# SVCurator: A Crowdsourcing app to visualize evidence of structural variants for the human genome

**Authors**

Lesley M Chapman[1], Noah Spies[1,2,26*], Patrick Pai[3], Chun Shen Lim[4], Andrew Carroll[5], Giuseppe Narzisi[6], Christopher M. Watson[7,8], Christos Proukakis[9], Wayne E. Clarke[6], Naoki Nariai[10], Eric Dawson[11,12], Garan Jones[13], Daniel Blankenberg[14], Christian Brueffer[15], Chunlin Xiao[16], Sree Rohit Raj Kolora[17-19], Noah Alexander[20], Paul Wolujewicz[21], Azza Ahmed[22], Graeme Smith[23], Saadlee Shehreen[24], Aaron M. Wenger[25], Marc Salit[1,2], Justin M. Zook[1]

1. Biosystems and Biomaterials Division, Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr, MS8312, Gaithersburg, MD 20899, USA
2. The Joint Initiative for Metrology in Biology, Stanford University, Stanford, CA, USA
3. University of Maryland, College Park
4. Department of Biochemistry, School of Biomedical Sciences, University of Otago, Dunedin, New Zealand
5. DNAnexus Inc, 1975 Mountain View STE 101., Mountain View, California, USA.
6. New York Genome Center, New York, NY 10013
7. School of Medicine University of Leeds Saint James's University Hospital Leeds LS9 7TF United Kingdom
8. Yorkshire Regional Genetics Service, The Leeds Teaching Hospitals NHS Trust, Saint James's University Hospital, Leeds LS9 7TF, United Kingdom
9. University College London, Institute of Neurology London, UK
10. Illumina, Inc. 5200 Illumina Way San Diego, CA 92122
11. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD, USA
12. Department of Genetics, University of Cambridge, Cambridge, UK
13. University of Exeter Medical School, Epidemiology and Public Health Group, Barrack Road, Exeter, Devon, EX2 5DW, UK
14. Genomic Medicine Institute Lerner Research Institute Cleveland Clinic, 9500 Euclid Avenue / NE50, Cleveland, OH 44195
15. Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden
16. National Institutes of Health/National Library of Medicine/NCBI, Bethesda, MD USA 20894
17. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany.
18. Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Leipzig, Germany.
19. Molecular Evolution and Systematics of Animals, Institute of Biology, University of Leipzig, Leipzig, Germany.
20. Molecular Biology Institute, University of California Los Angeles, Los Angeles, CA.

21. Weill Cornell, Belfer Research Building, 413 E. 69th St, New York, NY 10021
22. Center for Bioinformatics and Systems Biology, Faculty of Science, University of Khartoum and Department of Electrical and Electronic Engineering, Faculty of Engineering, University of Khartoum, Al Gamaa Avenue, PO Box 321, postal code 11111, Khartoum, Sudan
23. Guy's Hospital and St Thomas's NHS Foundation Trust Great Maze Pond, London, SE1 9RT
24. Department of Genetic Engineering & Biotechnology, University of Dhaka, Bangladesh
25. Pacific Biosciences, Menlo Park, California, 94025
26. Departments of Genetics and Pathology, Stanford University, Stanford, CA

* Current address: Celsius Therapeutics, Cambridge, MA

**Abstract**

A high quality benchmark for small variants encompassing 88 to 90% of the reference genome has been developed for seven Genome in a Bottle (GIAB) reference samples. However a reliable benchmark for large indels and structural variants (SVs) is yet to be defined. In this study, we manually curated 1235 SVs which can ultimately be used to evaluate SV callers or train machine learning models. We developed a crowdsourcing app - SVCurator - to help curators manually review large indels and SVs within the human genome, and report their genotype and size accuracy.

SVCurator is a Python Flask-based web platform that displays images from short, long, and linked read sequencing data from the GIAB Ashkenazi Jewish Trio son [NIST RM 8391/HG002]. We asked curators to assign labels describing SV type (deletion or insertion), size accuracy, and genotype for 1235 putative insertions and deletions sampled from different size bins between 20 and 892,149 bp. The crowdsourced results were highly concordant with 37 out of the 61 curators having at least 78% concordance with a set of ‘expert’ curators, where there was 93% concordance amongst ‘expert’ curators. This produced high confidence labels for 935 events. When compared to the heuristic-based draft benchmark SV callset from GIAB, the SVCurator crowdsourced labels were 94.5% concordant with the benchmark set. We found that curators can successfully evaluate putative SVs when given evidence from multiple sequencing technologies.

**Background**

Structural variants (SVs) are typically defined as DNA variants ≥ 50 base pairs (bp), and include: insertions, deletions, duplications, and inversions[1]. SVs have been linked to a number of human diseases[2]. The Genome in a Bottle Consortium developed benchmark small variants for seven human genomes[3-4], but the primary high-confidence SVs published for these genomes were deletions discovered from short reads[5]. Recent next generation sequencing technologies and analysis algorithms have substantially improved the discovery of SVs[6-7]. However, identifying SVs with high confidence remains a challenge as evidenced by inconsistent predictions of SVs across different methods[8]. Several groups have demonstrated that crowdsourcing applications

can be effective for generating labeled data for putative SVs. Greenside et al. used crowdsourcing to label 1781 deletions for the Personal Genomes Project Ashkenazi Jewish Trio son [HG002][4]. Recently, *SV-Plaudit* was used to evaluate 1350 SVs (97% deletions), and allowed participants to evaluate candidate SVs using samplot, which displays images representing short and long read sequencing technologies[5,6]. The web-based platform, Plotcritic, renders samplot images and provides users with an interface to evaluate putative SVs[5].

In the current study, we generated a list of SVs that contain SV type, size, and genotype labels which can ultimately be used to train machine learning models to characterize properties of a benchmark genome. These data were generated via a Python Flask-based web application (app) - SVCurator - that allows users to evaluate large indels and SVs from the one human's genome - the GIAB Ashkenazi Jewish Trio son [NIST RM 8391/HG002]. The platform allows users to inspect and classify large indels and SVs by providing a variety of IGV and svviz2 images from short, long, and linked read sequencing data for putative SVs randomly sampled from candidate calls. These were generated from over 30 variant callers using data produced from five different sequencing technologies. To evaluate the accuracy of curations, we discuss the levels of concordance with heuristic based labels assigned to events within the GIAB v0.6 sequence resolved SV calls for HG002.

## Results

### SVCurator platform overview

SVCurator is a Python Flask-based web platform (**Fig 1**) we developed to evaluate putative large indels ≥20bp and SVs from the union of callsets from diverse technologies and calling methods for the Genome in a Bottle (GIAB) Ashkenazi Jewish Trio son (HG002/NA24385) [ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenaziTrio/analysis/NIST_UnionSVs_12122017/SVmerge121217/].

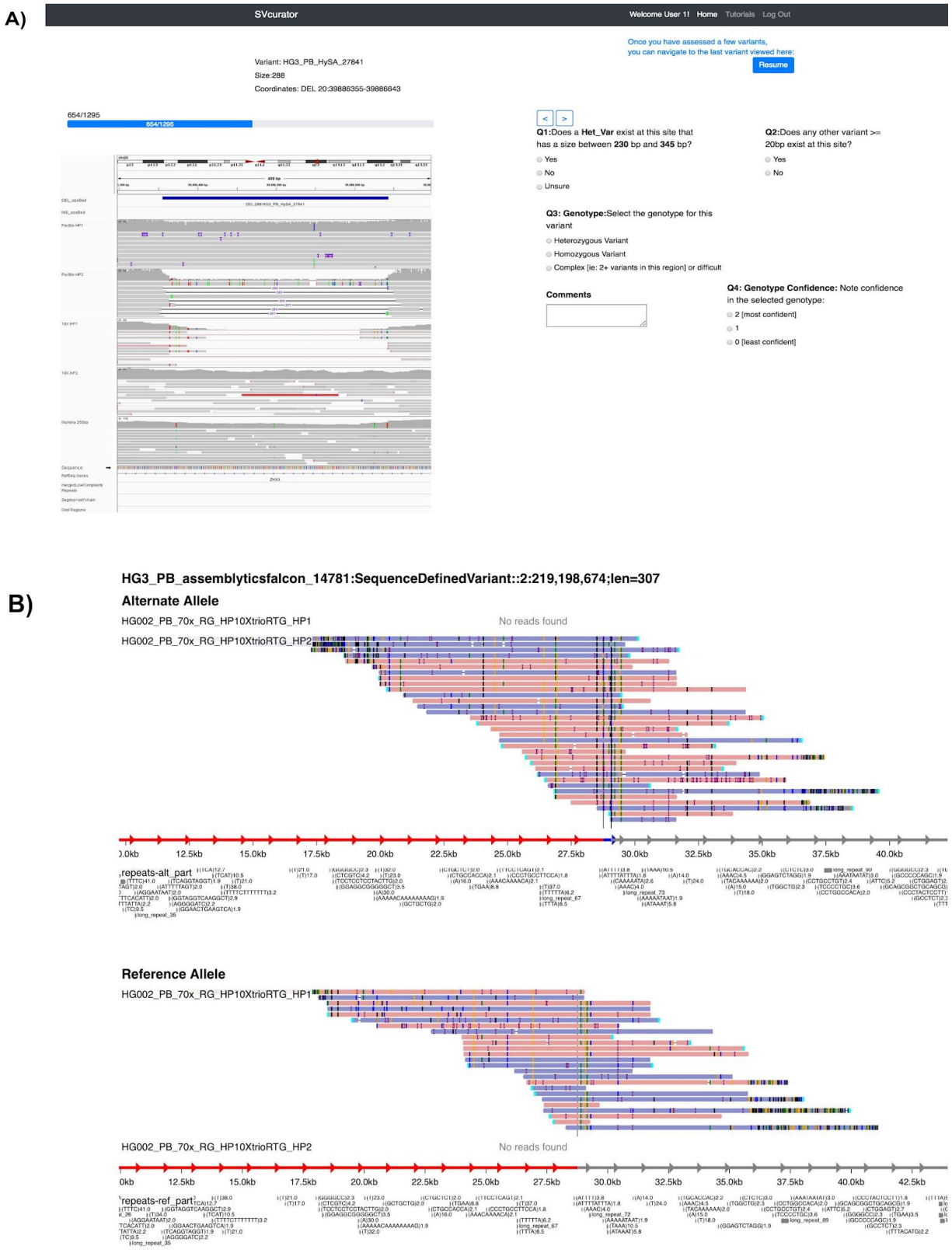



**Figure 1.** SVCurator web application interface. A) 13 read aligned and dotplot images from svviz2 for several technologies were also visible in addition to the IGV image shown here. B)

Svviz2 haplotype separated read aligned image for a 307bp insertion

Curators evaluated 1295 SV calls (579 deletions and 716 insertions) that were randomly selected from a pool of candidate variants binned by size (**Fig 2**). For each SV, SVCurator displays a number of images developed and recommended by experts from the GIAB consortium. Extensive data was generated from short, long, and linked-read whole genome sequencing technologies by the GIAB consortium. These data include Illumina 250bp paired end sequencing, Illumina 150bp paired end sequencing, Illumina 6kb mate-pair, haplotype-partitioned PacBio and haplotype-partitioned 10x Genomics (**Supplementary Fig** 1)[7]. Svviz2[8] was used to generate images of reads from each dataset aligned to the reference or alternate alleles. Svviz2 was also used to generate dotplots to visualize repetitive regions in the reference and alternate haplotypes and alignments of individual reads to the haplotypes. Images of Illumina 250x250bp paired end sequencing, haplotype-partitioned PacBio and haplotype-partitioned 10x Genomics in Integrative Genomics Viewer (IGV) were also included[7]. Participants were asked to evaluate each call and determine whether a SV exists at each site within 20% of the called size of the variant, assign a label describing the variant genotype [“Homozygous Reference”, “Heterozygous Variant”, “Homozygous Variant”, “Complex or difficult”] and a confidence score for the variant genotype (GT) assigned.

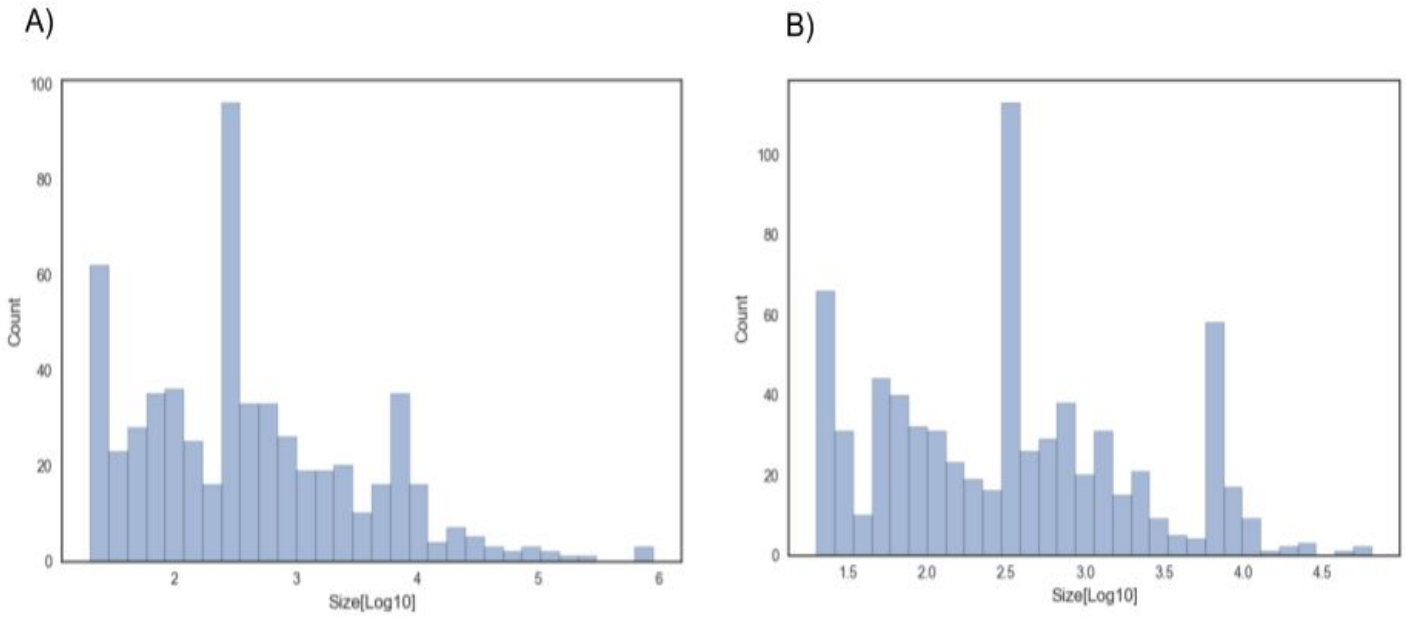



**Figure 2.** Events displayed in SVCurator randomly sampled from the union 171212 callset based on size. A) 579 deletions and B) 716 insertions.

Concordance amongst curators in evaluation of SVCurator Events

Curators were recruited from the GIAB analysis team and the genomics community through GIAB email lists and a GIAB Twitter account announcement. 136 participants registered to use the app, 61 of whom evaluated events. Of the 1295 events, 1290 events were curated at least 3 times (**Supplementary Fig** 2 [general distribution]). The average time to curate each event was 47.31 seconds (Supplementary Fig 3). To select curator responses for label evaluation, labels assigned by each curator were compared to labels assigned by a set of seven 'expert' curators from the GIAB Analysis Team who had experience curating SVs. The expert consensus label was assigned to each event by simple voting (i.e., from the label assigned by the most 'expert' curators). The percent concordance was defined as the ratio of 'expert' curators who agreed on the consensus label divided by the total number of expert curators who evaluated the event. On average, the 'expert' curators were 93% concordant on the labels assigned to each event. Each 'expert' was assigned a percent concordance score based on the level of concordance between their assigned label and the consensus label from the remaining experts.

Labels and concordance between 'experts' (percent and number of 'expert' curators that agreed on the final label) were found for all events. The concordance of each expert with the consensus expert label ranged from 77.7% to 100%. 541 events had at least 68% concordance with a consensus between at least 4 expert curators. All seven 'expert' curators agreed on the assigned label for 407 events. Overall, deletions averaged 86.36% concordance among 'experts' and insertions averaged 79.69% concordance. There were 298 deletions and 243 insertions where 'expert' curators had at least 68% concordance on the assigned label with 3 or more 'expert' curators who agreed on the assigned label.

There were 20 curators (including 5 'expert' curators) who evaluated more than 648 SVCurator events. Of these, on average 670 events per curator were available for further analysis after filtering responses where participants were unsure about an event existing at a particular site or were assigned low genotype confidence scores [Genotype confidence score = 0]. These curators had on average 86.92% concordance with 'expert' consensus labels (Fig 3).

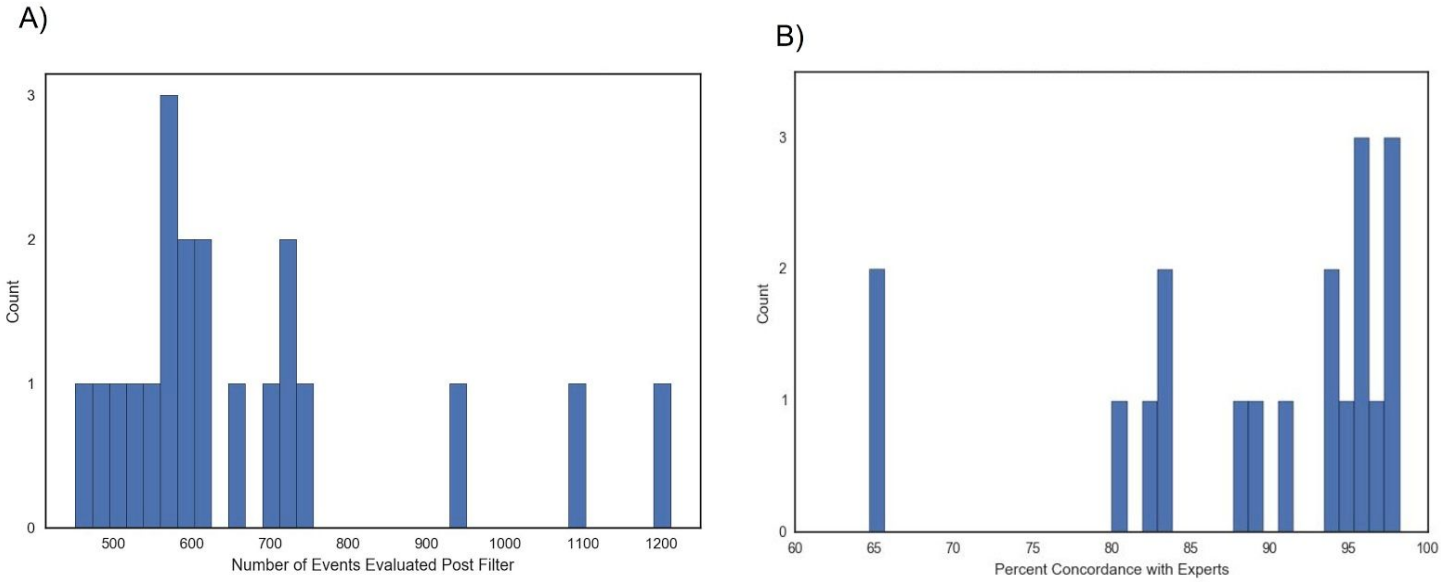



**Figure 3.** Responses from curators who evaluated over 648 events. A) Distribution of the number of events evaluated after filtering survey responses. B) Concordance of responses from curators that evaluated over 648 events with expert consensus genotype labels.

Because many curators were anonymous, we screened curators based on their concordance with the 'expert' consensus label for the 541 events. In order to filter the responses from curators that would be used to determine the final labels, responses from curators were filtered and binned into two threshold groups. Responses were placed into two groups of "top curators": 26 (out of 61) curators above Threshold 1 (90.9% or greater concordance, at least as concordant as the expert with the second lowest concordance), and 37 curators above Threshold 2 (77.7% or greater concordance, at least as concordant as the expert with the lowest concordance - see Supplementary Table 1). We filtered 133 out of 1295 sites because the consensus label of curators above Threshold 1 was different from the consensus label of curators above Threshold 2. 1162 events (527 deletions and 635 insertions) were retained (Supplementary Figure 4).

The responses from Threshold 1 and Threshold 2 top curators were highly concordant within each group (Fig 4). Threshold 1 top curators were more concordant than Threshold 2 top curators, particularly for insertions, but fewer Threshold 1 curators agreed on the assigned label. Complex events had the lowest levels of concordance for top curators within both groups with a mean concordance of 64% and 47% within top curators above Threshold 1 and 2, respectively.

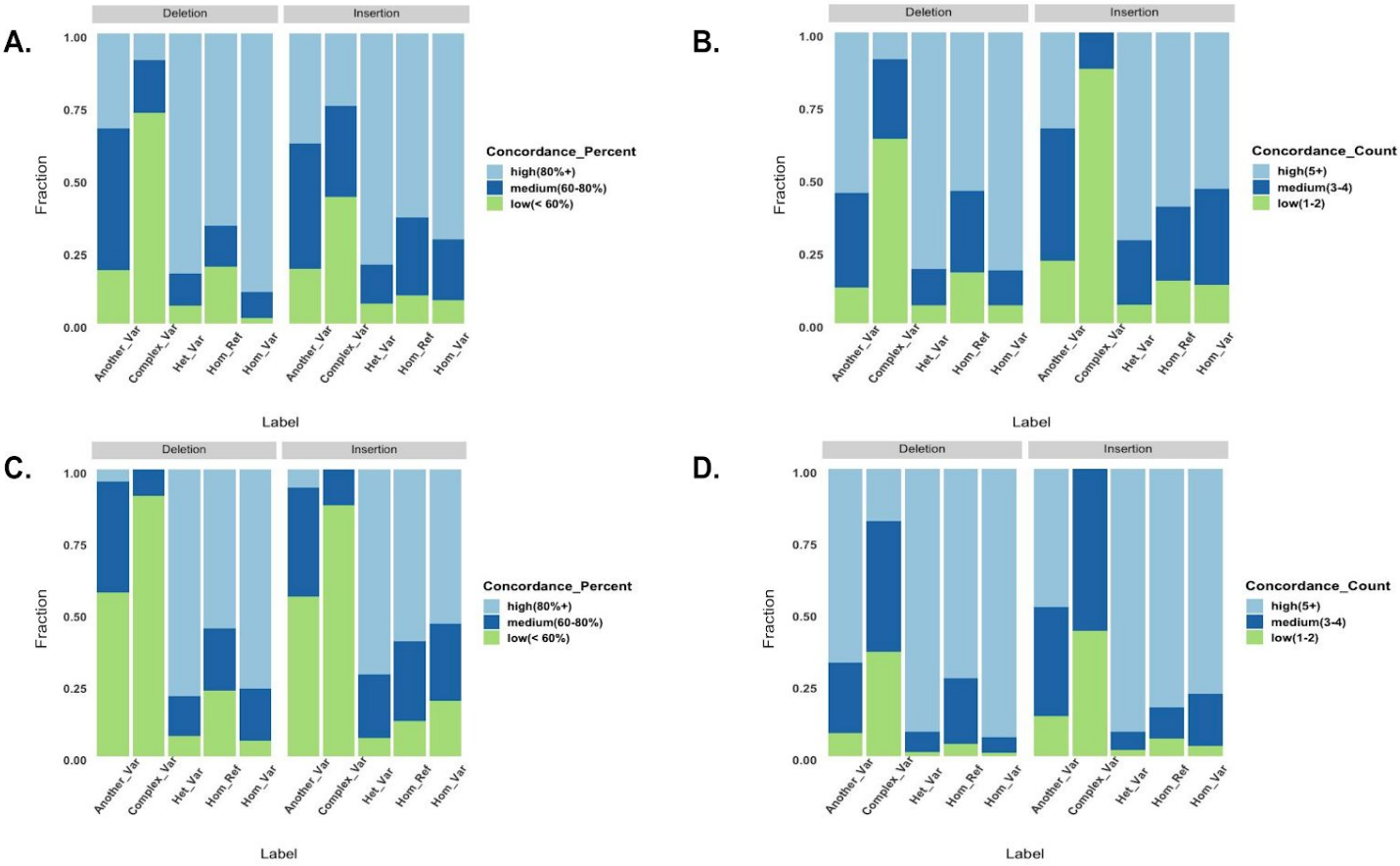



**Figure 4.** Concordance evaluation of labels assigned to SVCurator calls by top curators. A) Percent concordance amongst Threshold 1 top curators on assigned label. B) Fraction of top curators within Threshold 1 that agreed on the assigned label. C) Percent concordance amongst Threshold 2 top curators on assigned label. D) Fraction of Threshold 2 curators that agreed on the assigned label.**Concordance_Percent**: High (80% or more concordance); Medium (60-80% concordance); Low (60% or less concordance). **Concordance_Count**: High (5 or more curators agreed on the final label); Medium(3-4 curators agreed on the final label); Low(3 or fewer curators agreed on the final label assigned).

Label Evaluation

To evaluate the reliability of the top curators' labels for the 527 deletions and 635 insertions, they were compared to the GIAB v0.6 sequence resolved SV calls and benchmark regions for the Ashkenazi Jewish Trio son. 698 curated sites were inside the v0.6 benchmark regions, and the labels assigned by the top curators were 94.5% concordant with the v0.6 genotype labels (Fig 5).

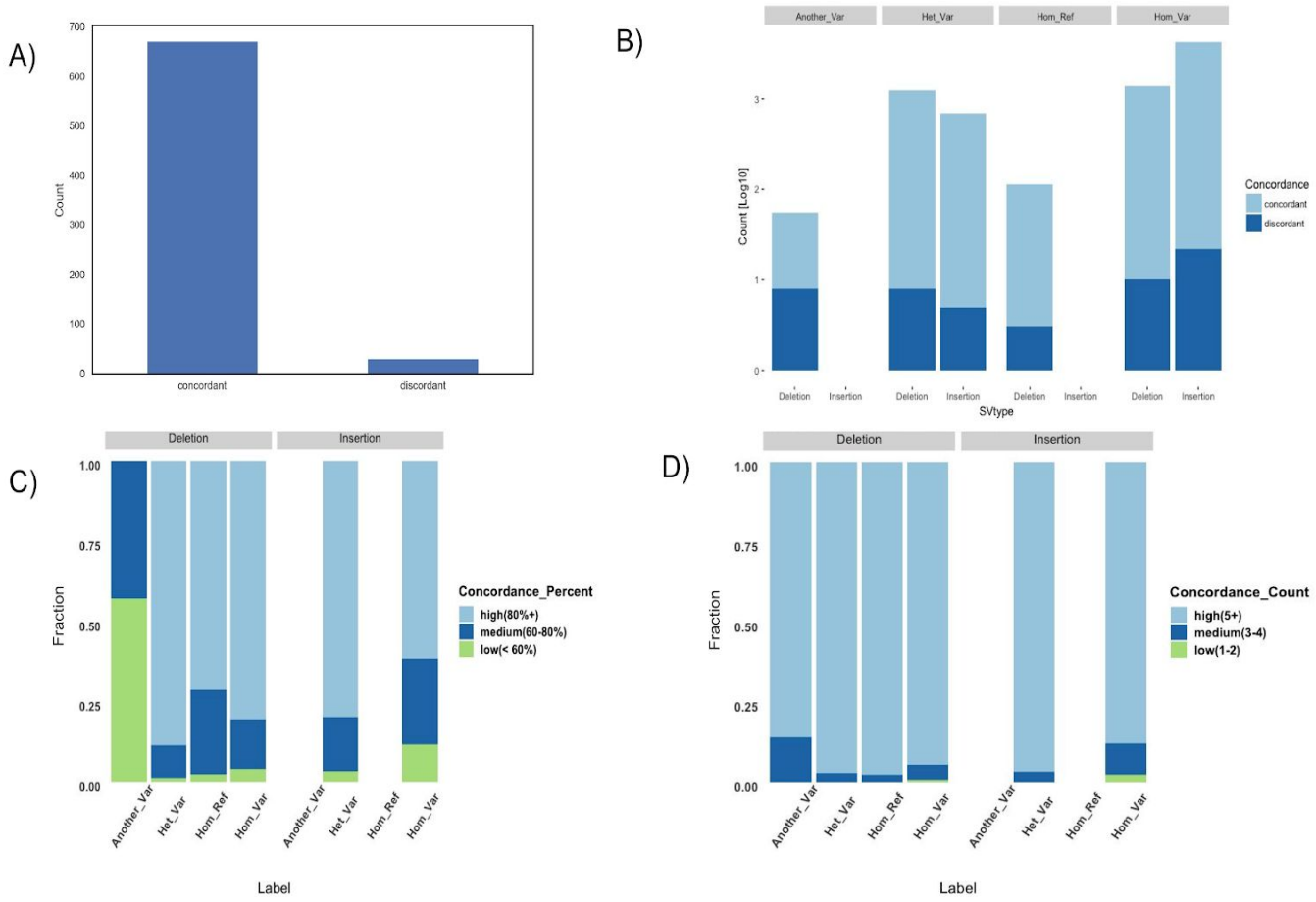



**Figure 5.** SVCurator labels assigned by top curators are supported by v0.6 GIAB high confidence genotypes. Crowdsourced labels were 94.5% concordant with the v0.6 labels. A) Count of the number of concordant and discordant sites between SVCurator crowdsourced labels and v0.6 heuristic assigned labels. B) Comparison of the number of concordant and discordant sites between the v0.6 GIAB genotype labels and labels assigned by top curators for the SVCurator SV calls on a log10 scale. C) v0.6 GIAB and SVCurator concordant labels showing the percent concordance amongst top curators on the label displayed. D) v0.6 GIAB and SVCurator concordant labels showing the count of the number of top curators that agreed on the final label. **Concordance_Percent:** High (80% or more concordance); Medium (60-80% concordance); Low (60% or less concordance). **Concordance_Count**: High (5 or more curators agreed on the final label); Medium(3-4 curators agreed on the final label); Low(3 or fewer curators agreed on the final label assigned).

The focus of the v0.6 sequence-resolved SV calls was on variants greater than 50bp in size, but we included the filtered v0.6 calls 20 to 49 bp in size in this comparison to help evaluate the reliability of top curators' labels in this size range. 10 of the 29 events discordant between the curators and v0.6 were 20 to 49 bp, and all but one of these appeared to be accurately labeled by curators or could be labeled in multiple ways. For instance, the event could be complex or

could contain two or more insertions of different sizes at the same loci. The v0.6 benchmark regions were designed to exclude complex events (i.e., regions with two or more SVs within 1000bp). 11 of 29 discordant events were labeled as complex variants by the top curators (2 of which were also 20 to 49bp in size). Fig 9 includes two examples of these events that were difficult to evaluate by the curators as shown by having 50% or less concordance amongst curators. Upon further curation, all but one or two of these appeared to be true complex variants. Of the remaining 10 discordant events, most appeared to be correctly classified by top curators. However, 2 events were classified as homozygous reference by curators even though another variant was in the same tandem repeat outside the IGV view displayed to curators. This difficulty in accurately classifying complex events in tandem repeat regions highlights the importance of expanding the view to display the entire tandem repeat region for variants overlapping them. Many of the differences between v0.6 and top curators were related to challenges in translating the v0.6 benchmark calls and regions into labels for the curated events. For example, because v0.6 focused on variants >49bp, v0.6 labels were often different if curators labeled a complex variant in which part of the variant was <50bp. There were also cases where multiple nearby variants could be combined into a single variant or separated into multiple variants. Figure 6 summarizes characteristics of the calls discordant between v0.6 and top curators.

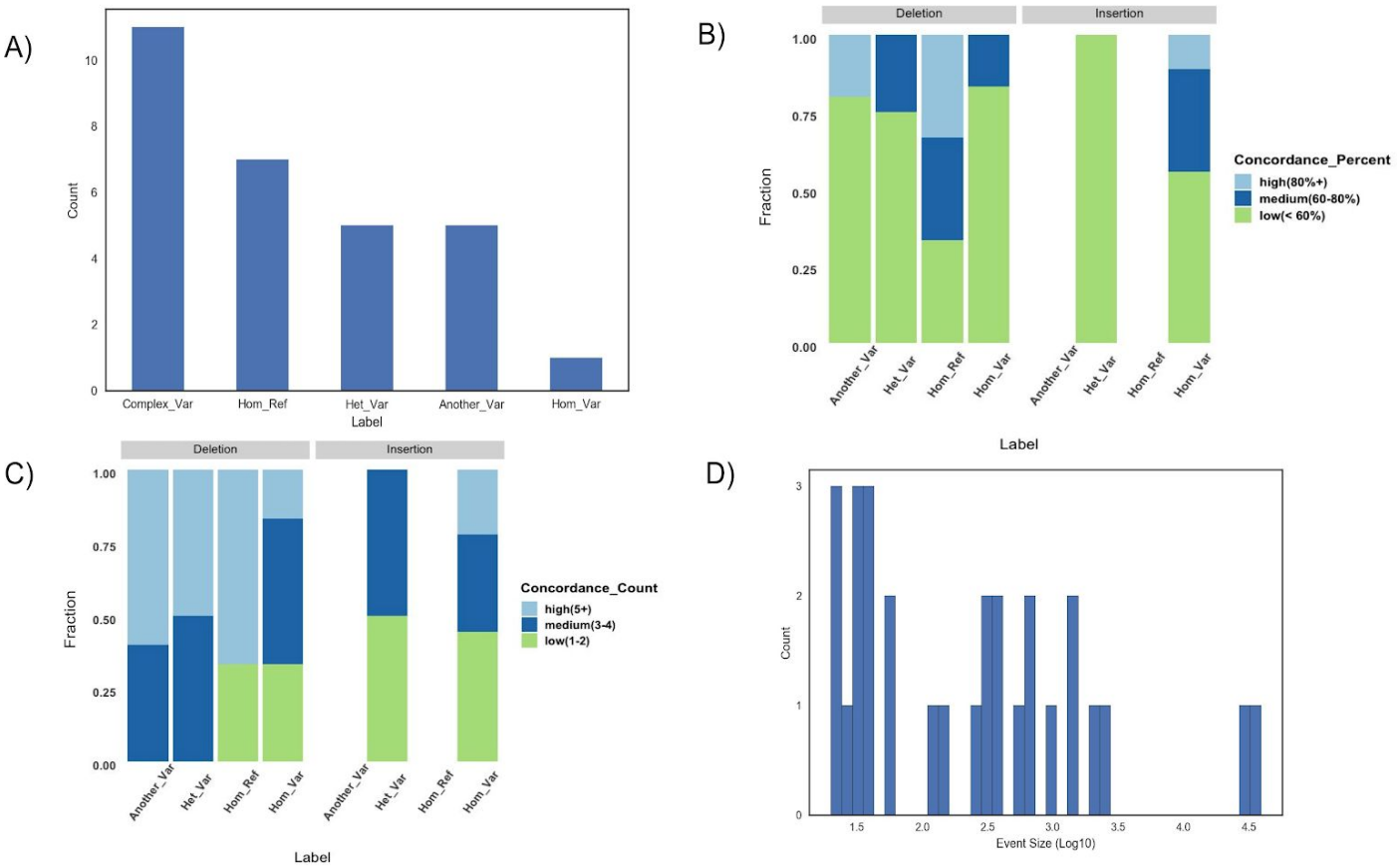


**Figure 6.** Evaluation of SVCurator and v0.6 high confidence genotypes discordant sites. A) SVCurator labels that were discordant with the v0.6 high confidence heuristics based genotype labels. The labels are as follows: Complex_Var: complex variant; Hom_Ref: homozygous reference; Het_Var: heterozygous variant; Another_Var: another variant; Hom_Var: homozygous variant. B) SVCurator labels that were discordant with v0.6 GIAB genotype labels. Figure shows the percent concordance amongst top curators for each SVCurator discordant label. C) SVCurator labels that were discordant with v0.6 GIAB genotype labels. Figure shows the fraction of the number of top curators that agreed. D) Size distribution of the discordant sites. **Concordance_Percent:** High (80% or more concordance); Medium (60-80% concordance); Low (60% or less concordance). **Concordance_Count**: High (5 or more curators agreed on the final label); Medium(3-4 curators agreed on the final label); Low(3 or fewer curators agreed on the final label assigned).

To assign final crowdsourced labels, a random sample of events were manually inspected. Events that were assigned labels with less than 50% concordance amongst all top curators were not included as final labels, which included 84 events. Upon manual inspection of 44 sites with only 50-60% concordance amongst all top curators, it was found that 61% of the events were assigned the correct label. Many of the incorrectly labeled events were not correctly classified as complex variants. Upon manual inspection of 28 sites with 60-70% concordance amongst all top curators, it was found that 85% of the events were assigned the correct label. Therefore, only events that were assigned labels with greater than 60% concordance amongst all top curators and at least 3 top curators agreed on the label were included in the final labeled callset. These sites included 496 insertions and 439 deletions with 94% of the events receiving labels of Homozygous Reference, Heterozygous Variant, or Homozygous Variant (Fig 7).

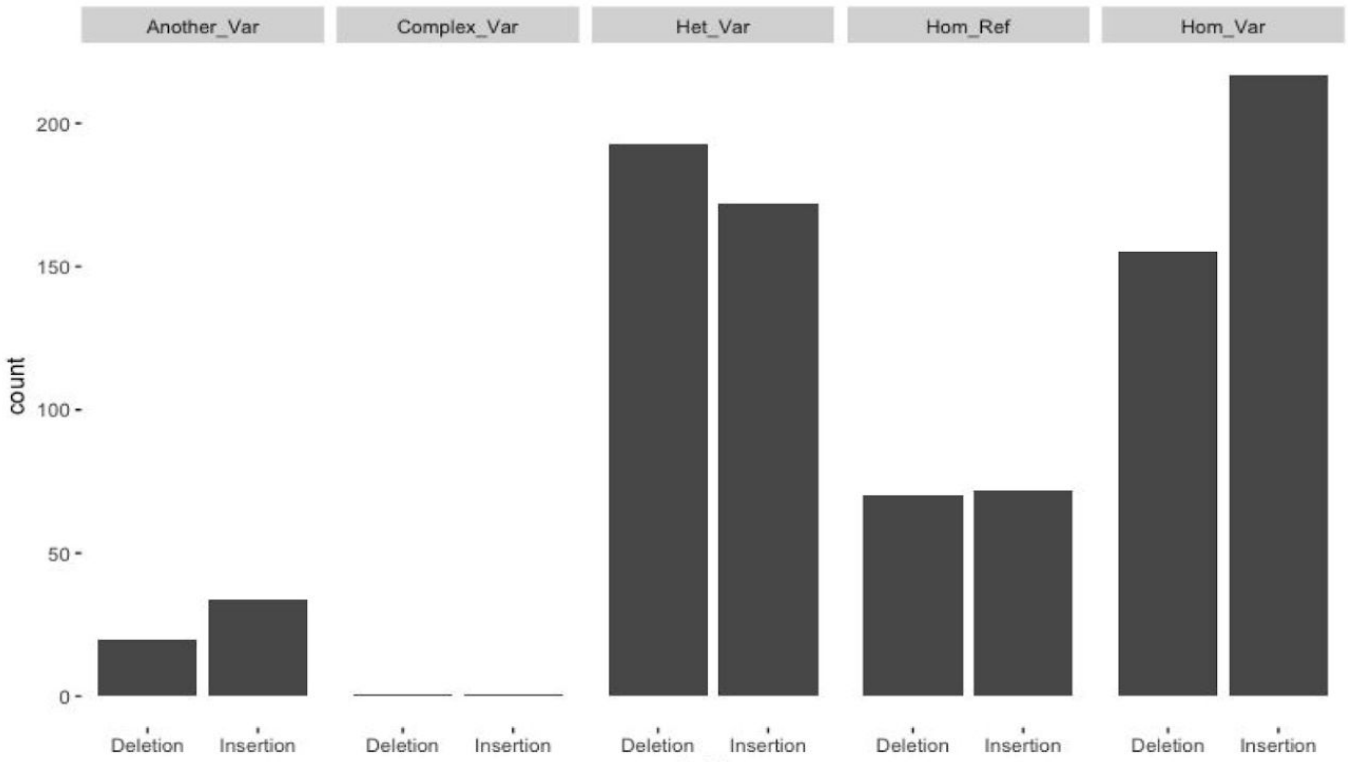



**Figure 7.** A summary of the final crowdsourced SVCurator labels.

We also used svviz2 to evaluate the curators' final labeled callset, including variants outside the v0.6 benchmark regions. svviz2 determines whether each read more closely matches the reference allele or the alternate allele or if it is ambiguous. svviz2 generates genotypes [Homozygous Reference, Heterozygous Variant, Homozygous Variant] based on the number of reads that align to the reference and alternate alleles, weighted by their mapping quality scores. We generated svviz2 genotypes from 5 datasets [Illumina 250bp paired end sequencing, Illumina 150bp paired end sequencing, Illumina mate-pair, haplotype-partitioned PacBio and haplotype-partitioned 10x Genomics], and genotypes from each technology were compared to the 879 SVCurator crowdsourced labels that were either: Homozygous Reference, Heterozygous Variant, Homozygous Variant. The crowdsourced label for 92.2% of the events were supported by at least 2 technologies (Fig 8) which included 811 out of 879 labels. There were also 58 events where only 1 technology supported the crowdsourced label; PacBio supported the majority of these events (26 events, mostly in tandem repeats) followed by Illumina Mate Pair (18 events). These results further support the accuracy of the crowdsourced labeled events, including those outside the v0.6 benchmark regions.

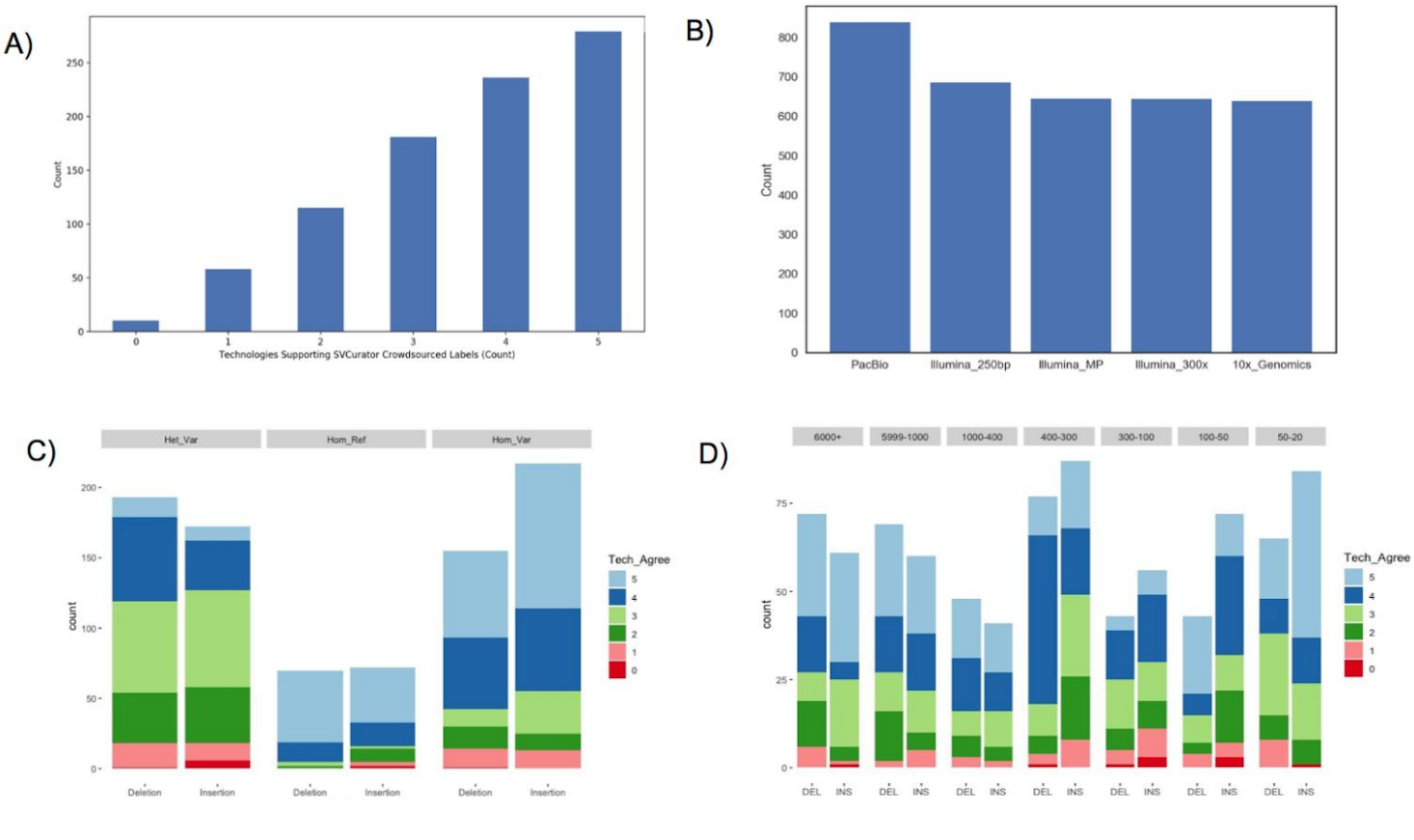



**Figure 8.** svviz2 genotypes support SVCurator crowdsourced labels. A) A summary of the number of technologies whose svviz2 genotypes support the SVCurator genotype label. 92.2% of the events were supported by at least 2 technologies. B) A count of the number of genotypes from each technology that match the SVCurator crowdsourced labels. C) A summary of the number of technologies that had genotype scores supporting the crowdsourced label as summarized based on label and variant type; and, D) by size of the event.

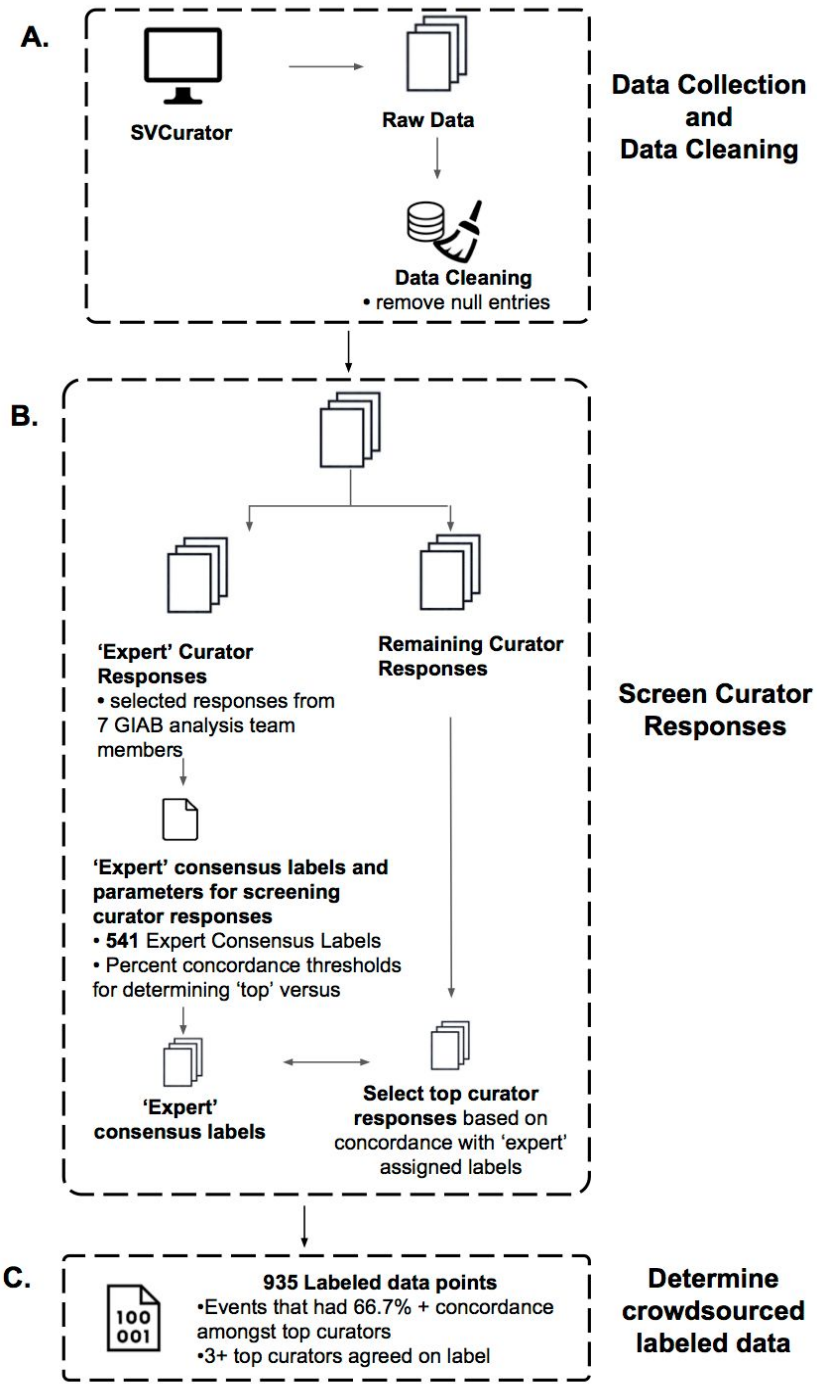



**Figure 9**. Schematic summarizing how SVCurator responses were processed to determine the final label for each event. A) Data Collection and Data Cleaning: Curators evaluated the 1295 events within SVCurator. After removing events that received a low confidence score for genotype assigned and an 'unsure' response for whether an event exists at a particular site, 1273 event remained for analysis. B) Screen Curator Responses: To determine the curator responses that were used to find final labels for the SVCurator events, first consensus labels assigned by 7 'expert' curators were determined. These 7 'expert' curators were members of the Genome in a Bottle (GIAB) analysis team. Of the 1273 events, 541 were assigned a consensus

label by the 'expert' curators, where each event had 68% or greater concordance on the assigned label and 4 or more experts that agreed on the assigned label. Using a leave-one-out strategy, a percent concordance score was found for each 'expert' curator, and the two lowest percent concordance scores (90.9% and 77.7%) were used as a threshold for screening top curators. To find the top curators, labels assigned by each curator were compared to the 541 events and percent concordance with experts was found for each curator. Curators that had 90.9% or greater concordance and 77.7% or greater concordance were considered top curators and their responses were placed in two threshold groups. The responses for these curators were used to find final labels for the SVCurator events. C) Determine crowdsourced labeled data: There were 935 events that were assigned final labels by top curators. These events had at least 66.7% concordance amongst top curators and at least 3 top curators that agreed on the final label assigned.

## Discussion

SVCurator is a crowdsourcing tool that incorporates read aligned images from multiple short, long and linked read sequencing data into an SV visualization tool that allow users to evaluate SV calls. SVCurator uniquely enables curators to evaluate multiple sources of evidence for each call in one app interface. We displayed svviz2 images of reads from 3 different Illumina sequencing methods, haplotype-partitioned PacBio, and haplotype-partitioned 10x Genomics aligned to reference and alternate alleles, as well as dotplots to visualize repetitive regions. The app also includes IGV images for comparison that display Illumina 250bp, PacBio and 10x Genomics reads. Both the IGV and svviz read aligned images include indicators of repetitive regions. Curators were also able to evaluate haplotype-partitioned PacBio and 10x Genomics data. These features allowed participants to more easily evaluate deletions and insertions, including repetitive regions and complex events.

The results of this study suggest that a group of participants can accurately curate SV calls by evaluating a variety of static images from multiple sequencing technologies. In general, simple heterozygous and homozygous variants and homozygous reference regions were accurately labeled, but complex variants were more challenging. To add additional support for these assigned labels, future work might include determining the Mendelian consistency for each event and completing PCR validation for a select group of events. A number of events with lower concordance scores were complex events that were assigned another label, and were often located in repetitive regions of the genome. Curators may not have taken into account the evidence within images that suggest a complex event. Crowdsourcing studies specifically focused on complex events could be conducted in the future to better characterize complex events. This would involve asking the participants to provide feedback on the way tutorials should be structured to facilitate the analysis of complex events.

The crowdsourced labels derived from this study will be useful training datasets for machine learning studies that evaluate SVs, and could be used as a resource to improve SV calling methods. The calls could also be used as a resource to help members of the clinical genomics community improve their evaluation of SVs. Crowdsourcing could also yield more reliable resources that could improve clinical interpretations of SVs as many of the guidelines are

qualitative[9]. Finally, this study demonstrates that crowdsourcing is a useful strategy for evaluating SV calls and the results of crowdsourcing could yield results that may be useful in improving SV tools and analyses approaches in multiple domains.

## Methods

### SVCurator Events

SVs and large indels from the Ashkenazi Jewish Trio son (NA24385/HG002) were randomly sampled from the Genome In A Bottle (GIAB) union callset [union_171212_refalt.sort.vcf]. The calls and distance metrics can be found at: ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/. The calls are a union set of 1+ million sequence-resolved calls ≥20bp from 5 short, long, and linked read technologies and over 30 SV discovery methods. The events were randomly sampled to obtain equal numbers of events in the following size bins: 20-49bp, 50-100bp, 100-300bp, 300-400bp, 400-1000bp, 1000-5999bp, 6000+ bp. 579 putative deletions and 716 putative insertions were included in the app.

### Participant Recruitment

Participants were recruited from the Genome in a Bottle Analysis Team (https://groups.google.com/forum/#!forum/giab-analysis-team) and from the genomics community via the @GenomeinaBottle Twitter account. SVCurator was made available to the public for one month to allow participants to evaluate the events within the app. An incentive of co-authorship on the current publication was offered for participants who curated at least half of the events (648 or more events).

### SVCurator App Interface

SVCurator (www.svcurator.com) is a Python Flask-based app (**Fig 1**) and uses SQLite3 as a database management system. User login was implemented using Google OAuth 2.0. The SVCurator app was deployed using pythonanywhere[www.pythonanywhere.com].

*App Images*: The interface consisted of four thumbnail images for each event and a set of four questions. The four thumbnail images consisted of the following: IGV image, svviz2 PacBio haplotype-partitioned read aligned image, svviz2 10X Genomics haplotype-partitioned read aligned image, svviz2 dotplot image representing reference versus alternate allele. A lightbox contained additional images to describe each event, and included the following: svviz2 read aligned image for haplotype and non-haplotype-partitioned PacBio data, 10X Genomics haplotype-partitioned data, Illumina 6kb Mate Pair data, Illumina HiSeq 250bp read length data, Illumina HiSeq 300x read depth data; svviz2 dotplots: represent reads with highest mapping quality versus reference and alternate allele, reference allele versus alternate allele, reference allele versus reference allele, and alternate allele versus alternate allele. Images included in the lightbox allowed curators to zoom in on sections of the SV call that required a more close evaluation. Each curator evaluated the same events for the first 43 events, and events 44-1295 were randomized for each user.

*Questions*: Participants were given the structural variant call: unique ID, size, chromosome number, start and end coordinates. For each event, curators evaluated the putative SV type, determined whether an event exists within 20% the size of the variant, and the genotype for each event. The questions included in SVCurator were designed to describe the size accuracy and genotype of each SV call. Members of the GIAB community helped structure and finalize the questions included in the app. Curators described each event by responding to the following questions:

Question 1 : Does a [insertion/deletion] exist at this site that has a size between [start coordinate] bp and [end coordinate]bp?

- Yes [Answer Q3 and Q4]
- No [Answer Q2 only]
- Unsure [move on to next variant]

Question 2: Does any other variant exist at this site?

- Yes
- No

Question 3: Select the genotype for this variant.

- Homozygous Reference
- Heterozygous Variant
- Homozygous Variant
- Complex [ie: 2+ variants in this region] or difficult
- Unsure

Question 4: Note confidence selected in the genotype

- 2 [most confident]
- 1
- 0 [least confident]

Comment Box: included to give curators the opportunity to add additional comments to describe each event or report any user interface issues (ie: images that may have not rendered properly)

Responses were collected over the course of one month after the app was made publically available. Participants were also provided with a tutorial that describes general guidelines for analyzing SV calls (https://lesleymaraina.github.io/svcurator_tutorial_2/) .

**Data used to generate images**

Aligned reads for the Ashkenazi Jewish Trio son (NA24385/HG002) were used to generate the images used within the app. The BAM files are publically available from the GIAB FTP site as follows:

| Sequencing | Library Type | Average | Average | BAM File Links |
|---|---|---|---|---|

| Technology | | Read Length [bp] | Read Depth | |
|---|---|---|---|---|
| Pacific Bioscience (PacBio) | Whole Genome Sequencing (WGS) Single End | 10-11Kb | 69x | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_MtSinai_NIST/Baylor_NGMLR_bam_GRCh37/HG002_PB_70x_RG_HP10XtrioRTG.bam |
| 10X Genomics Chromium Sequencing | WGS Linked Reads | 2x98 | 50x | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/10XGenomics_ChromiumGenome_LongRanger2.2_Supernova2.0.1_04122018/GRCh37/NA24385_300G/HG002_10x_84x_RG_HP10xtrioRTG.bam |
| Illumina HiSeq 2500 | WGS mate-pair | 2x100 [6000bp insert size] | 13-14x | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_Stanford_Illumina_6kb_matepair/ |
| Illumina HiSeq 2500 | WGS paired-end | 2x250 | 40-50x | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/ |
| Illumina HiSeq 2500 | WGS paired-end | 2x148 | 296.83x | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/ |

**svviz2 Images**

svviz2 (version 2.0a3, https://github.com/nspies/svviz2) aligned-read images and dotplots were generated for each event. Svviz2 is a SV visualization tool that identifies reads that support a reference allele, alternate allele (supports a SV call), or are ambiguous. 10X Genomics (10X) and PacBio sequencing images were haplotype-partitioned, PacBio reads were haplotype-partitioned using WhatsHap and reads for both 10X and PacBio were subsequently aligned to a reference and alternate allele using svviz2.

**IGV images**

Integrative Genomics Viewer (IGV) version 2.4.6 was also used to generate images for each event and displays tracks representing reads from haplotype-partitioned PacBio and 10X Genomics Chromium, Illumina Paired-End reads (250 base pair read length), and tracks describing repetitive regions of the genome (**Supplementary Fig** 1). Within each IGV image, the putative insertion or deletion was displayed along with flanking regions. For deletions, the flanking regions that were 20% of the size of the variant at the start and end position of each displayed event, and for insertions the flanking regions were 1.6 times the size of the variant at the start site and the region flanking the end position of the event was 70% of the size of the event.

**Crowdsourced Labels**

Each event was assigned one of the following genotype labels:

- Homozygous Reference [Hom_Ref]
- Heterozygous Variant [Het_Var]
- Homozygous Variant [Hom_Var]
- Complex [ie: 2+ variants in this region] or difficult [Complex_Var]
- Another variant exists at this site/A variant more than 20% different in size exists at this site [Another_Var]

Responses were processed as follows to generate the labels above:

| **Genotype Label** | **Question 1 : Does a [insertion/deletion] exist at this site that has a size between [start coordinate] bp and [end coordinate]bp?** | **Question 2: Does any other variant exist at this site?** | **Question 3: Select the genotype for this variant.** | **Question 4: Note confidence selected in the genotype** |
|---|---|---|---|---|
| Homozygous Reference | No | No | --- | --- |
| Heterozygous Variant | Yes | --- | Heterozygous Variant | 1 + |
| Homozygous Variant | Yes | --- | Homozygous Variant | 1+ |
| Another Variant Exists | No | Yes | --- | --- |
| Complex | Yes | --- | Complex | 1+ |

Responses were initially filtered as follows, if one of the following was true, responses were not included in the label assessment:

- curator provided no response to the questions
- curator selected 'Unsure' for the question 1
- curator selected 'least confident' for confidence in the genotype selected

To determine the curator responses that would be used to generate final labels for each event, curator responses were screened based on concordance with 'expert' consensus labels (**Fig 9**) since there are currently no comprehensive ground truth labels available for these events. Seven 'expert' curators were identified from the GIAB Analysis Team based on their known prior experience curating SVs. Each event was assigned one of the following labels ('Hom_Ref' [Homozygous Reference], 'Het_Var' [Heterozygous Variant], 'Hom_Var' [Homozygous Variant], 'Another_Var' [Another Variant Exists within 20% of the size of the variant], 'Complex_Var' [Complex Event]). The number of 'expert' curators that agreed on a label was determined as well as the percent concordance between 'experts' for each event. The percent concordance was determined based on the ratio of the number of 'expert' curators that agreed on a label versus the total number of 'expert' curators that evaluated each event. Consensus labels were assigned based on majority vote and events used for screening curators were those that had at least 3 expert curators agree on a label with at least 67% concordance. A leave-one-out strategy was used to determine the level of concordance between 'expert' curators. Two thresholds were set to determine remaining curators whos evaluations would be used for assigning crowdsourced labels. These thresholds were set based on the two lowest concordance levels between 'expert' curators.

Responses from curators with 77.7% concordance with experts and 90.9% concordance with 'expert' curators were included for further analysis. Only the events that had concordant labels between curators in the two threshold groups were used for further label analysis. Labels were determined for these events and events with at least 50% concordance amongst all top curators were evaluated further. Select events were manually inspected, and it was determined that sites with 60% or greater concordance with at least 3 curators that agreed on the label were included in the final labeled dataset.

**SVCurator Label Corroboration - GIAB v0.6 high confidence call genotype labels and svviz2 genotype labels**

*v0.6 Genotype Labels*
*The GIAB v0.6 Benchmark SV Set was generated using the following heuristics from the same union vcf sampled for the SVCurator variants above, which came from 30 callers and 5 technologies on all three members of the GIAB Ashkenazi trio at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_UnionSVs_12122017/union_171212_refalt.sort.vcf.gz*

1. *Sequence-resolved variants with at least 20% sequence similarity were merged into a single vcf line using SVanalyzer (https://github.com/nhansen/SVanalyzer)*
2. *Variants supported by at least two technologies (including BioNano and Nabsys) or by at least 5 callsets from a single technology had evidence for them evaluated and were genotyped using svviz2 with the four datasets in Table 2. Genotypes from Illumina and 10x were ignored in tandem repeats >100bp in length, and genotypes from PacBio were ignored in tandem repeats >10000bp. Genotypes from all datasets were ignored in segmental duplications >10000bp. If the genotypes from all remaining datasets were concordant, and PacBio supported a genotype of heterozygous or homozygous variant, then the variant was included in downstream analyses.*
3. *If two or more supported variants ≥50bp were within 1000bp of each other, they were filtered because they are potentially complex or inaccurate.*

*In addition, benchmark regions were formed with the following process:*

1. *Call variants from 3 PacBio-based and 1 10x-based assemblies*
2. *Compare variants from each assembly to our v0.5.0 PASS calls for HG2 allowing them to be up to 20% different in all 3 distance measures, and only keep variants not matching a v0.5.0 call.*
    a. *Cluster the remaining variants from all assemblies and keep any that are supported by at least one long read assembly*
3. *Find regions for each assembly that are covered by exactly one contig for each haplotype.*
4. *Find the number of assemblies for which both haplotypes cover each region*
5. *Subtract regions around variants remaining after #2, using svwiden's repeat-expanded coordinates, and expanded further to include any overlapping repetitive regions from Tandem Repeat Finder, RepeatMasker SimpleRepeats, and RepeatMasker LowComplexity, plus 50bp on each end.*
6. *High confidence regions are regions in #4 covered by at least 1 assembly minus the regions in #5.*
7. *Further exclude any regions in the Tier 2 bed file of unresolved and clusters of variants, unless the Tier 2 region overlaps a Tier 1 PASS call.*
8. 

**Table 2***:* Genotypes for HG002 were determined using a heuristics based strategy by determining cut-offs for weighted alternate and reference(REF) counts [ALT/(REF+ALT)]. The cut-offs were determined manually from looking at distributions for different size ranges. For each technology, a minimum ALT and REF count was set and genotypes were determined based on the ratio of REF to ALT counts. Variants that did not meet the criteria in this table were not included in the v0.6 comparison.

| Technology | Minimum ALT and REF count [ALT+REF≥n] | ALT and REF Count Ratio (x = ALT/(REF+ALT) ) | Genotype Label |
|---|---|---|---|

| | | | |
|---|---|---|---|
| PacBio | n≥8 | x<0.1 | Homozygous Reference (GT=0/0) |
| | | 0.25<x<0.75 | Heterozygous Variant (GT=0/1) |
| | | x>0.9 | Homozygous Variant (GT=1/1) |
| Illumina 250bp and MP Illumina | n≥8 | x<0.05 | Homozygous Reference (GT=0/0) |
| | | 0.1<x<0.9 | Heterozygous Variant (GT=0/1) |
| | | x>0.95 | Homozygous Variant (GT=1/1) |
| 10x Genomics | n≥5 | x1<0.05 AND x2<0.05 | Homozygous Reference (GT=0/0) |
| | | (x1>0.95 AND x2<0.05) OR (x2>0.95 AND x1<0.05) | Heterozygous Variant (GT=0/1) |
| | | x1>0.95 AND x2>0.95 | Homozygous Variant (GT=1/1) |

**Acknowledgments**

We thank many Genome in a Bottle Consortium Analysis Team members for helpful discussions about design of the SVCurator app and the experiment. We would especially like to thank Nancy Hansen of the National Human Genome Research Institute for advice on implementing SVAnalyzer as well as input on interpreting structural variant data. Certain commercial

equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.

## Author Contributions

L.M.C., M.S., and J.M.Z. designed the study. N.S. created svviz2. J.M.Z. and N.S. designed and performed svviz2 analysis. P.P. developed early iterations of the app. L.M.C. developed the final version of SVCurator and performed analysis of crowdsourced results. N.S., C.S.L., A.C., G.N., C.W., C.P., L.M.C., W.C., N.N., E.D., G.J., D.B., C.B., C.X., S.R.R.K., N.A., P.W., A.A., G.S., S.S., J.M.Z. curated at least 648 SVCurator events. L.M.C. and J.M.Z. wrote the manuscript with the assistance of members of the Genome in a Bottle Community.

## Supplementary Data

- SVCurator Final Labels [60% curator concordance with at least 3 curators agreed on the final label]
- SVCurator Final Labels with top curator statistics
- SVCurator Final Labels [includes labels assigned by each curator]
- List of sequencing technologies and variant callers used to discover the SV calls within SVCurator
- SVCurator App Code

## References

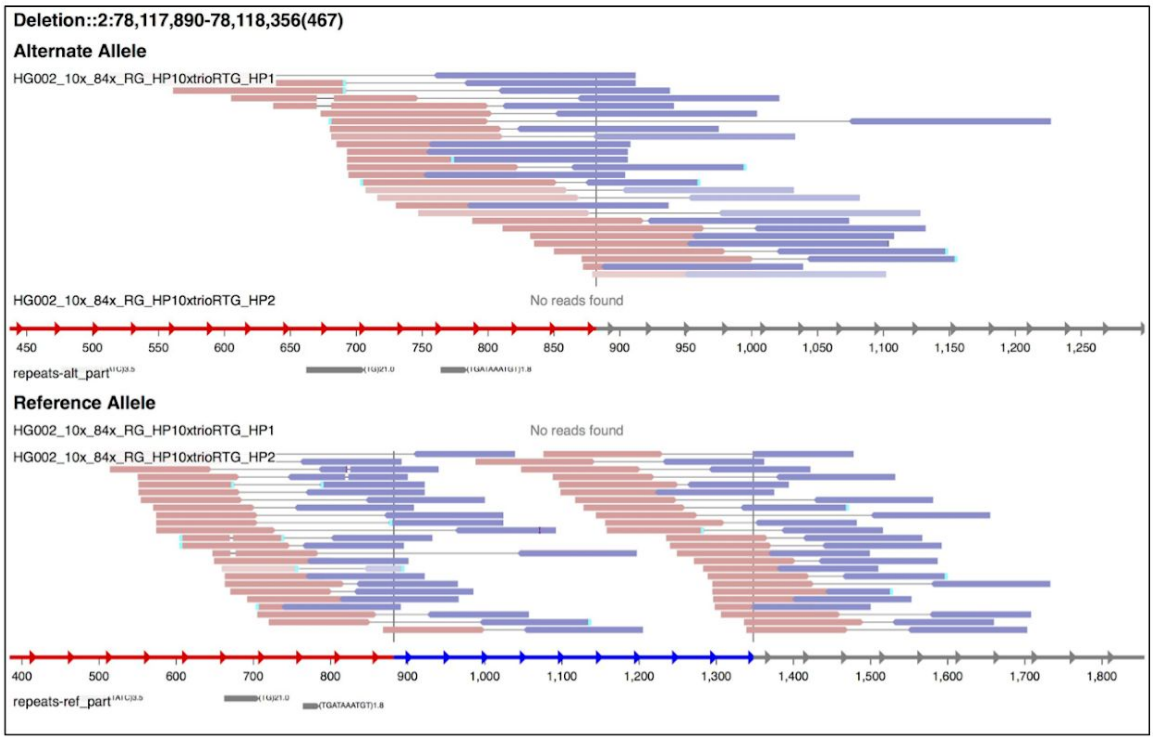
1. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75 (2015).
2. Stankiewicz, P. & Lupski, J. R. Structural Variation in the Human Genome and its Role in Disease. *Annu. Rev. Med.* **61,** 437–455 (2010).
3. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0054-x (2019).
4. Zook, J.M. *et al*. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0074-6 (2019).
5. Parikh, H. *et al*. svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*. **17**,64 (2016).
6. Chaisson, M.J.P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. BioRxiv. https://doi.org/10.1101/193144 (2018).
7. Audano, P.A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176,** 663 (2019).
8. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12,** 363 (2011).
9. Greenside, P. *et al.* CrowdVariant: a crowdsourcing approach to classify copy number variants. *bioRxiv* (2016).
10. Belyeu, J. R. *et al.* SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *Gigascience* **7,** giy064-giy064 (2018)
11. Muzzey, D. *et al.* Software-Assisted Manual Review of Clinical Next-Generation Sequencing Data: An Alternative to Routine Sanger Sequencing Confirmation with

Equivalent Results in >15,000 Germline DNA Screens. *The Journal of Molecular Diagnostics*. (2018)

12. Zook, J. M. *et al*. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Nature Scientific Data*, **3**, 160025 (2016).
13. Spies, N. *et al.* svviz: a read viewer for validating structural variants. *Bioinformatics.* **31,** 24 (2015).
14. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405 (2015).

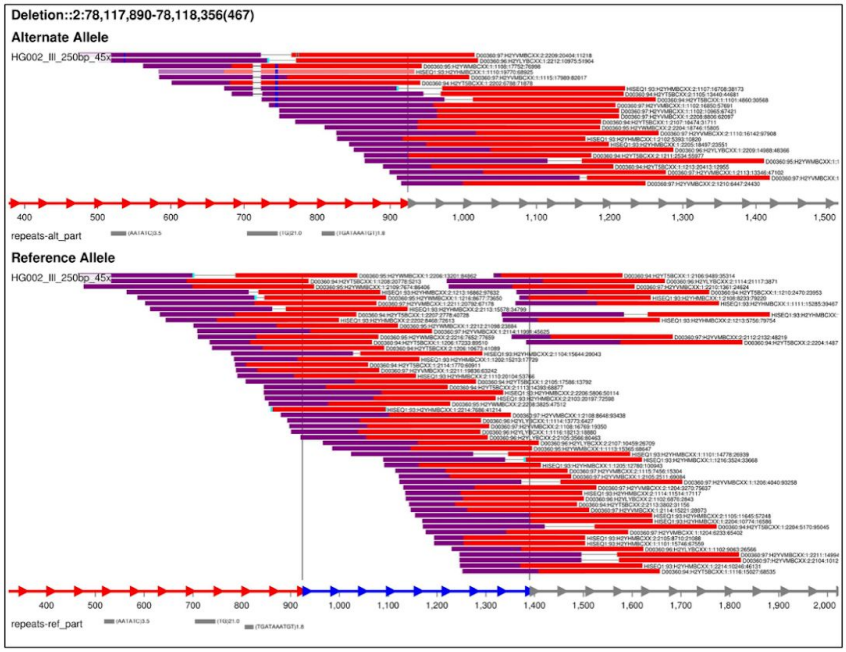
## Supplementary Figure 1-5

A)



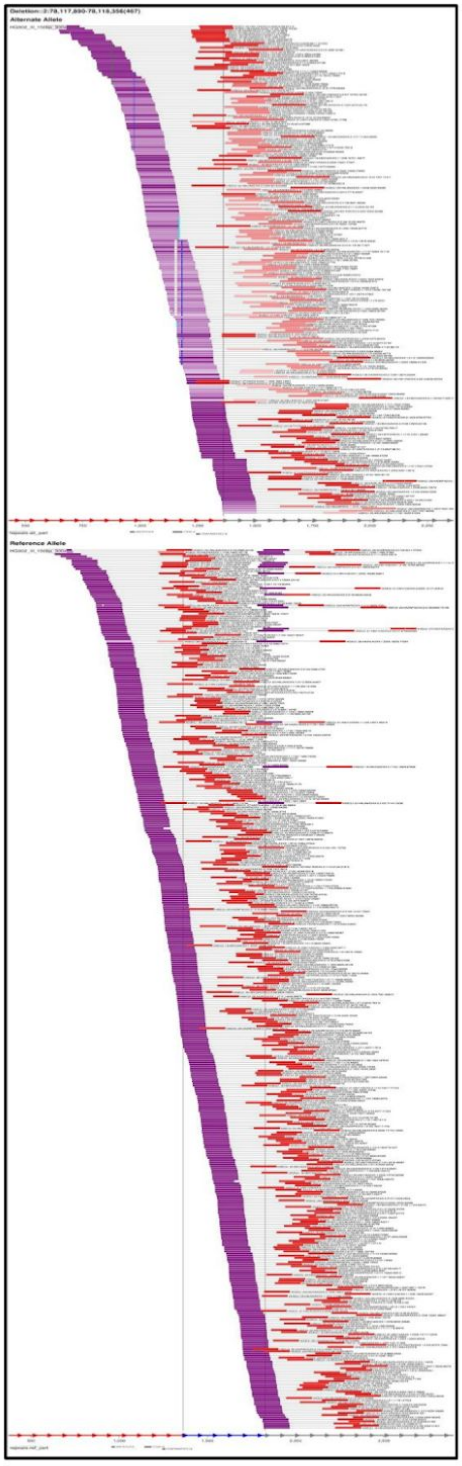

**Supplementary Figure 1**. Images were generated for each event from Integrated Genome Viewer, svviz2. A putative 467bp deletion is shown. Svviz2 generates read aligned images for each short read and long read sequencing technology. A) svviz2 read aligned image - 10x Genomics (read length = 98bp; read depth = 50x). Reads were aligned to reference and alternate allele by svviz2.

B)



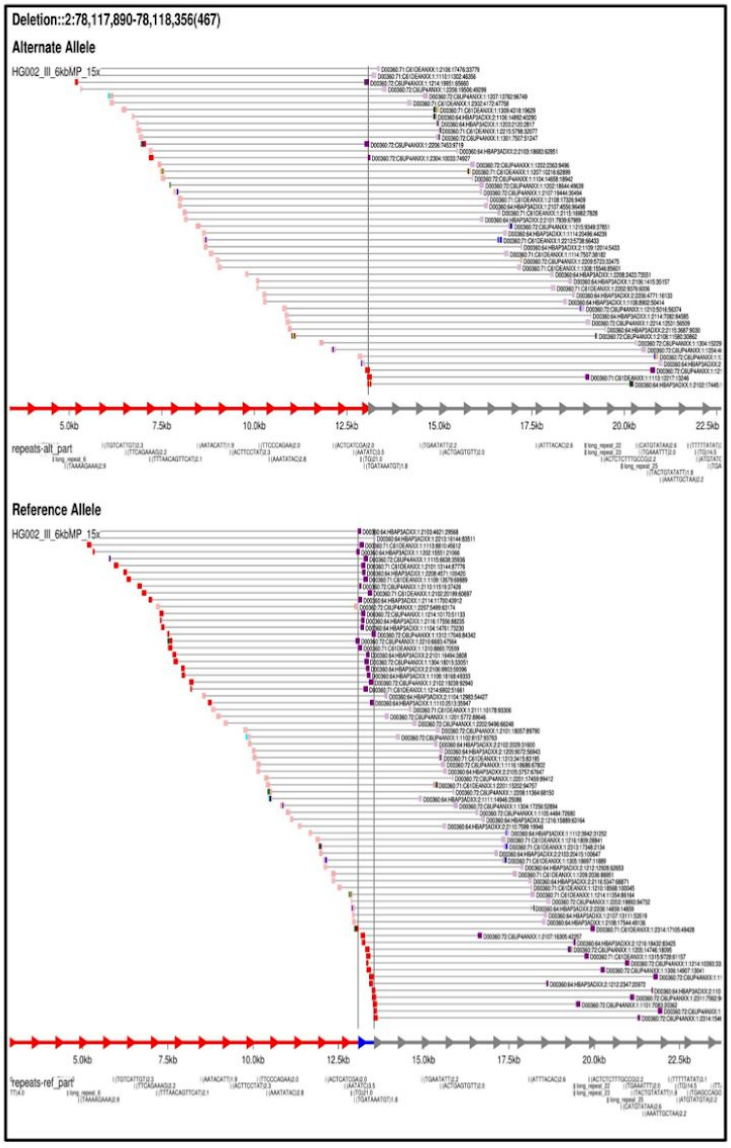

B) svviz2 read aligned image - Illumina HiSeq (read length = 250 bps; read depth = 40-50x)

C)



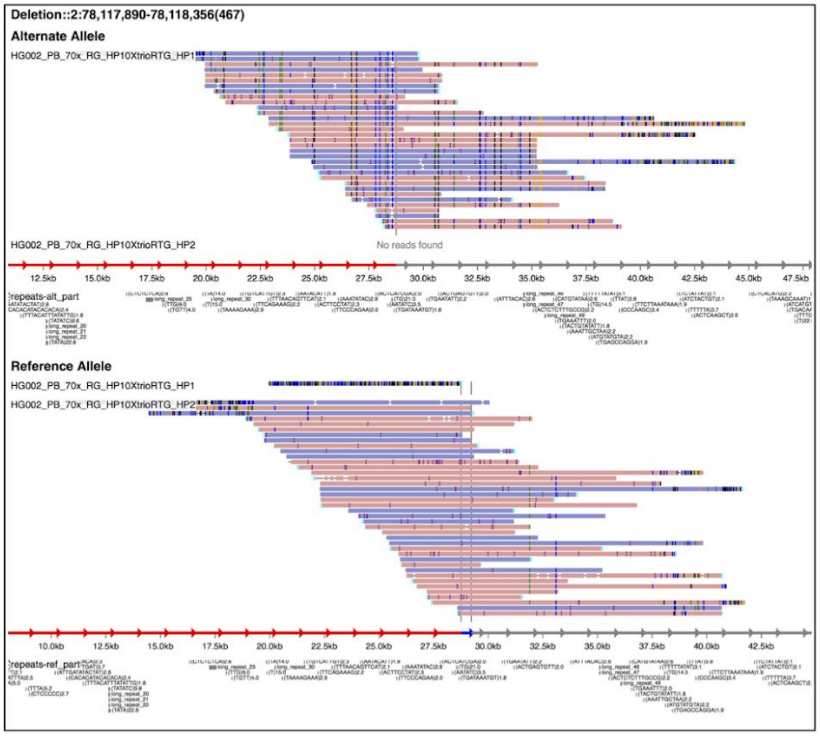

C) svviz2 read aligned image - Illumina HiSeq (read length = 148 bps; read depth = 296.83x)

D)



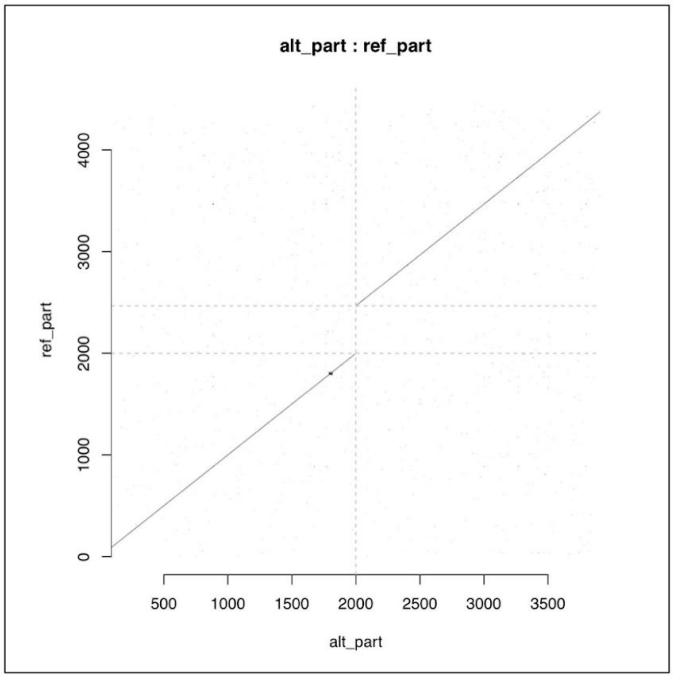

D) svviz2 read aligned image - Illumina Mate Pair (read length = 100 bps; insert size = 6000bp; read depth = 13-14x)

E)



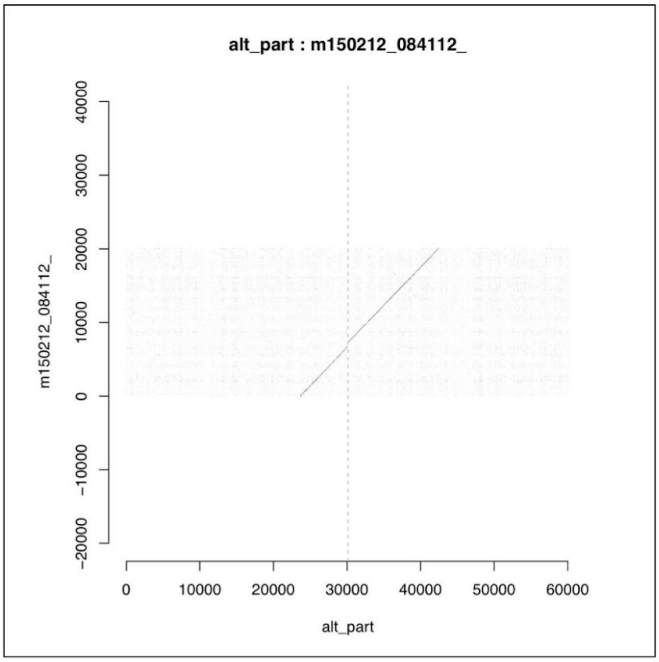

E) svviz2 read aligned image - Haplotype separated PacBio (read length = 10-11kb; read depth = 69x). Reads were haplotype separated using WhatsHap and aligned to reference and alternate allele by svviz2.

F)



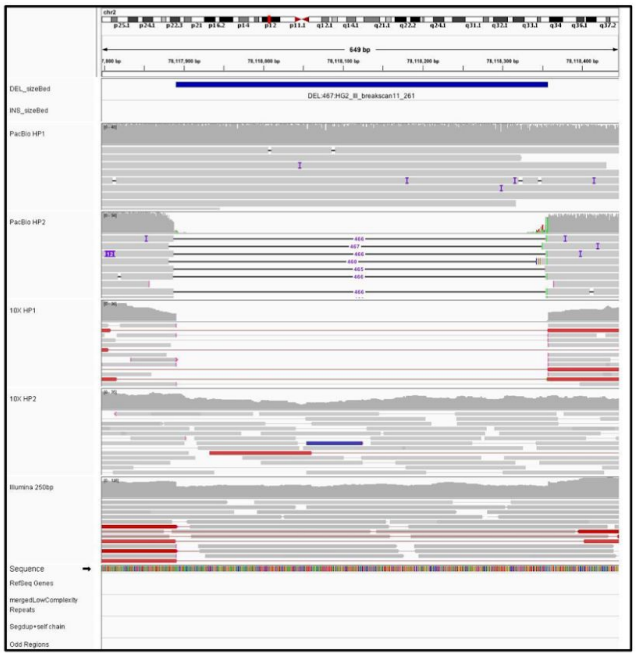

F) Svviz2 dotplot displaying reference versus alternate allele

G)



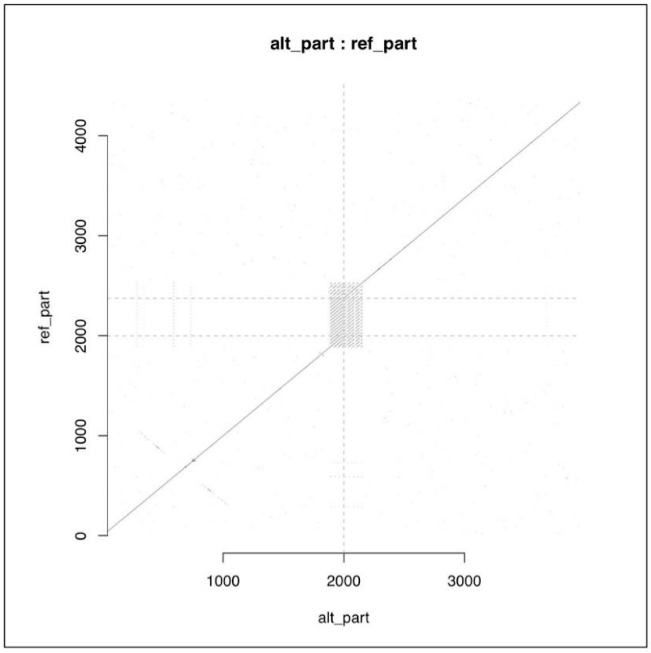

G) Svviz2 dotplot PacBio read with the highest mapping quality score versus the alternate allele.

H)

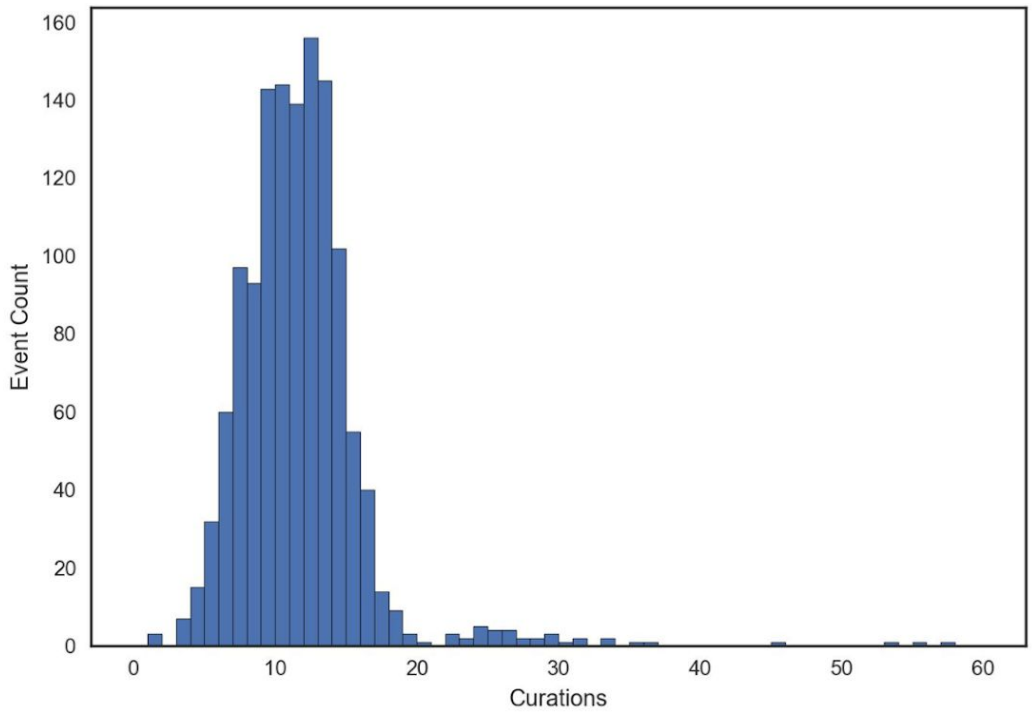



H) IGV image showing reads aligned to a putative variant. IGV tracks include: Haplotype separated PacBio reads, Haplotype Separated 10x Genomics reads, Illumina 250x250bp paired end sequencing, and tracks to describe repeat regions (low complexity repeats and segmental duplications). All reads were aligned to GRCh37 human reference genome.

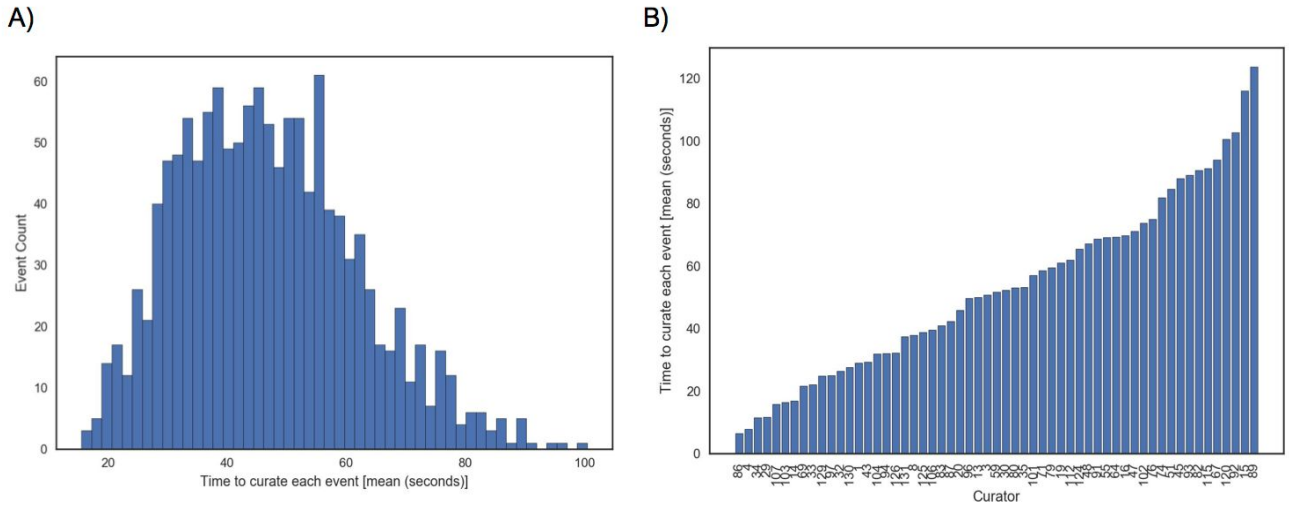
I)



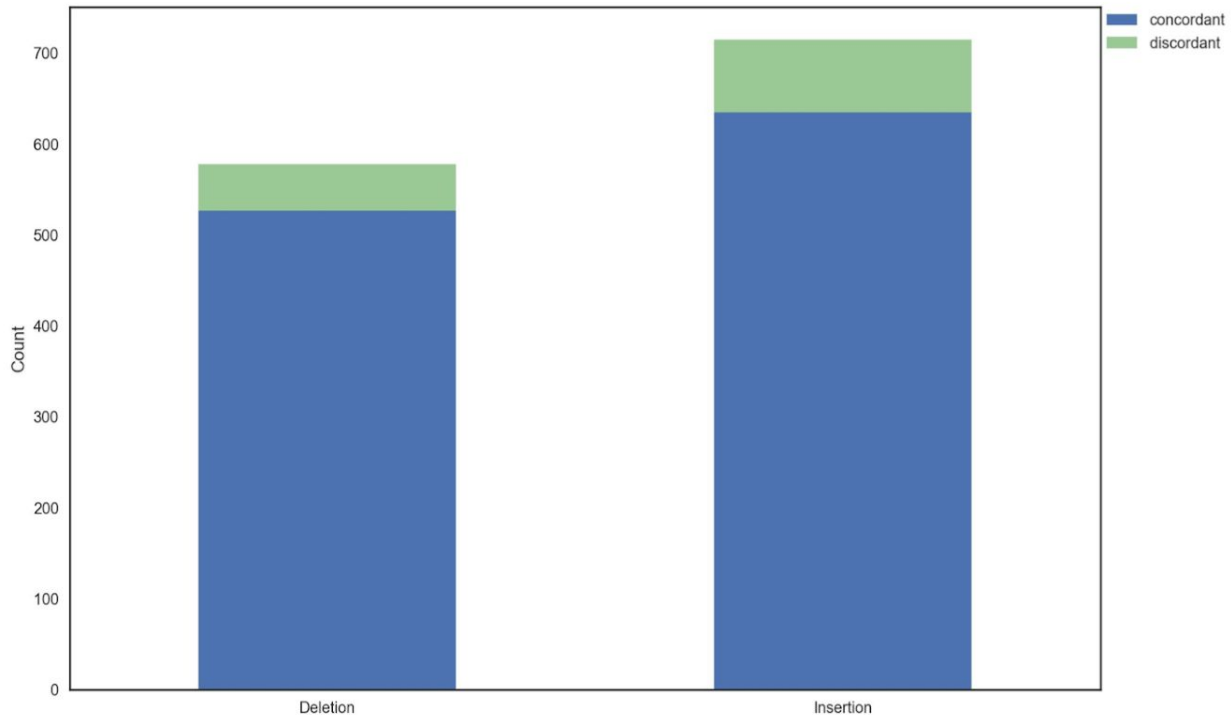

I) Svviz2 dotplot displaying a putative deletion in a tandem repeat.




**Supplementary Figure 2.** Summary of the number of curations for each event. 61 curators evaluated SVCurator events. Each of the 1295 sites were curated on average 11 times with 1290 events curated at least 3 times.

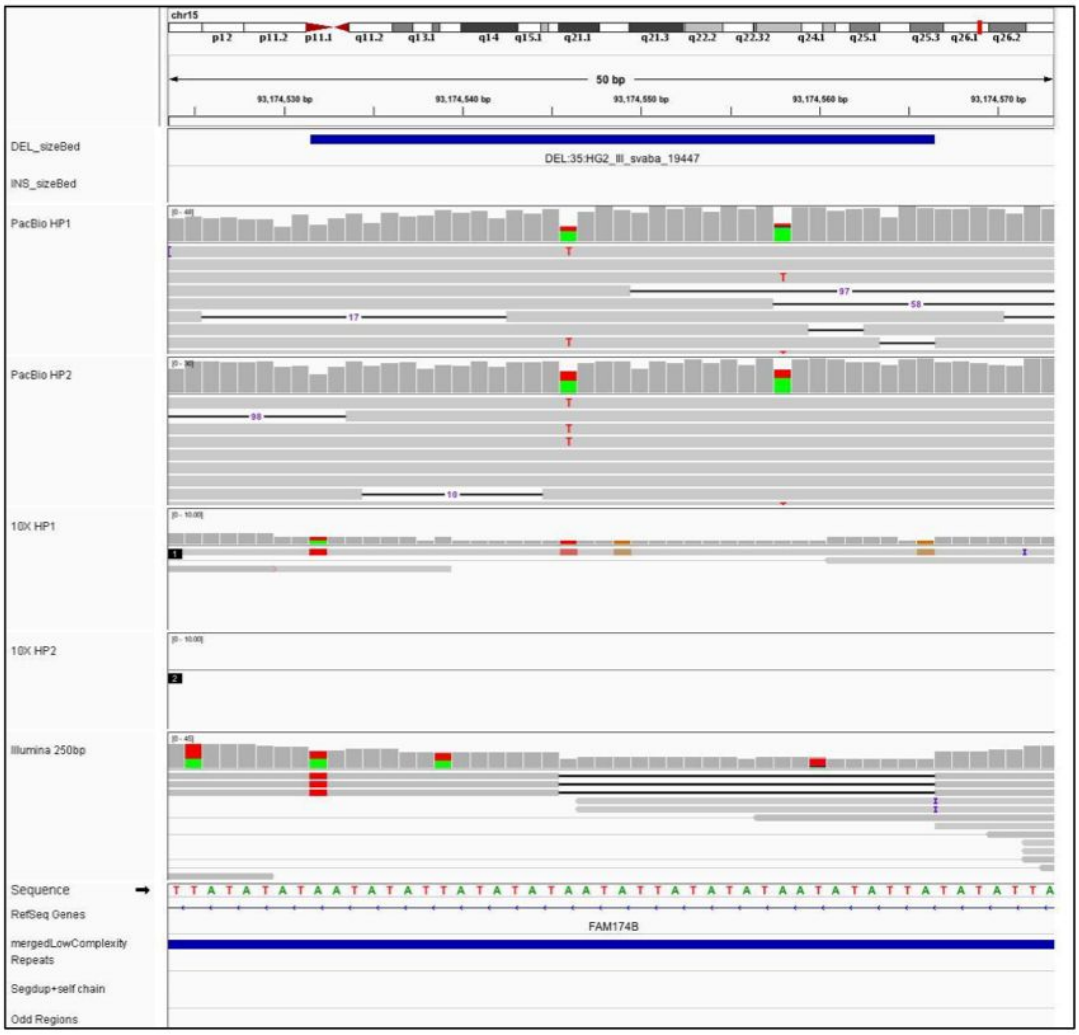



**Supplementary Figure 3.** An evaluation of the time to curate each SVCurator event. A) Overall distribution of the average time to curate events. B) Distribution of the average time to curate each event for each curator.

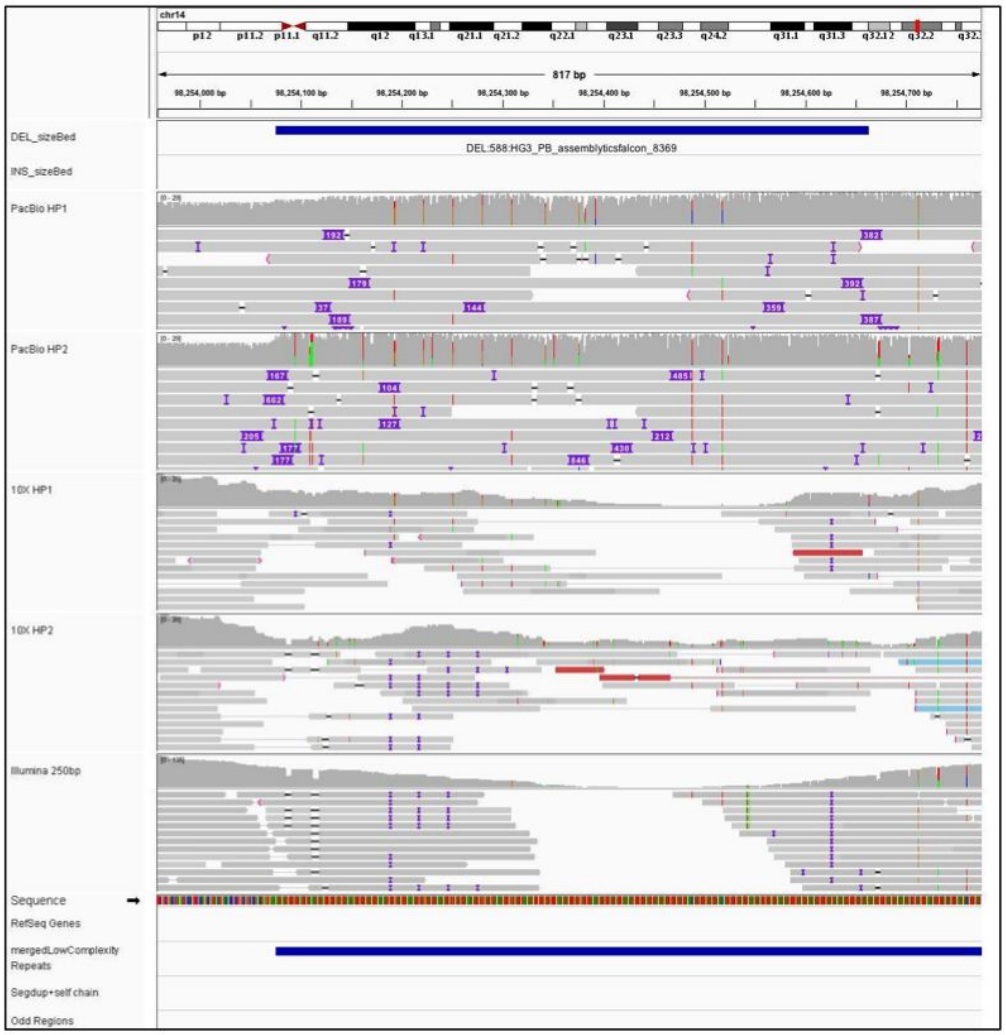



**Supplementary Figure 4.** Evaluation of concordance between Threshold 1 Top Curators (curators that had at least 90.9% concordance with experts) and Threshold 2 Top Curators (curators that had at least 77.7% concordance with experts).

A)



B)




**Supplementary Figure 5**. Examples of events that were discordant between consensus labels assigned by curators and the v0.6 high confidence genotypes discordant sites. IGV images showing examples of two events that had less than 50% concordance for the label assigned by the curators. A) small SV call. B) Large SV call.

**Supplementary Table 1**

| Expert Curator | Percent Concordance |
|---|---|
| 1 | 1.0000 |
| 2 | 0.9777 |
| 3 | 0.9743 |
| 4 | 0.9333 |
| 5 | 0.9318 |
| 6 | 0.9090 |
| 7 | 0.7777 |

**Supplementary Table 1**. Concordance scores amongst ‘expert’ curators.