Comprehensive analysis of structural variants in breast cancer genomes using single molecule sequencing

Supplemental Methods

Sergey Aganezov[1], Sara Goodwin[3], Rachel M. Sherman[1], Fritz J. Sedlazeck[2], Gayatri Arun[3], Sonam Bhatia[3], Isac Lee[1], Melanie Kirsche[1], Robert Wappel[3], Melissa Kramer[3], Karen Kostroff[4], David L. Spector[3], Winston Timp[1], W. Richard McCombie[3], Michael C. Schatz[1,3*]

1. Johns Hopkins University, Baltimore, MD, 21211
2. Baylor College of Medicine, Houston, TX 77030
3. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
4. Northwell Health, Lake Success, NY 11042

*corresponding author: mschatz@cs.jhu.edu

Analysis of short-read exclusive SVs.

First, we analyzed the origin and nature of short-read exclusive SV calls in samples 51T and SKBR3. For every short-read exclusive SV we considered two (one for insertions) 500bp windows, one centered on each breakpoint of the SV. We then computed the average read-depth coverage for every window from the ONT, PacBio, and 10xG/Illumina read alignments as well as the genome-wide average. We observe a large coverage increase compared to the genome-wide coverage in the 10xG/Ilumina alignments for 51T and SKBR3, both in terms of mean (from 30 to ~700 in 51T, 30 to ~800 in SKBR3) and median values (from 30 to ~75 in 51T, 30 to ~95 in SKBR3), whereas we observe more modest read-depth coverage changes in PacBio alignments (mean: 49 to 64, and 57 to 99; median: 50 to 52 and 50 to 59 in 51T and SKBR3, respectively) and ONT alignments (mean: 49 to 63 and 37 to 59; median: 51 to 52 and 33 to 36) in 51T and SKBR3 respectively).

To further investigate the coverage changes in short-read exclusive SV regions, we considered the whole reference genome G (GRCh38) with a set W of non-overlapping 500bp windows with BEDTools (Quinlan and Hall 2010) and computed an averaged read-depth coverage $c\,(w,\ t)$ for every window $w \in W$ for technology $t = \{10x,\ ONT,\ PB\}$ for 10xG/Illumina, ONT, and PacBio datasets, respectively. For each technology $t \in \{10x,\ ONT,\ PB\}$ and every window w we compute the mean-normalized coverage fraction $f(w,\ t)\ =\ c(w,\ t)\,/\,c(G,\ t)$, where $c(G,\ t)$ is the mean coverage across a whole genome G in technology t.

We then define a window $w \in W$ to have abnormally high-coverage of the 10xG/Illumina alignments if the window's coverage is greater than 1.5 times the genome wide average:

$$f(w, 10x)\ >\ 1.5 \quad (1)$$

and the coverage deviation of the window from the genome wide average in 10xG/Illumina is greater than twice that of the PacBio and ONT in the same window:

$$f(w,\ 10x)\,/\,f(w,\ ONT)\ >\ 2 \ (2) \quad \text{and} \quad f(w,\ 10x)\,/\,f(w,\ PB)\ >\ 2 \ (3).$$

For samples 51T and SKBR3 we observe that 71.7% (1,925/2,683) and 60.8% (892/1,467) of short read exclusive SVs have at least one breakpoint within a Illumina/10xG abnormally high-coverage window (**Supplemental Figure 9a,b**).

We further observe that 13.6% (366/2,683) and 15.7% (231/1,467) short read exclusive SVs have at least 1 breakpoint that lies outside of the Genome In A Bottle (GIAB) high confidence (HC) intervals for SV detection (Zook et al. 2020) in samples 51T and SKBR3 respectively (**Supplemental Figure 9c,d**).

Lastly, we consider how many short-read exclusive SVs in samples 51T and SKBR3 co-locate with a long-read specific SVs, which may indicate the SV type or size was mis-called from short-reads. For every long-read SV we created two 500bp breakpoint-centered intervals (one for insertions), as described for short-read SVs, above. Then, we used BEDTools to intersect breakpoint intervals for short- and long-read SVs. We observe that 12.7% (342/2,683) and 9.7% (240/1,467) of short-read exclusive SVs in samples 51T and SKBR3 had a nearby long-read SV, with the long-read SV type being an insertion most often, followed by deletion, duplication, inversion, and lastly translocation. We note that when considering sensitive, rather than specific long-read SVs, the number of short-read SVs with a nearby long-read SV increases to 50.2% (1,348/2,683) and 55.1% (808/1,467) for samples 51T and SKBR3, respectively (**Supplemental Figure 10**). An example of a short-read exclusive duplication (identified by both Lumpy and SvABA) with and without a nearby long-read SV is shown in **Supplemental Figure 11**.

Overall, we observe, in samples 51T and SKBR3 respectively, that 77.6% (2,082/2,683) and 71.7% (1,052/1,467) of short-read exclusive SVs that were not supported by long-read SVs of the same size/type can be explained by the combination of abnormally high-coverage 10xG/Illumina regions, SVs outside the GIAB HC intervals, and nearby long-read specific SVs of a different size or type. Additionally, considering nearby sensitive long-read SVs, as described above, increases the explainable variants to 87.4% (2,345/2,683) in 51T and 85.3% (1,251/1,467) in SKBR3.

Next, we used BEDTools to intersect short-read exclusive SVs with tandem repeats as determined by Tandem repeats finder (Benson 1999), downloaded from the UCSC Table Browser (Karolchik et al. 2004). We observe that 60.7% (1,628/2,683) and 72.1% (1,058/1,467) of short-read exclusive SVs in samples 51T and SKBR3 overlap simple repeat regions.

We then analyzed the SVs produced by individual short-read methods (both barcode aware and not), thus relaxing our 2+ short-read caller requirement for an SV to be considered, and compared them with specific SVs inferred with both ONT and PB (when available) long-read data. We merged the short-read SVs, per-method, with long-read SVs using SURVIVOR and the same merging strategy, as previously outlined (minimum SV size set to 30bp, breakpoint orientation considered, maximum distance between SVs set to 1000bp).

In **Supplemental Table 1** we report the number of specific SV calls produced by each short-read SV inference method (specific), and the number of short-read SVs that remained after intra-sample SV deduplication (dd) (i.e., two or more SV calls from the same caller were highly similar, based on our merging criteria outlined above, and we thus only count them once; this is consistent with other SV comparison merging approaches (Larson et al. 2019)). We then report the number of de-duplicated short-read SVs, per calling method, that had matching SVs called from the long-read dataset(s) (lr-supp). While counts of SVs are extraordinarily high from some of the short read callers, in some cases producing over 50k SV calls in one sample, these are reduced after deduplication. This is most prominent in the SV calls- produced by Lumpy, and appears to occur in abnormally high-coverage regions, with several examples shown in **Supplemental Figure 12**. However, even after intra-sample deduplication, the percentage of SVs produced by individual short-read SVs callers that are supported by long-read SVs, remains low, ranging from below 10% with the more permissive callers that produce tens of thousands of SV calls per sample, to ~77% with the most coservative caller, grocsv, which only calls up to a few hundred SVs per sample Overall, nearly 85% of the intra-sample deduplicated SVs do not have any long-read support.

Analysis of abnormal high-coverage regions in Illumina/10xG datasets

We performed additional analysis on the abnormally high coverage regions detected in the 10x Genomics sequencing datasets for samples 51T, 51N, SKBR3. We also included two additional, previously published, samples NA12878 (Jain et al. 2018; Zook et al. 2016) and HG002 (Zook et al. 2016; Shafin et al. 2020), for which we obtained ONT, 10x Genomics, and regular Illumina data. Long ONT reads were aligned with NGMLR as previously described. We used pre-existing alignments of Illumina reads, aligned with bwa-mem (Li and Durbin 2009) to GRCh38, and pre-existing 10x Genomics read alignments to GRCh38 performed with LongRanger. The alignments had a mean read-depth coverage of 33x, 70x, and 35x for sample NA12878 and 94x, 79x, and 80x for sample HG002, for ONT, 10xG, and Illumina read alignments, respectively.

Following the procedure outlined in ***Analysis of short-read exclusive SVs***, above, we computed 500bp-windows averaged read-depth alignment coverage of the main chromosomes for all 5 considered samples for ONT, 10x Genomics, and Illumina (when available) reads. We then considered, for each sample, all 500bp windows with abnormally high 10x Genomics coverage, as defined above, which we further intersected with GIAB HC genomic intervals. We report both the total number of high coverage windows and the overlap with GIAB HC regions observed in each sample **(Supplemental Figure 13a)** as well as how many of the GIAB HC high coverage regions are shared across observed samples **(Supplemental Figure 13b)**. While the majority (43,536) of identified high-coverage regions are sample-specific, we observe 7,228 regions, which we denote as *shared high coverage* (SHC), with abnormally high coverage in all 5 of observed samples.

We then measured the GC composition of the reference genome inside and outside of the SHC regions, finding that SHC regions have elevated GC composition relative to the rest of the reference genome **(Supplemental Figure 13c)**. It is likely the elevated GC contributes to uneven amplification within the 10xG barcoding protocol.

Furthermore, for every window w we computed the number of distinct verified 10x Genomics barcodes reads fro which are aligned in w. We observe an increase in the number of barcodes inside SHC regions as compared to the non-SHC regions **(Supplemental Figure 14)**. With the increase in the number of uniquely carcoded molecules coming from SHC regions, we further investigated the reads yields coming from molecules with distinct barcodes. For every window we computed the mean number $\bar{B}$ and the median number $\tilde{B}$ of reads having the same verified

10x Genomics barcode. We observe an increase in the mean number $\bar{B}$ of reads having the same barcode in SHC regions as compared to the counts in 20,000 non-SHC windows sampled uniformly at random from the GIAB HC genomic intervals **(Supplemental Figure 15a)**. However, we observe an almost negligible increase in the median number $\tilde{B}$ of reads having the same barcode when comparing SHC and non-SHC regions **(Supplemental Figure 15b)**.

With the increase in mean numbers $\bar{B}$ yet relatively unchanged median values $\tilde{B}$, we then investigated the molecule-read distribution tail. For every window w we identified the 5 most represented (via the number of associated reads) 10x Genomics barcodes, and then computed a number T_5 of all the reads coming from the 5 most represented barcodes in w. We observe a very sharp increase in the number T_5 of reads coming from 5 most frequently observed barcodes within SHC regions as compared to the non-SHC regions **(Supplemental Figure 16)**.

For samples NA12878 and HG002, for which we had both the 10x Genomics and regular paired-end Illumina short reads, we investigated if the SHC regions have similar coverage increase abnormalities in non-10xG Illumina alignments. We find that neither NA12878 **(Supplemental Figure 17a)** nor HG002 **(Supplemental Figure 17b)** have similar coverage increase abnormalities within 10xG SHC in the Illumina alignments, thus suggesting a 10xG-specific preparation/sequencing artifact. IGV screenshots of ONT, 10x Genomics, and Illumina (when available) alignments within several SHC regions are displayed in **Supplemental Figure 18**.

RCK's support for haplotype constraint groups

We assume that all the mutated cancer genomes evolve from a diploid reference genome R. Every chromosome in R is present in two homologous copies, which we label A and B respectively. A segment $s_H = C : [s^t, s^h]$, where $H \in \{A, B\}$ is a contiguous part of the chromosome C_H, and its endpoints s^t, s^h that determine the *tail* and the *head* of the segment are called *extremities*. We may omit chromosomal names in a segment's signature, when obvious and/or irrelevant, depending on the context. Segments are labeled $1, \ldots, m$ throughout the reference genome's chromosomes. Segments $(s_H, (s+1)_H)$ that are sequential on some chromosome determine an *adjacency* $\{s^h_H, (s+1)^t_H\}$. We denote by $A(G)$ a set of adjacencies present in genome G, and we denote by $A_N(G)$ a set of *novel adjacencies* (i.e., not present in the reference) present in genome G. We will omit the *tail* and *head* superscripts on extremities involved in adjacencies, when orientation is not important, just retaining the coordinates. Outermost segments' extremities on every chromosome are called *telomeres*, and we denote by $T(G)$ a set of telomeres in genome G.

We depict every large-scale rearrangement as a collection of double-strand breakages that destroy adjacencies, with possible amplification and/or loss of involved segments, and subsequent ligation of involved segments' extremities, which introduces novel adjacencies. We assume that somatic large-scale evolutionary history of observed cancer genomes complies with the Infinite Sites (IS) assumption, under which the same genomic coordinate (i.e., the same adjacency) on either A or B haplotype can be directly involved in at most one, however complex, genome rearrangement's breakage with subsequent ligation. We also assume that in rearranged genomes chromosomal telomeres are inherited from the reference, or more formally, $T(G) \subseteq T(R)$, as telomere genomic sequences play an important role in cells life cycles and they are required to contain specific sequences for the respective molecule replication to complete correctly.

For a genome G the inferred SVs correspond to set $\tilde{A}_N(G)$ of unlabeled novel adjacencies (i.e., for every novel adjacency $a = \left\{p_F^x,\ q_D^y\right\} \in A_N(G)$, where $F,\ D \in \{A,\ B\}$, and $x,\ y \in \{t,\ h\}$, we measure its unlabeled version $\tilde{a} = \{p^x,\ q^y\}$, or an SV, which is missing haplotype labels). Under the IS assumption every measured unlabeled novel adjacency $\tilde{a} = \{p^x,\ q^y\}$ has a unique haplotype-specific counterpart $a \in A_N(G)$ (i.e., unique haplotype labels $F,\ D \in \{A,\ B\}$ such that $a = \{p_F^x,\ q_H^y\} \in A_N(G)$). We also note, that we allow cases when multiple distinct measured SVs ($a = \{p,\ x\}$, $b = \{p,\ y\}$, $\ldots$) involve the same unlabeled extremity p and assume that all of the underlying true SVs involve the same labeled segment's extremity on one of the two haplotypes.

We call a set $P = \{p,\ q,\ \ldots\}$ of unlabeled extremities, or breakends, involved in measured SVs a *haplotype constraint group* if for every haplotype-specific version of the SVs involving extremities from P, the involved haplotype-specific extremities (e.g., $p_H,\ q_H$) belong to the same haplotype $H \in \{A,\ B\}$. We now describe how we obtain a set P of haplotype-constraint groups from the measurement data, as we extend the previous version of the RCK framework in which haplotype constraint groups were only determined for pairs of reciprocal (i.e., adjacent in the reference) breakends.

Given a set $\tilde{A}_N(G)$, of unlabeled novel adjacencies, or SVs, we call an inter-chromosomal SV $a = \{p,\ q\} \in \tilde{A}_N(G)$ *uninterrupted* (*uSV*) if $|q - p| < 5000$ and no other SV has a breakend overlapping with a, or, more formally, there does not exist an SV $b = \{u, v\} \in \tilde{A}_N(G)$, such that either $p \leq u \leq q$, or $p \leq v \leq q$, or both. We assume that every uninterrupted SV $a = \{p,\ q\}$ determines a haplotype-constraint group $P_a = \{p,\ q\}$, as the opposite will correspond to an unlikely event of 2+ double-strand breakages involving homologous copies of the same chromosome with breakage coordinates located very close to one another.

For a long-read r that spans a set S_r of SVs we consider the ordered sequence $O(S_r)$ of the SVs from S_r and an ordered sequence $O(B(S_r))$ of breakends $B(S_r)$ as determined by r's traversal of SVs in S_r. We note that since every SV determines a pair $\{p,\ q\}$ of breakends we naturally have $|O(B(S_r))| \equiv 0 (mod\, 2)$. For every consecutive pair $a,\ b$ of SVs from $O(S_r)$ let us observe the last breakend p of a and the first breakend q of b as determined by $O(B(S_r))$. Alternatively, we can say, that we observe every $2i$ th and $(2i+1)$ st elements p and q of $O(B(S_r))$, where $i = \overline{1,\ |O(S_r)|}$. Since both SVs a and b are spanned by the same read, that means that there exists a segment $s = \{p,\ q\}$ in the sequenced cancer genome, which was not altered by any rearrangements (as otherwise there would be another SV and another breakends between p and q). Under the IS assumption the long-read r a pair $\{p,\ q\}$ of breakends determines a haplotype constraint group $\{p,\ q\}$.

To obtain the set P of haplotype constraint groups for a set $\tilde{A}_N(G)$ of measured SVs we build a haplotype constraint graph $G_P = (V,\ E)$. The set $V = \{p,\ q \mid a = \{p,\ q\} \in \tilde{A}_N(G)\}$ of vertices is determined by breakends involved in SVs from $\tilde{A}_N(G)$. The set E of edges is constructed by adding edges $\{p, q\}$ that are determined by two-breakend haplotype constraint groups, which are either coming from reciprocal SVs' breakends, determined by unique SVs, or inferred from long-reads spanning multiple SVs. We note that if there exist two haplotype constraint groups P_1 and P_2 such that $P_1 \cap P_2 \neq \varnothing$, then the set $P = P_1 \cup P_2$ is also a haplotype constraint group. The desired set P then corresponds to connected components of size (i.e., number of vertices) greater than 1 in G_P.

RCK infers a edge multiplicity function μ on a corresponding Diploid Interval Adjacency Graph that is determined by the underlying cancer genomes, by solving an optimization problem formulated as a mixed integer linear program (MILP). In RCK 1.1 we extended the optimization problem formulation to include the constraints that haplotype constraint groups place on which haplotype-labeled versions of measured SVs can be present in the inferred karyotype graph.

Supplemental References

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338–345.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493–6.

Larson DE, Abel HJ, Chiang C, Badve A, Das I, Eldred JM, Layer RM, Hall IM. 2019. svtools: population-scale analysis of structural variation. *Bioinformatics* **35**: 4782–4787.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol*. http://dx.doi.org/10.1038/s41587-020-0503-6.

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

Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al. 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*. http://dx.doi.org/10.1038/s41587-020-0538-8.