UCDAVIS Bioinformatics Core

Towards Gapless, Chromosome Scale, Haplotype Assemblies

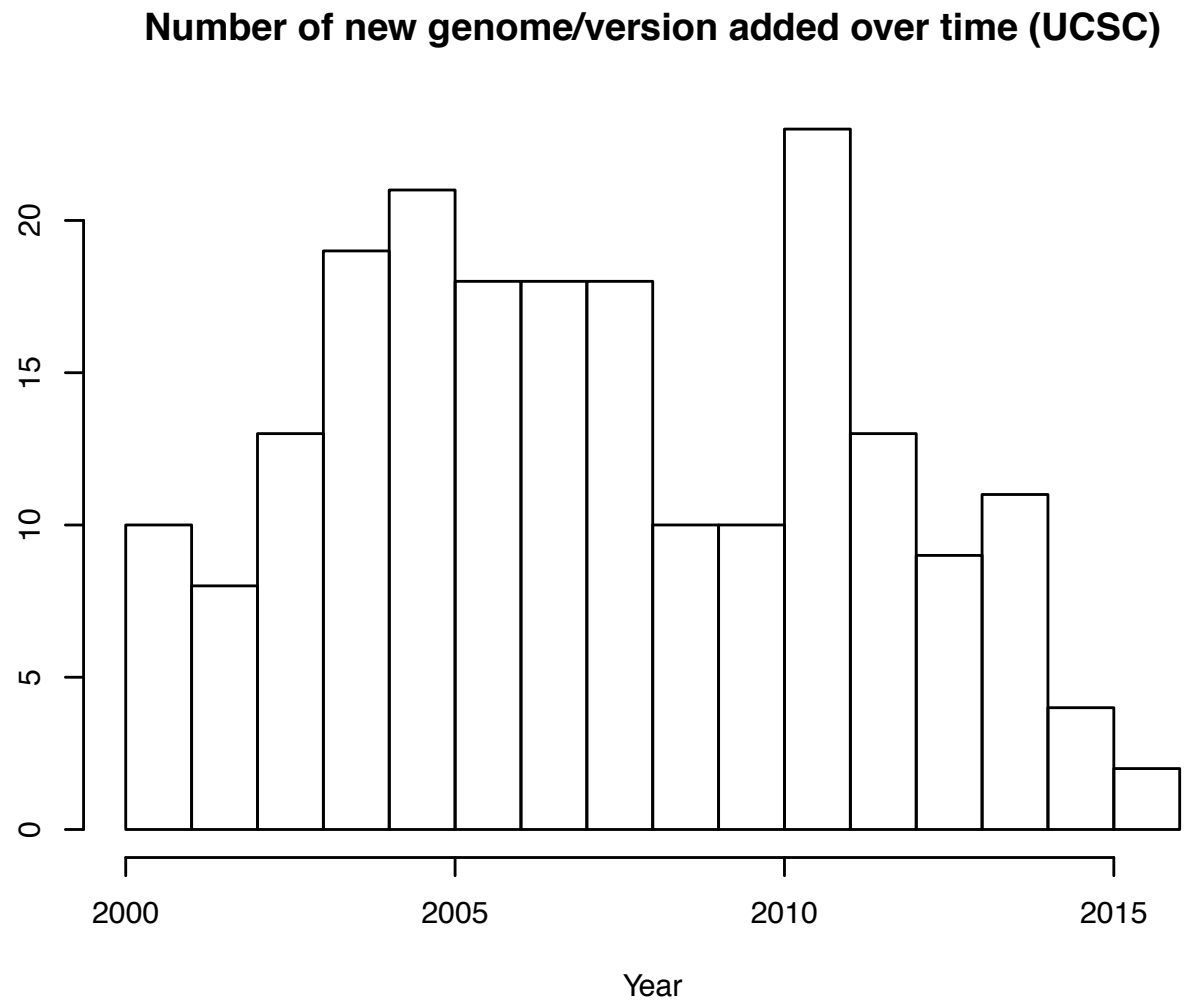
Matt Settles

UC Davis Bioinformatics Core

Human Genome

- In 1990, the National Institutes of Health (NIH) and the Department of Energy joined with international partners to sequence the human genome.
- In April 2003, researchers successfully completed the Human Genome Project, under budget ($2.7B) and more than two years ahead of schedule.
- Thousands of people contributed the Human Genome Project
- Even so, there remains ~400 gaps in the human reference sequence assembly representing hundreds of millions of bases.

Slow down at UCSC of adding genomes



Renewed focus on genomes

- Sequencing has become more democratic. For example, it took more than 50 people, around a dozen centers, $50 million and half a decade to generate a draft chimpanzee genome, published in 2005. This year, Eichler's lab completed a gorilla sequence for about $70,000. “That, to me, is a big deal,” he says.
- Also a big deal, says Eichler, is the quality of their sequences. An earlier version of a gorilla genome was published in **2012** but that was done with shorter pieces of DNA, and therefore left hundreds of thousands of gaps. His team used long-read technology, closed 90 percent of those gaps, and was able to complete many genes that were only partially sequenced in the first attempt.

Speed-reading the genome:
Cheaper methods of sequencing are opening up doors for new research and new career paths.
http://www.nature.com/naturejobs/science/articles/10.1038/nj0492

Gorilla Genome

Assembly	2012 Illumina Assembly	2016 Pacific Biosystems Assembly
Total length	3,041,976,159 bp	3,080,414,926 bp
Contigs	**465,847**	**16,073**
Total contig length	2,829,670,843 bp	3,080,414,926 bp
Placed contig length	2,712,844,129 bp	2,790,620,487 bp
Unplaced contig length	116,826,714 bp	289,794,439 bp
Max. contig length	**191,556 bp**	**36,219,563 bp**
Contig N50	**11.6 kb**	**9.6 mb**
Scaffolds	22,164	554
Max. scaffold length	10,247,101 bp	110,018,866 bp
Scaffold N50	914 Kb	23.1 Mb

2012 Assembly: ABI capillary sequence and short 35bp Illumina sequence + BAC PE data
2015 Assembly: PACBIO SMRT sequence + BAC PE data, INDEL corrected with Illumina sequence

Advances in high-noise, long-read assembly algorithms

- Summer of 2015
  - Pacific Biosystems Falcon assembler for SMRT assembly of large genomes
  - Canu fork of Celera Assembler for single-molecule high-noise sequences.

- Key features:
  - Discard all reads shorter than **X** bp to load into the overlapper, step significantly reduces the number of reads being analyzed.
  - Self correct reads from all-by-all overlaps (takes advantages of cluster env.)
  - Build a graph based on high quality, long corrected reads.
  - “Polish” the resulting assembly using all reads, 60x coverage produces high quality final contigs.

The 'Next, Next' Generation Sequencers (3rd Generation)

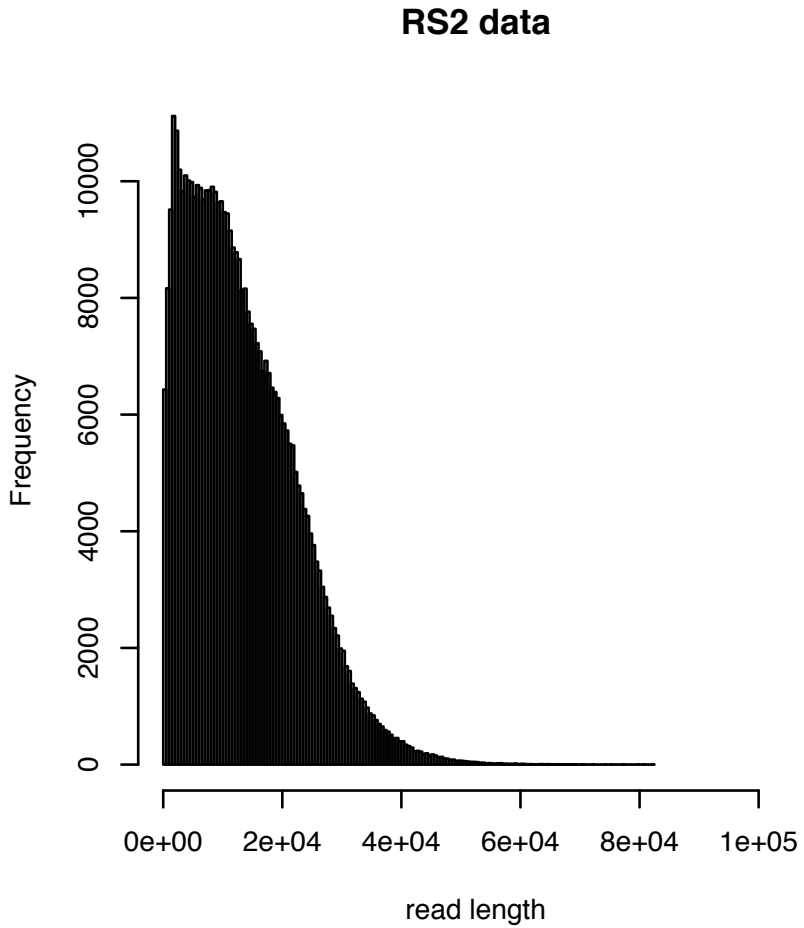
- 2009 – Single Molecule Read Time sequencing by Pacific Biosystems, most successful third generation sequencing platforms, RSII ~1Gb/zmw, new Pac Bio Sequel ~7Gb/zmw, near 100Kb possible read length.

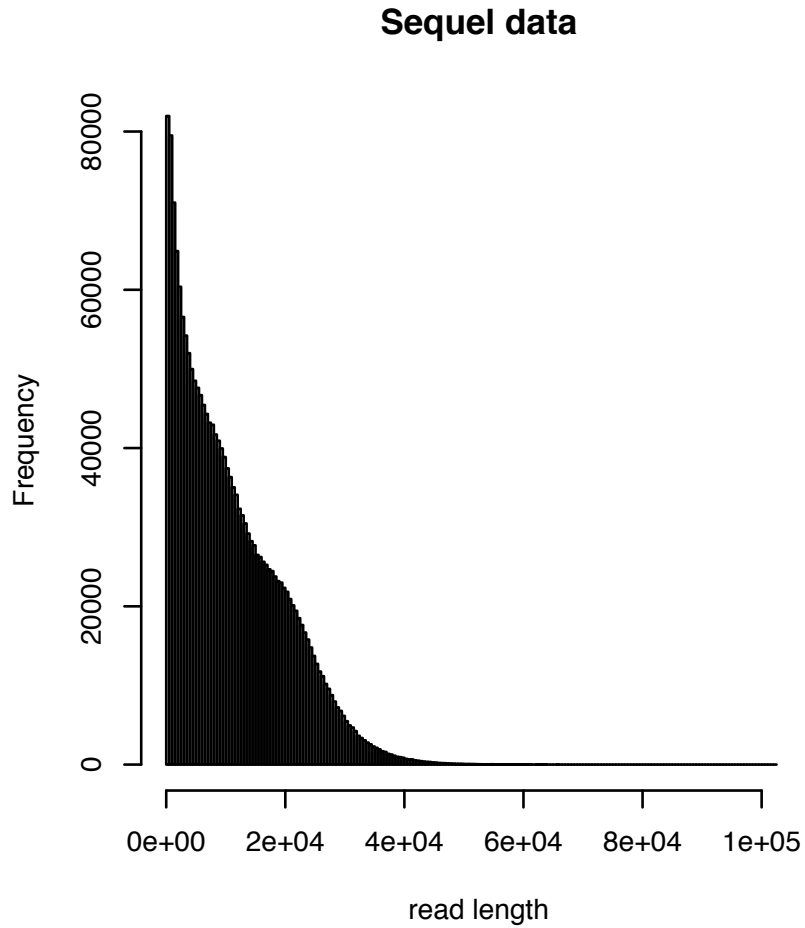
UCDAVIS Bioinformatics Core

Pac Bio Advances (RSII vs Sequel)

California Condor data (~1.2Gbp genome) based on 4 SMRT cell in Jan 2017

	RS2	Sequel
Read count	448,767	1,947,684
N50	10,426	4,293
Longest Read	82,366	102,310
# reads > 12Kb	217,691	754,157
Coverage > 12Kb	3.64	12.165





Assembly (60 SMRT cells)	Total assembly size	N90	N50	Number of contigs	Largest contig	Smallest_contig
Falcon + Quiver Polishing	1,239,863,868	1,106,390	17,286,884	1,164	77,968,233	2,802
Canu	1,240,661,679	1,080,915	14,278,087	1,004	45,704,690	1,812

UCDAVIS Bioinformatics Core

Oxford Nanopore

- 2015 – Another 3rd generation sequencer, founded in 2005. The sequencer uses nanopore technology developed in the 90's to sequence single molecules. Throughput is about 5-10Gb per flowcell, capable of near 200kb reads.

MinION

PromethION

Towards Gapless assemblies

❖ **Promise**

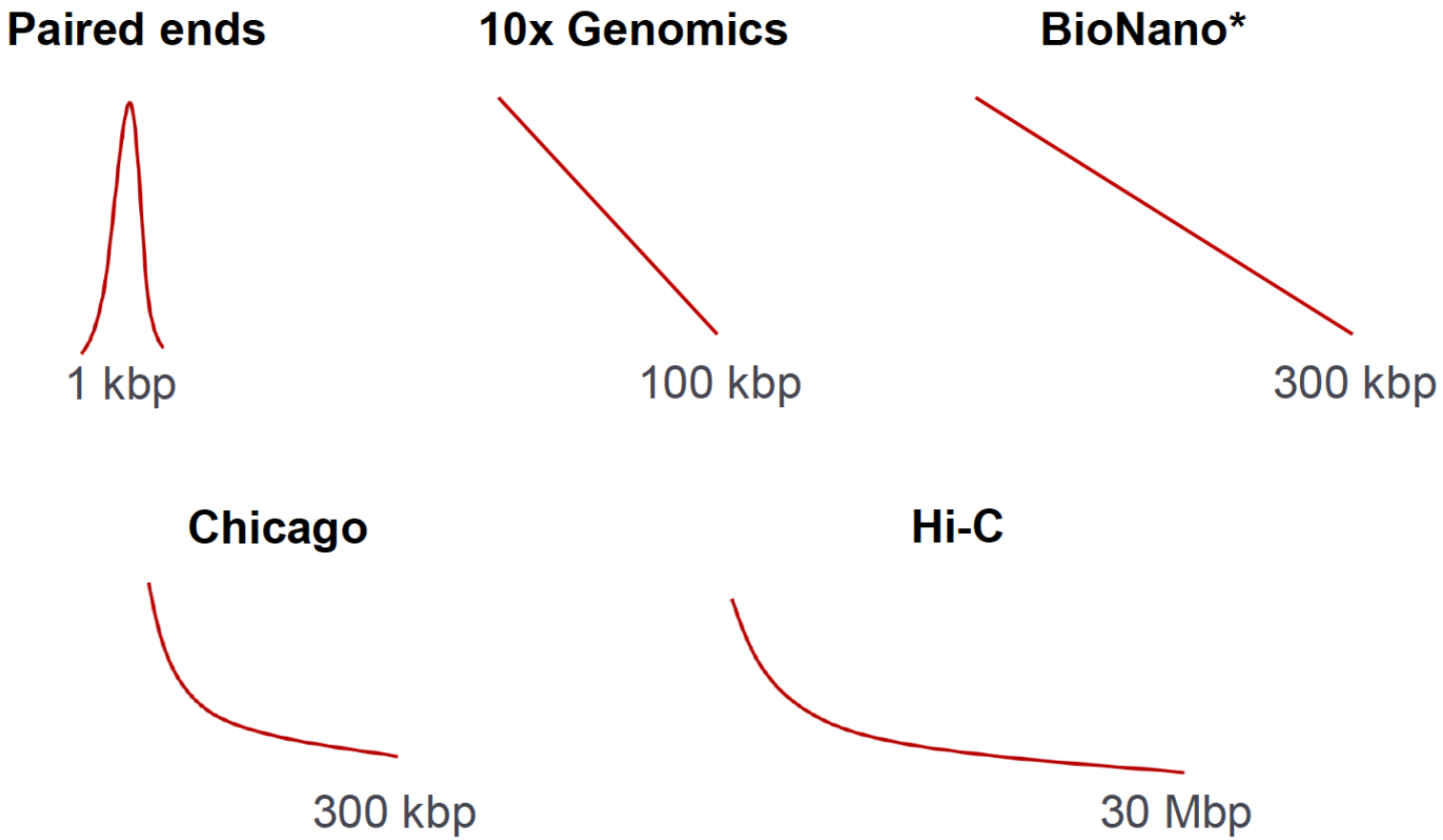
- Continued progress on DNA input and resulting PacBio/Oxford Nanopore Read Lengths and read depth will result in longer N50/N90 fewer resulting contigs.
- Algorithms are still young and have room for improvement.

❖ **Issues**

- Some mis-assemblies are still present, Chimeric reads are an issue
- Small INDELs are an issue and require cleanup (Illumina reads), especially within genes.

UCDAVIS Bioinformatics Core

Scaffolding Options

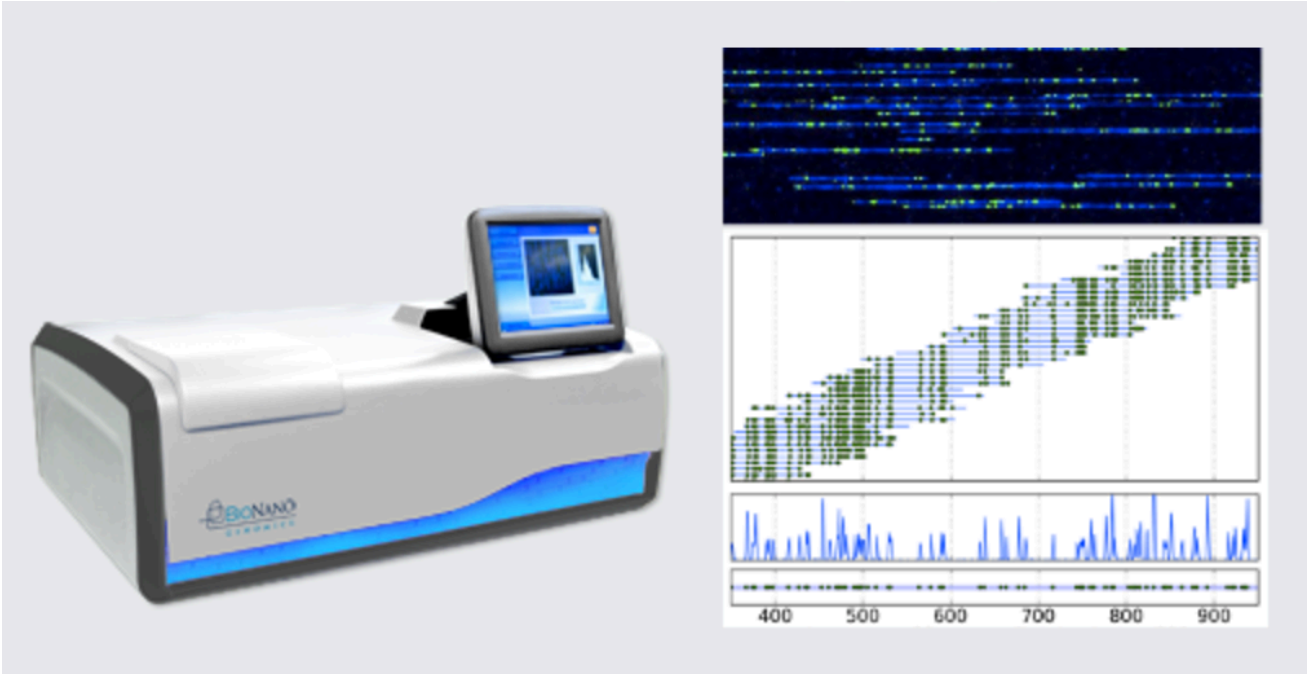


Borrowed from Sergy Koren talk from PacBio Informatics Developer Meeting in Jan 2017

Bionano Irys/Saphyr

- The Irys/Saphyr System puts the power of optical genome mapping. No more waiting for months to get a physical genome map. Bionano Next-Generation Mapping (NGM) provides long-range information to reveal true genome structure. Assists genomes assembles to near chromosomal arms.

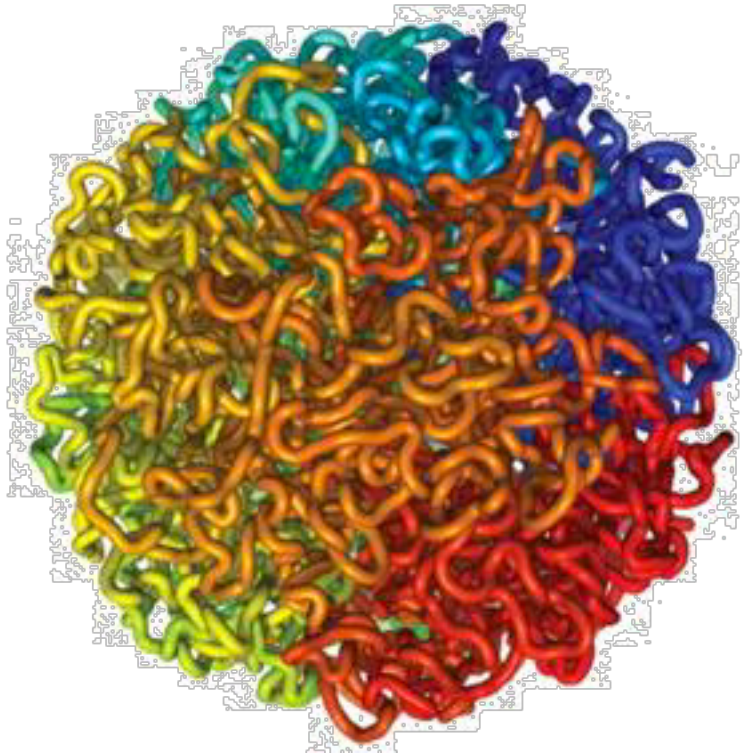
Not sequencing based



UCDAVIS Bioinformatics Core

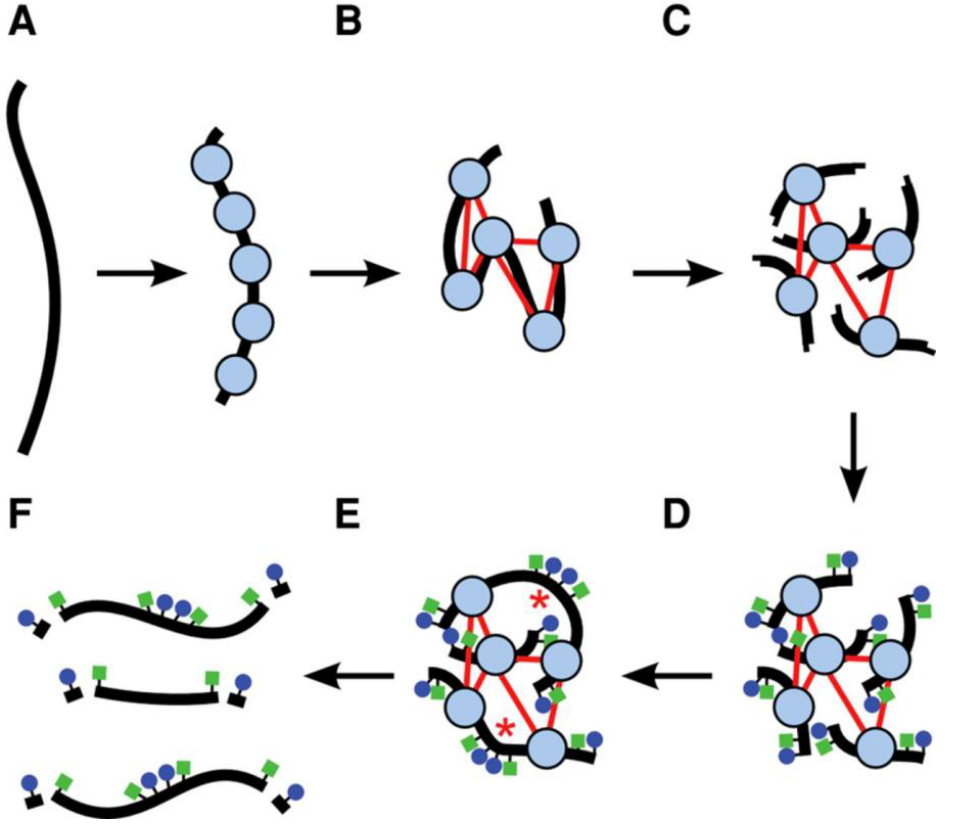
Dovetail and Hi-C (Cross Linking) on Illumina

Hi-C
Proximity Guided Scaffolding



Dovetail and Phase Genomics

Dovetail Chicago Libraries



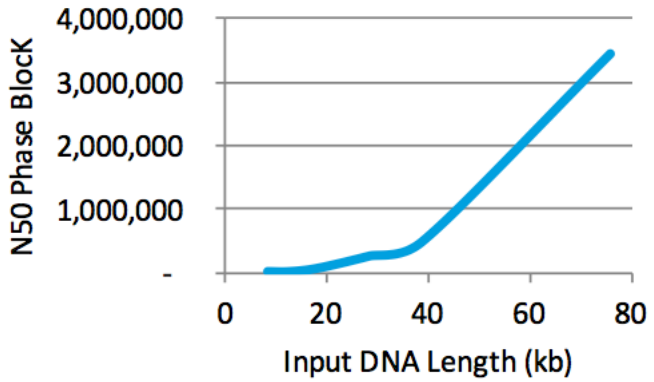
10x genomics on Illumina

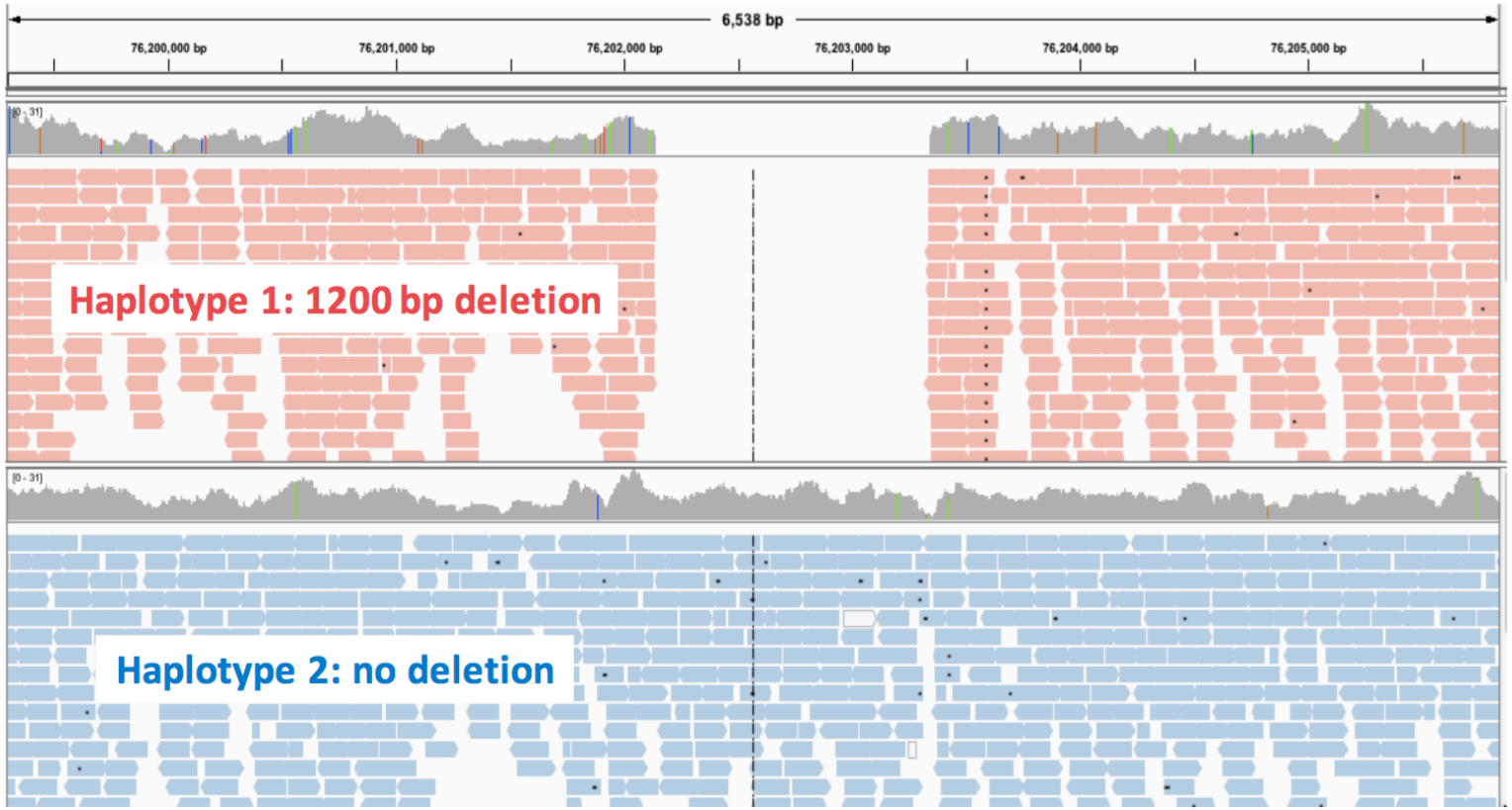
- 10x Genomics, Linked reads technology
- Illumina machines, Sequencing by Synthesis ~ 120Gb/lane, 2x150bp reads.

10x has its own assembler, Supernova

UCDAVIS Bioinformatics Core

10x Genomics phasing + high quality Illumina data





UCDAVIS Bioinformatics Core

10x Genomics, Supernova Genome Stats

Genome	Size (Gb)	DNA size(Kb)	N50 contig(Kb)	N50 scaffold(Mb)	N50 phase block (Mb)
NA12878	3.2	95.5	85.0	12.8	2.8
NA24385	3.2	111.3	90.0	10.4	3.9
HGP	3.2	138.8	104.9	19.4	4.6
Yoruban	3.2	126.9	100.5	16.1	11.4
Komodo dragon	1.8	85.4	95.3	10.2	0.4
Spotted owl	1.5	72.2	118.3	10.1	0.2
Hummingbird	1.0	86.2	87.6	12.5	10.1
Monk seal	2.6	92.3	93.8	14.8	0.6
Chili pepper	3.5	53.3	84.7	4.0	2.1
CowPea	**0.38**	**46.5**	**28.3**	**0.83**	**0.35**
Walnut	**0.89**	**55.0**	**48.0**	**0.60**	**0.25**
California Condor	**1.19**	**67.0**	**147.5**	**17.9**	**1.0**

Genome Assembly is converging on more standardized data models

- Trend is to consider sample, data generation and bioinformatics together.
  - ALLPATH-LG, started with specific requirement of sequencing libraries

Table 1. Provisional sequencing model for de novo assembly

Libraries, insert types*	Fragment size, bp	Read length, bases	Sequence coverage, ×	Required
Fragment	180†	≥100	45	Yes
Short jump	3,000	≥100 preferable	45	Yes
Long jump	6,000	≥100 preferable	5	No‡
Fosmid jump	40,000	≥26	1	No‡

  - Discovar

    250bp paired-end PCR-free Illumina reads. No other libraries are required.

The Kitchen Sink

- Available Technologies
  - Long Reads: Pacific Biosystems / Nanopore — Long Contigs
  - Optical Maps: BioNano — Scaffolding
  - Linked Reads: 10x Genomics — High base quality and phasing
  - Cross Linking: Hi-C / Dovetail Chicago — Scaffolding

- What the best combination, are all necessary? As algorithms improve, which become unnecessary
- Genome 10K project: Sequence 10,000 Invertebrates

Goat Genome

	CHIR_2.0 (BGI) - 2012	ARS1 - 2016
	14 Illumina PE libraries + Opgen	Pac Bio + Bionano + Hi-C
Coverage	175x	69x (@ 5.1Kb mean read length)
Assembly length	2.8 Gb	2.9Gb
Number of contigs	173,141	3,074
Contig N50	73.5 Kb	18.7 Mb
Number of scaffolds	103,494	31 (chromosomes)
Scaffold N50	9 Mb	87.3Mb

Adding in the optical maps from the Irys system reduced the total number of contigs to 1,780, with a contig N50 of 10.2 megabases. "The optical mapping increased the quality and confidence of the initial scaffolds," Phillippy said. The three technologies—PacBio, Bionano, and Hi-C—ended up being complementary to each other, he added. Finally, Illumina data is used to polish and make error corrections at the base level. **GenomeWeb** “Goat Genome Demonstrates Benefits of Combining Technologies for De Novo Assembly”, Mar 07, 2017

Focus of the Future

- To some extent we are limited by being able to generate enough high quality high molecular weight DNA.
- Continued improvement to sequencing chemistries for consistent and longer reads, quality improvement has become secondary.
- Incremental improvement of the computational algorithms, including improved alignment of error-prone reads (GFA2).
- Scaffolding algorithms, merge multiple data types/sources
- Polyploidy??
- Haplotyping – How to really use the data

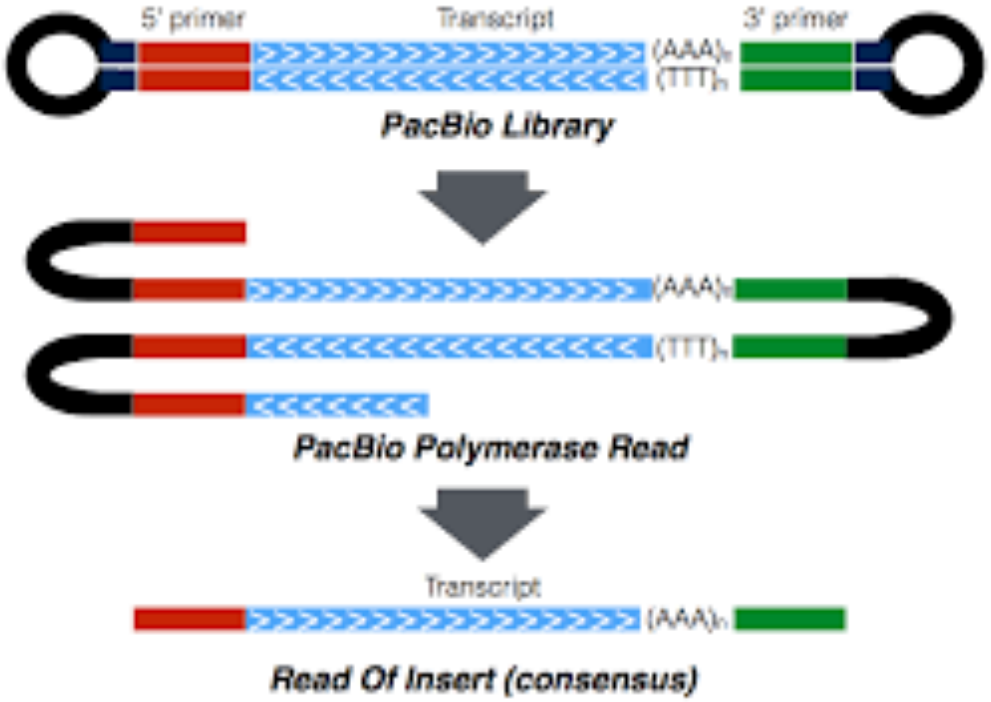
Graphical Format Assembly - GFA2

- Assembly is a pipeline
  - Overlap
  - Layout
  - Consensus
- With a common input (fastq) and common output (fasta), but no common intermediate file format, causes a duplication of effort.
- GFA2 - Common file format for assembly graph representation
  - Direct graph visualization, manipulation
  - Modular assembly tools (heterozygous/mis-assembled contigs)
  - Modular scaffolding tools
  - Graph aware annotation

Annotation – Pac bio Iso-seq

Produce full-length transcripts without assembly

The isoform sequencing (Iso-Seq) application generates full-length cDNA sequences — from the 5’ end of transcripts to the poly-A tail — After Circular consensus sequence (CCS) algorithm produces high quality isoforms.

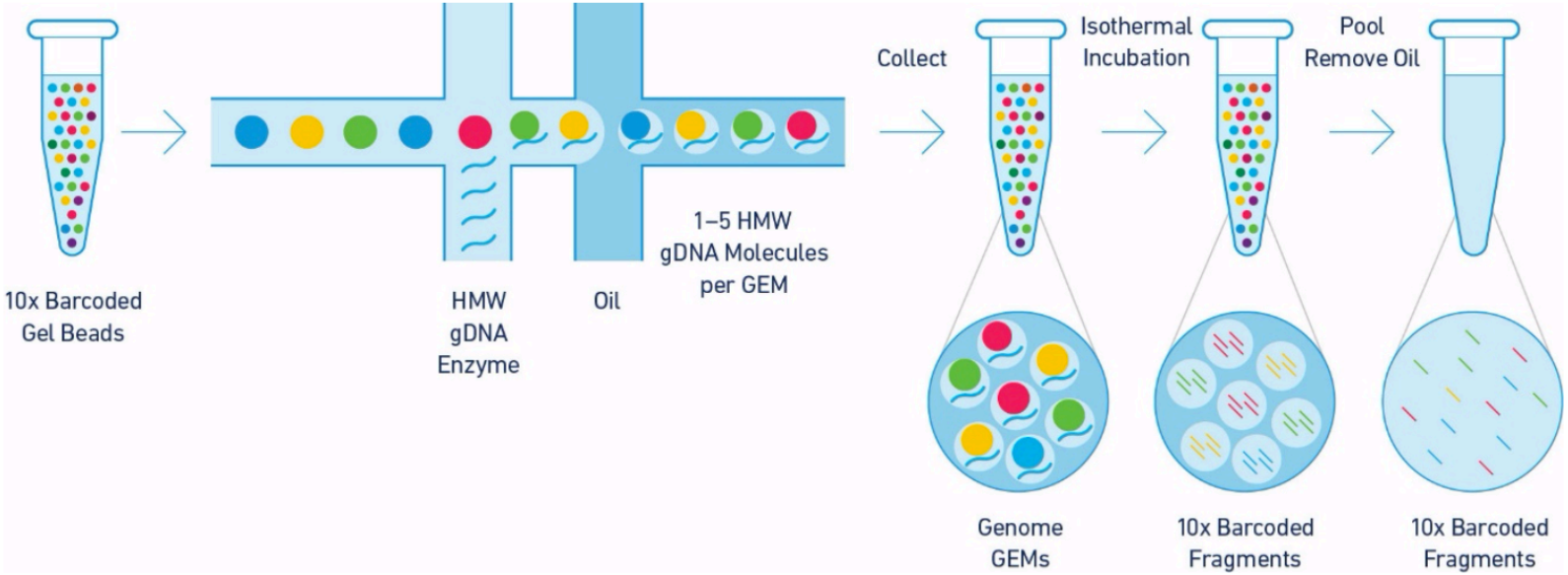


10x Genomes [Linked reads]

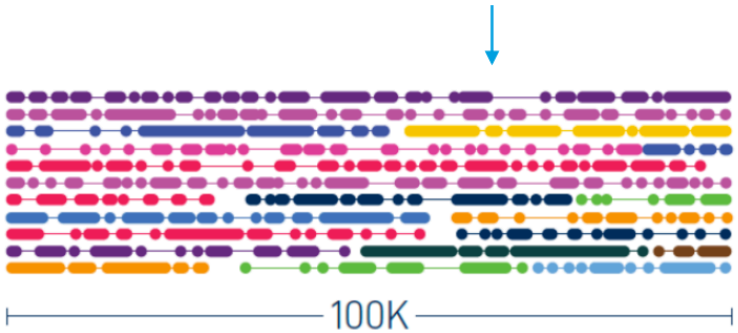
- **Genome** — Genome Resequencing
  - Call the full spectrum of variants (particularly long INDELS/CNV and structural variants) and unlock previously inaccessible regions from ***a single library at equivalent coverage as standard genome resequencing projects***
- **Exome** – Subselect reads using capture techniques (Agilent)
  - Enable phasing of genes and detection of structural and copy number variation
  - Agilent SureSelect baits improve gene phasing by closing gaps, and recovering hard-to-map loci in the genome (future kits to include previously failed regions)

UCDAVIS Bioinformatics Core

in a nut shell

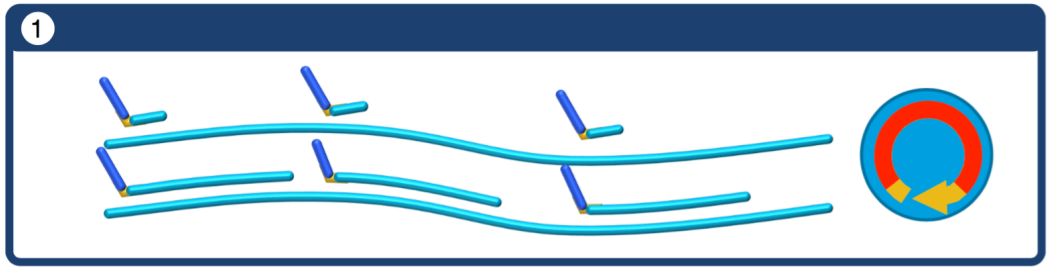


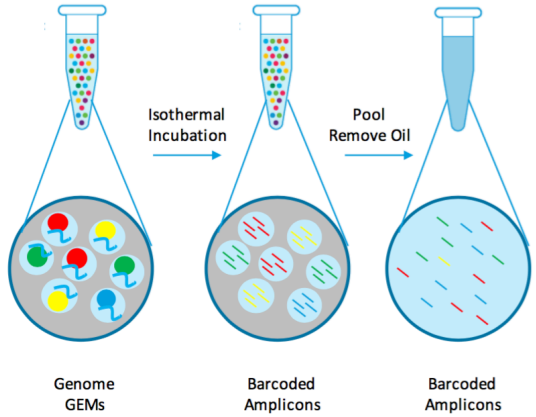
Linked-Reads



UCDAVIS Bioinformatics Core

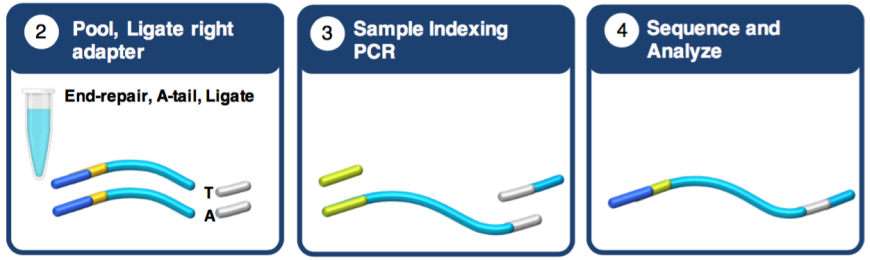
Laboratory Workflow



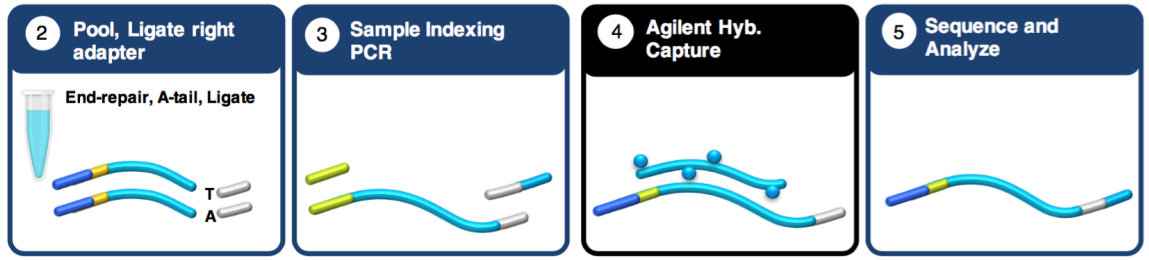


Shear?

Whole **Genome** Sequence

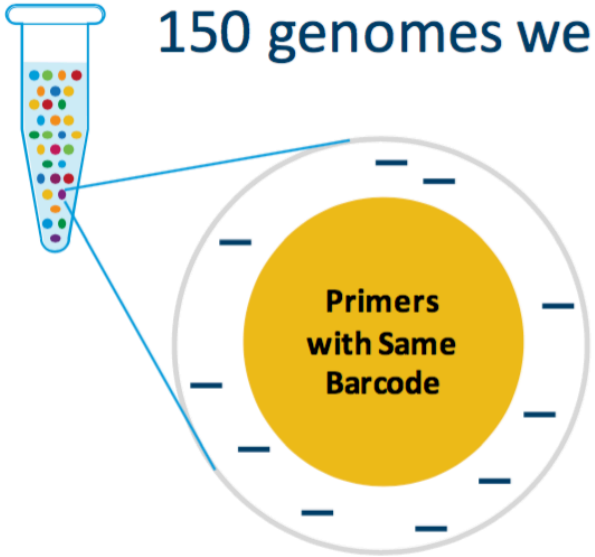


Whole **Exome** Sequence



The Math

1 ng Input DNA
= 300 genomes
copies of the
genome



150 genomes went into 1M partitions

Each GEM contains:

- One barcode (many copies)
- 1/6000 of the genome (500 Kb)
- At 50Kb length, 10 molecules

Chance that 2 molecules covering a locus are in same GEM:

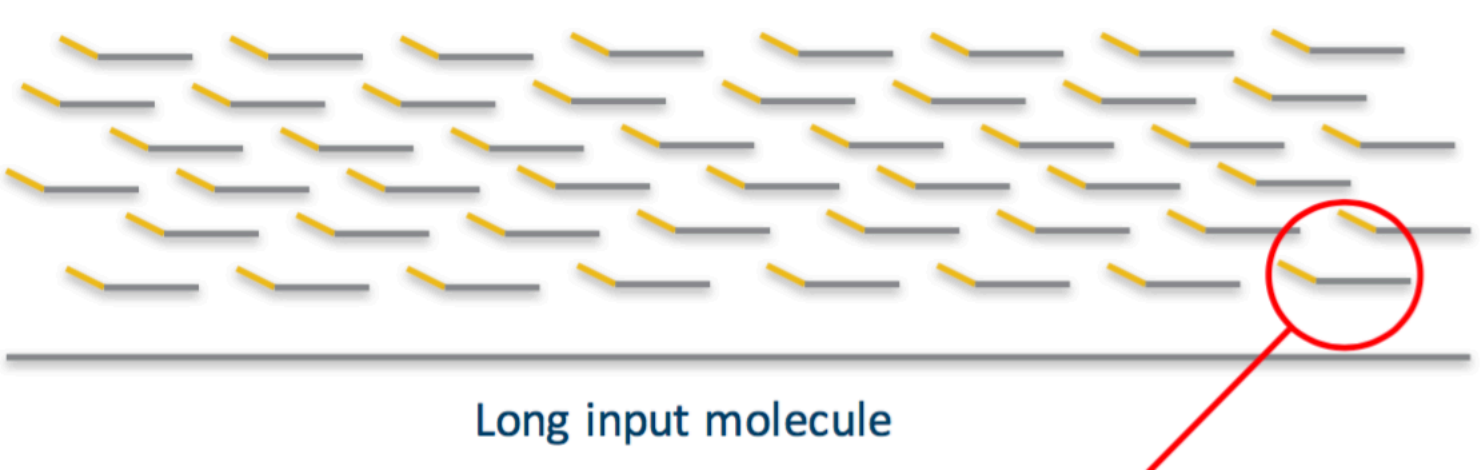
1 in 6000

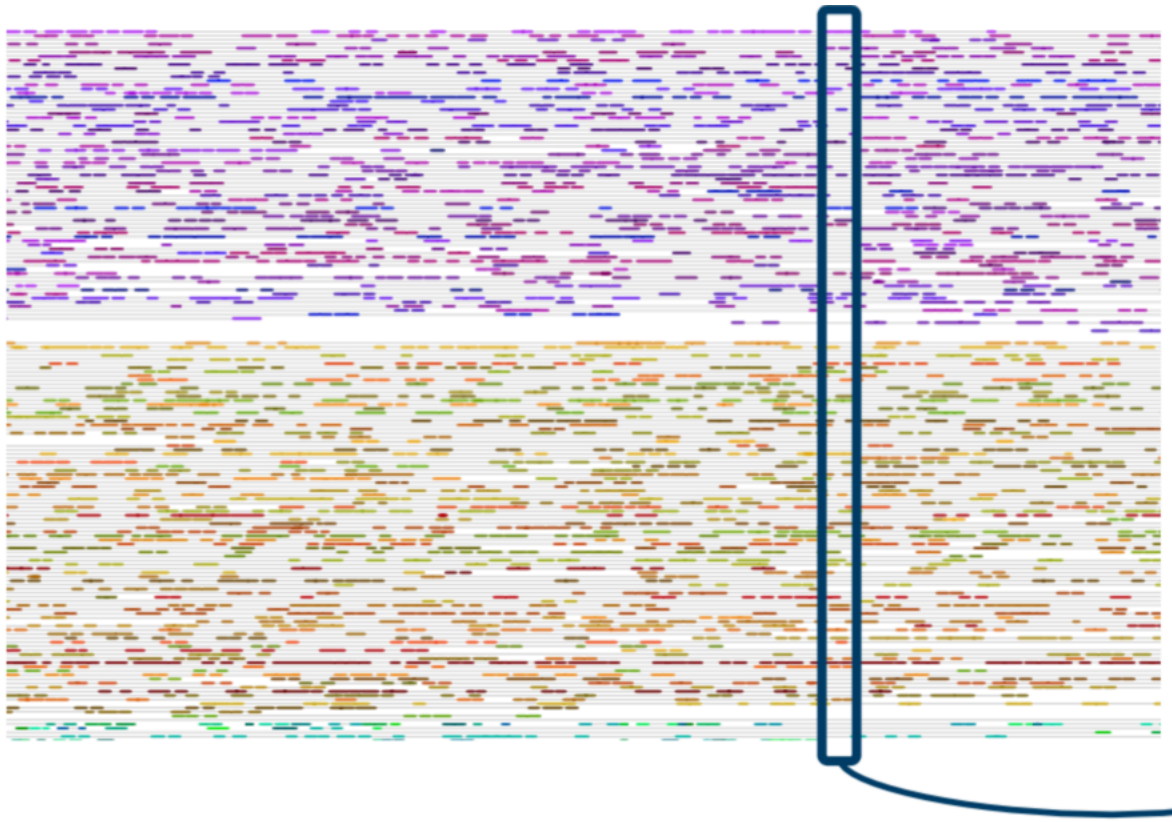
Percent unique barcodes at any genomic locus:

99.98%

Calculations imply that
about 50% of all possible
fragments end up in a bead

UCDAVIS Bioinformatics Core





150X avg *molecule* coverage

@ Recommended Loading

- Each locus will have 150 molecules
- Each locus will have 30x read depth

- ~35 fragments per molecule
  @50Kb molecules = 0.2x/molecule

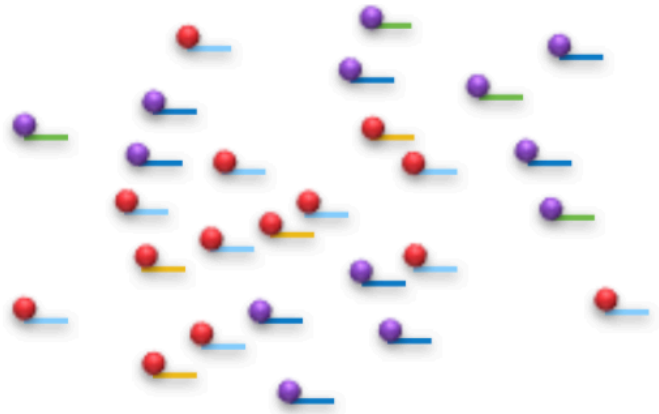
> 30X avg *read* coverage

Analysis – Biological Questions

- At recommended specs (for human genome)
  - Get ~30x coverage, adequate for standard variant analysis SNPs, small INDELS
  - Increased mapability to difficult regions [multi-mapped reads can be resolved by considering linked reads information], variants previously undetermined.
  - Detect large SV and CNV
  - Phased information

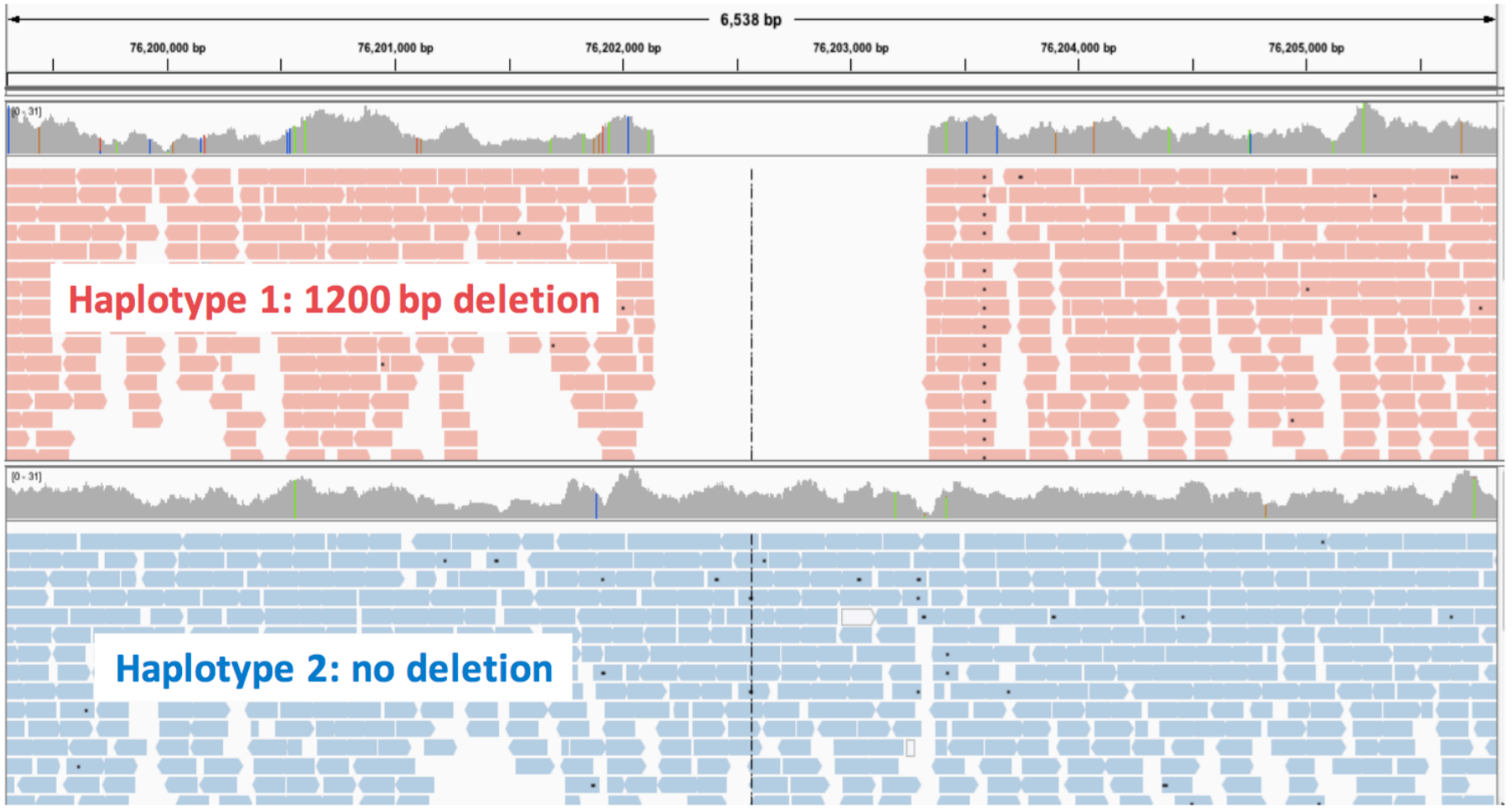
Detection of SV and CNV requires advanced computational techniques
Phasing has been used extensively in GWAS (usually imputed) to enhance analysis and inferences, slow to get to sequence based data
Potential applications for the technology are likely still yet to come

Increased Mapability



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

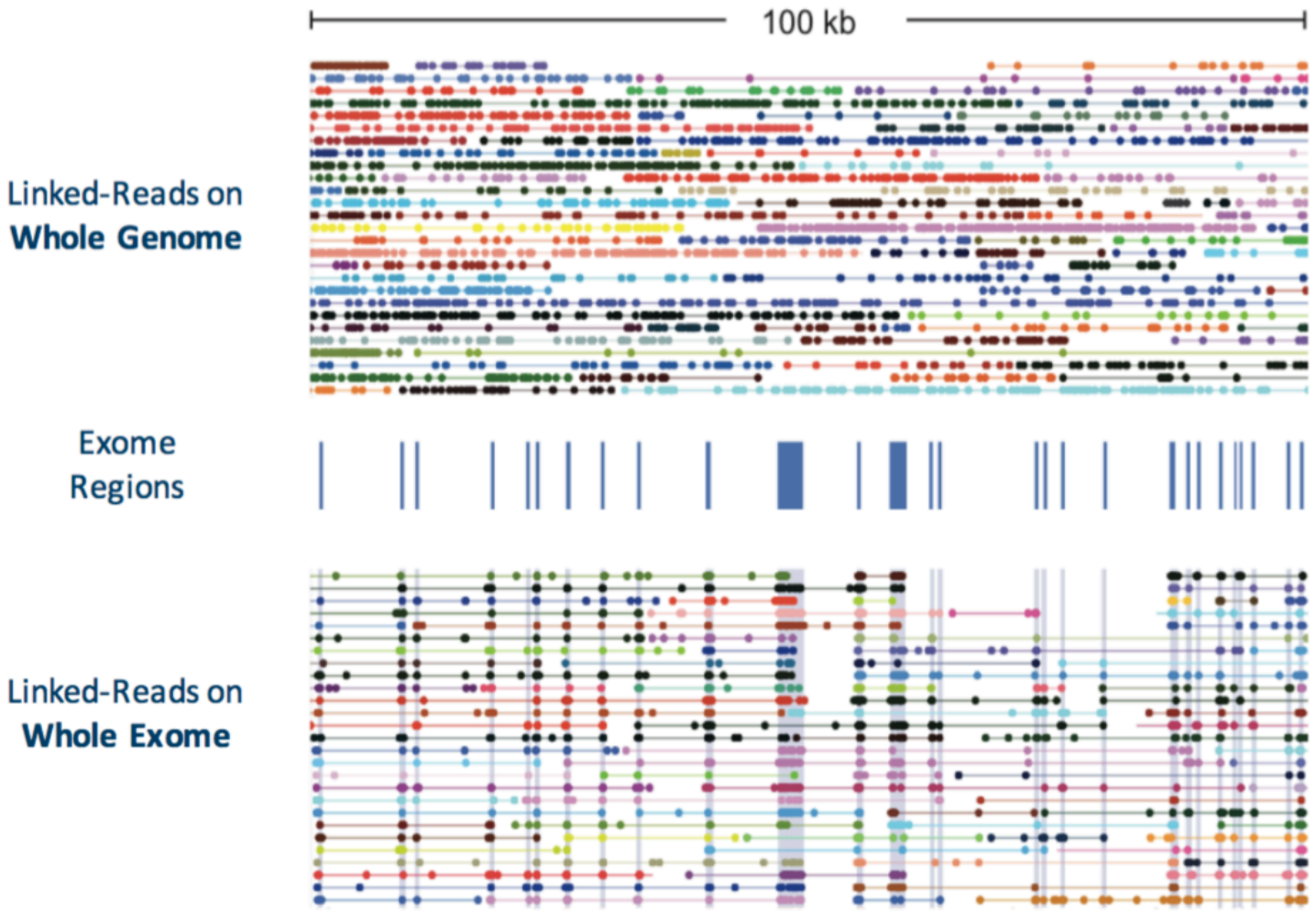
Linked Reads



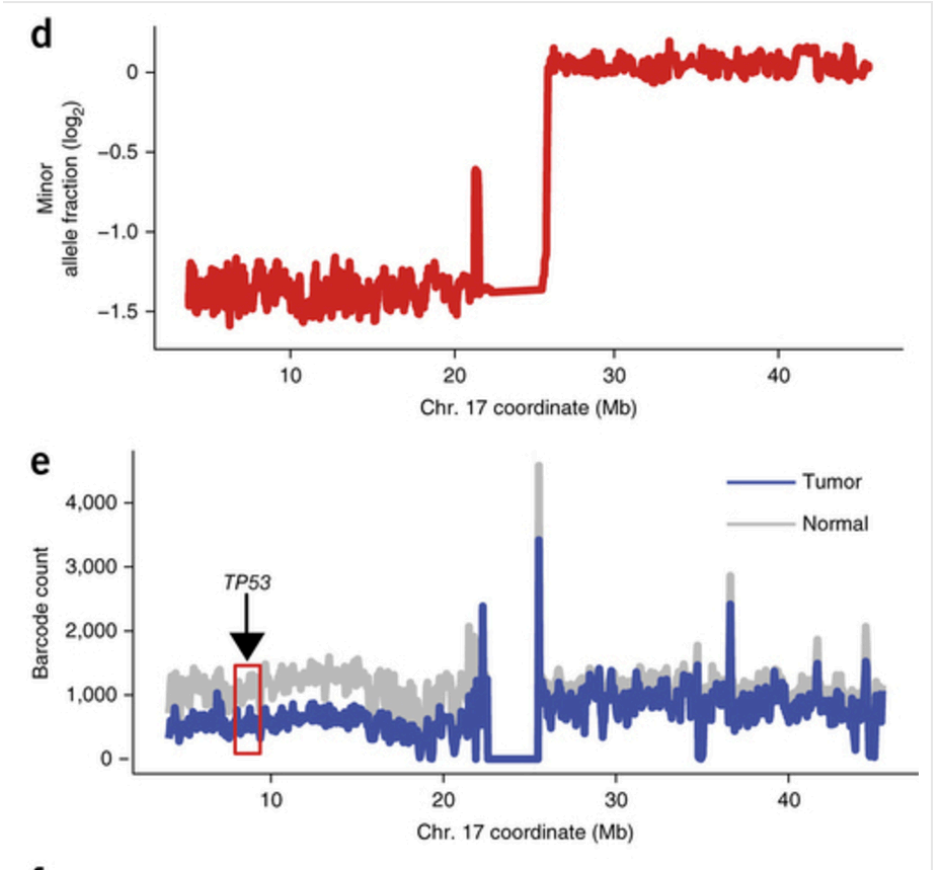
Capture – linked genes

Enrich reads of interest instead of random selection

Depending on size of capture, can pool more samples/lane



UCDAVIS Bioinformatics Core

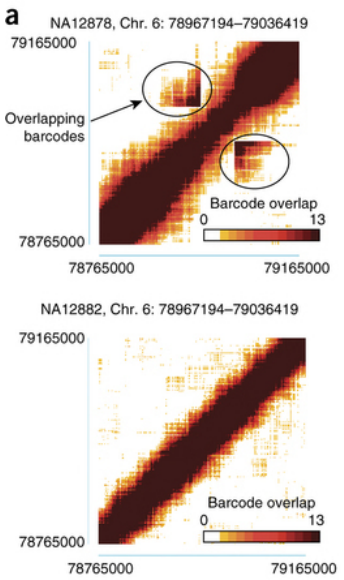


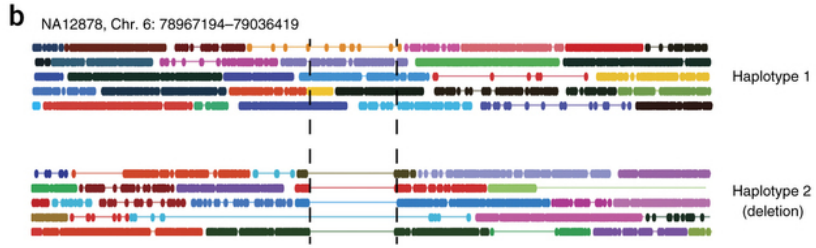
Thus, this TP53 C>T mutation is in trans with the deleted allele haplotype. As a result, the tumor contains only a single, inactivated copy of TP53. Taken together, this result shows the unambiguous biallelic inactivation of TP53

nature biotechnology

Haplotyping germline and cancer genomes with high-throughput linked-read sequencing

Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J Makarewicz, Yuan Li, Phillip Belgrader, Andrew D Price, Adam J Lowe, Patrick Marks, Gerard M Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong + *et al.*





c

Chr.	Location	Phase block	BCs intersecting Hap1	BCs intersecting Hap2	*P*	Deleted haplotype	Deletion inherited by NA12882	Phasing consistent with inheritance
1	189704509–189783359	50188502070	57	4	4.90×10^{-13}	1	No	Yes
3	162512134–162626335	190161712492	2	57	6.10×10^{-15}	2	Yes	Yes
5	104432113–104503673	27098341819	5	28	6.60×10^{-5}	2	Yes	Yes
6	78967194–79036419*	31076466627	45	7	7.00×10^{-8}	1	No	Yes
6	78967194–79036419*	31078967164	50	2	6.10×10^{-13}	1	No	Yes
8	39232074–39387229*	39030957010	45	0	5.70×10^{-14}	1	Yes	Yes
8	39232074–39387229*	39039370800	46	0	2.80×10^{-14}	1	Yes	Yes
Low-scoring SV candidates								
5	99400881–99715015	27098341819	97	87	0.51	FP	No	N/A
14	37631609–37771228	62033781205	2	8	0.11	Common	Common	N/A
14	106932640–107174931	640106692870	10	13	0.68	FP	No	N/A

*Two breakpoints of the deletion are on different phase blocks.

UCDAVIS Bioinformatics Core