Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Sequence analysis

VaPoR: a high-speed validation approach for structural variation using long-read sequencing technology.

Xuefang Zhao[1], Alexandra M. Weber[1], and Ryan E. Mills[1,2,*]

[1]Department of Computational Medicine and Bioinformatics and [2]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109 USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Although there are numerous algorithms that have been developed to identify structural variation (SVs) in genomic sequences, there is a dearth of approaches that can be used to evaluate their results. The emergence of new sequencing technologies that generate longer sequence reads can, in theory, provide direct evidence for all types of SVs regardless of the length of region through which it spans. However, current efforts to use these data in this manner require the use of large computational resources to assemble these sequences as well as manual inspection of each region. Here, we present VaPoR, a highly efficient algorithm that autonomously validates large SV sets using long read sequencing data. We assess of the performance of VaPoR on both simulated and real SVs with regards to various features including accuracy and sensitivity of breakpoint evaluation and report a high fidelity rate.
Availability: https://github.com/mills-lab/VaPoR
Contact: remills@umich.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Structural variants (SVs) are one of the major forms of genetic variation in humans and have been revealed to play important roles in various diseases including cancers and neurological disorders (Brand, et al., 2014; Stankiewicz and Lupski, 2010). Various approaches have been developed and applied to paired-end sequencing to detect SVs in whole genomes (Layer, et al., 2014; Rausch, et al., 2012; Zhao, et al., 2016), however individual algorithms often exhibit complementary strengths that can be leveraged in aggregate. Subsequently, investigators typically apply and compare multiple algorithms to their samples and design their own selection strategy according to the sensitivity and specificity requirements of specific research, while using orthogonal evidence from each approach as the only evidence that an actual structural rearrangement is present (Sudmant, et al., 2015). The emergence of long read sequencing technology, eg. Single Molecule Real-Time (SMRT) sequencing from Pacific Biosciences (PacBio), can provide direct evidence for the presence of an SV. Current strategies make use of de novo assembly to create large contigs that can be cross-referenced with a putative SV using manual inspection of the subsequent recurrence (dot) plot (Huddleston, et al., 2016). These types of dot plots have been used for decades to examine the specific features of sequence alignments (Gibbs and McIntyre, 1970), however they require manual curation and, coupled with the computational costs of sequence assembly, are time-consuming and inefficient at scale for the high throughput validation of large sets of SVs.

Here, we present a high-speed long read based assessment tool, VaPoR, that scores each SV prediction by autonomously analyzing the recurrence of windows within a local read against the reference genome in both their original and rearranged format according to the prediction. A positive score of each read on the altered reference, normalized against the score of the read on the original reference, supports the predicted structure. A baseline model is constructed as well by interrogating the reference sequence against itself at the query location. We show that our approach is able to quickly and accurately distinguish true from false positive predictions of both simple and complex SVs and is also able to assess the breakpoint accuracy of individual algorithms.

2 Methods

VaPoR takes in aligned sequence reads in BAM format and predicted SVs (>50bp) in various formats including VCF and BED. Evaluation of an SV is performed by comparing long reads that go through the event against reference sequences in two formats: (a) the original human reference to which the sample is aligned and (b) a modified reference sequence altered to match the predicted structural rearrangement. A recurrence matrix is then derived by sliding a fixed-size window with 1bp step through each read to mark positions where the read sequence and reference are identical. The matching patterns are then assessed as to the validity of the SV as described below and a validation score is reported. Given the large variance of SVs lengths, each SV is stratified into one of two groups, each with respective statistical models implemented: smaller SVs that can be completely encompassed within multiple (>10 by default) long sequences and larger events that are rarely covered by individual long reads. VaPoR workflow is briefly summarized in Fig. 1.

Small Variants Assessment:

For an SV k in sample s that is covered by n reads, the recurrence matrix between each read and the reference sequences in original (R_o) and altered (R_a) format is calculated. The vertical distance between each record ($x_{i,k,s,Rx}$, $y_{i,k,s,Rx}$) in matrix x and the diagonal ($x_{i,k,s,Rx}$, $x_{i,k,s,Rx}$) line is calculated as $d_{i,k,s,Rx}$= abs($x_{i,k,s,Rx}$ - $y_{i,k,s,Rx}$), and the average distance of all records would be exported as the score of each matrix:

$$Score_{k,s,Rx} = \sum_{i=1}^{m} d_{i,k,s,Rx} \; / \; m,$$

where m is the total number of records in the matrix. Sequences that share higher identity with the read shall have a lower $Score_{k,s,Rx}$, such that the score of each read is normalized as:

$$Score_{k,s,R} = Score_{k,s,R_o} \; / \; Score_{k,s,R_a} - 1,$$

where a positive $Score_{k,s,R}$ represents the superiority of the predicted structure versus the original and vise versa for negative $Score_{k,s,R}$, with one exceptional case where there exists a duplicated structure in the predicted SV such that the predicted structure would show higher $Score_{k,s,R}$ due to the multi-alignment of duplicated segments. To correct for duplications, VaPoR adopts the directed distance $d_{i,k,s,Rx}$ = $x_{i,k,s,Rx}$ - $y_{i,k,s,Rx}$ instead such that the distance contributed by centrosymmetric duplicated segments would offset each other.

Large Variants Assessment:

For larger SVs where there are few, if any, long reads that can transverse the predicted SV, VaPoR assesses the quality of each predicted junction instead using:

$$Score_{k,s,Rx} = \frac{\sum_{i=1}^{m} I = \begin{cases} 1, if\ abs\left(x_{i,k,s,Rx} - y_{i,k,s,Rx}\right) < 0.15 * x_{i,k,s,Rx} \\ 0, otherwise \end{cases}}{m},$$

where a larger $Score_{k,s,Rx}$ represents higher similarity between the read and the reference sequence. The normalized scores of each read is then defined as:

$$Score_{k,s,R} = Score_{k,s,R_a} \; / \; Score_{k,s,R_o} - 1,$$

VaPoR Score Calculation:

With a score assigned to each read spanning through the predicted structural variants, the VaPoR score ($Score_{k,s}$) is summarized as:

$$Score_{k,s} = \frac{\sum_{R=1}^{n} I = \begin{cases} 1, if\ \ Score_{k,s,R} > 0 \\ 0, otherwise \end{cases}}{n}$$

to represent the proportion of long reads supporting predicted structure. The highest supportive score (max($Score_{k,s,R}$)) is also reported as a reference for users to meet the specific requirement of their study design, for which we recommend 0.1 as the cutoff.

Flexible window size:

By default, VaPoR uses a window size of 10bp and requires an exact match between sequences, though these can be changed to user-defined parameters. However, many regions of the genome contain repetitive sequences resulting in an abundance of spurious matches in the recurrence matrix, thus introducing bias to the assessment. To address this, VaPoR adopts a quality control step by iteratively assessing the reference sequence against itself and tabulating the proportion of matches along the diagonal. The window size initially starts at 10bp and iteratively increases by 10bp until either (a) the proportion of matches on the diagonal exceeds 40% and the current window size is kept or (b) the window size exceeds 40bp whereby the event will be labeled as 'non-assessable and excluded from the evaluation.

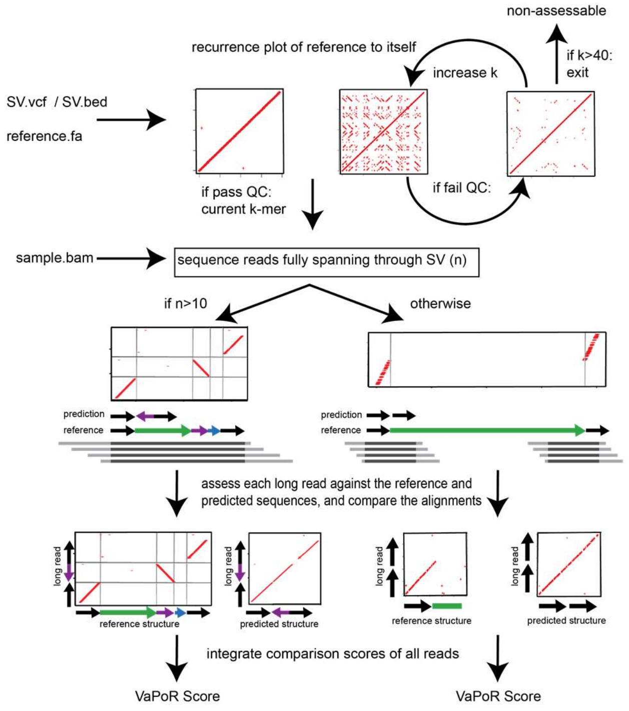

Fig 1. Flowchart describing the VaPoR algorithm. As input, the algorithm requires a set of structural variants in either VCF or BED format, a series of long reads and/or sequence contigs in BAM format, and the corresponding reference sequence. VaPoR then interrogates each variant individually at its corresponding reference location, assesses the quality of the region and assigns a score.

3 Results

We have assessed performance of VaPoR on both simulated and real genomes from the 1000 Genomes Project to assess the following characteristics: (a) sensitivity and false discovery rate on validating structural variants in simple and complex structures, (b) sensitivity of VaPoR on validating different levels of predicted breakpoint efficacy and (c) time and computational cost of VaPoR.

3.1 VaPoR on Simulated Data

Non-overlapping simple deletions, inversions, insertions and duplications as well as complex structural variants as previously categorized (Zhao, et al., 2016) were independently incorporated into GRCh38 in both heterozygous and homozygous states, excluding regions of the genome known to produce artifact signals as described from the ENCODE project blacklist (Consortium, 2012). Detailed descriptions of each simulated SV types simulated are summarized in Supplementary Tables 1- 3. We applied PBSIM (Ono, et al., 2013) to simulate the modified reference sequences to different read depth ranging from 2X to 70X with parameters difference-ratio 5:75:20. length-mean 12000, accuracy-mean 0.85 and model_qc model_qc_clr.

We applied VaPoR to the simulated SVs and first assessed the proportion of SVs that VaPoR is capable of interrogating (i.e. passed VaPoR QC) and found that VaPoR can successfully evaluate >80% of insertions, >85% deletion-duplications and >90% SVs in all other categories when the read depth is 10X or higher. We then assessed the sensitivity and false discovery rate at different VaPoR Score cutoffs and found that when looking into different types of SVs at a read depth of 30X, most of the SV types achieve a sensitivity >90% with false discovery rate <10% at VaPoR Score cutoff of 0.1 for heterozygous and 0.25 for homozygous events (Supplemental Figures 1-2). We further observed that there were no significant changes of sensitivity or false discovery rate once the read depth was at or above 20X (Supplemental Figures 3-4) and is consistent across different SV types (Supplemental Tables 1-3).

3.2 VaPoR on 1000 Genomes Project Samples

We also applied VaPoR to a set of diverse samples (HG00513 from CHS, HG00731 and HG00732 from PUR, NA19238 and NA19239 from YRI) that were initially sequenced by the 1000 Genomes Project (1KGP) and for which a high quality set of SVs were reported in the final phase of the project (Sudmant, et al., 2015). These samples were recently re-sequenced using PacBio and therefore provides a platform for assessing VaPoR on known data.

We examined SVs reported on chr1 of the 5 individuals to assess the sensitivity of VaPoR on real genomes (**Table 1**). We first observed that >95% of deletions and insertions could be successfully evaluated by VaPoR. For inversions there were a limited number of events reported but at maximum only 1 event failed the VaPoR quality control per individual. A sensitivity of >90% was achieved for deletions and >80% for insertions. To examine the false validation rate of VaPoR, we modified the reported chr2 events to appear at the same coordinates on chr1 and assessed them as though they were real events using the same sequence data set. VaPoR validated very few deletions or inversion and <10% of insertions.

3.3 Sensitivity to breakpoint accuracy

We next assessed the performance of VaPoR to validate SVs with varying degrees of breakpoint accuracy. Real coordinates were artificially shiftied each direction by -1000 to 1000 base pairs and re-assessed with VaPoR for both simulated and real samples. In both cases, VaPoR exhibited a robust validation score up to approximately 200bp overall, with some slight differences observed between different SV types (Supplemental Figures 6-8).

Table 1. Sensitivity and false discovery rate of different SV types

	deletion	insertion	inversion
Sample	Sens/FDR	Sens/FDR	Sens/FDR
HG00513	0.96/0.00 (0.94[1])	0.80/0.05 (0.93)	0.50/0.00 (0.71)
HG00731	0.94/0.00 (0.96)	0.85/0.07 (0.97)	0.60/0.00 (1.00)
HG00732	0.92/0.00 (0.98)	0.92/0.08 (0.96)	0.33/0.00 (0.86)
NA19238	0.90/0.00 (0.93)	0.88/0.10 (0.96)	1.00/0.00 (1.00)
NA19239	0.87/0.02 (0.95)	0.73/0.09 (0.96)	0.33/0.00 (1.00)

[1] Proportion of SVs that passed VaPoR QC, as determined for events on chr1 and chr2 together.

3.4 Runtime

The computation runtime of VaPoR was assessed using 2 Intel Xeon Intel Xeon E7-4860 processors with 4GB RAM each on both simulated and real genomes. The runtime of simulated event was observed to increase linearly with read depth (Supplemental Figure 9). For events sequenced up to 20X, VaPoR takes ~3 seconds to assess a simple SV and ~5s for a complex event. The assessment of real samples sequenced at 20X required ~1.4 seconds to assess a simple deletion or insertion and ~6 seconds for an inversion (Supplemental Table 4).

Acknowledgements

We thank the Human Genome Structural Variation Consortium (HGSVC) for generating and providing the deep PacBio sequencing. We also thank Yuanfang Guan and Kerby Shedden for discussions over specific statistical considerations. The 1000 Genomes Project Phase 3 data were obtained from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map/ and lifted over to GRCh38. PacBio sequence data were obtained from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/. Simulated date can be obtained from https://umich.box.com/v/vapor.

Funding

This work was supported by the National Institutes of Health [R01HG007068]. AMW was supported by the Genome Science Training Program at the University of Michigan [5T32HG000040]

Conflict of Interest: none declared.

References

Brand, H.*, et al.* Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am J Hum Genet* 2014;95(4):454-461.

Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57-74.

Gibbs, A.J. and McIntyre, G.A. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *European journal of biochemistry* 1970;16(1):1-11.

Huddleston, J.*, et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* 2016.

Layer, R.M.*, et al.* LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;15(6):R84.

Ono, Y., Asai, K. and Hamada, M. PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics* 2013;29(1):119-121.

Rausch, T.*, et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;28(18):i333-i339.

Stankiewicz, P. and Lupski, J.R. Structural variation in the human genome and its role in disease. *Annual review of medicine* 2010;61:437-455.

Sudmant, P.H.*, et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;526(7571):75-81.

Zhao, X.*, et al.* Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol* 2016;17(1):126.