Strong male bias drives germ line mutation in chimpanzees

Oliver Venn[1], Isaac Turner[1], Iain Mathieson[1,3], Natasja de Groot[2], Ronald Bontrop[2], Gil McVean[1*]

[1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

[2]Biomedical Primate Research Centre, Lange Kleiweg 161, 2288 GJ Rijswijk, The Netherlands

[3]Current address: Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115 USA

*Corresponding author: mcvean@well.ox.ac.uk

Germ line mutation determines rates of molecular evolution, genetic diversity and fitness load. In humans, the average point mutation rate is 1.2 x 10^{-8} per base pair per generation, with every additional year of father's age contributing 2 mutations across the genome and males contributing 3-4 times more mutation than females. To assess whether such patterns are shared with our closest living relatives we sequenced the genomes of a nine-member pedigree of Western chimpanzees, *Pan troglodytes verus*. Our results indicate a mutation rate of 1.2 x 10^{-8} per base pair per generation, but a male contribution 7-8 times that of females and a paternal age effect of 3 mutations per year of father's age. Thus mutation rates and patterns differ between closely related species.

Accurate determination of the rate of de novo mutation in the germ line of a species is central to the dating of evolutionary events. However, because mutations are rare events, efforts to measure the rate in humans have typically been indirect, calculated from the incidence of genetic disease or sequence divergence (*1-4*). However, high throughput sequencing technologies have enabled direct estimates of the mutation rate from comparison of the genome sequence of family members (*5-8*). Unexpectedly, these studies have indicated a mutation rate of, on average, approximately (c.) 1.2x10^{-8} per base pair per generation, or c. 0.5 x 10^{-9} per base pair per year, approximately half that inferred from phylogenetic approaches (*1, 9*). Moreover, they have demonstrated a substantial male-bias to mutation, such that 3-4 times as many autosomal mutations occur in the male compared to the female germ line (*6, 7*). Male-bias is largely caused by an increase in the rate of paternal but not maternal mutation with the age of the parent; approximately 2 additional mutations per year of father's age at conception (*7*). This difference is consistent with ongoing cell division in the male germ line but not in females (*10*).

An alternative approach for estimating the extent of male bias is to compare rates of sequence divergence on the autosomes (which spend equal time in the male and female germ lines) and the X chromosome (which spends 2/3 of the time in females) (*2, 11*). Such indirect approaches broadly agree with direct estimates in humans, but suggest that male bias may be stronger in chimpanzees (*12*). To test this hypothesis we sequenced the genomes of nine members of a three-generation pedigree of Western chimpanzees, *Pan troglodytes verus* (Figs 1A, S1). One trio was sequenced at high depth (average 51x), while other family members were sequenced to an average of 27x (Table S1). We inferred the structure of recombination and transmission across the pedigree (Fig. 1B) which enabled us to detect de novo point mutations in regions of high sequence complexity and to remove artifacts caused by mis-mapping, sequence that is absent from the reference genome and reference mis-assembly (*13*).

We used a probabilistic approach which, at a given site, compared the likelihood of different models for genetic variation inconsistent with the inferred transmission: genotyping error at a segregating variant, de novo mutation, single gene conversion event, segregating deletion and erroneous call

(Fig. 1C). The design was expected to enable haplotype phasing through transmission for 99.2% of sites that were heterozygous in the founders and 87.5% of de novo mutation events inherited by chimpanzee F (Fig. 1A). Read-based phasing was used to phase de novo events in other offspring and we performed independent validation to assess the accuracy of de novo variant calls. The false negative rate was estimated from allele-dropping simulations (*13*) .

Across the genomes of the nine pedigree members we called 4.1 million variants (SNPs and short indels) using a mapping-based approach and 3.0 million variants using an assembly-based approach (*14*). Genotype data confirmed expected pedigree relationships (Fig. S2). The intersection of call sets (1.6 million sites with a transition-transversion ratio of 2.2) established the underlying structure of recombination and transmission across the pedigree with a robust version of the Lander-Green algorithm (Fig. S3). Briefly, this is a two-stage strategy of identifying dominant inheritance vectors over 1 Mb intervals, followed by fine-mapping of cross-over breakpoints, which guards against problems caused by false positive variants and genotyping errors (*13*). Across the pedigree, we identified 375 cross-over events, with a distribution similar to human homologues, with the exception of human chromosome 2, which is a fusion of the chimpanzee chromosomes 2A and 2B (*15*) (Figs 2A, S4; Tables S2, S3).

Overall, we estimate the sex-averaged autosomal genetic map length to be 3,150 cM (95% Equal Tailed Probability Interval, ETPI, 2,850 – 3,490), compared to 3,505 cM in humans (*16*, *17*). On the X chromosome, we detected 9 cross-over events in the non-pseudoautosomal (non-PAR) region, indicating a female-specific genetic map length of 160 cM (95% ETPI 83 – 300), compared to 180 cM in humans. On the pseudoautosomal region (PAR) we detect four male cross-overs, giving a male specific estimate of 34 cM (95% ETPI 28 – 180; Tables S4, S5) in agreement with estimates in human (*13*). Males have 58% of the autosomal cross-over events of females and, unlike females, show an increase in cross-over frequency towards the telomere (Fig. 2B), similar to humans (Fig. S5). We also observed a decrease in cross-over frequency with maternal (2.65 cM per year, linear model P = 0.025), but not paternal age (Fig. 2C). However this observation could be explained by between-female variation (linear model P = 0.13 allowing for a maternal effect). The median interval size to which cross-over events can be localized is 7.0 kb, with 95% of all intervals localized to within 80 kb (excluding complex cross-over events), with cross-over events enriched in regions inferred to have high rates of recombination from patterns of linkage disequilibrium (*18*) (Fig. S6).

Conditional on the inferred transmission, we used a probabilistic approach to identify candidate de novo mutations among all variants called by the mapping approach, incorporating uncertainty in the inferred genotype through the use of genotype likelihoods (*13*). Across the pedigree, we identified 204 autosomal de novo mutations (2 of which are multi-nucleotide variants) which pass thresholds for evidence (Fig. S7), purity and consistency (Fig. S8, Table S6).

Several lines of evidence indicate a low false-positive rate. First, none of these sites were called as variants in the genomes of 10 unrelated chimpanzees from the same sub-species (*18*). Second, the transition-transverison ratio of the candidate de novo events (2.16) is comparable to that for segregating variants. Third, the transmission of candidate de novo events in chimpanzee F to her offspring is consistent with expectations. Finally, we used a genotyping platform to validate all de novo events identified in chimpanzees F and I. Of the 61 sites with valid assays (18 failed design), 1 is a false positive, indicating a false discovery rate of c. 2% (Table S7). To estimate false negative rates, we used allele-dropping simulations, with empirical distributions of coverage and allele-balance. Within the F1 generation, the false negative rate is estimated to be 3.4%. However, the F2 generation has a higher rate (23%), arising from lower coverage (25.6x) in founder chimpanzee C. False negative rates were used to correct subsequent regression analyses (Table S8). On the X chromosome, we identified 3 de novo point mutations.

As expected (*1*), a high fraction of de novo mutations are C>T transitions at CpG sites (24% of all point mutations, compared to 17% in humans (*7*); likelihood ratio test, LRT, P = 0.03) and even after accounting for such mutations we see a trend towards AT bases (73 G/C>A/T, 55 A/T>G/C; ratio 1.3:1; LRT P = 0.11; Fig. 3A). We also found that point mutations tend to cluster within individuals at nearby locations, similar to observations in humans (*8*). For example, 17% of all point mutations are within 1 Mb of at least one other mutation event in the pedigree and in 41% of such cases, these all occur in the same individual (compared to an expectation of 13%; permutation P = 0.001). Importantly, we validated all variants in clusters of 1Mb or less in E and F, indicating that these are not false positives. The excess of within-sample clustering extends up to c. 200kb (Fig. S9A) and does not correspond to a single mutation type (e.g. CpG mutation). Moreover, the effect remains after increasingly stringent filters for specificity are applied (Fig. S9B) (*13*). The finding of clustered point mutation events, which may potentially arise from correlated exposure to mutagens or variation in the efficacy of DNA repair, implies non-independence in the way novel variation enters a species and has consequences for interpretation of patterns of genetic variation. We do not observe enrichment around genes, repeat elements, or gaps in the assembly (Fig. S10).

To assess whether the rate of de novo mutation is affected by parental age, we used Poisson regression, allowing for family effects and separate linear relationships between age and mutation rate for males and females. Although the sample size is small, we find no evidence for either familial or maternal age effects (linear model P > 0.05) but evidence for a paternal age effect (linear model P = 0.006), and consistency in effect on repeat and non-repeat DNA backgrounds (Fig. S11). Although we cannot formally exclude the possibility of familial effects (*6*), our results are consistent with observations (*7*) that paternal age explains nearly all variation in mutation rate in humans. Bayesian linear regression, allowing for a paternal effect only and accommodating variation in false negative rate indicates a posterior mean paternal age of effect of 3.02 mutations per year (Fig. 3B;

95% ETPI 1.35 – 4.68). In contrast, the paternal age effect in humans is estimated to be 1.95 mutations per year (re-analysis of data from ref. 6; 95% ETPI 1.65 – 2.26).

We ascertained the parent of origin of de novo mutations for chimpanzee F through transmission to the F2 generation, finding that 30 of the 35 autosomal mutations occurred in the paternal lineage. We also found that 25% of de novo events could be phased directly from read-pairs spanning the mutation and a nearby heterozygous site that could be phased through transmission. Across the pedigree, we assigned 31 paternal and 6 maternal autosomal mutations (Fig. 3C). Overall, we estimate the aggregate male to female mutation ratio, $\alpha = 5.5$ (95% ETPI = 3.0 – 10). The point estimate is 40% higher than that reported for humans (7), though is close to estimates from chimpanzee-specific divergence rates on the X and autosomes (*12*). In contrast to indirect approaches (*3*), we find no evidence that different types of mutation have different values of α. For example, α at CpG sites is 5.3, compared to 5.6 at non-CpG sites. Combining data across all mutation types and using available parent of origin information in the Bayesian regression model indicates that, on average, mothers contribute 6.7 de novo mutations (95% ETPI = 3.5 – 10.3) and each additional year of paternal age generates 3.0 mutations (95% EPTI = 1.2 – 4.4; Fig. S12), with the onset of mutation occurring at 8.1 years of age (95% ETPI = 0 – 12 years), consistent with the onset of spermarche in chimpanzees of 7.5 years (*19*).

Within the pedigree studied, the average number of autosomal de novo mutations occurring each generation is 35, lower than current estimates for humans of 74.4 (*7*, *9*). However, the parental ages in the Western chimpanzee pedigree (averages of 18.9 for males and 15.0 for females) are lower than estimates of parental ages in the wild (24.3 for males and 26.3 for females), which are lower than estimates for humans (31.5 for males and 25.6 for females) (*20*). Using the fitted model for mutation rates, we predict that the average number of autosomal de novo mutations per offspring in the wild should be c. 69. We estimate the length of the autosomal genome accessible in our study to be 2,360 Mb across the autosomes (Table S9), indicating a mutation rate of 1.2×10^{-8} per base pair per generation and α to be 7.8 (Table S10).

Under a model in which the mutation rate increases linearly with parental age, the rate of neutral substitution is the ratio of the average number of mutations inherited per generation to the average parental age. We predict the neutral substitution rates to be c. 0.46×10^{-9} bp^{-1} year^{-1} in chimpanzees, compared to estimates in humans of c. 0.51×10^{-9} bp^{-1} year^{-1} (*9*). These results are consistent with near-identical levels of lineage-specific sequence divergence (*12*) but surprising given the differences in paternal age effect. In the intersection of the autosomal genome accessible in the current study and regions where human and chimpanzee genomes can be aligned with high confidence, the rate is slightly lower (0.45×10^{-9} bp^{-1} year^{-1}) and the level of divergence is 1.2% (*13*),

implying an average time to the most common ancestor (TMRCA) of 13 million years (95% ETPI 11 – 17 million years; Table S11).

Increased male bias can explain low levels of diversity on the chimpanzee X chromosome (*21*, *22*). Taking into account differences in generation time and effective population size, we predict that X chromosome diversity should be 56% that of autosomes (assuming equal and constant effective population sizes for males and females; Table S10) comparable to empirical estimates (*21*, *22*). Similarly, our results predict that the X chromosome rate of divergence is lower in chimpanzees than humans (74% of the autosomal rate in chimpanzees, 85% of the autosomal rate in humans). Previous explanations for unusual patterns of X chromosome diversity and divergence include a complex speciation event (*23*), extensive natural selection on the X chromosome (*22*), or, as supported by this study, by a greater male mutational bias in chimpanzees (*12*). This is likely related to differences in mating system between the species with chimpanzees showing higher levels of sperm competition through multiple mating and a higher relative testes mass than humans (0.27% of average adult male weight versus 0.079%) and higher levels of sperm production (*24*, *25*). We note that if differences in male mutational bias are to explain observed patterns of divergence, then gorillas would have a male mutational bias lower than humans arising from decreased sperm competition (*12*). Our results suggest that variation in mating patterns between species can impact the sex-bias of mutation and motivate the wider study of mutation rates and relationship to parental age across species.

Acknowledgements

Funded by Wellcome Trust grants 086786/Z/08/Z to OV and 090532/Z/09/Z to the WTCHG and MRC hub grant G0900747 91070. We thank M. Przeworski and D. Reich for discussion and comments on the manuscript, A. Kong for providing data on request from Ref. 6. Samples were provided through the Transnational Access Activity of the EUPRIM-Net under CITES authorization. Read-level data is accessible under SRA Study Accession Number PRJEB5937 from http://www.ebi.ac.uk/ena/data/view/PRJEB5937. All other project data are available from ftp://birch.well.ox.ac.uk.

References

1. M. W. Nachman, S. L. Crowell, Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297-304 (2000).
2. K. D. Makova, W. H. Li, Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624-626 (2002).
3. J. Taylor, S. Tyekucheva, M. Zody, F. Chiaromonte, K. D. Makova, Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol. Biol. Evol.* **23**, 565-573 (2006).
4. F. A. Kondrashov, A. S. Kondrashov, Measurements of spontaneous rates of mutations in the recent past and the near future. *Phil. Trans. R. S. B: Biological sciences* **365**, 1169-1176 (2010).
5. J. C. Roach *et al.*, Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-639 (2010).
6. D. F. Conrad *et al.*, Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712-714 (2011).
7. A. Kong *et al.*, Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).
8. J. J. Michaelson *et al.*, Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-1442 (2012).
9. A. Scally, R. Durbin, Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **13**, 745-753 (2012).
10. J. F. Crow, The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics* **1**, 40-47 (2000).
11. T. Miyata, H. Hayashida, K. Kuma, K. Mitsuyasu, T. Yasunaga, Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 863-867 (1987).
12. D. C. Presgraves, S. V. Yi, Doubts about complex speciation between humans and chimpanzees. *Trends in Ecol. Evol.* **24**, 533-540 (2009).
13. Supplementary Information.
14. Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, G. McVean, De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226-232 (2012).
15. J. W. IJdo, A. Baldini, D. C. Ward, S. T. Reeders, R. A. Wells, Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 9051-9055 (1991).
16. A. Kong *et al.*, Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-1103 (2010).
17. A. Kong *et al.*, A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241-247 (2002).
18. A. Auton *et al.*, A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**, 193-198 (2012).
19. J. Marson, S. Meuris, R. W. Cooper, P. Jouannet, Puberty in the male chimpanzee: progressive maturation of semen characteristics. *Biol. Reprod.* **44**, 448-455 (1991).
20. K. E. Langergraber *et al.*, Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 15716-15721 (2012).
21. G. H. Perry, J. C. Marioni, P. Melsted, Y. Gilad, Genomic-scale capture and sequencing of endogenous DNA from feces. *Mol. Ecol.* **19**, 5332-5344 (2010).
22. C. Hvilsom *et al.*, Extensive X-linked adaptive evolution in central chimpanzees. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 2054-2059 (2012).
23. N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103-1108 (2006).
24. R. V. Short, Sexual selection and its component parts, somatic and genital selection, as illustrated by man and the great apes. *Adv. Study. Behav.* **9**, 131-158 (1979).
25. A. P. Moller, Ejaculate quality, testes size and sperm competition in primates. *J. Hum. Evol.* **17**, 479-488 (1988).
26. Chimpanzee Sequencing Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).

27. G. Lunter, M. Goodson, Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936-939 (2011).
28. A. Rimmer *et al.*, Integrating mapping, assembly and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, (Accepted).
29. L. J. Peacock, C. M. Rogers, Gestation period and twinning in chimpanzees. *Science* **129**, 959 (1959).
30. S. B. Montgomery *et al.*, The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749-761 (2013).
31. Z. Iqbal, I. Turner, G. McVean, High-throughput microbial population genomics using the Cortex variation assembler. *Bioinformatics* **29**, 275-276 (2013).
32. E. S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 2363-2367 (1987).
33. G. R. Abecasis, S. S. Cherny, W. O. Cookson, L. R. Cardon, Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97-101 (2002).
34. A. G. Hinch *et al.*, The landscape of recombination in African Americans. *Nature* **476**, 170-175 (2011).
35. D. Kosambi, The estimation of map distances from recombination values. *Annals of Eugenics*, 3 (1944).
36. D. Karolchik *et al.*, The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, (2013).
37. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).
38. A. Henke, C. Fischer, G. A. Rappold, Genetic map of the human pseudoautosomal region reveals a high rate of recombination in female meiosis at the Xp telomere. *Genomics* **18**, 478-485 (1993).
39. D. C. Page *et al.*, Linkage, physical mapping, and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. *Genomics* **1**, 243-256 (1987).
40. F. Rouyer *et al.*, A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. *Nature* **319**, 291-295 (1986).

Figure legends

Figure 1. De novo mutations in a three-generation chimpanzee pedigree. A, Structure of the chimpanzee pedigree, indicating sequencing depth and date of birth (estimated for wild founders). Colors indicate expected inheritance proportions from founders. **B,** Visual representation of transmission across chromosome 21 following color assignment in part A; grandparental chromosomes distinguished by intensity. Red and black lines indicate position of female- and male-specific cross-over events respectively. **C,** Site classification at a candidate de novo mutation on chromosome 2A, showing, in the left-hand panel, the relative genotype likelihoods for each individual (as barplots), the resolved transmission structure (colored lines, following color assignment in part B) and, in the right-hand panel, the relative likelihoods for different models for the observed genetic variation.

Figure 2. Parameters of cross-over. A, Scatterplot showing the relationship between total genetic map length inferred for chromosomes and their homologues in humans. Black dotted line indicates the identity relationship; blue line indicates best linear fit (excluding chromosomes 2A and 2B). **B,** Relationship of recombination rate to relative telomere proximity demonstrates a telomeric effect in males, absent in females. **C,** Relationship between parental age and the number of cross-over events in mothers and fathers for the two family groups.

Figure 3. Characteristics of de novo mutations. A, Numbers of different types of de novo single base substitutions (multiple nucleotide events not shown). Panel below indicates relative abundance of each mutation type as a function of the adjacent nucleotides. **B,** Relationship between paternal age and total number of de novo mutations; grey points: observed counts, blue points: correction for false negative rate. The posterior mean and marginal 95% equal-tailed probability intervals for the Bayesian Poisson regression are shown as solid and dotted lines respectively. **C,** The number of de novo mutation events confirmed as paternal or maternal through either transmission to the F2 generation or from direct phasing within read-pairs. Offspring ordered left to right by increasing paternal age.

Strong male bias drives germ line mutation in chimpanzees

Oliver Venn[1], Isaac Turner[1], Iain Mathieson[1,3], Natasja de Groot[2], Ronald Bontrop[2], Gil McVean[1*]

Supplementary Information

Contents

1. Method summary

The genomes of a nine-member three-generation Western Chimpanzee pedigree were sequenced (100 bp paired-end reads with 450 bp average insert) from whole blood using Illumina HiSeq instruments (Illumina, CA) and mapped to the panTro3 reference (*26*). SNVs and small indels were called and filtered without incorporating pedigree information through both mapping- (*27*, *28*) and assembly-based (*14*) methods. Transmission vectors were inferred from the intersection of the call sets through a two-step process. First, a robust version of the Lander-Green algorithm was applied over 1Mb intervals, using the proportion of sites concordant with each transmission vector rather than likelihoods. Only single crossover events between intervals were permitted. Second, cross-over breakpoints were refined locally, using posterior decoding of the standard algorithm, incorporating uncertainty on the inferred genotype through genotype likelihoods. We defined breakpoints from the 95 percent credible interval for single event compatible transitions. Among all sites called by the mapping approach we identified positions where the reported genotypes were inconsistent with the inferred transmission vector. For such sites, we compared the likelihood of different models conditional on the transmission vector: segregating variant, de novo mutation, single gene conversion, segregating deletion, erroneous call. Candidates were then filtered on evidence (log-likelihood of de novo mutation model > 5 units higher than any other model), purity (no evidence for the variant allele in any other sample apart from offspring), and consistency (candidates segregate according to the inferred local transmission). The false positive rate was assessed through genotyping a subset of de novo mutations through a primer-extension assay (Sequenom, CA), and the false negative rate through allele-dropping simulations. The impact of paternal age on de novo mutation was calculated through Bayesian linear regression allowing for individual-specific false negative rates.

2. Samples

Nine individuals from a 3-generation Western chimpanzee (*Pan troglodytes verus*) pedigree were housed in a colony at the Biomedical Primate Research Centre (Rijswijk, Netherlands). Individuals C-I were born at the Centre; individuals A and B were caught from the wild. No disorders or syndromes are reported for the individuals. We verified reported genders by calculating the total heterozygous genotype call frequency on the X chromosome for male and female samples (P-value 0.007, unpaired t-test on observed X chromosome heterozygosity between genders). We also evaluated relatedness between individuals by calculating the Pearson correlation between individuals' genotype vectors excluding monomorphic sites and sites with missing genotypes. The observed pairwise correlation between individuals is concordant with their expected relatedness (Figure S2). Dates of birth were obtained from Centre records (estimates for A and B), and age at conception estimated using a gestation estimate of 0.621 years (*29*).

3. DNA extraction and sequencing

Blood samples were supplied under the EUPRIM-Net EU contract RII-026155 and CITES approval. DNA was extracted from peripheral blood mononuclear cells using affinity bead selection (Qiagen, Netherlands), and 100 bp paired-end shotgun libraries, with an average insert of 450 bp, were constructed. Libraries were sequenced on Illumina HiSeq 2000 instruments using standard chemistry and vendor provided base calling software, generating 0.9 Tb of sequence (Table S1).

4. Read-mapping, variant calling and filtering

Reads were aligned to the CGSC 2.1.3/panTro3 chimpanzee reference genome using Stampy (*27*) resulting in an average mapped coverage of 35x, and 90.8% of the aligned genome covered by at least 15 reads (Table S1).

Mapping-based variant calling

Variants were called using Platypus (version 0.2.1; www.well.ox.ac.uk/platypus) with default settings, treating samples as coming from unrelated individuals (family relationships used in downstream analyses). No adjustments were made for chromosome X. Variants were then filtered based on sequence context, with only variants satisfying all the following criteria retained:

- Variant does not lie in a homopolymer (run of the same base) of length 6 or more.
- Variant does not lie in a tandem repeat of length 10 or more.
- Not all samples are homozygous for the alternative allele.
- Fewer than 5 other variants are within +/- 25bp window.
- Variant has a single alternative allele.
- Variant does not lie in a predicted indel hotspot (*30*).

After filtering, we retained 3,957,360 single nucleotide variants (SNVs). Of these sites, 2,853,842 (66%) had been detected in an earlier study with partially overlapping samples (*18*). The overall transition/transversion ratio was 2.24 (2.28 and 2.18 for sites which had, or had not, been detected in (*18*)). We also called 169,387 small indels.

Assembly-based variant calling

Variants were called with the de novo assembler and variant caller Cortex (*14*). De Bruijn graphs were built for each sample with k=31, and in doing so PCR duplicates were removed. Reads were also broken at homopolymer runs longer than 8bp and at bases with base quality less than 5. Graphs were individually cleaned of low frequency contigs to remove sequencing error. Calling was performed with the 'Joint' pipeline (*31*). The reference genome was not used in calling. For the population filter, putative variants sites were filtered out if the relative likelihood of the variant

model was less than 10 times greater than repeat or error models. Variants whose 5' flank did not map with MAPQ >= 30 were discarded. Cortex called 3,400,416 bubbles, resulting in 3,727,902 variants of which 2,981,175 (80%) variants passed the filters.

5. Intersection call set

To identify a high quality subset, we intersected the mapping- and assembly-based calls (henceforth referred to as the intersection call set) leaving 1,595,502 sites with an overall transition/transversion ratio of 2.24. The genotypes (and genotype likelihoods) inferred through the mapping approach were assigned to the intersection call set.

6. Inferring recombination events and transmission across the genome

To reconstruct the patterns of descent of the parental chromosomes, we implemented a robust Lander-Green algorithm (*32*) to jointly infer the phase of parental genotypes and their transmission to children (the transmission vector). The standard Lander-Green approach is sensitive to mis-specified genotypes and false-positive errors (*33*), which cause artifactual switches between states. To guard against such errors, we used a two-stage approach. First, we inferred a transmission vector scaffold across non-overlapping intervals using a robust approach and allowing for at most one recombination event between intervals and none within. Informed by LD-based estimates of recombination in chimpanzees (*18*), we chose an interval size of 1 Mb, so that we would expect at most one recombination event per interval. Second, we fine-mapped the location of inferred cross-overs. Details of the inference method are given below.

The Lander-Green algorithm is a hidden Markov model (HMM) in which the hidden states are transmission vectors and emissions are the genotypes at each locus. Transmission vectors are binary vectors of length two times the number of children that describe which grandparental chromosome (grandpaternal or grandmaternal) is inherited for each child. Let the number of genotyped children be n. For each child i in $\{1,2,...,n\}$, let $V^{i,m}$ and $V^{i,p}$ be the maternal and paternal transmission indicators respectively, where $V^{i,m}(s)$ is 0 if the grand-maternal chromosome is inherited and 1 if the grand-paternal chromosome is inherited from the mother at site s, and similarly defined for the father. There are 2^{2n} possible transmission vectors.

To infer the transmission vector scaffold, we calculate the number of sites compatible with each potential underlying transmission vector. Let l represent the range of a chromosome interval and S the set of variant sites in l. We then calculate, C_j^l, the Mendel consistent proportion, defined as the

fraction of sites compatible with transmission vector V_j, by counting the potential emitted genotypes under V_j that match the genotype that maximises the genotype likelihood $g^*(s)$ at that site

$$C_j^l = \frac{1}{|S|} \sum_{s \in S} \delta\left(g^*(s) \in G(V_j)\right),$$

where $G(V_j)$ is the set of genotypes induced by projecting the set of all possible phased founder genotype combinations onto V_j and $\delta()$ is an indicator function taking the value 1 if g*(s) is within the set of possible genotypes and 0 otherwise.

To infer the underlying pattern of transmission across intervals, we implemented a finite state machine (FSM) that takes $\{C_j\}$ as the observation distribution. Transitions in the hidden states of a child represent a change in the chromosome inherited from a parent and thus represent a cross-over event. There are $2n$ possible transitions arising from a single cross-over event. Transitions in the mother and father are independent of each other and independent of the children's hidden states. We assume the probability of more than one cross-over between regions is small since coincident events are expected to be rare between the six children. Conceptually, a cross-over is equivalent to switching the chromosome inherited by a child, hence the size of the set of transmission vectors reached through one cross-over is equal to the number of transmitted chromosomes. To generate cross-over outcomes we specify a function $\rho(V_j(l))$ that generates all potential single cross-over events applied to transmission vector V_j at chromosome interval l.

Costs on transitions in the FSM were imposed to limit switches:

(1)	$p(i \to j) = 0$	if $V_i(l) = V_j(l+1)$,
(2)	$p(i \to j) = -0.25$	if $V_i(l) \in \rho(V_j(l+1))$,
(3)	$p(i \to j) = -\infty$	otherwise.

Case 1 represents no cross-over, Case 2 represents a single cross-over in a child, and Case 3 represents multiple cross-overs in the same interval, which are ignored. Note costs are in log units.

The transition costs were chosen heuristically, but we note that switches were largely insensitive to different penalty values. We used the Viterbi algorithm to identify the highest scoring transmission vector sequence across intervals for each chromosome.

7. Scaffold refinement

To improve the resolution of the cross-overs inferred across chromosome intervals, we implemented a modified HMM to use all available SNVs and analyzed each event separately (others have used similar constrained fine-mapping approaches (*34*)). We used the Forward-Backward algorithm applied to SNVs in the 1Mb surrounding cross-over regions identified at the transmission vector scaffold level.

To incorporate genotype uncertainty we computed transmission likelihoods from genotype likelihoods. Each family structure defines the set of transmission vectors $\mathcal{V}$ compatible with Mendelian transmission and consequently induces the set of potential genotypes $\mathcal{G}$. At a site, we compute the likelihood of a transmission vector V_j where *j* in $\{1,2,\ldots,2^{2n}\}$ at site *s* as,

$$L(V_j|s) \propto P(s|V_j) = \sum_{g \in G(V_j)} P(s|g)P(g),$$

where $G(V_j)$ is the set of genotypes induced by projecting the set of all phased founder genotype combinations on V_j, and $P(s|g)$ is the genotype likelihood at site *s* for genotype *g*. We use a uniform prior for P(g); i.e. each genotype vector that is compatible with the transmission vector is assumed to be equally likely. Because the data are typically strong, our inferences are only weakly affected by the choice of the prior.

Since each transmission vector induces a set of potential genotype vectors given the founder genotypes, we can use linear algebra and pre-calculated tables of the mapping between $\mathcal{G}$ and $\mathcal{V}$ to rapidly index the genotype likelihoods in the likelihood calculation.

We represent the full state matrix of transmission as,

(1) $p(i \to j) = 1 - 12/A$ if $V_i(s) = V_j(s+1)$,

(2) $p(i \to j) = 1/A$ if $V_i(s) \in \rho(V_j(s+1))$,

(3) $p(i \to j) = 0$ otherwise.

The probability of a switch, 1/A, was assessed empirically and a value of $A = 1000$ was selected. We incorporated the scaffold information by constraining the initiating and terminating states to be those inferred from the window-based analysis (see previous section). We then used posterior decoding to determine the location of the candidate cross-over event, assessed as the switch in posterior probability dominance for the two states. Cross-over breakpoints were defined to be the 95% credible interval between transmission vector change points between the scaffold-determined

initiating and terminating states. The credible interval was calculated as the interval containing a 95% change in the posterior decoding for the initiating and terminating states requiring symmetry in the change for both states and consistency within the interval. If the decoding failed these checks the region was flagged as 'complex' and was set to the transmission for the initiating state. Of the 375 cross-over events identified, 6 were defined to be 'complex' showing evidence for a double cross-over nearby to a single cross-over event (Table S2). Including the complex events, the median credible interval was 8,752 bp, spanning 5 variant sites (Table S3).

To visualize information about cross-over breakpoint resolution, we plotted the relative likelihoods of the five best ranking transmission vectors versus the inferred transmission vector after refinement (Figure S3). The majority of sites are well resolved, though there are both artifacts (e.g. chromosome 9 mis-assembly at c. 155 Mb) and possible missed events (e.g. potential double cross-over on chromosome 1 at 200 c. Mb). We also took the 192 cross-over intervals from the intersection set localization resolved to less than 10 kb and plotted the average LD-based recombination rates in Western chimpanzees (*18*) around their midpoints. Reassuringly we see strong (c. 3.7-fold) increase in LD-based recombination rates at these locations (Figure S6), serving to validate both datasets.

The total genetic map length was calculated by applying the Kosambi map function (*35*) for 12 meioses to the total number of detected crossovers.

Chromosome X

To analyse the X chromosome, we separated the pseudoautosomal region (PAR) from the rest of chromosome. Since the location of the PAR has not been defined in the chimpanzee reference genome, we defined the PAR as the human reference (GRCh37) PAR1 coordinates mapped to panTro3 coordinates (by mapping the GRCh37 PAR interval to panTro3 coordinates through syntenic alignments using the UCSC liftOver tool (*36*) this operation is known as liftOver). The panTro3 PAR coordinates were chrX:43,989-2,693,334. We analysed the PAR as for autosomes (described above) and modified the set of transmission vectors and the induced genotype vectors accordingly for the non-PAR to allow for hemizygous states in males. We retained apparent heterozygous sites on the male X.

Pseudoautosomal region (PAR)

The high repeat content of the PAR makes its assembly difficult. In the chimpanzee reference the PAR is covered by 214 contigs with a median length of 1,734 bp. There are almost no contigs between 0.8 Mb and 1.4 Mb. Nevertheless there are five contigs with length greater than 20 kb mapped to each end of the PAR. Inspection of alignments at these contigs indicated good mapping.

Given the contig spacing, we expect to have power to detect events separated by at least 0.13 Mb but may miss events occurring over shorter distances (Table S4). Since cross-overs are rare events, we therefore decided to infer the transmission patterns across the PAR using the five largest contigs.

Across the 5 contigs we detect 5 cross-overs; 4 paternal events in E, G, and two in I, and 1 maternal event in G (Table S4). To check for consistency, we also inferred transmission vectors over 1 Mb intervals, as was done for the autosomes. We recovered events in D and E, the only events detectable at a 1 Mb scale. We also inferred a cross-over in the grand-paternal chromosome transmitted to F which could indicate a double cross-over arising in the interval between the second and third 20 kb contigs, however this is also difficult to distinguish from a signal arising from mis-assembly. We therefore, do not include this event(s) in computing Table S5.

Non-PAR

To analyse the non-PAR region of chromosome X, we generated a set of X-specific emission and transmission probabilities. Through this approach we detected 9 female-specific cross-over events in the children: 3 in D, 1 in E, 1 in F, 1 in G, 1 in H, and 2 in I.

8. Classification of sites inconsistent with the inferred transmission

To distinguish between instances of segregating variant, *de novo* mutation, segregating deletion, gene conversion, and error, we calculate the likelihood of each model conditional on the transmission vector scaffold and the data observed at sites where the genotypes were inconsistent with the inferred transmission vector. We incorporate genotype uncertainty using the reported genotype likelihoods obtained from the mapping-based variant calling.

Here we define each of the five models. It is intuitive to think of the generative model for each scenario to inform the likelihood calculation under each model. We generate the potential genotypes conditional on the transmission vector scaffold at that site; this can be made efficient through pre-calculating lookup tables to index genotype likelihood values.

Segregating variant. We model a segregating variant as those genotypes compatible with the inferred transmission vector V_j, i.e. $\{G(V_j)\}$. Under this model, genotypes incompatible with the transmission vector only arising from genotyping error.

De novo mutation. We model non-recurrent mutations arising at sites where the founders are homozygous. To generate the potential genotypes under this mutation model we introduce mutations to each of the transmitted founder chromosomes and then allow the alleles to descend through the pedigree according to the inferred transmission. Although recurrent mutations do arise,

mutations of this form are rare and hard to distinguish from mis-alignments and are therefore not modelled. We also chose not to model mutations that arise on heterozygous backgrounds because such events are difficult to distinguish from mis-alignment and mis-calibrated genotype calls.

Segregating deletion. Segregating deletions result in unexpected homozygosity in the children. To model this effect, as with de novo mutations, we consider possible deletion genotypes in the founders and use the inferred transmission vector to define the set of possible genotypes in the children. To guard against assembly-driven false positives we require that least one founder has a reference compatible haplotype.

Gene conversion. We model single site gene conversions as site-specific changes in the chromosome inherited by one of the children allowing only one cross-over, i.e. requiring that a switch to a new transmission vector V_i satisfies $V_i \in \rho(V_{j,\text{scaffold}})$. Here, genotypes show no Mendel error, but are inconsistent with the inferred transmission vector.

Error. We model errors by identifying the maximum likelihood vector genotype irrespective of any constraint.

9. Identification and validation of de novo base substitutions

To guard against false positives arising through differential mapping between individuals, we constructed a genome mask across the alignments, we then identified de novo point mutation candidates in the alignment-based calls using filters on evidence, purity and consistency.

Mask

To avoid potential contamination from regions with sparse coverage that may induce false negatives and mis-assembled regions and/or mis-alignments, we generated a genome mask across individuals. First we applied the GATK CallableLoci module (*37*) to each sample's alignment using the following parameters:

- The fraction of reads with mapping quality ≤ 1 exceeds 0.1
- The number of aligned reads < 4, where reads must have a mapping quality ≥ 10 and base quality ≥ 20 to be counted

We then identified regions of the genome that passed filters in all individuals generating a single mask. Of the 3,307,943,878 bases in panTro3, we identified 2,537,740,297 bases (91.2% of the mappable reference genome, i.e. excluding bases called as N and gaps) across the individuals that satisfied the criteria (Table S9).

Next to incorporate the impact of the reference-based filters applied to the alignment-based calls, specifically the homopolymer, tandem repeat, and indel hotspot filters (see Mapping-based variant calling Section), we identified all regions in the reference sequence that would fail these filters and subtracted them from the mask resulting in 2,472,016,029 bases (89.8% of the mappable reference genome, Table S9).

Within the mask there were 25,212 sites with Mendel inconsistent genotypes on the autosomes, i.e. de novo mutation candidates. To distinguish de novo mutations from artefacts we applied three filters:

Evidence filter

To identify candidates relative to alternate models, we calculated the relative likelihood between the de novo point mutation model and the next best model. There was a lack of evidence for de novo mutation at the overwhelming majority of Mendel inconsistent sites; only 649 sites across the autosomes had relative log likelihood greater than zero for the de novo point mutation model. For sites with evidence, c. 25 % had a relative likelihood less than 5 (Figure S7). The threshold value was chosen by an iterative threshold-then-check procedure on the Mendel inconsistent calls. First, we identified a high-confidence subset from the candidates as a proxy to assess specificity under the threshold value. Second, we incremented the threshold value, filtered, and then subsampled the highest-ranking candidates that failed filters and manually inspected their alignments to confirm that they were indeed true negatives by checking for transmission consistency, read mapping, and purity. After converging to a threshold value between 4 and 5, we assessed a larger subsample of the candidates now also identifying false-positive candidates. Through this process we chose, a threshold value of 5, which removed all but 337 candidate sites.

Purity filter

To guard against potential false-positives arising through mis-genotyping, we required there be no trace of the mutant allele in any of the reads in founders or children where the variant allele was not called. We obtained counts of reads containing the mutant allele through the Platypus NV field. Note that some reads with reads containing a variant can be missed through this approach because Platypus applies upfront filtering to the reads before NV is calculated.

Consistency filter

Further, we required that all candidates segregate according to the inferred local transmission.

Application of these filters left 204 candidate de novo mutation sites across the autosomes (Table S6, Figure S8) and 3 on the X chromosome. No sites passed the de novo filters in the PAR (Table S6).

Validation against segregating variation in Western Chimpanzees

Previously, we sequenced 10 unrelated chimpanzees (*18*). Since mutations are rare events, de novo mutation candidates that are also detected as variation segregating in the unrelated individuals are likely to have arisen through false positive errors. To maximize the opportunity to detect errors, we mapped the raw, unfiltered variants called in the 10 chimpanzees (6,869,589 autosomal sites) to panTro3 (6,792,077 sites successfully mapped). None of the 207 candidates were called as variants in the data set.

Validation through re-genotyping

To estimate the false positive rate of the detected de novo mutation candidates, we attempted to validate 93 candidate de novo mutations using the Sequenom platform. The sites were composed of all the detected candidates in individuals E (45 sites) and F (35 sites), and 13 suspected false negatives across all individuals, which were the candidate de novo mutation sites that failed the purity filter, but had only one underlying conflicting read with a low assigned base quality.

Amplicons were designed using Sequenom Typer4.0 software (Sequenom, CA). One site failed amplicon design.

To guard against Sequenom errors arising from (a) PCR amplification based error and (b) measurement error, we duplicated each (a) amplification reactions and (b) assay spotting. Hence each site was genotyped four times in each individual.

Of 92 amplicons, 18 either failed, had inconsistencies across replicates, or mis-calling of the cluster plots. For the remaining sites, we calculated false positive rates under two scenarios. First, we considered a liberal case, where we only required genotype consistency within the trio relevant to the affected individual. Second, we required that the genotypes were consistent across the entire pedigree.

The false positive rate for the remaining sites (having sufficient information) was c. 1.67% (Table S7). Of the 15 potential false negative candidates included in the genotyping panel, 13 were true de novo mutations and 2 were not.

Manually curated set

Inevitably the set of de novo mutation candidates contain some small number of false positives, and a manually curated, conservative call set may be desired - though this comes with the undesirable addition of heuristic filtering strategies. To manually curate the calls, we individually assessed the properties of local alignments around each candidate. If we suspected potential mis-calling, or mis-alignment we marked these sites, this information is recorded in the *is.suspicious* indicator variable in

Table S6, there are 9 such sites. Excluding these sites has a minimal effect on estimates of male bias or the paternal age effect.

10. De novo mutation clustering

To assess the extent of within-sample clustering of de novo mutation events we calculated the number of de novo mutations that lie within a series of distance intervals from other de novo events. We then permuted the location of mutations, thus preserving the total number received by each individual. The permutation was run for 1000 iterations to given the null distribution and the expected counts of co-localised mutations shown in Fig. S9A. Moreover, the simulations were used to calculate a p-value for excess within-sample clustering over a scale of 1Mb. In no simulation was the extent of within-sample clustering as great as that observed).

To ensure that clustering is not the result of regions with poor reference genome quality we repeated the analysis using only the manually-curated variant set (Fig. S9B). This reduced the number of very-closely co-located mutations, but a strong excess of within-sample clustering is still observed ($p<0.001$).

11. Impact of sequence features on de novo mutation rates

To assess the impact of sequence context, potentially reflecting different mutational processes, on de novo mutation we considered repeat (rmsk), genic (refGene) and assembly-gap tracks downloaded from the UCSC Genome Browser (*36*).

First we looked for enrichment with proximity to each of the three features assessing significance through permutation. Specifically, we preserved the spacing between events on each chromosome by shifting all bases by a uniformly sampled perturbation distance modulo chromosome length. We found no deviation from the null expectation for each of the three tracks (Figure S10a, b, and c). Furthermore, we observe that point mutation rates show no significant distinction between repetitive and non-repetitive contexts with similar values of alpha and paternal-age effect (Figure S11).

12. Estimation of de novo mutation false negative rate

To estimate individual-specific false negative rates, we used allele-dropping simulations, using empirical distributions of coverage and allele-balance. For each simulated event, we introduced a mutation to the aligned genome at a uniformly sampled location and a founder chromosome, requiring that transmission respect the inferred local transmission. For the affected individual(s) we then allocated the aligned reads drawing from the empirical allele balance distribution, which was

generated from a sample of one million heterozygous sites. We then calculated genotype likelihoods across all the individuals at that site and recorded the true underlying mutation. The false negative rate was then calculated through applying our de novo mutation detection pipeline to the simulated mutations.

Though no transmission vector was enriched for false negatives, we did observe differential false negative rates across the pedigree (Table S8) arising from lower coverage in individual C that resulted in greater genotype uncertainty.

13. Determining the phase of de novo mutations

We used both direct phasing from read-pair data and phasing through transmission in F to assign paternal or maternal origin to de novo point mutations.

Direct phasing from read-pair data

To utilize information in reads and their mate pairs to phase variants in the filtered mapping-based calls, we first filtered low quality reads requiring bases at variant sites have quality > 20 and mapping quality > 30, then for each individual we assigned haplotypes and calculated strand bias based on the supporting reads.

To phase each de novo mutation, we extracted neighbouring heterozygous sites and used transmission to identify phase and assign maternal or paternal origin.

We assigned parental origin to 61 paternal and 11 maternal mutations through this approach (c. 25% of the candidates).

Phasing through transmission in individual F

We further phased mutations inherited by individual F through transmission to her offspring, assigning 30 events of paternal origin and 5 events of maternal. We note that there was exact agreement between phasing results through transmission and direct phasing for events detected in F.

14. Estimating male-bias (alpha)

We define α as the ratio of the total number of paternal to maternal germ line point mutations. To measure uncertainty in the estimate we use a Bayesian MCMC approach, modelling the observed total counts as independent Poisson-distributed random variables each with an independent uniform (0,1000) prior. We estimate 95% credible intervals from the posterior 95% Equal-Tailed Probability Interval (ETPI).

Using the 11 maternal and 65 paternal phased candidates we estimate $\alpha = 5.26$ (ETPI = 2.93 – 10.47). As a consistency check we also estimate α for individual F because the phase of mutations in this individual was confirmed through two independent approaches. For F we estimate $\alpha = 5.37$ (ETPI = 2.39 – 14.58). If we include the confirmed false negatives from the re-genotyping experiment this increases to 6.21 (ETPI = 2.78 – 16.62).

15. Estimating mutation rates

To estimate the de novo point mutation rate we used Bayesian linear regression to estimate the slope and intercept coefficients for maternal and paternal age effects using all 204 autosomal mutations. To convert to estimates of rates of neutral diversity and divergence we used published estimates of female- and male-specific generation times in the wild (*20*) and our estimate of the length of accessible genome across the autosomes. We estimated paternal-age effects both with and without using information on the parent of origin.

Estimating female- and male-specific mutation rates without parent-of-origin information

Initial linear model analysis indicated that there is no significant maternal age effect (P>0.05). To provide estimates of coefficients and measure uncertainty in these estimates we used Bayesian linear regression, allowing for a paternal effect only and accounting for variation in false negative rate across the children.

Specifically we analysed d_j, the total number of germ-line point mutations per child *j*, and the father's age at conception, t_j, to estimate the posterior mean and 95% credible interval for the posterior Equal-Tailed Probability Interval (ETPI) of intercept b_o and slope b_1. We modelled the number of point mutations as:

$$d_j \sim \text{Pois}(\lambda_j),$$

$$\lambda_j = (b_o + b_1 t_j)\beta_j,$$

where $j \in \{D, E, F, G, H, I\}$ and β_j is the power estimated for child *j* through the allele-dropping simulations. Priors for coefficients b_o and b_1 specified as uniform(-100,100), and the likelihood calculated as a Poisson probability.

We then estimated coefficients through rejection sampling, using updates

$$b_o^{i+1} = b_o^i + N(0, \sigma_0^2)$$

$$b_1^{i+1} = b_1^i + N(0, \sigma_1^2).$$

We found that setting σ_0^2 to 2 and σ_1^2 to 0.5 gave a good balance between efficient mixing and convergence. The MCMC was run for 10^6 iterations sampling every100 iterations and discarding the first 3000 samples.

Estimating female- and male-specific contributions using parent-of-origin information

To use parent of origin information, accounting for variation in power to detect de novo mutations across the pedigree and variable recovery of parent-of-origin information (through read-based and transmission-based phasing), we elaborated the Bayesian linear model above.

We modelled the number of point mutations detected in child *j* by allowing for separate male- and female-specific intercept terms.

$$d_j \sim \text{Pois}(\lambda_j),$$

$$\lambda_j = (b_{o,f} + b_{o,m} + b_{1,m}t_j)\beta_j, \text{ and}$$

$$n_{j,f} \sim \text{Binom}\left(n_{j,f} + n_{j,m}, \frac{b_{o,f}}{b_{o,f} + b_{o,m} + b_{1,m}t_j}\right).$$

Here, $n_{j,f}$ is the number of de novo mutations in child *j* of maternal origin and $n_{j,m}$ is the number of paternal origin. Priors for all coefficients were uniform(-100,100), and the likelihood was calculated as the product of a Poisson probability for the parental effect and Binomial probability for the male bias. We used an MCMC approach with rejection sampling to estimate coefficients, again using the Normal distribution as the proposal distribution. The algorithm was run for 10^6 iterations sampling every100 iterations and discarding the first 3000 samples.

Per year estimates of the mutation rate

To estimate the per year mutation rate we accounted for the proportion of the genome accessible to sequencing and estimated paternal and maternal generation times as the weighted mean of reported values for Western chimpanzees from (*20*); 26.3 years for females and 24.3 years for males. We then calculated the total number of mutations using the estimated female- and male-specific coefficients for the autosomes, and accounted for effective population size differences on chromosome X by assuming equal population sizes of males and females (i.e. the X chromosome has an effective population size ¾ of that of the autosomes). Human parameters were estimated from (*7*).

16. Estimates of human-chimpanzee divergence from genome alignments

To identify point substitutions between human and chimpanzee lineages, we downloaded the Ensembl EPO 6 primate alignment (release 75) and constructed GRCh37 aligned per chromosome

sequences for human, chimpanzee and orang-utan. We removed EPO alignments where the alignment segment mapped to multiple locations within the same species; we also recorded the locations of indels on the human lineage but removed these events to preserve GRCh37 coordinates. From these aligned sequences we identified substitutions between humans and chimpanzees.

We applied two filters to guard against false-positive substitutions arising from mis-assembly or mis-alignment. First, we filtered out substitutions within 10 bases of an indel or base assigned as N (unknown type). Second, we identified substitutions with more than 5 substitutions within the 12 flanking bases (>5 substitutions in 25 bases) and masked that window.

For the filtered regions we recorded a) the locations of N-bases, and b) the locations of insertions relative to the human lineage in the aligned-chimpanzee sequence, c) the 25 base windows with more than 5 clustered mutations, to generate a divergence mask for the species alignments.

To enable comparison with the observed spontaneous point mutation rate, we identified the regions accessible to both the pedigree sequencing and the species alignments in GRCh37 coordinates (combined mask). We then mapped the de novo mutation candidates (1 autosomal mutation failed liftOver, and 3 X chromosome mutations failed liftOver) and Mendel inconsistent sites from panTro3 coordinates to panTro4 coordinates using the liftOver tool. We applied the combined mask to the substitution candidates and the de novo mutation candidates.

Of the 33,287,505 substitution candidates identified on the autosomes, filter a) removed 65,781, b) removed 2,748,774, and c) removed 422,447 candidates. Furthermore 3,451,952 candidates were removed by applying the combined mask and 542 candidates were removed that overlapped Mendel inconsistent SNVs in the pedigree. We used the remaining 26,598,009 substitutions and 2,204,160,410 bases of unmasked sequence in divergence analyses.

For the X chromosome, we identified 1,295,698 substitution candidates. Filter a) removed 11,915, b) 122,360 and c) 29,236 candidates. The combined mask removed 316,879 candidates and 471 candidates overlapped Mendel inconsistent SNVs. We used the remaining 814,837 substitutions and 86,111,635 bases of unmasked sequence in divergence analyses.

Applying the combined mask to the de novo mutation candidates across the autosomes removed 12 candidates, we re-estimated paternal and maternal effects, values of alpha, and clustering of mutations using the remaining 192 mutation candidates.

To calculate an estimated time to most recent common ancestor (T) between humans and chimpanzees, we estimated mutation rates (μ , bp^{-1} $year^{-1}$) and divergence rates (k ,bp^{-1}) after applying the combined mask. The average number of mutations was calculated using the estimated average paternal and maternal ages for Western chimpanzees (*20*), accounting for differences in

effective population sizes between males and females on the X chromosome, and substituting in the estimated coefficients for males and females. We then calculated the divergence time as

$$T_i = \frac{k_i}{2\mu_i} \quad \text{where } i \in \{\text{autosome, chromosome X}\}.$$

17. Supplementary figure legends

Figure S1: Data analysis schematic. **A** One hundred base pair paired-end whole genome sequencing data (coloured rectangles) generated from blood samples taken from 9 member Western chimpanzee pedigree (individuals A – I). Hexagons: called variants (represented by hexagons) identified through assembly (Cortex: variants called independently; orange), and mapping (Platypus after mapping reads to panTro3 using Stampy: variants called and genotyped independently; purple). The intersection call set is identified between mapping- and alignment-based call sets. **B** The partitions of the reference genome where variation can be confidently called are identified across alignments, referred to as the accessible genome. **C** The transmission pattern across chromosomes is inferred from the intersection call set through a two-stage process. First, the transmission pattern is inferred between non-overlapping 1 Mb intervals through a robust implementation of the Lander-Green algorithm. Second, cross-over breakpoints are refined using posterior decoding on a constrained-state HMM using all available sites; grandparental chromosomes differentiated by opacity. **D** The probabilistic classification of sites inconsistent with the inferred transmission. Conditional on the inferred transmission, the likelihood of variation with error, gene conversion, de novo mutation and segregating deletion are calculated. The identified de novo mutation candidates are then filtered for false positive error modes. **E** The remaining candidates are validated through an independent technology to estimate the false positive rate. **F** The power to detect de novo mutations is calculated through simulation on the aligned genomes for each child. **G** Mutation rates, male bias and paternal age are calculated correcting for the estimated power to detect events and the length of the accessible genome.

Figure S2: Relatedness between pedigree individuals. The matrix of values represents the Pearson correlation between two individuals' genotype vectors at intersection sites across the autosomes; cells are coloured by their expected values indicated in the figure legend.

Figure S3: Relative likelihoods for the transmission scaffold. The relative likelihood between the inferred transmission vector (transmission scaffold) and the other five most likely transmission vectors at variant sites in the intersection call set; colours distinguish the relative likelihoods under each transmission vector (compared to the fitted vector). Dotted lines represent the midpoint of the inferred cross-over breakpoints. Values above the x-axis indicate sites for which one of the five alternative transmission vectors is more likely than the scaffold transmission vector. Points above

the y=0 axis indicate potential gene conversion events or various error types. Clustered regions (e.g. on chromosomes 2A, 9, 15 and 18) may indicate short cross-over events that have been missed or regions with extensive errors (e.g. arising through structural variation).

Figure S4: The distribution of cross-over events across the autosomes. For each chromosome, represented as grey rectangles, cross-over breakpoints are indicated by lollipops (female, below x-axis) and shovels (male, above x-axis) and the location of centromeres as brown rectangles. Each recipient child is assigned a different colour and distance from the x-axis (see legend). Barplots on the left hand side represent the total number of cross-overs in each individual stratified by grandparental origin.

Figure S5: The relationship between cross-overs and proximity to chromosome telomeres. **A** Represented are the mean recombination rate in humans (blue line) from (*7*) and Western chimpanzees (red line), excluding chromosomes 2, 2A and 2B respectively for **A** male-specific cross-overs and **B** female-specific cross-overs. Shaded areas represent the 95% confidence interval estimated through bootstrap resampling (N=100). Relative physical distance from the telomere was measured from the chromosome ends to centromere boundaries for each p- and q-arm (which were consequently superposed).

Figure S6: LD-based recombination rates around cross-over breakpoints inferred in the pedigree. Line represents the average recombination rate from (*18*) estimated across 1 kb intervals for 192 cross-overs whose breakpoints are resolved to less that 5 kb.

Figure S7: The effect of the relative likelihood filter threshold on the number of de novo mutation candidates passing filters. Each open circle represents number of de novo mutation candidates that pass the relative likelihood filter threshold (x-axis). The red line indicates the filter value used in this analysis.

Figure S8: The distribution of de novo mutation events across the autosomes. For each chromosome, represented as grey rectangles, the location of germ line point mutation events are indicated by dots. The child receiving the mutation is distinguished by colour and height above the x-axis. Note stacked events indicate the transmission of events from F (represented by green dots) to children.

Figure S9: The clustering of de novo point mutations within individuals. Points represent the number of clustered point mutations occurring in the same individual for increasing 100 kb distance intervals (red) and the null distribution estimated through permuting the location of mutations (blue, N=1000) for **A** all 207 candidate point mutations and **B** the 198 manually curated point mutations. Tick marks represent the 95% confidence intervals for the null distribution.

Figure S10: The relationship between de novo mutation events and different genome features. To assess whether the rate of events is affected by sequence context we calculated the distribution of mutation events with distance (red line, i.e. the observed distribution) from **A** repeat elements, **B** gene transcripts, **c** reference assembly gaps. The purple lines represent 100 permutation experiments where events were shifted across chromosomes (see Supplementary text).

Figure S11: The relationship between paternal age effect and the number of de novo mutations stratified by repeat and non-repeat DNA contexts. Points represent paternal age versus the number of point mutations in each child occurring on repeat (black) and non-repeat (red) DNA backgrounds. Linear regression coefficients are shown as dotted lines for events in repeat ($P = 0.119$) and non-repeat ($P = 0.014$). Counts were not corrected for the estimated false-negative rates.

Figure S12: The relationship between paternal and maternal age and the number of de novo point mutations incorporating phase information. The posterior mean estimates for separate paternal (blue) and maternal (red) mutation rates in Western chimpanzees are shown as solid lines with their estimated 95% ETPI envelopes, correcting for the estimated false negative rate in each offspring. Also shown as dotted lines is the estimate from humans for males (blue) and females (red); data from reference 13. Blue and red dots indicate paternal and maternal ages at birth within the pedigree respectively.

Table references

(*16*, *20*, *38-40*)