PopDel identifies medium-size deletions jointly in tens of thousands of genomes

Authors

Sebastian Roskosch[1,2], Hákon Jónsson[3], Eythór Björnsson[3,4,5], Doruk Beyter[3], Hannes P. Eggertsson[3], Patrick Sulem[3], Kári Stefánsson[3,4], Bjarni V. Halldórsson[3,6], Birte Kehr[1,2]

Affiliations

[1] Berlin Institute of Health (BIH), 10178 Berlin, Germany

[2] Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health

[3] deCODE genetics/Amgen Inc.

[4] Faculty of Medicine, School of Heath Sciences, University of Iceland

[5] Department of Internal Medicine, Landspítali – The National University Hospital of Iceland, Reykjavík, Iceland

[6] School of Science and Engineering, Reykjavik University

Abstract

Thousands of genomic structural variants segregate in the human population and can impact phenotypic traits and diseases. Their identification in whole-genome sequence data of large cohorts is a major computational challenge. Here we present PopDel, which identifies and genotypes deletions of about 500 to at least 10,000 bp in length in many genomes jointly. PopDel scales to tens of thousands of genomes as demonstrated by our evaluation on data of up to 49,962 genomes. Compared to previous tools, PopDel reduces the computational time needed to analyze 150 genomes from weeks to days. The deletions detected by PopDel in a single sample show a large overlap with high-confidence reference call sets. On data of up to 6,794 trios, inheritance patterns suggest a low false positive rate at a high recall. PopDel reliably reports common, rare and *de novo* deletions and the deletions reflect reported population structure. Therefore, PopDel enables routine scans for deletions in large-scale sequencing studies.

Introduction

Comprehensive and reliable collections of genetic variation are a foundation for research on human diversity and disease[1]. When the collections are available for a population or disease cohort, they facilitate a wide range of studies investigating mutation rates[2–4], mutational mechanisms[5–7], functional consequences of variants[8–10], ancestry relationships[11], disease risks[12], or treatment options[13]. Due to increased throughput and decreased cost, whole-genome sequencing (WGS) is now performed on cohorts of thousands of individuals. This includes sequencing at the population level in Iceland[14], the United Kingdom[15], or Crete[16] as well as sequencing of large cohorts for specific diseases, such as autism[17] or asthma[18], and in the general health research context in projects like GnomAD[19,20] or TopMed[21]. To create collections of genetic variation, the data from these large numbers of individuals needs to be integrated. The most direct way of achieving this is done in *joint* variant calling approaches, which analyze the data from many individuals together and infer the variants with genotypes directly from the input data.

For single nucleotide variants (SNVs) and small insertions/deletions (indels), joint calling has become the state of the art with tools that scale to tens of thousands of individuals[22,23]. For structural variants (SVs), the analysis of increasingly large numbers of individuals remains a major bioinformatic challenge[24]. Jointly detecting SVs in up to hundreds of individuals is a great achievement of previous projects and tools[25,26]. For larger cohorts, catalogues of SVs are generally created by first analyzing the data of each individual separately or in small subsets of individuals, subsequently merging the resulting call sets and, finally, determining genotypes for all individuals on the merged call set[27,28]. In this process, the aligned read data is typically accessed at least twice, for detecting and genotyping SVs, requiring substantial computational resources. In addition, the merging of SV call sets from different individuals is often arbitrary when the same SV is detected with shifted positions in several individuals [29,30]. Furthermore, variants that are only weakly supported by the data may not be discovered with this approach. A joint SV detection approach simplifies the calling process, is computationally more efficient if accessing the large amounts of input data only once, eliminates the need for an error-prone variant merging step, and may reveal weakly supported variants if carried by several individuals as the support accumulates across individuals.

To overcome current limitations of SV callers, we introduce a joint calling approach for deletions of a few hundred up to tens of thousands of base pairs in length. The approach is specifically designed for large cohorts and is, to our knowledge, the first joint approach for SV discovery that scales to tens of thousands of individuals. Nevertheless, it can also be applied to a single individual or small numbers of individuals, where it achieves comparable accuracy to widely-used deletion callers.

Results

Computational approach for joint deletion calling

Deletions can manifest themselves in the reference alignment of short-read sequences as local drops in read depth, changes in the distance between the alignment of two reads in a pair, and split-aligned reads[31]. The main focus of our joint deletion calling approach is on local changes in the read pair alignment distance compared to the genome-wide distribution of read pair alignment distances. This focus has the advantage that we can discover smaller deletions than read depth approaches while being more computationally efficient than approaches that consider split-aligned reads[32].

To achieve scalability to very large numbers of individuals, our approach implemented in the PopDel program consists of two steps: A profiling step, which reduces the aligned input sequencing read data per individual into a small read pair profile, followed by a joint calling step, which takes as input the read pair profiles and outputs deletion calls with genotypes across all individuals (Figure 1, Methods). This two-step design is reminiscent of joint calling of small variants in GATK[22] and CNV calling approaches that are based on read depth profiles[33,34]. Our read pair profiles contain an overall distribution of read pair distances as well as alignment start positions and distances of all read pairs that match certain quality criteria (**Supplementary Note**). The joint calling step processes these profiles of all individuals together in small genomic windows (default 30 bp) to discover and genotype deletions. For all windows, likelihood ratio tests are performed to test if deletions overlap the window in any of the jointly analyzed individuals. In the likelihood computation we use genotype weights to ensure that rare deletions can be found by boosting the signal in carriers and down-weighting the contribution of non-carriers dependent on the allele frequency. Finally, adjacent windows that support the same deletion are aggregated and output together with genotype likelihoods of all individuals.

The values of most parameters of PopDel are calculated from the input data. The input parameters for each likelihood ratio test are iteratively estimated (Figure 1, Methods): the deletion length, allele frequency, genotype weights and genotype likelihoods for all individuals for the three genotypes (non-carrier, heterozygous carrier and homozygous carrier). The minimum length of deletions that can be identified with our likelihood ratio test derives from the standard deviation of the insert sizes (**Supplementary Note**).

Assessment of scalability on simulated data

We simulated sequencing data of 1,000 diploid individuals each carrying 2,000 deletions with uniformly distributed allele frequencies, uniformly distributed lengths between 100 and 10,000 bp and uniformly distributed positions on chr21 (Methods). On these data, we compared the running time, memory consumption, recall and precision of PopDel to that of four popular SV callers that can identify SVs jointly in multiple samples (Delly[35], Lumpy[36] followed by SVTyper[37] for genotyping (LUMPY+SVTyper), Manta[38] and GRIDSS[39]). We note that PopDel only reports deletions while other callers look for other types of SVs, too.

With a running time of 397 minutes and a peak memory of 1.5 GB for profiling and joint calling on the simulated chr21 data of 1,000 individuals, PopDel is the fastest tool and among the tools that require the least memory (Figure 2A,B). It is also the only tool that can jointly discover SVs in the data of all 1,000 individuals. All other tools complete the calling on at most 500 individuals and fail at 300 or 600 individuals due to too many open files. For these tools, the user has to resort to single sample calling with subsequent merging and genotyping in order to analyze more individuals. The precision and recall of most tools, including PopDel, is high reflecting that the simulated data is easy to analyze (Figure 2C,D). Only GRIDSS's performance in precision and Manta's performance in recall drop significantly with increasing numbers of individuals indicating that these tools were primarily designed to analyze a single sample at a time.

Running times on public benchmarking data

Next, we assessed the running time of PopDel compared to Delly and Lumpy+SVTyper on short read WGS data for the well-studied genome of NA12878 (accession ERR194147) and the 150 unrelated genomes in the Polaris HiSeq X Diversity Cohort (accession PRJEB20654). With a total running time of approximately 29 minutes on a single core for profile creation and deletion calling of the NA12878 genome, PopDel is 7 times faster than Delly and 16 times faster than Lumpy+SVTyper (Table 1). On the data from the Polaris Diversity Cohort, PopDel completes deletion calling within less than three days of CPU time. As we were not able to perform joint SV detection with Delly and Lumpy with available computational resources (**Supplementary Note**), we created deletion call sets by running these tools on each sample individually with subsequent variant merging and sample-wise genotyping. Structural variants other than deletions were removed before merging. The total

CPU time needed by Delly for this task was almost three weeks (20 days), that of Lumpy+SVTyper more than three months (103 days, Table 1).

Comparison to reference deletion sets from the Genome in a Bottle (GiaB) consortium

The data of the NA12878 genome allowed us to compare the calls of the three tools to reference sets of deletion calls prepared by the GiaB Consortium[41]: the short read based reference set as well as a set of deletions called from PacBio long read data for this genome. In all our analyses, two deletions were considered the same if they had a reciprocal overlap of 50% or more. Deletion calls within centromeric regions were excluded in all analyses (Methods) as calls within these regions are generally less reliable[42].

On the NA12878 data, PopDel is competitive with Delly and Lumpy+SVTyper (Figure 3, **Supplementary Note**). All three tools succeed to identify the majority of deletions in the short read reference set (716/779, 91.9%) with PopDel identifying marginally more deletions (742, 95.3%) than Lumpy+SVTyper (732, 94%) and Delly (730, 93.7%). The fraction of PacBio deletions identified by all three tools is much lower (820/3,831, 21.4%). This is expected as the long PacBio reads reveal variants involving repeats that are invisible or hard to detect in short read data. Again, PopDel identifies a similar number of PacBio deletions (888, 23.2%) as Lumpy+SVTyper (892, 23.3%) and Delly (906, 23.6%). We note that PopDel reports fewer deletions that are not included in the two reference call sets than Delly and Lumpy+SVTyper, which can either be true or false positives.

Analysis of population structure based on deletions in the Polaris Diversity Cohort

In the Polaris Diversity Cohort, PopDel identifies an average of 1674 heterozygous and 205 homozygous deletions per individual (Methods). The cohort consists of three continental groups: Africans, East Asians and Europeans. As expected, Africans carry significantly (p-value $< 2.2 \cdot 10^{-16}$, two-sided t-test) more deletions than Europeans and East Asians (Figure 4A). Principal component analysis of PopDel's deletion calls (Methods) shows a clear separation between the three continental groups (Figure 4B) mirroring the well-known clustering resulting from small variants[25,43]. In particular, the first principal component separates the African samples from the other continental groups, while the second principal component additionally pulls apart the European and East Asian samples. These findings indicate that the deletions detected and genotyped by PopDel well reflect the biological differences between the continental groups. Similar results were obtained for the Delly and Lumpy+SVTyper deletions (**Supplementary Note**).

Performance evaluation using data of 49 Polaris trios

By combining the Polaris HiSeq X Kids Cohort (accession PRJEB25009) with the Polaris Diversity Cohort we obtain a set of 49 trios. In these trios, we evaluate the Mendelian inheritance error rate and transmission rate of reported deletions and their genotypes (Methods). The Mendelian inheritance error rate effectively assesses the genotyping of common variants. The transmission rate is also meaningful for rare variants measuring how often a deletion allele is inherited from a heterozygous parent (**Supplementary Note**). We calculated the transmission rate only for those deletions found in a single trio and where one parent is heterozygous and the other carries the reference allele on both haplotypes (Methods).

We determined filtering criteria for all three tools to reach a Mendelian inheritance error rate below 0.3%, which has been suggested as an acceptable error rate[42] (Figure 4C, Methods). With 1,712 deletions per trio that are consistent with Mendelian inheritance, the PopDel call set includes 15% more consistent deletions than Delly (1,485 consistent deletions per trio) and 40% more than Lumpy+SVTyper (1,222 consistent deletions per trio). The transmission rate of PopDel at 50.4% is very close to the expected 50%. The transmission rate of Delly is 49.2% and that of Lumpy is 47.5%. With 3363 deletions, the number of deletions reported by PopDel that we could use to calculate this transmission rate is 10% larger than for Delly (3064 deletions) and 7% larger than for Lumpy+SVTyper (3157 deletions). These results suggest that the joint approach implemented in PopDel identifies more deletions than other approaches at a similar or better accuracy.

Application to population-scale data from Iceland

We applied PopDel to whole-genome data of 49,962 Icelanders including 6,794 parent-child trios (Methods). The average number of deletions PopDel reports per Icelander on the 22 autosomes is 2,826 (without filtering on genotype quality). Using the same filtering criteria as for the Polaris trios, the Mendelian inheritance error rate in the 6,794 trios is 1.2% (1,859 consistent deletions on average per trio). The transmission rate for 4,180 deletions is 49.6%. While the Mendelian inheritance error rate is slightly higher in this large cohort than in the 49 Polaris trios, the transmission rate remains very close to the expected 50%. This implies that the majority of errors appear as common deletions shared by several individuals.

Identification of a* de novo *deletion in the Polaris data

In the deletions reported by PopDel for the Polaris Kids Cohort we searched for *de novo* deletions (Methods). We identified an 8901 bp deletion at chr6:93035858-93044759 in the Spanish individual HG01763 but not in her parents (Figure 5). Given that this deletion is intergenic and HG01763 is part of a cohort of healthy individuals[44], we expect the *de novo* deletion not to be of medical relevance. The closest transcript annotations in Gencode v29[45] are the lncRNA *AL138731.1* at a distance of 25.6 kb and the *EPHA7* gene in a distance of 195.3 kb. Interestingly, the deletion is close to a SNV that overlaps with deletion-supporting read pairs and allowed us to determine that the deletion haplotype was inherited from the mother (HG01762). Further evidence for this to be a true *de novo* event is given by 25 SNVs within the deletion that confirm the child to carry a single haplotype where both parents are heterozygous. All three individuals are heterozygous for numerous SNVs upstream of the deletion. Four of the SNVs within the deletion confirm that the event happened on a maternal haplotype.

Discussion

Identification and genotyping of structural variation in sequencing cohorts is a major computational challenge. To enable the analysis of the increasingly large cohorts that are being sequenced, we developed a novel joint deletion calling approach and implemented it in the tool PopDel. Compared to existing tools, the joint calling approach in PopDel greatly simplifies the analysis workflow and shows tremendous improvements in the required compute time. This indicates that PopDel scales to very large cohorts and our tests on population-scale data from Iceland substantiate its scalability.

PopDel consists of two steps: creation of read pair profiles per sample and joint deletion calling. The computational advantage of this two-step design is that the large input BAM files containing aligned read data need to be processed only once. The joint calling step takes all information needed for deletion detection and genotyping from the small read pair profiles. This implies that additional samples can be added to the analysis without the need to access all input BAM files again reducing the computational burden considerably.

The number of individuals studied does not severely affect the accuracy of PopDel. On data of a single individual, PopDel is competitive with previous tools. On the Polaris Diversity Cohort, the deletions called by PopDel recapitulate previous population genetic results showing that Africans carry more deletions than other continental groups and confirming that joint calling can be used to identify population structure. On the Polaris Kids Cohort, PopDel identifies more deletions at a better transmission rate for rare variants compared to other tools and reports a *de novo* deletion of about 9 kb. On Icelandic data, PopDel identifies deletions jointly in almost 50,000 genomes maintaining an excellent transmission rate for rare variants. All results confirm that the joint calling approach in PopDel is accurate across the frequency spectrum and the number of individuals analyzed.

The *de novo* deletion in the Polaris Kids Cohort together with the low transmission rate in the large number of Icelandic genomes demonstrates that PopDel provides a basis for studying rarely observed *de novo* deletion events. A previous study verified seven *de novo* deletions in the size range addressed by PopDel in 258 healthy trios[46]. Given their rate of *de novo* deletions, we expect to observe 1.33 medium-size *de novo* deletions in the 49 Polaris trios. This is well in line with our finding of a single *de novo* event.

When we tested an early version of PopDel on a selected 54 kb region covering the *LDLR* gene in 43,202 Icelanders, we identified a previously unknown 2.5 kb deletion in three closely related Icelanders shown to affect LDL levels (Björnsson et al. manuscript in preparation). This finding shows that PopDel is able to identify variants of biomedical interest even if they are present at a very low allele frequency in a population-scale cohort, and it showcases the importance of SVs in human health.

URLs

PopDel source code, https://github.com/kehrlab/PopDel; Scripts used for running the tools and evaluation, https://github.com/kehrlab/PopDel-scripts; Long read reference call set, ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz; Short read reference call set, ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/technical/svclassify_Manuscript/Supplementary_Information/Personalis_1000_Genomes_deduplicated_deletions.bed; NCBI genome remapping service, https://www.ncbi.nlm.nih.gov/genome/tools/remap;

Author contributions

B.K. and B.V.H. conceived the approach. S.R. and B.K. developed PopDel, designed the experiments and evaluated the results. S.R. simulated data, performed analyses on simulated and public data and tested on Icelandic data. B.V.H. applied PopDel on Icelandic data and assisted in testing. E.B., P.S. and K.S. studied the *LDLR* deletion. H.J assisted in the computation of transmission rates. H.P.E. and D.B. assisted in the analyses of Icelandic data. B.K and S.R. drafted the manuscript with feedback from B.V.H. All authors revised the draft and approved the final manuscript.

Competing financial interests

H.J., E.B., H.P.E., D.B., P.S, K.S. and B.V.H. are employees of deCODE Genetics/Amgen, Inc.

References

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
2. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).
3. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549,** 519–522 (2017).
4. Halldorsson, B. V *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363,** eaau1043 (2019).
5. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17,** 224–238 (2016).
6. Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat. Commun.* **6,** 7256 (2015).
7. Goldmann, J. M. *et al.* Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50,** 487–492 (2018).
8. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
9. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
10. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
11. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
12. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47,** D1005–D1012 (2019).
13. Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y. & Lauschke, V. M. Integrating rare genetic variants into pharmacogenetic drug response predictions. *Hum. Genomics* **12,** 26 (2018).
14. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).
15. The NIHR BioResource. Whole-genome sequencing of rare disease patients in a national healthcare system. *bioRxiv* 507244 (2019).
16. Gilly, A. *et al.* Cohort-wide deep whole genome sequencing and the allelic

architecture of complex traits. *Nat. Commun.* **9,** 4674 (2018).

17. Yuen, R. K. C. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20,** 602–611 (2017).
18. Mak, A. C. Y. *et al.* Whole-Genome sequencing of pharmacogenetic drug response in racially diverse children with asthma. *Am. J. Respir. Crit. Care Med.* **197,** 1552–1564 (2018).
19. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019).
20. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
21. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* 563866 (2019).
22. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2017).
23. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49,** 1654–1660 (2017).
24. Guan, P. & Sung, W.-K. Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* **102,** 36–49 (2016).
25. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
26. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Publ. Gr.* **43,** (2011).
27. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. *bioRxiv* 508515 (2018).
28. Larson, D. E. *et al.* svtools: population-scale analysis of structural variation. *bioRxiv* 494203 (2018).
29. Zarate, S. *et al.* Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers. *bioRxiv* 424267 (2018).
30. Mohiyuddin, M. *et al.* MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* **31,** 2741–4 (2015).
31. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12,** 363–376 (2011).
32. Escaramís, G., Docampo, E. & Rabionet, R. A decade of structural variants: description, history and methods to detect structural variation. *Brief. Funct. Genomics* **14,** 305–314 (2015).
33. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21,** 974–84 (2011).
34. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47,** 296–303 (2015).
35. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).
36. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic

framework for structural variant discovery. *Genome Biol.* **15,** R84 (2014).

37. Chiang, C. *et al.* SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12,** 966–968 (2015).

38. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32,** 1220–1222 (2016).

39. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27,** 2050–2060 (2017).

40. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. doi:10.1101/110387

41. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37,** 561–566 (2019).

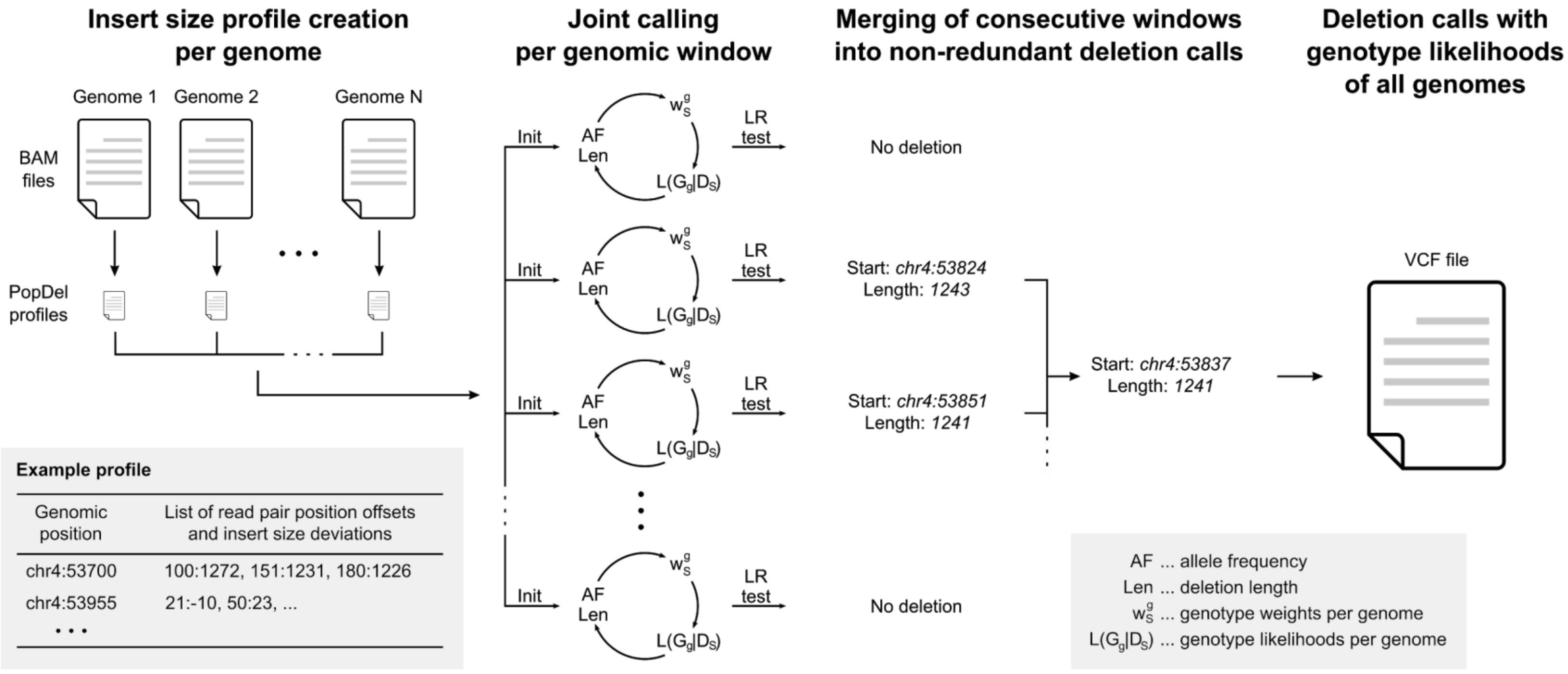
42. Zook, J. M. *et al.* A robust benchmark for germline structural variant detection. *bioRxiv* 664623 (2019).

43. Collins, R. L. *et al.* An open resource of structural variation for medical and population genetics. *bioRxiv* 578674 (2019).

44. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

45. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47,** D766–D773 (2019).

46. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25,** 792–801 (2015).



Example profile

Genomic position	List of read pair position offsets and insert size deviations
chr4:53700	100:1272, 151:1231, 180:1226
chr4:53955	21:-10, 50:23, ...
• • •	

AF ... allele frequency
Len ... deletion length
w_S^g ... genotype weights per genome
$L(G_g|D_S)$ ... genotype likelihoods per genome

Figure 1. The approach implemented in PopDel is divided into two steps. The first step reduces the BAM file of one individual at a time into a small profile. The second step processes the profiles of all individuals together by sliding a window (of size 30 bp by default) over the genome and assessing the likelihood of each window to overlap with a deletion in any individual. Sizes and allele frequencies of the deletions are estimated iteratively. Consecutive windows are combined into a single variant call and genotypes of all individuals are output. Init, Initialization; LR, likelihood ratio.

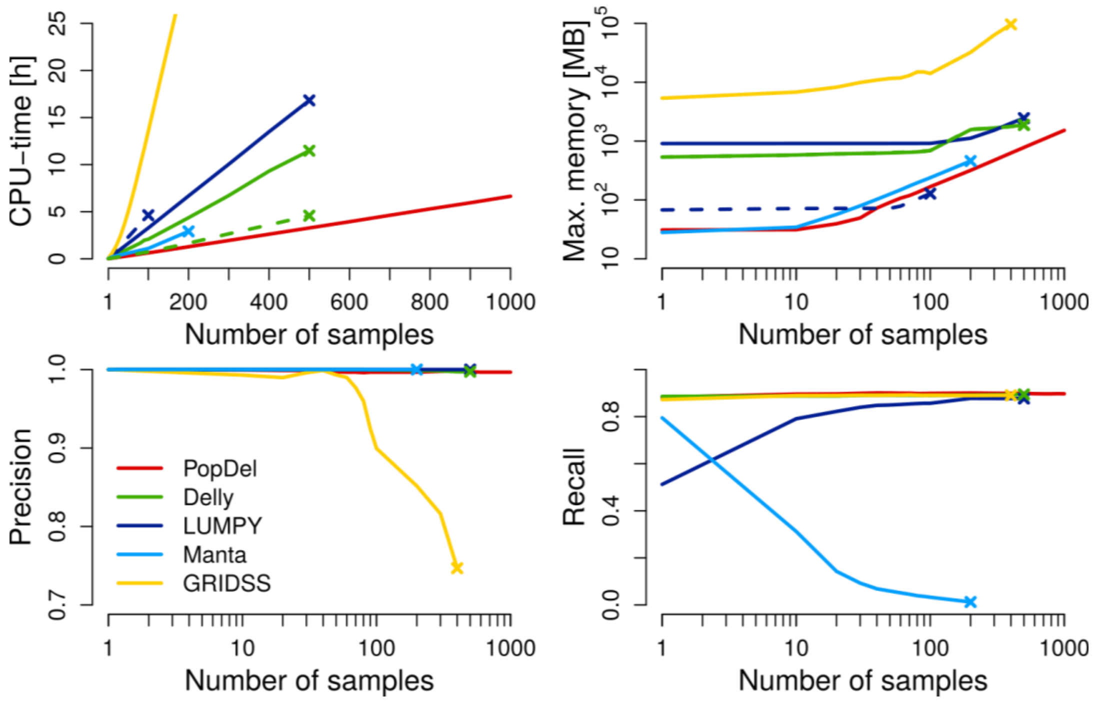


Figure 2. Performance on chromosome 21 data simulated for increasing numbers of samples. Delly and Lumpy were applied in two settings: Lumpy (solid line) and Lumpy-Express (dashed line), and Delly without (solid line) and with (dashed line) the option '--noIndels'. Lumpy's results include genotyping with SVTyper. Lines stopping with an 'x' before 1,000 samples signal that the respective tool crashes when trying to process an additional 100 samples.

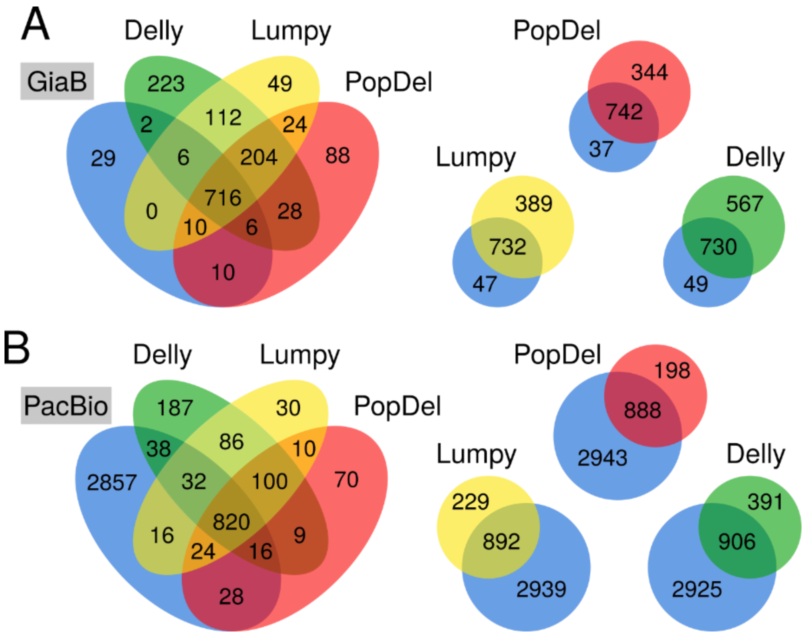


Figure 3. Call set overlap for PopDel, Delly and Lumpy on NA12878 with different reference call sets. (A) Overlap with GiaB short read call set. (B) Overlap with PacBio long read call set.

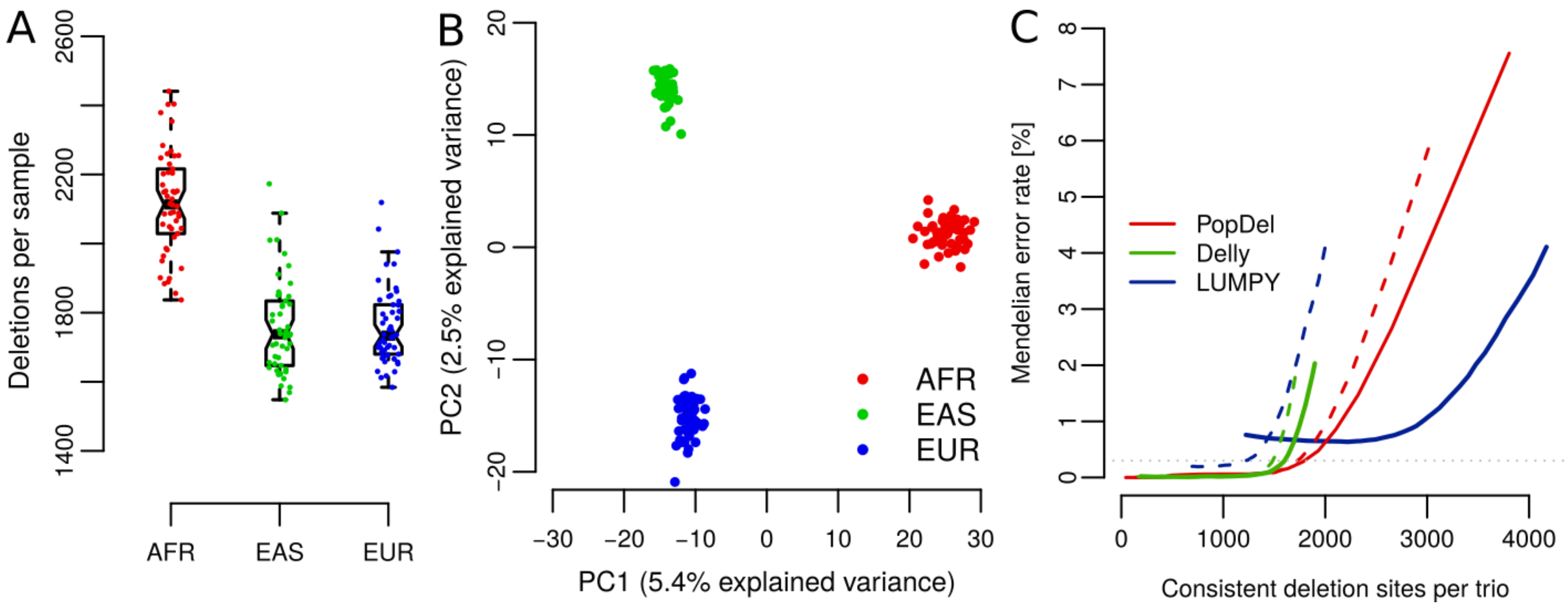


Figure 4. (A) Number of deletions per sample by continental group as called by PopDel on the Polaris Diversity Cohort. All data points are shown in color. Boxes indicate the first and third quartile, the center line indicates the median, whiskers extend to the most extreme data points that are not more than 1.5 interquartile ranges (IQR) away from the median, and notches end at +/-1.58 IQR/sqrt(n) where n is the number of data points. (B) Principal component analysis of PopDel's calls from the Polaris Diversity Cohort. (C) Mendelian inheritance error rate by number of consistent deletion sites per trio for the Polaris Kids Cohort. Dashed lines indicate a call sets where duplicate calls (≥ 50% reciprocal overlap) are removed.

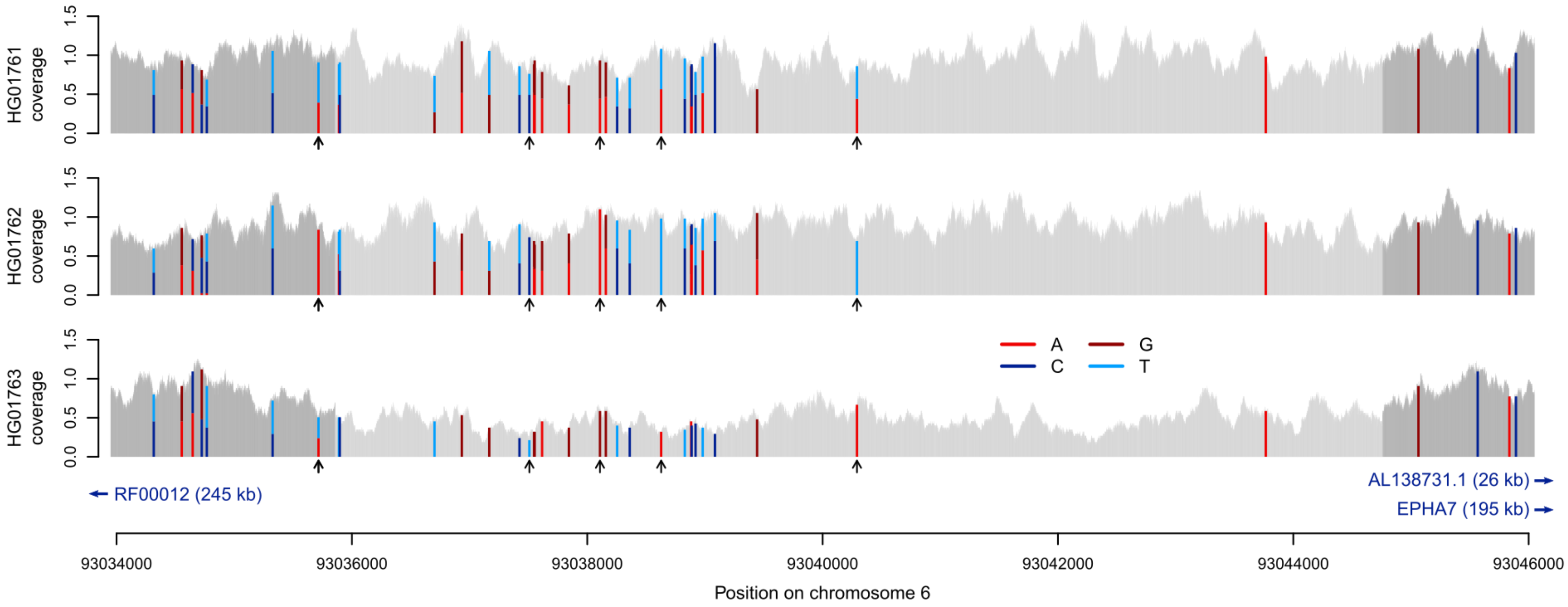


Figure 5. De-novo deletion identified by PopDel in one trio of the Polaris Kids Cohort. The child (HG0173, bottom) is carrier of a heterozygous deletion not present in either parent. The arrows mark the SNPs that allow the phasing of the haplotypes in the child.

Table 1. Running times of tested tools on NA1278 and Polaris Diversity Cohort.

	Running times (CPU hours)	
	NA12878 (single individual)	**Polaris Diversity Cohort (150 individuals)**
PopDel	0:29	60:17
Delly	3:24	483:35 (*)
Lumpy + SVTyper	8:03	2,467:37 (*)

(*) Single sample calling with subsequent variant merging and sample-wise genotyping.

Note that Delly and Lumpy report other types of SVs apart from deletions. These SVs were excluded before variant merging and genotyping whenever possible.

Methods

Read pair profile creation

PopDel reduces a coordinate-sorted BAM file of each sample into a read pair profile in a custom binary format (**Supplementary Note**). This profile stores positions and insert sizes of read pairs that align confidently (**Supplementary Note**) to the reference genome. In addition, the profile file contains meta information, including a distribution of insert sizes across the sample and an index, which allows for jumping to genomic positions in the profile. We define the insert size as the distance between the leftmost alignment position of the forward read to the rightmost alignment position of the reverse read in the pair extended by any clipped bases (**Supplementary Figure 1**). The null distribution of insert sizes is estimated by sampling the BAM file using pre-defined but user-configurable genomic regions with good mappability (**Supplementary Note**). If more than one library has been sequenced for a sample, PopDel writes separate profile data per read group to the profile file. An excerpt of an example profile is shown in the **Supplementary Note**. The profiling vastly reduces I/O during joint calling as the size of the profiles is on average only 1.76% of the original BAM file size (**Supplementary Note**).

Likelihood ratio test for joint deletion calling

The likelihood ratio test for a given genomic window compares the relative likelihood that a deletion of a certain length l overlaps the window against the relative likelihood of observing the reference haplotype:

$$\Lambda = \frac{\mathcal{L}(\text{del of length } l)}{\mathcal{L}(\text{no del})}$$

Let $S \in \mathcal{S}$ be a single sample from the set of all samples $\mathcal{S}$ and let I^S be the list of insert sizes for all the read pairs of S overlapping the given window (**Supplementary Figure 1**). Furthermore, let $\Delta^S = (i - \mu_S \mid i \in I^S)$ be the deviations of the insert sizes from the mean μ_S. We assume independence of samples and calculate the likelihood of the reference model as the product of the samples' likelihoods $\mathcal{L}(G_0|\Delta^S)$ for the reference genotype G_0

$$\mathcal{L}(\text{no del}) = \prod_{S \in \mathcal{S}} \mathcal{L}(G_0|\Delta^S)$$

For the likelihood of the deletion model we use the weighted sums of all three genotype likelihoods in a similar product

$$\mathcal{L}(\text{del of length } l) = \prod_{S \in \mathcal{S}} \sum_{g=0}^{2} \left(a_g^S \cdot \mathcal{L}\left(G_g | \Delta^S\right) \right)$$

where the a_g^S are sample- and genotype-specific weights (see below) with genotypes $g \in \{0,1,2\}$ corresponding to 0, 1 or 2 variant alleles and $a_0^S + a_1^S + a_2^S = 1$ for any $S \in \mathcal{S}$.

Assuming Wilks' theorem, which states that twice the logarithm of a likelihood ratio is asymptotically χ^2-distributed, we calculate a cutoff for Λ in PopDel using $2 \log \Lambda \sim \chi^2$ with a p-value threshold of 0.01 (one-tailed) and 1 degree of freedom in order to decide if the window overlaps a deletion.

Iterative estimation of parameters for the likelihood ratio test

The likelihood ratio test requires as input a deletion length, genotype likelihoods for all samples and sample- and genotype-specific weights. PopDel estimates these values for each window iteratively from the profiles together with an allele frequency that is needed for updating the weights (Figure 1). For simplicity, the following assumes one read group per sample but our implementation in PopDel also handles multiple read groups (**Supplementary Note**). To be able to detect deletions of different lengths from different haplotypes overlapping the same window, the iteration and likelihood ratio test are performed for several initializations of the deletion length. Initial lengths are estimated by identifying samples with similar third quartiles of Δ^S via greedy clustering (**Supplementary Note**). The initial allele frequencies f are set to the fraction of deletion-supporting reads pairs of all samples in the window (**Supplementary Note**). To calculate the *genotype likelihoods* of the three genotypes G_0, G_1 and G_2 of a single sample S, PopDel transforms the insert size histogram of S to reflect how many read pairs with a given insert size deviation $\delta \in \Delta^S$ are expected to overlap a window of size w (**Supplementary Note**). Let $H^S(\delta)$ be the resulting relative likelihood of observing a read pair with insert size deviation δ. We call a deviation *informative* if $H^S(\delta - \epsilon_S) \geq 2 \cdot H^S(\delta - l)$ or $2 \cdot H^S(\delta - \epsilon_S) \leq H^S(\delta - l)$, and define $\bar{\Delta}^S \subseteq \Delta^S$ to be the set of *informative* deviations. Furthermore, ϵ_S is a sample-specific reference shift (**Supplementary Note**) that accounts for local biases of the data such as GC-content[47,48].

PopDel calculates the likelihoods $\mathcal{L}\left(G_g|\Delta^S\right)$ as

$$\mathcal{L}(G_0|\Delta^S) = \prod_{\delta \in \bar{\Delta}^S} H^S(\delta - \epsilon_S)$$

$$\mathcal{L}(G_1|\Delta^S) = \binom{|\bar{\Delta}^S|}{k} p^k (1-p)^{(|\bar{\Delta}^S|-k)} \prod_{\delta \in \bar{\Delta}^S} \max\left(H^S(\delta - \epsilon_S), H^S(\delta - l)\right)$$

$$\mathcal{L}(G_2|\Delta^S) = \prod_{\delta \in \bar{\Delta}^S} H^S(\delta - l)$$

where we choose $p = 0.4$ and k amounts to the number of deviations $\delta \in \bar{\Delta}^S$ for which $H^S(\delta - l) > H^S(\delta - \epsilon_S)$.

The *sample- and genotype-specific weights* a_g^S are designed to give low weight to samples with a small likelihood for the genotype and a high weight to those with a good one and make it more likely to observe a carrier genotype when the allele frequency is high:

$$a_g^S = \frac{\mathcal{L}\left(G_g|\Delta^S\right) \cdot \mathcal{L}\left(f, G_g\right)}{\sum_{j=0}^{2}\left(\mathcal{L}\left(G_j|\Delta^S\right) \cdot \mathcal{L}\left(f, G_j\right)\right)}$$

with $\mathcal{L}\left(f, G_g\right)$ as

$$\mathcal{L}(f, G_0) = (1-f)^2$$

$$\mathcal{L}(f, G_1) = 2f(1-f)$$

$$\mathcal{L}(f, G_2) = f^2$$

Given the weights, the *allele frequency* f is updated using:

$$f^{new} = \frac{1}{2|\mathcal{S}|} \cdot \sum_{S \in \mathcal{S}} (a_1^S + 2a_2^S)$$

To update the *deletion length* l, probabilities $P_{l,\epsilon}^S(\delta)$ reflecting that a given insert size deviation δ resulted from a distribution shifted by l rather than by ϵ_S are calculated as

$$P_{l,\epsilon_S}^S(\delta) = a_1^S \cdot \frac{H^S(\delta - l)}{H^S(\delta - \epsilon_S) + H^S(\delta - l)} + a_2^S$$

and used to update l jointly across all samples as the weighted sum over all insert size deviations:

$$l^{new} = \frac{\sum_{S \in \mathcal{S}} \sum_{\delta \in \Delta^S} \delta \cdot P_{l,\epsilon_S}^S(\delta)}{\sum_{S \in \mathcal{S}} \sum_{\delta \in \Delta^S} P_{l,\epsilon_S}^S(\delta)}$$

The iteration for parameter estimation terminates when both the allele frequency and deletion length converge or additional termination conditions are met, e.g. reaching the maximum number of iterations (default 15) (**Supplementary Note**). A *start position* of the potential deletion is estimated during above calculations by keeping track of the rightmost ends of the forward reads of read pairs whose δ supports the deletion estimate (**Supplementary Note**).

Merging of consecutive deletion windows

To provide the user with a non-redundant list of deletion variants, adjacent windows that support the same deletion need to be merged. PopDel sorts all windows for which the null hypothesis of the likelihood ratio test can be rejected in ascending order of the predicted deletion start position, deletion length and descending deletion likelihood ratio. Moving over this sorted list of windows $w_0, w_1, \ldots$, a window $w_i, i \geq 0$ is merged with another window $w_{i+k}, k > 0$ if their start positions and deletion sizes are similar enough (**Supplementary Note**). When no more windows can be merged with w_i, a deletion is output with a start position and length calculated as the median over all merged windows. Merging continues with the next window w_{i+k+1} that has not been merged with any other window so far.

Deletion output

We report the genotype with the best mean PHRED-scaled genotype likelihood across the merged windows of one sample in the output. Samples without sufficient data or much higher than average coverage at the locus are not genotyped (**Supplementary Note**). The allele frequency is estimated by counting the number of alleles predicted to carry the variant, divided by the total number of genotyped alleles. We calculate a genotype quality as the difference of the best and second best PHRED-scaled genotype likelihoods.

Simulation of chr21 data

We simulated deletion variants with uniformly distributed length between 100 and 10,000 bp, uniformly distributed positions on chromosome 21 of GRCh38 and uniformly distributed allele frequency between 0 and 1. Regions containing ‘N’s were excluded and deletion were required to be at least 1000 bp apart. Using this set of deletions, we created 2,000 haplotypes by sampling deletions according to their allele frequency and inserting them into chromosome 21 of GRCh38. The haplotypes were combined into 1,000 diploid samples. The samples were subsequently used to simulate NGS reads with *art_illumina*[49] and the reads aligned to GRCh38 using *BWA-mem*[50].

Setup of SV callers on simulated data

PopDel (1.1.0) and Manta (1.4.0) were run with an option to limit the calling to chr21. Delly (0.7.8) was applied once with default parameters and once without small indel realignment (-n). GRIDSS (1.8.1) was provided a maximum heap size of 8 GB. Lumpy (0.2.13) was applied once only relying on *lumpyexpress* and once with prior extraction of split reads and discordant read pairs using samtools (1.7) (subsequently called *multi-step)*. Joint genotyping of the Lumpy calls was performed using SVtyper (0.6.0). All tools were applied on increasing numbers of BAM files, up to 1000 or until failure.

Evaluation on simulated data

Running time and memory consumption were measured on a dedicated work station (Intel Xeon E5-1630v3 8x3.5GHz, 64GB RAM) using '/usr/bin/time'. As PopDel and Lumpy consist of multiple steps, the running times are the sum of the time taken by all steps from the BAM files to the VCF file. The memory consumption of these tools is stated as the maximum memory consumption of all steps. As GRIDSS produces two break-ends per deletion, corresponding pairs of break-ends were collapsed into a single call and “LOW_QUAL” variants were removed. The calls of Lumpy and Delly were not filtered for variants that had the filter field set to “PASS” as this would have had a negative impact on their performance. A call and a simulated variant were considered to be the same if they had a reciprocal overlap of at least 50%. Each simulated variant is only allowed to be matched with one predicted variant. See **Supplementary Note** for results using alternative equality criteria.

Calling and filtering on real data

All samples were mapped to the human reference genome (GRCh38)[51] using *BWA-mem*[52,53]. If not stated otherwise, calling was performed jointly on all samples with default options and, if possible, limited to the reference sequence of the 22 autosomes. On NA12878 we further included chromosome X. Variants were filtered to the size range from 500 to 10,000 bp. The Polaris data and Icelandic data was filtered to high-confidence calls, where we kept only deletions with genotype quality scores above a fixed threshold. This threshold was chosen once per tool on the Polaris Kids Cohort such that the Mendelian inheritance error rate dropped below 0.3%[42]: 27 for PopDel, 28 for Delly and 104 for Lumpy+SVTYPER. The threshold for PopDel was set to 50 to search for *de novo* deletions.

The sample order of the Polaris Diversity and Kids Cohorts was shuffled but the same for all tools and no tool was provided pedigree information. Joint calling with Delly (-n) and Lumpy (multi-step) did not finish within 4 weeks. Therefore, they were run using single sample calling with merging and sample-wise re-genotyping (**Supplementary Note**). Calls other than deletions were removed early in Delly's and Lumpy's pipelines to reduce running time (**Supplementary Note**). Lumpy was run as lumpyexpress. Variants genotyped as 0/0 in all samples (as found in Lumpy+SVTyper's output) were removed. Delly (-n) variants were only considered if they had the 'FILTER' field set to 'PASS'. Significance of the difference in the number of variant alleles between continental groups was assessed using the R-function *t.test* with *var.equal=FALSE*.

Reference call sets for NA12878

GiaB short read and PacBio reference call sets were downloaded (see URLs), filtered for deletion variants and liftover from GRCh37 to GRCh38 was performed using the *NCBI Genome Remapping Service* (see URLs). All contigs except for chromosomes 1 to 22 and chromosome X were removed and VCF-to-BED conversion was performed for the PacBio callset.

Filtering of centromeric regions

Centromeric regions for GRCh38 were obtained through the UCSC table browser ("group: Mapping and Sequencing", "track: Centromeres"). Any variants in a reference set or call set having any overlap with a region in the BED file of centromeres were removed. Overlap was determined using *bedtools intersect*[54] (**Supplementary Note**).

Principal Component Analysis

Predicted genotypes of the Polaris Diversity Cohort were converted into a variant/sample matrix containing variant allele counts. Uninformative variants and those in linkage disequilibrium were removed (**Supplementary Note**). PCA was computed using the R-function *prcomp.*

Mendelian inheritance error rate and transmission rate

The Mendelian inheritance error rate was calculated on chromosomes 1 to 22. Duplicates within one call set were removed by removing all calls that had at least 50% reciprocal overlap with an upstream call. For all reported deletions, the three genotypes in each trio were inspected for Mendelian consistency (**Supplementary Note**). Trios with one or more missing genotypes and trios with all three samples genotyped as 0/0 were ignored. For calculating the transmission rate, we considered only deletions that were called in a single trio, one parent is a heterozygous carrier and the other parent carries the reference allele on both haplotypes. The transmission rate is the number of deletion alleles transmitted from the heterozygous parents to the children divided by the number of considered deletions.

Sequence data from 49,962 Icelanders

DNA was isolated from both blood and buccal samples. All participating subjects signed informed consent. The personal identities of the participants and biological samples were encrypted by a third-party system approved and monitored by the Data Protection Authority. The National Bioethics Committee and the Data Protection Authority in Iceland approved these studies. The Icelandic samples were whole-genome sequenced at deCODE Genetics using Illumina GAIIx, HiSeq, HiSeqX and NovaSeq sequencing machines, and sequences were aligned to the human reference genome (GRCh38)[51] using *BWA-mem*[52,53]. Details of the sample preparation, paired-end sequencing, read processing and alignment, and selection of the final set of BAM files have been previously described[55].

References

47. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** e72–e72 (2012).

48. Iakovishina, D., Janoueix-Lerosey, I., Barillot, E., Regnier, M. & Boeva, V. SV-Bay: structural variant detection in cancer genomes using a Bayesian approach with correction for GC-content and read mappability. *Bioinformatics* **32,** 984–992 (2016).

49. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28,** 593–594 (2012).

50. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

51. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27,** 849–864 (2017).

52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997 (2013).

54. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

55. Jónsson, H. *et al.* Data Descriptor: Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4,** (2017).