An Improved Genome Assembly of the European Aspen *Populus tremula*

Bastian Schiffthaler <bastian.schiffthaler@umu.se>[1], Nicolas Delhomme[2], Carolina Bernhardsson[4], Jerry Jenkins[3], Stefan Jansson[1], Pär Ingvarsson[4], Jeremy Schmutz[3], and Nathaniel Street[1]

[1]Umeå University, dept. of Plant Physiology, Umeå, Sweden
[2]Swedish University of Agricultural Sciences, dept. of Plant Physiology, Umeå, Sweden
[3]HudsonAlpha Institute for Biotechnology, Huntsville Al, USA
[4]Swedish University of Agricultural Sciences, dept. of Plant Biology and Forest Genetics, Uppsala, Sweden

1 Introduction

The *Populus* genus consists of about 30 species, which are commonly found in the Northern Hemisphere. They are an important model system for forest tree research, an ecological pioneer species, and of high commercial interest due to their rapid growth and ease of propagation (Stettler et al., 1996).

The genome assembly of the European aspen *Populus tremula* (Lin et al., 2018) proved difficult for a short-read based strategy due to high genomic variation. As a consequence, the fragmented sequence is impeding studies that benefit from highly contiguous data, particularly genome-wide association studies (GWAS) and comparative genomics.

Here we present an updated assembly based on long-read sequences, optical mapping and genetic mapping. This assembly - henceforth referred to as *Potra* V2 - is assembled into 19 contiguous chromosomes which provides a powerful tool for future association studies.

The genome sequence and any feature files are available from the PopGenIE resource (Sjödin et al., 2009).

2 Results and Discussion

2.1 Genome Assembly

The *P. tremula* genome assembled into 19 chromosomes and 1582 scaffolds with a combined length of 408834716bp. Aligning 95M Illumina reads (about 20x coverage) yields a 96.4% (97.77% in V1) map percentage with 94.19% (92.33% in V1) of these maps in proper pairs. The increase in proper pairs and a decrease in overall mapping reflects our expectation from an assembly with higher contiguity but lower per-base accuracy. Table 1 provides additional summary statistics for the raw assembly.

Table 1: Summary statistics for *P. tremula* version 2.

Statistic	Potra01	Potra02
# contigs (>= 0 bp)	204318	1601
# contigs (>= 1000 bp)	31632	1584
# contigs (>= 5000 bp)	7267	1339
# contigs (>= 10000 bp)	5151	986
# contigs (>= 25000 bp)	3209	491
# contigs (>= 50000 bp)	1789	255
Total length (>= 0 bp)	386236512	408834716
Total length (>= 1000 bp)	328536064	408824553
Total length (>= 5000 bp)	277117215	407999588
Total length (>= 10000 bp)	262322877	405364617
Total length (>= 25000 bp)	231504505	397478443
Total length (>= 50000 bp)	180499961	389097052
# contigs	12044	1489
Largest contig	418873	53234430
Total length	294670244	408605800
GC (%)	33.56	33.87
N50	69979	16928776
N75	29987	13637973
L50	1227	9
L75	2826	15
# N's per 100 kbp	5428.58	6573.91
Reads aligned (%)	97.77%	96.40%
Reads properly paired (%)	92.33%	94.19%

Analysis of the genome using BUSCO (Simão et al., 2015) with the `embryophyta_odb10` ortholog set showed 96.8% (96.8% in V1) complete BUSCOs, of which 81.7% (82.5% in V1) were single copy and 15.1% (14.3% in V1) duplicated. The first version of the assembly scored better in duplication and missing BUSCOs. Detailed values follow in table 2.

The long terminal repeat (LTR) index for the assembly (Ou et al., 2018) is 6.65, with 1.42% of intact LTRs 20.66% of total LTRs. This LTR index indicates a high-quality draft assembly comparable to apple v1.0 or cacao v1.0.
Analysis of structural variation in both genomes showed fewer overall variants in PotraV2, but a higher rate of insertions and inversions. The extremely high number of detected translocation events in PotraV1 is likely due to the overall fragmentation of the genome.

Synteny alignments between V1 and V2 showed that 59.6% of genomic regions on chromosomes (55.1% when including scaffolds) in V2 have a corresponding V1 alignment. The extent of the difference in these assemblies is suprising, especially given the high mapping rates of genomic shotgun sequence. It is plausible that the regions that are missed are comprised of sequence that is traditionally problematic for short-read assemblies, e.g.: repeats. Figure 2.1 shows a visual representation of the synteny alignments for the 19 chromosomes in V2.

Table 2: BUSCO genome statistics for both assemblies.

BUSCO	Potra01	Potra02
Complete	96.8%	96.8%
Single Copy	82.5%	81.7%
Duplicated	14.3%	15.1%
Fragmented	1.5%	0.9%
Missing	1.7%	2.3%

Table 3: Structural variants called after auto-alignment

Variant	Potra01	Potra02
Translocation	126172	23920
Copy Number Variant	5284	1332
Deletion	44770	35290
Insertion	22848	36864
Inversion	25	108
Split	965	853

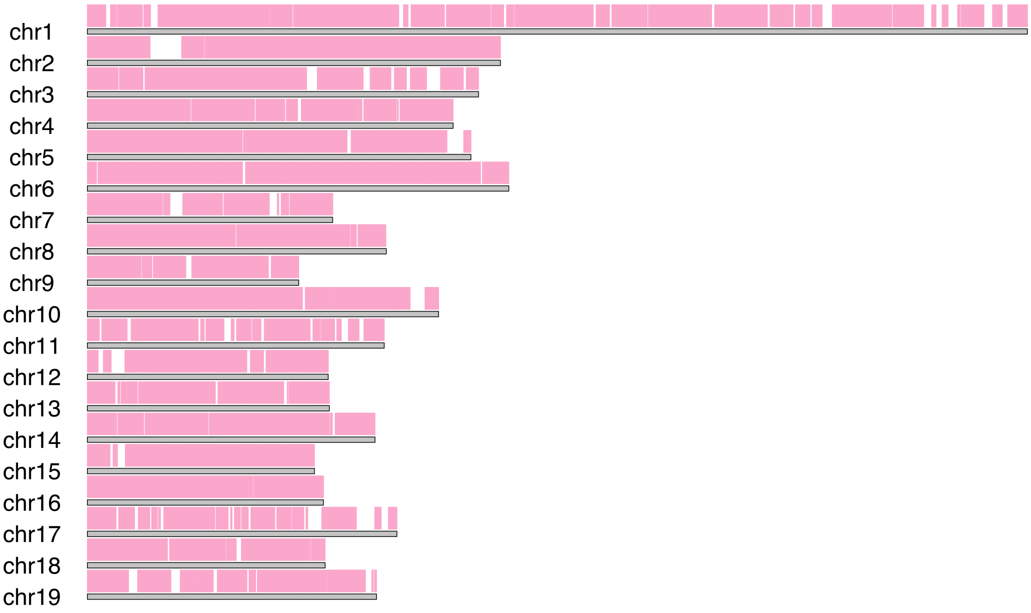

Figure 1: Synteny alignment of the version 1 assembly to the version 2 assembly. Blocks with a highly similar synteny alignment are shaded red.

Table 4: BUSCO transcript statistics for both assemblies.

BUSCO	Potra01	Potra02
Complete	96.8%	98.1%
Single Copy	30.2%	35.7%
Duplicated	66.6%	62.4%
Fragmented	2.3%	0.9%
Missing	0.9%	1.0%

2.2 Gene annotation

We identified in total 39894 gene models, 37184 of which on chromosomes and 2710 on scaffolds. In total, we detected 77949 transcripts, 73765 on chromosomes and 4184 on scaffolds (1.95 transcripts per gene). We found functional annotations for 73765 transcripts in 37184 genes.
Analysis of the predicted transcripts using BUSCO with the `embryophyta_odb10` ortholog set showed 98.1% (96.8% in V1) complete BUSCOs, of which 35.7% (30.2% in V1) were single copy and 62.4% (66.6% in V1) duplicated. Version 2 of the assembly performed slightly better in complete and single-copy BUSCOs. Detailed values follow in table 4.

3 Materials and Methods

If not otherwise specified, we omitted irrelevant arguments (s.a. file paths, parallelism) from command lines for the sake of clarity.
If no arguments are specified, we did not make any changes to the defaults.
Unless otherwise specified, we aligned genomic data with BWA mem v0.7.8-r455 (Li, 2013) and RNA-Seq data with STAR v2.6.1d (Dobin et al., 2013).
All scripts and config files can be found in the Git repository: https://github.com/bschiffthaler/aspen-v2

3.1 Data

Unless otherwise specified Science for Life Laboratory in Stockholm generated all sequence data. For genome assembly and correction, we generated two libraries:

- “PacBio data”: 28874072954 bases (filtered subreads, 60x coverage), Pacific Biosciences on the RSII platform. ENA: TBD
- “Illumina data”: 108353739802 bases (226x coverage), Illumina HiSeq2500. ENA: TBD

We also collected several RNA-Seq datasets for use in the genome annotation:

- “AspWood” (Sundell et al., 2017) (ENA: ERP016242)
- “Sex” (Robinson et al., 2014) (ENA: ERP002471)
- “SwAsp” (Mähler et al., 2017) (ENA: ERP014886)

- “Assembly version 1 tissue atlas” (Lin et al., 2018) (ENA: PRJEB23585)
- “Xylem/Leaf” (Lin et al., 2018) (ENA: PRJEB23585)
- “Leaf Development” (Unpublished data, sequenced by BGI Genomics)

3.2 Assembly

Initially, we assembled the genome using FALCON v0.3 (Chin et al., 2016). We include the FALCON config file in the Git repository. Subsequently, we aligned all the Illumina data to the initial assembly and used in-house scripts to correct homozygous SNPs and small INDEL issues. We then aligned the Illumina data to the fixed assembly and repeated the first round of fixing. For a third and final round of fixing, we used the Illumina data as input to Pilon (Walker et al., 2014) v2.11-1.18 to correct assembly issues per scaffold.

In order to reduce the presence of split haplotypes, we used HaploMerger2 (Huang et al., 2017) (retrieved: 2015-11-06). We include all HaploMerger2 scripts in the Git repository.

We subsequently created an optical map of the genome in collaboration with BioNano genomics, which we utilized to further scaffold and orient our current assembly.

Finally, we used the high-density genetic linkage map from Apuli et al. (2019) as input to ALLMAPS (Tang et al., 2015) to place the scaffolds into chromosomes.

3.3 Transcriptome Assembly

To provide evidences for the gene annotation process, we used trinity (Grabherr et al., 2011) to assemble the transcriptome of five RNA-Seq (Mortazavi et al., 2008) datasets from Populus tremula. Four of the datasets had already been used for the annotation of the previous genome version (Lin et al., 2018): exAtlas, exDiversity, Xylem.leaf and Leaf, while the fifth was derived from our AspWood resource (Sundell et al., 2017). These five datasets are available from the ENA (https://ebi.ac.uk/ena) under the accessions PRJEB5040, PRJEB1790, PRJEB28867, PRJEB28866, PRJEB14593, respectively. The reads were pre-processed as described in Lin et al. (2018) and Sundell Sundell et al. (2017). Briefly, the raw reads were filtered for rRNA using SortMeRNA (Kopylova et al., 2012) version 2.1 and trimmed for adapter sequences and lower quality using Trimmomatic (Bolger et al., 2014) v0.39. The filtered reads were then assembled using trinity (Haas et al., 2013) version 2.8.4 using default settings. The resulting transcript fasta files were then used as evidence for Maker-P.

3.4 Annotation

We first collected a set of diverse RNA-Seq datasets from previous studies. These, we aligned to the genome using STAR in 2-pass mode. For the first pass, we used the following arguments:

```
STAR --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 \
  --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 \
  --outFilterMismatchNoverReadLmax 0.1 --alignIntronMin 20 \
  --alignIntronMax 20000 --alignMatesGapMax 5000 \
  --outSAMtype BAM SortedByCoordinate --chimOutType WithinBAM
```

For the second pass, we provided the splice junctions from the first pass as `pass-1-SJ.out.tab` and used:

```
STAR --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 \
  --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 \
  --outFilterMismatchNoverReadLmax 0.1 --alignIntronMin 20 \
  --alignIntronMax 20000 --alignMatesGapMax 5000 \
  --outSAMtype BAM SortedByCoordinate --chimOutType WithinBAM \
  --limitSjdbInsertNsj 2000000 --sjdbFileChrStartEnd pass-1-SJ.out.tab
```

We then provided these alignments to BRAKER1 (Hoff et al., 2015) along with protein sequences of version 1 of the assembly, running BRAKER1 in hybrid mode with arguments:

```
braker.pl --genome=genome.fa --prot_seq=protein.fa \
  --prg=gth --softmasking --AUGUSTUS_ab_initio
```

In order to prepare the genome for annotation, we created a custom repeat library using RepeatModeler v1.0.11. We then concatenated the custom repeats with known repeats in *Viridiplantae* and the first assembly of the *P. tremula* genome. We masked the genome using RepeatMasker 4.0.8 [1].

We ran MAKER v2.31.10 (Campbell et al., 2014) on the masked genome in three passes. We include the MAKER config files in the Git repository. We used Trinity assemblies from all RNA-Seq datasets in conjunction with all transcripts from the v1 assembly as expressed sequence tag (EST) evidence. Further, we provided proteins from the v1 assembly and the v3.0 assembly of *P. trichocarpa* (Tuskan et al., 2006) as protein evidence. In order to train AUGUSTUS v3.0.2 (Stanke et al., 2008) and SNAP v2013-11-29 (Korf, 2004) we extracted confident predictions from the first run of MAKER using `maker2zff` from the MAKER suite and `zff2augustus_gbk.pl` from an external source[2]. We then proceeded with another round of MAKER including AUGUSTUS, SNAP and GeneMark-ES (Lomsadze et al., 2005). We repeated this process of training AUGUSTUS and SNAP once more for a third and final round of MAKER.

3.5 Functional Annotation

We aligned the transcripts and protein-coding sequences retrieved from MAKER to the NCBI nt (Wheeler et al., 2006) and UniRef90 (Consortium, 2018) databases, respectively. For transcripts, we used Blast+ version 2.6.0+ with the non-default parameters: `-evalue 1e-5` (Altschul et al., 1990). For proteins, we used Diamond version 0.9.26 with default parameters (Buchfink et al., 2015). We identified and extracted the sequences aligning solely to the NCBI nt database to complement the UniRef90 alignments using an ad-hoc script (available upon request). We then imported the resulting alignment files in Blast2GO (Götz et al., 2008) version 5.2. Finally, we used Blast2GO to generate the Gene Ontology (both GO and GO-Slim), PFAM (El-Gebali et al., 2018) and KEGG (Kanehisa and Goto, 2000) annotations.

3.6 Evaluation

To calculate summary statistics of the assembly, we used QUAST v5.0.2 (Gurevich et al., 2013), aligning a 20x coverage subset of the aspen V1 2x150 PE library data (ENA: PRJEB23581) to calculate mapping percentages.

[1] http://www.repeatmasker.org

[2] https://github.com/hyphaltip/genome-scripts/blob/master/gene_prediction/zff2augustus_gbk.pl

We ran BUSCO v3.0.2 for both the genomic and transcript sequences. We retrieved the "embryophyta_odb10" dataset from https://busco.ezlab.org/.

4 acknowledgements

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was supported by the Knut and Alice Wallenberg Foundation and the Umeå Plant Science Centre Berzelii Centre. N.R.S. was supported by the Trees and Crops for the Future project.

References

Altschul SF, et al. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Apuli RP, et al. Constructing a high-density linkage map to infer the genomic landscape of recombination rate variation in european aspen (populus tremula). *BioRxiv*, page 664037, 2019.

Bolger AM, et al. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

Buchfink B, et al. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59, 2015.

Campbell MS, et al. Genome annotation and curation using maker and maker-p. *Current Protocols in Bioinformatics*, 48(1):4–11, 2014.

Chin CS, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050, 2016.

Consortium U. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2018.

Dobin A, et al. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

El-Gebali S, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2018.

Götz S, et al. High-throughput functional annotation and data mining with the blast2go suite. *Nucleic acids research*, 36(10):3420–3435, 2008.

Grabherr MG, et al. Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data. *Nature biotechnology*, 29(7):644, 2011.

Gurevich A, et al. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

Haas BJ, et al. De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494, 2013.

Hoff KJ, et al. Braker1: unsupervised rna-seq-based genome annotation with genemark-et and augustus. *Bioinformatics*, 32(5):767–769, 2015.

Huang S, et al. Haplomerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, 33(16):2577–2579, 2017.

Kanehisa M and Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Kopylova E, et al. Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217, 2012.

Korf I. Gene finding in novel genomes. *BMC bioinformatics*, 5(1):59, 2004.

Li H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

Lin YC, et al. Functional and evolutionary genomic inferences in populus through genome and population sequencing of american and european aspen. *Proceedings of the National Academy of Sciences*, 115(46):E10970–E10978, 2018. ISSN 0027-8424. doi:10.1073/pnas.1801437115.

Lomsadze A, et al. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic acids research*, 33(20):6494–6506, 2005.

Mähler N, et al. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS genetics*, 13(4):e1006402, 2017.

Mortazavi A, et al. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621, 2008.

Ou S, et al. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*, 46(21):e126–e126, 2018. ISSN 0305-1048. doi:10.1093/nar/gky730.

Robinson KM, et al. Populus tremula (european aspen) shows no evidence of sexual dimorphism. *BMC plant biology*, 14(1):276, 2014.

Simão FA, et al. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015.

Sjödin A, et al. The populus genome integrative explorer (popgenie): a new resource for exploring the populus genome. *New phytologist*, 182(4):1013–1025, 2009.

Stanke M, et al. Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics*, 24(5):637–644, 2008.

Stettler R, et al. *Biology of Populus and its implications for management and conservation.* NRC Research Press, 1996.

Sundell D, et al. Aspwood: high-spatial-resolution transcriptome profiles reveal uncharacterized modularity of wood formation in populus tremula. *The Plant Cell*, 29(7):1585–1604, 2017.

Tang H, et al. Allmaps: robust scaffold ordering based on multiple maps. *Genome biology*, 16(1):3, 2015.

Tuskan GA, et al. The genome of black cottonwood, populus trichocarpa (torr. & gray). *science*, 313(5793):1596–1604, 2006.

Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963, 2014.

Wheeler DL, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 35(suppl_1):D5–D12, 2006.