

Supplementary Information for

Functional and evolutionary genomic inferences in *Populus* through genome and population sequencing of American and European aspen

Yao-Cheng Lin, Jing Wang, Nicolas Delhomme, Bastian Schiffthaler, Görel Sundström, Andrea Zuccolo, Björn Nystedt, Torgeir R. Hvidsten, Amanda de la Torre, Rosa M. Cossu, Marc P. Hoepfner, Henrik Lantz, Douglas G. Scofield, Neda Zamani, Anna Johansson, Chanaka Mannapperuma, Kathryn M. Robinson, Niklas Mähler, Ilia J. Leitch, Jaume Pellicer, Eung-Jun Park, Marc Van Montagu, Yves Van de Peer, Manfred Grabherr, Stefan Jansson, Pär K. Ingvarsson, Nathaniel R. Street.

Nathaniel Street

Email: nathaniel.street@umu.se

This PDF file includes:

Supplementary text

Figs. S1.1 to S8.6

Tables S1.1 to S8.2

References for SI reference citations

Supporting Information

1 Naming conventions

Throughout the manuscript, including table and figure legends, we use acronyms to refer to the three primary species of interest. The same acronyms are used in filenames and in tools at the PopGenIE.org web resource and FTP site:

- Potra: *Populus tremula* L.
- Potri: *Populus trichocarpa* Torr. & Gray
- Potrs: *Populus tremuloides* Michx.
- Podav: *Populus davidiana* Dode.

Similar acronyms are introduced for additional species as they appear.

Throughout this supplementary document, we refer to a number of scripts and analysis transcripts. These are available from our public git repository at <https://github.com/UPSCb/UPSCb/tree/master/manuscripts/Lin2018>. We also refer to the PopGenIE FTP resource, which is available at <ftp://plantgenie.org/Data/PopGenIE/>. Additional files are available from the FTP directory associated with this publication at <ftp://plantgenie.org/Publications/Lin2018>.

2 Genome sequencing and assembly

2.1 Material for Potra genome assembly

Root cuttings from an assumed wild-growing Potra individual growing on the Umeå University campus (63° 49'17"N, 20° 18'40"E) were collected in May 2009 and used to establish clonal replicates in the greenhouse. Root cuttings were established in five litre pots and fresh, young leaf material was sampled for all genomic DNA extractions, which were performed using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's standard protocol.

2.2 Sequence data generation

2.2.1 Potra

Shotgun 454 sequencing was performed by the Science for Life Laboratory (SciLifeLab; Stockholm, Sweden) using standard Roche (Roche 454 Life Sciences, Brandford, USA) library and sequencing protocols on a Genome Sequencer FLX Instrument. The sequencing data was produced in two batches, each using different Titanium chemistry versions that yielded mean reads length of 323 bp and 551 bp, respectively, with total depths of 2.38 Gbp and 2.55 Gbp, respectively.

Paired-end (PE; 2x101 bp) insert libraries of 150, 300 and 650 bp were produced and Illumina sequencing data was generated using standard protocols on the HiSeq platform by SciLifeLab (Stockholm, Sweden), sequenced to depths 36.26 Gbp, 20.52 Gbp and 12.83 Gbp.

Mate-pair (MP) reads were produced with target insert sizes of 3, 5 and 10 Kbp. The 3 and 5 Kbp libraries were prepared and sequenced at SciLifeLab (Stockholm, Sweden) using an in-house 454-derived circularisation protocol that resulted in a broad range of insert sizes, but very low fractions of PE reads. PE 2x101 bp sequencing was performed on an Illumina HiSeq 2000. The 10 Kbp library was prepared and sequenced by BGI (Beijing, China) using in-house protocols and sequenced using 2x75 bp reads on the Illumina HiSeq platform.

2.2.2 *Potrs*

P. tremuloides DNA for genome assembly was extracted from mature, freeze-dried leaves from genotype Dan2-1B7 using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). This sample is among those described in Wang et al. (2016)(1).

Paired-end (PE; 2x100 bp) Illumina sequencing data was generated using standard protocols on the HiSeq 2500 platform by SciLifeLab (Stockholm, Sweden). Sequencing libraries with target insert sizes of 150, 300 and 650 bp were produced and sequenced to depths of 17.85 Gbp, 23.77 Gbp and 22.01 Gbp, respectively.

2.2.3 *Podav*

Sequencing data from *Podav* individuals were obtained from NIFoS (National Institute of Forest Science, Korea). DNA was extracted from mature leaves sampled from mature trees in a common garden experiment in Suwon, Korea.

2.3 Quality control

Upon reception of all Illumina sequencing data, we performed initial quality control using FastQC (v0.10.1; <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). We used FastQC to analyse the data in this manner after every step which modified it. After the initial QC, we trimmed low quality reads with Trimmomatic (2) (v0.32) using parameters SLIDINGWINDOW:5:20 MINLEN:50 as well as removing adapters with ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10.

For the MP data, we used Picard tools (<http://broadinstitute.github.io/picard/>) to remove PCR duplicates, which are a common problem in MP libraries. After removing duplicated reads the two MP libraries represented ~38X and ~724X span coverage (the number of times the genome is covered by the insert size of the library) respectively.

Quality control of the 454 data was performed as an integral part of the GS-Assembler (Newbler) process.

Table S2.1 Cleaned data coverage statistics for sequencing data subsequently utilised for contig assembly.

Species	Sequencing platform	Library	Cleaned data (Gbp)
Potra	Illumina	150 PE	18.32
	Illumina	300 PE	8.81
	Illumina	650 PE	6.16
	454	350 SE	2.38
	454	550 SE	2.55
Potrs	Illumina	150 PE	8.15
	Illumina	300 PE	13.58
	Illumina	650 PE	10.91

PE; Paired-end; SE; single-end

2.4 Sequence data availability

We have deposited all raw sequencing data at the ENA resource (3) (www.ebi.ac.uk/ena) under the accession number PRJEB23585, except for Podav sequencing data, which are available from the PopGenIE.org Lin2018 FTP resource as accession pnas201801437. The Podav data will be described fully elsewhere.

Potra

The 454 datasets, the Illumina 150, 300 and 650 bp PE and the 3 and 5 Kbp MP datasets and the 10 Kbp MP dataset are available under the accession: PRJEB23583, PRJEB23581 and PRJEB23582, respectively.

Potrs

The Illumina 150, 300 and 650 bp PE libraries are available as accession: PRJEB23580.

Podav

The sequencing data is available from the PopGenIE.org Lin2018 FTP resource as accession pnas201801437.

2.5 Assembly-free genome characterisation

2.5.1 Genome size estimation using flow cytometry

Nuclear DNA contents were estimated by propidium iodide flow cytometry (4). The Potri genotype used to generate the reference genome assembly (5) was included for comparison. We did not have suitable leaf material from aspen species other than Potra. Briefly, fully expanded leaf tissue from each specimen (about 1 cm²) was chopped along with the appropriate internal standard (*Solanum lycopersicum* L. ‘Supiké polní rané’, 1C=958.44 Mb (6, 7)) using a new razor blade in 1 ml of LB01 buffer (8). A further 1 ml of LB01 was then added and the sample was passed through a 30 µm nylon filter. The homogenate was stained with 100 µl propidium iodide (1 mg/ml), and kept on ice for 20 min. For each species analysed, samples from three individuals were prepared and three replicates of each were run, recording at least 1,000 nuclei per fluorescence peak using a Partec Cyflow SL3 (Partec GmbH, Münster, Germany) flow cytometer fitted with a 100-mW green solid-state laser (Cobolt Samba). The resulting histograms were analysed with the FlowMax software (v. 2.7, Partec GmbH).

Table S2.2 Flow cytometry genome size estimates

Species	Genotype	2C (pg)	S.D.	1C (Mb)	CV (standard)	CV (target)
<i>P. tremula</i>	Asp201	0.98	0.03	479.22	2.46	3.85
<i>P. tremula</i>	SwAsp23	0.97	0.01	474.33	2.41	3.51
<i>P. tremula</i>	SwAsp51	0.98	0.01	479.22	2.11	3.1
<i>P. tremula</i>	SwAsp64	0.97	0.01	474.33	2.07	3.64
<i>P. tremula</i>	SwAsp110	0.98	0.01	479.22	2.31	3.37
<i>P. trichocarpa</i>	Nisqually-1	0.98	0.01	479.22	2.68	3.36

CV: coefficient of variation (must be < 5). 2C: Holoploid genome size (2n). SD: Standard deviation. 1pg = 978 Mbp.

2.5.2 Genome size estimation using kmer analysis

In order to enable *k*-mer based comparisons we generated *k*-mer hashes for each *k* in range 17-24 with jellyfish (9) (v2.1.3) as the shape of the *k*-mer frequency graph is usually heavily dependent on the value of *k*. For further analyses we selected a *k* of 24. We observed clear evidence of prevalent heterozygosity in all aspen species, as supported by the height of the half-maximal peak in the *k*-mer frequency spectra (Figure S2.1). We performed these analyses using the script: compare_k_17_24_final.R.

We additionally ran kmergenie (10) (v1.6950) with the “--diploid” option for all input PE libraries post quality trimming. In addition, we used *k*-mer frequency spectra to manually calculate the genome size (see script: estimate_size.R). Overall, the high heterozygosity prevented reliable *k*-mer based size estimates (Table S2.3), as is evident from the

difference in these k -mer based size estimate in comparison to the flow cytometry measures, which provide a far more accurate measure of true genome size.

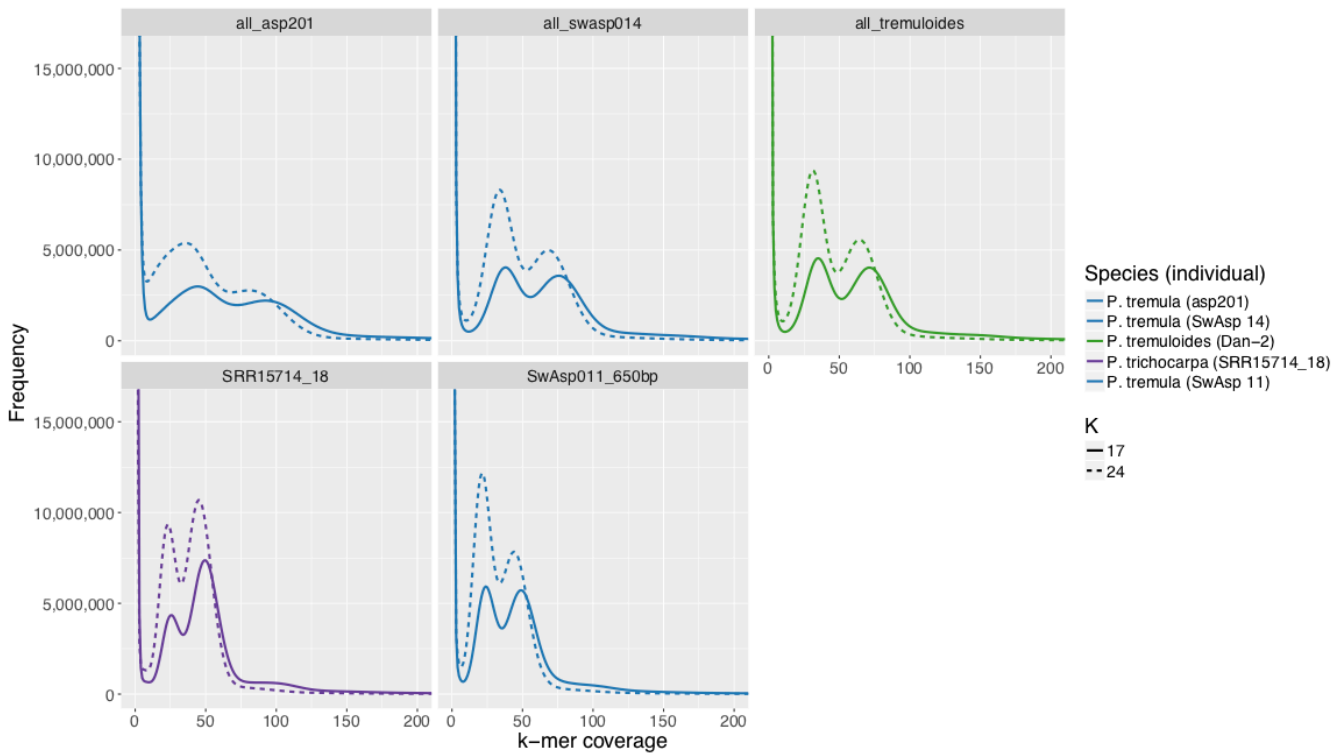


Figure S2.1 K-mer frequency distributions of all species.

Table S2.3 Genome size estimate results from various sources. Values indicate base pairs.

Library	k=17	k=21	k=24	kmergenie
<i>P. trichocarpa</i> (SRR15714_18_19)	504,154,946	536,335,049	560,172,162	402,806,141
<i>P. tremula</i> (Asp201)	408,920,404	442,596,202	464,452,804	312,646,087
<i>P. tremuloides</i> (Dan2-1B)	533,052,934	564,876,989	591,355,598	323,969,142

The columns K=x indicate results generated using k -mer frequency spectra, while kmergenie is the size estimated by the kmergenie tool.

2.5.3 Read alignment to the Potri genome

We explored the effect of alignment parameters, testing multiple alignment algorithms and allowing mismatch rates of up to 20%. Using *bwa-aln* (11) as an alignment tool, median coverage of coding (exon) and intergenic regions (excluding annotated repeats, introns, up- and down-stream regions) was 47X and 0.5X, respectively, with 47% of bases within intergenic regions having zero aligned reads. Changes to mismatch settings did not result in increased alignment rates, suggesting that low alignment rates were not the result of high Single Nucleotide Polymorphism (SNP) occurrence between the two genomes. We initially performed alignments before the availability of tools implementing maximal exact matching (MEM) algorithms (*e.g.* *bwa-mem* (11) and STAR (12)). Using *bwa* as an example allows comparison of alignment algorithms: *bwa-aln* performs global alignment of reads allowing for mismatches and for a single gap, with gap opening and extension strongly penalised; *bwa-mem* allows a read to be split into multiple sections with each section being able to contain mismatches and gaps. Allowing reads to be split in this way allows for alignment of a single read across multiple gaps, each of which can be of substantial length. In essence, *bwa-mem* can align reads where INDELS are common whereas *bwa-aln* cannot. Using *bwa-mem* resulted in a dramatic increase in read alignment rates to Potri, with 89% reads aligning and with coding region and intergenic coverage increasing to 54X and 8X respectively (Figure S2.2), with 27.4% of bases in intergenic regions having zero aligned reads.

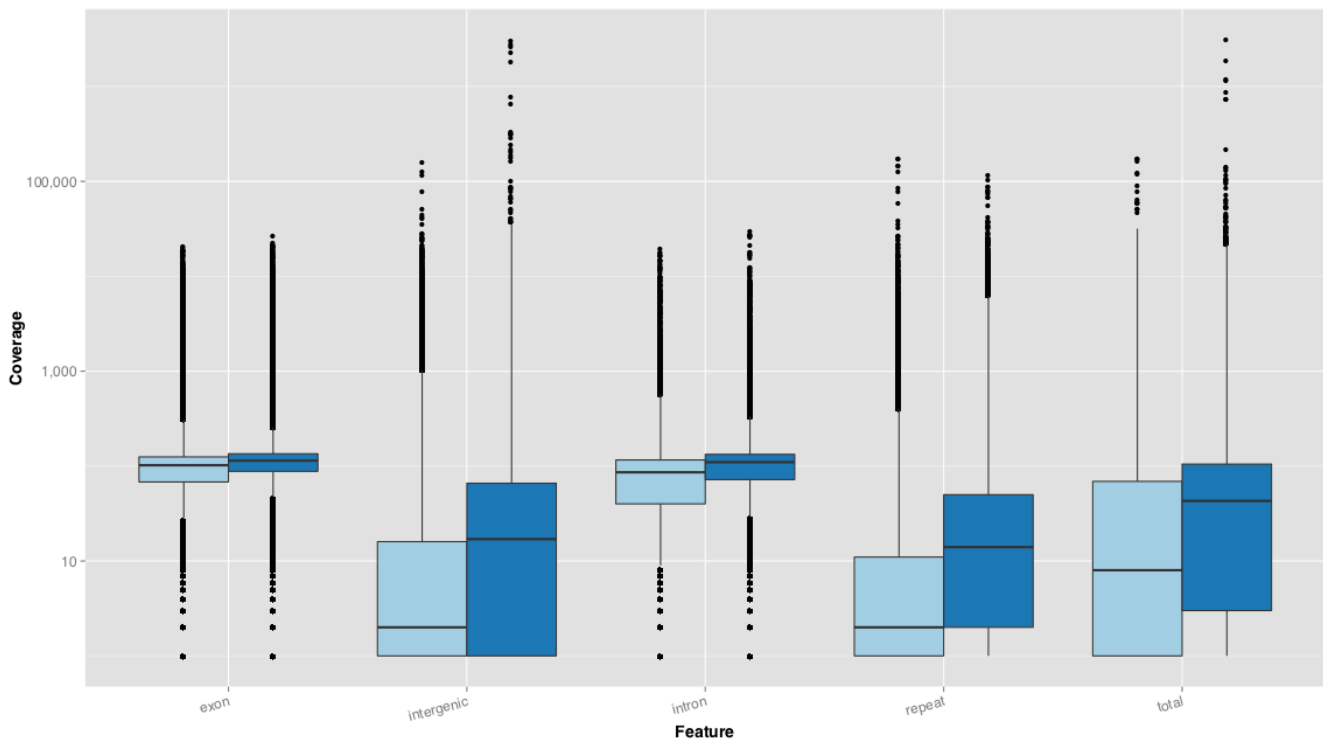


Figure S2.2 Genomic context of *bwa-aln* (light blue) and *bwa-mem* (dark blue) aligned reads for Potra. Reads from the assembled Potra individual (Asp201) were aligned to the reference Potri genome assembly (v3).

2.6 Genome assembly

2.6.1 Potra genome assembly

We used a hierarchical assembly approach that comprised separate contig assemblies of the 454 and Illumina data that were merged in a stepwise manner, with the final merged assembly subsequently scaffolded using the MP libraries. We tested alternatives to this approach, including hybrid assemblies using the CLC de novo tool (CLCBio, Aarhus, Denmark), ABySS (13) and GS-Assembler (Newbler) but all performed worse than the assembly produced using this hierarchical approach. An overview of the assembly pipeline is given in Figure S2.3 and details follow below.

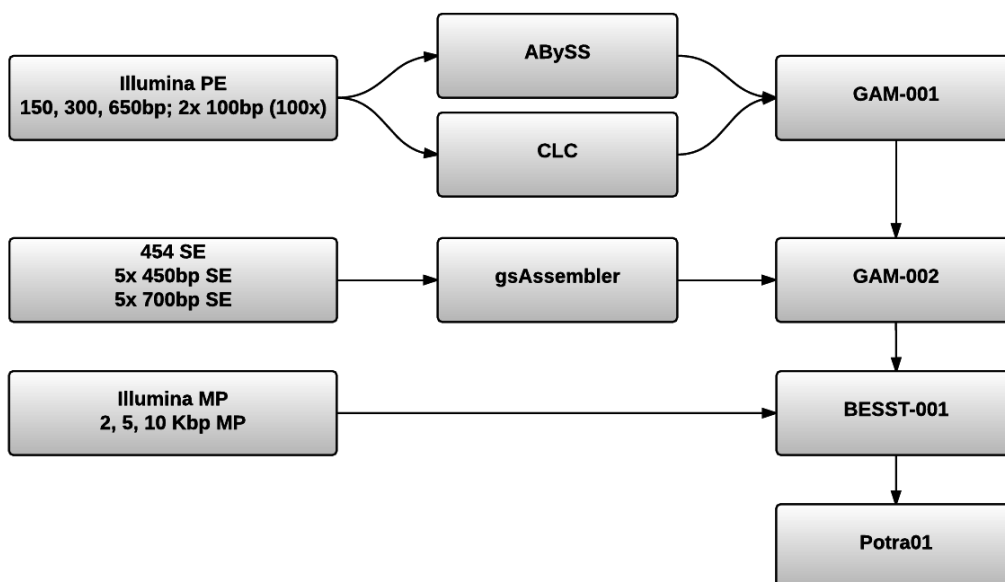


Figure S2.3 The assembly pipeline used to assemble the Potra genome. Sequencing library types are indicated as SE single end, PE paired-end and MP mate-pair. For Illumina PE libraries, the library target insert size is indicated followed by the read lengths. For 454 libraries, the mean read length is indicated. For Illumina MP libraries, the target insert size is indicated. The centre column of the figure indicates assembly tool names and the final column indicates the names assigned to the various assembly steps.

454 Assembly

All 454 data were assembled using GS Assembler v2.9 (A.K.A Newbler, Roche 454 Life Sciences, Brandford, USA) using the settings `-large -cpu 8 -minlen 45 -trim -het -sio -urt`. Assembly was performed using raw SFF files as input.

Illumina paired-end read assembly

We performed two separate assemblies of the Illumina data using the CLC de novo assembler (v4.2.0) and ABySS (v1.3.6). Before assembly, we checked sequence data quality for all raw data using FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and subsequently quality filtered the data to remove adapter contamination and low-quality bases as detailed above. The trimmomatic quality filtering step created some 'orphan' reads, *i.e.* where one of the two reads of a PE read failed to pass the quality criteria and hence was filtered out leaving behind a single forward or reverse mate. We used all such orphaned reads as single end reads during assembly.

We performed an initial exploration of the parameter space for each of the two assembly programs to determine the parameters used for the final assemblies. For each assembly, we examined contiguity statistics (primarily N50 length) and alignment rates of the Potri annotation (CDS sequences of primary transcripts per loci) to determine relative assembly success. We placed greater emphasis on gene alignment rates than in N50 length as a metric for assembly quality, although the two were in general agreement.

For both tools, we performed assemblies using k -mer sizes (referred to as word size by CLC) spanning the range 26-96, initially incrementing k in steps of five. For CLC we also tested assembly using default settings, where word size is determined automatically, for which a word size of 25 was selected. For CLC this was the optimal assembly. For ABySS, k -mers in the range of 40-50, with a peak at 45, were optimal and we therefore performed additional tests incrementing k in steps of one between 40-50. The N50 length was optimal at $k=49$ while gene alignment metrics were optimal at $k=47$. We therefore used a k -mer size of 47 for the final assembly. The ABySS assembly was performed using the settings `k=47 aligner=map np=60 n=10 N=3 s=500 b=10000 p=0.8 v=-v SIMPLEGRAPH_OPTIONS=---dist-error=30 POPBUBBLES_OPTIONS=---scaffold FAC_OPTIONS=---mmd FAC_OPTIONS=-e=450000000 lib='PE1 PE2 PE3'`. In both cases, no scaffolding was performed.

Assembly merging and scaffolding

We merged the various assemblies (Newbler for the 454 data, ABySS and CLC for the Illumina data) in two stages using the Genome Assembly Merging tool (GAM-NGS (14)). The first merge involved the two Illumina assemblies, where the CLC assembly was defined as the 'master' assembly during merging. In a second merging stage, the previously-merged Illumina assembly (GAM-001) was merged with the Newbler 454 assembly (GAM-002), with the Illumina assembly defined as the master. GAM-NGS uses sequencing read alignments to identify matching contigs between assemblies. We used the 300 bp PE Illumina library for this purpose. Quality trimmed PE reads from this library were aligned to all assemblies using `bwa-aln` with the settings `-l 28 -k 1 -n 3 -o 0`. The alignments were saved in BAM format and were coordinate sorted using `samtools` (15), as required by GAM-NGS.

The final GAM-merged assembly (GAM-002) was then scaffolded using all PE and MP libraries using BESST (16) (v1.0.4). For this purpose, all libraries were aligned to the GAM-002 assembly using bwa-aln with the settings -l 28 -k 1 -n 3 -o 0. This final, scaffolded assembly was used for all subsequent analyses and is referred to as Potr01 hereafter. An overview of the assembly pipeline is provided above in Figure S2.3.

2.7 Potrs genome assembly

As we only had PE Illumina data available for Potrs, assembly was performed using only ABySS, as detailed for Potra above. A final scaffolding of the ABySS assembly was performed using the three PE libraries using BESST, as detailed for Potra. The final, scaffolded assembly was named Potrs01.

2.8 Assembly characteristics and quality assessment

Assembly statistics for the Potra sub-stages are summarised in Table S2.4 and for the final Potra01 and Potrs01 assemblies in Table S2.5. In general, the final assemblies were highly fragmented and there was clear evidence of either assembly collapsing or of missing genome regions not being represented, as indicated by the final assembly size.

Table S2.4 Assembly statistics for the Potra assembly stages.

Assembly metric	ABySS	CLC	Newbler	GAM01	GAM02
Number of scaffolds	2,596,787	303,989	276,332	289,418	275,216
Total size of scaffolds	495,432,441	373,787,603	381,212,473	373,771,897	374,843,563
# scaffolds > 500 nt	92,967	121,577	164,556	113,763	108,776
% scaffolds > 500 nt	3.6	40	59.6	39.3	39.5
# scaffolds > 1K nt	68,773	75,153	93,883	71,587	70,503
% scaffolds > 1K nt	2.6	24.7	34	24.7	25.6
# > 10K nt	3,776	5,106	4,224	5,878	6,391
# > 100K nt	0	1	2	1	2
N50	1,326	4,008	3,056	4,520	4,925
scaffold %A	31.67	33.68	32.94	33.57	33.56
scaffold %C	16.87	16.62	16.68	16.58	16.58
scaffold %G	16.47	16.36	16.72	16.33	16.34
scaffold %T	31.38	33.34	33.65	33.26	33.26

Table S2.5 Assembly metrics for the final Potra01 and Potrs01 assemblies.

Assembly metric	Potra01	Potrs01
Number of scaffolds	216,318	164,504
Total size of scaffolds	390,124,095	377,489,497

# scaffolds > 500 nt	57,475	59,039
% scaffolds > 500 nt	26.6	35.9
# scaffolds > 1K nt	31,806	39,866
% scaffolds > 1K nt	14.7	24.2
# > 10K nt	5,161	10,248
# > 100K nt	687	28
N50	42,844	15,222
scaffold %A	32.18	33.16
scaffold %C	15.88	16.09
scaffold %G	15.75	16.02
scaffold %T	32.03	33.00

2.8.1 Gene space coverage

We evaluated gene space coverage of the Potra final assembly and assembly sub-stages by aligning the primary transcripts of the Potri reference genome to the various Potra assemblies using GMAP (17) (v14.12.29; -K 11000 –cross-species). The results were evaluated in R (script: parse_gmap.R, Figure S2.4). The scaffolded final assembly had the most unique transcripts aligned, least mismatches and highest coverage of all assemblies, with the results indicating that we achieved better representation of the gene space with each additional assembly merging step.

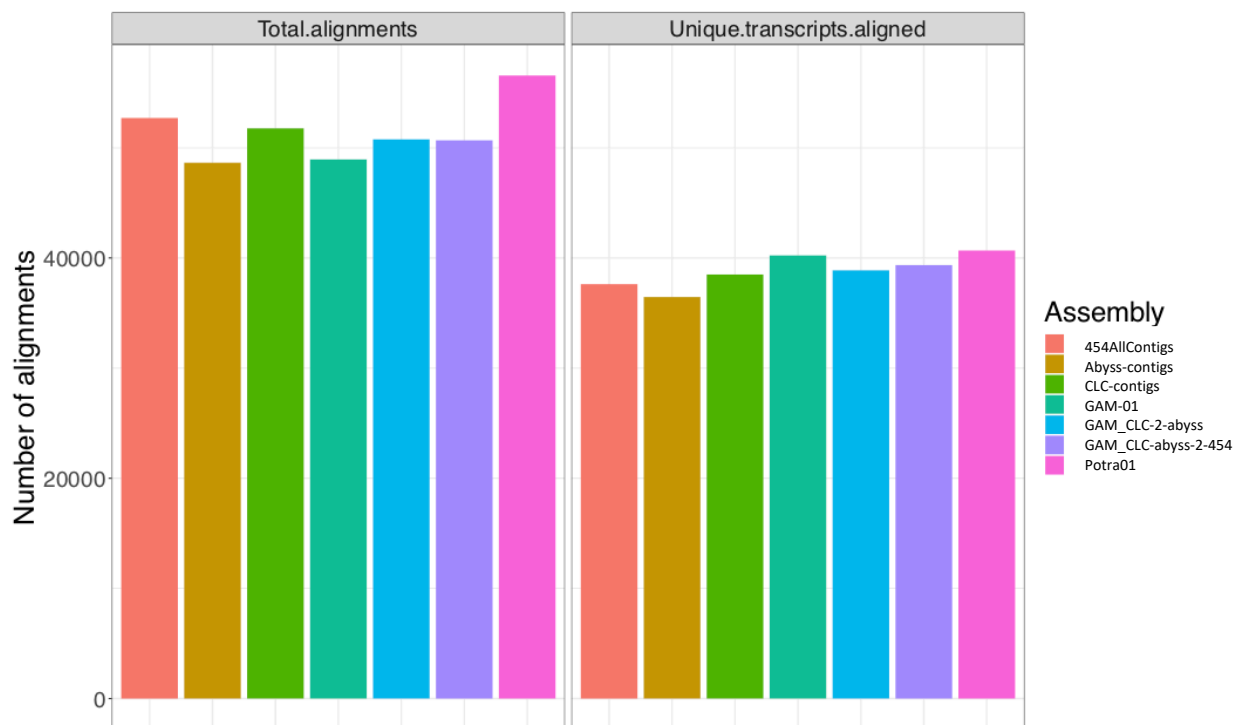


Figure S2.4 GMAP alignments of Potri primary transcripts to the Potra assembly sub-stages.

We similarly assessed the Potrs assembly, which revealed generally lower alignment rates and gene space contiguity than Potra, although the differences were not substantial. As such, we conclude that although the assemblies were all highly fragmented, the gene space is assembled contiguously and is well represented.

2.8.2 Genomic coverage

In order to evaluate the genomic coverage of the Potra and Potrs assemblies, we aligned all PE libraries to the unfiltered finished assemblies (Potra01, Potrs01 and Potri03) using `bwa-mem` and `bwa-aln` (v0.7.8). We calculated coverage values using the BEDtools (18) (v2.19.1) `genomecov` function at each position of the genome (`-d` option) for the aligned and sorted BAM file (`-ibam` option). We created the required BEDtools genome files in R from the genome FASTA files (see script: `bedtools_genome_from_fasta.R`).

Further, we converted the three annotation GFF files (see below for annotation details for Potra and Potrs) to BED format and used these to calculate per position coverage of the features in `genome.gff3`, `repeats.gff3` (repeatmasker annotation) and the `1Kb_regulatory.gff3`, the latter being all regions 1Kb up- and downstream of genes, unless other feature boundaries were encountered (another gene, end of scaffold). We created R data objects (scripts: `bedtools_tremula.R`, `bedtools_tremuloides.R`, `bedtools_trichocarpa.R`) for each alignment, resulting in 18 comparisons after post-filtering the scaffolds identified as contaminants. For the three assemblies, different genomic feature types were defined:

Potra: exon, gene, tRNA, mRNA, intron, repeat, downstream_regulatory, upstream_regulatory, miRNA

Potrs: exon, gene, tRNA, mRNA, intron, repeat, downstream_regulatory, upstream_regulatory

Potri: exon, gene, mRNA, intron, repeat, downstream_regulatory, upstream_regulatory

Boxplots (Figure S2.5) representing the data highlighted the generally poor coverage for aspen-to-*P.trichocarpa* alignments in intergenic regions. Gene regions (exon, intron, gene, mRNA), however, were well covered across all comparisons. Only the Potri assembly showed consistent coverage across all features when considering Potri reads aligned to the Potri assembly. The aspen assemblies suffered both in intergenic and repeat regions.

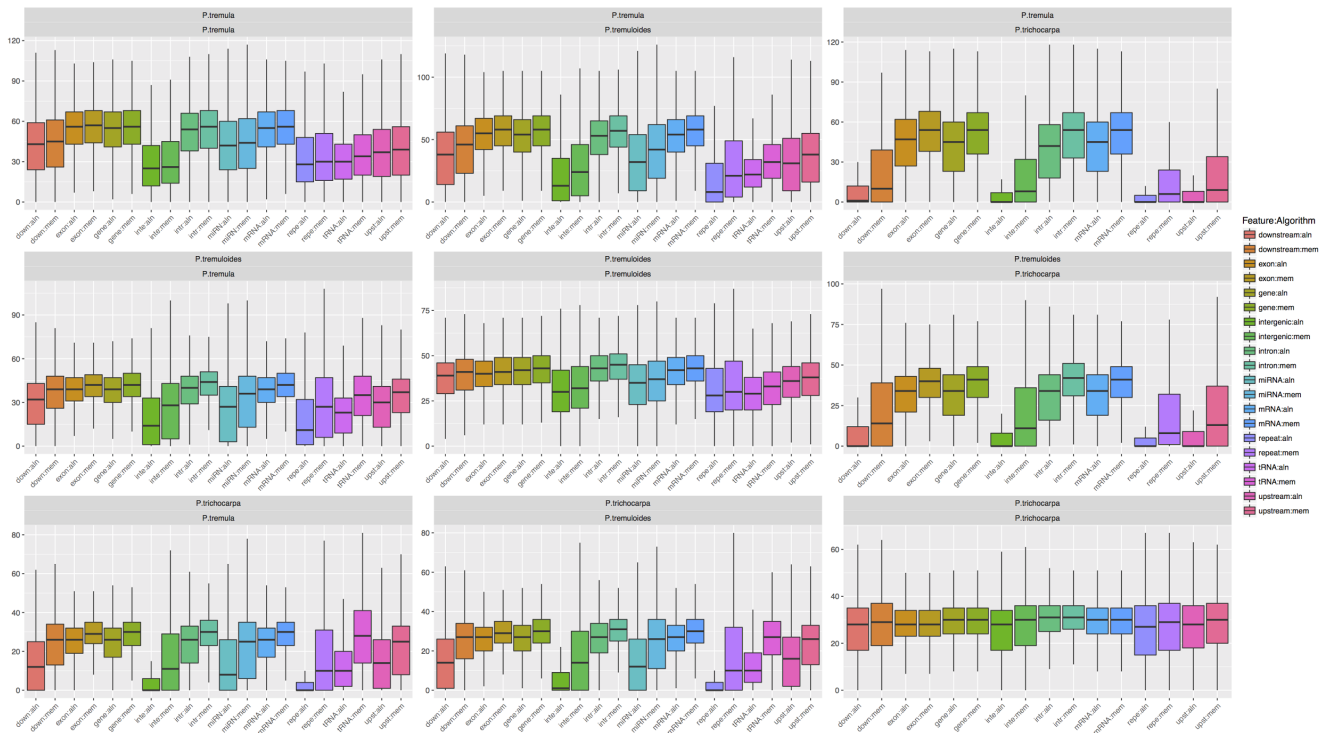


Figure S2.5 Genome coverage plots for reads from Potra, Potrs and Potri aligned to each genome in all pairwise combinations and to their own assemblies.

A visual comparison of bwa-aln alignments provided insight as to why the performance seems poor in comparison to bwa-mem when aligning across species (ex. Figure S2.6, Potra aligned to Potri). A high density of polymorphisms likely results in reads not being seeded in bwa-aln, whereas the bwa-mem algorithm employs seeding via a super-maximal exact match strategy, which performs better in difficult-to-align regions. We subsequently identified polymorphisms in the cross genome alignments using the HaplotypeCaller implemented in GATK (19) (v3.5).

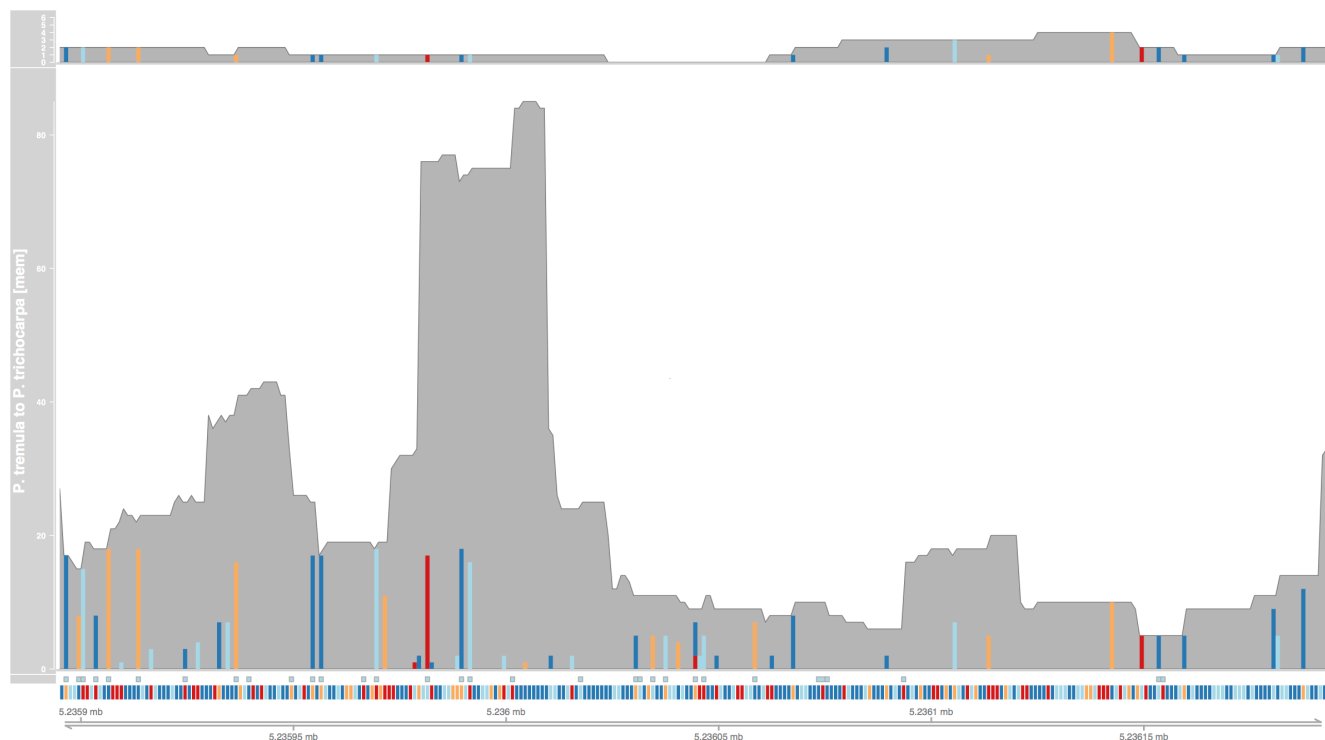


Figure S2.6 Alignments of all PE libraries of Potra to the Potri reference genome for a 300 bp region of Chr19. The bottom panel shows read alignments using bwa-mem, the top panel bwa-aln. Alignment depth is indicated by the grey region. The reference sequence is represented by colored bars below the depth coverage plot where A=light blue, C=light orange, G=red, T=dark blue. Within the coverage profile, the number of reads supporting an alternate base are indicated by colored bars. Figure created using script: GViz_coverage_aln_vs_mem.R

We sub-divided genomic coverage of unfiltered scaffolds by size groups of 500 bp, 1 Kbp, 2 Kbp, 5 Kbp, 10 Kbp, 20 Kbp, 40 Kbp, 80 Kbp, 160 Kbp, 320 Kbp, 640 Kbp, 1.28 Mbp and >1.28Mbp (Figure S2.7, S2.8). With the exception of the smallest two size groups, all self-alignments behaved as expected, showing a peak at the genome coverage. There was, however, an additional peak at 10x the expected read coverage (*i.e.* at 1,000X) in larger size groups, which might represent contaminants. See script size_group_simple.R. Further examination of the scaffold bins revealed that almost all scaffolds <2 Kbp represented split haplotypes, scaffolds 2–10 Kbp contained a mix of split and merged haplotypes and scaffolds >10 Kbp contained a greater proportion of merged than split haplotypes.

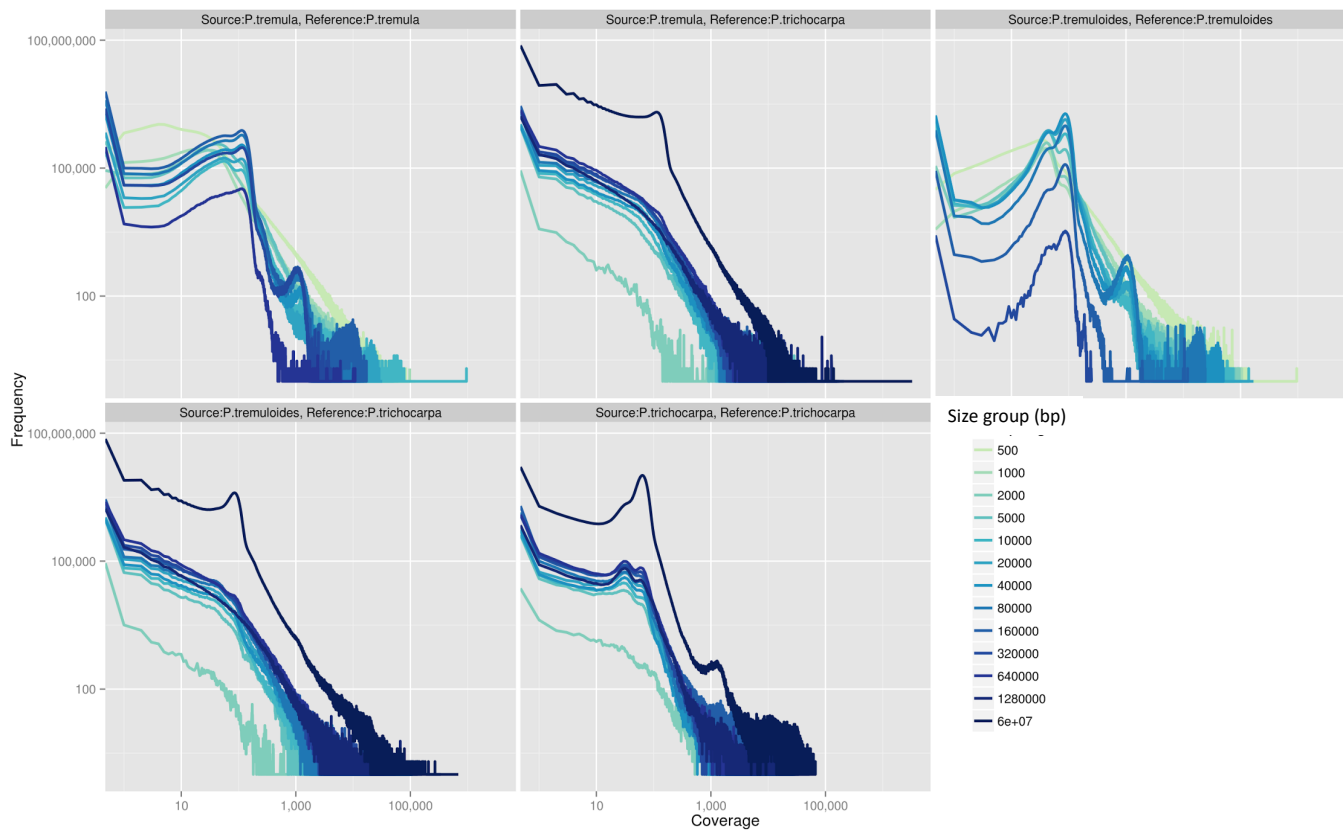


Figure S2.7 Genomic coverage for cross-genome and self-alignments broken down by scaffold size groups. Short scaffolds show different curves, indicating assembly issues, while longer scaffolds have a peak at the expected read coverage in self alignments. A second, smaller local maximum exists for self-alignments at roughly 10x the read coverage (~1000X), which might be contaminants.

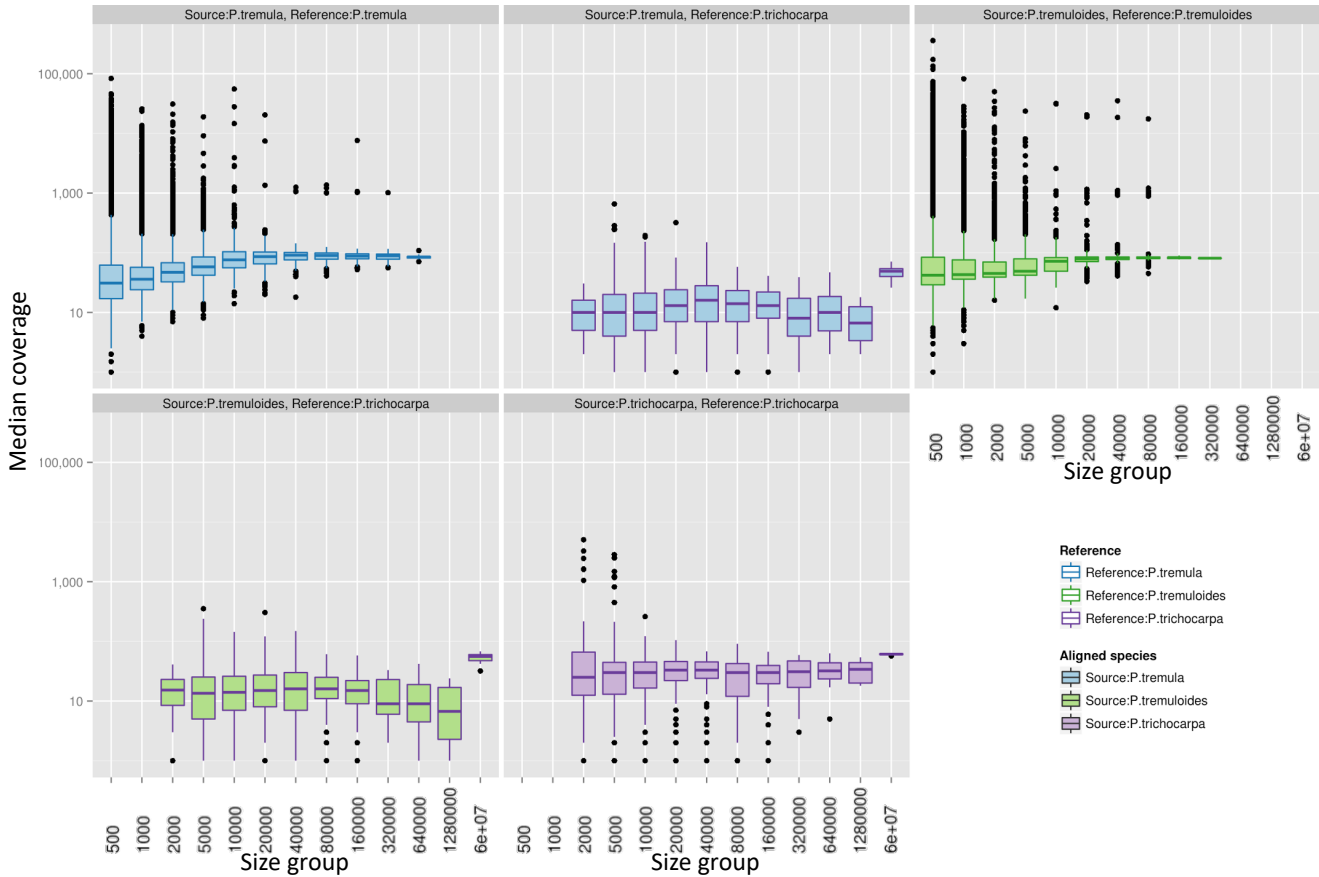


Figure S2.8 Distribution of scaffold coverage medians for cross-genome and self-alignments for scaffolds of various size classes. In the aspens, longer scaffolds approach the expected coverage while shorter ones show high variability.

2.8.3 Removal of redundant scaffolds

To remove redundant scaffolds from the Potra and Potrs assemblies we performed a self-self BLAST (20) (blast/2.2.29+, default settings). We removed scaffolds that were completely contained in another scaffold (cumulative coverage = 1) with 100% identity, and flagged scaffolds with >97% cumulative coverage and >97% percent identity as putatively redundant. These scaffolds represent assembly artefacts. See scripts `checkPtremulaRedundancy.R`, `checkPtremulooidesRedundancy.R` and `evaluateSelfSelfBLAST.R`.

2.8.4 Potra reads that did not align to the Potra assembly

We extracted Potra Illumina PE reads with no valid alignment to the Potra assembly (using `bwa-mem`) and aligned these to the NCBI nt database (downloaded 2015-03-18) using BLAST (v2.2.29+). The resulting data was interpreted in R (script: `self_self_BLAST_missing_taxonomy.R`). This analysis revealed that most unaligned reads originated from fungal contaminants, particularly of the *Melampsora* genus (Figure S2.9), which is a common fungal pathogen of *Populus* species. Other reads had high homology matches to *Populus*, but it is unclear whether these are missing in the assembly or represent unfiltered contaminants in the BLAST reference.

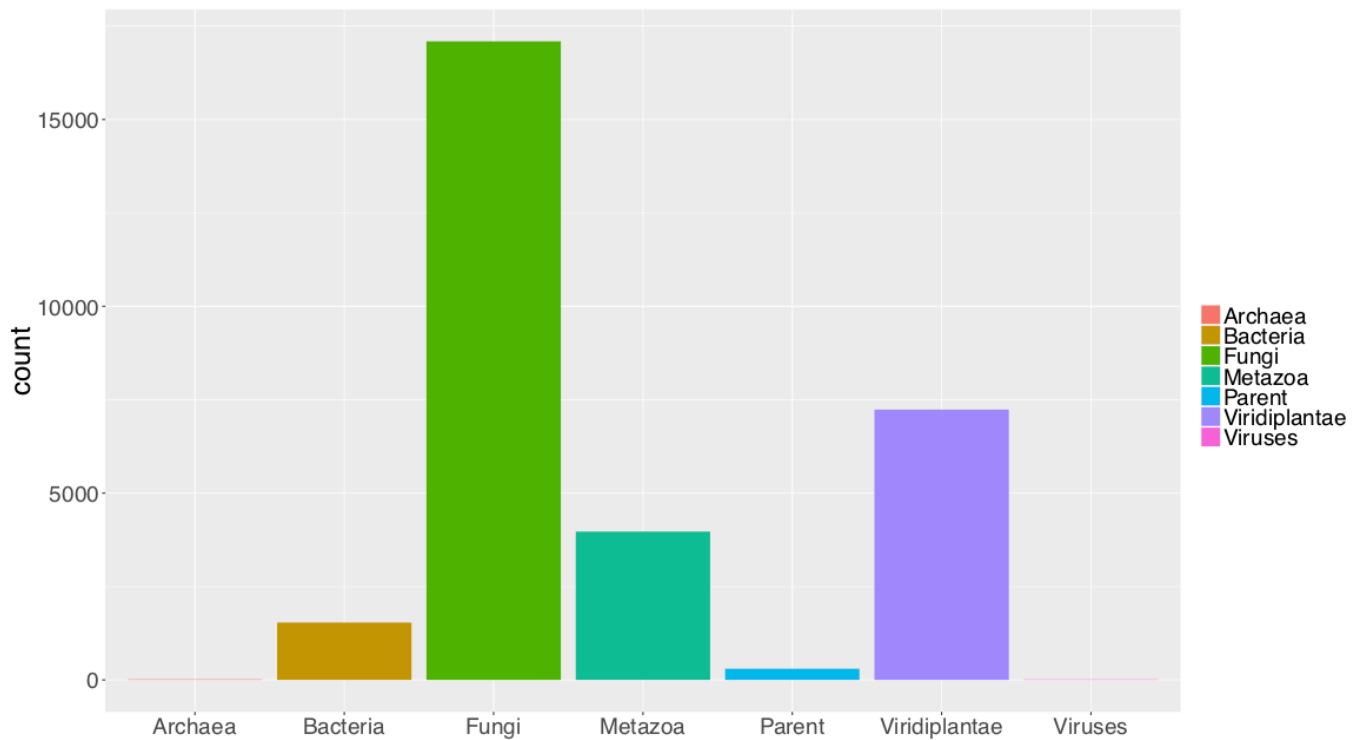


Figure S2.9 Read counts resulting from BLAST sequence homology searches of unaligned reads from Potra to the NCBI nt database. The data were tabulated by taxonomic division.

2.8.5 Cross-aspen EST alignment

We aligned 19,633 Potra and 5,730 Potrs ESTs to their own assemblies and across assemblies using GMAP (v14.12.29; -K 11000). The GMAP output was analysed in R (script: GMAP_EST.R). Overall, we observed no discernible difference when aligning ESTs to their corresponding aspen genome or to the foreign aspen genome (Table S2.6, Fig. S2.10):

Table S2.6 GMAP path number frequency for self- and cross-species alignment of ESTs. Both aspen genomes produced highly similar results.

Path number	Potra - Potra	Potra - Potrs	Potrs - Potra	Potrs - Potrs
1	16,228	4,706	15,843	4,737
2	1,043	656	1,142	572
3	132	75	163	58
4	46	18	52	17
5	67	42	99	26

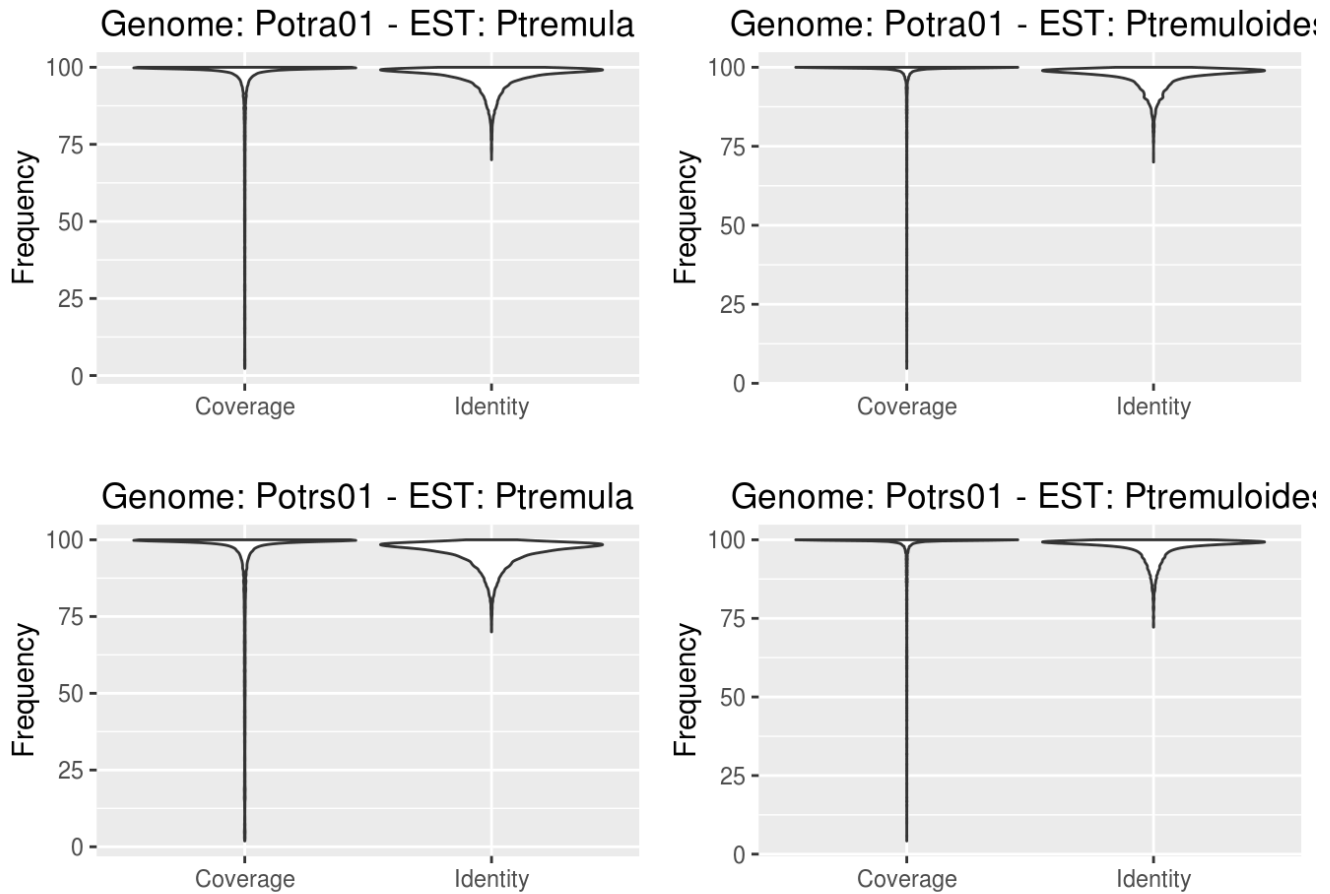


Figure S2.10 GMAP identity and coverage distribution when aligning ESTs to their own or to a foreign aspen genome. Both genomes produced highly similar results.

2.8.6 Assembly GC content statistics

We investigated the GC content of the Potra assembly using the script: `GC_length.R`, by plotting the relationships of GC to assembly length and coverage (Figure S2.11- S2.13). To compute coverage, we used BEDTools as previously described. We computed mean scaffold coverage using the script `streamMeanCtg.cpp`. There was a slight bias of GC, with contigs of higher overall GC having higher coverage. In addition, we analysed the unfiltered genome to compute the same statistics for the set of scaffolds that were filtered from the assembly as contaminants (Figure S2.14- S2.16).

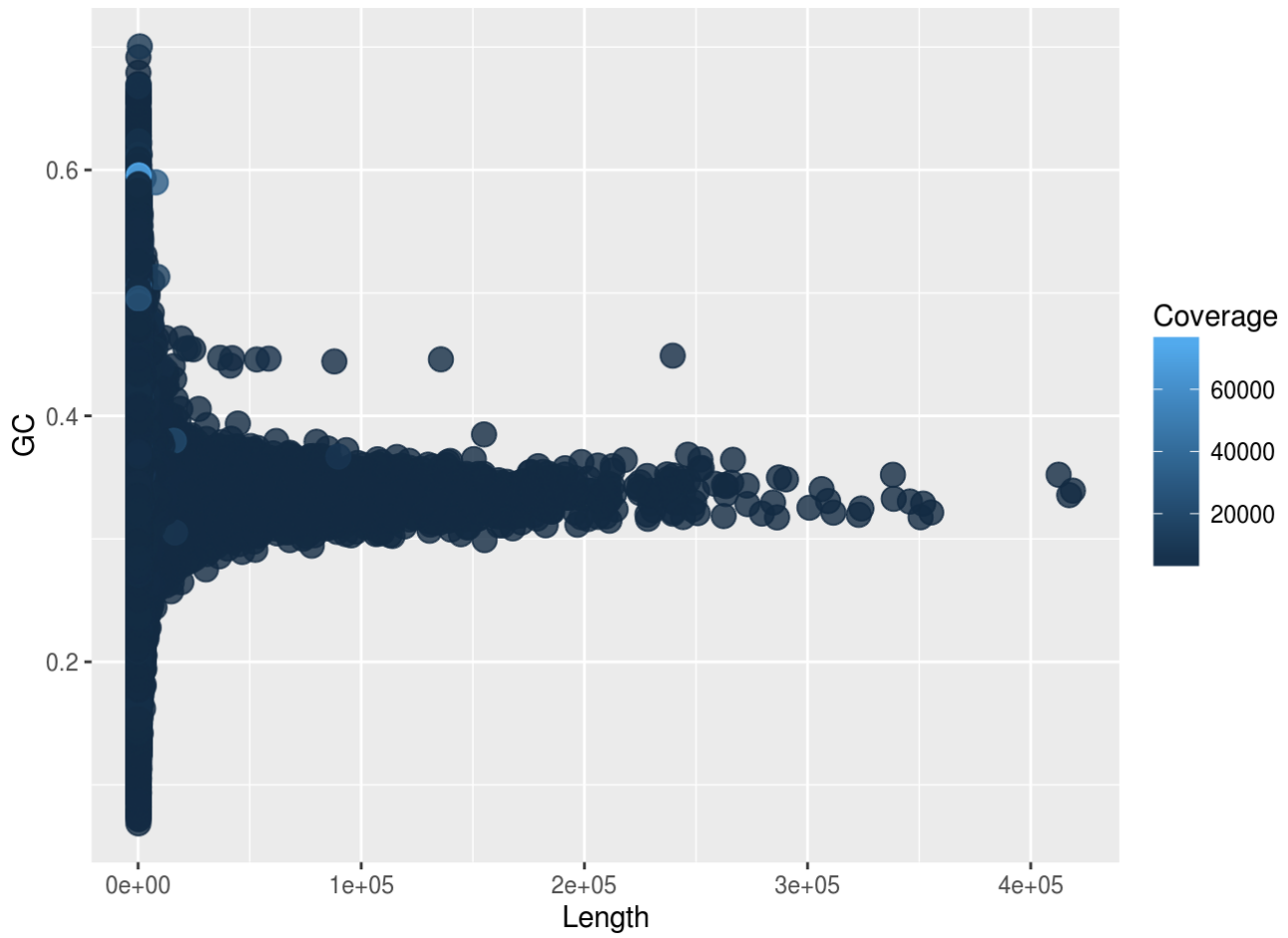


Figure S2.11 Final assembly GC ratio vs scaffold length. Points are shaded by coverage. There was no clear effect of GC content on the ability to construct long scaffolds. Data points at ~50% GC likely represent assembled plastid genome fragments.

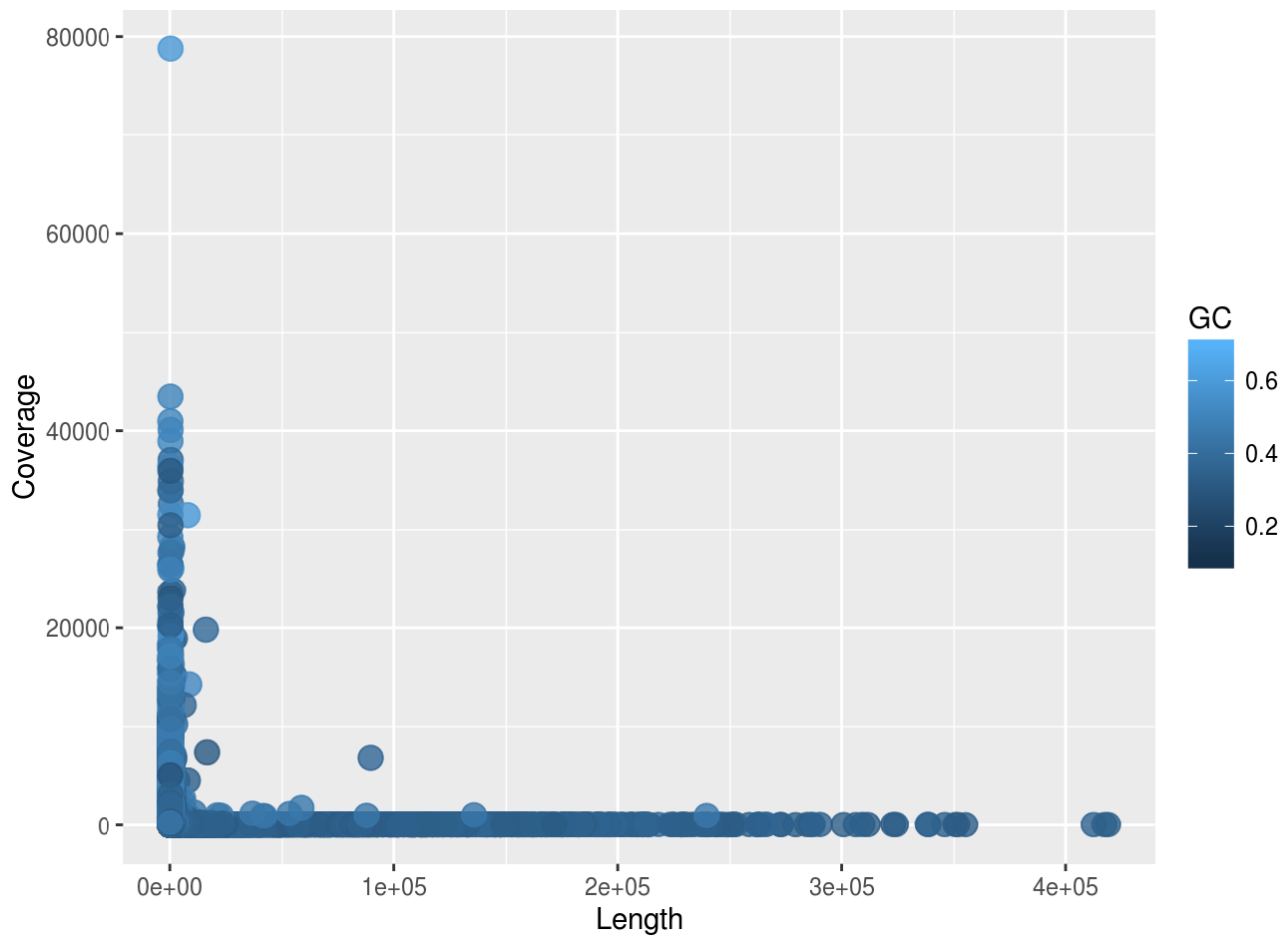


Figure S2.12 Final assembly Coverage vs scaffold length. Points are shaded by GC ratio. Shorter length scaffolds attract a larger number of reads. The extremes (short and high coverage) are rRNA gene loci, which are collapsed to a consensus copy in the assembly.

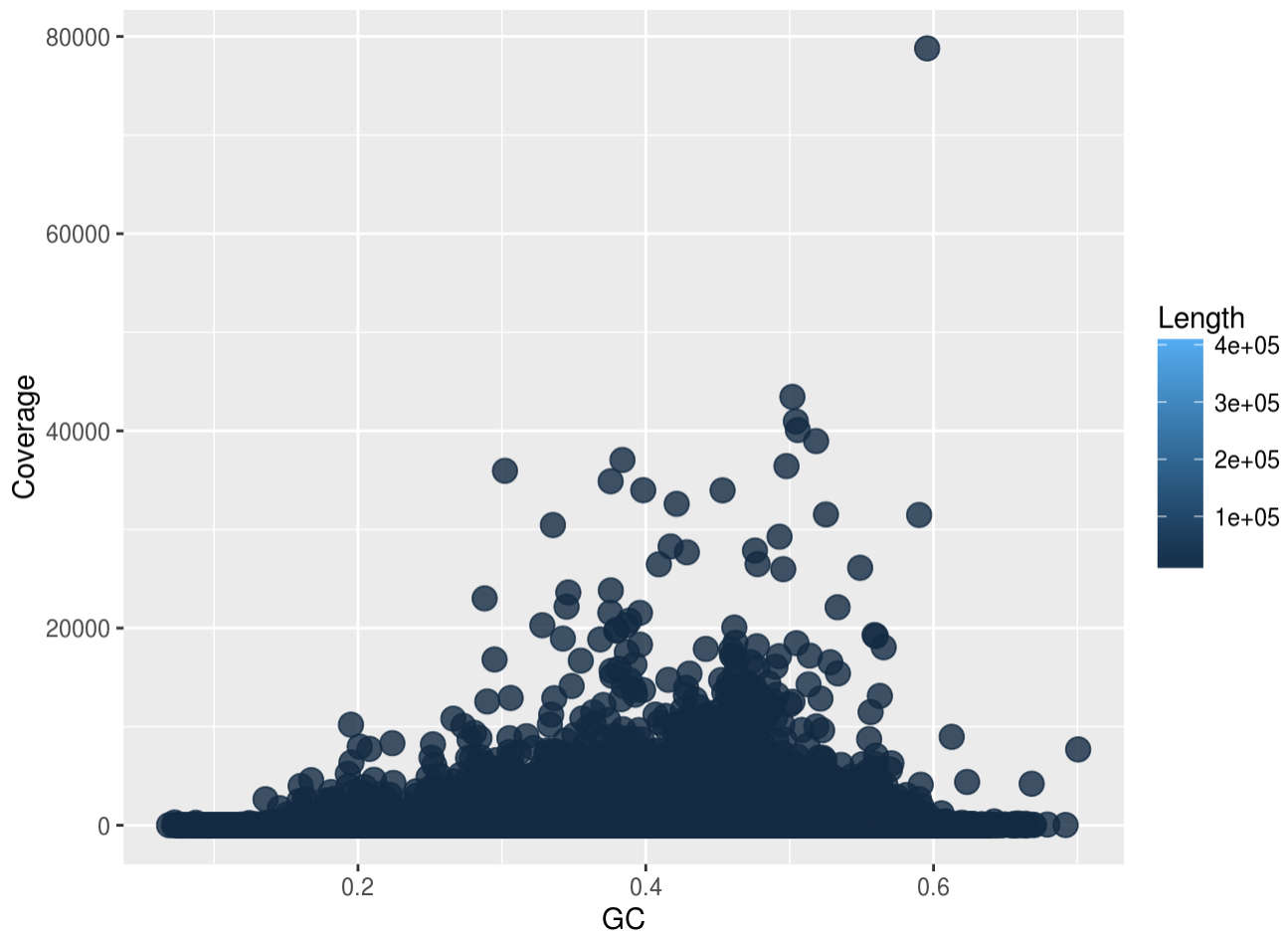


Figure S2.13 Final assembly coverage vs GC ratio. Points are shaded by length. There is a slight effect of GC content on coverage. This is possibly related to the known Illumina GC bias, esp. given the older chemistry of the sequencing data.

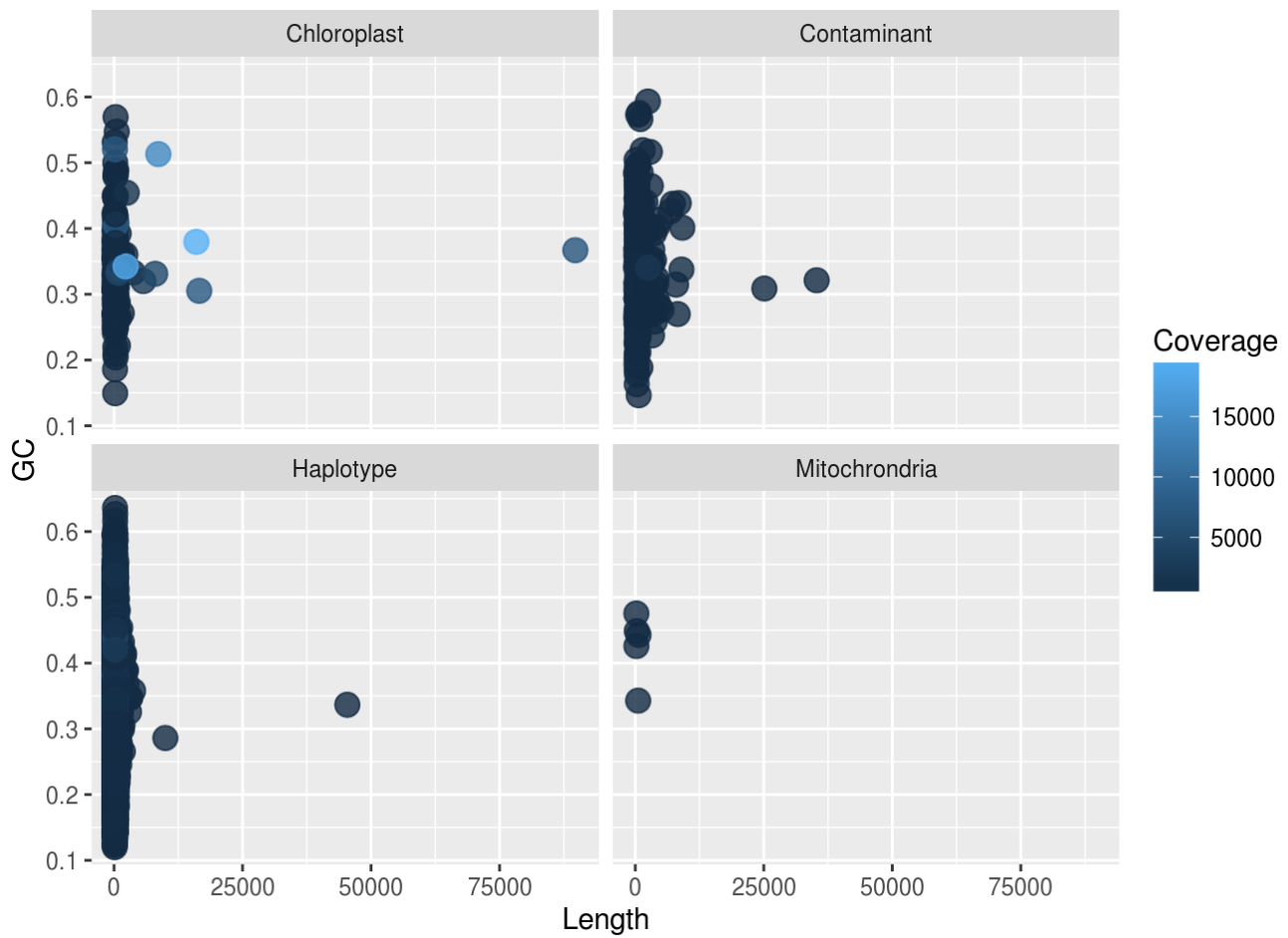


Figure S2.14 Contaminant GC ratio vs scaffold length. Points are shaded by coverage. In general, contaminants are represented by short scaffold lengths, but spread over the range of GC content. Surprisingly few mitochondrial reads were detected. Predicted haplotype scaffolds were identified using self-self sequence homology searches.

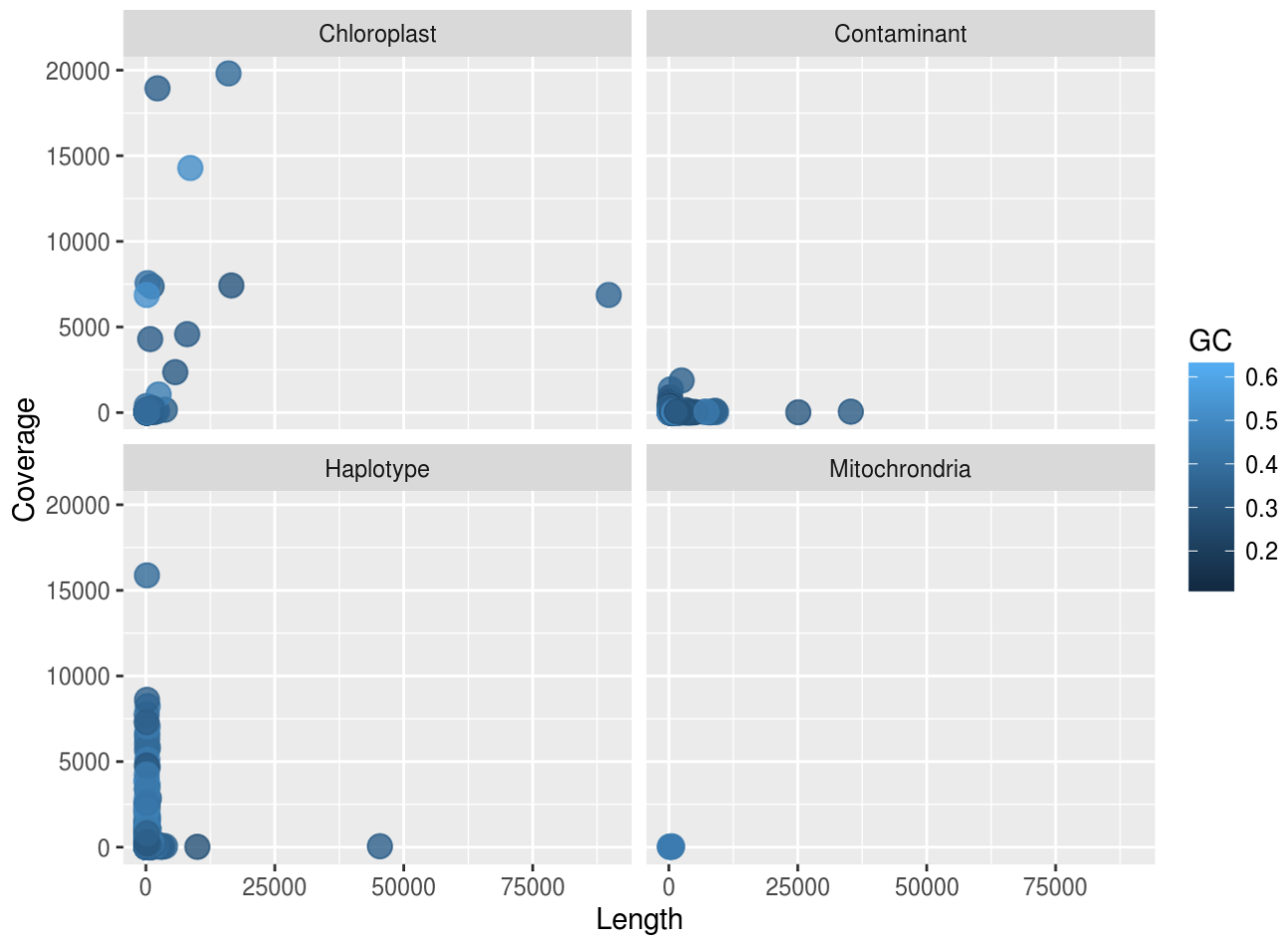


Figure S2.15 Contaminant Coverage vs scaffold length. Points are shaded by GC. Very few mitochondrial reads were identified, whereas short haplotypes and the chloroplast show high coverage. Predicted haplotype scaffolds were identified using self-self sequence homology searches.

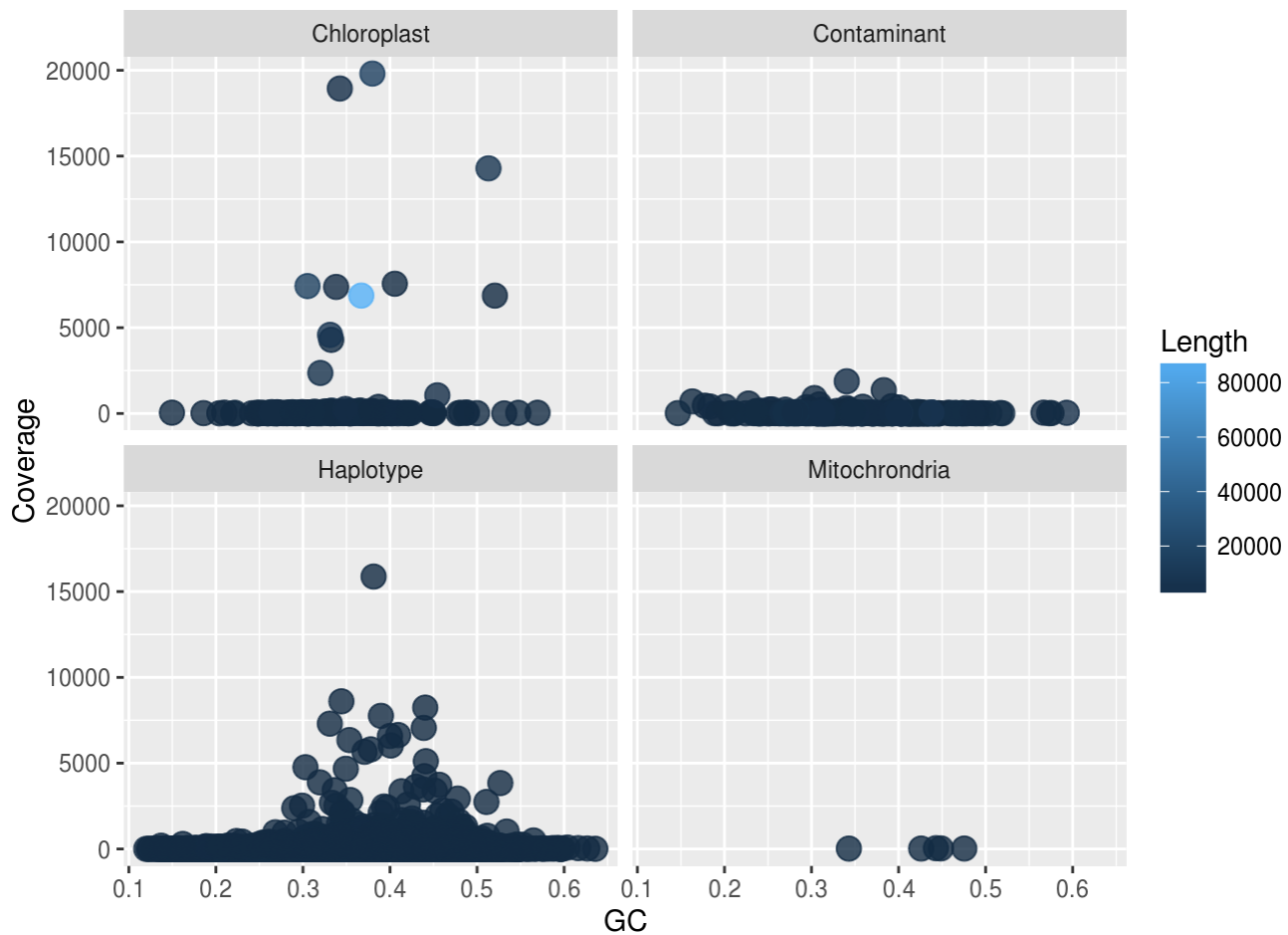


Figure S2.16 Contaminant coverage vs GC ratio and shaded by Length.

We performed the same analyses for the Potrs, which yielded comparable results.

2.8.7 Feature Response Curves

In order to evaluate potential miss-assemblies introduced by the hierarchical assembly merging approach that we used for the Potra assembly, we used FRC_align (21) (v1.2.0; https://github.com/vezzi/FRC_align). Alignments (bwa-mem) of one PE library (150 bp insert, post trimming) and one MP library (10Kb insert, post trimming, de-duplication and reverse complementation) were supplied, defining $\pm 25\%$ of the expected insert size as an acceptable standard deviation. The genome size was specified as 380 Mbp. The output curves per statistic were then plotted in R (script: plot_frc_stats.R). Details of the various metrics are available in the section “Multiple library statistic” of the FRC publication supplementary information (<http://www.nada.kth.se/~vezzi/publications/supplementary.pdf>). The FRC plots (Figure S2.17) showed that the final merged assembly (GAM_asp201-001) performed better than the input assemblies in SPAN and COMPR statistics and performs at least as well as the input in most other categories. In the STRECH and OUTIE categories, however, it performs worse than the input, suggesting artefactual expansion and miss-joins, most likely arising from the scaffolding step.

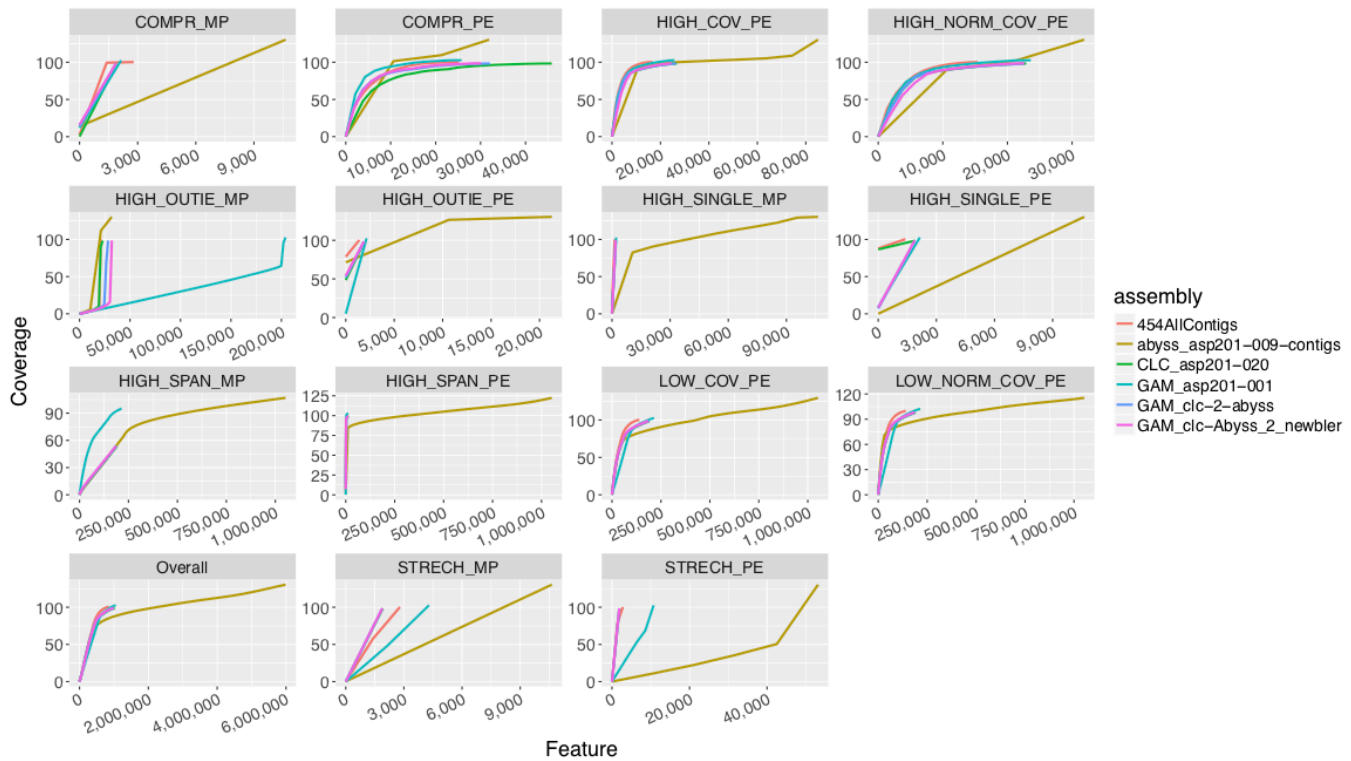


Figure S2.17 Feature Response Curve statistics comparison of input assemblies and all subsequent merging steps, as well as the final scaffolding. The input assemblies by CLC, ABySS and gsAssembler (Newbler) were merged via GAM-NGS and the final merged assembly scaffolded by BESST. Steeper curves indicate better performance in the FRC categories. In R, we then overlaid the FRCs with genomic features to investigate which features were prone to generate suspicious FRC statistics (Figure S2.18). Unsurprisingly, the vast majority of issues were detected in repetitive regions (script: `frc_feature_overlap.R`).

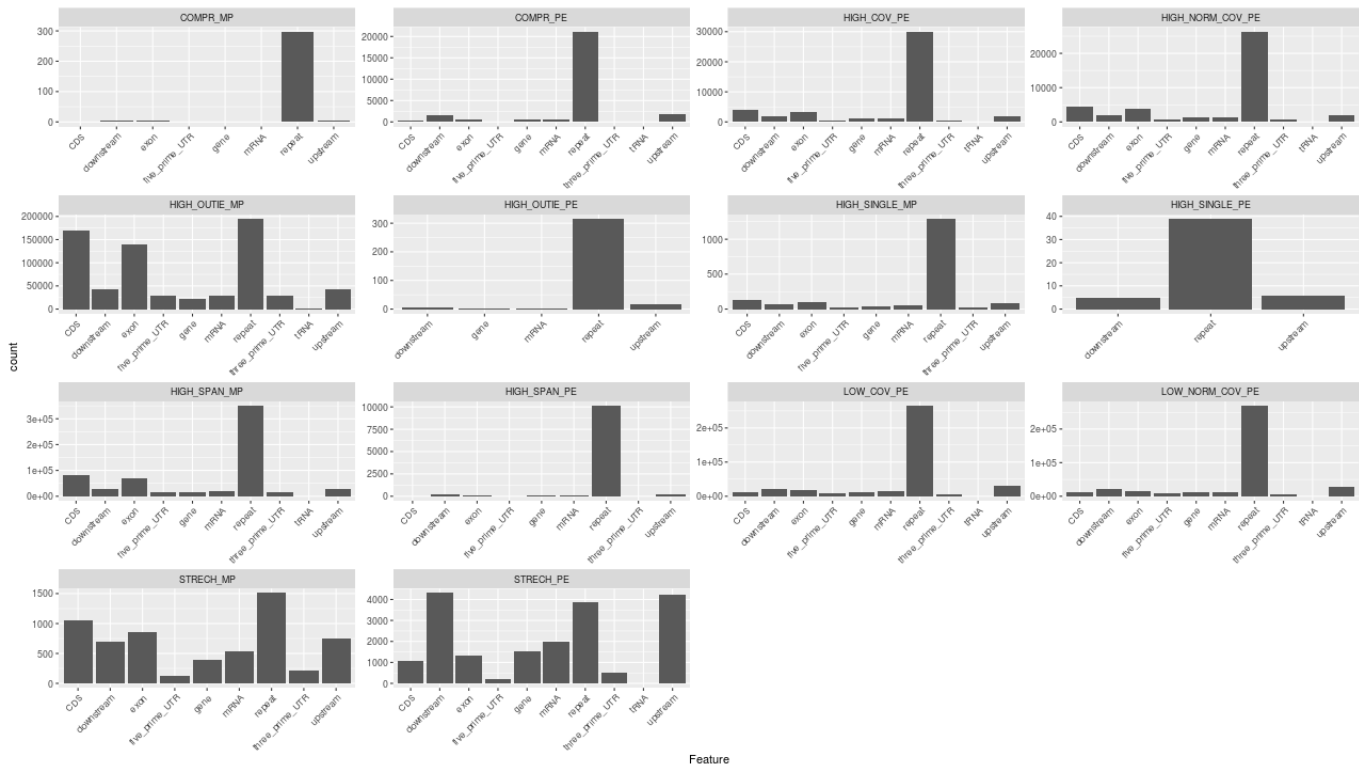


Figure S2.18 Overlapping suspicious regions from FRC with genomic features.

3 Genome and transposable element annotation

3.1 Repeat annotation

3.1.1 Methods

De novo repeats isolation and characterization

We searched the Potra and Potrs genome assemblies using RepeatScout (22). We retained repetitive sequences in each genome if <10 copies of identified repeats were present. In order to overcome the expected fragmentation of repeat consensus, we further assembled the output sequences by running cap3 (23) using relaxed settings (-o 30, -p 80, -s 500). We then characterised these assembled consensus sequences using the following pipeline:

- We used sequences as queries in tblastx homology searches against RepBase (24). If the sequence had a significant match (e-value < 10^{-10}) we classified it according to the subject it matched to, otherwise it was passed on to the next step;
- We used sequences as queries in blastX homology searches against NCBI or genbank. If the sequence had a significant match (e-value < 10^{-10}) with TE like sequences then we retained this hit and classified the sequence accordingly. If there were no significant hits we flagged it as “NHF” (“No hits found”). Importantly, if the

sequence had a significant hit with coding genes we discarded it as a putative member of a gene family. We additionally removed sequences of plastidial origin.

In doing so, we obtained 2,251 and 2,646 Potra and Potrs sequences, respectively. Further, we retrieved 148 Potri repetitive sequences from RepBase and added these to the two datasets. We then clustered this combined set of all these *Populus* spp. repetitive sequences using cd-hit, yielding a dataset composed of 3,678 sequences (Table S3.1).

Table S3.1 Identified repetitive sequences

TE class	#f sequences
Ty1-copia	616
Ty3-gypsy	500
Other LTR-RT	9
Non-LTR	90
SINE	10
DNA_TE (including MITEs)	111
NHF	2342
Total	3678

Phylogenetic analysis

We used 100 residues long amino acid sequences from the conserved Reverse Transcriptase domains of Ty1-copia, Ty3-gypsy and LINE elements to search the *Populus* spp. genome assemblies available. We retrieved all significant matches covering at least 80% of the query sequence length. We then collected a subset including 150 randomly selected paralogous per species sample and aligned these using the software Muscle (25). We used these alignments to build phylogenetic trees with the software MEGA6 (26), adopting the Neighbour Joining method and calculating bootstrap support values for 1000 replicates..

3.1.2 Results

TE abundance estimates across different Populus spp.

We used the repetitive library to estimate the Transposable Element (TE) content in the three *Populus* genomes. In the case of Potra, 21.54% of the genome was related to some kind of repetitive sequence, mostly TEs. The amount in Potrs was 22.09%. For comparative purposes, we also estimated the content in Potri, which was 29.42% (Table S3.2). From this analysis, we observed that Potri appeared to be enriched in TE-related sequences in comparison to the other two *Populus* spp. In particular Ty3-gypsy elements appeared to be more abundant in Potri (13.08%) than in the two aspens

(7.40% and 7.20% for Potra and Potrs, respectively). We did not observe the same for the other major TE classes. In order to rule out the possibility that the results were artefactual, i.e. due to some sort of technical driven TE depletion or enrichment in different assemblies, we took advantage of the availability of the large dataset of 454 random sheared reads from Potra. We assumed that this dataset represents a fair and unbiased sample of the complete Potra genome. When the screening for repetitive sequences was carried out on this sample, we observed that the overall TE estimate increased to 25.52% - a value that is fairly consistent with that calculated for Potri (Table S3.2). This evidence suggests that the difference in the TE content estimates is better explained by artefactual depletion of TEs in the two aspen genome assemblies than by an actual enrichment of Ty3-gypsy elements in the Potri genome. This depletion primarily involved Ty3-gypsy elements, which was the only TE class for which the difference in abundance was significant.

Table S3.2 Repeat element content as a percentage of total genome size.

Species	Total	Ty1-copia	Ty3-gypsy	Other-LTR	LINES	SINEs	DNA	NHF
Potra	21.54	4.00	7.4	0.13	0.38	0.36	3.28	5.99
Potra (454)	25.52	4.00	12.76	0.18	0.69	0.23	3.05	4.61
Potrs	22.09	4.02	7.2	0.11	0.35	0.38	3.43	6.60
Potri	29.42	4.19	13.08	0.18	0.70	0.39	4.72	6.16

Evolutionary dynamics of TEs amplification

In order to identify any evidence of recent species-specific, sustained TE activity in the three *Populus* species we focused on genome assembly regions that were unique to each single species. We identified a significant increase of TE-related sequences in the portion of the genome assembly unique to Potri (Table S3.3)

Table S3.3 Repeat element content as a percentage of total genome size for Potri vs aspens

Comparison	Total	Ty1-copia	Ty3-gypsy	Other-LTR	LINES	SINEs	DNA	NHF
<i>vs Potra</i>	48.28	7.84	17.26	0.32	1.29	0.91	10.73	9.93
<i>vs Potrs</i>	48.92	8.22	14.50	0.47	0.72	1.14	12.07	11.8
<i>Potri (all)</i>	29.42	4.19	13.08	0.18	0.70	0.39	4.72	6.16

In the case of the two aspen species we found that all the unique assembly fractions detected in cross species comparisons were underrepresented by TE-related sequences. This indicates that in Potra and Potrs there has been no appreciable TE activity since the speciation event separating them. Comparisons among species are shown for Potra and Potrs in Tables S3.4, S3.5

Table S3.4 Repeat element content as a percentage of total genome size for Potrs vs others

Comparison	Total	Ty1-copia	Ty3-gypsy	Other-LTR	LINEs	SINEs	DNA	NHF
<i>vs Potra</i>	16.65	3.00	4.16	0.08	0.20	0.45	2.60	6.16
<i>vs Potri</i>	13.41	2.28	2.88	0.05	0.09	0.44	2.21	5.46
<i>Potrs (all)</i>	22.09	4.02	7.2	0.11	0.35	0.38	3.43	6.60

Table S3.5 Repeat element content as a percentage of total genome size for Potra vs others

Comparison	Total	Ty1-copia	Ty3-gypsy	Other-LTR	LINEs	SINEs	DNA	NHF
<i>vs Potrs</i>	17.63	4.52	5.55	0.11	0.32	0.36	2.88	3.89
<i>vs Potri</i>	15.39	3.13	4.35	0.05	0.17	0.37	2.46	4.86
<i>Potra (all)</i>	21.54	4.00	7.4	0.13	0.38	0.36	3.28	5.99

Phylogenetic analysis of retroelements in Populus spp.

We found that Ty1-copia and Ty3-gypsy NJ phylogenetic trees were quite similar in terms of evolutionary insights provided (Figures S3.1, S3.2). In both cases we did not identify any Potri specific clade. Also, we found that the distribution of the paralogs retrieved from the Potra random sheared 454 library did not suggest that just a single LTR-RT family could explain the different abundance of the Ty3-gypsy elements in Potri. As such, we expect that any technically induced sampling bias in the Potra (and Potrs by extension) assembly would affect all LTR-RT families. For all three species, most of the paralogs were inter-mixed, demonstrating that the LTR-RTs complement is ancient and shared among them. In the case of LINE elements, we were able to identify a Potri abundant and species-specific clade, possibly representing a set of recently amplified elements.

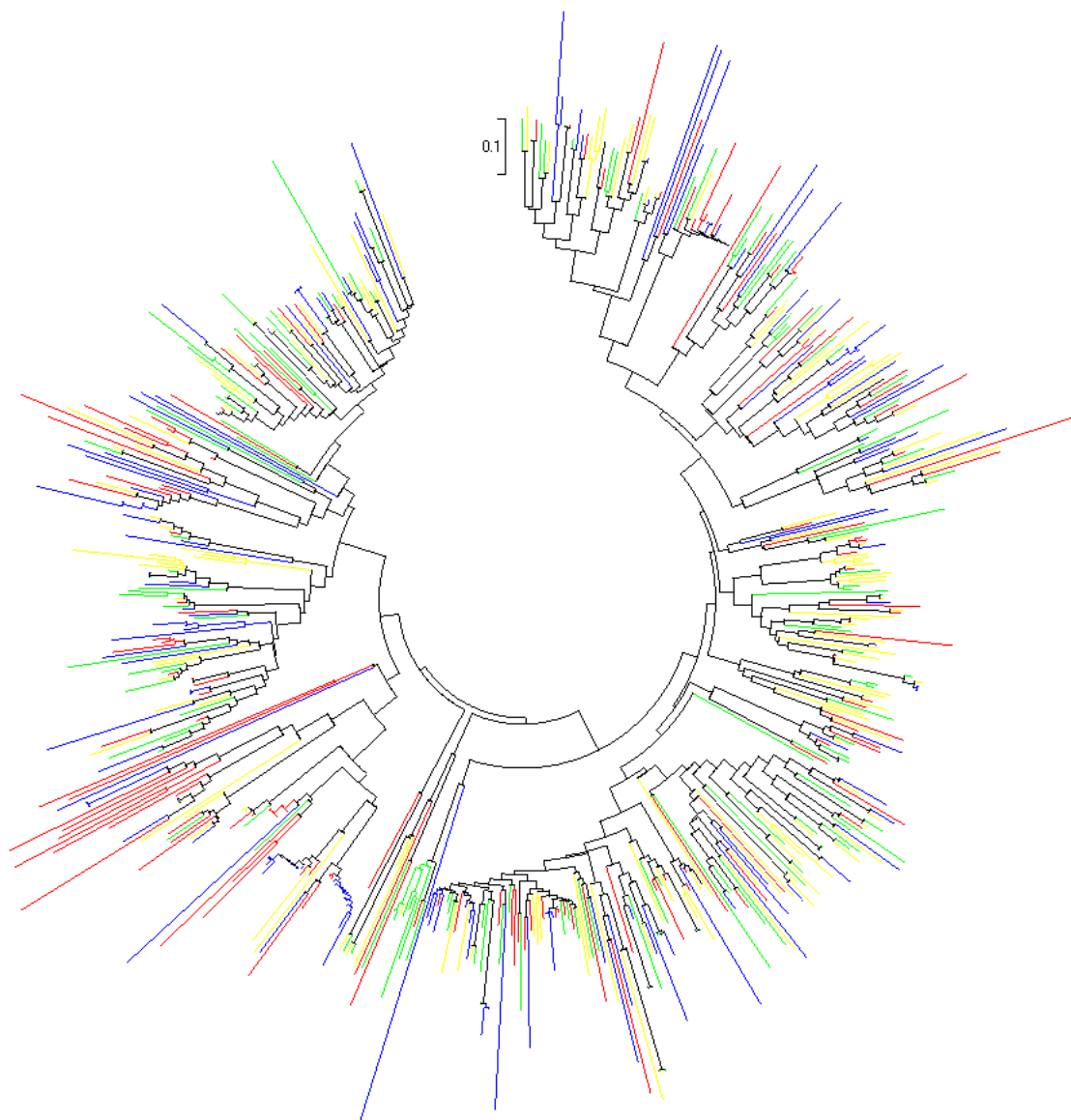


Figure S3.1 Ty1-copia. Red: Potri; yellow: Ptrs; green Ptra (assembly); blue Ptra (454 reads).

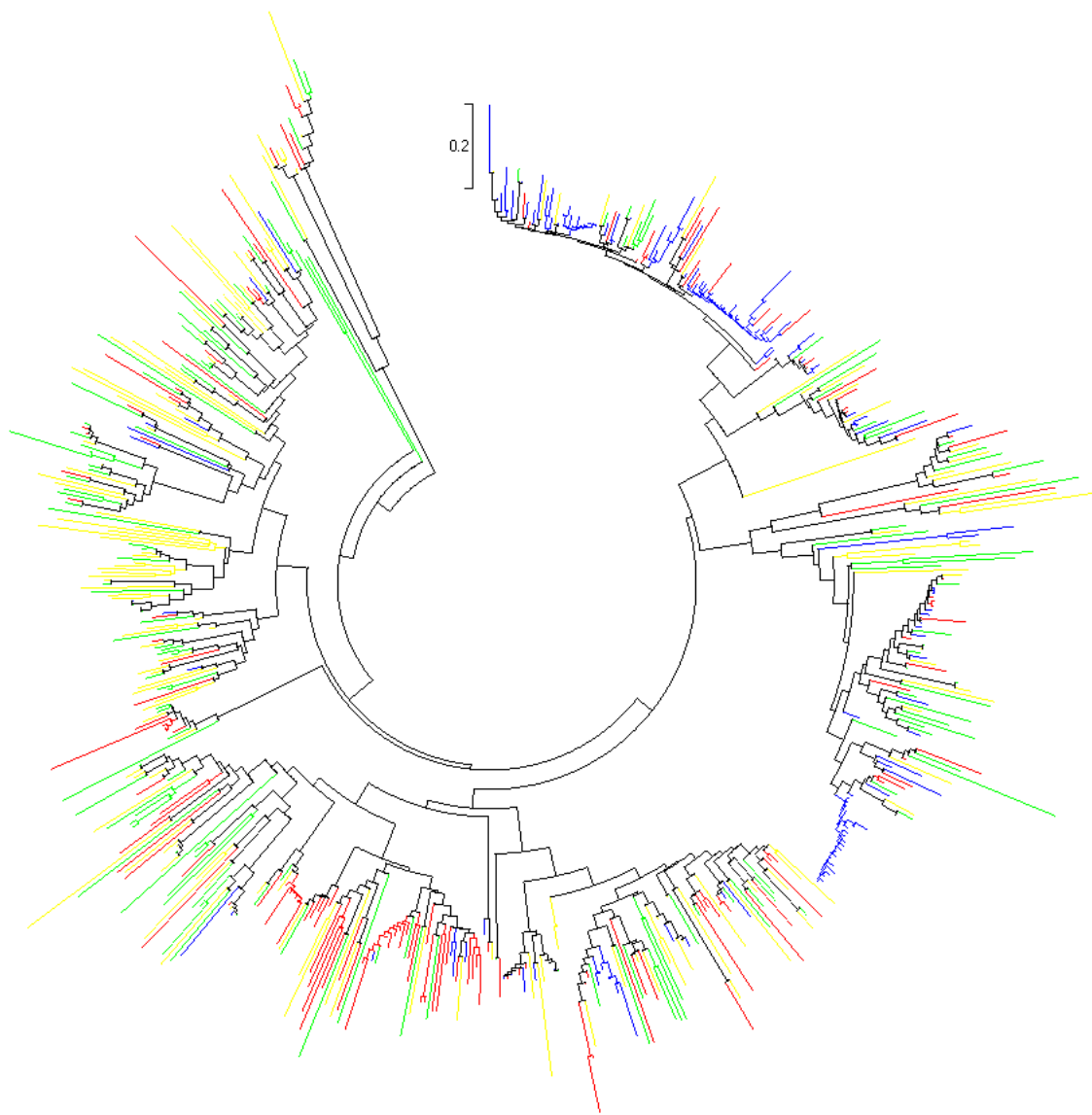


Figure S3.2 Ty3-gypsy. Red: Potri; yellow: Ptrs; green Ptrs (assembly); blue Ptrs (454 reads).

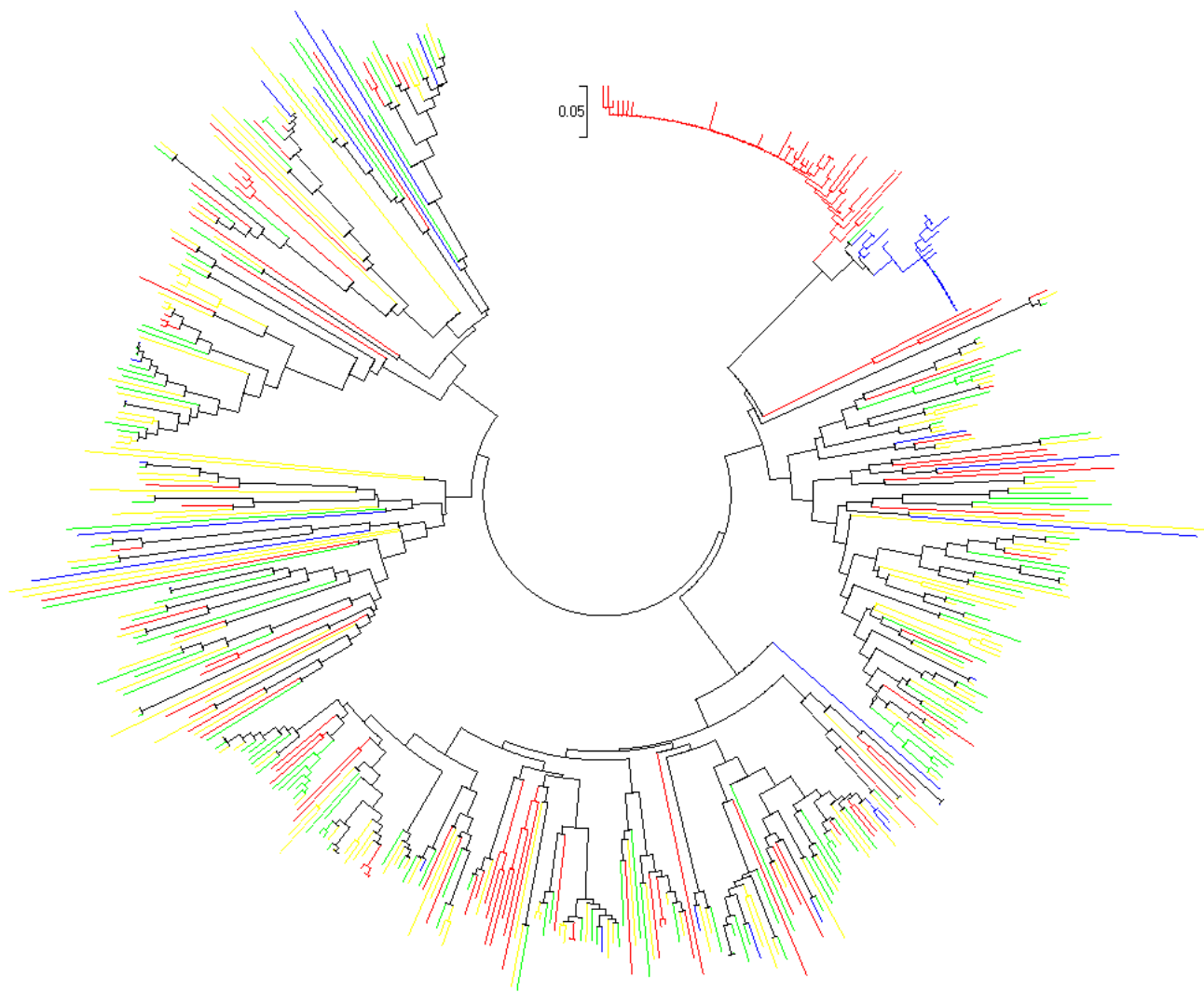


Figure S3.3 LINEs. Red: Potri; yellow: Ptrs; green Ptra (assembly); blue Ptra (454 reads).

3.2 Gene annotation

Briefly, we performed gene space annotation in three steps: first we used Maker (27) to generate gene annotations, which we iteratively refined using PASA (28) on the basis of the four transcriptome datasets detailed above before final manual curation (details below).

3.2.1 Gene annotation

We sourced input data for the annotation of the aspen assemblies from three complementary sources. First, we retrieved reviewed, full-length protein sequences from the uniprot database (29) for species belonging to the taxonomic group *spermatophyta*. This data set was combined with the more generic, curated uniprot-swissprot reference protein set. We then complemented our reference sequence set with RNA-Seq detailed below (4.2.2) and with publically available EST sequences for all poplar species having >5,000 ESTs, as available from the PlantGDB (30). In addition, to

ensure high specificity of the subsequent alignment and gene building steps, we compiled a library of modelled repeat sequences from the Potra assembly (see above) and combined it with curated repeat sequences for the genus *Populus* included with the RepeatMasker package (31).

We then computed the annotation of the aspen assemblies using the Maker package (32) (version 2.31.6) in two stages. First, we created an ‘evidence build’ directly from the provided reference sequences (Maker options `est2genome` and `protein2genome=1`). From this build, we manually selected and curated 1000 gene models for training the AUGUSTUS *ab-initio* gene predictor (33) (version 2.7). We used the resulting *ab-initio* profile to perform a second-pass annotation of the evidence-build to fill in missing gene loci and replace evidence-based models for which a transcript with a longer protein-product could be predicted from the data. We found this strategy yielded a gene build with a high degree of correspondence to the high-quality transcriptome data while still utilising the power of *ab-initio* gene predictions for cases where the evidence-based annotation yields unsatisfactory results due to missing or incomplete data.

For the functional inference of computed gene models, we extracted protein sequences for each annotated mRNA and subjected them to BLAST searches against the uniprot-swissprot reference database. In addition, we ran InterProScan (34) (version 5.7-48.0) to identify additional features, including Pfam domains, ProDom domains and information from the PIRSF and Superfamily databases.

3.2.2 Annotation refinement

We utilised four RNA-Seq datasets for the purpose of gene space annotation polishing using the PASA pipeline (Table S3.6). (i) The exAtlas resource (35), which comprises samples collected from a range of tissues at different developmental stages or under contrasting biotic and abiotic stresses. (ii) The exDiversity dataset, which comprises mature leaves sampled from a collection of wild-growing, sexually mature Potra individuals in the Umeå (Sweden) region (36). Both of these datasets are available within the PopGenIE.org expression visualisation tools. In addition, we utilised two in-house and currently unpublished datasets. The first profiles neo- and pre-formed leaf development in Potra, while the other profiles gene expression in xylem scrapes from a subset of the SwAsp genotypes (see below) grown under greenhouse conditions. Both datasets were generated using standard Illumina RNA-Seq protocols from polyA selected total RNA. The data was not strand-specific and was sequenced to a minimum depth of 10M PE reads per sample. These samples will be detailed elsewhere and were used here only to maximise coverage of the transcriptome to ensure as comprehensive an annotation as possible.

Table S3.6: Description of the datasets used for the PASA iterations

Dataset	Tissue	Number of samples	ENA Accession	Number of digitally
---------	--------	-------------------	---------------	---------------------

				normalised reads
exAtlas	Multiple	24	PRJEB5040	34,649,686
exDiversity	Mature leaves	17	PRJEB1790	27,409,045
Xylem/leaf	Xylem and mature leaves	8	PRJEB28867	58,532,489
Leaf	Developing leaves	58	PRJEB28866	24,673,292

We performed an *in-silico* normalisation (using trinity (37), with non-default arguments `--max_cov 50 --pairs_together`) on all four sets of RNA-Seq data to retain a non-redundant set of reads for each dataset, which were then assembled using both cufflinks (38) (with non-default parameters: `--max-intron-length 11000 --library-type fr-unstranded --multi-read-correct`) and trinity (with non-default parameters: `--min_kmer_cov 1`). We performed two separate trinity assemblies; one *de novo* and one genome-guided (with the extra parameter: `--genome_guided_max_intron 11000`). The assembly statistics are available in Table S3.7. We then iteratively used the three assemblies obtained per dataset to refine the annotation using PASA, following the order in Table S3.6. Every PASA run consisted of a double pass; a first iteration integrating novel annotations into the existing annotation and a second pass refining these updated annotations, as recommended by the PASA guidelines. After every iteration, we verified the refined gff3 annotation file and corrected errors using genomertools (39). Briefly, we first ran `$PASAHOME/scripts/launch_PASA_pipeline.pl` with non-default parameters: `-R -t transcripts.fasta.clean --ALIGNERS blat,gmap --TDN tdn.accs --cufflinks_gtf cufflinks -u transcripts.fasta -T --TRANSDECODER`; with the transcripts.fasta.clean file obtained using `seqclean` on the concatenation of the trinity assemblies, the tdn.accs obtained using: `$PASAHOME/misc_utilities/accession_extractor.pl` and the cufflinks file obtained previously. This was followed by the gff3 curation and a first run of `$PASAHOME/scripts/launch_PASA_pipeline.pl` with non-default parameters: `-A -t transcripts.fasta.clean -L --annots_gff3`. This process (curation and annotation) was repeated once, and the obtained gff3 curated for the next dataset iteration. The gff3 curation was done using genomertools `gt gff3` with non-default parameters: `-force -tidy yes -addintrons yes -addids yes -fixregionboundaries yes -retainids yes -sort yes -checkids yes`. The manual curation scripts are available in the Git repository: 3.2.2-Annotation-refinement directory. We finally manually corrected a small number of incoherent cases. We tracked all gene IDs across iterations to provide supporting evidence during the curation. Ultimately, after the final PASA run we standardised the IDs to ease future gene integration. The effect of every iteration in terms of modified or added genes is reported in Table S3.8.

Table S3.7 Cufflinks and Trinity assembly statistics

	Trinity	Trinity Genome-guided	Cufflinks
Number of contigs	313074	88622	110023
Total size of contigs	306002993	88774769	197778585
Longest contig	21967	16737	32882
Shortest contig	201	201	0
Number of contigs > 1K nt	102998	30794	71122
Number of contigs > 10K nt	26	20	171
Mean contig size	977	1002	1798
Median contig size	539	586	1464
N50 contig length	1743	1734	2610
L50 contig count	54913	16498	24863
contig %A	30.13	29.85	29.28
contig %C	19.78	19.94	19.05
contig %G	19.97	20.38	21.16
contig %T	30.13	29.84	30.49

Table S3.8 PASA iterative annotation updates

Dataset	Iteration	Gene updates	Novel splice variants	Novel genes	Number of annotated proteins changed
exAtlas	1	20497	17799	1268	8792
	2	1428	1821	7	1321
exDiversity	1	23956	11017	390	12055
	2	1289	1475	1	1118
Xylem/leaf	1	16705	3817	240	8205
	2	633	317	1	426
Leaf	1	31157	8733	208	13171
	2	50298	1066	2	50172

3.2.3 Manual annotation

We performed an intermediate analysis of the annotation, which revealed that a number of genes from Potri (and from the other aspen annotation) did not overlap any genes predicted by the Maker + PASA annotation pipeline but appeared to have a valid alignment to the genome assembly considered. Hence, Potri genes generating valid alignments using GMAP that did not overlap any existing feature in the PASA annotation were combined with the PASA annotation. We did the same for alignments of the Potrs/Potra initial gene annotation to the other species genome assembly. The

outcome of this resulted in the final annotation. We validated and extended this final set of annotations (e.g. to add intron features) using genomertools. Genes added in this final manual step were flagged as Low Confidence (LC), with all other genes flagged as High Confidence (HC).

3.2.4 Final annotation statistics

The final Potra annotation contained 35,984 genes: 34,184 protein coding; out of which 33,859 encode a protein longer than 29AA -, 517 miRNA, 598 ncRNA and 675 tRNAs. 29,252 and 6,057 genes were flagged as HC and LC, respectively (Table S3.9). The mRNA annotation contained 76,249 full-length (FL), 2,468 5'-partial, 2,468 3'-partial, 1,201 fragments and 1,107 non-coding transcripts.

As we did not have access to RNA-Seq data from Potrs we did not perform the PASA annotation steps. However, the final set of Potra annotated genes (before addition of Potrs genes) were aligned to the Potrs assembly and those genes showing evidence of alignment that were not represented by the MAKER annotation were manually added and flagged as LC. This resulted in the addition of 9,922 genes. The final Potrs annotation contained 36,830 genes (Table S3.9): 35,676 protein coding, 521 miRNA and 633 tRNAs. 26,842 and 8,852 genes were flagged as HC and LC, respectively. The mRNA annotation contained 34,005 full-length (FL), 5,270 5'-partial, 5,235 3'-partial, 2,835 fragments and 975 non-coding transcripts.

From these annotations, we extracted several subsets, from which we created files containing the corresponding mRNA, CDS and protein sequences. These are available from the PopGenIE Data FTP resource, which includes an explanatory Readme document. The filenames are largely self-explanatory, but briefly we assigned confidence as described above, i.e. it is LC (Low Confidence) for genes lifted-over from Potri/Potrs/Potra and for those that did not generate a coding primary mRNA and HC for the rest. Specifically, for the FASTA files, the representative set corresponds to the subset of coding sequences, one per gene, that encodes for the longest protein sequence. The FASTA header further indicates whether the sequence's confidence and the protein coding status (e.g. full-length, 5-prime partial, etc.), if relevant.

In addition to the subsets described above, we extracted protein sequences based on in-frame validity. We further subdivided that set into *fragment* (no start, no stop), *5p-partial* (start only), *3p-partial* (stop only) and complete (start and only one stop at the end).

Table S3.9 Final annotation statistics for the Potra01 and Potrs01 genome assemblies

Annotation feature	Potra01	Potrs01
High/Low confidence gene loci	29,252 / 6,057	26,842 / 8,852

High/Low confidence transcripts	76,557 / 8,312	34,439 / 13,899
High/Low gene loci expressed	27,825 / 4,833	NA / NA
Mean (\pm SD) transcripts per gene	2.4 (\pm 2.2)	1.3 \pm 0.9
Quartiles transcripts per gene (.25, .5, .75 and 1)	1 / 1 / 3 / 22	1 / 1 / 1 / 25
Median (\pm MAD) exon/intron length (bp)	159 (\pm 129) / 178 (\pm 138)	147 (\pm 110) / 171 \pm 130
Median (\pm MAD) gene/CDS length (bp)	2,644 (\pm 2,588) / 1,077 (\pm 699)	2,010 (\pm 2210) / 834 (\pm 720)
Median (\pm MAD) exon number	5 \pm 4	4 \pm 3
Single exon genes	7,294	8,355

3.2.5 Intron/exon length and count comparison

Using the annotation gff3 file, we compared the count and length of introns and exons across the three assemblies (script: SuppIntronExon.R). Overall, Potra had longer and slightly more exons, whereas intron statistics were very similar for all three species (Figure S3.4 - S3.6). The observed differences in exon length and number likely represent more comprehensive annotation of UTRs in Potra due to the extensive RNA-Seq resource utilised. In all three species UTRs remain relatively poorly annotated and, as such, should not be considered as a true representation of the total UTR region.

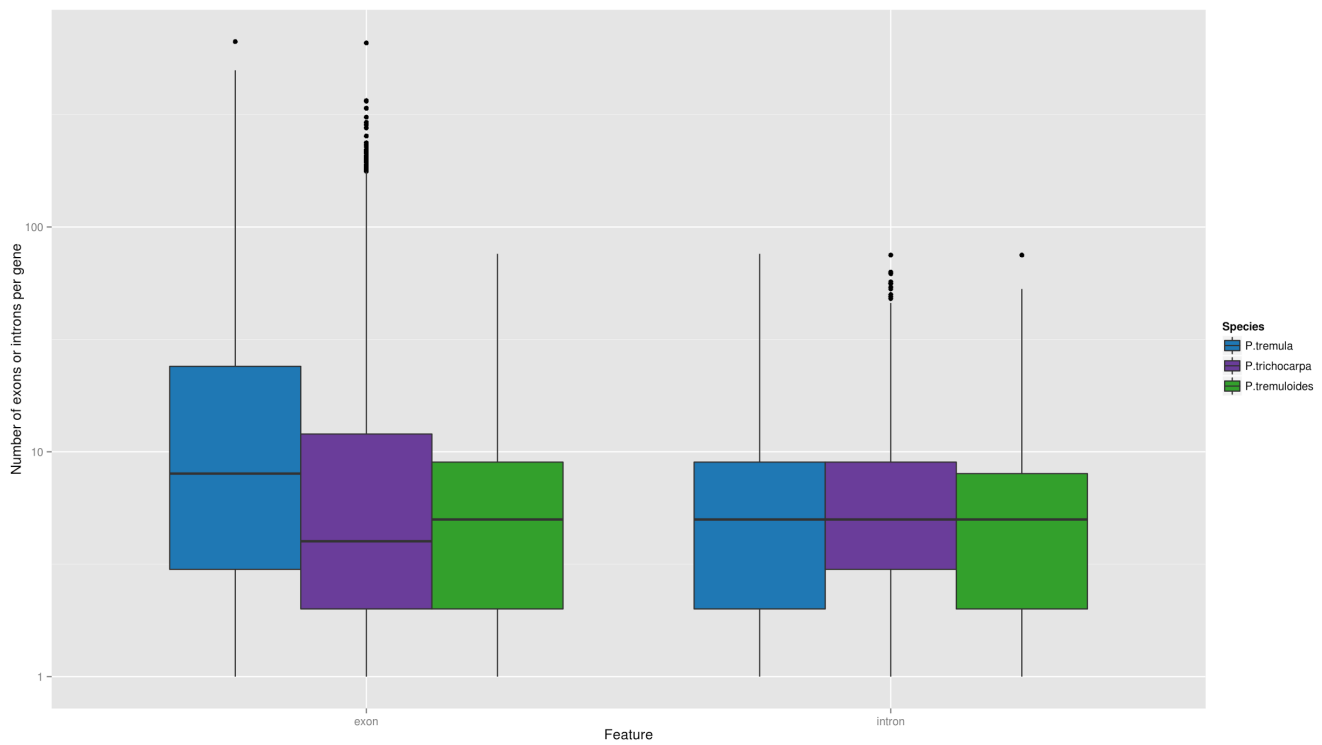


Figure S3.4 Number of introns and exons for the Potra (blue), Potrs (green), and Potri (purple) genome assemblies.

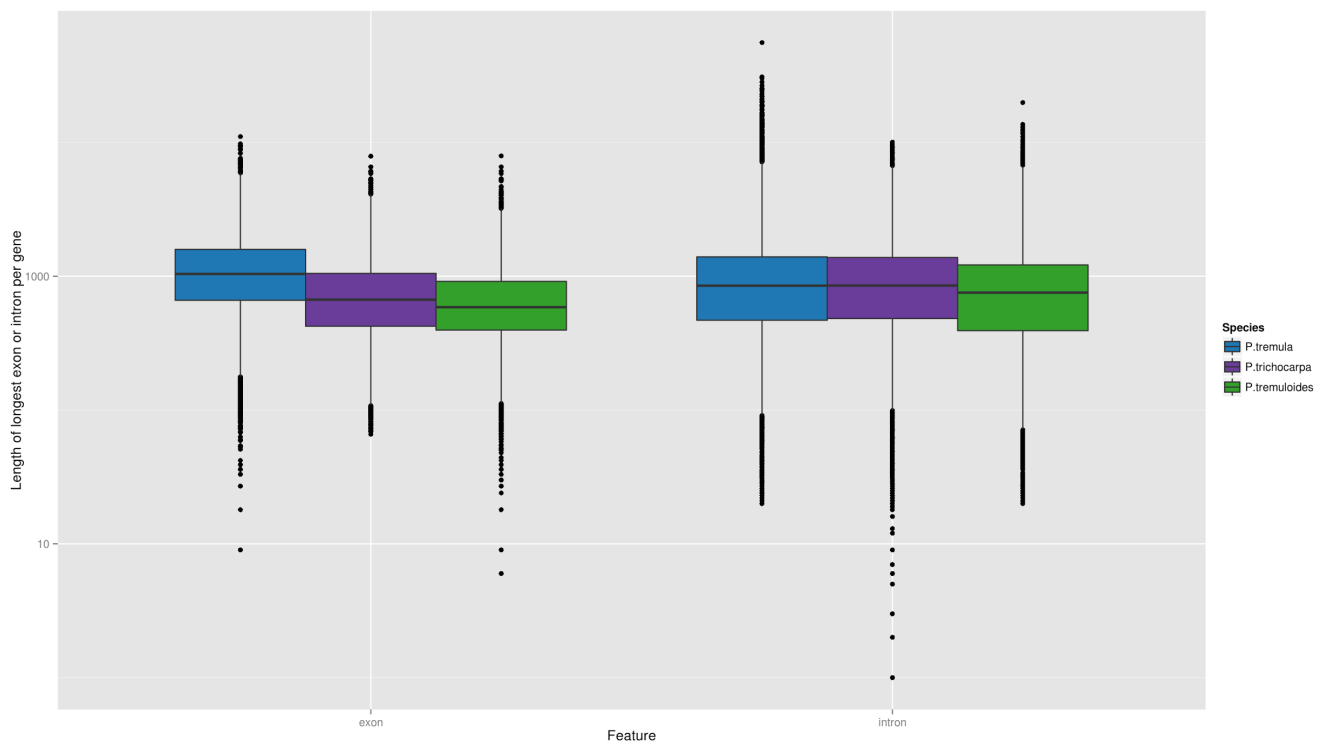


Figure S3.5 Length of longest intron and exon per gene for the Potra (blue), Potrs (green), and Potri (purple) genome assemblies.

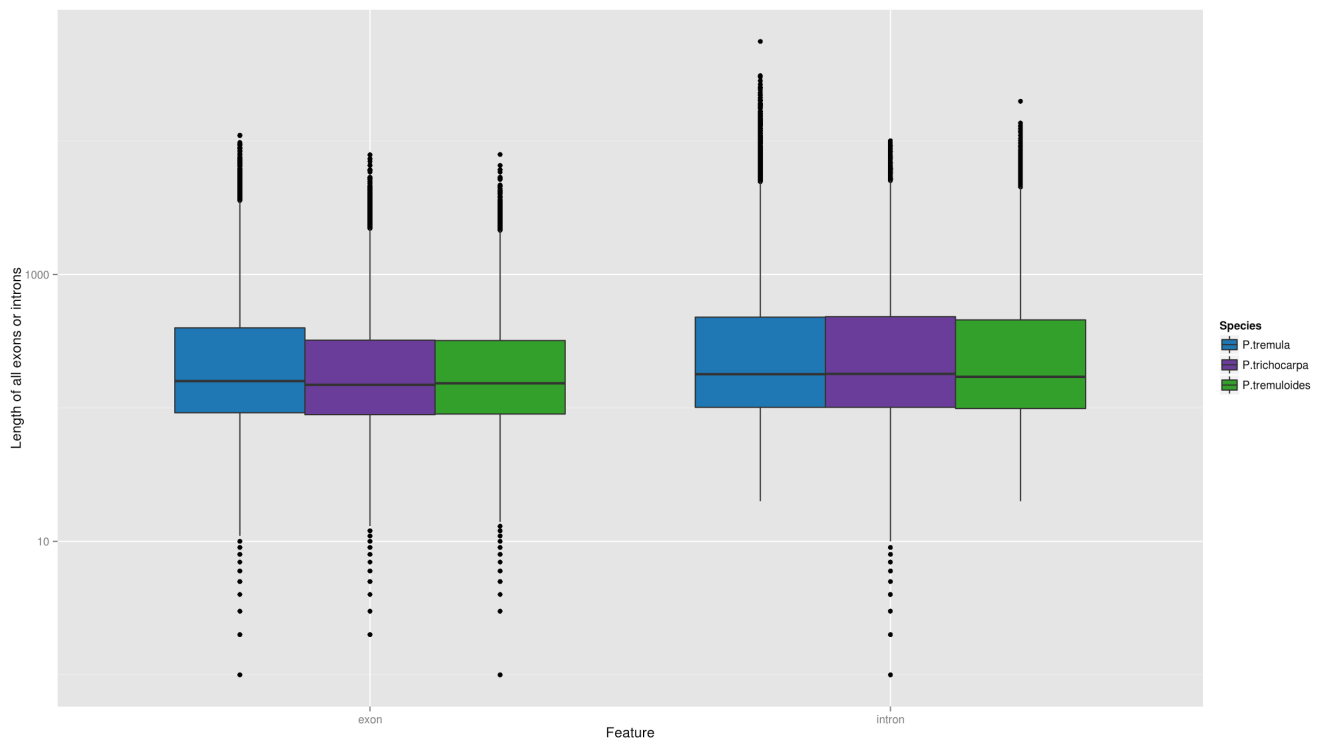


Figure S3.6 Length of all introns and exons for the Potra (blue), Potrs (green), and Potri (purple) genome assemblies.

3.3 Genome and gene annotation completeness assessment

3.3.1 CEGMA, PLAZA and BUSCO

We applied three complementary quality measurement methods to estimate the completeness and quality of genome assembly and gene prediction in Potra and Potrs.

CEGMA

The standard CEGMA pipeline (40) estimates completeness of a genome assembly. In summary, 93.55% of CEGMA genes were completely identified in the Potra genome, while in the Potrs genome the completeness was 91.94%. We further extended the genomic sequence-based pipeline into protein coding genes to examine the completeness of the gene prediction. The 458 hidden Markov model profiles from CEGMA were searched against predicted gene models and all except one of the 458 models were identified in both the Potra and Potrs gene prediction.

BUSCO

BUSCO (41) assesses genome assembly and gene prediction completeness based on the near universal single copy orthologs selected from OrthoDB. Using the pre-released BUSCO v1.0 plant database, more than 95% of BUSCO genes were completely recovered in Potra and Potrs genome assembly and gene prediction. Only 1.7% and 1.6% of BUSCO genes were missing in Potra and Potrs gene models, respectively.

PLAZA

Using the PLAZA 2.5 database, we identified 2,886 gene families that are single-copy in a majority of dicot genomes (with the exception of soybean) available in the PLAZA 2.5 database (PLAZA coreGF). We considered that these highly conserved gene families serve as useful markers to benchmark the completeness of the gene prediction (42). Due to differences in gene prediction quality between genomes and the N-terminal and C-terminal end of protein sequences tending to be less conserved across species, we used only the conserved protein regions of each coreGF for this analysis. In brief, we aligned protein sequences in each coreGF using MUSCLE, removing unconserved regions based on the MUSCLE multiple sequence alignment using trimAl (43). A hidden Markov Model (HMM) of the conserved region was subsequently built using HMMER 3.0. In total, 2,868 and 2,840 complete (>90% coverage) PLAZA coreGF were identified in the Potra and Potrs gene prediction, respectively.

3.3.2 Potri gene space alignment to the Potra01 and Potrs01 assemblies

We used GMAP to align Potri synthetic transcripts (see below) to the Potra genome. The transcript of the analysis can be found in our public Git repository: [3.3.2-Potri-gene-space-alignment-to-the-Potra01-and-Potrs01-assemblies/potriGenePotraGenomeAlignment.html](https://github.com/3.3.2-Potri-gene-space-alignment-to-the-Potra01-and-Potrs01-assemblies/potriGenePotraGenomeAlignment.html). Briefly, there were 80.28 % (33183 out of 41335) Potri genes aligning at a cutoff of 90% coverage and 70% identity; there were 92.7 % (38316 out of 41335) Potri genes aligning at a 80% coverage cut-off and finally there were 95.31 % (39398 out of 41335) Potri genes aligning at a 70% coverage cut-off. In all cases, 99% of the aligned genes had >50% of their length within a single scaffold. Finally, 605 genes showed no evidence of alignment. The rationale of using a 70% identity cutoff is based on the fact that we compare nucleotide sequences and anticipate the triplet 3rd base wobble. The actual sequence identity was higher (80-99%), as can be observed in the aforementioned html page.

Similarly, for alignments to the Potrs genome (Git: [3.3.2-Potri-gene-space-alignment-to-the-Potra01-and-Potrs01-assemblies/potriGenePotrsGenomeAlignment.html](https://github.com/3.3.2-Potri-gene-space-alignment-to-the-Potra01-and-Potrs01-assemblies/potriGenePotrsGenomeAlignment.html)), there were 77.77 % (32146 out of 41335) Potri genes aligning at a cutoff of 90% coverage and 70% identity; there were 91.34 % (37757 out of 41335) Potri genes aligning at a 80% coverage cut-off and finally there were 94.45 % (39042 out of 41335) Potri genes aligning at a 70% coverage cut-off. In all cases, 99% of the genes have >50% of their length within a single scaffold. Finally, 610 genes showed no evidence of alignment.

3.3.3 Potri gene alignment to the Potra01 and Potrs01 gene models

To determine how many gene models in the Potra01 and Potrs01 annotation were well represented in the Potri gene models and to indicate the quality of our gene prediction, we used a rather stringent filtering threshold to determine the representation of complete gene models (BLASTP e-value 1e-5, HSP coverage > 70% of the Potri protein). In total,

we identified 26,319 (of 33,858) Potra and 24,507 (of 35,694) Potrs gene models that were consistent with the Potri models. When we applied the same filtering threshold to self-self Potri alignments, 41,294 (of 41,335) models were selected.

4 Cross-species gene space analysis

4.1 Best BLAST hit

To generate initial cross-references across the three species, we performed pairwise protein sequence homology searches of the representative transcript sequences using BLAST+ (v2.2.29), filtered the results at a e-value threshold of $1e^{-5}$. We combined HSPs per query-subject pair and retained the pair maximising coverage as the cross-reference.

4.2 Potra gene expression

4.2.1 Synthetic transcript set

We constructed a set of non-redundant annotations (44) using the *createSyntheticTranscripts* function of the R/Bioconductor easyRNASeq package (45). Briefly, for every gene, we combined the transcripts per locus to form a single, synthetic transcript representative of all exonic sequences for that locus. We then extracted the corresponding mRNA sequences and used these in summarising gene expression. We also used the synthetic transcript sets to perform comparative cross-species alignments (see below).

4.2.2 Expression specificity

We calculated expression specificity as detailed in Delhomme *et al.*, (2015)(46). Briefly, the expression specificity score ranges from 0 (ubiquitous) to 1 (tissue/sample specific) and is calculated relative to the tissue with the highest mean expression e.g. a score of 0.8 indicates that the mean expression in other tissues is 20% of the expression in the tissue with the highest mean expression. The majority of genes tended to be ubiquitous, whereas a minority appeared to be very tissue/sample specific (Figure S4.1).

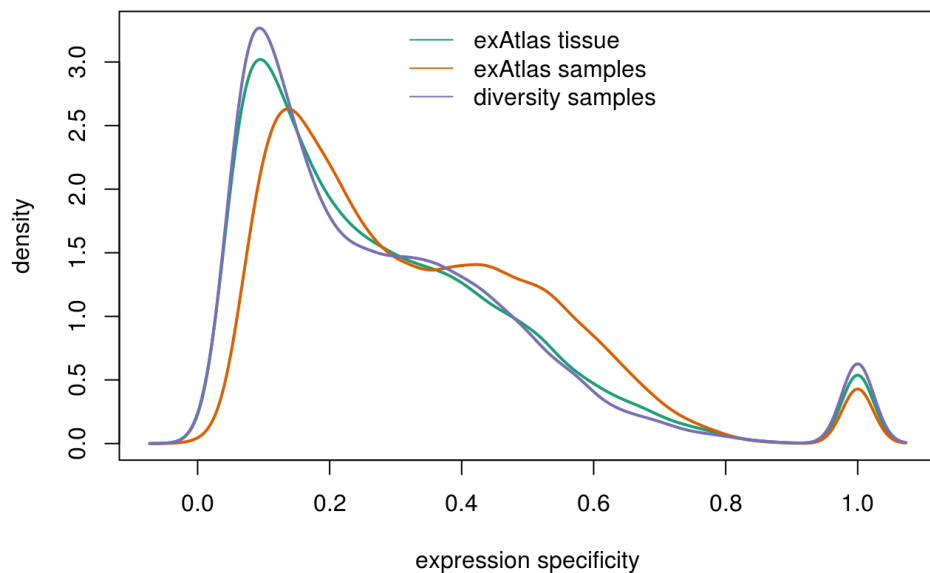


Figure S4.1 Density distribution of calculated expression specificity scores for Potra. The exAtlas dataset samples were grouped either on the basis of the tissue of origin or analysed separately per sample.

4.2.3 Silent genes

From the above analysis, we identified a number of genes (11-15%) that were not expressed in either the exAtlas or exDiversity datasets, with 3,780 genes ‘silent’ across both. We converted the obtained list of genes into a list of Potri gene IDs using the aforementioned cross-references, resulting in 2,128 Potra genes matching 1,799 Potri genes. Next, we used the PopGenIE resource(35) to identify Gene Ontology (GO (47)) category and Pfam (48) domain enrichment (Table S4.1). The GO Cellular Component *photosystem II* was the only enriched GO term while the Pfam domain enrichment revealed many domains of unknown function (DOFs) in addition to a diverse range of other domains. A KEGG (49) enrichment revealed a number of diverse enzymatic activities spanning a large range of cellular functions (including signalling, cell cycle, metabolism), most likely indicating that the datasets considered did not represent conditions or tissues in which many of these genes would be expressed. At the gene family level, these silent genes comprised 113 orphans, 1040 shared and 712 species-specific (see below for gene category information) genes.

Table S4.1 GO, PFAM and KEGG enrichment of silent genes. The enrichment test was performed using PopGenIE.org.

Datasource	Terms	p-valueFDR	Statistics	Description
GO	GO:0009523	2.11E-02	10/255 62/6017	photosystem II
PFAM	PF11820	2.59E-04	8/1227 21/29014	Protein of unknown function (DUF3339)
	PF05938	1.04E-02	7/1227 27/29014	Plant self-incompatibility protein S1
	PF00164	2.27E-02	4/1227 10/29014	Ribosomal protein S12

	PF02536	2.40E-02	9/1227 55/29014	mTERF
	PF02362	2.69E-02	14/1227 124/29014	B3 DNA binding domain
	PF00240	2.82E-02	13/1227 103/29014	Ubiquitin family
	PF00560	3.09E-02	13/1227 691/29014	Leucine Rich Repeat
	PF03087	4.44E-02	6/1227 33/29014	Arabidopsis protein of unknown function
	PF11919	4.65E-02	2/1227 2/29014	Domain of unknown function (DUF3437)
	PF04185	4.83E-02	3/1227 7/29014	Phosphoesterase family
	PF01466	5.24E-02	4/1227 14/29014	Skp1 family, dimerisation domain
	PF06916	6.38E-02	2/1227 3/29014	Protein of unknown function (DUF1279)
	PF11210	6.77E-02	2/1227 3/29014	Protein of unknown function (DUF2996)
	PF03468	7.00E-02	4/1227 18/29014	XS domain
	PF00033	7.23E-02	2/1227 3/29014	Cytochrome b(N-terminal)/b6/petB
	PF03650	7.74E-02	2/1227 3/29014	Uncharacterised protein family (UPF0041)
	PF06596	8.34E-02	2/1227 3/29014	Photosystem II reaction centre X protein (PsbX)
	PF00999	8.82E-02	6/1227 42/29014	Sodium/hydrogen exchanger family
	PF04054	9.03E-02	2/1227 3/29014	CCR4-Not complex component, Not1
	PF01158	9.16E-02	2/1227 4/29014	Ribosomal protein L36e
	PF00288	9.32E-02	4/1227 20/29014	GHMP kinases N terminal domain
	PF01406	9.58E-02	2/1227 4/29014	tRNA synthetases class I (C) catalytic domain
KEGG	K01883	8.45E-03	2/230 2/5582	cysteinyl-tRNA synthetase [EC:6.1.1.16]
	K01114	8.55E-03	3/230 7/5582	phospholipase C [EC:3.1.4.3]
	K06699	1.13E-02	2/230 2/5582	proteasome activator subunit 4
	K03364	1.69E-02	2/230 2/5582	cell division cycle 20-like protein 1, cofactor of APC complex
	K03094	2.40E-02	2/230 4/5582	S-phase kinase-associated protein 1

K00224	2.74E-02	2/230 4/5582	delta24(24(1))-sterol reductase [EC:1.3.1.71]
K02920	3.20E-02	2/230 4/5582	large subunit ribosomal protein L36e
K00872	3.38E-02	2/230 2/5582	homoserine kinase [EC:2.7.1.39]
K01183	6.19E-02	2/230 7/5582	chitinase [EC:3.2.1.14]
K04382	6.88E-02	2/230 7/5582	protein phosphatase 2 (formerly 2A), catalytic subunit [EC:3.1.3.16]
K13993	8.38E-02	2/230 9/5582	HSP20 family protein
K01376	9.14E-02	2/230 9/5582	cathepsin V [EC:3.4.22.43]

4.2.4 Cross-species comparisons

We observed that the majority of genes that did not align across species were short, non-protein-coding sequences. These ranged in number between 440 and 605, depending on the pairwise comparison considered. There were 136 Potra, 536 Potri and 146 Potrs genes with no identified ortholog in another species and that are therefore putatively species-specific. There were 1,127 genes that were unique to the two aspens, showing no alignment to Potri. Transcripts of these analyses are available from the Git repository to allow further exploration of the results. Out of these, 164 Potra and 128 Potrs genes were protein coding and encoded full length proteins. To assert whether these were truly species-specific we performed sequence homology searches of nucleotide or protein sequence (see section 3.2.4) against the NCBI nt (retrieved 14/01/2017) and UniRef90 (retrieved 01/2017) databases (Table S4.2). These results, together with the analysis transcript available in Git as 4.2.4-Cross-species-comparisons/aspenSpecificSequencesBlast.R, show that the majority of these genes are aspen-specific. An additional fraction had good protein but poor mRNA coverage, possibly indicative of very long unannotated UTR sequences.

Table S4.2 Aspen-specific genes without homology to nucleotide and protein databases for different cumulative coverage cutoffs. The aspen specific genes not showing alignments to Potri were aligned using BLAST+ 2.2.29 (non-default parameter e-value $1e-5$), to the NCBI nt database and their corresponding proteins against the Uniref90 database. The results are presented for reciprocal or query-only cumulative coverage cutoffs of 30% and 70%. Shown are the number of genes/proteins having no homology with a sequence in the database longer than the cutoff (NCBI nt and UniRef90 columns) or in either .

Database	Cutoff %	NCBI nt	UniRef90	Union
Potra	30	127	85	77

	30 reciprocal	127	85	77
	70	138	118	112
	70 reciprocal	145	119	115
Potrs	30	102	79	71
	30 reciprocal	102	79	71
	70	114	96	92
	70 reciprocal	117	97	95

5 Comparative analyses

5.1 Gene family comparison

To determine the orthologous relationship between three *Populus* species, we included *Arabidopsis thaliana* (Artha) as an outgroup species in the gene family comparison. In brief, we collated protein sequences from Artha (TAIR10 (50)), Potri (V3.0 (51)) and the two aspen species presented here and performed an all-against-all BLASTP sequence similarity search (Decypher TERA-BLASTP, e-value 1^{E-5} , max hits 500 run on a TimeLogic Decypher machine, Active Motif Inc., Carlsbad, CA). Based on the BLASTP result, we performed two rounds of clustering using TribeMCL (52) (MCL v10-201, inflation value 3.0) to delineate gene families. The first TribeMCL analysis generated the ‘Family’ level of orthologous information with the prefix of ‘F’ labelled in the gene family analysis. In order to obtain a deeper resolution of the orthologous information, we performed a second round of TribeMCL clustering based on the result at the ‘Family’ level. We performed a new all-against-all BLASTP search (Decypher TERA-BLASTP, e-value 1^{E-5} , max hit 500) of protein sequences within the same ‘Family’ and subjected the BLASTP result to the second round of graphic based clustering (TribeMCL, inflation value 5.0). The second round of the TribeMCL generated the ‘Group’ level of orthologous information with the prefix of ‘G’ labelled in the gene family analysis. We aggregated orphan genes from the second round of the TribeMCL clustering as the last ‘Group’ in each ‘Family’. In general, the two rounds of clustering were able to better identify the true orthologous genes in most of the large gene families.

Overall, >90% of genes in four species were assigned to ~22,000 gene families (Table S5.1 and Figure S5.1). On average, the Potri genome had more gene copies in each gene family (Figure S5.2). To further investigate the origin of the higher copy number genes in the Potri genome, we first updated the whole genome duplication (WGD (5)) on the Potri v3.0 genome (Phytozome 10). We used the same analysis method as the original genome paper (5) with the latest version of i-ADHoRe (53) (v3.0, gap_size = 30, cluster_gap = 35, tandem_gap = 10, q_value = 0.85, prob_cutoff = 0.001,

anchor_points = 5, alignment_method = gg4, level_2_only = true and multiple_hypothesis_correction = FDR). The updated Potri WGD analysis result is available from the PopGenIE.org Lin2018 FTP resource.

We selected the first 100 shared gene families (gene family containing all four species) with the largest gene family size variation (standard deviation > 14). In total, 2,627 Potri genes belonged to this category with 2,006, 1,667 and 1,634 genes in Artha, Potra and Potrs, respectively. A large fraction of the Potri genes belonged to tandem duplicated gene clusters (1,425 genes, p-value 3.2e-261) or anchor point genes derived from the recent WGD event (865 genes, p-value 1.5e-64). This result indicates that the relatively higher copy number of Potri genes compared to Potra and Potrs was strongly associated with genome assembly artefacts in the Potra and Potrs genomes. Due to the high sequence identity within the same tandem duplicate cluster, we observed numerous examples where almost identical tandem duplicate gene clusters were collapsed into few consensus copies during genome assembly. This is common to all such assemblies derived from short read sequencing technologies, and means that gene family analyses we report here must be considered preliminary until a mature, contiguous assembly becomes available.

Table S5.1 Summary of gene family comparison.

Type	Artha	Potra	Potri	Potrs
No. genes included in analysis	27,407	33,565	41,335	35,444
No. gene families (genes) (include species specific)	14,315 (24,931)	21,659 (32,131)	22,382 (38,422)	21,718 (31,991)
Share gene families (genes) (in more than 1 species)	13,321 (20,965)	21,577 (31,951)	21,951 (37,156)	21,526 (31,568)
Species specific gene families (genes)	994 (3,966)	82 (180)	431(1,266)	192 (423)
Orphans	2,475	1,434	2,913	3,453

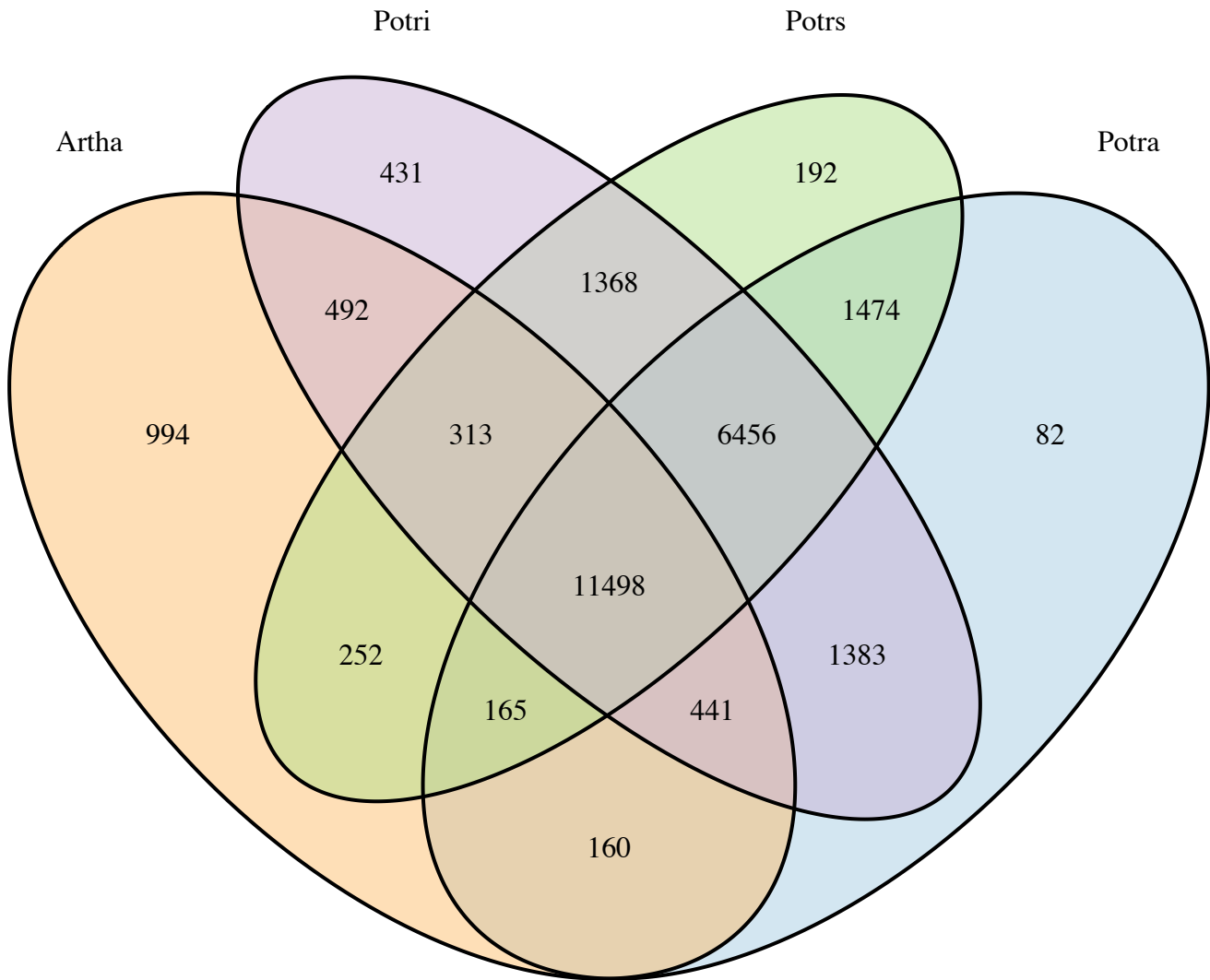


Figure S5.1 Overview of gene family comparison. The shared and unique number of gene families between four species.

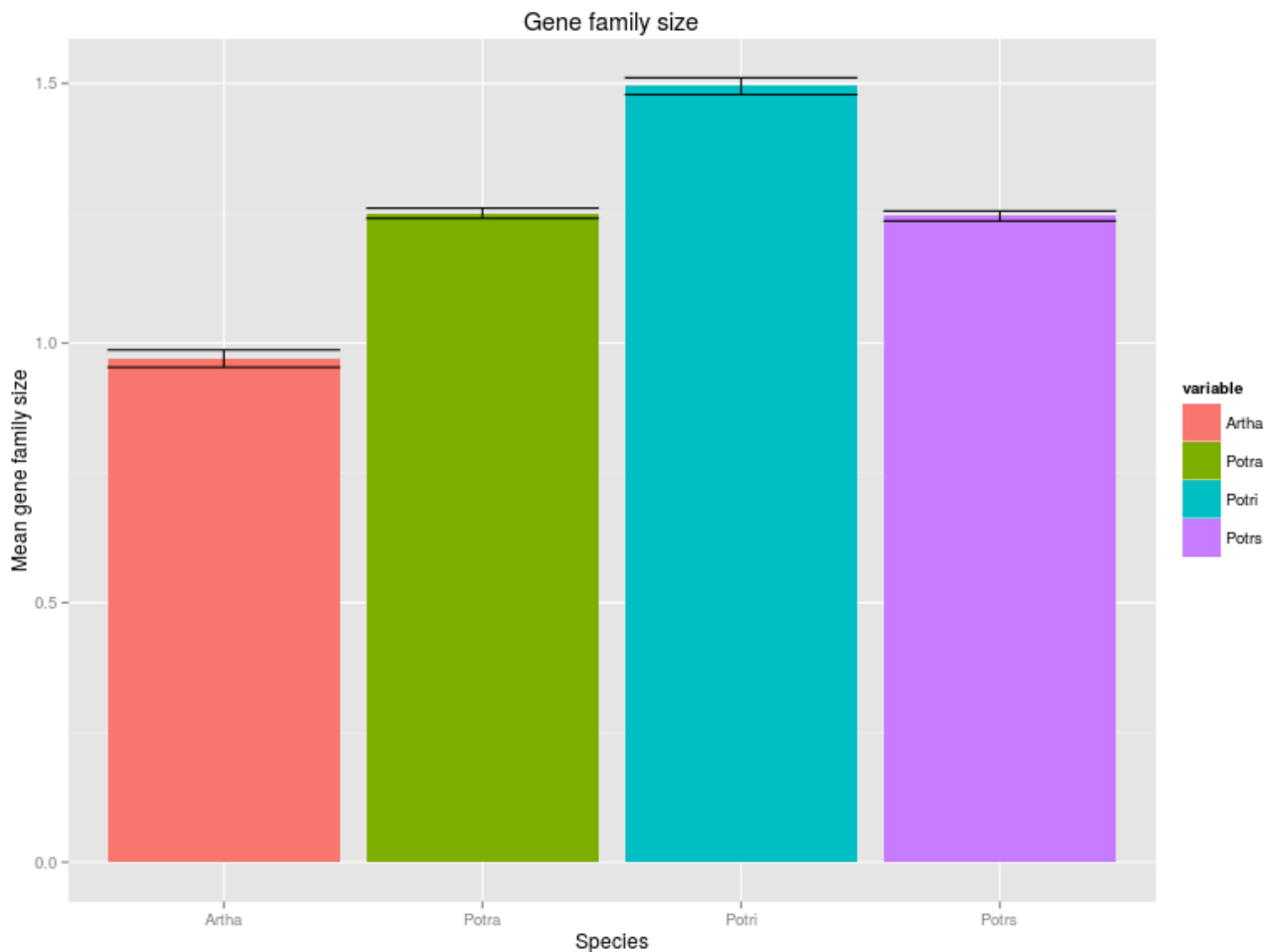


Figure S5.2 Gene family size distribution. The average gene family size of the shared gene families.

5.2 Estimation of Ks age distributions

We used the results of the TribeMCL orthologous assignment for the three *Populus* species and Artha (outgroup) to identify orthologous gene families among the species. We used only full-length protein sequences that were in frame (no internal codons). We did not consider gene families containing only one gene (orphans) or that were species specific with fewer than three genes per family. A total of 188,968 protein sequences were analysed.

We performed multiple sequence alignments for each gene family using MUSCLE (25). Following the completion of the alignment procedure, we used PAL2NAL (54) to remove alignment gaps, back translate the protein alignments to codon alignments, and format the files for PAML analyses. We eliminated sequences containing mismatches between the coding and the protein sequences.

We calculated the synonymous (Ks) and non-synonymous (Ka) number of nucleotide substitutions per site, and their rate ratio (Ka/Ks) using the maximum likelihood method of Golman & Yang (1994)(55) implemented in the Codeml

package in PAML (56) (v4.6). Codeml control files were constructed for each gene family using an in-house python script. Based on the TribeMCL result, we analysed 12,954 gene families containing at least one gene in three of the four species studied using pairwise estimations (runmode=-2, seqtype=1, model=0, NSsites=0). Codeml results were parsed and results were divided between orthologous comparisons (for all combinations of species) and paralogous comparisons (within species).

To estimate Ks age distributions, we analysed all orthologous and paralogous pairwise estimations with a Ks <5. We corrected for the redundancy of Ks values by removing duplicated pairwise comparisons. A total of 76,260 pairwise comparisons were used to plot the Ks age distribution graphs. Our results indicated the presence of three Ks peaks, the first one between *Arabidopsis* and *Populus*; the second one among *Populus* orthologs; and the third one among *Populus* paralogs (Supplementary Figures S5.3 and S5.4). These results suggest the presence of a WGD event common to *Populus* and *Arabidopsis* species, and another one specific to *Populus* species; confirming previous results comparing the genomes of Potri and Artha (5). We did not detect evidence for the presence of a more ancient WGD event, as suggested in previous studies (5, 57).

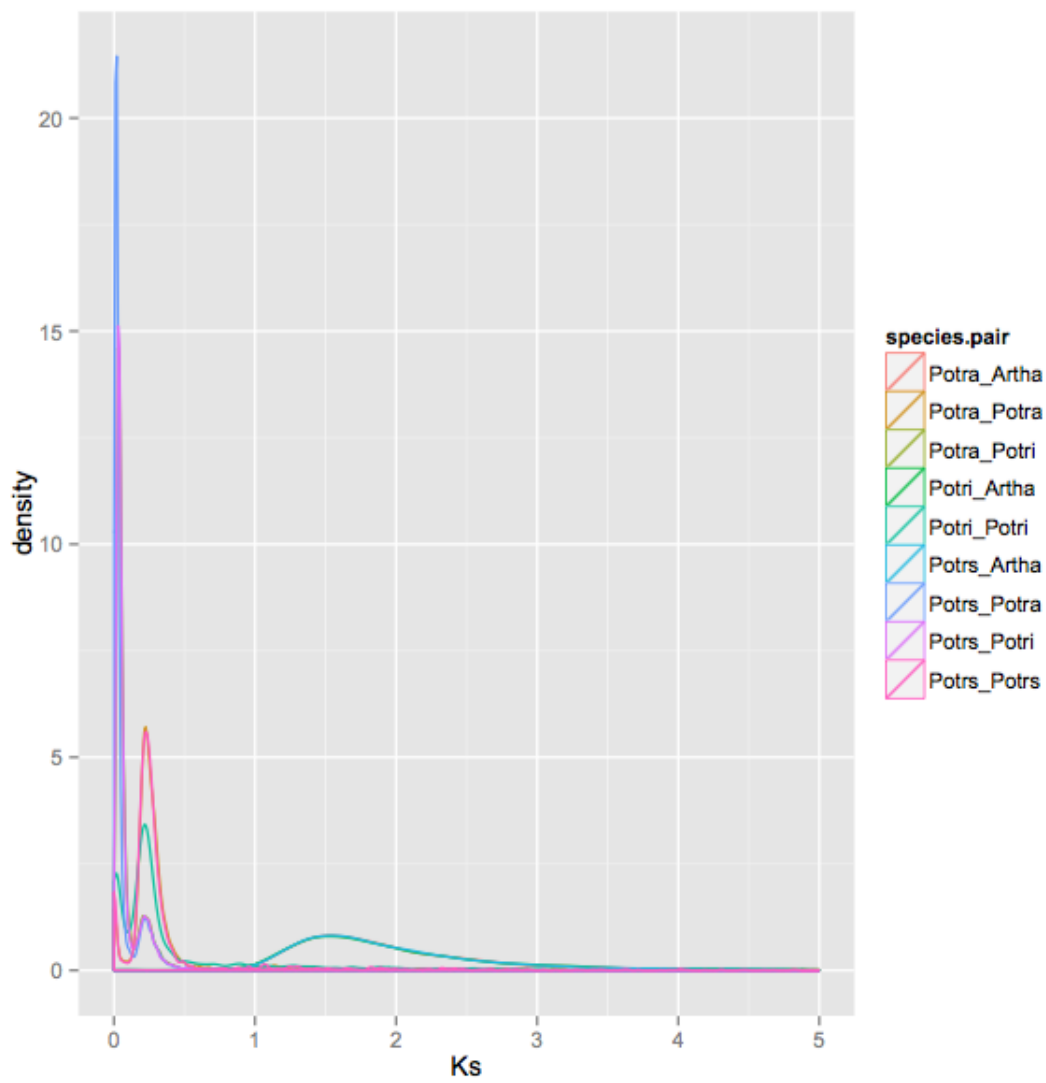


Figure S5.3. Ks age distributions for all species and within species pair combinations among Potri, Potra, Potrs and Artha.

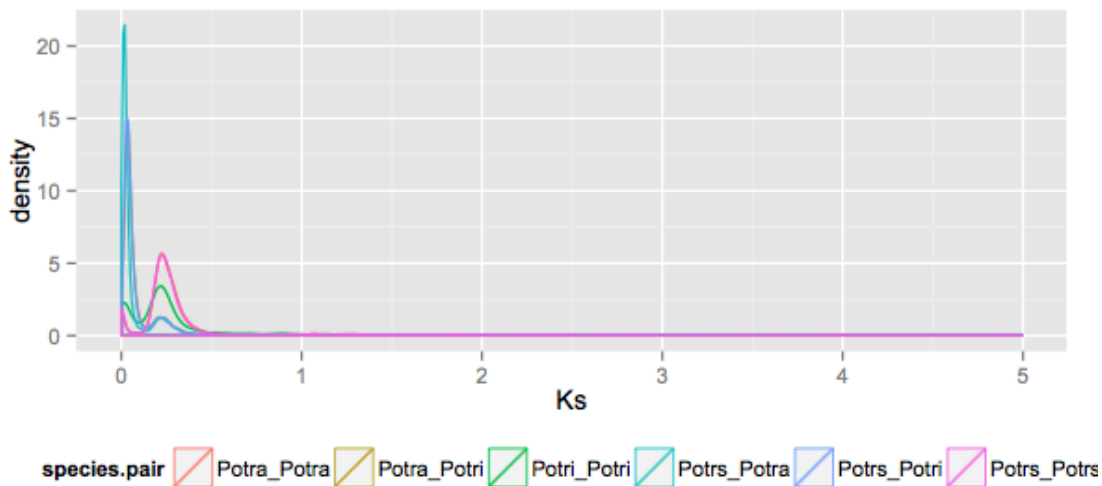
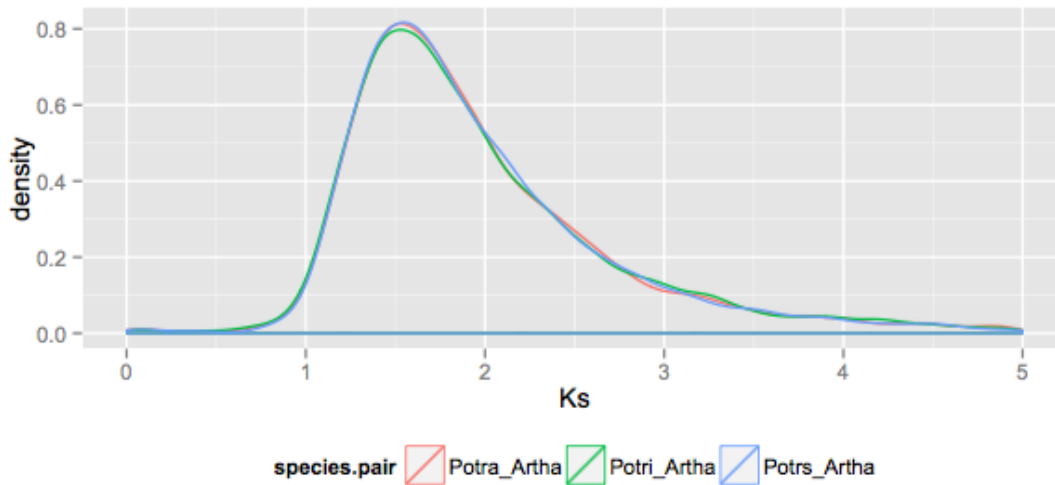


Figure S5.4 Ks age distributions for all species pair combinations among Potri, Potra, Potrs and Artha (top graph) and among *Populus* species (bottom graph).

5.3 Genes under diversifying selection

From the Codeml results, we identified the orthologous genes that had a Ka/Ks rate ratio of >1 , which are potentially under diversifying selection. We discarded genes with abnormally high Ka/Ks ratios ($Ka/Ks > 10$). After the removal of duplicated pairwise comparisons, we kept 2246 orthologous genes showing evidence of diversifying selection. Detailed results can be found in Dataset S1, worksheet 1. We then tested for enrichment of GO categories in the genes under selection using the BINGO plugin (v2.44)(58) for Cytoscape (59) using all annotated genes as the background. For Potri *P. trichocarpa* gene models were used, and for any analysis involving only aspen, *P. tremula* gene models were used. Corrections for multiple testing were performed using the Benjamini & Hochberg method (60) with a false discovery rate threshold of 0.05. Enriched categories included: regulation of transcription, regulation of gene expression, regulation of

biosynthetic process, regulation of metabolic process, among others in both Potra (Dataset S1 1, worksheet 2) and Potri (Dataset S1, worksheet 3).

We examined gene expression of 578 genes that have diverged between the two aspens and 282 genes that have diverged from Potri in both aspens. Of the genes that have diverged between the two aspens, 394 have variable expression in the Swedish Aspen (SwAsp) Collection (61). In the co-expression network, these genes have slightly lower connectivity both from a co-expression module perspective (k_{diff_norm} ; Mann-Whitney test, $p = 0.001$) as well as globally (k_{Total} in the plot below; Mann-Whitney test, $p = 0.01$, where k_{Total} is the total connectivity for a gene within the network and k_{diff_norm} is the difference between k_{Within} , which is connectivity of a gene within its assigned module, and k_{Out} , which is the difference between k_{Total} and k_{Within} , scaled for module size.). The same is also true for the genes that have diverged from Potri (174/282 have variable expression), but with a slightly more pronounced effect (Mann-Whitney k_{diff_norm} $p = 0.0006$ and k_{Total} $p = 0.03$).

Of the genes that have diverged between the aspens, 120 were also eGenes, that is, their expression is significantly associated with genetic variation (61). The absolute effect size for the associations involving these genes was significantly greater than for genes that have not diverged between the aspens (Mann-Whitney $p < 2.2e-16$). Of the genes that have diverged from Potri, 64 were eGenes and these showed the same trend with diverged genes having greater effect sizes than non-diverged eGenes (Mann-Whitney $p < 2.2e-16$).

We additionally examined the expression of genes diverged between the aspens and the from Potri using the aspen expression atlas (exAtlas), which comprises 24 tissues/conditions (35). Using the Tau score to indicate tissue specificity (62) showed that the genes that have diverged between the aspens do not have significantly different tissue specificity (571 genes; Mann-Whitney $p = 0.15$) while the genes that have diverged from Potri tend to be more tissue specific (269 genes; Mann-Whitney $p = 0.0003$).

5.4 Saguaro based genome comparisons

We ran the unsupervised genome-wide population analysis Saguaro (63) to segment the genomes into different local topologies based on reads from all populations (Table S7.1) mapped to Potri. Saguaro identified three distinct clades of cacti (i.e. regions exhibiting distinct local topologies) following species taxonomy, with two clades grouping individuals at equal distances (branch lengths), but one group indicating higher divergence in the Potri populations (Figure S5.5), covering about 0.093% of *the* Potri genome (0.54% when merging intervals closer than 10,000). When overlapping these regions with annotated protein coding genes, we found that out of the 12 longest regions in terms of SNP positions,

seven were directly involved in disease resistance (Table S5.2), potentially indicating distinct evolutionary pressure on regions responsible for immune response.

Table S5.2 The top 12 annotated genes overlapping Saguaro regions showing higher divergence between the aspens and Potri, sorted by the number of SNPs. Listed are the genomic regions in Potri, the genomic size of the regions in nucleotides, the number of SNP positions within the region, and the gene annotation in Potri.

Genomic location	Size (bp)	# of SNP positions	Gene annotation
Chr02: 14930867 .. 1494322	12357	136	S-adenosylmethionine synthetase, putative, expressed
Chr12: 2435272 .. 2437617	2345	132	Phytosulfokine receptor precursor, putative, expressed
Chr02: 20578451 .. 20579085	634	114	NBS-LRR type disease resistance protein, putative, expressed
Chr17: 15330162 .. 15330956	794	94	Disease resistance protein RPM1, putative, expressed
Chr01: 5086659 .. 5089719	3060	92	RNI-like superfamily protein/phytosulfokine receptor precursor, putative, expressed
Chr06: 27182372 .. 27182853	481	86	Disease resistance protein (TIR-NBS-LRR class) family
Chr19: 2695664 .. 2695961	297	84	Disease resistance protein RPS2, putative, expressed
Chr01: 4985561 .. 4988837	3276	84	Receptor like protein 42/Cf2/Cf5 disease resistance protein, putative, expressed
Chr05: 737262 .. 737518	256	78	NB-ARC domain-containing disease resistance protein
Chr17: 11270764 .. 11274825	4061	77	Receptor-like protein kinase At3g46290 precursor
Chr04: 10394169 .. 10394886	717	73	hAT dimerisation domain-containing protein
Chr18: 15610274 .. 15616543	6269	69	Receptor-like protein kinase 2, putative, expressed/disease resistance protein RGA2

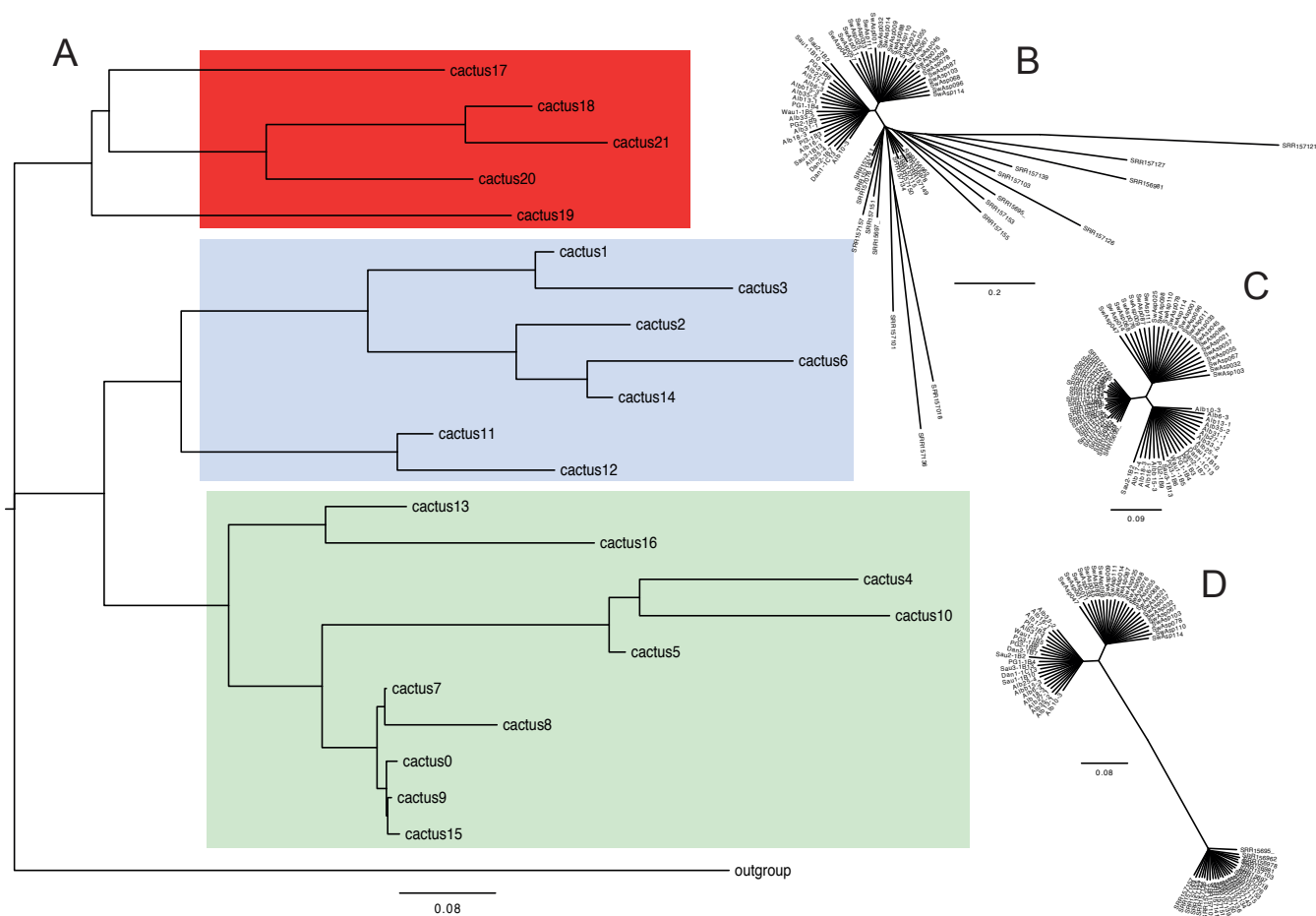


Figure S5.5 Topologies found by Saguaro grouped by similarity (A), and one representative topology per clade (B-D).

5.4 Divergence between Potri, Potra and Potrs

We aligned the aspen and *P. trichocarpa* genomes using Satsuma (64), and divided genomic regions according to the annotations into: exons; coding sequences (CDS); introns; 5' UTRs; 3' UTRs; 1000 bp upstream of the 5' UTR; 1000 bp downstream of the 3' UTR; and annotation-free (Neutral/intergenic). Overlaps between annotations (e.g. CDS and exon) were permissible, where differences were counted in two or more categories. INDELs were counted as the number of nucleotides present in one genome and absent from the other. Statistics for aspens versus *P. trichocarpa* are shown in Table S5.2, the differences between the aspens are listed in Table S5.3.

Table S5.2 Divergence between the aspen and Potri genomes.

Potrs sequence (Mbp)	Potrs SNP rate	Potrs INDEL rate	Potra sequence (Mbp)	Potra SNP rate	Potra INDEL rate
----------------------------	-------------------	---------------------	----------------------------	-------------------	---------------------

Neutral	68.1	0.104985	0.045370	63.9	0.111765	0.0442123
Exon	53.8	0.0264138	0.014841	75.0	0.0287441	0.0185746
CDS	38.6	0.0237439	0.010277	41.6	0.0242717	0.0108302
Intron	54.0	0.0312389	0.023925	60.8	0.0329523	0.0255413
5' UTR	4.4	0.0296111	0.023507	14.8	0.0317256	0.0253242
3' UTR	7.7	0.0317898	0.023966	24.5	0.0331336	0.0258717
Reg. up	83.6	0.0712776	0.060478	83.9	0.0717758	0.0555655
Reg. down	72.2	0.0648981	0.054503	60.8	0.0659404	0.050745

Reg. up/down indicated +/- 1000 bp from the start or stop codon of a gene CDS, respectively.

Table S5.3 Divergence between the Potra and Potrs genomes

	Sequence (Mbp)	SNP rate	INDEL rate
Non-coding	37.5	0.0564664	0.0305452
Exon	76.1	0.0178307	0.0129017
CDS	41.3	0.0161553	0.00905613
Intron	63.3	0.021562	0.0187214
5' UTR	15.2	0.0182138	0.0157666
3' UTR	25.5	0.0194221	0.0162154
5' Upstream	124.2	0.0480699	0.0362267
3' Downstream	63.3	0.0450085	0.0351881

5.5 Whole Genome Duplication and cross-species synteny

5.5.1 Whole Genome Duplication

We first ordered and oriented the scaffolds from both aspen genomes according to synteny using Chromosomeble from the Satsuma package (64), mapping 97.7% and 96.5% of sequence of the Potrs and Potra genomes respectively. We later used this pseudo-chromosome alignment to perform the population genetics and structural variant analyses that we present below. We then generated syntenic self-alignments using Satsuma, and visualised these using ChromosomePaint (Figure 2A; Figure S5.6), a tool based on the Whiteboard class (65).

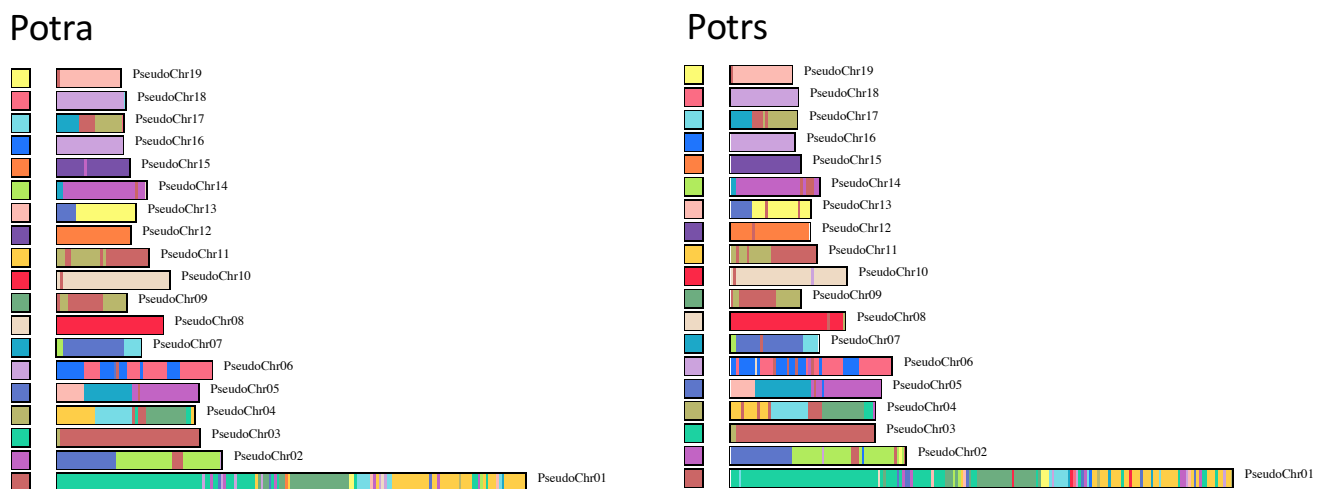


Figure S5.6 Self-alignments of nucleotide sequences for the Potra and Potrs genomes. Synteny matches following a Whole Genome Duplication event are shown by colour.

5.5.2 Example of local conserved synteny and paralogy

In order to identify the degree of local conserved synteny between the three species, we analysed some of the Potra scaffolds; we used reciprocal blast to identify orthologs and paralogs in all three genomes. We show an example of a Potra scaffold syntenic to a region of Potri chromosome 6, and to several scaffolds in Potrs (Figure S5.7). For some of the genes we could also identify retained duplicates from the whole genome duplication; these were located on a single scaffold in Potra, chromosome 16 in Potri and three scaffolds in Potrs.

Of note, the loss and retention of genes since the WGD appears to have occurred before the split of the three species. We did not observe any sign of species-specific losses in this region. This example demonstrates that it is possible to utilise the Potra genome assembly for gene family-oriented evolutionary analyses, both phylogenetic and syntenic, despite the limitations presented by the fragmentary nature of the current assembly version.

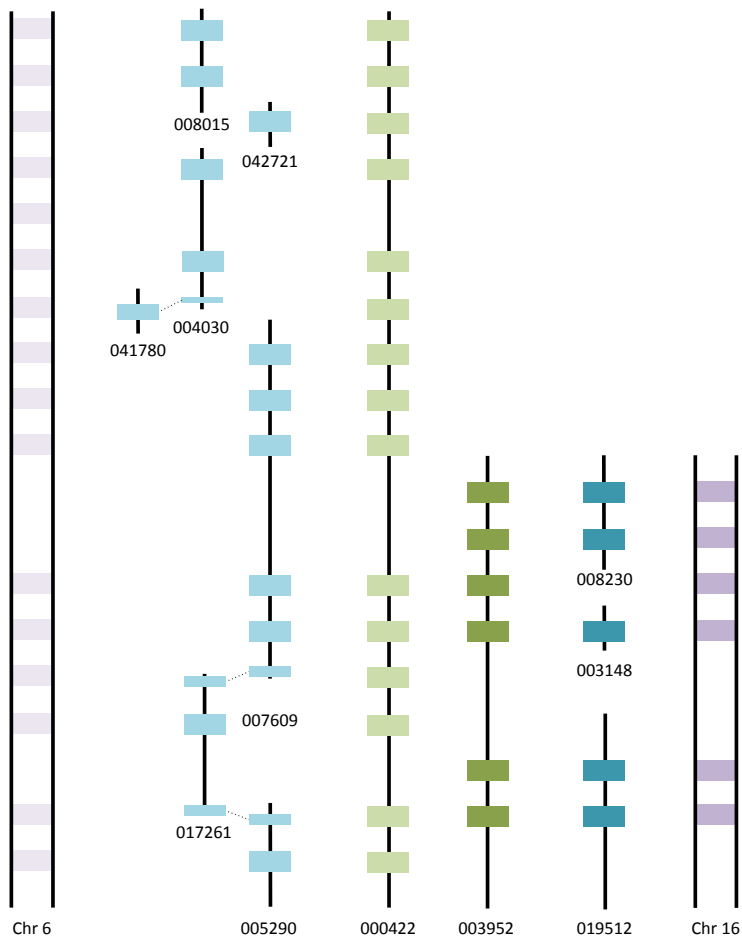


Figure S5.7 Schematic representation of scaffold Potra00422, with genes as light green boxes, and orthologs in Potrs, light blue, and Potri, light purple. Paralogs are coloured in darker shades of each colour (to the right side in the picture). Dotted lines between some of the genes in Potrs indicate that the gene is split between different scaffolds.

5.6 Promoter comparison

To obtain regulatory regions (up- and down-stream UTRs and promoters), we identified the first and last codon of each CDS and excluded those genes not associated with a valid start (ATG) or stop codon (TAG/TAA/TGA). UTRs were defined as the regions between the start/stop codon and the transcriptional start/stop site while promoters were defined as the regions up-/down-stream of the transcriptional start/stop site, extended until we either reached the edge of the scaffold or another gene.

Next, we used the extracted regulatory regions to compare up- and down-stream regions to Potri and between the two aspen assemblies. We reported the percentage of the orthologous region that could be aligned (BLAST E-value $< 1^{E-10}$, multiple hits were merged), with orthologous relationships defined using the above gene family analysis. If ortholog groups contained several genes from the same species, we only reported the most similar pair. There was high sequence similarity between *Populus* regulatory regions, even when compared to exon similarity, but similarity decreased as the regions were increasingly extended away from a gene (Figure S5.8). The aspens were clearly more similar to each other than to Potri. For comparison, virtually no regulatory regions could be aligned between Potri and Artha.

Genes with highly conserved regulatory regions (>90% alignable 500bp regions) were enriched for the GO category DNA binding ($P < 8^{E-5}$; Enrichment tool at PopGenIE.org). This result held when comparing Potri to both aspens using both up- and down-stream regions.

We found that the regulatory regions of paralogs derived from the whole genome duplication had diverged considerably (Figure S5.9).

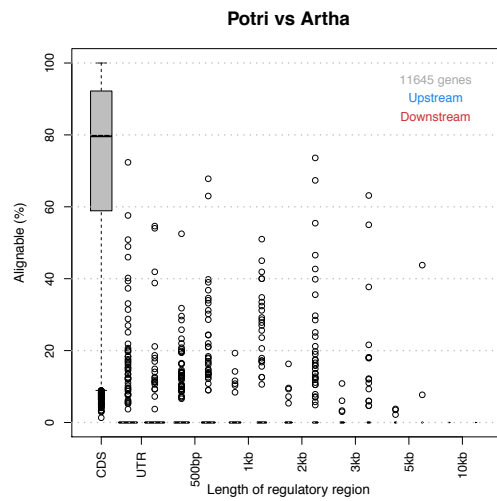
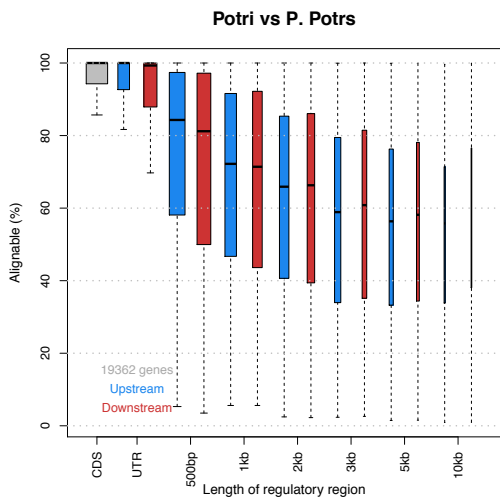
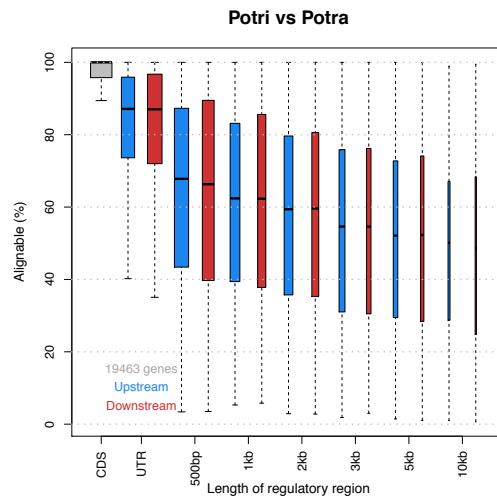
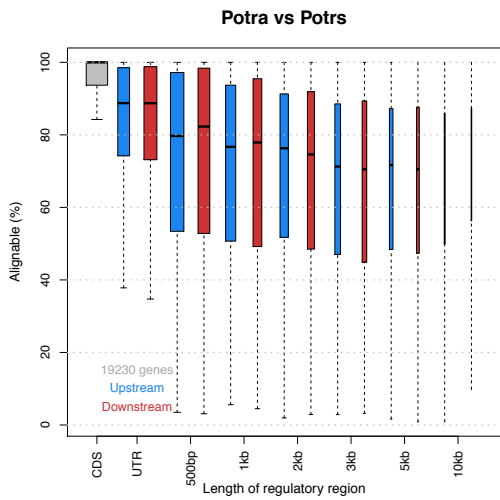


Figure S5.8 The percent of regulatory regions that could be aligned for different lengths. Corresponding numbers for coding sequence (exons) are added for comparison and the number of compared genes/ortholog groups are given in the legend (grey). The width of the boxes is proportional to the number of ortholog groups with at least one ortholog-pair with regions of the required length.

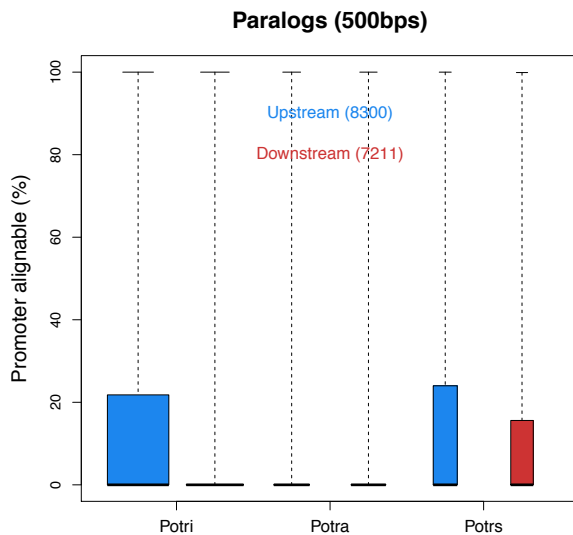


Figure S5.9 The percent of 500 bp paralogous regulatory regions that could be aligned. Only the most similar paralog-pair was reported from each ortholog group. Box widths are proportional to the number of compared paralogs. Potri: 8300/7211 upstream/downstream regulatory regions (given in the legend), Potra: 2799/2655 and Potrs: 1269/1117.

6 The sex determination region

Aspen species have, just like other members in the *Populus* genus, a genetic sex determination system (66), however in aspen this region has a centromeric location in contrast to the peritelomeric region identified in Potri (66). The *TOZ19* (*TORMOZEMBRYO DEFECTIVE*) gene was previously identified as a candidate within the aspen sex-determination region, and was shown to be male-specific in Potrs (67), with only the last exons present in Potra (36). We identified the remnant of *TOZ19* in the Potra genome and localized the orthologs to the neighbouring genes in Potri (see Figure S6.1) using both the synteny based nucleotide aligner Satsuma and BLAST. We were not able to determine the true orthologs in Potra and Potrs for several of the genes in the vicinity of *TOZ19* in Potri. The presence of gaps in the Satsuma alignment together with the fact that the identified orthologs were often located on single gene scaffolds indicates a low level of conserved synteny in this region. In order to compare with a region not suggested to be involved in sex determination, we repeated the analyses for the region surrounding the *TOZ19* paralog, *TOZ13*. In this region (Potri chromosome 13) the conserved synteny between the three species was high. The 25 genes analysed on chromosome 19 cover 1.67 Mbp, while the 25 genes on chromosome 13 cover only 322 Kbp. This together with the lack of conserved synteny between paralogs (from the salicoid WGD) are examples of how this region on chromosome 19 has a distinctive evolutionary history.

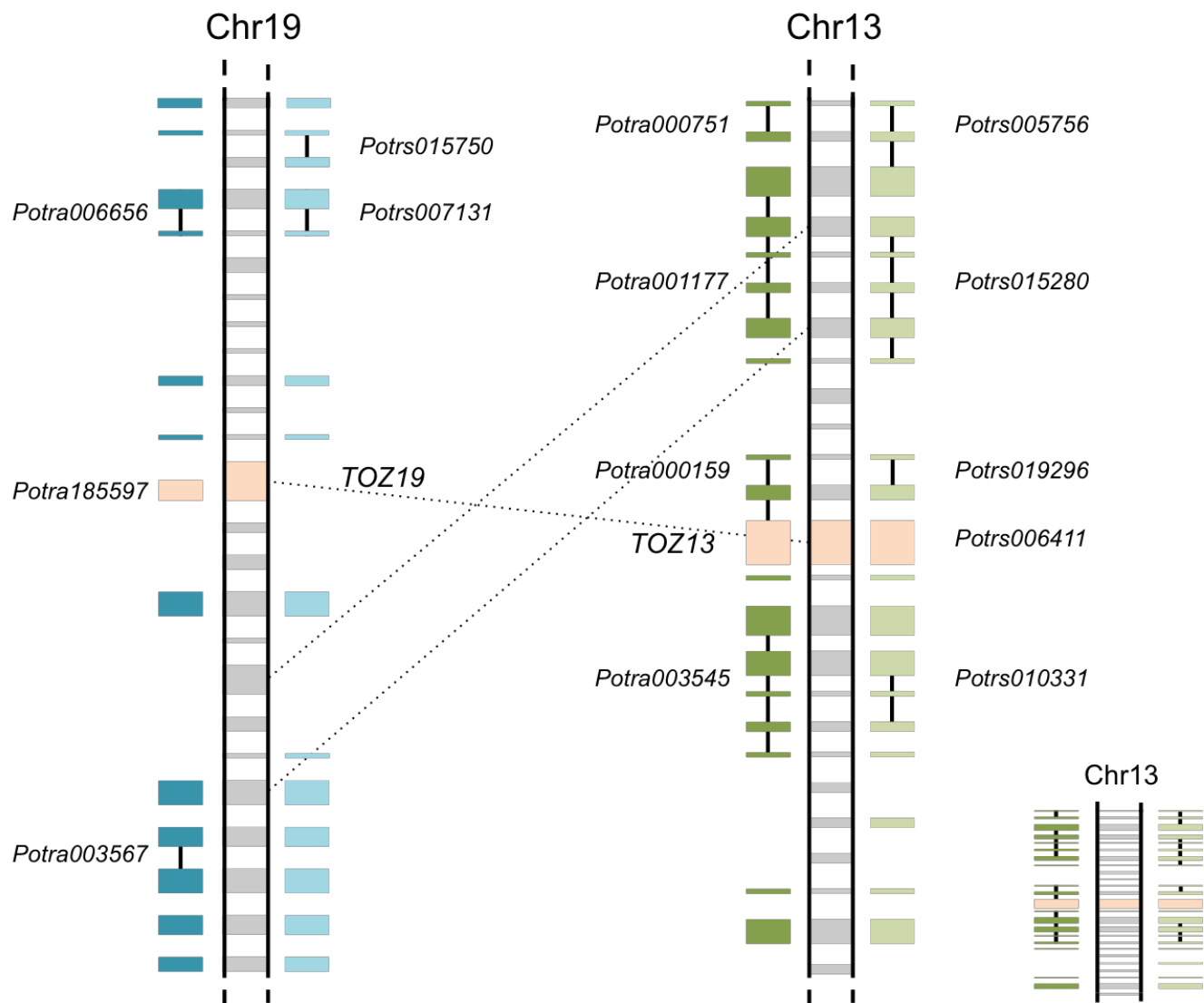


Figure S6.1 Schematic representation of the sex-determining region in aspen. To the left, the middle shows *TOZ19* and 12 genes up-stream and downstream. On both sides, the orthologs (detected using both BLAST and conserved synteny) in *Potra* and *Potrs* are depicted. The right shows the *TOZ19* paralog, *TOZ13*, accompanied by 12 genes on each side. Scaffold IDs are noted for the *TOZ19* and *TOZ13* gene, and in cases of more than one gene per scaffold. Dotted lines indicate paralogs in *Potri*. The small insert bottom right shows the region on chromosome 13 drawn to the same scale used for the region on chromosome 19.

7 Population genetics

7.1 Comparative population genomics statistics among species

All analyses in this section are based on the sequencing data that was mapped to its own assembly in all three species. This is a reanalysis of the data presented in Wang et al. (2016)(1), in which the read data was aligned only to the reference

Potri genome. As detailed in Wang et al. (2016)(1) this alignment approach masked substantial portions of the genome due to high sequence divergence rates among the species.

7.1.1 Methods

Samples and sequencing

Whole-genome re-sequencing data, as described in Wang et al. (2016)(1) from 24 genotypes of Potra, 22 genotypes of Potrs and 24 genotypes of Potri (Figure S7.1).

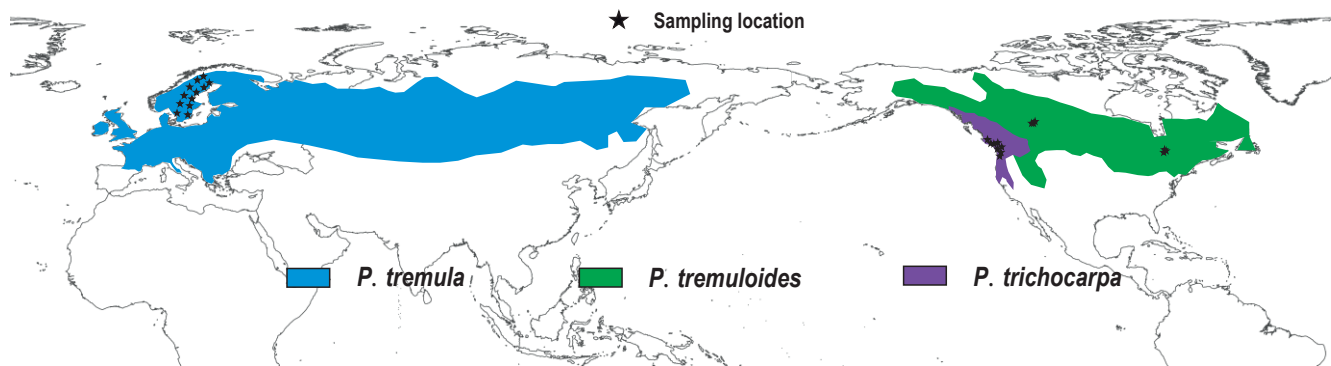


Figure S7.1. Sampling localities (details in Table S7.1, black star symbols) and distribution of *P. tremula* (blue areas), *P. tremuloides* (green areas) and *P. trichocarpa* (purple areas).

Raw read filtering, read alignment and post-processing alignment

Prior to read alignment, we used Trimmomatic to remove adapter sequences from reads and to cut off bases from the start and/or end of reads when the quality values were <20. If the length of the processed reads was reduced < 36 bp after trimming, reads were completely discarded. After quality control, all PE and orphaned SE reads from samples of Potra and Potrs were mapped to the new assembled genome presented here, whereas all reads from Potri samples were mapped to the Potri reference genome (v3.0). All read alignments were all performed using the bwa-mem algorithm with default parameters in (v0.7.10).

Several post-processing steps of alignments were performed to minimise the number of artefacts occurring in downstream analyses: First, we performed INDEL realignment since mismatching bases were usually found in regions with insertions and deletions (68). The RealignerTargetCreator in GATK was first used to find suspicious-looking intervals that likely required realignment. Then, the IndelRealigner was used to run the realigner over those intervals. Second, as reads resulting from PCR duplicates can arise during the sequencing library preparation, we used the MarkDuplicates

methods in the Picard package (<http://picard.sourceforge.net>) to remove those reads or read pairs having identical external coordinates and the same insert length. In such cases only we kept only the single read with the highest summed base qualities for downstream analysis. Third, in order to exclude genotyping errors caused by paralogous or repetitive DNA sequences where reads were poorly mapped to the reference genome, we removed sites with extremely low and extremely high read depths after investigating the empirical distribution of read coverage. We filtered out sites with a total coverage <100 or >2X the median coverage across all samples per species. When reads were mapped to multiple locations in the genome, they were randomly assigned to one location with a mapping score of zero. In order to account for such misalignment effects, we removed those sites if there were on average >1 mapped reads with mapping score equalling to zero per individual in each species. We used sites that passed all these filtering criteria in each species in downstream analyses.

Population genetic summary statistics

We used ANGSD (69) to estimate average pairwise nucleotide diversity (Θ_n) (70) and the neutrality statistic test Tajima's D (70) for different types of functional element (0-fold non-synonymous, 4-fold synonymous, intron, 3' UTR, 5' UTR, upstream and downstream regulatory regions and intergenic sites) over non-overlapping 1 Kbp windows in all three *Populus* species. Up- and Down-stream regulatory regions were defined as 1 Kbp regions up-/down-stream of the transcriptional start/stop site for genes. Only reads with mapping quality above 30 and the bases with quality score higher than 20 were included in these analyses.

Linkage disequilibrium (LD) and population-scaled recombination rate (ρ)

As estimates of LD and ρ require accurate SNP and genotype calls, we performed SNP calling with HaplotypeCaller of the GATK (v3.2.2), which called SNPs and INDELS simultaneously via local re-assembly of haplotypes for each individual and created single-sample gVCFs. We then used GenotypeGVCFs in GATK to merge multi-sample records together, correct genotype likelihoods, and re-genotype the newly merged record and perform re-annotation. We then performed several filtering steps to reduce the number of false positive SNPs and retain high-quality SNPs: (1) We removed all SNPs that overlapped with sites excluded by all previous filtering criteria. (2) We only retained bi-allelic SNPs with a distance of more than 5 bp away from any INDEL. (3) We treated genotypes with quality score (GQ) lower than 10 as missing and then removed those SNPs with genotype missing rate higher than 20%. (4) We removed SNPs that showed significant deviation from Hardy-Weinberg Equilibrium ($P < 0.001$). To estimate and compare the rate of LD decay among different annotation contexts in *Potra*, *Potrs* and *Potri*, respectively, we first used PLINK (71) (v1.9) to randomly thin the SNPs with minor allele frequency higher than 10% to 100,000 for each type of functional element. We then calculated the squared correlation coefficients (r^2) between all pairs of SNPs that were within a distance of 20 Kbp using PLINK.

To estimate the rate of decay of LD with physical distance, nonlinear regression was used to assess the relationship between the distance (in bp) and the degree of LD between all pairs of polymorphic sites (72). LD is assumed to be scored between pairs of polymorphic sites using the squared allele frequency correlations, r^2 . Under a simple drift-recombination model the expected value of r^2 is $E(r^2) = 1/(1+\rho)$, where $\rho=4Nc$ is the scaled recombination rate for the gene region. In the presence of mutations, the expectation changes to

$$E(r^2) = \left(\frac{10+\rho}{(2+\rho)(11+\rho)} \right) \left(1 + \frac{(3+\rho)(12+12\rho+\rho^2)}{n(2+\rho)(11+\rho)} \right) \quad (1)$$

where n is number of haplotypes sampled. Based on equation 1 and the decay of r^2 , we then estimated the scaled recombination rate ρ for each type of functional elements.

7.1.2 Results

The re-sequencing data for 24 Potra, 22 Potrs and 24 Potri individual is summarised in Table S7.1. Sequencing coverage varied from 15x to 53x for different individuals, with a median sequencing depth of 34.8x, 30.9x and 26.2x for Potra, Potrs and Potri, respectively. Mapping rates of reads exceeded 95% for Potra and Potrs but were a little lower for Potri (median mapping rate 96.3%).

Estimates of nucleotide diversity varied three to four-fold between aspen and Potri and also between 4-fold synonymous and 0-fold non-synonymous sites, with Potrs having the highest level of nucleotide diversity and Potri the lowest (Table S7.2). Nucleotide diversity also varied across non-coding regions, with 5'UTR showing stronger diversity reduction compared with 3'UTR and intronic regions. In accordance with inter-species sequence divergence (section 4.3 and 5.4), the level of intra-species diversity increased substantially with increasing proximity to genes (Table S7.2). Patterns of diversity in both coding and non-coding regions showed consistency across species.

We found that all three species showed a genome-wide excess of low-frequency polymorphisms, as evidenced by negative Tajima's D, although deviations were greater in Potra and Potrs compared to Potri (Table S7.3). In both aspens, 0-fold nonsynonymous sites showed the lowest Tajima's D values compared with other sites, suggesting they are under stronger effects of purifying selection than other genomic contexts. However, this pattern was not found in Potri.

Table S7.1 Summaries of whole-genome re-sequencing data for three *Populus* species that mapped to their respective genomes

SampleID	Site	Latitude	Longitude	Raw bases(Gb)	Filtered bases(Gb)	Uniquely mapped bases(Gb)	Mapping rate (%)	Proportion of covered genome (%)	Mean Coverage
<i>Populus tremula</i>									
SwAsp001	Simlang	56,6925	13,2147	21,39	12,02	10,8	98,97%	88,28%	30,23
SwAsp009	Simlang	56,7336	13,2517	16,95	13,73	12,66	98,95%	88,85%	35,12
SwAsp011	Ronneby	56,3478	15,025	34,59	29,08	26,32	98,83%	89,85%	72,59
SwAsp014	Ronneby	56,3081	15,1269	18,48	14,46	13,37	98,71%	89,02%	36,96
SwAsp021	Vargarda	57,9917	12,9119	13,80	10,73	9,94	98,77%	88,52%	27,68
SwAsp025	Vargarda	57,9869	12,9358	13,65	10,08	9,34	98,68%	88,63%	25,99
SwAsp032	Ydre	57,8492	15,3217	15,23	12,00	11,1	98,59%	88,90%	30,84
SwAsp033	Ydre	57,8281	15,3103	15,78	11,89	10,84	98,94%	88,64%	30,12
SwAsp045	Brunsborg	59,6425	12,9408	15,31	11,85	10,96	98,77%	88,83%	30,39
SwAsp047	Brunsborg	59,6308	12,9608	20,77	11,83	10,61	98,76%	88,85%	29,48
SwAsp055	Uppsala	59,8131	17,9817	17,01	13,14	12,22	98,71%	88,90%	33,77
SwAsp057	Uppsala	59,7761	17,9889	18,44	14,12	13,11	98,61%	89,11%	36,2
SwAsp067	Alvdalen	61,1978	13,8092	16,52	13,05	12,04	98,61%	88,98%	33,41
SwAsp068	Alvdalen	61,3017	13,7222	17,18	13,47	12,49	98,61%	88,88%	34,64
SwAsp076	Delsbo	61,7106	16,7311	23,91	21,32	19,6	98,87%	89,25%	54,27
SwAsp078	Delsbo	61,6925	16,6700	15,61	12,54	11,5	98,64%	88,89%	31,89
SwAsp087	Dorotea	64,3406	16,3992	17,27	14,21	13,15	98,90%	88,90%	36,49
SwAsp088	Dorotea	64,3358	16,3736	16,47	13,51	12,56	98,91%	88,69%	34,92
SwAsp096	Umea	63,9781	20,7056	16,99	13,66	12,53	98,97%	88,74%	34,88
SwAsp098	Umea	63,8656	20,4986	15,94	14,14	13,16	98,97%	88,57%	36,67
SwAsp103	Arjeplog	66,0247	18,5742	14,59	12,76	11,8	99,07%	88,46%	32,93
SwAsp110	Arjeplog	66,2592	18,0000	24,49	21,60	19,83	98,95%	91,07%	53,68
SwAsp111	Lulea	65,6703	21,8986	20,95	18,23	16,84	98,97%	88,86%	46,86
SwAsp114	Lulea	65,5544	22,3939	22,96	20,22	18,36	98,90%	88,91%	50,85

P. tremuloides

Alb10-3	Alberta	51,0718	-115,0044	14,99	12,66	11,03	98,06%	93,59%	29,15
Alb13-1	Alberta	51,0479	-115,0232	16,04	13,46	12,02	99,02%	91,70%	32,61
Alb16-1	Alberta	51,0838	-115,3892	13,94	11,45	10,25	98,45%	91,76%	27,63
Alb17-4	Alberta	51,0809	-115,3946	12,54	10,41	9,53	98,91%	91,78%	25,72
Alb18-3	Alberta	51,0686	-115,3516	14,60	12,38	11,29	98,89%	91,95%	30,64
Alb25-4	Alberta	51,0524	-114,9131	12,99	10,62	9,62	98,61%	91,67%	25,87
Alb27-1	Alberta	51,0405	-114,8939	20,07	17,27	15,2	99,01%	92,41%	41,16
Alb31-1	Alberta	51,0435	-114,8352	16,29	13,38	12,1	98,88%	91,99%	32,84
Alb33-2	Alberta	51,0431	-114,7568	19,44	16,09	14,63	98,93%	92,19%	39,49
Alb35-2	Alberta	51,0234	-115,0640	12,57	10,3	9,46	98,88%	91,48%	25,65
Alb6-3	Alberta	51,1324	-115,0664	19,47	16,67	14,99	99,05%	93,89%	39,65
Albb15-3	Alberta	51,0811	-115,3767	15,14	12,77	11,48	98,96%	91,72%	31,17
Dan1-1C13	UW Arboreturn	43,052	89,4242	16,42	14,23	12,91	99,02%	92,79%	34,59
Dan2-1B7	UW Arboreturn	43,0526	89,4253	22,00	18,48	16,61	99,43%	98,23%	42,15
PG1-1B4	Parfrey's Glenn	43,4249	89,6445	12,11	10,68	9,79	98,95%	91,74%	26,32
PG2-1B9	Parfrey's Glenn	43,4184	89,6417	16,53	14,13	12,8	98,96%	92,27%	34,34
PG3-1B6	Parfrey's Glenn	43,4198	89,6532	14,93	12,78	11,54	98,92%	92,91%	30,62
PI3-1B3	Pine Island Preserve area	43,5402	89,5666	16,75	13,37	12,22	98,78%	92,31%	32,71
Sau1-1B10	Boxter's Hollow	43,4036	89,8176	13,30	10,61	9,48	98,18%	92,45%	25,41
Sau2-1B2	Boxter's Hollow	43,4048	89,8243	13,74	6,82	6,20	98,81%	93,62%	16,34
Sau3-1B13	Boxter's Hollow	43,4053	89,8141	12,13	9,86	8,65	98,82%	91,65%	23,56
Wau1-1B5	Waushara country	44,1314	89,2082	26,40	21,73	17,41	98,89%	92,32%	48,26

P. trichocarpa

BESC-56	Talley_Way	46,099	-122,878	14,51	12,44	11,43	97,06%	94,08%	25,43
BESC-108	Rainier	46,114	-122,99	15,97	13,58	12,39	94,99%	93,75%	27,75
BESC-264	Monroe	47,842	-121,979	19,78	17,03	15,42	95,22%	94,01%	34,35
BESC-281	Monroe	47,851	-121,962	12,21	10,83	10,13	97,31%	93,48%	22,89

BESC-374	Skiou_Island	48,489	-122,16	19,63	16,48	14,94	95,33%	94,18%	33,64
BESC-840	Orting	47,042	-122,209	16,74	13,38	12,08	94,53%	94,03%	27,56
BESC-873	Sultan	47,856	-121,811	13,17	11,76	10,9	97,20%	93,68%	24,36
BESC-884	Gold_Bar	47,84	-121,691	14,32	12,46	11,52	96,66%	93,83%	25,89
DEND-17-2	DEND	52,817	-126,95	8,33	7,32	6,73	97,14%	93,08%	15,26
FNYI-28-4	Fanny_Bay_	52,817	-126,95	15,71	14,14	12,79	96,58%	93,82%	28,29
GW-11031	Corvallis	52,817	-126,95	19,01	16,86	15,24	96,35%	94,28%	34,23
GW-7983	Longview	46,117	-123	10,68	9,58	8,91	97,18%	93,22%	20,14
GW-9595	Turner	44,75	-122,867	16,17	13,77	12,12	93,13%	93,76%	27,18
GW-9598	Nisqually_River	47,067	-123,733	12,36	11,09	9,67	91,74%	93,69%	21,84
GW-9772	Acme	48,717	-122,2	15,55	13,35	12,12	96,31%	93,89%	26,9
GW-9792	Sedro_Woolley	48,717	-122,2	14,53	12,84	11,78	96,88%	93,66%	26,51
GW-9920	Orting	47,1	-122,2	12,67	11,1	9,36	88,71%	93,59%	21,23
GW-9959	Skamania	45,95	-121,95	14,41	9,75	8,3	89,35%	93,71%	18,64
HARC-26-1	Harrison	49,767	-122,217	12,07	10,2	9,56	97,41%	93,63%	21,76
HOMC-21-3	Homathko	51,233	-124,95	18,36	17,15	15,32	97,21%	93,63%	34,05
LILC-26-3	Harrison	51,233	-124,95	16,83	14,29	12,74	95,11%	93,89%	28,53
SLMB-28-4	Salmon	50,217	-125,817	13,43	11,06	9,84	92,43%	93,81%	22,33
SQMB-25-4	Squamish	50,217	-125,817	16	13,73	12,52	95,03%	94,03%	28,03
WHTE-28-1	Salmon	50,133	-126,05	12,75	11,24	10,28	96,69%	93,38%	23,13

Table S7.2 Estimates of nucleotide diversity (median and central 95% range) for various genomic contexts over 1 Kbp non-overlapping windows across the genome in Potra, Potrs and Potri.

	Potra	Potrs	Potri
0-fold	0.0024(0.0000-0.0155)	0.0032(0.0005-0.0164)	0.0008(0.0000-0.0145)
4-fold	0.0095(0.0006-0.0304)	0.0108(0.0017-0.0327)	0.0031(0.0000-0.0177)
UTR3'	0.0089(0.0009-0.0270)	0.0105(0.0021-0.0239)	0.0039(0.0000-0.0181)
UTR5'	0.0071(0.0006-0.0250)	0.0081(0.0015-0.0256)	0.0033(0.0000-0.0166)
Intronic	0.0087(0.0014-0.0298)	0.0095(0.0025-0.0279)	0.0037(0.0001-0.0215)
Upstream	0.0140(0.0023-0.0462)	0.0159(0.0037-0.0489)	0.0083(0.0005-0.0371)
Downstream	0.0142(0.0025-0.0471)	0.0157(0.0039-0.0486)	0.0082(0.0005-0.0372)
Intergenic	0.0167(0.0034-0.0493)	0.0188(0.0049-0.0494)	0.0108(0.0013-0.0382)

Table S7.3 Estimates of Tajima’s D (median and central 95% range) for various genomic contexts over 1Kbp non-overlapping windows across the genome in Potra, Potrs and Potri.

	Potra	Potrs	Potri
0-fold	-1.0531 (-2.2591-1.2452)	-1.9560 (-2.7356--0.1236)	-0.0384 (-1.7020-1.9754)
4-fold	-0.3454 (-1.9408-1.7627)	-1.2888 (-2.4273-0.6603)	-0.0274 (-1.6838-2.2592)
UTR3	-0.4210 (-2.0175-1.6020)	-1.4321 (-2.5187-0.2909)	-0.0650 (-1.8920-2.3140)
UTR5	-0.7019 (-2.1548-1.5376)	-1.6556 (-2.6067-0.2405)	-0.0805 (-1.8479-2.2268)
Intronic	-0.4101 (-2.0575-1.5959)	-1.5305 (-2.6069-0.2230)	-0.1176 (-1.9669-2.3226)
Upstream	-0.4990 (-2.0431-1.3092)	-1.3659 (-2.4729-0.2408)	-0.0071 (-1.8845-2.2410)
Downstream	-0.3613 (-1.9537-1.5138)	-1.2934 (-2.4455-0.3119)	-0.0109 (-1.8971-2.2345)
Intergenic	-0.4941 (-2.0064-1.2944)	-1.3916 (-2.4479-0.1278)	-0.0231 (-1.9146-2.1738)

Linkage disequilibrium was lowest in Potrs and highest in Potri, and generally decayed more rapidly in regions that were in close proximity to genes (e.g. Regulatory and UTRs) compared to genic regions and other regions far from genes in all three species (Figure S7.2). These results were also reflected in estimates of recombination rates, which were substantially higher in Potrs compared to the other two species (Figure S7.3). Similarly, population scaled recombination rates estimated from linkage disequilibrium decay were generally high in UTRs and Regulatory regions (1 kbp up-/downstream of the transcriptional start/stop site for genes) in all three species (Figure S7.3; Table S7.4).

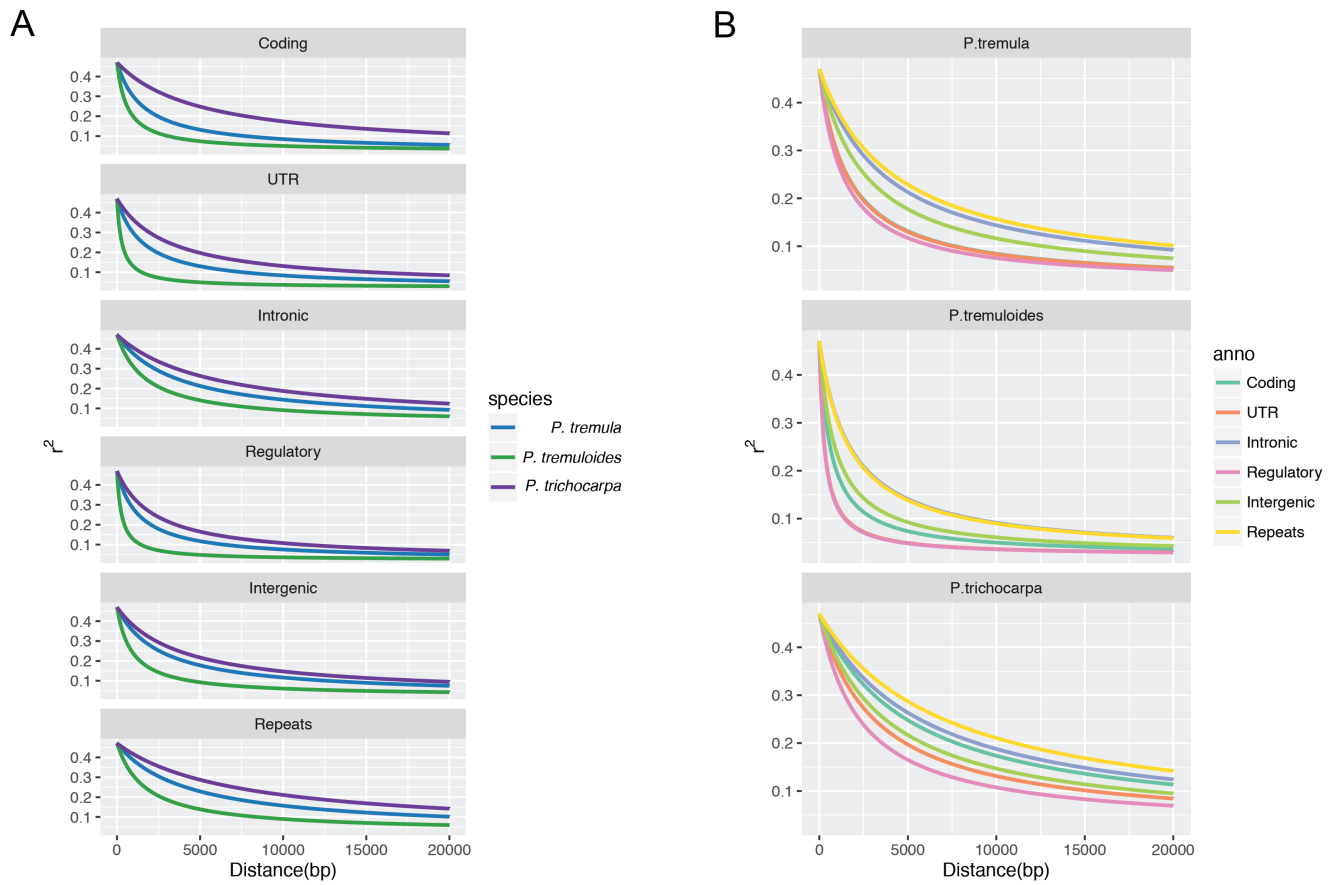


Figure S7.2 Decay of linkage disequilibrium (LD) with physical distance grouped by annotation class (A) and by species (B), respectively.

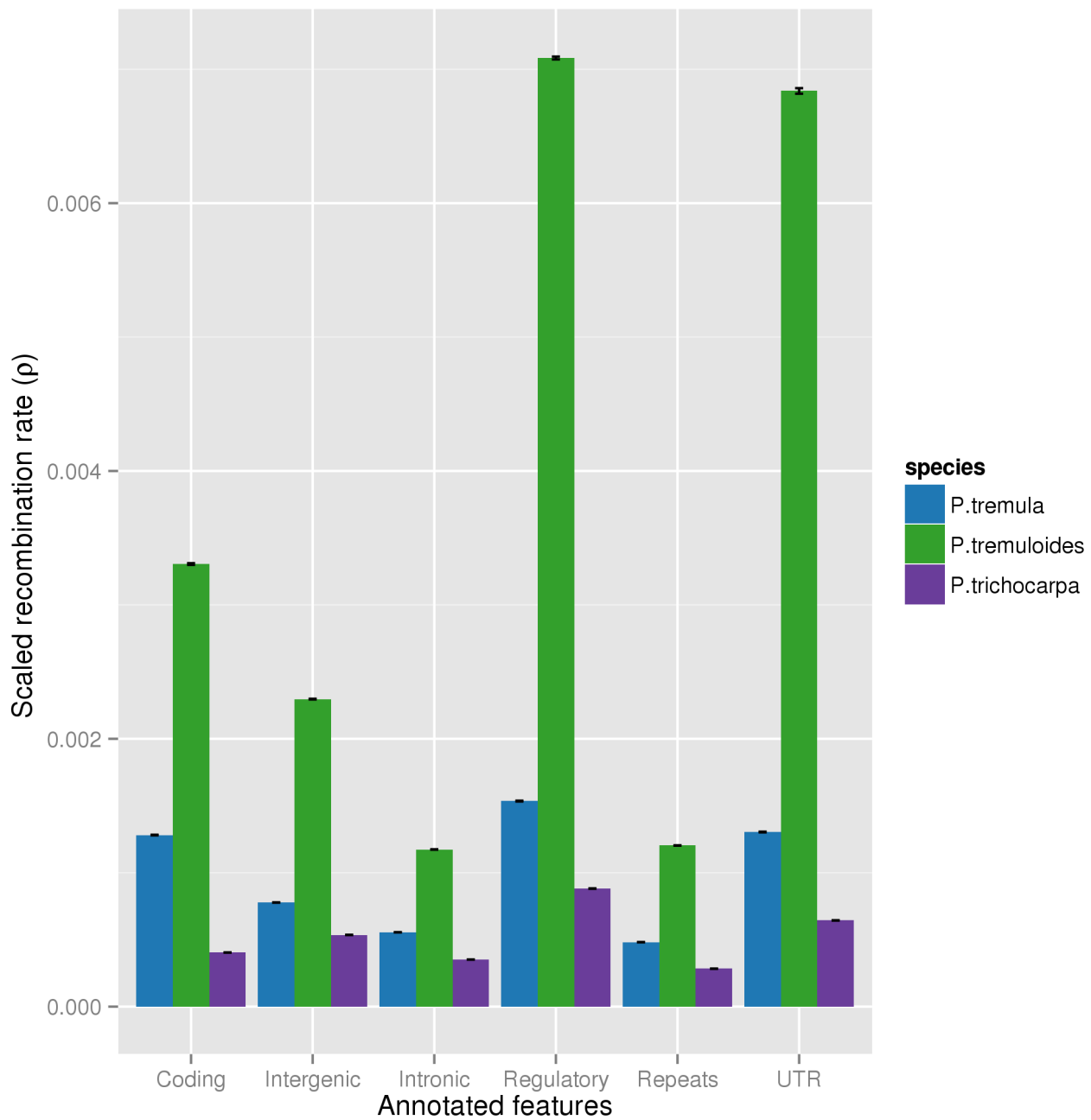


Figure S7.3 Estimates of population-scaled recombination rates among the three *Populus* species in both coding and noncoding regions.

Table S7.4 The population scaled recombination rate (ρ) estimated from LD decay, corresponding to Figure S7.3

	Potra_ ρ (SE)	Potrs_ ρ (SE)	Potri_ ρ (SE)
UTR	0.00130(2.541e-06)	0.00684(2.0699e-05)	0.00064(1.368e-06)
Coding	0.00128(2.225e-06)	0.00331(6.4160e-06)	0.00040(6.610e-07)
Intronic	0.00056(8.970e-07)	0.00117(1.4730e-06)	0.00035(5.580e-07)
Intergenic	0.00078(1.298e-06)	0.00230(2.6580e-06)	0.00054(9.130e-07)
Regulatory	0.00154(2.981e-06)	0.00708(1.0612e-05)	0.00088(1.535e-06)
Repeats	0.00048(7.740e-07)	0.00120(1.2930e-06)	0.00028(5.050e-07)

7.2 Genome-wide measures of negative and positive selection in aspens

In order to quantify the extent of negative selection acting on sites within different genomic contexts in the two aspen species, we used the methods of Keightley & Eyre-Walker (2007)(73) to compare the allele frequency spectrum (AFS) of various site categories to those for 4-fold synonymous sites (which were putatively neutral), and estimate the distribution of fitness effects (DFE). After taking into account the effect of slightly deleterious mutations, we further used the approach presented in Eyre-Walker & Keightley (2009)(74) to estimate the proportion of fixations driven by positive selection (α) and the rate of positive selection (ω) in both coding and noncoding regions. Potri was used as outgroup for both aspen species and, thus, all the analyses in this section (7.2) were based on the sequencing data mapped to Potri for the two aspen species.

7.2.1 Methods

Alignment and SNP calling

After quality control, all genomic reads from the 24 Potra and 22 Potrs were mapped to the Potri genome (v3.0) using the bwa-mem algorithm with default parameters (v0.7.10). We used the same post-processing steps of alignments as detailed above. We performed SNP and genotype calls using the GATK UnifiedGenotyper with default parameters across all sites. We excluded both variant and invariant sites from the allele frequency spectra (AFS) if the sites were: (1) located in regions that were excluded by all previous filtering criteria as shown in Wang et al. (2016)(1); (2) with a distance of more than 5 bp away from any INDELS; (3) with genotype missing rate higher than 20% after treating genotypes with quality score (GQ) lower than 10 among all samples as missing.

Estimating the distribution of fitness effects (DFE), α and ω

We generated folded AFS and divergence values from counts of sites using a custom Perl script for each category of functional elements: 4-fold synonymous sites, 0-fold non-synonymous sites, 3' UTR, 5' UTR, intronic or intergenic across all remaining sites after filtering. We used Potri as an outgroup to calculate between-species nucleotide divergence at 4-fold synonymous and potential selected sites as it is unlikely to be influenced by shared ancestral polymorphisms (1). Jukes-Cantor multiple hits correction was applied to the divergence estimates (75). Using 4-fold synonymous sites as the neutral reference, we estimated the distribution of fitness effects ($N_e s$), the proportion of fixations driven by positive selection (α) and the rate of positive selection (ω) for each category of functional elements using the approach of Eyre-Walker & Keightley (2009) (74) as implemented in the program DFE-alpha (73, 74). For the parameters of $N_e s$, α and ω , we generated 200 bootstrap replicates by resampling randomly across all sites in each class using R. We excluded the top and bottom 2.5% of bootstrap replicates and used the remainder to represent the 95% confidence intervals for each parameter.

7.2.2 Results

In both aspen species, negative selection was much stronger in coding than non-coding regions, which was most clearly seen in 0-fold non-synonymous sites, where > 40% of sites were subject to strong negative selection (strength of negative selection $N_e s > 100$; Figure S7.4A, B). This selection pattern observed in coding regions is also consistent with the patterns of polymorphism. The nucleotide diversity at 0-fold nonsynonymous sites showed the strongest reduction ($\pi_{P.tremula} = 0.0024$; $\pi_{P.tremuloides} = 0.0032$) and a more negative Tajima's D ($D_{P.tremula} = -1.053$; $D_{P.tremuloides} = -1.956$) compared to 4-fold synonymous sites ($\pi_{P.tremula} = 0.0095$, $\pi_{P.tremuloides} = 0.0108$; $D_{P.tremula} = -0.345$; $D_{P.tremuloides} = -1.288$) and other non-coding sites (Table S7.2; Table S7.3). Among non-coding regions, 5' UTRs showed stronger negative selection compared to 3' UTR, intronic and intergenic regions (Figure S7.4A, B; Table S7.5), and ~ 30% of 5' UTRs are under moderate levels of negative selection ($1 < N_e s < 100$) in both aspen species. While both UTRs and intronic sites showed a signal of negative selection, few were subject to strong selection ($N_e s > 100$) compared to 0-fold nonsynonymous sites. Additionally, we estimated that approximately 100% of intergenic sites were effectively neutral ($N_e s < 1$) (Figure S7.4A, B; Table S7.5), which implies a general lack of negative selection in most non-genic regions, and/or nearly equivalent, or even weaker, negative selection compared to synonymous sites (76). After taking into account the effect of slightly deleterious mutations we further used the approach presented in Eyre-Walker & Keightley (2009) (74) to estimate the proportion of fixations driven by positive selection (α) and the rate of positive selection (ω) in both coding and non-coding regions. We found that 0-fold nonsynonymous sites showed a high proportion ($\alpha = 30-40\%$) of divergence driven by positive selection in both Pota and Potrs (using Potri as an outgroup), corroborating results from earlier studies based on substantially smaller numbers of genes (76). Similarly, 5' UTRs showed a high proportion of divergence driven by positive selection ($\alpha = 30-40\%$) and exhibited evidence of stronger positive selection than 3' UTR, intronic and intergenic regions. Due to

the high sequence divergence and consequent alignment errors between aspen and Potri in intergenic regions, future efforts are required to understand selection patterns in these regions.

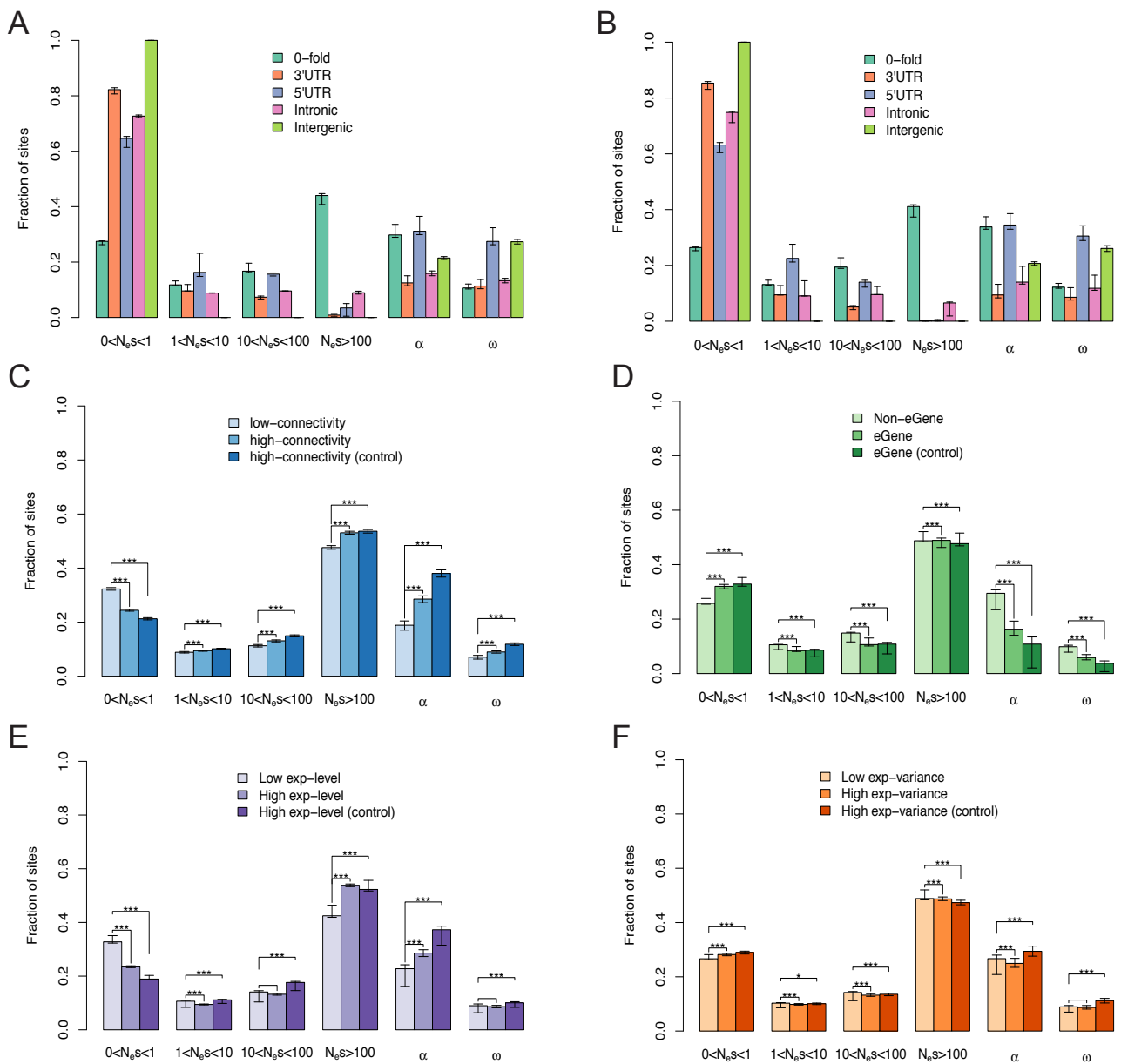


Figure S7.4 Estimates of negative and positive selection on coding and noncoding regions, separated by site type, in Potra (A) and Potrs (B). Error bars represent 95% bootstrap confidence intervals. More detailed information can be found in Table S7.5. (C-F) Estimates of negative and positive selection on 0-fold nonsynonymous sites in genes (C) with varying connectivity level in co-expression network; (D) with eQTLs (eGene) or not (non-eGene); (E) with varying expression level; (F) with varying expression variance. In addition, for gene categories of high connectivity, eGene, high expression level and high expression variance, the control group is to use the same neutral reference (4-fold synonymous sites from

the corresponding “low” gene categories) in order to control for the synonymous selection difference between these two categories of genes. N_e s categories, different bins of negative selection strength; α , the proportion of divergent sites fixed by positive selection; ω , the rate of adaptive substitution relative to neutral divergence.

Table S7.5 Estimates of the distribution of fitness effects of new mutations at 0-fold nonsynonymous sites, intronic sites, 5' UTR sites, 3' UTR sites and intergenic sites falling in different N_e s ranges, and proportion of divergence driven to fixation by positive selection (α) and the rate of adaptive substitution relative to neutral divergence (ω) in Potra and Potrs. 95% bootstrap confidence intervals are shown in parentheses.

Species	Category	Percentage of mutations in N_e s range				α	ω
		0-1	1-10	10-100	>100		
Potra	zero-fold	0.275 (0.263-0.277)	0.118 (0.115-0.132)	0.167 (0.163-0.196)	0.440 (0.408-0.447)	0.298 (0.289-0.335)	0.107 (0.103-0.121)
	UTR 3'	0.822 (0.807-0.829)	0.096 (0.096-0.119)	0.073 (0.067-0.077)	0.009 (0.003-0.011)	0.125 (0.115-0.151)	0.114 (0.104-0.137)
	UTR 5'	0.646 (0.614-0.653)	0.163 (0.148-0.232)	0.157 (0.150-0.161)	0.034 (0.005-0.050)	0.311 (0.299-0.365)	0.275 (0.263-0.324)
	Intronic	0.727 (0.722-0.731)	0.088 (0.088-0.089)	0.096 (0.096-0.096)	0.089 (0.085-0.095)	0.159 (0.151-0.167)	0.133 (0.125-0.141)
	Intergenic	1.000 (1.000-1.000)	0.000 (0.000-0.000)	0.000 (0.000-0.000)	0.000 (0.000-0.000)	0.215 (0.210-0.220)	0.273 (0.265-0.282)
Potrs	zero-fold	0.264 (0.251-0.266)	0.131 (0.128-0.147)	0.194 (0.190-0.227)	0.411 (0.376-0.416)	0.338 (0.330-0.374)	0.122 (0.118-0.135)
	UTR 3'	0.853 (0.834-0.858)	0.095 (0.094-0.120)	0.051 (0.046-0.056)	0.001 (0.000-0.002)	0.094 (0.084-0.132)	0.086 (0.076-0.120)
	UTR 5'	0.632 (0.601-0.643)	0.225 (0.207-0.278)	0.140 (0.123-0.147)	0.004 (0.000-0.007)	0.345 (0.329-0.385)	0.305 (0.289-0.342)
	Intronic	0.748 (0.713-0.752)	0.091 (0.090-0.140)	0.096 (0.095-0.124)	0.065 (0.023-0.069)	0.140 (0.133-0.197)	0.118 (0.111-0.165)
	Intergenic	1.000 (1.000-1.000)	0.000 (0.000-0.000)	0.000 (0.000-0.000)	0.000 (0.000-0.000)	0.207 (0.200-0.213)	0.261 (0.250-0.271)

7.3 Comparing patterns of negative and positive selection across genes in Potra

To investigate the relative contributions of negative and positive selection in driving differences in evolutionary rates for genes with different expression profiles, we used two complementary approaches derived from the McDonald-Kreitman (MK) test to quantify the strength of both negative and positive selection. The MK test, which compares the amount of polymorphism to divergence for categories of sites that are expected to evolve differently (e.g. synonymous vs.

nonsynonymous/noncoding), is robust to changes in population demography and holds promise for distinguishing between the relative roles of positive and negative selection in shaping divergence (77). In the first approach, we used a population-genetic model implemented in DFE-alpha as described above (73, 74) to explicitly model changes in population size (as reflected by the allele frequency spectrum of neutral sites) and the distribution of fitness effects (DFE) of new mutations at functional sites. After taking into account the effects of slightly deleterious mutations, we implemented the approach of Eyre-Walker and Keightley (2009)(74) to estimate the proportion of fixations driven by positive selection (α) and the rate of positive selection (ω). In the second approach, we used a generalized linear mixed model implemented in SnIPRE(78) to estimate mutation rates, divergence, constraint and selection effects simultaneously using genome-wide data without referring to any specific demographic model. This method explicitly incorporates the genome-wide effects as fixed effects and individual gene effects as random effects. For each gene, we focused our analyses on two key population genetics parameters: (1) the selection effect (γ), which quantifies the strength of selection per gene relative to neutrality, a low γ (<0) indicates that the gene evolved under negative selection and a high γ (>0) indicates the gene evolved under positive selection; (2) the constraint effect (f), which captures the proportion of nonsynonymous mutations that are non-lethal and quantifies the strength of selective constraint acting on a given gene (i.e. the removal of deleterious alleles from the general population). Therefore, f varies from 0 (all nonsynonymous mutations in the gene are lethal) to 1 (all nonsynonymous mutations in the gene are non-lethal).

7.3.1 Methods

Expression characterization

We used expression data from a natural population of Potra as described in Mähler et al. (2017), which generated paired-end RNA Seq expression data from winter buds at the point of spring bud flush for 219 individuals (clonal replicates) that represents 86 genotypes sampled from a common garden of the Swedish Aspen (SwAsp) collection. Based on the results from expression Quantitative Trait Locus (eQTL) mapping and co-expression network construction, five different gene expression features were measured for a total of 22,306 expressed genes (more detailed information of these gene features can be found in Mähler et al. (2017)(61): (1) gene expression level; (2) gene expression variance; (3) gene connectivity level measured from gene co-expression network; (4) the status of the gene (core vs. non-core) within the co-expression network modules; (5) genes for which the expression level is associated with at least one polymorphic locus (eGene) or not (non-eGene).

Genome resequencing and SNP calling

We used whole genome resequencing data from 94 Potra individuals (the same genotypes as for expression data shown above) from the Swedish Aspen collection, as described in Wang et al. (2018)(79). Briefly, this study conducted Illumina paired-end sequencing of individuals (with a mean read depth $\sim 30X$ per sample) from 12 locations spanning the latitudinal range in Sweden. Illumina reads were filtered by quality and then were mapped to Potra genome using *bwa-mem*. After initial alignment, duplicate reads were corrected by *MarkDuplicates* from *Picard* packages (<http://broadinstitute.github.io/picard/>), and indels were realigned using the GATK indel realigner and genotype and SNP calling were conducted using the GATK haplotypcaller under default parameters (80). Sites with extremely low (<400) or high (>4500) coverage, with a high number of reads (>200) with mapping score equalling zero, or being located in scaffolds with length smaller than 2kbp were removed (79). Only SNPs that were bi-allelic, with distance >5 bp away from indels, with available information derived from more than 70% of the sampled individuals with genotype quality score higher than 10, and not deviating from Hardy-Weinber equilibrium test were retained. Analyses were conducted on sites/SNPs that passed all previous filtering criteria. In order to determine the aspen lineage-specific divergence, we aligned reads from one Potri (used as outgroup species) individual (SRA ID: SRR1571343) to the Potra assembly and used *UnifiedGenotyper* in GATK to call SNPs at all sites. Sites with read coverage lower than 4 or higher than 70, with genotype quality lower than 30, with both alleles differing from Potra assembly, or with indels nearby were removed. In the end, 230,311,225 informative sites from Potri were retained for divergence measurements across the genome.

Distribution of fitness effects (DFE) and adaptive evolution

To estimate and compare the strength of positive and negative selection across gene categories with different expression profiles, we used SNP genotypes to estimate and summarize the allele frequency spectrum and divergence across all genes for each category of sites in the specific gene expression class. Firstly, for the 22,306 expressed genes, except for the status of eGene (6,241 eGenes and 16,065 non-eGenes) and core-genes (1,795 core genes and 20,511 non-core genes), which are binary variables, genes for another three features were sorted and split into two equally sized groups, which will be referred to as “low” (11,153 genes) and “high” (11,153 genes) gene class. Thereafter, site categories (4-fold synonymous, 0-fold nonsynonymous, 5' UTR, 3' UTR, Intronic, 1kbp upstream and downstream regulatory sites) were determined based on the gene annotation of the Potra assembly. Using 4-fold synonymous sites as a neutral reference, we used DFE-alpha to estimate the fraction of 0-fold nonsynonymous and other noncoding sites within specific gene classes that are under negative selection, and measure positive selection, α and ω . We estimated 95% confidence intervals (CI) using 200 bootstraps by sampling genes with replacement from each gene expression class. We determined significance between pairs of gene classes using Mann-Whitney tests. One concern for this type of comparison is the extent to which 4-fold synonymous sites are neutrally evolving, since synonymous site selection, if it exists, might differ between gene classes and will therefore influence the inference of both purifying and positive selection. To assess the effect of synonymous site selection on our DFE-alpha inference, for each gene category we

independently used the same neutral reference (4-fold synonymous sites) for “high” gene classes (corresponding to eGene, high connectivity, high gene expression level and high gene expression variance, respectively) as those for “low” gene classes (corresponding to non-eGene, low connectivity, low gene expression level and low gene expression variance, respectively), and use them as a control sets (see below). We then reran DFE-alpha on the control “high” gene class and test for whether there was evidence for a significant difference in the strength and direction of natural selection between different gene classes of each expression category.

SnIPRE approach

To further identify potential genes under selection, we analysed SNPs that were annotated as 4-fold synonymous and 0-fold nonsynonymous for each of the 22,306 expressed genes. We obtained d_s , d_N , p_s , p_N (divergence and polymorphism at 4-fold synonymous and 0-fold nonsynonymous sites) and the proportion of 4-fold synonymous and 0-fold nonsynonymous sites in all genes. SnIPRE was then used to estimate the selection coefficient γ and constraint coefficient f . The program was run in the fully Bayesian mode, with 100000 iterations of the MCMC sampler after 10000 iterations of burn-in and thinning parameter set to 4. The significant effects for the selection coefficient γ on genes are classified as being neutral, negative or positive, and the significant effects for the constraint coefficient f on genes are classified as neutral or constraint.

7.3.2 Results

DFE and adaptive divergence

Given that the results between the status of core gene and gene connectivity level in this section were highly similar, we only present the results from gene connectivity level here. We found that genes with higher connectivity had a significantly lower proportion of nearly neutral nonsynonymous mutations than genes with lower connectivity, and a significantly increased proportion of nonsynonymous mutations under strong negative selection (strength of negative selection $N_e s > 100$) (Figure S7.4C; *SI Appendix Table 2*). After accounting for the potential bias of different synonymous site selection between genes with lower and higher connectivity, we found negative selection acts more strongly on genes with higher connectivity (Figure S7.4C; *SI Appendix Table 2*), indicating that stronger negative selection also acts on synonymous sites of highly than lowly connected genes. In addition, we found genes with higher connectivity showed a significantly higher proportion of adaptive nonsynonymous substitutions (α) and a higher rate of positive selection (ω) than genes with lower connectivity, suggesting that there is stronger positive selection acting on highly connectivity genes (Figure S7.4C). Therefore, both stronger negative and positive selection on higher centrality of essential genes in the co-expression network may promote more rapid and efficient adaptation in long-lived forest trees such as aspen (81).

In contrast to the clear evidence for stronger negative and positive selection on nonsynonymous sites for genes with higher connectivity, weak differences in patterns of selection between high and low connectivity genes were found at non-coding regions (Figure S7.5-S7.7, *SI Appendix Table 2*). However, after accounting for synonymous site selection using the control dataset, there was evidence for stronger negative and positive selection at introns, 5' and 3' UTRs, suggesting that different patterns of selection at synonymous sites between genes with higher and lower connectivity may interfere with our inference of selection patterns in non-coding regions. Similar bias also applied to 1 Kbp up- and down-stream non-coding regions (Figure S7.8-S7.9, *SI Appendix Table 2*), with results for these regions indicating weaker negative selection on high than low connectivity genes. However, the differences in selection patterns between gene categories were much reduced after accounting for selection at synonymous sites, likely suggesting little actual difference in selection patterns for up- and down-stream regions of genes with high and low connectivity. This is in accordance with our above inference that there is a general lack of selection in intergenic regions. Considering selection on synonymous sites may widely bias our estimates and comparison of selection on non-coding regions between different categories of genes (not only for connectivity level but also in another three features listed above; Figure S7.5-S7.9, Dataset S2), we only focus on our results at 0-fold nonsynonymous sites in downstream text. More investigation of the action of selection on synonymous sites among various genes is needed in future studies, particularly given the evidence from our study that synonymous site selection acted differently in gene categories that reflect gene expression regulation (76).

We quantified the impact of negative and positive selection on eGenes and non-eGenes, finding that eGenes had a significantly higher proportion of nearly neutral nonsynonymous mutations compared to non-eGenes (Figure S7.4D, *SI Appendix Table 2*), suggesting weaker negative selection on nonsynonymous sites for eGenes. Moreover, there was a significantly lower proportion of adaptive nonsynonymous substitutions and a lower rate of positive selection among eGenes than non-eGenes (Figure S7.4D). Similar, but weaker, selection patterns were also found in non-coding regions after controlling for synonymous site selection, but not in 1kbp up- and downstream regions of these genes (Figure S7.5-S7.9; *SI Appendix Table 2*). In agreement with a recent study in *Capsella grandiflora* (82), our findings support the view that eGenes are under weaker negative selection and undergo less frequency positive selection compared to non-eGenes.

As gene expression level and variance are also important determinants of gene evolution in many species(83 – 85), we tested for whether the strength of selection differs between different expression categories. Consistent with many other studies (83 – 85), our results showed that highly expressed genes have a significantly smaller proportion of nearly neutral mutations (23.4%) compared to lowly expressed genes (32.8%), and are subject to stronger negative selection (Figure S7.4E, *SI Appendix Table 2*). In addition, there was higher proportion of adaptive nonsynonymous substitutions in high

expression genes, and the pattern became stronger after accounting for the synonymous site selection differences between highly and lowly expressed genes. In comparison, patterns of selection were inconsistent for non-coding regions between original and control datasets (Figure S7.5-S7.9, *SI Appendix Table 2*), likely indicating that patterns of selection at synonymous sites and the relative strength of selection at synonymous and noncoding sites also differed for genes with different expression levels. Different from the distinct differences on selection pattern between high and low expression genes, the selection difference between genes with different expression variance was subtle (Figure S7.4F, Figure S7.5-S7.9, *SI Appendix Table 2*). Genes with high expression variance were subject to slightly weaker, although significant, negative and positive selection compared to genes with low expression variance. The pattern of positive selection changed after accounting for the synonymous site selection, which could partially result from a lack of power to account for synonymous site selection when the selection patterns differ only slightly between gene categories.

Taken together, our results demonstrate that selection strength varied between genes of different functional importance and expression profiles, and future studies should explore the underlying driving evolutionary forces in greater detail across a wider range of plant species (61).

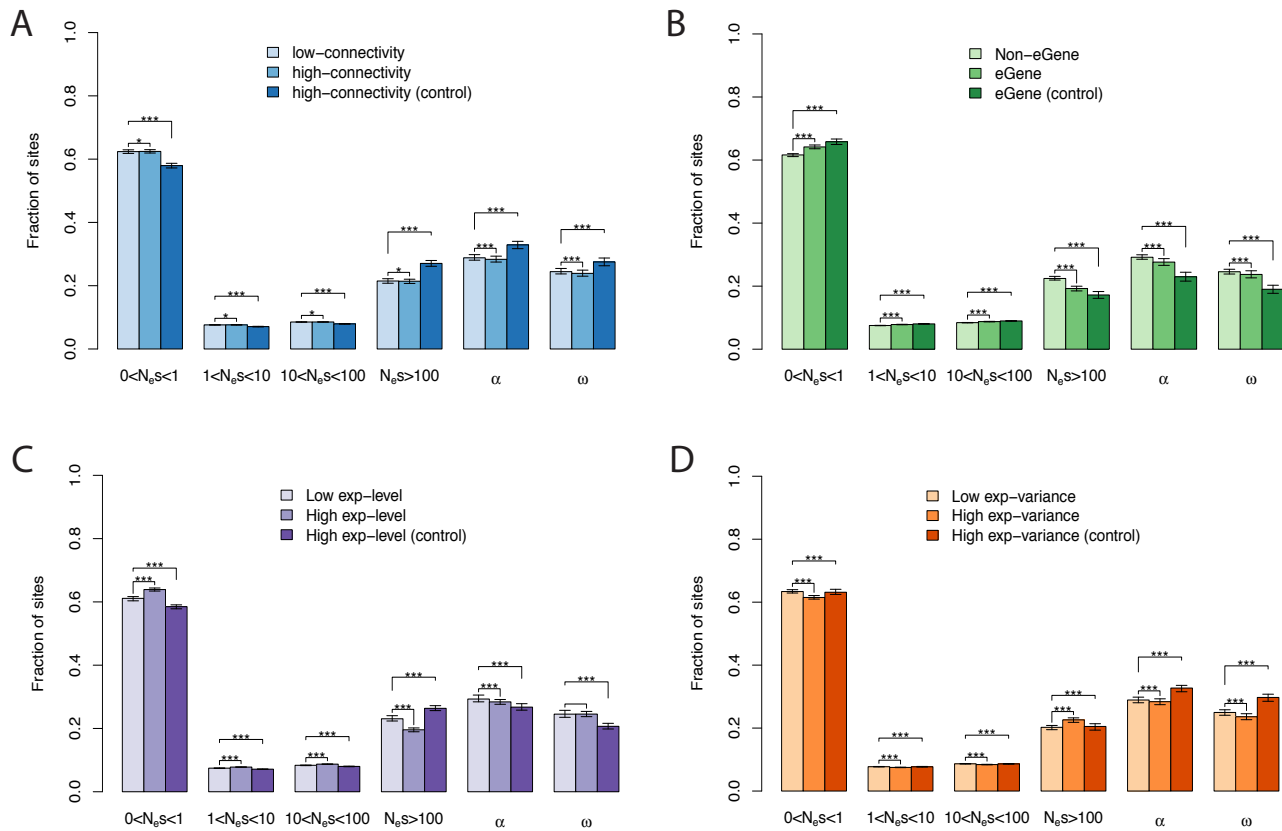


Figure S7.5 Estimates of negative and positive selection on intronic sites in genes **(A)** with varying connectivity level in co-expression network; **(B)** with eQTLs (eGene) or not (non-eGene); **(C)** with varying expression level; **(D)** with varying expression variance. For gene categories of high connectivity, eGene, high expression level and high expression variance, the control group is to use the same neutral reference (4-fold synonymous sites from the corresponding “low” gene categories) in order to control for the synonymous selection difference between these two categories of genes. N_e s categories, different bin of negative selection strength; α , the proportion of divergent sites fixed by positive selection; ω , the rate of adaptive substitution relative to neutral divergence.

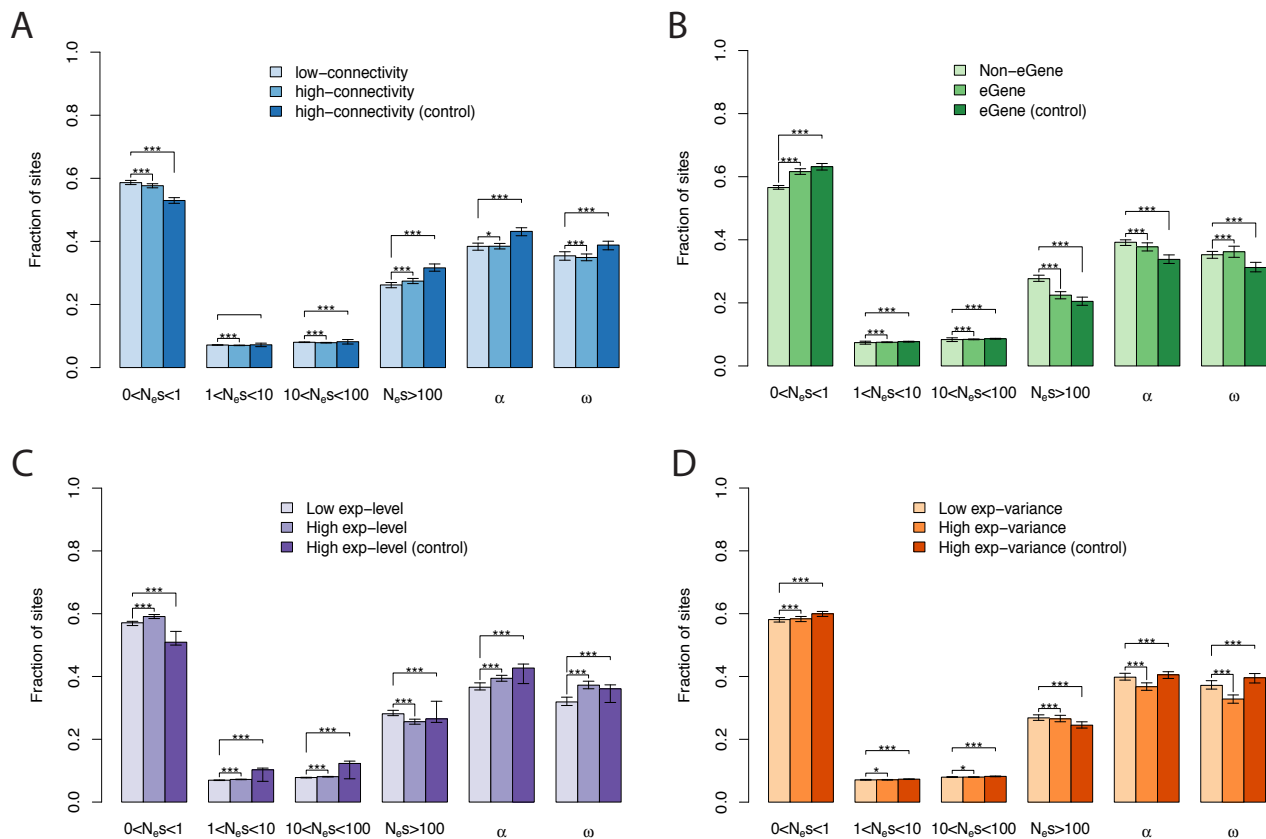


Figure S7.6 Estimates of negative and positive selection on 5' UTR sites in genes (A) with varying connectivity level in co-expression network; (B) with eQTLs (eGene) or not (non-eGene); (C) with varying expression level; (D) with varying expression variance. For gene categories of high connectivity, eGene, high expression level and high expression variance, the control group is to use the same neutral reference (4-fold synonymous sites from the corresponding "low" gene categories) in order to control for the synonymous selection difference between these two categories of genes. N_e s categories, different bin of negative selection strength; α , the proportion of divergent sites fixed by positive selection; ω , the rate of adaptive substitution relative to neutral divergence.

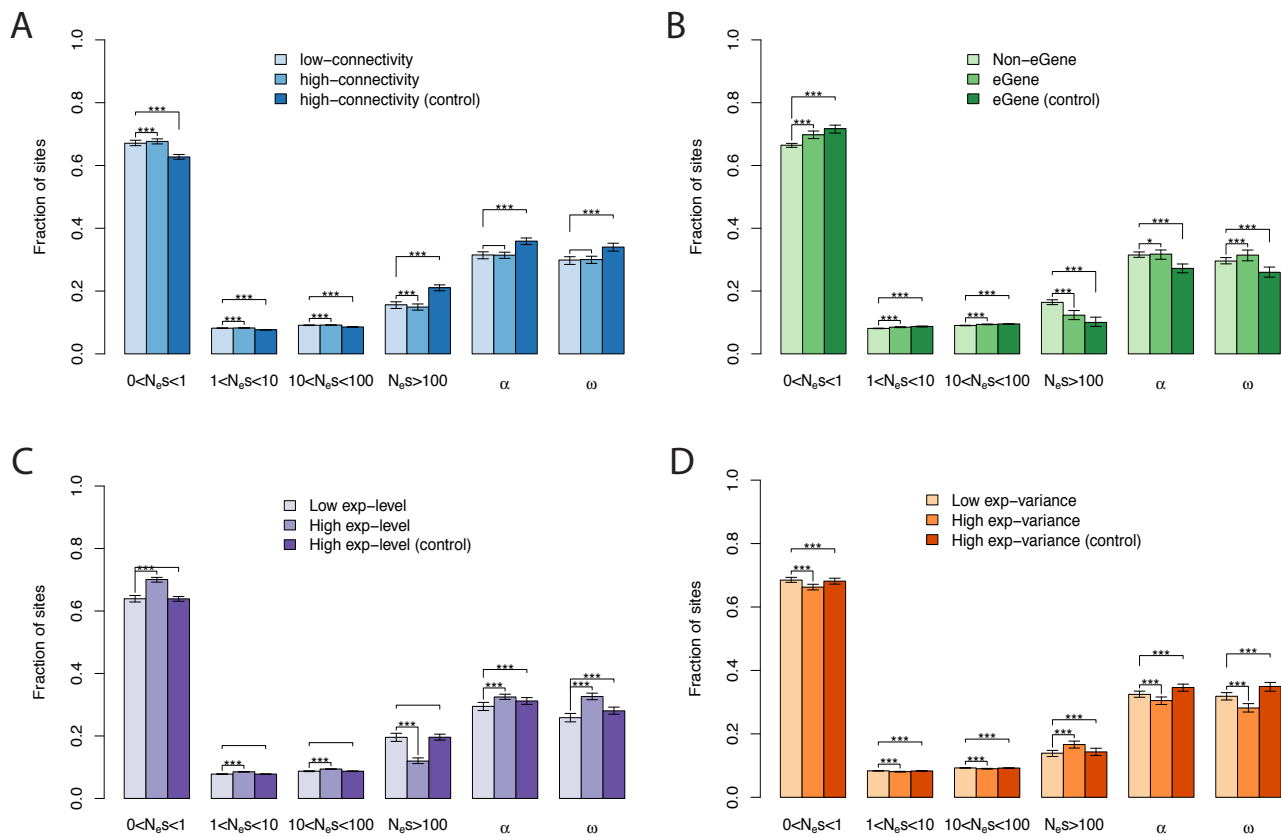


Figure S7.7 Estimates of negative and positive selection on 3' UTR sites in genes (A) with varying connectivity level in co-expression network; (B) with eQTLs (eGene) or not (non-eGene); (C) with varying expression level; (D) with varying expression variance. For gene categories of high connectivity, eGene, high expression level and high expression variance, the control group is to use the same neutral reference (4-fold synonymous sites from the corresponding "low" gene categories) in order to control for the synonymous selection difference between these two categories of genes. N_e s categories, different bin of negative selection strength; α , the proportion of divergent sites fixed by positive selection; ω , the rate of adaptive substitution relative to neutral divergence.

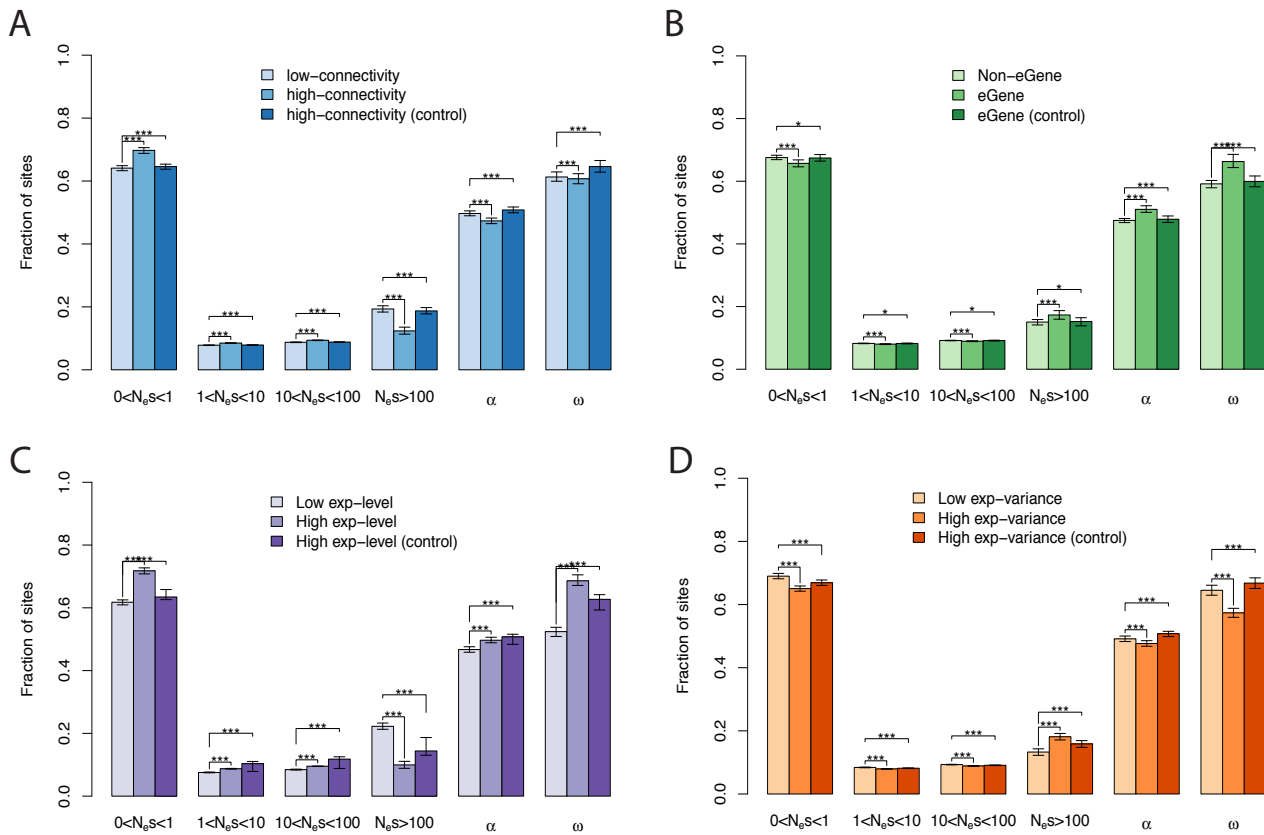


Figure S7.8 Estimates of negative and positive selection on 1kbp upstream sites in genes (A) with varying connectivity level in co-expression network; (B) with eQTLs (eGene) or not (non-eGene); (C) with varying expression level; (D) with varying expression variance. For gene categories of high connectivity, eGene, high expression level and high expression variance, the control group is to use the same neutral reference (4-fold synonymous sites from the corresponding “low” gene categories) in order to control for the synonymous selection difference between these two categories of genes. N_e s categories, different bin of negative selection strength; α , the proportion of divergent sites fixed by positive selection; ω , the rate of adaptive substitution relative to neutral divergence.

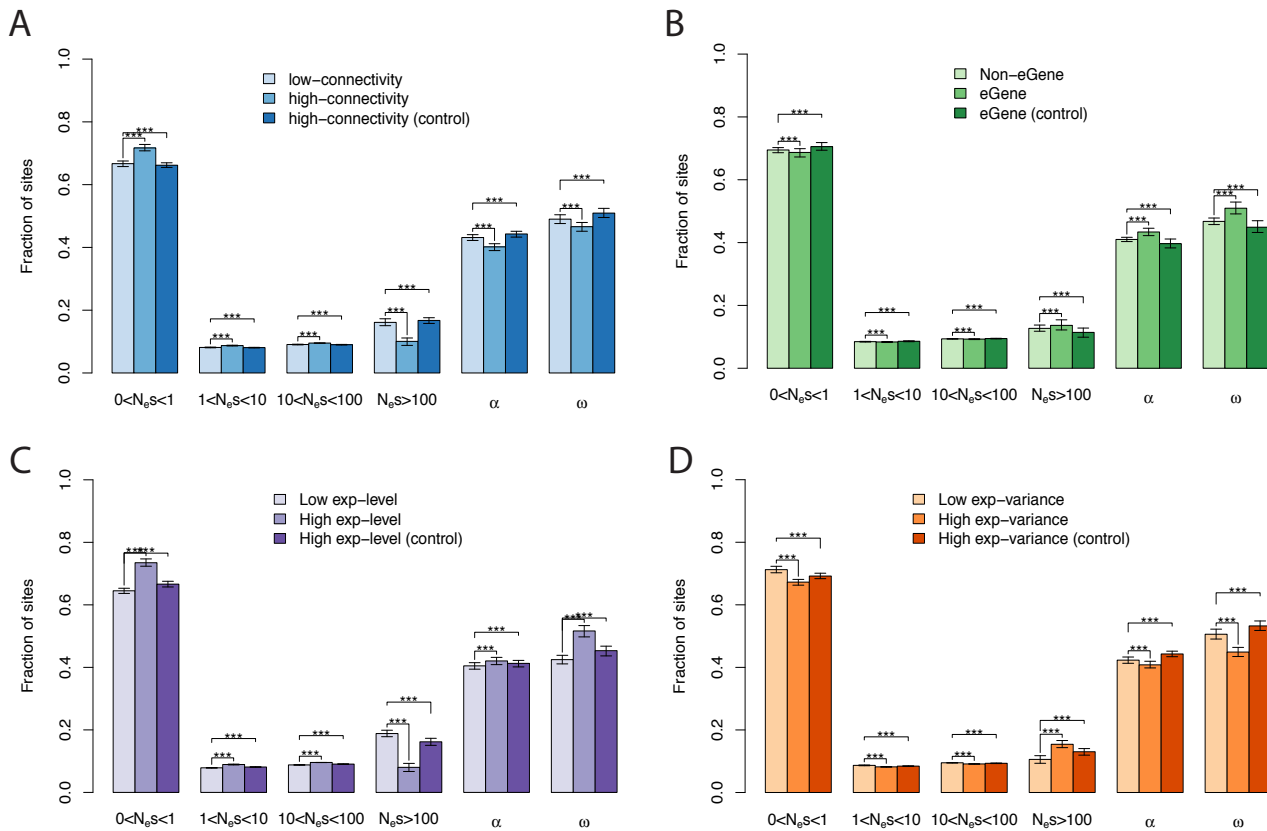


Figure S7.9 Estimates of negative and positive selection on 1 Kbp downstream sites in genes (A) with varying connectivity level in co-expression network; (B) with eQTLs (eGene) or not (non-eGene); (C) with varying expression level; (D) with varying expression variance. For gene categories of high connectivity, eGene, high expression level and high expression variance, the control group is to use the same neutral reference (4-fold synonymous sites from the corresponding “low” gene categories) in order to control for the synonymous selection difference between these two categories of genes. N_e s categories, different bin of negative selection strength; α , the proportion of divergent sites fixed by positive selection; ω , the rate of adaptive substitution relative to neutral divergence.

SnIPRE results

To identify specific genes under positive or negative selection, we applied the robust-to-demography SnIPRE method (Bayesian test) to estimate constraint and selection effects on *Populus* genes (22,306) that have been characterized by various expression features in a previous eQTL study (61). We found that more than 80% of genes (18,612 out of 22,306) evolved under selective constraint (Figure S7.10A), suggesting strong evolutionary constraints acted on non-synonymous variants in the *Populus* genome. The constraint effect on genes varied depending on the pattern of gene expression. Compared to neutrally evolved genes, the selectively constrained genes have significantly higher expression levels and lower expression variance (Figure S7.10B; Table S7.6). In addition, a higher proportion of core genes (hubs in co-

expression network modules) and a lower proportion of genes harbouring regulatory variation (with identified eQTLs) were found among selectively constrained genes (Figure S7.10B; Table S7.6).

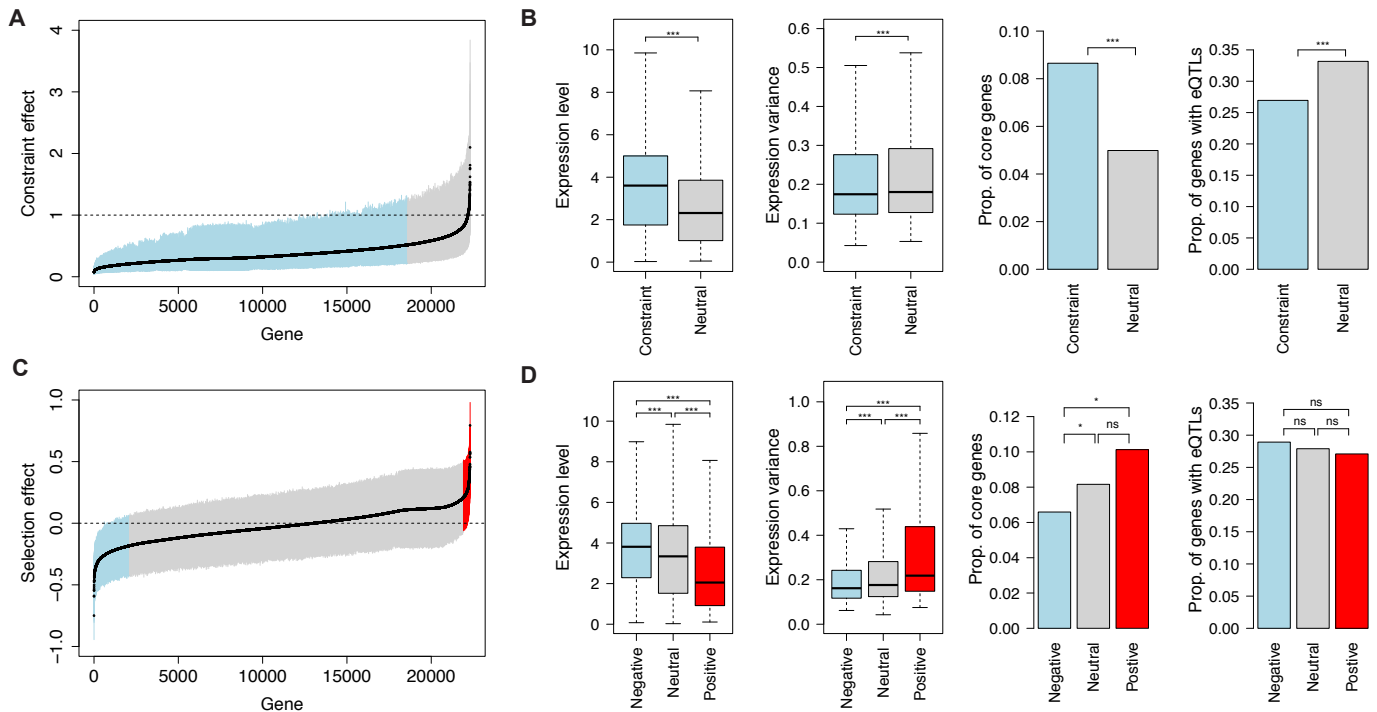


Figure S7.10 Selection and constraint effects in 22,306 expressed genes that have been characterized by various expression features. (A) sizes of constraint effect in expressed genes that are sorted in increasing order. The black line follows the average; the vertical line spans the Bayesian credibility intervals. The dashed line indicates neutrality. Genes with selective constraint effects are marked in blue, and those with neutrality are marked in grey. (B) Characterization of gene expression features (from left to right: the expression level, the expression variance, proportion of genes that are core genes, proportion of genes that harbour eQTLs) for genes evolved under selective constraint (blue) and evolved neutrally (grey). (C) Sizes of selective effects in expressed genes that are sorted in increasing order. The black line follows the average; the vertical line spans the Bayesian credibility intervals. The dashed line indicates neutrality. Genes with negative selection are marked in blue, genes with neutrality are marked in grey, and genes with positive selection are marked in red. (D) Characterization of gene expression features (from left to right: the expression level, the expression variance, proportion of genes that are core genes, proportion of genes that harbour eQTLs) for genes with negative (blue), neutrality (grey) and positive selection (red).

Without making any explicit assumptions about the distribution of fitness effect, we further used SnIPRE to estimate the selection effect on the 22,306 expressed genes as above (Figure S7.10C), where negative and positive values are consistent with negative selection and adaptive protein evolution during the divergence between aspen and poplar

(Potri was used as outgroup), and the magnitude reflects the strength of positive or negative selection. In general, most of the genes appear to be neutrally evolving (89%, 19,802 out of 22,306 genes) and we identify 395 genes under putative positive selection. Functional analysis using the Gene Ontology (GO) database found the genes under positive selection to be significantly enriched in organismal and system developmental processes including signal transduction, cell communication, protein metabolic process, and also stimulus and defense response (Table S7.7). Furthermore, we found that the genes under positive selection showed significantly reduced levels of gene expression and increased gene expression variance, and there is a significantly higher proportion of core-genes among genes under positive selection than genes under negative selection (Figure S7.10D). For the 40 positively selected core genes, we found they were enriched in nonstructural carbohydrate biosynthetic and metabolic process, response to biotic stimulus, nucleosome and chromatin assembly and organization (Table S7.8). In addition, there is no significant difference for genes with eQTL between positively and negatively selected genes (Figure S7.10D; Table S7.6).

Table S7.6 Summary of the correlation coefficient (Spearman's rank correlation coefficient) between sizes of constraint and selection effect and gene-expression related measures.

	Constraint effect		Selection effect	
	Pairwise	Partial	Pairwise	Partial
Expression level	-0.2266***	-0.1916***	-0.2318***	-0.2283***
Expression variance	-0.0075	-0.0068	0.1254***	0.1225***
Connectivity	-0.1543***	-0.0560***	-0.0324***	0.0368***
core vs Non-core	-0.0787***	-0.0163*	0.0057	-0.0042
eGene vs. Non-eGene	0.0502***	0.0417***	-0.0250***	-0.0125

* $P < 0.05$

** $P < 0.01$

*** $P < 0.001$

Table S7.7 Enriched Gene Ontology (GO) categories among the genes identified by SnIPRE as being under positive selection.

GO.ID	Term	Annotated	Significant	Expected	Fisher.p
GO:0007165	signal transduction	516	15	4.89	0.00010
GO:0023052	signaling	516	15	4.89	0.00010
GO:0007154	cell communication	619	16	5.86	0.00023
GO:0050896	response to stimulus	1050	22	9.94	0.00029
GO:0051716	cellular response to stimulus	658	16	6.23	0.00045
GO:0019538	protein metabolic process	2936	44	27.8	0.00046
GO:0006468	protein phosphorylation	1394	26	13.2	0.00047
GO:0044267	cellular protein metabolic process	2498	39	23.65	0.00052
GO:0016310	phosphorylation	1531	26	14.5	0.00190
GO:0043170	macromolecule metabolic process	4810	61	45.54	0.00196

GO:0044260	cellular macromolecule metabolic process	4341	56	41.1	0.00248
GO:0006952	defense response	64	4	0.61	0.00312
GO:0050789	regulation of biological process	1872	29	17.72	0.00398
GO:0050794	regulation of cellular process	1819	28	17.22	0.00519
GO:0065007	biological regulation	1919	29	18.17	0.00571
GO:0070588	calcium ion transmembrane transport	19	2	0.18	0.01368
GO:1901564	organonitrogen compound metabolic process	3668	46	34.73	0.01462
GO:0016598	protein arginylation	2	1	0.02	0.01885
GO:0006796	phosphate-containing compound metabolic process	1930	27	18.27	0.02078
GO:0006793	phosphorus metabolic process	1937	27	18.34	0.02170
GO:0017148	negative regulation of translation	3	1	0.03	0.02814
GO:0032269	negative regulation of cellular protein metabolic process	3	1	0.03	0.02814
GO:0034249	negative regulation of cellular amide metabolic process	3	1	0.03	0.02814
GO:0051248	negative regulation of protein metabolic process	3	1	0.03	0.02814
GO:0006816	calcium ion transport	28	2	0.27	0.02861
GO:0006807	nitrogen compound metabolic process	5457	62	51.67	0.02964
GO:0000160	phosphorelay signal transduction system	85	3	0.8	0.04665

Table S7.8 Enriched Gene Ontology (GO) categories among the core genes (hubs in co-expression network modules) identified by SnIPRE as being under positive selection.

GO.ID	Term	Annotated	Significant	Expected	Fisher.p
GO:0006006	glucose metabolic process	22	1	0.03	0.026
GO:0005992	trehalose biosynthetic process	23	1	0.03	0.027
GO:0005991	trehalose metabolic process	25	1	0.03	0.030
GO:0046351	disaccharide biosynthetic process	26	1	0.03	0.031
GO:0009607	response to biotic stimulus	27	1	0.03	0.032
GO:0009312	oligosaccharide biosynthetic process	30	1	0.04	0.036
GO:0006334	nucleosome assembly	38	1	0.05	0.045
GO:0031497	chromatin assembly	38	1	0.05	0.045
GO:0034728	nucleosome organization	38	1	0.05	0.045
GO:0006323	DNA packaging	40	1	0.05	0.047
GO:0006333	chromatin assembly or disassembly	40	1	0.05	0.047
GO:0065004	protein-DNA complex assembly	40	1	0.05	0.047
GO:0071824	protein-DNA complex subunit organization	40	1	0.05	0.047

8 Structural variation

8.1 Methods

Methods for detecting structural variants (SVs), such as micro-INDELS and larger variants as well as copy number variants, from short read re-sequencing samples are generally associated with a high level of both false positives and false negatives. To obtain as accurate variant calls as possible, we ran a number of different methods and combined the output. Variants were detected for the 24 individuals of Potra and 16 individuals of Potrs that were mapped both to their respective references but also to the Potri reference and also analysed for the presence of SNPs (7.3.2). A number of methods with different approaches to detect variants were used. Samtools (ver 0.1.19)(15) and Varscan (ver 2.3.7)(86) both detect short insertions and deletions within mapped reads. ControlFREC (ver 6.7)(87, 88) detects copy number variants based on sequence depth. Breakdancer (ver 1.4.5)(89) and Delly (ver 0.6.3)(90) detects SVs based on altered distance between mapped reads from PE data. Lumpy (ver 0.2.12)(91) integrates multiple SV signals jointly across multiple samples. All methods were run with default parameters. To increase sensitivity of the SV analyses we only included scaffolds greater than 2000 bp in further analysis.

We ran all methods for all individuals separately. The output from the different methods were combined using an in-house Perl script available through git. Only variants detected by at least two different methods were kept for further analyses. Based on these, we generated a complete set of reliable variants by combining all variants and scoring all individuals for the presence or absence of the variants in this set. Variants were also annotated using ANNOVAR (92) to assign them to genomic context and, where applicable, to associate them with an effect on protein coding sequence.

8.2 Results

Using the Potri reference to align reads from all individuals of Potra and Potrs, we identified 13,468 and 211 polymorphic deletions and insertions in Potra, respectively, and 11,004 and 115 polymorphic deletions and insertions in Potrs, respectively. Of these 1.0% and 0.9% are larger than 100 bp and 7,741 insertions and 70 deletions were in common to both aspens. The higher identification of deletions is most likely not a true biological signature, but rather reflects that the methods used for INDEL detection are better able to call deletions than insertions. As the majority of INDELS relative to Potri were in common among the two aspen species, this suggests that they occurred before the speciation event separating them.

When INDELS were called for Potra and Potrs using reads mapped to their respective species-specific assembly, we detected many more SVs: 67,2106 deletions and 58,8801 insertions for Potra and 48,9935 deletions and 42,0754 insertions for Potrs. The vast majority of INDELS identified from the Potra and Potrs assembly were shorter than 100 bp (median size of deletions and insertions were 2 bp and 1 bp for both Potra and Potrs) and only 0.45% and 0% of INDELS

were longer than 100 bp in Potra and Potrs, respectively (Figures S8.1 and S8.2). The inability to detect longer structural variants likely reflects the fragmented nature of these two assemblies as the presence of longer INDELS were observed when using Potri as a reference (see the preceding section).

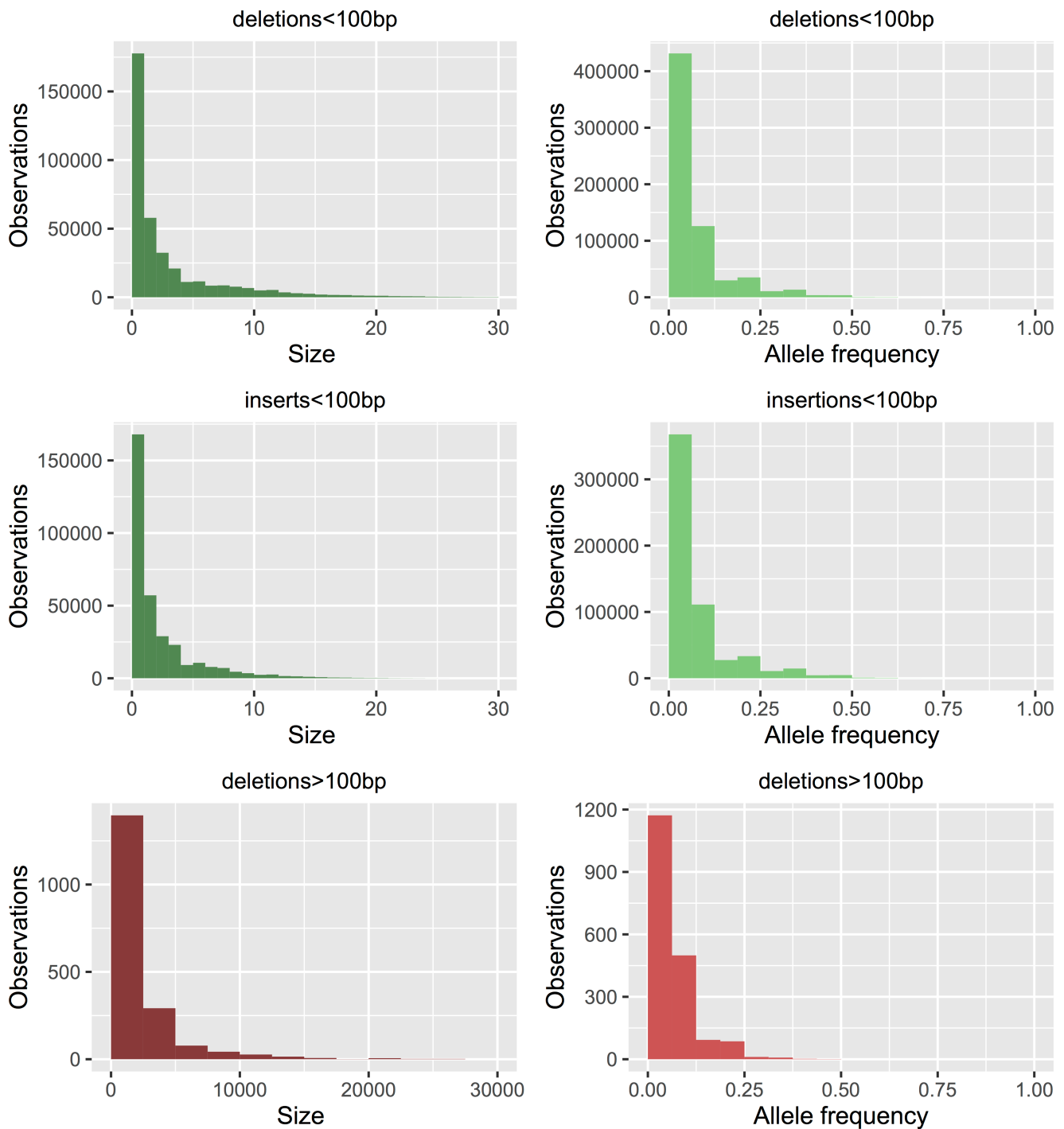


Figure S8.1 Size frequency distribution (left column) and site frequency spectrum (right column) of SVs in a sample of 24 Potra individuals, with reads mapped against the Potra assembly. No insertions >100 bp were identified.

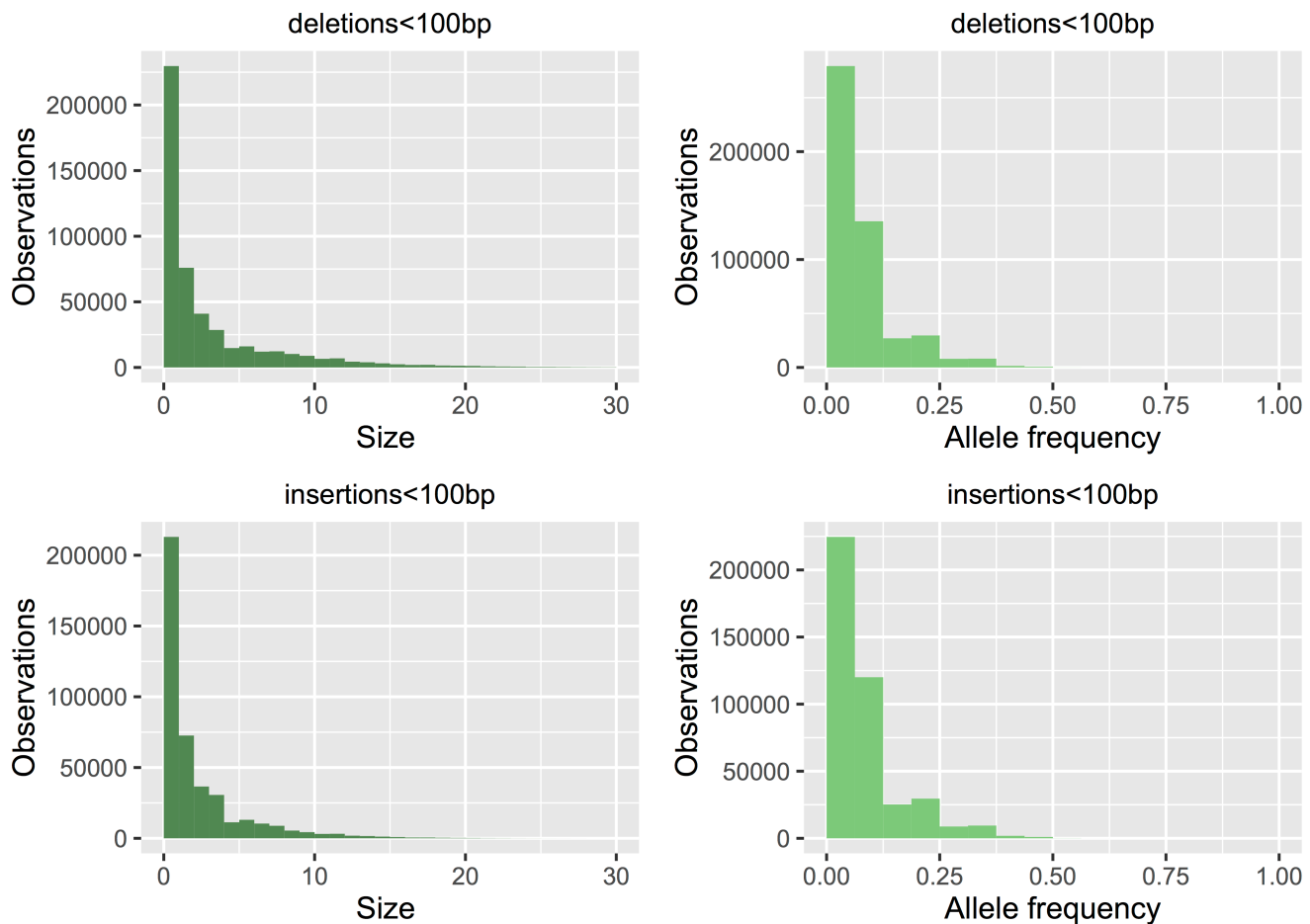


Figure S8.2 Size frequency distribution (left column) and site frequency spectrum (right column) of INDELs in a sample of 16 Potrs individuals, with reads mapped against the Potrs assembly. The analyses did not identify any INDELs larger than 100 bp.

We also used data on the occurrence of INDELs among the re-sequenced individuals for Potra (24 inds) and Potrs (16 inds) to calculate the frequency spectrum of segregating INDELs (right columns in Figures S8.1 and S8.2). Most INDELs were found to be rare, segregating at a mean frequency of 7% and 13% in the two species, respectively. We summarised the frequency spectrum for SVs of different sizes by calculating Tajima's D (Figure S8.3), showing that larger SVs are generally segregating at lower frequencies (more negative Tajima's D).

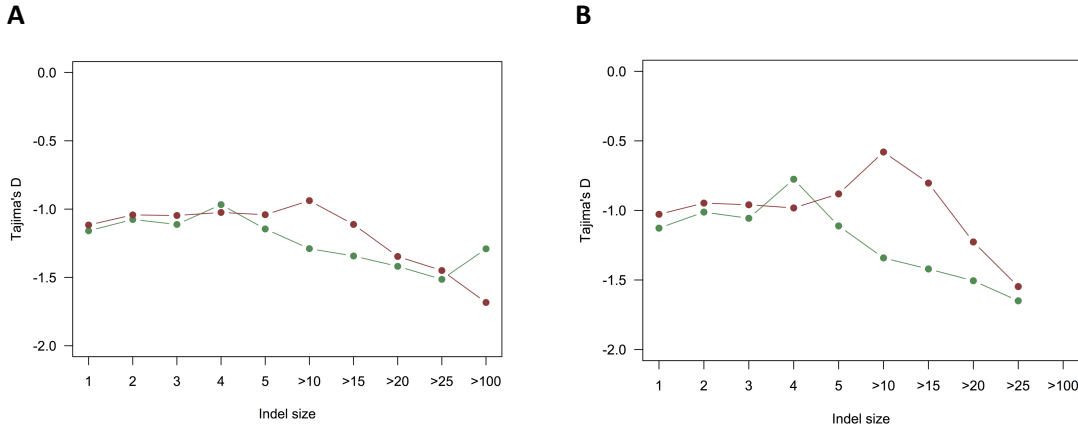


Figure 8.3 Tajima's D for SV of different sizes for (A) Potra and (B) Potrs. Data for deletions (green) and insertions (red) are shown separately for both species.

Variants were annotated using annovar and the results are summarized in Tables S8.1 – S8.2. More than 50% of INDELS were located in intergenic regions in both species, and about 30% were located in upstream, downstream and intronic regions (Tables S8.1 and S8.2). For shorter INDELS (<100bp) ~1-2% were located in exons and an additional 4-10% were located in UTRs. For longer INDELS (>100bp) a substantially greater fraction (15.2%) affect exonic regions.

Table S8.1 Summary of annotated effects of INDELS in Potra.

Type	Deletions<100bp				Deletions>100bp				Insertions<100bp			
	Numbers	Proportion	Median size (bp)	Mean size (bp)	Numbers	Proportion	Median size (bp)	Mean size (bp)	Numbers	Proportion	Median size (bp)	Mean size (bp)
UTR3	41632	0.062	2	4	178	0.058	1170	1750	33511	0.056	2	4
intergenic	387769	0.580	2	4	1446	0.477	1262	1907	349149	0.590	1	3
splicing	471	0.001	2	5	-	-	-	-	493	0.001	2	4
upstream	59274	0.089	2	4	350	0.116	1147	1783	54151	0.092	2	3
upstream;downstream	7389	0.011	2	4	47	0.016	1074	2088	6438	0.011	2	4
exonic	13004	0.019	3	6	461	0.152	2970	5259	9474	0.016	3	5
downstream	51446	0.077	2	4	280	0.092	1273	1951	45105	0.076	2	3
UTR5	23815	0.036	2	4	61	0.020	1210	2149	21432	0.036	2	4
intronic	83996	0.126	2	4	203	0.067	1064	1638	68839	0.117	2	3
UTR5;UTR3	281	0.000	2	5	-	-	-	-	193	0.000	2	4

Table S8.2 Summary of annotated effects of INDELS in Potrs

Type	Deletions<100bp				Deletions>100bp				Insertions<100bp			
	Numbers	Proportion	Median size (bp)	Mean size (bp)	Numbers	Proportion	Median size (bp)	Mean size (bp)	Numbers	Proportion	Median size (bp)	Mean size (bp)
UTR3	12467	0.025	2	4	-	-	-	-	10260	0.024	2	3
intergenic	303537	0.620	2	4	-	-	-	-	264442	0.628	1	3
splicing	346	0.001	3	5	-	-	-	-	331	0.001	3	5
upstream	49986	0.102	2	4	-	-	-	-	44252	0.105	2	3
upstream;downstream	4733	0.010	2	4	-	-	-	-	3935	0.009	2	4
exonic	5656	0.012	3	6	-	-	-	-	4472	0.011	3	5
downstream	41199	0.084	2	4	-	-	-	-	35109	0.083	2	3
UTR5	6661	0.014	2	4	-	-	-	-	5926	0.014	2	4
intronic	65223	0.133	2	4	-	-	-	-	51932	0.123	2	3
UTR5;UTR3	127	0.000	2	4	-	-	-	-	95	0.000	2	3

To further assess the genome-wide distribution of SVs in Potra, SV locations were mapped onto Potra pseudo-chromosomes (see section 5.5 above) and the number of structural variants were recorded in windows of 500 Kbp across the genome for small (1-9 bp) and large (>10 bp) insertions and deletions separately (Figure S8.4).

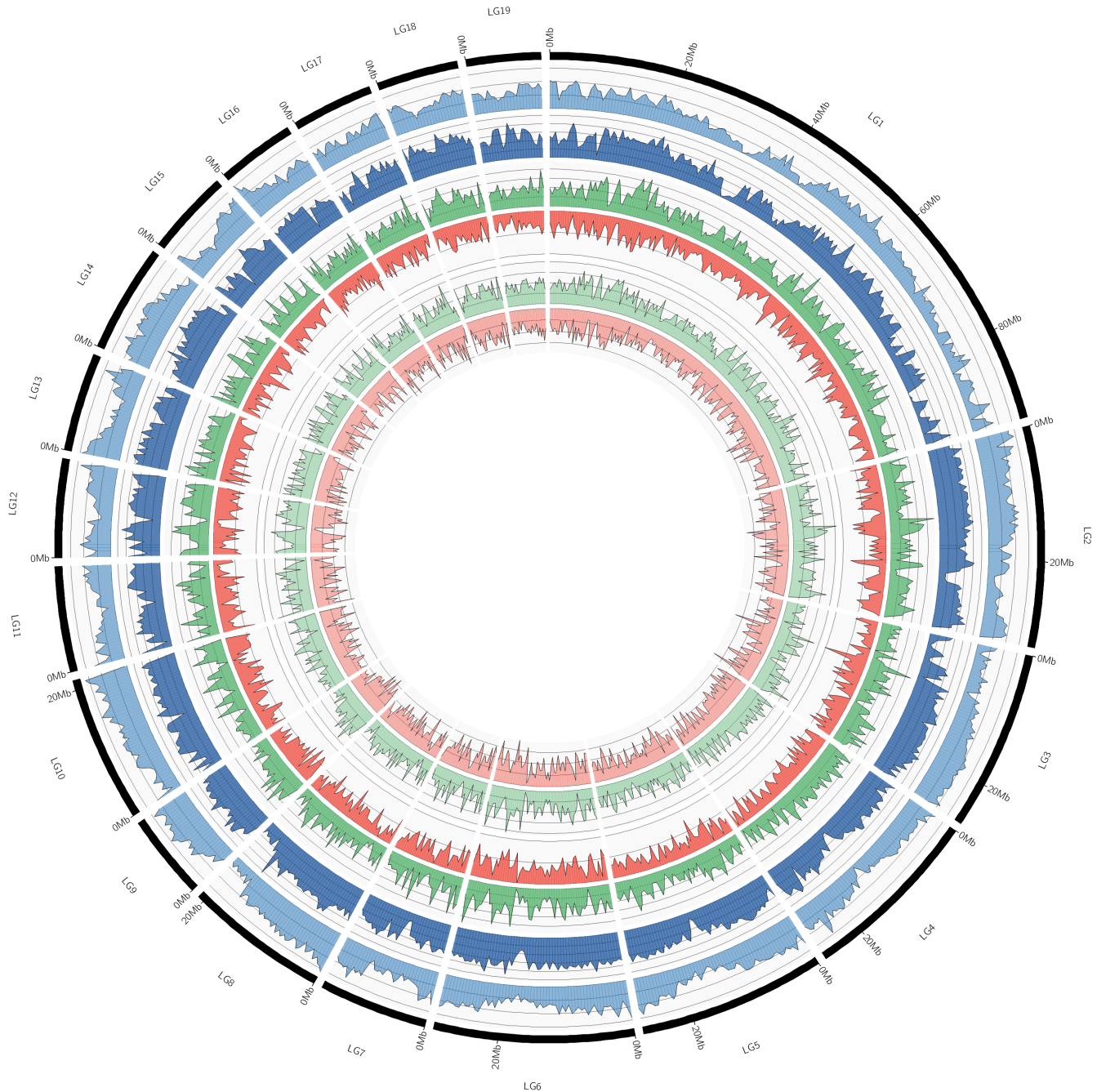


Figure S8.4 Genomic distribution of SVs across the Potra genome. Gene density (light blue track, y axis range 0-75), nucleotide diversity (dark blue, y axis range 0-6 ($\times 10^{-3}$)), deletions longer than 10bp (dark green track, y axis range 0-130), insertions longer than 10bp (dark red track, y axis range 0-55), deletions 2-9bp (light green track, y axis range 0-440), insertions 2-9bp (light red track, y axis range 0-390)

The fraction of long indels (> 100 bp) found in coding regions was higher than for short indels (<= 100 bp; 17.5% vs 1.70%). Correlating the INDELS against the expression of the gene that they are found (98 cases) show that they have a slightly wider distribution than flanking INDELS, *i.e.* INDELS that overlap with the 2 Kbp upstream/downstream regions of the gene (Fig S8.5). 1791/3032 long INDELS and 510640/1257875 short INDELS were associated with gene features, representing 2 Kbp upstream/downstream flanking regions, UTRs, coding exons, introns (Figure S8.5).

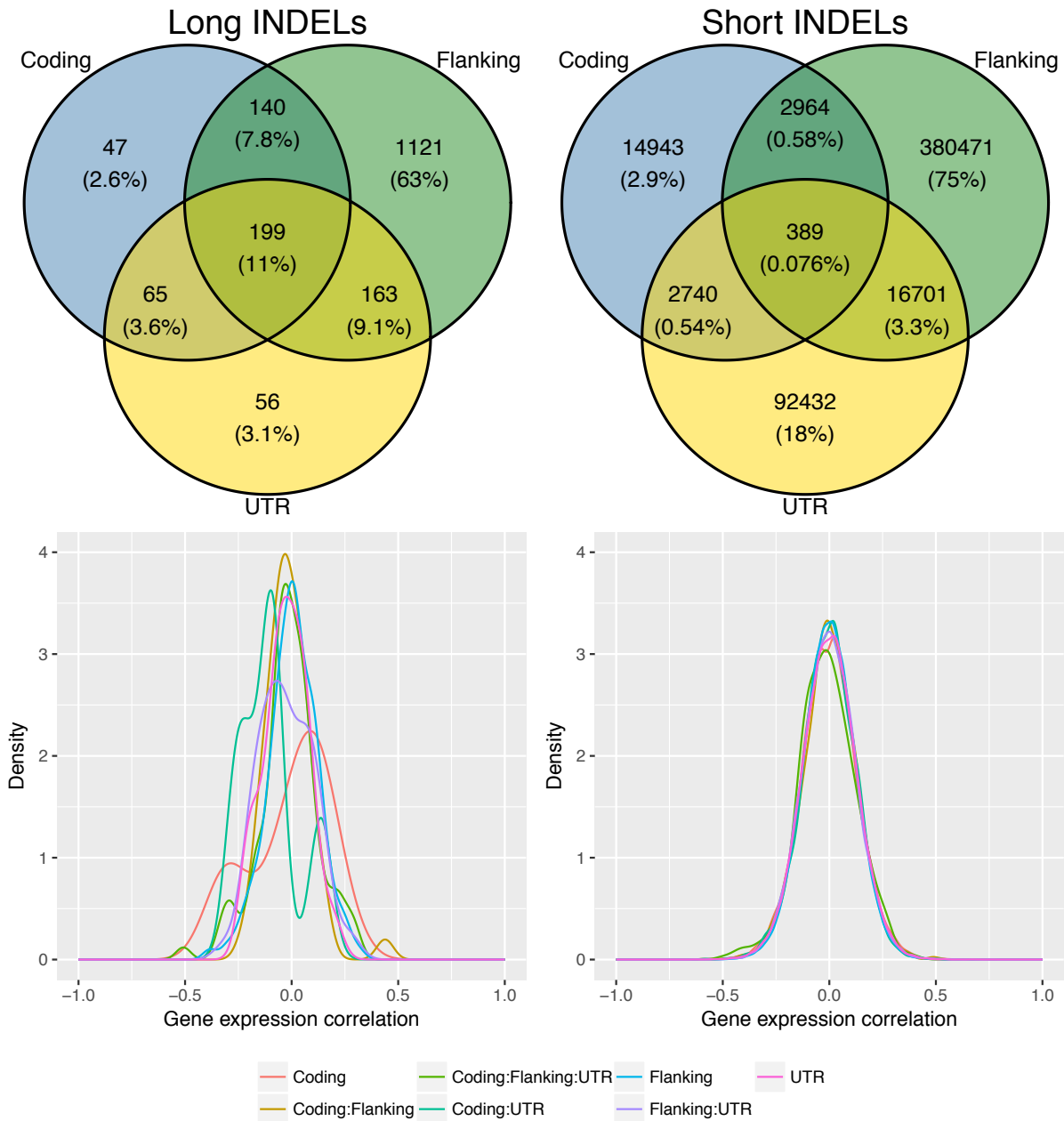


Figure S8.5 Venn diagram representations of the generic region contexts of long and short INDELs (above) and associated density distributions of the correlation between INDELs and the expression of the gene that they affect (below).

Even though a larger fraction of long INDELs affected coding regions compared to short INDELs, most of the long INDELs were non-coding (Figure S8.6). The total number of coding effects exceed the total number of coding INDELs due to INDELs with multiple effects being counted multiple times.

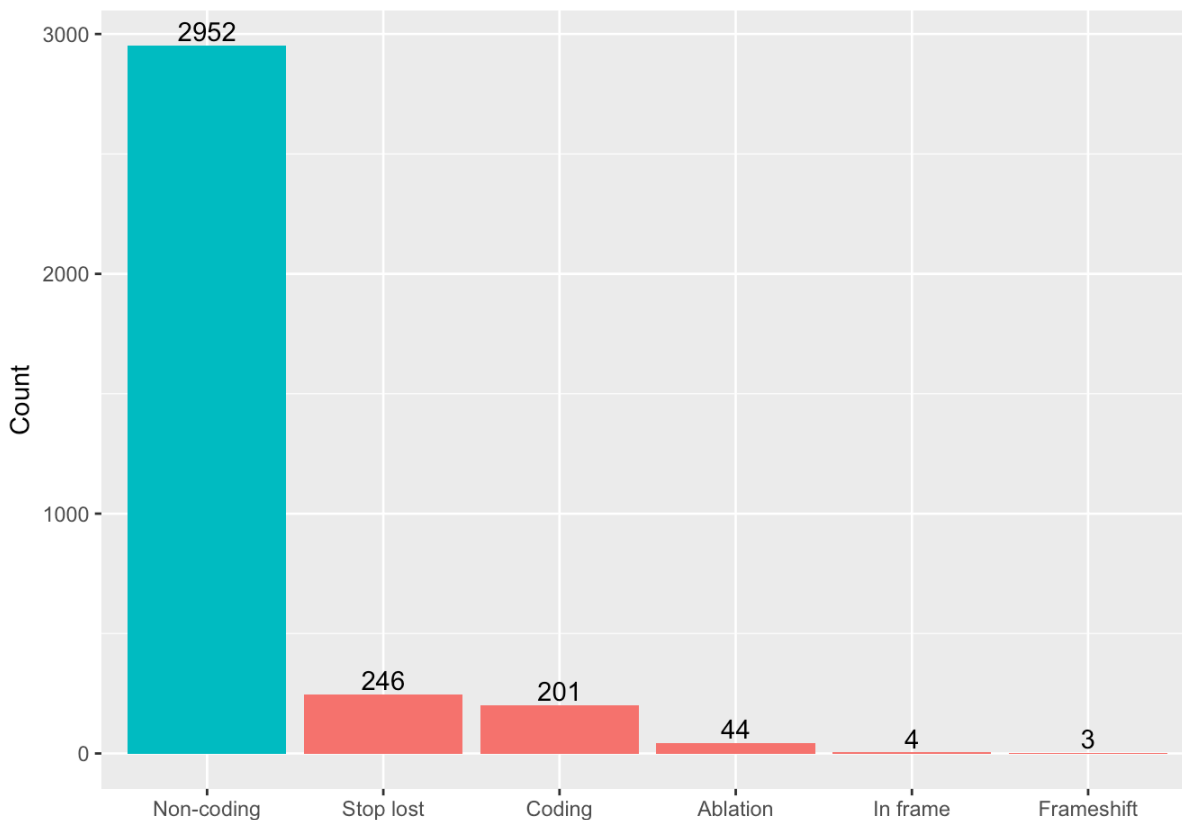


Figure S8.6 Number of long INDELs causing different classes of mutation effects. Where a single INDEL spans multiple genomic contexts, all contexts are counted. As such the total number of effects is greater than the number of INDELs.

References

1. Wang J, Street NRNR, Scofield DGDG, Ingvarsson PKPK (2016) Variation in Linked Selection and Recombination Drive Genomic Divergence during Allopatric Speciation of European and American Aspens. *Mol Biol Evol* 33(7):1754–1767.
2. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
3. Toribio AL, et al. (2017) European Nucleotide Archive in 2016. *Nucleic Acids Res* 45(D1):D32–D36.
4. Pellicer J, Leitch IJ (2014) The Application of Flow Cytometry for Estimating Genome Size and Ploidy Level in Plants (Humana Press, Totowa, NJ), pp 279–307.
5. Tuskan GA, et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313(5793):1596–604.
6. Dolezel J, Sgorbati S, Lucretti S (1992) Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol Plant* 85(4):625–631.
7. Dolezel J, Bartos J, Voglmayr H, Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. *Cytometry A* 51(2):127–8; author reply 129.
8. Dpooležel J, Binarová P, Lcretti S (1989) Analysis of Nuclear DNA content in plant cells by Flow cytometry. *Biol Plant* 31(2):113–120.
9. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764–770.
10. Chikhi R, Medvedev P (2013) Informed and Automated k-Mer Size Selection for Genome Assembly.
11. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
12. Dobin A, et al. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.
13. Simpson JT, et al. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19(6):1117–1123.
14. Vicedomini R, Vezzi F, Scalabrin S, Arvestad L, Policriti A (2013) GAM-NGS: genomic assemblies merger for next generation sequencing. *BMC Bioinformatics* 14(Suppl 7):S6.
15. Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
16. Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L (2014) BESST--efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* 15(1):281.
17. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859–1875.
18. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
19. McKenna A, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation

- DNA sequencing data. *Genome Res* 20(9):1297–1303.
20. Camacho C, et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10(1). doi:10.1186/1471-2105-10-421.
 21. Vezzi F, Narzisi G, Mishra B (2012) Reevaluating assembly evaluations with feature response curves: GAGE and assemblathons. *PLoS One* 7(12):e52210.
 22. Price AL, Jones NC, Pevzner PA (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics* 21(Suppl 1):i351–i358.
 23. Huang X, Madan A (1999) CAP3: A DNA Sequence Assembly Program. *Genome Res* 9(9):868–877.
 24. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1–4):462–467.
 25. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
 26. Tamura K, Stecher G, Peterson D, FilipSKI A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30(12):2725–9.
 27. Campbell MS, et al. (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol* 164(2):513–24.
 28. Haas BJ (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31(19):5654–5666.
 29. Schneider M, et al. (2009) The UniProtKB/Swiss-Prot knowledgebase and its Plant Proteome Annotation Program. *J Proteomics* 72(3):567–573.
 30. Dong Q, Schlueter SD, Brendel V (2004) PlantGDB, plant genome database and analysis tools. *Nucl Acids Res* 32(suppl_1):D354-359.
 31. Smit A, Hubley R, Green P (1996) RepeatMasker Open-3.0. <http://www.repeatmasker.org>.
 32. Cantarel BL, et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18(1):188–196.
 33. Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* 33(suppl 2):W465–W467.
 34. Zdobnov EM, Apweiler R (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9):847–848.
 35. Sundell D, et al. (2015) The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *New Phytol* 208(4):1149–1156.
 36. Robinson KMKM, et al. (2014) *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. *BMC Plant Biol* 14(1):276.
 37. Haas BJ, et al. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for

- reference generation and analysis. *Nat Protoc* 8(8):1494–1512.
38. Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27(17):2325–2329.
 39. Gremme G, Steinbiss S, Kurtz S (2013) GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Trans Comput Biol Bioinforma* 10(3):645–656.
 40. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061–7.
 41. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
 42. Veeckman E, Ruttink T, Vandepoele K (2016) Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. *Plant Cell* 28(8):1759–68.
 43. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
 44. Delhomme N, et al. (2015) Guidelines for RNA-Seq data analysis (prot 67). *Epigenesis*.
 45. Delhomme N, Padioleau I, Furlong EE, Steinmetz LM (2012) easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* 28(19):2532–3.
 46. Delhomme N, et al. (2015) Serendipitous Meta-Transcriptomics: The Fungal Community of Norway Spruce (*Picea abies*). *PLoS One* 10(9):e0139080.
 47. Ashburner M, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25–29.
 48. Finn R, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38(suppl 1):D211–D222.
 49. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28(1):27–30.
 50. Lamesch P, et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40(D1):D1202–D1210.
 51. Goodstein D, et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40(Database issue):D1178–D1186.
 52. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584.
 53. Proost S, et al. (2012) i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res* 40(2):e11–e11.
 54. Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* 34(Web Server):W609–W612.
 55. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5):725–736.

56. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24(8):1586–1591.
57. Vanneste K, Maere S, Van de Peer Y (2014) Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philos Trans R Soc Lond B Biol Sci* 369(1648):20130353-.
58. Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21(16):3448–3449.
59. Shannon P, et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* 13(11):2498–2504.
60. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B* 57(1):289–300.
61. Mähler N, et al. (2017) Gene co-expression network connectivity is an important determinant of selective constraint. *PLOS Genet* 13(4):e1006402.
62. Kryuchkova-Mostacci N, Robinson-Rechavi M (2016) A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* 18(2):bbw008.
63. Zamani N, et al. (2013) Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics* 14(1):347.
64. Grabherr MG, et al. (2010) Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* 26(9):1145–1151.
65. Sundström G, Zamani N, Grabherr MG, Mauceli E (2015) *Whiteboard*: a framework for the programmatic visualization of complex biological analyses. *Bioinformatics* 31(12):2054–2055.
66. Tuskan GAGA, et al. (2012) The obscure events contributing to the evolution of an incipient sex chromosome in *Populus*: a retrospective working hypothesis. *Tree Genet Genomes* 8(3):559–571.
67. Pakull B, Kersten B, Lüneburg J, Fladung M (2015) A simple PCR-based marker to determine sex in aspen. *Plant Biol* 17(1):256–261.
68. Wang J, Scofield D, Street NR, Ingvarsson PK (2015) Variant Calling Using NGS Data in European Aspen (*Populus tremula*). *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches* (Springer International Publishing, Cham), pp 43–61.
69. Korneliussen T, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15(1):356.
70. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3).
71. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
72. Remington DL, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A* 98(20):11479–84.

73. Keightley PD, Eyre-Walker A (2007) Joint Inference of the Distribution of Fitness Effects of Deleterious Mutations and Population Demography Based on Nucleotide Polymorphism Frequencies. *Genetics* 177(4):2251–2261.
74. Eyre-Walker A, Keightley PD (2009) Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Mol Biol Evol* 26(9):2097–2108.
75. JUKES TH, CANTOR CR (1969) CHAPTER 24 – Evolution of Protein Molecules. *Mammalian Protein Metabolism*, pp 21–132.
76. Ingvarsson PK (2010) Natural Selection on Synonymous and Nonsynonymous Mutations Shapes Patterns of Polymorphism in *Populus tremula*. *Mol Biol Evol* 27(3):650–660.
77. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351(6328):652–654.
78. Eilertson KE, Booth JG, Bustamante CD (2012) SnIPRE: Selection Inference Using a Poisson Random Effects Model. *PLoS Comput Biol* 8(12):e1002806.
79. Wang J, et al. (2018) A major locus controls local adaptation and adaptive life history variation in a perennial plant. *Genome Biol* 19(1):72.
80. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–8.
81. Luisi P, et al. (2015) Recent Positive Selection Has Acted on Genes Encoding Proteins with More Interactions within the Whole Human Interactome. *Genome Biol Evol* 7(4):1141–1154.
82. Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T (2017) Genomic analysis reveals major determinants of cis- regulatory variation in *Capsella grandiflora*. *Proc Natl Acad Sci* 114(5):1087–1092.
83. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158(2):927–31.
84. Gout J-F, Kahn D, Duret L (2010) The Relationship among Gene Expression, the Evolution of Gene Dosage, and the Rate of Protein Evolution. *PLoS Genet* 6(5):e1000944.
85. Williamson RJ, et al. (2014) Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding Regions of *Capsella grandiflora*. *PLoS Genet* 10(9):e1004622.
86. Koboldt DC, et al. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25(17):2283–2285.
87. Boeva V, et al. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28(3):423–425.
88. Boeva V, et al. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 27(2):268–269.
89. Chen K, et al. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6(9):677–681.
90. Rausch T, et al. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis.

Bioinformatics 28(18):i333–i339.

91. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 15(6):R84.
92. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164–e164.