Detecting Copy Number Variation on Low Coverage Whole Genome Sequencing

Lawrence S. Hon, Jeroen van den Akker, Danny De Sloover, Cynthia L. Neben, Alicia Y. Zhou, Ziga Mahkovec, Jeremy Ginsberg, Justin Lock, Scott Topper
Color Genomics, Burlingame, CA

Introduction

Next generation sequencing has become an integral tool in the study of human genetics, with whole genome sequencing (WGS) typically targeting 30X coverage. To support population-scale studies, low coverage WGS (lcWGS, approximately 1X coverage) can be used to survey genomes in a cost-effective manner and is not limited by pre-determined probe design as in array comparative genomic hybridization (CGH) and fluorescence in situ hybridization (FISH). Copy number variation (CNV) is an important component of genomic diversity and has been linked to a number of diseases in a dosage-sensitive manner. In addition, myelodysplastic cells can acquire genomic alterations that range in size and can extend to the size of full chromosomal aneuploidies. Here, we assess the feasibility and accuracy of using lcWGS to detect CNVs and explore CNV across a set of multi-allelic genes.

Methods

Laboratory procedures were performed at the Color laboratory. DNA was extracted from blood or saliva samples and sequenced using the NovaSeq 6000 instrument, at approximately 1X coverage. Using a bioinformatics pipeline developed for detection of structural variants in a clinical targeted NGS panel, we tuned read depth based calling (CNVkit) for use with lcWGS. To increase specificity for CNV calls, we used a window size of 20 kb; to increase sensitivity for multi-allelic regions we used a window size of 2 kb.

We utilized the resulting data in two ways. First, we assessed detection performance using a set of 65 clinical structural variants larger than 100 kb that had been previously confirmed present by array CGH, MLPA, or PCR. Using lcWGS, we were able to reliably call and detect deletions greater than 100 kb; duplications had lower sensitivity, with reliable detection achieved at over 300 kb. Notably, in one clinical sample a CNV near *NBN* was initially called from the targeted NGS panel data as a 3 MB copy number gain, but lcWGS and array CGH detected a chromosome 8 trisomy. Given the sensitivity in detecting these larger events, we further characterized rare, larger structural variants in an anonymized set of 27,049 samples, totaling 904 CNVs greater than one megabase.

Second, we explored CNV variation in the *LPA* KIV2 region, which is known to have copy number variation between populations[4,5].

Results

Figure 1. Detection of Confirmed Clinical CNVs using lcWGS

Assessment of recall across 65 clinical structural variants >100 kb previously confirmed by array CGH, MLPA, or PCR. All deletions >100 kb were detected, and duplications >300 kb were detected. An additional two CNVs >10 MB were detected but are not shown.

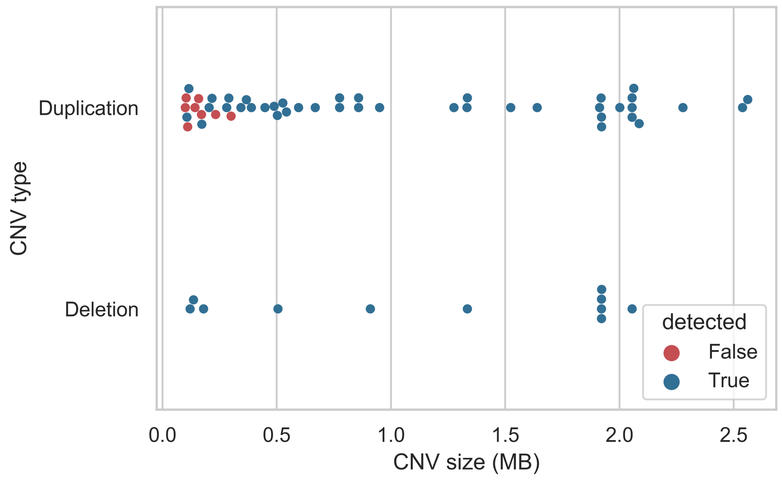
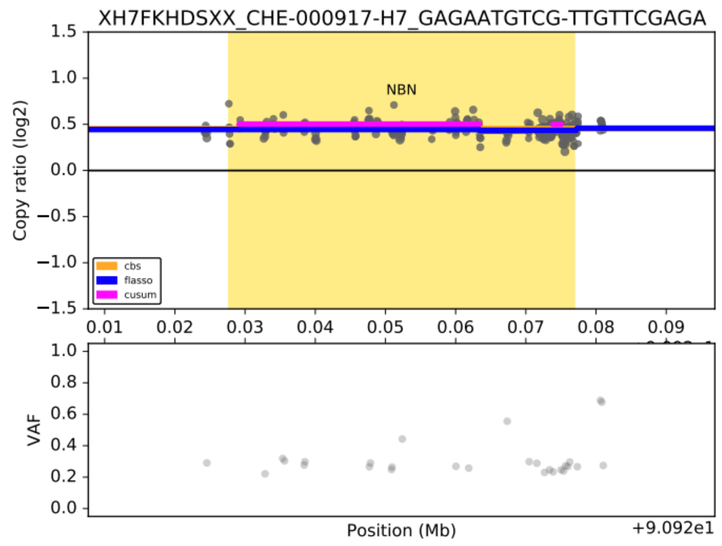


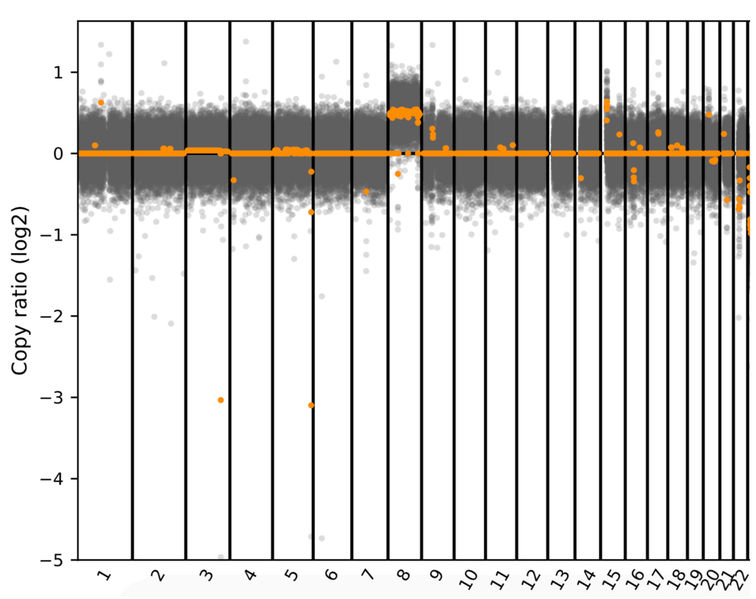
Figure 2. *NBN* Case Example

A notable clinical example was an apparently heterozygous CNV incorporating *NBN* that was initially called from the targeted NGS panel data as a 3 MB duplication (Fig.2a). lcWGS (Fig.2b) and array CGH (Fig. 2c) revealed that the signal was part of a chr8 trisomy. Trisomy 8 is a consistent finding in myelodysplastic syndromes, and further investigation confirmed this phenomenon in this participant.

(a) panel, *NBN* duplication



(b) lcWGS, view of whole genome with observable chr8 trisomy



(c) array CGH, chr8 trisomy

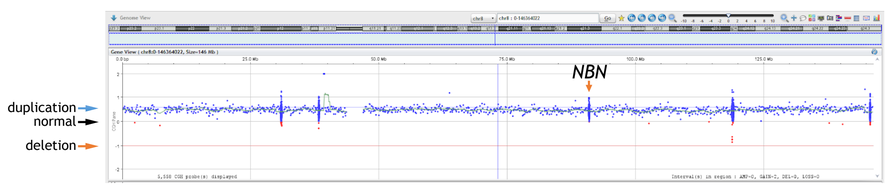


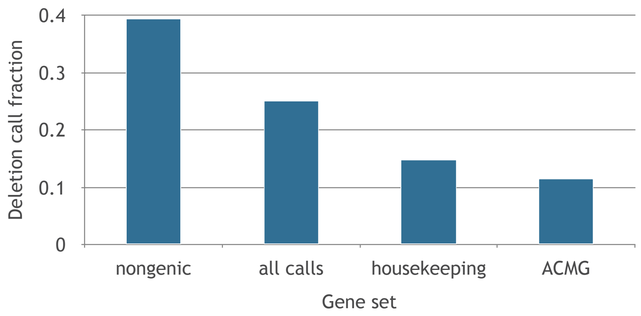
Figure 3. Characterization of Large CNVs (>1 MB)

CNVs greater than 1 MB were characterized among 27,049 samples. Table 3a summarizes the calls made, with gains being 3x more common than losses. Deletion call fraction (Fig.3b) and probability of loss-of-function intolerance (pLI) (Fig.3c) were analyzed as measures of haploinsufficiency.

(a) Summary statistics

Event type	Number of events	Percentage of samples with an event
Duplication	677	2.4%
Deletion	227	0.84%

(b) Fraction of all CNVs detected that were deletions. Comparing the fraction of deletions against detected CNVs across a gene set can be used as a measure of haploinsufficiency. Among nongenic CNVs, deletions are almost as common as duplications. Housekeeping genes[1] and the 59 genes on the ACMG secondary findings list[2] (ACMG) have lower fractions, suggesting a higher rate of haploinsufficiency across these genes.



(c) Genes that more frequently occur in large CNV deletions tend to have lower pLI[3]. For comparison, ACMG genes have a mean pLI of 0.45.

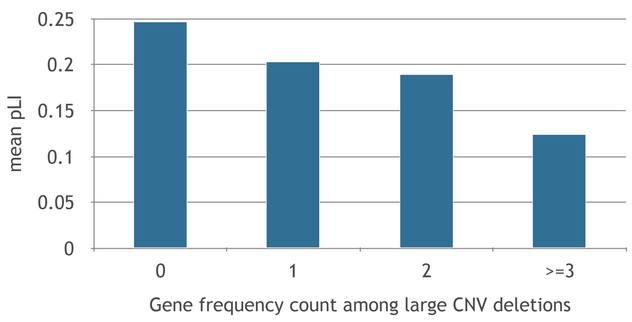
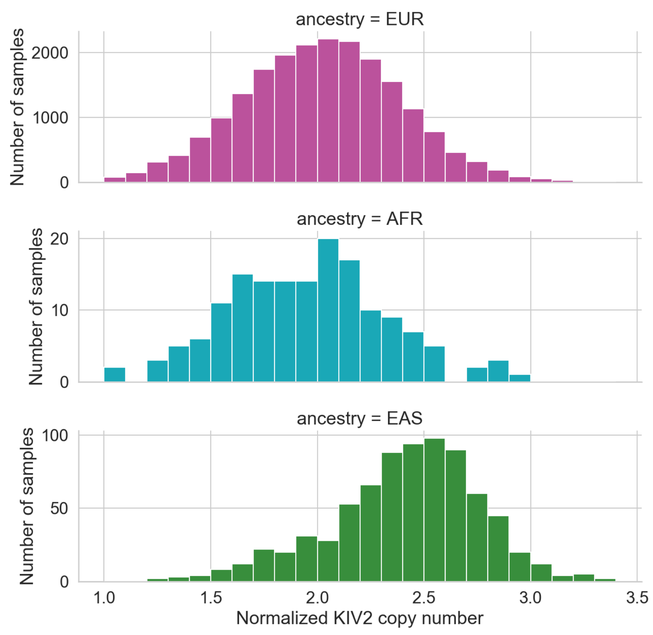


Figure 4. *LPA* KIV2 (kringle) copy number varies between populations

LPA KIV2 is a commonly varying intragene CNV that has been associated with atherosclerotic cardiovascular disease. Here we show *LPA* KIV2 copy number distribution (normalized to 2.0) across European (EUR), African (AFR), and East Asian (EAS) populations, as determined by genetic ancestry. As expected, the AFR mean is slightly shifted left relative to EUR[4], and the EAS mean is shifted right compared to EUR[5].



References

1. Eisenberg et al, 2013. doi: 10.1016/j.tig.2013.05.010
2. https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/
3. Lek et al, 2016. doi: 10.1038/nature19057
4. Zekavat et al., 2018. doi: 10.1038/s41467-018-04668-w
5. Noureen et al. 2016. doi: 10.1371/journal.pone.0121582

Conclusions

- The rich data generated by lcWGS can be harnessed for CNV detection.
- Large and rare structural events can be detected using this technology.
- Multi-allelic, within-gene CNV variation can be detected and characterized with lcWGS.
- Further research and validation is required to understand how lcWGS could be used for clinical applications.