

1
2
3
4 **A recurrence based approach for validating structural variation using long-read sequencing**
5 **technology.**
6
7

8
9 Xuefang Zhao¹, Alexandra M. Weber¹, and Ryan E. Mills^{1,2,#}
10

11
12 ¹Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI,
13 48109, USA, ²Department of Neurology, University of Michigan, Ann Arbor, MI, 48109, USA,
14

15
16 ³Veterans Affairs Medical Center, Ann Arbor, MI, 48105, USA, ⁴Department of Human Genetics,
17 University of Michigan, Ann Arbor, MI, 48109, USA
18

19
20
21 #Corresponding author: remills@umich.edu
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

ABSTRACT

Background

Although numerous algorithms 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. This is significant, as the accurate identification of structural variation is still an outstanding but important problem in genomics. 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 visual inspection of each region.

Results

Here we present VaPoR, a highly efficient algorithm that autonomously validates large SV sets using long read sequencing data. We assessed the performance of VaPoR on SVs in both simulated and real genomes and report a high-fidelity rate for overall accuracy across different levels of sequence depths. We show that VaPoR can interrogate a much larger range of SVs while still matching existing methods in terms of false positive validations and providing additional features considering breakpoint precision and predicted genotype. We further show that VaPoR can run quickly and efficiently without requiring a large processing or assembly pipeline.

Conclusions

VaPoR provides a long read based validation approach for genomic SVs that requires relatively low read depth and computing resources and thus will provide utility with targeted or low-pass sequencing coverage for accurate SV assessment. The VaPoR Software is available at: <https://github.com/mills-lab/vapor>.

Keywords: structural variation, copy number variation, sequence analysis

BACKGROUND

Structural variants (SVs) are one of the major forms of genetic variation in humans and have been revealed to play important roles in numerous 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 [3-6], however individual algorithms often exhibit complementary strengths that sometimes lead to disagreements as to the precise structure of the underlying variant. The emergence of long read sequencing technology, such as Single Molecule Real-Time (SMRT) sequencing from Pacific Biosciences (PacBio) [7, 8], can deliver reads ranging from several to hundreds of kilobases and provide direct evidence for the presence of an SV. Current strategies make use of *de novo* assembly to create long contigs with minimized error rate [9-11], and then predict SVs, typically with single base resolution, through direct comparison of the assembly against the reference. Though such approaches are powerful, they require both a very high sequencing depth and significant computing power and are currently impracticable for many ongoing research studies.

The additional information obtained from using long reads can still be leveraged to improve variant calling, however. Indeed, such approaches have already been implemented to combine high depth Illumina sequencing with lower depth PacBio reads to improve error correction and variant calling in the context of *de novo* genome assembly [12]. With structural variation, the current state of the art is to use long reads to manually assess potential SVs using subsequent recurrence (dot) plots [13], where the sequences are compared against the reference through a fixed size sliding window (k-mer) and the matches are plotted for visual inspection. The k-mer method is of higher robustness compared to direct sequences comparison [14], which is why these types of dot plots have been used for decades to examine the specific features of sequence alignments [15]. 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 investigates and scores each provided SV prediction by autonomously analyzing the recurrence of k-mers within a local read against both an unmodified reference sequence at that loci as well as a rearranged reference pertaining to the predicted SV structure. 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

1
2
3
4 constructed as well by interrogating the reference sequence against itself at the query location. We show
5 that our approach can quickly and accurately distinguish true from false positive predictions of both
6 simple and complex SVs as well as their underlying genotypes and is also able to assess the breakpoint
7 accuracy of individual algorithms.
8
9
10

11 **DATA DESCRIPTION**

12 *Simulated Data:*

13
14
15
16 Non-overlapping simple deletions, inversions, insertions and duplications as well as complex structural
17 variants as previously categorized [5] were independently incorporated into GRCh38 in both
18 heterozygous and homozygous states, excluding regions of the genome that are known to be difficult to
19 assess as described from the ENCODE project [16]. Detailed descriptions of each simulated SV types
20 simulated are summarized in Supplementary Tables 1- 2. We applied PBSIM [17] to simulate the
21 modified reference sequences to different read depth ranging from 2X to 70X with a parameters
22 difference-ratio of 5:75:20, length-mean 12000, accuracy-mean 0.85 and *model_qc model_qc_clr*.
23 Simulated data can be obtained from <https://umich.box.com/v/vapor>.
24
25
26
27
28
29
30
31
32
33
34
35
36
37

38 *Real Data*

39
40
41 We applied VaPoR to a set of diverse samples (HG00513 from CHS, HG00731 and HG00732 from
42 PUR, NA19238 and NA19239 from YRI) that were initially sequenced by the 1000 Genomes Project
43 and for which a high-quality set of SVs were reported in the final phase of the project [18]. These
44 samples were recently re-sequenced using PacBio to 20X coverage and therefore provides a platform for
45 assessing VaPoR on known data. The 1000 Genomes Project (1KGP) Phase 3 data were obtained from
46 ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/integrated_sv_map/ and lifted over to GRCh38.
47
48
49
50
51
52 PacBio sequence data were obtained from
53 http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/.
54
55
56

57 We have also compared VaPoR against the long read validation approach developed by Layer et al. [3],
58 which requires both PacBio and Moleculo long sequences for full evaluation of SVs. These comparisons
59
60
61
62
63
64
65

1
2
3
4 made use of NA12878, one of few samples that have been sequenced with various technologies
5 including Illumina NGS, PacBio and Moleculo with a truth SV set included in the 1KGP Phase 3 report.
6 The software for the long-read validation approach was obtained from: [https://github.com/hall-lab/long-](https://github.com/hall-lab/long-read-validation)
7 [read-validation](https://github.com/hall-lab/long-read-validation). The PacBio and the Moleculo sequences of this individual were obtained from :
8 http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20131209_na12878_pacbio/si/ and
9 [http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/NA12878/moleculo/ali-](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/NA12878/moleculo/alignment/)
10 [gnment/](http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/NA12878/moleculo/alignment/) respectively.
11
12
13
14
15
16
17
18
19
20

21 RESULTS

22
23
24 We assessed the performance of VaPoR on both simulated sequences and real genomes from the 1000
25 Genomes Project to assess the following characteristics: sensitivity and false discovery rate on
26 validating structural variants in simple and complex structures; sensitivity of VaPoR on validating
27 different levels of predicted breakpoint efficacy; stratification of VaPoR scores by genotype; and time
28 and computational cost of VaPoR.
29
30
31
32

33 VaPoR on Simulated Data

34
35
36
37 We applied VaPoR to simulated simple deletions, inversions, insertions and duplications as well as
38 complex structural variants and first assessed the proportion of SVs that VaPoR is capable of
39 interrogating (i.e. passed VaPoR QC). We found that VaPoR can successfully evaluate >80% of
40 insertions, >85% deletion-duplications and >90% SVs in all other categories when the read depth is 10X
41 or higher. We then assessed the sensitivity and false discovery rate (FDR) at different VaPoR score
42 cutoffs and found that a sensitivity >90% is achieved for most SV types across a wide range of read
43 depths while maintaining a false discovery rate <10% at a VaPoR score cutoff of 0.15 (Supplementary
44 Figures 1-2). We further observed that there were no significant changes of sensitivity or false discovery
45 rate once the read depth was at or above 20X and is consistent across different SV types (Figure 2,
46 Supplementary Figure 3-4, Supplementary Table 3).
47
48
49
50
51
52
53
54
55
56

57 VaPoR on 1000 Genomes Project Samples

58
59
60
61
62
63
64
65

1
2
3
4 We next examined SVs reported on chr1 of 5 diverse individuals from the 1000 Genomes Project [19] to
5 assess the sensitivity of VaPoR on real genomes (Table 1), with 197 – 258 SVs reported per individual
6 in Phase 3 of the project. We first observed that >95% of deletions and insertions could be successfully
7 evaluated by VaPoR. For inversions, there were a limited number of events reported but at maximum
8 only 1 event failed the VaPoR quality control per individual. Moreover, we observed 3-8% deletions and
9 insertions that are 10Kb or larger in size across the individuals. Such events were rarely fully covered by
10 long sequences according their length distribution (Supplementary Figure 5) and were assessed through
11 the ‘large variants assessment’ module implemented in VaPoR (Methods, Supplementary Figure 6), out
12 of which 100% were successfully evaluated. A sensitivity of >90% was achieved for deletions (Figure
13 3a) and >80% for insertions (Figure 3b) at the recommended cutoff of 0.15.
14
15
16
17
18
19
20
21
22
23

24 To examine the false validation rate of VaPoR, we modified reported events on chr2 to appear at the
25 same coordinates on chr1 and assessed them as though they were real events using the same sequence
26 data set. VaPoR validated very few deletions or inversion and <10% of insertions. We further compared
27 VaPoR against a long-read validation approach developed in conjunction with Lumpy [3] using SVs on
28 chr1 of NA12878 reported by the 1kGP Phase 3. VaPoR achieved a sensitivity of 72% for deletions and
29 86% for insertions, while the Lumpy-associated approach was only able to assess 11% and 0%
30 respectively. Both approaches exhibited a very low false validation rate when synthetically assigning the
31 variants to chr2, with 0 for all SV types by the Layer et al approach and varying between 0 and 2.5% for
32 VaPoR (Supplementary Table 4).
33
34
35
36
37
38
39
40

41 **SV breakpoint validation and accuracy**

42
43

44 One of the outstanding challenges of SV discovery is the precise determination of its location at
45 nucleotide resolution. Many short-read algorithms can correctly identify the presence of an SV but
46 report uncertainty at the breakpoints, as can be observed by the reported median confidence intervals of
47 +/-85bp across all events in the 1KGP Phase 3 set [18]. We therefore assessed the performance of
48 VaPoR to validate SVs with varying degrees of breakpoint accuracy by artificially shifting the
49 coordinates of simulated SVs (Supplementary Figures 7-8) and the Phase 3 SVs from the 1000 Genomes
50 samples (Figure 3c,d) by -1000 to 1000 base pairs and re-assessing the new positions with VaPoR.
51 Using default parameters, VaPoR exhibited a robust validation score up to approximately 200bp overall,
52 with some slight differences observed between different SV types. We note that this delineation is
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 partially dependent on the length of the flanking sequence selected, as larger flanking sequences would
5 allow for larger breakpoint offsets depending on user preference. SVs with confidence intervals
6 bounding expected breakpoint locations can be also be systematically assessed using subsequent VaPoR
7 application with offset breakpoints to identify the positions that exhibit the highest score.
8
9

10 11 12 **Discrimination of SV types and genotypes** 13

14
15 We identified a small number of SVs in the high quality 1000 Genomes set that did not validate with
16 VaPoR. Previous studies have shown that complex rearrangements are often misclassified as simple
17 structural changes [5, 13], and indeed upon manual inspection these appeared to consist of multiple
18 connected rearrangements. For example, we observed a reported inversion in HG00513 and NA19239
19 on chromosome 1 (chr1:239952707-239953529) that was invalidated by VaPoR; an investigation into
20 the long-reads aligned in the region showed the signature of an inverted duplication (Figure 4a) which,
21 when incorporated into a modified reference that location, matched almost exactly with the read
22 sequence (Figure 4b).
23
24
25
26
27
28
29

30
31 We further explored the distribution of VaPoR scores for this region and others across the sample set
32 and observed clear delineations between allelic copy number when fitted with a Gaussian mixture model
33 allowing for the generation of genotype likelihoods for each site (Figure 4c). These tracked with our
34 expected genotypes for the inverted duplication on chr1 across the 5 individuals queried while showing
35 no support for the originally predicted inversion (Figure 4d). This shows that VaPoR is not only able to
36 accurately genotype variants but can also distinguish between similar but distinct SV predictions in the
37 same region.
38
39
40
41
42
43
44

45 Using these data, we implemented a genotyping module as an option for users to assess predicted
46 genotypes with those derived using long reads. We compared the genotype of deletions and inversions
47 reported by the 1000 Genomes Phase 3 to the VaPoR genotypes at those loci and observed a non-
48 reference genotype concordance of 0.83 (Supplementary Table 5). The manual visual inspection of
49 regions with discordant genotypes using both the Illumina WGS and PacBio sequence alignments in
50 IGV [20] showed the VaPoR genotypes to be consistently correct in such cases. An updated non-
51 reference genotype concordance of 0.95 was achieved after we integrated these manual inspections into
52 the 1000 Genomes set.
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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 (Supplementary 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 (Supplementary Table 6), with a full genome analysis consisting of ~3,000 SVs larger than 50bp taking 2 CPU hours on average.

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.

METHODS

VaPoR Workflow

VaPoR takes in aligned sequence reads in BAM format and predicted SVs (>50bp) in various formats including VCF and BED. SVs are evaluated by comparing long reads that traverse the reported position of 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 (k-mer) 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 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 too big to fall within individual reads but for which the breakpoint regions can be assessed. Each class of SV is interrogated with different statistical models, as described below. 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. For each record i that corresponds to the fixed-size sequence window position and each format $R_x \in (R_o, R_a)$, we define a distance d_{i,k,s,R_x} as the vertical distance between each record ($X=x_{i,k,s,R_x}$, $Y=y_{i,k,s,R_x}$) in matrix x and the diagonal ($X=x_{i,k,s,R_x}$, $Y=x_{i,k,s,R_x}$) such that $d_{i,k,s,R_x} = \text{abs}(x_{i,k,s,R_x} - y_{i,k,s,R_x})$, and the average distance of all records would be assigned as the score of each matrix:

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

where m is the total number of records in the matrix. Sequences that share higher identity with the read will have a lower $Score_{k,s,R_x}$, 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 vice versa for negative $Score_{k,s,R}$. There exists one exceptional case where a duplicated structure resides within 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 these intrinsic duplications, VaPoR adopts the directed distance $d_{i,k,s,Rx} = x_{i,k,s,Rx} - y_{i,k,s,Rx}$ instead, and take the absolute value of their aggregation, such that the distance contributed by centrosymmetric duplicated segments would offset each other.

$$Score_{k,s,Rx}' = abs\left(\sum_{i=1}^m x_{i,k,s,Rx} - y_{i,k,s,Rx}\right) / m,$$

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:

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

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:

$$Score_{k,s,R} = Score_{k,s,Ra} / Score_{k,s,Ro} - 1,$$

VaPoR Score Calculation:

With a score assigned to each read spanning through the predicted structural variants, the VaPoR score is summarized as:

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

to represent the proportion of long reads supporting predicted structure.

1
2
3
4 The highest supportive score ($\max(\text{Score}_{k,s,R})$) is also reported as a reference for users to meet the
5 specific requirement of their study design, for which we recommend 0.1 as the cutoff.
6
7
8

9 Genotype Assessment:

10
11 The genotype and corresponding likelihood of a predicted SV is assessed by VaPoR using a method
12 previously described for SNP genotyping [21]. Based on the assumption of two alleles per genomic site
13 and k long reads adopted for the assessment, out of which j ($j \leq k$) reads were assigned with a non-
14 positive score, then the log likelihood of a particular genotype g can be estimate as:
15
16
17
18
19

20
21
22
$$l_g = -k * \log(2) + \sum_{i=1}^j \log((2 - g)\epsilon_i + g(1 - \epsilon_i)) + \sum_{i=j+1}^k \log((2 - g)(1 - \epsilon_i) + g\epsilon_i)$$

23
24
25

26 The error rate (ϵ_i) was estimated as the proportion of negative reads across the homozygous alternative
27 events and the positives across the homozygous reference, which is estimated to be 5% across the 1000
28 Genomes samples. The genotype with the highest likelihood is reported as the estimated genotype, with
29 the second largest likelihood in $-\log_{10}$ normalized scale reported as the genotype quality score.
30
31
32

33
34 Flexible window size:

35
36
37 By default, VaPoR uses a window size of 10bp and requires an exact match between sequences, though
38 these can be changed to user-defined parameters. However, many regions of the genome contain
39 repetitive sequences resulting in an abundance of spurious matches in the recurrence matrix, thus
40 introducing bias to the assessment. To address this, VaPoR adopts a quality control step by iteratively
41 assessing the reference sequence against itself and tabulating the proportion of matches along the
42 diagonal. The window size initially starts at 10bp and iteratively increases by 10bp until either (a) the
43 proportion of matches on the diagonal exceeds 40% and the current window size is kept or (b) the
44 window size exceeds 40bp whereby the event will be labeled as ‘non-assessable and excluded from the
45 evaluation.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 **FIGURE LEGENDS**
5

6 **Figure 1. Flowchart describing the VaPoR algorithm.** As input, the algorithm requires a set of
7 structural variants in either VCF or BED format, a series of long reads and/or sequence contigs in BAM
8 format, and the corresponding reference sequence. VaPoR then interrogates each variant individually at
9 its corresponding reference location, assesses the quality of the region and assigns a score.
10
11
12
13
14

15 **Figure 2. Accuracy of VaPoR on simulated heterozygous and homozygous SVs at varying degrees**
16 **of sequence coverage and VaPoR score cut-offs.** Receiver operator curves (ROC) are shown for
17 simple deletions, duplications and inversions (a,b) as well as complex rearrangements including inverted
18 duplications and deletion-inversion rearrangements (c,d).
19
20
21
22
23

24 **Figure 3. Validation rate and breakpoint accuracy of VaPoR on the 1000 Genomes Projects phase**
25 **3 calls.** VaPoR was applied on 5 individuals with reported SVs as a truth set: HG00513, HG00731,
26 HG00732, NA19238, NA19239. The validation rate of deletions (a) and insertions (b) are shown here
27 across different cutoff scores for VaPoR. Robustness to breakpoint accuracy was assessed by deviating
28 breakpoints from their actual positions across varying distances for deletions (c) and insertions (d).
29
30
31
32
33

34 **Figure 4. Validation and genotyping of assessed regions using VaPoR.** (a) Dot plot of reference
35 genome (GRCh38) to an aligned long read in NA19239
36 (m150208_160301_42225_c100732022550000001823141405141504_s1_p0/3831/0_12148) for a
37 reported inversion at position chr1:239952707-239953529. The signature is consistent with an inverted
38 duplication structure. (b) Dot plot of a different read
39 (m150216_212941_42225_c100729442550000001823151505141565_s1_p0/106403/0_13205) against
40 the same location, consistent with a non-variant (reference) structure. (c) Distribution of VaPoR scores
41 on all reported SVs on chr1 in samples HG00513, HG00731, HG00732, NA19238, NA19239, stratified
42 by color (solid) and modeled with a Gaussian mixture model (dashed). (d) VaPoR scores of SV above
43 now stratified by color as indicated in (c) for both reported inversion (red) and predicted inverted
44 duplication (blue).
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 **TABLES**
5
6
7

8 **Table 1.** Sensitivity and false discovery rate of different SV types
9

	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)

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28 ¹Proportion of SVs passed VaPoR QC, as listed in brackets, are counted
29 for events on chr1 and chr2 together.
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Availability and Requirements

Project name: VaPoR

Project home page: <https://github.com/millslab/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 [5T32HG000040]

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 RL, Eggert S, O'Dushlaine C, Braaten EB, et al. Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am J Hum Genet.* 2014;95 4:454-61. doi:10.1016/j.ajhg.2014.09.005.
2. Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, et al. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nature Genetics.* 2012;44 4:390-U195. doi:10.1038/ng.2202.
3. Layer RM, Chiang C, Quinlan AR and Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 2014;15 6:R84. doi:10.1186/gb-2014-15-6-r84.
4. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V and Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28 18:i333-i9. doi:10.1093/bioinformatics/bts378
bts378 [pii].
5. Zhao X, Emery SB, Myers B, Kidd JM and Mills RE. Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biology.* 2016;17 doi:10.1186/s13059-016-0993-1.
6. Chong Z, Ruan J, Gao M, Zhou W, Chen T, Fan X, et al. novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat Methods.* 2017;14 1:65-7. doi:10.1038/nmeth.4084.
7. Rhoads A and Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics.* 2015;13 5:278-89. doi:10.1016/j.gpb.2015.08.002.
8. Travers KJ, Chin CS, Rank DR, Eid JS and Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 2010;38 15:e159. doi:10.1093/nar/gkq543.
9. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature.* 2015;517 7536:608-11. doi:10.1038/nature13907.
10. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods.* 2015; doi:10.1038/nmeth.3454.
11. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and de novo assembly of a Chinese genome. *Nature communications.* 2016;7:12065. doi:10.1038/ncomms12065.
12. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol.* 2012;30 7:693-700. doi:10.1038/nbt.2280.
13. Huddleston J, Chaisson MJ, Meltz Steinberg K, Warren W, Hoekzema K, Gordon DS, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 2016; doi:10.1101/gr.214007.116.
14. Carvalho AB, Dupim EG and Goldstein G. Improved assembly of noisy long reads by k-mer validation. *Genome Res.* 2016;26 12:1710-20. doi:10.1101/gr.209247.116.
15. Gibbs AJ and McIntyre GA. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *European journal of biochemistry.* 1970;16 1:1-11.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

16. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489 7414:57-74. doi:10.1038/nature11247.

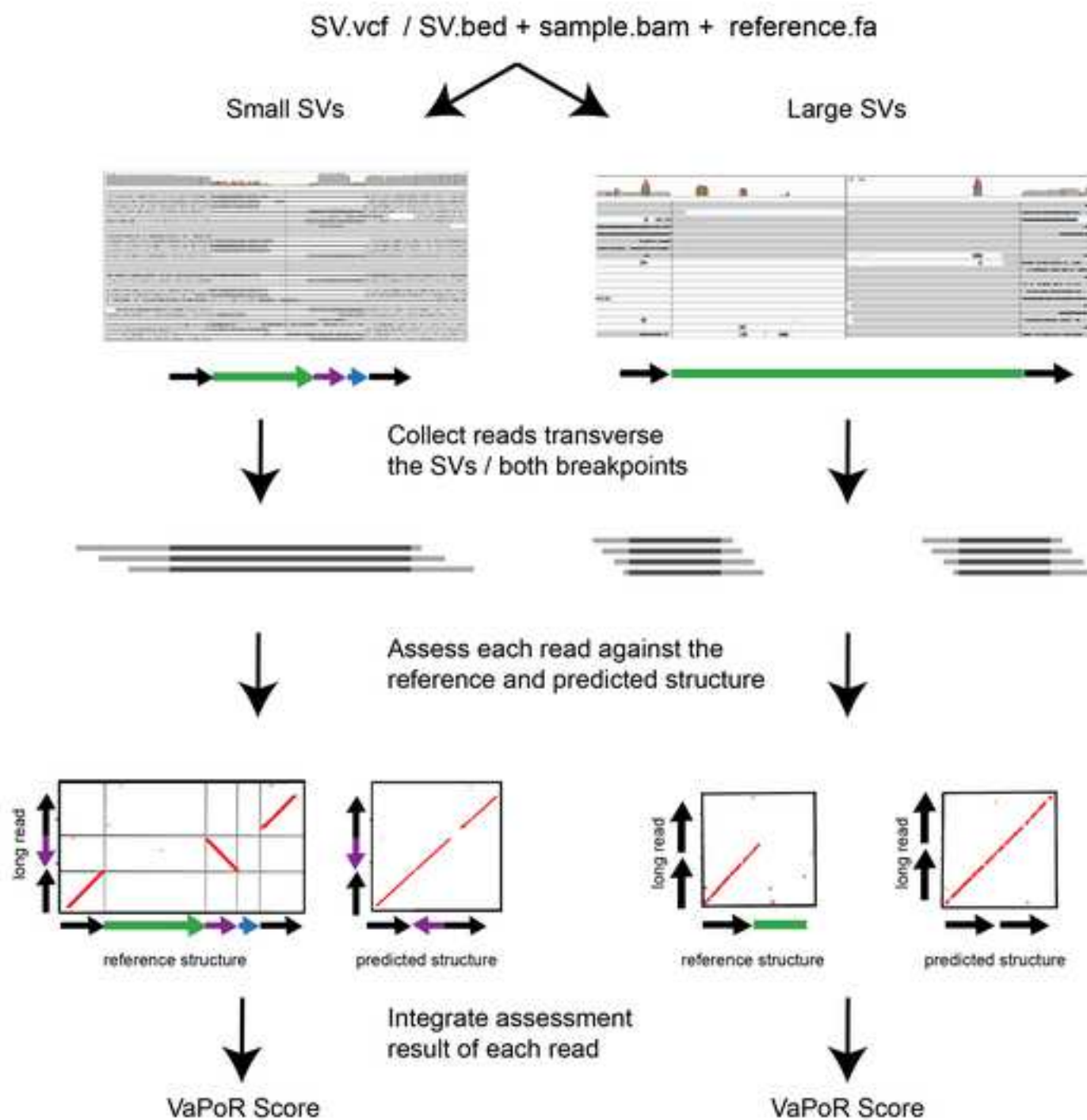
17. Ono Y, Asai K and Hamada M. PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics*. 2013;29 1:119-21. doi:10.1093/bioinformatics/bts649.

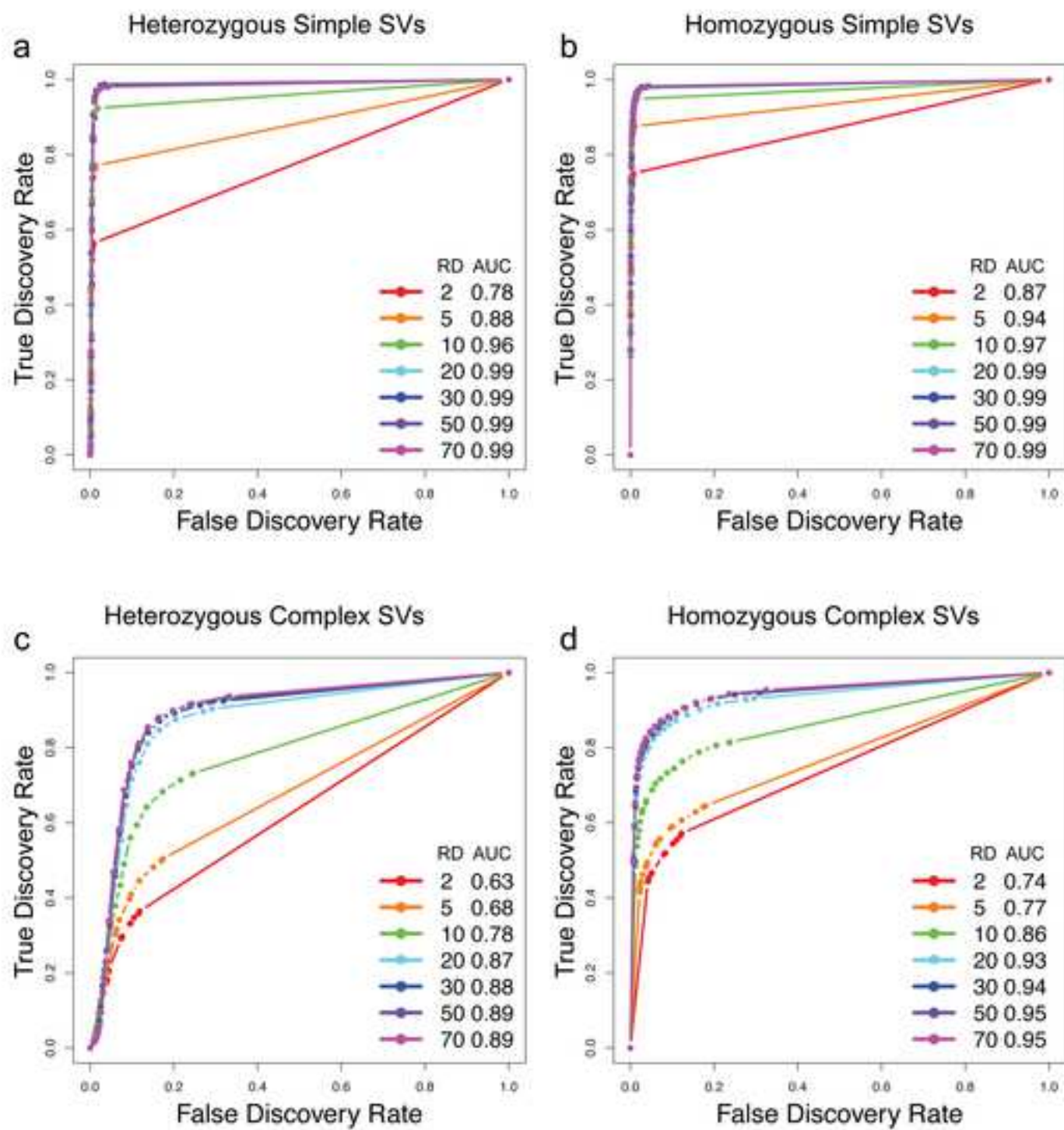
18. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526 7571:75-+. doi:10.1038/nature15394.

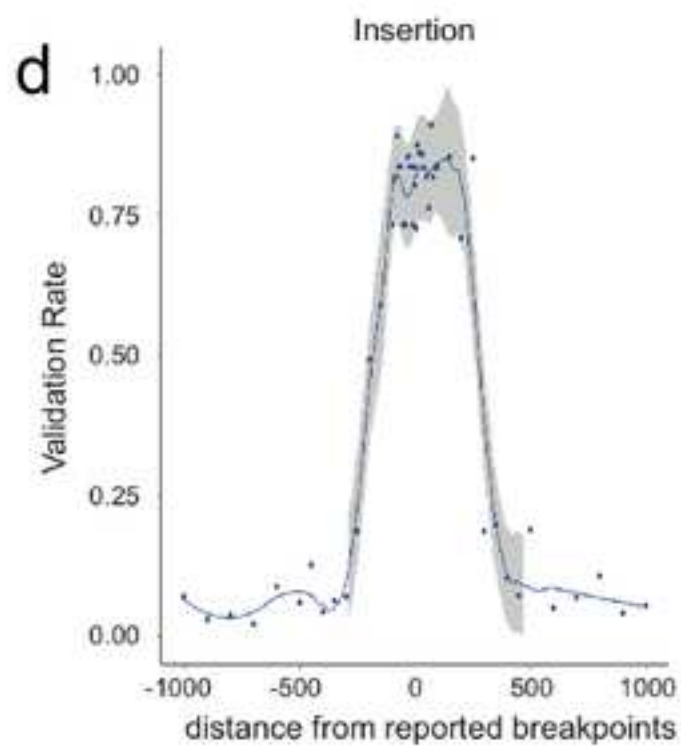
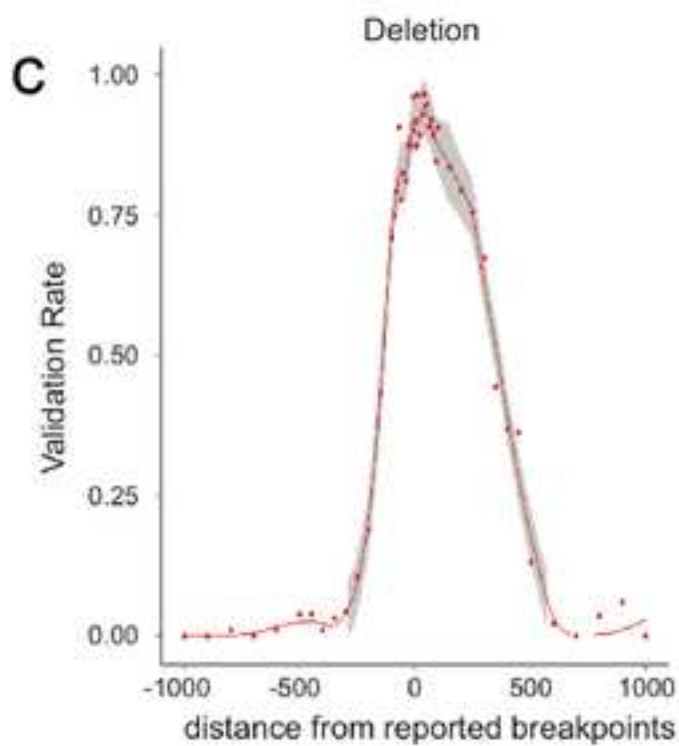
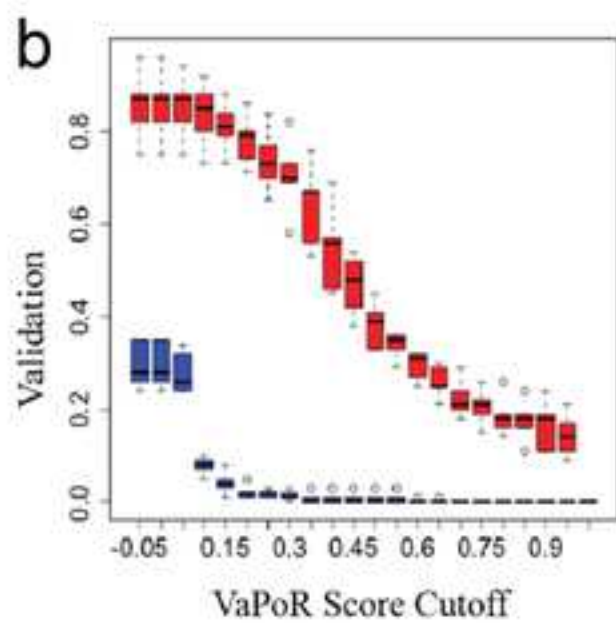
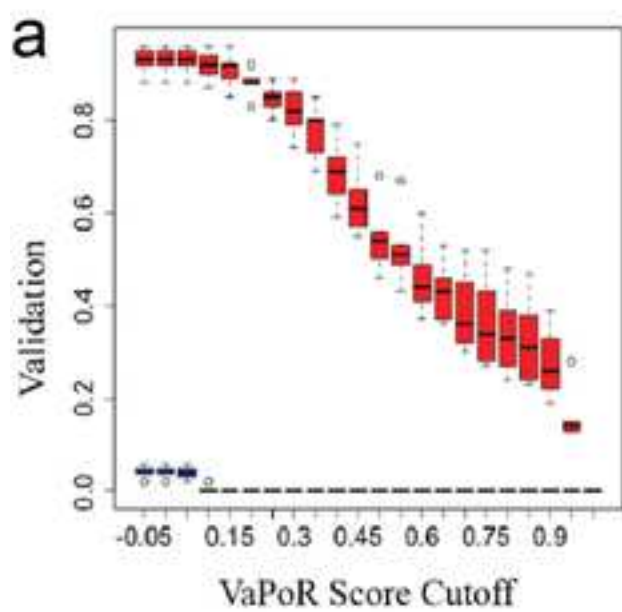
19. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature*. 2015;526 7571:68-74. doi:10.1038/nature15393.

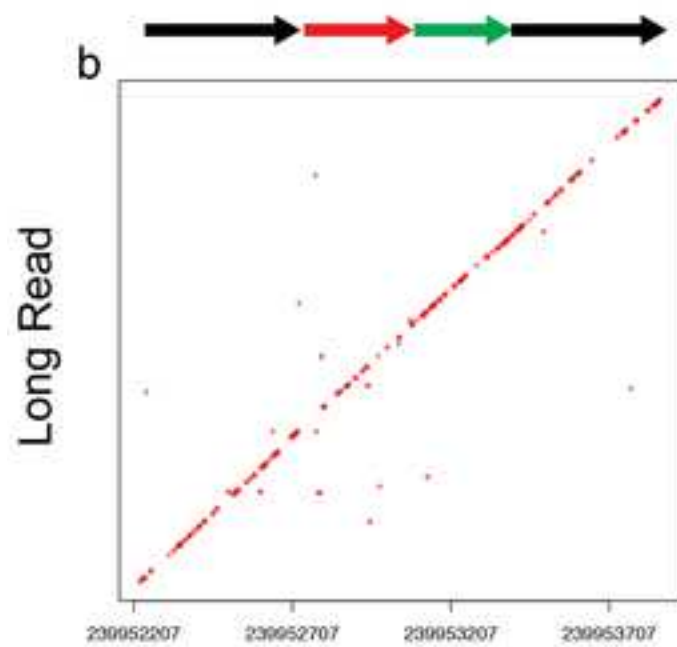
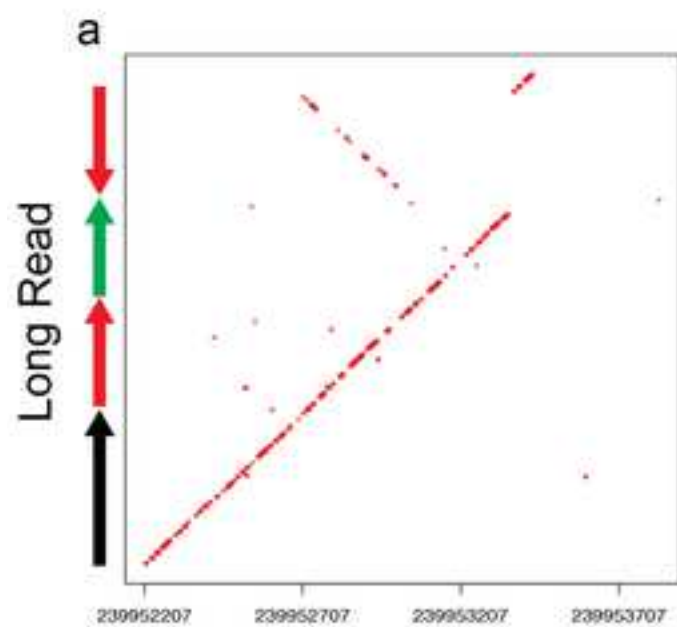
20. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29 1:24-6. doi:nbt.1754 [pii] 10.1038/nbt.1754.

21. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27 21:2987-93. doi:10.1093/bioinformatics/btr509 btr509 [pii].

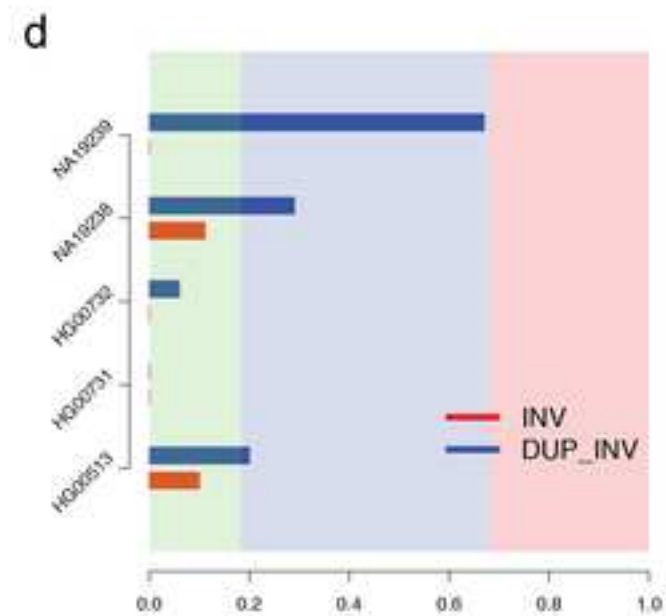
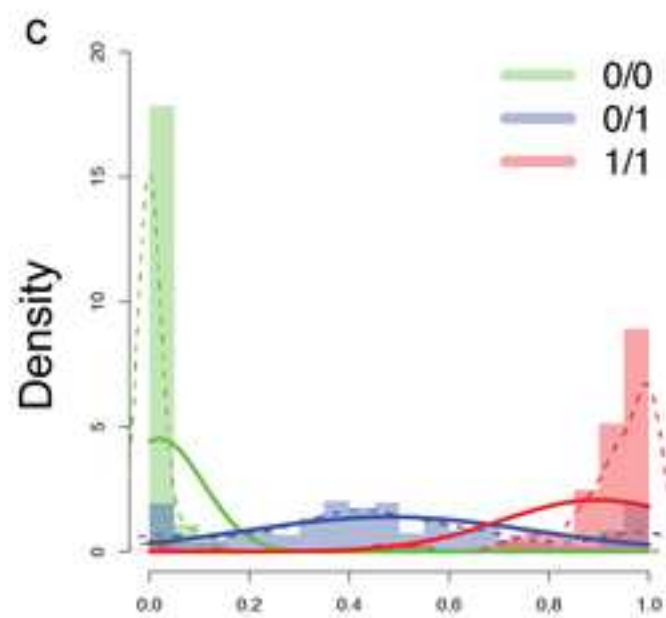








Chromosome 1



VaPoR Score



Click here to access/download
Supplementary Material
supplementary_figures.docx





Click here to access/download
Supplementary Material
supplementary_tables.xlsx