**SUPPLEMENTARY DISCUSSION**

***Problems associated with array-based detection of >100kb GS CNVs***

We show that using short-insert or mate-pair WGS ≥1x coverage can detect all seven GS deletions >100 kb whereas only one of these deletions is detectable with most arrays (Table 1)**.** As discussed in Haraksingh *et al* [1], all arrays are expected to detect these seven GS deletions based on total probe count and spacing. However, only one of the deletions was detected by most arrays. The chr3:162,514,471–162,625,647 deletion was only called in aCGH, but incorrectly called as duplication. This is likely because, based on the 1000 Genomes Project genotyping data [2], it is heterozygous in NA12878 and homozygous in the aCGH reference NA10851. On the other hand, four deletions, two on chromosome 4 and one each on chromosomes 6 and 8, are all heterozygous in both NA12878 and NA10851 and therefore, not expected to be found by aCGH. This is indeed the case for all CGH arrays with the exception of Agilent 2x400K-CNV, where these four deletions were all called as duplications. This artifact is likely to be platform-specific, but it is nevertheless puzzling as to how it could arise. We speculate that because both NA12878 and NA10851 harbor heterozygous deletions at these loci, the total amount of DNA available for hybridization at these loci is only half as much compared to other loci thus rendering the probe signals at these four loci more prone to experimental noise. Furthermore, a GS CNV set was also generated for NA10851 [1]. As mentioned in Haraksingh *et al* [1], 876 CNVs are common to both NA12878 and NA10851 and are not expected to be detectable by aCGH as long as the type of CNV (deletion or duplication) is the same. This also points to a general disadvantage of aCGH as it is highly dependent on the reference – it is possible to end up calling a CNV as a deletion that is in fact a duplication in the reference DNA or as a duplication in the test DNA where it is a deletion in the reference DNA [1].

***Challenges in CNV analysis for WGS and arrays***

The important question of how CNV detection using arrays and sequencing compare has remained in the field for quite some time [3–7], which is also the main motivation of our study. There are currently many bioinformatic challenges associated with sequencing-based CNV analysis with no well-established pipeline, protocol, and quality-control standards [8]. These challenges are mainly due to read mappability (particularly in repetitive regions), GC-content bias, read quality, and the difficulty in identifying duplications [8]. It is important to note that the problem of CNV detection in repetitive regions is not yet fully resolved for both sequencing and arrays [8, 9]. To best ameliorate these challenges and to improve CNV detection, we implemented many quality control measures in our WGS-CNV analysis. We removed CNV calls that overlapped reference gaps for read-depth analysis and those with supporting reads that map to repetitive regions for discordant paired-end CNV analysis. We also included the quality metric, *q0* [11], which indicates the fraction of supporting reads for a given CNV that have a mapping quality of zero for read-depth analysis so that readers can be guided with regards to the confidence of that CNV (Supplementary Table S2). Of the 15 currently available read depth-based CNV detection tools [10], we chose to use CNVnator [11], which corrects the read-depth value in a given genomic window based on GC-content. We trimmed reads from the 3’ end to retain >q20 quality score before the initial alignment, and we also ensured that singleton reads were not filtered out as they are still useful for read-depth analysis [8]. PCR duplicates were removed twice, using Samtools [12] and then Picard tools (see Methods). To further ensure quality after the initial alignment, we took additional steps to re-align reads around common indels and recalibrated the base-quality scores at loci known to have SNPs in the population following GATK best practice guidelines [13]. For discordant read-pair analysis, we ensured that the CNV events detected had multiple supporting reads as a function of sequencing coverage in order to minimize false positives, especially for duplications. False duplication calls may arise from read-depth analysis due to poor-quality alignments and PCR duplicates, which might yield artificially higher coverage in certain regions. Nevertheless, the detection of duplications using read-depth analysis is difficult, because while in non-repetitive diploid regions where reference copy number is 2 (CN=2), an additional duplication (CN=3) increases the read-depth signal by 50% (easily detectable, less susceptible to variance; in repetitive regions, however, where the reference copy number may be 6 (CN=6), an additional duplication (CN=7) will only yield a 17% increase in signal rendering it less detectable in the presence of signal variance [8, 11]. Notably, this challenge is also a feature of aCGH and arguably more prominent compared to sequencing [11, 14]. We believe that the lower validation rates for duplications in low-coverage WGS (Supplementary Figure S4b) are due to not only the difficulties of detecting duplications in general [8, 11] but also to the low number of duplications in GS CNVs (137 duplications vs 2,034 deletions).

***CNV analysis tools and usage***

In addition to CNVnator [11] and LUMPY [15], numerous openly available CNV detection tools have been developed for WGS data based on discordant read-pair analysis [16–20], read-depth analysis [3, 21–24], or combinations of these two principles [25–28]. We chose to use CNVnator and LUMPY mainly due to software availability and extensive previous usage experience on our part. Further investigation into the comparative performances of available tools and optimization of their operating parameters will very likely be beneficial. It is important to note that relatively few currently available tools can use WGS mate-pair data as their input, which somewhat narrows the choices available to investigators and practitioners of cytogenetics, but this is not a major drawback as even just the software packages used in this study already performed quite well for the purpose at hand.

We find that using a combination of read-depth analysis and discordant-read-pair analysis produces a more confident CNV call set. Generally, large CNVs called from discordant-read-pair analysis without corroboration from read-depth analysis are likely to be false positives since large CNVs are called more robustly using read-depth analysis (Figure 3d-i). However, when a CNV is called by both types of analyses, discordant-read-pair analysis has the ability to provide a more accurate genomic boundary of the CNV. Reference gaps are problematic for CNV calling using read-depth analysis, where the absence of read alignment in the gaps can be mis-interpreted as deletions, thus CNV calls that overlap these genomic regions should be flagged and looked at with extra scrutiny. (The same is true for array-based approaches). In addition, deletions and reference gaps that lie in close proximity are sometimes merged in read-depth analysis, particularly at low coverages. For example, the large NA12878 GS CNV on chromosome 19 (chr19:20,595,835-20,717,950, a 122,115bp deletion) is detected but called as a much larger 215kb deletion (chr19:20,505,001-20,720,000) when using read-depth analysis because, at low depth of coverage, the read-depth signal from this deletion is merged with a nearby reference gap (chr19:20,523,415-20,573,415) as well as with a Silver Standard deletion (chr19:20,504,101-20,586,400). However, this 122kb GS deletion on chromosome 19 is detected as a 122,419bp deletion (chr19:20,595,751-20,722,956) using discordant read-pair analysis from 5kb mate-pair WGS at 5x coverage.

In read-depth analysis, the “bin size” parameter for CNV calling dictates the lower size limit of CNV detection and is optimal if the ratio of average read depth signal to its standard deviation is approximately 4 to 5 [11]. We calculated the optimal bin size to be 5kb for our WGS data (see Methods). In discordant read-pair analysis, however, the library insert size determines in several ways the size range of detectable CNVs. In general, CNVs can be more easily detected at larger insert sizes due to larger physical coverage, i.e. the number of informative mate-pairs, but at the expense of losing detection power for smaller CNVs due to the increased variability in size selection for longer inserts. By combining read-depth and discordant-read-pair analysis, the 3kb mate-pair WGS library provides the widest size distribution of detected CNVs.

***Performance comparisons with 30x-coverage WGs***

To compare with the other end of the spectrum of sequence coverage, we applied our analysis pipeline to a 30x-coverage NA12878 short-insert WGS dataset. From this dataset, we detect 1,854 CNVs, and 922 are validated (Supplementary Figure S6, Supplementary Table S1). While considerably more CNVs are detected compared to low-coverage WGS as expected, interestingly, we find that the number of GS CNVs detected by 3kb-mate-pair WGS at 5x coverage comes surprisingly close (405 vs 555) to this short-insert 30x-coverage dataset. While sequencing at 30x coverage allow for reliable SNP genotyping as well, it also costs more, takes longer, and needs an off-site cluster for data processing which are impractical and not needed for clinical cytogenetic laboratories.

**REFERENCES**

1 Haraksingh RR, Abyzov A, Urban AE. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics* 2017;**18**:321.

2 Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine Mu X, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**:75–81.

3 Xi R, Hadjipanayis AG, Luquette LJ, Kim T-M, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, Kucherlapati R, Park PJ. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A* 2011;**108**:E1128-36.

4 Zhan B, Fadista J, Thomsen B, Hedegaard J, Panitz F, Bendixen C. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. *BMC Genomics* 2011;**12**:557.

5 de Ligt J, Boone PM, Pfundt R, Vissers LELM, de Leeuw N, Shaw C, Brunner HG, Lupski JR, Veltman JA, Hehir-Kwa JY. Platform comparison of detecting copy number variants with microarrays and whole-exome sequencing. *Genomics Data* 2014;**2**:144–6.

6 Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics* 2017;**18**:286.

7 Geistlinger L, da Silva VH, Cesar ASM, Tizioto PC, Waldron L, Zimmer R, Regitano LC de A, Coutinho LL. Widespread modulation of gene expression by copy number variation in skeletal muscle. *Sci Rep* 2018;**8**:1399.

8 Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* 2012;**28**:2711–8.

9 Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 2007;**39**:S16–21.

10 Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 2013;**14**:S1.

11 Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011;**21**:974–84.

12 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

13 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K V, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma* 2013;**43**:11.10.1-33.

14 Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AWC, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. Origins and functional impact of copy number variation in the human genome. *Nature* 2010;**464**:704–12.

15 Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;**15**:R84.

16 Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 2009;**6**:677–81.

17 Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009;**10**:R23.

18 Lee S, Hormozdiari F, Alkan C, Brudno M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods* 2009;**6**:473–4.

19 Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, Sahinalp SC. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 2010;**26**:i350-7.

20 Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, Schliep A, Schönhuth A. CLEVER: clique-enumerating variant finder. *Bioinformatics* 2012;**28**:2875–82.

21 Chiang DY, Getz G, Jaffe DB, O’Kelly MJT, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 2009;**6**:99–103.

22 Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavaré S. CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* 2010;**26**:3051–8.

23 Kim T-M, Luquette LJ, Xi R, Park PJ. rSW-seq: algorithm for detection of copy number alterations in deep sequencing data. *BMC Bioinformatics* 2010;**11**:432.

24 Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* 2011;**6**:e16327.

25 Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting copy number variation with mated short reads. *Genome Res* 2010;**20**:1613–22.

26 Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, Barillot E. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* 2010;**26**:1895–6.

27 Qi J, Zhao F. inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res* 2011;**39**:W567-75.

28 Sindi SS, Onal S, Peng LC, Wu H-T, Raphael BJ. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol* 2012;**13**:R22.