Male mutation bias is the main force shaping chromosomal substitution rates in monotreme mammals

Authors: Vivian Link[1], Diana Aguilar-Gómez[2], Ciro Ramírez[2], Laurence D. Hurst[3], and Diego Cortez[2*]

Institutions:

1) Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland.
2) Center for Genomic Sciences, UNAM, Cuernavaca, México.
3) The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, Somerset, BA2 7AY, UK.

Authors' email addresses:

Vivian Link: vivian.link@sunrise.ch

Diana Aguilar: daguilar@lcg.unam.mx

Ciro Ramírez: cramirez@lcg.unam.mx

Laurence D. Hurst: l.d.hurst@bath.ac.uk

Diego Cortez: dcortez@ccg.unam.mx

*__Author for Correspondence__: Diego Cortez, Center for Genomic Sciences, UNAM, Cuernavaca, México. + 52 (777) 313 4152. dcortez@ccg.unam.mx.

Abstract

In many species, spermatogenesis involves more cell divisions than oogenesis, and the male germline, therefore, accumulates more DNA replication errors, a phenomenon known as male mutation bias. The extent of male mutation bias (α) is estimated by comparing substitution rates of the X, Y and autosomal chromosomes, as these chromosomes spend different proportions of their time in the germlines of the two sexes. Male mutation bias has been characterized in placental and marsupial mammals as well as birds, but analyses in monotremes failed to detect any such bias. Monotremes are an ancient lineage of egg-laying mammals with distinct biological properties, which include unique germline features. Here, we sought to assess the presence and potential characteristics of male mutation bias in platypus and the short-beaked echidna based on substitution rate analyses of X, Y and autosomes. We established the presence of moderate male mutation bias in monotremes, corresponding to an α value of 2.12-3.69. Given that it has been unclear what proportion of the variation in substitution rates on the different chromosomal classes is really due to differential number of replications, we analyzed the influence of other confounding forces (selection, replication-timing, etc.) and found that male mutation bias is the main force explaining the between-chromosome classes differences in substitution rates. Finally, we estimated the proportion of variation at the gene level in substitution rates that is owing to replication effects and found that this phenomenon can explain >68% of these variations in monotremes, and in control species, rodents and primates.

Key words

Male mutation bias, Monotremes, Heterogametic sex chromosomes, Substitution rates

Introduction

The presence of male mutation bias was proposed by Haldane in 1947 (Haldane 1947) as an explanation for why hemophilia-causing mutations are more often inherited from the father than the mother. Haldane's prediction that the mutational rate would be higher in the male compared to the female germline was later supported by the observation that spermatogenesis involves a higher number of replication cycles than oogenesis, leading to an increase in the male mutation rate due to replication errors (Drost and Lee 1995; Ellegren 2007; Hurst and Ellegren 1998; Li et al. 2002; Makova and Li 2002). It has been well documented in great apes and rodents that the maternal gametes go through fewer genome replications than the paternal ones. This is because the maternal population of gametes has already been formed at birth of the future mother and maturation does not include further cell divisions besides meiosis (Chang et al. 1994; Drost and Lee 1995), while in males, by contrast, the formation of gametes (spermatogenesis) continues throughout adult life and necessitates a constant renewal of the initial spermatogenic cell (spermatogonia) pool through mitosis (Kanatsu-Shinohara and Shinohara 2013).

In 1987, Miyata and colleagues developed a framework to quantify male mutation bias (α) by contrasting the rates of neutral evolution on the sex chromosomes and autosomes (Miyata et al. 1987). In the presence of male mutation bias, the Y chromosome, which spends all its time in males, is expected to show the highest evolutionary rate, followed by the autosomes, which spend half their time in males and half in females, and finally the X chromosome, which spends only one-third of its time in males. Thus, α can be estimated using the following three equations, where A, X and Y are the rates of neutral evolution for the autosomes, X chromosome and Y chromosome, respectively:

(1) $\alpha\,(Y/X) = 2/((3X/Y)-1)$

(2) α (Y/A) = 1/((2A/Y)-1)

(3) α (X/A) = (4-(3X/A))/((3X/A)-2)

Miyata's framework assumes that a) the analyzed substitutions are selectively neutral, b) multiple substitutions are accounted for with appropriate correction methods, c) errors during replication are the unique source of genomic variation and d) replication errors are uniformly distributed throughout the genome. Provided that these conditions are met, the three equations should give the same estimate of α, which should furthermore equal the ratio of male over female germ cell divisions. Based on Miyata's equations, the male mutation bias has been estimated to be high in human and chimpanzee (α > 4) (Conrad et al. 2011; Kong et al. 2012; Makova and Li 2002; Presgraves and Yi 2009; Shimmin et al. 1993a; Venn et al. 2014), and moderate in mouse (α = 1-3.5) (Li et al. 2002; Malcom et al. 2003; McVean and Hurst 1997; Sandstedt and Tucker 2005; Smith and Hurst 1999; Wolfe and Sharp 1993). However, a number of diverging studies found that the three estimators of α are highly discordant and that autosomes and Y chromosomes evolved at similar rates in mouse and rat (McVean and Hurst 1997; Pink et al. 2009; Smith and Hurst 1999). Hence, these studies challenged previous results and suggested that Miyata's equations may provide incorrect estimates, as they do not consider other factors besides replication errors that might contribute to differences in substitution rates between chromosomal classes [*e.g.,* gene conversion, replication timing or recombination rates (Pink et al. 2009; Pink and Hurst 2010; Shimmin et al. 1993b)].

In a recent genome-wide study based on autosomal and X chromosome sequences, Wilson Sayres et al. (Wilson Sayres et al. 2011) detected strong signals of male mutation bias across 32 placental mammals, in particular in species with long generation times (Wilson Sayres et al. 2011), consistent with previous observations in birds (Bartosch-Harlid et al. 2003). However, when the authors repeated the analysis for the human and chimpanzee genomes while including the full sequence of the Y chromosome, they

obtained discrepant estimates of α with Miyata's three equations (Wilson Sayres et al. 2011). Similar observations have also been made for birds and rodents (Axelsson et al. 2004; Pink et al. 2009). These results further support the idea that replication errors might not explain all of the variation observed between chromosomal classes, confirming that other factors should be taken into consideration (Pink et al. 2009; Pink and Hurst 2010; Shimmin et al. 1993b). Given that the Y chromosome is more likely to be prone to background selection and hitchhiking effects, it has been assumed that these factors could influence the observed substitution rate of this chromosome, resulting in discrepant α estimates. These processes modulate effective population size of the entire non-recombining section of a Y chromosome thus affecting the fate of weakly deleterious mutations (but should have little influence on perfectly neutral sites, where the substitution rate should be equivalent to the mutation rate (Birky and Walsh 1988)). The effects on estimation of α of differential activity of processes such as gene conversion have been relatively little explored.

Given the exceptionalism of the Y chromosome it has been often assumed that the X/A comparison is more reliable (Wilson Sayres et al. 2011). However, the X chromosome is affected by selection due to the decay of the Y chromosome (strong purifying selection in males) (Delgado et al. 2009); has unusual replication timing (Pink and Hurst 2010), it being one of the last ones to be replicated which should increase its substitution rate; has lower content of CpG sites (Saxonov et al. 2006) that would diminish the substitution rate and possibly lower germline transcription-coupled repair, which may also modulate the substitution rate, as X-linked genes tend to be relatively tissue specific (Lercher et al. 2002). Hence, it is uncertain whether the X/A comparison would provide the most accurate α estimate. Thus comparisons employing Y we argue would be valuable. An important reason for the lack of the Y (and W) chromosomes from male mutation bias studies is because these chromosomes are often missing from whole-genome sequencing projects owing to their repeat-rich nature and because studies often prefer

to sequence the homogametic sex to maximize read counts for each chromosome.. The recurrent lack of information from these sex chromosomes has therefore limited the study of Miyata's equations and the study of other potential factors influencing α.

We recently reconstructed Y-linked transcripts across all three major mammalian clades (placentals, marsupials, and monotremes) (Cortez et al. 2014). Based on synonymous substitutions within X/Y gametologs, we were able to detect signatures of male mutation bias in placental mammals and marsupials, but not in monotremes, represented by the platypus and short-beaked echidna. However, the limited number of exonic sequences from 14 XY gametologs, together with potential negative selection acting at synonymous sites limited the statistical power of this analysis and prevented us from drawing firm conclusions regarding the presence of male mutation bias in monotreme mammals (Cortez et al. 2014).

Monotremes show several biological peculiarities that include egg-laying, spurs and venom production (only platypus (Wong et al. 2013)). Platypus also shows many genomic-specific features (Warren et al. 2008) that include an atypical sex system composed of ten different chromosomes (Rens et al. 2004; Rens et al. 2007), which originated independently from the X and Y chromosomes in other mammalian lineages (Veyrunes et al. 2008). Furthermore, even the germline in this lineage shows remarkable peculiarities since platypus lacks MSCI (Daish et al. 2015). Therefore, these unique features of monotremes may also include an unusual male mutation bias.

We, therefore, decided to perform an extended analysis of male mutation bias in platypus and echidna. By calculating substitution rates of the monotreme X, Y and autosomal chromosomes based on curated intronic alignments, we demonstrate the presence of male mutation bias in monotremes. We estimate

an α ranging from 2.12 to 3.69 for monotremes, which corresponds to a moderate bias. Moreover, our analyses are also useful to estimate the proportion of variation in substitution rates that is owing to replication effects. The results in monotremes and mammalian control species suggest that male replication bias might account for >68% of the observed differences at the gene level. We used the same type of short genomic reads for all species, thus allowing us to apply the same methodology. Finally, it is important to note that our methods can be applied to non-model species with poor genome assemblies and may be used to further illuminate patterns of male mutation bias across vertebrates.

Material and Methods

Genomic assemblies

In two previous studies (Cortez et al. 2014; Necsulea et al. 2014), we generated paired-end genomic reads for platypus, echidna, marmoset and rat using standard Illumina protocols (Truseq DNA) and HiSeq2000 sequencing platform. Male mutation bias has been well characterized in primates and rodents, thus allowing us to use them as control species. Supplementary Table 1 contains detailed information regarding the genomic data used in this study (number of reads, GenBank accession numbers, etc.).

In order to obtain male genomic assemblies for platypus, marmoset and rat that would increase the chances of having long Y-linked scaffolds and would reduce the chances of having chimeric Y sequences (sequences that combine both X and Y gametolog sequences), we applied a male-female subtraction approach similar to the one described by Cortez et al. (Cortez et al. 2014). First, we mapped the Illumina male genomic paired-end reads from platypus, marmoset and rat to their corresponding female reference genomes; reference genomes were downloaded from the Ensembl database (release 77)

(Flicek et al. 2014). We allowed two mismatches per read, and retained only those read-pairs of which none of the paired reads were mapped. We then used SOAP-de novo (Luo et al. 2012) (kmer = 31) to assemble the unmapped reads into scaffolds. As an alternative, we also assembled the unmapped reads using kmers of 21 and 25. However, the two genomes assembled with these alternative kmers showed significantly shorter introns (Supplementary Figure 1) and were thus not used for final analyses. Finally, we located Y-linked scaffolds by a targeted search (at >= 99% identity using blastn (Altschul et al. 1990)) of the exons of the Y-specific cDNAs that we reported previously (Cortez et al. 2014).

Echidna does not have a reference genome available. Thus, we assembled a female genome with SOAP-de novo (kmer = 31) using the entire set of female Illumina genomic reads. We also assembled a female genome for marmoset and rat with SOAP-de novo (kmer = 31) using the entire set of female Illumina genomic reads. For the male echidna assembly, we first mapped the Illumina male genomic reads (two mismatches allowed per read) to the newly assembled female genome. The unmapped paired-end reads were then used to assemble Y-specific scaffolds with SOAP-de novo (kmer = 31). We located Y-linked scaffolds by a targeted search (at >= 99% identity using blastn) of the exons of the Y-specific cDNAs that we reported previously (Cortez et al., 2014).

Since the echidna female assembly was highly fragmented, and thus the subtraction approach could have been inefficient in removing reads that are shared between males and females, we verified whether the Y-linked scaffolds in the echidna might be formed of chimeric sequences. We developed two approaches of increased stringency to generate male genomes: From the subset of reads that did not map to the assembled echidna female genome, we removed all the reads (and their pairs) that showed kmers of 30 or 40 nucleotides shared with any of the echidna female genomic reads. The remaining reads, which represent male-specific kmers, were then assembled into scaffolds using SOAP-

de novo (kmer = 31). The two kmer-derived genome assemblies were extremely fragmented (millions of contigs). However, all known Y exons and introns that we obtained with the less-stringent filtering approach (the one for which we did not use kmers) were found in these two alternative genome assemblies distributed however among various smaller contigs, thus confirming that our original Y scaffolds were not chimeric. The genome assembly that we obtained without the kmers filtering showed scaffolds with Y exons linked to longer intronic sequences (total length = 54,000nt; 30 kmer = 11,000nt and 40kmer = 14,000nt) and was thus used for all further analyses.

Collecting data from reference genomes

We studied 1-to-1 orthologous genes in primates, rodents, and monotremes (see Supplementary Table 2, for a detailed list of the genes we analyzed). In platypus we decided to work with the X-gametologs we previously identified (Cortez et al. 2014) and well-annotated X_5-linked genes; X_5 is the oldest X chromosome shared between monotremes and it is fully differentiated from the Y_5 chromosome (Cortez et al. 2014; Rens et al. 2007; Warren et al. 2008). For platypus, human and mouse we downloaded the protein-coding exonic and intronic sequences for the X and autosomes from the Ensembl database (release 77). Finally, we also downloaded all known Y protein-coding exons and intronic sequences for human and mouse Y genes from the Ensembl database.

Annotation of scaffolds

We selected all the scaffolds in the platypus, echidna, marmoset and rat male assemblies that mapped to known Y genes and Y transcripts. We then chose the best scaffolds based on three features: a) best match accuracy (>= 99% identity), b) the maximum amount of cDNA they covered and c) maximum length of the scaffold. We aligned the selected scaffolds to the cDNA and CDS of the Y genes/transcripts using MUSCLE (3.8.31) (Edgar 2004). We annotated the resulting concatenated alignments of the

scaffolds as follows: sequences that matched to the CDS were marked as exons, the sequences that mapped only to the cDNA as UTRs, and the sequences in between the exons as introns. We applied the same above-mentioned strategy to align the female scaffolds to the cDNAs sequences of annotated X and autosomal genes.

Obtaining orthologous intronic alignments

To limit the risk of including non-orthologous positions in the alignments, we considered only the intronic sequences that were located in the same scaffolds as conserved 1-to-1 orthologous exons in closely related species. Thus, for the Y sequences in monotremes, we first mapped the exons from platypus to the scaffolds of echidna with blastn and selected the best pairs of scaffolds with a minimum identity score of 90%. When we found more than one matching scaffold with the same identity score, we selected the longest one. We ordered the Y scaffolds of echidna following the structure and strand orientation of the platypus Y transcripts (gene annotations were based on Ensembl annotations). Then, we used Lagan20 (Brudno et al. 2003), an alignment program designed to work on non-coding sequences, to align the concatenated exonic and intronic sequences from platypus and echidna. We followed the same protocol to align the echidna X and autosomal scaffolds to the orthologous sequences in platypus, to align the marmoset Y, X and autosomal scaffolds to the orthologous sequences in human and to align the rat Y, X and autosomal scaffolds to the orthologous sequences in the mouse. We then aligned the 1-to-1 orthologous intronic sequences with Lagan20 and removed ambiguous positions using Gblocks (Talavera and Castresana 2007). Gblocks scans a multiple sequence alignment using a sliding window of 10 positions (minimum block) and removes segments that are misaligned and may represent non-orthologous regions. We excluded the first protein-coding exon and the following intron of all genes from the alignments because these introns often contain regulatory elements (Chamary and Hurst 2004).

Curating introns for hidden exons and other non-neutrally evolving positions

Since the annotation of exons in our introns was based on reported cDNAs that were either obtained from RNAseq data (Cortez et al. 2014) or derived from non-exhaustive database annotations (especially in the case of platypus), it is not unlikely that introns may contain hidden exons (Pink et al. 2009). Genuine intronic sequences are expected to show the same frequencies of single nucleotide changes along their entire sequences (Pink et al. 2009), as opposed to exons that are expected to have more nucleotide changes at the third codon position due to the redundancy in the genetic code. In order to remove hidden exons from the alignments, we calculated the frequencies of single nucleotide changes at the first, second and third codon positions from all annotated exons of Y, X and autosomal genes from monotremes, primates, and rodents (annotations were downloaded from the Ensembl database). We measured an average 10% increase in single nucleotide changes at the third codon position relative to the first and second position in the annotated exons. Based on these expected frequencies of single nucleotide changes in exons, we scanned the introns using an overlapping sliding windows of 51 nucleotides (our definition of the shortest exon) to account for all possible reading frames and removed those windows that: 1. did not have stop codons, 2. were on the same strand orientation as the other exons in that gene, and 3. showed at least a 10% increase frequency of substitutions in the third codon position. This method removed in average 2% of windows.

We later removed from the alignments the first 20 intronic nucleotides flanking the exons in order to remove regulatory sites such as splicing sites and splicing enhancers (Pink et al. 2009). Methylated cytosines followed immediately by a guanine have an increased likelihood of being transformed into a thymine, resulting in a C-to-T transition that is independent of replication (Jabbari and Bernardi 2004). Consequently, the effect of male mutation bias is obscured at CpG sites. We, therefore, removed all CpG sites from our alignments because male mutation bias is expected to be much lower at CpG sites than at

non-CpG sites (Taylor et al. 2006). In order to remove the signal contained at CpG sites, we screened both strands of the intronic sequence alignments and removed all CpG positions. Estimates were lower after this step, which we considered as an important indication that CpG sites were influencing the calculations.

The alignment program Lagan20 will correctly align the orthologous regions of two sequences, but it will also create gap-rich alignments of intronic segments that are not orthologous (lineage-specific insertions and deletions are common in intronic sequences). We used Gblocks (Castresana 2000) to extract the orthologous alignments and to remove the parts of our alignments that were gap-rich regions, representing spurious alignments. This step is especially important if the species divergence time is big, given that these alignments are expected to show a higher noise-to-signal ratio. We optimized the parameters of Gblocks for our study as follows*:* "*allowed gap positions*" *= all* and "*minimum length of blocks*" *= 30 nucleotides*. Finally, only those genes showing alignments >1,000bp were considered for further analyses. This threshold was defined to avoid extreme d_S values due to short sequences.

Calculating the substitution rates for Y, X, and autosomes

Although we excluded from our alignments all those positions that seemed to be under evolutionary constraints, we could not exclude the possibility that some other intronic regions could have low substitution rates (e.g., undetected micro-exons) or, alternatively, could represent mutation hotspots. In order to minimize the noise that could be introduced by the fluctuations in substitution frequencies within introns and within intronic positions, we used the nonparametric double bootstrap approach described by Axelsson et al. (Axelsson et al. 2004), which bootstraps the intronic alignments by both introns and sites: For a given chromosomal class (Y, X, and autosomes), we randomly resampled introns until we obtained the same intron number as in the original data set. From these, we randomly

resampled sites until we obtained the same amount of sites as the randomly sampled introns would have if they were concatenated. We repeated this procedure 1'000 times.

For each random alignment, we calculated the substitution rate using the Tamura-Nei model with the baseml program, implemented in the PAML44 package (Yang 1997). The Tamura-Nei is a model of DNA sequence evolution that corrects for multiple hits and takes into account the substitution rate differences between nucleotides and the inequality of nucleotide frequencies (Tamura and Nei 1993). Moreover, the Tamura-Nei model produces good estimates and outperforms other models in simulated data (Tamura and Kumar 2002) when GC content is stationary. Therefore, we verified that our sequences showed stationary GC content using the neighbor-dependent evolution model(Arndt et al. 2003).

Since Y, X and autosomal introns show important differences in GC content (Supplementary Figures 2 and 3; Supplementary Table 3), we decided to verify that the calculations of substitution rates and hence the pattern of male mutation bias was not affected by the differences in GC content. Thus, we adapted the double-bootstrap approach described above to re-analyze the intronic sequences in monotremes: We fixed the number of randomly chosen A/T positions to be equal to the number of randomly chosen G/C positions in the alignments. From these new alignments, with fixed GC% = 50 for all chromosomal classes (Supplementary Figures 2 and 3; Supplementary Table 3), we then calculated the substitution rates using the Tamura-Nei model and the baseml program implemented in the PAML44 package.

Analyzing male mutation bias and other confounding forces

In order to study the influence on chromosomal substitution rates of male mutation bias and other confounding forces we constructed a Generalized Linear Model (GLM). We tested whether differences in substitution rates between genes from the three chromosomal classes are associated with a variety of forces. We collected data for 698 1-to-1 orthologous genes between mouse and rat; rodents were the only species for which these variables were available. We thus worked with 6 Y-linked genes (maximum number of Y genes for which we could find data), 346 X-linked genes (maximum number of X genes for which we could find data) and 346 randomly selected autosomal genes. The predictors were: d_N/d_S ratios (as proxy of selection), germline expression levels (as proxy of transcription-coupled repair), replication timing and male mutation bias based on the time each chromosome spends in the male germline (Supplementary Table 4). We also gathered recombination data for mouse (Smagulova et al. 2011), but found insufficient overlap between the replication information and the genomic coordinates of the selected intronic sequences for which we collected all other variables.

For each of the 698 genes included in the model we obtained: (i) Pre-computed d_N/d_S ratios from the Ensembl database. (ii) Expression values were calculated for germline tissues, specifically, spermatids and spermatocytes (Soumillon et al. 2013). Briefly, reads were mapped to the reference genome (Ensembl version 83) using Hisat2 (Kim et al. 2015) and resulting FPKM values were then obtained and normalized with Cufflinks (Trapnell et al. 2013); we used $\log_2$ transformed median values across tissues. (iii) The www.replicationdomain.org database provides full sets of high-resolution maps of replication timing across the mouse genome. We calculated the median of replication timing based on the genomic coordinates of the 698 1-to-1 orthologous genes. We used replication timing data for early developmental cells (differentiation state *ESC*) because replication timing in male germline was

unavailable. We note, however, that as replication timing can change, our estimates of the impact of replication timing are likely to be minimum estimates. (iv) In order to include male replication bias as a predictor, we used the time each chromosome spends in the male germline, that is, Y-linked sequences spend 100% of their time in the male germline, whereas autosomes spend 50% of their time in the male germline and the X chromosome spends 1/3 of its time in the male germline. For this reason, we used the following values for the three chromosomal classes: Y = 1, X = 1/3 and A = 1/2. The response variable was both the mean substitution rate for the chromosomal classes and a windowed substitution rate for the chromosomal classes.

We worked with three alternative models. We first defined in the model the mean substitution rate as response variable and included (i) all values for the 698 genes for each predictor and (ii), the six central values (six values around the median) of each predictor, thus the three chromosomal classes contained the same number of values (the Y chromosome has six genes with available data). We tested a third model where we sorted the response' and the predictors' variables and then divided their values into 6 windows of equal size. We calculated the mean of each window and included these values in the model. We examined the three GLM using the following formula: glm(observed.substitution.rate.median.or.windowed ~ + dN.dS + expression.level + replication.timing + time.in.male.germline, family = gaussian). Variables followed an approximately normal distribution so a Gaussian distribution for the GLM was specified (Supplementary Figure 4).

Calculating α

We used the medians of the 1'000 bootstrapping rounds as the substitution rates for Y, X, and autosomes and we introduced these values into the three equations by Miyata (Miyata et al. 1987) in

order to obtain the empirical values of α. Moreover, after each of the 1'000 bootstrapping rounds three α values were calculated, and the resulting distributions served to calculate the 95% confident intervals of the median by selecting the 1 + n/2 + sqrt(n)/2 position as upper confidence level and the n/2 - sqrt(n)/2 position as the lower confidence level (n = sample size, *i.e.* 1000 values).

Estimating the proportion of variation in substitution rates due to male-biased mutation

The proportion of variation in substitution rates was obtained from 1,000 resampling rounds. For each round, we randomly selected six autosomal, six X-linked genes from the gene pool of 1-to-1 orthologous genes between human-marmoset, mouse-rat and platypus-echidna. The six Y-linked were always selected. We worked with six genes for each chromosomal class because we had only six Y genes/transcripts that we could use for all the species with >1,000bp of aligned and curated intronic sequences. We then estimated the variance in substitution rates of the six autosomal, six X-linked and six Y-linked genes. This value was considered as the variance from the initial gene set. Subsequently, we adjusted the Y and X rates using two correcting factors that were based on the time each chromosome spends in the male germline. The factor by which we adjusted the Y rates to correct for the acceleration of these sequences was a reduction of 50% of the original value. The factor by which we adjusted the X rates to correct for the slower rate of these sequences was an increase in 33% of the original value. We then calculated a second variance with the unchanged substitution rates of the six autosomal genes and the adjusted substitution rates of the X and Y genes. This value was considered as the variance from the adjusted gene set. Lastly, we defined an index of the amount of change as follows: (*variance from the initial gene set – variance of the adjusted gene set*) / *variance from the initial gene set*.

All statistical tests were performed using the R package, standard libraries. Data was plotted using the R package, 'ggplot2' library. Code used in this work can be downloaded from the following public FTP server: ftp://kanan.ccg.unam.mx/PDG/dcortez/Link_etal/.

Results

Assembly and alignment of intronic sequences

We estimated the degree of male mutation bias in monotremes (platypus and echidna) and, for comparison, in selected primates (human and marmoset) and rodents (mouse and rat) based on intronic sequences from the X, Y and autosomal chromosomes. Intronic sequences generally experience less purifying selection than synonymous sites, and intronic substitution rates therefore usually constitute better proxies for chromosomal mutation rates (Hurst and Ellegren 1998; Lercher et al. 2001). We selected Y and X sequences that are located outside the pseudoautosomal region because sex chromosomes still recombine at this region during meiosis.

Because there is no reference genome for the short-beaked echidna, and the five Y chromosomes are missing from the published platypus genome assembly (Warren et al. 2008), we devised a strategy based on RNA and DNA sequencing reads to assemble intronic sequences from the three chromosome classes (Methods). In brief, we first used Illumina short genomic reads (Cortez et al. 2014; Necsulea et al. 2014) to assemble the female echidna, marmoset and rat genomes, and identified X-linked and autosomal scaffolds guided by orthologous protein-coding sequences in the platypus, human and mouse genomes. Next, we assembled *de novo* male-specific DNA scaffolds from platypus, echidna, marmoset and rat using our previously published subtraction approach (Cortez et al. 2014) and extracted the introns of previously annotated Y-linked genes. Fully sequenced Y chromosomes were already available

for human and mouse (Church et al. 2009; Soh et al. 2014; Tilford et al. 2001). The primate and rodent species were selected because male mutation bias has been well characterized in these species. However, our choice of species was also influenced by two additional factors: First, the same type of short genomic reads that we used for echidna were available also for marmoset and rat, thus allowing us to apply the same methodology to all species pairs (while relying on the reference genomes of human, mouse and platypus). Second, given the uncertainty associated with the divergence time of echidna and platypus that ranges from 17 to more than 50 million years (Rowe et al. 2008; Warren et al. 2008), the mouse-rat (≈25 million years ago, Mya) and the human-marmoset (≈42.6 Mya) comparisons could help to assess whether the type and amount of data we collected for monotremes would be adequate to detect male mutation bias.

To avoid alignment of non-orthologous sequences, we only considered introns that were flanked by conserved orthologous exons from well-annotated genes (Methods). The alignments were trimmed to remove sequences that might otherwise bias our estimates of the neutral substitution rate, including annotated exons, first introns (that often contain regulatory sequences), potential hidden exons, non-orthologous positions, fast-evolving CpG sites, and sequences involved in splicing regulation (Methods). We estimated intronic substitution rates based on pairwise alignments (Axelsson et al. 2004; Chang et al. 1994; Makova and Li 2002; Miyata et al. 1987; Pink et al. 2009) of X-linked, Y-linked and autosomal 1-to-1 orthologous genes between platypus and echidna, human and marmoset, as well as mouse and rat. For monotremes, we obtained intronic sequences for 50 X-linked genes, 130 autosomal genes and 14 Y transcripts. For primates we obtained intronic sequences for 347 X-linked genes, 11758 autosomal genes and 6 Y genes, while for rodents we obtained intronic sequences for 330 X-linked genes, 9428 autosomal genes and 7 Y genes. A list of the genes used in this study can be found in the Supplementary Table 2. The elevated fragmentation of the platypus genome and the seemingly elevated number of

repeats in the echidna genome explains the lower number of genes for which we could recover sufficient intronic sequences. Nonetheless, we could work with hundreds of sequences in all species, which would allow obtaining balanced values.

Initial evidence of moderate male mutation bias in monotremes

We first analyzed the global autosomal variation in monotremes, primates, and rodents by calculating the substitution rates of all 1-to-1 orthologous genes between platypus-echidna, human-marmoset and mouse-rat (Figure 1). Notably, we observed that all autosomes in all species comparisons evolved more slowly than the Y chromosome and faster than the X, which fall as outliers in the distributions (Figure 1). This observation supports the notion of male mutation bias as a general determinant of chromosomal substitution rates in the three mammalian lineages investigated.

Although we filtered our alignments as stringently as possible, we could not completely exclude the possibility that some sites within them evolved under purifying selection (*e.g.*, as part of undetected exons) or represent mutation hotspots. To minimize the influence of such fluctuations on our estimates of chromosomal substitution rates and in order to correct for the different number of analyzed genes, we generated 1,000 intron alignments for each chromosome class and species pair by bootstrapping for both introns and sites (Axelsson et al. 2004) (Methods; Table 1). For each alignment, we then calculated one single substitution rate under the Tamura-Nei model, after confirming that GC content is stationary (Methods, Supplementary Tables 5 and 6).

The analysis was consistent with male mutation bias in monotremes (Figure 2a; Table 2), with the Y chromosomes evolving significantly faster than the autosomes, which in turn evolve significantly faster than the X chromosomes (Benjamini-Hochberg corrected $P < 0.05$, Welch Two Sample t-test). Although

the X, Y and autosomal chromosomes differ in terms of GC content (median for X-linked introns: 44.8%; Y-linked: 37.1%; autosomal: 40.1%), we did not find that this difference contributed to the observed differences in substitution rates between chromosome classes (Methods; Supplementary Figures 2 and 3; Supplementary Table 3), possibly because we removed the CpG mutations from the analyses.

We also detected signatures consistent with male mutation bias in primates and rodents (Chang et al. 1994; Li et al. 2002; Makova and Li 2002; Venn et al. 2014; Wilson Sayres et al. 2011) (Figure 2b and c; Table 2). Overall, sequence divergences are slightly higher between the two rodents than between the two primates, although these rodents diverged more recently (≈25 Mya) than the primates (≈42.6 Mya) (Hedges et al. 2006), which is consistent with the substantially higher genomic substitution rate per generation in rodents (Li et al. 1996). We observed higher substitution rates of autosomes and X sequences between monotremes and primates (Benjamini-Hochberg corrected $P > 0.05$, Welch Two Sample t-test), which, given the uncertainty of the time platypus and echidna split (17 to >50 Mya (Rowe et al. 2008; Warren et al. 2008)), could support a reduction in the genomic substitution rate on the monotreme lineage, as previously suggested (Warren et al. 2008).

Male mutation bias is the primary force shaping chromosomal substitution rates in monotremes and control species

Confounding forces may influence substitution rates and consequently cause discrepancies in male mutation bias estimates. For instance, purifying selection acting on the Y and X sequences might reduce the mutation load; the number of weakly deleterious mutations could increase on the Y chromosome by background selection and hitchhiking effects when the effective population sizes and recombination rates are low; transcription-coupled repair could reduce the mutation rate of the X chromosome because is more gene-rich than the Y chromosome; finally, late replication timing would increase the

mutation rates of the sex chromosomes. However, male mutation bias and these alternative forces have not been explored in a common statistical framework. We, therefore, decided to verify whether the observed substitution rates of Y, X and autosomes could be significantly associated with purifying selection, transcription-coupled repair, replication timing and time spent in male germline using a dedicated dataset which was only available for rodents (see Methods). We gathered these variables for 1-to-1 orthologous genes between mouse and rat. Next, we decided to explore the robustness of the associations between the predictors and the response variable using different sets of parameters.

We built three GLM, one using all available values and two with the same amount of values for the three chromosomal classes (see Methods). We defined as response variable both the mean substitution rates and a windowed substitution rate for each chromosomal class. We included as predictor variables all potential forces influencing substitution rates. The first GLM, which included all values, returned a highly significant relationship between time spent in male germline and the observed substitution rates ($P <$ 2e-16; odd ratio 1.16, CI-95%: 1.1622-1.1626). The two alternative GLM, which had the same number of values for each chromosomal class, also resulted in a significant relationship between time spent in male germline and the observed substitution rates (P<6.48e-13; odd ratio 1.16, CI-95%: 1.165-1.167). These models showed a significant association between d_N/d_S and the observed substitution rates too ($P <$ 0.00047; odd ratio 1.08, CI-95%: 0.99-1.11). We did not find any significant associations between the observed substitution rates and transcription in the male germline or replication timing. These results suggest that male replication bias is the primary force shaping substitution rates in rodents, although selection is playing a significant role as well. Detailed d_N/d_S patterns across chromosomal classes (Figure 3) reveal that Y sequences are under weaker purifying selection (higher d_N/d_S ratios), which was previously reported based on comparisons of X and Y gametologs (Wilson and Makova 2009). Given that the observed patterns between rodents, monotremes, and primates are remarkably similar (Figures 1

and 2), we can speculate that the results obtained for rodents are consistent with male mutation bias being the main force shaping substitution rates in monotremes and primates as well.

After establishing the relative importance of male mutation bias, we decided to quantify the degree of male mutation bias (α) using Miyata's equations (Equations 1-3). Although theory predicts that the three equations should give the same estimate of α, earlier studies in great apes, rodents and birds showed that this is not the case (Axelsson et al. 2004; Pink et al. 2009; Smith and Hurst 1999; Wilson Sayres et al. 2011). Consistent with these observations, although the estimates of α seem similar, they show non-overlapping confidence intervals for all three species pairs (Table 2). For monotremes, the median α values for the Y/X, Y/A and X/A comparisons are 2.99, 3.51 and 2.35, respectively (Table 2), suggesting moderate mutation bias in this lineage. Discrepancies between α estimates from Y/X, Y/A and X/A comparisons are likely due to confounding forces influencing the three estimates by Miyata, although male replication bias seems to be the main force shaping substitution rates. We calculated the average and range across all three estimates, X/A, Y/A and X/Y as a good indicators of α. Monotremes show moderate male mutation bias, corresponding to an average α value of 2.95 with values ranging from 2.12 to 3.69 (Table 2).

Our estimates for the control species show moderate male mutation bias in the human-marmoset comparison (average α: 2.46, range: 1.71-3.24, Table 2; previous estimates for human-chimpanzee: 4-6 (Kong et al. 2012; Makova and Li 2002; Presgraves and Yi 2009; Shimmin et al. 1993a; Venn et al. 2014); Table 2) and weak male mutation bias in rodents (average α: 1.75, range: 1.53-1.92, Table 2; previous estimates: 1-3.5 (Li et al. 2002; Malcom et al. 2003; McVean and Hurst 1997; Sandstedt and Tucker 2005; Smith and Hurst 1999; Wolfe and Sharp 1993); Table 2). Our human-marmoset value is considerably lower than what was previously been observed for human-chimpanzee (α = 4-6 (Makova

and Li 2002)). This may well reflect the possibility the chimpanzee has a longer generation time and more cell divisions in males, thus strong male bias (Venn et al. 2014), than does the marmoset and the ancestral species intermediate between human and marmoset. Whether the longer divergence times between human and marmoset (~40 million years ago) compared to human and chimpanzee (~6 millions years) is of itself of relevance is unclear.

So far we established that male replication bias seems to be the primary force shaping substitution rates of the three chromosomal classes. However, previous studies showed that substitution rates vary between and within chromosome (Lercher et al. 2001; Malcom et al. 2003; Matassi et al. 1999), which could reflect differential effects of confounding forces at the gene level. The variance across the substitution rates of autosomal, X-linked and Y-linked genes when analyzed together could be used as an indicator of the general variability within and between chromosomal classes. The influence of male replication bias as the primary force shaping substitution rates could be inferred from changes in variance after the substitution rates have been adjusted following the time each chromosomal class spends in the male germline (X-linked genes would accumulate 33% less substitutions than autosomes and Y-linked genes would accumulate 50% more substitutions than autosomes). When the new adjusted variance across individual genes is smaller than the original variance, in theory, this would suggest that substitution rates at the gene level are consistent with male replication bias, despite the initial within-chromosomal variability. On the other hand, when the adjusted values fall outside of the autosomal distribution, the new variance would be larger than the initial variance, and this would mean that substitution rates of the analyzed genes are not consistent with male replication bias and other confounding factors are playing a predominant role influencing substitution rates in this particular set of genes.

In order to estimate the proportion of variation in substitution rates that is owing to replication effects, we resampled 1,000 times the autosomal and X-linked gene pools of all 1-to-1 orthologous genes between the species pairs. For each of the 1,000 rounds we randomly selected six autosomal and six X-linked genes. We used the same Y-linked genes for all the analyses because this was the maximum number of Y genes/transcripts that could be used for the all the species. We calculated the substitution rates for all genes individually. For each initial and adjusted variance we can estimate an index of the amount of change (see Methods). As the new variance leads toward zero, the resulting value of this formula leads to 1, which would mean that 100% of the variation at the gene level is explained by male replication bias.. We found that in monotremes, primates and rodents the new variance is frequently smaller than the initial variance (Figure 4) and male replication bias explains ~68-83% of the differences at the gene level (monotremes median = 72%, rodents median = 68%, primates median = 83%). These values are in agreement with the results obtained in the GLM, which show male replication bias is the main force shaping substitution rates in rodents, although its relative contribution varies across species. In addition, the strength of male mutation bias is consistent with the α values: monotremes and primates show the highest α estimates (value range 2.12-3.69 and 1.71-3.24, respectively) and also present the highest percentages of the by-gene variation explained by male replication bias. On the other hand, rodents show the lowest α (value range 1.53-1.92) and also the lowest percentages of the by-gene variation explained by male replication bias.

Discussion

The overabundance of replication errors in the male germline has been proposed as the main force shaping global chromosomal substitution rates in placental mammals (Li et al. 2002; Makova and Li 2002; Miyata et al. 1987; Venn et al. 2014; Wilson Sayres et al. 2011). Its importance relies on the notion

that mutations would be primarily produced in males, a scenario that has been dubbed “male-driven evolution” (Li et al. 2002; Malcom et al. 2003; McVean and Hurst 1997; Sandstedt and Tucker 2005; Smith and Hurst 1999; Wolfe and Sharp 1993). Therefore, the strength of male mutation bias could be directly linked to the genomic variability of a lineage or species. Estimates of male mutation bias have been calculated in placental mammals (Wilson Sayres et al. 2011), with great apes showing the highest rates (Venn et al. 2014), and our previous phylogenetic assessments of synonymous substitution rates of Y- and X-linked genes (and autosomal orthologs from outgroup species) in marsupials suggest substantial male mutation bias in this major mammalian lineage as well (Cortez et al. 2014). Here, we examined whether male mutation estimates particular to placentals, marsupials and birds are also observed in monotremes, which have many biological and genomic peculiarities such as egg-laying, venom production (only platypus (Wong et al. 2013)), micro-chromosomes (Warren et al. 2008), an unique sex system composed of 9 or 10 different chromosomes (Rens et al. 2004; Rens et al. 2007) and an atypical germline that lacks MSCI (Daish et al. 2015). Our results predict that the male germline goes through approximately 2.95 times more rounds of cell divisions (DNA replications) per generation than does the female germline in monotremes.

A general caveat in our study is the assumption that most of the positions in the analyzed sequences are neutrally evolving, such that the observed substitution rates can be taken as proxies for the underlying mutation rate. Violations of this assumption can potentially introduce biases in the estimates of male mutation bias. A recent study of 29 mammalian genomes revealed that <30% of intronic positions are under evolutionary constraint (Lindblad-Toh et al. 2011). Thus, we sought to limit biases in our estimates by curating the intronic alignments (see Methods) and by applying a double bootstrapping approach that sub-sampled both introns and positions, taking advantage of the fact that constrained intronic positions are not randomly distributed (they tend to be closer to splicing sites).

Our work highlights the importance of using the three chromosome classes to evaluate the degree of male mutation bias. We examined whether substitution rates variations between chromosomes are a consequence of male mutation bias or alternative forces. Although we could not directly test this hypothesis in monotremes due to lack of information, we performed the analysis in rodents using a multivariate model. Our results suggest that substitution rates are mostly influenced by male replication bias (or relative time spent in the male germline more precisely) and that the Y chromosome is under weaker purifying selection, as also previously noted (Wilson and Makova 2009). Transcription and replication-timing seem to be not significant when included in the same statistical framework together with other factors (substitution rates were previously correlated with late-replication in Y-linked genes (Pink and Hurst 2010)). All three of Miyata's estimates are clearly influenced by confounding forces, although male replication bias stands as the main driver of substitution rates.

Our work represents a comprehensive effort to analyze the contribution of male mutation bias and its strength in monotremes and in control species, rodents and primates. Furthermore, our analyses may serve to estimate the proportion of variation in substitution rates that is owing to replication effects. The strength of male mutation bias seems to be specific to the species, that is, male mutation bias has less intensity (probably fewer male germline divisions) in rodents than primates and monotremes (Figure 3). These results confirm previous observations that showed limited influence of male mutation bias in rodents (Pink et al. 2009; Pink and Hurst 2010), but a strong effect of this phenomenon in primates (Makova and Li 2002; Venn et al. 2014). In the future, our methods can be applied to non-model vertebrate species with poor genome assemblies.

Acknowledgements

We thank I. Xenarios and the Vital-IT computational facility for computational support; H. Kaessmann, M. Warnefors, M. Cardoso-Moreira and R. Marin for helpful comments throughout the study; and the Kaessmann group in general for discussions. This research was supported by grants from the European Research Council (Starting Grant: 242597, SexGenTransEvolution; Consolidator Grant: 615253) and the Swiss National Science Foundation (Grants: 130287 and 146474) to Prof. Henrik Kaessmann.

References

Altschul, S. F., et al. (1990), 'Basic local alignment search tool', *J Mol Biol,* 215 (3), 403-10.

Arndt, P. F., Burge, C. B., and Hwa, T. (2003), 'DNA sequence evolution with neighbor-dependent mutation', *J Comput Biol,* 10 (3-4), 313-22.

Axelsson, E., et al. (2004), 'Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and Turkey', *Mol Biol Evol,* 21 (8), 1538-47.

Bartosch-Harlid, A., et al. (2003), 'Life history and the male mutation bias', *Evolution,* 57 (10), 2398-406.

Birky, C. W., Jr. and Walsh, J. B. (1988), 'Effects of linkage on rates of molecular evolution', *Proc Natl Acad Sci U S A,* 85 (17), 6414-8.

Brudno, M., et al. (2003), 'LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA', *Genome Res,* 13 (4), 721-31.

Castresana, J. (2000), 'Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis', *Mol Biol Evol,* 17 (4), 540-52.

Chamary, J. V. and Hurst, L. D. (2004), 'Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage', *Mol Biol Evol,* 21 (6), 1014-23.

Chang, B. H., et al. (1994), 'Weak male-driven molecular evolution in rodents', *Proc Natl Acad Sci U S A,* 91 (2), 827-31.

Church, D. M., et al. (2009), 'Lineage-specific biology revealed by a finished genome assembly of the mouse', *PLoS Biol,* 7 (5), e1000112.

Conrad, D. F., et al. (2011), 'Variation in genome-wide mutation rates within and between human families', *Nat Genet,* 43 (7), 712-4.

Cortez, D., et al. (2014), 'Origins and functional evolution of Y chromosomes across mammals', *Nature,* 508 (7497), 488-93.

Daish, T. J., Casey, A. E., and Grutzner, F. (2015), 'Lack of sex chromosome specific meiotic silencing in platypus reveals origin of MSCI in therian mammals', *BMC Biol,* 13, 106.

Delgado, C. L., et al. (2009), 'Physical mapping of the elephant X chromosome: conservation of gene order over 105 million years', *Chromosome Res,* 17 (7), 917-26.

Drost, J. B. and Lee, W. R. (1995), 'Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human', *Environ Mol Mutagen,* 25 Suppl 26, 48-64.

Edgar, R. C. (2004), 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Res,* 32 (5), 1792-7.
Ellegren, H. (2007), 'Characteristics, causes and evolutionary consequences of male-biased mutation', *Proc Biol Sci,* 274 (1606), 1-10.
Flicek, P., et al. (2014), 'Ensembl 2014', *Nucleic Acids Res,* 42 (Database issue), D749-55.
Haldane, J. B. (1947), 'The mutation rate of the gene for haemophilia, and its segregation ratios in males and females', *Ann Eugen,* 13 (4), 262-71.
Hedges, S. B., Dudley, J., and Kumar, S. (2006), 'TimeTree: a public knowledge-base of divergence times among organisms', *Bioinformatics,* 22 (23), 2971-2.
Hurst, L. D. and Ellegren, H. (1998), 'Sex biases in the mutation rate', *Trends Genet,* 14 (11), 446-52.
Jabbari, K. and Bernardi, G. (2004), 'Cytosine methylation and CpG, TpG (CpA) and TpA frequencies', *Gene,* 333, 143-9.
Kanatsu-Shinohara, M. and Shinohara, T. (2013), 'Spermatogonial stem cell self-renewal and development', *Annu Rev Cell Dev Biol,* 29, 163-87.
Kim, D., Langmead, B., and Salzberg, S. L. (2015), 'HISAT: a fast spliced aligner with low memory requirements', *Nat Methods,* 12 (4), 357-60.
Kong, A., et al. (2012), 'Rate of de novo mutations and the importance of father's age to disease risk', *Nature,* 488 (7412), 471-5.
Lercher, M. J., Williams, E. J., and Hurst, L. D. (2001), 'Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias', *Mol Biol Evol,* 18 (11), 2032-9.
Lercher, M. J., Urrutia, A. O., and Hurst, L. D. (2002), 'Clustering of housekeeping genes provides a unified model of gene order in the human genome', *Nat Genet,* 31 (2), 180-3.
Li, W. H., Yi, S., and Makova, K. (2002), 'Male-driven evolution', *Curr Opin Genet Dev,* 12 (6), 650-6.
Li, W. H., et al. (1996), 'Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis', *Mol Phylogenet Evol,* 5 (1), 182-7.
Lindblad-Toh, K., et al. (2011), 'A high-resolution map of human evolutionary constraint using 29 mammals', *Nature,* 478 (7370), 476-82.
Luo, R., et al. (2012), 'SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler', *Gigascience,* 1 (1), 18.
Makova, K. D. and Li, W. H. (2002), 'Strong male-driven evolution of DNA sequences in humans and apes', *Nature,* 416 (6881), 624-6.
Malcom, C. M., Wyckoff, G. J., and Lahn, B. T. (2003), 'Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity', *Mol Biol Evol,* 20 (10), 1633-41.
Matassi, G., Sharp, P. M., and Gautier, C. (1999), 'Chromosomal location effects on gene sequence evolution in mammals', *Curr Biol,* 9 (15), 786-91.
McVean, G. T. and Hurst, L. D. (1997), 'Evidence for a selectively favourable reduction in the mutation rate of the X chromosome', *Nature,* 386 (6623), 388-92.
Miyata, T., et al. (1987), 'Male-driven molecular evolution: a model and nucleotide sequence analysis', *Cold Spring Harb Symp Quant Biol,* 52, 863-7.
Necsulea, A., et al. (2014), 'The evolution of lncRNA repertoires and expression patterns in tetrapods', *Nature,* 505 (7485), 635-40.
Pink, C. J. and Hurst, L. D. (2010), 'Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents', *Mol Biol Evol,* 27 (5), 1077-86.
Pink, C. J., et al. (2009), 'Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates', *Genome Biol Evol,* 1, 13-22.
Presgraves, D. C. and Yi, S. V. (2009), 'Doubts about complex speciation between humans and chimpanzees', *Trends Ecol Evol,* 24 (10), 533-40.

Rens, W., et al. (2004), 'Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution', *Proc Natl Acad Sci U S A,* 101 (46), 16257-61.
Rens, W., et al. (2007), 'The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z', *Genome Biol,* 8 (11), R243.
Rowe, T., et al. (2008), 'The oldest platypus and its bearing on divergence timing of the platypus and echidna clades', *Proc Natl Acad Sci U S A,* 105 (4), 1238-42.
Sandstedt, S. A. and Tucker, P. K. (2005), 'Male-driven evolution in closely related species of the mouse genus Mus', *J Mol Evol,* 61 (1), 138-44.
Saxonov, S., Berg, P., and Brutlag, D. L. (2006), 'A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters', *Proc Natl Acad Sci U S A,* 103 (5), 1412-7.
Shimmin, L. C., Chang, B. H., and Li, W. H. (1993a), 'Male-driven evolution of DNA sequences', *Nature,* 362 (6422), 745-7.
Shimmin, L. C., et al. (1993b), 'Potential problems in estimating the male-to-female mutation rate ratio from DNA sequence data', *J Mol Evol,* 37 (2), 160-6.
Smagulova, F., et al. (2011), 'Genome-wide analysis reveals novel molecular features of mouse recombination hotspots', *Nature,* 472 (7343), 375-8.
Smith, N. G. and Hurst, L. D. (1999), 'The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate?', *Genetics,* 152 (2), 661-73.
Soh, Y. Q., et al. (2014), 'Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes', *Cell,* 159 (4), 800-13.
Soumillon, M., et al. (2013), 'Cellular source and mechanisms of high transcriptome complexity in the mammalian testis', *Cell Rep,* 3 (6), 2179-90.
Talavera, G. and Castresana, J. (2007), 'Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments', *Syst Biol,* 56 (4), 564-77.
Tamura, K. and Nei, M. (1993), 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees', *Mol Biol Evol,* 10 (3), 512-26.
Tamura, K. and Kumar, S. (2002), 'Evolutionary distance estimation under heterogeneous substitution pattern among lineages', *Mol Biol Evol,* 19 (10), 1727-36.
Taylor, J., et al. (2006), 'Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison', *Mol Biol Evol,* 23 (3), 565-73.
Tilford, C. A., et al. (2001), 'A physical map of the human Y chromosome', *Nature,* 409 (6822), 943-5.
Trapnell, C., et al. (2013), 'Differential analysis of gene regulation at transcript resolution with RNA-seq', *Nat Biotechnol,* 31 (1), 46-53.
Venn, O., et al. (2014), 'Nonhuman genetics. Strong male bias drives germline mutation in chimpanzees', *Science,* 344 (6189), 1272-5.
Veyrunes, F., et al. (2008), 'Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes', *Genome Res,* 18 (6), 965-73.
Warren, W. C., et al. (2008), 'Genome analysis of the platypus reveals unique signatures of evolution', *Nature,* 453 (7192), 175-83.
Wilson, M. A. and Makova, K. D. (2009), 'Evolution and survival on eutherian sex chromosomes', *PLoS Genet,* 5 (7), e1000568.
Wilson Sayres, M. A., et al. (2011), 'Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes', *Evolution,* 65 (10), 2800-15.
Wolfe, K. H. and Sharp, P. M. (1993), 'Mammalian gene evolution: nucleotide sequence divergence between mouse and rat', *J Mol Evol,* 37 (4), 441-56.

Wong, E. S., et al. (2013), 'Echidna venom gland transcriptome provides insights into the evolution of monotreme venom', *PLoS One,* 8 (11), e79092.
Yang, Z. (1997), 'PAML: a program package for phylogenetic analysis by maximum likelihood', *Comput Appl Biosci,* 13 (5), 555-6.

Supplementary files

Supplementary Figure 1: Lengths of intronic sequences obtained with alternative parameters to run SOAP *de novo*.

Supplementary Figure 2: Y, X and autosomal substitution rates and GC content in monotremes.

Supplementary Figure 3: Y, X and autosomal substitution rates with fixed GC content in monotremes.

Supplementary Figure 4: QQ-plots from the GLM.

Supplementary Table 1: Overview of genomic sequences used in this study. **Supplementary Table 2:** Overview of the genes used in this study. **Supplementary Table 3:** Substitution rates for all chromosomal classes with fixed GC content. **Supplementary Table 4:** Data used in the GLM. **Supplementary Table 5:** Substitution rates for all chromosomal classes after each of the 1,000 bootstrap rounds. **Supplementary Table 6:** Calculation of the 95% confident intervals of α. Supporting data are available in the online supplementary files and also upon request. **Supplementary Table 7:** Human-marmoset and mouse-rat homologous chromosomes.

Figure legends

Fig. 1: Substitution rates across all chromosomes in monotremes and control species. a-b) Substitution rates from the human-marmoset comparisons, sorted according to the human homologous chromosomes (a) or the marmoset homologous chromosomes (b). c-d) Substitution rates from the

mouse-rat comparisons, sorted according to the mouse homologous chromosomes (c) or the rat homologous chromosomes (d). e) Substitution rates from the platypus-echidna comparisons, sorted according to the platypus chromosomes. Data for the echidna could not be plotted because there is no reference genome for this species. Chromosomes are sorted according to their median substitution rate in descending order, from left to right. *P*-values were obtained by applying the Benjamini–Hochberg-corrected Welch Two Sample t-test. See Supplementary Table 7 for correspondence between human-marmoset and mouse-rat homologous chromosomes.

Fig. 2: Distribution of chromosomal substitution rates in monotremes and control species. Distributions of chromosomal substitution rates for Y, X, and autosomes in monotremes (a), primates (b) and rodents (c) obtained from 1,000 bootstrap rounds. Significance of the Welch Two Sample t-test Benjamini–Hochberg-corrected p-values are as follows: ***$P < 0.001$, *$P < 0.05$. Error bars, maximum and minimum values, excluding outliers. The red line at value 0.2 serves as visual aid.

Fig. 3: Selection, transcription and replication timing across chromosomal classes. Distributions of (a) d_N/d_S ratios, (b) expression levels and (c) replication timing for Y, X, and autosomes in rodents. Significance of the Welch Two Sample t-test Benjamini–Hochberg-corrected p-values are as follows: ***$P < 0.001$, *$P < 0.05$. Error bars, maximum and minimum values, excluding outliers.

Fig. 4: Male mutation bias at the gene level. Sorted values from 1,000 resampling rounds. Each value represents the difference between the initial and the adjusted variances. The derived percentage represents the proportion of variation at the gene level explained by male replication bias. Values for monotremes are in blue. Values for rodents are in pink. Values for primates are in yellow. The boxplot summarizes the data contained in three curves; the species color-code is the same. The horizontal red

line at value 70% serves as visual aid to show the number of resamplings in monotremes, primates and rodents above this proportion.

Table 1: Median values of the chromosomal substitution rates and variation coefficients in the three chromosomal classes.

Species	Chromosome	Median	Standard deviation	95% CI (lower - upper)
Monotremes (Platypus-Echidna)	Autosomes	0.1213	±0.03683	0.12052 - 0.12219
	X	0.105	±0.02164	0.10423 - 0.10568
	Y	0.1888	±0.0211	0.18715 - 0.19068
Primates (Human-Marmoset)	Autosomes	0.154	±0.0297	0.15245 - 0.15519
	X	0.1387	±0.0181	0.1373 - 0.14046
	Y	0.2317	±0.0189	0.22968- 0.23369
Rodents (Mouse-Rat)	Autosomes	0.1926	±0.011	0.19158 - 0.1934
	X	0.1777	±0.0136	0.17657 - 0.17901
	Y	0.2511	±0.0242	0.24973 - 0.25328

Table 2: Empirical, adjusted-fixed and final adjusted values for the chromosomal substitution rates and α estimates.

Species	Y.*emp*	A.*emp*	X.*emp*	α.X/Y	α.Y/A	α.X/A
Monotremes	0.1888	0.1213	0.105	2.99 (2.9-3.07)	3.51 (3.29-3.69)	2.35 (2.12-2.55)
Primates	0.2317	0.154	0.1387	2.52 (2.44-2.59)	3.03 (2.89-3.24)	1.84 (1.71-1.97)
Rodents	0.2511	0.1926	0.1777	1.78 (1.75-1.81)	1.87 (1.82-1.92)	1.6 (1.53-1.68)

emp is the empirical value directly obtained from the analysis of intronic sequences; ***α.X/Y, α.Y/A, α.X/A*** are the three α values with their respective 95% confident intervals.

Figure 1

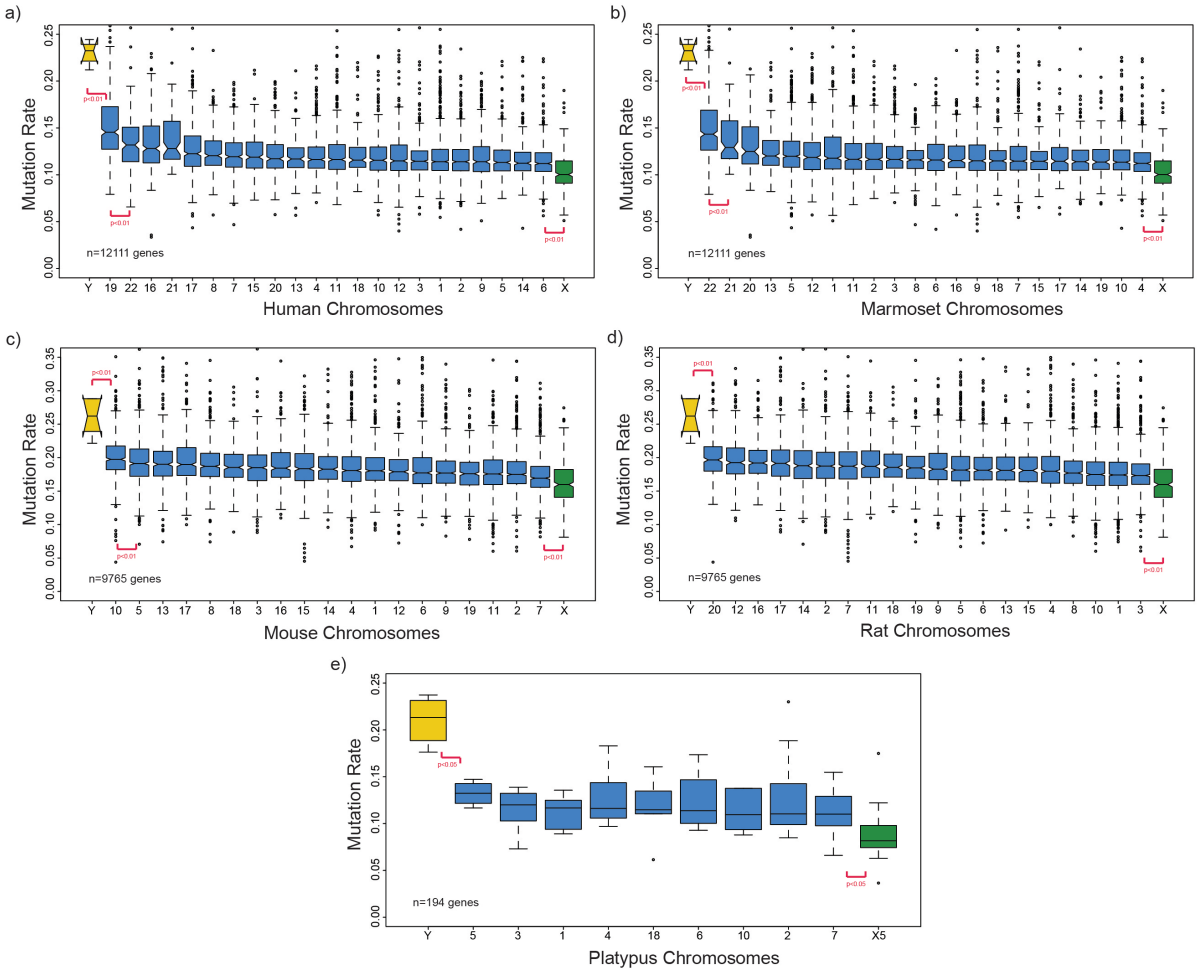

Fig. 1: Substitution rates across all chromosomes in monotremes and control species. a-b) Substitution rates from the human-marmoset comparisons, sorted according to the human homologous chromosomes (a) or the marmoset homologous chromosomes (b). c-d) Substitution rates from the mouse-rat comparisons, sorted according to the mouse homologous chromosomes (c) or the rat homologous chromosomes (d). e) Substitution rates from the platypus-echidna comparisons, sorted according to the platypus chromosomes. Data for the echidna could not be plotted because there is no reference genome for this species. Chromosomes are sorted according to their median substitution rate in descending order, from left to right. *P*-values were obtained by applying the Benjamini–Hochberg-corrected Welch Two Sample t-test. See Supplementary Table 7 for correspondence between human-marmoset and mouse-rat homologous chromosomes.

Figure 2

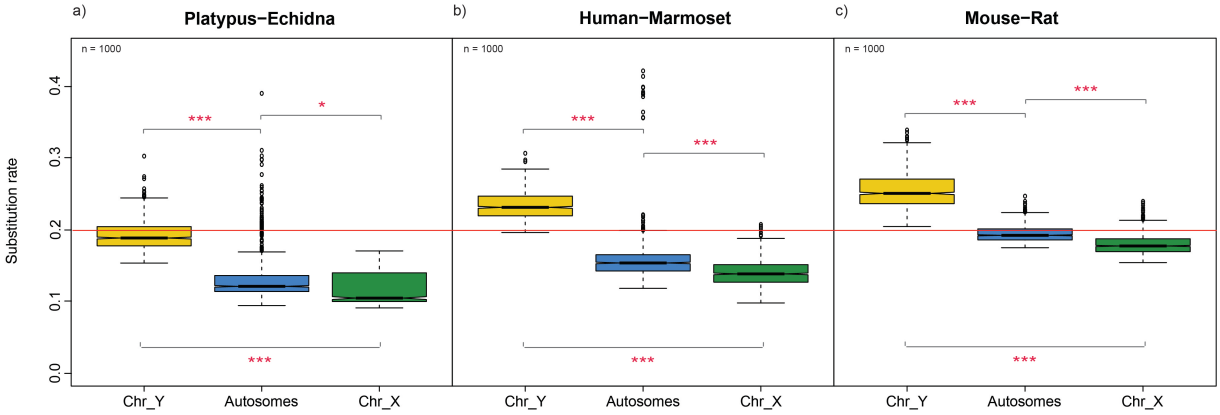

Fig. 2: Distribution of chromosomal substitution rates in monotremes and control species. Distributions of chromosomal substitution rates for Y, X, and autosomes in monotremes (a), primates (b) and rodents (c) obtained from 1,000 bootstrap rounds. Significance of the Welch Two Sample t-test Benjamini–Hochberg-corrected p-values are as follows: ***$P < 0.001$, *$P < 0.05$. Error bars, maximum and minimum values, excluding outliers. The red line at value 0.2 serves as visual aid.

Figure 3

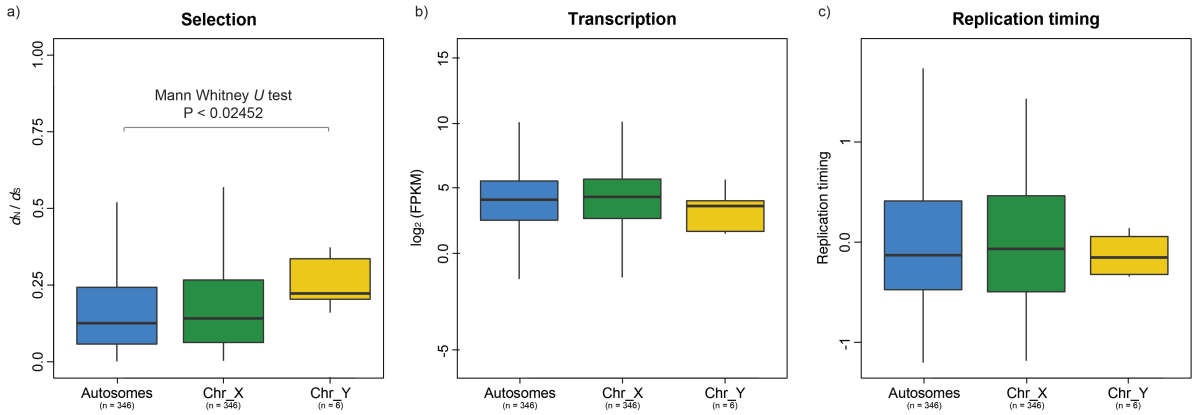

Fig. 3: Selection, transcription and replication timing across chromosomal classes. Distributions of (a) d_N/d_S ratios, (b) expression levels and (c) replication timing for Y, X, and autosomes in rodents. Significance of the Welch Two Sample t-test Benjamini–Hochberg-corrected p-values are as follows: ***$P < 0.001$, *$P < 0.05$. Error bars, maximum and minimum values, excluding outliers.

Figure 4

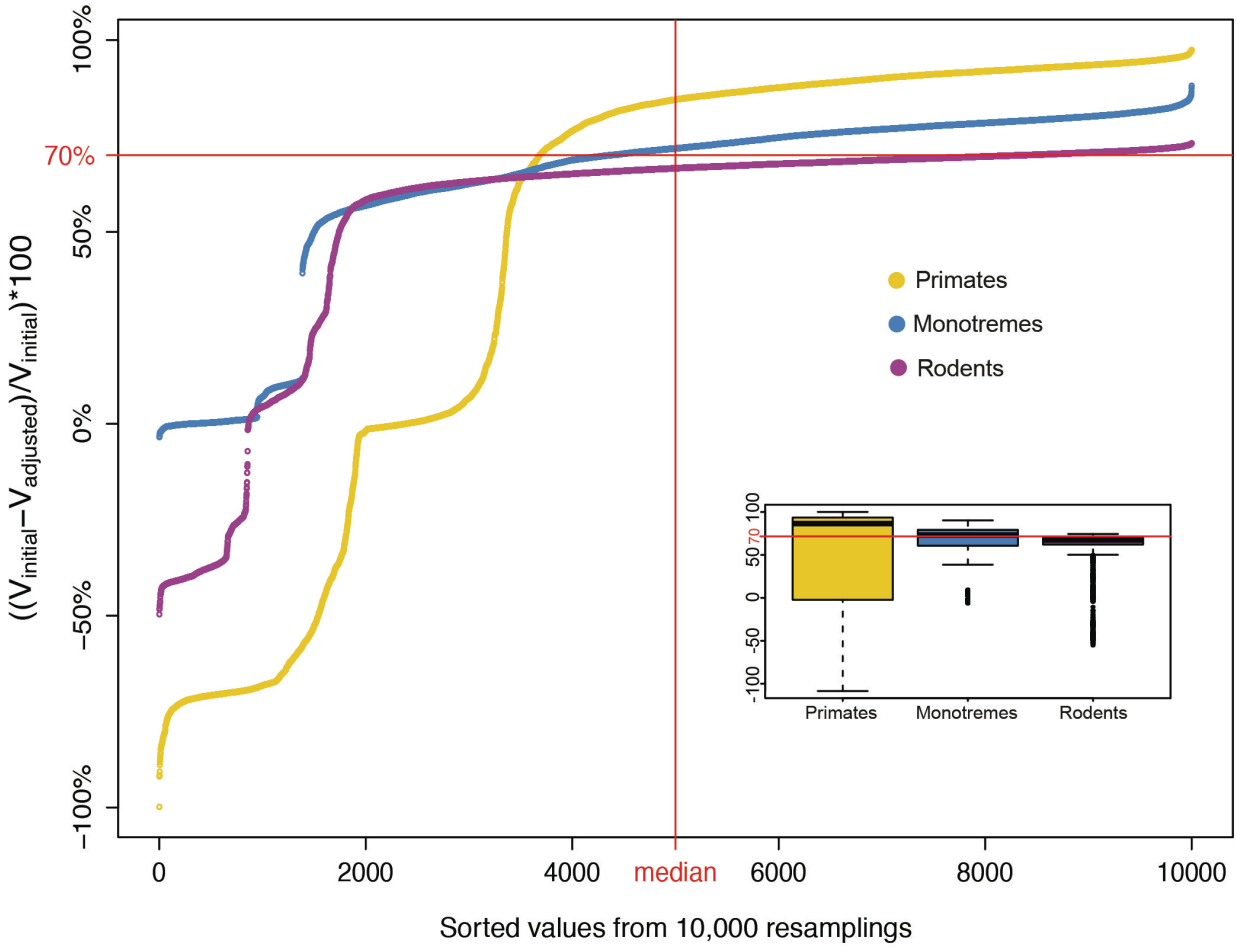

Fig. 4: Male mutation bias at the gene level. Sorted values from 1,000 resampling rounds. Each value represents the difference between the initial and the adjusted variances. The derived percentage represents the proportion of variation at the gene level explained by male replication bias. Values for monotremes are in blue. Values for rodents are in pink. Values for primates are in yellow. The boxplot summarizes the data contained in three curves; the species color-code is the same. The horizontal red line at value 70% serves as visual aid to show the number of resamplings in monotremes, primates and rodents above this proportion.