METHOD AND APPARATUS FOR DETECTING COPY NUMBER VARIATIONS IN A GENOME

Field

Replacing the current chromosomal microarray analysis (CMA) by a novel whole genome sequencing (WGS)-based assay

Abstract

While next-generation sequencing (NGS) has been considered as a standard technology for research applications across the life sciences, the conventional molecular cytogenetic methods, such as chromosomal microarray analysis (CMA) and fluorescent in situ hybridization (FISH) still remain the standard assays for detection of chromosomal aberrations at clinical laboratories. The major challenge of NGS as a first-tier assay is a lack of reliable bioinformatics tools. Some embodiments include an NGS-based copy number variation (CNV) calling algorithm, JAX-CNV, for detection of CNVs in the clinical setting. The performance of JAX-CNV was evaluated on ten samples from patients with constitutional disorders, which were examined in parallel by the clinically validated CMA assay at our CLIA-certified laboratory. In total, JAX-CNV identified all 54 CNVs reported by the CMA assay and 13 pathogenic CNVs that were previously detected by other clinical laboratories. The result demonstrated 100% concordance between JAX-CNV and the clinical CMA assay. Moreover, JAX-CNV detected additional 210 CNVs that were not captured by the CMA. To assess the false discovery rate of JAX-CNV, we selected 19 unique genomic regions from two samples for experimental validations using the droplet digital PCR (ddPCR). Four of the 19 regions were not validated and considered as false positive. Further analysis showed that all the four regions were located in segmental duplication or simple repeat regions. We also evaluated the robustness of the JAX-CNV pipeline by generating 30x, 20x, 15x, 10x, 9x, 8x, 6x and 4x coverage sequencing data and found 100% sensitivity of the detection at the coverage of 4x when >300Kb size cutoff was applied. This study indicates that NGS when paired with JAX-CNV can replace the current CMA as a first-tier clinical assay.

Introduction

Copy number variations (CNVs) have been suggested a key factor in human evolution (*1*, *2*), genomic diversity (*3*–*6*), and disease susceptibility (*7*–*9*). In addition, copy number gains (duplications) or losses (deletions) may cause microdeletion and microduplication syndromes as well as other genetic disorders, such as Williams syndrome (*10*), Prader-Willi syndrome (*11*), Angelman syndrome (*12*), Smith-Magenis syndrome (*13*), DiGeorge syndrome (*14*), and Pallister Killian syndrome (*15*). Numerous assays have been widely used in research and clinical laboratories for detection, including fluorescence in situ hybridization (FISH), PCR-based assays, chromosomal microarray analysis (CMA) and next-generation sequencing (NGS). In 2010, CMA was suggested by the American College of Medical Genetics and Genomics as a first-tier test for patients with unexplained developmental or intellectual disability, autism spectrum disorders, and congenital anomalies (*16*).

Over the past decade, the advancement of NGS technologies and computational tools has brought unprecedented advances in DNA sequencing throughput, speed, and cost. NGS allows whole genome sequencing (WGS) to be a possibility for researches in life sciences and healthcare. Recently, the NGS-based noninvasive prenatal testing (NIPT) of cell-free fetal DNA in the maternal blood has become a standard screening assay for chromosomal aneuploidy (*17*, *18*). A number of studies have examined the feasibilities and robustness of the WGS-based assays for discovering CNVs, and explored the possibility to replace the current CMA in the clinical laboratories (*19*–*24*). Although several WGS-based CNV calling algorithms (*25*–*31*) have been developed, none of them is widely accepted for pathogenic or high-risk CNV discrimination.

Rather than developing a new CNV calling algorithm, some researchers built pipelines using a combination of existing calling algorithms. Zhou *et al.* (*19*) combined calls from CNVnator (*29*) and LUMPY (*26*), and gave a conclusion that detecting copy number gains and losses on low-coverage NGS data outperforms array-based methods. Noll *et al.* proposed SKALD (*32*), which is based on consensus, filtered calls from BreakDancer (*33*) and GenomeSTRiP (*30*, *34*), while Trost *et al.* (*20*) employed CNVnator and ERDS (*31*) for CNV identification. However, none of those pipelines has been validated on multiple constitutional disease samples. Since those

pipelines consist of several calling algorithms and filters, replicating them can be difficult without a full release of pipelines. Moreover, using an ensemble of calling algorithms may improve accuracy, but it will add complications and thus take longer for making a diagnostic decision. Given the current limitations, it is not conceivable to meet the sensitivity, specificity, reproducibility, and speed requirements necessary for a true clinical grade pipeline without either a profound understanding of each algorithm used or the development of a brand new bioinformatics algorithm.

Here we present a newly developed NGS-based CNV algorithm, JAX-CNV, and the best practices of CNV detection for use on WGS data. We focus on large (>50Kb) deletions and duplications that are usually implicated to cause diseases and therefore, we can make a direct comparison to CMA. Ten Coriell samples associated with 13 pathogenic CNVs were selected. In addition to the pathogenic aberrations, CMA results in identifying 54 CNVs in the test samples. JAX-CNV successfully detects the 13 pathogenic CNVs and the 54 CNVs reported by microarray. Moreover, since a greater number of copy number gains and losses are detected by JAX-CNV, droplet digital PCR (ddPCR) is performed for the two selected samples for a comprehensive experimental validation. 4 out of the 19 regions were not validated false positive. Those false calls are all located in segmental duplications or simple repeats which are difficult genomic regions to resolve (*35*). In addition to the high sensitivity and specificity of the algorithm, JAX-CNV is light and fast for calling CNVs. In some embodiments it2takes less than an hour to complete analyses which, in turn, will accelerate diagnostic processing and lead to faster turnaround times. The exercise of this study shows the potential of WGS as a first-tier diagnostic assay and may replace CMA in clinic when paired with JAX-CNV.

Results

Development of a Clinical-Grade CNV Calling Algorithm

Currently, the microarray proficiency test offered by the College of American Pathologists (CAP) requests the participating clinical laboratories to report CNVs greater than 300Kb. This size-based reporting criteria guided the development of a clinical-grade CNV calling algorithm. We selected ten constitutional disease samples from the Coriell Institute (**Table 1**) which

contained at least one pathogenic CNV >300Kb for each sample. In total, there are 13 pathogenic CNVs (11 deletions and two duplications) ranging from 107.6Kb to 47.9Mb in size. Associated disorders are DiGeorge, Williams, Cri-du-chat, Smith-Magenis, Wolf-Hirschhorn, Miller-Dieker Lissencephaly, Tetralogy Fallot, 1p deletion, and Angelman syndromes. The 13 pathogenic CNVs set up the baseline of sensitivity analysis in the study.

In order to complete a comprehensive comparative study, the ten test samples were completed on both CMA and NGS technologies. Since Affymetrix CytoScan HD (CMA-based method; Affymetrix, Santa Clara, CA) is the current clinically validated microarray platform for the discovery of chromosomal aberrations at the Jackson Laboratory for Genomic Medicine (JAX-GM), all samples were processed following the clinical standard operating procedure of JAX-GM. The microarray analysis is completed using the vendor supplied software (**Material and Method: Affymetrix CytoScan HD analysis Flow**). We also completed WGS on the ten test samples by Illumina paired-end technology with read length 2x150bp and coverage ~45x (Supplementary Table S1). Short reads are mapped against GRCh38 human reference genome by BWA (*36*) followed by JAX-CNV for CNV identification (**Material and Method: NGS Analysis Flow**).

For eleven CNVs greater than 300Kb (CAP standard), both CMA and JAX-CNV identify them with 100% sensitivity (**Table 1**). For the other two small duplications (107.6Kb and 148.8Kb), CMA is unable to detect the 148.8Kb duplication at 22q11.21 due to the low resolution of the array caused by limited probe coverage. JAX-GM clinical microarray platform requests at least 50 array probes to ensure the high quality of CNV identification. Coordinates of the pathogenic, CMA, and JAX-CNV CNVs against GRCh38 human reference are given in Supplementary Tables S2-S5 while the plots of the calling regions are given in Supplementary files S1-S2. As a result, JAX-CNV proficiently identifies all 13 pathogenic chromosomal aberrations while JAX-GM clinical microarray platform leaves out one of them.

Sensitivities Assessment Compared to CMA-Based Method

Since Affymetrix CytoScan HD is a clinically validated platform at JAX-GM, all CNVs identified by the platform should also be recalled by JAX-CNV to show the potential of WGS as

a first-tier diagnostic assay. The CNV size cutoff of the JAX-GM clinical microarray platform is >50Kb. By this criterion, 61 CNVs are reported from the ten test samples including 12 pathogenic CNVs (Supplementary Table S4). Note that a pathogenic duplication at 22q11.21 is filtered due to low resolution of the array. Among the 61 CNVs, there are four deletions and three duplications with marginal qualities, and therefore ddPCR is necessary to resolve these aberrations. ddPCR assays for the targets are designed except for a 69Kb gain at 16p13 (chr16:14961449-15030399) due to complexity of the genomic region. The region is 99.9% identical with 16_KI270853v1_alt that makes design a unique primer impractical. The remaining aberrations (four deletions and two duplications) are confirmed falsely detected by CMA. The most interesting falsely identified CNV is the deletion at 6p25. The region is a common duplication region. The 1000 Genomes Project (*3*, *37*) which includes 2,504 samples shows 0.99 allele frequency of the duplication in the populations. Therefore, the deletion will be identified if a reference sample of CMA has this common duplication. After ddPCR validation, 54 CNVs including 39 deletions and 15 duplications remain (Supplementary Table S3).

JAX-CNV successfully identified all of 54 CNVs on WGS data (**Fig. 1**). The 50% reciprocal overlap is applied to evaluate CNV calls. Four deletions and two duplications do not overlap with CMA calls by 50% reciprocally, but they are in the regions either smaller or larger. The result suggests that NGS-based assay of the use of JAX-CNV is at least as accurate as CMA-based assay in sensitivity.

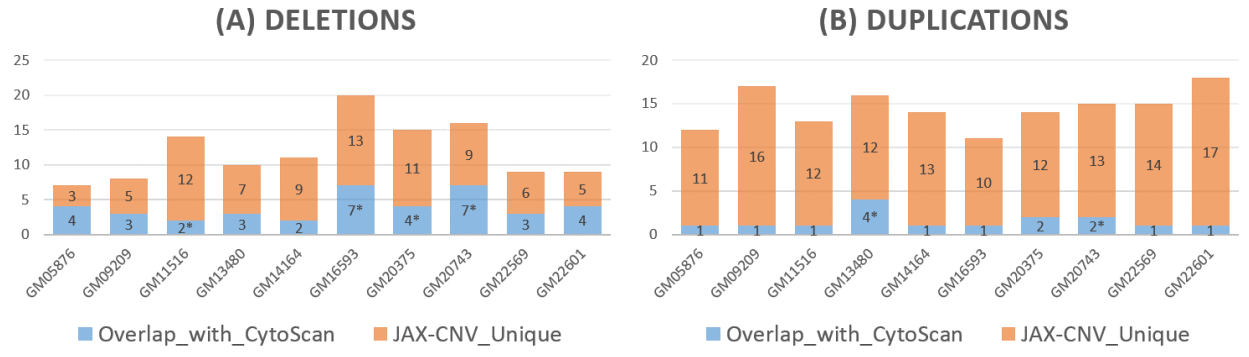


Fig. 1: The sensitivity assessment of JAX-CNV compared to calls reported by JAX-GM clinical microarray platform (CytoScan). * denotes that JAX-CNV identifies a CNV that not 50% reciprocal overlapping with CytoScan reported one, but recovered in manual review.

NGS False Discovery Rate Assessment

Fig. 1 not only shows that JAX-CNV identifies the same calls as CMA, but also indicates that JAX-CNV detects greater numbers of aberrations than microarray does. JAX-CNV detects an additional 210 aberrations (Supplementary Table S4), which results in almost a five-fold increase compared to CMA calls. To evaluate the accuracy of those additional aberrations, we selected 19 unique genomic regions from two samples, GM05876 and GM09209, for experimental validation (**Table 2**). A 224.9Kb loss at 21p11 of GM05876 is not conclusive in the ddPCR validation (**Material and Method: Droplet Digital PCR (ddPCR) Validation**) due to an unclear separation of positive and negative droplet clusters. 4 others, 14 are confirmed true and four are false. The four false calls are all duplications and located at either segmental duplication or simple repeat regions which are difficult genomic regions to resolve. Two duplications are even in the same cytoband, 16p11, which is disturbed by a large segmental duplication. As a result, 4 out of 19 CNVs for GM05876 (5 overlapping with CMA, 1 ddPCR primer undesignable, 1 ddPCR inconclusive, 8 ddPCR confirmed true, and 4 ddPCR confirmed false) and 4 out of 25 CNVs for GM09209 (5 overlapping with CMA, 1 ddPCR primer undesignable, 5 unreproducible on GRCh19, 10 ddPCR confirmed true, and 4 ddPCR confirmed false) are validated as false.

It is noteworthy that CMA cannot identify those 14 CNVs (7 deletions and 12 duplications) that are called by JAX-CNV and are validated true by ddPCR. Thus, among 13 and 15 validated true calls of GM05876 and GM09209, CMA only identifies 5 for each of them. The missing rates of CMA are then 61.5% and 66.7% for GM05876 and GM09209, respectively. Unlike WGS comprehensively sequences whole genomes, CMA relies on array probes for CNV detection. In other words, an aberration cannot be identified if there are not sufficient probes in the region, such as the pathogenic duplication at 22q11.21. For those unique CNVs detected on WGS data, they all lack of sufficient array probes for CMA to call them confidently (JAX-GM clinical microarray platform requests 50 consecutive probes for reporting an aberration).

CNV Detection on Low-Coverage WGS

Although, NGS cost drops rapidly, the price is still a concern as WGS being the first-tier assay in clinic. To tackle this issue, we down-sample WGS data and evaluated the sensitivities of different coverages. The samples are originally sequenced by the coverages ranging from 42x to 46x. The simulation of different coverages is done by SAMBAMBA (*38*) on the aligned BAM files. Based on the original coverage of 42-46x, we generated 30x, 20x, 15x, 10x, 9x, 8x, 6x and 4x WGS data. JAX-CNV is then applied on the different coverage data for CNV identification.

JAX-CNV is able to get reproducible results as low as 20x; that is a 50% of the original coverage (**Table 1**). At the lower coverage of 15x, JAX-CNV cannot identify a 148.8Kb duplication at 22q11.21 of GM14164, which is also the same one that CMA cannot detect. At the coverage of 10x, JAX-CNV cannot resolve the second 107.6Kb duplication at 9p24.1 of GM13480. Duplication detection is more sensitive to sequencing coverage than deletion. For deletions, even at the coverage of 4x, JAX-CNV still consummately identifies all pathogenic ones. However, low coverage leads to more noise and affects the quality of CNV identification. Two deletions on 17p11.2 (GM20743) and 4p16.3 (GM22601) are divided into a few small pieces rather than complete CNVs (Supplementary File S1). The other impact of low coverage is that more CNVs are detected which suggests that more false positive calls may be made. In summary, the sensitivity and specificity are maintained when down-sampling to 20x. For the widely accepted CAP standard, greater than 300Kb CNVs, JAX-CNV still gets 100% sensitivity on the coverage of 4x which leads lower cost for diagnosis.

To better understand the effect of sequencing coverages, we extended the comparison of using calls from JAX-GM clinical microarray platform. There are 54 CNVs (39 deletions and 15 duplications). 100% sensitivity is reproducible at the coverage of 20x (**Fig. 2**). However, as the coverage decreases the sensitivity also decreases and more CNVs do not 50% reciprocally overlap with CMA identified CNVs. At the coverage of 8x, JAX-CNV loses ability to detect one deletion and at the coverage of 4x, the algorithm misses six deletions (**Fig. 2A**). For duplications, at the coverages of 15x, 10x, 9x, 8x, 6x, and 4x, JAX-CNV misses two, nine, ten, eight, twelve, and eleven calls, respectively (**Fig. 2B**). The missing duplications at the coverage of 8x are fewer than 9x, but this finding does not suggest that 8x is better than 9x. On the contrary, it indicates

that the read depth signal for discovering duplications on low coverage samples is unstable. As a result, the sensitivities (for all deletions and duplications) decrease from 100% (42-46x, 30x and 20x), 96% (15x), 83% (10x-8x), 77% (6x) to 68% (4x). If we classify CNVs into deletions and duplications, sensitivity of deletions is much more prominent than duplications. Even down to 4X, the sensitivity of deletion detection remains at 85% while the sensitivity of duplication detection drops to 27%. This observation is consistent with the result in pathogenic CNV identification. The sensitivity of duplication detection is worse than the sensitivity of deletion detection.

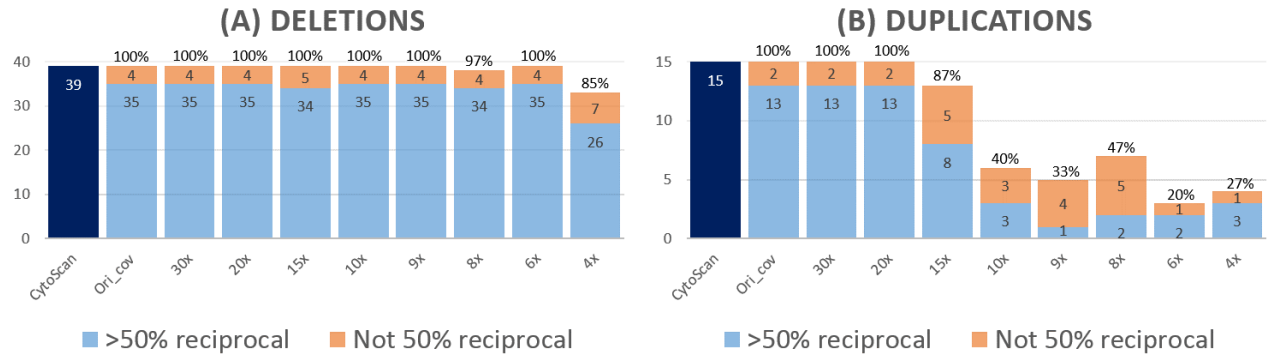


Fig. 2: CNV detection on nine different coverage WGS. Ori_cov is the original sequenced coverage ranging from 42x-46x for the ten test samples. SAMBAMBA is then applied to down-sample WGS from coverages 30x to 4x. The 100% sensitivity of calling CMA identified CNVs is reproducible at the coverages of 30x and 20x. Each bar is separated by >50% and not 50% reciprocal overlapping with CMA identified CNVs.

Discussion

Comparison with Other WGS-Based Algorithms

Over the past decade, several well-studied WGS-based CNV calling algorithms were invented for research purposes, such as Manta (*25*), Lumpy (*26*), Delly (*27*), CNVnator (*29*) and cn.MOPS (*28*). In addition, there are also combined methods, such as FusorSV (*39*) and MetaSV (*40*). We did not perform comparisons to FusorSV and MetaSV since they do not support for GRCh38 human reference genome. None of those methods was originally designed for a clinical

setting, which requires extraordinary sensitivity and specificity. For research purposes, sacrificing specificity for the maintenance of sensitivity is expedient as to not miss any potential CNVs. Nevertheless, reporting hundreds, or even thousands, of chromosomal aberrations is impractical for diagnosis due to the pressure of turnaround time in the clinic. Thus without careful clinical considerations, the sensitivity and specificity of those tools cannot meet the requirements for disease diagnosis.

For the 13 pathogenic CNVs used as a baseline for the current study, calling algorithms Manta, Lumpy, Delly, CNVnator and cn.MOPS identify 6, 6, 10, 13 and 4, respectively (Supplementary Table S7). It can be seen that read-depth-based algorithms, such as JAX-CNV and CNVnator, have relative high sensitivities. Other algorithms that use paired-end, split-read or combinations clearly show the weakness of low sensitivity of identifying those pathogenic chromosomal aberrations. For clinical applications, we expect that all reported variations are pathologically meaningful and a reasonable number of reported CNVs. JAX-CNV performs well in this respect. In contrast, Manta, Lumpy and Delly identify more than tens of thousands CNVs, which bring difficulties of decision making if using them for pathological diagnosis. The assessment of existing algorithms addresses an issue of using research-purpose algorithms in the clinic and suggests the necessity of tailored CNV algorithms.

Robustness Test

A clinical-grade pipeline not only requires high sensitivity and specificity, but also exceptional stableness and robustness. To perform a robustness test, we applied JAX-CNV on 934 samples from the Taiwan BioBank (TWB. https://www.twbiobank.org.tw/new_web_en/index.php). The read length is Illumina 2x151bp and coverage is ~30x. JAX-CNV is light and easy to run. For each sample, JAX-CNV requests 4.5G memory and can finish CNV detection in one hour without any program failures. The stableness and robustness of the algorithm are thus shown on those 934 samples.

For each sample, we also have data from the Affymetrix customized TWB genotype 653K array for comparative analyses. Due to the lack of nonpolymorphic copy-number probes in the TWB genotype array, its capacity of CNV detection is less sensitive than JAX-GM clinical microarray

platform. Thus, from TWB genotype array, we selected high confident CNVs that have more than 200 array probes and analyzed segments larger than 50Kb (Supplementary Table S8). JAX-CNV recalls 94.7% deletions and 83.9% duplications. Note that due to the lack of nonpolymorphic copy-number probes, there may be some false positive CNVs reported by the TWB genotype array, which will reduce the overlap with the result from NGS-based calling results. However, the high stableness, robustness, and efficiency of JAX-CNV can be seen from this experiment.

Conclusions and Ramifications

Chromosomal abnormalities involved in pathogenic diagnosis are not limited to CNVs, but other structural variants (SVs) as well including translocations and inversions. To identify those pathogenic translocations and inversions, we will develop new modules in the current pipeline. Detecting translocations and inversions is more difficulty than detecting CNVs. We cannot rely solely on read depth signal, which is the major signal we used for our current CNV caller. The paired-end alignment distance and orientation will be considered for identifying translocations and inversions. Breakpoints of inversions have been shown to likely be associated with deletions, which increases the difficulty to detect them. However, Dong *et al.* (*41*) show the potential to detect them on WGS data. Thus, we believe with a careful design and our knowledge of SVs, our pipeline will become a comprehensive SV caller for clinical applications.

We have developed a NGS-based CNV caller, JAX-CNV, which shows the potential of NGS replacing CMA as a first-tier diagnostic assay. The assessment on the ten constitutional disease samples shows 100% sensitivity that outperforms other calling algorithms. Besides, JAX-CNV is easy to run and so fast that it can complete calling CNVs for a 30x coverage sequenced sample in one hour. JAX-CNV meets the sensitivity, specificity, reproducibility, and speed requirements necessary in the clinic, and shows its potential to replace CMA-based methods as a first-tier diagnostic assay.

Material and Methods

Affymetrix CytoScan HD Analysis Flow

CNV microarray analysis was performed by the cytogenetics laboratory at JAX-GM, using the Affymetrix Cytoscan HD platform. The array consists of 2,696,550 probes that include 743,304 SNP probes and 1,953,246 nonpolymorphic copy-number probes. The average probe spacing for RefSeq genes is 880 bp, and 96% of genes are represented. DNA labeling, slide hybridization, washing, and scanning were performed following the manufacturer's protocol. CEL files were generated from scanned array image files by Affymetrix GeneChip Command Console software and were imported into Affymetrix Chromosome Analysis Suite (ChAS v3.3) software. Copy number data files (CYCHP files) were generated using Affymetrix CytoScan HD Array version NA36 (hg38) as a reference. Data were analyzed using the following filtering criteria: greater than 50Kb with a minimum of 50 consecutive markers.

NGS Analysis Flow

Pre-process

Fig. 3A shows the complete flow of WGS data analysis. The pre-process steps for a reference genome, such as GRCh19 or GRCh38, include BWA (*36*) index (v0.7.15) and JELLYFISH (*42*) count (v2.2.6). BWA index creates required files for BWA alignment that is the next step while JELLYFISH count calculates the counts of each 25-mer genome widely and generates k-mer DB. One of our developed modules takes k-mer DB and converts it to a FASTA-format log2 (25-mer) file. For example, if a 25-mer has only one position in the genome, the log2(1) is zero. For converting the zero to an ASCII code, we add 34 to make an ASCII code " (34 in decimal) represent zero. BWA index, JELLYFISH count, and k-mer DB converting may take 190, 105, and 403 minutes, respectively.

Alignment

Once the paired-end FASTQ files of a sample are received, FASTQC (v0.11.5) and BWA mem (v0.7.15) are applied for quality control and alignment either against GRCh19 or GRCh38. The

BAM is then generated by BWA as the alignment result is followed by SAMTOOLS (*43*) to sort alignments by coordinates. The resultant sorted BAM is the input file of JAX-CNV.

CNV calling

The first step of CNV calling is coverage calculation. JAX-CNV uses the log2(25-mer) FASTA-format file from the pre-process steps to scan the unique genome regions for each autosomal chromosome. A region is considered a unique genome region when each 25-mer count is one and the size of the region is larger than 20Kb. For each chromosome, we calculated a coverage based on 20 unique genome regions in the chromosome. We then applied the interquartile range to filter outlier coverages and calculated an overall coverage of the sequenced sample. Comparing the coverage of each chromosome with the overall coverage, we are able to detect aneuploidies. For those aneuploidies, we will not detect any smaller CNVs on the respective chromosomes in the further steps.

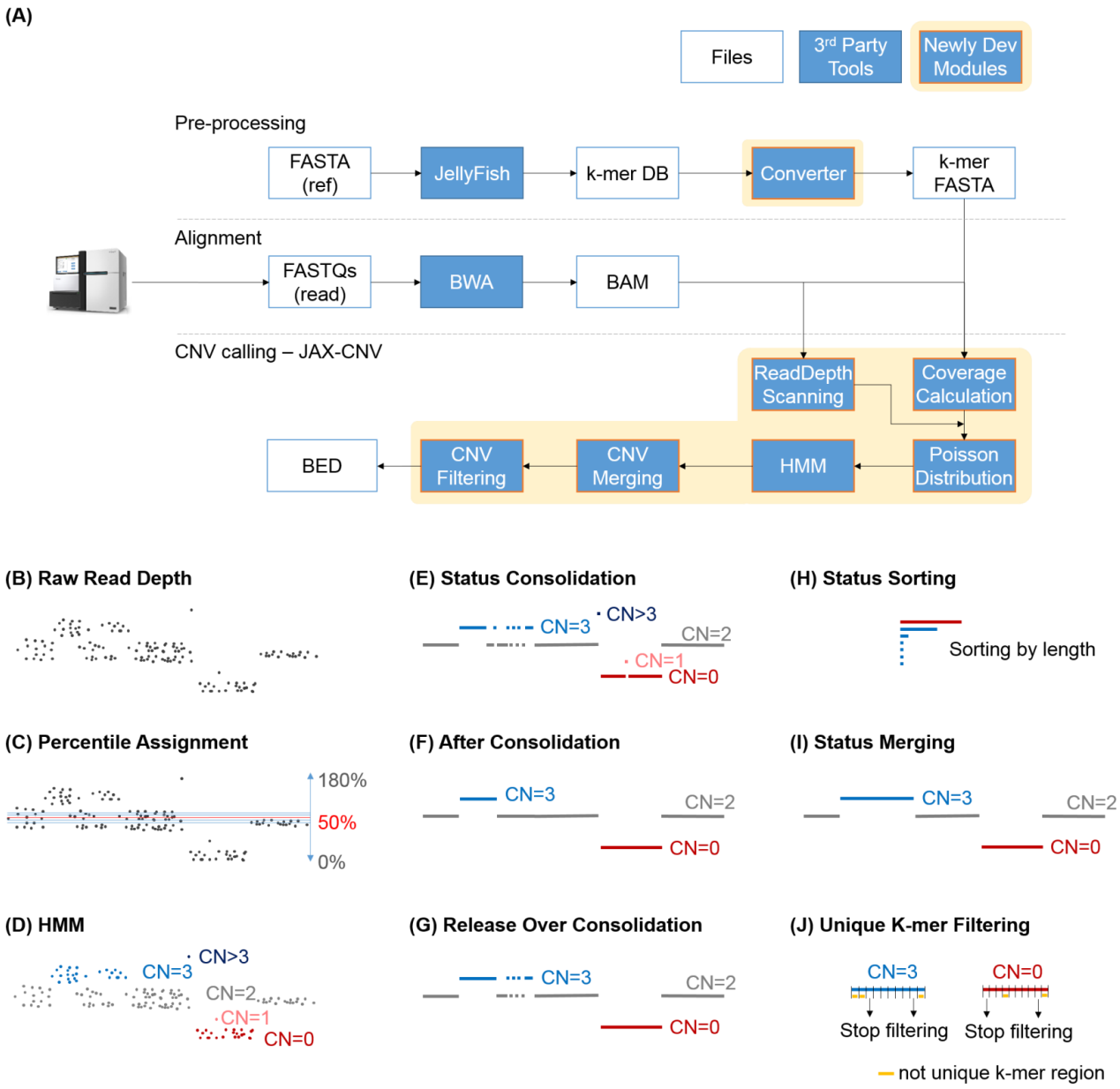
Once the overall coverage is calculated, we then scanned the BAM file by shifting bins (default size of a bin is 50bp and is user adjustable by --bin) for read depth calculation (**Fig. 3B**). According to the overall coverage as the baseline (at 50% percentile), the read depth of each bin can be translated to a percentile, from 0% to 180%. For example, if the overall coverage is 50 and a read depth of a bin is 100, the percentage tile of a bin will be 100% (100 / 50 * 50%; **Fig. 3C**). Then, a hidden Markov model (HMM) with a Poisson distribution of read depth is applied to convert the percentile of each bin to one of the five CNV statuses: CN=0 (loss), CN=1 (loss), CN=2 (normal), CN=3 (gain) and CN>3 (gain) (**Fig. 3D**). Afterwards, if the CNV statuses of two adjacent bins, we merge the bins (**Fig. 3E**).

Since the default bin size is set to 50bp, frequent CNV status oscillation will happen as shown in **Fig. 3E**. Using larger bin sizes, we can resolve the oscillation, but this change decreases the sensitivity as well. To maintain high sensitivity, we suggest 50bp as the bin size. Therefore, a merge step is then necessary to mitigate status oscillation. If the status' length is shorter than 5Kb then it will be absorbed by the previous status (**Fig. 3F**). This observation shows that the resolution of JAX-CNV is 5Kb. The status consolidation may be so aggressive that a region consists of too many other statuses. To resolve this, if the original status of the region covers less

than 80% in length, the merging will stop and reinstate the original statuses (**Fig. 3G**). After recognition of a complex region and the cease of merging, the CNV classifications are then sorted by their respective lengths (**Fig. 3H**). From the longest to the shortest, each CNV status will scan other statuses downstream and upstream by coordinates for further merging (**Fig. 3I**). This step allows larger CNVs to cross normal status and to merge smaller CNVs nearby. A larger CNV has more ability to cross normal statuses and merge others. Candidate CNVs are then generated.

For each candidate CNV, we divided it into ten regions of equal length. Each region is assigned a uniqueness value corresponding to the count of unique k-mer. Starting with the first region, we filtered it if the uniqueness value is low (percentage of unique k-mers is lower than 60% by default; user adjustable by --unique_kmer). Once we cannot filter, we stopped (**Fig. 3J**). The same procedure was performed from the last region to the first one as well. We reported CNVs in a BED-format file when the remaining regions are larger than 45Kb.

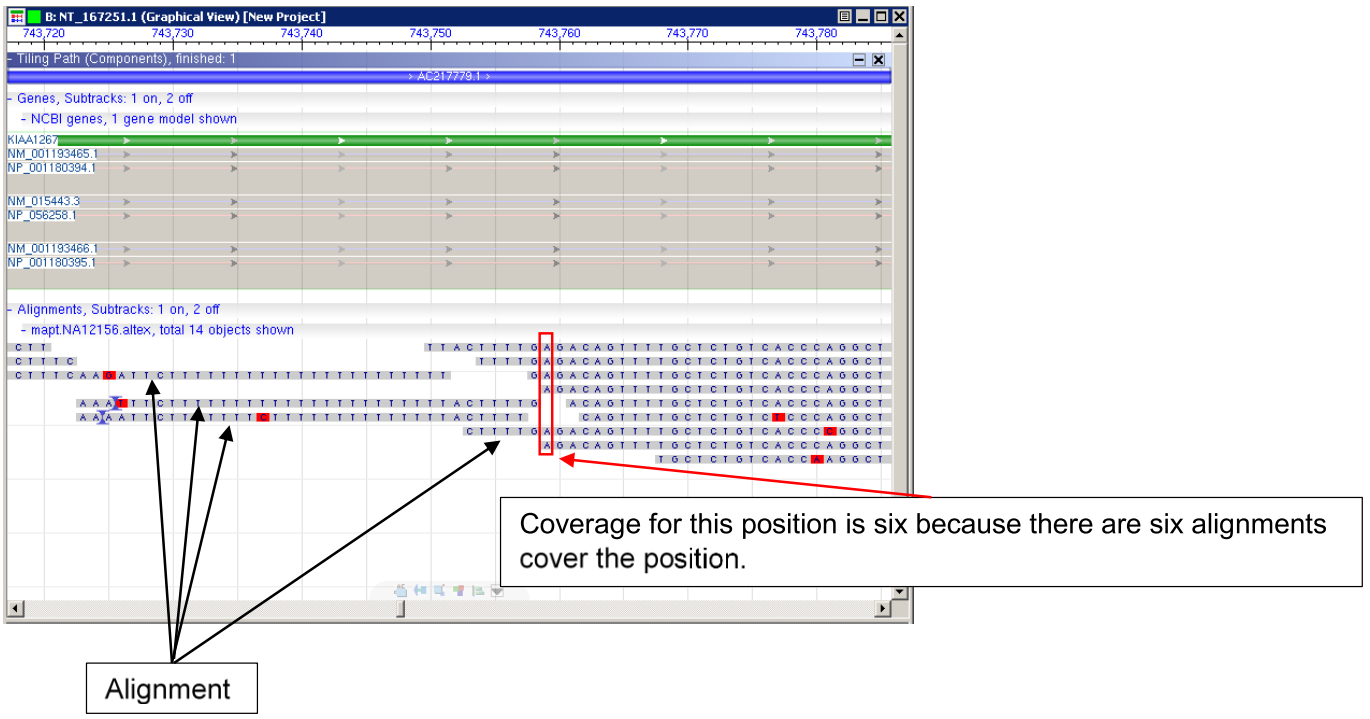
Fig. 3: (A) The flow of CNV detection comprises of three major steps, pre-processing, alignment and CNV calling. (B-J) The details of Poisson distribution, hidden Markov model (HMM), CNV merging and CNV filtering.



BAM

A BAM file (.bam) is the binary version of a SAM file. A SAM file (.sam) is a tab-delimited text file that contains sequence alignment data. These formats are described on the SAM Tools web site: http://samtools.github.io/hts-specs/.

Alignments look like the view below.



Coverage

The coverage is for each position in genome, how many alignments cover it.

Repetitive Region

As researchers estimate, there are approximate 50% of genomic regions are repetitive ones. Repetitive regions have similar sequences to other regions. For those regions, alignments may be not so confident since we will have several regions to place alignments and we don't know which one is correct. The k-mer FASTA is the file to tell how repetitive (or on the contrary how unique) a position in genome is.

Converter

Input: k-mer DB generated by JellyFish

Output: k-mer FASTA

The input is k-mer DB generated by JellyFish (k is 25 in our pipeline). K-mer DB is an encrypted lookup table that denotes how many times a sequence (with 25 letters/basepairs/nucleotides) appears. So, then for the first 25 basepairs in FASTA (ref), the lookup table illustrates how many times this sequence appears in the genome. Once we finish the first 25 basepairs, we move to the second 25 basepairs and repeat the actions until the end of the genome.

For example, if a 25-mer has only one position in the genome, the log2(1) is zero. For converting the zero to an ASCII code, we add 34 to make an ASCII code " (34 in decimal) represent zero.

" is kept in the output.

FASTA format: https://en.wikipedia.org/wiki/FASTA_format

ASCII code: http://www.pld.ttu.ee/~marek/PA_R4/ascii.html

Coverage Calculation

Input: k-mer FASTA and BAM

Output: a vector of coverages for all chromosome and an overall coverage

As we described in "what is repetitive region", we like to use unique (not repetitive) regions to calculate the coverage. So, one of the inputs of this module is k-mer FASTA generated by "Converter" to find unique regions.

A region is considered a unique genome region when each 25-mer count is one and the size of the region is larger than 20Kb. For each chromosome, we calculated a coverage based on 20 unique genome regions in the chromosome.

Once we identify the unique region, we check the alignments for coverage calculation so we need BAM as input as well. Figure 3B is showing an example of coverage in BAM.

We then applied the interquartile range to filter outlier coverages and calculated an overall coverage of the sequenced sample. Comparing the coverage of each chromosome with the

overall coverage, we are able to detect aneuploidies. For those aneuploidies, we will not detect any smaller CNVs on the respective chromosomes in the further steps.

Notice that the coverage calculation is done for each chromosome so this allows us to detect whole chromosome duplications or deletions (aneuploidies).

interquartile range: https://en.wikipedia.org/wiki/Interquartile_range

ReadDepthScanning

Input: BAM

Output: a vector of read depth (coverage) of each bin

Once the overall coverage is calculated, we then scanned the BAM file by shifting bins (default size of a bin is 50bp and is user adjustable by --bin) for read depth calculation (**Fig. 3B**).

Poisson Distribution

Input: a vector of read depth (coverage) of each bin and an overall coverage (as the baseline)

Output: a vector of percentile of read depth of each bin

According to the overall coverage as the baseline (at 50% percentile), the read depth of each bin can be translated to a percentile, from 0% to 180%. For example, if the overall coverage is 50 and a read depth of a bin is 100, the percentage tile of a bin will be 100% (100 / 50 * 50%; **Fig. 3C**).

HMM (Hidden Markov Model)

Input: a vector of percentile of read depth of each bin

Output: a vector of CNV status with regions sizes (sorted by coordinates)

Then, a hidden Markov model (HMM) with a Poisson distribution of read depth is applied to convert the percentile of each bin to one of the five CNV statuses: CN=0 (loss), CN=1 (loss), CN=2 (normal), CN=3 (gain) and CN>3 (gain) (**Fig. 3D**).

Now, for each bin (50bp) we assign a CNV status to it. If the adjacent bin assigned the same statuses, then we merge them together. Please notice that the regions are sorted by coordinates naturedly since we always process from the beginning to the end of the genome.

Afterwards, if the CNV statuses of two adjacent bins, we merge the bins (**Fig. 3E**).

CNV Merging

Input: a vector of CNV status with regions sizes (sorted by coordinates)

Output: a vector of CNV candidate regions

Since the default bin size is set to 50bp, frequent CNV status oscillation will happen as shown in **Fig. 3E**. Using larger bin sizes, we can resolve the oscillation, but this change decreases the sensitivity as well. To maintain high sensitivity, we suggest 50bp as the bin size. Therefore, a merge step is then necessary to mitigate status oscillation. If the status' length is shorter than 5Kb then it will be absorbed by the previous status (**Fig. 3F**). This observation shows that the resolution of JAX-CNV is 5Kb. The status consolidation may be so aggressive that a region consists of too many other statuses. To resolve this, if the original status of the region covers less than 80% in length, the merging will stop and reinstate the original statuses (**Fig. 3G**). After recognition of a complex region and the cease of merging, the CNV classifications are then sorted by their respective lengths (**Fig. 3H**). From the longest to the shortest, each CNV status will scan other statuses downstream and upstream by coordinates for further merging (**Fig. 3I**). This step allows larger CNVs to cross normal status and to merge smaller CNVs nearby. A larger CNV has more ability to cross normal statuses and merge others. Candidate CNVs are then generated.

CNV Filtering

Input: a vector of CNV candidate regions and k-mer FASTA

Output: a vector of CNV regions (final result)

For each candidate CNV, we divided it into ten regions of equal length. Each region is assigned a uniqueness value corresponding to the count of unique k-mer.

The k-mer information is from k-mer FASTA.

Starting with the first region, we filtered it if the uniqueness value is low (percentage of unique k-mers is lower than 60% by default; user adjustable by --unique_kmer). Once we cannot filter, we stopped (**Fig. 3J**). The same procedure was performed from the last region to the first one as well. We reported CNVs in a BED-format file when the remaining regions are larger than 45Kb.

BED format: https://useast.ensembl.org/info/website/upload/bed.html

Droplet Digital PCR (ddPCR) Validation

Droplet Digital PCR (ddPCR) assays were performed to examine the accuracy of the genomic aberrations detected by the JAX-GM clinical microarray platform and JAX-CNV. The customized assays were designed using Primer3Plus (*44*) based on hg38 assembly. All primer pairs were tested for their uniqueness across the human genome using In-Silico PCR from UCSC Genome Browser. A BLAT (*45*) search was also performed at the same time to make sure all primer candidates have only one hit in the human genome. Lastly, the NCBI 1000 Genome Browser was used to check if there were any SNPs in the primer or probe-binding region. All primers and probes used in this study were listed in Supplementary Table S6.

The ddPCR reactions were created following the Bio-Rad QX200™ system manufacturer protocol. A total of 10ng DNA template was mixed with a 2X ddPCR SuperMix for Probes (no dUTP), *Hind*III-HF enzyme (2U/reaction) (New England BioLabs, MA, USA), 20X primer/probe, (both FAM and HEX-labeled probes) and water to a final volume of 20 µL. Each reaction mixture was then loaded into the sample well of an eight-channel droplet generator cartridge. A volume of 70 µl of droplet generation oil was loaded into the oil well for each channel and covered with a gasket. The cartridge was placed into the Bio-Rad QX200™ Droplet Generator. After the droplets were generated in the droplet well, 40 µl was transferred into a 96-well PCR plate and then heat-sealed with a foil seal. PCR amplification was performed using a C1000 Touch thermal cycler with the following conditions for CNV detection: enzyme activation at 95°C for 10 minutes, denaturation and extension at 94°C for 30 seconds and 60°C for 1 minute for a total of 40 cycles, enzyme deactivation at 98°C for 10 minutes, finished with a 4°C hold. Once completed, the 96-well PCR plate was loaded on the QX200™ Droplet Reader.

All experiments had at least two normal controls, and a no-template control (NTC) with water. All samples and controls were run in duplicate, and data from any well with less than 8,000 droplets was treated as failed QC and excluded for downstream analysis. Analysis of the ddPCR data was utilized with QuantaSoft™ software.

1. G. H. Perry *et al.*, Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).

2. J. A. Bailey *et al.*, Recent segmental duplications in the human genome. *Science*. **297**, 1003–7 (2002).

3. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

4. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature*. **526**, 75–81 (2015).

5. R. Redon *et al.*, Global variation in copy number in the human genome. *Nature*. **444**, 444–54 (2006).

6. M. Zarrei, J. R. MacDonald, D. Merico, S. W. Scherer, A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).

7. C. Lee, S. W. Scherer, The clinical context of copy number variation in the human genome. *Expert Rev. Mol. Med.* **12** (2010), p. e8.

8. S. A. McCarroll, D. M. Altshuler, Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).

9. K. Inoue, J. R. Lupski, Molecular Mechanisms for Genomic Disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).

10. G. Merla, N. Brunetti-Pierri, L. Micale, C. Fusco, Copy number variants at Williams–Beuren syndrome 7q11.23 region. *Hum. Genet.* **128**, 3–26 (2010).

11. Y. Chen *et al.*, Copy number variations at the Prader-Willi syndrome region on chromosome 15 and associations with obesity in whites. *Obesity (Silver Spring).* **19**, 1229–34 (2011).

12. J. Clayton-Smith, T. Webb, X. J. Cheng, M. E. Pembrey, S. Malcolm, Duplication of chromosome 15 in the region 15q11-13 in a patient with developmental delay and ataxia with similarities to Angelman syndrome. *J. Med. Genet.* **30**, 529–31 (1993).

13. L. Potocki *et al.*, Molecular mechanism for duplication 17p11.2— the homologous recombination reciprocal of the Smith-Magenis microdeletion. *Nat. Genet.* **24**, 84–87 (2000).

14. P. J. Scambler, The 22q11 deletion syndromes. *Hum. Mol. Genet.* **9**, 2421–6 (2000).

15. R. Schubert, R. Viersbach, T. Eggermann, M. Hansmann, G. Schwanitz, Report of two new cases of Pallister-Killian syndrome confirmed by FISH: tissue-specific mosaicism and loss of i(12p) by in vitro selection. *Am. J. Med. Genet.* **72**, 106–10 (1997).

16. D. T. Miller *et al.*, Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).

17. P. Benn *et al.*, Position statement from the Aneuploidy Screening Committee on behalf of the Board of the International Society for Prenatal Diagnosis. *Prenat. Diagn.* **33**, 622–629 (2013).

18. American College of Obstetricians and Gynecologists Committee on Genetics, Committee Opinion No. 545. *Obstet. Gynecol.* **120**, 1532–1534 (2012).

19. B. Zhou *et al., J. Med. Genet.*, in press, doi:10.1136/jmedgenet-2018-105272.

20. B. Trost *et al.*, A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. *Am. J. Hum. Genet.* **102**, 142–155 (2018).

21. Z. Dong *et al.*, Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet. Med.* **18**, 940–948 (2016).

22. R. Truty *et al.*, Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genet. Med.*, 1 (2018).

23. A. C. Noll *et al.*, Clinical detection of deletion structural variants in whole-genome sequences. *npj Genomic Med.* **1**, 16026 (2016).

24. X. Zhu *et al.*, Identification of copy number variations associated with congenital heart disease by chromosomal microarray analysis and next-generation sequencing. *Prenat. Diagn.* **36**, 321–327 (2016).

25. X. Chen *et al.*, Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. **32**, 1220–1222 (2016).

26. R. M. Layer, C. Chiang, A. R. Quinlan, I. M. Hall, LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).

27. T. Rausch *et al.*, DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. **28**, i333–i339 (2012).

28. G. Klambauer *et al.*, cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 (2012).

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

30. R. E. Handsaker *et al.*, Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).

31. M. Zhu *et al.*, Using ERDS to Infer Copy-Number Variants in High-Coverage Genomes. *Am. J. Hum. Genet.* **91**, 408–421 (2012).

32. A. C. Noll *et al.*, Clinical detection of deletion structural variants in whole-genome sequences. *npj Genomic Med.* **1**, 16026 (2016).

33. X. Fan, T. E. Abbott, D. Larson, K. Chen, *Curr. Protoc. Bioinforma.*, in press, doi:10.1002/0471250953.bi1506s45.

34. R. E. Handsaker, J. M. Korn, J. Nemesh, S. A. McCarroll, Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–76 (2011).

35. J. Monlong *et al.*, Human copy number variants are enriched in regions of low mappability.

Nucleic Acids Res. **46**, 7236–7249 (2018).

36. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

37. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature*. **526**, 75–81 (2015).

38. A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, P. Prins, Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. **31**, 2032–4 (2015).

39. T. Becker *et al.*, FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.* **19**, 38 (2018).

40. M. Mohiyuddin *et al.*, MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*. **31**, 2741–2744 (2015).

41. Z. Dong *et al.*, Identification of balanced chromosomal rearrangements previously unknown among participants in the 1000 Genomes Project: implications for interpretation of structural variation in genomes and the future of clinical cytogenetics. *Genet. Med.* **20**, 697–707 (2018).

42. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. **27**, 764–770 (2011).

43. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–9 (2009).

44. A. Untergasser *et al.*, Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35**, W71-4 (2007).

45. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656–64 (2002).

Table 1: The comparison of Affymetrix CytoScan HD and JAX-CNV. The ten samples are selected associated with 13 pathogenic.

Coriell IDs	Clinical Disorder	Pathogenic CNV	Affymetrix CytoScan HD	JAX-CNV								
				Ori_cov (42-46x)	30x	20x	15x	10x	9x	8x	6x	4x
GM05876	DiGeorge Syndrome	22q11.21 (1.4Mb loss)	O	O	O	O	O	O	O	O	O	O
GM09209	Miller-Dieker Lissencephaly Syndrome	17p13.3 (5.9Mb loss)	O	O	O	O	O	O	O	O	O	O
GM11516	Angleman Syndrome	15q11.2q13.1 (7Mb loss)	O	O	O	O	O	O	O	O	O	O
GM13480	Williams Syndrome	7q11.23 (1.6Mb loss)	O	O	O	O	O	O	O	O	O	O
		9p24.1 (107.6Kb gain)	O	O	O	O	O					
GM14164	Tetralogy Fallot	13q14.2 (47.9Mb loss)	O	O	O	O	O	O	O	O	O	O
		22q11.21 (148.8Kb gain)	poor/low probe coverage	O	O	O						
GM16593	Cri-du-chat Syndrome	5p15.3 (14.7Mb loss)	O	O	O	O	O	O	O	O	O	O
		14q24.3 (2.7Mb loss)	O	O	O	O	O	O	O	O	O	O
GM20375	Angleman Syndrome	15q11.2q13.1 (4.9Mb loss)	O	O	O	O	O	O	O	O	O	O
GM20743	Smith-Magenis Syndrome	17p11.2 (2.1Mb loss)	O	O	O	O	O	O	O	O	O	*
GM22569	1p deletion Syndrome	1p36.33 (5.5Mb loss)	O	O	O	O	O	O	O	O	O	O
GM22601	Wolf-Hirschhorn Syndrome	4p16.3 (25.0Mb loss)	O	O	O	O	O	O	O	O	O	*

Note: O: denotes CNVs captured by the algorithm/method. *: CNVs are not 50% reciprocal overlapping, but recovered in manual review. Shadowed cells mean no call.

Table 2: ddPCR validation for NGS unique CNVs of GM05876 and GM09209.

Cytoband	Type	Size (Kb)	Samples	Validated	Remark
1p31	loss	55.3	GM05876	TRUE	
1p36	gain	53.9	GM05876 & GM09209	TRUE	
1p36	gain	68	GM05876	TRUE	
2p22	gain	45.1	GM05876 & GM09209	FALSE	Overlap with a SEGDUP
3q26	loss	114.2	GM09209	TRUE	
5q35	loss	52.3	GM05876 & GM09209	TRUE	
6p25	gain	50.4	GM05876 & GM09209	TRUE	
6p22	loss	86.4	GM09209	TRUE	
12p11	gain	57.7	GM05876	TRUE	
12p13	loss	87	GM05876 & GM09209	TRUE	
14q11	gain	185.6	GM09209	TRUE	
14q21	loss	54.3	GM09209	TRUE	
15q11.2	gain	84	GM05876	TRUE	
15q11.1	gain	91.3	GM09209	TRUE	
16p12	gain	68.95	GM09209	TRUE	
16p11	gain	59.4	GM05876 & GM09209	FALSE	Overlap with a simple repeats
16p11	gain	53.3	GM05876 & GM09209	FALSE	Overlap with a simple repeats
19p11	gain	50.3	GM05876 & GM09209	FALSE	Overlap with a SEGDUP
21p11	loss	224.9	GM05876	NA	Not conclusive

Supplementary Tables

S1

Illumina 2x150PE sequencing				
Sample	**ReadCount**	**ReadLength**	**EstimatedInsertSize**	**EstimatedCoverage**
GM05876	460,314,950	150bp PE	371	46
GM09209	455,583,105	150bp PE	397	46
GM11516	443,070,763	150bp PE	390	44
GM13480	422,056,408	150bp PE	383	42
GM14164	459,003,616	150bp PE	375	46
GM16593	459,768,472	150bp PE	381	46
GM20375	454,985,153	150bp PE	388	46
GM20743	440,830,304	150bp PE	388	44
GM22569	447,600,481	150bp PE	392	45
GM22601	448,272,730	150bp PE	383	45

S2

Pathogenic CNV coordinates					
Sample	**Chromosome**	**Begin**	**End**	**Length**	**Type**
GM05876	chr22	18890273	20324382	1,434,110	DEL
GM09209	chr17	150720	5918781	5,768,062	DEL
GM11516	chr15	23370621	28300455	4,929,835	DEL
GM13480	chr7	73245561	74727754	1,482,194	DEL
	chr9	5098561	5206183	107,623	DUP
GM14164	chr13	47227949	95062722	47,834,774	DEL
	chr22	18888903	19037700	148,798	DUP
GM16593	chr5	7554151	22285853	14,731,703	DEL
	chr14	78435797	81133167	2,697,371	DEL
GM20375	chr15	24215968	27830148	3,614,181	DEL
GM20743	chr17	16836955	18819510	1,982,556	DEL
GM22569	chr1	690077	6272609	5,582,533	DEL
GM22601	chr4	65773	25980331	25,914,559	DEL

S3

Affymetrix CytoScan HD reported CNV coordinates

Sample	Chromosome	Begin	End	Length	Type
GM05876	chr8	39368816	39529433	160,618	DEL
	chr8	46980782	47094982	114,201	DEL
	chr17	46135457	46215376	79,920	DUP
	chr19	20413299	20537899	124,601	DEL
	chr22	18929329	20325138	1,395,810	DEL
GM09209	chr6	94776948	94861597	84,650	DEL
	chr8	39389578	39529433	139,856	DEL
	chr17	150732	5977916	5,827,185	DEL
	chr17	46110125	46215376	105,252	DUP
GM11516	chr14	105780376	106771352	990,977	DEL
	chr15	23102646	28770762	5,668,117	DEL
	chr17	46135457	46215376	79,920	DUP
GM13480	chr7	72532206	72842895	310,690	DUP
	chr7	73179193	74881821	1,702,629	DEL
	chr9	5106684	5228343	121,660	DUP
	chr10	46417570	46583083	165,514	DUP
	chr17	36097968	36150098	52,131	DUP
	chr21	23120390	23267071	146,682	DEL
	chr22	22027193	22916199	889,007	DEL
GM14164	chr5	12578472	12678806	100,335	DEL
	chr13	47228260	95060365	47,832,106	DEL
	chr17	46135457	46215376	79,920	DUP
GM16593	chr1	91674697	92240952	566,256	DUP
	chr5	7554150	22290079	14,735,930	DEL
	chr7	65191248	65630056	438,809	DEL
	chr8	39389578	39529433	139,856	DEL
	chr11	55606698	55674829	68,132	DEL
	chr14	78433830	81133860	2,700,031	DEL
	chr14	105780448	106393766	613,319	DEL
	chr22	22374765	22868186	493,422	DEL
GM20375	chr7	69265509	69671863	406,355	DUP
	chr7	101325081	101498586	173,506	DUP
	chr8	136665652	136850192	184,541	DEL
	chr15	23370621	28300455	4,929,835	DEL
	chr18	72124126	72203343	79,218	DEL
	chr22	22217466	22509323	291,858	DEL
GM20743	chr1	248473949	248631917	157,969	DEL
	chr2	51127356	51290637	163,282	DEL
	chr7	45220233	46021123	800,891	DUP
	chr8	39389578	39529433	139,856	DEL

Sample	Chromosome	Begin	End	Length	Type
	chr8	136664977	136850192	185,216	DEL
	chr10	46400351	46583083	182,733	DUP
	chr11	55606698	55685520	78,823	DEL
	chr14	105757249	105862093	104,845	DEL
	chr17	16708434	18839821	2,131,388	DEL
GM22569	chr1	914086	6279099	5,365,014	DEL
	chr6	161981973	162032707	50,735	DEL
	chr11	55606542	55685520	78,979	DEL
	chr17	46110125	46215376	105,252	DUP
GM22601	chr4	68453	25980239	25,911,787	DEL
	chr11	55606542	55685520	78,979	DEL
	chr11	134283980	134347316	63,337	DEL
	chr17	46135457	46215376	79,920	DUP
	chr19	20412520	20537899	125,380	DEL

S4

JAX-CNV reported CNV coordinates

Sample	Chromosome	Begin	End	Length	Type	Pathogenic	Overlap_w_CytoScan	ddPCR Validated	Remark
GM05876	chr1	16594859	16649399	54,541	DUP			TRUE	
	chr1	16881249	16949249	68,001	DUP			TRUE	
	chr1	72289349	72348449	59,101	DEL			TRUE	
	chr2	37730949	37776099	45,151	DUP			FALSE	
	chr5	180949949	181004499	54,551	DEL			TRUE	
	chr6	296099	382499	86,401	DUP			TRUE	
	chr8	39374599	39533749	159,151	DEL		TRUE		
	chr8	46989499	47095299	105,801	DEL		TRUE		
	chr10	46514199	46561699	47,501	DUP			Cannot degin primers	
	chr12	31849049	31910299	61,251	DUP			TRUE	
	chr12	9482449	9569499	87,051	DEL			TRUE	
	chr15	22224249	22308299	84,051	DUP			TRUE	
	chr16	22613649	22699449	85,801	DUP			FALSE	
	chr16	33569079	33626049	56,971	DUP			FALSE	
	chr17	46135949	46219599	83,651	DUP		TRUE		
	chr19	20413049	20535199	122,151	DEL		TRUE		
	chr19	24334199	24385529	51,331	DUP			FALSE	
	chr21	9607449	9832429	224,981	DUP			Not conclusive	
	chr22	18900699	20351049	1,450,351	DEL	TRUE	TRUE		
GM09209	chr1	16595499	16662299	66,801	DUP			TRUE	
	chr2	37730499	37776099	45,601	DUP			FALSE	
	chr3	162794399	162908599	114,201	DEL			TRUE	
	chr5	180949049	181003749	54,701	DEL			TRUE	
	chr6	296099	382499	86,401	DUP			TRUE	
	chr6	29881499	29931899	50,401	DEL			TRUE	
	chr6	94776149	94871299	95,151	DEL		TRUE		
	chr8	39374549	39529899	155,351	DEL		TRUE		
	chr12	9482449	9569499	87,051	DEL			TRUE	
	chr14	19771049	19956649	185,601	DUP			TRUE	
	chr14	41140099	41194459	54,361	DEL			TRUE	
	chr15	20331469	20422849	91,381	DUP			TRUE	
	chr16	14961449	15030399	68,951	DUP			Cannot degin primers	
	chr16	22640049	22699449	59,401	DUP			TRUE	
	chr16	32552199	32665999	113,801	DUP				Not overlap with HG19 result
	chr16	33572719	33632299	59,581	DUP			FALSE	
	chr16	33632299	33692419	60,121	DUP			FALSE	
	chr16	33764649	33830249	65,601	DUP				Not overlap with HG19 result
	chr17	141699	5979599	5,837,901	DEL	TRUE	TRUE		
	chr17	46087949	46135899	47,951	DUP		TRUE		
	chr17	46135899	46204219	68,321	DUP		TRUE		
	chr19	24335899	24386209	50,311	DUP			FALSE	
	chr21	9591514	9784309	192,796	DUP				Not overlap with HG19 result
	chr21	10414749	10484004	69,256	DUP				
	chr21	10656739	10735669	78,931	DUP				

Sample	Chromosome	Begin	End	Length	Type	Pathogenic	Overlap_w_CytoScan	ddPCR Validated	Remark
GM11516	chr1	143586149	143698524	112,376	DEL				
	chr1	16600739	16664099	63,361	DUP				
	chr1	16881099	16949099	68,001	DUP				
	chr1	72290049	72348349	58,301	DEL				
	chr2	37730949	37775999	45,051	DUP				
	chr2	88860699	89031949	171,251	DEL				
	chr5	110226349	110278459	52,111	DEL				
	chr5	180949099	181004049	54,951	DEL				
	chr6	256649	382499	125,851	DUP				
	chr6	78257549	78327799	70,251	DEL				
	chr11	55597999	55681159	83,161	DEL				
	chr12	9482349	9580799	98,451	DEL				
	chr14	105776399	105867899	91,501	DEL				
	chr14	105881099	105991539	110,441	DEL				
	chr14	106027699	106082049	54,351	DEL				
	chr14	106188399	106323149	134,751	DEL				
	chr14	106369449	106629799	260,351	DEL				
	chr14	19771439	19956599	185,161	DUP				
	chr15	20351994	20422749	70,756	DUP				
	chr15	23381859	28234739	4,852,881	DEL	TRUE	TRUE		
	chr16	14955899	15030849	74,951	DUP				
	chr16	22613699	22699449	85,751	DUP				
	chr16	33562749	33839949	277,201	DUP				
	chr17	46135849	46220099	84,251	DUP		TRUE		
	chr21	10420409	10482744	62,336	DUP				
	chr21	10663639	10739119	75,481	DUP				
	chr21	9591514	9784309	192,796	DUP				
GM13480	chr1	16597524	16654899	57,376	DUP				
	chr2	88865049	89035599	170,551	DEL				
	chr3	162827599	162908649	81,051	DEL				
	chr4	34778249	34827699	49,451	DEL				
	chr6	296099	382449	86,351	DUP				
	chr7	72532649	72850349	317,701	DUP		TRUE		
	chr7	73302949	74597464	1,294,516	DEL	TRUE	TRUE		
	chr8	39374549	39533749	159,201	DEL				
	chr9	5098099	5231349	133,251	DUP	TRUE	TRUE		
	chr10	46513999	46561649	47,651	DUP				
	chr14	105786549	105864299	77,751	DEL				
	chr14	105864299	105939949	75,651	DEL				
	chr14	19771049	19956649	185,601	DUP				
	chr15	20332429	20429449	97,021	DUP				
	chr15	22255249	22302994	47,746	DUP				
	chr15	87281629	87328699	47,071	DEL				
	chr16	14955399	15030399	75,001	DUP				
	chr16	22613699	22699349	85,651	DUP				
	chr17	36109499	36157099	47,601	DUP		TRUE		
	chr18	60595899	60645999	50,101	DUP				
	chr19	24330399	24384009	53,611	DUP				

Sample	Chromosome	Begin	End	Length	Type	Pathogenic	Overlap_w_CytoScan	ddPCR Validated	Remark
	chr21	10420329	10482734	62,406	DUP				
	chr21	10655109	10734804	79,696	DUP				
	chr21	23120299	23268549	148,251	DEL		TRUE		
	chr21	9591559	9784339	192,781	DUP				
	chr22	22035099	22899799	864,701	DEL		TRUE		
GM14164	chr1	16591449	16656834	65,386	DUP				
	chr1	16878799	16954249	75,451	DUP				
	chr1	72289999	72348549	58,551	DEL				
	chr2	88832799	89034199	201,401	DEL				
	chr3	162827599	162914099	86,501	DEL				
	chr5	12578199	12688549	110,351	DEL		TRUE		
	chr5	180948999	181003649	54,651	DEL				
	chr6	297149	382449	85,301	DUP				
	chr6	29881524	29932124	50,601	DEL				
	chr6	78257499	78327549	70,051	DEL				
	chr11	55597999	55684399	86,401	DEL				
	chr13	47226699	63069399	15,842,701	DEL	TRUE	TRUE		
	chr13	63074949	95060649	31,985,701	DEL	TRUE	TRUE		
	chr15	20339709	20441649	101,941	DUP				
	chr15	22226799	22308299	81,501	DUP				
	chr16	14953999	15030449	76,451	DUP				
	chr16	22613699	22699499	85,801	DUP				
	chr17	36109499	36157099	47,601	DUP				
	chr17	46135449	46230599	95,151	DUP		TRUE		
	chr19	24330449	24384599	54,151	DUP				
	chr21	10414449	10483929	69,481	DUP				
	chr21	10666119	10740309	74,191	DUP				
	chr21	9591514	9784309	192,796	DUP				
	chr22	18947749	19025599	77,851	DUP	TRUE			
	chr22	22798299	22883899	85,601	DEL				
GM16593	chr1	143586629	143700454	113,826	DEL				
	chr1	16594299	16657209	62,911	DUP				
	chr1	248575999	248634899	58,901	DEL				
	chr1	72300349	72348799	48,451	DEL				
	chr1	91674349	92239949	565,601	DUP		TRUE		
	chr2	37730899	37776099	45,201	DUP				
	chr2	88862299	89083549	221,251	DEL				
	chr2	90035649	90118649	83,001	DEL				
	chr2	90169349	90226949	57,601	DEL				
	chr4	68574374	68625699	51,326	DEL				
	chr5	180949099	181004849	55,751	DEL				
	chr5	7553899	22293999	14,740,101	DEL	TRUE	TRUE		
	chr6	296499	382549	86,051	DUP				
	chr6	29881634	29931914	50,281	DEL				
	chr7	65228249	65462489	234,241	DEL		TRUE		
	chr8	39374649	39533699	159,051	DEL		TRUE		
	chr10	46519349	46564699	45,351	DUP				
	chr11	55597949	55681109	83,161	DEL		TRUE		
	chr14	105775299	105864299	89,001	DEL				

Sample	Chromosome	Begin	End	Length	Type	Pathogenic	Overlap_w_CytoScan	ddPCR Validated	Remark
	chr14	105864299	105990434	126,136	DEL				
	chr14	106019899	106083349	63,451	DEL				
	chr14	41139599	41194454	54,856	DEL				
	chr14	78433499	81133649	2,700,151	DEL	TRUE	TRUE		
	chr15	20351979	20422749	70,771	DUP				
	chr15	22225499	22308299	82,801	DUP				
	chr15	23371249	23428174	56,926	DEL				
	chr16	14954299	15030599	76,301	DUP				
	chr16	22613699	22699449	85,751	DUP				
	chr21	10413199	10484839	71,641	DUP				
	chr21	9591514	9784309	192,796	DUP				
	chr22	22387349	22906499	519,151	DEL		TRUE		
GM20375	chr1	16878299	16954349	76,051	DUP				
	chr1	248574599	248634849	60,251	DEL				
	chr2	37731049	37776099	45,051	DUP				
	chr2	88887849	89032499	144,651	DEL				
	chr5	180948999	181003799	54,801	DEL				
	chr5	97712149	97760549	48,401	DEL				
	chr6	78257549	78328499	70,951	DEL				
	chr7	101357409	101492849	135,441	DUP		TRUE		
	chr7	69256799	69669449	412,651	DUP		TRUE		
	chr8	136668099	136850199	182,101	DEL		TRUE		
	chr8	39374599	39533749	159,151	DEL				
	chr10	46514149	46561549	47,401	DUP				
	chr14	105883799	105992279	108,481	DEL				
	chr14	106027849	106082099	54,251	DEL				
	chr14	106116799	106320019	203,221	DEL				
	chr14	19771039	19956599	185,561	DUP				
	chr14	41139999	41194584	54,586	DEL				
	chr15	20332229	20429049	96,821	DUP				
	chr15	22224199	22308299	84,101	DUP				
	chr15	23367599	28347149	4,979,551	DEL	TRUE	TRUE		
	chr15	34431349	34517974	86,626	DEL				
	chr16	14960149	15030449	70,301	DUP				
	chr16	22613699	22699449	85,751	DUP				
	chr18	72123849	72196149	72,301	DEL		TRUE		
	chr19	24330449	24381179	50,731	DUP				
	chr21	10412599	10483789	71,191	DUP				
	chr21	10653949	10734274	80,326	DUP				
	chr21	9591514	9784309	192,796	DUP				
	chr22	22353359	22904999	551,641	DEL				
GM20743	chr1	143586659	143700434	113,776	DEL				
	chr1	16600809	16664199	63,391	DUP				
	chr1	16900899	16949049	48,151	DUP				
	chr1	248584099	248634849	50,751	DEL				
	chr1	72300699	72348499	47,801	DEL				
	chr2	51129299	51292149	162,851	DEL		TRUE		
	chr2	88863199	88947399	84,201	DEL				

Sample	Chromosome	Begin	End	Length	Type	Pathogenic	Overlap_w_CytoScan	ddPCR Validated	Remark
	chr2	88947399	89037399	90,001	DEL				
	chr3	195694799	195749999	55,201	DUP				
	chr5	180948949	181003699	54,751	DEL				
	chr6	256599	381949	125,351	DUP				
	chr6	78257549	78330249	72,701	DEL				
	chr7	45219199	46022249	803,051	DUP		TRUE		
	chr8	136668049	136850199	182,151	DEL		TRUE		
	chr8	39374649	39529799	155,151	DEL		TRUE		
	chr10	46218849	46332899	114,051	DUP				
	chr10	46514049	46561649	47,601	DUP				
	chr11	55597999	55681474	83,476	DEL		TRUE		
	chr12	9482399	9582449	100,051	DEL				
	chr14	105776449	105893649	117,201	DEL		TRUE		
	chr14	19770959	19956199	185,241	DUP				
	chr16	14960499	15030349	69,851	DUP				
	chr16	22613699	22699449	85,751	DUP				
	chr16	32557719	32655099	97,381	DUP				
	chr16	33613199	33693104	79,906	DUP				
	chr16	33763779	33830699	66,921	DUP				
	chr17	16852549	18391399	1,538,851	DEL	TRUE	TRUE		
	chr17	18624499	18721299	96,801	DEL	TRUE	TRUE		
	chr17	18763999	18824119	60,121	DEL	TRUE	TRUE		
	chr21	10413249	10483809	70,561	DUP				
	chr21	9591514	9784309	192,796	DUP				
GM22569	chr1	143585799	143697424	111,626	DEL				
	chr1	16594299	16662299	68,001	DUP				
	chr1	16880799	16948949	68,151	DUP				
	chr1	818499	6281549	5,463,051	DEL	TRUE	TRUE		
	chr2	88832849	89032549	199,701	DEL				
	chr6	161977199	162033149	55,951	DEL		TRUE		
	chr6	296649	382499	85,851	DUP				
	chr11	55598049	55664099	66,051	DEL		TRUE		
	chr12	9480799	9580649	99,851	DEL				
	chr14	105895199	106082149	186,951	DEL				
	chr14	106276549	106325049	48,501	DEL				
	chr15	20343919	20422429	78,511	DUP				
	chr16	14954499	15030649	76,151	DUP				
	chr16	22632849	22699449	66,601	DUP				
	chr16	32546974	32632249	85,276	DUP				
	chr16	33572059	33626149	54,091	DUP				
	chr16	33626149	33689959	63,811	DUP				
	chr16	33763419	33829299	65,881	DUP				
	chr17	46087949	46291899	203,951	DUP		TRUE		
	chr19	24330449	24387689	57,241	DUP				
	chr21	10412549	10483784	71,236	DUP				
	chr21	10663669	10739134	75,466	DUP				
	chr21	9591514	9784309	192,796	DUP				
	chr22	22817949	22899699	81,751	DEL				

Sample	Chromosome	Begin	End	Length	Type	Pathogenic	Overlap_w_CytoScan	ddPCR Validated	Remark
GM22601	chr1	16591449	16662299	70,851	DUP				
	chr1	16882149	16949149	67,001	DUP				
	chr2	37730799	37776099	45,301	DUP				
	chr2	88833099	89031949	198,851	DEL				
	chr3	162794349	162908749	114,401	DEL				
	chr4	68899	8529484	8,460,586	DEL	TRUE	TRUE		
	chr4	9479999	25982899	16,502,901	DEL	TRUE	TRUE		
	chr5	176128599	176226349	97,751	DUP				
	chr6	296099	382499	86,401	DUP				
	chr11	134281899	134344549	62,651	DEL		TRUE		
	chr11	49691599	49737699	46,101	DEL				
	chr11	55597949	55681604	83,656	DEL		TRUE		
	chr14	105883599	106504399	620,801	DEL				
	chr15	22224999	22308299	83,301	DUP				
	chr16	14963199	15030399	67,201	DUP				
	chr16	22621049	22700149	79,101	DUP				
	chr16	32555149	32645299	90,151	DUP				
	chr16	33577099	33626149	49,051	DUP				
	chr16	33626149	33688219	62,071	DUP				
	chr16	33764659	33830499	65,841	DUP				
	chr17	36109449	36157249	47,801	DUP				
	chr17	46135949	46204189	68,241	DUP		TRUE		
	chr19	20413049	20537899	124,851	DEL		TRUE		
	chr19	24330449	24384029	53,581	DUP				
	chr21	10414449	10483974	69,526	DUP				
	chr21	10663759	10739179	75,421	DUP				
	chr21	9591514	9784309	192,796	DUP				

S5

The sensitivity comparison of using Affymetrix CytoScan HD reported CNVs

	CytoScan (Cols 2-3)		JGM -- NGS									Other SV/CNV Algorithms				
Coriell IDs	Raw Calls Total (gain/loss)	After ddPCR Validation	Original 42-46x	30x	20x	15x	10x	9x	8x	6x	4x	Manta	Lumpy	Delly	CNVnator	cn.MOPS
GM05876	8(3/5)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	4(0/4)	4(0/4)	4(0/4)	4(0/4)	4(0/4)	2(0/2)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	0(0/0)
GM09209	5(2/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	3(0/3)	3(0/3)	3(0/3)	3(1/2)	3(1/2)	3(1/2)	4(1/3)	4(1/3)	1(0/1)
GM11516	3(1/2)	3(1/2)	3(1/2)	3(1/2)	3(1/2)	2(0/2)	2(0/2)	2(0/2)	2(0/2)	2(0/2)	2(0/2)	3(1/2)	3(1/2)	3(1/2)	3(1/2)	1(0/1)
GM13480	7(4/3)	7(4/3)	7(4/3)	7(4/3)	7(4/3)	7(4/3)	3(0/3)	4(1/3)	5(2/3)	3(0/3)	3(0/3)	7(4/3)	7(4/3)	7(4/3)	7(4/3)	1(0/1)
GM14164	3(1/2)	3(1/2)	3(1/2)	3(1/2)	3(1/2)	3(1/2)	2(0/2)	2(0/2)	2(0/2)	2(0/2)	2(0/2)	3(1/2)	3(1/2)	3(1/2)	3(1/2)	1(0/1)
GM16593	9(1/8)	8(1/7)	8(1/7)	8(1/7)	8(1/7)	8(1/7)	8(1/7)	8(1/7)	8(1/7)	7(0/7)	7(0/7)	8(1/7)	8(1/7)	8(1/7)	8(1/7)	3(0/3)
GM20375	7(2/5)	6(2/4)	6(2/4)	6(2/4)	6(2/4)	6(2/4)	4(0/4)	4(0/4)	3(0/3)	4(0/4)	4(0/4)	6(2/4)	6(2/4)	6(2/4)	6(2/4)	1(0/1)
GM20743	10(2/8)	9(2/7)	9(2/7)	9(2/7)	9(2/7)	9(2/7)	9(2/7)	8(1/7)	9(2/7)	8(1/7)	6(1/5)	9(2/7)	9(2/7)	8(2/6)	9(2/7)	3(0/3)
GM22569	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	3(1/2)	4(1/3)	4(1/3)	4(1/3)	4(1/3)	2(0/2)
GM22601	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	5(1/4)	1(0/1)
Total	61(18/43)	54(15/39)	54(15/39)	54(15/39)	54(15/39)	52(13/39)	45(6/39)	44(5/39)	45(7/38)	42(3/39)	37(4/33)	53(15/38)	53(15/38)	53(15/38)	54(15/38)	14(0/14)
Sensitivity (%)			100(100/100)	100(100/100)	100(100/100)	96(87/100)	83(40/100)	83(33/100)	83(47/97)	77(20/100)	68(27/85)	98(100/97)	98(100/97)	98(100/97)	100(100/100)	26(0/36)

The numbers of CNVs not 50% reciprocal overlapping with Affymetrix CytoScan HD results, but still in the regions.

	CytoScan		JGM -- NGS									Other SV/CNV Algorithms				
Coriell IDs	Raw Calls Total (gain/loss)	After ddPCR Validation	Original 42-46x	30x	20x	15x	10x	9x	8x	6x	4x	Manta	Lumpy	Delly	CNVnator	cn.MOPS
GM05876			0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	1(0/1)	1(0/1)	0(0/0)	0(0/0)	0(0/0)
GM09209			0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	1(0/1)	0(0/0)	1(1/0)	0(0/0)	0(0/0)	1(0/1)
GM11516			1(0/1)	1(0/1)	1(0/1)	1(0/1)	1(0/1)	1(0/1)	1(0/1)	0(0/0)	0(0/0)	1(0/1)	0(0/0)	0(0/0)	1(0/1)	0(0/0)
GM13480			1(1/0)	1(1/0)	1(1/0)	2(2/0)	0(0/0)	1(1/0)	1(1/0)	0(0/0)	0(0/0)	4(3/1)	4(3/1)	3(3/0)	1(1/0)	0(0/0)
GM14164			0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	1(0/1)	1(0/1)	1(0/1)	0(0/0)	1(0/1)
GM16593			1(0/1)	1(1/1)	1(1/1)	2(0/2)	3(1/2)	3(1/2)	3(1/2)	1(0/1)	2(0/2)	4(1/3)	4(1/3)	3(1/2)	2(0/2)	1(0/1)
GM20375			1(0/1)	1(0/1)	1(0/1)	3(2/1)	0(0/0)	0(0/0)	0(0/0)	1(0/1)	1(0/1)	1(0/1)	2(1/1)	1(0/1)	1(0/1)	1(0/1)
GM20743			2(1/1)	2(1/1)	2(1/1)	2(1/1)	3(2/1)	2(1/1)	3(2/1)	3(1/2)	4(1/3)	3(1/2)	4(1/3)	2(1/1)	1(0/1)	2(0/2)
GM22569			0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	1(1/0)	1(1/0)	0(0/0)	0(0/0)	3(1/2)	3(1/2)	2(1/1)	0(0/0)	1(0/1)
GM22601			0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	0(0/0)	1(0/1)	1(0/1)	1(0/1)	1(0/1)	1(0/1)
Total			6(2/4)	6(2/4)	6(2/4)	10(5/5)	7(3/4)	8(4/4)	9(5/4)	5(1/4)	8(1/7)	19(6/13)	21(8/13)	13(6/7)	7(1/6)	8(0/8)

S6

Region	Type	Left Primer	Right Primer	Internal Probe
1p31	DEL	CAGTCTGTTACCTGGTTCAC	GAGGTAATGGTAGGGTGCTA	CAAAGGCCTGAGCTTGGGAA
1p36	DUP	AGAAGCAGCCTCAGCAAT	CTCAGTGGTTTCTGGTCTTC	ATCTGGGTGGGCAGAACTTG
1p36	DUP	ACAGACAATGGAACAAGCAA	GGGAACTCTGAAATCCTTGG	ATGGGAGCAAGTCCTTTGCC
2p22	DUP	TATCTCTCATTGCCTCCGAA	CAAGGTGAGCAGAACATAGG	CTCACTGCTGAAATGATGAGGTC
3q26	DEL	TTGTGCTGTATCCTCCAAAC	CACTCCACCTGGTTACATTC	TGGAGTTGGGTTCAACGTGG
5q35	DEL	TGATAAAGGCGTCAAGGTTC	CTAGATAATGACAGCCAGCG	AGGGTAAAGGTGCTGCTTTGG
6p25	gain	GAGGGAGTTCATGTGGTAGA	AAGTAAGTGGAAGGAGTGGT	AGGTGTTTGCCTCTGAAGCC
6p25	gain	TGCTAGAAAGGTCCAAGTCT	TGCTGGAAACAACTGTCAAT	CCAGTTGCTGTTTGACCTGGA
6p25	gain	ATATGCCTTTAGAGGGAGCA	CTCCGCATGTACTTATTCCC	AACCCAAGGCTGACACAGTG
6p22	DEL	CGACTCCTCCAAGGAAG	CACTTGAGGGCTCCTTG	TGCCTTCCTAGACACTGGTGTTA
12p11	Gain	AATGCTTGTCCTCTGTAACG	TTAAATCACACTGCCATCCC	CCCTATTGGGATCTGGGCTAGT
12p13	DEL	GGGATAGGACACAGATGGAT	CTTATTCCCTTAACCCGCAG	AGCTGATTGCAGGTGCTTCC
14q11	DUP	CTAACAGAGCAGCATCACAA	TAAAGCAGCAAGATACAGCC	ACTCCTGCAGCCTAAGACAT
14q21	DEL	TCTGTGTTGTGTTGGATGTC	GTGCTGCTGGTTCTCTTATT	ACCTGGAAGTTCCCTAGGCA
15q11.2	DUP	TGTCTTAGGCTGAGTCTACC	AAACTCAAGGGCTCTAATGC	TCCAATATCCATCTTCTCATTCTTCCT
15q11.1	DUP	ATTGAACTGACAGCCAACAA	GCCTTACAGAGAACAGACAC	CTGCCTGAACGAAGCTCATCT
16p12	DUP	GCCACTATAACCTTTCCCAC	CAAGGACTCGCAAATTCTCT	CAGCATCCATCTCCAGTAACTTG
16p11	DUP	CCAACAGAGTGAGACTGTC	TGCAGAGGAGAACGTCATT	TGCTAGGGTTCATGCCACAC
16p11	DUP	GTTCATCACTTAAGCACCTG	ATCGGCAATTATGCAGAAGA	CGTGTTCCAGTCCAGTATCCC
19p11	DUP	ACCAGATGACACTAAGGGAA	CAGGAGAAACTCAGCCAAAT	AAAGAACCTACTAGAAATGTCGGG
21p11	DEL	TCTCTGACTTCCTGGTTCAA	ATTAGTTGGGCATGATGGTG	AGCCTCCTGAGTAGCTGGGA

S7

Comparison of other CNV calling algorithms on pathogenic CNVs

Sample	Length	Type	Manta	Lumpy	Delly	CNVnator	cn.MOPS
GM05876	1,434,110	DEL	O	O	O	O	
GM09209	5,768,062	DEL			O	O	
GM11516	4,929,835	DEL		O	O	O	O
GM13480	1,482,194	DEL			O	O	O
	107,623	DUP				O	
GM14164	47,834,774	DEL	O	O	O	O	
	148,798	DUP				O	
GM16593	14,731,703	DEL	O	O	O	*	
	2,697,371	DEL				O	
GM20375	3,614,181	DEL	*	*	O	O	O
GM20743	1,982,556	DEL	O		O	O	
GM22569	5,582,533	DEL			O	O	
GM22601	25,914,559	DEL	O	O	O	*	*
		Total CNVs	28,228	51,406	408,522	4,086	3,500

Note: O: denotes CNVs captured by the algorithm/method. *: CNVs are not 50% reciprocal overlapping, but recovered in manual review. The plots are given in Supplementary files S1-S2.

S8

Comparison of TWB Genotype Array and JAX-CNV on Taiwan BioBank Samples

	Deletion		Duplication	
	# Probes > 200	# Probes 100-200	# Probes > 200	# Probes 100-200
SNP Array	19	46	56	113
NGS recalled	18	32	47	71
Rate (%)	94.7	69.6	83.9	62.8

Development of a Next Generation Sequencing (NGS)-Based Platform for Detection of Copy Number Variations (CNVs) Associated with Constitutional Disorders

Wan-Ping Lee, Qihui Zhu, Silvia Liu, Eliza Cerveira, Mallory Ryan, Adam Mil-Homens, Lauren Bellfy, CZ Zhang, Charles Lee
The Jackson Laboratory for Genomic Medicine, Farmington, CT

ABSTRACT

Chromosomal abnormalities are known to be associated with a large number of constitutional disorders. Conventional cytogenetic analyses, such as karyotyping and fluorescence in situ hybridization (FISH), have been traditionally used for the detection of chromosomal aberrations in the clinic. Since 2010 chromosomal microarray (CMA) has replaced conventional cytogenetic analysis as the first-tier cytogenetic diagnostic test. As the advancement of NGS technologies, it is conceivable to replace the conventional and outdated molecular cytogenetic methods. However, the barrier is stable and robust bioinformatics tools. We then developed a new one to meet the requirement for clinical applications.

The newly developed NGS-based tool has been tested on ten Coriell samples associated with 13 known pathogenic CNVs, ranging from 107.6Kb to 47.9Mb. Preliminary results show 100% sensitivity, calling all 13 pathogenic CNVs. Interestingly, by using NGS technologies our pipeline is able to detect an additional 25 to 40 CNVs per sample which is a near six-fold increase in detection when compared to CMA resolution. The false discovery rates of using human reference genomes GRCh37 and GRCh38 are 89% and 100% respectively.

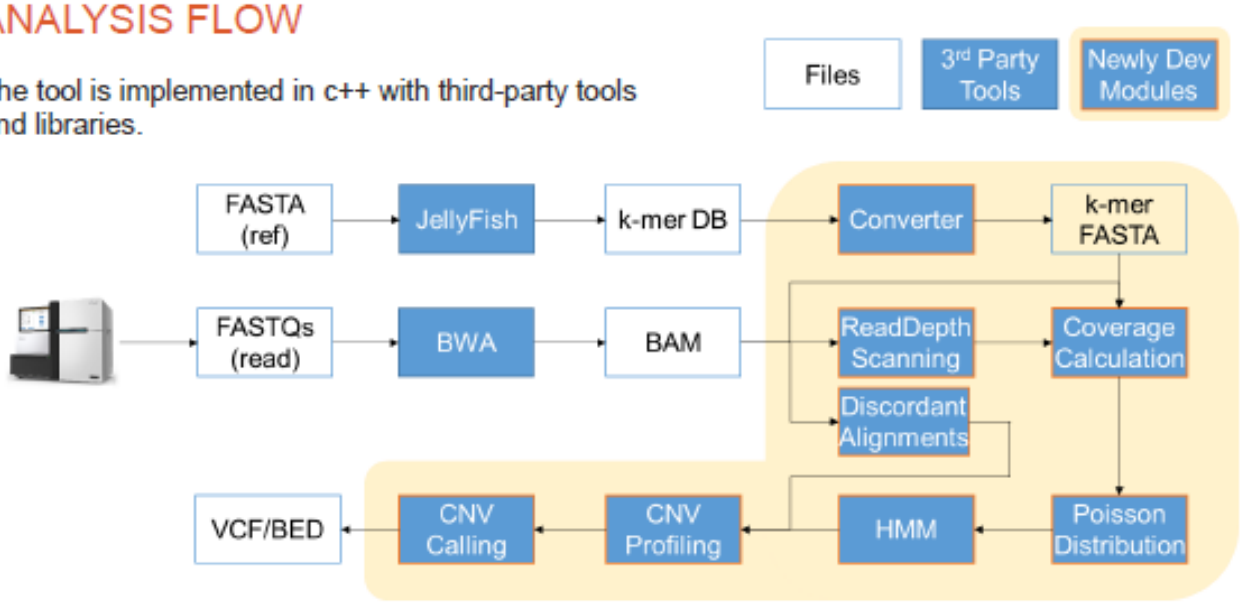
MILESTONES OF JGM CLIA SV PIPELINE

1	>300kb	DELs/DUPs (sensitivity >99.9%; no false positive)	
2	>50kb	DELs/DUPs (sensitivity >99.9%; FDR < 20%)	
3	>50kb	DELs/DUPs (sensitivity >99.9%; FDR < 2-5%)	
4	>100kb	INVs/TRAs (sensitivity >99.9%; FDR < 2-5%)	ETA: 06/2018
5	>50kb	INVs/TRAs (sensitivity >99.9%; FDR < 2-5%)	ETA: 09/2018

DEL: deletion;
DUP: duplication;
INV: inversion;
TRA: translocation

ANALYSIS FLOW

The tool is implemented in c++ with third-party tools and libraries.

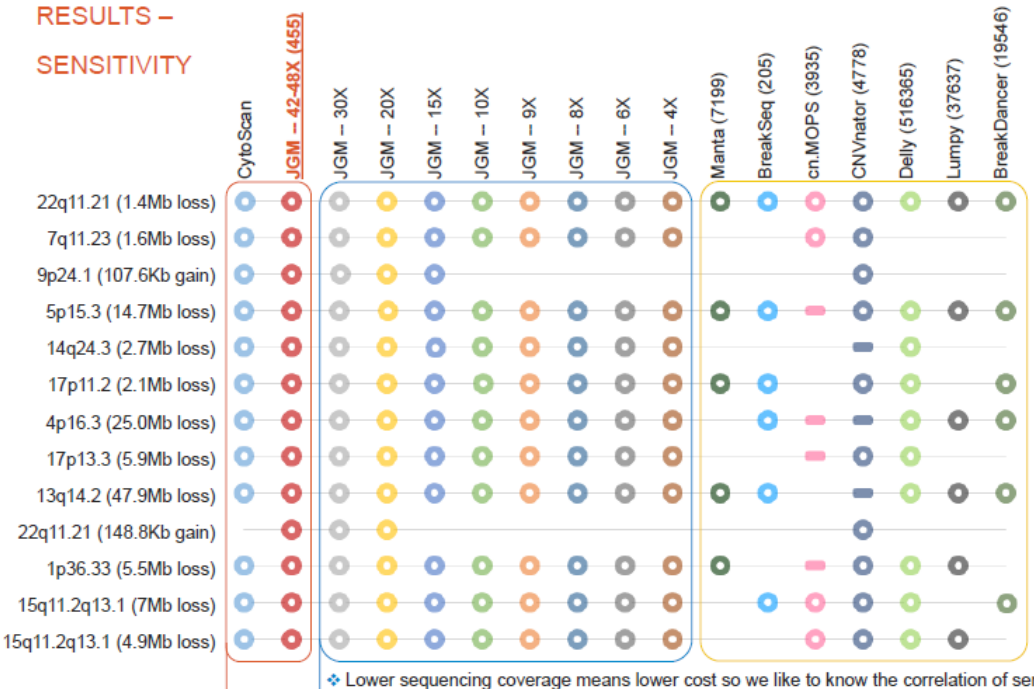


TEST SAMPLES & THEIR ASSOCIATED PATHOGENIC CNVS

Ten reference samples were selected as a benchmark and the 13 associated pathogenic CNVs were confirmed and reported by the Coriell Institute for Medical Research. The sizes of CNVs range from 107.6Kb to 47.9Mb while types are two gains and 11 losses.

Coriell IDs	Clinical Disorder	Pathogenic CNV		
		Location	Type	Size
GM05876	DiGeorge Syndrome	22q11.21	Loss	1.4Mb
GM13480	Williams Syndrome	7q11.23	Loss	1.6Mb
		9p24.1	Gain	107.6Kb
GM16593	Cri-du-chat Syndrome	5p15.3	Loss	14.7Mb
		14q24.3	Loss	2.7Mb
GM20743	Smith-Magenis Syndrome	17p11.2	Loss	2.1Mb
GM22601	Wolf-Hirschhorn Syndrome	4p16.3	Loss	25.0Mb
GM09209	Miller-Dieker Lissencephaly Syndrome	17p13.3	Loss	5.9Mb
GM14164	Tetralogy Fallot	13q14.2	Loss	47.9Mb
		22q11.21	Gain	148.8Kb
GM22569	1p deletion Syndrome	1p36.33	Loss	5.5Mb
GM11516	Angleman Syndrome	15q11.2q13.1	Loss	7.0Mb
GM20375	Angleman Syndrome	15q11.2q13.1	loss	4.9Mb

RESULTS – SENSITIVITY

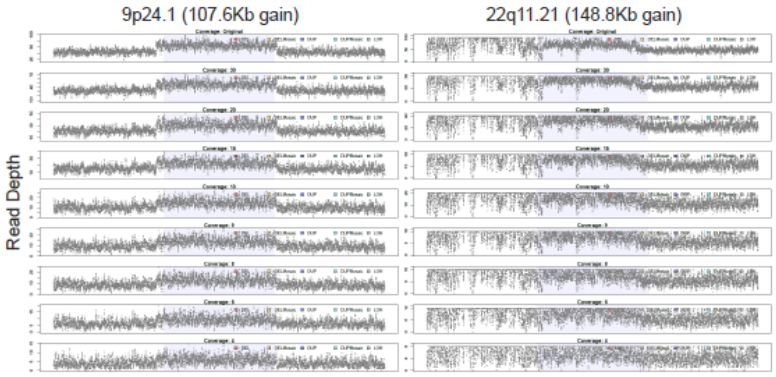


- The bars ▬ denote that reported CNVs are not 50% reciprocal overlapping with pathogenic ones, but recovered in manual review.
- All tools are performed by their default parameters on 42-48X BAM.
- The numbers in brackets show the total detected CNVs of each tool. Some tools try to report as many as possible to maintain sensitivity but it may cause difficulties for pathological diagnosis in clinic.

- Lower sequencing coverage means lower cost so we like to know the correlation of sensitivity and sequencing coverage.
- The sensitivity of deletions detection is not affected by sequencing coverage.
- The sensitivity of our newly developed NGS-based CNV detector is comparable with CytoScan (CMA-based).

SEQUENCING COVERAGE EFFECT

Two duplications are missing when we downsampled the sequencing coverage. Obviously duplication detection is more sensitive to sequencing coverage. Low coverage sequencing data leads to more noise and challenges of CNV detection.



FALSE DISCOVERY RATE ESTIMATION

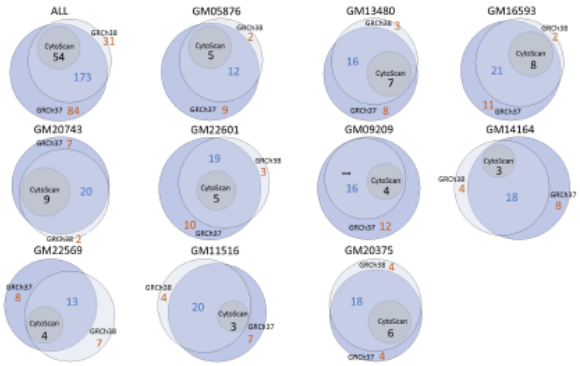
To estimate false discovery rate of our newly developed NGS-based CNV algorithm, we validated the four CNVs of GM05876 that are only detected by the NGS-based approach. Three of them are detected on both GRCh37 and GRCh38 while one is only detected on GRCh37. The one detected on GRCh37 is confirmed as false by Droplet Digital PCR (ddPCR).

Location	Type	Size (Kb)	GRCh37	GRCh38	ddPCR Validation
1p31	Loss	55.3	Yes	Yes	TRUE
4q35	Gain	140.5	Yes	No	FALSE
5q35	Loss	52.3	Yes	Yes	TRUE
12p11	Gain	57.7	Yes	Yes	TRUE

CYTOSCAN VS. NGS-BASED DETECTION GRCh37 VS. GRCh38 EFFECT

All CytoScan calling CNVs can be recalled by the newly developed NGS-based CNV algorithm.

GRCh38 as a more complete human reference genome and using it can help for filtering false CNVs.



CONCLUSIONS

- We developed a new CNV detector for the detection of chromosomal aberrations for pathological diagnosis due to no NGS-based tool was designed for clinical application.
- The sensitivity of the newly developed NGS-based tool is better than CytoScan (Considering Chromosomal Microarray CMA based) that is a standard assay at JGM CLIA which proves the potential of NGS as a first-tier cytogenetic diagnostic assay.

www.jax.org

CLAIMS

1. A computer implemented method for detecting copy number variations in a genome, comprising:

scanning at least one genome region to identify at least one autosomal chromosome;

performing a read depth calculation;

converting the read depth to a percentile representative of a coverage of each chromosome;

applying a hidden Markov model and a Poisson distribution to the percentile to provide at least one copy number variation status; and

filtering the at least one copy number variation status to identify at least one copy number variation in the genome.

2. A non-transitory computer readable storage medium, having computer readable instructions stored thereon that, when executed by a processor, cause the processor to execute a method to detect copy number variations in a genome, the method comprising the steps of:

scanning at least one genome region to identify at least one autosomal chromosome;

performing a read depth calculation;

converting the read depth to a percentile representative of a coverage of each chromosome;

applying a hidden Markov model and a Poisson distribution to the percentile to provide at least one copy number variation status; and

filtering the at least one copy number variation status to identify at least one copy number variation in the genome.