bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SV^2: Accurate Structural Variation Genotyping and *De Novo* Mutation Detection from Whole Genomes

Danny Antaki[1,2,3,4], William M Brandler[1,2,3], & Jonathan Sebat[1,2,3]

[1]Beyster Center for Genomics of Psychiatric Diseases, University of California San Diego, La Jolla, CA 92093 [2]Department of Psychiatry, University of California San Diego, La Jolla, CA 92093 [3]Department of Cellular and Molecular Medicine and Pediatrics, University of California San Diego, La Jolla, CA 92093 [4]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA 92093

Abstract:

Structural Variation (SV) detection from short-read whole genome sequencing is error prone, presenting significant challenges for analysis, *de novo* mutations in particular. Here we describe SV^2, a machine-learning algorithm for genotyping deletions and tandem duplications from paired-end whole genome sequencing data. SV^2 can rapidly integrate variant calls from multiple structural variant discovery algorithms into a unified callset with low rates of false discoveries and Mendelian errors with accurate *de novo* detection.

Introduction:

Structural Variation (SV) is a change of the structure of a chromosome larger than 50bp. SV is a major contributor to human genetic variation with 13% of the human genome defined as structurally variable[1], and is also implicated in a variety of human diseases[2,3]. *De novo* germline SV contribute risk for many congenital disorders, particularly where there is no family history,

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such as idiopathic autism or intellectual disability[4]. Putative *de novo* mutations are enriched for errors as they require only a single false SV genotype, either a false positive in the child or a false negative in the parent. Conversely, errors in inherited variants occur if both the parent and child have false positive genotypes at the same locus. Accurate genotyping is therefore particularly important for *de novo* mutation discovery. Also, given that SV can range in size from 50bp to 50Mb, typically multiple tools are required to fully capture SVs[1,5] with each operating as a standalone solution relying on read depth[6,7] or discordant paired-ends and split-reads[8,9]. Methods for harmonizing variant calls and scores from multiple methods into a unified set of SV genotypes are lacking.

Here we present SV^2 (support-vector structural-variant genotyper), a turn-key solution for unifying SV predictions into an integrated set of genotypes and likelihoods. SV^2 (https://www.github.com/dantaki/SV2) is an open source software written in Python that exploits read depth, discordant paired-ends, and split-reads in a supervised support vector machine classifier[10]. Required inputs include a BAM file with supplementary alignment tags (SA), a single nucleotide variant (SNV) VCF file with allelic depth, and either a BED or VCF file of deletions and tandem duplications to be genotyped. The final product is a VCF file with genotypes and annotations for genes, repeats, and other befitting statistics for SV analysis.

Main

The training set for the genotyping classifiers applies whole genome data and a gold standard of SV positions and genotypes with a reported false discovery rate (FDR) of 1-4%[1] from the 1000 Genomes Project (1KGP). SV^2 combines features from paired-end reads that are descriptive of the copy number state, each of which are implemented in SV prediction tools for

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

next generation sequencing and SNV microarrays[7-10]. Features extracted from variants in 27 unrelated high coverage (48x) samples include depth of coverage, discordant paired-ends, split-reads, and heterozygous allelic depth (HAD) ratio (Figure 1A&B). Given the small number of duplications in the high coverage samples, the duplication training set included 2,493 low coverage (7x) genomes, altogether employing over 32,000 deletions and 22,000 tandem duplications (Supplementary Table 1) in six classifiers (Methods).

We initially sought determination of SV[2] genotyping performance with cross-validation. We calculated the mean receiver operating characteristic (ROC) curve of 7 folds, maintaining the proportion of classes in the full training set. We found the average area under the curve (AUC) for deletions as 0.98 and for tandem duplications as 0.88 (Figure 1B&D). ROC curves for the remaining classifiers (Supplementary Figure 1&2) produced similar AUCs with the exception of the hemizygous deletion classifier. We suspect the suboptimal AUC of the hemizygous deletion classifier to be driven by incorrect gold standard genotypes suggested by mean of normalized coverage in each genotype class (Hemizygous deletion: REF=1.15, ALT= 0.91), in contrast to the hemizygous duplication training set (Hemizygous duplication: REF=0.99, ALT=1.99) (Supplementary Figure 2B&D).

We extended our evaluation of genotype performance using Illumina 2.5M SNV arrays taken from 17 previously published[4] families, totaling 57 individuals. In brief, SV calls were generated using LUMPY[9] and Manta[8] on high coverage whole genomes, merged according to 50% reciprocal overlap, and genotyped by SV[2] (Methods). False discovery rates (FDR) of SV[2] genotypes were estimated via the Intensity Rank Sum test, which produced a FDR of 40% for both unfiltered deletions and duplications (Figure 2A). We then formulated genotype likelihood

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

filters for standard SV analysis and *de novo* variant discovery (Methods & Supplementary Table 2) and found the FDR of SV^2 standard filters to be 1.24% for deletions and 4.41% for duplications. Unfiltered *de novo* variants carry a high FDR of 60% for deletions and 86% for duplications (Figure 2A), which fell to 0.54% for deletions and 0% for duplications with SV^2 *de novo* filters (Supplementary Figure 3).

Estimation of the FDR using SNV arrays may not provide accurate estimates for all classes of SV, because SNV probes overlap with a very small fraction of small SVs (<100 bp in length) and could be biased toward genomic regions with less repetitive sequence where SV detection by WGS is also optimal. Therefore, we also estimated SV^2 genotyping performance with PacBio Single Molecule Real-Time (SMRT) long read technology obtained from the 1KGP. SVs called with LUMPY and Manta were genotyped and merged with SV^2 using complementary paired-end high coverage (74x) genomes for three probands. We then queried PacBio SMRT alignments for supporting reads, defined as split-reads with breakpoints overlapping at least 80% to the paired-end prediction on chromosome 1. SVs were omitted if the PacBio coverage was less than 5 standard deviations from the chromosome mean; likewise, SVs intersecting our genome mask were removed (Methods). We defined true positive variants as those with at least one supporting read, which resulted in a FDR for unfiltered deletions at 20% and duplications at 35% (Figure 2A). With SV^2 standard genotype likelihood filters, the FDR for deletions fell to 12.5% and for duplications 20.1%. However, for *de novo* genotype likelihood filters, the FDR for deletions was 11.9% and 21.1% for duplications. These estimates are likely overestimating the FDR since we are limited by our ability to validate SV with PacBio SMRT sequences with the currently available tools. Other methods of validating SV from PacBio SMRT sequences, such as *de novo* assembly, have been described[11], but these data are not yet publicly available.

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Next, we further complemented our performance analysis by leveraging family-based inheritance, providing an alternative route for estimating genotyping accuracy[12]. We calculated the mean rate of Mendelian errors in 630 probands and their parents, resulting in a total 1,852 samples: 1,554 of which were provided by the Simons Simplex Collection (https://sfari.org/). Briefly, these samples were sequenced to high depths (>30x, read lengths > 100bp) and had SV called by ForestSV[7], LUMPY[9], and Manta[8] with accompanying SV[2] genotype scores. SVs were merged and then excluded if the overlap to repetitive elements and gaps exceeded 50% (Methods). For each proband, we determined the average Mendelian error rate at varying genotype likelihood cutoffs and allele frequencies. Rare variants with lower genotype likelihood scores were more prone to Mendelian inconsistencies, in contrast to rare variants with higher scores, suggesting that SV[2] reliably assigns low scores to false positives. Common variants had higher rates of Mendelian errors at higher genotype likelihoods, which has been observed previously for SVs[13] (Figure 2B). The average Mendelian error rate with standard filters was 0.026 (95% CI [0.025,0.027]) for deletions and 0.068 (95% CI [0.062,0.074]) for duplications, while stringent *de novo* filters produced a Mendelian error rate of 0.012 (95% CI [1.11×10^{-2},1.17×10^{-2}] for deletions and 0.031 (95% CI [0.03,0.033]) for duplications (Supplementary Figure 3).

Further validation incorporated the group-wise transmission disequilibrium test[14], a robust measure of specificity[15]. Bias towards under-transmission signifies either an abundance of type I errors in the parents, and/or of type II errors in the child, complicating tests of family-based association and *de novo* mutation calling. We calculated the average percent of transmitted variants in 630 probands described above. Unfiltered variants exhibited an under-transmission bias of 39.8% for deletions and 35.08 for duplications (deletions $P=9.61 \times 10^{-51}$, duplications $P=7.8 \times 10^{-18}$) (Figure 2C). However, standard genotype likelihood filters reduced under-

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transmission bias to 48.2% ($P=1.32x10^{-2}$) for deletions and 47.3% ($P=3.39x10^{-3}$) for duplications**.** SV^2 *de novo* filters further reduced under-transmission bias to 49.1% ($P=1.32x10^{-2}$) and 49.3% (P=1.0) for duplications (Supplementary Figure 3), confirming SV^2's capacity to mitigate false positives.

After confirming SV^2 genotypes deletions and tandem duplications accurately, we compared SV^2 filters to default filters for SVTyper and Manta and determined the FDR of rare variants filtered by either SVTyper, Manta, or SV^2 genotype likelihoods (Supplementary Figure 4). Deletions filtered by SVTyper and Manta carried a 1.0% and 1.6% FDR respectively, in contrast to SV^2 with 1.24% for standard and 0% *de novo* filters. Likewise, a 7.9% FDR was observed for duplications filtered by SVTyper and 9.6% for Manta, but with 4.4% and 0% FDR for standard and *de novo* SV^2 filters respectively. We then assessed the FDR of putative *de* novo mutations filtered by either method. SVTyper's filters returned 18 putative *de novo* mutations with a 22.2% FDR, Manta's filters produced 13 with 46.2% FDR, and SV^2 standard filtering resulted in 21 variants with 19% FDR. However, SV^2 stringent *de novo* filters produced 9 variants with 0% FDR demonstrating SV^2's ability to accurately resolve putative *de novo* mutations.

SV^2 compared to other SV genotyping software is noteworthy because of its exploitation of machine learning to reliably genotype and score deletion and tandem duplication predictions without compromising sensitivity. One advantage of SV^2 to comparable SV genotyping solutions is the ability to genotype breakpoints overlapping repetitive elements using read depth. Additionally, the incorporation of heterozygous allelic depth is better able to genotype tandem duplications, which are more prone to false positive genotypes due to fluctuations in read depth.

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

However, relying on the presence of SNVs tends to limit accurate genotyping to events larger than 3kbp. A caveat of SV^2 is that it cannot assign a copy number greater than 4, but this can be addressed with the addition of more gold standard examples. Ultimately, SV^2's strength is harmonizing genotypes and likelihoods from multiple callers and genotypers, simplifying analysis of SV and providing a much-needed tool for accurately resolving *de novo* mutations.

Figure Legends:

Figure 1: SV^2 Training Set and Cross Validation Performance

A: Kernel density estimates of 1000 Genomes phase 3 deletions less than 1000bp (left) and duplications (right) in 27 high coverage samples. Colors represent the gold standard phase 3 genotype with copy number on the X axis is a function of depth of coverage. **B:** Depicts the average ROC curve of 7-fold cross validation of the training sets in A, shaded areas represent the 95% confidence interval. The average AUC across all copy numbers for deletions was 0.98 and for duplications 0.88.

Figure 2: SV^2 Genotyping Evaluation

A: False discovery rate estimates from SNV microarrays and PacBio SMRT long reads for deletions (left) and tandem duplications (right). Black dotted line indicates 5% FDR. Unfiltered SV call sets have high rates of false positives (~40%) for both SNV arrays and PacBio SMRT reads. Likewise, unfiltered (raw) *de novo* mutations, had a FDR of 60% for deletions and 86% for duplications, estimated from SNV arrays. **B:** Mendelian error rates in 630 probands for deletions (left) and duplications (right). SV^2 reliably assigns poor genotype scores to false positives. Rare variants with high genotype scores tended to have fewer Mendelian

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

inconsistencies. **C:** Group-wise transmission disequilibrium tests in 630 probands for deletions (left) and duplications (right) with shaded regions representing 1 standard deviation. Unfiltered calls were biased towards under-transmission of variants with an average bias of 39.8% ($P=9.61x10^{-51}$) for deletions and 35.1% ($P=7.8x10^{-18}$) for duplications.

Supplementary Figure 1: Additional Training Set Cross Validation Performance

A: Cross validation performance of the >1000bp deletion training set was performed in similar fashion to Figure 1B. The Mean AUC of 7 folds for all copy number classes was determined to be 0.98. Shaded area represents 95% confidence interval. **B:** Depicts the cross validation performance of the paired-end duplication training set, with 2,494 low coverage samples. The mean AUC across all copy number classes was determined to be 0.84.

Supplementary Figure 2: Hemizygous Training Set Cross Validation Performance

A: Cross validation performance of the hemizygous deletion classifier with a mean AUC of 0.68 with shaded areas indicating 95% confidence intervals. We suspect the suboptimal results of cross validation to be possibly due to incorrect genotyping in the gold standard indicated in **B** where the distribution of coverage between the two classes is not as drastic in comparison to the hemizygous duplication training set in **D.** In contrast, the hemizygous duplication classifier performance in **C** had a mean AUC of 0.98 with distinct separation of copy number groups in **D.** Sample weights applied to each training method compensate for possible genotype errors in the gold standard (Methods).

Supplementary Figure 3: Performance of SV^2 Genotype Likelihood Filters

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A: Genotyping performance of unfiltered (top) and SV^2 filtered (bottom) of deletions at varying bins of SV lengths. False discovery rates (left), Mendelian error rates (center), and transmission bias (right) were estimated for variants called by ForestSV, LUMPY, and Manta and genotyped by SV^2. Unfiltered variants had high rates of false positives at all size bins. However, SV^2 filtering controlled for false positives with SVs<100bp contributing most of the errors. *De novo* filters consistently produced fewer errors than standard filters with the exception of SVs <100bp, which had a larger under-transmission bias attributed to the small number of variants (standard stringency: 3551, *de novo* stringency: 705). Error bars for Mendelian error rates represent 95% confidence intervals. Error bars for transmission disequilibrium test represent 1 standard deviation. **B:** Genotyping performance of unfiltered (top) and SV^2 filtered (bottom) tandem duplications. Similar to deletions, unfiltered duplications contain many false positives which are mitigated with SV^2 filters. *De novo* filters produce fewer false positives than standard filters, allowing for accurate discovery of *de novo* mutations.

Supplementary Figure 4: Comparison of Genotype Likelihood Filters and *De Novo* Prediction Performance

False discovery rates for deletions (left) and duplications (center) filtered by default filters for Manta, SVTyper, and SV^2. FDRs were estimated using SNV arrays and binned according to size. While the default filters for Manta and SVTyper perform well with FDRs less than 2% for deletions and 10% for duplications, SV^2 standard filters result in a 1.24% and 4.4% FDR for deletions and duplications respectively. *De novo* filters produced a 0% FDR for both deletions and duplications. For putative *de novo* mutations (right), SV^2 *de novo* filters produced a 0% FDR in contrast to standard filters.

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods:

SV² Workflow

SV² is a high-throughput SV genotyper that requires BAM alignments with supplementary reads (SA tags), a bgzipped and tabix indexed VCF with allelic depth for variants, and a BED or VCF file of deletion and tandem duplication positions to be genotyped. SV² first performs a preprocessing step that records basic statistics of each chromosome such as median coverage, insert size, and read length. Then SV² operates on each variant extracting informative features for genotyping with six support vector machine classifiers. The classifiers consist of three deletion classifiers and three duplication classifiers. One classifier of each SV type is dedicated to hemizygous variants, defined as those on male sex chromosomes, since there are only two states to classify. One of the two autosome deletion classifiers operates on SVs with lengths greater than 1000bp, where coverage was defined as the number of reads spanning a locus, and those smaller than 1000bp, where the median base-pair depth of coverage was considered. One autosome duplication classifier implements paired-end features: discordant paired-end and split-reads. When paired-end features were not available for duplications, the second duplication classifier instead relied on heterozygous allelic depth as a feature. After genotyping and scoring, a VCF is outputted with annotations for repeat elements, 1KGP phase 3 variant overlap, and genes.

Machine Learning Features of SV²

We sought to leverage SV genotyping with four orthogonal features: depth of coverage, discordant paired-ends, split reads, and heterozygous allelic depth (HAD) ratio. Coverage was

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

defined as either the number of reads spanning a locus or as the median base-pair depth for lengths <=1kbp. Reads were excluded if they aligned within our genome mask comprising of segmental duplications, short tandem repeats, assembly gaps, telomeres, and centromeres. Raw coverage values were normalized according the chromosome average, and then adjusted based on GC content with respect to PCR or PCR-free chemistries, adapted from CNVator[16]. We defined discordant paired-ends to have insert sizes greater than the chromosome median plus 5 times the median absolute deviation. To reduce noise, we limited the search for discordant paired-ends and split-reads to +/-500bp of the start and end positions of the SV. Likewise, only discordant paired-ends and split-reads were included if the mate-pair or the supplementary alignment mapped to the opposite side of the breakpoint. The resulting number of discordant paired-ends and split-reads was then normalized to the number of concordant reads within the locus. Akin to B-allele frequency on SNV microarrays[10], HAD was defined as the median ratio of coverage of the minor allele to the major allele for all heterozygous variants encompassing the SV.

SV^2 Training Set

Features were obtained from 27 PCR-free high coverage whole genomes (48x, 250bp read length) and 2,494 low coverage whole genomes (7x, 100bp read length) provided by 1KGP[17]. SV positions were obtained from the 1KGP phase three structural variation call set[1], retaining alleles with at least one alternate variant in the cohort. Due to the large number of samples for the paired-end duplication classifier, we randomly selected 100,000 homozygous reference examples for the final training set. Features were also excluded if the estimated copy number was greater than 10. Sample weights for training were defined as the inverse distance of expected coverage of the phase 3 genotype. The expected normalized coverage for homozygous

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reference was 1.0. The remaining expected coverages either add or subtract 0.5 from 1.0 according to the number of copies gained or lost. Training samples for the HAD classifier were weighted according to the inverse Euclidian distance of expected coverage and mean HAD value of each copy number class to compensate for allelic dropout.

SV² Classifier Parameter Selection

SV² genotypes SV with a support vector machine model with a radial basis function kernel from scikit-learn[18]. Support vector machine classifiers are governed by the parameters C and gamma, which represent the error of classification and the influence of training samples. Parameter sweeps of varying C and gamma values were performed with balanced class weights, with the exception of the paired-end duplication classifier which used heuristic class weights. Parameters were chosen by optimizing false discovery rate (SVtoolkit) and sensitivity of validated *de novo* variants in a previously published cohort[4].

Cross Validation

We assessed the performance of the training sets with seven-fold cross validation, where each fold maintained the proportion of copy number classes in the full training set. Using the 1KGP phase 3 SV genotypes as truth, the mean ROC and area under the curve was determined for each genotype class (Figure 1B&D, Supplementary Figure 2).

SV Genotyping Performance with SNV Arrays

We evaluated false discovery rates at varying genotype likelihood cutoffs using Illumina 2.5M SNV microarrays and SV calls from high coverage, paired-end whole genomes were obtained from 57 samples described previously[4]. Raw LUMPY[9] and Manta[8] calls were merged according to 50% reciprocal overlap, while removing any call that overlapped 50% of its length

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to a repeat element or an assembly gap. False discovery rates were obtained for the resulting call set using the IRS test from SVtoolkit.

SV Genotyping Performance with PacBio Single Molecule Real-Time Sequencing

We chose 3 probands sequenced using PacBio Single Molecule Real-Time (SMRT) from the 1KGP since they had higher coverage than the parents (proband mean depth = 38.9, parent mean depth=18.6). Raw reads (mean length= 8,345.2bp) were aligned to GRCh38 with bwa mem with the –x pacbio option. We then restricted our analysis to chromosome 1 to comply with 1KGP data release policy for unpublished data. SV calls from LUMPY and Manta were genotyped and merged with SV^2 using complementary Illumina paired-end whole genomes sequenced to deep depths (74.2X) with 125bp reads. We defined supporting reads as PacBio split-reads with breakpoints that reciprocally overlap 80% to SVs genotyped in the paired-end alignments. We omitted loci if the coverage of PacBio reads over a 10kb span of either the start or end position was less than 5 standard deviations from the mean chromosome coverage. Loci were also removed if either one of the breakpoints overlapped an element in our genome mask, in addition to removing calls if the region overlapped 50% to masked elements. False positives were defined as ALT genotypes without supporting PacBio spilt-reads, while true positives required 1 supporting read.

SV Genotyping Performance Leveraging Inheritance

We measured rates of Mendelian errors in 630 high coverage whole genome probands (1884 total individuals). 1,551 of the samples were obtained from the Simons Simplex Collection. SVs were called using ForestSV, LUMPY, and Manta. Raw calls were then genotyped by SV^2 and then merged after filtering. SVs with greater than 50% overlap to regions defined in our genome mask were removed. For each proband the number of inconsistent

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genotypes with respect to the parents was taken as a ratio to Mendelian consistent genotypes. We performed this analysis at varying alternate genotype likelihood cutoffs and allele frequencies and recorded the mean Mendelian error rate for the cohort. Rates of SV transmission were measured with group-wise transmission disequilibrium test (gTDT)[14].

Construction of Standard and De Novo Mutation Filters

Strict genotype likelihood filters were determined using the IRS test from SVtoolkit on previously mentioned sampled. For *de novo* filters, we leveraged variants previously validated by PCR and Sanger sequencing[4] as a guide in determining appropriate filters. We created a set of conditions that consider feature availability and the length of the SV to determine appropriate cutoffs, which can be found in Supplementary Table 2.

Comparison of Genotype Likelihood Filters

We compared SV[2] standard filters and stringent *de novo* filters to default filters from SVTyper and Manta. Variants were called by these two methods and filtered as described above. We restricted this comparison to rare variants defined as less than 1% allele frequency in parents. The FDR for each filter were determined using the IRS test from SVtoolkit while binning on the size of the SV (Supplementary Figure 3). Variants less than 100bp were omitted since genotype poorly on SNV arrays. Putative *de novo* variants were defined as those were both parents are homozygous reference with the proband genotyping as a gain or loss of one copy.

Acknowledgements

We would like to thank the 1000 Genomes Project and the Human Genome Structural Variation Consortium for providing PacBio SMRT alignments and complementary high coverage PCR-free Illumina short-read alignments for chromosome 1. D.A is supported by a T32 training grant

bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from the NIH(GM008666). Analysis of low coverage genomes was funded by a research grant through Amazon Web Services.

References:

1 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

2 Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712 (2010).

3 Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223 (2013).

4 Brandler, W. M. *et al.* Frequency and Complexity of De Novo Structural Mutation in Autism. *Am J Hum Genet* **98**, 667-679, doi:10.1016/j.ajhg.2016.02.018 (2016).

5 Consortium, G. o. t. N. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**, 818-825 (2014).

6 Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nature genetics* **47**, 296-303 (2015).

7 Michaelson, J. J. & Sebat, J. forestSV: structural variant discovery through statistical learning. *Nat Methods* **9**, 819-821, doi:10.1038/nmeth.2085 (2012).

8 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222, doi:10.1093/bioinformatics/btv710 (2016).

9 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).

10 Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-1674, doi:10.1101/gr.6861907 (2007).

11 Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, gr. 214007.214116 (2016).

12 Li, B. *et al.* A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet* **8**, e1002944, doi:10.1371/journal.pgen.1002944 (2012).

13 Zhang, D. *et al.* Accuracy of CNV detection from GWAS data. *PLoS One* **6**, e14511 (2011).

14 Chen, R. *et al.* A haplotype-based framework for group-wise transmission/disequilibrium tests for rare variant association analysis. *Bioinformatics* **31**, 1452-1459 (2015).

15 Yan, Q. *et al.* The impact of genotype calling errors on family-based studies. *Sci Rep* **6**, 28323, doi:10.1038/srep28323 (2016).

16 Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974-984, doi:10.1101/gr.114876.110 (2011).

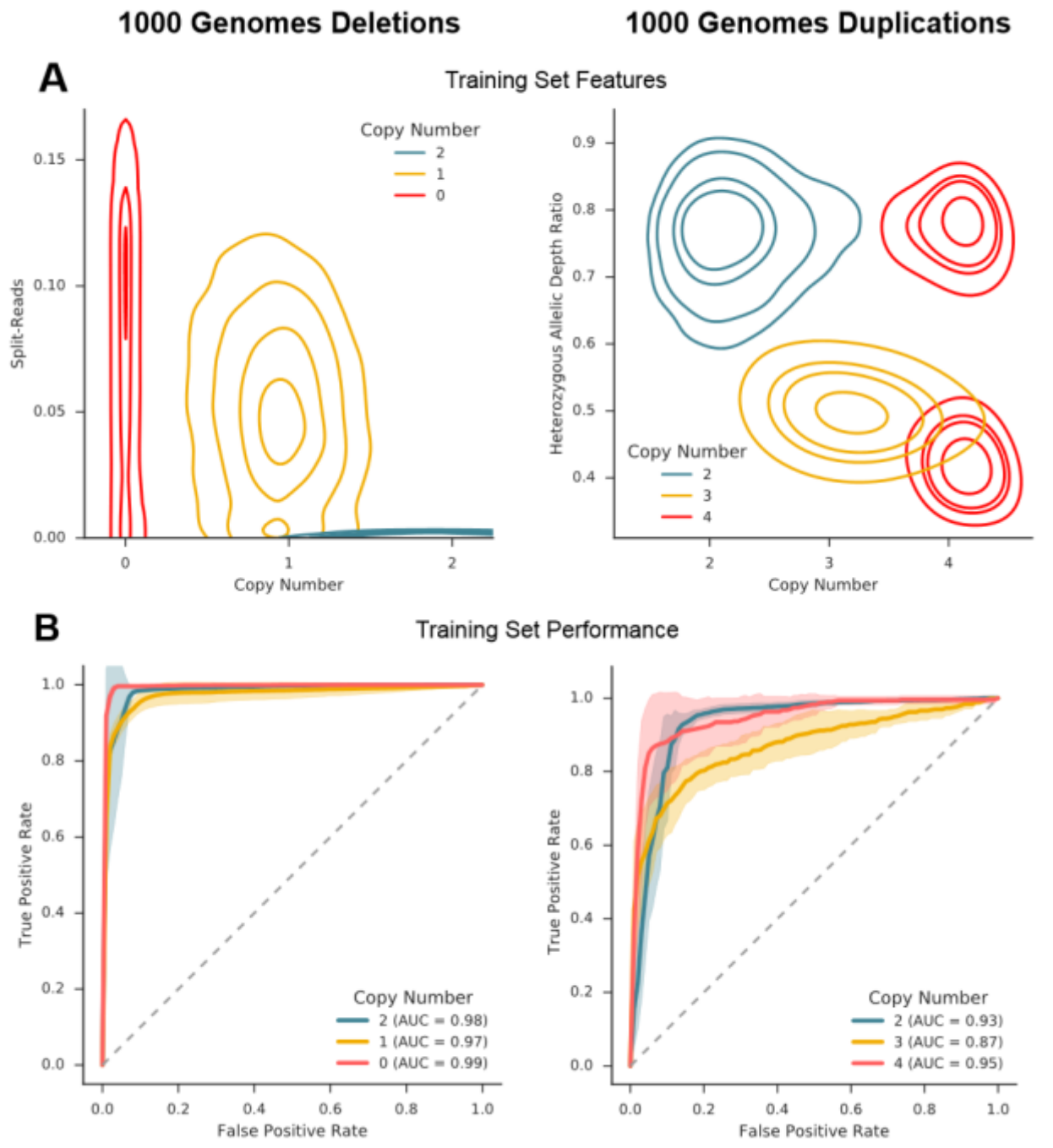
bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

18 Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238* (2013).

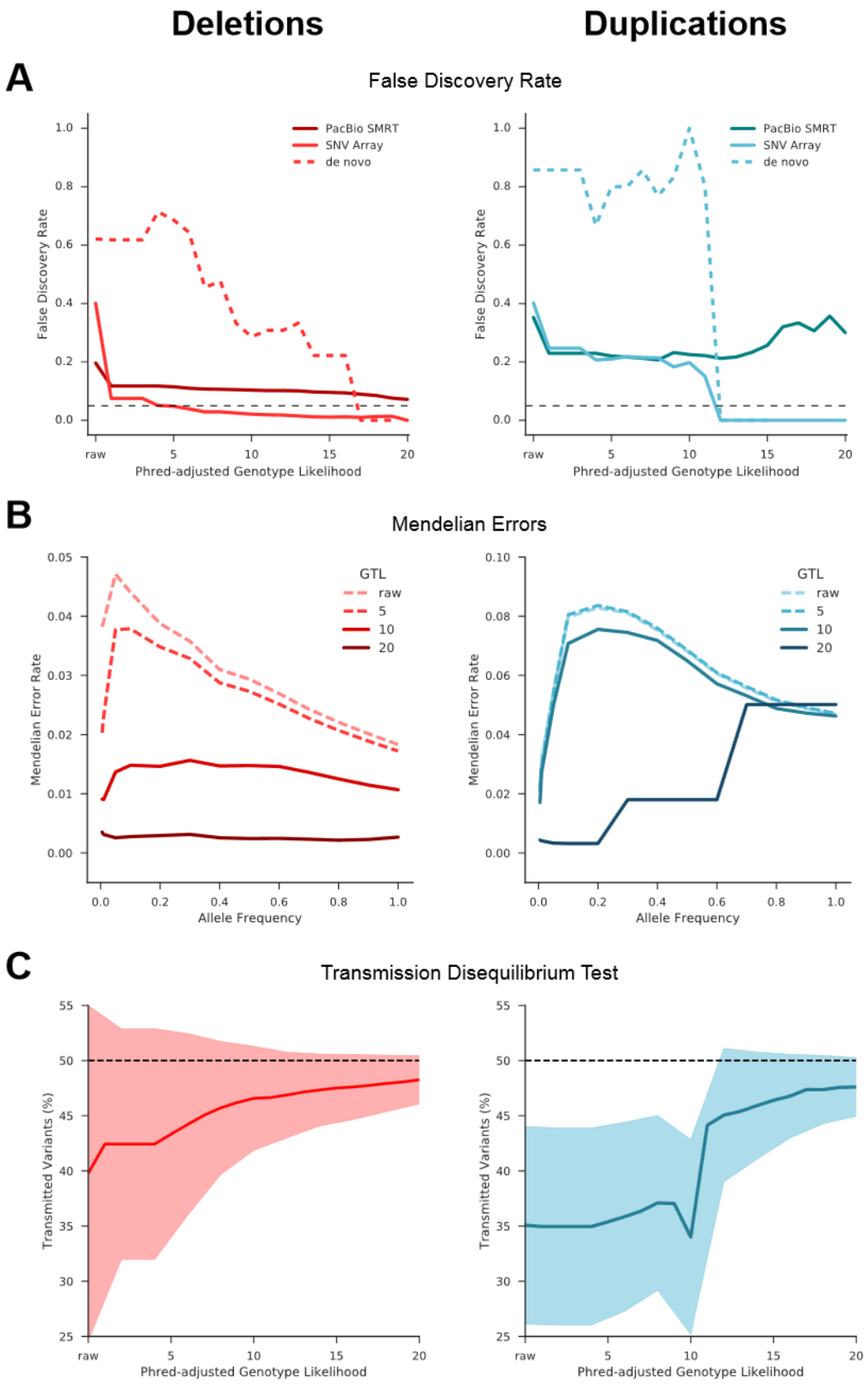
bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1



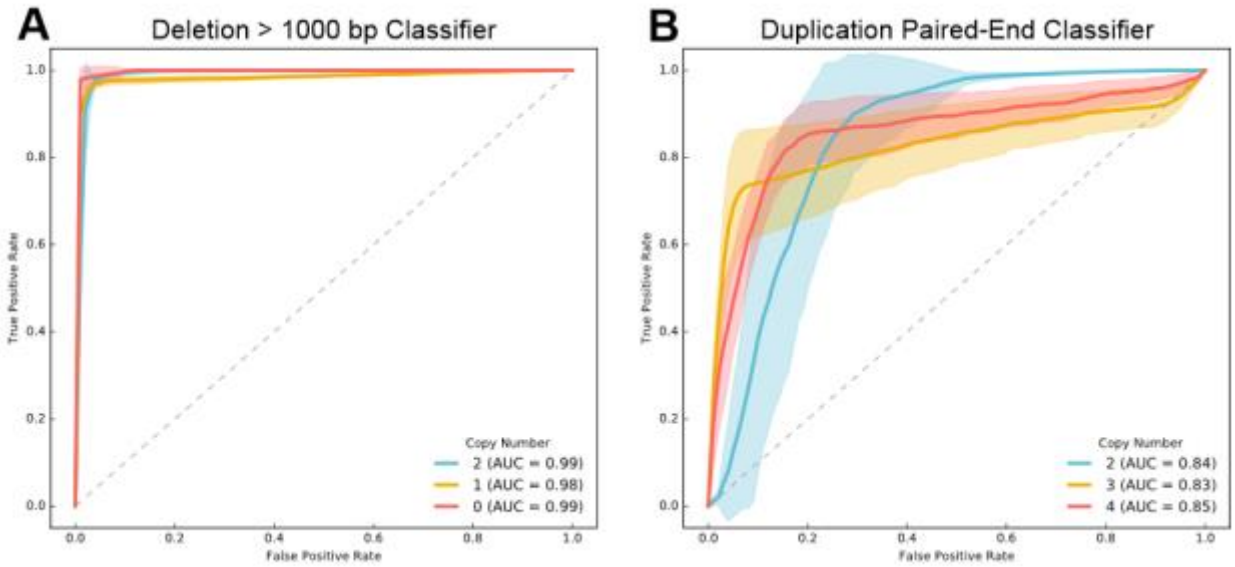
bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2

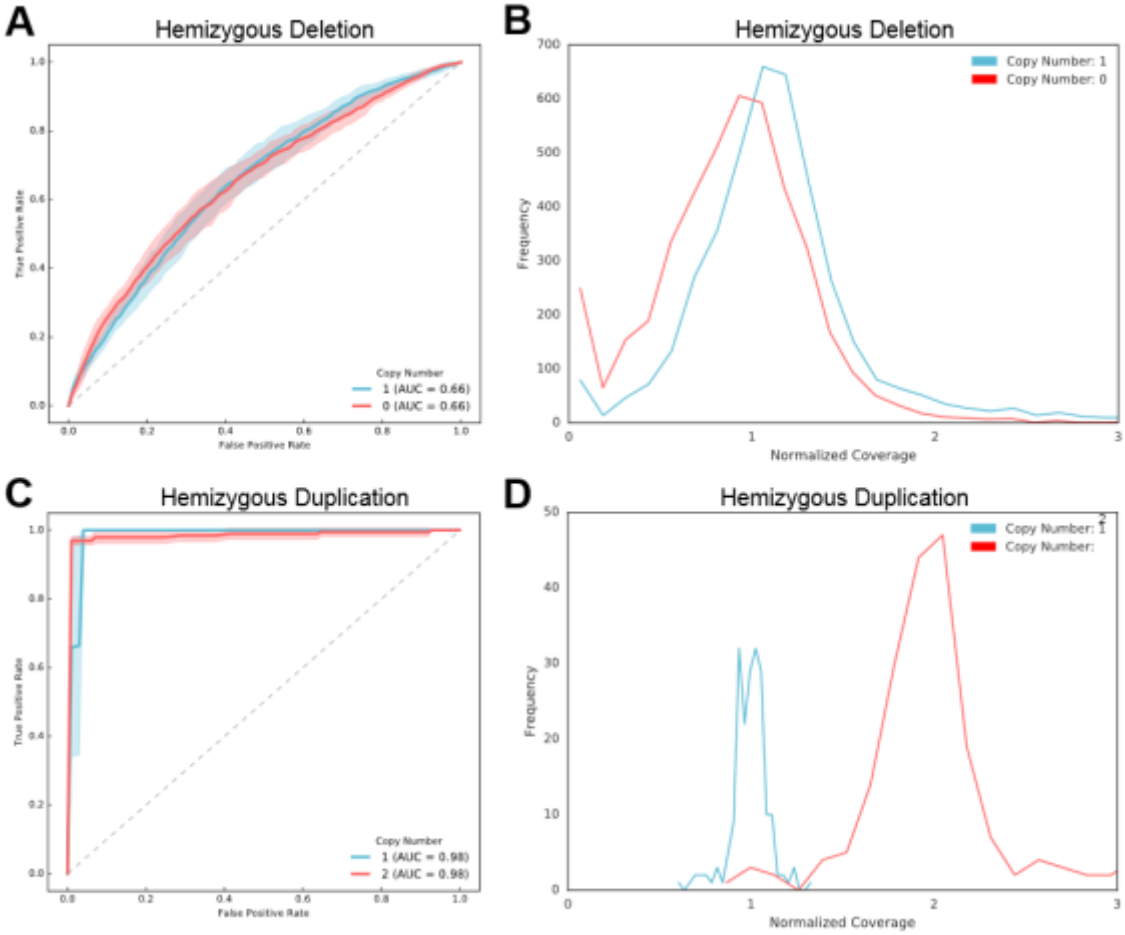


bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 1

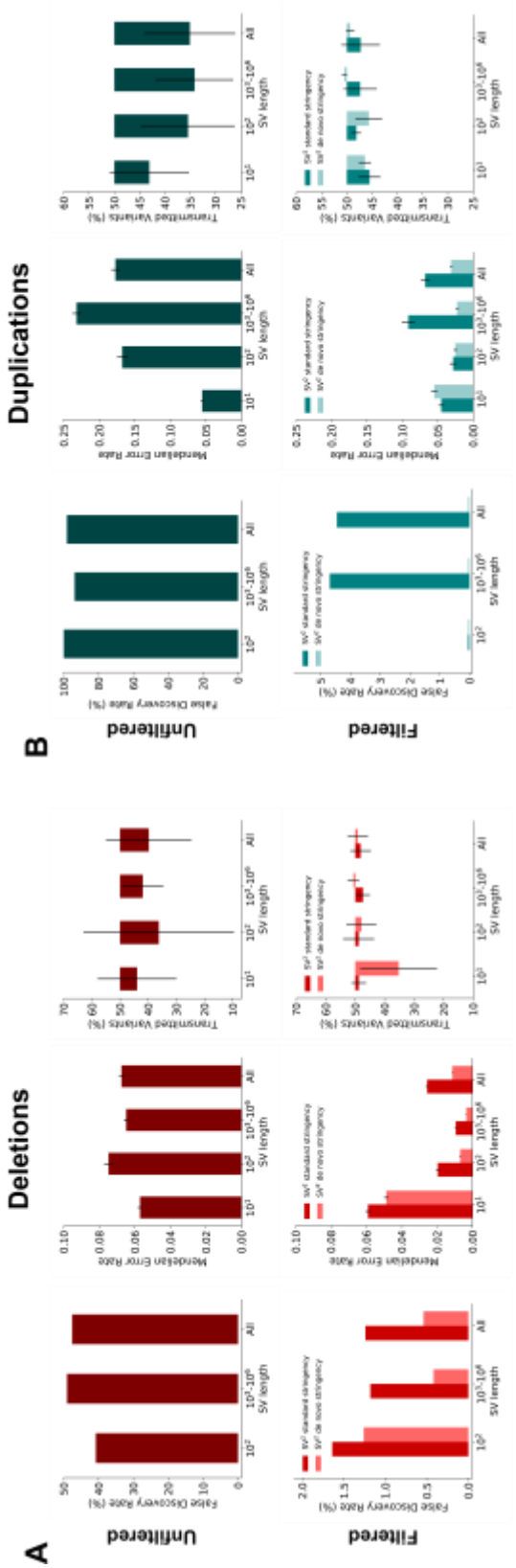


Supplementary Figure 2



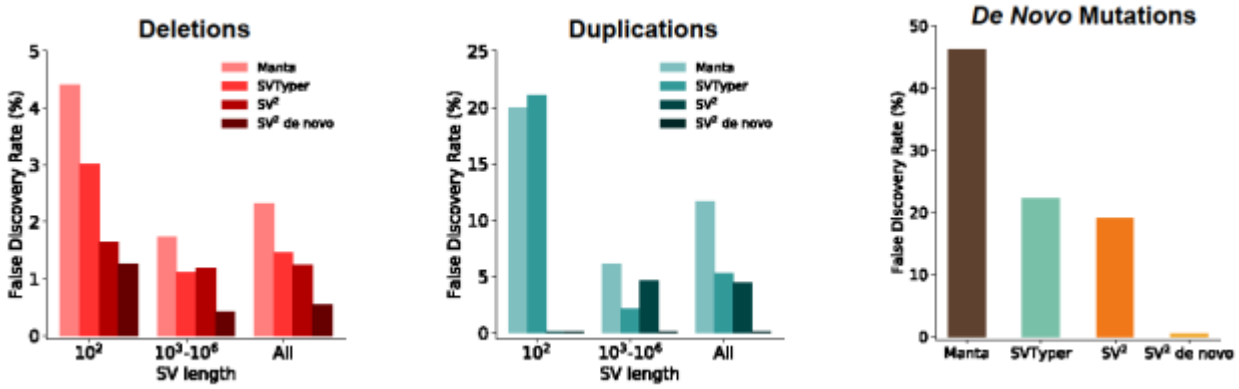
bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 3



bioRxiv preprint doi: https://doi.org/10.1101/113498; this version posted March 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 4



Supplementary Table 1: SV^2 Training Set Variant Counts
Supplementary Table 2: SV^2 Filters for Standard Analysis and *De Novo* Mutation Discovery