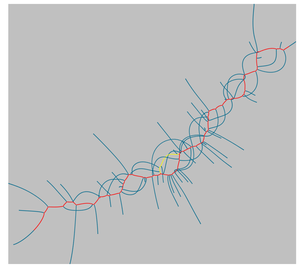
Scalpel

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

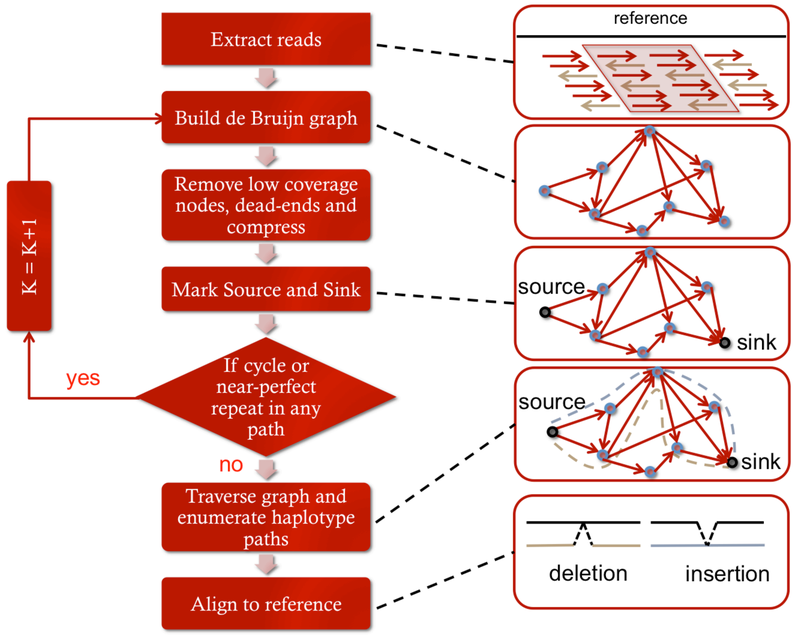
Giuseppe Narzisi, Jason A. O'Rawe, Ivan Iossifov, Yoon-ha Lee, Zihua Wang, Yiyang Wu,
Gholson J. Lyon, Michael Wigler, and Michael C. Schatz

Cold Spring Harbor Laboratory



Scalpel micro-assembly pipeline

Scalpel is a new open-source algorithm 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. 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.



Detecting variants in one single exome

We report anomalies for current INDEL detection tools: HaplotypeCaller and SOAPindel show a clear bias towards deletions and insertions respectively. Scalpel instead shows a well-balanced distribution in agreement with other studies of human INDEL mutations.

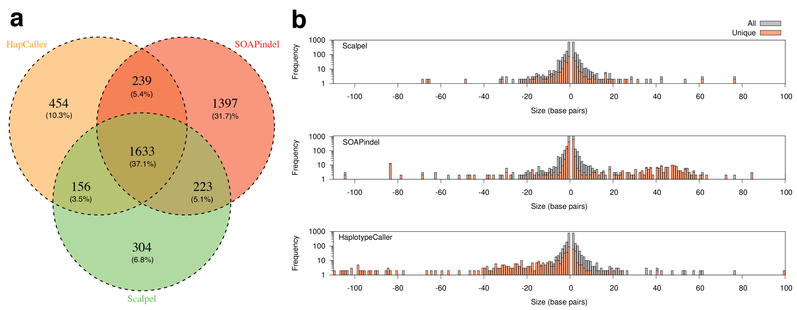


Figure 1: **Concordance of INDELs between pipelines.** (a) Venn Diagram showing the percentage of INDELs shared between the three pipelines. (b) Size distribution for INDELs called by each pipeline. The whole set of INDELs detected by the pipeline are colored in grey ("All"), while INDELs only called by the pipeline and not by the others are colored in orange ("Unique").

Targeted resequencing of 1000 representative candidate INDELs confirms the size bias. Analysis of the sequence composition correlates the high false-positive rates to genomic regions containing near-perfect repeats. The validation results also reveal the challenges to detect INDELs within microsatellites. This phenomenon is due to the high instability and higher error rates at microsatellite loci, where it is not unusual to have more than one candidate mutation.

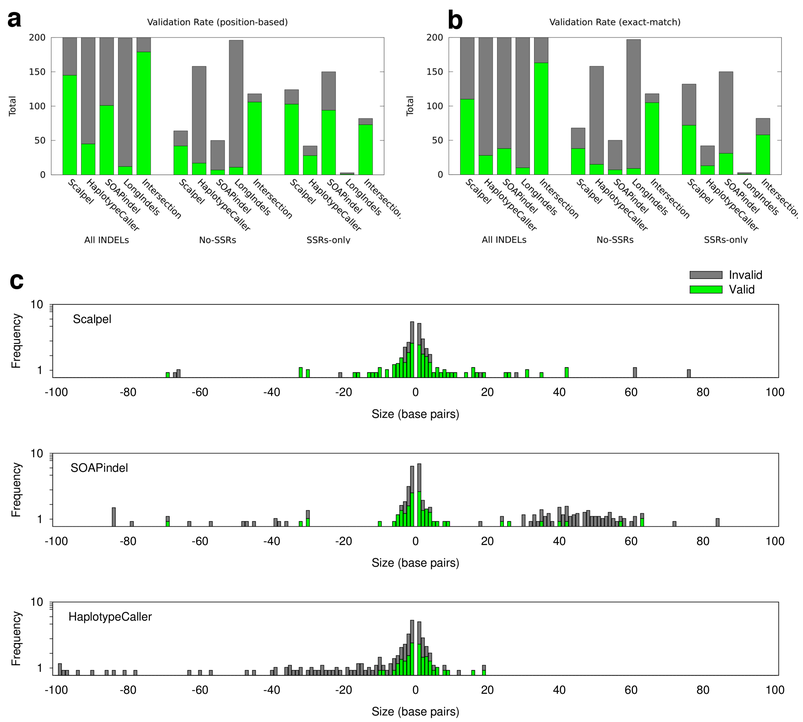


Figure 2: **Results of MiSeq validation.** (a) Validation rate for different INDEL categories using position-based match. (b) Validation rate for different INDEL categories using exact-match. Results are reported separately for each tool ("Scalpel", "HaplotypeCaller", and "SOAPindel"), for all INDELs 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").

De novo and transmitted INDELs in 593 families

Simons Simplex Collections (SSC): 593 families (2372 individuals), two parents each with exactly one affected child and one unaffected child.

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.

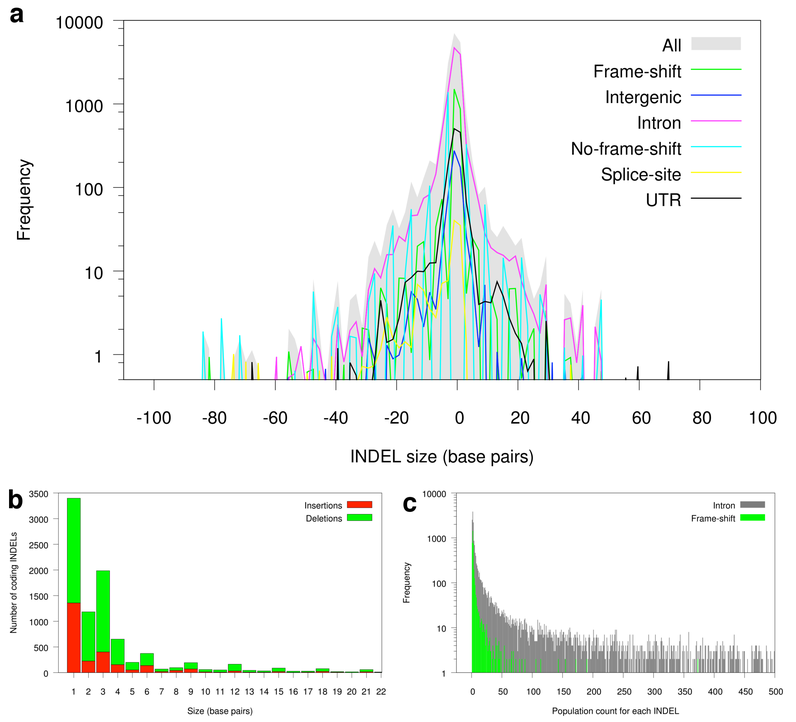


Figure 3: **Transmitted mutations in 593 families.** (a) Size distribution of insertions and deletions by annotation category. (b) Size distribution of INDELs within coding sequence (CDS). A spike is clearly visible for INDELs with size of multiple of three. (c) Histogram of INDELs frequency by annotation category showing how frame-shifts are typically found at low frequencies in the population.

De novo indels that are likely to severely disrupt the encoded protein - by causing frame-shifts, destroying splice sites, or introducing nonsense codons - are significantly more abundant in affected children than in unaffected siblings.

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

Availability: http://scalpel.sourceforge.net

Reference: Narzisi G., O'Rawe J.A., Iossifov I., Lee Y., Wang Z., Wu Y., Lyon G.J., Wigler M., Schatz M.C. Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly. **CSHL bioRxiv** (DOI: 10.1101/001370)