Detection of structural variants and indels within exome data

Emre Karakoc[1], Can Alkan[1,2], Brian J O'Roak[1], Megan Y Dennis[1], Laura Vives[1], Kenneth Mark[1], Mark J Rieder[1], Debbie A Nickerson[1] & Evan E Eichler[1,2]

We report an algorithm to detect structural variation and indels from 1 base pair (bp) to 1 Mbp within exome sequence data sets. Splitread uses one end–anchored placements to cluster the mappings of subsequences of unanchored ends to identify the size, content and location of variants with high specificity and sensitivity. The algorithm discovers indels, structural variants, *de novo* events and copy number–polymorphic processed pseudogenes missed by other methods.

Although the proportion of structural variants and small insertions and deletions (indels; <50 bp) detected in sequence databases has increased exponentially[1,2], recent comparisons of both experimental and computational methods suggest that the false negative rate remains high[3,4]. In addition to whole-genome sequencing, the widespread use of exome-capture technologies that target genomic protein-coding regions is a resource for discovering structural variants and indels associated with disease. The nature of the capture methods, limited size of coding regions and nonuniform distribution of the reads pose considerable computational challenges. As a result, variants >15 bp have rarely been reported in exome studies[5,6]. Discovery has been based largely on sequence alignment gaps limited to uniquely mapped regions of the genome (GATK[7] or SAMtools[8]). Here we describe a general combinatorial algorithm (Splitread) and validate it for discovery of indels and structural variants in exome data sets.

We developed Splitread to detect structural variants and indels on the basis of computational prediction of breakpoints (see Online Methods and **Supplementary Note** for details). Similar to Pindel[9], another split read–based approach for detecting breakpoints of indels via a regional search around the anchored reads within the maximum event size, our algorithm searches for clusters of mate pairs in which one end maps to the reference genome, but the other end does not because it traverses a breakpoint, creating a mapping inconsistency with respect to the reference sequence (**Fig. 1a**). We initially mapped reads using mrsFAST[10], which guarantees all possible placements within a given Hamming distance (reflecting the number of allowed mismatches). Next, we decomposed the unmapped end into subsequences of either equal length (balanced splits) or unequal length (unbalanced splits). Unlike Pindel, which uses pattern growth for optimal matching in the target region, we reiteratively searched for clusters of split reads using the balanced splits as seeds (**Fig. 1a**), which refine the location and size of the indel or structural variant event. We applied weighted set-cover approximation (**Supplementary Note**) to minimize the number of possible breakpoints, providing a maximum parsimony framework for all the mappings at the breakpoints.

We tested different thresholds for the number of balanced and unbalanced splits required to support a call. For each configuration, we plotted the proportion of events called by the 1000 Genomes Project that was predicted by Splitread for sample NA12891 (**Fig. 1b** and **Supplementary Table 1**). The slope provides the positive predictive value (PPV), and we maximized sensitivity (number of corroborated predictions) without any loss of specificity by selecting the local maximum of this line. At a threshold of at least two balanced and two unbalanced splits, we predicted 213 indel events <50 bp in the NA12891 exome, of which 69% (148) intersect with whole-genome sequence analysis (**Fig. 1c**) and 72% (154) intersect with dbSNP130 (ref. 2). As we expected for protein-coding sequence[11], indel sizes were predominantly in multiples of three, leading to no disruption of the protein-coding frame (47% or 100/213; **Fig. 1d**). If we exclude 1-bp indels, this fraction increases to 78% (100/129). We applied this threshold for the remainder of our analysis.

We identified an additional 63 structural variant events (>50 bp) after excluding annotated processed pseudogenes (**Supplementary Table 2**). Although only four of these were predicted by the 1000 Genomes Project, nine of the remaining events intersect with structural variants from dbSNP130, with sizes varying from 51 bp to 3,584 bp. We predict that 48 of these variants are common (observed in multiple HapMap samples we analyzed) and that only 21 variants are specific to NA12891. Several correspond to genes known to carry complex insertion and deletion polymorphisms or variable number of tandem repeats such as *MUC6*, *DSPP* and *MUC16* (ref. 12).

We compared Splitread with alternative indel detection methods Pindel[9] and GATK[7] (see **Supplementary Note** for comparison to CREST). Some 70% of Splitread calls are predicted by one of the other methods but a substantial fraction of calls are unique to each method. As we expected, indel events called by two or more methods had the highest concordance with dbSNP and 1000 Genomes calls (**Fig. 1e**). We selected 19 events uniquely called by Splitread and previously not reported by dbSNP or 1000 Genomes for PCR-based validation. We validated 13 of 19 events (**Supplementary Table 1**), giving an estimated PPV of 68%. Most map within

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

RECEIVED 15 JULY; ACCEPTED 16 NOVEMBER; PUBLISHED ONLINE 18 DECEMBER 2011; DOI:10.1038/NMETH.1810

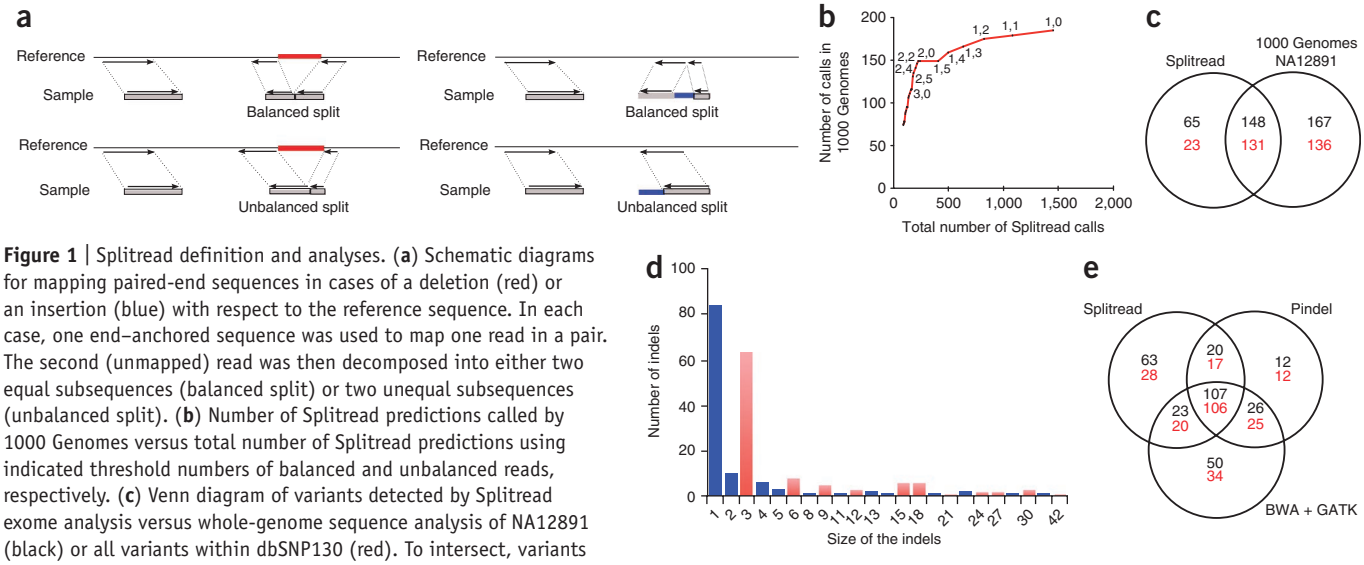

Figure 1 | Splitread definition and analyses. (**a**) Schematic diagrams for mapping paired-end sequences in cases of a deletion (red) or an insertion (blue) with respect to the reference sequence. In each case, one end–anchored sequence was used to map one read in a pair. The second (unmapped) read was then decomposed into either two equal subsequences (balanced split) or two unequal subsequences (unbalanced split). (**b**) Number of Splitread predictions called by 1000 Genomes versus total number of Splitread predictions using indicated threshold numbers of balanced and unbalanced reads, respectively. (**c**) Venn diagram of variants detected by Splitread exome analysis versus whole-genome sequence analysis of NA12891 (black) or all variants within dbSNP130 (red). To intersect, variants must be at same position and within 10 bp of the predicted size. (**d**) Length distribution of predicted insertions and deletions mapping within coding region of NA12891. Red, events with multiples of 3 bp; blue, events that would disrupt the frame. (**e**) Venn diagram of Pindel, GATK and Splitread call sets on NA12891. Black, total number of events; red, events previously detected as part of dbSNP130 and/or the 1000 Genomes Project.

low-complexity regions and correspond to repeat expansions and deletions (**Supplementary Table 1**). If we include previously reported events, Splitread accuracy was 87% (41/47).

We extended our analyses by generating exome sequence data from 11 HapMap samples whose genomes were sequenced at three- to fourfold coverage by the 1000 Genomes Project (**Supplementary Table 3**). Using Splitread, we observed an average of 325 events for each sample, including 286 indels and 39 structural variants (5:1 ratio). About 68% and 70% of the calls intersected 1000 Genomes and dbSNP130 predictions, respectively. From the 11 samples, we identified 192 previously unknown structural variants, 93 of which were observed two or more times; an average of nine events disrupting genes were unique to each individual (**Supplementary Tables 2** and **3**) .

As a final test, we applied Splitread to published exome data from 20 parent-child trios affected with sporadic autism-spectrum disorder[6]. We identified an average of 191 indels and 57 structural variants in this data set (**Supplementary Table 4**). To test the accuracy of our calling method, we randomly selected indels and structural variants not found in either dbSNP or the control individuals as part of the Exome Sequencing Project (http://esp.gs.washington.edu). We confirmed 10 of 12 events by PCR and sequencing, giving an estimated PPV of 83% (**Supplementary Table 5**). This included bona fide variation within repetitive and low-complexity regions such as a triplet and 12-mer insertion within a low-complexity coding portion of *SHROOM4* (**Supplementary Fig. 1**) missed by Pindel[9] and GATK[7].

A goal of parent-child trio sequencing is to discover potentially disruptive *de novo* events. This is challenging because the selection of potential *de novo* events either enriches false positives or represents inherited variants that were not detected (false negatives) in one of the parents. In this study, we detected and confirmed only one previously reported *de novo* variant, in *FOXP1* (ref. 6). The remaining events were either present in a parent or were false positives (**Supplementary Table 1**). We sought to increase our confidence in predicting *de novo* events by filtering via read depth. Because our method uses Hamming distance to align reads, structural variant and indel breakpoints should cause fewer reads to map in the affected child if the event is truly *de novo* (**Supplementary Note**). We added this functionality as a filter that normalizes the read depth of coding regions on the basis of coverage and then compares proband and parents to flag regions of reduced depth. We applied the filter specifically at predicted breakpoints to minimize false positives (**Supplementary Fig. 2**).

During our analysis of exome data sets, we routinely detected putative deletion events in which an intron was precisely removed such that flanking exons abutted perfectly. The structure of these events suggested the presence of uncharacterized processed pseudogenes as opposed to allelic deletions. These arise as a result of retrotransposition of spliced mRNA back into the genome. We discovered 25 such events in the 11 HapMap exomes (**Supplementary Table 6**), 14 of which could not be identified by BLAST searches against the reference genome (GRCh37). DNA amplification of flanking exons yielded 16 products consistent with a processed pseudogene in the affected individual, whereas the other 9 seemed to be polymorphic in the population (**Fig. 2**). Because pseudogenes can create potential Splitread artifacts, we created a modified exome reference for mapping that includes known processed pseudogenes, segmental duplications and copy-number polymorphic pseudogenes. Compared with a whole-genome reference, this modified exome reference increases mapping speed by a factor of 10 with only a 2% difference in the number of calls. Thus, Splitread can be applied to many exomes in a computationally efficient manner to generate a database of bona fide exonic indels and structural variants.

To test the applicability of Splitread to whole-genome data sets, we analyzed the genome of an individual (ND06769) with amyotrophic lateral sclerosis with frontotemporal dementia (ALS-FTD)[13] with a hexanucleotide repeat expansion (GGCCCC) in *C9orf72*. This is the causal variant of chromosome 9p21–associated ALS-FTD. This repeat expansion was missed by GATK and was discovered only through manual inspection of the read alignments[13]. Although the insertion is too long to be fully characterized by a split-read method (~1.5 kbp), our algorithm

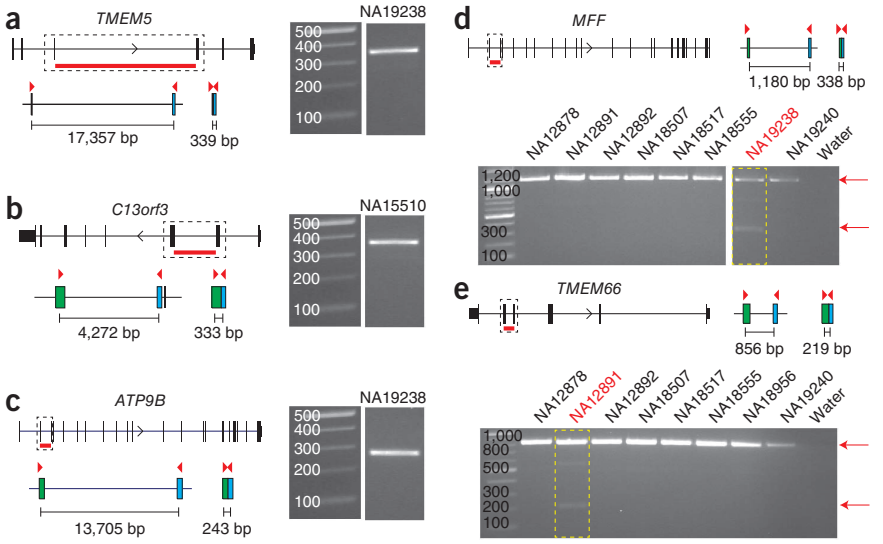

Figure 2 | Validation of processed pseudogenes. (**a**–**e**) Gene models and predicted intron deletions of processed pseudogenes. Primers (red triangles) were designed in coding regions. We detected the expected product size for processed pseudogenes for *TMEM5* (**a**), *C13orf3* (**b**), *ATP9B* (**c**), *MFF* (**d**) and *TMEM66* (**e**) in our PCR experiments. In **d**,**e** we genotyped the processed pseudogenes *MFF* and *TMEM66* within eight HapMap samples; each was amplified only in the predicted sample (boxed in yellow, NA19238 (*MFF*) and NA12891 (*TMEM66*)). All PCRs amplified the normal gene (signal on top), with only one sample each amplifying the processed gene.

discovered the approximate breakpoint of the expansion and supported the call with read-depth analysis. Splitread can detect insertions and deletions without size limitation. The size spectrum of the insertions that can be accurately characterized by Splitread is bound by the read length; however, the approximate breakpoints of larger insertions can be detected using one end–anchored reads.

Many validated events detected exclusively by Splitread involve microsatellite, low-complexity or polynucleotide tracts (**Supplementary Table 1** and **Supplementary Fig. 1**). Such regions are subject to higher mutation rates, owing in part to their greater potential for replication slippage[14]. Variation of this type, especially within coding regions, has frequently been associated with diseases including triplet repeat instability[14]. Our greater PPV for this class of variant stems from the fact that we considered multiple mappings frequently discarded by other methods. However, we clearly missed some genetic variation (**Fig. 1**), which emphasizes that no single approach comprehensively captures all genetic variation[3]. Splitread is limited by the dependence on balanced splits to seed an event, which directly depends on the coverage. Given 76-bp reads, the chance of detecting a heterozygous event is 55% at 20× coverage, but is >90% at 60× coverage. The sensitivity estimate is 79% at 20× coverage and 98% at 60× coverage. Such median sequence coverage is not uncommon in many exome sequencing projects.

In our exome analysis, we were surprised to discover many processed pseudogenes that are polymorphic but not represented in the human reference genome (**Supplementary Table 6**). We observed most of these variants more than once; they ranged in frequency from 3% to 72% on the basis of an assessment of 51 exomes (**Supplementary Table 6**). Using read-pair information, we mapped the location of all these polymorphisms using a one end–anchored mapping strategy[15]. A comprehensive catalog of the most common of these could be important for correctly interpreting disease-causing variants discovered in exome studies.

Because different methods vary in their sensitivity and specificity depending on the size, class and context of variants, multiple approaches should be considered to maximize variant discovery. Although most efforts are focused on detecting point mutations within coding sequence, there is an opportunity to explore the landscape of intermediate and larger genetic variation, especially because such variation is more likely to be gene-disruptive. It is critical to include this type of variation in future analyses to interpret the causes of disease. Re-examining exome data sets for larger and more complex variation may be particularly relevant when the causal variants for seemingly Mendelian diseases remain undiscovered.

METHODS

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

Accession codes. NCBI short-read archive: SRA039053.

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

ACKNOWLEDGMENTS

We thank T. Brown and S. Girirajan for helpful comments during manuscript preparation. This work was supported by Simons Foundation Autism Research Initiative award SFARI191889 (E.E.E.) and US National Institutes of Health grants HD065285 (E.E.E.), HHSN273200800010C (D.A.N.) and HL 102926 (D.A.N.). E.E.E. is funded by the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

E.K. designed and implemented the Splitread algorithm; E.K. and C.A. analyzed data; B.J.O., L.V., M.J.R. and D.A.N. generated sequencing data; M.Y.D. and K.M. carried out validation experiments and analyzed processed pseudogenes and E.K., C.A. and E.E.E. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

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

1. Church, D.M. *et al. Nat. Genet.* **42**, 813–814 (2010).
2. Sherry, S.T. *et al. Nucleic Acids Res.* **29**, 308–311 (2001).
3. Mills, R.E. *et al. Nature* **470**, 59–65 (2011).
4. Kidd, J.M. *et al. Cell* **143**, 837–847 (2010).
5. Ng, S.B. *et al. Nature* **461**, 272–276 (2009).
6. O'Roak, B.J. *et al. Nat. Genet.* **43**, 585–589 (2011).
7. Depristo, M.A. *et al. Nat. Genet.* **43**, 491–498 (2011).
8. Li, H. *et al. Bioinformatics* **25**, 2078–2079 (2009).
9. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. *Bioinformatics* **25**, 2865–2871 (2009).
10. Hach, F. *et al. Nat. Methods* **7**, 576–577 (2010).
11. Mills, R.E. *et al. Genome Res.* **21**, 830–839 (2011).
12. Nguyen, T.V. *et al. World J. Gastroenterol.* **12**, 6021–6025 (2006).
13. Renton, A.E. *et al. Neuron* **72**, 257–268 (2011).
14. Pearson, C.E., Nichol Edamura, K. & Cleary, J.D. *Natl. Rev.* **6**, 729–742 (2005).
15. Kidd, J.M. *et al. Nat. Methods* **7**, 365–371 (2010).

ONLINE METHODS

General approach. Similar to that of Pindel[16], our general approach considered mate pairs in which one end is anchored to the reference genome and the other end maps imprecisely owing to the presence of an underlying structural variant or indel breakpoint. We defined a paired-end split model for a short read that spans an insertion or deletion event as two subsequences of equal length (balanced split) or unequal length (unbalanced split) owing to noncontiguity with the reference sequence (**Fig. 1a**). Next, we developed a maximum parsimony-based general framework to efficiently identify all structural variants without restriction on size and search space. A summary of the algorithm follows with a detailed description in the **Supplementary Note**.

Mapping and breakpoint detection. All paired-end reads from the donor genome or exome were initially mapped to the reference genome or exome using mrsFAST[17]. mrsFAST is a seed-and-extend type algorithm that reports all map positions of a read to the reference genome given a set Hamming distance[18] (number of mismatches between two equal-length sequences without gaps). This preprocessing step rapidly identified breakpoints because any insertion or deletion created a frameshift, leading to a large Hamming distance. Reads that could not be mapped to the reference, by definition, were flagged as unmapped reads—if, for a read pair, one end was unmapped whereas its mate was mapped, we classified the pair as one-end anchored[19,20]. Hamming distance estimates compute linearly with respect to the size of the input sequences. We calculated the read depth of each base pair in our reference genome on the basis of all concordant and discordant map locations.

Split read definition. One end–anchored reads that span a simple breakpoint will map to the same location in the reference genome. We took advantage of this fact to remap the candidate one end–anchored reads from the preprocessing step back to the reference genome by decomposing these reads into two subsequences (**Fig. 1a**). If we split the unmapped read at the correct breakpoint, it maps to the reference genome around the breakpoint. In the case of a deletion, the distance between the split subsequences corresponds to the size of the deletion event (**Fig. 1a**). In the case of an insertion, the location is bracketed by the one end–anchored subsequences. The subsequences may overlap when the span size is less than the expected read length (such as in polynucleotide repeat expansions). We distinguished two types of split reads: (i) a balanced split, in which the unmapped read decomposed into two subsequences of equal length, and (ii) an unbalanced split, which partitioned into subsequences of unequal length. Regardless of the breakpoint location, for an unmapped mate of one end–anchored reads with a given length L, there always exists a subsequence with a length $\geq L/2$ (pigeonhole principle). We used this observation to detect structural variants. Given a one end–anchored pair, the unmapped mate is split into two equal-size subsequences paired together, which is defined as a split read. This approach reduced the search space by examining a small subset of the original one end–anchored read pairs enriched for those containing indel and structural variant breakpoints. Because Hamming distance estimates were used, our approach was sensitive to even 1-bp indels.

Clustering and set cover approximation. All possible mappings of the split reads were reported in which the insert size between the anchored read and the split unmapped read was within 3 s.d. of the initial distribution. For exome data, the insert size distribution is usually dictated by the library preparation and similar among commonly used methods. After the capture process, typically no size selection was carried out, thus the distribution was wider than a Gaussian distribution. Clusters were seeded on the basis of the presence of one or more balanced splits. Unbalanced split reads were assigned to a cluster if the unbalanced split supported the balanced split and if the mapped one end–anchored reads were concordant by position and orientation. Each split-read pair could be mapped to multiple clusters, indicating different events in the reference. Each cluster was associated with a set of split reads. We computed the minimum number of clusters such that all split reads were assigned to a unique cluster and the total level of support for each cluster was maximized. Each cluster was defined as a set of unique split reads and the cost of the set was defined as a function of the number of elements in the set. It is possible to use any type of cost function; in our method we used the number of mappings as the cost. This problem is equivalent to a weighted set cover problem[21] for which a simple greedy algorithm provides an $O(\log n)$ approximation[22], where n is the total number of clusters. The greedy algorithm was implemented iteratively: in each iteration it selected a set in which the cost per uncovered element was the minimum possible. After selecting the best set, all split reads belonging to the selected set were removed from the remaining sets. The costs were updated after the removal of the optimal set and iterated over the remaining sets. The algorithm terminated when all split reads were covered. Structural variants represented by the selected clusters were reported with their support value and the actual reads that map to their correct location in the reference genome. In this approach, the cost function can be defined in different ways. An alternative cost function can be defined as a combination of the split-read support and the read depth.

Splitread program. The corresponding program, Splitread, is implemented in C (available at http://splitread.sourceforge.net/) and requires as input paired-end mapping information generated by mrsFAST[17] from underlying raw sequence data (FASTQ format). The current version of Splitread was designed for the Illumina platform. Standard output includes the base pair–resolved location of the insertion or deletion, level of support (number of reads supporting each event) and the total Hamming distance of the read mappings. Final call sets can be filtered for the support and Hamming distance can be adjusted on the basis of exome or genome sequence coverage. Splitread may be used as a stand-alone program on a single CPU or can be run on a cluster with multiple nodes. Custom reference sequences can be generated for better performance or sensitivity.

Parameters for Pindel and GATK. We ran Pindel version 0.2.0 using insert size of 30 without BreakDancer results and with the maximum event size index set to 5 (8,092 bp) as recommended. GATK 1.0.5299 was used for indel calling using UnifiedGenotyper -glm DINDEL option.

Exome data sets. We analyzed two exome data sets in this study. First, we generated exome sequence data from 11 HapMap samples, NA12891, NA12892, NA19238, NA12878, NA15510, NA18507, NA18517, NA18555, NA18956, NA19129 and NA19240 (ref. 23),

doi:10.1038/nmeth.1810

most of which have been included as part of the 1000 Genomes Project (http://www.hapmap.org). In-solution exome capture was carried out using Roche NimbleGen EZ Exome SeqCap v2 (44 Mbp including 36 Mbp exon target). NA12891 and NA12892 were sequenced on Illumina GAIIx platform with 76-bp paired-end reads, and the remainder were sequenced using Illumina HiSequation (2000) platform (50-bp paired-end reads). The second collection comprised exome data from 20 trios (both parents and child) with a single child affected with autism spectrum disorder, primarily from the Simons Simplex Collection[24]. Exome capture was carried out using Roche NimbleGen EZ Exome SeqCap v1 probes and sequenced using the Illumina GAIIx platform primarily with paired-end 76-bp reads. Average coverage for this set was 196× with 90% of target exons covered with at least 8×.

16. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. *Bioinformatics* **25**, 2865–2871 (2009).
17. Hach, F. *et al. Nat. Methods* **7**, 576–577 (2010).
18. Hamming, R.W. *Bell Syst. Tech. J.* **29**, 147–160 (1950).
19. Kidd, J.M. *et al. Nat. Methods* **7**, 365–371 (2010).
20. Hajirasouliha, I. *et al. Bioinformatics* **26**, 1277–1283 (2010).
21. Karp, R.M. in *Complexity of Computer Computations* (J.W.T.R.E. Miller, ed.) 85–103 (Plenum, New York, 1972).
22. Chvatal, V. *Math. Oper. Res.* **4**, 233–235 (1979).
23. International HapMap Consortium.. *Nature* **437**, 1299–1320 (2005).
24. O'Roak, B.J. *et al. Nat. Genet.* **43**, 585–589 (2011).