Genome-wide predictability of restriction sites across the eukaryotic tree of life

Santiago Herrera (tiagohe@gmail.com)[1,2] Paula H. Reyes-Herrera[3], Timothy M. Shank [1]

[1] Biology Department, Woods Hole Oceanographic Institution, 266 Woods Hole Road, Woods Hole, MA 02543, USA
[2] Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[3] Facultad Ingeniería Electrónica y Biomedical, Universidad Antonio Nariño, Carrera 3 Este # 47a -15, Bloque 4, Bogotá, Colombia.

Abstract

High-throughput sequencing of reduced representation libraries obtained through digestion with restriction enzymes – generally known as restriction-site associated DNA sequencing (RAD-seq) – is now one most commonly used strategies to generate single nucleotide polymorphism data in eukaryotes. The choice of restriction enzyme is critical for the design of any RAD-seq study as it determines the number of genetic markers that can be obtained for a given species, and ultimately the success of a project.

In this study we tested the hypothesis that genome composition, in terms of GC content, mono-, di- and trinucleotide compositions, can be used to predict the number of restriction sites for a given combination of restriction enzyme and genome. We performed systematic *in silico* genome-wide surveys of restriction sites across the eukaryotic tree of live and compared them with expectations generated from stochastic models based on genome compositions using the newly developed software pipeline PredRAD (https://github.com/phrh/PredRAD).

Our analyses reveal that in most cases the trinucleotide genome composition model is the best predictor, and the GC content and mononucleotide models are the worst predictors of the expected number of restriction sites in a eukaryotic genome. However, we argue that the predictability of restriction site frequencies in eukaryotic genomes needs to be treated in a case-specific basis, because the phylogenetic position of the taxon of interest and the specific recognition sequence of the selected restriction enzyme are the most determinant factors. The results from this study, and the software developed, will help guide the design of any study using RAD sequencing and related methods.

Introduction

The use of restriction enzymes to obtain reduced representation libraries from nuclear genomes, combined with the power of next-generation sequencing technologies, is rapidly becoming one of the most commonly used strategies to generate single nucleotide polymorphism (SNP) data in both model and non-model organisms (Baird et al. 2008; Andolfatto et al. 2011; Elshire et al. 2011; Peterson et al. 2012). The hundreds, thousands or tens of thousands of SNPs embedded in the resulting restriction-site associated DNA (RAD) sequence tags (Baird et al. 2008) have a myriad of uses in biology ranging from genetic mapping (Wang et al. 2013; Weber et al. 2013), to population genomics (Hohenlohe et al. 2010; Andersen et al. 2012; White et al. 2013), phylogeography (Emerson et al. 2010; Reitzel et al. 2013), phylogenetics (Dasmahapatra et al. 2012; Eaton and Ree 2013), and marker discovery (Scaglione et al. 2012; Toonen et al. 2013).

The choice of appropriate restriction enzyme(s) is critical for the effective design of any study using RAD sequencing and related methods such as genotyping-by-sequencing (GBS) (Elshire et al. 2011), multiplexed shotgun genotyping (MSG) (Andolfatto et al. 2011), and double digest RAD-seq (ddRAD) (Peterson et al. 2012), among others. This choice determines the number of markers that can be obtained, the amount of sequencing needed for a desired coverage level, the number of samples that can be multiplexed, the monetary cost, and ultimately the success of a project. It has been widely suggested that the number of restriction sites in a genome, for a given enzyme, can be roughly predicted using simple probability, if one has an idea of the genome size and GC composition (Baird et al. 2008; Davey et al. 2011). Both of these parameters can be measured approximately in non-model organisms through sequencing-independent techniques such as flow cytometry (Vinogradov 1994; Vinogradov 1998; Šmarda et al. 2011). However, preliminary evidence has suggested that there can be significant departures from expectations for particular combinations of taxa and restriction enzymes (Davey and Blaxter 2011; Davey et al. 2011).

Type II restriction enzymes, endonucleases chiefly produced by prokaryotic microorganisms, cleave double stranded DNA (dsDNA) at specific unmethylated recognition sequences 4 to 8 base pairs long that are usually palindromic. These enzymes are thought to play an important role as defense systems against foreign phage dsDNA during infection or as selfish parasitic elements, and therefore have been the center of an evolutionary ‘arms race’ (Rambach and Tiollais 1974; Karlin et al. 1992; Rocha et al. 2001). Type II restriction enzymes are not known in eukaryotes and are not used as virulence factors by bacteria to infect eukaryotic hosts. Therefore there are no *a priori* reasons to believe that recognition sites in eukaryotic genomes are subject to selective pressures, but rather should be evolutionarily neutral. Eukaryotic genomes are known to have heterogeneous compositions with characteristic signatures at the level of di- and trinucleotides that are largely independent of coding status or function (Karlin and Mrázek 1997; Karlin et al. 1998; Gentles 2001). It is thus possible that genome composition at these levels has a large influence in the abundance of short sequence patterns, like recognition sequences of restriction enzymes, in eukaryotes.

The goal of this study is to test the hypothesis that genome composition can be used to predict the number of restriction sites for a given combination of restriction enzyme and taxon. For this we: i) performed systematic *in silico* genome-wide surveys of restriction sites for diverse kinds of type II restriction enzymes in 434 eukaryotic whole and draft genome sequences to determine their frequencies across taxa; ii) examined the composition of genomes at the level of di- and trinucleotides and determined patterns of compositional biases among taxa; iii) developed stochastic models based on GC content, mono-, di- and trinucleotide compositions to predict the frequencies of restriction sites across taxa and diverse kinds of type II restriction enzymes; iv) evaluated the accuracy of the predictive models by comparing the *in silico* observed frequencies of restriction sites to the expected frequencies predicted by the models. The number of restriction sites in a genome is not the only factor that determines the number of RAD tags that can be recovered experimentally. The architecture of each genome, and in particular the number of repetitive elements and gene duplicates, can contribute significantly. To quantify this contribution we assessed the proportion of restriction-site associated DNA tags that can potentially be recovered unambiguously after empirical sequencing. For this we performed *in silico* RAD sequencing and alignment experiments for all genome assembly-restriction enzyme combinations using a newly developed software pipeline, PredRAD (https://github.com/phrh/PredRAD).

Results

Observed frequencies of restriction sites

Observed frequencies of restriction sites were highly variable among broad taxonomic groups for the set of restriction enzymes here examined (Table 1) - except for *FatI* - with clear clustering patterns determined by phylogeny (Fig 1). For example for *NgoMIV* we observed 45.8 restriction sites per megabase (RS/Mb) ± 24.6 (mean ± SD) in core eudicot plants, compared to 277.4 ± 131.3 RS/Mb in commelinid plants (monocots). Among closely related species the frequency patterns were similar and variability generally small. Observed frequencies of restriction sites per megabase (RS/Mb) were inversely proportional to the length of the recognition sequence, with differences in orders of magnitude among 4-, 6-, and 8- cutters when compared within the same species, e.g. in the starlet anemone *Nematostella vectensis* there were 3917.6, 167.6, and 6.9 RS/Mb for the 4-cutter *FatI*, 6-cutter *PstI* and 8-cutter *SbfI*, respectively. Nucleotide composition of the recognition sequence did not show a clear correlation with the observed frequency of restriction sites, e.g. 83.6 RS/Mb ± 25.1 were observed in Neopterigii vertebrates for *KpnI* (GGTACC), compared to 622.6 RS/Mb ±119.1 observed for *PstI* (CTGCAG), both recognition sequences with a GC content of 66.7%.

Dinucleotide compositional biases

Dinucleotide odds ratios ($\bar{\rho}^*_{XY}$) (Burge et al. 1992), a measurement of relative dinucleotide abundances given observed component frequencies, revealed significant compositional biases for all possible dinucleotides (Fig 2). Both dinucleotides and trinucleotides are considered significantly underrepresented if the odds ratio is ≤ 0.78, significantly overrepresented if ≥ 1.23, and equal to expectation if =1 (Karlin et al. 1998). The dinucleotide compositional biases were highly variable among broad taxonomic groups but generally similar within. Two dinucleotide complementary pairs, CG/GC and AT/TA, had highly dissimilar relative frequencies between the members of each pair. The largest biases were for CG, being significantly underrepresented in groups like core eudicot plants ($\bar{\rho}^*_{XY}$=0.68 ± 0.11), gnathostomate vertebrates ($\bar{\rho}^*_{XY}$=0.32 ± 0.12), pucciniales fungi ($\bar{\rho}^*_{XY}$=0.66 ± 0.08), gastropods ($\bar{\rho}^*_{XY}$=0.68, SD=0.01), trebouxiophyceae green algae ($\bar{\rho}^*_{XY}$=0.61 ± 0.19) and saccharomycetales ($\bar{\rho}^*_{XY}$=0.78 ± 0.17). CG was significantly overrepresented in groups like apocritic insects ($\bar{\rho}^*_{XY}$=1.59 ± 0.18). The complementary dinucleotide GC was not particularly underrepresented in any broad taxonomic group, but tended towards overrepresentation in ecdyzosoan invertebrates ($\bar{\rho}^*_{XY}$=1.24 ± 0.12), being significant in several arthropod and nematode species. Other taxa that showed significant overrepresentation of GC included trebouxiophyceae ($\bar{\rho}^*_{XY}$=1.39 ± 0.04) and microsporidid fungi ($\bar{\rho}^*_{XY}$=1.28 ± 0.17). Relative abundances of the dinucleotide AT were within expectations for all eukaryotes, except for the fungus *Sporobolomyces roseus* (ρ^*_{XY}=0.78). Contrastingly, the TA dinucleotide tended towards underrepresentation throughout the eukaryotes ($\bar{\rho}^*_{XY}$=0.8 ± 0.13), except in a few hypocreomycetid fungi species for which it was significantly underrepresented. The TA dinucleotide was significantly underrepresented in groups like the trypanosomatidae ($\bar{\rho}^*_{XY}$=0.59 ± 0.03), choanoflagellida ($\bar{\rho}^*_{XY}$=0.43 ± 0.09), chlorophyta green algae ($\bar{\rho}^*_{XY}$=0.62 ± 0.15), and stramenopiles ($\bar{\rho}^*_{XY}$=0.70 ± 0.07), and marginally underrepresented in most euteleostei fish ($\bar{\rho}^*_{XY}$=0.77 ± 0.04), archosauria ($\bar{\rho}^*_{XY}$=0.76 ± 0.03) and basidiomycota ($\bar{\rho}^*_{XY}$=0.74 ± 0.09), among others.

The remaining dinucleotide complementary pairs had identical relative frequencies between the members of each pair. Dinucleotide pair GG/CC was marginally underrepresented in most eukaryotes ($\bar{\rho}^*_{XY}$=0.88 ± 0.15). In the sarcopterygii vertebrates ($\bar{\rho}^*_{XY}$=1.02 ± 0.06) and embryophyte plants ($\bar{\rho}^*_{XY}$=1.03 ± 0.07) GG/CC relative frequencies closely conformed to expectation. GG/CC was significantly overrepresented in handful of isolated ecdyzosoan, microsporidid and alveolate species, and significantly underrepresented in chlorophyta ($\bar{\rho}^*_{XY}$=0.72, SD=0.11), oomycetes ($\bar{\rho}^*_{XY}$=0.71 ± 0.05), and in several species of basidiomycota and dothideomycetes. Only the choanoflagellid *Salpingoeca* and the green alga *Asterochloris* presented a marginally significant bias for the dinucleotide pair AA/TT (ρ^*_{XY}=0.77 and 0.75 respectively). Similarly, *Salpingoeca* was the only taxon to show a significant bias for AC/GT (ρ^*_{XY}=1.42). Dinucleotide pair CA/TG was among the pairs with largest biases. Significant overrepresentation of CA/TG was found in several groups with large CG underrepresentation such as gnathostomates ($\bar{\rho}^*_{XY}$=1.31 ± 0.05), gastropods ($\bar{\rho}^*_{XY}$=1.29 ± 0.05), pucciniales ($\bar{\rho}^*_{XY}$=1.27 ± 0.02), trebouxiophyceae ($\bar{\rho}^*_{XY}$=1.62 ± 0.14), as well as several species of core eudicots and saccharomycetales. Other groups with significant CA/TG overrepresentation include onchocercid nematodes ($\bar{\rho}^*_{XY}$=1.26 ± 0.01), ustilaginomycotinid fungi ($\bar{\rho}^*_{XY}$=1.28 ± 0.05), trypanosomatids ($\bar{\rho}^*_{XY}$=1.25 ± 0.04), and amoebozoans ($\bar{\rho}^*_{XY}$=1.33 ± 0.06). Overrepresentation biases for the AG/CT dinucleotide pair were only present in amniotes ($\bar{\rho}^*_{XY}$=1.26 ± 0.02), sporidiobolales fungi ($\bar{\rho}^*_{XY}$=1.24 ± 0.01), and oxytrichid alveolates ($\bar{\rho}^*_{XY}$=1.24 ± 0.04), and other isolated species. Most of these taxa also had large CG underrepresentation. Lastly, most eukaryotes had GA/TC relative frequencies that conformed to expectations, except for few scattered species and small groups such as the microbotryomycetes fungi ($\bar{\rho}^*_{XY}$=1.45 ± 0.13), mamiellales green algae ($\bar{\rho}^*_{XY}$=1.40 ± 0.08), and eimeriorina alveolates ($\bar{\rho}^*_{XY}$=1.26 ± 0.02).

Triucleotide compositional biases

Trinucleotide odds ratios (γ^*_{XYZ}), a measurement of relative trinucleotide abundances given observed component frequencies, revealed compositional biases for most possible trinucleotides (Fig 3). However, most of these biases were only significant in scattered individual species (Fig 4). Among the trinucleotide pairs with significant underrepresentation, CTA/TAG and CGA/TCG showed the most definite broad taxonomic patterns. CTA/TAG was significantly underrepresented in most taxa, except for groups like commelinid plants (monocots) (γ^*_{XYZ}=0.87 ± 0.03), most core eudicots (γ^*_{XYZ}=0.81 ± 0.02), eleutherozoans (γ^*_{XYZ}=0.82 ± 0.01), molluscs (γ^*_{XYZ}=0.83 ± 0.01), and gnathostomates (γ^*_{XYZ}=0.82 ± 0.02) – exclusive of the chimaera *Callorhinchus milii*. Contrastingly the trinucleotide CGA/TCG was only significantly underrepresented in most tetrapod vertebrates (γ^*_{XYZ}=0.82 ± 0.02) – exclusive of muroid rodents, the bovidae and afrotheria.

The largest and more widespread overrepresentation biases were for the trinucleotide pair AAA/TTT, being significant in most eukaryotes, except for the majority of dikarya fungi (γ^*_{XYZ}=1.18 ± 0.07). The trinucleotide pairs TAA/TTA and AAT/ATT were significantly overrepresented in many metazoan taxa, particularly in neopterygii vertebrates (γ^*_{XYZ}=1.3 ± 0.05, and γ^*_{XYZ}=1.26 ± 0.05 respectively). AAG/CTT was significantly overrepresented in bacillariophytes (γ^*_{XYZ}=1.24 ± 0.03), oomycetes (γ^*_{XYZ}=1.28 ± 0.02), and saccharomycetales (γ^*_{XYZ}=1.26 ± 0.04). Lastly, CCA/TTG was significantly overrepresented in several tetrapod groups, including the laurasiatheria – exclusive of the chiroptera – (γ^*_{XYZ}=1.25 ± 0.02) and hominoidea (γ^*_{XYZ}=1.23 ± 0.004).

Expected frequencies of restriction sites

Trinucleotide composition models were in general a better predictor of the expected number of restriction sites than any of the other models, in terms of their accuracy and precision (Fig 5, Fig 6). The mononucleotide and GC content models produced undistinguishable predictions (Fig 5, Fig 6). In a few cases the other models outperformed the trinucleotide model, e.g. *EcoRI* (Fig 5, Fig 6, Fig 7). The fit of the predictions was highly variable among broad taxonomic groups but generally similar within, e.g. in Neopterigii vertebrates an average similarity index (*SI)* of 0.14 (SD 0.19) for *AgeI* with the dinucleotide model, compared to -0.31 (SD 0.19) in Sarcopterigii. The similarity index is defined as the quotient of the number of observed and expected restriction sites, minus one. A positive *SI* indicates that the number of observed restriction sites is greater than the expected, whereas a negative *SI* indicates a smaller number of observed sites than expected. If *SI* is equal to 0, then the number of observed sites is equal to the expectation. For example, a *SI* = 1 indicates that the number of observed restriction sites for a particular enzyme in a given genome is twice the number of expected sites predicted by a particular model.

Recovery of RAD-tags after in silico sequencing

In most cases the recovery of RAD-tags after *in silico* sequencing was very high, with a median percentage of suppressed alignments to the reference genome assembly of only 3%. (Fig 8). There was no evident recovery bias by restriction enzyme, but rather bias was pronounced in a few individual species, likely indicating an enrichment of repetitive regions or duplications.

Discussion

Genome-wide surveys of restriction sites

Observed cut frequencies for a given restriction enzyme are highly variable among broad eukaryotic taxonomic groups, but similar among closely related species. This is consistent with the hypothesis that the abundance of restriction sites is largely determined by phylogenetic relatedness. This pattern is most evident in groups that have a larger taxonomic representation, such as mammals. As more genome assemblies become available the pattern resolution will become clearer in many other underrepresented taxonomic groups, and through the use of comparative methods in a robust phylogenetic framework it would be possible to establish taxon-specific divergence thresholds diagnostic of significant evolutionary changes in genome architecture.

As expected, observed frequencies of restriction sites with shorter recognition sequences are generally higher than the observed frequencies with longer recognition sequences. However this pattern in not universal. There are several instances in which the frequency of restriction sites for a high-denomination cutter is higher than for a low-denomination cutter. For example, in primates the frequency of 8-cutter *SbfI* 24.6 RS/Mb (SD 1.7) is significantly higher than the frequency of 6-cutter *AgeI* 18.4 RS/Mb (SD 1.4). These deviations from expectation are indicative of enzyme-specific frequency biases for particular taxa, and, as illustrated in the results section, are not correlated with the base composition of recognition sequences.

Genomic compositional biases

Our analyses indicate that there are significant compositional biases for most dinucleotides and trinucleotides across the eukaryotes. Many of these biases are only significant in scattered individual species. However there are several particular dinuclotides and trinucleotides that show significant biases across the eukaryotic tree of life. Our observation that these biases are highly variable among broad taxonomic groups but generally similar within is congruent with findings from previous studies (Gentles 2001). The most obvious biases across taxa are observed in the gnatostomate vertebrates; however, this is most likely due to rampant undersampling in most other groups of eukaryotes (vertebrate genome assemblies represent 21% of all the taxa in this study).

The dinucleotides CG, GC, TA, and CA/TG show the most conspicuous bias patterns across the eukaryotic tree of life. Biases in most of these dinucleotides have been previously identified as likely linked to important biological processes. Notably the underrepresented dinucleotide CG is a widely known target for methylation related to transcriptional regulation (Bird 1980) and retrotransposon inactivation (Yoder et al. 1997) in vertebrates and eudicots. The corresponding overrepresentation of AG/CT fits the classic model of "methylation-deamination-mutation" by which a methylated cytosine in the CG pair tends to deaminate when unpaired and mutate into a thymidine with a corresponding CA complement. Interestingly CG, are GC, are significantly overrepresented in several groups of apocritic insects, as well as in some fungi and single-cell eukaryotes. CG is not a primary target for methylation in *Drosophila* (Lyko et al. 2000), instead CT, and in lesser degree CA and CC, are methylated in higher proportion. None of these dinucleotide pairs is significantly underrepresented in apocritic insects. The widespread TA underrepresentation has been traditionally attributed to stop codon biases, thermodynamic instability and susceptibility of UA to cleavage by RNAses in RNA transcripts (Beutler et al. 1989).

The trinucleotides CTA/TAG, AAA/TTT, TAA/TTA, CCA/TGG show the most conspicuous bias patterns across the eukaryotic tree of life. The biases in CTA/TAG have been widely attributed to the stop codon nature of UAG. However, the trinucleotides corresponding to the other stop codons (Burge et al. 1992), UAA and UGA, are overrepresented or not biased across eukaryotes. The reasons behind other cases of trinucleotide biases are less understood.

Predictability of restriction site frequencies

Our analyses indicate that in most cases the trinucleotide genome composition model is the best predictor, and the GC content and mononucleotide models are the worst predictors of the expected number of restriction sites in a eukaryotic genome. It is possible that the greater number of parameters in the trinucleotide model (64, compared to 16, 4 and 2 of the dinucleotide, mononucleotide and GC content model, respectively) is the cause of the better fit in general. However this trend is not universal. As illustrated in the results section, in a few cases the other models outperformed the trinucleotide composition model. Neither the GC content nor length of the recognition sequence can explain the observed discrepancies. It is not surprising that fit of the predictions made by the models is highly variable taxonomic groups, given the high variability observed in restriction sites frequencies and genetic compositions across the eukaryotic tree of life. We conclude that the predictability of restriction site frequencies in eukaryotic genomes needs to be treated in a case-specific basis, where the phylogenetic position of the taxon of interest and the specific recognition sequence of the selected restriction enzyme are the most determinant factors.

Implications for RAD-seq and related methodologies

For the design of a study using RAD-seq, or a related methodology, there are two general fundamental questions that researchers face: i) what is the best restriction enzyme to use to obtain a desired number of RAD tags in the organism of interest? And ii) how many markers can be obtained with a particular enzyme in the organism of interest? The results from this study, and the developed software pipeline PredRAD , will allow any researcher to obtain an approximate answer to these questions.

In a hypothetical best-case scenario for the design of a study using RAD-seq, or a related methodology, the species of interest is already included in the database presented here. In this case the best proxy for the number of RAD tags that could be obtained empirically would be twice the number of *in silico* observed restriction sites for each restriction enzyme (each restriction site is expected to produce two RAD tags, one in each direction from the restriction site) minus the number of suppressed read alignments to the reference genome assembly. For example, the a predicted number of RAD tags for *SbfI* in starlet anemone *Nematostella vectensis* is 3,370, being a close match to the range of RAD tags obtained empirically by Reitzel et al. (2013) of 2,300 – 2,800. If a new genome assembly becomes available for the species and/or the researcher wishes to evaluate an additional restriction enzyme, PredRAD can be re-run with these data to quantify the number of restriction sites, the recovery potential, as well as to estimate the probability of the new recognition sequence based on genome composition models.

In the scenario that the genome sequence of the species of interest is not available, the best alternative is to look at the closest relative with a genome assembly. A range of approximate values for the number of RAD tags can be obtained from i) the number of *in silico* observed restriction sites in the closely related species; ii) the frequency of restriction sites in the closely related species, and the genome size of the species of interest; and iii) the probability of the recognition sequence for the enzyme(s) based on the best-fit genome composition model (*SI* closest to 0) from the closely related species, and the genome size of the species of interest. The genome size of the species of interest can be estimated through sequencing-independent techniques such as flow cytometry (Vinogradov 1994; Vinogradov 1998; Šmarda et al. 2011).

For example, the predicted range in the number of RAD tags for *SbfI* in a thoracican barnacle is 10,000 – 30,000, based on the observed frequency of the SbfI recognition sequence and its probability using a trinucleotide composition model in the genome of the crustacean *Daphnia pulex* (ranges of genome size for barnacles were obtained from the Animal Genome Size Database, http://ww.genomesize.com). Herrera and Shank (**In prep.**) obtained *ca.* 18,000 RAD tags empirically. The possibility that frequency of restriction sites and genome composition can be accurately estimated from alternative datasets such as transcriptomes is worth evaluating.

Additional factors that can influence the actual number of RAD tag markers that can be obtained experimentally include: genome differences among individuals, level of heterozygosity, the amount of methylation in the genome, the number of repetitive regions and gene duplicates present in the target genome, the sensitivity of a particular restriction enzyme to methylation, the efficiency of the enzymatic digestion, the quality of library preparation and sequencing, the amount of sequencing, sequencing and library preparation biases, and the parameters used to clean, cluster and analyze the data, among others.

Conclusions

In this study we tested the hypothesis that genome composition can be used to predict the number of restriction sites for a given combination of restriction enzyme and genome. Our analyses reveal that in most cases the trinucleotide genome composition model is the best predictor, and the GC content and mononucleotide models are the worst predictors of the expected number of restriction sites in a eukaryotic genome. However, we argue that the predictability of restriction site frequencies in eukaryotic genomes needs to be treated in a case-specific basis, because the phylogenetic position of the taxon of interest and the specific recognition sequence of the selected restriction enzyme are the most determinant factors. The results from this study, and the software developed, will help guide the design of any study using RAD sequencing and related methods.

Methods

Observed frequencies of restriction sites

Assemblies from eukaryotic whole genome shotgun (WGS) sequencing projects available as of December 2012 were retrieved primarily from the U.S. National Center for Biotechnology Information (NCBI) WGS database (Table S1). Only one species per genus was included. Of the 434 genome assemblies included in this study 42% corresponded to fungi, 21% to vertebrates, 16% invertebrates, and 9% plants. Only unambiguous nucleotide calls were taken into account. Genome sequence sizes were measured as the number of unambiguous nucleotides in the assembly. A set of 18 commonly used palindromic restriction enzymes with variable nucleotide compositions was screened in each of the genome assemblies (Table 1). The number of restriction sites present in each genome was obtained by counting the number of unambiguous matches for each recognition sequence pattern. Under optimal experimental conditions each restriction site should produce two RAD tags, one in each direction from the restriction site. Therefore, we define the number of observed RAD tags in each genome assembly as twice the number of recognition sequence pattern matches.

Expected frequencies of restriction sites

To test the hypothesis that compositional heterogeneity in eukaryotic genomes can determine the frequency of restriction sites of each genome we characterized the GC content, as well as the mononucleotide, dinucleotide and trinucleotide compositions of each genome and developed probability models to predict the expected frequency of recognition sequences for each restriction enzyme. GC content was calculated as the proportion of unambiguous nucleotides in the assembly that are either guanine or cytosine, assuming that the frequency of guanine is equal to the frequency of cytosine. Mononucleotide composition was determined as the frequency of each one of the four nucleotides. Dinucleotide and trinucleotide compositions were determined as the frequency of each one of the 16 or 64 possible nucleotide combinations, respectively. The odds ratios proposed by Burge *et al.* (1992) were used to estimate compositional biases of dinucleotides (1) and trinucleotides (2) across genomes.

(1)

$$\rho^*_{XY} = \frac{f^*_{XY}}{f^*_X f^*_Y}$$

(2)

$$\gamma^*_{XYZ} = \frac{f^*_{XYZ} f^*_X f^*_Y f^*_Z}{f^*_{XY} f^*_{YZ} f^*_{XNZ}}$$

Where f_X^* is the relative frequency of the mononucleotide *X*, f_{XY}^* is the relative frequency of the dinucleotide *XY*, and f_{XYZ}^* is the relative frequency of the trinucleotide *XYZ*. All frequencies take into account the antiparallel structure of double stranded DNA. *N* represents any mononucleotide.

Mononucleotide and GC content sequence models were used to estimate the probability of a particular recognition sequence (3) assuming that each nucleotide is independent of the others and of its position on the recognition sequence. The GC content model assumes that the relative frequencies of guanine and cytosine in the genome sequence are equal. This model has only two parameters, the GC and AT frequencies. In the mononucleotide model there are four parameters, one for each of the four possible nucleotides.

(3)

$$p(s) = \prod_{i=1,\ldots,n(s)} p(s_i)$$

Here, $p(s_i)$ is the probability of nucleotide s_i at the position i of the recognition sequence. In the GC content model $p(s_i)$ can take the values of f_{GC} or f_{AT}. In the mononucleotide model $p(s_i)$ can take the values of f_A, f_G, f_C, or f_T.

Dinucleotide and trinucleotide sequence models were defined as first and second degree Markov chain transition probability models with 16 or 64 parameters, respectively (Karlin et al. 1992; Singh 2009). These models take into account the position of each nucleotide in the recognition sequence. Nucleotides along the recognition sequence are not independent from nucleotides in neighboring positions. The probability of a particular recognition sequence for these Markov chain models was calculated as:

(4)

$$p(s) = p(s_1) \prod_{i=2,\ldots,n(s)} p_c(s_i|s_{i-1},\ldots,s_{i-n})$$

Where $p(s_1)$ is the probability at the first position on the recognition sequence and p_c is the conditional probability of a subsequent nucleotide on the recognition sequence depending on the previous *n* nucleotides. In the dinucleotide sequence model $n = 1$ and in the trinucleotide sequence models $n = 2$.

Expectations versus observations

To assess the effectiveness of the predictive recognition sequence models we compared the number of observed restriction sites in the genome assemblies with the expected number. The expected number of restriction sites in a given genome was calculated as the product of the probability of a recognition sequence multiplied by the genome sequence size. To quantify the departures from expectation we define a similarity index (*SI)* as $FI = (O - E)/E$, where O and E are the observed and expected number of restriction sites, respectively. If $SI = 0$, then $E = O$. If $SI < 0$, then $E > O$, and *vice versa*.

Recovery of restriction-site associated DNA tags

To assess the proportion of restriction-site associated DNA tags that can potentially be recovered unambiguously after empirical sequencing we performed *in silico* sequencing experiments for all genome assembly-restriction enzyme combinations. For each restriction site located in the genome assemblies, 100 base pairs up- and down-stream of the restriction site were extracted. This sequence read length is typical of sequencing experiments performed with current Hi-Seq platforms (Illumina Inc.). The resulting RAD tags were aligned back to their original genome assemblies using BOWTIE v0.12.7 (Langmead et al. 2009). Only reads that produced a unique best alignment were retained. The analytical software pipeline here described and the output database files are available at https://github.com/phrh/PredRAD.

Acknowledgements

This research was supported by the Office of Ocean Exploration, National Oceanic and Atmospheric Administration (NA05OAR4601054), the National Science Foundation (OCE-0624627; OCE-1131620) and the Academic Programs Office (Ocean Ventures Fund award to SH), the Deep Ocean Exploration Institute (Fellowship support to TMS) and the Ocean Life Institute of the Woods Hole Oceanographic Institution. Adam Reitzel, Ann Tarrant, and Casey Dunn provided helpful discussions. We thank Ann Tarrant and Eleanor Bors for providing comments on this manuscript.

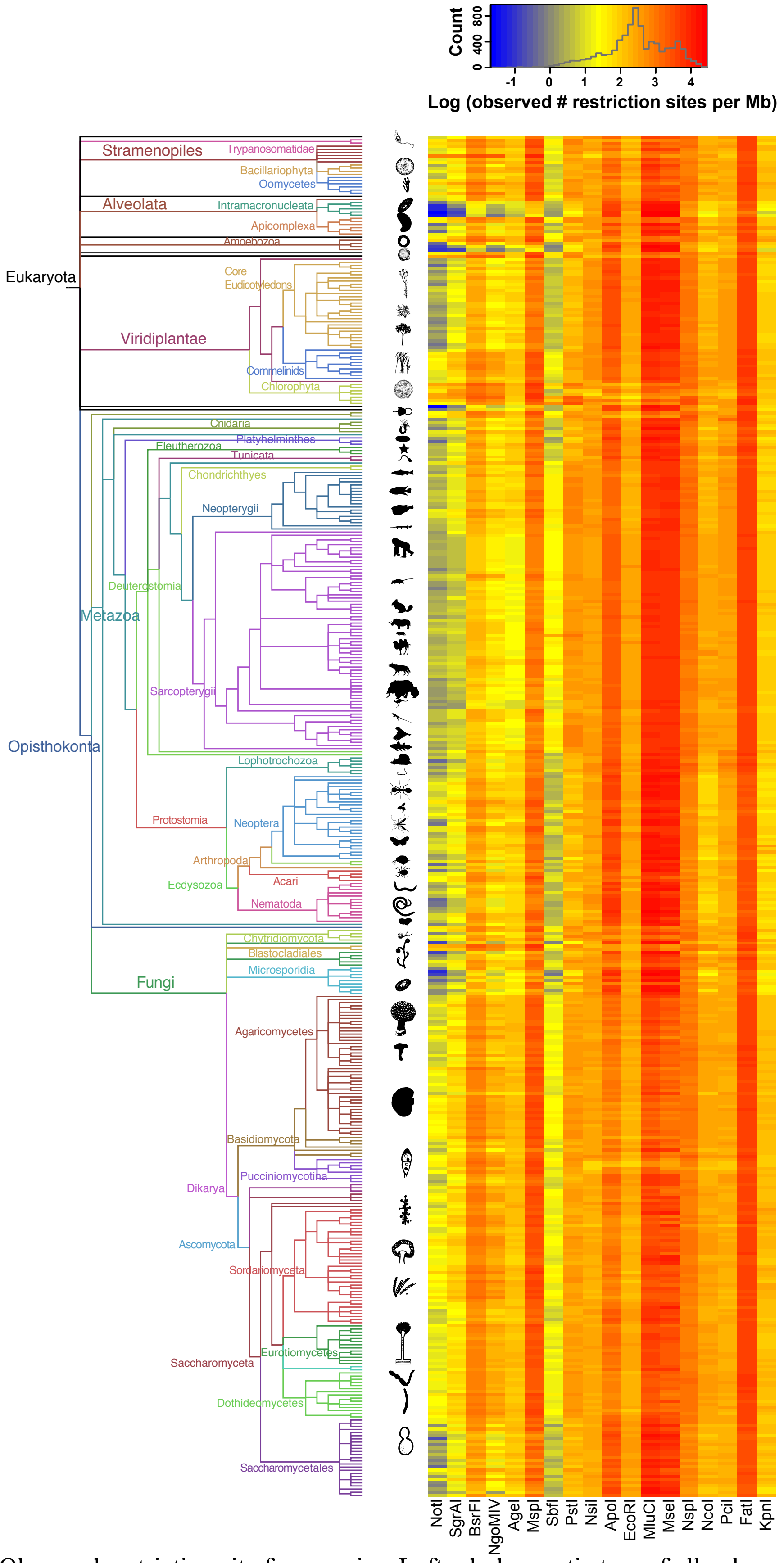

Figure 1. Observed restriction site frequencies. Left: phylogenetic tree of all eukaryotic taxa analyzed in

this study. The tree is based on the NCBI taxonomy tree retrieved on May 16, 2013 using the iTOL tool http://itol.embl.de (Letunic and Bork 2011). Branch colors and labels indicate broad taxonomic groups. Organism silhouettes and cartoons were created by the authors or obtained from http://phylopic.org/. Right: heatmap of the observed frequency of restriction sites. Each row corresponds to a species from the tree on the left, and each column corresponds to a different restriction enzyme. Gray line in the color-scale box shows the distribution histogram of all values.

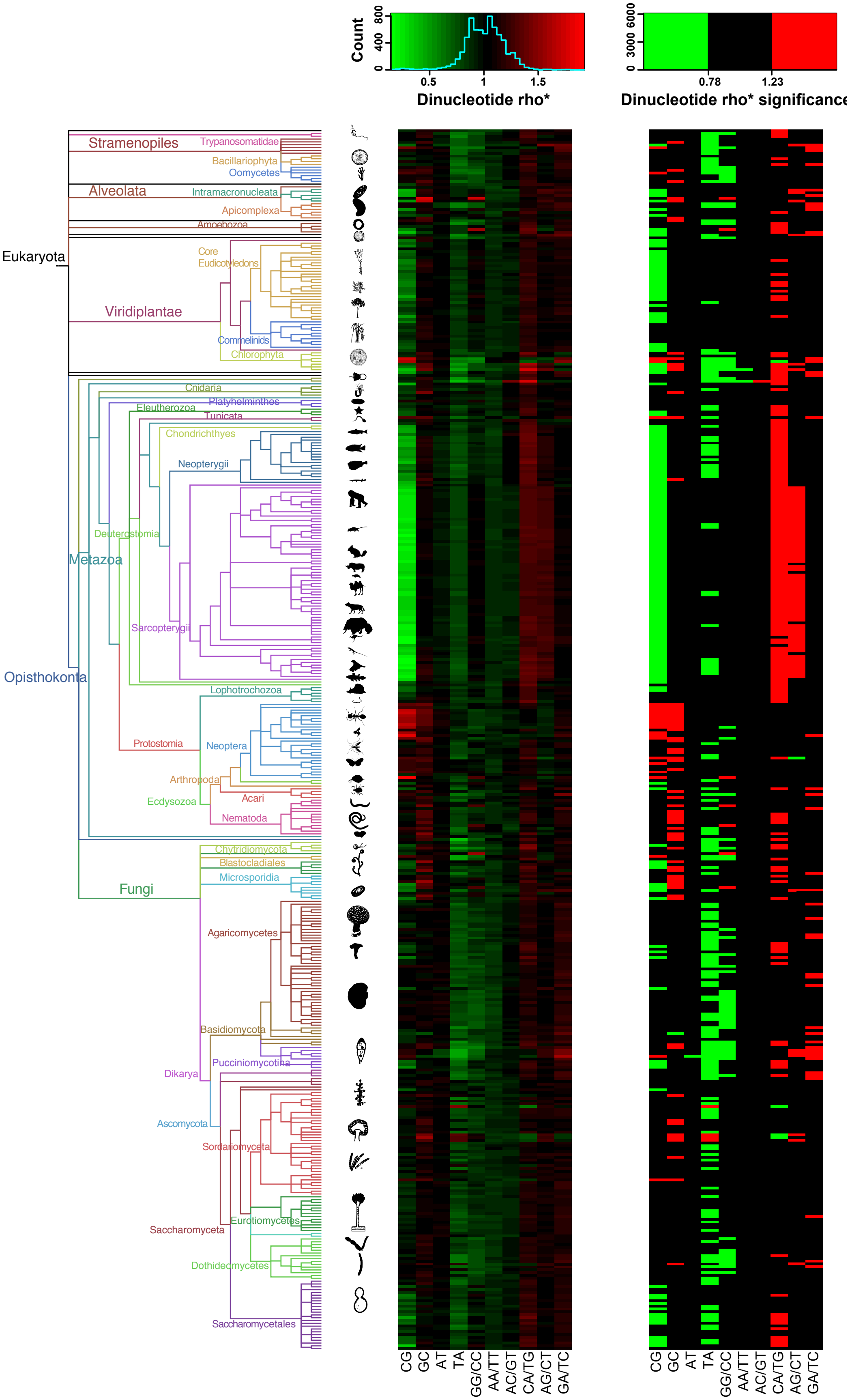

Figure 2. Dinucleotide compositional biases and significances. Left: phylogenetic tree as in Fig 1. Center:

heatmap of the ρ^*_{XY} odds ratio values. Right: heatmap of the ρ^*_{XY} odds ratio significant values $\rho^*_{XY}<0.78$ and $\rho^*_{XY}>1.23$. Each row corresponds to a species from the tree on the left, and each column corresponds to a different dinucleotide. Green indicates underrepresentation and red indicates overrepresentation. Cyan line in the color-scale box shows the distribution histogram of all values.

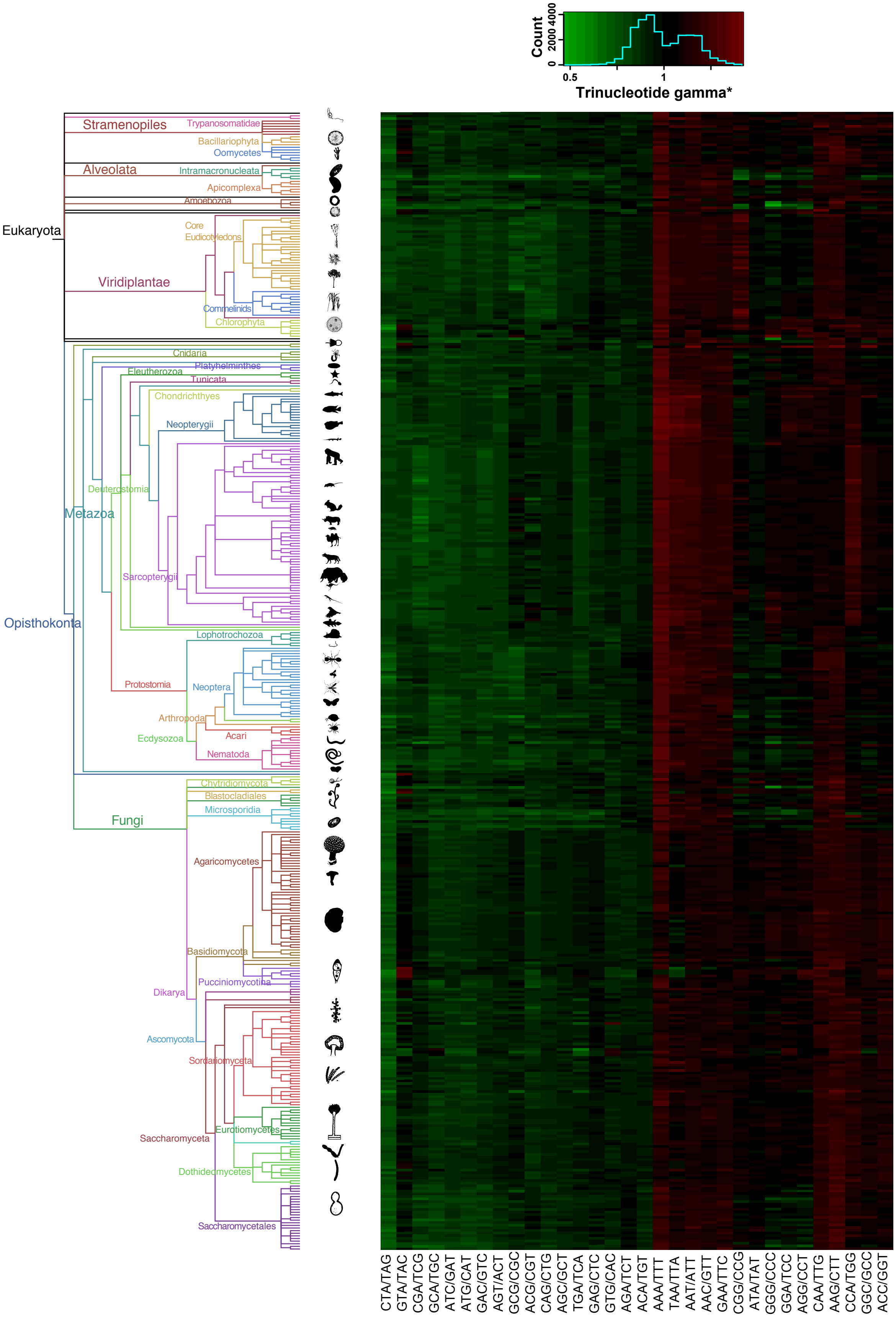

Figure 3. Trinucleotide compositional biases. Left: phylogenetic tree as in Fig 1. Right: heatmap of the

γ_{XYZ}^{*} odds ratio values. Each row corresponds to a species from the tree on the left, and each column corresponds to a different trinucleotide. Green indicates underrepresentation and red indicates overrepresentation. Cyan line in the color-scale box shows the distribution histogram of all values.

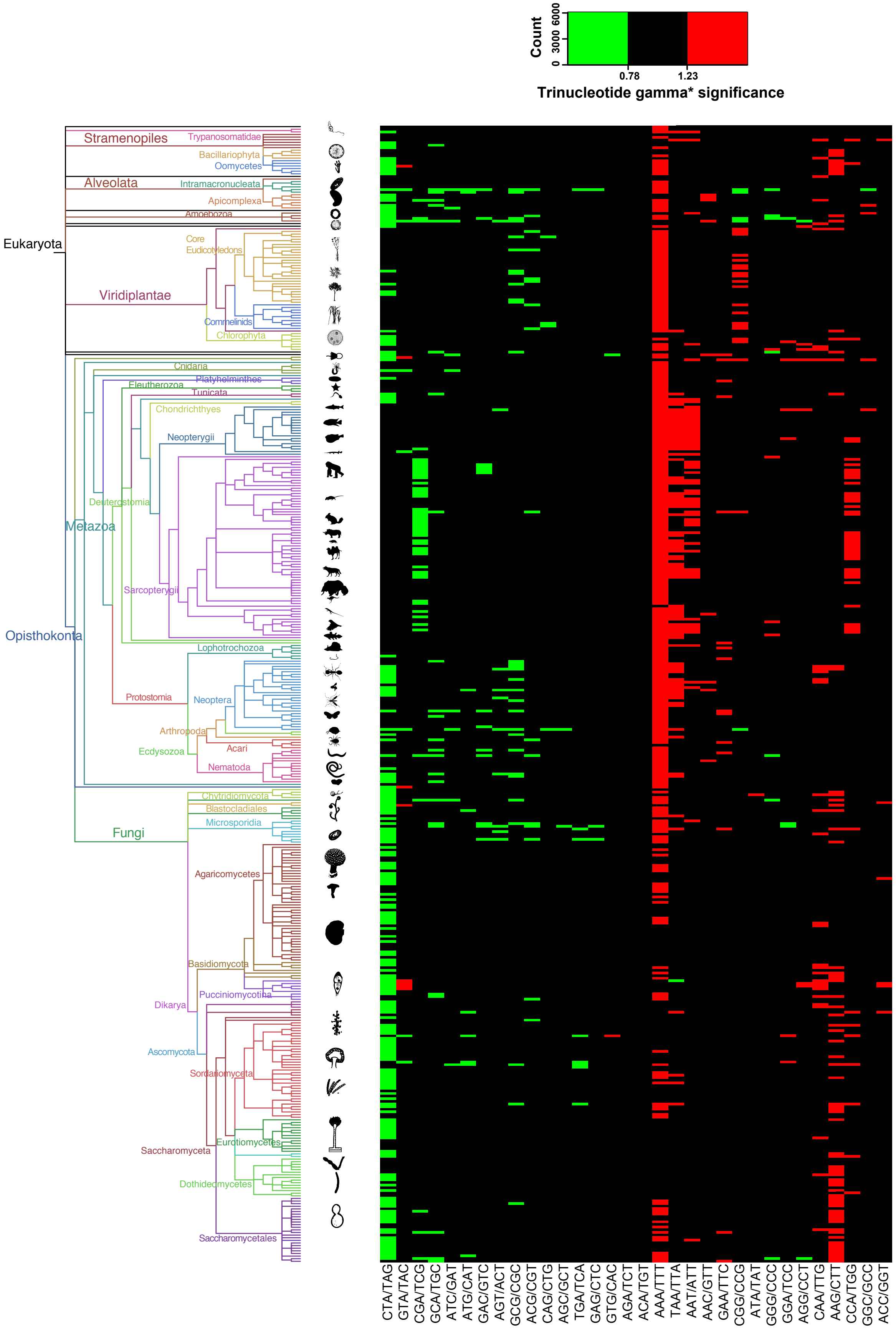

Figure 4. Trinucleotide compositional biases significances. Left: phylogenetic tree as in Fig 1. Right:

heatmap of the γ^*_{XYZ} odds ratio significant values ρ^*_{XY}<0.78 and ρ^*_{XY}>1.23. Each row corresponds to a species from the tree on the left, and each column corresponds to a different trinucleotide. Green indicates underrepresentation and red indicates overrepresentation. Cyan line in the color-scale box shows the distribution histogram of all values.

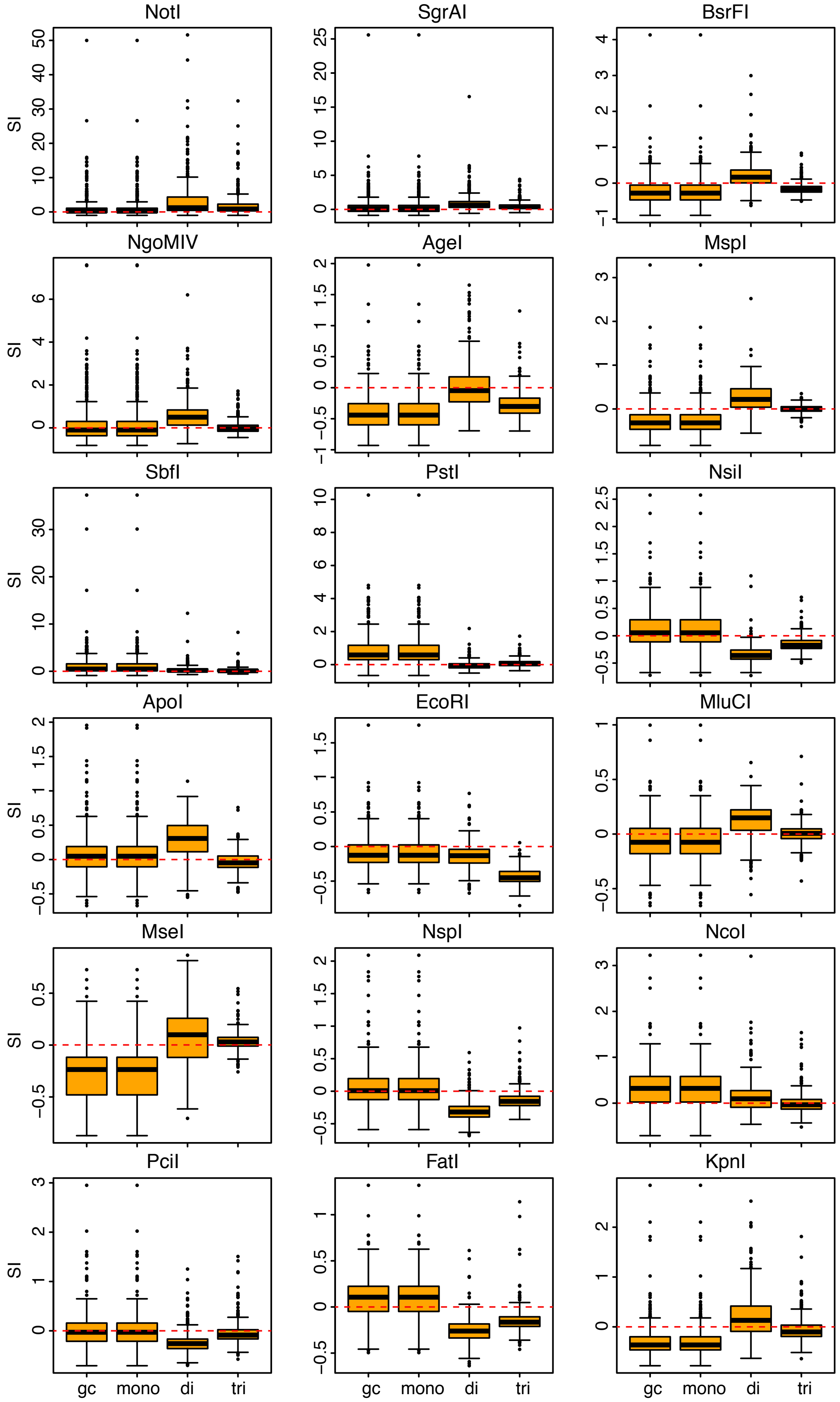

Figure 5. Overall fit of genome composition models per restriction enzyme. Vertical axes in the box and

whisker plots indicate the values of the similarity index (*SI*) for each species per enzyme. Horizontal axes in the box and whisker plots indicate the genome composition model: GC content (gc), mononucleotide (mono), dinucleotide (di), and trinucleotide (tri). Horizontal edges of range boxes indicate the first and third quartiles of the *SI* values under each composition model. The thick horizontal black line represents the median. Whiskers indicate the value of 1.5 times the inter-quartile range from the first and third quartiles. Outliers are defined as SI values outside the whiskers range and are represented by dots. Outlier value of *Entamoeba histoyitica* for *NotI* was excluded. Red dotted lines indicate *SI*=0.

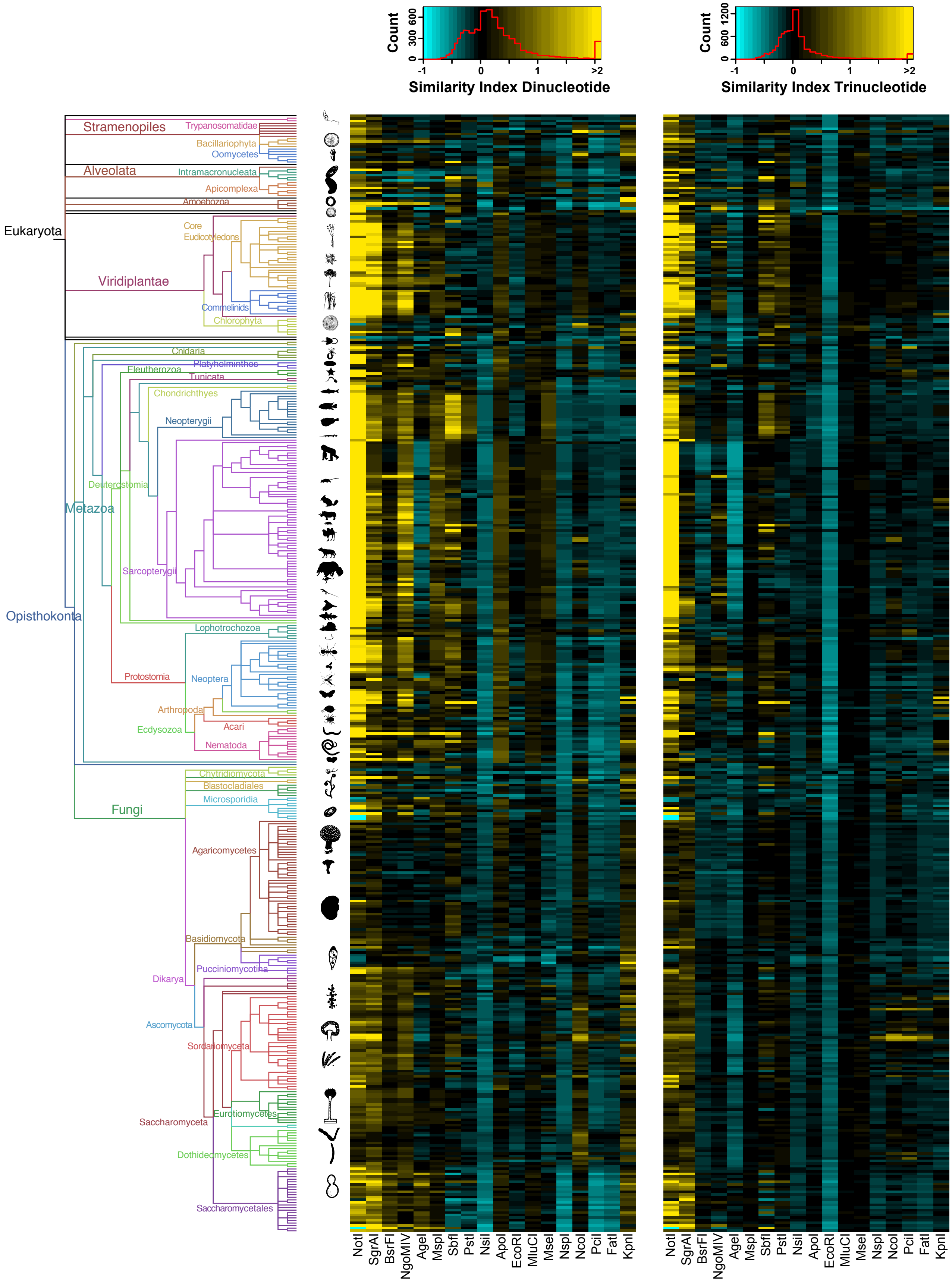

Figure 6. Similarity indexes for dinucleotide and trinucleotide genome composition models. Left:

phylogenetic tree as in Fig 1. Center: heatmap of the similarity indexes for the dinucleotide model Right: heatmap of the similarity indexes for the trinucleotide model**.** Each row corresponds to a species from the tree on the left, and each column corresponds to a different restriction enzyme. Cyan indicates *SI* < 0 and yellow indicates *SI* > 0. Red line in the color-scale box shows the distribution histogram of all values.

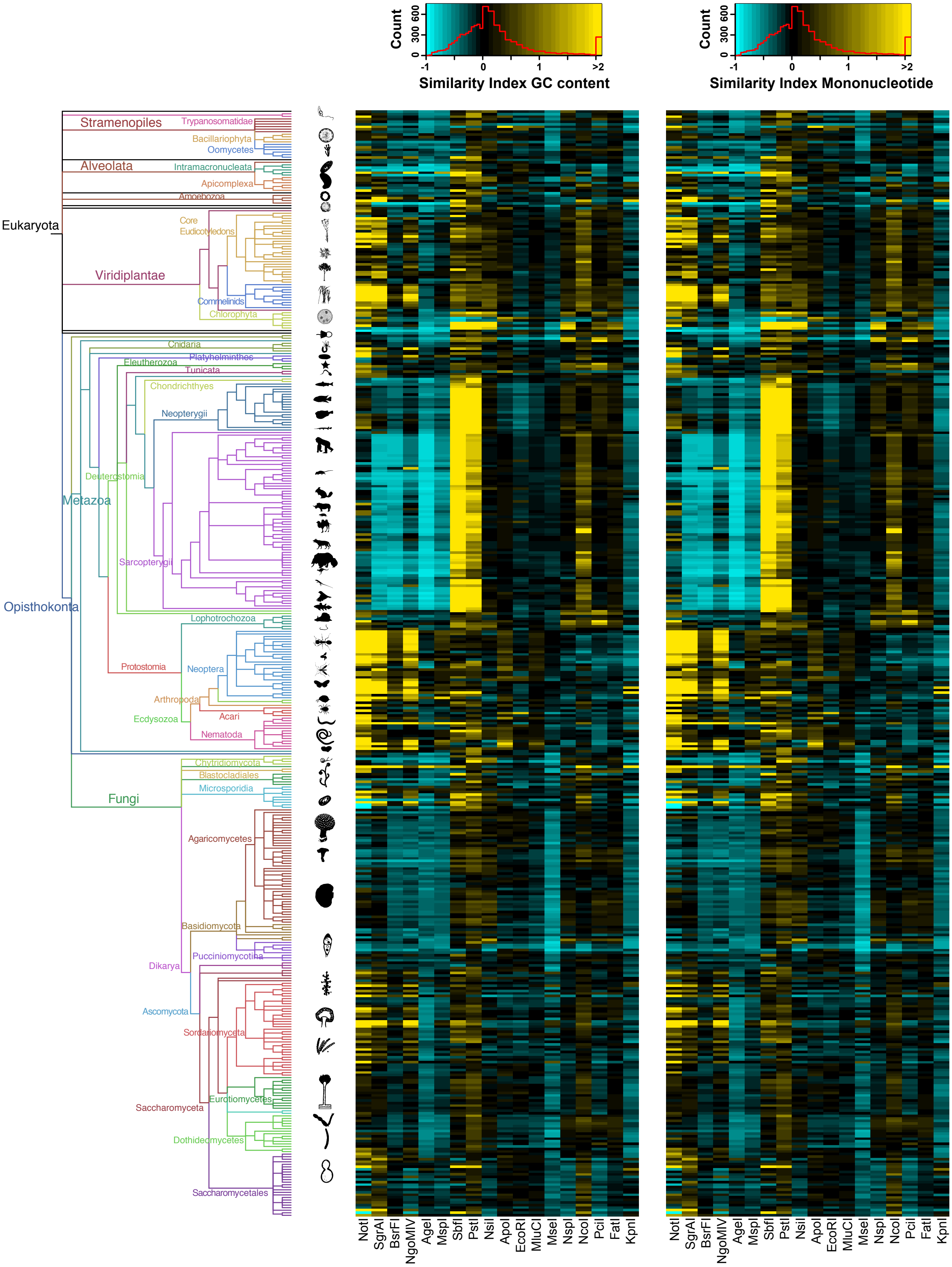

Figure 7. Similarity indexes for GC content and mononucleotide genome composition models. Left:

phylogenetic tree as in Fig 1. Center: heatmap of the similarity indexes for the GC content model Right: heatmap of the similarity indexes for the mononucleotide model**.** Each row corresponds to a species from the tree on the left, and each column corresponds to a different restriction enzyme. Cyan indicates *SI* < 0 and yellow indicates *SI* > 0. Red line in the color-scale box shows the distribution histogram of all values.

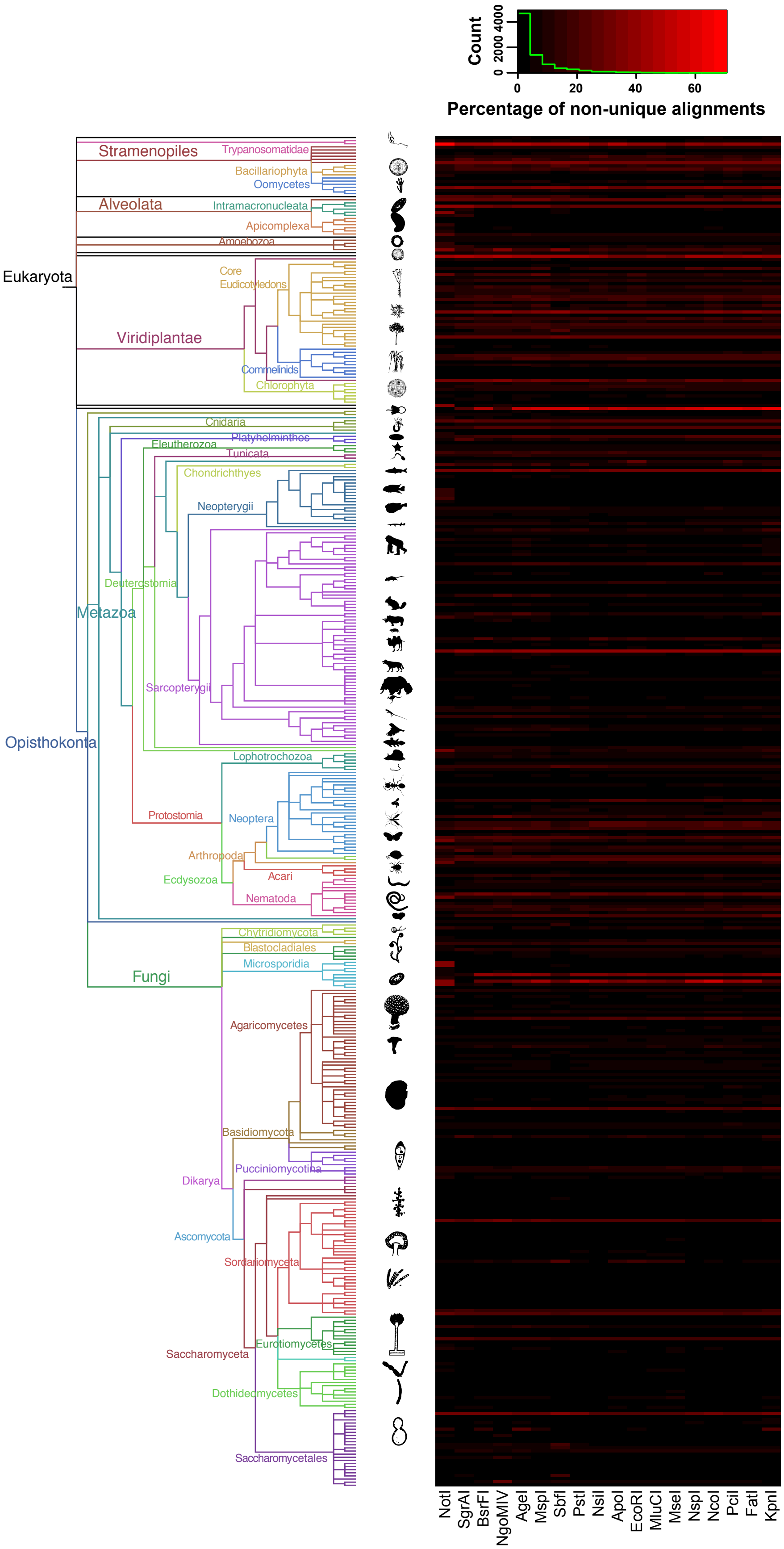

Figure 8. Recovery of RAD-tags after *in silico* genome digestion and sequencing. Left: phylogenetic tree

as in Fig 1. Right: heatmap of the percentage of RAD-tags that produced more than one unique alignment to their reference genome**.** Each row corresponds to a species from the tree on the left, and each column corresponds to a different restriction enzyme. Green line in the color-scale box shows the distribution histogram of all values.

Table 1. Restriction enzymes included in this study.

Core Sequence	Restriction Enzyme	Recognition Sequence	Recognition Sequence Length	GC Content of Recongition Sequence
GGCC				
	NotI	GCGGCCGC	8	100.0
CCGG				
	SgrAI	CRCCGGYG	8	87.5
	BsrFI	RCCGGY	6	83.3
	NgoMIV	GCCGGC	6	100.0
	AgeI	ACCGGT	6	66.7
	MspI	CCGG	4	100.0
TGCA				
	SbfI	CCTGCAGG	8	75.0
	PstI	CTGCAG	6	66.7
	NsiI	ATGCAT	6	33.3
AATT				
	ApoI	RAATTY	6	16.7
	EcoRI	GAATTC	6	33.3
	MluCI	AATT	4	0.0
TTAA				
	MseI	TTAA	4	0.0
CATG				
	NspI	RCATGY	6	50.0
	NcoI	CCATGG	6	66.7
	PciI	ACATGT	6	33.3
	FatI	CATG	4	50.0
GTAC				
	KpnI	GGTACC	6	66.7

Supplementary Figures and Tables

Figure S1. Poster-size figure including heatmaps from figures 1 to 7 and the phylogenetic tree of all eukaryotic taxa analyzed in this study including species names.

Table S1. Genome assemblies included in this study. Note that web addresses to individual assembly files, and the assembly files themselves, were as of December 2012 and may have changed.

References

Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, Félix M-A, Kruglyak L. 2012. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics* **44**: 285-290.

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, Stern DL. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research* **21**(4): 610-617.

Baird N, Etter P, Atwood T, Currey M, Shiver A, Lewis Z, Selker E, Cresko W, Johnson E. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**(10): 3376.

Beutler E, Gelbart T, Han J, Koziol J, Beutler B. 1989. Evolution of the genome and the genetic code: Selection at the dinucleo- tide level by methylation and polyribonucleotide cleavage. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **86**: 192-196.

Bird AP. 1980. DNA methylation and the frequency of Cpg in animal DNA. *Nucleic Acids Research* **8**(7): 1499-1504.

Burge C, Campbell AM, Karlin S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **89**(4): 1358-1362.

Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, Zimin AV, Hughes DST, Ferguson LC, Martin SH et al. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94-98.

Davey JW, Blaxter ML. 2011. RADSeq: next-generation population genetics. *Briefings in Functional Genomics and Proteomics* **9**(5-6): 416-423.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Publishing Group* **12**(7): 499-510.

Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology* **62**(5): 689-706.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) spproach for high diversity species. *PLoS One* **6**(5): e19379.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, Holzapfel CM. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **107**(37): 16196-16200.

Gentles AJ. 2001. Genome-scale compositional comparisons in eukaryotes. *Genome Research* **11**(4): 540-546.

Herrera S, Shank TM. **In prep.** Evolutionary history and biogeographical patterns of barnacles endemic to deep-sea hydrothermal vents.

Hohenlohe P, Bassham S, Etter P, Stiffler N, Johnson E, Cresko W. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**(2): e1000862.

Karlin S, Burge C, Campbell AM. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic acids research* **20**(6): 1363-1370.

Karlin S, Campbell AM, Mrázek J. 1998. Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**: 185-225.

Karlin S, Mrázek J. 1997. Compositional differences within and between eukaryotic genomes. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **94**(19): 10227-10232.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research* **39**: W475-W478.

Lyko F, Ramashoye BH, Jaenisch R. 2000. DNA methylation in *Drosophila melanogaster*. *Nature* **408**(538-540).

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* **7**(5): e37135.

Rambach A, Tiollais P. 1974. Bacteriophage ʹ having EcoRI endonucleases sites only in the nonessential sites of the genome. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **71**: 3927-3930.

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM. 2013. Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Molecular Ecology* **22**(11): 2953-2970.

Rocha EPC, Danchin A, Viari A. 2001. Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Research* **11**: 946-958.

Scaglione D, Acquadro A, Portis E, Tirone M, Knapp SJ, Lanteri S. 2012. RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *Bmc Genomics* **13**: 3.

Singh GB. 2009. Stochastic models for biological patterns. In *Bioinformatics for Systems Biology*, (ed. S Krawetz), pp. 151-162. Springer, New York.

Šmarda P, Bureš P, Šmerda J, Horová L. 2011. Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. *New Phytologist* **193**(2): 513-521.

Toonen RJ, Puritz JB, Forsman ZH, Whitney JL, Fernandez-Siva I, Andrews KR, Bird CE. 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* **1**: e203.

Vinogradov A. 1994. Measurement by flow cytometry of genomic AT/GC ratio and genome size. *Cytometry* **16**: 34-40.

-. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry* **31**: 100-109.

Wang J, Wurm Y, Nipitwattanaphon M, Riba-Grognuz O, Huang Y-C, Shoemaker D, Keller L. 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493**: 664–668.

Weber JN, Peterson BK, Hoekstra HE. 2013. Discrete genetic modules are responsible for complex burrow evolution in Peromyscus mice. *Nature* **493**(7432): 402-405.

White TA, Perkins SE, Heckel G, Searle JB. 2013. Adaptive evolution during an ongoing range expansion: the invasive bank vole (Myodes glareolus) in Ireland. *Molecular Ecology* **22**(11): 2971-2985.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13**(8): 335-340.