Microsoft Research

Sugarcane Genome De Novo Assembly Challenges

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

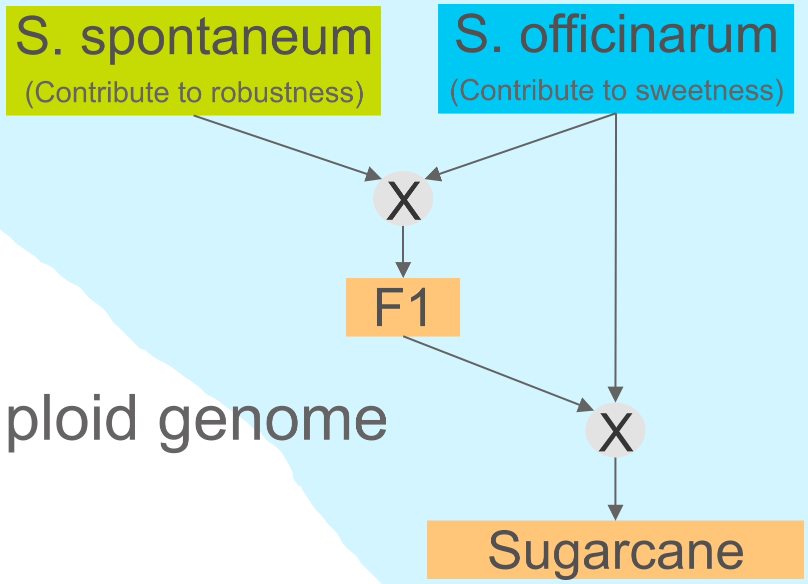
Acknowledgement

Ravi Pandya, Microsoft Research, Redmond, WA
Gabriel Rodrigues Alves Margarido , USP / ESALQ, Piracicaba, SP, Brazil
Bob Davidson , Microsoft Research, Redmond, WA
Jonas W. Gaiarsa , GaTE Lab - University of São Paulo, Sao Paolo, Brazil
Carolina G. Lembke , Institute of Chemistry - University of São Paulo, São Paulo, Brazil
Michael Schatz , Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
Marie-Anne Van Sluys , Universidade de São Paulo, Sao Paulo, Brazil
Glaucia M. Souza , Institute of Chemistry - University of São Paulo, São Paulo, Brazil
David Heckerman , Microsoft, Los Angeles, CA

Sugarcane Challenges

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



S. spontaneum

(Contribute to robustness)

S. officinarum

(Contribute to sweetness)

X

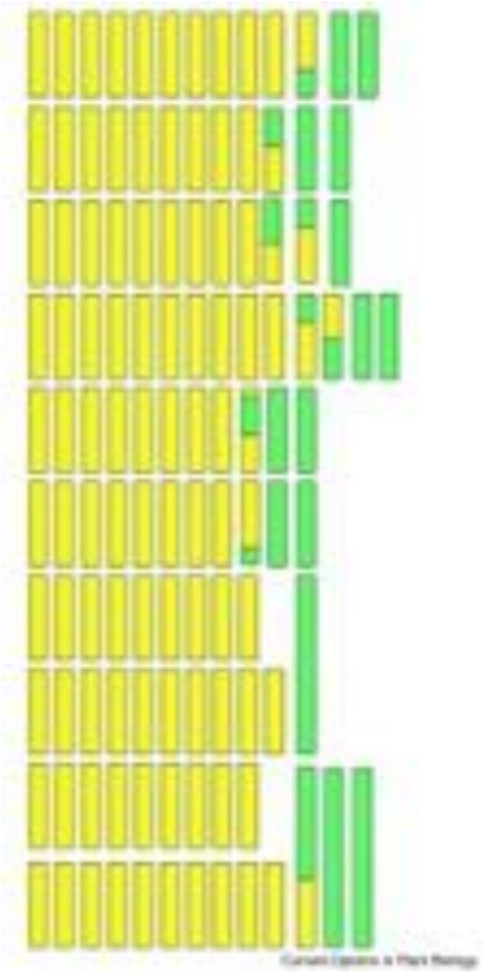
F1

X

Sugarcane

Why is sugarcane assembly harder? (1)

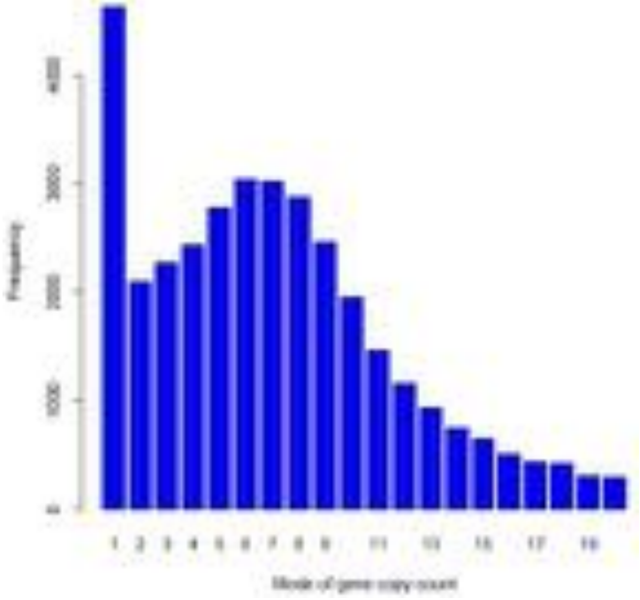
- Polyploidy/Aneuploidy
  - 10% of the chromosomes are inherited entirely from *S. spontaneum*, 80% are inherited entirely from *S. officinarum*
- Large scale recombination
  - 10% is the result of recombination between chromosomes from the two ancestral species, a few being double recombinants



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

Why is sugarcane assembly harder? (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

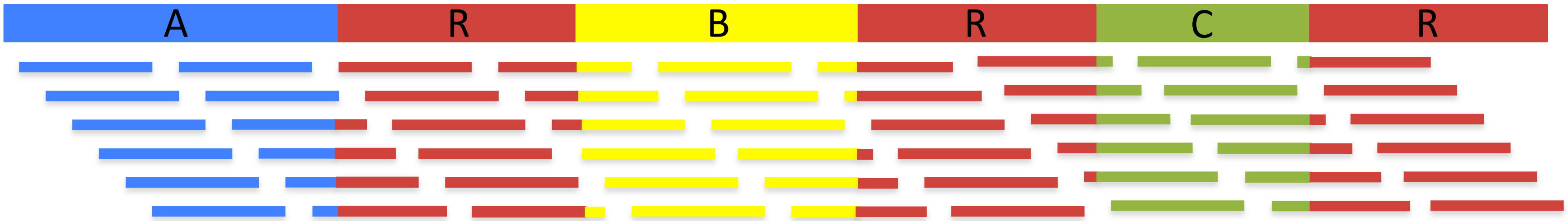
- Scaffold polyploid/aneuploid genome
  - How do we connect contigs/cluster contigs per chromosome/ fill gaps among contigs?
- Phasing haplotypes
  - Not solved in diploid genome yet
- Heterozygosity
  - How do we define/measure heterozygosity in polyploid/ aneuploid genome?
  - How do we quantify alleles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

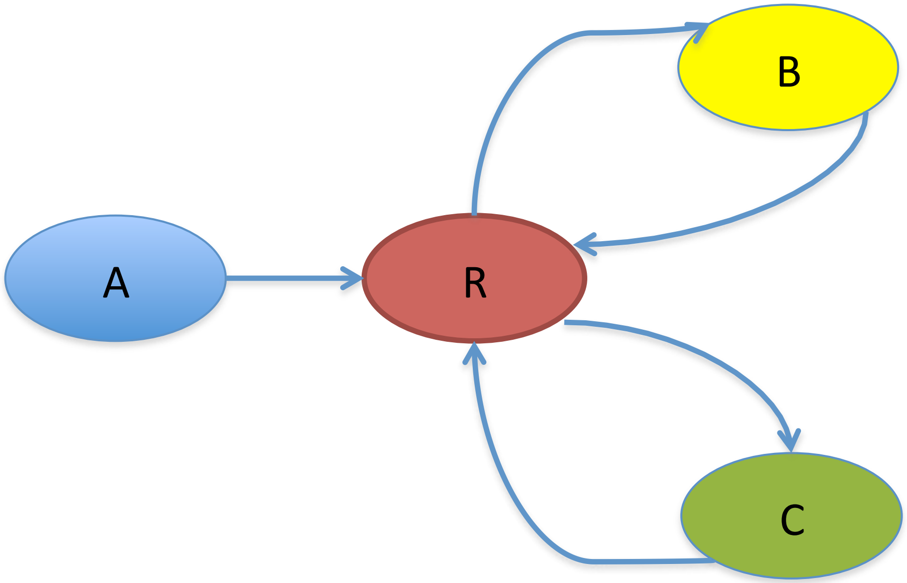
Gabriel Margarido et al. “*ConPADE: Genome assembly ploidy estimation from next-generation sequencing data*”, *under review*

Assembly Strategy

Data and Algorithms

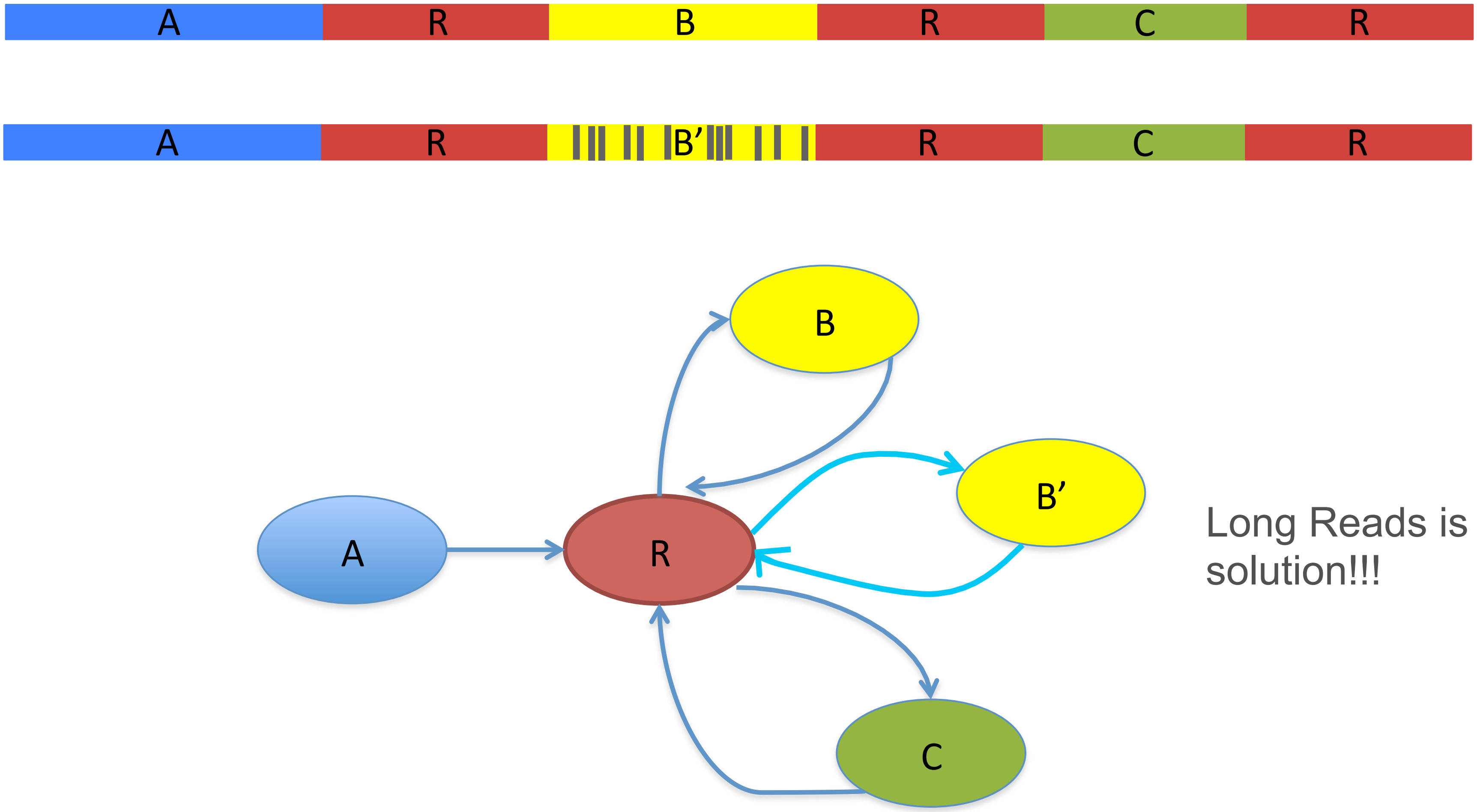
Assembly Complexity by Repeats





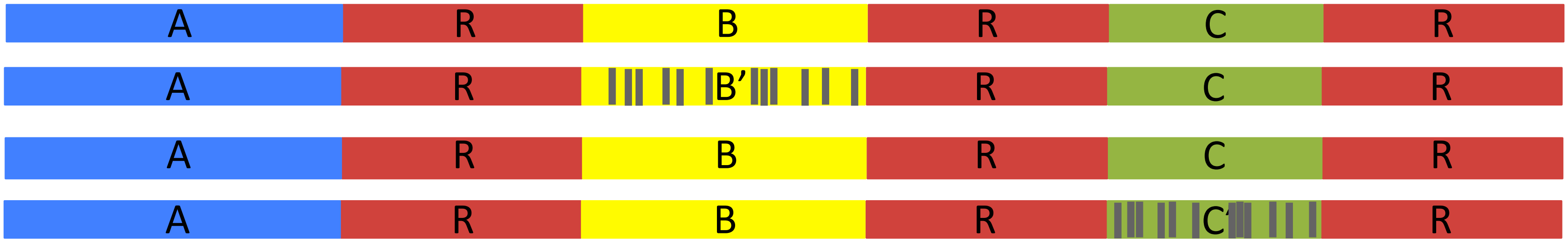
Long Reads is the solution!!!

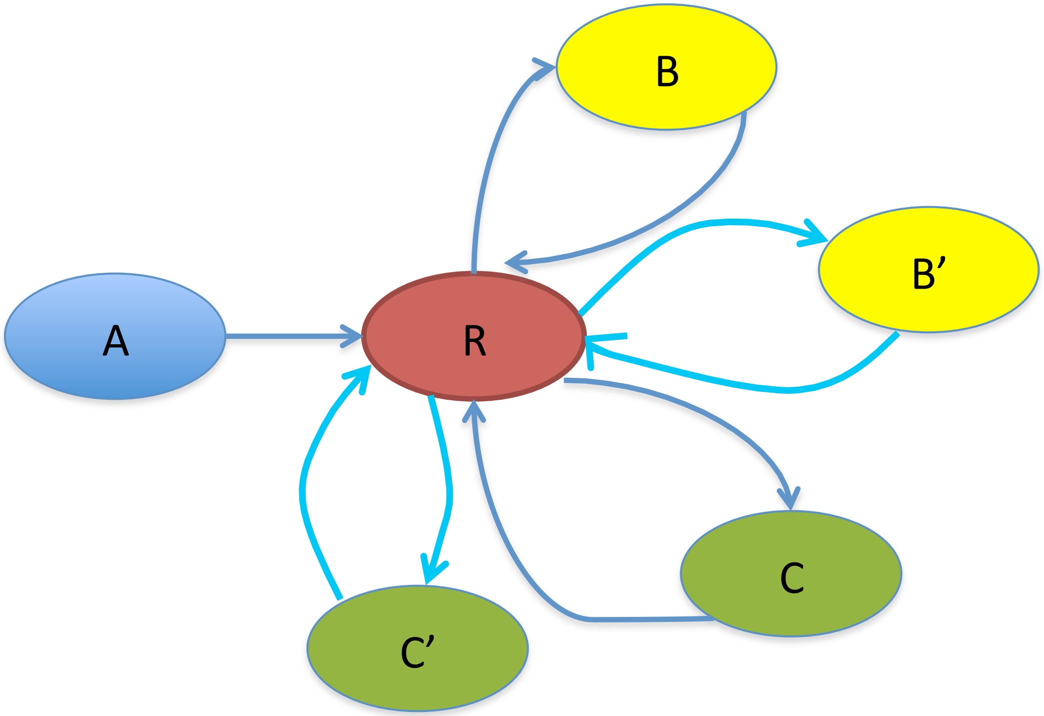
Assembly Complexity by Heterozygosity



Long Reads is the solution!!!

Assembly Complexity by Polyploidy





Choose the right data and the right method

DATA	Hiseq 2000 PE (2x100bp) - 575Gbp - **600x** of monoploid genome Roche454 - 9x of monoploid genome - [min=20 max=1,168] - Mean=332bp
Algorithm	SOAPdenovo (De Bruijn Graph)

RESULT	Max contig = **21,564** bp
	NG50=**823** bp
	Coverage=**0.86x**

Moleculo Reads

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

(2) Sheared fragments are then diluted

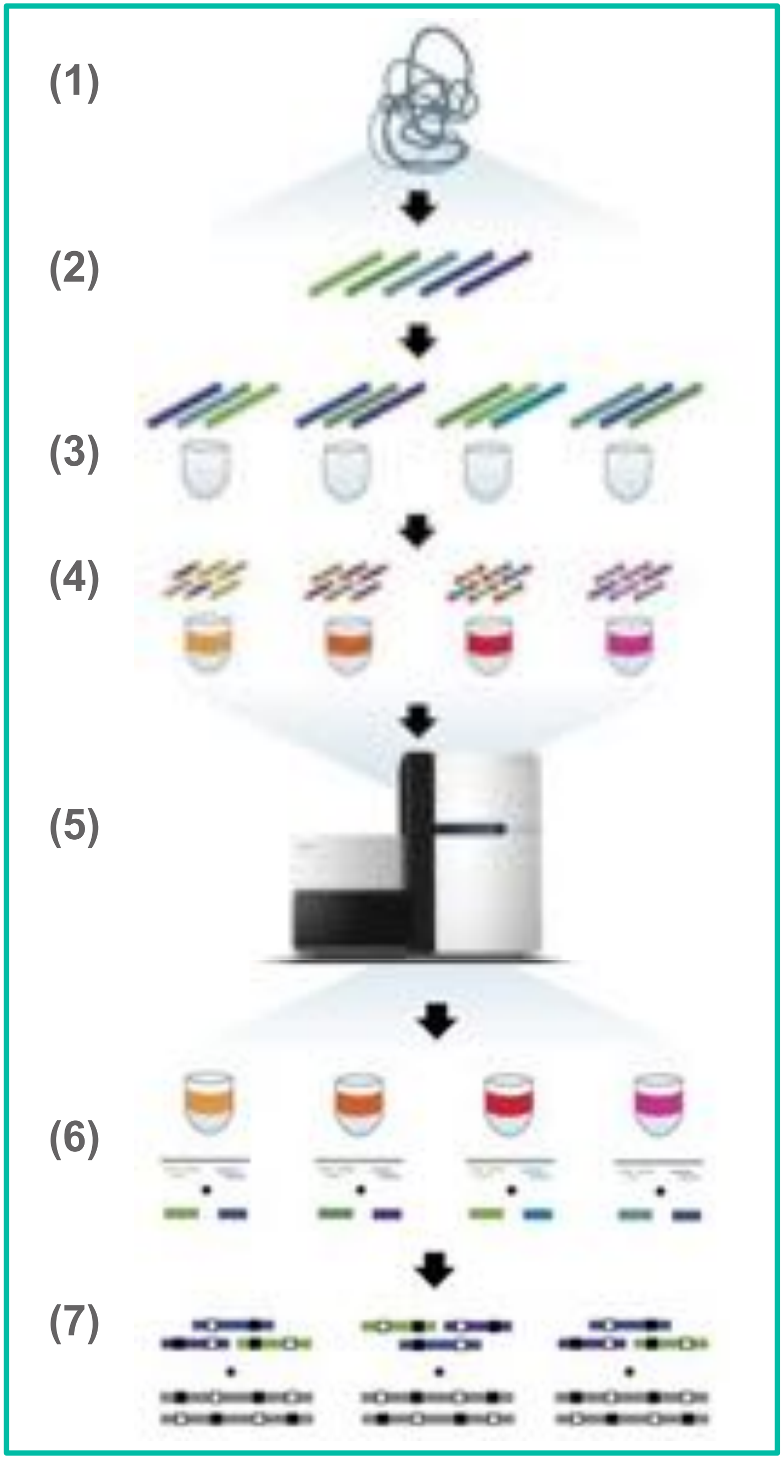
(3) and placed into 384 wells, at about 3,000 fragments per well.

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

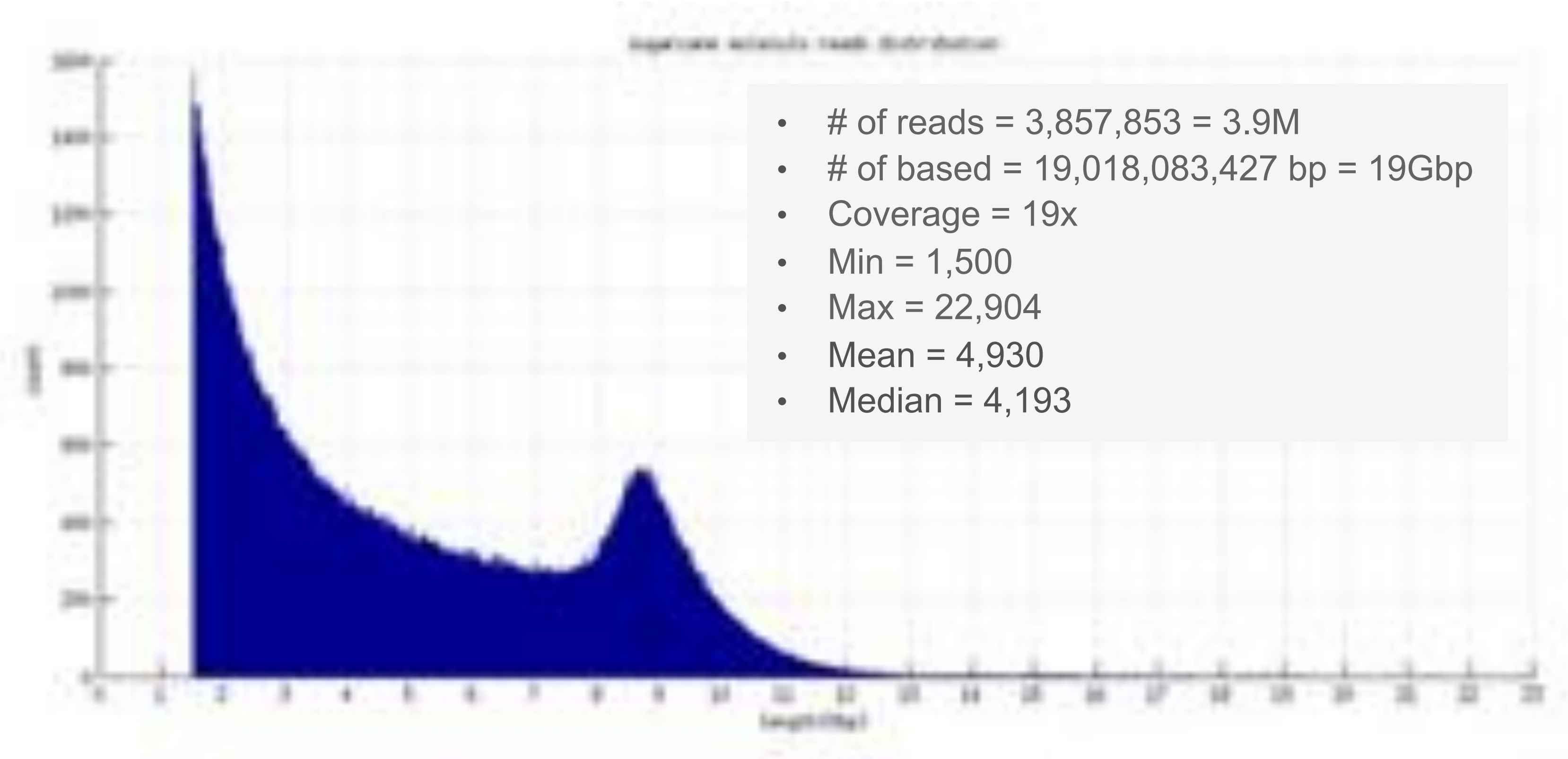
(5) before finally being pooled together and sequenced.

(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 method

DATA	Hiseq 2000 PE (2x100bp) - 575Gbp - **600x** of haploid genome Roche454 - 9x of haploid genome - [min=20 max=1,168] - Mean=332bp	Moleculo - 19Gbp - **19x** of haploid genome - [min=1,500 max=22,904] - Mean = 4,930bp
Algorithm	SOAPdenovo (De Bruijn Graph)	Celera Assembler (Overlap Graph)

RESULT	Max contig = **21,564** bp
	NG50=**823** bp
	Coverage=**0.86x**
		Max contig = **467,567** bp
		NG50=**41,394** bp
		Coverage=**3.59x**
		# of contigs = **450K**

De Bruijn vs. Overlap Graph

	De Bruijn Graph	**Overlap Graph** (Overlap-Layout-Consensus)
Unit	K-mer	Read
Information	Edge	Node
Algorithm	Eulerian path Visit every edge exactly once	Hamiltonian path Visit each node exactly once
Complexity	P (easy)	NP-Hard (hard)
Performance in reality	- Many Eulerian paths are possible. - Very sensitive to repeats - Very sensitive to errors - Uneven coverage - Performance is limited to k in k-mer	- Overlap and consensus are time and/or memory intensive jobs. + Repeats can be overcome by long reads + Performance depends on reads length
	Better choice for short reads	**Better choice for long reads**

CEGMA

- CEGs
  - Korf Lab in UC. Davis identified 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
  - 370K genes (duplicated counting possible)

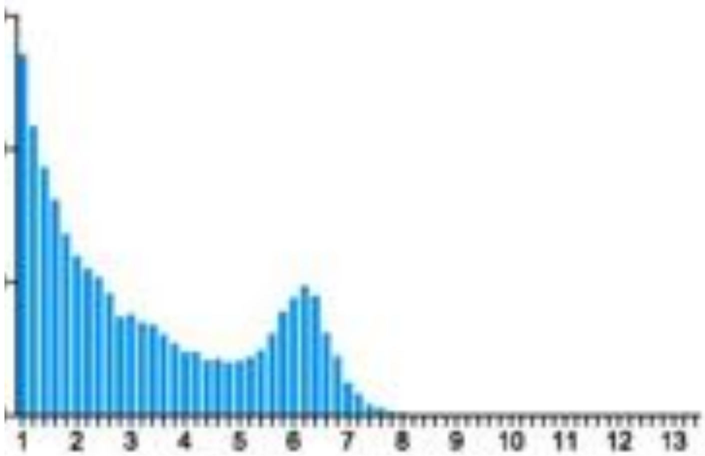
Next Steps

Extra Long Read

Scaffolding

Long Read Sequencing Technology

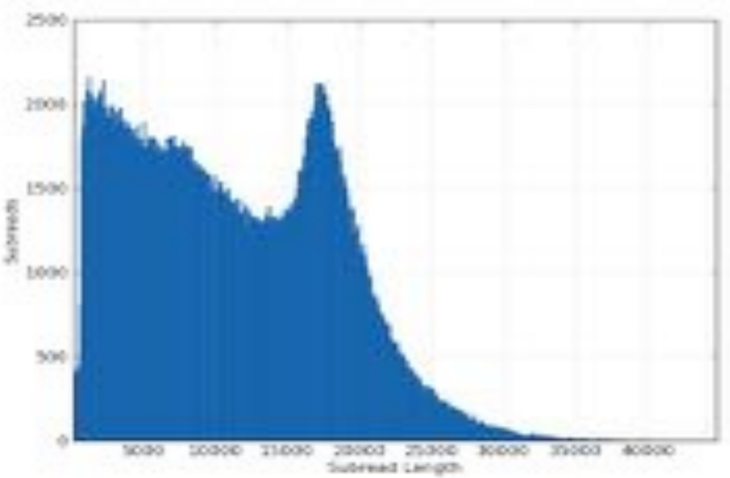
Moleculo



(Voskoboynik et al. 2013)

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

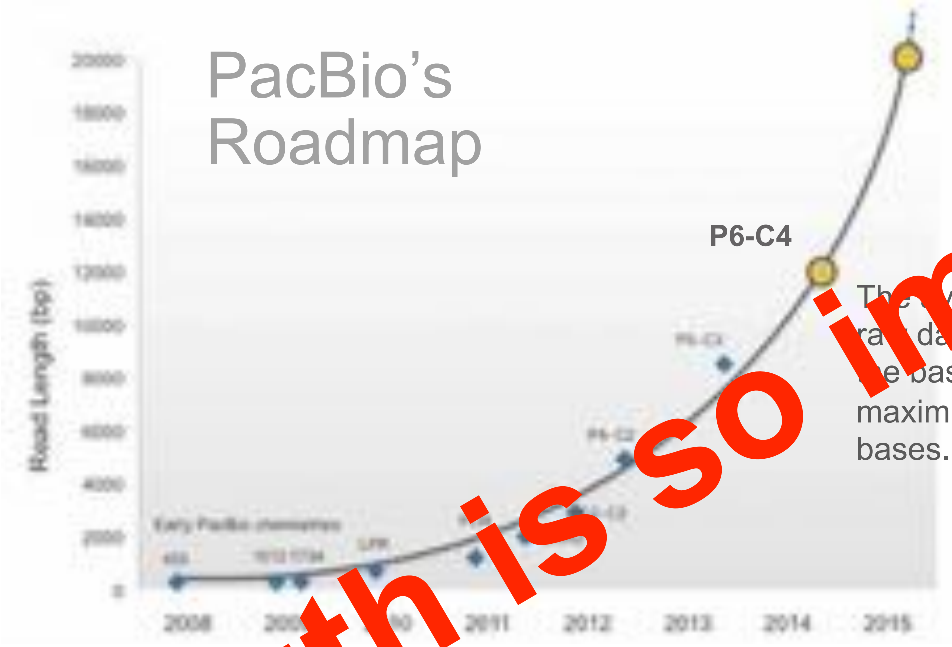
PacBio RS II



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

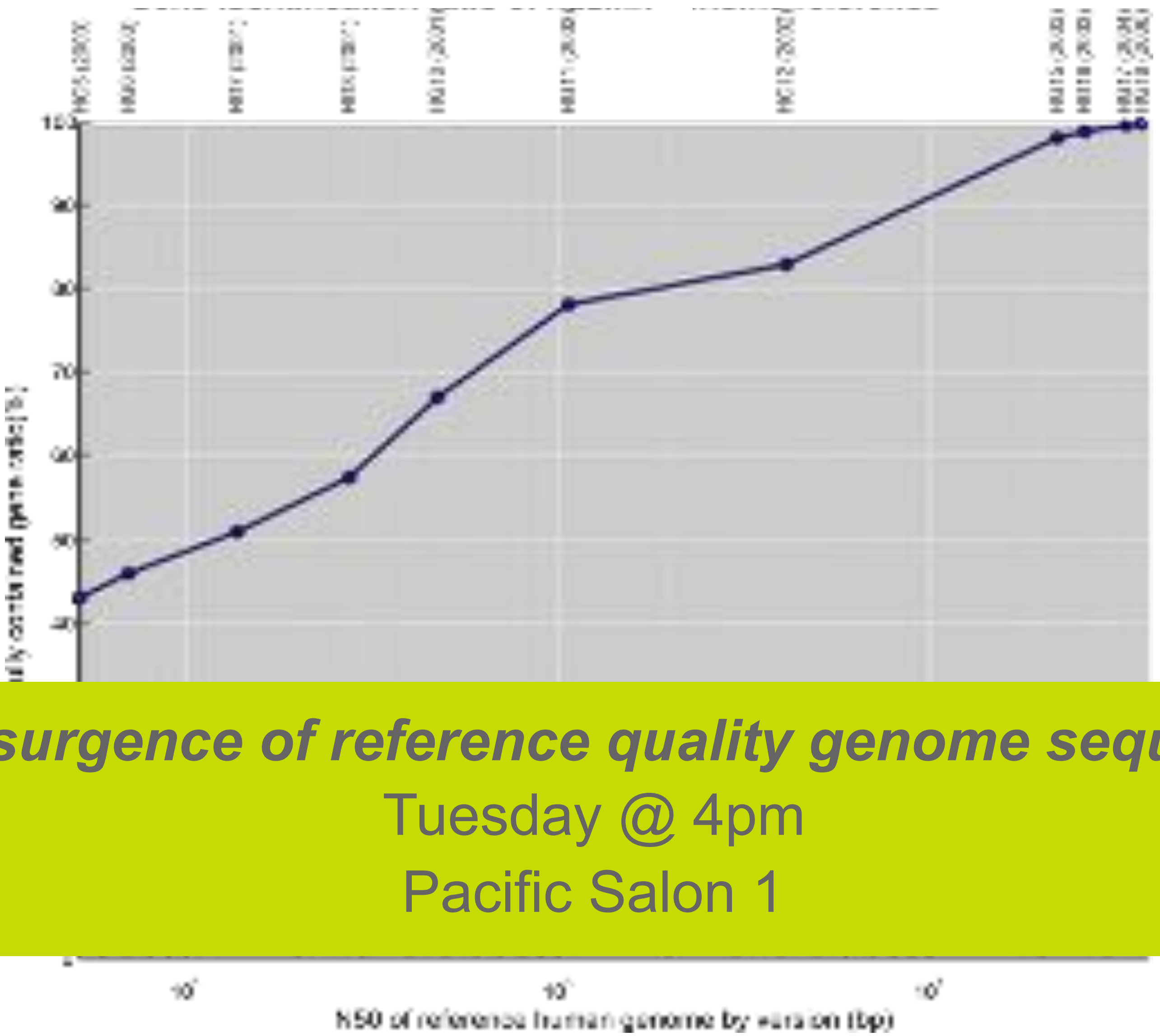


The average read length of the raw data set is >14 kb, with half of the bases in reads > 21 kb and the maximum read length of 64,500 bases.

- Read Length is increasing (PacBio)
- 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 (PacBio)
  - Cost
  - Accuracy
- Overall assembly quality will improve

Read length is so important

Human Reference Genome Quality



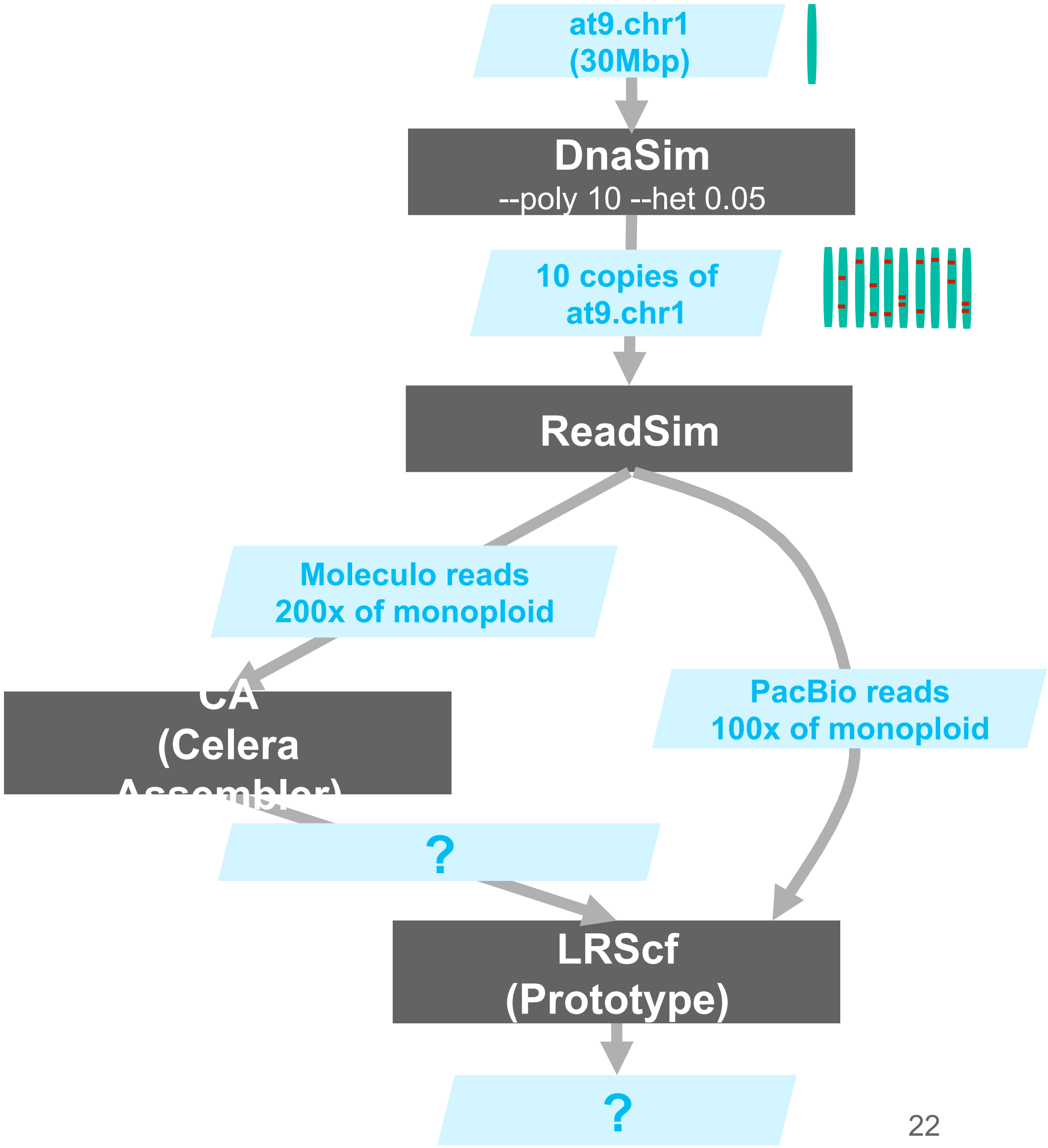
The resurgence of reference quality genome sequencing

Tuesday @ 4pm

Pacific Salon 1

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

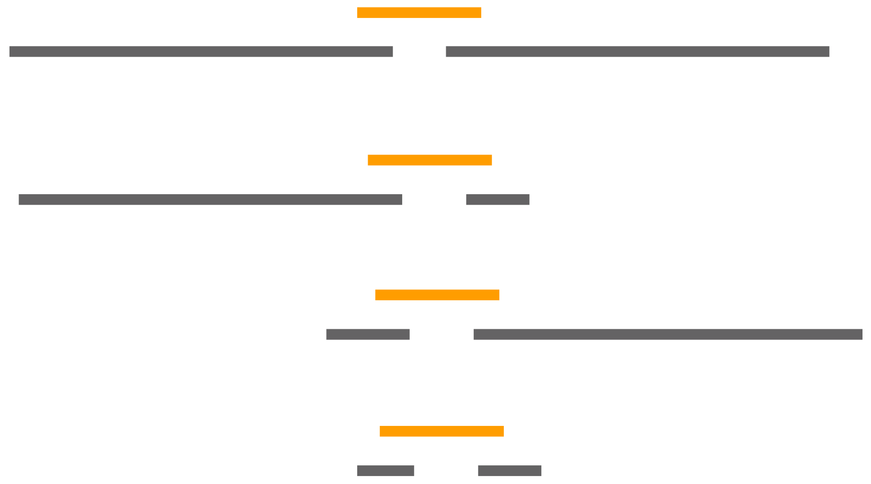
Prototype for scaffolding



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 lastest PacBio chemistry
4. Run Celera Assembler(CA) to assemble contigs with Moleculo reads
5. Run LRScf to scaffold the contigs with PacBio reads

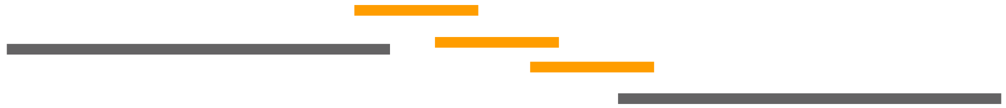
Preliminary Results

- 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 signals out of 40K PacBio Reads

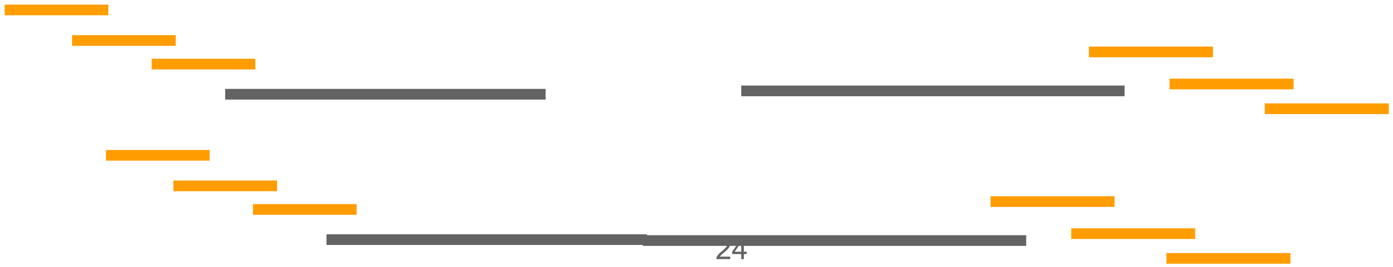


Sugarcane Scaffolding Challenges

- 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 and Recommendations

- Sugarcane de novo genome assembly
  - NG50 contig length improved 50 times
  - The longest contig extended 25 times to half million bp
- Prototype for scaffolding
  - We developed a pipe line for scaffolding where we (1) simulate heterozygous polyploidy genome. (2) simulate reads for long read sequencing technology such as Moleculo and PacBio. (3) assembly contigs with Moleculo reads and (4) scaffold with PacBio reads.
- **Recommendations for de novo genome assembly**
  - Use the longest possible reads for complex genomes
  - Use overlap graph for long reads to fully take advantage of information in long reads.
  - Use PacBio reads for scaffolding instead of Illumina jumping library for cost, effectiveness and accuracy.

Acknowledgement

Microsoft Research

Ravi Pandya
Bob Davidson
David Heckerman

Universidade de São Paulo

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

Cold Spring Harbor Laboratory

Michael Schatz

Brookhaven National Laboratory

Shinjae Yoo