## Methods

### Sequence Processing

A total of 7232 chromatogram files from two sets of replicate samples, two each for needle and cambium, were processed using an IPython (Pérez and Granger 2007) notebook and various utilities. The chromatograms were converted using Phred (Ewing and Green 1998) into FASTA sequences with a trimming cutoff probability of 0.01 (i.e., Phred score = 20), and sequences which were less than 100 bases long were excluded. These ESTs, from all replicates, were combined into a single file (where all of the sequence identifiers indicated the source and replicate of the EST). The combined file was processed, first using SeqClean (Chen et al. 2007), screening for both vectors using the Univec (Cochrane and Galperin 2010) database and contamination from *E. coli* K-12 substr. DH10B. The cleaned and trimmed ESTs were, in each case, assembled into unigenes using iAssembler (Zheng et al. 2011). The representation of the number of sequence fragments from each tissue and replicate were calculated from the output from iAssembler (Zheng et al. 2011), which tracks which individual FASTA sequences (decorated with the source tissue/replicate combination) were assembled into each unigenes. These counts were used to assess differential unigene expression.

The assembled unigenes were aligned to a custom local copy of the UniProtKB database. This database was created from the most current release of the Swiss-Prot as of 4/1/15. Briefly, the accessions from the plant taxonomic division were extracted and were used to extract FASTA sequences from the entire UniProt database. The top 10 blastx (Altschul et al. 1990) alignments for each unigene, having e-values < 1e-5, were retained for annotation using Blast2GO (Götz et al. 2008).

### Differential Gene Expression

To determine whether assembled unigenes were differentially expressed (DE) in either needle or cambium, counts across replicates were combined for each tissue. These cambium and needle counts were used as input to the online version of IDEG6 (Romualdi et al. 2003) at http://telethon.bio.unipd.it/bioinfo/IDEG6/. All test statistics were evaluated, but determination of differential expression was based on FDR-corrected p-values (Benjamini and Hochberg 1995) calculated from the general χ2 p-values from IDEG6. We chose to rely on the FDR (q < 0.05) method of correction rather than Bonferonni due the overly conservative nature of that method. We did not employ the false discovery rate correction of Storey and Tibshirani (2003) because the distribution of p-values from our data did not meet the asymptotic assumptions of that method.

### Gene Ontology

Blastx results were imported into Blast2GO using default parameters, keeping at most 10 hits with a high-scoring pair (HSP) length of at least 33 amino acids. Annotations were assigned using the default parameters (e-value: 1e-6, annotation cutoff: 55, GO weight: 5, HSP-hit coverage cutoff: 0, taxonomy filtering: none) and evidence code weights. The full suite of InterPro 5 (Jones et al. 2014) mappings were performed for each unigene and these results were merged with existing annotations. Unigene annotations were also augmented with ANNEX and mapped to enzyme code and KEGG pathways (Kanehisa and Goto 2000). The resulting annotations were exported to a text file for further analysis.

Differential GO term analysis was computed for needle and cambium using the topGO R package (Alexa and Rahnenführer 2010) which provides tests for GO terms while accounting for the GO graph topology. The GO hierarchy was trimmed to only include terms with two or more annotated unigenes (i.e., nodesize = 2). Each set of unigenes, which were deemed to be differentially expressed in either needle or cambium, was tested using a one-tailed Fisher’s exact test and assessed for significance at two different levels, either having an uncorrected p-value less than 0.05 or a corrected (Benjamini and Hochberg 1995) p-value < 0.05. In all, six tests for differentially expressed GO terms were conducted with each tissue type evaluated against each of three GO ontologies: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). Multiple test correction for GO term significance is often performed, and we include these results. However, due to the nested structure and non-independence of terms in the GO hierarchy, as well as the methodology employed in topGO, we question whether multiple test correction is appropriate in our case, so we only note where terms pass multiple test correction, but consider terms with raw p-values < 0.05 as significant.

All relevant analysis code can be found in IPython notebooks at http://www.github.com/cfriedline/black\_spruce. Raw data files can be accessed from the iPlant Data Store at https://de.iplantcollaborative.org/de/?type=data&folder=/iplant/home/cfriedline/pub\_data/black\_spruce, and the unigenes can be found in NCBI dbEST under accessions XXXX—XXXX.

## Results

### Sequence data

A total of 7232 chromatogram files were obtained from sequencing for two biological replicates

(e.g., P32, P40) for two tissue types: needle and cambium. Following processing with Phred and

filtering by length, 5996 raw ESTs were combined into a single file for downstream analysis. After processing with SeqClean, 5938 ESTs remained; 2842 were trimmed and 58 were removed from the dataset, either by mapping to E. coli (34), low complexity (1), or length/shortq (23). Assembly of the fragments resulted in 1945 unigenes with an average coverage of ~ ESTs per unigenes (range=[1, 274] ESTs/unigene). Descriptive information about the raw and assembled ESTs can be found in Table 1.



Table : EST statistics for needle and cambium ESTs. The number of singletons relates to the count from each sample that is the only EST present in an assembled unigene. Length distribution refers to the ESTs that were used to assemble unigenes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample** | **Tissue** | **Raw ESTs** | **Assembled ESTs** | **Singletons** | **Mean length/sd (min, max)** |
| P32 | Cambium | 1829 | 1526 | 189 | 435/157 (100, 850) |
| P40 | Cambium | 1628 | 1677 | 246 | 510/139 (101, 851) |
| P32 | Needle | 1926 | 1475 | 525 | 507/168 (102, 850) |
| P40 | Needle | 1849 | 1260 | 142 | 425/147 (100, 772) |

We find no significant differences between combined cambium and needle read lengths (Welch’s t = 1.16, df = 5633.37, p-value=0.25), which is important because length bias could influence annotation results with higher proportions of longer genes being associated with higher numbers of GO terms (Mi et al. 2012). Additionally the first two PCA axes of the counts for each sample accounts for a significant proportion of variance (69%) in the data, as shown in Figure 1. From the PCA, several observations can be made. PC1 accounts for the variance in EST counts/unigene largely in needle but also largely in unigenes that are found in both tissues. Accordingly, PC2 tracks the variance in cambium-specific unigene expression. There is also a strong relationship between replicates in both tissue types, lending credence to their biological replicability and ability to be combined for downstream analyses. Finally, the expression of unigenes in cambium appears to be nearly perfectly orthogonal to the expression of unigenes in the needles, highlighting the underlying biological differences between tissue types.Macintosh HD:Users:chris:Dropbox:Documents:science:postdoc:papers:black_spruce:manuscript:count_pca.pdf

Figure : PCA of unigene counts, colored by presence (count > 0) of a unigene in a tissue type

### Gene expression

The data in this study are largely organized into two main types: unigenes and ontology. Unigenes are derived from the assembly of quality-controlled and filtered ESTs. The count of individual ESTs that make up each unigene are taken directly as measures of expression of that unigene in the tissue under study, after combining replicates across tissue types. For example, unigene UN0003 was assembled from 1 EST in P32C and 1 EST in P40C, and was not found in either needle tissue sample; its expression, therefore, would be 2 in cambium and 0 in needle. For simplicity, the terms unigene and gene in this study, can be considered to be synonyms. Unigenes are also assigned to genes in UniProtKB using blastx, and these assignments are used to derive ontological terms from the GO database. As such, each unigene may be assigned to multiple GO terms, which can exist in any layer of the GO hierarchy, some of which may be nested into increasingly more general terms. This tree-like structure, therefore, creates a degree of non-independence between terms which must be accounted for in methods determining significant enrichment.

The 1945 unigenes were tested for differential expression (DE) using IDEG6. Expression levels across tissue types were largely similar, with 3203 ESTs in cambium and 2735 in needle. Overall, we found 221 significantly expressed unigenes uncorrected p-values < 0.05 and 57 with FDR-corrected (q-value) general χ2 p-values < 0.05.

When considering only those unigenes that are associated with one or more GO terms (n = 976 unigenes), this number is reduced to 112 significant DE unigenes assessed using uncorrected p-values, and 29 with q < 0.05. Of these 29, 12 are found in cambium and 17 are found in needle tissue. Blastx statistics for the DE unigenes can be found in Table 2 and Table 3. Across all ontologies, the cambium DE unigenes had 4 ± 8 GO terms, ranging from 1—30 terms per unigene; the needle differentially expressed unigenes had 6 ± 3 GO terms, ranging from 1—15 terms per unigene. More details on the GO terms assigned to needle and cambium DE unigenes can be found in Supplemental Table 1 and Supplemental Table 2.

Table : Blastx results against UniProtKB for 12 DE unigenes from cambium tissue. Length describes the length of the unigene, e-value the minimum e-value across all hits and similarity is the Mean sum. is the mean similarity between the unigene and its annotated sequence.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Unigene** | **UniProtKB Description** | **Length** | **E-Value** | **Mean sim.** |
| UN0192 | mt3\_picglmetallothionein-like protein | 585 | 2.10E-34 | 60.6 |
| UN0193 | mao2\_arathnad-dependent malic enzyme mitochondrial | 1257 | 1.20E-36 | 76.6 |
| UN0214 | phytochrome partial | 549 | 5.90E-20 | 81.8 |
| UN0215 | trxh\_picmathioredoxin h- | 716 | 5.40E-63 | 78.5 |
| UN0228 | nltp3\_horvunon-specific lipid-transfer protein 3 | 757 | 5.20E-25 | 62.6 |
| UN0283 | phytochrome partial | 610 | 3.00E+00 | 50 |
| UN0286 | mt3\_picglmetallothionein-like protein emb30 | 621 | 1.20E-26 | 61.6 |
| UN0297 | mt3\_picglmetallothionein-like protein emb30 | 709 | 2.30E-26 | 61.6 |
| UN0354 | mt3\_picglmetallothionein-like protein emb30 | 776 | 9.10E-36 | 59.89 |
| UN0357 | mt3\_picglmetallothionein-like protein emb30 | 641 | 8.30E-30 | 74 |
| UN0359 | mt3\_picglmetallothionein-like protein emb30 | 698 | 5.70E-26 | 62 |
| UN0362 | chi5\_orysjchitinase 5 | 647 | 9.40E-45 | 70.9 |

Table : Blastx results against UniProtKB for 17 DE unigenes from needle tissue. Length describes the length of the unigene, e-value the minimum e-value across all hits and similarity is the Mean sim. is the mean similarity between the unigene and its annotated sequence.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Unigene** | **UniProtKB Description** | **Length** | **E-value** | **Mean sim.** |
| UN0073 | act11\_arathactin-11 | 2077 | 0.00E+00 | 96.6 |
| UN0164 | rca\_orysjribulose bisphosphate carboxylase oxygenase chloroplastic | 573 | 1.20E-76 | 94.3 |
| UN0167 | pip21\_arathaquaporin | 647 | 5.30E-55 | 75.6 |
| UN0177 | rca\_arathribulose bisphosphate carboxylase oxygenase chloroplastic | 684 | 1.80E-24 | 85.4 |
| UN0195 | rbs\_larlaribulose bisphosphate carboxylase small chloroplastic | 606 | 1.00E-40 | 88 |
| UN0205 | rbs\_larlaribulose bisphosphate carboxylase small chloroplastic | 724 | 3.80E-41 | 85.4 |
| UN0240 | alf1\_peafructose-bisphosphate cytoplasmic isozyme 1 | 747 | 1.80E-62 | 76.8 |
| UN0243 | rbs\_larlaribulose bisphosphate carboxylase small chloroplastic | 616 | 1.00E-39 | 89.4 |
| UN0285 | psab\_phypaphotosystem i p700 chlorophyll a apoprotein a2 | 778 | 9.30E-115 | 99 |
| UN0323 | ml423\_arathmlp-like protein 423 | 587 | 5.10E-13 | 53 |
| UN0326 | rbs\_larlaribulose bisphosphate carboxylase small chloroplastic | 595 | 1.80E-40 | 87.7 |
| UN0338 | nltp2\_lencunon-specific lipid-transfer protein 2 | 749 | 3.50E-31 | 66.5 |
| UN0339 | rbs\_pinthribulose bisphosphate carboxylase small chloroplastic | 517 | 1.10E-39 | 71.7 |
| UN0340 | rca\_maldoribulose bisphosphate carboxylase oxygenase chloroplastic | 612 | 3.70E-91 | 91.7 |
| UN0345 | cb5\_arathchlorophyll a-b binding protein chloroplastic | 720 | 6.70E-58 | 79.5 |
| UN0360 | rbs\_larlaribulose bisphosphate carboxylase small chloroplastic | 684 | 9.70E-40 | 87.2 |
| UN0361 | rbs\_larlaribulose bisphosphate carboxylase small chloroplastic | 665 | 4.60E-40 | 87.9 |

### Gene ontology

As input to topGO, we considered the 12 cambium and 17 needle DE unigenes in order to test for GO term enrichment among the two tissue types. These data are summarized in

Table 4. For all ontologies, we find significant enrichment of GO terms for each ontology for each tissue, however, in cambium, none of these terms pass multiple test correction. As mentioned in the methods section, multiple test correction is not appropriate in this case because the features of the graph are not tested independently. Consequently, we consider terms with uncorrected p-values < 0.05 as significantly enriched.

Table 4: Counts of GO terms by tissue for each ontology for the set of 976 unigenes associated with at least one GO term. The first number indicates the number of significant unigenes and the second indicates the number of total possible unigenes testable for that ontology defined as a GO term associated with at least two unigenes. The third and fourth numbers indicate the number of significantly enriched GO terms with p < 0.05 and q < 0.05, respectively.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Ontology** | | |
| **Tissue** | **BP** | **CC** | **MF** |
| Cambium | 6/771/8/0 | 1/549/2/0 | 10/846/6/0 |
| Needle | 14/771/18/11 | 13/549/14/9 | 13/846/23/5 |

GO graphs were generated for four different sets of data (two each for all unigenes annotated with a GO term, and two each for the DE unigenes), keeping only nodes with at least 5 unigenes and a node score of 5 (alpha = 0.6) except for Biological Process ontology for cambium DE unigenes (unigenes >= 2, node score >= 0), these filters were relaxed in order to obtain data in this case. We did not find any Cellular Component annotations for cambium DE unigenes.

The important GO terms for both cambium and needle are shown in Figure 2, Supplemental Figure 1, and Supplemental Figure 2. We have chosen to show the important terms in two ways, first all significant terms are shown in the chart. The highest number of significant terms was found in the needle MF set (n = 23), so GO information for all tissues and ontologies was shown in this way for consistency. Therefore, the difference in number of terms between 23 and the number of significant terms for each tissue-ontology combination is made up the terms assigned to the most number of sequences, in decreasing order.

Overall, we find terms which are biologically consistent with known functions of each tissue type. In the needle, we see familiar terms associated with photosynthesis, chloroplasts, and metabolism while in cambium we see terms associated with metabolism and biosynthesis. There are several cases of term enrichment associated with each tissue type as well: photosynthesis-related terms in needle again, as well as ion-binding and stress response terms in cambium.

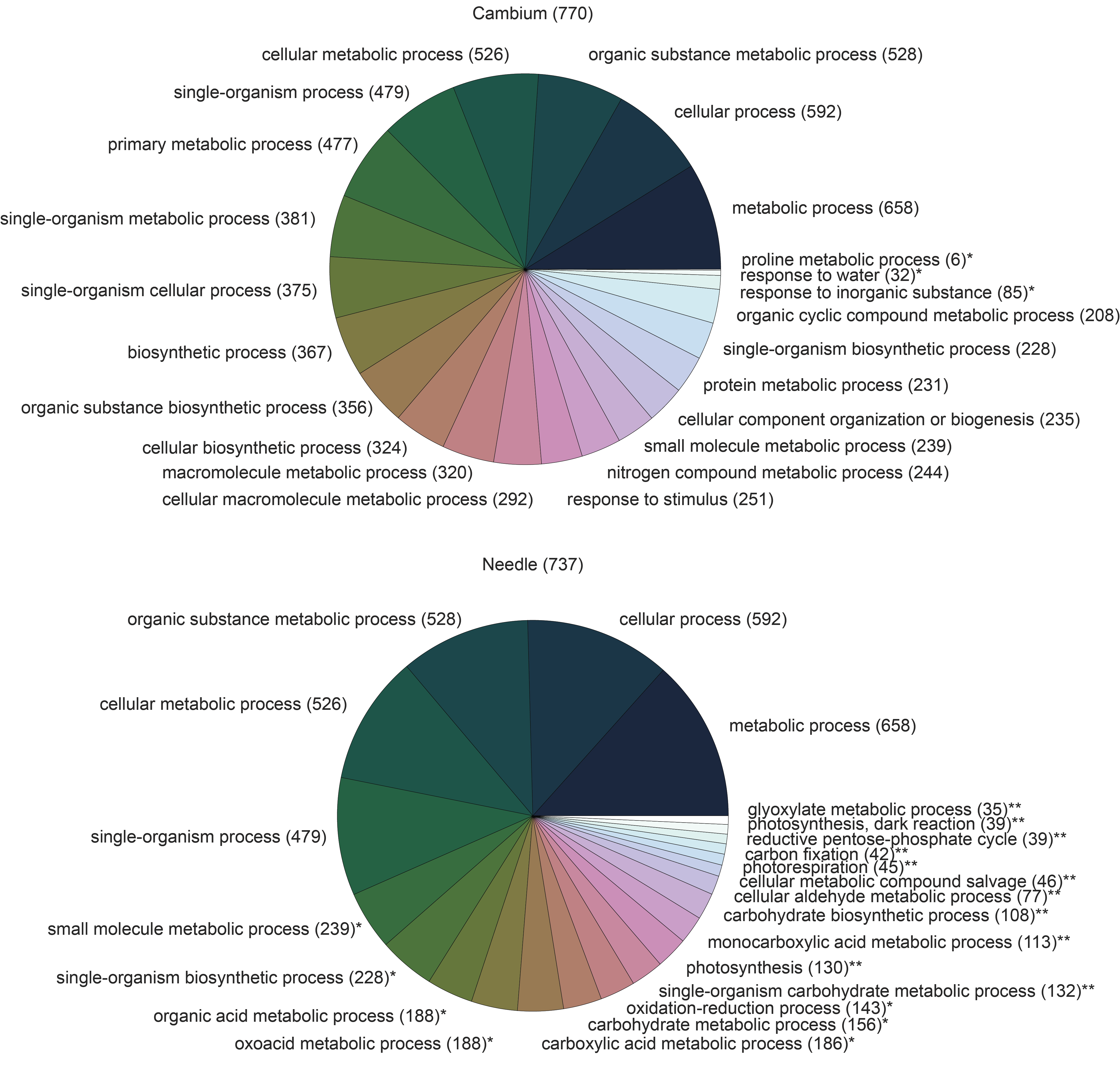


Figure : GO terms describing enrichment from the BP ontology. All significant terms are included, p < 0.05 (\*) and q < 0.05 (\*\*). The number in parentheses at the top of each chart represents the number of unique unigenes represented in the data, while the number at each term is the number of unigenes assigned to that term.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

## References

Alexa, A., and J. Rahnenführer. 2010. topGO: topGO: Enrichment analysis for Gene Ontology.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57:289–300.

Chen, Y.-A., C.-C. Lin, C.-D. Wang, H.-B. Wu, and P.-I. Hwang. 2007. An optimized procedure greatly improves EST vector contamination removal. BMC Genomics 8:416–416.

Cochrane, G. R., and M. Y. Galperin. 2010. The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. Nucleic Acids Research 38:D1–D4.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186–194.

Götz, S., J. M. García-Gómez, J. Terol, T. D. Williams, S. H. Nagaraj, M. J. Nueda, M. Robles, M. Talón, J. Dopazo, and A. Conesa. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36:3420–3435. Oxford University Press.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter. 2014. InterProScan 5: genome-scale protein function classification. J Gerontol 30:1236–1240.

Kanehisa, M., and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28:27–30. Oxford University Press.

Mi, G., Y. Di, S. Emerson, J. S. Cumbie, and J. H. Chang. 2012. Length bias correction in gene ontology enrichment analysis using logistic regression. PLoS ONE 7:e46128–e46128.

Pérez, F., and B. E. Granger. 2007. IPython: A System for Interactive Scientific Computing. Comput. Sci. Eng. 9:21–29. IEEE.

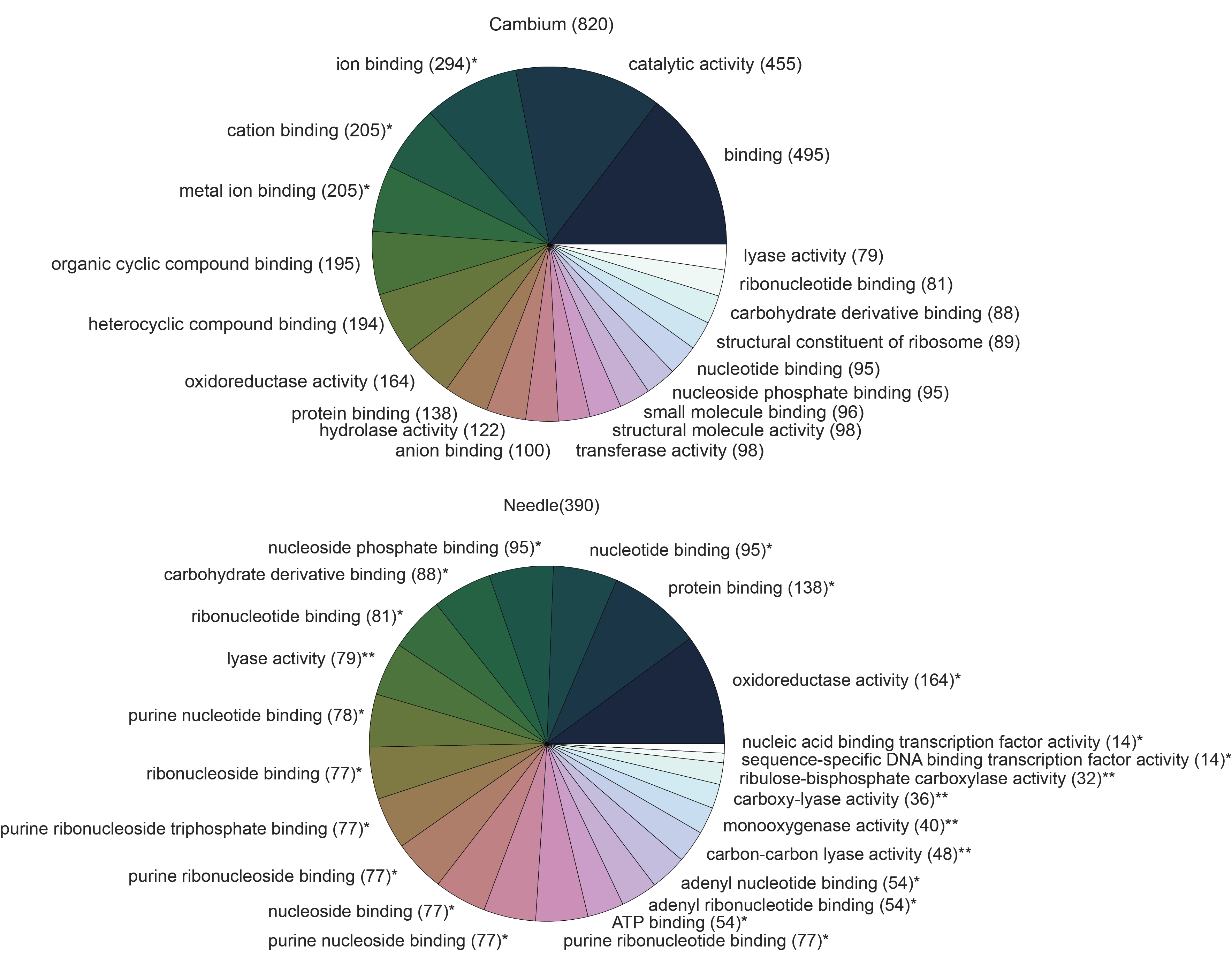
Romualdi, C., S. Bortoluzzi, and F. d'Alessi. 2003. IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. Physiological ….

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100:9440–9445.

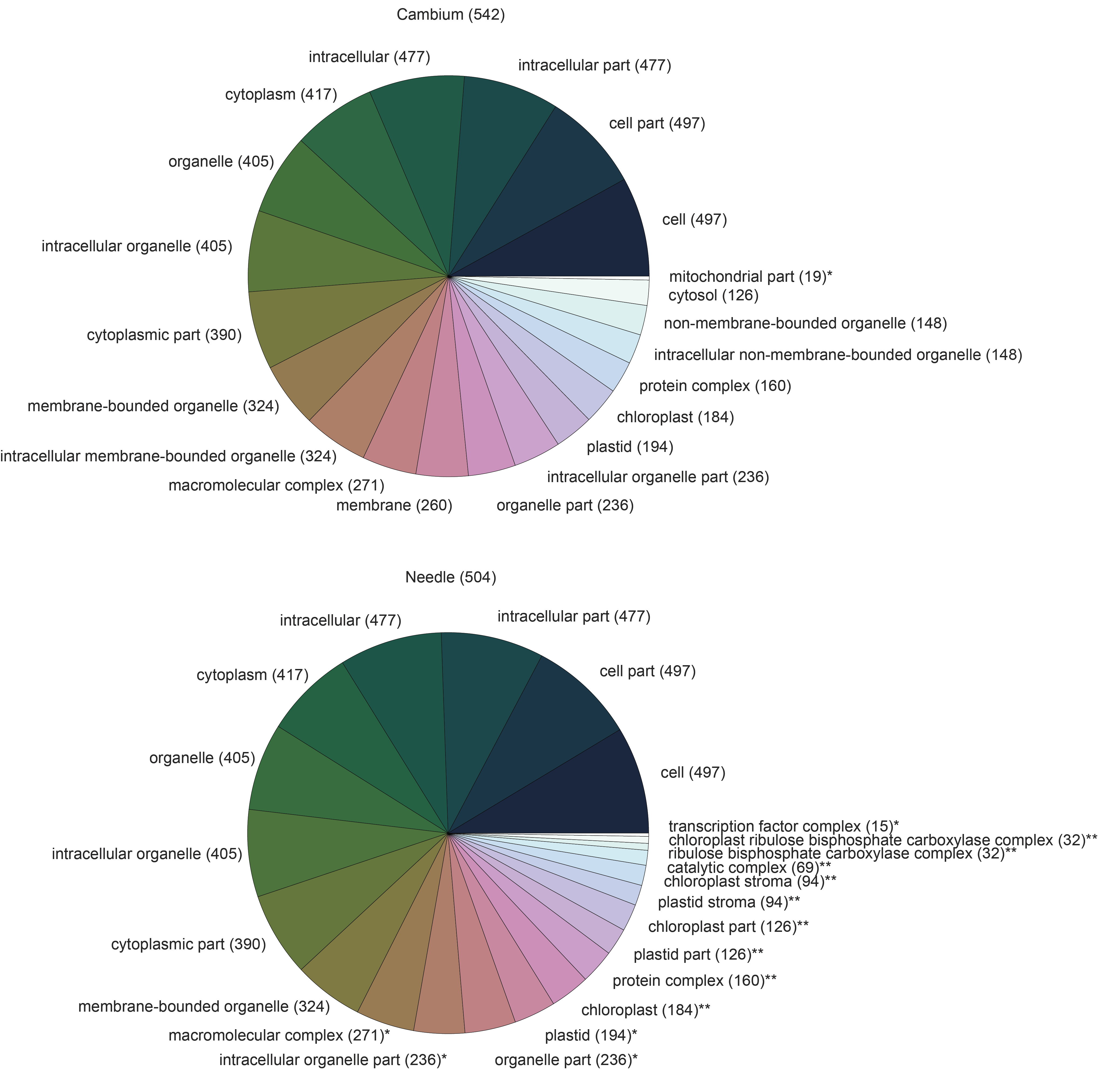
Zheng, Y., L. Zhao, J. Gao, and Z. Fei. 2011. iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences. BMC Bioinformatics 12:453–453. BioMed Central Ltd.

## Supplemental Information

### Figures



Supplemental Figure : GO terms describing enrichment from the MF ontology. All significant terms are included, p < 0.05 (\*) and q < 0.05 (\*\*). The number in parentheses at the top of each chart represents the number of unique unigenes represented in the data, while the number at each term is the number of unigenes assigned to that term.



Supplemental Figure : GO terms describing enrichment from the CC ontology. All significant terms are included, p < 0.05 (\*) and q < 0.05 (\*\*). The number in parentheses at the top of each chart represents the number of unique unigenes represented in the data, while the number at each term is the number of unigenes assigned to that term.

### Tables

Supplemental Table : GO terms associated with DE cambium genes with counts in each tissue.

|  |  |  |  |
| --- | --- | --- | --- |
| **Unigene** | **GO Terms** | **Cambium** | **Needle** |
| UN0192 | GO:0046872 | 30 | 0 |
| UN0193 | GO:0005759,GO:0005829,GO:0009507,GO:0004471,GO:0004473,GO:0005524,GO:0008270,GO:0008948,GO:0016652,GO:0042803,GO:0050897,GO:0051287,GO:0006094,GO:0006096,GO:0006108,GO:0006833,GO:0006972,GO:0007010,GO:0007030,GO:0009266,  GO:0009651,GO:0009827,GO:0009860,GO:0010498,GO:0046686,GO:0051260,GO:0006099,GO:0015976,GO:0006525,GO:0006560 | 21 | 0 |
| UN0214 | GO:0006950,GO:0009415 | 32 | 0 |
| UN0215 | GO:0015035,GO:0006662,GO:0045454,GO:0006118 | 20 | 0 |
| UN0228 | GO:0008289,GO:0006869 | 14 | 0 |
| UN0283 | GO:0006950,GO:0009415 | 38 | 0 |
| UN0286 | GO:0046872 | 16 | 0 |
| UN0297 | GO:0046872 | 18 | 0 |
| UN0354 | GO:0046872 | 51 | 0 |
| UN0357 | GO:0046872 | 37 | 0 |
| UN0359 | GO:0046872 | 12 | 0 |
| UN0362 | GO:0004568,GO:0006032,GO:0016998 | 19 | 0 |

Supplemental Table : GO terms associated with DE needle genes

|  |  |  |  |
| --- | --- | --- | --- |
| **Unigene** | **GO Terms** | **Cambium** | **Needle** |
| UN0073 | GO:0005618,GO:0005634,GO:0005739,GO:0005829,GO:0005856,GO:0005886,GO:0009506,GO:0009570,GO:0009941,GO:0048046,GO:0005200,GO:0005524,GO:0006094,GO:0010498,GO:0030036 | 0 | 24 |
| UN0164 | GO:0005524 | 0 | 10 |
| UN0167 | GO:0016020,GO:0005215,GO:0006810 | 0 | 10 |
| UN0177 | GO:0009570,GO:0003700,GO:0005524,GO:0005667,GO:0045449 | 1 | 14 |
| UN0195 | GO:0004497,GO:0005515,GO:0016984,GO:0009853,GO:0019253,GO:0055114,GO:0009573,GO:0046487 | 0 | 17 |
| UN0205 | GO:0004497,GO:0016984,GO:0009853,GO:0019253,GO:0055114,GO:0009573,GO:0046487 | 0 | 21 |
| UN0240 | GO:0004332,GO:0006096,GO:0006000,GO:0006013,GO:0006020,GO:0006094,GO:0006098,GO:0015976 | 0 | 9 |
| UN0243 | GO:0004497,GO:0005515,GO:0016984,GO:0009853,GO:0019253,GO:0055114,GO:0009573,GO:0046487 | 4 | 82 |
| UN0285 | GO:0009522,GO:0016021,GO:0015979 | 2 | 22 |
| UN0323 | GO:0006952,GO:0009607 | 0 | 25 |
| UN0326 | GO:0004497,GO:0005515,GO:0016984,GO:0009853,GO:0019253,GO:0055114,GO:0009573,GO:0046487 | 6 | 35 |
| UN0338 | GO:0006869,GO:0009414,GO:0009651,GO:0009737 | 4 | 36 |
| UN0339 | GO:0004497,GO:0016984,GO:0009853,GO:0019253,GO:0055114,GO:0009573,GO:0046487 | 0 | 9 |
| UN0340 | GO:0009570,GO:0003700,GO:0005524,GO:0005667,GO:0045449 | 2 | 15 |
| UN0345 | GO:0016020,GO:0009765 | 1 | 24 |
| UN0360 | GO:0004497,GO:0005515,GO:0016984,GO:0009853,GO:0019253,GO:0055114,GO:0009573,GO:0046487 | 1 | 54 |
| UN0361 | GO:0004497,GO:0005515,GO:0016984,GO:0009853,GO:0019253,GO:0055114,GO:0009573,GO:0046487 | 3 | 140 |