Rate of *de novo* mutations, father's age, and disease risk

Supplementary Information

Running title: Rate of *de novo* mutations

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 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,*].

***Corresponding authors:**

Augustine Kong, deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland.

kong@decode.is, Phone: 354-5701931, fax 354-5702850.

Kari Stefansson, deCODE genetics, Sturlugata 8, 101 Reykjavík, Iceland. kari.stefansson@decode.is, Phone:354-5701900, fax 354-5701901.

[1]deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland.

[2]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark.

[3]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL, United Kingdom.

[4]University of Iceland, Reykjavik, Iceland.

[5]Faculty of Medicine, University of Iceland, Reykjavik, Iceland.

1. Study samples.

A total of 2078 samples from a large sequencing project at deCODE were used in this study, 219 samples from 78 trios with two grandchildren who were not also members of other trios, along with 1859 population samples. For the offspring members of each trio, 44 were classified with Autism Spectrum Disorder (ASD) according to ICD-10 criteria using the Autism Diagnostic Interview-Revised (Lord, C., Rutter, M. & Le Couteur, A. 1994), and 21 were classified as having schizophrenia as diagnosed according to Research Diagnostic Criteria (Spitzer, R.L., Endicott, J. & Robins, E. 1978) using the Schedule for Affective Disorders and Schizophrenia Lifetime Version (Spitzer, R.L. & Endicott, J. 1977). The probands from the remaining 13 trios have neither diagnosis.

All biological samples used in this study were obtained according to protocols approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. Informed consent was obtained from all participants and all personal identifiers were encrypted with a code that is held by the Data Protection Commission of Iceland.

2. Preparation of samples for whole genome sequencing.

The TruSeq™ sample preparation kit (Illumina) was employed for the preparation of libraries for whole genome sequencing (WGS). In short, approximately 1 μg of genomic DNA, isolated from frozen blood samples, was fragmented to a mean target size of approximately 300-400 bp using a Covaris E210 instrument. The resulting fragmented DNA was end repaired using T4 and Klenow polymerases and T4 polynucleotide kinase with 10 mM dNTP followed by addition of an 'A' base at the ends using Klenow exo fragment (3′ to 5′-exo minus) and dATP (1 mM). Sequencing adaptors containing 'T' overhangs were ligated to the DNA products followed by agarose (2%) gel electrophoresis. Fragments of about 450-500 bp were isolated from the gels (QIAGEN Gel Extraction Kit), and the adaptor-modified DNA fragments were PCR enriched for ten cycles using Phusion DNA polymerase (Finnzymes Oy) and PCR primers PE 1.0 and PE 2.0 needed for paired-end sequencing. Enriched libraries were purified using AMPure XP beads. The quality and concentration of the

libraries were assessed with the Agilent 2100 Bioanalyzer using the DNA 1000 LabChip. Libraries were stored at −20 °C. All steps in the workflow were monitored using an in-house laboratory information management system (LIMS) with barcode tracking of all samples and reagents.

3. DNA whole genome sequencing.

Template DNA fragments were hybridized to the surface of paired-end (PE) flowcells (either for $GAII_x$ or HiSeq 2000 sequencing instruments) and amplified to form clusters using the Illumina cBot™. In brief, DNA (3–12 pM) was denatured, followed by hybridization to grafted adaptors on the flowcell. Isothermal bridge amplification using Phusion polymerase was then followed by linearization of the bridged DNA, denaturation, blocking of 3′ ends and hybridization of the sequencing primer.

Sequencing-by-synthesis (SBS) was performed on either Illumina $GAII_x$ or HiSeq 2000 instruments, respectively. Paired-end libraries were sequenced using 2x120 cycles of incorporation and imaging with Illumina SBS kits, TruSeq™ v5 for the GAIIx. For the HiSeq 2000, 2x101 cycles with SBS kits v2.5 or v3 were employed. Each library was initially run on a single lane on a $GAII_x$ for validation, assessing optimal cluster densities, insert size, duplication rates and comparison to chip genotyping data. Following validation, the desired sequencing depth (either 10X or 30X) was then obtained using either sequencing platform. Targeted raw cluster densities ranged from 500–800 K/mm^2, depending on the version of both the sequencing chemistry and the data imaging/analysis software packages (SCS.2.8/RTA1.8 or SCS2.9/RTA1.9 for the $GAII_x$ and HCS1.3.8. or HCS1.4.8 for HiSeq 2000). Real-time analysis involved conversion of image data to base-calling in real-time.

4. Sequence alignments and variants calling.

For each lane in the DNA sequencing output, the resulting qseq files were converted into fastq files using an in-house script. All output from sequencing was converted, and the Illumina quality filtering flag was retained in the output. The fastq files were then aligned against Build 36 of the human reference sequence using bwa version 0.5.9 (Li, H. & Durbin, R. 2009).

SAM file output from the alignment was converted into BAM format using samtools version 1.1.18 (Li, H. *et al* 2009), and an in-house script was used to carry the Illumina quality filter flag over to the BAM file. The BAM files for each sample were then merged into a single BAM file using samtools. Finally, Picard (versions from 1.17 to 1.55) (http://picard.sourceforge.net) was used to mark duplicates in the resulting BAM files.

GATK 1.2 (McKenna, A. *et al* 2010) was used for quality score recalibration and indel realignment. SNP/Indel discovery was then performed by GATK 1.2 on each sample separately using standard filtering parameters as recommended (http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3). For variant discovery, a confidence level threshold of 50.0 was used, which was slightly higher than was recommended for DEEP (>10X) coverage. The discovery set of SNPs and indels for the individuals, restricted to variants with lik(RR)/lik(RA) > 10^4 , max(lik(RA)/lik(AA), lik(AA)/lik(RA)) > 10^3 and local coverage less than three times the sample's average coverage, were merged using a combination of in-house scripts (similar to the CombineVariants tool in GATK) and the individuals were then recalled for the merged variant set. Variant sites were investigated as potential *de novo* mutations for each trio if none among the other sequenced individuals (excluding first degree relatives of the trio proband) had a lik(RR)/lik(RA) ratio greater than 10^4. All likelihoods evaluated here are based on the normalized Phred-scaled likelihoods calculated by the GATK variant caller (UnifiedGenotyper).

5. The reason for and the effect of applying filter (v).

As noted in the main text, after applying criteria (i) to (iv), there were 6,221 candidate *de novo* mutations remained. Two of these, on chromosome 6, were identical and seen in two siblings. Validation by Sanger sequencing revealed that the variant is actually also carried by the mother, and hence not really *de novo*. Removing these left us with 6,219 candidate mutations. (One of the 6,219, as noted in the main text, was revealed at a much later stage as a false positive by Sanger sequencing. This case is included in the analysis described here because we think this better reflects what led us to apply filter (v) in the first place. But, of course, removing it from this analysis would make very little difference.) For each of these called variants, among the quality reads, the

fraction of A (alternative allele) calls was calculated. **Supplementary Fig. 1** is a histogram of the 6219 fractions. The histogram has two modes, one at 50%, and one at 20% to 25%. This suggests a mixture of two distributions, one representing true heterozygotes with a mode at 50%, and one representing erroneous calls with a mode at a much lower percentage. Many of the cases contributing to the latter probably resulted from having reads from two or more different, but highly similar, regions mixed up together. For example, if two sites are mixed together, one is heterozygous and the other homozygous reference (RR), the fraction of A reads would be 25% in expectation. It could sometimes be one in 6, or 16.7%, if reads from three different sites were misaligned to one location. Out the 6219 candidates, fraction of A calls are at or below 30% for 1285 of them. Filtering these out from the set of 6219 gave a set 4,934. It is interesting to note that, if we did not eliminate the 1285 cases and performed the analysis with 6219 *de novo* mutations called, the estimated effect of father's age would be very similar, actually a little higher (2.30 mutations per year as opposed to 2.01), but the significance and fraction of variance explained after accounting for Poisson variation would be substantially reduced ($P = 7.6\times10^{-14}$ and variance explained = 67.7%, as opposed to $P = 3.9\times10^{-19}$ and variance explained = 93.9% when using the 4934 called mutations). That the estimate is higher is possibly because there are some true positives in the 1285 cases. The *P* value is less significant and variance explained is substantially lower because the 1285 cases are introducing a lot of noise. In particular, the 1285 cases exhibit substantial over dispersion, variance/mean = 3.9, and only a small fraction of that could be accounted for by father's age. We can get a rough estimate of how many true positives are in the 1285 cases in two ways. Firstly, with 30 reads, with a true heterozygote, the probability of having 9 or less A reads is 2.1%. Given that there are about 5000 *de novo* mutations in our trios, that corresponds to about 105 true positives filtered away, or 105 false negatives introduced. Secondly, from the histogram (**Supplementary Fig. 1**), there are 71 mutation calls with the fraction of A reads greater than or equal to 70%. Assuming symmetry, it would imply that about 71 true positives were filtered away by (v), and corresponds to a false negative rate of about 1.4%. Taking these two estimates into account, we believe that the filter (v) is likely to be responsible for about 2% of false negatives.

6. Models fitted, estimating fraction of variance explained and confidence intervals.

As noted in the main text, we fitted 3 models to evaluate the relationship between father's age and number of *de novo* mutations. Let Y denote the number of *de novo* mutations, and let X be the age of the father at conception of the child. The linear model was fitted by performing a simple regression of Y on X. The first exponential model fitted was done by regressing log(Y) on X. The second exponential model fitted was performed by regressing log(Y – 14.2) on X, noting that 14.2 was chosen because that is the mean number of maternal *de novo* mutations observed in the 5 trios for which parent of origin of the mutations could be determined. For the exponential fits, residual sum of squares and variance explained were calculated by converting the fitted values back to the original scale. Note that the same number of parameters, an intercept and a slope, were fitted in all 3 cases. The difference is just the scale under which the regression was performed. Because the 3 models are not nested, one cannot test one against another in a standard frequentist manner and compute *P*-values. But we note the following. If we add a quadratic term of X to the linear fit, the quadratic term is marginally significant with $P = 0.07$. But even with the quadratic term added, R^2, the fraction of variance explained, is still slightly lower than those resulting from fitting the two exponential models. Hence it is reasonable to say the exponential models fit the data better than the linear model.

For a Poisson distribution the variance is equal to the mean. Hence, using the data, a simple estimate of the fraction of variance explained after accounting for Poisson variation is

$R^2/[1 - \text{mean}(Y)/\text{var}(Y)]$ (*)

where R^2 is the fraction of total variance explained by father's age obtained from the model fit. For the linear fit, to slightly improve this estimate and to construct confidence intervals, we performed Monte Carlo simulations based on the following model:

$Y \sim \text{Poisson}(A + B*\text{age} + \text{Normal}(0,\text{SIGMA}))$.

We set A and B to the fitted values. By varying SIGMA, we could set the theoretical value of the fraction of systematic variance explained by father‘s age to any value we like. From the simulations, we found that the simple estimate (*) is slightly biased, in the sense that its sampling distribution has a mean/median that is a little higher (about 0.5%) than the actual value used to do the simulation. So we centered the estimate by choosing the value so that when it is used to perform the simulations, the median of the simulated values of (*) will correspond to the observed value calculated from the real data. Similarly, the lower bound of the 90% CI is the value so that, when it is used to perform the simulations, the 95th percentile of the simulated values of (*) will correspond to the observed value.

A similar method is used to obtain estimates and confidence intervals for the exponential fits.

7. Some details on the Sanger sequencing results.

One hundred and eleven of the *de novo* mutations called were randomly selected for validation using Sanger sequencing. Eleven failed primer design. For the 100 cases where we obtained primers, the first run generated reliable results for 86 of them, and all confirmed as *de novo* mutations. The 14 cases that failed to generate reliable results were rerun. Results were obtained for 8 cases, with 7 confirmations and one false positive identified where the putative variant was not observed in the proband. Hence, overall, we have 93 = 86+7 confirmations and one identified false positive. Among the other six cases, two of them had problems with the PCR and did not generate any useful results at all. For the other four cases, the mutation was seen in proband and not in the mother, but reliable results could not be obtained for the father due apparently to problems with low quality DNA. Hence, the data for these four cases, while not conclusive, are consistent with true *de novo* mutations.

8. The impact of false negatives on various analyses.

In the main text, we discussed how false positives of various types could impact the analyses. Here is a similar discussion on false negatives. Because of the limitations of current sequencing technology and that the methods used to call the variants are still far from perfect, we had to apply filters to limit the false positives. As a result, false

negatives are unavoidable. While not attempting to give a precise estimate of its overall magnitude, we note that the overall mutation rate observed here is not inconsistent with other estimates reported for trio data. For the analyses of father's age, false negatives that are Poisson in nature will bias the effect estimate of the linear model downwards, implying that the actual effect genomewide is very likely to be above the current estimate of 2.01 per year. However, the effect estimate for the exponential model, and, in general estimates of ratios, should not be substantially affected. Non-Poisson false negatives would add to the unexplained variance, and, following the same argument applied to false positives, their magnitude is likely to be modest.

9. Average number of *de novo* mutations in cases and in controls.

Suppose non-familial ASD/SZ cases are in each case caused by one (and only one) *de novo* mutation. Suppose, while father's age has an effect on the number of *de novo* mutations, its effect is in a multiplicative sense uniform over the genome. And suppose there is no other systematic factors influencing the number of mutations (or that their contributions are very small on a population level) other than Poisson variation. Then the chance of an individual being a case is essentially proportional to the number of *de novo* mutations they carry, e.g. a person carrying 120 *de novo* mutations will have 3 times the chance of being a case than those that carry 40. Of interest here is the reverse question --- what is the average number of *de novo* mutations in cases, and more specifically, on average how many more *de novo* mutations do cases have compared to population controls. The answer depends on the spread of the population distribution of the number of *de novo* mutations, the greater the spread the greater the difference. Mathematically, if X is a random variable having the mutation count distribution, then the difference is

$$[\text{mean}(X^2)/\text{mean}(X)] - \text{mean}(X)$$

But the spread of *de novo* mutation count distribution is driven, apart from Poisson variation, by father's age. Using the father's age distribution for 97,095 births in Iceland in the last century (mean = 31.7, SD = 6.31), and assuming the exponential model that was fitted for paternal mutations assuming that maternal mutation rate is fixed at 14.2, the above difference is estimated to be 4.70. However, a large fraction

of the effect is the consequence of the cases on average having older fathers than the controls. If we condition/adjust for father's age, that essentially means we are comparing cases and controls at a fixed age. In that case all variation comes from Poisson variation. So the difference above can be calculated by assuming that X has a Poisson distribution with some fixed mean. It happens that, regardless of the mean of the Poisson distribution, the above difference is 1.

10. Classifying *de novo* mutations by function and with respect to genes.

See **Supplementary Tables 1** and **2**.

11. Effective coverage of whole genome sequencing.

The effective genome coverage is based on the sum of the read depth over all 2,078 sequenced individuals. The initial coverage includes 2.628 billion non-CpG bases and 53.40 million CpG bases, a total of 2.681 billion. To calculate effective coverage, we applied one lower bound: (i) local coverage has to be above 50% of average genome coverage, and one upper bound: (ii) local coverage is no more than 3 times average genome coverage. After filtering using (i) and (ii), 2.583 billion non-CpG bases remained and 48.80 million CpG bases remained. Note that while less than 2% of non-CpG bases were filtered out, over 8% of CpG bases were filtered out. This is in part due to the fact that CpG bases are in regions that are GC rich, locations where current sequencing technology tends to have lower coverage (Wang et al. 2011). Note that according to (Keightley et al. 2009), mutation rate estimates stabilize at sites with a read depth above 4. Therefore, since our average sequencing depth is high, we expect only a small fraction of the genome needed to be removed when considering the coverage.

Criteria (i) and (ii) are chosen do deal with problematic regions for the sequencing technology we are employing. The boundaries were chosen after looking at the quality of SNP calls at different read depths. Variant calls in genomic regions with low coverage can both inflate the false positive rate (Keightley et al. 2009) and overlook mutations, which increases the false negative rate. Excessive coverage can correspond to regions with low complexity (e.g., in the vicinity of the centromeres

and telomeres) and sequence repeats, imposing challenges for read alignment and increasing the chance of calling false mutations.

12. The list of 4,933 *de novo* mutations

The attached excel file **Supplementary Table 1** contains information for each of the 4,933 *de novo* mutations individually. They correspond to the summary in **Supplementary Table 2**. The positions are based on Human Assembly Build 36.

References

Spitzer, R.L., Endicott, J. & Robins, E. Research diagnostic criteria: rationale and reliability. *Arch Gen Psychiatry* **35**, 773-782 (1978).

Spitzer, R.L. & Endicott J. *Schedule for Affective Disorders and Schizophrenia-Lifetime version (SADS-L)* 3rd edition, New York. New York State Psychiatric Institute (1977).

Lord, C., Rutter, M. & Le Couteur, A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* **24**, 659-695 (1994).

Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**,1297-1303 (2010).

Wang, W., Wei, Z., Lam T.-W., and Wange, J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Sci Rep.* 1:55 (2011).

Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar S., and Blaxter, M.L. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res.* 19: 1195-1201 (2009)

Supplementary Table 2. Breakdown by gene context

Gene Content	Count of Mutations
Non_synonymous coding	60
Stop_gained	2
Synonymous coding	11
UTR_3_prime	16
Upstream	175
Downstream	267
Intergenic	2589
Intron	1808
Transcript*	5

*It includes 4 pseudogenes and 1 Immunoglobulin gene.

Variants were annotated using SNP effect predictor (snpEff2.0.5, database hg36.5) and Genome Analysis Toolkit 1.4-9-g1f1233b with only the highest-impact effect (Cingolani, P. "snpEff:Variant effect prediction", http://snpeff.sourceforge.net, 2012).

Supplementary Figure 1.