

Research article

Open Access

# Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding

Steven G Ralph<sup>†1,6</sup>, Hye Jung E Chun<sup>†2</sup>, Dawn Cooper<sup>1</sup>, Robert Kirkpatrick<sup>2</sup>, Natalia Kolosova<sup>1,3</sup>, Lee Gunter<sup>4</sup>, Gerald A Tuskan<sup>4</sup>, Carl J Douglas<sup>3</sup>, Robert A Holt<sup>2</sup>, Steven JM Jones<sup>2</sup>, Marco A Marra<sup>2</sup> and Jörg Bohlmann<sup>\*1,3,5</sup>

Address: <sup>1</sup>Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada, <sup>2</sup>British Columbia Cancer Agency Genome Sciences Centre, Vancouver, British Columbia, V5Z 4E6, Canada, <sup>3</sup>Department of Botany, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada, <sup>4</sup>Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37831, USA, <sup>5</sup>Department of Forest Sciences, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada and <sup>6</sup>Department of Biology, University of North Dakota, Grand Forks, North Dakota, 58202-9019, USA

Email: Steven G Ralph - [steven.ralph@und.nodak.edu](mailto:steven.ralph@und.nodak.edu); Hye Jung E Chun - [echun@bcgsc.ca](mailto:echun@bcgsc.ca); Dawn Cooper - [dmcooper@sfu.ca](mailto:dmcooper@sfu.ca); Robert Kirkpatrick - [robertk@bcgsc.bc.ca](mailto:robertk@bcgsc.bc.ca); Natalia Kolosova - [kolosova@interchange.ubc.ca](mailto:kolosova@interchange.ubc.ca); Lee Gunter - [gunterle@ornl.gov](mailto:gunterle@ornl.gov); Gerald A Tuskan - [tuskanga@ornl.gov](mailto:tuskanga@ornl.gov); Carl J Douglas - [cdouglas@interchange.ubc.ca](mailto:cdouglas@interchange.ubc.ca); Robert A Holt - [rholt@bcgsc.ca](mailto:rholt@bcgsc.ca); Steven JM Jones - [sjones@bcgsc.ca](mailto:sjones@bcgsc.ca); Marco A Marra - [mmarra@bcgsc.ca](mailto:mmarra@bcgsc.ca); Jörg Bohlmann\* - [bohlmann@interchange.ubc.ca](mailto:bohlmann@interchange.ubc.ca)

\* Corresponding author †Equal contributors

Published: 29 January 2008

Received: 6 November 2007

BMC Genomics 2008, 9:57 doi:10.1186/1471-2164-9-57

Accepted: 29 January 2008

This article is available from: <http://www.biomedcentral.com/1471-2164/9/57>

© 2008 Ralph et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

**Background:** The genus *Populus* includes poplars, aspens and cottonwoods, which will be collectively referred to as poplars hereafter unless otherwise specified. Poplars are the dominant tree species in many forest ecosystems in the Northern Hemisphere and are of substantial economic value in plantation forestry. Poplar has been established as a model system for genomics studies of growth, development, and adaptation of woody perennial plants including secondary xylem formation, dormancy, adaptation to local environments, and biotic interactions.

**Results:** As part of the poplar genome sequencing project and the development of genomic resources for poplar, we have generated a full-length (FL)-cDNA collection using the biotinylated CAP trapper method. We constructed four FLcDNA libraries using RNA from xylem, phloem and cambium, and green shoot tips and leaves from the *P. trichocarpa* Nisqually-1 genotype, as well as insect-attacked leaves of the *P. trichocarpa* × *P. deltoides* hybrid. Following careful selection of candidate cDNA clones, we used a combined strategy of paired end reads and primer walking to generate a set of 4,664 high-accuracy, sequence-verified FLcDNAs, which clustered into 3,990 putative unique genes. Mapping FLcDNAs to the poplar genome sequence combined with BLAST comparisons to previously predicted protein coding sequences in the poplar genome identified 39 FLcDNAs that likely localize to gaps in the current genome sequence assembly. Another 173 FLcDNAs mapped to the genome sequence but were not included among the previously predicted genes in the poplar genome. Comparative sequence analysis against *Arabidopsis thaliana* and other species in the non-redundant database of GenBank revealed that 11.5% of the poplar FLcDNAs display no significant sequence similarity to other plant proteins. By mapping the poplar FLcDNAs against transcriptome data previously obtained with a 15.5 K cDNA microarray, we identified 153

FLcDNA clones for genes that were differentially expressed in poplar leaves attacked by forest tent caterpillars.

**Conclusion:** This study has generated a high-quality FLcDNA resource for poplar and the third largest FLcDNA collection published to date for any plant species. We successfully used the FLcDNA sequences to reassess gene prediction in the poplar genome sequence, perform comparative sequence annotation, and identify differentially expressed transcripts associated with defense against insects. The FLcDNA sequences will be essential to the ongoing curation and annotation of the poplar genome, in particular for targeting gaps in the current genome assembly and further improvement of gene predictions. The physical FLcDNA clones will serve as useful reagents for functional genomics research in areas such as analysis of gene functions in defense against insects and perennial growth. Sequences from this study have been deposited in NCBI GenBank under the accession numbers [EF144175](#) to [EF148838](#).

---

## Background

Poplars are keystone tree species in several temperate forest ecosystems in the Northern Hemisphere. Poplars are also intensively cultivated in plantation forestry for the production of wood, pulp, and paper. Fast growing poplars can serve functions in phytoremediation, as a sink for carbon sequestration, and as a feedstock for biofuel production. Poplar has also been firmly established as a model research system for long-lived woody perennials (reviewed in [1]). Advances in functional genomics of poplar have been greatly enhanced by the availability of a high-quality genome sequence from *P. trichocarpa* (Nisqually-1; [2]), combined with comprehensive genetic [3-6] and physical genome [7] maps, as well as the availability of several platforms for transcriptome analysis [8-11] and genetic transformation. Large collections of expressed sequence tags (ESTs) have also been developed from a variety of poplar species and hybrids focussing on gene discovery in wood formation, dormancy, floral development and stress response [9,11-20]. These short, single-pass EST reads have been a critical resource for gene discovery, genome annotation, and the construction of microarray platforms.

High-accuracy, sequence-verified FLcDNA sequences that span the entire protein-coding region of a given gene can advance comparative, functional, and structural genome analysis. For example, the accuracy of *ab initio* prediction of protein-coding regions in genome sequences is limited by the difficulty of finding islands of coding sequences within an ocean of non-coding DNA, and by the complexity of individual genes that may code for multiple peptides through alternative splicing. More robust approaches that unambiguously identify protein-coding regions in a genome sequence have used FLcDNA data, as demonstrated for example in *Arabidopsis thaliana* [21-23]. Despite their immense value, sequence-verified FLcDNA clones, where multiple passes verify the authenticity of reads, have not been generated in most plant species subjected to genomic analysis. Only a few large FLcDNA data

sets have been generated for plants; namely for rice [24], *Arabidopsis* [25], and maize [26,27]. In contrast, as of September 2007, there were only 1,409 complete sequences from individual poplar FLcDNA clones in the non-redundant (NR) division of GenBank, in addition to a larger number of putative full-length sequences assembled from EST reads of multiple cDNA clones.

Our poplar FLcDNA program in the areas of forest health genomics and wood formation has focused on mechanisms of defense and resistance against insects and genes associated with xylem development. The forest tent caterpillar (*Malacosoma disstria*; FTC; [28]) is a major insect pest that threatens the productivity of natural and plantation forests. Poplars deploy an array of combined defense strategies against herbivores that can be grouped as chemical and physical defenses, direct and indirect defenses, constitutive and induced defenses, as well as local and systemic defenses (reviewed in [29]). Several recent studies have been conducted on the molecular mechanisms underlying inducible defenses against herbivores in poplar [11,18,30-37].

In this paper, we report on the development of four FLcDNA libraries from poplar that served as the starting template for creating a substantial genomic resource of 4,664 sequence-verified FLcDNAs. We describe the overall structural features of these FLcDNA clones, annotation based on comparisons with other species, and the identification of 536 putative poplar-specific transcripts. Mapping the FLcDNA collection to the poplar genome sequence confirmed the overall high quality of the assembled genome sequence as well as the high quality of the FLcDNA resource, while also identifying 39 expressed poplar transcripts that appear to be derived from gap regions of the current genome sequence assembly and 173 new poplar genes that have not previously been identified in the genome assembly. By mapping 3,854 FLcDNAs to a poplar 15.5 K cDNA microarray platform and performing a comparison with existing transcriptome data, we identi-

fied 153 FLCDNAs that match transcripts differentially expressed following insect attack by FTC on poplar leaves.

## Results

### Selection and sequence finishing of FLCDNAs

FLcDNAs are defined as individual cDNA clones that contain the complete protein-coding sequence and at least partial 5' and 3' untranslated regions (UTRs) for a given transcript. This definition distinguishes *bona fide* FLCDNAs from *in silico* assembled EST sequences derived from multiple cDNA clones. In the latter case, it is possible that multiple, closely related genes or allelic variants of the same gene are assembled into a single consensus sequence. This problem is avoided when only sequences derived from the same physical FLCDNA clone are assembled. We prepared four FLCDNA libraries using the biotinylated CAP trapper method [38]. Three libraries constructed from xylem, phloem and cambium, and green shoot tips and leaves were derived from the *P. trichocarpa* Nisqually-1 genotype, for which the genome sequence has been reported [2]. An additional library was developed from the *P. trichocarpa* × *P. deltoides* hybrid H11-11 genotype using leaves subjected to FTC herbivory (Table 1).

To select candidate FLCDNAs for complete insert sequencing, we used a previously described bioinformatic pipeline for EST processing [11]. An initial set of 26,112 3' ESTs derived from FLCDNA libraries was combined with 81,407 3' ESTs from standard EST libraries [11] to generate a starting set of 107,519 3'-end ESTs, which resulted in 90,368 high-quality ESTs after filtering to remove sequences of low quality and contaminant sequences from yeast, bacteria and fungi. These sequences were then clustered using the CAP3 assembly program ([39]; assembly criteria: 95% identity, 40 bp window) to identify a set of 35,011 putative unique transcripts (PUTs; Figure 1). To maximize the capture of complete open reading frames (ORFs) and UTRs, only clones from full-length libraries were considered further. Using this strategy, we identified 5,926 cDNA candidate clones for full insert sequencing, which resulted in 4,664 sequence-verified poplar FLCDNA clones (see Additional file 1 and Figure 2). Inserts of 2,672

clones were completely sequenced using end reads only, with an average sequenced insert size of  $735 \pm 434$  bp (average  $\pm$  SD) and required an average of  $4.5 \pm 1.3$  end reads to finish to high sequence quality. Using a combination of end reads and primer walking, inserts of an additional 1,992 clones were completely sequenced, with an average insert size of  $1,308 \pm 567$  bp requiring  $5.9 \pm 2.8$  end reads and  $3.4 \pm 1.8$  internal primer reads per clone.

Analysis of the 4,664 FLCDNA sequences using the CAP3 clustering and assembly program ([39]; assembly criteria: 95% identity, 40 bp window) identified 3,505 FLCDNAs as unique singletons, with the remaining 1,159 grouping into 485 contigs, suggesting a total of 3,990 unique genes represented with finished FLCDNA sequences. The high percentage of unique transcripts (85.5%) within this set confirms the successful clone selection strategy (Figure 1) for establishing a low-redundancy clone set prior to sequence finishing.

### Sequence quality and "full-length" assessment of poplar FLCDNAs

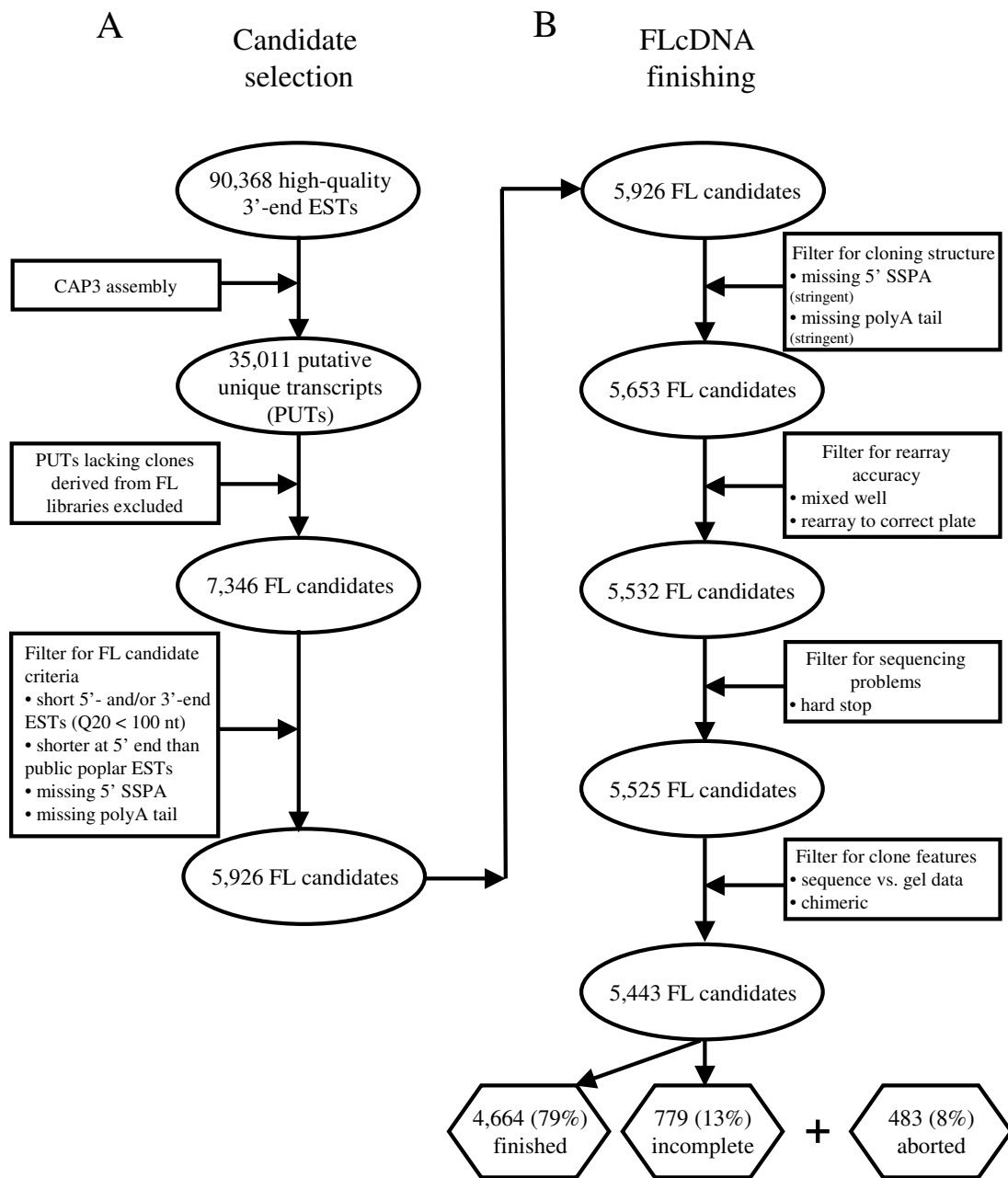
All 4,664 finished FLCDNAs achieved a minimum of Phred30 (i.e., one error in  $10^3$  bases) sequence quality at every base. The majority of FLCDNAs were of even higher quality with the minimum and average Phred values exceeding Phred45 (i.e., one error in  $3 \times 10^4$  bases) and Phred80 (i.e., one error in  $10^8$  bases), respectively (Figure 3). We predicted the complete protein-coding ORFs for all 4,664 FLCDNAs. The distribution of 5' UTR, ORF and 3' UTR lengths is illustrated in Figure 2 [also see Additional file 1]. The average sequenced FLCDNA length (from the beginning of the 5' UTR to the end of the polyA tail) was  $1,045 \pm 475$  bp (mean  $\pm$  SD), and ranged from 147 to 3,342 bp, whereas the average predicted ORF was  $649 \pm 429$  bp and ranged from 33 to 2,935 bp. ORFs could not be detected (i.e., 30 bp or less) for 96 FLCDNAs. The 5' and 3' UTRs averaged  $109 \pm 138$  bp and  $228 \pm 152$  bp, respectively. These results are comparable to CAP trapper FLCDNA collections from other plant species including maize (cDNA insert 799 bp, 5' UTR 99 bp, 3' UTR 206 bp; [27]), Arabidopsis (cDNA insert ca. 1.2 kb; [40]) and rice (5' UTR 259 bp, 3' UTR 398 bp; [24]). Similarly, the aver-

**Table 1: Libraries, tissue sources and species for sequences described in this study**

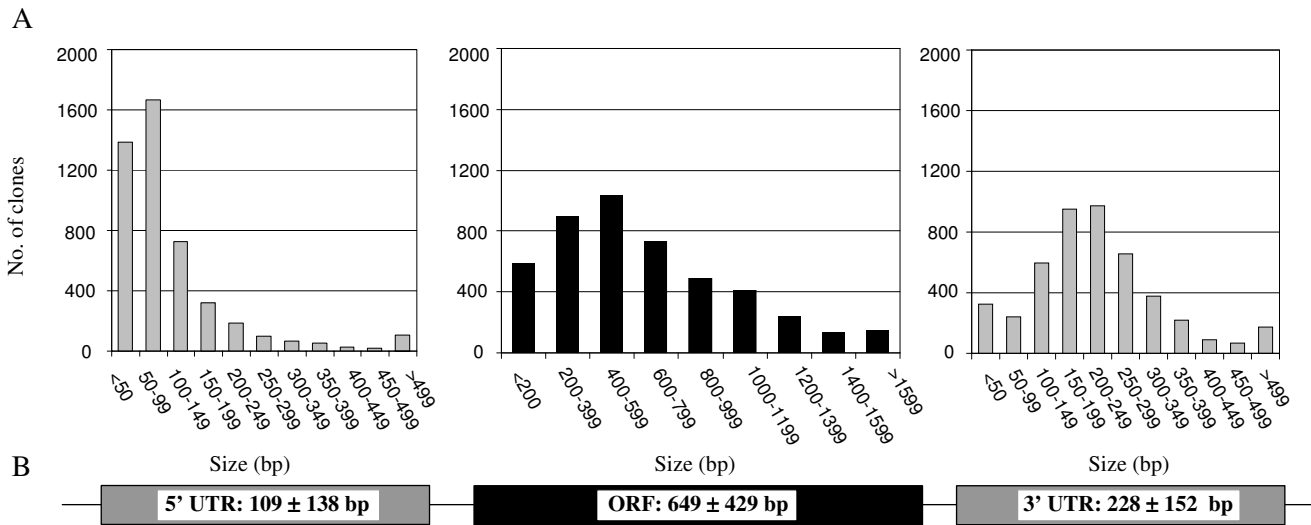
| cDNA Library   | Tissue/Developmental Stage   | Species (genotype)                                |
|----------------|--|---|
| PT-X-FL-A-1    | Outer xylem <sup>a</sup> .   | <i>Populus trichocarpa</i> (Nisqually-1)          |
| PT-P-FL-A-2    | Phloem and cambium <sup>a</sup> .  | <i>P. trichocarpa</i> (Nisqually-1)               |
| PT-GT-FL-A-3   | Young and mature leaves, along with green shoot tips <sup>a</sup> .  | <i>P. trichocarpa</i> (Nisqually-1)               |
| PTxD-IL-FL-A-4 | Local and systemic (above region of feeding) mature leaves harvested after continuous feeding by forest tent caterpillars, <i>Malacosoma disstria</i> . Local tissue was collected 4, 8 and 24 h post-treatment and systemic tissue 4, 12 and 48 h post-treatment <sup>b</sup> . | <i>P. trichocarpa</i> × <i>deltoides</i> (H11-11) |

<sup>a</sup>Harvested May 15th, 2001 from eight year old trees within the Boise Cascade region of Washington state.

<sup>b</sup>One or two year old saplings grown in potted soil under greenhouse conditions at the University of British Columbia.



**Figure 1**  
**Schematic of clone selection and complete insert sequencing of 4,664 FLcDNAs.** CAP3 assembly of 90,368 high-quality 3'-end ESTs identified 35,011 putative unique transcripts (PUTs) for the identification of candidate FLcDNAs. Only those PUTs containing at least one clone from a FLcDNA library were considered further. To maximize the number of FLcDNAs captured, candidate clones were excluded from further analysis if: (1) the 5' second strand primer adaptor (SSPA) was absent; (2) a polyA tail was absent; (3) 5'- and/or 3'-end ESTs had a Phred20 quality length (Q20) of < 100 nt; or (4) BLASTN ( $E < 1e^{-80}$ ) versus poplar ESTs in the public domain identified a candidate as potentially truncated (i.e., > 100 nt shorter) at the 5' end of the transcript relative to a matching EST. Among the 5,926 candidates selected for sequencing, only 483 (8%) were aborted at various stages of the sequence finishing pipeline due to: (1) missing cloning structures; (2) errors in re-array of glycerol stocks; (3) problematic sequencing such as hard stops; or (4) problematic clone features such as chimeric sequences. Through a combination of end reads and gap closing using primer walking, 4,664 (79%) sequence-verified FLcDNAs were completed. An additional 779 clones (13%) from the starting set of 5,926 will be finished in future work.



**Figure 2**  
**Distribution of open reading frame (ORF) and 5' and 3' untranslated region (UTR) sizes among the finished 4,664 FLcDNAs (A), and the mean ORF and UTR length ( $\pm$  standard deviation) (B).** Each finished FLcDNA sequence was examined for the presence of ORFs using either the EMBOSS getorf program (version 2.5.0; [55]) or an in-house BLAST-aided program. The getorf program identifies the longest stretch of uninterrupted sequence between a start (ATG) and stop codon (TGA, TAG, TAA) in the 5' to 3' direction for the predicted ORF. The BLAST-aided program detects ORFs by finding the starting methionine and stop codon in a poplar FLcDNA sequence relative to the same features in the most closely related Arabidopsis protein identified by BLASTX (E values  $< 1e^{-20}$ ). For this study, ORFs identified by the BLAST-aided method were utilized except in cases where the FLcDNA sequence did not show high similarity to an Arabidopsis protein, in which case the ORF identified by the getorf program was chosen. The presence and coordinates of the 5' second strand primer adaptor sequence (SSPA) and polyA tail were also noted. The regions between the 5'SSPA and the predicted ORF start and between the predicted ORF stop and the polyA tail were taken to be the 5' and 3' UTRs, respectively. The 5' SSPA and 3' polyA tail lengths were not included when determining UTR length.

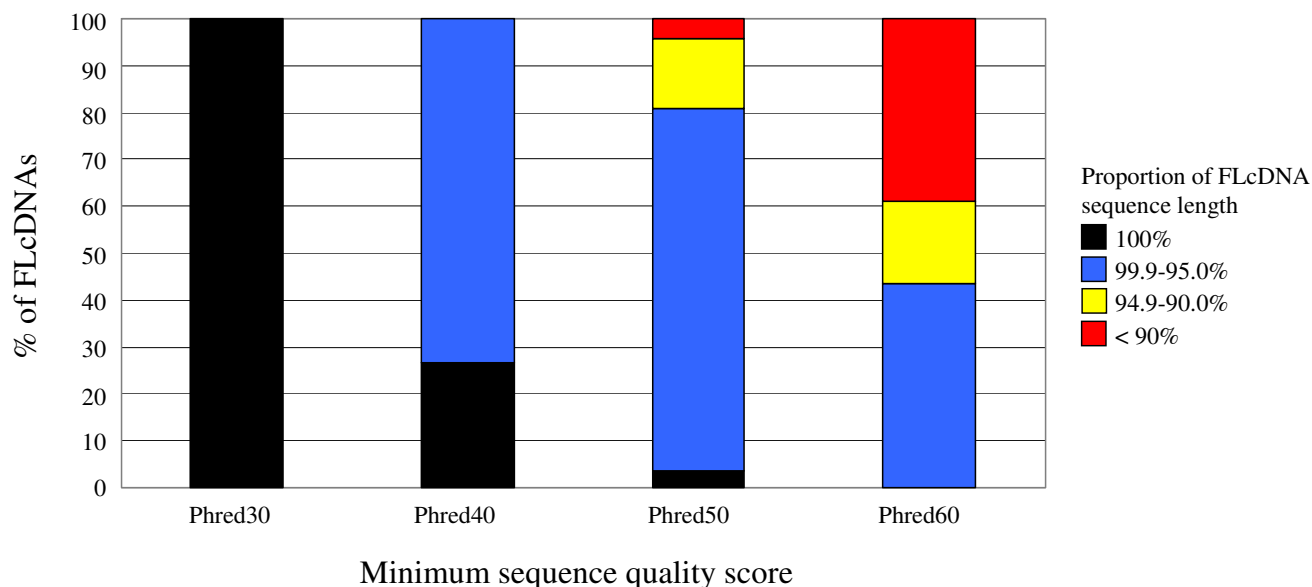
age transcript length of the 45,555 poplar reference genes predicted *ab initio* from the genome sequence was 1,079 bp and 5' and 3' UTRs averaged 92 bp [2], in close agreement with our results obtained with FLcDNAs.

To further assess the quality of the 4,664 poplar FLcDNAs, we performed reciprocal BLAST analysis against peptide sequences in The Arabidopsis Information Resource (TAIR) and against a set of 1,409 poplar sequences previously identified to be full-length (collected from the NR division of GenBank). Reciprocal BLAST analysis was performed with a stringent similarity threshold [% identity  $\geq 50\%$ ; expect (E) value  $\leq 1e^{-20}$ ] and identified 2,774 and 288 pairs, respectively, with Arabidopsis and previously published poplar FLcDNAs (Figure 4). Of the 288 homologous poplar transcript pairs (i.e., previously published poplar sequences with high sequence similarity to FLcDNAs reported in this study), 228 (79.2%) agreed well with regard to their ORF lengths and position of their start and stop codons ( $\pm$  ten amino acids; Figure 4). For the remaining pairs, the predicted 5' and/or 3' ORF ends did not match suggesting alternative start or stop codons, splice variants, or the possibility that one of the pair members

was either truncated or had an incorrectly predicted ORF. When comparing the poplar FLcDNA collection to reciprocal matches from TAIR Arabidopsis peptides, we observed a similar number of 2,151 (77.5%) pairs with similar ORF lengths and positions of their starting methionine and stop codons ( $\pm$  ten amino acids; Figure 4). These results indicate the majority of the 4,664 poplar FLcDNAs represent true full-length transcripts with complete ORFs and correctly annotated start and stop codons.

**Mapping FLcDNAs to the poplar genome sequence to reassess gene prediction and to identify possible gaps in the genome assembly**

As part of the poplar genome sequencing project [2], the poplar FLcDNAs were used to train a series of gene prediction algorithms to identify coding regions in the genome sequence. To reassess the effectiveness of gene prediction in the current genome assembly and to search for possible genome sequence gaps, we took two approaches: 1) BLAT [41] was utilized to map FLcDNAs to the assembled genome sequence, and 2) BLASTN was applied to align FLcDNAs with the 45,555 protein-coding gene loci predicted from the poplar genome sequence. Using BLAT, we

**Figure 3**

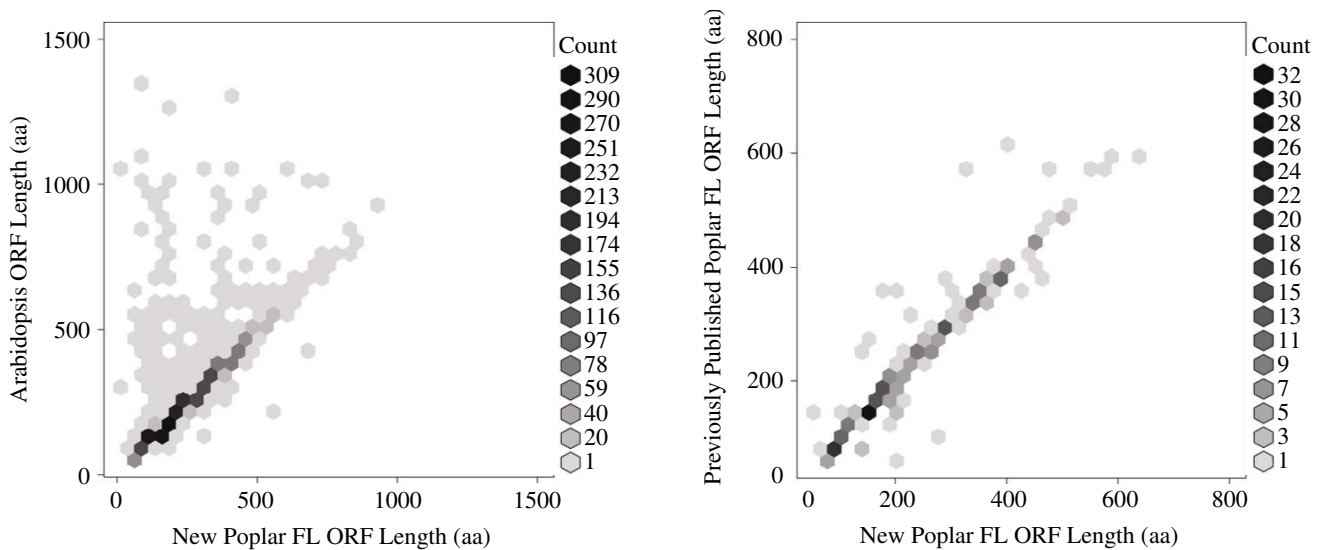
**Validation of sequence quality of FLcDNAs.** Sequence accuracy was measured as the percentage of the 4,664 FLcDNAs which, with 100%, 95.0–99.9%, 90.0–94.9% or < 90.0% of their sequence length, exceeded Phred30, Phred40, Phred50 or Phred60 sequence quality thresholds. All 4,664 FLcDNAs exceeded the Phred30 quality thresholds (calculated as less than 1 error in  $10^3$  sequenced nucleotides) over 100% of their sequence length. Even at the threshold level of Phred60 (calculated as less than 1 error in  $10^6$  sequenced nucleotides) the majority (61.2%) of the FLcDNA sequences met this very high sequence quality score over > 90.0% of their length.

mapped 4,642 poplar FLcDNAs (99.5%) to the genome at a minimum threshold (tile match length  $\geq 11$  bp, score  $\geq 30$ , sequence identity  $\geq 90\%$ ; Figure 5). From this set, 3,847 (82.9%) mapped to the 19 linkage groups (i.e., chromosomes) whereas the remainder mapped to scaffold segments that were not incorporated into the poplar genome sequence assembly. Examination of the linkage group location of FLcDNAs suggests a pattern of random distribution when grouped by cDNA library/tissue of origin, with an approximately even distribution of FLcDNAs throughout the genome (Figure 5). When we applied a more stringent similarity threshold (sequence identity  $\geq 95\%$ , alignment coverage  $\geq 95\%$ ), the number of poplar FLcDNAs matching to the genome was only slightly reduced to 4,487 (96.2%).

In addition to BLAT analysis, we also compared the FLcDNAs with the 45,555 predicted protein-coding gene loci identified in the genome sequence using BLASTN and observed 4,452 (95.5%) matched at an E value  $< 1e^{-50}$  (see Additional file 1). In order to identify possible sequence gaps in the  $7.5\times$  coverage genome, we searched for FLcDNAs lacking a stringent BLAT to the genome match and a BLASTN match (E value  $\geq 1e^{-50}$ ) to the predicted gene models. This approach identified only 39 candidates, of which 20 (0.4%) FLcDNAs also had a strong match by

BLASTN (E value  $< 1e^{-50}$ ) to one or more poplar ESTs in the public domain, excluding ESTs reported in this study (Table 2 and see Additional file 1), suggesting that these FLcDNAs represent expressed poplar genes that likely map to gap regions within the current genome draft. We cannot exclude the possibility that the remaining 19 FLcDNAs represent sequences from bacterial, fungal or insect species present on poplar tissues harvested for cDNA library construction, which were not filtered as contaminant sequences in our EST and FLcDNA processing procedures.

To identify expressed genes that were not predicted in the original genome annotation [2], we searched among the set of 4,487 FLcDNAs with a stringent BLAT match to the genome that did not match to any of the 45,555 predicted gene models (E value  $\geq 1e^{-50}$ ). This analysis revealed 173 FLcDNAs, 79 of which also showed strong similarity (E value  $< 1e^{-50}$ ) to one or more poplar ESTs in the public domain (see Additional file 1), suggesting that these 79 FLcDNAs represent expressed genes and possibly non-coding RNAs, that were missed by gene prediction software during the annotation of the poplar genome. The fact that these poplar transcripts had been missed could be due in part to the relatively short lengths of these 79 FLcDNAs (average FLcDNA and predicted ORF length of 555 bp and 67 bp, respectively; see Additional file 1).



**Figure 4**

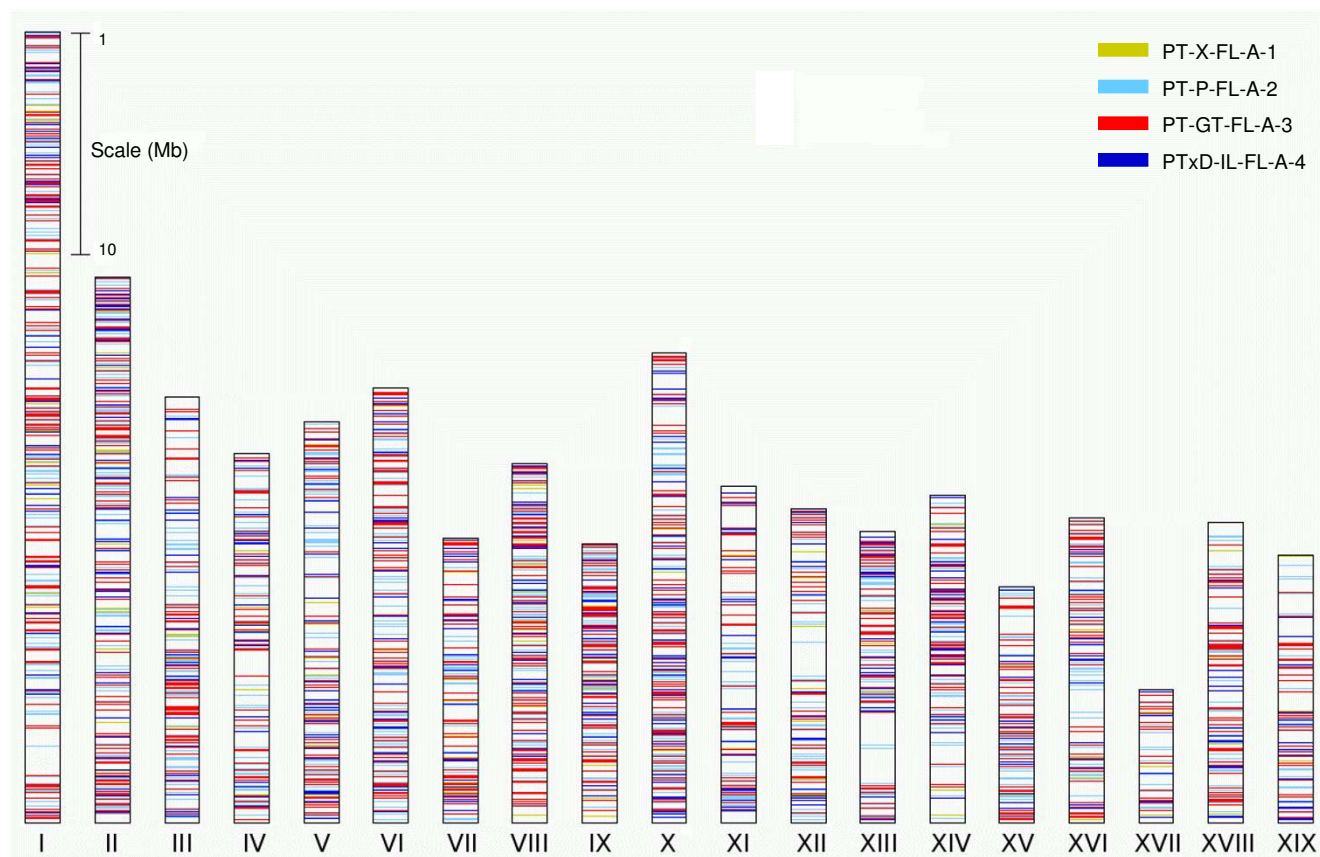
**Validation of poplar FLcDNAs by comparison to reciprocal BLAST matches against Arabidopsis peptides and previously published poplar FLcDNAs.** The set of 4,664 poplar FLcDNAs were compared using BLASTX to both The Arabidopsis Information Resource (TAIR) non-redundant Arabidopsis peptide set (28,952 sequences [56]) and a collection of 1,409 previously published poplar sequences from the non-redundant (NR) division of GenBank ([57], the NR release of December 19<sup>th</sup>, 2006) annotated as full-length (excluding predicted proteins derived from genomic DNA). FLcDNAs were excluded from the analysis when the in-house BLAST-aided ORF detection software identified a FLcDNA as problematic according to the following categories: truncation at the 5'-end (319), truncation at the 3'-end (50), frameshift (12), stop codon in the middle of an ORF (9), or inverted insert (3) [see Additional file 1]. No problematic features were identified in the remaining 4,271 FLcDNAs. This comparison identified 2,774 homologous Arabidopsis-poplar pairs and 288 homologous poplar transcript pairs. A FLcDNA pair was considered homologous if (1) the top BLASTX match exceeded a stringent threshold (% identity  $\geq 50\%$ ; expect value  $\leq 1e^{-20}$ ) and (2) the reciprocal TBLASTN analysis identified the same poplar FLcDNA with a score value equal to or within 10% of the top match. ORF lengths for Arabidopsis and public poplar sequences were extracted from the TAIR and NR records, respectively, and poplar ORF lengths from this study were predicted using either the EMBOSS getorf or in-house BLAST-aided programs (see Figure 2 legend). The greyscale shading of each hexagon represents poplar FLcDNA abundance. ORF lengths for three Arabidopsis-poplar pairs and eight homologous poplar transcript pairs differed by more than 500 aa and are not included in the figure.

**Comparative sequence annotation of poplar FLcDNAs against Arabidopsis and other plants identifies proteins unique to poplar**

Despite the growing research interest in poplar as a model angiosperm tree species and the recent completion of the poplar genome sequence, poplar still represents a difficult experimental system with relatively few functionally characterized proteins, compared to other established model systems such as Arabidopsis. Therefore, our effort of *in silico* annotation of poplar FLcDNAs was largely based on comparison with Arabidopsis together with the NR database of GenBank containing sequences from all plants, among other species. Using BLASTX, we found that the proportion of FLcDNAs with similarity to TAIR Arabidopsis proteins was 87.5% (4,081) at E value  $< 1e^{-05}$  and 55.5% (2,590) at E value  $< 1e^{-50}$  (Figure 6A). Similar values were obtained when using BLASTX to compare against peptides from other species in the NR division of GenBank (88.0% matches at E value  $< 1e^{-05}$  and 56.9%

matches at E value  $< 1e^{-50}$ ) (Figure 6A). As expected, the proportion of poplar FLcDNAs with sequence similarity to previously published poplar ESTs (i.e., ESTs available in the dbEST division of GenBank, excluding ESTs from this study) by BLASTN was very high, with 96.3% (4,496) and 94.3% (4,401) of FLcDNAs having matches with E values  $< 1e^{-05}$  and  $< 1e^{-50}$ , respectively (Figure 6A).

To identify genes that are potentially unique to poplar, we next examined the relationship of sequence similarity among the poplar FLcDNAs and best matching sequences in the TAIR Arabidopsis proteins, other NR database proteins (which includes all plant species), and previously published poplar EST datasets. Of the 4,664 poplar FLcDNAs, 3,994 (85.6%) had at least low sequence similarity to sequences in all three databases (E values  $< 1e^{-05}$ ; Figure 6B). Only 95 FLcDNAs had no similarity (E values  $\geq 1e^{-05}$ ) to sequences in any of these databases; however, 87 of these strongly matched to the poplar genome using BLAT



**Figure 5**

**Mapping FLCDNAs to the poplar genome.** 4,664 poplar FLCDNAs were aligned to the genome using BLAT with default parameters (match length  $\geq 11$  bp, BLAT score  $\geq 30$ , sequence identity  $\geq 90\%$ ). Prior to alignment, the 5' second strand primer adaptor sequences (SSPA) and polyA tails were removed. Among 4,642 poplar FLCDNAs that exceeded the minimal criteria for a match to the genome, 3,847 mapped to chromosomes whereas the remainder mapped to scaffold segments. Colored bars indicate the cDNA library of origin for those FLCDNAs mapping to one of the 19 poplar chromosomes. Applying a higher stringency threshold (sequence identity  $\geq 95\%$ , alignment coverage  $\geq 95\%$ ), 4,487 or 96.2% of poplar FLCDNAs could be mapped to the genome.

(sequence identity  $\geq 95\%$ , alignment coverage  $\geq 95\%$ ). Our results suggest that these 87 genes that are represented with FLCDNAs and with poplar genomic sequences are new genes that have not previously been identified in other poplar EST collections or among genes in Arabidopsis and other plant species (see Additional file 1).

In addition, we also identified 536 poplar FLCDNAs (including the 95 FLCDNAs with no similarity to sequences in the three databases examined) with no similarity to Arabidopsis or NR proteins ( $E$  values  $\geq 1e^{-05}$ ), of which 346 FLCDNAs matched with high similarity to both the poplar genome by BLAT and to previously published poplar ESTs by BLASTN ( $E$  values  $< 1e^{-50}$ ; Figure 6B and see Additional file 1). These poplar FLCDNAs could represent genes that were gained and then rapidly diverged in sequence since the recent whole genome duplication in

poplar, or they may also represent non-coding RNAs or small peptides in poplar that share limited sequence similarity with other plants. The fact that these putative poplar-specific FLCDNAs do not share similarity with existing plant sequence data may also reflect the limited availability of sequence data from Salicaceae species closely related to poplar in the current NR database. To test these putatively poplar-specific FLCDNAs for known functional domains, we performed a search of the Pfam database [42]. At a threshold of  $E$  values  $< 1e^{-05}$ , we identified 2,908 (62.3%) poplar FLCDNAs with similarity to a Pfam domain; however, among the collection of 346 putatively poplar-specific genes only 8 FLCDNAs in this set matched a Pfam domain (see Additional file 1). Domain matches included PF05162.3/ribosomal protein L41 (WS0112\_A21, WS0116\_F12, WS0124\_J06, WS01230\_B01, and W01118\_I11), PF05160.3/DSS1/



**Table 2: Expressed FLcDNAs that identify possible gaps in the genome sequence assembly**

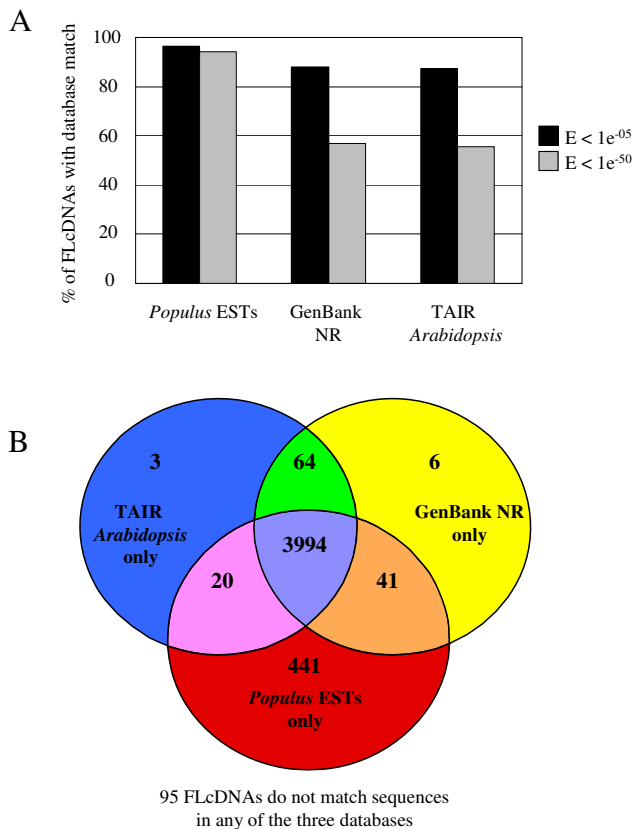
| Clone ID    | GenBank ID               | FLcDNA length (bp) | FL status/ORF size (aa) | NR BLASTP best match  |             | dbEST BLASTN best match                                  |             |
|-------------|--------------------------|--------------------|-------------------------|---|-------------|--|-------------|
|             |                          |                    |                         | GenBank accession, gene name, species   | BLAST Score | GenBank accession, species                               | BLAST Score |
| WS0138_J20  | <a href="#">EF148816</a> | 1444               | FL/340                  | AAB39877.1, NMT1 protein, <i>Uromyces fabae</i>                               | 1572        | DN493922.1, <i>Populus tremula</i>                       | 770         |
| WS01313_D10 | <a href="#">EF148323</a> | 1439               | FL/363                  | At3g20790, oxidoreductase, <i>Arabidopsis thaliana</i>                        | 1233        | DN501083, <i>P. trichocarpa</i>                          | 1318        |
| WS0127_P01  | <a href="#">EF148143</a> | 1237               | FL/299                  | AAD01907, methenyltetrahydrofolate dehydrogenase, <i>Pisum sativum</i>        | 1213        | CV131075.1, <i>P. deltoides</i>                          | 1511        |
| WS01231_K20 | <a href="#">EF147482</a> | 1207               | FL/256                  | At5g20060, phospholipase/carboxylesterase family, <i>A. thaliana</i>          | 1026        | DV464443.2, <i>P. fremontii</i> × <i>P. angustifolia</i> | 1479        |
| WS0135_G15  | <a href="#">EF148633</a> | 992                | n.a.                    | No matches  | n.a.        | BU891205, <i>P. tremula</i>                              | 240         |
| WS01312_F21 | <a href="#">EF148269</a> | 946                | n.a.                    | No matches  | n.a.        | BI122644.1, <i>P. tremula</i> × <i>P. tremuloides</i>    | 729         |
| WS01315_I11 | <a href="#">EF148467</a> | 836                | n.a.                    | No matches  | n.a.        | BU824948.1, <i>P. tremula</i> × <i>P. tremuloides</i>    | 339         |
| WS01312_H02 | <a href="#">EF148274</a> | 835                | n.a.                    | No matches  | n.a.        | BU791223.1, <i>P. trichocarpa</i> × <i>P. deltoides</i>  | 779         |
| WS01212_B01 | <a href="#">EF146690</a> | 821                | FL/88                   | BAB68268.1, drought-inducible protein, <i>Saccharum officinarum</i>           | 147         | BU879805.1, <i>P. trichocarpa</i>                        | 595         |
| WS0122_E05  | <a href="#">EF147284</a> | 739                | FL/131                  | CAB80775.1, proline-rich protein, <i>A. thaliana</i>                          | 340         | BU866461.1, <i>P. tremula</i>                            | 890         |
| WS0122_O15  | <a href="#">EF147357</a> | 736                | FL/162                  | At4g10300, hypothetical protein, <i>A. thaliana</i>                           | 444         | CX181869.1, <i>Populus × canadensis</i>                  | 1215        |
| WS0113_C11  | <a href="#">EF145750</a> | 722                | FL/136                  | At3g12260, complex I/LVR family protein, <i>A. thaliana</i>                   | 426         | BU879375.1, <i>P. trichocarpa</i>                        | 1223        |
| WS0125_PI8  | <a href="#">EF147919</a> | 596                | 3' trunc./70            | AAF71823.1, pumilio domain protein, <i>P. tremula</i> × <i>P. tremuloides</i> | 167         | CX187487.1, <i>Populus × canadensis</i>                  | 722         |
| WS01123_K15 | <a href="#">EF145357</a> | 483                | n.a.                    | No matches  | n.a.        | CK319617.1, <i>P. deltoides</i>                          | 268         |
| WS01231_G04 | <a href="#">EF147458</a> | 416                | 5' trunc./62            | At3g18790, hypothetical protein, <i>A. thaliana</i>                           | 200         | CX184264.1, <i>Populus × canadensis</i>                  | 543         |
| WS0124_L22  | <a href="#">EF147751</a> | 360                | n.a.                    | No matches  | n.a.        | BI128250.1, <i>P. tremula</i> × <i>P. tremuloides</i>    | 494         |
| WS0126_O09  | <a href="#">EF148027</a> | 342                | n.a.                    | No matches  | n.a.        | CF228572.1, <i>P. tremula</i> × <i>P. alba</i>           | 410         |
| WS01118_P04 | <a href="#">EF144846</a> | 300                | n.a.                    | No matches  | n.a.        | CX184524.1, <i>Populus × canadensis</i>                  | 242         |
| WS0136_N09  | <a href="#">EF148717</a> | 278                | n.a.                    | No matches  | n.a.        | CX179364.1, <i>Populus × canadensis</i>                  | 458         |
| WS0138_I14  | <a href="#">EF148811</a> | 231                | n.a.                    | No matches  | n.a.        | CX170421.1, <i>P. deltoides</i>                          | 228         |

SEM1 family (WS0123\_P21), PF06376.2/unknown function (WS0112\_B13), and PF04689.3/DNA binding protein S1FA (WS01110\_K04).

#### **Annotation of poplar FLcDNA transcripts affected by FTC herbivory**

A major emphasis of the program that motivated the development and analysis of poplar FLcDNAs is the discovery of genes affected by insect attack. To identify herbivore-responsive genes among the poplar FLcDNAs, we first mapped the FLcDNA set onto a poplar 15.5 K microarray based on BLASTN comparison to ESTs spotted on the array. This microarray platform was previously used for profiling of the poplar leaf transcriptome affected by FTC larvae feeding [11]. Using a stringent similarity threshold of  $\geq 95\%$  identity over  $\geq 95\%$  alignment coverage, we identified 3,854 FLcDNAs that matched with 3,974 EST elements on the array (see Additional file 2). Although we did observe some cases of individual FLcDNAs mapping to multiple array elements, as well as mul-

multiple FLcDNAs mapping to the same array element, it should be noted that the *in silico* match stringency applied here is likely higher than the capability of cDNA microarrays to discriminate among highly similar transcripts by actual DNA hybridization. Next, we identified poplar FLcDNAs with a role in the response to insect attack by screening the 3,854 FLcDNAs against existing transcriptome data of differentially expressed (DE) genes in leaves that were exposed for 24 hours to FTC feeding [11]. This approach resulted in the identification of 129 and 24 FLcDNAs that were induced or repressed, respectively, in FTC-treated leaves compared to untreated control leaves (Tables 3 and 4) using the DE criteria of fold-change  $\geq 2.0$ -fold,  $P$  value  $< 0.05$  and  $Q$  value  $< 0.05$ . A complete list of expression data is provided [see Additional file 2]. Each of the 153 FLcDNAs was translated and evaluated for the presence of ORFs, and annotation was assigned based on manual examination of the highest scoring and most informative BLASTX matches in NR.



**Figure 6**  
**Sequence annotation of 4,664 high-quality poplar FLCDNAs against published databases.** Panel A shows the percentage of FLCDNAs with similarity to entries in three databases using expect (E) value thresholds of  $< 1e^{-05}$  and  $< 1e^{-50}$ : matches to previously published poplar ESTs (i.e., ESTs available in GenBank, excluding ESTs from this study) identified by BLASTN; amino acid sequences in the non-redundant (NR) division of GenBank identified by BLASTX; and The Arabidopsis Information Resource (TAIR) non-redundant Arabidopsis peptide matches identified by BLASTX. Panel B shows a Venn diagram of distinct and overlapping patterns of sequence similarity against the three databases (public poplar ESTs, TAIR, NR) at a BLAST E value threshold of  $< 1e^{-05}$ . At this threshold, 95 poplar FLCDNAs had no similarity to sequences in any of the databases examined.

Among FTC-induced transcripts represented with FLCDNAs, we identified a large number of defense-related and stress response proteins such as chitinases, Kunitz protease inhibitors, dehydrins, beta-1,3-glucanases, pathogenesis related protein PR-1, and glutathione-S-transferase (Table 3). Several classes of transcription factors (TFs) were also strongly affected by FTC feeding such as bZIP domain TFs, NAC domain TFs, NAM domain TFs and ethylene response factor TFs. A number of genes asso-

ciated with signaling were also strongly affected by FTC feeding, including allene oxide cyclase involved in jasmonate formation and calreticulin associated with calcium signaling. We also observed a substantial number of FLCDNAs annotated as involved in phenolic metabolism, particularly flavonoid biosynthesis, including isoflavone reductase, EPSP synthase, flavonoid 3-O-glycosyl transferase and flavanone 3-hydroxylase, along with several cytochrome P450s of unknown function (Table 3). Among the FTC-repressed transcripts represented with FLCDNAs, we observed photosystem II proteins associated with photosynthesis, malate dehydrogenase and thiamine biosynthesis enzyme associated with primary metabolism, several zinc finger TFs, and stress-responsive proteins such as small heat shock and universal stress proteins (Table 4). Twenty two of the 153 FTC-responsive genes represented with FLCDNAs matched to hypothetical proteins of unknown function and nine have no obvious similarity to any proteins in the NR database.

## Discussion

Previous studies using the biotinylated CAP trapper method for FLCDNA library construction have demonstrated this technique to be highly effective for capturing predominantly true full-length clones in large-scale projects [24,25,27]. In this study, we generated a set of 4,664 FLCDNAs, which represents the third largest plant FLCDNA resource published to date, behind only Arabidopsis and rice. CAP3 clustering and assembly indicates that more than 85% of the FLCDNAs are non-redundant within this collection. The average sequence length, ORF and UTR sizes of the poplar FLCDNAs were comparable to those observed with the CAP trapper-derived FLCDNA collections for maize [27], Arabidopsis [40] and rice [24], and were also very similar to the *ab initio* predicted reference genes in the poplar genome sequence [2]. Applying a reciprocal BLAST strategy, we demonstrated that among FLCDNAs with high sequence similarity to known Arabidopsis peptides and/or previously published poplar FLCDNAs, nearly 80% had similar ORF lengths and starting methionine and stop codon positions. Collectively, these data show that the poplar FLCDNA libraries are of high quality and that our clone selection strategy combined with the CAP trapper method was effective in capturing *bona fide* FLCDNAs from poplar.

Comparison of poplar FLCDNAs and the poplar genome sequence assembly confirmed both the overall high accuracy of the current genome assembly, as well as the quality of the FLCDNA resource described here. However, as has been previously demonstrated with efforts to identify the complete catalogue of genes in Arabidopsis and rice, gene prediction and genome assembly is an iterative process. The results reported here for the mapping of FLCDNAs to the poplar genome sequence reveal opportunities for

**Table 3: FLcDNAs corresponding to transcripts most strongly induced by forest tent caterpillar (FTC) feeding [fold-change (FC) ≥ 2.0, P value < 0.05, Q value < 0.05]**

| 15.5 K Array ID | Matching FLcDNA ID       | GenBank ID               | FL status/ORF size (aa) | NR BLASTP best match  |             | FTC feeding @ 24 h |        |        |
|-----------------|--------------------------|--------------------------|-------------------------|---|-------------|--------------------|--------|--------|
|                 |                          |                          |                         | GenBank accession, gene name, species   | BLAST score | FC                 | P      | Q      |
| WS0151_M13      | WS0131_K04 <sup>a</sup>  | <a href="#">EF148503</a> | FL/202                  | BAB85998.1, Kunitz trypsin inhibitor, <i>Populus nigra</i>                          | 396         | 60.4               | <0.001 | <0.001 |
| WS0132_F23      | WS0133_O14 <sup>a</sup>  | <a href="#">EF148554</a> | FL/202                  | BAB85997.1, Kunitz trypsin inhibitor, <i>P. nigra</i>                               | 380         | 50.2               | <0.001 | <0.001 |
| WS0134_B13      | WS0134_B13               | <a href="#">EF148557</a> | FL/212                  | AAQ84217.1, Kunitz trypsin inhibitor, <i>Populus trichocarpa</i> × <i>deltoides</i> | 387         | 46.2               | <0.001 | <0.001 |
| WS0133_N23      | WS0133_N23               | <a href="#">EF148553</a> | FL/197                  | CAJ21341.1, Kunitz trypsin inhibitor, <i>P. nigra</i>                               | 383         | 38.8               | <0.001 | <0.001 |
| WS0124_G12      | WS0124_G12               | <a href="#">EF147703</a> | FL/159                  | AAQ08196.1, translation initiation factor 5A, <i>Hevea brasiliensis</i>             | 316         | 29.0               | <0.001 | <0.001 |
| WS01223_D01     | WS01223_D01              | <a href="#">EF146918</a> | FL/359                  | At1g74320, choline kinase, <i>Arabidopsis thaliana</i>                              | 537         | 28.4               | <0.001 | <0.001 |
| WS0134_E16      | WS0134_E16               | <a href="#">EF148571</a> | 5' trunc./124           | AAA16342.1, vegetative storage protein, <i>P. trichocarpa</i> × <i>deltoides</i>    | 239         | 27.4               | <0.001 | <0.001 |
| WS01120_O24     | WS01120_O24              | <a href="#">EF145143</a> | 3' trunc./56            | At4g07960, putative glucosyltransferase, <i>A. thaliana</i>                         | 72          | 26.4               | <0.001 | <0.001 |
| WS01211_H19     | WS01211_H19              | <a href="#">EF146657</a> | FL/337                  | CAN72815, hypothetical protein, <i>Vitis vinifera</i>                               | 253         | 26.0               | <0.001 | <0.001 |
| WS0121_J16      | WS0122_N13               | <a href="#">EF147347</a> | FL/339                  | AAK01124.1, vegetative storage protein, <i>P. trichocarpa</i> × <i>deltoides</i>    | 509         | 25.4               | <0.001 | <0.001 |
| WS0141_P05      | WS0132_K10 <sup>a</sup>  | <a href="#">EF148516</a> | FL/202                  | AAQ84216.1, Kunitz trypsin inhibitor, <i>Populus trichocarpa</i> × <i>deltoides</i> | 386         | 22.7               | <0.001 | <0.001 |
| WS01118_D16     | WS01118_D16              | <a href="#">EF144781</a> | n.a.                    | No protein matches  | n.a.        | 16.8               | <0.001 | <0.001 |
| WS0168_C17      | WS01119_J20              | <a href="#">EF144899</a> | FL/285                  | AA43790.1, hypothetical protein, <i>Gossypium hirsutum</i>                          | 77          | 16.0               | <0.001 | <0.001 |
| WS01119_E18     | WS01119_E18              | <a href="#">EF144877</a> | 3' trunc./67            | At5g61770, brix domain-containing protein, <i>A. thaliana</i>                       | 85          | 15.7               | <0.001 | <0.001 |
| WS0133_B24      | WS0133_K20 <sup>a</sup>  | <a href="#">EF148543</a> | FL/202                  | CAH59150.1, Kunitz trypsin inhibitor, <i>Populus tremula</i>                        | 351         | 15.5               | <0.001 | <0.001 |
| WS0155_D02      | WS0138_H02 <sup>a</sup>  | <a href="#">EF148810</a> | FL/251                  | BAB21610.2, mangrin/allene oxide cyclase, <i>Bruguiera sexangula</i>                | 336         | 14.4               | <0.001 | <0.001 |
| WS0152_M24      | WS0128_J15               | <a href="#">EF148194</a> | FL/91                   | At5g24165, hypothetical protein, <i>A. thaliana</i>                                 | 72          | 13.7               | <0.001 | <0.001 |
| WS01118_N14     | WS01118_N14              | <a href="#">EF144837</a> | frameshift/47           | At4g27960, ubiquitin conjugating enzyme 9, <i>A. thaliana</i>                       | 96          | 13.2               | <0.001 | <0.001 |
| WS01212_M19     | WS0128_D22               | <a href="#">EF148166</a> | FL/509                  | ABA01477.1, cytochrome P450, <i>Gossypium hirsutum</i>                              | 726         | 12.3               | <0.001 | 0.002  |
| WS01211_N06     | WS0118_O23 <sup>a</sup>  | <a href="#">EF146529</a> | FL/225                  | ABS12347.1, dehydrin, <i>P. nigra</i>   | 167         | 11.8               | <0.001 | <0.001 |
| WS0132_A15      | WS01313_N19              | <a href="#">EF148368</a> | FL/396                  | At4g18550, lipase class 3 family protein, <i>A. thaliana</i>                        | 385         | 11.6               | <0.001 | 0.001  |
| WS01212_B20     | WS0128_L03               | <a href="#">EF148205</a> | FL/318                  | CAA73220.1, isoflavone reductase, <i>Citrus</i> × <i>paradise</i>                   | 469         | 10.4               | <0.001 | <0.001 |
| WS0122_C03      | WS0122_C03               | <a href="#">EF147271</a> | FL/133                  | CAN82925.1, hypothetical protein, <i>V. vinifera</i>                                | 114         | 9.2                | <0.001 | 0.001  |
| WS0113_H20      | WS0113_H20               | <a href="#">EF145803</a> | n.a.                    | No protein matches  | n.a.        | 8.8                | <0.001 | <0.001 |
| WS0134_J14      | WS0134_J14 <sup>a</sup>  | <a href="#">EF148597</a> | FL/202                  | AAQ84216.1, Kunitz trypsin inhibitor, <i>P. trichocarpa</i> × <i>deltoides</i>      | 380         | 7.9                | <0.001 | <0.001 |
| WS01120_N21     | WS01120_N21              | <a href="#">EF145138</a> | n.a.                    | No protein matches  | n.a.        | 6.9                | <0.001 | <0.001 |
| WS0114_H12      | WS0114_H12               | <a href="#">EF145947</a> | FL/252                  | At4g01470, major intrinsic family protein, <i>A. thaliana</i>                       | 364         | 6.3                | <0.001 | <0.001 |
| WS0126_E15      | WS0126_E15               | <a href="#">EF147963</a> | FL/325                  | At1g30910, molybdenum cofactor sulfuryase family protein, <i>A. thaliana</i>        | 444         | 6.2                | <0.001 | <0.001 |
| WS0168_F14      | WS01123_O20              | <a href="#">EF145380</a> | FL/217                  | At3g18030, phosphopantothoenyl cysteine decarboxylase, <i>A. thaliana</i>           | 350         | 6.2                | <0.001 | <0.001 |
| PX0019_C05      | PX0019_C05               | <a href="#">EF144379</a> | FL/214                  | AAF64453.1, heat-shock protein 90, <i>Euphorbia esula</i>                           | 330         | 5.7                | <0.001 | <0.001 |
| WS0205_K16      | WS01214_G11              | <a href="#">EF146815</a> | FL/387                  | CAN71454.1, hypothetical protein, <i>V. vinifera</i>                                | 682         | 5.6                | <0.001 | <0.001 |
| WS0152_N17      | WS0114_F10 <sup>a</sup>  | <a href="#">EF145928</a> | FL/70                   | BAA03527.1, ATP synthase epsilon subunit, <i>Ipomoea batatas</i>                    | 120         | 5.6                | <0.001 | 0.001  |
| WS01118_A11     | WS0113_M04               | <a href="#">EF145848</a> | FL/97                   | At1g77710, ubiquitin-fold modifier precursor, <i>A. thaliana</i>                    | 150         | 5.5                | <0.001 | <0.001 |
| WS0132_L23      | WS0132_L23               | <a href="#">EF148518</a> | FL/372                  | AAP87281.1, beta-1,3-glucanase, <i>Hevea brasiliensis</i>                           | 540         | 5.4                | <0.001 | 0.002  |
| WS0124_C22      | WS0124_C22               | <a href="#">EF147658</a> | 5' trunc./142           | CAA42660.1, luminal binding protein, <i>Nicotiana tabacum</i>                       | 213         | 5.4                | <0.001 | <0.001 |
| WS01116_C06     | WS01123_N20              | <a href="#">EF145376</a> | FL/250                  | At4g38210, expansin A20 precursor, <i>A. thaliana</i>                               | 351         | 5.2                | <0.001 | <0.001 |
| WS0114_D04      | WS01211_M02 <sup>a</sup> | <a href="#">EF146676</a> | FL/414                  | AAB71419.1, calreticulin, <i>Ricinus communis</i>                                   | 556         | 5.0                | <0.001 | <0.001 |
| WS01117_O15     | WS01117_O15              | <a href="#">EF144759</a> | FL/230                  | At4g11150, Vacuolar ATP synthase subunit E1, <i>A. thaliana</i>                     | 295         | 4.7                | <0.001 | <0.001 |
| WS0133_J24      | WS0133_J24               | <a href="#">EF148541</a> | FL/177                  | At1g01250, AP2 transcription factor, <i>A. thaliana</i>                             | 303         | 4.6                | 0.001  | 0.004  |
| WS0148_P02      | WS0127_F13               | <a href="#">EF148073</a> | 5' trunc./283           | At1g64660, methionine gamma-lyase, <i>A. thaliana</i>                               | 424         | 4.5                | <0.001 | 0.001  |
| WS02010_D02     | WS0126_C10 <sup>a</sup>  | <a href="#">EF147943</a> | FL/68                   | NP_001066879.1, hypothetical protein, <i>Oryza sativa</i>                           | 175         | 4.4                | <0.001 | <0.001 |
| WS0155_H06      | WS0125_E23               | <a href="#">EF147828</a> | FL/215                  | CAN69111.1, glutathione-S-transferase, <i>V. vinifera</i>                           | 415         | 4.3                | <0.001 | <0.001 |

**Table 3: FLcDNAs corresponding to transcripts most strongly induced by forest tent caterpillar (FTC) feeding [fold-change (FC) ≥ 2.0, P value < 0.05, Q value < 0.05] (Continued)**

|             |                          |                          |                |   |      |     |        |        |
|-------------|--------------------------|--------------------------|----------------|---|------|-----|--------|--------|
| WS01119_L18 | WS01119_L18              | <a href="#">EFI44906</a> | FL/56          | NP_001068325.1, 40S ribosomal protein, <i>O. sativa</i>                           | 182  | 4.3 | <0.001 | <0.001 |
| WS0134_F23  | WS0134_F23               | <a href="#">EFI48579</a> | FL/312         | CAN79077.1, annexin, <i>V. vinifera</i>   | 575  | 4.2 | <0.001 | <0.001 |
| WS0117_C05  | WS0124_M24               | <a href="#">EFI47756</a> | FL/538         | AAA80588.1, calnexin, <i>Glycine max</i>  | 1231 | 4.1 | <0.001 | <0.001 |
| WS0175_A23  | WS01125_H02 <sup>a</sup> | <a href="#">EFI45504</a> | FL/181         | AAT08648.1, ADP-ribosylation factor, <i>Hyacinthus orientalis</i>                 | 587  | 4.0 | 0.004  | 0.014  |
| WS0153_O15  | WS0135_A12               | <a href="#">EFI48616</a> | FL/388         | At4g24220, vein patterning 1, <i>A. thaliana</i>                                  | 711  | 4.0 | <0.001 | <0.001 |
| WS0141_G12  | WS01312_A02              | <a href="#">EFI48234</a> | FL/273         | At1g19180, hypothetical protein, <i>A. thaliana</i>                               | 160  | 4.0 | <0.001 | 0.003  |
| WS0168_D23  | WS01230_E07              | <a href="#">EFI47385</a> | FL/420         | ABD32854.1, hypothetical protein, <i>Medicago truncatula</i>                      | 670  | 4.0 | <0.001 | 0.001  |
| WS0154_B02  | WS01228_N21              | <a href="#">EFI47184</a> | 5' trunc./186  | At5g07340, calnexin, <i>A. thaliana</i>   | 251  | 3.9 | <0.001 | <0.001 |
| WS01116_D23 | WS01116_D23              | <a href="#">EFI44634</a> | FL/84          | At3g60540, sec61 beta family protein, <i>A. thaliana</i>                          | 92   | 3.8 | <0.001 | <0.001 |
| WS0117_O22  | WS0117_O22 <sup>a</sup>  | <a href="#">EFI46403</a> | FL/68          | At1g27330, hypothetical protein, <i>A. thaliana</i>                               | 103  | 3.5 | <0.001 | <0.001 |
| WS0122_A01  | WS01227_N20              | <a href="#">EFI47117</a> | FL/399         | At1g74210, glycerophosphodiester phosphodiesterase, <i>A. thaliana</i>            | 606  | 3.5 | <0.001 | <0.001 |
| WS0144_K08  | WS01119_H21              | <a href="#">EFI44889</a> | FL/358         | ABQ10199.1, cysteine protease, <i>Actinidia deliciosa</i>                         | 594  | 3.5 | <0.001 | <0.001 |
| WS0147_I02  | WS0125_D08               | <a href="#">EFI47814</a> | FL/444         | AAS79603.1, prephenate dehydratase, <i>Ipomoea trifida</i>                        | 653  | 3.3 | <0.001 | 0.001  |
| WS0111_C18  | WS0125_B22 <sup>a</sup>  | <a href="#">EFI47800</a> | FL/395         | P47916, S-adenosyl methionine synthetase, <i>P. deltoides</i>                     | 785  | 3.3 | <0.001 | 0.001  |
| WS0151_N14  | WS0127_M05               | <a href="#">EFI48121</a> | FL/485         | Q01781, S-adenosylhomocysteine hydrolase, <i>Petroselinum crispum</i>             | 939  | 3.3 | <0.001 | <0.001 |
| WS01212_P09 | WS01212_P09              | <a href="#">EFI46734</a> | FL/161         | ABC47922.1, pathogenesis-related protein 1, <i>Malus × domestica</i>              | 236  | 3.2 | 0.005  | 0.016  |
| PX0015_M10  | PX0015_M10               | <a href="#">EFI44335</a> | n.a.           | No protein matches  | n.a. | 3.2 | <0.001 | <0.001 |
| WS0111_A20  | WS0111_A20               | <a href="#">EFI44935</a> | FL/360         | CAN67616.1, cupin family protein, <i>V. vinifera</i>                              | 474  | 3.2 | <0.001 | <0.001 |
| WS0117_P18  | WS0117_P18               | <a href="#">EFI46411</a> | FL/93          | NP_001047293.1, hypoxia-responsive family protein, <i>O. sativa</i>               | 122  | 3.2 | <0.001 | <0.001 |
| WS0131_J08  | WS0131_J08               | <a href="#">EFI48502</a> | FL/452         | AAA70334.1, omega-3 fatty acid desaturase, <i>Sesamum indicum</i>                 | 708  | 3.1 | <0.001 | <0.001 |
| WS0173_J22  | WS01229_P15              | <a href="#">EFI47254</a> | frameshift/441 | CAH05011.1, alpha-dioxygenase, <i>Pisum sativum</i>                               | 679  | 3.1 | <0.001 | 0.002  |
| WS0151_H21  | WS01314_F07 <sup>a</sup> | <a href="#">EFI48393</a> | FL/505         | AAB05641.1, protein disulphide isomerase, <i>R. communis</i>                      | 786  | 3.1 | <0.001 | <0.001 |
| WS0141_E06  | WS0128_M17               | <a href="#">EFI48216</a> | FL/338         | CAN79663.1, hypothetical protein, <i>V. vinifera</i>                              | 284  | 3.0 | <0.001 | <0.001 |
| WS01211_D15 | WS01211_D15              | <a href="#">EFI46643</a> | FL/258         | NP_001061550.1, 60S ribosomal protein L7A, <i>O. sativa</i>                       | 398  | 3.0 | 0.004  | 0.012  |
| WS01110_A05 | WS01110_A05              | <a href="#">EFI44530</a> | 5' trunc./46   | AAT45244.1, EPSP synthase, <i>Conyza canadensis</i>                               | 87   | 3.0 | <0.001 | <0.001 |
| WS0122_A21  | WS0122_A21               | <a href="#">EFI47261</a> | FL/349         | At3g62600, DNAJ heat shock family protein, <i>A. thaliana</i>                     | 542  | 3.0 | <0.001 | <0.001 |
| WS0154_D16  | PX0019_K19               | <a href="#">EFI44475</a> | FL/172         | ABL67655.1, cyclophilin, <i>Citrus cv. Shiranuhi</i>                              | 303  | 3.0 | <0.001 | <0.001 |
| WS0114_N12  | WS0114_N12               | <a href="#">EFI46003</a> | 5' trunc./243  | AAU08208.1, chloroplast ferritin precursor, <i>Vigna angularis</i>                | 357  | 3.0 | 0.001  | 0.007  |
| WS0153_O16  | WS0136_K07 <sup>a</sup>  | <a href="#">EFI48708</a> | FL/113         | CAA40072.1, hypothetical protein, <i>P. trichocarpa × deltoides</i>               | 225  | 2.9 | <0.001 | <0.001 |
| WS01117_D04 | WS01117_D04              | <a href="#">EFI44703</a> | FL/137         | CAN73155.1, hypothetical protein, <i>V. vinifera</i>                              | 110  | 2.9 | <0.001 | <0.001 |
| WS01120_A02 | WS01120_A02              | <a href="#">EFI45080</a> | 5' trunc./105  | At1g03010, phototropic-responsive NPH3 family protein, <i>A. thaliana</i>         | 177  | 2.8 | <0.001 | 0.001  |
| WS0178_L06  | WS01211_M01              | <a href="#">EFI46675</a> | FL/415         | NP_001064428.1, no apical meristem transcription factor, <i>O. sativa</i>         | 98   | 2.8 | <0.001 | 0.001  |
| WS0143_C23  | WS01228_M23 <sup>a</sup> | <a href="#">EFI47179</a> | FL/212         | ABB89210.1, dehydroascorbate reductase, <i>S. indicum</i>                         | 343  | 2.7 | <0.001 | <0.001 |
| WS0127_I09  | WS0127_I09               | <a href="#">EFI48095</a> | FL/235         | CAB77025.1, Rho GDP dissociation inhibitor, <i>N. tabacum</i>                     | 294  | 2.7 | 0.003  | 0.012  |
| PX0015_K10  | PX0015_K10               | <a href="#">EFI44326</a> | 3' trunc./65   | At2g15590, hypothetical protein, <i>A. thaliana</i>                               | 39   | 2.7 | 0.001  | 0.004  |
| WS0152_M05  | WS01111_A23              | <a href="#">EFI44570</a> | FL/125         | At1g69230, nitrilase-associated protein, <i>A. thaliana</i>                       | 80   | 2.7 | 0.001  | 0.006  |
| WS0134_H19  | WS0134_H19               | <a href="#">EFI48589</a> | FL/461         | At5g28237, tryptophan synthase, <i>A. thaliana</i>                                | 579  | 2.7 | <0.001 | 0.001  |
| WS0122_P22  | WS0122_P22               | <a href="#">EFI47367</a> | 5' trunc./46   | AAS89832.1, flavonoid 3-O-glucosyltransferase, <i>Fragaria × ananassa</i>         | 47   | 2.6 | 0.009  | 0.023  |
| WS0113_E03  | WS0113_E03               | <a href="#">EFI45764</a> | 5' trunc./130  | At1g73600, phosphoethanolamine N-methyltransferase, <i>A. thaliana</i>            | 198  | 2.6 | <0.001 | 0.001  |
| WS02012_L20 | WS01212_L02 <sup>a</sup> | <a href="#">EFI46720</a> | FL/440         | AAV50009.1, N-hydroxycinnamoyl/benzoyltransferase, <i>Malus × domestica</i>       | 451  | 2.5 | <0.001 | 0.001  |
| WS0116_I22  | WS01119_O01 <sup>a</sup> | <a href="#">EFI44919</a> | FL/212         | ABB89210.1, dehydroascorbate reductase, <i>S. indicum</i>                         | 360  | 2.5 | <0.001 | 0.001  |
| WS0128_C01  | WS0128_C01               | <a href="#">EFI48156</a> | FL/205         | CAC85245.1, salt tolerance protein, <i>Beta vulgaris</i>                          | 246  | 2.5 | 0.001  | 0.005  |
| PX0011_E19  | PX0011_C19               | <a href="#">EFI44204</a> | FL/341         | At1g10840, eukaryotic translation initiation factor subunit 3, <i>A. thaliana</i> | 573  | 2.5 | <0.001 | 0.002  |
| WS0128_M01  | WS0128_M01               | <a href="#">EFI48209</a> | 5' trunc./197  | ABN08481.1, homeodomain-related, <i>M. truncatula</i>                             | 103  | 2.4 | <0.001 | 0.003  |
| WS01126_B13 | WS01126_B13              | <a href="#">EFI45551</a> | 3' trunc./136  | CAN77060.1, ubiquitin activating enzyme, <i>V. vinifera</i>                       | 239  | 2.4 | 0.017  | 0.035  |
| WS01125_E14 | WS01125_E14 <sup>a</sup> | <a href="#">EFI45493</a> | FL/207         | NP_001058535.1, cyclophilin, <i>O. sativa</i>                                     | 340  | 2.4 | <0.001 | 0.001  |

**Table 3: FLcDNAs corresponding to transcripts most strongly induced by forest tent caterpillar (FTC) feeding [fold-change (FC) ≥ 2.0, P value < 0.05, Q value < 0.05] (Continued)**

|             |                          |                          |                |   |      |     |        |        |
|-------------|--------------------------|--------------------------|----------------|---|------|-----|--------|--------|
| WS01218_P22 | WS01120_G07 <sup>a</sup> | <a href="#">EF145102</a> | FL/170         | NP_001050870.1, glycine-rich RNA-binding protein, <i>O. sativa</i>      | 144  | 2.4 | 0.004  | 0.013  |
| WS01117_L06 | WS01117_L06              | <a href="#">EF144744</a> | frameshift/136 | NP_001046690.1, ribosomal protein L10A, <i>O. sativa</i>                | 171  | 2.4 | <0.001 | <0.001 |
| WS01117_E15 | WS01117_E15              | <a href="#">EF144711</a> | n.a.           | No protein matches  | n.a. | 2.4 | <0.001 | 0.001  |
| WS01110_A14 | WS0122_K19               | <a href="#">EF147330</a> | FL/476         | AAF18411.1, integral membrane protein, <i>Phaseolus vulgaris</i>        | 897  | 2.4 | <0.001 | <0.001 |
| WS0156_A21  | WS0127_G12 <sup>a</sup>  | <a href="#">EF148080</a> | n.a.           | No protein matches  | n.a. | 2.4 | 0.017  | 0.035  |
| WS0127_G19  | WS0127_G19               | <a href="#">EF148082</a> | frameshift/251 | At4g11640, serine racemase, <i>A. thaliana</i>                          | 354  | 2.4 | <0.001 | 0.002  |
| WS0112_O04  | WS0112_O04               | <a href="#">EF145713</a> | 5' trunc./566  | ABS01352.1, methionine synthase, <i>Carica papaya</i>                   | 1073 | 2.4 | <0.001 | 0.001  |
| WS0155_E17  | WS01212_I06 <sup>a</sup> | <a href="#">EF146705</a> | FL/363         | ABM67589.1, flavanone 3-hydroxylase, <i>V. vinifera</i>                 | 645  | 2.4 | 0.003  | 0.012  |
| WS0168_M07  | WS0137_HI3 <sup>a</sup>  | <a href="#">EF148760</a> | FL/62          | ABF98145.1, hypothetical protein, <i>O. sativa</i>                      | 57   | 2.4 | <0.001 | 0.003  |
| WS0119_H18  | WS0117_P08               | <a href="#">EF146405</a> | 5' trunc./188  | CAN83141.1, hypothetical protein, <i>V. vinifera</i>                    | 218  | 2.3 | <0.001 | 0.003  |
| WS0157_L22  | WS0128_B17               | <a href="#">EF148154</a> | 5' trunc./388  | CAN76057.1, glucosyltransferase, <i>V. vinifera</i>                     | 411  | 2.3 | 0.002  | 0.008  |
| WS0185_E12  | WS0124_A18               | <a href="#">EF147646</a> | FL/285         | CAH60723.1, aquaporin, <i>P. tremula</i> × <i>tremuloides</i>           | 488  | 2.3 | 0.001  | 0.007  |
| WS0125_I01  | WS0125_I01               | <a href="#">EF147858</a> | FL/477         | BAA36972.1, flavonoid 3-O-galactosyl transferase, <i>Vigna mungo</i>    | 442  | 2.3 | 0.003  | 0.011  |
| PX0019_C07  | PX0019_C07               | <a href="#">EF144380</a> | 5' trunc./222  | CAN74465.1, hypothetical protein, <i>V. vinifera</i>                    | 369  | 2.3 | 0.015  | 0.033  |
| WS01111_E24 | WS0113_P06               | <a href="#">EF145877</a> | FL/290         | AAN32641.1, short-chain alcohol dehydrogenase, <i>Solanum tuberosum</i> | 399  | 2.3 | <0.001 | 0.003  |
| WS01212_B14 | WS01214_D06 <sup>a</sup> | <a href="#">EF146806</a> | FL/363         | ABM67589.1, flavanone 3-hydroxylase, <i>V. vinifera</i>                 | 644  | 2.3 | 0.003  | 0.011  |
| WS0181_A04  | WS01312_M14              | <a href="#">EF148294</a> | frameshift/232 | CAN74806, bZIP transcription factor, <i>V. vinifera</i>                 | 152  | 2.3 | 0.002  | 0.009  |
| WS0116_F22  | WS0116_F22               | <a href="#">EF146228</a> | frameshift/239 | At3g05290, mitochondrial substrate carrier protein, <i>A. thaliana</i>  | 283  | 2.3 | 0.004  | 0.013  |
| WS01121_C12 | WS01121_C12              | <a href="#">EF145159</a> | FL/216         | At2g25110, MIR domain-containing protein, <i>A. thaliana</i>            | 349  | 2.3 | <0.001 | <0.001 |
| WS01214_P11 | WS01214_P11              | <a href="#">EF146849</a> | FL/219         | ABL84692, glutathione S-transferase, <i>V. vinifera</i>                 | 345  | 2.3 | 0.002  | 0.009  |
| WS0128_G16  | WS01228_N10              | <a href="#">EF147182</a> | FL/207         | AAN03471.1, hypothetical protein, <i>G. max</i>                         | 99   | 2.2 | <0.001 | <0.001 |
| WS0209_J01  | WS0135_O22               | <a href="#">EF148667</a> | FL/318         | AAG23965.1, endochitinase, <i>Vigna sesquipedalis</i>                   | 461  | 2.2 | 0.001  | 0.004  |
| WS01119_M12 | WS01110_H18              | <a href="#">EF144553</a> | FL/118         | At5g04750, F1F0-ATPase inhibitor protein, <i>A. thaliana</i>            | 52   | 2.2 | <0.001 | <0.001 |
| WS0205_L05  | WS01228_D08              | <a href="#">EF147142</a> | frameshift/233 | AAX85981.1, NAC4 protein, <i>G. max</i>                                 | 362  | 2.2 | 0.019  | 0.038  |
| WS0123_D13  | WS0137_E08               | <a href="#">EF148737</a> | FL/533         | At5g58270, STARK1 ATPase, half ABC transporter, <i>A. thaliana</i>      | 642  | 2.2 | <0.001 | <0.001 |
| WS0112_P02  | WS0116_L21               | <a href="#">EF146273</a> | FL/145         | At5g27670, histone 2A, <i>A. thaliana</i>                               | 196  | 2.2 | <0.001 | 0.002  |
| WS01214_A14 | WS01225_E15              | <a href="#">EF146945</a> | FL/330         | At5g07010, sulfotransferase family protein, <i>A. thaliana</i>          | 394  | 2.2 | 0.002  | 0.009  |
| WS01211_G15 | WS01211_G15              | <a href="#">EF146653</a> | FL/507         | AAL24049.1, cytochrome P450, <i>Citrus sinensis</i>                     | 677  | 2.2 | <0.001 | 0.002  |
| WS0123_E09  | WS0123_E09               | <a href="#">EF147535</a> | FL/210         | ABB89210.1, dehydroascorbate reductase, <i>S. indicum</i>               | 332  | 2.2 | <0.001 | <0.001 |
| WS0114_N11  | WS0114_N11               | <a href="#">EF146002</a> | 5' trunc./313  | AAF73006.1, NADP-dependent malic enzyme, <i>R. communis</i>             | 450  | 2.1 | <0.001 | <0.001 |
| WS0154_G22  | WS0122_L10               | <a href="#">EF147335</a> | 5' trunc./381  | CAN74204.1, hypothetical protein, <i>V. vinifera</i>                    | 535  | 2.1 | 0.001  | 0.005  |
| WS0181_N15  | WS0133_H05               | <a href="#">EF148536</a> | FL/283         | ABG73415.1, chloroplast pigment-binding protein, <i>N. tabacum</i>      | 496  | 2.1 | <0.001 | 0.001  |
| WS0131_L08  | WS0137_P12 <sup>a</sup>  | <a href="#">EF148792</a> | FL/214         | NP_001060368.1, emp24/gp25L/p24 transmembrane protein, <i>O. sativa</i> | 288  | 2.1 | <0.001 | <0.001 |
| WS0124_N24  | WS0124_N24               | <a href="#">EF147765</a> | FL/584         | NP_001048852.1, acyl-activating enzyme 11, <i>O. sativa</i>             | 750  | 2.1 | 0.017  | 0.036  |
| WS0116_E14  | WS0116_E14               | <a href="#">EF146213</a> | n.a.           | No protein matches  | n.a. | 2.1 | 0.001  | 0.004  |
| WS0128_N06  | WS0128_N06               | <a href="#">EF148221</a> | FL/257         | At4g18260, cytochrome b-561, <i>A. thaliana</i>                         | 294  | 2.1 | 0.005  | 0.016  |
| WS01122_N10 | WS01122_N10              | <a href="#">EF145286</a> | FL/91          | At1g62440, leucine-rich repeat extensin, <i>A. thaliana</i>             | 107  | 2.0 | 0.010  | 0.025  |
| WS01214_M13 | WS01214_M13              | <a href="#">EF146841</a> | FL/378         | At5g45670, GDSL-motif/hydrolase family protein, <i>A. thaliana</i>      | 298  | 2.0 | <0.001 | 0.001  |
| WS01213_H17 | WS01213_H17              | <a href="#">EF146756</a> | FL/597         | At4g34200, phosphoglycerate dehydrogenase, <i>A. thaliana</i>           | 884  | 2.0 | <0.001 | 0.003  |
| WS01122_N02 | WS01231_J04 <sup>a</sup> | <a href="#">EF147472</a> | FL/196         | XP_001334748.1, hypothetical protein, <i>Danio rerio</i>                | 59   | 2.0 | 0.003  | 0.010  |
| WS0156_F12  | WS0118_O10               | <a href="#">EF146525</a> | FL/102         | At2g18400, ribosomal protein L6, <i>A. thaliana</i>                     | 165  | 2.0 | <0.001 | <0.001 |

<sup>a</sup>Multiple FLcDNAs match to the same microarray EST, a complete list of matching FLcDNAs is provided elsewhere [see Additional file 2].

**Table 4: FLCDNAs corresponding to transcripts most strongly repressed by forest tent caterpillar (FTC) feeding [fold-change (FC) ≥ 2.0, P value < 0.05, Q value < 0.05]**

| 15.5 K Array ID | Matching FLCDNA ID       | GenBank ID               | FL status/ORF size (aa) | NR BLASTP best match   |             | FTC feeding @ 24 h |        |        |
|-----------------|--------------------------|--------------------------|-------------------------|--|-------------|--------------------|--------|--------|
|                 |                          |                          |                         | GenBank accession, gene name, species  | BLAST score | FC                 | P      | Q      |
| WS0162_B18      | WS01227_D07              | <a href="#">EF147075</a> | FL/465                  | AAX84673.1, cysteine protease, <i>Manihot esculenta</i>                          | 782         | 0.33               | <0.001 | <0.001 |
| WS0112_D20      | WS0112_D20               | <a href="#">EF145637</a> | FL/99                   | At1g67910, hypothetical protein, <i>Arabidopsis thaliana</i>                     | 69          | 0.34               | <0.001 | 0.001  |
| WS0126_C06      | WS0126_C06               | <a href="#">EF147942</a> | FL/121                  | At2g45180, protease inhibitor/lipid transfer protein, <i>A. thaliana</i>         | 108         | 0.34               | 0.018  | 0.038  |
| WS0131_P03      | WS0131_P03 <sup>a</sup>  | <a href="#">EF148510</a> | FL/303                  | CAN63090.1, zinc finger transcription factor, <i>Vitis vinifera</i>              | 135         | 0.36               | <0.001 | 0.001  |
| WS0178_F11      | WS01228_M08              | <a href="#">EF147174</a> | 5' trunc./106           | At1g22770, gigantea protein, <i>A. thaliana</i>                                  | 150         | 0.38               | <0.001 | 0.002  |
| WS0127_F15      | WS0127_F15               | <a href="#">EF148074</a> | FL/173                  | CAN68427.1, hypothetical protein, <i>V. vinifera</i>                             | 207         | 0.40               | <0.001 | 0.001  |
| WS0121_B24      | WS0128_M21               | <a href="#">EF148217</a> | FL/139                  | AAU03358.1, acyl carrier protein, <i>Lycopersicon esculentum</i>                 | 119         | 0.41               | <0.001 | <0.001 |
| WS0147_J04      | WS0134_M10               | <a href="#">EF148605</a> | n.a.                    | No protein matches   | n.a.        | 0.41               | 0.004  | 0.014  |
| WS0158_G10      | WS0128_E13               | <a href="#">EF148173</a> | 5' trunc./628           | At1g56070, elongation factor, <i>A. thaliana</i>                                 | 1239        | 0.41               | 0.001  | 0.005  |
| WS0152_E14      | WS0112_O08 <sup>a</sup>  | <a href="#">EF145715</a> | FL/252                  | ABH09330.1, aquaporin, <i>V. vinifera</i>  | 375         | 0.42               | <0.001 | 0.003  |
| WS0143_B24      | WS01227_O15              | <a href="#">EF147121</a> | FL/267                  | At1g06460, small heat shock protein, <i>A. thaliana</i>                          | 146         | 0.42               | <0.001 | 0.001  |
| WS0127_G18      | WS0127_G18               | <a href="#">EF148081</a> | n.a.                    | No protein matches   | n.a.        | 0.43               | <0.001 | <0.001 |
| WS0182_D02      | WS01226_N23              | <a href="#">EF147055</a> | FL/335                  | CAN75691.1, methyltransferase, <i>V. vinifera</i>                                | 534         | 0.43               | 0.001  | 0.005  |
| WS0124_D16      | WS0124_D16               | <a href="#">EF147668</a> | FL/164                  | At3g62550, universal stress protein, <i>A. thaliana</i>                          | 188         | 0.44               | <0.001 | 0.001  |
| WS0163_G24      | WS0115_E02               | <a href="#">EF146059</a> | FL/341                  | AAD56659.1, malate dehydrogenase, <i>Glycine max</i>                             | 566         | 0.45               | 0.003  | 0.010  |
| WS0175_O14      | WS01313_J01 <sup>a</sup> | <a href="#">EF148349</a> | FL/239                  | CAN63226.1, hypothetical protein, <i>V. vinifera</i>                             | 313         | 0.45               | <0.001 | 0.001  |
| WS0178_N22      | WS01111_H24              | <a href="#">EF144589</a> | FL/161                  | ABG27020.1, SKP1-like ubiquitin-protein ligase, <i>Medicago truncatula</i>       | 219         | 0.46               | <0.001 | <0.001 |
| WS0121_H19      | WS0121_H19               | <a href="#">EF146882</a> | FL/350                  | AAW66657.1, thiamine biosynthetic enzyme, <i>Picrorhiza kurroa</i>               | 539         | 0.48               | 0.005  | 0.016  |
| WS0206_B21      | WS0131_B11               | <a href="#">EF148494</a> | FL/133                  | CAA59409.1, photosystem II reaction center protein, <i>Spinacia oleracea</i>     | 140         | 0.48               | 0.001  | 0.006  |
| WS0155_M12      | WS0136_E20               | <a href="#">EF148683</a> | FL/234                  | CAN60736.1, hypothetical protein, <i>V. vinifera</i>                             | 313         | 0.48               | 0.001  | 0.007  |
| WS0152_F02      | WS01117_K24              | <a href="#">EF144742</a> | FL/384                  | CAN83255.1, CCH-type zinc finger protein, <i>V. vinifera</i>                     | 432         | 0.49               | <0.001 | 0.002  |
| WS01224_P10     | WS0124_L08 <sup>a</sup>  | <a href="#">EF147742</a> | FL/137                  | CAA28450.1, photosystem II 10 kDa polypeptide, <i>Solanum tuberosum</i>          | 191         | 0.49               | <0.001 | 0.003  |
| WS0115_N05      | WS0115_N05               | <a href="#">EF146146</a> | FL/250                  | AAM21317.1, auxin-regulated protein, <i>Populus tremula</i> × <i>tremuloides</i> | 449         | 0.50               | 0.005  | 0.016  |
| WS0125_F02      | WS0125_F02               | <a href="#">EF147829</a> | FL/516                  | At1g60590, polygalacturonase, <i>A. thaliana</i>                                 | 715         | 0.50               | 0.001  | 0.005  |

<sup>a</sup>Multiple FLCDNAs match to the same microarray EST, a complete list of matching FLCDNAs is provided elsewhere [see Additional file 2].

improvement of the genome sequence assembly (i.e., targeting apparent gaps for re-sequencing), as well as opportunities to further improve tools for the *in silico* prediction of genes. To address the discovery of apparent gaps in the genome assembly, the availability of 39 FLCDNAs that are not covered in the current assembly could be used to target BAC clones for re-sequencing and filling of gap regions. Similarly, the discovery of 173 FLCDNAs that do not have corresponding gene predictions in the current genome annotation may provide an opportunity to further improve gene prediction tools for poplar. Algorithms used for gene prediction in the poplar genome sequence assembly could be tested with these 173 FLCDNAs to find out why they may have initially been missed. If this leads to an improvement of prediction tools, the assembled genome sequence could be tested with the modified tools to identify additional genes.

The comparative sequence annotation of poplar FLCDNAs against Arabidopsis, the NR database, and previously pub-

lished poplar ESTs revealed that *ca.* 88% of poplar FLCDNAs showed similarity to sequences in Arabidopsis or other plants. Many of the *ca.* 11.5% of poplar FLCDNAs without significant sequence similarity in Arabidopsis or other plants are supported with evidence of gene expression in the form of previously published poplar ESTs and matching the poplar genome sequence, thus excluding the possibility that they are artifacts of cDNA library construction. The discovery of poplar FLCDNAs without matches in other plant species is also in agreement with previous analysis of the poplar genome sequence where 11% of predicted proteins had no similarity to proteins in the NR database and 12% had no similarity to Arabidopsis proteins [2]. For comparison, only 64% of the 28,444 ORFs derived from rice FLCDNAs showed significant similarity to coding sequences predicted from the Arabidopsis genome and conversely, only 75% of Arabidopsis coding sequences had similarity to rice FLCDNAs [24]. These findings suggest that a substantial proportion of protein-coding sequences are not conserved among all plant species.

The putative poplar-specific genes could be the product of past local or whole genome duplications in the lineage that led to extant poplar species [2,43] followed by sequence divergence [44,45]. Furthermore, *ca.* 2% of poplar FLcDNAs did not contain a predicted ORF suggesting these putative poplar-specific genes likely encode non-coding RNAs (i.e., rRNAs, tRNAs, snoRNAs etc.).

## Conclusion

We developed a large FLcDNA resource of high sequence quality and low-level redundancy that facilitated the discovery of a substantial number of genes not present among the published sequences of other plant species, and that also facilitated the discovery of several hundred insect-affected genes in the poplar leaf transcriptome that were represented by FLcDNAs. The newly established poplar FLcDNA resource will be valuable for further improvement of the poplar genome assembly, annotation of protein-coding regions, and for functional and comparative analysis of poplar genes. Specifically, the identification of FLcDNAs that are not covered in the current genome assembly or that were not predicted during the genome annotation provides opportunities to further refine the current genome assembly. The availability of a large collection of FLcDNAs that show altered gene expression following insect herbivory affords more rapid characterization of the role of these genes in poplar biotic interactions.

## Methods

### Full-length cDNA libraries

Plant materials used in the construction of cDNA libraries are described in Table 1. Isolation of total and poly(A)<sup>+</sup> RNA are described elsewhere (see Additional file 3). FLcDNA libraries were directionally constructed (5' *Sst*I and 3' *Xho*I) according to published methods [46,47], with modifications described in detail elsewhere (see Additional file 3).

### DNA sequencing and sequence filtering

Details of bacterial transformation with plasmids, clone handling, DNA purification and evaluation, and DNA sequencing are provided elsewhere (see Additional file 3). Sequences from each cDNA library were closely monitored to assess library complexity and sequence quality. DNA sequence chromatograms were processed using the PHRED software (versions 0.000925.c and 0.020425.c) [48,49]. Sequences were quality-trimmed according to the high-quality (hq) contiguous region determined by PHRED and vector-trimmed using CROSS\_MATCH software [50]. Sequences with less than 100 quality bases (Phred 20 or better) after trimming and sequences having polyA tails of  $\geq 100$  bases were removed from analysis. Also removed were sequences representing bacterial, yeast or fungal contaminations identified by BLAST searches

[51,52] against *E. coli* K12 DNA sequence (GI: 6626251), *Saccharomyces cerevisiae* [53], *Aspergillus nidulans* (TIGR ANGI.060302), and *Agrobacterium tumefaciens* (custom database generated using SRS, Lion Biosciences). Sequences were also compared to the GenBank NR database using BLASTX. Top ranked BLAST hits involving other non-plant species and with E values  $< 1e^{-10}$  were classified as contaminants and removed prior to EST assembly.

### Selection of candidate FLcDNA clones and sequencing strategy

All 3'-end ESTs remaining after filtering were clustered and assembled using CAP3 [39] (assembly criteria: 95% identity, 40 bp window). The resulting contigs and singletons were defined as the PUT set. PUTs with a cDNA clone from a FLcDNA library were selected as candidates for complete insert sequencing (Figure 1). Candidate clones from FLcDNA libraries were single-pass sequenced from both 3'- and 5'-ends and both sequences were used for subsequent clone selection. Next, clones were screened for the presence of a polyA tail (3'-end EST) and the second-strand primer adaptor (SSPA; 5'-ACTAGTTTAATTAATTAATCCCCCCCCCCC-3'; 5'-end EST). Clones lacking either of these features were eliminated. A polyA tail was defined as at least 12 consecutive, or 14 of 15 "A" residues within the last 30 nt of the 3'-end EST (5' to 3'). The presence of the SSPA was detected using the Needleman-Wunsch algorithm limiting the search to the first 30 nt of the 5'-end EST (5' to 3'). The SSPA was defined as eight consecutive "C" residues and a  $> 80\%$  match to the remaining sequence (5'-ACTAGTTTAATTAATTAATTAAT-3'). In each case, the algorithms used to detect the 5' and 3' clone features were set to produce maximal sensitivity while maintaining a 0% false positive rate, as determined using test data sets. Candidate clones for which either of the initial 5'-end or 3'-end EST reads had a Phred20 quality length of  $< 100$  nt were also excluded. Finally, candidate clones were compared to poplar ESTs in the public domain (excluding ESTs from this collection; BLASTN match E  $< 1e^{-80}$ ) to identify candidate FLcDNAs potentially truncated at the 5' end of the transcript relative to a matching EST. Any clone with a 5' end that was  $> 100$  nt shorter than the matching public EST was excluded. For each PUT represented by multiple candidate clones after filtering, the clone with the longest 5' sequence was selected for complete insert sequencing. Insert sizing performed on 4,848 of 5,926 candidate clones using colony PCR with vector primers and standard gel electrophoresis revealed an average insert size of *ca.* 1,085 bp. Based on this information, a sequencing strategy emphasizing the use of end reads was chosen.

### Sequence finishing of FLCDNA clones

FLcDNA clones selected for complete sequence finishing were rearranged into 384-well plates, followed by an additional round of 5'-end and 3'-end sequencing using vector primers. All end reads from an individual clone were then assembled using PHRAP (version 0990329) [48-50]. To meet our sequence quality criteria, the resulting clone consensus sequence was required to achieve a minimum average score of Phred35, with each base position having a minimum score of Phred30. Each base position also required at least two sequence reads, of minimum Phred20, that were in agreement with the consensus sequence (i.e., no high-quality discrepancies). Clones that did not meet these finishing criteria after two rounds of end read sequencing were then subjected to successive rounds of sequencing using custom primers designed using the Consed graphical tool version 14 [54] until the required quality levels were achieved. Regardless of the finishing strategy, all clones that did not meet the minimum finishing criteria according to an automated pipeline were flagged for manual examination. Clones were aborted if they were manually verified to lack the minimum finishing criteria after three rounds of custom primer design, were identified as chimeric sequences, or were refractory to sequence finishing due to the presence of a "hard-stop". FLCDNA sequences have been deposited in the NR division of GenBank [[EF144175](#) to [EF148838](#)].

### Gene expression meta-analysis of FLCDNAs

Poplar FLCDNA sequences were mapped to a cDNA microarray containing 15,496 poplar ESTs [11]; Gene Expression Omnibus (GEO) platform number GPL5921] using BLASTN with a stringent threshold of  $\geq 95\%$  identity over  $\geq 95\%$  of alignment coverage. To identify FLCDNAs that were DE following FTC feeding, FLCDNAs mapping to the microarray were matched to an existing microarray dataset that examined gene expression in hybrid poplar leaves 24 hours after continuous FTC feeding [11]; GEO series number [GSE9522](#)).

### Authors' contributions

This study was conceived and directed by SGR, CJD and JB. Full-length cDNA libraries were developed by SGR, DC and NK. Data was analyzed by SGR, HJEC and RK with assistance from the coauthors. LG conducted DNA sequencing at the ORNL under the direction of GAT. RAH, SJM and MM directed sequencing and bioinformatics work at the GSC. SGR, HJEC and JB wrote the paper. All authors read and approved the final manuscript.

## Additional material

### Additional file 1

Full-length cDNA inventory. Predicted protein-coding features and annotation for the poplar full-length cDNA collection.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-57-S1.xls>]

### Additional file 2

Microarray dataset. Poplar FLCDNAs mapped to the genome-wide transcript profile of poplar leaves 24 h after the onset of forest tent caterpillar feeding using a 15.5 K array.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-57-S2.xls>]

### Additional file 3

Supplemental methods. Poplar methods for RNA isolation, full-length cDNA library construction, bacterial transformation with plasmids, clone handling, DNA purification and evaluation, and DNA sequencing are provided.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-9-57-S3.doc>]

## Acknowledgements

We thank Diana Palmquist, Brian Wynhoven, Jerry Liu, Yaron Butterfield and Asim Siddiqui of the Genome Sciences Centre for assistance with bioinformatic analyses; Jeff Stott, George Yang and many other staff at the Genome Sciences Centre for assistance with DNA sequencing; Claire Oddy and Sharon Jancsik of the University of British Columbia for assistance with clone insert sizing; Bob McCron from the Canadian Forest Service for access to forest tent caterpillars; and David Kaplan for greenhouse support. The work was supported by Genome British Columbia, Genome Canada and the Province of British Columbia (Treenomix Conifer Forest Health grant to J.B., and Treenomix grant to J.B. and C.J.D.), and by the Natural Science and Engineering Research Council of Canada (NSERC, grant to J.B.). Salary support for J.B. has been provided, in part, by the UBC Distinguished University Scholar Program and an NSERC Steacie Memorial Fellowship.

## References

1. Jansson S, Douglas CJ: **Populus: a model system for plant biology**. *Annu Rev Plant Biol* 2007, **58**:435-458.
2. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroove S, Déjardin A, dePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehling J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P,



- Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)**. *Science* 2006, **313**:1596-1604.
3. Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA: **Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map**. *Theor Appl Genet* 2004, **109**:451-463.
  4. Zhang D, Zhang Z, Yang K, Li B: **Genetic mapping in (*Populus tomentosa* × *Populus bolleana*) and *P. tomentosa* Carr. using AFLP markers**. *Theor Appl Genet* 2004, **108**:657-662.
  5. Cervera MT, Storme V, Soto A, Ivens B, Van Montagu M, Rajora OP, Boerjan W: **Intraspecific and interspecific genetic and phylogenetic relationships in the genus *Populus* based on AFLP markers**. *Theor Appl Genet* 2005, **111**:1440-1456.
  6. Woolbright SA, DiFazio SP, Yin T, Martinsen GD, Zhang X, Allan GJ, Whitham TG, Keim P: **A dense linkage map of hybrid cottonwood (*Populus fremontii* × *P. angustifolia*) contributes to long-term ecological research and comparison mapping in a model forest tree**. *Heredity* 2008, **100**:59-70.
  7. Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin T, DiFazio SP, Ali J, Asano JK, Chan S, Cloutier A, Girn N, Leach S, Lee D, Mathewson CA, Olson T, O'Connor K, Prabhur AL, Smailus DE, Stott JM, Tsai M, Wye NH, Yang GS, Zhuang J, Holt RA, Putnam NH, Vrebalov J, Giovannoni JJ, Grimwood J, Schmutz J, Rokhsar D, Jones SJM, Marra MA, Tuskan GA, Bohlmann J, Ellis BE, Ritland K, Douglas CJ, Schein JE: **A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation**. *Plant J* 2007, **50**:1063-1078.
  8. Andersson A, Keskitalo J, Sjödin A, Bhalerao R, Sterky F, Wissel K, Tandre K, Aspeborg H, Moyle R, Ohmiya Y, Bhalerao R, Brunner A, Gustafsson P, Karlsson J, Lundeberg J, Nilsson O, Sandberg G, Strauss S, Sundberg B, Uhlen M, Jansson S, Nilsson P: **A transcriptional timetable of autumn senescence**. *Genome Biol* 2004, **5**:R24.1-R24.13.
  9. Brosché M, Vinocur B, Alatalo ER, Lamminmäki A, Teichmann T, Ottow EA, Djilianov D, Afif D, Bogeat-Triboulot MB, Altman A, Polle A, Dreyer E, Rudd S, Paulin L, Auvinen P, Kangasjärvi J: **Gene expression and metabolite profiling of *Populus euphratica* growing in the Negev desert**. *Genome Biol* 2005, **6**:R101.1-R101.17.
  10. Harding SA, Jiang H, Jeong ML, Casado FL, Lin HW, Tsai CJ: **Functional genomics analysis of foliar condensed tannin and phenolic glycoside regulation in natural cottonwood hybrids**. *Tree Physiol* 2005, **25**:1475-1486.
  11. Ralph S, Oddy C, Cooper D, Yueh H, Jancsik S, Kolosova N, Philippe RN, Aeschliman D, White R, Huber D, Ritland CE, Benoit F, Rigby T, Nantel A, Butterfield YSN, Kirkpatrick R, Chun E, Liu J, Palmquist D, Wynnoven B, Stott J, Yang G, Barber S, Holt RA, Siddiqui A, Jones SJM, Marra MA, Ellis BE, Douglas CJ, Ritland K, Bohlmann J: **Genomics of hybrid poplar (*Populus trichocarpa* × *deltoides*) interacting with forest tent caterpillars (*Malacosoma disstria*): Normalized and full-length cDNA libraries, expressed sequence tags, and a cDNA microarray for the study of insect-induced defences in poplar**. *Mol Ecol* 2006, **15**:1275-1297.
  12. Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, Villarroel R, Van Montagu M, Sandberg G, Olsson O, Teeri TT, Boerjan W, Gustafsson P, Uhlén M, Sundberg B, Lundeberg J: **Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags**. *Proc Natl Acad Sci USA* 1998, **95**:13330-13335.
  13. Bhalerao R, Keskitalo J, Sterky F, Erlandsson R, Björkbacka H, Birve SJ, Karlsson J, Gardeström P, Gustafsson P, Lundeberg J, Jansson S: **Gene expression in autumn leaves**. *Plant Physiol* 2003, **131**:430-442.
  14. Kohler A, Delaruelle C, Martin D, Encelot N, Martin F: **The poplar root transcriptome: analysis of 7000 expressed sequence tags**. *FEBS Lett* 2003, **542**:37-41.
  15. Ranjan P, Kao YY, Jiang H, Joshi CP, Harding SA, Tsai CJ: **Suppression subtractive hybridization-mediated transcriptome analysis from multiple tissues of aspen (*Populus tremuloides*) altered in phenylpropanoid metabolism**. *Planta* 2004, **219**:694-704.
  16. Schrader J, Moyle R, Bhalerao R, Hertzberg M, Lundeberg J, Nilsson P, Bhalerao RP: **Cambial meristem dormancy in trees involves extensive remodelling of the transcriptome**. *Plant J* 2004, **40**:173-187.
  17. Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, Sundberg B, Gustafsson P, Uhlén M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S: **A *Populus* EST resource for plant functional genomics**. *Proc Natl Acad Sci USA* 2004, **101**:13951-13956.
  18. Christopher ME, Miranda M, Major IT, Constabel CP: **Gene expression profiling of systemically wound-induced defenses in hybrid poplar**. *Planta* 2004, **219**:936-947.
  19. Nanjo T, Futamura N, Nishiguchi M, Igasaki T, Shinozaki K, Shinohara K: **Characterization of full-length enriched sequence tags of stress-treated poplar leaves**. *Plant Cell Physiol* 2004, **45**:1738-1748.
  20. Rishi AS, Munir S, Kapur V, Nelson ND, Goyal A: **Identification and analysis of safener-inducible expressed sequence tags in *Populus* using a cDNA microarray**. *Planta* 2004, **220**:296-306.
  21. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O: **Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies**. *Nucleic Acids Res* 2003, **31**:5654-5666.
  22. Castelli V, Aury JM, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quéfier F, Scarpelli C, Schächter V, Temple G, Caboche M, Weissenbach J, Salanoubat M: **Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation**. *Genome Res* 2004, **14**:406-413.
  23. Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA: **Features of *Arabidopsis* genes and genome discovered using full-length cDNAs**. *Plant Mol Biol* 2006, **60**:69-85.
  24. Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu R, Ueda M, Matsubara K, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashidume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y: **Collection, mapping and annotation of over 28,000 cDNA clones from japonica rice**. *Science* 2003, **301**:376-379.
  25. Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K: **Functional annotation of a full-length *Arabidopsis* cDNA collection**. *Science* 2002, **296**:141-145.
  26. Lai J, Dey N, Kim CS, Bharti AK, Rudd S, Mayer KFX, Larkins BA, Becraft P, Messing J: **Characterization of the maize endosperm transcriptome and its comparison to the rice genome**. *Genome Res* 2004, **14**:1932-1937.
  27. Jia J, Fu J, Zheng J, Zhou X, Huai J, Wang J, Wang M, Zhang Y, Chen X, Zhang J, Zhao J, Su Z, Lv Y, Wang G: **Annotation and expression profile analysis of 2073 full-length cDNAs from stress-induced maize (*Zea mays* L.) seedlings**. *Plant J* 2006, **48**:710-727.
  28. Fitzgerald TD: *The Tent Caterpillars* Ithaca, New York: Cornell University Press; 1995.
  29. Philippe RN, Bohlmann J: **Poplar defense against insect herbivores**. *Canadian Journal of Botany* 2007, **85**:1111-1126.
  30. Constabel CP, Yip L, Patton JJ, Christopher ME: **Polyphenol oxidase from hybrid poplar. Cloning and expression in response to wounding and herbivory**. *Plant Physiol* 2000, **124**:285-295.
  31. Haruta M, Major IT, Christopher ME, Patton JJ, Constabel CP: **A Kunitz trypsin inhibitor gene family from trembling aspen (*Populus tremuloides* Michx.): cloning, functional expression, and induction by wounding and herbivory**. *Plant Mol Biol* 2001, **46**:347-359.
  32. Peters DJ, Constabel CP: **Molecular analysis of herbivore-induced condensed tannin synthesis: cloning and expression**

- of dihydroflavonol reductase from trembling aspen (*Populus tremuloides*). *Plant J* 2002, **32**:701-712.
33. Arimura G, Huber DPW, Bohlmann J: **Forest tent caterpillars (*Malacosoma disstria*) induce local and systemic diurnal emissions of terpenoid volatiles in hybrid poplar (*Populus trichocarpa* × *deltoides*): cDNA cloning, functional characterization, and patterns of gene expression of (-)-germacrene D synthase *PtdTPS1*.** *Plant J* 2004, **37**:603-616.
  34. Wang J, Constabel CP: **Polyphenol oxidase overexpression in transgenic *Populus* enhances resistance to herbivory by forest tent caterpillar (*Malacosoma disstria*).** *Planta* 2004, **220**:87-96.
  35. Lawrence SD, Dervinis C, Novak N, Davis JM: **Wound and insect herbivory responsive genes in poplar.** *Biotechnol Lett* 2006, **28**:1493-1501.
  36. Major IT, Constabel CP: **Molecular analysis of poplar defense against herbivory: comparison of wound- and insect elicitor-induced gene expression.** *New Phytol* 2006, **172**:617-635.
  37. Miranda M, Ralph SG, Mellway R, White R, Heath MC, Bohlmann J, Constabel CP: **The transcriptional response of hybrid poplar (*Populus trichocarpa* × *P. deltoides*) to infection by *Melampsora medusae* leaf rust involves induction of flavonoid pathway genes leading to the accumulation of proanthocyanidins.** *Mol Plant-Microbe Interac* 2007, **20**:816-831.
  38. Carninci P, Kvan C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, Schneider C: **High-efficiency full-length cDNA cloning by biotinylated CAP trapper.** *Genomics* 1996, **37**:327-336.
  39. Huang X, Madan A: **CAP3: a DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
  40. Seki M, Carninci P, Nishiyama Y, Hayashizaki Y, Shinozaki K: **High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper.** *Plant J* 1998, **15**:707-720.
  41. Kent WJ: **BLAT-the BLAST-like alignment tool.** *Genome Res* 2002, **12**:656-664.
  42. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonhammer ELL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**:D247-D251.
  43. Sterck L, Rombauts S, Jansson S, Sterky F, Rouzé P, Van de Peer Y: **EST data suggest that poplar is an ancient polyploid.** *New Phytol* 2005, **167**:165-170.
  44. Hughes AL: **The Evolution of Functionally Novel Proteins after Gene Duplication.** *Proc R Soc Lond B* 1994, **256**:119-124.
  45. Ku HM, Vision T, Liu J, Tanksley SD: **Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny.** *Proc Natl Acad Sci USA* 2000, **97**:9121-9126.
  46. Carninci P, Hayashizaki Y: **High-efficiency full-length cDNA cloning.** *Methods Enzymol* 1999, **303**:19-44.
  47. Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y: **Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes.** *Genome Res* 2000, **10**:1617-1630.
  48. Ewing B, Green P: **Base-calling of automated sequencer traces using phred II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
  49. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.
  50. **Laboratory of Dr. Phil Green: software resources** [<http://www.phrap.org>]
  51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
  52. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
  53. **FTP directory yeast genome** [<ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/yeast.nt.gz>]
  54. Gordon D, Abajian C, Green P: **Consed: A graphical tool for sequence finishing.** *Genome Res* 1998, **8**:195-202.
  55. **EMBOSS** [<http://emboss.sourceforge.net/>]
  56. **The Arabidopsis Information Resource** [<http://www.arabidopsis.org>]
  57. **FTP directory GenBank** [<ftp://ftp.ncbi.nih.gov/blast/db/>]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

