Microsoft Research

Microsoft

Sugarcane Genome De Novo Assembly Challenges and Modeling

Hayan Lee

Ph.D. student in Computer Science Dept., Stony Brook University
Research Assistant, Schatz Lab, Cold Spring Harbor Laboratory
Research Intern in eScience Group, Microsoft Research

Sugarcane History

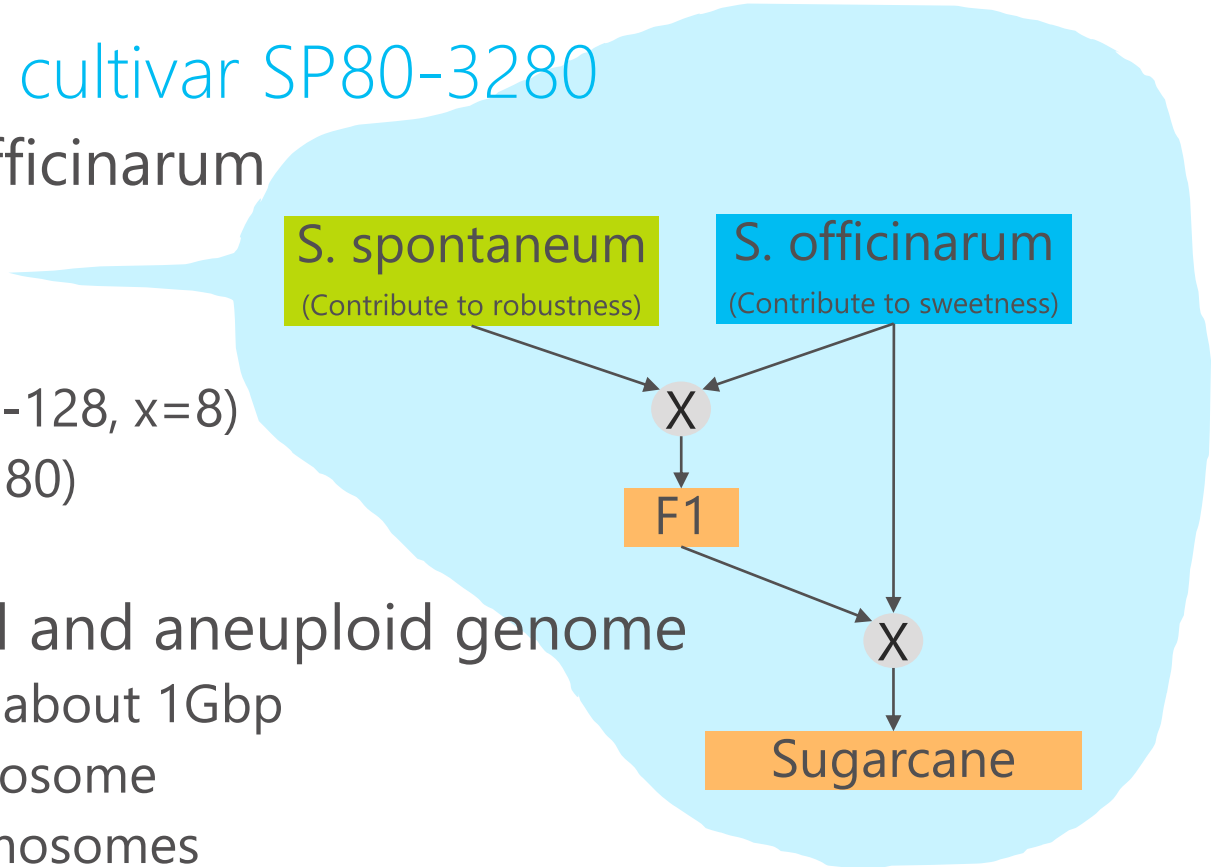
Sugarcane for food and biofuel

- 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.

http://sugarcane.org/sustainability/producing-food-and-fuel

Sugarcane

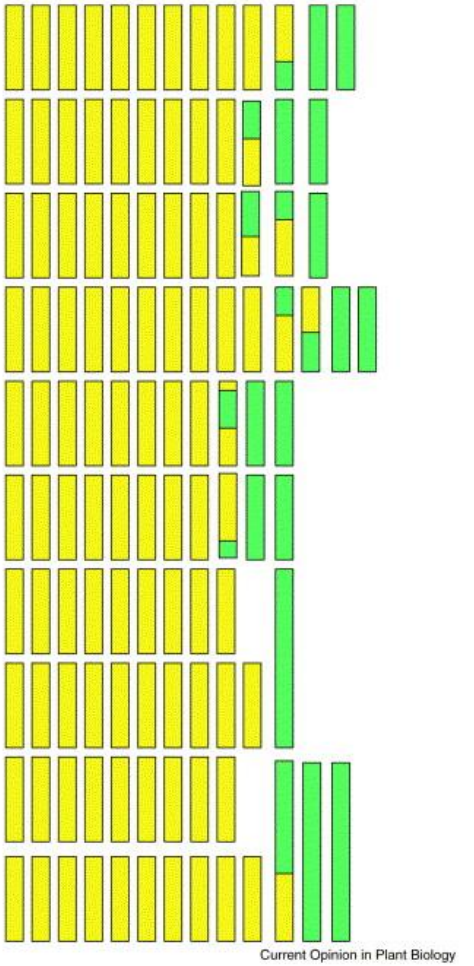
- A hybrid sugarcane cultivar SP80-3280
  - S.spontaneum x S.officinarum
  - A century ago….
  - Saccharum genus
    - S. spontaneum (2n=40-128, x=8)
    - S. officinarum (2n=8x=80)
  - Big, highly polyploid and aneuploid genome
    - Monoploid genome is about 1Gbp
    - 8-12 copies per chromosome
    - In total, 100-130 chromosomes
    - Total size is about 10Gbp




Sugarcane Challenges

Why is sugarcane assembly harder? (1)

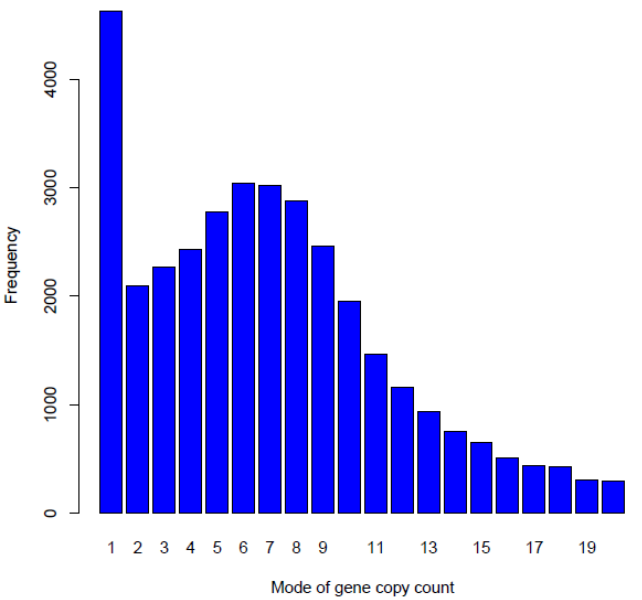
- Polyploidy/Aneuploidy
  - 10% of the chromosomes are inherited in their entirety from *S. spontaneum*, 80% are inherited entirely from *S. officinarum*
- Large scale recombination
  - 10% is the result of recombination between chromosomes from the two ancestral species, a few being double recombinants



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

Why is sugarcane assembly hard? (2)

- Heterozygosity
  - The most heterozygous region has 5% of differences
- Repeats
  - Polyploidy will boost repeats across copies of chromosomes
  - Haploid genome has many repeats
  - Polyploidy causes even more copies



Four Important Questions in Sugarcane

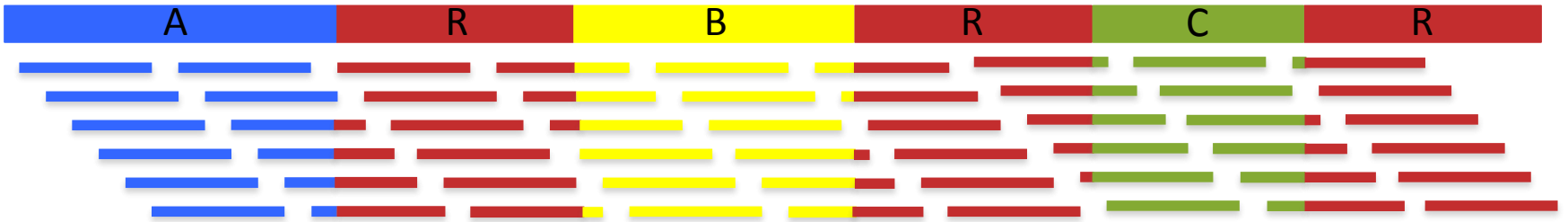
- Scaffold polyploidy/aneuploidy genome
  - How do we connect contigs/cluster contigs per chromosome/fill gaps among contigs?
- Phasing haplotypes
  - Not solved in diploid genome yet
- Heterozygosity
  - How do we measure heterozygosity in polyploidy/aneuploidy genome?
  - How do we quantify alleles 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?

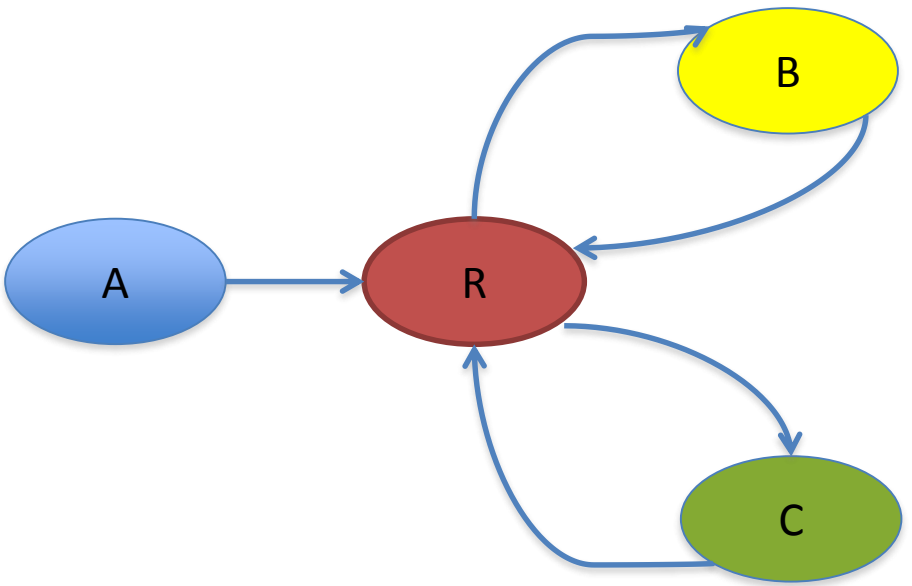
Gabriel Margarido et al. "*ConPADE: Genome assembly ploidy estimation from next-generation sequencing data*", *under review*

Assembly

Data and Tools

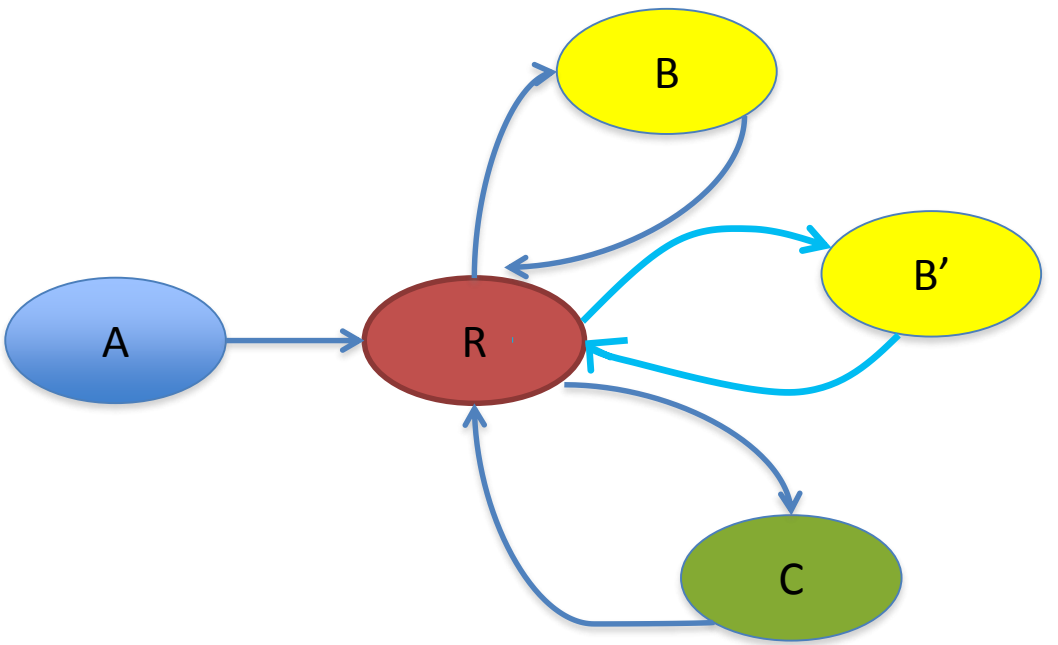
Assembly Complexity by Repeats



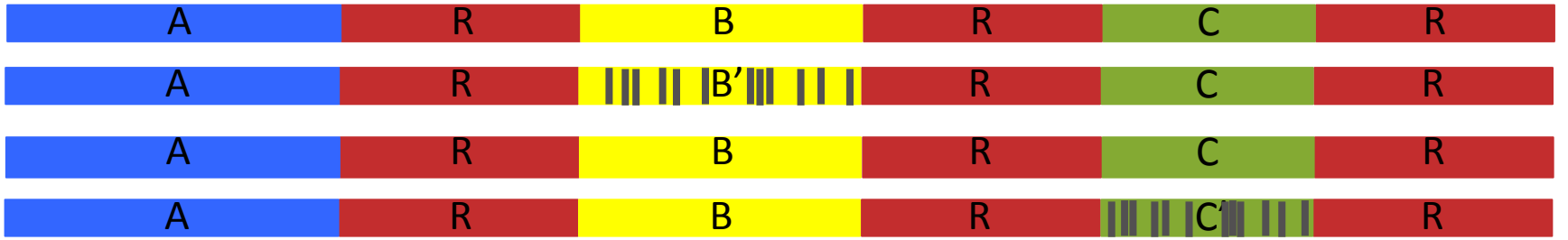


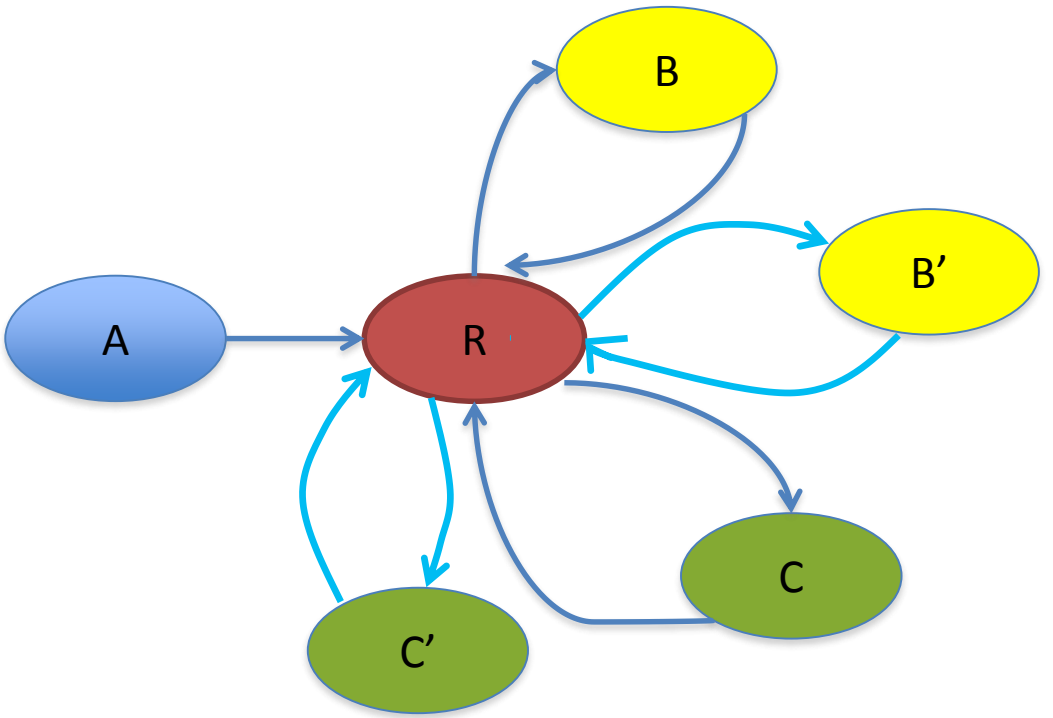
Long Reads is the solution!!!

Assembly Complexity by Heterozygosity



Assembly Complexity by Polyploidy





Long Reads is the solution!!!

Moleculo Reads

(1) The DNA is sheared into fragments of about 10Kbp

(2) Sheared fragments are then diluted

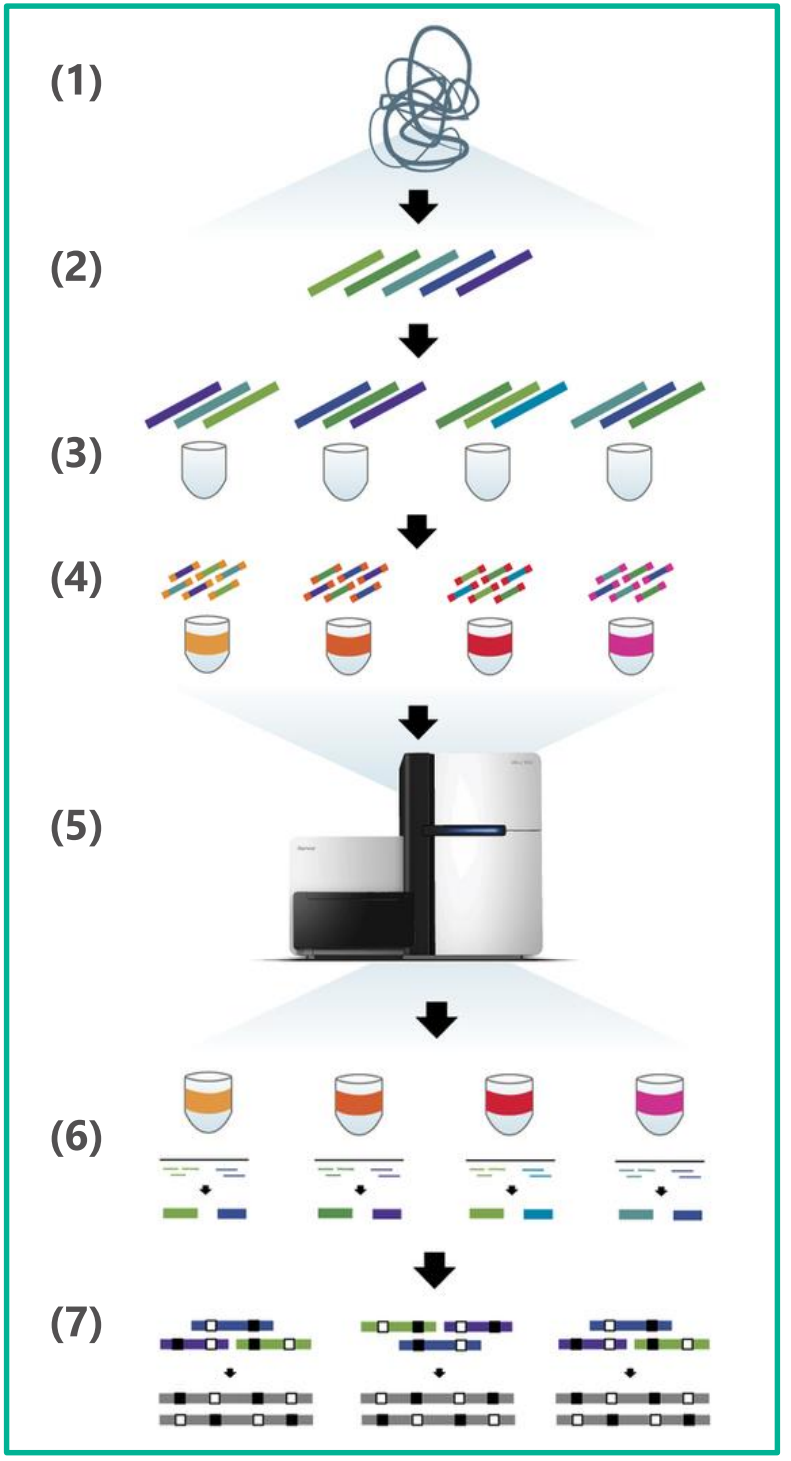
(3) and placed into 384 wells, at about 3,000 fragments per well.

(4) Within each well, fragments are amplified through long-range PCR, cut into short fragments and barcoded

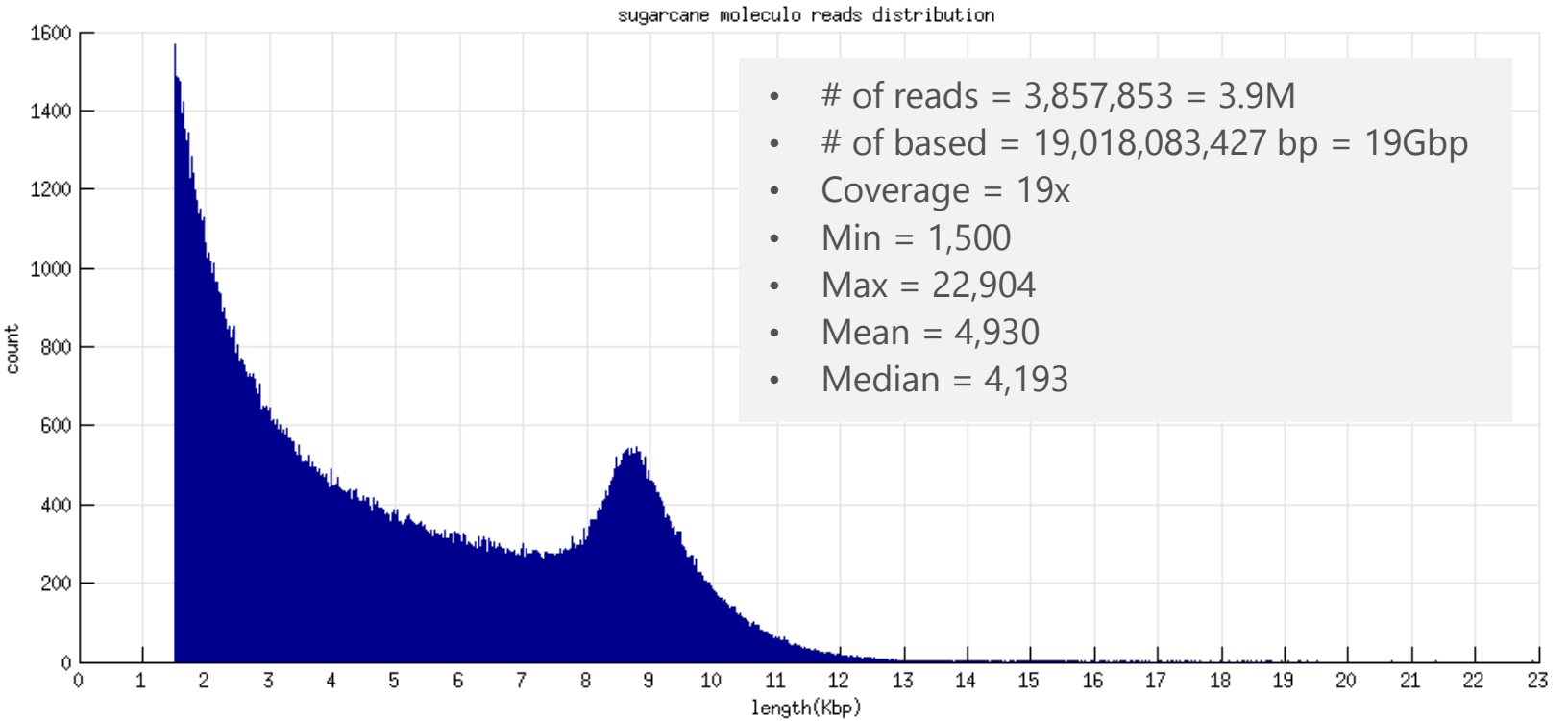
(5) before finally being pooled together and sequenced.

(6) Sequenced short reads are aligned and mapped back to their original well using the barcode adapters.

(7) Within each well, reads are grouped into fragments, which are assembled to long reads.



Moleculo Reads



- # of reads = 3,857,853 = 3.9M
- # of based = 19,018,083,427 bp = 19Gbp
- Coverage = 19x
- Min = 1,500
- Max = 22,904
- Mean = 4,930
- Median = 4,193

Choose the right data and the right tool

DATA	Hiseq 2000 PE (2x100bp) - 575Gbp - **600x** of monoploid genome Roche454 - 9x of monoploid genome - [min=20 max=1,168] - Mean=332bp	Moleculo - 19Gbp - **19x** of monoploid genome - [min=1,500 max=22,904] - Mean = 4,930bp
SOFTWARE	SOAPdenovo	Celera Assembler
RESULT	Max contig = **21,564** bp NG50=**823** bp Coverage=**0.86x**	Max contig = **467,567** bp NG50=**41,394** bp Coverage=**3.59x**

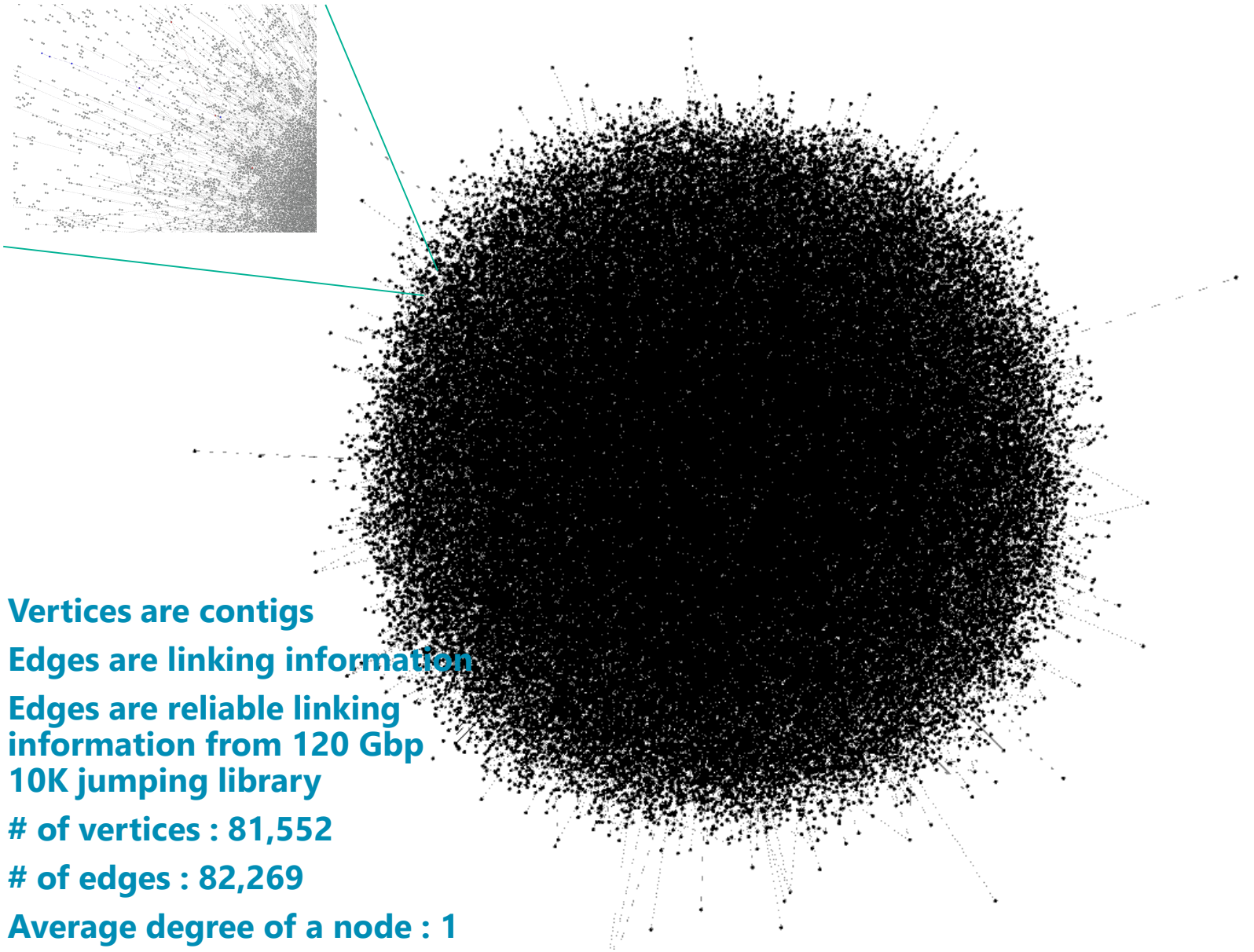
CEGMA

- 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

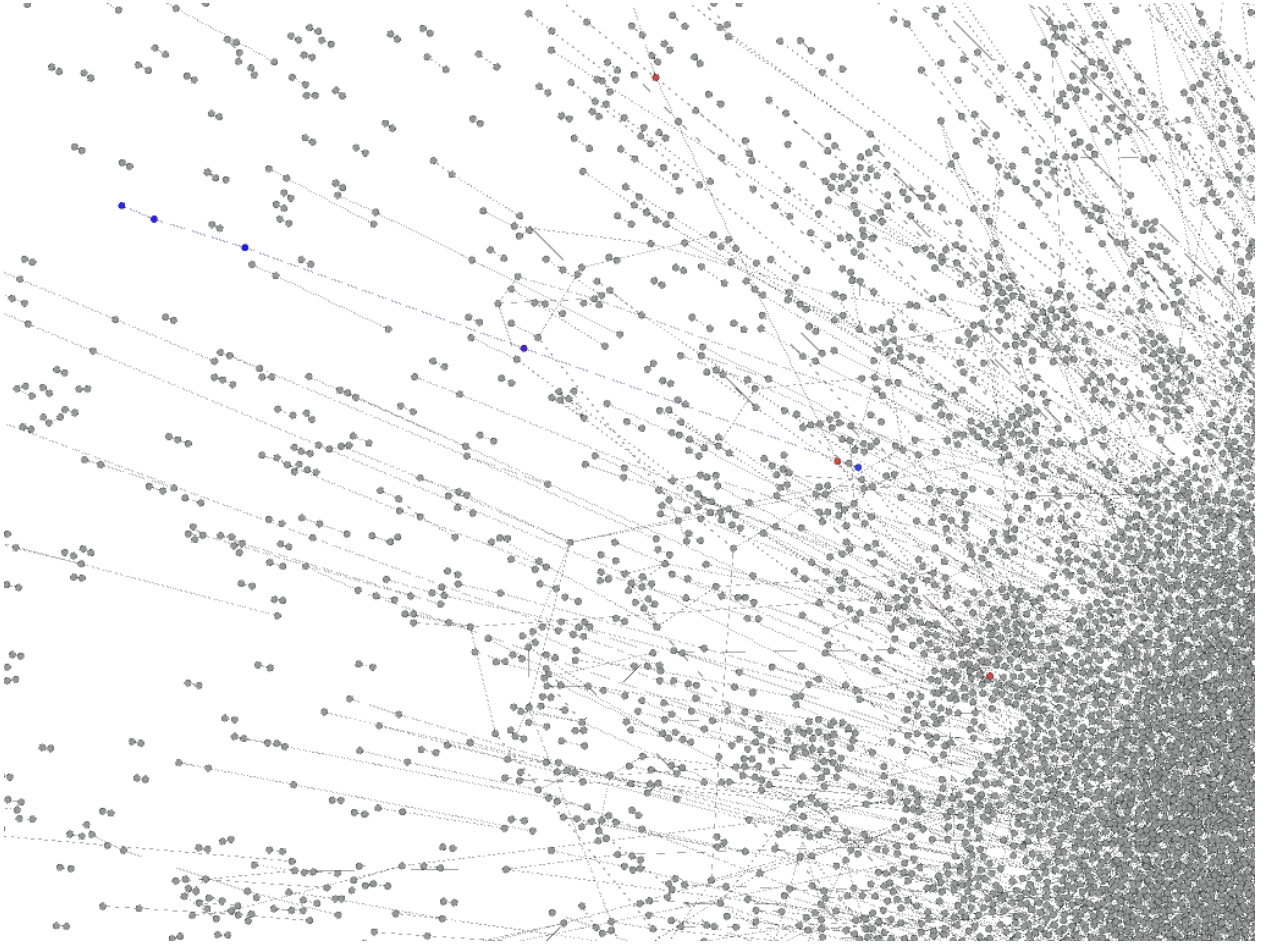
- Gene prediction aided by sorghum gene model
  - In progess…
  - 39k sorghum genes were found in sugarcane contigs at least partially

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**

NP-Hard Hairball of Sugarcane



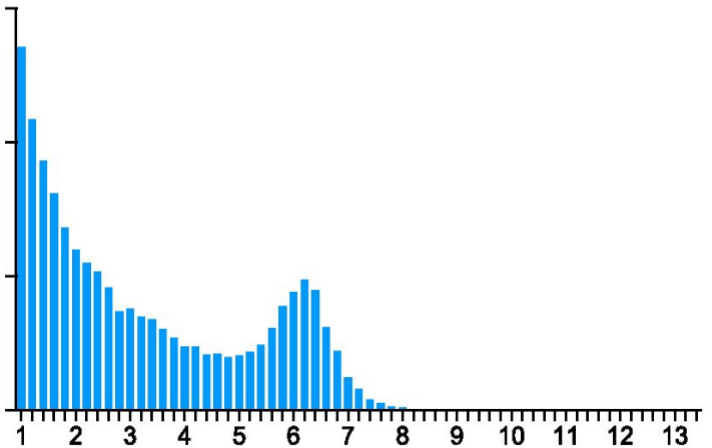
- **# of connected components = 17,919**
- **Average number of vertices per CC= 2.54**
- **The biggest CC has 25 vertices**

The Future of Long Read Sequencing

Modeling genome assembly performance

Long Read Sequencing Technology

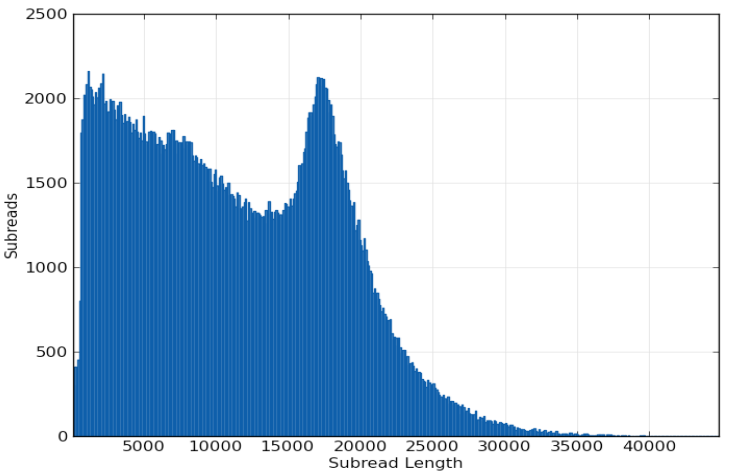
Moleculo



(Voskoboynik et al. 2013)

- Mean is around 5Kbp
- Very accurate (< 0.1% of error rate)

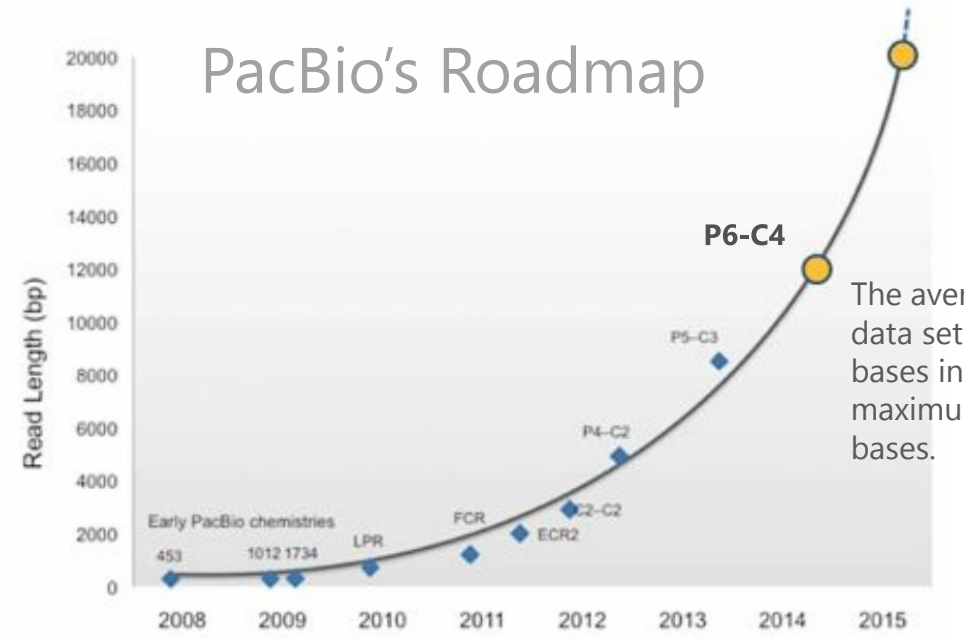
PacBio RS II



PacBio (2014)

- Mean is over 14Kbp
- High error rate 10-15%, but can be corrected down to 1% by short reads or contigs

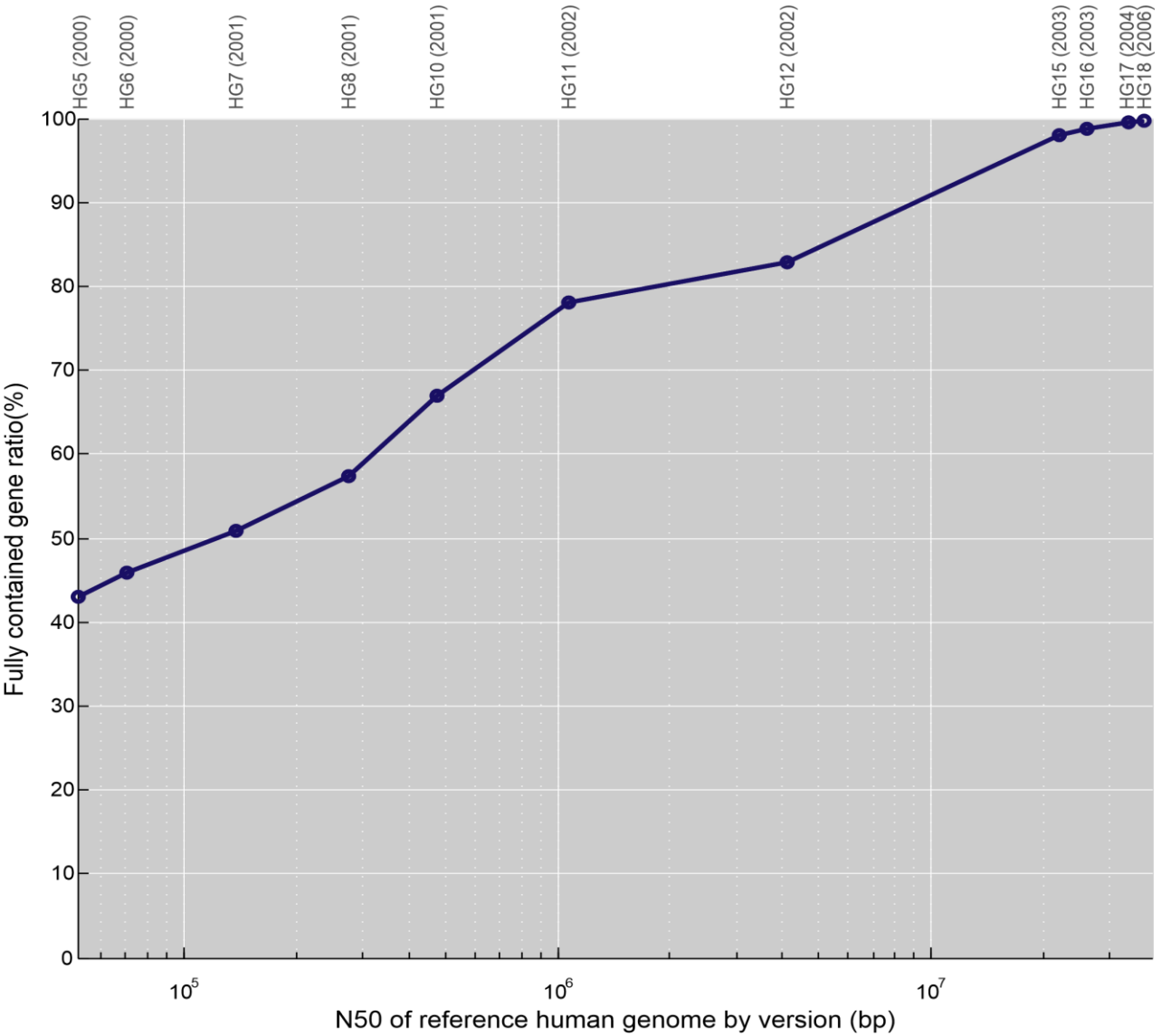
Benefits of Long Reads



The average read length of the raw data set is >14 kb, with half of the bases in reads > 21 kb and the maximum read length of 64,500 bases.

- Read Length is increasing
- Very informative whether it has high error rate or not
- More repeats resolved
- Assembly graph will be simpler
- We can get longer contigs without scaffolding
- Better scaffolding solution than long jumping library
- Overall assembly quality will improve

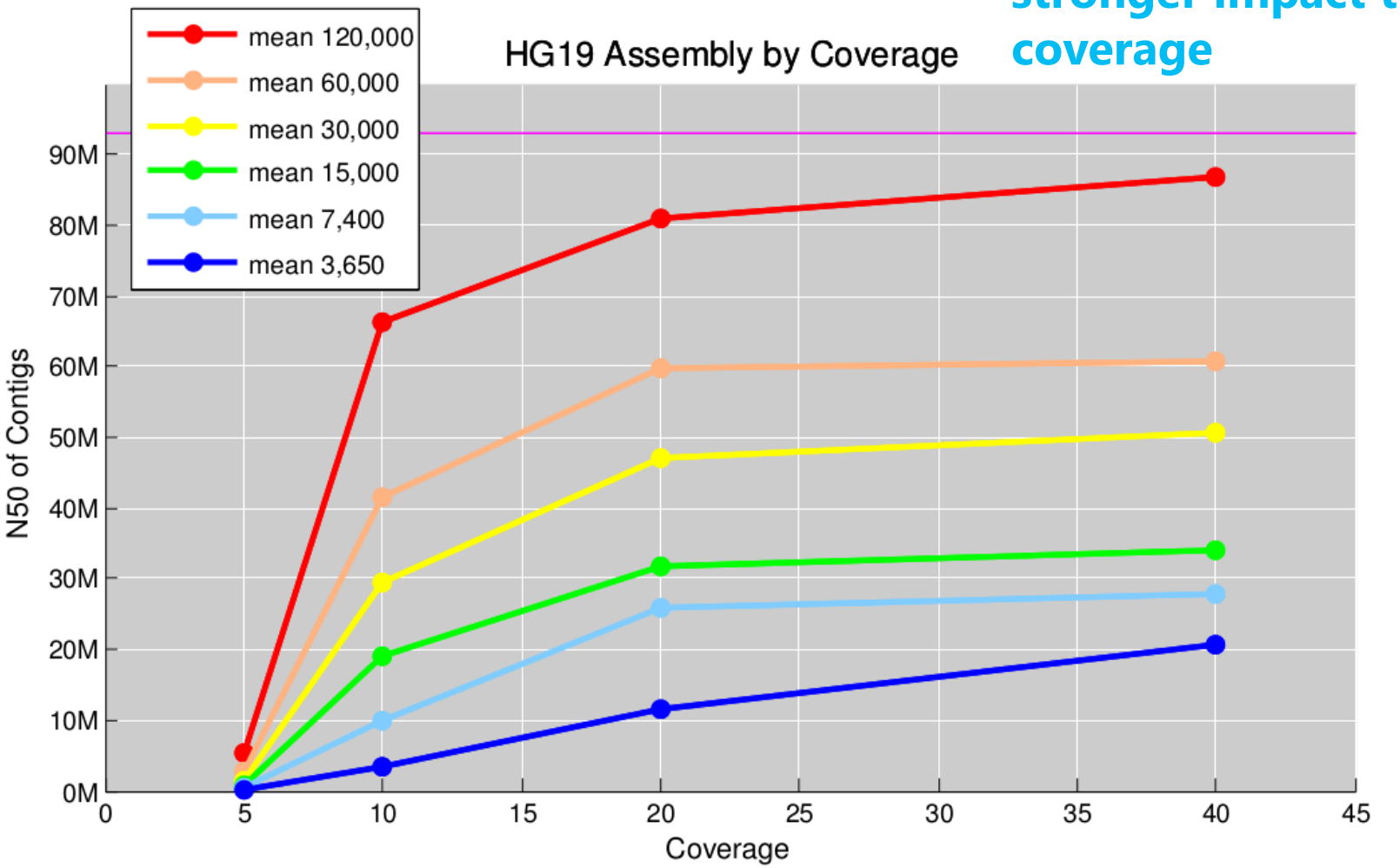
Human Reference Genome Quality



Hayan Lee et al. *"How long is long enough?"*, (in preparation)

HG19 Genome Assembly Performance by Our Simulation

Read Length has stronger impact than coverage



Modeling Genome Assembly Performance

- To predict genome assembly performance

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

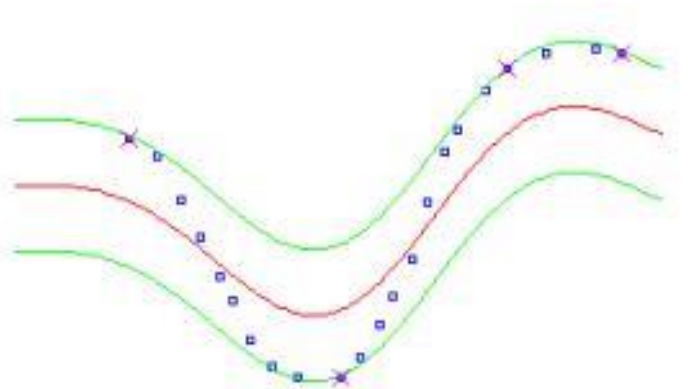
- Using four features

$$\text{Performance}(\%) \approx f\begin{pmatrix} Read\ Length \\ Coverage \\ Repeats \\ Genome\ Size \end{pmatrix}$$

Support Vector Regression (SVR)

- Epsilon insensitive loss function

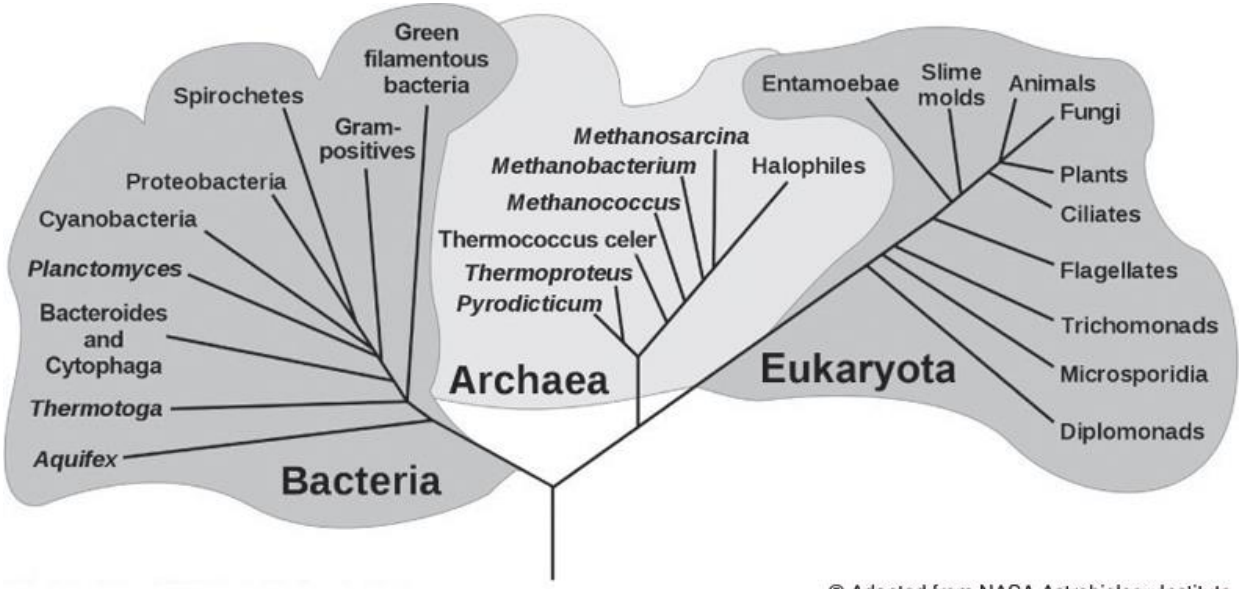
$$L(y, f(x,w)) = \begin{cases} 0\,, & if\ |y - f(x,w)| \le \varepsilon \\ |y - f(x,w)| - \varepsilon, & otherwise \end{cases}$$



- Benefits
  - Simplest fit
  - Robust to outliers

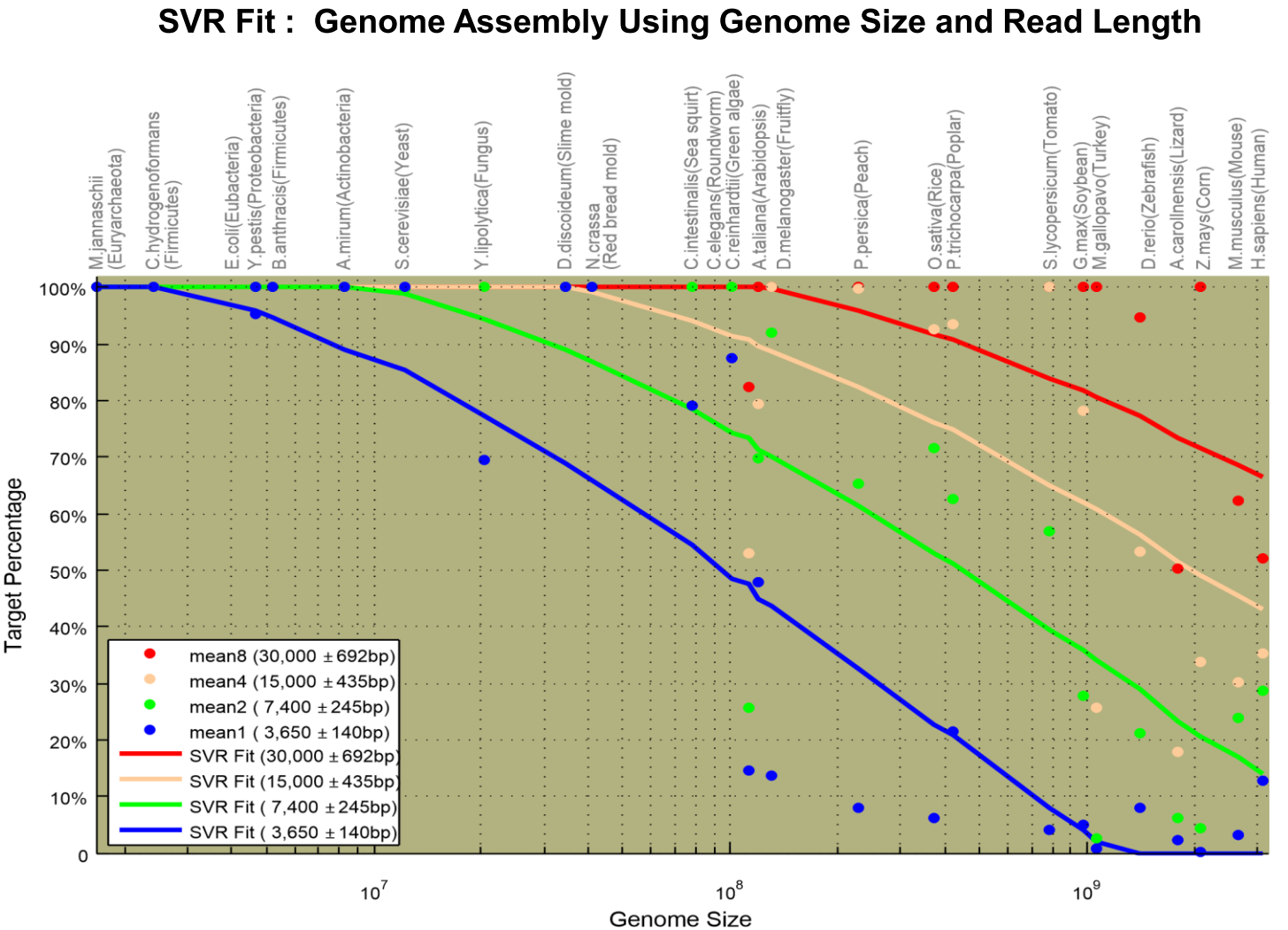
26 Species Across Tree of Life

Model Organism	ID	Genome Size
M.jannaschii	1	1,664,970
C.hydrogenoformans	2	2,401,520
E.coli	3	4,639,675
Y.pestis	4	4,653,728
B.anthracis	5	5,227,293
A.mirum	6	8,248,144
yeast	7	12,157,105
Y.lipolytica	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



© Adapted from NASA Astrobiology Institute

Reference Genome Quality



Hayan Lee et al. *"How long is long enough?"*, (in preparation)

Web Service

Genome Assembly Performance Prediction - Mozilla Firefox

Http://qb.cshl.edu/asm_model/predict.html

Genome Assembly Performance Prediction

This is the Genome Assembly Performance Prediction Service. If you have any queries please email Hayan Lee(hlee@cshl.edu).

Although assembly performance is a function of genome size, read length, coverage and repeats, in this prediction model, we only used 3 features; genome size, read length and coverage for the simplicity.

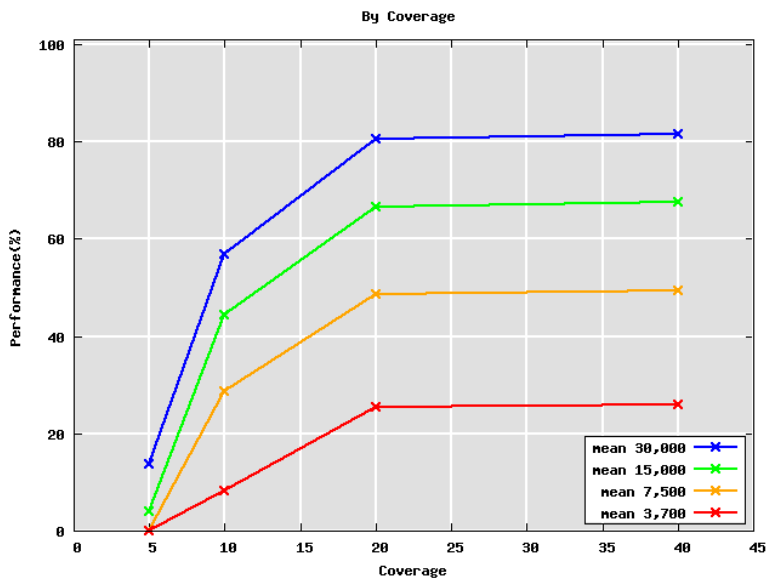
Given genome size, we internally set read lengths and coverages for you. With 3 features, our model predicts the expected performance of assembly. Performance is defined as follows:

Performance(%) = N50 of assembly / N50 of chromosome segments

Genome size : 1000123456

Submit

Assembly Prediction of Genome Size 1000123456



Hayan Lee et al. *"How long is long enough?"*, (in preparation)

Contributions

- Sugarcane de novo genome assembly
  - N50 contig improved 50 times
  - The longest contig extended to half million bp
- Modeling
  - We made a new model that predicts genome assembly performance in 15% of residue boundary.
- Recommendations for de novo genome assembly
  - Use the coverage >20x at least for perfect reads but note that more and more coverage does not guarantee that your N50/contigs get longer and longer
  - Use the longest reads possible
- Machine Learning & Big Data
  - Contact <hlee@cshl.edu>

Acknowledgement

Microsoft Research

Ravi Pandya
Bob Davidson
David Heckerman

Universidade de São Paulo

Gabriel Margarido
Marie-Anne Van Sluys
Glaucia Souza

Cold Spring Harbor Laboratory

Michael Schatz

Brookhaven National Laboratory

Shinjae Yoo