ARTICLE

doi:10.1038/nature11396

Rate of *de novo* mutations and the importance of father's age to disease risk

Augustine Kong[1], Michael L. Frigge[1], Gisli Masson[1], Soren Besenbacher[1,2], Patrick Sulem[1], Gisli Magnusson[1], Sigurjon A. Gudjonsson[1], Asgeir Sigurdsson[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Wendy S. W. Wong[3], Gunnar Sigurdsson[1], G. Bragi Walters[1], Stacy Steinberg[1], Hannes Helgason[1], Gudmar Thorleifsson[1], Daniel F. Gudbjartsson[1], Agnar Helgason[1,4], Olafur Th. Magnusson[1], Unnur Thorsteinsdottir[1,5] & Kari Stefansson[1,5]

Mutations generate sequence diversity and provide a substrate for selection. The rate of *de novo* mutations is therefore of major importance to evolution. Here we conduct a study of genome-wide mutation rates by sequencing the entire genomes of 78 Icelandic parent–offspring trios at high coverage. We show that in our samples, with an average father's age of 29.7, the average *de novo* mutation rate is 1.20×10^{-8} per nucleotide per generation. Most notably, the diversity in mutation rate of single nucleotide polymorphisms is dominated by the age of the father at conception of the child. The effect is an increase of about two mutations per year. An exponential model estimates paternal mutations doubling every 16.5 years. After accounting for random Poisson variation, father's age is estimated to explain nearly all of the remaining variation in the *de novo* mutation counts. These observations shed light on the importance of the father's age on the risk of diseases such as schizophrenia and autism.

The rate of *de novo* mutations and factors that influence it have always been a focus of genetics research[1]. However, investigations of *de novo* mutations through direct examinations of parent–offspring transmissions were previously mostly limited to studying specific genes[2,3] or regions[4–7]. Recent studies that used whole-genome sequencing[8,9] are important but too small to address the question of diversity in mutation rate adequately. To understand the nature of *de novo* mutations better we designed and conducted a study as follows.

Samples and mutation calls

As part of a large sequencing project in Iceland[10–12] (Methods), we sequenced 78 trios, a total of 219 distinct individuals, to more than 30× average coverage (Fig. 1). Forty-four of the probands (offspring) have autism spectrum disorder (ASD), and 21 are schizophrenic. The other 13 probands were included for various reasons, including the construction of multigeneration families. The probands include five cases in which at least one grandchild was also sequenced. In addition, 1,859 other Icelanders, treated as population samples, were also whole-genome sequenced (all at least 10×, 469 more than 30×). These were used as population samples to help to filter out artefacts. Sequence calling was performed for each individual using the Genome Analysis Toolkit (GATK) (Methods). The focus here is on single nucleotide polymorphism (SNP) mutations. The investigation was restricted to autosomal chromosomes.

Criteria for calling a *de novo* SNP mutation were as follows. (1) All variants that have likelihood ratio: lik(AR)/lik(RR) or lik(AA)/lik(RR) $> 10^4$, in which R denotes the reference allele and A the alternative allele, in any of the 1,859 population samples, were excluded. Some recurrent mutations could have been filtered out, but the number should be small. The *de novo* mutation calls further satisfy the conditions that (2) there are at least 16 quality reads for the proband at the mutated site; (3) the likelihood ratio lik(AR)/lik(RR) is above 10^{10}; and (4) for both parents, the ratio lik(RR)/lik(AR) is above 100. Applying criteria (1) to (4) gave 6,221 candidate mutations. Further examination led us to apply extra filtering (5) by including only SNPs in which the number of A allele calls is above 30% among the quality sequence reads of the proband. This was considered necessary because there was an abnormally high number of putative mutation calls in which, despite extremely high lik(AR)/lik(RR) ratios for the proband, the fraction of A calls was low (Supplementary Fig. 1). Applying (5) eliminated 1,285 candidate mutations (Supplementary Information). With high coverage, the false negatives resulting from (5) is estimated to be a modest 2% (Supplementary Information). After three more candidates were identified as false

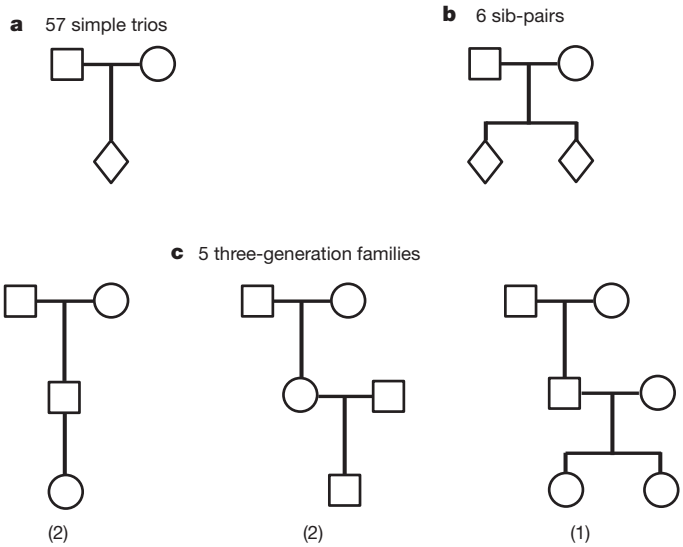

Figure 1 | A summary of the family types. **a**, Fifty-seven simple trios. **b**, Six sib-pairs accounting for 12 trios. **c**, Five three-generation families accounting for nine trios.

[1]deCODE Genetics, Sturlugata 8, 101 Reykjavik, Iceland. [2]Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark. [3]Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, UK. [4]University of Iceland, 101 Reykjavik, Iceland. [5]Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland.

positives by Sanger sequencing (see section on validation), a total of 4,933 *de novo* mutations, or an average of 63.2 per trio, were called. (The *de novo* mutations are listed individually in Supplementary Table 1.)

Parent of origin and father's age

For the five trios in which a child of the proband was also sequenced, the parent of origin of each *de novo* mutation called was determined as follows. If the paternal haplotype of the proband was transmitted to his/her child, and the child also carries the mutation, then the mutation was considered to be paternal in origin. If the child carrying the paternal haplotype of the parent does not have the mutation, then it is inferred that the mutation is on the maternal chromosome of the proband. Similar logic was applied when the child inherited the maternal haplotype of the proband. In the five trios, the average number of paternal and maternal mutations is 55.4 and 14.2, respectively (Table 1). If mutations were purely random with no systematic difference between trios, their number should be Poisson distributed with the variance equal to the mean. The data, however, show overdispersion (Table 1). This is much more notable for the paternal mutations (variance = 428.8, $P = 1.2 \times 10^{-5}$) than the maternal mutations (variance = 48.7, $P = 0.016$). Moreover, the number of paternal mutations has a monotonic relationship with the father's age at conception of the child. Here, the mean number of paternal mutations is substantially higher than the mean number of maternal mutations (ratio = 3.9), but the difference is even greater for the variance (ratio = 8.8). Hence, variation of *de novo* mutation counts in these individuals is mostly driven by the paternal mutations.

Relationships between parents' age and the number of mutations (paternal and maternal combined, as they could not be reliably separated without data from a grandchild) were examined using all 78 trios (Fig. 2). The number of mutations increases with father's age ($P = 3.6 \times 10^{-19}$) with an estimated effect of 2.01 mutations per year (standard error = 0.17). Mother's age is substantially correlated with father's age ($r = 0.83$) and, not surprisingly, is also associated with the number of *de novo* mutations ($P = 1.9 \times 10^{-12}$). However, when father's age and mother's age were entered jointly in a multiple regression, father's age remained highly significant ($P = 3.3 \times 10^{-8}$), whereas mother's age did not ($P = 0.49$). On the basis of existing knowledge about the mutational mechanisms in sperm and eggs[2], the results support the notion that the increase in mutations with parental age manifests itself mostly, maybe entirely, on the paternally inherited chromosome.

Given a particular mutation rate, due to random variation, the number of actual mutations is expected to have a Poisson distribution. After taking Poisson variation into account, with a linear fit (effect = 2.01 mutations per year), father's age explains 94.0% (90% confidence interval: 80.1%, 100%) (Supplementary Information) of the remaining variation in the observed mutation counts. When an exponential model is fitted (red curve in Fig. 2), the number of paternal and maternal mutations combined is estimated to increase by 3.23% per year. This model explains 96.6% (90% confidence interval: 83.2%, 100%) of the remaining variation. A third model fitted (blue curve in Fig. 2) assumes that the maternal mutation rate is constant at 14.2 and paternal mutations increase exponentially. This explains 97.1% (90% confidence interval: 84.3%, 100%) of the remaining variation and the rate of paternal mutations is estimated to increase by 4.28% per year, which corresponds to doubling every 16.5 years and increasing by 8-fold in 50 years. Seventy-six of the 78 trios have father's ages between 18 and 40.5, a range in which the differences between the three models are modest. Hence, although it seems that the number of paternal *de novo* mutations increases at a rate that accelerates with father's age, more data at the upper age range are needed to evaluate the nature of the acceleration better.

Table 1 | *De novo* mutations observed with parental origin assigned

	Father's age (yr)	Mother's age (yr)	Number of *de novo* mutations in proband		
			Paternal chromosome	Maternal chromosome	Combined
Trio 1	21.8	19.3	39	9	48
Trio 2	22.7	19.8	43	10	53
Trio 3	25.0	22.1	51	11	62
Trio 4	36.2	32.2	53	26	79
Trio 5	40.0	39.1	91	15	106
Mean	29.1	26.5	55.4	14.2	69.6
s.d.	8.4	8.8	20.7	7.0	23.5
Variance	70.2	77.0	428.8	48.7	555.3

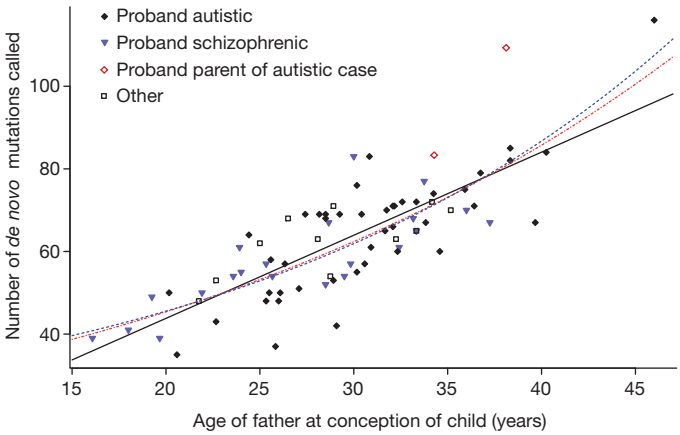

Figure 2 | Father's age and number of *de novo* mutations. The number of *de novo* mutations called is plotted against father's age at conception of child for the 78 trios. The solid black line denotes the linear fit. The dashed red curve is based on an exponential model fitted to the combined mutation counts. The dashed blue curve corresponds to a model in which maternal mutations are assumed to have a constant rate of 14.2 and paternal mutations are assumed to increase exponentially with father's age.

Validation and the nature of errors

Among the *de novo* mutations originally called, two were observed twice, both in siblings, one on chromosome 6 and one on chromosome 10. These cases were examined by Sanger sequencing. The mutation on chromosome 6 is not actually *de novo* as it was seen in the mother also. The one on chromosome 10 was confirmed, that is, it was observed in both siblings, who share the paternal haplotype in this region, but not the parents. This supports the theory that *de novo* mutations in different sperms of a man are not entirely independent[2]. Our trios include seven sib-pairs with 921 *de novo* mutations called. A false *de novo* mutation call for one sib resulting from a missed call in the parent would also show up in the other sib about 50% of the time. Only one such false positive was detected, indicating that this type of error accounts for a small percentage (2/(920/2) = 0.43%) of the called mutations. To evaluate the overall number of false positives, 111 called *de novo* mutations were randomly selected for Sanger sequencing. Eleven failed primer design. Six did not produce results of good quality in at least one member of the corresponding trio (Supplementary Information). For the remaining 94 cases, 93 were confirmed as *de novo* mutations—that is, the mutated allele was observed in the proband but not in the parents. One false positive, in which the putative mutation was not observed in the proband, was identified. The 17 cases that could not be verified are more likely to be located in genomic regions that are more difficult to analyse and hence probably have higher false-positive rates than average. Even so, the overall false-positive rate for the *de novo* mutation calls cannot be high.

The variance of the number of false positives is as important as the mean. False positives that are Poisson distributed, although adding noise, would not create bias for the effect estimates in either the linear

or the exponential models for father's age, nor would they bias the estimate of the fraction of variance explained after accounting for Poisson variation. In general, they do not create substantial bias for analyses of differences and ratios. However, if the variance of the false positives is higher than the mean, resulting from systematic effects that affect trios differently, such as DNA quality and library construction, it would increase the unexplained variance and reduce the fraction of variance explained by father's age. The candidates filtered out by criterion (5), if kept, would have introduced false positives of this kind (Supplementary Information). Because father's age explains such a high fraction of the systematic variance of the currently called *de novo* mutations, false positives with this property cannot be common. A similar discussion about false negatives[13] is in Supplementary Information.

Father's age and diseases

Consistent with other epidemiological studies[14,15], in Iceland, the risk of schizophrenia increases significantly with father's age at conception ($n = 569$, $P = 2 \times 10^{-5}$). Father's age is also associated with the risk of ASD. The observed effect is limited to non-familial cases ($n = 631$, $P = 5.4 \times 10^{-4}$), defined as those in which the closest ASD relative is farther than cousins. The epidemiological results, the effect of father's age on *de novo* mutation rate shown here, together with other studies that have linked *de novo* mutations to autism and schizophrenia, including three recent studies of autism through exome sequencing[4–6], all point to the possibility that, as a man ages, the number of *de novo* mutations in his sperm increases, and the chance that a child would carry a deleterious mutation (not necessarily limited to SNP mutations) that could lead to autism or schizophrenia increases proportionally. However, this model does not indicate that the relationship observed here between mutation rate and father's age would have been much different if the probands studied were chosen to be all non-ASD/schizophrenic cases instead. For example, assume that autism/schizophrenia is in each case caused by only one *de novo* mutation. Then autism/schizophrenia cases would on average have more *de novo* mutations than population samples. The magnitude could be substantial if the distribution of father's age has a large spread in the population, but then most of the difference would be caused by the cases having older fathers. If we control for the age of the father at the conception of the individual, then this difference in the average number of *de novo* mutations between control individuals and those with autism/schizophrenia would be reduced to approximately one (Supplementary Information).

Mutations by type and by chromosome

Examination of the 4,933 *de novo* mutations showed that 73 are exonic, including two stop-gain SNPs and 60 non-synonymous SNPs (Supplementary Table 2). One non-familial schizophrenic proband carries a *de novo* stop-gain mutation (p.Arg113X) in the neurexin 1 (*NRXN1*) gene, previously associated with schizophrenia[16–20]. One non-familial autistic proband has a stop-gain *de novo* mutation (p.R546X) in the cullin 3 (*CUL3*) gene. *De novo* loss of function mutations in *CUL3* have been reported to cause hypertension and electrolyte abnormalities[21]. Recently, a separate stop-gain *de novo* mutation (p.E246X) in *CUL3* was reported in an autistic case[5]. Another one of our mutations is a non-synonymous variant (p.G900S) two bases from a splice site in the EPH receptor B2 (*EPHB2*), a gene implicated in the development of the nervous system. A *de novo* stop-gain mutation (p.Q858X) in this gene has recently been described in another autistic case[6]. Given the small number of loss of function *de novo* mutations we and others have reported (approximately 70 genes in the three autism exome scans[4–6]), the overlap is unlikely to be a coincidence. Hence, *CUL3* and *EPHB2* can be added to the list of genes that are relevant for ASD. Effective genome coverage, computed by discounting regions that have either very low (less than half genome average) or very high (more than three times genome average) local coverage, the latter often a symptom of misaligning reads, was estimated to be 2.63 billion base pairs (Supplementary Information). From that, 4,933 mutations correspond to a germline mutation rate of 1.20×10^{-8} per nucleotide per generation, falling within the range between 1.1×10^{-8} and 3.8×10^{-8} previously reported[3,7,8,22,23]. Tables 2 and 3 summarize the nature of the *de novo* mutations with respect to sequence context. Approximately two-thirds (3,344/4,933 = 67.8%) are transitions. Moreover, there is a clear difference between mutation rates at CpG and non-CpG sites. CpG dinucleotides are known to be mutational hotspots in mammals, ostensibly because spontaneous oxidative deamination of methylated cytosines leads to an increase in transition mutations[24]. The observed rate of transitions here is 18.2 times that at non-CpG sites, higher than but not inconsistent with previous estimates of 13.3 (ref. 23) and 15.4 (ref. 3). The transversion rate is also higher at CpG sites, 2.55-fold that at non-CpG sites. Most of this increased transversion rate at CpG sites is presumably due to general mutation bias favouring mutations that decrease G+C content. The rate of mutations that change a strong (G:C) base pair to a weak (A:T) one is 2.15-times higher than mutations in the opposite direction. This mutational pressure in the direction of A+T is observed for both transitions (ratio = 2.24) and transversions (ratio = 1.82), and cannot be solely explained by CpG mutations. The father's age does not seem to affect the ratios between the rates of these different classes of mutations, that is, as a man ages rates of all mutation types increase by a similar factor.

Table 2 | Germline mutation rates at CpG and non-CpG sites

Type of mutation	*n*	Rate per base per generation
Transition at non-CpG	2,489	6.18×10^{-9}
Transition at CpG	855	1.12×10^{-7}
Transversion at non-CpG	1,516	3.76×10^{-9}
Transversion at CpG	73	9.59×10^{-9}
All	4,933	1.20×10^{-8}

Mutation rates are per generation per base. For non-CpG sites, the effective number of bases examined is taken as 2.583 billion, whereas for CpG sites the number is 48.8 million. These numbers take into account the variation of local coverage in sequencing (Supplementary Information).

The average number of mutations for each chromosome separately and the effect of father's age are displayed in Fig. 3. The effect of father's age is significant ($P < 0.05$) for 14 of the 22 chromosomes when evaluated individually. The solid line in the figure corresponds to a model in which the linear effect of father's age is proportional to the mean number of mutations on the chromosome, or that father's age has a uniform multiplicative effect across the chromosomes. All 22 95% confidence intervals overlap the line, indicating that the results are consistent with the model.

Discussion

The recombination rate is higher for women than men, and children of older mothers have more maternal recombinations that those of young mothers[25]. However, men transmit a much higher number of mutations to their children than women. Furthermore, even though our data also show some overdispersion in the number of maternal *de novo* mutations, it is the age of the father that is the dominant factor in determining the number of *de novo* mutations in the child. Seeing an association between father's age and mutation rate is not surprising[2], but the large linear effect of more than two extra mutations per year, or the estimated exponential effect of paternal mutations doubling every 16.5 years, is striking. Even more so is the fraction of the

Table 3 | Strong-to-weak and weak-to-strong mutation rates

Mutation type	S→W (*n*) rate	W→S (*n*) rate	S→W rate/W→S rate
Transition	(2,025) 1.21×10^{-8}	(1,319) 5.42×10^{-9}	2.24
Transversion	(446) 2.67×10^{-9}	(358) 1.47×10^{-9}	1.82
All	(2,471) 1.48×10^{-8}	(1,677) 6.89×10^{-9}	2.15

n denotes observed mutation counts, and mutation rates are calculated per generation per base. For strong (S; G:C) to weak (W; A:T), the effective number of sites examined is taken as 1.071 billion, and for weak to strong the number is 1.56 billion.

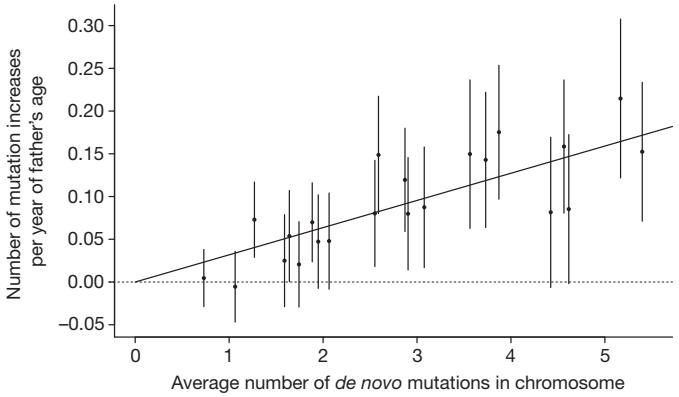

Figure 3 | Effect of father's age by chromosome. By chromosome, the estimated increase in the number of *de novo* mutations per year of father's age is plotted against the average number of mutations observed. The 95% confidence intervals are given. The solid straight line corresponds to the model in which the additive effect of father's age on the number of *de novo* mutations is assumed to be proportional to the mean number of mutations on the chromosome. From left to right, the points correspond to chromosome 21, 22, 19, 20, 15, 17, 18, 14, 16, 13, 12, 9, 10, 11, 8, 7, 6, 3, 5, 4, 2 and 1.

variation it explains, which limits the possible contribution by other factors, such as the environment and the genetic and non-genetic differences between individuals, to mutation rate on a population level. Given the results, it may no longer be meaningful to discuss the average mutation rate in a population without consideration of father's age. Also, even though factors other than father's age do not seem to contribute substantially to the mutation rate diversity in our data, it does not mean that hazardous environmental conditions could not cause a meaningful increase in mutation rate. Rather, the results indicate that, to estimate such an effect for a specific incident, it is crucial to take the father's age into account.

It is well known that demographic characteristics shape the evolution of the gene pool through the forces of genetic drift, gene flow and natural selection. With the results here, it is now clear that demographic transitions that affect the age at which males reproduce can also have a considerable effect on the rate of genomic change through mutation. There has been a recent transition of Icelanders from a rural agricultural to an urban industrial way of life, which engendered a rapid and sequential drop in the average age of fathers at conception from 34.9 years in 1900 to 27.9 years in 1980, followed by an equally swift climb back to 33.0 years in 2011, primarily owing to the effect of higher education and the increased use of contraception (Fig. 4). On the basis of the fitted linear model, whereas individuals born in 1900 carried on average 73.7 *de novo* mutations, those born in 1980 carried on average only 59.7 such mutations (a decrease of 19.1%), and the mutational load of individuals born in 2011 has increased by 17.2% to 69.9. Demographic change of this kind and magnitude is not unique to Iceland, and it raises the question of whether the reported increase in ASD diagnosis lately is at least partially due to an increase in the average age of fathers at conception. Also, the observations here are likely to have important implications for the use of genetic variation to estimate divergence times between species or populations, because the mutation rate cannot be treated as a constant scaling factor, but rather must be considered along with the paternal generation interval as a time-dependent variable.

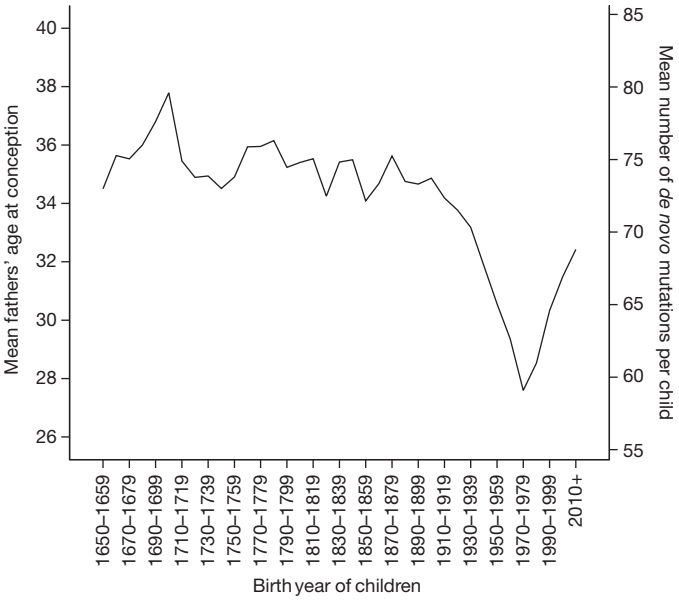

Figure 4 | Demographics of Iceland and *de novo* mutations. The deCODE Genetics genealogy database was used to assess fathers' age at conception for all available 752,343 father–child pairs, in which the child's birth year was ≥1650. The mean age of fathers at conception (left vertical axis) is plotted by birth year of child, grouped into ten-year intervals. On the basis of the linear model fitted for the relationship between father's age and the number of *de novo* mutations, the same plot, using the right vertical axis, shows the mean number of expected mutations for each ten-year interval.

METHODS SUMMARY

Whole-genome sequence data for this study were generated using the Illumina $GAll_x$ and HiSeq2000 instruments. The sequencing reads were aligned to the hg18 reference genome with Burrows–Wheeler aligner (BWA)[26] and duplicates were marked with Picard (http://picard.sourceforge.net/). Quality score recalibration, indel realignment and SNP/indel discovery were then performed on each sample separately, using GATK 1.2 (ref. 27). Likelihoods presented are based on the normalized Phred-scaled likelihoods that are calculated by the GATK variant calling. Statistical analysis was performed in part using the R statistical package. Estimates and confidence intervals for the fraction of variance explained after accounting for Poisson variation were calculated using Monte Carlo simulations (Supplementary Information). Variants were annotated using SNP effect predictor (snpEff2.0.5, database hg36.5) and GATK 1.4-9-g1f1233b with only the highest-impact effect (P. Cingolani, 'snpEff:Variant effect prediction', http://snpeff.sourceforge.net, 2012). More details are in Supplementary Information.

Received 28 February; accepted 4 July 2012.

1. Keightley, P. D. Rates and fitness consequences of new mutations in humans. *Genetics* **190,** 295–304 (2012).
2. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1,** 40–47 (2000).
3. Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21,** 12–27 (2003).
4. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
5. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485,** 246–250 (2012).
6. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).
7. 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).
8. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43,** 712–714 (2011).
9. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328,** 636–639 (2010).
10. Holm, H. *et al.* A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nature Genet.* **43,** 316–320 (2011).
11. Rafnar, T. *et al.* Mutations in *BRIP1* confer high risk of ovarian cancer. *Nature Genet.* **43,** 1104–1107 (2011).
12. Sulem, P. *et al.* Identification of low-frequency variants associated with gout and serum uric acid levels. *Nature Genet.* **43,** 1127–1130 (2011).
13. Keightley, P. D. *et al.* Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* **19,** 1195–1201 (2009).
14. Malaspina, D. Paternal factors and schizophrenia risk: de novo mutations and imprinting. *Schizophr. Bull.* **27,** 379–393 (2001).
15. Croen, L. A., Najjar, D. V., Fireman, B. & Grether, J. K. Maternal and paternal age and risk of autism spectrum disorders. *Arch. Pediatr. Adolesc. Med.* **161,** 334–340 (2007).

16. Duong, L. *et al.* Mutations in *NRXN1* in a family multiply affected with brain disorders: *NRXN1* mutations and brain disorders. *Am. J. Med. Genet.* **159B,** 354–358 (2012).
17. Gauthier, J. *et al.* Truncating mutations in *NRXN2* and *NRXN1* in autism spectrum disorders and schizophrenia. *Hum. Genet.* **130,** 563–573 (2011).
18. Kirov, G. *et al.* Comparative genome hybridization suggests a role for *NRXN1* and *APBA2* in schizophrenia. *Hum. Mol. Genet.* **17,** 458–465 (2008).
19. Levinson, D. F. *et al.* Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and *VIPR2* duplications. *Am. J. Psychiatry* **168,** 302–316 (2011).
20. Rujescu, D. *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum. Mol. Genet.* **18,** 988–996 (2009).
21. Boyden, L. M. *et al.* Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. *Nature* **482,** 98–102 (2012).
22. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107,** 961–968 (2010).
23. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
24. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274,** 775–780 (1978).
25. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nature Genet.* **36,** 1203–1206 (2004).
26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
27. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was partly funded by The National Institutes of Health grant MH071425 (K.S.); the European Community's Seventh Framework Programme, PsychCNVs project, grant agreement HEALTH-F2-2009-223423, and NextGene project, grant agreement IAPP-MC-251592; The European Community IMI grant EU-AIMS, grant agreement 115300.

Author Contributions A.K. and K.S. planned and directed the research. A.K. wrote the first draft and together with K.S., S.B., P.S., A.H. and U.T. wrote the final version. O.T.M. and U.T. oversaw the sequencing and laboratory work. G. Masson, G. Magnusson and G.S. processed the raw sequencing data. A.K. and M.L.F. analysed the data, with W.S.W.W., H.H., G.B.W., S.S., G.T. and D.F.G. providing assistance. P.S. and S.A.G. performed functional annotations. S.B. analysed the mutations with respect to sequence content. A.S., Aslaug J. and Adalbjorg J. did the Sanger sequencing. A.H. investigated the contribution of demographics.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.K. (kong@decode.is) or K.S. (kari.stefansson@decode.is).