# Best Practices for Diploid Assembly of Complex Genomes Using PacBio: A Case Study of Cascade Hops

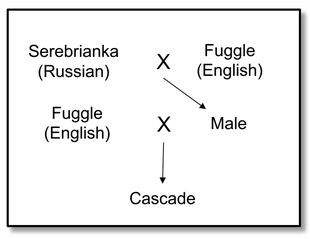
Sarah B. Kingan[1], Paul Peluso[1], David R. Rank[1], John A. Henning[2], David Hendrix[3]
1. PacBio, Menlo Park, CA; 2. USDA-ARS, Hop Breeding & Genetics program, Corvallis, OR;
3. Department of Biochemistry and Biophysics, Oregon State University

## Summary

A high quality reference genome is an essential resource for plant and animal breeding and functional and evolutionary studies. The common hop (*Humulus lupulus,* Cannabaceae) is an economically important crop plant used to flavor and preserve beer. Its genome is large (flow cytometry-based estimates of diploid length >5.4Gb[1]), highly repetitive, and individual plants display high levels of heterozygosity, which make assembly of an accurate and contiguous reference genome challenging with conventional short-read methods.

We present a contig assembly of Cascade Hops using PacBio long reads and the diploid genome assembler, FALCON-Unzip[2]. The assembly has dramatically improved contiguity and completeness over earlier short-read assemblies[3,4]. The genome is primarily assembled as haplotypes due to the outbred nature of the organism. We explore patterns of haplotype divergence across the assembly and present strategies to deduplicate haplotypes prior to scaffolding.

## Sample Prep and Sequencing




- Cascade hops widely used in American craft brewing.
- Backcross hybrid of English and Russian varieties.

**Figure 1. Genetic background of Cascade Hops[5].**

- Cascade clones grown in greenhouse under controlled conditions. ~100 ug young leaves harvested and placed immediately on ice.
- DNA extracted with modified Qiagen DNAeasy mini-prep kit. Chemical precipitation and glass hooks substituted for spin columns to prevent shearing.
- 2 SMRTbell libraries constructed from unsheared DNA with 10kb size selection (BluePippin). Quality assessed with pulse field gels.

**Table 1. PacBio SMRT Sequence Data.** Data were generated from Nov. 2016-July 2017 on PacBio RS II with P6-C4 chemistry, and Sequel 1.2.1 – 2.0 chemistry.

| PacBio SMRT Sequence Data | |
|---|---|
| Sequel Cells | 83 |
| PacBio RS II Cells | 126 |
| Number Subreads | 50,416,927 |
| Total Subread Bases | 342,733,627,375 |
| Prop. Sequel Data (bp) | 74 % |

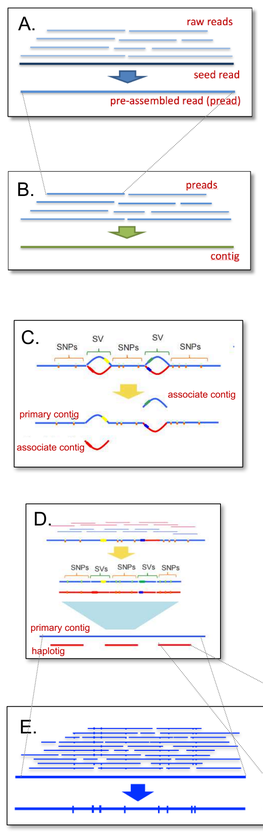
## Pre-Assembly Results

- Pre-assembly ("error-correction") generates high accuracy pre-assembled reads ("preads") by raw read overlap and consensus (Fig 2A).
- Pre-assembled yield = 75% = (total length preads / total length seed reads), used as metric for data quality and coverage.

| Statistic | Raw Reads | Seed Reads | PReads |
|---|---|---|---|
| Read Count | 35.2 M | 12.1 M | 13.8 M |
| Total Bases | 288 Gb | 182 Gb | 135 Gb |
| Coverage | 107 fold | 67 fold | 50 fold |
| Mean Length | 8 kb | 15 kb | 12 kb |
| N50 Length | 12 kb | 15 kb | 12 kb |

**Table 2. Summary Statistics of Reads. Raw reads**: filtering: subreads >500 bp, single subread from each ZMW. **Seed reads**: Raw reads longer than 10 kb (manual cut- off)**. Preads** may be shorter than seed reads due to low coverage regions that break or truncate seed reads. Coverage based on 2.7 Gb 1N genome size.

## Assembly Stages and Compute Resources




| Stage | Wall Time | Max N Jobs |
|---|---|---|
| A. Pre-Assembly | 79.57 h | 2,400 |
| B. Overlap | 20.57 h | 633 |
| C. FALCON | 2.45 h | 1 |
| D. Unzip | 80.93 h | 11,866 |
| E. Phased Polish | 130.16 h | 49,765 |

**Table 3. Wall times and number of jobs for five major stages of assembly.**

| Stage | Total CPU Time |
|---|---|
| Pre-Assembly DAligner | 37,704 h |
| Overlap DAligner | 9,470 h |
| Unzip Read Tracking | 8,087 h |
| Phased Polish (BLASR + arrow) | 9,236 h |

**Table 4. CPU time for selection of compute-intensive stages.** Job was run on SGE-managed cluster with 30 nodes, each with 512 Gb of RAM distributed over 64 slots. Up to 384 jobs were scheduled at a time.

**Figure 2. Five major stages of assembly.** See Table 3 for name and compute time for each stage.

## Haplotype Deduplication

Purge haplotigs[6] was run on FALCON-Unzip primary contigs with the goal of generating haploid version of the assembly prior to scaffolding. Contigs with >80% haploid coverage depth by length were flagged as "candidate haplotigs". Six iterations of homology assignment and duplicate haplotig removal was performed with BLAST and LASTZ alignments. Curated assembly length and BUSCO duplicate gene percentage indicate that highly divergent haplotypes have been retained.

| Statistic | Draft PacBio Primary | Draft PacBio Haplotigs | Curated PacBio Primary | Teamaker Short Read Asm |
|---|---|---|---|---|
| Total Length | 4.24 Gb | 1.35 Gb | 3.79 Gb | 1.77 Gb |
| No. Contigs | 11,705 | 38,060 | 8,652 | 861,745 |
| Contig NG50 | 867 kb | NA | 866 kb | 1.4 kb |
| Longest Contig | 8.25 Mb | 1.77 Mb | 8.25 Mb | 69 kb |

**Table 5. Genome Assembly Stats for PacBio assembly before and after Purge Haplotigs.** The Illumina-based Teamaker v1.1[4] assembly is shown for comparison. Flow cytometry 1N size is 2.7 Gb[1] and is the genome size used for NG50 statistics.

| BUSCO Status | Draft Assembly | Draft Primary Only | Curated Assembly |
|---|---|---|---|
| Complete | 89.3 % | 86.7 % | 85.0 % |
| Duplicated | 55.3 % | 37.7 % | 30.1 % |
| Missing | 8.8 % | 10.9 % | 12.8 % |
| Fragmented | 1.9 % | 2.2 % | 2.2 % |

**Table 6. Genome completeness assessment for PacBio assembly with BUSCO.** Embryophyta gene set (N=1440) run with BUSCO3[7].

## Assembled Genome Size Estimate

Purge haplotigs pipeline defines four coverage bins from raw read coverage histogram for the full draft assembly. The proportion of each contig with haploid and diploid coverage was used to estimate assembled genome length.

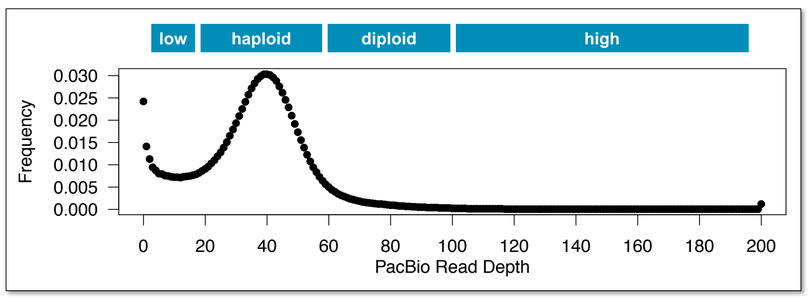



**Figure 3. Coverage of raw PacBio reads for draft FALCON-Unzip assembly.** Lack of diploid coverage peak indicates genome largely assembled as haplotypes (not collapsed diploid haplotypes, Fig 4). Raw reads >2500 bp were mapped to draft FALCON-Unzip assembly as part of phased polishing (Fig 2E).

| Coverage Bin | Total Length |
|---|---|
| Diploid | 297 Mb |
| Haploid | 4.16 Gb |
| 1N Length | 2.38 Gb |

**Table 7. Length of diploid- and haploid-assembled genome.**
1N length calculated as:
[(2 * Diploid Length) + Haploid Length] / 2.

## Haplotype Divergence

Haplotypes may be assembled as single primary contigs, primary-haplotig pairs, or primary contig pairs, depending on the level of heterozygosity in the genomic region.

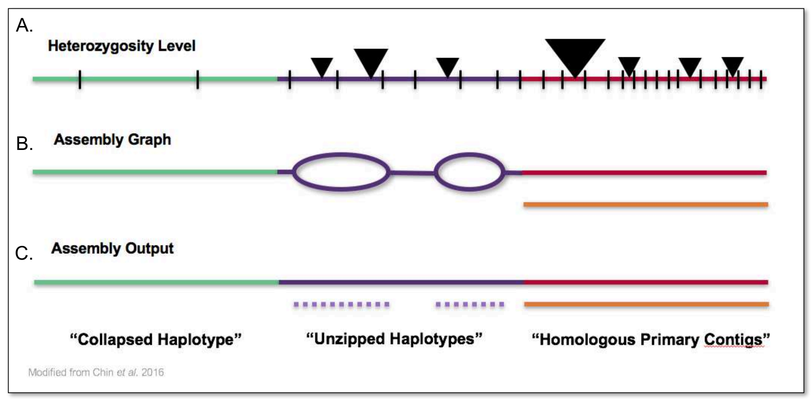



**Figure 4. Impact of heterozygosity on assembly output.** A. Low (green), moderate (purple), and high heterozygosity regions (red). SNPs: tic marks, structural variants (SVs): triangles. B. Overlap graph for contig assembly. C. Resulting contigs.

Haplotype divergence was compared for primary-haplotig pairs ("unzipped haplotypes") and homologous primary contigs. Haplotypes were aligned with nucmer[8] and analyzed with show-coords and Assemblytics[9]. Homologous primary contig pairs identified with purge haplotigs[6].

**Table 8. Summary statistics for classes of assembled haplotypes.** Mean alignment proportion: ratio of total alignment length to length of haplotig or shorter primary contig.

| Statistic | Unzipped Haplotypes | Homol. Prim. Contigs |
|---|---|---|
| Number Contig Pairs | 31,654 | 2,491 |
| Mean Alignment Proportion | 0.952 | 0.873 |

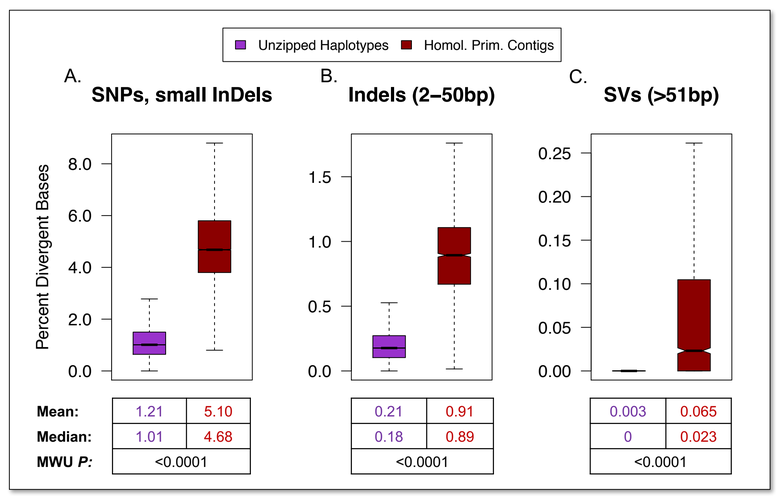



**Figure 5. Unzipped haplotypes display lower variation than homologous primary contigs.** A. Percent divergence in SNPs and small insertion-deletions. B. Percentage of bases in small indels (2-50 bp). C. Percentage of bases in SVs (>50 bp).

## Conclusion

- The PacBio FALCON-Unzip diploid assembly is both more complete (>2-fold increase in assembly length) and more contiguous (>600-fold increase in contig NG50) than previous short-read assemblies.
- Most assembled contigs are haplotypes due to high heterozygosity, with haplotype divergence ranging from 1-5% for SNPs and small indels.
- Haplotype deduplication removed ~450 Mb of haplotigs prior to scaffolding.

## References

1. Grabowska-Joachimiak et al. (2006) Genome size in *Humulus lupulus L.* and *H. japonicus* Siebold and Zucc. (Cannabaceae). ACTA Soc. Bot. Pol. 75(3), 207.
2. Chin et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 13(12), 1050.
3. Natsume et al. (2015). The Draft Genome of Hop ( Humulus lupulus ), an Essence for Brewing. Plant and Cell Phys. 56(3), 428.
4. Hendrix et al. unpublished. Teamaker v1.1 http://hopbase.cgrb.oregonstate.edu/resources
5. Lemmens and Gerard. (1998) The Breeding and Parentage of Hop Varieties. Brewers Digest. 73(5), 16.
6. Roach M (2017) https://bitbucket.org/mroachawri/purge_haplotigs (accessed Dec 4, 2017)
7. Simão et al. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 31(19), 3210-3212.
8. Marcais G (2017) Mummer4 https://github.com/mummer4/mummer
9. Nattestad M, Schatz MC. (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 32(19), 3021.

Acknowledgements

The authors would like to thank Emily Hatas, Greg Concepcion, and Richard Hall at PacBio.