Supporting Information Methods S1 and Notes S1

Article title: The Plant Genome Integrative Explorer Resource: PlantGenIE.org

Authors: David Sundell, Chanaka Mannapperuma, Sergiu Netotea, Nicolas Delhomme, Yao-Cheng Lin, Andreas Sjödin, Yves Van de Peer, Stefan Jansson, Torgeir R. Hvidsten and Nathaniel R. Street

Article acceptance date: 8 June 2015

Notes S1 An overview of PlantGenIE.org tools and associated implementation details.

Table 1 An overview of PlantGenIE.org tools and associated implementation details

Tool	Available in	Function	Language(s)	Database	db Table	Tool Source	Link
Drupal CMS	P, C, A	Framework	PHP, JS, HTML	PostgreSQL	Drupal DB	Drupal	https://www.drupal.org/
exHeatmap	P, C, A	Expression HeatMap	Python, PHP, JS, HTML	MySQL	P/C/A-db1	PlantGenIE	
exImage	P, C, A	Pictographic view of expression	PHP, Perl, JS, HTML	MySQL	P/C/A-db1	PlantGenIE	
exPlot	P, C, A	Expression Profiles	PHP, Perl, JS, HTML	MySQL	P/C/A-db1	PlantGenIE	
exNet	P, C, A	Expression Network visualization	Python, PHP, JS, HTML	MySQL	P/C/A-db1	PlantGenIE	
Chromosome Diagram	P	Plots gene location	PHP, JS, HTML	MySQL	P/C/A-db2	PlantGenIE	
JBrowse	P,C	Genome Viewer	PHP, JS, HTML	MySQL	JBrowse DB	GMOD JBrowse	http://gmod.org/wiki/JBrowse
GBrowse	P, C, A	Genome Viewer	Perl, JS, HTML	MySQL	GBrowse DB	GBrowse	
BLAST	P, C, A	Sequence homology search	PHP, Perl, JS, XSL, HTML	PostgreSQL	Chado DB	GMOD BLAST	https://www.drupal.org/project/biosoftware_bench
WebApollo	P, C	Community Annotation Editor	JSP, Perl, JS, HTML	PostgreSQL	Chado DB	WebApollo	http://genomearchitect.org/
Gene Pages	P, C, A	Gene Information page	PHP, Perl, JS, XSL, HTML	MySQL/ PostgreSQL	P/C/A-db2	PlantGenIE	
GeneList	P, C, A	Search PlantGenIE	PHP, JS, HTML	MySQL/ PostgreSQL	P/C/A-db2	PlantGenIE	

		Databases					
Enrichment	P, C, A	Statistical over-enrichment	Python, HTML, JS	MySQL	P/C/A-db1	PlantGenIE	
Galaxy	P, C, A	Integration and analysis platform	Python, HTML, JS	PostgreSQL	Galaxy DB	Galaxy Project	http://galaxyproject.org/
ComPlEX	P, C, A	Comparative Network visualization	PHP, JS, HTML	MySQL	db3	PlantGenIE	

Available In codes: P, PopGenIE; C, ConGenIE; A, AtGenIE.

Fig. 1 Expression database schema. This database stores all expression data related to expression tools – exImage, exNet, exHeatmap and exPlot tools**.**

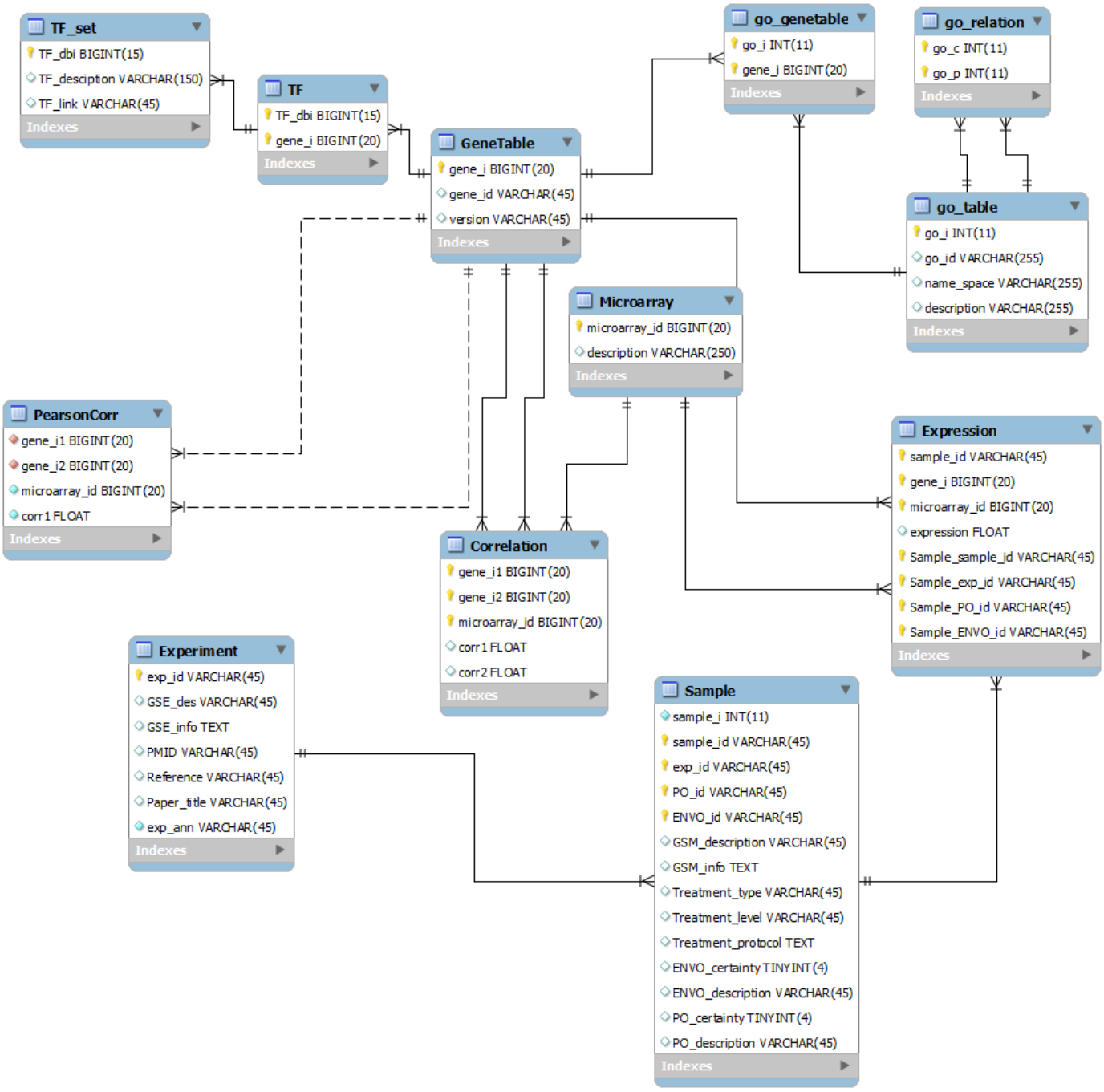


Fig. 2 Backend database schema. This database stores all data for Genome tools, especially the GeneList, Gene pages and Chromosome Diagram tools.

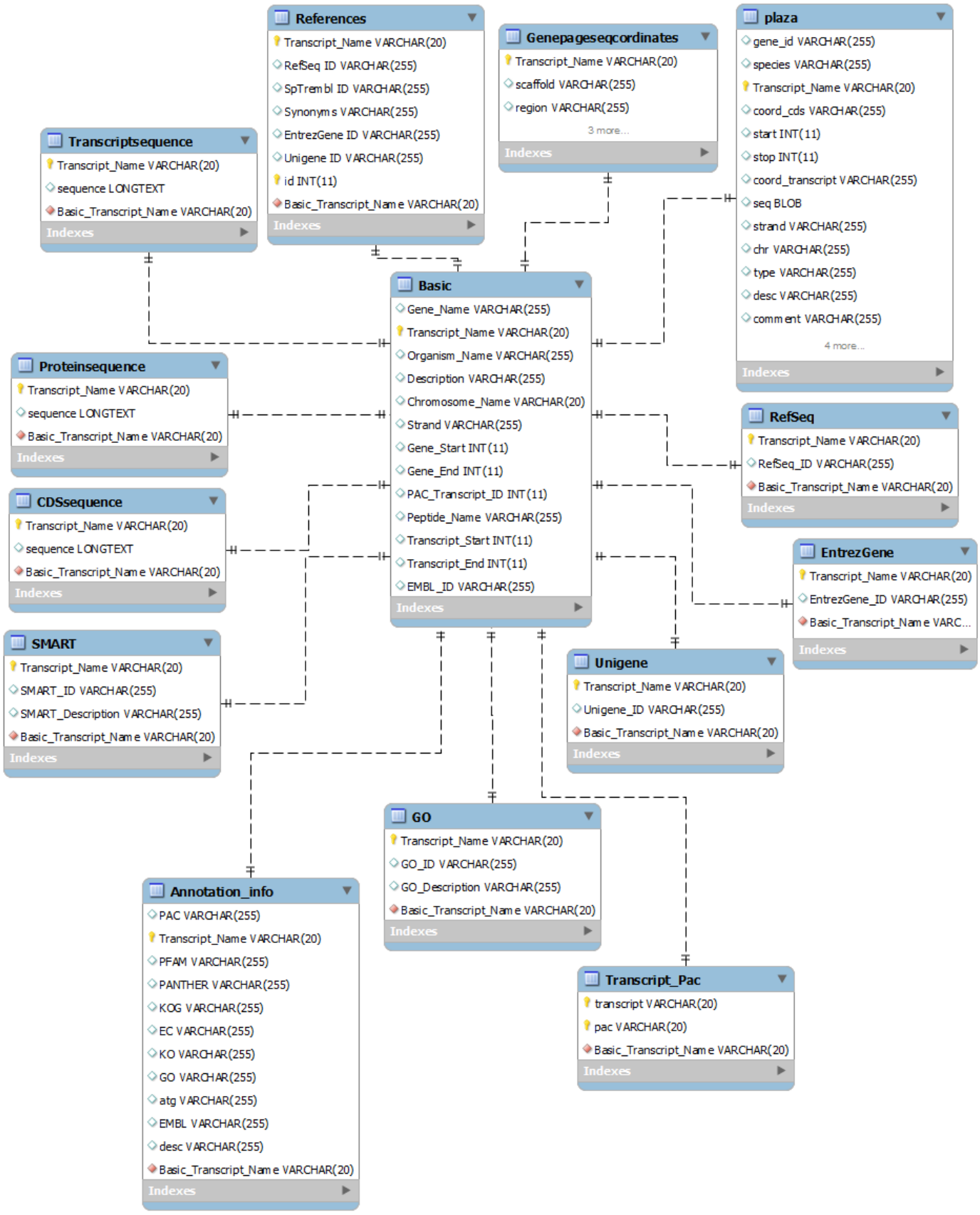
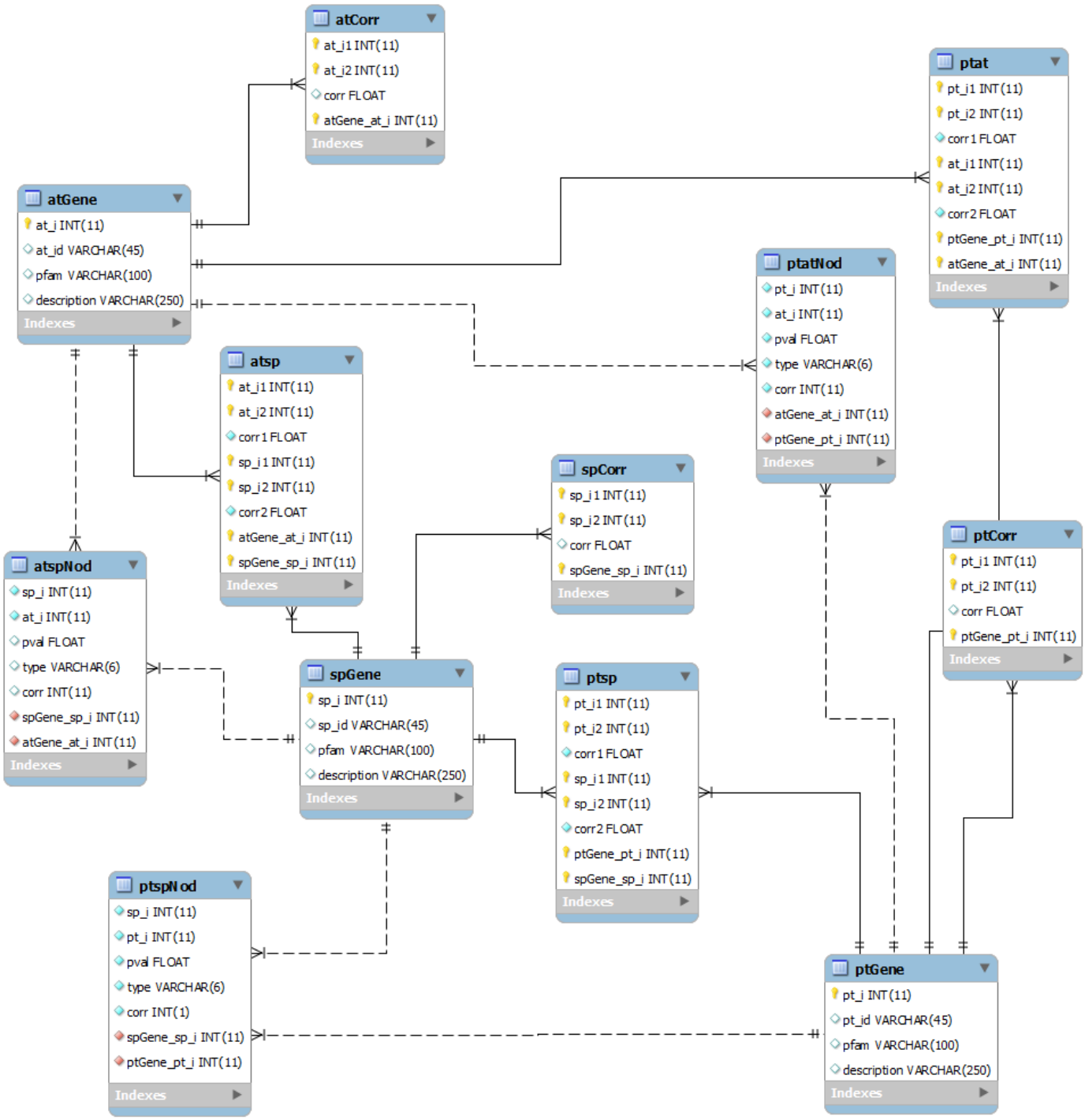


Fig. 3 Database schema for ComPlEx 2.0.



Methods S1 Sampling, RNA extraction and RNASeq analysis details for samples comprising the *Populus tremula* expression atlas.

Total RNA was extracted from 0.5 g tissue using a modified version of the CTAB method (Chang *et al.*, 1993) as described in Street *et al.* (2006). Briefly, the 10 sampled leaves were ground under liquid nitrogen using a pestle and mortar and 0.5 g of ground material was then used for RNA extraction. Precipitated RNA was further purified using an RNeasy Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. RNA concentration and purity were measured using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and integrity was analyzed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany). Total RNA preparations were sent to the Science for Life Laboratory (SciLifeLab, Stockholm, Sweden) for sequencing. Paired-end (2 × 100 bp) RNA-Seq data were generated using standard Illumina protocols and kits (TruSeq SBS KIT-HS v3, FC-401-3001; TruSeq PE Cluster Kit v3, PE-401-3001) and all sequencing was performed using the Illumina HiSeq 2000 platform. Samples were multiplexed by the addition of a unique barcode sequence and all samples were profiled on two lanes of the same flowcell. Briefly, the sequencing protocol involved DNase 1 digestion of total RNA, mRNA isolation by use of oligo(dT) beads, mRNA fragmentation, first and second strand cDNA synthesis, end-repair, A-tailing, barcoded adapter ligation and PCR amplification. Sequencing libraries were quality checked using an Agilent 2100 Bioanalyzer (Agilent Technologies) before sequencing. The quality of the raw sequence data was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Data were then filtered to remove adapters and trimmed for quality using Trimmomatic (v0.32; (Bolger *et al.*, 2014); settings TruSeq3-PE-2.fa`:2:30:10 LEADING:3 SLIDINGWINDOW:5:20 MINLEN:50`). Residual ribosomal RNA (rRNA) contamination was assessed and filtered using SortMeRNA (v1.9; (Kopylova *et al.*, 2012); settings `-n 6 -a 8 -v`) using the rRNA sequences provided with SortMeRNA (rfam-5s-database-id98.fasta, rfam-5.8s-database-id98.fasta, silva-bac-16s-database-id85.fasta, silva-euk-18s-database-id95.fasta, silva-bac-23s-database-id98.fasta and silva-euk-28s-database-id98.fasta). After both filtering steps, FastQC was run again to ensure that no technical artefacts had been introduced. Filtered reads were aligned to v3.0 of the *P. trichocarpa* genome (retrieved from the Phytozome resource, Goodstein *et al.*, 2012) using STAR (v2.3.1e, Dobin *et al.*, 2013) using the non-default settings: --OutQSconversion -31 --outReadsUnmapped Fastx --alignIntronMax 11000. The annotations obtained from the *P. trichocarpa* v3.0 GFF annotation file were modified to generate 'synthetic' gene models; that is, for each gene a nonredundant set of all exons from all transcripts was defined, with overlapping exons merged where necessary. This gene-model GFF file and the STAR read alignments were used as input to the HTSeq (http://www-huber.embl.de/users/anders/HTSeq/) htseq-count python utility to calculate exon-based read count values. The htseq-count utility utilises only uniquely mapping reads. Read counts were imported into R (v3.1.0, R-Core-Team, 2012) using the

Bioconductor (v2.14, Gentleman *et al.*, 2004) package (v1.4.5, Love *et al.*, 2014) to calculate normalized read counts. Normalized read counts were used for all subsequent expression analyses and for populating the relevant database tables for display in associated tools at PopGenIE.org.

Identifying tissue specific genes

Samples were grouped into similar tissue types based on annotation and specificity was then indicated using a score calculated as: Gene *G* is potentially specific in a tissue *T* if:

1. *G*'s max expression across all samples is in *T*
2. *G*'s highest average expression in a tissue is in *T*
3. *G*'s average expression in *T* is at least 2.0 (variance stabilised normalised expression)

All genes passing these criteria are then assigned a score calculated as average expression in tissue *T* divided by average expression in the tissue with the second highest average. The 10 highest scoring genes where not all genes returned a score >2 or all genes with a score >2 were returned (Table S2). These results are also available from the PopGenIE.org FTP site (ftp://popgenie.org/popgenie).

References

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2144–2120.

Chang S, Puryear J, Cairney J. 1993. A simple and efficient method for isolating RNA from pine trees. *Plant Molecular Biology Reporter* **11:** 113–116.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al.* 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5:** R80.

Goodstein D, Shu S, Howson R, Neupane R, Hayes R, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N *et al.* 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40:** D1178–D1186.

Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28:** 3211–3217.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology* **15:** 550.

R Core Team. 2014. *R: a language and environment for statistical computing*, *v3.2.0*. Vienna, Austria. URL http://www.r-project.org/.

Street NR, Skogström O, Sjödin A, Tucker J, Rodríguez-Acosta M, Nilsson P, Jansson S, Taylor G. 2006. The genetics and genomics of the drought response in *Populus*. *The Plant Journal* **48:** 321–41.