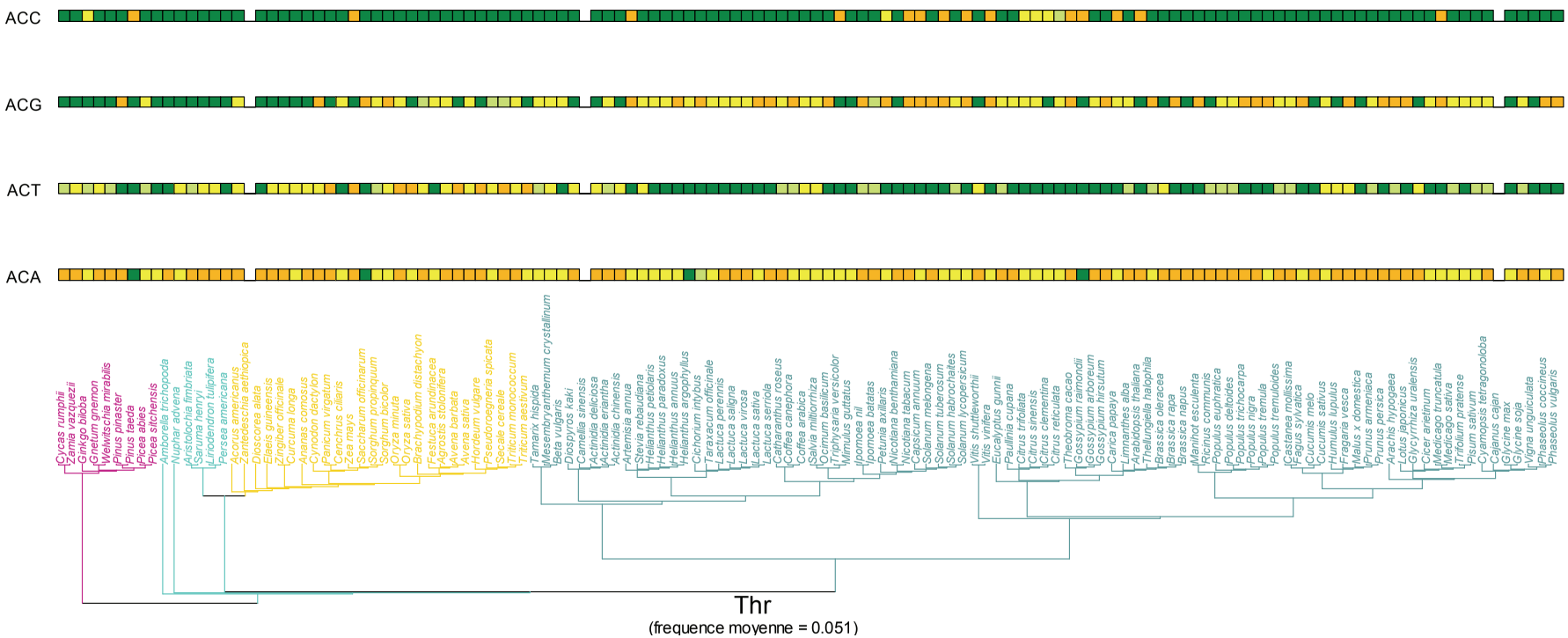

ATC
ATT
ATA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus jonaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Ocimum basilicum
Salvia miltiorrhiza
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Ile
(frequence moyenne = 0.0529)

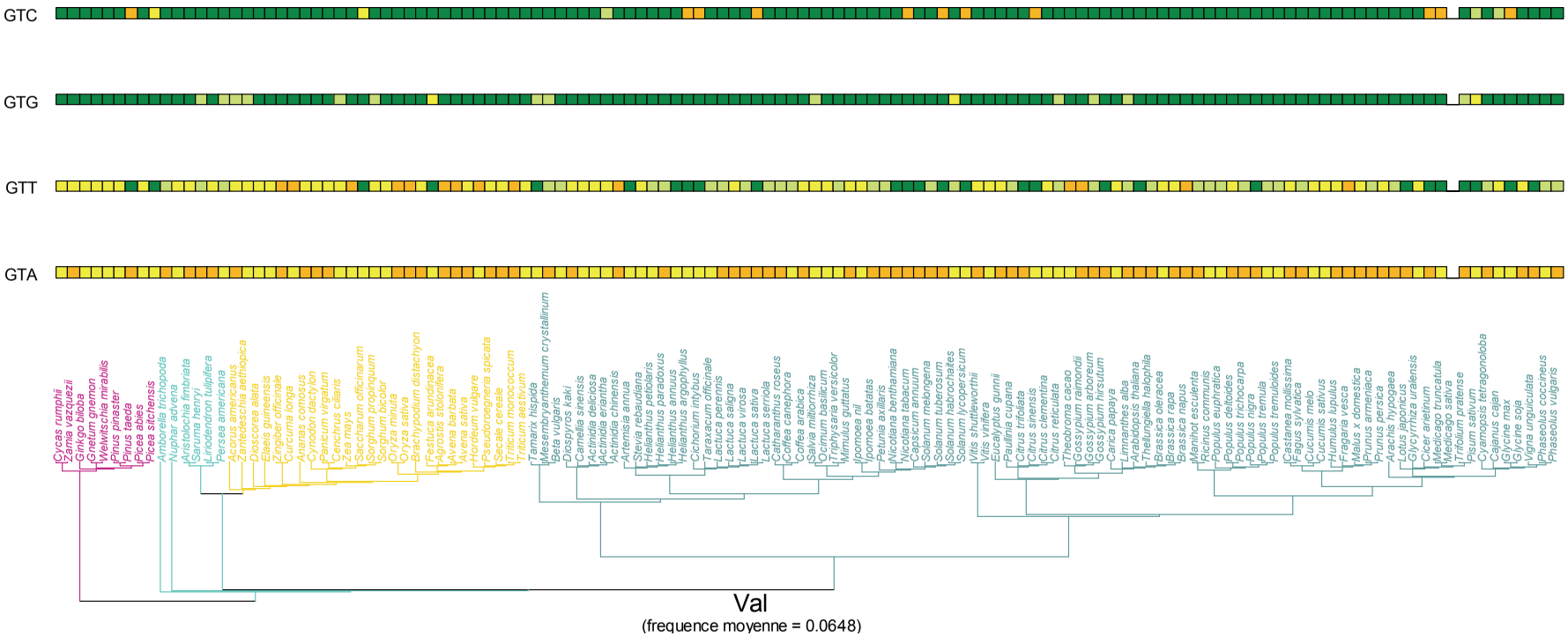

TAG
TGA
TAA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Lanchnus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Tripnysaira versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Stop
(frequence moyenne = 9e-04)

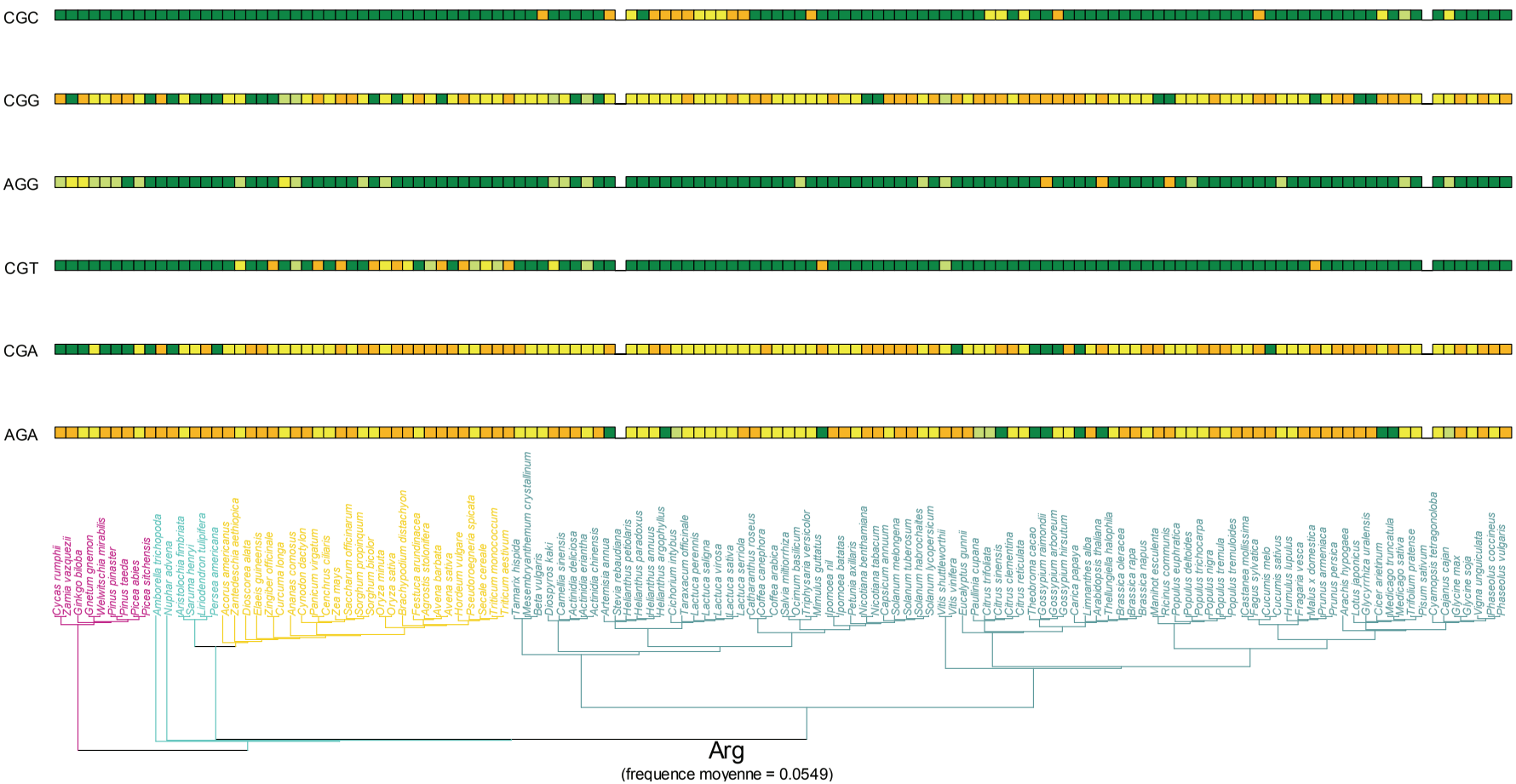

GCC
GCG
GCT
GCA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Lanchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus annuus
Helianthus paradoxus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Tripnysaira versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Glycyrrhiza uralensis
Lotus japonicus
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Ala
(frequence moyenne = 0.0716)

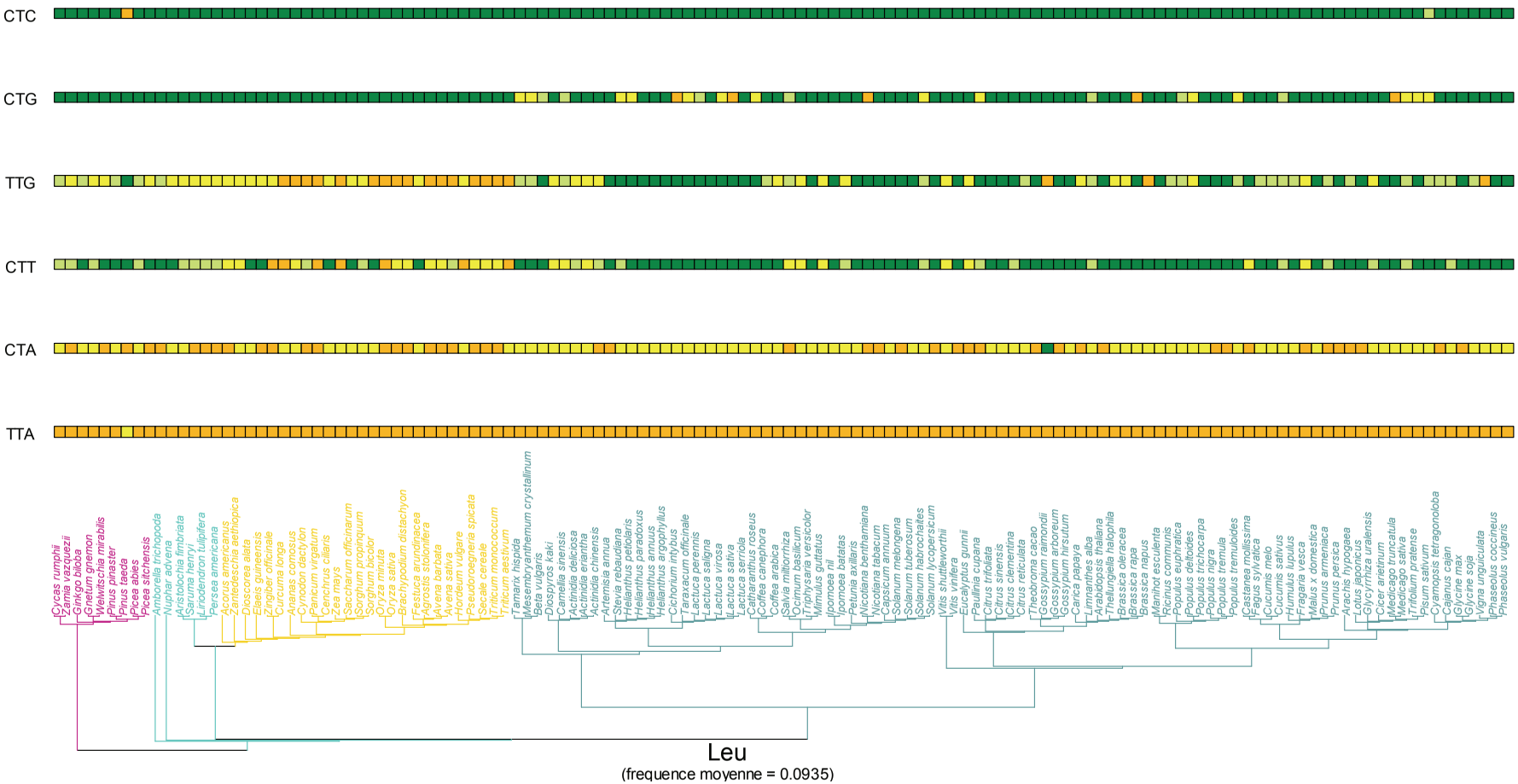

GGC
GGG
GGT
GGA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Gly
(frequence moyenne = 0.0695)

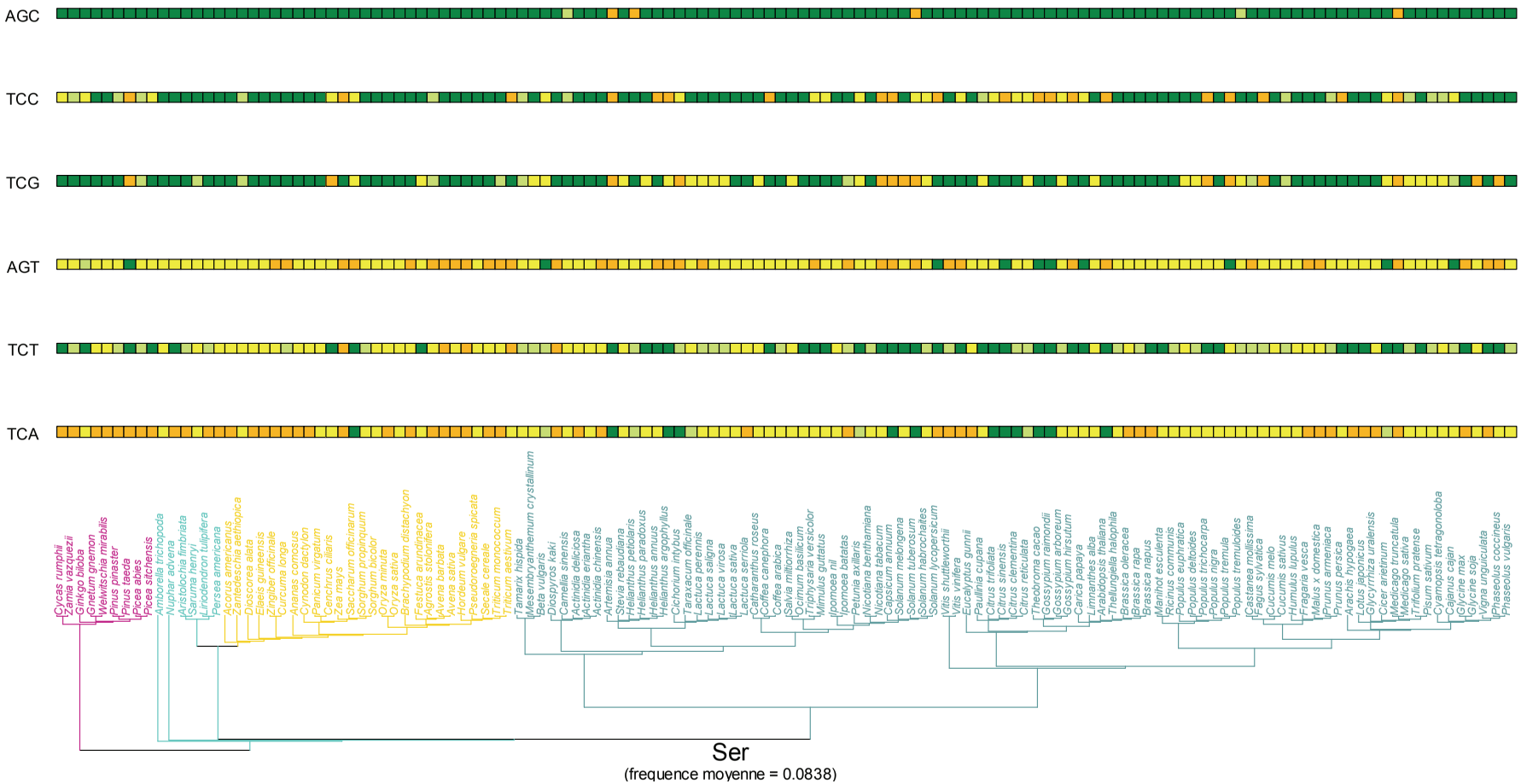

CCC
CCG
CCT
CCA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Glycyrrhiza uralensis
Lotus japonicus
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Pro
(frequence moyenne = 0.0547)


ACC
ACG
ACT
ACA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Pinus abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Glycyrrhiza uralensis
Lotus japonicus
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Thr
(frequence moyenne = 0.051)

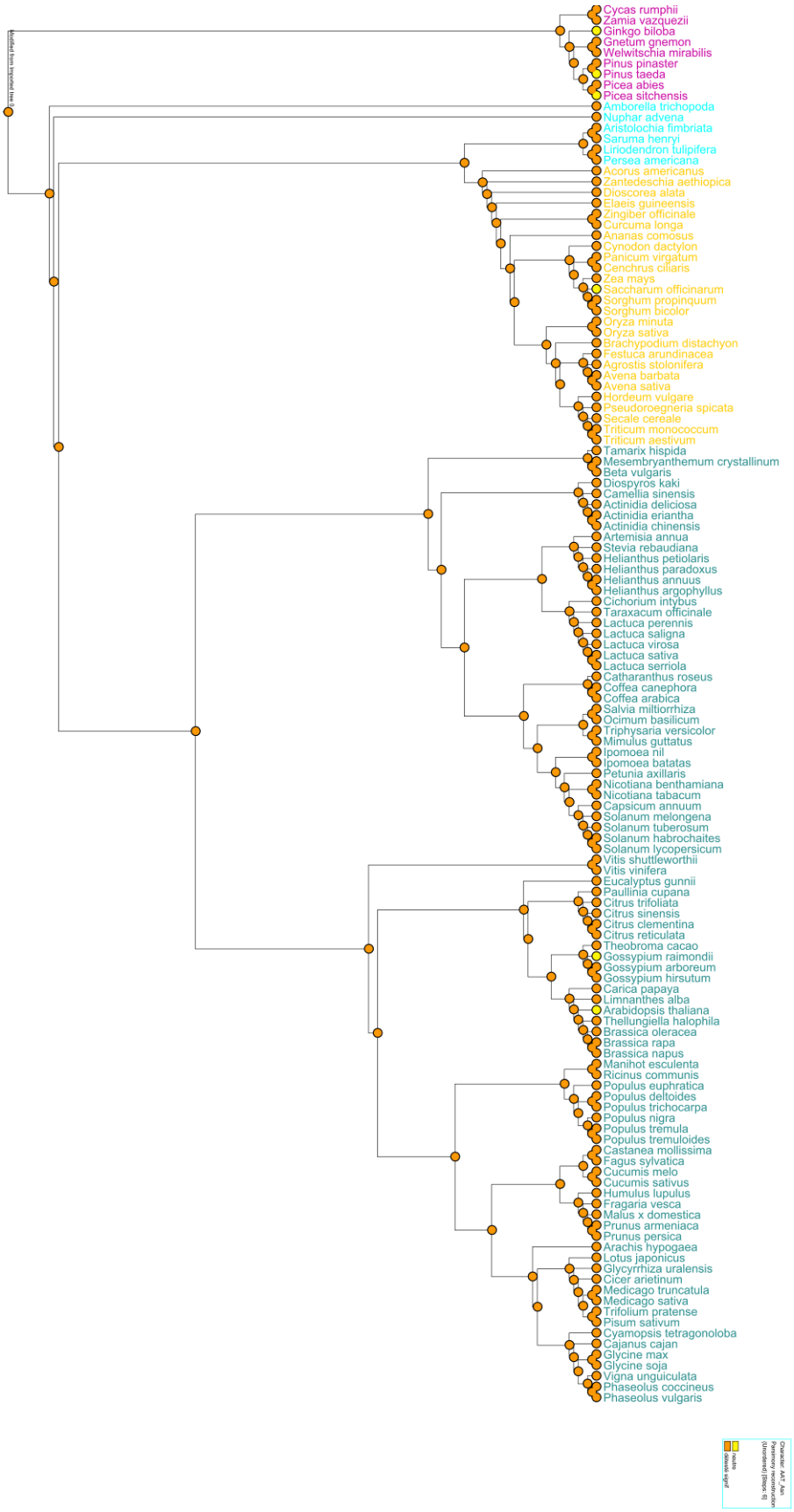

GTC
GTG
GTT
GTA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Pinus abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Glycyrrhiza uralensis
Lotus japonicus
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Val
(frequence moyenne = 0.0648)


CGC
CGG
AGG
CGT
CGA
AGA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus banksiana
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Commelina diffusa
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Arg
(frequence moyenne = 0.0549)

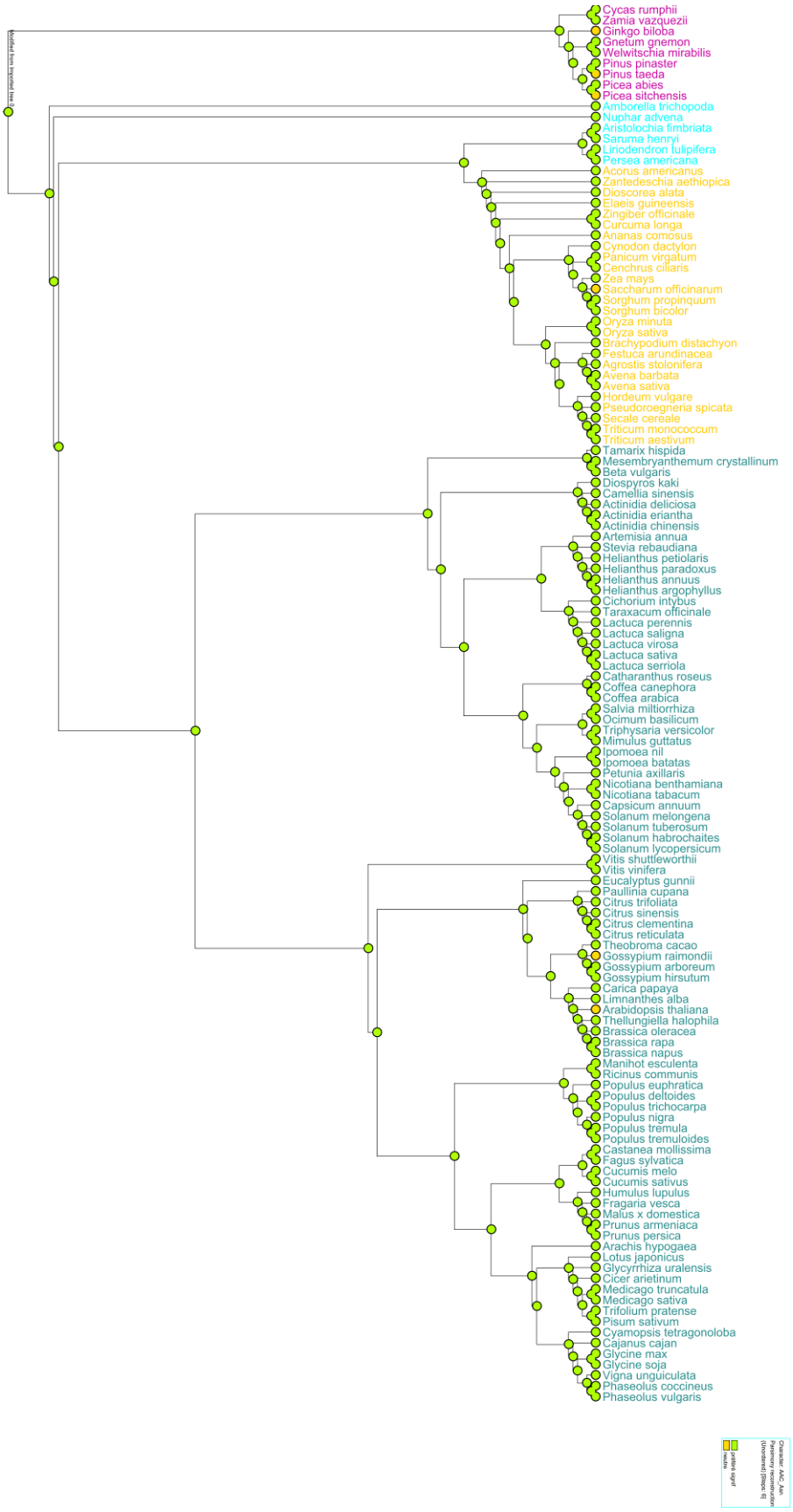

CTC
CTG
TTG
CTT
CTA
TTA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus taeda
Pinus abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Lanchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca officinale
Lactuca saligna
Lactuca sativa
Lactuca virosa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Leu
(frequence moyenne = 0.0935)


AGC
TCC
TCG
AGT
TCT
TCA
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Phoenix dactylifera
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Lactuca perennis
Lactuca saligna
Lactuca sativa
Lactuca virosa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paulinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Quercus mongolica
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Glycyrrhiza uralensis
Lotus japonicus
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Ser
(frequence moyenne = 0.0838)

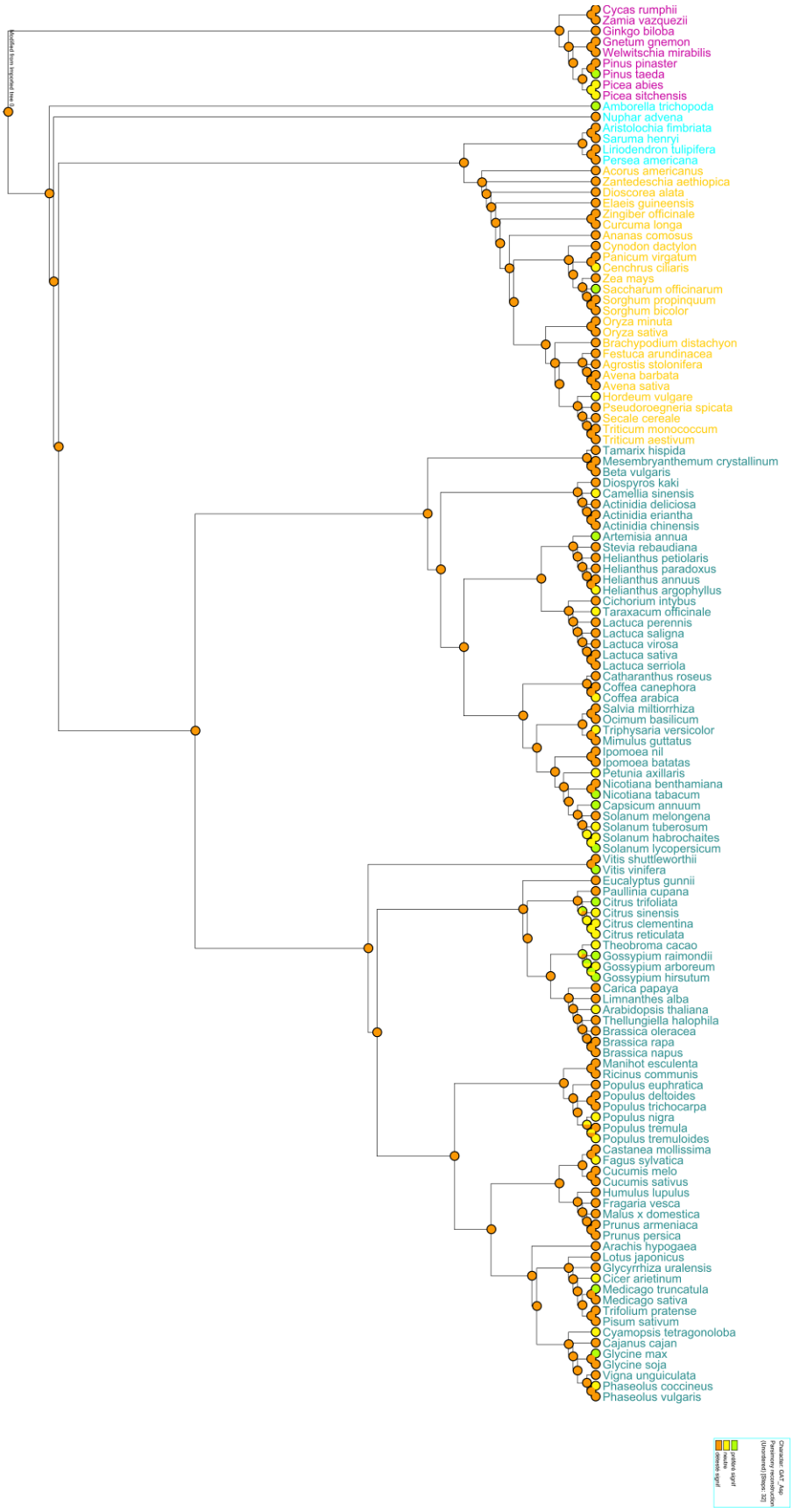
Asn AAT


Modified from imported tree (1
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: AAT_Asn
Parsimony reconstruction
(Unordered) [Steps: 6]
Inputs
Steps: 6
Unknown
Steps: 6

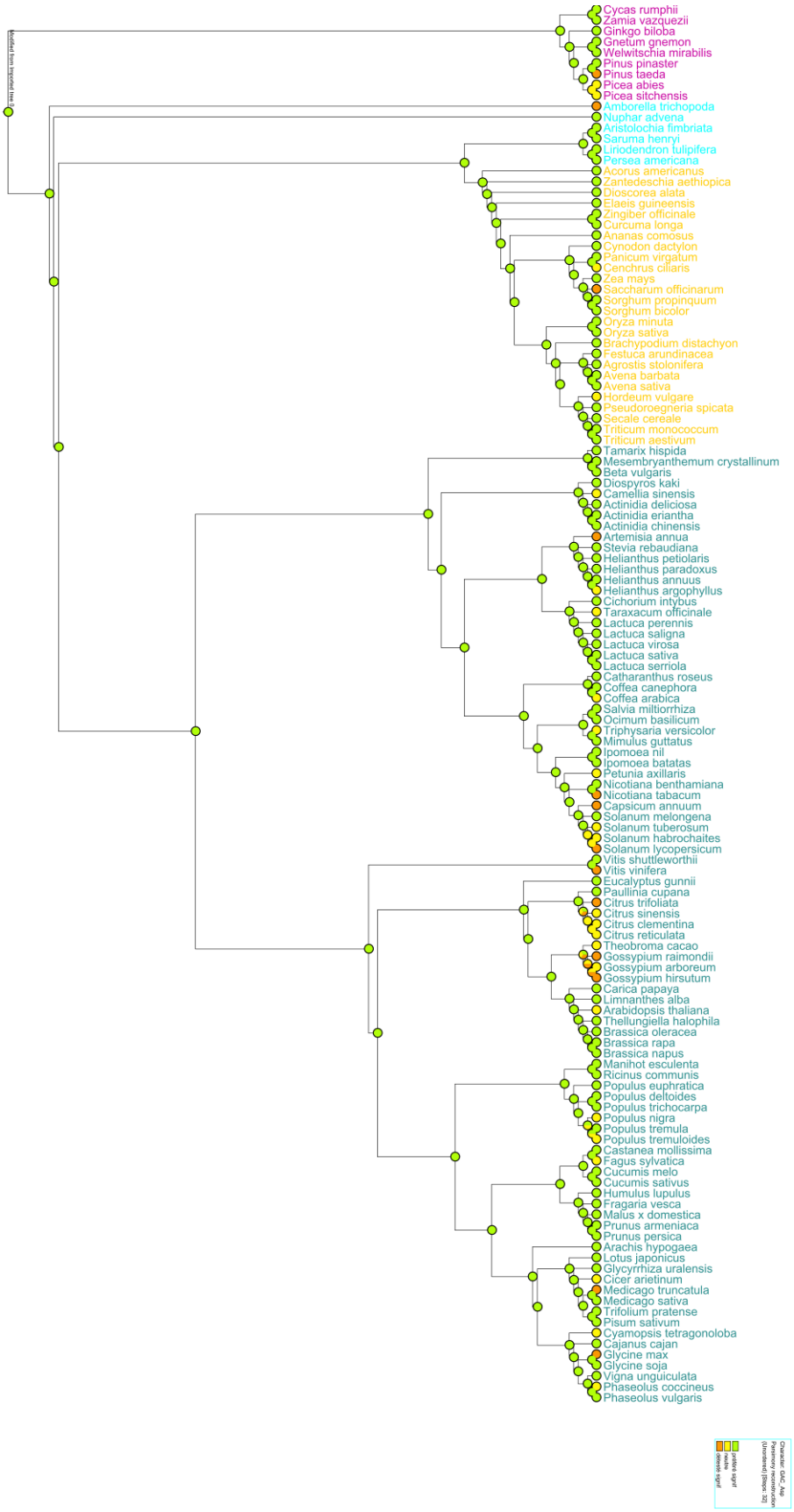
Asn AAC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: AAC_Asn
Parsimony reconstruction
(Unordered) [Steps: 5]
Ambiguous state
Active
Inactive

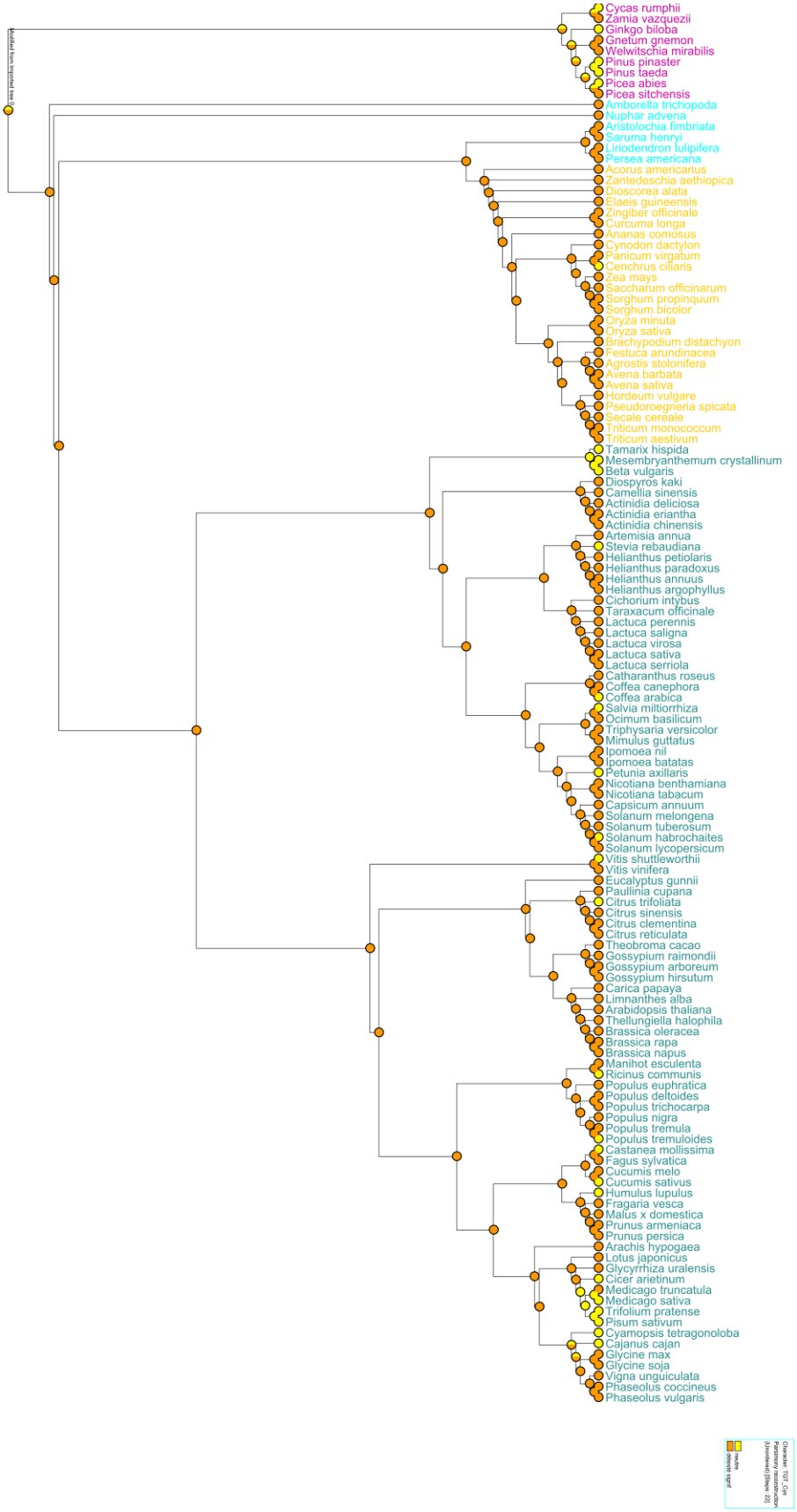
Asp GAT


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

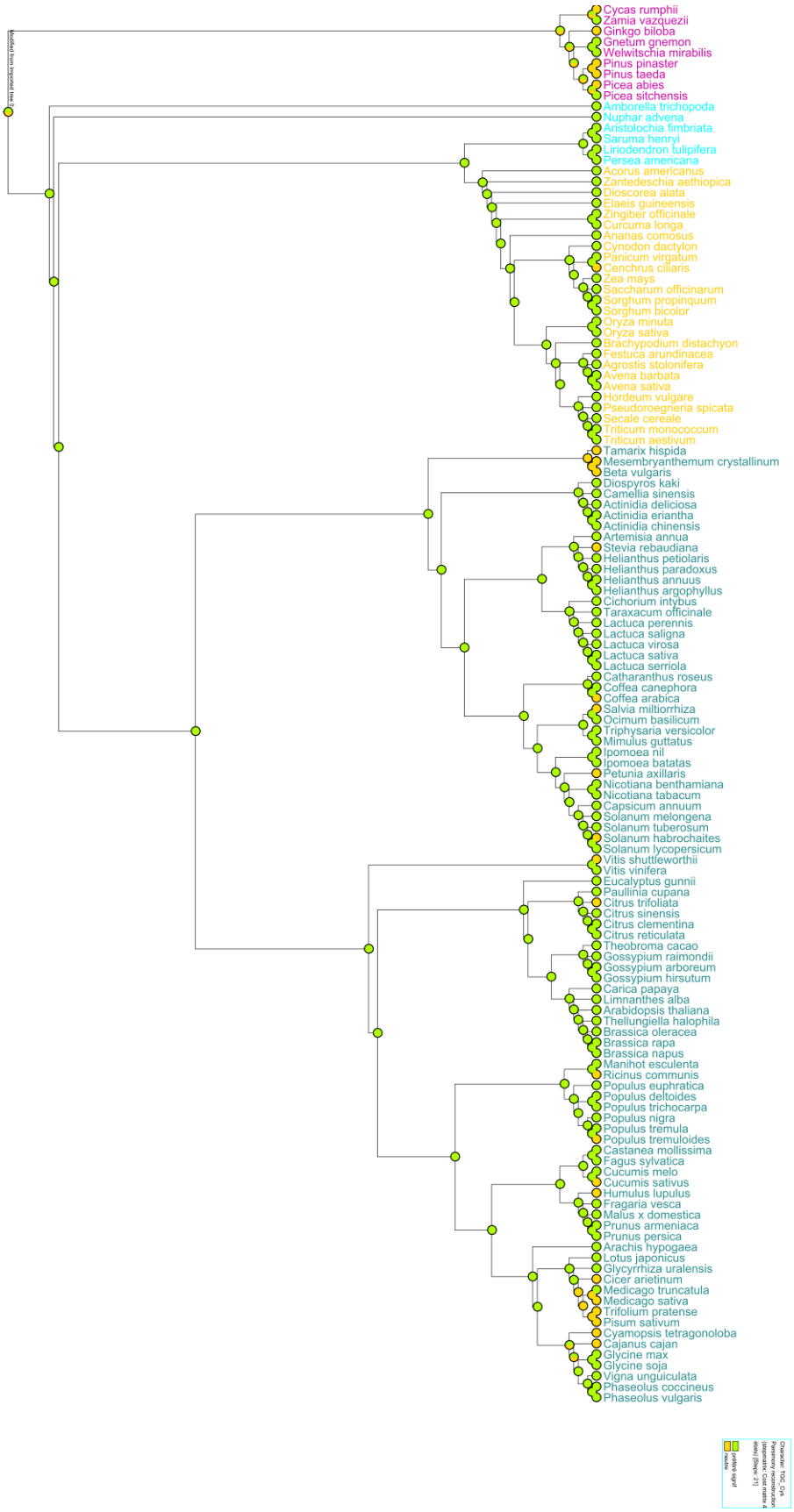
Asp GAC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GAC_Asp
Parsimony reconstruction
(Unordered) [Steps: 32]
Ambiguous signal
Auxin
Absence signal

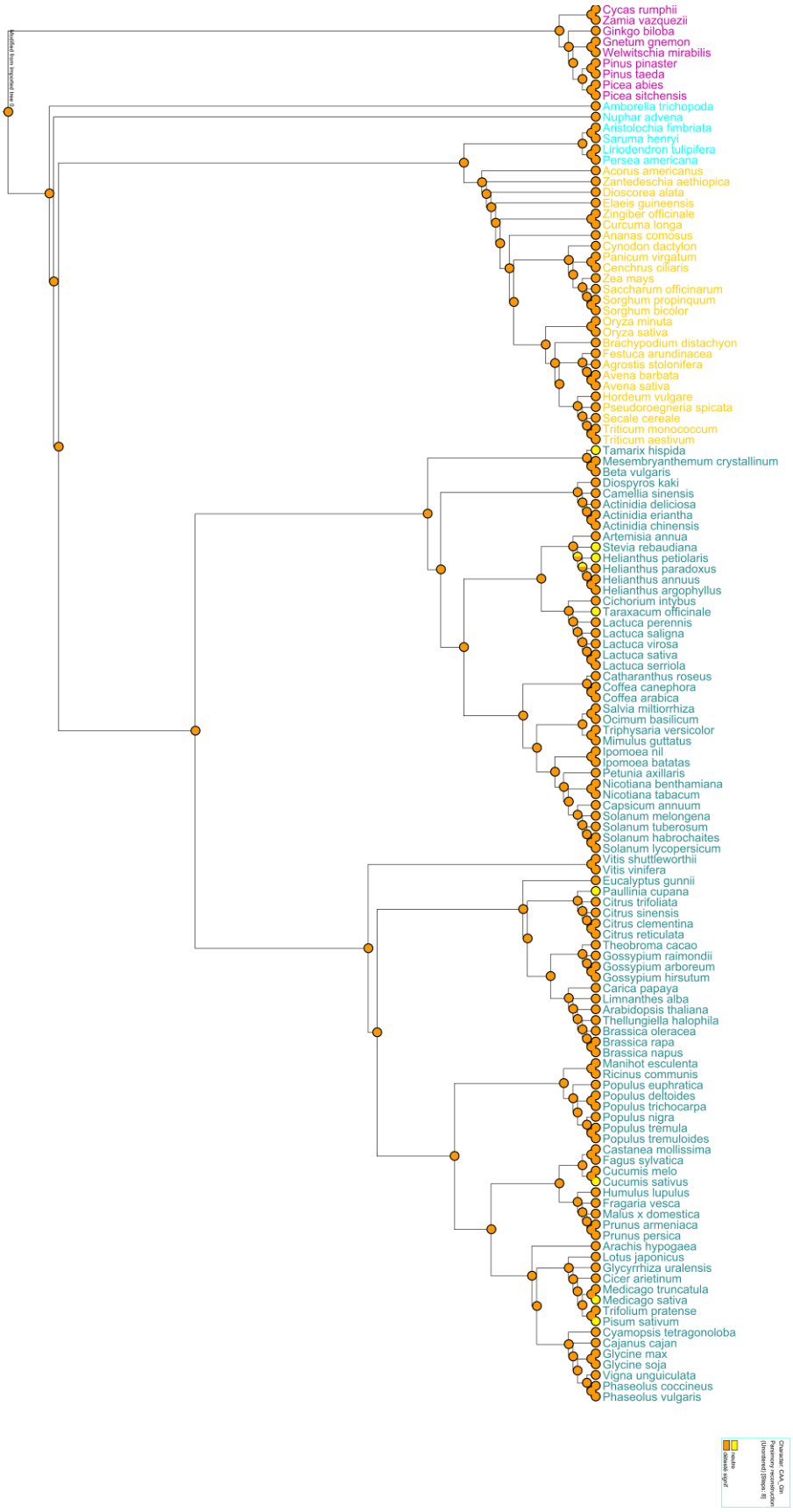
Cys TGT


Modified from Unrooted Tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TGT_Cys
Parsimony reconstruction
(Unordered) [Steps: 22]
Gains
Losses
Unchanged
Ambiguous

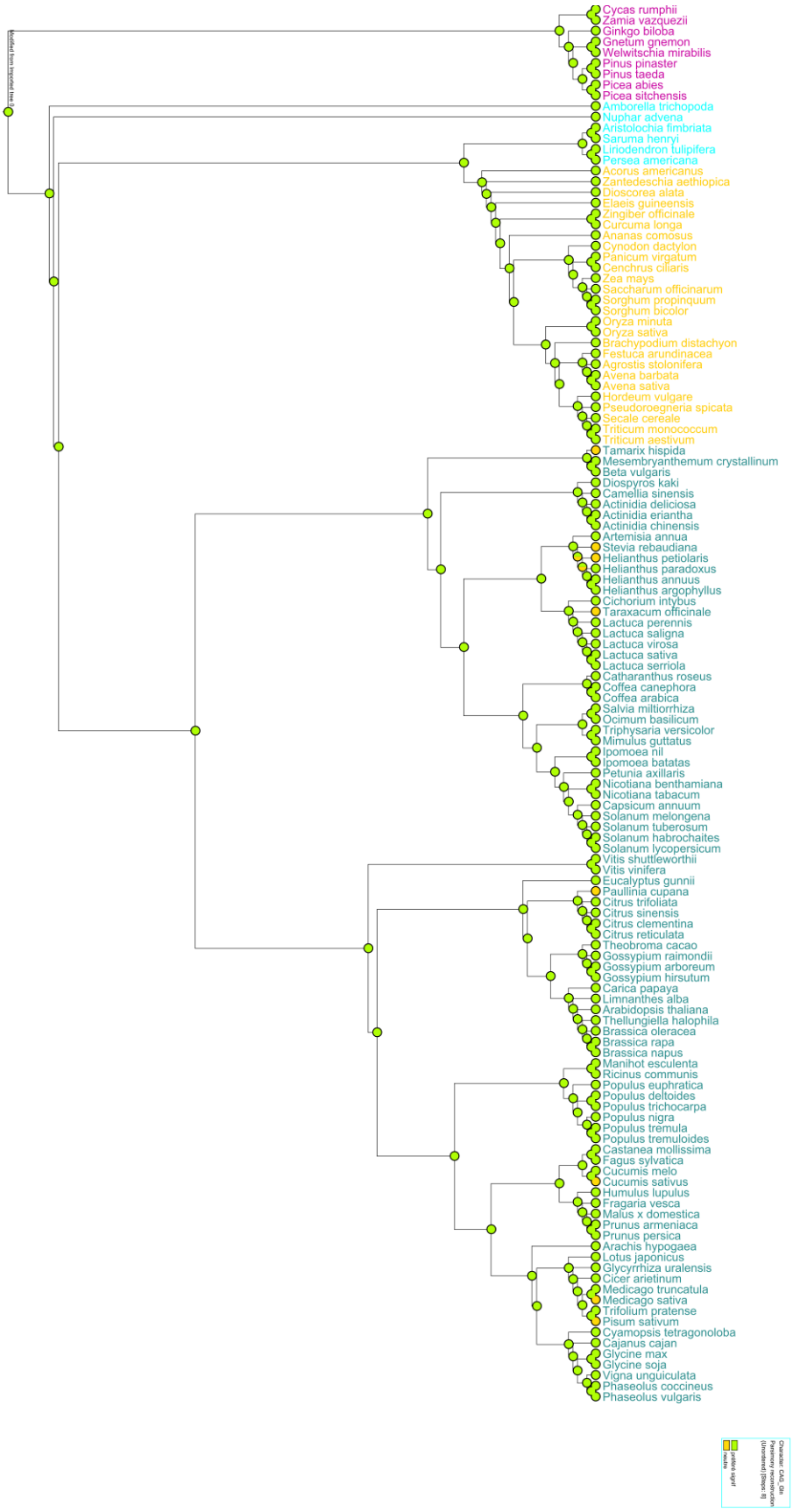
Cys TGC


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

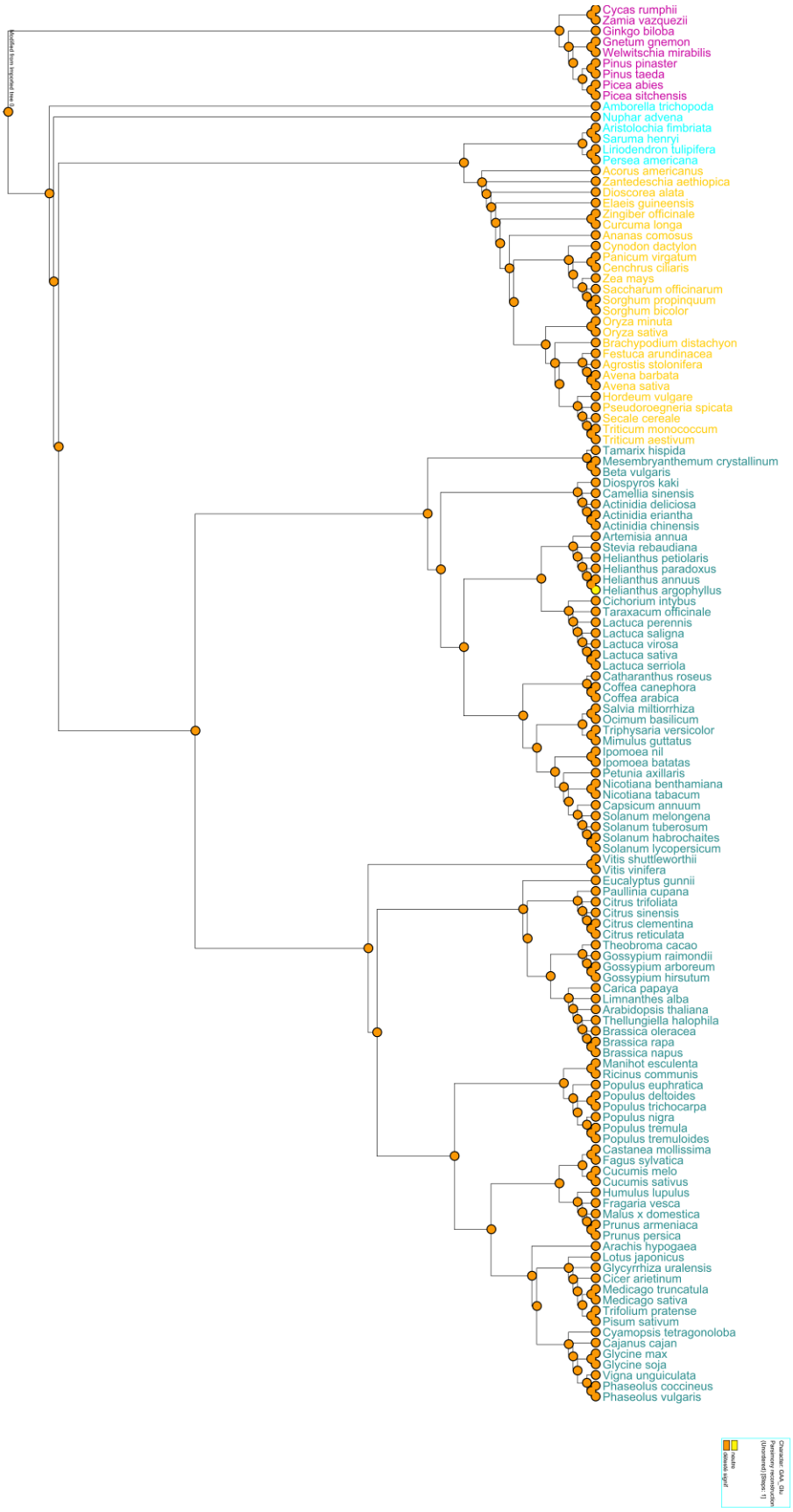
Gln CAA


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

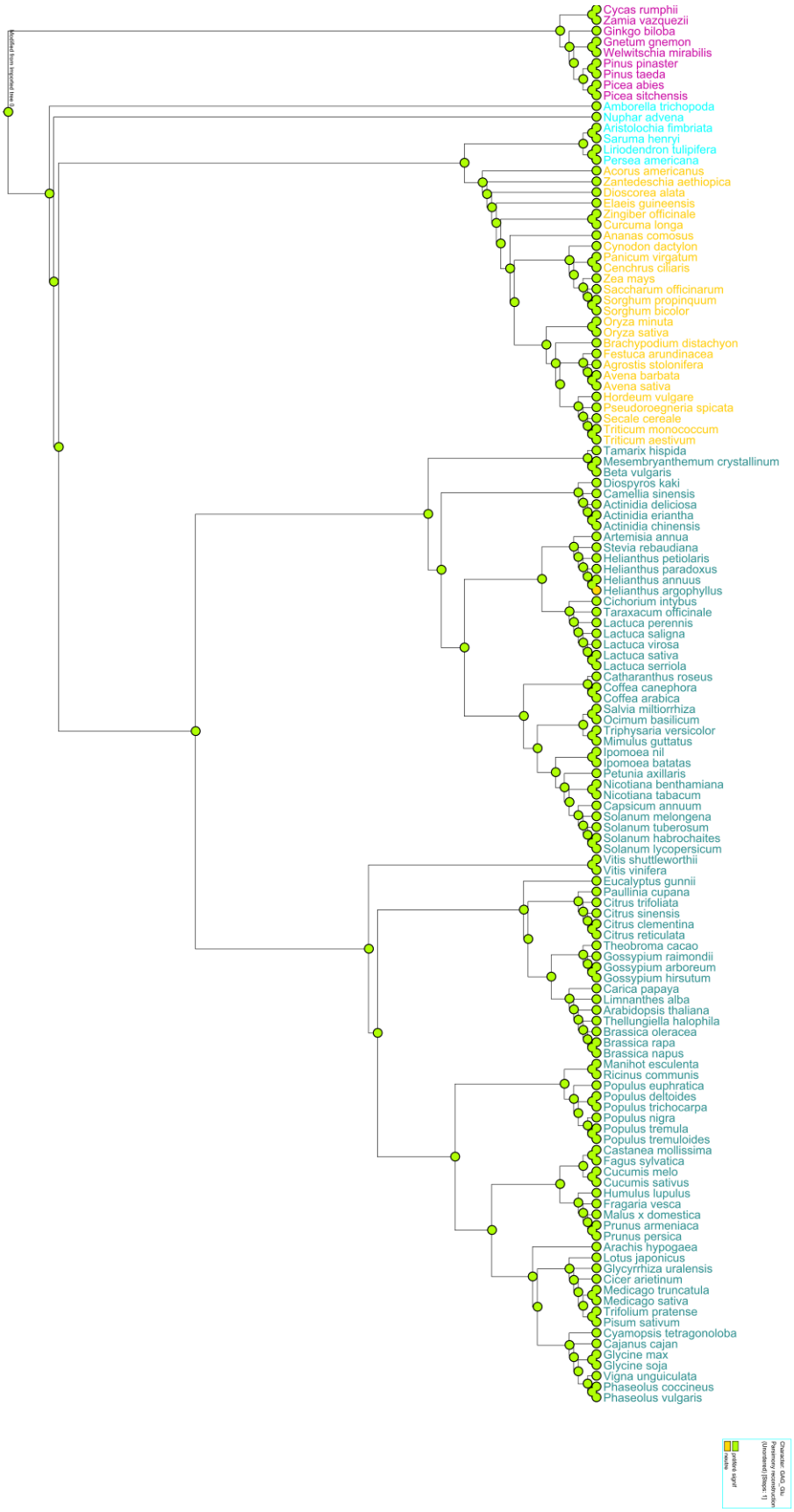
Gln CAG


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

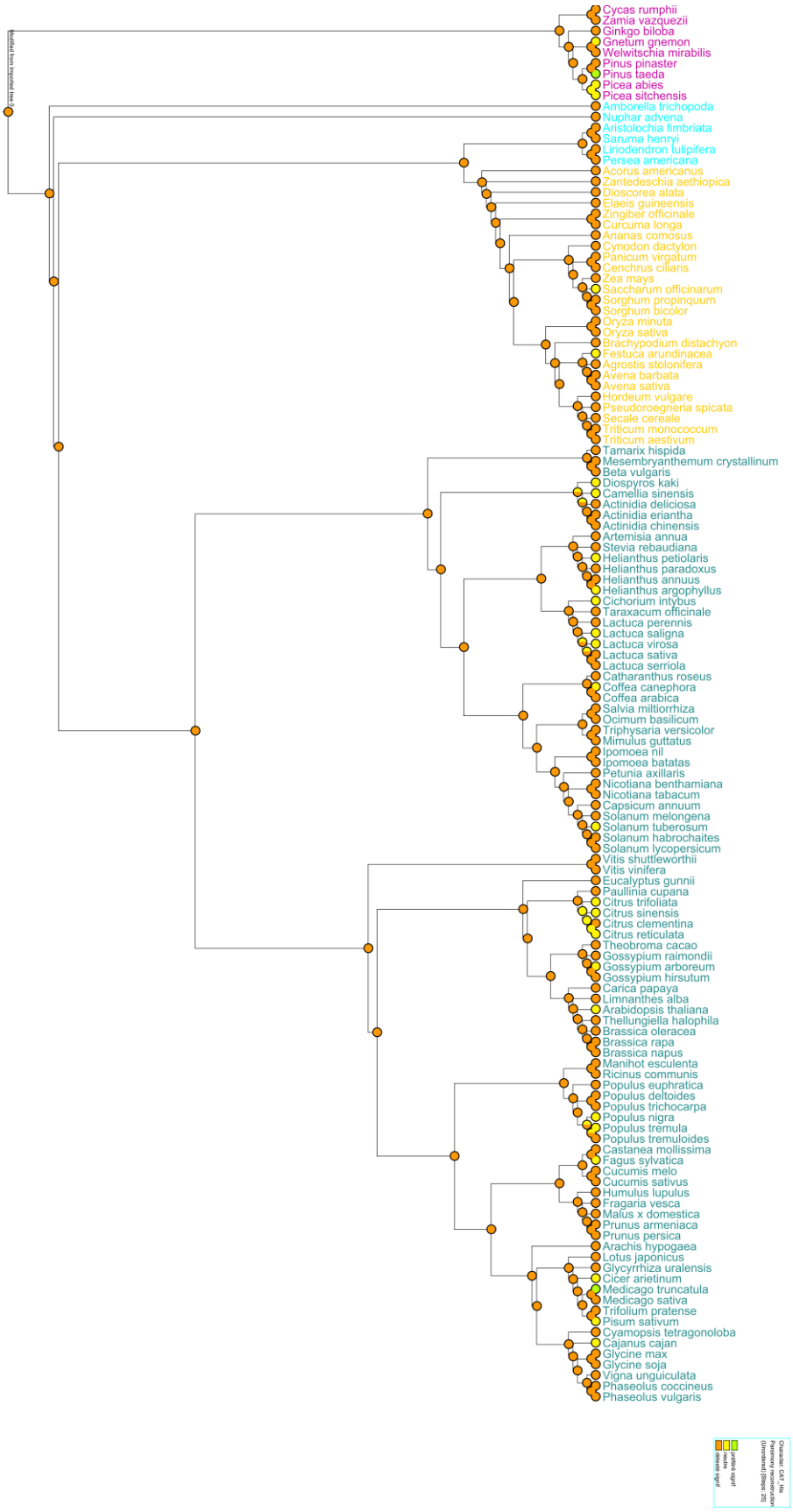
Glu GAA


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

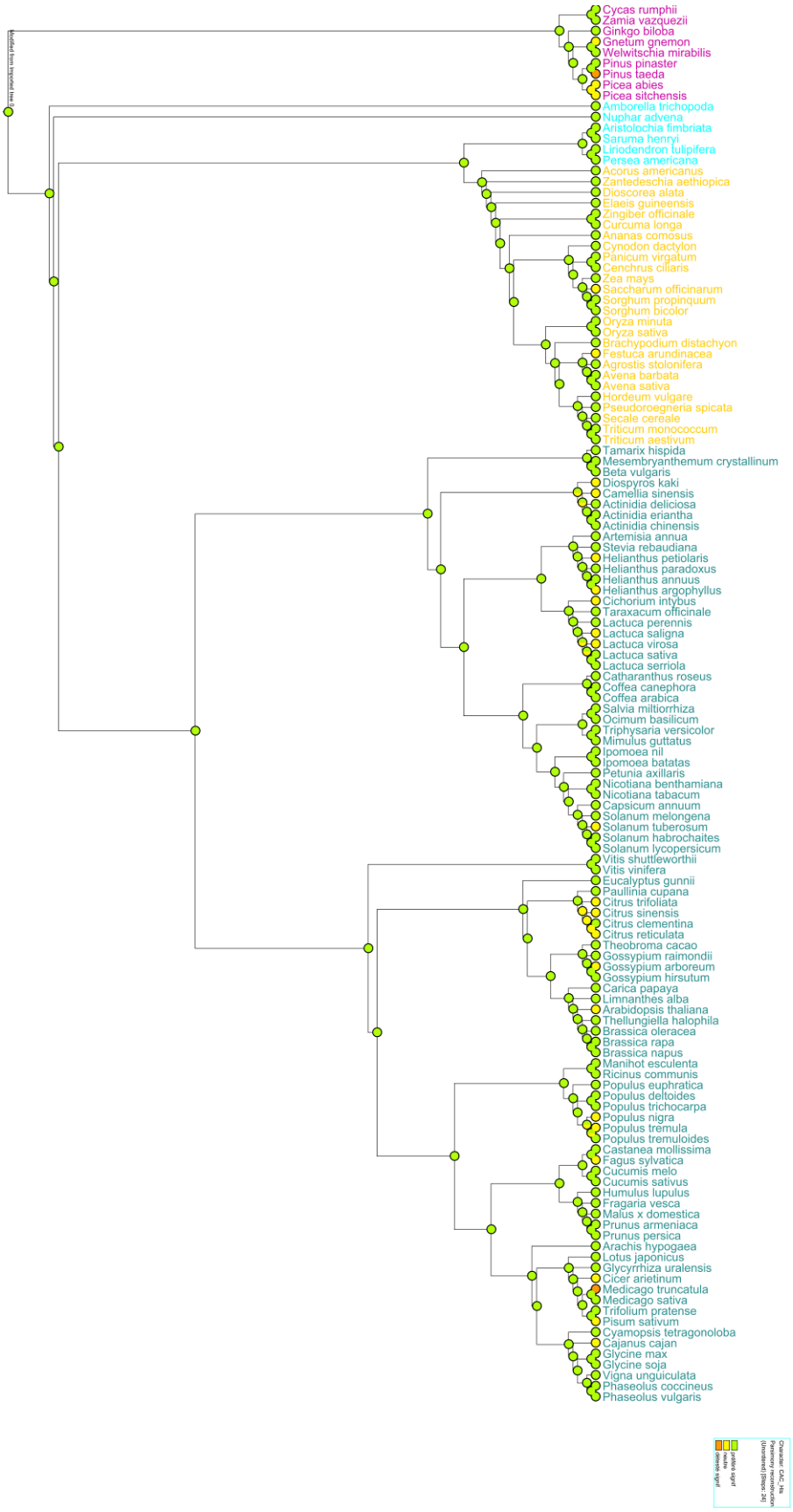
Glu GAG


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GAG_Glu
Parsimony reconstruction
(Unordered) [Steps: 1]
present equivocal
absent

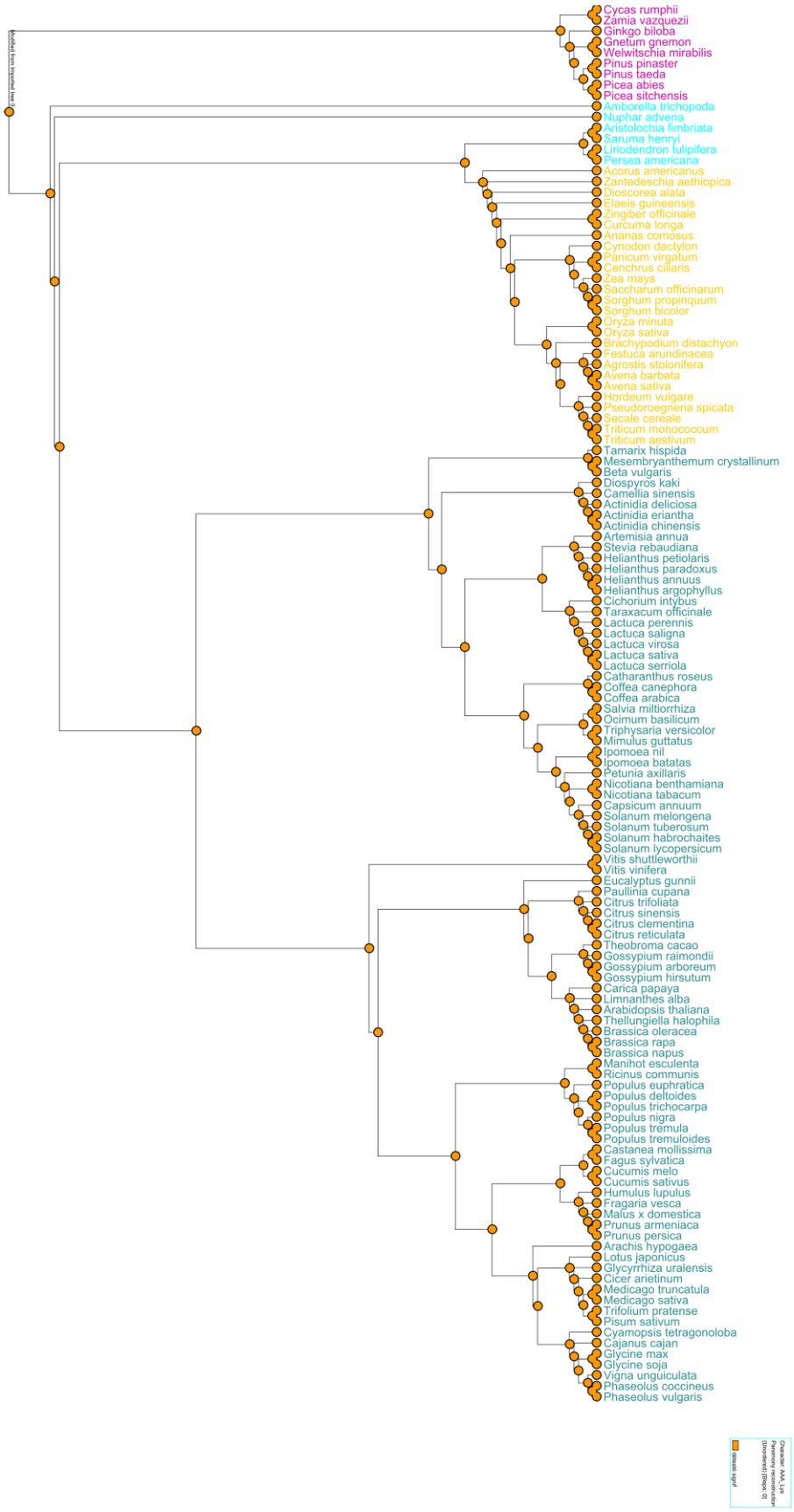
His CAT


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CAT_His
Parsimony reconstruction
(Unordered) (Steps: 28)
yellow signif
ancstr
orange signif

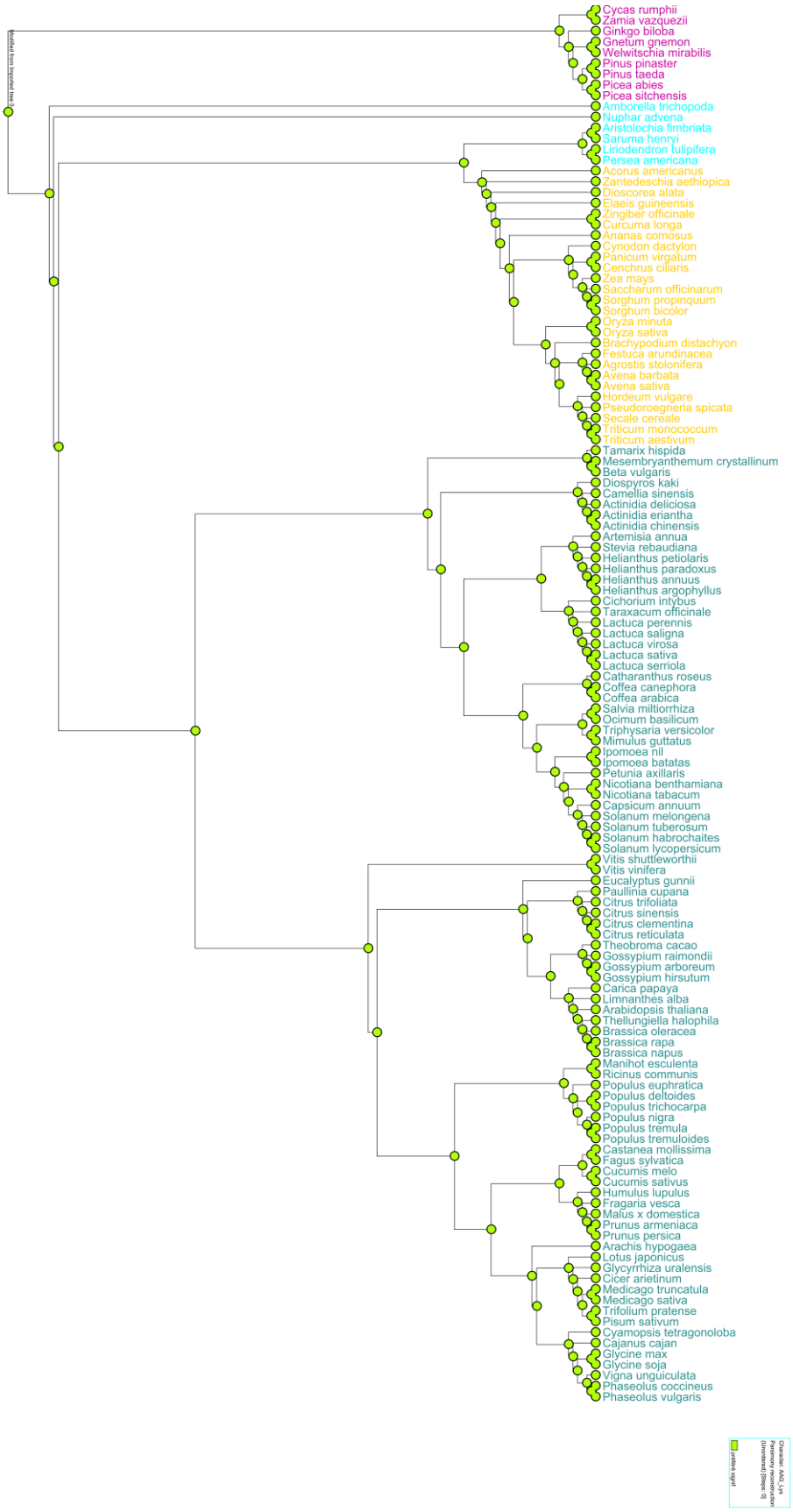
His CAC


Modified from imported tree 1
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CAC_His
Parsimony reconstruction
(Unordered) [Steps: 24]
present state
absence state
inferred state

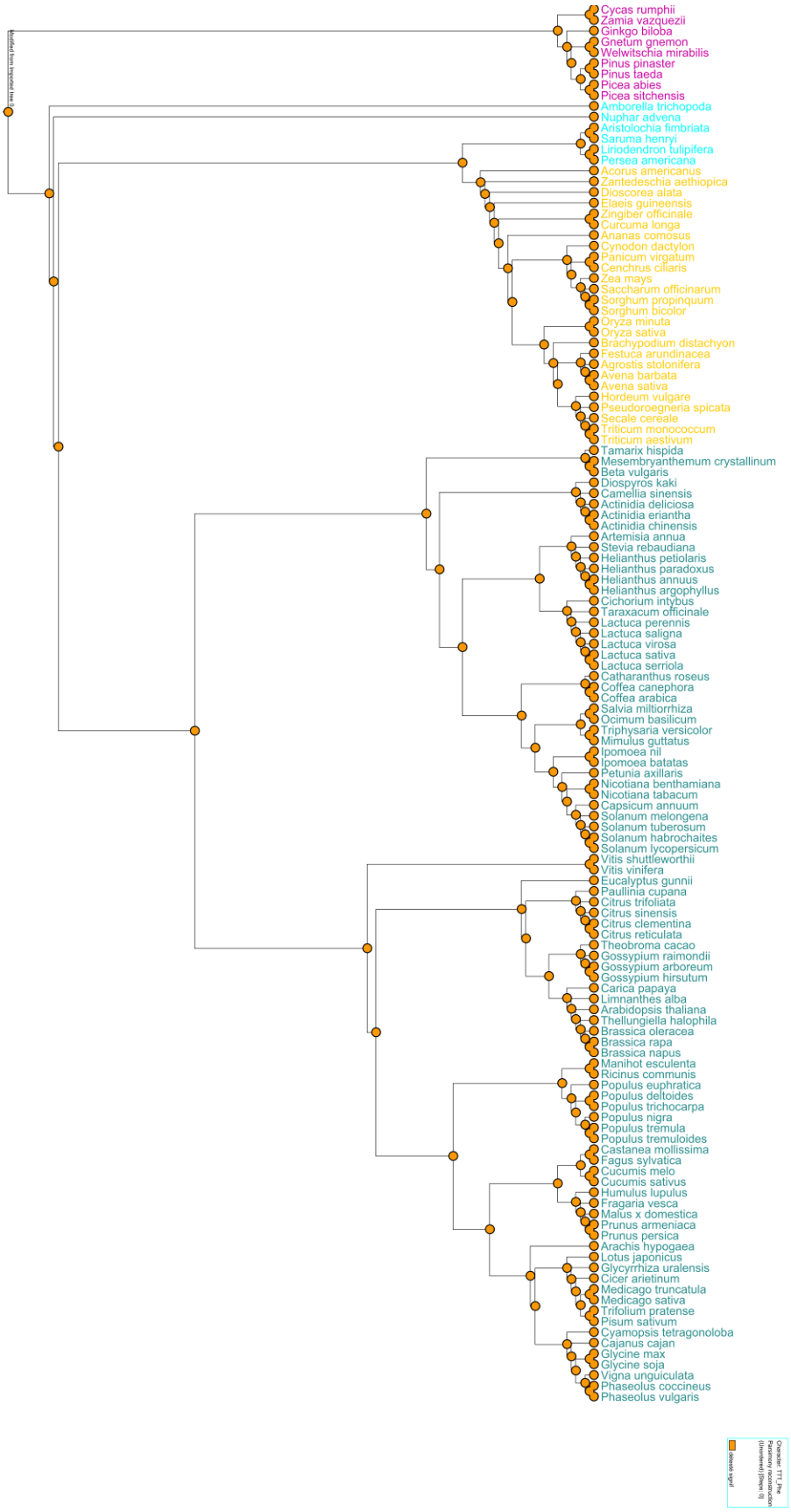
Lys AAA


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: AAA_Lys
Parsimony reconstruction
(Unordered) [Steps: 0]
absolute signif

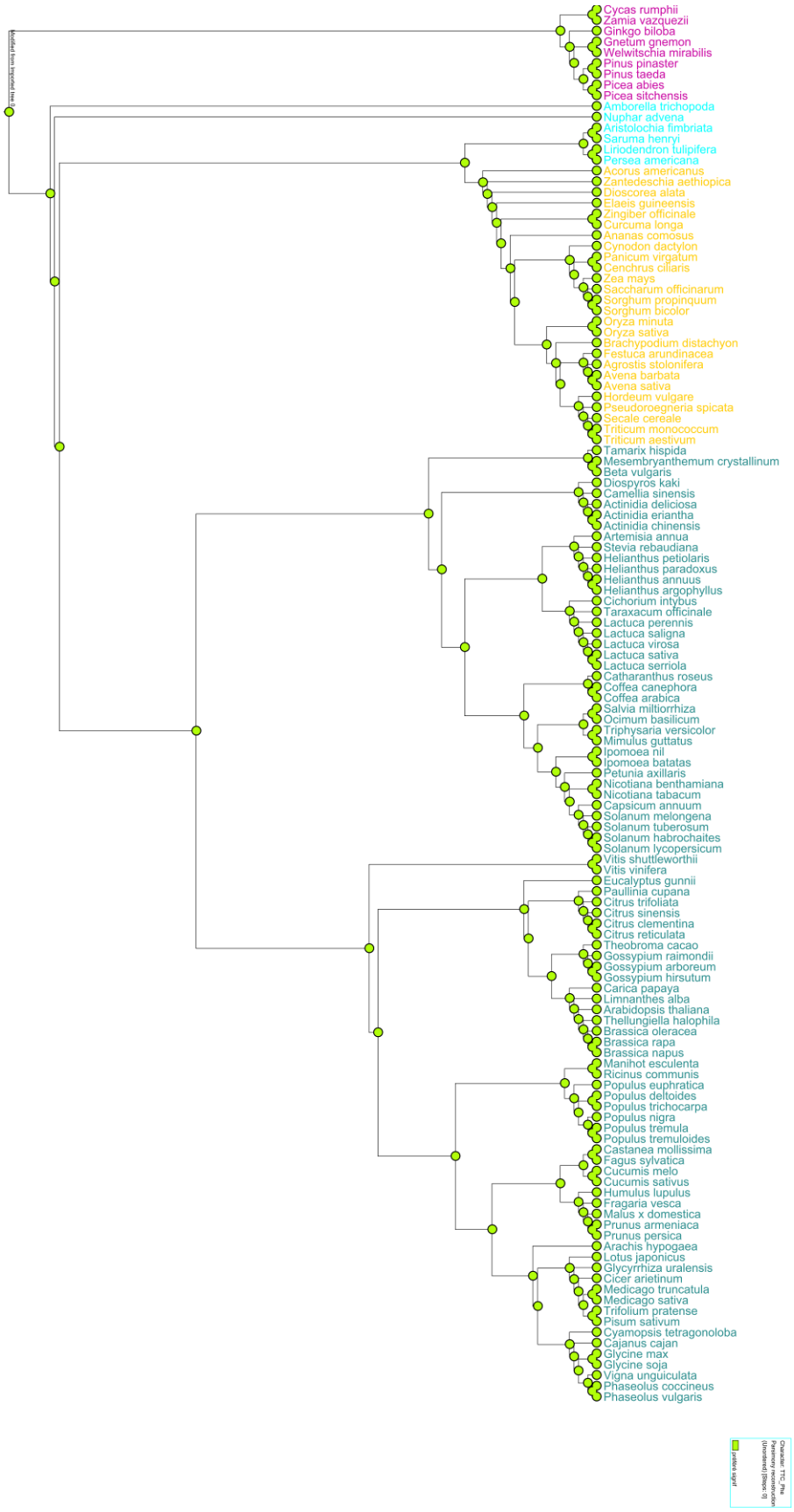
Lys AAG


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: AAG Lys
Parsimony reconstruction
(Unordered) [States: 0]
evidence equal

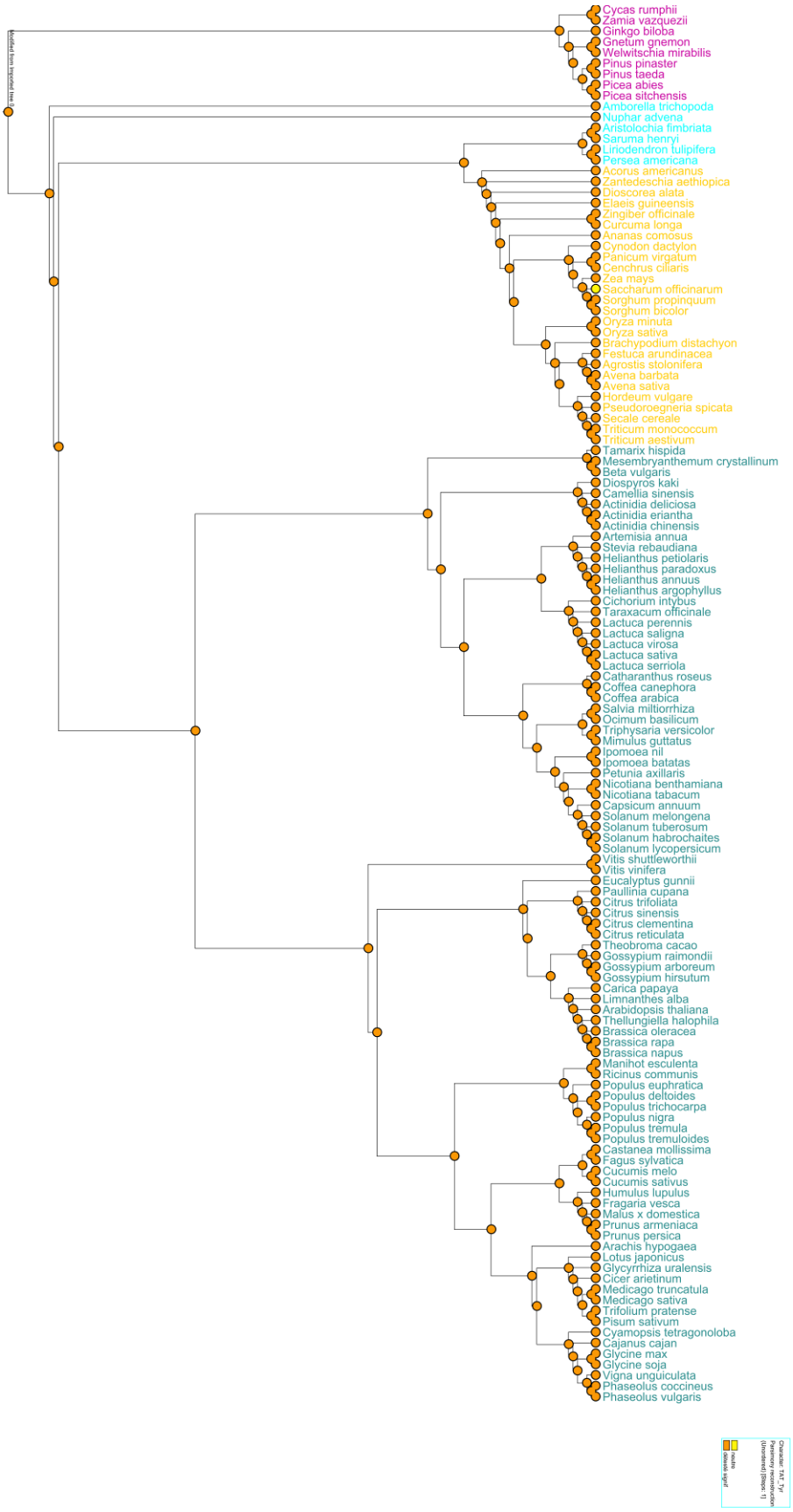
Phe TTT


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TTT, Phe
Parsimony reconstruction
(Unordered) [Steps: 0]
absent
present

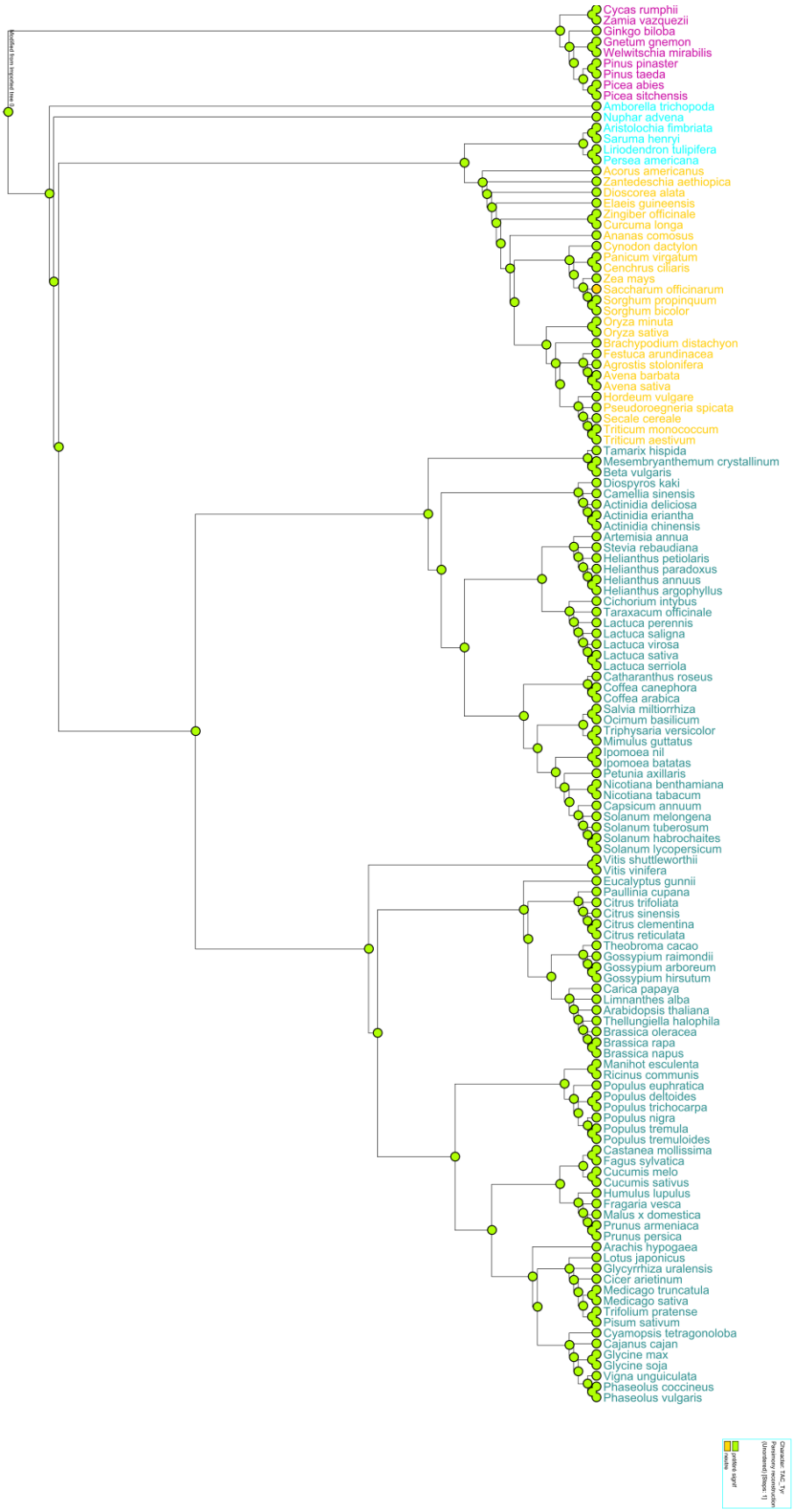
Phe TTC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TTC_Phe
Parsimony reconstruction
(Unordered) (Steps: 0)
additive sig=1

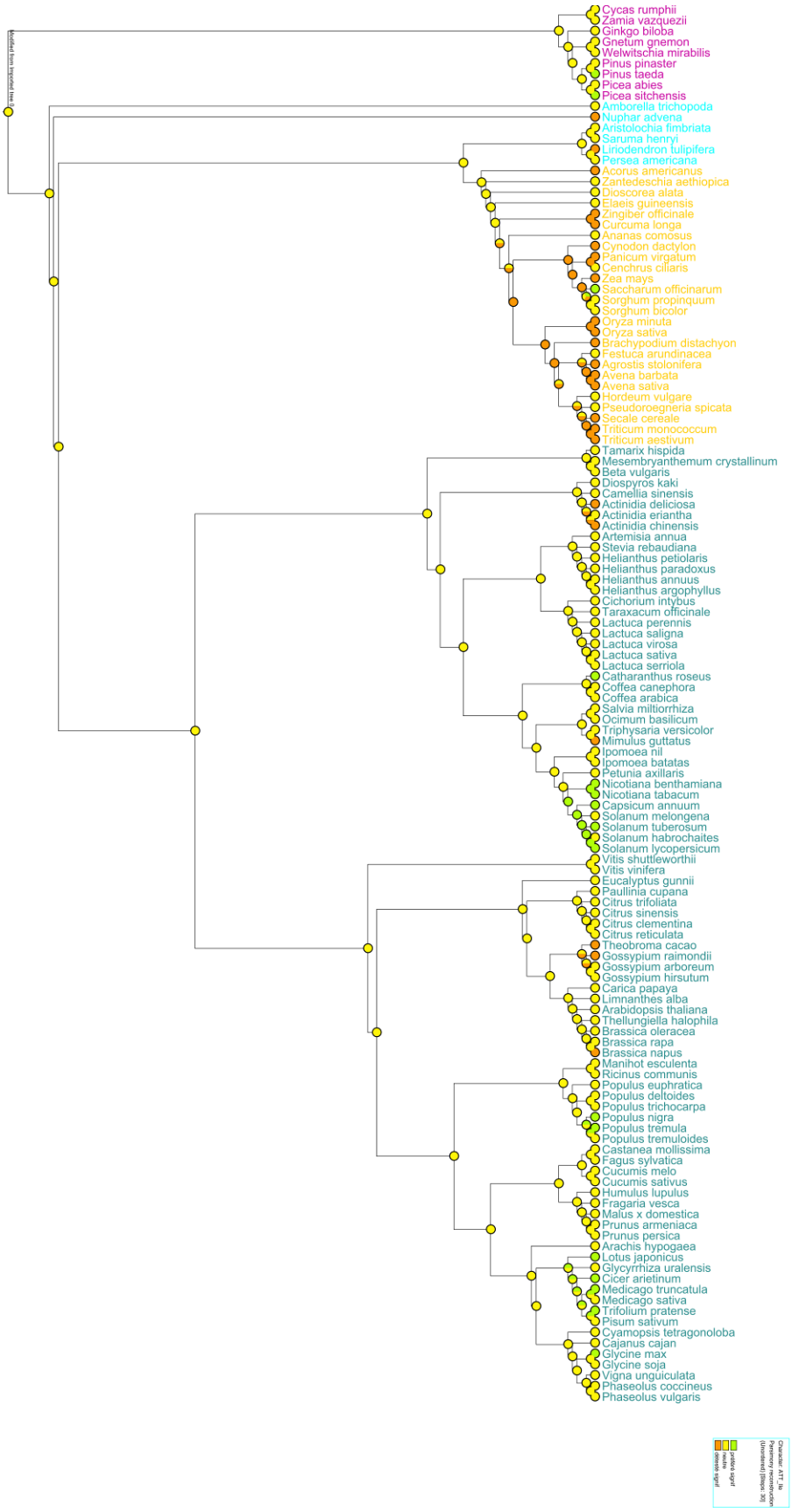
Tyr TAT


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

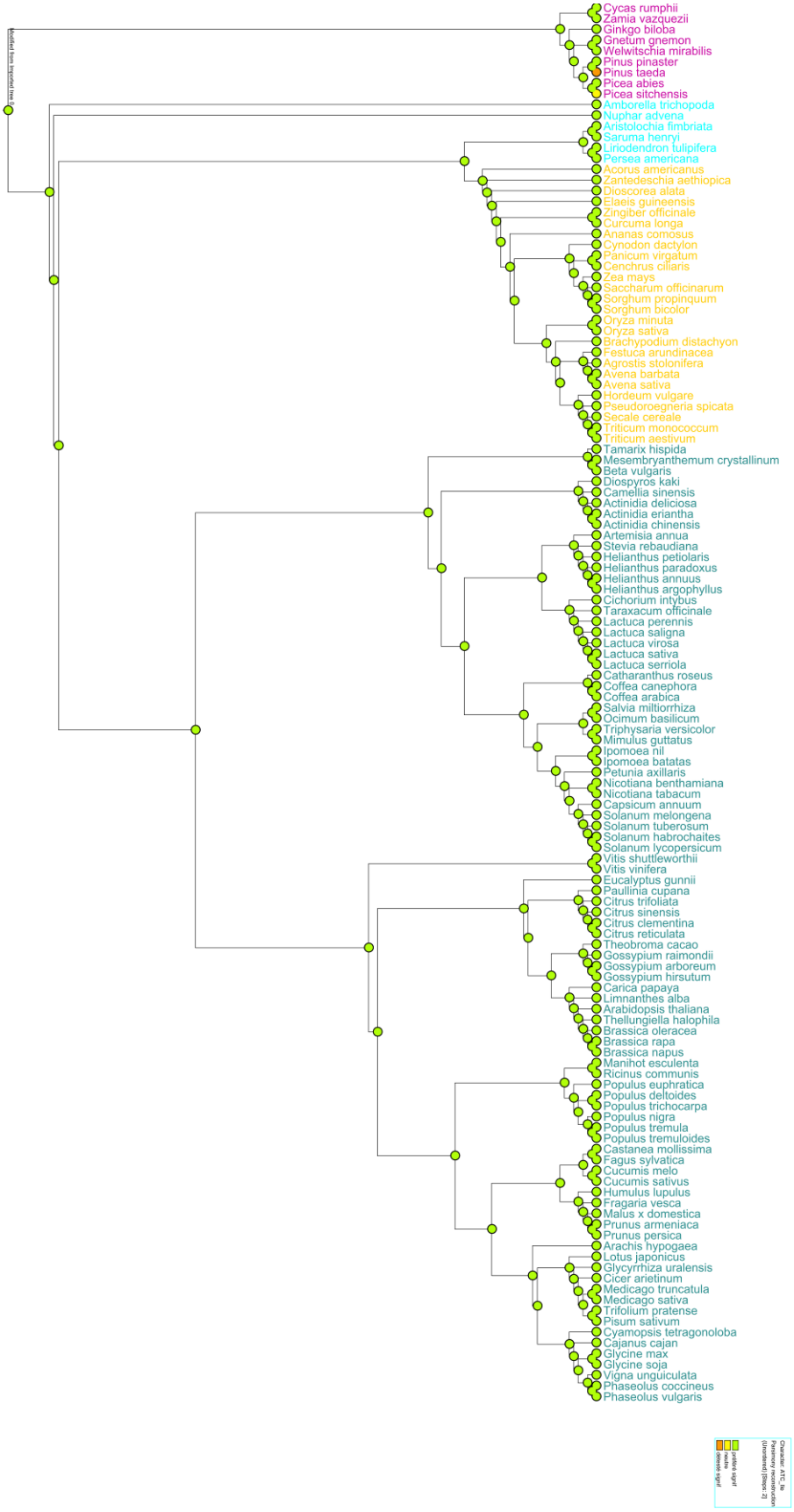
Tyr TAC


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

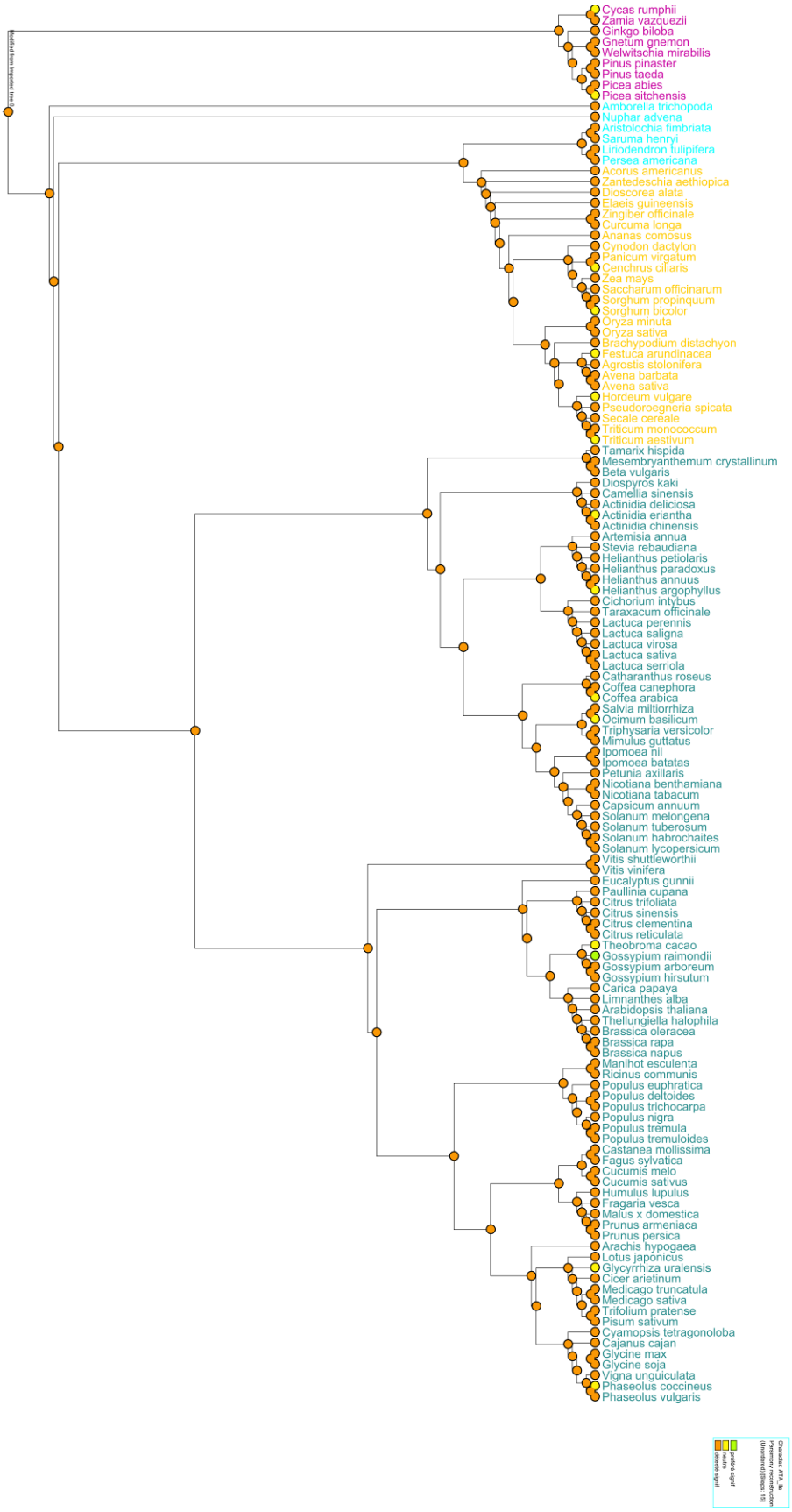
Ile ATT


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

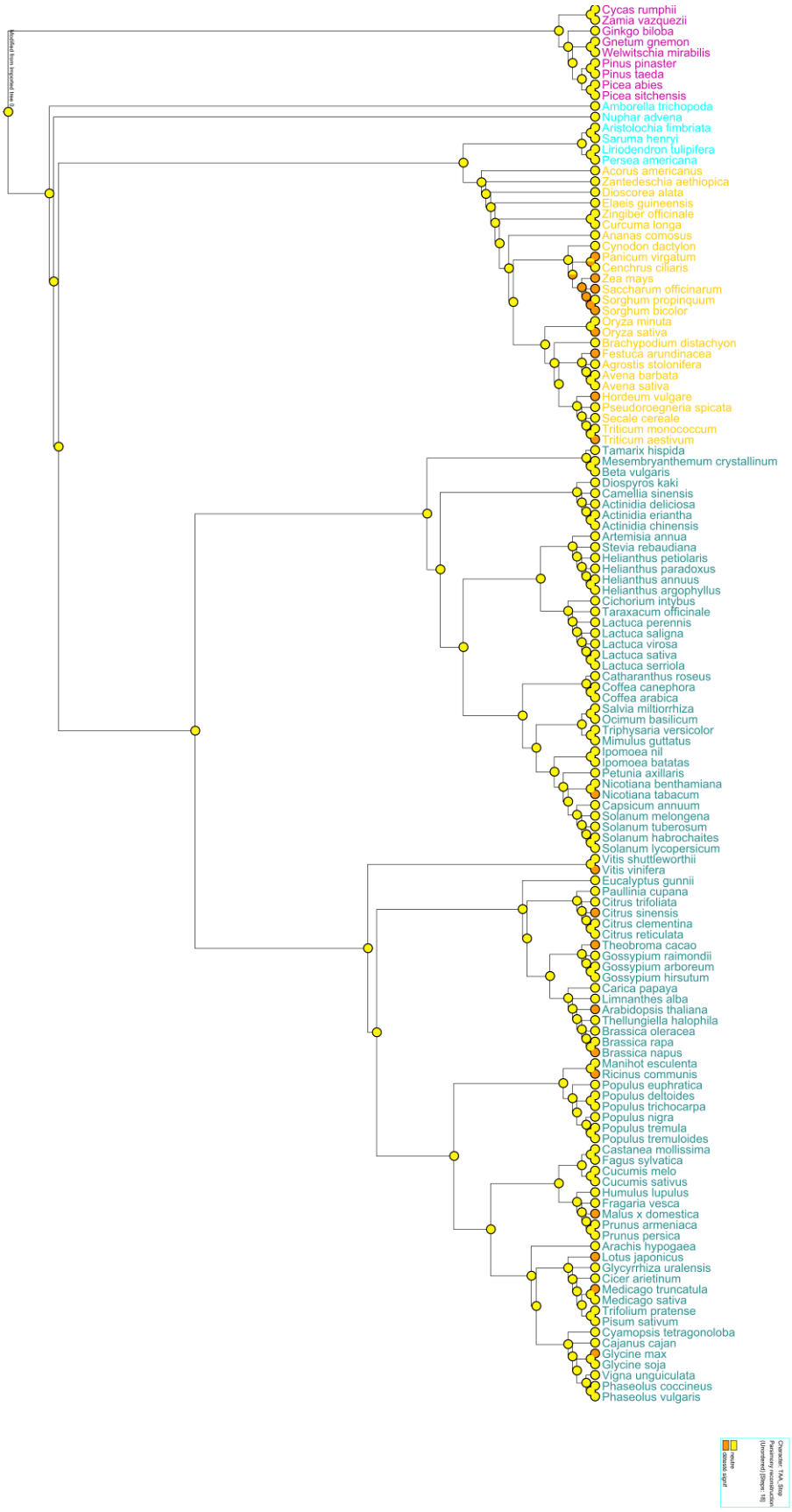
Ile ATC


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

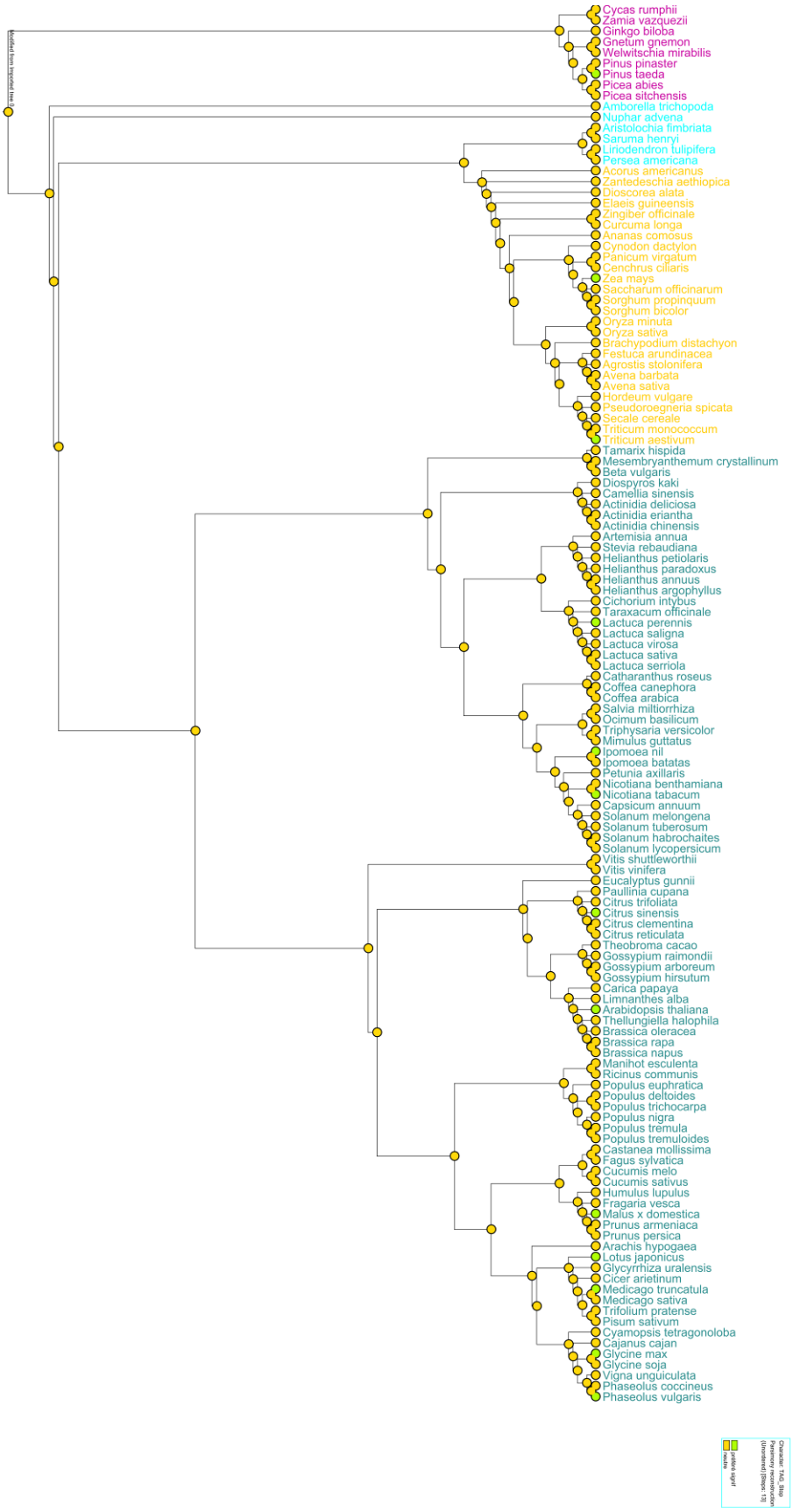
Ile ATA


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

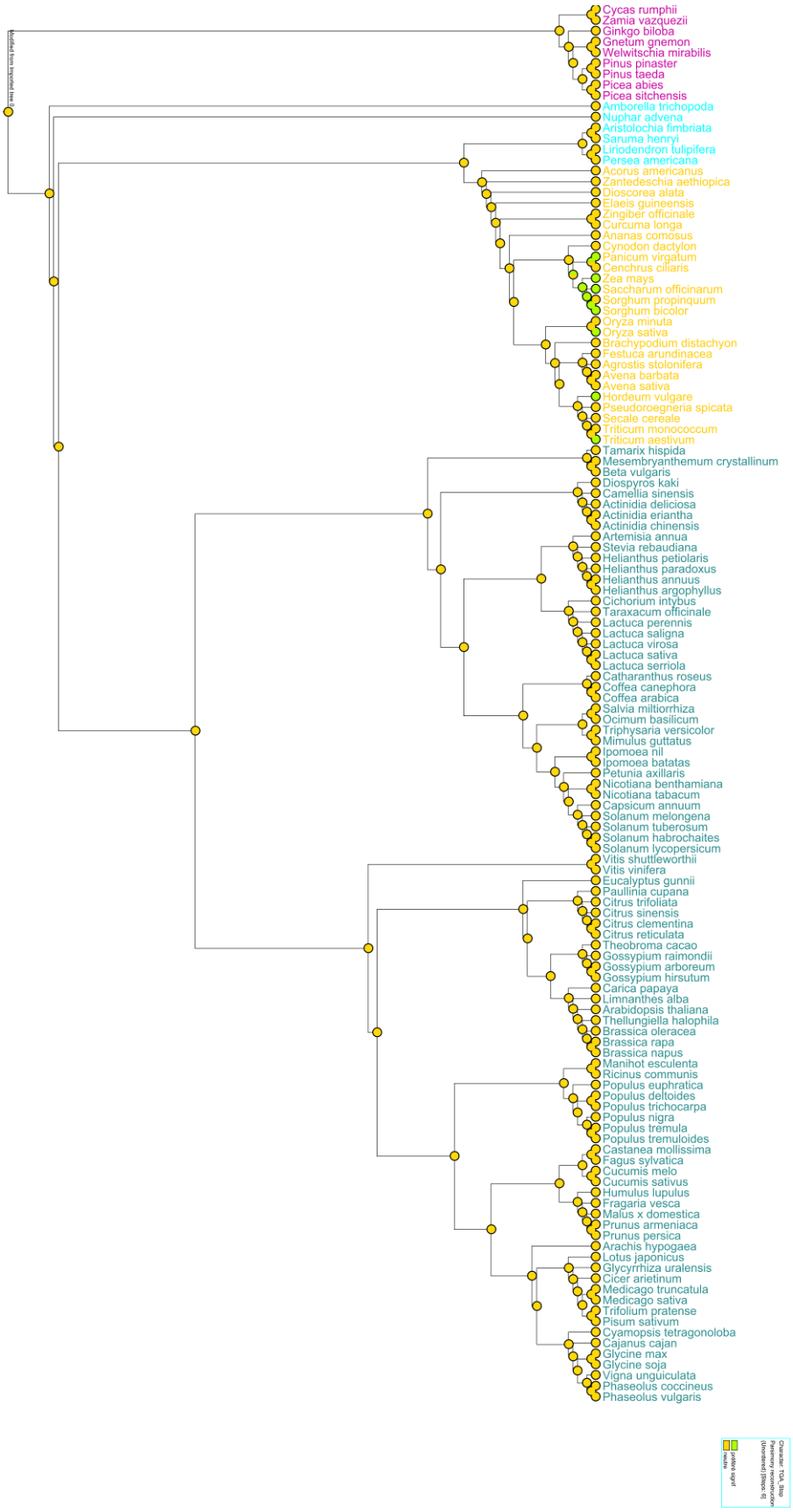
Stop TAA


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TAA_Stop
Parsimony reconstruction
(Unordered) (Steps: 16)
multi
ambig
absent
present

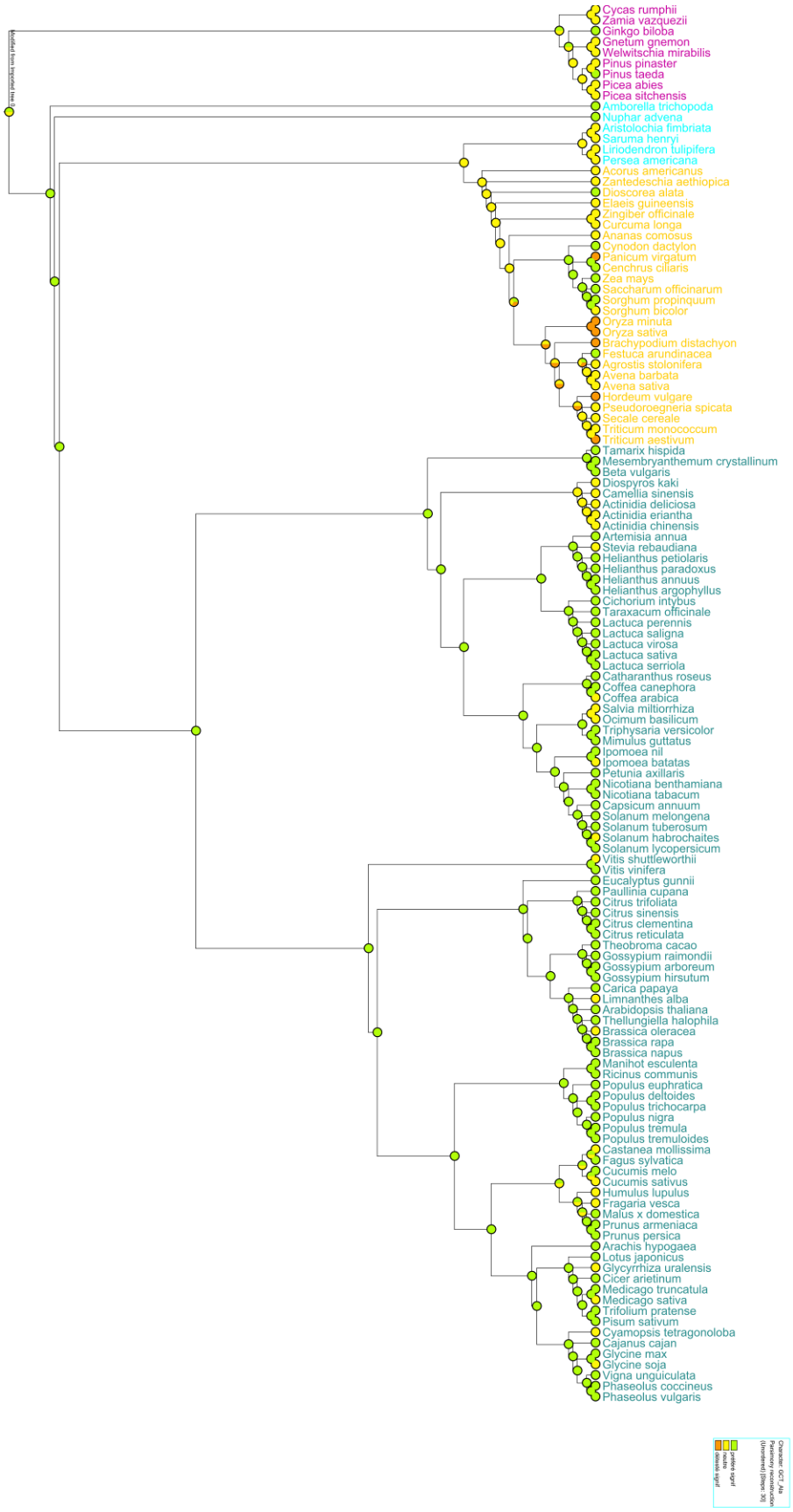
Stop TAG


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

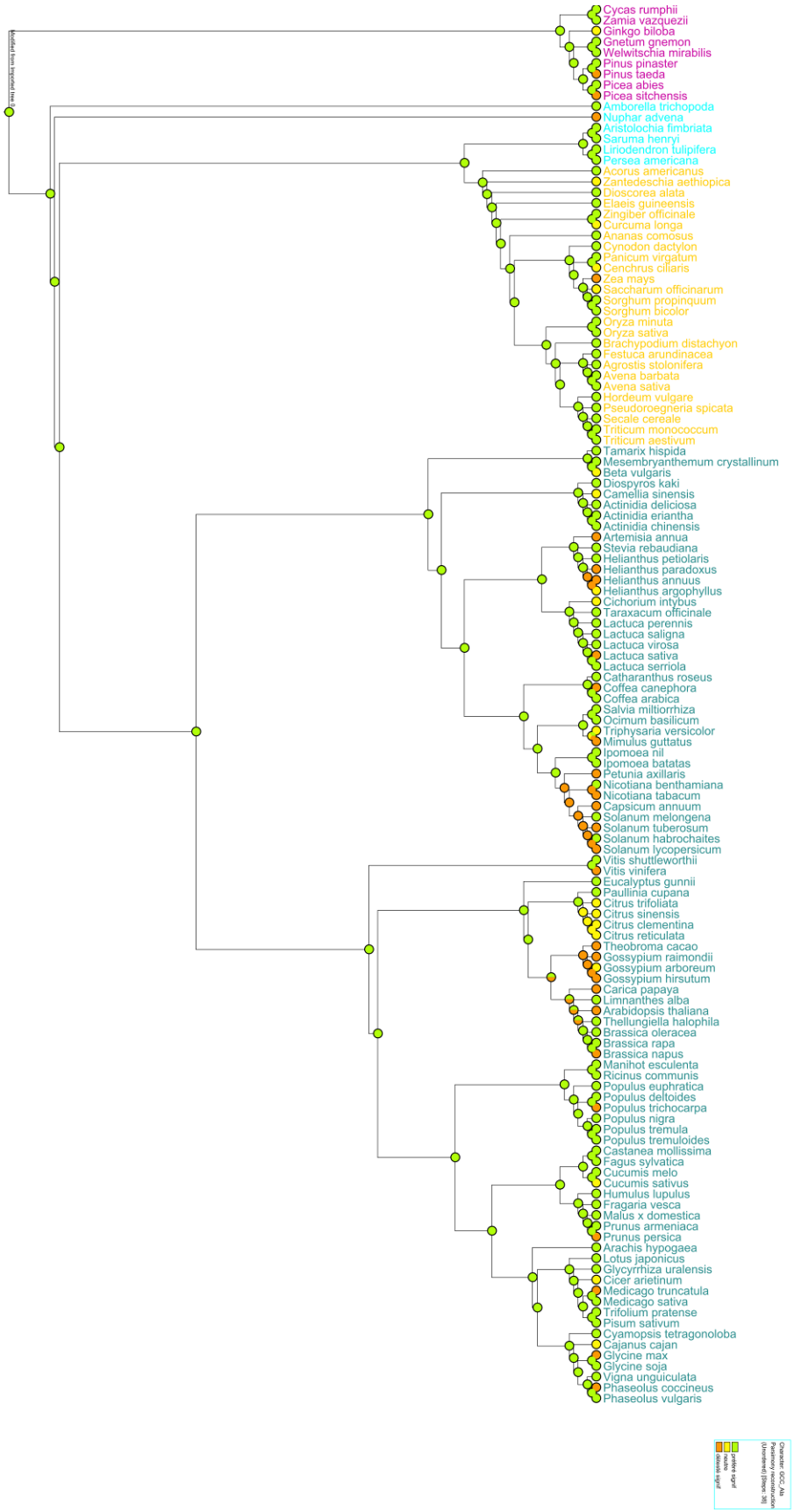
Stop TGA


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TGA_Stop
Parsimony reconstruction
(Unordered) [States: 0]
Ambiguous state
Native

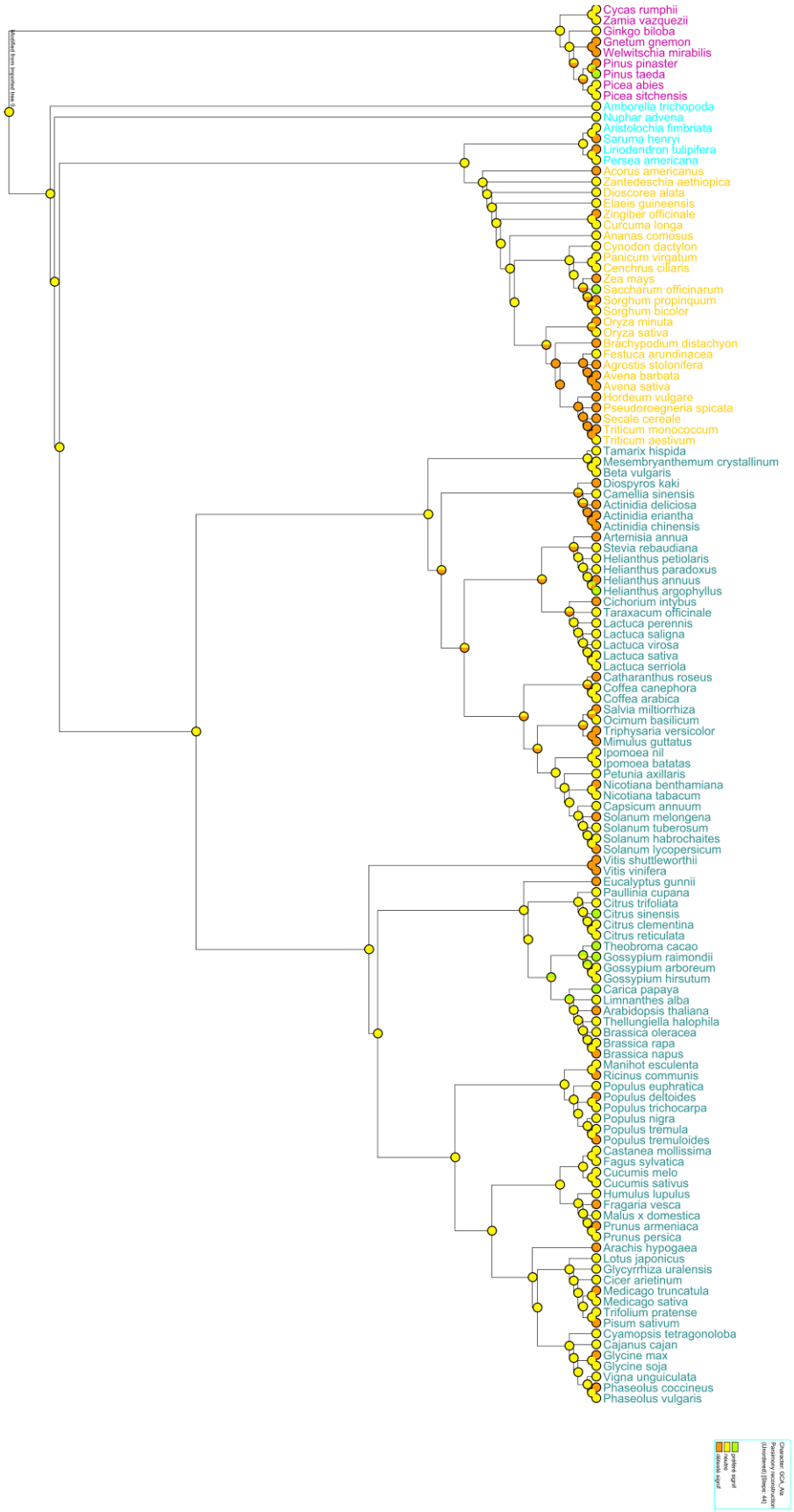
Ala GCT


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

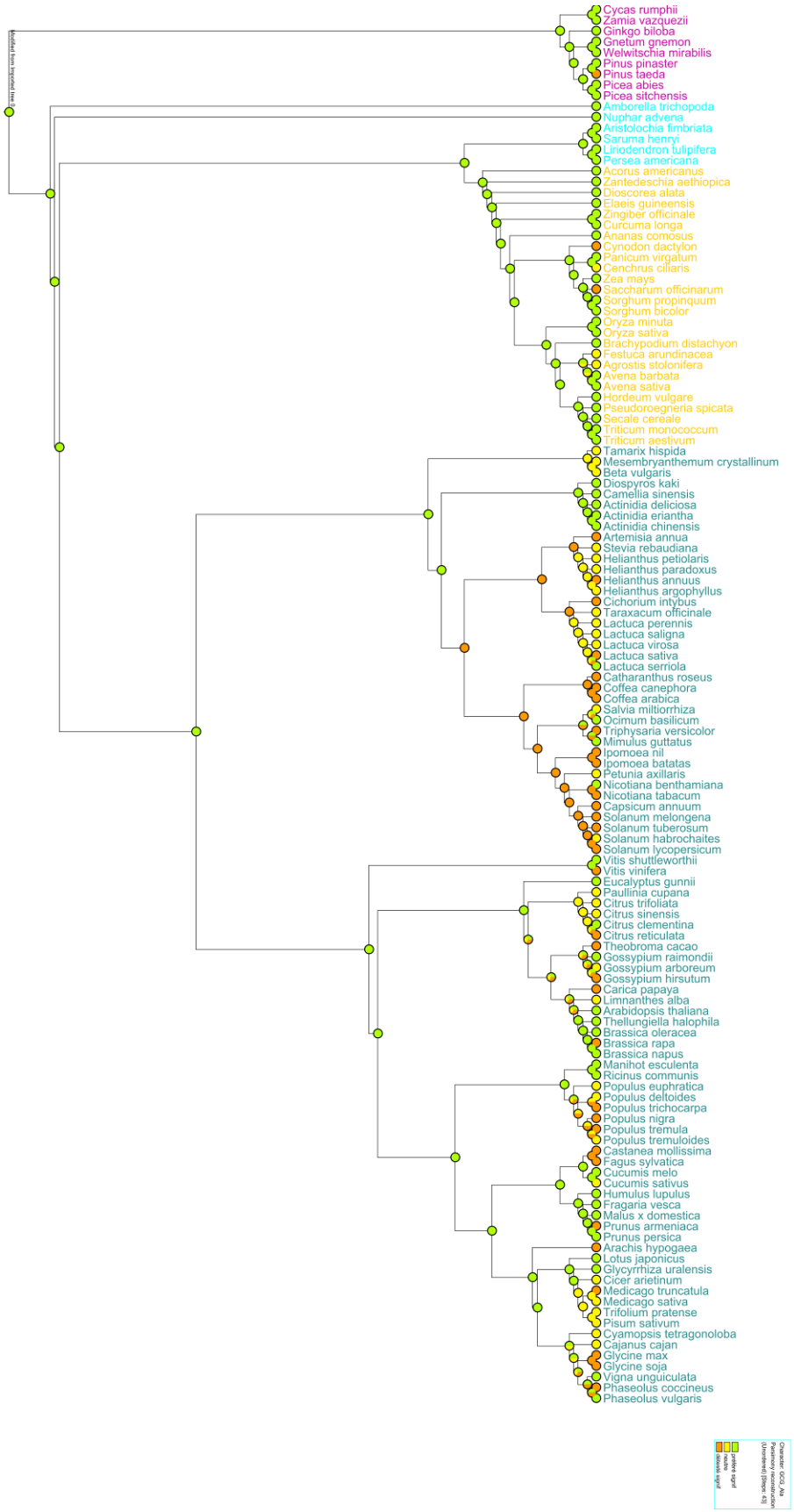
Ala GCC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GCC_Ala
Parsimony reconstruction
(Unordered) [Steps: 38]
Ambiguous signal
Absent signal

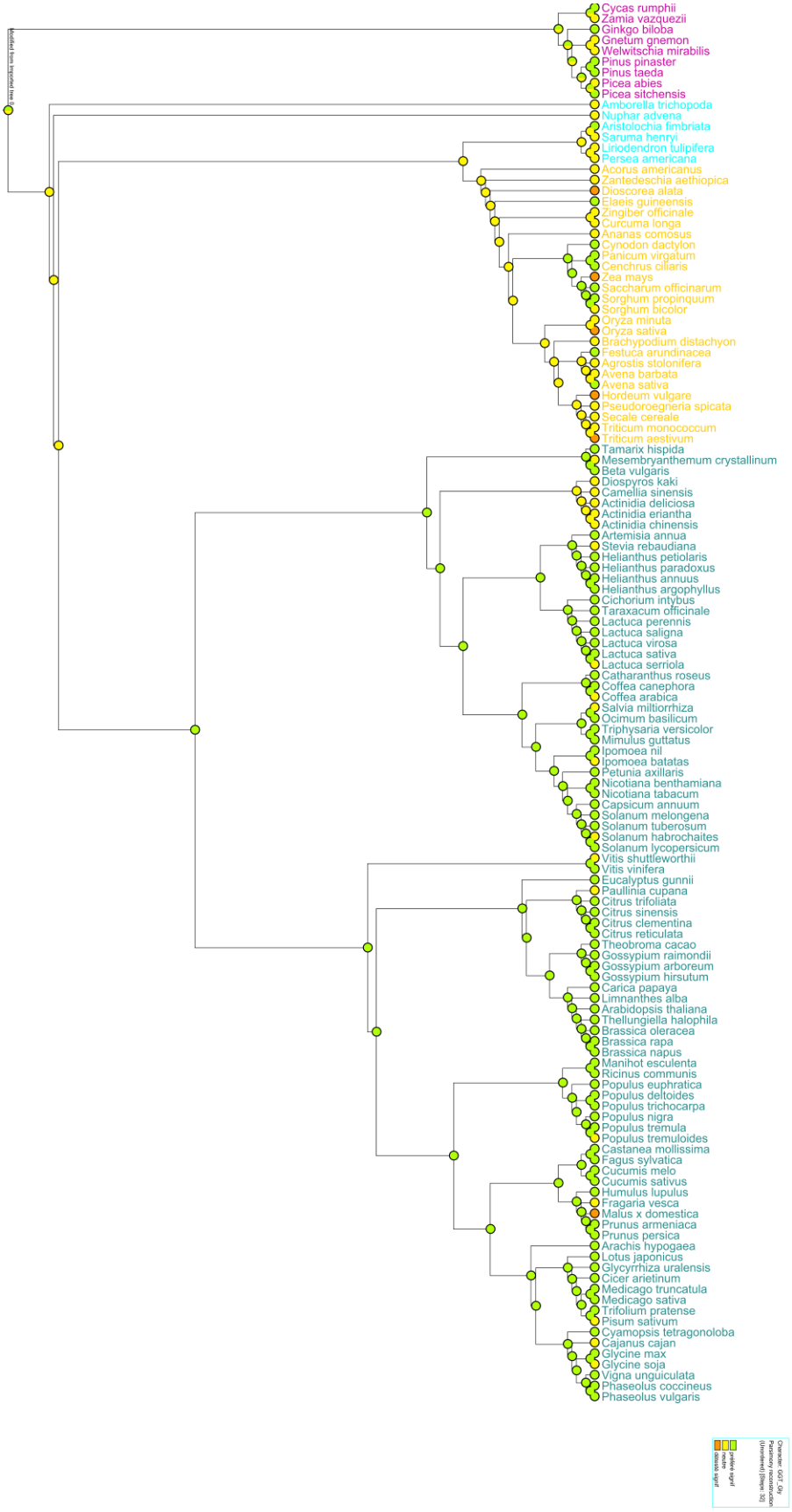
Ala GCA


Modified from Imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GCA_Ala
Parsimony reconstruction
(Unordered) [Steps: 44]
ambiguous
absence signif.
presence signif.

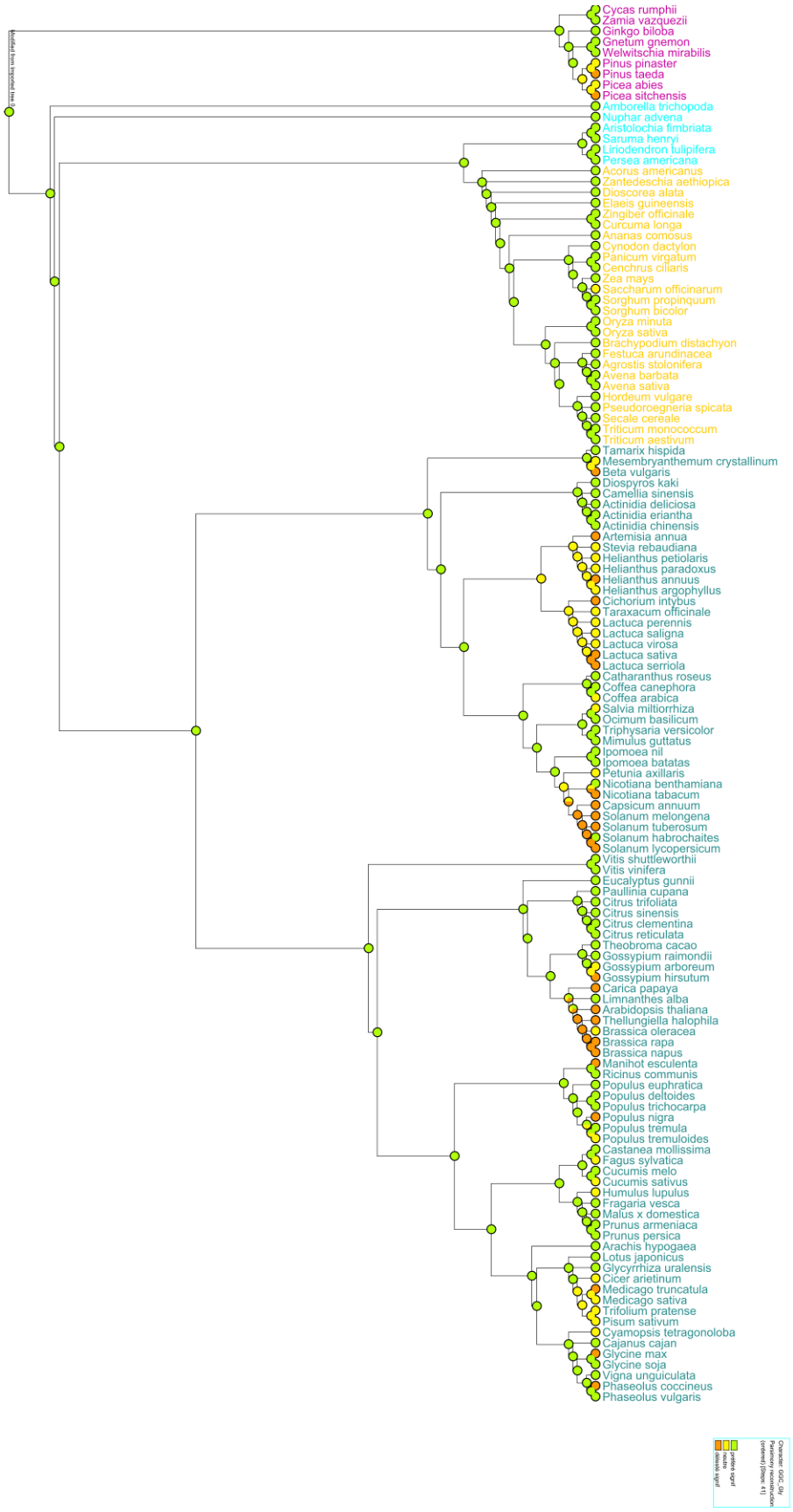
Ala GCG


Modified from Imported tree 1
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GCG_Ala
Parsimony reconstruction
(Unordered) [Steps: 43]
ambiguous
absent signif
present signif

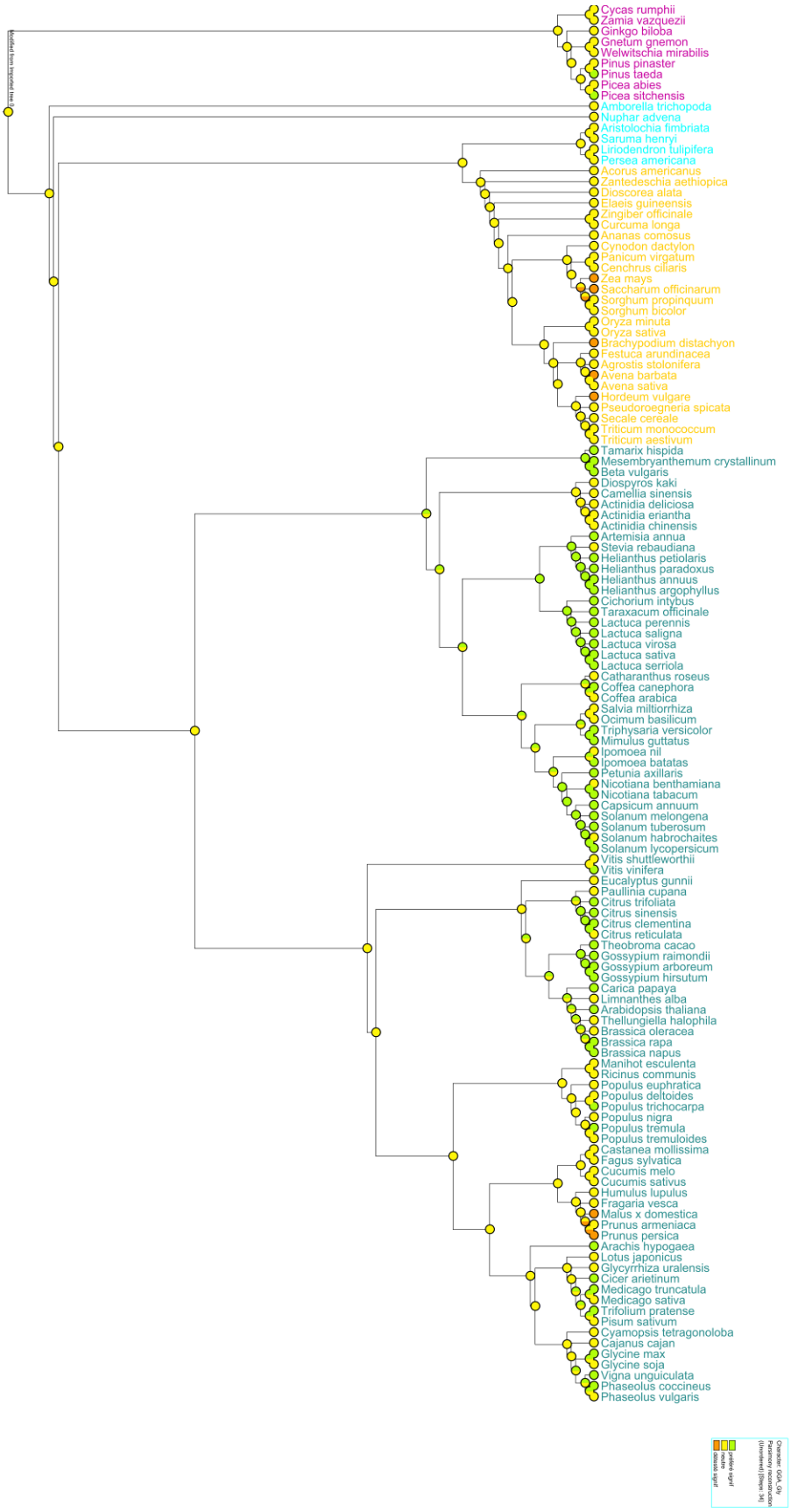
Gly GGT


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GGT_Gly
Parsimony reconstruction
(Unordered) [Steps: 32]
ambiguous signal
absence signal
presence signal

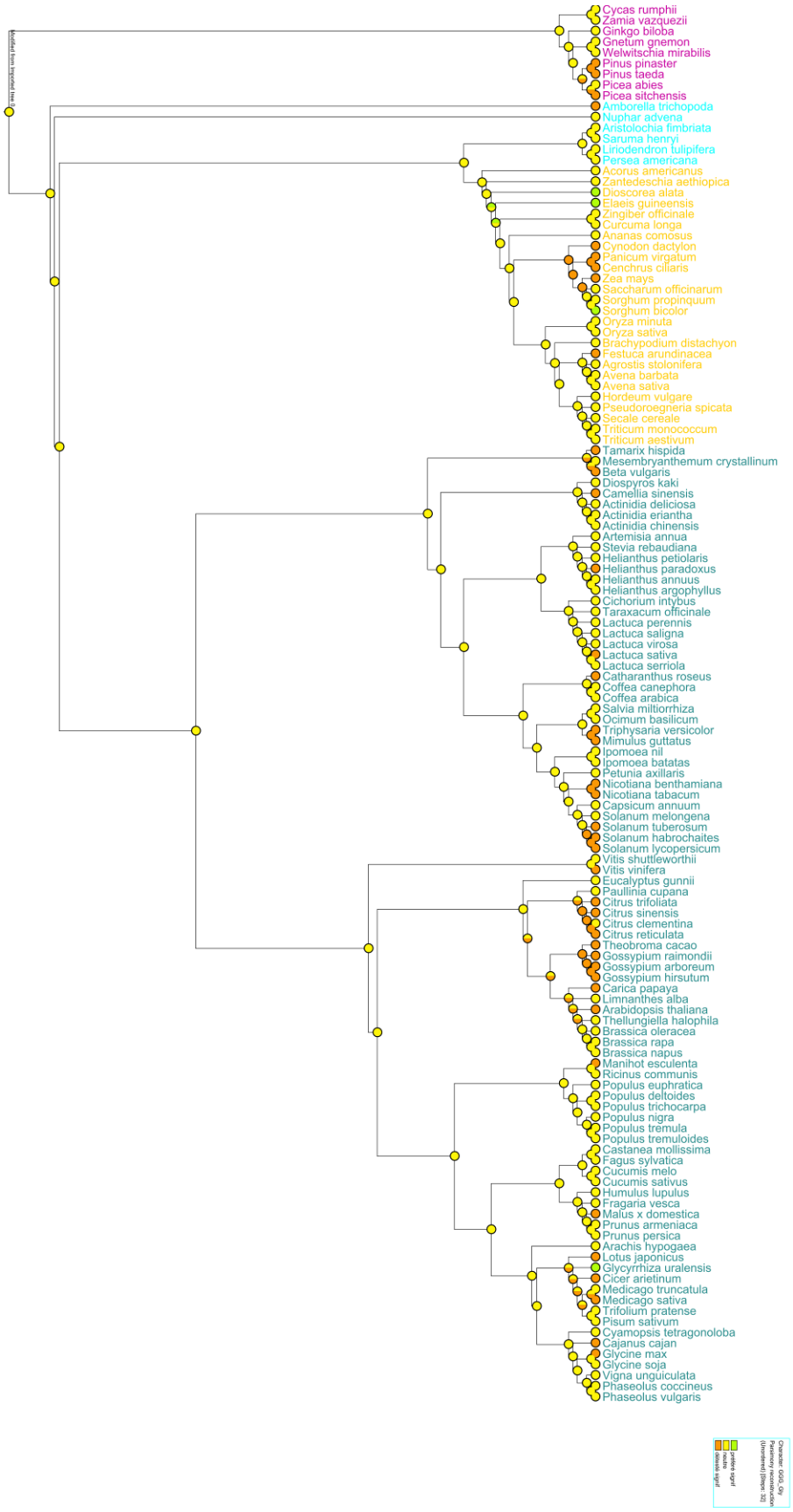
Gly GGC


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GGC_Gly
Parsimony reconstruction
(Unknown) [Steps: 41]
ambiguous signal
no state
absolute signal

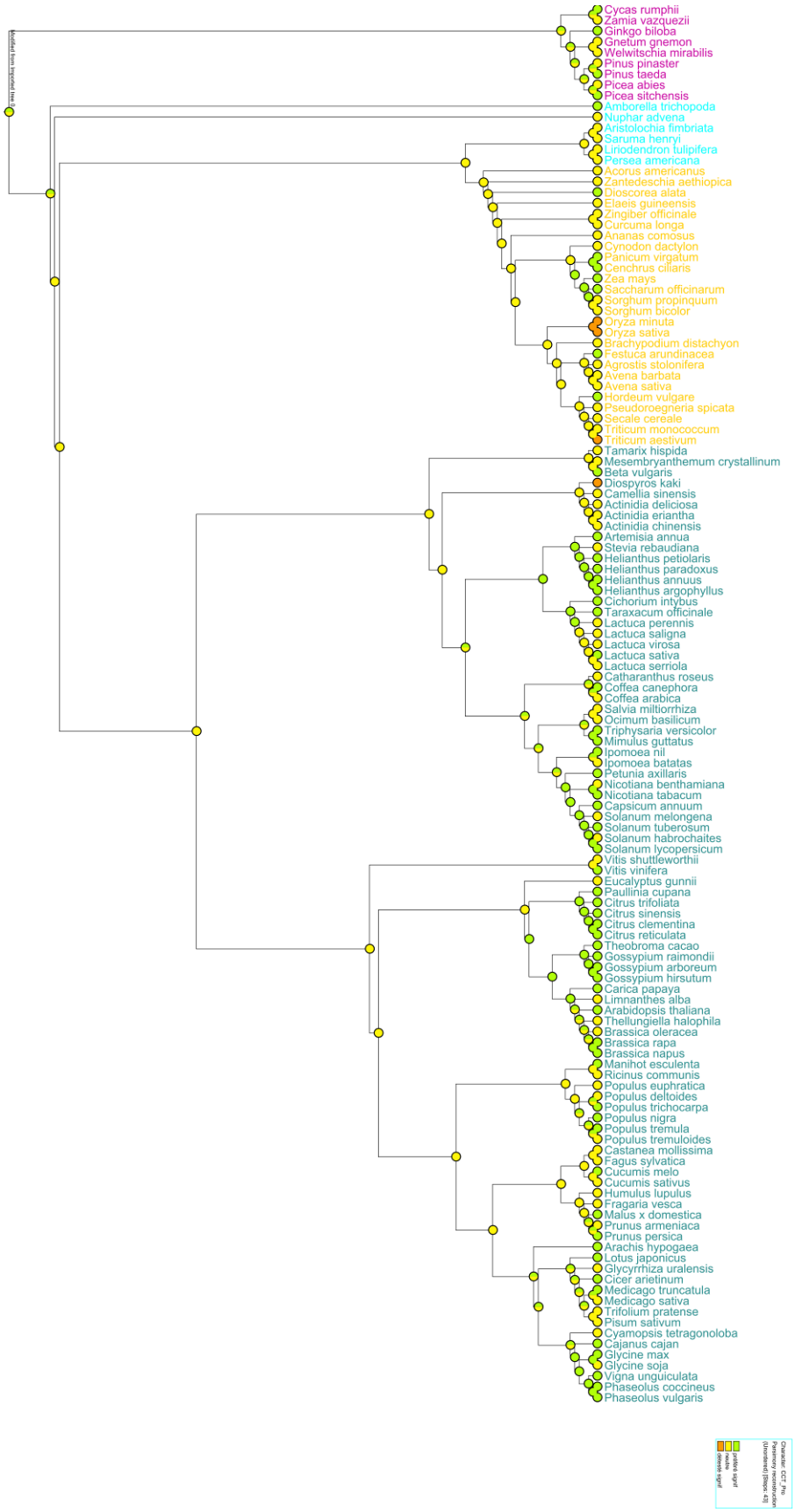
Gly GGA


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

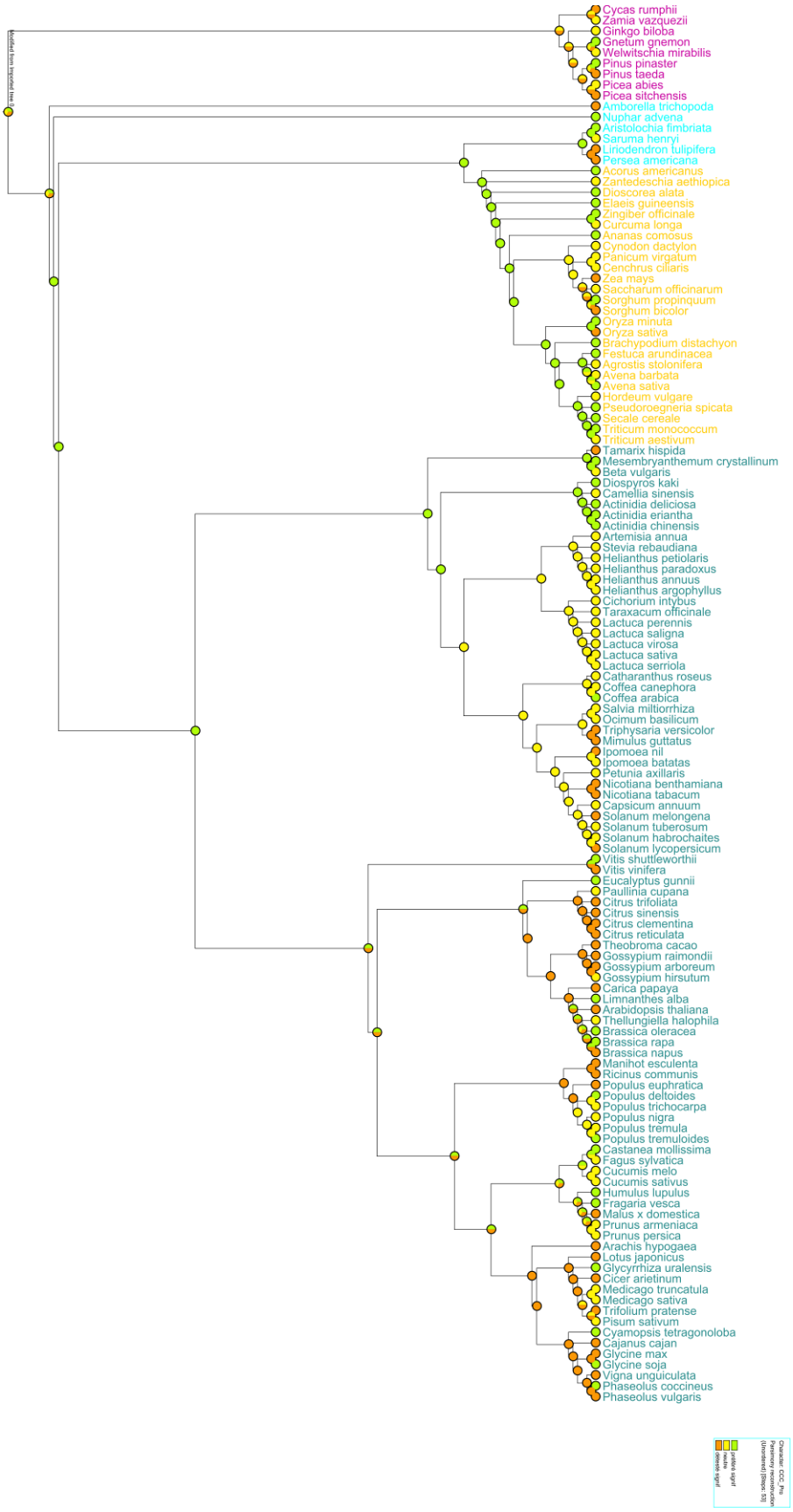
Gly GGG


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

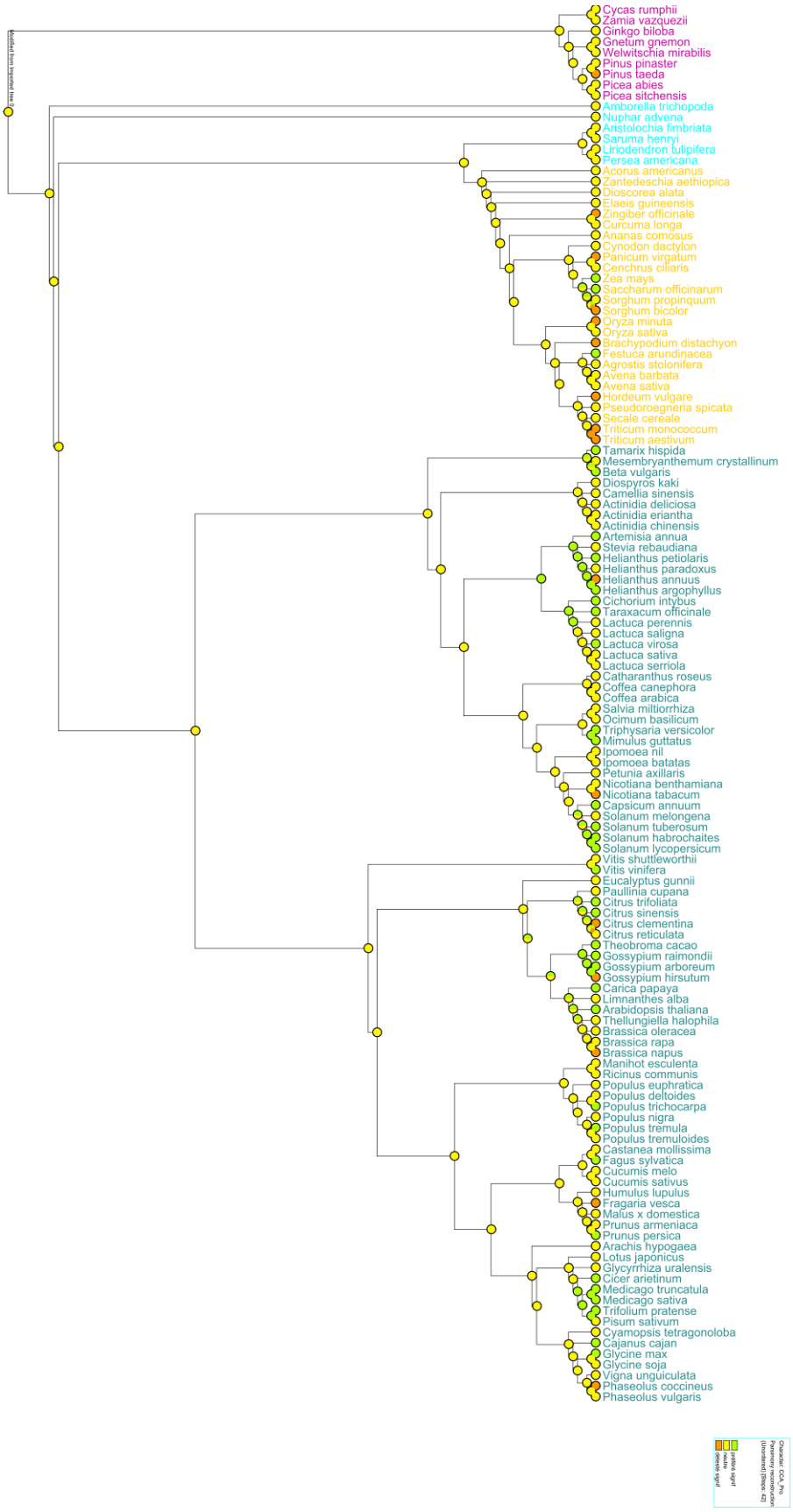
Pro CCT


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CCT_Pro
Parsimony reconstruction
(Unordered) (Steps: 43)
additive signal
ancestral
inactive
inactive signal

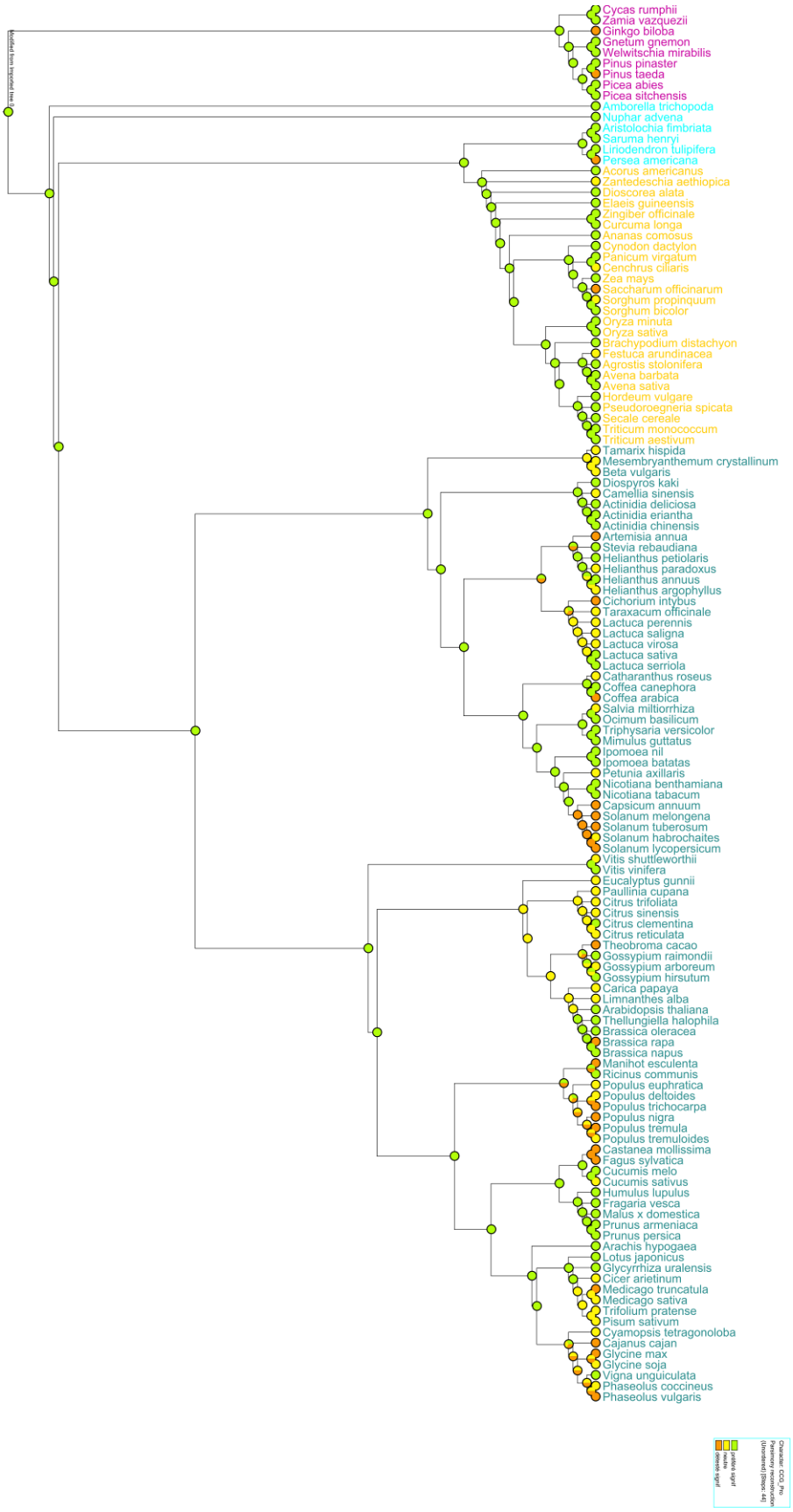
Pro CCC


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CCC_Pro
Parsimony reconstruction
(Unordered) [Steps: 53]
additive signal
ACCTRAN
DELTRAN signal

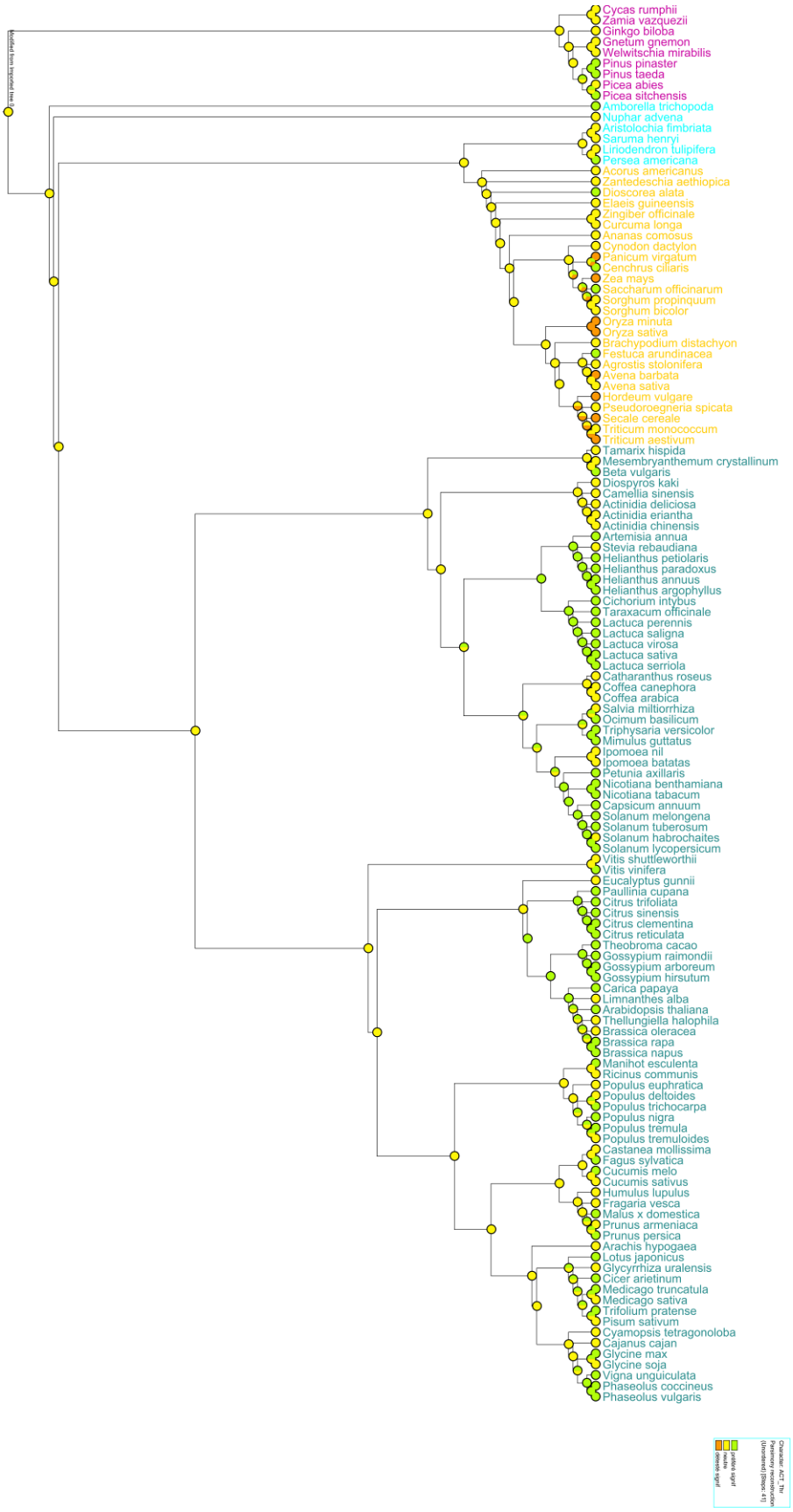
Pro CCA


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CCA_Pro
Parsimony reconstruction
(Unordered) (Steps: 42)
additive signal
ancestral
parsimony signal

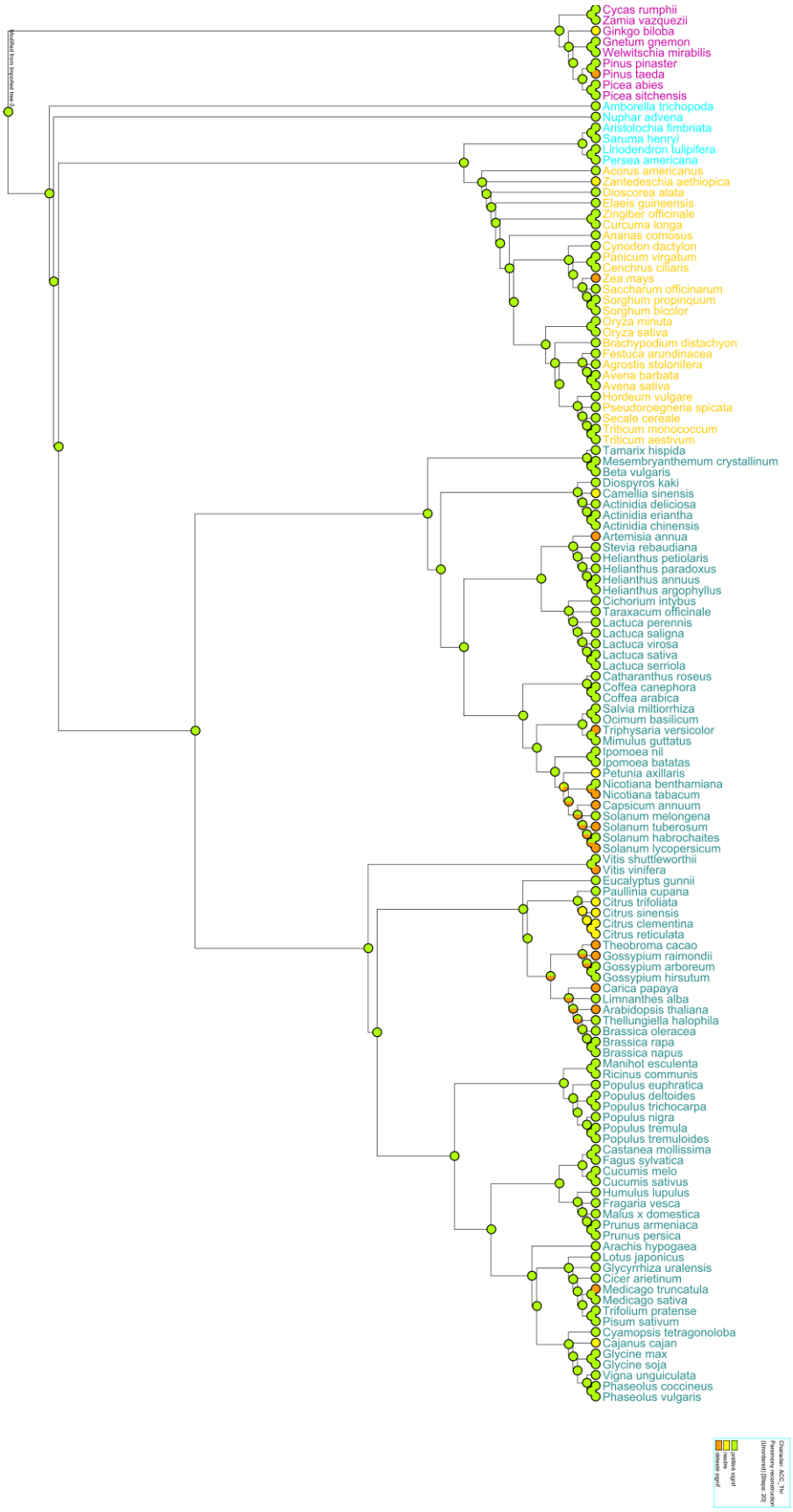
Pro CCG


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

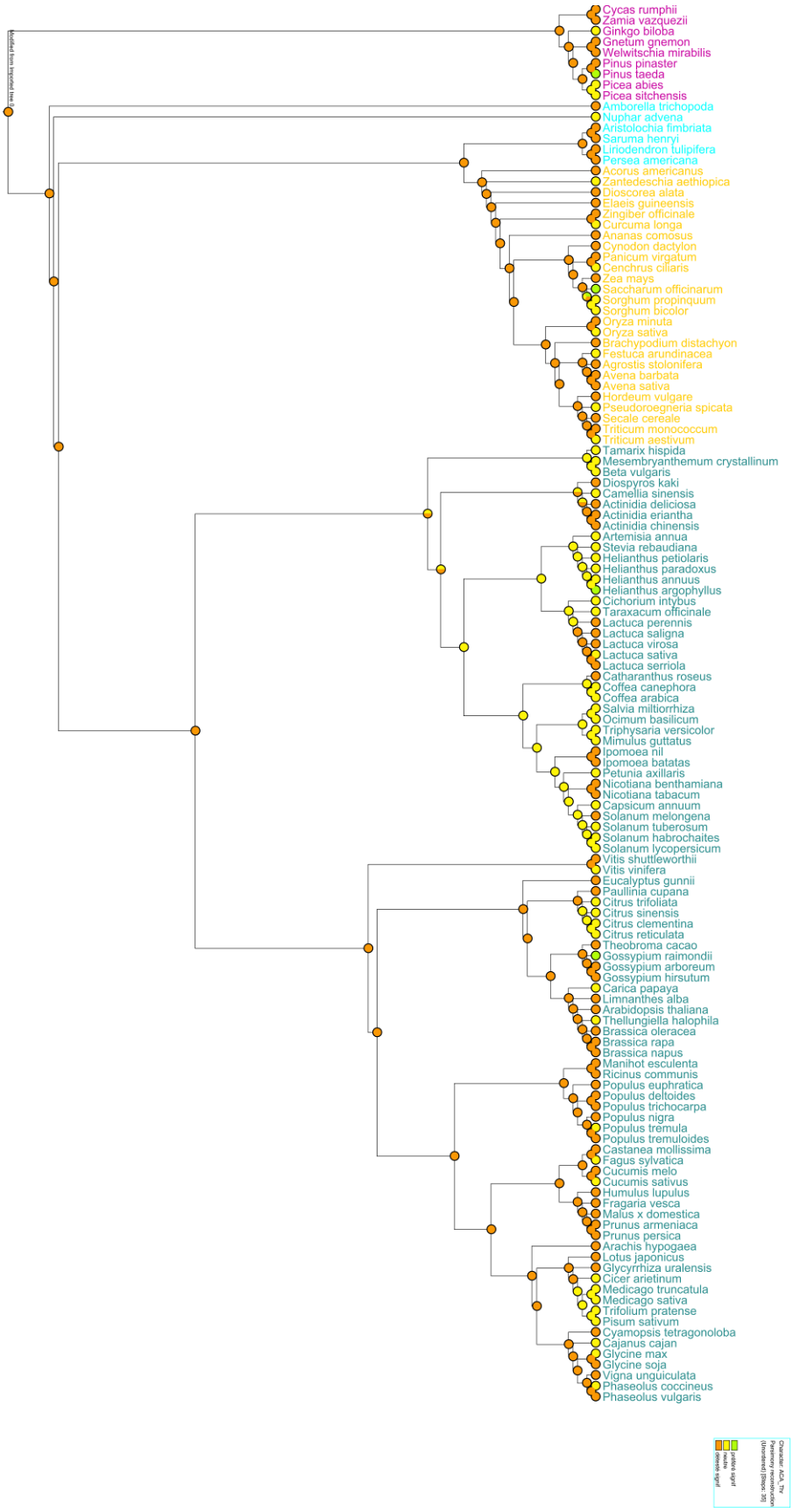
Thr ACT


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: ACT_Thr
Parsimony reconstruction
(Unordered) [Steps: 41]
ambiguous signif
multiple signif
absence signif

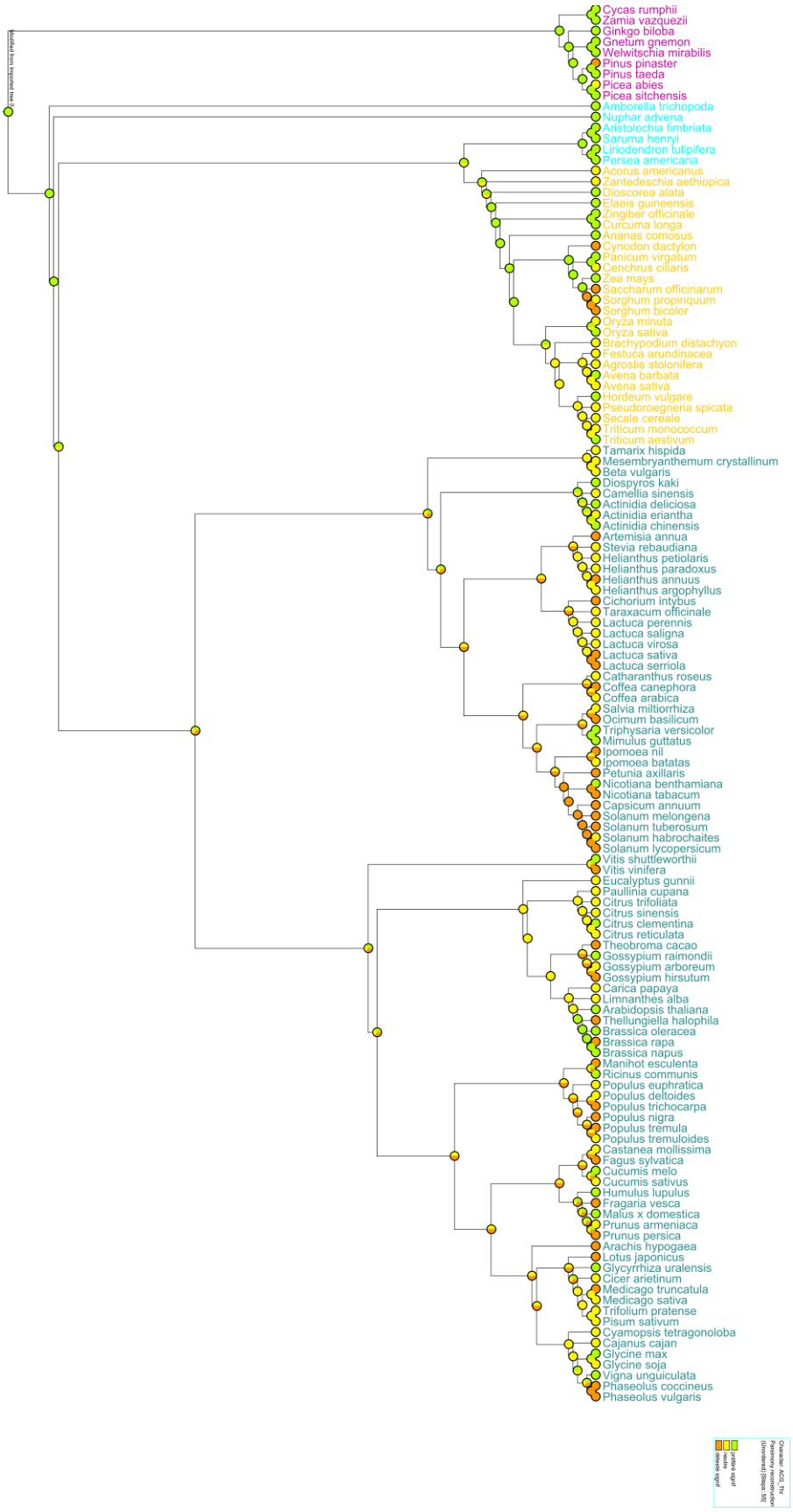
Thr ACC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: ACC_Thr
Parsimony reconstruction
(Unordered) [Steps: 20]
ambiguous signal
absence
presence signal

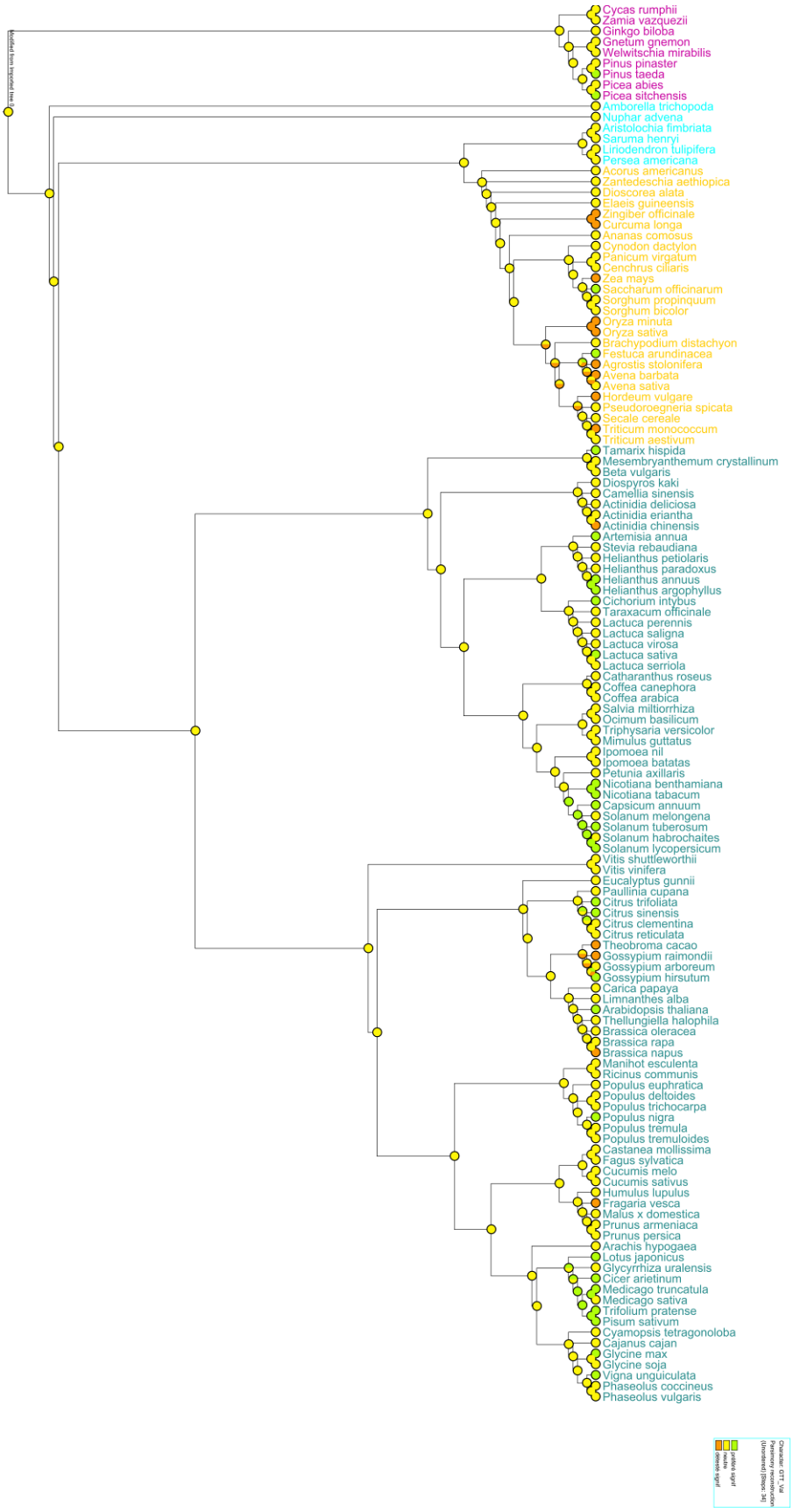
Thr ACA


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: ACA_Thr
Parsimony reconstruction
(Unordered) [Steps: 26]
ambiguous signal
inactive
absence signal

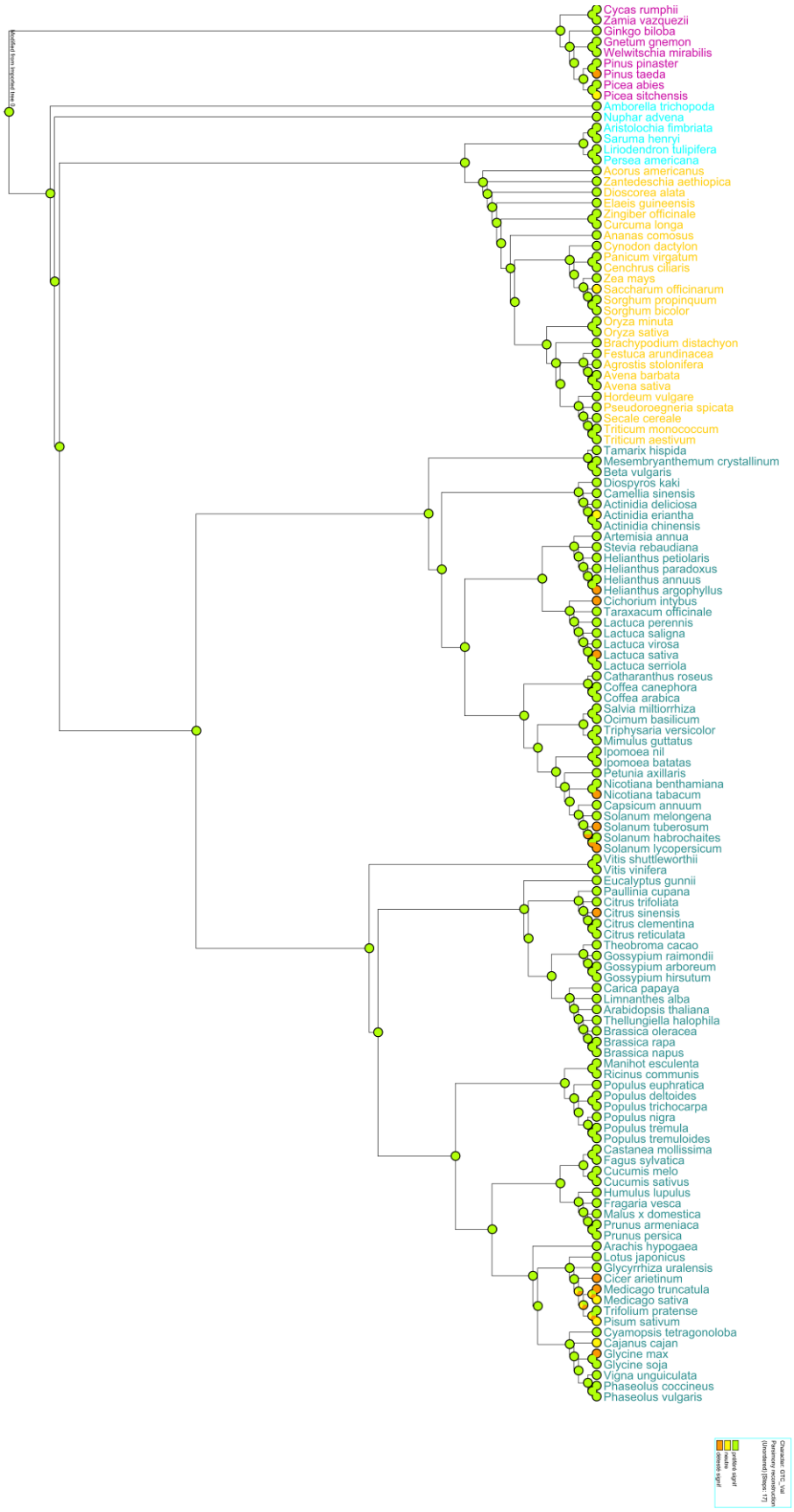
Thr ACG


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: ACG, Thr
Parsimony reconstruction
(Unordered) [Steps: 59]
ambiguous signal
no data
unique signal

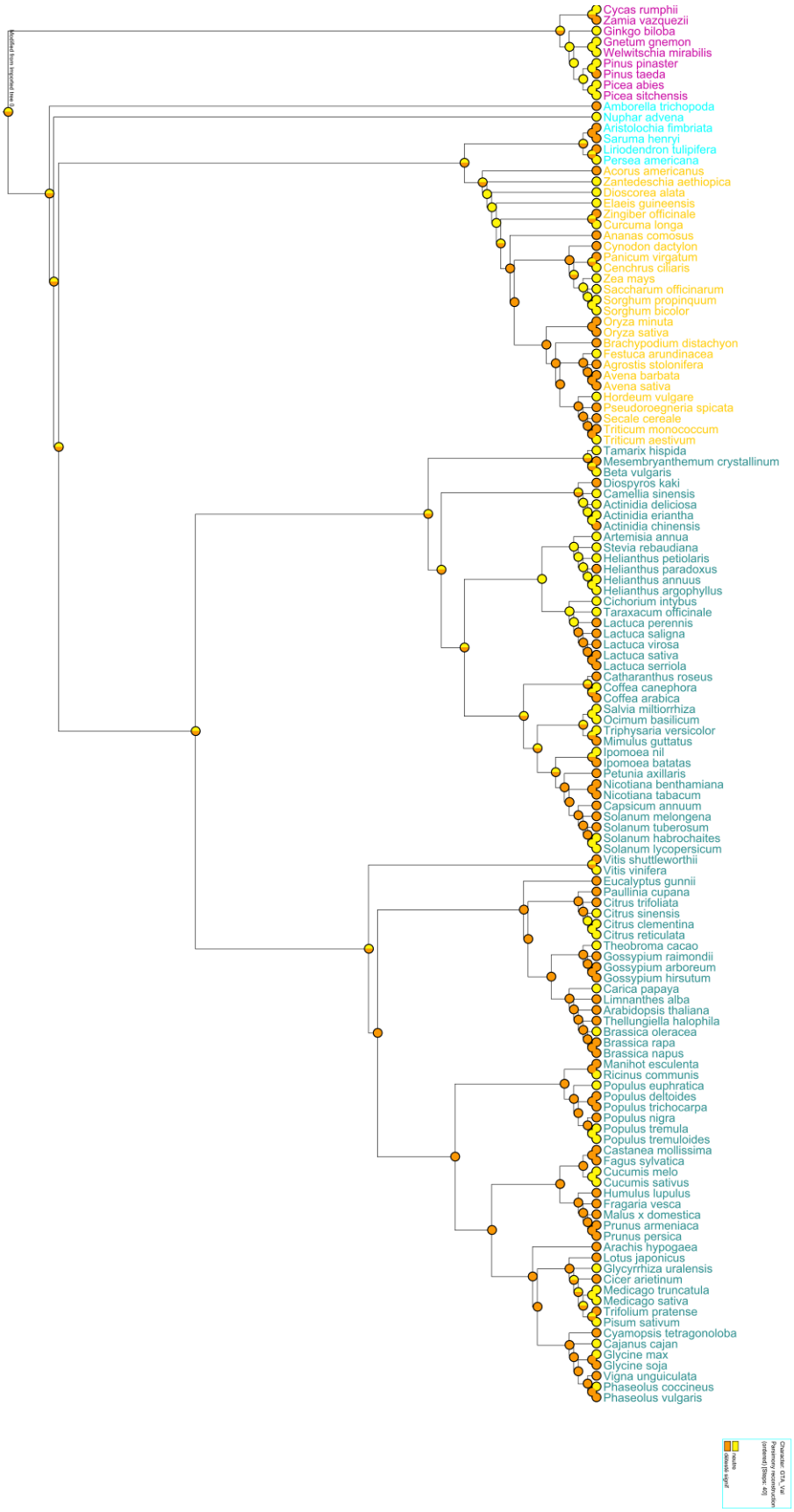
Val GTT


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GTT_Val
Parsimony reconstruction
(Unordered) [Steps: 34]
additive signal
ACCTRAN
DELTRAN signal

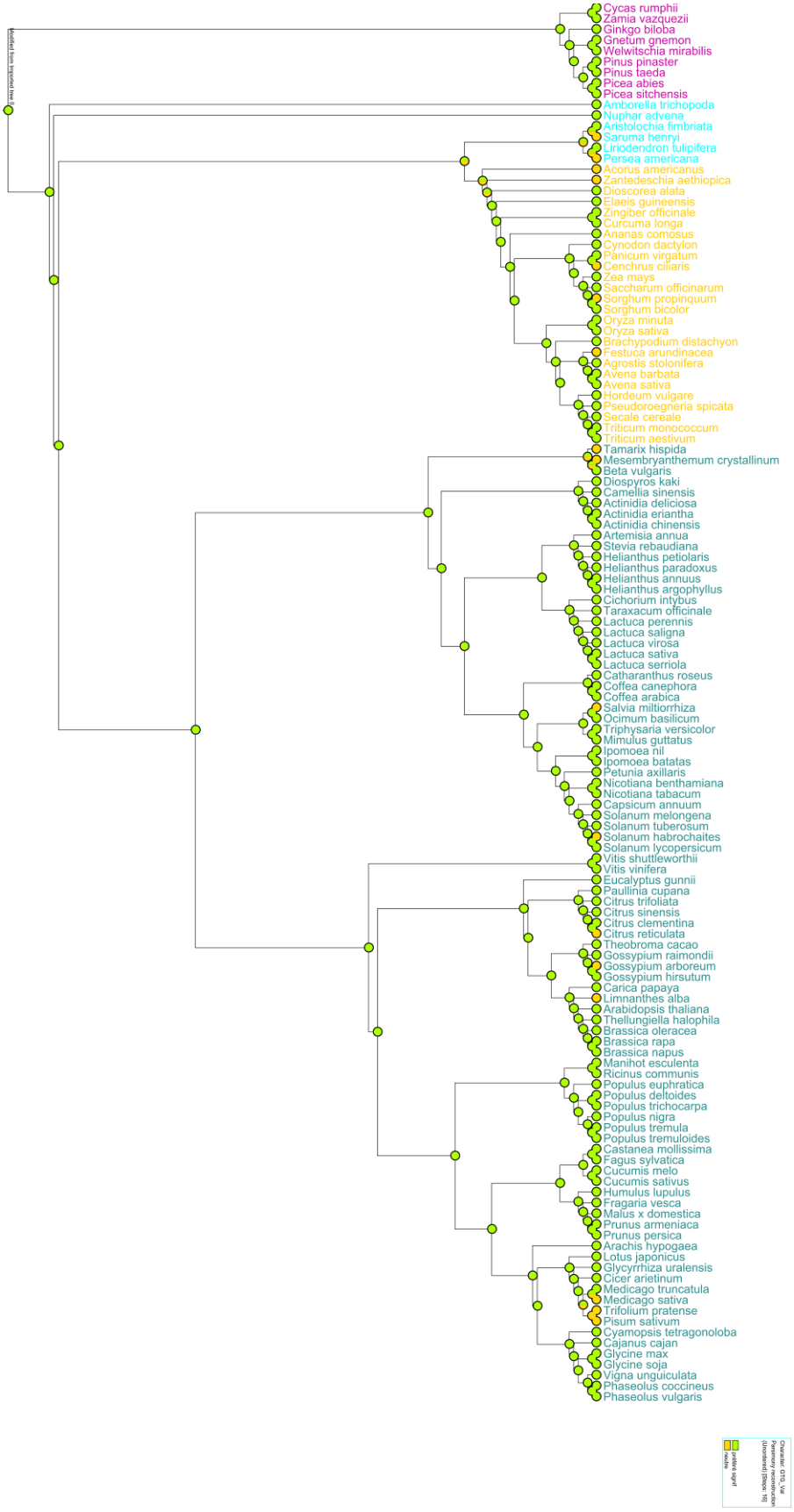
Val GTC


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GTC, Val
Parsimony reconstruction
(Unordered) [Steps: 17]
additive signif
ancillo
selective signif

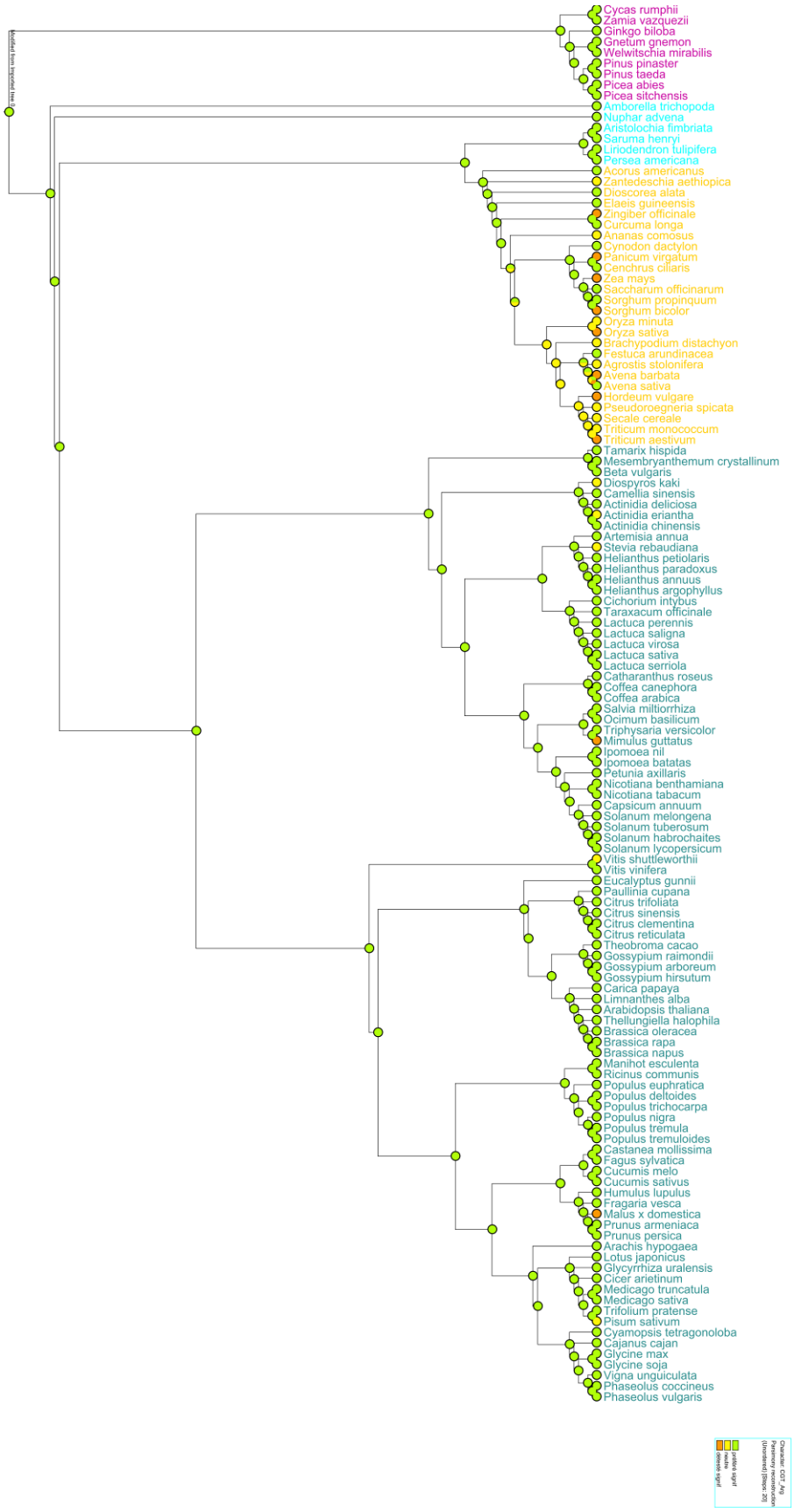
Val GTA


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GTA_Val
Parsimony reconstruction (ordered) [Steps: 43]
Absent
Present

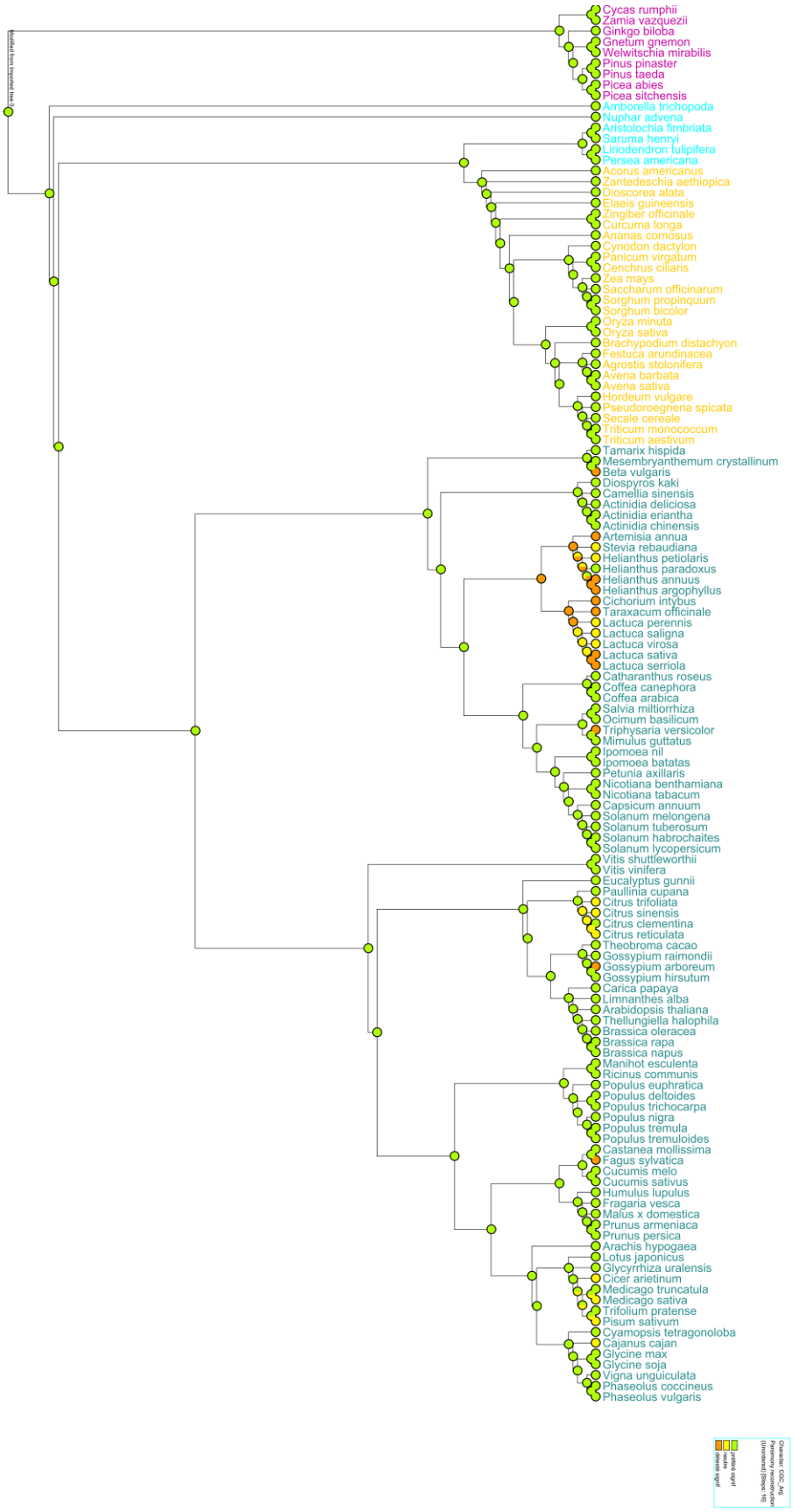
Val GTG


Modified from imported tree 1
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: GTG_Val
Parsimony reconstruction
(Unordered) [Steps: 16]
additive signif.
none

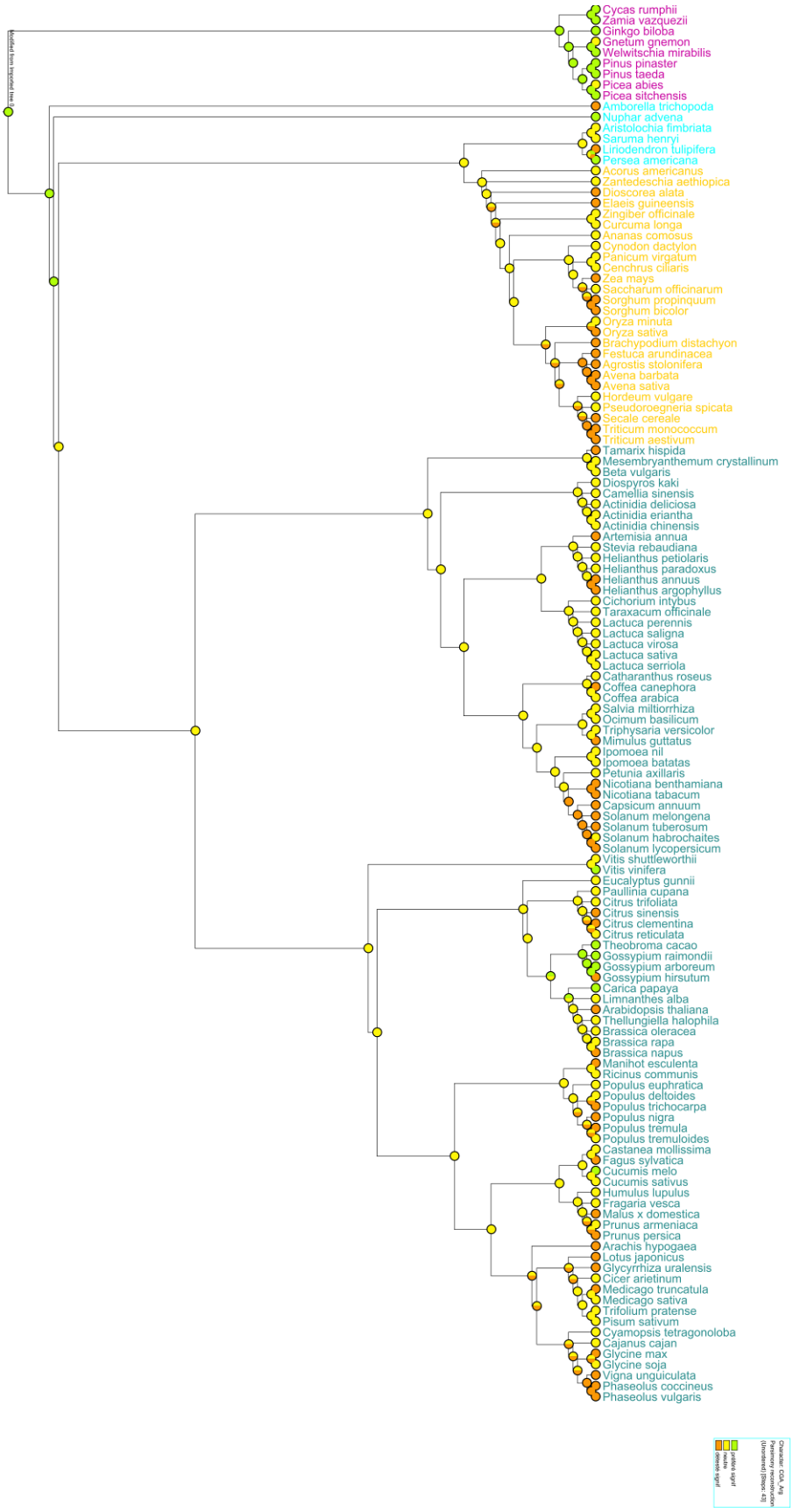
Arg CGT


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CGT_Arg
Parsimony reconstruction
(Unordered) [Steps: 20]
additive signal
ancillo
weasle signal

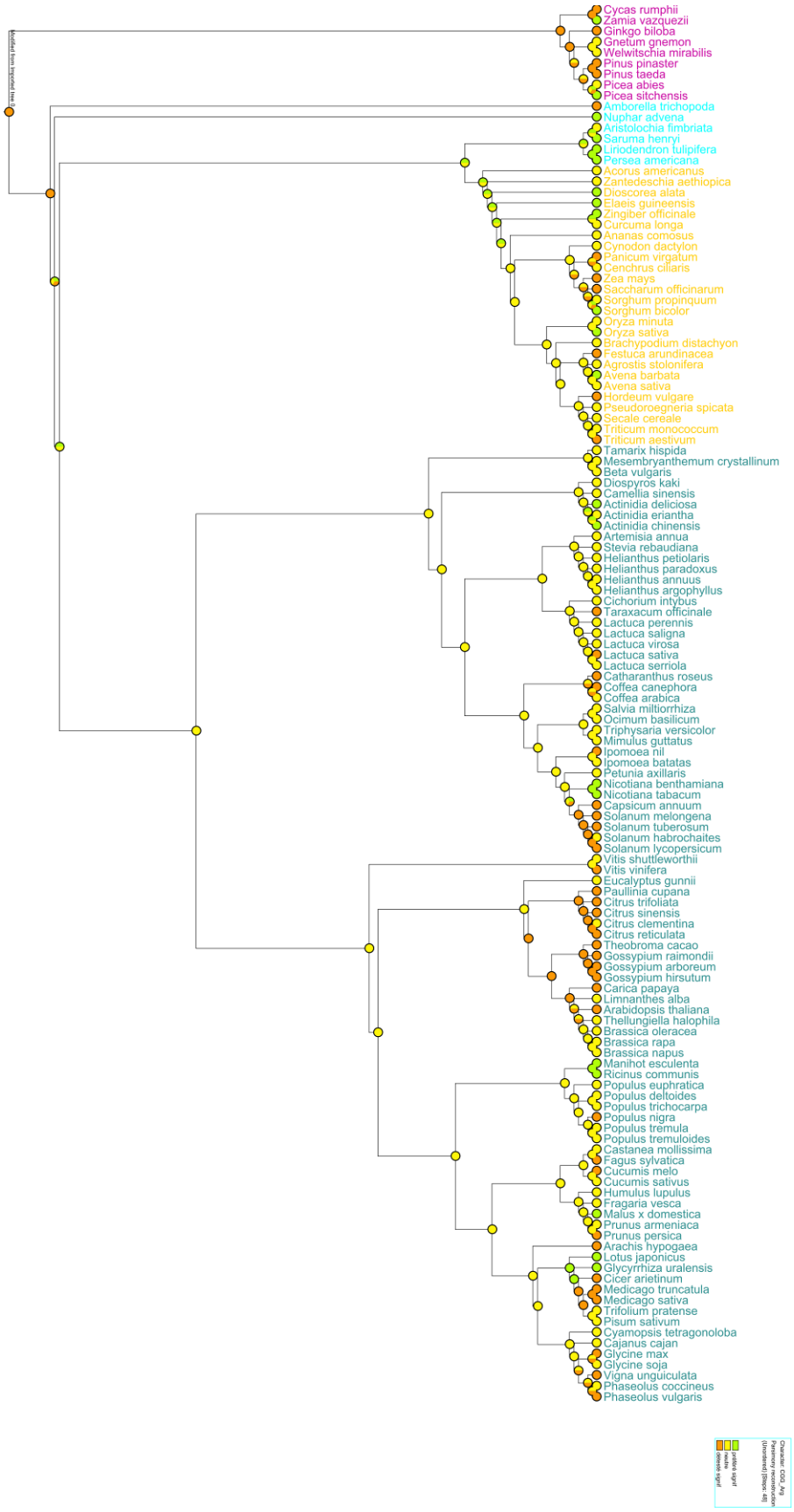
Arg CGC


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CGC_Arg
Parsimony reconstruction
(Unordered) [Steps: 16]
additive signal
ACCTRAN
DELTRAN signal

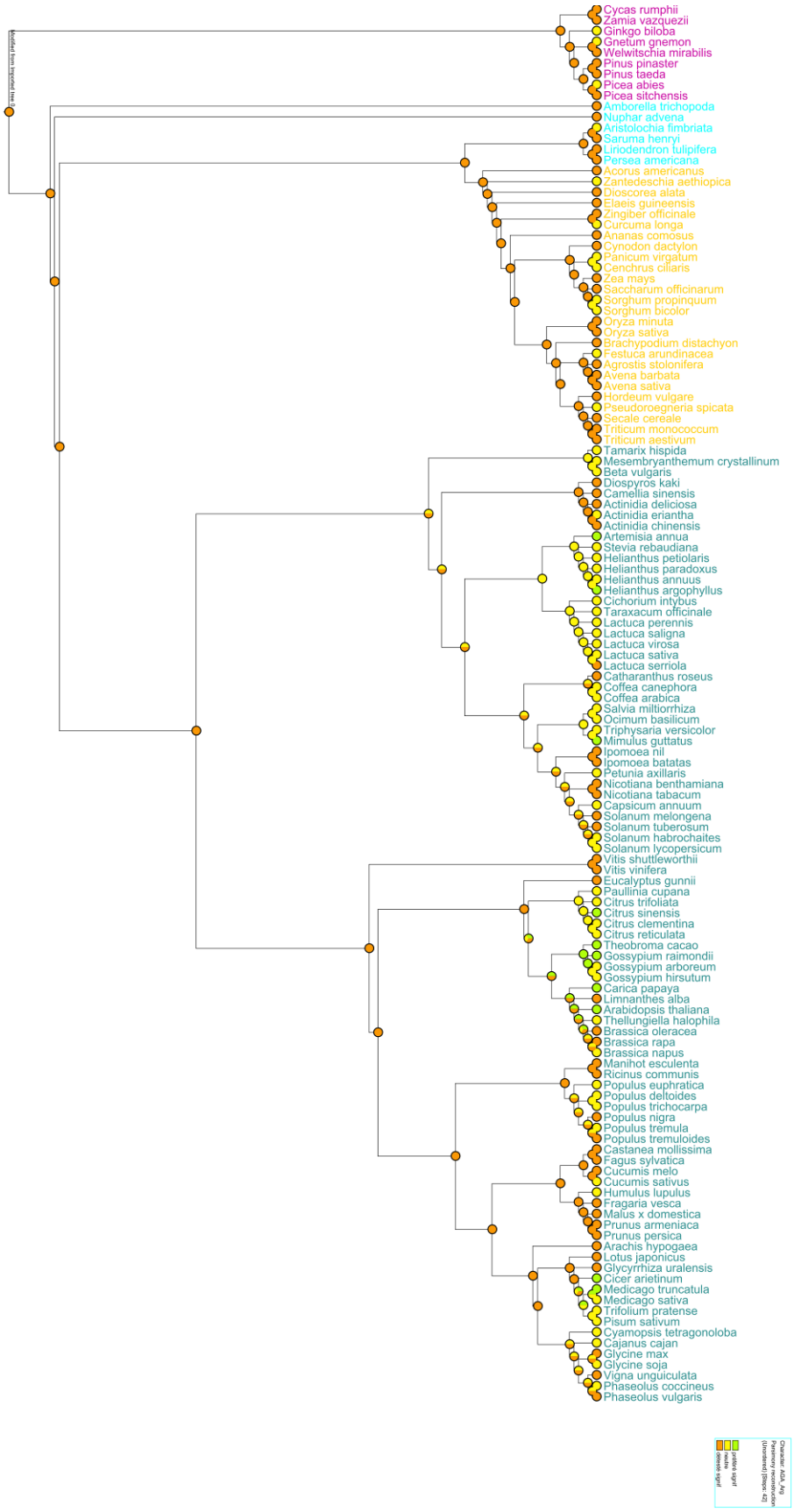
Arg CGA


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CGA_Arg
Parsimony reconstruction
(Unordered) [Steps: 43]
additive signal
ACCTRAN
DELTRAN signal

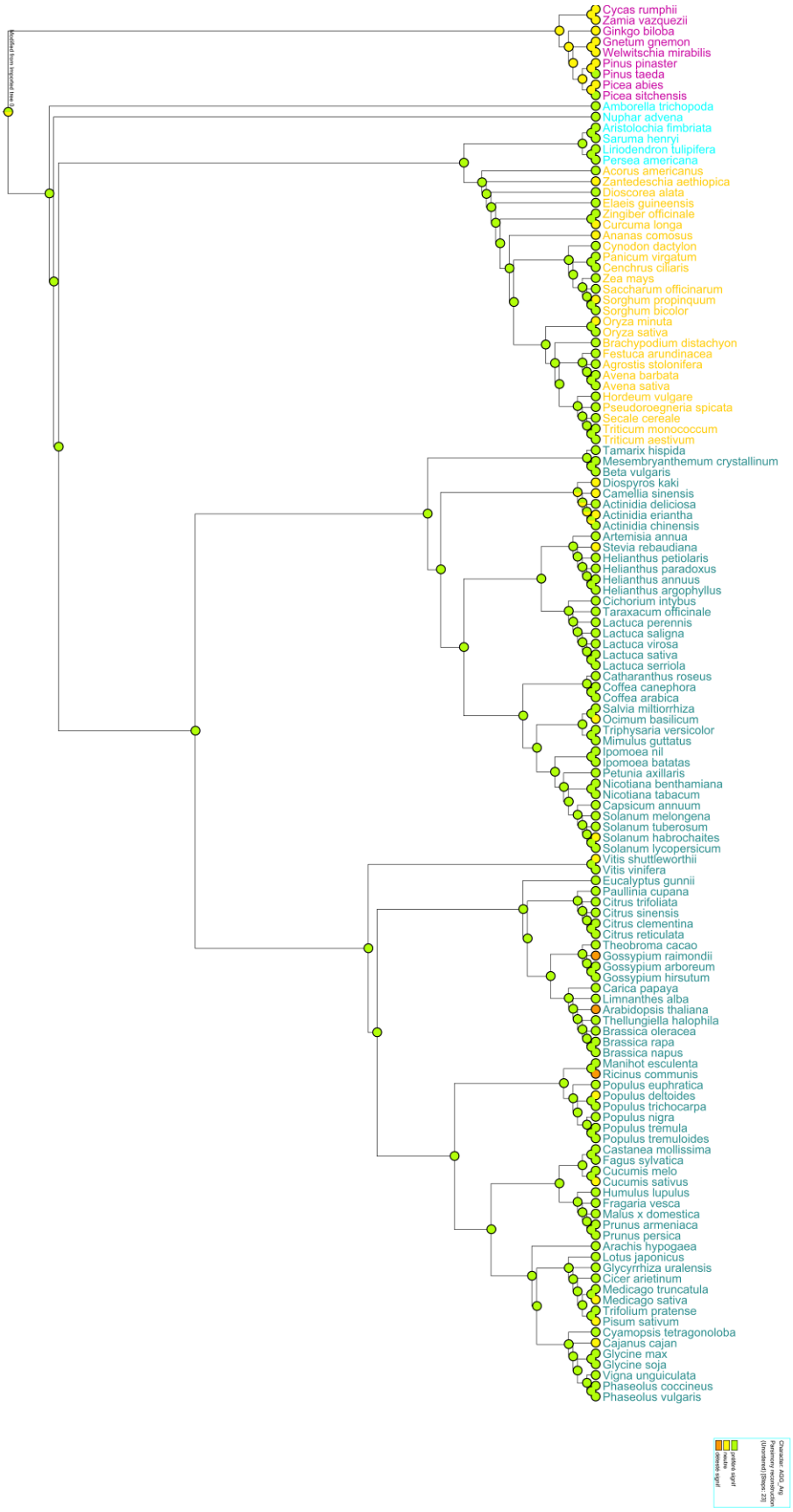
Arg CGG


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CGG_Arg
Parsimony reconstruction
(Unordered) [Steps: 48]
ambiguous signal
no data
unique signal

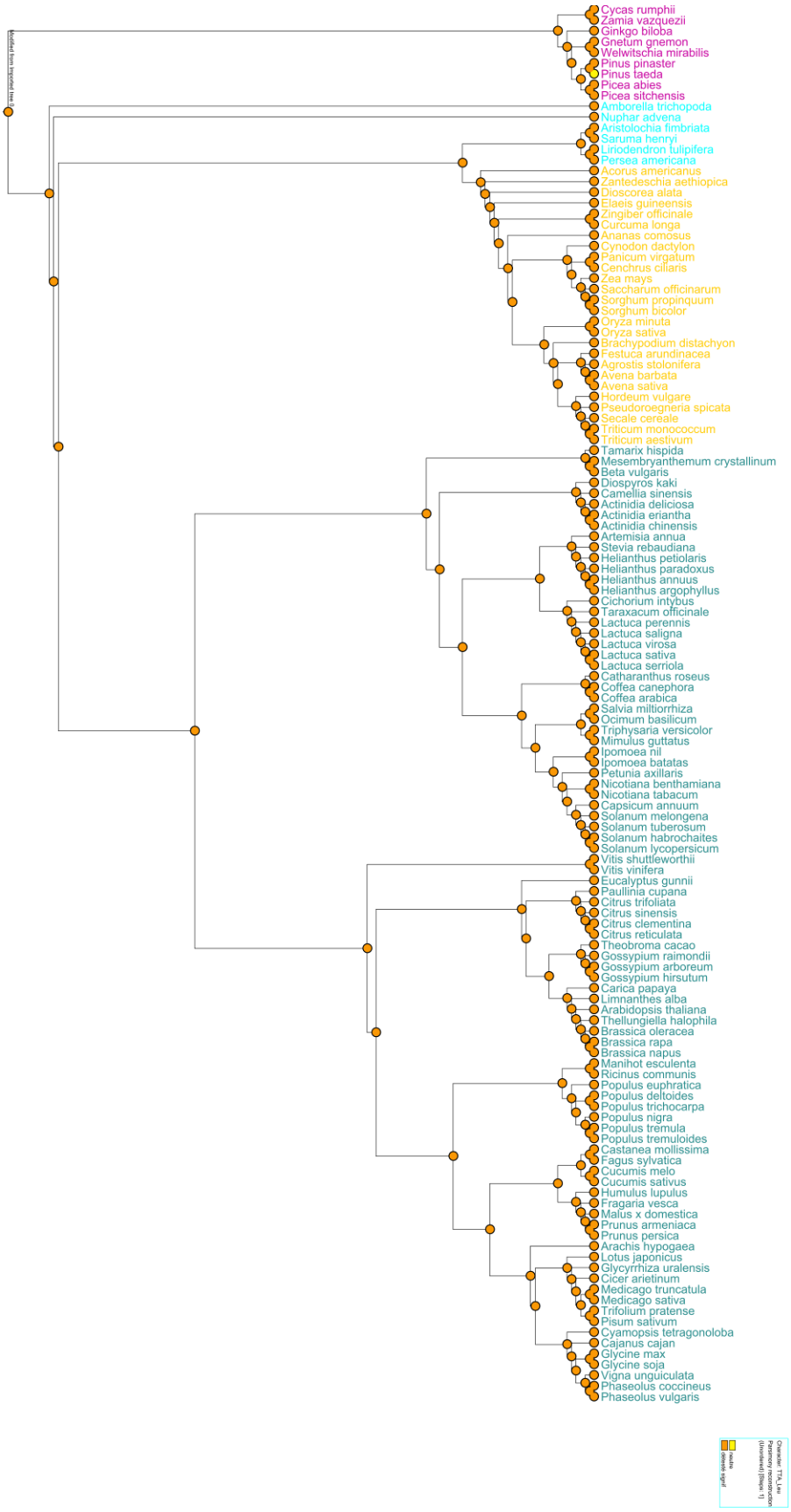
Arg AGA


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: AGA_Arg
Parsimony reconstruction
(Unordered) [Steps: 42]
additive signal
ancillo
weasle signal

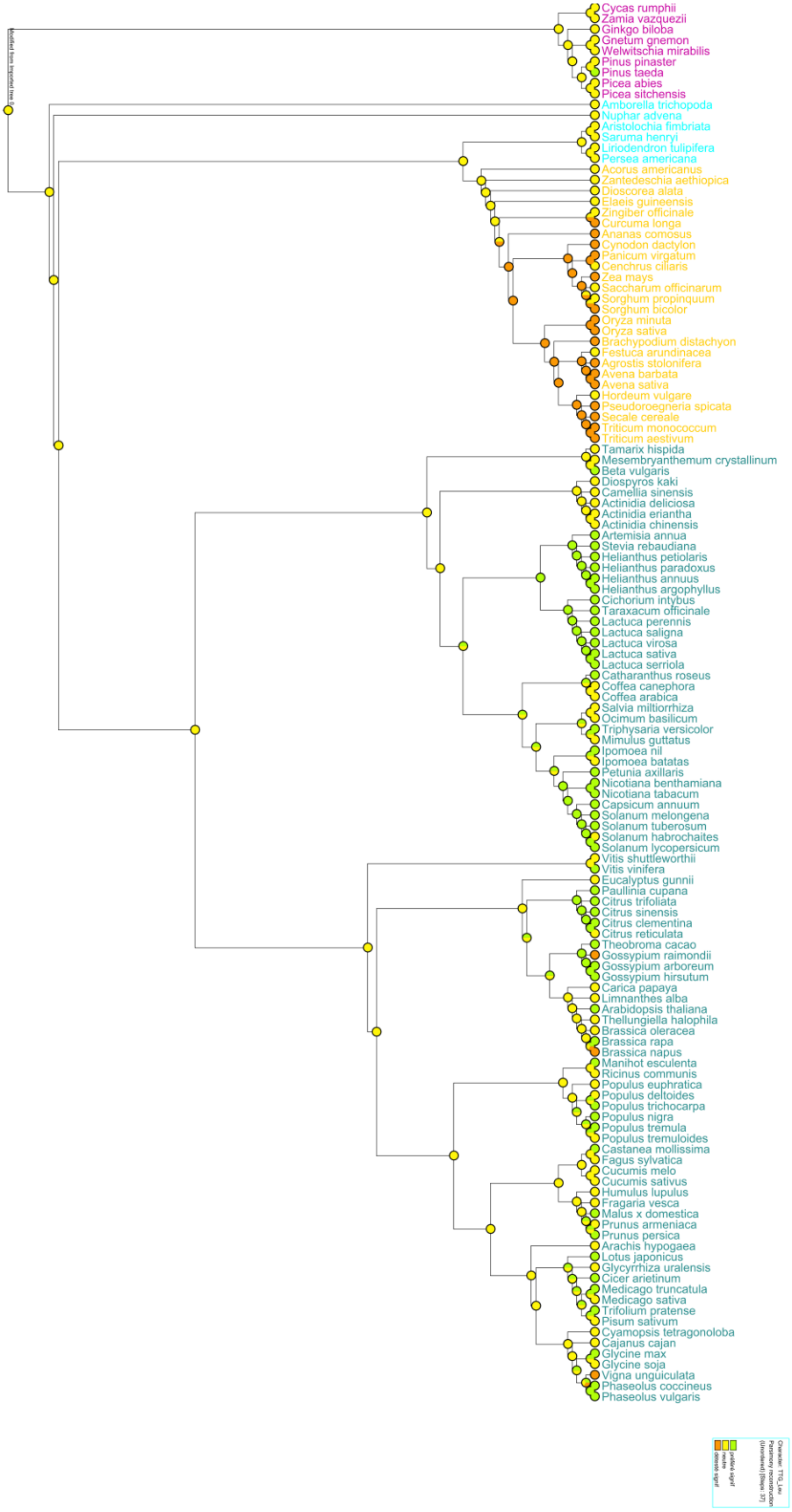
Arg AGG


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: AGG_Arg
Parsimony reconstruction
(Unordered) [Steps: 23]
additive signif
ancillo
ambiguous signif

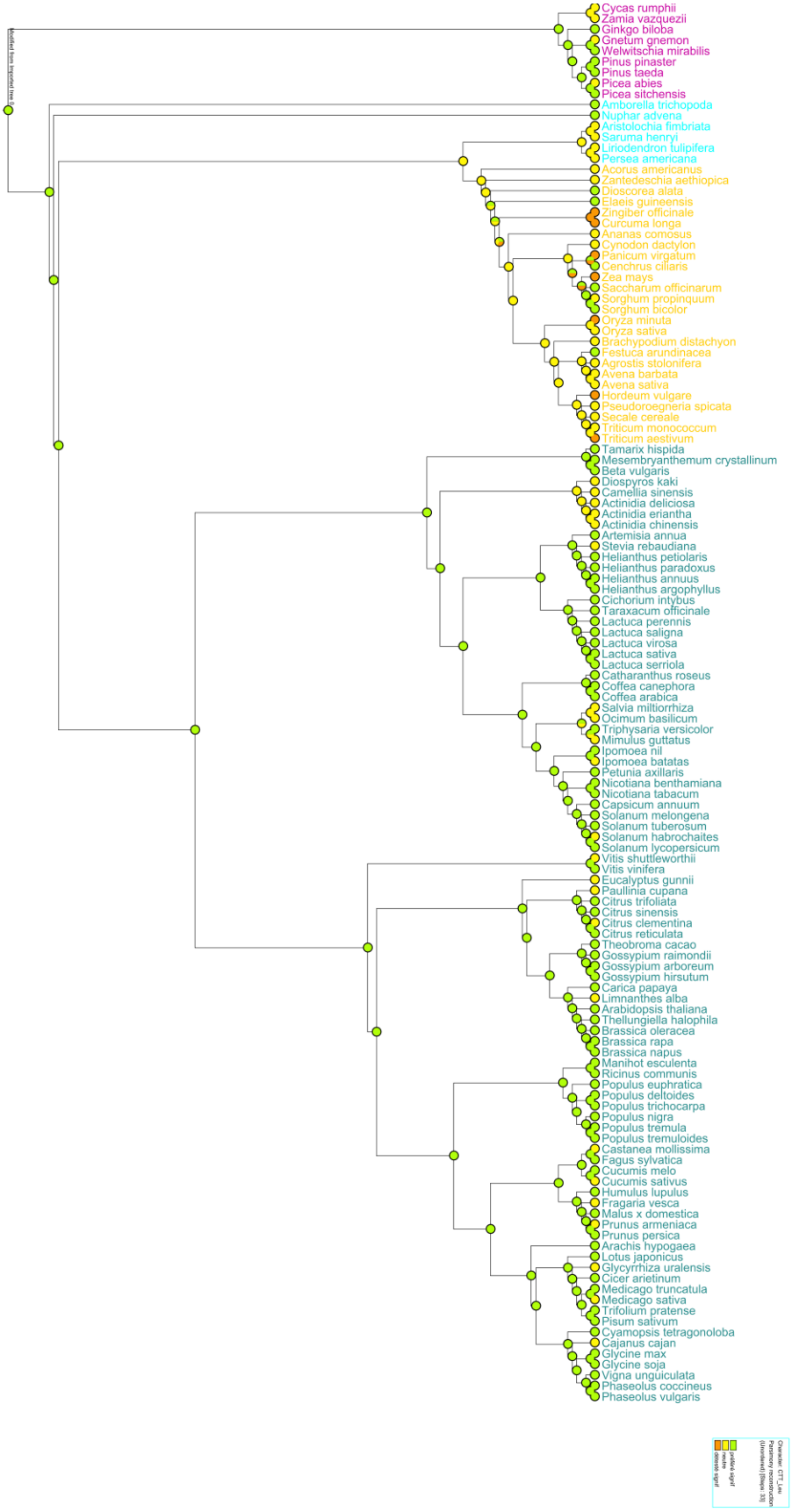
Leu TTA


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

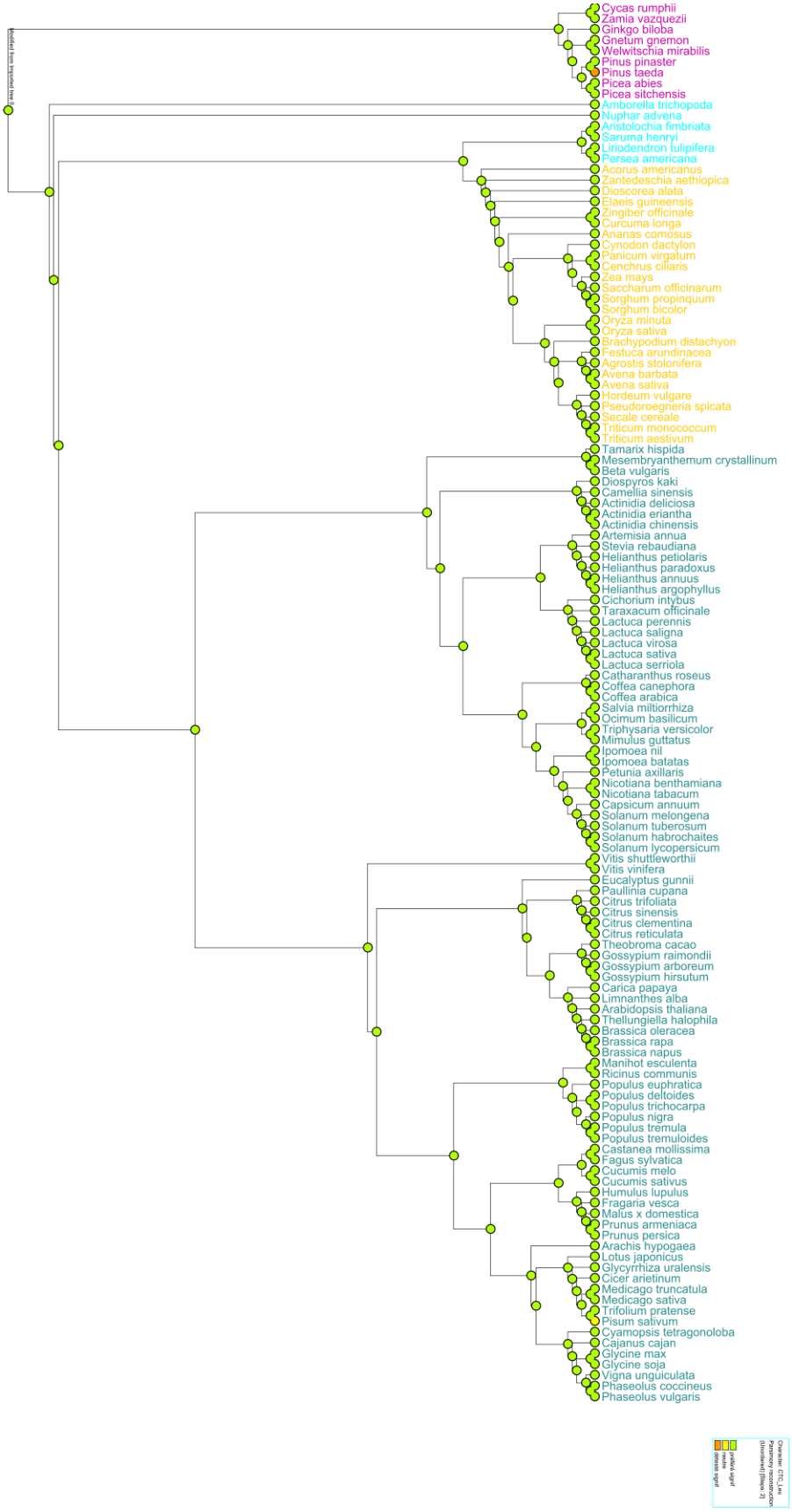
Leu TTG


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

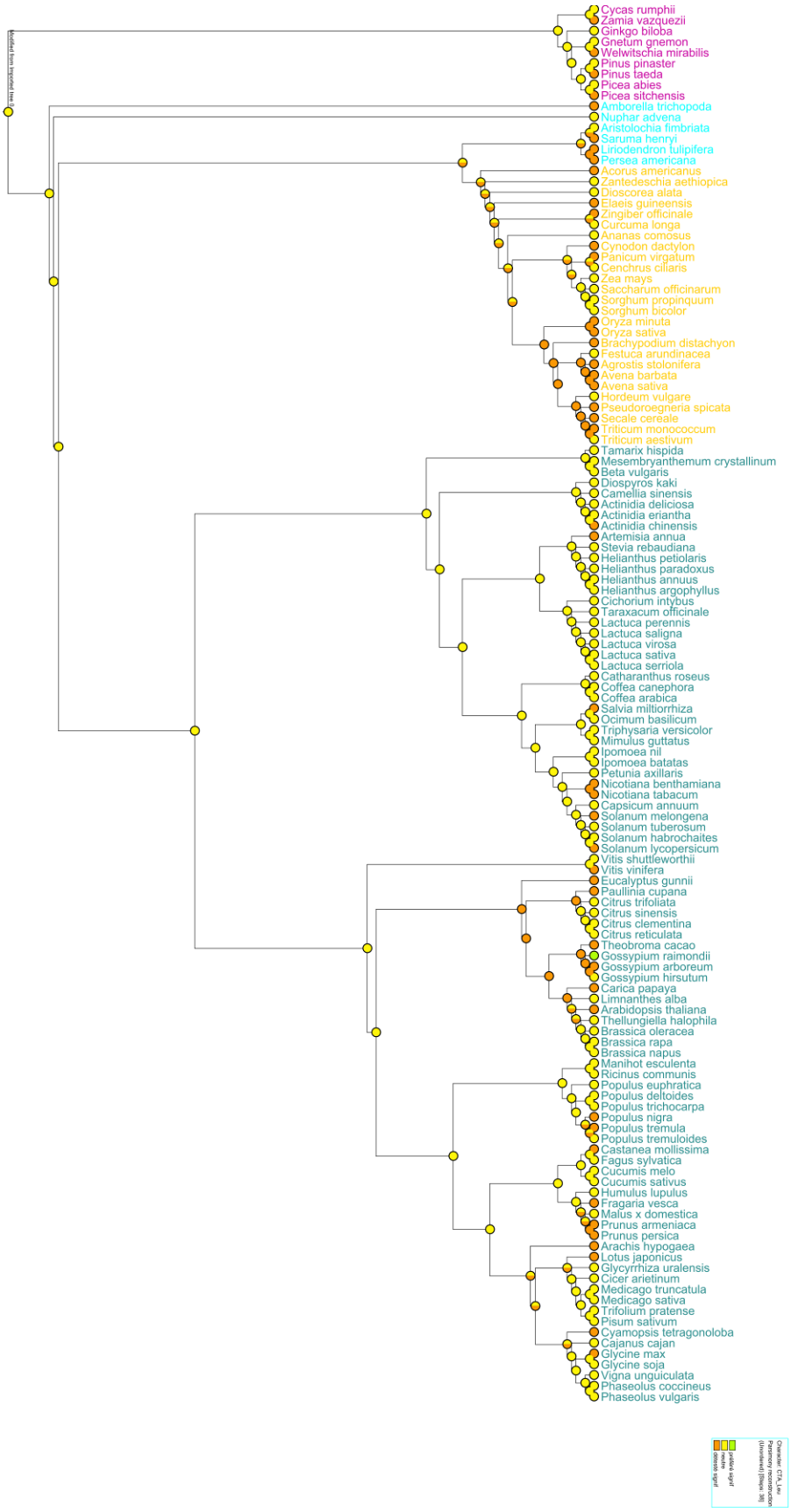
Leu CTT


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus peticolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

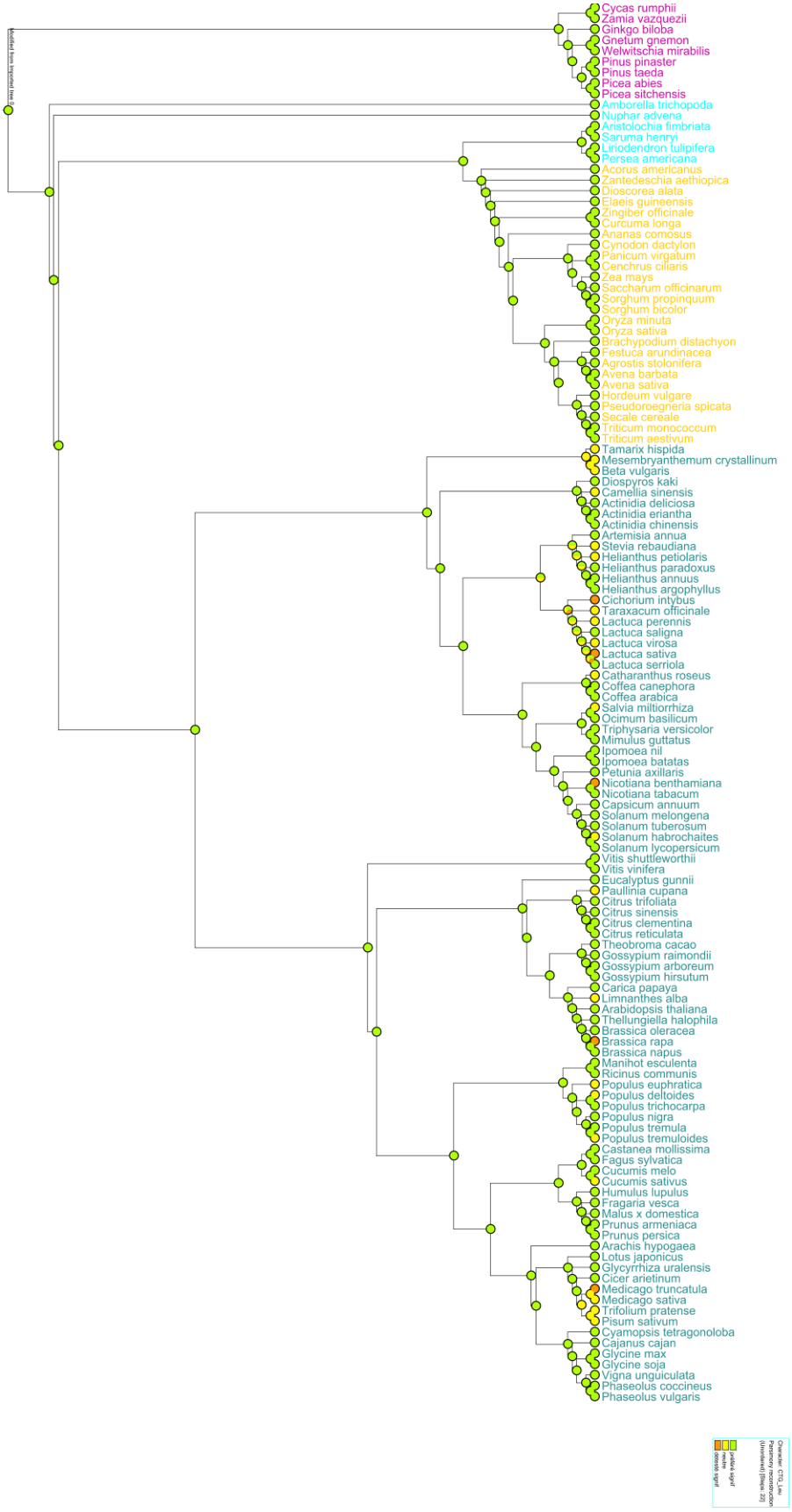
Leu CTC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus peticolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CTC_Leu
Parsimony reconstruction
(Unordered) [Steps: 2]
ambiguous
absence signif
no data
presence signif

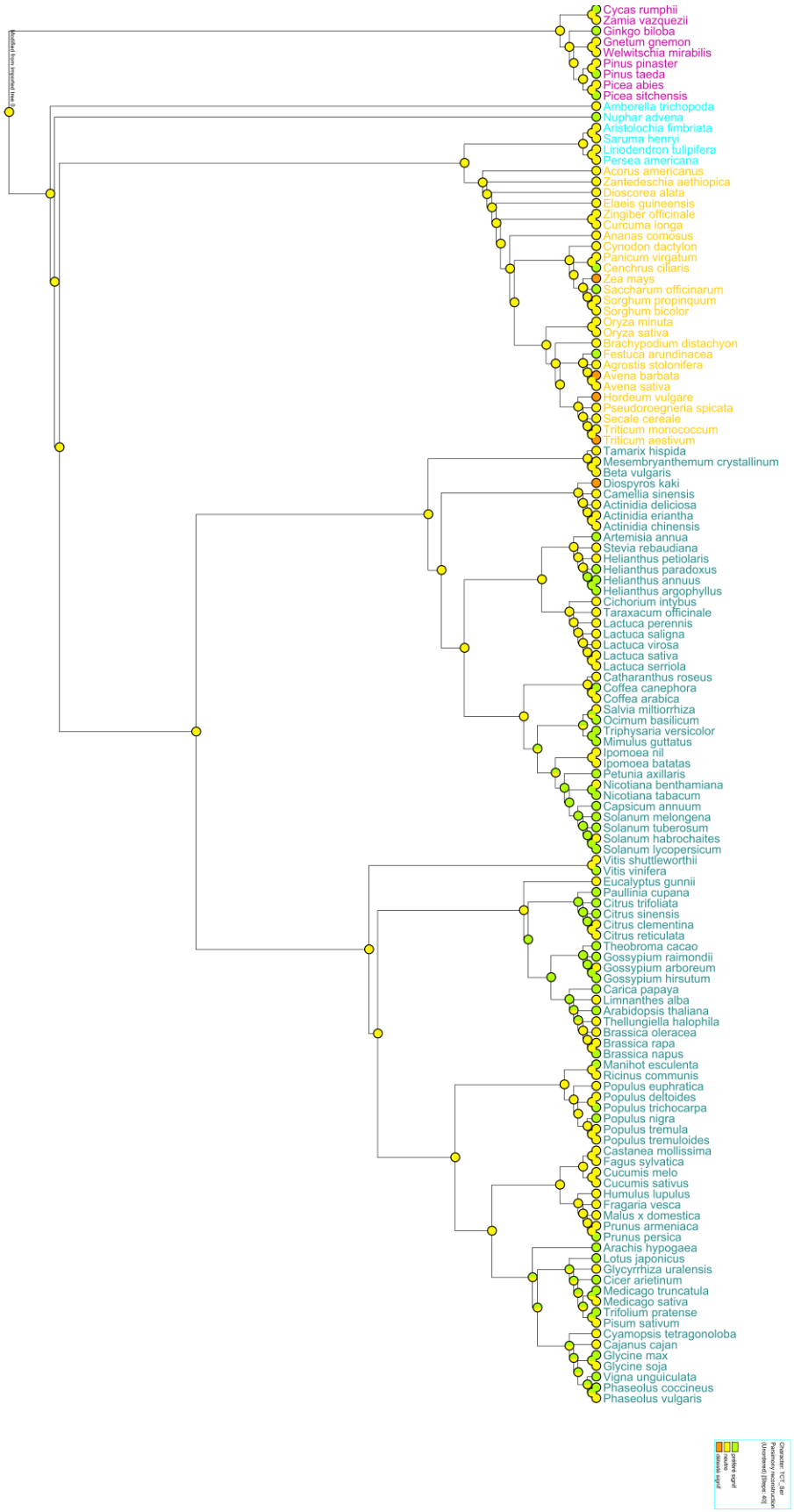
Leu CTA


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus peticolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

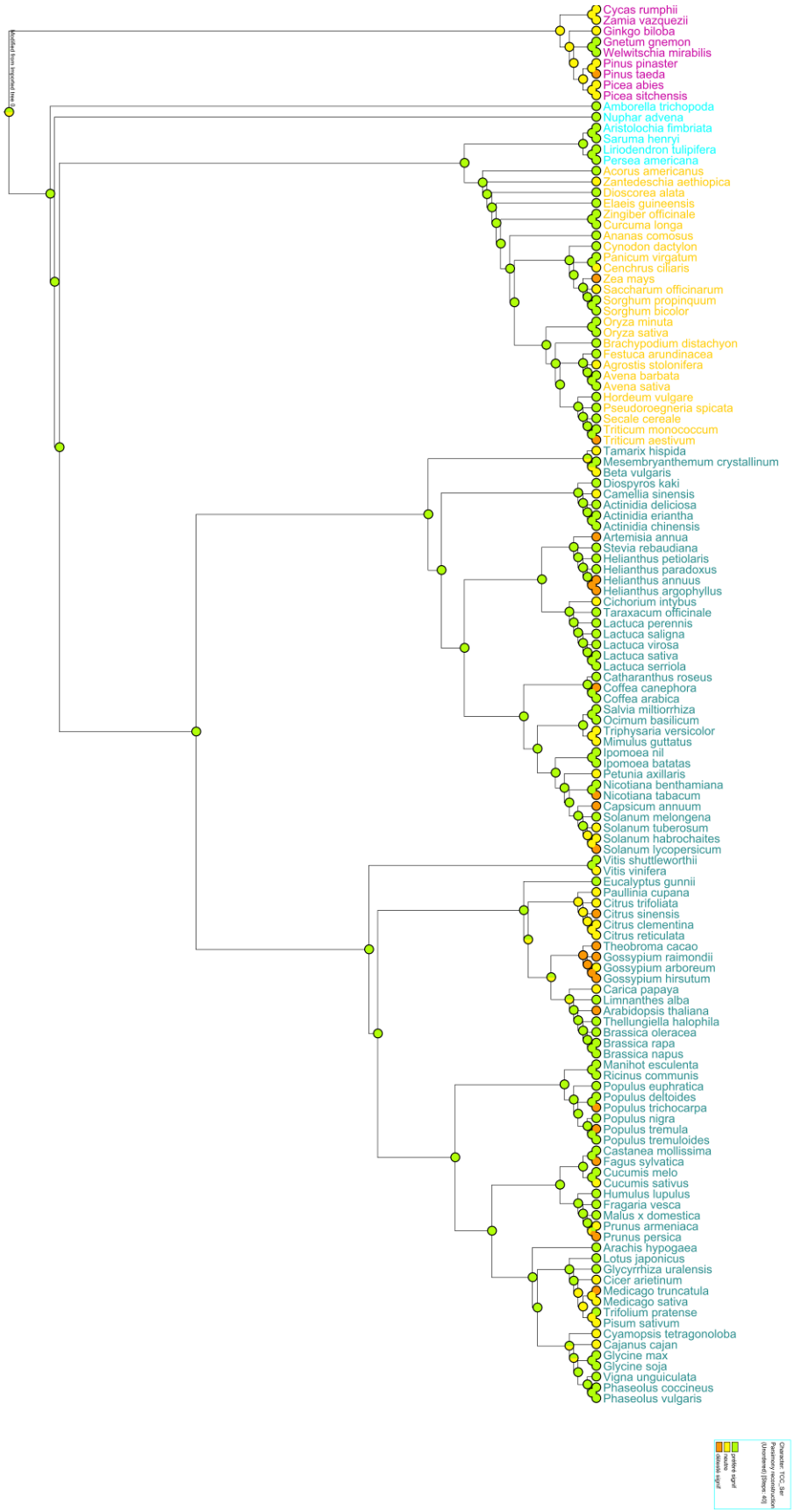
Leu CTG


Modified from Imported tree 1
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: CTG_Leu
Parsimony reconstruction
(Unordered) (Steps: 22)
ambiguous signal
no data
absence signal

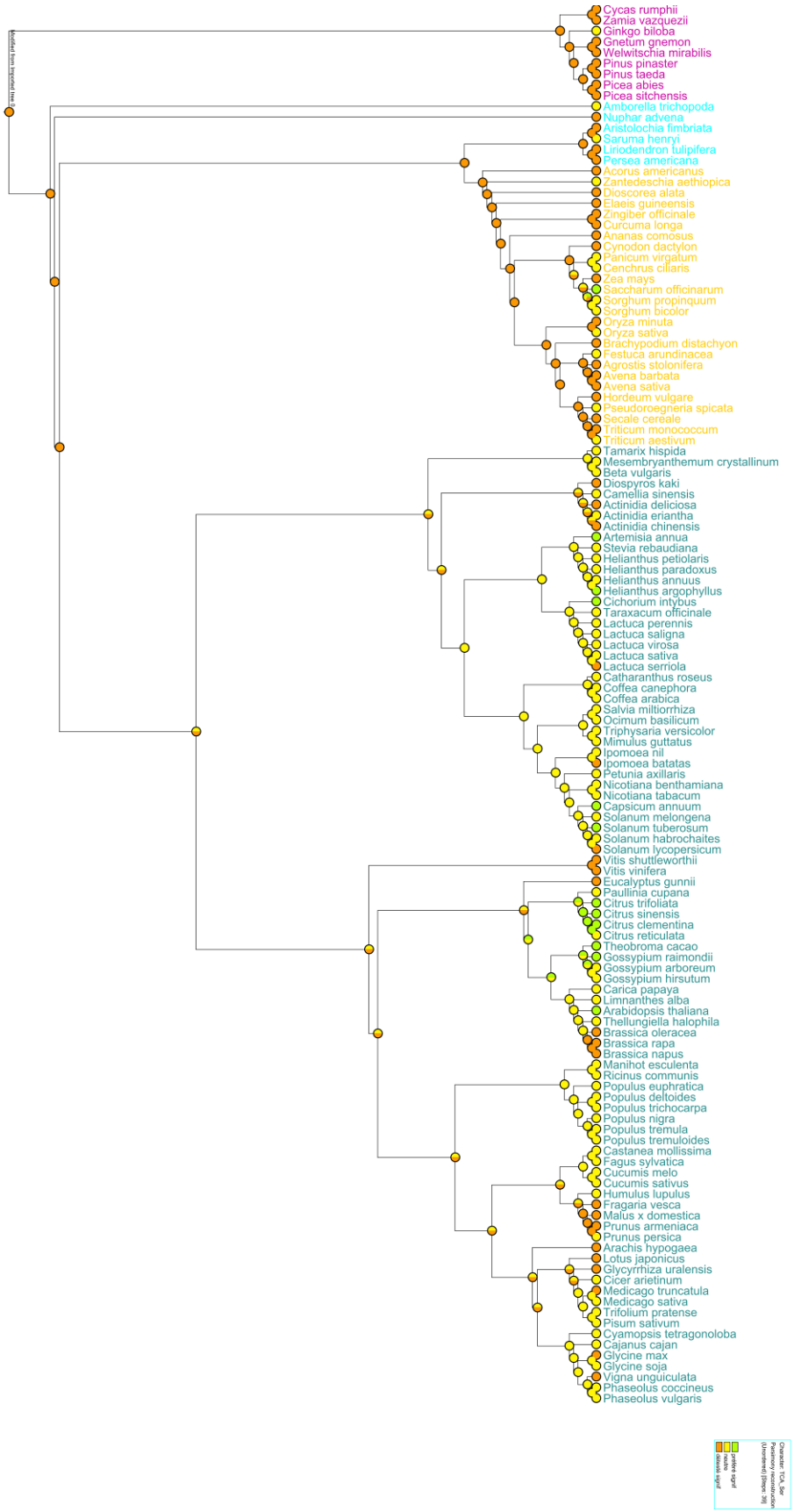
Ser TCT


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TCT_Ser
Parsimony reconstruction
(Unordered) [Steps: 40]
ambiguous
absence
presence

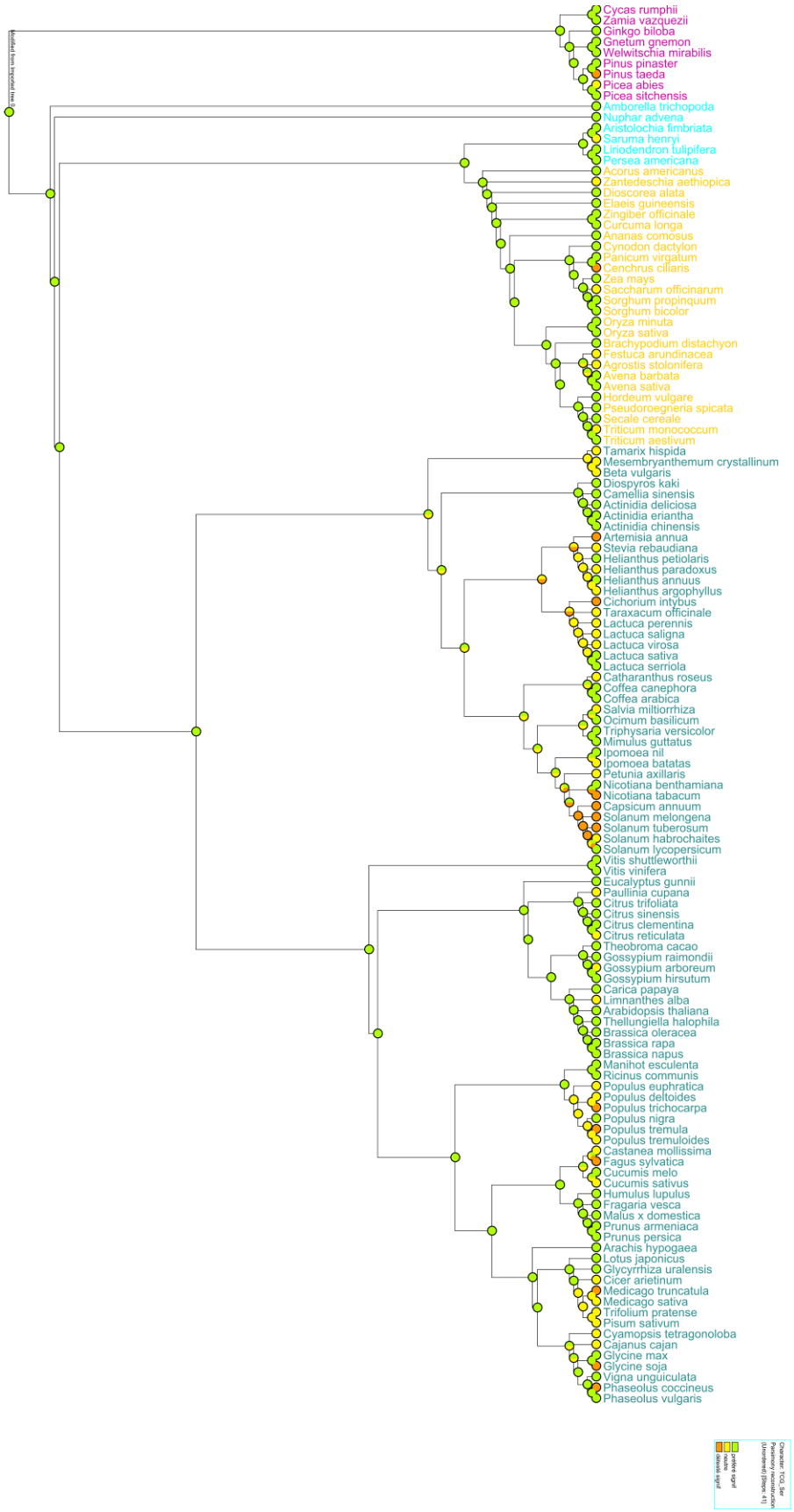
Ser TCC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TCC_Ser
Parsimony reconstruction
(Unordered) [Steps: 40]
absence apparent
equivocal
absence ambiguous

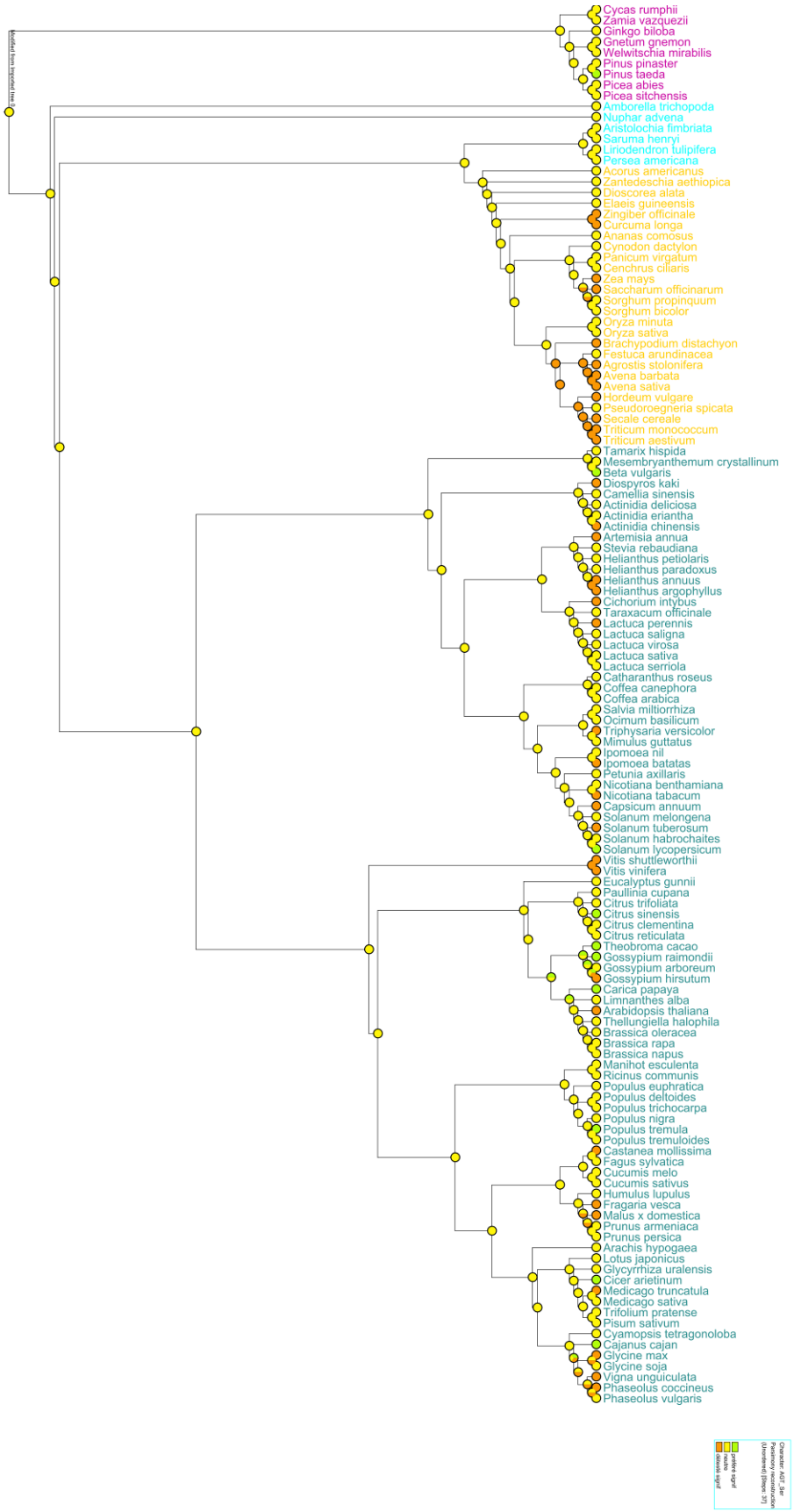
Ser TCA


Modified from imported tree 1
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TCA_Ser
Parsimony reconstruction
(Unordered) [Steps: 39]
ambiguous signal
absent
derived signal

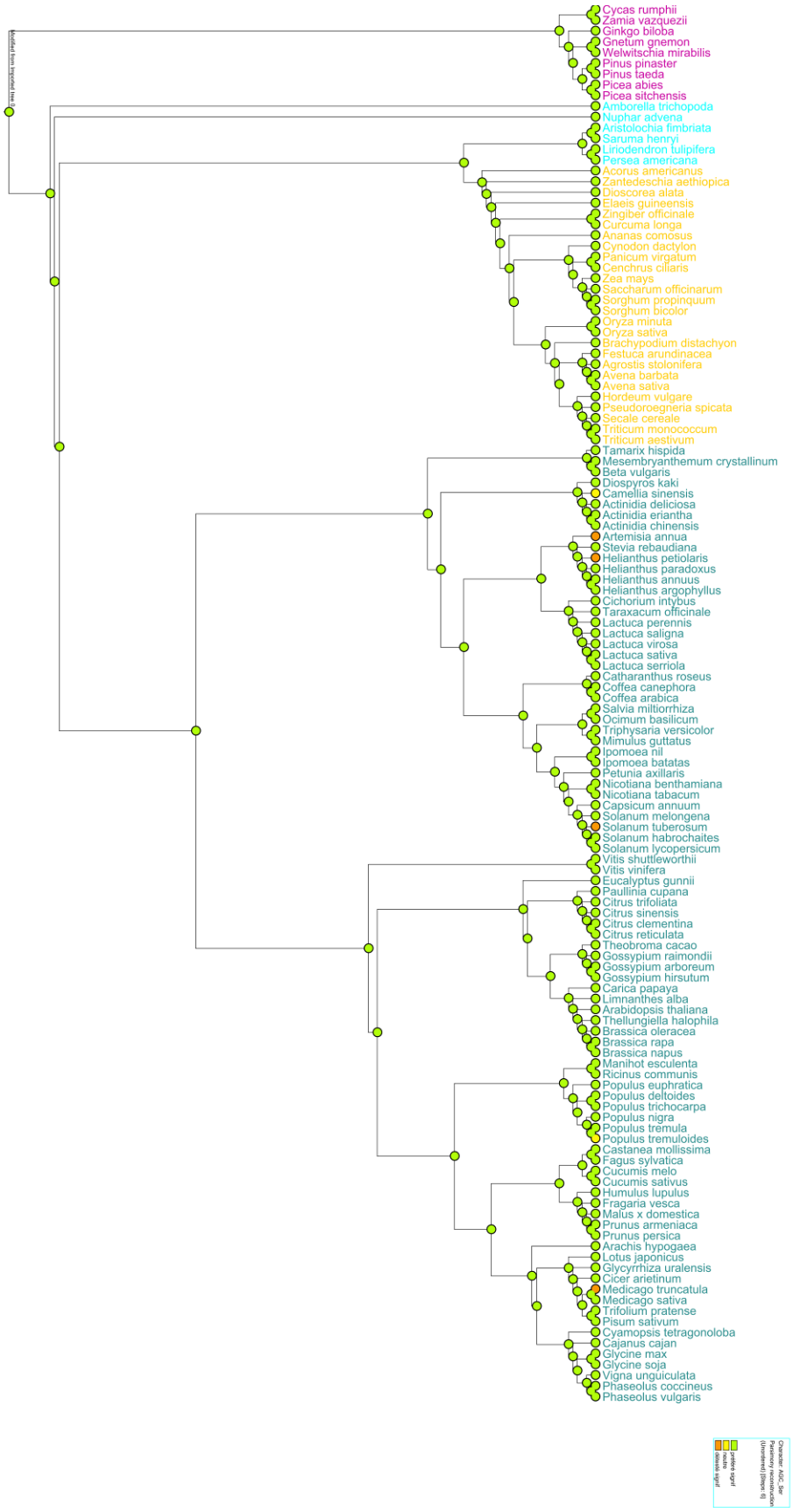
Ser TCG


Modified from imported tree 3
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: TCG_Ser
Parsimony reconstruction
(Unordered) [Steps: 41]
ambiguous
absent
present

Ser AGT


Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris

Ser AGC


Modified from imported tree 0
Cycas rumphii
Zamia vazquezii
Ginkgo biloba
Gnetum gnemon
Welwitschia mirabilis
Pinus pinaster
Pinus taeda
Picea abies
Picea sitchensis
Amborella trichopoda
Nuphar advena
Aristolochia fimbriata
Saruma henryi
Liriodendron tulipifera
Persea americana
Acorus americanus
Zantedeschia aethiopica
Dioscorea alata
Elaeis guineensis
Zingiber officinale
Curcuma longa
Ananas comosus
Cynodon dactylon
Panicum virgatum
Cenchrus ciliaris
Zea mays
Saccharum officinarum
Sorghum propinquum
Sorghum bicolor
Oryza minuta
Oryza sativa
Brachypodium distachyon
Festuca arundinacea
Agrostis stolonifera
Avena barbata
Avena sativa
Hordeum vulgare
Pseudoroegneria spicata
Secale cereale
Triticum monococcum
Triticum aestivum
Tamarix hispida
Mesembryanthemum crystallinum
Beta vulgaris
Diospyros kaki
Camellia sinensis
Actinidia deliciosa
Actinidia eriantha
Actinidia chinensis
Artemisia annua
Stevia rebaudiana
Helianthus petiolaris
Helianthus paradoxus
Helianthus annuus
Helianthus argophyllus
Cichorium intybus
Taraxacum officinale
Lactuca perennis
Lactuca saligna
Lactuca virosa
Lactuca sativa
Lactuca serriola
Catharanthus roseus
Coffea canephora
Coffea arabica
Salvia miltiorrhiza
Ocimum basilicum
Triphysaria versicolor
Mimulus guttatus
Ipomoea nil
Ipomoea batatas
Petunia axillaris
Nicotiana benthamiana
Nicotiana tabacum
Capsicum annuum
Solanum melongena
Solanum tuberosum
Solanum habrochaites
Solanum lycopersicum
Vitis shuttleworthii
Vitis vinifera
Eucalyptus gunnii
Paullinia cupana
Citrus trifoliata
Citrus sinensis
Citrus clementina
Citrus reticulata
Theobroma cacao
Gossypium raimondii
Gossypium arboreum
Gossypium hirsutum
Carica papaya
Limnanthes alba
Arabidopsis thaliana
Thellungiella halophila
Brassica oleracea
Brassica rapa
Brassica napus
Manihot esculenta
Ricinus communis
Populus euphratica
Populus deltoides
Populus trichocarpa
Populus nigra
Populus tremula
Populus tremuloides
Castanea mollissima
Fagus sylvatica
Cucumis melo
Cucumis sativus
Humulus lupulus
Fragaria vesca
Malus x domestica
Prunus armeniaca
Prunus persica
Arachis hypogaea
Lotus japonicus
Glycyrrhiza uralensis
Cicer arietinum
Medicago truncatula
Medicago sativa
Trifolium pratense
Pisum sativum
Cyamopsis tetragonoloba
Cajanus cajan
Glycine max
Glycine soja
Vigna unguiculata
Phaseolus coccineus
Phaseolus vulgaris
Character: AGC_Ser
Parsimony reconstruction
(Unordered) [Steps: 6]
ambiguous signal
absent signal

Diversity and Evolution of Nucleotide Landscapes in Plants

The nucleotide landscape – the way base composition varies along a genome – is a striking feature of genome organization and is highly variable between species. The evolutionary causes of such heterogeneity in GC content have been much debated. Biased gene conversion towards G and C (BGC) and selection on codon usage (SCU) are thought to be main forces. BGC is a neutral process associated with recombination favouring G and C alleles over A and T ones. SCU is a selection process favouring the so-called "preferred" codons, i.e., those whose translation is the most efficient. Contrary to vertebrates, plant nucleotide landscapes are still poorly known. Most studies focused on the GC-poor and homogeneous *Arabidopsis thaliana* genome and on the GC-rich and heterogeneous rice genome. The contrast between these two genomes was often generalized as a dicot/monocot dichotomy but this vision is clearly phylogenetically biased.

The objectives of this study are to characterize angiosperm nucleotide landscapes on a wide phylogenetic scale and to better understand the evolutionary mechanisms acting upon the evolution of nucleotide landscapes. To what extent do nucleotide landscapes vary across angiosperm phylogeny? Are nucleotide landscapes shaped by BGC and SCU? Are taxa affected with the same intensity?

To tackle these issues, I used a comparative genomic approach relying on EST data analysis on over 230 angiosperm and gymnosperm species. Through the nucleotide landscape survey for such a wide range of species I found a continuum of GC-heterogeneity patterns across phylogeny, some taxa such as Poaceae being strikingly GC-rich and heterogeneous. My results suggest that nucleotide landscapes could have been shaped by BGC and, to a lesser extent, by SCU. GC-content enrichment and impoverishment are likely to have occurred several times independently during plant evolution and could be explained by intensity variations of BGC and SCU. Using a proxy for EST expression level, I also characterized the diversity of preferred codons between species. Codon usage preferences were shown to be evolutionarily more unstable for four- and six-fold degenerate codon families. Finally, I could link the evolution of codon usage preferences to the evolution of genome base composition. My results suggest that genome base composition, partially shaped by BGC, seems to drive the coevolution between codon usage preferences and tRNAs.

Keywords: base composition, biased gene conversion towards G and C, selection on codon usage, codon usage preferences, Gymnosperms, Angiosperms, Expressed Sequence Tags.

Diversité et évolution des paysages nucléotidiques des plantes

Le paysage nucléotidique – la manière dont la composition nucléotidique varie le long du génome – est une caractéristique marquante de l'organisation des génomes et varie fortement entre espèces. Plusieurs hypothèses émergent des nombreux débats autour des mécanismes évolutifs à l'origine de ces hétérogénéités du taux de GC, parmi lesquelles la conversion génique biaisée vers G et C (BGC) et la sélection sur l'usage du code (SUC). La BGC est un processus neutre associé à la recombinaison qui favorise les allèles en G ou C au détriment des allèles en A ou T. La SUC est une force de sélection qui favorise les codons dits « préférés », ceux dont la traduction serait la plus efficace. Contrairement à ceux des vertébrés, les paysages nucléotidiques des plantes sont peu connus. La plupart des études ont été consacrées au génome d'*Arabidopsis thaliana*, pauvre en GC et homogène, et à celui du riz, riche en GC et hétérogène. Le contraste entre ces deux génomes a souvent été généralisé comme une dichotomie entre dicotylédones et monocotylédones, mais cette vision est clairement phylogénétiquement biaisée.

Les objectifs de ce travail de thèse sont de caractériser les paysages nucléotidiques des angiospermes à une large échelle phylogénétique et de mieux comprendre quels sont les mécanismes évolutifs jouant sur l'évolution de ces paysages nucléotidiques. Comment varient les paysages nucléotidiques le long de la phylogénie des angiospermes ? SUC et BGC façonnent-elles ces paysages nucléotidiques ? Les différents taxons sont-ils affectés avec la même intensité ?

Pour répondre à ces questions, j'ai utilisé une approche de génomique comparative basée sur l'analyse de données EST chez plus de 230 espèces d'angiospermes et de gymnospermes. L'exploration des paysages nucléotidiques de ce large éventail de plantes a montré que les patrons d'hétérogénéité des paysages nucléotidiques suivent un continuum le long de la phylogénie, avec des groupes particulièrement riches et hétérogènes en GC, les graminées par exemple. Mes résultats suggèrent que les paysages nucléotidiques des plantes pourraient avoir été façonnés par la BGC et, dans une moindre mesure, par la SUC. Des épisodes indépendants d'enrichissement et d'appauvrissement en GC ont vraisemblablement eu lieu au cours de l'évolution des plantes, et pourraient être expliqués par des variations d'intensité de ces mécanismes. En utilisant une prédiction du degré d'expression des EST, j'ai également mis en évidence une diversité dans les codons préférés entre espèces. Les préférences d'usage des codons se sont révélées plus labiles au cours de l'évolution pour les codons des acides aminés au code quatre et six fois dégénéré. J'ai pu lier l'évolution des préférences d'usage des codons à l'évolution de la composition nucléotidique des génomes. Mes résultats suggèrent que la composition en base des génomes, affectée en partie par la BGC, orienterait la coévolution entre préférence d'usage du code et ARNt.

Mots-clés : composition nucléotidique, conversion génique biaisée vers G et C, sélection sur l'usage du code, préférences d'usage du code, gymnospermes, angiospermes, *Expressed Sequence Tags*.