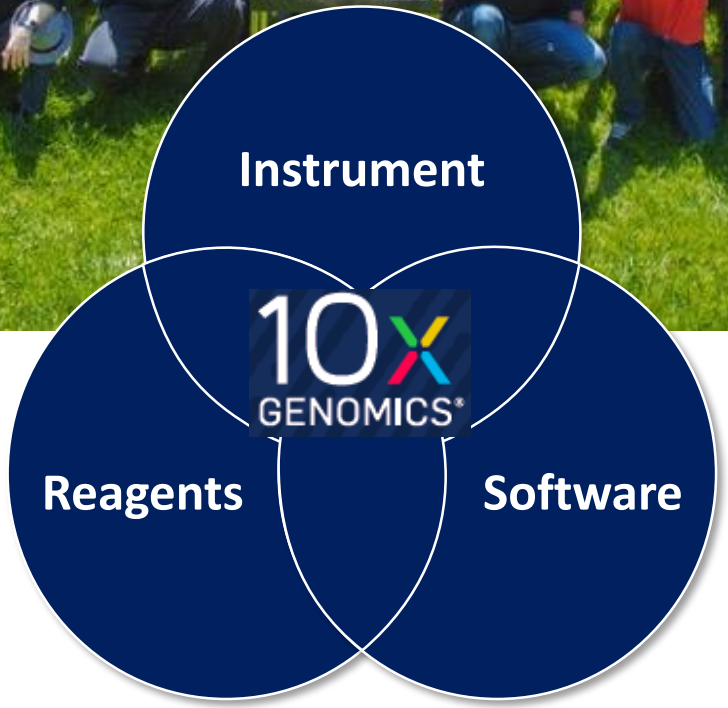
Everyday *de novo* Assembly with the Supernova™ Assembler

Anup Parikh
Director of Product Marketing
10x Genomics

November 2016

Changing the definition of sequencing

Our Solutions. Your Sequencer. Powerful Discovery



The Chromium System

One system, one workflow, powerful new sequencing applications

Single Cell 3′ Solution

Perform deep profiling of complex cell populations with high-throughput digital gene expression on a cell-by-cell basis. Trace expression profiles to individual cells to ensure biologically relevant signals are not masked by bulk average measurements.

Genome Solution

Use the power of Linked-Reads to fully resolve haplotypes, structural variation, and detect variants in previously inaccessible and complex regions of the genome. Ensure biologically relevant variants are not masked by averaging over chromosomes.

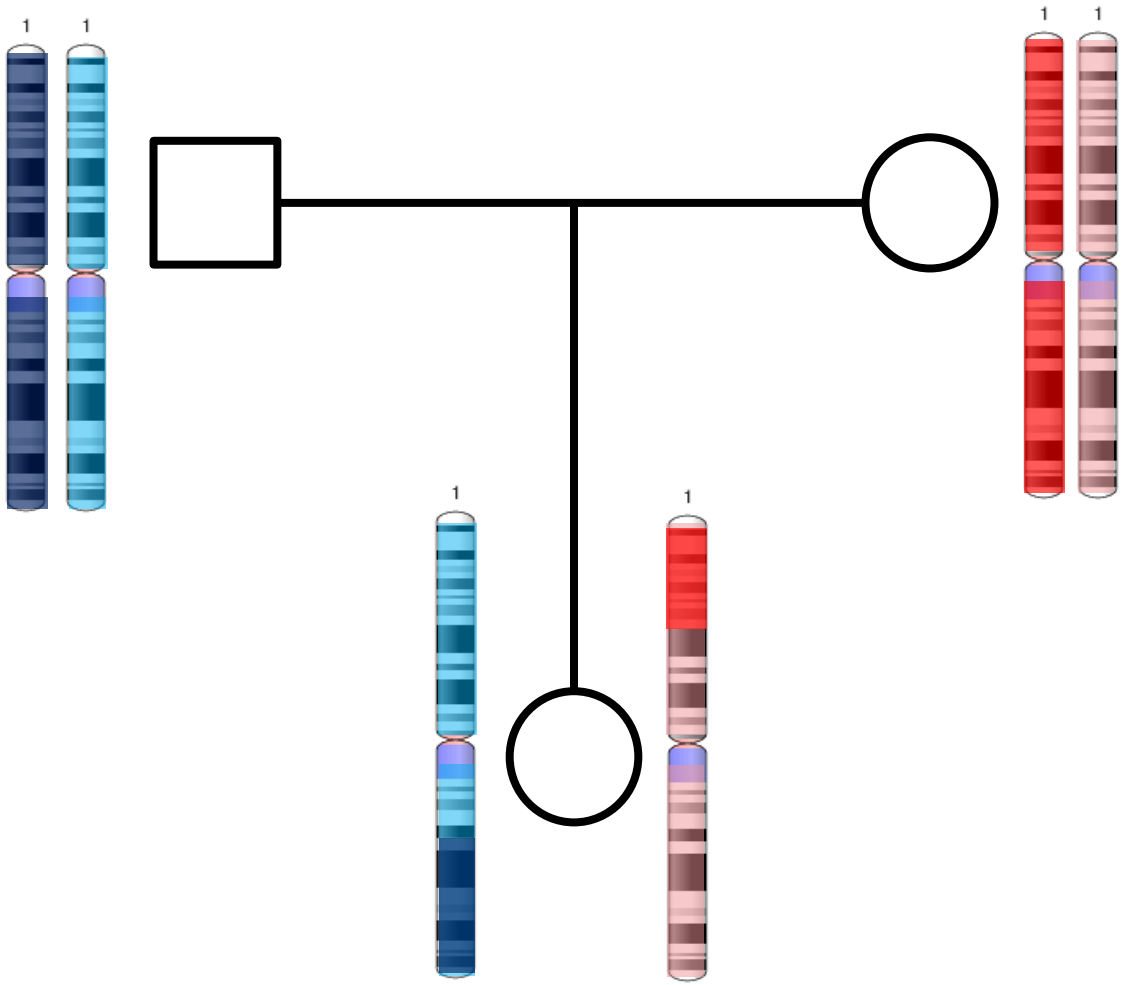
De Novo Solution

Discover the true genome with the Supernova™ Assembler and open the door to low-cost, everyday diploid assemblies. Unlock sample-specific sequence, probe diploid genome structure with a single workflow starting with 1ng DNA input.

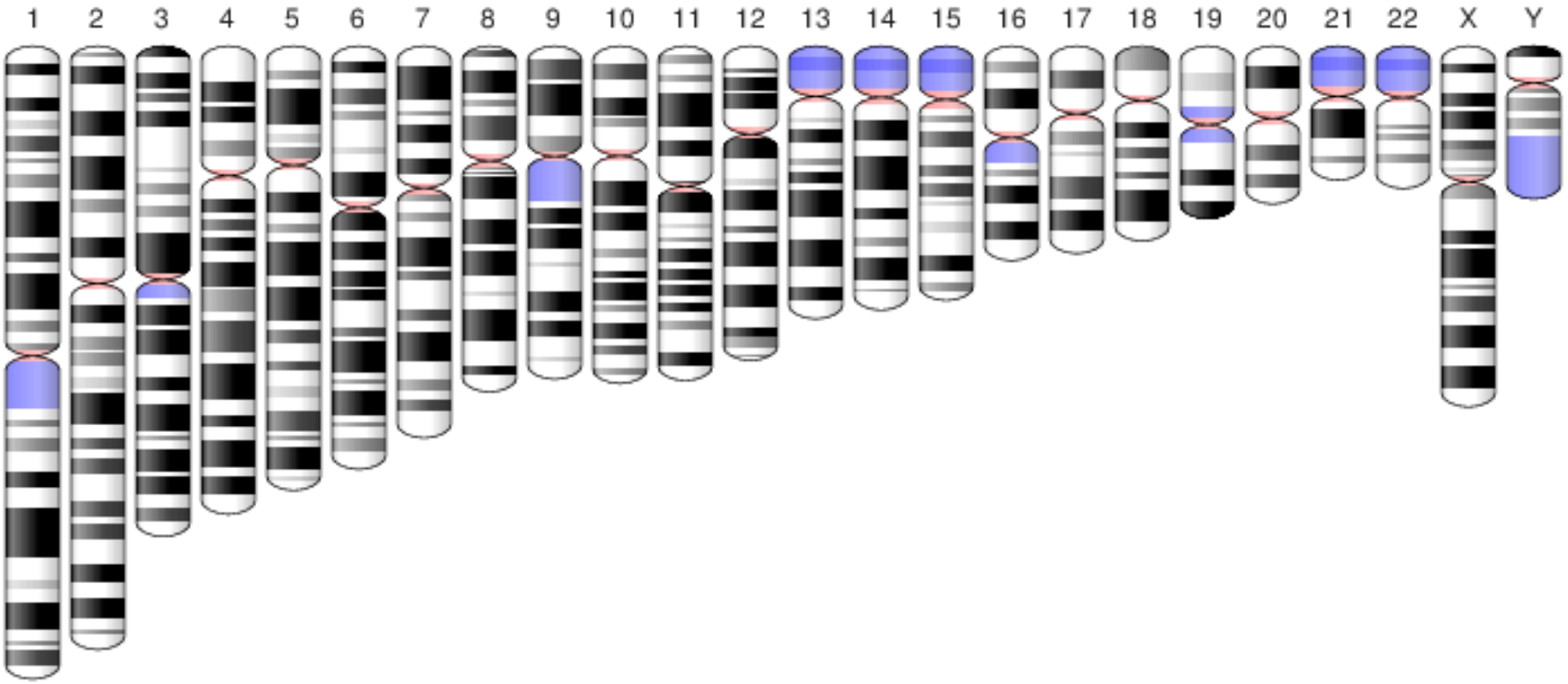
Targeted Solution

Maintain long range information in existing targeted panels with Linked-Reads. Resolve haplotypes, structural variation, and detect variants in previously inaccessible and complex regions without sequencing the whole genome.

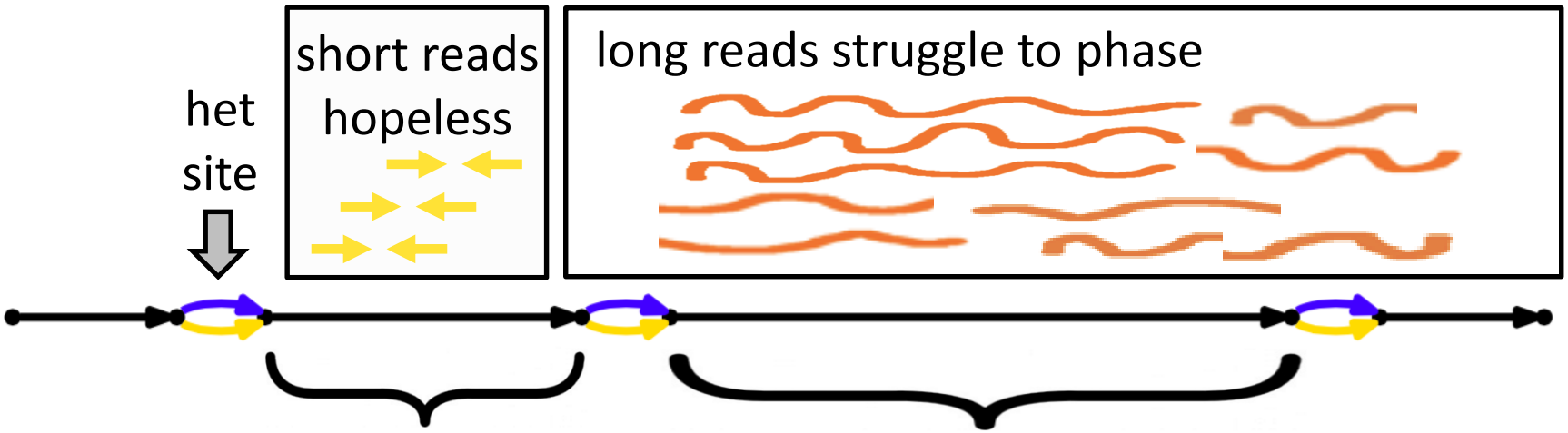
Our actual genome: diploid



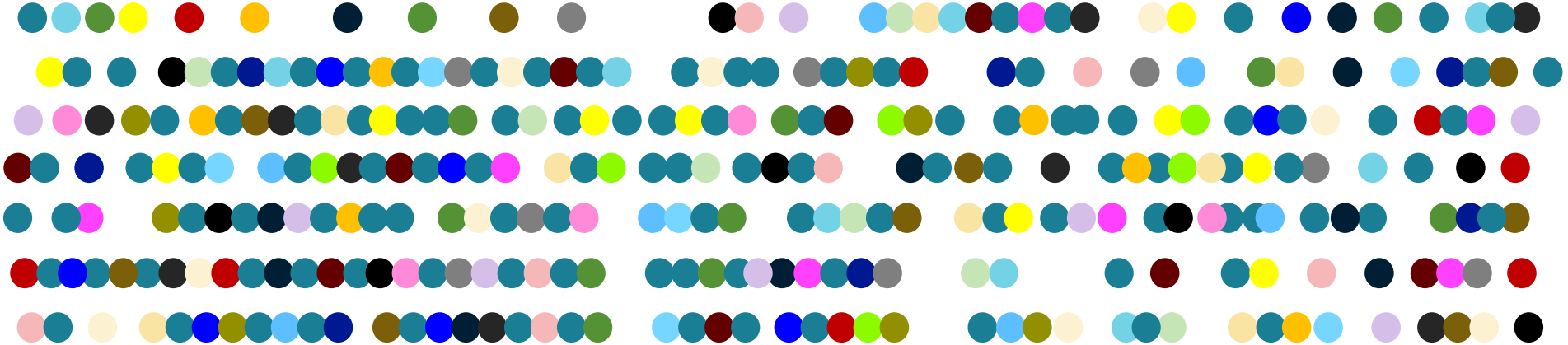
How we represent our genome: haploid



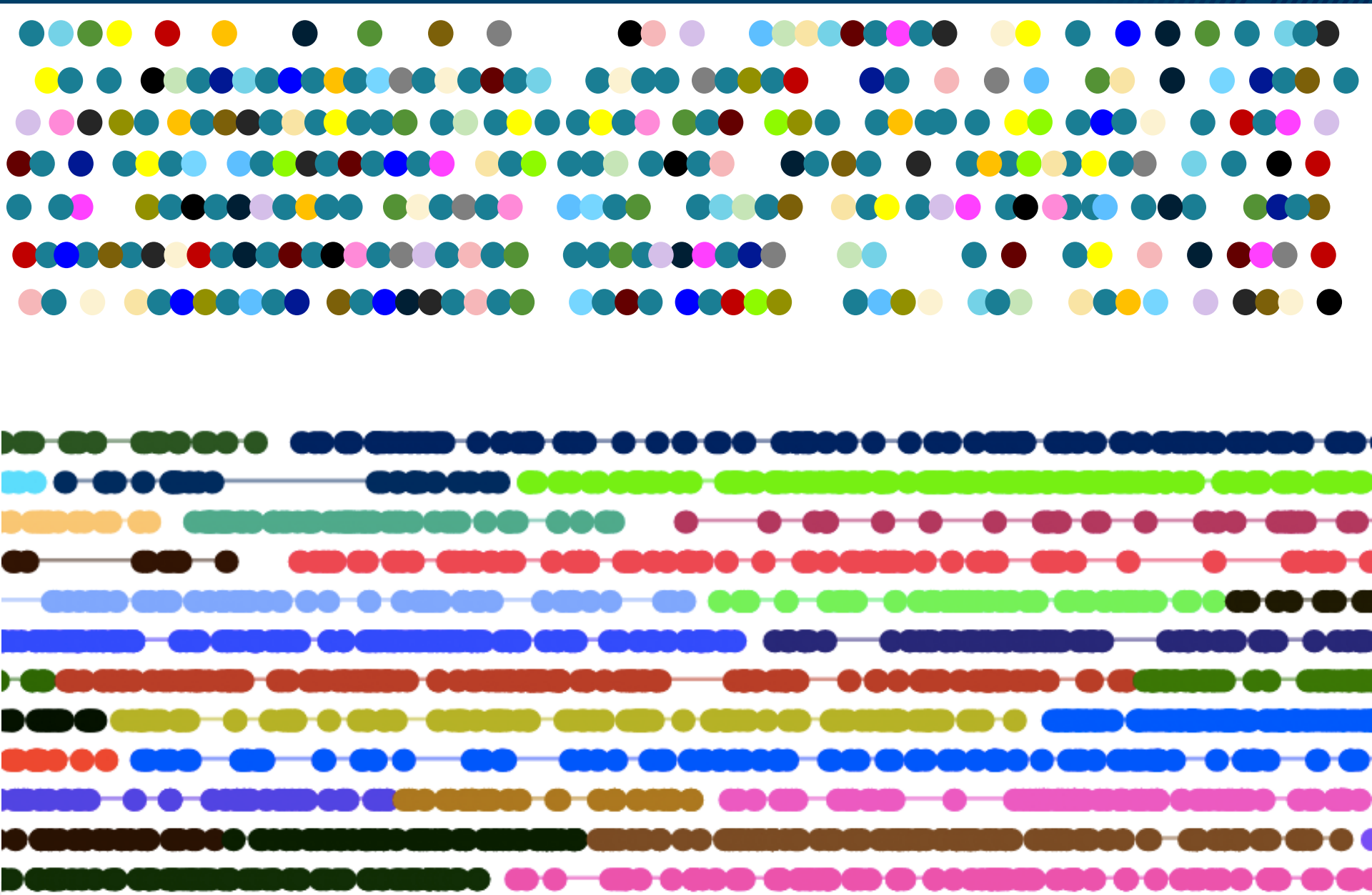
Single reads too short to phase large genomes



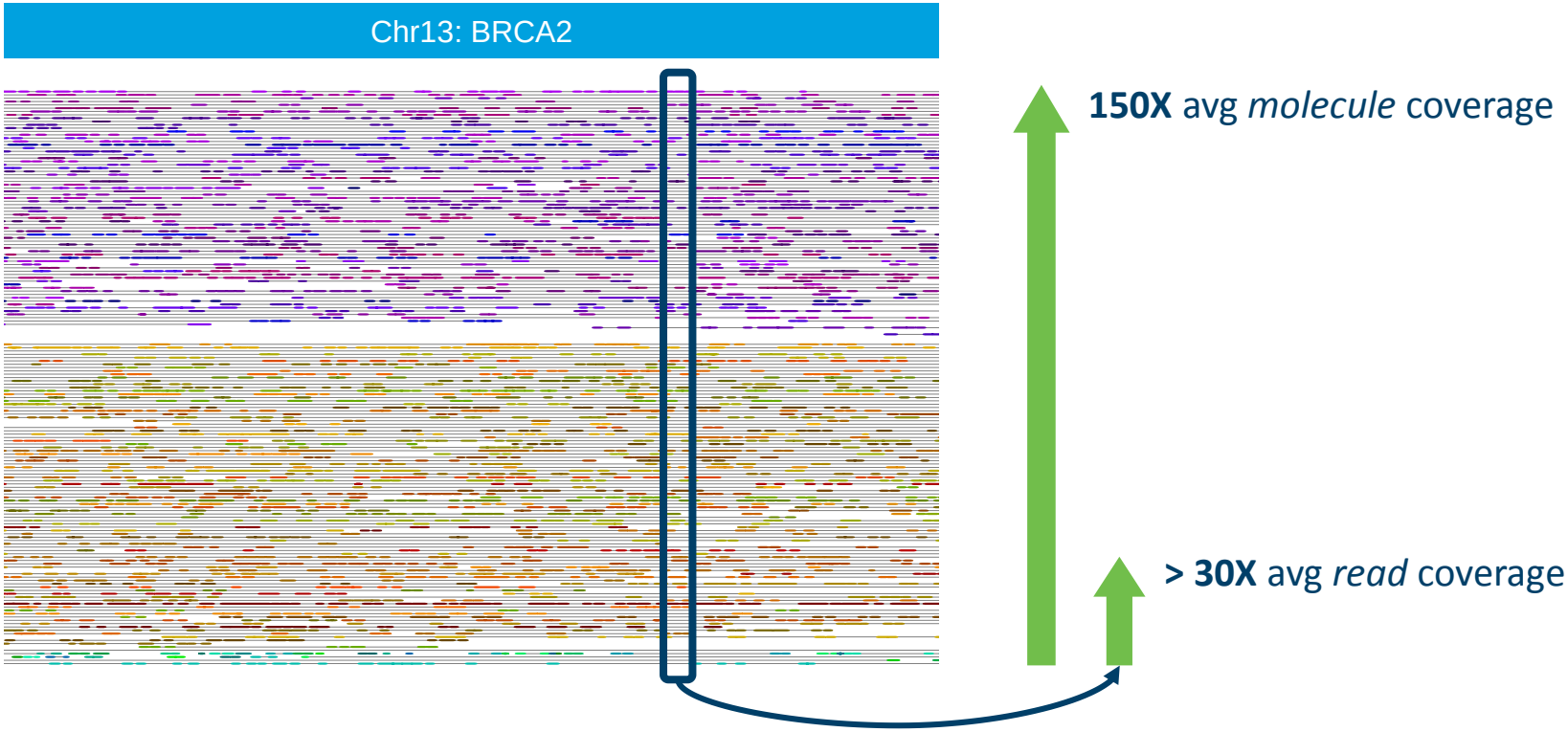
Un-Linked-Reads: short range information



Linked-Reads: long range information



Example – Molecule vs Read Coverage



A given genomic locus will have

150X avg molecule depth, and **30X** avg read depth

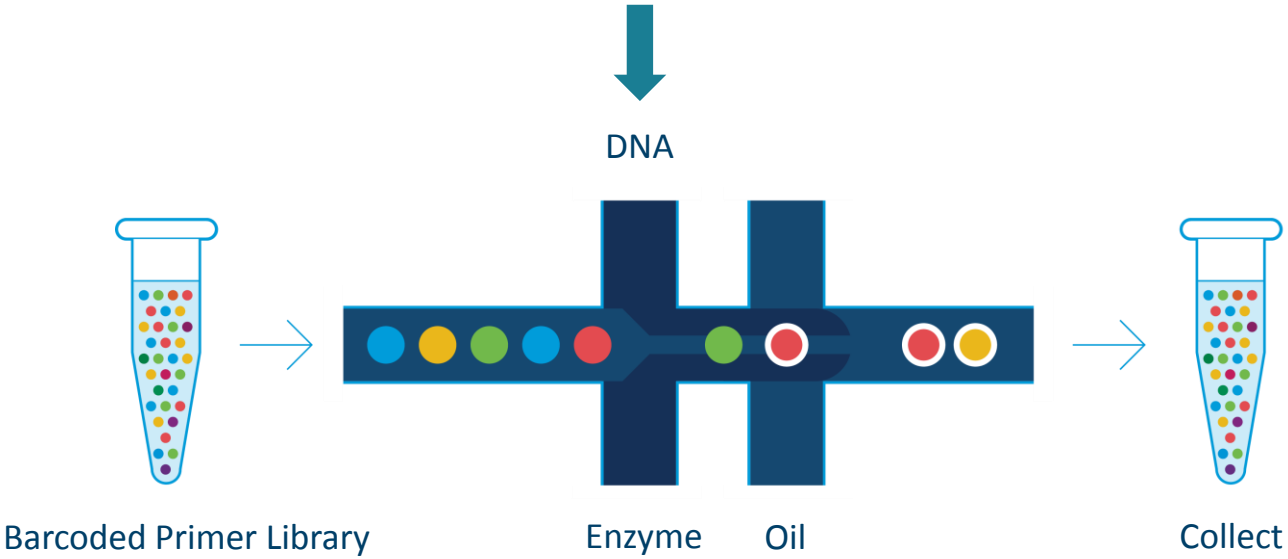
(150X molecule depth) *x* (0.2X read/m) = 30X read depth

Making Linked-Reads: Partitioning

Start with:

HMW gDNA, 100Kb+ molecules

1.0 ng input DNA = 300 copies of the genome



0.5ng DNA = **150** copies of the genome,

partitioned into > 1M GEMs

Mapping difficult in repetitive regions of genomes

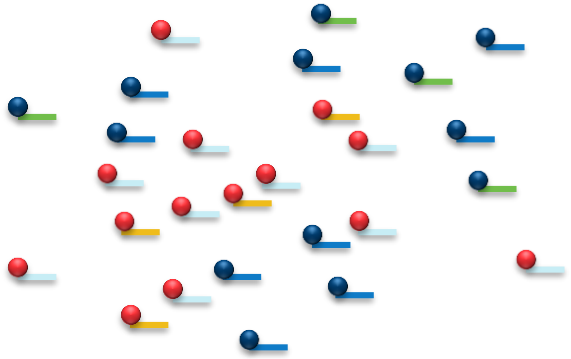
10X GENOMICS

Short Reads

Short Read Aligners Cannot Place Reads Correctly

Long Range information Rescues Critical Content

Linked-Reads



1. *Confident* mapping provides anchors
2. *Barcodes* recruit short reads into paralogous loci

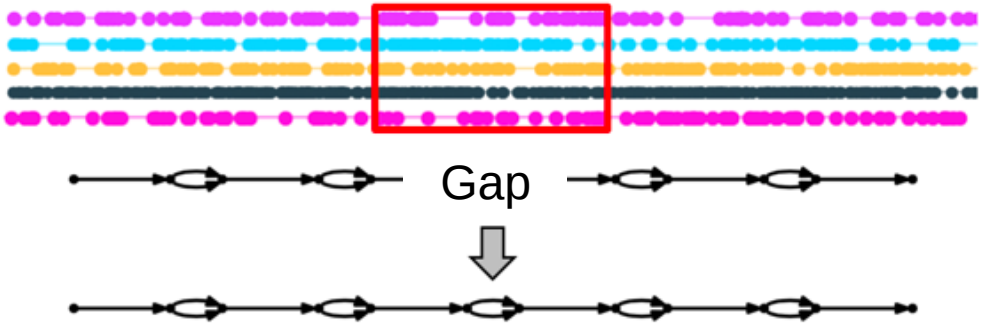
Lariat™ Aligner Correctly Places Short Reads Even in Paralogous Loci

Linked-Reads help resolve repeat gaps and phase assemblies

10X GENOMICS

Linked-Reads enable assembly to repeat regions previously inaccessible

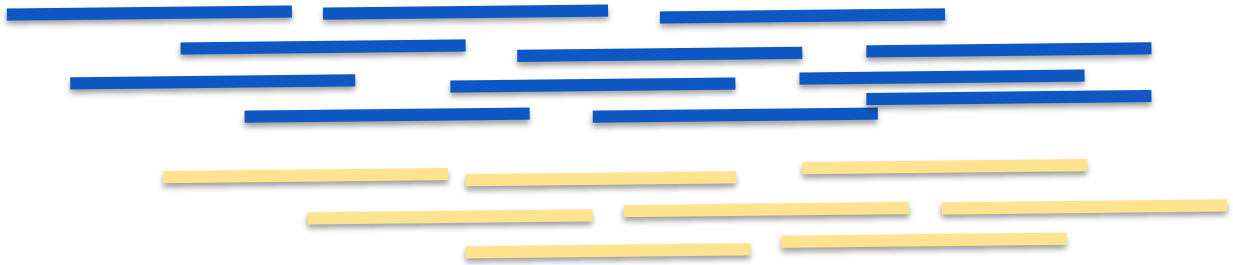
We use the barcodes to locate and fill gaps (labeled 'stuff' below). Briefly, the barcodes allow us to infer a pool of reads that should cover the gap, consisting of all the reads in certain barcodes. After creating a local assembly from this pool, it can be reinserted into the global graph.



Old Assembly: Compress Haplotypes

Sequences from haplotype 1

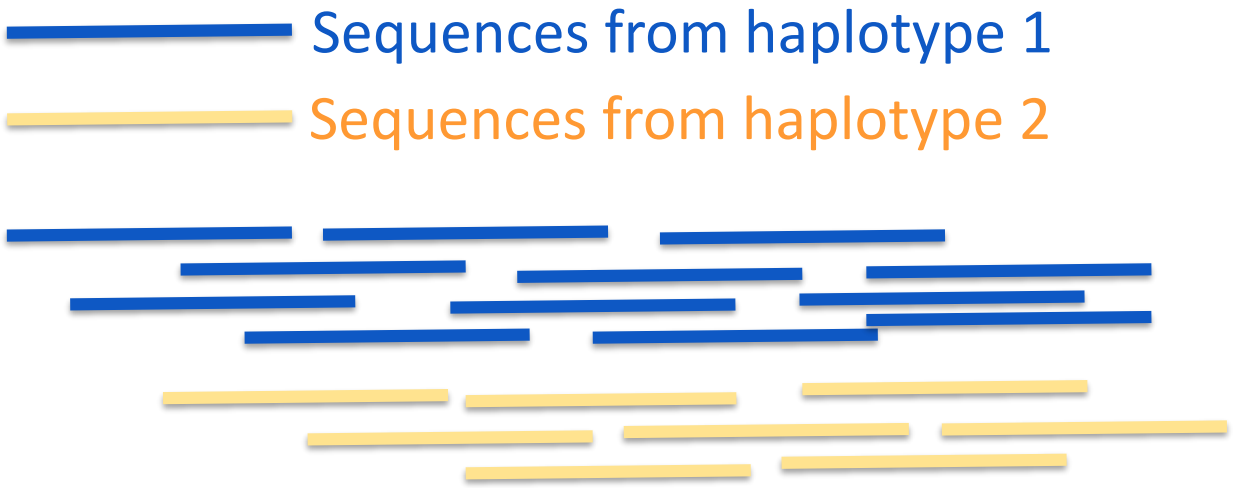
Sequences from haplotype 2



Old Assembly model: compress into a consensus

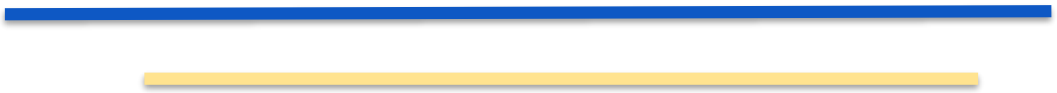
Church et al., 2011 PLoS Biology

Chromium™ Genome and Supernova™ enable true diploid assembly



Old Assembly model: compress into a consensus

New Assembly model: represent both haplotypes



Church et al., 2011 PLoS Biology

Megabubbles Capture High Heterozygosity

Small chunk of a megabubble from an inbred organism

```
         * *                                        *
CAAGTAAACCTATCCATAATGCATACAACCATACATATATATAACTACGCAGGAAAAGGGAATCGGGTTTAGTATGCATT
CAAGTAAAACCATCCATAATGCATACAACCATACATATATATAACCACGCAGGAAAAGGGAATCGGGTTTAGTATGCATT

            *             *       *                              *
TAAAATCTTTACCTGGCTTGTGTAGCAATTTTCGTCCTAAACCACCCACGGGCTATATGGGTTAGCTCCCCCTGCTGAAA
TAAAATCTTTACATGGCTTGTGTAGTAATTTTCATCCTAAACCACCCACGGGCTATATGGGTCAGCTCCCCCTGCTGAAA

                                        **
GGGCCCCAATACATGAAGTTACACGATTAGAGAAAATAAGTTAGAATGACTAGTACATGCCCCAATATAAAGTGTGATTT
GGGCCCCAATACATGAAGTTACACGATTAGAGAAAATAACCTAGAATGACTAGTACATGCCCCAATATAAAGTGTGATTT

          *                                          * *     **
CATGCACACGATTACCAAGACCAACCTCATATCGGCAAATATGAGTTTCCAGTTTTAAGCCCCTGGAGACCTCCACCTTC
CATGCACACGGTTACCAAGACCAACCTCATATCGGCAAATATGAGTTTCCAGTTTTGATCCCCCTGAGACCTCCACCTTC

                *                                    *                          *
CACGACTTTGCTACACTTTCCCCTTTAATGCACTACAAGAAATTACCCCTATTGCAATGAATCTTTTGGCAATGGTTCCA
CACGACTTTGCTACATTTTCCCCTTTAATGCACTACAAGAAATTACCCTTATTGCAATGAATCTTTTGGCAACGGTTCCA
```

Let's Assess Phasing Accuracy

0 1

NA24385

Some bases in the child's genome can be phased

0 **0** **1** **0**

ACTTAGCATTAC ••• CTCCACGT ••• TACGGGTA ••• GGTAATAATTAC

0010 *should be all zeros or all ones*

Now do this for a three megabase region

Long Read Contigs Crisscross Chromosomes

ftp-trace.ncbi.nih.gov/giab/ftp/data/AshkenazimTrio/analysis

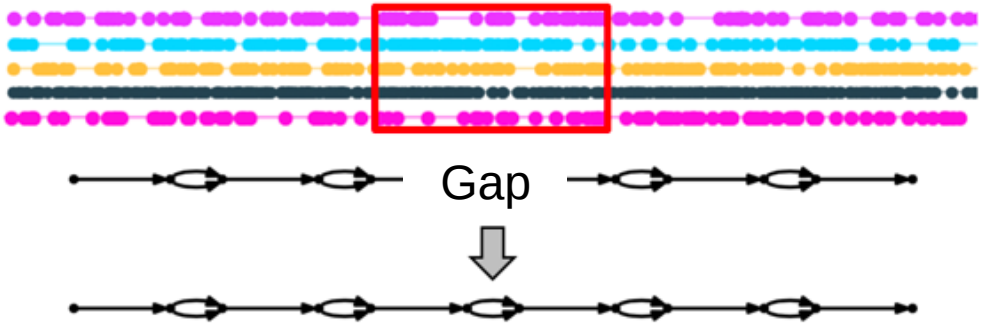
Supernova™ Mirrors Actual DNA (3 Mb)

Supernova™ Mirrors Actual DNA (3 Mb)

Linked-Reads help resolve repeat gaps and phase assemblies

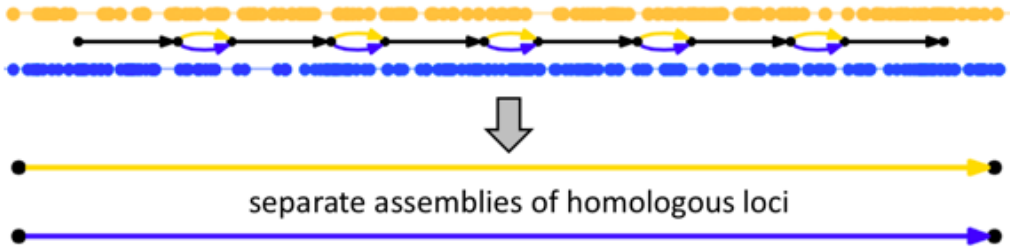
Linked-Reads enable assembly to repeat regions previously inaccessible

We use the barcodes to locate and fill gaps (labeled 'stuff' below). Briefly, the barcodes allow us to infer a pool of reads that should cover the gap, consisting of all the reads in certain barcodes. After creating a local assembly from this pool, it can be reinserted into the global graph.

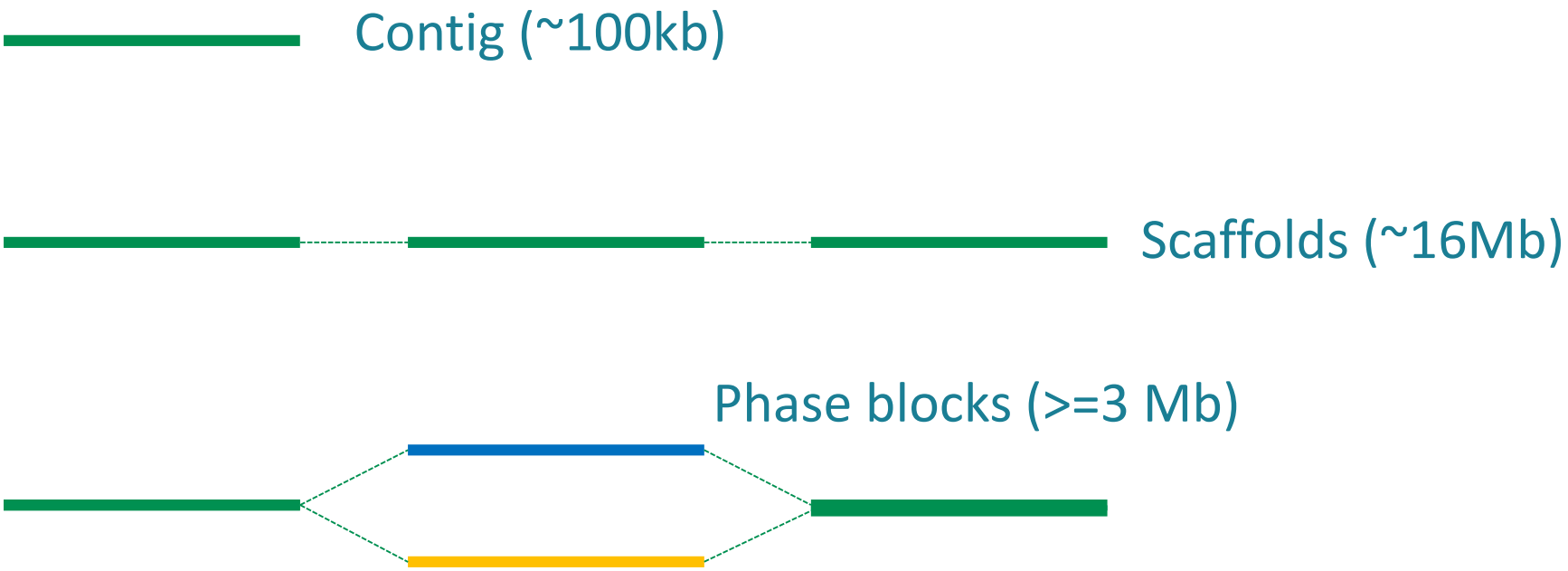


Supernova™ utilizes Linked-Reads to create phased assemblies

In the simplified diagram below, one parental allele is gold, while the other is blue. We show one barcoded molecule landing on all the gold alleles, and one landing on all the blue alleles, thus pulling them apart. In reality, many barcoded molecules collude to execute this phasing operation.



Key Assembly structure and metrics



- **Contig**: an ungapped sequence.
- **Prefect Stretch**: prefect base level sequence accuracy
- **Scaffold**: a gapped sequence containing multiple contigs for which the order and orientation is asserted. Scaffold length driven by molecule length.
- **Phase Block**: True diploid sequences. Dependent on the # of HET variant and the length of the input molecules. Phase block length driven by diversity and molecule length

Released De Novo assembly of NA12878 human genome

Sample input and Sequencing	Supernova™ Compute Requirements	Supernova™ Assembly	
• Input DNA: 1.25ng • Molecule size: 80.3 kb • 2x150bp on Illumina HiSeq X Ten • 1200M reads (56x coverage)	• 28 cores (1 server) running for 48 hours • 1,344 total core-hours • 2TB of data generated	# of Scaffold > 10Kb	1651
		N50 contig	103.67 kb
		N50 scaffold size	14.06 Mb
		Assembly size (scaffolds >= 10 kb)	2.74 Gb
		N50 phase block size	1.96 Mb

View and download at
http://software.10xgenomics.com/de-novo-assembly/overview/datasets

Supernova™ – System Requirements

Local Mode

- Run on single, standalone Linux system
- CentOS/RedHat or Ubuntu
- At least 24 cores, 384GB RAM, and 2TB disk

Runtime

- 180Gb genome: 1,300 core-hours (48hrs on 28 cores)

2 commands, 1 minute to install.
No make. No compile. No dependencies.

```
$ wget http://10xgenomics.com/Supernova™-1.0.tar.gz
$ tar xf Supernova™-1.0.tar.gz
$ Supernova demux --run=/mnt/hiseq/160123_SL-HLA_1234_BHAD58ADXX
$ Supernova run   --id=NA12878 --fastqs=HAD58ADXX/
```

Lots of Wild Stuff to Assemble

Characteristics of supported genomes

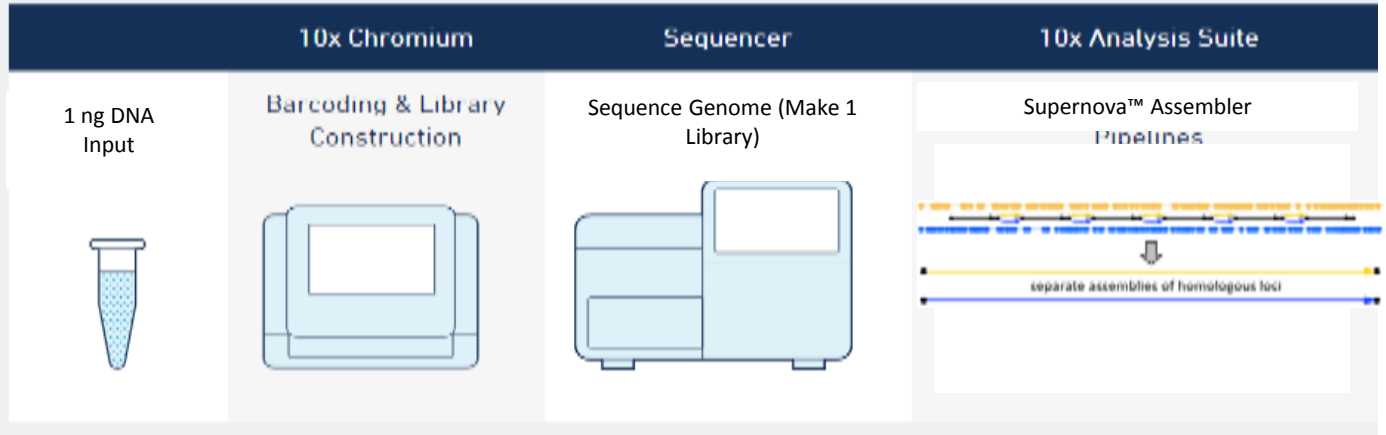
- Genome size: We recommend that the genome size be in the range 1-3.2 GB total.
- Supernova™ well tested on diploid genomes. Not tested on haploid but should work well. Polyploid genomes have not tested.
- We strongly recommend that DNA be obtained from an individual organism or clonal population.
- Supernova™ has not been tested on genomes having repeat content far greater than human, nor on genomes having extreme GC content.
- Visit support.10xgenomics.com for more information on sample and compute guidance.

High Quality Supernova™ Assemblies from diverse genomes

	sample	Size (Gb)	DNA size (kb)	N50 contig (kb)	N50 scaffold (Mb)	HET SNP spacing (kb)	N50 phase block (Mb)
human	NA12878	3.2	95.5	85.0	12.8	1.7	2.8
human	NA24385	3.2	111.3	90.0	10.4	1.5	3.9
human	HGP*	3.2	138.8	104.9	19.4	1.5	4.6
human	Yoruban	3.2	126.9	100.5	16.1	1.1	11.4
nonhuman	Komodo dragon	1.8	85.4	95.3	10.2	10.3	0.4
nonhuman	Spotted owl	1.5	72.2	118.3	10.1	7.1	0.2
nonhuman	Hummingbird	1.0	86.2	87.6	12.5	0.4	10.1
nonhuman	Monk seal	2.6	92.3	93.8	14.8	17.7	0.6
	Chili pepper	3.5	53.3	84.7	4.0	0.4	2.1

* HGP shows 19Kb N50 perfect stretch

End to end solution for high quality diploid assembly in <2 weeks



DNA extraction (1hr-3 days)	Library prep (2 days)	Sequencers (<3 days)	Supernova™ Assembly (2 days)
• Blood (1.5 hours) • Cells (1.5 hours) • Gel plug (3 days) • DNA size > 50 kb acceptable >100 kb Preferred	• Single library from just 1ng of DNA	• HiSeq 2000/2500 • HiSeq X • 2x150 reads	• Run on single, standalone CentOS/RedHat or Ubuntu system • At least 24 cores, 384GB RAM, and 2TB disk • Output: diploid assembly