Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly

Giuseppe Narzisi[1], Jason A. O'Rawe[2,3], Ivan Iossifov[1], Yoon-ha Lee[1], Zihua Wang[1], Yiyang Wu[2,3], Gholson J. Lyon[2,3], Michael Wigler[1], and Michael C. Schatz[1]

[1]Simons Center for Quantitative Biology, One Bungtown Road, Cold Spring Harbor Laboratory, NY, USA, 11724
[2]Stanley Institute for Cognitive Genomics, One Bungtown Road, Cold Spring Harbor Laboratory, NY, USA, 11724
[3]Stony Brook University, 100 Nicolls Rd, Stony Brook, NY, USA, 11794

Abstract

We present a new open-source algorithm, Scalpel, for sensitive and specific discovery of INDELs in exome-capture data. By combining the power of mapping and assembly, Scalpel searches the de Bruijn graph for haplotype-specific sequence paths (contigs) that span each exon. The algorithm reports a single path for homozygous exons, two paths for heterozygous exons, and multiple paths for more exotic variations. A detailed repeat composition analysis coupled with a self-tuning *k*-mer strategy allows Scalpel to outperform other state-of-the-art approaches for INDEL discovery. We extensively compared Scalpel with a battery of >10000 simulated and >1000 experimentally validated INDELs between 1 and 100bp against two recent algorithms for INDEL discovery: GATK HaplotypeCaller and SOAPindel. We report anomalies for these tools in their ability to detect INDELs, especially in regions containing near-perfect repeats which contribute to high false positive rates. In contrast, Scalpel demonstrates superior specificity while maintaining high sensitivity. We also present a large-scale application of Scalpel for detecting de novo and transmitted INDELs in 593 families with autistic children from the Simons Simplex Collection. Scalpel demonstrates enhanced power to detect long (≥20bp) transmitted events, and strengthens previous reports of enrichment for de novo likely gene-disrupting INDEL mutations in children with autism with many new candidate genes. The source code and documentation for the algorithm is available at http://scalpel.sourceforge.net.

Introduction

Enormous advances in next-generation sequencing technologies and computational variation analysis have made it feasible to study human genetics in unprecedented detail. The analysis of Single Nucleotide Variations (SNVs) has become a standard technique and high quality software is available for discovering SNVs with high confidence [1,2]. However, the same level of performance and reliability is not yet available for detecting of *INsertions* and *DELetions* in DNA sequences (INDELs) [3]. INDELs are the second most common sources of variation in human genomes and the most common structural variant [4]. Many INDELs map within human genes at functional loci, and have been shown to influence many human traits and diseases by

introducing frame-shifts or otherwise interrupting the protein coding sequence. When located within microsatellites (simple sequence repeats, SSRs, of 1 to 6bp motifs), INDELs typically alter the length of the repeat motif and have been linked to more than 40 neurological diseases in humans [5]. INDELs have been also shown to have an important genetic component in autism spectrum disorders [6]: *de novo* INDELs that are likely to severely disrupt the encoded protein are significantly more abundant in affected children than in their unaffected siblings.

Although INDELs play such an important role in human genetics, detecting them within next-generation sequencing data is still problematic for several reasons: (1) reads overlapping the INDEL sequence are more difficult to map [7] and reads supporting INDEL events may be aligned with multiple mismatches rather than with a gap; (2) irregularity in capture efficiency near the edges of the coding region and non-uniform read distribution across the target region increase the number of false positives called at these sites in whole exome data; (3) increased error rates makes detection of INDELs very difficult within microsatellites; and, as shown in this study, (4) the presence of localized, near identical repetitive sequences can confound the analysis, creating a high rate of false positives. For these and other reasons the size of INDELs detectable by available software tools has been relatively small, rarely more than a few dozen base pairs [8]. Consequently our understanding of INDEL origins and functional effects lags behind SNVs.

Two major paradigms are currently used for detecting INDELs in next-generation sequencing data. The first, and most common approach is to first map all the input fragments to the reference genome using any of the available read mappers (BWA, Bowtie, Novoalign, etc.). Mutations are then revealed by differences between the reference and the reads mapped at the particular location, although the available algorithms are not as effective for mapping across INDELs of more than a few bases. Advanced approaches exploit the information in paired-end reads to perform local realignments to detect longer mutations, e.g., GATK UnifiedGenotyper [1, 2] and Dindel [9]. In principle these methods can identify INDELs of size up to half the read length, although in practice their sensitivity is greatly reduced for variants of more than 20bp. Split-read methods, such as Pindel [10] and Splitread [11], can theoretically find deletions of any size, but their power is limited by the short read length of current next-generation sequencing technologies and suffer the same drawbacks of mapping approaches for long insertions.

The second paradigm consists of performing de novo whole-genome assembly of the input reads. Variations are then detected by computing differences between the assembled contigs and the reference genome [12,13]. In principle this paradigm has the potential to detect larger mutations since it is unbiased by the reference sequence, but in practice is less sensitive: whole-genome sequence assemblers have been designed with the goal of reconstructing the high level structure of a genome, while detecting structural variations requires a fine-grained and localized analysis of the sequence composition to report homozygous and heterozygous mutations.

Recently three approaches have been developed that use de novo assembly specifically for variation discovery: GATK HaplotypeCaller, SOAPindel [14] and Cortex [15]. According to the

GATK documentation (http://www.broadinstitute.org/gatk/), the HaplotypeCaller calls SNPs and INDELs simultaneously via local de-novo assembly of the haplotypes. Although this is the recommended tool using the GATK software package, the details have not been published and remain unknown. SOAPindel is the variation caller from the Short Oligonucleotide Analysis Package (SOAP). Our new method shares some similarity with this algorithm, although it differs in several important ways to enhance sensitivity and specificity of the results (see Methods for details). Finally Cortex is a de-novo sequence assembler that uses colored de Bruijn graphs for detecting and genotyping simple and complex genetic variants in an individual or population. However Cortex was tailored for whole-genome sequencing data and not demonstrated to account for the wide coverage fluctuations in exome-capture data. Thus, it is not used for comparison in this paper.

We present a novel DNA sequence micro-assembly pipeline, Scalpel, for detection of SNPs, insertion, and deletions within exome-capture data (**Fig. 1**). By combining the power of mapping and assembly, Scalpel searches for haplotype-specific sequence paths (contigs) that span each exon. The assembler reports a single path for homozygous exons, two paths for heterozygous exons, and multiple paths for more exotic variations. For example, if the sample has an insertion in just one of the two haplotypes, the assembler would discover the INDEL and also the unmodified reference sequence. After the sequences are assembled, they are aligned using a sensitive gapped sequence aligner to the reference exon to identify variants. Thanks to this approach, Scalpel is able to accurately discover and validate larger and more complex mutations with increased power compared to standard mapping methods. Moreover, the algorithm includes an on-the-fly analysis of the repeat composition of each exon, coupled with a self-tuning *k*-mer strategy, which allows Scalpel to outperform current state-of-the-art approaches for INDEL discovery in exons with complex repeat structure.

We extensively compared Scalpel with >10,000 simulated and >1,000 experimentally validated INDELs between 1 and 100bp against GATK HaplotypeCaller and SOAPindel. We report anomalies for these tools in the ability to detect INDELs, especially in regions containing near-perfect repeats, which leads to a high false positive rate. In contrast, Scalpel demonstrates superior specificity while maintaining high sensitivity without bias towards insertions or deletions. We also present a large-scale application of Scalpel to analyze de novo and transmitted INDELs in 593 (2372 individuals) families with autistic children from the Simons Simplex Collection [16]. We demonstrate more power to detect long transmitted events, and confirm a strong enrichment for de novo gene disrupting INDEL mutations in children with autism, and we identify many new candidates.

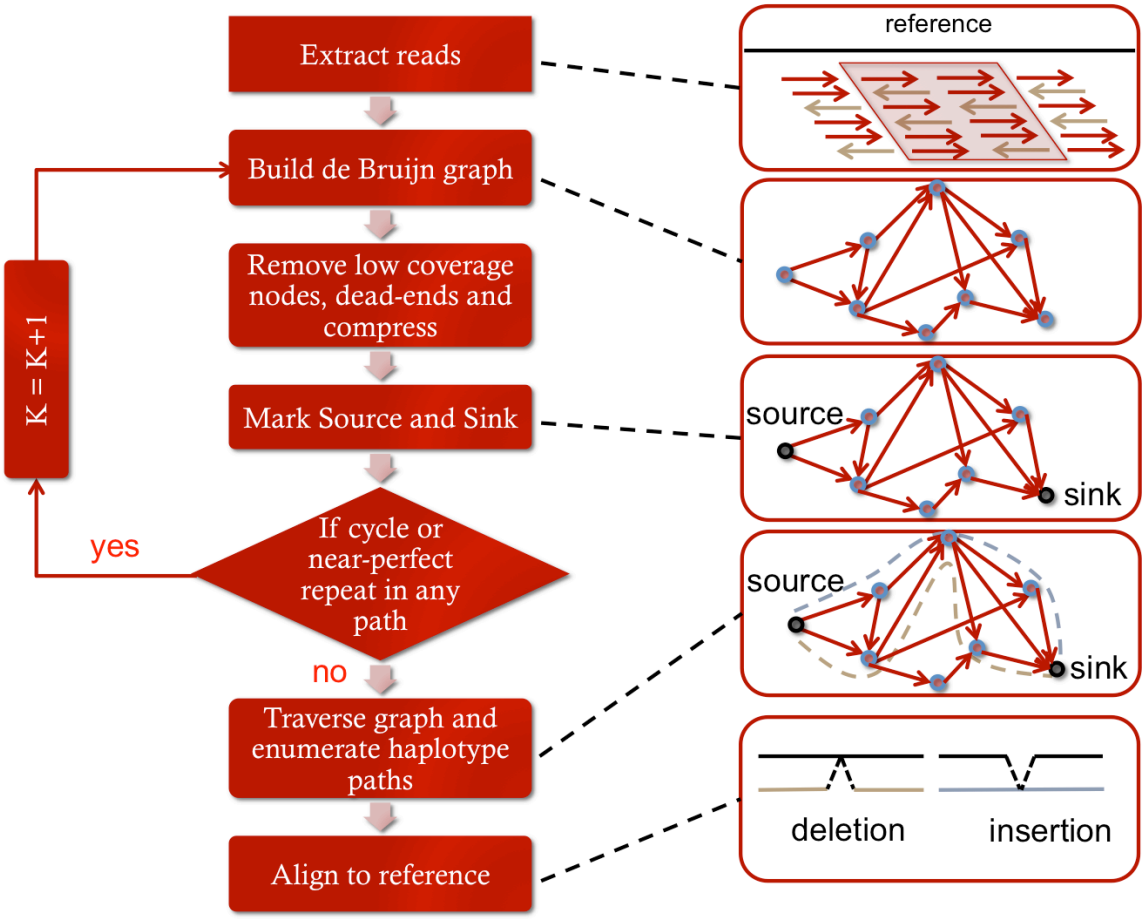

Figure 1. Overview of the Scalpel algorithm workflow.

Results

Experimental validation of variants in one single exome

It is now established that standard mapping methods have reduced power to detect large (≥20bp) INDELs [14, 15] and we confirm this result in this paper using simulated reads (**Supplementary Note 1**). However, the performance of variation discovery tools changes dramatically when applied to real data. In order to elucidate these anomalies, we performed a large-scale validation experiment involving ~1000 INDELs from one single exome. The individual has a severe case of Tourette Syndrome and obsessive compulsive disorder (sample id: K8101) and was sequenced to ≥20 reads per base pair over 80% of the exome target using the Agilent 44MB SureSelect capture protocol and Illumina HiSeq2000 paired-end reads, averaging 90bp after trimming. INDELs were called using three different pipelines according to their best practices: Scalpel, SOAPindel and GATK HaplotypeCaller (see online Methods). Interestingly, there is only ~37% concordance between all the pipelines, and each method reports a variable number of INDELs unique to that pipeline (**Fig. 2a**). Note that such low concordance is in close agreement with the recent analysis reported by O'Rawe *et al.* [3], and is much lower than the concordance for SNVs (~60%).

From this analysis alone, it is hard to judge the quality of INDELs unique to each pipeline, as these could either represent superior sensitivity or poor specificity. Interestingly, the size distribution of all INDELs called by each pipeline (**Fig. 2b**) has a clear bias towards deletions for the HaplotypeCaller and towards insertion for SOAPindel. Scalpel instead shows a well-balanced distribution between insertions and deletions, in agreement with other studies of human INDEL mutations [8].

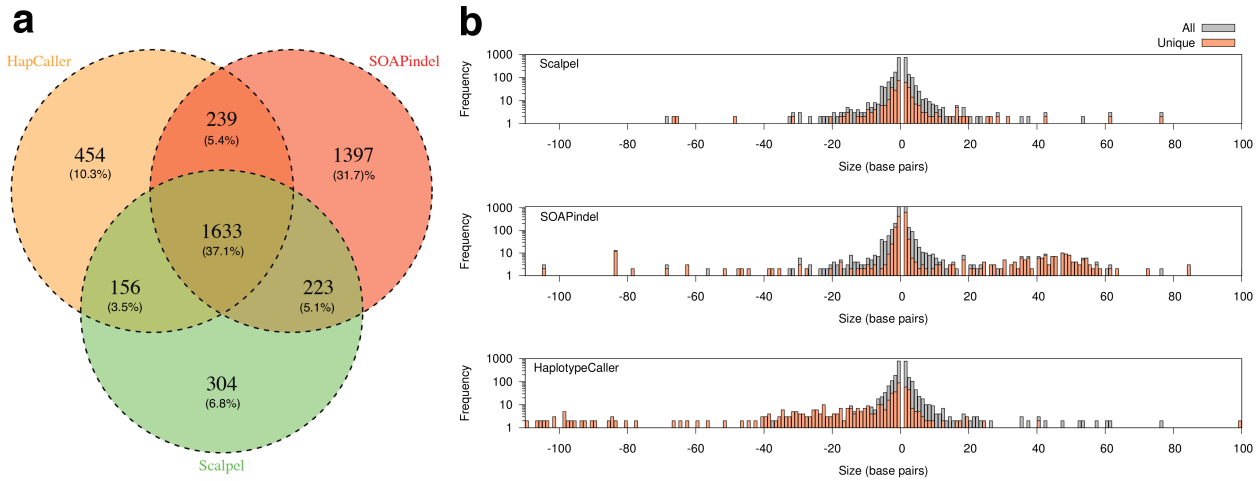

Figure 2. Concordance of INDELs between pipelines. (a) Venn Diagram showing the percentage of INDELs shared between the three pipelines. (b) Size distribution for INDELs called by each pipeline. The whole set of INDELs detected by the pipeline are colored in grey ("All"), while INDELs only called by the pipeline and not by the others are colored in orange ("Unique").

We further investigated the performance of the algorithms by a focused resequencing of a representative sample of the candidate INDELs. Specifically we performed deep re-sequencing of selected INDELs from all the tools using the more recent 250bp Illumina MiSeq sequencing protocol (see online Methods). Based on the data depicted in **Figure 2a**, we selected a total of 1000 INDELs according to the following five categories:

1. 200 random INDELs from the intersection of all pipelines.
2. 200 random INDELs only found by HaplotypeCaller.
3. 200 random INDELs only found by SOAPindel
4. 200 random INDELs only found by Scalpel.
5. 200 random INDELs of size ≥ 30bp from the union of all three algorithms.

Figure 3 shows the validation results for each INDEL category. Due to possibly ambiguous representation of an INDEL we "left-normalize" the coordinates of the reported INDELs using the approach of O'Rawe *et al.* [3]. However, some ambiguity can still remain, especially inside microsatellites, so we computed validation rates using two different approaches for comparing INDELs. (1) *Position-based:* an INDEL is considered valid if there is a mutation with the exact same starting position in the validation data (**Fig. 3a**). (2) *Exact-match*: an INDEL is considered valid if there is a mutation in the validation data that not only starts at the same coordinate but also has the same sequence composition (**Fig. 3b**).

As reported in prior studies, INDELs that are detected by all pipelines have a high validation rate and their sizes follow a lognormal distribution, and thus dominated by the smallest events

(**Supplementary Fig. 1**). However, the rate of INDELs passing validation varies dramatically for each tool. Respectively, only 22% and 55% of the HaplotypeCaller and SOAPindel INDELs could be validated even when the less strict position-based approach was used, whereas 77% of Scalpel's specific INDELs are true positive. Even worse is the outcome for the long INDELs: overall less than 10% passed validation, with SOAPindel and HaplotypeCaller calling the majority of these as erroneous INDELs (**Table 1**). The high false-positive rate for long deletions is also highlighted in **Figure 3c** where the validation rate by INDEL size for each variant caller is reported. HaplotypeCaller and SOAPindel show bias towards erroneous long deletions and insertions respectively.

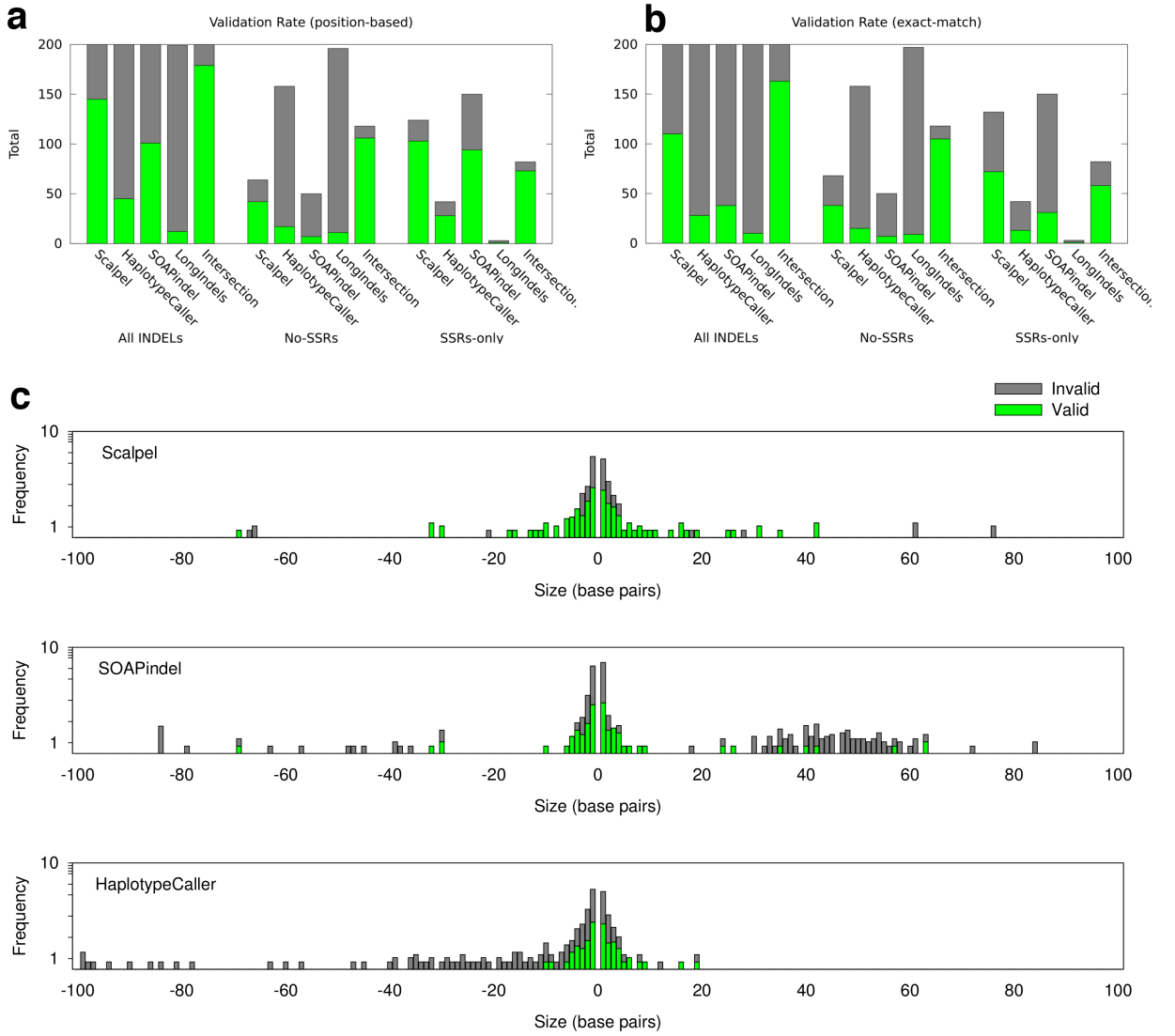

Figure 3. Results of MiSeq validation. (a) Validation rate for different INDEL categories using position-based match. (b) Validation rate for different INDEL categories using exact-match. Results are reported separately for each tool ("Scalpel", "HaplotypeCaller", and "SOAPindel"), for all INDELs of size ≥ 30bp from the union of the mutations detected by all three pipelines ("LongIndels"), and for INDELs in the intersection ("Intersection"). Validation results are further organized into three groups: validation for all INDELs ("All INDELs"), validation only for INDELs within microsatellites ("SSRs-only"), and validation for INDELs that are not within microsatellites ("No-SSRs"). (c) Stacked histogram of validation rate by INDEL size for each variant caller. INDELs that passed validation are marked with green color ("Valid"), while INDELs that did not pass validation are marked with grey color ("Invalid").

Since detecting INDELs within microsatellites is particularly challenging, we further divide the results to report the relative validation rate for this class of mutations separately. SOAPindel shows an appreciably higher rate of false-positives when only INDELs within microsatellites are considered ("SSRs-only" in **Fig. 3a** and **3b**). When microsatellites are excluded ("no-SSRs" in **Fig. 3a** and **3b**), the performance of SOAPindel and HaplotypeCaller drops significantly, while the Scalpel validation rate is only slightly reduced. This result is in agreement with the size biases highlighted in **Figure 2b**: microsatellites mutations are typically short in size, and when removed from the whole set, the remaining longer INDELs belong to the set that generates the size bias. Also note that **Figure 3a** and **Figure 3b** illustrates the relative abundance of INDELs called within microsatellites for each tool, although HaplotypeCaller seems to filter against these. Finally, it is important to observe that, when switching from position-based to exact-match, INDELs within microsatellites show significant reduction in validation rate. This phenomenon is due to their high instability and higher error rates, and as a result it is not unusual to have more than one candidate mutation at a microsatellite locus.

Table 1. Validation rate for INDELs specific to each pipeline. PPV is the positive predictive value computed as #TP/(#TP+#FP), where #TP is the number of true-positive calls and #FP is the number of false-positive calls.

Tool	Valid (all)	Invalid (all)	PPV (%) (all)	Valid (≥30bp)	Invalid (≥30bp)	PPV (%) (≥30bp)
Scalpel	145	43	77.1	13	1	92.8
SOAPindel	101	99	50.5	8	129	5.8
HaplotypeCaller	45	155	22.5	7	62	11.3

We further investigated the low validation rate for the long INDELs category by inspecting the sequence composition of all false-positive long INDELs. Specifically, we selected all 129 SOAPindel long mutations that did not pass validation and reassembled them using Scalpel. The majority of these mutations (115) overlap repeat structures where the reference contains a perfect or near-perfect repeat structure (**Supplementary Fig. 2**), and the associated assembly graph contained a repeat induced cycle. In contrast, of the 62 false-positive long INDELs from HaplotypeCaller, only 16 overlap a repeat structure. The remaining false positive deletions appear to be due to an aggressive approach used by the algorithm when processing soft-clipped sequences. On careful inspection of these data, the soft-clipped reads in false positive INDELs for HaplotypeCaller have a different form from the validation data, are highly variable, and are conjectured to be mapping artifacts of reads from different genomic locations (**Supplementary Fig. 3**).

Detecting de novo and transmitted INDELs in the Simons Simplex Collection

The Simons Simplex Collection (SSC) is a permanent repository of genetic samples from 2,700 families operated by SFARI (**http://sfari.org**) in collaboration with 12 university-affiliated research clinics. Each simplex family has one child affected with autism spectrum disorder, and unaffected siblings. Each genetic sample also has an associated collection of phenotype measurements and assays. The results presented in this section are based on a subset of the SSC composed of 593 families (2372 individuals). Specifically this subset of the SSC collection corresponds to families that have been examined in three recent studies: 343 families from

Iossifov *et al.* [6] (CSHL), 200 families from Sanders *et al.* [17] (Yale), and 50 families from O'Roak *et al.* [18] (University of Washington). We selected only family units of four individuals (father, mother, proband, one unaffected sibling), referred to as "quads," for all analyses in this study.

Transmitted mutations

Using Scalpel we detected a total of 3.3 million INDELs in 593 families, corresponding to an average of ~1400 (=3388139/(4*593)) mutations per individual. Accounting for population frequencies of each INDEL, there were 27795 distinct transmitted INDELs across the exomes. **Figure 4a** shows the histogram of INDELs sizes by annotation category. Although we detected INDELs only within the exome-capture target regions, we find close agreement between the size distributions reported by us and the one reported by Montgomery *et al.* [8] on low coverage whole-genome data from 179 individuals from 3 different population groups (YRI, CEU, CHB/JPT). We also compared the set of INDELs detected by Scalpel with the GATK-UnifiedGenotyper based mapping pipeline used by Iossifov *et al.* [6], and observe superior power to detect longer insertions, which supports the results obtained on simulated data (**Supplementary Fig. 4**). This result is also in agreement with the observation that insertions are harder to detect than deletions when using short reads.

Despite targeting exons, INDELs are more abundant in introns than other genic locations in the collection [4,19] (**Supplementary Fig. 5**). Within the coding sequence (CDS), frame-preserving INDELs are more abundant than frame shifts (**Fig. 4b)**. Also, in agreement with MacArthur *et al.* [19], we detected a large number of so-called loss-of-function (LOF) variants in protein-coding genes. Since these are all transmitted events in healthy parents and transmitted equally to autistic and non-autistic individuals, they demonstrate a high level of variation in functional gene content between healthy humans. Mutations are found at lower frequency in the population when located in protein-coding sequences compared to intronic regions (**Fig. 4c**). Finally, for each annotation category, we observe an enrichment of deletions over insertions (**Supplementary Table 1**), with an overall 2:1 ratio across all the classes. Similar trends were reported in previous studies [8,20].

To estimate the positive predictive rate of Scalpel to discover transmitted mutations, we performed targeted re-sequencing of 31 long (≥29bp) transmitted INDELs. Excluding INDELs that failed to sequence (4), 21 passed validation (out of 24), which gives an 87% true positive rate. For three INDELs we could not judge the result of the validation because they were either too long (≥70bp) for being validated using 143bp reads or they were located in a very complex region that was not possible to confidently align.

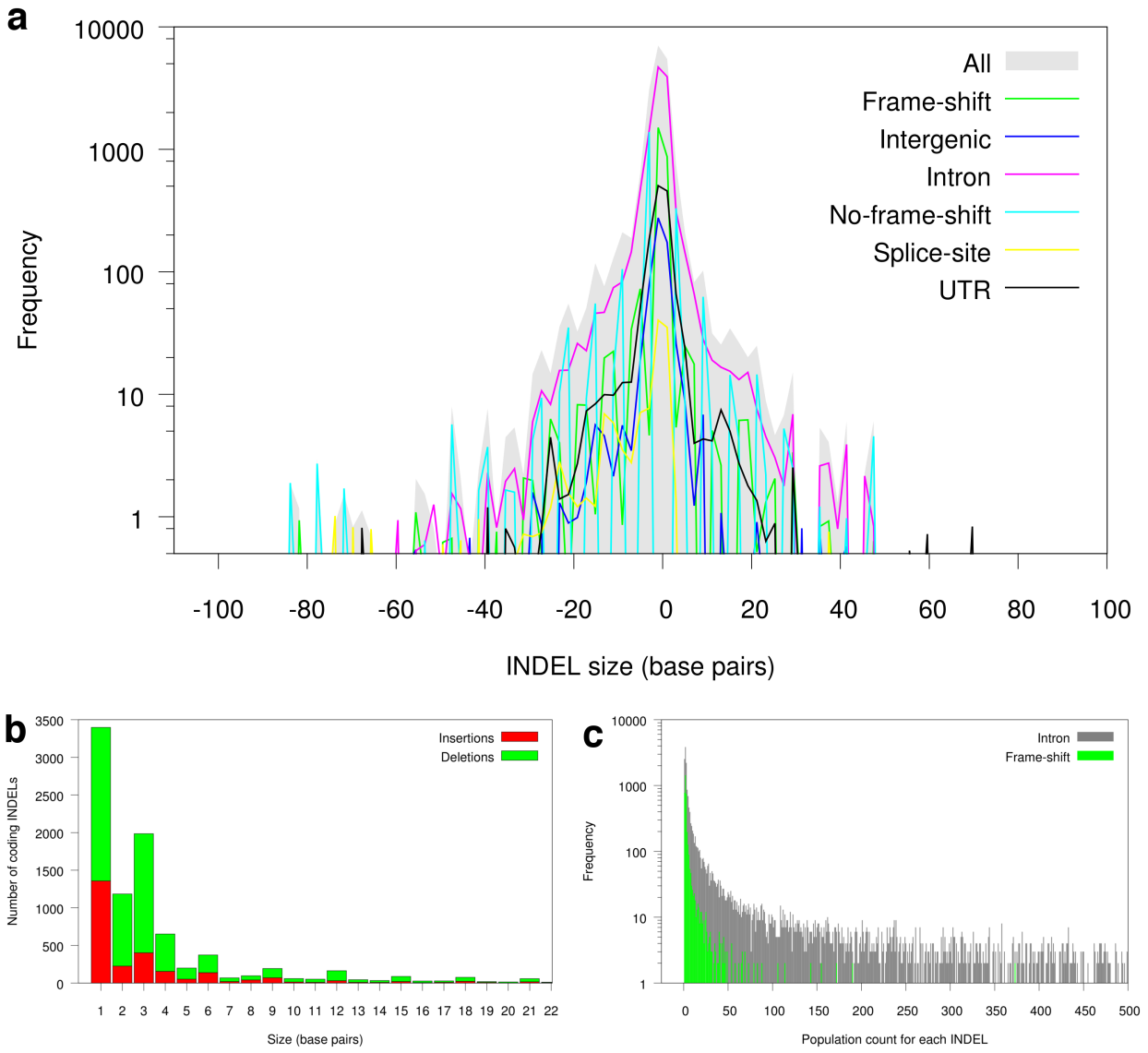

Figure 4. Transmitted mutations in 593 families. (a) Size distribution of insertions and deletions by annotation category. (b) Size distribution of INDELs within coding sequence (CDS). A spike is clearly visible for INDELs with size of multiple of three. (c) Histogram of INDELs frequency by annotation category showing how frame-shifts are typically found at low frequencies in the population.

De novo mutations

In our prior work, we had reported that de novo *likely gene disruptions (LGD)* mutations, including frame-shift INDELs, are significantly more abundant in affected children with autism than in unaffected siblings by nearly a 2:1 ratio [6]. Other smaller studies came to similar conclusions [17,18,21]. Here we reanalyzed these data with Scalpel to detect de novo INDELs with the goals of confirming such signal, predict additional candidates that could have been missed by the older pipeline, and extend the analysis to a larger number of families. In order to narrow down the list of candidate genes as best as possible, we excluded all mutations that are common in the population, and, used stringent coverage filters (see online Methods) to select a total of 97 high quality de novo INDELs. **Table 2** shows that, even after extending the population

size from 343 [6] to 593, the same 2:1 enrichment for LGD mutations is confirmed: 35 frame shifts in autistic children vs. 16 in siblings (p-value 0.01097).

Table 2. Summary of De Novo INDELs in 593 SSC Families in different contexts. "Aut" stands for autistic child and "Sib" for his/her sibling; "M" stands for males and "F" for females.

INDEL effect	Aut	Sib	Aut M	Aut F	Sib M	Sib F	Total
Frame shift	35	16	25	10	12	4	51
Intron	13	16	11	2	6	10	29
Intergenic	2	0	2	0	0	0	2
No frame shift	4	5	4	0	1	4	9
Splice-site	2	0	2	0	0	0	2
UTR	2	2	2	0	0	2	4
Total	58	39	46	12	19	20	97

In addition to finding novel mutations, in a few cases Scalpel was able to correct the size of the INDEL reported by other algorithms. Notably a 2 bp intronic deletion reported by the GATK-based pipeline was revised and confirmed to be a 33 bp deletion (**Supplementary Note 2** and **Supplementary Figs. 6-8**). Similarly two closely located candidate LGD deletions of 2 bp and 4 bp respectively, turned out to be a single longer non-deleterious 6 bp deletion (**Supplementary Note 2** and **Supplementary Figs. 9** and **10**). Finally, we performed targeted re-sequencing of 102 candidate INDELs; 84 were confirmed as de novo mutations, 11 were invalid and 7 failed to sequence, giving an 88% positive predictive rate for Scalpel.

Table 3 shows the list of de novo frame-shift LGD mutations in autistic children (the complete list of 97 de novo INDELs in all 593 families is reported in **Supplementary Table 2**). Sanders *et al.* [17] had previously only reported SNVs in their study, but our analysis of only these 200 families reports the same strong enrichment for LGD INDEL mutations in autistic children: 11 frame-shift LGDs INDEL mutations in autistic children compared to only 4 in their healthy siblings. This reanalysis reveals an important novel component of deleterious INDEL mutations and their associated genes that had been previously missed. Over the entire set of 593 families, we find a significantly high overlap between the LGD target genes and the set of 842 FMRP-associated genes [22], in agreement with the previously reported results [6]. Specifically 8 out of 35 LGDs in autistic children overlap with the 842 FMRP-associated genes.

Finally we compared the list of de novo variants detected by Scalpel with the list of de novo mutations discovered by the GATK-UnifiedGenotyper based mapping pipeline used by Iossifov *et al.* [6]. We found 21 mutations that were reported by GATK but not found by Scalpel. After manual inspection of these loci, we observed that 7 of these mutations where below the coverage thresholds used by Scalpel, while the rest of the variants were located in regions hard to assemble either because of the presence of a complex repeat or because of coverage dropping below the threshold required by Scalpel to generated a complete assembly.

Table 3. Likely Gene-Disrupting (LGD) frame-shift INDELs in children affected with autism. The "Family ID" column indicates the ID of the relevant family. The "Study" column shows the study in which the family was previously analyzed: CSHL, YALE or University of Washington (WASH). Under "Gender," M stands for males and F for females. The "Location" column reports the location of the variant in chr:position format. The "Variant"column shows detail for reconstructing the haplotype around the de novo variant relative to the reference genome as follows:"ins(seq)"indicates an insertion of the provided sequence "seq"; and "del(N)" denotes a deletion of N nucleotides. The "Gene" column reports the affected gene. The "Amino Acid Position" column shows the position of the first incorrectly encoded amino acid within the encoded protein/the length of the protein. When a mutation affects multiple isoforms of a transcript, the earliest proportionate coordinate is given. "FMRP target" indicates whether the corresponding gene's RNA was found to physically associate with FMRP [22].

Family ID	Study	Gender	Location	Variant	Gene	Amino Acid Position	FMRP Target
13548	CSHL	F	11:11314680	del(8)	*GALNTL4*	522/608	no
12858	CSHL	F	9:37015071	del(1)	*PAX5*	111/392	no
12952	CSHL	M	7:104748101	del(1)	*MLL5*	1066/1859	yes
13646	CSHL	M	9:35060456	del(5)	*VCP*	515/807	no
13548	CSHL	F	11:11314690	del(1)	*GALNTL4*	521/608	no
12673	CSHL	M	22:40661587	del(4)	*TNRC6B*	451/1834	yes
12939	CSHL	M	17:42399124	del(2)	*SLC25A39*	112/360	no
13018	CSHL	M	7:100201680	del(1)	*PCOLCE*	101/450	no
13176	CSHL	F	14:68272015	del(1)	*ZFYVE26*	397/2540	no
12950	CSHL	M	7:138968840	del(4)	*UBN2*	1063/1348	no
13096	CSHL	M	7:150164232	del(1)	*GIMAP8*	149/666	no
12653	CSHL	M	6:170593076	ins(A)	*DLL1*	431/724	no
13616	CSHL	M	4:47571001	ins(G)	*ATP10D*	1001/1427	no
13092	CSHL	M	19:49004781	ins(AGGTCAG)	*LMTK3*	307/1490	yes
13162	CSHL	M	6:72889392	ins(A)	*RIMS1*	196/1693	no
13439	CSHL	M	10:53458250	ins(A)	*CSTF2T*	354/617	no
13590	CSHL	M	15:80137554	ins(A)	*MTHFS*	147/147	no
13398	CSHL	M	1:151377904	ins(CGTCATCA)	*POGZ*	1194/1402	no
13552	CSHL	M	21:38877834	del(1)	*DYRK1A*	496/764	no
13168	CSHL	F	11:119214625	del(1)	*MFRP*	342/580	no
12323	CSHL	M	9:96439930	del(1)	*PHF2*	1088/1097	no
13471	CSHL	M	1:152286920	del(2)	*FLG*	147/4062	no
12705	CSHL	M	10:428609	ins(C)	*DIP2C*	657/1557	yes
12235	YALE	M	13:99100553	del(2)	*FARP1*	1040/1046	no
12099	YALE	M	21:38845117	del(2)	*DYRK1A*	48/764	no
12507	YALE	M	16:89350772	del(4)	*ANKRD11*	725/2664	yes
13618	YALE	F	8:37702146	del(2)	*BRF2*	374/420	no
12383	YALE	M	3:52454425	del(1)	*PHF7*	129/382	no
11808	YALE	F	6:31525440	ins(TG)	*NFKBIL1*	124/367	no
13739	YALE	F	X:153135039	del(2)	*L1CAM*	401/1258	no
13000	YALE	M	17:8424205	del(1)	*MYH10*	755/2008	yes
11712	YALE	M	1:53416507	del(2)	*SCP2*	70/524	no
11282	YALE	M	1:155317483	ins(CTTG)	*ASH1L*	2589/2965	yes

13618	YALE	F	15:93524061	del(4)	*CHD2*	965/1829	no
13447	WASH	F	6:157527665	del(4)	*ARID1B*	1784/2237	yes

Discussion

Assembly is the missing link towards high accuracy and increased power for INDEL mutation discovery for two reasons: (1) it allows the algorithm to break free from the expectations of the reference and (2) extends the power of the method to detect longer mutations. These features are crucial for the analysis of inherited and somatic mutations. Although these ideas have been explored recently in the literature, currently available tools for variant detection suffer from a high error rate. Scalpel is a powerful new method for detecting INDELs in NGS data that combines the power of assembly and mapping into a single unified framework. While featuring an enhanced power to detect longer mutations, Scalpel does not lose specificity thanks to a detailed repeat composition analysis combined with a self-tuning *k*-mer strategy. We demonstrate the superior specificity of Scalpel by comparing it with state-of-the-art INDEL callers on 1000 validated INDELs from one single exome. Such a large-scale re-sequencing experiment was fundamental to explain the sources of errors of current variant detection software, especially in regions containing near-perfect repeats. A large-scale application to detect de novo and transmitted INDELs in families with autistic children from the Simons Simplex Collection reveals the enhanced power of Scalpel to detect long (≥20bp) transmitted events and confirms a strong enrichment for de novo likely gene-disrupting (LGD) INDELs mutations in children with autism. Such results also hold for a larger collection of 1303 SSC families (not presented in this study). Although in this paper we show results only for exome-capture data, the Scalpel algorithm is agnostic to the sequencing protocol and can be used for whole-genome data as well. We envision that Scalpel will play an important role in the near future for the analysis of inherited and somatic mutation in human studies.

Acknowledgments

The project was supported in part by National Institutes of Health award (R01-HG006677) to M.C.S., the CSHL Cancer Center Support Grant (5P30CA045508), the Stanley Institute for Cognitive Genomics, and the Simons Foundation (SF51 and SF235988) to M.W. The DNA samples used in this work are included within SSC release 13. Approved researchers can obtain the SSC population dataset described in this study by applying at https://base.sfari.org. We thank S. Eskipehlivan for the technical assistance with the MiSeq validation experiments. We thank H. Fang, M. Bekritsky, S. Neuburgerand, M. Ronemus, D. Levy, B. Yamron, and B. Mishra for helpful discussions and comments on the paper. We thank H. Fang and R. Aboukhalil for testing the software.

Authors contributions

G.N. developed the software and conducted the computational experiments. G.N. and M.C.S. designed and analyzed the experiments. Y.W. performed the MiSeq validation experiments. J.A.O. designed the primers and analyzed the MiSeq data. G.J.L. planned and supervised the

experimental design for INDEL validation. Z.W. deigned the primers and performed experiments for the validation of de novo and transmitted INDELs in the SSC. I.I., Y.L., and M.W. assisted with the analysis of the SSC. G.N. and M.C.S. wrote the manuscript with input from all authors. All of the authors have read and approved the final manuscript.

References

1. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-303 (2010).
2. DePristo, M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498 (2011).
3. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* **5**:28 (2013).
4. Mullaney, J.M., Mills, R.E., Pittard, W.S. & Devine, S.E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**:R, 131–6 (2010).
5. Pearson, C.E., Edamura, N.K. & Cleary, J.D. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Gen.* **6** (10): 729–742 (2005).
6. Iossifov, I. *et al*. De novo gene disruptions in children on the autism spectrum. *Neuron* **74**, 2 285-299 (2012).
7. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome. Res.* **18**, 1851–1858 (2008).
8. Montgomery, S.B. *et al*. The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res*. **23**, 749-761 (2013).
9. Albers, C.A. *et al.* Dindel: Accurate indel calls from short-read data. *Genome Res.* **21**, 961–973 (2011).
10. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
11. Karakoc, E., *et al*. Detection of structural variants and indels within exome data. *Nat Methods* **9(2)**, 176-8 (2011).
12. Li, Y. *et al*. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nature Biotechnology* **29**, 723–730 (2011).
13. Heng, L. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics* **28** (14): 1838-1844 (2012).
14. Li, S. *et al*. SOAPindel: Efficient identification of indels from short paired reads. *Genome Res.* **23**, 195-200 (2012).
15. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat, Genet.* **44**, 226–232 (2012).
16. Fischbach G.D. & Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors, *Neuron* **68**, 192–195 (2010)
17. Sanders, S.J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
18. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
19. MacArthur, D.G. & Chris, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans**.** *Hum. Mol. Genet*. **19** (R2), R125-R130 (2010).
20. Sjödin, P., Bataillon, T. & Schierup, M.H. Insertion and Deletion Processes in Recent Human History. *PLoS ONE* **5**(1): e8650 (2010).
21. Neale, B.M. *et al*. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).

22. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, **146**, 247–261 (2011).

Online Methods

The Scalpel pipeline

Scalpel is designed to perform localized micro-assembly of specific regions of interest in a genome with the goal of detecting insertions and deletions with high accuracy. It is based on the de Bruijn graph assembly paradigm where the reads are decomposed into overlapping *k*-mers, and directed edges are added between *k*-mers that are consecutive within any read [23]. **Figure 1** shows the high level structure of the pipeline. (1) The pipeline begins with a fast alignment of the reads to the reference genome using BWA [24]. Importantly, these alignments are not directly used to call variations, but only to localize the analysis by identifying all the reads that have similarity to a given locus. Reads are then extracted in the region of interest (e.g., exon) including: (i) *well-mapped* reads, (ii) *soft-clipped* reads, and (iii) reads that *fail to map*, but are anchored by their mate. The latter two classes correspond to locations where the mapper encountered trouble aligning the reads, especially because of the large INDELs present, so it's necessary to include them in the assembly. (2) Once localized, the algorithm computes an on-the-fly assembly of the reads in the current region using the de Bruijn graph paradigm, specifically, reads are decomposed into overlapping *k*-mers (starting with a default *k*=25) and the associated graph is constructed. (3) Using the reference sequence, one source node and one sink node are then selected according to the procedure described later in the "Graph traversal" section. (4) An on-the-fly analysis of the repeats in each region is used to automatically select the *k*-mer size to be used for the assembly, described in section "Repeat analysis". (5) The graph is then exhaustively examined to find end-to-end haplotypes paths that span the region. (6) After the sequences are assembled, they are aligned to the reference to detect candidate mutations using a sensitive gapped sequence aligner based on the Smith Waterman algorithm [25] targeted at the reference window. Finally, the above assembly process is applied using a sliding window approach over each target region. By default a window size of 400bp is used with a sliding factor of 100bp. The sliding window strategy is fundamental to handle the highly non-uniform read distribution across the target (see **Supplementary Fig. 11**). A window size of 400bp is large enough to assemble the majority of the exons into a single contig since ~95% of the human exon-targets are shorter than 400bp (see **Supplementary Fig. 12**), however each assembly task is small enough for using in-depth techniques to optimize the assembly.

Graph construction

Two critical components of the Scalpel algorithm are (i) construction of the de Bruijn graph and (ii) detection of haplotype paths spanning the targeted region. Reads aligning to the region are extracted and decomposed into overlapping *k*-mers. In order to model the double stranded nature of the DNA, a bidirected de Bruijn graph is constructed [26, 27]. The graph is then compressed by merging all non-branching chains of *k*-mers into a single node. Tips and low coverage nodes are removed according to input threshold parameters to remove obvious sequencing errors. Note that, differently from traditional de Bruijn graph assemblers, Scalpel does not use any threading strategy to resolve collapsed repeats. Threading allows resolution of

repeats whose lengths are between *k* and the read length. However we observed in both real and simulated data that, due to the localized graph construction, if a bubble were not covered end-to-end by the reads, threading would either disconnect the graph or introduce errors. Repeats are instead handled differently as explained in the next section.

Repeat Analysis

Due to the highly non-uniform read depth distribution across the targeted region and the presence of near-perfect repeats that can mislead the assembly (**Supplementary Note 3** and **Supplementary Fig 13**), Scalpel implements a detailed repeat composition analysis coupled with a self-tuning *k*-mer strategy. Specifically, when assembling a window, Scalpel inspects both the base pair composition of the corresponding reference sequence as well as the resulting de Bruijn graph for the presence of cycles in the graph or near-perfect repeats in the assembled sequences. If a repeat structure is detected, the graph is discarded and a larger *k*-mer is selected. This process continues until a maximum *k*-mer length is reached, which is a function of the read length. If no *k*-mer value can be chosen to avoid the presence of repeats, the region is skipped and the next available region from the sliding window scheme is analyzed. This conservative strategy reduces the number of false-positive calls in highly repetitive regions, and, according to our experiments, skips less than 2% of possible windows in the human exome.

The proposed self-tuning *k*-mer strategy is similar to the dynamic approach used by SOAPindel to reconnect a broken path in low coverage regions. However, SOAPindel searches for unused reads with gradually shorter *k*-mers until a path is formed or the lower bound on *k*-mer length has been reached. Scalpel instead starts from a small *k*-mer value (input parameter) first and then gradually increases it, such that the smallest possible *k*-mer value is used for each region. This strategy has the advantage of better handling of repetitive sequences, highly polymorphic regions, and sequencing errors: source and sink have higher chance to be selected (see section "Graph traversal") and a smaller *k*-mer reduces the chance of fragmented assembly in low coverage regions.

Graph traversal

Once a valid de Bruijn graph is constructed, Scalpel examines the graph to find end-to-end haplotype sequence-paths that span the target window. Because the coverage from exome capture data is highly non-uniform, a special selection algorithm is used to find the edges of each window where coverage is present. First, two nodes in the graph are labeled as *source* and *sink* according to the following procedure: the reference sequence of the target region is scanned left-to-right to detect the first sequence of *k* bases that exactly matches one of the *k*-mers from the nodes in the graph, this node will be marked as the source. In a similar fashion the sink node is detected scanning the reference sequence right-to-left. Since every region is first inspected for repeats, source and sink can be safely selected at this stage. After the source and sink nodes are identified, all possible source-to-sink paths are enumerated up to a max number (default 100,000) using a depth-first search (DFS) traversal of the graph, similarly to Sutta assembly algorithm [28]. Note that since the regions to assemble are very small, time and space computational complexities associated with large-scale whole-genome assembly are not relevant and an exact brute-force strategy can be efficiently applied. If there are no repeat

structures in the graph, all the candidate haplotype paths are enumerated and aligned to the portion of the reference sequence delimited by source and sink *k*-mers using the standard Smith-Waterman-Gotoh alignment algorithm with affine gap penalties. The list of candidate mutations is then generated. Under typical conditions, the assembler reports a single path for homozygous mutations and two paths for heterozygous mutations. For example, if the sample has an insertion in only one of the two haplotypes, the assembler would discover the INDEL and also the unmodified reference sequence. Note that a traditional sequence assembler would have selected only one of these two paths (usually with higher coverage) and discarded the other one. Scalpel instead examines both paths to distinguish, for example, between homozygous and heterozygous mutations. However, in practice, various factors in real data complicate the detection process and, sometimes, multiple paths are reported in the case of more exotic variations. For example, the Illumina sequencing platform is particularly error prone around microsatellites (e.g., homopolymer runs) and, as a consequence, multiple candidate alleles are elucidated by the data at these loci. Highly polymorphic regions are also prone to generate multiple haplotype paths and could be computationally demanding: if the distance between multiple nearby mutations is too large to infer phasing information, each of the associated bubbles in the graph will give rise to two different paths.

Exome capture data

Exome capture for the sample K8101 was carried out using the Agilent 44MB SureSelect protocol and then sequenced on Illumina HiSeq2000 with average read length of 100bp. More than 80% of the target region was covered with depth of 20 reads or more. All of the HiSeq data have been submitted to the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under project accession number SRX265476.

MiSeq validation

1000 INDELs were selected for MiSeq validation (see the Results section for selection criteria). PCR primers were designed using Primer 3 (http://primer3.sourceforge.net) to produce amplicons ranging in size from 200 to 350 bp, with INDELs of interest located approximately in the center of each amplicon. Primers were obtained from Sigma-Aldrich® in 96-well mixed plate format, 10 µmol/L dilution in Tris per Oligo. Upon arrival, all primers were tested for PCR efficiency using a HAPMAP DNA sample (Catalog ID NA12864, Coriell Institute for Medical Research, Camden, NJ, USA) and LongAmp® Taq DNA Polymerase (New England Biolabs, Beverly, MA, USA). PCR products were visually inspected for amplification efficiency using agarose gel electrophoresis. For the validation experiment, the same PCR protocol as above was performed using sample K8101-49685 genomic DNA as template. PCR product was verified on E-Gel® 96 gels (Invitrogen Corp., Carlsbad, CA, USA) and subsequently pooled for ExoSAP-IT® (Affymetrix Inc., Santa Clara, CA, USA) cleanup. The cleanup product was further purified using QIAquick PCR Purification Kit (QIAGEN Inc., Valencia, CA, USA) and quantified by Qubit® dsDNA BR Assay Kit (Invitrogen Corp.). Library construction for the MiSeq Personal Sequencer platform (Illumina Inc.) was performed based on the TruSeq DNA Sample Prep LS protocol (Illumina®), omitting the DNA fragmentation step. Finally, before being loaded onto the MiSeq machine, the quality and quantity of the sample was again verified using the Agilent DNA 1000 Kit on the Agilent Bioanalyzer and with quantitative PCR (Kapa Biosystems Inc., Woburn,

MA, USA). This protocol generated high quality 250 bp reads (paired end) with an average coverage of 47018X over the validated INDELs (see **Supplementary Fig. 14**). All of the MiSeq data have been submitted to the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under project accession number SRX386284.

Alignment

Sequencing reads from K8101 exome-capture data were aligned using BWA (v0.6.2-r126) with default parameters to the human reference HG19. Alignments were converted from SAM format to sorted, indexed BAM files with SAMTools (v0.1.18). The Picard tool (v1.91) was used to remove duplicate reads. These BAM files were used as input for all the INDEL callers used in this study. Reads coming from the re-sequencing experiments were also aligned using BWA. However, if the INDEL approaches half the size of the read length, even after target re-sequencing, mapping the reads containing the INDEL is problematic. The problem is emphasized if the INDEL is located towards the ends of the read (instead of in the middle). To avoid this problem we aligned sequencing reads containing long INDELs (≥30 bp) using Bowtie2 [29] instead of BWA. Bowtie2 offers an end-to-end alignment mode that searches for alignments involving all of the read characters, also called an "untrimmed" or "unclipped" alignment. Specifically, we used the following parameter settings: “--end-to-end --very-sensitive --score-min L,-0.6,-0.6 –rdg 8,1 –rfg 8,1 --mp 20,20”.

Variant Calling

INDELs for K8101 were called using Scalpel, GATK HaplotypeCaller and SOAPindel as follows:

Scalpel. Scalpel (v0.1.1 beta) was run on the indexed BAM using the following parameter setting: “--single --lowcov 1 --mincov 3 –outratio 0.1 --intarget”. INDELs showing high coverage unbalance were then removed (chi-square *k*-mer score > 20).

GATK. GATK software tools (v2.3-9) were used for improvement of alignments and genotype calling and refining with recommended parameters. BAM files were re-aligned with the GATK IndelRealigner, and base quality scores were re-calibrated by the GATK base quality recalibration tool. Genotypes were called by the GATK UnifiedGenotyper and HaplotypeCaller. GATK was used to filter high-quality INDELs by hard criteria: "QD < 2.0, ReadPosRankSum < -20.0 FS > 200.0”. We also tested the most recent release of the GATK toolkit (v2.7-4) and we observed the same bias towards long deletion reported in the Results section of the paper.

SOAPindel. SOAPindel (v 2.0.1) was run on the indexed BAM file using default parameters. According to SOAPindel documentation, putative INDELs are initially assumed to be located near the unmapped reads whose mates mapped to the reference genome. SOAPindel then executes a local assembly (kmer=25 by default) on the clusters of unmapped reads. The assembly results were aligned to reference in order to find the potential INDELs. To distinguish true and false positive INDELs, SOAPindel generates Phred quality scores, which take into consideration the depth of coverage, INDEL size, number of neighboring variants, distance to the edge of contig, and position of the second different base pair. Only those INDELs filtered by internal threshold are retained in the final INDEL call set.

Finally, for all pipelines we selected only INDELs located within the regions targeted by the exome capture protocol.

Analysis of De Novo INDELs related to Autism

After eliminating all candidate positions that are common in the population, and thus unlikely to be related to the disorder, we re-assembled each region associated with the candidates INDELs across the family members using a more sensitive parameter setting for Scalpel. Specifically we reduce the starting *k*-mer value to 10 and turned off the removal of low coverage nodes. This step was important to adjust for possible allele imbalance favoring the reference allele over the mutation in the parents, but was impractical to do initially for the whole collection: lowering the *k*-mer and keeping all the nodes in the graph significantly increase the computation complexity of the algorithm. Then we selected de novo INDELs with chi-square *k*-mer score ≤ 10.84.The chi-square *k*-mer score is computed using the standard formula for the chi-square test statistics (χ^2) but applied to the *k*-mer coverage of the reference and alternative alleles for the mutation according to the following formula:

$$\mathrm{X}^2 = \frac{(C_o^R - C_e^R)^2}{C_e^R} + \frac{(C_o^A - C_e^A)^2}{C_e^A}$$

where C_o^R and C_o^A are the observed *k*-mer coverage for the reference and alternative alleles, and C_e^R and C_e^A are the expected coverage such that $C_e^R = C_e^A = totCov/2$. Finally we enforced parents to have at least a *k*-mer coverage of 15 over the assembled region.

System requirements and software availability

Scalpel is written in Perl and C++. The source code is freely available as an open-source software project on the SourceForge website at https://sourceforge.net/projects/scalpel/. It usually takes 2-3 hours to process one exome-capture data set (80% of target at ≥ 20X) using 10 CPUs and requiring a minimum of 3GB of RAM.

References

23. Nagarajan, N, & Pop, M. Sequence assembly demystified. *Nature Reviews Genetics* **14**, 157-167 (2013).
24. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **26** (5), 589-595 (2010).
25. Smith, T.F. & Waterman, M.S. Identification of Common Molecular Subsequences. *Journal of Molecular Biology* **147**: 195–197 (1981).
26. Medvedev, P., Georgiou, K., Myers, G. & Brudno, M. Computability of models for sequence assembly. *Lecture Notes in Computer Science* **4645**, 289–301 (2007).
27. Jackson, B.G. & Aluru, S. Parallel Construction of Bidirected String Graphs for Genome Assembly. *Parallel Processing, 2008. ICPP '08. 37th International Conference on*, 346-353 (2008).
28. Narzisi, G. & Mishra, B. Scoring-and-Unfolding Trimmed Tree Assembler: Concepts, Constructs and Comparisons. *Bioinformatics*, **27** (2), 153-160 (2011).
29. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).