IDENTIFICATION AND CLINICAL ASSESSMENT OF DELETION STRUCTURAL VARIANTS IN WHOLE GENOME SEQUENCES OF ACUTELY ILL NEONATES

By

Aaron C. Noll

Submitted to the graduate degree program in Clinical Research and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Science

_____

Chairperson Edward F. Ellerbeck, MD, MPH

_____

Stephen F. Kingsmore, MB, BAO, ChB, DSc, FRCPath

_____

Peter G. Smith, PhD

Date Defended: April 7, 2014

The Thesis Committee for Aaron C. Noll
certifies that this is the approved version of the following thesis:

IDENTIFICATION AND CLINICAL ASSESSMENT OF DELETION STRUCTURAL VARIANTS IN WHOLE GENOME SEQUENCES OF ACUTELY ILL NEONATES

_____

Chairperson Edward F. Ellerbeck: Chair, MD

Date approved: April 17, 2014

ABSTRACT

Background

Effective management of acutely ill newborns with genetic conditions requires rapid and comprehensive identification of causative haplotypes. It has been previously shown that whole genome sequencing (WGS) can identify small variants contributing to the genetic illness of such patients in less than 50 hours. Deletion structural variants (SVs) >50 nucleotides are implicated in many genetic diseases and with WGS data can now be identified with a performance and timeframe sufficient for diagnosis in neonatal intensive care units. Here we describe the development of a solution that combines consensus calls from two SV detection tools (Breakdancer [BD] and GenomeStrip [GS]) with a novel filtering strategy.

Results

WGS simulation data demonstrated BD and GS consensus calls had 83% sensitivity and 99% positive predictive value with high precision. Through raw data inspection in the integrated genome viewer (IGV) consensus calls overlapping with SNP arrays were found to be 95% true positive and were subsequently used for filter parameterization. Consensus calling and filtering were implemented as a computational pipeline. IGV evaluation of pipeline results in a tetrad demonstrated calls were over 80% true positive but insensitive. Pipeline usage in 10 proband family sets revealed a possibly causative deletion SV in the MMP21 gene for two siblings. MMP21 is thought to play a role in embryogenesis in humans and may be responsible for the heterotaxy phenotype in humans. Further studies are needed to confirm these results.

Conclusions

The identification of deletion SVs has the potential to increase the diagnostic yield of WGS data. The methods described in this study may be useful in the research of disease detection in acutely ill neonates.

ACKNOWLEDGEMENTS

I would like to thank my mentor Stephen Kingsmore for the many helpful comments, dedication and direction in support of my master's thesis project. I would like to thank the many others at Children's Mercy for their efforts, without which this project could not have been completed. I am appreciative of my committee members who provided helpful insights. Finally, I am grateful to Paola, Isabella, Markus, my mother, and siblings for their enduring love and encouragement and above all the strength that comes from enduring faith in God.

TABLE OF CONTENTS

ABSTRACT…… iii
ACKNOWLEDGEMENTS…… iv
TABLE OF CONTENTS…… v
INTRODUCTION……... 1
Burden of genetic disease in neonates……. 1
Whole genome sequencing and types of mutations in genetic diseases……. 1
CNV characteristics……. 1
Deletion SVs known to cause genetic disease…… 2
Methods for diagnosis of genetic disease in neonates..…….. 2
Goals for CNV (SV) detection in WGS…… 3
RESULTS…… 4
Whole genome sequences of acutely ill neonates…..……. 4

Structural variation detection tool identification.…..……..... 4
Evaluation of tools using simulated Chromosome 1 deletions.…..……. 5
Whole genome simulations.…..……. 8
Evaluation of deletion prediction tools with whole genome simulations……..... 10

Evaluation of BreakDancer and GenomeStrip with experimental WGS…….…..... 11
Establishment of filtering criteria via WGS deletion SV predictions
overlapping with arrays……...……...……...……..... 12
Development of WGS deletion detection pipeline……...……..... 13
Analysis of precision in replicate sets before and after filtering……..……. 14
Analysis of deletion SV prediction segregation in trios and a tetrad……..… 16
Analysis of deletion SVs predicted to overlap with genes…… 18
Assessment of deletion SV predictions overlapping with exons and

identified to be only in proband……..... 19

Assessment of deletion SV predictions overlapping with exons and

identified to be in proband and one parent……..... 20

Extent and impact for deletions in 73 individuals by WGS…….. .. 23

MATERIALS AND METHODS…… 26
Study Participants…… 26
Whole Genome Sequencing……..… 26

Next Generation Sequencing Analysis..... …. 27

Affymetrix SNP/CNV array materials and methods..... 27

DISCUSSION……... 27
CONCLUSION……. 30
TABLES
Table 1. Ten software tools were evaluated for performance in detection
of deletion SVs in genome sequences.....…… 5
Table 2. Five of the 10 tools yielded SV predictions that overlapped a
simulated Chr 1 deletion by at least one nucleotide……..……..... 6
Table 3. Filter criteria derived from 112 likely true positive

BD∩$^{90\%}$GS ∩$^{90\%}$AR deletion predictions……..……..... 13

Table 4. Deletion predictions in three sets of WGS performed with each of samples UDT173 and pg96, showing the effect of filtering……..……...….. 16

Table 5. Statistics for deletion SV predictions from 10 trios that overlapped with exons…… 22

FIGURES

Figure 1. Simulated versus expected values for Chr 1 and whole genome simulation……..……..…….……..... 6

Figure 2. An example of a homozygous deletion SV (1:26,950,977-26,954,445) in the Chr1 simulation that features both reduced depth of coverage and stretched read pairs……..... 7

Figure 3. Heterozygous (Chr 2:84,271,069-84,277,071) and homozygous (Chr 11:73,476,280-73,478,163) deletion SVs from a WGS simulation, displayed in IGV…….…..... 9

Figure 4. Performance measures for five CNV detection tools as determined by reciprocal overlap of predictions from three iterations on one of three WGS simulations ……..……..…..... 11

Figure 5. A cartoon of the deletion detection pipeline…… 14

Figure 6. Comparison of deletion predictions in experimental WGS replicates before and after filtering…….. 15

Figure 7. Summary results for overlap of deletion SV predictions in 9 trios pre- and post-filtering…… 17

Figure 8. Venn diagram showing overlap of deletion predictions in a tetrad after filtering…… 18

Figure 9. Distribution of sizes for unique deletion SV predictions found in trios and overlapping with exons……..……. 21

Figure 10. Data shown for tetrad and presumed causative mutation MMP21 for siblings CMH184 and 185……. 22

Figure 11. BD∩$^{90\%}$GS deletion SV predictions in 73 samples were filtered and merged to achieve a unique set…… 24

Figure 12. Deletions per chromosome from unique set……..….. 25

Figure 13. A trio analysis flow diagram for isolation of putative disease causing deletions SVs…… 26

REFERENCES……..…….. 31

INTRODUCTION

Burden of genetic disease in neonates

Genetic conditions are uncommon individually but as a whole consume substantial healthcare resources.[20,21,22] Genetic diseases from chromosomal aberrations, and congenital malformations or deformations contribute to at least 1/5 of infant mortalities in the US.[25-27] Recent advances in genomics technologies and computational analysis have yielded unprecedented progress towards understanding the relationship of genomic variation to human morbidity and mortality.[23]

Whole genome sequencing and types of mutations in genetic diseases

By detecting disease causing single nucleotide (nt) variants (SNVs) and small nucleotide (< 50nt) insertions and deletions (indels), whole-genome sequencing (WGS) has been shown to accelerate and improve the sensitivity for the diagnosis of genetic illness in neonates.[25] In certain situations the causative mutations may be of classes other than small nucleotide variants and a patient will remain undiagnosed. A subset of such complex cases may be potentially explained through the impact of genomic structural variants (SVs). SVs (insertions, deletions, translocations and inversions > 50bp in length) are known to contribute to birth defects and other spontaneous traits, Mendelian diseases, and complicated genetic conditions.[24] Insertion and deletion SVs or copy number variants (CNVs) are the most common form of SV.[35]

CNV characteristics

It is estimated that the average diploid human genome differs from the human genome reference by ~5M SNVs (Kingsmore et al., unpublished) and one thousand CNVs >500nt [36] SNVs are thought to arise at a constant rate (~2.5×10^{-8} per nt per generation)[37] but CNV rates can vary widely at different loci (1.7×10^{-6} to 1×10^{-4} per locus per generation)[38]. CNVs are thought to cover 30% of the human genome versus < 2% for SNVs, albeit our knowledge of genomic CNV burden is primitive [34]. The mutation mechanism and selection pressure differ between insertions and deletions[31]. Meiotic deletion SV rates in human sperm at

4 hotspots were shown to be at least two-fold higher for deletions than insertions[1]. *De novo* CNV mutations arise via genomic rearrangements, in 'cis' (via intrachromosomal events) or in 'trans' (interchromosomal), and through non-allelic homologous recombination and non-homologous recombination. CNV length varies from a few hundred to several million nucleotides[39,40] and the size distribution for deletions is right skewed with smaller variants being the most frequent[34,31]. There is a strong purifying selection for deletions in exons and introns due to their potential for deleterious traits[31]. The impact of deletion SVs can range from having no discernible outcome to being incompatible with life, and the threshold for disease is most likely somewhere between these two extremes[35].

Deletion SVs known to cause genetic disease

Single gene deletions have been implicated in Duchenne Muscular Dystrophy, Neurofibromatosis type 1, Tuberous Sclerosis, Sotos syndrome, CHARGE syndrome, Spinal Muscular Atrophy, and Pelizaeus-Merzbacher Disease[35]. Medelian inheritance for deletions is seen in Williams-Beuren syndrome, Smith Magenis syndrome, Hereditary Neuropathy with liability to Pressure Palsy, Miller-Dieker lissencephaly, 22q11.2 deletion syndrome, Gaucher disease, Pituitary dwarfism, Thalassemia, Ichthyosis, juvenile Batten disease, and red green color blindness[34]. A high frequency of *de novo* deletion SVs is thought to be a risk factor for autism spectrum disorder[41] and rare deletion SVs have been associated with Schizophrenia[42]. Deletions near genes are thought to contribute to Crohn's disease, psoriasis and osteoporosis[31]. Routine tests are available for many of these conditions; however, the next logical step is to uncover the entire spectrum of deletion SVs for routine genotyping and to better understand the contribution of deletion SVs to human disease.[31]

Methods for diagnosis of genetic disease in neonates

Karyotyping and FISH (Fluorescent In Situ Hybridization) were used to generate initial views of common and rare genome structural variation[26]. Both techniques are restricted to small sample sizes and large SVs (0.5 to 5Mb) because of their low throughput and poor resolution.[28] Array comparative genomic

hybridization (array CGH) and single nucleotide polymorphism (SNP) arrays have been extensively used to discover CNVs[32]. Array CGH requires a test and reference sample to be labeled for comparative hybridization. CNV gains or losses are then inferred from this signal ratio. A single sample is first hybridized for SNP arrays, and at each SNP probe log ratios for clustering intensities are used to detect CNV gains or losses[33]. Both have low breakpoint resolution. Most arrays and analysis algorithms detect large CNVs equally well but unless custom designed for specific loci, are insensitive to events less than 10Kb[30]. WGS data can also be used to detect insertion and deletion SVs[30]. Paired End Mapping (PEM), Depth of Coverage (DOC), Split Read Mapping (SRM), and local assembly are the four principal methods used to detect SVs in WGS data[4]. The DOC approach identifies CNVs by interrogating read depth in sequential genomic windows that is greater or less than a predefined (e.g. using a parametric model) or dynamically determined background level. Paired ends are nucleotide sequences read from the ends of DNA fragments, as is typical for Illumina next generation sequencing. In the PEM model, pairs with a mapping distance congruent with the intended DNA fragment size and expected orientation are deemed concordant. Non-concordant pair mapping signatures are used to infer if an event is an insertion (mapping distance less than expected), deletion (mapping distance greater than expected), inversion (mapping orientation opposite to expected) or translocation (pairs map to different chromosomes). In PEM the maximum detectable insertion size is limited by the library fragment length but inversions, translocations and other complex rearrangements (e.g. fusion genes) are not detectable by arrays. The SRM method assumes CNV breakpoints occur within reads. At least one of the segments resulting from read bifurcation must align to a unique genome location. Of these four modes of SV discovery, whole genome assembly may hold the greatest, long-term promise for accurately typing all SV forms[34]. Unfortunately none of these approaches is comprehensive. In the interim, for a typical WGS sample, large proportions of validated SVs will be unique to each method. Although only DOC accurately predicts absolute gains or losses it does not resolve breakpoints well. PEM requires consistent fragment sizes and performs poorly in repetitive loci. Similarly, SRM is unreliable in non-unique areas, and short read sequence assembly is heavily biased to repeats and falls apart over such regions[30]. Numerous

permutations of these four paradigms and other novel methods have been implemented as computational tools yet there is a lack of consensus regarding which programs are the best at accurately detecting SVs.

Goals for CNV (SV) detection in WGS

Clinical grade SV detection is critical for the diagnosis of certain genetic disorders, and broadly for the assessment of missing causative haplotypes. Our goal was to formally compare existing tools designed to detect SVs in WGS data, and to develop a SV detection pipeline. We focused on the deletion subset of SVs since as a class they are thought to be the most numerous[1,2], deleterious[3], and readily detectable in paired end WGS data[4].

Recently we have been performing WGS of familial triads or tetrads in which the proband is an acutely ill neonate or infant receiving care in an intensive care unit. The goal of these studies has been to develop methods for molecular diagnosis in a time-frame consistent with clinical management decisions and that are comprehensive for genetic diseases. The focus of the current study was to expand the range of causative mutations to include deletion SVs.

RESULTS

Whole genome sequences of acutely ill neonates

74 WGS samples were analyzed. 14 were unrelated, 2 were a pair, 54 were trios, and 4 were part of a tetrad. One sample (NA12878) was used as the control in trio and tetrad comparisons. Human genome GRCh37.p5 was used as the reference version for alignments and simulations. Most samples were sequenced as 2 x100 nucleotide reads with a fragment size of 200-400nt and a mean read depth of 40X (40 fold).

Structural variation detection tool identification

A literature survey identified 50 whole genome sequence (WGS) structural variation (SV) software programs. Ten of these did not require substantial effort for installation or execution, did not require a control sample, supported the widely used .bam format[11] for genome sequence data, and could be run concurrently on multiple processors (Table 1). The performance of the ten tools for SV deletion identification was evaluated using simulated genomic data.

	Publication	Method	Chr 1 sim	WGS sim
Breakdancer	Chen et al. 2009	PEM	-	x
Clever	Marschall et al. 2012	Read alignment graph and max cliques	x	
Cn.MOPS	Klambauer et al. 2011	DOC and Poission distribution		
Control-freec	Boeva et al. 2012	SNP B allele frequencies and DOC	x	
Dindel	Albers et al. 2011	Realignment with probablistic indel calls		
ERDS	Zhu et al. 2012	DOC and paired HMM	x	
GasvPRO	Sindi et al. 2012	DOC, PEM and probabalistic model		
Genomestrip	Handsaker et al. 2011	DOC, PEM and SRM	x	x
Lumpy	Layer et al. 2014	PEM and DOC (SRM with special aligner)	x	
SVDetect	Zeitouni et al. 2010	DOC, PEM		

Table 1. Ten software tools were evaluated for performance in detection of deletion SVs in genome sequences. ('x' indicates a tool passed a simulation evaluation. '–' indicates Breakdancer did not meet criteria to pass the Chr1 sim but was retained on account of prior experience with the tool on a different project. [DOC – depth of coverage, PEM – paired end mapping, SRM – Split read mapping, HMM – hidden markov model, SNP – single nucleotide polymorphism])

Evaluation of tools using simulated Chromosome 1 deletions

Tool performance was initially evaluated using synthetic data. 270 homozygous deletion SVs of size 500 – 10,000 nucleotides (nt) were created in a representation of human chromosome (Chr) 1 with 40X coverage by 2x100nt paired reads. Reads were simulated from this modified Chr 1 GRCh37.p5 sequence file using wgsim 0.3.0[11] (with default parameters). Simulated reads were aligned to the human reference

GRCh37.p5 using GSNAP[10] version 2012-07-12, and sam files were converted to the bam form using samtools[11] 0.1.18.

Read depth, repetitive regions, and GC content can influence the accuracy of deletion predictions.[9] To ensure that the simulated data set was comparable to the reference sequence, values for these attributes in the simulated chromosomes were compared to the Chr 1 reference. GC content and repeat feature frequency for the simulated deletions differed < 10% from the Chr 1 reference, and simulation mean read depth was found to be nearly identical to the target 40X (Figure 1).

When visually inspected in the integrated genome viewer (IGV), simulated deletions were found to have stretched read pairs spanning breakpoints with uniform inner and outer read depths which are critical for the PEM and DOC detection methods, respectively (Figure 2). Together these findings demonstrated that the simulated Chr 1 dataset was suitable for an initial evaluation of deletion SV detection tools.

Five of the 10 tools yielded SV predictions that overlapped a simulated Chr 1 deletion by at least one nucleotide (genome coordinate overlaps were determined by standard Linux utilities and bedtools 2.17.0[12]; Table 2). Breakdancer was retained due to prior success with the tool on another project. These six tools were evaluated further.

Tool	TP	FP	FN	SENS	PPV	F1
Breakdancer	0	82	270	0	0	0
Clever	32	1683	238	0.118519	0.018659	0.032242
Controlfreec	5	449	265	0.018519	0.011013	0.013812
ERDS	149	1204	121	0.551852	0.110126	0.183611
GenomeStrip	146	673	124	0.540741	0.178266	0.268136
Lumpy	247	526524	23	0.914815	0.000469	0.000937

Table 2. Five of the 10 tools yielded SV predictions that overlapped a simulated Chr 1 deletion by at least one nucleotide.

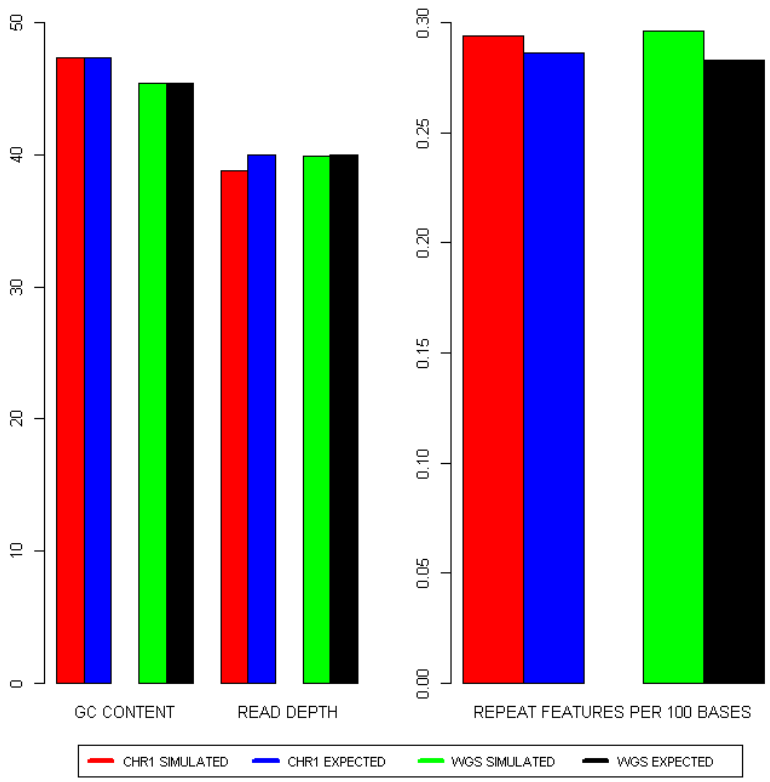

Figure 1. Simulated versus expected values for Chr 1 and whole genome simulation. Substantial simulation bias for GC content, repeat features and mean depth all have the potential to influence tool prediction performance. Curated repetitive regions found by RepeatMasker[5], Tandem Repeat Finder[6], and Dust[7] for Chr 1 were obtained from Ensembl[8] and used to calculate repeat feature frequency. Less than 10% difference was observed between expected and simulated values.

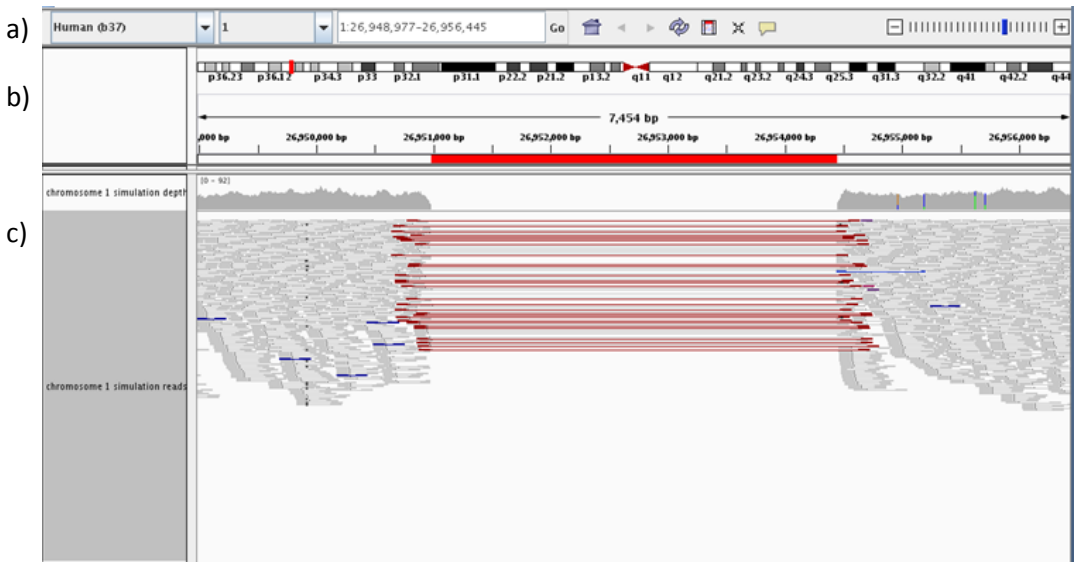

Figure 2. An example of a homozygous deletion SV (1:26,950,977-26,954,445) in the Chr1 simulation that features both reduced depth of coverage and stretched read pairs. The top panel (a) shows the currently chosen reference (GRCh37), chromosome (1) and coordinates for the chromosome region being displayed. The middle panel (b) has a chromosome ideogram with a red box indicating the section of the chromosome being displayed. The ruler reflects the visible portion of the chromosome. Tick marks indicate chromosome locations. The span lists the number of nucleotides currently displayed. The bottom panel (c) shows depth and read tracks. Read pairs are displayed as dyads of rectangles each joined by a line. Pairs with normal separation are shown in grey. Stretched pairs are shown in red. Compressed pairs are shown in blue.

Whole genome simulations

There is not yet a "gold standard" set of known deletion SVs for a reference WGS[13,14]. However, the 1,000 genomes project (1KGP) structural variation analysis group (SVAG) has constructed a putative and partially validated SV deletion set from 185 human genomes[15]. To evaluate the six tools at genome scale, three WGS samples with known deletions were simulated with parameters derived from SVAG analyses. Deletion SVs were simulated with random length (from 600 to 8000 nt) and intra-chromosomal placement, while being distributed proportionally to chromosomes by their size at a rate of one per 400Kb (~7500 per sample). Previous WGS experience was used to establish rates for SNPs (1 per Kb), small insertions and deletions (0.1 per Kb), and nucleotide errors (5 per Kb). Sequence fragment size was 400nt, read length was 2 x 100 nt, and read depth was 40X. Read simulation, alignment, and bam file creation were as before. Differences between WGS simulation and expected genome reference values for GC content, repetitive feature frequency and target depth were < 10%, similar to the Chr 1 simulation (Figure 1). Three independent WGS samples were simulated to reduce potential tool deletion position advantages occurring by chance via random placement. To compare sensitivity for homozygous and heterozygous deletion SVs, the simulated deletion SVs were 98% heterozygous and 2% homozygous. A subset of deletion SVs were visually inspected for each sample using IGV (Figure 3).

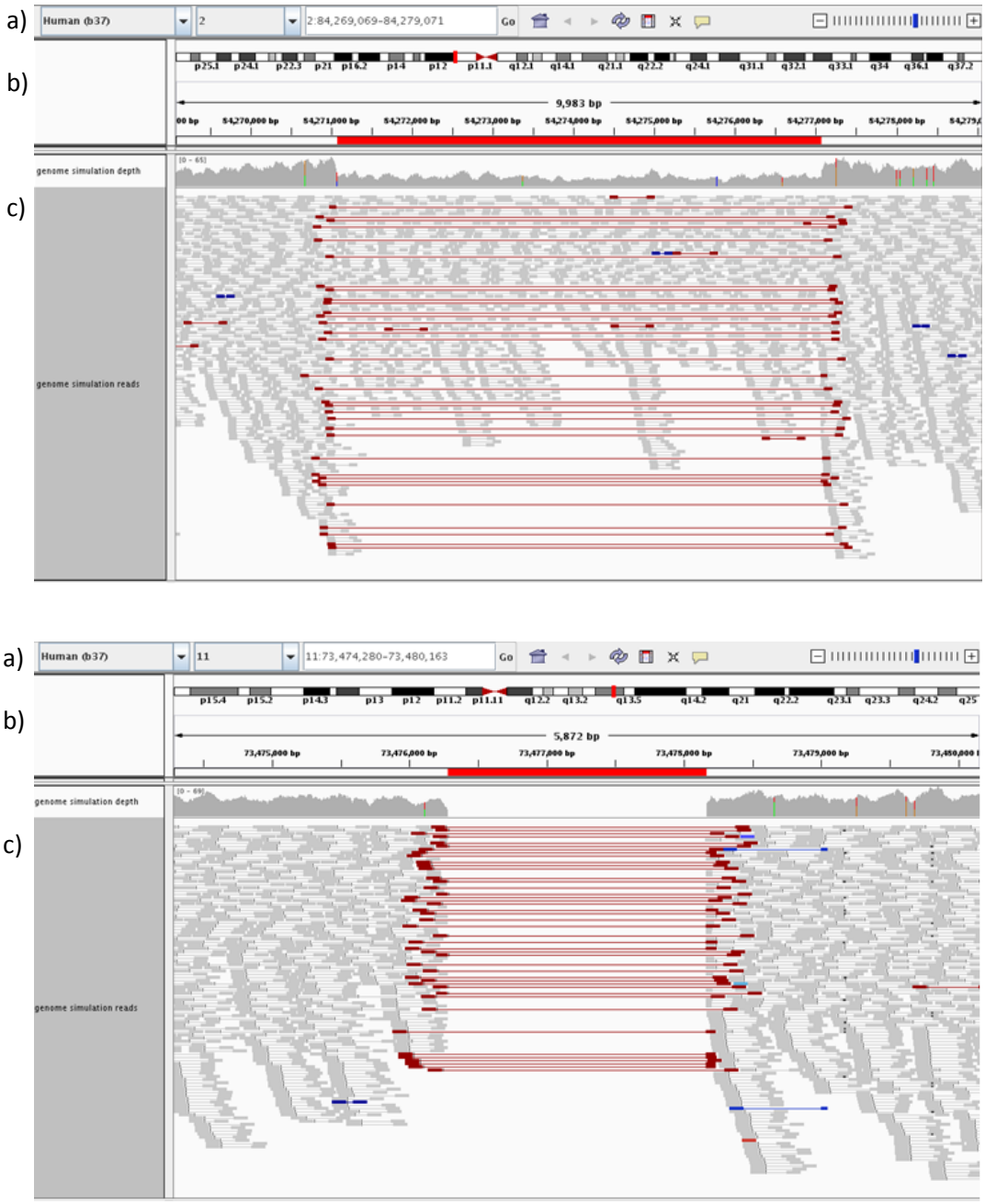

Figure 3. Heterozygous (Chr 2:84,271,069-84,277,071) and homozygous (Chr 11:73,476,280-73,478,163) deletion SVs from a WGS simulation, displayed in IGV. Most WGS simulated deletion SVs had stretched pairs with appropriate read depth. The top panel (a) shows the currently chosen reference (GRCh37), chromosome (2;11) and coordinates for the chromosome region being displayed. The middle panel (b) has a chromosome ideogram with a red box indicating the section of the chromosome being displayed. The ruler reflects the visible portion of the chromosome. Tick marks

indicate chromosome locations. The span lists the number of nucleotides currently displayed. The bottom panel (c) shows depth and read tracks. Read pairs are displayed as dyads of rectangles each joined by a line. Pairs with normal separation are shown in grey. Stretched pairs are shown in red. Compressed pairs are shown in blue.

Evaluation of deletion prediction tools with whole genome simulations

Nine evaluation sets, comprising three iterations for each of the three independent samples, were attempted for each SV deletion detection tool. ERDS failed repeatedly and was removed from further analysis. Tool predictions were compared to simulated deletion SVs (depicted as set notation where intersection is $\cap$, not intersecting is $\overline{\cap}$, and union is U) at 6 discrete reciprocal overlap values (1bp, 1%, 25%, 50%, 90% and 99%) since, to our knowledge, no standard SV coordinate overlap criteria yet exist. Performance measures were true positives (TPs), false positives (FPs), false negatives (FNs), sensitivity (SENS), positive predictive value (PPV) and F1-measure. With an unknown quantity of true negatives the F1-measure substituted for specificity. TPs, FPs and FNs were counted and SENS, PPV, and F1 were calculated for each tool (SENSITIVITY = TP/(TP+FN); PPV= TP/(TP+FP); F1 = 2*(SENS*PPV)/(SENS+PPV)). Tools were iteratively executed for each sample three times to evaluate the precision of predictions. The difference in sensitivity between homozygous and heterozygous deletion predictions was less than 1%.

Tool performance measures were obtained for the combined homozygous and heterozygous set for each of the simulated WGS samples. Breakdancer (BD) and GenomeStrip (GS) exhibited excellent sensitivity, F1, precision and PPV (Figure 4). In general, prediction performance decreased as the reciprocal overlap criterion increased, although BD and GS performance was excellent from 1bp - 90% overlap ($BD \cap^{90\%} GS$, Figure 4). Lumpy's performance dropped significantly at overlap criteria > 50%. Control-Freec predicted many large deletions that had poor overlap with TPs. Clever had the most imprecision (variability between iterations) and consistently predicted the fewest deletions. WGS simulation results indicated that TP deletion SVs were detected by PEM alone (BD), although combining PEM, DOC, and

SRM (GS) was optimal. Compared to BD or GS alone, a 90% (BD∩$^{90\%}$GS) prediction overlap (reciprocal overlap was required for all comparisons) was 3.8% less sensitive but yielded a 3.3% reduction in the ratio of FPs to TPs. We concluded that BD and GS exhibited the best deletion detection in simulated WGS, and that the subset of BD and GS predictions with 90% overlap (BD∩$^{90\%}$GS) appeared to achieve an optimal balance of sensitivity and specificity.

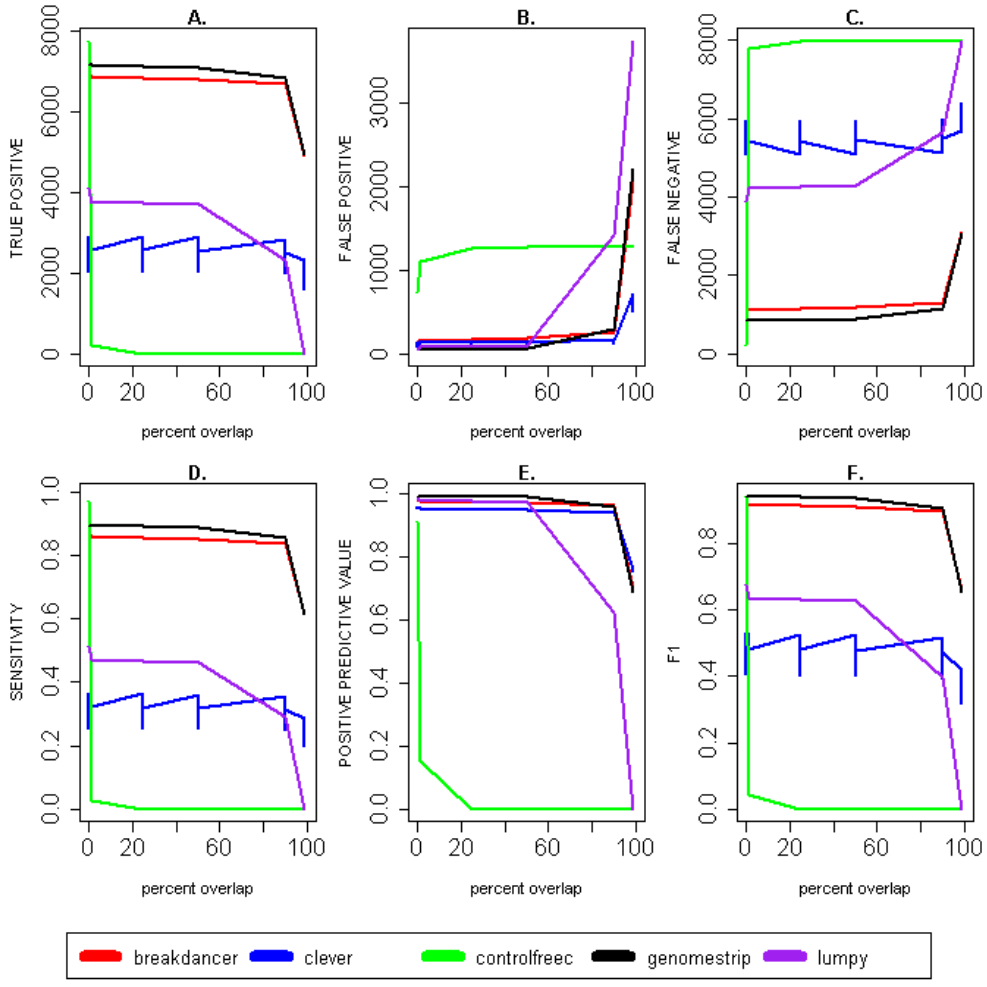

Figure 4. **Performance measures for five CNV detection tools as determined by reciprocal overlap of predictions from three iterations on one of three WGS simulations.** Shown are WGS simulation results as measured by TPs (A), FPs (B), FNs (C), SENS (D), PPV (E), and F1 (F). BD and GS had the best performance overall and were selected for analysis of experimental WGS. Similar results were observed for two other WGS simulations. (Supplementary figures S1 and S2)

Evaluation of BreakDancer and GenomeStrip with experimental WGS

NA12878, a HapMap CEU trio child[16] was chosen for initial experimental testing, as it had been extensively sequenced by the 1KGP[17], and has been selected as a future SV benchmark by the National Institutes of Standards and Technology[18]. We generated an experimental NA12878 dataset comprising a 2 x 100 nt, 40X WGS. BD and GS WGS deletion predictions were compared with two Affymetrix SNP array technical replicates (array replicate 1 [AR_1] and array replicate 2 [AR_2]), for which we also generated NA12878 experimental data. While it had been assumed that the SNP array data would serve as a 'gold standard' deletion SV set, the $AR_1 \cap^{90\%} AR_2$ replicate overlap was only 22% of $AR_1 U AR_2$ (67/(131+175)). There were significant differences in quality metrics between the array replicates, which could explain this lack of overlap. Thus it did not appear that Affymetrix SNP array data had sufficient precision to be a 'gold standard' comparator.

Nevertheless, we compared $BD \cap^{90\%} GS$ predictions to array replicate predictions ($AR_1 \cap^{90\%} AR_2$) using a 10% overlap cutoff ($(BD \cap^{90\%} GS) \cap^{10\%} (AR_1 \cap^{90\%} AR_2)$). Predictions were further validated by visual inspection in IGV. We observed a TP ratio of 39% for BD alone (n=28), 63% for GS alone (n=24), 67% for $BD \cap^{90\%} GS$ (n=21), and 95% for $(BD \cap^{90\%} GS) \cap^{10\%} (AR_1 \cap^{90\%} AR_2)$ (n=16). Although GS had outperformed BD only slightly on simulated WGS, this difference was substantial for NA12878 WGS. Consistent with simulated WGS, NA12878 WGS $BD \cap^{90\%} GS$ slightly outperformed GS alone, and appeared possibly to have sufficient specificity for use as a screening test. However, the much larger number of deletions predicted by GS or BD in WGS than by array hybridization suggested the potential for substantial FPs. Therefore we sought to develop a quality filter that retained TPs while reducing possible FPs.

Establishment of filtering criteria via WGS deletion SV predictions overlapping with arrays

Our filtering strategy focused on common WGS deletion characteristics used by detection algorithms, tool-generated quality metrics, and knowledge gained from NA12878 WGS FP visualizations. The depth of coverage (DOC) model assumes deletions have significantly less read depth than flanking regions.

Paired end mapping (PEM) predicts a deletion when distances between aligned read pairs bordering the deletion were greater than expected[19]. When viewed in IGV, many FPs were very large or associated with ineffective unique read mapping to repetitive regions. For each prediction, GS provided a breakpoint confidence interval (GSCI [0 to #] where lower is better) and BD generated a quality score (BDS [0-99] where higher is better) that appeared to be useful as quality measures. Thus, a novel quality filter was developed that combined a maximum depth ratio (ratio of read depth within a deletion to that flanking the deletion), GSCI, overlapping repeat feature count and prediction size, with minimum cutoff values for BDS and quantity of supporting pairs. Based on the results from NA12878 WGS $(BD \cap^{90\%} GS) \cap^{10\%} (AR_1 \cap^{90\%} AR_2)$, we assumed a high TP ratio for all $BD \cap^{90\%} GS \cap^{95\%} AR$. Thus, filter values were parameterized on $BD \cap^{90\%} GS \cap^{95\%} AR$ predictions (n=112) from sequencing and array assays for 14 samples (Table 3).

Filter attribute	Parameterized value
Max depth ratio:	0.7
Max GS CI:	40
Min BD Score:	30
Max deletion prediction size:	500,000 nt
Min supporting read pairs:	2
Max overlapping repeat features:	1500

Table 3. Filter criteria derived from 112 likely true positive $BD \cap^{90\%} GS \cap^{90\%} AR$ deletion predictions.

Development of WGS deletion detection pipeline

BD, GS, filter parameters, and overlapping genomic features were incorporated into a computational pipeline that ran in parallel starting from a bam file (Figure 5). Computation was completed on most WGS samples within eight hours. Coverage plots of deletion predictions were automatically generated with IGV to facilitate initial validation by visual inspection.

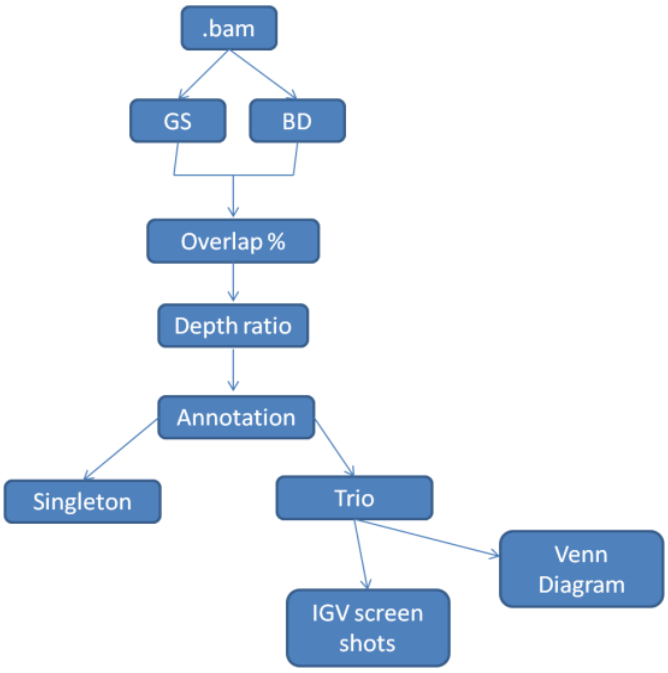

Figure 5. A cartoon of the deletion detection pipeline. GS and BD are executed concurrently on a bam file. Filter attributes, overlap %, and annotations are obtained for each prediction. For family sets additional steps are performed.

Analysis of precision in replicate sets before and after filtering

Three datasets were used to assess the effects of filtering on the sensitivity and specificity of BD$\cap^{90\%}$GS. The first was an assessment of run-to-run precision of deletion predictions in two samples (UDT173 and pg96), each with three WGS experimental replicates (r1, r2, and r3). While the replicates utilized the same sample, alignment and prediction tools, the sequencing methods differed among them. UDT173, for example, was sequenced with: r1. NextSeq 500 2 x 120 nt reads with version 4 (v4) chemistry, r2. HiSeq 2500 2 x 100 nt reads with v3 chemistry and a 26 hour recipe, and r3. HiSeq 2500 2 x 100 nt reads with v4 chemistry and an 18 hour recipe. Pg96 was sequenced with: r1. HiSeq 2500 2 x 250 nt reads r2. HiSeq 2500 2 x 120 nt reads, 11 day protocol, and v3 chemistry, and r3. HiSeq 2500 2 x 100 nt reads, v4 chemistry and 18 hour recipe. Given these methodological differences, we expected greater similarity for r2$\cap^{90\%}$r3 than for r3$\cap^{90\%}$r1 or r2$\cap^{90\%}$r1, which matched actual results (Figure 6 and Table 4). Application of the filters decreased the total number of deletion predictions by 11.7-fold and 3.2 fold in UDT173 and

pg96, respectively (Table 4). In contrast, filtering decreased the r1∩$^{90\%}$r2∩$^{90\%}$r3 class by only 2.2-fold and 1.3-fold, respectively, indicating that filtering improved prediction quality (Table 4).

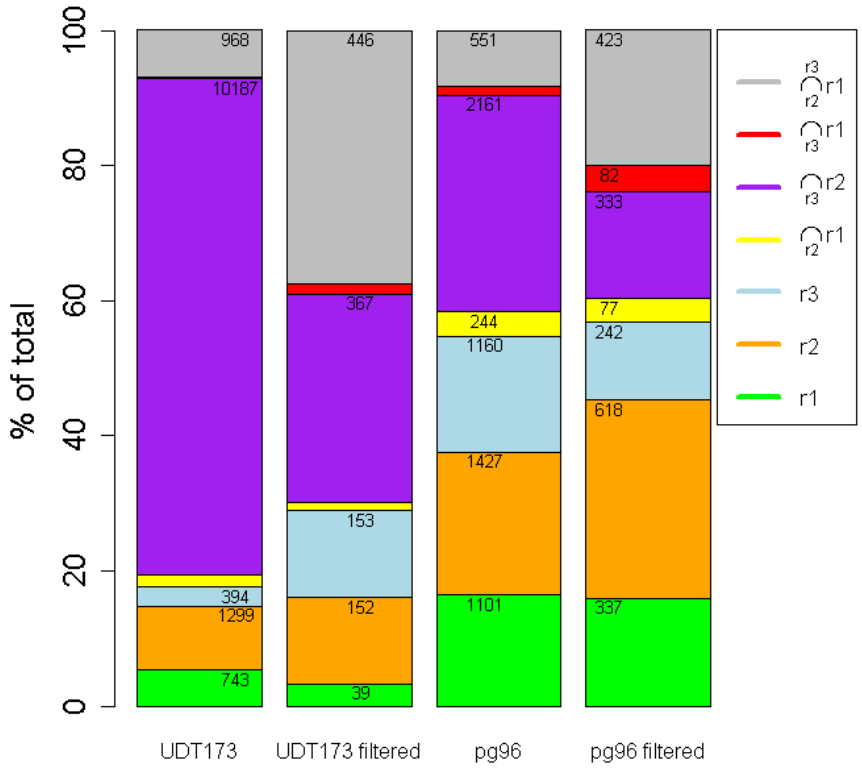

Figure 6. Comparison of deletion predictions in experimental WGS replicates before and after filtering. Filtering resulted in the greatest increase in the r1∩$^{90\%}$r2∩$^{90\%}$r3 class as a proportion of total.

Sample	r1	r2	r3	r1∩$^{90\%}$r2	r2∩$^{90\%}$r3	r1∩$^{90\%}$r3	r1∩$^{90\%}$r2∩$^{90\%}$r3
UDT173	743	1,299	394	266	10,187	24	968
UDT173 filtered	39	152	153	15	367	19	446
pg96	1,101	127	1,160	244	2,161	87	551
pg96 filtered	337	618	242	77	333	82	423

Table 4. Deletion predictions in three sets of WGS performed with each of samples UDT173 and pg96, showing the effect of filtering.

Analysis of deletion SV prediction segregation in trios and a tetrad

The vast majority of nuclear genome variations should obey the rules of Mendelian inheritance in WGS of familial trios and tetrads. Thus, inheritance information is very useful in assessing the veracity of deletion predictions. The segregation of predicted deletion SVs was assessed in WGS of one tetrad and 9 trios, each comprising an acutely ill infant proband and both parents. BD$\cap^{90\%}$GS for parent 1 (p1), parent2 (p2), and proband (pb) (with family member 1 (f1) or affected sibling (pb2) for tetrads) were compared to identify familial deletion SV segregation. IGV inspection of several thousand filtered BD$\cap^{90\%}$GS deletion SV predictions indicated that those with 0.1% reciprocal overlap in a trio were overwhelmingly identical, and therefore were merged. For example [pb (BD$\cap^{90\%}$GS)]$\cap^{0.1\%}$[p1(BD$\cap^{90\%}$GS)] $\cap^{0.1\%}$[p2(BD$\cap^{90\%}$GS)] were designated as trio shared predictions. Furthermore, IGV visualization of 1,384 deletion SV predictions in a familial tetrad revealed 100% Mendelian inheritance of genotypes (Figure 8). Thus deletions that did not exhibit Mendelian inheritance in trios represented false negative categorizations that stemmed from tool insensitivity. Filter constraints (depth ratio, GSCI or BDS), overlap requirements or lack of a deletion SV prediction from one or both tools contributed most to insensitivity.

In 9 trios filtering increased the total proportion for pb$\cap$p1$\cap$p2, pb$\cap$p1, and pb$\cap$p2 by 8.95%, 1.92%, and 4.97% but decreased pb unique prediction by 15.84% respectively. These results are concordant with expected Mendelian inheritance where the majority of deletions will be shared with parents and few will be *de novo*. (Figure 7). Gratifyingly, 78% of 1,384 filtered deletion predictions in a tetrad were true positives, as assessed by IGV. IGV tetrad analysis demonstrated a TP% and SENS estimate for pb1$\cap^{90\%}$pb2$\overline{\cap}^{90\%}$p1$\overline{\cap}^{90\%}$p2 to be 72% and 50% respectively. For pb1$\cap^{90\%}$pb2$\cap^{90\%}$p1$\cap^{90\%}$p2 TP% and SENS estimates were 67% and 25% respectively. The TP% for pb1 alone was 81% (Figure 8). Although filter and overlap criterion reduced FPs for each individual sample, insensitivity yielded FNs which established incorrect familial associations for deletion SV predictions.

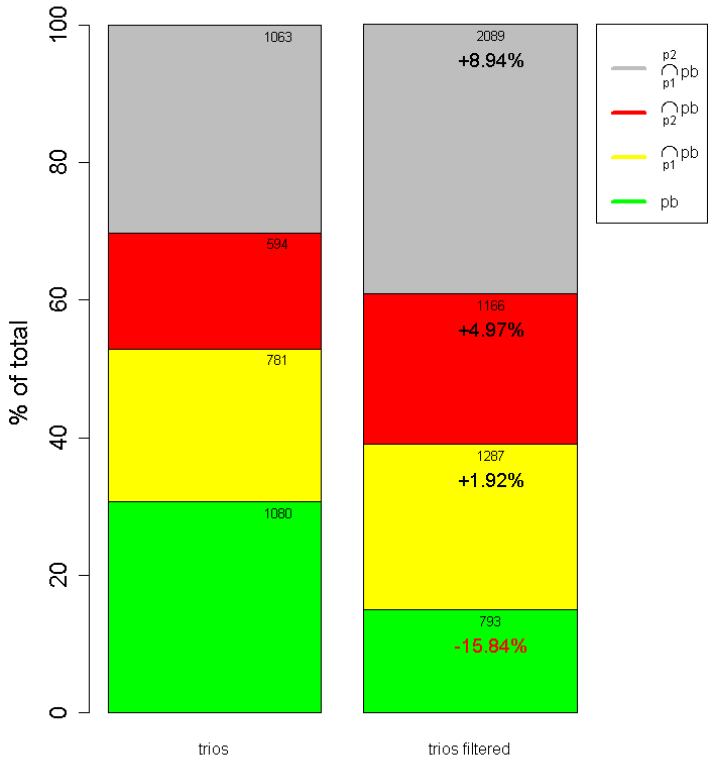

Figure 7. Summary results for overlap of deletion SV predictions in 9 trios pre- and post-filtering

Filtering increased the total proportion for pb∩p1∩p2, pb∩p1, and pb∩p2 by 8.95%, 1.92%, and 4.97% but decreased pb unique predictions by 15.84% respectively. Each trio had an acutely ill infant proband.

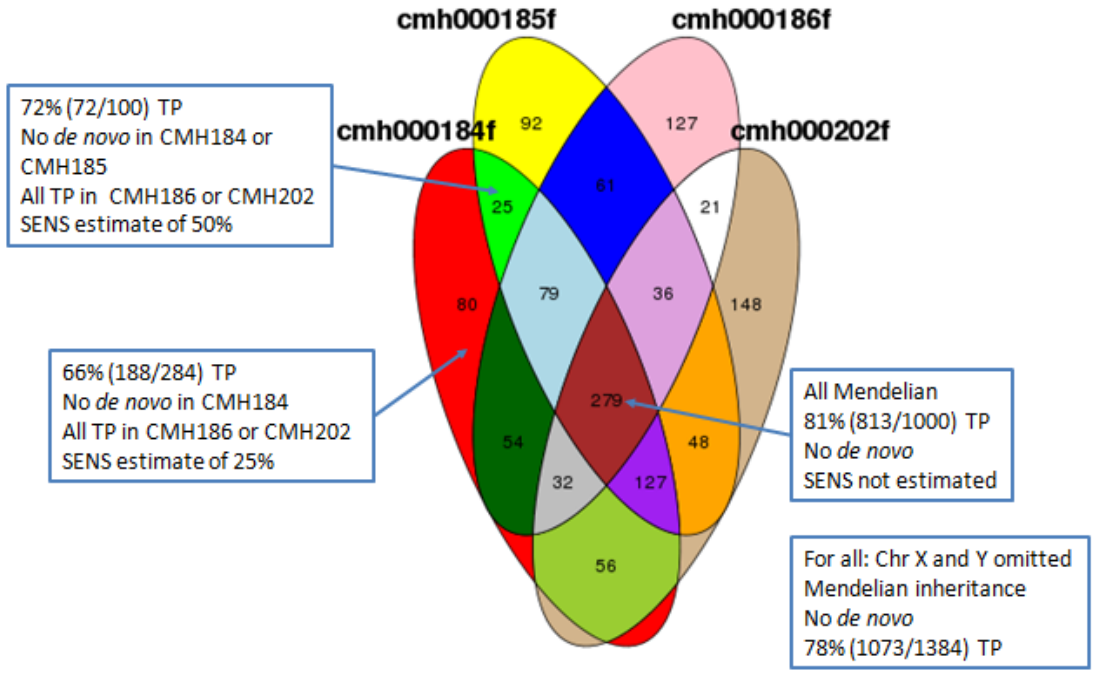

Figure 8. Venn diagram showing overlap of deletion predictions in a tetrad after filtering.

CMH000184 and CMH000185 were affected sibling stillbirths. CMH000186 and CMH000202 were parents. TP rates were promising but sensitivity was lower than desired. (f = filtered)

Analysis of deletion SVs predicted to overlap with genes

We next sought deletion SVs that overlapped genes and were likely to be associated with disease phenotypes. Of 8872 filtered BD$\cap^{90\%}$GS deletions identified in WGS of 30 individuals, 2,119 overlapped genes. They ranged in size from 563 - 484,514 nt and affected 0 - 373 exons each. On average, an individual WGS sample had 71 gene-associated deletion SVs, encompassing 701,233 nt, that affected a total of 226 exons. The 30 individuals represented 10 trios, each with an acutely ill infant proband. After merging overlapping deletions in each of the ten familial trios, 1,165 unique deletion SVs remained. Proximity of start and stop sites, however, indicated that many of these were recurrent among trios. After merging recurrent deletions, 466 distinct deletions remained. Of these, 281 were unique to single trios (Table 5). A frequency distribution of these 281 deletion SVs demonstrated that most ranged between 1Kp to 10Kbp in size (Figure 9). Since rare diseases cannot be caused by common mutations, the subset

of those 281 deletions that was present in each proband contained candidates for disease causality for each proband. These 281 deletions were visually inspected with IGV and zygosity/TP status in the proband and inheritance pattern in the trio were recorded (TP=84%; Hom=0%; all Mendelian). Pseudogenes, olfactory or taste receptors, and MHC genes were omitted. Finally, the biological plausibility of the genes for the phenotype was assessed.

Assessment of deletion SV predictions overlapping with exons and identified to be only in proband

Deletions were assessed in IGV and identified as occurring in either the proband alone, in the proband and parent or sibling, or in the proband and both parents. If an exon deletion was identified as heterozygous, that gene was further interrogated to determine if any mutations (e.g. SNPs, small insertions, or deletions) might be present on the remaining allele.

All 50 exon overlapping deletion SVs occurring in the proband were heterozygous. CMH222 and CMH223 were unique in that the trio consisted of two affected siblings and a parent. Two exonic heterozygous deletion SV predictions were found in both siblings but not in the parent. The first was in the *OSBPL2* gene which encodes an oxysterol binding protein with unknown function. The second was found in *SNORD116-12* which has been implicated in Prader-Willi syndrome although phenotypes for CMH222 and CMH223 did not match this disorder. No additional mutations were identified in CMH222 and CMH223 for either gene.

A heterozygous deletion of exon 3 in the *RCC1* gene was identified in the proband CMH725 and one parent. No additional mutations were identified in CMH725 for RCC1. A heterozygous deletion overlapping the terminal exon of *CTC-260E6* and *ZNF486* was found in the proband CMH531 and both parents. No additional mutations were identified for CMH531 in either gene.

Two heterozygous deletions were identified in the proband CMH436 and parents. The first deleted the entire coding sequence of the heat shock protein *HSP90AA4*. The second impacted a region of

chromosome 1 encompassing the genes *MSTO1*, *MSTOP2P* and *GON4L*. No additional mutations were identified for CMH436 in these genes.

Assessment of deletion SV predictions overlapping with exons and identified to be in proband and one parent

Of 79 putative deletion SVs predicted to exist in the proband and one parent, 6 were of interest. A heterozygous deletion of exon 1 in *GRID2IP* was identified in CMH222, 223 and 224. No additional mutations were identified for either sibling in this gene.

A heterozygous deletion in *RGS17* was found in CMH396 and one parent. No additional mutations were identified for CMH396 in this gene.

A heterozygous deletion of *MMP21* gene was identified in CMH184, CMH185, and the unaffected parent CMH186. Interestingly, a small deletion induced frameshift was also identified in *MMP21* for CMH184, CMH185, and the other unaffected parent CMH202 (Figure 10). *MMP21* encodes a matrix metalloproteinase and is reported to be involved in the normal physiologic breakdown of the extracellular matrix during embryonic development and tissue remodeling[45]. Mice homozygous for an ENU-induced mutation in MMP21 have been reported to exhibit heterotaxy and related cardiovascular defects[46] which is a phenotype similar to that of both CMH184 and 185.

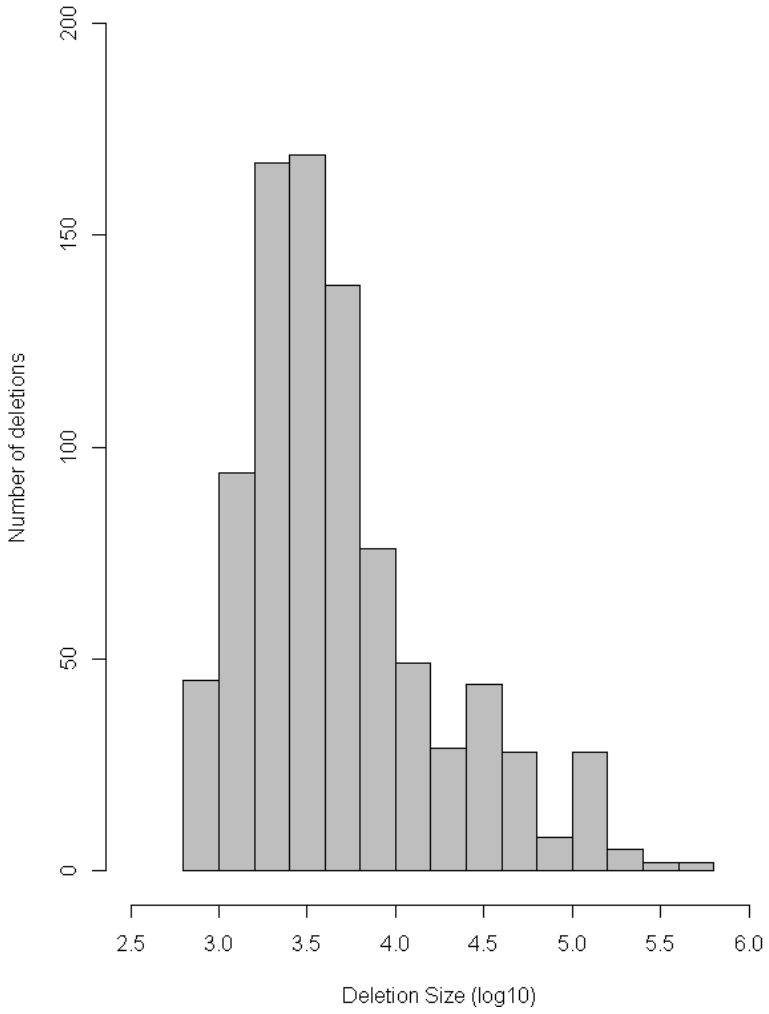

Figure 9. Distribution of sizes for unique deletion SV predictions found in trios and overlapping with exons.

Feature	Nucleotides	Exons	Repeats	Number
# predicted filtered deletions (merged by trios)				1,165
# predicted filtered deletions	21,037,001	6,770		2,119
# unique deletions (present in only 1 trio)				281
Max	484,514	373	1,404	30

Min	563	0	0	1
Average	701,233	226		71
average size	9928			

Table 5. Statistics for deletion SV predictions from 10 trios that overlapped with exons

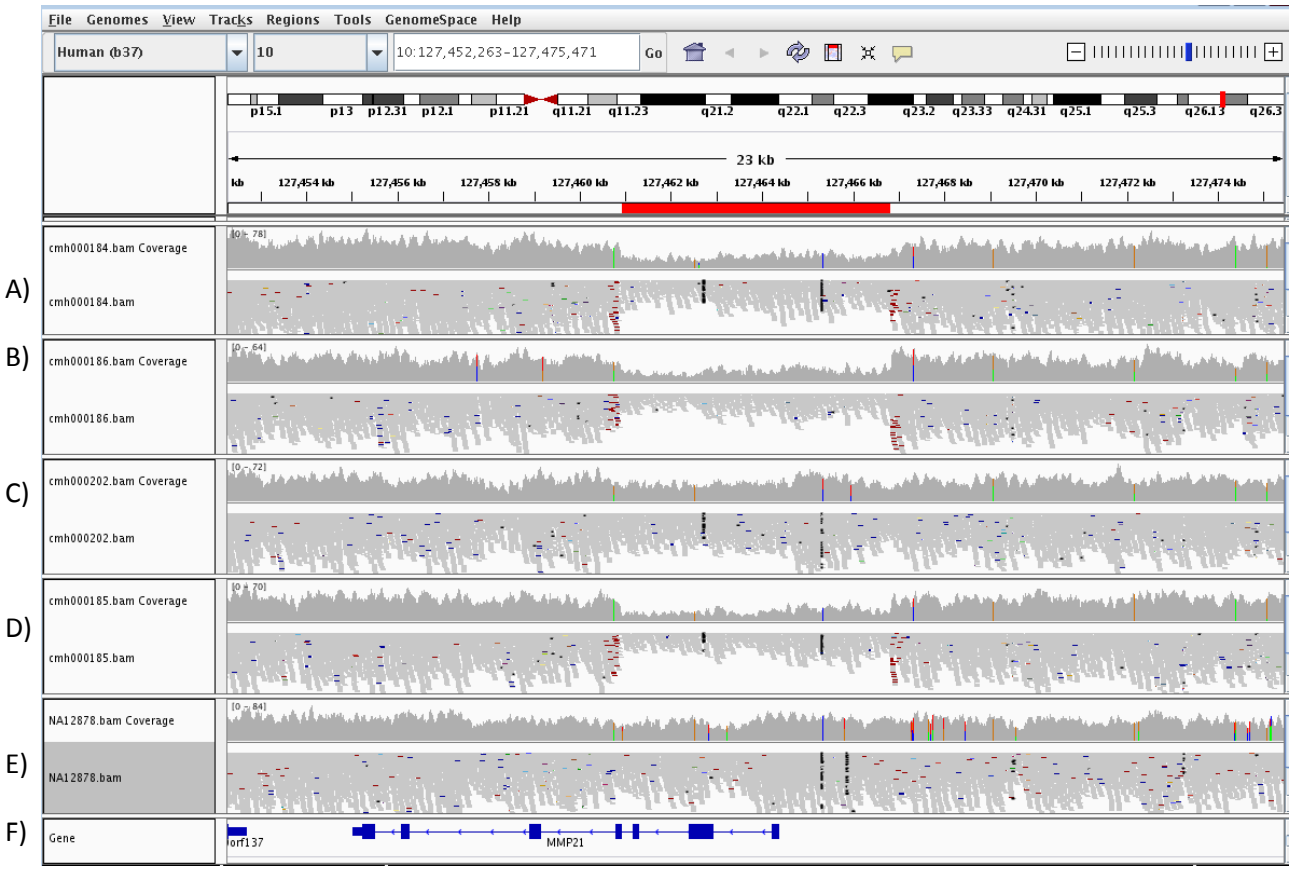

Figure 10. Data shown for tetrad and presumed causative mutation MMP21 for siblings CMH184 and 185. A) CMH184 (proband 1) Upper region demonstrates that coverage over predicted region (horizontal red bar) follows a heterozygous deletion SV pattern. Lower region depicts a dark vertical line over the sixth MMP21 exon which corresponds to small deletison that induced a frame shift. Red stretch pairs flanking the prediction region provide additional evidence for a deletion SV.

B) CMH186 (unaffected parent 1) Upper region demonstrates that coverage over predicted region follows a heterozygous deletion SV pattern. Lower region shows a normal haplotype in reads. Red stretch pairs flanking prediction region provide additional evidence for a deletion SV.

C) CMH202 (unaffected parent 2) Upper region demonstrates normal coverage over predicted region without signs of deletion. Lower region depicts a dark vertical line over sixth MMP21 exon which corresponds to small deletion which induced a frame shift.

D) CMH185 (proband 2) Upper region demonstrates that coverage over predicted region follows a heterozygous deletion SV pattern. Lower region depicts dark vertical line over sixth MMP21 exon which corresponds to small deletion which induced frame shift. Red stretch pairs flanking prediction region provide additional evidence for a deletion SV.

E) NA12878 (control sample) Upper region demonstrates normal coverage over predicted region without signs of deletion. Lower region demonstrates no abnormalities over sixth MMP21 exon.

F) MMP21 gene with blue rectangles corresponding to exons and line corresponds to untranslated regions of gene.

Extent and impact for deletions in 73 individuals by WGS

Finally, we assessed the size distribution and impact of the unique non-overlapping BD$\cap^{90\%}$GS deletion SVs predicted for 73 individuals (13 were unrelated, 2 were a pair, 54 were trios, and 4 were part of a tetrad). BD$\cap^{90\%}$GS predictions were filtered as before, and merged if overlapped by at least one nucleotide resulting in 5107 unique deletions. (median size=2,067nt, mean size=8,111nt) Most deletions were under 10Kb in size, and those that overlapped with genes had a similar frequency distribution to the overall set while those that overlapped with exons tended to be larger. Of the 5107 unique deletions, 2579 (50%) and 701 (13%) overlapped by at least one nucleotide with a gene or exon respectively (Figure 11). The frequency of this unique set was highest on chromosomes 6, 7, 17, and 18 and was the lowest for chromosome Y (Figure 12). SVAG found 22,025 deletions in 185 unrelated individuals with a median

SV size of 729 nt and mean of 8Kb with 9381 (43%) and 1093 (5%) overlapping with a gene or exon respectively[15]. Despite differences in the relatedness of individuals sequenced, our findings were similar to those of the SVAG in terms of mean size and overlap with genes and exons.

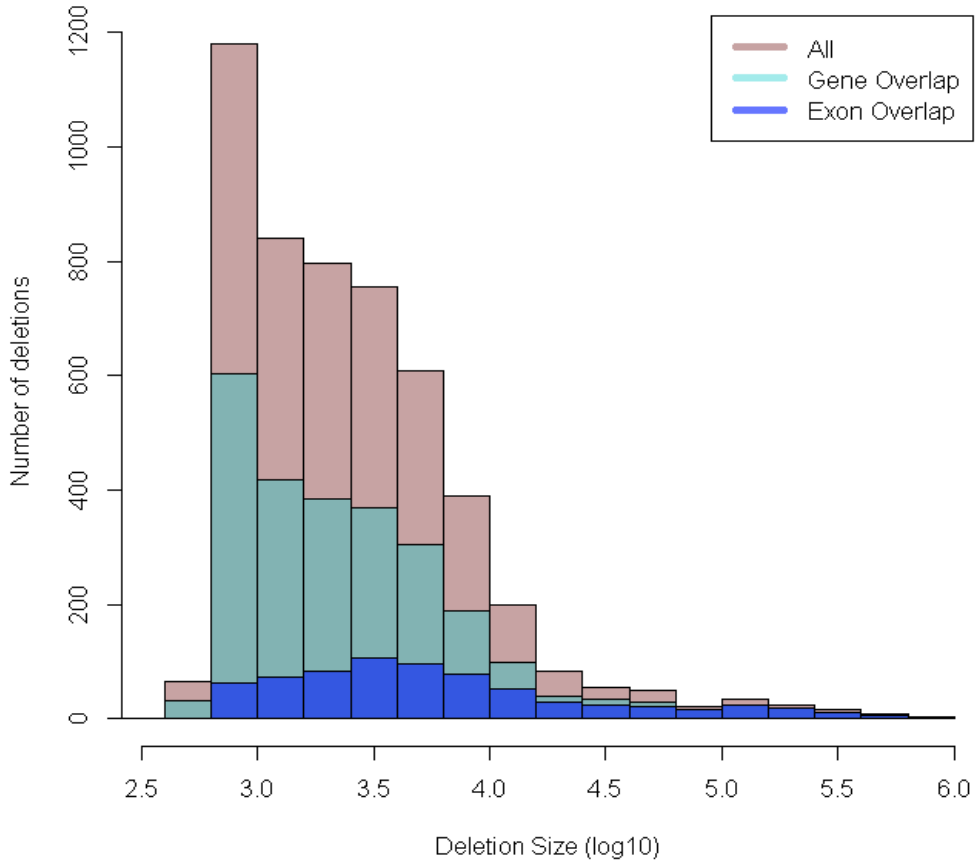

Figure 11. BD∩[90%]GS deletion SV predictions in 73 samples were filtered and merged to achieve a unique set. 5107 total unique deletions were found with 50% overlapping with genes and 13% overlapping with exons.

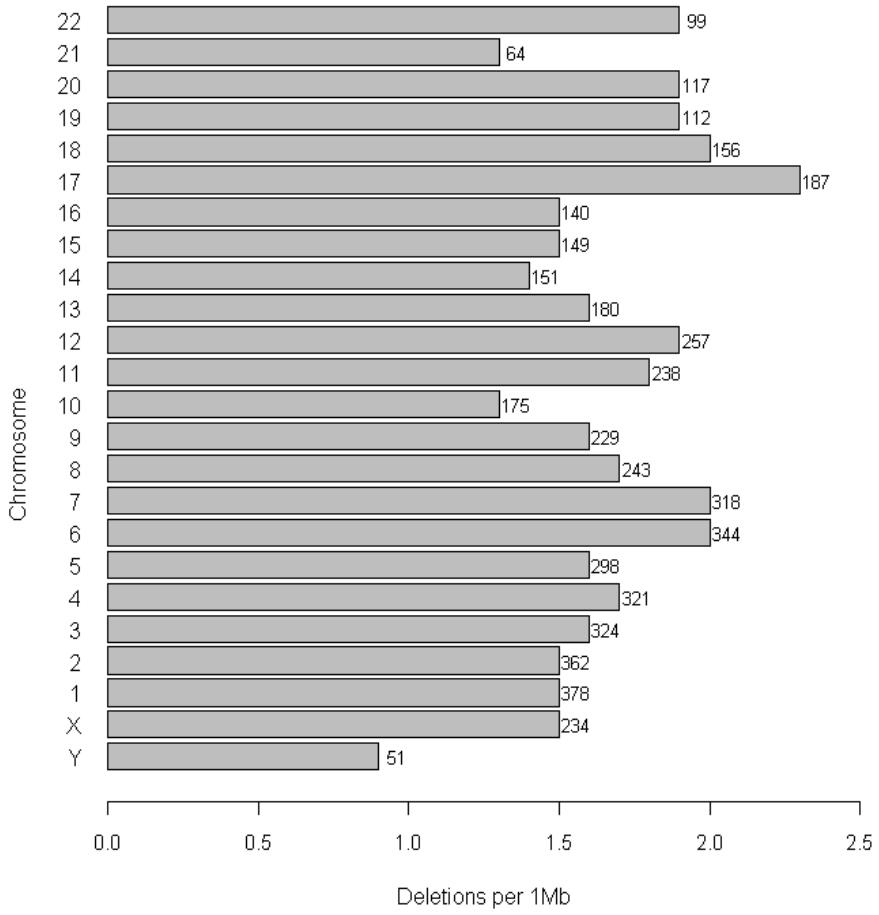

Figure 12. BD∩$^{90\%}$GS deletion SV predictions in 73 samples were filtered and merged to achieve a unique set. The frequency of this unique set was highest on chromosomes 6, 7, 17, and 18 and the lowest for chromosome Y.

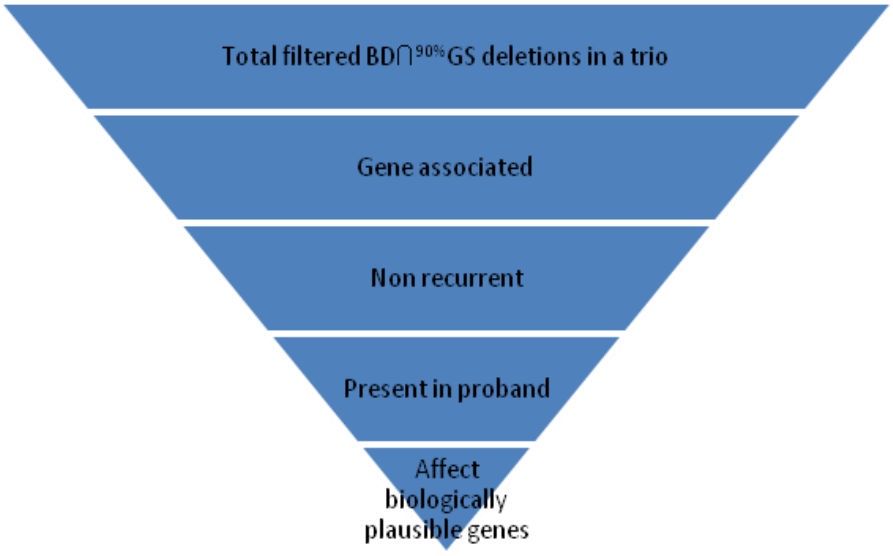

Figure 13. Shown is a trio analysis flow diagram for isolation of putative disease causing deletions SVs.

MATERIALS and METHODS

Study Participants

WGS data were drawn from individuals previously enrolled in the Center for Pediatric Genomic Medicine research repository and approved by the institutional review board at Children's Mercy Hospital Kansas City, MO.

Whole Genome Sequencing

Isolated genomic DNA was prepared for WGS with a modification of the Illumina TruSeq sample preparation. Briefly, 500 ng of DNA was sheared with a Covaris S2 Biodisruptor, end-repaired, A-tailed, and adaptor-ligated. Polymerase chain reaction (PCR) was omitted. Libraries were purified with SPRI beads (Beckman Coulter). Quantitation was carried out by real-time PCR. Libraries were denatured with 0.1 M NaOH and diluted to 2.8 pM in hybridization buffer. Samples for rapid WGS were each loaded onto two flowcells, followed by sequencing on Illumina HiSeq 2500 instruments. Cluster generation, followed by 2 × 100 cycle sequencing reads, separated by paired-end turnaround, were performed automatically by the instrument.

Next Generation Sequencing Analysis

Sequence data was generated with Illumina RTA 1.12.4.2 & CASAVA-1.8.2, aligned to the human reference GRCh37.p5 using GSNAP[10]. Sequence analysis employed FASTQ files, the compressed binary version of the Sequence Alignment/Map format (bam, a representation of nucleotide sequence alignments). Analysis programs were either written in Perl, R, make or the shell scripting language.

Affymetrix SNP/CNV array materials and methods

Isolated genomic DNA was prepared using the four day workflow using the standard 8 step Affymetrix cytoscan assay protocol. Arrays were then washed, stained and scanned. Raw .cel and .dat files were converted to .cychp files using CytoScan® HD Array. Chromosome Analysis Suite 2.0 NetAffx 32.3 (hg19) was used for final analysis and exporting of deletion calls.

DISCUSSION

If outcomes are to be improved for acutely ill neonates suspected of having a genetic condition, it is paramount to confirm or refute a molecular diagnosis in a timely manner. The full potential for WGS data to inform care providers in this regard has not yet been fully realized. Indeed, as more and more rich genomic datasets are generated, the impediment preventing comprehensive SV identification is more analytical than technical[34]. The abundance of SV detection tools reflects the complexity of this challenge, and also the absence of a robust benchmark that all tools can be measured by. In this study much effort was devoted to constructing an alternative source for SV tool evaluation. Poor overlap between NA12878 Affymetrix array technical replicates indicated arrays could not substitute as a standard SV set, since they appeared to lack sufficient sensitivity or precision. It is critical to evaluate SV detection tools formally since there is no default commonly employed program, as currently exists for WGS nucleotide variant detection (i.e. GATK[43]). SV detection tool run times can vary from hours to several days for a WGS data set, and our initial Chr 1 (~8% of entire genome) test was appropriately sized to efficiently assess programs, while still resembling a whole genome test. It should be noted that for clinical use, SV

detection tools must run as quickly as tools used for small variant detection. Our whole genome simulation was based on characteristics from a validated subset of CNVs from 185 1KGP samples. This analysis led us to select the 90% reciprocally overlapping predictions (BD$\cap^{90\%}$GS) from the top two performing tools Breakdancer (BD) and GenomeStrip (GS) as our consensus set, since this approach decreased the ratio of FPs to TPs, while not dramatically altering sensitivity. From NA12878 WGS and array comparisons we confirmed BD$\cap^{90\%}$GS decreased the ratio of FPs to TPs compared to either tool alone, and concluded that BD$\cap^{90\%}$GS$\cap^{90\%}$AR would represent a high quality set of predictions.

A filter for deletion predictions was parameterized from 112 BD$\cap^{90\%}$GS$\cap^{90\%}$AR. A computational pipeline was implemented to overlap, filter and annotate putative BD and GS deletion SVs, while completing in less than 8 hours per WGS dataset. We assessed our pipeline and filtering though analysis of WGS experimental replicates, trio inheritance patterns, and inspection of raw data for a tetrad in IGV. Experimental replicate analysis demonstrated that filtering enriched for predictions shared between replicates. Our two NA12878 array technical replicates yielded 22% overlap yet r1$\cap^{90\%}$r2$\cap^{90\%}$r3/r1Ur2Ur3 for the two WGS experimental replicate sets pg96 and UDT173 equaled 20% and 37% respectively. When considering the fundamental differences between WGS experimental replicates and that the consensus for each set was based on a three way comparison there is evidence that our WGS pipeline performance is equal to or better than arrays which are the current SV detection standard. Trio analysis indicated that filtering enhanced Mendelian inheritance patterns and reduced the quantity of *de novo* predictions unique to probands. Inspection in IGV of selected regions from a tetrad Venn diagram (Figure 8) demonstrated high specificity but relative insensitivity. WGS from 73 individuals yielded a deletion SV size distribution similar to the structural variation analysis group (SVAG) results in terms of mean size, and the proportion that overlapped with genes and exons. Our higher proportion of deletion SVs overlapping with genes and exons could be due to a greater disease burden in our sample or a more extensive set of annotations.

Our pipeline prioritized specificity over sensitivity, for the greatest current utility, and the resultant limitations of a hard filter became most apparent in familial analysis. A similar challenge had faced early SNP detection methods, and was, in part, solved through the use of training set population data to improve analytical performance[44]. This is, unfortunately, not yet an effective solution for SV detection since few validated sets of SVs exist. Using a more flexible filter and or only one tool to increase sensitivity would currently decrease the specificity of results unacceptably. Additional problems remain for deletion SV detection, such as inaccurate breakpoints around repetitive regions and multiple overlapping predictions. We had expected to find a higher rate of *de novo* deletions which empirically appears to be close to zero.

Our clinical approach for effective diagnosis of deletion SV disorders in newborns will be restricted to familial sets where predictions from each tool are 1) intersected, 2) filtered, and 3) annotated, with all common deletions being removed. The latter is a powerful filter for genetic disease diagnosis, since rare diseases cannot be causally associated with common SVs[25]. Finally, the last step in our clinical computational pipeline will be to focus on those genes predicted to be associated with a newborn's specific phenotype[25]. (Figure 13) This also is a powerful filter, given the availability of tools to nominate disease genes comprehensively on the basis of symptoms. However, it is not useful when the patient's disease features are atypical or reflective of an early phase in disease evolution, or if the patient has a novel genetic disease. The combination of allele frequency and phenotypic filters allows greater tolerance of non-specificity, which, in turn, enables parameterization for greater analytic sensitivity. The end goal of this clinical strategy is to find homozygous gene-overlapping deletion SVs or heterozygous deletion SVs where a disease causative nucleotide mutation exists on the other DNA strand. Using WGS to clinically diagnose genetic conditions is a powerful strategy since both SNPs and deletion SVs can be interrogated in the same dataset.

The next immediate step will be to validate a large number of potentially-disease causing putative deletion SVs though either qPCR, Sanger sequencing, arrays or RNA-seq. We expect that additional

experience from the routine use of the deletion SV detection pipeline for WGS neonatal disease diagnosis will yield insights for ways to increase sensitivity (e.g. filter modifications). By changing our upstream alignment approach, GS is expected to gain additional ability to identify deletion SVs within reads and increase sensitivity though its SRM functionality. Finally, validated deletion SVs from each new sample sequenced will be added to an internal reference database of TP pathogenic and benign deletion SVs.

CONCLUSIONS

Identification of deletion SVs, when combined with small variant detection in a rapid software pipeline, can prove to be an effective tool for identifying disease causing mutations in neonates. Results from this study indicate that deletion SVs in WGS data can be expeditiously detected with high specificity using open source tools and special criterion. We expect our computational pipeline to be continuously improved through routine clinical use, Mendelian inheritance patterns and molecular validation techniques.

REFERENCES

1. Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M., Hurles, M. E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat Genet, 40(1), 90-95. doi: 10.1038/ng.2007.40
2. Handsaker, R. E., Korn, J. M., Nemesh, J., & McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet, 43(3), 269-276. doi: 10.1038/ng.768
3. Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet, 10, 451-481. doi: 10.1146/annurev.genom.9.081307.164217
4. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., & Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics, 28(21), 2711-2718. doi: 10.1093/bioinformatics/bts535
5. Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*. 1996-2010 <http://www.repeatmasker.org>
6. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 27(2), 573-580.
7. Morgulis, A., Gertz, E. M., Schaffer, A. A., & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol, 13(5), 1028-1040. doi: 10.1089/cmb.2006.13.1028
8. Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren,

Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J.P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbino, and Stephen M.J. Searle Ensembl 2014
Nucleic Acids Research 2014 42 Database issue:D749-D755 doi: 10.1093/nar/gkt1196

9. Zhao, M., Wang, Q., Wang, Q., Jia, P., & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics, 14 Suppl 11, S1. doi: 10.1186/1471-2105-14-S11-S1
10. Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics, 26(7), 873-881. doi: 10.1093/bioinformatics/btq057
11. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079. doi: 10.1093/bioinformatics/btp352
12. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842. doi: 10.1093/bioinformatics/btq033
13. Valsesia, A., Mace, A., Jacquemont, S., Beckmann, J. S., & Kutalik, Z. (2013). The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. Front Genet, 4, 92. doi: 10.3389/fgene.2013.00092
14. Zhang, D., Qian, Y., Akula, N., Alliey-Rodriguez, N., Tang, J., Bipolar Genome, S., Liu, C. (2011). Accuracy of CNV Detection from GWAS Data. PLoS One, 6(1), e14511. doi: 10.1371/journal.pone.0014511
15. Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Genomes, P. (2011). Mapping copy number variation by population-scale genome sequencing. Nature, 470(7332), 59-65. doi: 10.1038/nature09708

16. International HapMap, C., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature, 449(7164), 851-861. doi: 10.1038/nature06258
17. Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061-1073. doi: 10.1038/nature09534
18. Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. doi: 10.1038/nbt.2835
19. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., & Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics, 28(21), 2711-2718. doi: 10.1093/bioinformatics/bts535
20. McCandless, S. E., Brunger, J. W., & Cassidy, S. B. (2004). The burden of genetic disease on inpatient care in a children's hospital. *Am J Hum Genet, 74*(1), 121-127. doi: 10.1086/381053
21. Dye, D. E., Brameld, K. J., Maxwell, S., Goldblatt, J., Bower, C., Leonard, H., . . . O'Leary, P. (2011). The impact of single gene and chromosomal disorders on hospital admissions of children and adolescents: a population-based study. *Public Health Genomics, 14*(3), 153-161. doi: 10.1159/000321767
22. Kumar, P., Radhakrishnan, J., Chowdhary, M. A., & Giampietro, P. F. (2001). Prevalence and patterns of presentation of genetic disorders in a pediatric emergency department. *Mayo Clin Proc, 76*(8), 777-783. doi: 10.1016/S0025-6196(11)63220-5
23. Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature, 470*(7333), 187-197. doi: 10.1038/nature09792
24. Lu, X. Y., Phung, M. T., Shaw, C. A., Pham, K., Neil, S. E., Patel, A., . . . Beaudet, A. L. (2008). Genomic imbalances in neonates with birth defects: high detection rates by using chromosomal microarray analysis. *Pediatrics, 122*(6), 1310-1318. doi: 10.1542/peds.2008-0297

25. Saunders, C. J., Miller, N. A., Soden, S. E., Dinwiddie, D. L., Noll, A., Alnadi, N. A., . . . Kingsmore, S. F. (2012). Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med, 4*(154), 154ra135. doi: 10.1126/scitranslmed.3004041
26. Hauck, F. R., Tanabe, K. O., & Moon, R. Y. (2011). Racial and ethnic disparities in infant mortality. *Semin Perinatol, 35*(4), 209-220. doi: 10.1053/j.semperi.2011.02.018
27. M. C. Lynberg, M. J. Khoury, Contribution of birth defects to infant mortality among racial/ethnic minority groups, United States, 1983. MMWR CDC Surveill. Summ. 39, 1–12 (1990).
28. Kochanek, K. D., Kirmeyer, S. E., Martin, J. A., Strobino, D. M., & Guyer, B. (2012). Annual summary of vital statistics: 2009. *Pediatrics, 129*(2), 338-348. doi: 10.1542/peds.2011-3435
29. Trask, B. J., Massa, H., Brand-Arpon, V., Chan, K., Friedman, C., Nguyen, O. T., . . . Giorgi, D. (1998). Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum Mol Genet, 7*(13), 2007-2020.
30. Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet, 12*(5), 363-376. doi: 10.1038/nrg2958
31. Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., . . . Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature, 464*(7289), 704-712. doi: 10.1038/nature08516
32. Park, H., Kim, J. I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S., . . . Seo, J. S. (2010). Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet, 42*(5), 400-405. doi: 10.1038/ng.555
33. Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E., & Nickerson, D. A. (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet, 40*(10), 1199-1203. doi: 10.1038/ng.236

34. Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet, 10*, 451-481. doi: 10.1146/annurev.genom.9.081307.164217

35. Buchanan, J. A., & Scherer, S. W. (2008). Contemplating effects of genomic structural variation. *Genet Med, 10*(9), 639-647. doi: 10.1097GIM.0b013e318183f848

36. Gonzaga-Jauregui, C., Lupski, J. R., & Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annu Rev Med, 63*, 35-61. doi: 10.1146/annurev-med-051010-162644

37. Kondrashov, A. S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat, 21*(1), 12-27. doi: 10.1002/humu.10147

38. Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. *Nat Genet, 39*(7 Suppl), S43-47. doi: 10.1038/ng2084

39. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., . . . Sellers, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A, 104*(50), 20007-20012. doi: 10.1073/pnas.0710052104

40. Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., . . . Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science, 318*(5853), 1108-1113. doi: 10.1126/science.1145720

41. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., . . . Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science, 316*(5823), 445-449. doi: 10.1126/science.1138659

42. International Schizophrenia, C. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature, 455*(7210), 237-241. doi: 10.1038/nature07239

43. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20(9), 1297-1303. doi: 10.1101/gr.107524.110

44. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 43(5), 491-498. doi: 10.1038/ng.806
45. Marchenko, G. N., Marchenko, N. D., & Strongin, A. Y. (2003). The structure and regulation of the human and mouse matrix metalloproteinase-21 gene and protein. Biochem J, 372(Pt 2), 503-515. doi: 10.1042/BJ20030174
46. Mmp21 matrix metallopeptidase 21. from http://www.informatics.jax.org/marker/MGI:2664387