Sequencing and assembly of the 22-gigabase genome of loblolly pine

Daniela Puiu1, Steven L. Salzberg1 , Aleksey Zimin2, James Yorke2

1 Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland.
2 Institute for Physical Science and Technology,University of Maryland College Park.

Abstract

The **loblolly pine** is the most widespread and commercially important pine of the Southeast US. The **Loblolly Pine Genome Project (LPGP)** is part of the USDA-funded **PineRefSeq** project whose aim is the sequencing the three largest genomes ever: **loblolly pine** (22Gb), **sugar pine** (33Gbp) and **Douglas-fir** (18Gbp). The large genome size and high repeat content makes conifer sequencing very challenging. The PineRefSeq goal is the development of high quality reference genome sequences and model approaches for sequencing other large complex genomes.

Introduction

The **loblolly pine** 22Gbp diploid genome (n=12) has been sequenced using the Illumina technology from a combination of whole genome shotgun and pooled fosmid libraries. 13 billion reads and 1.7 trillion bases have been generated. Six universities and research institutes have been involved in this project. The assembly has been done in collaboration by the **University of Maryland** and **Johns Hopkins University**. Our main goal were the library evaluation and selection as well as the development of a high quality assembly. In this poster we are going to present the loblolly pine whole genome assembly results.

Background

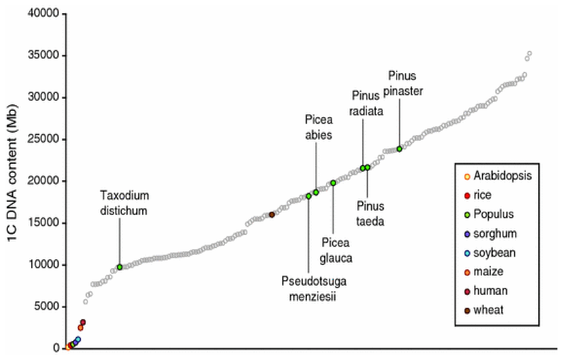


Figure 1: Genome sizes of 181 conifers (Murray et al. 2004) and select angiosperms (Bennett and Leitch 2005) with complete- or partially-sequenced genomes. Angiosperms are represented by colored circles. Conifers are represented by gray circles. A few socioeconomically important conifer species, are represented by black circles with labels.

Data analysis

Quality&base composition: fastx toolkit

- Kmer countig: jellyfish, kmerfreq =>genome size estimation
- Adapter & low quality trimming: ea-utils
- Contamination removal: bwa alignment to contaminant db.
- Error correction: QuORUM (part of MaSURCA)
- Library insert & complexity estimation

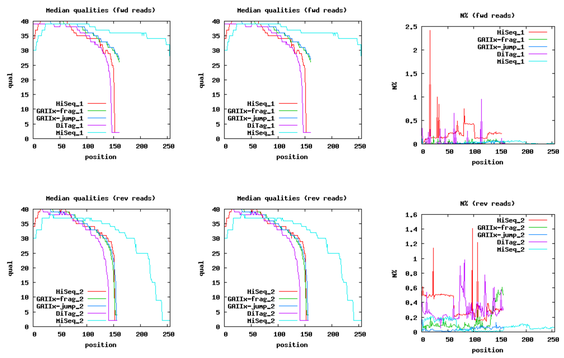


Figure 2: Quality& Base Composition of 100K sampled mate-pairs

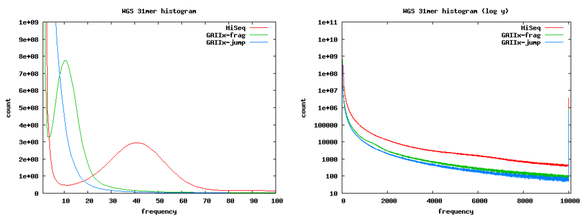


Figure 3: Kmer count analysis ; peak values => coverage; long tail=> repeats

Assembly results

When the PineRefSeq project has started in 2011, no genome assembler was able to handle 13 billion reads . The **MaSuRCA** assembler has been developed by the PineRefSeq team at UMD specifically for assembling such large genomes. **MaSuRCA** is based on the **Celera assembler** and uses an **overlap-layout-consensus** approach with **K-unitigs** and **superReads** which allows for a 100-fold data reduction

Input: QuORUM error-corrected reads
Parameters: K=79
Runtime & resources:3 months on a 64-core 1 TB machine

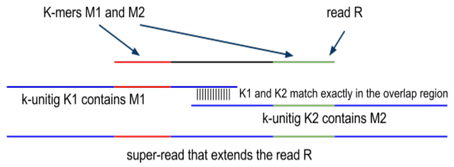


Figure 4: An example of a read whose super-read has two k-unitigs

	Number 10^6	N50 10^3	Max 10^6	Sum 10^9
Contigs	18.6	8.2	0.2	20.1
Scaffolds	16.4	30.7	5.0	22.5
Super-scaffolds	8.3	86	5.1	22.3

Table 2; MaSuRCA v1.0 assembly statistics; N50 was computed based on a 22Gbp genome size; The scaffolds were assembled into super-scaffolds using uncorrected long insert mates

SOAPdenovo2, released in 2012 has been used as an alternative assembler. It uses a **sparse** implementation of the **de-Bruijn graph.**

Input: QuORUM error-corrected reads
Parameters: K=79, map_len=63
Runtime & resources: 1 month on a 24-core 0.5 TB machine

	Number 10^6	N50 10^3	Max 10^6	Sum 10^9
Edges	113.1	0.4	0.05	16.6
Contigs	83.1	0.6	0.06	22.6
Scaffolds	18.7	54.7	0.9	19.5
Contigs	20.2	4.4	0.1	17.9
Scaffolds	17.1	54.7	0.9	19.5

Tables 3a.b; SOAPdenovo2 assembly statistics; before& after gap closing

Optimization

SOAPdenovo2 pregraph_sparse which only takes ~3 days to run, it has been used for K-mer optimization. SOAPdenovo2 scaffolder was used for super-scaffoling the MaSuRCA scaffolds.

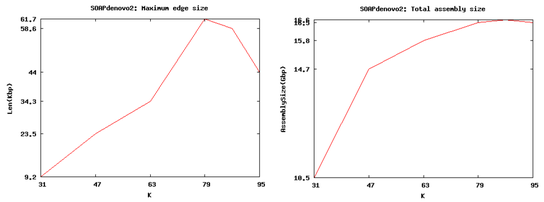


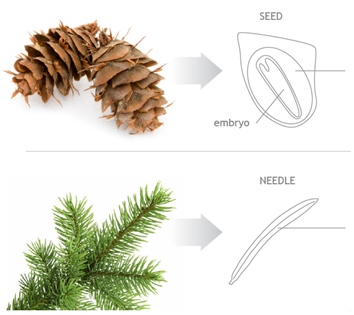
Figure 5:K-mer optimization; K~=79 => longest edges & largest assembly

Conclusions

Using the MaSuRCA assembler we have been able to generate a 22.5Gbp loblolly pine assembly with an N50 contig size of 8.2Kbp and N50 scaffold size of 30.7Kbp.

This assembly is by far superior to the 19.5Gbp one generated by SOAPdenovo2 whose original N50 contig size was only 0.6Kbp. Some of faster SOAPdenovo2 steps were used in parameter optimization and assembly improvement.

Sequencing



	#Lanes	Mate pairs 10^6	InsertLen bp	Bases 10^9	Cvg
HiSeq*	33	4845	273-565	1,177	49
GAIIx-frag*	35	975	209-637	303	13
GAIIx-jump	50	863	1,304-5,466	272	11
DiTag	45	52	38000	16	0.7
MiSeq*	4	14	400-700	7	0.3
Total	167	6,749	209-38,000	1,775	74

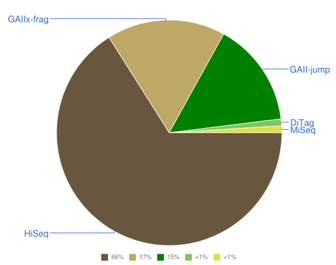


Table 1: Dataset composition: 83% of reads come from short insert haploid libraries (pine nut);
17% of the reads come from long insert diploid libraries (needle); <1% MiSeq data were used for assembly evaluation.

References

1. http://www.pinegenome.org/pinerefseq/
2. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, Luo et al. GigaScience 2012
3. The MaSuRCA Genome Assembler, Aleksey Zimin et all, under review

Acknowledgements

UC DAVIS PLANT SCIENCES
INDIANA UNIVERSITY
WASHINGTON STATE UNIVERSITY
UC DAVIS COLLEGE OF BIOLOGICAL SCIENCES
C·H·O·R·I Children's Hospital Oakland Research Institute
TEXAS A&M UNIVERSITY
USDA NIFA The National Institute of Food and Agriculture