nature genetics

Estimating the human mutation rate using autozygosity in a founder population

Catarina D Campbell[1], Jessica X Chong[2], Maika Malig[1], Arthur Ko[1], Beth L Dumont[1], Lide Han[2], Laura Vives[1], Brian J O'Roak[1], Peter H Sudmant[1], Jay Shendure[1], Mark Abney[2], Carole Ober[2,3] & Evan E Eichler[1,4]

Knowledge of the rate and pattern of new mutation is critical to the understanding of human disease and evolution. We used extensive autozygosity in a genealogically well-defined population of Hutterites to estimate the human sequence mutation rate over multiple generations. We sequenced whole genomes from 5 parent-offspring trios and identified 44 segments of autozygosity. Using the number of meioses separating each pair of autozygous alleles and the 72 validated heterozygous single-nucleotide variants (SNVs) from 512 Mb of autozygous DNA, we obtained an SNV mutation rate of 1.20×10^{-8} (95% confidence interval $0.89–1.43 \times 10^{-8}$) mutations per base pair per generation. The mutation rate for bases within CpG dinucleotides (9.72×10^{-8}) was 9.5-fold that of non-CpG bases, and there was strong evidence ($P = 2.67 \times 10^{-4}$) for a paternal bias in the origin of new mutations (85% paternal). We observed a non-uniform distribution of heterozygous SNVs (both newly identified and known) in the autozygous segments ($P = 0.001$), which is suggestive of mutational hotspots or sites of long-range gene conversion.

Various approaches have provided a wide range of SNV mutation rate estimates ($1–3 \times 10^{-8}$ mutations per base pair per generation). Early studies of mutation rates in humans focused on specific loci or the *de novo* incidence of disease[1–4]. More recent studies have leveraged whole-genome sequencing data on a total of three nuclear families to estimate *de novo* mutation rates for SNVs of approximately 1×10^{-8} mutations per base pair per generation[5,6]. Comparative studies of chimpanzee and human genomes provided higher estimates (for instance, 2.5×10^{-8}) but are highly contingent on uncertainty about the number of generations since human-chimpanzee divergence[7].

In contrast to studies that are focused on identifying new mutations arising in a single generation, the examination of populations with a small number of founding individuals is ideal for estimating mutation rates across a small number of generations. The Hutterites are a population of Anabaptist farmers living on the plains of the United States and Canada who are descended from a small group of founders (<90 individuals). The genealogy of this group is completely known, and genome-wide SNP genotype data have been collected from over 1,400 individuals who are related to each other in a 13-generation pedigree descended from 64 founders[8,9]. Due to increased levels of consanguinity, Hutterite individuals carry large segments of the genome that are autozygous or homozygous by recent decent[10]. The alleles in an autozygous segment are descended from a recent common ancestor and have accumulated mutations in the generations since transmission from this individual.

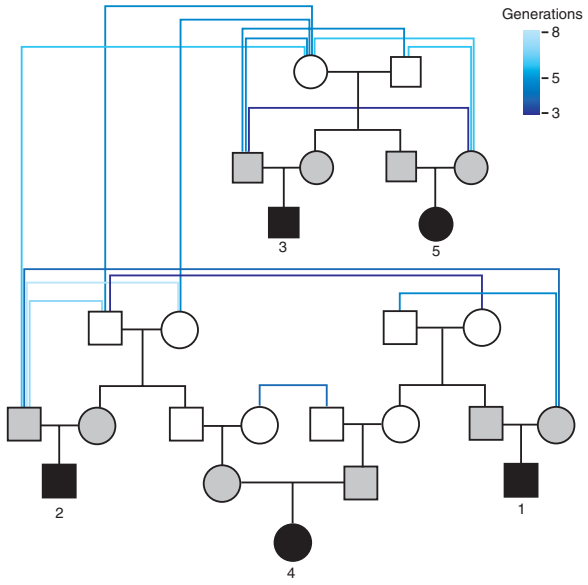

Figure 1 Relationship of sequenced individuals. Simplified pedigree showing the relationship between the 15 sequenced individuals. Black symbols represent the children in the five trios, and gray symbols represent their parents. Founders are connected by blue lines, with the shade of blue indicating the number of generations separating the connected individuals. For clarity, only the shortest relationships between each individual and the parents of that individual are shown. The color scale represents the number of generations separating the individuals, where darker blue indicates fewer generations and lighter blue indicates more generations.

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Human Genetics, The University of Chicago, Chicago, Illinois, USA. [3]Department of Obstetrics and Gynecology, The University of Chicago, Chicago, Illinois, USA. [4]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 6 June; accepted 30 August; published online 23 September 2012; doi:10.1038/ng.2418

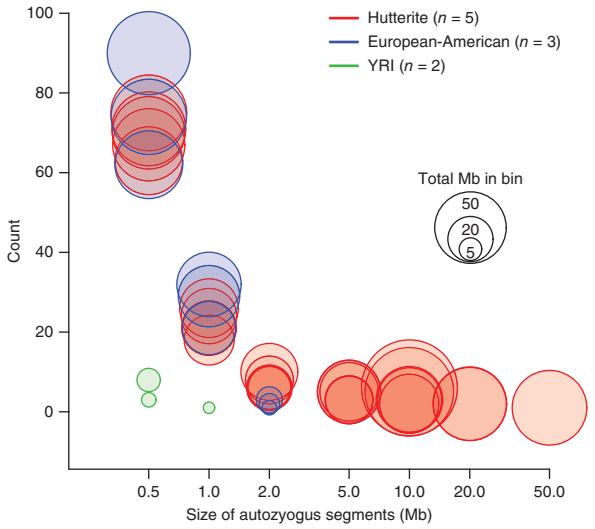

Figure 2 Elevated autozygosity in the Hutterite individuals. Autozygous segments were binned by size for the five Hutterite individuals, three European-American individuals and two Yoruba individuals (YRI). The *x* axis represents bins of autozygous segments of different size, and the *y* axis shows the number of segments in each bin. In each bin, individuals are represented by 'bubbles', with the size of the bubble denoting the total amount of genomic sequence in that bin.

We selected five Hutterite parent-offspring trios for whole-genome sequencing, with the parents in each trio being related to each other by 6–8 (mean of 6.6) meiotic transmissions (**Fig. 1**). We performed whole-genome sequencing of DNA isolated from whole blood using Illumina paired-end sequencing, generating 775 Gb of sequence with an average of 13-fold coverage per individual (**Supplementary Table 1**). The sequencing reads for each sample were aligned to the human reference genome (NCBI Build 36). We identified a total of 5.4 million SNVs on the basis of the intersection of variant calls from 2 different algorithms[11,12] (**Supplementary Table 2**). The SNP genotypes from whole-genome sequencing were highly concordant to those generated by SNP microarray (mean genotype concordance of 99.7%) (**Supplementary Table 2**).

We identified extended regions of homozygosity in the offspring of the five trios and in five previously sequenced genomes (three European-Americans and two Yoruba)[13] (Online Methods). The extent of homozygosity was correlated to the inbreeding coefficients of the Hutterite individuals (**Supplementary Fig. 1**, **Supplementary Table 3** and **Supplementary Note**). As expected, the five Hutterite probands showed significantly greater autozygosity (223 Mb on average per individual) than other European-American individuals (95 Mb) or the Yoruba individuals (4 Mb) (**Fig. 2** and **Supplementary Table 3**). Although the amount of short homozygous segments was

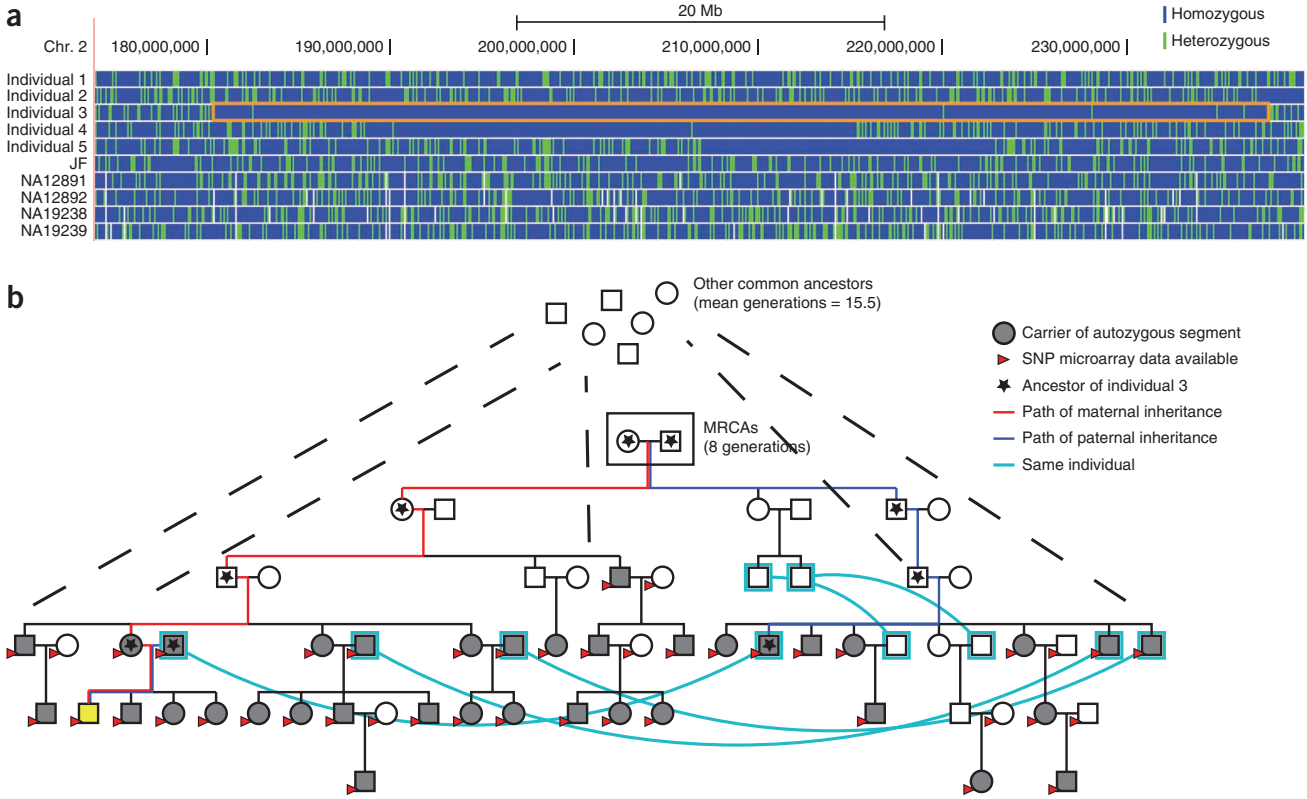

Figure 3 Determination of the MRCA for an autozygous segment. (**a**) A 54-Mb autozygous segment on chromosome 2 in individual 3. Genomic coordinates (hg18) are given on the horizontal axis, and each individual is represented on the vertical axis, including the five Hutterite individuals, the three European-Americans and the two Yoruba. Each SNV is represented by a vertical bar that is colored blue if the variant is homozygous and green if it is heterozygous. The autozygous segment in individual 3 is boxed in orange. (**b**) Determination of the MRCA for the autozygous segment in individual 3. The pedigree containing all the haplotype carriers of the autozygous haplotype is shown. Individual 3 is shown in yellow. Haplotype carriers have two MRCAs (boxed) as well as additional common ancestors further up the pedigree. The paths from these individuals to the autozygous subject are shown in red for the maternal ancestors and blue for the paternal ancestors; all ancestors of the individual are marked with a star. Black dashed lines represent relationships to common ancestors further back in the pedigree.

Table 1 SNV mutation rates determined from segments of autozygosity

Individual	Segments (>5 Mb)	Total callable (Mb)[a]	Mean meioses (MRCA)[b]	SNVs[c]	SNV μ	95% CI[d]
1	7	63.4	13.8	13	1.51×10^{-8}	$0.62–2.28 \times 10^{-8}$
2	6	55.9	13.8	7	0.92×10^{-8}	$0.17–1.72 \times 10^{-8}$
3	9	124.8	9.9	13	1.07×10^{-8}	$0.45–1.63 \times 10^{-8}$
4	10	147.6	12.0	19	1.09×10^{-8}	$0.56–1.55 \times 10^{-8}$
5	12	120.8	12.0	20	1.40×10^{-8}	$0.73–1.96 \times 10^{-8}$
All	44	512.4	11.9	72	1.20×10^{-8}	$0.89–1.43 \times 10^{-8}$

[a]Non-segmental duplication, non-simple repeat and non-dbSNP132 variants with at least six mapped reads. [b]Weighted by length of segment. [c]Validated as newly identified, heterozygous variants. [d]Based on a Poisson distribution.

similar in the Hutterite individuals and the other European-American individuals, we observed 33-fold more autozygous base pairs in segments of greater than 2 Mb in length in the Hutterite individuals (**Fig. 2** and **Supplementary Note**). We further refined and validated segments that were longer than 5 Mb by comparing to autozygous segments identified in SNP microarray data for the same samples[10,14] to obtain a final list of 44 regions of autozygosity (6–12 segments per individual; 5–54 Mb in length). We restricted subsequent analyses to these 512 Mb of autozygous DNA (**Supplementary Fig. 2** and **Supplementary Table 4**).

We determined the number of meioses separating each allele within each autozygous segment. The small founding population and complex genealogy of the Hutterite population (**Fig. 1**) made this potentially problematic because of the large number of shared common ancestors and multiple paths of descent between any ancestor-descendant pair. To resolve the ancestry of the autozygous segments, we combined the pedigree structure and genome-wide SNP genotype data[9] to identify the most recent common ancestors (MRCAs) on the basis of segregation within the Hutterite genealogy[8] (**Fig. 3**, **Supplementary Fig. 3** and **Supplementary Note**). Using the identified MRCAs, we estimated that the 2 haplotypes of the 44 autozygous segments were separated by 8–18 meioses (**Supplementary Table 4**).

To calculate the SNV mutation rate, we identified heterozygous SNVs within each autozygous segment, excluding regions of common repeats, segmental duplication and known SNPs (found in dbSNP132). We validated 72 SNVs as heterozygous by Sanger-based capillary sequencing (**Table 1**, **Supplementary Table 5** and **Supplementary Note**). We calculated an SNV mutation rate (μ) of 1.20×10^{-8} (95% confidence interval (CI) = $0.89–1.43 \times 10^{-8}$) mutations per base pair per generation. We observed consistent μ values across the five trios, with values ranging between 0.92×10^{-8} and 1.51×10^{-8} (**Fig. 4** and **Table 1**). Among these mutations, we observed an excess of transitions relative to transversions, resulting in a Ti/Tv ratio of 1.64 that was not significantly different from the genome-wide SNV ratio of 2.17 (two-tailed χ-squared $P = 0.27$). Twelve of the 72 validated heterozygous SNVs (16.7%) mapped to CpG dinucleotides. We calculated a μ value for CpG sites of 9.72×10^{-8} mutations per CpG base pair per generation, which is 9.5× greater than the μ value for non-CpG bases (1.02×10^{-8}). We also estimated the mutation rate on the basis of *de novo* mutations in the most recent generation (**Supplementary Table 6** and **Supplementary Note**). Using 176 validated *de novo* SNVs, we calculated a mutation rate of 0.96×10^{-8} (95% CI = $0.82–1.09 \times 10^{-8}$) mutations per base pair per generation; although this rate is lower than the one calculated using autozygosity, the confidence intervals of these rates overlap (**Fig. 4**).

We identified and validated one potential gene conversion event involving paralogs of segmental duplications containing the genes *C4A* and *C4B* in a region where lower copy number has been associated with lupus[15]. Although individual 4 had a total diploid copy number of six for this CNV, we determined that the sequence content of the two alleles differed (**Supplementary Fig. 4**), likely as a result of gene conversion between paralogous copies of, at a minimum, the *TNXA* and *TNXB* genes (6 kb).

Both theoretical and experimental analyses have predicted that the male germline contributes disproportionately to *de novo* mutations compared to the female germline[7,16,17]. However, a recent analysis on two parent-offspring trios reported a paternal bias in mutation in one trio and a maternal bias in the other[5]. Given the complexity of the Hutterite pedigree and transmissions through multiple female and male ancestors, we focused on the putative genome-wide *de novo* mutations in the most recent generation. We used molecular phasing[5,12,17] to determine the parental origin of 26 of the 176 validated *de novo* SNVs and found that 84.6% (22 of 26; 95% CI = 70.8–98.5%) of *de novo* SNPs originated on the paternal haplotype, confirming a male bias for new SNVs (two-tailed binomial $P = 2.67 \times 10^{-4}$).

One advantage of using autozygosity in the identification of recent mutations is the ability to identify potential gene conversion events between homologous chromosomes. Such events could lead to clusters of heterozygous SNVs (including known SNPs) within regions of autozygosity, and we identified four clusters (with two or more SNVs mapping within 10 kb of each other) (**Table 2**). One of these clusters is 309 kb in length, suggesting that it most likely arose as a product of crossover events[18]. Excluding this large cluster, the average distance between heterozygous SNPs in the remaining three clusters was 2,723 bp (range of 7–7,839 bp). We tested this distribution by simulation (n = 10,000 replicates) and determined that there was a significant excess of 'clustered' SNVs compared to that expected with a random distribution of variants (empirical $P = 0.001$).

We also tested whether the *de novo* SNVs in the most recent generation were uniformly distributed in the genome. Notably, we observed three clusters of validated *de novo* variants (**Table 2** and **Supplementary Table 5**) and a significant excess (empirical $P = 6 \times 10^{-6}$) of *de novo* SNVs in close proximity (<10 kb) using simulations (n = 1,000,000 replicates).

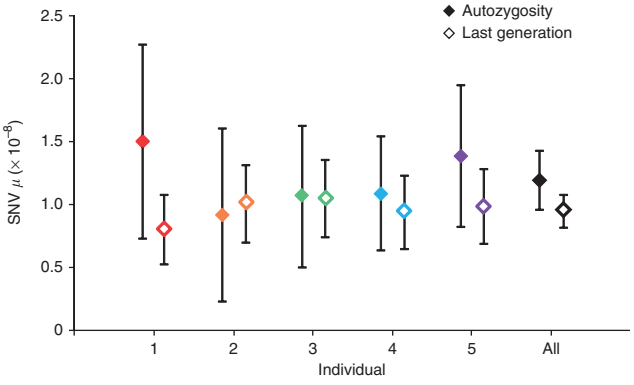

Figure 4 SNV mutation rate estimates. The SNV mutation rate point estimates are shown for each individual and all five individuals combined, with the error bars representing the 95% CIs that were generated on the basis of a Poisson distribution. SNP μ is the number of SNV mutations per base pair per generation. Filled diamonds represent estimates from autozygous segments, and open diamonds represent estimates from SNVs identified in the most recent generation.

Table 2 Clusters of heterozygous SNVs

Individual	Chr.	Start	End	Length (bp)	Heterozygous SNVs (new)[a]	CpG SNVs[b]	GC (%)[c]	Meioses[d]
1	1	4403120	4414313	11,193	3 (1)	0	0.46	15.5
1	1	74906779	74909810	3,031	2 (2)	0	0.36	14.0
5	1	10788610	10793450	4,840	5 (0)	1	0.50	9.0
5	2	211397873	211706890	309,017	66 (1)	8	0.35	9.0
1	7	45928832	45930883	2,051	2 (2)	0	0.47	2.0
1	16	77252309	77256230	3,921	2 (2)	0	0.43	2.0
2	1	189717619	189717626	7	2 (2)	0	0.31	2.0

The first four clusters are in autozygous segments followed by the three clusters of *de novo* SNVs. Chr., chromosome. [a]Total number of SNVs in the cluster with the number of new (non-dbSNP132) SNVs in parentheses. [b]SNVs in CpG dinucleotides. [c]Percentage of G and C bases in the heterozygous cluster. [d]Number of meioses in which event(s) occurred, calculated based on the MRCA.

There has recently been much interest in using massively parallel sequencing data to obtain an accurate estimate of the mutation rate using nuclear families[5,6]. We developed an approach using extended regions of autozygosity to discover new mutations that have emerged within a few generations. Compared to analyses focused on *de novo* mutations in a single generation, our approach significantly reduces the number of false positives and somatic mutations, as most mutations in autozygous segments are transmitted from one of the parents. In addition, given the relationship between paternal age and the number of *de novo* mutations[17,19], our approach reduces this confounding effect by yielding an average mutation rate over 8–18 meioses. Disadvantages include uncertainty about the ancestry of the autozygous segments (**Supplementary Note**), the smaller genomic 'search space' (512 Mb), the potential to confound new mutation and gene conversion events and increased potential for purifying selection to eliminate a small fraction of new mutations, although the fraction of such events should be negligible[20]. We have tried to reduce the confounding effect of gene conversion by limiting our analysis to newly identified SNVs. We estimated an SNV mutation rate of 1.20×10^{-8} mutations per base pair per generation using autozygous segments, which is higher than the rate of 0.96×10^{-8} that we estimated for the most recent generation and the rate of 1.1×10^{-8} that was previously published for the whole genome[5,6] yet lower than the rate estimated in a recent resequencing study[21].

While this manuscript was under review, two additional studies characterizing the human mutation rate were published. First, a sequence mutation rate of 1.2×10^{-8} was calculated from an analysis of whole-genome sequencing of over 70 trios[19], which is equal to the rate obtained in our analysis, suggesting the accuracy of using autozygosity to estimate the mutation rate. Notably, the reported quantification of the correlation between the number of mutations and paternal age[19] suggests that the relatively young age of the fathers of the trios analyzed here (21–30 years at the time of childbirth) may provide an explanation for the lower mutation rate we observed in the most recent generation. In a second publication, an inferred sequence mutation rate of 1.82×10^{-8} was calculated by modeling population genetic parameters on the basis of the mutational properties of microsatellites[22]; the differences between this estimate and our estimates are likely due to differences in methodology.

We observed a non-uniform distribution of a small fraction of mutations within autozygous segments that seem to provide evidence of recent allelic gene conversion. We observed three clusters of variants that were unlikely to have been generated by crossover mechanisms and might represent potential allelic gene conversion events in the autozygous segments, although one cluster (11 kb) was larger than expected for typical gene conversion events[23]. Only one of the ten SNVs in these clusters was in a CpG dinucleotide, and the GC content (0.36–0.50) of these three regions was not consistent with a model of recurrent mutation due to CpG methylation and demethylation. The average distance between heterozygous SNPs in these clusters was 2,723 bp, ruling out compound mutation[24] as a likely mechanism. Notably, one of these clusters of heterozygous SNVs is comprised of two new SNVs (not in dbSNP) and could be further evidence of a non-uniform distribution of new mutations similar to what we observed for *de novo* mutations. In addition, we observed an excess of heterozygous bases at dbSNP positions in autozygous segments ($n = 22$), most of which were not clustered with other heterozygous variants (16 of 22) but may also be evidence of recent gene conversion events.

Notably, we observed a non-uniform distribution (three clusters with two *de novo* SNVs within 10 kb; range 7–3,921 bp) (empirical $P = 6 \times 10^{-6}$) among the validated *de novo* events. One of these clusters contained SNVs that were 7 bp apart, suggesting a compound mutational event[24]; on the basis of this event, 0.97% of *de novo* mutations were calculated to be part of multinucleotide mutations (95% CI (Wilson method) = 0.27–3.5%). Although this estimate is somewhat lower than the estimate of 2–3% of *de novo* mutations in compound mutational events, which was estimated on the basis of whole-genome sequencing data from two trios[24], the confidence intervals do overlap. The remaining two clusters contained SNVs greater than 2 kb apart, suggesting that, even at greater distances, mutational mechanisms do not produce uniform distributions of new mutations.

We have presented a new approach for estimating the sequence mutation rate in humans. We based our analysis on autozygosity and whole-genome sequencing data from whole blood–derived DNA to remove the effects of somatic mutations and cell line artifacts. In addition, we were able to detect other recent changes in the genome, including gene conversion events. Furthermore, we believe that the application of this approach to additional families has the potential to elucidate the dynamics of other forms of mutation, including CNVs and indels.

URLs. Picard, http://picard.sourceforge.net/; sequence data for additional European-American, http://aws.amazon.com/datasets/3357.

METHODS

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

Accession codes. Sequence data have been deposited at the database of Genotypes and Phenotypes (dbGAP) as a substudy of Genetic Studies in the Hutterites (accession phs000185.v1.p1).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to M. Przeworski for thoughtful comments on the manuscript. We thank C. Lee, B. Paeper, J. Smith and M. Rieder for assistance with sequence data generation and J. Huddleston for technical advice. We are grateful to T. Brown for assistance with manuscript preparation. C.D.C. was supported by a US National Institutes of Health (NIH) Ruth L. Kirschstein National Research Service Award (NRSA; F32HG006070). P.H.S. is supported by a Howard Hughes Medical Institute International Student Research Fellowship. This work was supported by an American Asthma Foundation Senior Investigator Award to

E.E.E., by US NIH grants R01 HD21244 and R01 HL085197 to C.O. and by US NIH grant R01 HG002899 to M.A. E.E.E. is an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

C.D.C., J.X.C., C.O. and E.E.E. designed the study. C.D.C. performed the genome sequencing analysis, molecular inversion probe (MIP)-targeted resequencing analysis and mutation rate calculations. J.X.C. performed analyses to determine the ancestry of the autozygous segments. M.M. performed and analyzed validation experiments, including Sanger sequencing, microarray hybridization and MIP capture. A.K. and P.H.S. performed read-depth copy-number analysis. B.L.D. identified and analyzed the clusters of heterozygous SNVs in the autozygous segments. L.H. and M.A. performed autozygosity analysis with SNP microarray data. L.V. and B.J.O. created the sequencing libraries. B.J.O. designed the MIP oligonucleotides. L.V., along with M.M., performed MIP capture. E.E.E., C.O., M.A. and J.S. supervised the project. C.D.C. and E.E.E. wrote the manuscript with input and approval from all coauthors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/ng.2418.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Haldane, J.B.S. The rate of spontaneous muation of a human gene. *J. Genet.* **31**, 317–326 (1935).
2. Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2003).
3. Drake, J.W., Charlesworth, B., Charlesworth, D. & Crow, J.F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
4. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
5. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).
6. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
7. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
8. Chong, J.X. *et al.* A common spinal muscular atrophy deletion mutation is present on a single founder haplotype in the US Hutterites. *Eur. J. Hum. Genet.* **19**, 1045–1051 (2011).
9. Cusanovich, D.A. *et al.* The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum. Mol. Genet.* **21**, 2111–2123 (2012).
10. Abney, M., Ober, C. & McPeek, M.S. Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am. J. Hum. Genet.* **70**, 920–934 (2002).
11. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
12. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
13. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
14. Han, L. & Abney, M. Identity by descent estimation with dense genome-wide genotype data. *Genet. Epidemiol.* **35**, 557–567 (2011).
15. Yang, Y. *et al.* Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.* **80**, 1037–1054 (2007).
16. Haldane, J.B. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann. Eugen.* **13**, 262–271 (1947).
17. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
18. Fledel-Alon, A. *et al.* Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet.* **5**, e1000658 (2009).
19. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
20. Khalak, H.G. *et al.* Autozygome maps dispensable DNA and reveals potential selective bias against nullizygosity. *Genet. Med.* **14**, 515–519 (2012).
21. Awadalla, P. *et al.* Direct measure of the *de novo* mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* **87**, 316–324 (2010).
22. Sun, J.X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161–1165 (2012).
23. Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C. & Patrinos, G.P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**, 762–775 (2007).
24. Schrider, D.R., Hourmozdi, J.N. & Hahn, M.W. Pervasive multinucleotide mutational events in eukaryotes. *Curr. Biol.* **21**, 1051–1054 (2011).

ONLINE METHODS

DNA samples and whole-genome sequencing. We selected 5 parent-offspring trios for whole-genome sequencing from a 13-generation pedigree of Hutterites from South Dakota (**Fig. 1**). Detailed phenotyping data are available for all individuals in this study design. None of the individuals studied have a Mendelian disorder. All individuals consented to participation in the study, and the project was approved by the institutional review boards (IRBs) at The University of Chicago and the University of Washington.

One library for each individual was constructed from DNA isolated from whole blood using Illumina-recommended protocols. Briefly, 1–3 μg of DNA was fragmented using sonication. The resulting fragments were end repaired, adenosine overhangs were added and adaptors were ligated. Size selection was performed by running the library on an agarose gel and excising a band of approximately 400 bp in size. Size-selected libraries were amplified using quantitative PCR. The resulting libraries were sequenced using an Illumina HiSeq 2000 to generate 51- and 101-bp paired-end reads. We generated 36–68 Gb of sequencing data for each of the 15 individuals (10–17× coverage of the human genome) (**Supplementary Table 1**).

Sequence data analysis and SNV identification. We aligned the paired-end reads to the NCBI Build 36 reference of the human genome using Burrows-Wheeler Aligner (BWA, version 0.5.9) with standard parameters[25]. The quality scores for the mapped reads were recalibrated, and PCR duplicates were removed using Picard (version 1.43). Then, we realigned the reads around potential insertions and/or deletions (indels) to reduce spurious SNV calls due to misalignment in these regions using Genome Analysis Toolkit (GATK) software (version 1.0.5777)[12]. We used both GATK and SAMtools (version 0.1.8)[11] to identify SNVs. After generating an initial list of SNVs, we used the following approaches to generate a high-confidence variant list. First, we applied variant recalibration to the variants identified with GATK and used a variant quality score recalibration (VQSR) threshold of 2.30 (99% of known high-quality SNPs identified) to generate a final list of GATK-called variants[26]. In addition, we applied the standard recommended GATK filters and filtered out SNVs located near indels. For the SAMtools call set, we filtered out calls with a quality score less than 10. We used the intersection of the GATK and SAMtools call sets for further analysis.

Concordance with SNP microarray genotypes and false negative rate estimation. To assess the quality of our SNV genotypes from whole-genome sequencing, we compared these data to genotype data generated for these same individuals on Affymetrix SNP microarrays (versions 500K and 6.0)[9]. We assessed the genotype concordance for the 211,438–468,681 SNPs that passed quality control metrics on the microarray for each individual (**Supplementary Table 1**).

We estimated the false negative rate for heterozygous variants because our calculations of mutation rate are based on heterozygous SNVs and these variants are more likely to be missed in moderate-coverage whole-genome sequencing data. For each individual, we determined the number of heterozygous SNPs genotyped by SNP microarray and then determined how many of these variants were missed by whole-genome sequencing (called as homozygous). We calculated false negative rate as the number of missed heterozygous SNPs divided by the total number of heterozygous SNPs (**Supplementary Table 1**).

Identification and definition of autozygous segments. We identified long stretches of homozygosity in the genomes of the five children of the Hutterite trios, two European-Americans from the Centre d'Etude du Polymorphisme Humain collection (CEU)[13], two Yoruba from Ibadan, Nigeria (YRI)[13] and one additional European-American male. We ran PLINK[27] on all filtered SNV genotypes, with variants in segmental duplications removed to avoid artifacts due to paralogous sequence variation. We required a minimum homozygosity length of 600 kb, a maximum gap in homozygosity of 100 kb and a maximum of three heterozygous SNPs per 5 Mb. We merged all segments within 50 kb, as it seemed that some regions, especially in the Hutterites, were erroneously split. We did not consider regions with >20% gaps or segmental duplications in the reference assembly.

We compared the resulting list of large autozygous segments (>5 Mb) to regions of likely autozygosity determined with SNP microarray genotypes and the 3,555-member pedigree using the program IBDLD[14]. The intersection of regions determined by genome sequence data and those determined by SNP microarray data gave us a final list of 44 autozygous regions of greater than 5 Mb in length. We trimmed all 44 segments by 100 kb from each edge because we observed an excess of heterozygous variants at the edges (**Supplementary Fig. 5**).

Determination of common ancestors for autozygous segments. The 15 sequenced individuals are part of a larger 13-generation pedigree of Hutterites, and we made use of this pedigree information to trace each autozygous segment to the MRCA of all carriers, as previously described[8]. We used genome-wide SNP genotypes of 1,415 individuals from this large pedigree[8,9] to identify all individuals who carried at least 1 allele that was identical by state (IBS) to sequenced individuals with the autozygous segment across the majority of the SNPs in the autozygous region. We considered individuals to be a carrier of the autozygous haplotype if they had IBS of ≥1 with the sequenced individual at >99% of the genotyped SNPs in the autozygous segment; the median number of SNPs analyzed was 1,204 (range of 84–5,493). After we identified all haplotype carriers for an autozygous segment, we examined the pedigree to identify all common ancestors of these individuals. We defined the MRCA as the individual with the smallest mean number of meioses to all haplotype carriers. To estimate mutation rate, we used the mean number of meioses in all paths from the MRCA(s) to the autozygous individual (**Supplementary Note**).

Identification of heterozygous SNVs in regions of autozygosity. We intersected our list of autozygous segments in each sample with the list of quality-filtered SNVs identified in that sample. Then, for each heterozygous SNV identified in the autozygous sample, we applied an additional read-depth filter, requiring a minimum of six sequencing reads at that location in the heterozygous individual. We did not consider SNVs in simple repeats (based on the Tandem Repeats Finder track for hg18 in the UCSC Genome Browser) or segmental duplications or those that were found in dbSNP132. We identified a total of 85 unreported heterozygous SNVs in the refined autozygous regions (**Supplementary Table 5**) and attempted to validate these variants with Sanger sequencing (**Supplementary Note**).

Determination of SNV mutation rate using autozygosity. For mutation rate calculations, we determined the number of base pairs in each autozygous segment that we were able to test for heterozygous variants. We counted the number of unique (not in segmental duplications), non-simple repeat, non-dbSNP132 bases with a read depth of at least six that were callable by GATK; these callable bases served as the denominator in our mutation rate calculation. To calculate mutation rate for each individual and across all five individuals, we applied the formula

$$\mu = (N \times (1 + \mathrm{FNR}))/(G \times L)$$

where N was the number of validated heterozygous SNVs in the autozygous regions for that individual, FNR was the false negative rate for heterozygous SNVs, G was the weighted average of the number of meioses separating the alleles of autozygous segments (weighted by length of segment) for that individual and L was the sum of callable base pairs in autozygous segments in that individual. The same formula was applied to calculate the mutation rate across all 5 individuals, where N was the 72 validated heterozygous SNVs, G was 11.9—the weighted (by segment length) mean number of meioses separating the two alleles of all autozygous segments—and L was the 512.4 Mb of total callable base pairs in autozygous segments in the 5 individuals. To calculate the mutation rate at CpG base pairs, we considered only the estimated number of true heterozygous SNVs at CpG bases divided by callable CpG base pairs and the number of generations to the MRCA.

Determination of mutation rate using *de novo* mutations in the most recent generation. We identified putative *de novo* SNVs across the whole genome that were observed only in a single individual (and were not observed in the parents of that individual or in any of the other Hutterite genomes). We did not consider variants in segmental duplications, simple repeats or known

doi:10.1038/ng.2418

SNPs (present in dbSNP132), and we required a read depth of at least six for all individuals in the trio. After applying these filters, we obtained a list of 632 putative *de novo* variants (**Supplementary Table 7**). Using molecular inversion probes (MIPs)[28,29], we attempted to capture and resequence these putative *de novo* variants (**Supplementary Note**). We calculated the mutation rate using the equation

$$\mu = (N \times (1 - \mathrm{FDR}) \times (1 + \mathrm{FNR}))/(2 \times L)$$

where N was the total number of putative *de novo* mutations, FDR was the false discovery rate for putative *de novo* SNVs (**Supplementary Table 6** and **Supplementary Note**), FNR was the false negative rate and L was the total callable base pairs with read depth of at least six in all members of the trio in the genome.

Simulations of mutation clusters. To determine whether there was a non-random distribution of heterozygous SNVs in autozygous segments, we performed the following simulations. We randomly permuted the positions of heterozygous SNVs in autozygous blocks that contained more than one heterozygous SNV and determined the number of clusters of SNVs with less than 10 kb of sequence between them. We repeated this process 10,000 times and determined an empirical P value on the basis of the number of simulations that gave at least 3 heterozygous SNV clusters (heterozygous SNVs within 10 kb or each other). We performed a similar analysis for the *de novo* SNVs by permuting 1,000,000 times the locations of the estimated number of true *de novo* SNVs in each sample within the callable regions of the genome.

25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
26. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
27. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
28. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
29. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6**, 315–316 (2009).