A recurrence based approach for validating 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, University of Michigan, Ann Arbor, MI, 48109, USA, [2]Department of Human Genetics, University of Michigan, Ann Arbor, MI, 48109, USA

[#]Corresponding author: remills@umich.edu

ABSTRACT

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 and report a high-fidelity rate for various features including overall accuracy, sensitivity of breakpoint precision, and predicted genotype.

Keywords: structural variation, copy number variation, sequence analysis

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 (1,2). Various approaches have been developed and applied to paired-end sequencing to detect SVs in whole genomes (3-5), however individual algorithms often exhibit complementary strengths that sometimes lead to disagreements as to the underlying variant. 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 (6). These types of dot plots have been used for decades to examine the specific features of sequence alignments (7), 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 per 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 can quickly and accurately distinguish true from false positive predictions of both simple and complex SVs as well as their underlying genotypes and is also able to assess the breakpoint accuracy of individual algorithms.

MATERIALS AND METHODS

VaPoR Workflow

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: 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, with different statistical model applied. The VaPoR workflow is briefly summarized in Figure 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.

VaPoR Accuracy Assessment

Simulated Data:

Non-overlapping simple deletions, inversions, insertions and duplications as well as complex structural variants as previously categorized (3) were independently incorporated into GRCh38 in both heterozygous and homozygous states, excluding regions of known difficult regions of the genome as described from the ENCODE project (8). Detailed descriptions of each simulated SV types simulated are summarized in Supplementary Tables 1- 3. We applied PBSIM (9) 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. Simulated data can be obtained from https://umich.box.com/v/vapor.

Real Data

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 (10). These samples were recently re-sequenced using PacBio and therefore provides a platform for assessing VaPoR on known data. 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/.

RESULTS

We assessed the performance of VaPoR on both simulated sequences and real genomes from the 1000 Genomes Project to assess the following characteristics: sensitivity and false discovery rate on validating structural variants in simple and complex structures; sensitivity of VaPoR on validating different levels of predicted breakpoint efficacy; stratification of VaPoR scores by genotype; and time and computational cost of VaPoR.

VaPoR on Simulated Data

We applied VaPoR to simulated simple deletions, inversions, insertions and duplications as well as complex structural variants and first assessed the proportion of SVs that VaPoR is capable of

interrogating (i.e. passed VaPoR QC). We 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 (FRR) at different VaPoR score cutoffs and found that when considering different types of SVs across various read depths of, most of the SV types can achieve a sensitivity >90% with false discovery rate <10% at a VaPoR score cutoff of 0.15 (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 and is consistent across different SV types (Figure 2, Supplemental Tables 1-3).

VaPoR on 1000 Genomes Project Samples

We next examined SVs reported on chr1 of 5 individuals from the 1000 Genomes Project (11) 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 (Figure 3a) and >80% for insertions (Figure 3b). To examine the false validation rate of VaPoR, we modified reported events on chr2 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.

We next assessed the performance of VaPoR to validate SVs with varying degrees of breakpoint accuracy. Real coordinates were artificially shifted 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 (Figure 3c,d).

Discrimination of SV types and genotypes

We identified a small number of SVs in the high quality 1000 Genomes set that did not validate with VaPoR. Previous studies have shown that complex rearrangements are often misclassified as simple structural changes (3,12), and indeed upon manual inspection these appeared to consist of multiple connected rearrangements. For example, we observed a reported inversion in HG00513 and NA19239 on chromosome 1 (chr1:239952707-239953529) that was invalidated by VaPoR; an investigation into the long-reads aligned in the region showed the signature of an inverted duplication (Figure 4a) which, when incorporated into a modified reference that location, matched almost exactly with the read sequence (Figure 4b).

We further explored the distribution of VaPoR scores for this region and others across the sample set and observed clear delineations between allelic copy number when fit with a Gaussian mixture model allowing for the generation of genotype likelihoods for each site (Figure 4c). These tracked with our expected genotypes for the inverted duplication on chr1 across the 5 individuals queried while showing no support for the originally predicted inversion (Figure 4d). This shows that VaPoR is not only able to accurately genotype variants but can also distinguish between similar but distinct SV predictions in the same region.

Runtime and efficiency

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 7). 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).

DISCUSSION

Here we present an automated assessment approach, named VaPoR, for exploring various features of predicted genomic structural variants using long read sequencing data. VaPoR directly compares the input reads with the reference sequences with relatively straightforward computational metrics, thus achieving high efficiency in both run time and computing cost. VaPoR exhibits high sensitivity and specificity in both simulated and real genomes, with the capability of discriminating partially resolved SVs either consisting of similar but incorrect SV types at the same location or correct SVs with offset breakpoints. Furthermore, we show that VaPoR performs well at low read depths (5-10X), thus providing the option of systematically assessing large-scale SVs with a lower sequencing cost.

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

	deletion
Sample	Sens/FDR
HG00513	0.96/0.00 (0.94[1])
HG00731	0.92/0.00 (0.96)
HG00732	0.92/0.00 (0.98)
NA19238	0.89/0.00 (0.94)
NA19239	0.85/0.00 (0.95)

[1]Proportion of SVs passed VaPoR QC, as listed in brackets are counted for events on chr1 and chr2 together.

Figure Legends:

Figure 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.

Figure 2. Accuracy of VaPoR on simulated heterozygous and homozygous SVs at varying degrees of sequence coverage and VaPoR score cut-offs. Receiver operator curves (ROC) are shown for simple deletions, duplications and inversions (a,b) as well as complex rearrangements including inverted duplications and deletion-inversion rearrangements (c,d).

Figure 3. Validation rate and breakpoint accuracy of VaPoR on the 1000 Genomes Projects phase 3 calls. VaPoR was applied on 5 individuals with reported SVs as a truth set: HG00513, HG00731, HG00732, NA19238, NA19239. The validation rate of deletions (a) and insertions (b) are shown here across different cutoff scores for VaPoR. Robustness to breakpoint accuracy was assessed by deviating breakpoints from their actual positions across varying distances for deletions (c) and insertions (d).

Figure 4. Validation and genotyping of assessed regions using VaPoR. (a) Recurrence plot of reference genome (GRCh38) to an aligned long read in NA19239 (m150208_160301_42225_c100732022550000001823141405141504_s1_p0/3831/0_12148) for a reported inversion at position chr1:239952707-239953529. The signature is consistent with an inverted duplication structure. (b) Recurrence plot of a different read (m150216_212941_42225_c100729442550000001823151505141565_s1_p0/106403/0_13205) against

the same location, consistent with a non-variant (reference) structure. (c) Distribution of VaPoR scores on all reported SVs on chr1 in samples HG00513, HG00731, HG00732, NA19238, NA19239, stratified by color (solid) and modeled with a Guassian mixture model (dashed). (d) VaPoR scores of SV above now stratified by color as indicated in (c) for both reported inversion (red) and predicted inverted duplication (blue).

Availability and Requirements

Project name: VaPoR

Project home page: https://github.com/mills-lab/vapor

Operating systems: Linux, OS X

Programming languages: Python, R

Other requirements: Python v2.7.8+, rpy2, HTSeq, samtools v0.19+, pyfasta v0.5.2+, and pysam 0.9.1.4+.

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.

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 [T32HG000040]

Authors' contributions

XZ designed the algorithm, wrote the program, comparatively benchmarked the different algorithms, and wrote the manuscript. AMW generated simulated data, aided in assessment testing, and revised the manuscript. REM conceived the study, modified the algorithm, and revised the manuscript. All authors read and approved the final manuscript.

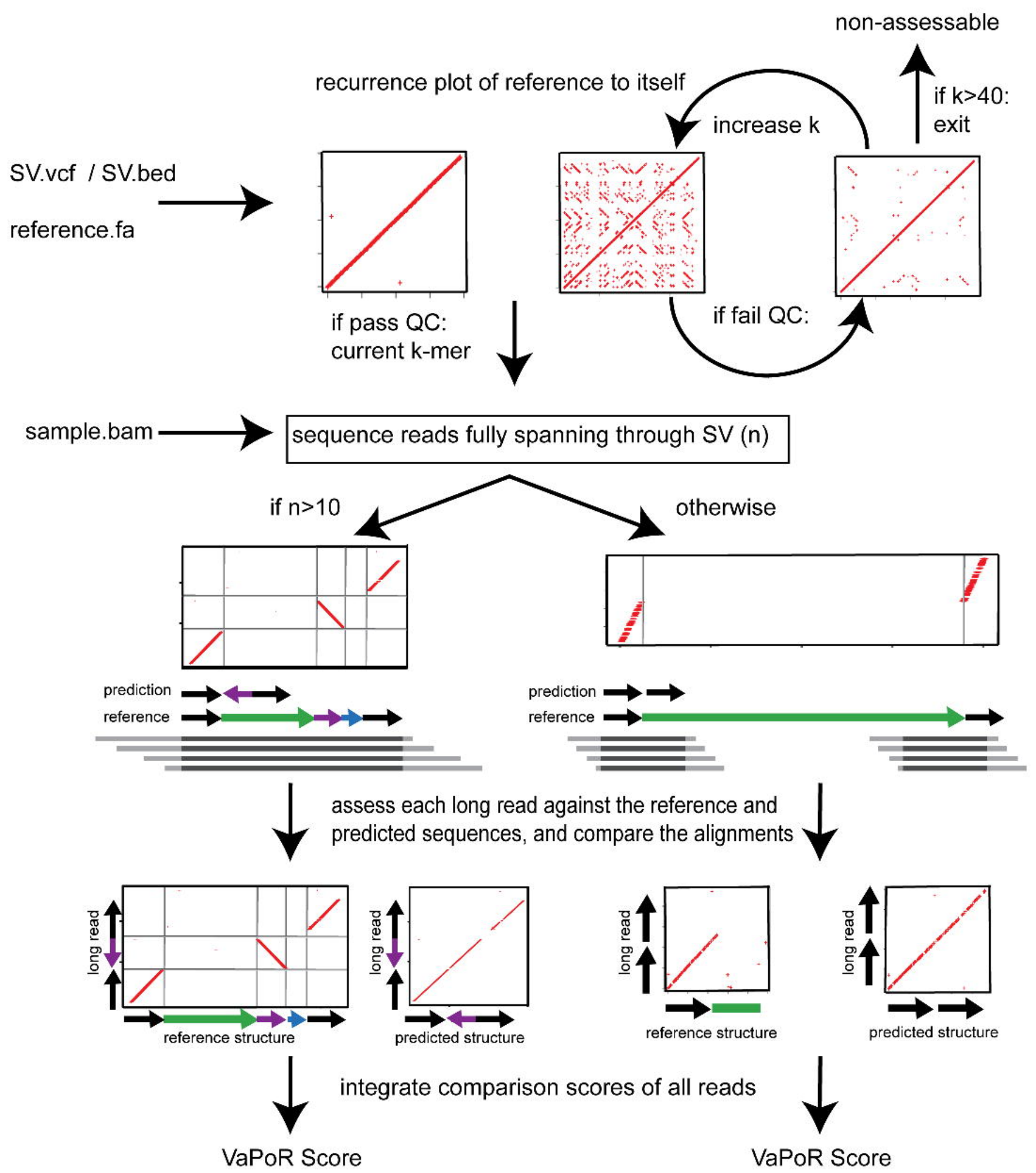
Competing interests

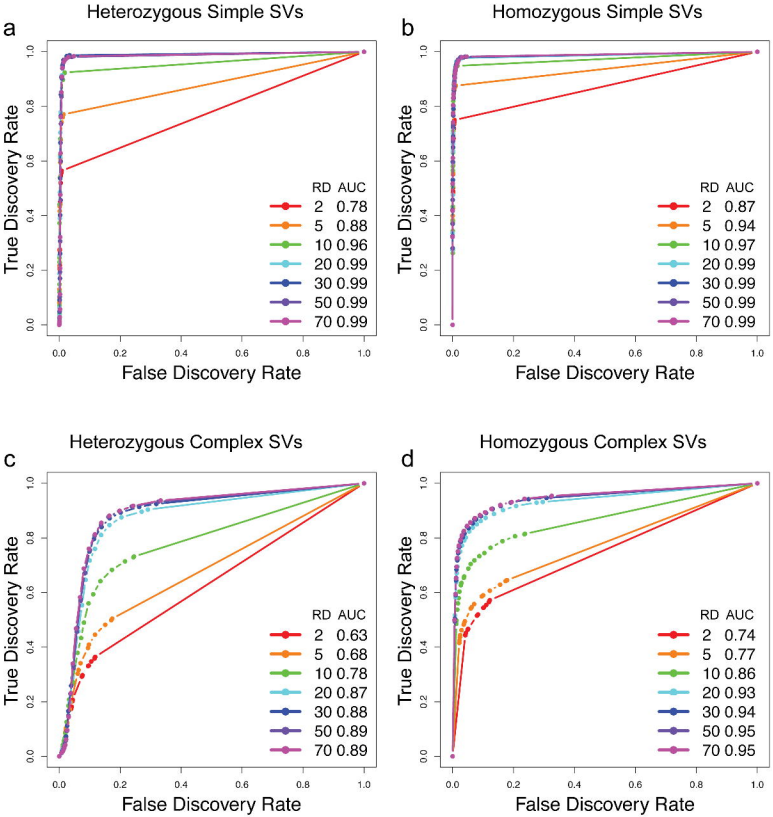
None declared.

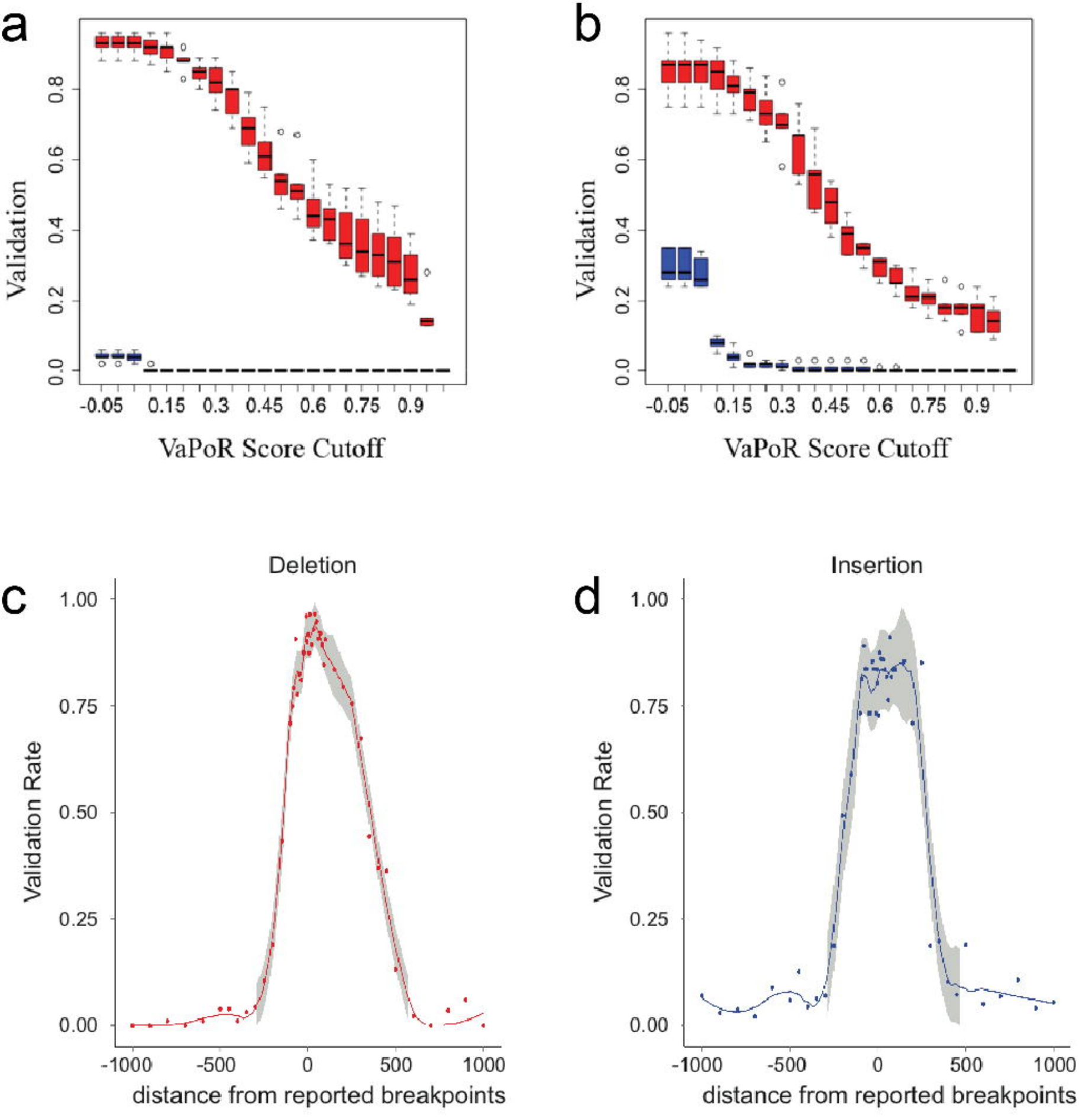
References

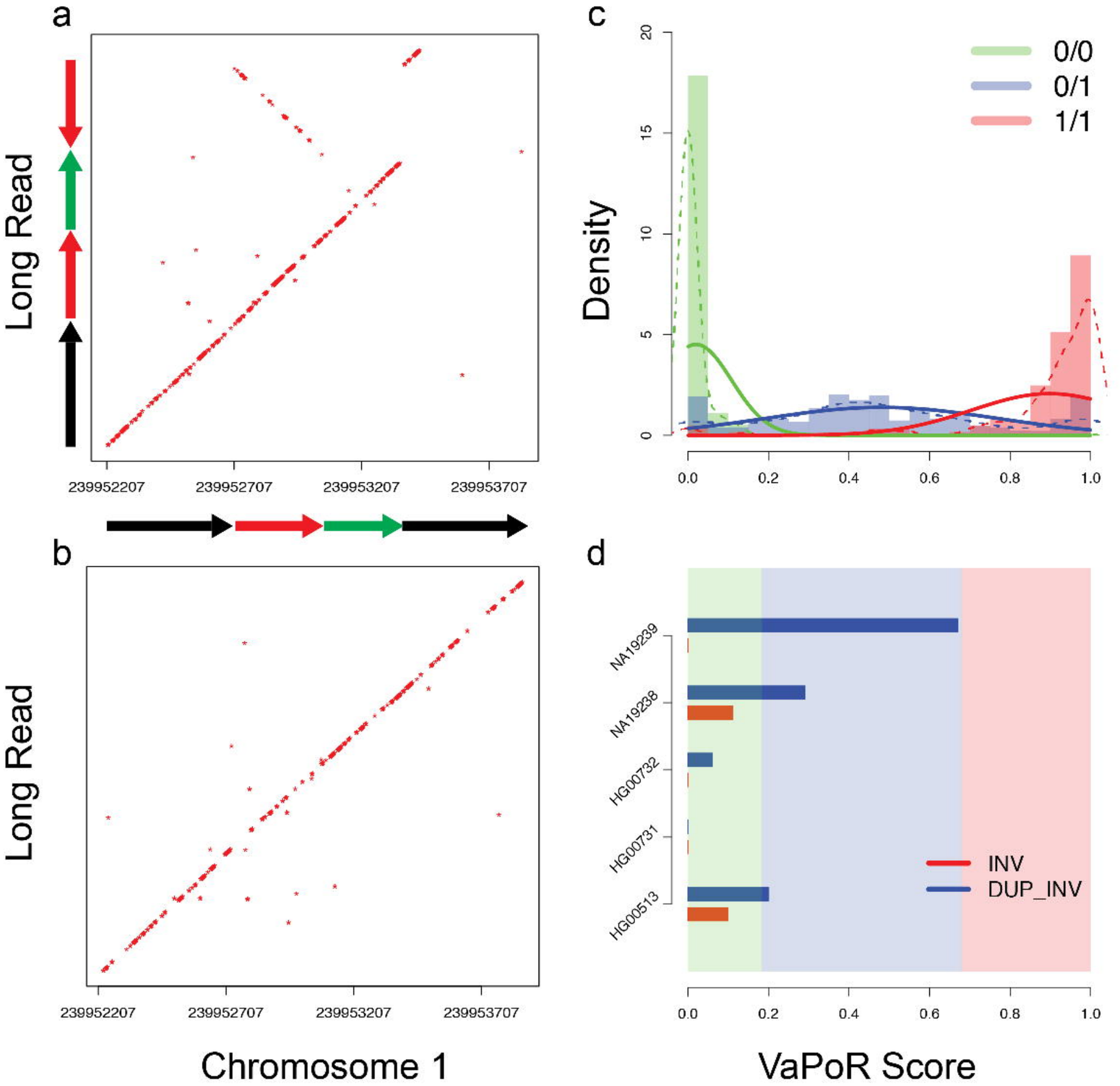
1. Brand, H., Pillalamarri, V., Collins, R.L., Eggert, S., O'Dushlaine, C., Braaten, E.B., Stone, M.R., Chambert, K., Doty, N.D., Hanscom, C. *et al.* (2014) Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am J Hum Genet*, **95**, 454-461.
2. Stankiewicz, P. and Lupski, J.R. (2010) Structural variation in the human genome and its role in disease. *Annual review of medicine*, **61**, 437-455.
3. Zhao, X., Emery, S.B., Myers, B., Kidd, J.M. and Mills, R.E. (2016) Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol*, **17**, 126.
4. Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V. and Korbel, J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333-i339.
5. Layer, R.M., Chiang, C., Quinlan, A.R. and Hall, I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*, **15**, R84.
6. Huddleston, J., Chaisson, M.J., Meltz Steinberg, K., Warren, W., Hoekzema, K., Gordon, D.S., Graves-Lindsay, T.A., Munson, K.M., Kronenberg, Z.N., Vives, L. *et al.* (2016) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res*.
7. Gibbs, A.J. and McIntyre, G.A. (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *European journal of biochemistry*, **16**, 1-11.
8. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
9. Ono, Y., Asai, K. and Hamada, M. (2013) PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics*, **29**, 119-121.
10. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75-81.

11. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68-74.
12. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75-81.


non-assessable
recurrence plot of reference to itself
increase k
if k>40:
exit
SV.vcf / SV.bed
reference.fa
if pass QC:
current k-mer
if fail QC:
sample.bam
sequence reads fully spanning through SV (n)
if n>10
otherwise
prediction
reference
prediction
reference
assess each long read against the reference and
predicted sequences, and compare the alignments
long read
reference structure
long read
predicted structure
long read
reference structure
long read
predicted structure
integrate comparison scores of all reads
VaPoR Score
VaPoR Score


a
Heterozygous Simple SVs
True Discovery Rate
False Discovery Rate
RD AUC
2 0.78
5 0.88
10 0.96
20 0.99
30 0.99
50 0.99
70 0.99
b
Homozygous Simple SVs
True Discovery Rate
False Discovery Rate
RD AUC
2 0.87
5 0.94
10 0.97
20 0.99
30 0.99
50 0.99
70 0.99
c
Heterozygous Complex SVs
True Discovery Rate
False Discovery Rate
RD AUC
2 0.63
5 0.68
10 0.78
20 0.87
30 0.88
50 0.89
70 0.89
d
Homozygous Complex SVs
True Discovery Rate
False Discovery Rate
RD AUC
2 0.74
5 0.77
10 0.86
20 0.93
30 0.94
50 0.95
70 0.95


a
Validation
0.0
0.2
0.4
0.6
0.8
-0.05
0.15
0.3
0.45
0.6
0.75
0.9
VaPoR Score Cutoff
b
Validation
0.0
0.2
0.4
0.6
0.8
-0.05
0.15
0.3
0.45
0.6
0.75
0.9
VaPoR Score Cutoff
c
Deletion
Validation Rate
0.00
0.25
0.50
0.75
1.00
-1000
-500
0
500
1000
distance from reported breakpoints
d
Insertion
Validation Rate
0.00
0.25
0.50
0.75
1.00
-1000
-500
0
500
1000
distance from reported breakpoints


a
Long Read
239952207
239952707
239953207
239953707
b
Long Read
239952207
239952707
239953207
239953707
Chromosome 1
c
Density
0/0
0/1
1/1
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
d
NA19239
NA19238
HG00732
HG00731
HG00513
INV
DUP_INV
0.0
0.2
0.4
0.6
0.8
1.0
VaPoR Score