

GigaScience

SV-plaudit: A cloud-based framework for manually curating thousands of structural variants

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00103	
Full Title:	SV-plaudit: A cloud-based framework for manually curating thousands of structural variants	
Article Type:	Research	
Funding Information:	National Human Genome Research Institute (K99HG009532)	Dr Ryan Layer
	National Human Genome Research Institute (R01HG006693)	Dr Aaron R Quinlan
	National Human Genome Research Institute (R01GM124355)	Dr Aaron R Quinlan
	National Cancer Institute (US) (U24CA209999)	Dr Aaron R Quinlan
Abstract:	SV-plaudit is a framework for rapidly curating structural variant (SVs) predictions. For each SV, we generate an image that visualizes the coverage and alignment signals from a set of samples. Images are uploaded to our cloud framework where users assess the quality of each image using a client-side web application. Reports can then be generated as a tab-delimited file or annotated VCF. As a proof of principle, nine researchers collaborated for one hour to evaluate 1,350 SVs each. We anticipate that SV-plaudit will become a standard step in variant calling pipelines and the crowd-sourced curation of other biological results.	
Corresponding Author:	Ryan Layer UNITED STATES	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Jonathan R Belyeu	
First Author Secondary Information:		
Order of Authors:	Jonathan R Belyeu	
	Thomas J Nicholas, PHD	
	Brent S Pedersen, PHD	
	Thomas A Sasani	
	James M Havrilla	
	Stephanie N Kravitz	
	Megan E Conway	
	Brian K Lohman, PHD	
	Aaron R Quinlan, PHD	
Ryan Layer		
Order of Authors Secondary Information:		

Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<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>	Yes
<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>	Yes
<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>	Yes

1 **SV-plaudit: A cloud-based framework for manually curating thousands of structural** 2 3 4 **variants**

5
6
7
8
9 Jonathan R. Belyeu^{1,2}, Thomas J. Nicholas^{1,2}, Brent S. Pedersen^{1,2}, Thomas A. Sasani^{1,2}, James M. Havrilla^{1,2},
10
11 Stephanie N. Kravitz^{1,2}, Megan E. Conway¹, Brian K. Lohman^{1,2}, Aaron R. Quinlan^{1,2,3+}, Ryan M. Layer^{1,2+}
12

13
14
15
16 1. Department of Human Genetics, University of Utah, Salt Lake City, UT

17
18 2. USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT

19
20 3. Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

21
22
23
24 + *To whom correspondence should be addressed*
25

26 27 28 29 **ABSTRACT**

30
31
32 *SV-plaudit* is a framework for rapidly curating structural variant (SVs) predictions. For each SV, we generate an
33
34 image that visualizes the coverage and alignment signals from a set of samples. Images are uploaded to our
35
36 cloud framework where users assess the quality of each image using a client-side web application. Reports can
37
38 then be generated as a tab-delimited file or annotated VCF. As a proof of principle, nine researchers collaborated
39
40 for one hour to evaluate 1,350 SVs each. We anticipate that *SV-plaudit* will become a standard step in variant
41
42 calling pipelines and the crowd-sourced curation of other biological results.
43
44

45
46
47
48 Code available at <https://github.com/jbelyeu/SV-plaudit>
49

50
51
52
53 Demonstration video available at <https://www.youtube.com/watch?v=ono8kHMKxDs>
54

54 55 **KEYWORDS**

56
57 Structural variants; Visualization; Manual curation
58

59 60 61 **BACKGROUND**

1 1 Large genomic rearrangements, or structural variants (SVs), are an abundant form of genetic variation within the
2
3 human genome^{1,2}, and they play an important role in both species evolution^{3,4} and human disease phenotypes⁵⁻⁹.
4
5 While many methods have been developed to identify SVs from whole-genome sequencing (WGS) data¹⁰⁻¹⁴, the
6
7 accuracy of SV prediction remains far below that of single-nucleotide and insertion-deletion variants¹.
8
9 Improvements to SV detection algorithms have, in part, been limited by the availability and applicability of high-
10
11 quality truth sets. While the Genome in a Bottle¹⁵ consortium has made considerable progress toward a gold-
12
13 standard variant truth set, the incredibly high quality of the data underlying this project (300X and PCR-free) calls
14
15 into question the generality of the accuracy obtained in typical quality WGS datasets (30X with PCR-
16
17 amplification).
18
19

20
21
22
23
24 Given the high false positive rate of SV calls from genome and exome sequencing, manual inspection is a critical
25
26 quality control step, especially in clinical cases. Scrutiny of the evidence supporting an SV is considered to be a
27
28 reliable "dry bench" validation technique, as the human eye can rapidly distinguish true SV signal from alignment
29
30 artifacts. In principle, we could improve the accuracy of SV call sets by visually validating every variant. In
31
32 practice, however, current genomic data visualization methods¹⁶⁻²¹ were designed primarily for spot checking a
33
34 small number of variants and are difficult to scale to the thousands of SVs in typical call sets. Therefore, a curated
35
36 set of SVs requires a new framework that scales to thousands of SVs, minimizes the time needed to adjudicate
37
38 individual variants, and manages the collective judgment of large and often geographically dispersed teams.
39
40

41
42
43
44 Here we present *SV-plaudit*, a fast, highly-scalable framework enabling teams of any size to collaborate on the
45
46 rapid, web-based curation of thousands of SVs. In the web interface, users answer a curation question (e.g. is
47
48 this variant a somatic variant, a germline variant, or a false positive) for a series of pre-computed images (**Fig 1**)
49
50 that contain the coverage, paired-end alignments, and split-read alignments for the region surrounding a
51
52 candidate SV for a set of relevant samples (e.g., tumor and matched normal samples). Responses are collected
53
54 and returned as a report which can be used to identify high-quality variants.
55
56

57
58
59
60 While a team of curators is not required, collecting multiple opinions for each SV allows *SV-plaudit* to report the
61
62 consensus view (i.e., a "curation score") of each variant. This consensus is less susceptible to human error and
63
64

1 1 does not require expert users to score variants. With *SV-plaudit*, it is practical to inspect and score every variant
2
3 in a call set, thereby improving the accuracy of SV predictions in individual genomes, and curating high quality-
4
5 truth sets for SV method tuning.
6
7
8
9

10 RESULTS

11
12 To assess *SV-plaudit's* utility for curating SVs, nine researchers in the Quinlan laboratory at the University of
13
14 Utah manually inspected and scored the 1,350 SVs (1,310 deletions, 8 duplications, 4 insertions, and 28
15
16 inversions) that the 1000 Genomes Project identified in the NA12878 genome (**Supplemental File 1**). Since we
17
18 expect trio analysis to be a common use case of *SV-plaudit*, we included alignments from NA12878 and her
19
20 parents (NA12891 and NA12892), and participants considered the curation questions “The SV in the top sample
21
22 (NA12878) is:” and answers “GOOD”, “BAD”, or “DE NOVO”. In total, the full experiment took less than two hours
23
24 with Amazon costs totaling less than \$0.05. The images (**Supplemental File 2**) were generated in 3 minutes (20
25
26 threads, 2.7 seconds per image) and uploading to S3 required 5 minutes (full command list in **Supplemental**
27
28 **File 3**). The mean time to score all images was 60.1 minutes (2.67 seconds per image) (**Fig 2A**, reports in
29
30 **Supplemental Files 4,5**). In the scoring process, no de novo variants were identified. 40 images did not render
31
32 correctly due to issues in the alignment files (e.g., coverage gaps) and were removed from the subsequent
33
34 analysis (**Supplemental File 6**).
35
36
37
38
39
40
41

42 For this experiment, we use a curation score that mapped “GOOD” and “DE NOVO” to the value one, “BAD” to
43
44 the value zero, and the mean as the aggregation function (**Fig 2B**). Most (70.5%) of variants were scored
45
46 unanimously, with 67.1% being unanimously “GOOD” (score = 1.0, e.g., **Fig 1A**) and 3.4% being unanimously
47
48 “BAD” (score = 0.0, e.g. **Fig 1B**). Since we had nine scores for each variant, we expanded our definition of
49
50 “unambiguous” variants to be those with at most one dissenting vote (score <0.2 or >0.8), which accounts for
51
52 87.1% of the variants. The 12.9% of SVs that were “ambiguous” (more than one dissenting vote, $0.2 \leq \text{score}$
53
54 ≤ 0.8) were generally small (median size of 310.5bp versus 899.5bp for all variants, **Fig 2C**) or contained
55
56 conflicting evidence (e.g., paired-end and split-read evidence indicated an inversion and the read-depth evidence
57
58 indicated a deletion, e.g., **Fig 1C**).
59
60
61
62
63
64
65

1 1 Other methods, such as SVTYPER²³ and CNVNATOR²⁴, can independently assess the validity of SV calls.
2
3 3 SVTYPER genotypes SVs for a given sample by comparing the number of discordant paired-end alignments
4
5 and split-read alignments that support the SV to the number of pairs and reads that support the reference allele.
6
7 CNVNATOR uses sequence coverage to estimate copy number for the region affected by the SV. Both of these
8
9 methods confirm the voting results (**Fig 2D**). Considering the set of “unambiguous” deletions, SVTYPER and
10
11 CNVNATOR agree with the *SV-plaudit* curation score in 92.3% and 81.7% of cases, respectively. Here,
12
13 agreement means that unambiguous false SVs (curation score < 0.2) have a CNVNATOR copy number near
14
15 two (between 1.4 and 2.4) or an SVTYPER genotype of homozygous reference. Unambiguous true SVs (curation
16
17 score > 0.8) have a CNVNATOR copy number near one or zero (less than 1.4), or an SVTYPER genotype of
18
19 non-reference (heterozygous or homozygous alternate).
20
21

22
23
24
25
26 Despite this consistency, using either SVTYPER or CNVNATOR to validate SVs can lead to false positives or
27
28 false negatives. For example, CNVNATOR reported a copy number loss for 44.2% of the deletions that were
29
30 scored as unanimously BAD, and SVTYPER called 30.7% of the deletions that were unanimously GOOD as
31
32 homozygous reference. Conversely, CNVNATOR had few false negatives (2.4% of unanimously GOOD
33
34 deletions were called as copy neutral), and SVTYPER had few false positives (0.2% of non-reference variants
35
36 were unanimously BAD).
37
38

39
40
41
42 These results demonstrate that, with *SV-plaudit*, manual curation can be a cost-effective and robust part of the
43
44 SV detection process. While we anticipate that automated SV detection methods will continue to improve, due
45
46 in part to the improved truth sets that *SV-plaudit* will provide, directly viewing SVs will remain an essential
47
48 validation technique. By extending this validation to full call sets, *SV-plaudit* not only improves specificity but can
49
50 also enhance sensitivity by allowing user to relax quality filters and rapidly screen large sets of calls. Beyond
51
52 demonstrating *SV-plaudit*'s utility, our curation of SVs for NA12878 is useful as a high-quality truth set for method
53
54 development and tuning. A VCF of these variants annotated with their curation score is available in
55
56

57 **Supplementary File 5.**
58

59
60
61
62 **DISCUSSION**
63
64
65

1 1 *SV-plaudit* is an efficient and scalable framework for the manual curation of large-scale SV call sets. Backed by
2
3 Amazon S3 and DynamoDB, *SV-plaudit* is easy to deploy and scales to teams of any size. Each instantiation of
4
5 *SV-plaudit* is completely independent and can be deployed locally for private or sensitive datasets, or be
6
7 distributed publicly to maximize participation. By rapidly providing a direct view of the raw data underlying
8
9 candidate SVs, *SV-plaudit* delivers the infrastructure to manually inspect full SV call sets. This functionality is
10
11 vital to a wide range of WGS experiments, from method development to the interpretation of disease genomes.
12
13 We are actively working on machine learning methods that will leverage the curation scores for thousands of SV
14
15 predictions as training data.
16
17
18
19
20

21 **CONCLUSIONS**

22
23 *SV-plaudit* was designed to judge how well the data in an alignment file corroborate a candidate SV. The question
24
25 of whether a particular SV is a false positive due to artifacts from sequencing or alignment is a broader issue
26
27 that must be answered in the context of other data sources such as mappability and repeat annotations. While
28
29 this second level of analysis is crucial, it is beyond the scope of this paper, and we argue this analysis be
30
31 performed only for those SVs that are fully supported by the alignment data. While *SV-plaudit* combines *samplot*
32
33 and *PlotCritic* to enable the curation of structural variant images, we emphasize that the *PlotCritic* framework
34
35 can be used to score images of any type. Therefore, we anticipate that this framework will facilitate "crowd-
36
37 sourced" curation of many other biological images.
38
39
40
41
42
43

44 **METHODS**

45
46 **Overview.** *SV-plaudit* (**Fig 3**) is based on two software packages: *samplot* for SV image generation, and
47
48 *PlotCritic* for staging the Amazon cloud environment and managing user input. Once the environment is staged,
49
50 users log into the system and are presented with a series of SV images in either a random or predetermined
51
52 order. For each image, the user answers the curation question and responses are logged. Reports on the
53
54 progress of a project can be quickly generated at any point in the process.
55
56
57
58
59
60
61
62
63
64
65

1 1 **Samplot.** *Samplot* is a Python program that uses *pysam*²² to extract alignment data from a set of BAM or CRAM
2
3 files, and *matplotlib*²³ to visualize the raw data for the genomic region surrounding a candidate SV (**Fig 3A**). For
4
5 each alignment file, *samplot* renders the depth of sequencing coverage, paired-end alignments, and split-read
6
7 alignments where paired-end and split-read alignments are color-coded based by the type of SV they support
8
9 (e.g., black for deletion, red for a duplication, etc.) (**Fig 1**). Alignments are positioned along the x-axis by genomic
10
11 location and along the left y-axis by the distance between the ends (insert size), which helps users to differentiate
12
13 normal alignments from discordant alignments that support an SV. Depth of sequencing coverage is also
14
15 displayed on the right y-axis to allow users to inspect whether putative copy number changes are supported by
16
17 the expected changes in coverage. To improve performance for large events, we downsample “normal” paired-
18
19 end alignments (a +/- orientation and an insert size range that is within Z standard deviations from the mean; by
20
21 default Z = 4). Plots for each alignment file are stacked and share a common x-axis that reports the chromosomal
22
23 position. By convention, the sample of interest (e.g., proband or tumor) is displayed as the top track, followed by
24
25 the set of related reference genomes tracks (e.g., parents and siblings, matched normal sample). Users may
26
27 specify the exact order by using command line parameters to *samplot*. A visualization of all genes and exons
28
29 within the locus is displayed below the alignment plots to provide context for assessing the SV's relevance to
30
31 phenotypes. Rendering time depends on the number of samples and the size of the SV, but most images will
32
33 require less than 5 seconds, and *samplot* rendering can be parallelizable by SV call.
34
35
36
37
38

39
40
41 **PlotCritic.** *PlotCritic* (**Fig 3B**) provides a simple web interface for scoring images and viewing reports that
42
43 summarize the results from multiple users and SV images. *PlotCritic* is both highly scalable and easy to deploy.
44
45 Images are stored on Amazon Web Services (AWS) S3 and DynamoDB tables store project configuration
46
47 metadata and user responses. These AWS services allow *PlotCritic* to dynamically scale to any number of users.
48
49 It also precludes the need for hosting a dedicated server, thereby facilitating deployment.
50
51
52

53
54
55 After *samplot* generates the SV images, *PlotCritic* manages their transfer to S3 and configures tables in
56
57 DynamoDB based on a JSON configuration file (`config.json` file in **Fig 3B**). In this configuration file, one
58
59 defines the curation questions posed to reviewers, as well as the allowed answers and associated keyboard
60
61 bindings to allow faster responses (`curationQandA` field in **Fig 3B**). In turn, these dictate the text and buttons
62
63
64
65

1 1 that appear on the resulting web interface. As such, it allows the interface to be easily customized to support a
2
3 wide variety of curation scenarios. For example, a cancer experiment may display a tumor sample and matched
4
5 normal sample and ask users if the SV appears in both samples (i.e., a germline variant) or just in the tumor
6
7 sample (i.e., a somatic variant). To accomplish this, the curation question (`question` field in **Fig 3B**) could be
8
9 “In which samples does the SV appear?”, and the answer options (`answers` field in **Fig 3B**) could be “TUMOR”,
10
11 “BOTH”, “NORMAL”, “NEITHER”. Alternatively, in the case of a rare disease, the interface could display a
12
13 proband and parents and ask if the SV is only in the proband (i.e., de novo) or if it is also in a parent (i.e.,
14
15 inherited). Since there is no limit to the length of a questions or number of answers options, *PlotCritic* can support
16
17 more complex experimental scenarios.
18
19
20

21
22
23
24 Once results are collected, *PlotCritic* can generate a tab-delimited report or annotated VCF that, for each SV
25
26 image, details the number of times the image was scored and the full set of answers it received. Additionally, a
27
28 curation score can be calculated for each image by providing a value for each answer option and an aggregation
29
30 function (e.g., mean, median, mode, standard deviation, min, max). For example, consider the cancer example
31
32 from above where the values three, two, one, and zero mapped to the answers “TUMOR”, “BOTH”, “NORMAL”,
33
34 and “NEITHER”, respectively. If "mode" were selected as the curation function, then the curation score would
35
36 reflect the opinion of a plurality of users. The mean would reflect the consensus among all users, and the
37
38 standard deviation would capture the level of disagreement about each image. While we expect mean, median,
39
40 mode, standard deviation, min, and max to satisfy most use cases, users can implement custom scores by
41
42 operating on the tab-delimited reported.
43
44
45

46
47
48 Each *PlotCritic* project is protected by AWS Cognito user authentication, which securely restricts access to the
49
50 project website to authenticated users. A project manager is the only authorized user at startup and can
51
52 authenticate other users using Cognito’s secure services. The website can be further secured using HTTPS and
53
54 additional controls, such as IP restrictions, can be put in place by configuring AWS IAM access controls directly
55
56 for S3 and DynamoDB.
57
58
59

60
61
62 **AVAILABILITY OF SOURCE CODE AND REQUIREMENTS**
63
64
65

1 1 Project name: SV-Plaudit
2
3 3 Project home page: <https://github.com/jbelyeu/SV-plaudit>
4
5 5 Operating systems: Mac OS and Linux
6
7
8 8 Programing language: Python, bash
9
10 License: MIT
11

12
13
14
15 **AVAILABILITY OF SUPPORTING DATA AND MATERIAL**
16

17 The datasets generated and/or analyzed during the current study are available in the 1000 Genomes Project
18 repository, <ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/>
19
20

21
22
23
24 All data generated during this study are included in this published article and its supplementary information files.
25

26
27
28 **DECLARATIONS**
29
30

31
32
33 **List of abbreviations**
34

35 SV: Structural Variant
36
37

38
39
40 **Ethics approval and consent to participate**
41

42 Not applicable
43
44

45
46
47 **Consent for publication**
48

49 Not applicable
50
51

52
53
54 **Competing interests**
55

56 The authors declare that they have no competing interests.
57

58
59
60
61 **Funding**
62
63
64
65

1 1 This research was supported by a US National Human Genome Research Institute awards to RML (NIH
2
3 K99HG009532) and ARQ (NIH R01HG006693 and NIH R01GM124355), as well as a US National Cancer
4
5 Institute award to ARQ (NIH U24CA209999).
6
7
8
9

510 **Authors' contributions**

612 JRP and RML developed the software. JRB, TJN, BSP, TAS, JMH, SNK, MEC, BKL, and RML scored variants
613
14 for the experiment. JRP, ARQ, and RML wrote the manuscript. ARQ and RML conceived the study.
715
16
17
18

919 **REFERENCES**

- 1022 1. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–
23 81 (2015).
1124
- 1226 2. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
1227
- 1328 3. Newman, T. L. *et al.* A genome-wide survey of structural variation between human and chimpanzee.
1329
30 *Genome Res.* **15**, 1344–1356 (2005).
1431
- 1533 4. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease.
1534
1635 *Nat. Rev. Genet.* **7**, 552–564 (2006).
1636
- 1737 5. Payer, L. M. *et al.* Structural variants caused by Alu insertions are associated with risks for many human
1738
1839 diseases. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E3984–E3992 (2017).
1840
- 1942 6. Schubert, C. The genomic basis of the Williams-Beuren syndrome. *Cell. Mol. Life Sci.* **66**, 1178–1197
1943
2044 (2009).
2045
- 2146 7. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome.
2147
2248 *Nature* **463**, 191–196 (2010).
2249
- 2351 8. Venkitaraman, A. R. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**, 171–182
2352
2453 (2002).
2454
- 2555 9. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and
2556
2657 evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).
2658
2759
2860
2961
3062
3163
3264
3365

- 1 1 10. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break
2
3 points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**,
4
5 2865–2871 (2009).
6
7
- 4 8 11. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis.
9
10 *Bioinformatics* **28**, i333–i339 (2012).
11
- 6 12 12. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome
13
14 structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
15
16
- 8 17 13. Kronenberg, Z. N. *et al.* Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput.*
18
19 *Biol.* **11**, e1004572 (2015).
20
- 10 21 14. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural
22
23 variant discovery. *Genome Biol.* **15**, R84 (2014).
24
- 12 25 15. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel
26
27 genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
28
29
- 14 30 16. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance
31
32 genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
33
34
- 16 35 17. Fiume, M., Williams, V., Brook, A. & Brudno, M. Savant: genome browser for high-throughput sequencing
36
37 data. *Bioinformatics* **26**, 1938–1944 (2010).
38
- 18 39 18. Munro, J. E., Dunwoodie, S. L. & Giannoulatou, E. SVPV: a structural variant prediction viewer for paired-
40
41 end sequencing datasets. *Bioinformatics* **33**, 2032–2033 (2017).
42
- 20 43 19. O'Brien, T. M., Ritz, A. M., Raphael, B. J. & Laidlaw, D. H. Gremlin: an interactive visualization model for
44
45 analyzing genomic rearrangements. *IEEE Trans. Vis. Comput. Graph.* **16**, 918–926 (2010).
46
47
- 22 48 20. Wyczalkowski, M. A. *et al.* BreakPoint Surveyor: a pipeline for structural variant visualization.
49
50 *Bioinformatics* **33**, 3121–3122 (2017).
51
- 24 52 21. Spies, N., Zook, J. M., Salit, M. & Sidow, A. svviz: a read viewer for validating structural variants.
53
54
55 *Bioinformatics* **31**, 3994–3996 (2015).
56
- 26 57 22. [PDF]pysam documentation - Read the Docs.
58
- 27 59 23. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9**, 90–95 (2007).
60
61
62
63
64
65

1 1 24. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**,
 2 2
 3 3 966–968 (2015).
 4 4
 5 5 25. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and
 6 6
 7 7 characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**,
 8 8
 9 9 974–984 (2011).
 10 10
 11 11
 12 12

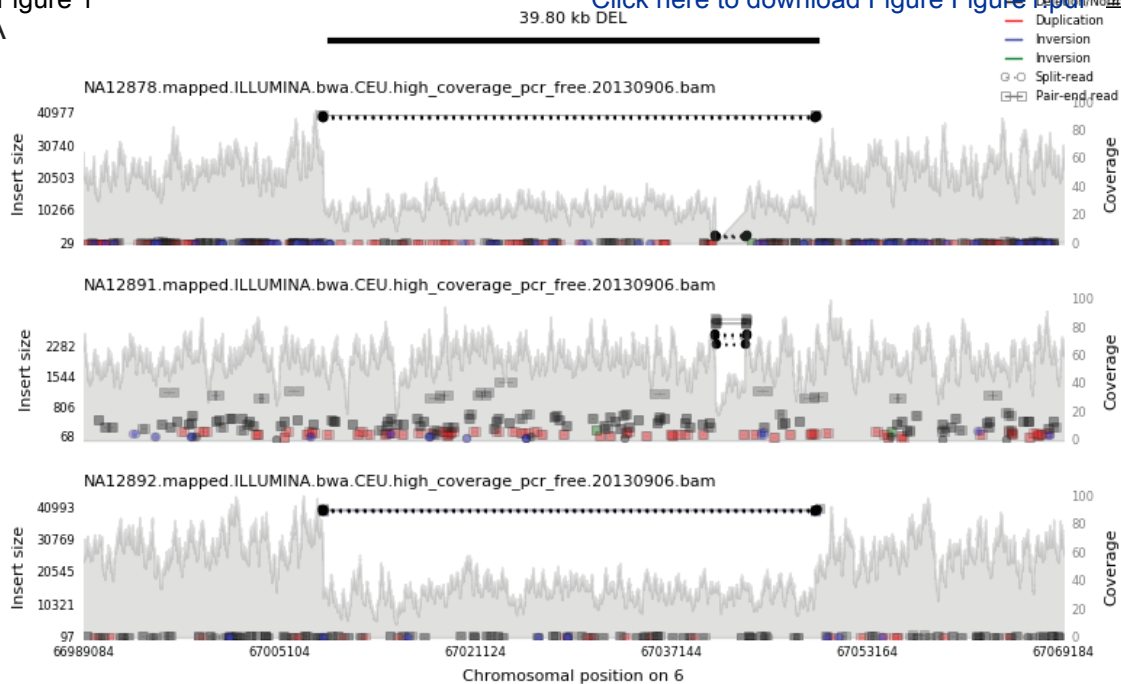
13 13
 14 14
 15 15
 16 16 **FIGURE LEGENDS**

17 17
 18 18 **Figure 1.** Example *samplot* images of putative deletion calls that were scored as **A)** unanimously GOOD, **B)**
 19 19
 20 20 unanimously BAD, and **C)** ambiguous with a mix of GOOD and BAD scores. The black bar at the top of the figure
 21 21
 22 22 indicates the genomic position of the predicted SV, and the following subfigures visualize the alignments and
 23 23
 24 24 sequence coverage of each sample. Subplots report paired-end (square-ends connected by a solid line) and
 25 25
 26 26 split-read (circle-ends connected by a dashed line) alignments by their genomic position (x-axis) and the distance
 27 27
 28 28 between mapped ends (insert size, left y-axis). Colors indicate the type of event the alignment supports (black
 29 29
 30 30 for deletion, red for duplication, and blue and green for inversion) and intensity indicates the concentration of
 31 31
 32 32 alignments. The grey filled shapes report the sequence coverage distribution in the locus for each sample (right
 33 33
 34 34 y-axis).
 35 35
 36 36
 37 37
 38 38
 39 39

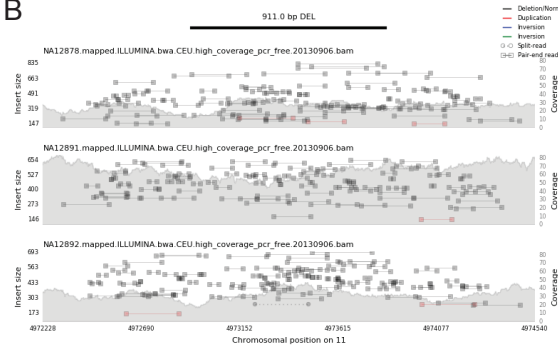
40 40 **Figure 2.** A) The distribution of the time between when an image was presented and when it was scored. B) The
 41 41
 42 42 distribution of curation scores. C) The SV size distribution for all, unanimous (score 0 or 1), unambiguous (score
 43 43
 44 44 <0.2 or >0.8) and ambiguous (score ≥ 0.2 and ≤ 0.8) variants. D) A comparison of predictions for deletions
 45 45
 46 46 between CNVNATOR copy number calls (y-axis), SVTYPER genotypes (color, “Ref.” is homozygous reference
 47 47
 48 48 and “Non-ref.” is heterozygous or homozygous alternate), and curation scores (x-axis). This demonstrates a
 49 49
 50 50 general agreement between all methods with a concentration of reference genotypes and copy number two (no
 51 51
 52 52 evidence for a deletion) at curation score less than 0.2, and non-reference and copy number one or zero events
 53 53
 54 54 (evidence for a deletion) at curation score greater than 0.8. There are also false positives for CNVNATOR (copy
 55 55
 56 56 number less than 2 at score = 0), and false negatives for SVTYPER (reference genotype at score = 1).
 57 57
 58 58
 59 59
 60 60
 61 61
 62 62
 63 63
 64 64
 65 65

1 **Figure 3.** The *SV-Plaudit* process. **A)** *Samplot* generates an image for each SV from VCF considering a set of
2 alignment (BAM or CRAM) files. **B)** *PlotCritic* uploads the images to an Amazon S3 bucket and prepares
3 DynamoDB tables. Users select a curation answer (“GOOD”, “BAD”, or “DE NOVO”) for each SV image.
4 DynamoDB logs user responses and generates reports. Within a report, a curation score function can be
5 specified by mapping answer options to values and selecting an aggregation function. Here “GOOD” and “DE
6 NOVO” were mapped to one, “BAD” to zero, and the mean was used. One useful output option for a report is a
7 VCF annotated with the curation scores (shown here in bold as a **SVP**).

A



B



C

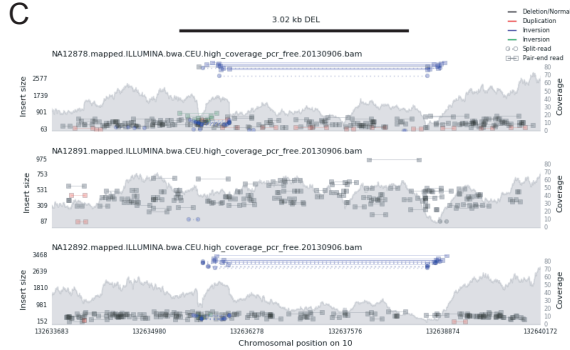


Figure 2

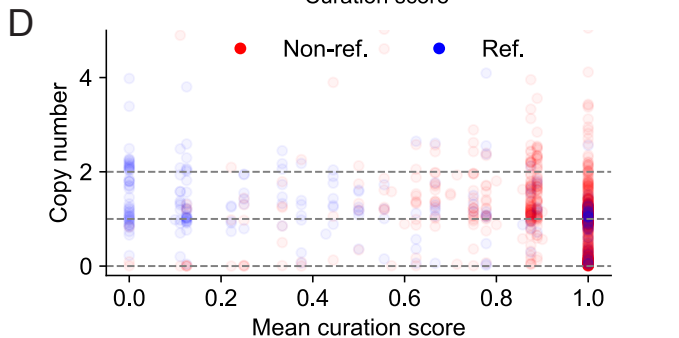
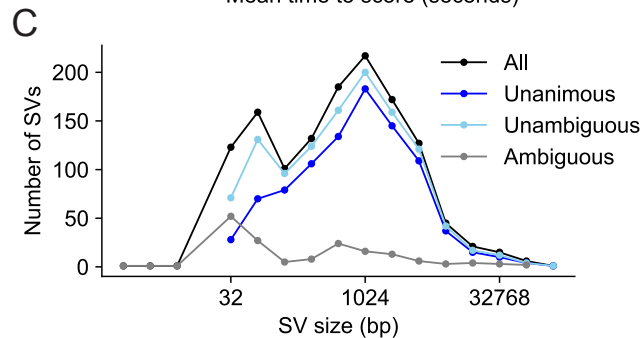
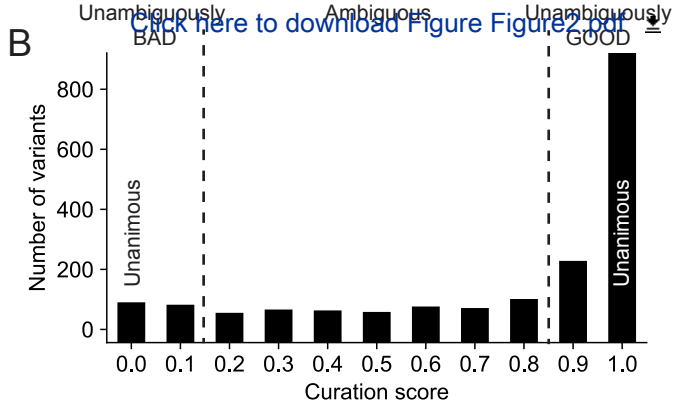
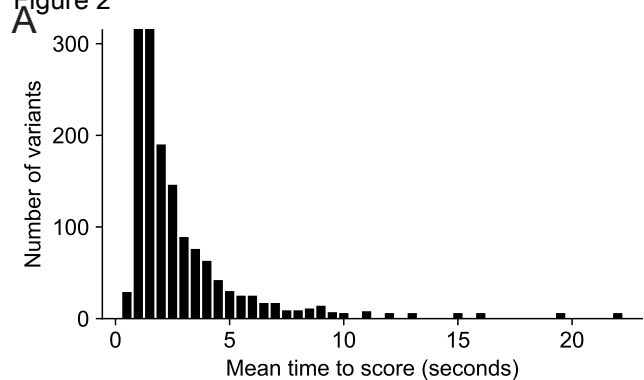
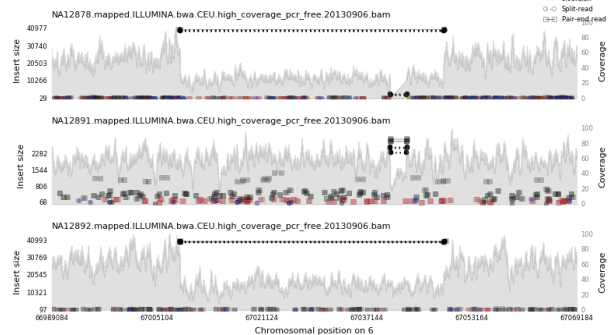


Figure 3

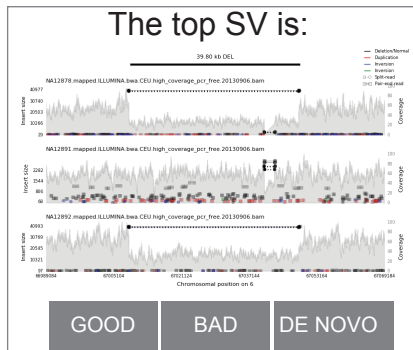
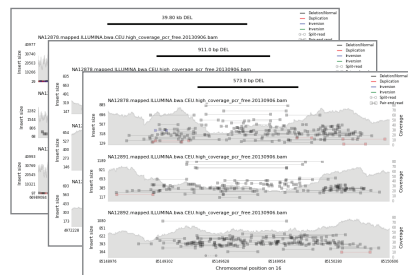
[Click here to download Figure Figure3.pdf](#)

#CHROM	POS	INFO
6	67009228	SVTYPE=DEL;END=67049033 → samplot →
11	4972926	SVTYPE=DEL;END=4973937
16	85149501	SVTYPE=DEL;END=85150074

BAMs
 NA12878.bam
 NA12891.bam
 NA12892.bam



B



User input

Type: DEL
 Chrom: 6
 Start: 67009228
 End: 67049033
 User: jon@utah.edu
 Result: GOOD

config.json

```
{ "curationVariables": {
  "curationQandA": {
    "question": "The top SV is:",
    "answers": {
      "g", "GOOD",
      "b", "BAD",
      "d", "DE NOVO" } },
  "AWSValues": { ... } }
```



S3 DynamoDB

PlotCritic


VCF report

#CHROM	POS	INFO
6	67009228	SVTYPE=DEL;END=67049033; SVP=1.0
11	4972926	SVTYPE=DEL;END=4973937; SVP=0.0
16	85149501	SVTYPE=DEL;END=85150074; SVP=0.5



Click here to access/download
Supplementary Material
NA12878.vcf





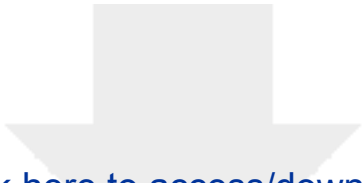
Click here to access/download
Supplementary Material
Supplemental_File_3.sh



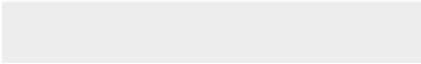



Click here to access/download
Supplementary Material
Supplemental_File_4.csv





Click here to access/download
Supplementary Material
Supplemental_File_6.txt





Click here to access/download
Supplementary Material
Supplemental_File_5.vcf