Whole-genome sequencing analysis of genomic copy number variation (CNV) using low-coverage and paired-end strategies is highly efficient and outperforms array-based CNV analysis

Bo Zhou,[1,2] Steve S. Ho,[1,2] Xianglong Zhang,[1,2] Reenal Pattni,[1,2] Rajini R. Haraksingh,[1,2,3] Alexander E. Urban[1,2,4,*]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California 94305, USA
[2]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA
[3]Current affiliation: Department of Life Sciences, The University of the West Indies, Saint Augustine, Trinidad and Tobago
[4]Program on Genetics of Brain Function, Stanford Center for Genomics and Personalized Medicine, Tasha and John Morgridge Faculty Scholar, Stanford Child Health Research Institute, Stanford University School of Medicine, Stanford, California 94305, USA
*Correspondence: aeurban@stanford.edu

ABSTRACT

CNV analysis is an integral component to the study of human genomes in both research and clinical settings. Array-based CNV analysis is the current first-tier approach to analyzing CNV in clinical cytogenetics. Decreasing costs in high-throughput sequencing and cloud computing have opened doors for the development of sequencing-based CNV analysis pipelines with fast turnaround times. We carry out a systematic and quantitative comparative analysis of several low-coverage whole-genome sequencing (WGS) strategies to detect copy number variation (CNV) in the human genome and demonstrate that low-coverage WGS is an efficient and comprehensive tool for CNV analysis, in particular when compared to the array-based analysis used in clinical settings. We compared the CNV detection capabilities of WGS strategies (short-insert, 3kb-, and 5kb-insert mate-pair) each at 1X, 3X, and 5X coverages relative to each other and to all current high-density oligonucleotide arrays. As a benchmarking device, we used the Gold Standard and Silver Standard CNVs generated for the genome of 1000-Genomes-Project CEU subject NA12878. Overall, low-coverage WGS detects drastically more Gold Standard CNVs compared to arrays and is accompanied with many fewer CNV calls without secondary validation. Furthermore, we also show that WGS (at ≥1X coverage) is able to detect all seven validated deletion CNVs >100 kb in the NA12878 genome whereas only one of these deletions is detected by most arrays. Lastly, we show that the much larger 15 Mbp Cri-du-chat deletion can be clearly seen at even just 1X genomic sequencing coverage with short-insert paired-end WGS.

Keywords: copy number variation; CNV; array CGH; SNP array; read-depth analysis; discordant read-pair analysis; low-coverage whole-genome sequencing; mate-pair sequencing; Cri-du-Chat deletion

INTRODUCTION

CNV in the human genome sequence contributes a large portion of human genetic diversity [1–5]. Many CNVs, typically those deletions or duplications that are on the smaller end of the size spectrum, are common, i.e. they are present at an overall frequency of more than 1% in the general human population [3–6]. CNVs on the large side of the size spectrum are relatively rare, and many of these large CNVs are associated with human disease, particularly birth defects, neuropsychiatric disorders and with cancer [7–14]. Having technologies available for the reliable and accurate detection and characterization of CNVs in a given human genome is highly relevant for both clinical diagnostics and basic research. Microarray-based CNV analysis has become a first-tier clinical cytogenetics procedure in patients with, for example, unexplained developmental delay/intellectual disability [15], autism spectrum disorder [16], multiple congenital anomalies [17], and cancer [13,14].

The highest sensitivity and resolution in CNV detection is achieved through deep-coverage, paired-end whole-genome sequencing (WGS) [5]. However, the cost for what is currently the standard for deep-coverage WGS (>30x coverage using short-insert paired-end reads) is still considerably higher than for array-based analysis; turnaround time is also much longer since the samples have to go through an offsite core, and the computational requirements are very substantial regarding hardware, time, and personnel. Analysis by current deep-coverage WGS methods can not only detect CNVs but also SNPs, short insertions and deletions (indel), as well as, with some limitations, sequence variants that are quite challenging to parse out such as inversions and retro-transposition events. However, for cytogenetic applications, whether in clinical

diagnostics or basic research, such types of variants are for the most part not yet interpretable as to their clinical effects.

With the advent of bench-top high-throughput DNA sequencers it is now possible to perform low-coverage WGS on-site instead of through a sequencing core facility. To make the most beneficial use of this option, i.e. to control per-sample-costs as well as turnaround times, it seems beneficial to use a strategy of lower sequencing coverage (i.e. 1x-5x genomic coverage) with sample multiplexing to be cost-effective while carefully weighing the options of short-insert versus long-insert paired-end (mate-pair) library preparation.

While other recent studies have demonstrated that WGS, including low-coverage WGS, is effective for CNV detection in clinical samples [18–20], a quantitative and systematic direct performance comparison for CNV analysis, between various low-coverage WGS strategies, relative to deep-coverage WGS, and relative to array-based analysis is needed to fully assess the feasibility of replacing arrays with low-coverage WGS and to guide researchers in their choices for specific settings. Here, we carried out this comparative analysis of the abilities of several different low-coverage WGS approaches to detect and resolve genomic CNVs. We determined the CNV detection capabilities of standard 350 bp short-insert WGS, 3kb-insert mate-pair WGS, and 5kb-insert mate-pair WGS, each at 1x, 3x, and 5x depths of genomic sequencing. Our benchmarking strategy was to use a Gold Standard (GS) set of validated CNVs present in the genome of CEU subject NA12878 (currently probably the best-studied genome in the world) [3,21,22] and to determine the fraction of these GS-CNVs that could be detected by each low-coverage WGS option. For this comparison against GS-CNVs we used the

same approach that was used in Haraksingh *et al* [23] for the benchmarking of high-density oligomer arrays, and we then also compared the low-coverage WGS performance against the results for arrays from this earlier study.

MATERIALS AND METHODS

NA12878 and NA16595 library construction and sequencing

NA12878 and NA16595 LCL DNA aliquots were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (Camden, NJ, USA). DNA aliquots were diluted 10-fold and concentrations were quantified using the Qubit dsDNA HS Assay Kit (Invitrogen Life Technologies, Waltham, MA). DNA was then verified to be pure (OD260/280 > 1.8; OD260/230 > 1.5) by using Nanodrop analysis and of high molecular weight (mean size >30 kb) by using pulse-field gel-electrophoresis. NA16595 DNA was first sheared to 350 bp using the Covaris S2 instrument (Covaris, Woburn, MA), and a NA16595 short-insert WGS library was constructed using the Kapa Hyper Prep Kit (KR0961) from Kapa Biosystems (Woburn, MA, USA), using 100 ng of sheared DNA as input. The library was PCR amplified for 5 cycles. For NA12878, 3kb and 5kb mate-pair sequencing libraries were constructed using the Nextera Mate-pair Library Prep Kit (FC-132-1001) from Illumina following the standard protocol and using 4 μg of DNA as input for each library. Insert size selection (3kb and 5kb) was carried out on the BluePippin instrument from Sage Science (Beverly, MA, USA), specifically the S1 selection marker protocol (tight selection) with a 0.75% cassette. Size selection was then verified using the 2100 Bioanalyzer DNA 12000 Kit (Agilent Technologies, Santa Clara, CA, USA). Mate-pair libraries (3kb and 5kb) were PCR amplified for 10 cycles.

Library sequencing and alignment

NA12878 350 bp short-insert (2X151 bp) sequencing data (32X genomic coverage) was available and obtained from Illumina Public Data (https://basespace.illumina.com/datacentral) (Run ID: NextSeq 500 v2: TruSeq Nano350, (NA12878)_H3GYCBGXX). The NA16595 short-insert (350 bp) paired-end WGS library and the NA12878 mate-pair libraries were sequenced on the Illumina NextSeq 500 platform (2X151bp) using NextSeq Mid-Output 300-cycle Kits (FC-404-2003). For mate-pair libraries, the reverse complement of the Nextera circularized single junction adapter sequence (AGATGTGTATAAGAGACAG) was trimmed from the 3' end of sequences followed by another trimming of the Nextera circularized single junction adapter sequence (CTGTCTCTTATACACATCT) using the FASTQ Toolkit (version 1.0) application on the Illumina Basespace platform (basespace.illumina.com). Illumina external read adapters were trimmed for all libraries before alignment. Trimmed reads were aligned to the hg19 human reference using the BWA Whole Genome Sequencing application version 1.0 from Illumina Basespace [24]. Duplicates were marked and removed using Samtools (version 1.2) [25] followed by Picard tools version 1.52 (http://broadinstitute.github.io/picard/). Alignment BAM files were sub-sampled to coverage depths of 1X, 3X, and 5X using Samtools (version 1.2) [25]. For libraries that underwent discordant read-pair analysis (NA12878 libraries), indels were then locally realigned against the Mills & 1000 Genomes Gold Standard Indels and base scores were recalibrated using Genome Analysis Tool Kit following the standard best practices workflow [26,27]. For CNV analysis, all alignment bam files were subsampled to 1X, 3X, and 5X genome coverage using Samtools (version 1.2) using the "samtools view –s"

option [25], and filtered for reads with MAPQ > 60. Coverage for alignment files was computed using igvtools (version 2.3.82; count, -w 5000 –z 7) with Integrative Genomics Viewer [28].

CNV analysis by read-depth

CNV calling from sequencing data following the principle of read-depth analysis was carried out using CNVnator (version 0.2.7) [29]. Indel-realigned and recalibrated BAM files (see above) were used as input to extract read alignments using CNVnator. A bin size of 5000 bp was used for the intermediate steps involved in running CNVnator: histogram generation, statistics calculation and signal partitioning. Final CNV calling was run using the "-unique" option. The resulting CNVs were filtered for reference gaps (hg19) downloaded from UCSC Genome Browser [30,31], by 50% overlap by genomic coordinate. CNVs were then filtered for calls greater than 300kb that overlapped 50% with either a gap region or the major histocompatibility complex on chromosome six.

CNV analysis by discordant read-pairs

CNV analysis of sequencing data using discordant read-pairs was carried out using LUMPY (version 0.6.11) [32]. Discordantly-mapped reads were first extracted and sorted from the processed alignment file as described in [32]. The *lumpyexpress* command was issued to obtain pre-filtered variant calls. Segmental duplications (hg19) and reference gaps (hg19) downloaded from UCSC Genome Browser [30,31] were excluded from the analysis through the "-x" option. Only deletion and duplication calls were used for analysis (i.e. inversion or translocation calls were not considered). The minimum number of supporting paired-ended reads required to make a CNV call was

set to equal to *c* for short-insert WGS and *c* + 3 for mate-pair WGS where *c* is the fold number of genomic depth of coverage. CNVs larger than 150kb were filtered out.

CNV merging and benchmarking

CNV calls from read-depth and discordant-read-pair analysis from WGS data were merged. CNVs called by read-depth were replaced by CNV calls by discordant read-pairs if the calls overlapped with at least 50% reciprocal overlap in the merged call set. Using the same method of comparison as described in Haraksingh *et al* [23], this merged set of CNVs was compared against the Gold Standard CNV set for the NA12878 genome and split into two groups: (1) greater than or equal to 50% reciprocal overlap, and (2) greater than or equal 10% but less than 50% reciprocal overlap. Merged sets were then compared to Silver Standard CNVs and split into two groups: (1) greater than 50% reciprocal overlap, and (2) less than 50% reciprocal overlap. The sensitivity of CNV calling methods was determined by dividing the number of detected Gold Standard CNVs with >50% evidence by the total number of Gold Standard CNVs.

Array data processing

All NA12878 array data was obtained from [23]. NA16595 LCL DNA ordered from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research (Camden, NJ, USA) was sent to Marcrogen (Rockville, MD, USA) for array hybridization on the MEGA array design (Illumina Infinium platform). Two replicates of NA16595 DNA were hybridized on the Infinium Multi-Ethnic Global-8 arrays (i.e. MEGA array, Illumina, Inc., San Diego, CA, USA). Raw probe intensity data were normalized and processed by Illumina's GenomeStudio software to export the log R ratio (LRR) and B allele frequency (BAF) for all markers. Genotyping coordinates were based on hg19.

QuantiSNP [33] was used to call CNVs. Only CNVs based on signal from > 10 consecutive probes and a Max. Log BAF > 10 were included in the analysis.

Accession Numbers

The accession numbers for sequencing data reported in this paper can be are NCBI Sequencing Read Archive (SRA): bioproroject accession PRJNA407971, biosample accession SAMN07671981 & SAMN07671982, and study SRP118709. The accession numbers for NA16595 array data reported in this paper are NCBI Gene Expression Omnibus (GEO): GSE105092, GSM2817863, and GSM2817864.

RESULTS

CNV detection from array and WGS

We obtained the NA12878 GS-CNVs from Haraksingh *et al* [23]. These GS-CNVs consist of CNV calls curated from high-confidence CNV calls that had been generated by the 1000-Genomes-Project and that had been experimentally validated with orthogonal methods.

From short-insert and long-insert paired-end (i.e. mate-pair) WGS for genome NA12878, we performed CNV detection using both read-depth and discordant-read-pair analysis (Figure 1a). For read-depth analysis, we used CNVnator [29] at a maximum resolution of 5 kb. Discordant-read-pair analysis was performed using LUMPY [32], after removing segmental duplication regions from analysis. CNVs that overlap hg19 reference gaps by greater than 50% were filtered out. After filtering, the union of CNV calls from both analyses, read-depth and read-pair (Figure 1a), was used as the final call set for comparison with the GS-CNVs as well as with the array calls from Haraksingh *et al* [23]. At sequencing depths of 1x, 3x, and 5x genomic coverage, short-

insert WGS detected 182, 405, and 535 autosomal CNVs respectively (Supplementary Table S1, Figure 1b); 3kb mate-pair WGS detected 452, 689, and 748 CNVs respectively (Supplementary Table S1, Figure 1b); 5kb mate-pair WGS detected 496, 572, and 726 CNVs respectively (Supplementary Table S1, Figure 1b)

Performance comparison of CNV detection between WGS and arrays

We obtained from Haraksingh *et al*. [23] the CNVs detected in the genome of subject NA12878 by each of the 17 currently commercially available high-density oligonucleotide arrays, representing three different technologies: aCGH from Agilent (*n*=5), SNP genotyping arrays from Illumina (*n*=10), and aCGH/SNP combination arrays from Affymetrix (*n*=2). Two technical replicates had been performed for each array hybridization, and CNVs were called using both array platform-specific software as well as platform-agnostic software Nexus from Biodiscovery.

We assessed CNV detection by short-insert and mate-pair WGS at genome coverages of 1x, 3x, and 5x using the same benchmarks as described in Haraksingh *et al* [23], where the capability of each array platform was assessed by the numbers of detected CNVs in the NA12878 genome that reciprocally overlap a GS-CNV set of NA12878 CNVs. The GS set of CNVs for the NA12878 genome consists of a subset of the CNVs released by the 1000 Genomes Project [5]. This subset of CNVs was compiled using multiple computational discovery methods from 8x coverage population-scale sequencing (data available 1000genomes.org) and analysis of 2,504 individual genomes [34]. The false positive rate is estimated to be 3.1%. The CNVs in the GS set ranges from 50 bp to 453,312 bp in size with 1,941 and 135 autosomal deletions and duplications respectively [23]. While most GS-CNV duplications are larger than 10 kb, all

CNVs smaller than 1kb are deletions. Seven deletions are larger than 100 kb (Table 1). A NA12878 Silver Standard set of CNVs was also used (also obtained from Haraksingh *et al* [23]) which consisted of CNVs generated using only the CNVnator algorithm on CNVs 2×250 bp short-insert sequencing data at 60x coverage from the 1000-Genomes-Project. We filtered out those CNVs in the Silver Standard set that overlapped hg19 reference gaps (downloaded from UCSC Genome Browser [30,31]) by greater than 50% so that they can be excluded from our analysis. The size distribution of the filtered Silver Standard CNVs is shown in Supplementary Figure S1.

As was done for each of the CNV call sets from each of the 17 array platforms in Haraksingh *et al* [23], the CNV call set for each WGS library at each sequencing coverage depth was assessed by first determining the number of CNVs whose boundaries overlap the boundaries of a GS-CNV by greater than or equal to 50% reciprocally, the number of CNVs whose boundaries overlap boundaries of a GS-CNV by greater than or equal to 10% reciprocally but less than 50%, and from the CNVs whose boundaries do not overlap those of a GS-CNV by > 10% reciprocally, and the number of CNVs whose boundaries overlap those of a Silver Standard CNV by greater than or equal to 50% reciprocally (Supplementary Table S1, Figure 1b). Since more CNV calls and GS-CNV overlaps resulted from the array platform-specific analysis overall and not all arrays had the second replicate, we used the array CNV calls from platform-specific analysis of the first replicate of each array for our comparison. With the exception of short-insert WGS at 1x coverage, all WGS CNV call sets were able to detect significantly more CNVs and NA12878 GS CNVs than any of the arrays and these calls were also accompanied by fewer CNV calls without secondary validation (i.e. CNV calls that are not of high

confidence) (Figure 1b, Supplementary Table S1). At sequencing depths of 1x, 3x, and 5x and by > 50% reciprocal overlap, short-insert WGS detected 49, 165, and 245 GS autosomal CNVs respectively; 3kb mate-pair WGS detected 251, 381, and 408 GS autosomal CNVs respectively; and 5kb mate-pair WGS detected 271, 276, and 337 GS autosomal CNVs respectively. As expected, read-depth analysis resulted in higher rates of CNV detection with increasing sequencing coverage but showed similar numbers of CNV calls across different WGS libraries (Supplementary Figure S2a, Supplementary Table S2). Discordant read-pair analysis resulted in consistently more CNV calls than read-depth analysis for mate-pair libraries at all coverages and short-insert library at 5x coverage (Supplementary Figure S2b, Supplementary Table S3).

Of the arrays, the Illumina HumanOmni1 Quad array detected the highest number of GS CNVs (131); however, even at 1x coverage, 3kb- and 5kb-mate-pair WGS were able to detect approximately 2 fold more GS-CNVs (251 and 271 respectively) (Supplementary Table S1, Figure 1b). While Agilent 2x400K CNV and Illumina HumanOmni5Exome-4v1 arrays could detect comparable numbers of CNV to that of short-insert WGS at 3x coverage and to that of mate-pair WGS at 1x coverage, the vast majority of CNVs detected in these two arrays have no secondary support, i.e. are not of high confidence, in contrast to low-coverage WGS based results where the majority of CNVs detected for all insert-lengths and at all sequencing coverages have secondary support (Figure 1b, Supplementary Figure S3, Supplementary Table S1).

At 5x coverage, 3kb-mate-pair WGS detected the most GS NA12878 CNVs (408), and also had the second lowest portion of CNV calls with no secondary evidence (22.7%) compared to the other WGS libraries (Figure 1b, Supplementary Figure S3,

Supplementary Table S1). The portion of CNV calls with no secondary evidence ranges from 22%-28% for mate-pair WGS and 29%-35% for short-insert WGS (Supplementary Figure S3, Supplementary Table S1). For all mate-pair WGS CNV call sets, the portion of CNVs without secondary support is lower than that of all array platforms with the exception of Affymetrix SNP 6.0 (10.6%), and for short-insert WGS CNV call sets this portion is lower than all array platforms except for the Affymetrix SNP 6.0 (10.6%) and the now discontinued Illumina HumanOmni1Quad platforms (29.7%).

Sensitivity of GS CNV detection

Overall, low-coverage WGS is able to detect approximately 5-fold more GS-CNVs on average compared to arrays (Figure 2a). For example the most sensitive array (as measured by the number of GS CNVs detected) Illumina HumanOmni1 Quad, detects less than one third as many GS CNVs than the most sensitive WGS method (3kb-mate-pair at 5x coverage). The second least sensitive WGS method (short-insert at 3x coverage) is still more sensitive than Illumina HumanOmni1 Quad. Even the least sensitive WGS method (short-insert at 1x coverage) is more sensitive than ten out of the 17 of arrays. It is also important to note that all high-throughput methods of CNV detection left more than 80% of GS NA12878 CNVs undetected (Figure 2a). Mate-pair WGS, however, was able to detect more than 50% of GS CNVs in the 5kb-10kb size range (Figure 2b-d).

While CNV detection increases with sequencing depth for all WGS libraries, the increase is non-linear. The highest increases were observed from 1x to 3x coverage for short-insert and 3kb-mate-pair WGS (Figure 1). While more CNVs were consistently detected from mate-pair WGS compared to short-insert WGS, interestingly, at 3x and 5x

sequencing coverage more total CNVs and GS CNVs were detected from 3kb-mate-pair WGS than from 5kb-mate-pair WGS at 3x and 5x coverages, respectively (Figure 1b, Figure 2a). In addition, while an increase in coverage is associated with an overall increase in the detection of GS-CNVs in the NA12878 genome (Figure 1b, Figure 2a), adding additional sequencing coverage seems to have less of an effect in detecting additional GS-CNVs as CNV sizes increase above 50kb (Figure 2b-d). In short-insert WGS, as sequencing depth increases, the most drastic increases in detected GS=CNVs are between 5kb and 50kb (Figure 2b). For mate-pair WGS, this change over size range is similar, though less pronounced (Figure 2c, d). Both short-insert and mate-pair libraries show a negligible increase in the detection of very large GS-CNVs (50kb-500kb) with increasing coverage (Figure 2b-d).

Size distribution of CNV calls

The size of CNVs detected in short-insert and mate-pair WGS ranges from 100bp to 500kb and 1kb to 500kb respectively (Figure 3, Tables S2-S4). Read depth and discordant read pair analysis detect CNVs in different size ranges (Figure 3d-I, Supplementary Figure S4). Overall, WGS detects CNVs in a wider size range compared to arrays (Supplementary Figure S4). Using a CNVnator [29] bin size of 5 kb, all resulting CNVs detected using read depth analysis are greater than 5 kb (Figure 3d,e,f). As expected, the size distributions for all CNVs called by WGS based read-depth analysis are very similar (Figure 3d-f), whereas discordant read-pair detection shows more variability in the sizes of detected CNVs (Figure 3g-i) due to experimental variability in library insert size. A greater proportion of CNVs below 5kb are detected in 3kb mate-pair WGS compared to 5kb mate-pair WGS (Figure 3b,c) since for larger

insert sizes the experimental variability in the size selection step during library preparation makes for increased uncertainty in calling of smaller CNVs (at the same time a larger insert size will yield a greater ability to span CNV boundaries with multiple constructs, thus increasing overall detection power, especially for CNVs in the medium to larger size range). Furthermore, the increases in the numbers of CNVs detected as a result of increases in sequencing depth for both read-depth and discordant-read-pair analyses are mainly for CNVs less than 50kb in size (Figure 3a-i).

All seven GS deletions larger than 100kb are detected by mate-pair WGS at or above 3x coverage

All seven large deletions were detected with all WGS libraries and with greater than 50% reciprocal overlap with the exception of the chr6:78,892,808-79,053,430 deletion (mean = 44% overlap) (Supplementary Table S5). Previously it had been shown that only one of seven NA12878 GS deletions larger than 100kb is detected by most arrays [23]. Affymetrix SNP6.0 and Affymetrix CytoScanHD performed the best out of the 17 arrays, detecting six and five of these deletions respectively (Table 1). Four out of these seven deletions were detected, but called as high copy duplications in the Agilent 2x400 CGH array. The NA12878 GS deletion on chromosome 3 (chr3: 162,514,471-162,625,647) was only detected in the Agilent arrays but was consistently mis-called as a duplication CNV.

Detection of the Cri-du-chat deletion by read-depth analysis only

We also demonstrate that a deletion of a much larger scale, the 15 Mbp Cri-du-chat deletion on chromosome 5 (DNA sample NA16595 from the Coriell Institute), can be readily detected at sequencing coverages of 1x, 3x, and 5x from a short-insert WGS

library, by using read-depth analysis only. The large Cri-du-chat deletion was detected at all coverages, as indicated by the substantial drop in read-depth coverage when viewed in the Integrative Genomics Viewer [28], spanning from 5p15.31 to 5p14.2 (~15Mbp) (Figure 4 and Supplementary Table S6). The Cri-du-chat deletion from this NA16595 DNA sample was also confirmed using the Illumina Multi-Ethnic Genotyping Array in two technical replicates (Supplementary Table S6).

DISCUSSION

Using the CNVs in the genome of the 1000-Genomes-Project CEU genome NA12878 as a benchmark, we systematically compared the CNV detection performances of low-coverage WGS relative to each other and relative to deep-coverage WGS (by virtue of the Silver Standard list of CNVs). Furthermore, we compared the performance of the sequencing based strategies to that of the various array-based analysis platforms that are currently in routine-use for cytogenetics. CNVs were called using standard methods for both low-depth WGS and arrays, and the detected CNVs were then compared to a list of Gold Standard CNVs in the genome of NA12878 that had been distilled from the 1000-Genomes Project, as well as to a set of Silver Standard CNVs from NA12878 WGS 2×250 bp short-insert sequencing at 60x coverage. In almost all scenarios, the WGS-based analysis shows a considerably higher sensitivity at detecting GS-CNVs than even the best performing arrays, which is furthermore accompanied by many fewer CNV calls with no secondary support (Figure 1b, Figure 2a, Supplementary Figure S3). We also show that short-insert and mate-pair WGS at or above 1x coverage were able to detect all seven GS deletions larger than

100 kb in the genome of NA12878 whereas only one of these deletions is detectable with most arrays (Table 1).

Although the methods for CNV detection from WGS data have been available for up to a decade [1,29,35,36], the cost of sequencing as well as long turnaround times for deep-coverage WGS and heavy computational requirements for analysis have been major obstacles that have prevented the adoption of WGS-based methods for cytogenetic applications. Our comparative analysis here shows that these obstacles can be overcome by adopting low-coverage WGS strategies. This means that a cytogenetics laboratory can now avail itself of a technology with a CNV-detection and resolving power that compares very favorably with existing standard methodologies such as array-based analysis or karyotyping while not having to accept an increased burden in terms of cost per sample or turnaround time. Our WGS libraries were sequenced on the Illumina NextSeq 500, in multiplexed fashion, at a cost in sequencing consumables of approximately $150 per 1x coverage. The short-insert libraries were prepared using the Kapa Hyper Prep kit from Kapa Biosystems (Santa Clara, CA). The cost per library is approximately $40-$50 with 50 ng of genomic DNA needed as input for a library of high complexity. The mate-pair library construction reagents cost approximately $300 per library when using the Illumina Mate Pair Library Prep kit and there is a requirement of 2 - 4 μg of high molecular weight genomic DNA (mean size > 20 kb). The mate-pair library construction costs can be decreased further to about $50-$80 per library if mechanical DNA shearing is employed instead of the enzymatic shearing that is used in the Illumina Mate Pair Library Prep kit [37]. The time frame for the entire WGS process (library construction, sequencing, and data analysis) takes

approximately 4-5 days for mate-pair libraries and 2.5-3 days for WGS libraries. The amount of input genomic DNA required for mate-pair WGS is approximately 2-fold more than for most arrays. However, as long as this amount of DNA is available it could be reasoned that it is preferable to use mate-pair WGS for CNV analysis instead of arrays, considering for example that mate-pair WGS even at just 1x genomic coverage is much more sensitive at detecting genomic CNVs than all currently used arrays. For samples for which DNA is limited, short-insert WGS at just 3x genomic sequence coverage (at a cost of circa $350 per sample using bench-top instruments such as the Illumina NextSeq 500) is still a WGS based approach that is as effective, if not more so, in terms of cost and effort required than arrays while easily outperforming these same arrays in terms of ability to detect and resolve CNVs. It should also be taken into consideration that sequencing costs will only continue to decrease thus rendering the use of WGS for CNV detection even more cost effective over the foreseeable future.

In addition to CNVnator [29] and LUMPY [32], There are currently numerous computational CNV detection tools openly available that have been developed for the use with WGS data based on the principles of discordant read-pair analysis [38–42], read depth analysis [43–47], or combinations of these two principles [48–51]. The choice of algorithm used for CNV analysis from WGS data is likely to greatly impact the number and accuracy of CNV calls that can be made [36,52] and a comprehensive comparative analysis of these algorithms was beyond the scope of this present study. We chose to use CNVnator and LUMPY mainly due to software availability and extensive previous usage experience on our part. Further investigation into the comparative performances of computational tools and optimization of their operating parameters will very likely be

beneficial. It is important to note that relatively few of the currently available software tools can use WGS mate-pair data as their input, which somewhat narrows the choices available to investigators and practitioners of cytogenetics, but this is not a major drawback as even just the software packages used in this study already performed quite well for the purpose at hand.

We have found that using a combination of read-depth analysis and discordant-read-pair analysis produces a more confident CNV call set. Generally, large CNVs called from discordant-read-pair analysis without corroboration from read-depth analysis are likely to be false positives since large CNVs are called more robustly using read-depth analysis (Figure 3d-i). However, when a CNV is called by both types of analyses, discordant-read-pair analysis has the ability to provide a more accurate genomic boundary of the CNV. Gaps in the reference genome are problematic for CNV calling using read-depth analysis, where the absence of read alignment in the gap-regions can be interpreted as deletions by the read-depth analysis algorithm, thus CNV calls that overlap these genomic regions should be flagged and looked at with extra scrutiny (the same is true for array-based approaches). In addition, deletions and reference gaps that lie in close proximity are sometimes merged in read-depth analysis, particularly at low depth of sequencing. For example, the large NA12878 GS CNV on chromosome 19 (chr19:20,595,835-20,717,950, a 22,115bp deletion) is detected but called as a much larger 215kb deletion (chr19:20,505,001-20,720,000) when using read-depth analysis because, at low depth of coverage, the read-depth signal from this deletion is merged by the CNVnator algorithm with a nearby reference gap (chr19:20,523,415-20,573,415) as well as with a Silver Standard deletion (chr19:20,504,101-20,586,400).

However, this 122kb GS deletion on chromosome 19 is detected as a 122,419bp deletion (chr19:20,595,751-20,722,956) using discordant read pair analysis from 5kb mate-pair WGS at 5x coverage.

In read-depth analysis, the bin size chosen for CNV calling dictates the lower size limit of CNV detection. Since our WGS data is of low coverage, we chose a relatively large bin size (5kb) in order to gain higher confidence for the read-depth based CNV calls. In discordant read pair analysis, however, the library insert size determines in several ways the size range of detectable CNVs. In general, CNVs can more easily be detected at larger insert sizes due to the increase in physical coverage, i.e. the number of informative mate-pairs, but at the expense of losing detection power for smaller CNVs because of an increase in uncertainty due to the increased experimental variability during fragment size selection. By combining read-depth and discordant-read-pair analysis, the 3kb mate-pair WGS library provides the widest size distribution of detected CNVs.

In summary, we show that overall low-coverage WGS is drastically more sensitive at detecting CNVs compared to the best performing arrays currently available commercially and that WGS-based detection is accompanied with fewer calls without secondary validation. We also show that the prospect of replacing arrays with low-coverage WGS in a cytogenetic context seems promising and essentially at hand. These results taken together will contribute to the discussion on when and via which route it will be plausible to transition from array-based methods to WGS-based approaches in cytogenetics practice.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported by the Stanford Medicine Faculty Innovation Program and from the National Institutes of Health NHGRI grant P50 HG007735.

REFERENCES

1. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007;318(5849):420-426. doi:10.1126/science.1149504.
2. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*. 2009;10(4):241-251. doi:10.1038/nrg2554.
3. Consortium T 1000 GP. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073. doi:10.1038/nature09534.
4. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65. doi:10.1038/nature11632.
5. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393.
6. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004;305(5683):525-528. doi:10.1126/science.1098918.
7. de Vries BBA, Pfundt R, Leisink M, et al. Diagnostic genome profiling in mental retardation. *Am J Hum Genet*. 2005;77(4):606-616. doi:10.1086/491719.
8. Sharp AJ, Hansen S, Selzer RR, et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet*. 2006;38(9):1038-1042. doi:10.1038/ng1862.
9. Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism. *Science*. 2007;316(5823):445-449. doi:10.1126/science.1138659.
10. Walsh T, McClellan JM, McCarthy SE, et al. Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science (80-)*. 2008;320(5875):539-543. doi:10.1126/science.1155174.
11. Girirajan S, Campbell CD, Eichler EE. Human Copy Number Variation and Complex Genetic Disease. *Annu Rev Genet*. 2011;45(1):203-226. doi:10.1146/annurev-genet-102209-163544.
12. Zhang Y, Haraksingh R, Grubert F, et al. Child development and structural variation in the human genome. *Child Dev*. 2013;84(1):34-48. doi:10.1111/cdev.12051.
13. Michels E, De Preter K, Van Roy N, Speleman F. Detection of DNA copy number

alterations in cancer by array comparative genomic hybridization. *Genet Med*. 2007;9(9):574-584. doi:10.1097/GIM.0b013e318145b25b.

14. Uddin M, Thiruvahindrapuram B, Walker S, et al. A high-resolution copy-number variation resource for clinical and population genetics. *Genet Med*. 2015;17(9):747-752. doi:10.1038/gim.2014.178.
15. D'Arrigo S, Gavazzi F, Alfei E, et al. The Diagnostic Yield of Array Comparative Genomic Hybridization Is High Regardless of Severity of Intellectual Disability/Developmental Delay in Children. *J Child Neurol*. 2016;31(6):691-699. doi:10.1177/0883073815613562.
16. Stobbe G, Liu Y, Wu R, Hudgings LH, Thompson O, Hisama FM. Diagnostic yield of array comparative genomic hybridization in adults with autism spectrum disorders. *Genet Med*. 2014;16(1):70-77. doi:10.1038/gim.2013.78.
17. Gijsbers AC, Lew JY, Bosch CA, et al. A new diagnostic workflow for patients with mental retardation and/or multiple congenital abnormalities: test arrays first. *Eur J Hum Genet*. 2009;17(11):1394-1402. doi:10.1038/ejhg.2009.74.
18. Liang D, Peng Y, Lv W, et al. Copy number variation sequencing for comprehensive diagnosis of chromosome disease syndromes. *J Mol Diagn*. 2014;16(5):519-526. doi:10.1016/j.jmoldx.2014.05.002.
19. Liu S, Song L, Cram DS, et al. Traditional karyotyping vs copy number variation sequencing for detection of chromosomal abnormalities associated with spontaneous miscarriage. *Ultrasound Obstet Gynecol*. 2015;46(4):472-477. doi:10.1002/uog.14849.
20. Dong Z, Zhang J, Hu P, et al. Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med*. 2016;18(9):940-948. doi:10.1038/gim.2015.199.
21. Zook JM, Chapman B, Wang J, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*. 2014;32(3):246-251. doi:10.1038/nbt.2835.
22. Eberle MA, Fritzilas E, Krusche P, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res*. 2017;27(1):157-164. doi:10.1101/gr.210500.116.
23. Haraksingh RR, Abyzov A, Urban AE. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics*. 2017;18(1):321. doi:10.1186/s12864-017-3658-x.
24. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324.
25. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.
26. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013;43:11.10.1-33. doi:10.1002/0471250953.bi1110s43.
27. DePristo M a, Banks E, Poplin R, et al. A framework for variation discovery and

genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498. doi:10.1038/ng.806.

28. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26. doi:10.1038/nbt.1754.
29. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011;21(6):974-984. doi:10.1101/gr.114876.110.
30. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006. doi:10.1101/gr.229102. Article published online before print in May 2002.
31. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32(Database issue):D493-6. doi:10.1093/nar/gkh103.
32. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84. doi:10.1186/gb-2014-15-6-r84.
33. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35(6):2013-2025. doi:10.1093/nar/gkm076.
34. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81. doi:10.1038/nature15394.
35. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14(Suppl 11):S1. doi:10.1186/1471-2105-14-S11-S1.
36. Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*. 2014;15(2):256-278. doi:10.1093/bib/bbs086.
37. Park N, Shirley L, Gu Y, Keane TM, Swerdlow H, Quail M a. An improved approach to mate-paired library preparation for Illumina sequencing. *Methods Next Gener Seq*. 2013;1(July):10-20. doi:10.2478/mngs-2013-0001.
38. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6(9):677-681. doi:10.1038/nmeth.1363.
39. Korbel JO, Abyzov A, Mu XJ, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol*. 2009;10(2):R23. doi:10.1186/gb-2009-10-2-r23.
40. Lee S, Hormozdiari F, Alkan C, Brudno M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods*. 2009;6(7):473-474. doi:10.1038/nmeth.f.256.
41. Hormozdiari F, Hajirasouliha I, Dao P, et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*.

2010;26(12):i350-7. doi:10.1093/bioinformatics/btq216.
42. Marschall T, Costa IG, Canzar S, et al. CLEVER: clique-enumerating variant finder. *Bioinformatics*. 2012;28(22):2875-2882. doi:10.1093/bioinformatics/bts566.
43. Chiang DY, Getz G, Jaffe DB, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*. 2009;6(1):99-103. doi:10.1038/nmeth.1276.
44. Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavaré S. CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics*. 2010;26(24):3051-3058. doi:10.1093/bioinformatics/btq587.
45. Kim T-M, Luquette LJ, Xi R, Park PJ. rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics*. 2010;11:432. doi:10.1186/1471-2105-11-432.
46. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*. 2011;6(1):e16327. doi:10.1371/journal.pone.0016327.
47. Xi R, Hadjipanayis AG, Luquette LJ, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A*. 2011;108(46):E1128-36. doi:10.1073/pnas.1110574108.
48. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. *Genome Res*. 2010;20(11):1613-1622. doi:10.1101/gr.106344.110.
49. Zeitouni B, Boeva V, Janoueix-Lerosey I, et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*. 2010;26(15):1895-1896. doi:10.1093/bioinformatics/btq293.
50. Qi J, Zhao F. inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res*. 2011;39(Web Server issue):W567-75. doi:10.1093/nar/gkr506.
51. Sindi SS, Onal S, Peng LC, Wu H-T, Raphael BJ. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol*. 2012;13(3):R22. doi:10.1186/gb-2012-13-3-r22.
52. Tattini L, D'Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol*. 2015;3. doi:10.3389/fbioe.2015.00092.
53. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 2011;12(5):363-376. doi:10.1038/nrg2958.

Figure Legends

Figure 1. CNV Calls by WGS and Comparisons with Arrays. (a) Schematic diagram of detection of CNVs (deletions and duplications) using discordant read-pair and read-depth analysis [1,29,53]. Using discordant read-pair analysis, deletions are detected when the distance of alignment to the reference genome between read pairs are closer than the expected insert size of the library, and duplications are detected when the orientation of the aligned read pairs are inversed. Using read-depth analysis, deletions and duplications are detected when there is a pronounced decrease and increase, respectively, in alignments of reads spanning a genomic region relative to the average number of alignments over the genome. **(b)** Number of autosomal CNVs in the genome of subject NA12878 called from short-insert, 3kb mate-pair, and 5kb mate-pair libraries sequenced at 1X, 3X, and 5X coverage compared against array platforms as previously described [23]. Array-based CNV calls were made according to platform-specific algorithms, and WGS CNV calls were made by combining discordant read pair and read-depth analysis. Gold: validated autosomal CNVs (overlap >50% reciprocally with NA12878 GS CNVs). Green: 10%-50% reciprocal overlap with NA12878 GS CNVs. Blue: <10% overlap with GS CNVs, >50% reciprocal overlap with NA12878 Silver Standard CNVs. Red: no overlap (<10% overlap with GS CNVs and <50% overlap with Silver Standard CNVs).

Figure 2. Sensitivity of WGS Detection of NA12878 GS CNVs. (a) Sensitivity of GS CNV detection across WGS libraries and array platforms, as determined by the ratio of detected autosomal GS CNVs to total number of autosomal GS CNVS. Array-based CNV calls were made according to platform-specific algorithms, and WGS CNV calls were made by combining discordant read-pair and read-depth analysis. Green: autosomal CNVs that overlap >50% reciprocally with NA12878 GS CNVs. Blue: total number of NA12878 GS autosomal CNVs. Sensitivity of GS CNV detection in different size ranges from **(b)** short-insert, **(c)** 3kb mate-pair, and **(d)** 5kb mate-pair libraries at sequencing depths 1X, 3X, and 5X. CNVs were called by combining discordant read-pair and read-depth analysis.

Figure 3. **Size Distributions of NA12878 Detected by WGS.** Size distribution of CNVs detected in the NA12878 genome from **(a)** short-insert, **(b)** 3kb mate-pair, and **(c)** 5kb mate-pair libraries called by combining discordant read-pair and read-depth analysis. CNVs called from **(d)** short-insert, **(e)** 3kb mate-pair, and **(f)** 5kb mate-pair libraries by read-depth analysis only. CNVs called from **(g)** short-insert, **(h)** 3kb mate-pair, and **(i)** 5kb mate-pair libraries by discordant read-pair analyses only.

Figure 4. **NA16595 Cri-du-Chat Deletion.** IGVscreenshot[28] of the detection of NA16595 Cri-du-Chat deletion on chromosome 5 by short-insert WGS at 1X, 3X, and 5X coverage. The vertical axis represents the magnitude of coverage, and blue dots represent respective coverage genomic position. Two areas within the deletion show unusually high coverage due to overlap with segmental duplications resulting in cross-mapping of sequencing reads.

Supplementary Figure S1. Size Distribution of NA12878 Silver Standard CNVs. Histogram of size distribution of NA12878 Silver Standard CNVs with hg19 reference gaps filtered out.

Supplementary Figure S2. Overlap with NA12878 GS CNVs from WGS Read-Depth and Discordant Read-Pair Analysis. Number of autosomal CNVs in the

NA12878 genome from short-insert, 3kb mate-pair, and 5kb mate-pair libraries sequenced at 1X, 3X, and 5X detected by **(a)** read-depth analysis only and **(b)** discordant read-pair analysis only. Gold: validated autosomal CNVs (overlap >50% reciprocally with NA12878 GS CNVs). Green: 10%-50% reciprocal overlap with NA12878 GS CNVs. Blue: <10% overlap with GS CNVs, >50% reciprocal overlap with NA12878 Silver Standard CNVs. Red: no overlap (<10% overlap with GS CNVs and <50% overlap with Silver Standard CNVs).

Supplementary Figure S3. **CNV Calls from WGS and Arrays without Overlap with NA12878 Gold and Silver Standard CNVs.** Ratios of autosomal CNV calls with no overlap with NA12878 Gold and Silver Standard CNVs from short-insert, 3kb mate-pair, and 5kb mate-pair WGS libraries sequenced at 1X, 3X, and 5X compared against those called from array platforms as previously described [23]. CNVs with <10% overlap with GS CNVs and <50% overlap with Silver Standard CNVs are defined as "no overlap" calls. Array calls were made according to platform specific algorithms, and WGS CNVs calls were made by combining discordant read pair and read-depth analysis. Red: autosomal CNVs without overlap with GS or Silver Standard CNVs. Gray: remaining proportion of total autosomal CNV calls.

Supplementary Figure S4. **Violin plots of CNV Size Distribution Across WGS and Array Platforms.** Violin plots of CNV sizes in the NA12878 genome called from short-insert, 3kb mate-pair, and 5kb mate-pair libraries sequenced at 1X, 3X, and 5X, and arrays with >150 CNV calls. **(a)** CNVs called by read-depth analysis only **(b)** CNVs called by discordant read-pair analyses only **(c)** CNVs called by combining discordant read pair and read-depth analysis **(d)** CNVs called by array platform-specific algorithms with >150 calls. Y-axis is log scaled.

Table 1. Detection of GS Deletions > 100 kb.

Supplementary Table S1. Summary of WGS CNV calls in the in the NA12878 genome and their overlaps with NA12878 Gold and Silver Standard CNVs

Supplementary Table S2. List of CNVs in the NA12878 genome called using short-insert, 3kb mate-pair, and 5kb mate-pair WGS at sequencing coverage depths 1X, 3X, and 5X by combining discordant read pair and read-depth analysis.

Supplementary Table S3. List of CNVs in the NA12878 genome called using short-insert, 3kb mate-pair, and 5kb mate-pair WGS at sequencing coverage depths 1X, 3X, and 5X by discordant read-pair analysis only.

Supplementary Table S4. List of CNVs in the NA12878 genome called using short-insert, 3kb mate-pair, and 5kb mate-pair WGS at sequencing coverage depths 1X, 3X, and 5X by read-depth analysis only.

Supplementary Table S5. Percent overlap between GS deletions > 100kb and those called by WGS.

Supplementary Table S6. List of CNVs in the NA16595 genome called Illumina MEGA array platform and by short-insert WGS at sequencing depths 1X, 3X, and 5X by read-depth analysis.

Table 1 Detection of Gold Standard deletions > 100 kb

Gold Standard Deletion	Size (bp)	WGS Detection	Array Detection (Haraksingh et al. 2017)
chr3:162,514,471-162,625,647	111,176	All WGS	All Agilent arrays[a]
chr4:69,375,591-69,491,543	115,952	All WGS	Affymetrix Cytoscan HD Affymetrix SNP 6.0 Agilent 2x400K CNV[a] Illumina HumanOmni1Quad-v1 Illumina HumanOmni2.5-8v1 Illumina HumanOmni25Exome-8v1 Illumina HumanOmni5-4v1 Illumina HumanOmni5Exome-v1
chr4:70,122,981-70,231,746	108,765	All WGS	Affymetrix SNP 6.0 Agilent 2x400K CNV[a] Illumina HumanOmni1Quad-v1 Illumina HumanOmni2.5-8v1 Illumina HumanOmni25Exome-8v1
chr6:78,892,808-79,053,430	160,622	All WGS	Affymetrix Cytoscan HD Affymetrix SNP 6.0 Agilent 2x400K CNV[a] Illumina HumanOmni1Quad-v1 Illumina HumanOmni2.5-8v1 Illumina HumanOmni25Exome-8v1 Illumina HumanOmni5-4v1 Illumina HumanOmni5Exome-v1 Illumina HumanOmniExpressExome 1.2
chr8:39,195,825-39,389,230	193,405	All WGS	Affymetrix Cytoscan HD Affymetrix SNP 6.0 Agilent 2x400K CNV[a] Illumina HumanOmni1Quad-v1 Illumina HumanOmni2.5-8v1

chr11:4,248,265-4,353,229	104,964	All WGS	Affymetric SNP6.0 Agilent 2X400K CNV Agilent 1X1M HR
chr19:20,595,835-20,717,950	122,115	All WGS	All arrays except Illumina Psych Array

[a]Called as high copy duplication

717

Figure 1 - CNV Calls by WGS and Comparisons with Arrays

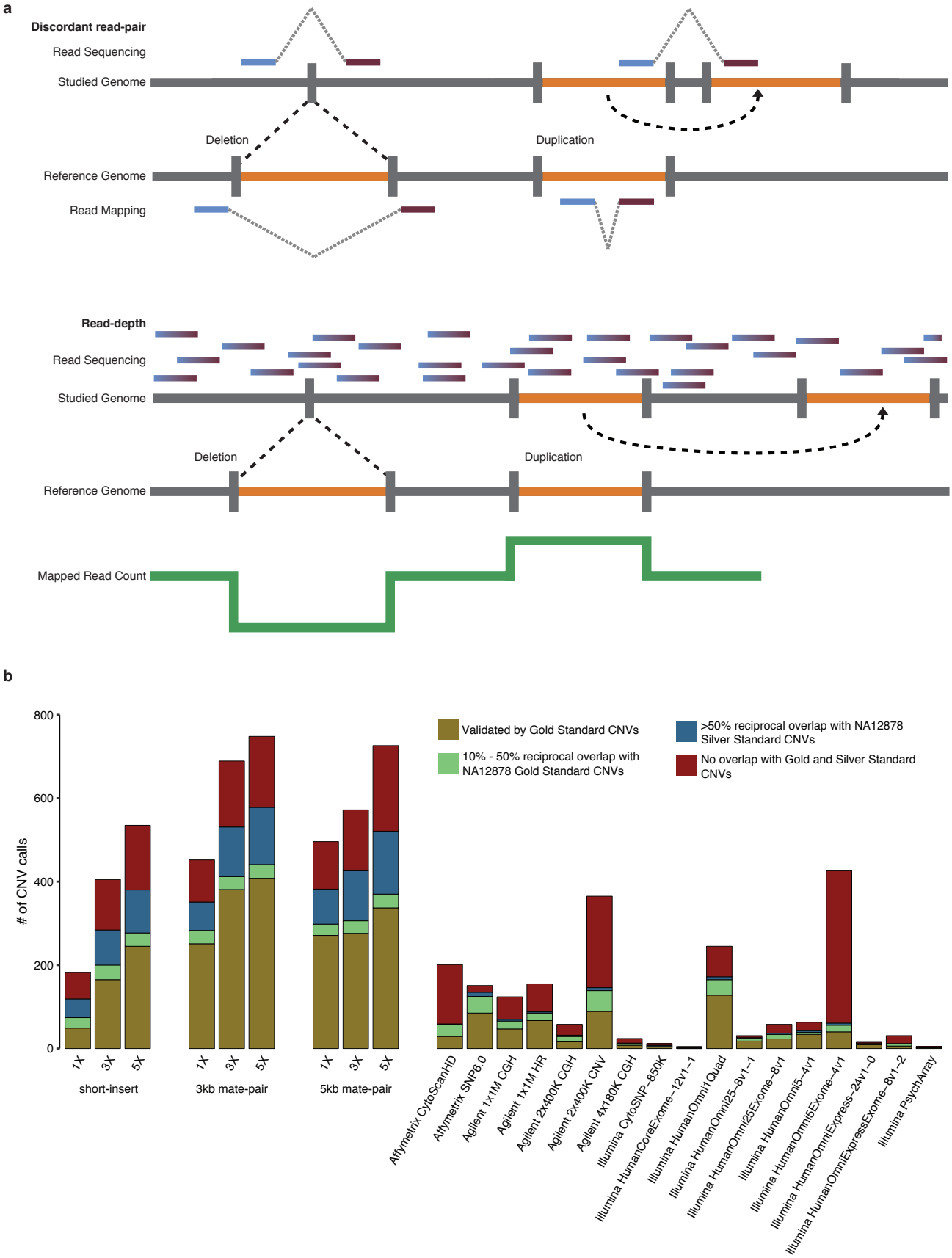


Figure 2 - Sensitivity of WGS Detection of NA12878 GS CNVs

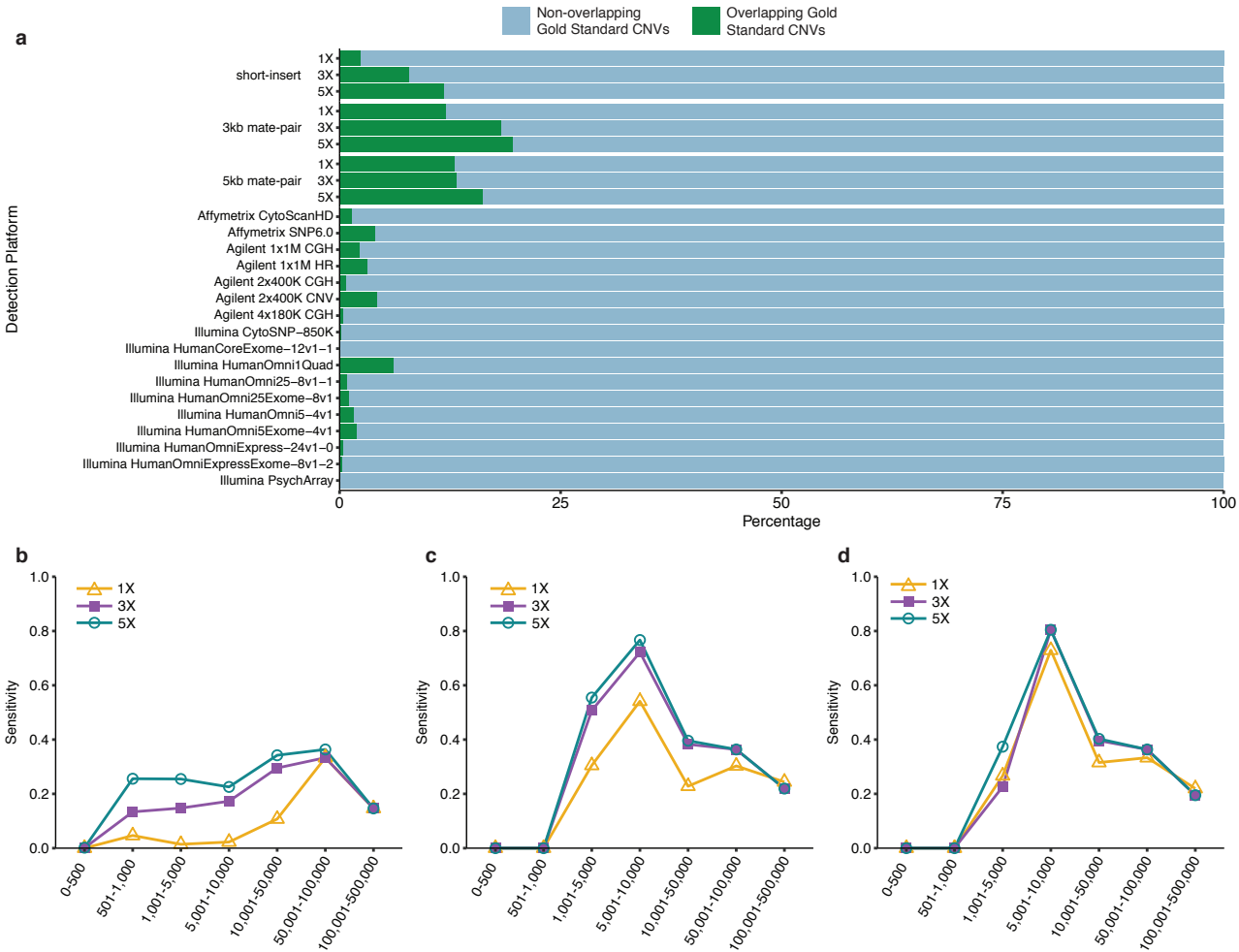


Figure 3 - Size Distributions of NA12878 Detected by WGS

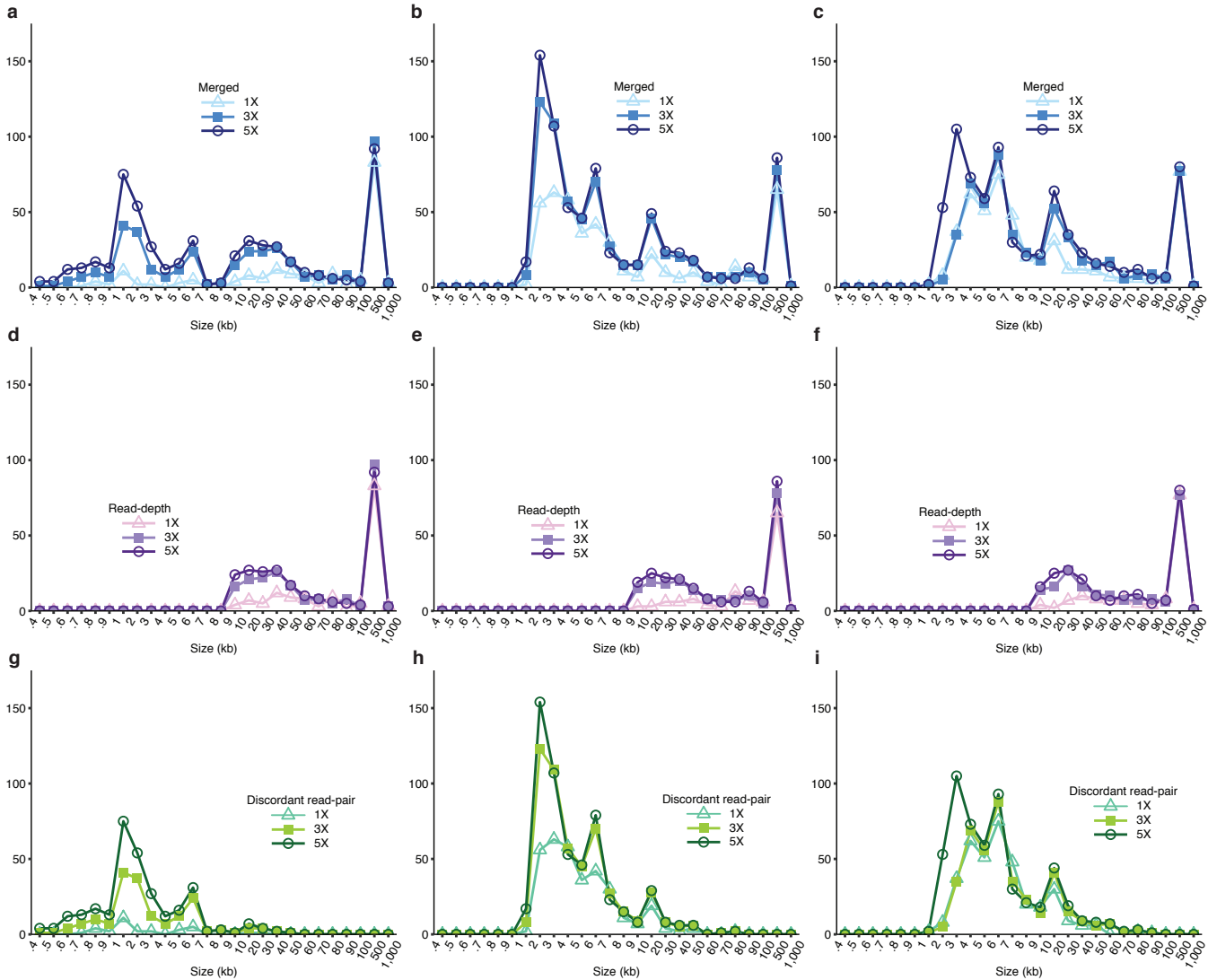
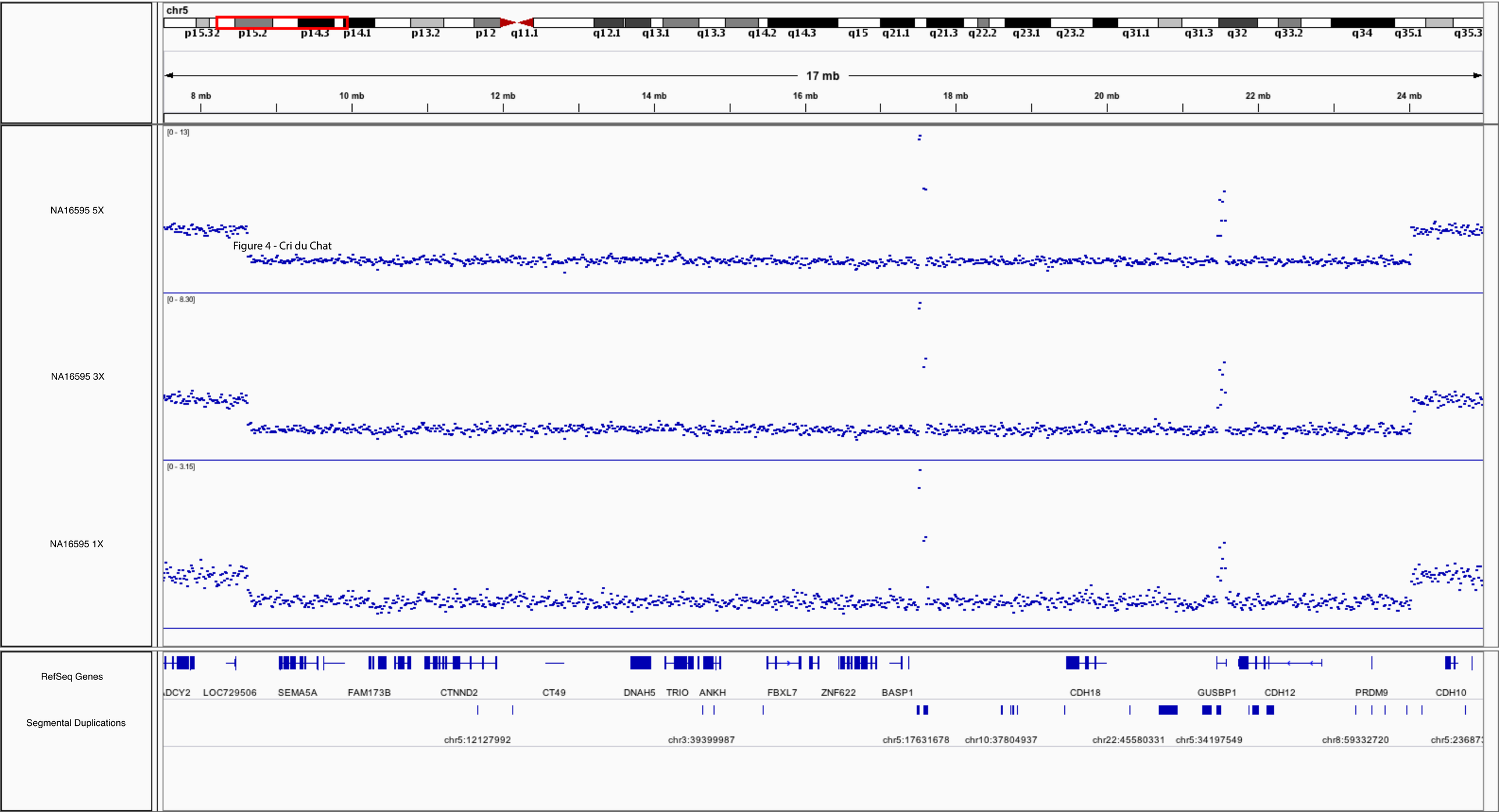
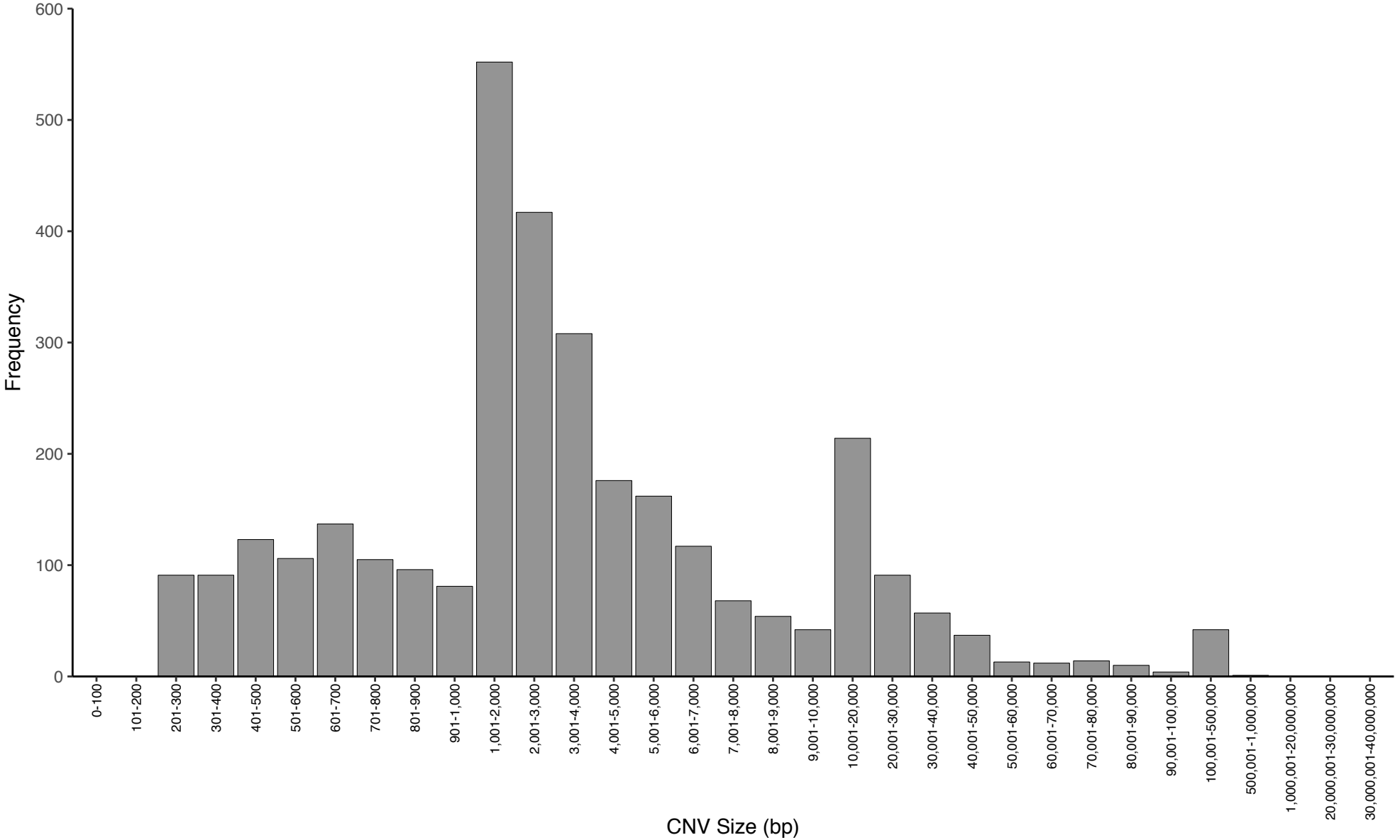


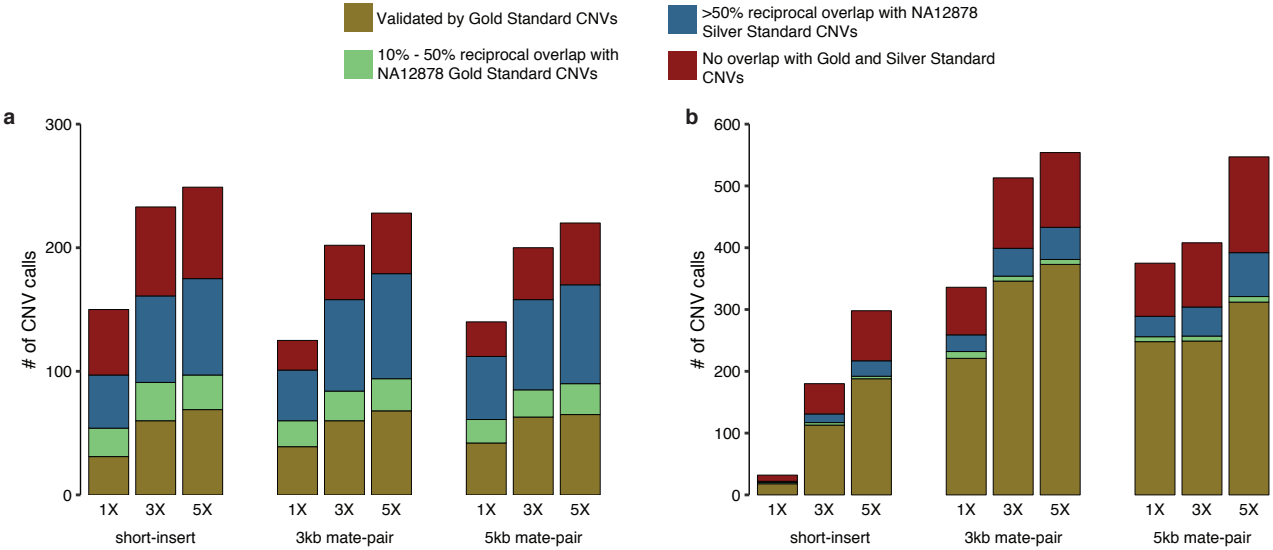
Figure 4 - NA16595 Cri-du-Chat Deletion



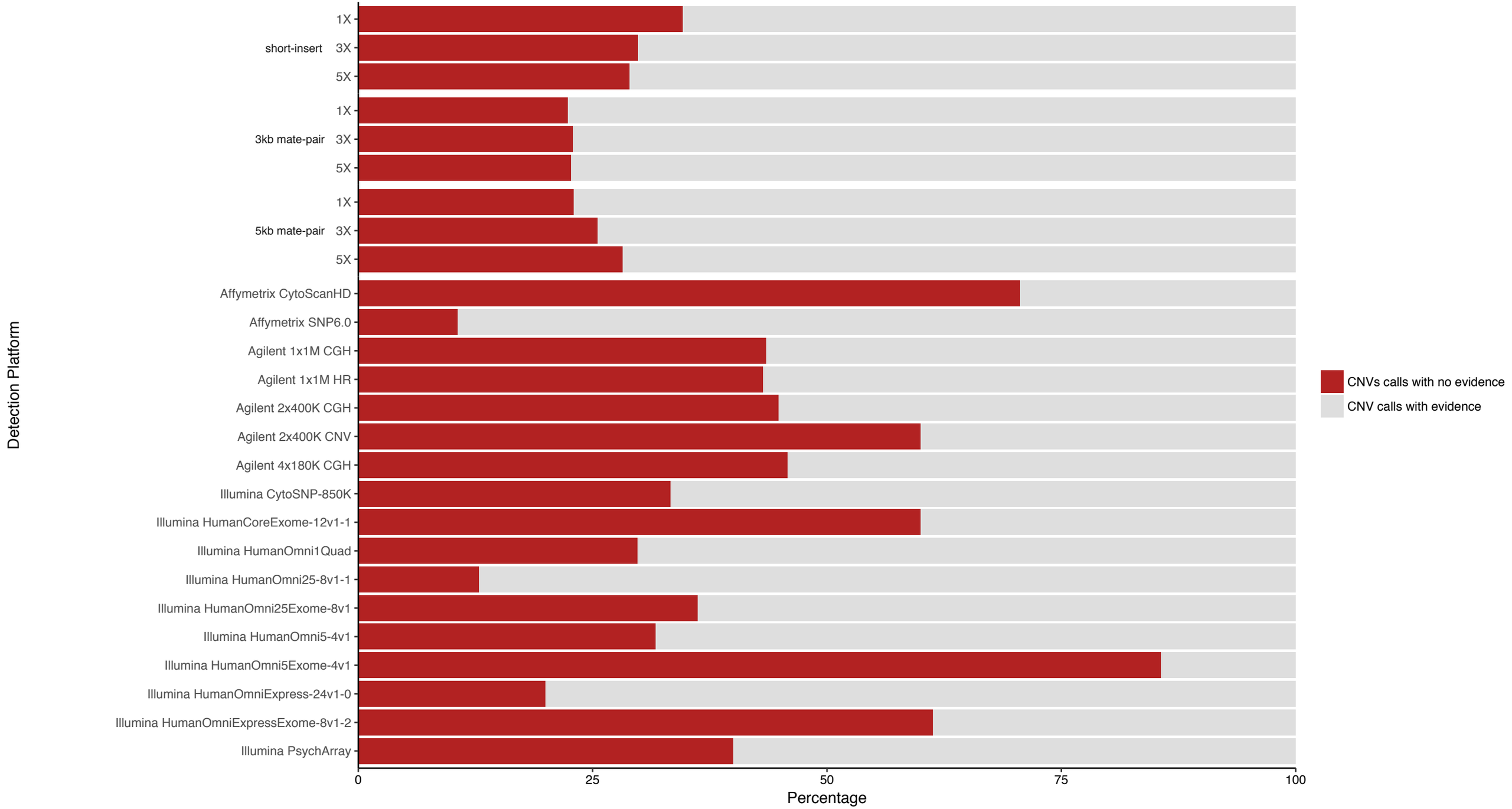
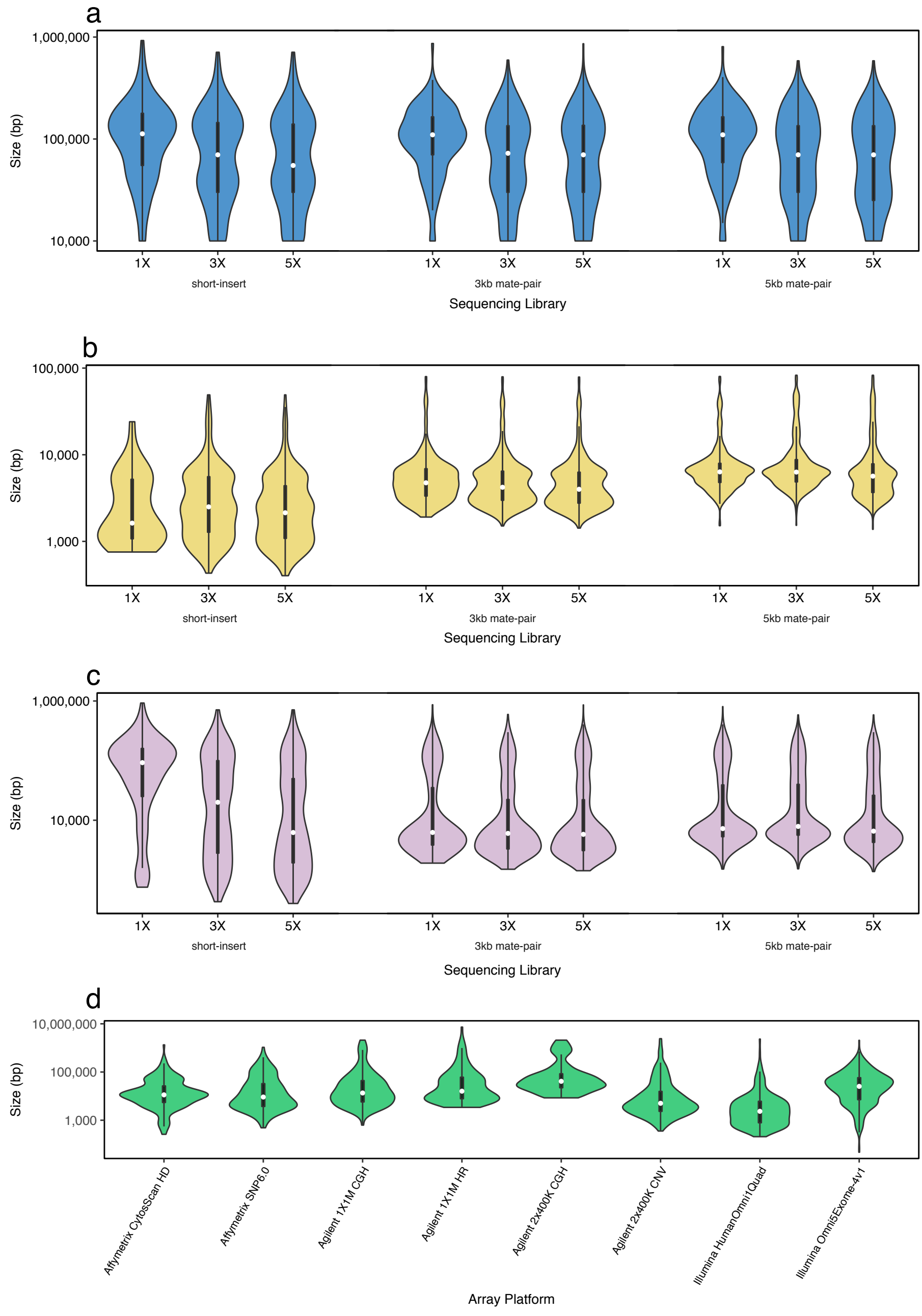
Supplementary Figure S1 - Size Distribution of NA12878 Silver Standard CNVs



Supplementary Figure S2 - Overlap with NA12878 GS CNVs from WGS Read-Depth and Discordant Read-Pair Analysis



Supplementary Figure S3 - CNV Calls from WGS and Arrays without Overlap with NA12878 Gold and Silver Standard CNVs



Supplementary Figure S4 - Violin plots of CNV Size Distribution Across WGS and Array Platforms