

Manuscript Number:	GIGA-D-19-00035	
Full Title:	Evaluation of computational genotyping of Structural Variations for clinical diagnoses	
Article Type:	Research	
Funding Information:	National Human Genome Research Institute (UM1 HG008898)	Dr. Richard A. Gibbs
Abstract:	<p>Background</p> <p>In recent years, many Structural Variations (SVs) have been identified as having a pivotal role in causing genetic disease. Nevertheless, the discovery of SVs based on short DNA sequence reads from next-generation DNA sequence methods is still error-prone, suffering from low sensitivity and high false discovery. These shortcomings can be partially overcome with the use of long reads, but the current expense precludes their application for routine clinical diagnostics. SV genotyping, on the other hand, offers cost-effective application as diagnostic tool in the clinic, with potentially no false positives and low occurrence of false negatives.</p> <p>Results</p> <p>We assess five state-of-the-art SV genotyping software methods that test short read sequence data. These methods are characterized based on their ability to genotype certain SV types and size ranges. Furthermore, we analyze their applicability to parse different VCF file sub-formats, or to rely on specific meta information that is not always at hand. We compare the SV genotyping methods across a range of simulated and real data including SVs that were not found with Illumina data alone. We assess their sensitivity and ability to filter out initial false discovery calls to assess their reliability.</p> <p>Conclusion</p> <p>Our results indicate that, although SV genotypers have superior performance to SV callers, there are performance limitations that suggest the need for further innovation.</p>	
Corresponding Author:	Fritz Sedlazeck UNITED STATES	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Varuna Chander	
First Author Secondary Information:		
Order of Authors:	Varuna Chander Richard A. Gibbs Fritz J. Sedlazeck	
Order of Authors Secondary Information:		
Additional Information:		
Question	Response	
Are you submitting this manuscript to a	No	

<p>special series or article collection?</p>	
<p>Experimental design and statistics</p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	<p>Yes</p>
<p>Resources</p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

--	--

1 Evaluation of computational genotyping of Structural Variations for 2 clinical diagnoses.

3 Varuna Chander¹, Richard A. Gibbs¹, Fritz J. Sedlazeck^{1*}

4 1: Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston
5 TX 77030

6
7 VC: Varuna.Chander@bcm.edu

8 RG: agibbs@bcm.edu

9 FJS: fritz.sedlazeck@bcm.edu

10 * Correspondence: fritz.sedlazeck@bcm.edu

11 Abstract

12 Background:

13 In recent years, many Structural Variations (SVs) have been identified as having a pivotal role in
14 causing genetic disease. Nevertheless, the discovery of SVs based on short DNA sequence reads
15 from next-generation DNA sequence methods is still error-prone, suffering from low sensitivity
16 and high false discovery. These shortcomings can be partially overcome with the use of long
17 reads, but the current expense precludes their application for routine clinical diagnostics. SV
18 genotyping, on the other hand, offers cost-effective application as diagnostic tool in the clinic,
19 with potentially no false positives and low occurrence of false negatives.

21 Results:

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

22 We assess five state- of-the- art SV genotyping software methods that test short read sequence
23 data. These methods are characterized based on their ability to genotype certain SV types and
24 size ranges. Furthermore, we analyze their applicability to parse different VCF file sub-formats,
25 or to rely on specific meta information that is not always at hand. We compare the SV
26 genotyping methods across a range of simulated and real data including SVs that were not
27 found with Illumina data alone. We assess their sensitivity and ability to filter out initial false
28 discovery calls to assess their reliability.

29

30 **Conclusion:**

31 Our results indicate that, although SV genotypers have superior performance to SV callers,
32 there are performance limitations that suggest the need for further innovation.

33

34 [Keywords](#)

35 Structural Variations, Genotyping, clinical diagnosis, Next Generation Sequencing

36

37 [Background](#)

38 With the continuous advancement of sequencing technologies, our understanding of the
39 importance of Structural Variations (SVs) is increasing[1]. Structural Variations play critical roles
40 in evolution[2], genetic diseases (e.g. mendelian or Cancer) [2, 3] and the regulation of
41 cells/tissues[4]. Furthermore, SVs compromise a substantial proportion of genomic differences
42 between cell types, individuals, populations and species [1, 4-8]. Structural Variations are
43 generally identified as being 50bp or longer genomic variations and categorized into five types:

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

44 Insertions, Deletions, Duplications, Inversions and Translocations [9]. They are most often
45 identified by leveraging paired-end, split read signals and coverage information[8].

46
47 Methods for the detection of SVs are still in their infancy, with some procedures reporting high
48 (up to 89%) levels of false discovery[7, 8, 10-12] (i.e. SVs that are inferred due to artifacts, but
49 not truly present in the sample) and between 10% to 70% false negatives[5, 7] (i.e. missing
50 present SVs in the samples). Although the performance of these methods can improve by the
51 use of long reads, this is often not practical due to high sequencing costs [13-15]. Therefore,
52 using short reads alone significantly hinders SV discovery for routine clinical diagnosis [16].

53
54 An additional challenge is the interpretation of the possible functional consequences of SVs.
55 Despite the availability of existing methods to compare SVs (e.g. SURVIVOR [5]) and to study
56 the potential impact of SVs on genes (VCFanno [17], SURVIVOR_ant [18]), there is still a paucity
57 of methods to assess their allele frequency among the human populations. These issues lead to
58 problems that hinder routine screening for SVs in patient data and limit their proper
59 recognition and characterization for clinical diagnoses.

60
61 The identification of SVs that have been previously identified in other, different samples is in
62 principle, easier than *de novo* detection. For known SVs it is possible to computationally detect
63 SVs directly from short read DNA sequence data in individual per patient samples, guided by
64 the expected position of split reads and discordant paired reads that can confirm breakpoints.
65 This is less demanding than calling *de novo* SVs, since we focus only on specific genomic

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

66 regions. This approach reduces the false discovery rates and therefore, reduces the potential
67 misdiagnosis of patients. In addition, the false negative rate can be reduced as it is easier to
68 genotype a variant than to identify a new SV. Genotyping known SVs has further the advantage
69 that previous studies are likely to have annotated the event, and its possible association with
70 certain diseases. Such events are recorded in SV databases (e.g. dbVar [19]).

71

72 In this paper, we review the current state of SVs genotyping methods and investigate their
73 potential applicability for clinical diagnosis. In particular, we address whether these SV
74 genotypers can re-identify SVs that short read callers were initially blind to (over GiaB [20] call
75 sets) and how they perform for initially falsely inferred SVs. We precisely map out which
76 genotypers operate on which types of SVs and their ability to genotype SVs based on sizes.

77

78 [Analyses](#)

79 Existing methods

80 Here we assess SVs genotyping methods: DELLY [21], Genome STRiP [22], STIX[23], SV2 [24] and
81 SVTyper [25]. They share a common feature in which they require a bam file of the mapped
82 reads and a VCF file that will be genotyped for SVs as inputs. **Table 1** lists their dependencies
83 and their ability to genotype certain types of SVs.

84

85 Overall, they can be divided into groups that support only two SV types (e.g. Genome STRiP) up
86 to methods that support all SV types (SVTyper and DELLY), but require specific meta-

87 information to do so. In the following, we give a brief description of each method that we
 88 assessed. Further insights can be obtained from their respective publications or manuals.

Genotyper	Approach	SV Type					Inputs	Dependencies
		Del	Ins	Inv	Dup	Tra		
Delly	RD, PR, SR	✓		✓	✓	✓	BAM, VCF, Ref	Bcftools [26]
Svtyper	SR, PR	✓		✓	✓	✓	BAM, VCF, Ref	
SV2	RD, PR, SR	✓			✓		BAM, SNV VCF, VCF, Ref, PED file	
STIX	PR,SR	✓		✓			BAM compressed, PED file, VCF, Ref	Excord, Giggle [27]
Genome StRiP	RD, PR, SR	✓			✓		BAM, VCF, Ref	GATK[28]

93 *Table 1: Overview of the SV genotypers assessed here and their ability to assess different SV types. RD: read depth, SR: split*
 94 *reads, PR: paired end reads*

95 DELLY[21] is originally an SV caller that includes a genotype mode to redefine multi-sample
 96 VCFs. It operates on split and paired-end reads to genotype deletions, duplications, inversions

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

97 and translations. However, for all types except the deletions, DELLY requires a sequence
98 resolved call in its own format to be able to estimate the genotype.

99
100 Genome STRiP[22] genotypes only deletions and duplications. The unique aspect of Genome
101 STRiP is that it was designed to genotype multiple samples simultaneously. It requires the GATK
102 pipeline and prepackaged reference metadata bundles.

103
104 STIX[23], which is the most recently developed method included here, is designed the other
105 way around. First, it extracts the discordant read pairs and split reads and generates a
106 searchable index per sample. This index can then be queried if it supports a particular variant.
107 Noteworthy, STIX in the current form only provides information on how many reads support a
108 variant rather than the genotype itself. This is done with a flag describing whether the reads are
109 supported by a particular variant and the number of reads supporting it.

110
111 SVTyper[25] uses a Bayesian likelihood model that is based on discordant paired-end reads and
112 split reads. It was designed to genotype deletions, duplications, inversions and translocations.
113 However, for the latter, it requires specific ID tags provided by Lumpy[29] to genotype them.

114
115 SV2[24] uses a support vector machine learning to genotype deletions and duplications based
116 on discordant paired-end, split read and coverage. Furthermore, it was the only SV genotyper
117 assessed here that leverages SNP calls for its prediction.

118

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

119 Evaluation of SVs computational genotypers based on simulated data

120 To first assess the performance of genotyping methods for SVs, we simulated data sets with
121 100bp Illumina like paired-end reads. Each data set includes 20 SVs simulated for a certain SVs
122 type (duplications, indels, inversions and translocation) and a certain size range (100bp, 250bp,
123 500bp, 1kbp, 2kbp, 5kbp, 10kbp, 50kbp). For each of the data sets, we called SVs using
124 SURVIVOR[5] based on a union set of DELLY, Manta[30], Lumpy[29] calls to include true positive
125 as well as false positive SVs calls (see methods). Given the nature of the simulated data, we only
126 observed 17 false discoveries while we were missing 17.25% of the simulated SVs.
127 **Supplementary Table 1** shows the results for the discovery set over the 32 simulated data sets
128 based on 640 simulated SVs on chr21 and chr22.

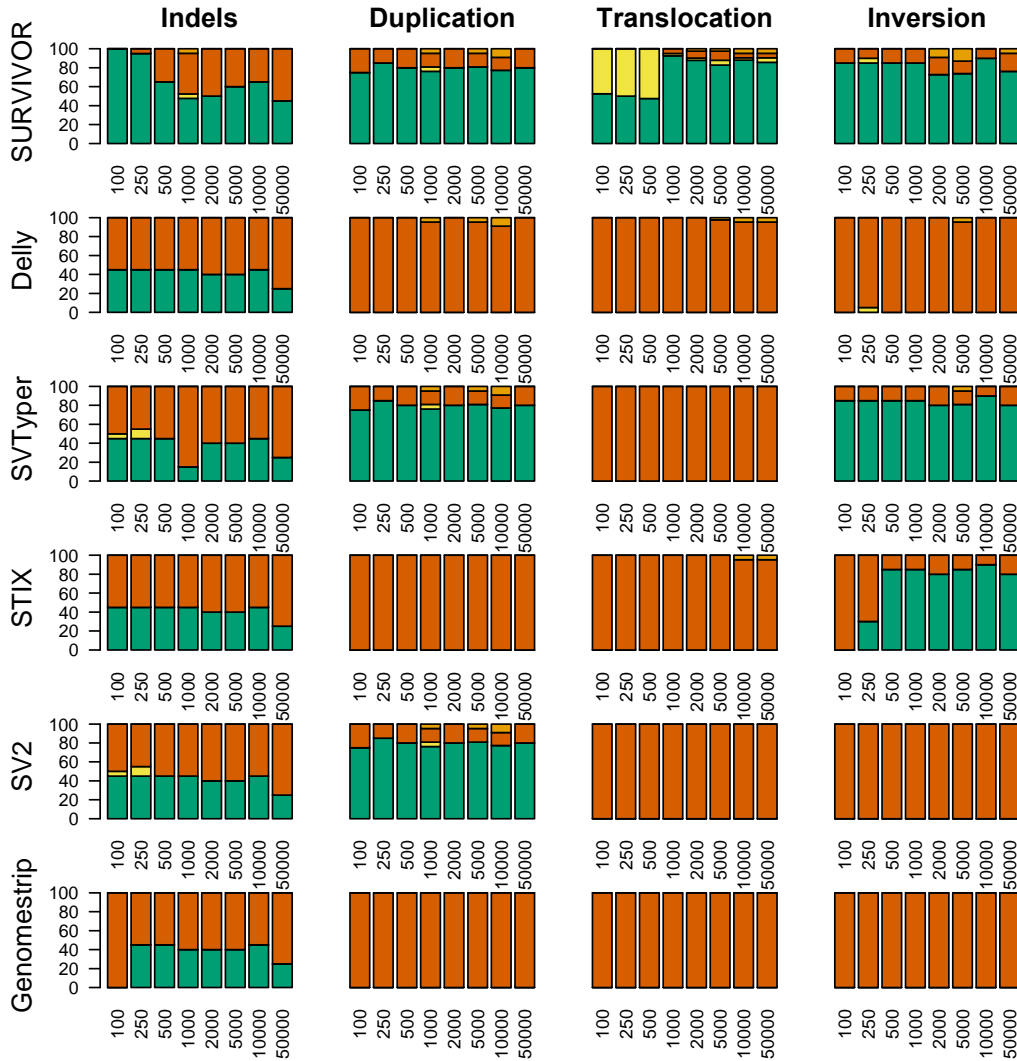


Figure 1: Evaluation of Illumina like reads to assess the SV genotyper ability to re-identify certain types and size ranges.

SURVIVOR is a union set of *Delly*, *Lumpy* and *Manta* to generate the VCF file as an input for the SV genotypers. The evaluation is based on the ability to discover SVs in the first place (*SURVIVOR*). Noteworthy, *Delly* and *SVTyper* can genotype more SVs given their costume information provided from their caller which are *Delly* and *SVTyper*, respectively.

The generated VCF files were taken as input for five SV genotyper callers: *DELLY*, *Genome STRiP*, *SV2*, *STIX* and *SVTyper*. **Figure 1** provides an overview with respect to the ability to discover SVs in the first place (*SURVIVOR*). **Supplementary Table 1** shows the result for all genotypers over the 32 simulated data sets. Interestingly, we observed that not all methods

1
2
3
4 138 accept a standardized VCF file and certain SV types require unique information. For example,
5
6
7 139 while SVTyper is able to genotype deletions, inversion and duplications, it will just work on BND
8
9
10 140 (translocation) events if the ID pairs provided by Lumpy are included in the VCF file. Also DELLY,
11
12 141 which is also capable to infer deletions, inversion, duplications and translocations types of SVs
13
14
15 142 is only able to genotype deletions given a standardized VCF without the extra information.

16
17 143
18
19
20 144 The overall performance of each method was evaluated based on the input VCF generated by
21
22 145 SURVIVOR. Thus, if all of the short read based SV callers were not able to resolve the insertions
23
24
25 146 of 5kbp, then it would count as wrong/missed identified SV. In addition, we assessed the ability
26
27
28 147 for the SV genotyper to filter out falsely called SVs. SVTyper (64.70%) had the highest rate of
29
30 148 correctly genotyping SVs to be present, followed by SV2 (41.57%). Importantly, SV2 was able to
31
32
33 149 genotype deletions and duplications, while SVTyper assessed deletions, duplications and
34
35 150 inversions. Genome STRiP had the lowest (13.40%) success rate of all methods because it can
36
37
38 151 only genotype deletions and duplications. One possible reason to keep in mind for this is that
39
40
41 152 Genome STRiP was designed for population-based genotyping. SVTyper improved marginally
42
43 153 (86.26%) when BND events, which represented translocations, were ignored, followed by the
44
45
46 154 second best method SV2 (83.15%) when focused on deletions and duplications.

47
48 155
49
50
51 156 Next, we assessed the ability of the SV genotypers to reduce the false positives, i.e. initially
52
53
54 157 wrongly inferred SVs. This will represent the scenario of accidentally genotyping a SV that is not
55
56 158 represented in the sample due to sequencing or mapping biases. Over the 32 call sets,
57
58
59 159 SURVIVOR had only 17 false positive calls for the simulated data. Genome STRiP performed best

1
2
3
4 160 with filtering out all falsely detected SVs, but suffers from the lowest ability to genotype SV
5
6
7 161 variations. STIX performed better as it can filter out 13 (76.47%) of the false positive SV calls. In
8
9
10 162 contrast, STIX also achieved a higher (71.76%) performance for correctly identifying SVs.
11
12 163 Although SVTyper had the highest accurately genotyped SVs it performed poorly by filtering out
13
14 164 81.82% of the false positives obtained during the discovery phase.
15
16

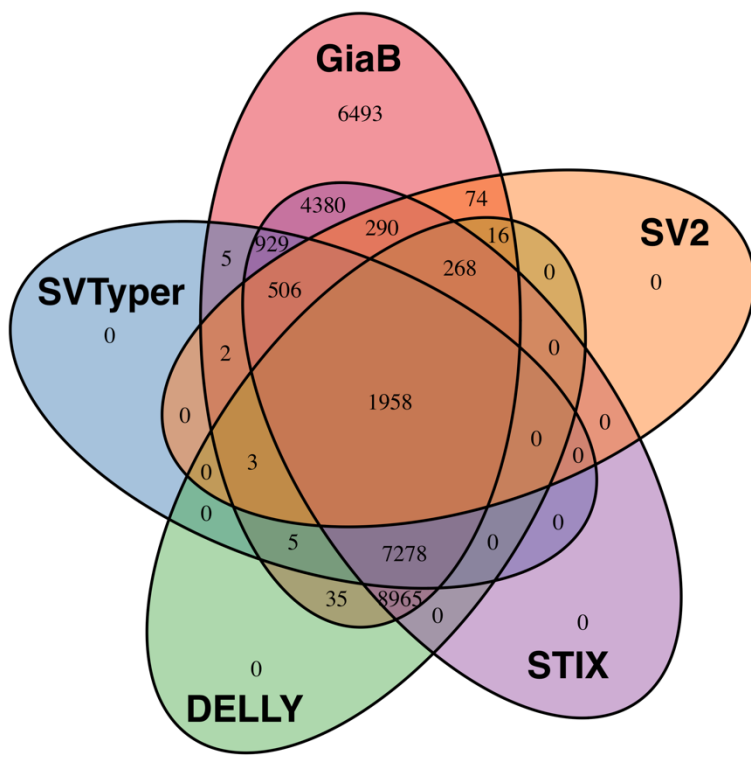
17 165
18
19
20 166 In summary, we observed that none of the methods is clearly superior for correctly genotyping
21
22 167 and correctly filtering/non-reporting variations. Especially non of the program were able to
23
24
25 168 genotype insertions in the simulated data sets. Nevertheless, STIX and SV2 showed a good
26
27
28 169 performance with a good balance of sensitivity and being able to correctly discard false
29
30 170 positives.
31

32
33 171
34
35 172 Evaluation of SVs computational genotypers based on GiaB Ashkenazy Son
36

37
38 173 We further assessed the ability to genotype SVs calls based on the long read DNA sequence
39
40
41 174 data from the Ashkenazy Son (HG002). We are using the current released call set (v0.5.0) from
42
43 175 GiaB, which was generated using sequence resolved calls from multiple technologies such as
44
45
46 176 Illumina, PacBio, BioNano etc. and multiple SV callers and *de novo* assemblies based on these
47
48 177 technologies or a combination of them [20]. Here we are giving this call set the benefit of
49
50
51 178 doubt. It is important to note that 8,195 of these SV calls could not be initially discovered with
52
53
54 179 any Illumina assembly or caller but originated from PacBio based calls or BioNano based calling.
55
56 180 We are using this call set to genotype the SVs based on a 300x Illumina bam file for HG002 and
57
58
59 181 compare the obtained SV genotype predictions to the genotypes reported by GiaB. The first
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

182 observation was that most of the SV genotypers were unable to process the VCF file provided
183 by GiaB. We used SURVIVOR to reduce the information included in the GiaB VCF file. Next, we
184 filtered out the reported INS and complex events from this call set as most SV genotypers
185 crashed while assessing these entries. Unfortunately, we were not able to run GenomeSTRiP
186 successfully as it kept crashing even on a subset of these calls.



189

190 *Figure 2: Evaluation based on GiaB call set v0.5.0 deletions only.*

191

192 **Figure 2** displays the detectable deletions based on the GiaB call set (v0.5.0) per SV genotyper.

193 STIX performed the best among all methods identifying 24,574 (78.74%) of the provided

1
2
3
4 194 deletions. It is important to note that STIX does not currently report genotypes. Thus, we relied
5
6
7 195 only on the information if STIX found reads that support this event rather than genotype
8
9
10 196 information. DELLY performed as the second best identifying 18,528 (59.37%) deletions
11
12 197 followed by SVTyper (34.24%) and SV2 (9.99%). Only 6.27% of the deletion calls from GiaB call
13
14
15 198 set were genotyped by all SV genotype methods. Although this is a very low percent, it is
16
17 199 positive that up to 78.74% of the deletions could be successfully identified out of 62,676
18
19
20 200 deletions (20bp+) in total. Noteworthy, 4921 deletions out of this set were never observed by
21
22 201 any Illumina based caller or assembly. This highlights the potential benefit of using SV
23
24
25 202 genotypers.

26
27 203
28
29
30 204 Next, we assessed the size regimes that SVs genotypers were able to recognize SVs
31
32
33 205 (**Supplementary Table 2**). The deletions from GiaB call set 0.5.0 ranged from 20bp up to
34
35 206 997kbp with a median size of 36bp. All of the SV genotypers were able to identify deletions
36
37
38 207 down to a size of 20bp. Interestingly we observed different median sizes of genotyped
39
40
41 208 deletions, which represents the ability of specific methods to genotype small or large events.
42
43 209 DELLY (31bp) had the lowest median SV size followed by SVTyper (32bp), STIX (35bp) and SV2
44
45
46 210 (116bp). Furthermore, DELLY (816kbp) genotyped also the longest SVs followed by STIX
47
48 211 (694kbp), SV2 (656kbp) and SVTyper (656kbp).

49
50
51 212
52
53 213 When assessing the genotype concordance (**Supplementary Table 3**), DELLY performed the
54
55
56 214 best with an agreement rate of 87.08% given that it identified the variant in the first place. SV2
57
58
59 215 achieved a 78.59% of genotype agreement, however it had one of the lowest recall rates

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

237 We were able to establish the first assessments of sensitivity and false discovery rate for SV
238 genotypers not only focusing on Illumina detectable SVs, but further for SVs that were
239 impossible to be discovered by Illumina alone. The latter becomes important as the field is
240 continuously identifying SVs that based on long read technologies such as PacBio or Oxford
241 Nanopore [13, 15]. These technologies often enable the detection of more complex SVs and
242 also the detection of variations in regions that are hard to assess by Illumina alone. Thus, we
243 rely on SV genotyper to assess if a particular SVs is present either in experiments with large
244 sample sizes to evaluate the allele frequency or in clinical screenings to ease the diagnosis of
245 patients [16].

246
247 Unfortunately, our study also highlights multiple general issues of SV genotyping methods.
248 First, we observed that the methods tested here suffer from a limitation of SV types that they
249 are able to assess. None of the methods were able to assess novel insertions that also
250 represent repeat expansions, which is a subclass of SVs recognized to for their impact in cancer
251 and other phenotypes. Second, most of the methods suffer from a very strict VCF formatting
252 requirement ignoring the current standard and relied further on individual flags that are hard or
253 impossible to regenerate.

254
255 Overall STIX performed well on simulated and GiaB based SVs calls. It showed a good balance of
256 sensitivity vs false discovery and was able to be run on standard VCF files. Nevertheless, the
257 lack of genotype estimations is a clear limitation since it is often relevant to know if a variation
258 is heterozygous or homozygous. Overall, the current methods, although limited in performance,

1
2
3
4 259 represents an advantage to diagnose patients with SVs compared to the discovery of SVs simply
5
6
7 260 because of their much reduced false discovery rate.
8

9 261

12 262 [Potential implications](#)

14 263 SVs genotyping represents a possibility to infer SVs in clinical diagnosis by solving the problems
15
16
17 264 of false discovery and false negative called SVs, compared to the discovery SV methods.

19 265 However, genotyping SVs methods seem to require additional development to improve their
20
21
22 266 ability to operate on different size regimes and all types of SVs (including insertions). Here we
23
24
25 267 presented an overview of the current state-of-the-art methods, which highlights the need to
26
27
28 268 improve upon the current state of the art to enable SV diagnosis of patients in clinical setups.

30 269

33 270 [Methods](#)

35 271 **Simulated datasets**

38 272 We simulated 20 SVs per dataset each for a certain type (indel, inversions, duplication and
39
40
41 273 translocation) and a certain size (100bp, 250bp, 500bp, 1kbp, 2kbp, 5kbp, 10kbp, 50kbp) for chr
42
43 274 21 and 22 using SURVIVOR simSV. These simulations included a 1% SNP rate. After the
44
45
46 275 simulation of the sample genomes we simulated reads using Mason [31] with the following
47
48
49 276 parameter “Illumina -ll 500 -n 100 -N 39773784 -sq -mp -rn 2 “ to generate 100bp paired-end
50
51 277 Illumina like reads. The reads were mapped with BWA MEM[32] using the -M option to mark
52
53
54 278 duplicated reads to the entire genome (GRCh38-2.1.0). Subsequently, we ran Manta (v1.2.1),
55
56 279 DELLY (v0.7.8) and Lumpy (v0.2.13) to call SVs over the simulated datasets. For each data set
57
58
59 280 we generated a union call set based on all 3 callers using SURVIVOR merge (v1.0.3) allowing

1
2
3
4 281 1kbp distance and allowing only the same SV type to be merged. This union set, as well as the
5
6
7 282 SV genotyper output, was evaluated with SURVIVOR eval for the following categories:
8
9 283 Precise: calling an SVs within 10bp and inferring the correct type. Indicated: allowing a
10
11 284 maximum of 1kbp between the simulated and the called breakpoints and ignoring the
12
13 285 predicted type of SVs. Missing: a simulated SVs but not re identified. Additional: a SVs that was
14
15 286 called, but not simulated. The results were summarized using a costume R script operating on
16
17 287 the output of SURVIVOR available on request.
18
19
20
21

22 288
23
24
25 289 **SV genotyping: simulated data**

26
27 290 For genotyping the simulated data set we used the union call VCF based on the SURVIVOR
28
29 291 output as described above. We used DELLY (v0.7.8) specifying the output (-o), the vcf to be
30
31 292 genotyped (-v) and the reference file (-g) as fasta and the bam file. We ran DELLY with the VCF
32
33 293 file from SURVIVOR over the SV discovery caller. The obtained output from DELLY was
34
35 294 converted using bcftools view (v1.7 (using htslib 1.7)) [26] to obtain a VCF file and was filtered
36
37 295 to ignore genotyped calls with 0/0. SVTyper (v0.1.4) was used on the VCF generated from
38
39 296 SURVIVOR based on the discovery phase. We filtered the obtained VCF for genotypes that could
40
41 297 not have been accessed by SVTyper. SV2 (version 1.4.3) was run on the SURVIVOR generated
42
43 298 VCF file for SVs genotyping but required also a SNV file. We generated this SNV file using
44
45 299 Freebayes (v1.1.0-46-g8d2b3a0-dirty) [33] with the default parameters. The resulting SNV file
46
47 300 from Freebayes was compressed and indexed by bgzip and tabix -p vcf [34], respectively. SV2
48
49 301 report their result in three folders (sv2_preprocessed, sv2_features and sv2_genotypes) from
50
51 302 which we used the result reported in sv2_genotypes to benchmark the method. Genome
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 303 STRIP(v2.00.1774) was used following the suggested parameters and the VCF file generated by
5
6
7 304 SURVIVOR. STIX (early version available over GitHub on April 6th 2018) was used to index the
8
9
10 305 bam file using giggle (v0.6.3) [27], excord (v0.2.2) and samtools (v1.7) [26] following the
11
12 306 suggested pipeline. Next, we run STIX with “-s 500” on the VCF files from SURVIVOR and
13
14
15 307 ignoring output VCF entries with "STIX_ZERO=1", which filters out entries where STIX does not
16
17 308 find any evidence for the SV.

19
20 309
21
22 310 SV genotyping: GiaB

23
24
25 311 We obtained the GiaB SV call set (v0.5.0) from [ftp://ftp-](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/)
26
27 312 [trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/) .

28
29
30 313 The SNV calls were taken from here [ftp://ftp-](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh37/)
31
32 314 [trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh37/](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh37/)

33
34
35 315 and the corresponding bam file from [ftp://ftp-](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam)
36
37 316 [trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG00](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam)
38
39 317 [2_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.b](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam)

40
41 318 [am](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam) . The SVs call set needed to be filtered and reduced for just one sample (HG002) using cat
42
43
44 319 and SURVIVOR and was subsequently filtered for deletions only. We ran all SV genotyping
45
46 320 methods like described above. Subsequently, we filtered the results for genotypes: 0/1 and 1/1
47
48
49 321 with the exception of STIX. STIX was filtered base on if it reports reads to support the SVs or
50
51 322 not. This was necessary since STIX does currently not report genotypes. After filtering we
52
53
54 323 merged all data sets together including the original VCF provided using SURVIVOR with a
55
56 324 maximum distance of 10bp and requiring the same types. We analyzed these merged calls
57
58
59
60
61
62
63
64
65

1
2
3
4 325 based on if the original call set reported a genotype to be heterozygous or homozygous
5
6
7 326 alternative. The Venn diagram was generated based on the support vector reported by
8
9
10 327 SURVIVOR and the R package Venn.diagram. The length of the SVs that were able to be
11
12 328 genotyped were extracted using awk filtering for existing calls.
13

14 329

16
17 330 Availability of data and materials

18
19
20 331 The data sets used in this study are available here [ftp://ftp-](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/)
21
22 332 [trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/) and
23
24
25 333 from [ftp://ftp-](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam)
26
27 334 [trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG00](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam)
28
29
30 335 [2_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.b](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam)
31
32
33 336 [am](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.hs37d5.300x.bam). The simulated data sets are available on request.
34

35 337

36
37
38 338 Funding

39
40
41 339 This research was supported by National Institutes of Health award (UM1 HG008898).
42

43 340

44
45
46 341 Authors' contributions

47
48
49 342 VC and FS performed the analysis. VC, FS and RG wrote the manuscript. FS and RG directed the
50
51 343 project.
52

53 344

54
55
56 345 Ethics approval and consent to participate

57
58
59 346 Not applicable
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

347

348 [Consent for publication](#)

349 Not applicable

350

351 [Competing interests](#)

352 F.J.S. has participated in PacBio sponsored meetings over the past few years and have received
353 travel reimbursement and honoraria for presenting at these events

354

355 [References](#)

- 356 1. Weischenfeldt J, Symmons O, Spitz F and Korbel JO. Phenotypic impact of genomic
357 structural variation: insights from and for human disease. Nat Rev Genet. 2013;14
358 2:125-38. doi:10.1038/nrg3373.
- 359 2. Lupski JR. Structural variation mutagenesis of the human genome: Impact on disease
360 and evolution. Environ Mol Mutagen. 2015;56 5:419-36. doi:10.1002/em.21943.
- 361 3. Macintyre G, Ylstra B and Brenton JD. Sequencing Structural Variants in Cancer for
362 Precision Therapeutics. Trends Genet. 2016;32 9:530-42. doi:10.1016/j.tig.2016.07.002.
- 363 4. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical
364 Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, et al. Genetic effects on
365 gene expression across human tissues. Nature. 2017;550 7675:204-13.
366 doi:10.1038/nature24277.

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

367 5. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural
368 variations have strong effects on quantitative traits and reproductive isolation in fission
369 yeast. *Nat Commun.* 2017;8:14061. doi:10.1038/ncomms14061.

370 6. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy
371 number polymorphism in the human genome. *Science.* 2004;305 5683:525-8.
372 doi:10.1126/science.1098918.

373 7. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An
374 integrated map of structural variation in 2,504 human genomes. *Nature.* 2015;526
375 7571:75-81. doi:10.1038/nature15394.

376 8. Tattini L, D'Aurizio R and Magi A. Detection of Genomic Structural Variants from Next-
377 Generation Sequencing Data. *Front Bioeng Biotechnol.* 2015;3:92.
378 doi:10.3389/fbioe.2015.00092.

379 9. Alkan C, Coe BP and Eichler EE. Genome structural variation discovery and genotyping.
380 *Nat Rev Genet.* 2011;12 5:363-76. doi:10.1038/nrg2958.

381 10. English AC, Salerno WJ and Reid JG. PBHoney: identifying genomic variants via long-read
382 discordance and interrupted mapping. *BMC Bioinformatics.* 2014;15:180.
383 doi:10.1186/1471-2105-15-180.

384 11. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy
385 number variation by population-scale genome sequencing. *Nature.* 2011;470 7332:59-
386 65. doi:10.1038/nature09708.

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

387 12. Teo SM, Pawitan Y, Ku CS, Chia KS and Salim A. Statistical challenges associated with
388 detecting copy number variations with next-generation sequencing. *Bioinformatics*.
389 2012;28 21:2711-8. doi:10.1093/bioinformatics/bts535.

390 13. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al.
391 Accurate detection of complex structural variations using single-molecule sequencing.
392 *Nat Methods*. 2018; doi:10.1038/s41592-018-0001-7.

393 14. Goodwin S, McPherson JD and McCombie WR. Coming of age: ten years of next-
394 generation sequencing technologies. *Nat Rev Genet*. 2016;17 6:333-51.
395 doi:10.1038/nrg.2016.49.

396 15. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al.
397 Resolving the complexity of the human genome using single-molecule sequencing.
398 *Nature*. 2015;517 7536:608-11. doi:10.1038/nature13907.

399 16. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read
400 genome sequencing identifies causal structural variation in a Mendelian disease. *Genet*
401 *Med*. 2018;20 1:159-63. doi:10.1038/gim.2017.86.

402 17. Pedersen BS, Layer RM and Quinlan AR. Vcfanno: fast, flexible annotation of genetic
403 variants. *Genome Biol*. 2016;17 1:118. doi:10.1186/s13059-016-0973-5.

404 18. Sedlazeck FJ, Dhroso A, Bodian DL, Paschall J, Hermes F and Zook JM. Tools for
405 annotation and comparison of structural variation. *F1000Res*. 2017;6:1795.
406 doi:10.12688/f1000research.12516.1.

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

407 19. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. DbVar and
408 DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* 2013;41
409 Database issue:D936-41. doi:10.1093/nar/gks1213.

410 20. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of
411 seven human genomes to characterize benchmark reference materials. *Sci Data.*
412 2016;3:160025. doi:10.1038/sdata.2016.25.

413 21. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V and Korbel JO. DELLY: structural
414 variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.*
415 2012;28 18:i333-i9. doi:10.1093/bioinformatics/bts378.

416 22. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large
417 multiallelic copy number variations in humans. *Nat Genet.* 2015;47 3:296-303.
418 doi:10.1038/ng.3200.

419 23. Layer RM: <https://github.com/ryanlayer/stix> (2018).

420 24. Antaki D, Brandler WM and Sebat J. SV2: accurate structural variation genotyping and
421 de novo mutation detection from whole genomes. *Bioinformatics.* 2018;34 10:1774-7.
422 doi:10.1093/bioinformatics/btx813.

423 25. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq:
424 ultra-fast personal genome analysis and interpretation. *Nat Methods.* 2015;12 10:966-8.
425 doi:10.1038/nmeth.3505.

426 26. Li H. A statistical framework for SNP calling, mutation discovery, association mapping
427 and population genetical parameter estimation from sequencing data. *Bioinformatics.*
428 2011;27 21:2987-93. doi:10.1093/bioinformatics/btr509.

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

27. Layer RM, Pedersen BS, DiSera T, Marth GT, Gertz J and Quinlan AR. GIGGLE: a search engine for large-scale integrated genome analysis. *Nat Methods*. 2018;15 2:123-6. doi:10.1038/nmeth.4556.

28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20 9:1297-303. doi:10.1101/gr.107524.110.

29. 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.


30. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32 8:1220-2. doi:10.1093/bioinformatics/btv710.

31. Holtgrewe M. *Mason-A Read Simulator for Second Generation Sequencing Data*. 2010. Institut für Mathematik und Informatik, Freie Universität Berlin.

32. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv e-prints*. 2013; doi: arXiv:1303.3997 [q-bio.GN].

33. Garrison E and Marth G. Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints*. 2012.

34. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. 2011;27 5:718-9. doi:10.1093/bioinformatics/btq671.



Click here to access/download
Supplementary Material
Tables.xlsx

February 1, 2019

To the editors,

We are pleased to share our manuscript with you for consideration as a manuscript in GigaScience. Our work entitled “Evaluation of computational genotyping of Structural Variations for clinical diagnoses.”, represents the first systematically assessment of Structural Variation (SV) genotyping methods and their individual advances and disadvantages.



Human Genome Sequencing Center

One Baylor Plaza
Houston, Texas 77030-3498

TEL: (713)798-6539
FAX: (713)798-5741

As you may know, calling SVs using short reads is often discussed as error-prone and lacking of sensitivity. On the other hand, more and more papers appear highlighting the impact of SVs on different human diseases and across phenotypes among many species ranging from bacteria to humans. Here we suggest to utilize SV genotyping as a way to overcome this paradigm and to scan for annotated SVs in short read based sequencing data sets on a routine basis. This reduces the false positive rate significantly compared to short read based de novo SV calling.

In our manuscript, we further assess the ability and state of the art of these methods to recall SVs. This is first done over simulated data but then extended using the new gold standard from GIAB. Here we focus on variants that short read based calling is blind to and that were only found with long read technologies. We could show that our proposed strategy enables the detection of SVs also with short read based methods. This allows a broader screen for these important variant class for multiple studies and even a path forward for clinical screenings.

In the end, we conclude giving advice to which method performed the best and which method to choose to accomplish these goals. We further highlight general problems to the currently existing methods and make suggestions for future developments. Thus, we think our manuscript is of interest to the readers of GigaScience and in general for people that are currently interested in SVs and genetic diversity, but are cautious about the low performance of SV calling based on short read data.

We welcome any comments that you and your reviewers may have. We would recommend: Ryan Layer (ryan.layer@gmail.com), Aaron Quinlan (aaronquinlan@gmail.com), Justin Zook (justin.zook@nist.gov) and Marc Salit ([msalit@stanford.edu](mailto:mсалit@stanford.edu)).

Sincerely,

A handwritten signature in black ink, appearing to read "Fritz J. Sedlazeck".

Fritz J. Sedlazeck, Ph.D.

Assistant Professor,
Human Genome Sequencing Center,
Baylor College of Medicine,
Houston, TX, 77030