Sequencing

Kristian Stevens
Mark Crepeau
Charis Cardeno
Charles H. Langley

University of California, Davis
Evolution & Ecology

- Two parallel and complementary approaches to the *P. taeda* genome.
  - WGS (short fragments from a haploid & large fragment end-sequencing from the diploid).
  - Fosmid pools (lower complexity and effective haploidy.
- Shared methods – Illumina GAIIx and HiSeq.
- Different pipelines and timelines
- Short reports of progress on both.

Selecting the Megagametophyte

- Goal: deep (>40X) representative short insert libraries from a single haploid ($1N$) segregant.
- Libraries from DNA preps of 22 megagametophytes were prepared, sized and analyzed.
- DNA samples were genotyped to verify parentage.

Most of the tissue in a pine seed is the haploid megagametophyte.

Selecting the Megagametophyte

- 10 size-selected libraries from megagametophytes were sequenced.
- Quality control metrics were derived from alignments to three reference sequences:
  - *P. taeda* Chloroplast (Parks, *et al*. BMC-Biology, 2009)
  - BAC Sequences (Kovach, *et al.* 2010; Clemson U. and JGI-HAGSC)
  - Sequenced Transcription Units (dendrome.ucdavis.edu)

Comparing Libraries

- Interrogated libraries were all sequenced in the same flowcell lane with equivalent fragment sizes, read lengths, and error properties.
- To examine genomic sampling variance the BWA-aligned fragments were normalized to equivalent concentrations using random sampling.
- Only one read of each DNA fragment was used to avoid autocorrelation.

Comparing Libraries

- The variances in the number of reads aligned in non-overlapping windows were examined.
- Coverage of the BACs were highly sensitive to non-homologous repeat content.
- Coverage of the chloroplast exhibit substantial variation.
- Greatest weight was given to the coverage of the TU targets.

Summary of Results

Limited variation in the standard deviation of the number of reads aligned in non-overlapping windows.

	Chloroplast	TUs	BACs
Library ID	**StdDev**	**StdDev**	**StdDev**
MGP_2_5	4.0	3.5	32.4
MGP_10_5	4.5	3.3	32.0
MGP_11_6	5.2	4.7	32.8
MGP_7_5	4.2	3.3	32.7
MGP_4_5	4.7	3.4	32.8
MGP_3_400	6.2	3.1	32.2
MGP_8_400	8.1	3.3	33.7
MGP_5_6	4.6	3.3	32.2
MGP_12_5	6.7	3.3	32.2
MGP_9_400	6.3	3.4	32.2

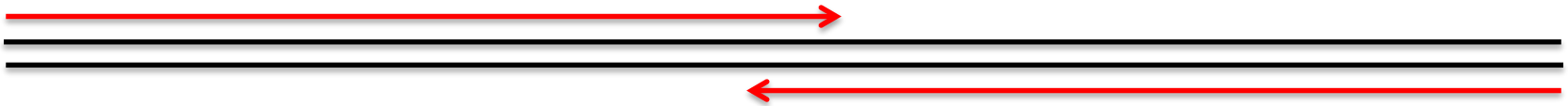
Summary of Results

Library ID	Insert Size Mean	Std	CV	Chloroplast StdDev	%	TUs Covered	BACs %	G+C %
MGP_2_5	277	14	5%	6.05	1.7	388	43	38.34
MGP_7_5	273	12	4%	6.21	1.2	417	45	39.05
MGP_5_6	275	13	5%	6.04	2.3	347	41	38.29
MGP_10_5	270	12	4%	6.81	3	206	41	38.83
MGP_11_6	269	12	4%	7.55	4.2	285	42	39.37
MGP_4_5	271	12	4%	7.61	1.7	356	45	39.39
MGP_3_400	272	13	5%	8.76	1.7	328	44	39.17
MGP_12_5	265	12	5%	10.33	1.4	364	44	39.25
MGP_9_400	267	13	5%	9.24	1.3	328	44	39.37

Orientation & Size OK > 99% for all Libraries

Overlapping reads

- Genomic assemblers perform better with overlapping reads.



- Factors affecting the yield of overlapping reads:
  - GAIIx high quality read length.
  - Mean and variance of fragment size.

Overlapping reads

- We determined that our instrument can routinely deliver >Q20 over 160 bp (read 1) and 156 bp (read 2).
- Repeatedly produce libraries with fragment size CVs < 4% .
- With the chosen library and these read lengths the yield of overlapping reads is > 98% .

Megagametophyte WGS Coverage

HiSeq Coverage

	Run 1 Lane 6	*Run 2 Lanes 1-8*	*Run 3 Lanes 1-8*	*Totals*
Read Pairs x 10^6	126	1135	1304	**2565**
BP x 10^9	38.6	229	326	**594**
Fold Coverage	1.6	9.6	13.6	**25**

GA2X Coverage

	Totals
Lanes Sequenced	***23.5***
Read Pairs x 10^6	**907**
BP x 10^9	**284.9**
Fold Coverage	**11.87**

Prototypic Fosmid Pool

Our current working *P. taeda* fosmid pool size is 500 clones. This is already optimal in the sense that

- Complexity of individual pools is well within the specification of available assemblers.
- Effectively haploid facilitating assembly ($2N$ fosmid library but $1N$ pool).
- Near (if not within) budget.

Prototypical Fosmid Pool

- With an expected mean insert size of 37 kbp, the prototypical 500 fosmid pool consists of 18.5 Mbp of *P. taeda* genomic DNA.
- Additional complexity due to fosmid vector and *E. coli* host is filtered prior to assembly.
- This presents a modest-to-standard sized target genome for available assemblers.

Prototypical Fosmid Pool

- *E. coli* and fosmid vector contamination is manageable.
  - Across three libraries estimated *E. coli* contribution to sequenced DNA ranged from 3.64% to 3.90% with a mean estimate of 3.75% .
  - Fosmid vector contribution ranged from 14.5% to 15.8% with a mean estimate of 15.3%

Prototypical Fosmid Pool

- The current total non-target overhead is a low 19.1%.
- Fosmid Pool Libraries Constructed and Sequenced (*)
  - Short insert sizes (bp):

    250, 260, 270,* 280, 290, 400*, 500, 600*, 700;

  - And a large-fragment (≈ 3 kbp) “jumping” library insert size: ~ 3 kbp

Assembly Results for the first of twelve 500 fosmid pools

Assembler	stat	count	quartiles q1	quartiles q2	quartiles q3	n50	sum
Allpaths-LG	scf	987	2499	7781	30271	26298	14×10^6
	ctg	1524	2355	6031	12509	10324	14×10^6
	scf30K+	248	33595	35682	38361	30114	9×10^6
MSR-CA	scf	2162	506	1375	9224	14753	15×10^6
	ctg	3519	503	1339	5000	6826	14×10^6
	scf30K+	136	32603	35087	38119	30147	5×10^6
SOAP	scf	3251	123	185	495	33389	15×10^6
	ctg	23873	76	175	348	1515	15×10^6
	scf30K+	322	33907	35766	38683	33389	12×10^6

Daniela Puiu and Steven Salzburg

500 x 38 x10 kbp = 19×10^6 bp

Fosmid Pool Sequencing Pipeline

- Fosmid pool DNA received from CHORI
- Quantify DNA by fluorescent dye binding assay

Short Insert

- Aliquot 5 ul DNA
- Sonicate to fragment DNA
- End repair fragments
- A-tail fragments
- Ligate Illumina multiplex adapter
- Size select adapter-ligated fragments by agarose gel electrophoresis
- QC and quantitate using Agilent Bioanalyzer
- Enrich 10 ng size selected fragments using 10 cycles PCR
- QC and quantitate enriched library using Agilent Bioanalyzer
- Sequence

Long Insert

- Pool DNAs to create equimolar pool of pools
- Aliquot 10 ul DNA
- Fragment DNA using HydroShear
- End repair and biotinylate fragments
- Size select by agarose gel electrophoresis
- QC and quantitate using Agilent Bioanalyzer
- Circularize size selected fragments
- Digest un-circularized DNA
- Sonicate to fragment circularized DNA
- Bind biotinylated fragments to streptavidin beads
- End repair fragments
- A-tail fragments
- Ligate Illumina multiplex adapter
- Enrich adapter-ligated fragments using 18 cycles PCR
- Remove enriched library from beads
- Size select enriched library by agarose gel electrophoresis
- Sequence

FP Pipeline Development

- Reducing the amount of non-target overhead.
- Choosing the best assembler
- Choosing the optimal mix of short and long inserts for a fixed cost per pool
- Increasing pool size without degrading assembly quality.
  - Nested experiment of 500, 1000, 2000 fosmids.

The End