nature genetics

Genome-wide patterns and properties of *de novo* mutations in humans

Laurent C Francioli[1,15], Paz P Polak[2,15], Amnon Koren[3,15], Androniki Menelaou[1], Sung Chun[2], Ivo Renkens[1], Genome of the Netherlands Consortium[4], Cornelia M van Duijn[5], Morris Swertz[6,7], Cisca Wijmenga[6,7], Gertjan van Ommen[8], P Eline Slagboom[9], Dorret I Boomsma[10], Kai Ye[9,11], Victor Guryev[12], Peter F Arndt[13], Wigard P Kloosterman[1], Paul I W de Bakker[1,14,16] & Shamil R Sunyaev[2,16]

Mutations create variation in the population, fuel evolution and cause genetic diseases. Current knowledge about *de novo* mutations is incomplete and mostly indirect[1–10]. Here we analyze 11,020 *de novo* mutations from the whole genomes of 250 families. We show that *de novo* mutations in the offspring of older fathers are not only more numerous[11–13] but also occur more frequently in early-replicating, genic regions. Functional regions exhibit higher mutation rates due to CpG dinucleotides and show signatures of transcription-coupled repair, whereas mutation clusters with a unique signature point to a new mutational mechanism. Mutation and recombination rates independently associate with nucleotide diversity, and regional variation in human-chimpanzee divergence is only partly explained by heterogeneity in mutation rate. Finally, we provide a genome-wide mutation rate map for medical and population genetics applications. Our results provide new insights and refine long-standing hypotheses about human mutagenesis.

Understanding rates and patterns of human mutation is important for analyzing relationships among species and populations[1,2], for detecting natural selection[3,4] and for mapping genes underlying complex traits[5]. The properties of mutations have traditionally been studied using model organisms[6], fully penetrant, dominant mendelian diseases[7,8], and comparative genomics and population genetics approaches[9,10]. However, these approaches are limited in scope, indirect and influenced by other factors such as natural selection. Using high-throughput sequencing technologies, recent pedigree sequencing studies have provided whole-genome observations of germline *de novo* mutations and showed that mutation rate increases with paternal age[11–13], varies along the genome in weak correlation with various epigenetic properties and is higher in conserved genomic regions, including exons[11].

We identified *de novo* mutations in 250 Dutch parent-offspring families (231 trios, 11 families with monozygotic twins and 8 families with dizygotic twins) by whole-genome sequencing of blood-derived DNA to 13-fold coverage. We considered dizygotic twins as distinct and included one twin from each monozygotic twin pair, resulting in a total of 258 offspring. We identified 11,020 *de novo* mutations, with an estimated sensitivity of 68.9% and specificity of 94.6% (ref. 13). By comparing 350 validated mutations in monozygotic twins, we estimate that ~97% of the mutations in our data are germline and ~3% are somatic. To account for the mutation calling biases inherent to sequencing data, we simulated *de novo* mutations, taking into account fluctuations in sequence coverage (Online Methods), and used this simulated set as a 'null' baseline against which we compared observed *de novo* mutations to characterize their patterns and properties. We also corrected for variation in the sequencing coverage of different family trios.

Paternal age explains about 95% of the variation in global mutation rate in the human population[12]. Specifically, there is an increase of one to two mutations per year of paternal age[11–13], which is thought to stem from continuous cell divisions in the paternal germ line, beginning in the embryonic development of primordial germ cells and continuing in spermatogenesis throughout a man's life. A key question is whether changes in the global mutation rate are accompanied by a shift in the mechanisms of spontaneous mutagenesis. If so, such a shift might be reflected in the genomic distribution of *de novo* mutations.

[1]Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands. [2]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [4]A full list of members and affiliations appears in the **Supplementary Note**. [5]Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands. [6]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [7]Genomics Coordination Center, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [8]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [9]Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands. [10]Department of Biological Psychology, VU University Amsterdam, Amsterdam, the Netherlands. [11]Genome Institute, Washington University, St. Louis, Missouri, USA. [12]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [13]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. [14]Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands. [15]These authors contributed equally to this work. [16]These authors jointly supervised this work. Correspondence should be addressed to S.R.S. (ssunyaev@rics.bwh.harvard.edu) or P.I.W.d.B. (pdebakker@umcutrecht.nl).

Received 6 October 2014; accepted 7 April 2015; published online 18 May 2015; doi:10.1038/ng.3292

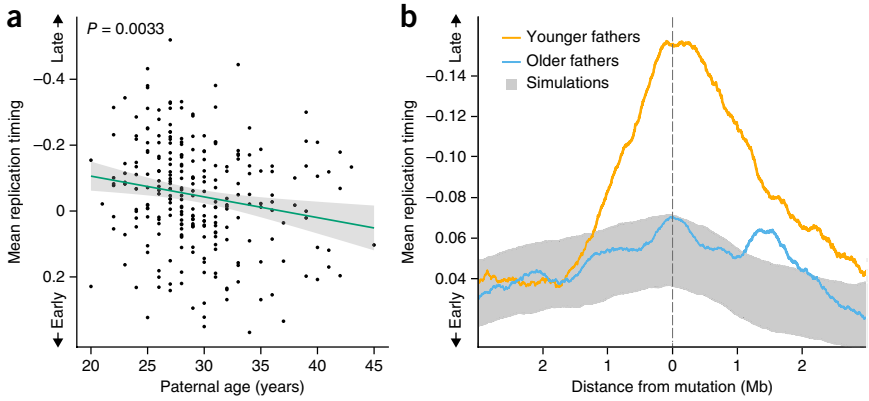

Figure 1 Mutations in the offspring of younger fathers are biased toward later-replicating regions. (**a**) Mean replication timing of *de novo* mutations in each of the 258 offspring as a function of their father's age. The green line shows the least-squares regression line ($P = 0.0033$), and the gray area represents the 95% confidence interval. The downward slope of the regression line indicates a shift of mutations toward earlier-replicating regions with advancing paternal age. Replication timing is represented by *z* score, with higher values indicating earlier replication. (**b**) The mean replication timing profile around *de novo* mutations, stratified by paternal age (orange: <28 years, $n = 3{,}697$; blue: ≥28 years, $n = 7{,}323$). The gray area shows the null expectation based on simulations (mean ± 1 s.d.). The age of the split between younger and older fathers was chosen to maximize the difference between the groups ($P = 5.7 \times 10^{-4}$, 23 tests). Mutations in younger fathers tend to be located in large (~2 Mb) regions of late-replicating DNA. In contrast, the replication timing distribution of mutations in older fathers is similar to that of simulated mutations.

Because previous studies have suggested that the epigenetic landscape varies with age[14], we investigated whether paternal age was associated with the location of *de novo* mutations with respect to various epigenetic variables (Online Methods). Using a linear regression model, we found that the replication timing of *de novo* mutations was significantly associated with paternal age ($P = 0.0022$; **Fig. 1a**), whereas chromatin accessibility, chromatin modifications and recombination rate were not ($P > 0.098$; **Supplementary Fig. 1**). Mutations in the offspring of younger fathers (<28 years old) were strongly enriched in late-replicating genomic regions ($P = 4.9 \times 10^{-4}$; **Fig. 1b**), whereas there was no significant replication timing bias for mutations in the offspring of older fathers (≥28 years old; $P = 0.68$; **Fig. 1b**). The age groups were chosen to maximize the difference in replication timing between the mutations in the offspring of younger and older fathers ($P = 5.7 \times 10^{-4}$; **Supplementary Fig. 2**). The age-dependent association between the distribution of mutations and DNA replication timing was not specific to the replication timing data set nor to the cell type in which it was measured (**Supplementary Fig. 3** and **Supplementary Table 1**). Together, these data show that *de novo* mutations in the offspring of younger fathers are biased toward late-replicating regions, whereas those in the offspring of older fathers are not.

To confirm that our results are due to paternal rather than maternal age (which are highly correlated with each other within families; $r = 0.81$), we restricted our analysis to mutations with unambiguous paternal or maternal origin (Online Methods). Consistent with the results above, paternal age was significantly associated with replication timing ($n = 1{,}991$ mutations; $P = 0.032$), but maternal age was not ($n = 630$ mutations; $P = 0.26$). The difference between paternal and maternal age was significant ($P = 0.0019$), with the comparison controlled for the different numbers of mutations in the paternal and maternal lines (**Supplementary Fig. 4**).

Replication timing itself correlates with chromatin structure and gene activity: early-replicating regions of the genome have a higher gene density and elevated gene expression levels in comparison to late-replicating genomic regions[15]. Therefore, the paternal age effects described above are likely to have functional consequences. Indeed, we found that the proportion of *de novo* mutations in genic regions increased by 0.26% with each additional year of paternal age ($P = 0.0085$; **Fig. 2**). On average, the offspring born to 40-year-old fathers harbored twice as many genic mutations as the offspring of 20-year-old fathers (19.06 versus 9.63 mutations) but only 55% more intergenic mutations (35.24 versus 22.68). An important

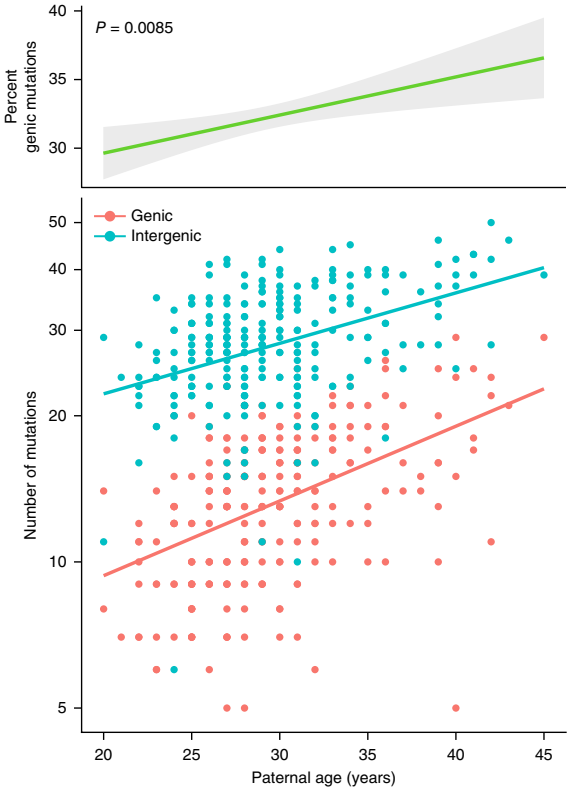

Figure 2 The offspring of older fathers harbor a higher percentage of *de novo* mutations in genes. Top, the percentage of *de novo* mutations within genic regions as a function of paternal age at conception ($P = 0.0085$; slope = 0.26% per year of paternal age). The green line shows the least-squares regression line ($P = 0.0085$), and the gray area represents the 95% confidence interval. Bottom, the number of genic (red) and intergenic (blue) *de novo* mutations in offspring (on a logarithmic scale) as a function of paternal age. The red line shows the least-squares regression for genic mutations ($P < 2 \times 10^{-16}$); the blue line shows the least-squares regression for intergenic mutations ($P = 3.7 \times 10^{-14}$). The steeper slope of the regression line for genic mutations indicates a faster relative increase in the rate of genic mutations in comparison to intergenic mutations with paternal age.

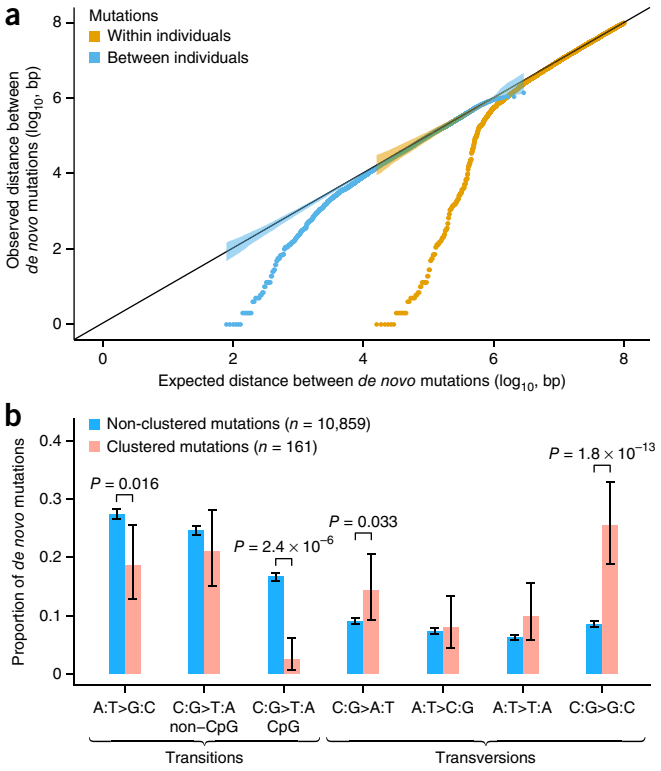

Figure 3 Mutation clusters exhibit a unique mutational spectrum. (**a**) The distances between adjacent *de novo* mutations (observed) as compared to a uniform distribution of mutations across the genome (expected). Closely spaced mutations are enriched both across individuals (golden yellow) and within individuals (blue). The strength of this effect is greatest within individuals, where 78 mutation clusters of up to 20 kb in length are observed. In fact, 1.5% of all *de novo* mutations in our study are in such clusters. Shaded areas represent the 95% confidence intervals. (**b**) Comparison of the mutation spectra between clustered (pink) and non-clustered (blue) *de novo* mutations (error bars, 95% confidence intervals). We defined mutation clusters as regions with two or more mutations within 20 kb of each other in the same individual. Mutations within clusters show a significantly reduced number of transitions ($P = 1.2 \times 10^{-12}$ for all transitions; $P = 4.1 \times 10^{-6}$ when excluding C>T transitions at CpG sites) and a strongly elevated number of C>G transversions ($P = 1.8 \times 10^{-13}$).

implication of this result is that mutations in older fathers are not only more numerous, but they are also individually more likely to have functional consequences.

Together, these observations suggest that the increase in the number of *de novo* mutations with paternal age is accompanied by a change in their mechanism of formation related to DNA replication timing and, consequently, their chromosomal distribution and functional impact. Although the source of these differences is unclear, they could reflect variations in replication timing or mutagenesis between the symmetrical mitoses that occur during the generation of paternal germ cells and the asymmetrical mitoses of the spermatogonia during spermatogenesis.

Irrespective of paternal age, mutation rates are higher in functional genomic regions[11]. Indeed, 1.22% of *de novo* mutations were exonic, which represents a 28.7% enrichment in comparison to our simulated null baseline ($P = 0.008$). Similarly, mutation rates in regulatory regions marked by DNase I–hypersensitive sites (DHSs) were elevated ($P = 0.005$). The elevated mutation rate for both exons and DHSs appeared to be driven by CpG dinucleotides, as after excluding CpGs we observed no significant difference from the null expectation. Methylated CpGs represent highly mutable sequences in humans. The increased mutation rates at CpG sites are thought to have evolved recently (around the time of mammalian radiation)[16]. Thus, whereas sequences in neutrally evolving regions of the genome have had sufficient time to equilibrate with respect to dinucleotide contexts, purifying selection has maintained hypermutable CpGs in functional regions[17,18].

Unlike observations in cancer somatic mutation[19,20] and in comparative genomics studies[9], we did not detect a reduction in mutation rates in transcribed and DHS regions after correcting for sequence context. However, we note that our study was only adequately powered to find a depletion of at least 17.4% in these regions (90% power; **Supplementary Fig. 5**).

The distribution of *de novo* mutations along the genome was non-random, both within and across individuals (**Fig. 3a**), extending beyond correlations with epigenetic variables and functional elements. At the extreme, we observed clusters of nearby mutations in an individual. This clustering was particularly strong for distances of up to 20 kb ($P < 1 \times 10^{-6}$), in which range there were a total of 78 clusters of 2–3 mutations. These observations are consistent with and expand on previous studies based on more limited data[11,21]. We did not find a significant difference between the 161 clustered mutations and the 10,859 non-clustered mutations with respect to recombination rate ($P = 0.52$) or replication timing ($P = 0.059$). Interestingly, mutations within clusters exhibited a unique mutational spectrum ($P = 9.7 \times 10^{-16}$), with reduced numbers of transitions and strongly elevated numbers of C>G nucleotide changes (**Fig. 3b**), suggesting a specific underlying mechanism. The nucleotide contexts of these clustered mutations are distinct from the previously observed same-strand TCW>TTW or TCW>TGW mutations (where W corresponds to either A or T), reminiscent of the activity of the APOBEC cytosine deaminases that leads to clustered mutations in cancer cells[22,23]. Although not caused by APOBEC activity, C>G mutations may result from deaminated cytosines in single-stranded DNA that would be converted to apurinic sites by base-excision repair DNA glycosylases and subsequently subjected to error-prone translesional DNA synthesis[24].

Comparative genomics studies have predicted that the mutation rate is variable on the megabase scale[9,25]. However, the extent to which the mutation rates and patterns predicted by comparative genomics studies reflect the true underlying properties of germline mutations is unknown. Here we sought to separate the intrinsic properties of mutational processes from other population processes, such as background selection, hitchhiking and biased gene conversion.

Previous studies have shown that nucleotide diversity within populations (π) is correlated with the local recombination rate, but it is unclear whether this correlation is due to a mutagenic effect of recombination[26] or to background selection and hitchhiking mechanisms[27,28]. In our study, local recombination rates[29] were significantly associated with *de novo* mutation rates ($P = 0.0015$), when controlling for CpG sites and GC content. Despite this association, we found that the rates of both mutation ($P < 2 \times 10^{-16}$) and recombination ($P < 2 \times 10^{-16}$) independently contributed to nucleotide diversity. Thus, recombination appears to influence nucleotide diversity above and beyond any mutagenic effect.

We next estimated the extent to which human-chimpanzee sequence divergence is influenced by mutation and recombination rates (**Fig. 4**). The correlation between the substitution rates from a human-chimpanzee comparative genomics (HCCG) model[30] and observed *de novo* mutation rates was significant ($r = 0.18$; $P = 1.3 \times 10^{-15}$). When

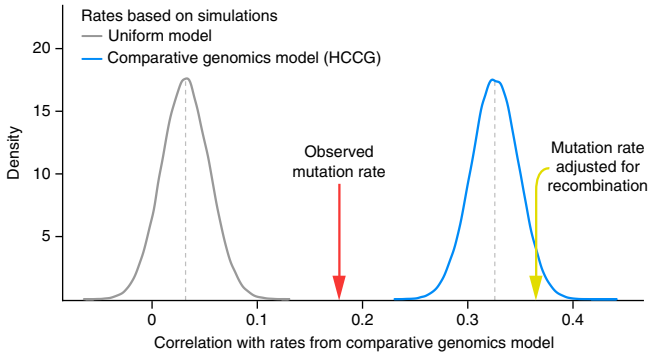

Figure 4 Influence of mutation and recombination rates on human-chimpanzee divergence. The correlation to substitution rates computed from a human-chimpanzee comparative genomics (HCCG) model is plotted for mutation rates inferred from a uniform mutation rate model (gray distribution) and for mutation rates inferred from the HCCG model itself (blue distribution), both based on 100,000 simulations of $n = 11{,}020$ mutations and binned in 1-Mb windows. By sampling the same number of mutations, comparisons between rate estimates are meaningful. The effect of sampling is illustrated by the mean correlation of the HCCG with itself at only 0.33, with this correlation asymptotically reaching unity with infinite sampling. The correlation is also given for observed *de novo* mutation rates (red arrow; $n = 11{,}020$) and observed *de novo* mutation rates adjusted for local recombination rates[31] (yellow arrow; $n = 11{,}020$). Correlation with observed *de novo* mutation rates ($r = 0.18$) is stronger than correlations with rates based on the uniform model (mean $r = 0.032$; $P < 1 \times 10^{-5}$), indicating that the HCCG model partly captures regional variations in mutation rate. However, the correlation with observed *de novo* mutation rates is weaker than correlations with rates based on the HCCG model itself (mean $r = 0.33$; $P < 1 \times 10^{-5}$), suggesting that there are other contributing factors. When adjusting observed *de novo* mutation rates for local recombination rates, the correlation is 0.37, illustrating that substitution rates computed from the HCCG model capture both mutation rates and orthogonal evolutionary forces associated with local recombination rates.

compared to substitution rates based on sampling the HCCG itself for the same number of mutations (mean $r = 0.33$), we found that *de novo* mutation rates explained about one-third of the human-chimpanzee sequence divergence along the genome (Online Methods). However, observed mutation rates adjusted for local recombination rates[31] were more strongly correlated with the HCCG model ($r = 0.37$) than observed mutation rates alone (**Fig. 4**). This illustrates that the comparative genomics model captures both variation in mutation rate and other, orthogonal evolutionary forces associated with recombination rate, as has been suggested by others[27,32].

In contrast to the large-scale regional variation, we found that the influence of flanking nucleotides on *de novo* mutations was in excellent agreement with results based on comparative genomics[33] ($r^2 = 0.993$; **Supplementary Fig. 6**), suggesting that the mutational spectrum has been relatively constant in recent evolution. We also observed a previously predicted[26–28] strand asymmetry for mutations in transcribed regions (**Supplementary Fig. 7**), especially for A>G mutations ($P = 5.9 \times 10^{-5}$). This asymmetry is likely a byproduct of the action of transcription-coupled repair. We found a modest 2.8% depletion of mutations in transcribed regions relative to intergenic regions ($P = 0.047$). This is in sharp contrast with somatic cancer mutations, where a similar strand asymmetry was accompanied by a strong reduction in the frequency of mutations in transcribed regions[20].

Having a well-calibrated mutation model is essential for evaluating the significance of *de novo* mutation patterns observed in pedigree sequencing studies (especially in the absence of appropriate controls in disease studies)[34]. Previous mutation models have been based on comparative genomics, but, as shown above, these models are not representative of germline mutation rates alone, as they also incorporate other evolutionary forces. To bridge this gap, we used the empirical distribution of *de novo* mutations along the genome to refine a mutation model based on human-chimpanzee divergence rates, considering flanking sequence context, local recombination rates, mutation type and the transcribed strand in coding regions (Online Methods). In addition to genome-wide rates, we also calculated gene-level mutation rates, separately estimating synonymous, missense and nonsense mutation rates. This mutation rate map can be used for evolutionary inferences based on human mutation rates and for the identification of disease-associated genes with recurrent *de novo* mutations.

We describe here the most extensive catalog thus far of *de novo* germline mutations in healthy individuals, identifying several mechanisms influencing the distribution of mutations along the genome. In particular, clustered mutations suggest the existence of a new mutagenic mechanism, and the effect of replication timing on germline mutations depends on paternal age. Heterogeneity in mutation rate substantially influences genomic variation in the rate of sequence evolution, adding to the effects of evolutionary forces acting at the population level.

URLs. Picard tools, http://broadinstitute.github.io/picard/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under accession EGAS00001000644. The mutation rate map can be found on the Genome of the Netherlands (GoNL) website at http://www.nlgenome.nl/.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Gordenin for very helpful comments. The Genome of the Netherlands (GoNL) Project is funded by the Biobanking and Biomolecular Research Infrastructure (BBMRI-NL), which is financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007). S.R.S., P.P.P. and S.C. are funded by US National Institutes of Health grants 1 R01 MH101244 and 1 R01 GM078598.

AUTHOR CONTRIBUTIONS

S.R.S. and P.I.W.d.B. planned and directed the research. L.C.F. called and filtered the mutations. W.P.K. and I.R. validated candidate mutations. L.C.F. designed and executed the simulations. A.K., L.C.F. and A.M. performed replication timing analyses. P.P.P., L.C.F. and A.M. analyzed factors influencing regional mutation rates and spectra. L.C.F. and P.P.P. analyzed mutation clusters. P.P.P. and P.F.A. computed the comparative genomics model and compared it against observed mutation rates. S.C. and P.P.P. created the mutation rate map. L.C.F., P.P.P., A.K., P.I.W.d.B. and S.R.S. wrote the manuscript. A.M., S.C., C.M.v.D., M.S., C.W., G.v.O., P.E.S., D.I.B., K.Y., V.G., P.F.A. and W.P.K. provided critical feedback on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Sawyer, S.A. & Hartl, D.L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
2. Felsenstein, J. & Churchill, G.A. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**, 93–104 (1996).
3. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
4. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
5. Veltman, J.A. & Brunner, H.G. *De novo* mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
6. Friedberg, E.C., Walker, G.C. & Siede, W. *DNA Repair and Mutagenesis* (ASM Press, 1995).
7. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
8. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
9. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).
10. Schaibley, V.M. *et al.* The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res.* **23**, 1974–1984 (2013).
11. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell* **151**, 1431–1442 (2012).
12. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
13. Genomes of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
14. Jenkins, T.G., Aston, K.I., Pflueger, C., Cairns, B.R. & Carrell, D.T. Age-associated sperm DNA methylation alterations: possible implications in offspring disease susceptibility. *PLoS Genet.* **10**, e1004458 (2014).
15. Koren, A. DNA replication timing: coordinating genome stability with genome regulation on the X chromosome and beyond. *Bioessays* **36**, 997–1004 (2014).
16. Arndt, P.F., Petrov, D.A. & Hwa, T. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol. Biol. Evol.* **20**, 1887–1896 (2003).
17. Schmidt, S. *et al.* Hypermutable non-synonymous sites are under stronger negative selection. *PLoS Genet.* **4**, e1000281 (2008).
18. Subramanian, S. & Kumar, S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. *Genome Res.* **13**, 838–844 (2003).
19. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
20. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
21. Campbell, C.D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277–1281 (2012).
22. Roberts, S.A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
23. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
24. Chan, K., Resnick, M.A. & Gordenin, D.A. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair* (*Amst.*) **12**, 878–889 (2013).
25. Arndt, P.F., Hwa, T. & Petrov, D.A. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J. Mol. Evol.* **60**, 748–763 (2005).
26. Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S. & Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**, 1527–1535 (2003).
27. Begun, D.J. & Aquadro, C.F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
28. Lercher, M.J. & Hurst, L.D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**, 337–340 (2002).
29. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
30. Duret, L. & Arndt, P.F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4**, e1000071 (2008).
31. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
32. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
33. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. & Sunyaev, S. Analysis of sequence conservation at nucleotide resolution. *PLOS Comput. Biol.* **3**, e254 (2007).
34. Gratten, J., Visscher, P.M., Mowry, B.J. & Wray, N.R. Interpreting the role of *de novo* protein-coding mutations in neuropsychiatric disease. *Nat. Genet.* **45**, 234–238 (2013).

ONLINE METHODS

The Genome of the Netherlands data. This study uses *de novo* mutation data from the Genome of the Netherlands (GoNL) project, for which all data generation and processing steps were detailed in a previous publication[13]. A brief version is included here.

The GoNL project includes 250 Dutch parent-offspring families (231 trios, 8 quartets with dizygotic twins and 11 quartets with monozygotic twins) sampled throughout the Netherlands without phenotypic ascertainment. For this study, we used all 250 sets of parents as well as 258 genetically unique offspring, removing one of the two twins (chosen randomly) in each of the monozygotic twin pairs.

Samples were sequenced using 91-bp paired-end reads for libraries with an insert size of 500 bp on an Illumina HiSeq 2000. Alignment and variant calling were devised on the basis of Genome Analysis Toolkit (GATK) best practices v2 (refs. 35,36). Sequence data were mapped to human reference genome Build 37 using bwa 0.5.9-r16 (ref. 37), duplicate reads were removed using Picard tools, local indel realignment was performed around indels using GATK IndelRealigner and base qualities were recalibrated using GATK BaseQualityScoreRecalibration. Variants were called using GATK UnifiedGenotyper v1.4 on all samples simultaneously and filtered using GATK VariantQualityScoreRecalibration.

Detection of *de novo* mutations was performed using the trio-aware genotype caller GATK PhaseByTransmission, which leverages familial, population and mutation rates, and mutations were filtered using a random forest machine learning classifier (trained on 592 true positive and 1,630 false positive putative *de novo* mutations validated experimentally). We obtained a set of 11,020 high-confidence mutations in the 269 children of the GoNL project with an estimated 92.2% accuracy[13]. All putative *de novo* mutations found in the 11 monozygotic twin quartets were subjected to validation in both twins. Of the 680 mutations detected and validated in either twin, 660 were shared by both twins and 20 were unique to a single child. We therefore estimate that 97% of the mutations in our data are germline and 3% are somatic. Using GATK ReadBackedPhasing, we assigned parental origin to 1,991 paternal and 630 maternal *de novo* mutations on the basis of phase-informative reads.

Simulation of *de novo* mutations. We simulated *de novo* mutations at the read level to create a null distribution (uniform) while accounting for the effect of fluctuation in coverage inherent to high-throughput sequencing. We generated 264,768 random positions throughout the GoNL accessible genome[13] (i.e., ~1 for every 1,000 bp), excluding any position that was polymorphic in GoNL or outside the accessible genome. For each of these positions, we generated a random non-reference allele to be used as a decoy mutation. For each trio separately, we extracted reads from the children overlapping each of the positions to insert the decoy mutation. Because *de novo* mutations are always heterozygous, each read had a 50% probability of being selected to carry the mutation. For all reads selected to carry the mutation, we replaced the reference base with the decoy mutated base. Base and mapping qualities were kept intact under the assumption that altering a single base in 90 would not significantly affect these values. We then applied our entire *de novo* mutation calling pipeline to each decoy mutation.

Using these simulations, our *de novo* mutation calling pipeline had an average sensitivity of 67.9%. The sensitivity was heavily influenced by coverage across the entire trio (r = 0.87). One outlier sample showed abnormally low sensitivity (−5.8 s.d.) but was kept in the study as there were no quality concerns based on earlier quality control processing[13].

On the basis of these simulations, we estimated the power to call a *de novo* mutation as a function of coverage in each individual in the trio. We found that simulated mutations covered by at least 9 reads in each parent, by at least 4 reads in the child and by at least 30 reads across the entire trio were detected with 92.5% sensitivity. On average, 68.8% of the genome was covered in each trio using these thresholds. We considered all bases covered by these thresholds as high-confidence bases in our analyses.

To derive a null distribution for *de novo* mutations on the basis of the simulations above, we randomly sampled a single child at each of the 264,768 sites at which we inserted decoy mutations. Sampling was performed regardless of whether the simulated mutation was called in the child and led to a total of 179,845 called mutations against which we compared our *de novo* mutations.

Influence of paternal age on the genomic location of mutations. We annotated each *de novo* mutation with replication timing measured in lymphoblastoid cell lines (LCLs)[38], expression levels in LCLs[39], recombination rates[31], and DHSs and histone marks (H3K27ac, H3K4me1 and H3K4me3) measured in lymphocytes (GM12878) from the Encyclopedia of DNA Elements (ENCODE) Project[40]. We then used a linear regression model to investigate possible relationships between paternal age and the localization of *de novo* mutations with respect to the above epigenetic variables, while correcting for GC content, CpG sites and sequencing coverage. We used a stepwise model selection by Akaike information content (AIC), starting with a saturated model including all variables and their interactions, to derive a parsimonious model. The resulting parsimonious model only contained DNA replication timing (P = 0.0022) and H3K4me3 levels (P = 0.35), owing to a weakly significant interaction between the two (P = 0.035). This interaction is possibly caused by the correlation between replication timing and H3K4me3 levels. We estimated the significance of the other epigenetic variables by adding each one by one into the model and comparing the resulting models against the parsimonious model using ANOVA tests (**Supplementary Fig. 1**).

We dichotomized our database on the basis of the age of the father to contrast the replication timing profiles of younger and older fathers. We ran an exhaustive search for an age threshold that would maximize the difference between the two groups. For each of the 23 possible thresholds, we used a Kolmogorov-Smirnoff test to compare the replication timing profile of the younger and older fathers (**Supplementary Fig. 2**). We found a peak around 28 years of age ($P = 5.7 \times 10^{-4}$) and therefore used this as the age threshold. Hereafter, we will refer to fathers who were <28 years old at conception as 'younger fathers' and fathers who were ≥28 years old at conception as 'older fathers'.

We compared the distributions of replication timing for mutations from younger and older fathers using a Mann-Whitney test and found that the distributions for younger fathers were significantly shifted toward later-replicating regions ($P = 1.3 \times 10^{-4}$). We also compared the distributions of replication timing for simulated mutations against distributions in the offspring of younger and older fathers and found the latter to be shifted toward later-replicating regions ($P = 4.9 \times 10^{-4}$) and similar (P = 0.68), respectively.

We repeated the same analyses using independent replication timing data[41] in four cell types (lymphoblastoid cells, neural precursor cells, embryonic stem cells (of four separate lines) and induced pluripotent stem cells (of two separate lines)) and observed consistent results across all cell types (**Supplementary Fig. 3**).

To delineate whether the effect we observed was paternal, maternal or both, we used mutations for which we could unambiguously determine parental origin and ran the linear regression model using the father's age for the 1,991 paternally inherited mutations (β = 0.0092; P = 0.038) and using the mother's age for the 630 maternally inherited mutations (β = −0.0096; P = 0.26) separately. Because of the difference in sample size between the mutation sets, we resampled 10,000 sets of 630 mutations from the paternally inherited mutations and ran the linear regression on these sets. We found that the expected paternal effect was significantly greater than the maternal one with the same number of mutations (P = 0.0019; **Supplementary Fig. 4**).

We next ran a robust linear regression model between the percentage of genic mutations in each of the 258 offspring and paternal age, correcting for coverage, and found a significant association (β = 0.0026; P = 0.0085). We used a robust linear regression model to account for a single sample that showed an abnormally high percentage of genic mutations (>8 s.d. from the mean). This sample was no different from the others in terms of quality metrics such as coverage, SNP heterozygosity, proportion of known and new SNPs and possible contamination.

We used linear regression models to compute the increase in the number of mutations with paternal age for genic (β = 0.52; $P < 2 \times 10^{-16}$) and intergenic (β = 0.32; $P = 3.7 \times 10^{-14}$) mutations separately while correcting for coverage. On the basis of these computations, we estimated that an offspring born to a 20-year-old father would receive on average 9.63 genic and 22.68 intergenic mutations, whereas an offspring of a 40-year-old father would receive on average 19.06 genic and 35.24 intergenic mutations.

Mutation clusters. We tested whether the intra- and interindividual distributions of *de novo* mutations deviated from a simulated uniform distribution

across the genome, correcting for detection power by assigning a probability equal to the average number of high-confidence bases (see "Simulations of *de novo* mutations") across all trios on the kilobase scale. We used a Kolmogorov-Smirnoff test and found that both intraindividual ($P = 3.3 \times 10^{-4}$) and interindividual ($P = 5.8 \times 10^{-5}$) distributions were enriched for more closely spaced mutations than expected (**Fig. 3**). The strongest enrichment was for intraindividual mutations separated by up to ~20 kb, and we therefore defined mutation clusters as regions of 20 kb or less containing two or more *de novo* mutations in the same sample. We observed 73 clusters of 2 mutations and 5 clusters of 3 mutations. We ran 1mln permutations to test whether these clusters were due to generally hypermutated regions. In each permutation round, we permutated the samples to which each mutation belonged and counted the number of clusters obtained. The maximum number of clusters we found under this permutation scheme was 18, far fewer than the 78 we observed in total, indicating that clustered mutations are likely co-occurring rather than independent. We then looked at the substitution types for clustered *de novo* mutations and compared them against those of non-clustered *de novo* mutations using χ^2 tests. We also looked for differences in a larger context (considering multiple flanking nucleotides) but did not detect any additional signature.

Mutation rates in exonic regions. We annotated all observed and simulated *de novo* mutations with their coding status (exonic, intronic or intergenic) using the UCSC CCDS track[42]. We used a χ^2 test to investigate differences in the numbers of observed and simulated mutations for exonic and non-exonic regions and found a 28% enrichment of observed mutations in exonic regions ($P = 0.008$). When considering non-CpG sites only, there was no significant difference ($P = 1.0$). Using a bootstrapping approach (randomly removing mutations regardless of their coding status), we computed that we would have 83.5% power to detect the above enrichment when removing exonic mutations if this enrichment were present.

Mutation rates in DNase I–hypersensitive sites. Using ENCODE[40] measurements of DHSs, we defined a set of conserved peaks present in at least two cell types and annotated all observed and simulated mutations as within or outside a DHS peak (DHSstatus was 0 if the mutation was outside a DHS peak or 1 if it was within a DHS peak). We performed logistic regression with a dummy variable, DNMstatus (*de novo* mutation status; 0 for simulated mutations and 1 for observed mutations), as the response variable and distance to DHSs as the explanatory variable. Under this model, DHSstatus was significantly associated with DNMstatus ($\beta = 0.11$; $P = 0.0041$). When a CpG covariate (0 for mutations outside CpGs and 1 for those at CpGs) was added to the model, CpG status was strongly associated with DNMstatus ($\beta = 2.74$; $P < 2 \times 10^{-16}$) and the distance to DHSs was no longer significant ($\beta = 0.038$; $P = 0.34$).

Influence of mutation and recombination rates on nucleotide diversity. We annotated all observed and simulated mutations with recombination rates from the DECODE recombination map[31]. We compared the distributions of recombination rates at mutation sites for observed and simulated mutations using a logistic regression model with a dummy variable, DNMstatus (0 for observed mutations and 1 for simulated mutation), as the response variable and recombination rate, correcting for CpGs and GC content (1 kb up- and downstream of the mutation site, as the explanatory variable. We found a significant positive association between recombination rates and DNMstatus ($\beta = 0.01$; $P = 0.0015$).

We then computed the nucleotide diversity (π) in the 10-kb region centered on each observed or simulated mutation using VCFTools[43]. We ran a linear regression model with π as the response variable and DNMstatus and recombination rate as explanatory variables. We found that, under this model, both DNMstatus ($\beta = 3.1 \times 10^{-4}$; $P < 2 \times 10^{-16}$) and recombination rate ($\beta = 7.64 \times 10^{-6}$; $P < 2 \times 10^{-16}$) were independently associated with π. We repeated the analysis with π computed for the 100-kb region centered on each mutation site and found similar results. Correcting for local GC content (computed over the same region as π) and CpG status also did not influence these associations.

Influence of mutation and recombination rates on human-chimpanzee divergence. To estimate the influence of mutation rates on human-chimpanzee divergence, we studied the correlations between a human-chimpazee comparative genomics (HCCG) model and (i) the mutation rates observed for 11,020 *de novo* mutations, (ii) the substitution rates obtained by sampling 11,020 mutations on the basis of the HCCG model and (iii) the substitution rates obtained by sampling 11,020 substitution on the basis of a null context-dependent mutation rate model. All rate computations were performed for 1-Mb non-overlapping regions. Because our power to call *de novo* mutation varied along the genome, the mutation rate for each region was corrected for the average fraction of high-confidence bases per trio in that region (determined from simulations).

The HCCG substitution model was computed using genome-wide three-sequence alignments derived from the Pecan 10 amniotes multiple-sequence alignments available at the Ensembl database (version 56)[44,45] (restricted to human, chimpanzee and macaque; excluding exons and CpG islands). We inferred substitution rates $r_{t,i}$ for each substitution type t in {A>G, A>C, A>T, C>A, C>G, C>T, CpG>TpG (to account for the hypermutability of CpG sites)} in each region i using a maximum likelihood–based method as described elsewhere[30]. We assumed the rates of complementary substitution processes to be equal but did not assume that the substitution process was time reversible.

We next computed the total substitution rate per window using the rate inferred above for all bases b in {A, C} using the following formula

$$n_{b,i}\sum_{t_b} r_{t_b,i} + 2n_{\mathrm{CpG},i}r_{\mathrm{CpG>TpG},i}$$

where $n_{b,i}$ is the number of high-confidence bases b in window i and t_b is the set of substitutions in t where the ancestral base is b (for example, for b = A, t_b = {A>C, A>G, A>T}).

The genome-wide averaged model was computed assuming a uniform substitution rate matrix, defined as the mean of the substitution rate matrices over all 1-Mb regions. We then applied the same procedure as above to obtain a uniform rate but context-dependent substitution model.

Using the above HCCG substitution rate model and the genome-wide averaged model, we drew 100,000 genomic profiles of 11,020 mutations (the number of observed *de novo* mutations) using the Poisson random number generator in R[46]. For each of these simulated substitution profiles, as well as for the observed *de novo* mutation rates, we computed the correlation with the HCCG model (**Fig. 4**).

To investigate the effect of local sex-averaged recombination rates[31] on the HCCG model, we computed the substitution rate s_i for each region i using the following Poisson regression with both local recombination rate ρ_i and observed mutation count n_i

$$\log(s_i) = \beta_0 + \beta_\rho \rho_i + \beta_n n_i$$

We then computed the Pearson correlation between the HCCG model and the above Poisson regression.

Strand asymmetry in transcribed regions. Using the UCSC CCDS track, we annotated all genic *de novo* mutations with the direction of transcription. We annotated each *de novo* mutation with its corresponding strand-dependent substitution type (A>G, G>A, A>C, A>T, C>A or C>G). We used χ^2 tests to evaluate strand differences for each substitution type in transcribed regions (**Supplementary Fig. 6**).

We used a χ^2 test to compare observed and simulated mutation counts in intronic and intergenic regions and found a modest 2.8% depletion for observed mutations in intronic regions ($P = 0.047$).

Trinucleotide context dependency. We used the context-dependent substitution matrix from fixed differences in human, chimpanzee and baboon (from multiple-sequence alignments of the three species) available on the UCSC Genome Browser[42,47,48] to empirically calculate the 'directed' 64 × 3 mutation matrix using the model implemented in SCONE[33]. We accounted for multiple mutational events, restricted our analysis to non-exonic regions and removed CpG islands. We then computed the same mutation matrix on the basis of *de novo* mutations and computed the Pearson correlation coefficient for the two matrices ($r^2 = 0.993$).

doi:10.1038/ng.3292

Mutation rate map. Although our set of *de novo* mutations is the largest available thus far, it is still a relatively sparse sampling across the genome. For this reason, we decided to use a human-chimpanzee-macaque primate substitution model[30] and refined it using our observed mutations.

The local mutation rates derived from a human-chimpanzee-macaque primate substitution model[30] were corrected for biases due to local recombination rate[31], the type of mutation and the direction of transcription along the strand. This correction was applied for 2,339 1-Mb non-overlapping windows across the autosomes after excluding windows where (i) the sex-averaged recombination rate was unavailable for more than 10% of the window, (ii) the sex-averaged recombination rate across the window was 0 or greater than 3 cM/Mb, (iii) the primate local substitution rate was estimated to be 0 or extremely high or (iv) more than 800 kb on average were below our high-confidence calling threshold for each trio (as determined by simulations). All mathematical details and computed correction factors are described in the **Supplementary Note** and in **Supplementary Tables 2–4**.

35. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
36. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
37. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
38. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
39. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
40. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
41. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).
42. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
43. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
44. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
45. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
46. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
47. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
48. Murphy, W.J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).