UCDAVIS Bioinformatics Core

Towards Error-free, Gapless, Chromosome Scale, Haplotype Assemblies

Matt Settles, PhD

UC Davis Bioinformatics Core

December 16, 2020

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.

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 2016

The Earth Biogenome Project

Ambitious project to sequence all Euk life on Earth (~1.5M species) in 20 years time, for about similar total cost as the Human Genome Project ($4.7B).
https://www.pnas.org/content/115/17/4325

Umbrella consortium of many existing genome projects including:

- **5,000 Insect Genomes (i5K)**
- **10,000 Bird Genomes (B10K)**
- **10,000 Plant Genomes (10KP)**
- **California Conservation Genomics Project (Cal CGP)**
- **Darwin Tree of Life (60K organisms on the British Isles)**
- **Fish 10,000 Genomes (Fish 10K)**
- **Genome 10K/VGP (1 genome from each Vertebate genus, ~70K)**
- **Many More**

https://www.earthbiogenome.org/

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

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.

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.

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

Oxford Nanopore

Pacific Biosciences

Towards Gapless assemblies

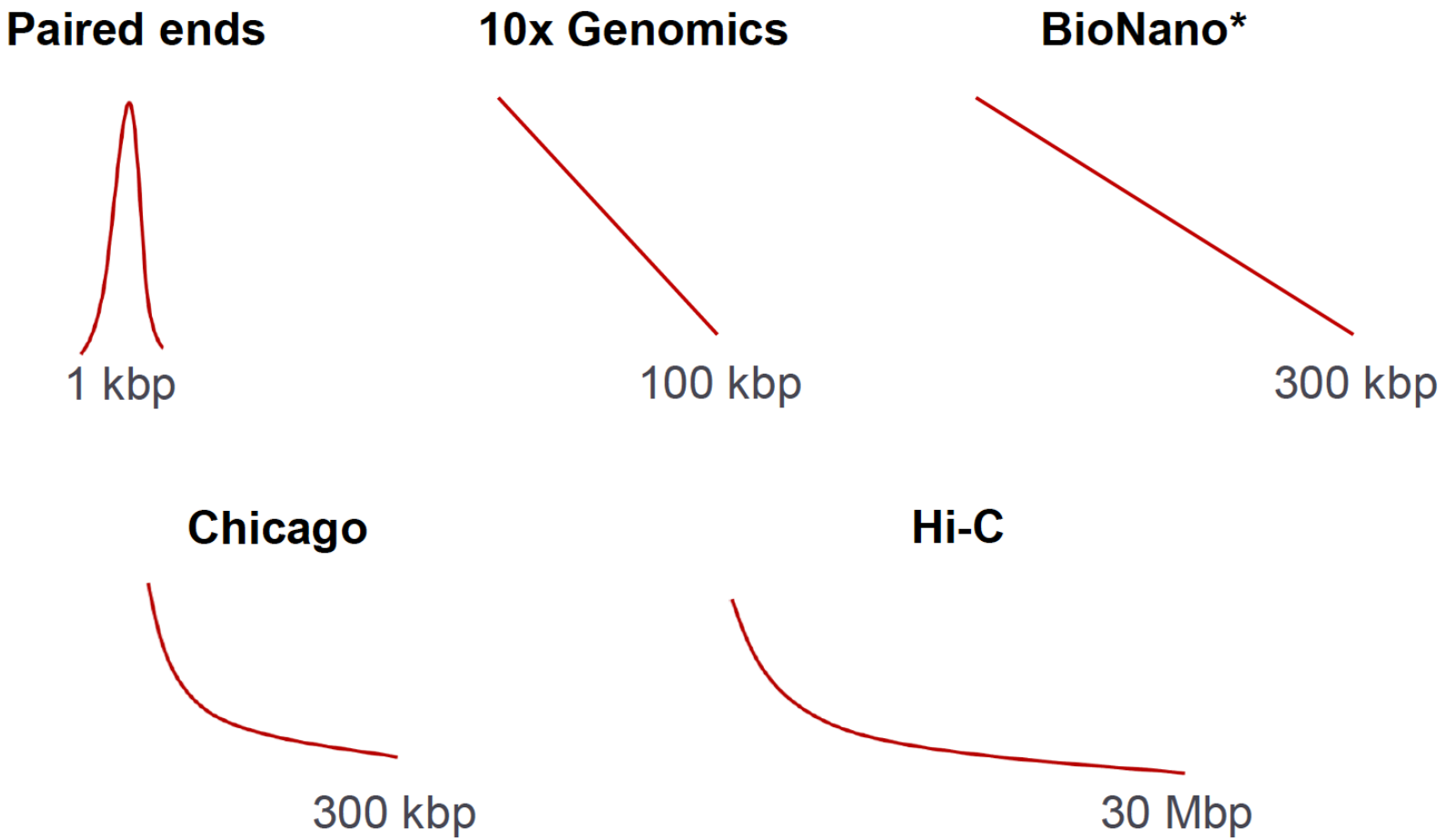
❖ **Promise**

- Continued progress on DNA input (HMW) and resulting PacBio/Oxford Nanopore read lengths, read depth, and quality will result in longer N50/N90 fewer resulting contigs.
- Algorithms are starting to mature, but still have room for improvement.

❖ **Issues**

- Some mis-assemblies are still present, Chimeric reads (PacBio) and sequence bias (ONT) are an issue
- Small INDELs, especially homopolymers 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

Linked Reads Technology - 10x genomics

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

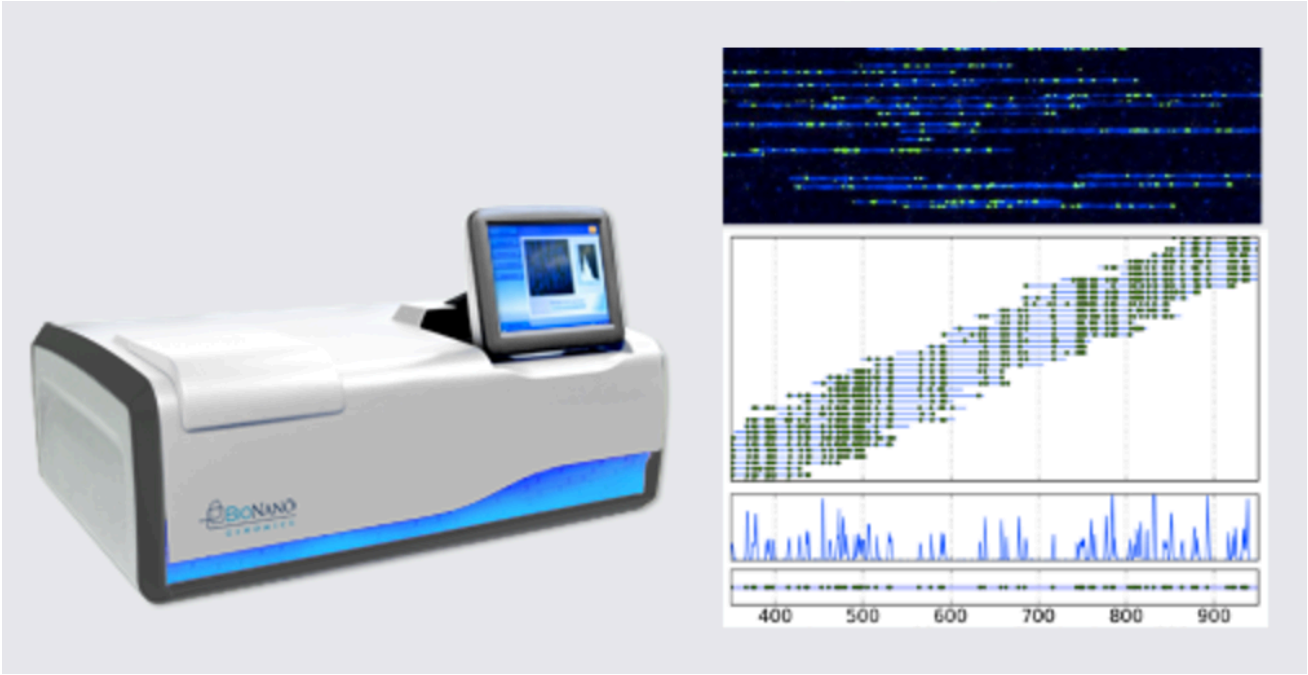
ARCS - https://github.com/bcgsc/arcs/tree/binomialx2

10x has its own assembler, Supernova

Bionano Genomics

- 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.

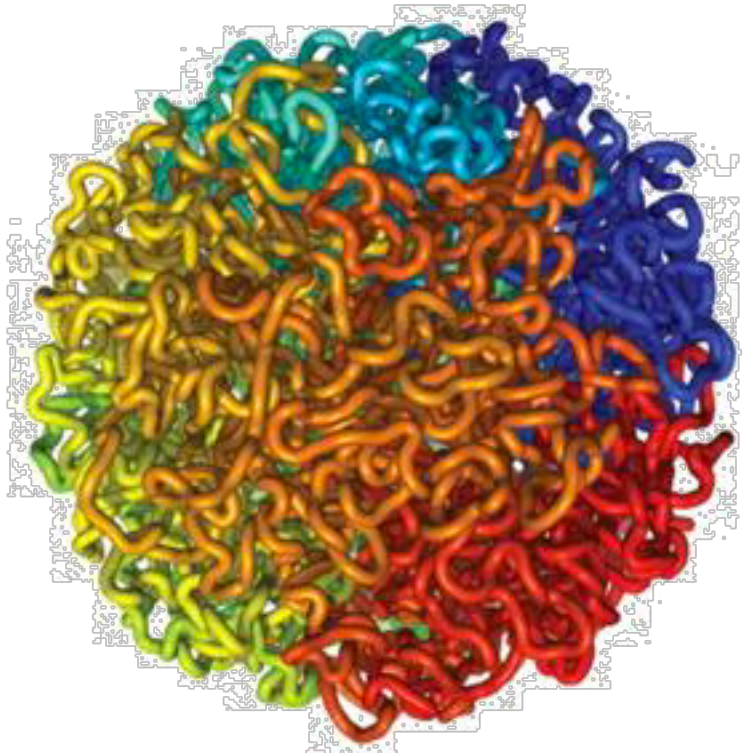
Not sequencing based



UCDAVIS Bioinformatics Core

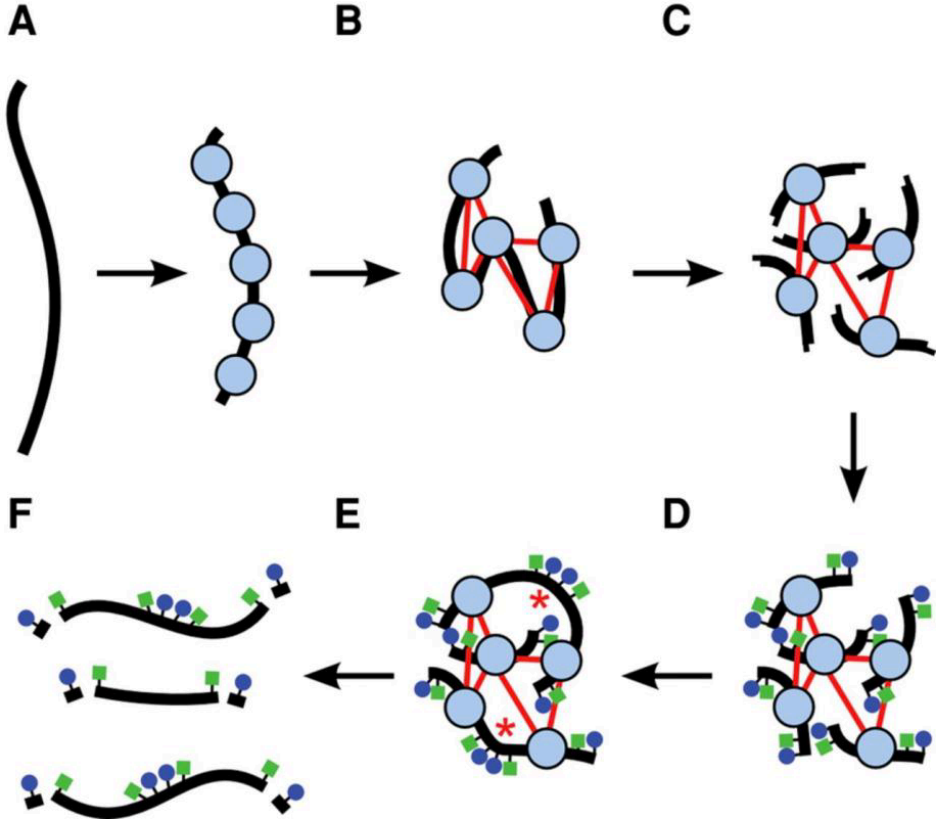
Dovetail Chicago and Hi-C (Cross Linking) on Illumina

Hi-C
Proximity Guided Scaffolding



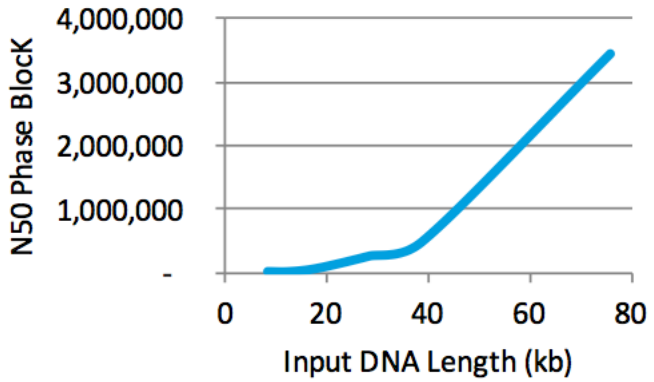
Dovetail and Phase Genomics

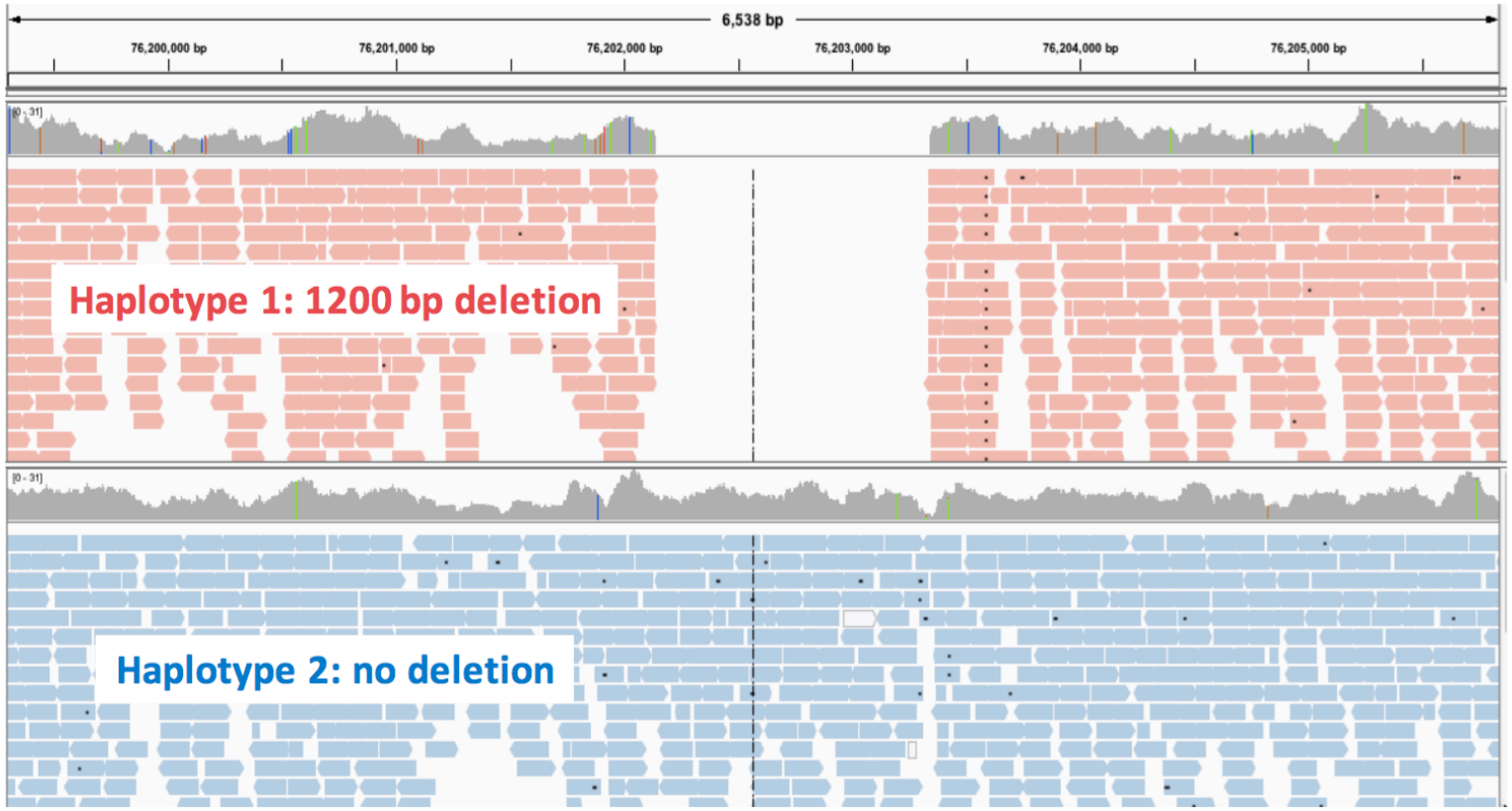
Dovetail Chicago Libraries



UCDAVIS Bioinformatics Core

Phasing: Linked Reads + high quality Illumina data





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

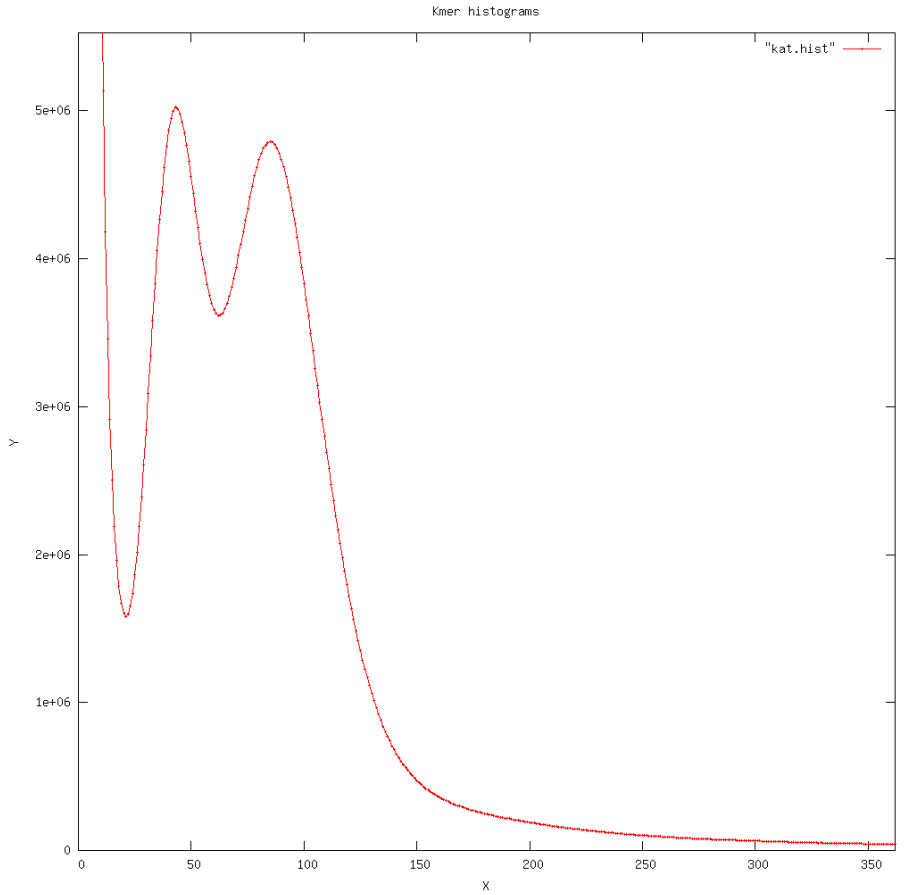
Adding in the optical maps from the Irys system reduced the total number of contigs to 1,780, with a contig N50 of 10.2 megabases. "The optical mapping increased the quality and confidence of the initial scaffolds," Phillippy said. The three technologies—PacBio, Bionano, and Hi-C—ended up being complementary 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

Order: The Kitchen Sink

- Available Technologies
  - Long Reads: — Long Contigs — **2 or 3**
    - Pacific Biosystems (HiFi), Oxford Nanopore (Possibly UL for scaffolding)
  - Optical Maps: — Scaffolding — **4**
    - BioNano — **1**
  - Linked Reads: ~~10x Genomics~~ — High base quality and phasing — **2**
    - TELL-Seq, stLFR (BGI), CPTv2-seq (Illumina)
  - Cross Linking: Hi-C — Scaffolding
    - Phase, Arima, Dovetail (OmniC)
- 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

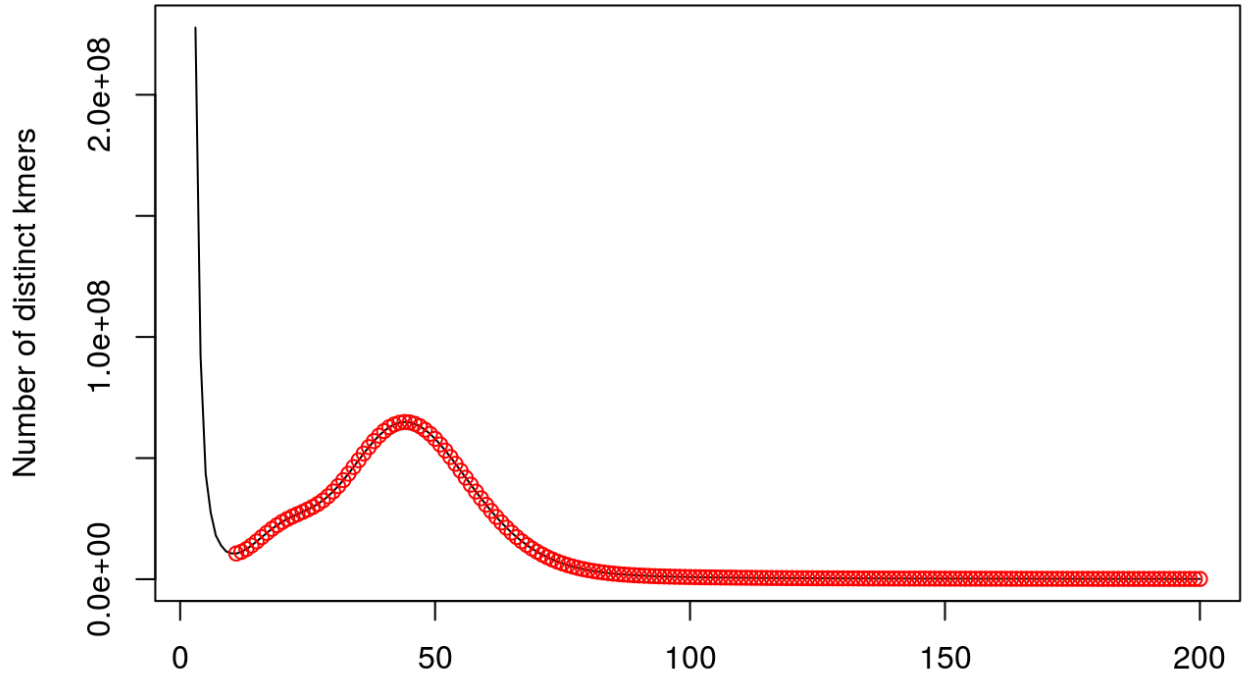
UCDAVIS Bioinformatics Core

Recommend to start with Linked Reads



- 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

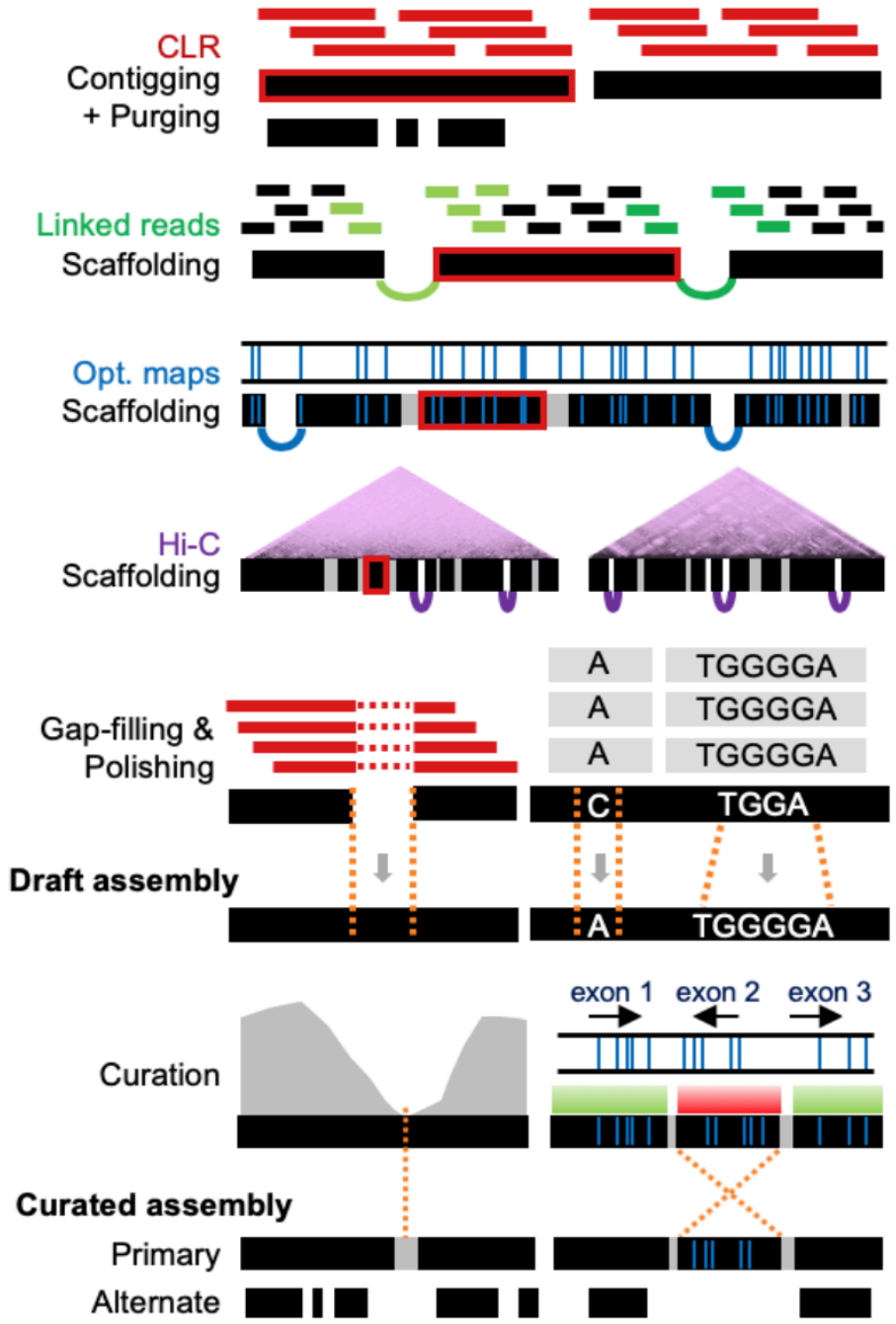
VGP Paper

Towards complete and error-free genome assemblies of all vertebrate species

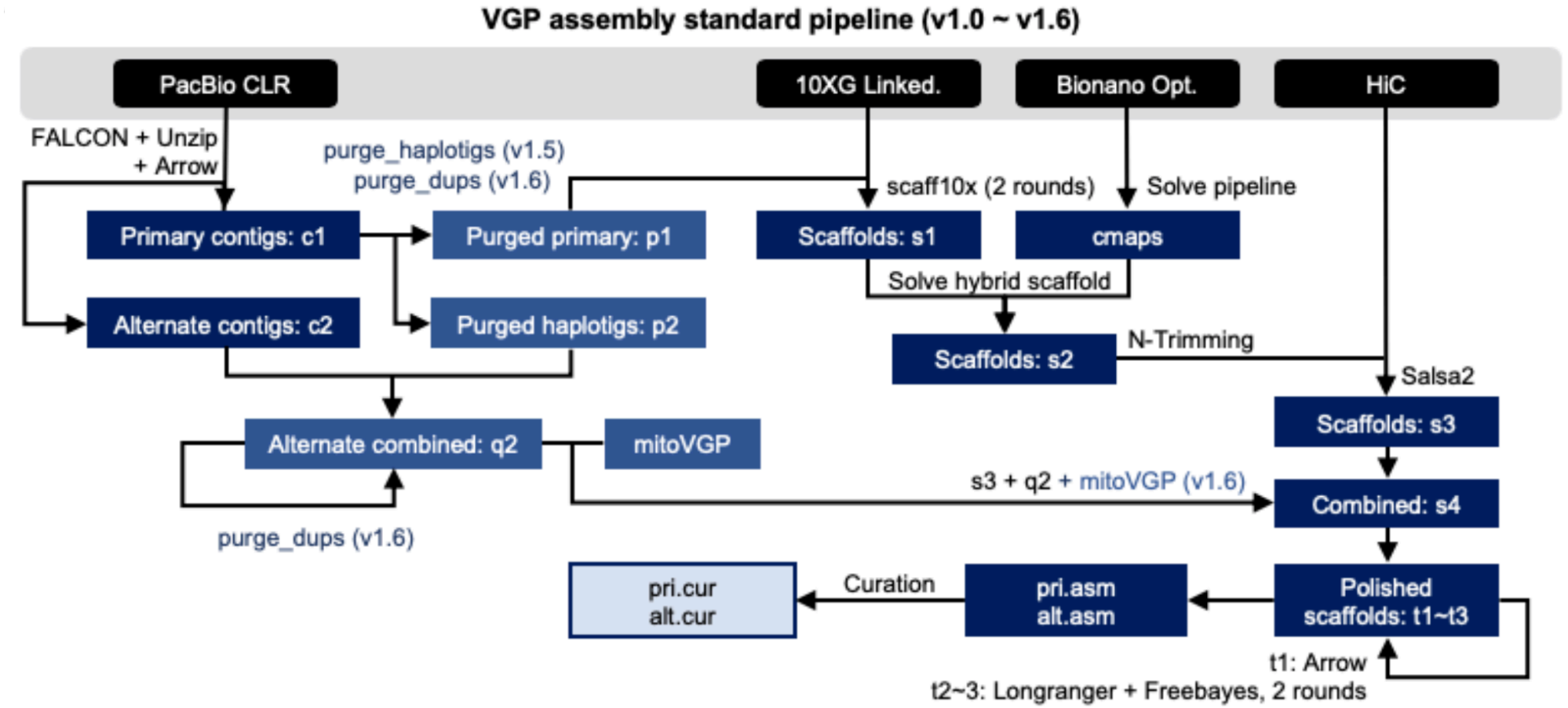
May 23, 2020

https://www.biorxiv.org/content/10.1101/2020.05.22.110833v1

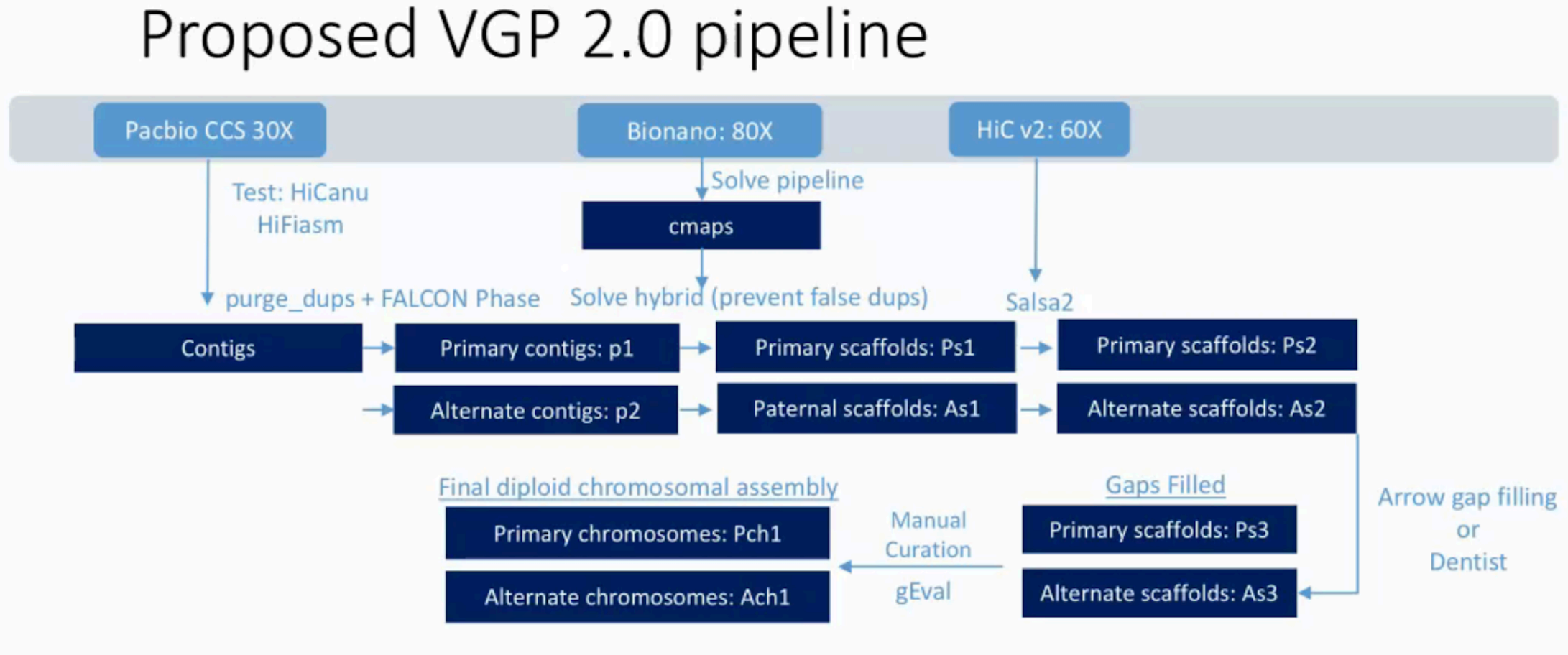
- Evaluation of data types and algorithms
- Pipeline development
- Establishment of assembly standards



VGP Paper (Pipeline v1.0 - v1.6)



VGP (Proposed Pipeline V2.0)



T2T Consortium

https://sites.google.com/ucsc.edu/t2tworkinggroup

- The first end-to-end (no gaps) Human Chromosome (X Chromosome), published July 14, 2020. https://www.nature.com/articles/s41586-020-2547-7

- 50X coverage of ultra-long Oxford Nanopore sequencing, including 44 Gb of sequence in reads 100 kb+ and a maximum read length exceeding 1 Mb.
- A de novo assembly combining this nanopore data with 70X of existing PacBio data achieved an NG50 contig size of 75 Mb (compared to 56 Mb for GRCh38), with some chromosomes broken only at the centromere.
- Using this assembly as a basis, T2T chose to then manually finish the X chromosome. The few unresolved segmental duplications were assembled using ultra-long reads spanning the individual copies, and the ~2.8 Mbp X centromere was assembled by identifying unique variants within the array and using these to anchor overlapping ultra-long reads.

Focus of the Future

- To some extent we are limited by being able to generate enough (quantity) high quality, high molecular weight DNA.
- Continued improvement to sequencing chemistries for consistent (PacBio Hifi CCS Reads) and longer reads, read quality improvement has become secondary.
- Incremental improvement of the computational algorithms.
- Scaffolding algorithms, algorithms merge multiple data types/sources (GFA2).
- Polyploidy is now doable and an active area of research.
- Haplotyped genomes – But 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.