Comparative Transcriptomic Approaches Exploring Contamination Stress Tolerance in *Salix* sp. Reveal the Importance for a Metaorganismal de Novo Assembly Approach for Nonmodel Plants[1][OPEN]

Nicholas J. B. Brereton[2]*, Emmanuel Gonzalez[2], Julie Marleau, Werther Guidi Nissim, Michel Labrecque, Simon Joly, and Frederic E. Pitre

Institut de recherche en biologie végétale, University of Montreal, Montreal QC H1X 2B2, Canada (N.J.B.B., E.G., J.M., M.L., S.J., F.E.P.); and Montreal Botanical Garden, Montreal, QC H1X 2B2, Canada (W.G.N., M.L., S.J., F.E.P.)

ORCID ID: 0000-0003-3672-1398 (E.G.).

Metatranscriptomic study of nonmodel organisms requires strategies that retain the highly resolved genetic information generated from model organisms while allowing for identification of the unexpected. A real-world biological application of phytoremediation, the field growth of 10 *Salix* cultivars on polluted soils, was used as an exemplar nonmodel and multifaceted crop response well-disposed to the study of gene expression. Sequence reads were assembled de novo to create 10 independent transcriptomes, a global transcriptome, and were mapped against the *Salix purpurea* 94006 reference genome. Annotation of assembled contigs was performed without a priori assumption of the originating organism. Global transcriptome construction from 3.03 billion paired-end reads revealed 606,880 unique contigs annotated from 1588 species, often common in all 10 cultivars. Comparisons between transcriptomic and metatranscriptomic methodologies provide clear evidence that nonnative RNA can mistakenly map to reference genomes, especially to conserved regions of common housekeeping genes, such as actin, α/β-tubulin, and elongation factor 1-α. In *Salix*, Rubisco activase transcripts were down-regulated in contaminated trees across all 10 cultivars, whereas thiamine thizole synthase and CP12, a Calvin Cycle master regulator, were uniformly up-regulated. De novo assembly approaches, with unconstrained annotation, can improve data quality; care should be taken when exploring such plant genetics to reduce de facto data exclusion by mapping to a single reference genome alone. *Salix* gene expression patterns strongly suggest cultivar-wide alteration of specific photosynthetic apparatus and protection of the antenna complexes from oxidation damage in contaminated trees, providing an insight into common stress tolerance strategies in a real-world phytoremediation system.

Coppiced willows have the ability to produce high biomass yields in temperate regions under challenging conditions and have positive impacts on biodiversity (Labrecque et al., 1995; Hasselgren, 1999; Kuzovkina and Quigley, 2005; Anderson and Fergusson, 2006; Sage et al., 2006; Haughton et al., 2009; Kuzovkina and Volk, 2009). The biomass from certain cultivars can be sugar rich and highly accessible, permitting reduced severity pretreatment for high cell wall glucose release (Ray et al., 2012), which can be both economically and environmentally beneficial for downstream bioenergy applications such as lignocellulosic biofuel production (Stephenson et al., 2010). However, the use of high-grade agricultural land for biomass cultivation has the potential to displace nutrient-demanding food crops as well as negate the environmental benefits of bringing degraded land into production by using the efficient nutrient cycling physiology of biomass crops such as willow (Bollmark et al., 1999; Weih and Nordh, 2002; Black et al., 2011; Graham-Rowe, 2011; Murphy et al., 2011; Brereton et al., 2014). One of the current strategies to bring degraded land back into productivity using low input agriculture is to actively rejuvenate contaminated or polluted land in a process termed phytoremediation. Numerous studies have demonstrated willow's capacity to tolerate contaminated soils beyond the majority of agricultural crops (Robinson et al., 2000; Volk et al., 2006; Pitre et al., 2010; Grenier et al., 2015), presenting an industrially pertinent opportunity to reduce high biomass production costs (Huang et al., 2009;

[1] This work was financially supported by the GenoRem Project (Genome Canada and Genome Québec) as well as BioFuelNet Canada and NCE (Networks of Center of Excellence)

[2] These authors contributed equally to the article.

* Address correspondence to nicholas.brereton@umontreal.ca; nicholas.brereton@transcriptomics.org.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Nicholas Brereton (nicholas.brereton@umontreal.ca).

F.E.P., S.J., and M.L. designed the study; F.E.P., S.J., M.L., J.M., and W.N.G. established the field trial and sample preparation; N.J.B.B., E.G., M.L., F.E.P., and S.J. interpreted the data and drafted the manuscript. All authors read and approved the final manuscript.

[OPEN] Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.16.00090

Black et al., 2011; Yue et al., 2014) through added-value cultivation. However, the genetic mechanisms behind such tolerance are poorly understood.

A complex trait such as contamination tolerance may require a diverse array of developmental alterations and responses relating to contaminant immobilization, transport or metabolism, oxidoreduction, drought tolerance, xylem alteration (i.e. hydraulic architecture), and biotic stress resistance (Newman et al., 1998; Pulford and Watson, 2003; Liu et al., 2009; Gill and Tuteja, 2010). High throughput RNA sequencing (RNA-seq) provides the opportunity to assess the complex genetic interplay of strategies to achieve these traits by providing a snapshot of expressed RNA within a plant tissue at a given moment (the transcriptome). However, interpretation of transcriptomic data toward revealing common or variant genetic responses between closely related cultivars remains difficult due to the extraordinarily high complexity of the biology involved. Such transcriptomic sequence analysis is still quite distant from the gold-standard of functional analysis through perturbation of gene expression level directly, either by gene knockout, overexpression, or RNAi. Use of these techniques is either not applicable or often ineffective for complex trait analysis in most nonmodel organisms, although a number of studies have made impressive inroads along these lines in (model) woody crops (Rugh et al., 1998; Pilate et al., 2002; Coleman et al., 2008; Wang et al., 2010).

There is a tendency toward functional prediction in genome-wide expression studies. Frustratingly, this is proving somewhat perilous (especially in crop biology), with research suggesting transcript levels are often not directly correlated with protein levels or, importantly, rates of protein function (Greenbaum et al., 2003; Le Roch et al., 2004; Maier et al., 2009; Vogel and Marcotte, 2012). So while RNA-seq remains a powerful tool of choice for contemporary exploration of complex traits at a genetic level, care must be taken when making assumptions regarding the impact of differential expression alone. In this regard, high resolution RNA-seq analysis at such large scales is potentially better suited for hypothesis generation than hypothesis-driven research. Nevertheless, in terms of what could be predicted to be differentially expressed (DE) in leaves of willow due to petroleum hydrocarbon contamination in soil, the following would be expected: general stress responses (potentially including oxidative stress from overproduction of reactive oxygen species; Yurekli and Porgali, 2006), indirect treatment-specific interactions (such as salinity and drought response interactions; Popko et al., 2010; Bauddh and Singh, 2012), and direct responses to petroleum hydrocarbons. In terms of a direct response, there is little evidence suggesting similar organic contaminates are often absorbed and mobilized to above-ground tissue (Alkio et al., 2005; Watts et al., 2006; El Amrani et al., 2015; Shiri et al., 2015) or metabolized to any degree by willow directly. There is, however, a growing body of evidence pertaining to metaorganismal interactions whereby a multitude of organisms collectively exploit these unique environmental conditions (Weyens et al., 2009; Kang et al., 2012; Bell et al., 2014a; Yergeau et al., 2014; Gonzalez et al., 2015).

Current estimates suggest that there are approximately 11 million distinct species that exist globally (Mora et al., 2011); of these, only around the order of 0.001% had been genome sequenced and annotated as of April 2013 (Ellegren, 2014). The estimated proportion of plant species having been sequenced and annotated is of the order of 0.01%. It would therefore seem prudent to consider RNA-seq data interpretation methodologies that are designed to derive useful information from the 99.99% of nonmodel organisms comprising the majority of unknown biological information on earth. There is a growing engagement with the intricate interdependent relationships between multiple species in nature (metaorganismal complexity; Bosch and McFall-Ngai, 2011; Bordenstein and Theis, 2015) in a wide range of biological fields; however, such complexity is challenging (Hanage, 2014). Recent work into the human microbiome is proving an essential element in health research (Gill et al., 2006; Turnbaugh et al., 2007; Nelson et al., 2010; Davids et al., 2016), particularly concerning bacterial diversity within gastrointestinal studies. Such diversity in the microbiome is now well established as an essential factor in root-soil interactions, often termed the rhizosphere (Luo et al., 2009; He et al., 2013; Sullivan et al., 2013; Yergeau et al., 2014; Bell et al., 2015; de Vrieze, 2015). Less work has extended the metaorganism beyond bacteria, or singular target pathogens, herbivores, or symbiotes. Recent evidence, such as the seemingly ubiquitous presence of mites in human epidermis (Thoemmes et al., 2014), opens a door toward a relatively unexplored, more inclusive strategy indicating the value of an organism-blind approach to interpretation of RNA-seq data. This is particularly important for phytoremediation tree systems where rhizospheric bacteria and fungi, as hypothesized from expression profiles, seem essential to understanding organic contamination tolerance (Bell et al., 2014a; Yergeau et al., 2014). Above ground, less research has been conducted with a metaorganismal approach in trees, although endophytes have been demonstrated interacting within the system (Doty et al., 2005; Kang et al., 2012; Khan et al., 2014; Delhomme et al., 2015).

In light of these findings pertaining to the metaorganism and in the absence of strong evidence suggesting that tissue in higher eukaryotes is ever sterile of foreign organisms, it would also seem useful to develop RNA-seq data interpretation methodologies to allow for observation of foreign organism-derived RNA sequences. Here, a strategy for annotating and interpreting RNA-seq in nonmodel plant species without constraint to a single reference genome is assessed that accepts the assumption that foreign organisms will always be present in plant tissue outside of strongly selective laboratory conditions. Three different RNA-seq interpretation approaches are compared: (1) Ten individually assembled de novo transcriptomes with unconstrained annotation (allowing nonnative organisms);

(2) reference genome mapping of each cultivar, which assumes the presence of just a single organism; (3) de novo assembly of a single, global transcriptome (including all 10 cultivars) with unconstrained annotation.

The poorly understood, yet likely multifaceted, crop trait of phytoremediation is used as an exemplar treatment in place of simulated data, as it involves unknown variables and only superficially understood genetic systems.

RESULTS

Hydrocarbon Contamination Phenotype of 10 Willow Cultivars

The contamination concentration in the soil considered contaminated was an average of 837.5 mg kg^{-1} C10-C50, 62.5 mg kg^{-1} PAHs, and 0.2 mg kg^{-1} PCB, whereas noncontaminated soil had no detectable C10-C50, PAH, or PCB. Predicted biomass yields, calculated from plot level harvest yields, had significantly higher biomass yields on the noncontaminated land, ranging from 10 FW t ha^{-1} (S05) to 60 FW t ha^{-1} (SV1). Biomass yields of trees grown on contaminated land from all cultivars were relatively high, varying from 7 FW t ha^{-1} (cultivar SV1) to 18 FW t ha^{-1} (Millbrook; Fig. 1).

Mapping Extracted RNA to Independently de Novo Assembled Transcriptomes

We acquired an average of 303 million paired-end reads per cultivar (a total of 0.6 trillion nucleotides from all 60 trees). The percentage of reads mapping to the independent de novo assemblies ranged from 87% (Fish Creek) to 82% (S05; Fig. 2). These mapping rates were obtained using slightly more stringent alignment criteria than the default for increased confidence. Mapping of the same data to the global transcriptome was on average 10.22% less effective for each cultivar; this is unsurprising given the nature of the assembly process (the drop corresponds to the loss of some cultivar-specific contigs due to cultivar amalgamation). Mapping of RNA-seq data to the *S. purpurea* 94006 reference genome ranged from 87.77% in Fish Creek (which was equivalent to de novo mapping rates) to 45.75% in S44. The two cultivars with the highest genome mapping rates, Fish Creek and Millbrook, are the closest cultivars in terms of species lineage: *S. purpurea* and *S. purpurea* x *S. miyabeana*, respectively. The constructed cultivar phylogeny, from 42 sets of commonly annotated DE contigs from (independently assembled) de novo transcriptomes, was also used to estimate distance of each cultivar from the *S. purpurea* 94006 reference genome (Fig. 2). The distance calculated by the differences between these de novo assembled contigs closely matched the genome mapping efficacy.

Differential Expression of Genes (Independent de Novo)

An average of 314,133 unique contigs were assembled per cultivar (mean N50: 2173), from which an average of 14,139 of those contigs were identified as DE (4.82%) due to contamination, ranging from 24,968 DE contigs in SX67 to 8,602 DE contigs in SX61 (Supplemental Files S1 and S7). Of these DE unique contigs, 82% were best annotated as *Salix* in origin on average, while the rest were best annotated by non-*Salix* organisms or had no confident BLASTx hit in either NCBI nr, SwissProt, TrEMBL, or the *Salix purpurea*

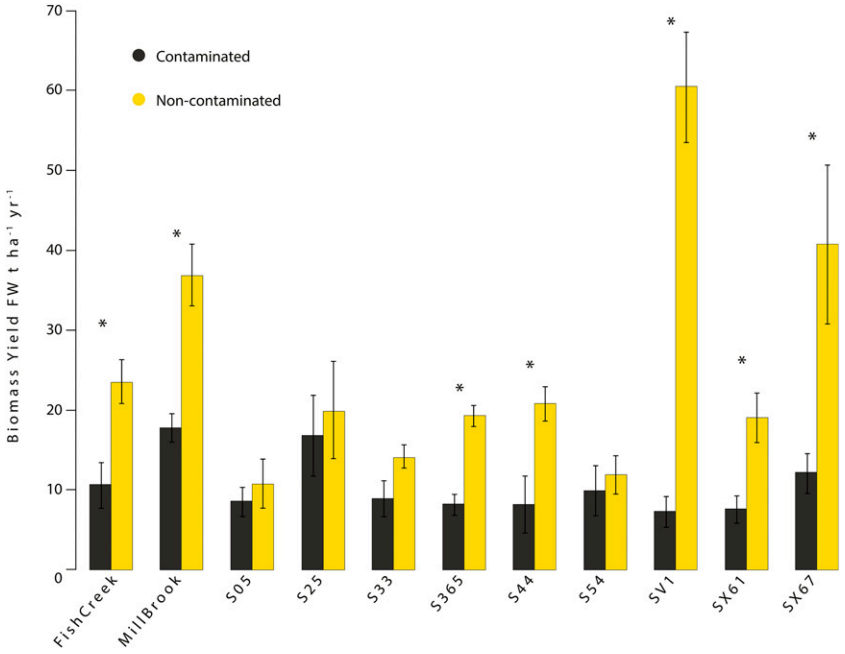

Figure 1. Biomass yields. Biomass yield from for all 10 cultivars grown on either contaminated or noncontaminated land. Mean biomass yields were measured as the total fresh weight of all above-ground harvested biomass in the second year of growth for each of four trees per cultivar per treatment. Yields per hectare were projected based on planting density. *Significant difference between treatment (*t* test $P < 0.05$). Error bars represent SE ($n = 4$ trees).

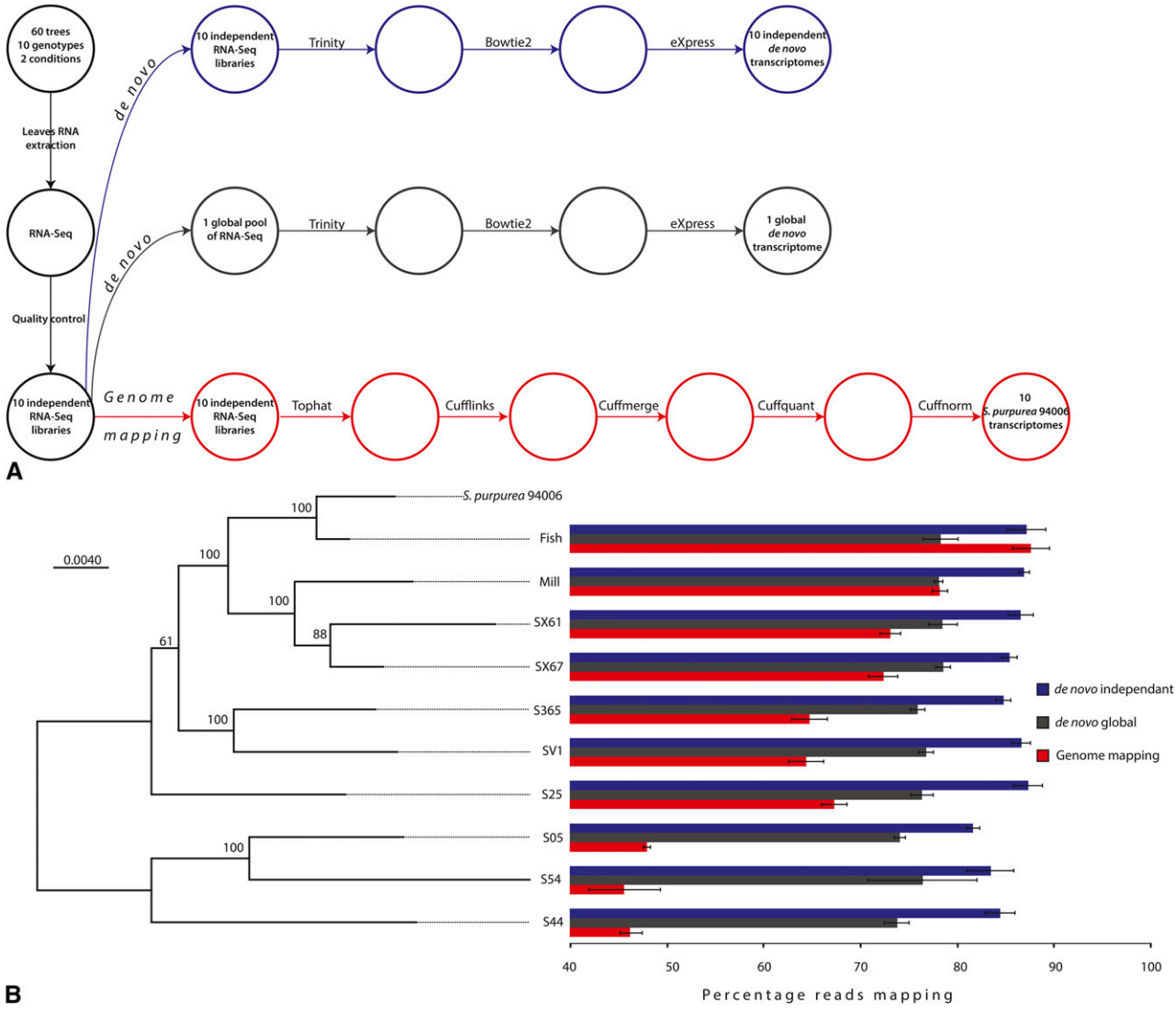

Figure 2. Schematic transcriptomic approaches and mapped efficacy. A, A diagrammatic representation of the different transcriptomic strategies tested: 1. Independent de novo assemblies, 2. a single global assembly (including all 10 cultivars) and standard mapping of reads to a reference genome. B, Left, a phylogeneic tree constructed from individual de novo assembled DE contigs sharing common *S. purpurea* 94006 annotation (Sapur, at the top of the tree, represents the reference *S. purpurea* 94006). Branch length is proportional to sequence divergence, and 1,000 bootstrap replications were performed to estimate percentage branch support. Right, the percentage of illumina sequence reads mapping from each cultivar using the different assembly approaches: independent de novo assemblies, a single global assembly, and mapping to the *S. purpurea* 94006 reference genome. Error bars represent SE ($n = 3$ trees).

94006 genome (no hit, bitscore <50 or e-value $>10^{-4}$, were classified as unknown).

Differential Expression of Contigs Annotated as Plant (Independent de Novo)

An average of 11,636 contigs annotated as *Salix* were identified as DE transcripts per cultivar, with SX67 having the highest number (20,956) and SX61 the lowest (7,349; Supplemental File S1). While this large variation in the number of unique DE transcripts was identified *between cultivars*, very few *Salix* DE transcripts were unique to treatment *within each cultivar* (≤2% in all cases), with variation instead being predominantly in relative abundance. DE Salix transcripts that were uniformly expressed in very high abundance (fragments per kilobase of transcript per million [FPKM]) across a number of cultivars in contaminated trees (in the top 1% in terms of transcript abundance) encoded: thiamine thiazole synthase (TTS; *SapurV1A.0229s0310.x.p*, *SapurV1A.0345s0250.x.p*, *SapurV1A.0722s0220.x.p*), thiamine biosynthesis protein ThiC (*SapurV1A.1685s0060.x.p*, *SapurV1A.0041s0760.x.p*), universal stress A-like protein (*SapurV1A.0088s0440.x.p*, *SapurV1A.0180s0440.x.p*, *SapurV1A.2622s0030.x.p*),

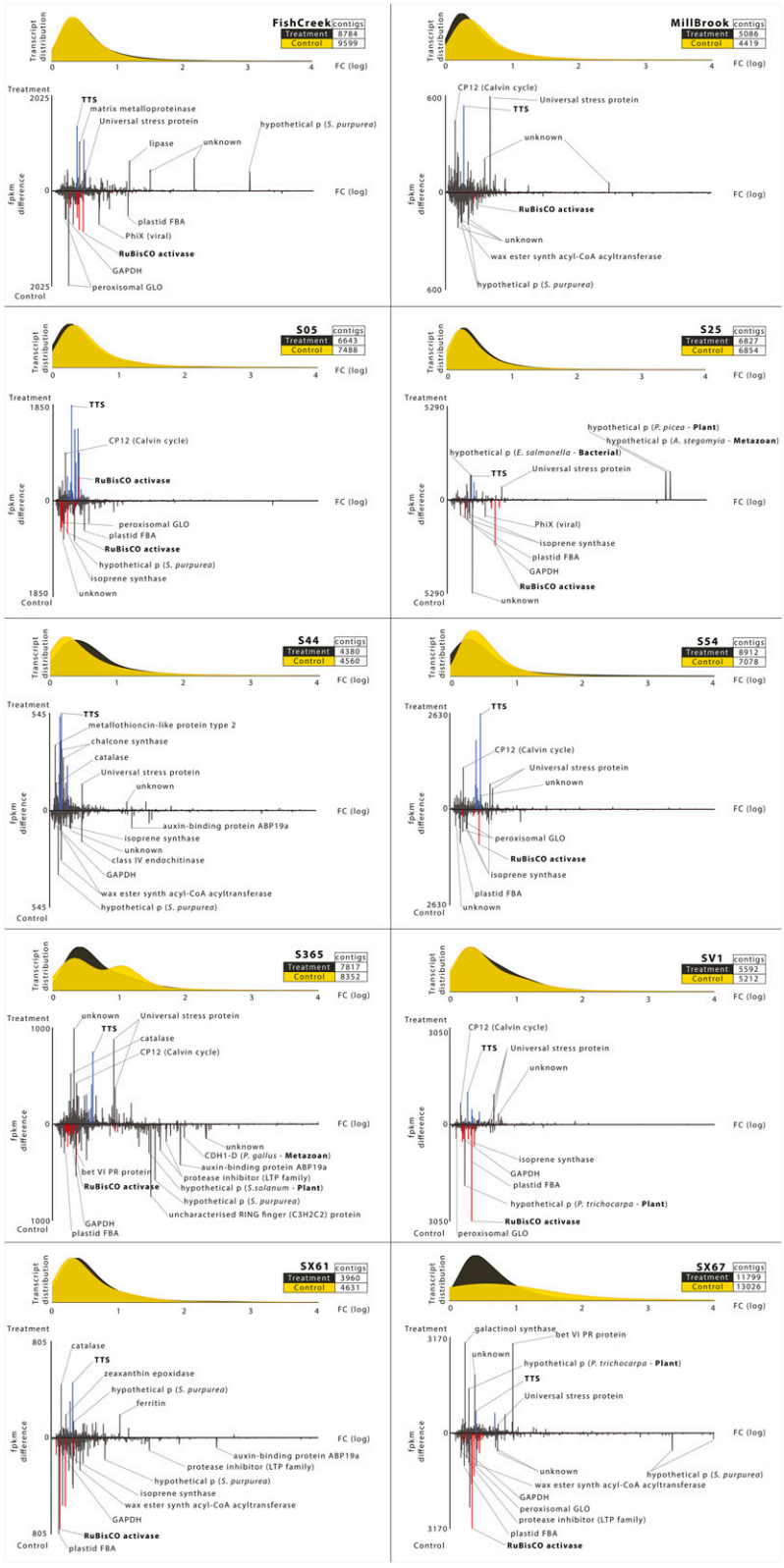

Figure 3. DE gene distribution and abundance (FPKM) weighted fold change. DE transcripts from individual de novo assemblies of each of the 10 cultivars. Top: fold change (FC $-\log_{10}$) distribution of DE genes per treatment. Bottom: individual (normalized mean) transcript counts (FPKM difference) per DE gene are segregated by fold change (for a weighted view of differential expression). Treatment represents contaminated trees, whereas control represents noncontaminated trees: all TTS and Rubisco

and Calvin cycle protein CP12 (*SapurV1A.0158s0210.x.p*, *SapurV1A.0180s0320.x.p*; Fig. 3). As well as these, other consistently high abundance transcripts common in contaminated trees across cultivars encoded RNA binding proteins RBM24 (Splicing factor 3b, subunit 4), RBM42 (Alternative splicing factor SRp20/9G8 [RRM superfamily]), and a light harvesting chlorophyll *a/b* binding protein.

Transcripts encoding Rubisco activase proteins (*SapurV1A.0281s0180.x.p*, *SapurV1A.0214s0360.x.p*) were the most consistently high abundance DE genes in noncontaminated trees across all the cultivars (except S44). Other DE *Salix* transcripts uniformly expressed (in a number of cultivars; Fig. 3) in extraordinary abundance in noncontaminated trees (in the top 1% in terms of transcript abundance) encoded Plastid Fru 1,6-bisphosphate aldolase (plastid FBA; *SapurV1A.0091s0210*), peroxisomal (S)-2-hydroxy-acid/glycolate oxidase (peroxisome GLO; *SapurV1A.0207s0030.x.p*, *SapurV1A.0617s0030.x.p*), a specific glyceraldehyde-3-phosphate dehydrogenase (GAPDH; *SapurV1A.0053s0730.x.p*), and isoprene synthase (*SapurV1A.0312s0290.x.p*).

Differential Expression of Contigs Annotated as Non-Plant (Independent de Novo)

Differentially expressed sequences, annotated as foreign to *Salix* and as non-plant, were present in all cultivars (Fig. 4). A total of 1,283 unique DE contigs was best annotated as deriving from metazoa across all the cultivars encompassing 141 different species of origin. The most abundant metazoan species of origin, in terms of numbers of unique transcripts, was *Drosophila sophophora* with 201 DE unique transcripts identified. Of these 201 sequences, 155 were in greater abundance in noncontaminated trees (transcript abundance, FPKM) whereas only 46 were in greater abundance in contaminated trees, totaling 293.61 and 70.38 FPKM, respectively. One of the other most abundant metazoan species of origin was *Tetranychus urticae*. Differential expression of these sequences, present in the cultivars S365 and SV1, showed the same pattern as *D. sophophora* with 106 transcripts having greater expression in noncontaminated trees, whereas only 9 transcripts had greater expression in contaminated trees, totaling 234.76 and 28.00 FPKM, respectively.

Only 66 unique DE transcripts were best annotated as of bacterial origin and interacting with treatment via differential expression across the 10 cultivars (Supplemental File S1; Fig. 4). These originated from 33 distinct bacteria. Compared to metazoa, this number of unique transcripts was relatively low, yet some were in very high abundance; for example, an uncharacterized *Salmonella enterica* protein was extraordinarily high in contaminated S25 trees (2482.76 FPKM). *Escherichia coli* was the most omnipresent species responding to treatment in terms of differential expression, present in 8 of the 10 cultivars.

The largest kingdom represented in non-plant DE transcripts was fungi. A total of 1,663 unique transcripts was identified as DE, spanning all 10 cultivars and putatively originating from 101 distinct species. The most highly represented species was *Pyrenophora tritici-repentis*, comprising 365 unique DE transcripts over five cultivars (S05, SV1 S365, S44, and SX61). Unlike the arthropod patterning of high expression in noncontaminated trees, all *P. tritici-repentis* annotated transcripts had higher abundance in contaminated trees. Once this global presence of foreign organism-derived RNA is recognized across all the cultivars, it is interesting to analyze the annotated function of DE genes from foreign organisms. One clear example of interest was the large increase in expression of a *Parastagonospora nodorum* TTS gene, parallel in regulation to that of *Salix* TTS in treated trees. The gene in question (SNOG_05965; UniProt unique identifier Q0UQJ9) was highly up-regulated in treated trees of three independently assembled cultivars: S44, S365, and SV1. A high bitscore and poBit provides good confidence in homology of the translated protein hit as well as being the best hit in the queried databases. It is also observable that the sequence was present in trees under both treatments.

To further explore the foreign organism expression and diversity of the system, the complete annotation of all assembled contigs needed to be performed, including those that were not DE; this was achieved through the construction of a global transcriptome.

Differential Expression of Unknown Contigs (Independent de Novo)

An average of 13.03% of DE contigs currently (2015) have no confident annotation (no BLASTx hit: bitscore <50 or e-value $>10^{-4}$) in the major protein repositories (nr, SWISS-Prot, TrEWBL) or the *Salix* genome. While some of these sequences could be artifacts of the de novo assembly process, many are not only identified as DE due to treatment but are some of the most prominent sequences within each cultivar in terms of fold change and abundance (FPKM; Fig. 3). All cultivars had unknown DE sequences in the top 50 most abundant transcripts in contaminated and noncontaminated trees. Because of this extensive scale of expression, acknowledgment, and quantification of these sequences, often discarded out-of-hand during early bioinformatics steps, is technically important in order to prevent issues with library scaling and proportionality. Direct comparison of unknown DE contigs from independent de novo assembly cannot be made with confidence (via annotation) and so is instead performed using the global transcriptome.

Figure 3. (*Continued.*)
activase transcripts are represented in blue and red respectively. The most abundant transcripts within the system are labeled, including annotation species of origin if not *Salix purpurea* 94006. Viral PhiX 174 sequence is the control spike used in Illumia kits. PPDE $\geq$ 0.95.

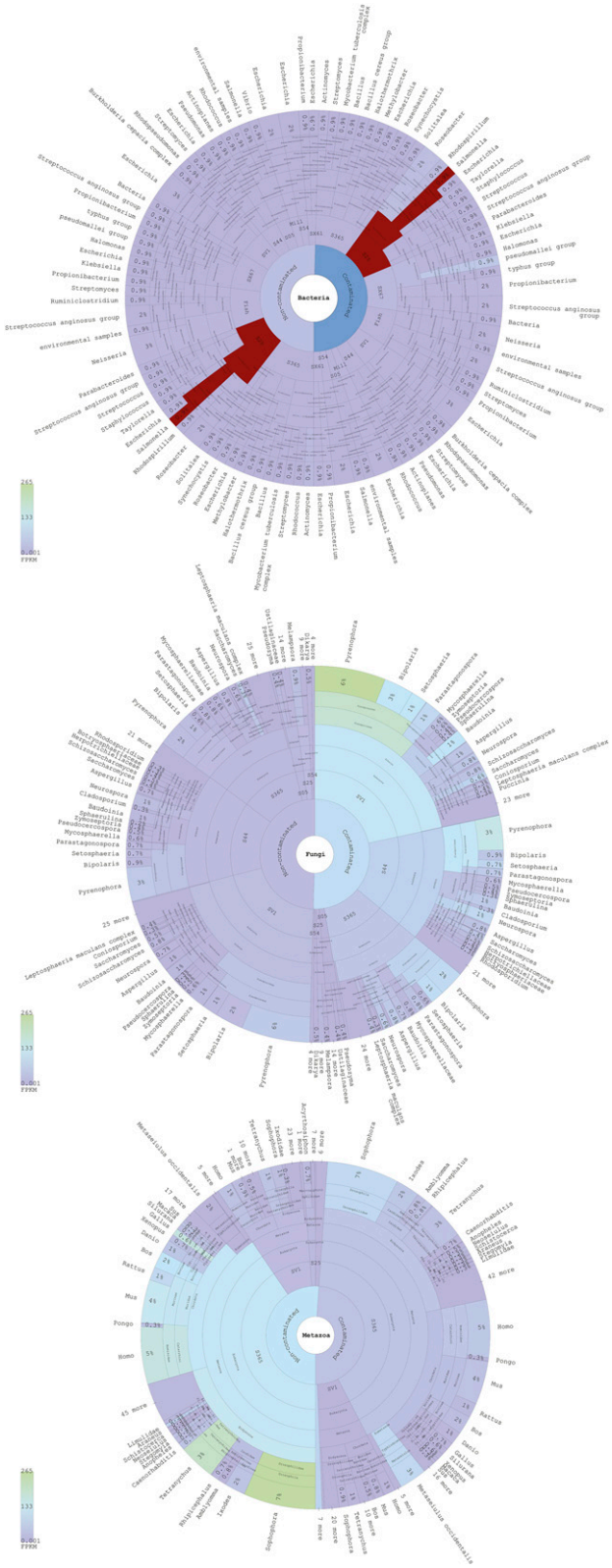

Figure 4. Origin of unique DE contigs (independent de novo assemblies). Krona charts presenting an overview of all transcripts DE in independent de novo assemblies that were annotated as originating from bacteria, fungi, or metazoa. The proportion of each taxonomic grouping is defined by the number of unique transcripts, whereas the color represents the relative abundance (FPKM) of transcripts in each taxon (upper color boundary was limited to improve differentiation by abundance). All 10 cultivars are included and split into noncontaminated (left) and contaminated trees (right). This allows the lack of unique transcript *absence*, but the strong transcript abundance treatment effect, between cultivars to be visualized. Completely interactive charts are available at https://github.com/gonzalezem/Figure4/blob/master/README.md.

Mapping RNA to a Reference Genome

To compare the de novo assembly results, which include identification of foreign organism-derived sequences, we used contemporary genome mapping (align-and-assemble approach) against the *S. purpurea* 94006 reference genome. No foreign organism (or unknown) sequences present within RNA extracted from plant material can be identified as such during the reference genome mapping, but *Salix* genes of sufficient nucleotide homology to the reference map should be identified. An average of 5,657 DE unique transcripts was identified across the cultivars, ranging from 11,276 in SX67 to 2,377 in S44. An average of 60.86% fewer transcripts was identified as DE when compared to de novo assemblies, the most being lost in cultivars S44 and S54. While this could be a product of additional construction of false isoforms, sequence investigation (Supplemental Data S1) indicates a substantial number of highly expressed isoforms are true splice or cultivar variants lost as either unpredicted by splice site analysis or direct sequence diversity from the reference map. Rubisco activase and TTS DE genes, identified by de novo methods as the most extreme in terms of abundance in response to treatment, were successfully recognized using genome mapping (and were of uniform regulation to de novo approaches; Supplemental Files S2 and S8).

Forced Mapping Test

The possibility for the RNA expressed by foreign organisms (present in plant tissue RNA extractions) to map to the reference genome, and therefore be mistakenly characterized as native *S. purpurea* 94006 expression, was tested. Publicly available RNA-seq data from four separate species were mapped directly onto the *S. purpurea* 94006 reference genome, and these comprised: *Homo sapiens* (human), *Carassius auratus* (goldfish), *Pleurocybella porrigens* (angel wing fungi), and *Oryza sativa* (rice). This "forced mapping" process revealed that sequence reads from each of these organisms did indeed map onto the reference genome and would be falsely identified as native to *Salix* using standard genome mapping. Only 0.09% of reads mismapped from human, 0.01% from goldfish, 0.02% from angel wing fungi, and 0.21% from rice (Fig. 5). The impact of these reads could potentially be large because of high mapping events. In an attempt to reduce this

impact (of foreign RNA as a technical confounding variable), a quick and simplified mapping strategy to the *S. purpurea* 94006 transcriptome was devised. The transcriptome mapping reduced mapping events by 95% in human, 50% in goldfish, 50% in angel wing fungi, and 75% in rice.

The *Salix* genes hosting these foreign reads varied in number: 98 (153 transcripts) for human, 59 (86 transcripts) for goldfish, 83 (119 transcripts) for angel wing fungi, and 641 (1082 transcripts) for rice (Supplemental File S5). Out of curiosity, we compared the fate of mismapped reads from each organism and found mapping events in a number of common *Salix* genes (Fig. 5; Supplemental File S4). Reads from all four organisms directly mapped to *Salix* myosin H chain-like protein (*SapurV1A.0019s0450*), actin (*SapurV1A.0231s0320*, *SapurV1A.0655s0050*, *SapurV1A.0018s0700*, *SapurV1A.0251s0180*), α-Tubulin (*SapurV1A.0019s0610*), elongation factor 1-alpha (*SapurV1A.0023s0330*, *SapurV1A.0023s0340*), and polyubiquitin (*SapurV1A.0779s0090*), and would thus be treated as native to *Salix.* Sometimes mapping was within coding regions while others were highly repetitive and within the 5′ untranslated region (Supplemental File S4). The evolutionary distance of these species from *Salix* suggests these highly conserved sequence regions may be ancient. As a different test of this, we repeated the transcriptome mapping using DNAseq from 20,000- to 60,000-year old *Mammuthus primigenius* (woolly mammoth) DNA samples. Woolly mammoth reads mapped to 81 *Salix* genes (118 transcripts); these included 3 of the 5 common mismapped *Salix* genes: actin (*SapurV1A.0018s0700*, *SapurV1A.0251s0180*), myosin H chain-like protein (*SapurV1A.0019s0450*), and elongation factor 1-alpha (*SapurV1A.0023s0330*, *SapurV1A.0023s0340*).

Mapping Extracted RNA to a de Novo Assembled Global Transcriptome

To directly compare de novo assembled contigs between cultivars (as opposed to comparison via annotation), we assembled a single, global transcriptome using all 60 trees and then compared DE genes from each cultivar. This used the total of 3.03 billion paired-end reads (0.6 trillion bases) to assemble 612,041 unique contigs (N50 of 913), which back-mapped an average of 76.83% of reads per cultivar. These were then filtered, removing those with zero abundance to leave 606,880 contigs in total.

Common DE *Transcripts (Global de Novo)*

An average of 10,004 contigs per cultivar was identified as DE due to contamination (an average of 7,662 was annotated as *Salix* transcripts). S44 stood out as the most distinct cultivar in terms of common DE genes, being present in only 51% of the contigs shared by 9/10 cultivars (so, in most cases, the first outlier; Supplemental File S6). Rubisco activase and TTS contigs, the exemplar treatment-related genes investigated in detail here, were prevalent in their commonality throughout the *Salix* cultivars but also in their sheer level of abundance. The construction of a global transcriptome also allows the identification of common DE genes that were better annotated outside of the *Salix* reference genome. Sixty of these contigs were shared by all 10 cultivars (Fig. 6; Supplemental File S6), with 2 TTS isoforms being the most abundant shared transcripts in contaminated trees. By using secondary annotation, included in the unconstrained annotation strategy, it is possible to confirm that the majority of these contigs have no homologous sequence within 10% of poBit of the primary annotation.

Full Annotation (Including non-DE) (Global de Novo)

The global transcriptome assembly had a greatly reduced number of contigs in total, 606,880 contigs vs 3,141,329 for all 10 cultivars assembled independently, because of common genes shared between cultivars. This reduction allowed complete assembly annotation (including non-DE); 359,360 contigs (60%) were confidently annotated. Of the 606,880 global transcriptome contigs, those unique to a cultivar ranged from 1,062 to 3,197 except in those cultivars with a high presence of foreign organisms (primarily fungi and metazoa), which contained a high number of unique contigs: S25 (30,311), S365 (31,830), S44 (12,393), and SV1 (33,192; Figs. 4, 7, 8). Only 118,738 unique contigs were shared as common to all 10 cultivars. An average of 112,548 unique contigs was annotated as *Salix* per cultivar, ranging from 96,851 (Fish Creek) to 131,494 (S365; Supplemental File S3a).

The total of 359,360 unique contigs was annotated from 1,588 species in the global transcriptome. A total of 1,445 non-plant species was identified as being the origin of annotation of 190,224 unique contigs, while 142 distinct (non-*Salix*) plant species were identified as being the origin of annotation of 14,506 unique contigs (shared across all cultivars; Figs. 7 and 8, half being poplar). In total, 66 distinct non-plant species were identified as being the origin of annotation for at least 1,000 unique contigs each (so were very highly represented), while 105 non-plant species were identified as having relatively high expression (>100 FPKM within at least one cultivar of the global transcriptome; Supplemental File S3a). The most prominent annotation species in terms of unique contigs numbers included fungi: *Cryptococcus neoformas* (15,432 unique contigs, syn. *Filobasidiella neoformans*); bacteria: *E. coli* (266); metazoa: *D. melanogaster* (5,305); Amoebozoa: *Dictyostelium discoideum* (1,066, slime mold); Alveolates: *Tetrahymena thermophile* (515, ciliated protist). The non-*Salix* plant contigs represented 8.6% of all the contigs annotated as plant in total (this corresponds to the same proportion of DE plant genes annotated as non-*Salix*). No distance relationship in terms of recognized phylogeny was evident.

Contigs best annotated from the bacteria *Salmonella enterica* were identified by total annotation of the global transcriptome (including non-DE contigs). These contigs were present in every cultivar in both contaminated and noncontaminated trees. Although the number of

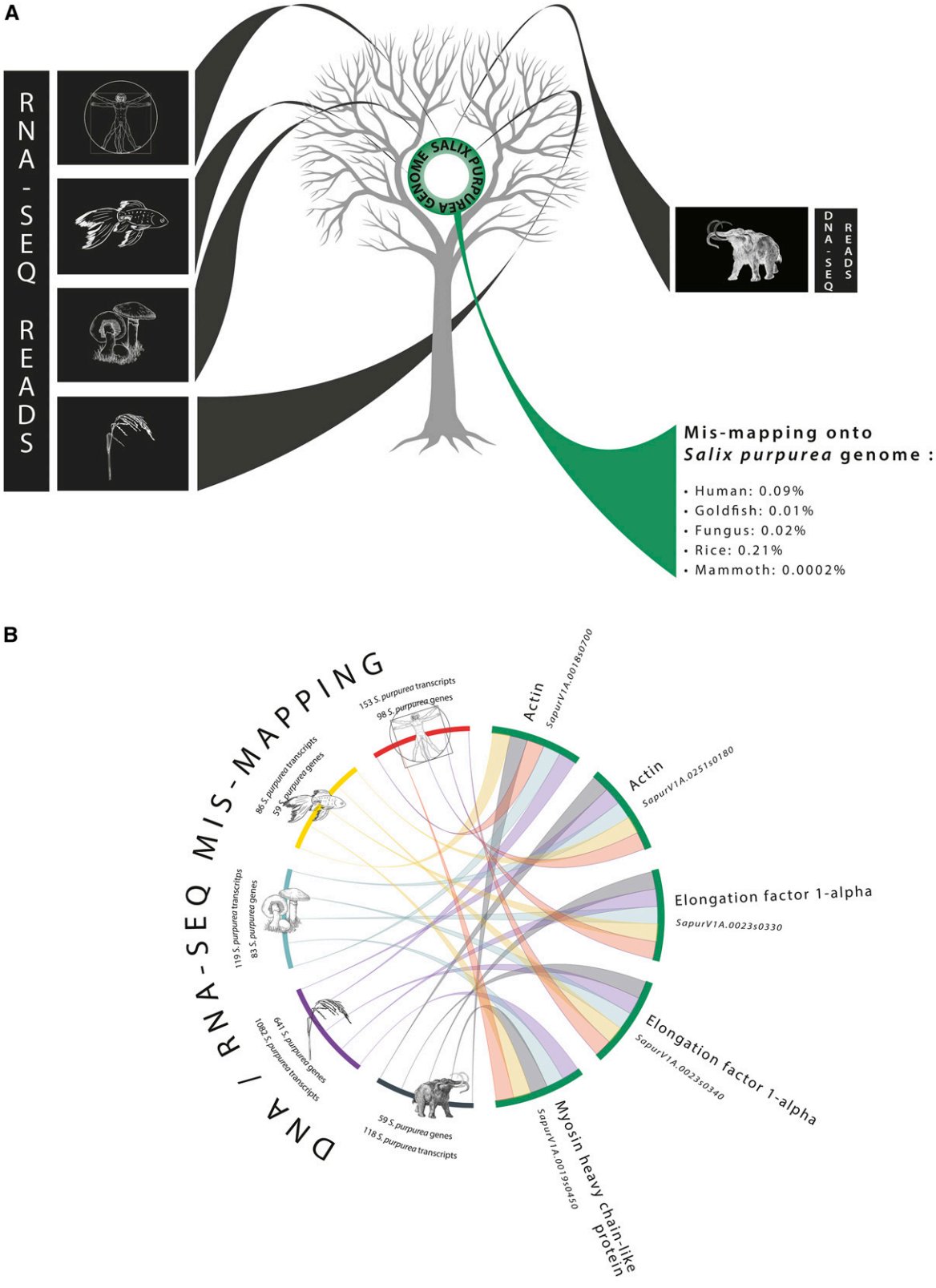

Figure 5. Forced mapping. Forced mapping was performed using the reference *S. purpurea* 94006 genome and the reference *S. purpurea* 94006 transcriptome. The publically available RNA-seq data mapped against the references was derived from RNA extracted from human, goldfish, angel wing fungi, rice, and woolly mammoth. A, When mapped against the *S. purpurea* 94006 genome, *Salix* genes hosted some foreign reads (mismapping) in all cases. B, When reads were mapped against the reference

unique transcripts was very low in all cultivars regardless of treatment, the abundance of transcripts (FPKM) was extraordinarily high (the fourth most abundant transcript origin in every cultivar; Table III; Fig. 7).

DISCUSSION

Hydrocarbon Contamination Effect on Biomass Yields

Reduced biomass yields were observed in contaminated trees, but substantial variation existed in the extent of this reduction in different cultivars. While some cultivars had very high yields on noncontaminated land, no clear relationship between noncontaminated and contaminated yields between cultivars was observed. This suggests elite phytoremediation cultivars most likely need to be directly selected by their response to a specific contaminant as opposed to by previous yield performance outside of phytoremediation systems. Extraordinarily high yields, as projected up from plot level harvests of SV1, are often seen in scientific field trials and are potentially an overestimate of true yields at larger scale cultivation. However, the SV1 cultivar is a native (North American) willow species (*Salix dasyclados*) that has previously been shown to produce very high yields at field scale, often up to 30 t ha^{-1} of oven weight biomass (Kopp et al., 2001; Labrecque and Teodorescu, 2003, 2005). What can be reasoned from these yields, and brought later into interpretation of comparative leaf transcriptome data, is that trees grown on contaminated land may be differentially expressing genes directly reflecting contamination tolerance but may also reflect the indirect contamination effect of the reduced growth phenotype resulting from the whole organism tolerance response.

Mapping Extracted RNA to Independently de Novo Assembled Transcriptomes

The independent cultivar de novo assembly and annotation approach captured data from foreign organism-derived RNA that should not map to the *S. purpurea* 94006 reference genome sequence. We also considered the potential of capturing data in terms of translated protein sequences of sufficient homology to the *S. purpurea* 94006 reference genome or other plant species that potentially would not map directly using conventional genome mapping (as too divergent at a nucleotide level, this is directly assessed during genome mapping). Very high mapping rates were maintained across all cultivars (Fig. 2), and the number of unique DE contigs, predominantly *Salix* but also identified from a diverse array of organisms (Fig. 4), showed no relationship or observable patterning reflective of phylogenetic distance from the most genetically well-characterized willow, *S. purpurea* 94006. The number of contigs identified as DE (approximately 5%) is lower than could be expected based on other studies in similar systems, such as that performed by Pang et al. (2013) of 39.53%, but such rates may be specific to each trait under investigation and the nature of the assembly.

Plant Genes *(Independent de Novo)*

Of the large number of organisms represented in the annotation of DE transcripts, highly characterized model organism plants were overrepresented (Supplemental File S1). It is unlikely that RNA was contaminated with these foreign plants, and we assumed throughout the analysis that these contigs were *Salix* transcripts yet to be characterized as such, or more specifically, were not present in, not annotated, or too divergent from *S. purpurea* 94006 to be recognized. An interesting feature of the additional secondary annotation, which corroborates this, is that when a particular contig is best represented by plant species other than the species of interest (here being *S. purpurea*), there is increased confidence that there was no close *Salix* homolog (Supplemental Files S1, S2 and S3). The dominance of these few, well-characterized model organisms (principally *Populus* and Arabidopsis) is perhaps revealing evidence of how much of this natural metatranscriptomic world is currently unknown and, correspondingly, how many organisms and species are poorly characterized or entirely unknown. These contigs, best annotated by non-*Salix* plants, represented an average of 3.11% of the DE plant transcripts (11.7% in the global transcriptome including non-DE transcripts), a similar proportion of 8% of plant annotated DE genes were identified as non-*Salix* by a similar trial using pot-grown Fish Creek willow (Gonzalez et al., 2015).

Salix Genes (Independent de Novo)

Rubisco activase, GAPDH, and plastid FBA were all strongly up-regulated in noncontaminated trees, suggesting comparatively higher Calvin cycle activity (Somerville et al., 1982; Parry et al., 2008). This is unsurprising given the stressful environment created by petroleum hydrocarbon contamination. Increased FBA

Figure 5. (*Continued.*)
S. purpurea 94006 transcriptome, mismapping genes varied in number: 98 (153 transcripts) in human, 59 (86 transcripts) in goldfish, 83 (119 transcripts) in angel wing fungi, 641 (1082 transcripts) in rice, and 59 (118 transcripts) in woolly mammoth. All five organisms commonly mismapped to two transcripts of *S. purpurea* 94006 actin (*SapurV1A.0018s0700, SapurV1A.0251s0180*), one transcript encoding a myosin H chain-like protein (*SapurV1A.0019s0450*) and two elongation factor 1-alpha transcripts (*SapurV1A.0023s0330, SapurV1A.0023s0340*). The transcript regions of this common mismapping are illustrated in Supplemental File S4.

Genotype number	No. shared DE contigs	% DE higher in non-contaminated trees	% DE higher in contaminated trees	% DE non-uniform DE regulation
X 10	60	46.67	53.33	0.00
X 9	149	35.57	64.43	0.00
X 8	233	39.06	58.80	2.15
X 7	406	37.93	58.37	3.69
X 6	657	34.40	59.51	6.09
X 5	1024	38.18	51.76	10.06

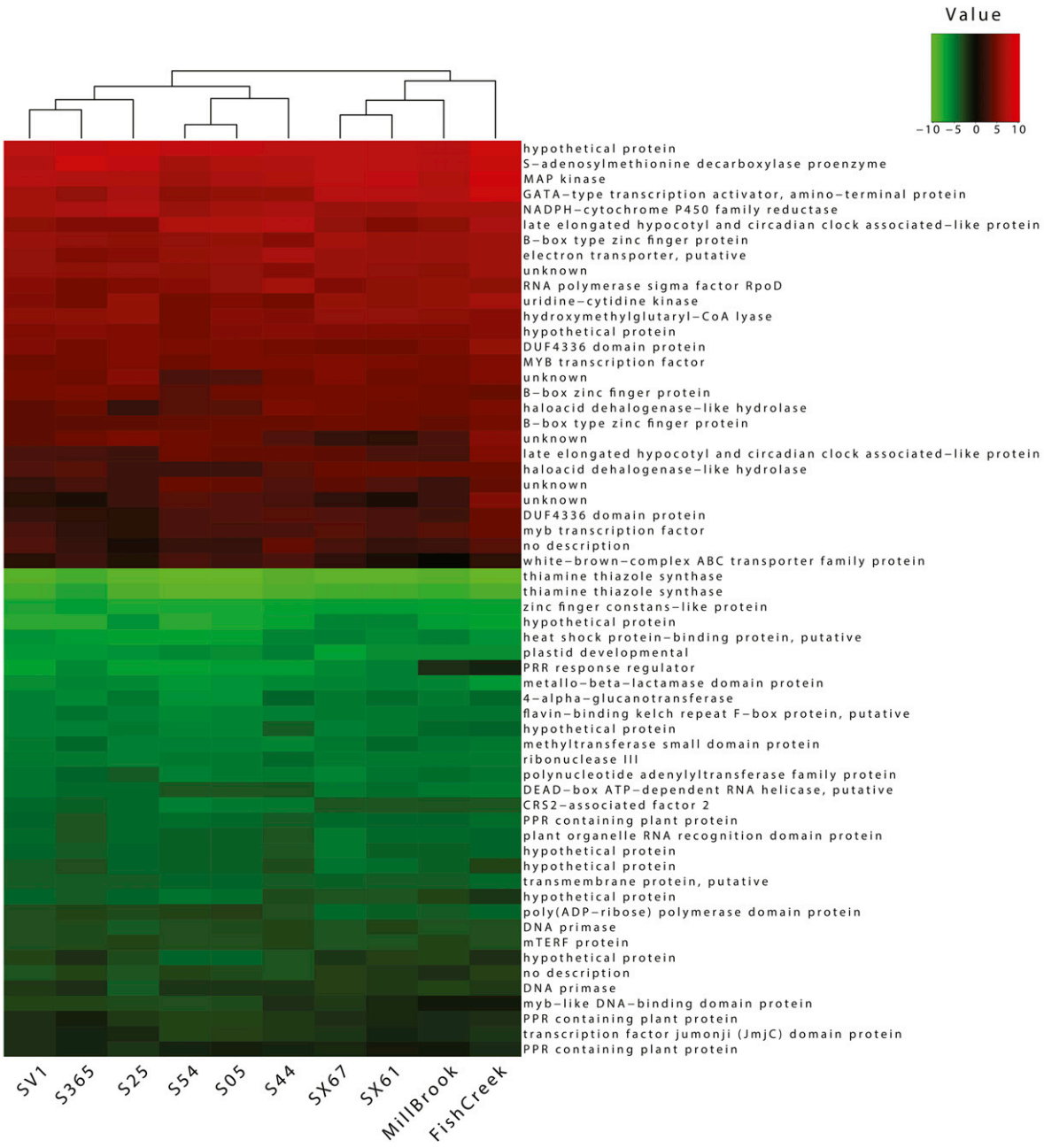

Figure 6. DE genes common to all 10 cultivars. The global transcriptome allows comparison of cultivars directly based on contigs (as opposed to annotation). Sixty DE contigs were shared by all 10 cultivars; as illustrated by the heatmap, all shared uniform regulation in each cultivar. The phylogeny above the heatmap is the same is based on the constructed phylogeny. Green indicates DE contigs with greater transcript abundance in contaminated trees, while red indicates those with greater abundance in

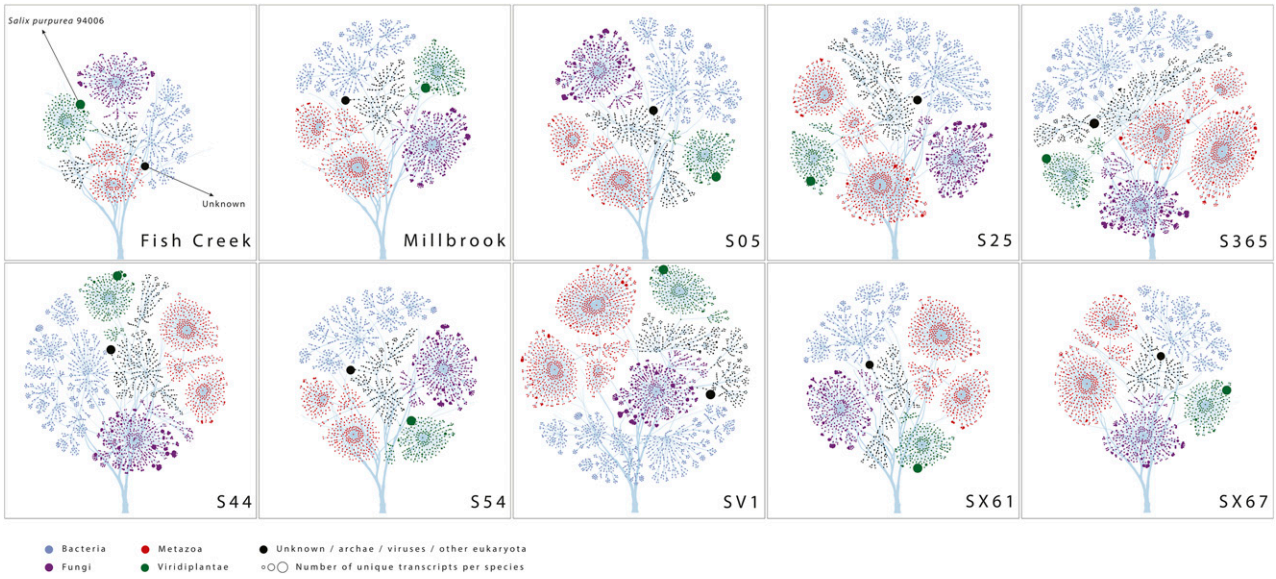

Figure 7. Flower graphs of the global transcriptome separated by cultivar. All transcripts (including those "unknown" in having confident annotation hit) for each of the 10 cultivars within the de novo assembled global transcriptome. The size of the distal node (i.e. species taxon) is proportional to the total number of unique transcripts. Kingdom and phylum wide patterning (consistent across cultivars) can be visualized by color (e.g. fungi consistently represented roughly one-third of unique transcripts across every cultivar). *Salix* RNA was by far the most *abundant* relative transcript amount (as opposed to number of unique transcripts) comprising approximately 90% FPKM. Completely interactive charts for all (and just DE) transcripts, including each taxon name and unique transcript count, are available at https://github.com/gonzalezem/Figure7/blob/master/README.md.

has been shown to enhance both biomass yield and photosynthesis in Arabidopsis (Uematsu et al., 2012). Peroxisomal (S)-2-hydroxy-acid oxidase (GLO1 and GLO4), also in higher abundance in noncontaminated trees, are known to regulate glycolate oxidase protein levels in leaves (Zhang et al., 2012) and, more recently, suppression of glycolate oxidase has been shown to deactivate Rubisco, inhibiting photosynthesis in rice (Lu et al., 2014).

It is well documented that thiamine biosynthesis, consistent with the high abundance of TTS in contaminated trees here, is up-regulated by persistent stress (Rapala-Kozik et al., 2008; Rapala-Kozik et al., 2012), specifically in leaves (Lingua et al., 2012), and that up-regulation of thiazole moiety precursor transcripts (HET-P) can confer improved tolerance to oxidative stress and drought conditions (Rizhsky et al., 2004; Rapala-Kozik et al., 2008). There have also been suggestions that such transcripts may play a role in DNA repair and as a potential signaling molecule for abiotic stress (Goyer, 2010). Importantly, in light of the strong evidence of comparatively reduced Calvin cycle activity in contaminated trees, thiamine diphosphate is integral to *RuBP* regeneration in the Calvin cycle (in terms of gene expression; Lindqvist et al., 1992). Concurrent to this, a single light-harvesting chlorophyll *a*/*b* binding protein transcript was very highly up-regulated in contaminated trees. Previous research has demonstrated that these proteins, in the antenna complexes of the photosynthetic apparatus, can be up-regulated in response to abiotic stress, perhaps due to a vulnerability to oxidative damage (Kurepin et al., 2015). In particular, light-harvesting chlorophyll *a*/*b* binding protein has been implicated as interacting with alterations to redox homeostasis via ABA signaling (Xu et al., 2012a) and in response to lead contamination (Pradeep Kumar et al., 2011). Calvin cycle CP12 (Fig. 3), a small chloroplast protein increasingly recognized as master regulator of the Calvin cycle, complexes to down-regulate GAPDH under oxidative conditions to form the binary complex A4-GAPDH/CP12, which strongly suppresses GAPDH activity (see GAPDH expression above; Wedel et al., 1997; Gontero and Maberly, 2012; Michelet et al., 2013; López-Calcagno et al., 2014) and is essential to the leaves' response to abiotic/oxidative stress (Yoo et al., 2011). The consistent up-regulation of universal stress protein and isoprene synthesis also agrees with the high abiotic stress conditions potentially produced by hydrocarbon contamination (Maqbool

Figure 6. (*Continued.*)
noncontaminated trees. Intensity of color is scaled by $\log_2$ of relative FPKM. Two TTS transcripts where the most abundant shared across all cultivars in contaminated trees, whereas a hypothetical protein (of poor functional characterization) was the highest in noncontaminated trees. Heat maps for contigs that were DE in at least nine and at least eight cultivars are included in Supplemental File S6. PPDE ≥ 0.95

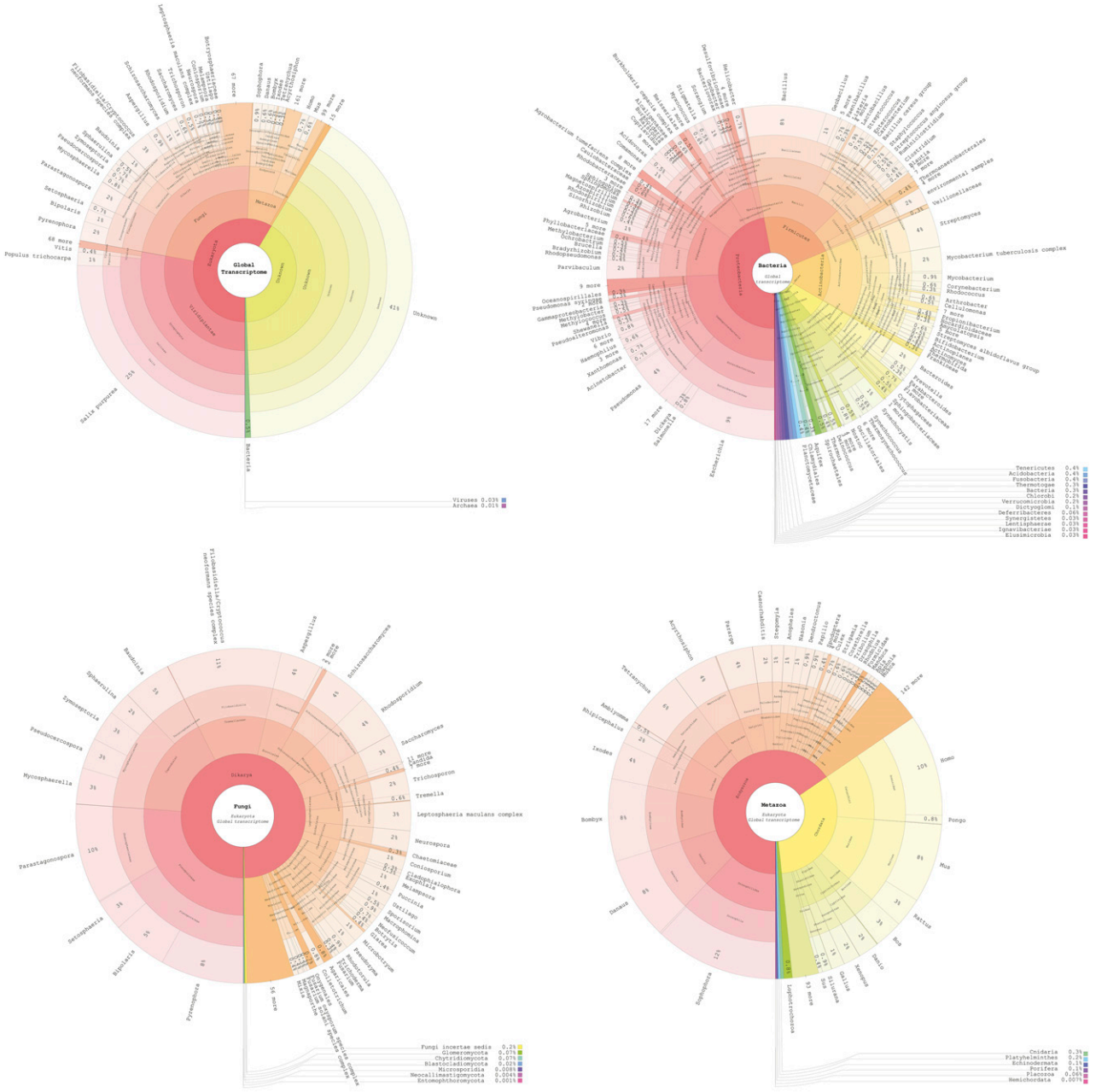

Figure 8. Krona charts of all contigs in the global transcriptome. Krona charts representing the taxonomic origin of annotation of all contigs present in the global transcriptome of all 10 cultivars and including nondifferentially expressed contigs. The proportion of each taxonomic grouping is defined by the number of unique contigs (606,880 in total). Bacteria, fungi, and metazoa are presented from the same charts to illustrate the depth and complexity of life in willow leaves. A completely interactive chart, including unique transcript counts (separated by cultivar) at each taxon, is available at https://github.com/gonzalezem/Figure8/blob/master/README.md.

et al., 2009; Loukehaich et al., 2012) and particularly in response to oxidative stress (Nachin et al., 2005; Vickers et al., 2009).

RNA biding proteins are emerging as regulators of plant responses to environmental stress (Ambrosone et al., 2012). RBM24 (Splicing factor 3b, subunit 4; high homologous to RNA chloroplast RNA binding protein CP29, also a subunit of the antenna complex) was shown to be up-regulated in response to abiotic stress, such as cold stress (Andersson et al., 2001; Yakushevska et al., 2003; Amme et al., 2006). RBM42 (Alternative splicing factor SRp20/9G8, RRM superfamily) has been characterized as a stress responsive spliceosome protein (Cavaloc et al., 1999; Duque, 2011).

The highly consistent pattern, observed repeatedly across multiple cultivars, suggests these are common elements of *Salix*-wide phytoremediation equipment and, as such, could be good expression markers for similar abiotic stress responses.

Non-Plant RNA (Independent de Novo)

One of the assumptions accepted in the design of de novo assembly annotation here was that RNA originating from foreign organisms is likely present in RNA extracted from plant tissue. More precisely in terms of methodological approach, unless sequence origin (beyond the target organism of interest) is permitted and foreign RNA presence is directly assessed, there is a risk that any subsequent data interpretation could be dangerously confounded. These dangers come from both potential technical and biological confounding variables; technical issues relating to proportionality in library normalization and biological uncertainty surrounding presence of foreign organisms (not observable) during experimentation (Thoemmes et al., 2014; Gonzalez et al., 2015). Once the genes available to respond to treatment are acknowledged as variable, the paradigm relating to treatment comparison to control (up- or down-regulation) is undermined. It is therefore helpful to instead view the metagenome as dynamic with respect to treatment. Given this complexity of extra-laboratory biological systems, we considered expression in terms of relative abundance in each treatment.

Contigs best annotated as non-plant RNA were identified in every tree of every cultivar. The ubiquitous presence of bacteria is not surprising within the meta-organism (although often incorrectly considered absent from polyA-enriched mRNA; Cao and Sarkar, 1992; Sarkar, 1996; Kushner, 2004; Slomovic et al., 2005, 2006; Mohanty et al., 2008; Mohanty and Kushner, 2011) considered absent from polyA-enriched mRNA) but does confirm the value of unconstrained annotation. The entirety of the bacterial mRNA present in the tissue is likely not represented because of the polyA enrichment and, while the presence of bacterial sequence within the biological system is of very high certainty here, the polycistronic nature of the transcriptional unit in prokaryotes is currently difficult to assess with operon prediction still in its infancy (Güell et al., 2011; Fortino et al., 2014; Mao et al., 2015). Technology is being developed to more confidently identify the transcriptional unit from de novo assembled contigs, such as Rockhopper (Tjaden, 2015) and Trinity's jaccard-clip (Haas et al., 2013), but is not explored here. High numbers of unique contigs and high abundance of transcripts were observed as DE from fungi and metazoa in a cultivar-specific manner (Fig. 4). This strongly suggests a potential to confound *Salix* expression if observed in isolation (Gonzalez et al., 2015). In terms of direct biological interaction of these foreign organisms with the *Salix* response to contamination, there was extensive and complex cross-talk between these organisms. For example, it is interesting to note that a TTS isoform (*SNOG_05965*) of the fungi *Phaeosphaeria nodorum* was DE and uniformly in higher abundance in contaminated trees in three independent cultivars: S44, S365, and SV1. The role of TTS here, in both *Salix* and *Phaeosphaeria*, is poorly understood in this context. Such data allow simple and promising hypothesis generation for future research. The *E. coli* protein Streptomycin 3′′-adenylyltransferase was also identified as DE and in greater abundance in contaminated trees in six cultivars: Millbrook, S44, S54, SV1, SX61, and SX67. The protein has previously been isolated as present in 3% of organisms from a wastewater metagenomic assembly (Gomez-Alvarez et al., 2012).

The identification of unknown contigs, often DE and in high abundance, was universal to all cultivars here and represented the second largest group of annotation (or lack thereof) after *Salix*. Maintaining (not discarding) such unknown sequence aligns with the understanding that only a small fraction of the genes and isoforms that exist on earth have yet been sequenced. The identification of these sequences may not be of immediate importance (other than the technical advantages of tracking read fate) but, as all sequences are maintained and associated to this large scale field trail, such information may become relevant and even help prevent experimental repetition in the future. Direct comparison of unknown contigs between cultivars was made using the global transcriptome.

Mapping Extracted RNA to a Reference Genome

As the scale of the 10 de novo assemblies was relatively large compared to previous research, parity of results with contemporary genome mapping was tested using a reference genome, *S. purpurea* 94006. Similar results in terms of the exemplar genes, selected as having consistent expression interaction with contamination treatment (across multiple cultivars), were observed as from de novo assembled transcriptomes. Such parity of the major findings also supports the use of de novo assembly and unconstrained annotation methods when a reference genome is not available as well as confirming that previous genome mapping research in the field is not necessarily confounded by foreign organisms or genetic divergence from the reference genome.

The efficacy with which each cultivar mapped to the *S. purpurea* 94006 reference genome followed the phylogenetic relationships predicted from independent de novo assemblies very strongly (Fig. 2; Supplemental Files S1 and S2). This illustrated the extent to which larger proportions of data are lost during genome mapping as cultivars or species become more divergent from the reference genome. The comparative read mapping rates also suggest that these lost data can potentially be retained via de novo assembly. The very substantial shortfall in the percentage of reads mapping in all cultivars other than Fish Creek (the closest relative to the reference genome) is potentially worrying in terms of the biology being overlooked. The separation

of assembly and annotation in a de novo approach allows the high resolution assembly of nucleotide sequences but also a more forgiving, and informative, separate annotation via translated protein sequence. This loss of data, unless a sterilized clone of a reference genome is being sequenced, is important to consider for researchers mapping a nonmodel organism to a related reference genome directly.

The information retention in de novo assembly (where >90% of additional reads mapped compared to standard mapping of the most distant cultivars to the reference genome) was further improved by additional annotation provided by major available protein repositories not present in the reference genome. Often when comparing different genetic backgrounds, the very aim is to identify genetics underpinning variation, identifying only gene expression that is common to the reference genome could limit the identification of these important factors of variation. The use of translated nucleotide sequence instead of nucleotide sequence for annotation (or assembly and annotation) may seem counter-intuitive in terms of confidence in annotation; however, the imperative contemporary need is to exploit RNA-seq technology for the nonmodel organismal world, representing approximately 99.99% of what is currently unknown (Mora et al., 2011; Ellegren, 2014). A long-term benefit of this approach is that annotation can be improved as major repository databases become more populated, but also, in the short term, that differences between *Salix* trees here are captured as opposed to discarded out of hand. Such differences are important to explore if the genetics beyond model organisms are to be understood (and phenotypes improved).

Foreign organism gene expression within the tissue was lost from the genome mapping analysis. Recent research has demonstrated how this foreign organism gene expression can be indicative of fundamentally confounding biological variation. In *Salix*, the potential for strong biological interaction was demonstrated with a crop pest, *Tetranychus urticae* (the two-spotted spider mite), having greater expression in 99% of its genes in trees cultivated on noncontaminated soil than on contaminated soil (Gonzalez et al., 2015). Without the assessment of foreign organism gene expression, the very strong corresponding up-regulation of *Salix* resistance genes in noncontaminated trees would have been attributed erroneously (as down-regulation of resistance genes) to contamination response. Given a system of unknown complexity, it would seem prudent to allow observation of such interactions. Interestingly, one of the cultivars (S365) showed the same pattern of differential spider mite gene expression, with 92% of genes being up-regulated in noncontaminated trees.

As well as foreign organisms having the potential to act as biological confounding variables, the potential for all the identified non-plant RNA to technically interact with genome mapping by being mistakenly mapped (or mismapped) to the *S. purpurea* 94006 reference genome (and therefore adversely interacting with statistical differential expression analysis) was considered.

Forced Mapping

To test whether foreign organism-derived RNA-seq reads can potentially mismap directly to a reference genome, we performed "forced mapping" of non-*Salix* RNA-seq datasets to the salix reference genome. Some foreign RNA did mismap to the *S. purpurea* 94006 reference genome from all organisms tested. It is difficult to quantify the degree to which such mismapping would affect statistical analysis of differential expression in standard genome mapping of RNA extracted from plant tissue; however, while the rates of mismapping were relatively low, the potential for DE data to be compromised is clearly established. Given this potential, the test was repeated using transcriptome mapping designed to deliberately reduce mismapping events. The transcriptome mapping successfully reduced mismapping at the expense of isoform resolution, somewhat less essential in mapping here owing to de facto loss of data resolution in the form of cultivar variants.

RNA-seq has helped reveal a high degree of protein sequences conservation across eukaryotes. For example, Daetwyler et al. (2014) recently identified SMC2 as having conserved sequence in a broad spectrum of eukaryotes. We compared the fate of the mismapped reads from the four external organisms tested (human, goldfish, angel wing fungi, and rice) to see if any *Salix* genes were promiscuous in hosting reads from more than one organism. Five *Salix* genes were identified as common hosts to RNA-seq reads from all four organisms. There were stretches of high conservation in these sequences at the amino acid level between all species; third base codon degeneracy was often present but not great enough to prevent seed mapping of reads (using standard, default alignment criteria).

The same forced mapping was then performed using DNA sequencing of an extinct organism (representing some of the oldest DNA available to us), the woolly mammoth (Lynch et al., 2015), where three of the five ubiquitous genes with regions of high conservation also mismapped. As the number of *Salix* genes hosting reads from each organism (mapped independently) was relatively small, such commonality is rather surprising. Some of those common regions were highly repetitive, but others encode protein regions that have seemingly changed little for over 500 million years; it is likely they play important roles for cell and organism integrity as a whole. Many more additional mismaping genes were

Table I. *Cultivar information (Lauron-Moreau et al., 2015)*

Fish Creek	*S. purpurea*
SX67	*S. miyabeana*
SX61	*S. miyabeana*
S05 (clone: 5005)	*S. nigra*
S25	*S. eriocephala*
S365	*S. caprea*
SV1	*S.* × *dasyclados*
S54	*S. acutifolia "wild"*
S44 (clone:5044)	*S. alba*
Millbrook	*S. purpurea* × *S. miyabeana*

Table II. *De novo assembly information*

	Individual Transcriptomes										Global Transcriptome
Cultivar	Fish	Mill	S05	S25	S365	S44	S54	SV1	SX61	SX67	All cultivars
Number of reads	563M	599M	576M	632M	566M	589M	578M	558M	617M	590M	6,052M
Contigs assembled	229,892	284,102	345,260	280,401	408,700	391,709	307,747	407,182	240,622	245,704	612,041
Number of Trinity genes	63,434	78,211	106,237	109,119	206,887	154,119	93,073	198,086	74,223	72,740	512,657
N50	2,456	2,147	1,628	2,656	2,063	2,040	1,831	2,206	2,376	2,323	913

shared between four of the five organisms. All of the ubiquitously common genes (as well as the majority of genes common to 4/5 organisms) detected as highly likely to mismap between eukaryotic life were involved in the cytoskeleton. When mismapping reads were blasted back against their respective genome sequences, gene homologs were usually returned. For example, the human reads mismapping *Salix* elongation factor 1 alpha returned the human homolog (EEF1A2, nr). However, in the case of highly repetitive sequence ($[tc]_{14}[ac]_{26}$) of the *Salix* myosin H chain-like protein (within the 5′ untranslated region and so not present in the *Salix* duplicate gene), some mismapping human reads seemingly come from sequence inside a membrane-associated guanylate kinase (MAGUK), interestingly characterized by an ability to form protein-protein interactions with cytoskeleton proteins and microtubule/actin machinery (specifically, *DLG2*, which has structural and functional roles associated with the cortical actin cytoskeleton; Handa et al., 2007). As housekeeping genes, such as these (Nicot et al., 2005), are often used as comparative reference controls for techniques such as qPCR, we recommend care be taken to ensure primers are designed within regions unique to an organism of interest. This potential for mismapping is important given that fungi seem present in the majority of plant tissue, less so for goldfish and woolly mammoths (unlikely to be present in the leaves of trees).

Mapping Extracted RNA to a de Novo Assembled Global Transcriptome

The ability to directly compare contigs at a sequence level using the de novo assembled global transcriptome, as opposed to comparison via annotation, improved confidence that foreign plant annotated contigs do represent currently uncharacterized *Salix* genes (as of August 2015), as many examples were shared throughout *Salix* but are not present in the reference genome (Supplemental File S3b). Commonly DE contigs confirmed that the high abundance contigs annotated with the same *S. purpurea* 94006 genes in independent de novo assembly, and highlighted as genes of interest, shared sequence similarity as well as annotation. Interestingly, all DE contigs shared by every cultivar were commonly regulated in relation to contamination (Fig. 6) and also included a hypothetical protein (poorly characterized) as the most abundant in noncontaminated trees shared across all cultivars (deserving further study). The secondary annotation (retention of other very good sequence annotation hits) was useful in navigating the large number of contigs that were distinct from the *S. purpurea* 94006 reference genome. The most abundant DE genes that were conserved between cultivars (yet distant from the reference genome) and annotated as plants other than *Salix* derived from poplar, Arabidopsis, and common grape vine. Unknown DE contigs represented an average of 13% of all DE contigs in independent cultivar de novo assemblies; these could be artifacts of the de novo assembly process or genuine uncharacterized genes. We directly compared unknown DE contigs across the cultivars using the global transcriptome. Those shared between all or multiple cultivars (Fig. 6; Supplemental File S3b) and of uniform expression are likely genuine uncharacterized genes as opposed to artifacts, and remain valuable candidates for future investigation and functional characterization.

*Total Annotation of the Global Transcriptome (Including non-*DE *Contigs)*

We found no published examples of RNA-seq of any tree organ being free of RNA from foreign organisms if such a finding was permitted (Doty, 2008, 2009; Bosch and McFall-Ngai, 2011; Bell et al., 2014b; Khan et al., 2014). Whether such large amounts of foreign organism-derived RNA is in great enough abundance to act as a confounding variable in terms of technical quantification of RNA from the organism of interest, or in terms of the biological question being posed, is currently impossible to know without first performing some degree of analysis allowing observation (Thoemmes et al., 2014). We therefore suggest the necessity of performing at least cursory de novo assembly and unconstrained annotation before mapping RNA-seq data to a reference genome in experimental

Table III. *Salmonella enterica FPKM and unique transcripts in each cultivar*

Genotype	Number of Unique Transcripts	Total FPKM Noncontaminated	Total FPKM Contaminated
Fish Creek	5	4,171	12,303
Millbrook	5	4,573	8,194
S05	10	8,605	8,901
S25	6	2,451	7,131
S365	17	3,826	8,271
S44	12	2,478	10,892
S54	5	936	7,783
SV1	16	3,211	5,135
SX61	7	5,942	11,689
SX67	5	2,143	12,831

systems save those in model organisms grown under very high selection in controlled laboratory conditions.

All contigs in the system were treated as unknown, with annotation of assembled sequence only representing the best information currently available for describing unknown sequence. In terms of confident presence or absence of a given organism within the system, confidence can be somewhat improved by the high number of independent contigs annotated by a given species. Many of these organisms are considered tree symbionts or pathogens and were identified as being the origin of annotation of over 1,000 unique contigs (Figs. 7 and 8), including organisms such as *T. uricae* (spider mite; Grbić et al., 2011), *Melampsora larici-populina* (poplar leaf rust fungi; Rinaldi et al., 2007), *Rhodotorula graminis* (pigmented yeast; Khan et al., 2012), and *Acanthamoeba castellanii* (Herdler et al., 2008). The highly complex cross-talk indicated by such expression is best assessed through differential expression analysis because of the scale of information and uncertainty present within the system. Very high numbers of unique contigs were assembled and best annotated as *Homo sapiens* (human) or *Mus musculus* (mouse) in every tree (representing the second and third most abundant metazoa after *D. melanogaster*). It is important to note that this could be contamination of each tree sample but, given the general overrepresentation of extensively studied organisms in the system, more likely demonstrates the extent of uncharacterized genes or organisms within the biosphere (Ekblom and Galindo, 2011).

One of the complications that arises when foreign organism RNA is acknowledged in RNA-seq data is a change to the paradigm of up- or down-regulation with respect to control, as an organism with the capacity to alter expression can be present or absent in either a control or treatment group. This presence or absence could be a biological interaction with treatment in itself (Gonzalez et al., 2015) and so reveals new difficulty in interpretation of differential expression. The complete annotation of the global transcriptome allowed preliminary investigation into whether differential expression was driven by presence/absence of contigs in each treatment group or instead reflected an expression interaction with treatment. By not treating the genome's capacity to respond as static, the system may potentially better reflect the natural, nonlaboratory world; however, it is noted that there was a surprisingly small difference in unique contigs between treatments *within* any cultivar (Figs. 3 and 4). Variation was very much in the levels of expression as opposed to the potential for expression due to absence of foreign genes within a given treatment (as reflected in *T. urticae* differential expression here and in other recent research that also explored tissue variation; Gonzalez et al., 2015). There were, however, large differences between cultivars in terms of the unique contigs present. In light of these findings, it seems likely that each tissue, within each different cultivar, acts as a unique ecological environment providing variant niches, thus resulting in large changes to species population demography. This agrees more closely with the niche theory of the shaping of metaorganismal communities than the more stochastic, neutral theory (Smillie et al., 2011; Jeraldo et al., 2012).

CONCLUSIONS

From the data generated here and available literature, we suggest that all major plant organs may contain foreign organisms and foreign organism-derived RNA outside of artificially controlled laboratory experiments. Further to this, for transcriptomic studies, a de novo assembly and metatranscriptomic annotation step should be conducted before the possibility of fundamentally confounding variables, in the form of foreign organism-derived RNA, can be discounted or sufficiently controlled for. The metatranscriptomic approach reveals it can be useful to consider each plant tissue, of each specific cultivar, as a potentially unique environment or habitat, resulting in a diverse local population of organisms potentially spanning all domains and numerous kingdoms; the metagenome being dynamic with respect to gene expression in the organism as a whole.

We also put forward the explanation that low read mapping rates often observed in crop transcriptomic studies are likely due to two factors: first, foreign organism-derived RNA (as well as uncharacterized sequence) that mismap at a low rate to the organism of interest and second, mapping to a reference genome may be difficult in nonmodel organisms due to the divergence of the accession of interest to the reference genome unless clonal.

Extensive variation was present in *Salix* gene expression between cultivars; however, some clear commonalities in gene expression were shared between all, suggesting the global toolkit of phytoremediation machinery necessary for tolerance to petroleum hydrocarbon induced stress in *Salix* includes consistent alteration of photosynthesis regulation and protection of photosynthetic equipment from oxidative stress. Conversely, such commonalities are unlikely to drive the natural variation observed between cultivars, variation that is independent of the biomass yields achieved when cultivated under more favorable conditions.

RNA-seq provides an incredibly powerful tool to uncover that which is currently obscure in the biological world; efforts should be made to ensure the close integration and iterative interaction of classical biology and the rapidly advancing field of bioinformatics, both of which are mutually dependent and necessary for future headway into what seems increasingly obvious is a metaorganismal world.

MATERIALS AND METHODS

Site, Cultivar, and Contamination Composition

The trial was established on the site of a former petrochemical plant in Varennes, southern Quebec, Canada (45° 46′ N, 73° 22′ W). The site included one area contaminated with petroleum hydrocarbons (C10-C50s at an average of

837.5 mg kg^{-1}, PAHs 62.5 mg kg^{-1}, and PCBs 0.2 mg kg^{-1}) and one that was noncontaminated (C10-C50s <100 mg kg^{-1}, PAHs <0.1 mg kg^{-1}, and PCBs < 0.017 mg kg^{-1}; Yergeau et al., 2014; Gonzalez et al., 2015; Grenier et al., 2015). Field design was previously published (Bell et al., 2014a; Hassan et al., 2014; Grenier et al., 2015) with each area covering 300 m^2 with 75 2-year-old trees per cultivar. Cuttings were established at a density of 30,000 ha^{-1}, consisting of rows planted 1 m apart and trees within rows planted 30 cm from each other. Only 10 of 11 cultivars were sampled for RNA extraction (the site also included one non-planted control; Table I). Ten leaves (between the fifth and fifteenth fully unfurled leaves) from the tip of the tallest stem were harvested from four trees and flash frozen as individual tree replicates for RNA extraction. Soil samples were analyzed using ICP-MS by AGAT Laboratories (Montreal, QC).

RNA Extraction and Illumina Sequencing

RNA was extracted using a modified CTAB protocol (Chang et al., 1993; Gambino et al., 2008) with RNA quantity and quality assessed with a Bio-Analyser (Agilent). After initial characterization of RNA quality, only the three best extractions per cultivar per treatment were sequenced. Genome Quebec Innovation Centre performed library construction (TrueSEquation 100 bp paired-ends libraries, Illumina TruSeq RNA Sample Preparation Kit). PolyA containing mRNA was purified using polyT magnetic beads before random hexamer pairing for the cDNA synthesis. The samples were sequenced (four per lane) using an Illumina HiSEquation 2000 sequencing system. Sequencing information and quality control data are provided in Supplemental File S7 (raw data are provided in Supplemental Data S1). The viral sequence phiX174 is used as a spike control within the Illumina HiSEquation 2000 sequencing system; this was discarded from data interpretation.

De Novo Assembly

Data were filtered using Trimmomatic (Lohse et al., 2012) to trim poor quality nucleotides at the beginning and the end of each sequence. Reads shorter than 40 bp after quality control were removed from the pool. Reads were assembled de novo individually by cultivar and into a single global transcriptome (Table II; Fig. 2) using Trinity software with default parameters (Haas et al., 2013). Transcripts shorter than 200 bp were discarded. Sequences qualified as a Trinity "gene" were the union of transcripts similar enough to be considered by Trinity as putative isoforms of the same gene.

Bowtie2 software (Langmead, 2010; Langmead and Salzberg, 2012) was used to map RNA-seq reads back to de novo transcriptomes with the additionally stringent alignment criteria to improve confidence (these included enabling rejection of discordant alignment and mixed alignment; Supplemental File S7). This yielded an average mapping efficiency of 86% of reads for individual de novo transcriptome assemblies and 77% for the global transcriptome (Fig. 2). This is a lower mapping efficiency than can be obtained using default parameters (tested at an average of 93% read mapping for cultivar S365; Supplemental File S7). Raw and normalized transcript abundance was calculated using eXpress (Roberts and Pachter, 2013) with default parameters. EBSeq (Leng et al., 2013) was used to identify DE transcripts between the two experimental conditions in each cultivar. EBSeq can be less prone to adjust perceived outliers or to discard data due to FDR control than some programs (Soneson and Delorenzi, 2013). Significance is identified and expressed as posterior probability of differential expression (PPDE) ≥0.95 (Leng et al., 2013; Supplemental Files S1–S3, S6). RT-qPCR was not used to validate gene expression due to the prerequisite for highly characterized sequence data (Unamba et al., 2015). A number of criteria need to be met for qPCR reference genes to be regarded as reliable (Chervoneva et al., 2010), importantly, traditional "housekeeping" genes (such as β-actin and 18S) have been extensively acknowledged as unstable in numerous biological systems (Gorzelniak et al., 2001; Solanas et al., 2001; Glare et al., 2002; Raaijmakers et al., 2002; Brunner et al., 2004; Gonçalves et al., 2005; Nicot et al., 2005; González-Verdejo et al., 2008; Barsalobres-Cavallari et al., 2009; Paolacci et al., 2009; Xu et al., 2012b; Jiang et al., 2014; Llanos et al., 2015). While many of these cited studies establish expression stability of more appropriate nontraditional reference genes, the scale and complexity of multiple nonmodel plant cultivars (in this extra-laboratory research) makes such establishment problematic without using RNA-Seq itself to establish transcriptome-wide expression.

Phylogenetic relationships (Fig. 2) were estimated using DE contigs from independent de novo assemblies that shared the same *S. purpurea* 94006 annotation (in eight or more cultivars). All retained contigs were aligned by Muscle (Edgar, 2004) in Geneious (Kearse et al., 2012) and then concatenated in a single alignment. Regions of alignment with >30% missing data as well as ambiguously aligned regions (that represented <0.4% of the total alignment) were removed. The resulting alignment of 66,535 nucleotides was analyzed with phyml vers. 3 (Guindon et al., 2010) with 1,000 bootstrap replications to estimate percentage branch support.

Annotation

The nonmodel organism metatranscriptomic (*unconstrained*) annotation strategy, which queries a broad range of protein sequence repositories, was performed as outlined by Gonzalez et al. (Gonzalez et al., 2015). Briefly, the de novo assembled contigs were annotated using three major protein databases (nr, SwissProt, and TrEMBL) as well as the *S. purpurea* 94006 reference genome. A novel method for selecting annotation from BLASTx returns was used.

The annotation selection procedure aims at improving homology inference compared to selections based simply on an Expect-value (e-value) and/or a score for similar sequence (bitscore), which can, in some cases, lead to a poor choice of the best hit for a given query. Specifically, BLAST was not designed to calculate protein homology but uses a heuristic method to produce an e-value and bitscore; what can be derived from BLAST output has been usefully discussed by Pearson and Sierk (2005): "if a similarity score is not random, then the sequences must be not unrelated." In other terms, every alignment that passes a reasonable e-value test denotes statistically significant sequence similarity, suggesting the sequences are related. E-values $< 10^{-6}$ for nucleotide BLAST (BLASTn, megaBLAST) and e-value $< 10^{-3}$ for protein BLAST (BLASTp, BLASTx; Altschul et al., 1990) are often considered appropriate statistical cut-offs in annotation strategies. However, it seems then unreasonable to pick e-value, a true statistical value, as an indicator to differentiate our best BLAST hits. Moreover, e-value is database (or library) size dependent (Karlin and Altschul, 1990) and an e-value threshold becomes less reliable as the size of the database decreases (Wood-Charlson et al., 2015). Normalized scores (or bitscores), derived from substitution scoring matrices, are library size independent (Karlin and Altschul, 1990) and thus directly comparable. Due to the relatively complex theory behind heuristic methods, BLAST-ranked output results may be seen as a windfall; however, a number of articles urge biologists to treat BLAST hits with caution, such as that by Pertsemlidis and Fondon (2001): "Although normalized scores allow comparison of the results of searches using different scoring systems, they are an extreme reduction of the rich information available in an alignment."

Given multiple high scoring alignments for a single sequence (generated due to the complexity present in nonsimulated biological data), all statistically characterized as nonrandom, a method to select a best alignment is necessary. BLAST returns hits sorted by lowest e-value and, for returns with common e-value, sorted by highest bitscore. While this is often not a problem for very high scoring hits, for sequences that are less well represented in a database (common when nonmodel organisms are investigated), the best alignments can potentially be lost. We developed a method that allows us to confidently choose the best alignment beyond highest bitscore alone.

Percentage Optimal Bitscore

The scoring matrix BLOSUM62, which is the default in BLAST, was chosen as a good compromise for scoring protein sequences of unknown divergence in the attempt to consider the samples without any prejudice regarding species diversity. Bitscores <50 are generally considered very unreliable (Roux et al., 2013; Wright et al., 2014) so were removed and an e-value requirement of $<10^{-4}$ was applied. Hits with similar bitscores do not necessarily align to the same part of the protein or have similar alignment length, underlining the risk in considering them as "similar." We believe this apparent pitfall, sometimes made due to a misconception regarding the nature of bitscore in BLAST output (Pertsemlidis and Fondon, 2001; Pearson and Sierk, 2005), could be avoided by considering an *optimal bitscore*, representing the highest possible bitscore generated by a given alignment. Simply aligning the part of the protein involved in the alignment with itself gives the optimal bitscore (previous scoring parameters are maintained).

The actual bitscores obtained can then be compared to the optimal bitscore to yield a percentage: bitscore / optimal bitscore * 100. We called this percentage optimal bitscore (poBit; Supplemental File S7). Because there is an inherent bias favoring short alignments, a weighted value for each annotation was assigned, or a *confidence coefficient*, based on the highest scoring alignment. For a set of alignments for a given unique contig, we defined the confidence coefficient as bitscore / highest alignment bitscore. By weighting the poBit with the confidence coefficient, we obtain a *corrected* poBit able to differentiate annotation hits

of similar bitscores with increased confidence. The highest poBit reveals the best scoring alignment for a given contig, hence is considered as the best annotation provided by BLAST (Supplemental File S7, example table).

Using the poBit filter for annotation, a substantial number of DE transcripts, often present across multiple of our *Salix* cultivars and of uniform regulation, were best annotated by non-*Salix* plant species (Supplemental File S1). These sequences, because of their consistent assembly, annotation, and differential expression between multiple independent cultivars, likely represent a large number of uncharacterized *Salix* genes of potential influence in the advanced stress tolerance mechanisms present in willow. The capacity to annotate these genes greatly increased the number of DE isoforms available for downstream hypothesis generation relating to the biological trait of interest. Separate but related to this, we extended the methodological approach (driven toward the biological unknown and acknowledging the uncertainty present in these bioinformatics methods) by also retaining those BLAST hits that were not selected but have a high comparable poBit (within 10%). By including this extra information within a *secondary* set of annotation (Supplemental Files S1–S3), a biologist can gain confidence in a particular annotation as well as, importantly, reminding the biologist that BLAST annotation of an unknown sequence is, in almost all cases, uncertain. The number of primary annotation hits in the global transcriptome was 359,360, while the number of secondary annotation hits was 6,002,308.

Mapping to a Reference Genome

We also mapped reads to the *S. purpurea* 94006 reference genome (*Salix purpurea* v1.0, DOE-JGI, http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Spurpurea). The Tuxedo protocol (Trapnell et al., 2012) was used to assemble the transcriptome from each cultivar against the reference genome: Tophat, Cufflinks, Cuffmerge, Cuffquant, Cuffnorm, Cuffdiff, and CummeRbund were all executed with default parameters. To compare these results with the de novo transcriptome assembly approach, we also calculated differential expression using EBSeq with abundance extracted from Cuffdiff output.

Forced mapping, the mapping of RNA known to be foreign to a reference map, was performed using the reference *Salix purpurea* 94006 genome (using Tophat as above) and reference *Salix purpurea* 94006 transcriptome (using Bowtie2 as above). Publically available non-*Salix* RNA-seq data used for forced mapping were acquired from *Homo sapiens* (human, EBI: PRJEB6971); *Pleurocybella porrigens* (angel wing fungi, EBI: DRR003995; EBI: DRR003996; Suzuki et al., 2013); *Carassius auratus* (goldfish, EBI: DRR014529; Abe et al., 2014); *Oryza sativa* (rice, EBI: SRR611648); and *Mammuthus primigenius* (woolly mammoth; DNA-seq, EBI: SRR2009641; EBI SRR2009644; Lynch et al., 2015).

Image and Data Analysis

Custom scripting (in Python, R, Shell, Javascript) and Krona (Ondov et al., 2011) were used to generate images and figures as well as to navigate and query transcriptomic data.

Accession Numbers

Sequence data from this article can be found in the at the ENA online repository (PRJEB11540, ena-STUDY-IRBV). See Supplemental Data S1.

Supplemental Data

The following supplemental materials are available.

Supplemental File S1. Independent de novo data.

Supplemental File S2. Reference genome mapping data.

Supplemental File S3.Global de novo transcriptome data: a) separated cultivars; b) shared contigs.

Supplemental File S4. Common mis-mapping alignments.

Supplemental File S5. All mis-mapping genes.

Supplemental File S6. Common genes in 8 and 9 cultivars, heatmap.

Supplemental File S7. Sequencing data and quality control data.

Supplemental File S8. Reference genome mapping: FPKM weighted fold-change.

Supplemental Data S1. Raw data available at the ENA online repository (PRJEB11540, ena-STUDY-IRBV).

ACKNOWLEDGMENTS

We are grateful to Pétromont Inc. for allowing access to the Varennes site. We thank the Genome Quebec Innovation Centre for support and Calcul Quebec for computing resources.

Received January 22, 2016; accepted March 20, 2016; published May 2, 2016.

LITERATURE CITED

Abe G, Lee SH, Chang M, Liu SC, Tsai HY, Ota KG (2014) The origin of the bifurcated axial skeletal system in the twin-tail goldfish. Nat Commun **5:** 3360

Alkio M, Tabuchi TM, Wang X, Colón-Carmona A (2005) Stress responses to polycyclic aromatic hydrocarbons in Arabidopsis include growth inhibition and hypersensitive response-like symptoms. J Exp Bot **56:** 2983–2994

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

Ambrosone A, Costa A, Leone A, Grillo S (2012) Beyond transcription: RNA-binding proteins as emerging regulators of plant response to environmental constraints. Plant Sci **182:** 12–18

Amme S, Matros A, Schlesier B, Mock HP (2006) Proteome analysis of cold stress response in Arabidopsis thaliana using DIGE-technology. J Exp Bot **57:** 1537–1546

Anderson GQA, Fergusson MJ (2006) Energy from biomass in the UK: sources, processes and biodiversity implications. Ibis **148:** 180–183

Andersson J, Walters RG, Horton P, Jansson S (2001) Antisense inhibition of the photosynthetic antenna proteins CP29 and CP26: implications for the mechanism of protective energy dissipation. Plant Cell **13:** 1193–1204

Barsalobres-Cavallari CF, Severino FE, Maluf MP, Maia IG (2009) Identification of suitable internal control genes for expression studies in Coffea arabica under different experimental conditions. BMC Mol Biol **10:** 1

Bauddh K, Singh RP (2012) Growth, tolerance efficiency and phytoremediation potential of Ricinus communis (L.) and Brassica juncea (L.) in salinity and drought affected cadmium contaminated soil. Ecotoxicol Environ Saf **85:** 13–22

Bell TH, Cloutier-Hurteau B, Al-Otaibi F, Turmel M-C, Yergeau E, Courchesne F, St-Arnaud M (2015) Early rhizosphere microbiome composition is related to the growth and Zn uptake of willows introduced to a former landfill. Environmen Microbiol **17:** 3025–3038

Bell TH, El-Din Hassan S, Lauron-Moreau A, Al-Otaibi F, Hijri M, Yergeau E, St-Arnaud M (2014a) Linkage between bacterial and fungal rhizosphere communities in hydrocarbon-contaminated soils is related to plant phylogeny. ISME J **8:** 331–343

Bell TH, Joly S, Pitre FE, Yergeau E (2014b) Increasing phytoremediation efficiency and reliability using novel omics approaches. Trends Biotechnol **32:** 271–280

Black MJ, Whittaker C, Hosseini SA, Diaz-Chavez R, Woods J, Murphy RJ (2011) Life Cycle Assessment and sustainability methodologies for assessing industrial crops, processes and end products. Ind Crops Prod **34:** 1332–1339

Bollmark L, Sennerby-Forsse L, Ericsson T (1999) Seasonal dynamics and effects of nitrogen supply rate on nitrogen and carbohydrate reserves in cutting-derived Salix viminalis plants. Can J For Res **29:** 85–94

Bordenstein SR, Theis KR (2015) Host biology in light of the microbiome: ten principles of holobionts and hologenomes. PLoS Biol **13:** e1002226

Bosch TCG, McFall-Ngai MJ (2011) Metaorganisms as the new frontier. Zoology (Jena) **114:** 185–190

Brereton NJB, Pitre FE, Shield I, Hanley SJ, Ray MJ, Murphy RJ, Karp A (2014) Insights into nitrogen allocation and recycling from nitrogen elemental analysis and 15N isotope labelling in 14 genotypes of willow. Tree Physiol **34:** 1252–1262

Brunner AM, Yakovlev IA, Strauss SH (2004) Validating internal controls for quantitative plant gene expression studies. BMC Plant Biol **4:** 14

Cao GJ, Sarkar N (1992) Identification of the gene for an Escherichia coli poly(A) polymerase. Proc Natl Acad Sci USA **89:** 10380–10384

Cavaloc Y, Bourgeois CF, Kister L, Stévenin J (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA **5:** 468–483

Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep **11:** 113–116

Chervoneva I, Li Y, Schulz S, Croker S, Wilson C, Waldman SA, Hyslop T (2010) Selection of optimal reference genes for normalization in quantitative RT-PCR. BMC Bioinformatics **11:** 253

Coleman HD, Park JY, Nair R, Chapple C, Mansfield SD (2008) RNAi-mediated suppression of p-coumaroyl-CoA 3′-hydroxylase in hybrid poplar impacts lignin deposition and soluble secondary metabolism. Proc Natl Acad Sci USA **105:** 4501–4506

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet **46:** 858–865

Davids M, Hugenholtz F, Martins Dos Santos V, Smidt H, Kleerebezem M, Schaap PJ (2016) Functional profiling of unfamiliar microbial communities using a validated de novo assembly metatranscriptome pipeline. PLoS One **11:** e0146423

de Vrieze J (2015) The littlest farmhands. Science **349:** 680–683

Delhomme N, Sundström G, Zamani N, Lantz H, Lin YC, Hvidsten TR, Höppner MP, Jern P, Van de Peer Y, Lundeberg J, Grabherr MG, Street NR (2015) Serendipitous meta-transcriptomics: the fungal community of Norway spruce (Picea abies). PLoS One **10:** e0139080

Doty SL (2008) Enhancing phytoremediation through the use of transgenics and endophytes. New Phytol **179:** 318–333

Doty SL, Dosher MR, Singleton GL, Moore AL, Van Aken B, Stettler RF, Strand SE, Gordon MP (2005) Identification of an endophytic Rhizobium in stems of Populus. Symbiosis **39:** 27–35

Doty SL, Oakley B, Xin G, Kang JW, Singleton G, Khan Z, Vajzovic A, Staley JT (2009) Diazotrophic endophytes of native black cottonwood and willow. Symbiosis **47:** 23–33

Duque P (2011) A role for SR proteins in plant stress responses. Plant Signal Behav **6:** 49–54

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:** 1792–1797

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb) **107:** 1–15

El Amrani A, Dumas AS, Wick LY, Yergeau E, Berthomé R (2015) "Omics" Insights into PAH Degradation toward Improved Green Remediation Biotechnologies. Environ Sci Technol **49:** 11281–11291

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol **29:** 51–63

Fortino V, Smolander OP, Auvinen P, Tagliaferri R, Greco D (2014) Transcriptome dynamics-based operon prediction in prokaryotes. BMC Bioinformatics **15:** 145

Gambino G, Perrone I, Gribaudo I (2008) A rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochem Anal **19:** 520–525

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE (2006) Metagenomic analysis of the human distal gut microbiome. Science **312:** 1355–1359

Gill SS, Tuteja N (2010) Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. Plant Physiol Biochem **48:** 909–930

Glare EM, Divjak M, Bailey MJ, Walters EH (2002) beta-Actin and GAPDH housekeeping gene expression in asthmatic airways is variable and not suitable for normalising mRNA levels. Thorax **57:** 765–770

Gomez-Alvarez V, Revetta RP, Santo Domingo JW (2012) Metagenome analyses of corroded concrete wastewater pipe biofilms reveal a complex microbial system. BMC Microbiol **12:** 122

Gonçalves S, Cairney J, Maroco J, Oliveira MM, Miguel C (2005) Evaluation of control transcripts in real-time RT-PCR expression analysis during maritime pine embryogenesis. Planta **222:** 556–563

Gontero B, Maberly SC (2012) An intrinsically disordered protein, CP12: jack of all trades and master of the Calvin cycle. Biochem Soc Trans **40:** 995–999

Gonzalez E, Brereton NJB, Marleau J, Guidi Nissim W, Labrecque M, Pitre FE, Joly S (2015) Meta-transcriptomics indicates biotic cross-tolerance in willow trees cultivated on petroleum hydrocarbon contaminated soil. BMC Plant Biol **15:** 246

González-Verdejo CI, Die JV, Nadal S, Jiménez-Marín A, Moreno MT, Román B (2008) Selection of housekeeping genes for normalization by real-time RT-PCR: analysis of Or-MYB1 gene expression in Orobanche ramosa development. Anal Biochem **379:** 176–181

Gorzelniak K, Janke J, Engeli S, Sharma AM (2001) Validation of endogenous controls for gene expression studies in human adipocytes and preadipocytes. Horm Metab Res **33:** 625–627

Goyer A (2010) Thiamine in plants: aspects of its metabolism and functions. Phytochemistry **71:** 1615–1624

Graham-Rowe D (2011) Agriculture: beyond food versus fuel. Nature **474:** S6–S8

Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, et al (2011) The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature **479:** 487–492

Greenbaum D, Colangelo C, Williams K, Gerstein M (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol **4:** 117

Grenier V, Pitre FE, Guidi Nissim W, Labrecque M (2015) Genotypic differences explain most of the response of willow cultivars to petroleum contaminated soil. Trees **15:** 871–881

Güell M, Yus E, Lluch-Senar M, Serrano L (2011) Bacterial transcriptomics: what is beyond the RNA horiz-ome? Nat Rev Microbiol **9:** 658–669

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59:** 307–321

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc **8:** 1494–1512

Hanage WP (2014) Microbiology: microbiome science needs a healthy dose of scepticism. Nature **512:** 247–248

Handa K, Yugawa T, Narisawa-Saito M, Ohno S, Fujita M, Kiyono T (2007) E6AP-dependent degradation of DLG4/PSD95 by high-risk human papillomavirus type 18 E6 protein. J Virol **81:** 1379–1389

Hassan SED, Bell TH, Stefani FOP, Denis D, Hijri M, St-Arnaud M (2014) Contrasting the community structure of arbuscular mycorrhizal fungi from hydrocarbon-contaminated and uncontaminated soils following willow (Salix spp. L.) planting. PLoS One **9:** e102838

Hasselgren K (1999) Utilization of sewage sludge in short-rotation energy forestry: a pilot study. Waste Manag Res **17:** 251–262

Haughton AJ, Bond AJ, Lovett AA, Dockerty T, Sunnenberg G, Clark SJ, Bohan DA, Sage RB, et al (2009) A novel, integrated approach to assessing social, economic and environmental implications of changing rural land-use: a case study of perennial biomass crops. J Appl Ecol **46:** 315–322

He J, Li H, Luo J, Ma C, Li S, Qu L, Gai Y, Jiang X, Janz D, Polle A, Tyree M, Luo ZB (2013) A transcriptomic network underlies microstructural and physiological responses to cadmium in Populus x canescens. Plant Physiol **162:** 424–439

Herdler S, Kreuzer K, Scheu S, Bonkowskia M (2008) Interactions between arbuscular mycorrhizal fungi (Glomus intraradices, Glomeromycota) and amoebae (Acanthamoeba castellanii, Protozoa) in the rhizosphere of rice (Oryza sativa). Soil Biol Biochem **40:** 660–668

Huang HJ, Ramaswamy S, Al-Dajani W, Tschirner U, Cairncross RA (2009) Effect of biomass species and plant size on cellulosic ethanol: a comparative process and economic analysis. Biomass Bioenergy **33:** 234–246

Jeraldo P, Sipos M, Chia N, Brulc JM, Dhillon AS, Konkel ME, Larson CL, Nelson KE, Qu A, Schook LB, et al (2012) Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. Proc Natl Acad Sci USA **109:** 9692–9698

Jiang Q, Wang F, Li MY, Ma J, Tan GF, Xiong AS (2014) Selection of suitable reference genes for qPCR normalization under abiotic stresses in Oenanthe javanica (BI.) DC. PLoS One **9:** e92262

Kang JW, Khan Z, Doty SL (2012) Biodegradation of trichloroethylene by an endophyte of hybrid poplar. Appl Environ Microbiol **78:** 3504–3507

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA **87:** 2264–2268

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28:** 1647–1649

Khan Z, Guelich G, Phan H, Redman R, Doty S (2012) Bacterial and yeast endophytes from poplar and willow promote growth in crop plants and grasses. ISRN Agronomy **2012:** 11

Khan Z, Roman D, Kintz T, delas Alas M, Yap R, Doty S (2014) Degradation, phytoprotection and phytoremediation of phenanthrene by endophyte Pseudomonas putida, PD1. Environ Sci Technol **48:** 12221–12228

Kopp RF, Abrahamson LP, White EH, Volk TA, Nowak CA, Fillhart RC (2001) Willow biomass production during ten successive annual harvests. Biomass Bioenergy **20:** 1–7

Kurepin LV, Ivanov AG, Zaman M, Pharis RP, Allakhverdiev SI, Hurry V, Hüner NP (2015) Stress-related hormones and glycinebetaine interplay in protection of photosynthesis under abiotic stress conditions. Photosynth Res **126:** 221–235

Kushner SR (2004) mRNA decay in prokaryotes and eukaryotes: different approaches to a similar problem. IUBMB Life **56:** 585–594

Kuzovkina YA, Quigley MF (2005) Willows beyond wetlands: uses of Salix L. species for environmental projects. Water Air Soil Pollut **162:** 183–204

Kuzovkina YA, Volk TA (2009) The characterization of willow (Salix L.) varieties for use in ecological engineering applications: co-ordination of structure, function and autecology. Ecol Eng **35:** 1178–1189

Labrecque M, Teodorescu T, Daigle S (1995) Effect of wastewater sludge on growth and heavy metal bioaccumulation of two Salix species. Plant Soil **171:** 303–316

Labrecque M, Teodorescu TI (2003) High biomass yield achieved by Salix clones in SRIC following two 3-year coppice rotations on abandoned farmland in southern Quebec, Canada. Biomass Bioenergy **25:** 135–146

Labrecque M, Teodorescu TI (2005) Field performance and biomass production of 12 willow and poplar clones in short-rotation coppice in southern Quebec (Canada). Biomass Bioenergy **29:** 1–9

Langmead B (2010) Aligning short sequencing reads with Bowtie. *In* editorial board, A D Baxevanis, et al, Current Protocols in Bioinformatics. doi: 10.1002/0471250953.bi1107s32

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods **9:** 357–359

Lauron-Moreau A, Pitre FE, Argus GW, Labrecque M, Brouillet L (2015) Phylogenetic relationships of American willows (Salix L., Salicaceae). PLoS One **10:** e0121965

Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ, et al (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Res **14:** 2308–2318

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics **29:** 1035–1043

Lindqvist Y, Schneider G, Ermler U, Sundström M (1992) Three-dimensional structure of transketolase, a thiamine diphosphate dependent enzyme, at 2.5 A resolution. EMBO J **11:** 2373–2379

Lingua G, Bona E, Todeschini V, Cattaneo C, Marsano F, Berta G, Cavaletto M (2012) Effects of heavy metals and arbuscular mycorrhiza on the leaf proteome of a selected poplar clone: a time course analysis. PLoS One **7:** e38662

Liu H, Weisman D, Ye YB, Cui B, Huang YH, Colon-Carmona A, Wang ZH (2009) An oxidative stress response to polycyclic aromatic hydrocarbon exposure is rapid and complex in Arabidopsis thaliana. Plant Sci **176:** 375–382

Llanos A, François JM, Parrou JL (2015) Tracking the best reference genes for RT-qPCR data normalization in filamentous fungi. BMC Genomics **16:** 71

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res **40:** W622–627

López-Calcagno PE, Howard TP, Raines CA (2014) The CP12 protein family: a thioredoxin-mediated metabolic switch? Front Plant Sci **5:** 9

Loukehaich R, Wang T, Ouyang B, Ziaf K, Li H, Zhang J, Lu Y, Ye Z (2012) SpUSP, an annexin-interacting universal stress protein, enhances drought tolerance in tomato. J Exp Bot **63:** 5593–5606

Lu Y, Li Y, Yang Q, Zhang Z, Chen Y, Zhang S, Peng XX (2014) Suppression of glycolate oxidase causes glyoxylate accumulation that inhibits photosynthesis through deactivating Rubisco in rice. Physiol Plant **150:** 463–476

Luo ZB, Janz D, Jiang X, Göbel C, Wildhagen H, Tan Y, Rennenberg H, Feussner I, Polle A (2009) Upgrading root physiology for stress tolerance by ectomycorrhizas: insights from metabolite and transcriptional profiling into reprogramming for stress anticipation. Plant Physiol **151:** 1902–1917

Lynch VJ, Bedoya-Reina OC, Ratan A, Sulak M, Drautz-Moses DI, Perry GH, Miller W, Schuster SC (2015) Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the arctic. Cell Reports **12:** 217–228

Maier T, Güell M, Serrano L (2009) Correlation of mRNA and protein in complex biological samples. FEBS Lett **583:** 3966–3973

Mao X, Ma Q, Liu B, Chen X, Zhang H, Xu Y (2015) Revisiting operons: an analysis of the landscape of transcriptional units in E. coli. BMC Bioinformatics **16:** 356

Maqbool A, Zahur M, Husnain T, Riazuddin S (2009) GUSP1 and GUSP2, two drought-responsive genes in Gossypium arboreum have homology to universal stress proteins. Plant Mol Biol Rep **27:** 109–114

Michelet L, Zaffagnini M, Morisse S, Sparla F, Pérez-Pérez ME, Francia F, Danon A, Marchand CH, Fermani S, Trost P, et al (2013) Redox regulation of the Calvin-Benson cycle: something old, something new. Front Plant Sci **4:** 470

Mohanty BK, Giladi H, Maples VF, Kushner SR (2008) Analysis of RNA decay, processing, and polyadenylation in Escherichia coli and other prokaryotes. RNA turnover in bacteria. Archaea Organelles **447:** 3–29

Mohanty BK, Kushner SR (2011) Bacterial/archaeal/organellar polyadenylation. Wiley Interdiscip Rev RNA **2:** 256–276

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean? PLoS Biol **9:** e1001127

Murphy R, Woods J, Black M, McManus M (2011) Global developments in the competition for land from biofuels. Food Policy **36:** S52–S61

Nachin L, Nannmark U, Nyström T (2005) Differential roles of the universal stress proteins of Escherichia coli in oxidative stress resistance, adhesion, and motility. J Bacteriol **187:** 6265–6272

Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, et al; Human Microbiome Jumpstart Reference Strains Consortium (2010) A catalog of reference genomes from the human microbiome. Science **328:** 994–999

Newman LA, Doty SL, Gery KL, Heilman PE, Muiznieks I, Shang TQ, Siemieniec ST, Strand SE, et al (1998) Phytoremediation of organic contaminants: a review of phytoremediation research at the University of Washington. J Soil Contam **7:** 531–542

Nicot N, Hausman JF, Hoffmann L, Evers D (2005) Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. J Exp Bot **56:** 2907–2914

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinformatics **12:** 385

Pang T, Ye CY, Xia X, Yin W (2013) De novo sequencing and transcriptome analysis of the desert shrub, Ammopiptanthus mongolicus, during cold acclimation using Illumina/Solexa. BMC Genomics **14:** 488

Paolacci AR, Tanzarella OA, Porceddu E, Ciaffi M (2009) Identification and validation of reference genes for quantitative RT-PCR normalization in wheat. BMC Mol Biol **10:** 11

Parry MAJ, Keys AJ, Madgwick PJ, Carmo-Silva AE, Andralojc PJ (2008) Rubisco regulation: a role for inhibitors. J Exp Bot **59:** 1569–1580

Pearson WR, Sierk ML (2005) The limits of protein sequence comparison? Curr Opin Struct Biol **15:** 254–260

Pertsemlidis A, Fondon III JW (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol **2:** S2002

Pilate G, Guiney E, Holt K, Petit-Conil M, Lapierre C, Leplé JC, Pollet B, Mila I, Webster EA, Marstorp HG, et al (2002) Field and pulping performances of transgenic trees with altered lignification. Nat Biotechnol **20:** 607–612

Pitre FE, Teodorescu TI, Labrecque M (2010) Brownfield phytoremediation of heavy Metals using Brassica and Salix supplemented with EDTA: results of the first growing season. J Environ Sci Engineering **4:** 51–59

Popko J, Hänsch R, Mendel RR, Polle A, Teichmann T (2010) The role of abscisic acid and auxin in the response of poplar to abiotic stress. Plant Biol (Stuttg) **12:** 242–258

Pradeep Kumar S, Arun Mozhi Varman P, Ranjitha Kumari BD (2011) Identification of differentially expressed proteins in response to Pb stress in Catharanthus roseus. Afr J Environ Sci Technol **5:** 689–699

Pulford ID, Watson C (2003) Phytoremediation of heavy metal-contaminated land by trees: a review. Environ Int **29:** 529–540

Raaijmakers MHGP, van Emst L, de Witte T, Mensink E, Raymakers RAP (2002) Quantitative assessment of gene expression in highly purified hematopoietic cells using real-time reverse transcriptase polymerase chain reaction. Exp Hematol **30:** 481–487

Rapala-Kozik M, Kowalska E, Ostrowska K (2008) Modulation of thiamine metabolism in Zea mays seedlings under conditions of abiotic stress. J Exp Bot **59:** 4133–4143

Rapala-Kozik M, Wolak N, Kujda M, Banas AK (2012) The upregulation of thiamine (vitamin B1) biosynthesis in Arabidopsis thaliana seedlings under salt and osmotic stress conditions is mediated by abscisic acid at the early stages of this stress response. BMC Plant Biol **12:** 2

Ray M, Brereton N, Shield I, Karp A, Murphy R (2012) Variation in cell wall composition and accessibility in relation to biofuel potential of short rotation coppice willows. BioEnergy Res **5:** 685–698

Rinaldi C, Kohler A, Frey P, Duchaussoy F, Ningre N, Couloux A, Wincker P, Le Thiec D, Fluch S, Martin F, et al (2007) Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust Melampsora larici-populina. Plant Physiol **144:** 347–366

Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R (2004) When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. Plant Physiol **134:** 1683–1696

Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods **10:** 71–73

Robinson BH, Mills TM, Petit D, Fung LE, Green SR, Clothier BE (2000) Natural and induced cadmium-accumulation in poplar and willow: implications for phytoremediation. Plant Soil **227:** 301–306

Robinson DG, Wang JY, Storey JD (2015) A nested parallel experiment demonstrates differences in intensity-dependence between RNA-seq and microarrays. Nucleic Acids Res **43:** e131

Roux S, Krupovic M, Debroas D, Forterre P, Enault F (2013) Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol **3:** 130160

Rugh CL, Senecoff JF, Meagher RB, Merkle SA (1998) Development of transgenic yellow poplar for mercury phytoremediation. Nat Biotechnol **16:** 925–928

Sage R, Cunningham M, Boatman N (2006) Birds in willow short-rotation coppice compared to other arable crops in central England and a review of bird census data from energy crops in the UK. Ibis **148:** 184–197

Sarkar N (1996) Polyadenylation of mRNA in bacteria. Microbiology **142:** 3125–3133

Shiri M, Rabhi M, Abdelly C, El Amrani A (2015) The halophytic model plant Thellungiella salsuginea exhibited increased tolerance to phenanthrene-induced stress in comparison with the glycophitic one Arabidopsis thaliana: application for phytoremediation. Ecol Eng **74:** 125–134

Slomovic S, Laufer D, Geiger D, Schuster G (2005) Polyadenylation and degradation of human mitochondrial RNA: the prokaryotic past leaves its mark. Mol Cell Biol **25:** 6427–6435

Slomovic S, Portnoy V, Liveanu V, Schuster G (2006) RNA polyadenylation in prokaryotes and organelles; different tails tell different tales. Crit Rev Plant Sci **25:** 65–77

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature **480:** 241–244

Solanas M, Moral R, Escrich E (2001) Unsuitability of using ribosomal RNA as loading control for Northern blot analyses related to the imbalance between messenger and ribosomal RNA content in rat mammary tumors. Anal Biochem **288:** 99–102

Somerville CR, Portis AR, Ogren WL (1982) A mutant of Arabidopsis thaliana which lacks activation of RuBP carboxylase in vivo. Plant Physiol **70:** 381–387

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics **14:** 91

Stephenson AL, Dupree P, Scott SA, Dennis JS (2010) The environmental and economic sustainability of potential bioethanol from willow in the UK. Bioresour Technol **101:** 9612–9623

Sullivan TS, McBride MB, Thies JE (2013) Rhizosphere microbial community and Zn uptake by willow (Salix purpurea L.) depend on soil sulfur concentrations in metalliferous peat soils. Appl Soil Ecol **67:** 53–60

Suzuki T, Igarashi K, Dohra H, Someya T, Takano T, Harada K, Omae S, Hirai H, Yano K, Kawagishi H (2013) A new omics data resource of Pleurocybella porrigens for gene discovery. PLoS One **8:** e69681

Thoemmes MS, Fergus DJ, Urban J, Trautwein M, Dunn RR (2014) Ubiquity and diversity of human-associated Demodex mites. PLoS One **9:** e106265

Tjaden B (2015) De novo assembly of bacterial transcriptomes from RNA-seq data. Genome Biol **16:** 1

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc **7:** 562–578

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The human microbiome project. Nature **449:** 804–810

Uematsu K, Suzuki N, Iwamae T, Inui M, Yukawa H (2012) Increased fructose 1,6-bisphosphate aldolase in plastids enhances growth and photosynthesis of tobacco plants. J Exp Bot **63:** 3001–3009

Unamba CIN, Nag A, Sharma RK (2015) Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. Front Plant Sci **6:** 1074

Vickers CE, Gershenzon J, Lerdau MT, Loreto F (2009) A unified mechanism of action for volatile isoprenoids in plant abiotic stress. Nat Chem Biol **5:** 283–291

Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet **13:** 227–232

Volk TA, Abrahamson LP, Nowak CA, Smart LB, Tharakan PJ, White EH (2006) The development of short-rotation willow in the northeastern United States for bioenergy and bioproducts, agroforestry and phytoremediation. Biomass Bioenergy **30:** 715–727

Wang YC, Qu GZ, Li HY, Wu YJ, Wang C, Liu GF, Yang CP (2010) Enhanced salt tolerance of transgenic poplar plants expressing a manganese superoxide dismutase from Tamarix androssowii. Mol Biol Rep **37:** 1119–1124

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10:** 57–63

Watts AW, Ballestero TP, Gardner KH (2006) Uptake of polycyclic aromatic hydrocarbons (PAHs) in salt marsh plants Spartina alterniflora grown in contaminated sediments. Chemosphere **62:** 1253–1260

Wedel N, Soll J, Paap BK (1997) CP12 provides a new mode of light regulation of Calvin cycle activity in higher plants. Proc Natl Acad Sci USA **94:** 10479–10484

Weih M, Nordh NE (2002) Characterising willows for biomass and phytoremediation: growth, nitrogen and water use of 14 willow clones under different irrigation and fertilisation regimes. Biomass Bioenergy **23:** 397–413

Weyens N, van der Lelie D, Taghavi S, Vangronsveld J (2009) Phytoremediation: plant-endophyte partnerships take the challenge. Curr Opin Biotechnol **20:** 248–254

Wood-Charlson EM, Weynberg KD, Suttle CA, Roux S, van Oppen MJH (2015) Metagenomic characterisation of viral communities in corals: Mining biological signal from methodological noise. Environ Microbiol Rep doi: 10.1111/1758-2229.12275

Wright JJ, Mewis K, Hanson NW, Konwar KM, Maas KR, Hallam SJ (2014) Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. ISME J **8:** 455–468

Xu YH, Liu R, Yan L, Liu ZQ, Jiang SC, Shen YY, Wang XF, Zhang DP (2012a) Light-harvesting chlorophyll a/b-binding proteins are required for stomatal response to abscisic acid in Arabidopsis. J Exp Bot **63:** 1095–1106

Xu Y, Zhu X, Gong Y, Xu L, Wang Y, Liu L (2012b) Evaluation of reference genes for gene expression studies in radish (Raphanus sativus L.) using quantitative real-time PCR. Biochem Biophys Res Commun **424:** 398–403

Yakushevska AE, Keegstra W, Boekema EJ, Dekker JP, Andersson J, Jansson S, Ruban AV, Horton P (2003) The structure of photosystem II in Arabidopsis: localization of the CP26 and CP29 antenna complexes. Biochemistry **42:** 608–613

Yergeau E, Sanschagrin S, Maynard C, St-Arnaud M, Greer CW (2014) Microbial expression profiles in the rhizosphere of willows depend on soil contamination. ISME J **8:** 344–358

Yoo KS, Ok SH, Jeong BC, Jung KW, Cui MH, Hyoung S, Lee MR, Song HK, Shin JS (2011) Single cystathionine beta-synthase domain-containing proteins modulate development by regulating the thioredoxin system in Arabidopsis (vol 23, pg 3577, 2011). Plant Cell **23:** 3577–3594

Yue DJ, You FQ, Snyder SW (2014) Biomass-to-bioenergy and biofuel supply chain optimization: overview, key issues and challenges. Comput Chem Eng **66:** 36–56

Yurekli F, Porgali ZB (2006) The effects of excessive exposure to copper in bean plants. Acta Biol Cracov Ser; Bot **48:** 7–13

Zhang Z, Lu Y, Zhai L, Deng R, Jiang J, Li Y, He Z, Peng X (2012) Glycolate oxidase isozymes are coordinately controlled by GLO1 and GLO4 in rice. PLoS One **7:** e39658