Genome architecture impacts on reduced representation population genomics

Carles Galià-Camps[1], Cinta Pegueroles[2], Xavier Turon[3], Carlos Carreras[4], and Marta Pascual[1]

[1]Universitat de Barcelona
[2]Facultat de Biologia, Universitat de Barcelona
[3]Center for Advanced Studies of Blanes (CEAB, CSIC)
[4]Universitat de Barcelona Facultat de Biologia

June 24, 2023

Abstract

Genomic architecture is a key evolutionary trait for living organisms. Due to multiple complex adaptive and neutral forces which impose evolutionary pressures on genomes, there is a huge disparity of genomic features. However, existing genome architecture studies are taxon biased, and thus a wider picture should be obtained by expanding the taxonomic scope. Moreover, the extent to which genomic architecture determines the typology of loci recovered in reduced representation sequencing techniques with digestion enzymes is largely unexplored. Here, we observed that whereas plants mostly increase their genome size by expanding their intergenic regions, animals expand both intergenic and intronic regions, although the expansion patterns differ between deuterostomes and protostomes. We found positive correlations between the percentage of loci obtained with in-silico digestion using 2b-enzymes mapping in introns, exons and intergenic categories and the percentage of these regions in the genome. However, exonic regions showed a significant enrichment regardless of the enzyme used. Moreover, the percentage of loci retained after secondary reductions varied with selective-adaptors and genome GC content. In summary, we show that genome architecture has an impact on the markers obtained in reduced representation sequencing that should be considered in conservation genomics for correct wildlife management.

Genome architecture impacts on reduced representation population genomics

Running title: Genome architecture impacts RADseq studies

Carles Galià-Camps[1,2,*], Cinta Pegueroles[1,2], Xavier Turon[3,ǂ], Carlos Carreras[1,2,ǂ], Marta Pascual[1,2,ǂ]

(1) Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, SPAIN

(2) Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), Barcelona, Spain.

(3) Department of Marine Ecology, Centre d'Estudis Avançats de Blanes (CEAB-CSIC), Accés Cala Sant Francesc 14, 17300 Blanes, SPAIN

*Corresponding author: cgaliacamps@gmail.com

ǂ: These authors jointly supervised this work

Abstract

Genomic architecture is a key evolutionary trait for living organisms. Due to multiple complex adaptive and neutral forces which impose evolutionary pressures on genomes, there is a huge disparity of genomic features. However, existing genome architecture studies are taxon biased, and thus a wider picture should be obtained by expanding the taxonomic scope. Moreover, the extent to which genomic architecture determines the typology of loci recovered in reduced representation sequencing techniques with digestion enzymes is largely unexplored. Here, we observed that whereas plants mostly increase their genome size by expanding their intergenic regions, animals expand both intergenic and intronic regions, although the expansion patterns differ between deuterostomes and protostomes. We found positive correlations between the percentage of loci obtained with *in-silico* digestion using 2b-enzymes mapping in introns, exons and intergenic categories and the percentage of these regions in the genome. However, exonic regions showed a significant enrichment regardless of the enzyme used. Moreover, the percentage of loci retained after secondary reductions varied with selective-adaptors and genome GC content. In summary, we show that genome architecture has an impact on the markers obtained in reduced representation sequencing that should be considered in conservation genomics for correct wildlife management.

Keywords

Genome evolution, 2b-RAD, Secondary reduction, Genomic categories, Exon enrichment

Introduction

The availability of genomes is blooming. In the last five years, methodological advantages in the sequencing of long fragments have enhanced exponentially the quantity and quality of genomic resources, and several initiatives from global to regional scope have arisen aiming to produce genomes of all biodiversity (Formenti et al., 2022; Lewin et al., 2022). In this context, genome availability provides an unprecedented opportunity to dig deep into the genome architecture of living organisms (Campbell et al., 2018; Hotaling et al., 2021) including genome size (Hidalgo et al., 2017), repeated (Platt et al., 2018; Wu & Lu, 2019) and duplicated regions (Heckenhauer et al., 2022; Li et al., 2018), GC content (Amit et al., 2012; Haerty & Ponting, 2015), and percentage of intergenic and genic regions (Francis & Wörheide, 2017; Zhu et al., 2009), among others. Although previous studies on genome architecture focused on certain taxonomic groups and genomic traits (Kapusta et al., 2017; Mueller & Jockusch, 2018; Platt et al., 2018; Wu & Lu, 2019), a general picture is still missing. Genome evolutionary processes are complex, and involve many mechanisms which are heterogeneous among taxa. The current availability of chromosome-level genomes across taxonomic groups allows identifying broad patterns of genomic architecture, which might impact on population genomic studies, as a key element when assessing genomic structural variants and performing SNP calling (Rhie et al., 2021). Thus, it is important to know beforehand the genomic architecture of the study taxon, since it might affect the category of the loci being analyzed, and therefore influence the results.

Population genomic studies is a fast-expanding field. Reduced genome sequencing techniques using restriction site digestion enzymes (RAD) are widely used to obtain genome-wide markers of targeted species, rendering population genomic analyses feasible, especially when working with species with big genome sizes or without reference genomes (Guo et al., 2021; Manuzzi et al., 2019; Peterson et al., 2012; Torrado et al., 2020). These methods allow working with many individuals without compromising SNP calling accuracy, since high sequencing depth is required for reliable genotyping (Davey & Blaxter, 2010; Galià-Camps et al., 2022). Restriction enzymes presumably cleave the genome randomly and the resultant fragments are assumed to mirror the genomic structure of the original genome (Davey & Blaxter, 2010; Wang et al., 2012). Consequently, the percentage of loci in a genomic category should be representative of the percentage of the genome in the same genomic category, although this has not yet been tested empirically. In consequence, there is an urgent need to evaluate whether this assumption holds true for major taxonomic groups.

Among these methods, 2b-RAD uses 2b-enzymes that identify the recognition site and cleave DNA upstream and downstream at a given length generating small fragments of 32–34 bp, with sticky ends including a few random nucleotides (Marshall & Halford, 2010). Thanks to the small fragments produced, this technique allows working with degraded DNA (Barbanti et al., 2020). All generated fragments can be sequenced following standard protocols with ligation of fully degenerate adaptors for library building. Nonetheless, this enzyme family allows using base-selective adaptors, which select fragments with desired

nucleotides in their sticky ends (Barbanti et al., 2020; Galià-Camps et al., 2022; Wang et al., 2012). This capacity provides to this technique the capacity to further reduce the number of loci, making studies cheaper and therefore allowing to work with species with large genomes, as well as to include many more individuals given a locked budget. In comparison to the first reduction, which is produced by the enzymes' target sites, secondary reduction via base selection could be directed to specific categories by setting which nucleotides are fixed in the adaptors. The use of this technique, however, depends crucially on the absence of biases in the number of loci being retained due to base selection, whose lack of bias has not been formally tested.

Here, we demonstrate that there is a genomic architecture trichotomy among plants, protostomes and deuterostomes, and corroborate that the general assumption that reduced representation sequencing with 2b-enzymes reflects the overall genome architecture holds true, albeit with a slight enrichment in exonic regions. Additionally, we show that building genomic libraries with base-selective adaptors efficiently reduces the number of loci without compromising the percentage of genomic categories recovered, and can therefore be used for correct wildlife genomic management and conservation. Nonetheless, we detect a mild differential enrichment on the number of loci in each selection type according to the genome GC content. Our results can guide adaptor selection in future genomic studies according to the tackled taxon and research question.

Material and Methods

Reference genome datasets

We downloaded 80 chromosome-level genome assemblies from GeneBank, ranging from 102Mb to 4.7Gb belonging to plants and animals (Data S1). Information on the GC content for each genome was also retrieved (Data S1). The clusterings of the selected taxa (herein designated as supergroup and group) were based on the phylogenetic relationships obtained from Timetree web server (http://www.timetree.org/) (Kumar et al., 2017). We defined 3 different supergroup clusters, composed of 13 plants, 18 protostomes and 49 deuterostomes (Data S1) (Figure 1a). Additionally, a total of 12 different groups were defined: plants (13), molluscs (5), nematodes (1), arthropods (12), echinoderms (1), tunicates (1), fishes (14), amphibians (6), mammals (12), lepidosaurs (3), testudines (2) and birds (10). The groups with more than six species were further evaluated separately (Figure 1a). We retrieved the annotation files of the same genomes in GFF format, obtaining the annotation for 44 genomes: 10 plants, 2 molluscs, 1 nematode, 9 arthropods, 1 tunicate, 7 fishes, 2 amphibians, 5 mammals, 1 lepidosaur, 2 testudines and 4 birds (Data S1).

Genomic architecture and functional categories

For every annotated genome assembly, we calculated the number of base pairs in each genomic category (intergenic, intronic and exonic) using the genomecov function with -d -split options from BEDTools (Data S1) (Quinlan & Hall, 2010). To do so, we first converted the GFF files to bed12 format using the gff3_file_to_bed.pl utility from Transdecoder (https://github.com/TransDecoder/TransDecoder/blob/master/util/gff3_file_to_be

d.pl). After this step, we calculated the relative proportion of the three genomic categories for each genome.

Genomic in-silico digestions

We computationally digested the 80 genomes with 2b-enzymes, which cleave at both sides of the recognition site generating fragments of uniform length (Wang et al., 2012), using the program Phyper.pl (Seetharam & Stuart, 2013). This program recognizes 2b-enzyme targets, cleaves the DNA, and exports the obtained fragments. We carried out the analyses with three 2b-enzymes with different GC content in their recognition sites: AlfI ([10/12]GCA[N6]TGC[12/10], 66% GC), CspCI ([11/13]CAA[N5]GTGG[12/10], 57% GC) and BaeI ([10/15]AC[N4]GTAYC[12/7], 50% GC). For every *in-silico* digestion we obtained a summary file with the total number of loci (all fragments) and only the unique loci (fragments that were present only once in the genome), and two fasta files corresponding to the total and unique sequences (Seetharam & Stuart, 2013). Although the number of total and unique loci are informative, these can be deeply influenced by the genome size and the enzyme used. In order to reduce the effect of these two factors, we calculated the percentage of unique loci to standardize the data for comparisons.

In-silico digestions simulating the use of base-selective adaptors for secondary reduction

Base-selective adaptors can further reduce the number of loci obtained when using 2b-enzymes, and should ideally optimize the costs on population genomic studies without compromising the genomic information (Galià-Camps et al., 2022;

Wang et al., 2012). We simulated the effect of using base-selective adaptors with the bash script select_bases_fasta_2.0.sh (Barbanti et al., 2020). We performed in-silico base selections of the unique loci with GC (S) sticky ends (G-G, G-C, C-C and C-G) and AT (W) sticky ends (A-A, A-T, T-T and T-A) in order to determine the percentage of unique loci that would be retained with this secondary selection using the three enzymes for each of the 80 downloaded genomes. Finally, we calculated the percentage of retained sequences after the selection with S and W compared with the initial number of unique loci for each species and enzyme.

Categorical profiling of 2b-RAD loci

To calculate which proportion of loci correspond to intergenic, intronic and exonic categories, we first selected the sequences of the unique loci resulting from the *in-silico* digestion with the three enzymes for the annotated genomes. We then compared them against their corresponding reference genome using BLAST (Altschul et al., 1990) and kept only the coordinates of those assignments with a match of 100% (same size, 100% of identity, e-value=1^10^{-16}). Afterwards, we used the in-house script classifyBlastOut.py pipeline to classify unique hits as genic (exonic and intronic) or intergenic (https://github.com/EvolutionaryGenetics-UB-CEAB/classifyBlastOut/). The blast hits that included both exonic and intronic regions were classified as exonic, independently if they belonged to the same gene or to different overlapping genes. Finally, we estimated the percentage of unique loci corresponding to each genomic category in S-selected and W-selected datasets for each annotated genome and enzyme.

Graphics and statistical analyses

Dispersion plots and violin plots were drawn with the R package “ggplot2” (Wickham et al., 2016), and regression formulas and their R^2 and p-values were calculated using the “stats” package from R. General Linear Mixed-Effects Models (GLMMs) were conducted with the R packages “lme4” (Bates, 2010), and “car” (Fox et al., 2012) was used to assess statistically significant effects of the explanatory factors. For statistical requirements, data were transformed for normalization. For the factors with frequency values (percentage of unique loci, percentage of genome in genomic category and percentage of unique loci in genomic category) normalization was achieved through an arcsine-square root transformation. For the factors with count values (number of total loci, number of unique loci and genome size) normalization was achieved through a logarithmic transformation. The package “rsq” (D. Zhang, 2018) was used to check the proportion of the variance explained by the whole model and by the fixed factors included in it. Tukey post-hoc comparisons for levels of significant factors of the GLMM were carried out with the package “emmeans” (Lenth et al., 2020), and plots were generated with the function *emmip* from the same package.

Results

Evolutionary trends on genome architecture

The compilation of the 80 genomes (Figure 1a) highlighted significant different trends among supergroups regarding how the three considered genomic categories change related to genome size (Figure 1b). In all three taxonomic supergroups, species with small genome sizes proportionally had higher amounts of exonic regions, as shown by the negative slope values of their regression equations and a high coefficient of determination (Table S1). The percentage of intergenic regions in plants increased with genome size, while keeping the percentage of intronic regions at low levels with a significant negative regression (Figure 1b, Table S1). On the other hand, animal genomes increased in size by expanding both intergenic and intronic regions (Figure 1b). However, the two animal supergroups differed in the abundance of intergenic regions (Figure 1b), which only increased significantly with genome size in deuterostomes (Table S1). General Linear Mixed-Effects Models (GLMM) on the percentage of each genomic category as the dependent variable detected significant differences for the interactions considering the three factors, Genomic category, Supergroup and Genome size (Table S2). For the double interaction for the categorical factors Genomic category and Supergroup, no differences among Supergroups for exonic regions were indicated by the post-hoc tests (Table S3). However, the percentage of intronic and intergenic regions was significantly different between plants and animals (Fig1b, Table S3).

In-silico genome digestions using 2b-RAD enzymes

Our results showed that the total number of loci (all fragments) and unique loci (those fragments whose sequence was present only once in the genome) obtained after in-silico digestions were highly correlated (Figure S1). Both the total and unique loci significantly increased with genome size in all enzymes, regardless of taxonomic level (Figure 2, Table S4). The phylogenetic relationships of the species included in the analysis determined the regressions' equations and coefficients (Figure 2, Table S4). When considering species separated by supergroups (plants, protostomes and deuterostomes), or groups with six or more analyzed species (plants, arthropods, fishes, amphibians, mammals and birds), all the regression equations had significant positive slopes but varied according to the species being included and the enzyme used (Table S4). Mammals were the exception to the global trends (Table S4), since two genomes in this group (platypus and red deer, the smallest and largest genomes analyzed, respectively) had low numbers of loci (Data S1). In the three GLMM models tested considering all species together (Total model), split by supergroup (Supergroup model) or by group (Group model), the proportion of the variance explained by the fixed factors was high (Table S5). For the Total model, significant differences were found among enzymes (AlfI, CspCI, BaeI), with genome sizes and their interaction (Table S5). Differences were due to the higher number of loci obtained with AlfI, being this number intermediate for CspCI and smallest for BaeI, and to the increase of the number of loci with genome size, with different slopes for each enzyme (Figure 2, Table S4). For the Supergroup model, the fixed factors explained 92% of the variance for total loci and 90% for unique loci, and supergroup, enzyme and genome size presented significant

differences, as well as the interaction enzyme*supergroup in both total and unique loci (Table S5). Tukey's post-hoc pairwise comparisons indicated major significant differences between deuterostomes and the other two taxa for all enzymes but BaeI (Table S6). Similar results were obtained when considering the group model, with the highest proportion of variance explained by the fixed factors (Table S5). As in the Supergroup model, no differences when using BaeI were found between taxonomic groups (Table S7). However, AlfI presented significant differences in total and unique loci when comparing plants and arthropods against the other groups as assessed with Tukey's post-hoc tests (Table S7). For CspCI, significant differences were only found when comparing arthropods' unique loci with other groups.

The percentage of unique loci varies across taxa

Unique loci are of ultimate interest in population genomic analysis, since loci found in multiple locations are removed in the filtering steps. However, they are deeply influenced by genome size and enzyme, as previously shown. Consequently, we used the percentage of unique loci for comparison and to reduce the effect of these two factors. As expected, our GLMM showed that the percentage of unique loci recovered, in relation to the total number of loci, was not different across enzymes and did not change with genome size when considering all species together (Fig 3a, Table S8). However, these percentages were dependent on the taxa analyzed, since the proportion of the variance of the full model explained by the fixed factors increased when the species were combined in lower level phylogenetic groups, suggesting lineage-specific variation (Table S8). On the Supergroup model, deuterostomes displayed

279 significantly higher percentages of unique loci than plants and protostomes
280 (Figure 3a, Table S9). Finally, the Group model showed different behaviors
281 depending on the groups, since mammals (0.958±0.036, mean±SE) and birds
282 (0.972±0.027) presented a higher percentage of unique loci (Figure 3a), although
283 this effect was only significant in mammals when compared to plants, arthropods
284 or fishes (Table S10). Surprisingly, birds did not show significant values despite
285 their high percentage of unique loci and low dispersion values (Figure 3a).
286 However, the model presented a high 95% confidence interval on the percentage
287 of unique loci in birds, which overlapped with all other groups (Figure S2)
288
289 ***2b-RAD digestions slightly enrich exonic loci***
290 In all supergroups, the percentage of unique loci in a given category significantly
291 increased with the percentage of the genome that is in the same category (Figure
292 3b, Table S11). Overall, the percentage of variation explained by all regression
293 equations was very good, as indicated by the coefficient of determination of the
294 full model (R^2), although in deuterostomes and plants the values were lower for
295 the intronic region (Table S11). Loci mapping in exonic regions were significantly
296 more frequent than expected, since the values falled above the dotted line (Figure
297 3b) that represents the percentage of loci in a genomic category expected under
298 the null hypothesis of random distribution of loci. Moreover, the slopes of their
299 regression equations were significantly above one in all three taxonomic
300 supergroups and for the three enzymes (Table S11). On the contrary, loci in
301 intronic regions presented regression slopes smaller than one, although only
302 significant in deuterostomes for the three enzymes (Table S11). The regressions
303 demonstrated that the percentage of loci in intergenic regions had a good fit with

the percentage of the intergenic fraction in the genome, and did not differ significantly from one with the only exception of enzyme CspCI in plants (Table S11). Nevertheless, the observed values were inferior to the expected ones as they always fall below the dotted line (Figure 3b). The GLMM using as the dependent variable the ratio between the percentage of loci in a genomic category and the percentage of the same category in the genome, identified significant differences among enzymes, supergroups, genomic categories and their pairwise interactions (Table S12). There were significant differences between genomic categories for all enzymes with the exception of the comparison between intergenic and intronic categories for AlfI and BaeI (Table S13). The percentage of loci in intergenic regions did not differ between supergroups, although it was significantly different between plants and all animals for the exonic category and plants and protostomes for the intronic ones. All genomic categories were significantly different in all supergroups but between intergenic and intronic regions for plants and deuterostomes (Table S13, Figure S3).

Base selection performance depends on genome GC content

As expected, our results showed that base-selective adaptors efficiently reduced the number of loci, and that this reduction was highly dependent on the selection performed (Figure 4). The percentage of unique loci retained was significantly different between selection strategies, with W-selection providing a significantly higher number of loci than S-selection, as shown in the GLMM analyses (Table S14, Figure 4a). Additionally, significant differences were found for some of the interactions when simulating adaptor selection (Table S14) for the Total,

Supergroup and Group models. For the Total model, BaeI displayed significantly different values for both S and W selection when compared to AlfI and CspCI, and selection type always showed significant differences independently of the enzyme used (Table S15). Differences were not found between enzymes within selection type for the Supergroup and Group models, but S-selection presented a lower percentage of loci in all taxonomic groups (Table S16, Table S17). Another significant common interaction for the Supergroup and Group models was found for the interaction between selection and supergroup/group. For the Supergroup model, all contrasts were significantly different (Table S17). On the other hand, the Group model showed that both S and W selection behaved differently when comparing mammals to all other groups but birds. Furthermore, S-selection provided a significantly lower percentage of loci for all groups but birds (Figure 4a, Table S17, Figure S4). The number of loci retained could be highly influenced by the nucleotide content of the genomes, generally richer in AT than in GC (Data S1), and thus returning more loci when using W-selection (32.7±0.029%) than S-selection (19.2±0.025%). We observed that the percentage of loci retained with S-selection within each supergroup was positively correlated with the species genome GC content, while the percentage of loci retained with W-selection was negatively correlated with the species GC content (Table S18, Figure 4b). Thus, more loci were retained with S selection when the genome GC content was higher, being the opposite situation with W selection (Figure 4b). Overall, considering the species GC content, which is on average 39.93±3.9% (Supplementary Data), the expected mean number of loci for the 80 analyzed species would be 15.94% (% of $GC^2/100$). Similarly, the W-selection expected number of loci given an average AT content of 60.07±3.9%

would be 36.08% after applying the same equation (% of $AT^2/100$). Thus, the number of S-selected and W-selected loci matches the probability of finding a “G” or a “C” at both ends of the enzyme’s cleavage site according to the species nucleotide content.

Base selection significantly enrich genomic categories

After base-selection, we mapped back the selected loci to their reference genomes and identified to which genomic category they belonged. Loci in exons were enriched by both W and S selections since the values falled above the dotted line (Figure 5), and their regression slopes were higher than one, an effect that was significant in all cases except for protostomes with S-AlfI and S-CspCI (Figure 5, Table S19). Introns had slopes below one, although only significantly different in plants in the S-selection. On the other hand, intergenic regions showed a deficit of selected loci (Figure 5) but did not display differences between S and W treatments, and their slopes were not different from one in all cases but for plants with S-selection, where it was significantly higher (Table S19). We carried out General Linear Mixed-Effects Models (GLMM) using the ratio between the percentage of loci in a genomic category and the percentage of the same genomic category in the genome as the dependent variable, and enzyme, selection, genomic category and supergroup as fixed factors (Table S20). All pairwise interactions were significant, except for the enzyme*selection interaction (Table S20). The only significant triple interaction was for enzyme, supergroup and genomic category (Table S20). The post-hoc tests for the two pairwise interactions involving selection were carried out since selection was the only factor that was not included in the significant three way interaction. All the post-

hoc tests for the selection and genomic category interaction resulted in significant values with the exception of the comparison between S-W selection in the intergenic regions (Table S21). The selection type directly affected the percentage of unique loci in a given genomic category with a higher percentage of intronic loci in W-selection and higher for exonic regions in S-selection (Table S21). All genic regions showed significantly different behaviors regardless of the selection performed with loci enriched in exonic regions and depleted in intergenic regions (Table S21). Regarding the interaction between supergroup and selection, only deuterostomes showed significant differences between S and W selection (Table S21). When comparing the different supergroups, only plants for W-selection were different to both protostomes and deuterostomes (Table S21). The three way interaction showed that all supergroups presented significant differences in the dependent variable between genomic categories in the same direction regardless of the enzyme used, with the exception of protostomes which did not show significant differences between intergenic and intronic regions with any of the three enzymes (Table S22, Figure S5). Plants showed a significantly higher percentage of loci in exons compared to animals with all three enzymes (Table S22, Figure S5). Protostomes showed significantly lower values in introns than plants when using the enzymes AlfI and BaeI. Finally, there were no significant differences between enzymes with the exception of AlfI recovering more loci in the exons of plants (Table S22, Figure S5).

Discussion

In population genomic studies it is of crucial importance to have a good understanding on the genomic architecture of the study taxon, to design an

optimal study as we demonstrate that it has a direct impact on the results and subsequent interpretations. Our results show that plants and animals increase genome size by differently expanding intergenic and intronic genomic categories. The number of markers in the different genomic categories with 2b-enzymes, used in reduced representation genome sequencing, positively correlates with the percentage of each region in the genome with an enrichment of loci in exons. Moreover, secondary reduction techniques, allowed in library construction when using 2b-enzymes, shows how the GC genome content is influencing the number of loci retained upon the selective-adaptors used. Finally, with the trends detected in the present study, the number of total and unique loci in addition to the percentage of intergenic, intronic and exonic regions, and loci within them, can be estimated for new study taxa from our specific regression lines provided their genome size.

Genomic architecture is shaped by multiscale evolutionary processes

Our results shed light on a genome architecture main dichotomy between plants and animals, with the second ones further differentiating protostomes and deuterostomes in how they increase genome size. Since genome architecture is subjected to evolutionary processes, an amalgam of constraints have shaped the different supergroup genomes which should not be neglected when designing population genomic studies. In the case of plants, we have shown that they likely expanded their intergenic regions in order to increase their genome. Allopolyploidy has been a main evolutionary trigger for plants, since 87.5% to 99.5% of them have been subjected to hybridization at some point during their evolutionary history, with a posterior rediploidization (Qiao et al., 2019). This

process involves many genomic changes, with fast gene deletion being one of the predominant mechanisms (Li et al., 2021). As a result, homeolog gene loss after polyploidization may allow plants to solve dosage-balance constraints explaining the evolutionary success of allopolyploidy in this group (Soltis et al., 2015). Retained homeologs in plants enhance protein family diversity without relying on introns to create different isoforms through alternative splicing, resulting in a higher number of genes (Kress et al., 2022; Qiao et al., 2019; Wang et al., 2019). Thus, allopolyploidy may help to maintain the exonic and intronic regions at low proportion despite increasing genome sizes, as observed in our study. Furthermore, the high percentage of intergenic regions in plants compared to animals is in agreement with plants increasing their genomes by expansions of transposable elements that can be activated by hybridization and polyploidization altering silencing mechanisms (Ågren & Wright, 2015; Wendel et al., 2016).

Conversely, intronic regions were highly abundant in animals across genome sizes suggesting an alternative strategy to enhance protein diversity. The abundance of intronic regions has been proposed to facilitate alternative splicing as the principal mechanism of gene family structural enrichment in animals (Grau-Bové et al., 2018). On the other hand, the percentage of intergenic regions in animals also increased, especially in deuterostomes. In the origin of vertebrates, two ancient rounds of whole genome duplications 450 Mya occurred (Sacerdot et al., 2018), whose signal has been diluted by transposable element (TE) expansions (Kapusta et al., 2017; Naville et al., 2019). The duplication event followed by TE activity might have increased the proportion of intergenic regions, since the effect of TE could inactivate former duplicated genes and thus

contribute to the expansion of intergenic regions at expense of ancient genes (Kapusta et al., 2017; Naville et al., 2019). Furthermore, regulatory elements modulating gene expression, highly abundant in vertebrates, have been identified in intergenic regions (Borys & Younger, 2020; Elkon & Agami, 2017). For instance, genes with large intergenic regions are preferentially expressed in neural tissues in vertebrates, suggesting not only regulation through cis-regulatory elements but also structural chromatin variation mediated by elements in intergenic regions for these organisms (Jaura et al., 2022). In fact, intergenic regions contain a wide range of long non-coding RNA families that act regulating gene expression in specific environmental or physiological contexts (Marlétaz et al., 2023). Consequently, the increase of intergenic regions with genome size in deuterostomes, as detected in our study, might facilitate species evolution through regulatory networks. However, the information on regulatory elements is limited to a few species due to the lack of comprehensive annotations in most organisms, highlighting the need for correct annotation for the increasing number of available reference genomes.

Reduced sequencing techniques reflect genomic architecture traits

The absence of biases in the number of loci and their genome composition being retained by reduced genome representations on population genomic studies is a daring prior to be assumed without evidence. With using three different 2-enzymes in the three major eukaryotic lineages, we have been able to demonstrate that the usage of this technique generally mirrors genome structure with few considerations to take into account. As expected, we found that the number of total loci increased altogether with genome size in all enzymes

regardless of taxonomic level. However, significant differences were found for the interaction between enzyme and supergroup, indicating that enzyme selection determines the number of markers recovered in different species according to their taxonomic groups, as previously found empirically but with a very limited number of taxa (Barbanti et al., 2020). Moreover, no significant interactions were detected between taxonomic categories and genome sizes, indicating that the number of loci increases with size in the same way regardless of the species' phylogenetic placement. The number of total and unique loci are key parameters for RADseq studies, as only unique loci will pass the filtering process. Therefore it is of great interest to calculate in advance which is the expected loci number for the study species in order to adapt the number of loci needed for the study and optimize sequencing effort while minimizing missing data (Barbanti et al., 2020; Galià-Camps et al., 2022). Similarly to the percentage of regions in a genomic category, the number of total and unique loci can be best approximated for a new species of interest from the taxon specific regression equation presented here, if an estimation of the genome size is available, since differences have been observed among taxonomic groups.

Although the number of unique loci positively increases with genome size, not all taxonomic groups behaved similarly when considering the percentage of unique loci according to its total number. In this scenario, plants and protostomes showed lower percentage of unique loci in comparison to deuterostomes. Abundance of recent transposable elements in protostomes and polyploidy in plants might determine the lower proportion of unique loci in these groups (Belser et al., 2018; Chueca et al., 2021; Li et al., 2018; Wang et al., 2019; Wu & Lu, 2019). Mammals and birds are well studied taxa that share many genomic

evolutionary traits, such as an active expansion of transposable elements but large DNA deletions (Feng et al., 2020; Kapusta et al., 2017) that might determine the higher percentage of unique loci found in these two groups in the present study. Thus, it would be reasonable to find significant differences among these taxa and all the other ones, as we found for mammals. However, birds did not show significant differences compared to the other groups despite their high percentage of unique loci and low dispersion values. Birds are known for having compact genome sizes, ranging from 0.9 to 1.6Gb, thought to be driven by flight constraints (Feng et al., 2020; Kapusta et al., 2017). Since the statistical model for birds integrates genome size as a continuous variable, it needs to estimate values from 0 to 5Gb. This effect adds uncertainty to the model as shown by the presence of large confidence intervals and, as a result, it generates a model for birds whose values are not significantly different from the other taxa. Fishes present high dispersion values. In addition to the two rounds of whole genome duplications in the base of the vertebrate lineage, fishes suffered a third genome duplication ca. 350 Mya (teleost's genome duplication), and a fourth recent one (5.6 to 11.4 Mya) occurred in cyprinids (Berthelot et al., 2014; Chen et al., 2019). Consequently, the lower levels of unique loci in fish, and specifically the three outliers with an extremely low proportion corresponding to the cyprinid species, are coherent with the statement that loci obtained from digestion with 2b-enzymes reflect genome evolution.

When assessing whether 2bRADseq is biased on genome composition, we found that there was a mild enrichment in exonic regions. Actually, other studies have recently proved that, indeed, in RADseq studies exonic enrichment is found depending on the enzyme used (López et al., 2022). The enzymes used in this

study (AlfI, CspCI and BaeI) have a percentage of GC content over 50% in their recognition sites, which is coherent with the higher percentage of 2b-RAD loci in the GC enriched exonic regions (Amit et al., 2012; Glémin et al., 2014; Schwartz et al., 2009) rather than intronic or intergenic. Thus, the effect of different enzymes preferentially targeting certain genomic regions, specifically exons, should be considered in study design. Digestion enzymes are defensive molecules synthesized mostly by bacteria to neutralize pathogens by targeting a determined sequence of an exogenous genome and breaking it (Loenen et al., 2014; Samson et al., 2013). As consequence, the higher percentage of loci in exonic regions found is probably a result of the restriction enzyme functionality, which by natural selection should have evolved to target coding regions of pathogenic elements in order to fastly inactivate them to ensure survival of the bacterial threatened organism (Hampton et al., 2020). Thus, the enrichment of loci in exonic regions when using 2b-enzymes validate exons being the genomic category with the highest percentage of GC content, especially in plants (Amit et al., 2012; Glémin et al., 2014). Our results, therefore, support that the GC content of enzyme recognition sites influences the genomic categories being recovered in RADseq genomic studies, as previous studies suggested (López et al., 2022).

Base-selection unravels GC biases in taxa and genomic categories

2b-RADseq is a highly interesting technique, since it allows working with degraded samples and is the only one that permits a secondary reduction by using base-selective adaptors to further reduce sequencing costs (Barbanti et al., 2020). As expected, we have shown that base-selective adaptors efficiently

553 reduced the number of loci recovered, but the number of loci retained depended
554 on the selection performed and the genome GC content.
555 Our results demonstrate that the target GC content of each enzyme might drive
556 the differences on the number of loci being retained by each base-selection.
557 Accordingly, AlfI, the enzyme with the richest GC content in the recognition site
558 (66% GC), recovers a higher percentage of S-selected loci than BaeI (50% GC).
559 This effect is accentuated by the intrinsic features of each genome, since the
560 higher is the GC content of the genome, the higher percentage of the loci are
561 recovered using S-selection, whereas lowering those recovered if using W-
562 selective adaptors. Thus, the percentage of genome GC content can be used to
563 predict the performance of both S-selective and W-Selective secondary
564 reductions. The number of S-selected and W-selected loci should match the
565 probability of finding a “G” or a “C” at both ends of the enzyme’s cleavage site
566 according to the species nucleotide content. Our results provided values close to
567 the predicted ones based on the species GC content, validating the feasibility to
568 know beforehand the secondary reduction performance depending on each
569 species genome size and selection type, and supporting the impact of genome
570 architecture in determining the percentage of base-selective loci obtained.
571 Loci in exons were enriched by both W and S selections, although their
572 regression slopes were above 1 and higher in S selection. This pattern is
573 coherent with exonic regions across eukaryotes being GC enriched, and
574 therefore further overrepresented when using S-selective adaptors. On the other
575 hand, the percentage of loci in introns had slopes below one, although only
576 significant in plants in the S-selection. Similarly, slopes below one reflect that
577 intronic regions are overall slightly enriched in AT nucleotides (Amit et al., 2012;

Zhu et al., 2009). Adenine and thymine are bonded by only two hydrogen bridges, while guanine and cytosine are paired by three. Consequently, AT generates less persistent secondary structures in pre-mRNA, which are easier to be removed by alternative splicing (J. Zhang et al., 2011) than GC ones. Overall, the high R^2 values in exonic and intergenic regions suggest that the selection of loci by 2b-enzymes combined with base-selective adaptors mirrors the percentage of these two categories in the genome.

Concluding remarks

The integrative approach adopted in the present work has demonstrated that the wide genomic architecture richness across the eukaryotic tree of life has been shaped by multiple complex adaptive and neutral forces leading their evolution. Evolutionary trends demonstrated here open a study field on inter-specific major lineages, which have been mostly understudied due to the precedent unavailability of genomes and the lack of taxon wide studies. We demonstrated that species-specific genome architecture plays a key role in reduced representation population genomic studies, since the typology of the loci recovered by these methodologies mirrors the genomic structure, although with slight enrichments of some categories depending on the enzyme, selection type and tackled taxon. Accordingly, depending on the enzyme, secondary selection and species genomic architecture, a fair representation of the genome will be recovered, a crucial requirement in population genomic studies which aim for accurate management and conservation across species with diverse genomic architectures.

Acknowledgments

CG thanks the members of his family for providing animal pictures for Figure 1a, in particular to his brother's dog Keni, shaped under the vertebrates' clade.

This research was funded by MarGech (PID2020-118550RB, MCIN/AEI/10.13039/501100011033) from the Spanish Government. CG was funded with the predoctoral contract [PRE-2018-085227 - MCIN/AEI/10.13039501100011033] by the Spanish Ministry of Science, Innovation and Universities, and by ERDF "A way of making Europe". The authors CG, CP, CC and MP are members of the research group 2021 SGR 01271, and XT is member of the research group 2021 SGR 00405, from the Generalitat de Catalunya (AGAUR).

References

Ågren, J. A., & Wright, S. I. (2015). Selfish genetic elements and plant genome size evolution. *Trends in Plant Science*, *20*(4), 195–196.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., Pupko, T., & Ast, G. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Reports*, *1*(5), 543–556.

Barbanti, A., Torrado, H., Macpherson, E., Bargelloni, L., Franch, R., Carreras, C., & Pascual, M. (2020). Helping decision making for reliable and cost-effective 2b-RAD sequencing and genotyping analyses in non-model species. *Molecular Ecology Resources*, *20*(3), 795-806.

Bates, D. (2010). lme4 : linear mixed-effects models using S4 classes. *http://cran.r-project.org/package=lme4*.

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A.-M., Delourme, R., Deniot, G., Denoeud, F., Duffé, P., Engelen, S., Lemainque, A., Manzanares-Dauleux, M., Martin, G., Morice, J., Noel, B., … Aury, J.-M. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, *4*(11), 879–887.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., Aury, J.-M., Louis, A., Dehais, P., Bardou, P., Montfort, J., Klopp, C., Cabau, C., Gaspin, C., Thorgaard, G. H., … Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, *5*, 3657.

Borys, S. M., & Younger, S. T. (2020). Identification of functional regulatory elements in the human genome using pooled CRISPR screens. *BMC Genomics*, *21*(1), 107.

Campbell, C. R., Poelstra, J. W., & Yoder, A. D. (2018). What is Speciation Genomics? The roles of ecology, gene flow, and genomic architecture in the formation of

644 species. *Biological Journal of the Linnean Society. Linnean Society of London*,
645 *124*(4), 561–583.
646 Chen, Z., Omori, Y., Koren, S., Shirokiya, T., Kuroda, T., Miyamoto, A., Wada, H.,
647 Fujiyama, A., Toyoda, A., Zhang, S., Wolfsberg, T. G., Kawakami, K., Phillippy, A.
648 M., NISC Comparative Sequencing Program, Mullikin, J. C., & Burgess, S. M.
649 (2019). De novo assembly of the goldfish () genome and the evolution of genes after
650 whole-genome duplication. *Science Advances*, *5*(6), eaav0547.
651 Chueca, L. J., Schell, T., & Pfenninger, M. (2021). De novo genome assembly of the land
652 snail Candidula unifasciata (Mollusca: Gastropoda). *G3* , *11*(8), jkab180
653 Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics.
654 *Briefings in Functional Genomics*, *9*, 416–423.
655 Elkon, R., & Agami, R. (2017). Characterization of noncoding regulatory DNA in the
656 human genome. *Nature Biotechnology*, *35*(8), 732–746.
657 Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G.,
658 Guo, C., Faircloth, B. C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W.,
659 Andreu-Sánchez, S., Margaryan, A., Howard, J. T., Parent, C., … Zhang, G. (2020).
660 Dense sampling of bird diversity increases power of comparative genomics. *Nature*,
661 *587*(7833), 252–257.
662 Formenti, G., Theissinger, K., Fernandes, C., Bista, I., Bombarely, A., Bleidorn, C., Ciofi,
663 C., Crottini, A., Godoy, J. A., Höglund, J., Malukiewicz, J., Mouton, A., Oomen, R.
664 A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Svardal, H.,
665 Theofanopoulou, C., … European Reference Genome Atlas (ERGA) Consortium.
666 (2022). The era of reference genomes in conservation genomics. *Trends in Ecology*
667 *& Evolution*, *37*(3), 197–202.
668 Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly,
669 M., Gorjanc, G., Graves, S., & Others. (2012). Package “car.” *Vienna: R Foundation*
670 *for Statistical Computing*, *16*. https://cran.uni-
671 muenster.de/web/packages/car/car.pdf.
672 Francis, W. R., & Wörheide, G. (2017). Similar Ratios of Introns to Intergenic Sequence
673 across Animal Genomes. *Genome Biology and Evolution*, *9*(6), 1582–1598.
674 Galià-Camps, C., Carreras, C., Turon, X., & Pascual, M. (2022). The impact of adaptor
675 selection on genotyping in 2b-RAD studies. *Frontiers in Marine Science*, *9*, 1079839
676 Glémin, S., Clément, Y., David, J., & Ressayre, A. (2014). GC content evolution in coding
677 regions of angiosperm genomes: a unifying hypothesis. *Trends in Genetics: TIG*,
678 *30*(7), 263–270.
679 Grau-Bové, X., Ruiz-Trillo, I., & Irimia, M. (2018). Origin of exon skipping-rich
680 transcriptomes in animals driven by evolution of gene architecture. *Genome Biology*,
681 *19*(1), 135.
682 Guo, C., Ma, P.-F., Yang, G.-Q., Ye, X.-Y., Guo, Y., Liu, J.-X., Liu, Y.-L., Eaton, D. A. R.,
683 Guo, Z.-H., & Li, D.-Z. (2021). Parallel ddRAD and Genome Skimming Analyses
684 Reveal a Radiative and Reticulate Evolutionary History of the Temperate Bamboos.
685 *Systematic Biology*, *70*(4), 756–773.
686 Haerty, W., & Ponting, C. P. (2015). Unexpected selection to retain high GC content and
687 splicing enhancers within exons of multiexonic lncRNA loci. *RNA* , *21*(3), 333–346.
688 Hampton, H. G., Watson, B. N. J., & Fineran, P. C. (2020). The arms race between
689 bacteria and their phage foes. *Nature*, *577*(7790), 327–336.
690 Heckenhauer, J., Frandsen, P. B., Sproul, J. S., Li, Z., Paule, J., Larracuente, A. M.,
691 Maughan, P. J., Barker, M. S., Schneider, J. V., Stewart, R. J., & Pauls, S. U. (2022).
692 Genome size evolution in the diverse insect order Trichoptera. *GigaScience*, *11*,
693 giac011.
694 Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A. R., & Leitch, I. J.
695 (2017). Is There an Upper Limit to Genome Size? *Trends in Plant Science*, *22*(7),
696 567–573.
697 Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2021). Toward a genome sequence for
698 every animal: Where are we now? *Proceedings of the National Academy of*

Sciences of the United States of America, *118*(52), e2109019118.
Jaura, R., Yeh, S.-Y., Montanera, K. N., Ialongo, A., Anwar, Z., Lu, Y., Puwakdandawa, K., & Rhee, H. S. (2022). Extended intergenic DNA contributes to neuron-specific expression of neighboring genes in the mammalian nervous system. *Nature Communications*, *13*(1), 2733.
Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(8), E1460–E1469.
Kress, W. J., Soltis, D. E., Kersey, P. J., Wegrzyn, J. L., Leebens-Mack, J. H., Gostel, M. R., Liu, X., & Soltis, P. S. (2022). Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2115640118.
Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819.
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2020). emmeans: estimated marginal means. R package version 1.4. 4. *The American Statistician*.
Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M. L., Cai, J., Caperello, N. D., Carlson, K., Castilla-Rubio, J. C., Chaw, S.-M., Chen, L., Childers, A. K., Coddington, J. A., … Zhang, G. (2022). The Earth BioGenome Project 2020: Starting the clock. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2115635118.
Li, Z., McKibben, M. T. W., Finch, G. S., Blischak, P. D., Sutherland, B. L., & Barker, M. S. (2021). Patterns and Processes of Diploidization in Land Plants. *Annual Review of Plant Biology*, *72*, 387–410.
Li, Z., Tiley, G. P., Galuska, S. R., Reardon, C. R., Kidder, T. I., Rundell, R. J., & Barker, M. S. (2018). Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(18), 4713–4718.
Loenen, W. A. M., Dryden, D. T. F., Raleigh, E. A., Wilson, G. G., & Murray, N. E. (2014). Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Research*, *42*(1), 3–19.
López, A., Carreras, C., Pascual, M., & Pegueroles, C. (2022). Evaluating restriction enzyme selection for genome reduction in conservation genomics. *bioRxiv,* 2022-11. https://doi.org/10.1101/2022.11.26.518029
Manuzzi, A., Zane, L., Muñoz-Merida, A., Griffiths, A. M., & Veríssimo, A. (2019). Population genomics and phylogeography of a benthic coastal shark (*Scyliorhinus canicula*) using 2b-RAD single nucleotide polymorphisms. *Biological Journal of the Linnean Society. Linnean Society of London*, *126*(2), 289–303.
Marlétaz, F., Couloux, A., Poulain, J., Labadie, K., Da Silva, C., Mangenot, S., Noel, B., Poustka, A. J., Dru, P., Pegueroles, C., Borra, M., Lowe, E. K., Lhomond, G., Besnardeau, L., Le Gras, S., Ye, T., Gavriouchkina, D., Russo, R., Costa, C., … Lepage, T. (2023). Analysis of the sea urchin genome highlights contrasting trends of genomic and regulatory evolution in deuterostomes. *Cell Genomics*, *3*(4), 100295.
Marshall, J. J. T., & Halford, S. E. (2010). The Type IIB restriction endonucleases. *Biochemical Society Transactions*, *38*(2), 410–416.
Mueller, R. L., & Jockusch, E. L. (2018). Jumping genomic gigantism [Review of *Jumping genomic gigantism*]. *Nature Ecology & Evolution*, *2*(11), 1687–1688.
Naville, M., Henriet, S., Warren, I., Sumic, S., Reeve, M., Volff, J.-N., & Chourrout, D. (2019). Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. *Current Biology: CB*, *29*(7), 1161–1168.e6.
Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (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.
Platt, R. N., 2nd, Vandewege, M. W., & Ray, D. A. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *26*(1-2), 25–43.
Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biology*, *20*(1), 38.
Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , *26*(6), 841–842.
Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746.
Sacerdot, C., Louis, A., Bon, C., Berthelot, C., & Roest Crollius, H. (2018). Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biology*, *19*(1), 166.
Samson, J. E., Magadán, A. H., Sabri, M., & Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nature Reviews. Microbiology*, *11*(10), 675–687.
Schwartz, S., Meshorer, E., & Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology*, *16*(9), 990–995.
Seetharam, A. S., & Stuart, G. W. (2013). Whole genome phylogeny for 21 Drosophila species using predicted 2b-RAD fragments. *PeerJ*, *1*, e226.
Soltis, P. S., Marchant, D. B., Van de Peer, Y., & Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development*, *35*, 119–125.
Torrado, H., Carreras, C., Raventos, N., Macpherson, E., & Pascual, M. (2020). Individual-based population genomics reveal different drivers of adaptation in sympatric fish. *Scientific Reports*, *10*(1), 12683.
Wang, J., Qin, J., Sun, P., Ma, X., Yu, J., Li, Y., Sun, S., Lei, T., Meng, F., Wei, C., Li, X., Guo, H., Liu, X., Xia, R., Wang, L., Ge, W., Song, X., Zhang, L., Guo, D., … Wang, X. (2019). Polyploidy Index and Its Implications for the Evolution of Polyploids. *Frontiers in Genetics*, *10*, 807.
Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, *9*(8), 808–810.
Wendel, J. F., Jackson, S. A., Meyers, B. C., & Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biology*, *17*, 37.
Wickham, H., Chang, W., & Wickham, M. H. (2016). Package “ggplot2.” *Create Elegant Data Visualisations Using the Grammar of Graphics. Version*, *2*(1), 1–189.
Wu, C., & Lu, J. (2019). Diversification of Transposable Elements in Arthropods and Its Impact on Genome Evolution. *Genes*, *10*(5), 338.
Zhang, D. (2018). rsq: R-squared and related measures. *R Package Version*.
Zhang, J., Kuo, C. C. J., & Chen, L. (2011). GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics*, *12*, 90.
Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.-Q., & Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, *10*, 47.

Data Accessibility Statement

Supplementary data is available for this paper, including genome assemblies' accession numbers and all values needed to replicate the study (Data S1).

Benefit-Sharing Statement

No Nagoya Protocol agreement nor national Sampling Permits were necessary to conduct the study. Benefits from this research accrue from the sharing of our data and results on public files as in the 'Data Accessibility Statement'.

Author Contributions

CG: Conceptualization of the study, Data retrieval, In-silico digestion analyses, Statistical analyses, Data curator, Graphic design, Manuscript drafting, Manuscript revision.
CP: Conceptualization of the study, Genomic category analyses, Manuscript drafting, Manuscript revision.
XT: Conceptualization of the study, Manuscript drafting, Manuscript revision.
CC: Conceptualization of the study, Manuscript drafting, Manuscript revision.
MP: Conceptualization of the study, Statistical analyses, Data curator, Manuscript drafting, Manuscript revision.

Conflict of Interest

Authors declare they have no conflict of interests.

Figures

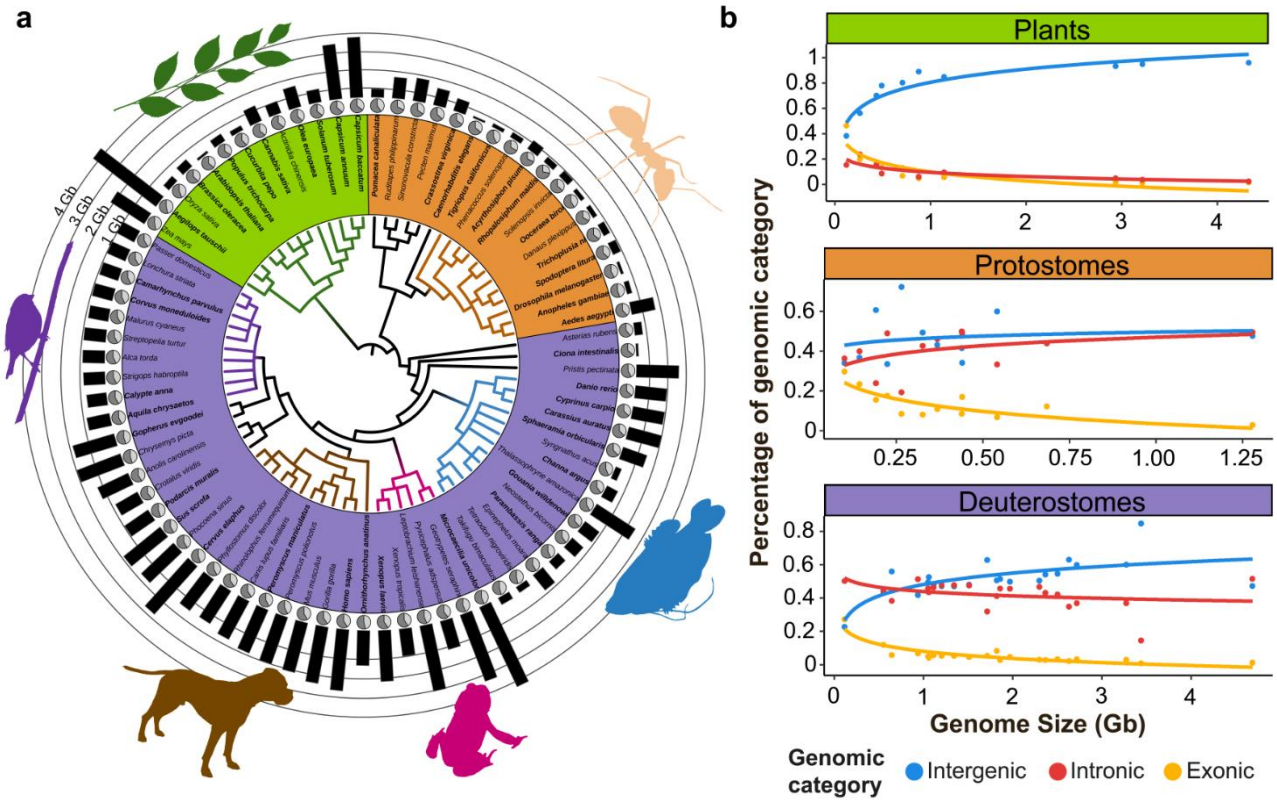
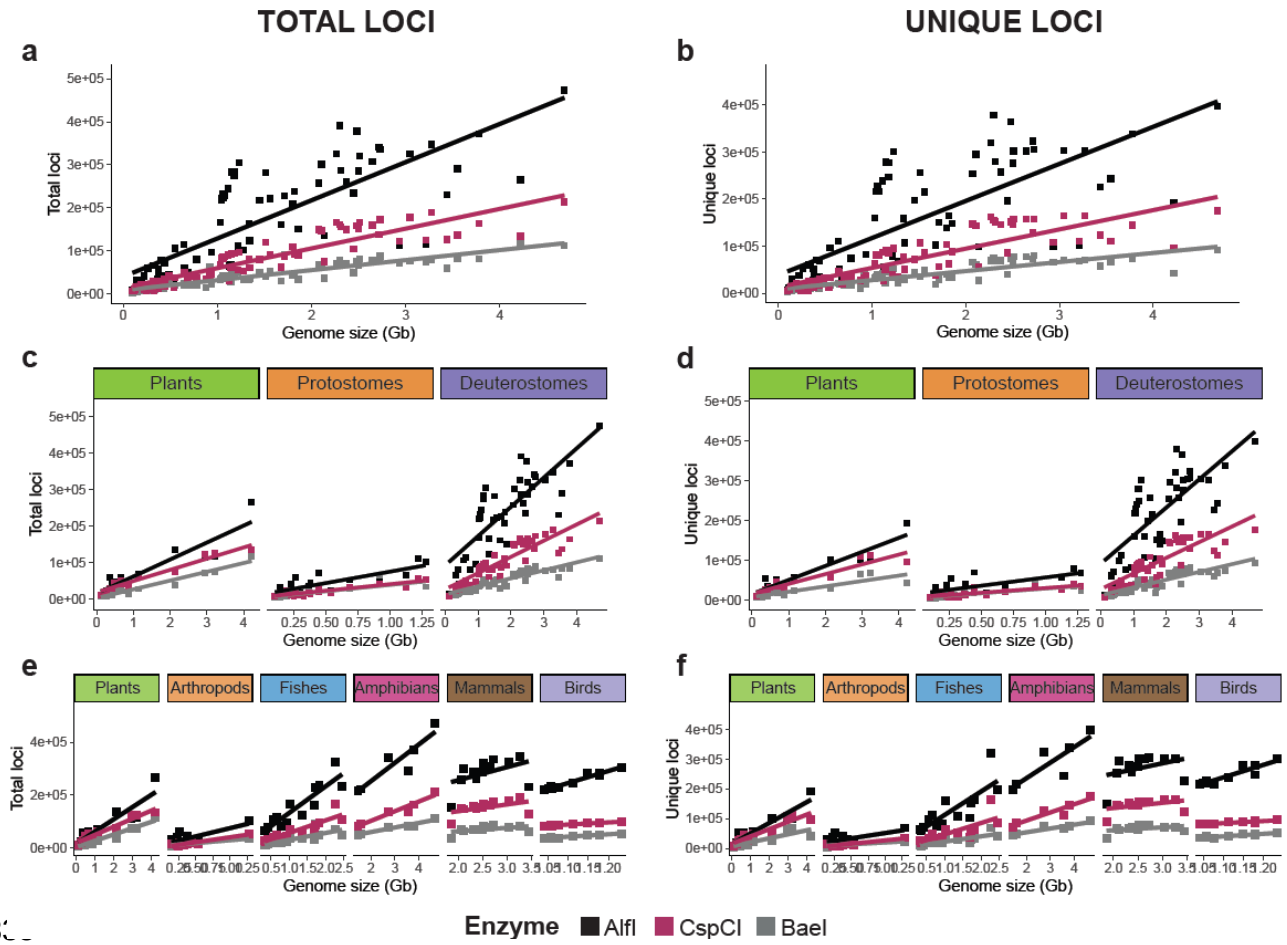


Figure 1: Phylogenetic representation of the 80 genomes used and their genomic architecture. **a:** Phylogenetic tree, in which monophyletic groups with six or more species are indicated by colored branches and identified with a shape (Plants=green, Arthropods=orange, Fishes=blue, Amphibians=pink, Mammals=brown and Birds=violet). The species names are highlighted with background color according to the three Supergroups considered (Plants=green, Protostomes=orange, Deuterostomes=purple). Species names in bold indicate genomes with annotation information. Pie diagrams indicate the GC content of each species (GC=Light gray, AT=Dark gray), and bars their genome sizes. **b:** Percentage of intergenic, intronic and exonic regions related to genome size for the 44 species with annotated genomes belonging to the three supergroups.



839

840 **Figure 2: Linear regressions of the loci yielded by each enzyme (AlfI, CspCI,**
841 **BaeI) according to each species' genome size for total and unique loci. a-**
842 **b:** Linear regressions considering all 80 genomes. **c-d:** Linear regressions for
843 each supergroup independently (plants, protostomes and deuterostomes). **e-f:**
844 Independent linear regressions for those groups with six or more species (plants,
845 arthropods, fishes, amphibians, mammals and birds). Regression equations can
846 be found in Table S4.

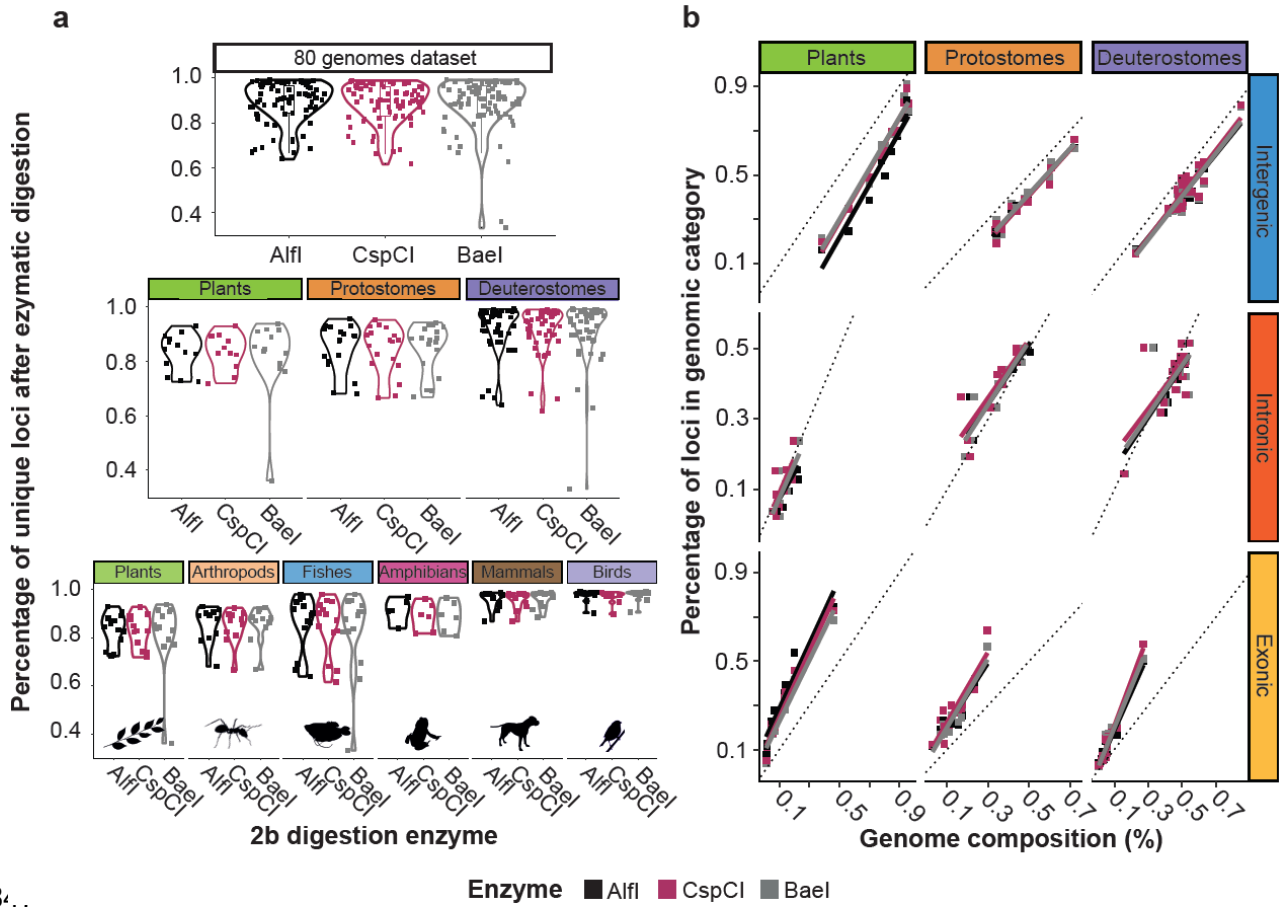


Figure 3: Typology and distribution of 2b-RAD loci in the genomes across taxa. a: Violin plots of the percentage of unique loci, in relation with the total number of loci, obtained after *in-silico* genome digestion with 2b-enzymes. The percentage of unique loci increases from plants to deuterostomes, regardless of enzyme. **b:** Percentage of loci assigned to each genomic category (intergenic, intronic and exonic) compared to the percentage of the same genomic category in the genome. Dotted lines indicate the percentage of loci in a genomic category expected under the null hypothesis of random distribution of loci.

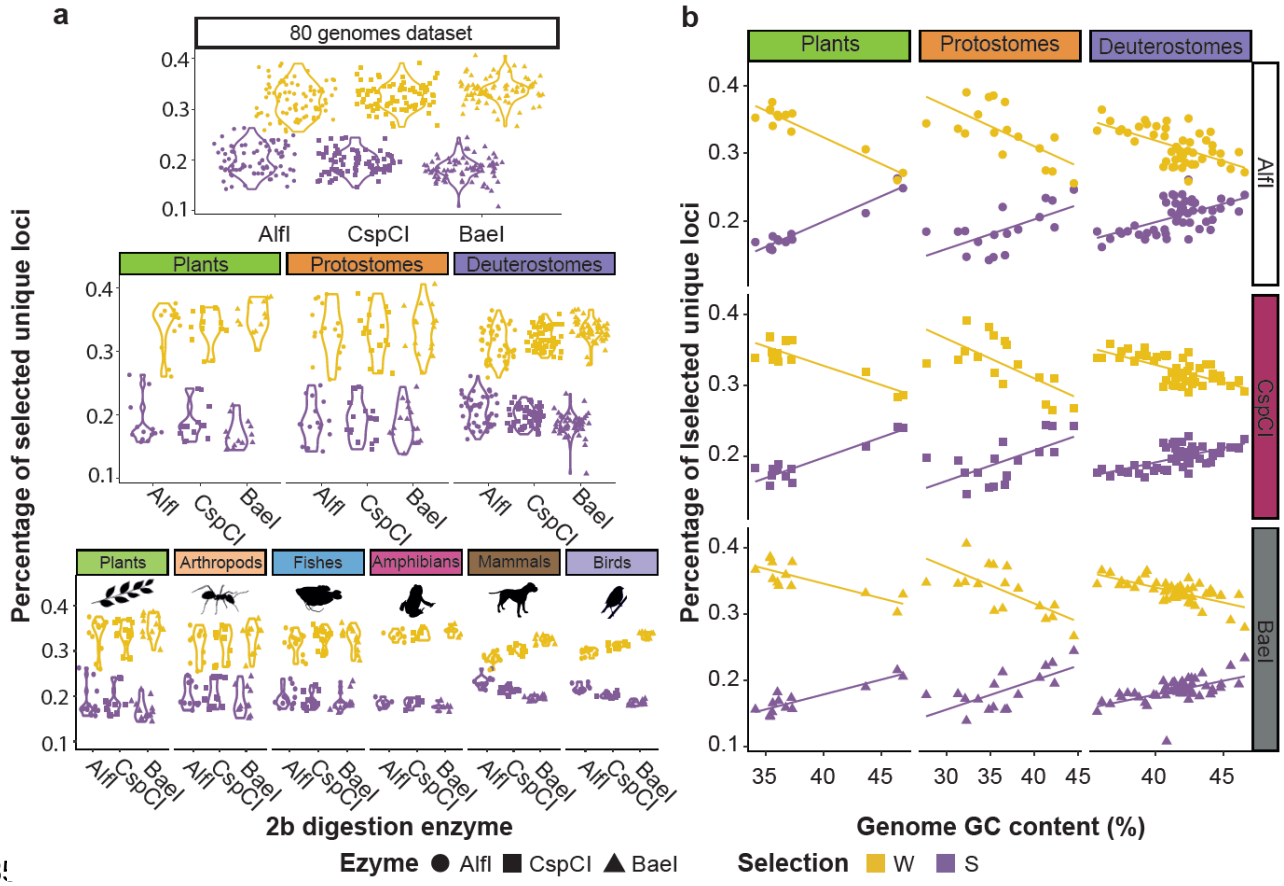
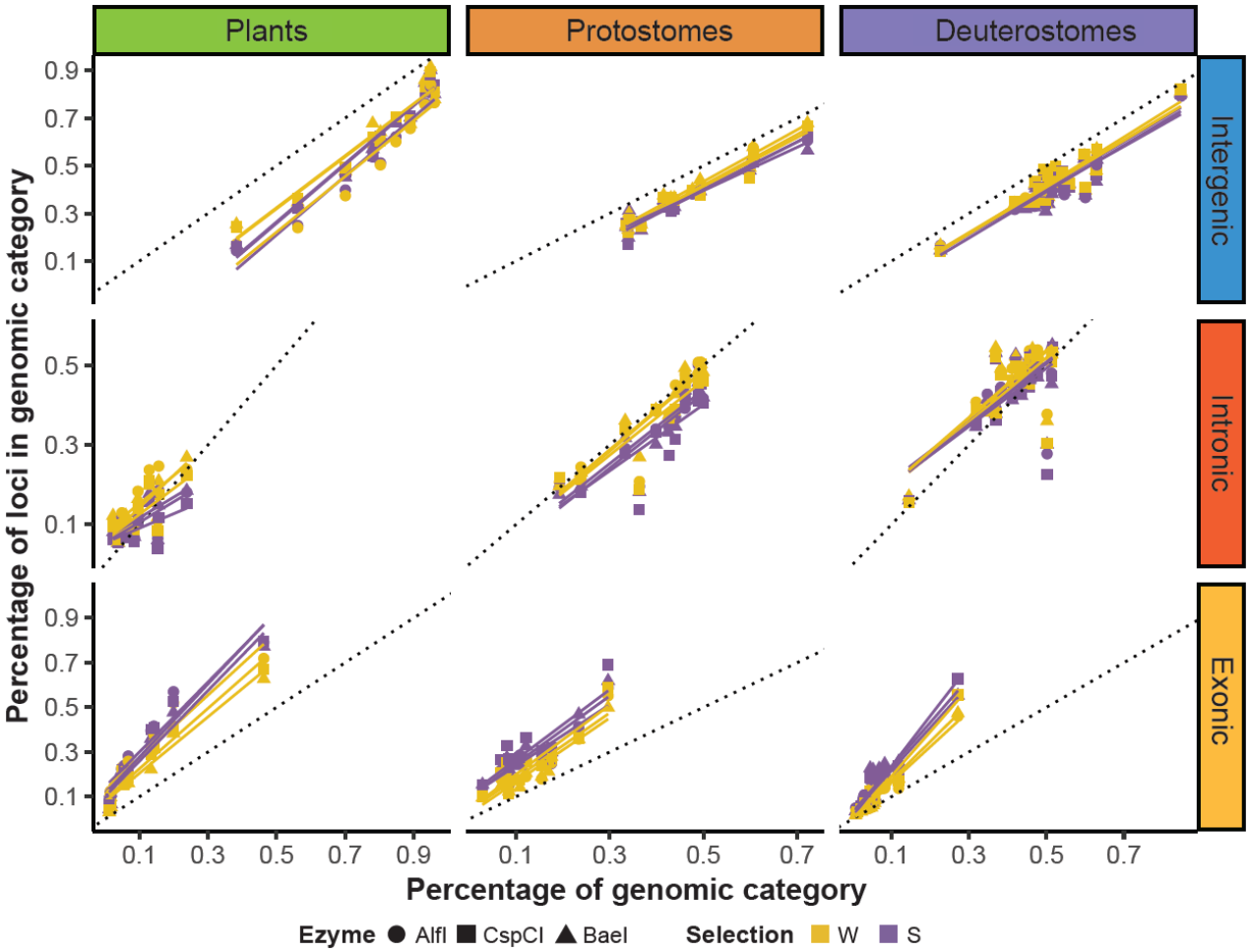


Figure 4: Typology and distribution of 2b-RAD loci in the genomes across taxa when using base selective adaptors. a: Percentage of unique loci retrieved when S (GC) and W (AT) base-selective adaptors are *in-silico* applied for each enzyme (AlfI, CspCI, BaeI). **b:** Percentage of selected loci according to genome's GC content. Linear regressions of the percentage of unique loci according to each species' genome GC content for each enzyme (AlfI, CspCI, BaeI) with *in-silico* base selection (S, W).



867

868 **Figure 5: Percentage of base-selective loci assigned to each genomic**
869 **category compared to the percentage of the same genomic category in the**
870 **genome.** Dotted lines indicate the percentage of loci in a genomic category
871 expected under the null hypothesis of random distribution of loci.

SUPPLEMENTARY TABLES AND FIGURES

Genome architecture impacts on reduced representation population genomics

Carles Galià-Camps[1,2], Cinta Pegueroles[1,2], Xavier Turon[3], Carlos Carreras[1,2], Marta Pascual[1,2]

Table S1: Logarithmic regression equations and their coefficients of determination (R^2) of the percentage of each genomic category (y) with genome size (x). The three taxonomic supergroups include only the species with annotated genomes. Significant p-values are in bold.

Supergroup	Genomic Category	Regression equation	R^2	p-value
Plants	Intergenic	y=0.806+0.150log(x)	0.87	**<0.001**
	Intronic	y=0.095-0.048log(x)	0.69	**0.003**
	Exonic	y=0.099-0.103log(x)	0.75	**0.001**
Protostomes	Intergenic	y=0.495+0.029log(x)	0.03	0.612
	Intronic	y=0.469+0.061log(x)	0.17	0.178
	Exonic	y=0.035-0.090log(x)	0.68	**0.001**
Deuterostomes	Intergenic	y=0.481+0.099log(x)	0.49	**<0.001**
	Intronic	y=0.439-0.038log(x)	0.13	0.097
	Exonic	y=0.080-0.061log(x)	0.80	**<0.001**

Table S2: General Linear Mixed-Effects Models (GLMM) of the percentage of each genomic category. Fixed factors are genomic category (intergenic, intronic, exonic), supergroup (plants, protostomes and deuterostomes) and genome size, using only the species with annotated genomes. Species is considered a random factor. For each factor, we provide their degrees of freedom (DF), chi-square (χ^2) and p-value, and coefficient of determination of the full model and their fixed factors (R^2). Significant p-values are in bold.

Factor	DF	χ2	p-value	R^2 model	R^2 fixed
Intercept	1	56.02	**<0.001**	0.93	0.93
Genomic Category	2	95.38	**<0.001**		
Supergroup	2	2.50	0.286		
Genome Size	1	46.94	**<0.001**		
Genomic Category*Supergroup	4	25.28	**<0.001**		
Genomic Category*Genome Size	2	123.67	**<0.001**		
Supergroup*Genome Size	2	2.46	0.293		
Genomic Category*Supergroup*Genome Size	4	30.57	**<0.001**		

Table S3: Tukey's post-hoc pairwise contrasts for the interaction Genomic Category*Supergroup. The column contrast indicates the factor categories being compared by the post-hoc test and the columns before contrast indicate which factors are being tested (*) or fixed. For each comparison we provide its t-ratio and p-value. Significant p-values are in bold.

Genomic Category	Supergroup	Contrast	t-ratio	p-value
Exonic	*	Plants - Protostomes	0.96	0.999
Exonic	*	Plants - Deuterostomes	-0.02	1.000
Exonic	*	Protostomes - Deuterostomes	-1.04	0.998
Intergenic	*	Plants - Protostomes	6.75	**<0.001**
Intergenic	*	Plants - Deuterostomes	12.36	**<0.001**
Intergenic	*	Protostomes - Deuterostomes	-0.14	1.000
Intronic	*	Plants - Protostomes	-8.20	**<0.001**
Intronic	*	Plants - Deuterostomes	-13.36	**<0.001**
Intronic	*	Protostomes - Deuterostomes	1.11	0.996
*	Plants	Exonic - Intergenic	-24.70	**<0.001**
*	Plants	Exonic - Intronic	-1.06	0.998
*	Plants	Intergenic - Intronic	23.64	**<0.001**
*	Protostomes	Exonic - Intergenic	-8.03	**<0.001**
*	Protostomes	Exonic - Intronic	-7.81	**<0.001**
*	Protostomes	Intergenic - Intronic	0.21	1.000
*	Deuterostomes	Exonic - Intergenic	-23.16	**<0.001**
*	Deuterostomes	Exonic - Intronic	-19.57	**<0.001**
*	Deuterostomes	Intergenic - Intronic	3.59	**0.011**

Table S4: Linear regressions on the number of total and unique loci (y) with genome size (x) considering three models: the 80 genomes altogether (Total model), split by supergroup (plants, protostomes and deuterostomes) and using only the groups with more than six species analyzed (plants, arthropods, fishes, Amphibia, mammals and birds). Note that plants, in both the supergroup and group models, include information for the same species but has been included twice to facilitate cross comparison in the two levels (supergroup and group). For each enzyme, we provide the regression equation, R^2 and p-value. Significant p-values are in bold.

Model	Dataset	Enzyme	TOTAL LOCI			UNIQUE LOCI		
			Regression equation	R^2	p-value	Regression equation	R^2	p-value
Total	80 Genomes	AlfI	y=39516.4+88857.2x	0.64	**<0.001**	y=38512.3+78742.3x	0.56	**<0.001**
		CspCI	y=13363.2+45990.9x	0.82	**<0.001**	y=12914.2+40901.9x	0.73	**<0.001**
		BaeI	y=8141.9+23321.8x	0.84	**<0.001**	y=8646.2+19230.6x	0.76	**<0.001**
Supergroup	Plants	AlfI	y=14100.4+46636.8x	0.84	**<0.001**	y=14474.8+35294.9x	0.89	**<0.001**
		CspCI	y=17441.3+30791.9x	0.95	**<0.001**	y=15360.2+24556.6x	0.87	**<0.001**
		BaeI	y=1351.5+24113.0x	0.95	**<0.001**	y=6601.6+13544.5x	0.74	**<0.001**
	Protostomes	AlfI	y=11517.6+62585.7x	0.64	**<0.001**	y=14691.2+41316.6x	0.50	**0.001**
		CspCI	y=4030.0+35755.0x	0.77	**<0.001**	y=5870.8+23690.5x	0.68	**<0.001**
		BaeI	y=2776.3+34910.2x	0.69	**<0.001**	y=4440.0+23671.1x	0.64	**<0.001**
	Deuterostomes	AlfI	y=91001.7+80646.0x	0.65	**<0.001**	y=89993.0+70892.1x	0.55	**<0.001**
		CspCI	y=25412.8+44657.3x	0.80	**<0.001**	y=26415.4+39487.1x	0.70	**<0.001**
		BaeI	y=11710.2+22039.7x	0.79	**<0.001**	y=11385.8+19545.9x	0.75	**<0.001**
Group	Plants	AlfI	y=14100.4+46636.8x	0.84	**<0.001**	y=14474.8+35294.9x	0.89	**<0.001**
		CspCI	y=17441.3+30791.9x	0.95	**<0.001**	y=15360.2+24556.6x	0.87	**<0.001**
		BaeI	y=1351.5+24113.0x	0.95	**<0.001**	y=6601.6+13544.5x	0.74	**<0.001**
	Arthropods	AlfI	y=11038.6+62495.6x	0.55	**0.01**	y=15815.8+35696.6x	0.35	0.056
		CspCI	y=1000.4+38025.5x	0.84	**<0.001**	y=4011.5+22832.6x	0.75	**0.001**
		BaeI	y=6143.9+21652.7x	0.78	**<0.001**	y=7424.8+12540.4x	0.52	**0.012**
	Fishes	AlfI	y=30979.4+101071.6x	0.83	**<0.001**	y=33929.7+78683.4x	0.65	**<0.001**
		CspCI	y=7737.7+47845.7x	0.79	**<0.001**	y=8690.0+38334.5x	0.58	**0.002**
		BaeI	y=3122.6+26218.0x	0.64	**0.00**	y=4568.8+17930.0x	0.67	**<0.001**
	Amphibians	AlfI	y=113053.3+69735.4x	0.84	**0.01**	y=131130.8+52317.7x	0.73	**0.031**
		CspCI	y=28376.3+36616.9x	0.93	**0.00**	y=38453.5+27438.5x	0.86	**0.007**
		BaeI	y=24232.0+17361.0x	0.94	**0.00**	y=28353.1+12828.4x	0.87	**0.006**
	Mammals	AlfI	y=155748.6+49977.4x	0.20	0.15	y=182134.5+34503.3x	0.13	0.248
		CspCI	y=86161.0+25885.4x	0.24	0.11	y=100447.7+17500.3x	0.16	0.197
		BaeI	y=35476.3+13169.7x	0.21	0.14	y=41153.8+9805.3x	0.15	0.218
	Birds	AlfI	y=-262154.0+460870.9x	0.95	**<0.001**	y=-228179.8+424133.5x	0.85	**<0.001**
		CspCI	y=-3285.3+82365.5x	0.96	**<0.001**	y=6260.4+71250.0x	0.78	**0.001**
		BaeI	y=-60995.7+92940.4x	0.91	**<0.001**	y=-55092.6+86644.1x	0.86	**<0.001**

Table S5: General Linear Mixed-Effects Models of the number of total and unique loci including the 80 genomes**.** Fixed factors are enzyme (AlfI, CspCI, BaeI), genome size, supergroup (plants, protostomes and deuterostomes) and group (plants, arthropods, fishes, amphibians, mammals and birds). Species are considered a random factor. Three models have been tested including all 80 species combined (Total model), separating species by supergroup (Supergroup model), and separating species by group (Group model). For each factor we provide the degrees of freedom (DF), chi-square (χ^2) and p-value, and coefficient of determination of the full model and their fixed factors (R^2). Significant p-values are in bold.

			TOTAL LOCI				UNIQUE LOCI			
Model	**Factor**	**DF**	**χ2**	**p-value**	**R^2 model**	**R^2 fixed**	**χ2**	**p-value**	**R^2 model**	**R^2 fixed**
Total	Intercept	1	73226.85	**<0.001**	0.93	0.87	54824.25	**<0.001**	0.94	0.84
	Enzyme	2	845.88	**<0.001**			879.81	**<0.001**		
	Genome Size	1	460.58	**<0.001**			350.37	**<0.001**		
	Enzyme*Genome Size	2	8.97	**0.011**			11.87	**0.003**		
Supergroup	Intercept	1	16333.53	**<0.001**	0.97	0.92	12588.44	**<0.001**	0.97	0.90
	Enzyme	2	135.03	**<0.001**			151.35	**<0.001**		
	Supergroup	2	94.41	**<0.001**			100.50	**<0.001**		
	Genome Size	1	78.07	**<0.001**			57.36	**<0.001**		
	Enzyme*Supergroup	4	137.81	**<0.001**			134.13	**<0.001**		
	Enzyme*Genome Size	2	2.27	0.322			0.16	0.922		
	Supergroup*Genome Size	2	0.72	0.697			1.71	0.425		
	Enzyme*Supergroup*Genome Size	4	7.17	0.127			7.28	0.122		
Group	Intercept	1	19875.90	**<0.001**	0.97	0.93	17237.06	**<0.001**	0.97	0.93
	Enzyme	2	154.69	**<0.001**			173.24	**<0.001**		
	Group	5	80.50	**<0.001**			90.91	**<0.001**		
	Genome Size	1	95.00	**<0.001**			78.54	**<0.001**		
	Enzyme*Group	10	98.05	**<0.001**			102.79	**<0.001**		
	Enzyme*Genome Size	2	2.60	0.273			0.19	0.911		
	Group*Genome Size	5	2.91	0.714			4.82	0.438		
	Enzyme*Group*Genome Size	10	4.15	0.940			5.83	0.830		

Table S6: Tukey's post-hoc pairwise contrasts for the interaction Enzyme*Supergroup for total and unique loci. The column contrast indicates the variables being compared with the post-hoc test and the columns before contrast indicate which factors are being tested (*) or fixed. For each comparison we provide its t-ratio and p-value. Significant p-values are in bold.

			TOTAL LOCI		UNIQUE LOCI	
Supergroup	**Enzyme**	**Contrast**	**t-ratio**	**p-value**	**t-ratio**	**p-value**
Plants	*	AlfI - CspCI	2.51	0.212	2.52	0.210
Plants	*	AlfI - BaeI	9.73	**<0.001**	10.59	**<0.001**
Plants	*	CspCI - BaeI	7.22	**<0.001**	8.08	**<0.001**
Protostomes	*	AlfI - CspCI	4.70	**<0.001**	4.79	**<0.001**
Protostomes	*	AlfI - BaeI	3.37	**0.017**	3.47	**0.012**
Protostomes	*	CspCI - BaeI	-1.33	0.975	-1.32	0.976
Deuterostomes	*	AlfI - CspCI	20.41	**<0.001**	20.35	**<0.001**
Deuterostomes	*	AlfI - BaeI	38.50	**<0.001**	38.59	**<0.001**
Deuterostomes	*	CspCI - BaeI	18.09	**<0.001**	18.24	**<0.001**
*	AlfI	Plants - Protostomes	-0.26	1.000	0.23	1.000
*	AlfI	Plants - Deuterostomes	-8.26	**<0.001**	-8.32	**<0.001**
*	AlfI	Protostomes - Deuterostomes	-5.35	**<0.001**	-5.96	**<0.001**
*	CspCI	Plants - Protostomes	2.09	0.510	2.41	0.275
*	CspCI	Plants - Deuterostomes	-2.58	0.179	-3.25	**0.027**
*	CspCI	Protostomes - Deuterostomes	-4.15	**0.001**	-4.97	**<0.001**
*	BaeI	Plants - Protostomes	-2.51	0.212	-2.10	0.499
*	BaeI	Plants - Deuterostomes	-1.69	0.828	-3.03	0.052
*	BaeI	Protostomes - Deuterostomes	1.72	0.812	0.33	1.000

Table S7: Tukey's post-hoc pairwise contrasts for the interaction Enzyme*Group for total and unique loci. The column contrast indicates the variables being compared with the post-hoc test and the columns before contrast indicate which factors are being tested (*) or fixed. For each comparison we provide its t-ratio and p-value. Significant p-values are in bold.

			TOTAL LOCI		UNIQUE LOCI	
Group	**Enzyme**	**Contrast**	**t-ratio**	**p-value**	**t-ratio**	**p-value**
Plants	*	AlfI - CspCI	2.66	0.437	2.66	0.435
Plants	*	AlfI - BaeI	10.21	**<0.001**	11.14	**<0.001**
Plants	*	CspCI - BaeI	7.55	**<0.001**	8.49	**<0.001**
Arthropods	*	AlfI - CspCI	2.75	0.355	2.79	0.323
Arthropods	*	AlfI - BaeI	1.74	0.996	1.62	0.999
Arthropods	*	CspCI - BaeI	-1.01	1.000	-1.17	1.000
Fishes	*	AlfI - CspCI	10.09	**<0.001**	10.25	**<0.001**
Fishes	*	AlfI - BaeI	17.66	**<0.001**	18.67	**<0.001**
Fishes	*	CspCI - BaeI	7.57	**<0.001**	8.42	**<0.001**
Amphibians	*	AlfI - CspCI	5.26	**<0.001**	5.30	**<0.001**
Amphibians	*	AlfI - BaeI	8.09	**<0.001**	8.14	**<0.001**
Amphibians	*	CspCI - BaeI	2.83	0.296	2.84	0.290
Mammals	*	AlfI - CspCI	2.38	0.704	2.37	0.713
Mammals	*	AlfI - BaeI	6.04	**<0.001**	6.00	**<0.001**
Mammals	*	CspCI - BaeI	3.66	0.025	3.63	**0.027**
Birds	*	AlfI - CspCI	3.14	0.129	3.13	0.133
Birds	*	AlfI - BaeI	3.99	**0.008**	3.96	**0.009**
Birds	*	CspCI - BaeI	0.85	1.000	0.83	1.000

Table S7 (Continued)

			TOTAL LOCI		UNIQUE LOCI	
Group	**Enzyme**	**Contrast**	**t-ratio**	**p-value**	**t-ratio**	**p-value**
*	AlfI	Plants - Arthropods	1.29	1.000	1.81	0.992
*	AlfI	Plants - Fishes	-5.82	**<0.001**	-5.52	**<0.001**
*	AlfI	Plants - Amphibians	-4.36	**0.002**	-4.81	**<0.001**
*	AlfI	Plants - Mammals	-3.10	0.147	-3.71	**0.022**
*	AlfI	Plants - Birds	-3.61	**0.030**	-3.71	**0.021**
*	AlfI	Arthropods - Fishes	-4.17	**0.004**	-4.54	**0.001**
*	AlfI	Arthropods - Amphibians	-4.10	**0.005**	-4.84	**<0.001**
*	AlfI	Arthropods - Mammals	-3.35	0.070	-4.20	**0.004**
*	AlfI	Arthropods - Birds	-3.90	**0.011**	-4.25	**0.003**
*	AlfI	Fishes - Amphibians	-0.94	1.000	-1.56	1.000
*	AlfI	Fishes - Mammals	-0.43	1.000	-1.18	1.000
*	AlfI	Fishes - Birds	-2.00	0.955	-2.19	0.864
*	AlfI	Amphibians - Mammals	0.25	1.000	0.03	1.000
*	AlfI	Amphibians - Birds	-1.46	1.000	-1.36	1.000
*	AlfI	Mammals - Birds	-1.53	1.000	-1.30	1.000
*	CspCl	Plants - Arthropods	2.85	0.286	3.33	0.075
*	CspCl	Plants - Fishes	-0.84	1.000	-0.68	1.000
*	CspCl	Plants - Amphibians	-0.90	1.000	-1.48	1.000
*	CspCl	Plants - Mammals	-1.79	0.993	-2.47	0.620
*	CspCl	Plants - Birds	-1.21	1.000	-1.43	1.000
*	CspCl	Arthropods - Fishes	-3.26	0.091	-3.66	**0.026**
*	CspCl	Arthropods - Amphibians	-2.98	0.205	-3.78	**0.017**
*	CspCl	Arthropods - Mammals	-3.45	0.051	-4.32	**0.002**
*	CspCl	Arthropods - Birds	-2.52	0.573	-2.96	0.217
*	CspCl	Fishes - Amphibians	-0.40	1.000	-1.08	1.000
*	CspCl	Fishes - Mammals	-1.40	1.000	-2.16	0.883
*	CspCl	Fishes - Birds	-0.97	1.000	-1.24	1.000
*	CspCl	Amphibians - Mammals	-0.92	1.000	-1.12	1.000
*	CspCl	Amphibians - Birds	-0.73	1.000	-0.69	1.000
*	CspCl	Mammals - Birds	-0.11	1.000	0.05	1.000
*	Bael	Plants - Arthropods	-0.51	1.000	-0.30	1.000
*	Bael	Plants - Fishes	-0.79	1.000	-0.68	1.000
*	Bael	Plants - Amphibians	-1.42	1.000	-2.32	0.765
*	Bael	Plants - Mammals	-0.83	1.000	-1.84	0.989
*	Bael	Plants - Birds	-1.80	0.992	-2.18	0.867
*	Bael	Arthropods - Fishes	0.13	1.000	-0.03	1.000
*	Bael	Arthropods - Amphibians	-0.56	1.000	-1.36	1.000
*	Bael	Arthropods - Mammals	-0.27	1.000	-1.22	1.000
*	Bael	Arthropods - Birds	-1.36	1.000	-1.81	0.992
*	Bael	Fishes - Amphibians	-0.96	1.000	-1.92	0.977
*	Bael	Fishes - Mammals	-0.47	1.000	-1.53	1.000
*	Bael	Fishes - Birds	-1.58	1.000	-2.00	0.957
*	Bael	Amphibians - Mammals	0.24	1.000	-0.03	1.000
*	Bael	Amphibians - Birds	-1.06	1.000	-1.02	1.000
*	Bael	Mammals - Birds	-1.15	1.000	-0.94	1.000

Table S8: General Linear Mixed-Effects Models (GLMM) of the percentage of unique loci (in relation to the total number of loci) for the 80 genomes. Fixed factors are: enzyme (AlfI, CspCI, BaeI), genome size, supergroup (plants, protostomes and deuterostomes) and group (plants, arthropods, fishes, amphibians, mammals and birds). Species are considered a random factor. Three models have been tested including all species combined (Total model), separating species by supergroup (Supergroup model), and separating species by group (Group model). For each factor we provide the degrees of freedom (DF), chi-square (χ^2) and p-value, and coefficient of determination of the full model and their fixed factors (R^2). Significant p-values are in bold.

Model	Factor	DF	χ2	p-value	R^2 model	R^2 fixed
Total	Enzyme	2	0.98	0.6141	0.85	0.00
	Genome Size	1	0.24	0.6238		
	Enzyme*Genome Size	2	1.45	0.4852		
Supergroup	Intercept	1	937.30	**<0.001**	0.91	0.23
	Enzyme	2	1.17	0.558		
	Genome Size	1	0.88	0.349		
	Supergroup	2	26.22	**<0.001**		
	Enzyme*Genome Size	2	6.86	**0.032**		
	Enzyme*Supergroup	4	1.03	0.906		
	Genome Size*Supergroup	2	3.90	0.142		
	Enzyme*Genome Size*Supergroup	4	5.23	0.265		
Group	Intercept	1	1413.66	**<0.001**	0.94	0.49
	Enzyme	2	1.65	0.438		
	Genome Size	1	1.32	0.250		
	Group	5	28.66	**<0.001**		
	Enzyme*Genome Size	2	9.71	**0.008**		
	Enzyme*Group	10	7.20	0.706		
	Genome Size*Group	5	5.33	0.377		
	Enzyme*Genome Size*Group	10	5.93	0.821		

Table S9: Tukey's post-hoc pairwise contrasts between supergroups (plants, protostomes and deuterostomes) on the percentage of unique loci. For each comparison, we provide its t-ratio and p-value. Significant p-values are in bold.

Contrast	t-ratio	p-value
Plants - Protostomes	1.22	0.539
Plants - Deuterostomes	-4.08	**<0.001**
Protostomes - Deuterostomes	-4.18	**<0.001**

Table S10: Tukey's post-hoc pairwise comparisons between groups (plants, arthropods, fishes, amphibians, mammals and birds) on the percentage of unique loci. For each comparison, we provide its t-ratio and p-value. Significant p-values are in bold.

Contrast	t-ratio	p-value
Plants - Arthropods	1.47	0.909
Plants - Fishes	-0.41	1.000
Plants - Amphibians	-2.57	0.179
Plants - Mammals	-3.76	**0.006**
Plants - Birds	-1.28	0.968
Arthropods - Fishes	-1.67	0.796
Arthropods - Amphibians	-3.00	0.060
Arthropods - Mammals	-3.99	**0.003**
Arthropods - Birds	-1.89	0.627
Fishes - Amphibians	-2.32	0.306
Fishes - Mammals	-3.57	**0.011**
Fishes - Birds	-1.17	0.986
Amphibians - Mammals	-1.49	0.898
Amphibians - Birds	-0.07	1.000
Mammals - Birds	0.87	0.999

Table S11: Linear regressions between the percentage of loci in a genomic category (y) and the percentage of the same genomic category in the genome (x). The regressions are carried out only considering the annotated genomes for each supergroup (plants, protostomes, deuterostomes), genomic category (intergenic, intronic, exonic), and enzyme (AlfI, CspCI, BaeI) independently. For each combination, we provide the regression equation, the coefficient of determination of the model (R^2) and p-values. We provide the significance of the slope differing from 1 indicated with an asterisk, since 1x is the expected value of the percentage of loci in a given category when it mirrors the percentage of the same category in the genome. Significant p-values are given in bold.

Supergroup	Genomic Category	Enzyme	Regression equation	R^2	p-value	slope ≠ 1
Plants	Intergenic	AlfI	y=-0.164+1.186x	0.94	**<0.001**	
		CspCI	y=-0.293+1.165x	0.97	**<0.001**	*
		BaeI	y=-0.273+1.146x	0.94	**<0.001**	
	Intronic	AlfI	y=0.001+0.735x	0.45	**0.033**	
		CspCI	y=-0.009+0.996x	0.53	**0.017**	
		BaeI	y=-0.020+0.955x	0.67	**0.004**	
	Exonic	AlfI	y=0.139+1.465x	0.92	**<0.001**	*
		CspCI	y=0.098+1.482x	0.95	**<0.001**	*
		BaeI	y=0.089+1.384x	0.94	**<0.001**	*
Protostomes	Intergenic	AlfI	y=-0.104+1.029x	0.98	**<0.001**	
		CspCI	y=-0.108+1.032x	0.95	**<0.001**	
		BaeI	y=-0.092+1.013x	0.94	**<0.001**	
	Intronic	AlfI	y=0.088+0.849x	0.84	**<0.001**	
		CspCI	y=0.112+0.827x	0.75	**<0.001**	
		BaeI	y=0.072+0.897x	0.88	**<0.001**	
	Exonic	AlfI	y=0.064+1.433x	0.87	**<0.001**	*
		CspCI	y=0.065+1.592x	0.80	**<0.001**	*
		BaeI	y=0.045+1.558x	0.90	**<0.001**	*
Deuterostomes	Intergenic	AlfI	y=-0.064+0.940x	0.82	**<0.001**	
		CspCI	y=-0.079+0.987x	0.84	**<0.001**	
		BaeI	y=-0.082+0.971x	0.84	**<0.001**	
	Intronic	AlfI	y=0.089+0.722x	0.58	**<0.001**	*
		CspCI	y=0.139+0.634x	0.49	**<0.001**	*
		BaeI	y=0.102+0.699x	0.56	**<0.001**	*
	Exonic	AlfI	y=0.014+1.724x	0.93	**<0.001**	*
		CspCI	y=0.002+2.055x	0.95	**<0.001**	*
		BaeI	y=0.014+1.848x	0.92	**<0.001**	*

Table S12: General Linear Mixed-Effects Models of the ratio between the percentage of loci in a genomic category and the percentage of genome in the same genomic category. Factors considered for evaluation in the GLMM are enzyme (AlfI, CspCI, BaeI), supergroup (plants, protostomes, deuterostomes), and genomic category (intergenic, intronic, exonic) and their pairwise interactions. For each factor we provide the degrees of freedom (DF), chi-square (χ^2) and p-value, and coefficient of determination of the full model and their fixed factors (R^2). Significant p-values are in bold.

Factor	DF	χ2	p-value	R^2 model	R^2 fixed
Intercept	1	1295.55	**<0.001**	0.79	0.77
Enzyme	2	18.65	**<0.001**		
Supergroup	2	84.76	**<0.001**		
Genomic Category	2	363.66	**<0.001**		
Enzyme*Supergroup	4	15.67	**0.003**		
Enzyme*Genomic Category	4	18.32	**0.001**		
Supergroup*Genomic Category	4	94.62	**<0.001**		
Enzyme*Supergroup*Genomic Category	8	14.65	0.066		

Table S13: Tukey's post-hoc pairwise contrasts for the interactions Supergroup*Enzyme, Genomic Category*Enzyme and Genomic Category*Supergroup for the ratio between the percentage of loci in a genomic category and the percentage of the same genomic category in the genome. The column contrast indicates the variables being compared with the post-hoc test and the columns before contrast indicate which factors are being tested (*) or fixed. For each comparison we provide its t-ratio and p-value. Significant p-values are in bold.

Interaction	Supergroup	Enzyme	Contrast	t-ratio	p-value
Supergroup * Enzyme	Plants	*	AlfI - CspCI	0.59	1.000
	Plants	*	AlfI - BaeI	1.66	0.843
	Plants	*	CspCI - BaeI	1.08	0.998
	Protostomes	*	AlfI - CspCI	-0.79	1.000
	Protostomes	*	AlfI - BaeI	0.09	1.000
	Protostomes	*	CspCI - BaeI	0.88	1.000
	Deuterostomes	*	AlfI - CspCI	-0.30	1.000
	Deuterostomes	*	AlfI - BaeI	-0.04	1.000
	Deuterostomes	*	CspCI - BaeI	0.26	1.000
	*	AlfI	Plants - Protostomes	1.74	0.790
	*	AlfI	Plants - Deuterostomes	2.77	0.107
	*	AlfI	Protostomes - Deuterostomes	0.86	1.000
	*	CspCI	Plants - Protostomes	0.50	1.000
	*	CspCI	Plants - Deuterostomes	1.93	0.641
	*	CspCI	Protostomes - Deuterostomes	1.45	0.944
	*	BaeI	Plants - Protostomes	0.24	1.000
	*	BaeI	Plants - Deuterostomes	0.97	0.999
	*	BaeI	Protostomes - Deuterostomes	0.74	1.000

Table S13: (Continuation)

	Genomic Category	Enzyme	Contrast	t-ratio	p-value
Genomic Category * Enzyme	Exonic	*	AlfI - CspCI	1.51	0.922
	Exonic	*	AlfI - BaeI	2.91	0.068
	Exonic	*	CspCI - BaeI	1.40	0.960
	Intergenic	*	AlfI - CspCI	-0.62	1.000
	Intergenic	*	AlfI - BaeI	-0.66	1.000
	Intergenic	*	CspCI - BaeI	-0.04	1.000
	Intronic	*	AlfI - CspCI	-1.28	0.983
	Intronic	*	AlfI - BaeI	-0.28	1.000
	Intronic	*	CspCI - BaeI	0.99	0.999
	*	AlfI	Exonic - Intergenic	21.52	**<0.001**
	*	AlfI	Exonic - Intronic	18.60	**<0.001**
	*	AlfI	Intergenic - Intronic	-2.92	0.066
	*	CspCI	Exonic - Intergenic	19.39	**<0.001**
	*	CspCI	Exonic - Intronic	15.81	**<0.001**
	*	CspCI	Intergenic - Intronic	-3.57	**0.007**
	*	BaeI	Exonic - Intergenic	17.95	**<0.001**
	*	BaeI	Exonic - Intronic	15.41	**<0.001**
	*	BaeI	Intergenic - Intronic	-2.54	0.189
	Genomic Category	**Supergroup**	**Contrast**	**t-ratio**	**p-value**
Genomic Category * Supergroup	Exonic	*	Plants - Protostomes	8.26	**<0.001**
	Exonic	*	Plants - Deuterostomes	9.77	**<0.001**
	Exonic	*	Protostomes - Deuterostomes	0.52	1.000
	Intergenic	*	Plants - Protostomes	-1.66	0.846
	Intergenic	*	Plants - Deuterostomes	-2.12	0.474
	Intergenic	*	Protostomes - Deuterostomes	-0.28	1.000
	Intronic	*	Plants - Protostomes	-4.12	**0.001**
	Intronic	*	Plants - Deuterostomes	-1.97	0.604
	Intronic	*	Protostomes - Deuterostomes	2.82	0.093
	*	Plants	Exonic - Intergenic	24.34	**<0.001**
	*	Plants	Exonic - Intronic	22.52	**<0.001**
	*	Plants	Intergenic - Intronic	-1.81	0.733
	*	Protostomes	Exonic - Intergenic	15.21	**<0.001**
	*	Protostomes	Exonic - Intronic	10.38	**<0.001**
	*	Protostomes	Intergenic - Intronic	-4.82	**<0.001**
	*	Deuterostomes	Exonic - Intergenic	19.55	**<0.001**
	*	Deuterostomes	Exonic - Intronic	17.07	**<0.001**
	*	Deuterostomes	Intergenic - Intronic	-2.48	0.220

Table S14: General Linear Mixed-Effects Models of the percentage of selected unique loci after secondary reduction including the 80 genomes**.** Fixed factors are enzyme (AlfI, CspCI, BaeI), selection (S, W), genome size, supergroup (plants, protostomes, deuterostomes), and group (plants, arthropods, fishes, amphibians, mammals and birds). Species are considered a random factor. Three models have been tested including all species combined (Total model), separating species by supergroup (Supergroup model), and separating species by group (Group model). For each factor we provide the degrees of freedom (DF), chi-square (χ^2) and p-value, and coefficient of determination of the full model and their fixed factors (R^2). Significant p-values are in bold.

Model	Factor	DF	χ2	p-value	R^2 model	R^2 fixed
Total	Intercept	1	18650.34	**<0.001**	0.87	0.87
	Enzyme	2	22.41	**<0.001**		
	Selection	1	829.63	**<0.001**		
	Genome Size	1	4.60	**0.032**		
	Enzyme*Selection	2	38.72	**<0.001**		
	Enzyme*Genome Size	2	5.03	0.081		
	Selection*Genome Size	1	7.04	**0.008**		
	Enzyme*Selection*Genome Size	2	8.91	**0.012**		
Supergroup	Intercept	1	3428.80	**<0.001**	0.90	0.90
	Enzyme	2	5.83	0.054		
	Selection	1	253.03	**<0.001**		
	Supergroup	2	23.79	**<0.001**		
	Genome Size	1	2.96	0.086		
	Enzyme*Selection	2	9.60	**0.008**		
	Enzyme*Supergroup	4	2.71	0.608		
	Selection*Supergroup	2	50.01	**<0.001**		
	Enzyme*Genome Size	2	1.22	0.544		
	Selection*Genome Size	1	5.52	**0.019**		
	Supergroup*Genome Size	2	17.74	**<0.001**		
	Enzyme*Selection*Supergroup	4	4.71	0.318		
	Enzyme*Selection*Genome Size	2	2.41	0.300		
	Enzyme*Supergroup*Genome Size	4	0.56	0.967		
	Selection*Supergroup*Genome Size	2	36.78	**<0.001**		
	Enzyme*Selection*Supergroup*Genome Size	4	1.10	0.894		
Group	Intercept	1	5059.08	**<0.001**	0.92	0.92
	Enzyme	2	8.61	**0.014**		
	Selection	1	373.33	**<0.001**		
	Group	5	8.59	0.127		
	Genome Size	1	4.36	**0.037**		
	Enzyme*Selection	2	14.16	**0.001**		
	Enzyme*Group	10	4.38	0.929		
	Selection*Group	5	18.71	**0.002**		
	Enzyme*Genome Size	2	1.80	0.407		
	Selection*Genome Size	1	8.15	**0.004**		
	Group*Genome Size	5	19.54	**0.002**		
	Enzyme*Selection*Group	10	8.21	0.609		
	Enzyme*Selection*Genome Size	2	3.55	0.169		
	Enzyme*Group*Genome Size	10	0.82	1.000		
	Selection*Group*Genome Size	5	40.14	**<0.001**		
	Enzyme*Selection*Group*Genome Size	10	1.72	0.998		

Table S15: Tukey's post-hoc pairwise contrasts for the interaction Selection*Enzyme for the percentage of selected unique loci after secondary reduction on the Total model. The column contrast indicates the variables being compared with the post-hoc test and the columns before contrast indicate which factors are being tested (*) or fixed. For each comparison we provide its t-ratio and p-value. Significant p-values are in bold.

Selection	Enzyme	Contrast	t-ratio	p-value
S	*	AlfI - CspCI	1.57	0.679
S	*	AlfI - BaeI	5.00	**<0.001**
S	*	CspCI - BaeI	3.43	**0.006**
W	*	AlfI - CspCI	-1.89	0.427
W	*	AlfI - BaeI	-4.45	**<0.001**
W	*	CspCI - BaeI	-2.56	0.094
*	AlfI	S - W	-25.97	**<0.001**
*	CspCI	S - W	-29.42	**<0.001**
*	BaeI	S - W	-35.41	**<0.001**

Table S16: Tukey's post-hoc pairwise contrasts for the interactions Selection*Enzyme and Selection*Supergroup for the percentage of selected unique loci after secondary reduction on the Supergroup model. The column contrast indicates the variables being compared with the post-hoc test and the columns before contrast indicate which factors are being tested (*) or fixed. For each comparison we provide its t-ratio and p-value. Significant p-values are in bold.

Interaction	Selection	Enzyme	Contrast	t-ratio	p-value
Selection*Enzyme	S	*	AlfI - CspCI	0.34	1.000
	S	*	AlfI - BaeI	2.55	0.096
	S	*	CspCI - BaeI	2.22	0.221
	W	*	AlfI - CspCI	-0.70	0.997
	W	*	AlfI - BaeI	-2.27	0.197
	W	*	CspCI - BaeI	-1.56	0.680
	*	AlfI	S - W	-23.47	**<0.001**
	*	CspCI	S - W	-24.51	**<0.001**
	*	BaeI	S - W	-28.29	**<0.001**
	Selection	**Supergroup**	**Contrast**	**t-ratio**	**p-value**
Selection*Supergroup	S	*	Plants - Protostomes	3.90	**0.001**
	S	*	Plants - Deuterostomes	-3.07	**0.021**
	S	*	Protostomes - Deuterostomes	-6.54	**<0.001**
	W	*	Plants - Protostomes	-3.90	**0.001**
	W	*	Plants - Deuterostomes	3.76	**0.002**
	W	*	Protostomes - Deuterostomes	7.01	**<0.001**
	*	Plants	S - W	-26.25	**<0.001**
	*	Protostomes	S - W	-23.51	**<0.001**
	*	Deuterostomes	S - W	-46.13	**<0.001**

Table S17: Tukey's post-hoc pairwise contrasts for the interactions Selection*Enzyme and Selection*Group for the percentage of selected unique loci after secondary reduction on the Group model. The column contrast indicates the variables being compared with the post-hoc test and the columns before contrast indicate which factors are being tested (*) or fixed. For each comparison we provide its t-ratio and p-value. Significant p-values are in bold.

Interaction	Selection	Enzyme	Contrast	t-ratio	p-value
Selection*Enzyme	S	*	AlfI - CspCI	0.99	0.970
	S	*	AlfI - BaeI	2.41	0.142
	S	*	CspCI - BaeI	1.41	0.789
	W	*	AlfI - CspCI	-1.14	0.930
	W	*	AlfI - BaeI	-2.18	0.242
	W	*	CspCI - BaeI	-1.04	0.959
	*	AlfI	S - W	-11.32	**<0.001**
	*	CspCI	S - W	-13.45	**<0.001**
	*	BaeI	S - W	-15.91	**<0.001**

Table S17: (Continued)

Interaction	Selection	Group	Contrast	t-ratio	p-value
Selection*Group	S	*	Plants - Arthropods	1.80	0.937
	S	*	Plants - Fishes	0.38	1.000
	S	*	Plants - Amphibians	1.14	1.000
	S	*	Plants - Mammals	-3.81	**0.007**
	S	*	Plants - Birds	-2.58	0.323
	S	*	Arthropods - Fishes	-1.61	0.985
	S	*	Arthropods - Amphibians	-0.70	1.000
	S	*	Arthropods - Mammals	-4.26	**0.001**
	S	*	Arthropods - Birds	-3.22	0.053
	S	*	Fishes - Amphibians	0.91	1.000
	S	*	Fishes - Mammals	-3.98	**0.004**
	S	*	Fishes - Birds	-2.68	0.252
	S	*	Amphibians - Mammals	-3.97	**0.004**
	S	*	Amphibians - Birds	-2.91	0.138
	S	*	Mammals - Birds	-0.25	1.000
	W	*	Plants - Arthropods	-1.68	0.973
	W	*	Plants - Fishes	0.32	1.000
	W	*	Plants - Amphibians	-0.88	1.000
	W	*	Plants - Mammals	3.80	**0.007**
	W	*	Plants - Birds	2.60	0.305
	W	*	Arthropods - Fishes	1.83	0.923
	W	*	Arthropods - Amphibians	0.77	1.000
	W	*	Arthropods - Mammals	4.17	**0.002**
	W	*	Arthropods - Birds	3.18	0.060
	W	*	Fishes - Amphibians	-1.07	1.000
	W	*	Fishes - Mammals	3.65	**0.012**
	W	*	Fishes - Birds	2.51	0.372
	W	*	Amphibians - Mammals	3.80	**0.007**
	W	*	Amphibians - Birds	2.82	0.176
	W	*	Mammals - Birds	0.27	1.000
	*	Plants	S - W	-31.28	**<0.001**
	*	Arthropods	S - W	-14.24	**<0.001**
	*	Fishes	S - W	-31.09	**<0.001**
	*	Amphibians	S - W	-15.82	**<0.001**
	*	Mammals	S - W	-4.98	**<0.001**
	*	Birds	S - W	-2.48	0.392

Table S18: Linear regressions on the number of percentage of selected loci by W and S adaptors (y) with GC content (x) considering independently each enzyme (AlfI, CspCI, BaeI) used on each supergroup (plants, protostomes and deuterostomes). For each enzyme, we provide the regression equation, R^2 and p-value. Significant p-values are in bold.

		W-selection			S-selection		
Supergroup	**Enzyme**	**Regression equation**	**R^2**	**p-value**	**Regression equation**	**R^2**	**p-value**
Plants	**AlfI**	y=0.639-0.008x	0.90	**0.000**	y=-0.1+0.007x	0.93	**0.000**
	CspCI	y=0.545-0.005x	0.73	**0.000**	y=-0.033+0.006x	0.83	**0.000**
	BaeI	y=0.523-0.004x	0.62	**0.001**	y=-0.002+0.004x	0.80	**0.000**
Protostomes	**AlfI**	y=0.548-0.006x	0.45	**0.003**	y=0.028+0.004x	0.38	**0.008**
	CspCI	y=0.534-0.006x	0.44	**0.004**	y=0.036+0.004x	0.39	**0.008**
	BaeI	y=0.536-0.006x	0.47	**0.002**	y=0.023+0.004x	0.47	**0.003**
Deuterostomes	**AlfI**	y=0.562-0.006x	0.36	**0.000**	y=-0.018+0.005x	0.33	**0.000**
	CspCI	y=0.526-0.005x	0.45	**0.000**	y=0.029+0.004x	0.41	**0.000**
	BaeI	y=0.522-0.005x	0.46	**0.000**	y=0.012+0.004x	0.32	**0.000**

Table S19: Linear regressions between the percentage of unique loci in a genomic category (y) and the percentage of the same category in the genome (x) for the annotated genomes. Each supergroup (plants, protostomes and deuterostomes), genomic category (intergenic, intronic, exonic), enzyme (AlfI, CspCI, BaeI) and selection (W-selection, S-selection) has been considered independently. For each combination, we provide the regression equation, the coefficient of determination (R^2) and p-values. Significant p-values are given in bold. Slopes significantly different from one are indicated with an asterisk, since 1x is the expected value of the percentage of loci in a given category when it mirrors the percentage of the same category in the genome.

Supergroup	Genomic Category	Enzyme	W-selection				S-selection			
			Regression equation	R^2	p-value	slope ≠ 1	Regression equation	R^2	p-value	slope ≠ 1
Plants	Intergenic	AlfI	y=-0.362+1.167x	0.93	**<0.001**		y=-0.407+1.235x	0.95	**<0.001**	*
		CspCI	y=-0.226+1.091x	0.95	**<0.001**		y=-0.358+1.232x	0.97	**<0.001**	*
		BaeI	y=-0.222+1.091x	0.92	**<0.001**		y=-0.349+1.226x	0.96	**<0.001**	*
	Intronic	AlfI	y=0.083+0.701x	0.51	**0.020**		y=0.065+0.520x	0.39	0.056	*
		CspCI	y=0.061+0.648x	0.58	**0.011**		y=0.055+0.364x	0.33	0.081	*
		BaeI	y=0.061+0.814x	0.70	**0.003**		y=0.054+0.517x	0.50	**0.023**	*
	Exonic	AlfI	y=0.133+1.405x	0.92	**<0.001**	*	y=0.143+1.566x	0.92	**<0.001**	*
		CspCI	y=0.09+1.351x	0.95	**<0.001**	*	y=0.114+1.635x	0.93	**<0.001**	*
		BaeI	y=0.078+1.263x	0.94	**<0.001**	*	y=0.101+1.580x	0.94	**<0.001**	*
Protostomes	Intergenic	AlfI	y=-0.114+1.064x	0.97	**<0.001**		y=-0.090+0.987x	0.98	**<0.001**	
		CspCI	y=-0.099+1.025x	0.94	**<0.001**		y=-0.109+1.013x	0.95	**<0.001**	
		BaeI	y=-0.109+1.087x	0.96	**<0.001**		y=-0.059+0.910x	0.91	**<0.001**	
	Intronic	AlfI	y=-0.019+1.01x	0.82	**<0.001**		y=-0.026+0.929x	0.80	**<0.001**	
		CspCI	y=-0.008+0.942x	0.78	**<0.001**		y=-0.015+0.833x	0.67	**0.001**	
		BaeI	y=-0.012+1x	0.90	**<0.001**		y=-0.039+0.937x	0.80	**<0.001**	
	Exonic	AlfI	y=0.037+1.459x	0.85	**<0.001**	*	y=0.100+1.372x	0.82	**<0.001**	
		CspCI	y=0.045+1.552x	0.84	**<0.001**	*	y=0.108+1.570x	0.74	**<0.001**	
		BaeI	y=0.023+1.431x	0.90	**<0.001**	*	y=0.096+1.515x	0.83	**<0.001**	*
Deuterostomes	Intergenic	AlfI	y=-0.07+0.967x	0.84	**<0.001**		y=-0.061+0.917x	0.80	**<0.001**	
		CspCI	y=-0.088+1.013x	0.86	**<0.001**		y=-0.079+0.972x	0.83	**<0.001**	
		BaeI	y=-0.079+0.978x	0.85	**<0.001**		y=-0.095+0.973x	0.81	**<0.001**	
	Intronic	AlfI	y=0.115+0.848x	0.65	**<0.001**		y=0.136+0.745x	0.47	**<0.001**	
		CspCI	y=0.114+0.81x	0.55	**<0.001**		y=0.127+0.731x	0.41	**0.001**	
		BaeI	y=0.115+0.841x	0.62	**<0.001**		y=0.122+0.764x	0.49	**<0.001**	
	Exonic	AlfI	y=0.004+1.595x	0.94	**<0.001**	*	y=0.026+1.911x	0.91	**<0.001**	*
		CspCI	y=-0.006+1.977x	0.96	**<0.001**	*	y=0.011+2.249x	0.93	**<0.001**	*
		BaeI	y=0.003+1.717x	0.94	**<0.001**	*	y=0.032+1.998x	0.86	**<0.001**	*

Table S20: General Linear Mixed-Effects Models for the ratio between the percentage of loci in a genomic category and the percentage of the same genomic category in the genome after selection using the annotated genomes**.** Fixed factors are enzyme (AlfI, CspCI, BaeI), selection (S, W), supergroup (plants, protostomes, deuterostomes), and genomic category (intergenic, intronic, exonic). For each factor we provide the degrees of freedom (DF), chi-square (χ^2) and p-value, and coefficient of determination of the full model and their fixed factors (R^2). Significant p-values are in bold.

Factor	DF	χ2	p-value	R2 model	R2 fixed
Intercept	1	1151.05	**<0.001**		
Enzyme	2	15.45	**<0.001**		
Selection	1	2.34	0.126		
Supergroup	2	52.81	**<0.001**		
Genomic Category	2	467.27	**<0.001**		
Enzyme*Selection	2	2.23	0.328		
Enzyme*Supergroup	4	18.95	**0.001**		
Enzyme*Genomic Category	4	13.93	**0.008**	0.86	0.75
Selection*Supergroup	2	6.41	**0.041**		
Selection*Genomic Category	2	9.70	**0.008**		
Supergroup*Genomic Category	4	74.71	**<0.001**		
Selection*Supergroup*Genomic Category	4	4.70	0.319		
Enzyme*Selection*Supergroup	4	3.24	0.518		
Enzyme*Supergroup*Genomic Category	8	16.06	**0.042**		
Enzyme*Selection*Genomic Category	4	1.96	0.744		
Enzyme*Selection*Supergroup*Genomic Category	8	2.82	0.945		

Table S21: Tukey's post-hoc pairwise contrasts on the 2-way interactions (selection*genomic category, selection*supergroup). The column contrast indicates the variables being compared with the post-hoc test and the first two columns indicate which factors are fixed (the factor being tested is represented with asterisk). For each comparison we provide its p-value. Significant p-values are in bold.

Interaction	Genomic Category	Selection	Contrast	t-ratio	p-value
Selection*Genomic Category	Exonic	*	S - W	12.27	**<0.001**
	Intergenic	*	S - W	-1.12	0.935
	Intronic	*	S - W	-4.96	**<0.001**
	*	S	Exonic - Intergenic	48.18	**<0.001**
	*	S	Exonic - Intronic	38.80	**<0.001**
	*	S	Intergenic - Intronic	-9.38	**<0.001**
	*	W	Exonic - Intergenic	34.79	**<0.001**
	*	W	Exonic - Intronic	21.57	**<0.001**
	*	W	Intergenic - Intronic	-13.22	**<0.001**
	Supergroup	**Selection**	**Contrast**	**t-ratio**	**p-value**
Selection*Supergroup	Plants	*	S - W	-0.28	1.000
	Protostomes	*	S - W	2.41	0.137
	Deuterostomes	*	S - W	5.17	**<0.001**
	*	S	Plants - Protostomes	2.76	0.072
	*	S	Plants - Deuterostomes	2.19	0.267
	*	S	Protostomes - Deuterostomes	-0.97	0.975
	*	W	Plants - Protostomes	3.66	**0.006**
	*	W	Plants - Deuterostomes	3.72	**0.005**
	*	W	Protostomes - Deuterostomes	-0.41	1.000

Table S22: Tukey's post-hoc pairwise contrasts on the significant 3-way interaction (enzyme*supergroup*genomic category). The column contrast indicates the variables being compared with the post-hoc test and the first three columns indicate which factors are fixed (the factor being tested is represented with asterisk). For each comparison we provide its p-value. Significant p-values are in bold.

Supergroup	Genomic Category	Enzyme	Contrast	t-ratio	p-value
Plants	*	AlfI	Exonic - Intergenic	28.83	**<0.001**
Plants	*	AlfI	Exonic - Intronic	17.50	**<0.001**
Plants	*	AlfI	Intergenic - Intronic	-11.34	**<0.001**
Plants	*	CspCI	Exonic - Intergenic	22.67	**<0.001**
Plants	*	CspCI	Exonic - Intronic	15.65	**<0.001**
Plants	*	CspCI	Intergenic - Intronic	-7.02	**<0.001**
Plants	*	BaeI	Exonic - Intergenic	20.47	**<0.001**
Plants	*	BaeI	Exonic - Intronic	11.94	**<0.001**
Plants	*	BaeI	Intergenic - Intronic	-8.53	**<0.001**
Protostomes	*	AlfI	Exonic - Intergenic	14.70	**<0.001**
Protostomes	*	AlfI	Exonic - Intronic	12.90	**<0.001**
Protostomes	*	AlfI	Intergenic - Intronic	-1.79	0.998
Protostomes	*	CspCI	Exonic - Intergenic	16.75	**<0.001**
Protostomes	*	CspCI	Exonic - Intronic	15.57	**<0.001**
Protostomes	*	CspCI	Intergenic - Intronic	-1.18	1.000
Protostomes	*	BaeI	Exonic - Intergenic	14.24	**<0.001**
Protostomes	*	BaeI	Exonic - Intronic	12.77	**<0.001**
Protostomes	*	BaeI	Intergenic - Intronic	-1.47	1.000
Deuterostomes	*	AlfI	Exonic - Intergenic	19.49	**<0.001**
Deuterostomes	*	AlfI	Exonic - Intronic	13.77	**<0.001**
Deuterostomes	*	AlfI	Intergenic - Intronic	-5.72	**<0.001**
Deuterostomes	*	CspCI	Exonic - Intergenic	19.58	**<0.001**
Deuterostomes	*	CspCI	Exonic - Intronic	14.84	**<0.001**
Deuterostomes	*	CspCI	Intergenic - Intronic	-4.74	**<0.001**
Deuterostomes	*	BaeI	Exonic - Intergenic	20.29	**<0.001**
Deuterostomes	*	BaeI	Exonic - Intronic	14.56	**<0.001**
Deuterostomes	*	BaeI	Intergenic - Intronic	-5.73	**<0.001**

Table S22: (Continued)

Supergroup	Genomic Category	Enzyme	Contrast	t-ratio	p-value
*	Exonic	AlfI	Plants - Protostomes	8.20	**<0.001**
*	Exonic	AlfI	Plants - Deuterostomes	9.29	**<0.001**
*	Exonic	AlfI	Protostomes - Deuterostomes	0.09	1.000
*	Exonic	CspCI	Plants - Protostomes	4.03	**0.009**
*	Exonic	CspCI	Plants - Deuterostomes	5.68	**<0.001**
*	Exonic	CspCI	Protostomes - Deuterostomes	1.23	1.000
*	Exonic	BaeI	Plants - Protostomes	3.96	**0.012**
*	Exonic	BaeI	Plants - Deuterostomes	4.05	**0.009**
*	Exonic	BaeI	Protostomes - Deuterostomes	-0.43	1.000
*	Intergenic	AlfI	Plants - Protostomes	-1.76	0.999
*	Intergenic	AlfI	Plants - Deuterostomes	-2.09	0.962
*	Intergenic	AlfI	Protostomes - Deuterostomes	-0.13	1.000
*	Intergenic	CspCI	Plants - Protostomes	-0.74	1.000
*	Intergenic	CspCI	Plants - Deuterostomes	-1.19	1.000
*	Intergenic	CspCI	Protostomes - Deuterostomes	-0.38	1.000
*	Intergenic	BaeI	Plants - Protostomes	-0.86	1.000
*	Intergenic	BaeI	Plants - Deuterostomes	-0.88	1.000
*	Intergenic	BaeI	Protostomes - Deuterostomes	0.10	1.000
*	Intronic	AlfI	Plants - Protostomes	4.51	**0.002**
*	Intronic	AlfI	Plants - Deuterostomes	3.34	0.094
*	Intronic	AlfI	Protostomes - Deuterostomes	-1.83	0.997
*	Intronic	CspCI	Plants - Protostomes	3.10	0.187
*	Intronic	CspCI	Plants - Deuterostomes	1.58	1.000
*	Intronic	CspCI	Protostomes - Deuterostomes	-2.02	0.979
*	Intronic	BaeI	Plants - Protostomes	3.78	**0.022**
*	Intronic	BaeI	Plants - Deuterostomes	2.51	0.678
*	Intronic	BaeI	Protostomes - Deuterostomes	-1.85	0.997

Table S22: (Continued)

Supergroup	Genomic Category	Enzyme	Contrast	t-ratio	p-value
Plants	Exonic	*	AlfI - CspCI	4.74	**<0.001**
Plants	Exonic	*	AlfI - BaeI	6.82	**<0.001**
Plants	Exonic	*	CspCI - BaeI	2.08	0.957
Protostomes	Exonic	*	AlfI - CspCI	-1.89	0.993
Protostomes	Exonic	*	AlfI - BaeI	0.28	1.000
Protostomes	Exonic	*	CspCI - BaeI	2.16	0.922
Deuterostomes	Exonic	*	AlfI - CspCI	-0.36	1.000
Deuterostomes	Exonic	*	AlfI - BaeI	-0.61	1.000
Deuterostomes	Exonic	*	CspCI - BaeI	-0.26	1.000
Plants	Intergenic	*	AlfI - CspCI	-1.42	1.000
Plants	Intergenic	*	AlfI - BaeI	-1.54	1.000
Plants	Intergenic	*	CspCI - BaeI	-0.12	1.000
Protostomes	Intergenic	*	AlfI - CspCI	0.16	1.000
Protostomes	Intergenic	*	AlfI - BaeI	-0.18	1.000
Protostomes	Intergenic	*	CspCI - BaeI	-0.34	1.000
Deuterostomes	Intergenic	*	AlfI - CspCI	-0.27	1.000
Deuterostomes	Intergenic	*	AlfI - BaeI	0.19	1.000
Deuterostomes	Intergenic	*	CspCI - BaeI	0.46	1.000
Plants	Intronic	*	AlfI - CspCI	2.89	0.273
Plants	Intronic	*	AlfI - BaeI	1.26	1.000
Plants	Intronic	*	CspCI - BaeI	-1.63	1.000
Protostomes	Intronic	*	AlfI - CspCI	0.78	1.000
Protostomes	Intronic	*	AlfI - BaeI	0.15	1.000
Protostomes	Intronic	*	CspCI - BaeI	-0.63	1.000
Deuterostomes	Intronic	*	AlfI - CspCI	0.71	1.000
Deuterostomes	Intronic	*	AlfI - BaeI	0.17	1.000
Deuterostomes	Intronic	*	CspCI - BaeI	-0.53	1.000

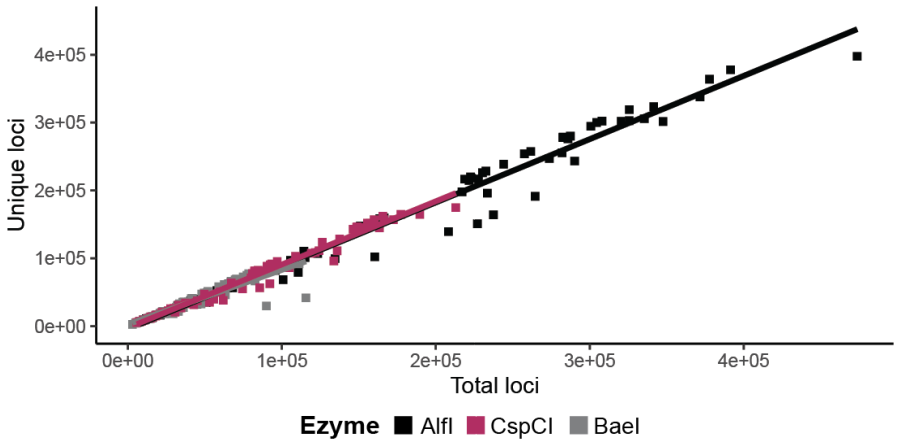


Figure S1: Linear regression between total and unique loci in the 80 analyzed genomes. AlfI (r = 0.99, p <0.001), CspCI (r = 0.99, p <0.001), BaeI (r = 0.92, p <0.001).

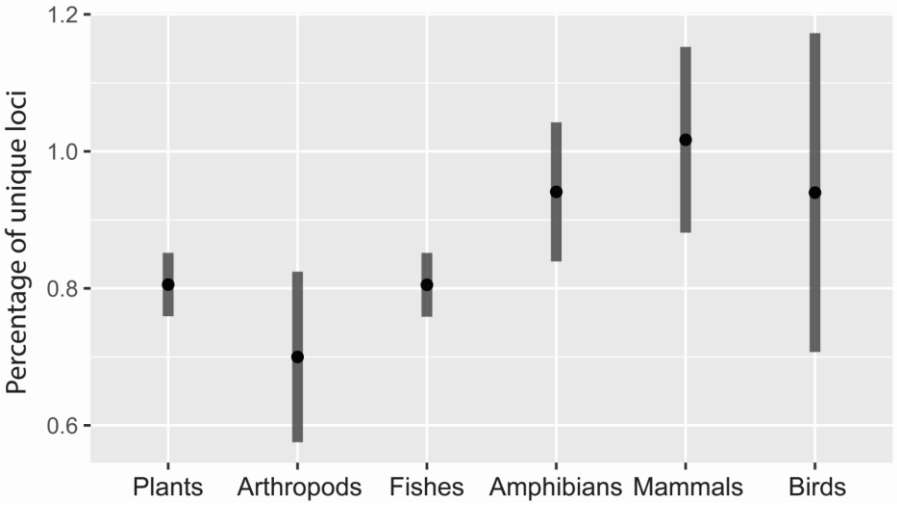


Figure S2: Predicted values of the percentage of unique loci in plants, arthropods, fishes, amphibians, mammals and birds with the group model (Table S8). Mean values are marked with a dot and their 95% confidence intervals are represented with the lines.

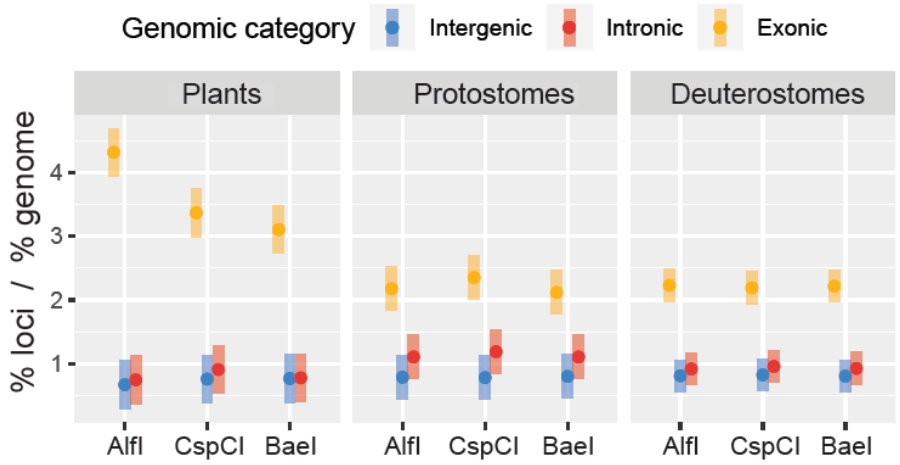


Figure S3: Predicted values of the ratio between the percentage of loci in a genomic category and the percentage of genome in the same genomic category with the GLMM provided in Table S13. Mean values are marked with a dot and their 95% confidence intervals are represented with lines.

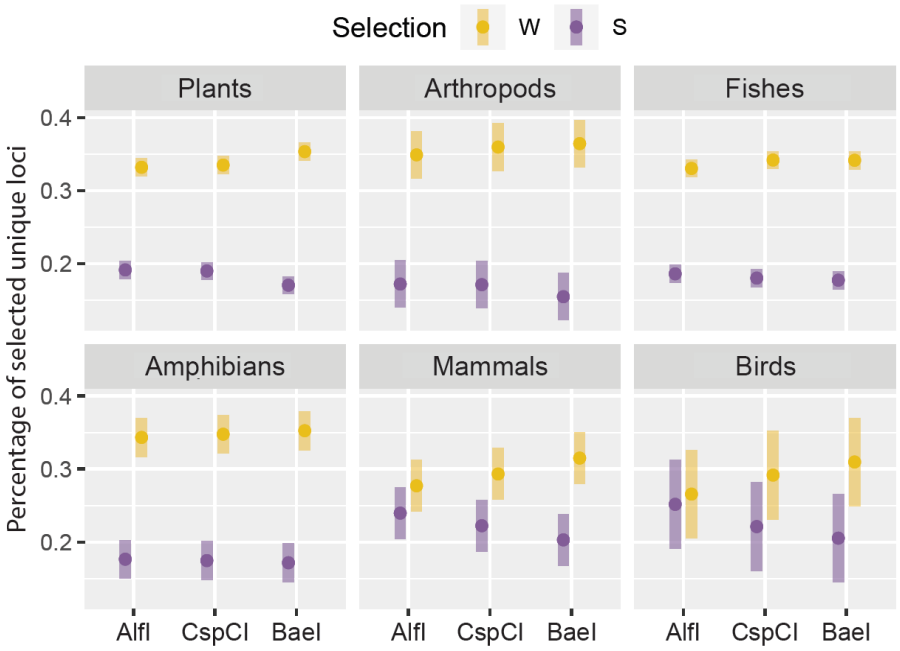


Figure S4: Predicted values of the selected unique loci after secondary reduction with the Group model provided in Table S17.Mean values are marked with a dot and their 95% confidence intervals are represented with lines.

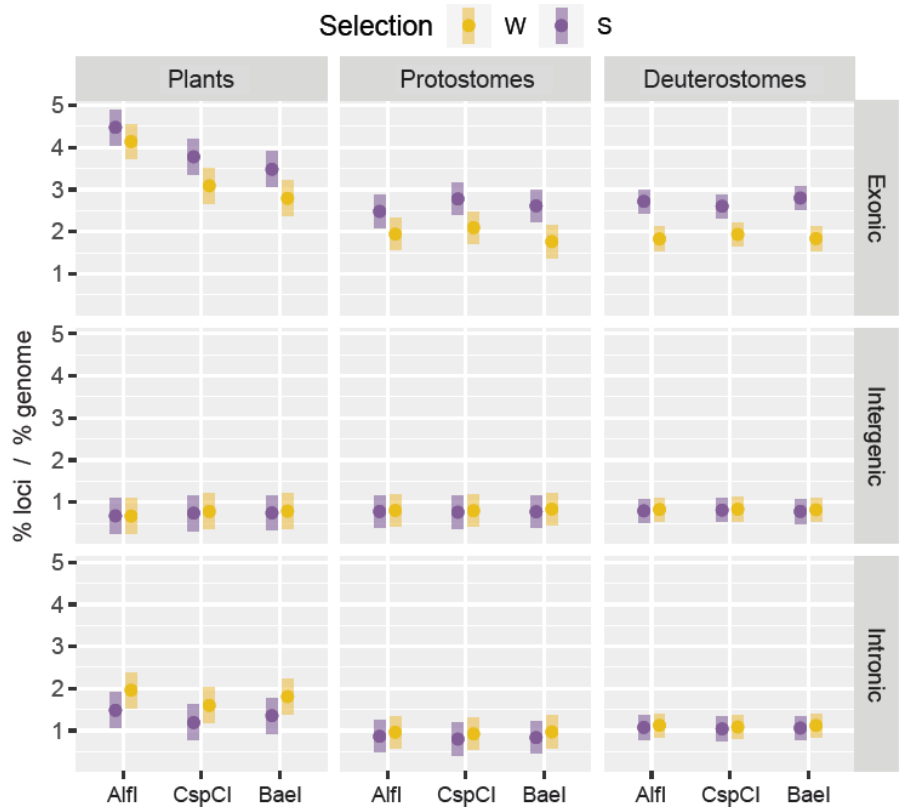


Figure S5: Predicted values, of the ratio between the percentage of loci in a genomic category and the percentage of the same genomic category in the genome after selection, from the GLMM provided in Table S22. Mean values are marked with a dot and their 95% confidence intervals are represented with lines.