UCDAVIS Bioinformatics Core

Towards Gapless, Chromosome Scale, Haplotype Assemblies

Jie Li, PhD

UC Davis Bioinformatics Core

August 30, 2018

Slides courtesy of Matt Settles

UCDAVIS Bioinformatics Core

The **mission** of the Bioinformatics Core facility is to facilitate outstanding omics- scale research through these activities:

Data Analysis

The Bioinformatics Core promotes experimental design, advanced computation and informatics analysis of omics scale datasets that drives research forward.

Research Computing

Maintain and make available high-performance computing hardware and software necessary for todays data-intensive bioinformatic analyses.

Training

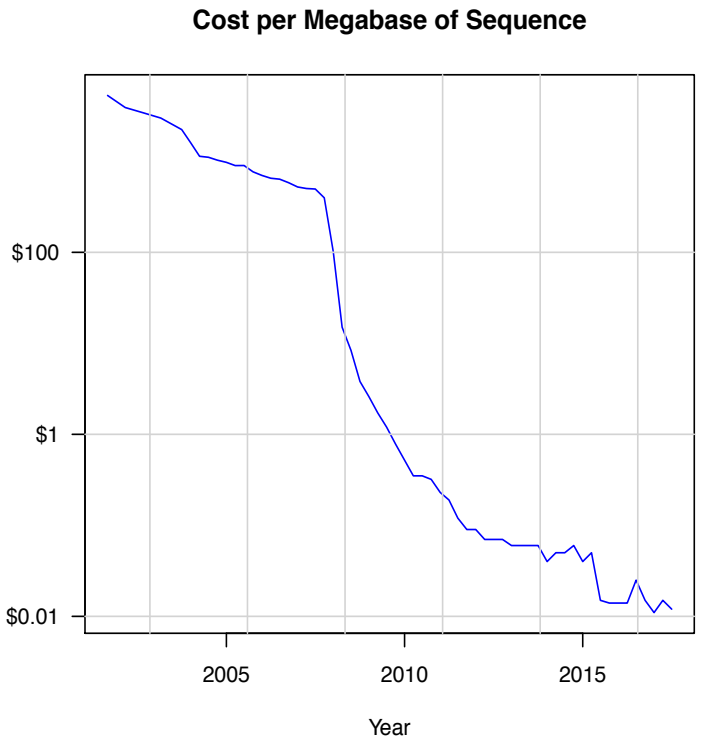
The Core helps to educate the next generation of bioinformaticians through highly acclaimed training workshops, seminars and through direct participation in research activities.

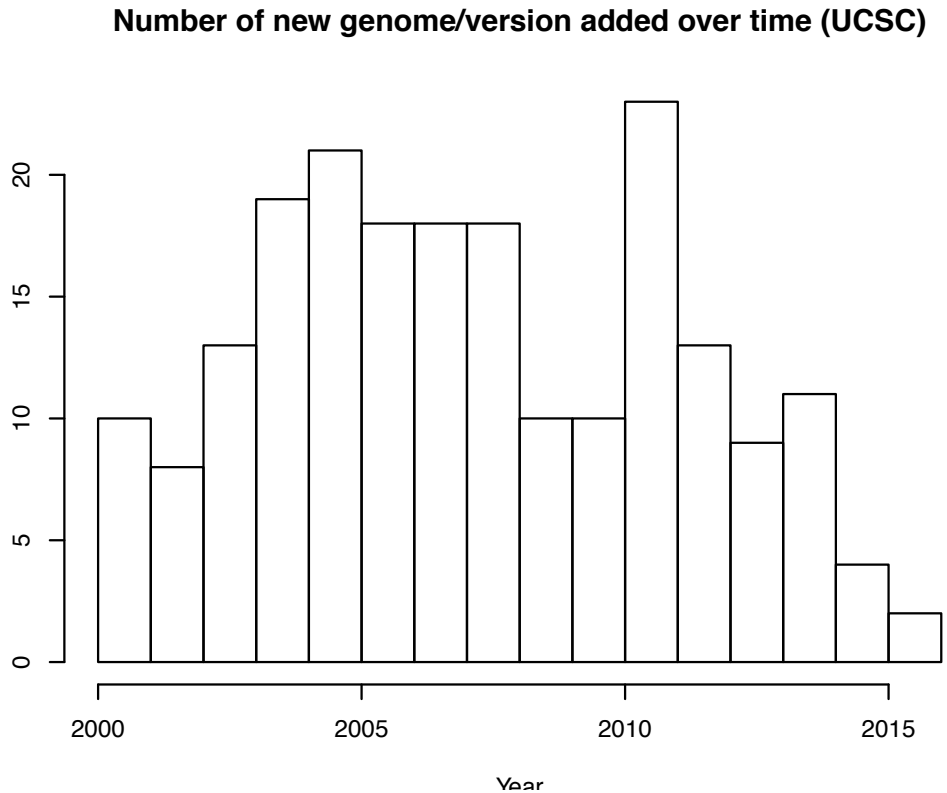
Human Genome

- 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.

UCDAVIS Bioinformatics Core

The 3rd phase of Genome Assemblies





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 genome for about $70,000. “That, to me, is a big deal,” he says.
- Also a big deal, says Eichler, is the quality of their assembly. 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 2016

UCDAVIS Bioinformatics Core

Gorilla Genome

Assembly	2012 Illumina Assembly	2016 Pacific Biosciences 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
2016 Assembly: PACBIO SMRT sequence + BAC PE data, INDEL corrected with Illumina sequence

UCDAVIS Bioinformatics Core

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 de novo

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

UCDAVIS Bioinformatics Core

Advances in high-noise, long-read assembly algorithms

- Summer of 2015
  - Pacific Biosciences 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.

UCDAVIS Bioinformatics Core

Gapless: The ‘Next, Next’ Generation Sequencers (single molecule, long reads)

Oxford Nanopore

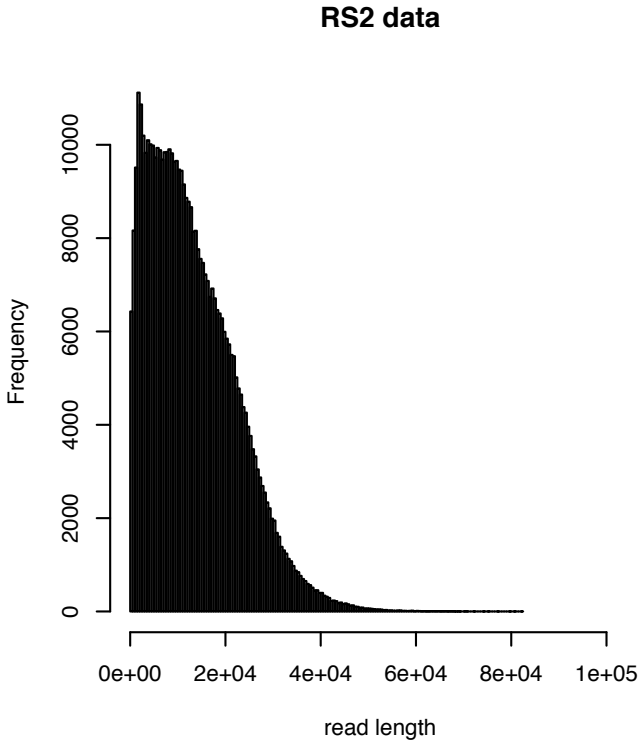
Pacific Biosciences

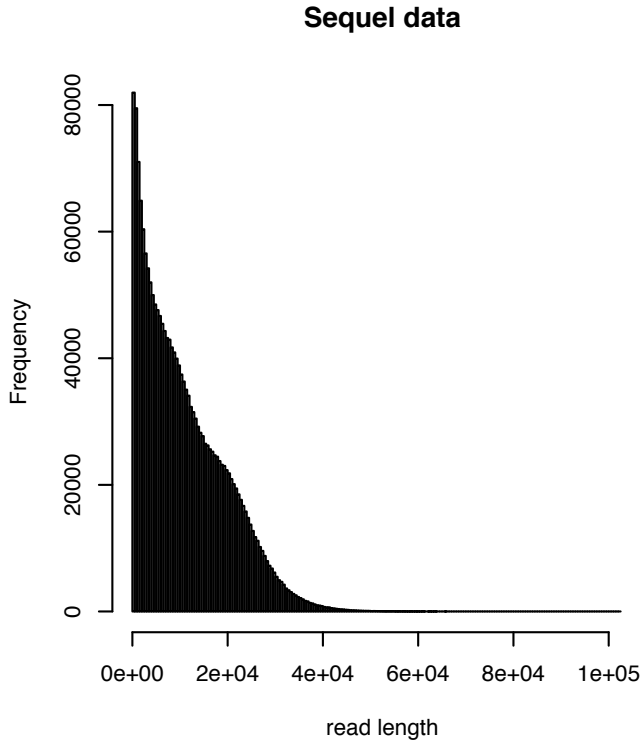
UCDAVIS Bioinformatics Core

Pac Bio Advances (RSII vs Sequel)

California Condor data (~1.2Gbp genome) based on 65 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 (65 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

Towards Gapless assemblies

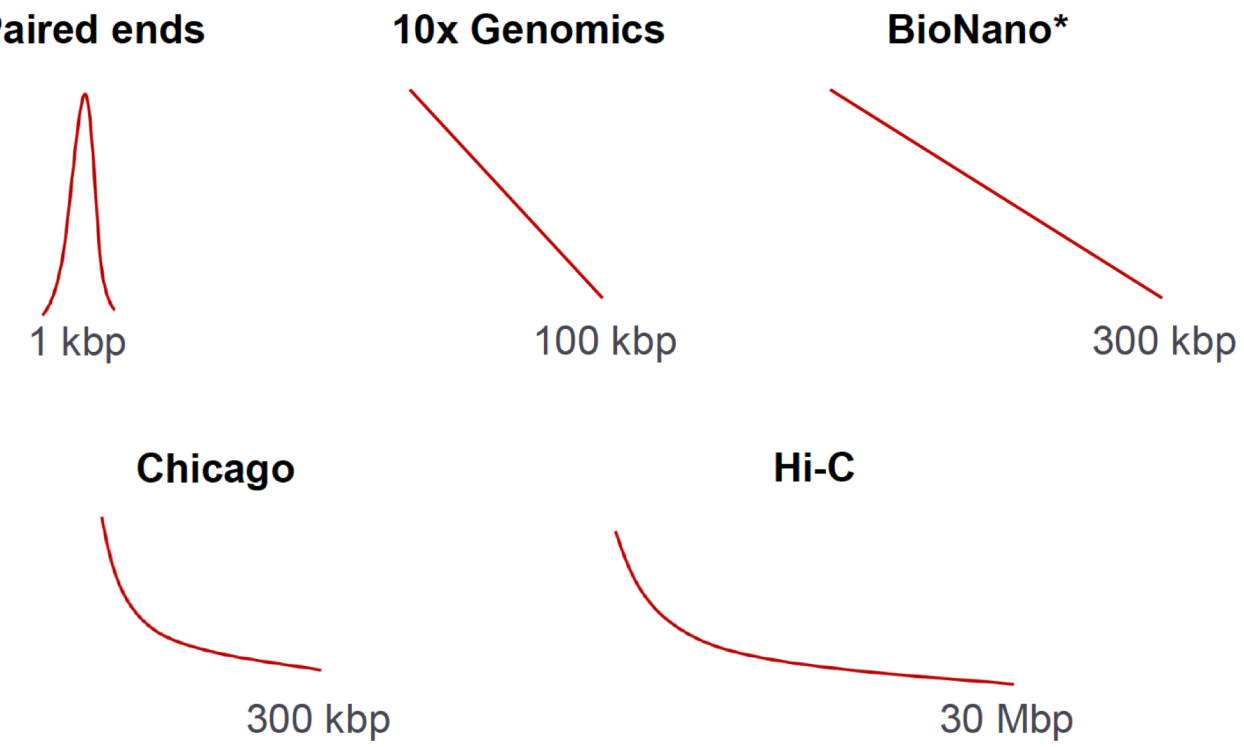
❖ **Promise**

- Continued progress on DNA input and resulting PacBio/Oxford Nanopore Read Lengths and read depth will result in longer N50/N90 and fewer 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.

Chromosome Scale: Scaffolding Options



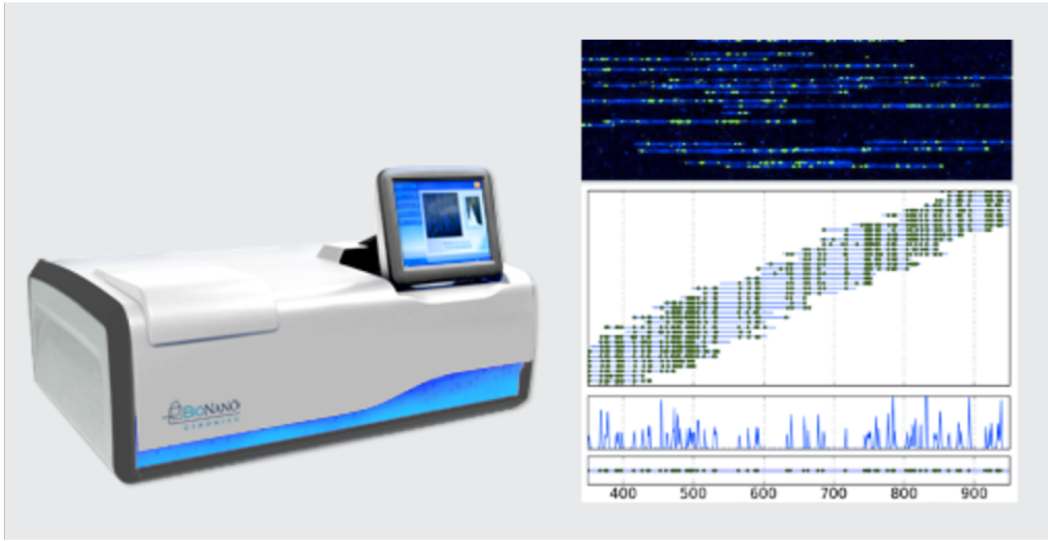
'Borrowed' from Sergy Koren talk from PacBio Informatics Developer Meeting in Jan 2017

UCDAVIS Bioinformatics Core

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.

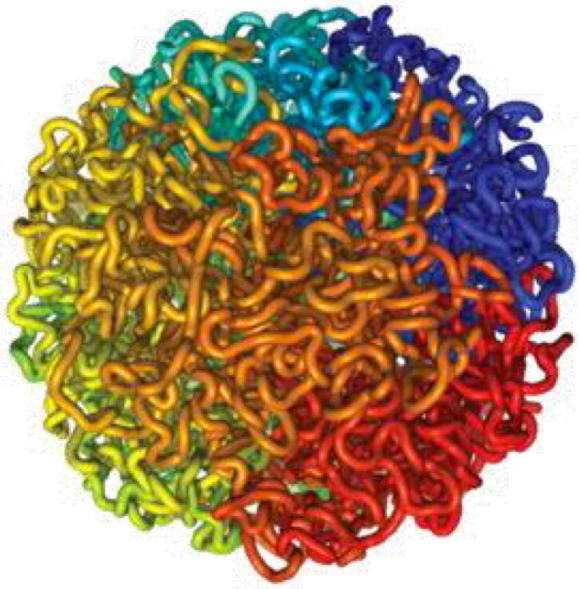
Not sequencing based



UCDAVIS Bioinformatics Core

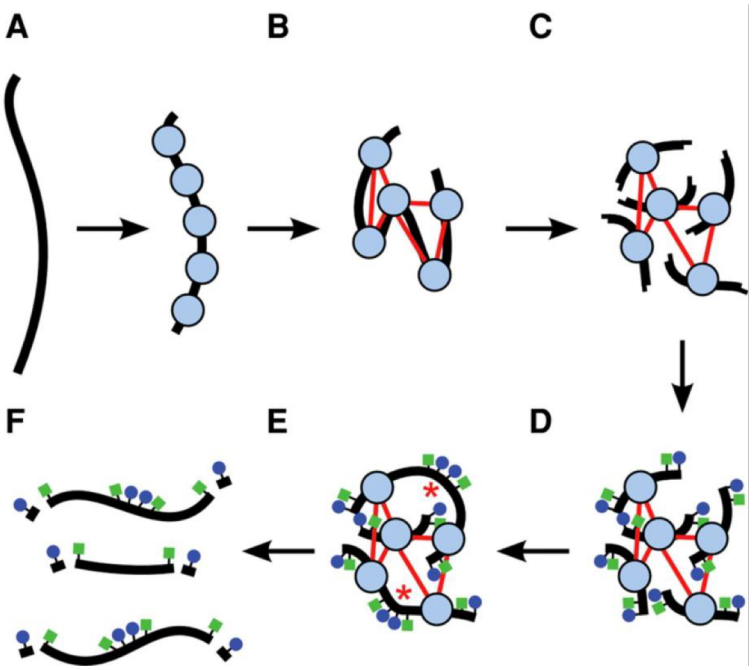
Dovetail and Hi-C (Cross Linking) on Illumina

Hi-C
Proximity Guided Scaffolding



Dovetail and Phase Genomics

Dovetail Chicago Libraries



UCDAVIS Bioinformatics Core

10x genomics on Illumina

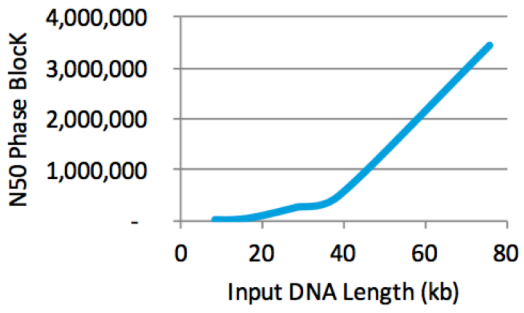
- 10x Genomics, Linked reads technology
- Illumina machines, Sequencing by Synthesis 2x150bp reads.

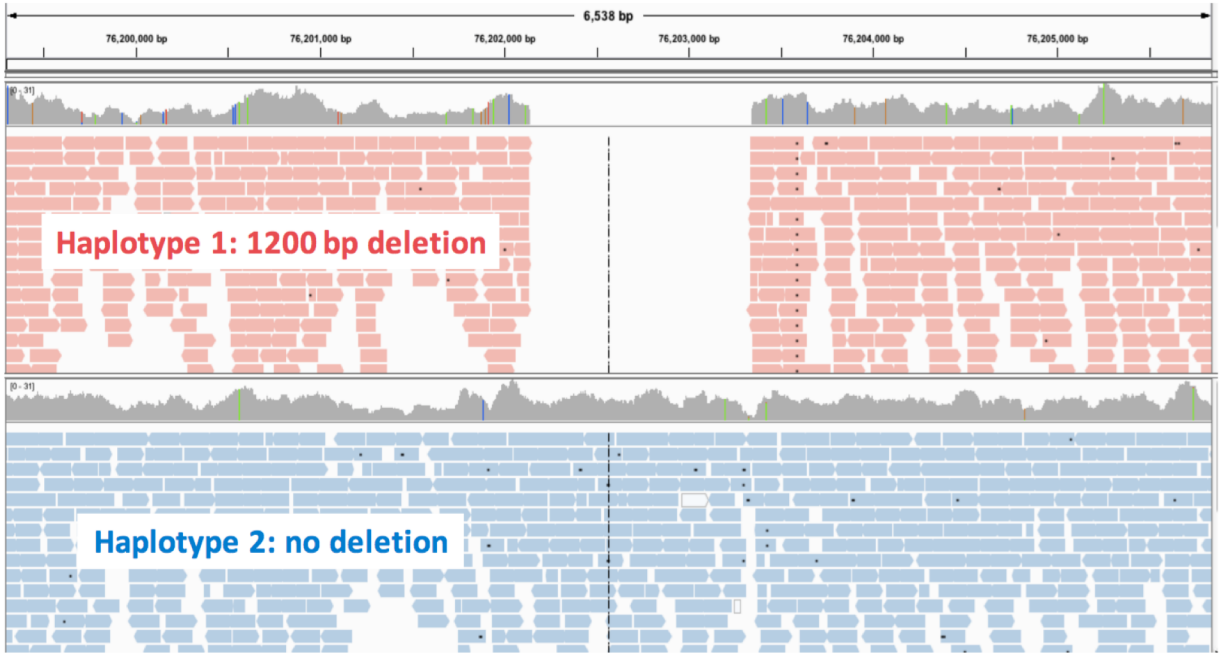
ARCS - https://github.com/bcgsc/arcs/tree/binomialx2

10x has its own assembler, Supernova

UCDAVIS Bioinformatics Core

Phasing: 10x Genomics + high quality Illumina data, draft genomes??





UCDAVIS Bioinformatics Core

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 and Vertebrate Genome project: Sequence ~66,000 vertebrates

UCDAVIS Bioinformatics Core

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

UCDAVIS Bioinformatics Core

Order: The Kitchen Sink

- Available Technologies

▪ Long Reads: Pacific Biosystems / Nanopore	Long Contigs	**2 or 3**
▪ Optical Maps: BioNano	Scaffolding	**4**
▪ Linked Reads: 10x Genomics	High base quality and phasing	**1**
▪ Cross Linking: Hi-C ~~/ Dovetail Chicago~~	Scaffolding	**2**

- Make sure you have enough sample at the start of the project to add techniques over time
- Algorithms to improve combining data will improve over time

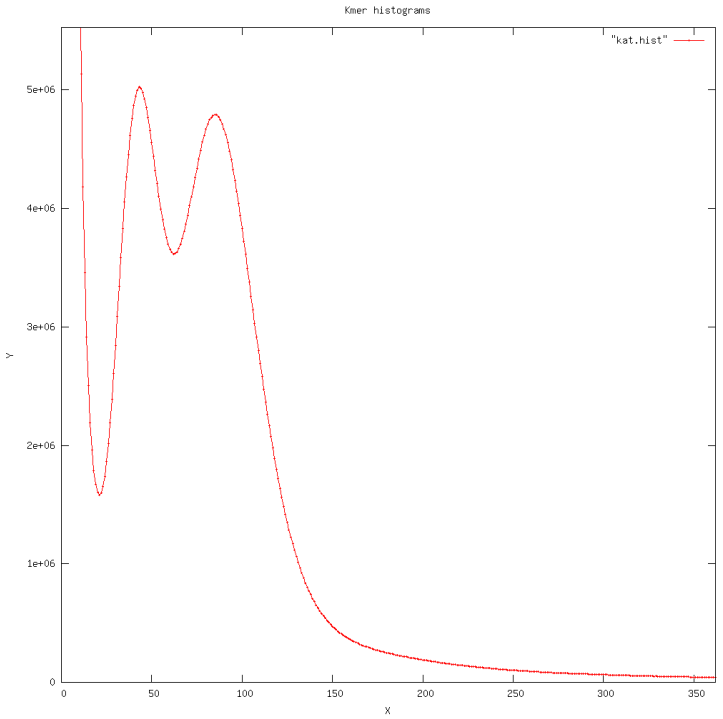
UCDAVIS Bioinformatics Core

10x Genomics, Supernova Genome Stats

Genome	Size (Gb)	DNA size(Kb)	N50 contig (Kb)	N50 scaffold (Mb)	N50 phase block (Mb)
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
Menidia+	0.40	34.4	60.0	10.0	6.5
Holbrookia Lizard#	1.70	37.6	47.2	1.34	0.83
Sceloporus Lizard	1.56	50.4	59.3	1.38	1.10
Sturgeon	0.40	76.0	15.4	0.16	0.12
Black Tailed Deer#+	2.47	40.8	293.3	32.4	6.61
Green Plant1	0.37	64.7	16.6	0.90	0.83
Green Plant2	0.32	40.0	15.0	0.12	0.13
Euk1#+	2.31	33.8	260.7	22.8	0.74

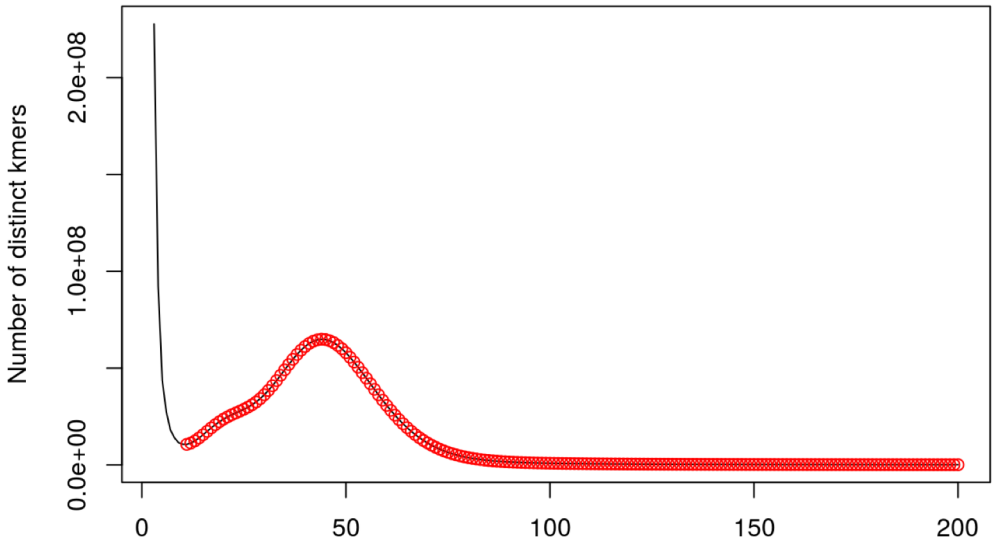
UCDAVIS Bioinformatics Core

Recommend to start with 10x genomics



- Kmer profiles and estimate genome size
- High quality Illumina data for polishing long reads
- Linked read data for scaffolding and haplotyping

- Relatively Cheap
- Best case scenario, adequate genome and can stop



UCDAVIS Bioinformatics Core

California Condor – PacBio vs 10x

$70K of PacBio - $4,000/MB of N50

NXX <chr>	LXX <int>	Length <chr>
N10	2	69,465,997 bp
N20	4	44,079,833 bp
N30	8	32,030,892 bp
N40	12	24,344,512 bp
N50	18	17,286,884 bp
N60	26	12,594,230 bp
N70	38	8,238,335 bp
N80	58	4,692,950 bp
N90	113	1,106,390 bp

$4K of 10X genomics - $220/MB of N50

NXX <chr>	LXX <int>	Length <chr>
N10	2	66,331,765 bp
N20	5	40,547,711 bp
N30	9	27,738,616 bp
N40	14	23,855,415 bp
N50	20	18,014,548 bp
N60	29	13,383,059 bp
N70	39	10,933,856 bp
N80	55	5,730,001 bp
N90	86	2,390,190 bp

The assembly contained 2.82% (35,325,300bp) uncharacterized ‘N’ basepair.

UCDAVIS Bioinformatics Core

Black Tailed Deer

10x Genomics Linked Reads Sequencing, Assembled with SuperNova (v2.0.0)

The 10X Supernova assembly resulted in 35,253 contigs for a total final genome size of 2,824,399,154bp. The assembly contained 1.06% (29,905,230bp) uncharacterized 'N' basepair. The GC content of the assembly was 41.57%.

N50, L50 contig values

The N50 length is defined as the shortest sequence length at 50% of the genome. It can be thought of as the point of half of the mass of the distribution; the number of bases from the N50 contig and all contigs longer than the N50 will be close to the number of bases from all contigs shorter than the N50. The summary of the assembly NXX defined similarly is as below.

NXX	LXX	Length
N10	4	60,839,816 bp
N20	10	41,600,771 bp
N30	17	37,902,543 bp
N40	26	27,877,142 bp
N50	37	23,084,682 bp
N60	53	15,190,433 bp
N70	78	8,993,699 bp
N80	117	6,031,357 bp
N90	187	2,541,180 bp

UCDAVIS Bioinformatics Core

Black Tailed Deer

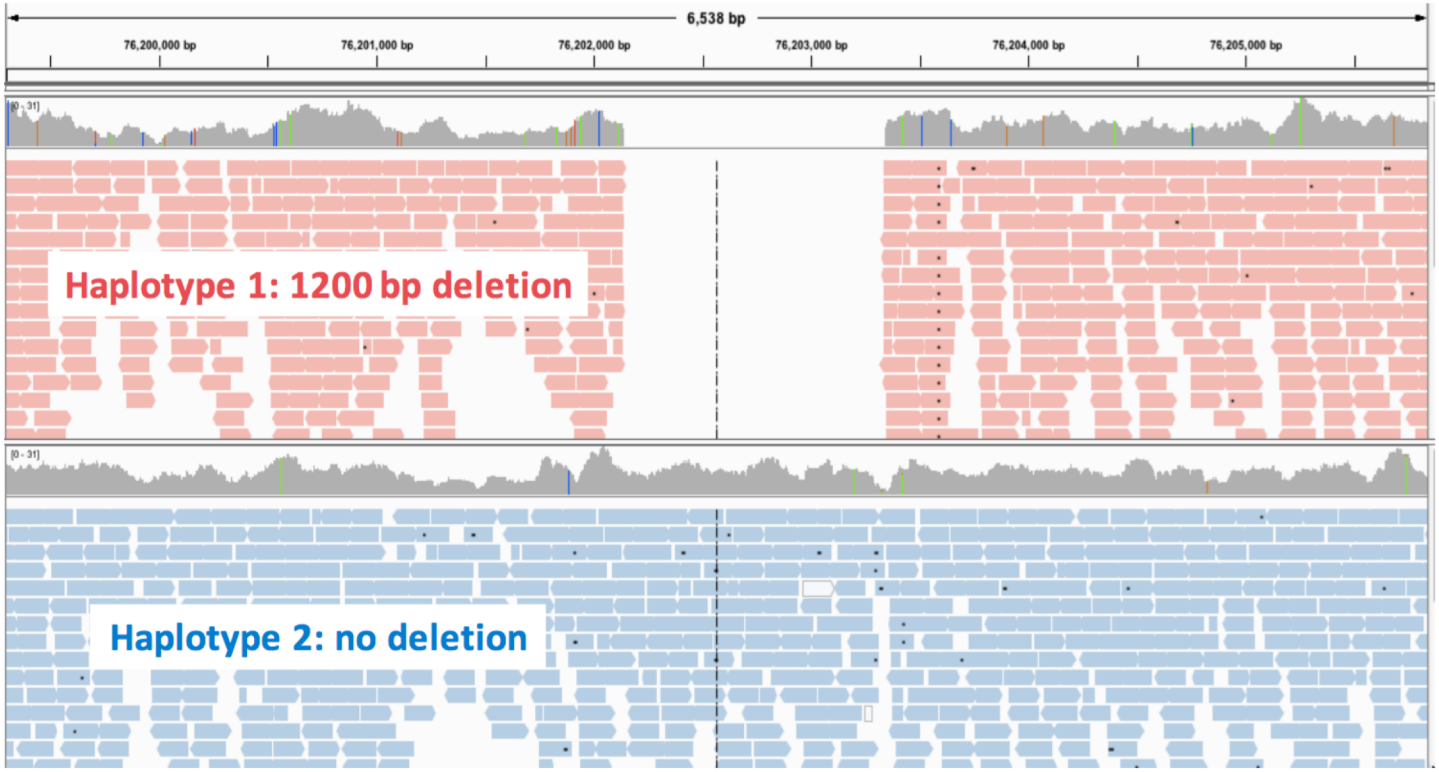
NXX	LXX.X	Length.X	LXX.nova	Length.nova
N10	4	60,839,816 bp	5	51,704,845 bp
N20	10	41,600,771 bp	11	40,871,444 bp
N30	17	37,902,543 bp	19	33,282,318 bp
N40	26	27,877,142 bp	29	26,153,834 bp
N50	37	23,084,682 bp	41	19,671,835 bp
N60	53	15,190,433 bp	57	15,290,162 bp
N70	78	8,993,699 bp	78	11,622,033 bp
N80	117	6,031,357 bp	110	7,314,267 bp
N90	187	2,541,180 bp	166	2,916,021 bp

Illumina NovaSeq system is comparable to Illumina X system:

- Higher percentage of duplicates (~5%)

UCDAVIS Bioinformatics Core

Linked Reads allows better phasing



UCDAVIS Bioinformatics Core

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, algorithms merge multiple data types/sources (GFA2).
- Polyploidy is now doable
- Haplotyped genomes – How to really use the data

UCDAVIS Bioinformatics Core

Informatics Workshops

Bioinformatics: Genome Assembly and Analysis Workshop [Pac Bio and 10x Genomics]

Dec. 12, 2016, 9:30 a.m. - Dec. 14, 2016, 4:30 p.m.

Organizer - Bioinformatics Core

Contact - UC Davis Bioinformatics Core, training.bioinformatics@ucdavis.edu

Dec, 2018

Bioinformatics: Single Cell RNA-Seq Workshop @ UCSF

March 19, 2018, 9:30 a.m. - March 21, 2018, 4:30 p.m.

Organizer - Bioinformatics Core

Contact - UC Davis Bioinformatics Core, training.bioinformatics@ucdavis.edu

Location - UCSF, Genentech Hall, Mission Bay (Room TBA)

https://registration.genomecenter.ucdavis.edu/

UCDAVIS Bioinformatics Core

Questions?