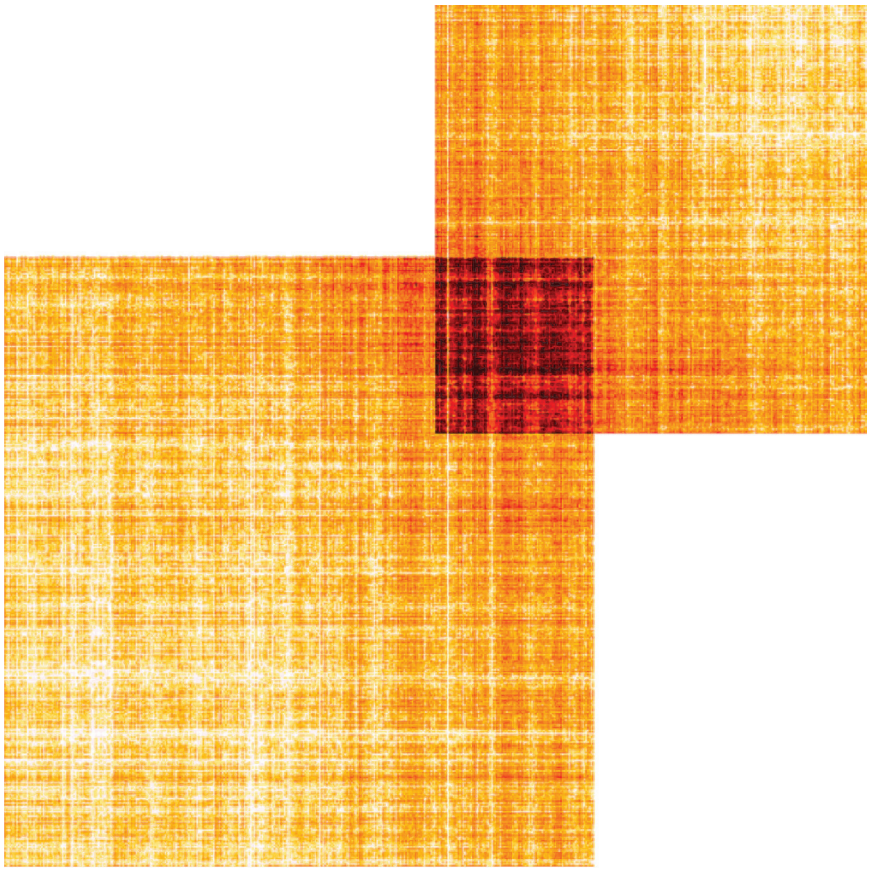
Application Note



CHROMIUM™

Structural Variant Analysis with Linked-Reads

10xGenomics.com/products/#genome

CHROMIUM™ GENOME AND EXOME SOLUTIONS

More Complete Structural Variant Analysis

Highlights

- Linked-Reads provide important haplotype information that can be used for more complete structural variant analysis
- Haplotype phasing can improve the confidence of SV-calls by removing "noise" from un-phased data
- Linked-Reads enable the mapping of reads to repetitive regions of the genome, where structural variant breakpoints often cluster
- Long Ranger™ utilizes the Linked-Read data to reliably detect structural variants in genome and exome data

Introduction

Despite the widespread recognition of the contribution of structural variants (SVs) and copy number variants (CNVs) to health[1,2], they remain one of the most difficult types of variation to accurately ascertain from genomic data, in part because they tend to be clustered in duplicated and repetitive regions of the genome that are typically not accessible by short read sequencing[3,4]. Limitations in sequencing technology and the human reference have led the field to approach human genome analysis as if the genome were haploid. This makes it necessary to average the signal between the two haplotypes, diluting the variant signal and making it difficult to separate true variation from the background noise of depth variability, stochastic changes in allelic representation, probe failure, and alignment artifacts.

The primary SV-calling methods of read depth (RD), split read (SR), read pair (RP) and re-assembly are frequently used in combination to overcome the drawbacks and variation blind spots of each method and still only result in overall sensitivity and Positive Predictive Value of 30-84% and 27-85% respectively on genome[5]. Because of the targeted nature of exome sequencing, the higher depth variability and the statistical unlikelihood that reads spanning the breakpoint will exist and be mapped correctly, the necessary signal is further reduced. Although attempts have been made to call structural variation in exomes, performance is typically reduced and limited to certain variant types and sizes[6-9]. Copy-neutral variation remains difficult to detect because read depth alone is not informative.

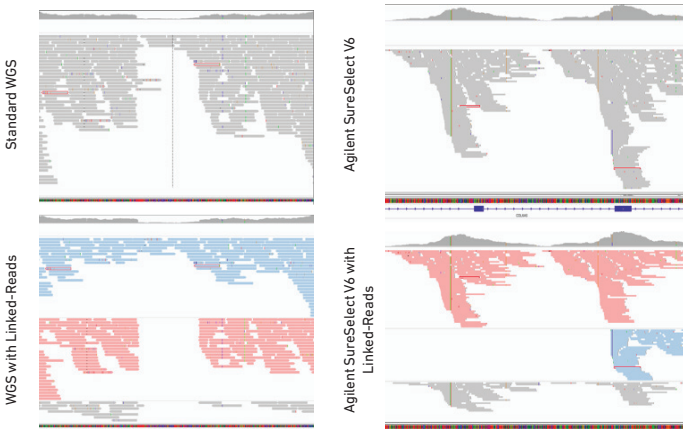
The assessment of SV methods rely on ground truth data composed primarily of variation ascertained by short read sequencing. Thus, sensitivity and PPV estimates are useful for comparison between methods but do not represent true biological estimates of accuracy. The scale of the missing data problem was demonstrated by Chaisson *et al.* in their characterization of the CHM1 complete hydatidiform mole, an effectively haploid sample. They identified 26,079 insertions and deletions ≥50bp, 85% of which were novel, with insertions being particularly under-represented in datasets[4]. Thus, for this study, long-read sequencing of a single haploid genome was capable of discovering more structural variation than the short read sequencing of >175 diploid individuals. Without access to the full breadth of variation present in genomes and exomes, our biological understanding of variation and effect on phenotype will remain limited.

Using Linked-Read Haplotyping to Provide SV-Calling Confidence

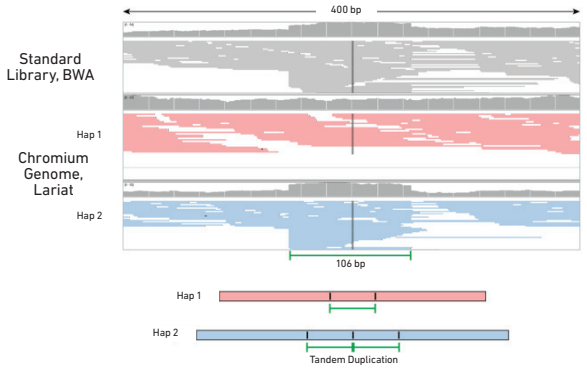
Haplotype information, provided by Linked-Reads, allows exploitation of the diploid nature of the genome. Assessing for the presence of variation haplotype-by-haplotype provides a much more favorable signal-to-noise ratio. Examples of un-phased and phased genome and exome data over regions with known intermediate deletions or duplications from NA12878 are shown below. Because heterozygous variants are expected to be contained on a single haplotype, the noise of the unaltered haplotype can be ignored, amplifying the signal. Linked-Reads are colored when they are phased (blue for one haplotype, red for the other), and grey when they are un-phased, as shown below.

325bp GIAB deletion on standard and Linked-Read WGS.

875bp GIAB deletion standard and Linked-Read exome sequencing.

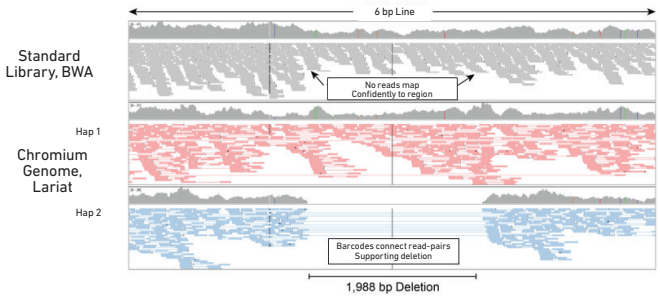


106bp GIAB duplication on standard and Linked-Read WGS.



Using Linked-Read Alignments to Access Difficult Regions

Linked-Reads enable the mapping of reads to repetitive regions of the genome where structural variant breakpoints often cluster. Multiply mapping reads can be rescued if they share barcodes with uniquely mapping reads. Thus, regardless of sequence identity, repetitive regions can be mapped if they do not share barcodes with their homologous sequence. Practically, this occurs when the two repetitive sequences are located far enough away from each other in genomic space that they are unlikely to reside on the same high molecular weight molecule utilized in partitioning. This enables variant calling, including structural variants within repetitive elements. For example, a 1,988bp deletion (known from GIAB) shown below can be called within a 6kb LINE element due to the rescue of repetitively mapped reads that share barcodes with unique reads lying outside the LINE element.



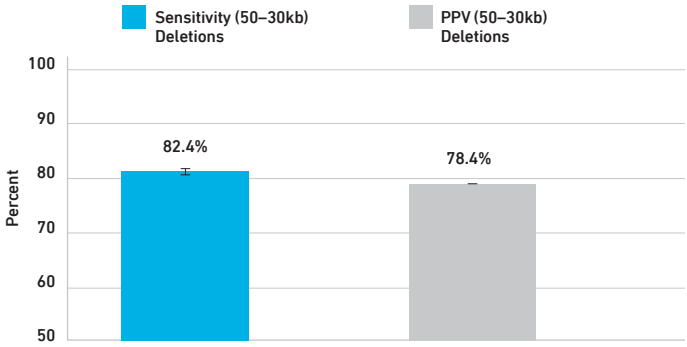
Methods

High molecular weight DNA from the reference cell line NA12878 was obtained, and libraries were prepared (Chromium Genome from 1.25ng input DNA) and sequenced to ~35x depth for genome and ~65x depth for exome (12Gb Linked-Reads + Agilent SureSelect^{XT} V6 with 'bridging baits'). Reads were aligned to GRCh37. Phasing and variant calling was performed with Long Ranger v2.116. Truth sets for structural variant accuracy were created using published call sets[13-18], PacBio evidence, and manual curation.

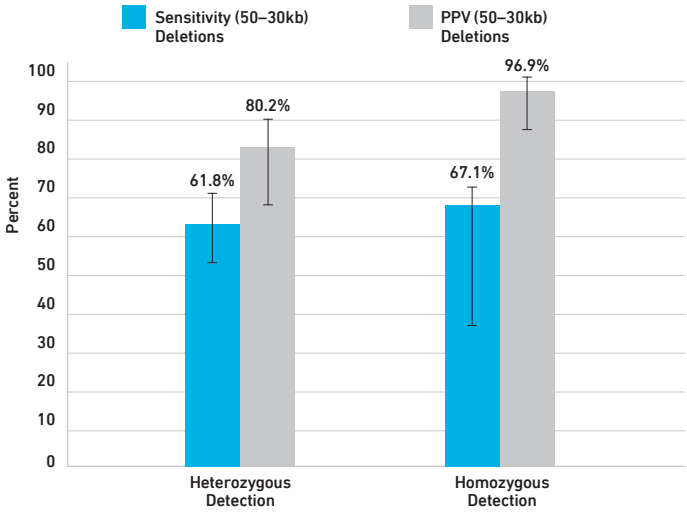
Performance of Long Ranger™ 2.1 on NA12878

The overall performance of Linked-Reads run on the Long Ranger 2.1 software for deletion detection across all size ranges is shown below. There are only two known inversions in NA12878, both of which can be reliably detected by Linked-Reads on both genome and exome.

Sensitivity and PPV of variant detection on NA12878 Chromium Genome (128Gb) (n=6).



Sensitivity and PPV of variant detection on Chromium Exome + SureSelect V6 (9Gb) (n=14).



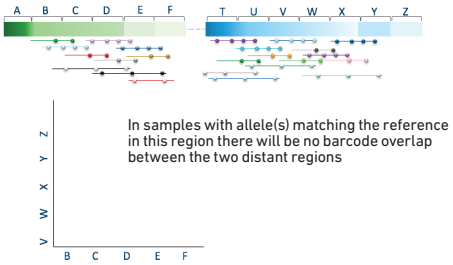
Using Linked-Read Barcodes to Detect Anomalous Barcode Sharing

Barcode information can be used beyond phasing to identify regions located abnormally close to or far from each other. This method proves to be particularly powerful in the detection of copy-neutral events such as inversions and balanced translocations. Genomic regions that, in the reference genome, occur further apart from each other than the average molecule length used in the Chromium Genome library preparation can be assessed for their co-involvement in a structural variant.

The illustrations below show the concept behind the Matrix Plots utilized to demonstrate structural variation in the 10x Genomics visualization tool Loupe™. Two regions of the genome A-F and T-Z are separated by a distance such that long input molecules will rarely or never contain both loci. Therefore, the reads mapping to each locus are expected to rarely or never share barcodes with reads mapped to the other locus. However, when a structural variant alters the two loci, in this case by inversion, the locations closest to the breakpoints (C&D with X&W on one end; F&E with Y&Z on the other) will show significant and unexpected sharing of barcodes. The degree of sharing is shown in the matrix plot as a heat plot with black being the most sharing and white being no sharing.

Illustration of molecules, barcodes, and matrix plot for a sample with no inversion and with inversion relative to the reference sequence.

No Inversion



With Inversion

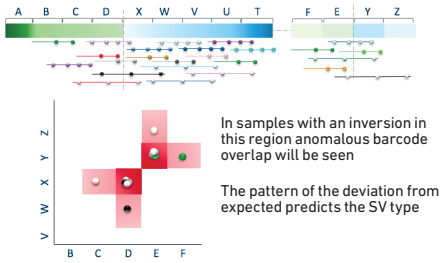


Illustration of a balanced translocation between chromosomes 8 and 10 ascertained by karyotype in a clinical research sample.

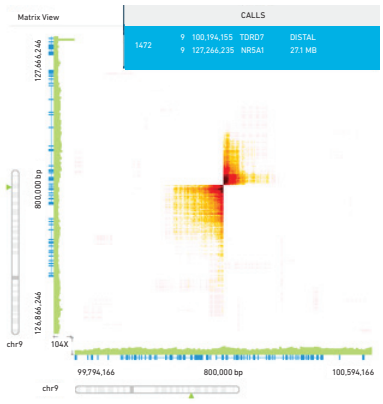
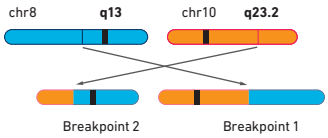
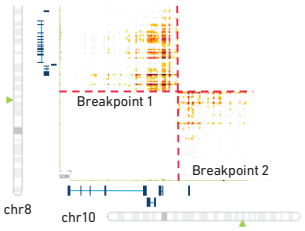


Illustration of a balanced translocation between chromosomes 8 and 10 ascertained by karyotype in a clinical research sample.



Matrix plot for exome sequencing (Agilent V6 with Linked-Read bridging baits) of a clinical research sample with a previously known balanced translocation between chromosomes 8 and 10.



Chromium™ Exome Solution datasets available for download:

- NA 12878 and other Datasets

Files available at: support.10xgenomics.com/genome-exome/datasets

Download Long Ranger and Loupe Genome Browser at: support.10xgenomics.com/genome-exome/software/downloads/latest

Conclusion

Algorithms developed by 10x Genomics to make use of Linked-Reads are open-source. Thus, they can be incorporated into existing workflows together with other variant calling algorithms or used as a stand-alone variant-calling resource.

1. SMFM, Am J Obstet Gynecol. 2016 Jul 15.
2. Manning and Hudgins Genet Med. 2010 Nov;12(11):742-5.
3. (Quinlan and Hall, Trends Genet. 2012 Jan;28(1):43-53;
4. Chaisson et al., Nature. 2015;517:608–611.
5. Layer et al. Genome Biol. 2014; 15(6): R84.
6. de Ligt J, Boone PM, Pfundt R, Vissers LE, Richmond T, Geoghegan J, et al. Hum. Mutat. 2013;34(10):1439–48. doi: 10.1002/humu.22387.
7. Challis D, Antunes L, Garrison E, Banks E, Evani US, Muzny D, et al. BMC Genomics. 2015;16:143. doi: 10.1186/s12864-015-1333-7
8. Du et al. BMC Med Genomics. 2016 Aug 27;9(1):56
9. Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, et al. Bioinformatics. 2012;28(21):2747–54. doi: 10.1093/bioinformatics/bts526.
10. https://clinicalgenome.org/site/ assets/files/1460/ mandelker_comprehensiveexome.pdf
11. Weise et al. J Histochem Cytochem. 2012 May; 60(5): 346–358.
12. http://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger
13. Parikh et al. BMC Genomics. 2016 Jan 16;17:64.
14. ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ ftp/data/NA12878/ NA12878_PacBio_MtSinai/
15. Kidd et al. Nature. 2008 May 1; 453(7191): 56–64.
16. Mills et al. Nature. 2011 Feb 3;470(7332):59-65.
17. Yang et al. Cell. 2013 May 9;153(4):919-29.
18. Layer et al. Genome Biol. 2014 Jun 26;15(6):R84.

Additional Resources

support.10xgenomics.com/genome-exome

10xGenomics.com | info@10xgenomics.com

10x GENOMICS®