Supplementary Information

Estimating 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†]

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195
[2]Department of Human Genetics, The University of Chicago, Chicago, IL 60637
[3]Department of Obstetrics and Gynecology, The University of Chicago, Chicago, IL 60637
[4]Howard Hughes Medical Institute, Seattle, WA 98195

Contents:

Nature Genetics: doi:10.1038/ng.2418

Supplementary Tables

Supplementary Table 1: Genome sequence generated

Sample	Total sequence*	Mean coverage†	Concordance to SNP array‡	Het. FNR§	paired end distribution$	
					median	s.d.
1	53 Gb	13.02	99.68%	1.95%	178	43
father1	43 Gb	11.84	99.75%	2.08%	233	64
mother1	47 Gb	12.47	99.57%	2.38%	228	67
2	69 Gb	17.07	99.90%	1.68%	219	46
father2	46 Gb	12.06	99.70%	2.20%	221	65
mother2	36 Gb	9.72	99.44%	2.90%	223	61
3	67 Gb	16.24	99.89%	1.53%	217	46
father3	39 Gb	10.51	99.55%	2.46%	209	59
mother3	46 Gb	12.21	99.67%	2.22%	235	64
4	59 Gb	15.21	99.78%	1.74%	209	46
father4	57 Gb	11.90	99.74%	2.10%	222	63
mother4	45 Gb	10.72	99.63%	2.22%	218	63
5	63 Gb	13.06	99.77%	1.78%	208	42
father5	51 Gb	13.28	99.74%	2.03%	208	57
mother5	54 Gb	14.31	99.65%	2.08%	218	60

*Based on the total number of sequencing reads generated. †Effective coverage of the genome after mapping with BWA. ‡Genotype concordance to SNPs genotyped on Affymetrix 500K and 6.0 microarrays. §False negative rate for heterozygous variants based on comparisons to SNP microarray data. $Median insert size and standard deviation (s.d.) of insert sizes for the paired end libraries sequenced.

Nature Genetics: doi:10.1038/ng.2418

Supplementary Table 2: Statistics on identified SNVs

Sample	SNVs	Novel SNVs*	Ti/Tv†	heterozygosity‡
1	2,733,690	55,064	2.14	5.63×10^{-4}
father1	2,742,745	53,787	2.15	5.74×10^{-4}
mother1	2,776,905	55,468	2.15	5.95×10^{-4}
2	2,767,370	58,264	2.15	5.76×10^{-4}
father2	2,697,671	52,752	2.15	5.37×10^{-4}
mother2	2,698,797	53,221	2.15	5.36×10^{-4}
3	2,751,698	58,344	2.15	5.61×10^{-4}
father3	2,762,920	55,502	2.15	5.78×10^{-4}
mother3	2,761,303	56,159	2.15	5.75×10^{-4}
4	2,748,639	57,525	2.15	5.72×10^{-4}
father4	2,728,510	52,606	2.15	5.59×10^{-4}
mother4	2,764,670	56,469	2.15	5.87×10^{-4}
5	2,761,982	57,495	2.15	5.74×10^{-4}
father5	2,740,348	53,760	2.15	5.66×10^{-4}
mother5	2,781,887	57,301	2.15	5.95×10^{-4}
Mean	2,747,942	55,581	2.15	5.70×10^{-4}
Standard deviation	25,172	1,993	0.0026	0.17×10^{-4}
All	5,427,490	341,415	2.17	91.4×10^{-4}

*Not reported in dbSNP132. †Transition to transversion ratio for all SNVs. ‡Heterozygous SNVs per basepair.

Nature Genetics: doi:10.1038/ng.2418

Supplementary Table 3: Properties of autozygosity

Individual	Population	inbreeding coef.*	emp. inbreeding coef.†	Total Mbp‡	Percent of genome	N§	average length (Mbp)	Max (Mbp)	N > 5 Mbp$
1	Hutterite	0.0308	0.0408	175.78	0.06	106	1.66	17.78	7
2	Hutterite	0.0279	0.0375	160.49	0.05	100	1.60	15.59	6
3	Hutterite	0.0567	0.0666	234.01	0.08	104	2.25	54.40	9
4	Hutterite	0.046	0.0741	280.20	0.09	114	2.46	28.65	10
5	Hutterite	0.0479	0.0662	263.70	0.09	113	2.33	27.95	12
NA12891	CEU	NA	NA	96.33	0.03	105	0.92	2.07	0
NA12892	CEU	NA	NA	74.55	0.02	84	0.89	2.05	0
JF	Eu. Am.	NA	NA	114.69	0.04	125	0.92	2.58	0
NA19238	YRI	NA	NA	2.05	0.00	3	0.68	0.70	0
NA19239	YRI	NA	NA	6.61	0.00	9	0.73	1.29	0

*Determined based on pedigree structure. †Estimated from SNP microarray data. ‡In autozygous segments. §Number of autozygous segments greater than 600 kbp. $Number of autozygous segments greater than 5 Mbp.

Nature Genetics: doi:10.1038/ng.2418

Supplementary Table 4: Large (>5 Mbp) autozygous segments identified

Individual	chr	start	end	length*	callable bp†	meioses (MRCA)‡	meioses (CAs)§	min meioses$	max meioses$	SNVs#	m**
1	1	4380204	12659029	8278825	7266637	15.50	16.77	15.00	19.00	0	0.00
1	1	74496036	91845959	17349923	15506278	14.00	15.95	14.00	18.00	3	1.39
1	1	206788618	212685969	5897351	5461825	15.67	16.45	15.00	18.00	1	1.23
1	6	139817172	155954044	16136872	14782891	13.75	15.97	13.00	18.00	0	0.00
1	12	12727386	20305297	7577911	7022499	14.00	16.30	14.00	19.00	0	0.00
1	17	3889240	13138519	9249279	8287289	12.00	15.52	12.00	18.00	4	4.05
1	17	69596127	75783594	6187467	5108872	12.00	15.58	12.00	18.00	0	0.00
2	1	46683429	61942991	15259562	14389324	12.50	16.53	12.00	21.00	2	1.16
2	2	13099145	24889124	11789979	11120912	14.50	16.96	14.00	21.00	0	0.00
2	4	143724532	150737339	7012807	6382121	13.75	16.60	13.00	21.00	0	0.00
2	6	125369705	131673895	6304190	6074399	15.00	17.30	14.00	21.00	0	0.00
2	7	22773232	34951868	12178636	11108343	12.50	16.57	12.00	21.00	1	0.75
2	9	83512725	90998603	7485878	6787189	16.33	17.65	16.00	20.00	4	3.69
3	2	181128523	235344001	54215478	51515397	8.00	13.95	8.00	20.00	0	0.00
3	5	89100412	108539956	19439544	18245947	8.00	14.24	8.00	20.00	0	0.00
3	10	10104714	21180824	11076110	9891093	12.00	15.86	12.00	20.00	1	0.84
3	10	126392926	133403037	7010111	6508885	12.00	15.18	12.00	20.00	0	0.00
3	12	19932243	24984425	5052182	6959256	12.00	15.42	12.00	20.00	0	0.00
3	12	120076450	127703723	7627273	4822316	12.00	15.84	12.00	20.00	1	1.73
3	13	27284672	36897091	9612419	9060958	12.00	16.05	12.00	20.00	0	0.00
3	15	53023254	65145693	12122439	11547507	14.25	16.46	14.00	20.00	1	0.62
3	19	55832228	63685051	7852823	6215216	10.00	15.18	10.00	20.00	0	0.00
4	2	153274937	168517066	15242129	14312509	15.83	18.09	15.00	22.00	2	0.94
4	2	190744833	214609471	23864638	22483717	9.00	16.43	9.00	21.00	1	0.50
4	3	134042594	139153842	5111248	4779996	16.08	18.65	15.00	22.00	0	0.00
4	4	142301120	154091355	11790235	10798243	18.06	18.98	17.00	22.00	4	2.19
4	5	111959448	140150943	28191495	26466806	14.07	17.56	13.00	22.00	1	0.29
4	5	149363205	164853808	15490603	14643582	11.50	17.23	11.00	22.00	0	0.00

Nature Genetics: doi:10.1038/ng.2418

4	6	16268962	42045145	25776183	23040104	8.00	16.55	8.00	22.00	2	1.09
4	7	97505884	107290667	9784783	8267674	15.83	18.52	15.00	22.00	0	0.00
4	8	122314901	129257931	6943030	6582873	10.00	16.68	10.00	22.00	1	1.52
4	15	72973557	92472189	19498632	16175531	9.00	16.42	9.00	21.00	1	0.69
5	1	9511993	22505336	12993343	10473394	9.00	14.76	9.00	21.00	0	0.00
5	2	117639769	131392339	13752570	12085789	14.00	16.80	14.00	21.00	1	0.59
5	2	207136840	222465357	15328517	13950466	9.00	14.76	9.00	21.00	1	0.80
5	5	54388582	74104801	19716219	16374856	9.00	15.49	9.00	21.00	0	0.00
5	6	13456825	19824836	6368011	5932392	14.80	17.07	14.00	21.00	1	1.21
5	7	70484909	77330588	6845679	3562381	16.35	17.38	15.00	19.00	0	0.00
5	8	101861306	125652223	23790917	22187376	14.75	16.93	14.00	21.00	1	0.32
5	10	14102886	25516498	11413612	9962741	10.67	16.28	10.00	21.00	2	2.02
5	16	53425894	60525840	7099946	6420164	10.67	16.08	10.00	21.00	1	1.56
5	X	40468899	48875515	8406616	7394528	12.75	16.34	12.00	21.00	0	0.00
5	X	61917570	67970017	6052447	5271973	16.35	17.63	15.00	19.00	1	1.27
5	X	98092685	107017870	8925185	7194854	12.75	16.75	12.00	21.00	1	1.16

*Based on genomic positions (hg18). †Bases with at least six mapped reads and not in segmental duplications, simple repeats, dbSNP132. ‡Mean across all potential paths from MRCA to proband. §Mean meioses based on all paths from all common ancestors to proband. $Based on all common ancestors. #Novel and validated. **x10^{-8}

Supplementary Table 5: Summary of Sanger sequencing validations performed

See Excel file.

Nature Genetics: doi:10.1038/ng.2418

Supplementary Table 6: Summary of *de novo* SNV mutation rate

Trio	de novo calls*	MIPs†	FP proband‡	Inherited§	de novo$	ND#	Callable (Gbp)**	FDR††	μ ($x10^{-8}$)‡‡	95% CI ($x10^{-8}$)§§
1	117	114	9	60	29	19	2.169	0.69	0.84	0.56-1.12
2	147	139	2	86	35	24	2.102	0.72	1.01	0.69-1.32
3	167	158	3	103	39	22	2.158	0.73	1.06	0.73-1.36
4	122	115	2	70	36	14	2.182	0.67	0.95	0.64-1.24
5	79	75	2	28	37	12	2.243	0.45	0.99	0.68-1.28
All	632	601	18	347	176	91	10.855	0.67	0.96	0.82-1.09

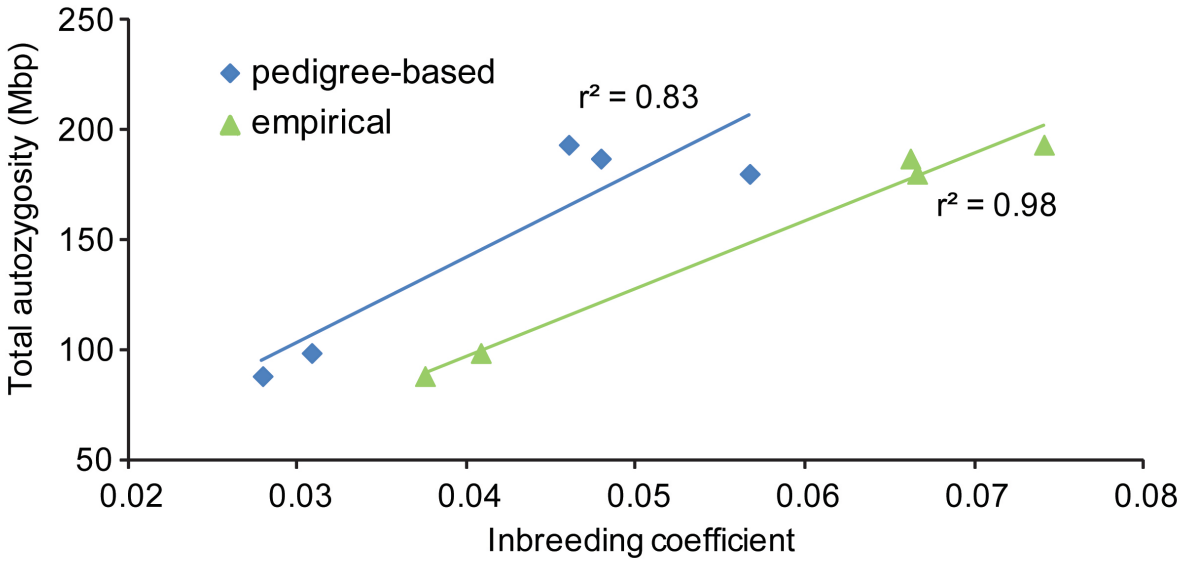
*Putative *de novo* SNVs identified in WGS. †Molecular inversion probes (MIPs) successfully designed. ‡False positive (i.e. homozygous reference genotype) in the proband. §SNV is inherited from one of the parents. $Validated in proband and not observed in the parents. #Could not be determined from validation experiments. **Total bases were "callable" by GATK in all three members of the trio with a read depth of at least six but did not overlap dbSNP132, tandem repeats, or segmental duplications. ††False discovery rate based on the number of true positive *de novo* SNVs as a fraction of the total putative *de novo* SNVs with validation data. ‡‡Mutation rate calculated as described in the methods. §§95% confidence intervals around mutation rate estimate based assuming a Poisson process.

Supplementary Table 7: Summary of putative *de novo* mutations and validation

See Excel file.

Nature Genetics: doi:10.1038/ng.2418

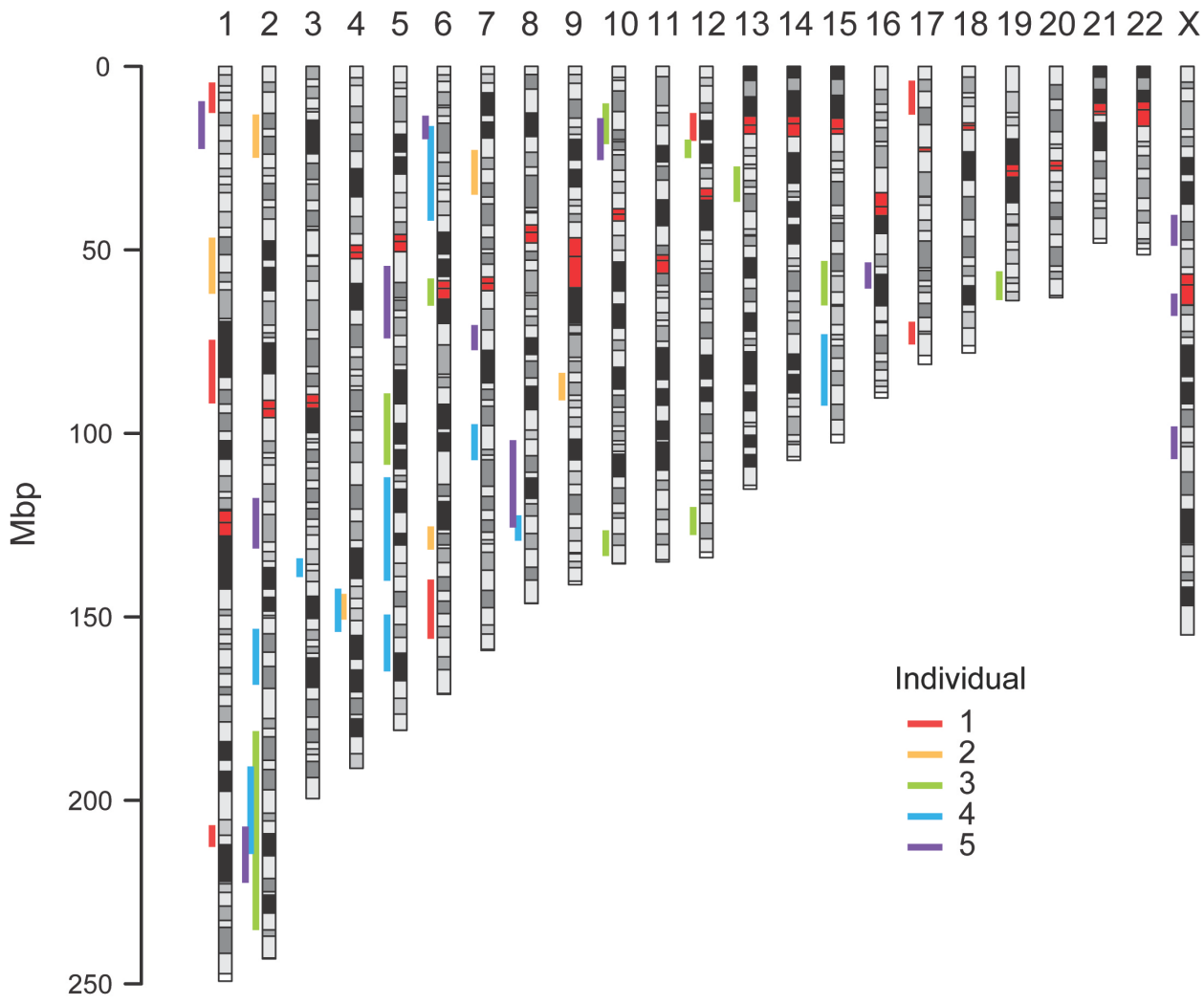
Supplementary Figures



Supplementary Figure 1 – Autozygosity is correlated with inbreeding coefficient.

The total amount of autozygosity in the five Hutterite individuals is correlated to the inbreeding coefficient determined from the pedigree (blue; $r^2 = 0.83$) and determined empirically using SNP microarray data (green; $r^2 = 0.98$).

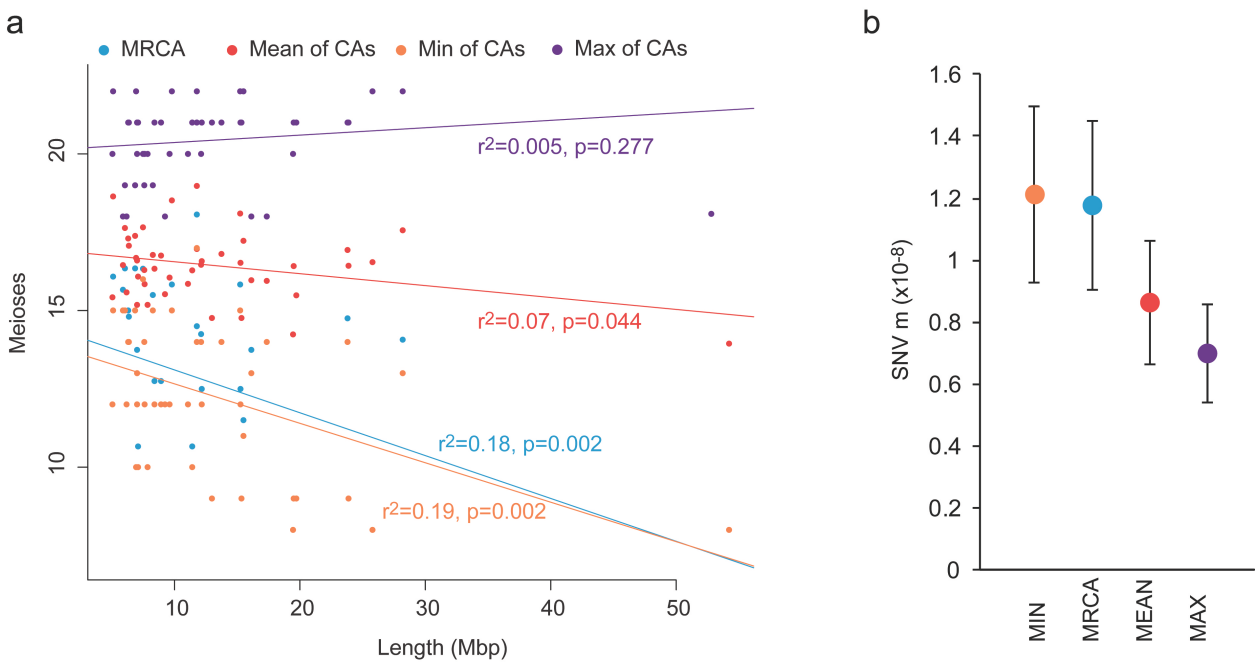
Nature Genetics: doi:10.1038/ng.2418



Supplementary Figure 2 – Distribution of autozygous segments in the genome.

Chromosome ideograms for autosomes and chromosome X are shown. The 44 autozygous segments greater than 5 Mbp are shown for all five Hutterites, each in a different color. There are autozygous segments on most chromosomes and these segments do not appear to be clustered in any way.

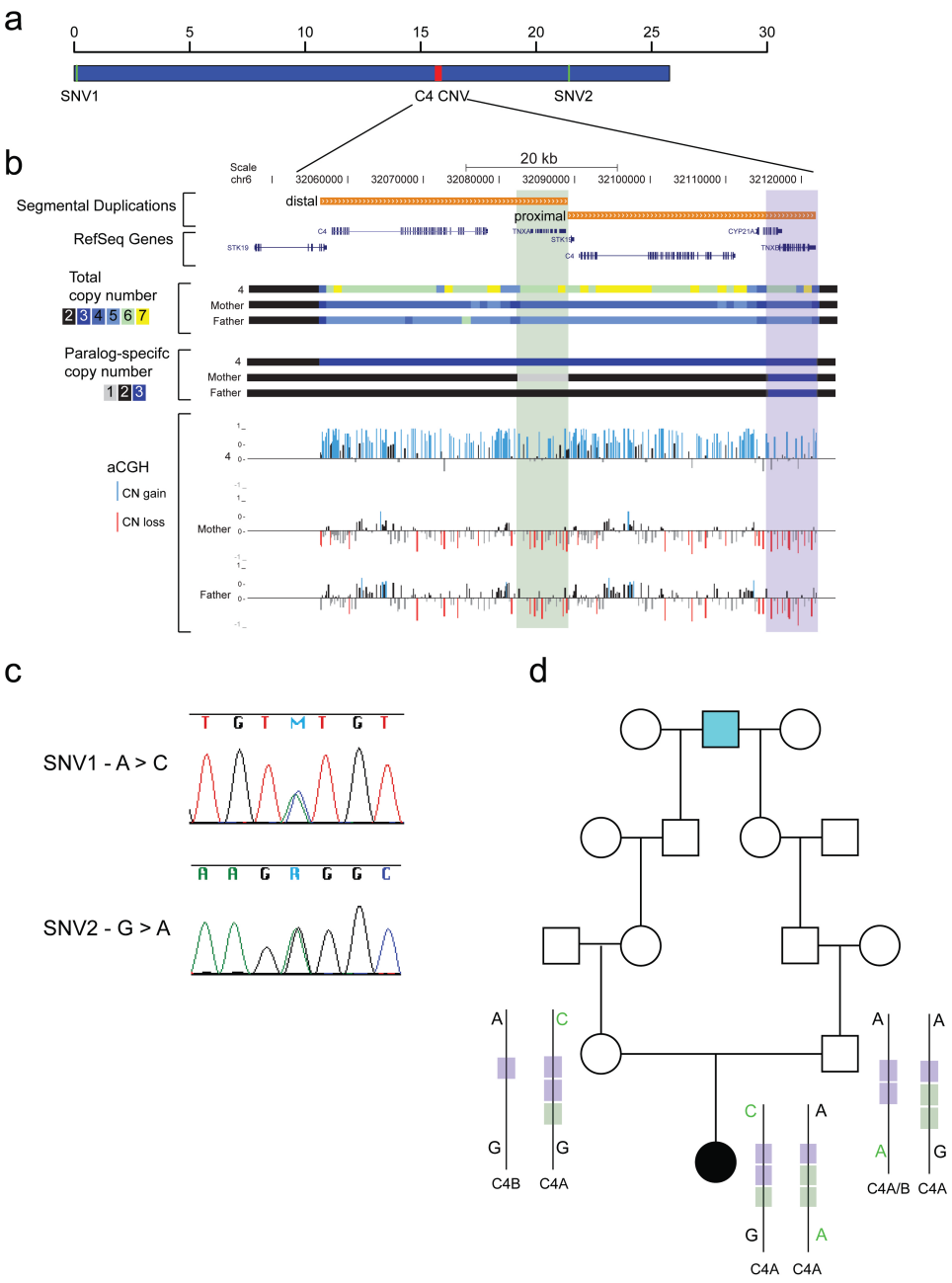
Nature Genetics: doi:10.1038/ng.2418



Supplementary Figure 3 – Comparison of multiple methods for estimating the number of meioses separating the alleles of autozygous segments.

a) Longer autozygous segments should be, on average, younger. The length of each of the 44 autozygous segments is plotted versus the number of meioses separating the alleles assuming the MRCA(s) as the origin(s) (blue), the mean meioses separating the alleles based on all CA of the haplotype carriers in the pedigree (red), the minimum meioses from all CAs (orange), and the maximum meioses from CAs (purple). b) Mutation rate point estimates and 95% confidence intervals based on each estimate of meioses.

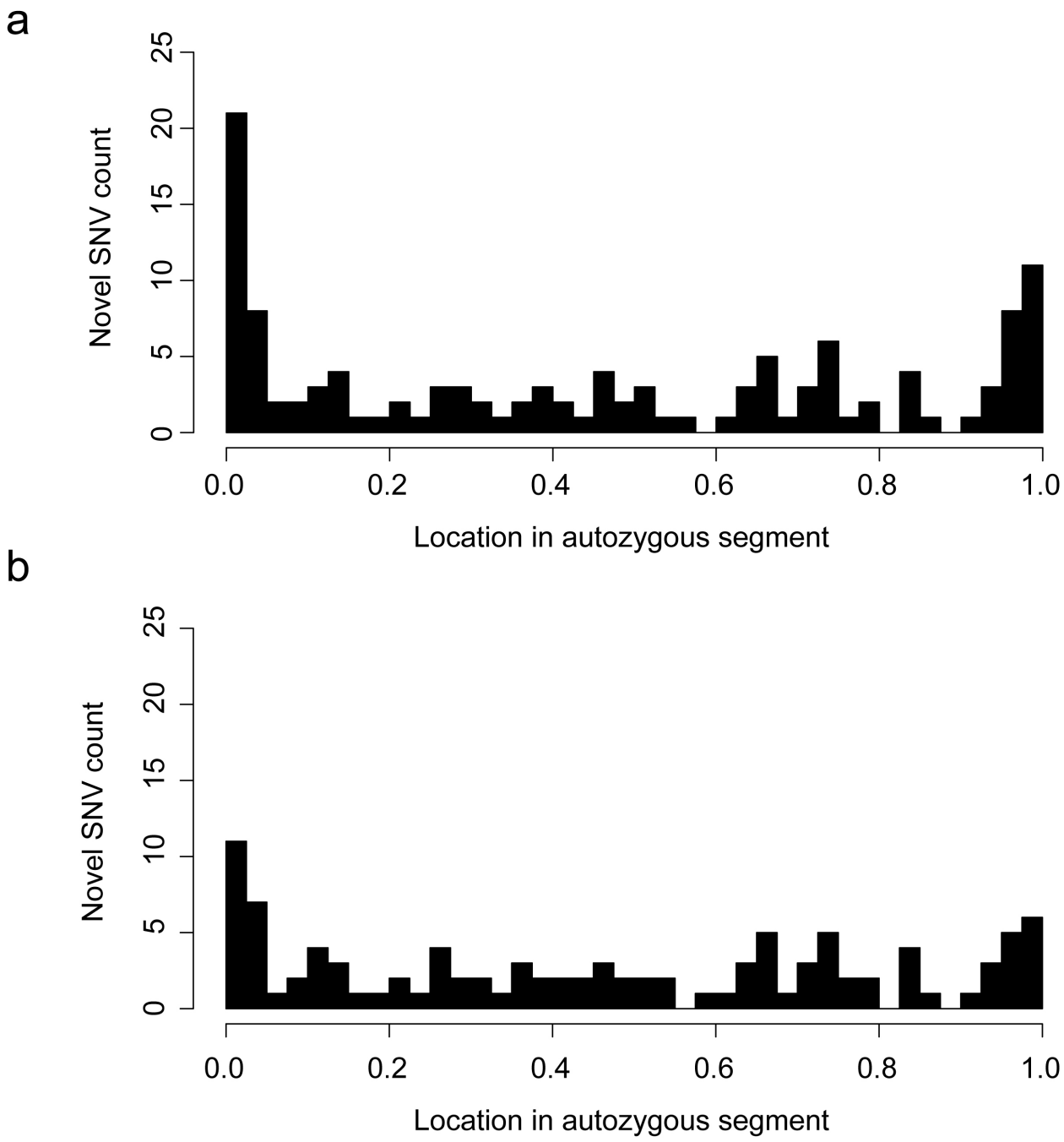
Nature Genetics: doi:10.1038/ng.2418



Supplementary Figure 4 – Putative recent copy number change in the *C4* locus.

a) A 26 Mbp segment of homozygosity on chromosome 6 represented in blue with heterozygous SNVs in green and a heterozygous structural variant in red. b) A pair of tandem segmental duplications shown in orange is located in a 26 Mbp autozygous segment on chromosome 6 in Individual 4. The paralog-specific copy numbers are shown for the individual and both parents in 1 kbp windows of non-repeat masked sequence. The two highly identical paralogs can be distinguished by paralogous sequence variants between *TNXA* and *TNXB* as highlighted in green and purple. Microarray CGH data supports these results and is shown below where each probe is represented by a vertical bar colored blue if the log ratio is greater than 0.5 and red if the log ratio is less than -0.5. c) Confirmation of both heterozygous SNVs by Sanger sequencing. d) The pedigree relating the sequenced individuals to the MRCA of this segment; the parents of the autozygous individual are half-third cousins. The C4 genotypes were determined previously using serological typing [1].

Nature Genetics: doi:10.1038/ng.2418



Supplementary Figure 5 – Distribution of heterozygous SNVs in autozygous segments.

Histograms of the number of SNVs in 40 bins across all autozygous segments are shown for a) original autozygous segments and b) autozygous segments with 100 kbp trimmed from each end.

Nature Genetics: doi:10.1038/ng.2418

Supplementary Note

Comparison of the extent of autozygosity

As expected, we observed an increase in autozygosity among the Hutterites when compared to Yoruba and other European-American individuals. The total amount of autozygous basepairs was more correlated with the inbreeding coefficient estimated empirically than that estimated genealogically. The difference in these correlations is likely due to random deviations in segment length and/or cryptic relatedness between individuals that is not represented in the pedigree. Some of the founders of the Hutterite populations were likely related to one another prior to emigration to the United States [2]. If we focus on the background level of autozygosity, we observe shorter segments than has been estimated previously from "outbred" European populations [3]. This may be a result of the increased resolution provided by WGS compared to SNP-microarray platforms.

Determination of the MRCA for autozygous segments

We leveraged SNP microarray data from 1415 Hutterites and identified all individuals sharing at least one allele identical by state (IBS) with the sequenced individual at ≥99% of the SNPs in a particular autozygous segment. We inferred that IBS ≥1 at many consecutive SNPs constitutes a region that is identical by descent (IBD) [4-5]. We identified all CAs of these haplotype carriers and iterated over all possible paths between each of the haplotype carriers and the CAs to calculate the mean number of meiosis separating those individuals. Due to the large family sizes in the Hutterite pedigree, we could use a conservative threshold for identifying haplotype carriers. For example, if we failed to identify a carrier in a family, then we were likely to identify that individual's siblings as carriers and these individuals would have the exact same CAs. Therefore, we required haplotype carriers to have at least one allele IBS to the autozygous haplotype for >99% of the SNPs in the region. Due to the complexity of the Hutterite pedigree, there were multiple "paths" connecting a CA to a haplotype carrier. We calculated the mean number of meioses from each CA to each of the haplotype carriers. We considered the CA(s) with the lowest mean meioses to all haplotype carriers to be the MRCA(s).

Based on the expectation that longer segments will have a shorter divergence time, we compared different estimates of divergence time in meioses and segment length. The mean number of meioses between the original individual with the autozygous segment and the MRCA(s) was significantly correlated with segment length in the expected direction and we used this value in our calculations of mutation rate (Figure 3; Supplementary Fig. 3). To assess the upper bound on the mutation rate, we determined the minimum meioses separating the alleles of each segment based on all CAs, which is by definition the shortest path from the MRCAs. To assess the lower bound on the mutation rate estimate, we determined the maximum divergence path based on all CAs. In addition, we also calculated the mean of the divergence for all CAs for

Nature Genetics: doi:10.1038/ng.2418

each segment. We then recalculated mutation rate based on these values. As expected, our original estimate based on the MRCA ($1.20x10^{-8}$) is very similar to that based on the shortest path ($1.22x10^{-8}$). The mutation rate based on the longest path is lower ($0.69x10^{-8}$) and the rate based on the mean of the paths to all CAs is intermediate ($0.87x10^{-8}$). Therefore, we estimated that the uncertainty in estimating the age of the segments yields a range of mutation rates from $0.69x10^{-8}$ to $1.22x10^{-8}$ (Supplementary Fig. 3).

SNV validation with Sanger sequencing

We designed PCR primers flanking all 85 novel, heterozygous SNVs using Primer3 [Ref. 6] (Supplementary Table S5). The resulting amplicons were approximately 400 bp in size, and the primers were tagged with M13 primers for sequencing. We performed PCR using 5 pmol each of forward and reverse primers, 25 ng of genomic DNA, and 1X PCR master mix (Roche). Standard cycling was performed: 95° for 5 minutes; then 35 cycles of 95° for 30 seconds, 58° for 30 seconds, and 72° for 45 seconds; followed be a final extension at 72° for 7 minutes. We checked the resulting PCR products on an agarose gel for the presence of a single band at the expected size. PCR products were then cleaned and sequenced using both forward and reverse M13 primers using Sanger sequencing. We used Sequencer (http://genecodes.com/) to align the resulting sequences to the targeted loci in the human genome reference assembly and looked for the presence of the expected heterozygous SNV (i.e. the presence of two peaks of the expected nucleotides) using both the forward and reverse primers. Of the 85 tested SNVs, we obtained interpretable sequencing for 78, and of these, 72 were validated as heterozygous. We considered these 72 as the recent mutations in autozygous segments.

Analysis of reported SNPs in regions of autozygosity

In addition to the analysis of novel SNVs, which we used to estimate mutation rate, we also examined reported SNPs (dbSNP132) with heterozygous genotypes. We also sought to estimate the recurrent mutation rate by analyzing heterozygous SNVs that have been reported in dbSNP. We validated 22 heterozygous SNPs by Sanger Sequencing (Supplementary Table 6), and calculated that we had obtained 5.62 Mbp of callable dbSNP basepairs in autozygous segments. Using the equation reported in the methods section with the 11.9 as the number of meioses, we obtained a rate of $3.27x10^{-7}$. This is over ten times higher than our estimate for SNV mutation rate suggestion that most of these variants are not recurrent mutations.

Validation of putative *de novo* SNVs

In order to quickly validate as many of the 632 putative *de novo* SNVs as possible, we designed 70 bp molecular inversion probes (MIPs) [7-8]. We selected a single MIP per variant and excluded MIPs where both arms aligned to more than 10 locations in the reference genome. Based on these criteria, we successfully targeted to 601 of the 632 putative *de novo* variants. We created an equimolar pool of MIPs for each trio except we added 50-fold excess of MIPs with low scores, high GC content (>60% of gap filled bases), or low GC content (<25%). We

Nature Genetics: doi:10.1038/ng.2418

performed the capture on all 15 individuals as previously described [9]. We phosphorylated the MIPs, performed the gap-fill and ligation reactions using 50 ng of genomic DNA for each individual, removed unreacted probes with exonuclease, and amplified and barcoded the libraries. The resulting libraries were cleaned with magnetic beads (Agencourt) to remove unreacted reagents. The libraries for all individuals were pooled and sequenced using Illumina MiSeq 150bp paired-end reads to a median coverage of 2252 of the targeted regions.

Of the 601 variants for which we attempted capture, we successfully captured (median coverage of target >100), sequenced and genotyped 530 (88%) (Supplementary Table 7). To determine the accuracy of this approach, we also performed Sanger sequencing on 61 putative *de novo* variants including 11 variants that failed MIP design or capture (Supplementary Table 6). Of the 50 variants for which we obtained Sanger and Illumina sequencing, 44 were concordant between the two methods (88%). Of the six discordant variants, we determined that two were missed by Sanger due to SNPs in the primer binding sites, one variant had a homozygous alternate allele call for the proband with Illumina but a heterozygous call with Sanger due to a potential capture bias, one variant was called heterozygous in the Illumina data with very low allele balance in the father (only 8.6% of reads contained the alternate allele) suggesting a potential mosaic event, and the remaining two discordancies remain unexplained.

Based on a combination of MIP and Sanger sequencing, we analyzed data for 541 putative *de novo* variants. Of these, 176 were validated as *de novo*, 347 were inherited from one of the parents, and 18 were false positive variant calls in the proband (Supplementary Tables 6 and 7). The high rate of undercalls in the WGS data from the parents is not surprising given that the average false negative rate for the parents was about double that in the probands (1.73% vs. 2.31%) (Supplementary Table 1). We calculated a false discovery rate for each trio and for all samples combined based on the number of invalidated *de novo* SNVs as a proportion of the total number of SNVs analyzed (Supplementary Table 6).

Assignment of *de novo* mutations to a parental haplotype

For each of the 632 putative *de novo* SNVs, we identified all SNPs within 500 bp. For each of these nearby SNPs, we examined whether the variant was heterozygous in the individual and whether the parents had different genotypes. These variants were considered as informative. For each of the 202 putative *de novo* SNVs with an informative SNP within 500 bp, we performed molecular phasing using GATK [10]. We were able to assign parental haplotype origin to 26 validated *de novo* variants (Supplementary Table 7).

Nature Genetics: doi:10.1038/ng.2418

References

1. Weitkamp, L.R. & Ober, C. Ancestral and recombinant 16-locus HLA haplotypes in the Hutterites. *Immunogenetics* **49**, 491-7 (1999).
2. Pichler, I. *et al.* Drawing the history of the Hutterite population on a genetic landscape: inference from Y-chromosome and mtDNA genotypes. *Eur J Hum Genet* **18**, 463-70 (2010).
3. McQuillan, R. *et al.* Runs of homozygosity in European populations. *Am J Hum Genet* **83**, 359-72 (2008).
4. 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**, 1045-51 (2011).
5. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068-75 (2008).
6. You, F.M. *et al.* BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**, 253 (2008).
7. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat Methods* **4**, 931-6 (2007).
8. 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**, 315-6 (2009).
9. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
10. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).

Nature Genetics: doi:10.1038/ng.2418