

Long Read Sequencing Technology

- Algorithms and its applications -

Hayan Lee@Schatz Lab

May 18, 2015

Graduate Student Symposium

Outline




- **Background**
 - Long read sequencing technology
- **The Resurgence of reference quality genome (3Cs)**
 - The next version of Lander-Waterman Statistics (Contiguity)
 - Historical human genome quality by gene block analysis (Completeness)
 - The effectiveness of long reads in de novo assembly (Correctness)
- **Sugarcane de novo genome assembly challenge**
 - The effectiveness of **accurate long reads** in de novo assembly especially for highly heterozygous aneuploidy genome
 - Pure long read de novo assembly, combine with Moleculo and PacBio reads.
- **Contributions**

Background


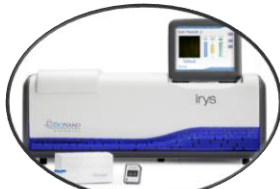
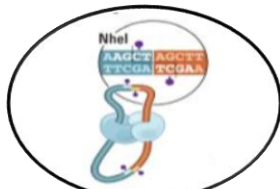
- **BAC-by-BAC + Sanger Era (~ 2007)**
 - Very high quality reference genomes for human, mouse, worm, fly, rice, Arabidopsis and a select few other high value species.
 - Contig sizes in the megabases, but costs in the 10s to 100s of millions of dollars
- **Next-Gen Era (2007 to current)**
 - Costs dropped, but genome quality suffered
 - Genome finishing almost completely abandoned; “exon-sized” contigs
 - These low quality draft sequences are (1) missing important sequences, (2) lack context to discover regulatory elements or evolutionary patterns, and (3) contain many errors
- **Third-Gen Era (current)**
 - New biotechnologies (single molecule, chromatin assays, etc) and new algorithms (MHAP, LACHESIS, etc) are leading to a *Resurgence of Reference Quality Genomes*
 - *De novo* assemblies of human and other large genomes with contig sizes over 1Mbp.

Third-Gen Sequencing Technology

- Long Read Sequencing: De novo assembly, SV analysis, phasing

<p>Illumina/Moleculo</p>  <p>3-5kbp (Kuleshov et al. 2014)</p>	<p>Pacific Biosciences</p>  <p>10-15kbp (Berlin et al, 2014)</p>	<p>Oxford Nanopore</p>  <p>5-10kbp (Quick et al, 2014)</p>
--	--	--

- Long Span Sequencing: Chromosome Scaffolding, SV analysis, phasing

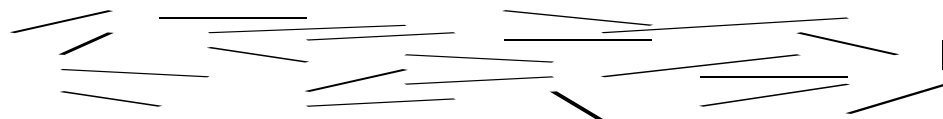
<p>Molecular Barcoding</p>  <p>30-60kbp (10Xgenomics.com)</p>	<p>Optical Mapping</p>  <p>100-150kbp (Cao et al, 2014)</p>	<p>Chromatin Assays</p>  <p>25-100kbp (Putnam et al, 2015)</p>
--	--	---

Outline

- **Background**
 - Long read sequencing technology and algorithms
- **The Resurgence of reference quality genome (3Cs)**
 - The next version of Lander-Waterman Statistics (Contiguity)
 - Historical human genome quality by gene block analysis (Completeness)
 - The effectiveness of long reads in de novo assembly (Correctness)
- **Sugarcane de novo genome assembly challenge**
 - The effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploidy genome
 - Pure long read de novo assembly, combine with accurate long reads and erroneous long reads
- **Contributions**

De novo genome assembly

1. Shear & Sequence DNA



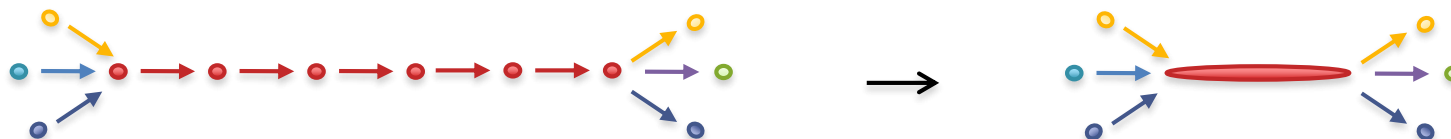
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

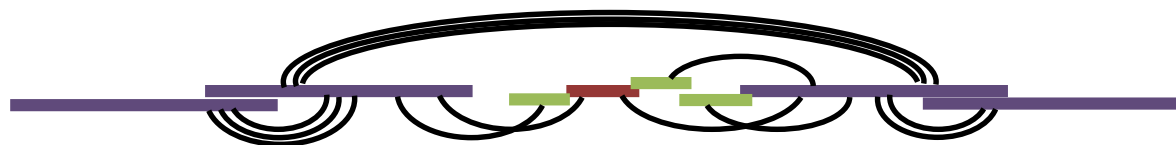
GGATGCGCGACACGT CGCATATCCGGTTTGGT CAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph



4. Detangle graph with long reads, mates, and other links



Many Genomes Are Sequenced...

Many Questions Are Raised...

But...

- How long should the read length be?
- What coverage should be used?

Given the read length and coverage,

- **How long are contigs? <- Contiguity prediction**
- How many contigs?
- How many reads are in each contigs?
- How big are the gaps?

Lander-Waterman Statistics

GENOMICS 2, 231-239 (1988)

Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis

ERIC S. LANDER^{*†} AND MICHAEL S. WATERMAN[‡]

^{*}Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142; [†]Harvard University, Cambridge, Massachusetts 02138; and [‡]Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089

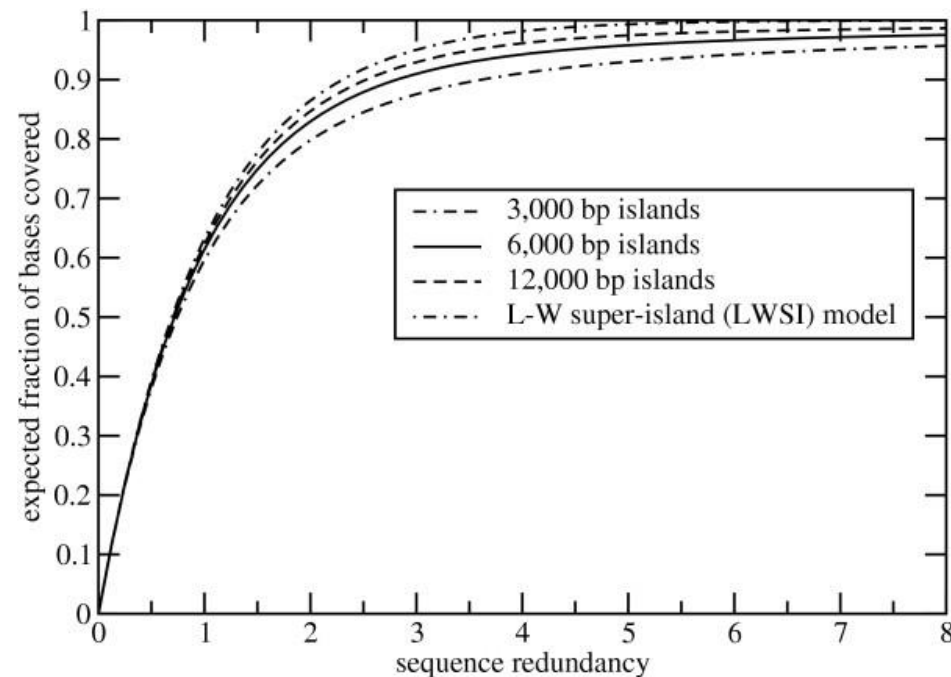
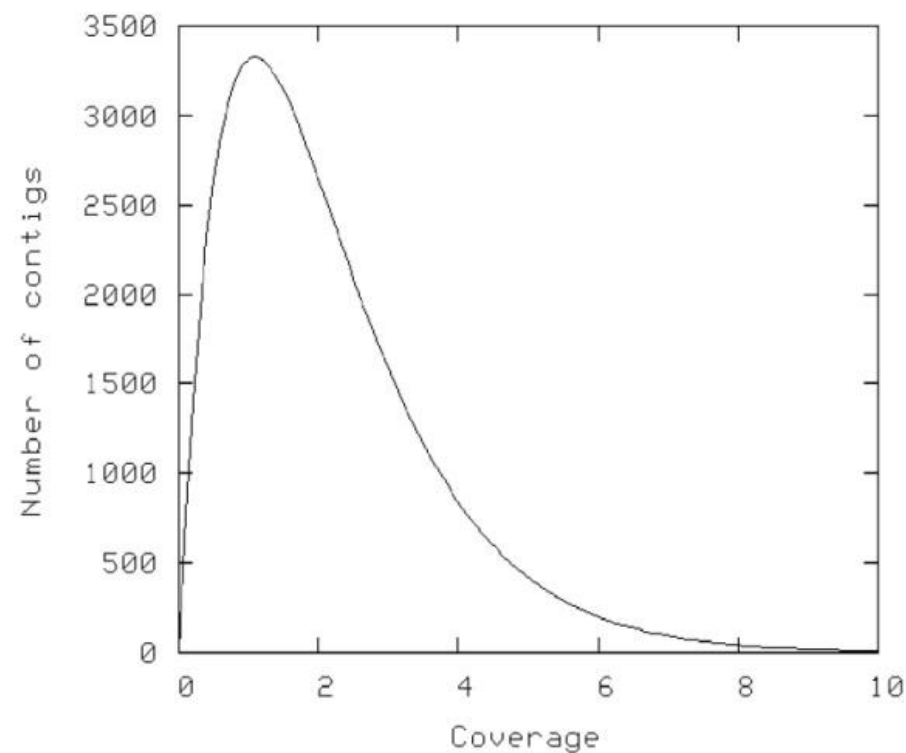
Received January 13, 1988; revised March 31, 1988

Results from physical mapping projects have recently been reported for the genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, and similar projects are currently being planned for other organisms. In such projects, the physical map is assembled by first "fingerprinting" a large number of clones chosen at random from a recombinant library and then inferring overlaps between clones with sufficiently similar fingerprints.

available region of up to several megabases and of studying its properties. In addition, the overlapping clones comprising the physical map would constitute the logical substrate for efforts to sequence an organism's genome.

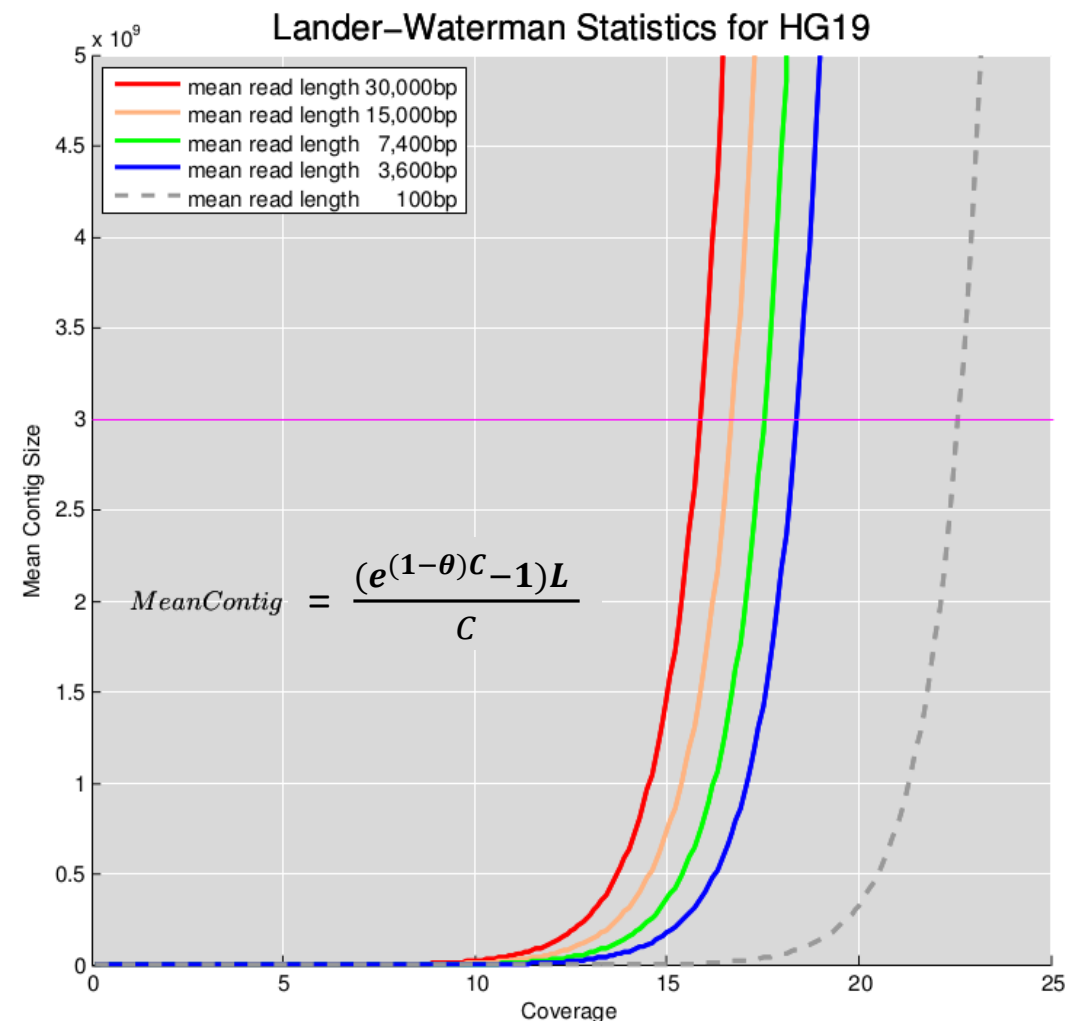
Recently, three pioneering efforts have investigated the feasibility of assembling physical maps by means of "fingerprinting" randomly chosen clones. The fingerprints consisted of information about restriction fragment lengths. Overlaps between clones were in-

Lander-Waterman Statistics



In practice, it's useful only in low coverage (3-5x) but becomes nonsensical in high coverage.

HG19 Genome Assembly Performance by Lander-Waterman Statistics



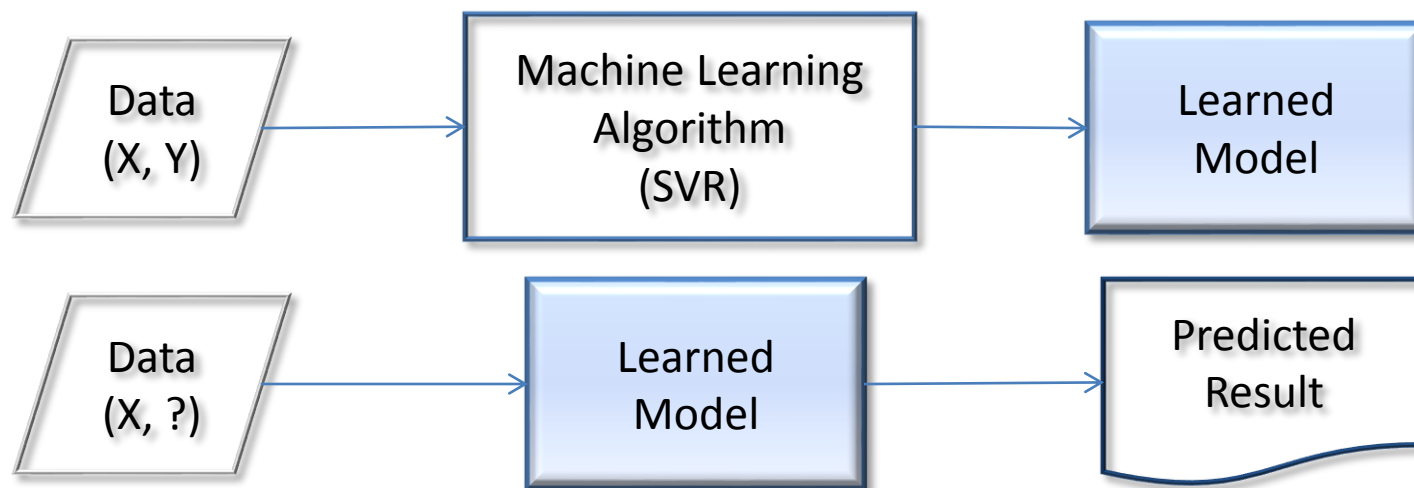
Two key observations

1. Contig over genome size
2. Read Length vs. Coverage

Technology vs. Money

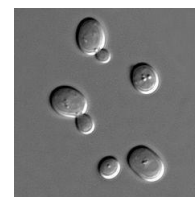
Empirical Data-driven Approach

- We selected 26 species across tree of life and exhaustively analyzed their assemblies using simulated reads for 4 different length (6 for HG19) and 4 different coverage per species
- For the extra long reads, we fixed the Celera Assembler(CA) to support reads up to 0.5Mbp

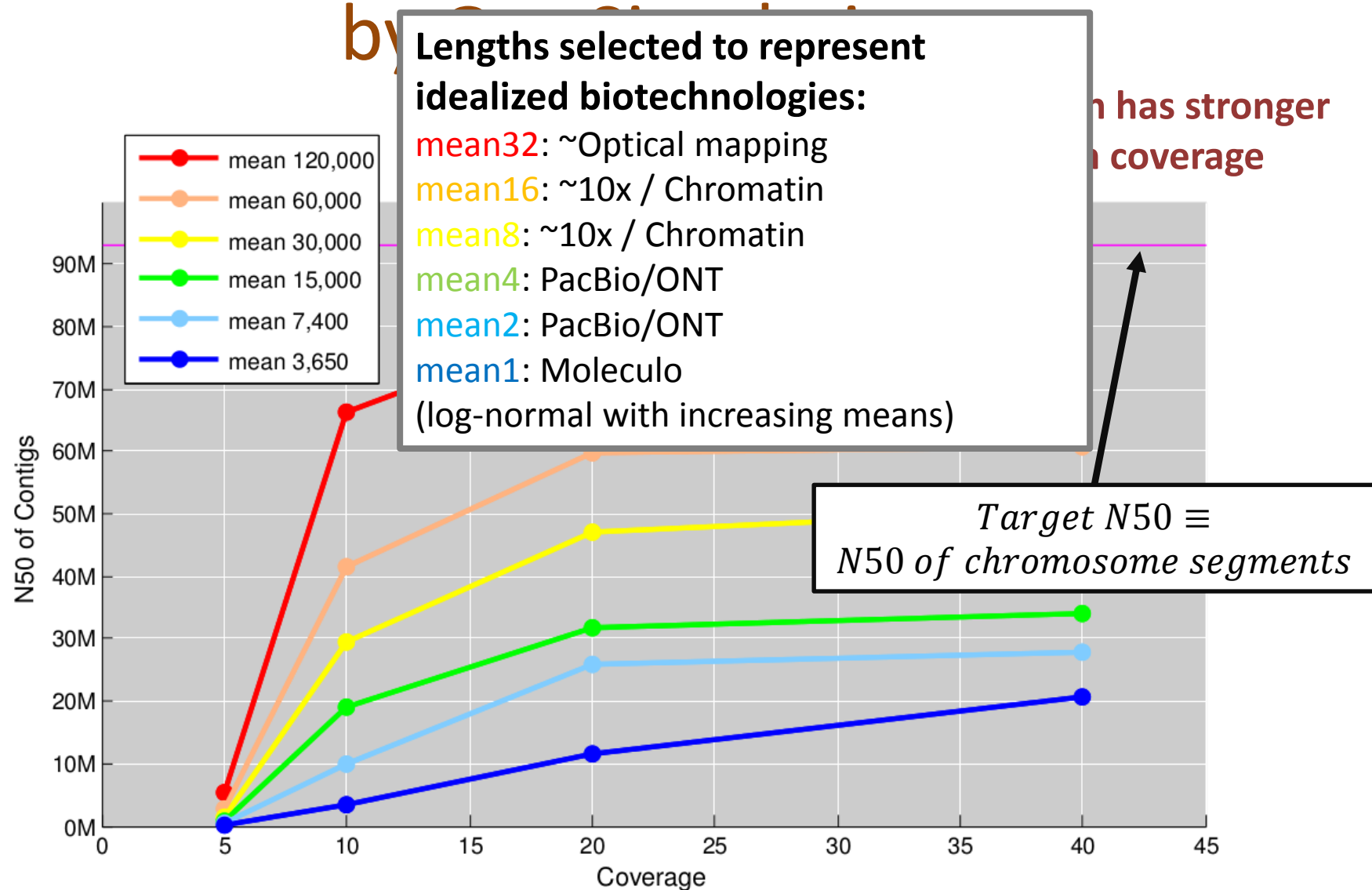


26 Species Across Tree of Life

Model Organism	ID	Genome Size
<i>M.jannaschii</i>	1	1,664,970
<i>C.hydrogenoformans</i>	2	2,401,520
<i>E.coli</i>	3	4,639,675
<i>Y.pestis</i>	4	4,653,728
<i>B.anthraxis</i>	5	5,227,293
<i>A.minum</i>	6	8,248,144
yeast	7	12,157,105
<i>Y.lipolytica</i>	8	20,502,981
slime mold	9	34,338,145
Red bread mold	10	41,037,538
sea squirt	11	78,296,155
roundworm	12	100,272,276
green alga	13	112,305,447
arabidopsis	14	119,667,750
fruitfly	15	130,450,100
peach	16	227,252,106
rice	17	370,792,118
poplar	18	417,640,243
tomato	19	781,666,411
soybean	20	973,344,380
turkey	21	1,061,998,909
zebra fish	22	1,412,464,843
lizard	23	1,799,126,364
corn	24	2,066,432,718
mouse	25	2,654,895,218
human	26	3,095,693,983



HG19 Genome Assembly Performance



Why?

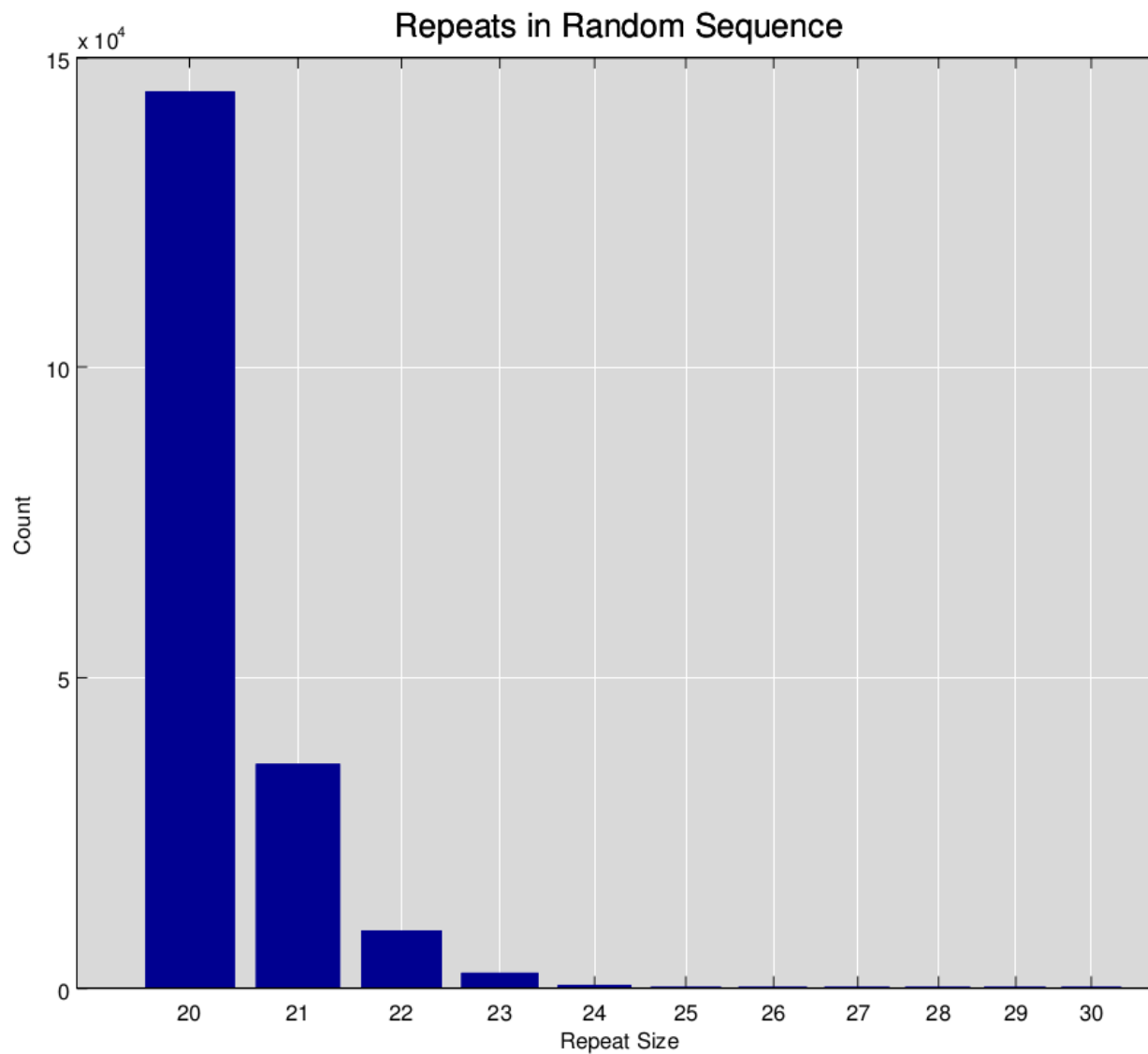
Lander-Waterman Statistics

- **Assumptions!!!**
- **If genome is a random sequence, it will work**
- **It works only in low coverage 3-5x**
- **It works for small genomes (< yeast)**

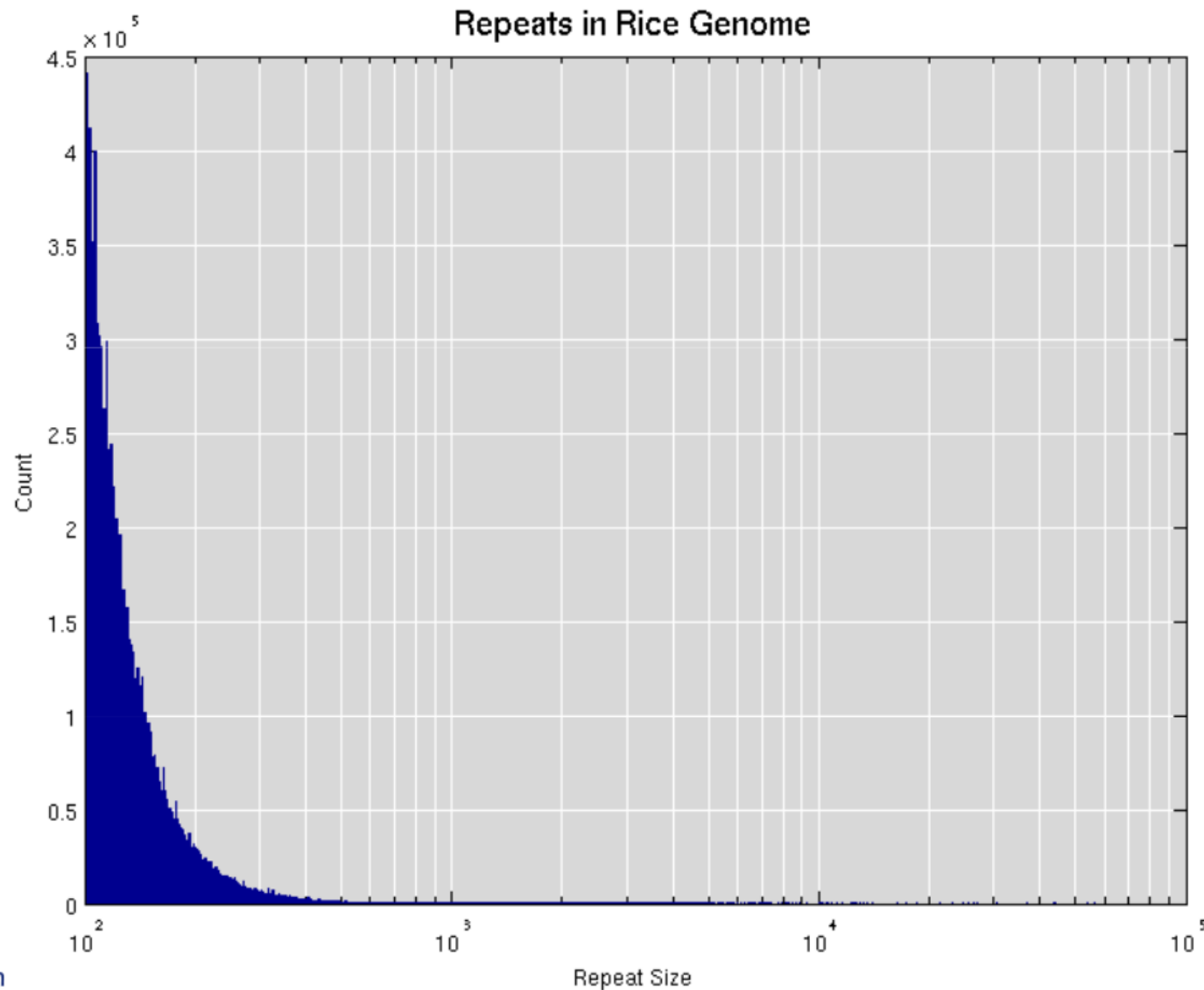
Our Approach

- Stop assuming that we cannot guarantee!!!
- We tried to assume as least as possible.
- Instead of building on top of assumptions, we let the model learn from the data
- Empirical data-driven approach

Repeats



Repeats in Rice



Our Goal

- To **predict** genome assembly **contiguity**

$$\text{Performance}(\%) \equiv \frac{N50 \text{ from assembly}}{N50 \text{ of chromosome segments}} \times 100$$

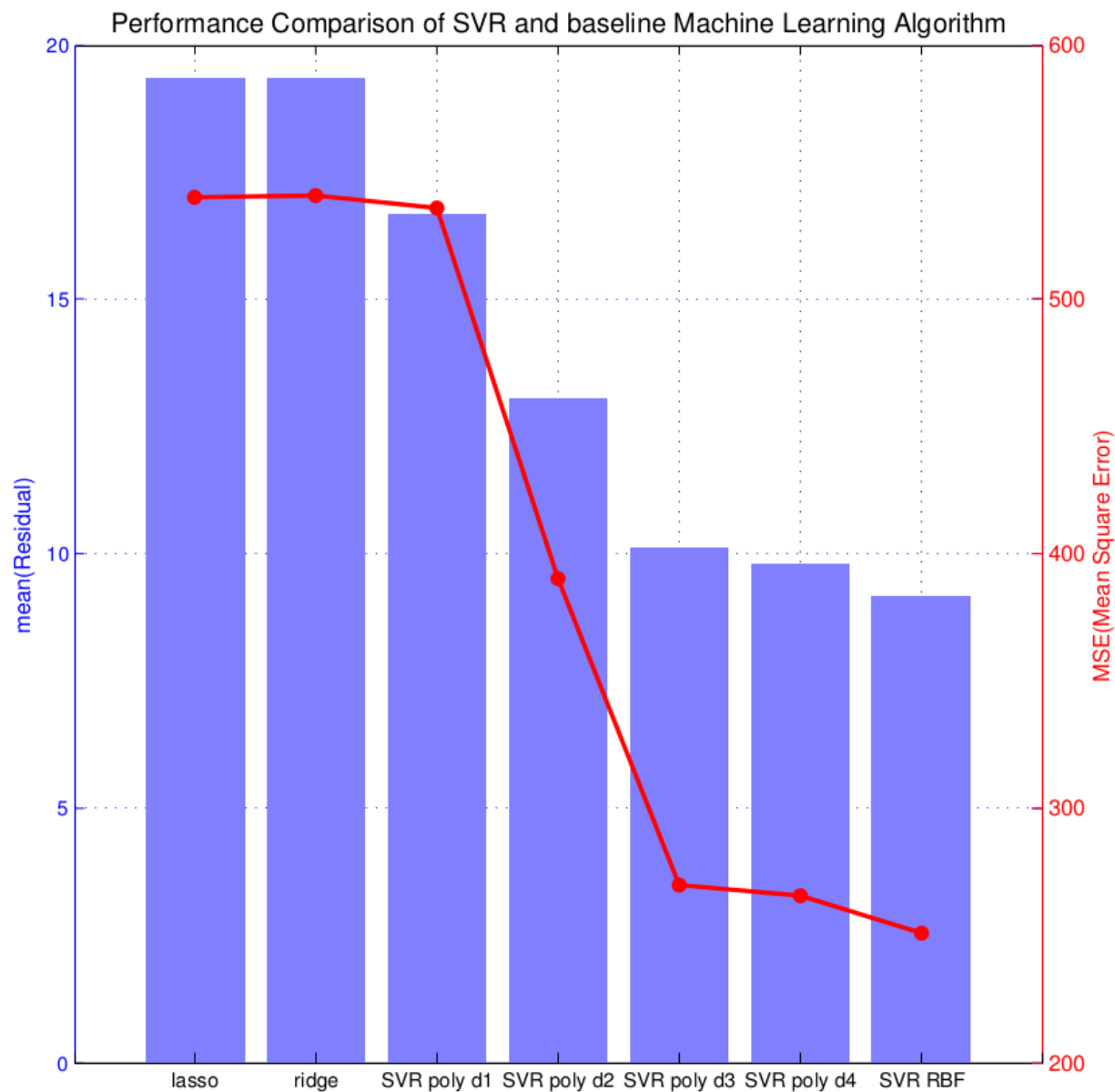
$$\approx f \left(\begin{array}{c} \text{Read Length} \\ \text{Coverage} \\ \text{Repeats} \\ \text{Genome Size} \end{array} \right)$$

Challenges for Prediction

- Sample size is small
- Quality is not guaranteed
- Predictive Power
- Overfitting

Support Vector Regression (SVR)
Cross Validation

The diagram consists of two text labels at the bottom: 'Support Vector Regression (SVR)' in dark red and 'Cross Validation' in dark blue. From 'SVR', three red arrows point upwards to the first three list items: 'Sample size is small', 'Quality is not guaranteed', and 'Predictive Power'. From 'Cross Validation', two blue arrows point upwards to the last two list items: 'Predictive Power' and 'Overfitting'.



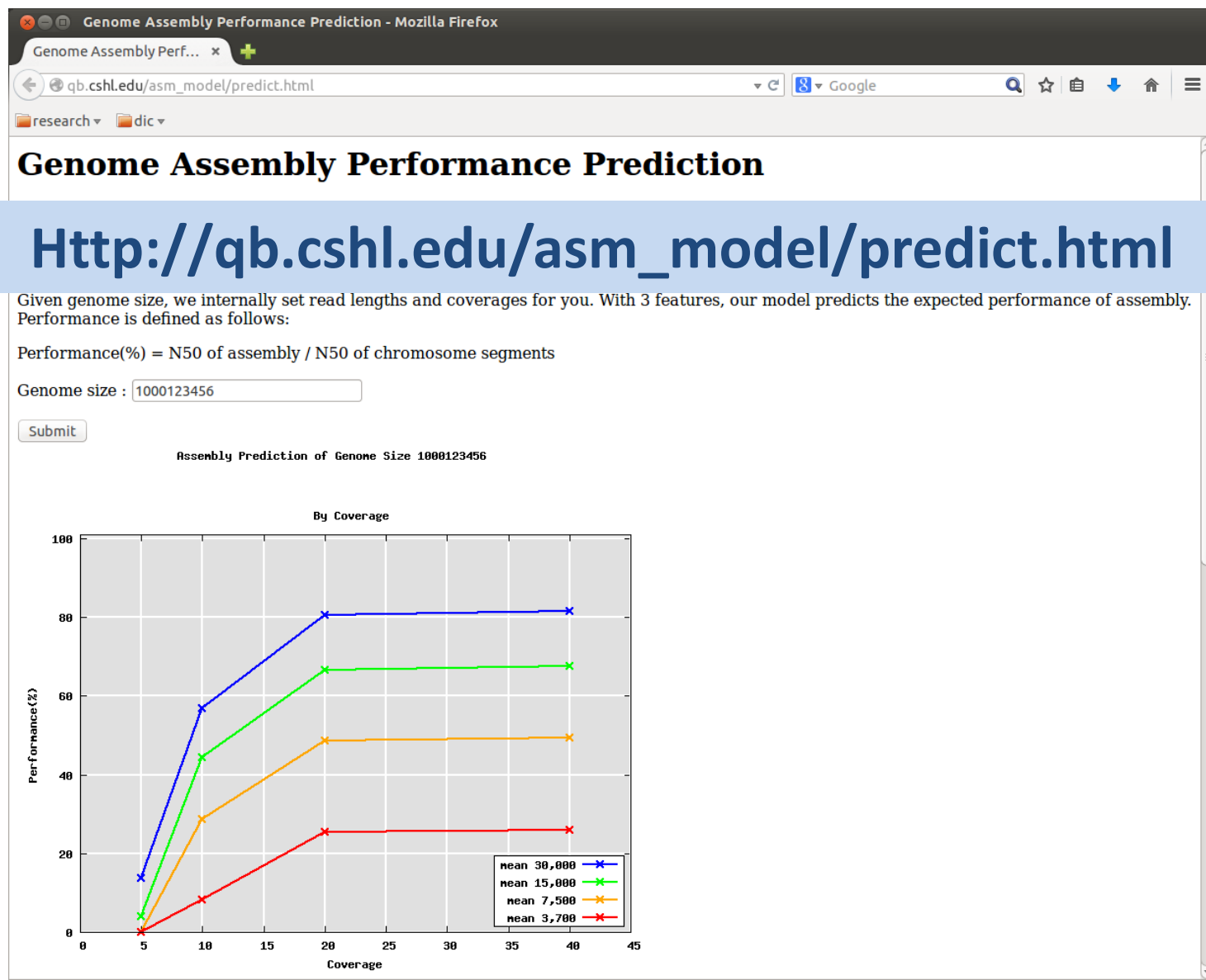
The resurgence of reference genome gaultiy

Lee, H, Gurtowski, J, Yoo, S, Marcus, S, McCombie, WR, Schatz MC *et al.* (2015) *In preparation*

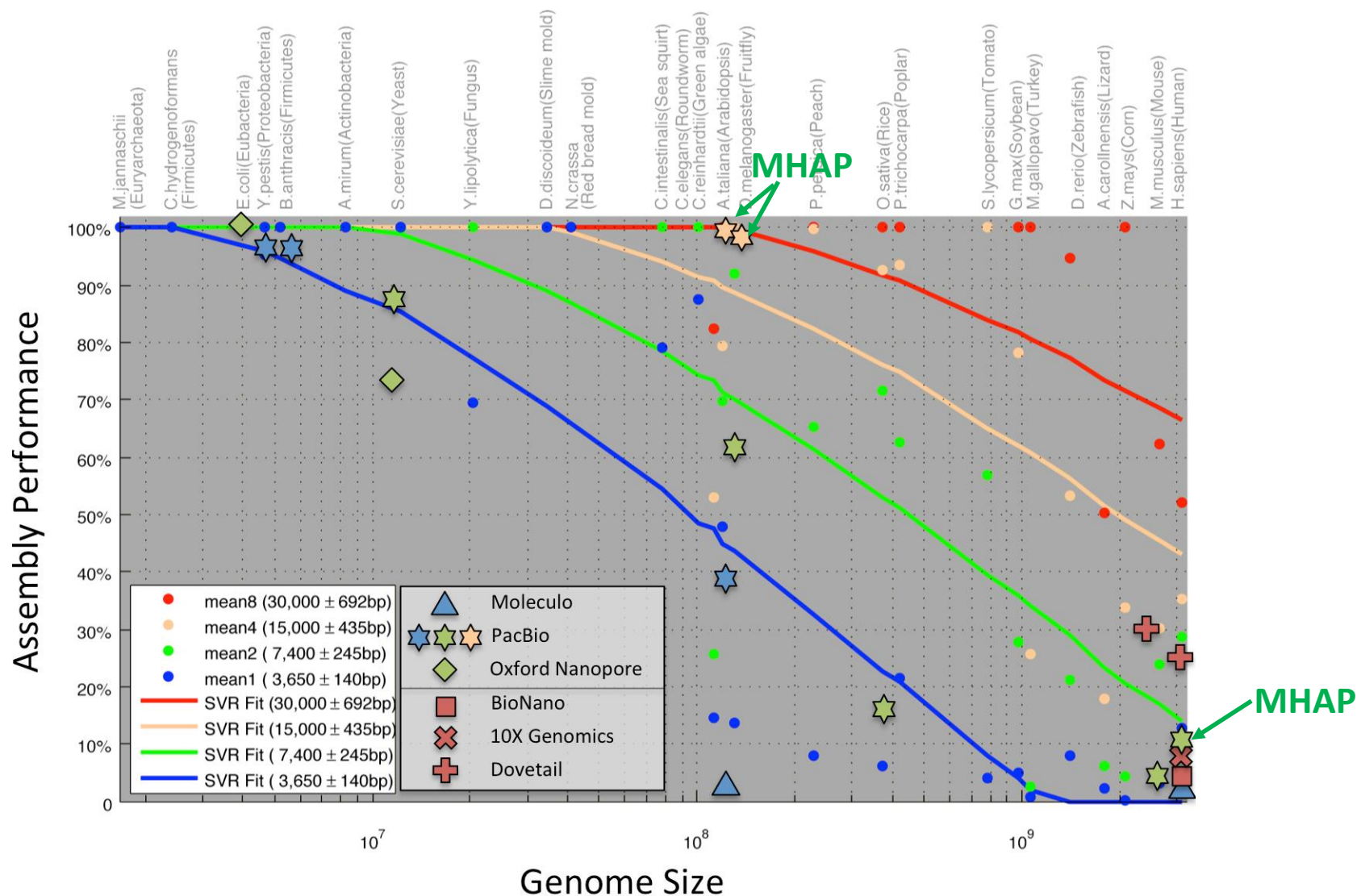
Predictive Power

- **Average of residual is 15%**
- **We can predict the new genome assembly performance in 15% of error residual boundary**
- **Genome size, read length and coverage used explicitly**
- **Repeats are included implicitly**

Web Service for Contiguity Prediction



Reference Genome Quality



Preprint



bioRxiv
beta
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS

Search

New Results

Error correction and assembly complexity of single molecule sequencing reads.

Hayan Lee , James Gurtowski , Shinjae Yoo , Shoshana Marcus , W. Richard McCombie , Michael Schatz

doi: <http://dx.doi.org/10.1101/006395>

Abstract

Info/History

Metrics

Data Supplements

☐ Preview PDF

☐ Previous

Posted June 18, 2014.

☐ Download PDF

☐ Email

Tweet 61

Like 8

Abstract

Third generation single molecule sequencing technology is poised to revolutionize genomics by enabling the sequencing of long, individual molecules of DNA and RNA. These technologies now routinely produce reads exceeding 5,000 basepairs, and can achieve reads as long as 50,000 basepairs. Here we evaluate the limits of single molecule sequencing by assessing the impact of long read sequencing in the assembly of the human genome and 25 other important genomes across the tree of life. From this, we develop a new data-driven model using support vector regression that can accurately predict assembly performance. We also present a novel hybrid error correction algorithm for long PacBio sequencing reads that uses pre-assembled Illumina sequences for the error correction. We apply it several prokaryotic and eukaryotic genomes, and show it can achieve near-perfect assemblies of small genomes (< 100Mbp) and substantially improved assemblies of larger ones. All source code and the assembly model are available open-source.

Subject Area

Bioinformatics

Subject Areas

All Articles

Animal Behavior and Co

Biochemistry

Bioengineering

Bioinformatics

Biophysics

Cancer Biology

Cell Biology

Developmental Biology

Ecology

Validated by MHAP

Add results



bioRxiv
beta
THE PREPRINT SERVER FOR BIOLOGY

[HOME](#) | [ABOUT](#) | [SUBMIT](#) | [ALERTS / RSS](#)



[Advanced Search](#)

New Results

Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing

Konstantin Berlin , Sergey Koren , Chen-Shan Chin , James Drake , Jane M Landolin , Adam M Phillippy
doi: <http://dx.doi.org/10.1101/008003>

Abstract

[Info/History](#)

[Metrics](#)

[Data Supplements](#)

[Preview PDF](#)

[Previous](#)

Posted August 14, 2014.

[Download PDF](#)

[Email](#)

[Share](#)

[Citation Tools](#)

[Tweet](#) 167

[Like](#) 13

[g+1](#) 2

Subject Area

Bioinformatics

Subject Areas

All Articles

[Animal Behavior and Cognition](#)

[Biochemistry](#)

[Bioengineering](#)

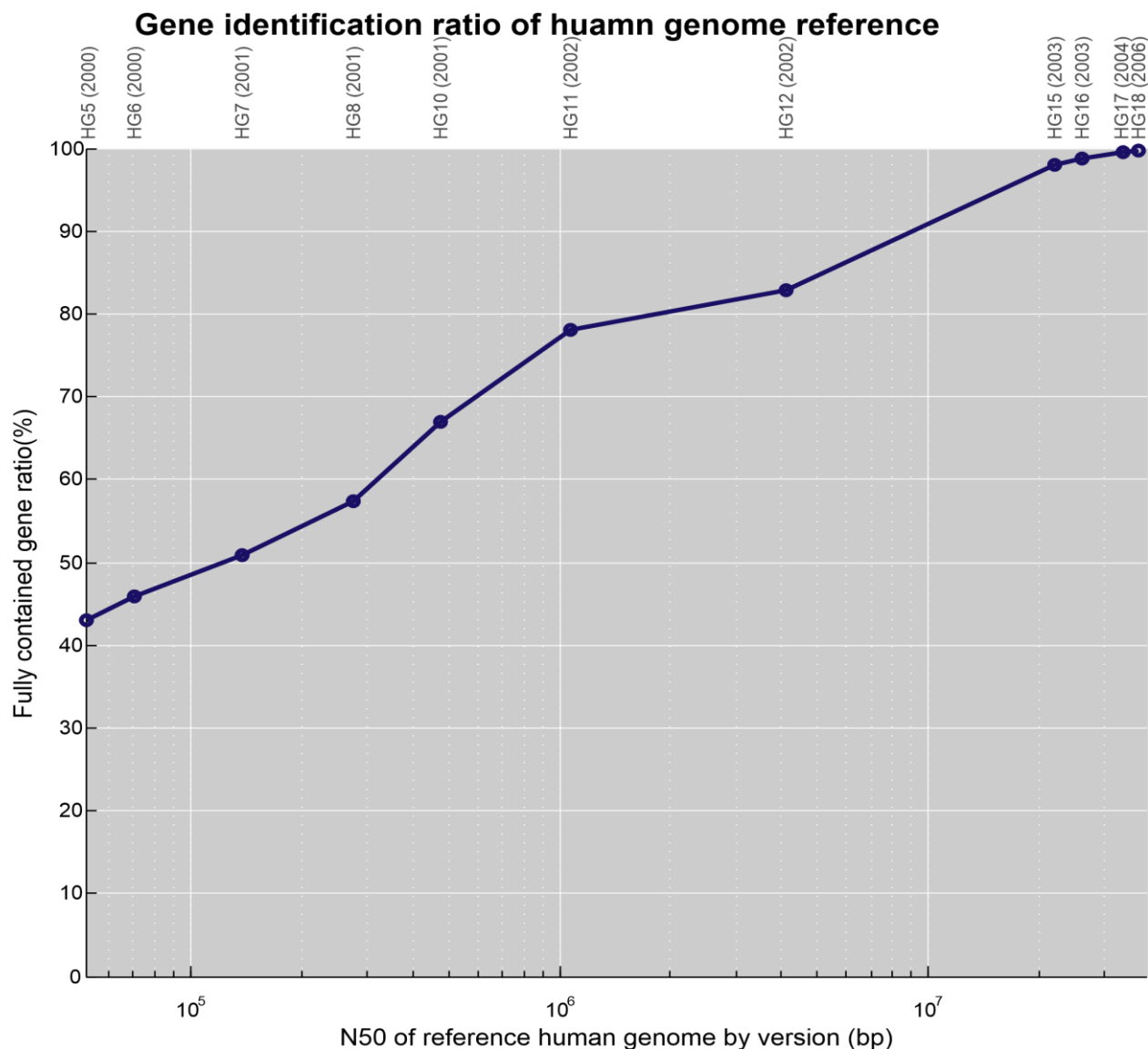
[Bioinformatics](#)

[Biophysics](#)

We report reference-grade de novo assemblies of four model organisms and the human genome from single-molecule, real-time (SMRT) sequencing. Long-read SMRT sequencing is routinely used to finish microbial genomes, but the available assembly methods have not scaled well to larger genomes. Here we introduce the MinHash Alignment Process (MHAP) for efficient overlapping of noisy, long reads using probabilistic, locality-sensitive hashing. Together with Celera Assembler, MHAP was used to reconstruct the genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and human from high-coverage SMRT sequencing. The resulting assemblies include fully resolved chromosome arms and

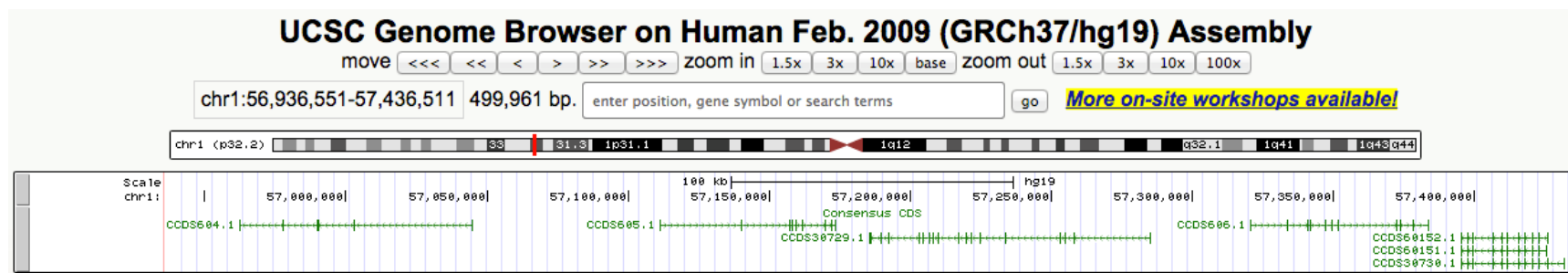
Completeness

Human Reference Genome Quality by gene block analysis



Completeness

Human Reference Genome Quality by gene block analysis



gene1

gene2

gene1 - Gene

gene2

gene5

gene10

gene20

Regulatory elements

gene50

gene100

gene200

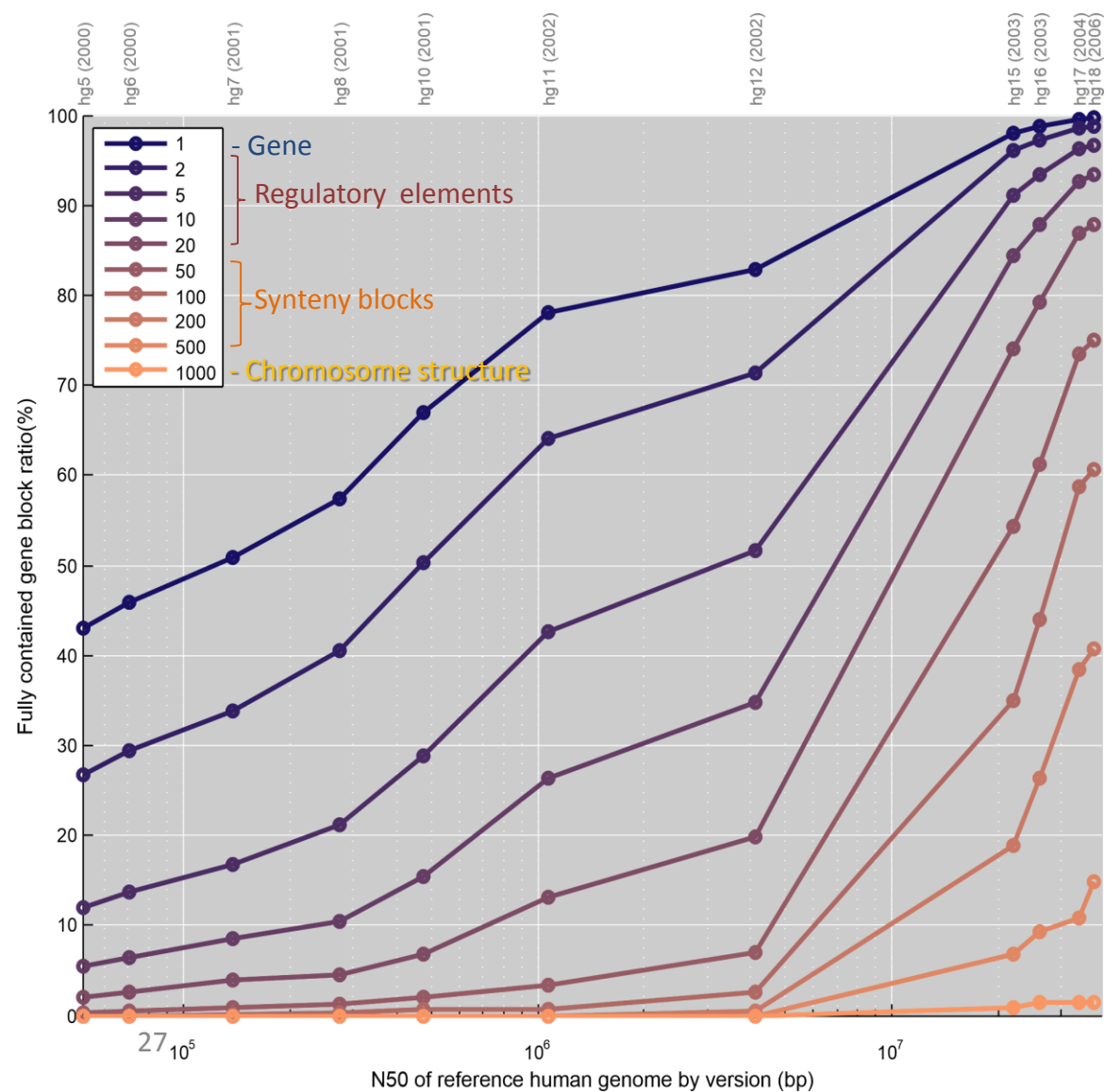
gene500

Synten blocks

gene1000 - Chromosome structure

Completeness

Human Reference Genome Quality by gene block analysis



Larger contigs and scaffolds empowers analysis at every possible level.

- SNPs (~10k clinically relevant)
- Genes
- Regulatory elements
- Synteny blocks
- Chromosome structure

Correctness Summary in HG19

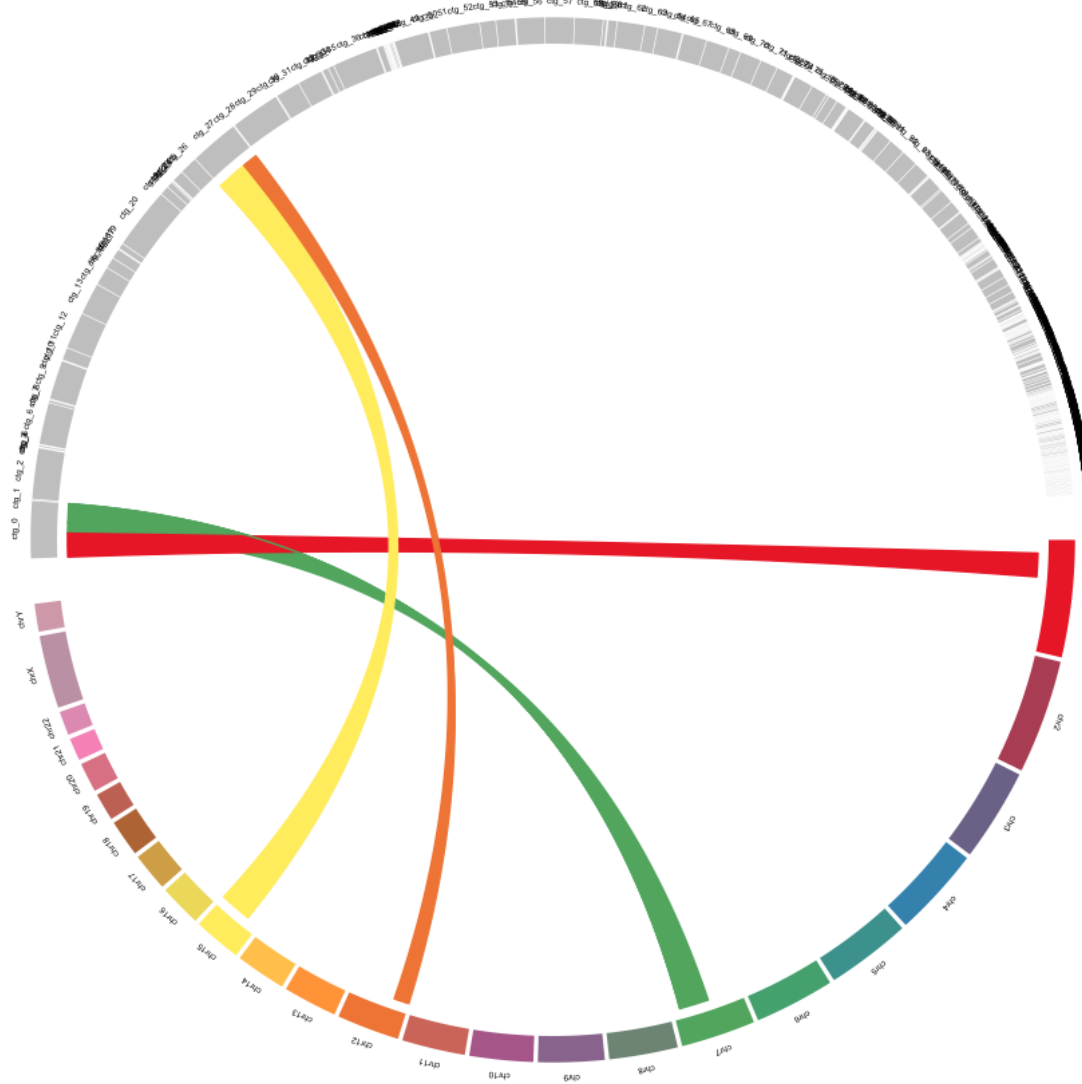
N50 misleading

HG19	(major) misassembly	(major) breaks
	False Positive	False Negative
	Increase N50 (falsely lengthen contiguity)	Decrease N50 (shorten contiguity)
	Mislead us in biological meaning	Negatively impact on downstream research
Mean1	209	4069
Mean2	70	462
Mean4	49	296
Mean8	33	197
Mean16	9	42
Mean32	7	5

Misassembly

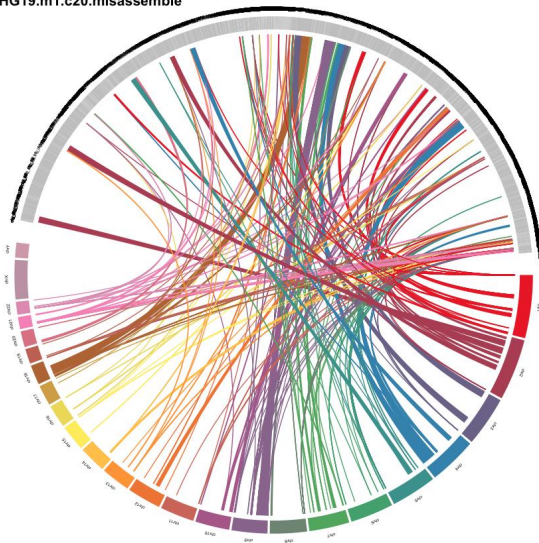
A critical error in de novo assembly

HG19.m8.c20.misassemble

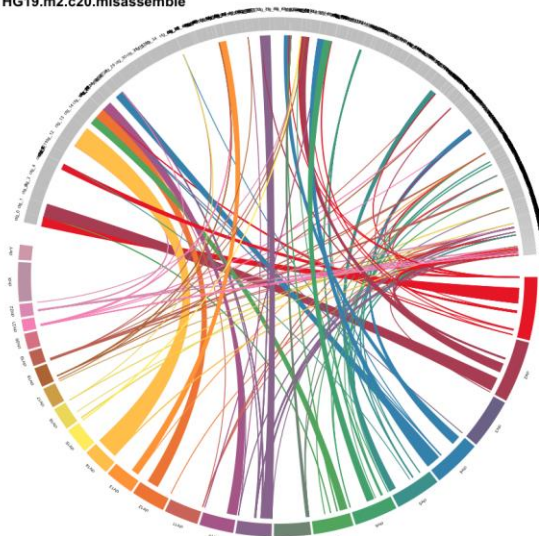


Misassembly Analysis in HG19

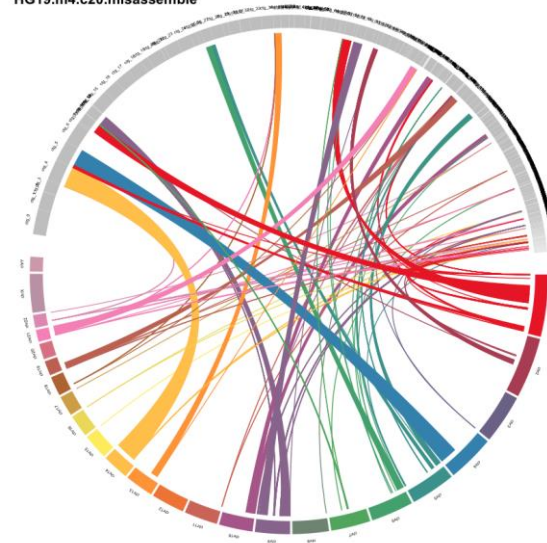
HG19.m1.c20.misassemble



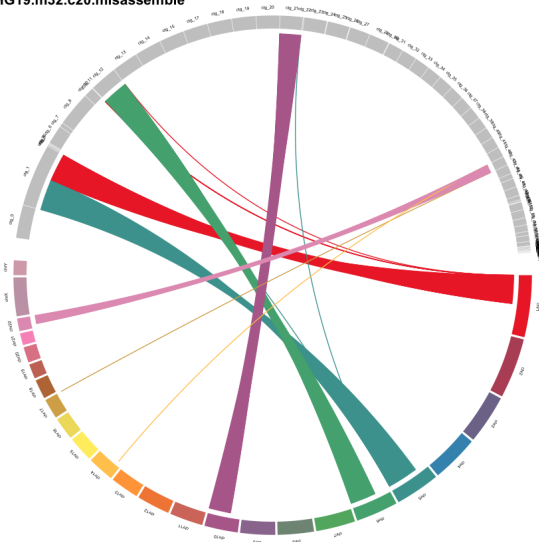
HG19.m2.c20.misassemble



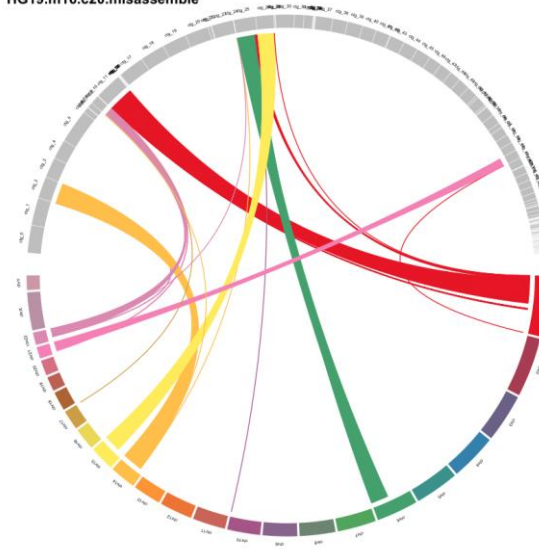
HG19.m4.c20.misassemble



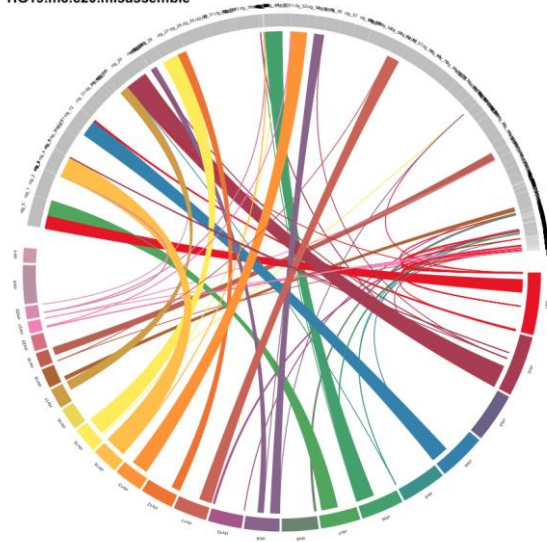
HG19.m32.c20.misassemble



HG19.m16.c20.misassemble

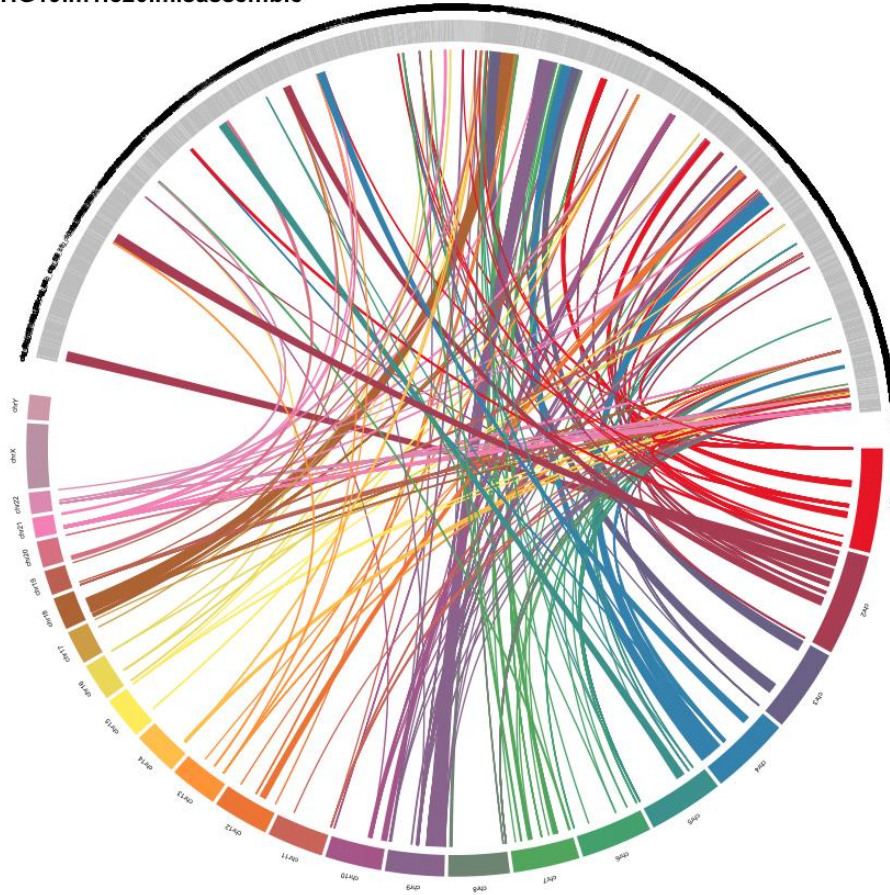


HG19.m8.c20.misassemble

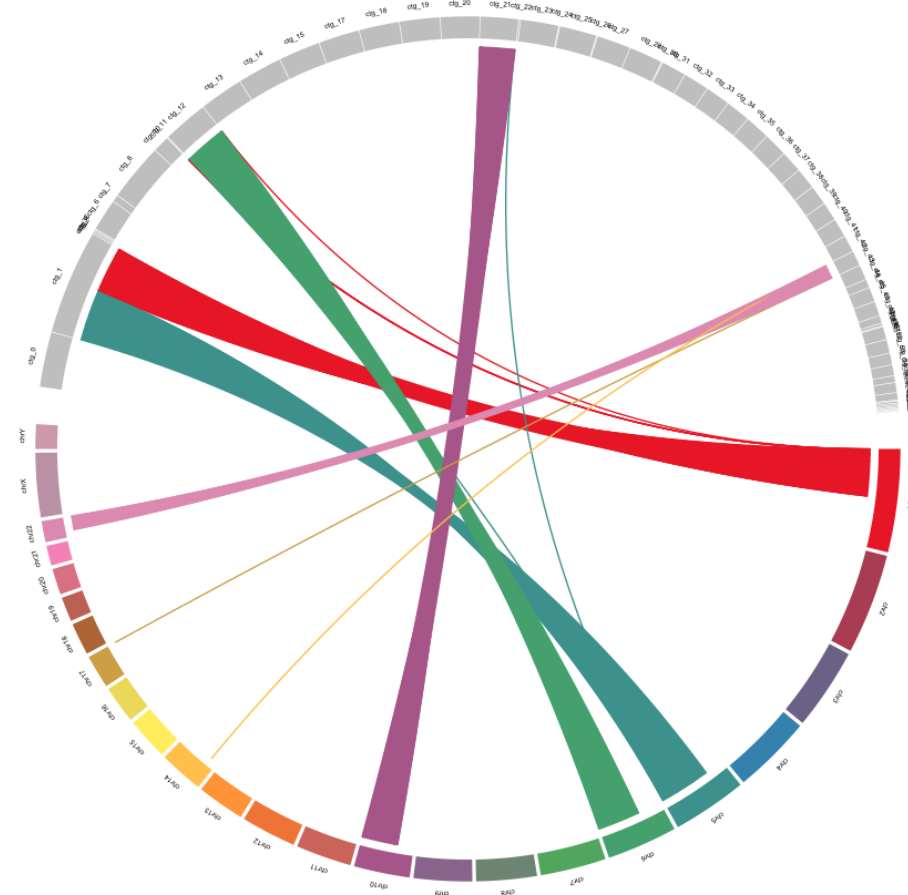


Misassembly Analysis in HG19

HG19.m1.c20.misassemble



HG19.m32.c20.misassemble



Long read sequencing technology helps to reduce both misassembly and breaks thus increase correctness of de novo genome assembly

Summary & Recommendations

Reference quality genome assembly is here

- Use the longest possible reads and spans for the best assembly
- Coverage and algorithmics overcome most random errors

Megabase N50 improves the analysis in every dimension

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization

Need to develop methods to jointly analyze multiple high-quality references at once

Outline

- **The Resurgence of reference genome quality (3Cs)**
 - The next version of Lander-Waterman Statistics (Contiguity)
 - Historical human genome quality by gene block analysis (Completeness)
 - The effectiveness of long reads in de novo assembly (Correctness)
- **Sugarcane de novo genome assembly challenge**
 - The effectiveness of **accurate long reads** in de novo assembly especially for highly heterozygous aneuploidy genome
 - Pure long read de novo assembly, combine with Moleculo and PacBio reads.
- **Contributions**

Sugarcane for food and biofuel

- **Food**

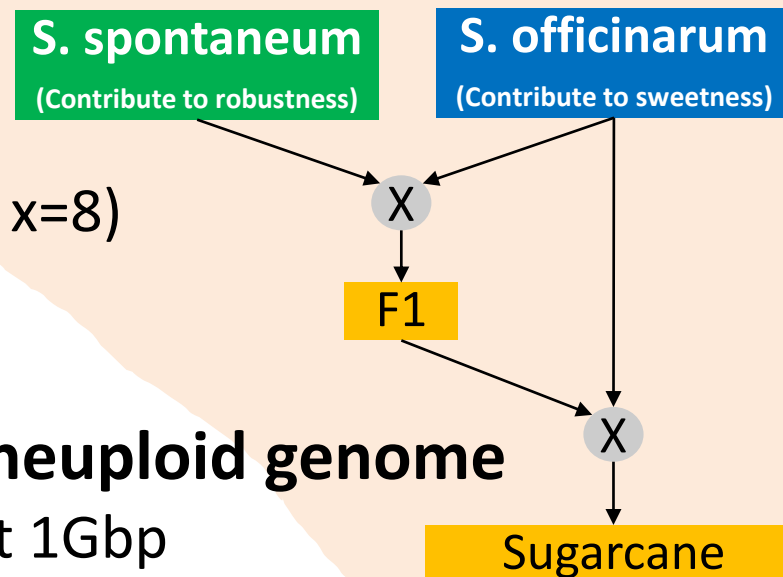
- By 2050, the world's population will grow by 50%, thus another 2.5 billion people will need to eat!
- Rapidly rising oil prices, adverse weather conditions, speculation in agricultural markets are causing more demand

- **Biofuel**

- By 2050, global energy needs will double as will carbon dioxide emission
- Low-carbon solution
- Sugarcane ethanol is a clean, renewable fuel that produces on average 90 percent less carbon dioxide emission than oil and can be an important tool in the fight against climate change.

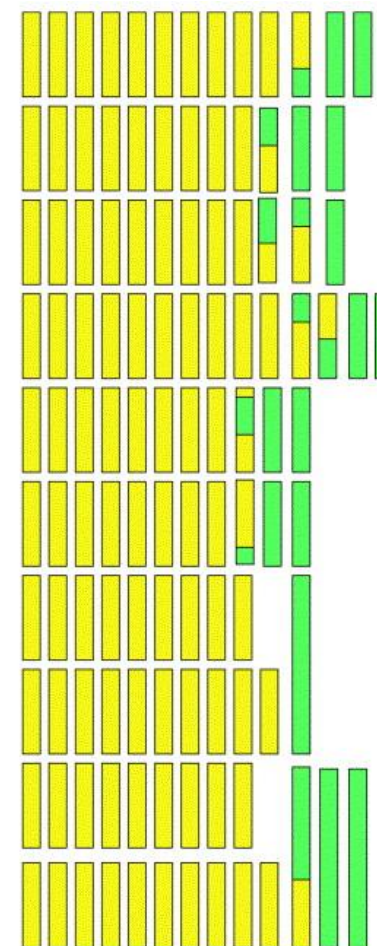
A hybrid sugarcane cultivar SP80-3280

- **S.spontaneum x S.officinarum**
- **A century ago....**
- **Saccharum genus**
 - S. spontaneum ($2n=40-128$, $x=8$)
 - S. officinarum ($2n=8x=80$)
- **Big, highly polyploid and aneuploid genome**
 - Monoploid genome is about 1Gbp
 - 8-12 copies per chromosome
 - In total, 100-130 chromosomes
 - Total size is about 10Gbp



Why is sugarcane assembly harder? (1)

- **Polyploidy/Aneuploidy**
 - 10% of the chromosomes are inherited in their entirety from *S. spontaneum*, 80% are inherited entirely from *S. officinarum*
- **Large scale recombination**
 - 10% is the result of recombination between chromosomes from the two ancestral species, a few being double recombinants



(source) <http://ars.elsa-cdn.com/content/image/1-s2.0-S1369526602002340-gr1.jpg>

Four Important Questions in Sugarcane

- **Scaffold polyploidy/aneuploidy genome**
 - How do we connect contigs/cluster contigs per chromosome/fill gaps among contigs?
- **Phasing haplotypes**
 - Not solved in diploid genome yet
- **Heterozygosity**
 - How do we measure heterozygosity in polyploidy/aneuploidy genome?
 - How do we quantify alleles and get ratio?
- **Inference of polyploidy/aneuploidy estimation**
 - How do we infer the number of copies per chromosome in aneuploidy genome, especially in the large scale of recombination?

Margarido GRA, Heckerman D (2015) ConPADE: Genome Assembly Ploidy Estimation from Next-Generation Sequencing Data. *PLoS Comput Biol* 11(4): e1004229. doi: 10.1371/journal.pcbi.1004229

Choose the right data and the right method

DATA	<p>Hiseq 2000 PE (2x100bp)</p> <ul style="list-style-type: none">- 575Gbp- 600x of haploid genome <p>Roche454</p> <ul style="list-style-type: none">- 9x of haploid genome- [min=20 max=1,168]- Mean=332bp	<p>Moleculo</p> <ul style="list-style-type: none">- 19Gbp- 19x of haploid genome- [min=1,500 max=22,904]- Mean = 4,930bp
Algorithm	SOAPdenovo (De Bruijn Graph)	Celera Assembler (Overlap Graph)
RESULT	<p>Max contig = 21,564 bp</p> <p>NG50=823 bp</p> <p>Coverage=0.86x</p>	<p>Max contig = 467,567 bp</p> <p>NG50=41,394 bp</p> <p>Coverage=3.59x</p> <p># of contigs = 450K</p>

CEGMA

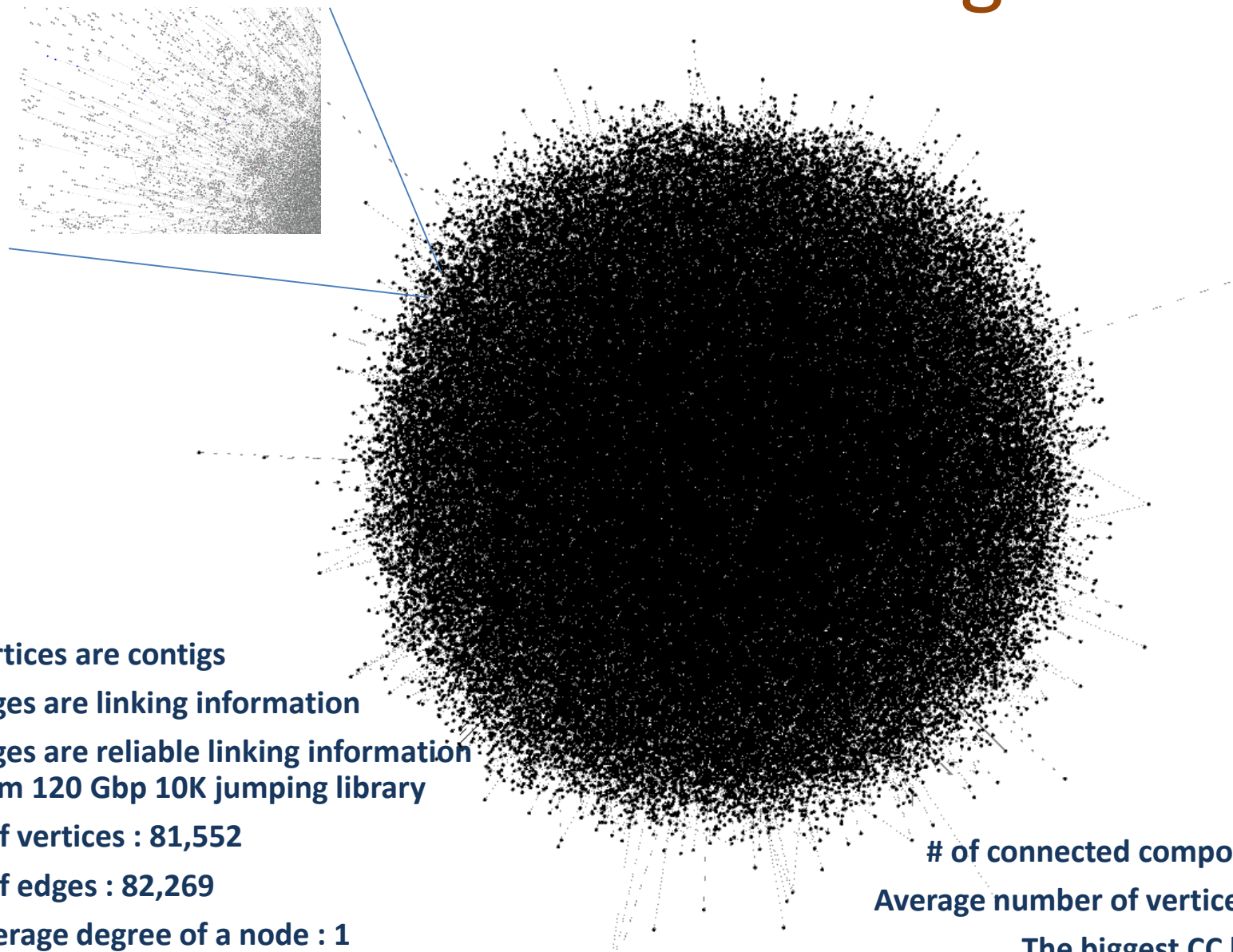
- **CEGs**
 - Korf Lab in UC. Davis selected 248 core eukaryotic genes

- **Statistics of the completeness**

	Prots	%Completeness	Total	Average	%Ortho
Complete	219	88.31	827	3.78	89.04
Partial	242	97.58	1083	4.48	95.45

- **Gene prediction aided by sorghum gene model**
 - In progress...
 - 39k sorghum genes were found in sugarcane contigs at least partially

NP-Hard Hairball of Sugarcane



Vertices are contigs

Edges are linking information

Edges are reliable linking information
from 120 Gbp 10K jumping library

of vertices : 81,552

of edges : 82,269

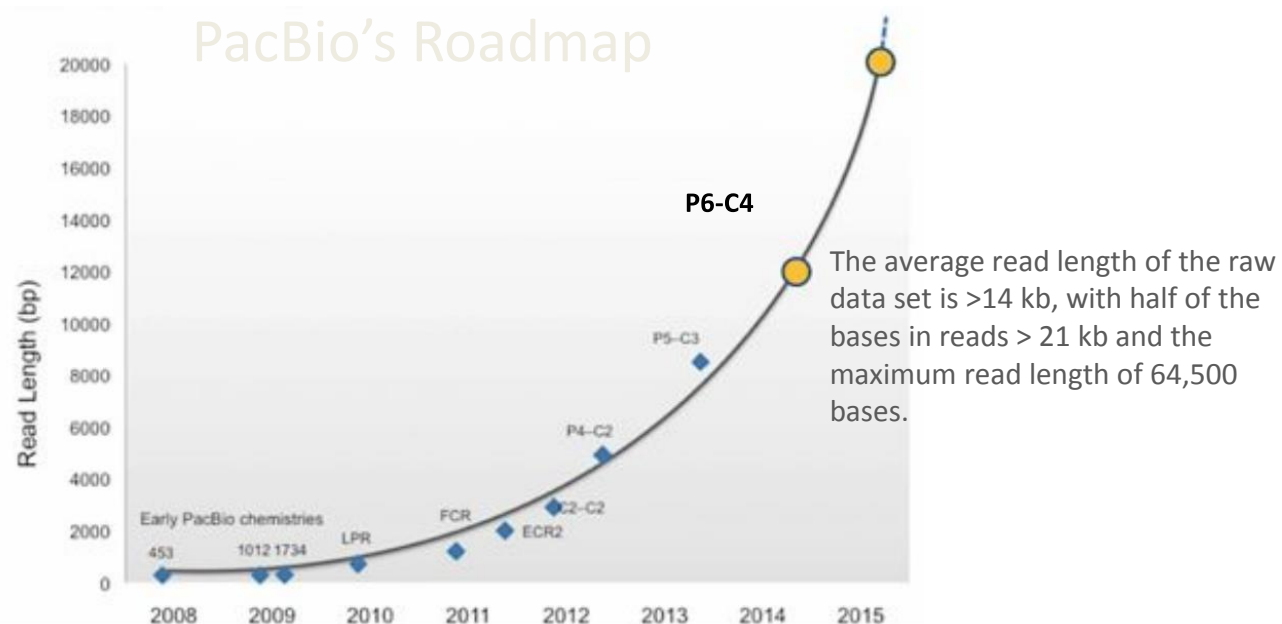
Average degree of a node : 1

of connected components = 17,919

Average number of vertices per CC= 2.54

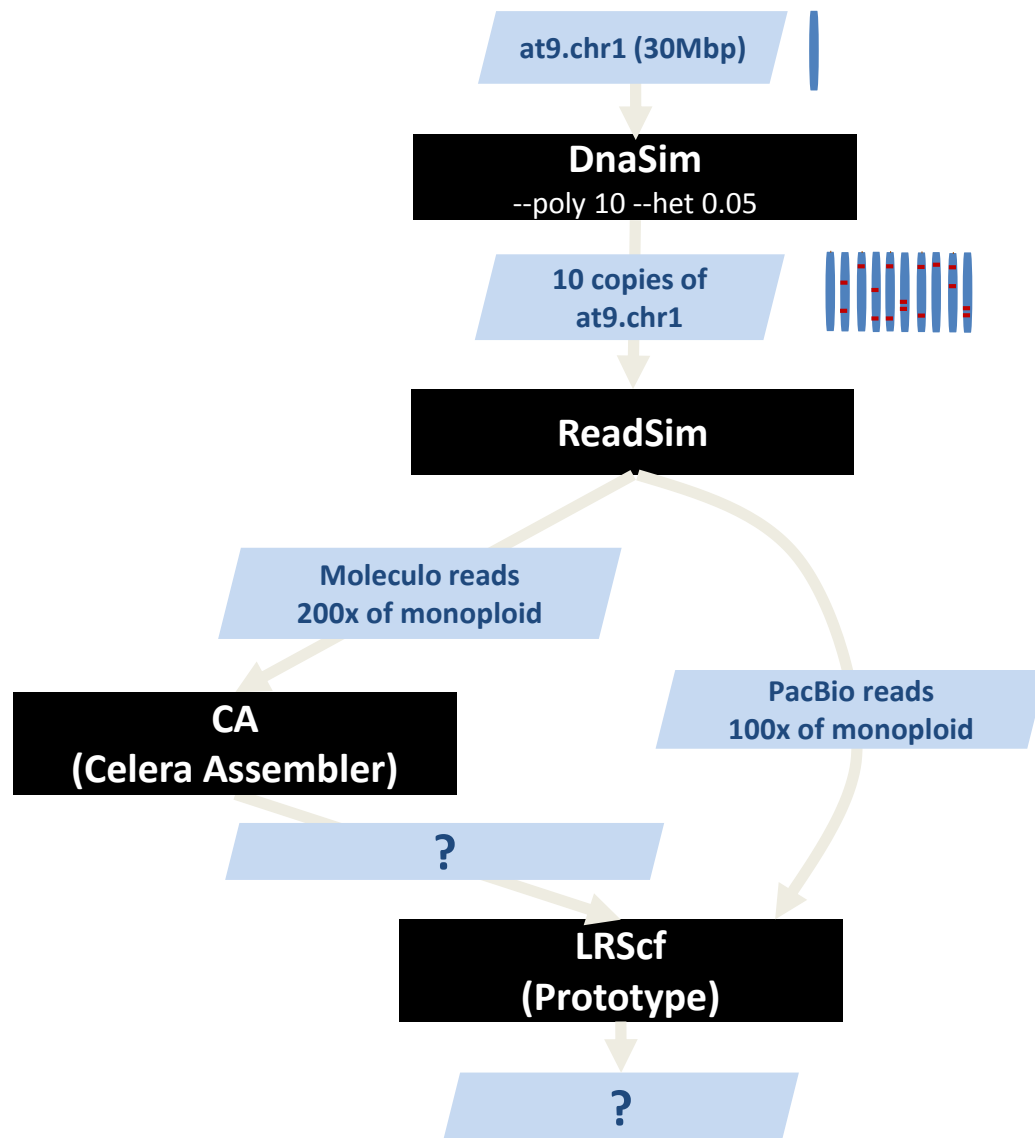
The biggest CC has 25 vertices

Benefits of Long Read Scaffolding



- Read Length is increasing, the cost is decreasing
- Very informative whether it has high error rate or not
- More repeats resolved
- Better scaffolding solution than long jumping library
- We don't have to approximate insert size by MLE or so.
- It's much better to fill gaps with some base information rather than just NNNNNN.

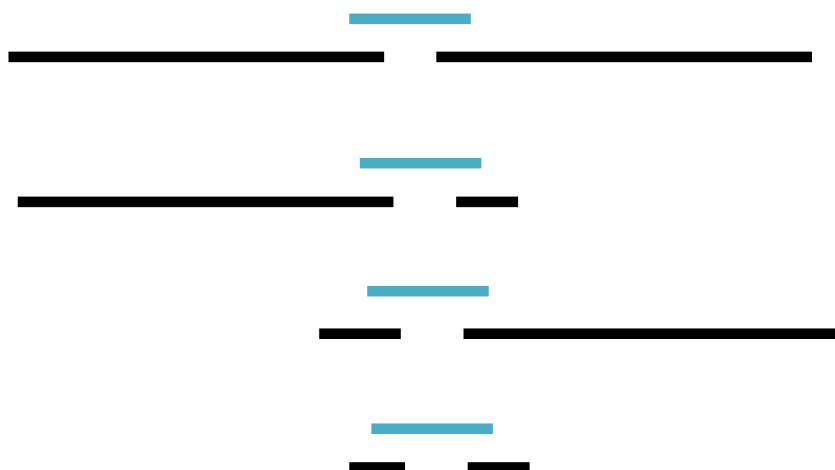
Prototype for scaffolding



1. **Simulate heterozygous polyploidy genome**
 - 10 copies with 5% of difference from original chromosome
2. **Simulate Moleculo reads from polyploidy genome**
 - Read length distribution follows exactly real molecule read distribution
3. **Simulate PacBio reads from polyploidy genome**
 - Simulate P6-C4, the latest PacBio chemistry
4. **Run Celera Assembler(CA) to assemble contigs with Moleculo reads**
5. **Run LRScf to scaffold the contigs with PacBio reads**

Preliminary Results

- **Moleculo-based contigs from CA**
 - Around 700 contigs
- **Long Read Scaffolding**
 - Align PacBio reads to all contigs
 - Find PacBio reads that link between two contigs
 - Around 1600 alignments out of 40K PacBio Reads

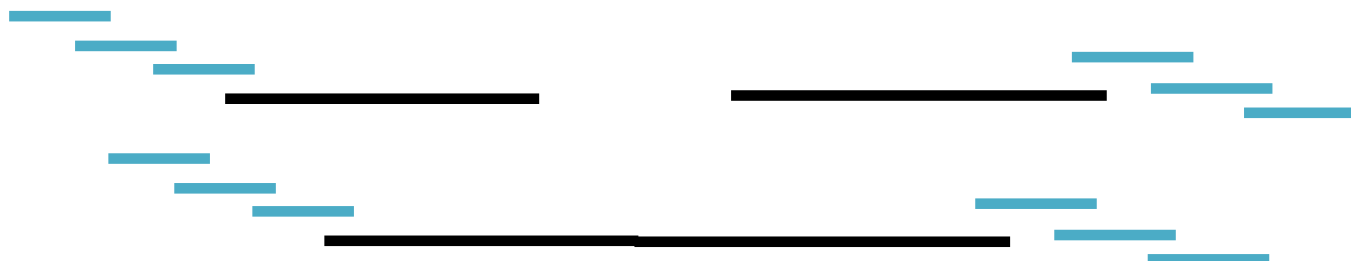


Sugarcane Scaffolding Challenges

- **How to represent aneuploidy genome?**
- **How to screen out false positive link information?**
 - # Weakly connected components 5
 - # Strongly connected components 61
 - True value $5 < 10 < 61$
- **How to assemble PacBio reads across gaps?**



- **How to extend contigs with PacBio reads?**



Contributions

- **The Resurgence of reference genome quality (3Cs)**
 - Provide the data-driven model, a.k.a. the next version of Lander-Waterman Statistics to predict contiguity of de novo genome assembly project
 - Analysis of completeness and correctness in historical human genome assembly
- **Sugarcane de novo genome assembly challenge**
 - Showed the effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploidy genome
 - NG50 contig length improved 50 times
 - The longest contig extended 25 times to half million bp
 - Pure long read de novo assembly for both contigs and scaffolding

Acknowledgements



Schatz Lab

Michael Schatz
Fritz Sedlazeck
James Gurtowski
Sri Ramakrishnan
Han fang
Maria Nattestad
Rob Aboukhalil
Tyler Garvin
Mohammad Amin
Shoshana Marcus



Shinjaee Yoo

Microsoft®

Research

Ravi Pandya
Bob Davidson
David Heckerman



University of São Paulo

Gabriel Rodrigues Alves Margarido
Jonas W. Gaiarsa
Carolina G. Lembke
Marie-Anne Van Sluys
Glaucia M. Souza



Stony Brook University

The State University of New York

McCombie Lab

Dick McCombie
Sara Goodwin



Cold Spring Harbor Laboratory



Thank You

Q & A

