INVESTIGATIONS

Sex determination in *Ceratopteris richardii* is accompanied by transcriptome changes that drive epigenetic reprogramming of the young gametophyte: supplementary material

Nadia M. Atallah[*,1], **Olga Vitek**[†], **Federico Gaiti**[‡], **Milos Tanurdzic**[§,1] **and Jo Ann Banks**[**,1]

[*]Purdue University Center for Cancer Research, Purdue University, West Lafayette, IN 47907, United States, [†]College of Computer and Information Science, Northeastern University, Boston, MA 02115, United States, [‡]Department of Medicine, Weill Cornell Medicine and New York Genome Center, New York, NY 10013, United States, [§]School of Biological Sciences, The University of Queensland, St. Lucia, Australia, [**]Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, 47907, United States

ABSTRACT The fern *Ceratopteris richardii* is an important model for studies of sex determination and gamete differentiation in homosporous plants. Here we use RNA-seq to *de novo* assemble a transcriptome and identify genes differentially expressed in young gametophytes as their sex is determined by the presence or absence of the male-inducing pheromone called antheridiogen. Of the 1,163 consensus differentially expressed genes identified, the vast majority (1,030) are up-regulated in gametophytes treated with antheridiogen. Our GO term enrichment analyses of these DEGs reveals that a large number of genes involved in epigenetic reprogramming of the gametophyte genome are up-regulated by the pheromone. Additional hormone response and development genes are also up-regulated by pheromone. This *C. richardii* transcriptome and gene expression dataset will prove useful for studies focusing on sex determination and differentiation in plants.

KEYWORDS
Sex determination
RNA-seq
Ceratopteris
gametophyte
epigenetics
gibberellin
antheridiogen
transcriptome

SUPPLEMENTAL TABLES

Table S1. A list of qRT-PCR primers used in expression validation.

Table S2. Run metrics for Trinity assembly and differential expression analysis.

Table S3. Excel spreadsheet of ABA and GA primary biosynthetic and signaling components. Components were identified by using the *Ceratopteris* gametophyte transcriptome as a query and the *Arabidopsis* proteome (TAIR10) database.

Table S4. Excel spreadsheet of all 1,163 differentially expressed genes and statistical support.

Table S1 qRT PCR Primers

Gene	Forward Sequence	Reverse Sequence
$CrEF1\alpha$	5'CAGACCAGTCGGAGCAAAAGT'3	5'TCCTGTGGGAAGGGTGGAA'3
comp39080	5'CGCAAGGGATAGCCAAATTA'3	5'CGATCTCAACGCGATCTACA'3
comp82638	5'CTGCTGCCTCTCAGTGTGAC'3	5'ATCACGCGCTTGTAGGACTT'3
comp114251	5'AGCTCAAATGCCACCACTTT'3	5'ACATAGCCGCTGCTGTTCTT'3
comp38095	5'ATGCCGAATGGAAGACTGTT'3	5'TTCATATTCGGCGACTCCTT'3
comp82048	5'GGTATGACGCCACAGAACCT'3	5'TGCAGACATTGCAGGATACC'3
comp103387	5'TCGAAAGAGAGGCAACACCT'3	5'ACTTTCCGAGAAGCAGTGGA'3
comp46913	5'TGGGCAAACTTCAGGTAAGG'3	5'TGAGGCTGTGTCAGAGATGC'3
comp105977	5'AGGAAATCGCTGGACGTAGA'3	5'CCTCATCCTTCCAACATCGT'3
comp110703	5'GAGGTAAGGCAAGCGCTCTA'3	5'CCAACGGCCATGAGAAGTAT'3
comp109704	5'GGCGAAATACCTGCAAATGT'3	5'TCACGACACACAACCACAGA'3
comp84184	5'ATGGGCAGATGGTGGAAATA'3	5'TGACCATTGTCTCCCTCAGA'3

Table S2 Run Metrics, assembly and analysis statistics for the combined, $-A_{CE}$ and $+A_{CE}$ datasets

	$-A_{CE}1$	$-A_{CE}2$	$-A_{CE}3$	$+A_{CE}1$	$+A_{CE}2$	$+A_{CE}3$
Run Metrics						
Total bases	3,062,485,438	1,576,208,121	3,371,604,624	3,779,164,066	3,303,557,793	3,387,042,171
Total reads	30,321,638	15,606,021	33,382,224	37,417,466	32,708,493	33,535,071
Average GC%	47	47	47	46	46	46
% reads Phred scores >20	95	95	93	96	96	96
% reads Phred scores >30	71	73	65	76	86	86
Number reads aligned	10,934,932	12,955,923	18,336,178	30,072,430	26,072,076	26,617,968
% reads aligned	88	87	89	88	88	88
Analysis						
	Combined Data	$-A_{CE}$	$+A_{CE}$			
DESeq DEGs	1,183	140	1,043			
edgeR DEGs	3,700	1,585	2,115			
EBSeq DEGs	3,065	1,065	2,000			
Consensus DEGs	1,163	133	1,030			
Assembly						
Total transcripts assembled	206,059					
Total genes assembled	111,977					
N50	1,988					
Min length	151					
Max length	17,306					
Average length	867					
Genes with read support	82,820					

SUPPLEMENTAL FIGURES

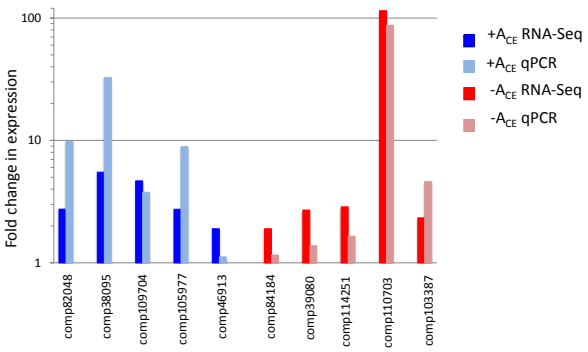

Figure S1 Comparison of gene expression from qRT-PCR vs. RNA-Seq. The fold changes for the qRT-PCR data were calculated using the ΔCt method (Livak and Schmittgen 2001). Blue bars show fold change for genes with higher expression in $+A_{CE}$ samples, red bars show fold changes for genes more highly expressed in $-A_{CE}$ samples. The genes comp46913 and comp84184 did not exhibit significant differences between $+A_{CE}$ and $-A_{CE}$ samples.

LITERATURE CITED

Livak, K. J. and T. D. Schmittgen, 2001 Analysis of relative gene expression data using real-time quantitative pcr and the 2- $\delta\delta$ct method. methods **25**: 402–408.

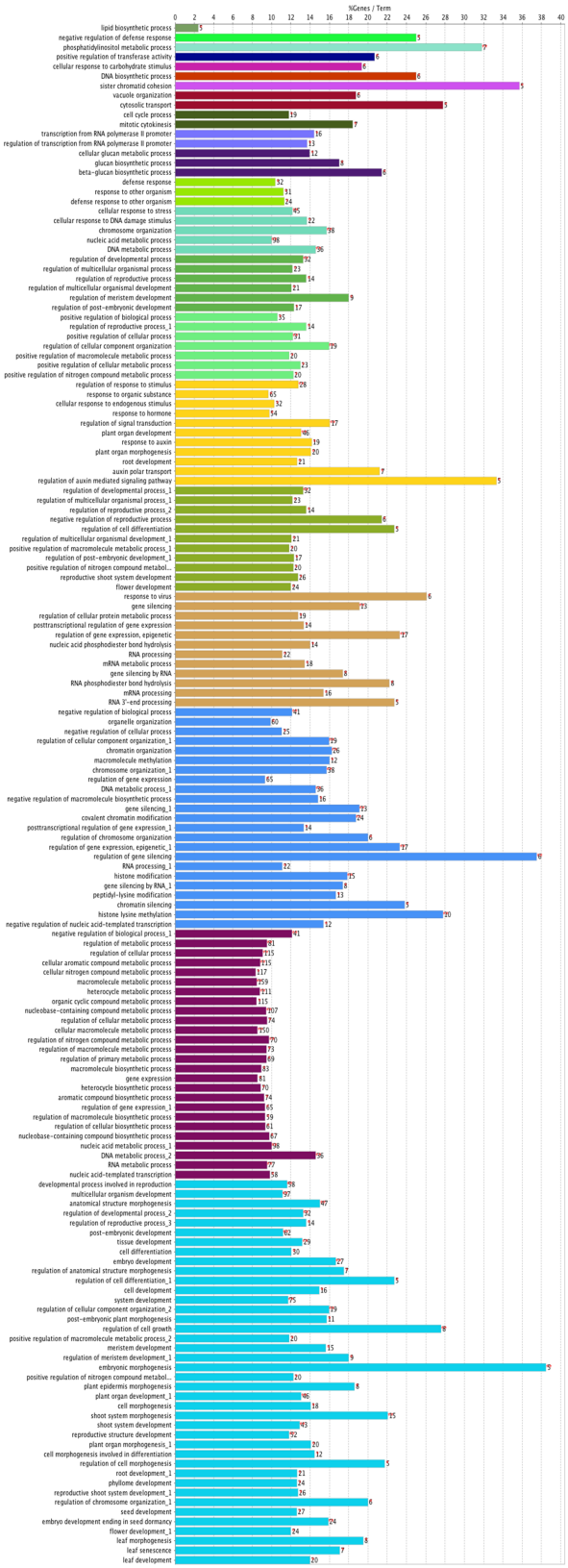

Figure S2 GO terms associated with each node in Figure 3. Numbers above bars indicate number of genes mapped to the category and red asterisks indicate degree of statistical significance.