CrowdVariant: a crowdsourcing approach to classify copy number variants

Peyton Greenside

Biomedical Informatics, Stanford University,
Stanford, CA 94305, USA
pgreens@stanford.edu

Justin Zook and Marc Salit

National Institute of Standards and Technologies (NIST),
Material Measurement Laboratory, 100 Bureau Dr., Gaithersburg, MD 20899 USA
justin.zook@nist.gov, salit@nist.gov

Madeleine Cule

Verily Life Sciences, Calico,
269 E Grand Ave, South San Francisco, CA 94080 USA
cule@calicolabs.com

Ryan Poplin, Mark DePristo*

Google Brain, Verily Life Sciences,
1600 Amphitheatre Parkway, Mountain View, CA USA
rpoplin@google.com, mdepristo@google.com
**Corresponding Author*

Copy number variants (CNVs) are an important type of genetic variation that play a causal role in many diseases. The ability to identify high quality CNVs is of substantial clinical relevance. However, CNVs are notoriously difficult to identify accurately from array-based methods and next-generation sequencing (NGS) data, particularly for small ($<$ 10kbp) CNVs. Manual curation by experts widely remains the gold standard but cannot scale with the pace of sequencing, particularly in fast-growing clinical applications. We present the first proof-of-principle study demonstrating high throughput manual curation of putative CNVs by non-experts. We developed a crowdsourcing framework, called CrowdVariant, that leverages Google's high-throughput crowdsourcing platform to create a high confidence set of deletions for NA24385 (NIST HG002/RM 8391), an Ashkenazim reference sample developed in partnership with the Genome In A Bottle (GIAB) Consortium. We show that non-experts tend to agree both with each other and with experts on putative CNVs. We show that crowdsourced non-expert classifications can be used to accurately assign copy number status to putative CNV calls and identify 1,781 high confidence deletions in a reference sample. Multiple lines of evidence suggest these calls are a substantial improvement over existing CNV callsets and can also be useful in benchmarking and improving CNV calling algorithms. Our crowdsourcing methodology takes the first step toward showing the clinical potential for manual curation of CNVs at scale and can further guide other crowdsourcing genomics applications.

Keywords: copy number variation, precision medicine, crowdsourcing

1. Introduction

Copy number variation is a type of structural variation that involves large-scale duplications or deletions of parts of a chromosome. Copy number variants can have substantial effects on cell and organism phenotype and are associated with many kinds of human disease (Redon et al., 2006) (Feuk, Carson, & Scherer, 2006) (Sudmant et al., 2015). Identifying CNVs is an important component of clinical pipelines for assessing genetic mutations that contribute to disease progression. Numerous algorithms have been developed to characterize these variants from genotyping arrays and next-generation sequencing data (English et al., 2015) (Tattini, D'Aurizio, & Magi, 2015) (Mills et al., 2011) (Kidd et al., 2008). However, these algorithms often have poor concordance on both the location and the type of copy number variant, particularly for small-scale (< 10kbp) CNVs (Scherer et al., 2007) (Pinto et al., 2011), leading experts to rely heavily on manual curation. One key challenge in further developing and assessing these algorithms is the lack of a large set of "gold standard" or reference copy number variants.

Crowdsourcing has been used successfully to obtain gold standard labels in projects such as Galaxy Zoo (Raddick et al., 2010), ClickWorkers (Ishikawa, ST and Gulick, 2012), FoldIt (Cooper et al., 2010), and Zooniverse (Prather et al., 2013), but little investigation has been done to understand how crowdsourcing can be best utilized to analyze genomic variation (Haghighi et al., 2018). Basic questions include whether or not any domain expertise is truly needed, how large the crowd should be, and how to best train and display genetic variation to workers. We investigated the use of crowdsourcing platforms to classify copy number variants, focusing on deletions, and to address these basic questions. Google has developed the Crowd Compute platform to facilitate large-scale crowdsourcing problems, and we developed our framework with this platform to enable high throughput classifications. In this work we show proof of principle in a well characterized reference genome, an essential first step before deploying the method on more variable genomes such as from clinical samples. In a similar vein, we focus on deletions as the most frequent and also likely easiest to classify type of structural variation before focusing on more complex applications. CrowdVariant can be used to develop high confidence CNV sets, to benchmark new CNV detection algorithms, and to enable high throughput manual curation of CNVs using both experts and non-experts.

2. Results

2.1. *The CrowdVariant Framework*

The CrowdVariant framework uses a crowdsourcing platform to display putative copy number variant sites to workers and aggregates classifications from a pool of workers to determine the copy number state. Using this framework, we first ran an experiment to compare non-expert and expert classifications on a pilot set of putative CNV sites and then expanded our classi-

fications to curate a genome-wide set of high confidence CNVs [Figure 1].

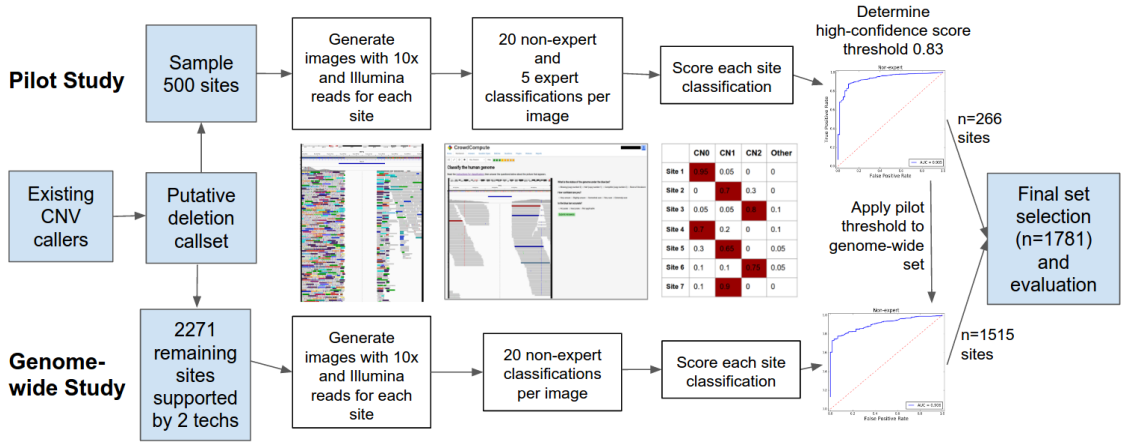

Fig. 1. The experimental design was constructed to first evaluate a pilot set of sites with both experts and non-experts before applying the same framework to a genome-wide set of sites using non-experts only.

CrowdVariant displays pileup images of putative copy number variant sites using the Integrative Genomics Viewer (IGV), showing all reads aligned to the site and the flanking regions [Supplementary Figure 1] (Thorvaldsdóttir, Robinson, & Mesirov, 2013). Workers classify the site, assess break point accuracy and report their confidence based on seeing one image at a time.

We selected a set of 500 putative deletion sites for the pilot phase of our study. We first called putative sites using an ensemble approach from multiple sequencing technologies (Illumina, PacBio, Complete Genomics and BioNano) and corresponding algorithms (see Supplementary Methods for details) (Abyzov, Urban, Snyder, & Gerstein, 2011) (Garrison & Marth, 2012) (Mohiyuddin et al., 2015) (Hormozdiari, Hajirasouliha, McPherson, Eichler, & Sahinalp, 2011) (Iqbal, Caccamo, Turner, Flicek, & McVean, 2012) (Mak et al., 2016) (Chaisson et al., 2014) (Nattestad & Schatz, 2016) (Drmanac et al., 2010). We then randomly selected from all putative sites 500 pilot sites ranging from 100bp to 3000bp with varying levels of support from existing algorithms [Supplementary Table 1].

We used aligned 10X Genomics (10X) and Illumina paired-end (Illumina) reads from the reference Ashkenazim trio made available by the Genome In A Bottle (GIAB) Consortium (Zook et al., 2016). For each putative copy number variant site, we generated an image for each member of the trio (son/mother/father) using Illumina reads, one image for the son's diploid reads and one image for each haplotype of the son's reads using 10X reads. Although workers potentially saw multiple images of the same site, we did not disclose to workers the experimental design, the sequencing technology, the individual or the site being shown in an

effort to most fairly compare experts and non-experts.

In our pilot study, 20 non-experts each classified all 6 images for the 500 pilot sites. We launched a global recruitment for self-reported experts curators with over 110 individuals from several dozen institutions signing up to classify variants. The participation rate was highly variable with an average of 76 questions per expert [Supplementary Figure 2]. We ensured that all 6 images for at least 100 sites were classified by 5 experts each.

2.2. *Non-experts can curate high quality copy number variants*

Both experts and non-experts agreed on a consensus classification for the majority of sites [Supplementary Figure 3]. We visualized the responses for non-experts [Figure 2] and experts [Figure 3] by weighting each copy number classification and clustering workers and sites to reveal performance differences across sequencing platforms and individuals. We kept the identity of each non-expert worker separate, but we merged the expert answers into artificial workers 1 through 5 as experts did not answer enough questions individually to be meaningfully compared. For 86% of images, at least 70% of non-expert workers agreed on the classification, showing that non-experts can be trained to interpret copy number variants in a consistent manner [Supplementary Table 2]. Non-experts primarily had difficulty classifying haplotype images and systematically confused CN2s as CN1s for haplotype images only (see Fig. 8 haplotype heatmaps). Beyond these systematic errors, there were several non-experts that deviated from the majority either from lack of effort or understanding. Improving the documentation by showing more than 2 examples of each copy number type could further improve non-expert performance.

Agreement among workers was used to assign a final classification and confidence score to each putative site. We defined the CrowdVariant score as the proportion of workers that voted in favor of the most popular classification (CN0/CN1/CN2/None of the Above), with higher scores reflecting more confident classifications. We incorporated worker classifications for all images of the same site, but classified each site for each individual in the trio independently. We counted all diploid classifications but only those haploid classifications where the pair of haplotype images was consistent with a diploid classification [Supplementary Methods]. We assign the most likely copy number state to each site by selecting the classification with the largest proportion of votes.

Non-experts performed similarly to experts when comparing the rate of Mendelian violations among the trio (classifications that would not plausible from Mendelian inheritance) for each site [Supplementary Methods] [Table 1]. We found that 89% and 90% of all sites were classified without a Mendelian violation for experts and non-experts, respectively. The sites with Mendelian violations had lower scores and could largely be filtered out of the high quality set. The CrowdVariant scores discriminated Mendelian violations from genetically plausible trio classifications with an AUC of 0.89 for non-experts and an AUC of 0.87 for experts [Supplementary Methods] [Supplementary Figure 4]. For comparison, we randomized all answers

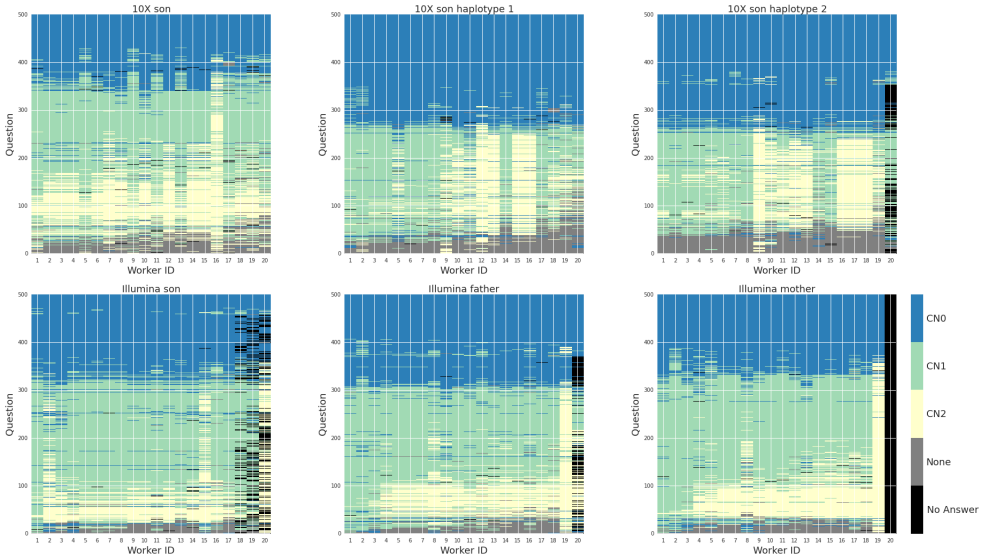

Fig. 2. Non-expert classifications for 500 sites were color coded, weighted and clustered (see Supp. Methods for details). Rows represent a question (i.e. an image of a putative site using a particular sequencing technology) and columns represent workers. Clockwise from top left: 10X son, 10X son haplotype 1 only, 10X son haplotype 2 only, Illumina mother, Illumina father, Illumina son.

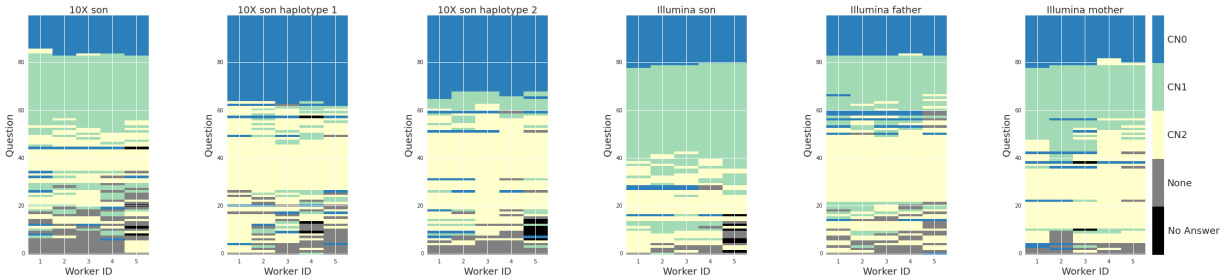

Fig. 3. Expert classifications for 100 sites were color coded, weighted and clustered (see Supp. Methods for details). Rows represent a question (i.e. an image of a site using a particular sequencing technology) and columns represent workers. Left to right: 10X son, 10X son haplotype 1 only, 10X son haplotype 2 only, Illumina son, Illumina father, Illumina mother.

by re-sampling the entire worker by classification matrices for experts and non-experts and re-computed the rate of Mendelian violations [Supplementary Table 3]. The AUCs for expert and non-expert randomized answers were 0.47 and 0.50, respectively, and both 95% confidence intervals overlapped a random AUC of 0.5.

We curated a high confidence set of CNVs for the son (NA24385) with high probability of correctness and no Mendelian violations [Supplementary Materials]. We initially intended to use self-reported confidence to filter lower quality classifications, but most non-experts consistently reported medium to high confidence despite minimal training [Supplementary Figure 5]. To avoid relying on self-reported confidence, we ranked all 500 sites by their CrowdVariant score and selected all sites with a higher score than the site with the first Mendelian violation.

Metric \ Data Set	Expert	Non-expert
Percent of sites without violation	89/100 (89%)	448/500 (90%)
ROC AUC	0.87	0.89
ROC AUC 95% confidence interval	[0.79, 0.95]	[0.86, 0.92]
Average violation probability	0.15	0.14

This violation occurred at score 0.83 and resulted in discarding approximately half of the sites for a total of 266 high confidence sites. The high confidence set of sites contains 122 CN0, 138 CN1, 5 CN2 and 1 "None of the above" classification. 252 out of 266 are supported by at least two other technologies. Importantly, for all sites in the high quality set that were classified by both experts and non-experts, there was 100% agreement (n=56 sites) between experts and non-experts.

2.3. *CrowdVariant can classify CNVs with variable support or unclear breakpoints*

CrowdVariant agrees with consensus classifications from existing algorithms, while also classifying variants that are challenging for existing algorithms. CrowdVariant scores assigned to each site are correlated with the number of technologies underlying the original calls [Figure 4]. CrowdVariant classifications also show strong agreement with svviz (Spies, Zook, Salit, & Sidow, 2015), a semi-automated visualization tool that determines whether each read supports the reference allele, alternate allele, or is ambiguous. We used a preliminary heuristic method to classify copy number variants based on the read counts supporting the reference and alternate alleles as determined by svviz for each dataset, and required agreement across all datasets that had clear support for a genotype [Supplementary Methods]. When comparing all high confidence classifications, agreement with svviz was 82%. CrowdVariant was able to resolve 26 sites that were uncertain for svviz, explaining part of the discrepancy. When we removed sites that were classified as "None of the Above" in CrowdVariant or uncertain in svviz, agreement was 91% between the two methods. Agreement with svviz also increased with the number of supporting technologies [Figure 5].

The true power of incorporating many data types is clear when all 6 images of the same site are viewed together [Figure 6]. We find in multiple cases the crowd is able to resolve copy number state where other methods cannot, particularly when the boundary points are incorrect or ambiguous [Figure 7, Supplementary Figure 7]. While non-experts make some mistakes, we find that they do so in a consistent manner, such as mistaking a difficult-to-sequence region for a deletion, and they could likely be trained to recognize other features in the image that would clarify these mistakes. Phased data is particularly powerful for classifying heterozygous CNVs that are otherwise ambiguous and provides visual confirmation of the CrowdVariant

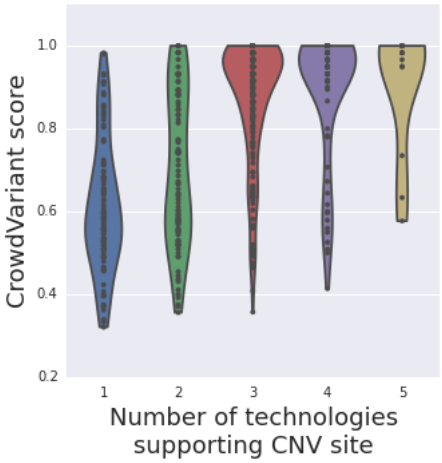

Fig. 4. CrowdVariant scores determined by non-expert workers stratified by the number of supporting technologies from existing CNV callers.

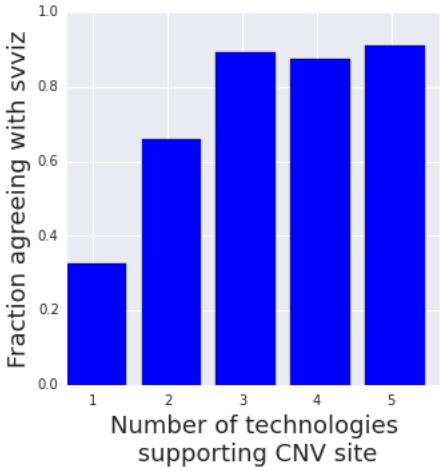

Fig. 5. Agreement (within each bin) with svviz classifications for sites with varying support from orthogonal technologies. We only compare sites with CN0, CN1 or CN2 classifications from both methods.

results in conjunction with all other images for the site.

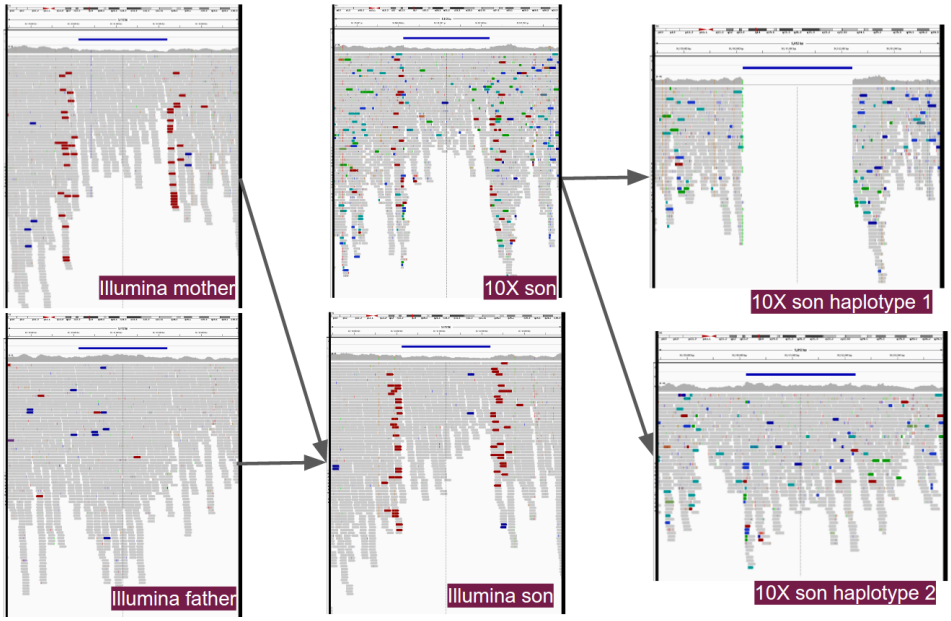

Fig. 6. Viewing all image types together shows the power of combining familial and phasing information in different sequencing platforms. This variant (chr15:36160125-36162210) was classified as copy number 1 in the son with CrowdVariant score 1.0 and is part of the high quality set. The variant is visible in the mother, both diploid son images and one of the haplotype images. Clockwise from top left: Illumina mother, 10X son, 10X son haplotype 1, 10X son haplotype 2, Illumina son, Illumina father.

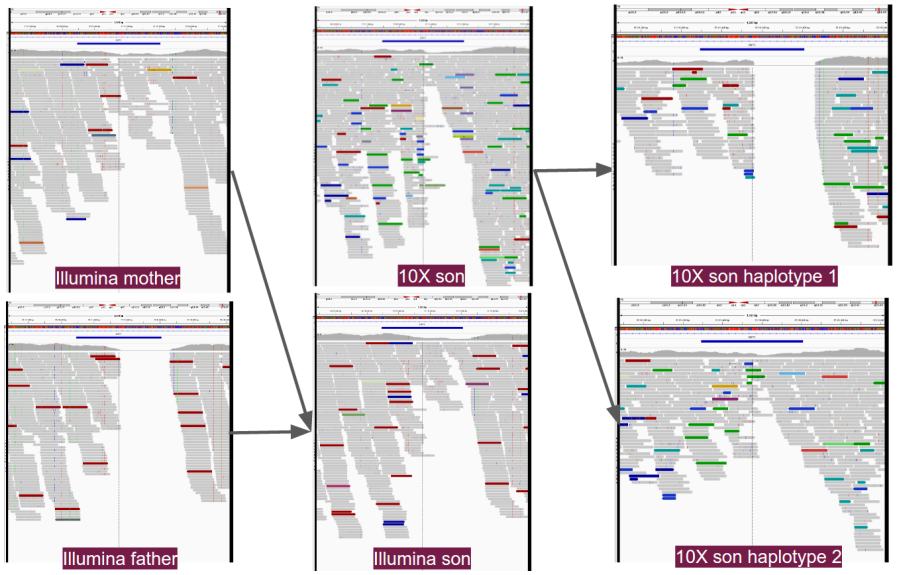

Fig. 7. Viewing all image types together shows the power of combining familial and phasing information in different sequencing platforms. This variant (chr19:57111292-57111809) was classified as CN1 in the son with score 0.89 and is part of the high quality set. Svviz classified this example as CN2 due to the imprecise breakpoints. Clockwise from top left: Illumina mother, 10X son, 10X son haplotype 1, 10X son haplotype 2, Illumina son, Illumina father. Mother appears to share CNV with the son, while the father is wildtype. Visualizations produced by default IGV settings.

2.4. *CrowdVariant can be used to curate a genome-wide high quality set of copy number variants*

Having demonstrated that we can use non-expert workers to curate a high quality set of copy number variants, we expanded our classifications genome-wide. We took all putative CNV sites that were supported by GIAB callsets from at least 2 technologies and had not been classified in the pilot set (n=2271) and recruited 20 non-expert classifications for each site for all 6 image types. Due to the larger volume of images, not every worker classified all images in the genome-wide set. Consistent with the pilot study, we observed strong agreement among non-expert workers in the genome-wide set. Again, the primary inconsistencies were classifications for the haplotype images [Figure 8].

We scored each site by the proportion of workers voting for each classification and applied the threshold determined by the first 500 sites to curate high quality genome-wide classifications. This resulted in 1,515 new high confidence sites for the son (NA24385). The CrowdVariant scores for these sites correlate with the number of supporting technologies [Figure 9]. Likely due to requiring 2 supporting technologies, these sites were in even stronger agreement with svviz with 97.2% agreement among sites given CN0/CN1/CN2 classifications with both methods [Figure 10]. The high quality genome-wide set includes calls for 93 sites that svviz found uncertain. The additional genome-wide set includes 959 CN1, 552 CN0, 3 CN2 and 1 None of the Above. The CrowdVariant scores for the genome-wide set of CNVs also demonstrate similar concordance with orthogonal technologies [Figure 9] and classify Mendelian violations in

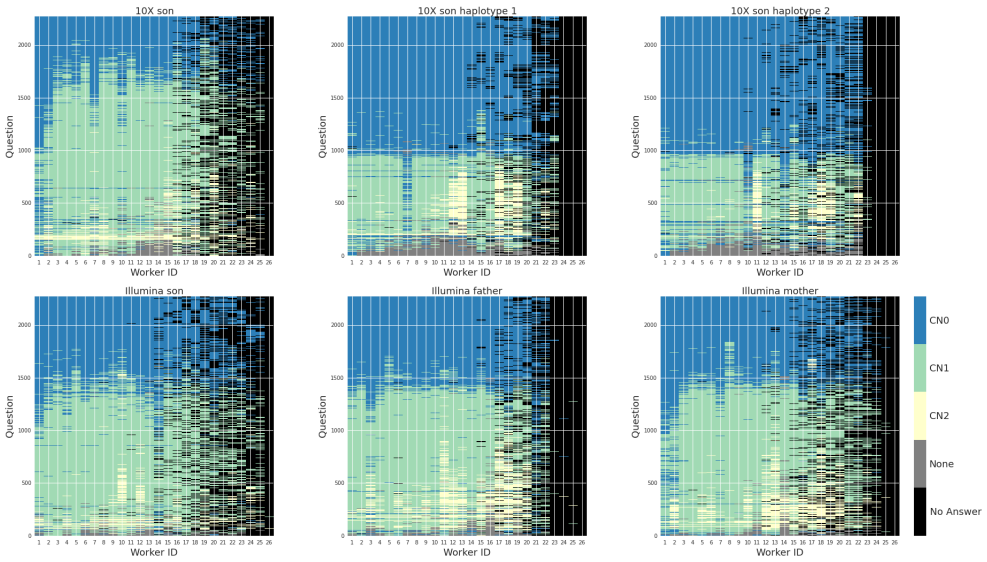

Fig. 8. Non-expert classifications for genome-wide sites in Phase 3 were color coded, weighted and clustered. Rows represent a question (i.e. an image of a particular site using a particular sequencing technology) and columns represent workers. Clockwise from top left: 10X son, 10X son haplotype 1 only, 10X son haplotype 2 only, Illumina mother, Illumina father, Illumina son.

the trio with auROC 0.94 [Supplementary Figure 8]. Above the threshold for high confidence determined from the pilot study, there was only one Mendelian violation in the genome-wide set occurring at a score of 0.94 [Supplementary Figure 9]. Combining with the 266 high quality sites from the pilot set, we finalized a set of 1,781 high confidence CNVs.

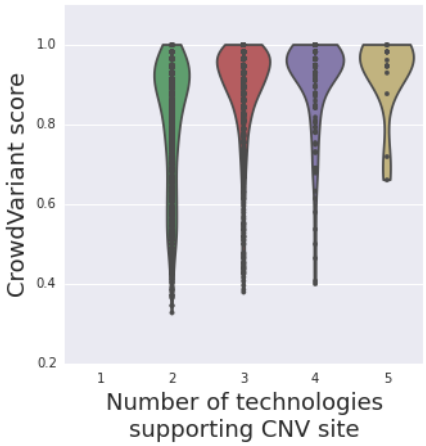

Fig. 9. CrowdVariant scores for all genome-wide sites determined by non-expert workers stratify by the number of supporting technologies from existing CNV callers.

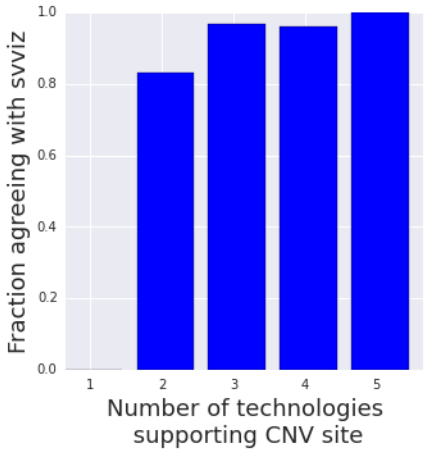

Fig. 10. Agreement with svviz classifications for genome-wide sites with varying support from orthogonal technologies. We only compare sites with CN0, CN1 or CN2 classifications from both methods.

3. Discussion

We show that individuals with no background in genomics can be trained to accurately classify and thereby curate copy number variants. This is possible because the classification of CNVs based on images of aligned NGS reads is ultimately a pattern recognition problem, and even non-experts with limited training can excel at recognizing these patterns. As soliciting expert participation is prohibitively more difficult than non-expert participation (evident in the small amount of expert data we were able to collect), the ability to use non-experts enables crowdsourcing on a substantially larger scale. Deployment of manual curation on the ever growing body of clinical samples would likely require this adaptation as the volume will quickly exceed the capacity of experts. In this study, the larger scale afforded by non-expert workers allowed us to curate thousands of putative CNVs across the entire genome of a single individual from the Genome In A Bottle reference collection.

We are able to use non-expert classifications by using confidence scores to recognize the limit of their abilities. For many applications, such as deriving gold standard labels to improve machine learning methods, it is more critical to determine which classifications are trusted than to classify everything correctly. As machine learning approaches are increasingly adopted to solve genomic problems, crowdsourcing can provide an avenue to derive trusted training sets at high throughput for low cost.

While we have shown that crowdsourcing can be used to generate high confidence labels for CNVs, there are several limitations to our study. First, the set of CNVs we present is not a complete set for the GIAB Ashkenazim son (NA24385), but instead a set of the highest confidence sites. Further, we only know that a CNV is segregating at the site, but we do not know its exact position or size. One broader limitation of crowdsourcing is that people can be consistent but wrong, however this limitation is shared by other approaches such as ensemble-based computational methods. In the current framework, our high confidence classifications are also enriched for sites that are overall easier to classify. However, there are many ways to increase confidence for more difficult questions by scaling the number of workers, augmenting training schemes, improving confidence metrics or considering alternative experimental designs such as those that incorporate both experts and non-experts depending on the particular question's difficulty. Nevertheless, we are confident that our crowdsourced, genome-wide set of curated CNVs will prove valuable to methods developers working to improve CNV calling algorithms.

Many possibilities exist for improving and expanding on this proof-of-concept study demonstrating the crowdsourcing curation of genomic variants. Incorporating images from additional technologies, such as long-read sequencing, could likely identify additional high confidence sites and remove some errors from using only short reads. Additional work might also use input from users about the precision of breakpoints. Other types of images could also be used, such as dot plots from assembly-assembly alignments and svviz images with reads mapped to reference and alternate alleles. These additional methods may help non-experts classify more difficult types of structural variants, like complex changes, insertions, inversions, and translo-

cations, as well as variants in difficult, repetitive regions of the genome.

We use Google's high throughput crowdsourcing platform, but as additional crowdsourcing platforms become available at low cost, soliciting participation from the crowd will become progressively easier. By using strategic experimental design, crowdsourcing can be a productive avenue to compete with and improve upon computational methods in difficult areas of genomics. Copy number variation, a domain where many experts still use manual inspection, is just one of these many areas. We provide a resource of high quality copy number variant classifications for a reference genome as a result of our study but ultimately see the potential expand far beyond these results.

Data Access

All Supplementary Methods, Figures and Data are available at `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/technical/CrowdVariant_SupplementaryInfo/`. We provide the scores for each putative copy number variant site and label the high quality sites. All raw worker answers for both non-experts and experts are available as well.

References

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011, jun). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, *21*(6), 974–984.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., ... Eichler, E. E. (2014, nov). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., ... Players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988, sep). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–45.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., ... Reid, C. A. (2010, jan). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (New York, N.Y.)*, *327*(5961), 78–81.

English, A. C., Salerno, W. J., Hampton, O. A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D. I., ... Gibbs, R. A. (2015). Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC genomics*, *16*(1), 286.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, *7*(2), 85–97.

Garrison, E., & Marth, G. (2012, jul). Haplotype-based variant detection from short-read sequencing.

Haghighi, A., Krier, J. B., Toth-Petroczy, A., Cassa, C. A., Frank, N. Y., Carmichael, N., ... Womens Hospital Project, B. G. M. (2018, aug 13). An integrated clinical program and crowdsourcing strategy for genomic sequencing and mendelian disease gene discovery. *NPJ Genom Med*, *3*, 21. Retrieved 2018-09-30, from `http://www.nature.com/articles/s41525-018-0060-9` doi: 10.1038/s41525-018-0060-9

Hormozdiari, F., Hajirasouliha, I., McPherson, A., Eichler, E. E., & Sahinalp, S. C. (2011, dec). Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome research*, *21*(12), 2203–2212.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012, jan). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, *44*(2), 226–232.

Ishikawa, ST and Gulick, V. (2012). Clickworkers interactive: towards a robust crowdsourcing tool for collecting scientific data. In *Lunar and planetary science conference* (pp. 2–3).

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., ... Eichler, E. E. (2008, may). Mapping and sequencing of structural variation from eight human genomes. *Nature*, *453*(7191), 56–64.

Mak, A. C. Y., Lai, Y. Y. Y., Lam, E. T., Kwok, T.-P., Leung, A. K. Y., Poon, A., ... Kwok, P.-Y. (2016). Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. *Genetics*, *202*(1).

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., ... Korbel, J. O. (2011, feb). Mapping copy number variation by population-scale genome sequencing. *Nature*, *470*(7332), 59–65.

Mohiyuddin, M., Mu, J. C., Li, J., Bani Asadi, N., Gerstein, M. B., Abyzov, A., ... Lam, H. Y. K. (2015, aug). MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics (Oxford, England)*, *31*(16), 2741–2744.

Nattestad, M., & Schatz, M. C. (2016, oct). Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics (Oxford, England)*, *32*(19), 3021–3023.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., ... Feuk, L. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature biotechnology*, *29*(6), 512–20.

Prather, E. E., Cormier, S., Wallace, C. S., Lintott, C., Jordan Raddick, M., & Smith, A. (2013). Measuring the conceptual understandings of citizen scientists participating in zooniverse projects: A first approach. *Astronomy Education Review*, *12*(1).

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., ... Vandenberg, J. (2010). Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review*, *9*(1), 010103.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., ... Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–54.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011, jan). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 77.

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., ... Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature genetics*, *39*(7 Suppl), S7–S15.

Spies, N., Zook, J. M., Salit, M., & Sidow, A. (2015, jul). Svviz: A read viewer for validating structural variants. *Bioinformatics*, *31*(24), 3994–3996.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., ... Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81.

Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in bioengineering and biotechnology*, *3*(June), 92.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013, mar). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192.

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., ... Salit, M. (2016, jun). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, *3*, 160025.