nature genetics

Variation in genome-wide mutation rates within and between human families

Donald F Conrad[1,2], Jonathan E M Keebler[3,4], Mark A DePristo[5], Sarah J Lindsay[1], Yujun Zhang[1], Ferran Casals[3], Youssef Idaghdour[3], Chris L Hartl[5], Carlos Torroja[1], Kiran V Garimella[5], Martine Zilversmit[3], Reed Cartwright[6], Guy A Rouleau[7], Mark Daly[5], Eric A Stone[4,6], Matthew E Hurles[1] & Philip Awadalla[3] for the 1000 Genomes project[8]

J.B.S. Haldane proposed in 1947 that the male germline may be more mutagenic than the female germline[1]. Diverse studies have supported Haldane's contention of a higher average mutation rate in the male germline in a variety of mammals, including humans[2,3]. Here we present, to our knowledge, the first direct comparative analysis of male and female germline mutation rates from the complete genome sequences of two parent-offspring trios. Through extensive validation, we identified 49 and 35 germline *de novo* mutations (DNMs) in two trio offspring, as well as 1,586 non-germline DNMs arising either somatically or in the cell lines from which the DNA was derived. Most strikingly, in one family, we observed that 92% of germline DNMs were from the paternal germline, whereas, in contrast, in the other family, 64% of DNMs were from the maternal germline. These observations suggest considerable variation in mutation rates within and between families.

Mutation underlies all heritable genetic variation, and the observation that a mutation has arisen *de novo* can be highly discriminating in identifying causal pathogenic variation in individuals[4–6]. Attempts to measure mutation rates in humans fall into two broad categories: direct methods that estimate the number of mutations that have occurred in a known number of generations[7,8] and indirect methods that infer mutation rates from levels of genetic variation within or between species. Previous estimates of germline base substitution rates range from 1.1 to 3×10^{-8} per base per generation[7,9–14]. This variation is caused, in part, by uncertainty or assumptions in key parameters, such as divergence times between species, generation times and ancestral population sizes. Furthermore, all previous estimates represent an average across multiple generations and/or an average of male and female mutation rates. Consequently, the previous studies provided no information on how mutation rates vary between individuals of either the same or different sexes or between gametes within an individual. It has been proposed that the mammalian male germline may be more mutagenic than the female germline because of the greater number of cell divisions in the former[1]. Subsequent studies[2,3] based on whole-genome sequences of human and chimpanzee have indicated a sixfold difference in mutation rate between male and female germlines, averaged across ~5–7 million years of independent evolution of the two lineages.

High-throughput sequencing enables whole-genome analysis of mutation rates in human pedigrees[7] and promises to revolutionize our understanding of how mutation rates vary between sexes, individuals and families. We analyzed lymphoblastoid cell lines from two parent-offspring trios (from the HapMap CEU and YRI populations) sequenced genome wide to greater than a 22-fold mapped depth using three different sequencing platforms during the pilot phase of the 1000 Genomes Project[15] (Online Methods). We developed three independent probabilistic algorithms to identify candidate DNMs from these sequence data (**Supplementary Note**). From the union of candidate DNMs identified by the three algorithms, we selected 3,236 and 2,750 potential DNMs for experimental validation from the CEU and YRI trios, respectively, which is far in excess of the expected number of true germline DNMs, to maximize our sensitivity to detect DNMs.

We attempted validation of every candidate DNM using two new experimental approaches and additional resources from each family to unambiguously distinguish germline DNMs from somatic or cell-line DNMs (Online Methods, **Fig. 1** and **Supplementary Tables 1–3**). For the CEU trio, we performed these validation experiments on LCL-derived DNA from both the original trio and a third generation from the same family. For the YRI trio, we performed these validation experiments on LCL-derived DNA from the trio as well as on whole-genome–amplified blood-derived DNA from the same individuals. Using these validation data, we classified each putative DNM into one of five categories: (i) germline DNM, (ii) non-germline (somatic or arising in cell culture) DNM, (iii) inherited variant, (iv) false positive

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [2]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA. [3]Ste Justine Hospital Research Centre, Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal, Quebec, Canada. [4]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA. [5]Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [6]Department of Genetics, North Carolina State University, Raleigh, North Carolina, USA. [7]Ste Justine Hospital Research Centre, Department of Medicine, Faculty of Medicine, University of Montreal, Montreal, Quebec, Canada. [8]A full list of members is provided in the **Supplementary Note**. Correspondence should be addressed to P.A. (philip.awadalla@umontreal.ca) or M.H. (meh@sanger.ac.uk).

Received 14 October 2010; accepted 19 May 2011; published online 12 June 2011; doi:10.1038/ng.862

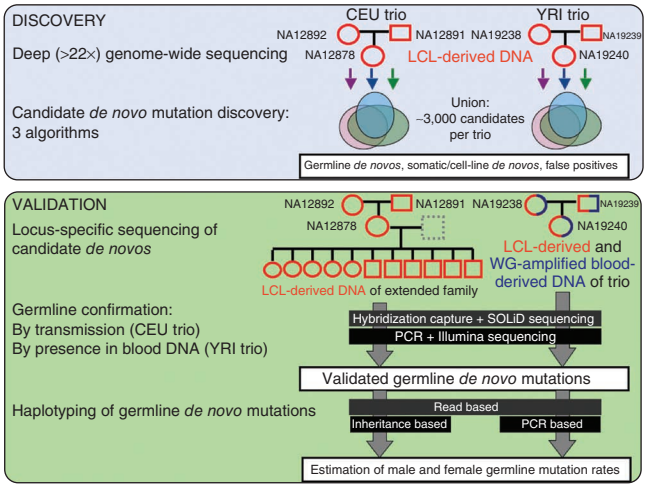

Figure 1 Overview of the study design. The two phases of the project, discovery and validation, are shown schematically, including the samples from each family that were used in each phase. LCL, lymphoblastoid cell line; WG, whole genome.

or (v) inconclusive (**Table 1**, **Supplementary Note**, **Supplementary Table 1** and **Supplementary Figs. 1**,**2**). We identified 49 and 35 germline DNMs and 952 and 643 non-germline DNMs in the CEU and YRI trios, respectively. The observed ~20:1 ratio of non-germline DNMs to germline DNMs was substantially larger than the 1:1 ratio published previously[4]. This difference could be caused by the age of the cell lines (number of passages), the mutagenicity of the cell culture conditions and/or the clonality of the cell lines. We observed differences in the mutational characteristics of germline DNMs, non-germline DNMs and inherited germline variants in terms of the ratio of transitions and transversions, the proportion of CpG mutations, the clonality of mutations, their occurrence at sites under selective constraint and the evidence for transcription-coupled repair (**Table 1**, **Supplementary Note** and **Supplementary Figs. 3**,**4**).

By estimating the false negative rates in the discovery and validation of DNMs and quantifying the proportion of the genome that we were able to scrutinize reliably for DNMs (**Supplementary Note**), we estimated the germline DNM rates in each trio to be 1.17×10^{-8} (95% CI 0.88×10^{-8} to 1.62×10^{-8}) and 0.97×10^{-8} (95% CI 0.67×10^{-8} to 1.34×10^{-8}) for the CEU and YRI trios, respectively. The sex-averaged germline mutation rate estimates we derived agree very closely with three other recent studies focusing on sex-averaged mutation rates in the most recent generation[4,7,13]. Averaging across these four studies gave a more precise sex-averaged mutation rate of 1.18×10^{-8} ($\pm 0.15 \times 10^{-8}$ (s.d.)), which is less than half of the frequently cited sex-averaged mutation rate derived from the human-chimpanzee sequence divergence of 2.5×10^{-8} (ref. 14). These apparently discordant estimates can be largely reconciled if the age of the human-chimpanzee divergence is pushed back to 7 million years, as suggested by some interpretations of recent fossil finds[16], and by considering more recent (and slightly lower) robust genome-wide estimates of sequence divergence[17]. These considerations suggest a plausible range for the divergence-derived mutation rate of 1.12×10^{-8} to 2.05×10^{-8}, which encompasses the averaged contemporary mutation rate above. Moreover, by considering that the distribution of mutation rates in the population could contain a long tail of relatively rare individuals with considerably higher mutation rates (perhaps as a result of genetic or environmental factors), it can be appreciated that the mean rate across many generations could be considerably greater than the modal rate within a generation.

We ascertained for most germline DNMs whether they arose on a paternal or maternal haplotype using three alternative methods (Online Methods, **Supplementary Note** and **Supplementary Table 1**). Where more than one haplotyping method could be applied to the same DNM ($N = 17$), the results were 100% concordant. Male and female germline mutation rates in the two trios (**Fig. 2**) were significantly different ($P < 3 \times 10^{-6}$, Fisher exact test). In one family, 92% of the germline DNMs were from the paternal germline, whereas, in the other family, only 36% of the DNMs were paternal in origin. Although the confidence intervals of some of the parent-specific rates

Table 1 Mutational properties of different classes of validated sites

	Germline DNMs	Non-germline DNMs	False positives	Inherited variant	No call
	Total (CEU/YRI)	Total (CEU/YRI)	Total (CEU/YRI)	Total (CEU/YRI)	Total (CEU/YRI)
Count	84 (49/35)	1,586 (952/634)	2,360 (1,304/1,065)	464 (129/335)	1,483 (802/681)
Posterior prob					
FIGL	0.98 (1/0.96)	0.95 (0.96/0.94)	0.87 (0.9/0.83)	0.78 (0.79/0.77)	0.83 (0.86/0.78)
FPIR	0.96 (0.97/0.96)	0.9 (0.93/0.85)	0.63 (0.72/0.52)	0.38 (0.46/0.35)	0.59 (0.68/0.47)
SIMTG	13.38 (13.3/13.55)	9.76 (9.68/9.91)	10.75 (10.42/11.17)	11.16 (10.28/11.59)	11.06 (10.97/11.2)
Ts:Tv	2.82 (2.5/3.38)	0.98 (0.92/1.06)	0.76 (0.8/0.72)	1.83 (1.43/2.02)	1.1 (0.99/1.25)
CpG	0.13 (0.14/0.11)	0.07 (0.09/0.06)	0.06 (0.05/0.06)	0.13 (0.14/0.13)	0.07 (0.06/0.1)
GERP	−0.3 (−0.28/−0.32)	−0.18 (−0.22/−0.11)	−0.21 (−0.2/−0.22)	−0.25 (−0.45/−0.18)	−0.22 (−0.25/−0.18)
Function					
Coding-missense	0 (0/0)	16 (6/10)	15 (10/5)	2 (1/1)	0 (0/0)
Coding-synonymous	1 (1/0)	1 (0/1)	9 (8/1)	0 (0/0)	4 (1/3)
NMD_transcript	0 (0/0)	2 (1/1)	7 (4/3)	3 (0/3)	4 (0/4)
Splice site	0 (0/0)	3 (3/0)	0 (0/0)	0 (0/0)	0 (0/0)
5′ UTR	0 (0/0)	3 (3/0)	4 (0/4)	0 (0/0)	4 (2/2)
3′ UTR	0 (0/0)	5 (4/1)	20 (14/6)	3 (2/1)	11 (6/5)
Non-coding gene	5 (4/1)	122 (75/46)	197 (96/102)	36 (11/25)	100 (54/46)
Intronic	36 (19/17)	521 (319/197)	806 (437/374)	177 (54/123)	552 (305/247)
Intergenic	42 (25/17)	925 (544/378)	1,302 (735/570)	243 (61/182)	808 (434/374)

Posterior prob, average support metric (posterior probability or $\log_{10}$ odds score) of calls reported by each discovery method; Ts:Tv, ratio of transitions to transversions; CpG, proportion of sites within CpG; GERP, average genomic evolutionary rate profiling (GERP) score; no call, insufficiently informative data to make a high confidence call; UTR, untranslated region; FIGL, family-aware Illumina genotype-likelihood–based method; FPIR, family-aware probabilistic Illumina-read–based method; SIMTG, sample-independent multiple technology genotype–based method.

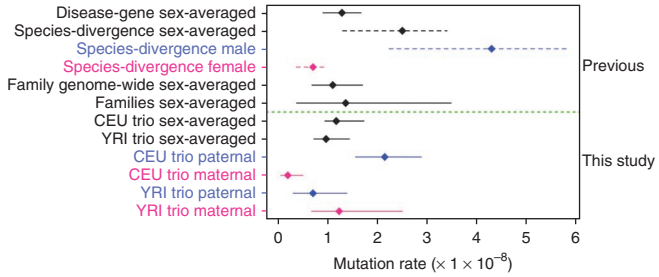

Figure 2 Comparison of mutation rate estimates. Mutation rates estimated from previous studies are shown above the dashed green line. Solid lines encompassing point estimates represent 95% confidence intervals. Dashed lines encompassing point estimates represent reported plausible ranges. The disease-gene sex-averaged rate comes from reference 13, with 95% confidence intervals calculated as 1.96 times the standard error. The species-divergence sex-averaged rate comes from reference 14, which specifies the plausible range shown here. The species-divergence sex-specific rates come from scaling the sex-averaged point estimate and the upper and lower bounds by the ratio of male to female mutation rate of 6.11 estimated in reference 3. The family genome-wide sex-average comes from reference 7 and the families sex-average comes from reference 4.

overlap, the paternal rates in the two trios do not overlap and neither do the maternal rates. These differences could be caused by extensive variation in the number of DNMs in gametes from the same individual or by considerable variation between individuals in their underlying DNM rate. With only a single offspring per family, we could not distinguish between these two alternatives, but either would give rise to substantial variation in the number of DNMs between offspring of different families. The potential scale of this variation can be appreciated by simply considering that exchanging the paternal gamete in the CEU trio for that in the YRI trio would have resulted in a fivefold difference in the number of mutations seen in the two offspring.

Some of this variation in mutation rates between families might be explained by differences in parental ages and a dependency of mutation rate on age. Unfortunately, parental ages at conception for these two trios were not available, but nevertheless, the analysis of larger sibling relationships would be required to disentangle fully the effects of parental age from genetic and environmental factors that might also differ between families. Variation in mutation rates between individuals could also be partly explained by a recent relaxation of the selective constraint on mutation rates resulting from the lower efficiency of selection in humans as compared to the most recent common ancestor of humans and chimpanzees[16] because of our small effective population size[17]. Mutation is a random process and, as a result, considerable variation in the numbers of mutations is to be expected between contemporaneous gametes within an individual. If modeled as a Poisson process, the 95% confidence intervals on a mean of ~30 DNMs per gamete (as expected from a mutation rate of $\sim 1 \times 10^{-8}$) ranges from 20 to 41, which is a twofold difference. Truncating selection might act to remove the most mutated gametes and thus reduce this variation among gametes that successfully reproduce, however, any additional heterogeneity in stem-cell ancestry or environment (for example, variation in the number of cell divisions leading to contemporaneous gametes) would likely increase inter-gamete variation in the number of mutations.

In summary, whereas there may be growing concordance in the estimates of the average mutation rates in contemporary generations, we present evidence of substantial variance in sex-specific mutation rates between families. The variation in mutation rates that we observed is of potential clinical importance, as it suggests that the risk of misdiagnosing a DNM as being pathogenic could vary substantially between individuals.

Advances in sequencing technologies that lower costs and increase fidelity (**Supplementary Note**) will empower further studies of mutational processes by applying the framework we have established here for estimating sex-specific mutation rates in families. These future studies promise to revolutionize our understanding of mutation processes and how they vary between individuals and between families as a result of age, genetic background and environmental exposures.

URLs. UCSC Genome Browser LiftOver web tool, http://genome.ucsc.edu/cgi-bin/hgLiftOver.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank A. Kernytsky, G. McVean, T. Massingham, J. Thorne, J. Hussin, A. Motsinger, Coriell Cell Repositories and members of the 1000 Genomes analysis group for their help and support. D.F.C., S.J.L., Y.Z., C.T. and M.E.H. were funded by the Wellcome Trust (grant number 077014/Z/05/Z). J.E.M.K., F.C., Y.I., M.Z., G.A.R. and P.A. were funded by the Ministry of Development, Exploration and Innovation (grant number PSR-SIIRI-195) in Quebec and a Genome Quebec Award for Population and Medical Genomics to P.A.

AUTHOR CONTRIBUTIONS

M.E.H. and P.A. conceived the study. D.F.C., J.E.M.K., M.A.D., M.D., R.C., E.A.S. and P.A. developed statistical methodologies. D.F.C., J.E.M.K., M.A.D., C.L.H., K.V.G., E.A.S., M.E.H. and P.A. analyzed the data. F.C., Y.I., G.A.R., C.T., M.Z., S.J.L. and Y.Z. generated validation data. D.F.C., P.A. and M.E.H. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Haldane, J.B.S. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann. Eugen.* **13**, 262–271 (1947).
2. Bohossian, H.B., Skaletsky, H. & Page, D.C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406**, 622–625 (2000).
3. Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F. & Makova, K.D. 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. Awadalla, P. *et al.* Direct measure of the *de novo* mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* **87**, 316–324 (2011).
5. Gauthier, J. *et al. De novo* mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. *Proc. Natl. Acad. Sci. USA* **107**, 7863–7868 (2010).
6. Lee, C., Iafrate, A.J. & Brothman, A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat. Genet.* **39**, S48–S54 (2007).
7. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
8. Xue, Y. *et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.* **19**, 1453–1457 (2009).
9. Crow, J.F. How much do we know about spontaneous human mutation rates? *Environ. Mol. Mutagen.* **21**, 122–129 (1993).
10. Haldane, J.B.S. The rate of spontaneous mutation of a human gene. *J. Genet.* **31**, 317–326 (1935).
11. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
12. Kondrashov, A.S. & Crow, J.F. A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.* **2**, 229–234 (1993).
13. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
14. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
15. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
16. Chen, F.C. & Li, W.H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
17. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).

ONLINE METHODS

Deep genome-wide sequencing. Each trio was sequenced using Illumina, 454 and SOLID technologies, as described elsewhere[15]. The total mapped sequence coverage from all platforms (or from Illumina only, in parentheses), was 22.9 (22.9), 28.4 (28.4) and 69 (36.6) reads per base for NA19238, NA19239 and NA19240, respectively, and 24.5 (24.5), 28.2 (28.2) and 66.8 (35.0) reads per base for NA12891, NA12892 and NA12878, respectively.

Candidate *de novo* mutation discovery. Three different algorithms, the family-aware probabilistic Illumina-read–based method, the family-aware Illumina genotype-likelihood–based method and the sample-independent multiple technology genotype–based method, were developed for DNM discovery (**Supplementary Note**). The first two algorithms are probabilistic approaches that use the Illumina data and jointly analyze data from all three family members simultaneously, although they differ in their underlying statistical methodology. The third approach considers data from each sample independently but jointly considers data from all three sequencing platforms. The subsequent genotypes of family members were compared to identify apparent new alleles present in the offspring, and the three genotype confidence values were summarized to rank candidate DNMs.

Filtering. Filters were applied to exclude genomic regions in which we expected that false positive DNM calls might be enriched (**Supplementary Note**). These filters were derived from genome annotations, sites of known polymorphisms and properties of the pilot 2 data, the union of which covers approximately 470 Mb of sequence in each trio. As calling was not attempted in these regions, we are agnostic about the nature of DNMs in these sites.

Experimental validation. Two independent validation experiments were performed on the cell-line DNA from each member of each trio as well as on cell-line–derived DNA from the 11 offspring of the CEU trio offspring and blood-derived DNA from the YRI trio members. The first experiment comprised nested PCR amplification of putative DNMs followed by read-pair sequencing of pooled PCR products on the Illumina platform. The second experiment comprised hybridization capture of putative DNMs using Agilent SureSelect technology followed by sequencing on the SOLID3 platform. The validation data were analyzed jointly using a mixture model framework to classify each site into one of four categories: (i) germline DNM, (ii) somatic or cell-line DNM, (iii) inherited variant or (iv) false positive, and any site that could not be confidently assigned into any one of the four classes was defined 'inconclusive'. Raw results of this validation experiment are shown in **Supplementary Table 1**.

Haplotyping *de novo* mutations. Validated germline DNMs were assigned to parental haplotypes by (i) inspection of the phase information inherent in the original sequencing data, (ii) using linkage of the DNM to nearby variants in the third generation of the CEU trio and (iii) PCR co-amplification of the DNM and the nearest informative heterozygous site, followed by cloning and end sequencing of multiple clones per DNM (**Supplementary Table 1**).

Rate estimation. The sex-average mutation rate for each trio was estimated by correcting for the false negative rate in the DNM discovery by the combination of the three DNM discovery algorithms, correcting for the false negative rate in the DNM validation (that is, the proportion of putative DNMs that were classified as being inconclusive after validation) and dividing by the number of bases that passed the genome filters, and thus, had been scrutinized for DNMs. Uncertainty is estimated from the Poisson confidence intervals on the number of DNMs observed. The sex-specific rates were estimated by scaling the sex-average rates by the proportion of haplotyped DNMs that were ascribed to paternal and maternal gemlines, with the uncertainty being estimated from the Poisson confidence intervals on the numbers of haplotyped DNMs ascribed to either parental germline.

Functional constraints. GERP scores measure conservation as the difference between the expected and observed rates of nucleotide substitution at a given human base[18]. GERP scores are position specific and were estimated from aligned orthologous sequences, in this case from genomic alignments of 16 amniote species in Ensembl 58 (Compara.16_amniota_vertebrates_Pecan). GERP scores were extracted for sites of polymorphism identified by the 1000 Genomes trio pilot, as well as sites from the combined candidate *de novo* mutation lists from the CEU and YRI families (**Supplementary Table 1**). The positions of these sites were converted from the hg18 coordinate system to hg19 using the LiftOver web tool available at the UCSC Genome Browser website (see URLs).

18. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).