New Phytologist

Research

The Plant Genome Integrative Explorer Resource: PlantGenIE.org

David Sundell[1,2]*, Chanaka Mannapperuma[1]*, Sergiu Netotea[2], Nicolas Delhomme[1], Yao-Cheng Lin[3,4], Andreas Sjödin[2,5], Yves Van de Peer[3,4,6], Stefan Jansson[1], Torgeir R. Hvidsten[1,7] and Nathaniel R. Street[1,2]

[1]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-907 81 Umeå, Sweden; [2]Computational Life Science Cluster (CLiC), Umeå University, SE-907 81 Umeå, Sweden; [3]Department of Plant Systems Biology, VIB, Ghent, Belgium; [4]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium; [5]Department of Chemistry, Umeå University, SE-907 81 Umeå, Sweden; [6]Genomics Research Institute, University of Pretoria, Hatfield Campus, 0028 Pretoria, South Africa; [7]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1432 Ås, Norway

Author for correspondence:
Nathaniel R. Street
Tel: +46725372003
Email: nathaniel.street@umu.se

Received: *19 March 2015*
Accepted: *8 June 2015*

New Phytologist (2015)
doi: 10.1111/nph.13557

Key words: annotation, coexpression, conifer, database, genome browser, *Populus*, transcriptomics, web resource.

Summary

- Accessing and exploring large-scale genomics data sets remains a significant challenge to researchers without specialist bioinformatics training.
- We present the integrated PlantGenIE.org platform for exploration of *Populus*, conifer and Arabidopsis genomics data, which includes expression networks and associated visualization tools. Standard features of a model organism database are provided, including genome browsers, gene list annotation, BLAST homology searches and gene information pages. Community annotation updating is supported via integration of WebApollo.
- We have produced an RNA-sequencing (RNA-Seq) expression atlas for *Populus tremula* and have integrated these data within the expression tools. An updated version of the COMPLEX resource for performing comparative plant expression analyses of gene coexpression network conservation between species has also been integrated.
- The PlantGenIE.org platform provides intuitive access to large-scale and genome-wide genomics data from model forest tree species, facilitating both community contributions to annotation improvement and tools supporting use of the included data resources to inform biological insight.

Introduction

Advances in genomics, particularly developments in high-throughput sequencing, have enabled biologists to sequence the genome and transcriptome and to assay an expanding range of functional genomic processes in a rapidly expanding number of species. This is an exciting liberation from previous reliance on a small number of model species, often now enabling use of the most appropriate species to address a biological question rather than the best compromise model system. However, such advances require concomitant development of resources to enable the wider community to both access and extract knowledge from this wealth of data.

Populus has matured as a model system for forest tree and plant science research, with numerous genetic and genomics resources having been created (Wullschleger *et al.*, 2002; Jansson & Douglas, 2007). Curating these disparate resources, and their subsequent community dissemination, is a challenging yet essential undertaking to ensure maximal knowledge gain and new insight. These needs originally motivated us to develop PopGenIE.org (*Populus* Genome Integrative Explorer) as an easily accessible web resource focused primary on providing visualization tools for *Populus* genomics data (Sjödin *et al.*, 2009). We have subsequently continued to develop the resource by integrating new tools and data, making improved versions of popular tools and removing tools that were little used or worked ineffectively in order to streamline the user experience.

The availability of 'next'-generation sequencing (NGS) technologies continues to revolutionize the field of biology and has enabled novel insights and exciting new discoveries in the fields of RNA, transcriptomics and epigenetics in addition to facilitating an explosion in the number of genome sequencing projects being undertaken. One such example was the recent publication of three conifer genomes (Birol *et al.*, 2013; Nystedt *et al.*, 2013; Neale *et al.*, 2014; Wegrzyn *et al.*, 2014; Zimin *et al.*, 2014), which filled an important gap in the evolutionary history of green plants. With genomes of *c.* 20 Gbp, these also represented the largest genomes to have been sequenced to date. As part of the Norway spruce (*Picea abies*) genome project, Nystedt *et al.* (2013) established the ConGenIE.org (Conifer Genome Integrative Explorer) web resource to host the Norway spruce genome and associated RNA-sequencing (RNA-Seq) expression atlas data.

Netotea *et al.* (2014) recently presented the COMPLEX (comparative plant expression) comparative regulomics web resource as part of a study exploring cross-species conservation and divergence of gene coexpression networks. COMPLEX enables users to visualize the conservation or divergence of coexpression

*These authors contributed equally to this work.

New Phytologist

neighbours of orthologous genes, which can be used as an important additional source of information for identifying 'functional orthologues' or 'expressologues' rather than the most sequence-similar (i.e. highest sequence homology) ,orthologue. Initially, ComPlEx included coexpression networks for Arabidopsis (*Arabidopsis thaliana*) poplar (*Populus* spp.) and rice (*Oryza sativa*) derived from data assayed using Affymetrix© (Santa Clara, CA, USA) expression arrays and available at GEO (Edgar *et al.*, 2002). In order to integrate the resource within PlantGenIE.org we have developed an updated version of ComPlEx that includes coexpression data from the three dedicated PlantGenIE.org sites: *A. thaliana*, *Populus* species and Norway spruce.

Here we introduce the new umbrella PlantGenIE.org platform, providing details of tools and data resources that are integrated within the various species-specific PlantGenIE.org subdomains. As *A. thaliana* remains the central model system for plant science research, the entry point of many users to any plant-focused resource is often the identifiers for the set of *A. thaliana* genes of interest. We have therefore included the AtGenIE.org (*Arabidopsis thaliana* Genome Integrative Explorer) resource to provide a consistent user interface alongside the *Populus* (PopGenIE.org) and conifer (ConGenIE.org) subdomains. We have extended the ConGenIE.org resource to include the white spruce (*Picea glauca*) (Birol *et al.*, 2013) and loblolly pine (*Pinus taeda*) (Neale *et al.*, 2014; Wegrzyn *et al.*, 2014; Zimin *et al.*, 2014) genomes and have updated and integrated the ComPlEx resource. We provide worked examples, representing common usage scenarios, and highlight example use of the available tools for exploring gene function within and across species.

Description

The Plant Genome Integrative Explorer (http://PlantGenIE.org) domain serves as an umbrella site linking the (currently) three dedicated subdomains: *Populus* Genome Integrative Explorer (PopGenIE.org); Conifer (ConGenIE.org); and *Arabidopsis thaliana* (AtGenIE.org). A common set of core tools is available at each of these three sites (Fig. 1a), which are briefly detailed below. Full details, including implementation information, are available at the associated online help pages and in Supporting Information Notes S1. The tools are based either on existing open-source resources (GBrowse, JBrowse, Blast, exNet and Galaxy) or on original tools that we have developed (exImage, exPlot, exHeatmap, Chromosome Diagram, Enrichment, GeneList and ComPlEx). All code that we have developed is available freely from the associated file transfer protocol (FTP) resource (ftp://plantgenie.org).

Later, new and updated tools refer to additions or updates relative to the original PopGenIE.org resource (Sjödin *et al.*, 2009).

New tools

GeneList is the central gene identifier (ID) and annotation search tool, and the most likely entry point for many users. Users can

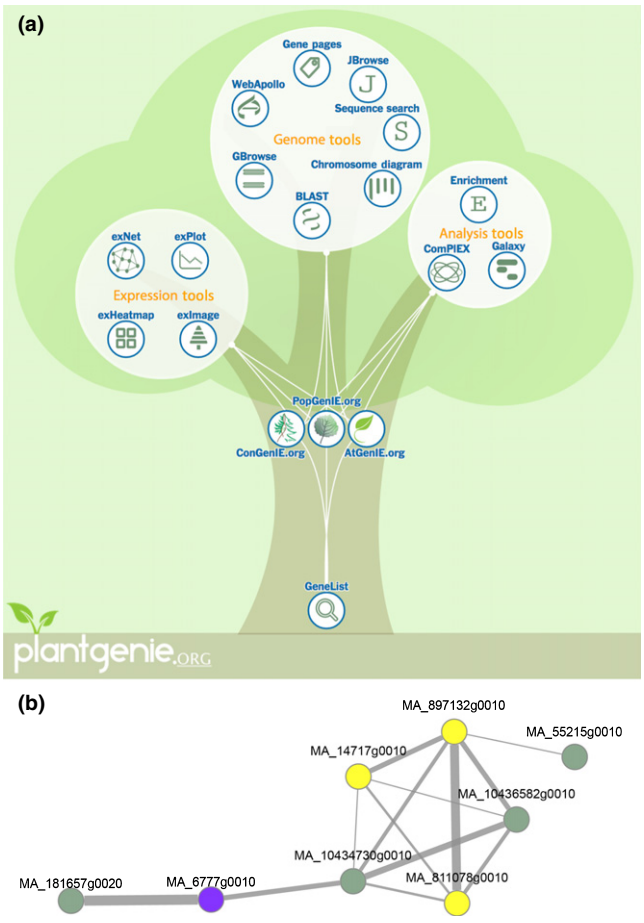

Fig. 1 (a) An overview of the tools and species available at the PlantGenIE.org platform, which comprises of *Populus* Genome Integrative Explorer (*Pop*GenIE), Conifer (ConGenIE) and *Arabidopsis thaliana* (*At*GenIE). Within the Expression Tools, exNet displays per-species expression networks; exPlot generates line graphs of expression profiles; exImage visualizes single gene expression profiles using an electronic fluorescent pictograph (eFP); exHeatmap represents expression profiles for genes in the active gene list using a heatmap representation that can optionally be clustered. Within the Genome Tools, GBrowse and JBrowse are genome browsers; Chromosome Diagram plots the location of genes in the active gene list in chromosomes (and is therefore only available at PopGenIE and AtGenIE); Gene Pages display structural and functional annotations, expression overview, and gene family and community annotation information; Blast performs sequence homology searches; WebApollo is a community annotation platform (only available at PopGenIE and ConGenIE); Sequence Search extracts sequence information for genes in the active gene list. Within the Analysis Tools, Enrichment performs overenrichment analysis for gene ontology (GO) categories, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Protein family (Pfam) domains within the active gene list; Galaxy is a platform for intensive data analysis; ComPlEx (comparative analysis of plant coexpression networks) allows analysis of cross-species coexpression conservation. (b) exNet representation of conserved *A. thaliana*–*Picea abies* coexpression neighbours of the single *P. abies* orthologue (MA_6670g0010) of the *A. thaliana* Vascular Related Nac-Domain Protein 4 (VND4) gene (AT1G12260) as presented in Nystedt *et al.* (2013). Coexpression neighbours of VND4 were first selected in ComPlEx using a threshold of four. *Picea abies* genes with conserved coexpression to MA_6670g0010 were then selected and added with a gene list. These genes were then visualized in the exNet tool at ConGenIE.org using a threshold of four. Edge thickness indicates coexpression strength. The 'colour genes' feature of exNet was used to identify MA_6670g0010 (shown in purple) and genes in the GO category 'oxidoreductase activity' (GO:0016491; shown in yellow).

input IDs from any of the species included at PlantGenIE.org or can perform free text searches (for example cellulose synthase genes can be searched for by entering 'CesA' to find genes with annotation information including this term). Matching genes are then displayed and can be added to the currently active gene list or saved to a new gene list. Gene lists can be named and multiple lists can be stored for return use. A user can additionally share a gene list by sending a generated sharing uniform resource locator (URL) web address. By default, shared URLs remain active for 1 month; however, users can optionally request that a shared URL be available long-term as a stable resource, making them a suitable alternative method for linking to annotated gene lists within publications or presentations, for example.

exNet provides visual exploration of expression networks within precalculated transcriptional expression networks (Fig. 1b) using the Cytoscape Web network browser (Lopes *et al.*, 2010). ComPlEx can be used to explore expression network conservation between the PlantGenIE.org species. This tool is implemented as detailed in Netotea *et al.* (2014) but using an updated visual interface that includes automated gene list creation of orthologues for the PlantGenIE.org species. exHeatmap generates an expression-based heat map representation for genes in the currently active gene list within a selected gene expression data set, such as the expression atlases included at the three PlantGenIE.org subdomains. 'Enrichment' performs overenrichment analyses for gene ontology (GO) categories (Ashburner *et al.*, 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa & Goto, 2000) or protein family (Pfam) domain presence (Finn *et al.*, 2010) for the active gene list. Users can select a multiple testing correction method in addition to setting *P*-value and minimum gene count filtering thresholds. 'Chromosome Diagram' plots the location of genes within the active gene list along chromosomes. As such, this tool is only offered for species where chromosome-level assemblies are available. 'Sequence Search' extracts the corresponding genomic, transcript, coding DNA sequence (CDS) or peptide sequences in FASTA format given a list of gene IDs. GBrowse (Stein *et al.*, 2002) remains as the primary genome browser tool; however, we also now provide JBrowse (Skinner *et al.*, 2009) as an alternative as this is often a more suitable option when browsing large-scale data sets such as NGS read alignments for RNA-Seq profiles. WebApollo (Lee *et al.*, 2013) has been implemented at both PopGenIE.org and ConGenIE.org to facilitate community annotation updating. This resource allows a user to submit both structural and annotation-based corrections and information within a simple web-based graphical user interface (GUI). Submitted updates are immediately incorporated into the associated gene information web page including submitter information. WebApollo is not available for *A. thaliana* as there is already a suitable and stable resource for community annotation available (Lamesch *et al.*, 2012). Galaxy is an open-source project (Goecks *et al.*, 2010) that aims to provide a user-friendly, web-based GUI to perform computationally intensive analyses of, for example, high-throughput second-generation sequencing data such as RNA-Seq. We provide a custom instance of Galaxy that includes the genomes and associated annotations hosted at PlantGenIE.org, as these are not commonly available at other public resources. We are implementing a number of workflows with integrated links from various components of PlantGenIE.org. For example, we include workflows constructed using the Osiris suite, a set of wrappers around a number of phylogenetic tools (Oakley *et al.*, 2014), to allow users to generate phylogenetic trees from selected sequences imported from the GeneList tool or using gene family information from a gene information web page.

Updated tools

exImage replaces the eFP browser (Winter *et al.*, 2007) as a tool to provide a pictorial representation of single gene expression levels within predefined collections of samples, such as the expression atlases. For each PlantGenIE.org species, we include an image representing the associated expression atlas data set and, where suitable, additional images collating samples from large-scale expression profiling experiments. exPlot is an updated version of our previous plotting tool for generating line graph representations of expression across samples from a selected experiment for genes in the currently active gene list. Blast homology searches are available using an updated server interface that includes integrated links to display matching homologous sequence regions within the GBrowse genome browser.

In addition to the analysis tools described above, we have also updated the gene information pages, which are largely accessed via links provided within the GeneList and Expression analysis tools or via browsing of the genome in GBrowse or JBrowse. The updated information pages include basic structural information, available associated functional annotations, display of sequence data (genomic, transcript, CDS, and peptide regions, and custom selectable upstream and downstream flanking regions), an expression overview, cross-species gene family information (as detailed in Nystedt *et al.*, 2013) and details of associated publications and user-submitted annotations. Using the included gene family information, a desired set of species can be selected and the corresponding protein sequences then imported directly either to the phylogeny.fr resource (Dereeper *et al.*, 2008) or to the PlantGenIE.org Galaxy instance or exported in FASTA format.

We provide a number of web services to allow programmatic access to the tools by external web resources, details of which are available at http://api.plantgenie.org. All data stored in the back-end databases are provided on the FTP site (ftp://plantgenie.org/). A public git repository containing generic versions of the pipelines for analysis of NGS data (such as the expression atlas RNA-Seq data detailed below), in addition to project-specific code and data, can be cloned from https://microasp.upsc.se/root/upscb-public.

An expression atlas for *Populus tremula*

To provide a comprehensive gene expression resource generated using RNA-Seq, both as a community resource and to support

gene prediction in the *P. tremula* genome project, we generated an expression atlas comprising 24 samples collected from various tissues under a range of abiotic, biotic and seasonal conditions (Tables 1, S1; Fig. 2a; Methods S1). All samples collected from glasshouse conditions represent pools collected from multiple biological replicates (min $n = 3$). From each biological replicate, multiple leaves/roots and so on were sampled (min $n = 3$ per replicate tree). Pooling was performed after RNA extraction from each biological replicate to ensure equimolar representation. In the case of samples collected from the mature 'parent' tree (i.e. the tree from which the glasshouse plants were cloned), which is growing as a wild aspen on the Umeå University campus (Bhalerao *et al.*, 2003), each sample represents a pool of multiple leaves or flowers (min $n = 5$) collected from multiple branches (min $n = 3$) or of multiple roots ($n = 15$). We refer to this data set as the '*P. tremula* expression atlas' or using the abbreviation 'Potra exAtlas'. The raw data are available from the European Nucleotide Archive (ENA) as accession ERP004398 and the matrix of normalized expression values is available from the PlantGenIE.org FTP site. Full experimental details are provided in Methods S1.

We used the expression atlas together with all existing RNA-Seq data generated at the Umeå Plant Science Centre (both published and unpublished) to generate a set of the most tissue-representative genes (Table S2).

Table 1 Details of samples comprising the *Populus tremula* expression atlas

ENA ID	Sample type	Condition	Location	Date	% unique	% multiple	% no feature
ERS374199	Twigs	Nongirdled	Outdoor	30 July 2010	85.91	5.83	3.37
ERS374200	Flowers	Dormant	Outdoor	3 May 2010	86.11	5.12	2.69
ERS374201	Leaves	Drought (3 d)	Glasshouse	30 July 2010	88.28	6.20	2.68
ERS374202	Flowers	Expanded (not mature)	Outdoor	30 May 2010	81.72	4.76	2.37
ERS374203	Leaves	Beetle-damaged (3 d)	Glasshouse	5 August 2010	87.76	5.76	3.02
ERS374204	Leaves	Mechanical damage (3 d)	Glasshouse	5 August 2010	86.81	6.35	2.74
ERS374205	Leaves	Mature	Outdoor	3 August 2010	88.74	5.35	3.18
ERS374206	Leaves	Mature	Outdoor	30 July 2010	85.69	5.31	3.07
ERS374207	Flowers	Expanding	Outdoor	18 May 2010	88.42	4.99	2.38
ERS374208	Suckers	Whole suckers	Outdoor	5 August 2010	88.14	5.91	2.38
ERS374209	Petiole	Mature	Outdoor	30 July 2010	87.88	5.58	3.41
ERS374210	Leaves	Healthy (mature)	Glasshouse	5 August 2010	88.01	6.08	3.02
ERS374211	Buds	Pre-chilling	Outdoor	30 July 2010	88.92	5.33	3.17
ERS374212	Buds	Dormant	Outdoor	3 May 2010	88.20	4.80	3.38
ERS374213	Leaves	Drought (3 d)	Glasshouse	30 July 2010	87.84	6.25	2.74
ERS374214	Leaves	Freshly expanded	Outdoor	25 May 2010	88.14	6.04	2.66
ERS374215	Leaves	Nongirdled	Outdoor	3 September 2010	89.37	5.09	3.64
ERS374216	Roots	Control	Glasshouse	30 July 2010	88.12	5.73	3.20
ERS374217	Leaves	Girdled	Outdoor	3 September 2010	89.35	5.30	3.54
ERS374218	Seeds	Mature	Outdoor	21 June 2010	69.71	4.08	2.54
ERS374219	Leaves	Young (expanding)	Outdoor	18 May 2010	89.81	5.39	2.02
ERS374220	Leaves	Control	Glasshouse	30 July 2010	88.88	5.86	2.68
ERS374221	Phloem/cambium	Dormant	Outdoor	3 May 2010	89.18	5.84	3.33
ERS374222	Roots	Drought (3 d)	Glasshouse	30 July 2010	88.59	5.17	3.29

The associated RNA-Seq data are available from the European Nucleotide Archive (ENA) as accession ERP004398 and ENA ID refers to the sample ID assigned. % unique refers to the percentage of filtered reads that aligned uniquely to the *P. trichocarpa* genome and % multiple refers to reads aligning nonuniquely (i.e. equally good alignment to > 1 position). % no feature refers to reads aligning but not overlapping a known gene.

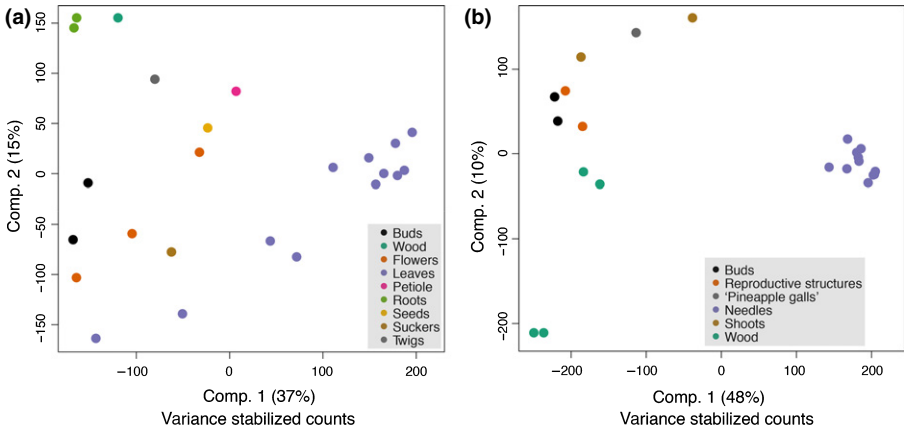

Fig. 2 Overview of the *Populus tremula* and *Picea abies* expression atlas data sets. Principal component analysis (PCA) visualization of normalized expression values are shown for genes expressed within the (a) *P. tremula* and (b) *P. abies* expression atlas data sets. Samples are classified by tissue type, as indicated in the key.

Results

Below we present the Potra exAtlas in addition to worked examples of how the PlantGenIE.org resource can be used to perform representative common tasks, such as updating and annotating gene lists, exploring potential gene function on the basis of expression and exploring gene coexpression conservation.

To provide a resource for exploring potential gene function on the basis of expression, we have developed the *P. tremula* Expression Atlas (Potra exAtlas) comprising 24 samples profiled using RNA-Seq (Table 1) with an average 18.4 M high-quality paired-end reads per sample, representing > 77 Gbp (441.6 M reads) in total. These samples were used to calculate an expression network, available in exNet, to identify sets of tissue-representative genes (Table S2), a subset of which are represented in exImage. We generated normalized expression values by aligning quality filtered and trimmed RNA-Seq reads to the *Populus trichocarpa* reference annotation with, on average, 87% reads aligning uniquely and 92.6% aligning in total (Table 1). This exemplifies that the coding space is well conserved between the two species and that the reference genome assembly can be utilized for RNA-Seq-based expression quantification in additional *Populus* species. The data set contains reads aligning to 90% of genes expressed in at least one sample (37 267 out of 41 335) and 86% of genes expressed in at least three samples (35 533). An average 2.6% of reads aligned but did not overlap annotated features (i.e. protein-coding genes) and, as such, may represent novel genes (detailed characterization of these will be presented elsewhere). Principal component analysis (PCA) representation of the data (Fig. 2a) reveals clear separation of tissue types on the basis of expression. This basic separation of tissues types is equally reflected in the *P. abies* expression atlas data available in ConGenIE.org (Fig. 2b; Nystedt *et al.*, 2013), where clear separation of needle and wood samples was observed in

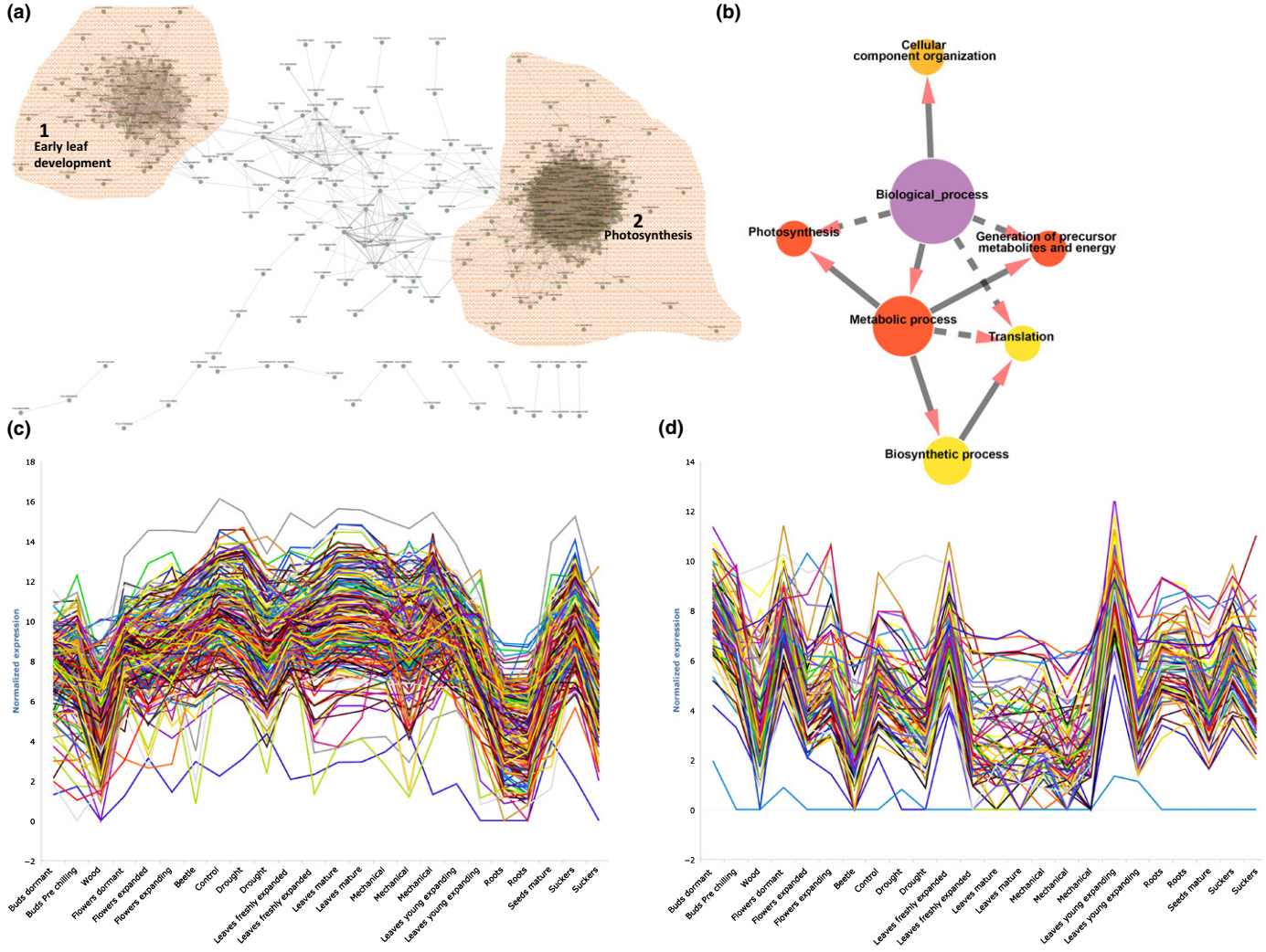

Fig. 3 Use of PopGenIE.org tools to update a list of 1116 *Populus trichocarpa* genes identified as being highly representative of leaves (Street *et al.*, 2008) using the V1.1 genome annotation. There are matches to 681 V3 gene IDs. (a) The coexpression network of the 352 genes that were coexpressed (connected by an edge in the network) at default thresholds within the '*Populus* Affymetrix' network available in exNet. Two user-selected subsets of connected genes are highlighted and labelled as 1 and 2. (b) Gene ontology (GO) Enrichment results using the GO full ontology for all genes in (a). (c, d) exPlot line graphs showing expression within the *P. tremula* expression atlas data set for the two user-identified subclusters highlighted and labelled as 1(c) and 2(d) in (a).

addition to the intermediate nature of the sexual reproductive and immature bud samples.

One point of frustration, and often a significant barrier to the use of previously published results from genome-scale analyses, is the changing of gene names between different assembly and annotation releases of a genome. For *P. trichocarpa*, we use liftover information provided by the phytozome.org resource (which is the official provider of annotation information for *P. trichocarpa*) to allow conversion of an input set of V1 or V2 gene IDs to matching V3 genes. Correspondence is generally high between V2 and V3 but substantially lower for V1 to V3. As an example use of this functionality, we used the GeneList tool to convert a list of genes identified from a large-scale analysis of cDNA microarray data as being highly representative of leaves (Street *et al.*, 2008). These 1116 V1 gene IDs returned matches to only 681 V3 gene IDs (Table S3). The GO ENRICHMENT tool was then utilized to test for functional over-enrichment within these 681 genes. In agreement with the results presented in Street *et al.* (2008), significant overenrichment for GO terms associated with developmental processes and photosynthesis was identified. We next used the exNet tool to select the 352 genes that were coexpressed (connected by an edge in the network) at the default thresholds within the '*Populus* Affymetrix' network (Fig. 3a). This coexpression network comprises 462 Affymetrix microarray samples representing a diverse set of developmental samples and condition perturbations (see Netotea *et al.*, 2014 for further details). The expression and expand thresholds represent context likelihood of relatedness (CLR) values, which are based on mutual information (MI); a pair-wise measure of mutual dependence. The CLR approach transforms each MI value into a z-score indicating how much higher (in SD) that MI value is than the average MI value in the network neighbourhood (see Netotea *et al.*, 2014). Hence, the CLR value indicates the significance of the coexpression between two genes.

This revealed two distinct subclusters, which we manually selected and saved as new gene lists. Cluster 1 (Fig. 3a) contained 195 genes (Table S4) and was enriched for photosynthesis genes (Fig. 3b). Cluster 2 (Fig. 3a) contained 77 genes (Table S5) and was enriched for GO Plant Slim categories including the Biological Process 'DNA metabolic process' and the Cellular Component category 'Nucleus'. Plotting expression profiles for each cluster within the 'Potra exAtlas' data set using the EXPLOT tool (Fig. 3c,d, respectively) revealed cluster-specific expression profiles. Genes in cluster 1 had high expression in mature leaf samples, while cluster 2 had distinct expression peaks in dormant leaf and flower buds, expanding and freshly expanding leaves, that is, were active during early leaf development, which reflects the GO ENRICHMENT results. Expression of genes in cluster 2 was generally higher in roots than expression of genes within cluster 1, possibly suggesting that this cluster contains genes more generally involved in development rather than being leaf specific.

In the same publication, Street *et al.* (2008, Table 2 therein) detailed the most highly connected (central) genes within an identified set of coexpression modules. We used the GENELIST tool to identify the corresponding V3 IDs. There were six genes, four of which had corresponding V3 matches (Potri.005G257500 corresponding to estExt_fgenesh4_pm.C_LG_V0721, Potri.001G167800 corresponding to estExt_fgenesh4_pg.C_1580005, Potri.003G038300 corresponding to estExt_fgenesh4_pg.C_440087, and Potri.010G123400 corresponding to estExt_fgenesh4_pg.C_LG_X1117). We again used EXNET to select the coexpression neighbours of these four genes within the *Populus* Affymetrix network at default thresholds. Only Potri.005G257500 returned a cluster of coexpressed genes at this high threshold, which, on the basis of GO ENRICHMENT results, corresponded to the cluster annotated as 'DNA replication and structure' (blue cluster) in Street *et al.* (2008). The other four genes, in particular Potri.003G038300 (the 'chloroplast/photosynthesis', or turquoise, cluster from Street *et al.* (2008)), also had coexpression neighbours that return GO ENRICHMENT results in agreement with those in the original paper, but required use of lower coexpression thresholds.

As detailed above, we have updated the COMPLEX resource (Netotea *et al.*, 2014) for investigating expression network conservation and divergence to include expression networks for the three current PlantGenIE.org focal species. COMPLEX has been integrated within the PlantGenIE.org infrastructure, for example through the use of a common set of gene lists that appear in all three subdomains. As an example use of this resource, we explored the expression conservation of the two *PHOTOSYSTEM I SUBUNIT D* (*PSA-D*) genes (At4g02770 and At1g03130) used as worked examples in Mutwil *et al.* (2011). The protein products of these genes are required for assembly of the photosystem I complex (Ihnatowicz *et al.*, 2004). Entering these two gene IDs identified a single orthologue in Norway spruce (*P. abies*). We selected the coexpression neighbours of the two *A. thaliana* genes at a coexpression threshold of seven and a conservation threshold of three (see Netotea *et al.*, 2014 for further details of these thresholds). A tightly connected set of 13 genes were present in the AtGenExpress (Schmid *et al.*, 2005) Development network, all but one of which has an identified orthologue in Norway spruce (Fig. 4a). The corresponding set of 10 orthologues in Norway spruce was also highly connected, having conserved expression edges and revealing a highly conserved set of coexpressed genes. The gene family information presented as a 'tab' within the gene information page for the single Norway spruce orthologue (MA_9169733g0010, highlighted in red in Fig. 4a) of the two starting *A. thaliana* genes was used to select all listed gene family members in the included species. A phylogenetic tree based on these sequences was then produced using the provided link to the PlantGenIE.org GALAXY instance (Fig. 4b).

Discussion

The explosion in genomics data available, in both volume and range, from an ever-increasing range of species can be both exciting and intimidating. For those who do not specialize in bioinformatics, or lack even basic bioinformatics skills, not being able to access and extract new insight from this wealth of data can be frustrating. There is therefore a need to design easy to access,

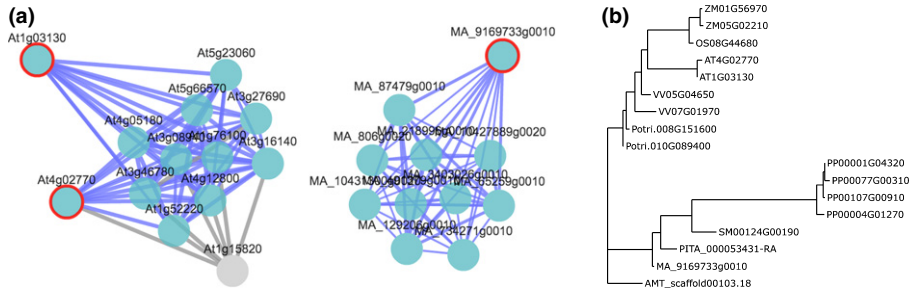

Fig. 4 (a) Conservation of the expression neighbourhood between *Arabidopsis thaliana* (left) and *Picea abies* (right) for genes coexpressed with the *A. thaliana* genes *PHOTOSYSTEM I SUBUNIT D1* (*At-PSA-D1*; At4g02770) and *At-PSA-D2* (At1g03130). The *P. abies* orthologue of the two *A. thaliana* genes is highlighted in red. (b) An unrooted phylogenetic tree for gene family members of the *P. abies* orthologue (MA_9169733g0010) represented in (a). Family members were selected using the Gene Information page with a subset of the available species selected and exported to the PlantGenIE Galaxy platform. The shared Galaxy workflow 'Create a Phylogenetic Tree' was applied and the final step of the workflow exported as a Scalable Vector Graphic (SVG) image. The included species are identified by gene ID prefixes: AT, *Arabidopsis thaliana*; VV, *Vitis vinifera*; Potri, *Populus trichocarpa*; PITA, *Pinus taeda*; PP, *Physcomitrella patens*; SM, *Saleginela moellendorffii*; ZM, *Zea mays*; OS, *Oryza sativa* ssp. Japonica; AMT_scaffold, *Amborella trichopoda*.

publicly available and free web resources and tools to overcome this limitation. As a result of the increasing range of species for which transcriptomics data are available, the potential of comparative analyses such as comparative regulomics is also growing. Such analyses can provide a powerful means of transferring knowledge of gene function between species, of increasing understanding of evolutionary changes to expression control and of informing selection of not only the most sequence-similar homologue but also the most expression conserved 'expressologue' (Netotea *et al.*, 2014).

With these challenges and opportunities in mind, we have developed the PlantGenIE.org resource as an integrated set of tools for exploring genomics data within and among the model forest tree systems of *Populus* and conifers (Fig. 1a). The tools included allow intuitive and easy to interpret access to large-scale transcriptomics data sets and analysis of coexpression conservation and divergence between species, in addition to providing community resources for refining and updating genome annotations.

Our intention is that PlantGenIE.org should complement existing resources, where they exist, and not compete with these. We have therefore integrated links to such resources and we obtain and re-disseminate genomic and annotation information from the official data sources for included genomes (currently TAIR for *A. thaliana* and Phytozome for *P. trichocarpa*). Our emphasis at PlantGenIE.org is to provide access to, and visualization of, genomics data as this is a service not currently offered by other genome resource sites for the included species. The resource currently comprises data primarily from Norway spruce and *Populus*, although we envisage expanding the range of species available as adequate data resources are developed. Alternatively, we would welcome involvement from the communities of additional plant species.

We anticipate that the PlantGenIE.org resource will enable the plant community to extract and support new biological insights through use and visualization of the included large-scale and genome-wide data sets. To help guide new users, we have included a number of example workflows that are detailed at the PlantGenIE.org homepage and that represent the most common usage scenarios of our past visitors, as identified using (anonymous) Google Analytics tracking of how users progress between pages (i.e. tools). We also provide quick tours and example data sets for all available tools and help pages that include basic usage overviews and screen cast videos showing example usage, in addition to more technical and detailed information on tool implementation and the available data sets.

We hope that the WebApollo resource will be utilized and embraced by the community to provide improvements and corrections to the current genome annotations. This will enable the wider community to benefit from the expert knowledge of specialists in addition to informing future annotation updates. Such annotation information need not be restricted to corrections of gene model structures but can include links to published work and associated expression support that may inform biological function. All such submitted annotation information will be immediately disseminated via the associated gene information page with full credit given to the submitter.

We will continue to expand and develop the resource by inclusion of additional species (although we intend to maintain a focus on perennial, woody species), data sets and types and through development of new tools. The included version of genomes and annotations will be updated as these are released by ourselves or the community. Currently, the resource is focused largely on transcriptomics data and genome assemblies. Extending the range of included transcriptional data sets available for the conifer species, largely through engagement with the associated communities, is a target of current effort to ensure maximal value for biological insight. However, we anticipate incorporating additional data types including chromatin immunoprecipitation sequencing (ChIP-Seq), small RNA-Seq, genome-wide sequence variant information (e.g. single nucleotide polymorphisms) and phenotypes in the near future.

Acknowledgements

This work was supported by funds from the Alice Wallenberg Foundation, the Swedish Research Council (VR), the Swedish Governmental Agency for Innovation Systems (Vinnova), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas), and the Swedish

Foundation for Strategic Research (SSF), in part through the UPSC Berzelii Centre for Forest Biotechnology. NRS is supported by the Trees and Crops for the Future (TC4F) project. We thank Simon Birve for extensive support establishing and maintaining IT infrastructure.

References

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J *et al.* 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**: 25–29.

Bhalerao R, Keskitalo J, Sterky F, Erlandsson R, Bjorkbacka H, Birve SJ, Karlsson J, Gardestrom P, Gustafsson P, Lundeberg J *et al.* 2003. Gene expression in autumn leaves. *Plant Physiology* **131**: 430–442.

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Saint Yuen MM, Keeling CI, Brand D, Vandervalk BP *et al.* 2013. Assembling the 20 Gb white spruce (*Picea glauca)* genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**: 1492–1497.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M *et al.* 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**: W465–W469.

Edgar R, Domrachev M, Lash AE. 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**: 207–210.

Finn R, Mistry J, Tate J, Coggill P, Heger A, Pollington J, Gavin L, Gunasekaran P, Ceric G, Forslund K *et al.* 2010. The Pfam protein families database. *Nucleic Acids Research* **38**: D211–D222.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* **11**: R86.

Ihnatowicz A, Pesaresi P, Varotto C, Richly E, Schneider A, Jahns P, Salamini F, Leister D. 2004. Mutants for photosystem I subunit D of *Arabidopsis thaliana*: effects on photosynthesis, photosystem I stability and expression of nuclear genes for chloroplast functions. *Plant Journal* **37**: 839–852.

Jansson S, Douglas CJ. 2007. *Populus*: a model system for plant biology. *Annual Review of Plant Biology* **58**: 435–458.

Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**: 27–30.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M *et al.* 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**: D1202–D1210.

Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biology* **14**: R93.

Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. 2010. Cytoscape Web: an interactive web-based network browser. *Bioinformatics* **26**: 2347–2348.

Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S. 2011. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* **23**: 895–910.

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD *et al.* 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**: R59.

Netotea S, Sundell D, Street NR, Hvidsten TR. 2014. ComPlEx: conservation and divergence of co-expression networks in *A. thaliana, Populus* and *O. sativa*. *BMC Genomics* **15**: 106.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A *et al.* 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**: 579–584.

Oakley TH, Alexandrou MA, Ngo R, Pankey MS, Churchill CK, Chen W, Lopker KB. 2014. Osiris: accessible and reproducible phylogenetic and phylogenomic analyses within the Galaxy workflow management system. *BMC Bioinformatics* **15**: 230.

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of *Arabidopsis thaliana* development. *Nature Genetics* **37**: 501–506.

Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S. 2009. The *Populus* Genome Integrative Explorer (PopGenIE): a new resource for exploring the *Populus* genome. *New Phytologist* **182**: 1013–1025.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. *Genome Research* **19**: 1630–1638.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al.* 2002. The generic genome browser: a building block for a model organism system database. *Genome research* **12**: 1599–1610.

Street NR, Sjödin A, Bylesjö M, Gustafsson P, Trygg J, Jansson S. 2008. A cross-species transcriptomics approach to identify genes involved in leaf development. *BMC Genomics* **9**: 589.

Wegrzyn JL, Liechty JD, Stevens KA, Wu L-S, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martínez-García PJ *et al.* 2014. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* **196**: 891–909.

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. 2007. An 'electronic fluorescent pictograph' browser for exploring and analyzing large-scale biological data sets. *PLoS ONE* **2**: e718.

Wullschleger SD, Jansson S, Taylor G. 2002. Genomics and forest biology: *Populus* emerges as the perennial favorite. *Plant Cell* **14**: 2651–2655.

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ *et al.* 2014. Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics* **196**: 875–890.

Supporting Information

Additional supporting information may be found in the online version of this article.

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

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

Table S1 Details of RNA-Seq samples comprising the *Populus tremula* expression atlas

Table S2 Tissue-specific genes identified from *Populus tremula* RNA-Seq expression data

Table S3 Leaf representative genes from Street *et al.* (2008)

Table S4 Identified coexpressed genes within 'cluster 1', as detailed in the main text

Table S5 Identified coexpressed genes within 'cluster 2', as detailed in the main text

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.