Analysis of Cambridge isolates of *Marchantia polymorpha*

Bernardo Pollak, Mihails Delmans and Jim Haseloff
Department of Plant Sciences
University of Cambridge
Downing Street.
Cambridge. United Kingdom. CB2 3EA.

Summary

1. *The Cam-1 and Tak-1 isolates of Marchantia polymorpha differ by approximately 0.0015% within predicted exon regions. The amount of SNPs and indels increase to 0.0045% and 0.0066% in introns and intragenic regions, respectively. A total of 1,010,971 SNPs and indels were found in the Marchantia isolates. This can be compared to the situation in Arabidopsis, with 349,171 SNPs found when comparing the Landsberg erecta and Columbia accessions (Lu et al., 2012), and 4,902,039 SNPs found across 80 Arabidopsis strains (Cao et al., 2011).*

2. *Marchantia spores provide a unique system of synchronised germination for examining the genetic and cellular processes that underpin chloroplast differentiation, cell division and expansion and differentiation of a plant body plan. We have mapped transcriptomic changes over the first 96 hours of spore germination.*

3. *We have used the annotated Tak-1 genome as a reference and for consistency have adopted the metrics used by Tom Yamato and colleagues. We have included an extra requirement of minimum FPKM expression to reduce false positive assignment.*

4. *There are notable and coordinated shifts in accumulation of gene transcripts associated with chloroplast and cell wall functions. Using the same criteria, we identify 366 gene transcripts that are >10x enriched in sporeling tissues, higher than the existing estimate of 59. The measurements of gene transcription levels are consistent across the finely spaced data set.*

5. *The data provides a list of differentially regulated genes that will a source for identifying new promoters and genes active in early growth and development in Marchantia.*

Sequence comparison of the Cam-1 and Tak-1 isolates.

Cam1 genome sequencing reads were aligned to Tak1 genome assembly using the Burrows - Wheeler Aligner (Li and Durbin, 2009). Duplicated reads were removed from the alignment using MarkDuplicates tool from Picard (http://picard.sourceforge.net).

SNP and indels were called using The Genome Analysis Toolkit (McKenna *et al.*, 2010) featuring a variant quality score recalibration algorithm described in (De Wit et al., 2012).

Alignment Statistics	
Total number of reads	127,336,086
Average phred quality score	34.7
Number of mapped reads	105,733,522 (80%)
bp covered	204,022,363 (90%)
Average number of reads supporting a bp	51

SNP/indel Analysis					
Region	**SNP/Indels**	**% of all loci**	**Length of region (bp)**	**% of total length**	**SNP/indels per 1kb**
Intergenic	1,010,971	82%	153,390,713	68%	6.59
Intragenic	224,762	18%	72,370,426	32%	3.11
- Exons	48,423	4%	32,752,581	15%	1.48
- Introns	176,339	14%	39,617,845	18%	4.45
Total	1,235,733	100%	225,761,139	100%	5.47

References:

Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis.

Pingli Lu, Xinwei Han, Ji Qi, Jiange Yang, Asela J. Wijeratne, Tao Li and Hong Ma (2012) Genome Research 22:508–518.

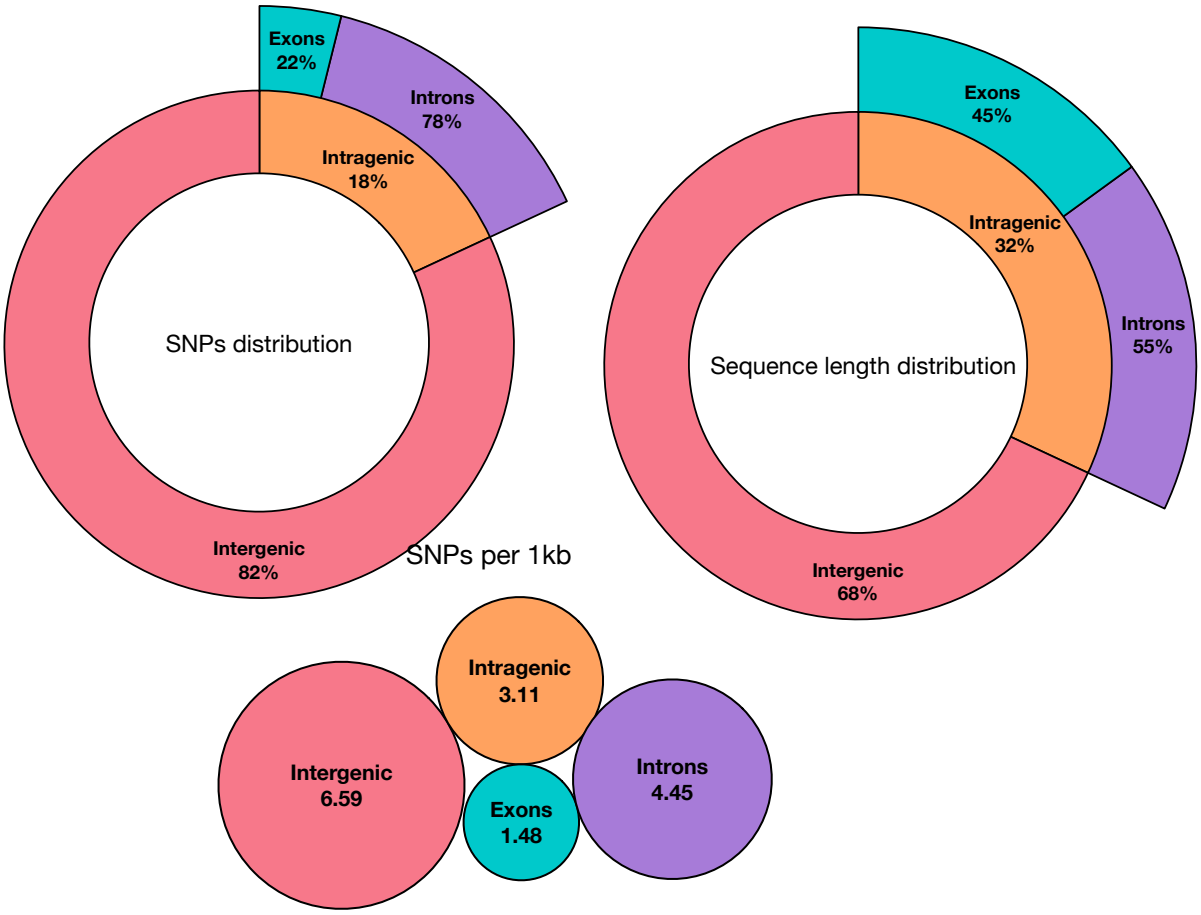
Jun Cao, Korbinian Schneeberger, Stephan Ossowski, Torsten Günther, Sebastian Bender, Joffrey Fitz, Daniel Koenig, Christa Lanz, Oliver Stegle, Christoph Lippert, Xi Wang, Felix Ott, Jonas Müller, Carlos Alonso-Blanco, Karsten Borgwardt, Karl J Schmid & Detlef Weigel. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics. 43:956-965.

De Wit P,. Pespeni M., Ladner J., Barshis D., Seneca F., Jaris H., Therkildsen N., Morikawa M., Palumbi S. (2012) The simple fool's guide to population genomics via RNA-Seq: An introduction to high-throughput sequencing data analysis. Molecular Ecology Resources 12:1058-1067.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. (2010) The Genome Analysis Toolkit: a

MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-303.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.



Transcriptome analysis of germinating sporelings

Methods

Crosses

Marchantia polymorpha Cam-1 & Cam-2 gametophytes were grown in soil and induced by far-red LED exposure for sexual organ development. Sperm was collected from Cam-1 antheridia and sprayed onto Cam-2 archegoniophores every 2-3 days for 1 week. After 3 weeks, sporangia began emerging and spore heads (archegoniophores bearing sporangia) were collected and dried with silica gel on 50mL Falcon tubes for 3 days. Spore heads were stored at -80ºC.

Spore germination and axenic growth

Twenty frozen spore heads were crushed and and resuspended with 1mL of sterile water per spore head. Resuspended spores were filtered through a 40μM cell strainer and spun down at 13.000 RPM for 1 minute. Spores were sterilised by adding 1mL of Milton's sterilising solution (Milton Mini Sterilising Tablet (active ingredient: sodium dichloroisocyanurate CAS: 2893-78-9: 19.5% (w/w), Laboratoire Rivadis, www.milton-tm.com) in 25mL of sterile water, and incubated at RT for 20 minutes in a rotating shaker at 100RPM. Spores were then spun down and washed once with 1mL of water. Finally they were resuspended in 100uL per spore head and sown onto 1/2 Gamborg's B5 plates with 1% agar in decreasing amounts in respect to the expected mass obtained through germination (1/3 of total spores were kept for time 0h, 1/4 of total spores for time 24h, 1/5 for time 48h, 1/8 for time 72h and 1/10 for time 96h). Plates were sealed with micropore tape and placed in the culture chamber on a 16:8 day/night cycle at 21ºC, with 60 µmol photons m^{-2} s^{-1} of illumination.

Sample collection and sequencing

Spores were collected from plates at the same hour of the day, every 24 hours for 4 days using sterile water with an L-spreader, and placed on 1.5mL Eppendorf tubes. Excess liquid was removed and tubes stored at -80ºC. Three biological replicates were performed for each time point. Total RNA was extracted using Qiagen RNeasy mini kit and RNA concentration was measured with a Qubit fluorimeter using the RNA BR assay kit. RNA integrity was assessed with a Bioanalyzer 2100 machine using the Agilent RNA 6000 Nano kit and samples were sent to BGI for further processing. Library preparation was performed with the Illumina TruSeq RNA library preparation kit v.2 and 100bp paired-end sequencing was carried out on a Illumina HiSeq 2000 machine, with samples multiplexed on 2 lanes.

Microscopy

During sample collection, small aliquots were used for microscopy to assess spore germination. Ten μL of resuspended sporelings were placed on a glass slide and a 22 by 22 mm coverslip was placed on top of the droplet. Imaging was carried out in a Leica DMI6000 B inverted fluorescence microscope with a 63X oil immersion objective.

Results

Imaging of spores showed distinct developmental transitions during the time span studied (Figure 1). During early gemination, between 0 and 48 hours, imbibition and expansion of the sporeling occurs where massive swelling of the sporeling can be seen. As early as 24 hours, chloroplast differentiation and proliferation can already be observed and chlorophyll autoflorescence can be detected, consecutively an increase in size and number of chloroplasts follows. Rhizoids start emerging after 48 hours of germination and the photosynthetic cell begins dividing between 48 and 72 hours, polarising the sporeling and defininig a clear axis. Further divisions follow at 96 hours where expansion of the photosynthetic region occurs, increasing considerably the photosynthetic region.

RNA sequencing produced > 24M clean paired-end (PE) reads per time point (Table 1) on average. Reads were mapped to the Marchantia polymorpha v3.1 genome using the STAR alignment software (Dobin, *et al*., 2013) and mapped reads were processed through Cuffdiff (Trapnell, et al., 2013). Sporeling FPKM normalised read counts were then loaded onto the FPKM matrix as sporeling datasets and enrichment analysis for sporelings was carried out by obtaining the ratio between the maximum value of any sporeling dataset over the value of any other tissue. Genes were defined as enriched in sporeling tissue if the ratio was > 10. Also, genes having infinity ratios (where all other tissues had 0 FPKM values) were assigned as enriched if the maximum expression value of the sporeling datasets was > 0.3 FPKM, as to avoid extremely low counts from being assigned as enriched. This approach produced 366 genes enriched in sporeling tissue, which can be found in the supplementary enriched genes table.

To create graphs of differential expression analysis, mapped reads were counted using the featureCount (Liao et al., 2014) function of the subread package, yielding a raw count for reads mapped to exons. Counts were added to produce raw counts for isoforms and a table containing all counts to genes from each sample was created. Filtering by the sporeling enriched gene set was carried out and differential expression analysis was performed using the DESeq2 package (Love et al., 2014) using only time as a factor and 3 biological replicates per time point and an Benjamini-Hochberg adjusted p value > 0.05. Transcripts with < 10 counts were removed from further analysis. A heatmap of r-log transformed counts for the top 60 regulated genes is shown in Figure 2. This dataset provides insight into sporeling specific developmental processes occurring during early germination and allows identification of genes regulated at specific stages of sporeling development. Number of genes regulated between time points is shown in Table 2.

Top gene ontology (GO) terms are shown in Table 3 and a full list is included in the supplementary enriched genes table. Among the terms found in the GO list, a number of metabolic processes such as chloroplast related functions and cell wall related functions appear to be overrepresented. Furthermore, we performed protein BLAST searches agains a Viridiplantae filtered Uniprot database using an e-value cutoff of $10e^{-4}$ to identify putative orthologs. From the 366 sporeling enriched genes, we found 186 unique hits. From this putative ortholog list, we found proteins such as transcription factors, cell cycle regulators, hormonal signalling, nitrogen metabolism, transcription-related proteins, translation, and among the most highly represented, cell wall remodelling and expansion proteins (expansins, xyloglucan endotransglycosylases), and photosynthesis (photosystem proteins), carbon fixation (RuBisCO) and chloroplast replication

proteins (ACR). Selected putative orthologs for chloroplast-related functions can be found in Table 4 and for cell wall-related functions in Table 5. Complete putative ortholog list is included in the supplementary enriched genes table.

References:

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15-21. (2013).

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 31:46-53. (2013).

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30:923-30. (2014).

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550. (2014).

Figures

Figure 1. Assessment of germination by fluorescence microscopy. Autofluorescence is shown in the red channel overlaid over bright field images. Images were taken at the time of sample collection. Scale bar shown at bottom right corner represents 10 um.

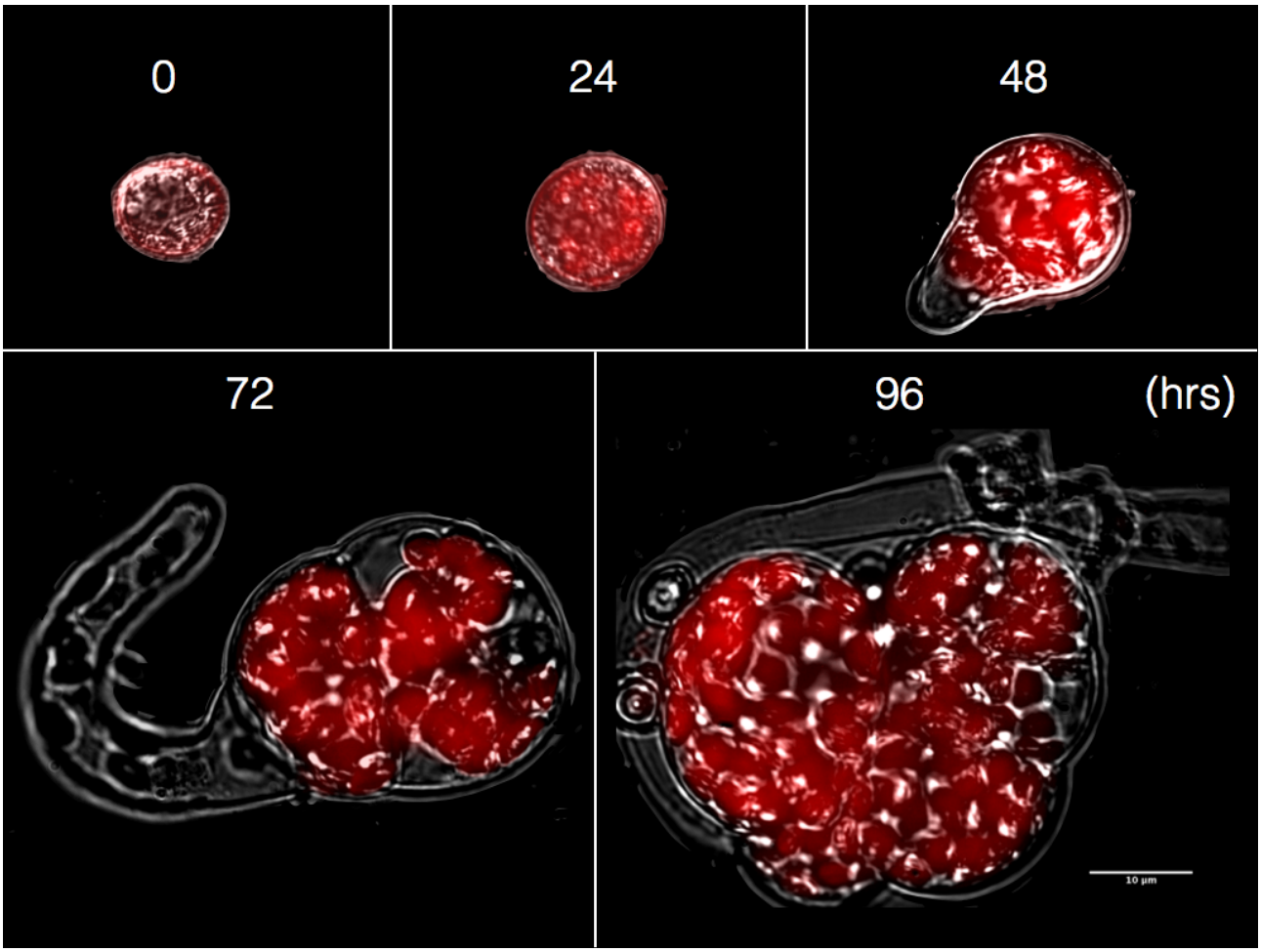


Figure 2. Heatmap for the top 60 regulated genes resulting from differential expression analysis of the sporeling-enriched gene set by DEseq2 pipeline. Heat map shows clusters for sets of genes following similar regulation dynamics across time course.

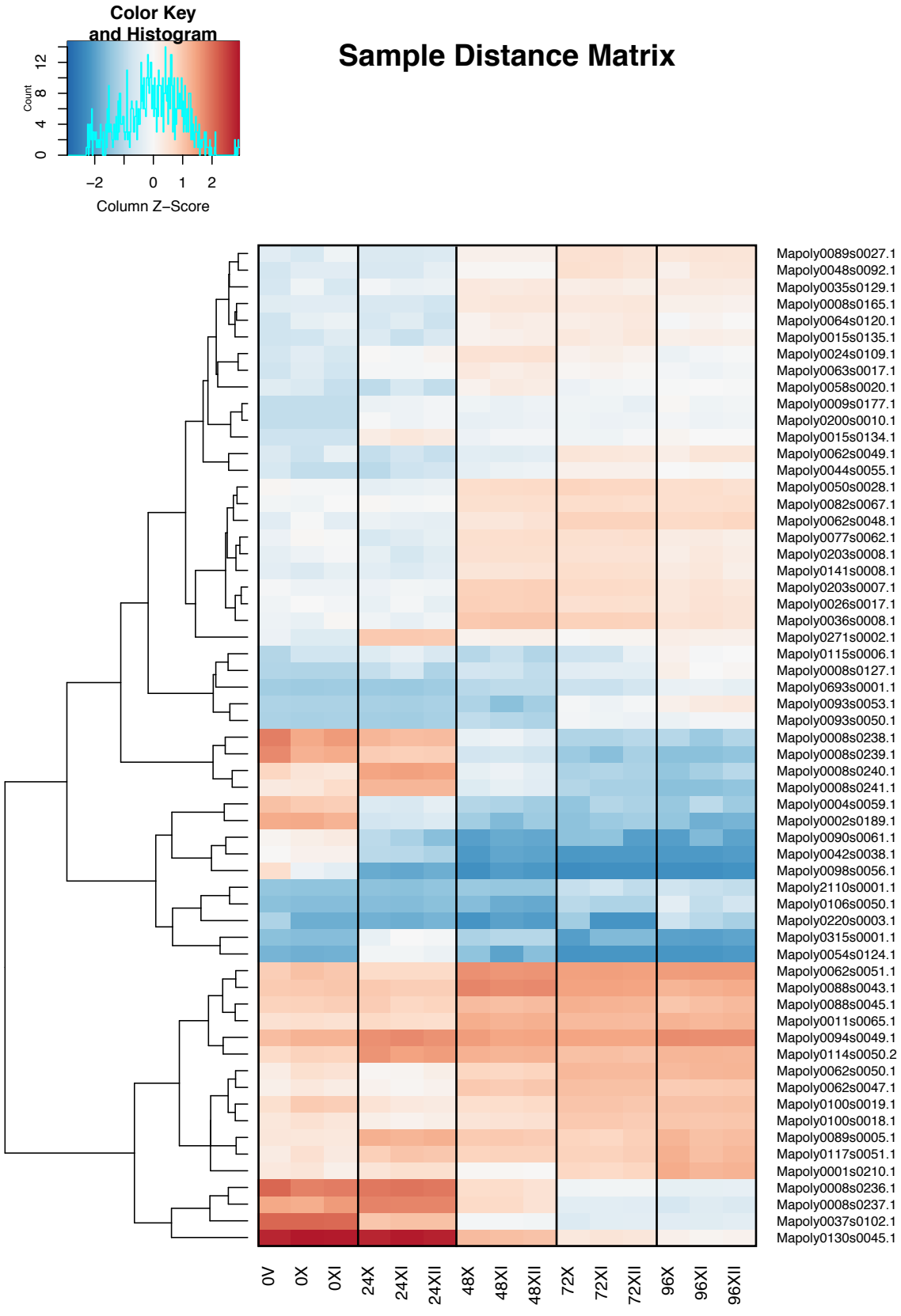


Table 1. Reads obtained from sequencing. Reads represent clean filtered reads, stripped of adapter contamination and filtered by quality values > 20.

Sample	Replicate 1 reads (PE)	Replicate 2 reads (PE)	Replicate 3 reads (PE)
0h	31,341,564	23,292,343	24,838,123
24h	26,186,219	26,485,860	27,220,829
48h	27,052,956	25,589,507	20,184,573
72h	24,697,479	26,233,384	23,261,433
96h	26,028,608	25,222,677	24,214,899

Table 2. Number of sporeling-enriched regulated genes between time points during sporeling germination.

Time points (hrs)	Up-regulated	Down-regulated	Total
0 vs 24	52	63	115
24 vs 48	47	50	97
48 vs 72	22	52	74
72 vs 96	39	16	55
0 vs 96	82	69	151

Table 3. GO terms for sporeling enriched genes.

ID	Term	Nº
GO:0005515	Protein binding	23
GO:0055114	Oxidation-reduction process	18
GO:0016020	Membrane	16
GO:0004672	Protein kinase activity	11
GO:0005524	ATP binding	11
GO:0006468	Protein phosphorylation	11
GO:0020037	Heme binding	10
GO:0004601	Peroxidase activity	8
GO:0006979	Response to oxidative stress	8
GO:0008270	Zinc ion binding	8
GO:0016021	Integral to membrane	8
GO:0016491	Oxidoreductase activity	8
GO:0003676	Nucleic acid binding	7
GO:0003723	RNA binding	6
GO:0003824	Catalytic activity	6
GO:0004553	O-glycosyl hydrolase activity	6
GO:0005199	Structural constituent of cell wall	6
GO:0005618	Cell wall	6
GO:0005975	Carbohydrate metabolic process	6
GO:0009664	Plant-type cell wall organization	6
GO:0055085	Transmembrane transport	6
GO:0006073	Cellular glucan metabolic process	5
GO:0006396	RNA processing	5
GO:0015979	Photosynthesis	5
GO:0048046	Apoplast	5
GO:0005515	Protein binding	4
GO:0005525	GTP binding	4
GO:0008152	Metabolic process	4
GO:0009055	Electron carrier activity	4
GO:0009522	Photosystem I	4
GO:0016762	Xyloglucan:xyloglucosyl transferase activity	4

Table 4. Putative orthologs for sporeling enriched genes with chloroplast-related functions.

Isoform id	Source organism	Ortholog	Full name	E-value
Mapoly0006s0012.1	Brassica napus	LPAT1	1-acyl-sn-glycerol-3-phosphate acyltransferase 1, chloroplastic	4.00E-06
Mapoly0036s0031.2	Spinacia oleracea	RR1	30S ribosomal protein S1, chloroplastic	2.00E-29
Mapoly0090s0058.1	Acutodesmus obliquus	RR4	30S ribosomal protein S4, chloroplastic	1.00E-22
Mapoly0118s0046.1	Acorus calamus	RK16	50S ribosomal protein L16, chloroplastic	2.00E-08
Mapoly0058s0064.1	Arabidopsis thaliana	NTH1	Endonuclease III homolog 1, chloroplastic	5.00E-40
Mapoly0120s0016.1	Arabidopsis thaliana	HF101	Fe-S cluster assembly factor HCF101, chloroplastic	3.00E-04
Mapoly0054s0124.1	Arabidopsis thaliana	PHT21	Inorganic phosphate transporter 2-1, chloroplastic	5.00E-58
Mapoly0001s0001.1	Adansonia digitata	MATK	Maturase K	4.00E-04
Mapoly0015s0004.1	Acorus americanus	NU5C	NAD(P)H-quinone oxidoreductase subunit 5, chloroplastic	3.00E-07
Mapoly0033s0050.1	Oryza sativa subsp. japonica	PWD	Phosphoglucan, water dikinase, chloroplastic	4.00E-06
Mapoly0085s0066.1	Acorus calamus	YCF3	Photosystem I assembly protein Ycf3	5.00E-12
Mapoly0033s0038.1	Acorus calamus	PSAA	Photosystem I P700 chlorophyll a apoprotein A1	2.00E-34
Mapoly0149s0029.1	Araucaria araucana	PSAA	Photosystem I P700 chlorophyll a apoprotein A1 (Fragment)	2.00E-26
Mapoly0043s0109.1	Acorus calamus	PSAB	Photosystem I P700 chlorophyll a apoprotein A2	4.00E-20
Mapoly0125s0033.1	Arabidopsis thaliana	PSAG	Photosystem I reaction center subunit V, chloroplastic	6.00E-05
Mapoly0124s0050.3	Arabidopsis thaliana	PSY	Phytoene synthase, chloroplastic	3.00E-04
Mapoly0013s0012.1	Oryza sativa subsp. japonica	DXS2	Probable 1-deoxy-D-xylulose-5-phosphate synthase 2, chloroplastic	2.00E-07
Mapoly0034s0092.4	Oryza sativa subsp. japonica	1-Apr	Probable 5'-adenylylsulfate reductase 1, chloroplastic	2.00E-15
Mapoly0049s0016.1	Arabidopsis thaliana	CGEP	Probable glutamyl endopeptidase, chloroplastic	1.00E-136
Mapoly0090s0057.1	Arabidopsis thaliana	CRTSO	Prolycopene isomerase, chloroplastic	2.00E-05
Mapoly0003s0272.8	Arabidopsis thaliana	DEGP1	Protease Do-like 1, chloroplastic	4.00E-04
Mapoly0064s0104.2	Arabidopsis thaliana	ARC6	Protein Accumulation and Replication of Chloroplasts 6, chloroplastic	1.00E-162
Mapoly0062s0051.1	Arabidopsis thaliana	CHUP1	Protein CHUP1, chloroplastic	8.00E-05
Mapoly0172s0015.3	Chlamydomonas reinhardtii	MBB1	PsbB mRNA maturation factor Mbb1, chloroplastic	7.00E-04
Mapoly0046s0082.2	Nicotiana tabacum	KPYA	Pyruvate kinase isozyme A, chloroplastic	7.00E-131
Mapoly0007s0124.1	Chlamydomonas reinhardtii	RBL	Ribulose bisphosphate carboxylase large chain	1.00E-15

Isoform id	Source organism	Ortholog	Full name	E-value
Mapoly0114s0050.2	Chlamydomonas reinhardtii	RBS1	Ribulose bisphosphate carboxylase small chain 1, chloroplastic	9.00E-42
Mapoly0032s0002.1	Arabidopsis thaliana	VPS54	Vacuolar protein sorting-associated protein 54, chloroplastic	8.00E-16

Table 5. Putative orthologs for sporeling enriched genes with cell wall-related functions.

Isoform id	Source organism	Ortholog	Full name	E-value
Mapoly0062s0047.1	Nicotiana tabacum	EXTN	Extensin	2.00E-08
Mapoly0062s0048.1	Arabidopsis thaliana	EXTN1	Extensin-1	2.00E-12
Mapoly0062s0050.1	Arabidopsis thaliana	EXTN3	Extensin-3	1.00E-12
Mapoly0009s0002.2	Olea europaea	ALL9	Glucan endo-1,3-beta-D-glucosidase	4.00E-25
Mapoly0203s0008.1	Arabidopsis thaliana	XTH10	Probable xyloglucan endotransglucosylase/hydrolase protein 10	2.00E-80
Mapoly0094s0049.1	Arabidopsis thaliana	XTH11	Probable xyloglucan endotransglucosylase/hydrolase protein 11	2.00E-43
Mapoly0089s0005.1	Arabidopsis thaliana	XTH16	Probable xyloglucan endotransglucosylase/hydrolase protein 16	3.00E-86
Mapoly0130s0045.1	Arabidopsis thaliana	XTH1	Putative xyloglucan endotransglucosylase/ hydrolase protein 1	5.00E-68
Mapoly0203s0007.1	Arabidopsis thaliana	XTH14	Xyloglucan endotransglucosylase/ hydrolase protein 14	2.00E-79