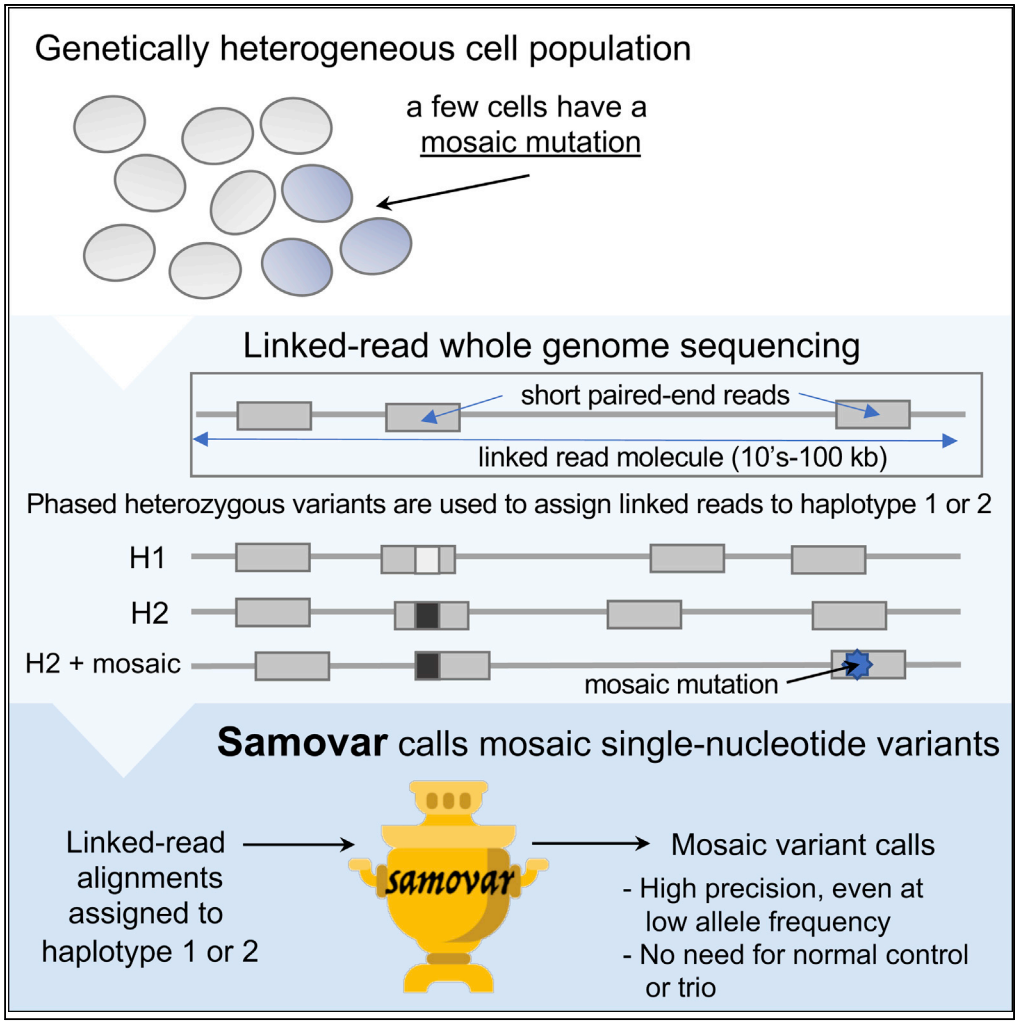
iScience

CellPress

Special Issue: RECOMB-Seq 2019

Article

Samovar: Single-Sample Mosaic Single-Nucleotide Variant Calling with Linked Reads



Charlotte A. Darby, James R. Fitch, Patrick J. Brennan, ..., Peter White, Ben Langmead, Michael C. Schatz

langmea@cs.jhu.edu (B.L.)
mschatz@cs.jhu.edu (M.C.S.)

HIGHLIGHTS

Samovar uses haplotype-specific features from linked reads to call mosaic variants

Samovar quickly evaluates candidates with a random forest over 33 features

Only one sample is needed, with accuracy comparable with paired samples or trios

Samovar finds somatic variants in cancer driving genes in 13 pediatric cancer cases

Darby et al., iScience *18*, 1–10
August 30, 2019
https://doi.org/10.1016/j.isci.2019.05.037

iScience

CellPress

Special Issue: RECOMB-Seq 2019

Article

Samovar: Single-Sample Mosaic Single-Nucleotide Variant Calling with Linked Reads

Charlotte A. Darby,[1] James R. Fitch,[2] Patrick J. Brennan,[2] Benjamin J. Kelly,[2] Natalie Bir,[2] Vincent Magrini,[2,3] Jeffrey Leonard,[3,4] Catherine E. Cottrell,[2,3] Julie M. Gastier-Foster,[2,3] Richard K. Wilson,[2,3] Elaine R. Mardis,[2,3] Peter White,[2,3] Ben Langmead,[1,*] and Michael C. Schatz[1,5,6,7,*]

SUMMARY

Linked-read sequencing enables greatly improves haplotype assembly over standard paired-end analysis. The detection of mosaic single-nucleotide variants benefits from haplotype assembly when the model is informed by the mapping between constituent reads and linked reads. Samovar evaluates haplotype-discordant reads identified through linked-read sequencing, thus enabling phasing and mosaic variant detection across the entire genome. Samovar trains a random forest model to score candidate sites using a dataset that considers read quality, phasing, and linked-read characteristics. Samovar calls mosaic single-nucleotide variants (SNVs) within a single sample with accuracy comparable with what previously required trios or matched tumor/normal pairs and outperforms single-sample mosaic variant callers at minor allele frequency 5%–50% with at least 30X coverage. Samovar finds somatic variants in both tumor and normal whole-genome sequencing from 13 pediatric cancer cases that can be corroborated with high recall with whole exome sequencing. Samovar is available open-source at https://github.com/cdarby/samovar under the MIT license.

INTRODUCTION

Genomic mosaicism results from postzygotic *de novo* mutations, ranging from single-nucleotide changes to larger structural variants and whole chromosome aneuploidy. Mosaic mutations are present in some of the cells belonging to the offspring but in none of either parents' cells (Biesecker and Spinner, 2013; Cohen et al., 2015). The distribution and prevalence of cells with a mosaic mutation depend on a combination of the developmental cell lineage, stage at which the mutation occurred, selection for or against cells with the mutation (Youssoufian and Pyeritz, 2002), and cell migration (Freed et al., 2014). Somatic mosaicism refers to genetic heterogeneity among non-germ cells, which accrue in normally dividing cells throughout the human lifetime (Gajecka, 2016; Laurie et al., 2012; Kennedy et al., 2012) corroborated by monozygotic twin studies (Ouwens et al., 2018). Mosaicism also plays an important role in many genetic diseases. Pathologically, cancer is characterized by an overall increased mutational load in tumor cells as well as a high level of intra-tumor genetic heterogeneity (Vogelstein et al., 2013; Watson et al., 2013). Mosaicism has also been implicated in autism (Freed and Pevsner, 2016) and is being explored in connection to other neurological disease (Poduri et al., 2013; McConnell et al., 2017; D'Gama and Walsh, 2018). Causal mosaic mutations have also been found for Sturge-Weber syndrome (Shirley et al., 2013), McCune-Albright syndrome (Weinstein et al., 1991), and Proteus syndrome (Lindhurst et al., 2011), among others.

Mosaic variants can be detected by whole-genome or targeted sequencing of affected tissue. Samovar operates on linked reads, which are sets of sequencing reads deriving from a longer fragment such as those from the 10X Genomics Chromium instrument (Pleasanton, CA, USA). Although the individual ("constituent") reads are typical short Illumina reads, the longer fragments can be tens or hundreds of kilobases long. The mapping from constituent reads to fragments of origin is established by molecular barcodes added in the Chromium library preparation step. The average sequencing coverage per long fragment is usually low: around 0.1-fold (Zheng et al., 2016; Marks et al., 2019). Since constituent reads can be paired-end, we use the term "long fragment" for the longer fragment from which a linked read is derived and "short fragment" for fragments from which paired-end reads are derived.

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

[2]The Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA

[3]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, USA

[4]Department of Neurosurgery, Nationwide Children's Hospital, Columbus, OH, USA

[5]Department of Biology, Johns Hopkins University, Baltimore, MD, USA

[6]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

[7]Lead Contact

*Correspondence: langmea@cs.jhu.edu (B.L.), mschatz@cs.jhu.edu (M.C.S.)

https://doi.org/10.1016/j.isci.2019.05.037

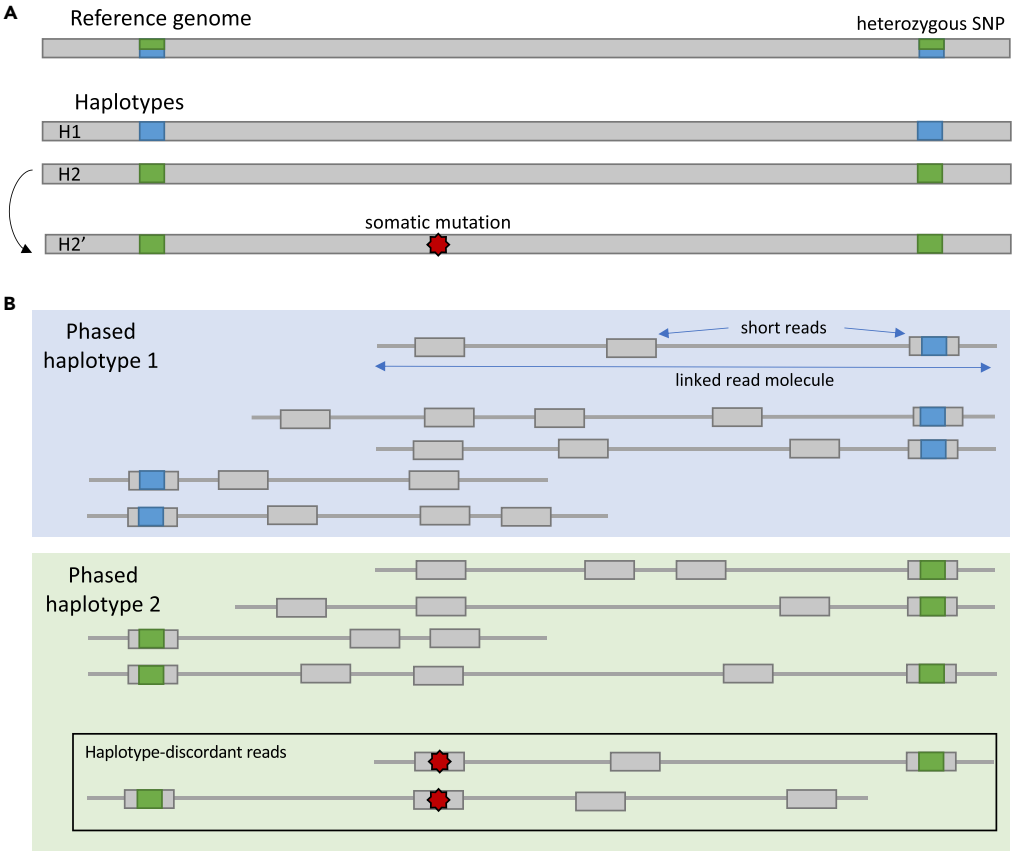


Figure 1. Schematic Representation of Somatic Mutations within a Phased Sample

(A) A mosaic mutation occurs on haplotype H2.

(B) Therefore, in linked-read sequencing, where short reads can be phased when linked reads overlap phased heterozygous variants, mosaic mutations manifest on reads from only one haplotype, here H2. Adapted from Figure 3 of Dou et al., 2018.

The properties of linked reads enable many potential improvements in variant detection and related analyses (Sedlazeck et al., 2018). For example, a constituent read that would align repetitively by itself might align uniquely when alignments of other reads from the same long fragment are accounted for (Bishara et al., 2015; Shajii et al., 2018). Linked-read-based algorithms have been developed for *de novo* assembly (Kuleshov et al., 2016; Weisenfeld et al., 2017; Mostovoy et al., 2016), *de novo* mutation calling (Zhou et al., 2018), assembly error correction (Jackman et al., 2018), and structural variant calling (Elyanow et al., 2018; Xia et al., 2018; Spies et al., 2017; Eslami Rasekh et al., 2017; Fang et al., 2018). Also, linked reads enable more accurate and contiguous assembly of haplotypes (Zheng et al., 2016; Marks et al., 2019; Edge et al., 2017) since constituent reads can be phased even when only some overlap heterozygous variants (Figure 1B).

Although downstream tools benefit automatically from some linked-read properties, e.g., improved alignment accuracy, other benefits require specialized methods to exploit. In particular, the detection of a somatic mosaic single-nucleotide variant (SNV) can benefit from haplotype assembly when the variant detection model is informed by the mapping between constituent reads and linked reads. As an example, in a diploid sample with haplotypes H1 and H2, suppose a mosaic mutation occurs on haplotype H2 yielding a collection of reads (labeled H2′) that have the mosaic allele but otherwise match H2 (Figure 1A). The mosaic mutation will likely be tolerated by the haplotype assembler, and the reads will still be assigned to H2 (Figure 1B). The fact that all the mosaic-carrying reads fall on the same haplotype is a hallmark of post-zygotic mosaicism (Freed and Pevsner, 2016) and contrasts with sequencing error, which would tend to distribute the "mosaic" alleles evenly across haplotypes (Usuyama et al., 2014). Reads with the mosaic allele

are called haplotype-discordant reads, and these are the most reliable kind of evidence we can gather in support of mosaic variants.

The mosaic variant caller's task is to distinguish the signature of a mosaic variant from that of a germline variant after it has been affected by sequencing errors, alignment errors, copy-number changes, and other confounders. Most methods employ statistical tests on the sequencing reads aligned to a particular site, comparing allele frequency between "tumor" and "normal" (or between the observed and expected value for a germline variant). HapMuC (Usuyama et al., 2014) uses haplotype phasing of nearby heterozygous germline variants in conjunction with a tumor-normal pair to call somatic variants, but local phasing is limited by read length of paired-end short reads. In single-cell linked-read data, LiRA (Bohrson et al., 2019) leverages heterozygous germline variants and the additional locality information of linked reads to call mosaic SNVs. See Dou et al., 2018 for a review of methods to detect such mutations in scenarios other than cancer and Wang et al., 2013 for a comparison of several tools in the cancer context. Samovar is unique in that it is the first to evaluate haplotype-discordant reads identified through linked-read sequencing, thus enabling phasing and mosaic variant detection across essentially the entire genome. It also evaluates the statistical characteristics of the haplotypes, depth of coverage, and potential confounders such as alignment errors to robustly identify mosaic variants from a single sample.

RESULTS

Samovar Pipeline

We present Samovar, a single sample mosaic SNV caller designed for 10X Genomics linked-read whole-genome sequencing (WGS) data. Samovar takes as input phased variants in VCF format and linked-read alignments in BAM format. These are both output by 10X Genomics' Long Ranger pipeline, which preprocesses reads, aligns linked reads, calls variants, and assembles haplotypes.

The Samovar workflow is shown in Figure 2 and proceeds in six major steps. In step 1, Samovar identifies all genomic sites where there are sufficient data to apply our model. This is done by filtering based on features such as depth of coverage, fraction of reads that are phased, frequency of the candidate mosaic allele, and related data characteristics. In step 2, Samovar modifies the input BAM file to introduce synthetic mosaic variants to be used as sample-specific training data. Specifically, these variants are used as positive examples for training our model, whereas real homozygous/heterozygous variants, as called by Long Ranger, are used as negative examples. In step 3, Samovar trains a random forest model containing an ensemble of 100 individual decision trees that scores sites according to their resemblance to the synthetic-mosaic sites. In step 4, Samovar scores all sites that passed the initial filter using this model. In step 5, complex repeat regions and non-diploid copy-number regions are optionally filtered out. In step 6, a final filter removes false positives resulting from alignment errors to produce scored mosaic variant calls.

Simulated Dataset

To benchmark Samovar, we used bamsurgeon (Ewing et al., 2015) to insert synthetic mosaic variants into the NA24385 10X Genomics Chromium BAM file from the Genome in a Bottle (GIAB) project (Zook et al., 2016). Training and testing occurred using sites on the autosomal chromosomes only since NA24385 is male, and the training used an independent set of synthetic variants from those used for the evaluation. The mean inferred linked-read length is 16,176 bp with standard deviation 54,387 bp. To evaluate performance at lower coverage and in other tools' tumor/normal "paired" mode, the original BAM file (mean coverage 61.8; median 60 at bamsurgeon-modified sites, excluding reads marked duplicate) was split in half based on read group tag and we subsequently modified only one-half with bamsurgeon (mean coverage 30.6, median 29 at bamsurgeon-modified sites). Splitting by read group tag ensures that an entire linked read will be placed into the derivative BAM file. Experiments with the original BAM file are referred to as "60X coverage" and those with the subsample as "30X coverage."

Samovar Model Comparison

To measure the specific advantage conferred by linked reads, we also implemented two reduced Samovar models that incorporate less of the variant phasing information. The "short-only" model redefines the fragment-level model features so that they use information summarized over the shorter, paired-end-level

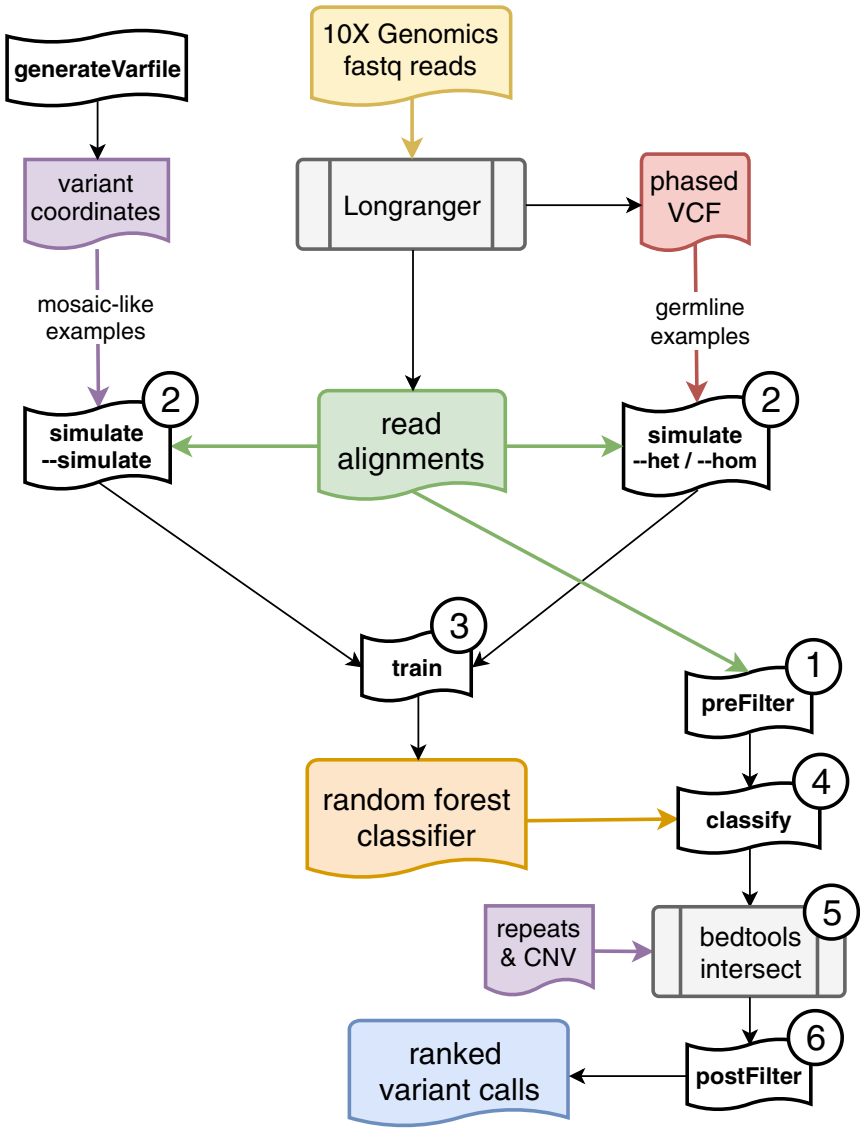


Figure 2. Samovar Workflow

fragments rather than the longer linked-read-level fragments. In this model, a paired-end read is assigned to a haplotype only if one of the ends overlaps a heterozygous variant phased by Long Ranger. Past work showed that even the phasing information from short fragments can improve mosaic variant calling accuracy (Usuyama et al., 2014). We find that, although the precision is comparable with that of the Samovar full model, the number of variant calls is much lower, resulting in a genome-wide recall of 2.0% at 30X and 60X, because there are few sites for which adequate phasing information can be compiled from short reads alone (Figure S4, Table S4).

We also created a "no-phasing" Samovar model that used no fragment phasing information at all. This was accomplished simply by omitting the fragment-level features from the model. When stratified by mosaic allele frequency (MAF), precision in every bin is near zero, although genome-wide recall is 68.3%, underscoring the importance of phasing features to our approach (Figure S4, Table S4).

MosaicHunter and MuTect2 Comparison

We compared Samovar with MosaicHunter v. 1.1 (Huang et al., 2017). We ran MosaicHunter in "tumor-only mode" analyzing only the bamsurgeon-mutated BAM file from NA24385, as well as in "trio mode" where

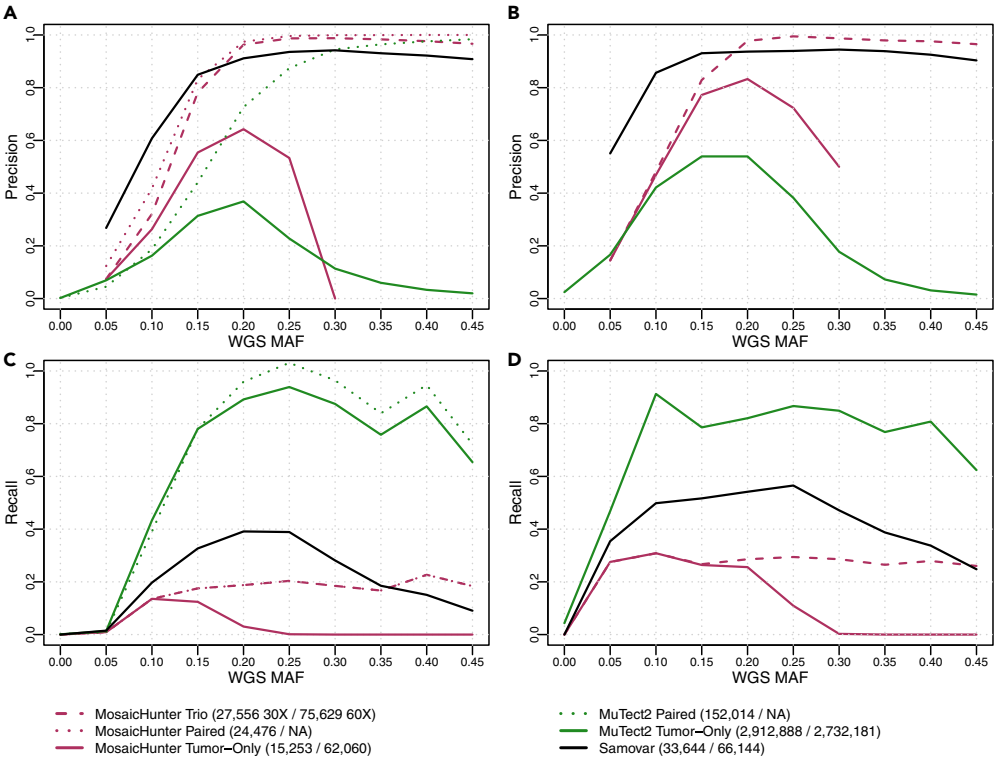


Figure 3. Precision and Recall Calculated for Samovar, MuTect2, and MosaicHunter Variant Calls Stratified by Mosaic Allele Fraction (MAF) in the Whole-Genome Sequencing Data (WGS)

(A–D) (A) 30X coverage, precision; (B) 60X coverage, precision; (C) 30X coverage, recall; (D) 60X coverage, recall.

the unaltered GIAB 10X Genomics Chromium BAM files from the mother (NA24143) and father (NA24149) were also provided. The parental BAM files were similarly produced by Long Ranger but not modified by bamsurgeon. Although Samovar does not use trio information, we hypothesized that its modeling of linked reads would allow it to have competitive accuracy. The modified and unmodified halves of the BAM file split by read group were provided when MosaicHunter was run in "paired-mode" as tumor and normal, respectively.

We also compared Samovar with MuTect2 from GATK v. 4.0.12.0 (Cibulskis et al., 2013). We ran MuTect2 in "tumor-only mode" and tumor/normal "paired-mode" on the same data described earlier. Tumor-only mode calls mosaic and germline mutations simultaneously but does not differentiate between the categories; hence the number of calls is much higher and the precision suffers at higher MAF where germline heterozygous variants comprise most of the call set.

Figure 3 shows each tool's precision and recall, stratified by MAF in the tumor WGS. Precision is calculated as the fraction of variant calls made that were bamsurgeon synthetic mutations, and recall is calculated as the fraction of bamsurgeon synthetic mutations that were in each tool's variant call set. Samovar achieves consistently higher precision than the tumor-only modes of MuTect2 and MosaicHunter. Importantly, Samovar's precision is also comparable with that of those tools in their trio and paired modes, with MosaicHunter's paired and trio modes achieving slightly higher precision at MAFs $\geq$ 0.2 and MuTect2's paired mode achieving higher precision at MAFs $\geq$ 0.3.

Note that in all cases, the original 10X Genomics BAM file was used. This means that all three Samovar models (as well as MuTect2 and MosaicHunter) benefited from the improved alignment accuracy of the linked-read-aware Lariat aligner, giving the short-only and no-phasing models and the other two methods a somewhat artificial advantage.

In addition to performance genome-wide we evaluated precision and recall (i.e., TPR) across different annotated genomic regions: genes, exons, all repeats, Alu repeats, segmental duplications, enhancers,

30X Coverage	Samovar			MuTect2						MosaicHunter								
				Tumor-Only			Paired			Tumor-Only			Paired			Trio		
	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*
Autosomes	84.0	30.1	44.4	3.0	83.2	5.7	60.8	91.4	73.0	31.5	5.1	8.8	79.2	20.7	32.8	70.4	20.7	32.0
Exons	84.0	28.3	42.4	3.6	85.3	7.0	60.1	92.0	72.7	35.0	7.1	11.8	82.1	30.8	44.8	73.7	30.8	43.4
Genes	84.9	30.1	44.4	3.2	84.4	6.2	63.0	92.0	74.8	32.6	5.7	9.7	79.9	22.7	35.4	71.2	22.7	34.5
Enhancer	88.5	31.0	45.9	3.9	86.7	7.5	72.9	92.3	81.4	37.8	5.9	10.1	85.5	29.5	43.8	80.2	29.5	43.1
Promoter	83.3	26.1	39.8	3.0	83.2	5.8	59.4	90.9	71.9	35.3	6.1	10.4	80.5	25.1	38.3	73.7	25.1	37.5
Alu	82.0	28.6	42.4	2.3	78.2	4.5	54.5	88.4	67.4	8.6	0.0	0.1	56.5	0.3	0.6	53.1	0.3	0.6
RepeatMasker	84.2	29.6	43.9	2.8	81.5	5.3	58.9	90.1	71.2	20.2	0.3	0.6	72.3	1.4	2.7	61.3	1.4	2.7
Seg. Dup.	25.6	10.4	14.8	1.3	56.9	2.5	18.4	62.8	28.5	6.6	0.5	0.9	39.3	1.7	3.2	29.1	1.7	3.2
60X coverage	Prec	Rec	*F*	Prec	Rec	*F*				Prec	Rec	*F*				Prec	Rec	*F*
Autosomes	84.6	43.0	57.1	3.6	76.0	7.0				32.4	15.5	20.9				46.8	27.2	34.4
Exons	84.3	41.8	55.9	4.7	79.6	8.8				38.5	25.3	30.5				54.0	45.5	49.4
Genes	85.6	43.4	57.6	3.9	77.2	7.5				33.1	17.0	22.4				47.7	30.0	36.8
Enhancer	90.8	47.8	62.6	4.8	77.9	9.0				36.9	22.7	28.1				51.6	40.0	45.1
Promoter	85.4	40.7	55.2	4.0	76.8	7.6				38.5	21.1	27.3				56.4	40.5	47.2
Alu	81.1	42.9	56.1	3.0	68.0	5.7				16.5	0.2	0.5				31.7	0.5	1.0
RepeatMasker	84.2	42.2	56.2	3.4	74.1	6.4				24.7	1.0	1.9				38.3	1.8	3.4
Seg. Dup.	28.0	13.1	17.8	1.6	48.5	3.1				9.8	1.5	2.6				18.5	2.7	4.7

Table 1. Precision (Prec), Recall (Rec), and F Score of Each Tool for the Synthetic Mosaic Variants Inserted by Bamsurgeon

and promoters listed in the UCSC Genome Browser and Ensembl, shown in Table 1. Recall is calculated as the fraction of bamsurgeon synthetic mutations with at least four mosaic allele reads that were in the variant call set since both Samovar and MosaicHunter require at least four reads to support a variant call. In practice, many tools including Samovar and MosaicHunter apply filters that exclude portions of the genome that lack sufficient evidence or that are inherently difficult to analyze, such as highly repetitive portions, which particularly contributes to MosaicHunter's poor performance in these genomic regions (see "Genomic regions and filters"). Furthermore, 66% of the Samovar false-negative sites over which recall was evaluated in the 30X coverage experiment and 38% of false negatives in the 60X experiment had fewer than four haplotype-discordant reads, which is the default requirement for Samovar. Relaxing this parameter can boost recall, although it may also impact precision.

Pediatric Cancer Dataset

We next studied a collection of 13 pediatric cancer cases that we sequenced—both tumor and normal—using 10X Genomics Chromium WGS and Whole-Exome Sequencing (WES). One of these cases was studied previously (Miller et al., 2018), and the other twelve are novel to this work. We ran Samovar, MosaicHunter (in both paired and tumor-only modes), and MuTect2 (in both paired and tumor-only modes) on each of the 13 tumor WGS datasets. When running MosaicHunter or MuTect2 in paired mode, we also provided the paired normal WGS.

To estimate accuracy of the different approaches, we used the WES sequencing as a validation dataset as it provides independent and deeper coverage over candidate variants within the exome. We first identified the calls from each tool within the exome capture region. The number and precision of the exome-coincident calls made by each tool are shown in Table 2.

Case	Samovar		MuTect2				MosaicHunter			
	Full Model		Tumor-Only		Paired		Tumor-Only		Paired	
	Calls	Prec	Calls	Prec	Calls	Prec	Calls	Prec	Calls	Prec
1	22	**0.71**	23,216	0.03	406	0.45	202	0.63	144	0.62
2	23	**0.75**	23,960	0.02	341	0.20	258	0.25	124	0.27
3	42	**0.74**	23,866	0.02	359	0.34	177	0.45	68	0.66
4	37	**0.72**	24,317	0.02	285	0.28	159	0.46	81	0.59
5	21	**0.91**	24,036	0.01	321	0.33	170	0.45	69	0.70
6	50	**0.95**	23,978	0.01	265	0.36	234	0.41	108	0.56
7	23	**0.80**	23,905	0.02	245	0.29	88	0.63	58	0.78
8	28	**0.74**	23,949	0.02	322	0.24	187	0.44	86	0.47
9	25	**0.62**	24,893	0.02	276	0.31	185	0.46	78	0.56
10	29	**0.53**	25,290	0.01	313	0.28	344	0.33	144	0.49
11	22	0.70	24,043	0.02	284	0.41	105	0.75	83	**0.80**
12	21	0.58	23,875	0.02	278	0.48	178	0.58	72	**0.81**
13	15	0.71	23,663	0.02	268	0.35	112	0.76	66	**0.80**
Total	358		312,991		3,963		2,399		1,181	

Table 2. Number of Variant Calls in the Exome Capture Regions and Precision (Prec) Based on Supporting Reads Found in WES

Samovar has the highest validation rate in 10 of the 13 cases. Bold indicates the highest precision for each pediatric case.

We then examined the corresponding WES tumor data for evidence of the mosaic call made in the WGS data. We considered a mosaic variant call to be "validated" if (1) the corresponding WES tumor sample had at least 50 aligned reads at the locus with at least 4 reads supporting the mosaic allele, and (2) the mosaic variant was not found to be germline by Long Ranger in both the tumor and normal WGS data from that patient. Figure 4 stratifies the validation rate by MAF in the WGS data, and Table 2 shows each tool's overall precision for the calls in the exome capture region. The bar graph shows the number of variants in each MAF bin. MosaicHunter paired called three times as many variants as Samovar, and MuTect2 paired called eleven times as many variants. This is because Samovar requires phasing-based evidence to make a call, which makes it more stringent, and because tumor/normal callers can identify variants that are homozygous or heterozygous in the tumor sample but have a different genotype compared with normal. Additionally, MuTect2 does not filter out CNV regions like MosaicHunter and Samovar, allowing it to call variants in a larger region of the genome. However, Samovar's validation rate is comparable with the paired callers across a range of MAF, indicated by the comparable precision of Samovar in Figure 4E compared with other tools' paired modes in a and c. Against tumor-only modes of other tools, Samovar has superior precision especially at MAF $\geq$ 0.15: MuTect2 tumor-only mode is not designed to differentiate heterozygous from high-MAF mosaic variants, and MosaicHunter makes few calls with a low validation rate.

As Samovar demonstrated high single-sample precision in simulation, comparable with the other tools' paired analysis, we are also able to run it on the normal control available for each of these cases. Sensitivity was measured in the same fashion using WES of the normal sample; across all 13 samples, 732 variants were in the exome capture region and the validation rate was 65% (see Table S9 for per-sample statistics). More mutations were found in normal samples because a larger fraction of the genome was excluded by CNVNATOR calls in tumor samples, as shown in Table S2. Interestingly, using ANNOVAR (Wang et al., 2010), we determined 11 of these mosaic mutations across 7 cases were nonsynonymous (amino-acid-changing) in one of the 299 cancer driver genes identified in Bailey et al., 2018. The extent of mosaicism in normal tissue and how this may relate to pediatric cancer are interesting avenues of future study now possible with Samovar.

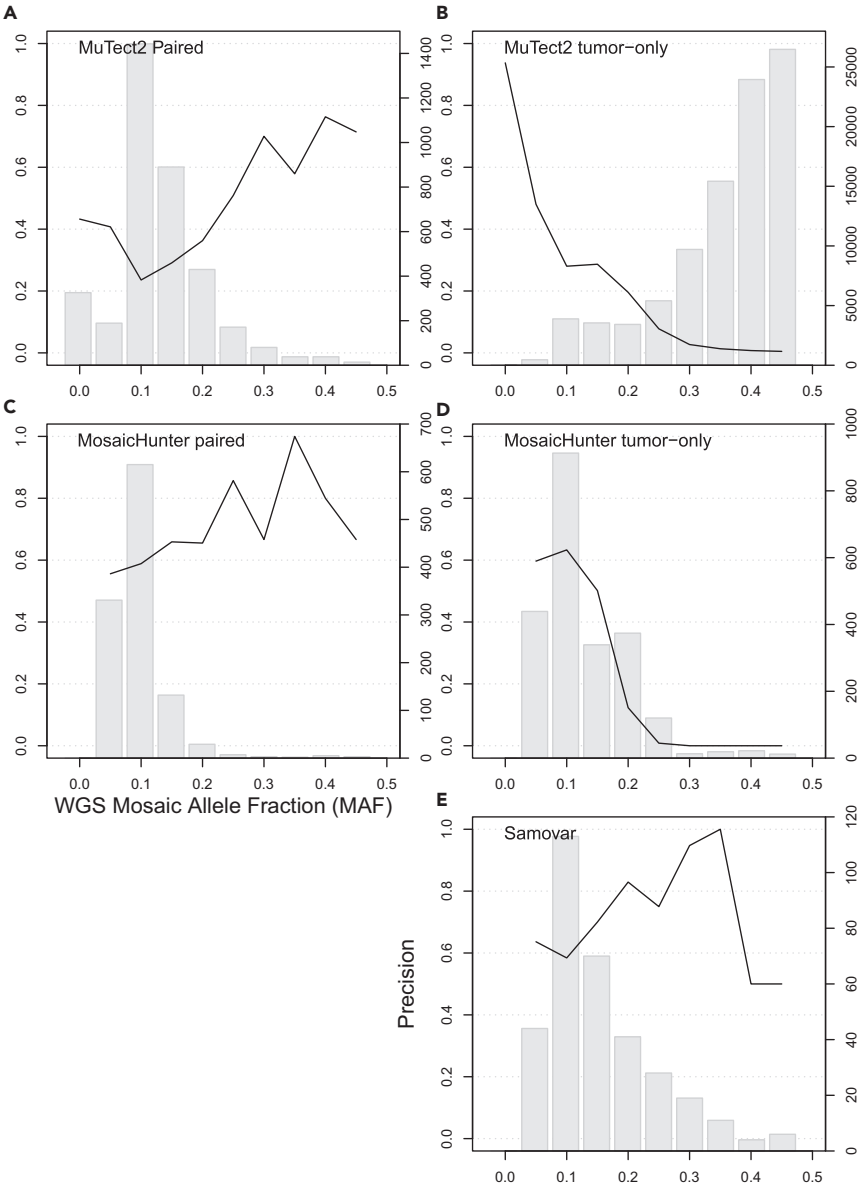


Figure 4. WES Support for Pediatric Cancer Somatic Variant Calls

Plots show fraction of variant calls in exome capture region supported by WES data (black line, left axis ticks) and number of variant calls (gray bars, right axis ticks) stratified by mosaic allele fraction (MAF), combined for the 13 pediatric cancer cases studied. The panels show results for (A) MuTect2 Paired, (B) MuTect2 tumor-only, (C) MosaicHunter paired, (D) MosaicHunter tumor-only, and (E) Samovar.

DISCUSSION

Genomic mosaicism is an important characteristic of many human diseases and conditions. Accurately identifying mosaic variants has previously relied on paired samples or trio analysis, which increases study costs and complexity of studies and may not be possible in many situations. By taking advantage of linked-read properties, particularly the ability to accurately assemble haplotypes, Samovar is able to call mosaic SNVs for a single sample at a level of precision that is comparable with that of paired and trio-based methods. Samovar also achieves substantially higher precision at low MAFs (<15%) and higher recall in more difficult-to-analyze portions of the genome such as segmental duplications and repetitive elements. This opens the door to a wider range of discoveries than are possible with current methods.

Although Samovar already compares favorably to tools that use matched-normal and trio data, in the future it will be important to investigate whether Samovar's recall and precision can be further improved by incorporating trio and matched-normal data directly into its model. Based on the results collected here, we expect that a key benefit of this would be to improve recall at all MAFs and to extend the high precision achieved by the existing paired- and trio-based methods into the low end of the MAF spectrum.

Limitations of Study

Samovar requires 10X Genomics linked-read data, which currently adds approximately 15% to the cost of a standard paired-end Illumina sequencing experiment. We demonstrate that limited phasing information is available from paired-end reads, but that experiment still used the haplotype phasing information of individual variants from the linked reads and could not be replicated from paired-end reads alone. Finally, although the Samovar model detects only SNPs, it could theoretically be extended to small indels that display the same pattern of haplotype-discordant reads. For this analysis additional indel-related features would also be needed to discriminate true indels from sequencing and alignment errors.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND SOFTWARE AVAILABILITY

The 10X Genomics linked-read whole-genome sequencing (WGS) and whole exome sequencing (WES) data described for the thirteen pediatric cancer cases are available within dbGaP under accession phs001820.v1.p1. The GIAB BAM files with simulated mutations are available at http://share.schatz-lab.org/samovar/simulation.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.05.037.

ACKNOWLEDGMENTS

This work was supported by the Nationwide Children's Hospital Foundation. It was also supported by NIH grants U01MH106884 to BL, R01GM118568 to BL, R01-HG006677 to MCS, and R21-CA220411 to MCS; and NSF grant DBI-1350041 to MCS. Part of this research project was conducted using computational resources at the Maryland Advanced Research Computing Center (MARCC).

AUTHOR CONTRIBUTIONS

Conceptualization, C.A.D., C.E.C., J.M.G.-F., R.K.W., E.R.M., P.W., B.L., and M.C.S.; Methodology, C.A.D., B.L., and M.C.S.; Investigation, C.A.D., J.R.F., P.J.B., B.J.K., N.B., V.M., and J.L.; Resources, J.L.; Data Curation, J.R.F., P.J.B., and B.J.K.; Writing - Original Draft, C.A.D., J.R.F., P.J.B., B.J.K., P.W., B.L., and M.C.S.; Writing - Review & Editing, C.A.D., P.W., B.L., and M.S.; Visualization, C.A.D.; Supervision, P.W., B.L., and M.S.; Funding Acquisition, B.L. and M.C.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 9, 2019
Revised: May 6, 2019
Accepted: May 24, 2019
Published: August 30, 2019

REFERENCES

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18.

Biesecker, L.G., and Spinner, N.B. (2013). A genomic view of mosaicism and human disease. Nat. Rev. Genet. *14*, 307–320.

Bishara, A., Liu, Y., Weng, Z., Kashef-Haghighi, D., Newburger, D.E., West, R., Sidow, A., and Batzoglou, S. (2015). Read clouds uncover variation in complex regions of the human genome. Genome Res. *25*, 1570–1580.

Bohrson, C.L., Barton, A.R., Lodato, M.A., Rodin, R.E., Luquette, L.J., Viswanadham, V.V., Gulhan, D.C., Cortés-Ciriano, I., Sherman, M.A., Kwon, M., et al. (2019). Linked-read analysis identifies

mutations in single-cell DNA-sequencing data. Nat. Genet. *2019*, 1.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Cohen, A.S.A., Wilson, S.L., Trinh, J., and Ye, X.C. (2015). Detecting somatic mosaicism: considerations and clinical implications. Clin. Genet. *87*, 554–562.

D'Gama, A.M., and Walsh, C.A. (2018). Somatic mosaicism and neurodevelopmental disease. Nat. Neurosci. *21*, 1504–1514.

Dou, Y., Gold, H.D., Luquette, L.J., and Park, P.J. (2018). Detecting somatic mutations in normal cells. Trends Genet. *34*, 545–557.

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res. *27*, 801–812.

Elyanow, R., Wu, H.T., and Raphael, B.J. (2018). Identifying structural variants using linked-read sequencing data. Bioinformatics *34*, 353–360.

Eslami Rasekh, M., Chiatante, G., Miroballo, M., Tang, J., Ventura, M., Amemiya, C.T., Eichler, E.E., Antonacci, F., and Alkan, C. (2017). Discovery of large genomic inversions using long range information. BMC Genomics *18*, 65.

Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y., et al. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic singlenucleotide- variant detection. Nat. Methods *12*, 623–630.

Fang, L., Kao, C., Gonzalez, M.V., da Silva, R.P., Li, M., Hakonarson, H., and Wang, K. (2018). LinkedSV: detection of mosaic structural variants from linked-read exome and genome sequencing data. bioRxiv, 409789, https://doi.org/10.1101/409789.

Freed, D., Stevens, E., and Pevsner, J. (2014). Somatic mosaicism in the human genome. Genes (Basel) *5*, 1064–1094.

Freed, D., and Pevsner, J. (2016). The contribution of mosaic variants to autism spectrum disorder. PLoS Genet. *12*, e1006245, https://doi.org/10.1371/journal.pgen.1006245.

Gajecka, M. (2016). Unrevealed mosaicism in the next-generation sequencing era. Mol. Genet. Genomics *291*, 513–530.

Huang, A.Y., Zhang, Z., Ye, A.Y., Dou, Y., Yan, L., Yang, X., Zhang, Y., and Wei, L. (2017). MosaicHunter: accurate detection of postzygotic single nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. Nucleic Acids Res. *45*, 1–10.

Jackman, S.D., Coombe, L., Chu, J., Warren, R.L., Vandervalk, B.P., Yeo, S., Xue, Z., Mohamadi, H., Bohlmann, J., Jones, S.J.M., et al. (2018). Tigmint: correcting assembly errors using linked reads from large molecules. BMC Bioinformatics *19*, 393.

Kennedy, S.R., Loeb, L.A., and Herr, A.J. (2012). Somatic mutations in aging, cancer and neurodegeneration. Mech. Ageing Dev. *133*, 118–126.

Kuleshov, V., Snyder, M.P., and Batzoglou, S. (2016). Genome assembly from synthetic long read clouds. Bioinformatics *32*, i216–i224.

Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R., McHugh, C.P., Ling, H., Hetrick, K.N., Pugh, E.W., Amos, C., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. *44*, 642–650.

Lindhurst, M.J., Sapp, J.C., Teer, J.K., Johnston, J.J., Finn, E.M., Peters, K., Turner, J., Cannons, J.L., Bick, D., Blakemore, L., et al. (2011). A mosaic activating mutation in AKT1 associated with the proteus syndrome. N. Engl. J. Med. *365*, 611–619.

Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2019). Resolving the full spectrum of human genome variation using linked-reads. Genome Res. *29*, 635–645.

McConnell, M.J., Moran, J.V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J.A., Fasching, L., Flasch, D.A., Freed, D., et al. (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: the Brain Somatic Mosaicism Network. Science *356*, https://doi.org/10.1126/science.aal1641.

Miller, K.E., Kelly, B., Fitch, J., Ross, N., Avenarius, M.R., Varga, E., Koboldt, D.C., Boué, D.R., Magrini, V., Coven, S.L., et al. (2018). Genome sequencing identifies somatic BRAF duplication c.1794_1796dupTAC;p.Thr599dup in pediatric patient with low-grade ganglioglioma. Cold Spring Harb. Mol. Case Stud. *4*, https://doi.org/10.1101/mcs.a002618.

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E.T., Hastie, A.R., Marks, P., Lee, J., Chu, C., Lin, C., Džakula, Ž., et al. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. Nat. Methods *13*, 587–590.

Ouwens, K.G., Jansen, R., Tolhuis, B., Slagboom, P.E., Penninx, B.W.J.H., and Boomsma, D.I. (2018). A characterization of postzygotic mutations identified in monozygotic twins. Hum. Mutat. *39*, 1393–1401.

Poduri, A., Evrony, G.D., Cai, X., and Walsh, C.A. (2013). Somatic mutation, genomic variation, and neurological disease. Science *341*, 1237758.

Sedlazeck, F.J., Lee, H., Darby, C.A., and Schatz, M.C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. *19*, 329–346.

Shajii, A., Numanagić, I., Whelan, C., and Berger, B. (2018). Statistical binning for barcoded reads improves downstream analyses. Cell Syst. *7*, 219–226.e5.

Shirley, M.D., Tang, H., Gallione, C.J., Baugher, J.D., Frelin, L., Cohen, B., North, P.E., Marchuk, D.A., Comi, A.M., and Pevsner, J. (2013). Sturge–Weber syndrome and port-wine stains caused by somatic mutation in *GNAQ*. N. Engl. J. Med. *368*, 1971–1979.

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J.M., Salit, M., West, R.B., Batzoglou, S., and Sidow, A. (2017). Genome-wide reconstruction of complex structural variants using read clouds. bioRxiv *14*, 915–920.

Usuyama, N., Shiraishi, Y., Sato, Y., Kume, H., Homma, Y., Ogawa, S., Miyano, S., and Imoto, S. (2014). HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. Bioinformatics *30*, 3302–3309.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K.B., Pao, W., and Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. *5*, 91.

Watson, I.R., Takahashi, K., Futreal, P.A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. Nat. Rev. Genet. *14*, 703–718.

Weinstein, L.S., Shenker, A., Gejman, P.V., Merino, M.J., Friedman, E., and Spiegel, A.M. (1991). Activating mutations of the stimulatory G protein in the McCune–Albright syndrome. N. Engl. J. Med. *325*, 1688–1695.

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. Genome Res. *27*, 757–767.

Xia, L.C., Bell, J.M., Wood-Bouwens, C., Chen, J.J., Zhang, N.R., and Ji, H.P. (2018). Identification of large rearrangements in cancer genomes with barcode linked reads. Nucleic Acids Res. *46*, e19.

Youssoufian, H., and Pyeritz, R.E. (2002). Mechanisms and consequences of somatic mosaicism in humans. Nat. Rev. Genet. *3*, 748–758.

Zheng, G.X., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linkedread sequencing. Nat. Biotechnol. *34*, 303–311.

Zhou, X., Batzoglou, S., Sidow, A., and Zhang, L. (2018). HAPDeNovo: a haplotype-based approach for filtering and phasing de novo mutations in linked read sequencing data. BMC Genomics *19*, 467.

Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025.

ISCI, Volume *18*

Supplemental Information

Samovar: Single-Sample Mosaic Single-Nucleotide Variant Calling with Linked Reads

Charlotte A. Darby, James R. Fitch, Patrick J. Brennan, Benjamin J. Kelly, Natalie Bir, Vincent Magrini, Jeffrey Leonard, Catherine E. Cottrell, Julie M. Gastier-Foster, Richard K. Wilson, Elaine R. Mardis, Peter White, Ben Langmead, and Michael C. Schatz

Transparent Methods

1 Samovar pipeline

Samovar is implemented in Python 3 and operates on the alignment (BAM) and variant (VCF) files produced by 10x Genomics' Long Ranger pipeline. See "Samovar requirements" below for software dependency and input file requirements.

(1) preFilter Samovar first scans the genome calculating the features listed in Figure S8 at each site. Each feature has a numerical threshold, and if all filters are passed the site is considered in step 4 (classify) as a candidate variant site. These filters examine measurements such as depth, number of haplotype-discordant reads, quality of the alignments and credibility of the read phasing.

(2) simulate Simulated mosaic training examples are generated at regular intervals across the genome at a range of mosaic allele frequency (MAF) from 0.025 to 0.475 at increments of 0.025. Such sites are called "simulation sites." Sites harboring germline variant calls can be excluded by specifying them in a VCF. For each phased alignment having the reference allele at the simulation site, the reference allele is randomly changed to the mosaic base with probability equal to the target MAF. For an unphased alignment having the reference allele, the reference allele is randomly changed to the mosaic base with probability $\frac{\text{MAF}}{2}$, on the principle that unphased reads are equally likely to originate from either haplotype. The features listed in Figure S5 are computed for the simulation sites to obtain true-mosaic training examples. The same features are computed for FILTER=PASS phased heterozygous (GT=0|1 or GT=1|0) and homozygous (GT=1|1 or GT= 0|0) variant sites from the VCF to get true-non-mosaic examples.

(3) train A random forest model is trained with an equal number of simulation sites and non-mosaic sites. Non-mosaic sites are selected to have equal amounts of heterozygous and homozygous calls in the VCF. We use the RandomForestClassifier module from the scikit-learn library (Pedregosa et al. 2011) with `max_leaf_nodes 50` and `n_estimators 100`, though Samovar allows the user to customize these hyperparameters. The random forest features described in Table S5 take into account the abundance and consistency of evidence for a mosaic variant, including the number of haplotype discordant reads, mosaic allele fraction, base quality, alignment score, amount of soft clipping, presence of indels, etc.

After cross-validation at a variety of sequencing depths (Table S1), we found that using 20,000 mosaic, 10,000 heterozygous and 10,000 homozygous training examples achieved a balance of computational efficiency and accuracy. We subsampled the NA24385 BAM file used for the simulation experiment and ran the Samovar simulate and train steps. For each number of training examples, average performance statistics are reported for ten independent train/validation splits; 0.5 and 0.9 refer to the random forest probability that the example is in the mosaic class.

(4) classify Genomic sites passing the preFilter are classified by the trained random forest model, yielding the predicted probability that the site is mosaic. Sites with probability above a cutoff are reported in BED format. Based on cross-validation at a variety of sequencing depths, we found that a probability cutoff of 0.5 balances false positive rate and true positive rate (Table S1), though this can be adjusted to trade between sensitivity and precision.

(5) region-based filter As Illumina sequencing is known to have high error rates within microsatellites and simple repeat sequences (Fang et al. 2014), we exclude candidate mosaic variants identified in these regions. Specifically, we exclude variants within +/- 2bp from 1,2,3,4-bp repeats at least 4bp long with at least 3 copies of the unit. Within hg19, 72.0% of autosomes and 71.4% of autosomes+X+Y will remain after this region filter, and within GRCh38 73.8% of autosomes

and 73.1% of autosomes+X+Y remain. We also exclude any CNV regions +/- 5bp identified by CNVNATOR (Abyzov et al. 2011) because polymorphism among the copies of a repeated region would be misconstrued as mosaicism.

(6) postFilter Our expectation is that mosaic variants are isolated events. Samovar applies a final test to distinguish an isolated, likely mosaic variant from the situation where there are many nearby variants co-occurring on the same reads. The latter pattern is usually caused by alignment errors in the presence of repetitive DNA and copy number variation. Specifically, we examine each base within a fixed distance of the mutative mosaic locus. At each base we conduct a Fisher's exact test, testing if the alleles observed at the query base associate with the haplotype-discordant reads. This is diagrammed in Figure S2. If the most significant p-value among all the statistical tests is less than the threshold, the site is filtered out. Based on simulations, we find that the p-value threshold can be set to 0.005 (default) or lower based on the desired balance between precision and recall. There is an option to avoid particular sites when calculating the minimum p-value among all nearby sites and it is recommended to use the germline VCF of variant calls here.

The final mosaic variant calls are reported in VCF format. VCF INFO tags are used to record depth, allele frequency, fraction of reads phased by Long Ranger, number of haplotype-discordant reads, the model-predicted probability, and the minimal p-value obtained by the postFilter.

2 Simulated dataset

Input data We downloaded the 10x Genomics Chromium datasets for the A/J trio processed with Long Ranger version 2.1 and GRCh38 from the GIAB project: FTP Link We use the BAM file from sequencing the son's genome (NA24385) as the basis for this simulation experiment, but MosaicHunter uses the BAM files for the mother (NA24143) and father (NA24149) in trio mode.

We use a custom fork of bamsurgeon (Ewing et al. 2015) to edit the reads in the BAM file. Given a target MAF, a 2 $\times$ MAF fraction of reads with tag HP=1, and a MAF fraction of reads with no HP tag are selected to mutate. The alternate allele is chosen randomly among the three non-reference bases.

Simulated mosaic mutations were introduced at evenly spaced intervals every 20,000 bp on the autosomes with target MAF between 0.025 and 0.475 in increments of 0.025. Reads were realigned with BWA-MEM after mutations were introduced. To compute precision, the denominator is sites with at least 4 alt-allele reads and 16 total reads (not marked duplicate or QC fail). This is because the parameters we chose for Samovar and MosaicHunter require at least 4 reads to call a mosaic variant, and Samovar's depth filter threshold is 16 (MosaicHunter's minimum depth is 25, which we keep, so technically fewer sites are visible to MosaicHunter).

Samovar We use 20,000 simulated mosaic, 10,000 heterozygous and 10,000 homozygous training examples to train each random forest model described. Table S3 has the feature importances of the Samovar model, with abbreviation and number as in Figure S5.

Samovar Short-read phasing model Samovar is designed to take advantage of the long-range phasing information given by linked reads. Previous methods similarly took advantage of the shorter-range phasing information given by paired-end sequencing. We can simulate the paired-end strategy in Samovar, allowing us to compare to the linked-read strategy while holding the rest of the pipeline constant. We begin by creating a "short-read phasing" Samovar model that breaks down the linked reads into their constituent paired-end reads and considers only these shorter fragments when compiling linked-read-related features such as haplotype-discordant reads.

Supposing that we have the complete haplotype phasing from Long Ranger, we assign a haplotype to a pair of reads if either mate overlaps at least SNP with a phased genotype in the VCF. Out of 1.91 billion reads, 9.76% of reads could be phased. Only 0.006% of reads overlapped variants but had alleles for conflicting haplotypes – these were not phased. Table S6 has the feature importances of this limited model, with abbreviation and number as in Figure S6.

Samovar No-phasing model While we do not advocate this approach, for the purposes of comparison, we remove all phasing-related features from Samovar to create a "no-phasing" model. Table S7 has the feature importances of this limited no-phasing model, with abbreviation and number as in Figure S7. Filters use the default parameters described in the preFilter feature list (Figure S8).

MosaicHunter Version 1.1. We used the default recommended parameters when possible, except we did not use the misaligned_reads_filter because it was extremely slow. In addition, because we have simulated far more mosaic sites than would be expected in a normal genome, we do not want to penalize MosaicHunter because it deliberately filters mosaic sites that are close to each other so we changed the following parameters:

- `clustered_filter.inner_distance=2000` [default 20000]
- `clustered_filter.outer_distance=2000` [default 20000]

We also adjusted MosaicHunter's supporting read threshold since Samovar requires at least 4 minor (mosaic) allele reads using: `base_number_filter.min_minor_allele_number=4` [default 3]

We used liftOver to transfer the provided `WGS.error_prone.b37.bed` and `all_repeats.b37.bed` to GRCh38 coordinates, and downloaded `dbsnp_human_9606_b150_GRCh38p7` bed files for the `common_site_filter`, `repetitive_region_filter`, `mosaic_filter.dbsnp_file` respectively. CNVNATOR was used to predict regions of copy number variation and this BED file was provided as the `indel_region_filter.bed_file` parameter.

Note that the `homopolymers_filter`, `common_site` and `repetitive_region` BED files leave visible only 32.2% of bases in the GRCh38 autosomes (34.4% including X and Y) to call mosaic variants. For comparison, Samovar considers about 73% of GRCh38 visible.

MuTect2 Version 4.0.12.0. We executed the standard GATK workflow of the Mutect2 program followed by FilterMutectCalls.

Genomic feature analysis knownGene, knownGene exons, RepeatMasker, RepeatMasker Alu, Segmental duplications are from UCSC Table Browser (GRCh38, accessed 10/02/18). Ensembl Enhancer, Ensembl Promoter + flanking are from Release 94. Ensembl FTP Site

3 Pediatric cancer dataset

Genomic DNA samples Peripheral blood and paired tumor samples were obtained from patients enrolled onto the "Nationwide Children's Neuro-Oncology Tumor and Epilepsy Tissue Bank" protocol (IRB16-00777) at Nationwide Children's Hospital. 13 cases with paired blood and tumor derived DNA were extracted following the manufacturers recommendation using the AllPrep Kit for tumors (Qiagen) and Gentra Purgene or QIAamp Kit (Qiagen) for blood samples. Genomic DNA was quantified with the Qubit dsDNA HS Assay Kit (Life Technologies) and diluted to approximately 1 ng/μL final concentration. DNA source and input mass into sample preparation is described in Supplementary File 1.

Sample preparation and sequencing Linked-read whole genome sequencing (WGS) and whole exome sequencing (WES) libraries were generated (Weisenfeld et al. 2017). Partitioning and barcoding high molecular weight (HMW) DNA was performed using a Chromium Controller Instrument (10x Genomics, CA), and Illumina sequencing libraries were prepared following protocols described in the manufacturer's user guide (Chromium Genome Reagent Kits v2 - Rev A). For WES, 250 ng of each 10x linked-reads library was hybridized in pools (see Supplementary File 1) with 3 pmol of the xGEN Exome Research Panel v1.0 (Integrated DNA Technologies, Coralville, IA) per the manufacturers protocol. Post WES enrichment used standard Illumina P5 and P7 primers (Griffith et al. 2015), and PCR cycling is highlighted in Supplementary File 1. Final libraries were quantified by qPCR (KAPA Biosystems Library Quantification Kit for Illumina platforms), diluted to 3 nM and sequenced using a paired-end recipe on the Illumina HiSeq 4000 next-generation sequencing instrument.

Bioinformatic Analysis Cases using reference genome GRCh38 2.1.0 (1, 2, 7, 10, 11) were processed with Long Ranger 2.1.6 and GATK HaplotypeCaller 3.8-0. Samples using reference genome b37 2.1.0 (3, 4, 5, 6, 8, 9, 10, 12) were processed with Long Ranger 2.1.3 and GATK HaplotypeCaller 3.5-0. The sequencing coverage and fraction of the genome identified by the CNVNATOR (Abyzov et al. 2011) calls is recorded in Table S2, and the oncology diagnosis of each case in Table S8.

4 Computational efficiency

We report timing results for the 30X GIAB sample. Samovar completed in 7 hours with 48 parallel threads for the "filter" step and up to 4 parallel threads for other steps. MuTect2 paired mode completed in 136 hours with 48 parallel threads. MosaicHunter tumor-only and trio modes completed in 29 hours each and paired mode completed in 7 hours. Note MosaicHunter does not offer parallelism options. (See "Command line arguments" for details.)

5 Samovar requirements

Samovar is implemented in Python 3 (also compatible with Python 2). It uses several libraries, including pyfaidx, scikit-learn, simplesam, and fisher. As input, Samovar requires the alignment (BAM) and variant (VCF) files produced by 10x Genomics' Long Ranger pipeline. Long Ranger processes the raw Illumina reads and performs linked read-aware alignment with Lariat (Bishara et al. 2015), small variant calling with Freebayes (Garrison and Marth 2012) or GATK (DePristo et al. 2011), structural variant calling, and haplotype assembly. Specifically, Samovar requires that the BAM have the HP (molecule haplotype), AS (Lariat best alignment score), and XS (Lariat second-best alignment score) extra fields, and requires that the VCF have the FILTER column and GT field. Information on the BAM file tags and phased VCF file are available at 10x Genomics.

6 Computational performance

Running time for each tool was listed for the 30X simulation experiment described in the text. Samovar was run on a single machine with 48 cores for the "filter" step and 4 cores for other parallelizable steps, with pypy when possible. Maximum memory usage was 19.2 GB, and the filter step reported 4200% CPU usage when allocated 48 cores. MosaicHunter and MuTect2 were run on a cluster in a scatter-gather format where each chromosome was computed independently and the results were merged. MosaicHunter does not offer parallelism options, although slightly greater

than 100% average CPU usage was seen. On chromosome 1, paired mode used maximum 25.6 GB memory; tumor-only mode used 25.3 GB; trio mode used 25.0 GB. MuTect2 was run with 48 cores for the "native pair HMM," although only 600% CPU usage was seen on average. On chromosome 1, paired mode used maximum 5.7 GB memory; tumor-only mode used 5.5 GB.

7 Command line arguments

MosaicHunter Version 1.1
The tumor-only, paired, and trio configuration file templates provided with the software distribution were used, containing default parameters.

```
java -jar mosaichunter.jar -C [configuration file] -P output_dir=[output directory]
```

MuTect2 - Paired Mode Version 4.0.12.0

```
gatk Mutect2 -R [reference genome] -I [tumor BAM file] -tumor [tumor sample name] -I
[normal BAM file] -normal [normal sample name] -O [MuTect2 VCF file]
--native-pair-hmm-threads 48
gatk FilterMutectCalls -V [MuTect2 VCF file] -O [Filtered VCF file]
grep -v "multiallelic" [Filtered VCF file] | grep -v "0/1/2" | vcftools --vcf -
--remove-indels --remove-filtered-all --recode --recode-INFO-all --out [Final MuTect2
VCF file]
```

MuTect2 - Tumor-only mode Version 4.0.12.0

```
gatk Mutect2 -R [reference genome] -I [tumor BAM file] -tumor [tumor sample name] -O
[MuTect2 VCF file] --native-pair-hmm-threads 48
gatk FilterMutectCalls -V [MuTect2 VCF file] -O [Filtered VCF file]
grep -v "multiallelic" [Filtered VCF file] | grep -v "0/1/2" | vcftools --vcf
- --remove-indels --remove-filtered-all --recode --recode-INFO-all --out [Final MuTect2
VCF file]
```

Samovar

```
samovar generateVarfile --out out.varfile --vcf [Sample VCF] --fai [Reference genome
FAI]
samovar simulate --bam [Sample BAM] --varfile [Sample VCF] --het --max 15000 --nproc
4 > het.features.tsv
samovar simulate --bam [Sample BAM] --varfile [Sample VCF] --hom --max 15000 --nproc
4 > hom.features.tsv
samovar simulate --bam [Sample BAM] --varfile out.varfile --simulate --nproc 4
> mosaic.features.tsv
samovar train --mosaic mosaic.features.tsv --het het.features.tsv --hom hom.features.tsv
samovar preFilter --bam [Sample BAM --nproc 48 > vectors.txt 2> intervalsComplete.txt
samovar classify --clf clf.pkl --vectors vectors.txt > predictions.tsv
bedtools intersect -v -a predictions.tsv -b [Samovar repeat BED file] | bedtools
intersect -v -a stdin -b [CNVNATOR BED file] > regionfiltered.tsv
samovar postFilter --bam [Sample BAM] --bed regionfiltered.tsv --ref [Reference genome]
--vcfavoid [Sample VCF] --nproc 4 --p 0.005 > samovar.vcf
```

8 Genomic regions and filters

In Table 1 and Figure 3 Samovar and MosaicHunter use their respective default filters but we have treated the tools as though they are interrogating roughly the same portion of the genome. Table S5 and Figure S3 attempt to normalize the differences by reporting just those sites that pass both tools' filters. In GRCh38, this is 32.8% of the autosomal sequence, containing MosaicHunter's simple sequence repeat filter and repetitive region bed files, and Samovar's simple sequence repeat filter, as well as any CNV regions identified by CNVNATOR.

References

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Fang, H., Y. Wu, G. Narzisi, J. A. O'Rawe, L. T. Barron, J. Rosenbaum, M. Ronemus, I. Iossifov, M. C. Schatz, and G. J. Lyon (2014). "Reducing INDEL calling errors in whole genome and exome sequencing data". In: *Genome Med* 6.10, p. 89.

Abyzov, Alexej, Alexander E Urban, Michael Snyder, and Mark Gerstein (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." In: *Genome Research* 21.6, pp. 974–84. DOI: 10.1101/gr.114876.110.

Ewing, Adam D, Kathleen E Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, J Christopher Bare, Christine P'ng, Daryl Waggott, Veronica Y Sabelnykova, et al. (2015). "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection". In: *Nature Methods* 12.7, pp. 623–630. DOI: 10.1038/nmeth.3407.

Weisenfeld, Neil I, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe (2017). "Direct determination of diploid genome sequences." In: *Genome Research* 27.5, pp. 757–767. DOI: 10.1101/gr.214874.116.

Griffith, Malachi, Christopher A. Miller, Obi L. Griffith, Kilannin Krysiak, Zachary L. Skidmore, Avinash Ramu, Jason R. Walker, Ha X. Dang, Lee Trani, David E. Larson, et al. (2015). "Optimizing Cancer Genome Sequencing and Analysis". In: *Cell Systems* 1.3, pp. 210–223. DOI: 10.1016/j.cels.2015.08.015.

Bishara, Alex, Yuling Liu, Ziming Weng, Dorna Kashef-Haghighi, Daniel E Newburger, Robert West, Arend Sidow, and Serafim Batzoglou (2015). "Read clouds uncover variation in complex regions of the human genome." In: *Genome Research* 25.10, pp. 1570–80. DOI: 10.1101/gr.191189.115.

Garrison, Erik and Gabor Marth (2012). *Haplotype-based variant detection from short-read sequencing*. eprint: arXiv:1207.3907.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, et al. (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nat. Genet.* 43.5, pp. 491–498.

Supplementary Figures and Tables

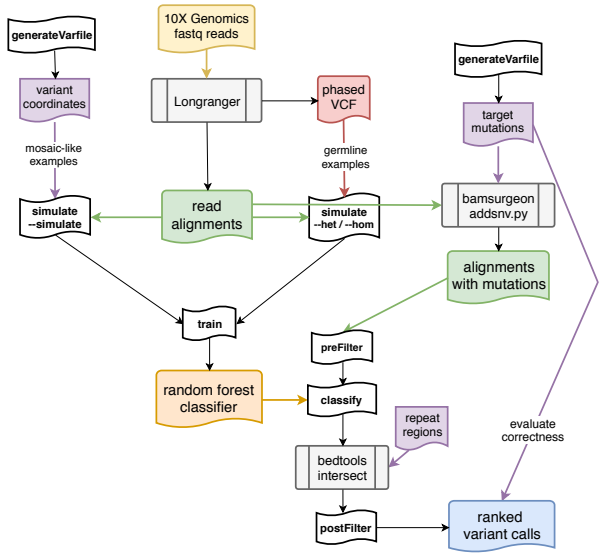
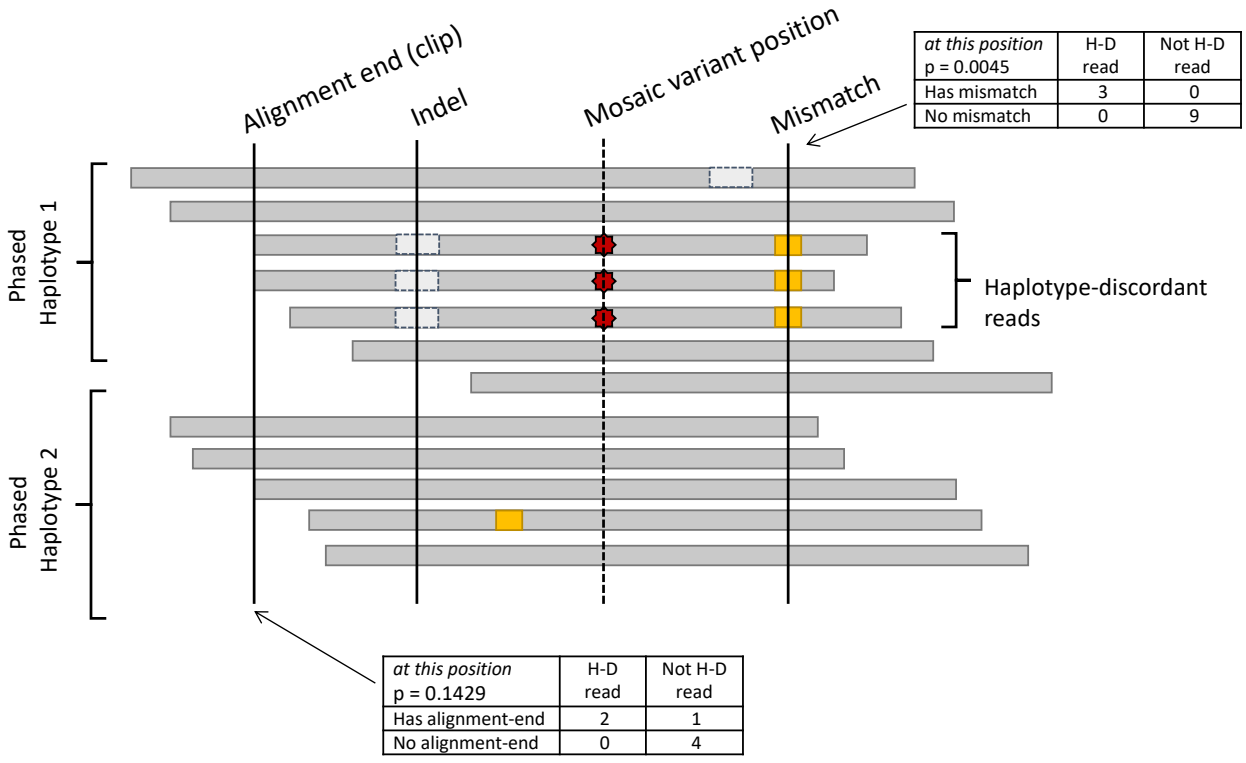


Figure S1: Simulation experiment workflow (left) additionally evaluates correctness of the calls based on mutations generated with bamsurgeon. Related to Figure 2.



at this position p = 0.0045	H-D read	Not H-D read
Has mismatch	3	0
No mismatch	0	9

at this position p = 0.1429	H-D read	Not H-D read
Has alignment-end	2	1
No alignment-end	0	4

Figure S2: The postFilter step calculates statistical association between haplotype-discordant reads and alignment features such as start/end position, indel or mismatch. Related to Figure 2.

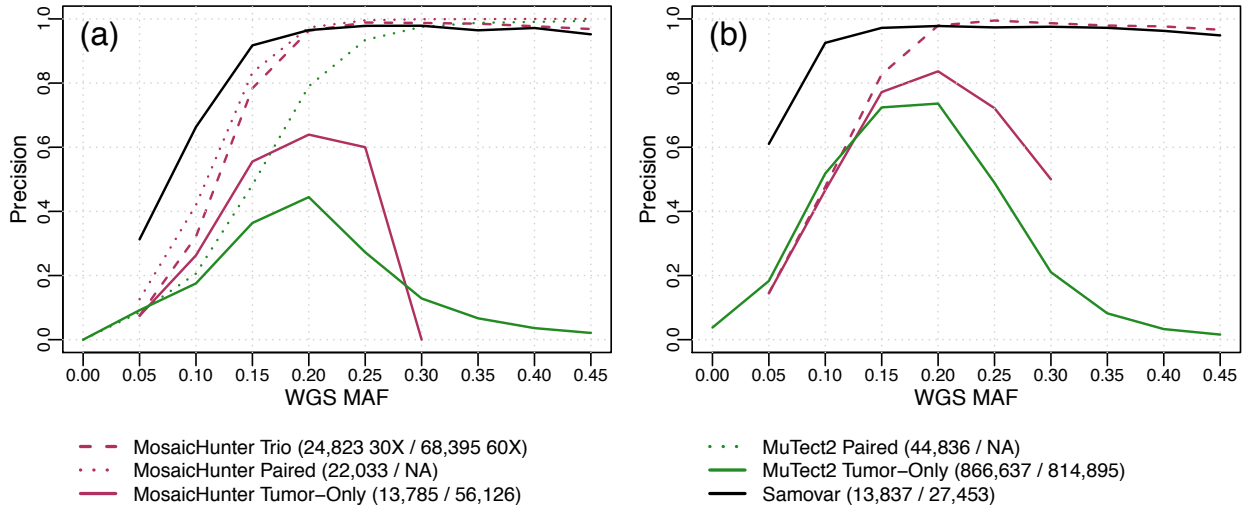


Figure S3: *Precision calculated in the genomic region not filtered by MosaicHunter or Samovar's region filters, calculated for Samovar, MuTect2, and MosaicHunter variant calls stratified by mosaic allele fraction (MAF) in whole genome sequencing data (WGS). (a) 30X coverage (b) 60X coverage. Related to Figure 3.*

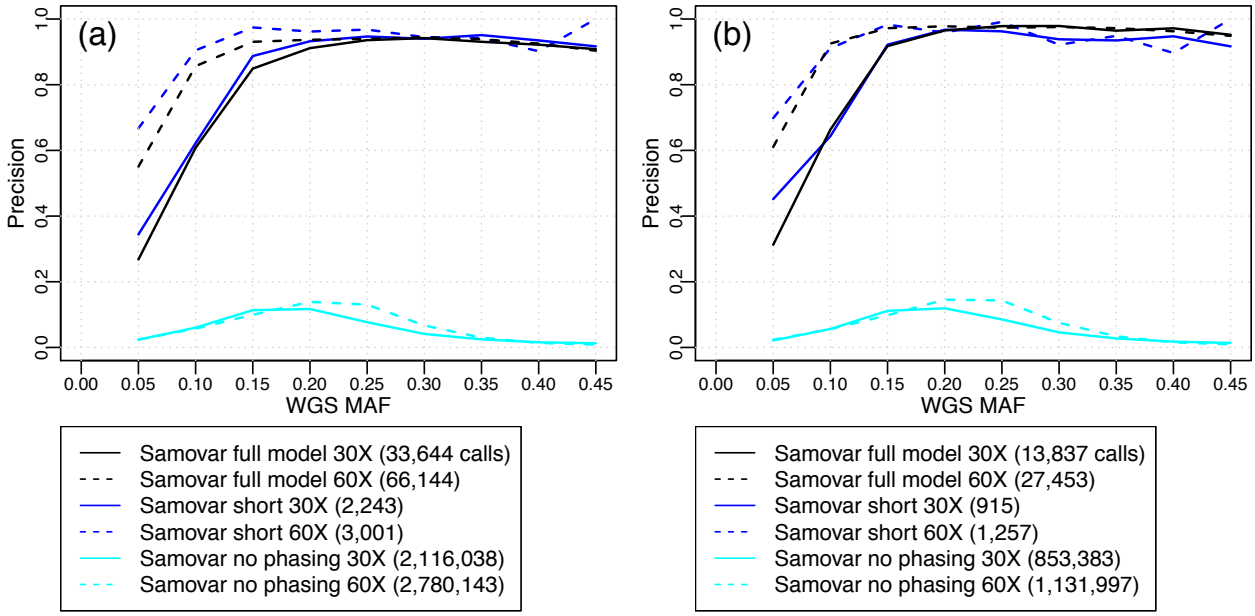


Figure S4: *Precision calculated for variant calls made by Samovar's full model and the "short-only" and "no-phasing" models created for illustration, stratified by mosaic allele fraction (MAF) in whole genome sequencing data (WGS). (a) Autosomes (b) Genomic region not filtered by MosaicHunter or Samovar's region filters. Related to Figure 3.*

1. Depth [excluding marked duplicates, QC fail, secondary and supplementary alignments]
2. Fraction of reads phased [HP tag assigned by Long Ranger]
3. Fraction of reads on the more common haplotype [max(number of HP=1 reads, number of HP=2 reads)]
4. MAF
5. MAF of phased reads
6. Number of haplotype-discordant (HD) reads
7. Fraction of phased reads that are HD
8. Fraction of HD reads on the more common haplotype [max(number of HP=1 HD reads, number of HP=2 HD reads)]
9. MAF of HD reads
10. Average base quality of HD reads
11. Average position from the closer end of the alignment on HD reads of the site being classified
12. Average number of soft-clipped bases on HD reads
13. Average number of indels in alignment of HD reads
14. Average value of AS – XS (Lariat alignment scores) of HD reads
15-21. Features 8-14 for the set of phased reads that are not HD
22-26. Features 10-14 for the set of mosaic-allele reads
27-31. Features 10-14 for the set of reference-allele reads
32. "weighted" HD read base quality: sum of HD read base quality / sum of all phased reads base quality
33. "weighted" mosaic-allele read base quality: sum of mosaic-allele read base quality / sum of reference- and mosaic-allele read base quality

Figure S5: Samovar random forest features. Related to Figure 2.

1. Depth [excluding marked duplicates, QC fail, secondary and supplementary alignments]
2. Fraction of reads phased [computed based on read or its mate overlapping phased variants]
3. Fraction of reads on the more common haplotype [max(number of HP=1 reads, number of HP=2 reads)]
4. MAF
5. MAF of phased reads
6. Number of haplotype-discordant [HD] reads
7. Fraction of phased reads that are HD
8. Fraction of HD reads on the more common haplotype [max(number of HP=1 HD reads, number of HP=2 HD reads)]
9. MAF of HD reads
10. Average base quality of HD reads
11. Average position from the closer end of the alignment on HD reads of the site being classified
12. Average number of soft-clipped bases on HD reads
13. Average number of indels in alignment of HD reads
14-19. Features 8-13 for the set of phased reads that are not HD
20-23. Features 10-13 for the set of mosaic-allele reads
24-27. Features 10-13 for the set of reference-allele reads
28. "weighted" HD read base quality: sum of HD read base quality / sum of all phased reads base quality
29. "weighted" mosaic-allele read base quality: sum of mosaic-allele read base quality / sum of reference- and mosaic-allele read base quality

Figure S6: Random forest features used in the "short-only" model. Related to Figure 2.

1. Depth [excluding marked duplicates, QC fail, secondary and supplementary alignments]
2. MAF
3. Average base quality of mosaic-allele reads
4. Average position from the closer end of the alignment on mosaic-allele reads of the site being classified
5. Average number of soft-clipped bases on mosaic-allele reads
6. Average number of indels in alignment of mosaic-allele reads
7-10. Features 3-6 for the set of reference-allele reads
11. "weighted" mosaic-allele read base quality: sum of mosaic-allele read base quality / sum of reference- and mosaic-allele read base quality

Figure S7: Random forest features used in the "no-phasing" model. Related to Figure 2.

Median depth of mosaic sites							
13	# training examples	**1000**	**2000**	**5000**	**10000**	**20000**	
(mindepth = 14)	Mosaic >0.5	0.9011	0.9013	0.90304	0.90524	0.90627	
	Mosaic >0.9	0.75662	0.7672	0.77045	0.77511	0.77566	
	Het <0.5	0.94845	0.9491	0.94826	0.94843	0.94747	
	Het <0.9	0.98975	0.9895	0.98991	0.99036	0.9898	
	Hom <0.5	0.95197	0.9553	0.95669	0.95762	0.95648	
	Hom <0.9	0.99244	0.9917	0.99265	0.99315	0.99284	
25	# training examples	**1000**	**2000**	**5000**	**10000**	**20000**	**30000**
	Mosaic >0.5	0.92024	0.9185	0.92068	0.92388	0.92355	0.92352
	Mosaic >0.9	0.80002	0.8101	0.81813	0.81823	0.81985	0.81635
	Het <0.5	0.9598	0.9608	0.96178	0.96032	0.95997	0.96035
	Het <0.9	0.99254	0.9912	0.99108	0.99168	0.99144	0.99174
	Hom <0.5	0.95935	0.9636	0.96356	0.96434	0.96431	0.96501
	Hom <0.9	0.99462	0.9938	0.99384	0.99436	0.99528	0.99523
37	# training examples	**1000**	**2000**	**5000**	**10000**	**20000**	**30000**
	Mosaic >0.5	0.93072	0.933	0.93332	0.93238	0.93283	0.93262
	Mosaic >0.9	0.82435	0.8366	0.84427	0.84729	0.84162	0.84557
	Het <0.5	0.96298	0.9658	0.96576	0.96782	0.96777	0.96736
	Het <0.9	0.99293	0.9924	0.99243	0.99187	0.99252	0.9924
	Hom <0.5	0.96414	0.9685	0.96835	0.97036	0.96997	0.97156
	Hom <0.9	0.99545	0.9951	0.99521	0.99496	0.99551	0.99578
50	# training examples	**1000**	**2000**	**5000**	**10000**	**20000**	**30000**
	Mosaic >0.5	0.94165	0.9393	0.94284	0.94151	0.94403	0.94246
	Mosaic >0.9	0.84033	0.858	0.86627	0.86898	0.87017	0.86613
	Het <0.5	0.97008	0.9709	0.97114	0.97159	0.97119	0.97202
	Het <0.9	0.99502	0.994	0.99365	0.99311	0.99345	0.99341
	Hom <0.5	0.96994	0.9733	0.97322	0.97372	0.97408	0.97418
	Hom <0.9	0.99629	0.996	0.99599	0.99566	0.99611	0.99631
62	# training examples	**1000**	**2000**	**5000**	**10000**	**20000**	**30000**
	Mosaic >0.5	0.95066	0.9508	0.95103	0.94975	0.9523	0.95019
	Mosaic >0.9	0.86295	0.8777	0.88112	0.88244	0.88564	0.88039
	Het <0.5	0.97127	0.9725	0.97229	0.97518	0.9736	0.97379
	Het <0.9	0.99523	0.9938	0.99411	0.99414	0.99471	0.99444
	Hom <0.5	0.96965	0.9746	0.97586	0.9779	0.97693	0.97807
	Hom <0.9	0.99668	0.9963	0.99674	0.99675	0.99694	0.99707

Table S1: Cross-validation to evaluate the number of training examples and the random forest score threshold. Related to Figure 2.

1. Minimum depth (excluding marked duplicates, QC fail, secondary and supplementary alignments) [at least 16]
2. Minimum fraction of reads phased [at least 0.5]
3. Minimum fraction of reads on less-prevalent haplotype [at least 0.3]
4. Maximum fraction of reads that have neither reference nor mosaic allele [at most 0.05]
5. Minimum mosaic allele frequency [at least 0.05]
6. Minimum number of haplotype-discordant reads [at least 4]
7. Maximum number of haplotype-discordant reads on the less-prevalent haplotype [at most 0.1]
8. Minimum average position from end of alignment of haplotype-discordant reads [at least 10]

Filters that can be "on" or "off":
1. At least one haplotype-discordant read, one haplotype-concordant read, one reference-allele read and one mosaic-allele read must be aligned in proper pair orientation
2. At least one haplotype-discordant read, one haplotype-concordant read, one reference-allele read and one mosaic-allele read must have an alignment that is not soft-clipped
3. At least one haplotype-discordant read, one haplotype-concordant read, one reference-allele read and one mosaic-allele read must be aligned on the plus and on the minus strand

Figure S8: preFilter features [default value to pass filter in brackets]. Related to Figure 2.

	Tumor			Normal		
Case	WES coverage	WGS coverage	CNVNATOR %	WES coverage	WGS coverage	CNVNATOR %
1	549	45	9.3	617	42	8.7
2	504	41	16.8	529	41	9.3
3*	271	35	23.6	255	34	11.0
4*	223	34	12.4	232	34	11.9
5*	207	34	15.1	268	35	10.9
6*	226	40	11.8	223	38	11.5
7	472	35	10.3	445	38	8.4
8*	330	35	11.1	319	34	10.9
9*	411	36	16.1	346 (Blood) 400 (Tissue)	36 (Blood) 36 (Tissue)	10.8 (Blood) 11.0 (Tissue)
10*	500	40	11.0	392	37	22.0
11	669	37	10.5	579	35	10.5
12*	618	37	11.4	726	37	10.9
13	777	37	20.1	681	37	8.7

Table S2: Cases using reference genome GRCh38 2.1.0 (1, 2, 7, 10, 11) were processed with Long Ranger 2.1.6 and GATK HaplotypeCaller 3.8-0. Samples using reference genome b37 2.1.0 (3, 4, 5, 6, 8, 9, 10, 12) were processed with Long Ranger 2.1.3 and GATK HaplotypeCaller 3.5-0. Related to Table 2.

Importance	Abbreviation	Number in Figure S5
0.206699	weightedMbq	33
0.136303	MAF	4
0.115912	MAF_phased	5
0.101952	weightedCbq	32
0.078008	fracC	7
0.075791	CMAF	9
0.065965	nC	6
0.058420	Mavgbq	22
0.050114	Cavgbq	10
0.028026	NMAF	16
0.016379	Mavgclip	24
0.009496	MavgASXS	26
0.008695	Mavgind	25
0.007776	NavgASXS	21
0.006754	JavgASXS	31
0.006130	CavgASXS	14
0.003744	Cfrach	8
0.003665	Cavgind	13
0.003250	Cavgclip	12
0.002759	Navgbq	17
0.002578	Navgind	20
0.002276	Javgind	30
0.001835	Javgbq	27
0.001569	Javgclip	29
0.001236	fracphased	2
0.001149	depth	1
0.000996	Navgpos	18
0.000904	Mavgpos	23
0.000504	Cavgpos	11
0.000476	Javgpos	28
0.000405	Navgclip	19
0.000148	frach	3
0.000086	Nfrach	15

Table S3: Samovar model feature importances in simulation experiment. Related to Figure 3.

	Samovar									MuTect2						MosaicHunter								
	Full Model			Short			No Phasing			Tumor-Only			Paired			Tumor-Only			Paired			Trio		
30X Coverage	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*
Autosomes	84.0	30.1	44.4	83.7	2.0	3.9	3.4	68.3	6.4	3.0	83.2	5.7	60.8	91.4	73.0	31.5	5.1	8.8	79.2	20.7	32.8	70.4	20.7	32.0
Exons	84.0	28.3	42.4	85.5	1.8	3.5	4.6	70.7	8.6	3.6	85.3	7.0	60.1	92.0	72.7	35.0	7.1	11.8	82.1	30.8	44.8	73.7	30.8	43.4
Genes	84.9	30.1	44.4	84.5	1.8	3.6	3.9	69.2	7.5	3.2	84.4	6.2	63.0	92.0	74.8	32.6	5.7	9.7	79.9	22.7	35.4	71.2	22.7	34.5
Enhancer	88.5	31.0	45.9	90.9	2.1	4.1	4.4	61.8	8.2	3.9	86.7	7.5	72.9	92.3	81.4	37.8	5.9	10.1	85.5	29.5	43.8	80.2	29.5	43.1
Promoter	83.3	26.1	39.8	76.9	1.4	2.7	4.0	65.2	7.5	3.0	83.2	5.8	59.4	90.9	71.9	35.3	6.1	10.4	80.5	25.1	38.3	73.7	25.1	37.5
Alu	82.0	28.6	42.4	81.1	2.3	4.4	2.7	73.1	5.3	2.3	78.2	4.5	54.5	88.4	67.4	8.6	0.0	0.1	56.5	0.3	0.6	53.1	0.3	0.6
RepeatMasker	84.2	29.6	43.9	82.3	2.0	3.9	2.9	67.0	5.5	2.8	81.5	5.3	58.9	90.1	71.2	20.2	0.3	0.6	72.3	1.4	2.7	61.3	1.4	2.7
Seg. Dup.	25.6	10.4	14.8	51.9	0.8	1.5	0.6	25.5	1.2	1.3	56.9	2.5	18.4	62.8	28.5	6.6	0.5	0.9	39.3	1.7	3.2	29.1	1.7	3.2
60X Coverage	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*				Prec	Rec	*F*				Prec	Rec	*F*
Autosomes	84.6	43.0	57.1	87.8	2.0	4.0	3.2	67.9	6.1	3.6	76.0	7.0				32.4	15.5	20.9				46.8	27.2	34.4
Exons	84.3	41.8	55.9	87.3	1.7	3.2	4.6	69.4	8.7	4.7	79.6	8.8				38.5	25.3	30.5				54.0	45.5	49.4
Genes	85.6	43.4	57.6	89.1	2.0	3.8	4.0	68.9	7.5	3.9	77.2	7.5				33.1	17.0	22.4				47.7	30.0	36.8
Enhancer	90.8	47.8	62.6	93.3	2.2	4.2	4.4	61.1	8.1	4.8	77.9	9.0				36.9	22.7	28.1				51.6	40.0	45.1
Promoter	85.4	40.7	55.2	83.1	1.5	2.9	4.0	64.5	7.5	4.0	76.8	7.6				38.5	21.1	27.3				56.4	40.5	47.2
Alu	81.1	42.9	56.1	84.6	2.5	4.8	2.6	72.7	5.0	3.0	68.0	5.7				16.5	0.2	0.5				31.7	0.5	1.0
RepeatMasker	84.2	42.2	56.2	87.1	2.1	4.1	2.6	66.6	5.0	3.4	74.1	6.4				24.7	1.0	1.9				38.3	1.8	3.4
Seg. Dup.	28.0	13.1	17.8	64.3	0.7	1.3	0.5	23.6	1.0	1.6	48.5	3.1				9.8	1.5	2.6				18.5	2.7	4.7

Table S4: Precision (Prec), recall (Rec), and F score of each tool for the synthetic mosaic variants inserted by bamsurgeon. This table includes the Samovar "short" and "no-phasing" models, engineered to demonstrate the importance of linked reads for recall and phasing information for precision. Related to Table 1.

	Samovar									MuTect2						MosaicHunter								
	Full Model			Short			No Phasing			Tumor-Only			Paired			Tumor-Only			Paired			Trio		
30X Coverage	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*
Autosomes	89.6	42.1	57.3	86.9	2.7	5.2	3.6	94.1	7.0	3.2	85.6	6.2	66.1	93.2	77.4	31.6	14.8	20.2	79.3	59.4	67.9	70.5	59.4	64.5
Exons	93.8	39.6	55.7	83.7	2.1	4.2	5.6	94.7	10.6	4.1	87.2	7.9	64.7	93.4	76.4	34.9	12.5	18.4	82.4	54.2	65.3	73.9	54.2	62.5
Genes	90.9	42.4	57.8	87.9	2.5	4.8	4.6	94.8	8.7	3.4	86.7	6.6	67.1	93.7	78.2	32.7	15.1	20.6	80.0	60.1	68.6	71.2	60.2	65.2
Enhancer	94.2	42.0	58.1	91.7	2.5	4.9	5.5	96.1	10.4	4.0	87.7	7.7	70.1	93.6	80.2	36.4	11.3	17.3	85.9	58.7	69.7	80.1	58.7	67.8
Promoter	91.4	36.3	52.0	76.7	1.7	3.4	4.6	93.9	8.8	3.2	85.8	6.2	60.5	92.4	73.1	35.4	11.6	17.5	80.7	48.7	60.7	74.2	48.7	58.8
Alu	30.2	18.6	23.0	25.0	1.2	2.4	0.9	60.5	1.8	1.4	61.4	2.8	27.1	61.4	37.6	9.7	4.3	5.9	52.6	28.6	37.0	50.0	28.6	36.4
RepeatMasker	68.5	33.1	44.7	73.7	2.3	4.4	0.7	73.9	1.3	2.6	72.3	5.0	45.6	75.8	57.0	24.4	10.1	14.2	75.5	45.0	56.4	65.3	45.0	53.3
Seg. Dup.	6.8	4.4	5.3	28.6	0.6	1.2	0.1	17.0	0.2	0.8	42.5	1.6	11.1	39.6	17.4	7.8	4.1	5.4	37.6	11.9	18.1	27.7	12.3	17.0
60X Coverage	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*	Prec	Rec	*F*				Prec	Rec	*F*				Prec	Rec	*F*
Autosomes	89.7	60.3	72.1	89.3	2.7	5.3	3.4	94.0	6.6	4.0	78.7	7.6				32.4	44.5	37.5				46.8	78.3	58.5
Exons	91.7	58.5	71.4	89.2	2.1	4.1	5.9	95.0	11.2	5.4	81.8	10.1				38.6	44.7	41.4				54.0	80.6	64.7
Genes	90.8	60.8	72.8	91.3	2.7	5.2	5.0	95.1	9.4	4.3	79.8	8.1				33.1	45.0	38.1				47.6	79.4	59.5
Enhancer	94.5	63.3	75.8	94.4	2.8	5.5	5.8	94.8	10.9	4.9	79.6	9.3				36.5	45.1	40.4				51.1	79.6	62.3
Promoter	91.2	57.3	70.4	92.2	2.3	4.5	4.9	94.7	9.4	4.2	78.2	8.0				38.6	40.5	39.5				56.5	78.1	65.6
Alu	45.9	30.0	36.3	57.1	3.1	5.8	0.7	53.1	1.4	2.1	56.2	4.1				17.0	20.8	18.7				32.4	43.8	37.3
RepeatMasker	73.6	43.9	55.0	82.7	2.5	4.8	0.4	69.4	0.8	3.2	62.8	6.0				27.5	31.2	29.2				41.5	54.9	47.3
Seg. Dup.	9.2	6.0	7.3	15.4	0.4	0.7	0.1	13.3	0.2	1.0	32.9	2.0				10.1	10.3	10.2				18.8	18.5	18.7

Table S5: Precision (Prec), recall (Rec), and F score of each tool for the synthetic mosaic variants inserted by bamsurgeon in the region of the genome not filtered out by MosaicHunter or Samovar. This table includes the Samovar "short" and "no-phasing" models, engineered to demonstrate the importance of linked reads for recall and phasing information for precision. Related to Table 1.

Importance	Abbreviation	Number in Figure S6
0.20520657	weightedMbq	33
0.16749462	MAF	4
0.13410861	weightedCbq	32
0.08401046	fracC	7
0.07369292	MAF_phased	5
0.06894606	Mavgbq	22
0.06203258	nC	6
0.02828793	Cfrach	8
0.02817725	Mavgclip	24
0.02439597	Cavgbq	10
0.01783998	MavgASXS	26
0.01466872	JavgASXS	31
0.01386411	CMAF	9
0.01311324	CavgASXS	14
0.01274496	NMAF	16
0.00876721	NavgASXS	21
0.008098	Mavgind	25
0.00757443	fracphased	2
0.00331775	Navgind	20
0.00318227	Javgind	30
0.00307518	Cavgind	13
0.00297741	Mavgpos	23
0.00291929	Cavgpos	11
0.00285171	Javgbq	27
0.00209531	Javgclip	29
0.00169362	Navgbq	17
0.00123443	Javgpos	28
0.00070641	Navgclip	19
0.00070569	Nfrach	15
0.00069773	Navgpos	18
0.00069729	depth	1
0.00052806	frach	3
0.00029423	Cavgclip	12

Table S6: Short-read phasing model feature importances in simulation experiment. Related to Figure 3.

Importance	Abbreviation	Number in Figure
0.42298002	weightedMbq	11
0.29348068	MAF	2
0.14507064	Mavgbq	3
0.07905771	Mavgclip	5
0.03207496	Mavgind	6
0.00746506	Javgind	10
0.00448625	Javgpos	8
0.00438944	Javgbq	7
0.00403807	depth	1
0.0038253	Mavgpos	4
0.00313186	Javgclip	9

Table S7: No-phasing model feature importances in simulation experiment. Related to Figure 3.

Case	Diagnosis	Sex*
1	Indeterminate, most consistent with oligodendroglioma	M
2	Pilocytic astrocytoma	M
3	Medulloblastoma, WHO grade IV, most consistent with non-WNT, non-SHH subgroup	M
4	Pilocytic astrocytoma	F
5	Glioblastoma (recurrence)	F
6	Pilocytic astrocytoma	F
7	Ewing-like sarcoma	M
8	Ganglioglioma, WHO grade 1	M
9	Diffuse Midline Glioma, H3 K27M-mutant, WHO grade IV	M
10	Indeterminate, high grade glioma/astrocytoma	M
11	Ganglioglioma, WHO grade 1	F
12	Glioma (low grade)	M
13	Clival chordoma	F

Table S8: Metadata for each case. * Sex determined from alignments to Y-chromosome. Related to Table 2.

Case	Calls	Sensitivity
1	58	0.70
2	85	0.61
3	70	0.57
4	73	0.68
5	51	0.58
6	73	0.58
7	61	0.63
8	43	0.62
9	39	0.48
10	50	0.45
11	30	0.73
12	59	0.78
13	70	0.88
Total	762	

Table S9: Samovar analysis of normal WGS dataset for pediatric cancer cases. Number of calls shown is for the WES capture region, and validation performed as described in main text. Related to Figure 4.