Generating Complex High Quality Plant Reference Genomes with PacBio

Jeremy Schmutz[1,2], Jerry Jenkins[1], Chris Daum[2], Christopher Plott[1], Dave Flowers[1], Kerrie Barry[2], David Goodstein[2], Jane Grimwood[1,2]

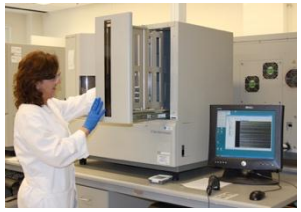
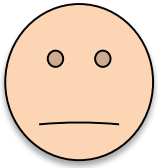
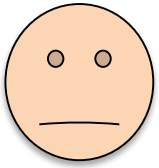
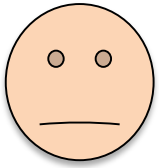
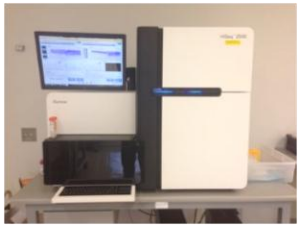
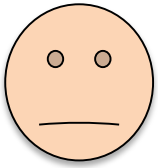
[1] HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
[2] DOE Joint Genome Institute, Walnut Creek, CA, USA

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY

1. Model: Understand how a genome and its genes interact with their environment
2. Crop: Build resources to link genetic variation to trait differences and apply these improve agriculture

Timeline of plant genomes

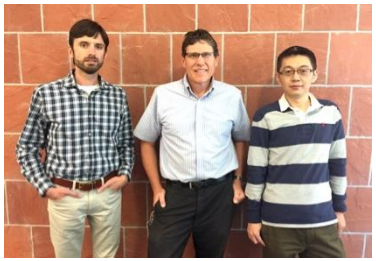
		Cost	Completeness	Contigs	Scaffolds
2005					
2009					
2012					
2016					

Overview

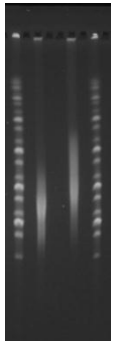
- How to sequence a plant genome with PacBio
- Results on JGI Flagship genomes
- Allotetraploid references on the RS
- Complex genomes with Sequel

How to sequence a plant genome with PACBIO

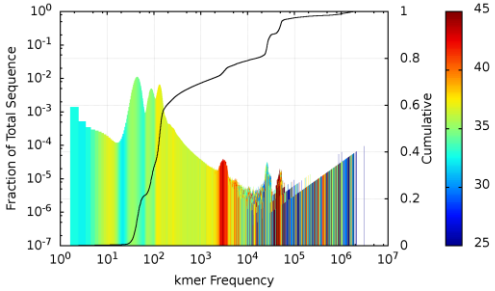
1. Get good DNA (50-150kb)



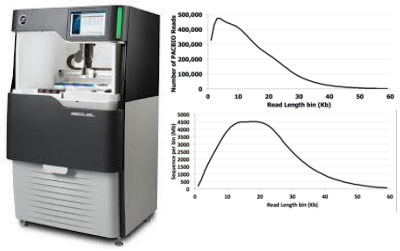
AGI- Wing Group



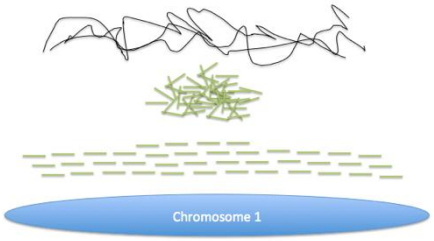
2. Assess the genome: Ploidy, het rate, repeat, organelle, contamination



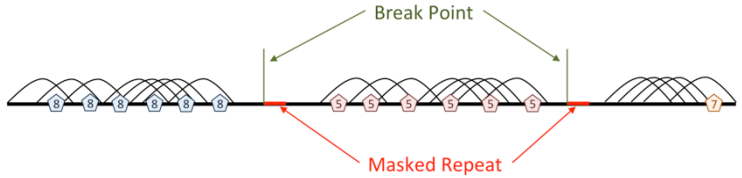
3. Generate long, high quality PACBIO libraries and collect data, either 40-50x or 70-80x, + outbreds



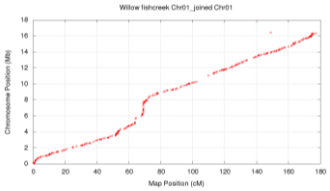
4. Assemble the genome (Falcon, MECAT, Canu), quiver or arrow polish



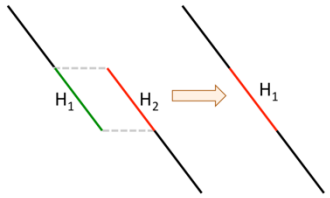
5. Break false joins



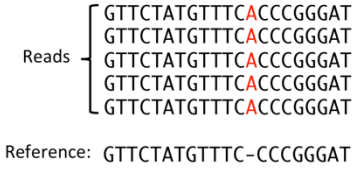
6. Integrate data and order into chromosomes



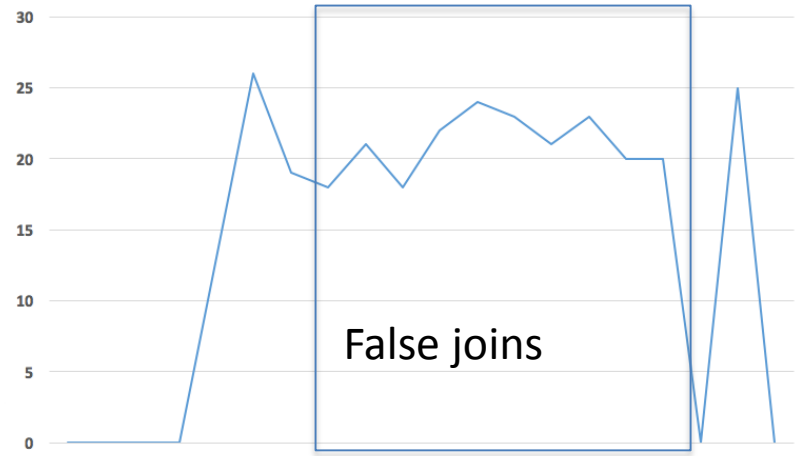
7. Address haplotype & assembly overlap

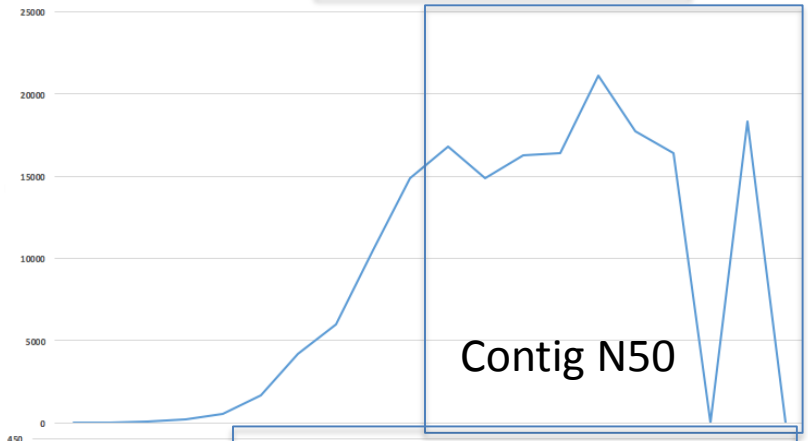


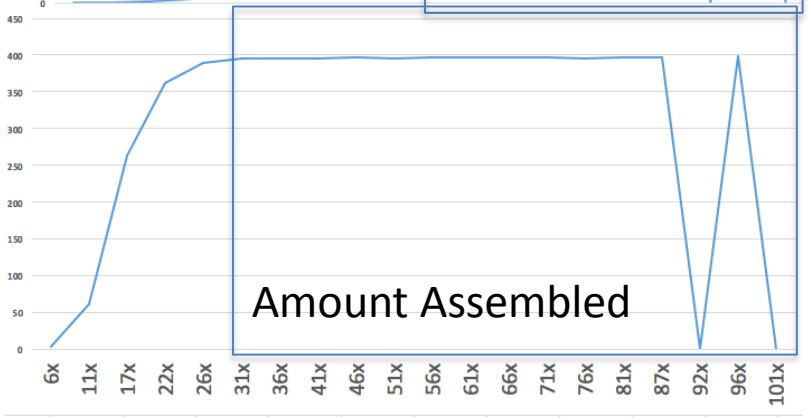
8. Final polish with Illumina to reduce homozygous errors

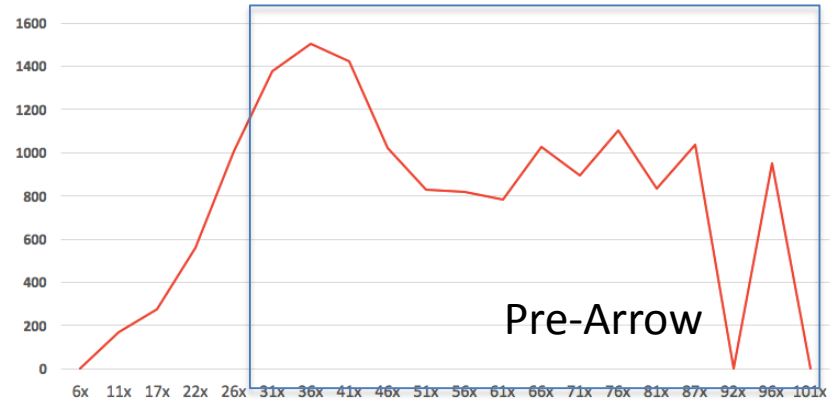


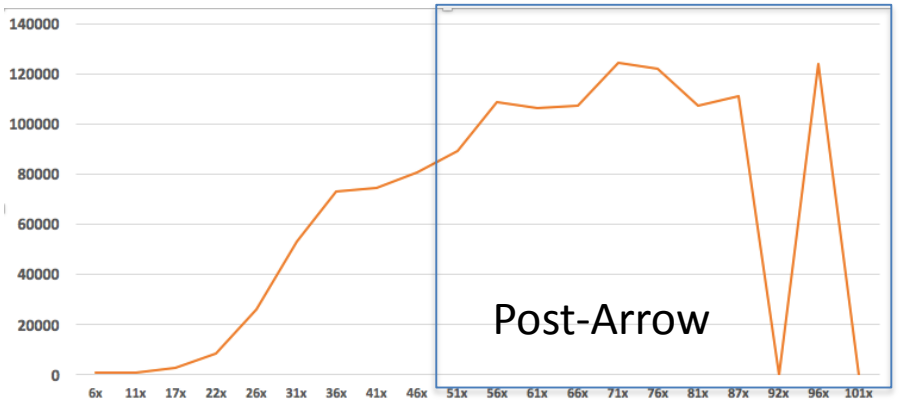
How much to sequence?











60x-70x raw is ideal for inbred
30x+ to survey the genome

MECAT and Canu

MECAT: fast mapping, error correction, de novo assembly for single molecule sequencing reads

- Superior performance compared to current tools with improved results
  - 10-70 X faster on reference genome, 5-20X on pairwise alignment, 5-20 X faster on error correction
- Collaboration between Sun Yat-sen University and Clemson
- Available at https://github.com/xiaochuanle/MECAT

Feng Luo

CLEMSON UNIVERSITY

中山大學 SUN YAT-SEN UNIVERSITY

Canu: Modified CELERA that uses MinHash to assess overlaps & localizes repeats to increase scaffolding accuracy

- auto-grid configuration and general ease of use
- GFA graph output
- Easy, flexible and accurate
- Available at https://github.com/marbl/canu

Adam Phillippy

Brian Walenz

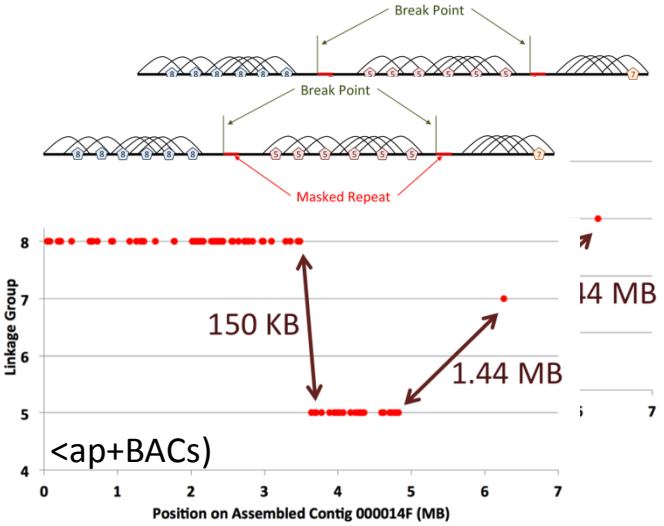
Sergey Koren

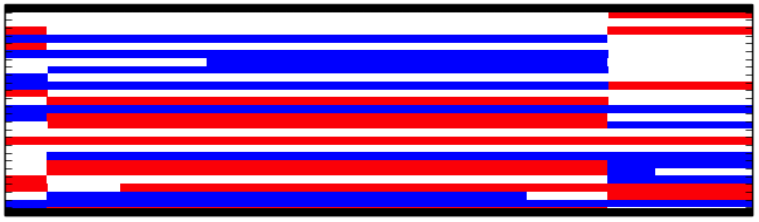
NIH National Human Genome Research Institute

Koren, Genome Research (2017)

Resolving false joins

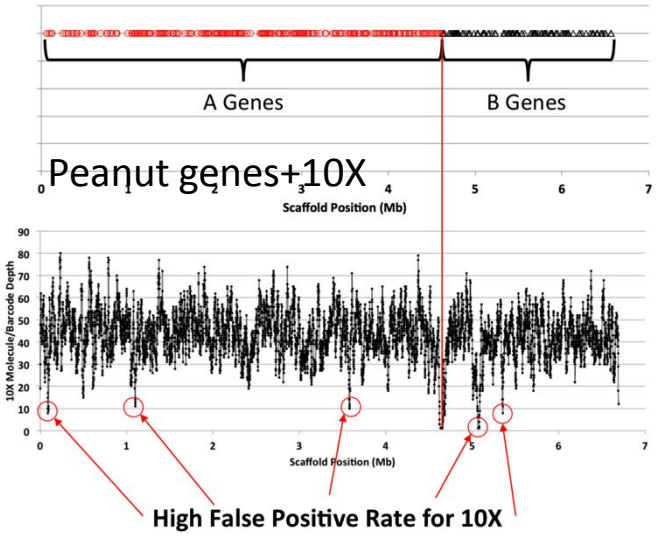
- False joins occur in every plant assembly!
- Common Challenges:
  - In genetic maps the distance between markers can be large
  - Illumina pair information has a high false positive rate due to repeats
- Best practice is to combine 2 or more resources





000006F, nMarkers=6943, size=3,647,542

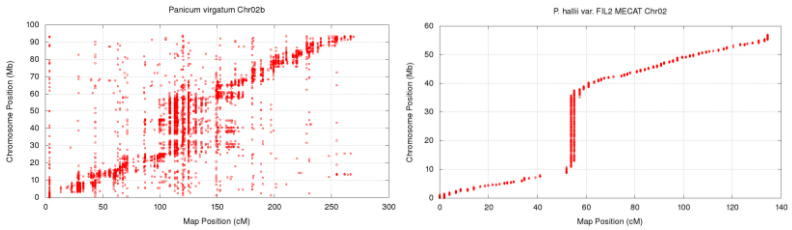
POP-seq markers indicate abrupt change in calling at misjoins.



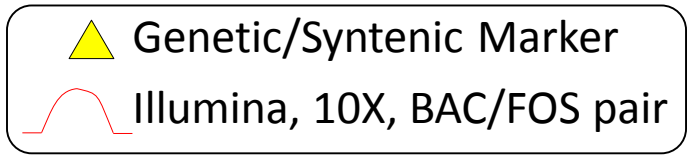
Mascher: Plant Journel 2013

Constructing chromosomes

- Leverage genetic map, BAC/FOS, 10X, HI-C, POP-seq, and/or synteny to accurately order and orient contigs/scaffolds into chromosomes



Maps vary in fidelity and density

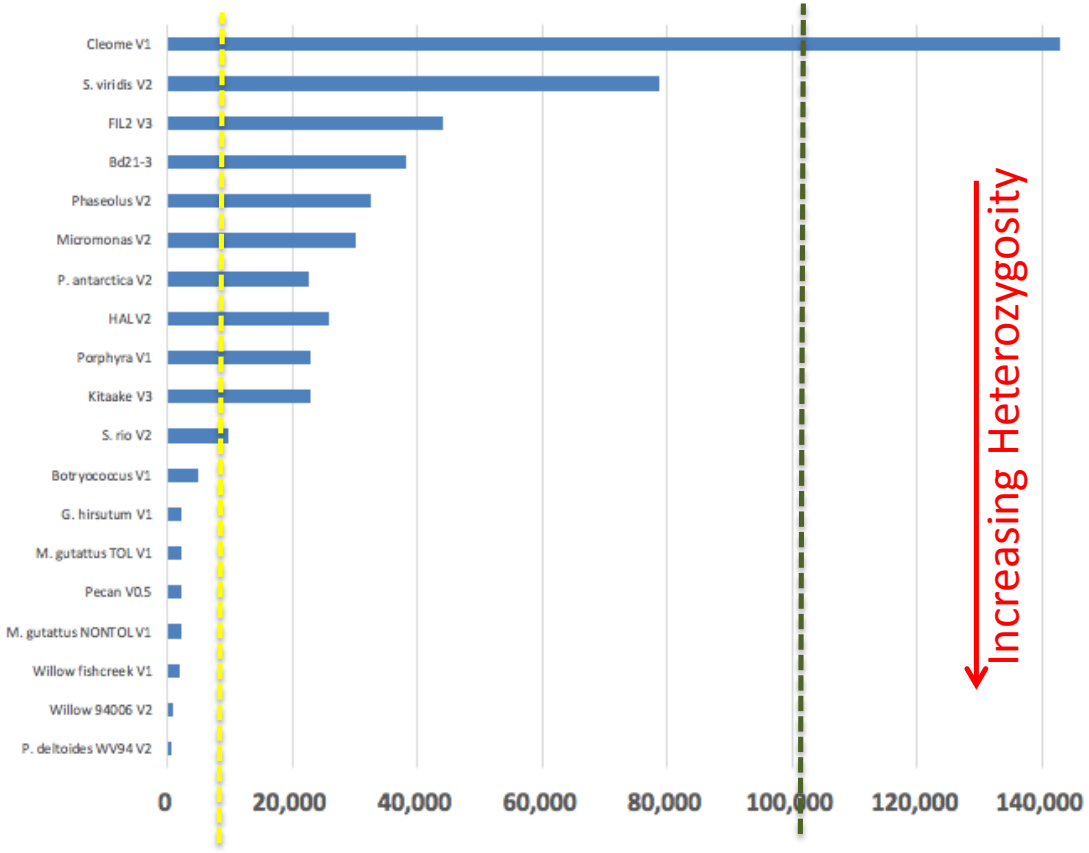


Sequence with a single marker can be oriented using pairing information

Unanchored sequence can be oriented and ordered using pairs as well.

Polishing away remaining InDels

- We polish with Illumina resequencing against alignable bases by patching in correct consensus
- Extremely important for non-inbreds!



Worst: P. deltoides 1 in 764
Willow 1 in 1,018

Best: Cleome 1 in 142,791
S. viridis 1 in 78,689

Data sources

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY

Chris Daum

Jane Grimwood

Upgrading and expanding Flagships

Panicum virgatum

Panicum hallii

Sorghum bicolor

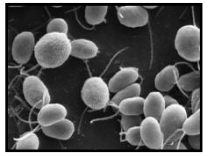
Brachypodium distachyon

Miscanthus sinensis

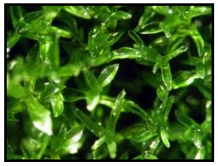
Setaria italica

Glycine max

Populus trichocarpa

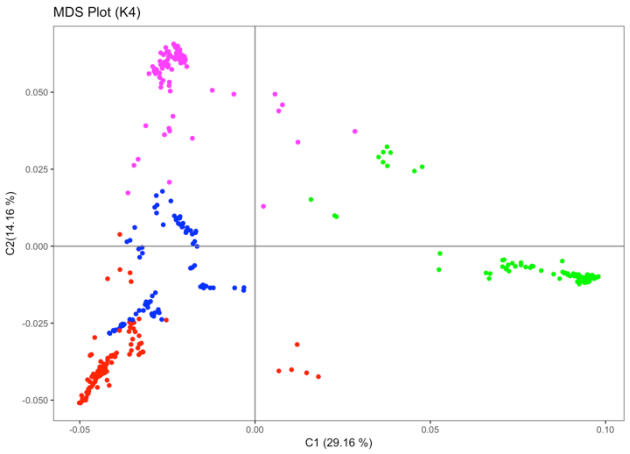


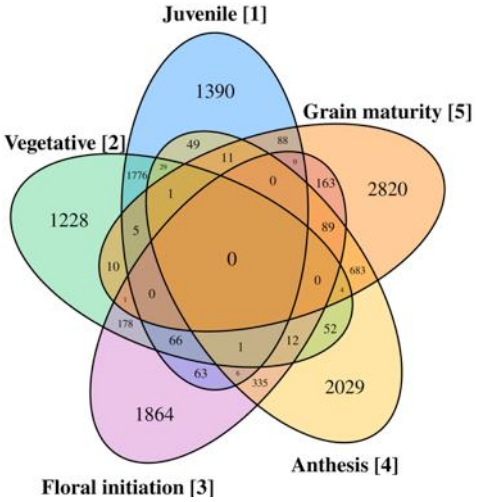
Chlamydomonas reinhardtii

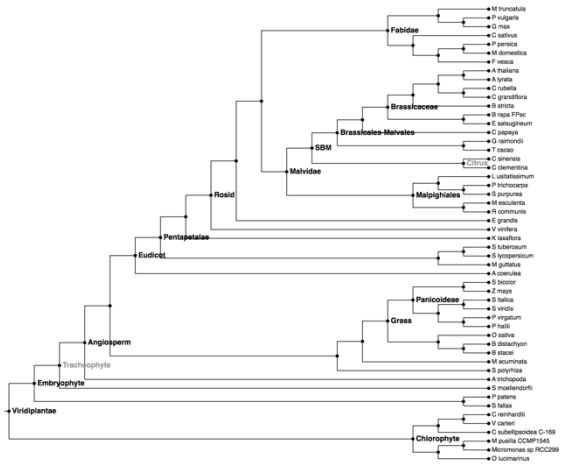


Physcomitrella patens

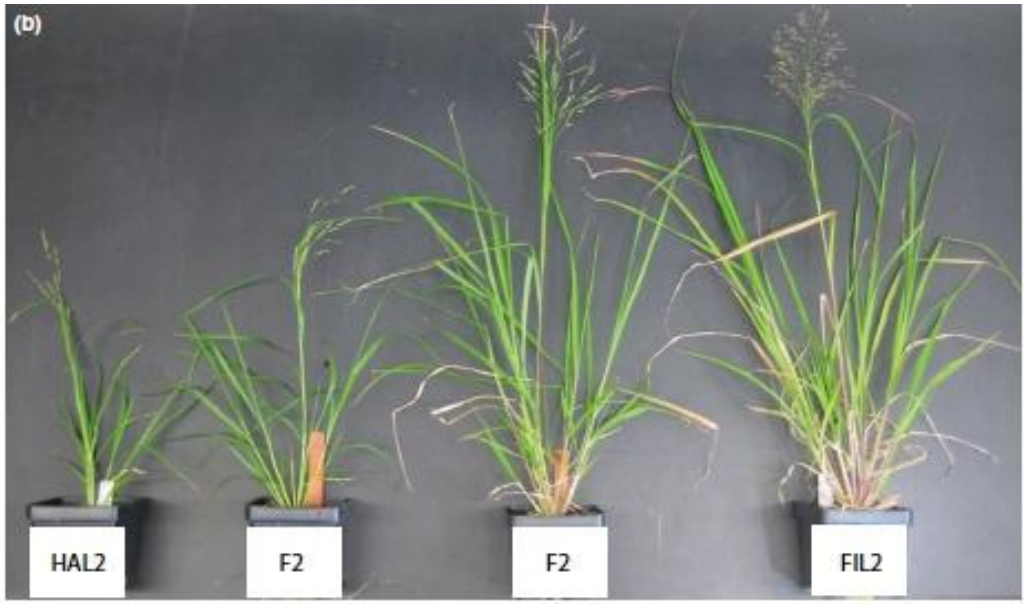
JGI Plant Flagship Genomes

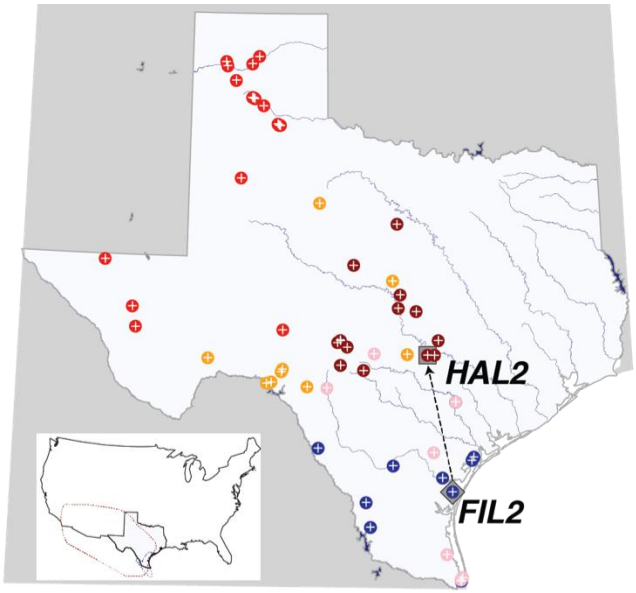






Panicum hallii: switchgrass model





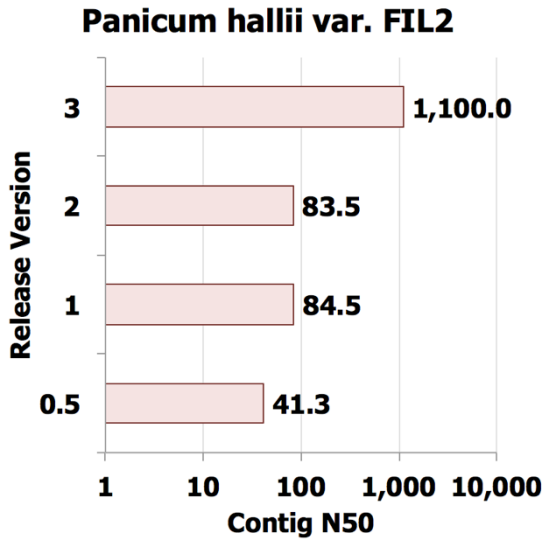
- Desert and coastal adapted varieties (HAL and FIL)
- Diploid and crossable = genetics
- Developed by Juenger Lab (UT-Austin)
- 550 Mb

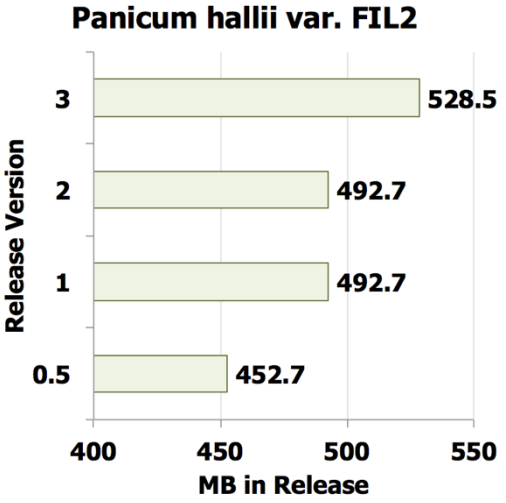
National Science Foundation
Directorate for Biological Sciences (BIO)

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

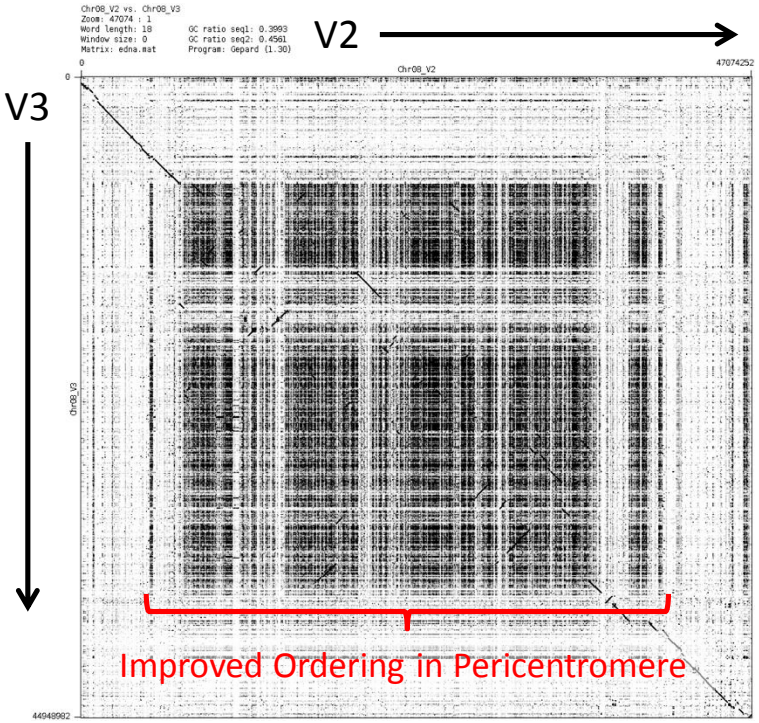
FIL2

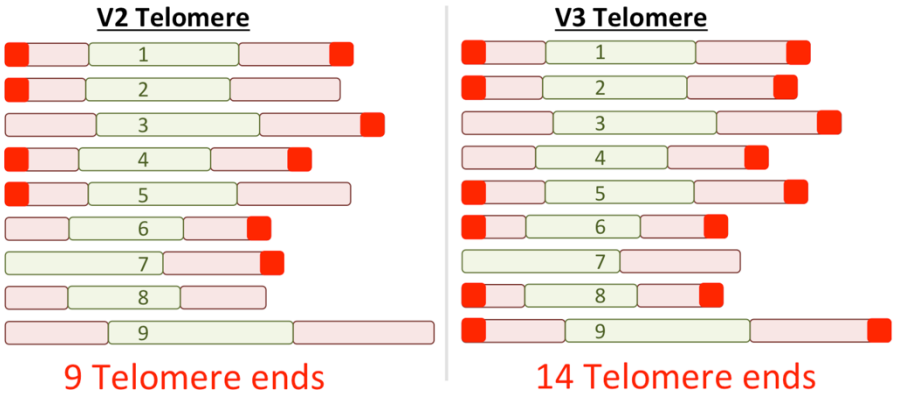
- A short list of technologies thrown at FIL
  - Sanger BES & FES
  - Short read and long read 454
  - Illumina fragments & Mate pairs (2x100,2x150,2x250)
  - Moleculo reads
  - RAD-seq maps, eQTL based maps
  - Illumina clone based pools
  - 10x genomics
  - PacBio P5/C3, P6/C4
  - Abyss, ALLPATHS, Arachne, Mercaulous, Falcon, MECAT





Improved FIL V3

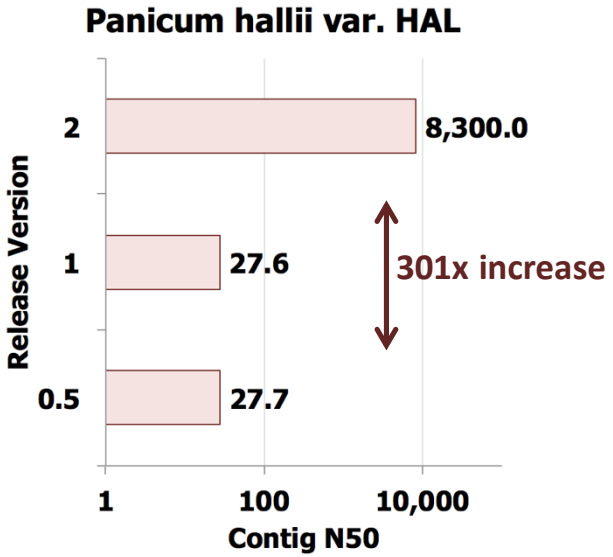


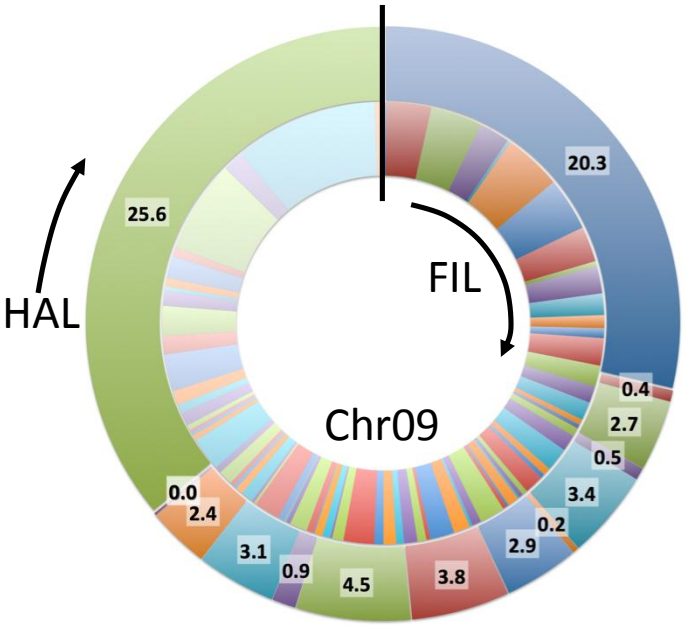


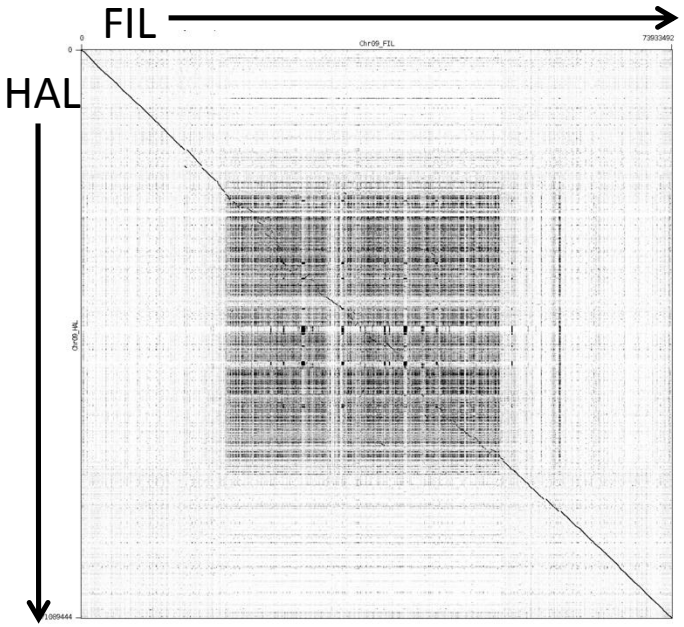
Version	1.0	3.0
Primary	37,254	40,232
Alternate	12,627	10,387
Total	49,881	50,619

HAL awesomeness

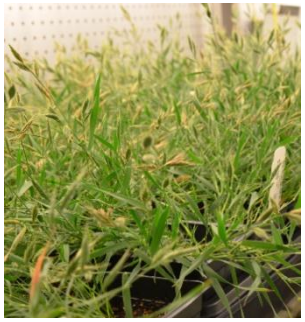
- P6/C4 RS2
- 1 month
- 3 days for assembly
- End to end telomeres







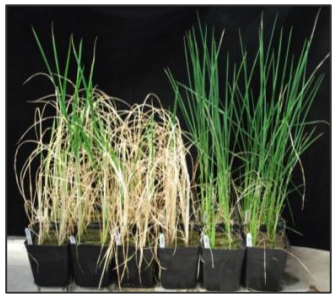
Alternative reference genotypes



Brachypodium BD21-3

Debbie L Chingcuanco
USDA

	Bd21-1 V3	Bd21-3
Size	271 Mb	273 Mb
Contig N50	22.0 Mb	12.6 Mb
Contigs	34	62



Oryza sativa var. Kitaake

Pam Ronald
UC Davis

	Nipponbare*	Kitaake
Size	374 Mb	378 Mb
Contig N50	7.7 Mb	1.4 Mb
Contigs	302	476

Sorghum bicolor var. Rio

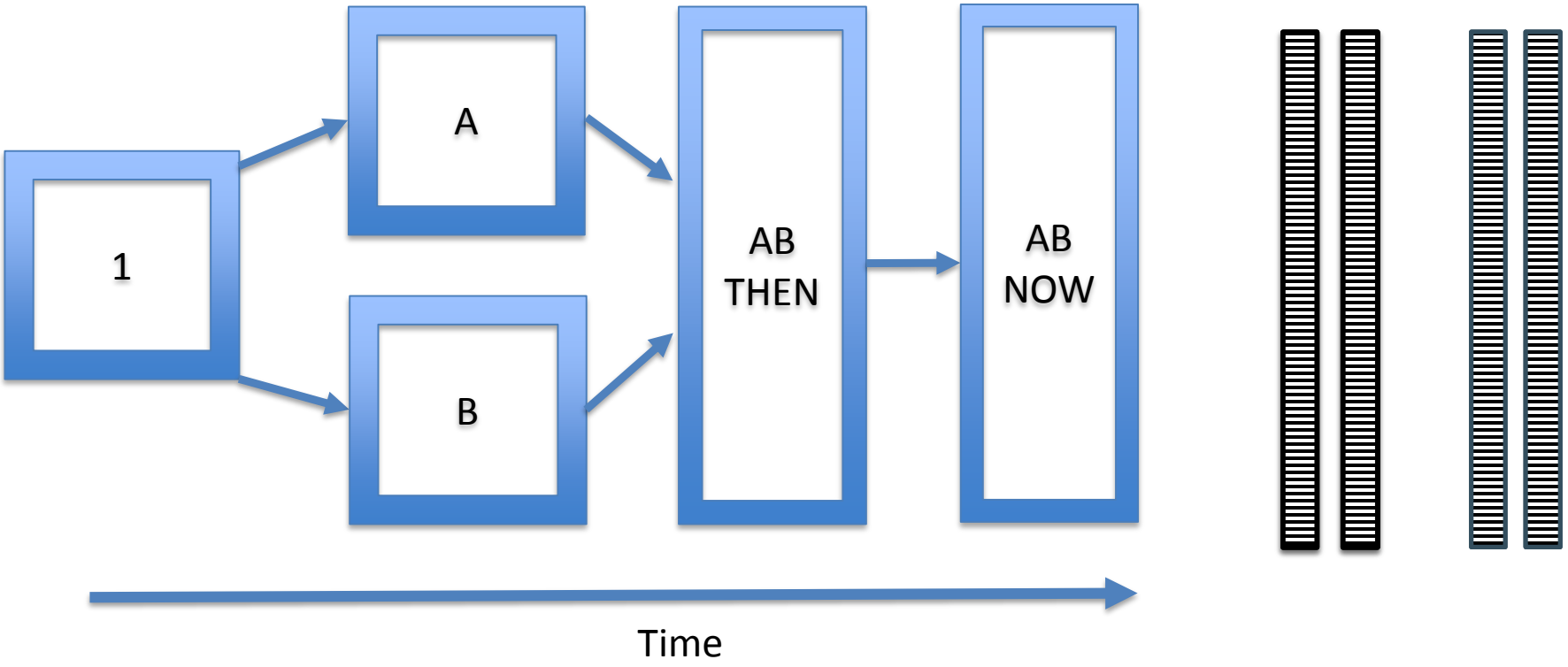
Elizabeth Cooper
Clemson

	BTx623 v4	Rio
Size	635 Mb	694 Mb
Contig N50	1.2 Mb	0.4 Mb
Contigs	4,941	3,830

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

*Kawahara, Rice 2013

Allopolyploids: Why are they difficult to sequence?



Produce high quality reference genome for tetraploid cotton & investigate cotton polyploids

National Science Foundation
Directorate for Biological Sciences (BIO)

Cotton Incorporated

USDA

Jeff Chen
UT-Austin

Brian Scheffler
USDA-ARS

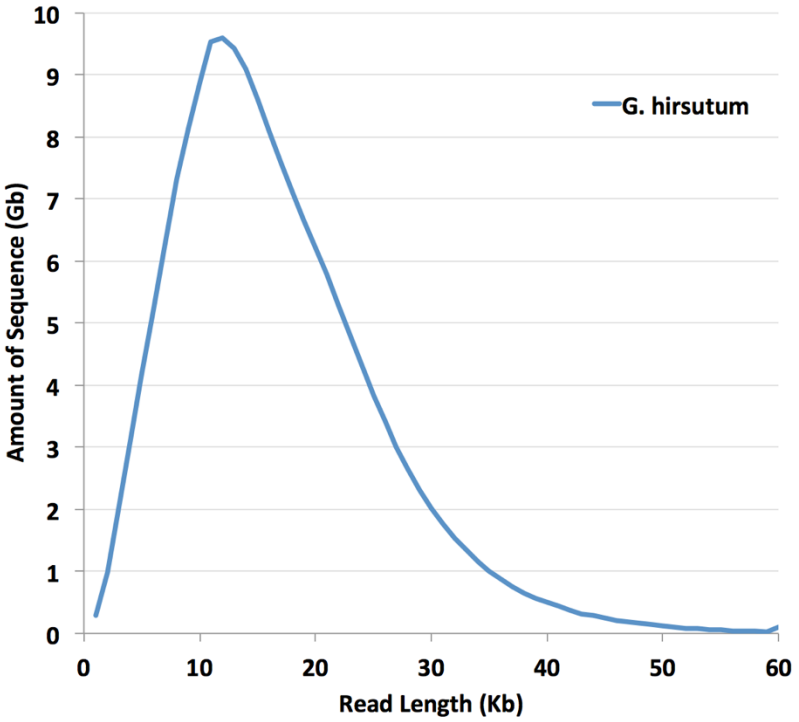
Chris Saski
Clemson University

David Stelly
Texas A&M

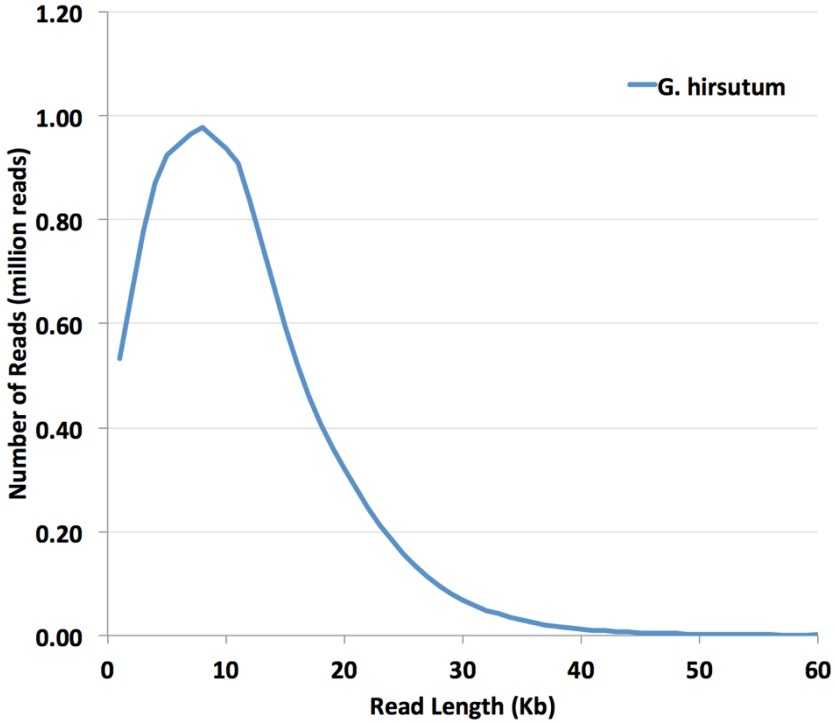
Dan Peterson
MS State

Jane Grimwood
HudsonAlpha

Gossypium hirsutum TM-1



Sequence 179GB

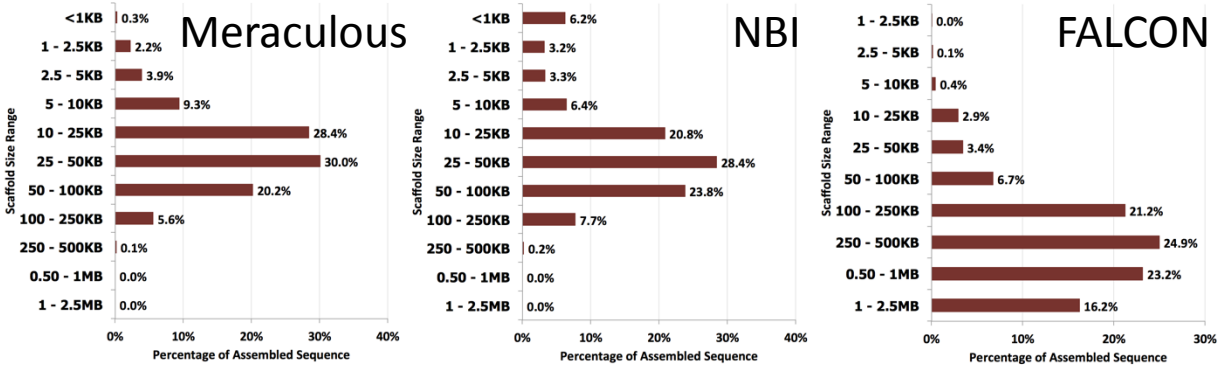


Reads 16.3M: 9.6kb average

295 RS cells worth, 606Mb average single pass!

G. hirsutum assemblies

	Coverage	Assembled Sequence (Gb)	Total Contigs	Contig N50 (Kb)	Annotated Genes Found
Merac[1]	48x illumina	1.81	1,224,135	26	64,190 (96.4%)
NBI[2]	245x illumina	2.16	444,799	58	66,279 (99.6%)
V1 FALCON	78x PacBio (9.5 Kb)	2.26	13,583	389	66,574 (100%)



[1]Chapman, PlosOne, 2011

[2]Zhang, et al. Nature Biotechnology, 2015

http://phytozome.jgi.doe.gov/pz/portal.html

Peanuts

Peggy Ozias-Akins & Scott Jackson
University of Georgia

- Develop tetraploid genome reference for peanut and peanut improvement
- American “orphan crop” with little investment in crop improvement
- Complicated and recent tetraploid genome

THE PEANUT FOUNDATION

Current peanut assemblies

		Sequence (MB)	Contigs	CN50 (kb)
A	Illumina	940	135k	23
B	Illumina	1,257	122k	23
Tetraploid	Illumina	2,042	153k	47
Tetraploid	PACBIO	2,525	3k	1,600

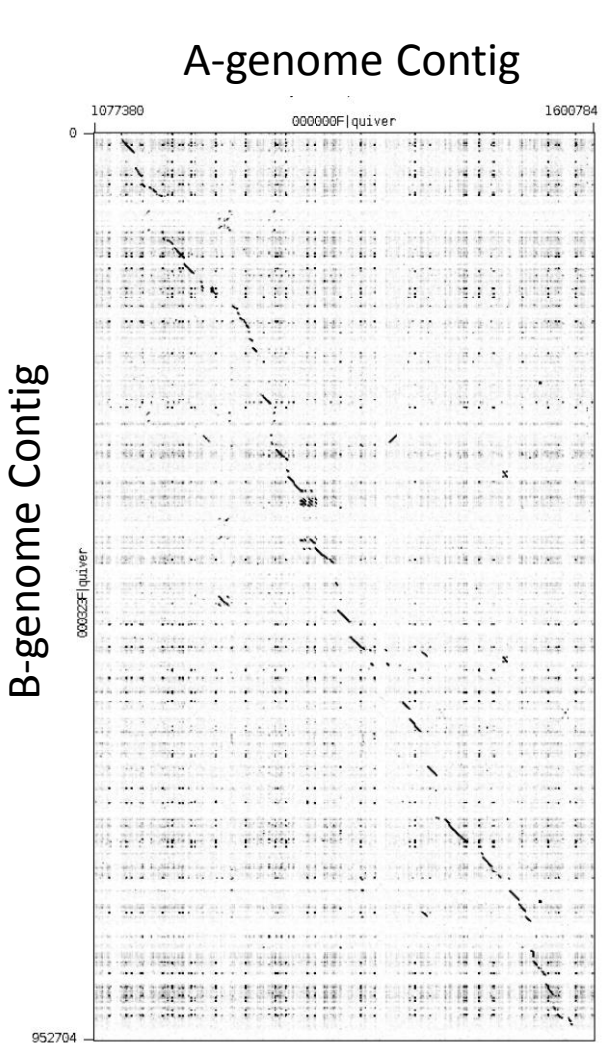
NATURE GENETICS | ARTICLE OPEN

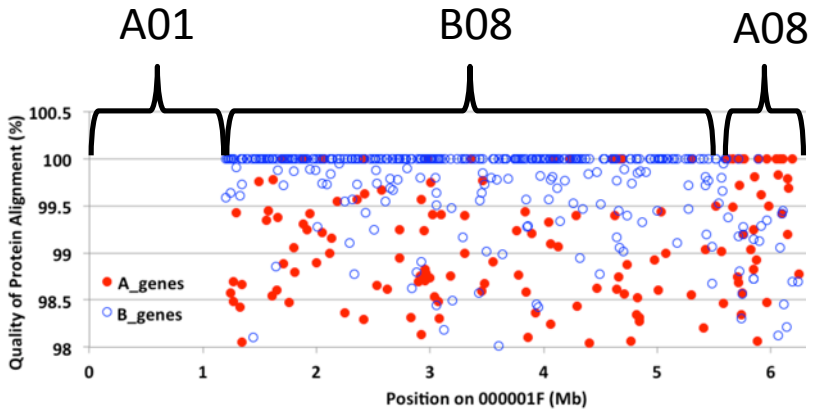
日本語要約

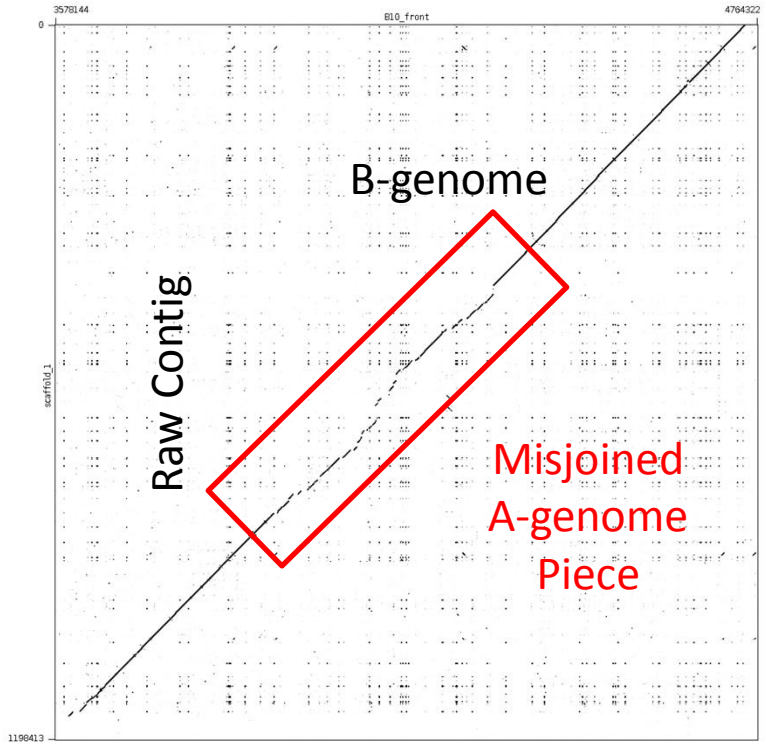
The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut

Bertioli, 2016

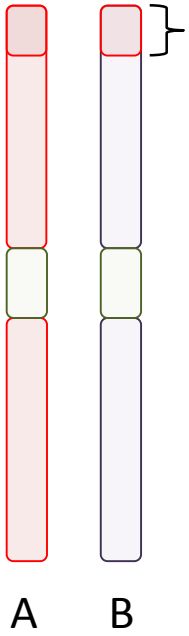
Splitting A & B is challenging



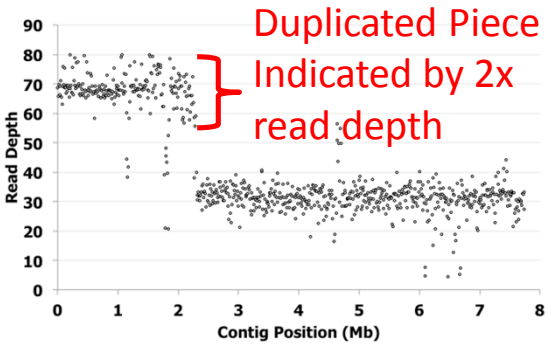


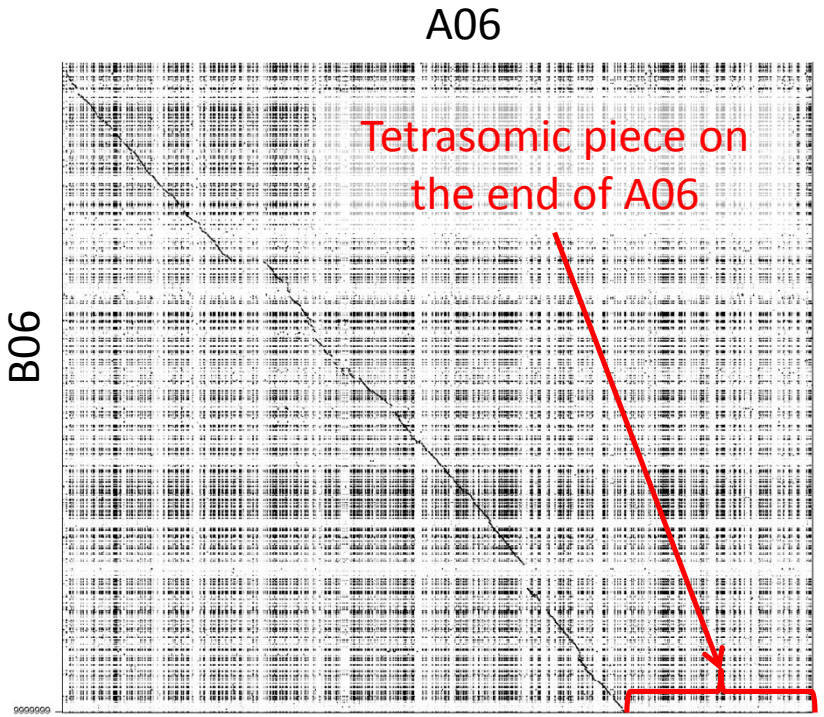


A &B duplications indicate tetrasomy



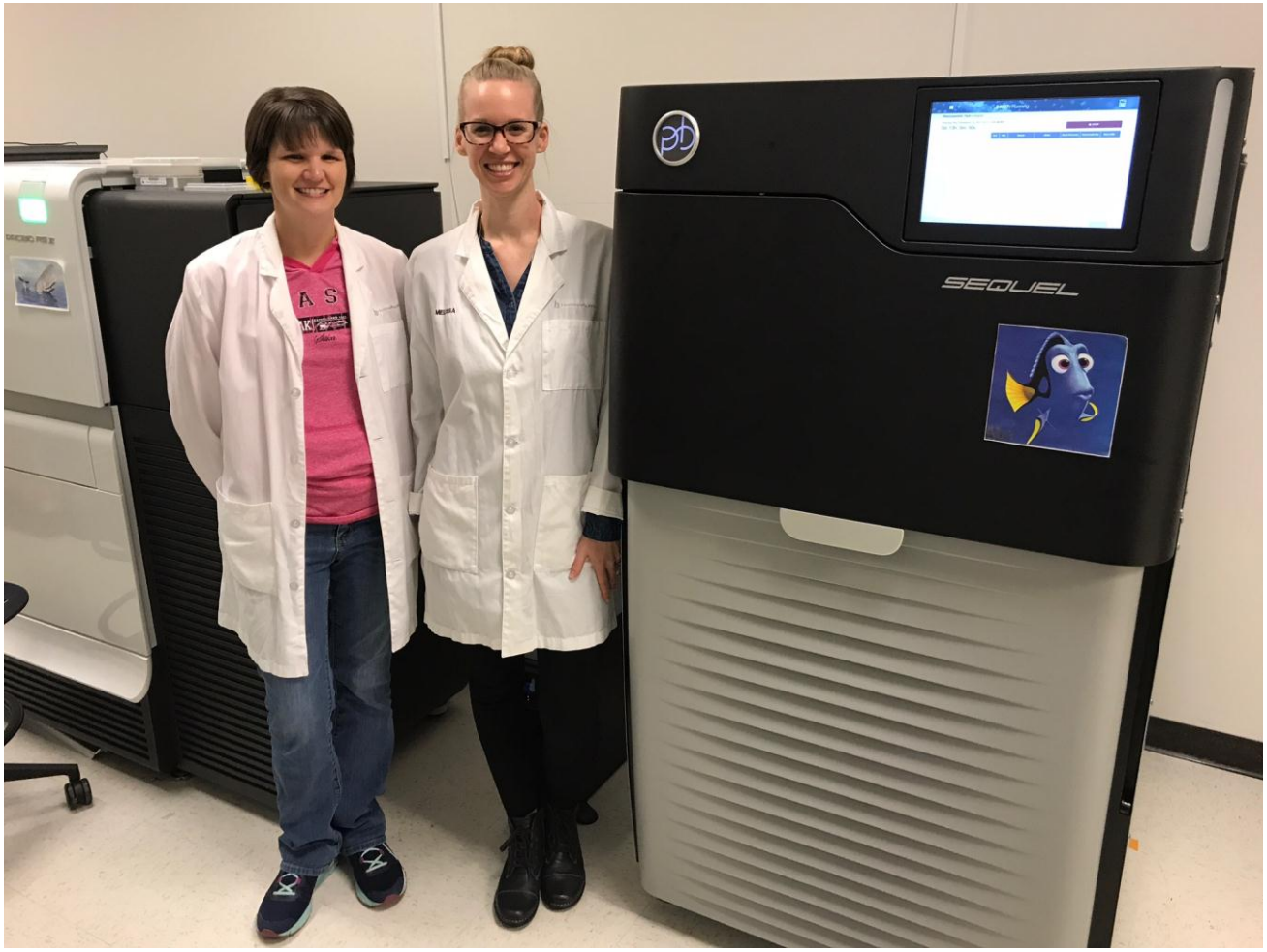
Tetrasomic region where segment of A has been duplicated in B





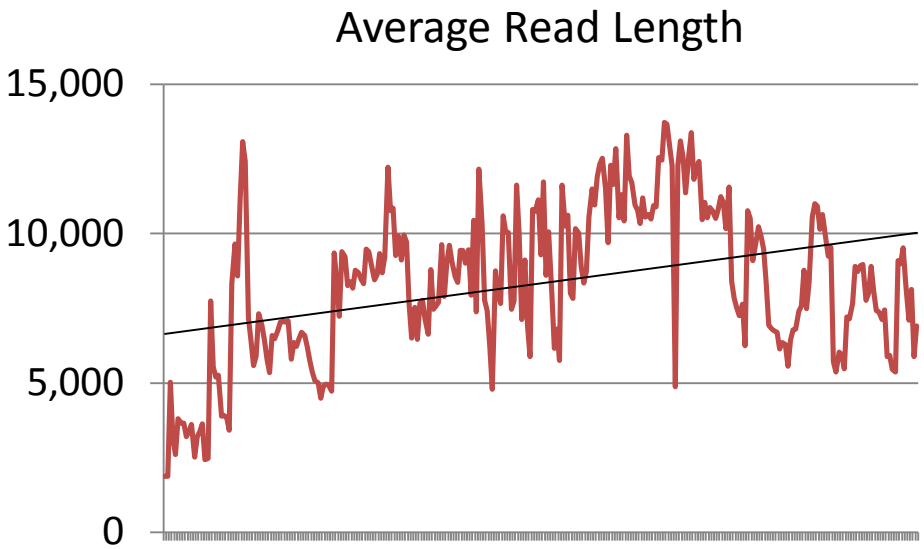
We will duplicate these regions in the final peanut assembly…

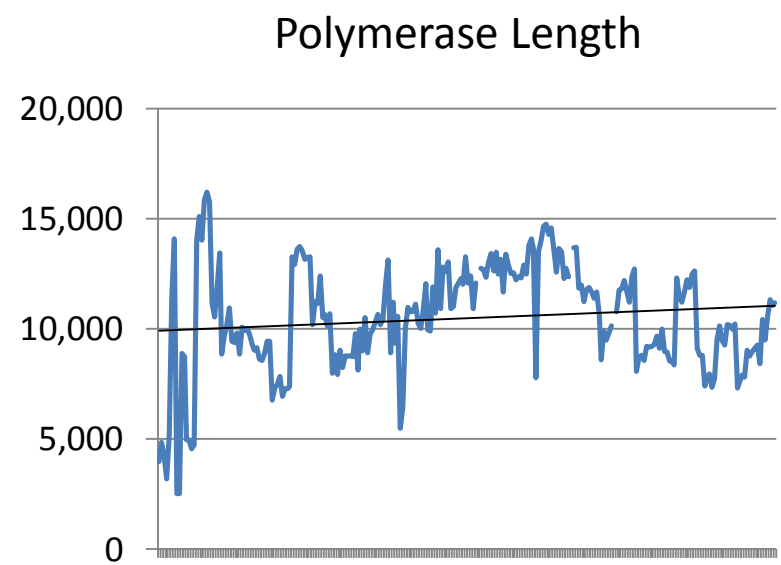
Sequel progress

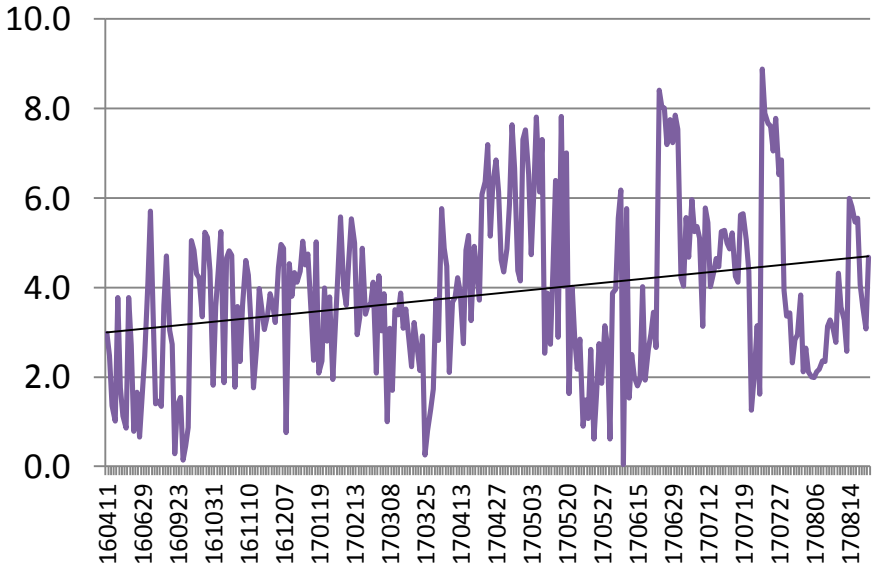


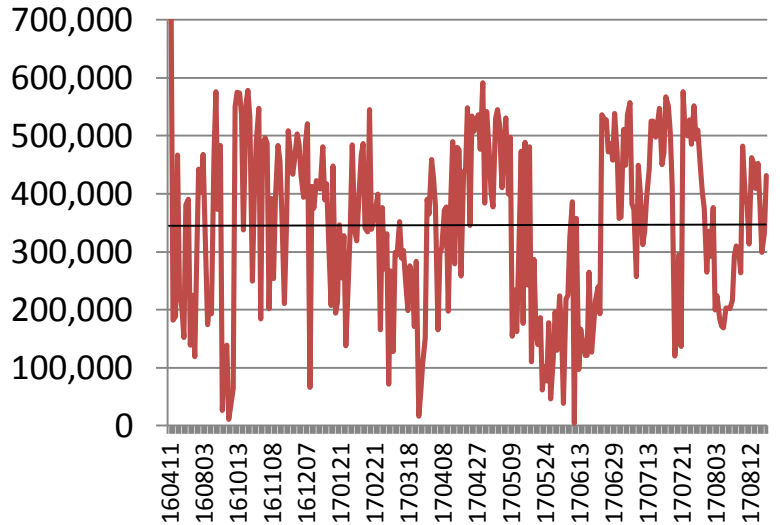
- Test shear DNA
- Carefully titrate DNA, for Sequel less DNA can make a better library
- Increase bead bind durations for difficult samples
- Run one cell from polymerase bind, then estimate for the rest

Sequel performance @ HA

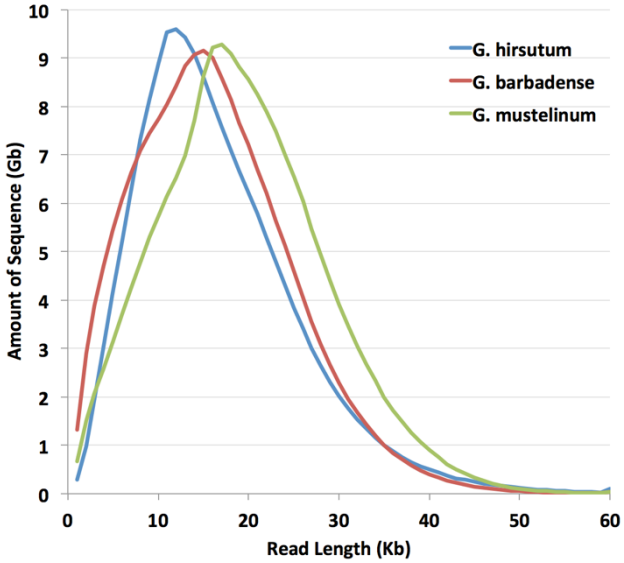


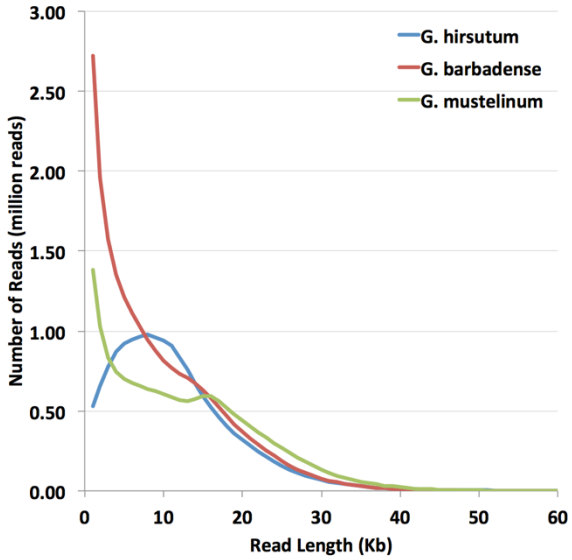






Sequel Data Collection - Cotton





	Total Reads	Coverage	Average Read Length
G. hirsutum	16,328,337	66.3x	9,589
G. barbadense	21,649,137	71.6x	6,874
G. mustelinum	16,607,922	77.1x	10,729

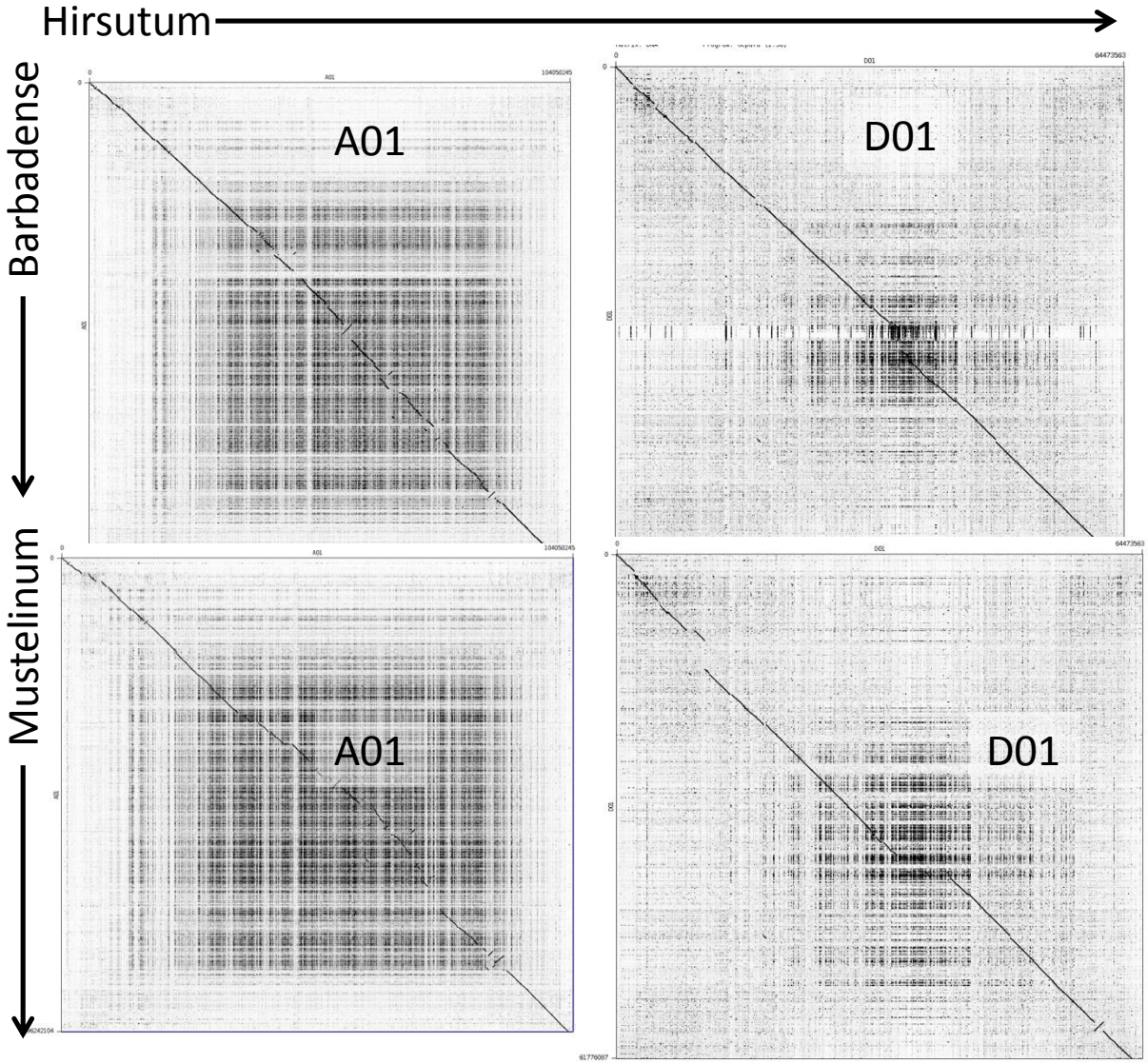
Sequel based polyploid assemblies

- 2,493 contigs
- 2.2 Gb assembled
- 2.3 Mb contig N50
- 58 cells

GB

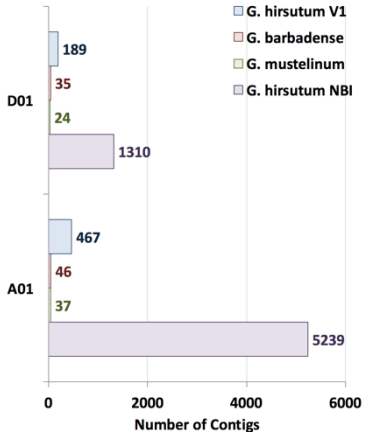
- 2,267 contigs
- 2.3 Gb assembled
- 2.5 Mb contig N50
- 42 cells!

GM



GB

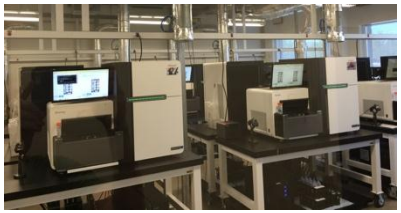
GM



2100-3000 kb contigs, orders of magnitude less contigs!

Better Tools

New Plant References



Diversity & Transcriptomics

1

10ky

N & H_20

Normal Way

Min N & H_20
Pest Resistance
Perennials

Better Way 2

[1]Doebley 1995 [2]Land Institute

How to work with the JGI

- Community Science Program (CSP): http://bit.ly/CSP_JGI
  - Large-scale due Spring 2018
  - Small-scale microbial and metagenome proposals due **September 4:** http://bit.ly/JGI-Small-Scale
  - DNA Synthesis: http://bit.ly/JGI-DNA-Synthesis
- Facilities Integrating Collaborations for User Science (FICUS)
  - Taps the power of genomics at JGI *and* molecular characterization at EMSL to accomplish one research project. Due Spring 2018: http://bit.ly/FY18-FICUS

- **Strategic Partnership Projects (SPPs): E**nable research funded by an Industry partner to perform a defined scope of work using JGI's unique facilities, equipment, and personnel. Sponsor may elect to obtain intellectual property.
- **Cooperative Research and Development Agreements (CRADAs):** Enable research jointly sponsored by the Berkeley Lab and one or more partners for shared benefit.

Join us in San Francisco...

FOR THE 13TH ANNUAL

Genomics of Energy & Environment Meeting

HOSTED BY THE

U.S. Department of Energy Joint Genome Institute

Hilton San Francisco Union Square

MARCH 13–16, 2018

Registration opens September 12, 2017

- Partnering Sessions
- Sponsorship Opportunities
- Workshops
- Poster Sessions
- Short Talks

VEGA Viral EcoGenomics & Applications

Symposium hosted by the DOE Joint Genome Institute

Funding Sources

U.S. DEPARTMENT OF ENERGY | Office of Science

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

NSF National Science Foundation
Directorate for Biological Sciences (BIO)

BILL & MELINDA GATES foundation

USB UNITED SOYBEAN BOARD
Progress Powered by U.S. Farmers

USDA United States Department of Agriculture
National Institute of Food and Agriculture

cotton Cotton Incorporated

THE LAND INSTITUTE

THE PEANUT FOUNDATION

HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY