Characterization of segmental duplications and large inversions using Linked-Reads

Fatih Karaoglanoglu [1], Camir Ricketts [2,4], Marzieh Eslami Rasekh [3], Ezgi Ebren [1], Iman Hajirasouliha [4,5,*] and Can Alkan [1,6,7,*]

[1]Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey

[2]Tri-Institutional Computational Biology & Medicine Program, Cornell University, NY, USA

[3]Graduate Program in Bioinformatics, Boston University, Boston, MA, 02215, USA.

[4]Inst. for Computational Biomedicine, Dept. of Physiology and Biophysics, Weill Cornell Medicine, NY, USA

[5]Englander Institute for Precision Medicine, The Meyer Cancer Center, Weill Cornell Medicine, NY, USA

[6]Bilkent-Hacettepe Health Sciences and Technologies Program, Ankara, 06800, Turkey and

[7]Department of Computer Science, ETH Zürich, 8006, Switzerland.

Abstract

Many algorithms aimed at characterizing genomic structural variation (SV) have been developed since the inception of high-throughput sequencing. However, the full spectrum of SVs in the human genome is not yet assessed. Most of the existing methods focus on discovery and genotyping of deletions, insertions, and mobile elements. Detection of balanced SVs with no gain or loss of genomic segments (e.g., inversions) is particularly a challenging task. Long read sequencing has been leveraged to find short inversions but there is still a need to develop methods to detect large genomic inversions. Furthermore, currently there are no algorithms to predict the insertion locus of large interspersed segmental duplications.

Here we propose novel algorithms to characterize large (>40Kbp) interspersed segmental duplications and (>80Kbp) inversions using Linked-Read sequencing data. Linked-Read sequencing provides long range information, where Illumina reads are tagged with barcodes that can be used to assign short reads to pools of larger (30-50 Kbp) molecules. Our methods rely on *split molecule* sequence signature that we have previously described [11]. Similar to the split read, split molecules refer to large segments of DNA that span an SV breakpoint. Therefore, when mapped to the reference genome, the mapping of these segments would be discontinuous. We redesign our earlier algorithm, VALOR, to specifically leverage Linked-Read sequencing data to discover large inversions and characterize interspersed segmental duplications. We implement our new algorithms in a new software package, called VALOR$_2$.

Availability: VALOR$_2$ is available at https://github.com/BilkentCompGen/valor.

* Joint corresponding authors. calkan@cs.bilkent.edu.tr, imh2003@med.cornell.edu

1 Introduction

Alterations of DNA content and organization larger than 50 bp, commonly referred to as genomic structural variations (SVs) [2], are among the major drivers of evolution [24, 29], and diseases of genomic origin [38]. Despite decades of research they remain difficult to accurately characterize contributing to our lack of full understanding of the etiology of complex diseases, termed *missing heritability* [9].

High-throughput sequencing (HTS) technologies are widely employed to discover and genotype various classes of SVs since their inception [18, 13, 26, 34, 12, 19, 36]. However, effectiveness has been limited by either very short read lengths (e.g., Illumina), or high error rates and prohibiting cost (e.g., PacBio and Oxford Nanopore). The human genome complexity further contributes to our lack of full characterization of structural variants, especially large-scale duplications and balanced rearrangements due to the repetitive and duplicated sequence at the SV breakpoints [17]. Despite high error rates, long reads offer improvement in complex SV discovery, either used alone [10, 16], or when integrated with standard short-read sequencing data [32].

Recently Linked-Read sequencing methods such as the 10x Genomics system (10xG) was introduced as an alternative method to generate highly accurate Illumina short reads data with additional long-range information [27]. In the 10xG system, large DNA molecules (typically 10-100 Kbp) are barcoded and randomly separated into over a million partitions (here we term these partitions "pools"). Each pool contains roughly 2-30 large molecules. These pools are then sequenced at very low coverage (∼0.1X) using the standard Illumina platform. Shared barcodes among Illumina read pairs show them as generated from the same pool. Since each pool is diluted to contain only a very small fraction of the input DNA, the probability of barcode collision is negligible [45]. For example, assuming 20 molecules per pool and an average size of 30 Kbp per molecule, each pool on average contains only $\frac{1}{5{,}000}$ of the haploid human genome. Linked-Reads then can be used to "reconstruct" large molecules that originate from the same haplotype. Furthermore, Linked-Read sequencing makes it possible to obtain very high physical coverage with the cost of generating moderate sequence coverage data[1].

The ability of extracting long range information from accurate and inexpensive but short read sequencing data makes Linked-Read sequencing attractive for various applications. It has been used for genome scaffolding [47], haplotype-aware assembly [27, 33, 43], metagenomics [8], single cell transcriptome profiling [35, 44] and regulatory network clustering [1], haplotype phasing [27, 33, 48], and genome structural variation discovery [11, 23, 37, 45].

Linked-Read techniques for genomic structural variation discovery include VALOR [11], Long Ranger [23] and GROC-SVs [37]. VALOR was the first algorithm that used "split molecule" signature, similar to the commonly used split read signature [46], together with traditional read pair signature [42, 26, 2] to characterize large (>500 Kbp) inversions. Split molecules are defined as large molecules that span an SV breakpoint, and therefore mapped as two disjoint intervals to the reference genome.

Long Ranger[23] is a comprehensive software package developed by 10X Genomics, for the purpose of barcode-aware read alignment and resolving full-scale human germline genome variation, while GROC-SVs is an optimized tool for somatic and complex SVs in cancer genomes. Both Long Ranger and GROC-SVs employ a novel idea to utilize discordance in expected "barcode coverage" as well as barcode similarities across distant locations for potential large-scale SV signals. In addition,

[1] e.g., 30X sequence coverage corresponds to 150X physical coverage.

GROC-SVs [37] performs local assembly on barcoded reads to detect large complex events that are between 10-100 Kbp with breakpoint resolution.

Despite the aforementioned advances in SV discovery using various technologies, detecting both balanced rearrangements (i.e., inversions and translocations) and segmental duplications (SDs) remain challenging due to mapping ambiguity. Note that it is still possible to identify increase in SD copy number using read depth signature [3, 40], however, no method yet exists to *anchor* a new SD (i.e. find their insertion locations).

Here we present **novel algorithms** to discover large (> 40Kbp) direct and inverted interspersed SDs using Linked-Read sequencing data. We redesign and extend upon VALOR and use split molecule and read pair signatures to detect SDs and estimate the insertion sites of the new SD paralogs, and further include read depth signature to filter potential false positives caused by incorrect mappings. We implemented our new algorithms as the VALOR$_2$ software package. Briefly, VALOR$_2$ differs from the former version of VALOR through: 1) it can characterize segmental duplications in both direct and inverted orientation, 2) incorporates read depth information to improve predictions and reduce false calls, and 3) provides full support to alignment files (i.e., BAM) generated from 10xG Linked-Read data sets.

Using simulated data sets we show that VALOR$_2$ achieves high precision and recall (94% and 82%, respectively) for segmental duplications, and 98% and 76% for large inversions. We also applied VALOR$_2$ to the genome of NA12878 sequenced with the 10xG platform [27] and identified 5 direct, and 9 inverted segmental duplications.

2 Methods

We have previously described an earlier version of VALOR$_2$ that uses split molecules and read pair signature to detect inversions [11]. Here we describe novel formulations, algorithms and optimizations to characterize large (> 80Kbp) inversions and (> 40Kbp) *segmental duplications* in both direct and inverted orientation. We depict the split molecule and read pair sequence signatures for these types of large SVs in Figure 1.

2.1 Glossary

Here we define several terms that we use in this manuscript:

- *molecule*: a large molecule (30-50 Kbp) that was barcoded and pooled using the 10xG platform. Here we refer to the physical entity.
- *submolecule*: a molecule identified *in silico* by the VALOR$_2$ algorithm by analyzing read map locations.
- *candidate split*: a pair of submolecules with the same barcode that potentially signal a SV event.
- *split molecule pair*: a pair of candidate splits with different barcodes that potentially signal the same SV event.

2.2 Overview of the VALOR$_2$ algorithm

VALOR$_2$ depends on only the alignment files (i.e., BAM) with the necessary barcode information generated with Long Ranger, BWA-MEM, or a similar read mapper. Briefly, VALOR$_2$ first tries to identify the underlying large molecules separately for each barcode, which we call *submolecules*. In this step, we do not consider reads that map to satellite regions, and we discard very short submolecules. Two identified submolecules are paired together (called *candidate splits*) if the summation of their span is $\leq \mu_{\text{molecule}} + 3\sigma_{\text{molecule}}$ where μ_{molecule} is the average and σ_{molecule} is the standard deviation of the inferred submolecule sizes. Next, VALOR$_2$ removes those candidate splits with no read pair support. It also discards those that signal a duplication event without read depth support. Additionally, any candidate splits that span assembly gaps are removed from consideration. VALOR$_2$ then 1) pairs candidate splits with different barcodes that likely signal the same SV event (*split molecule pairs*), and 2) models the split molecule pairs as vertices in a graph and solves the maximal quasi clique problem [6]. In this graph, edges represent overlap (i.e., "agreement") between two split molecule pairs. Finally, VALOR$_2$ reports SVs that are supported by at least two pairs of split molecules.

Below, we present the details for each step in the VALOR$_2$ algorithm.

2.3 Molecule Recovery

The first step of the VALOR$_2$ algorithm involves identification (or, recovery) of the large molecules from mapped data. Initially, we call the intervals returned by this recovery as *submolecules*. For this purpose we use a sliding window approach to greedily group reads with the same barcode that are mapped in close proximity (Algorithm 1). Here we only consider concordantly mapped read pairs, and we take the full span of a read pair as a *fragment*. For each barcode, we scan each chromosome and merge together fragments if they are within a user-defined distance T, or if a new fragment is within distance Q from the leftmost fragment in a re-identified submolecule. We use $Q = 80,000$ and $T = 10,000$ by default[2], determined by parameter sweeping. Finally, we remove very short submolecules (<3 Kbp by default) that correspond to less than 1/10 of expected average molecule size from consideration.

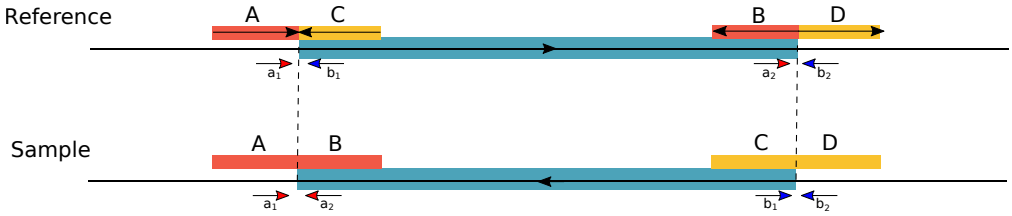
2.4 Clustering using SV graph

We first record all pairs of submolecules that share the same barcode and map to the same chromosome as *candidate splits*, and then compare all possible pairs of candidate splits across different barcodes to find those that signal an inversion or a duplication, termed *split molecule pairs* (see Figure 1 for the depiction of candidate splits and split molecule pairs). We limit inversion predictions and the duplication size by the largest inversion size we can find in the literature [4] (≈7 Mbp). Next, we test whether the split molecule pairs are supported by read pair signature (Figure 1). Here we require at least 3 read pairs to signal the same SV event and we remove candidate splits with insufficient support from consideration.

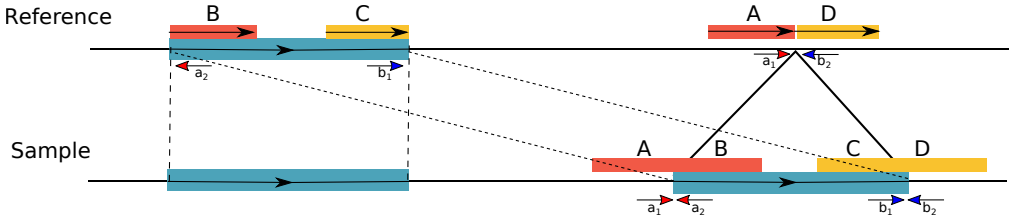
We construct an SV graph G as follows (Figure 2). We denote each remaining split molecule pair as a vertex in G, and we create an edge between two vertices if their corresponding split molecule pairs signal the same SV event. Finally on the resulting graph we find clusters of read pair supported split molecule pairs by approximately solving the maximal clique problem using the

[2]Corresponds to $2 \cdot \mu_{\text{molecule}}$ and $\mu_{\text{molecule}}/2$, respectively.

a)



b)



c)

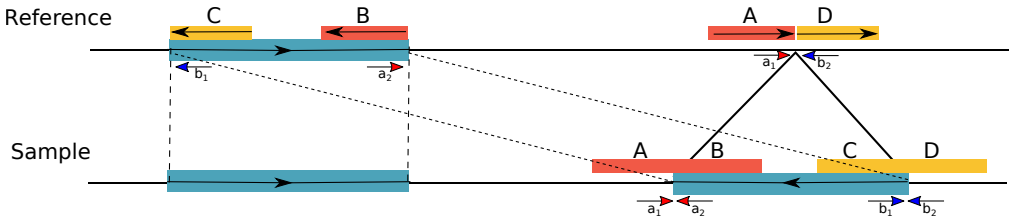


Figure 1: Split molecule and read pair sequence signatures used in VALOR$_2$. a) Inversion, b) interspersed duplication in direct orientation, c) inverted duplication. In each case, the large molecules that span the SV breakpoints are split into two mapped regions. Note that, it is not possible to determine the mapped strand of the split molecules shown here. From the perspective of the reference genome (i.e., mapping), A,B,C,D are defined as *submolecules*, A/B and C/D pairs are *candidate splits*, and A/B-C/D quadruple is a *split molecule pair*.

quasi-clique formulation [6]. Here a quasi clique is defined as an approximate clique with V vertices and $\gamma \cdot \binom{|V|}{2}$ edges, where γ is a user-defined parameter, which we set to $\gamma = 0.6$ by default. Each quasi clique defines a putative SV event.

Algorithm 1 Molecule recovery.

Require: Alignments in BAM format with barcodes, look-ahead parameter (Q), extend parameter (T).

Ensure: Set of submolecules $S_M = \{M_1, M_2 \ldots, M_k\}$ (value of k is unknown and will be determined by the algorithm)

```
S_M ← ∅
i ← 1
for each chromosome c do
    for each barcode b do
        M_i = ∅
        for l = 1 to length(c) do
            if short fragment f with barcode b maps to c[l] then
                if M_i = ∅ then
                    M_i ← f
                    s(M_i) ← s(f)
                    e(M_i) ← e(f)
                else if (s(f) < s(M_i) + Q) or (s(f) < e(M_i) + T) then
                    M_i ← M_i ∪ f
                    e(M_i) ← e(f)
                else
                    S_M ← S_M ∪ M_i
                    i ← i + 1
                end if
            end if
        end for
    end for
end for
return S_M
```

$s(f)$ denotes the map start location and $e(f)$ denotes the map end location of fragment f.

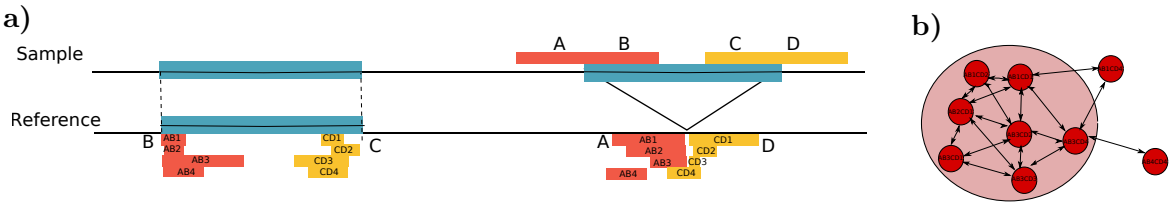


Figure 2: Building the SV graph from split molecule pairs. a) Four pairs of split molecules that signal a segmental duplication. b) Corresponding SV graph, where each vertex denotes a pair of submolecules that signal the duplication, and edges show "agreement" between pairs. The shaded area corresponds to the quasi-clique selected as representative of the putative SV.

We identify inversions breakpoints with two coordinates, and duplications with three coordinates. The third coordinate for a duplication event denotes the insertion breakpoint, for which we

provide a confidence interval.

2.5 Molecule depth filtering

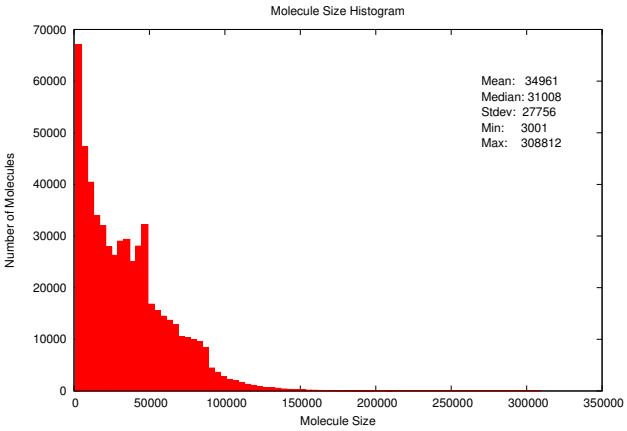


Figure 3: Molecule size histogram mapped to chromosome 16 as observed in the Linked-Read sequencing data generated from the genome of the NA12878 sample [27].

Although there are only a small number of molecules that share the same barcode (2-30), it is still possible that two or more different molecules originate from the same chromosome. Additionally, the molecule sizes do not follow Gaussian, Poisson, or a similar distribution (Figure 3), thus it is not possible to distinguish true split molecules from "normal" but short molecules. The read pair sequence signature is not entirely reliable either due to the mis-mapping artifacts within or around repeats and duplications. We, therefore, apply additional filtering on duplication calls based on "molecule depth". We reason that the number of molecules that originate from segmental duplications must be higher than the genome-wide average, similar to the traditional read depth signature [5, 3]. In this step, we first calculate the average molecule depth (μ_{depth}) and standard deviation (σ_{depth}) in the entire genome. We then discard segmental duplication predictions with molecule depth $< \mu_{\text{depth}} + \sigma_{\text{depth}}$.

3 Results

We tested VALOR$_2$ using both simulated and real data sets to compare the precision and recall rates of VALOR$_2$ with one other tool that use Linked-Read sequencing (Long Ranger [23], and two tools that use only WGS data sets (DELLY [31] and LUMPY [19]). However, VALOR$_2$ is the only tool that can characterize interspersed duplications, therefore we limit our comparison to only inversions, and evaluate VALOR$_2$'s performance on duplications using simulations. We find that VALOR$_2$ is complementary to other methods in inversion calls as VALOR$_2$ aims to find larger (>80Kbp) inversions, while the other tools focus on smaller (<100 Kbp) SVs.

3.1 Simulation experiments

We used VarSim [28] to generate a simulated diploid human genome. Our simulation included variants of various lengths and types: 2.8 million SNPs, ≈195,000 indels, and ≈ 5,000 SVs (>50 bp, up to 6 Mbps). We found that VarSim only generates tandem duplications, therefore we randomly changed a subset of simulated tandem duplications to interspersed in the simulated VCF file, assigned random insertion breakpoints, and then applied changes to the reference. We then generated Illumina WGS reads at 40X depth of coverage using ART [14], and 10xG Linked-Reads at 50X coverage using LRSim [22]. The 10xG Linked-Reads simulation has extra coverage to account for the barcode sequences that are part of the read and other losses as also described in [23].

We mapped the simulated reads to the human reference genome (GRCh37) using BWA-MEM [20] for WGS, and Long Ranger for 10xG data sets. We then applied the standard BAM processing that includes sorting with SAMtools [21] and marking duplicates with Sambamba [41]. We used VALOR$_2$ and Long Ranger to generate SV call sets from the 10xG simulation, and DELLY and LUMPY to call variants using the WGS simulation. We limited our comparison to only large SVs (>80Kbp for inversions, >40Kbp for duplications) and we required >50% reciprocal overlap between the simulation and the prediction for inversions and the duplicated segments using BEDtools [30]. We also require the inferred insertion breakpoint is within a distance of $\mu_{molecule}/2$ (in simulation experiments $\mu_{molecule} = 40$ Kbp) of the simulation breakpoint to consider a duplication to be correctly predicted.

Table 1: Prediction performance evaluation using simulated structural variants.

Variant	Tool	# Simulated	# Predicted	True	False	Precision	Recall
Duplication (direct)	VALOR$_2$	78	66	61	5	0.92	0.78
Duplication (inverted)	VALOR$_2$	56	51	49	2	0.96	0.88
Inversion	VALOR$_2$	94	65	64	1	0.98	0.76
	LUMPY	94	42	44	4	0.90	0.47
	DELLY	94	896	79	761	0.15	0.84
	Long Ranger	94	92	68	27	0.71	0.72

We evaluate prediction performance of only large (>80Kbp for inversions, >40Kbp for duplications) SVs. Note that LUMPY, DELLY, and Long Ranger are not able to call interspersed duplications, thus we provide only the inversion prediction benchmark. Precision is calculated as $\frac{TP}{TP+FP}$, and recall is defined as $\frac{TP}{TP+FN}$, where TP: true positive, FP: false positive, FN: false negative.

We present the prediction performance of the tools we tested in Table 1. We found that VALOR$_2$ is able to correctly predict >82% of large duplications (inverted and direct combined), and 78% of large inversions, while maintaining $92 - 96\%$ precision for duplications and 82% precision for inversions. Long Ranger, the other algorithm that used Linked-Reads, correctly predicted 72% of the inversions with 71% precision. Of the WGS-based tools, DELLY achieved high sensitivity for inversions and it was able to correctly predict 84% of large inversions, however it suffered from very low precision (15%). On the contrary, LUMPY achieved high precision (90%), but it was able to discover only 47% of the simulated inversions. This is likely because neither DELLY nor LUMPY were optimized to find such large inversion events. Overall, VALOR$_2$ performed the best in terms of precision and recall balance in the simulation experiment.

Table 2: Inversion prediction performance evaluation in the NA12878 genome using InvFEST database.

	Called	InvFEST-Valid.	InvFEST-Pred.	InvFEST-All
$VALOR_2$	135	6	5	17
Long Ranger	476	1	10	14
LUMPY	7	0	0	0
DELLY	2,340	1	6	24

Here we only focus on large (> 80Kbp) inversions in the NA12878 genome. InvFEST-Valid.: validated inversions in the genome of NA12878, InvFEST-Pred.: predicted inversions in the genome of NA12878, InvFEST-All: all inversions reported in the InvFEST database [25], except those that are annotated as *unreliable prediction*.

Table 3: Segmental duplications predicted in the NA12878 genome using $VALOR_2$.

Chr	**Start**	**End**	**Type**	**Target**	**No. of genes**
1	120,600,786	120,692,870	Direct	1q21.1	1
1	144,832,884	145,751,706	Direct	1p22.3	25
1	145,062,336	145,116,024	Direct	1p11.2	
16	86,451,165	86,498,200	Direct	16q11.2	
17	21,522,544	21,551,840	Direct	17p11.2	
1	17,019,657	17,111,181	Inverted	1q42.3	4
1	145,983,326	146,027,347	Inverted	1p22.3	3
4	15,160	67,199	Inverted	4q35.2	2
8	2,189,297	2,290,508	Inverted	8p23.2	
10	46,965,140	47,022,150	Inverted	10q11.22	2
11	4,250,956	4,331,367	Inverted	11p15.4	
16	21,542,145	21,593,639	Inverted	16p12.2	
16	22,543,245	22,709,969	Inverted	16p12.2	2
X	153,423,995	153,485,001	Inverted	Xq28	3

3.2 NA12878 genome

We also compared the performance of $VALOR_2$ with that of Long Ranger on the NA12878 germline genome, along with other commonly used SV callers (DELLY AND LUMPY). NA12878 phased variant calls were obtained from 10X Genomics on their Chromium platform. From these we extracted 476 large inversions, 14 of which were also present in the InvFEST database (Table 2) but only one was experimentally validated. When given the same data, $VALOR_2$ was able to call 135 inversions, a higher percentage of which were found in the InvFEST database that also included six experimentally validated inversions. Of the four tools we tested, $VALOR_2$ had the largest number of validated inversions within its call set while predicting the second lowest number of total inversions (only LUMPY, which only called 7 inversions, has fewer). This result further highlights the IH: superior precision and recall of $VALOR_2$. DELLY was able to identify 24 inversions in the NA12878 genome which were also in the InvFEST database but called a total of 2,340 inversions. A majority of these calls were only predicted by DELLY and due to a lack of precision, may signify an over-representation of false positives (Figure 4). $VALOR_2$ was very useful in identifying

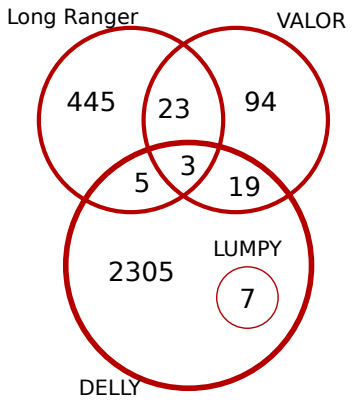


Figure 4: Comparison of the inversion predictions (> 80 Kb) by VALOR$_2$, Long Ranger, DELLY, and LUMPY in the NA12878 genome.

large scale duplications by exploiting linked read information in the NA12878 sequencing data. We predicted multiple direct segmental duplications and inverted duplications with chromosomes 1 and 16 containing both classes of duplications (Table 3).

3.3 CHM1 genome

Finally we tested VALOR$_2$ using Linked-Read data set of a haploid human genome cell line (CHM1 [15, 39, 7]). We used VALOR$_2$ to find inversions and segmental duplications. Overall, VALOR$_2$ characterized 133 inversions (>80 Kbp), 14 inverted and 22 direct segmental duplications (>40 Kb). Unfortunately there are no gold standard data sets for SDs for this genome available in the literature, and the largest previously reported inversion in [7] is 36 Kbp, which is less than the smallest inversion that VALOR$_2$ predicts. We therefore compared only with the large inversions in the InvFEST database, and we found that 10% (16/117) of VALOR$_2$ predictions were present in InvFEST (Figure 5).

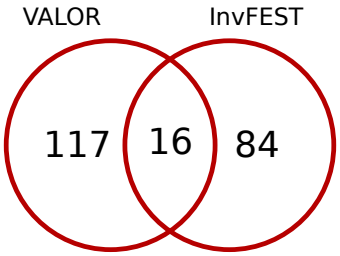


Figure 5: Intersection of all inversions reported by InvFEST (validated or predicted) with VALOR$_2$ predictions on CHM1 genome.

4 Discussion

In this work, we presented novel algorithms to effectively utilize the encoded long-range information in Linked-Read data for the purpose of characterizing large-scale structural variations. The current state of the art SV detection techniques using Linked-Read such as Long Ranger or GROC-SV are optimized for certain range of SV sizes. For example, GROC-SVs achieves the best sensitivity for events in the range of (30 Kb-100 Kb). However, our technique, VALOR can detect events of a size larger than 100 Kb, including segmental duplications. A future direction for our study is to integrate additional techniques such as local assembly to characterize smaller-scale SVs (i.e. starting from only 50 bp) and to resolve SV breakpoints more precisely. Although single molecule techniques such as Oxford Nanopore (ONT) promise to generate reads in the order of 100 Kb in length, the current error rate, the cost, and the low throughput make them prohibitive in practice. As these techniques get developed and more cost effective, we can explore them not only for the purpose of further validation of our method but also for devising integrative techniques to fully resolve the complexity of repetitive DNA common in mammalian genomes. Another future direction to get more confirmation of our call sets will be using techniques such as fluorescent in situ hybridization (FISH).

Acknowledgements

We thank H. İ. Özercan, A. Soylev and D. Meleshko for computational support.

Funding This work was supported by a grant by TÜBİTAK (215E172), an EMBO Installation Grant (IG-2521), and a Marie Curie Career Integration Grant (303772) to C.A. This work was also supported by start-up funds (Weill Cornell Medicine) to I.H. C.R. received support from the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant 1T32GM083937). The authors also acknowledge the Computational Genomics Summer Institute funded by NIH grant GM112625 that fostered international collaboration among the groups involved in this project.

Conflict of Interest None to declare.

Data availability

The NA12878 genome sequenced with the 10x Genomics Platform is available via the Genome in a Bottle Project [49] FTP site at `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/10Xgenomics_ChromiumGenome_LongRanger2.1_09302016/NA12878_hg19/` Short read sequencing data for the same genome can be downloaded from the Illumina Platinum Genomes Project at https://www.illumina.com/platinumgenomes.html. The CHM1 genome generated with 10xG Linked-Reads is available at https://support.10xgenomics.com/de-novo-assembly/datasets/2.0.0/chm.

References

[1] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan

Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature methods*, 14:1083–1086, November 2017.

[2] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, May 2011.

[3] Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S. Cenk Sahinalp, Richard A Gibbs, and Evan E Eichler. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*, 41(10):1061–1067, Oct 2009.

[4] Francesca Antonacci, Jeffrey M Kidd, Tomas Marques-Bonet, Mario Ventura, Priscillia Siswara, Zhaoshi Jiang, and Evan E Eichler. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet*, 18(14):2555–2566, Jul 2009.

[5] Jeffrey A Bailey, Zhiping Gu, Royden A Clark, Knut Reinert, Rhea V Samonte, Stuart Schwartz, Mark D Adams, Eugene W Myers, Peter W Li, and Evan E Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, Aug 2002.

[6] Mauro Brunato, Holger H. Hoos, and Roberto Battiti. *On Effectively Finding Maximal Quasi-cliques in Graphs*, pp. 41–55. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[7] Mark J P. Chaisson, John Huddleston, Megan Y. Dennis, Peter H. Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M. Landolin, John A. Stamatoyannopoulos, Michael W. Hunkapiller, Jonas Korlach, and Evan E. Eichler. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517:608–611, Jan 2015.

[8] David C. Danko, Dmitry Meleshko, Daniela Bezdan, Christopher Mason, and Iman Hajirasouliha. Minerva: an alignment and reference free approach to deconvolve linked-reads for metagenomics. *bioRxiv*, 2017.

[9] Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*, 11(6):446–450, Jun 2010.

[10] Adam C. English, William J. Salerno, and Jeffrey G. Reid. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, 15:180, 2014.

[11] Marzieh Eslami Rasekh, Giorgia Chiatante, Mattia Miroballo, Joyce Tang, Mario Ventura, Chris T Amemiya, Evan E Eichler, Francesca Antonacci, and Can Alkan. Discovery of large genomic inversions using long range information. *BMC Genomics*, 18:65, January 2017.

[12] Iman Hajirasouliha, Fereydoun Hormozdiari, Can Alkan, Jeffrey M Kidd, Inanc Birol, Evan E Eichler, and S. Cenk Sahinalp. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, 26(10):1277–1283, May 2010.

[13] Fereydoun Hormozdiari, Can Alkan, Evan E Eichler, and S. Cenk Sahinalp. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res*, 19(7):1270–1278, Jul 2009.

[14] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, Feb 2012.

[15] John Huddleston, Swati Ranade, Maika Malig, Francesca Antonacci, Mark Chaisson, Lawrence Hon, Peter H. Sudmant, Tina A. Graves, Can Alkan, Megan Y. Dennis, Richard K. Wilson, Stephen W. Turner, Jonas Korlach, and Evan E. Eichler. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res*, 24(4):688–696, Apr 2014.

[16] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36:338–345, April 2018.

[17] Jeffrey M Kidd, Gregory M Cooper, William F Donahue, Hillary S Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague, Can Alkan, Francesca Antonacci, Eric Haugen, Troy Zerr, N. Alice Yamada, Peter Tsang, Tera L Newman, Eray Tüzün, Ze Cheng, Heather M Ebling, Nadeem Tusneem, Robert David, Will Gillett, Karen A Phelps, Molly Weaver, David Saranga, Adrianne Brand, Wei Tao, Erik Gustafson, Kevin McKernan, Lin Chen, Maika Malig, Joshua D Smith, Joshua M Korn, Steven A McCarroll, David A Altshuler, Daniel A Peiffer, Michael Dorschner, John Stamatoyannopoulos, David Schwartz, Deborah A Nickerson, James C Mullikin, Richard K Wilson, Laurakay Bruhn, Maynard V Olson, Rajinder Kaul, Douglas R Smith, and Evan E Eichler. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, May 2008.

[18] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, Bruce E Taillon, Zhoutao Chen, Andrea Tanzer, A. C Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P Carter, Matthew E Hurles, Sherman M Weissman, Timothy T Harkins, Mark B Gerstein, Michael Egholm, and Michael Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, Oct 2007.

[19] Ryan M. Layer, Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*, 15(6):R84, 2014.

[20] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013.

[21] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[22] Ruibang Luo, Fritz J Sedlazeck, Charlotte A Darby, Stephen M Kelly, and Michael C Schatz. LRSim: a linked-reads simulator generating insights for better genome partitioning. *Computational and structural biotechnology journal*, 15:478–484, 2017.

[23] Patrick Marks, Sara Garcia, Alvaro Martinez Barrio, Kamila Belhocine, Jorge Bernate, Rajiv Bharadwaj, Keith Bjornson, Claudia Catalanotti, Josh Delaney, Adrian Fehr, et al. Resolving the full spectrum of human genome variation using linked-reads. *BioRxiv*, p. 230946, 2017.

[24] Tomas Marques-Bonet, Jeffrey M Kidd, Mario Ventura, Tina A Graves, Ze Cheng, LaDeana W Hillier, Zhaoshi Jiang, Carl Baker, Ray Malfavon-Borja, Lucinda A Fulton, Can Alkan, Gozde Aksay, Santhosh Girirajan, Priscillia Siswara, Lin Chen, Maria Francesca Cardone, Arcadi Navarro, Elaine R Mardis, Richard K Wilson, and Evan E Eichler. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, 457(7231):877–881, Feb 2009.

[25] Alexander Martínez-Fundichely, Sònia Casillas, Raquel Egea, Miquel Ràmia, Antonio Barbadilla, Lorena Pantano, Marta Puig, and Mario Cáceres. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res*, 42(Database issue):D1027–D1032, Jan 2014.

[26] Paul Medvedev, Monica Stanciu, and Michael Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, 6(11 Suppl):S13–S20, Nov 2009.

[27] Yulia Mostovoy, Michal Levy-Sakin, Jessica Lam, Ernest T Lam, Alex R Hastie, Patrick Marks, Joyce Lee, Catherine Chu, Chin Lin, Željko Džakula, Han Cao, Stephen A Schlebusch, Kristina Giorda, Michael Schnall-Levin, Jeffrey D Wall, and Pui-Yan Kwok. A hybrid approach for de novo human genome sequence assembly and phasing. *Nature methods*, 13:587–590, July 2016.

[28] John C. Mu, Marghoob Mohiyuddin, Jian Li, Narges Bani Asadi, Mark B. Gerstein, Alexej Abyzov, Wing H. Wong, and Hugo Y K. Lam. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, 31(9):1469–1471, May 2015.

[29] Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, August E. Woerner, Timothy D. O'Connor, Gabriel Santpere, Alexander Cagan, Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch, Asger Hobolth, Anders E. Halager, Maika Malig, Jessica Hernandez-Rodriguez, Irene Hernando-Herraez, Kay Prüfer, Marc Pybus, Laurel Johnstone, Michael Lachmann, Can Alkan, Dorina Twigg, Natalia Petit, Carl Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo, Marc Dabad, Michael L. Wilson, Laurie Stevison, Cristina Camprubí, Tiago Carvalho, Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela Kondova, Ronald E. Bontrop, Anne Pusey, Felix Lankester, John A. Kiyang, Richard A. Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pascal Gagneux, David Comas, Hans Siegismund, Julie Blanc, Lidia Agueda-Calpena, Marta Gut, Lucinda Fulton, Sarah A. Tishkoff, James C. Mullikin, Richard K. Wilson, Ivo G. Gut, Mary Katherine Gonder, Oliver A. Ryder, Beatrice H. Hahn, Arcadi Navarro, Joshua M. Akey, Jaume Bertranpetit, David Reich, Thomas Mailund,

Mikkel H. Schierup, Christina Hvilsom, Aida M. Andrés, Jeffrey D. Wall, Carlos D. Bustamante, Michael F. Hammer, Evan E. Eichler, and Tomas Marques-Bonet. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, Jul 2013.

[30] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

[31] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, Sep 2012.

[32] Anna Ritz, Ali Bashir, Suzanne Sindi, David Hsu, Iman Hajirasouliha, and Benjamin J. Raphael. Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, 30(24):3458–3466, Dec 2014.

[33] Jeong-Sun Seo, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, Han Cao, Ji-Young Yun, Jihye Kim, Junho Kuk, Gun Hwa Park, Juhyeok Kim, Hanna Ryu, Jongbum Kim, Mira Roh, Jeonghun Baek, Michael W Hunkapiller, Jonas Korlach, Jong-Yeon Shin, and Changhoon Kim. De novo assembly and phasing of a Korean human genome. *Nature*, 538:243–247, October 2016.

[34] Suzanne Sindi, Elena Helman, Ali Bashir, and Benjamin J Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25:i222–i230, June 2009.

[35] Daniel A Skelly, Galen T Squiers, Micheal A McLellan, Mohan T Bolisetty, Paul Robson, Nadia A Rosenthal, and Alexander R Pinto. Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. *Cell reports*, 22:600–610, January 2018.

[36] Arda Soylev, Can Kockan, Fereydoun Hormozdiari, and Can Alkan. Toolkit for automated and rapid discovery of structural variants. *Methods*, 129:3–7, 2017.

[37] Noah Spies, Ziming Weng, Alex Bishara, Jennifer McDaniel, David Catoe, Justin M Zook, Marc Salit, Robert B West, Serafim Batzoglou, and Arend Sidow. Genome-wide reconstruction of complex structural variants using read clouds. *Nature methods*, 14:915–920, September 2017.

[38] Paweł Stankiewicz and James R. Lupski. Structural variation in the human genome and its role in disease. *Annu Rev Med*, 61:437–455, 2010.

[39] Karyn Meltz Steinberg, Valerie A. Schneider, Tina A. Graves-Lindsay, Robert S. Fulton, Richa Agarwala, John Huddleston, Sergey A. Shiryev, Aleksandr Morgulis, Urvashi Surti, Wesley C. Warren, Deanna M. Church, Evan E. Eichler, and Richard K. Wilson. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res*, 24(12):2066–2076, Dec 2014.

[40] Peter H Sudmant, Jacob O Kitzman, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, Laurakay Bruhn, Jay Shendure, 1000 Genomes Project, and Evan E Eichler. Diversity of human copy number variation and multicopy genes. *Science*, 330(6004):641–646, Oct 2010.

[41] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of ngs alignment formats. *Bioinformatics*, 31(12):2032–2034, Jun 2015.

[42] Eray Tuzun, Andrew J Sharp, Jeffrey A Bailey, Rajinder Kaul, V. Anne Morrison, Lisa M Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, Maynard V Olson, and Evan E Eichler. Fine-scale structural variation of the human genome. *Nat Genet*, 37(7):727–732, Jul 2005.

[43] Neil I Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe. Direct determination of diploid genome sequences. *Genome research*, 27:757–767, May 2017.

[44] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19:15, February 2018.

[45] Li C Xia, John M Bell, Christina Wood-Bouwens, Jiamin J Chen, Nancy R Zhang, and Hanlee P Ji. Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic acids research*, November 2017.

[46] Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, Nov 2009.

[47] Sarah Yeo, Lauren Coombe, René L Warren, Justin Chu, and Inanç Birol. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, 34:725–731, March 2018.

[48] Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J Makarewicz, Yuan Li, Phillip Belgrader, Andrew D Price, Adam J Lowe, Patrick Marks, Gerard M Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E Birch, Steven W Short, Keith P Bjornson, Pranav Patel, Erik S Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K Lockwood, David Stafford, Joshua P Delaney, Indira Wu, Heather S Ordonez, Susan M Grimes, Stephanie Greer, Josephine Y Lee, Kamila Belhocine, Kristina M Giorda, William H Heaton, Geoffrey P McDermott, Zachary W Bent, Francesca Meschi, Nikola O Kondov, Ryan Wilson, Jorge A Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N Fehr, Adrian Chan, Serge Saxonov, Kevin D Ness, Benjamin J Hindson, and Hanlee P Ji. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*, 34:303–311, March 2016.

[49] Justin M. Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. *Nat Biotechnol*, 32(3):246–251, Mar 2014.