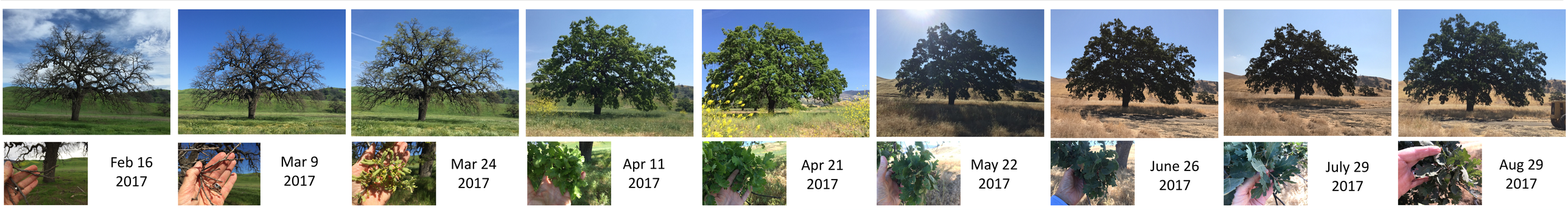
High quality genome assembly of a California endemic oak, *Quercus lobata*

Victoria L. Sork[1], Aleksey Zimin[2, 3], Daniela Puiu[2], Sorel Fitz-Gibbon[1], Paul F. Gugger[4], Matteo Pellegrini[1], Steven L. Salzberg[2]

[1]University of California, Los Angeles [2]Johns Hopkins University, Baltimore, MD [3]University of Maryland, College Park

[4]University of Maryland Center For Environmental Science, Frostburg. Email: vlsork@ucla.edu



1. Motivation & Goals

Oaks contribute significant economic and ecosystem benefits, comprising more biomass than any other tree genus in North American and being the most speciose genus across the Northern Hemisphere. Resource managers and forests geneticists need access to a high-quality genome for this genus to develop management, conservation, and restoration strategies.

Quercus lobata is also a great model system for study of adaptive genetic variation that will generate useful insight relevant to other tree species because populations have persisted for at least several hundred thousands years due to a lack of glaciation, allowing a strong signature of selection.

Goals:

1. To produce a high quality chromosomal assembly of *Quercus lobata* Née (Fagaceae)
2. To create a reference annotated genome to be used for in-progress evolutionary studies investigated genes involved in local adaptation.

2. Sequencing Data

Table 1. Data used for genome assemblies.

PacBio RS II reads: 80× coverage	Illumina short reads: ~200× coverage
12 runs; 53 flowcells; P6-C4 chemistry	Pair-Ends: ~150× cvg; 250 bp reads, 500 bp inserts
Read length: mean = 9 kb, max = 80 kb	Mate-Pairs: ~50× cvg; 3-12 k inserts

3. Assembly Strategy

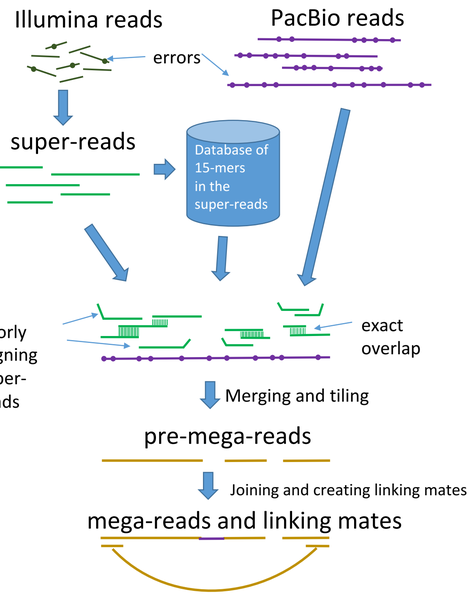


Figure 1. Initial assembly of Illumina and PacBio reads by MaSuRCA 3.2.1 mega-reads strategy (Zimin et al., 2017) to produce **Hybrid Primary and Alternative** haplotype assemblies (see Table 2).

- Low-error rate Illumina reads used to build longer super-reads (green lines).
- 15mers from super-reads aligned to PacBio reads (purple lines).
- Inconsistent super-reads are discarded
- Remaining super-reads merged along PacBio templates.
- Further independent merging to produce mega-reads (yellow) and generate linking mates across gaps.

Further assembly strategy

1. Aligned 83,644 assembled transcripts (Cokus et al.). Moved 317 uniquely aligned scaffolds from alternative to primary to produce **Hybrid+Transcript Primary and Alternative.**
2. Additional scaffold merging based on the unique sequence alignments on Hybrid+Transcript Primary (custom script) to produce **Hybrid+Transcript Primary Merged or Version 2** assembly.
3. Dovetail Genomics Scaffolding.
4. Removal of additional haplotype redundancy identified via Dovetail scaffolding, see Panel 4.

References

1. Bodénès C, Chancerel E, Ehrenmann F, Kremer A, Plomion C. 2016. High-density linkage mapping and distribution of segregation distortion regions in the oak genome. DNA Research, 23(2):115
2. Cokus SJ, Gugger PF, Sork VL. 2015 Evolutionary insights from de novo transcriptome assembly and SNP discovery in California white oaks. BMC genomics. 16:552.
3. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.;31(19):3210-2.
4. Sork VL, Fitz-Gibbon ST, Puiu D, Crepeau M, Gugger PF, Sherman R, Stevens K, Langley CH, Pellegrini M, Salzberg SL. 2016. First draft assembly and annotation of the genome of a California endemic oak. *Quercus lobata* Née (Fagaceae). G3: Genes | Genomes | Genetics. 11:3485-3495.
5. Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research.27:787-92.

Acknowledgments

Funded by National Science Foundation Plant Genome Research Program (IOS-1444611)
We thank Krista Beckly for lab work and Andy Luntz for field work. Sequencing and scaffolding was performed by: Genomics Sequencing Laboratory at The California Institute for Quantitative Biosciences, Berkeley; UC Davi Genome Center DNA Technologies and Expression Analysis Cores; and Dovetail Genomics., Santa Cruz, CA.

4. Dovetail Genomics Scaffolding

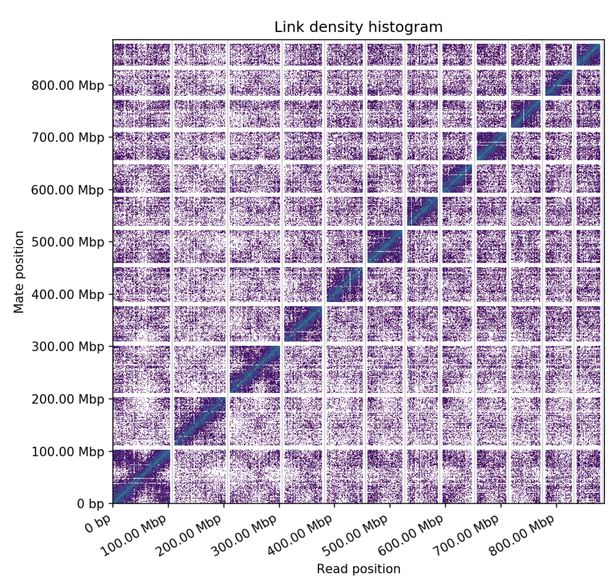


Figure 2. Dovetail assembly of the 12 largest scaffolds, likely corresponding to 12 chromosomes (covering 96% of sequence).

- Dovetail Genomics Hi-C (Chicago) library and HiRise scaffolding increased NG50 scaffold size from 1.9 Mbp to 74.6 Mbp.
- Directly adjacent contigs with > 50% syntenic alignment are considered to be redundant haplotypes leading to 14 Mbp moved to **Alternative** haplotype assembly.
- 6 of 12 longest scaffolds contain telomere sequences on one end.

5. PacBio and Illumina Assembly Results

Table 2. Assembly comparison (est. genome size = 730 Mbp)

Assembly	Scaffolds	Max (Mb)	Sum (Mb)	NG50 (Kb)
Hybrid Primary (Alternative)	3,599 (16,729)	6.7 (1.2)	818 (466)	1,228 (69)
Hybrid+Transcript Primary (Alternative)	3,916 (16,412)	6.7 (0.8)	872 (412)	1,228 (59)
Version 2: H+T Primary Merged	3,258	10.2	861	1,945
Version 3 clean final	2,028	104	847	74,600

6a. Genome Validation – *Q. robur* linkage groups

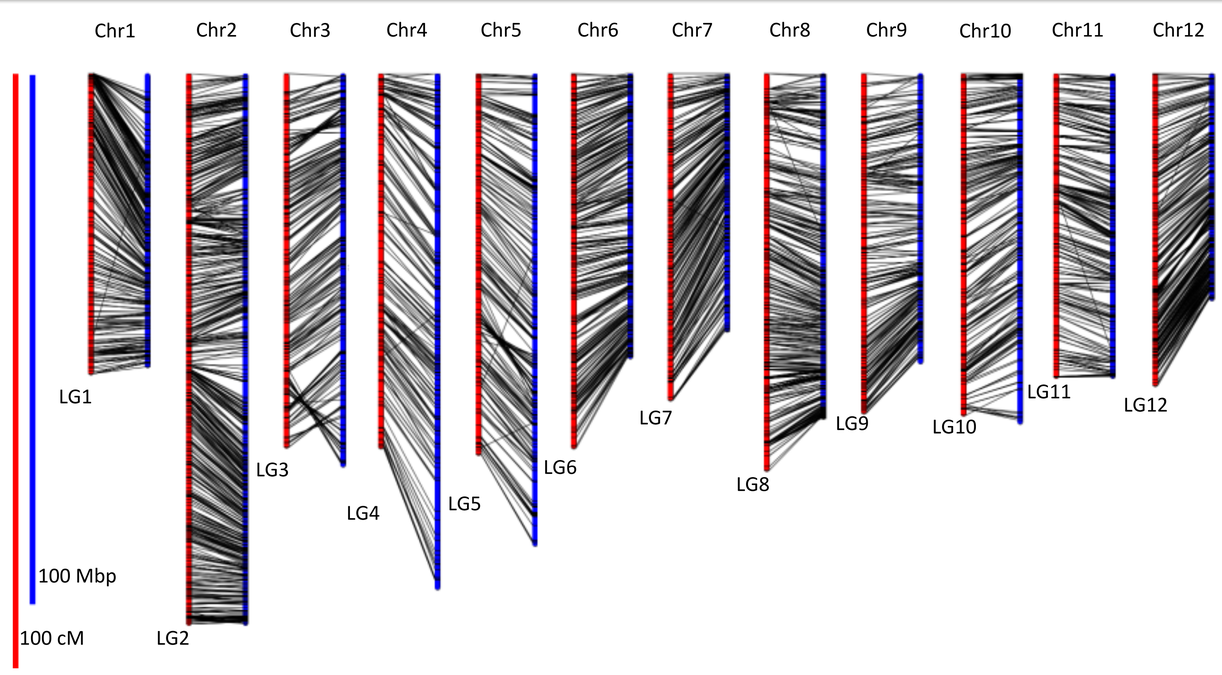
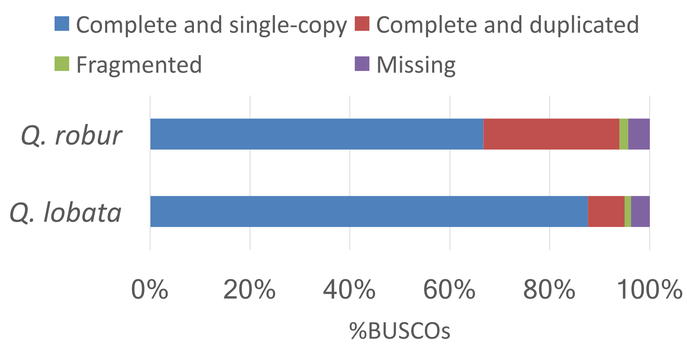


Figure 3. Twelve assembled scaffolds (blue; 813 Mbp; 96% of the assembled genome sequence) were tethered to twelve *Q. robur* linkage groups (red, Bodénès et al.) via 3,958 uniquely mapped genetic markers.

6b. Genome Validations – BUSCO Assessment

Figure 4. BUSCO (v3.0.2) search for 1440 **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs within the embryophyte (v9) database (Simão et al. 2015).

→ 90% of orthologs are complete and single copy.



Summary

1. Valley oak Version 3.0 is a high quality genome assembly with twelve large scaffolds corresponding to twelve chromosomes.
2. Correspondence of scaffolds to chromosomes is shown using *Q. robur* linkage maps.
3. Completeness and successful removal of haplotype redundancy is demonstrated by BUSCO analysis.

Next steps: Annotation of genome, and utilization with ongoing evolutionary studies.

Valley oak version 3.0 is available at valleyoak.ucla.edu