10x GENOMICS®

TECHNICAL NOTE

An Introduction to Linked-Read Technology for a More Comprehensive Genome and Exome Analysis

INTRODUCTION

Currently, the predominant method for genome analysis involves sequencing an individual genome with short reads without retaining haplotype knowledge and then aligning those reads to a haploid consensus reference assembly. While this approach provides sufficient power to call single nucleotide variants (SNVs) across most of the genome, a complete analysis of the genome is not possible. Because haplotype information is not retained for the sequenced genome or the reference, the reconstruction of long range haplotypes is challenging using only short read data from a single genome. Additionally, most analytical methods struggle to call large structural variants, particularly balanced events such as inversions and translocations using short reads alone. Finally, due to the widespread presence of high identity repeats and paralogs, entire regions of the genome are inaccessible to most short read analytical methods. Fueled by Chromium™ technology, 10x Linked-Reads amplify the power of short read sequencing and enable the analysis of a more complete genome.

Linked-Reads are generated from short read sequences created with an in-line barcode (Fig 1). Because limiting DNA amounts are utilized, reads that share a barcode can be grouped as deriving from a single long input molecule. In this way, long range information can be assembled from short reads.

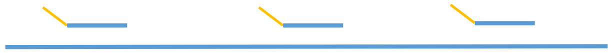


Fig 1. Reads (short blue lines) are generated from each high-molecular weight gDNA molecule (long blue line). Reads from the same molecule will share the same barcode (shown in gold).

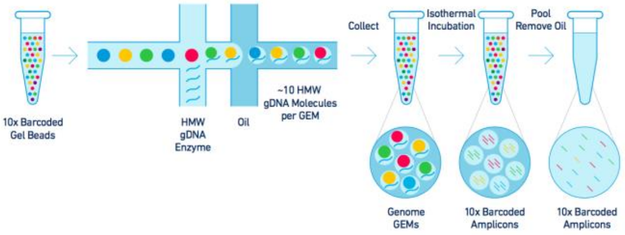


Fig 2. Chromium™ Technology mixes functionalized gel beads containing unique barcodes with enzymes and a limiting amount of genomic DNA to create >1,000,000 uniquely addressable partitions in minutes. Using a limiting dilution of molecules allows the correct mapping of reads to their corresponding molecules.

GENERATING LINKED-READS

A high efficiency microfluidic device mixes functionalized gel beads containing unique barcodes with enzymes and a limiting amount of genomic DNA (Fig 2). These components are encapsulated in oil to produce a GEM, **G**el-bead in **EM**ulsion. With approximately 4 million unique barcodes available, a typical run will partition individual barcodes into 1.4 million GEMs. Within each GEM, there are, on average, 10 gDNA molecules. The primer contained on the gel bead is configured such that read1 is adjacent to the 16 bp unique barcode, which is followed by a random 6mer (Fig 3A). An excess of barcoded sequencing template is produced, with only a subset of it typically sequenced in order to sample more molecules rather than more reads per molecule (Fig 3B). In an average run, each molecule has approximately 0.2x sequence coverage.

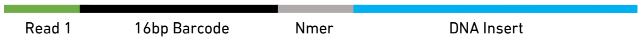


Fig. 3A. Each gel bead contains a primer that is configured to align read1 with the 16 bp unique barcode, which is followed by a random 6mer and the DNA insert.

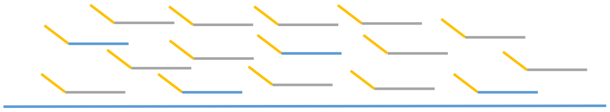


Fig 3B. Although an excess of barcoded sequencing template is produced (short grey and blue lines), only a subset of it is actually sequenced (short blue lines).

HOW DO LINKED-READS DIFFER FROM SYNTHETIC LONG READS?

Linked-Reads are not synthetic long reads. Synthetic long reads attempt to fully reconstruct the long molecule by over-sequencing and then performing local assembly to reconstruct the molecule. This leads to increased template sequencing at the expense of physical coverage. Less physical coverage lessens power to link distant loci (Fig 4A).

By contrast, Linked-Reads allow for significantly increased physical coverage with only a slight increase in standard sequencing. It is this increased physical coverage that provides the power to link distant loci and reconstruct long range haplotypes (Fig 4B). On average, a standard Chromium™ genome run provides 150x physical coverage and 30x sequence coverage at a given locus.

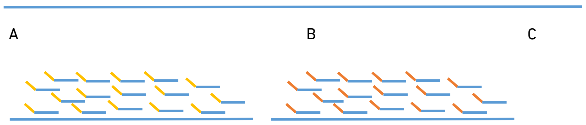


Fig 4A. Synthetic long reads illustrated above demonstrate increased sequencing at the expense of decresased physical coverage. There is an ample amount of read coverage, but without sufficient physical coverage, the three loci (A, B and C) cannot be linked.

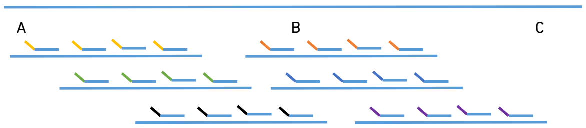


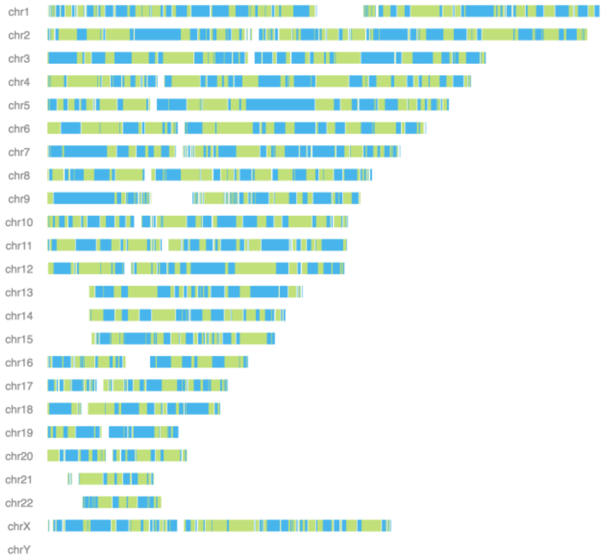
Fig 4B. Linked-Reads, with only a slight increase in standard sequencing, allow increased physical coverage that provides the power to link distant loci and reconstruct long range haplotypes. In the figure above, with the superior physical coverage of Linked-Reads, the three loci (A, B and C) can be linked.

THE POWER OF LINKED-READS

Linked-Reads expand the current potential of standard short read sequencing, opening up previously challenging applications.

Long range haplotype reconstruction

Linked-Reads enable large scale haplotype reconstruction. Fig 5 shows a standard run of the NA12878 genome. Alternating colors delineate phase blocks. At standard sequencing depths, phase block lengths are determined primarily by the length of the input DNA and the diversity of the sample. For the run shown in this figure, the N50 phase block length is 4.6 Mb, and the longest phase block is 31.2 Mb. The input molecule length was 80 Kb.



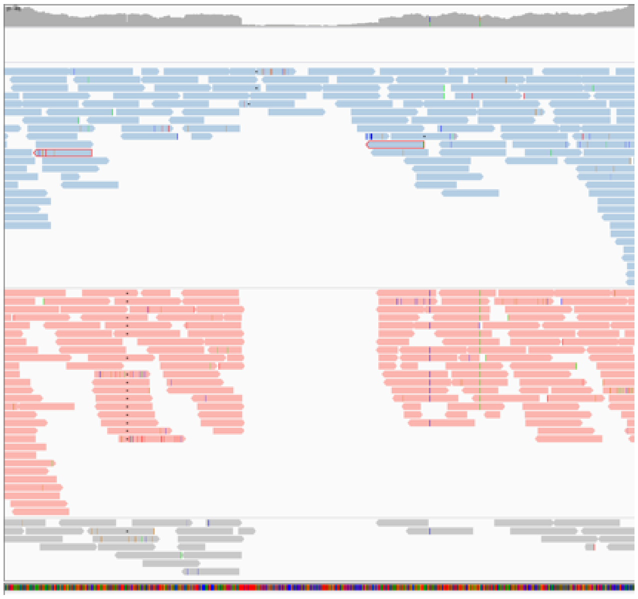


Fig 5. Standard run of the NA12878 genome. Alternating colors delineate phase blocks.

Fig 6. A 325 bp heterozygous deletion detected using Linked-Reads. Reads are partitioned into distinct Haplotypes. Haplotype1 shown in blue, haplotype 2 in pink.

Robust variant calling

Linked-Reads enable improved SV calling. A variety of algorithms allow us to detect a wide range of variants, including SNVs, deletions, inversions and translocations. Fig 6 shows an example of a 325 bp heterozygous deletion. By partitioning the reads into distinct haplotypes (haplotype1 is shown in blue and haplotype 2 is shown in pink), there is no need to average read depth across both haplotypes, thus improving sensitivity to such events.

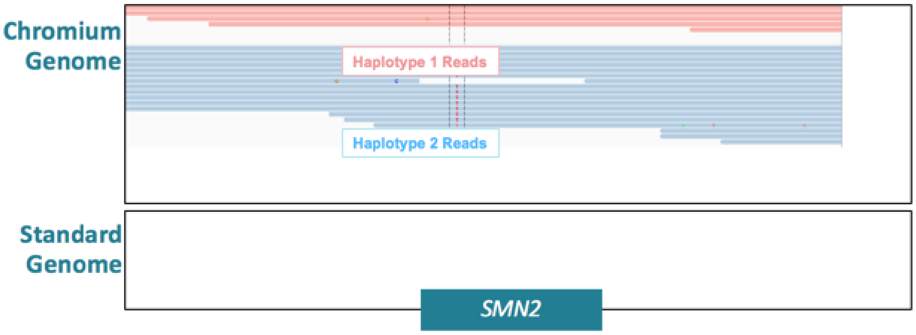


Fig 7. An alignment to the *SMN2* gene, using Chromium™ Genome in conjunction with the Lariat aligner. As *SMN2* has a closely related paralog, *SMN1*, standard short read methods fail to align reads uniquely in this region. Haplotype1 shown in pink, haplotype 2 in blue.

Improved access to inaccessible parts of the genome

The Lariat™ aligner uses barcode information to provide high quality alignments in regions of the genome typically inaccessible to standard short read approaches. Fig 7 shows reads aligned to the *SMN2* gene. *SMN2* has a closely related paralog, *SMN1*. Standard short read methods fail to align reads uniquely in this region. Here we see that Chromium™ Genome, coupled with the Lariat™ aligner, provides both high quality alignments as well as phasing of reads (haplotype 1 is shown in pink, haplotype 2 in blue). The highlighted variant has been validated using an orthogonal technology.

De novo diploid assembly

The Supernova™ assembler utilizes the power of Linked-Reads to reconstruct individual haplotypes independent of a reference. In addition to producing highly contiguous assemblies, Supernova™ does not produce a haploid consensus assembly. Instead, regions of diversity are separated into their individual sequences in order to capture the diploid nature of these samples. Table 1 shows statistics for 7 human assemblies.

Sample	Ethnicity	Sex	Coverage	Fragment	N50 Contig	N50 Scaffold	N50 Phase Block	Gap
NA24385	AJ	M	56	120	106.4	15.1	4.2	2.6
NA19240	YRI	F	56	125	118.8	16.4	9.3	2.3
NA19238	YRI	F	56	115	114.6	18.7	8	2.1
NA12878	EUR	F	56	92	118.5	16.4	2.8	2.9
HGP	EUR	M	56	139	120.2	18.6	4.5	2.5
HG00733	PR	F	56	106	123.6	17.8	3.4	2.0
HG00512	HAN	M	56	102	113.2	15.4	2.7	2.2

Table 1. Statistics for 7 human assemblies

Notices

Document Number

CG00044 Rev A *Technical Note*

Legal Notices

Nothing contained herein shall constitute any warranty, express or implied, as to the performance of any products described herein. Any and all warranties applicable to any products are set forth in the applicable terms and conditions of sale accompanying the purchase of such product. 10x Genomics provides no warranty and hereby disclaims any and all warranties as to the use of any third party products or protocols described herein. The use of products described herein is subject to certain restrictions as set forth in the applicable terms and conditions of sale accompanying the purchase of such product. “10x”, “10x Genomics”, “Changing the Definition of Sequencing”, “Chromium”, “GemCode”, “Loupe”, “Long Ranger”, “Cell Ranger” and “Supernova” are trademarks of 10x Genomics, Inc. All other trademarks are the property of their respective owners. All products and services described herein are intended FOR RESEARCH USE ONLY and NOT FOR USE IN DIAGNOSTIC PROCEDURES.

The use of 10x Product(s) in practicing the methods set forth herein has not been validated by 10x, and such non-validated use is NOT COVERED BY 10X STANDARD WARRANTY, AND 10X HEREBY DISCLAIMS ANY AND ALL WARRANTIES FOR SUCH USE.

Nothing in this document should be construed as altering, waiving or amending in any manner 10x Genomics, Inc., terms and conditions of sale for the Chromium™ Controller, consumables or software, including without limitation such terms and conditions relating to certain use restrictions, limited license, warranty and limitation of liability, and nothing in this document shall be deemed to be Documentation, as that term is set forth in such terms and conditions of sale. Nothing in this document shall be construed as any representation by 10x Genomics, Inc that it currently or will at any time in the future offer or in any way support any application set forth herein.

Customer Information and Feedback

For technical information or advice, please contact our Customer Technical Support Division online at any time.

Email: support@10xgenomics.com

10x Genomics 7068 Koll Center Parkway

Suite 401

Pleasanton, CA 94566 USA