Assembling Crop Genomes with Single Molecule Sequencing

Michael Schatz

Feb 22, 2013
AGBT, Marco Island, FL

Jason Chin @infoecho Feb 18

What is the longest single contig that one has ever seen from a de Bruijn graph assembler without PE or jumping library?

Expand

Michael Schatz @mike_schatz Feb 18

@infoecho around 100kbp without looking very hard. I suspect you could get >1Mbp from a well behaved microbe. infinite from a random genome.

Expand Reply Delete Favorite More

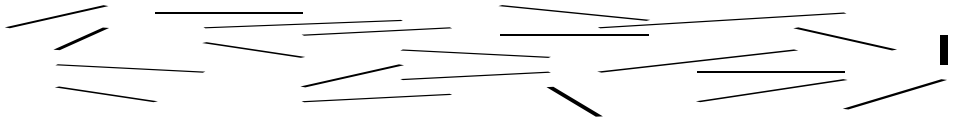
```
$ perl -e 'print ">random\n"; @D=split //,"ACGT"; \
   for (1...100000000){print $D[int(rand(4))];} \
   print "\n"' | fold > random.fa
```

```
$ wgsim –r 0 -e 0 -N 50000000 -1 100 -2 1 \
   random.fa random.reads.fq /dev/null
$ SOAPdenovo-63mer all –s random.cfg -K 63 -o random.63
```

```
$ getlengths random.63.contig
1 99999990
```

Assembling a Genome

1. Shear & Sequence DNA



2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

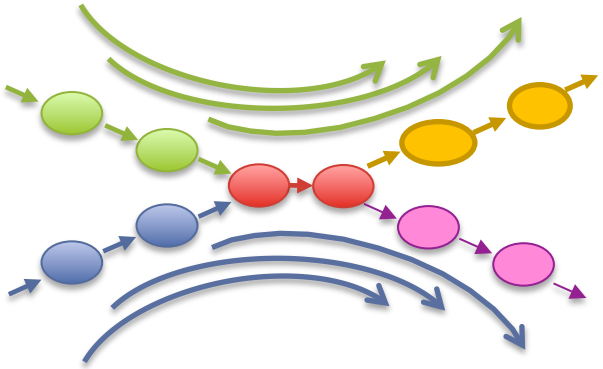
3. Simplify assembly graph

random genome

real genome

Ingredients for a good assembly

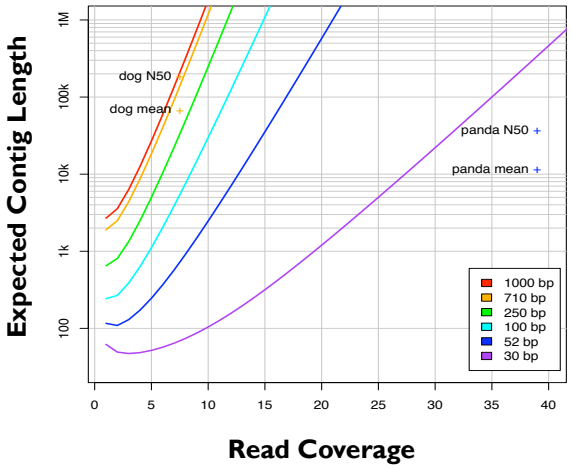
Read Length



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

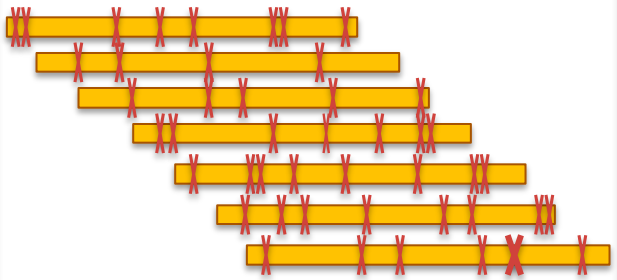
Coverage



High coverage is required

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

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

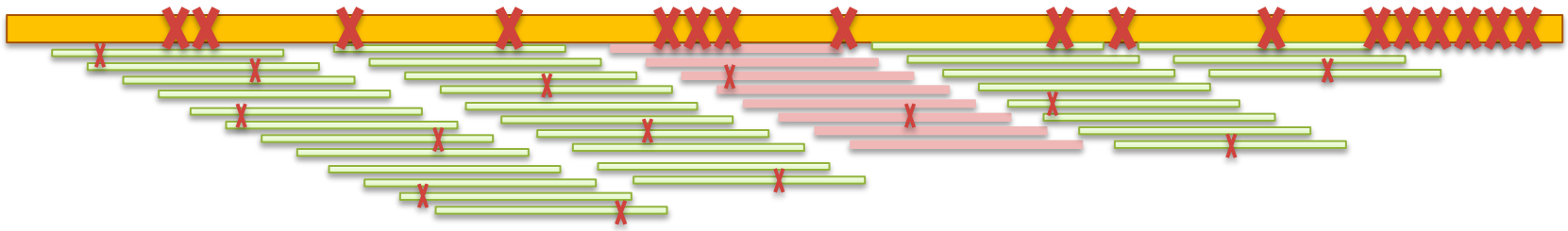
Lower throughput (600Mbp/day)
Lower accuracy (~90%)
Long reads (2-5kbp+)

PacBio Error Correction

http://wgs-assembler.sf.net

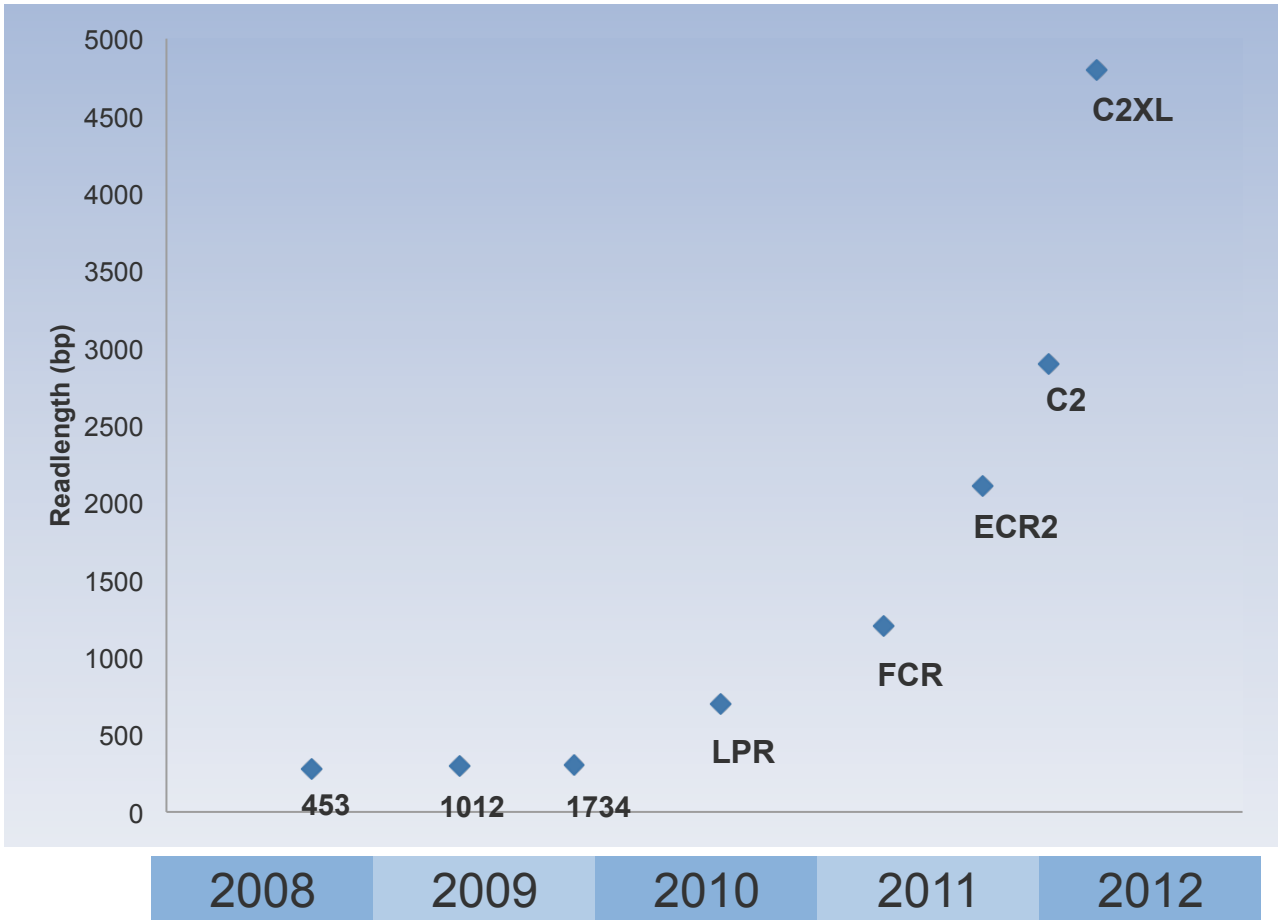
1. Correction Pipeline
    1. Map short reads to long reads
    2. Trim long reads at coverage gaps
    3. Compute consensus for each long read

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

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:

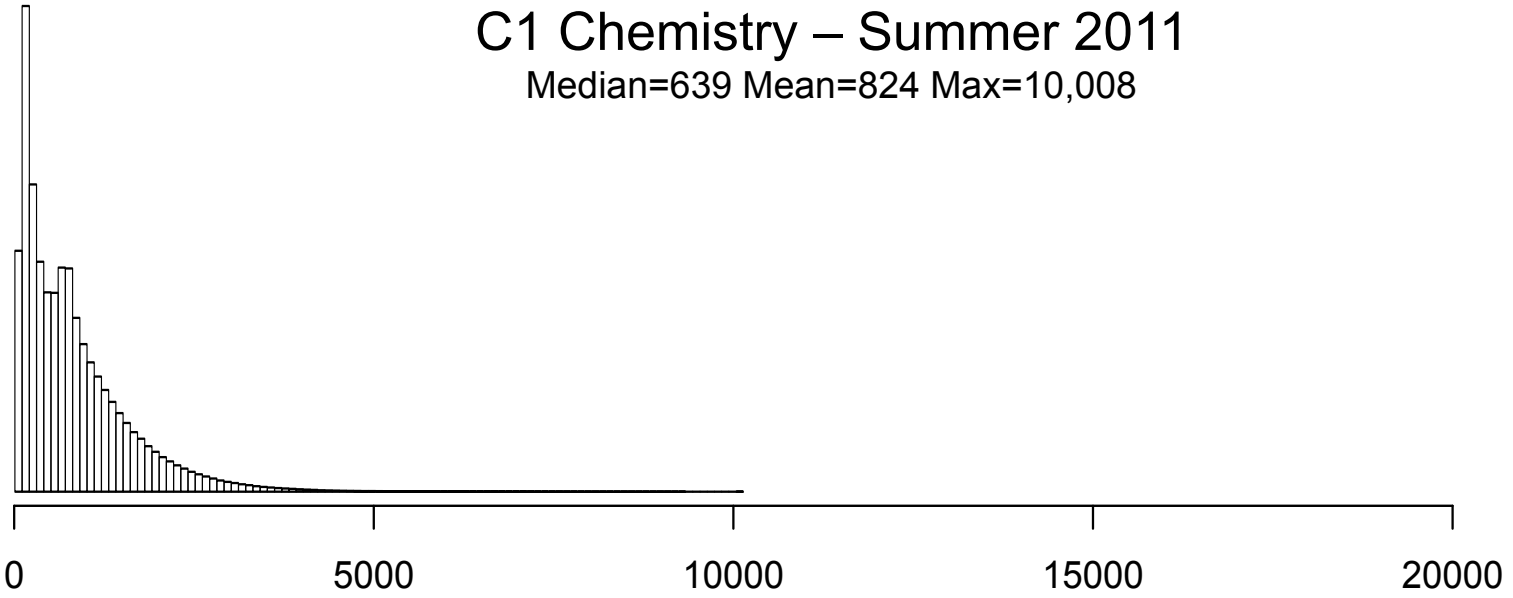
   Maximize subreads

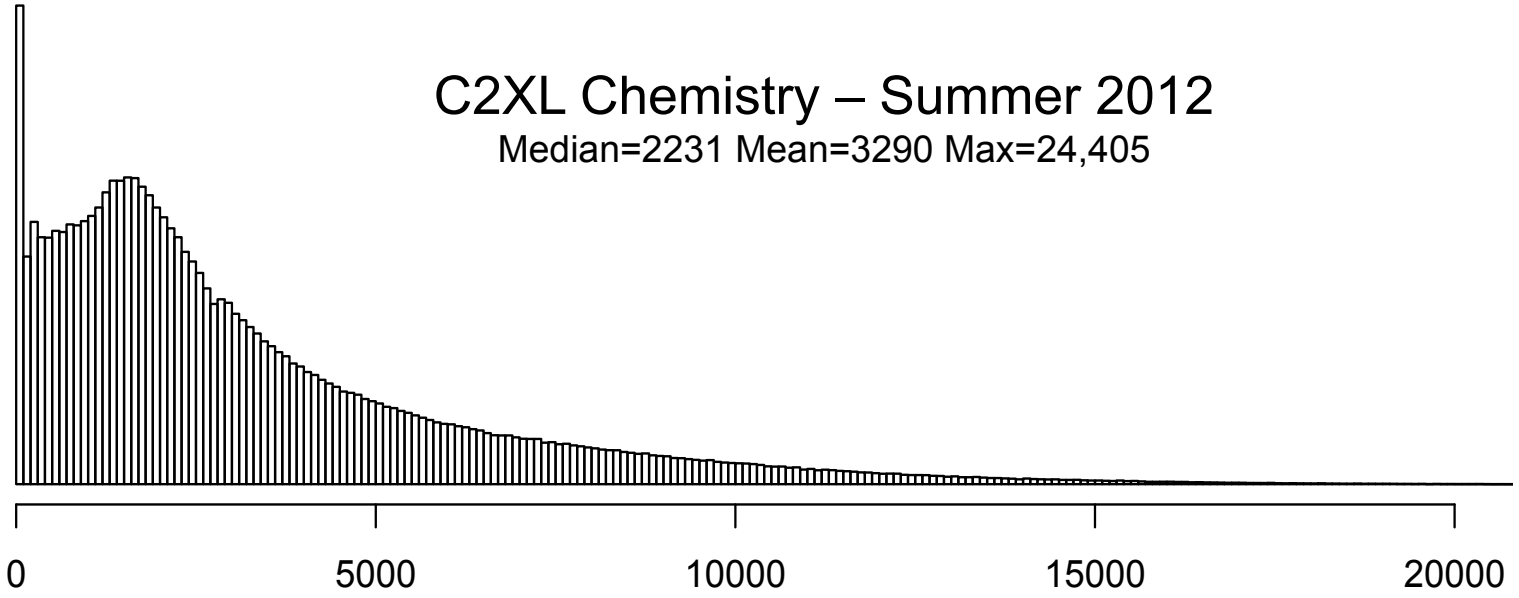
3. MagBead loading:

   Load longest fragments

See Eric Antonio's talk tomorrow at 9:30 for details

PacBio Long Read Rice Sequencing





Plant Genomics

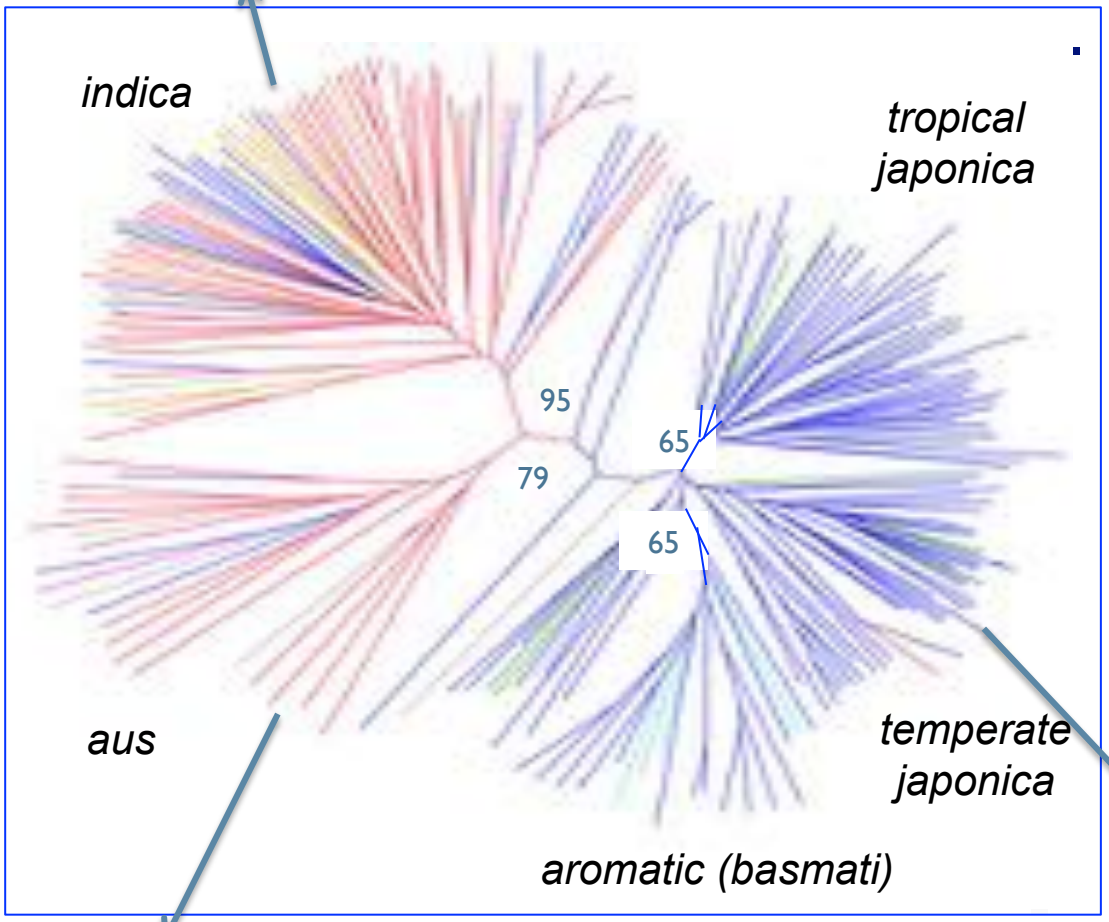
- Motivations
  - 15 crops provide 90% of the world's food
  - Responsible for maintaining the balance of the carbon cycles, soil from erosion
  - Promising sources of renewable energy
  - Plant byproducts used in many medicines
  - Model organisms for studying biological systems

- Challenges
  - Very large genomes, some many times larger than human
  - High repeat content, especially high copy retrotransposons
  - High ploidy, high heterozygosity

Population structure in *Oryza sativa*

3 varieties selected for *de novo* sequencing

IR64



DJ123

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

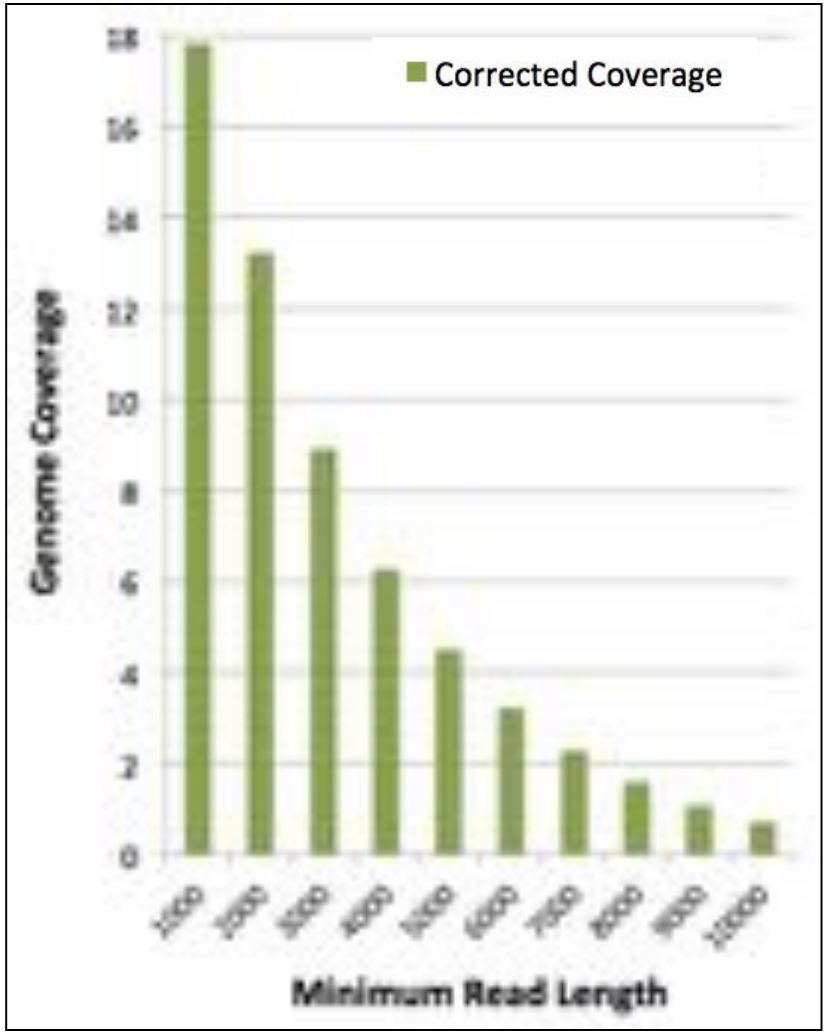
High quality BAC-by-BAC reference

- ~370 Mbp genome in 12 chromosomes
- About 40% repeats:
  - Many 4-8kbp repeats
  - 300kbp max high identity repeat (99.99%)
- Useful model for other cereal genomes

Nipponbare

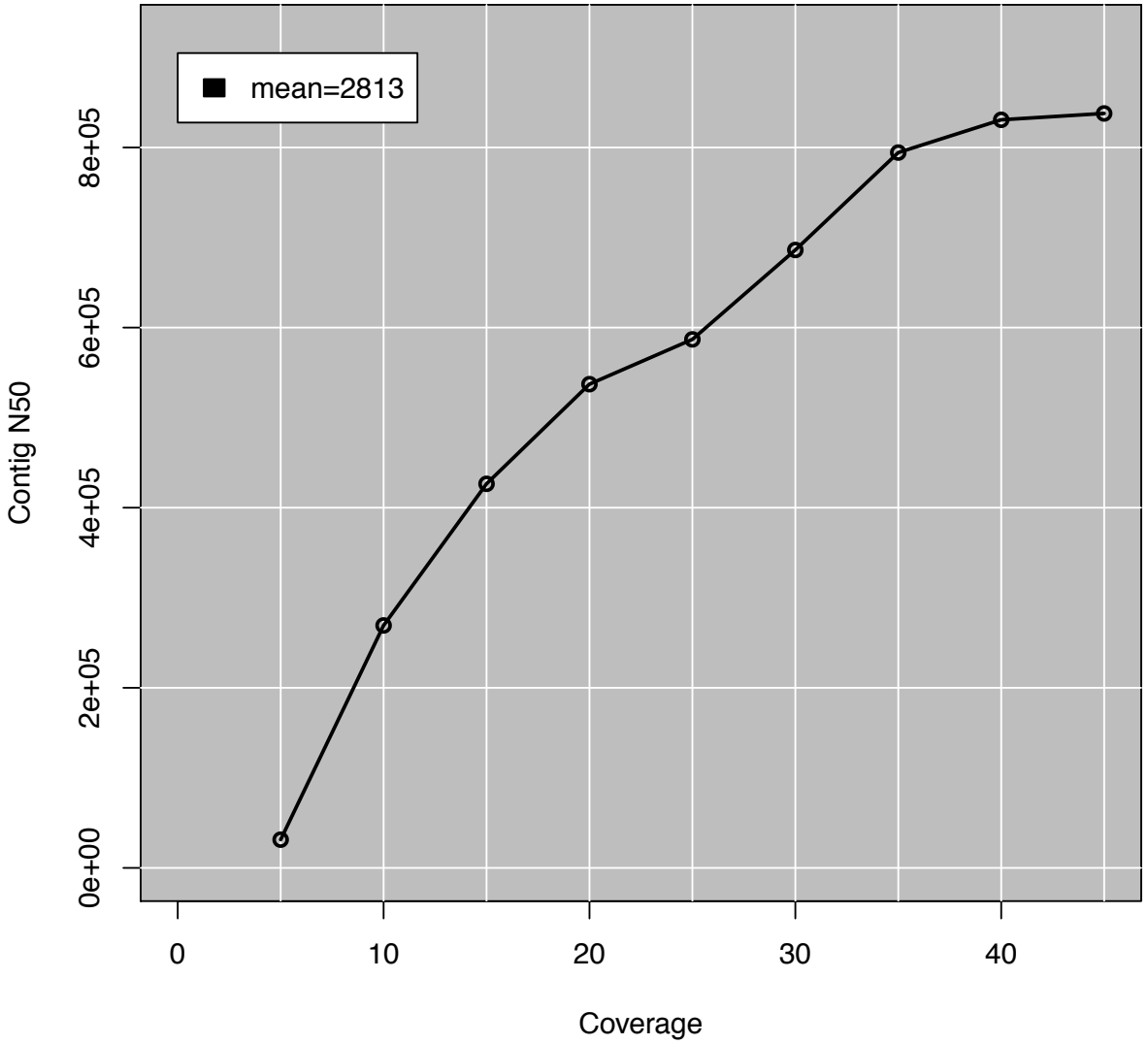
Preliminary Rice Assemblies

Assembly	Contig NG50
HiSeq Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248
PBeCR Reads 7x @ 3500 ** MiSeq for correction	25,724
PBeCR + Illumina Shred 7x @ 3500 ** MiSeq for correction 5x @ 3000bp shred	36,127



In collaboration with McCombie & Ware labs @ CSHL

Assembly Coverage Model

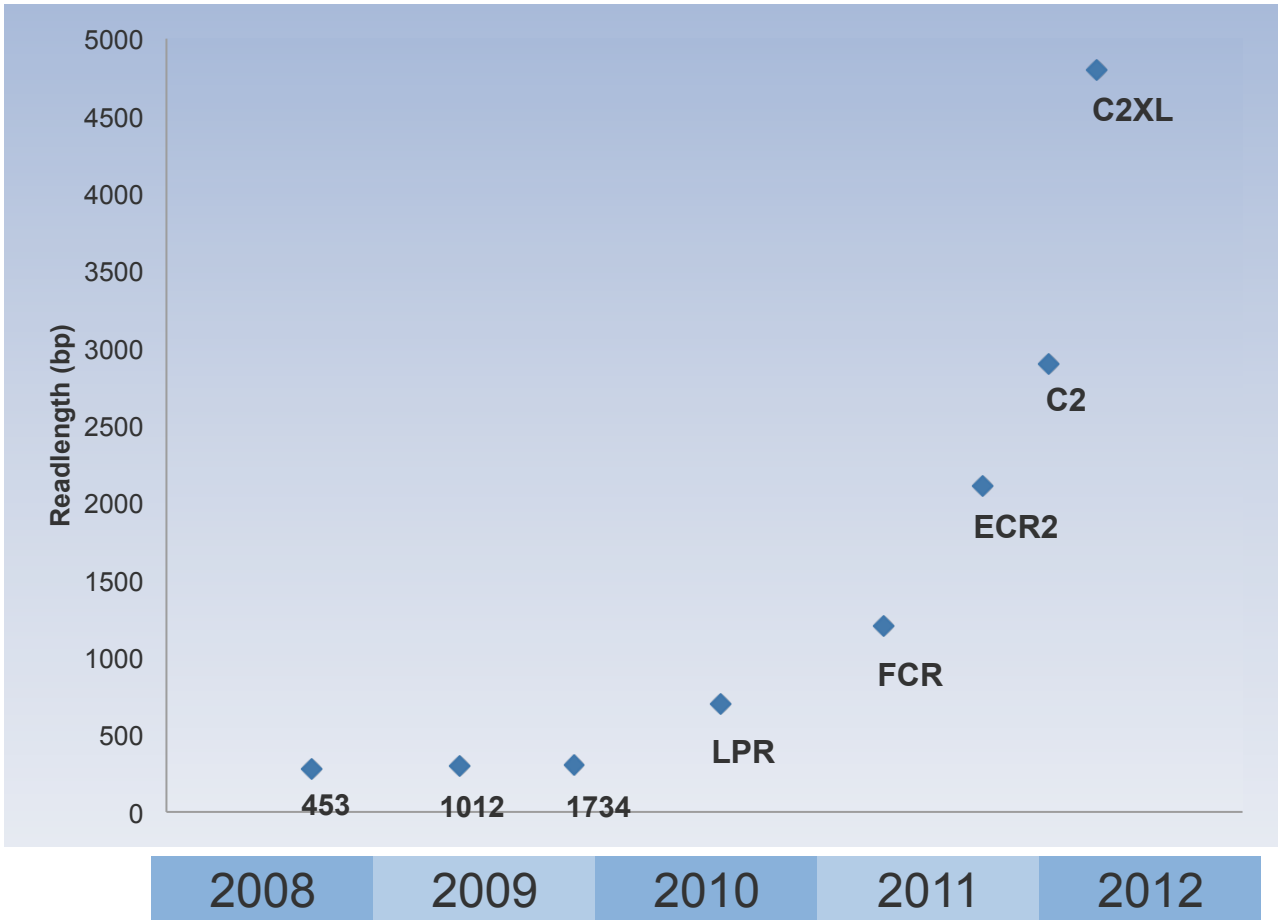


Simulate PacBio-like reads to predict how the assembly will improve as we add additional coverage

Only 8x coverage is needed to sequence every base in the genome, but 40x improves the chances repeats will be spanned by the longest reads

Assembly complexity of long read sequencing
Marcus, S, Lee, H, et al. (2013) *In preparation*

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:

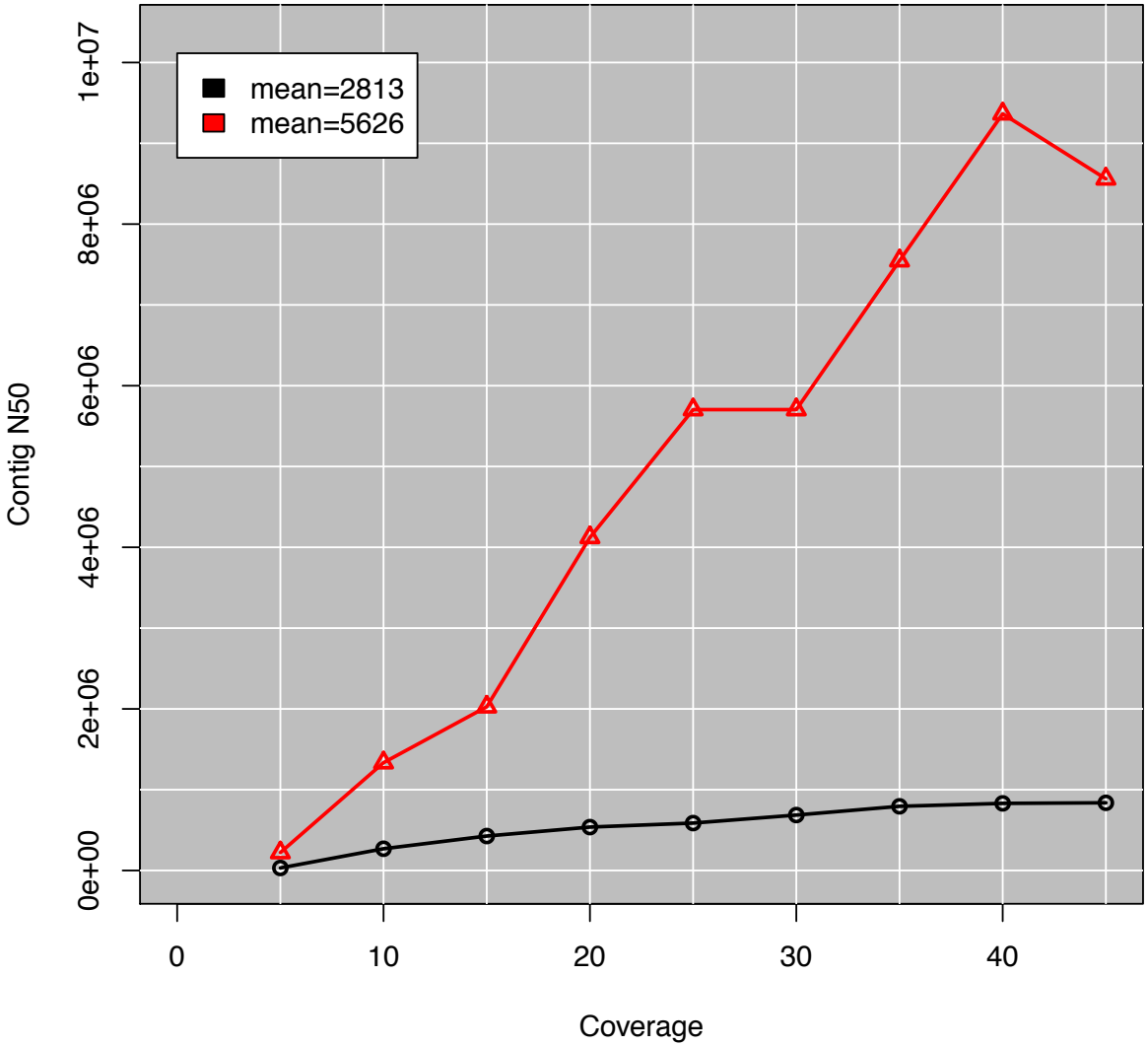
   Maximize subreads

3. MagBead loading:

   Load longest fragments

See Eric Antonio's talk tomorrow at 9:30 for details

Speculation for AGBT14

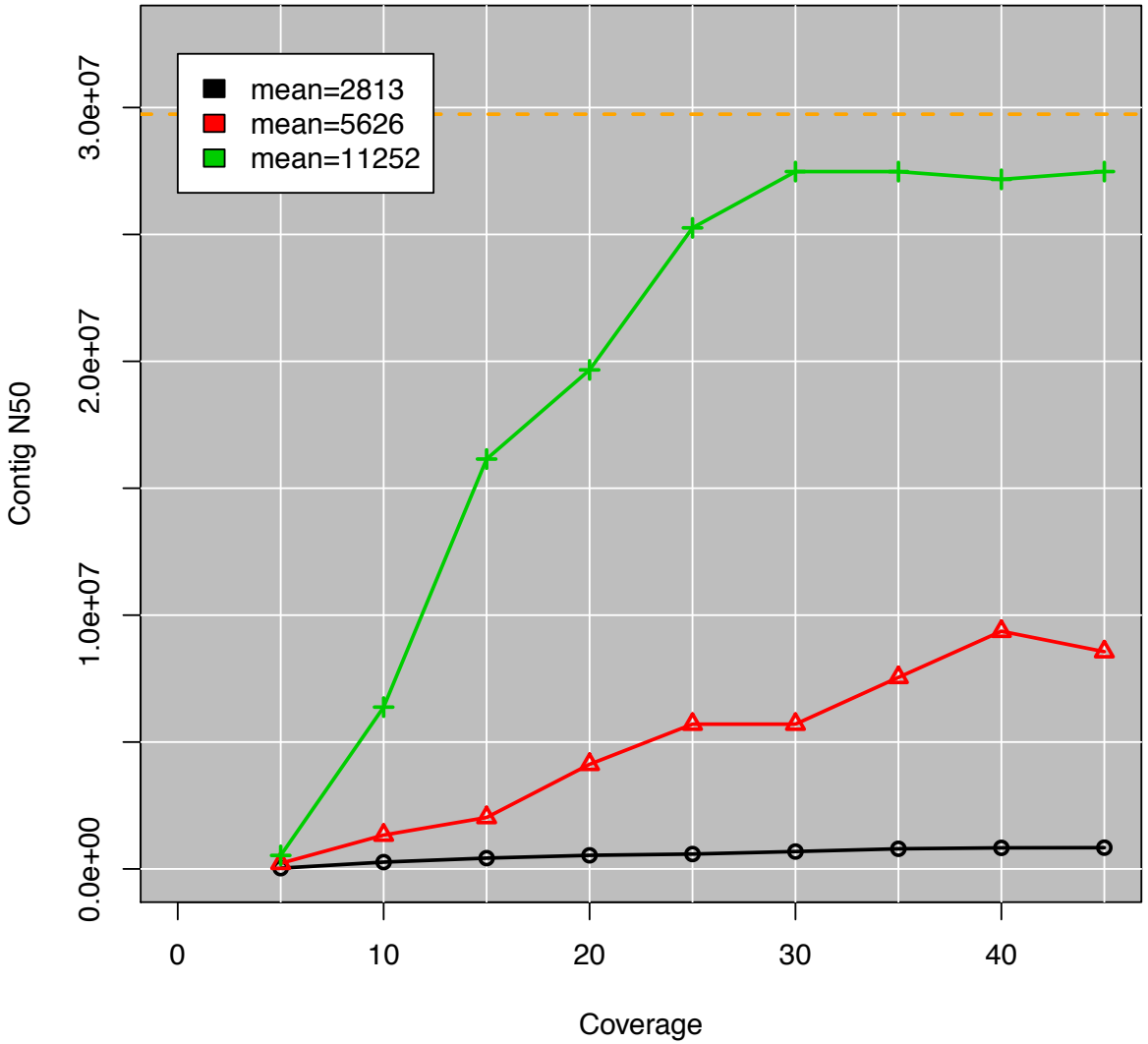


Doubling the average read length dramatically improves the assembly quality

- Able to span a larger repeats and lock contigs together

Expect to see contig N50 values over 1Mbp very soon, even in very complicated plant and animal species

Speculation for AGBT14



With PacBio-like reads averaging 11.2kbp (4x current), we should be able to assemble almost every chromosome arm of rice into single contigs

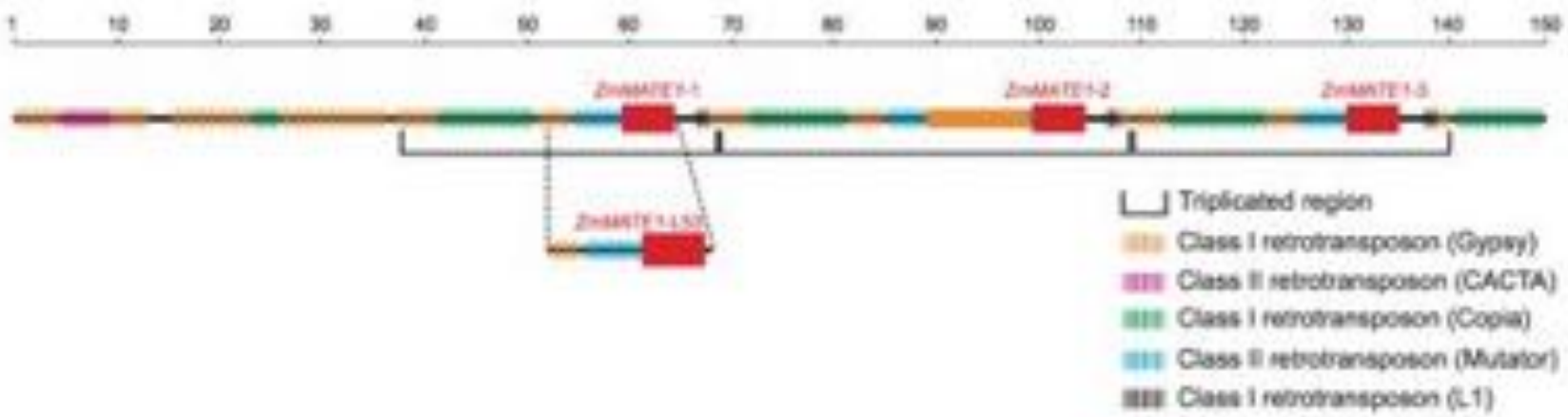
- The 300kbp near perfect repeat is the only exception

Even with the current assembly, we are seeing new genes and other sequences missing in the "high quality" BAC-by-BAC reference genome

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) *PNAS. In press*

Assembly of Complex Crop Genome

- Hybrid assembly let us combine the best characteristics of 2nd and 3rd gen sequencing
- Long reads and good coverage are the keys to a good assembly
  - With good coverage, we can "polish" out errors
  - Single contig de novo assemblies of entire microbial chromosomes is now routine
  - Single contig de novo assemblies of entire plant and animal chromosomes is on the horizon
- We are starting to apply these technologies to discover significant biology that is otherwise impossible to measure

Acknowledgements

Schatz Lab
Shoshana Marcus
Hayan Lee
James Gurtowski
Alejandro Wences

CSHL
McCombie Lab
Ware Lab

NBACC
Adam Phillippy
Sergey Koren

Cornell
Lyza Maron

Everyone at PacBio

NSF

National Human Genome Research Institute

Thank You!

http://schatzlab.cshl.edu
@mike_schatz
#agbt13