De novo assembly of complex crop genomes

Michael Schatz

Oct 18, 2012

PacBio Users Meeting, Menlo Park, CA

@mike_schatz

Plant Genomics

- Motivations
  - 15 crops provide 90% of the world's food
  - Responsible for maintaining the balance of the carbon cycles, soil from erosion
  - Promising sources of renewable energy
  - Plant byproducts used in many medicines
  - Model organisms for studying biological systems

- Goals
  - Understand basis of differences among subpopulations and varieties (duplications, CNVs, etc.) that lead to important phenotypes
  - Many of these differences relate to ability to grow in less than optimum conditions
    - Drought, aluminum tolerance, etc

Why are plant genomes hard to assemble?

1. ***Biological***:
   - (Very) High ploidy, heterozygosity, repeat content

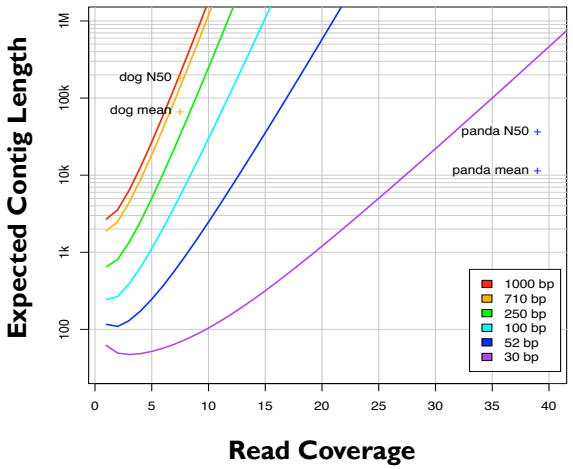
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

3. ***Computational***:
   - (Very) Large genomes, complex structure

4. ***Accuracy***:
   - (Very) Hard to assess correctness

Ingredients for a good assembly

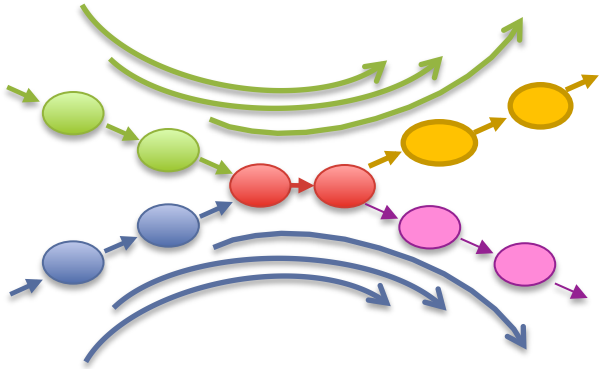
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

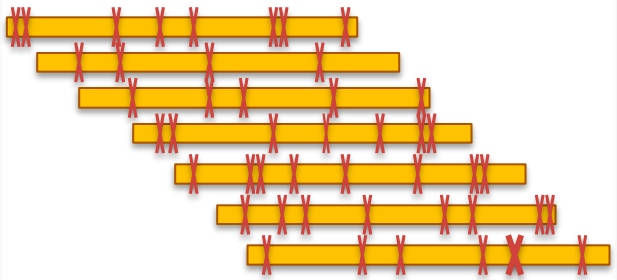
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Hybrid Sequencing

Illumina

Sequencing by Synthesis

High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

Pacific Biosciences

SMRT Sequencing

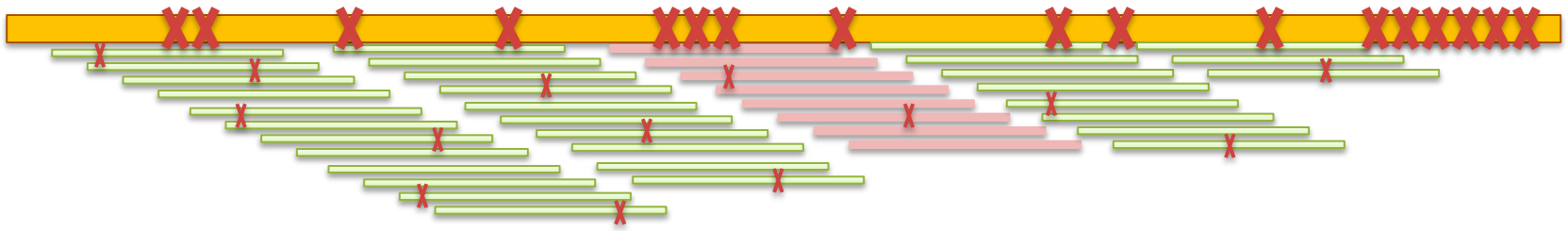
Lower throughput (600Mbp/day)
Lower accuracy (~85%)
Long reads (2-5kbp+)

PacBio Error Correction

http://wgs-assembler.sf.net

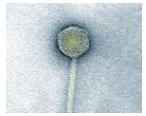
1. Correction Pipeline
    1. Map short reads to long reads
    2. Trim long reads at coverage gaps
    3. Compute consensus for each long read

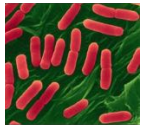
2. Error corrected reads can be easily assembled, aligned

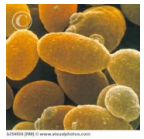


Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, et al. (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

SMRT-Assembly Results







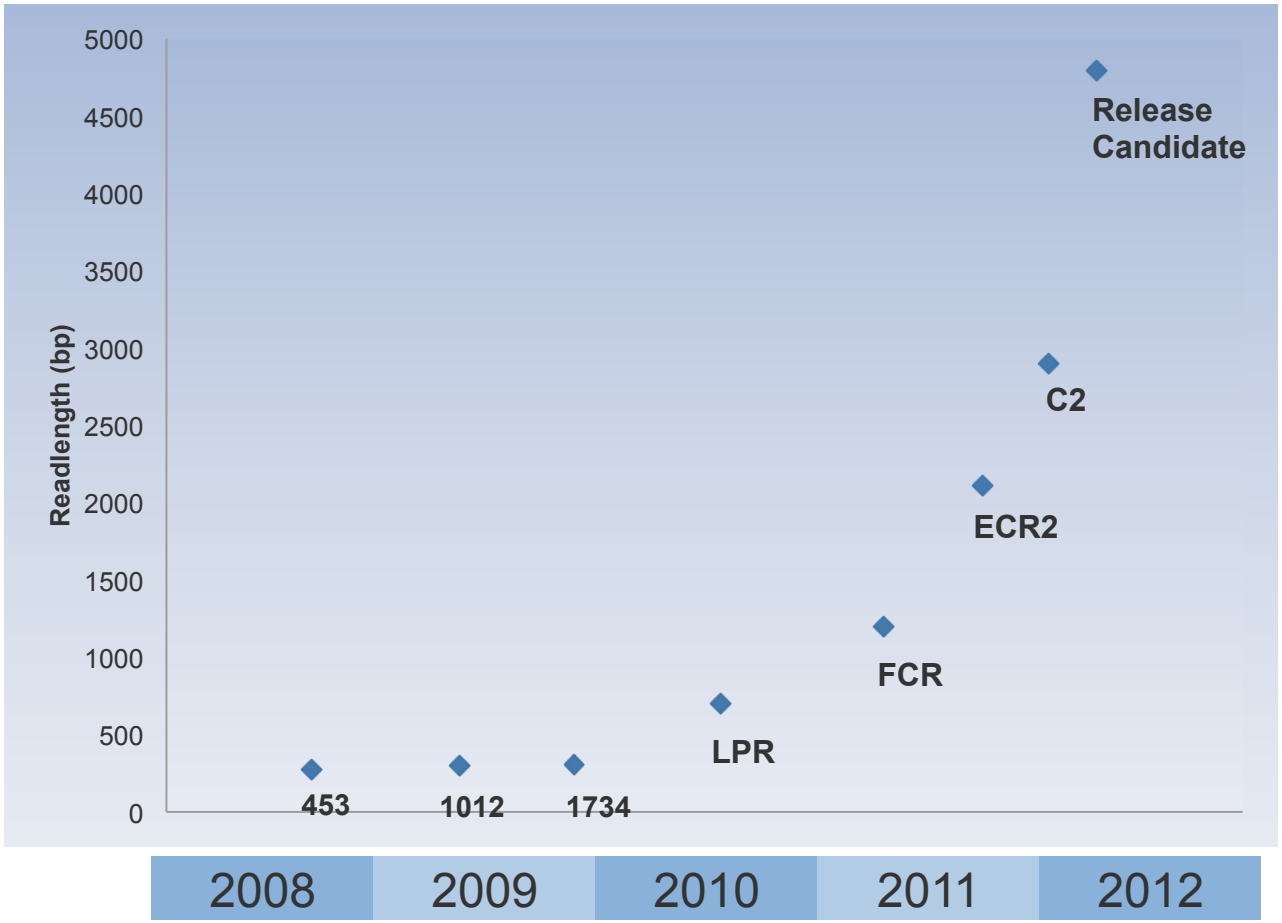
Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
Lambda NEB3011	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
(median: 727 max: 3 280)	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
E. coli K12	Illumina 100X 500bp	4 639 675	4 462 836	61	221 615 / 221 553	100 338 / 83 037 (82.76%) *
(median: 747 max: 3 068)	PacBio PBcR 18X		4 465 533	77	239 058 / 238 224	71 479 / 68 309 (95.57%) *
	Both 18X PacBio PBcR + Illumina 50X 500bp		4 576 046	65	238 272 / 238 224	93 048 / 89 431 (96.11%) *
E. coli C227-11	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322
(median: 1 217 max: 14 901)	PacBio 25X PBcR (corrected by 25X CCS)		5 207 946	80	357 234	98 774
	Both PacBio PBcR 25X + CCS 25X		5 269 158	39	647 362	227 302
	PacBio 50X PBcR (corrected by 50X CCS)		5 445 466	35	1 076 027	376 443
	Both PacBio PBcR 50X + CCS 25X		5 453 458	33	1 167 060	527 198
	Manually Corrected ALLORA Assembly[9]		5 452 251	23	653 382	402 041
S. cerevisiae S228c	Illumina 100X 300bp	12 157 105	11 034 156	192	266 528 / 227 714	73 871 / 49 254 (66.68%) *
(median: 674 max: 5 994)	PacBio PBcR 13X		11 110 420	224	224 478 / 217 704	62 898 / 54 633 (86.86%) *
	Both PacBio PBcR 13X + Illumina 50X 300bp		11 286 932	177	262 846 / 260 794	82 543 / 59 792 (72.44%) *
Melopsittacus undulatus	Illumina 194X (220/500/800 paired-end 2/5/10Kb mate-pairs)	1.23 Gbp	1 023 532 850	24 181	1 050 202	47 383
	454 15.4X (FLX + FLX Plus + 3/8/20Kbp paired-ends)		999 168 029	16 574	751 729	75 178
(median 997, max 13 079)	454 15.4X + PacBio PBcR 3.75X		1 071 356 415	15 081	1 238 843	99 573

Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case

Transcript Alignment

- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
  - Raw reads and raw alignments (red) have many spurious indels inducing false frameshifts and other artifacts
  - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing
- New collaboration with Gingeras Lab looking at splicing in human

PacBio Technology Roadmap



Internal Roadmap has made steady progress towards improving read length and throughput

Very recent improvements:

1. Improved enzyme:

   Maintains reactions longer

2. "Hot Start" technology:

   Maximize subreads

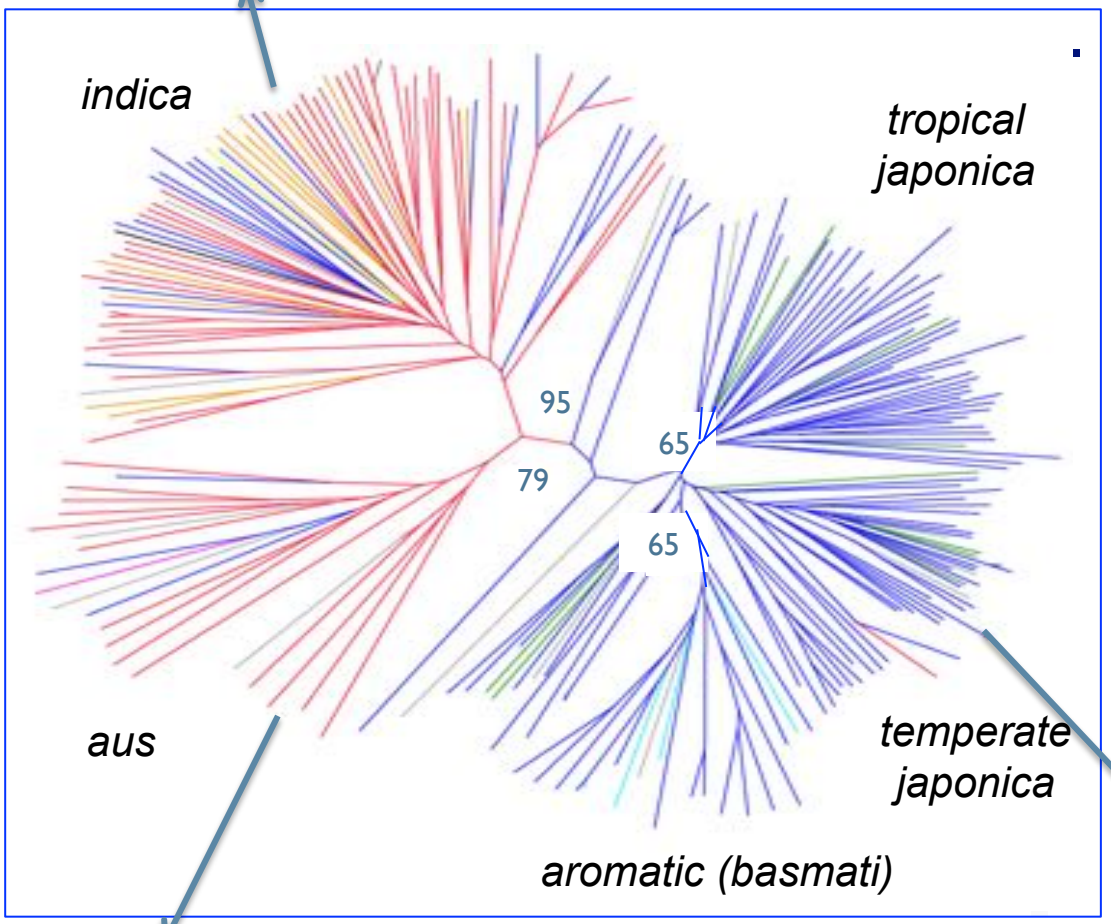
3. MagBead loading:

   Load longest fragments

Population structure in *Oryza sativa*

3 varieties selected for *de novo* sequencing

IR64



DJ123

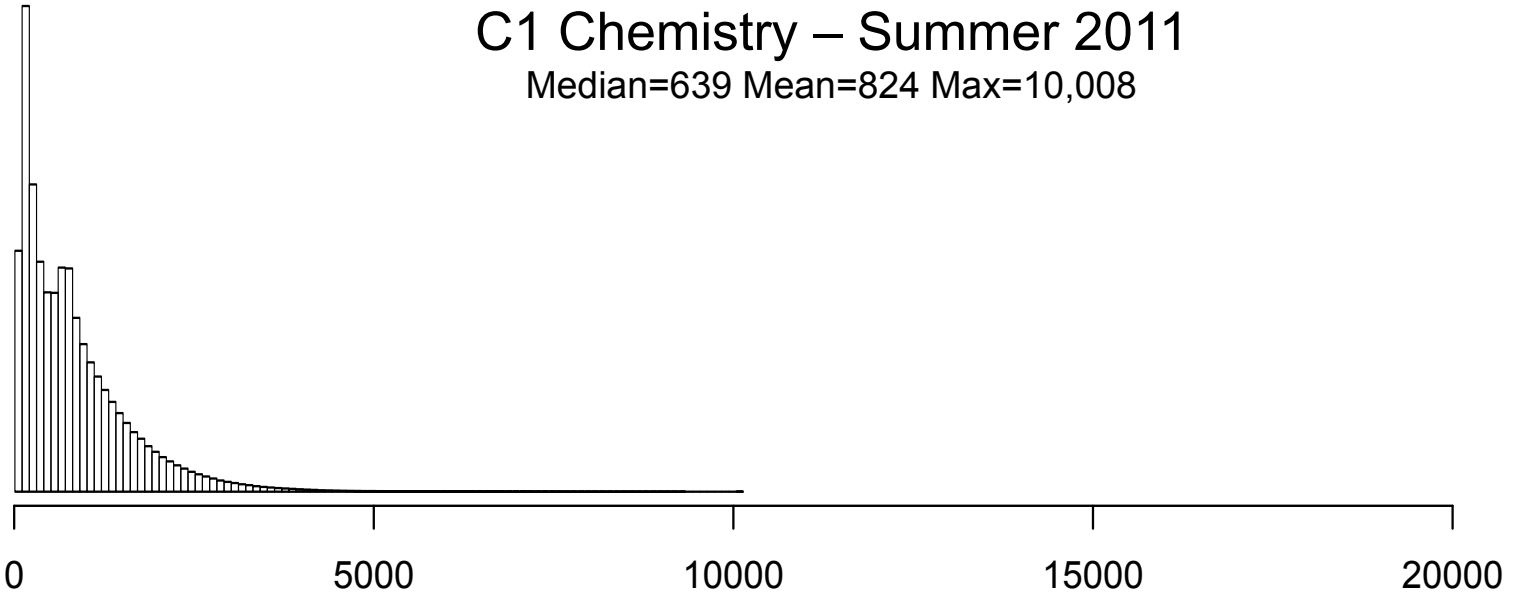
Nipponbare

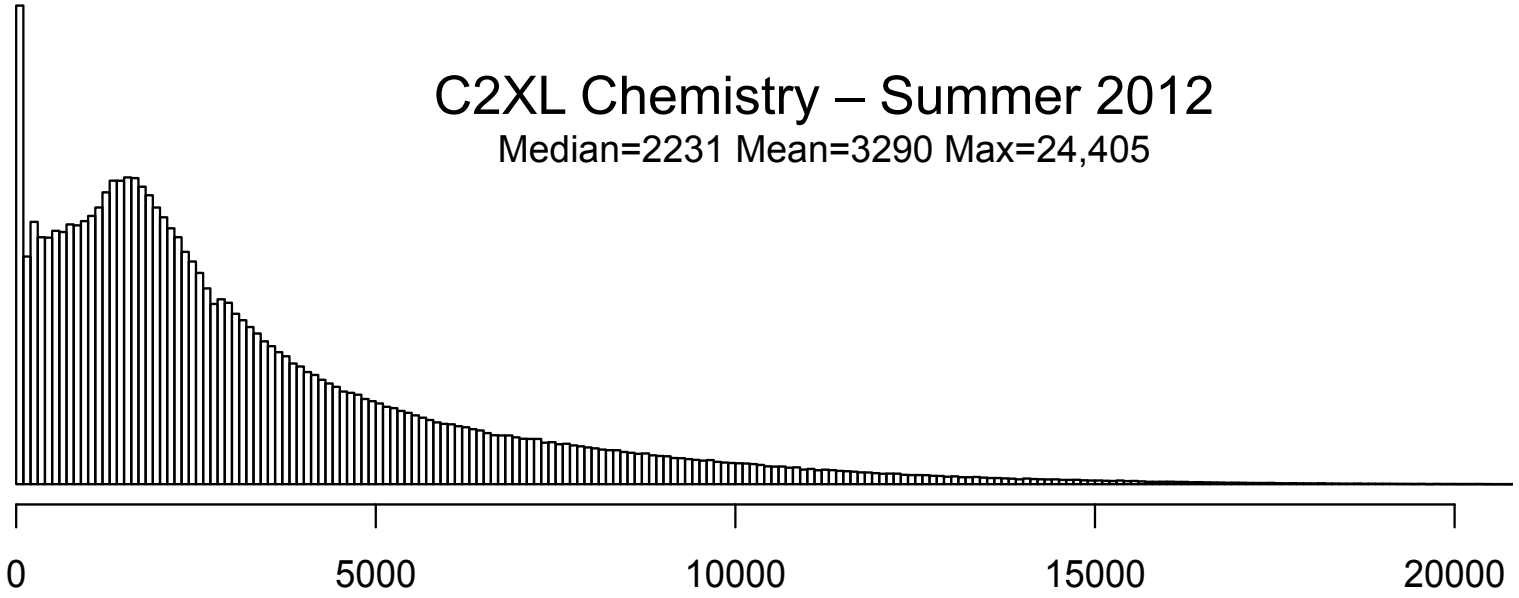
Garris et al. (2005)
Genetics 169: 1631–1638

Genome Characteristics

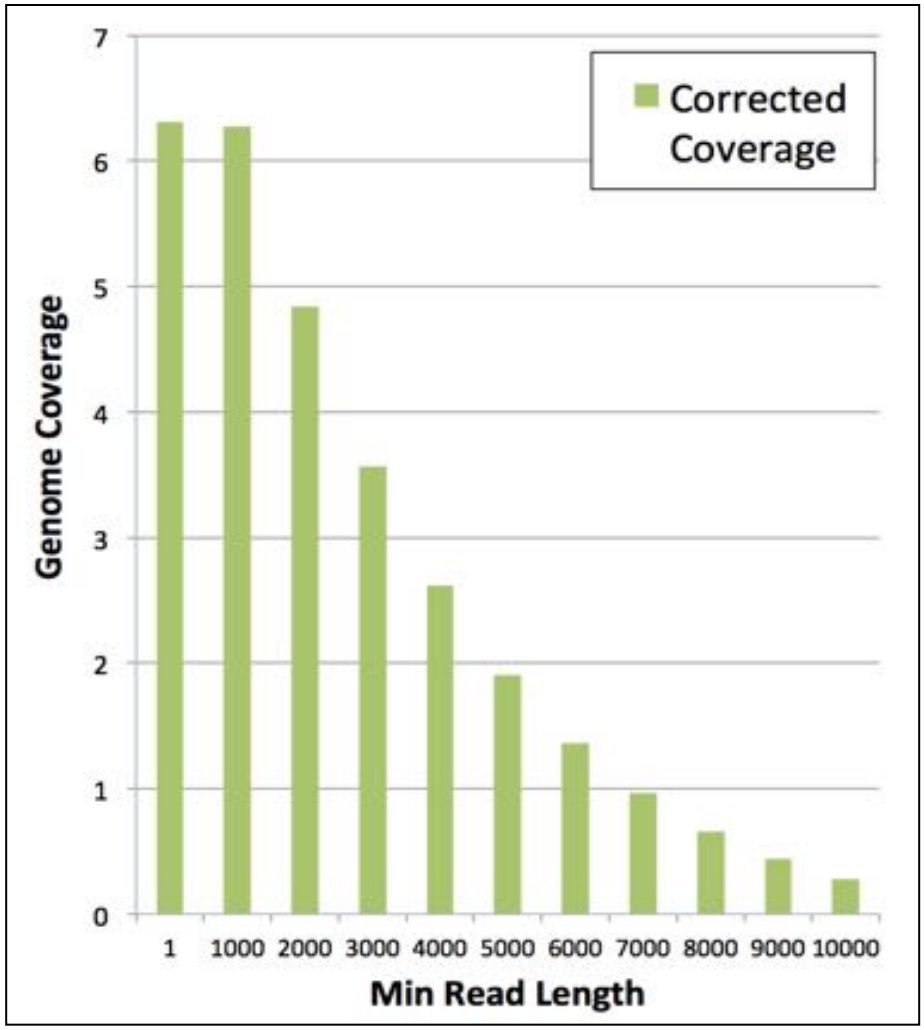
- 440 Mbp genome
- About 40% repeats
- Relatively easy to get high quality DNA
- High quality, BAC by BAC reference available for Nipponbare
- Useful model for other cereal genomes

PacBio Long Read Rice Sequencing





Preliminary Rice Assemblies



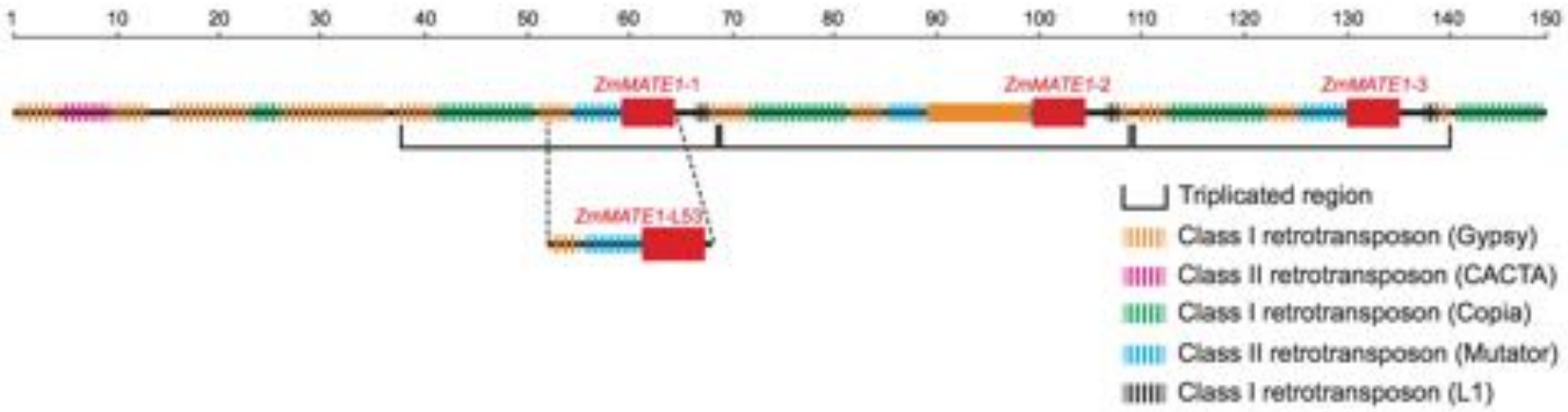
Assembly	Contig N50
Illumina Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,444
PBeCR Reads 6.3x 2146bp ** MiSeq for correction	13,600
Illumina Mates 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	13,696
PBeCR + Illumina Shred 6.3x 2146bp ** MiSeq for correction 51x 2x50bp @ 4800	25,108

In collaboration with McCombie & Ware labs @ CSHL

Long Read CNV Analysis

Aluminum tolerance in maize is important for drought resistance and protecting against nutrient deficiencies

- Segregating population localized a QTL on a BAC, but unable to genotype with Illumina sequencing because of high repeat content and GC skew
- Long read PacBio sequencing corrected by CCS reads revealed a triplication of the ZnMATE1 membrane transporter

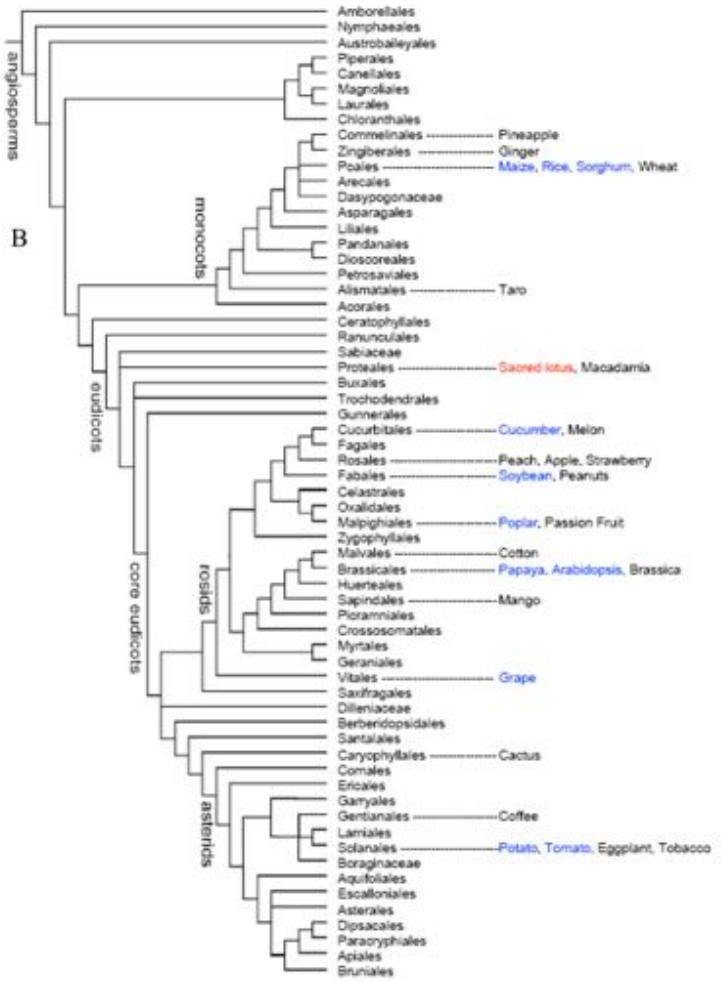


A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils

Maron, LG *et al.* (2012) *Under review.*

Wheat Sequencing

Aegilops tauschii



- One of the most important cereal crops in the world
- A. tauschii is one of the three ancestral species (DD) in modern bread wheat (*Triticum aestivum*)
  - Also looking to sequence other 2 species, and bread wheat
  - ~4.5Gbp Genome Size

In Collaboration with McCombie and Ware labs

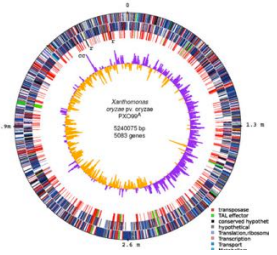
Wheat Sequencing & Assembly

Technology	Read Length	Fragment Length	Coverage
Illumina	100 bp	180 bp	69x
	100 bp	300 bp	50x
	35 bp	2,000 bp	6.6x
	35 bp	5,000 bp	6.5x

Assembly	Count	Max	N50	Sum
Scaffolds	97,313	2.76 Mbp	23,193	1.36 Gbp (30%)
Contigs	556,767	165 kbp	4,623	928 Mbp (20%)

- Poor coverage of the genome due to extreme repeat content
  - Had to downsample reads to fit into RAM
  - Randomly discard reads covered by kmers that occur more than 500 times
- Ramping up for PacBio long reads

Assembly Summary



Assembly quality depends on

1. ***Coverage***: low coverage is mathematically hopeless
2. ***Read length***: longer reads resolve repeats and complex regions
3. ***Read Quality:*** need clean libraries, clean reads

- PacBio RS has capabilities not found in any other technology
  - Substantially longer reads -> span repeats
  - Unbiased sequence coverage -> close sequencing gaps
  - Single molecule sequencing -> haplotype phasing, alternative splicing

- PacBio enables highest quality de novo assembly
  - Longer reads have fundamentally more information than shorter reads
  - Because the errors are random we can compensate for them informatically
  - Software available open source at http://wgs-assembler.sf.net

Acknowledgements

Schatz Lab

Giuseppe Narzisi
Shoshana Marcus
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
James Gurtowski
Alejandro Wences
Hayan Lee
Rushil Gupta
Avijit Gupta
Shishir Horane
Deepak Nettem
Varrun Ramani
Eric Biggers

CSHL

Hannon Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

NBACC

Adam Phillippy
Sergey Koren

Cornell

Lyza Maron

Everyone at PacBio

Thank You!

http://schatzlab.cshl.edu/