ARTICLE

doi:10.1038/nature13151

Origins and functional evolution of Y chromosomes across mammals

Diego Cortez[1,2], Ray Marin[1,2], Deborah Toledo-Flores[3], Laure Froidevaux[1], Angélica Liechti[1], Paul D. Waters[4], Frank Grützner[3] & Henrik Kaessmann[1,2]

Y chromosomes underlie sex determination in mammals, but their repeat-rich nature has hampered sequencing and associated evolutionary studies. Here we trace Y evolution across 15 representative mammals on the basis of high-throughput genome and transcriptome sequencing. We uncover three independent sex chromosome originations in mammals and birds (the outgroup). The original placental and marsupial (therian) Y, containing the sex-determining gene *SRY*, emerged in the therian ancestor approximately 180 million years ago, in parallel with the first of five monotreme Y chromosomes, carrying the probable sex-determining gene *AMH*. The avian W chromosome arose approximately 140 million years ago in the bird ancestor. The small Y/W gene repertoires, enriched in regulatory functions, were rapidly defined following stratification (recombination arrest) and erosion events and have remained considerably stable. Despite expression decreases in therians, Y/W genes show notable conservation of proto-sex chromosome expression patterns, although various Y genes evolved testis-specificities through differential regulatory decay. Thus, although some genes evolved novel functions through spatial/temporal expression shifts, most Y genes probably endured, at least initially, because of dosage constraints.

In most mammals, Y chromosomes are required to override the program underlying development of the default sex, females[1]. Extant mammals possess an XY (male heterogametic) sex chromosome system, with rare exceptions that experienced secondary XY loss[2], but sex chromosomes evolved from different autosomes in therians (placentals and marsupials) and egg-laying monotremes (Fig. 1). Therians share the same XY system, whereas monotremes have multiple X and Y chromosomes that are partially homologous to bird sex chromosomes[3–5], where females are heterogametic (ZW system). Sex chromosome differentiation occurred through recombination arrests along the Y, leading to reduced selection and associated gene decay and repeat accumulation[6,7]. Consequences of Y deterioration for X chromosome evolution were recently investigated[8], but the fact that this chromosome is refractory to assembly owing to its repeat-rich nature[7] have hindered evolutionary studies of the Y itself. Nevertheless, dedicated efforts determined complete Y sequences in three primates[9–11] (human, chimpanzee, macaque) and large portions of two carnivore (dog and cat) Y chromosomes[12]. Together with smaller scale work, these studies provided initial clues to Y evolution, such as the stabilization of Y gene content through ampliconic sequence accumulation[10,11,13]. However, our understanding of mammalian Y chromosome evolution remains limited owing to the restricted amount and phylogenetic representation of available Y chromosome data.

Mammalian Y gene repertoires

To explore Y evolution, we developed a subtraction approach that directly targets and assembles exons of the male-specific (non-pseudoautosomal) Y chromosome (MSY) on the basis of high-throughput sequencing of transcriptomes and genomes from both sexes (Extended Data Fig. 1 and Methods). In brief, male transcripts were assembled from male-specific RNA sequencing (RNA-seq) reads not mapping onto female reference genomes. Y identity was confirmed using the whole-genome sequence data; that is, true Y transcripts are supported by genomic reads unique to males. Genes with no or low expression were detected by screening male-specific genomic data with Y orthologues from other species. The genomic data also served to support the absence of ancestral Y genes (that is, their evolutionary loss). We validated our approach using large-scale PCR/Sanger sequencing-based screening of male/female genomic DNA and published Y chromosomes, thus confirming that complete coding sequences of all Y genes (with the potential exception of those not expressed in the sampled tissues and lacking a known Y orthologue) could be deduced for a given species.

We applied our procedure to sequencing data that we collected for ten mammals (Supplementary Tables 1–4 and Methods). Together with available Y sequences[9–12], we could thus investigate Y evolution in 15 species covering all major lineages of the class Mammalia[14] (placentals or eutherians, marsupials, monotremes), all but one (Xenarthra) of the eutherian superorders[14], the two marsupial superorders (American and Australian marsupials), both extant monotreme families (platypus, echidna), and all major groups of 'higher primates' (that is, simians: the great apes, Old World monkeys (OWMs) and New World monkeys (NWMs)). For comparison, we produced similar data for the bird (chicken) W chromosome.

We identified a total of 134 different Y protein-coding genes in the ten new species, thus approximately doubling the number of previously known Y genes (Fig. 1 and Extended Data Fig. 3; note that the letter Y is added at the end of gene symbols for all genes for which Y-linkage is detected, following common practice), and 214 distinct pseudogenes and noncoding RNAs (Supplementary Tables 5–17 and Supplementary Data 1). Our read coverage analysis (Methods) predicts 0–6 multi-copy protein-coding genes per species, some of which are shared across species and thus presumably belong to conserved ampliconic Y regions[11] (Fig. 1 and Supplementary Tables 5–17). Most of the 155 distinct noncoding RNAs seem to stem from ampliconic regions, with two or more identical copies per locus (Supplementary Tables 5–17). We recovered all of the recently identified chicken W protein-coding genes[15,16] and added 11 distinct noncoding RNAs (Fig. 1 and Supplementary Table 18).

[1]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [3]The Robinson Research Institute, School of Molecular and Biomedical Science, University of Adelaide, Adelaide, South Australia 5005, Australia. [4]School of Biotechnology and Biomolecular Sciences, UNSW Australia, Sydney, New South Wales 2052, Australia.

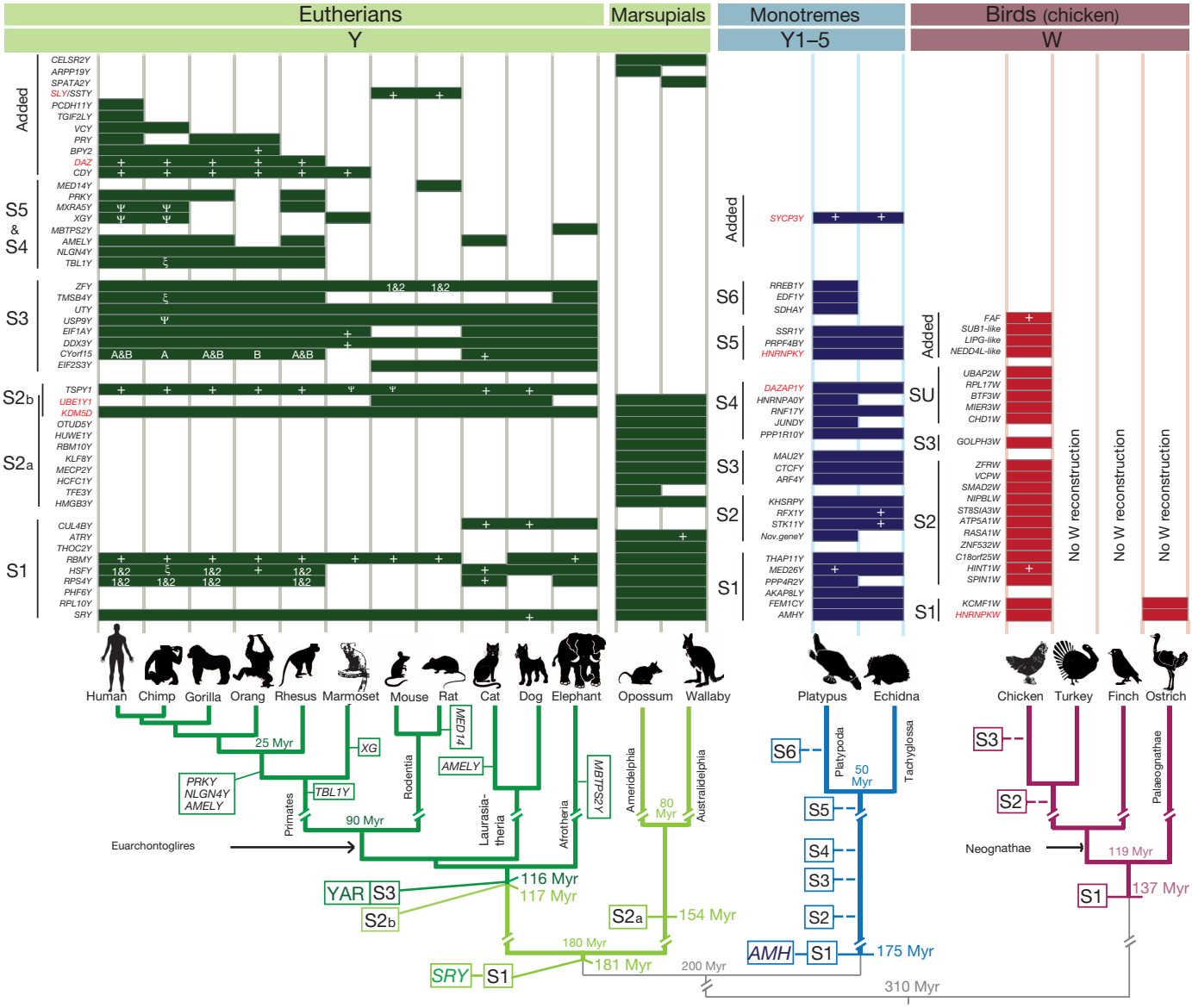


Figure 1 | Y (W) protein-coding gene repertoires and their origins. Top, Y/W genes and strata (genes are ordered within strata based on d_S). Genes added to differentiated sex chromosomes ('Added') and independently recruited genes (red text) are indicated. For echidna and ostrich, only genes shared with sister species (platypus and chicken, respectively) were identified on the basis of genomic sequencing data. For bird strata age estimations, turkey and zebra finch Z gametologues were included. Bottom, phylogenetic tree of the 19 species. Lineages, divergence times and origins of strata and genes are indicated. For therian strata, monotreme S1 and bird S1, refined age estimates on branches are based on d_S trees (Extended Data Fig. 5a–f). More approximate placement on branches for platypus S2–S5 and chicken S2–S3 (genes for which strata could not be assigned with confidence are listed as undefined, SU) are based on pairwise (median) d_S values of stratum genes (dashed lines). Note that although actual Y chromosome assemblies underlie the previously published[12] dog/cat Y repertoires shown here, the repertoires may be incomplete, because these assemblies (in particular for cat) are incomplete[12]. +, multi-copy genes; ψ, pseudogenes; ξ, intact genes in chimpanzee formerly annotated as pseudogenes; A&B 1&2, known duplicates; YAR, Y-added region.

Origins of therian Y genes

We next assessed the time at which Y genes began to evolve independently of their X counterpart (gametologue), or when they were recruited (through transposition/translocation) to already differentiated Y regions from autosomes or the X. We thus reconstructed phylogenetic trees for all Y genes, X gametologues, and autosomal orthologues from outgroup species with different sex chromosomes (Extended Data Fig. 4, Supplementary Data 1 and Methods). In these trees, the location of nodes reflecting the divergence of Y and X clades, relative to lineage divergence nodes, mark the phylogenetic position at which a given Y gene emerged. We assessed divergence rates at synonymous sites (d_S) before/after differentiation of Y/X gametologues for further refinement of ages (Extended Data Fig. 5a–f).

Y chromosomes consist of distinct regions ('strata') that stopped recombining at specific time points[10,17]. Our phylogenetic dating of the original eutherian stratum 1 (S1), which, like S2, stems from the same ancestral proto-sex chromosome portion (the 'Y/X-conserved' region, YCR/XCR; ref. 18), shows that it originated in the last common therian ancestor (Fig. 1 and Extended Data Fig. 4a; see the following external link for individual trees: ftp://ftp.vital-it.ch/papers/kaessmann/Nature-Cortez/Cortez_etal_Nature_YX_gene_trees_alignments.zip), confirming that sex chromosomes of humans and other therians emerged after the therian–monotreme split[3,4] and were never common to all mammals, as implied by some previous work[10,19]. Notably, therian sex chromosome origination occurred just before the eutherian–marsupial split ~180 million years (Myr) ago (Fig. 1 and Extended Data Fig. 5a), consistent with our previous hypothesis[3,8]. S1 contains a common set of four genes, including the sex-determining gene *SRY*[1], which has been maintained since sex chromosome inception in most therians, although some S1 genes apparently became dispensable in some species (Fig. 1). Interestingly,

marsupials retained twice the number of ancestral S1 genes. Finally, it is noteworthy that d_S for Y gametologues substantially increased relative to that of X counterparts in therians following S1 differentiation (Benjamini–Hochberg-corrected $P < 0.05$, Mann–Whitney *U*-test; Extended Data Figs 5g and 7a–d), suggesting that male mutation bias (that is, a higher mutation rate in males than in females owing to a higher number of germline cell divisions), previously reported for eutherians[20], has shaped genomes of therians since their emergence.

S2 arose ~117 Myr ago in the eutherian ancestor, after the split from the marsupial lineage (Extended Data Figs 4b and 5b), consistent with some previous studies[21] but contrary to others[19]. In marsupials an independent S2, containing twice the number of eutherian S2 genes, arose in the marsupial ancestor ~37 Myr before the eutherian S2 (Extended Data Figs 4e and 5c). Thus, the therian pseudoautosomal region (PAR) shrank at different rates in eutherians and marsupials, involving the convergent differentiation of two genes (*KDM5D*, *UBE1Y1* (also known as *UBA1Y*)), as previously proposed[21] (Fig. 1).

Our analysis of S3, a large autosomal addition to the already established sex chromosomes[18] (the 'Y/X-added region'), shows that it contained seven genes and was already defined in the common eutherian ancestor (Fig. 1 and Extended Data Fig. 4c), as could previously only be surmised[10,17]. S3 only differentiated just before the placental mammal radiation ~116 Myr ago, at about the same time or perhaps even concomitantly with S2 (for example, as part of the same inversion event; Extended Data Fig. 5d), implying that the eutherian MSY consisted merely of S1 for ~60 Myr following the eutherian–marsupial split. As for S1/S2, S3 gene content has been remarkably stable during evolution (Fig. 1). However, curiously, similarly to S1 (*HSFY*, *RPS4Y*), rodents and marmosets independently lost the same two S3 genes (*TMSB4Y*, *Cyorf15Y*), whereas *TMSB4Y* was independently also lost in Laurasiatheria[22] (Fig. 1). It is also noteworthy that the elephant retained six more ancestral S1–S3 genes (total 15) than the marmoset (total nine), the eutherian with the smallest detected S1–S3 gene repertoire.

Further analyses show that all genes of S4 and S5, as previously defined[10], originated in the common catarrhine (OWM/great ape) ancestor ~25–40 Myr ago, except for one gene (*TBL1Y*), which presumably emerged in the common simian ancestor >40 Myr ago (Fig. 1). The marmoset Y acquired a unique gene, *XG* (Fig. 1), which spans the human PAR boundary[23] and remains in the elephant PAR region[24]. Furthermore, *AMELY*, also residing near the PAR boundary, probably emerged independently in primates and Laurasiatheria[21]. *MBTPS2Y* arose as part of an independent S4 in the afrotherian lineage, and *MED14Y* represents a recent acquisition in rats (Fig. 1). Primate Y chromosomes recruited additional genes through various (retro)transpositions/translocations from autosomes during recent evolution[10,11] (Fig. 1). These results illustrate the dynamic recent, partly convergent, evolution of the eutherian Y.

Although extant gene numbers are small, the above results illustrate that many Y genes endured over long evolutionary time periods. Modelling of Y gene loss dynamics revealed that S1–S3 gene decay proceeded rapidly upon differentiation, at similar rates as previously estimated[10], and then markedly levelled off when small but apparently essential gene repertoires were defined (Extended Data Fig. 8c).

Monotreme Y biology and evolution

Gene content and evolution of monotreme Y chromosomes (five in platypus, four in echidna owing to a Y_5–Y_3 fusion[25]) previously remained largely unknown, and the X chromosomes of platypus, the only monotreme with an assembled genome[26], are only partly reconstructed. To assess monotreme sex chromosome evolution, we dated the 25 identified monotreme Y protein-coding genes (Fig. 1) and X gametologues. Because no X gametologue is assigned to any currently assembled X, we confirmed X identity for 23 out of 25 gametologues and assessed probable locations on the basis of male/female genomic read coverage and synteny mapping (Extended Data Fig. 6, Supplementary Tables 19, 20 and Methods), revealing that monotreme gametologues stem from various ancestral synteny blocks (Supplementary Tables 19, 20) that assembled during evolution through genomic rearrangements[4].

We detected six potential Y chromosome strata, five of which emerged in the platypus–echidna ancestor (Fig. 1 and Extended Data Fig. 5h). Thus, nearly all (22 out of 25) platypus Y genes differentiated at least ~50 Myr ago in the common monotreme ancestor (Supplementary Data 1, external data link). Importantly, monotreme S1 originated early during monotreme evolution ~175 Myr ago (Extended Data Figs 4f and 5e), which, together with our dating of therian sex chromosome origination, implies two independent yet essentially concomitant sex chromosome origination events in mammals, rules out the possibility of a turnover of one mammalian sex chromosome system into the other, and raises the question of the nature of sex determination in the common mammalian ancestor more than ~200 Myr ago.

Notably, one of the S1 genes is *AMH*, encoding the anti-Müllerian hormone, a key component of sex determination cascades across vertebrates[1,27], which in eutherians blocks the development of female reproductive organs upon activation by SOX9, which in turn is activated by SRY[1]. Our observation that *AMH* is Y-linked in monotremes and part of the oldest stratum, S1 (Fig. 1 and Extended Data Fig. 5h), together with findings that *AMH* may precede *SOX9* expression and act as the primary sex-determining trigger in some vertebrates[27–29], render this gene a prime candidate for being a principal monotreme sex-determining gene. We predicted *AMHX* to be located on chromosome X_1 (Supplementary Tables 19, 20) and *AMHY* to be part of Y_5, on the basis of our synteny approach and the recent mapping[30] of the S1 gene *MED26Y* (Fig. 1), which similarly to *AMHY* stems from an ancestral linkage group corresponding to chicken chromosome 28. Notably, we confirmed that *AMHY* is located on Y_5 by physical mapping (Extended Data Fig. 7e, f and Methods). Our results favour a scenario in which a fusion of an ancestral chromosome segment containing *AMH*, presumably to the original proto-Y_5-X_5, formed the initial monotreme proto-sex chromosomes. Their differentiation was followed by various translocations/fusions with autosomes[25] and subsequent differentiation events, which ultimately led to the current Y genes, of which only two genes (*FEM1CY* and *HNRNPKY*) were derived from the initial Z-homologous proto-Y_5 portion. Curiously, in parallel to eutherians, monotreme Y chromosomes also acquired *DAZ* and *SLY* homologous genes (Fig. 1).

Finally, contrary to our therian observations, we could not detect significantly different d_S values between monotreme Y and X gametologues (Extended Data Fig. 5g), which indicates that male mutation bias[20] is lacking or limited in monotremes.

W origin

Next, we performed the first phylogenetic dating of chicken W strata (Fig. 1 and Supplementary Data 1, external data link), taking advantage of Z orthologues from recent bird genome sequences[31,32] and male/female RNA-seq data from ostriches[33] (allowing the extraction of ostrich W genes), representing the most basal avian lineage, ratites. Contrary to previous work[34], this analysis reveals an initial stratum with two genes (*HNRNPK*, *KCMF1*) that emerged in the common bird ancestor ~140 Myr ago (Extended Data Figs 4g and 5f), whereas the remaining chicken W strata seem to have originated later during avian evolution (Fig. 1), although a more precise phylogenetic dating of these strata will require W sequences from other birds. *KCMF1* and *HNRNPK*, which was independently retained on both the W and monotreme Y (reflecting their shared ancestry; Fig. 1), are widely expressed housekeeping genes[35,36] that do not represent obvious sex-determination candidates but rather dosage-sensitive genes that were preserved to maintain proto-Z/W gene dose. Our results are consistent with a Z dosage-based mechanism of sex determination in birds[15,37].

Functional evolution of Y (W) genes

To understand why only small specific subsets of 9–25 Y/W protein-coding genes have been preserved from the original proto-sex chromosome repertoires, we assessed whether they possess characteristic

functions using simulations, which started with the ancestral sets of proto-sex-linked genes and then randomly removed genes until current Y/W gene numbers were reached (Methods). We then compared the functions of the simulated gene sets to those observed using Gene Ontology (GO) annotations[38]. Notably, this analysis revealed that highly non-random gene sets with similar functions were maintained across the different Y/W chromosomes (corrected $P < 0.01$, one-tailed alpha test). Current Y/W genes are enriched for genes involved in transcription/transcription regulation (GO Biological Process) and specific DNA binding/transcription factor activity (GO Molecular Function) (Supplementary Tables 21, 22), which is consistent with recent observations in fruitflies[39] and indicates that current Y genes were, at least initially, preserved to maintain ancestral gene dosage[39], given that regulatory genes and genes with binding functions are often haploinsufficient[40].

Expression levels of autosomal orthologues of Y/X genes in species with different sex chromosome systems, unaffected by sex-related selection, may serve to gauge proto-sex chromosome expression levels[8]. Expression levels inferred for proto-sex chromosomal precursors of Y/X gametologues using this procedure are higher than those of other proto-sex chromosome genes in all amniote lineages (Extended Data Fig. 8a), conforming to the above notion that current Y genes derive from highly expressed genes with universal (dosage-sensitive) functions. Consistently, X/Z gametologues, which are expected to generally preserve ancestral functions[8], show higher expression levels than other X/Z-linked genes (Extended Data Fig. 8a).

Expression levels of therian Y genes decreased since sex chromosome differentiation (corrected $P < 0.05$, Mann–Whitney *U*-test), with current (median) expression levels that are between 3.1 (opossum) to 15 (rodents) times lower than inferred ancestral expression levels of single proto-sex chromosome alleles (Fig. 2), reflecting partial regulatory decay of Y genes and/or evolution towards new functions. Further analyses show that many therian Y genes maintained ancestral patterns of usually ubiquitous expression, with interesting exceptions such as *SRY* and *AMHY*, which appear to have been sex-biased already on the proto-sex chromosomes (Fig. 3). Nevertheless, a larger number of genes evolved new expression patterns on the Y than on the X or autosomes (corrected $P < 0.05$, Fisher's exact test, Supplementary Table 23). All of these Y genes evolved testis-specific expression (Fig. 3), consistent with the male-limited transmission of the Y[11]. Notably, therian Y genes evolved testis specificity by experiencing substantially stronger expression reduction in somatic tissues than in testis ($P < 0.05$, Mann–Whitney *U*-test; Extended Data Fig. 8b). Thus, testis specificity evolved through differential regulatory decay rather than upregulation in testis, as observed in fruitflies[39]. Contrary to therians, platypus Y and chicken W genes show no overall expression reduction relative to their proto-sex precursors (Fig. 2), and most of them also preserved ancestral spatial expression patterns (Fig. 3). Thus, Y/W expression preservation is particularly pronounced in monotremes and birds.

To explore cellular functions of therian testis-specific Y genes, we used the mouse as a model, assessing expression patterns of mouse Y genes across all major testicular cell types using dedicated data[41]. Contrary to ubiquitously expressed Y genes, testis-specific Y transcripts are restricted to three germ-cell types: mitotic germ cells (spermatogonia), meiotic cells (spermatocytes) and, especially, post-meiotic cells (round spermatids) (Extended Data Fig. 9a). Although the high expression levels of these genes in spermatids may partly reflect 'promiscuous' transcription, facilitated by overall open chromatin[41], it may also indicate post-meiotic functions, consistent with the finding that the spermatid-specific *SLY* gene (Extended Data Fig. 9a) is a sex chromosome regulator in spermatids[42]. Detected noncoding and pseudogenic Y transcripts are often testis specific across species and sometimes have many genomic copies (Supplementary Tables 5–17). Notably, nearly all noncoding mouse Y transcripts are specifically expressed in spermatids/spermatocytes (Extended Data Fig. 9b), probably mainly due to the permissive chromatin environment[41]. The lower abundance of transcripts in spermatocytes relative to spermatids is presumably due to meiotic sex chromosome inactivation, which counteracts transcriptional promiscuity[41].

Overall Y evolution and selective pressures

To further characterize selective signatures of Y evolution, we calculated branch-specific rates of nonsynonymous (d_N) and synonymous (d_S) substitutions, which revealed substantially increased d_N/d_S values for Y/W genes compared to both X/Z gametologues and autosomal precursors, consistent with previous work[43,44] (Extended Data Fig. 10a). Y branch d_N/d_S is always statistically significantly below 1 and we find no evidence for positive selection, neither for the most basal Y gene branches

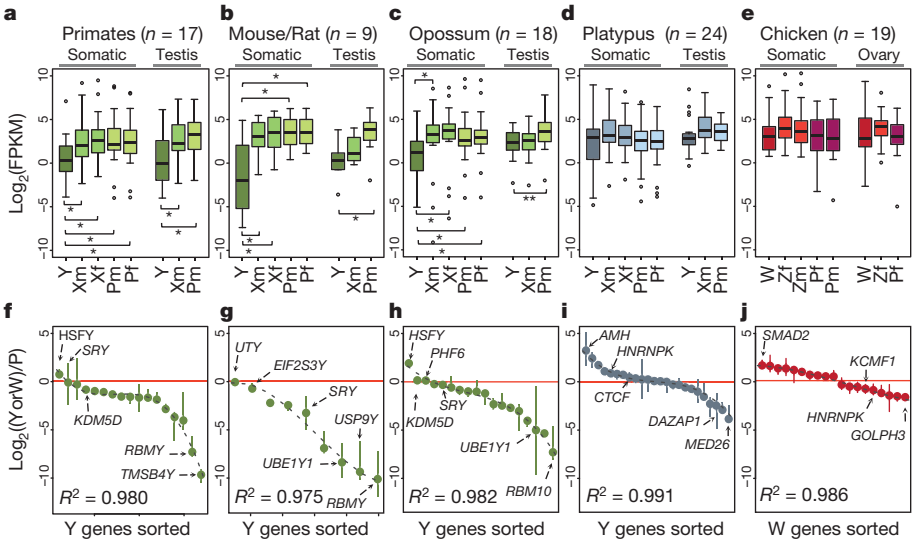


Figure 2 | Expression level evolution on amniote sex chromosomes. **a**–**d**, Expression level distributions (based on medians across somatic tissues or testis) of Y genes (Y); X genes in males (Xm) and females (Xf); precursors of X/Y genes on proto-sex chromosomes in males (Pm) and females (Pf). **e**, Similar distributions for (proto-)sex chromosomes in chicken. Note: for proto-sex chromosome plots, inferred expression output values were calculated per single gene copy/allele, to assess conservation of ancestral expression levels in current single Y (W) chromosomes. Significant differences (Mann–Whitney *U*-test): Benjamini–Hochberg-corrected $*P < 0.05$, $**P < 0.01$. Error bars, maximum and minimum values, excluding outliers. **f**–**j**, Median current/ancestral expression level ratios for individual genes. Error bars, 95% confidence intervals. Ratios are plotted on a $\log_2$ scale, that is, a ratio of 0 (non-$\log_2$ ratio of 1, red line) indicates that current and ancestral expression levels are similar. R^2 statistics represents the best fit to a third-order exponential curve. Gene numbers (*n*) underlying the data are indicated.

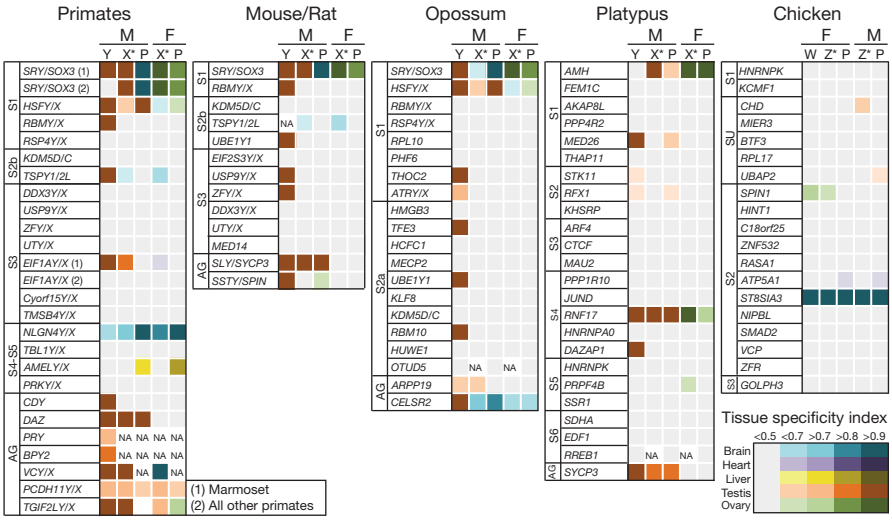


Figure 3 | Spatial expression pattern evolution on amniote sex chromosomes. Expression patterns of Y/X genes and proto-sex chromosome (P) precursors (inferred from 1:1 autosomal orthologues in outgroups), or 'added genes' (AG) and duplicate/precursor genes (asterisk), in males and females, as assessed by the tissue-specificity index. Grey indicates ubiquitous expression; darker tones of colours indicate increasing specificity of expression in a given tissue. For some genes, homologues could not be analysed (for example, owing to lack of X copy or missing data; not applicable, NA).

nor when combining descendant branches following gametologue differentiation (Methods). Thus, although some Y genes may have been shaped by positive selection in certain species[12,45], Y chromosome evolution is generally characterized by relaxed purifying selection, consistent with the impaired selection on the MSY[7,43,46]. Thus, Y genes probably generally preserved ancestral protein functions, consistent with the observation that even *SRY* may be functionally replaced by its X counterpart and functional precursor *SOX3* (ref. 47) and that most X genes likely underlying Turner syndrome (impaired development due to X chromosome monosomy) not only escape X inactivation and are under strong purifying selection but also have conserved Y gametologues[48]. Together with our expression and simulation results, these considerations suggest that most Y genes were, at least initially, preserved because of (regulatory) dosage constraints. Some Y genes then evolved new functions in spermatogenesis or development, primarily through temporal and/or spatial expression shifts, as exemplified by the therian sex-determining gene, *SRY*.

METHODS SUMMARY

We collected Illumina RNA-seq data (polyadenylated RNA fraction) for 166 tissue samples from nine mammals and two birds, and high-throughput Illumina paired-end genomic sequencing data for males and females from nine mammals and chicken. We developed a subtraction approach based on male/female RNA-seq data, Illumina genomic data and available genomes to identify and assemble Y (W) transcripts. The approach was validated using large-scale PCR and Sanger sequencing. We used genomic read coverage to predict multi-copy Y (W) genes, and male/female read coverage differences to identify X-linked genes in platypus. Phylogenetic tree reconstruction of sex-chromosome-linked genes and 1:1 autosomal orthologues from outgroup species, and associated synonymous substitutions rate analyses, were used to estimate ages of sex chromosome stratification events. We performed Monte Carlo simulations to assess non-random overlaps of Y (W) gene functions across different sex chromosome systems. RNA-seq read mapping and expression level estimation were performed using standard procedures and reference genomes to which newly identified Y (W) genes were added. To infer ancestral expression levels of Y (W) and X (Z) genes, we used 1:1 autosomal orthologues from outgroup species with different sex chromosome systems. All newly established Y and W sequences are provided in Supplementary Data 1.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 August 2013; accepted 17 February 2014.

1. Kashimada, K. & Koopman, P. *Sry*: the master switch in mammalian sex determination. *Development* **137,** 3921–3930 (2010).
2. Wilson, M. A. & Makova, K. D. Genomic analyses of sex chromosome evolution. *Annu. Rev. Genomics Hum. Genet.* **10,** 333–354 (2009).
3. Potrzebowski, L. *et al.* Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol.* **6,** e80 (2008).
4. Veyrunes, F. *et al.* Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.* **18,** 965–973 (2008).
5. Grützner, F. *et al.* In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature* **432,** 913–917 (2004).
6. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Phil. Trans. R. Soc. Lond. B* **355,** 1563–1572 (2000).
7. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Rev. Genet.* **14,** 113–124 (2013).
8. Julien, P. *et al.* Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol.* **10,** e1001328 (2012).
9. Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463,** 536–539 (2010).
10. Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483,** 82–86 (2012).
11. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).
12. Li, G. *et al.* Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res.* **23,** 1486–1495 (2013).
13. Marais, G. A., Campos, P. R. & Gordo, I. Can intra-Y gene conversion oppose the degeneration of the human Y chromosome? A simulation study. *Genome Biol. Evol.* **2,** 347–357 (2010).
14. Murphy, W. J., Pevzner, P. A. & O'Brien, S. J. Mammalian phylogenomics comes of age. *Trends Genet.* **20,** 631–639 (2004).
15. Ayers, K. L. *et al.* RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome. *Genome Biol.* **14,** R26 (2013).
16. Moghadam, H. K., Pointer, M. A., Wright, A. E., Berlin, S. & Mank, J. E. W chromosome expression responds to female-specific selection. *Proc. Natl Acad. Sci. USA* **109,** 8207–8211 (2012).
17. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434,** 325–337 (2005).
18. Murtagh, V. J. *et al.* Evolutionary history of novel genes on the tammar wallaby Y chromosome: implications for sex chromosome evolution. *Genome Res.* **22,** 498–507 (2012).
19. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286,** 964–967 (1999).
20. Wilson Sayres, M. A. & Makova, K. D. Genome analyses substantiate male mutation bias in many species. *Bioessays* **33,** 938–945 (2011).
21. Pearks Wilkerson, A. J. *et al.* Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. *Genomics* **92,** 329–338 (2008).
22. Chang, T. C., Yang, Y., Retzel, E. F. & Liu, W. S. Male-specific region of the bovine Y chromosome is gene rich with a high transcriptomic activity in testis development. *Proc. Natl Acad. Sci. USA* **110,** 12373–12378 (2013).
23. Weller, P. A., Critcher, R., Goodfellow, P. N., German, J. & Ellis, N. A. The human Y chromosome homologue of XG: transcription of a naturally truncated gene. *Hum. Mol. Genet.* **4,** 859–868 (1995).
24. Rodríguez-Delgado, C. L., Waters, P. D., Gilbert, C., Robinson, T. J. & Graves, J. A. Physical mapping of the elephant X chromosome: conservation of gene order over 105 million years. *Chromosome Res.* **17,** 917–926 (2009).
25. Rens, W. *et al.* The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biol.* **8,** R243 (2007).
26. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453,** 175–183 (2008).
27. Cutting, A., Chue, J. & Smith, C. A. Just how conserved is vertebrate sex determination? *Dev. Dyn.* **242,** 380–387 (2013).

28. Western, P. S., Harry, J. L., Graves, J. A. & Sinclair, A. H. Temperature-dependent sex determination in the American alligator: AMH precedes SOX9 expression. *Dev. Dyn.* **216,** 411–419 (1999).
29. Hattori, R. S. *et al.* A Y-linked anti-Mullerian hormone duplication takes over a critical role in sex determination. *Proc. Natl Acad. Sci. USA* **109,** 2955–2959 (2012).
30. Tsend-Ayush, E. *et al.* Identification of mediator complex 26 (*Crsp7*) gametologs on platypus X1 and Y5 sex chromosomes: a candidate testis-determining gene in monotremes? *Chromosome Res.* **20,** 127–138 (2012).
31. Dalloul, R. A. *et al.* Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol.* **8,** e1000475 (2010).
32. Warren, W. C. *et al.* The genome of a songbird. *Nature* **464,** 757–762 (2010).
33. Adolfsson, S. & Ellegren, H. Lack of dosage compensation accompanies the arrested stage of sex chromosome evolution in ostriches. *Mol. Biol. Evol.* **30,** 806–810 (2013).
34. Nam, K. & Ellegren, H. The chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata. *Genetics* **180,** 1131–1136 (2008).
35. Matunis, M. J., Michael, W. M. & Dreyfuss, G. Characterization and primary structure of the poly(C)-binding heterogeneous nuclear ribonucleoprotein complex K protein. *Mol. Cell. Biol.* **12,** 164–171 (1992).
36. Wu, Z. *et al.* Targeted ubiquitination and degradation of G-protein-coupled receptor kinase 5 by the DDB1–CUL4 ubiquitin ligase complex. *PLoS ONE* **7,** e43997 (2012).
37. Smith, C. A. *et al.* The avian Z-linked gene *DMRT1* is required for male sex determination in the chicken. *Nature* **461,** 267–271 (2009).
38. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25,** 25–29 (2000).
39. Zhou, Q. & Bachtrog, D. Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. *Science* **337,** 341–345 (2012).
40. Kondrashov, F. A. & Koonin, E. V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* **20,** 287–290 (2004).
41. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3,** 2179–2190 (2013).
42. Cocquet, J. *et al.* The multicopy gene *Sly* represses the sex chromosomes in the male mouse germline after meiosis. *PLoS Biol.* **7,** e1000244 (2009).
43. Wilson, M. A. & Makova, K. D. Evolution and survival on eutherian sex chromosomes. *PLoS Genet.* **5,** e1000568 (2009).
44. Berlin, S. & Ellegren, H. Fast accumulation of nonsynonymous mutations on the female-specific W chromosome in birds. *J. Mol. Evol.* **62,** 66–72 (2006).
45. Hughes, J. F., Skaletsky, H. & Page, D. C. Sequencing of rhesus macaque Y chromosome clarifies origins and evolution of the *DAZ* (*Deleted in AZoospermia*) genes. *Bioessays* **34,** 1035–1044 (2012).
46. Charlesworth, B. Model for evolution of Y chromosomes and dosage compensation. *Proc. Natl Acad. Sci. USA* **75,** 5618–5622 (1978).
47. Sutton, E. *et al.* Identification of *SOX3* as an XX male sex reversal gene in mice and humans. *J. Clin. Invest.* **121,** 328–341 (2011).
48. Park, C., Carrel, L. & Makova, K. D. Strong purifying selection at genes escaping X chromosome inactivation. *Mol. Biol. Evol.* **27,** 2446–2450 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. Harshman and the Lausanne Genomics Technology Facility for high-throughput sequencing support; I. Xenarios and the Vital-IT computational facility for computational support; S. Pääbo for great ape DNA samples; C. Roos for marmoset DNA samples; P. Jensen for chicken samples; E. Tsend-Ayush for help in determining the complete sequence of the *AMHY* gene in platypus; P. Gonzalez-Rubio for help with figure designs; M. Cardoso-Moreira, F. Carelli, A. Necsulea and M. Warnefors for comments on the manuscript; the Kaessmann group in general for discussions; and the Marmoset Genome Sequencing Consortium for making the marmoset genome assembly and annotation data available and for granting permission to use them for the analyses described in this study before publication. D.T.F. was supported by the Mexican National Council for Science and Technology (CONACyT). P.D.W. was supported by an ARC fellowship. F.G. was supported by an ARC fellowship. This research was supported by grants from the European Research Council (Starting Independent Grant: 242597, SexGenTransEvolution) and the Swiss National Science Foundation (Grant: 130287) to H.K.

Author Contributions D.C. performed most data processing and biological analyses. R.M. processed platypus genomic data and assessed X identity of contigs and the most likely chromosomal location of X gametologues in this species. D.T.-F. performed FISH experiments in platypus. L.F. and A.L. prepared samples and generated RNA-seq and genomic sequencing libraries. L.F. performed the large-scale PCR/Sanger sequencing validation experiments. P.D.W. provided elephant and tammar wallaby fibroblast samples and advised on these species' sex chromosome biology. F.G. provided platypus and echidna samples, supervised FISH experiments, and provided advice on the sex chromosome biology of these species. The project was supervised and originally designed by H.K. The paper was written by D.C. and H.K. with input from all authors.

Author Information RNA and DNA sequencing data as well as reconstructed Y/W sequences have been deposited in the Gene Expression Omnibus (GEO), Sequence Read Archive (SRA) and Transcriptome Shotgun Assembly (TSA) Database under the accession codes GSE50747, SRP029216, SRP026469 and PRJNA236159. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.C. (diegoclaudio.cortezquezada@unil.ch) or H.K. (henrik.kaessmann@unil.ch).

METHODS

RNA-seq data collection. For this study, we collected RNA-seq data (polyadenylated RNA fraction) for 166 tissue samples from eight mammals (human, chimpanzee, gorilla, orangutan, macaque, mouse, opossum and platypus) and birds (chicken and ostrich) previously generated by us[41,49,50] and other groups[33,51] (Illumina Human Body Map 2.0 and ostrich) (Supplementary Tables 1–3). These data include additional deep-coverage RNA-seq data that we generated using our original libraries[50] for male brain and testis, as well as chicken ovary, for the purpose of this and parallel projects[41,49] (in which these data are initially published), in order to increase the probability to detect Y (W) transcripts. In addition, we generated RNA-seq data for male and female marmoset and rat (brain, cerebellum, heart, kidney, liver, testis and ovary), male and female elephant (fibroblasts) and male wallaby (liver and testis) (Supplementary Tables 1–3). Overall, we used a total of 6.78 billion RNA-seq reads in this study, with 3.14 billion reads generated by us in the framework of this and the parallel projects. RNA-seq libraries were prepared using standard Illumina protocols and sequenced on Illumina GA IIx or HiSeq 2000 platforms (Supplementary Tables 1–3). Comparability of data derived from different library procedures was validated as described elsewhere[49]. RNA-seq data generated in the framework of our previous study by Brawand *et al.*[50] was processed using Ibis[52] to increase the number of usable reads and reduce the error rate (see Supplementary Note in Brawand *et al.*[50]). All the other samples were processed using more recent Illumina base callers.

Genomic sequencing data generation. We sequenced genomic DNA from two individuals (1 male and 1 female, same individuals as used for RNA-seq) from gorilla, orangutan, marmoset, rat, elephant, opossum, wallaby, echidna, platypus and chicken using standard Illumina protocols (Truseq DNA) for short insert size (target size: 400–450-base pair (bp)) paired-end libraries. Each genome was sequenced in one or (approximately) half of an Illumina HiSeq 2000 lane (~150–300 million 100-bp paired-end reads per genome) (Supplementary Tables 1–3).

Assembly of Y- and W-linked transcripts. Several previous studies developed approaches that contrast male and female genome and/or transcriptome data and allowed for successful reconstruction of Y/W genes or contigs[15,16,53–55]. To assemble Y (W) transcripts for the purpose of this study, we developed a specific male/female subtraction approach initially based on the RNA-seq data and available female reference genomes (Extended Data Fig. 1). First, to enrich for male-specific (that is, Y chromosome) sequences, RNA-seq reads from male tissues were mapped using TopHat 1.4.0 (ref. 56) to female reference genomes downloaded from the Ensembl database[57] (release 67); in the case of human, chimpanzee and mouse (reference genomes based on DNA from males), Y chromosomes and unassigned contigs were removed from the genomic data before RNA-seq read mapping. Similarly, the W and Unknown chromosomes of chicken were also removed from the (female) reference genome. In the case of elephant, for which no published genome is available, we mapped the male RNA-seq reads to the genomic scaffolds assembled with SOAPdenovo2 (ref. 58) from our Illumina female genomic sequencing data for this species. For each species, we then assembled the transcriptomes of all male (female for chicken) tissues individually and all male (female) tissues combined on the basis of unmapped reads from the previous step using SOAPdenovo-Trans[59] (http://soap.genomics.org.cn) (seeds: k-mers of 21 bp, 23 bp, 25 bp, 27 bp, 29 bp) and Trinity[60] (seed: default k-mer of 25 bp). Because resulting contigs represent consensus sequences that may have resolved RNA-seq sequencing errors (potentially preventing individual RNA-seq reads to be correctly mapped to the reference genomes; step 1 above), we sought to map these contigs against the female reference genomes using BLAT[61]. Contigs mapping over 90% of their sequence to the reference genome at 100% identity were discarded from further analyses. Also, contigs onto which RNA-seq reads from female tissues could be mapped (using Bowtie 0.12.5, ref. 62) over 90% of the sequence with 100% identity were discarded, thus yielding a final list of potential Y/W-linked transcript sets (Supplementary Table 4). Performing *de novo* transcript reconstructions with different k-mers and different bioinformatics tools (SOAP and Trinity) resulted in several contigs of various lengths and exon numbers for each individual gene. To obtain transcripts of maximum length for each gene, we merged overlapping contigs into single consensus transcript assemblies. Specifically, we first used TGICL (http://compbio.dfci.harvard.edu/tgi/software), which implements Megablast and CAP3, to group the contigs on the basis of sequence similarity and merge them. We then confirmed the transcript assemblies on the basis of multiple alignments of the initial contigs generated with MUSCLE[63].

To connect split consensus sequences (due to low local coverage for lowly expressed genes), we used BLASTn to detect reads that map to the ends of contig fragments and CAP3 (ref. 64) to generate a new consensus sequences; this process was repeated until maximum extension was achieved. Specifically, contigs were initially trimmed (20 nt on each end), due to the overall lower sequence quality at the end of contigs. In each round of extension, we required at least five male RNA-seq reads to align at 100% identity and over a minimum of 30 nt at the ends of the contig and then obtained new extended sequences. We note that only for 38% of genes gaps needed to be closed using this approach (adding, on average 8% of coding sequence and 15% of UTR sequence) and that, in the remaining cases, full coding sequences with UTR sequences were already reconstructed during the previous steps. We validated the power and accuracy of this approach by performing simulations on fully sequenced Y chromosomes from human, chimpanzee, macaque and mouse (Extended Data Fig. 2). For these species, we repeatedly re-sampled an increasing number of RNA-seq reads from male tissues and applied our Y reconstruction pipeline to each subset of reads. We then calculated the lengths of our Y gene reconstructions with respect to the sequences of annotated protein-coding genes on the Y chromosomes at ≥ 99% of identity. The results from these analyses (Extended Data Fig. 2) show that even for human and chimpanzee, for which we produced no data in addition to those from Brawand *et al.*[50], nearly all Y genes are detected (Extended Data Fig. 2a, b). These genes are known to be lowly expressed or not to be expressed at all (*AMELY*) in the sampled tissues. Furthermore, many genes (in the case of chimpanzee) or most genes (in the case of humans, for which more data are available from Brawand *et al.*[50]) are reconstructed over most of their length (for example, 19 of 26 human Y genes with more than 95% reconstructed sequence lengths). For macaque and mouse, for which we produced additional testis and brain data, the detection/reconstructions are even more complete, such that, for example, all Y genes (macaque) or all but one Y gene (mouse) are detected with large percentages of reconstructed sequence lengths (Extended Data Fig. 2c, d). Notably, repeated resampling analyses (Extended Data Fig. 2e) show that for mouse and macaque, for which we produced additional data, saturation in Y gene reconstruction is reached more rapidly than for chimp and human. Importantly, the fact that saturation in Y reconstruction is reached with 226 million RNA-seq reads, which corresponds to the number of reads we have for gorilla (the species with the overall smallest amount of data among the species for which we perform *de novo* reconstructions in our study and for which we have complete tissues sets) suggests that our RNA-seq data are sufficient to detect/reconstruct Y genes in an optimal way given our initially transcriptome-based approach (that is, the amount of RNA-seq data we produced is not limiting our capacity of predicting Y genes). However, we note that it is more likely that we underestimate Y gene repertoires for orangutan, elephant, wallaby and echidna than for the other species in our study, given that we could not obtain complete tissues sets and/or only generate genomic sequencing data for these rare species (Supplementary Tables 1–3) (that is, we might miss species-specific genes not expressed in the sampled tissues).

We finally note that we used our own Y gene/transcript predictions (derived from our subtraction approach) for all analyses in this study, except for human, chimpanzee and macaque. Ensembl mouse Y gene annotations are available, but have not yet been published. We only used the assembled mouse Y chromosome sequence[65] (GRCm38) to map our contigs (Extended Data Fig. 9).

Validation of Y/W-specific contigs using high-throughput male/female genomic sequencing data. To validate true Y/W transcripts (Extended Data Fig. 1), we aligned male and female Illumina genomic reads to all contigs using BLASTn, requiring 100% of identity. In the case of contigs that are not Y-linked, male and female reads would have the same probability to map along the sequence of the contig, which is what we observed for X/Z gametologue sequences used as controls (Supplementary Tables 5–18). Therefore, we considered contigs to likely be Y/W-specific when male (female in chicken) reads mapped along most of the sequence of a contig (>60%) and the coverage obtained using female (male) genomic reads was lower than 20% of the contig length. However, in general, Y-linked contigs are supported by >90% of male reads and <10% for female reads (Supplementary Tables 5–18).

Large-scale PCR-based validation of subtraction approach. To validate our transcriptome/genome-based subtraction approach for the detection/assembly of Y (W) transcripts, we screened male/female genomic DNA (extracted using a standard phenol/chloroform protocol) using a large-scale (~4,000 reactions in total) PCR approach (Extended Data Fig. 1), performed with JumpStart REDTaq ReadyMix (Sigma-Aldrich) according to the manufacturer's protocol in 96-well plates containing pre-filled oligonucleotide mixes (Invitrogen). We thus assessed all putative Y transcripts (>300 nt) and, as a control, 100 other contigs not supported in the genomic read coverage validation step that were absent from the reference genome. For primer design, putative Y transcripts were mapped against female reference genomes with BLAT[61] (mapping location with low identity in homologous regions, for example, X gametologues or autosomal precursors), and we then designed two pairs of PCR primers for each transcript using Primer3 (ref. 66), avoiding regions of high sequence similarity. For other contigs, optimal primers were designed using Primer3 directly based on the contig sequence. All primer sequences are available upon request. Negative (reagents without DNA template) and positive (autosomal β-actin and previously known Y genes) controls were included in each PCR plate. Technical replicate experiments were performed for each plate to confirm male-specific amplifications of predicted Y transcripts. PCR reactions were run on agarose gels, and male-specific DNA bands (in both

replicates) were extracted and purified using Gen Elute Agarose Spin Columns (Sigma-Aldrich) and sequenced using standard Sanger sequencing. Resulting sequences were then compared to reconstructed Y transcripts. All transcripts predicted using the male/female transcriptome/genome-based subtraction approach could be validated using this PCR approach (that is, they could only be amplified in PCR reactions with male genomic DNA as template, and sequenced PCR products matched reconstructed Y transcripts with >99% identity). Control contigs could always be amplified using both male and female DNA, suggesting that these contigs represent autosomal sequences missing from reference genomes.

Reconstruction of Y-linked genes using genomic data (independent of transcriptome data). Given that our subtraction approach depends on transcription information, we used a genomic approach to detect and reconstruct Y genes for species for which we did not generate RNA-seq data or lacked testis data, and to verify that Y genes were not missed or incomplete by our transcriptome-based subtraction approach due to lack of expression in sampled tissues. This approach seeks to identify orthologous genes of known Y genes in a given species on the basis of male/female high-throughput genomic sequencing data. Specifically, we used known Y coding sequences from closely related species and assemble Y gene candidates by aligning reads with the highest similarity from the male genomic library to these coding sequences. We then discarded contigs onto which reads from the female genomic library mapped over >50% of the sequence (for example, representing X-linked gametologues). To assess the power and accuracy of this approach, we used known human Y genes as templates for detecting orthologous Y genes in simulated male/female genomic chimpanzee, macaque and mouse genomic read data sets. All orthologous Y genes were detected in this test (that is, there were no false negatives) and human genes known to not be present in these species (for example, *AMELY* in mouse) were not erroneously identified (that is, there were no false positives). We thus applied this method to retrieve missing/partial orthologous Y genes, particularly in orangutan (missing testis RNA-seq data), elephant (only fibroblast RNA-seq data) and echidna (no RNA-seq data), where we used known genes from human, mouse and platypus, respectively (Supplementary Tables 5–18). We also applied this approach to detect orthologues of human pseudogenes across primates. For Y genes that could not be detected/reconstructed using the transcriptome data in a given species, we assessed whether genomic Illumina reads unique to males corresponding to these genes could be identified for that species, using as templates Y orthologues from the most closely related outgroup species. If no such reads could be identified, the Y gene was considered absent/lost/pseudogenized beyond recognition. Thus, the only genes that could be missed using the combination of our detection approaches are genes that are not expressed in the sampled tissues and/or for which there is no known Y orthologue. We note that given that echidna does not have a reference genome, we also assembled echidna X gametologues using platypus X gametologues as templates (Supplementary Tables 5–18). Finally, we assembled W sequences from female ostrich RNA-seq data using chicken protein-coding and non-coding contigs as templates, applying an approach that is analogous to that based on genomic sequencing data and known orthologues (see above). Z gametologues of the two identified ostrich W genes were reconstructed using chicken Z genes as templates.

Definition of Y gene names and X gametologues. To establish Y (W) gene identity and to identify and extract X gametologue sequences, we searched Ensembl-annotated gene sequences and NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/) for the closest homologue using BLASTn and BLASTx (matches were confirmed using visual inspection of sequence alignments). Contigs without any significant match in the databases were considered to be noncoding. We delimited coding sequences in the Y/W transcripts based on MUSCLE[63] and PRANK[67] alignments with known (annotated) gametologues or orthologous genes. All Y, W, X and Z contig and coding sequences are available in Supplementary Data 1. We note that our RNA-seq-based Y transcript reconstructions are expected to yield the most frequent isoform for a given Y gene.

Prediction of multi-copy genes. The amount of genomic Illumina read data for the different species is generally too low to allow for reliable Y gene copy number estimation directly based on genomic read coverage. We thus developed an approach based on simulations, which establishes the behaviour of the read coverage from genomic sequencing libraries as a function of gene copy number as a statistical framework for estimating Y gene copy numbers (Extended Data Fig. 10b). Specifically, we constructed mock genomes (length: 3.3 Gb) that contain one mock gene (drawn from a random pool of distinct human cDNAs of different lengths: >1 kb and <10 kb) with a defined number of copies (remainder of the genome represented by 'Ns'). For every round of simulation, we then constructed a standard (mock) Illumina genomic read data set (100 bp paired-end reads, 5× genome coverage) including reads for a given copy number of the mock gene, mapped the genomic reads onto the gene in the genome using BLASTn (100% identity), and calculated the median coverage along the gene's sequence. We then repeated this process 1,000 times for different copy numbers (1–20) and for 1,000 mock genes. We finally obtained a theoretical function of the coverage distributions with respect to copy number differences (Extended Data Fig. 10b). In a second step, we selected from the respective Ensembl reference genomes cDNAs from 1,000 different genes (>1 kb), mapped the Illumina read data from this study or Ensembl-derived genomic data onto these cDNAs, and then calculated the median coverage for each cDNA. The median coverage value across all cDNAs was taken as the baseline that represents a single-copy gene with two alleles. Together with the function obtained from the simulations (see above), we used this baseline to estimate copy numbers in our set of Y (W) transcripts. Notably, known members of the ampliconic families were detected as being multi-copy (Supplementary Tables 5–18).

Prediction and characterization of X-linked contigs and genes in platypus. We used the genomic reads obtained for male (M) and female (F) platypus to evaluate potential X identity of the 24 contigs that harbour the homologues of identified Y genes (one homologue/contig could not be identified for 1 of the 25 detected Y genes, that is, *RREB1Y*). Specifically, given that the read coverage along the non-recombinant regions of the X chromosomes in males is expected to be half of the coverage observed in females for similar depth of sequencing (M:F $\log_2$ ratio = −1), direct comparison of the coverage between sexes allows identification of putative X-linked contigs. Thus, genomic reads were first aligned to the reference genome using Bowtie[62]. Uniquely aligned reads (one mismatch) were used to compute the median coverage for male (C_M) and female (C_F), in windows of 100 bp. The C_M/C_F ratios obtained for each window were averaged along the contig to obtain one mean C_M/C_F ratio per contig. Additionally, the $C_F - C_M$ difference was computed for each window and averaged along the contig to obtain the mean difference in coverage between female and male. Given that the coverage obtained for the autosomes is ∼14×, the expected mean difference for non-recombinant contigs is ∼7×; however, changes in coverage can lead to considerable local deviations from this value. To statistically support X-linked contigs, we tested whether M:F coverage ratios show statistical deviations from two reference values (M:F $\log_2$ ratio = 0, that is, no difference in coverage; M:F $\log_2$ ratio = −1, that is, twofold higher coverage in females) using the one-sample Wilcoxon signed rank test (Benjamini–Hochberg-corrected, $P < 0.05$). For a fully X-linked contig, a significant M:F $\log_2$ ratio deviation from 0 and a non-significant deviation from −1 is expected. To extend the list of known X-linked genes for the proto-sex chromosome analyses and Gene Ontology (GO) simulations (see below), we identified all genes in contigs that showed a M:F ratio ≤ 0.75 and a non-zero genomic read coverage over ≥ 50% of the contig.

To narrow down the possible location of the contigs containing the X gametologues to one of the five X chromosomes, we relied on the synteny between platypus and two other species: human and chicken. Our approach assumes that when most regions of a given contig *C* align close to regions that are orthologous to one of the assembled X chromosomes, this is because contig *C* is most likely located in that particular X chromosome. Thus, we implemented an algorithm that scans for a given contig all the individual alignments (that is, orthologous regions) between platypus and the reference genome and seeks to identify the X origin of each alignment in the platypus. Denoting by N_C the set of alignments involving contig *C*, and by N_X the set of alignments involving X chromosomes, for each alignment *A* in N_C the algorithm finds the alignment *B* in N_X that maps closest to *A*. Thus, alignments *A* would be assigned to the particular X chromosome to which alignment *B* belongs. In the end, the most likely location of contig *C* would be the X chromosome with the highest fraction of *C* sequence assigned to it. The estimated location for each of the 23 contigs according to the orthology with human and chicken is shown in Supplementary Tables 19, 20. We note that to define the set N_X, we considered both the partially assembled X chromosomes (that is, X1, X2, X3 and X5) and known X-linked contigs[4], which provide an additional 26.9 Mb of sequence. Only alignments in which the length in platypus is ≥ 500 nt were included in the set N_X. Alignments for platypus/human (ornAna1.hg19.net; http://hgdownload-test.cse.ucsc.edu/goldenPath/ornAna1/vsHg19/) and platypus/chicken (ornAna1.monDom5.net; http://hgdownload-test.cse.ucsc.edu/goldenPath/ornAna1/vsMonDom5/) were downloaded from the UCSC genome browser website (http://genome.ucsc.edu/).

Phylogenetic analysis. For phylogenetic tree reconstructions, we aligned coding sequences of Y/X (W/Z) gametologues and their orthologous sequences in other vertebrates using PRANK[67] based on encoded amino acids sequences, except for therian S1 and marsupial-specific S1 trees, for which PRANK amino acid sequence alignments were used. Poorly aligned regions were removed using BMGE[68]. The most likely phylogenetic tree and associated bootstrapping values for each gene or concatenated gene set were obtained using PhyML[69] (control file parameters: -i gene_file.phy -d nt -q -f m -t e -v e -a e–use median -b 1000). Note that in the case of multi-copy genes we included all different gene copies (where these were available or could be reconstructed) in our concatenation-based tree analyses (that is, phylogenetic tree reconstructions, d_S analyses) by repeatedly (100 times) and randomly selecting gene copies for the concatenations, and then reconstructing phylogenetic

trees (100 bootstrap replicates) and performing d_S analyses (as detailed in the following section) on each concatenated alignment.

We carefully inspected all individual gene trees for potential evidence of gene conversion. We could not identify such cases, suggesting that gene conversion is not frequent in sex chromosome evolution. Consistently, a previous placental sex chromosome study[43] only found evidence for gene conversion in a single exon from one gene (*ZFY*) that we had excluded in our tree reconstruction for *ZFY*. We also considered the possibility of gene conversion for the *UBE1Y1* and *KDM5D* genes, for which partial gene conversion was previously suggested[70]. However, the trees obtained in our reconstructions for the full-length alignments (which are based on many more Y and autosomal orthologues from outgroup species than in the previous study[70]) are very similar to those obtained when removing the previously proposed gene conversion region, and very clearly show that these gametologues differentiated independently (as part of the two independent strata S2a and S2b) in placentals and marsupials (Extended Data Fig. 4).

Nonsynonymous and synonymous substitution analyses. Pairwise alignments of coding sequences of X/Y (Z/W) gametologues were obtained using PRANK[67] based on encoded amino acid sequences. d_S values were then calculated using codeml (pair-wise option) as implemented in PAML[71] (Supplementary Tables 5–18). Genes with the same phylogenetic position in monotremes (where multiple strata formed in the common ancestral branch leading to platypus and echidna) were subsequently grouped into strata based on both the phylogenetic information and statistical partitioning of d_S values using a consensus of algorithms (Hartigan–Wong, Lloyd, Forgy and MacQueen) as implemented in the R software package (http://www.r-project.org).

Branch-specific d_N, d_S and d_N/d_S values in phylogenetic trees were estimated using the codeml free-ratio model as implemented in the PAML[71] for individual gene and concatenated sequence alignments, using the PhyML coding sequence trees (see above) as input. To assess whether Y (W) coding sequences have evolved under the influence of positive selection, we used a comparative branch-site test for the most basal branch following stratum formation or combining all descendant branches following Y/X (W/Z) gametologue divergence (foreground branches). We compared the likelihood of a model that allows for $d_N/d_S > 1$ at a subset of sites (that is, d_N/d_S is estimated from the data) for the foreground branches to that of a null model where d_N/d_S of this site class was fixed to 1 for these branches. Statistical significance was assessed using likelihood ratio tests[72].

To assess the age at which sex chromosome strata originated, we first calculated 95% confidence intervals of d_S values for each stratum (branches just before and after each stratification event) on the basis of bootstrapping analyses (100 replicates each) of concatenated alignments for each stratum. To account for sequence variability multi-copy gene sequences, we repeatedly selected random copies for these genes in the concatenation-based d_S inferences (see above, Extended Data Fig. 5). To perform these analyses, genes included in the concatenations needed to be present in relevant ingroup and outgroup species. Given that the mutation rate was previously reported to be higher in male eutherians and found to be higher in all therians (that is, including marsupials) in our study, and that different chromosomes spend different proportions of time in males (autosomes, 50%; Y, 100%; X, ~33%), we needed to correct observed d_S values of branches following stratification (Y and X, respectively) relative to preceding branches (proto-sex chromosomes, where genes evolved as autosomes). We thus adjusted therian Y/X d_S branch values to match proto-sex chromosome values using the following formula that is based on the different proportion of times chromosomes spend in the two sexes:

$$F * (Xd_S + (Xd_S * 1/3)) = (Yd_S*0.5) / F$$

where Xd_S is the observed X chromosomal d_S for a given post-stratification branch, Yd_S is the observed Y chromosomal d_S for a given post-stratification branch, and F is a factor that corrects for the uncertainty in observed d_S values due to differences in selective constraint at synonymous sites[73] (that is, relaxation of selection for Y genes[43] and this study, relative to X-linked sequences) and stochastic effects. We then used both the pre- and corrected post-stratification d_S values and the sum of these values together with known lineage divergence times (retrieved from http://www.timetree.org/) surrounding a given stratification event, to estimate actual ages of the stratification event (Extended Data Fig. 5a–f, 95% confidence intervals for d_S and age estimates).

Monotremes do not seem to have male mutation bias (this study). For birds, we do not detect male-biased mutation, potentially owing to compensation of a higher rate observed for Z-linked genes by relaxation of constraint for W genes. Notably, male mutation bias in birds could previously only be detected on the basis of long intronic sequences[44]. Thus, to estimate ages of stratification events in monotremes and birds, we performed similar analyses as described for therians (see above) but used median Y/X (Z/W) d_S values instead of corrected values.

GO simulations. We performed Monte Carlo simulations to assess non-random overlaps of Y (W) gene functions across different sex chromosome systems using as an initial input ancestral sets of proto-sex chromosomal genes defined as: human (502 XCR and X-added region genes with 1:1 orthologues in opossum), opossum (354 XCR genes with 1:1 orthologues in humans), platypus (624 annotated X_{1-5}-linked genes and X-linked predictions from this study that have 1:1 orthologues in human/chicken), chicken (421 Z-linked genes that have 1:1 orthologues in human/platypus). We assigned GO terms[38] to each gene on the basis of human annotations (http://www.geneontology.org/GO.annotation.shtml). In each round of simulation, we randomly removed genes from the ancestral set of proto-sex chromosome genes pools until currently observed Y (W) gene numbers were reached (human, 16; opossum, 19; platypus, 22; chicken, 19). We then compared the overlap of Y (W) gene functions for all pairs of species, where the overlap is defined as the smallest number observed in a given comparisons (for example, if a given GO term is found 5 times in one simulated extant gene set of one species and 1 time in another, then the overlap is counted as 1). This process was repeated 100,000 times. We then compared the obtained simulated distributions of GO overlaps with those observed across species. To assess statistically significance of potentially non-random overlaps, we applied one-tail alpha tests (Benjamini–Hochberg-corrected $P < 0.01$).

Expression levels and spatial patterns on current and ancestral (proto) sex chromosomes. We added Y (W) coding genes and noncoding sequences to the reference genomes to assess their expression levels. We then mapped all RNA-seq reads with TopHat 1.4.0 (ref. 56) and then used Cufflinks 2.0.0 (ref. 74) (all mapped reads, embedded multi-read and fragment bias correction) to calculate the FPKM (fragments per kilobase of transcript per million mapped reads) values for all genes in the genomes with our refined annotations[50]. We normalized expression levels across samples and species with a median scaling procedure[50]. In case multiple samples from different male individuals were available for a given tissue, the median expression value across these samples was used for further analyses. Similarly, frontal cortex and cerebellum expression values were combined into a single median 'brain' value. The tissue specificity index (TSI) for a given gene was calculated as the expression level (FPKM) in the tissue with the highest expression level divided by the sum of expressions values in all tissues[8]. To infer ancestral expression levels, we exploited the fact that the current sex chromosomes are derived from ancestral autosomes and therefore have autosomal counterparts in species with non-homologous sex chromosomes, which are informative with respect to proto-sex chromosome expression patterns[8]. We thus calculated ancestral sex chromosome expression levels as median expression levels for autosomal 1:1 orthologues of Y/X (W/Z) genes in outgroup species with different sex chromosomes systems: median expression across platypus, chicken and *Xenopus tropicalis* (data published in ref. 49) for therian sex chromosomes; and median expression across therians and *Xenopus* for platypus or chicken (1:1 orthologous gene set numbers: therians, 132; chicken, 294; platypus, 424; including known and predicted X-linked contigs). The TSI was also calculated for inferred proto-sex gene expression. Note that to assess the extent of conservation of ancestral expression levels in the current single Y (W) chromosome (Fig. 2 and Extended Data Fig. 8b), inferred expression output values were calculated per single gene copy (that is, expression levels of 1:1 orthologues on autosome pairs from outgroup species with different sex chromosome systems were divided by 2).

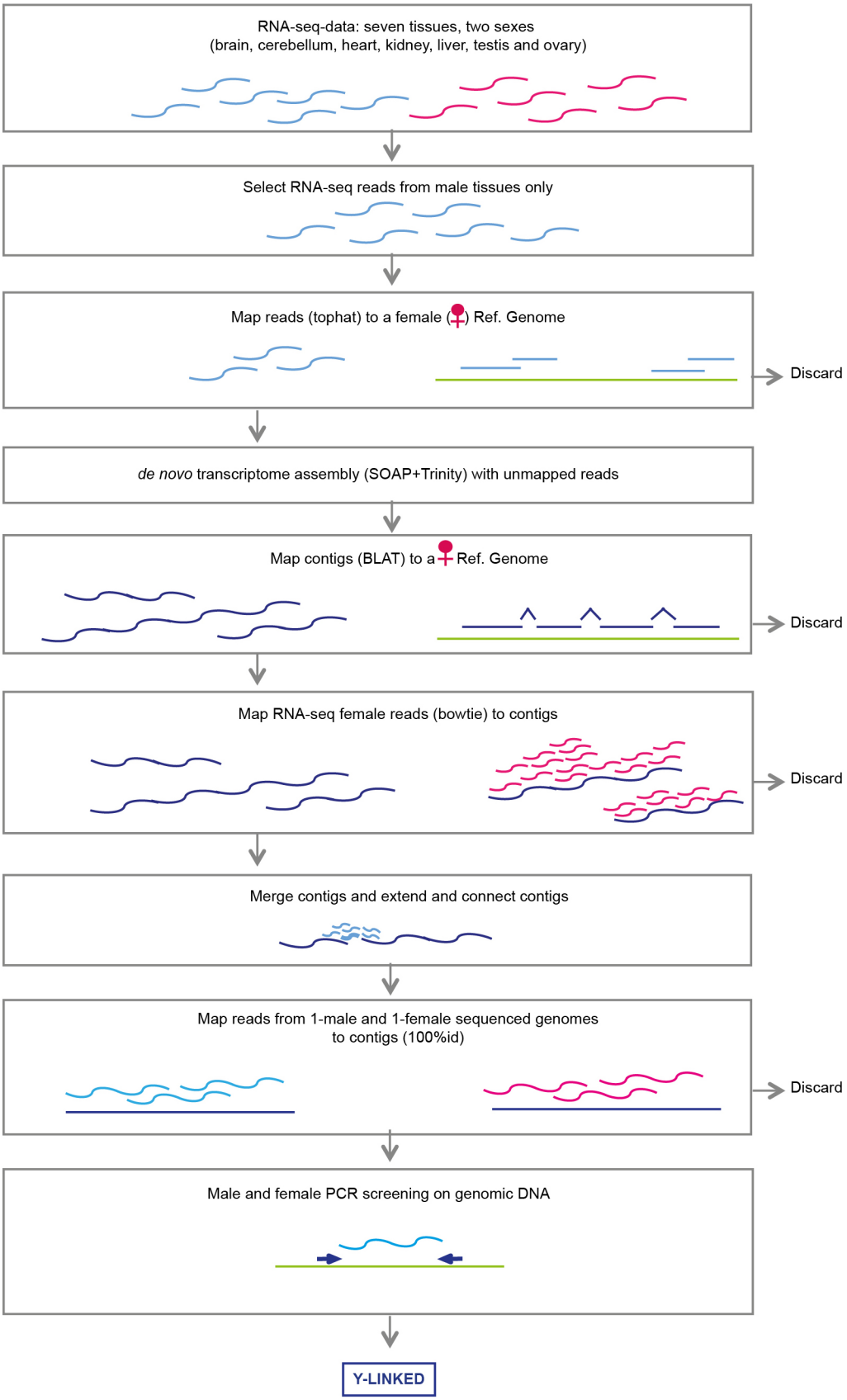
General information about statistical analyses and tools. All statistical analyses, graphical representations and most of the simulations were carried out using the R software package (http://www.r-project.org). Other analyses and simulations were performed with custom in-house Perl scripts (Supplementary Data 1). Multiple test corrections were performed with Biobase and multtest packages from the Bioconductor software package[75]. All tests are two-sided except for the GO simulations (see above). All reported P values were corrected for multiple tests using the Benjamini–Hochberg procedure. We chose the tests (non-parametric) on the basis of distribution of the variables. We used non-parametric statistics or randomization tests for non-normal distributions. The variations in the data are presented in each figure.

Fluorescence *in situ* hybridization. *AMHY* probe was generated by PCR using the following conditions. Each reaction was performed in 25 µl volume containing 25–50 ng platypus testis cDNA, 5× PCR reaction buffer with $MgCl_2$ (Promega), 5 U μl^{-1} of *Taq* DNA polymerase, 0.4 µM of each forward (5′-GGAGAGTCAAAGGTTCAAATCTGG-3′) and reverse (5′-AGCCACCATATTAGGCATGAGG-3′) primer and 0.1 mM dNTPs. Initial denaturation was carried out at 96 °C for 3 min followed by 35 cycles of: denaturation at 96 °C for 30 s, annealing at 59 °C for 1 min and extension at 72 °C for 2 min. Final extension was performed at 72 °C for 7 min. The product was then gel-purified using PureLink Quick Gel Extraction Kit, Invitrogen, cloned and sequenced.

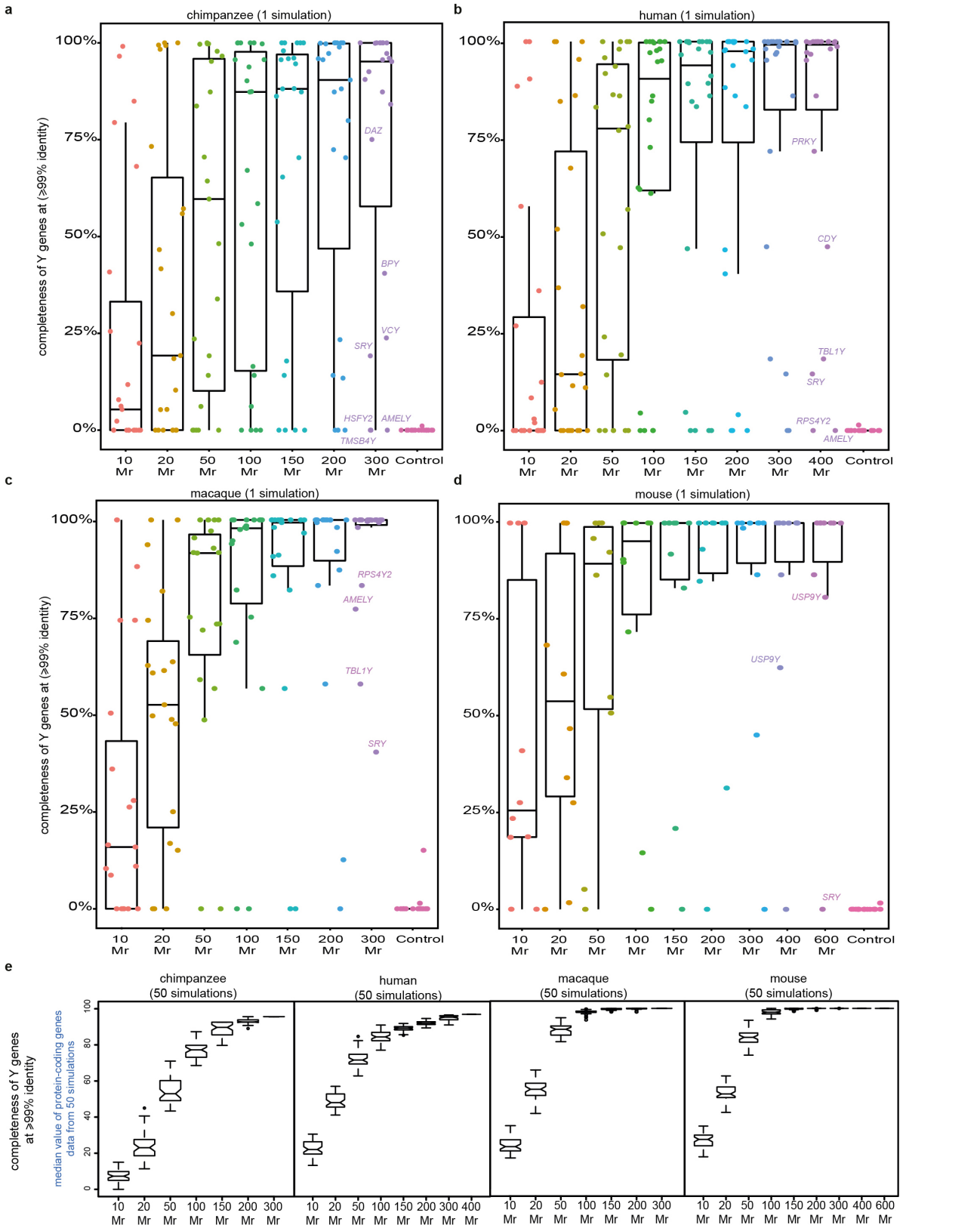
We physically mapped platypus *AMHY* by hybridizing the labelled *AMHY* PCR product onto male platypus mitotic metaphase chromosomes through TSA-FISH[76] (Extended Data Fig. 7e, f) (note that standard FISH using BAC clones could not be performed, as no BACs containing *AMHY* could be identified in the relevant available

libraries). We co-localized *AMHY* with Y_5-specific BAC 243O9 using standard FISH as described before[30,77], which confirmed Y_5 localization (not shown). For the TSA-FISH the probe was labelled with biotin-16-dUTP using gene-specific primers. PCR conditions were as described above except that the dNTP mix contained 0.2 mM of each dATP, dCTP and dGTP, 0.1 mM dTTP (Bioline) and 0.1 mM biotin-16-dUTP (Roche). The labelled probe was then gel-purified and partially digested for 10 min at 37 °C to ensure a fragment length of around 300 bp. The digestion reaction of 52 µl contained 1 µl DNase I (10 U μl^{-1}; diluted 1:3,000) (New England Biolabs), 10× DNase I Reaction Buffer and 350 ng gel-purified probe DNA. The probe was cleaned up using Micro Bio-Spin 6 Chromatography Columns (Bio-Rad) and precipitated with 2 µl salmon sperm DNA and 500 µl ethanol. The probe was then re-suspended in 5 µl of 50% formamide, 10% dextran sulphate and 2 × SSC and denatured at 80 °C for 10 min. Slides were pretreated with 100 µg ml RNase A/2 × SSC at 37 °C for 30 min and with 3% hydrogen peroxide (H_2O_2) in 1 × PBS for 30 min followed by 0.01% pepsin in 10 mM HCl at 37 °C for 10 min. Finally, slides were re-fixed in 1 × PBS/50 mM $MgCl_2$/1% formaldehyde, followed by dehydration in an ethanol series. Pre-treated slides were denatured in a 70% formamide/2 × SSC solution at 70 °C for 3 min and dehydrated. Hybridization was performed overnight at 37 °C. TSA Biotin System Kit (Perkin Elmer) was used for detection according to manufacturer's instructions. Slides were washed three times for 5 min in 50% formamide/2 × SSC, 5 min in 2 × SSC at 42 °C and 5 min in 0.1 × SSC at 60 °C, followed by incubation in 300 µl of TNB blocking buffer (0.1 M TRIS-HCl, pH 7.5, 0.15 M NaCl, 0.5% Blocking Reagent supplied in kit). Biotin label was detected with SA-HRP diluted 1:100 in TNB buffer (30 min incubation). Slides were then washed 3 times for 5 min in TNT buffer (0.1 M TRIS-HACL, pH 7.5, 0.15 M NaCl, 0.05% Tween20). Amplification was done by a 10-min incubation with 300 µl Tyramide-Biotin conjugate diluted 1:50 in Amplification Diluent (both supplied in kit), then washed 3 times for 5 min in TNT buffer. To visualize Tyr-Bio, Alexa Fluor 488–Streptavidin (Jackson ImmunoResearch) diluted 1:100 in TNB buffer was applied for 45 min at 37 °C. Slides were washed three times for 5 min each in TNT buffer and then counterstained with DAPI (0.2 µg ml^{-1} in 2 × SSC) for 1 min, and washed with water, air dried and mounted (Vectashield, Vector Laboratories). Images were taken with a Zeiss AxioImager Z.1 epifluorescence microscope equipped with a CCD camera and Zeiss Axiovision Software.

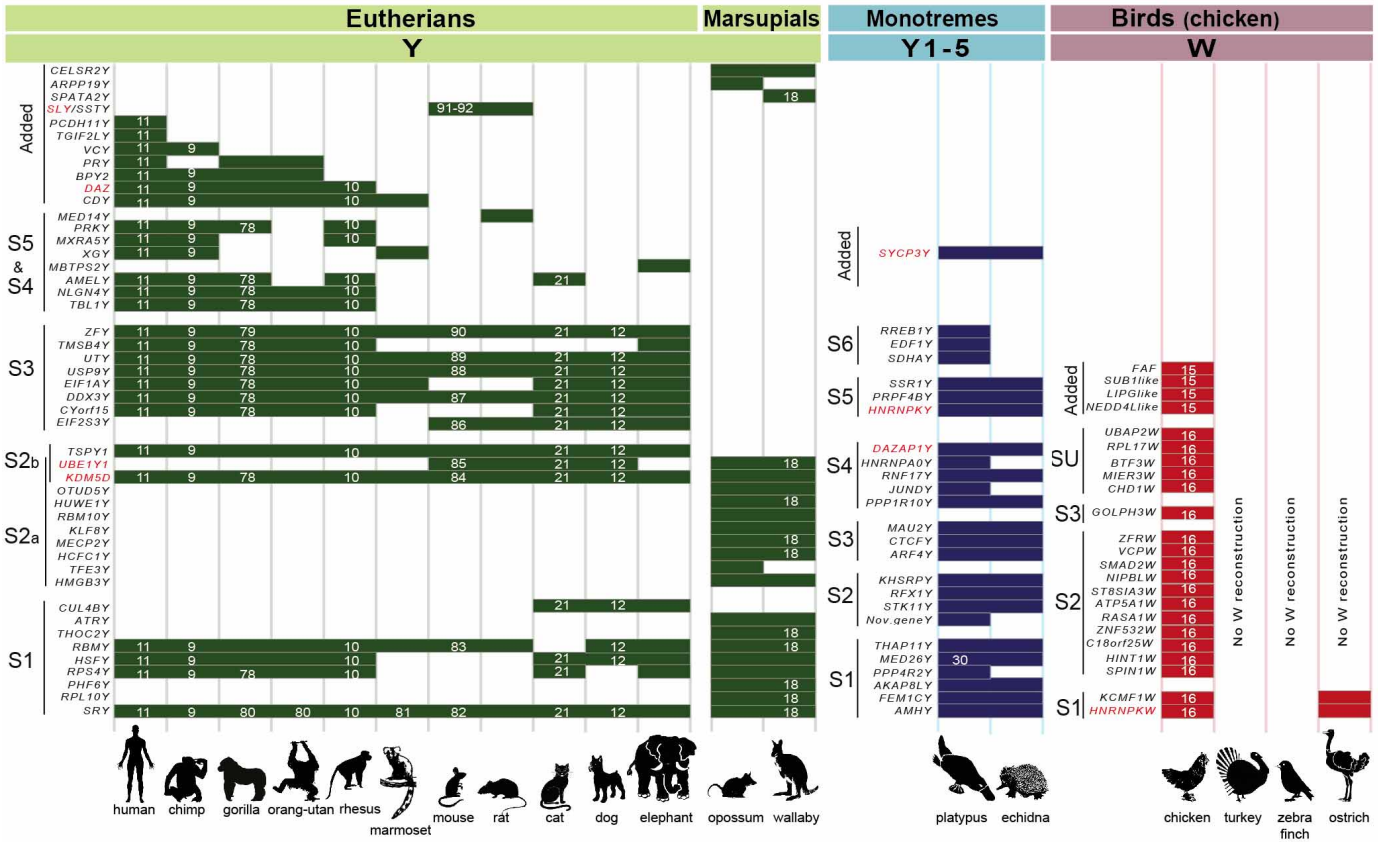
49. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–640 (2014).
50. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).
51. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).
52. Kircher, M., Stenzel, U. & Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* **10,** R83 (2009).
53. Chen, N., Bellott, D. W., Page, D. C. & Clark, A. G. Identification of avian W-linked contigs by short-read sequencing. *BMC Genomics* **13,** 183 (2012).
54. Carvalho, A. B., Dobo, B. A., Vibranovski, M. D. & Clark, A. G. Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **98,** 13225–13230 (2001).
55. Carvalho, A. B. & Clark, A. G. Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res.* **23,** 1894–1907 (2013).
56. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
57. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40,** D84–D90 (2012).
58. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1,** 18 (2012).
59. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24,** 713–714 (2008).
60. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol.* **29,** 644–652 (2011).
61. Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
62. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
63. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
64. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9,** 868–877 (1999).
65. Church, D. M. *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol.* **7,** e1000112 (2009).
66. Untergasser, A. *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).
67. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl Acad. Sci. USA* **102,** 10557–10562 (2005).
68. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10,** 210 (2010).
69. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
70. Katsura, Y. & Satta, Y. No evidence for a second evolutionary stratum during the early evolution of mammalian sex chromosomes. *PLoS ONE* **7,** e45488 (2012).
71. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
72. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15,** 568–573 (1998).
73. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Rev. Genet.* **7,** 98–108 (2006).
74. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
75. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
76. Schriml, L. M. *et al.* Tyramide signal amplification (TSA)-FISH applied to mapping PCR-labeled probes less than 1 kb in size. *Biotechniques* **27,** 608–613 (1999).
77. Kortschak, R. D., Tsend-Ayush, E. & Grutzner, F. Analysis of SINE and LINE repeat content of Y chromosomes in the platypus, *Ornithorhynchus anatinus*. *Reprod. Fertil. Dev.* **21,** 964–975 (2009).
78. Goto, H., Peng, L. & Makova, K. D. Evolution of X-degenerate Y chromosome genes in greater apes: conservation of gene content in human and gorilla, but not chimpanzee. *J Mol Evol.* **68,** 134–144 (2009).
79. Kim, H. S. & Takenaka, O. Evolution of the X-linked zinc finger gene and the Y-linked zinc finger gene in primates. *Mol. Cells* **10,** 512–518 (2000).
80. Whitfield, L. S., Lovell-Badge, R. & Goodfellow, P. N. Rapid sequence evolution of the mammalian sex-determining gene SRY. *Nature* **364,** 713–715 (1993).
81. Moreira, M. A. SRY evolution in Cebidae (Platyrrhini: Primates). *J. Mol. Evol.* **55,** 92–103 (2002).
82. Gubbay, J. *et al.* A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature* **346,** 245–250 (1990).
83. Ma, K. *et al.* A Y chromosome gene family with RNA-binding protein homology: candidates for the azoospermia factor AZF controlling human spermatogenesis. *Cell* **75,** 1287–1295 (1993).
84. Agulnik, A. I. *et al.* A mouse Y chromosome gene encoded by a region essential for spermatogenesis and expression of male-specific minor histocompatibility antigens. *Hum. Mol. Genet.* **3,** 873–878 (1994).
85. Mitchell, M. J. *et al.* Homology of a candidate spermatogenic gene from the mouse Y chromosome to the ubiquitin-activating enzyme E1. *Nature* **354,** 483–486 (1991).
86. Ehrmann, I. E. *et al.* Characterization of genes encoding translation initiation factor from X-inactivation and evolution. *Hum. Mol. Genet.* **7,** 1725–1737 (1998).
87. Strausberg, R. L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99,** 16899–16903 (2002).
88. Hall, N. M. *et al. Usp9y* (ubiquitin-specific protease 9 gene on the Y) is associated with a functional promoter and encodes an intact open reading frame homologous to *Usp9x* that is under selective constraint. *Mamm. Genome* **14,** 437–447 (2003).
89. Mazeyrat, S. *et al.* The mouse Y chromosome interval necessary for spermatogonial proliferation is gene dense with syntenic homology to the human *AZFa* region. *Hum. Mol. Genet.* **7,** 1713–1724 (1998).
90. Mardon, G. & Page, D. C. The sex-determining region of the mouse Y chromosome encodes a protein with a highly acidic domain and 13 zinc fingers. *Cell* **56,** 765–770 (1989).
91. Touré, A. *et al.* Identification of novel Y chromosome encoded transcripts by testis transcriptome analysis of mice with deletions of the Y chromosome long arm. *Genome Biol.* **6,** R102 (2005).
92. Touré, A. *et al.* A protein encoded by a member of the multicopy *Ssty* gene family located on the long arm of the mouse Y chromosome is expressed during sperm development. *Genomics* **83,** 140–147 (2004).



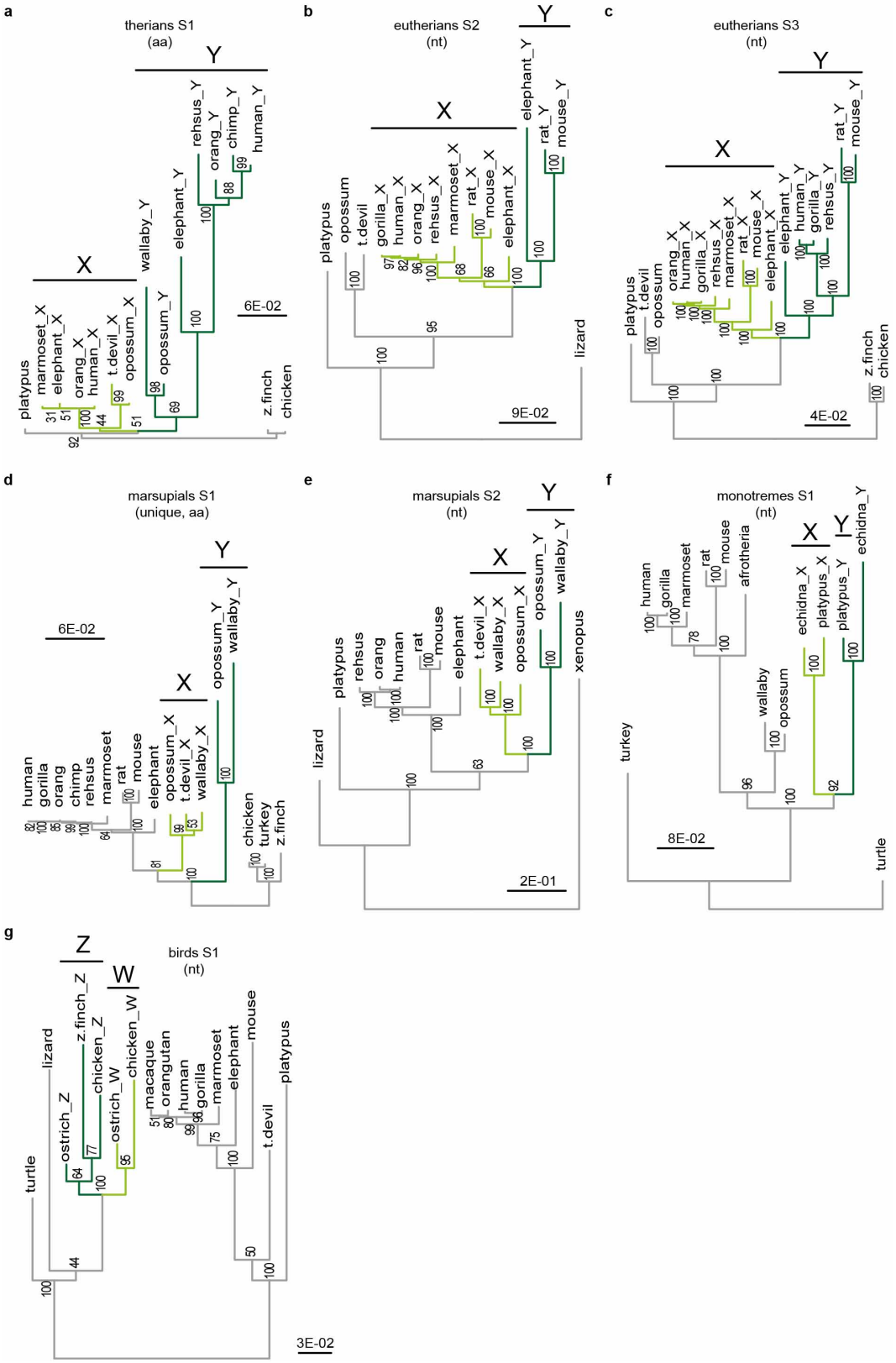
Extended Data Figure 1 | Overview of subtraction approach to detect and assemble Y (W) chromosome genes/transcripts.



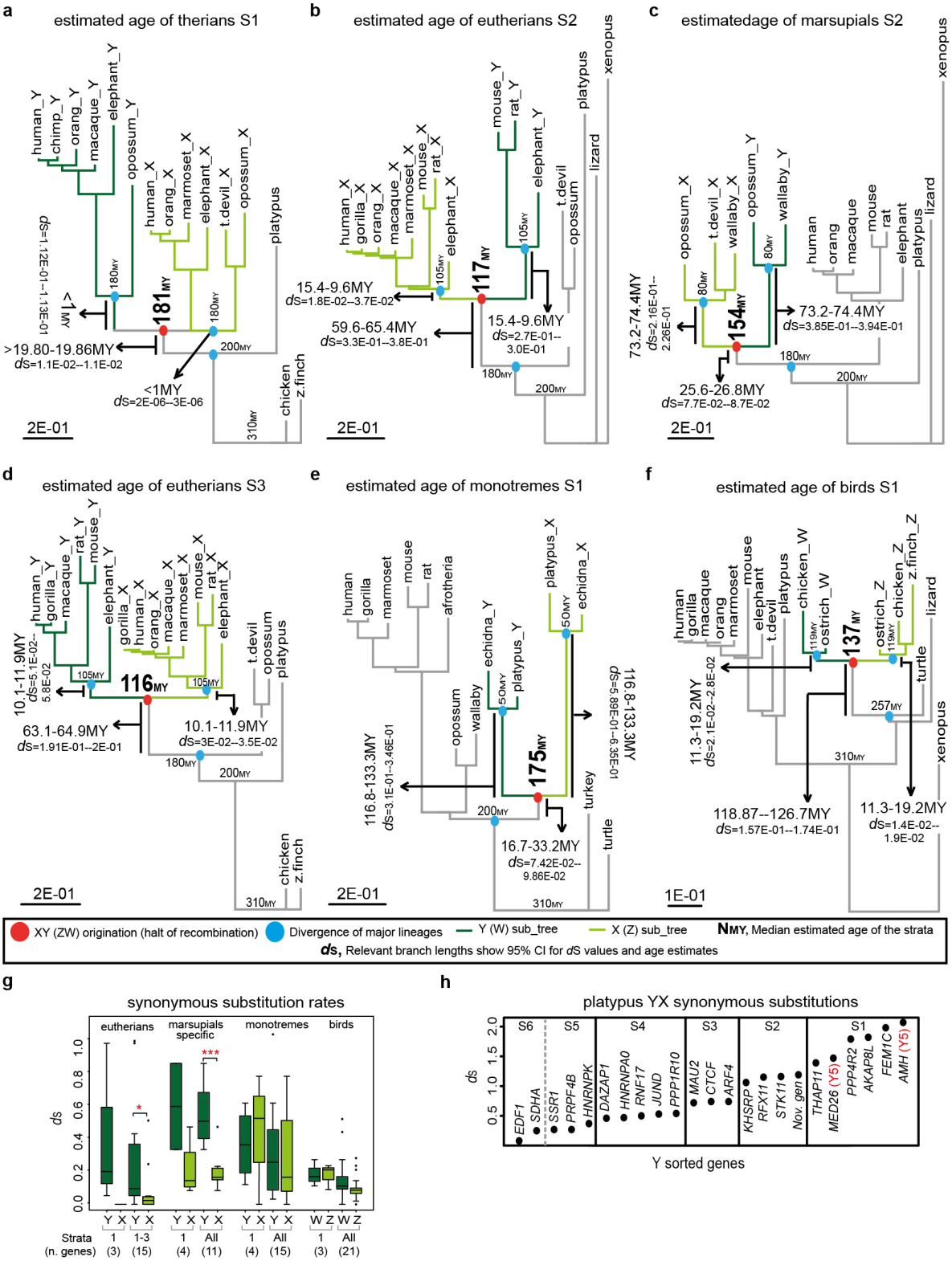
Extended Data Figure 2 | Resampling simulations to evaluate the performance of the Y transcript reconstruction approach. a–d, Results of one simulation showing the completeness (length of reconstructed sequences with respect to annotated sequences at ≥ 99% identity) of Y protein-coding transcripts using increasing numbers of randomly sampled male reads (million reads, Mr) when performing reconstructions. Simulations were carried out for human, chimpanzee, macaque and mouse where Y chromosomes have been fully sequenced. Twenty randomly selected genes from chicken were included in the simulations to control for false positive reconstructions (the median of the completeness of these control genes is zero). **e**, Distribution of the completeness of Y transcripts for repeated resampling analyses for different numbers of randomly sampled male reads (50 resampling replicates for each read number). Error bars, maximum and minimum values, excluding outliers.



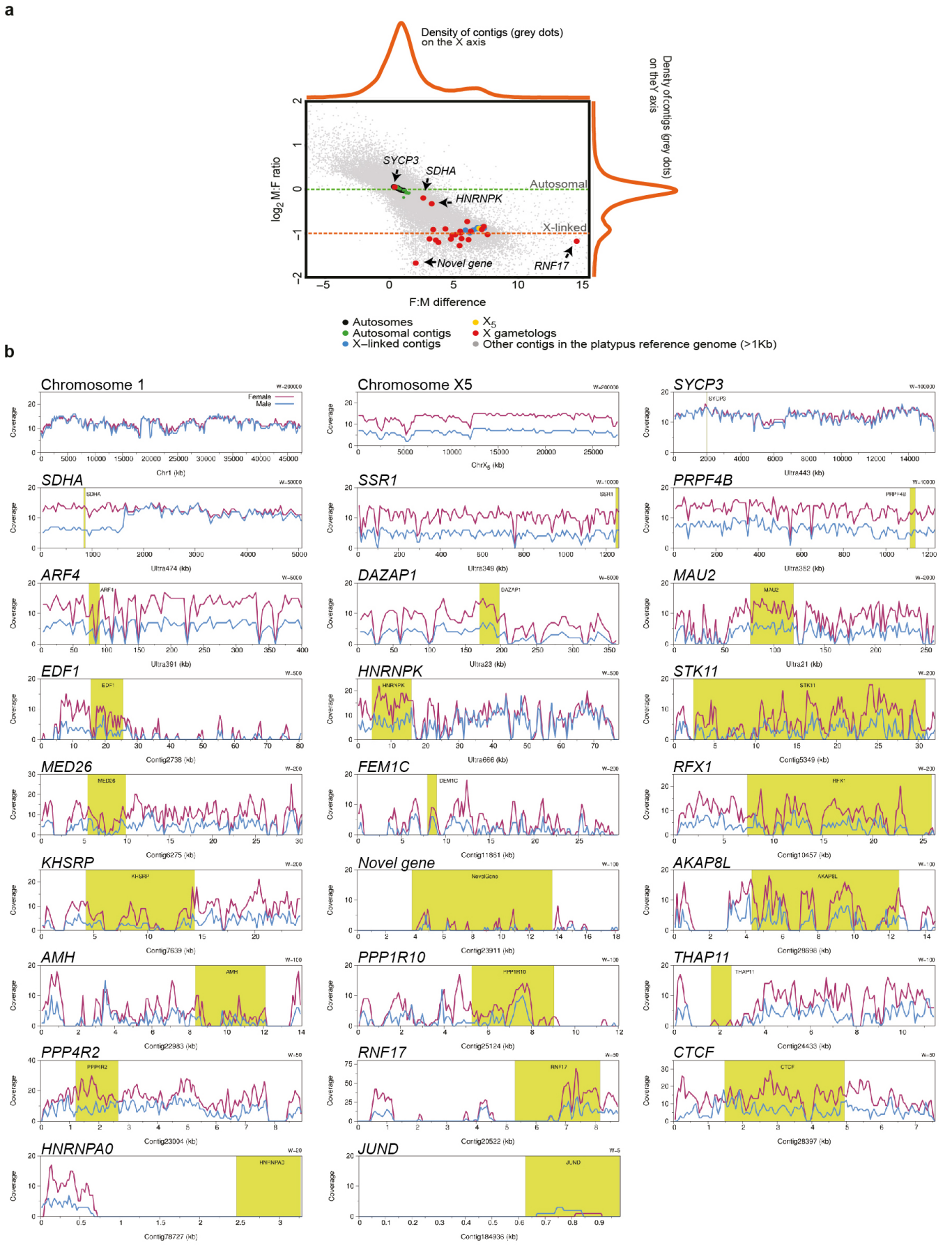
Extended Data Figure 3 | Y (W) genes previously described in the literature. See Fig. 1 for detailed legend. Numbers indicate the references in which Y and W genes are described.



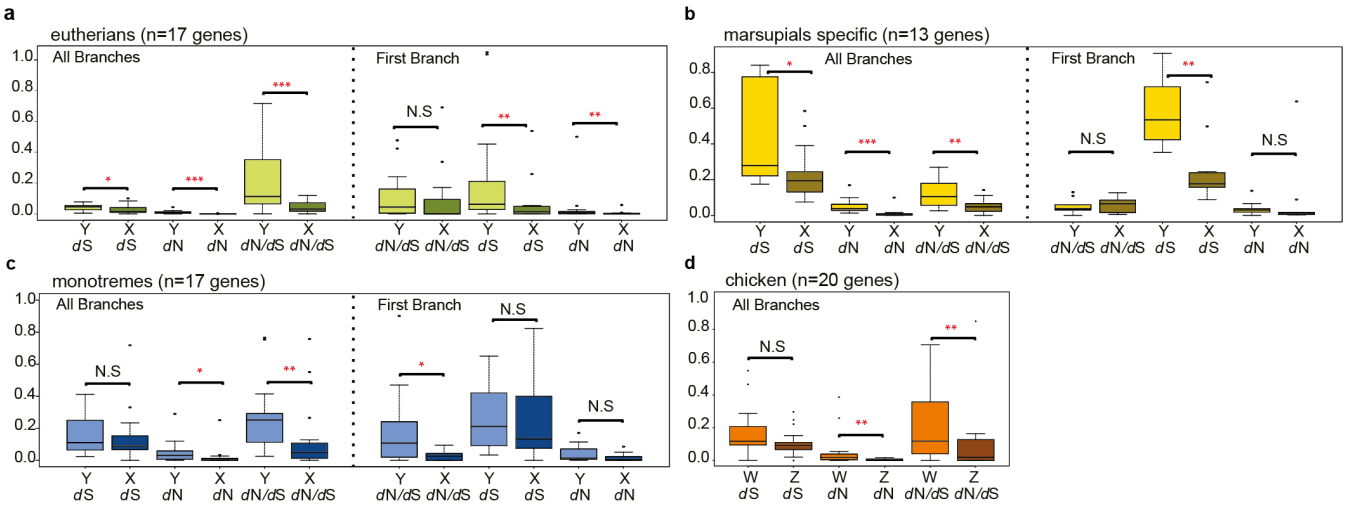
Extended Data Figure 4 | Phylogenetic trees constructed on the basis of alignments of concatenated genes for a given stratum. a–g, All trees are based on coding nucleotide sequences, except for therian S1 and marsupials S1 trees, which are based on an amino acid alignment, given that coding sequences are highly diverged, leading to a tree that is less well supported at key nodes. Bootstrap values are based on 100 bootstrap replicates. In the specific case of therian S1 and eutherian S3, we included all different gene copies (where these were available from previous work or could be reconstructed) in the analyses for the multi-copy genes (*RBMY*, *RPS4Y*, *HSFY*, *Cyorf15*) by repeatedly (100 times) and randomly selecting gene copies for the concatenations, and then reconstructing phylogenetic trees (100 bootstrap replicates). Reported bootstrap values at each node thus represent the median value for 100 trees obtained with different combinations of multi-copy sequences. Trees for individual genes are available at ftp://ftp.vital-it.ch/papers/kaessmann/Nature-Cortez/Cortez_etal_Nature_YX_gene_trees_alignments.zip.

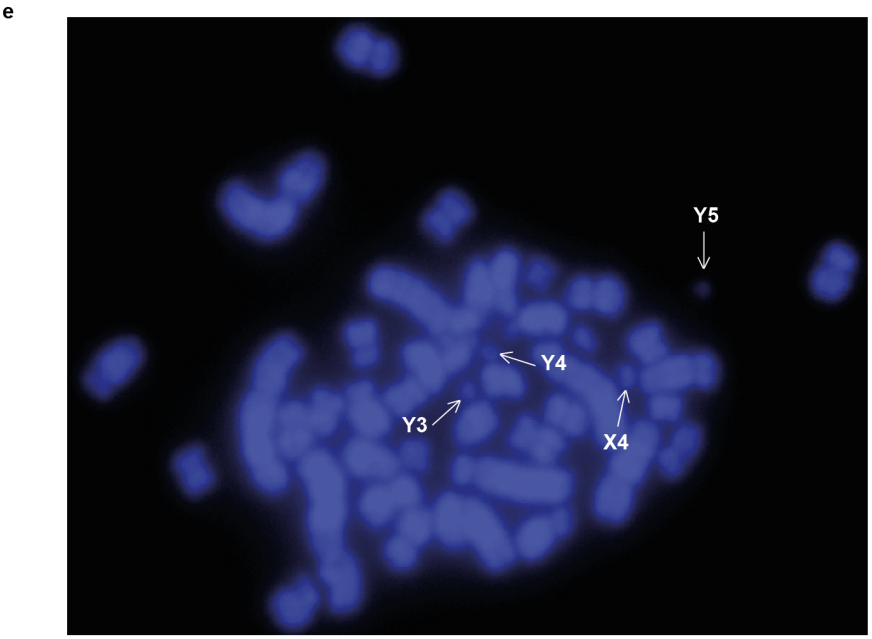


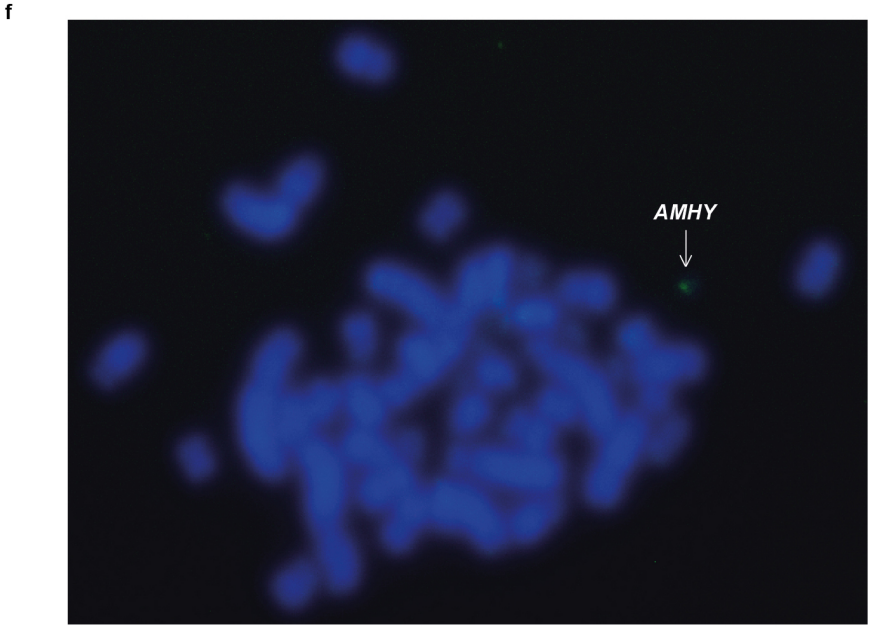
Extended Data Figure 5 | Strata ages and synonymous substitution rate evolution. **a**–**f**, Phylogenetic trees based on synonymous site divergences (d_S) of concatenated Y (W) genes for different strata; key branch lengths (d_S and inferred corresponding millions of years, Myr) as well as divergence times and inferred strata ages (in green) are indicated. Branch lengths and corresponding age ranges are given as 95% confident intervals calculated with simulated data (Methods). The following genes were used in underlying concatenated alignments (other genes were lacking orthologues in some species and were thus not included): *SRY*, *RBMY* and *RPS4Y* (total alignment length: 2,607 nucleotides (nt)) for therian S1; *KDM5D* (4,794 nt) for eutherian S2; *KDM5D*, *HCHC1Y*, *MECP2Y*, *UBE1Y1* and *HUWE1Y* (24,657 nt) for marsupial S2; *USP9Y*, *UTY* and *DDX3Y* (12,408 nt) for eutherian S3; *AMHY*, *FEM1CY*, *AKAP8LY* and *MED26Y* (5,469 nt) for monotreme S1; *HNRNPKW* and *KCMF1W* (2,268 nt) for bird S1. **g**, d_S values for the first (most basal) branches that follow different stratification events and lead to Y and X clades, respectively. Statistically significant differences (Mann–Whitney *U*-test): Benjamini–Hochberg-corrected $^*P < 0.05$, $^{***}P < 0.001$. Error bars, maximum and minimum values, excluding outliers. **h**, Pairwise d_S values for Y and X gametologues in platypus. Genes were grouped into strata on the basis of phylogenetic information and d_S value clustering (Methods).



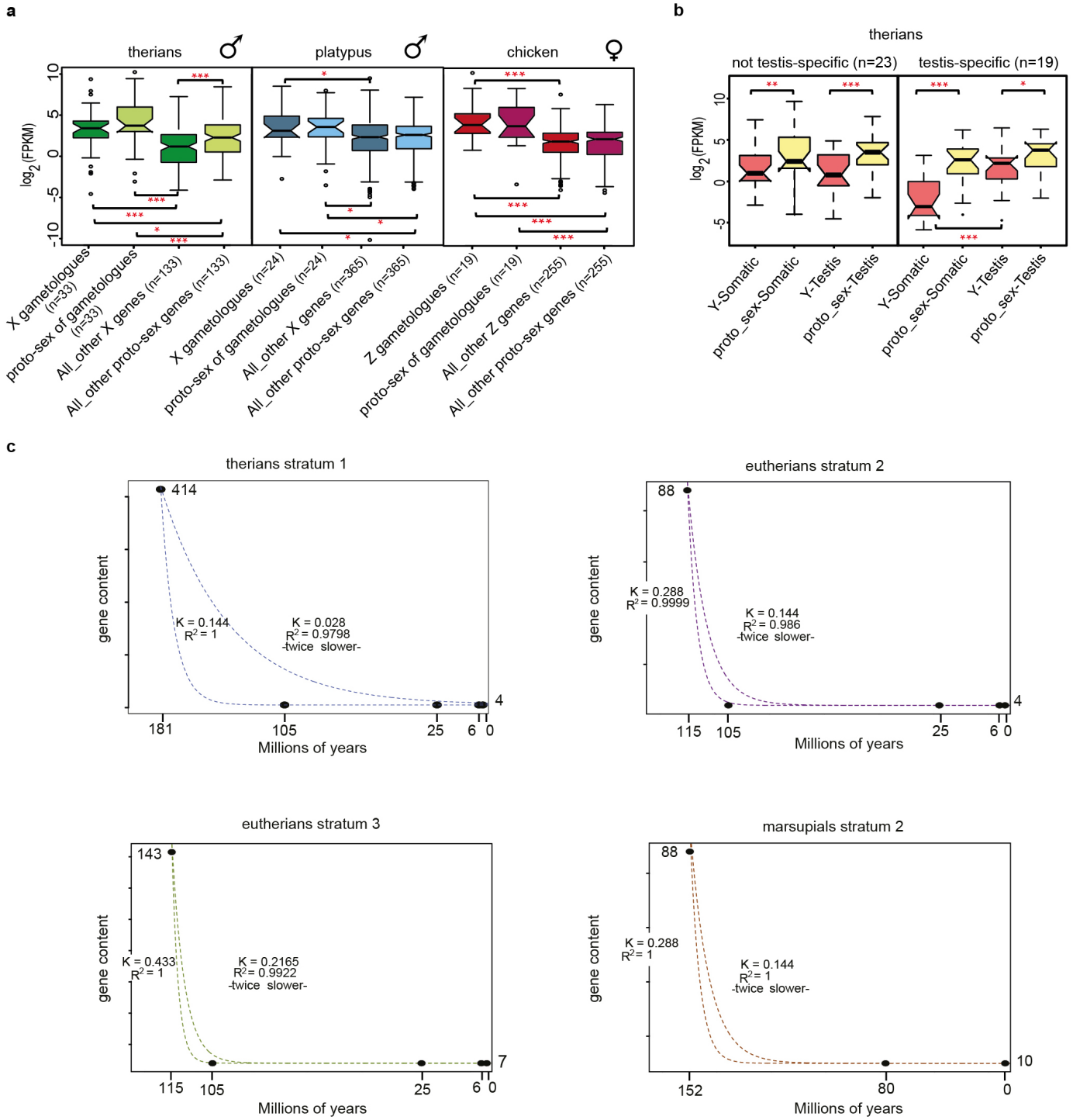
Extended Data Figure 6 | Prediction of X contigs using male/female genomic read coverage analyses. **a**, Male (M) to female (F) genomic read coverage ratio/difference for 23 unassembled contigs (red dots) containing the closest homologues of all platypus Y genes except for *RREB1Y* (no homologue could be traced) and *JUND* (low genomic read coverage), fully differentiated part of X_5 (yellow), assembled autosomes (black), and autosomal contigs (green), previously experimentally mapped X-linked contigs (blue), and all other unassembled contigs (grey). The density plots describe the distribution of unassembled contigs; the major peak in the *y*-axis distribution reflects similar coverage in males and females, as expected for autosomal contigs and observed for assembled autosomes. The minor peak reflects a twofold higher coverage in females, as expected for X-linked sequences and observed for the assembled X_5 chromosome. **b**, Coverage profiles for chromosome 1, X_5, and the 24 contigs containing gametologues of reconstructed Y genes (*RREB1Y* contig is missing, as no gametologue could be identified for this gene). Plots are sorted by contig sequence lengths. Consistent with the expectation for X gametologues, 19 out of 24 contigs show an overall twofold higher coverage in females (M:F $\log_2$ ratio significantly different from 0 but not from -1; one-sample Wilcoxon signed-rank test; Benjamini–Hochberg-corrected, $P < 0.05$; Methods and Supplementary Tables 19, 20). The contig containing the 'Novel Gene', which has a particularly low M:F ratio, has low coverage but shows twofold higher coverage in females compared to males in regions with mapped reads, consistent with X-linkage. Contigs containing *SDHAY* and *HNRNPKY* homologues show a twofold higher coverage in females in the region containing the genes, suggestive of a location in a non-recombinant X region (these genes are likely located close to pseudoautosomal boundaries). The closest homologue of *SYCP3Y* shows a profile typical of autosomes, which indicates that *SYCP3Y* was recruited directly to the Y from an autosome. *JUND* cannot be analysed owing to the limited genomic read coverage.



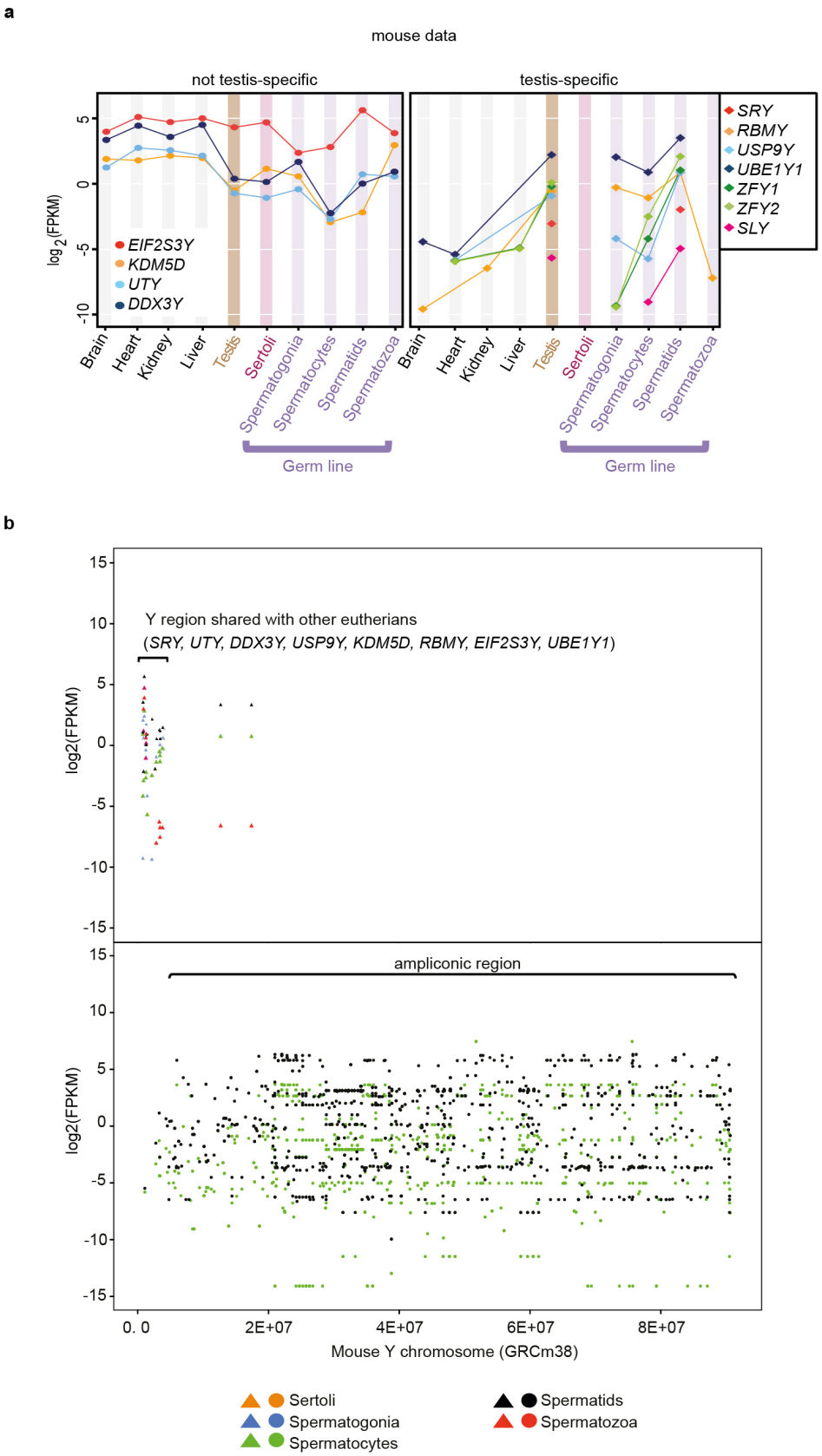




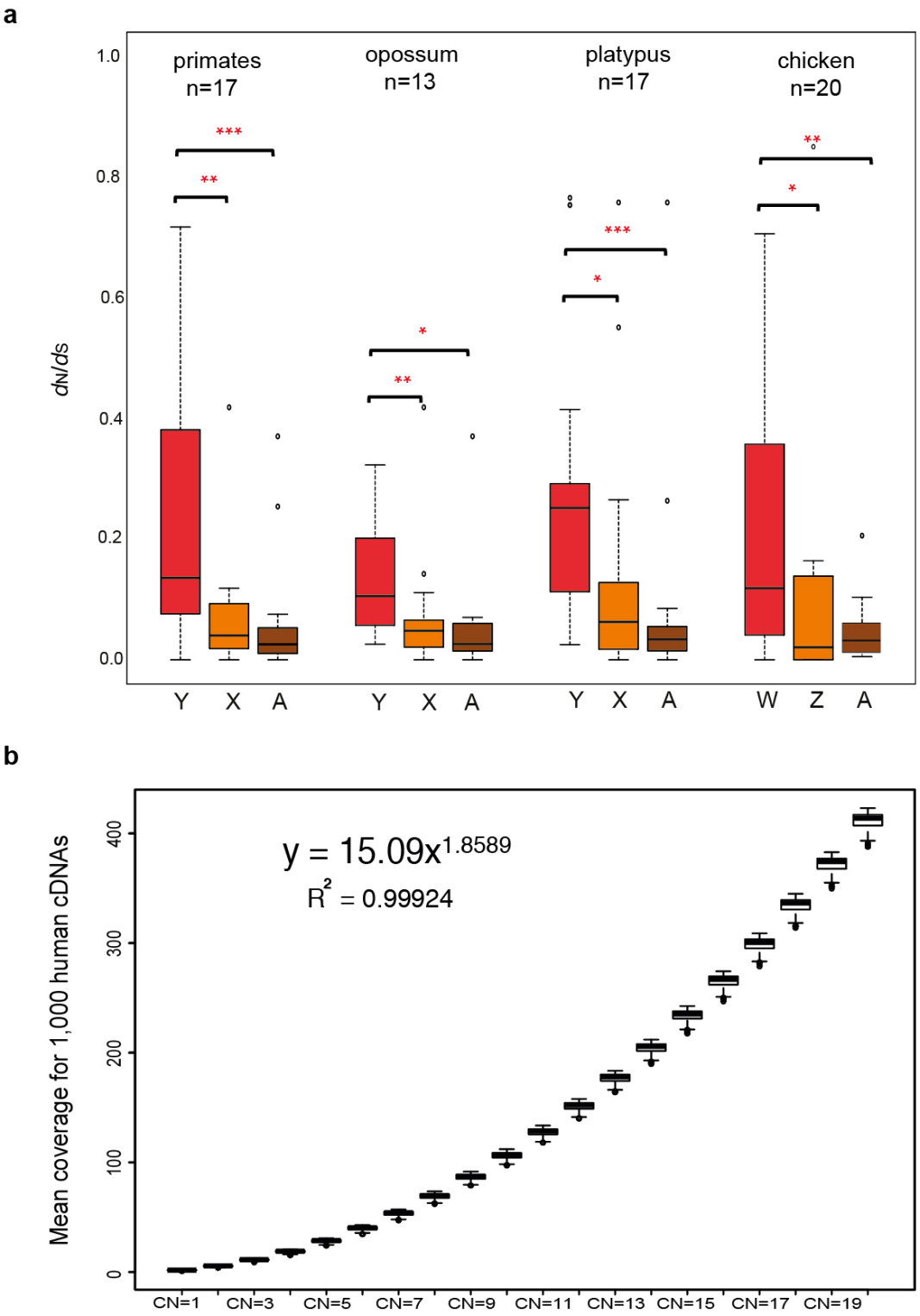
Extended Data Figure 7 | Synonymous and nonsynonymous substitution rates across Y/X (W/Z) branches and physical mapping of *AMHY*. **a**–**d**, d_S, d_N, d_N/d_S values for all Y/X branches (medians across branches for listed Y genes and their gametologues), or the first (most basal) branches that follow different stratification events and lead to Y and X clades, respectively. The Mann–Whitney *U*-test was used to compare distributions of Y and X rates. Significant (Benjamini–Hochberg-corrected) *P* values are shown in red. **a**, Eutherians, 17 Y genes (*SRY*/*SOX3*, *RBMY*/*X*, *RPS4Y1*/*X*, *AMELY*/*X*, *DDX3Y*/*X*, *EIF1AY*/*X*, *PRKY*/*X*, *TSPY1*/*X*, *USP9Y*/*X*, *ZFY*/*X*, *UTY*/*X*, *EIF2S3Y*/*X*, *NLGN4Y*/*X*, *KDM5D*/*C*, *UBE1Y1*/*UBA*, *TMSB4Y*/*X* and *Cyorf15*/*TBL1X*). **b**, Marsupials, 13 genes (*ATRY*/*X*, *HCFC1Y*/*X*, *MECP2Y*/*X*, *HUWE1Y*/*X*, *RBM10Y*/*X*, *RPL10Y*/*X*, *TFE3Y*/*X*, *THOC2Y*/*X*, *KLF8Y*/*X*, *HMGB3Y*/*X*, *PHF6Y*/*X*, *KDM5D*/*C* and *UBE1Y1*/*UBA*). **c**, Monotremes, 17 genes (*AMHY*/*X*, *FEM1CY*/*X*, *AKAP8LY*/*X*, *PPP4R2Y*/*X*, *MED26Y*/*X*, *THAP11Y*/*X*, *STK11Y*/*X*, *RFX1Y*/*X*, *KHSRPY*/*X*, *ARF4Y*/*X*, *CTCFY*/*X*, *MAU2Y*/*X*, *PPP1R10Y*/*X*, *HNRNPKY*/*X*, *DAZAP1Y*/*X*, *PRPF4BY*/*X* and *SSR1Y*/*X*). **d**, Chicken, 20 genes (*NIPBLW*/*Z*, *ATP5A1W*/*Z*, *BTF3W*/*Z*, *C18orf25W*/*Z*, *CHD1W*/*Z*, *GOLPH3W*/*Z*, *HINT1W*/*Z*, *KCMF1W*/*Z*, *MIER3W*/*Z*, *NEDD4LW*/*Z*, *RASA1W*/*Z*, *RPL17W*/*Z*, *SMAD2W*/*Z*, *SPIN1W*/*Z*, *ST8SIA3W*/*Z*, *UBAP2W*/*Z*, *VCPW*/*Z*, *ZFRW*/*Z*, *ZNF532W*/*Z* and *HNRNPKW*/*Z*). **e**, DAPI (4′,6-diamidino-2-phenylindole) staining of male platypus metaphase chromosomes. Chromosome Y_5 is minute and the smallest chromosome in platypus. The next smallest chromosomes (Y3, Y4 and X4) are also indicated, for comparison. **f**, Localization of *AMHY* gene using FISH. The *AMHY* PCR probe (green signal) hybridized specifically to chromosome Y_5. Error bars, maximum and minimum values, excluding outliers.



Extended Data Figure 8 | Expression levels of X gametologues and proto-sex chromosome genes, expression evolution towards testis-specificity, and kinetics of Y gene loss. **a**, Expression level characteristics (somatic tissues) of gametologues on current X and proto-sex chromosomes. Expression level distributions for current X gametologues (X), precursors of current X/Y genes on proto-sex chromosomes as inferred from 1:1 autosomal orthologues in outgroup species, all current X-linked genes, all proto-sex chromosomal genes. Similar distributions for chicken Z-linked genes and proto-sex chromosome precursors. Statistically significant differences (Mann–Whitney *U*-test): Benjamini–Hochberg-corrected $*P < 0.05$, $***P < 0.001$. **b**, Current (Y) and inferred ancestral (proto-sex) expression levels of Y genes that gained testis-specific expression during evolution (right) and those that did not (left) in somatic tissues and testis (as indicated on *x* axis). Note: for proto-sex chromosome plots, inferred expression output values were calculated per single gene copy (see Fig. 2 legend). **c**, Kinetics of ancestral gene decay on the therian Y. Gene numbers are plotted on the *y* axis (ancestral gene numbers are indicated at the top), and time (in Myr) is plotted on the *x* axis. Dots indicate minimum inferred or observed gene numbers at different time points of mammalian evolution based on our d_S inferences (see Fig. 1 and Extended Data Fig. 5) and known lineage divergence times: 181 Myr (S1 formation), 152 Myr (S2a formation), 115 Myr (S2b and S3), 105 Myr (afrotherian split from other placentals), 80 Myr (American–Australian marsupial split), 25 Myr (Old-World monkey–ape split) and 6 Myr (human–chimp split). Dashed lines represent best-fit curves to data points using each of the decay models as indicated. Ancestral gene numbers and decay rates (K, Myr $-$ 1) were taken from Hughes *et al.*[10] except where alternative rates are indicated. Error bars, maximum and minimum values, excluding outliers.



Extended Data Figure 9 | Spermatogenic expression patterns of Y protein-coding genes and along the mouse Y chromosome. **a**, Left, expression levels of 4 mouse Y genes with ubiquitous spatial expression profiles across different organs and individual spermatogenic cell types. Right, expression levels of 7 mouse Y genes with testis-specific expression across organs and cell types. **b**, Top, spermatogenic expression of protein-coding genes located in the Y-conserved region (YCR). Lower panel: Expression of 31 reconstructed Y-linked transcript contigs (including unknown presumably noncoding sequences, the known protein-coding genes *SLY* and *SSTY*; all with more than 10 copies) along the ampliconic region of the mouse Y chromosome. Genes and contigs were mapped to 1,452 positions using BLASTn to the assembled mouse Y chromosome from genome version GRCm38 (ref. 65).



Extended Data Figure 10 | Nonsynonymous over synonymous substitution rates across Y/X (W/Z) branches and simulated multi-copy genes. **a**, Median d_N/d_S values for all Y/X branches and branches leading to autosomal orthologues in outgroup species with different sex chromosome systems (A). Statistically significant differences (Mann–Whitney *U*-test): Benjamini–Hochberg-corrected $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **b**, Mean coverage values from 1,000 different complementary DNAs from human (>1 kb <10 kb) that were introduced to a mock genome having different copy-numbers (CN), that is, from 1 copy to 20 identical copies. Error bars, maximum and minimum values, excluding outliers.