Chromosome size affects sequence divergence between species through the interplay of recombination and selection

Anna Tigano [1,2*], Ruqayya Khan[3], Arina D. Omer[3], David Weisz[3], Olga Dudchenko[3,4], Asha S. Multani[5], Sen Pathak[5], Richard R. Behringer[5], Erez L. Aiden[3,4,6,7,8], Heidi Fisher[9], Matthew D. MacManes[1,2]

[1]University of New Hampshire, Molecular, Cellular, and Biomedical Sciences Department, Durham, NH 03824, USA

[2]Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH 03824, USA

[3]The Center for Genome Architecture, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

[4]Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77030, USA

[5]Department of Genetics, University of Texas, M.D. Anderson Cancer Center, Houston, TX 77030, USA

[6]Center for Theoretical and Biological Physics, Rice University, Houston, TX 77030, USA

[7]Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai 201210, China

[8]School of Agriculture and Environment, University of Western Australia, Perth, WA 6009, Australia

[9]Department of Biology, University of Maryland, College Park, MD 20742, USA

*corresponding author: anna.tigano@unh.edu

Keywords: mammal, genome assembly, *Peromyscus*, *Mus*, great apes, genome evolution

Abstract

The structure of the genome, including the architecture, number, and size of its chromosomes, shapes the distribution of genetic diversity and sequence divergence. Importantly, smaller chromosomes experience higher recombination rates than larger ones. To investigate how the relationship between chromosome size and recombination rate affects sequence divergence between species, we adopted an integrative approach that combines empirical analyses and evolutionary simulations. We estimated pairwise sequence divergence among 15 species from three different Mammalian clades - *Peromyscus* rodents, *Mus* mice, and great apes - from chromosome-level genome assemblies. We found a strong significant negative correlation between chromosome size and sequence divergence in all species comparisons within the *Peromyscus* and great apes clades, but not the *Mus* clade, demonstrating that the dramatic chromosomal rearrangements among *Mus* species masked the ancestral genomic landscape of divergence in many comparisons. Moreover, our evolutionary simulations showed that the main factor determining differences in divergence among chromosomes of different size is the interplay of recombination rate and selection, with greater variation in larger populations than in smaller ones. In ancestral populations, shorter chromosomes harbor greater nucleotide diversity. As ancestral populations diverge and eventually speciate, diversity present at the onset of the split contributes to greater sequence divergence in shorter chromosomes among daughter species. The combination of empirical data and evolutionary simulations also revealed other factors that affect the relationship between chromosome size and divergence, including chromosomal rearrangements, demography, and divergence times, and deepen our understanding of the role of genome structure on the evolution of species divergence.

Introduction

Chromosomes are the fundamental unit of inheritance of the nuclear DNA in all eukaryotic species and their evolution goes arm in arm with organismal evolution. Not only the sequence, but also the size, shape, structure and number of chromosomes can vary between species, populations, and even individuals within a population (Hauffe and Searle 1993; Graphodatsky et al. 2011; Dion-Côté et al. 2017; Moura et al. 2020). Chromosome evolution is therefore crucial to our understanding of the evolution and maintenance of biodiversity. The rapidly increasing number of chromosome-level genome assemblies has started to shed light on the role of chromosomes in the genomic distribution of genetic diversity within and between species. For example, chromosome structure, including the location of telomeres or centromeres, can be a strong predictor of the position of dips and peaks in nucleotide diversity (π) and sequence divergence (d) within a chromosome (Butlin 2005; Smukowski and Noor 2011; Burri et al. 2015; Sardell et al. 2018; Tigano, Jacobs, et al. 2020). The heterogeneous distribution of π and d in the genome is apparent among chromosomes too (Dutoit et al. 2017; Henderson and Brelsford 2020), but the role of genome structure in generating this distribution, including the number and size of chromosomes in a genome, is less clear.

To avoid the production of aberrant gametes, the correct segregation of chromosomes during meiosis requires that chromosomes undergo at least one cross-over per event (Mather 1938; Hassold and Hunt 2001), which results in shorter chromosomes experiencing overall proportionally higher recombination rates. In fact, a significant relationship between chromosome size and recombination rate has been reported in many species (but not all) from fungi to mammals (Kaback et al. 1992; Jensen-Seaman et al. 2004; Pessia et al. 2012; Farré et al. 2013; Kawakami et al. 2014; Haenel et al. 2018). Chromosome size is also inversely correlated with π in some species of birds and mammals (Dutoit et al. 2017; Murray et al. 2017; Tigano, Colella, et al. 2020), but not others (Pessia et al. 2012; Dutoit et al. 2017). Further, higher d in microchromosomes (< 20 Mb) relative to macrochromosomes (> 40 Mb) has been observed in several bird species (Delmore et al. 2018). These studies show the intricate relationship between chromosome size, recombination rate, nucleotide diversity, and sequence divergence: although

recombination rate tends to explain differences in diversity and divergence among chromosomes of different length, other factors may contribute to this relationship.

Investigating the factors shaping the levels and patterns of sequence divergence between species is fundamental to understand the molecular mechanisms underlying the process of adaptation and speciation. Nonetheless, the relative contribution of recombination to divergence among species has rarely been directly investigated, especially at the chromosome scale, in species other than humans (Phung et al. 2016). A recent study used chromosome size as a proxy for recombination rate to test how genome structure affected divergence among eight avian sister species pairs, and reported significantly higher divergence in microchromosomes than in macrochromosomes (Delmore et al. 2018), but it did not address the mechanisms underpinning this relationship. It is unlikely that linked selection influences the accumulation of divergence at neutral sites. Under selective sweeps or background selection, the probability that a particular neutral variant goes to fixation is simply its allele frequency, which is the same as in a neutral model (Birky and Walsh 1988). However, selection in ancestral populations will have led to variation in genetic diversity across the genome. As populations diverge, the initial differences between daughter populations simply reflect patterns of diversity present in the parent population. Heterogeneous levels of diversity across the genome of ancestral populations may give rise to variation in divergence between daughter populations. It is therefore crucial to understand what shapes the distribution of diversity across the genome.

Although the direct correlation between recombination rate and π is well understood (Begun and Aquadro 1992; Nachman 2001; Cutter and Choi 2010), the factors and their relative roles in determining this relationship are less clear (Ellegren and Galtier 2016). Non-crossover gene conversion (gene conversion hereafter) and selection, including linked selection, are among the factors most commonly invoked to explain the correlation between recombination and π. Gene conversion is the process by which double-strand DNA breaks during meiosis are repaired using homologous sequence as template without crossing-over, and although it affects shorter sequences than crossing-over events do, it can increase diversity and affect divergence among populations and species (Korunes and Noor 2017). Under a purely

neutral model of evolution, π is determined by the effective population size (N_e) and the mutation rate (μ), as expressed by the equation $\pi=4N_e\mu$ (Tajima 1983). In this model, higher recombination rate may increase π in smaller chromosomes if recombination was mutagenic through gene conversion (Coop and Przeworski 2007; Arbeithuber et al. 2015). For example, a study on humans showed a correlation between recombination, diversity, and divergence to the chimpanzee and the baboon, and explained these relationships with a pure neutral model entailing recombination-associated variation in mutation rates (Hellmann et al. 2003). In contrast, under a non-neutral evolutionary model, selection reduces diversity in the genomic regions surrounding beneficial or deleterious mutations via selective sweeps and background selection, respectively, with a greater diversity-reducing effect in areas of low recombination (Smith and Haigh 1974; Wiehe and Stephan 1993; Hudson and Kaplan 1995) - even though purifying selection can also counteract the loss of diversity due to background selection by associative overdominance if the deleterious variant is recessive (Ohta 1971; Gilbert et al. 2020). Support for the role of selection comes from another study on humans, showing how background selection in the ancestral population affects neutral divergence by reducing diversity in the sites close to selected sites (Phung et al. 2016). At the chromosome level, both gene conversion and linked selection may therefore contribute to the higher diversity reported in smaller chromosomes.

The analysis of chromosome-level patterns of diversity and divergence are now possible thanks to the increasing number of high-quality, chromosome-level assemblies available for several closely-related species within a clade. Great apes and *Mus* mice were the first mammalian clades with enough genomes to enable comparative genomics analyses (Thybert et al. 2018) and a wealth of information is available on the genomes of these species. For example, while within humans and among great apes, fine-scale diversity and divergence seem correlated with recombination, and recombination with chromosome size, (Hellmann et al. 2003; Jensen-Seaman et al. 2004), these relationships are weaker or not present in the house mouse *Mus musculus* (Jensen-Seaman et al. 2004; Kartje et al. 2020). Rodents of the genus *Peromyscus* represent an ideal third clade to understand how common the relationships between chromosome size, recombination, diversity, and divergence are in mammals. For example, even if

Peromyscus are rodents like *Mus*, the cactus mouse (*Peromyscus eremicus*) shows a strong inverse correlation between chromosome size and π (Tigano, Colella, et al. 2020) like in humans (Hellmann et al. 2003), and conserved synteny, recombination rates, and crossover patterning among *Peromyscus* species (Peterson et al. 2019; Smalec et al. 2019) suggest that this relationship between chromosome size and π may be common among species in this genus. In light of these observations on the correlation (or lack thereof) between π and chromosome size, we hypothesize that also *d* will show a negative relationship with chromosome size in great apes and *Peromyscus* but not in *Mus*.

By combining the analysis of chromosome-level genome assemblies from three different mammalian clades - *Mus* spp., *Peromyscus* spp., and great apes - and individual-based evolutionary simulations, we aim to understand whether chromosomes of different sizes show different levels of sequence divergence among species within a clade and to disentangle the evolutionary, demographic, and molecular factors linking recombination, diversity within species, and divergence among species. Through evolutionary simulations, we tested the role of recombination, effective population size N_e, severity of bottleneck associated with population splitting, gene conversion, selection, divergence time, and the interplay of these factors, and discussed the role that they may play in the generation and maintenance of genetic variation and divergence, which is foundational to our understanding of the process of speciation and adaptation in the mammals examined here and other taxa.

Methods

Analyses of divergence

We examined chromosome-level reference genomes for four *Mus* species (*M. musculus*, *M. spretus*, *M. caroli*, and *M. pahari*), five great apes (*Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, and *Pongo abelii*), and six *Peromyscus* species (*P. maniculatus*, *P. polionotus*, *P. eremicus*, *P. crinitus*, *P. nasutus*, and *P. californicus*; Accession numbers in Table S1), of which all but two *Peromyscus* reference genomes were publicly available. We *de novo* assembled the reference genomes of *P. nasutus* and *P. californicus* using a combination of sequencing approaches and final chromosome-scaffolding with Hi-C

data. For *P. nasutus*, we obtained a tissue sample from a female individual collected at El Malpais National Conservation Area (New Mexico, USA) and stored at the Museum of Southwestern Biology (MSB:Mamm:299083). We extracted high molecular weight DNA with the MagAttract HMW DNA Kit (QIAGEN) and selected long fragments (> 10 kb and progressively above 25 kb) using a Short Read Eliminator Kit (Circulomics Inc.). A 10X Genomics linked-read library was generated using this high quality DNA sample at Dartmouth Hitchcock Medical Center (New Hampshire, USA) and sequenced at Novogene (California, USA) using one lane of 150 bp paired-end reads on an Illumina HiSeq X sequencing platform. We produced a first draft of the *P. nasutus* genome assembly using *Supernova 2.1.1* (Weisenfeld et al. 2017) with default settings and these linked reads as input. To order and orient scaffolds in chromosomes we generated and sequenced a proximity-ligation library (Hi-C) from the same sample used for the 10X library as part of the DNA Zoo consortium effort. The Hi-C data were mapped to the 10X assembly with Juicer (Durand et al. 2016) and scaffolds were ordered and oriented in chromosomes with the *3D-DNA* pipeline (Dudchenko et al. 2017) and Juicebox Assembly Tools (Dudchenko et al.). The Hi-C data are available on www.dnazoo.org/assemblies/Peromyscus_nasutus visualized using Juicebox.js, a cloud-based visualization system for Hi-C data (Robinson et al. 2018). For the *P. californicus* genome, high molecular weight DNA was extracted from liver tissue from a captive female individual from a colony maintained at the University of Maryland (USA) and sequenced using 10X Genomics technology at the UC Davis Genome Center (California, USA). A first draft genome for *P. californicus* was based on these 10X linked reads and assembled using *Supernova* as for *P. nasutus*. Then, Chicago and Dovetail Hi-C libraries were created by Dovetail Genomics (California, USA) and used to scaffold the draft assembly with the HiRise pipeline. The Chicago data were used first and the resulting improved assembly was used as input for a second round of scaffolding with the Hi-C data only. The alignment of several *Peromyscus* genomes revealed some assembly errors in the existing *P. eremicus* assembly. Although those did not greatly affect estimates of sequence divergence at the chromosome-level, we generated an additional Hi-C library from a primary fibroblast collection at the T.C. Hsu Cryo-Zoo at the University of Texas MD Anderson Cancer Center. Using the new data we performed misjoin

correction and re-scaffolding using 3D-DNA (Dudchenko et al., 2017:) and Juicebox Assembly Tools (Dudchenko et al., 2018:) . The new Hi-C data for *P. eremicus* are available on www.dnazoo.org/assemblies/Peromyscus_eremicus visualized using Juicebox.js, a cloud-based visualization system for Hi-C data (Robinson et al. 2018).

We generated pairwise alignments and estimated sequence divergence (*d*) with *Mummer4* (Marçais et al. 2018) and custom scripts (https://github.com/atigano/mammal_chromosome_size). First, we aligned pairs of genomes in each clade with *nucmer*, randomly choosing one as the reference and the other as the query, and the settings *--maxgap 2000* and *--mincluster 1000*. We retained a global set of alignments (*-g*) longer than 10 kb using *delta-filter* and converted the output into 'btab' format using *show-coords*. To identify and exclude N-to-N matches from downstream analyses we based our analyses on the estimated 'percent similarity' rather than 'percent identity'. As percent similarities were calculated for alignments of different lengths, we calculated weighted mean chromosome-level *d* (= 1 - (percent similarities)) for each chromosome correcting for alignment length. For the purpose of this study, we focused on autosomes and excluded estimates for sex chromosomes, when present in the genome assembly, because sex chromosomes experience a different combination of evolutionary forces than do autosomes. We tested the ability of the $\log_{10}$-transformed chromosome size in bp (explanatory variable) to predict mean chromosome-level divergence (response variable) separately for each species pairwise alignment using linear models (simple linear regressions), and plotted these relationships in R version 3.6.2 (R core team).

Evolutionary simulations

To disentangle the factors contributing to the relationship between chromosome size, recombination, diversity and divergence, we performed individual-based time-forward evolutionary simulations in SLiM3 (Haller and Messer 2019). We set up a simple Wright-Fisher model, where an ancestral population (pop_A) splits into two populations (pop_1 and pop_2) after $20N_e$ generations (Fig. S1A). We selected this time as a burn-in to allow for coalescence, to generate diversity and to reach stable allele

frequencies. To test for the effect of population size and its changes over time, we simulated ancestral populations pop_A of 10,000, 40,000 and 160,000 individuals, which grossly encompass variation in N_e among great apes, *Peromyscus* and *Mus* (Lack et al. 2010; Phifer-Rixey et al. 2012; Prado-Martinez et al. 2013; Harris et al. 2016; Colella et al. 2020) and modeled bottlenecks of different severity associated with the split of pop_A into two daughter populations pop_1 and pop_2: individuals from pop_A were either sorted into two daughter populations of equal size (N_e in pop_1 and pop_2 were $0.5N_e$ of pop_A) or an additional bottleneck further reduced N_e in pop_1 and pop_2 to $0.1N_e$ of pop_A. As our working hypothesis was that chromosome size affects diversity and divergence due to higher recombination rates r in smaller chromosomes, we simulated chromosomes of fixed length (1 Mb) with varying r to account for chromosome size variation, while keeping everything else the same. Assuming one crossover/chromosome on average, mean chromosome-wide r was 10^{-8} (1/sequence length), so to encompass variation in chromosome size in the mammals examined we simulated nine different recombination rates, spanning $0.33r$ to $3r$, which extends beyond the variation in recombination rates expected to occur in these species based on variation in chromosome size. Note that recombination rates are even across the chromosome and constant through time. We calculated mean gene size (including introns and exons) and mean distance between genes from the gene annotation of the *P. eremicus* genome and built the chromosome structure based on these values, resulting in each chromosome having 9 coding genes of 20.5 kb separated by 94.5 kb of intergenic sequence (Fig. S1B). We used a fixed germline mutation rate for all models as estimated in *M. musculus* ($5.7*10^{-9}$, (Milholland et al. 2017)). We modeled gene conversion rate ($r/3$) and gene conversion tract length (440 bp), when included in the model, based on estimates in *Drosophila melanogaster* (Miller et al. 2016). In neutral models all mutations were neutral, whereas in models with selection mutations in coding genes could be neutral, deleterious or advantageous at a relative frequency of 0.3/1/0.0005, with the non-neutral mutations being always codominant. The fitness effects of the non-neutral mutations were drawn from a gamma distribution with a mean selection coefficient s of $\pm 15.625*10^{-3}$ and a shape parameter alpha of 0.3 based on the parameter space explored by (Campos and Charlesworth 2019; Stankowski et al. 2019). We scaled N_e, μ, and r by a

factor of 25 to expedite simulations and ran 30 unique simulation replicates for each combination of parameters. All the different parameters used in the simulations are summarized in Table 1.

To investigate the factors affecting levels of diversity in chromosomes of different sizes, we sampled 30 individuals when pop_A reached $20N_e$ generations (i.e. right before the split) for each simulation, output variant sites in a VCF file, and calculated π across the chromosome, in coding genes only, and in intergenic areas only, using VCFtools (Danecek et al. 2011). To obtain estimates of sequence divergence from simulations comparable to those from pairwise alignment of genome assemblies, we calculated *d* as the proportion of unmatched bases between two haploid genomes sampled randomly from each of the two diverging populations pop_1 and pop_2. We output these estimates right after the split and every 250,000 generations afterwards, up to 10 million generations. To further disentangle the effect of direct and linked selection, we estimated *d* also in coding genes and intergenic areas separately between the same genomes sampled above one generation after the split, when *d* is highest and not affected by decay yet (see Results and discussion).

Results and discussion

Empirical data show a strong, inverse relationship between chromosome size and divergence between species, with a few exceptions

Chromosome size (log-transformed) was a strong significant predictor of mean sequence divergence *d* in each of the species pairwise comparisons in *Peromyscus* and great apes (all comparisons had $p < 0.001$ using linear models; Fig. 1A and 1C for example comparisons, Fig. S2 and S3 for all comparisons). Chromosome size showed a negative relationship to mean *d* and explained 62-89% and 46-65% of the variance in mean *d* across chromosomes (all R^2 are adjusted hereafter) in *Peromyscus* and great apes, respectively. Among *Mus* spp., we found a significant, negative relationship between chromosome size and *d* only between *M. pahari* and *M. spretus* ($p < 0.001$; Fig. 1E), which explained 42% of the variance in *d* across chromosomes. We hypothesized that the discrepancy between results from *Mus* and the other two clades examined could be explained by relative poor genome structure conservation in *Mus* so we

investigated this further. Among the *Mus* genome alignments, the *M. pahari*/*M. spretus* comparison was the only one where *M. pahari* was used as a reference genome. *M. pahari* is the most divergent (3-6 million years ago) and differs from the other *Mus* species in that it shows a karyotype with 24 chromosomes, while *M. spretus, M. musculus* and *M. caroli* exhibit karyotype with only 20 chromosomes (Thybert et al. 2018). Further, fewer synteny breaks between *M. pahari* and the rat (*Rattus norvegicus*) relative the other *Mus* species analyzed here (19 versus 35; (Thybert et al. 2018)) demonstrate that the *M. pahari* karyotype is the most similar to the ancestral karyotype of the *Mus* species included here. As the reference and query genomes in each pairwise comparison were chosen randomly, we produced new alignments and calculated *d* for all the possible pairwise combinations within each clade to test for the effect of the reference genome to chromosome-level *d* estimates. While in the *Peromyscus* and great apes clades all reference-query combinations, including the reciprocal of the comparisons first analyzed (Fig. 1B and 1D) showed a strong negative relationship between chromosome size and *d* (R^2= 0.59-0.91 among *Peromyscus* and 0.46-0.65 among great apes, all $p < 0.001$; Fig. S2 and S3), in the *Mus* clade this relationship emerged only when *M. pahari* was used as reference genome ($p < 0.001$; Fig. 1F and S4) and explained 52 and 54% of the variance in *M. musculus* and *M. caroli*, respectively (Fig. S4). While a significant positive correlation between neutral human-primate divergence and human recombination rate has been reported at smaller scales (i.e. in 100 kb sliding windows across the genome; (Phung et al. 2016)), our results show that these relationships between recombination, diversity, and divergence are strong at a large, chromosome scale within three different highly divergent clades (~30-90 MYA), hence also in *Mus*, where, at least in *M. musculus*, the relationship between π and recombination at smaller scales is not always significant (Kartje et al. 2020).

The fact that chromosome size in the most ancestral karyotype (*M. pahari*), but not the derived ones (*M. musculus*, *M. spretus*, and *M. caroli*), is a strong predictor of levels of *d* between species with different genome structures indicates that these patterns evolve and are maintained across long evolutionary scales. The retention of ancestral patterns suggest that recombination hotspots could be conserved in rearranged chromosomes despite the evolution of a different genome structure, or that a

different genomic landscape of recombination has not been sufficient to redistribute variation in sequence divergence expected based on differences in chromosome size over this time scale. Though the human genome underwent a chromosomal fusion compared to the other great apes, the correlation between chromosome size and *d* among great apes did not seem affected by the use of the human genome as reference (i.e. chromosome size did not explain a lower proportion of the variance in these comparisons; Fig. 1C, 1D, S3).

The choice of a model species is largely based on its potential to provide insights that are generalizable to other organisms, yet our results builds on previous work in *M. musculus* (Jensen-Seaman et al. 2004; Kartje et al. 2020) showing that the *Mus* clade is rather an outlier, and does not serve as a good model to analyze general patterns with regards to how genome and chromosome structure affect and shape heterogenous levels and patterns of diversity and divergence across the genome. However, our analysis of the *Mus* genomes provides substantial insight into the reasons a species may deviate from expectations, showing that the apparent lack of a relationship between chromosome size and recombination rate, π, and *d* in *M. musculus* is well explained by the rearrangements of chromosome fragments playing a major confounding factor.

Evolutionary simulations reveal the factors driving the empirical patterns.

Simulations helped generate a mechanistic understanding of most of the empirical patterns reported in this and other studies (Dutoit et al. 2017; Murray et al. 2017; Kartje et al. 2020; Tigano, Colella, et al. 2020). In neutral simulations, we did not observe variation in π among chromosomes with different recombination rates whether the model included gene conversion or not (ANOVA, $p > 0.05$). These results indicate that recombination alone does not explain variation in π among chromosomes and that gene conversion does not contribute substantially to increasing levels of π, at least at the rates that we assumed and over relatively short evolutionary times ($20N_e$ generations). Gene conversion occurs at a fraction of the recombination rate and affects only a small segment of DNA (100-2000 bp) at a time (Korunes and Noor 2017; Korunes and Noor 2019), hence its effect on chromosome-wide levels of π may

be detectable only over long evolutionary times. Recombination could also be mutagenic *per se* by promoting *de novo* mutations at the DNA breaks caused by crossovers, but the mechanism underlying this phenomenon is not clear (Hodgkinson and Eyre-Walker 2011). We did not model crossover mutagenesis in our simulations, but a recent study in humans found the mutation rate associated with crossovers to be ~4% , i.e. one de novo mutation every ~23 crossovers (Halldorsson et al. 2019) suggesting that crossover mutagenesis could contribute to, but not entirely account for, the variation in π among chromosomes of different sizes over long evolutionary times, similarly to gene conversion. In contrast, in models with selection, recombination rate was a significant ($p < 0.001$) predictor of differences in π among chromosomes across populations of three vastly different N_e (Fig. 2A). However, Δπ - the difference in π between the chromosomes with the highest and lowest recombination rates - spanned over two orders of magnitude when comparing the smallest and the largest simulated ancestral populations ($\Delta\pi = 3.16*10^{-5}$-$2.69*10^{-3}$; Fig. 3). Also the proportions of variance in π explained by recombination rate increased with N_e: they were 13, 60, and 85% in populations of 10,000, 40,000 and 160,000 individuals respectively in models without gene conversion (estimates were similar for models with gene conversion, except for the smallest N_e where *r* explained 27% of the variation), suggesting that while selection is the main determinant of the relationship between recombination and π in large populations, genetic drift prevails in small populations.

The comparison of π across models with and without selection shows that diversity is lower overall, regardless of the recombination rate, in chromosomes affected by selection (Fig. 2A). The reduction in π is strongest at the coding genes, which experience both positive and negative selection directly, and indirectly through linked selection (Fig. 2A). Chromosome-wide estimates are more similar to those based on the analysis of intergenic areas subjected to linked selection only (Fig. 2A), which is at least partly due to the relatively much larger proportion of non-coding over coding regions. Nonetheless, a positive relationship between recombination rate and π is evident across the different genomic areas and N_e considered (Fig. 2A). The comparison of π estimates from across the chromosome, coding genes only, and intergenic areas only, corroborate that differences in diversity among chromosomes of different sizes

are due to the balance between selection and recombination, with selection reducing diversity and recombination reducing linkage disequilibrium, thus reducing the effect of linked selection. As recombination increases, it more strongly counteracts linked selection in the intergenic areas, to the point of almost restoring levels of diversity similar to those expected under a neutral model (Fig. 2A). In other words, in our simulations the interplay between recombination and selection is the main factor driving the inverse correlation between chromosome size and π, a pattern that is described in many species.

The fact that recombination was a stronger contributor to variation in π in larger populations than in smaller ones could be due to one or the combination of two factors: 1) as the effect of selection in the genome depends on both effective population size and the strength of selection ($N_e s$), larger populations will have proportionally more selective sweeps than smaller populations - because more mutations in small populations will have scaled coefficients so small that they will actually behave as neutral - and these sweeps will be more strong and efficient at removing diversity, and 2) larger populations will have higher population recombination rates ($\rho = 4N_e r$), which will break linkage disequilibrium even more efficiently in chromosomes with high recombination rates. As smaller population will have proportionally more mutations behaving as neutral ones, more mutations will be more predominantly governed by stochasticity rather than by the deterministic effect of selection in these small populations (Charlesworth 2009), which is consistent with selection explaining less variation in π in smaller populations than in larger ones (see above).

In neutral models, *d* did not vary across chromosomes with different recombination rates (ANOVA, $p >> 0.05$; Fig. 3B), while in models with selection differences in *d* among chromosomes one generation after the split were significantly different from zero (ANOVA, $p < 0.001$; Fig. 2B) and strongly correlated with recombination rate ($p < 0.001$; Fig. 2B), i.e. chromosomes with lower recombination rates had lower *d* in populations, across all three simulated N_e. Sequence divergence right after the split reflected the levels of diversity within the ancestral population before the split across all models, with higher Δd - the difference in *d* between the chromosomes with the highest and lowest recombination rates - in larger populations (Fig. 2B). Similarly to the patterns observed for π, *d* was

lowest in coding regions and highest in intergenic areas, with chromosome-wide estimates lower than, but similar to, the latter (Fig. 2B). Testing empirically whether species with higher ancestral N_e show higher Δd would support the contribution of demography in the relationship between recombination and d. However, across the three clades examined here, N_e and divergence times between species seem to covary, for example with great apes having not only the smallest N_e but also the most recent species divergence, so that the actual relative contributions of N_e and divergence times cannot be disentangled presently.

In neutral models, d increases linearly with time ($4N_e\mu + 2T\mu$, where T is the number of generations) so the severity of the bottleneck at the time of the split does not have any effect on divergence between isolated populations in our neutral simulations, even at the smallest N_e. Although genetic drift was *de facto* the only evolutionary force driving changes in allele frequency in these neutral models, its effect size ($= 1/2N_e$) in our simulated populations was nonetheless very small ($6.25*10^{-6}$ - 10^{-4}), which indicates that new mutations are the main source of d over time in neutral models. In fact, our d estimates encompass both fixed and segregating mutations in each of the two compared populations.

In models with selection, d increased over time, though at a slower pace than in neutral models, and faster in smaller populations relative to larger populations (Fig. 3). Further, while larger populations showed higher overall divergence than smaller ones in the early stages of divergence, the trend reversed with time: after 10 million generations d between the smallest populations (N_e = 5,000 individuals each) surpassed d both between the medium-sized (N_e = 20,000 individuals each) and between the largest populations (N_e = 80,000 individuals each; Fig. 3). This pattern was even more pronounced when populations pop_1 and pop_2 were affected by a stronger bottleneck at the time of the split from pop_A: populations of all sizes accumulated d faster than in the models with a weaker bottleneck, and even faster between the smallest populations (N_e = 1,000 individuals each) compared to estimates from larger populations (N_e = 4,000 and 16,000 individuals each, respectively; Fig. S5). These results show how genetic drift is much stronger in small populations, but only in the models with selection (Fig. S5). Neutral mutations fix at a much faster rate than those under selection because the fixation probability of a

383 locus under selection depends also on the strength of selection acting on its linked sites (Hill-Robertson
384 interference; (Hill and Robertson 1966; Felsenstein 1974)). Given the distribution of fitness effects (DFE)
385 we simulated, the probability of a beneficial mutation to effectively act as a neutral mutation (i.e. that $N_e s$
386 < 1) is higher in smaller populations than in larger ones (Charlesworth 2009). Therefore, in smaller
387 population a higher proportion of beneficial or deleterious mutations will act as neutral and their fixation
388 probability will be higher and depend only on the combination of genetic drift and linked selection, which
389 in turn will depend on the strength of selection acting on the linked mutation and the rate of
390 recombination affecting linkage disequilibrium between the two mutations.

391 We found that Δd decreased with divergence time in all models with selection (Fig. 4), suggesting
392 that either divergence rate is relatively accelerated in large chromosomes or slowed down in small ones.
393 Based on what discussed above, larger chromosomes, where recombination is lower, should experience
394 stronger Hill-Robertson interference, and hence lower fixation probabilities and slower divergence rate,
395 which is in contrast with what observed. Alternatively, lower recombination in larger chromosomes could
396 strengthen the effect of linked selection, resulting in local chromosome-wide reductions in N_e, and thus
397 stronger genetic drift and faster rate of sequence divergence than smaller chromosomes with higher
398 recombination rates. This is clearly illustrated by an empirical study on Bornean and Sumatran
399 Orangutans (*P. pygmaeus* and *P. abelii*) showing that estimates of ancestral N_e vary among chromosomes
400 and that chromosome size is a strong predictor of variation in both inferred ancestral N_e and
401 recombination rate, which in turn suggests a direct relationship between these two (Mailund et al. 2011).
402 Additionally, (Phung et al. 2016) showed that the window-based correlation between recombination and
403 divergence rate in simulated genomes decreased with splitting time, and attributed this decrease to a
404 concurrent decrease in levels of ancestral variation. At the chromosome level, these observations suggest
405 that chromosomes of different sizes may lose ancestral variation at different rates and thus may account
406 for the decay in Δd.

407 These different rates of divergence explain why over time Δd becomes negative in our
408 simulations, i.e. the chromosome with the lowest recombination rate becomes more divergent than the

chromosome with the highest recombination rate, in the populations with the smaller N_e (Fig. 4). Initially, the differences in Δd will be determined by $\Delta\pi$ in the ancestral population at the time of the split, but with time the different divergence rates among chromosomes caused by the interplay of recombination, selection, and thus drift, will erode Δd, and even reverse it (Fig. 4). Furthermore, the severity of the bottleneck at the species split affected the Δd decay rate, with a faster decay in less severely reduced N_e, regardless of ancestral N_e (Fig. 4). Also gene conversion seemed to accelerate the Δd decay overall, but not in the largest population experiencing the weaker bottleneck (Fig. 4). That Δd decay was generally faster both in populations experiencing milder bottlenecks, which therefore had larger N_e, and in models with gene conversion compared to those without, suggests that higher mutation rates in these cases can accelerate Δd decay. Although the ultimate mechanism is not clear, faster Δd decay may be explained by the rate of loss of ancestral polymorphism (Phung et al. 2016), which should be higher with higher mutation rates. Importantly, the effect of gene conversion becomes more evident after 5 million generations (Fig. 4), showing that, although gene conversion is not the main determinant of differences in π and d among chromosomes of different sizes, it can contribute to these patterns over long evolutionary times. The evolutionary simulations provide clear insights into why the variation in divergence among chromosomes of different sizes decreases with time, and suggest that it would require either small N_e and/or long divergence times to observe a negative Δd. We did not observe any negative Δd in our species pairwise comparisons. In the future, the empirical test of these observations will require the inclusion of additional clades and larger number of species comparisons within clades to verify how commonly this occurs empirically and to disentangle the factors promoting, or hindering, this pattern.

Our simulations were based on reasonably realistic parameters, except for the absence of neutrally evolving introns in simulated genes, to reach an acceptable compromise between capturing the complexity of the evolutionary processes, and their interactions, while maintaining enough statistical power to understand the relative roles of the many factors at play. Notwithstanding that the "real" values for the factors included in our simulations are not available for most species or nonetheless difficult to estimate, excluding introns provided more target sequence for selection to act on and a good trade-off in

terms of computational time and resources. With the exception of a clear effect of N_e and divergence time in the magnitude of the variation in *d* observed among chromosomes, our evolutionary simulations well illustrate the processes driving the empirical patterns of divergence between species described in three different clades of mammals: in the absence of chromosomal rearrangements, the interplay of recombination and selection determines levels of π in the ancestral species; higher π in the ancestral species results in higher *d* among haplotypes sorted into the daughter species, thus explaining the differences in π and *d* among chromosomes of different sizes.

Empirical analyses and simulations highlight the rule and the exceptions

Species showing an inverse relationship between recombination rate and chromosome size are found among mammals, birds, yeast, worms, and plants (Pessia et al. 2012), which highlights that this relationship is not an idiosyncratic feature of a particular taxon but rather a common trend. Our evolutionary simulations show that varying recombination rates across chromosomes should result in differences in π and *d* among chromosomes of different sizes, but empirical support for this prediction is mixed. In *M. musculus*, for example, chromosome size is not a good predictor of variation in π within the species (Pessia et al. 2012) or *d* to other *Mus* spp. (our study). Our analyses here have shown that this lack of correlation is due to dramatic changes in the genome structure of *M. musculus* and other congeneric species relative to their common ancestor (most similar to *M. pahari*), and hence stress the importance of the choice of the reference genome in this type of analysis. Not only the genome we use as reference could mask existing relationships due to the evolution of different genome structure, as we have shown here, but also its quality is crucial to obtain high-quality genome alignments, to calculate *d* accurately, and to estimate chromosome sizes from sequence length in the absence of cytological data.

No significant relationship between chromosome size and divergence between human and chimpanzee was found in another study (Patterson et al. 2006), but these results were based on only 20 Mb of aligned sequences. Moreover, different avian species have shown a positive (Dutoit et al. 2017), a negative (Manthey et al. 2015), or no relationship (Callicrate et al. 2014) between chromosome size and

π. Dramatic chromosomal rearrangements can be excluded in these examples (Ellegren 2010), begging the question: what other factors could explain these deviations from our model? First, given the variation in *d* within chromosomes, incomplete genome sampling may confound these chromosome-level relationships. Second, Dutoit and colleagues (2017) argue that in the collared flycatcher (*Ficedula albicollis*) a positive relationship between chromosome size and π, which is opposite to expectations, could be explained by the density of targets of selection, higher in smaller chromosomes than in larger ones in this species. However, given the high degree of synteny conservation among birds (Ellegren 2010), all avian species should show a similar pattern, which is not the case. For example, the comparison of genome-wide patterns in π in the passenger pigeon (*Ectopistes migratorius*), known as the most abundant bird in North America before it went extinct, and the band-tailed pigeon (*Patagioenas fasciata*), with a current population size three orders of magnitude smaller, not only shows higher π in smaller chromosomes in both species, but also the effect of N_e on Δπ, as per our predictions (Murray et al. 2017). Our analyses highlight the contribution of demography (i.e. N_e, severity of bottleneck and genetic drift) in affecting, and even reversing, the relationship between recombination, π, and *d*, which could potentially explain the opposite correlation reported in the collared flycatcher. The importance of historical demography has been demonstrated also in the divergence of the sex chromosome Z in *Heliconius* butterflies using a combination of empirical data and evolutionary simulations (Van Belleghem et al. 2018). Alternatively, the strength of selection, rather than the density of targets of selection, could disrupt the correlation between chromosome size and π in case of strong selective sweeps preferentially occurring in small chromosomes. Finally, limited variation in recombination, π, and *d* among chromosomes could be simply due to lack of variation in chromosome size. The analysis of 128 eukaryotic and prokaryotic genomes has shown that variation in chromosome size is directly proportional to genome size (Li et al. 2011), suggesting that variation in recombination, π, and *d* among chromosomes should decrease with genome size.

Conclusions

Variation in recombination across the genome affects the evolution and maintenance of traits relevant to adaptation and speciation, the genomic architecture of the loci underlying those traits, and our ability to detect those loci (Yeaman and Otto 2011; Yeaman 2013; Cruickshank and Hahn 2014; Burri et al. 2015; Lotterhos 2019; Booker et al. 2020). We have shown strong evidence from empirical analyses and evolutionary simulations that the inverse relationship between recombination rate and chromosome size can result in significant differences in π and *d* among chromosomes of different sizes, and this relationship is known to affect estimates of ancestral N_e in chromosomes of different sizes (Mailund et al. 2011). In the clades included in this study, N_e covaries with divergence time scales, thus it is not possible to disentangle the relative effect of these two factors on patterns of divergence at this time. Future analyses of species from with different combinations of N_e and divergence times will help address this gap. Nonetheless, our study shows that chromosome size should be considered in the study of the genomic basis of adaptation and speciation. Do smaller chromosomes play a proportionally more prominent role that larger chromosomes in adaptation and speciation? Or are these differences in π and *d* strong enough to confound signals of selection in the genome? As chromosome-level assemblies and population whole genome resequencing data of closely-related species become available for an increasing number of taxa, the combination of empirical and theoretical investigations will help address these outstanding questions and generate new ones on chromosome and genome evolution.

Acknowledgements

We would like to thank the Museum of Southern Biology for the *P. nasutus* sample, and Tom Booker for helpful feedback on the simulations and a previous version of the manuscript. All analyses and simulations were performed on the UNH Premise Cluster. This work was funded by the National Institute of Health National Institute of General Medical Sciences to M.D.M. (1R35GM128843). All scripts for sequence divergence analyses, evolutionary simulation, and figures are stored at https://github.com/atigano/mammal_chromosome_size. Pawsey Supercomputing Centre with funding

513 from the Australian Government and the Government of Western Australia for computational support of
514 the DNA Zoo assembly effort. E.L.A. was supported by an NSF Physics Frontiers Center Award
515 (PHY1427654), the Welch Foundation (Q-1866), a USDA Agriculture and Food Research Initiative
516 Grant (2017-05741), and an NIH Encyclopedia of DNA Elements Mapping Center Award
517 (UM1HG009375). DNA Zoo sequencing effort is supported by Illumina, Inc. The Hi-C data generated by
518 the DNA Zoo Consortium is available on www.dnazoo.org and via the DNA Zoo SRA BioProject
519 #PRJNA512907.

520

521 **Author contributions**

522 A.T. conceived the study, performed analyses and simulations, and wrote the first version of the
523 manuscript. R.K., A.D.O. generated the Hi-C data for *P. nasutus* and *P. critinus* as part of the DNA Zoo
524 effort, and R.K., D.W., O.D. and E.L.A. used these data to generate the corresponding chromosome-
525 length assemblies. A.M., S.P., and R.R.B. provided species-authenticated *Peromyscus* primary
526 fibroblasts. A.T., H.F., and M.D.M. generated data, assembled genomes, and reviewed and edited the
527 paper. All authors approved the final version of the manuscript.

References

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci. U. S. A.* 112:2109–2114.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519–520.

Birky CW Jr, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 85:6414–6418.

Booker TR, Yeaman S, Whitlock M. 2020. Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Mol. Ecol.* 29:4274-4279.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res.* 25:1656–1665.

Butlin RK. 2005. Recombination and speciation. *Mol. Ecol.* 14:2621–2635.

Callicrate T, Dikow R, Thomas JW, Mullikin JC, Jarvis ED, Fleischer RC, NISC Comparative Sequencing Program. 2014. Genomic resources for the endangered Hawaiian honeycreepers. *BMC Genomics* 15:1098.

Campos JL, Charlesworth B. 2019. The Effects on Neutral Variability of Recurrent Selective Sweeps and Background Selection. *Genetics* 212:287–303.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* 10:195–205.

Colella JP, Tigano A, Dudchenko O, Omer AD, Khan R, Bochkov ID, Aiden EL, MacManes MD. 2020. Multiple evolutionary pathways to achieve thermal adaptation in small mammals. *bioRxiv* [Internet]:2020.06.29.178392. Available from: https://www.biorxiv.org/content/10.1101/2020.06.29.178392v2.abstract

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

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23:3133–3157.

Cutter AD, Choi JY. 2010. Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Res.* 20:1103–1111.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

Delmore KE, Lugo Ramos JS, Van Doren BM, Lundberg M, Bensch S, Irwin DE, Liedvogel M. 2018. Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. *Evol. Lett.* 2:76–87.

564 Dion-Côté A-M, Symonová R, Lamaze FC, Pelikánová Š, Ráb P, Bernatchez L. 2017. Standing
565 chromosomal variation in Lake Whitefish species pairs: The role of historical contingency and
566 relevance for speciation. *Mol. Ecol.* 26:178–192.

567 Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander
568 ES, Aiden AP, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields
569 chromosome-length scaffolds. *Science* 356:92–95.

570 Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, Pham M, St Hilaire BG, Yao
571 W, Stamenova E, et al. 2018. The Juicebox Assembly Tools module facilitates de novo assembly of
572 mammalian genomes with chromosome-length scaffolds for under $1000. Available from:
573 http://dx.doi.org/10.1101/254797

574 Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox
575 Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* 3:99–101.

576 Dutoit L, Burri R, Nater A, Mugal CF, Ellegren H. 2017. Genomic distribution and estimation of
577 nucleotide diversity in natural populations: perspectives from the collared flycatcher (*Ficedula*
578 *albicollis*) genome. *Mol. Ecol. Resour.* 17:586–597.

579 Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol. Evol.* 25:283–291.

580 Ellegren H, Galtier N. 2016. Determinants of genetic diversity. *Nat. Rev. Genet.* 17:422–433.

581 Farré M, Micheletti D, Ruiz-Herrera A. 2013. Recombination Rates and Genomic Shuffling in Human
582 and Chimpanzee—A New Twist in the Chromosomal Speciation Theory. *Mol. Biol. Evol.* 30:853–
583 864.

584 Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* 78:737–756.

585 Gilbert KJ, Pouyet F, Excoffier L, Peischl S. 2020. Transition from Background Selection to Associative
586 Overdominance Promotes Diversity in Regions of Low Recombination. *Curr. Biol.* 30:101–107.e3.

587 Graphodatsky AS, Trifonov VA, Stanyon R. 2011. The genome diversity and karyotype evolution of
588 mammals. *Mol. Cytogenet.* 4:22.

589 Haenel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale crossover rate
590 variation in eukaryotes and its significance to evolutionary genomics. *Mol. Ecol.* 27:2477–2497.

591 Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B,
592 Oddsson A, Halldorsson GH, Zink F, et al. 2019. Characterizing mutagenic effects of recombination
593 through a sequence-level genetic map. *Science* [Internet] 363. Available from:
594 http://dx.doi.org/10.1126/science.aau1043

595 Haller BC, Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model.
596 *Mol. Biol. Evol.* 36:632–637.

597 Harris SE, Xue AT, Alvarado-Serrano D, Boehm JT, Joseph T, Hickerson MJ, Munshi-South J. 2016.
598 Urbanization shapes the demographic history of a native rodent (the white-footed mouse,
599 *Peromyscus leucopus*) in New York City. *Biology Letters* [Internet] 12:20150983. Available from:
600 http://dx.doi.org/10.1098/rsbl.2015.0983

601 Hassold T, Hunt P. 2001. To err (meiotically) is human: the genesis of human aneuploidy. *Nat. Rev.*

Genet. 2:280–291.

Hauffe HC, Searle JB. 1993. Extreme karyotypic variation in a Mus musculus domesticus hybrid zone: the tobacco mouse story revisited. *Evolution* 47:1374–1395.

Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* 72:1527–1535.

Henderson EC, Brelsford A. 2020. Genomic differentiation across the speciation continuum in three hummingbird species pairs. *BMC Evol. Biol.* 20:113.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetical Research* [Internet] 8:269–294. Available from: http://dx.doi.org/10.1017/s0016672300010156

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* 12:756–766.

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–1617.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen C-F, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* 14:528–538.

Kaback DB, Guacci V, Barber D, Mahon JW. 1992. Chromosome size-dependent control of meiotic recombination. *Science* 256:228–232.

Kartje ME, Jing P, Payseur BA. 2020. Weak Correlation between Nucleotide Variation and Recombination Rate across the House Mouse Genome. *Genome Biol. Evol.* 12:293–299.

Kawakami T, Smeds L, Backström N, Husby A, Qvarnström A, Mugal CF, Olason P, Ellegren H. 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology* [Internet] 23:4035–4058. Available from: http://dx.doi.org/10.1111/mec.12810

Korunes KL, Noor MAF. 2017. Gene conversion and linkage: effects on genome evolution and speciation. *Mol. Ecol.* 26:351–364.

Korunes KL, Noor MAF. 2019. Pervasive gene conversion in chromosomal inversion heterozygotes. *Mol. Ecol.* 28:1302–1315.

Lack JB, Pfau RS, Wilson GM. 2010. Demographic history and incomplete lineage sorting obscure population genetic structure of the Texas mouse (*Peromyscus attwateri*). *J. Mammal.* 91:314–325.

Li X, Zhu C, Lin Z, Wu Y, Zhang D, Bai G, Song W, Ma J, Muehlbauer GJ, Scanlon MJ, et al. 2011. Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. *Mol. Biol. Evol.* 28:1901–1911.

Lotterhos KE. 2019. The Effect of Neutral Recombination Variation on Genome Scans for Selection. *G3* 9:1851–1867.

Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH. 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent

639 hidden Markov model. *PLoS Genet.* 7:e1001319.

640 Manthey JD, Klicka J, Spellman GM. 2015. Chromosomal patterns of diversity and differentiation in
641 creepers: a next-gen phylogeographic investigation of *Certhia americana*. *Heredity* 115:165–172.

642 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and
643 versatile genome alignment system. *PLoS Comput. Biol.* 14:e1005944.

644 Mather K. 1938. Crossing-over. *Biol. Rev. Camb. Philos. Soc.* 13:252–292.

645 Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. 2017. Differences between germline and somatic
646 mutation rates in humans and mice. *Nat. Commun.* 8:15183.

647 Miller DE, Smith CB, Kazemi NY, Cockrell AJ, Arvanitakis AV, Blumenstiel JP, Jaspersen SL, Hawley
648 RS. 2016. Whole-Genome Analysis of Individual Meiotic Events in *Drosophila melanogaster*
649 Reveals That Noncrossover Gene Conversions Are Insensitive to Interference and the Centromere
650 Effect. *Genetics* 203:159–171.

651 Moura MN, Cardoso DC, Lima Baldez BC. 2020. Intraspecific variation in the karyotype length and
652 genome size of fungus-farming ants (genus *Mycetophylax*), with remarks on procedures for the
653 estimation of genome size in the Formicidae by flow cytometry. *PLoS One* [Internet]. Available
654 from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0237157

655 Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da
656 Fonseca RR, Fulton TL, et al. 2017. Natural selection shaped the rise and fall of passenger pigeon
657 genomic diversity. *Science* 358:951–954.

658 Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.*
659 17:481–485.

660 Ohta T. 1971. Associative overdominance caused by linked detrimental mutations. *Genet. Res.* 18:277–
661 286.

662 Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of
663 humans and chimpanzees. *Nature* 441:1103–1108.

664 Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for widespread GC-
665 biased gene conversion in eukaryotes. *Genome Biol. Evol.* 4:675–682.

666 Peterson AL, Miller ND, Payseur BA. 2019. Conservation of the genome-wide recombination rate in
667 white-footed mice. *Heredity* 123:442–457.

668 Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK, Nachman MW. 2012.
669 Adaptive evolution and effective population size in wild house mice. *Mol. Biol. Evol.* 29:2949–2955.

670 Phung TN, Huber CD, Lohmueller KE. 2016. Determining the Effect of Natural Selection on Linked
671 Neutral Divergence across Species. *PLoS Genet.* 12:e1006199.

672 Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner
673 AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history.
674 *Nature* 499:471–475.

675 Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.js

676 Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* 6:256–258.e1.

677 Sardell JM, Cheng C, Dagilis AJ, Ishikawa A, Kitano J, Peichel CL, Kirkpatrick M. 2018. Sex
678 Differences in Recombination in Sticklebacks. *G3* 8:1971–1983.

679 Smalec BM, Heider TN, Flynn BL, O'Neill RJ. 2019. A centromere satellite concomitant with extensive
680 karyotypic diversity across the *Peromyscus* genus defies predictions of molecular drive.
681 *Chromosome Res.* [Internet]. Available from: http://dx.doi.org/10.1007/s10577-019-09605-1

682 Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet. Res.* 23:23–35.

683 Smukowski CS, Noor MAF. 2011. Recombination rate variation in closely related species. *Heredity*
684 107:496–508.

685 Stankowski S, Chase MA, Fuiten AM, Rodrigues MF, Ralph PL, Streisfeld MA. 2019. Widespread
686 selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS*
687 *Biol.* 17:e3000391.

688 Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–
689 460.

690 Thybert D, Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, Kolmogorov M,
691 Janoušek V, Akanni W, et al. 2018. Repeat associated mechanisms of genome evolution and
692 function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Res.* 28:448–459.

693 Tigano A, Colella JP, MacManes MD. 2020. Comparative and population genomics approaches reveal
694 the basis of adaptation to deserts in a small rodent. *Mol. Ecol.* [Internet]. Available from:
695 http://dx.doi.org/10.1111/mec.15401

696 Tigano A, Jacobs A, Wilder AP, Nand A, Zhan Y, Dekker J, Therkildsen NO. 2020. Chromosome-level
697 assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and
698 structural genetic variation. *Cold Spring Harbor Laboratory* [Internet]:2020.10.27.357293.
699 Available from: https://www.biorxiv.org/content/10.1101/2020.10.27.357293v1.abstract

700 Van Belleghem SM, Baquero M, Papa R, Salazar C, McMillan WO, Counterman BA, Jiggins CD, Martin
701 SH. 2018. Patterns of Z chromosome divergence among Heliconius species highlight the importance
702 of historical demography. *Mol. Ecol.* 27:3852–3872.

703 Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome
704 sequences. *Genome Res.* 27:757–767.

705 Wiehe TH, Stephan W. 1993. Analysis of a genetic hitchhiking model, and its application to DNA
706 polymorphism data from Drosophila melanogaster. *Mol. Biol. Evol.* 10:842–854.

707 Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc.*
708 *Natl. Acad. Sci. U. S. A.* 110:E1743–E1751.

709 Yeaman S, Otto SP. 2011. Establishment and maintenance of adaptive genetic divergence under
710 migration, selection, and drift. *Evolution* 65:2123–2129.

711

Figures and Tables

Fig.1. Plots showing the relationship between $\log_{10}$-transformed chromosome size (bp) and sequence divergence among species within the *Peromyscus*, Hominidae and *Mus* clades. On the left panel (A, C, E), one representative comparison from each of the *Peromyscus*, Hominidae, and *Mus* clades are displayed (see Figure S2, S3, and S4 for all comparisons). The comparison of the same species pairs are represented on the right panel but the query and reference species are inverted in plots B, D, F to highlight that in the *Mus*, but not in *Peromyscus* and *Hominidae* clades, the choice of the reference genome affects the correlation between chromosome size and *d*. In the bottom panel, the comparison between *Mus spretus* and *M. pahari* is shown, with *M. pahari* as reference on the left (E) and with *M. spretus* as reference on the right (F).

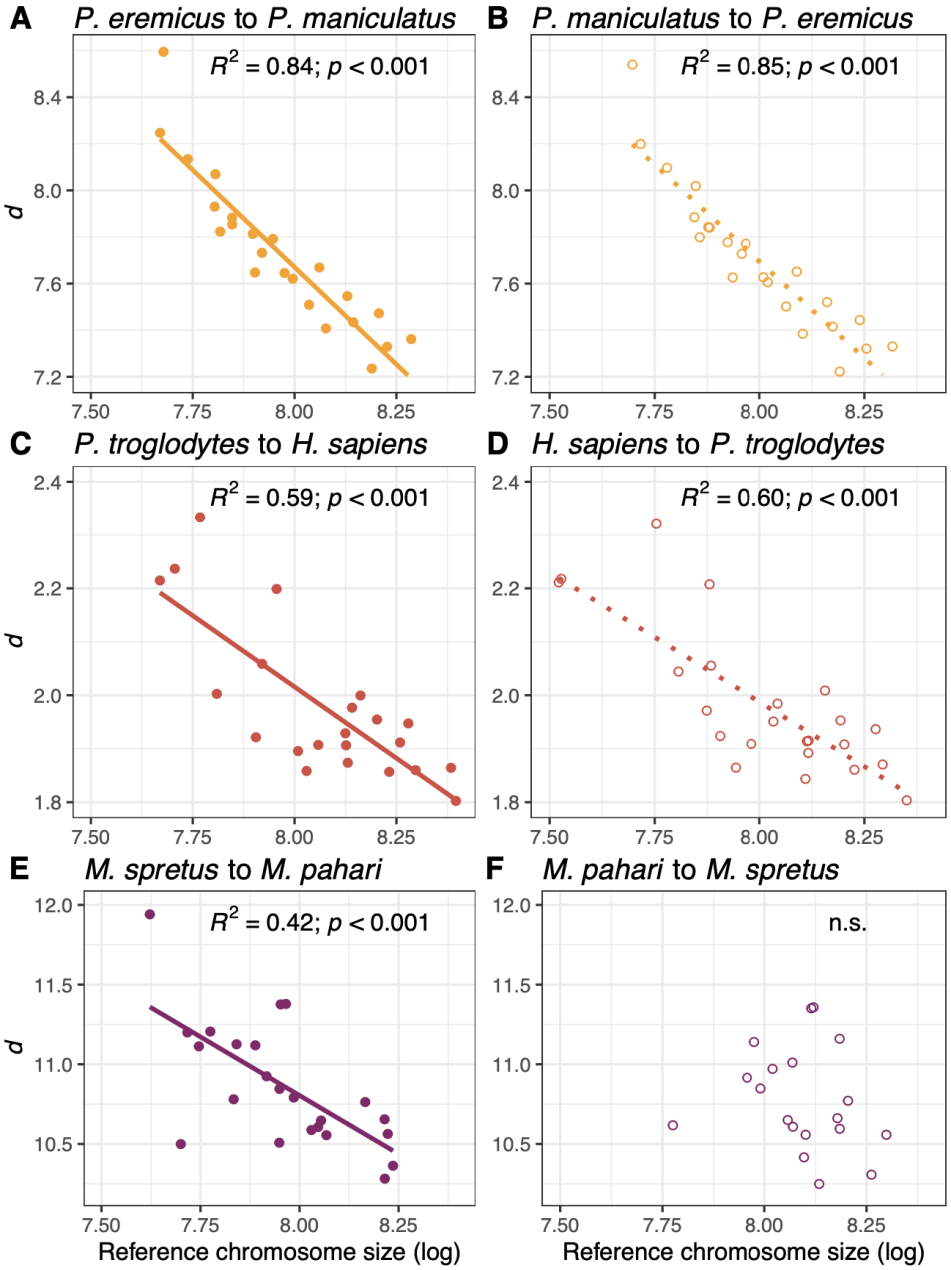


Fig. 2. Boxplots summarizing results from evolutionary simulations on the relationship between recombination rate and π after $20N_e$ generations in the ancestral population (panel A on the left) and d one generation after the split after a mild bottleneck (0.5 of ancestral N_e) between the two daughter populations (panel B on the right) in each of three simulated ancestral N_e. Boxplots refer to the results from the models with selection and the dashed line shows the results from the neutral models. Here are the results with models without gene conversion as no significant differences were found between models with and without gene conversion.

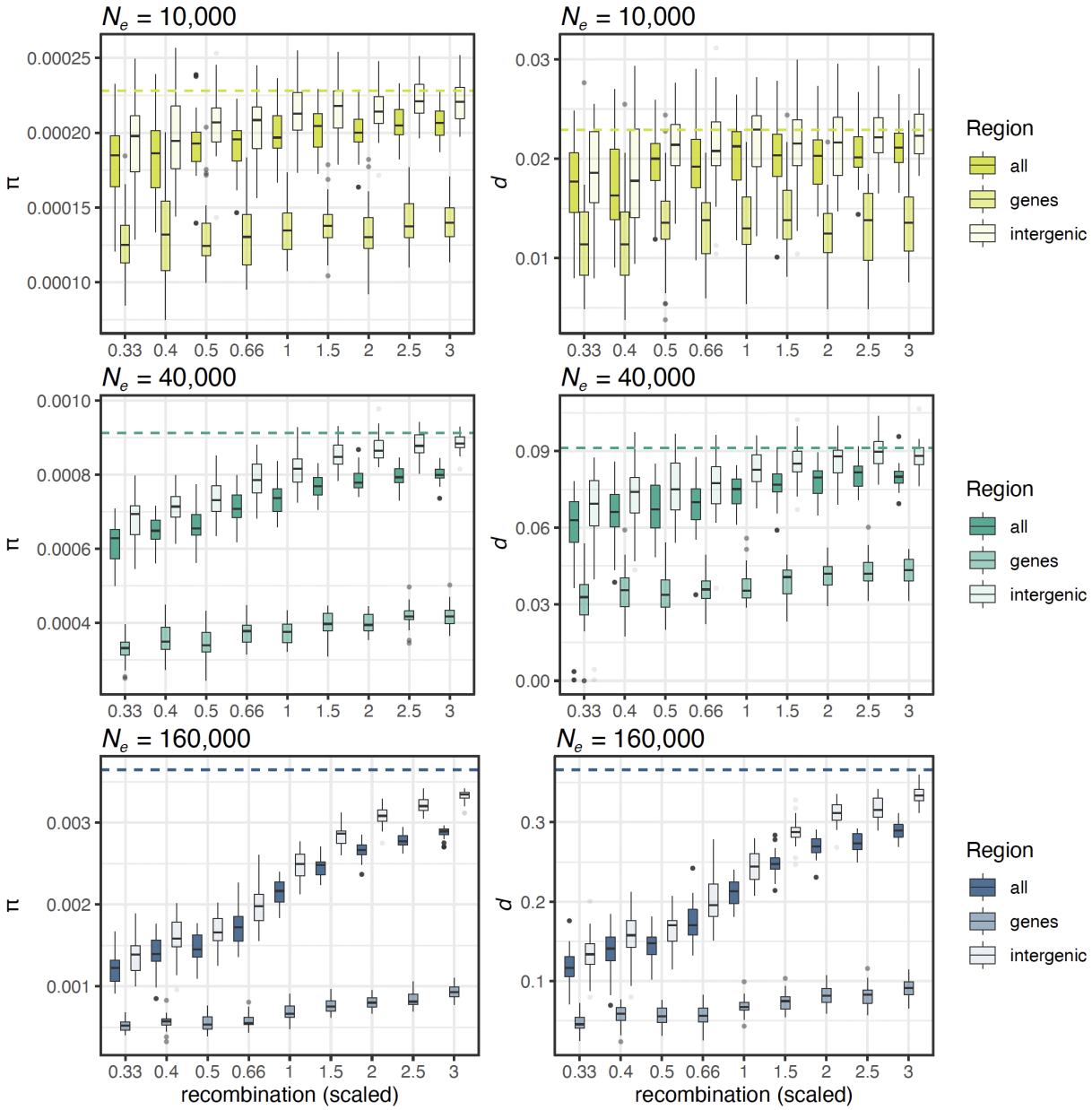
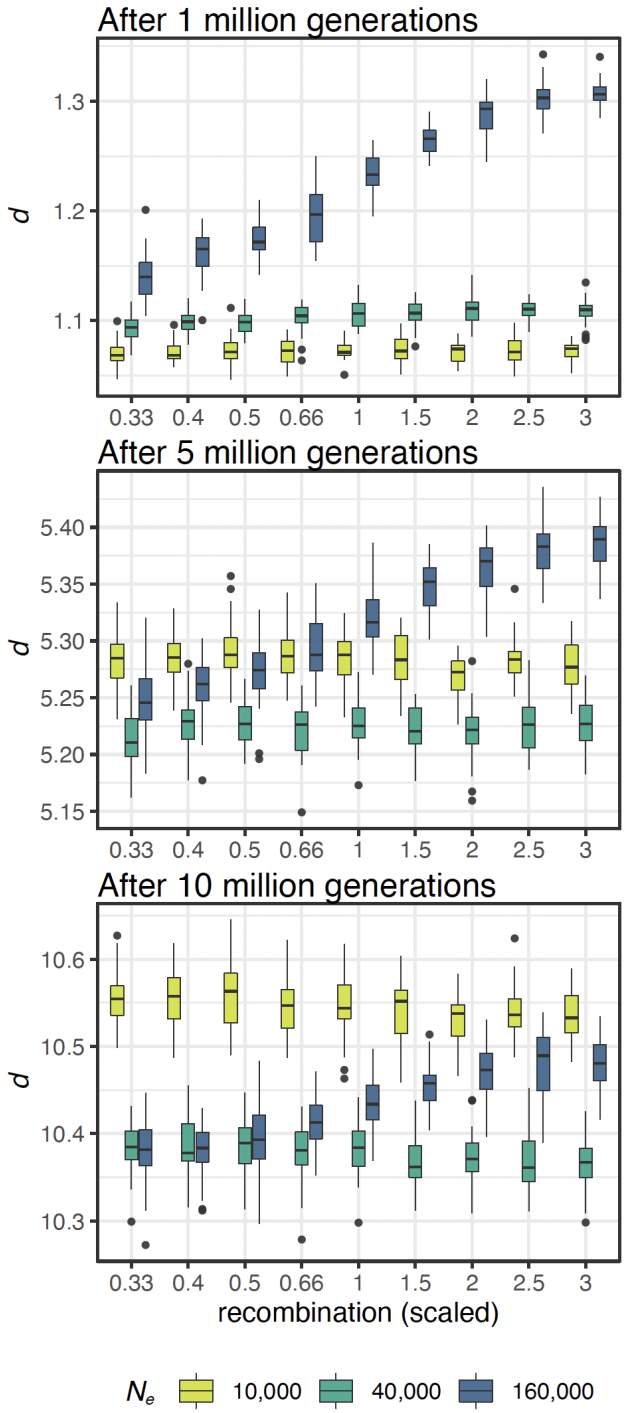
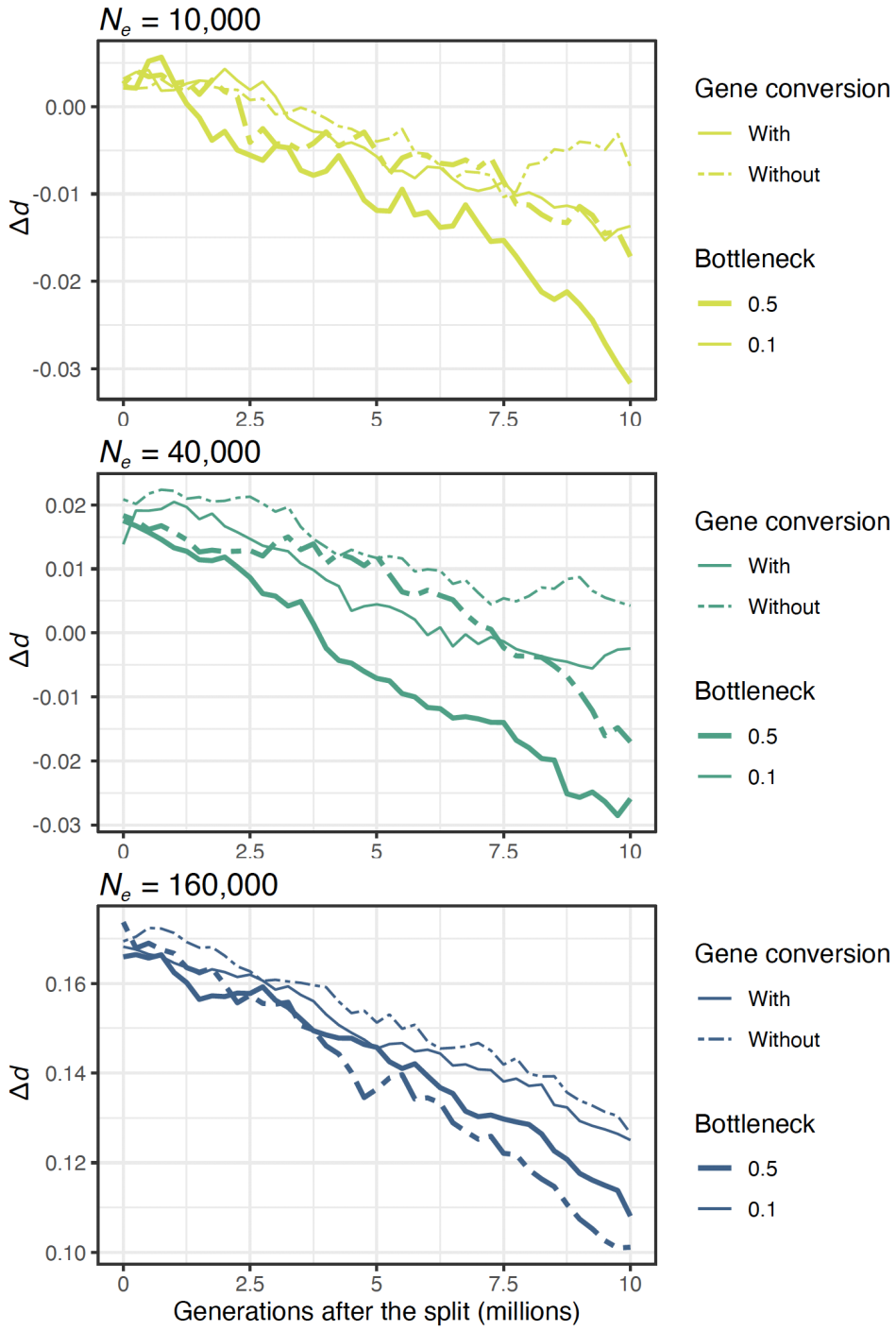


Fig. 3. Boxplots summarizing results from evolutionary simulations on the relationship between recombination rate and *d* in models with selection and without gene conversion in each of three simulated ancestral N_e and three time points after the split from the ancestral pop_A and a mild bottleneck. Gene conversion was not included in these models as no significant differences were found between models with and without gene conversion. (See Fig. S5 for comparisons with neutral models and models with a more severe bottleneck).



742 **Fig. 4.** Plots showing the decay of Δd over time in the evolutionary simulations based on the models with
743 and without gene conversion, and with a mild (0.5) and a severe bottleneck (0.1) for each of the three
744 simulated ancestral N_e.



745
746

747

748

749 **Table 1.** Summary of parameters used in evolutionary simulations.

Variable	**Values**	**Scaled values as in simulations (x25)**
Mutation rate	$5.7*10^{-9}$	$1.42*10^{-7}$
Mean recombination rate *r*	10^{-8}	$2.5*10^{-7}$
Gene conversion rate	*r*/3	*r*/3
Gene conversion tract length	440 bp	440 bp
Selection coefficient *s*	$\pm 15.625*10^{-3}$	$\pm 15.625*10^{-3}$
Relative frequency of neutral, deleterious, and advantageous mutations	0.3, 1, and 0.0005	0.3, 1, and 0.0005
Selection model	neutral/with selection	neutral/with selection
N_e of pop_A	10,000/40,000/160,000	400/1,600/6,400
Reduction of N_e of pop_1 and pop_2 relative to pop_A	0.5/0.1	0.5/0.1
Recombination rates	0.33*r*/0.4*r*/0.5*r*/0.66*r*/*r*/1.5*r*/2*r*/2.5*r*/3*r*	0.33*r*/0.4*r*/0.5*r*/0.66*r*/*r*/1.5*r*/2*r*/2.5*r*/3*r*
Gene conversion	yes/no	yes/no

750