The resurrection genome of *Boea hygrometrica*: A blueprint for survival of dehydration

Lihong Xiao[a], Ge Yang[a], Liechi Zhang[a], Xinhua Yang[b], Shuang Zhao[b], Zhongzhong Ji[a], Qing Zhou[b], Min Hu[b], Yu Wang[a], Ming Chen[b], Yu Xu[a], Haijing Jin[a], Xuan Xiao[a], Guipeng Hu[a], Fang Bao[a], Yong Hu[a], Ping Wan[a], Legong Li[a], Xin Deng[c], Tingyun Kuang[d], Chengbin Xiang[e], Jian-Kang Zhu[f,g,1], Melvin J. Oliver[h,1], and Yikun He[a,1]

[a]School of Life Sciences, Capital Normal University, Beijing 100048, China; [b]Beijing Genomics Institute-Shenzhen, Shenzhen 518083, China; [c]Key Laboratory of Plant Resources and [d]Key Laboratory of Photobiology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China; [e]School of Life Sciences, University of Science and Technology of China, Hefei 230022, China; [f]Shanghai Center for Plant Stress Biology, Chinese Academy of Sciences, Shanghai 200032, China; [g]Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907; and [h]Plant Genetics Research Unit, Midwest Area, Agricultural Research Service, United State Department of Agriculture, University of Missouri, Columbia, MO 65211

Contributed by Jian-Kang Zhu, March 26, 2015 (sent for review February 10, 2015; reviewed by Sagadevan G. Mundree and Andrew J. Wood)

"Drying without dying" is an essential trait in land plant evolution. Unraveling how a unique group of angiosperms, the Resurrection Plants, survive desiccation of their leaves and roots has been hampered by the lack of a foundational genome perspective. Here we report the ∼1,691-Mb sequenced genome of *Boea hygrometrica*, an important resurrection plant model. The sequence revealed evidence for two historical genome-wide duplication events, a compliment of 49,374 protein-coding genes, 29.15% of which are unique (orphan) to *Boea* and 20% of which (9,888) significantly respond to desiccation at the transcript level. Expansion of early light-inducible protein (ELIP) and 5S *rRNA* genes highlights the importance of the protection of the photosynthetic apparatus during drying and the rapid resumption of protein synthesis in the resurrection capability of *Boea*. Transcriptome analysis reveals extensive alternative splicing of transcripts and a focus on cellular protection strategies. The lack of desiccation tolerance-specific genome organizational features suggests the resurrection phenotype evolved mainly by an alteration in the control of dehydration response genes.

vegetative desiccation tolerance | resurrection plant | *Boea hygrometrica* | drought tolerance enhancement | genome

Resurrection plants constitute a unique cadre within the angiosperms: they alone have the remarkable capability to survive the complete dehydration of their leaves and roots. How the dry and visually "dead" plants come alive when water becomes available has long fascinated plant biologists and the lay public alike. The majority of plants, including all our crops, can rarely survive tissue water potentials of less than −4 Mpa. Resurrection plants can, in contrast, survive tissue water potentials of −100 MPa (equilibration to air of 50% relative humidity) and below. The ability to desiccate and resurrect vegetative tissues is considered a primal strategy for surviving extensive periods of drought (1). Desiccation tolerance (DT) has played a major role in plant evolution (1): Postulated as critical for the colonization of terrestrial habitats. DT, as it relates to seed survival and storage, is also arguably the primary plant trait that governs global agriculture and food security. Vegetative DT was lost early in the evolution of tracheophytes (1) and is rare in the angiosperms, but has since reappeared within several lineages, at least 13 of which belong to the angiosperms (2).

Vegetative DT is a complex multigenic and multifactorial phenotype (3–5), but understanding how DT plants respond to and survive dehydration has great significance for plant biology and, more directly, for agriculture. Resurrection plants offer a potential source of genes for improvement of crop drought tolerance (5, 6) as the demand for fresh water grows (7).

In recent decades, efforts have been focused on exploring the structural, physiologic, and molecular aspects of DT in a number of plant species (4). Although a functional genomic approach has been fruitful in revealing the intricacies of DT in resurrection plants (5, 8), and a system approach is contemplated (4), efforts are hampered by the lack of a sequenced genome for any of the resurrection plants. To fill this critical gap, we sequenced the genome of one of the important DT models (9), *Boea hygrometrica*.

B. hygrometrica is a homiochlorophyllous dicot in Gesneriaceae that grows in rocky areas throughout most of China (10). Not only is the whole plant DT (Fig. 1*A*), but a detached leaf or leaf segment retains the DT phenotype and can regenerate a new "seedling" even after several dehydration and rehydration cycles (Fig. 1*B* and *SI Appendix*, Fig. S1 *A* and *B*) (11). Drying leaf tissues exhibit classical dehydration-associated structural changes (12), including a folded cell wall and condensed cytoplasm (*SI Appendix*, Fig. S1 *C–E*).

Here we present a high-quality draft genome of *B. hygrometrica*, along with a full assessment of the changes in the leaf transcriptomes that occur during desiccation and that relate to the resurrection phenotype.

Results

Whole-Genome Features. The whole-genome shotgun sequenced draft genome of *B. hygrometrica* delivers a ∼1,548-Mb assembly,

Significance

The genome analysis presented here represents a major step forward in the field of desiccation tolerance and a much-anticipated resource that will have a far-reaching effect in many areas of plant biology and agriculture. We present the ∼1.69-Gb draft genome of *Boea hygrometrica*, an important plant model for understanding responses to dehydration. To our knowledge, this is the first genome sequence of a desiccation-tolerant extremophile, offering insight into the evolution of this important trait and a first look, to our knowledge, into the genome organization of desiccation tolerance. The underpinning genome architecture and response in relation to the hydration state of the plant and its role in the preservation of cellular integrity has important implications for developing drought tolerance improvement strategies for our crops.

Author contributions: L.X., T.K., and Y. He designed research; L.X., G.Y., L.Z., and Z.J. performed research; Y.W., Y.X., H.J., X.X., G.H., F.B., Y. Hu, L.L., X.D., and C.X. contributed new reagents/analytic tools; L.X., X.Y., S.Z., Q.Z., M.H., M.C., P.W., J.-K.Z., M.J.O., and Y. He analyzed data; and L.X., J.-K.Z., M.J.O., and Y. He wrote the paper.

Reviewers: S.G.M., Queensland University of Technology; and A.J.W., Southern Illinois University.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE48671), and the BioSample database, www.ncbi.nlm.nih.gov/biosample (accession no. SAMN02215335).

[1]To whom correspondence may be addressed. Email: jkzhu@purdue.edu, Mel.Oliver@ars.usda.gov, or yhe@cnu.edu.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1505811112/-/DCSupplemental.

PLANT BIOLOGY

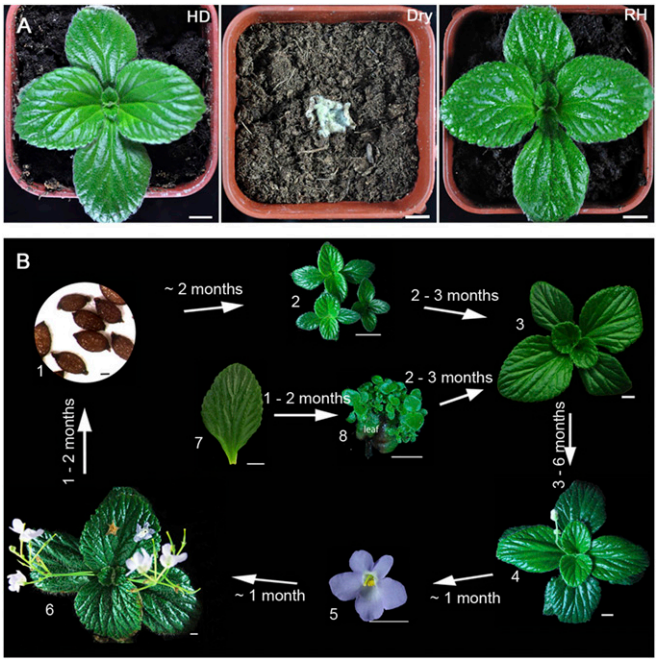

Fig. 1. Phenotypes during the dry–rehydration cycle and life cycle of *B. hygrometrica*. (*A*) Vegetative phenotypes of hydrated (HD), dry (2 weeks withholding water), and rehydrated for 48 h (RH) *B. hygrometrica*. (*B*) Life cycle of *B. hygrometrica* from seed germination or leaf regeneration to mature plant. (Scale bar for seed morphology, 1 mm; scale bar for plants, 1 cm.)

generated from 4.74×10^{11} high-quality reads (*SI Appendix*, Table S1), and represents 91.52% of the ~1,691-Mb estimated genome size (*SI Appendix*, Table S2) predicted from 17-nucleotide depth distribution (*SI Appendix*, Fig. S2 and Table S3). The assembly was generated by an iterative hybrid approach (Table 1 and *SI Appendix*, Fig. S3). Approximately 85.86% of the assembly is nongapped sequence. The quality of the assembly was assessed by alignment to Sanger-derived fosmid sequences, allowing only a limited potential for misassemblies (*SI Appendix*, Table S4 and Dataset S1). The extent of sequence coverage was confirmed by the mapping of 2,360 sequenced expressed sequence tags (*SI Appendix*, Table S5 and Dataset S2).

The fourfold degenerate synonymous site of the third codon position (4DTv) values for coding regions for each of the duplicate gene pairs in the pairwise orthologous segments within *B. hygrometrica* genome revealed two whole-genome duplication events (4DTv ~0.5 and ~1.0; Fig. 2*A*). The species divergence event between *B. hygrometrica* and *Solanum tuberosum* or *Solanum lycopersicum* (4DTv ~0.54 or 0.49) that occurred around the most recent duplication event in the *B. hygrometrica* genome (4DTv ~0.5) likely reflects the divergence of the Lamiales from the Solanales (Fig. 2*A*). The ancient duplications, composed of several intermittent small duplication events (4DTv ~0.9 to ~1.3), may explain the large genome size, high level of repetitive sequences, and multicopy genes in the *B. hygrometrica* genome. The *B. hygrometrica* genome possessed a higher guanine-cytosine (GC) content (42.30%) than *S. tuberosum*, *S. lycopersicum*, or *Arabidopsis thaliana* (Table 1 and *SI Appendix*, Fig. S4), which is close to the upper limit for dicots (13). More than three fourths of the genome is composed of repeat sequences (75.75% of the assembled genome; Table 1 and *SI Appendix*, Fig. S5 and Table S6), which is similar to other dicots (14) but somewhat higher than *S. tuberosum* (62.2%) (15). Much of the unassembled genome is also composed of repetitive sequences, and the majority of the repetitive sequences could not be associated with known transposable element families. Plant transposable elements (TEs) are a significant source of small RNAs that function to epigenetically regulate TE and gene activity and are known to regulate DT in dicots (16). A recently discovered retroelement expressed in *B. hygrometrica*, osmotic and alkaline resistance 1, strengthens the possible role for LTRs in stress tolerance, and perhaps DT (17).

The draft genome also encodes 196 microRNA (miRNA), 538 tRNA, 1,512 *rRNA*, and 151 snRNA genes (*SI Appendix*, Table S7). In comparison with other dicot genomes (18), the *B. hygrometrica* genome encodes a large number of *rRNA* genes, especially 5S *rRNA* genes. Apart from their obvious structural role in ribosomes, large numbers of *rRNA* repeats (rDNA) have been linked with DNA stability, at least in yeast (19): a function that would be advantageous for surviving desiccation. There are 1,119 5S *rRNA* genes interspersed throughout the genome. This is 25–50 times the number contained in the only two other Asterid genomes that have been sequenced: *S. lycopersicum* (47 5S *rRNA* genes) and *S. tuberosum* (23 5S *rRNA* genes). The majority of the 5S *rRNA* genes are interspersed throughout the genome (Dataset S3); only 34 were clustered in four scaffolds (*SI Appendix*, Fig. S6).

Gene prediction protocols revealed 49,374 protein-coding genes, 40.68% of which are supported by RNA-Seq data and 23,250 (47.09%) of which had sufficient similarity to database entries to tentatively assign gene function (see *SI Appendix*, Table

Table 1. Overview of assembly and annotation for the *B. hygrometrica* draft genomes

Item	Features
Genome size (predicted and assembled)	1,691 and 1,548 Mb
Assembled in predicted genome	91.52%
No gap sequences in assembled genome	85.86%
Number of scaffolds (>100 bp)	520,969
Total length of scaffolds	1,547,684,042
N50 (scaffolds)	110,988
Longest scaffold	1,434,191
Number of contigs (>100 bp)	659,074
Total length of contigs	1,328,817,553
N50 (contigs)	11,187
Longest of contigs	691,061
GC content	42.30%
Number of predicted gene models	49,374
Mean transcript length (mRNA)	2,535.41
Mean coding sequence length	977.30
Mean number of exons per gene	3.58
Mean exon length	273.12
Mean intron length	604.33
Number of genes annotated	23,250
Number of genes unannotated	47.09%
Number of miRNA genes	196
Mean length of miRNA genes	112.4 bp
miRNA genes share in genome	0.00142%
Number of *rRNA* fragments	1512
Mean length of *rRNA* fragments	101.6 bp
rRNA fragments share in genome	0.00988%
Number of tRNA genes	538
Mean length of tRNA genes	76.2 bp
tRNA genes share in genome	0.00264%
Number of snRNA genes	151
Mean length of snRNA genes	117.0 bp
snRNA genes share in genome	0.00114%
Total size of repeat sequences	1,172,433,882
Repeat sequences share in genome	75.75%
Total size of transposable elements	1,163,296,466
TEs share in genome	75.16%
Total size of tandem repeats	62,678,253
Tandem repeats share in genome	4.05%

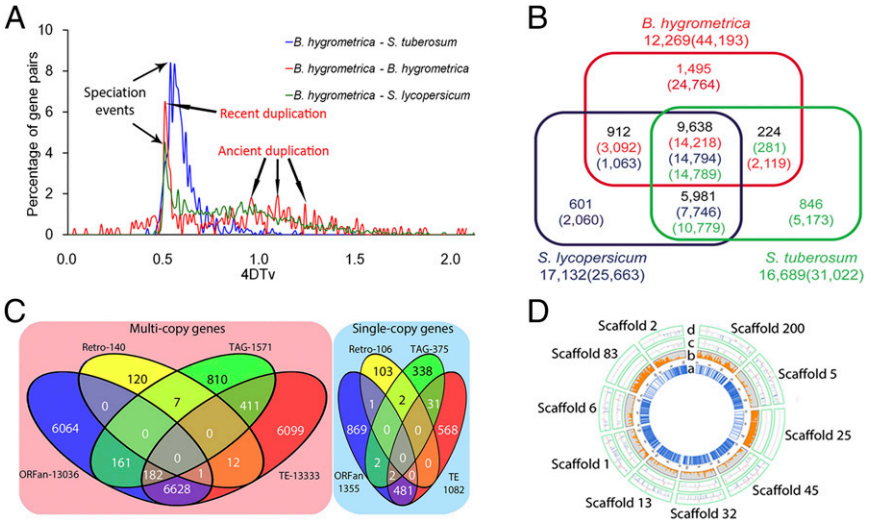

Fig. 2. *B. hygrometrica* genome features. (*A*) Genome duplication in genomes of *B. hygrometrica*, *S. tuberosum*, and *S. lycopersicum*, as revealed through 4DTv analyses. (*B*) A Venn diagram illustrating shared and specific gene families and genes (within brackets) in *B. hygrometrica*, *S. tuberosum*, and *S. lycopersicum*. The gene family and its related number of genes are listed in each of the components. (*C*) A Venn diagram of gene set. (*D*) Profiles integrating genome structures with DEGs of the longest 10 scaffolds. (*a*–*d*) Scaffolds indicating the distribution of ORFs (*a*, in blue), repetitive sequences with DNA II and RNA transposon (*b*, in yellow and orange), and DEG distribution on scaffolds in HD vs. 70% RWC and HD vs. 10% RWC (*c*, pink, accumulating DEGs; *d*, green, declining DEGs).

S8 and *SI Appendix*, *Results* for details). The structural features of the protein-coding gene complements for *B. hygrometrica* were closely comparable to those reported for *S. tuberosum* and *S. lycopersicum* but differed substantially from those reported for *Arabidopsis* (*SI Appendix*, Fig. S7 and Table S8). Of the predicted 12,269 potential gene families, 9,638 (~78.56%), involving 14,218 genes, are shared with *S. tuberosum* and *S. lycopersicum* genomes, reflecting the common origin between Lamiales and Solanales in asterids (Fig. 2*B*).

Predicted genes were functionally annotated by a consensus approach, using InterPro (20), Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) (21), Swissprot, and Translated EMBL Nucleotide Sequence Data Library (TrEMBL) (22). The largest number of genes exhibited homology with proteins in the TrEMBL (46.12%) and InterPro (37.71%) databases (*SI Appendix*, Table S9). In total, 23,250 genes (47.09%) had sufficient similarity to database entries to tentatively assign gene function. Of the annotated protein-coding genes, multicopy genes outnumber single-copy genes by a factor of two (Fig. 2*C* and Dataset S4). Both categories contain an almost equal number of genes contained in TEs and genes classified as orphans [genes that are not a member of a gene family and have no significant sequence similarity to any entry in protein databases outside the taxon of interest (23)]. Up to 97% of the orphan genes originated from duplication events (*SI Appendix*, Table S10).

Of the genes that are historically associated with DT, in the *Boea* genome, only the early light-inducible protein (ELIP) gene family exhibits evidence of expansion. *B. hygrometrica* has seventeen ELIP genes (15 *ELIP1* and two *ELIP2*). One of the Asterid sequenced genomes, the *S. tuberosum* genome, reports a single ELIP gene (15), similar to the pea and tobacco genome (24), and *S. lycopersicum* has two ELIP genes (ELIP1 and ELIP2) (25), similar to *Arabidopsis* and barley.

The Genome and Desiccation Tolerance. To examine the response of the genome to desiccation, and to understand the architecture of its tolerance mechanisms within the genome, we profiled the dehydration-induced alteration of gene expression (Dataset S5). We constructed a genome-wide dehydration response profile by integrating the scaffold protein-coding and repetitive sequence mapping analysis with 9,888 differentially expressed genes (DEGs; identified as greater than twofold change in transcript abundances from that for hydrated controls, at a P value of < 0.05) during drying (Fig. 2*D* and Dataset S5). There was no obvious clustering of DEGs, the majority of which are located, as expected, predominantly in scaffolds that contain few repetitive sequences and that are gene-rich (Dataset S6). The lack of clustering of any significant number of DEGs with their scattered location among a large number of contigs suggests DT was not acquired in a recent evolutionary or restructuring event (sufficient time for dispersal of genes throughout the genome) but, rather, as a retooling of existing genetic elements to deliver the DT phenotype in vegetative tissues.

Gene Expression and Desiccation. The majority of genes expressed in the leaves of *B. hygrometrica* belong to gene families. The large number of orphan genes, ~29% of all annotated genes and 8.51–10.48% of expressed annotated genes, was within the expected range for orphan gene content of eukaryotic genomes (*SI Appendix*, Table S11) (23), of which only a small number (a maximum of 128) were significantly responsive to dehydration (*SI Appendix*, Table S11). Of the 9,888 DEGs, 58.18% responded to moderate dehydration [70% relative water content (RWC)] and 87.47% responded to dehydration to 10% RWC (Fig. 3*A* and Dataset S5). There were 1,239 DEGs that only responded to moderate dehydration (769 increase and 470 decline), and 4,135 specifically responded during desiccation (2,188 increase and 1,947 decline).

The assignment of GO terms for 7,716 DEGs (Dataset S5) focuses on membrane components and organelle structure, biopolymer molecular processes and intermediary metabolism, and metal binding, hydrolytic, and oxidoreductase activities (Fig. 3*B* and *SI Appendix*, Table S12). Enrichment analysis of the 7,758 DEGs with KEGG annotation (Fig. 3*C* and Datasets S5 and S7) revealed that glycerophospholipid metabolism and soluble *N*-ethylmaleimide sensitive fusion attachment protein receptor interactions in vesicular trafficking (both processes involved in membrane maintenance) are favored during dehydration. Dehydration also favored transcripts involved in the pathogen defense system, a common observation for abiotic stress responses, and one often brokered by plant hormones [e.g., abscisic acid (ABA) (26)]. As tissues approach desiccation, transcripts that populate the mRNA surveillance pathway appear and accumulate, indicating a need to remove damaged transcripts from the drying cells. Dehydration also resulted in depletion of transcripts that represent a wide range of metabolic processes (Fig. 3*C*), primarily for pathways involved in growth (photosynthesis and nitrogen metabolism). A more focused clustering of 734 high-level DEGs revealed three major clusters ($\log_2$ base mean value in one sample is more than fourfold higher than that in any other sample; Fig. 3 *D* and *E*, *SI Appendix*, *Results*, and Dataset S8), offering a broad assessment of the response to desiccation and a broad comparison with similar transcriptomes of other resurrection dicots (5).

This and other studies of vegetative dehydration/desiccation transcriptomes (27) point toward a central core of genes and gene products associated with the ability to survive drying: ABA metabolism and signaling, phospholipid signaling, late

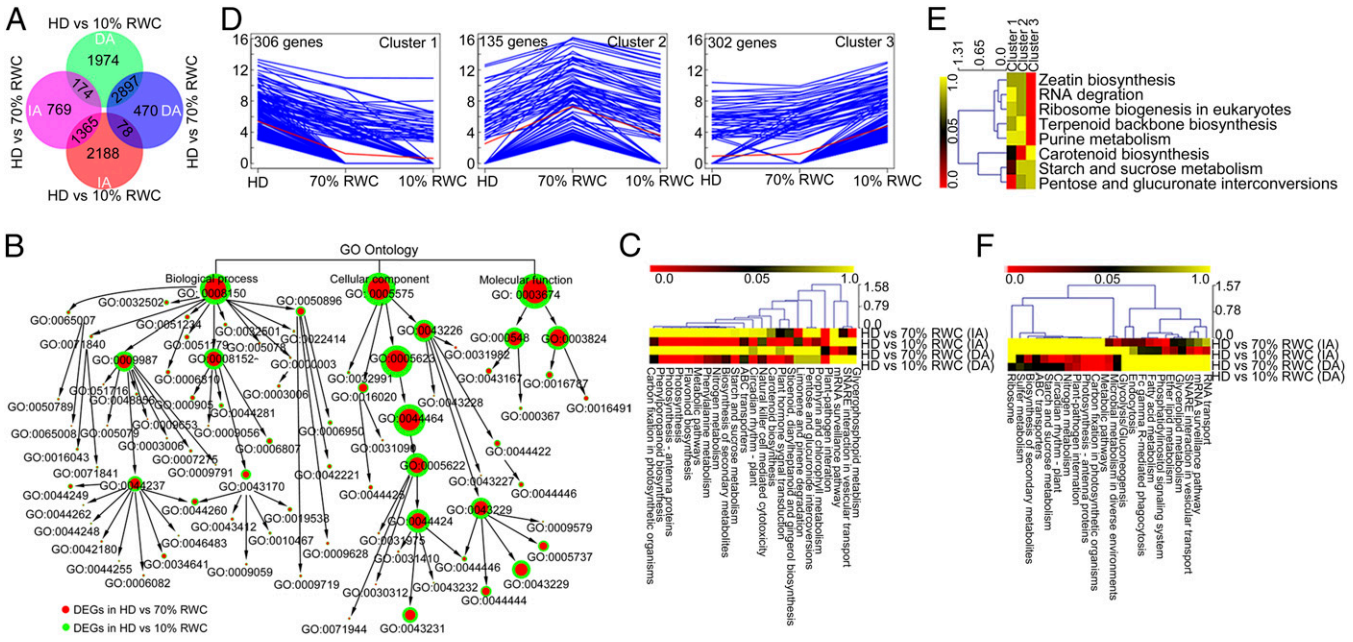

Fig. 3. Transcriptional responses during dehydration. (*A*) Venn diagrams show the number of differentially expressed genes during dehydration and rehydration: hydrated (HD), dehydration (70% RWC), and desiccation (10% RWC). IA, increased abundance; DA, decline in abundance. (*B*) GO classifications of the DEGs respond to dehydration and desiccation. Only GO terms with a gene number larger than 150 are shown. (*C*) Heat maps of significantly enriched pathways in DEGs during dehydration. The yellow and red colors indicate the Q-value for significantly enriched pathways. (*D*) Clusters of high-level ($\log_2$ fold change > 4) DEGs during dehydration. The *y* axis gives the normalized expression level by DESeq software (on a log scale) of DEGs. Each blue line represents a different gene, and the red line indicates the gene expression trend of DEGs in each cluster. (*E*) The heat map describes the significantly enriched pathways in each cluster. The yellow and red color delineates the Q-value for significantly enriched pathways. (*F*) Significantly enriched pathways for those DEGs for which alternative splicing occurred during dehydration. The yellow and red color shows the Q-value for significantly enriched pathways.

embryogenesis abundant proteins (LEAs) (protective proteins), components of reactive oxygen species (ROS) protection and detoxification pathways, and ELIPs (Dataset S9).

Of the 21 DEGs associated with ABA metabolism, eight positive DEGs encode enzymes directly involved in ABA biosynthesis and catabolism, indicating tight control of ABA levels during dehydration (*SI Appendix*, *Results*). A single phospholipase D gene, *PLD-1α*, controlled, in part, desiccation response of the resurrection dicot *Craterostigma plantagineum* (28). This may also be the case for *B. hygrometrica*, as evidenced by the increased abundance of transcripts from one of the two PLD-1α genes during dehydration (Dataset S9). Other *PLDs* (three *PLD-γs*, a *PLD-β*, and a *PLD-P1/Z1*) also responded positively to dehydration, indicating that phospholipid signaling may be more complex in *B. hygrometrica*.

The *B. hygrometrica* genome contains a plethora of LEA protein genes [65 with 51 expressed and 47 DEGs (Dataset S9)], which is a much greater number than reported for the transcriptomes of *C. plantagineum* (27) or *Haberlea rhodopensis* (29). The greater number of expressed LEA genes may reflect the length and severity of the seasonal dehydration periods experienced by *Boea* compared with the other resurrection species (*SI Appendix*, *Results*). The proteins derived from two *LEA1s*, *Bhs4_093* and *Bhs4_094*, have been demonstrated to stabilize the photosynthetic proteins (such as LHCs) in transgenic tobacco seedlings during dehydration and rehydration (30).

The response to dehydration for genes involved in ROS protection and mitigation of oxidative damage is a complex one. Early studies revealed the importance of glutathione metabolism in the dehydration response of *Boea* species (11). Specific members of the GST gene family responded to dehydration stress, along with several peroxidases (Dataset S9), indicative of a need for detoxification and repair of oxidative damage (*SI Appendix*, *Results*).

The increase in abundance of ELIP transcripts is a common feature of the response of DT plants to dehydration (4), as observed in *H. rhodopensis* and *C. plantagineum* (27, 29). Thirteen of the 17 ELIP orthologs in the *B. hygrometrica* genome were ranked as positive DEGs (*SI Appendix*, *Results* and Dataset S9). It thus appears that the protection of photosystem II is a major aspect of the DT mechanism for *B. hygrometrica*.

Relating the RNA-Seq data for dehydrating to the draft genome revealed that 7,127 of the genes represent two or more alternative splicing (AS) products, delivering more functional variation than specified by the annotated gene complement alone (Dataset S10). Of the DEGs, 4,491 (45.42%) exhibited AS during dehydration (*SI Appendix*, Table S13 and Dataset S11). Alternative 5′ splice sites dominated the four major AS patterns. Pathway enrichment of AS-DEGs favored an increase in abundance of transcripts related to endocytosis and Fc gamma R-mediated phagocytosis, fatty acid metabolism, and peroxisomal functions, suggestive of needs for membrane component and protein removal or recycling as cells lose water, as well as an ongoing repair of membranes and removal of ROS. AS was also involved in transcript selections for the processes that were revealed in the overall analysis of DEGs mentioned previously (Fig. 3*F*) (31).

Discussion

Vegetative DT most likely evolved in certain angiosperm lineages from selection pressures exerted by an environment that delivered lengthy periods of little or no soil water. The lack of DT-specific genome organizational features in *B. hygrometrica*, such as clustering of DEGs, supports the contention that vegetative DT evolved primarily from an alteration in the regulation of preexisting genetic modules. This most likely involved those genetic components that deliver developmentally controlled DT to seeds and pollen (32). A portion of that alteration in the regulation of gene expression in *B. hygrometrica* clearly involves AS of transcripts and the plant hormone ABA.

The *B. hygrometrica* genome offers some important insights into the genetic strategies used for accomplishing vegetative DT and its evolution in this resurrection species. The large number of orphan genes housed within the genome, ∼10% of expressed

genes, reflects the somewhat unique nature of this resurrection species. Orphan genes are thought to represent lineage-specific adaptations and, in some plant species, to be linked to stress responses (e.g., rice) (33). This may also be true for the expressed orphan genes of *B. hygrometrica*, but only a small number (128) can, at this point, be associated with the resurrection phenotype and probably represent species-specific aspects of the DT mechanism.

The apparent expansion of 5S *rRNA* genes in the *Boea* lineage may reflect the need for a supply of active ribosomes during the rapid resumption of protein synthesis (and recovery) on rehydration. Because ribosomal 5S *rRNA* transcripts can only be amplified by transcription, it would seem reasonable to suggest the 5S *rRNA* gene expansion in *B. hygrometrica* evolved to meet the protein synthesis burden inherent in the resurrection phenotype. As this is the first resurrection genome, to our knowledge, to be sequenced, it remains to be seen whether this is a common genotypic feature of resurrection species.

The genome sequence and transcriptome also revealed an expansion of the ELIP gene family in *B. hygrometrica* concomitant with enhanced transcript abundance for 13 of the 17 gene family members. ELIP proteins are postulated to protect the photosynthesis machinery from photooxidative damage by preventing the accumulation of free chlorophyll by binding pigments and preserving the chlorophyll-protein complexes (34). ELIP proteins (and transcripts) have been reported to increase in abundance in a linear fashion with the amount of photoactivation and photodamage to the photosystem II reaction centers, D1 protein degradation, and changes in pigment level (24). Photooxidative damage is a primary stressor for resurrection species, as they spend a considerable amount of time in the dried state and under high-light conditions (35). Thus, it appears that *B. hygrometrica* has evolved a strategy of ELIP gene expansion to aid in its ability to protect its photosynthetic apparatus, particularly photosystem II, from oxidative damage: an essential and perhaps central aspect of its DT mechanism. The transcriptomic analysis provides a broader perspective on the nature of the cellular protection aspects of vegetative DT, highlighted by the increase in transcript abundance for LEA protein genes, GST gene family, and peroxidases.

The draft genome offers a unique opportunity to construct a systems approach to understanding the mechanistic aspects of DT and resurrection in plants. Such an approach can help influence our understanding of the evolution of the land plants and our attempts to design strategies for the improvement of the dehydration tolerance of our major crops as food security issues increase in importance globally.

Materials and Methods

The original accessions for *B. hygrometrica* were collected from a dry rock crack in Fragrant Hills in a Beijing suburb in China. The genome was sequenced using the whole-genome shotgun approach, using Illumina HiSeq and Roche 454 platforms. Whole-genome shotgun data were used to assemble the draft genome, using the hybrid assembly strategy by Newbler, SSPACE, and SOAP de novo algorithm. Genes were annotated using a combined approach on the repeat masked genome with ab initio gene predictions, protein similarity, and transcripts to build optimal gene models. Repeat sequences were identified by both de novo approach and sequence similarity at the nucleotide and protein levels. Detailed information of materials, methods, and any associated references are available in the *SI Appendix, Materials and Methods*.

ACKNOWLEDGMENTS. We thank the Beijing Genetics Institute staff members and Capital Normal University graduate students for their assistances on genome sequencing, assembling, and bioinformatic analyses. This study was supported by funds from the Chinese Ministry of Agriculture (2014ZX08009-23B, 2009ZX08009-058B), Chinese Ministry of Science and Technology (2007AA021405), and the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (Y. He). We are also thankful for the special financial support from the National Key Disciplines of China for this project.

1. Oliver M-J, Tuba Z, Mishler B-D (2000) The evolution of vegetative desiccation tolerance in land plants. *Plant Ecol* 151(1):85–100.
2. Proctor M-C-F, Pence V-C (2002) Vegetative tissues: Bryophytes, vascular resurrection plants, and vegetative propogules. *Desiccation and survival in plants: Drying without dying*, eds Black M, Pritchard H-W (CABI, Wallingford, Oxon), pp 207–267.
3. Moore J-P, Le N-T, Brandt W-F, Driouich A, Farrant J-M (2009) Towards a systems-based understanding of plant desiccation tolerance. *Trends Plant Sci* 14(2):110–117.
4. Oliver MJ, Cushman JC, Koster KL (2010) Dehydration tolerance in plants. *Methods Mol Biol* 639:3–24.
5. Gechev T-S, Dinakar C, Benina M, Toneva V, Bartels D (2012) Molecular mechanisms of desiccation tolerance in resurrection plants. *Cell Mol Life Sci* 69(19):3175–3186.
6. Oliver M-J, et al. (2011) A sister group contrast using untargeted global metabolomic analysis delineates the biochemical regulation underlying desiccation tolerance in Sporobolus stapfianus. *Plant Cell* 23(4):1231–1248.
7. Gerten D, et al. (2011) Global water availability and requirements for future food production. *J Hydrometeorol* 12(5):885–900.
8. Cushman JC, Oliver MJ (2011) Understanding vegetative desiccation tolerance using integrated functional genomics approaches within a comparative evolutionary framework. *Plant desiccation tolerance in plants,* eds. Luttge U, Beck E, Bartels D (Springer: Heidelberg), pp 307–338.
9. Deng X, Hu Z, Wang H (1999) mRNA differential display visualized by silver staining tested on gene expression in resurrection plant *Boea hygrometrica*. *Plant Mol Biol Rep* 17(3):279–279.
10. Wilson C-L (1974) Floral anatomy in Gesneriaceae. I. Cyrtandroideae. Bot Gaz 135:247–268.
11. Jiang G, et al. (2007) Proteome analysis of leaves from the resurrection plant *Boea hygrometrica* in response to dehydration and rehydration. *Planta* 225(6):1405–1420.
12. Vicré M, Farrant J-M, Driouich A (2004) Insights into the cellular mechanisms of desiccation tolerance among angiosperm resurrection plant species. *Plant Cell Environ* 27(11):1329–1340.
13. Matassi G, Montero L-M, Salinas J, Bernardi G (1989) The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. *Nucleic Acids Res* 17(13):5273–5290.
14. Flavell R-B, Bennett M-D, Smith J-B, Smith D-B (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem Genet* 12(4):257–269.
15. Xu X, et al.; Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475(7355):189–195.
16. Hilbricht T, et al. (2008) Retrotransposons and siRNA have a role in the evolution of desiccation tolerance leading to resurrection of the plant *Craterostigma plantagineum*. *New Phytol* 179(3):877–887.
17. Zhao Y, et al. (2014) Identification of a retroelement from the resurrection plant *Boea hygrometrica* that confers osmotic and alkaline tolerance in *Arabidopsis thaliana*. *PLoS ONE* 9(5):e98098.
18. Gorman SW, Teasdale RD, Cullis CA (1992) Structure and organization of the 5S rRNA genes (5S DNA) in *Pinus radiata* (*Pinaceae*). *Plant Syst Evol* 183(3-4):223–234.
19. Kobayashi T (2011) Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cell Mol Life Sci* 68(8):1395–1403.
20. Mulder N-J, et al. (2007) New developments in the InterPro database. *Nucleic Acids Res* 35(Database issue):D224–D228.
21. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25(1):25–29.
22. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30.
23. Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. *Nat Rev Genet* 12(10):692–702.
24. Yurina NP, Mokerova DV, Odintsova MS (2013) Light-inducible stress plastid proteins of phototrophs. *Russ J Plant Physiol* 60(5):577–588.
25. Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635–641.
26. Lee S-C, Luan S (2012) ABA signal transduction at the crossroad of biotic and abiotic stress responses. *Plant Cell Environ* 35(1):53–60.
27. Dinakar C, Bartels D (2013) Desiccation tolerance in resurrection plants: New insights from transcriptiome, proteome, and metabolome analysis. *Front Plant Sci* 4(482):1–14.
28. Frank W, Munnik T, Kerkmann K, Salamini F, Bartels D (2000) Water deficit triggers phospholipase D activity in the resurrection plant *Craterostigma plantagineum*. *Plant Cell* 12(1):111–124.
29. Gechev T-S, et al. (2013) Molecular mechanisms of desiccation tolerance in the resurrection glacial relic *Haberlea rhodopensis*. *Cell Mol Life Sci* 70(4):689–709.
30. Liu X, et al. (2009) LEA 4 group genes from the resurrection plant *Boea hygrometrica* confer dehydration tolerance in transgenic tobacco. *Plant Sci* 176(1):90–98.
31. The Angiosperm Phylogeny Group (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot J Linn Soc* 141(4):399–436.
32. Galf DF, Oliver MJ (2013) The evolution of desiccation tolerance in angiosperm plants: A rare yet common phenomenon. *Funct Plant Biol* 40(4):315–328.
33. Guo W-J, Li P, Ling J, Ye S-P (2007) Significant comparative characteristics between orphan and nonorphan genes in the rice (*Oryza sativa L.*) genome. *Comp Funct Genomics* 2007:21676.
34. Alamillo J-M, Bartels D (2001) Effects of desiccation on photosynthesis pigments and the ELIP-like dsp 22 protein complexes in the resurrection plant Craterostigma plantagineum. *Plant Sci* 160(6):1161–1170.
35. Farrant J-M, Willigen V-C, Lofell D-A, Bartsch S, Whittaker A (2003) An investigation into the role of light during desiccation of three angiosperm resurrection plants. *Plant Cell Environ* 26(8):1275–1286.

SUPPLEMENTAL INFORMATION

Xiao et al.: The resurrection genome of *Boea hygrometrica*: a blueprint for survival of dehydration

Supplemental Materials and Methods

Plant materials

The desiccation-tolerant homoiochlorophyllous resurrection plant *Boea hygrometrica* (Bunge) R. Br (Gesneriaceae), a dry habitat extremophile, is a perennial, out-breeding, rosette-forming herb found above 500 m in elevation predominantly in Central, South, East and Southeast China (www.efloras.org). Although it as an outcrossing species, *B. hygrometrica* exhibits a relatively low level of heterozygosity indicating either that it is a facultative self pollinator or there is low genetic variability within the population of the region where it was collected. To establish a genome sequence and maintain the genetic source we chose to isolate and establish a line, derived from a single seed collected by Dr. Lihong Xiao in the Fragrant Hills suburb of Beijing, China. The line was vegetatively cloned at least three times and it is these clones that serve as the original source of DNA. Clones are maintained for future genomic studies. For the genomic aspects of this study, thirty-day old seedlings from this line were divided into three sub-populations: one grown in the dark for DNA extraction for sequencing, a second transplanted into soil-filled pots and grown in a greenhouse under conditions of 16/8h light/dark, 25ºC and 70% humidity for dehydration and desiccation treatments, and a third used for successive subcultures and strain maintenance.

Cytological and physiological experiments

Microscopic and ultrastructural studies: Three-month-old plants, grown in soil-filled pots, were subjected to a drying event by withholding water under conditions of 16/8h light/dark, 25°C and 30% humidity in a growth chamber. For scanning electron microscopy (SEM), the hydrated and fully rehydrated samples were fixed in 3% glutaraldehyde in 0.1 M phosphate buffer (pH 7.4). Samples were dehydrated using a graded ethanol series, critical-point dried with liquid carbon dioxide, mounted on aluminum stubs, sputter coated with gold palladium (1), and analyzed using a Hitachi S – 4800 scanning electron microscope (Hitachi, http://www.hitachi.com/). To avoid the cell wall expansion during aqueous fixation, desiccated samples were directly subjected to the critical-point drying without fixation or ethanol treatments.

Samples for light microscopy and transmission electron microscopy (TEM) were processed as follows: sliced tissues were fixed in 3% gluteraldehyde in 0.1 M phosphate buffer (pH 7.4) containing 0.5% caffeine under vacuum for approximately 10 min until the sample ceased to float. Samples were left in the fixative for at least 2 hours at room temperature or overnight at 4°C and then post-fixed (for contrast and to avoid tissue expansion or shrinkage) in a 1% osmium tetroxide solution in phosphate buffer for 2 hours. After passage through a graded ethanol series for dehydration, the material was infiltrated with and embedded in epoxy resin. Sections for analysis were obtained using a Leica EM UC6 microtome (Leica, http://www.leica-microsystems.com/). For cellular organization analysis, sections were stained with 1% toluidine blue O (TBO) for five minutes before examination using a Leica DMRE2 microscope. For ultrastructural observations, sections were further stained with 1% uranyl acetate and 1% lead citrate1 and examined using a Hitachi 7500 transmission electron microscope (Hitachi, http://www.hitachi.com/).

All images were processed for publication using Adobe Photoshop CS5 (Adobe Systems).

Dehydration, desiccation treatment and the measurement of RWC: To establish a drying curve, five groups of six mature healthy leaves per group were randomly selected every 12 hours from 30 three-month-old plants to determine the representative relative water content (rRWC) of the population. The rRWC was calculated according to formula: RWC% = (FW – DW)/(FTW – DW) • 100%, where FW was the fresh weight, FT was the weight at full turgor, and DW was the weight of the same sample dried at 65°C for 12 hours. The full turgor weight of the sample was achieved by submersion in deionized water overnight at 4°C in the dark. For experimental samples, leaves were collected from individual plants, at the same time of day for well-watered (WW), 70% RWC and dried (10%RWC) treatments. Leaves were ground to a fine powder in liquid nitrogen and stored frozen at -80°C for transcriptome and methylome analyses.

Genome sequencing and assembly

High-molecular-weight DNA preparation: High-quality genomic DNA for *de novo* genome sequencing was prepared from 30-day-old axenic etiolated seedling tissues, grown in the dark for 2 weeks before collection to simplify extraction and minimize chloroplast DNA (cpDNA) contamination. DNA was extracted using a

phenol/chloroform method (2) and treated with RNase A and proteinase K, to reduce RNA and protein contamination, respectively and further precipitated in 95% ethanol and rinsed in 75% ethanol.

Illumina library construction and sequencing: Short paired-end (PE) insert DNA libraries, with insert sizes of from 170 bp to 800 bp, were prepared following the manufacturer's standard protocol (Illumina, San Diego, CA). For long (2 - 40 Kbp) mated-pair libraries, we used the Illumina's mate pair library kit, which included several steps of DNA circularization, digestion of linear DNA, fragmentation of circularized DNA, and purification of biotinylated DNA fragments prior to adapter ligation. After library preparation and quality control of DNA samples, template DNA fragments were hybridized to the surface of flow cells on an Illumina HiSeq™ 2000 sequencer, isothermally amplified to form clusters, and sequenced following the standard manufacture's protocols (Illumina).

454 pyrosequencing library construction and sequencing: For Roche 454 GS FLX and GS FLX+ sequencing, 600 bp and 1,000 bp shotgun libraries were prepared by using protocols provided by the manufacturer (Roche Applied Science, Mannheim, Germany). In brief, quantified DNA fragments were polished to create blunt ends for adaptor ligation and a single A overhang added to the ends of the DNA fragments. Adaptors containing fluorescent molecules were ligated onto the polished fragments. Sequencing was performed following the recommendations of the manufacturer (Roche Applied Science, Mannheim, Germany).

Filtering processes of raw data: To reduce the impact of sequencing errors and sample contamination on the genome assembly, we subjected the Illumina HiSeq 2000and Roche 454 raw data to a stringent filtering process. The raw reads generated from the Illumina pipeline contain contaminating reads from chloroplast and mitochondrial DNA as well as artificial reads generated by base-calling duplicates and adapter contamination. To remove these contaminants we first aligned the raw reads to the chloroplast and plant mitochondrial sequences deposited in the NCBI database using the SOAP 2.21 software. Reads with significant homology to organellar sequences in the NCBI database were discarded. Artificial reads were removed as described for the panda genome (3) sequence project. Raw reads generated from 454 sequencing were similarly filtered to remove organellar sequence contamination using Newbler 2.6 software set to default parameters (Roche, http://www.roche.com/). After filtration we retained a total of 474.36 Gb high quality filtered sequence, of which, 458.74 Gb resulted from Illumina and 15.62 Gb from 454 sequencing.

Estimation of the Genome Size with 17-mer Analysis: A K-mer refers to an artificial sequence division of K nucleotides. A raw sequencing read with L bp contains (L – K + 1) K-mers if the length of each K-mer is K bp. The frequency of each K-mer can be calculated from the raw genome sequencing reads. The K-mer frequencies along the sequencing depth gradient follow a Poisson distribution in a given data set. During deduction, the genome size G = K_num/peak_depth, where the K_num is the total number of K-mer, and Peak_depth is the expected value of K-mer depth. Typically, K = 17. Lower quality reads were filtered and removed prior to 17-mer frequency assessments.

Draft genome assembly: WGS data from three platforms, Illumina HiSeq™ 2000, Roche 454 GS FLX, and Roche 454 GS FLX+, were used to assemble the *B. hygrometrica* genome using the hybrid assembly strategy by Newbler, SSPACE (4) and SOAP *de novo* algorithms (5). The filtered 454 reads were first used to construct contigs with Newbler 2.6 software. Long mate-paired reads were used step by step to link the contigs corrected by Illumina small paired-end fragment reads to scaffolds with SSPACE software. To fill gaps inside constructed scaffolds, the majority of which were composed of repeats masked during the scaffold construction, we used the Illumina small paired-end fragments and 454 data to retrieve read pairs that had one read well-aligned on the contigs and another read located in the gap region, and then conducted a local assembly for the collected reads with SOAPdenovo software.

Genome assembly quality assessments

Fosmid sequences versus draft genome comparison: To evaluate the quality of the assembled genome, 36 fosmids with insert size of 27.8 – 43.5 kb from randomly selected genomic DNA regions were sequenced to a minimum of six-fold coverage using Sanger shotgun sequencing with an ABI3730xl DNA Analyzer (Applied Biosystems, USA, www.appliedbiosystems.com). Comparisons between sequenced fosmids and their equivalent scaffold regions (including inserted N-gaps) were conducted based on the BLASTN (cutoff of identity 0.05) method (3) to check the coverage rate.

EST library construction and end sequencing and sequence comparisons: To evaluate the fidelity of the assembly of gene containing regions in the draft genome, EST libraries were constructed and end sequenced using a standard Sanger sequencing protocol. The filtered and cleaned sequences were mapped to the draft genome using

Blat (6) with cutoff for identity of 0.9 and N included in scaffolds to validate the coverage of the gene containing regions.

Genome annotation

Annotation of DNA repeat sequences: Repeat sequences were identified by both *de novo* approach (7) and sequence similarity at the nucleotide and protein level (8). Transposable elements were identified at both the DNA and protein levels, based on known sequences contained within the DNA repeat database (9), using both RepeatMasker 3.3.0 (8) and RepeatProteinMask 3.3.0 (the same package with RepeatMasker) software respectively. We used the *de novo* prediction programs RepeatModeler 1.0.5 and LTR-FINDER 1.0.5 (10) to build a *de novo* repeat library based on the sequenced genome. Contaminating sequences and multi-copy genes in the library were removed. LTR-FINDER was used to search the whole genome for the characteristic structure of full-length long terminal repeat retrotransposons (LTR). Using the generated *de novo* library as a database, RepeatMasker was used to find and classify repeat sequences in the genome.

Gene prediction and function annotation: Genes were annotated using a combined approach on the repeat masked genome with *ab initio* gene predictions, protein similarity and transcripts to build optimal gene models.The *de novo* genepredictionwas performed on the repeat-masked genome using hidden Markov model (HMM) based Augustus (11) and Genscan (12) software with parameters trained for *A. thaliana*. For homology-based gene prediction, protein sequences of six different species (*Arabidopsis thaliana*, *Carica papaya*, *Cucamis sativus*, *Fragaria vesca*, *Glycine max* and *Vitus vinifera*) were mapped onto the genome using TblastN with an E-value cutoff 1×10-5, the aligned sequences as well as their corresponding query proteins were then filtered and passed to GeneWise (13) to search for accurate spliced alignments. Source evidences generated from the three approaches were integrated by GLEAN (version 1.1) (14) to produce a consensus gene set.

A combination of RNA sequencing (RNA-Seq) and GLEAN based analysis was employed to improve the integrity and fidelity of the gene predictions. Transcriptomic clean reads from five samples, representing hydrated, dehydrating and desiccation conditions, with three biological replicates for each sample, were mixed and aligned to the assembled genome using TopHat (15) software to identify candidate exon regions and the donor and acceptor sites of introns. Mismatches of no more than 2 bases were allowed in the alignment. The Cufflinks (16) protocol, using default parameters, was performed to assemble the alignments into transcripts. Based on these assembled potential transcript sequences, open reading frames (ORFs) were predicted using the HMM-based training parameters, to obtain reliable transcript predictions. These transcript predictions were combined with those from the GLEAN analysis to generate a gene set with a greater degree of confidence. The final gene set contains 49,374 predicted genes, all of which were retained for further analysis.

Gene functions were assigned according to the best match of the alignments using BlastP to the SwissProt/TrEMBL databases (http://www.uniprot.org/). The motifs and gene domains were assigned by an InterProScan (17) comparison against all available protein databases, e.g., ProDom, PRINTS, Pfam, SMART, PANTHER and PROSITE. Gene Ontology (18) IDs for each gene were obtained from the corresponding InterPro generated entries. All genes were aligned against KEGG (19) proteins, and the metabolic pathway predictions were derived from matched genes in the KEGG database.

Identification of non-coding RNA genes: The tRNAscan-SE (20) aligorithms, set with eukaryote parameters, were used to identify tRNA positions. The snRNA and miRNA sequences were predicted using a two-step method: alignment with Blast followed byan INFERNAL (http://infernal.janelia.org/) search against the Rfam database (Release 9.1) (21). The rRNAs were annotated by aligning the BlastN data, with E-value 1×10-5, against a ref rRNA sequence from *B. hygrometrica* or a closely related species.

Identification of ORFan, tandem repeat, and TE-contained genes: ORFan genes were identified using a BLAST filtering approach (BLASTP, e-value < 0.01), following the method used for pigeonpea ORFan gene predition (22). To identify tandem repeat genes (TAGs), paralogous genes were identified by BLASTP (Identity > 40%, E-value < 1e-15, Match length> 100) and a tandem duplication event was defined as a genome region in which at least two paralogous genes occur in one location (separated by no more than one other gene) (23-24).

Identification of retrogenes: All protein sequences that were used as queries for searching the genome were identified by TblastN (25). The exons that generated a high score for paired sequences (Hsps) within the TblastN data were linked using a dynamic algorithm. A gene as defined as homologous when the query sequence had greater than 70% homology within a homologous chain contained in the genome combined with a greater than 50% identity/sequence similarity. Candidate genes were selected that had less than a 40bp gap or intron within the homologous genes as generated by GeneWise (13). Alignments were constructed between candidate genes

and protein sequence as determined by FASTA (26). The sequences were retained when the alignment length was longer than 40 amino acids and primary protein sequence similarity was greater than 40%. A gene was defined as a retrocopy when the best-aligned protein with a candidate gene contained at least one 70-base intron and more than one exon per gene. Retrocopies were divided into intact retrocopies and retropseudogenes according to the existence or absence of a frame shift and early termination codon, compared to their parent genes. Ka (nonsynonymous substitution), Ks (synonymous substitution) and Ka/Ks between the aligned retrocopy and its parent gene were calculated according to a method described by Li and Pamilo and Bianchi (27-28), using the KaKs calculator 1.2 (29) software. Functional retrogenes were determined when Ka/Ks significantly less 0.5 ($p < 0.5$) in an intact retrocopy by using codeml program in PAML4 (30).

Comparative genome analyses

All-versus-all BLASTP (E-value less than1×10-5) was used to detect orthologous or paralogous genes between *A. thaliana, B. hygrometrica, S. tuberosum* and *S. lycopersicum*. An orthologous gene was defined as a reciprocal BLASTP hit between species. Syntenic blocks (>5 genes per block) were identified using MCscan (31) (-a, -e:1e-5, -u:1, -s:5). Long blocks were chosen for illustration using Circos (http://circos.ca/) (32-33). To show relative block size, the Ribbon option of Circos50 was used to draw thick lines, which at the start and end points have a thickness that directly corresponds to the size of the duplicated block.

4DTv (fourfold degenerate synonymous sites of the third codon) distribution is used to indicate the likelihood of a whole genome duplication (WGD) event. If a WGD event had occurred, 4DTv is also used to confirm speciation before or after the WGD. In the intraspecies alignment each aligned block represents paralogous segment pairs that arose from the genome duplication whereas, in the interspecies alignment each aligned block represents the orthologous pair derived from the shared ancestor. We calculated the 4DTv for each gene pair from the aligned block to generate a distribution for the 4DTv values to estimate the speciation or WGD event that occurred during the evolutionary history of the plant.

Transcriptome analysis during dehydration

RNA-Seq and identification of differentially expressed genes (DEGs): The RNA samples for transcriptome analyses were collected from adult leaf tissues that grew in soil-filled pot with or without dehydration treatment. RNA was isolated from the leaf tissues with three biological replicates for each of the well watered, dehydration and desiccation treatments. Oligo (dT) magnetic beads were used to enrich for mRNAs and cDNA libraries were prepared for Illumina HiSeqTM 2000 sequencing platform in the single-end (SE) mode. High quality filtered reads were mapped to the draft reference genome version 1.0 with SOAP aligner (Soap2.21) (30) (mismatches >2 bases). If there was more than one transcript for a single gene, the longest was used to calculate expression level and coverage. Gene expression was normalized (BaseMean) for each sample and differentially expressed genes (DEGs) were identified by DESeq (34) for each compared group by using "P-adj (adjusted p value) < 0.05 and the |log2Ratio|> 1" as the threshold.

GO and KEGG pathway enrichment: To obtain the significantly enriched GO term for DEGs, all DEGs were mapped to GO terms in the GO database (http://www.geneontology.org/) and the gene numbers for every term were calculated. The significantly enriched GO terms were selected using a hypergeometric test to develop hierarchical clusters of a sample tree by Euclidean Distance. The color scale limits were set as: Red shows Q = 0, Black is Q = 0.05, Yellow is Q ≥ 1.0.

To further clarify the biological functions of DEGs, a pathway-based analysis was conducted using the public pathway-related database (35). Main biochemical pathways and signal transduction pathways with Qvalue <0.05 were considered as significantly enriched in DEGs. We first select he significant pathways based on the hypergeometric distribution of Q value (< 0.05), and hierarchical clustering by using was as described for the GO enrichment.

Expression pattern analysis: Differentially expressed genes (DEGs) with similar expression patterns can indicate a functional correlation. DEGs, that had an expression level (BaseMean) higher than fourfold above the control in any treatment, were used to perform a clustering using the MEV (36) software. Each abscissa denotes an experimental condition, and the value of y-coordinate corresponds to the log2baseMean (Fig. 4).

Supplemental results on genome features

De novo transposable element (TE) annotation indicated that long terminal repeat retrotransposons (LTRs) occupied 72.99% of the assembled genome (*SI Appendix*, Table S6), and only 18.44% of the assembled genome was annotated LTRs in Repbase (*SI Appendix*, Table S14). The subcategories of *gypsy* were the most abundant LTRs, second to *copia*.

Structurally, 348 syntenic blocks, distributed throughout 113 scaffolds encompassing 2,420 syntenic genes (6.9 genes/block), on a sequence basis, representing approximately 5% of the predicted genes are contained in regions of conserved local gene arrangements (microsynteny) (*SI Appendix*, Table S15). We identified 560 or 3,951 syntenic blocks between the *B. hygrometrica* scaffolds and the genomes of *Solanum tuberosum* or *Solanum lycopersicum*, the first sequenced genomes in the Asterids (euasterids I) (37-40), respectively. These syntenic blocks encompass 5,568 (*B. hygrometrica* vs *S. tuberosum*) or 29,655 (*B. hygrometrica* vs *S. lycopersicum*) *B. hygrometrica* genes (*SI Appendix*, Table S16; Dataset S12). Microsyntenic profiles, established using all the syntenic gene block-containing scaffolds within *B. hygrometrica* as well as the 30 longest scaffolds of *B. hygrometrica* aligned with *S. tuberosum* or *S. lycopersicum* scaffolds indicated that the paired syntenic regions within and between species were distributed in both repetitive-poor and gene-clustered regions in the draft *B. hygrometrica* genome (*SI Appendix*, Fig. S8).

A combination of *de novo* gene prediction protocols and homology-based methods defined the gene complement of *B. hygrometrica* to consist of 48,915 unique protein-coding genes (*SI Appendix*, Table S6). To assist annotation and address associated biological questions, we utilized 8.64 MB of RNA-Seq data from independent libraries representing several stages in a desiccation-rehydration cycle. The independent assembly (41-42) of RNA-Seq data generated 20,087 unique transcripts that led to the prediction of a protein-coding gene complement of 49,374 for the *B. hygrometrica* genome (Version1.0). GO enrichment analysis indicated that the annotation of *B. hygrometrica* genes had a similar distribution to that of *S. tuberosum* and *S. lycopersicum* (*SI Appendix*, Fig. S9).

Supplemental results on transcriptome and desiccation

Of the 9,888 DEGs, the more focused cluster analysis of 734 high level DEGs (> 4 Fold), revealed three major clusters (Fig. 3*D*; and Dataset S8). Cluster 1, transcript accumulation only occurred in the hydrated tissues, primarily encoding proteins associated with photosynthesis. Cluster 2, transcripts that accumulated under moderate stress and then depleted, primarily of the carotenoid biosynthesis pathway including both antioxidant production and abscisic acid (ABA) biosynthesis. Cluster 3, transcripts that accumulated as leaves desiccate, primarily encoding proteins of nucleic acid metabolism including RNA degradation, purine metabolism, zeatin phytohormones (cytokinins) metabolism and terpenoid biosynthesis. The cluster analysis offers a broad assessment of the response to desiccation and a broad comparison to similar transcriptomes of other resurrection dicots (43-44).

We analyzed, in detail, the expression patterns of a central core of genes and gene products associated with the ability to survive drying: including ABA metabolism and signaling, Late Embryogenesis Abundant proteins (LEAs) (protective proteins) and components of ROS protection and detoxification pathways (Dataset S9).

There were 26 genes in the genome associated with ABA metabolism; 22 were expressed and 21 of those were DEGs during dehydration (Dataset S9). Four of the eight positive DEGs encode enzymes directly involved in ABA biosynthesis, primarily three putative 9-cis-epoxycarotenoid dioxygenase (NCED) genes: two *NCED3* and one *NCED4* homolog. Three of these genes appeared to be activated during moderate dehydration, and perhaps specific to the early response, as transcripts were barely detectable in the hydrated tissues. Other *NCED* members, *NCED1* (2 genes) and *NCED4* (2 genes) were negative DEGs and thus are specific to ABA metabolism during normal growth. The remaining four positive DEGs represent *CYP707A* genes that encode ABA 8'-hydroxylases, the primary enzyme for ABA catabolism, indicating tight control of ABA levels during dehydration. Two of the *CYP707A* and one of two *CYP707A1* genes expressed in the hydrated state, were activated by moderate dehydration and accumulated transcripts following severe dehydration. The remaining *CYP707A1* and *CYP797A4* and *CYP707A2*, only accumulated transcripts during drying.

Of the eight genes encoding core elements of the ABA receptor complex (*PYLs*: Dataset S9), seven were classified as DEGs. However only one of the *PYLs*, a *PYL9*, accumulated transcript in response to both moderate and severe dehydration suggesting that there is sufficient receptor available to mediate the ABA signaling pathway during dehydration. A single *PYL5* was a positive DEG only under severe dehydration but transcript abundance was so low that it may not be biologically relevant. Of greater significance is that 10 of the 11 expressed group A protein phosphatases type 2C (PP2C) genes are positive DEGs, all 10 under moderate dehydration and 8 during desiccation. Transcripts for three of the *PP2Cs* were only present in dehydrating tissues. However, this result is somewhat enigmatic as type 2C protein phosphatases are known as negative regulators

of ABA signaling (45). In the presence of ABA, the PYLs interact with and inhibit the PP2Cs, thus relieving the protein kinase SnRK2s from inhibition to phosphorylate downstream effectors (46). There were five *SnRK2s* in *B. hygrometrica*, four were expressed and classified as DEGs.

Of the 47 LEA-DEGs (Dataset S9), twenty-nine exhibited an increase in abundance during dehydration and almost half of these were *LEA2*s, *LEA1s* or *Dehydrins*. Several of the *LEA* DEGs had barely detectable transcript abundance under hydrated conditions and so their accumulation appeared to be dehydration specific, of note are the transcripts of two *LEA1s* (*Bhs222_060*, *Bhs4_093*) and one *LEA2* (*Bhs31748_001*), that accumulate to very high levels. The most abundant *LEA* transcripts during dehydration encoded 5 LEA1s, 5 LEA4s, one LEA2, and one Dehydrin, indicating their importance in the dehydration response. Transcripts of the Dehydrin gene, *Bhs1119_057*, were highly abundant under all conditions.

Specific members of the 52-member glutathione-S-transferase (GST) gene family responded to dehydration stress, along with several peroxidases, indicative of a need for detoxification and repair of oxidative damage. Transcripts of the responsive GST gene, *Bhs63_020V1.1*, encoding a Phi group GST was highly abundant under hydrated conditions and accumulated during desiccation, suggesting that the maintenance of the redox state of the target for this GST is relatively important to the cell. Two of the 7 expressed superoxide dismutase (SOD) genes, *Bhs7173_001* and *Bhs109_028*, also responded to dehydration by accumulating transcripts from relatively high levels in the hydrated tissues, presumably to combat a buildup of hydrogen peroxide in the cells under both normal and stressful conditions. In concordance with early findings of the importance of glutathione metabolism in the dehydration response of *Boea* species (47-48), we observed a significant increase in transcript abundance for several of the members of gene families that encode the enzymes of these pathways. The genome encodes as many as 85 peroxidases (PODs), but only half of them were expressed in leaves during dehydration (and *SI Appendix*, Table S20). Several PODs are represented as high abundance transcripts under both hydrated and dehydrating conditions. Of the eight PODs that are classified as positive DEGs, two (*Bhs211_058* and *Bhs4_048*) putative glutathione peroxidases (GPX) responded relatively dramatically to moderate dehydration indicating a rapid need to reduce hydrogen peroxide or organic hydroperoxides early in the dehydration process.

Of the AS-DEGs (4,491), Alternative 5'- splice sites (A5SS) dominated the four major alternative splicing patterns, followed by the alternative 3' - splice site (AS3SS) category (and *SI Appendix*, Table S13). Pathway enrichment analyses of AS-DEGs were evident in the overall analysis of DEGs (Fig. 4*F*) and the GO analysis of those DEGs identified as targets for AS (and *SI Appendix*, Fig. S10; Dataset S13).

Accession code

The genome data generated by the whole project are available in GenBank of National Center for Biotechnology Information as Bioproject ID PRJNA182117. The RNA-Seq data are available in the GEO datasets under accession number GSE48671.

Supplemental References

1. Reynolds ES (1963) The use of lead citrate at high pH as electron opaque stain in electron microscopy. *J. Cell. Biol.* 17: 208–212.
2. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF (2006) A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nature Protocols* 1: 2320–2325.
3. Li R, Fan W, Tian G, Zhu H, He L et al. (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463: 311–317.
4. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27: 578–579.
5. Li R, Zhu H, Ruan J, Qian W, Fang X et al. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 20: 265–272.
6. Kent WJ (2002) BLAT – The BLAST–Like Alignment Tool. Genome Res. 12: 656–664.
7. Price AL, Jones NC, Pevzner PA (2005) *De novo* identification of repeat families in large genomes. *Bioinformatics* 21: i351–i358.
8. Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker (http://www. repeatmasker.org).
9. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogentic and Genome Research* 110: 462–467.
10. Xu Z, Wang H (2007) LTR_finder: an efficient tool for the predition of full-length ltr retrotransposons. *Nucl. Acids Res.* 35: W265–268.
11. Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19: ii215–225.
12. Salamov AA, Solovyev VV (2000) *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res.* 10: 516–522.
13. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res. 14: 985–995.
14. Mackey AJ, Pereira FCN, Roos DS (2007) GLEAN: improved eukaryotic gene prediction by statistical consensus of gene evidence. Genome Biol. 8: R13.
15. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.
16. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28: 511–515.
17. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A et al. (2007) New developments in the InterPro database. *Nucleic Acids Research* 35: D224–228.
18. Ashburner M, Bal, CA, Blake JA, Botstein D, Butleret H et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25: 25–29.
19. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28: 27–30.
20. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* 25: 955–964.
21. Griffiths-Jones S, Moxon X, Marshall M, Khanna A, Eddy SR (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucl. Acids Res.* 33: D121–124.
22. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK et al. (2012) Draft genome sequence of pigeonpea (Cajanuscajan), an orphan legume crop of resource-poor farmers. *Nature Biotech.* 30: 83–89.
23. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell* 16: 1667–1678.
24. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815.
25. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
26. Pearson WR, Wood T, Zhang Z, Miller W (1997) Comparison of DNA sequences with protein sequences. *Genomics* 46: 24–36.
27. Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy, genes: Rates and interdependence between the

genes. *Molecular Biology and Evolution* 10: 271–281.
28. Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* 36: 96–99.
29. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J (2006) KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4: 259–263.
30. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24: 1586–1591.
31. The Brassica rapa Genome Sequencing Project Consortium. (2011) The genome of the mesopolyploid crop species Brassica rapa. *Nature Genetics* 919: 1–6.
32. Tang H, Bowers J, Wang X, Ming R, Alam M et al. (2008) Synteny and Collinearity in Plant Genomes. *Science* 320: 486–488.
33. Tang H, Wang X, Bowers JE, Ming R, Alam M et al. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Research* 18: 1944–1954.
34. Li R, Yu C, Li Y, Lam T, Yiu S et al. (2009) SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* 25: 1966–1967.
35. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M et al. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36 (Database issue): D480–4.
36. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA et al. (2006) TM4 microarray software suite. Methods in Enzymology 411: 134–93.
37. Albach DC, Soltis PS, Soltis DE (2011) Patterns of embryological and biochemical evolution in the Asterids. *Syst. Bot.* 26: 242–262.
38. The Angiosperm Phylogeny Group (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141: 399–436.
39. The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475: 189–195.
40. The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635–641.
41. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.
42. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28: 511–515.
43. Rodriguez MCS, Edsgard SD, Hussain SS, Alquezar D, Rasmussen M, *et al.* (2010) Transcriptomes of the desiccation-tolerant resurrection plant *Craterostigma plantagineum*. *The Plant Journal* 63: 212–228.
44. Gechev TS, Benina M, Obata T, Tohge T, Sujeeth N, *et al.* (2013) Molecular mechanisms of desiccation tolerance in the resurrection glacial relic *Haberlea rhodopensis*. *Cell Mol. Life Science* 70 (4): 689–709.
45. Komatsu K, Suzuki N, Kuwamura M, Nishikawa Y, Nakatani M, *et al*. (2013) Group A PP2Cs evolved in land plants as key regulators of intrinsic desiccation tolerance. *Nature Communications* 4: 2219. DOI: 10.1038/ncomms3219.
46. Cutler SR, Rodriguez PL, Finkelstein RR, Abrams S (2010)Abscisic acid: emergence of a core signaling network. *Annu. Rev. Plant Biol.* 61: 651–679.
47. Jiang GQ, Wang Z, Shang H, Yang W, Hu Z, *et al.* (2007) Proteome analysis of leaves from the resurrection plant *Boea hygrometrica* in response to dehydration and rehydration. *Planta* 225: 1405–1420.
48. Navari-Izzo F, Meneguzzo S, Loggini B, Vazzana C, Sgherri CLM (1997) The role of the glutathione system during dehydration of *Boea hygroscopica*. *Physiologia Plantarum* 99: 23–30.

Supplemental Figure 1

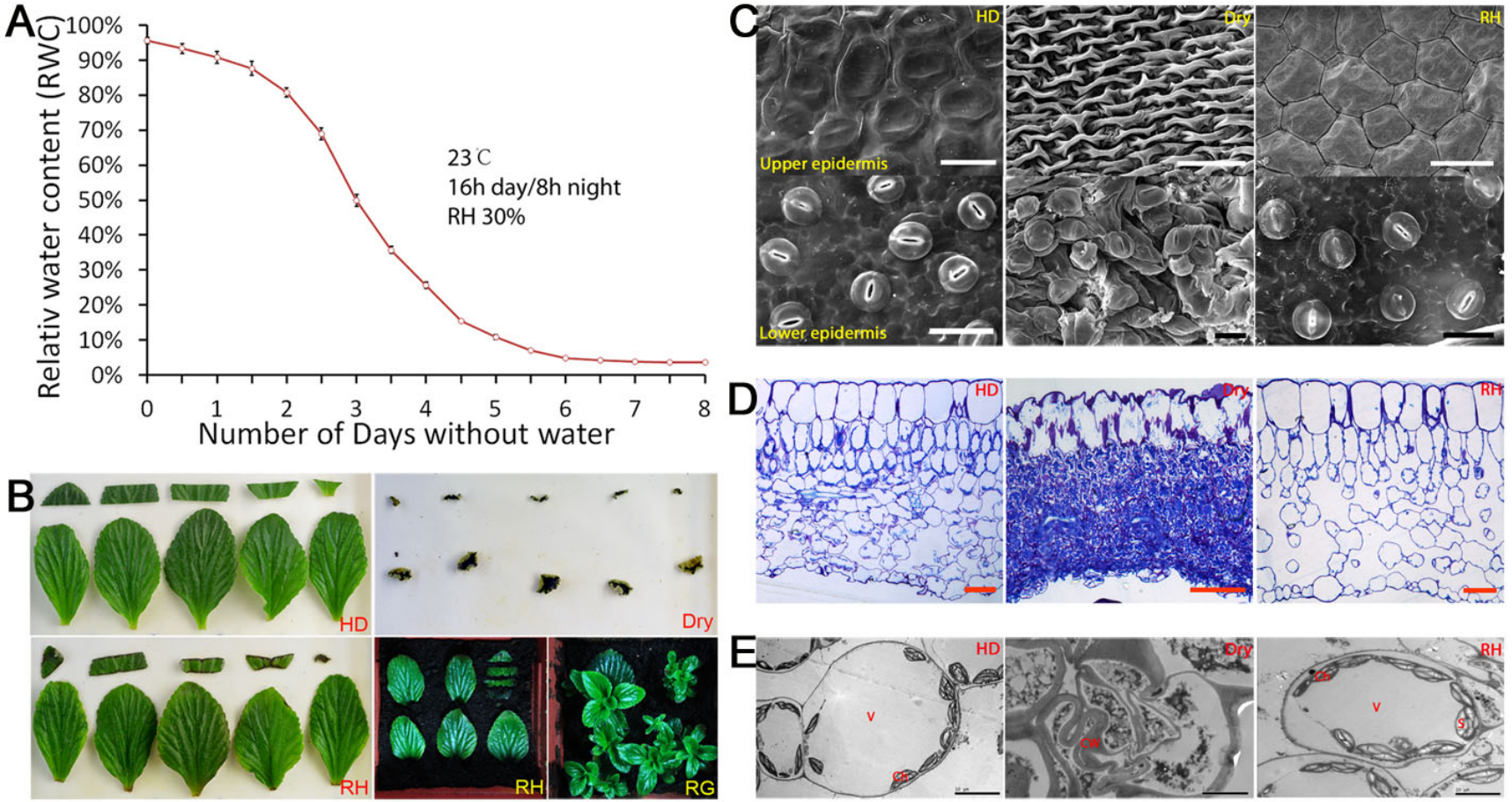

Fig. S1. Dehydration-rehydration cycle of detached leaves and their regeneration in soil.
A. Drying curve during dehydration of three month old plants growth in soil-filled pots, at 25℃, in 16 hrs day / 8 hrs night, and 30% relative humidity (RH).B. Phenotypes of hydrated (HD), dry (5 days), rehydrated for 48 hours (RH), transferred to soil-filled pots after rehydration for 48 hours, and two-month old regeneration seedlings after potting. C. Scanning Electron Micrographs of leaf surfaces (showing upper and lower epidermises) of HD, dry and RH. Scale bar = 50 µm. D. Leaf transection of HD, dry and RH by Toluidine Blue O (TBO) staining. Scale bar = 50 µm. E. Transmission electron micrographs of leaf surfaces (showing upper and lower epidermis) of HD, dry and RH.

Supplemental Figure 2

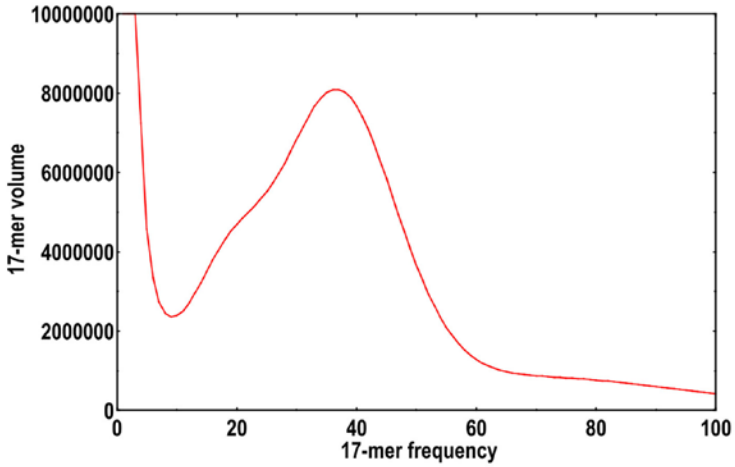

Fig. S2. Illumina 17-mer volume of B. hygrometrica. The volume of K-mers is plotted against the frequency at which they occur. The left-hand, truncated, peak at low frequency and high volume represents K-mers containing essentially random sequencing errors, while the right-hand distribution represents proper (putatively error-free) data. The total K-mer number is 62,569,613,891, and the volume peak is 37. The genome size can be estimated as (total K-mer number)/(the volume peak), which is 1691.0 Mb for *B. hygrometrica*.

Supplemental Figure 3

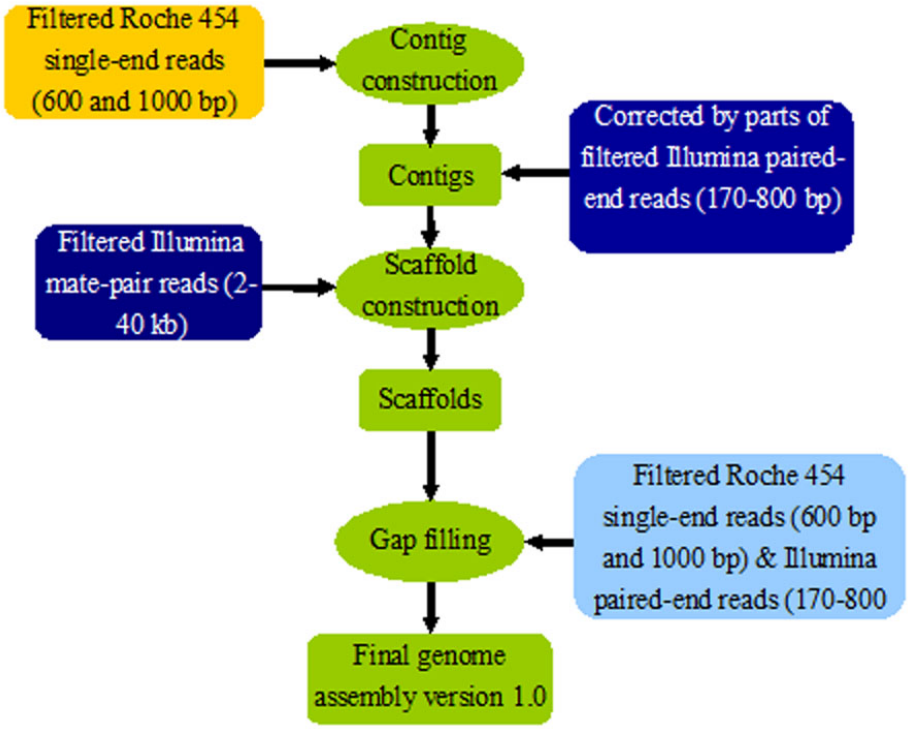

Fig. S3. Pipeline of genome sequencing and assembly.

Supplemental Figure 4

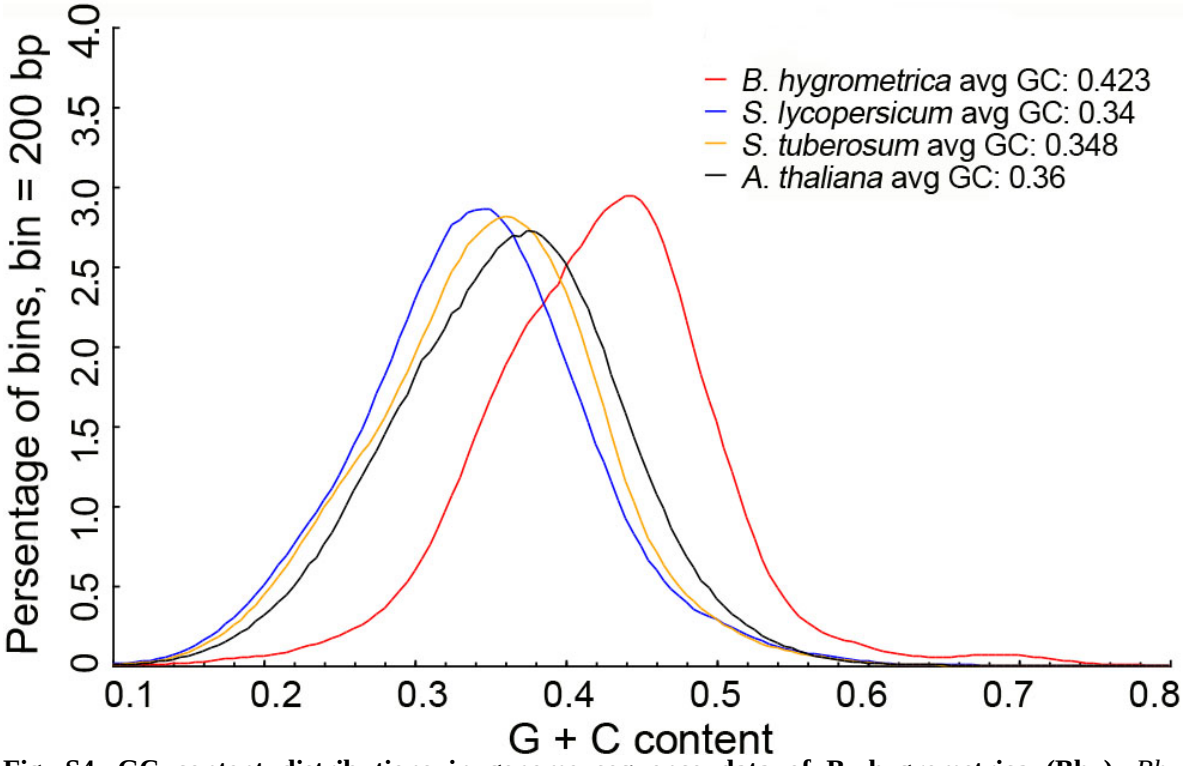

Fig. S4. GC content distributions in genome sequence data of B. hygrometrica (Bhy). *Bhy = B. hygrometrica*, *Stu = S. tuberosum* and *Ath = Arabidopsis thaliana*. The x-axis is GC content percent and the y-axis is the proportion of the windows number divided by the total windows. 200 bp non-overlapping sliding windows have been used along the genomes.

Supplemental Figure 5

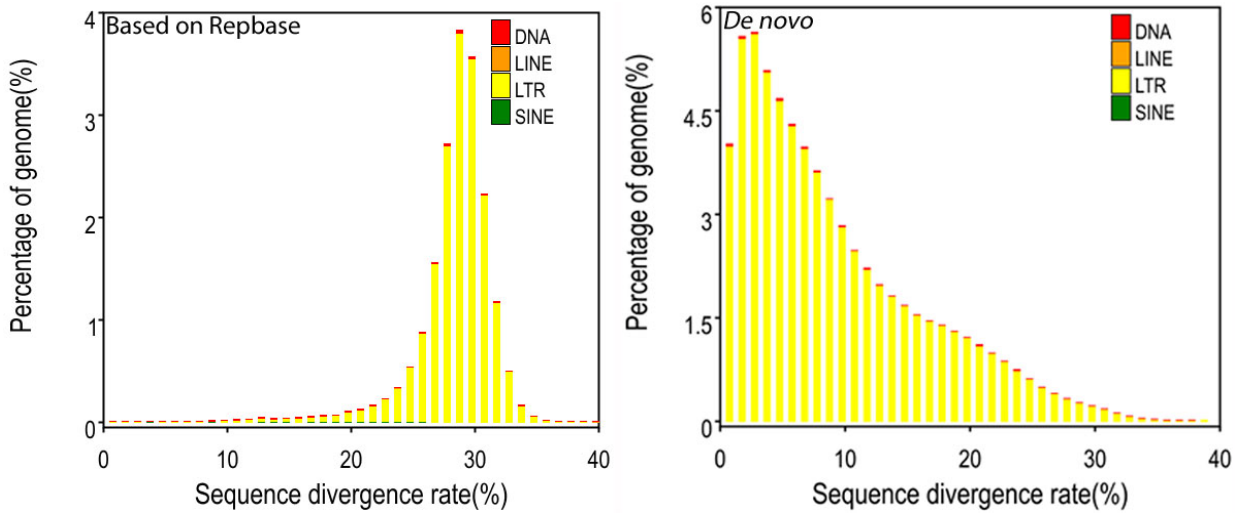

Fig. S5. Distribution of TE sequence divergence in the B. hygrometrica genome.

Supplemental Figure 6

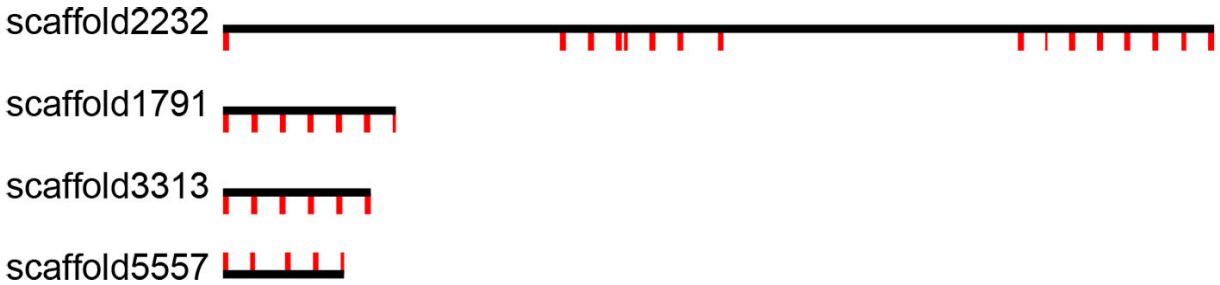

Fig. S6. Scaffolds with 5S rRNAs clustered. Black lines represent scaffolds and the red short lines show 5S rRNA genes. 5S rRNA under scaffolds indicate that they locate at "-" chain of scaffolds, otherwise, they locate at "+" chain.

Supplemental Figure 7

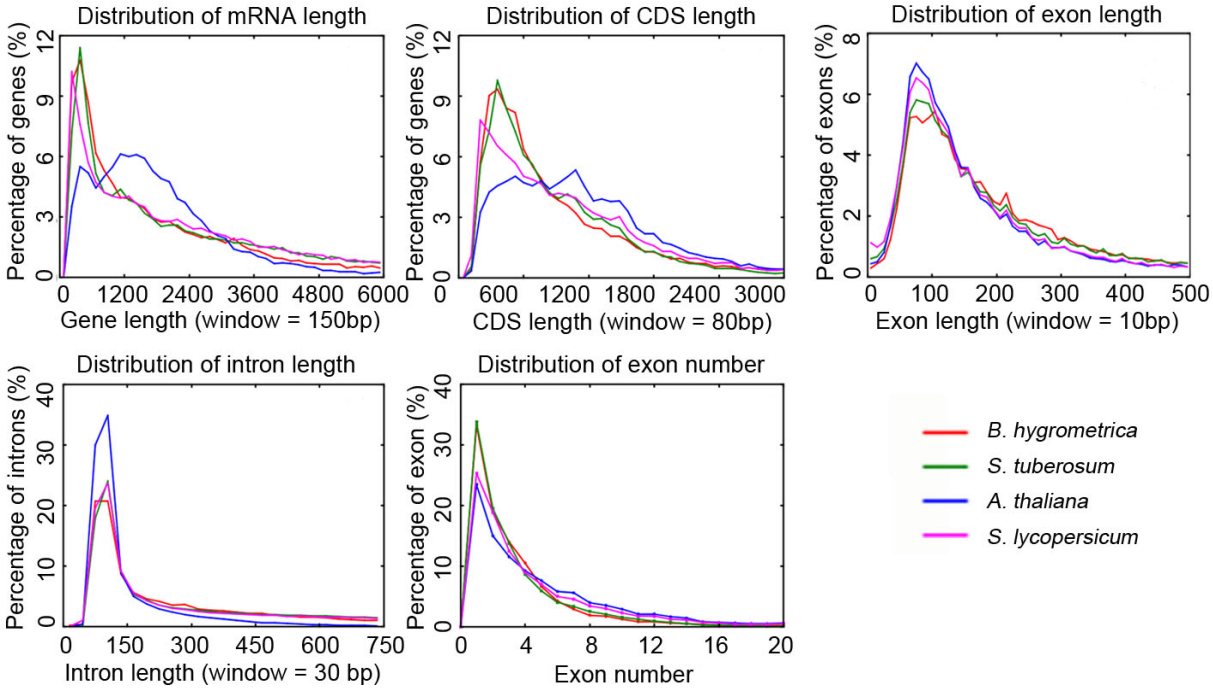

Fig. S7. Comparison of gene parameters of the B. hygrometrica genome, to the S. tuberosum and A.thaliana genomes. No unexpected differences were observed between the *B. hygrometrica*, *S. tuberosum* and *S. lycopersicum* genomes, reflecting the close phylogenetic relationship between the two species, which is indicative of the high quality of the gene structure annotation. Significant differences, however, were observed between *B. hygrometrica* and *A. thaliana* genomes, as expected from the relatively unrelated phylogenetic context for these two species.

Supplemental Figure 8

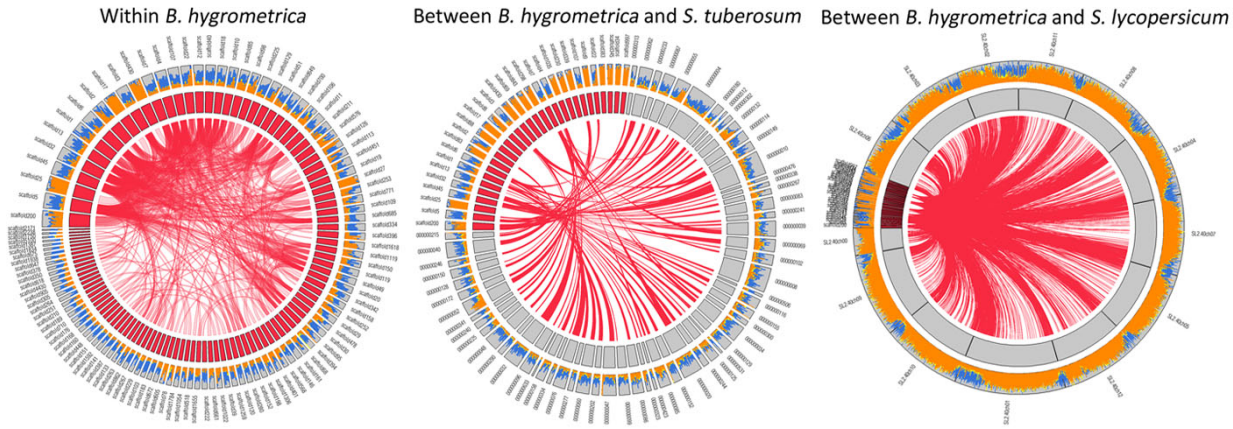

Fig. S8. Micro-synteny analysis. The left, micro-synteny within *B. hygrometrica* scaffolds (only the scaffolds with syntenic relationship are shown). The middle, micro-synteny between scaffolds of *B. hygrometrica* and *S. tuberosum* (comparison of the 30 longest scaffolds in *B. hygrometrica* with the syntenic scaffolds in *S. tuberosum*). The right, micro-synteny between scaffolds of *B. hygrometrica* and *S. lycopersicum* (comparison of the 30 longest scaffolds in *B. hygrometrica* with the syntenic scaffolds in *S. lycopersicum*). *B. hygrometrica* scaffolds are represented by the red block and S. tuberosum by gray (inner circle).

Supplemental Figure 9

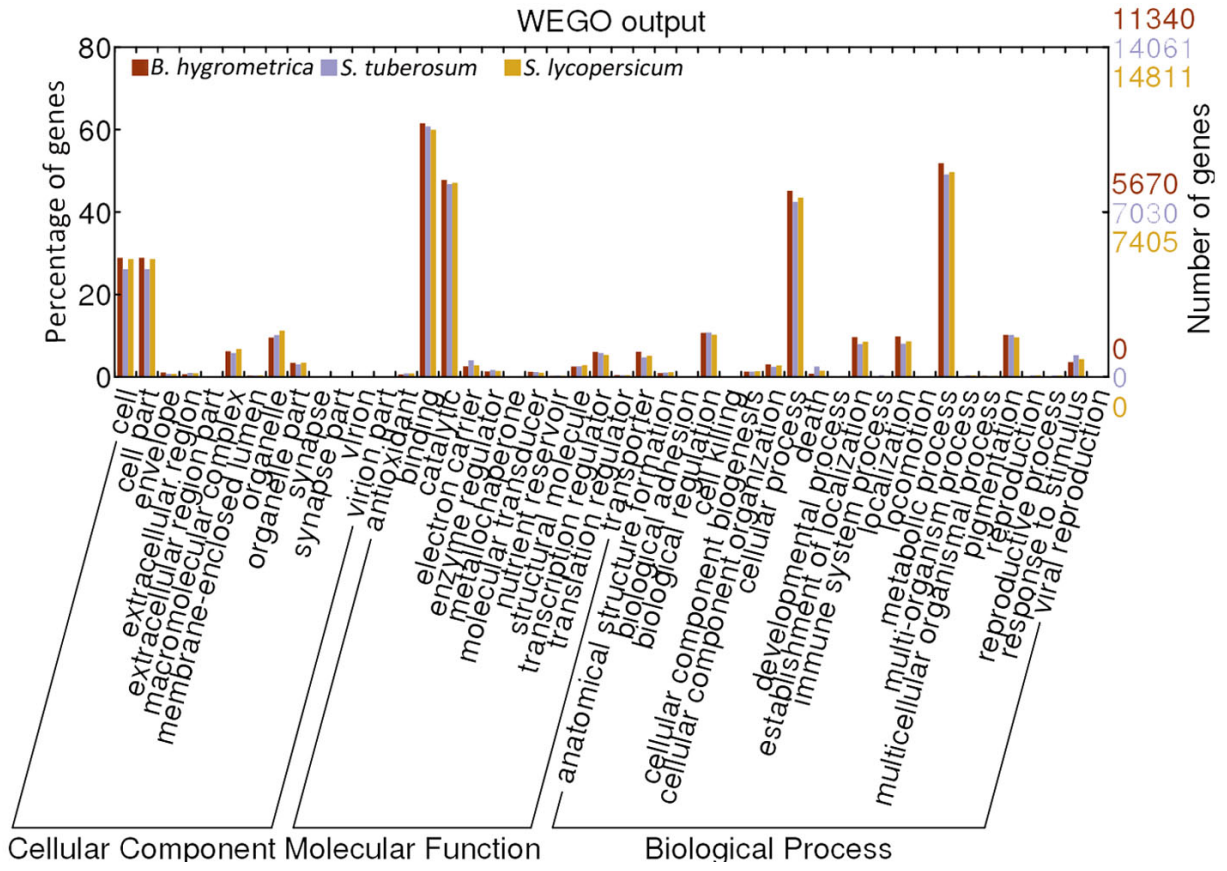

Fig. S9. Gene Ontology enrichment analysis of annotated *B. hygrometrica* genes. The x axis indicates GO terms; left y axis shows the percentage of genes and the right is the number of gene for each GO term involved.

Supplemental Figure 10

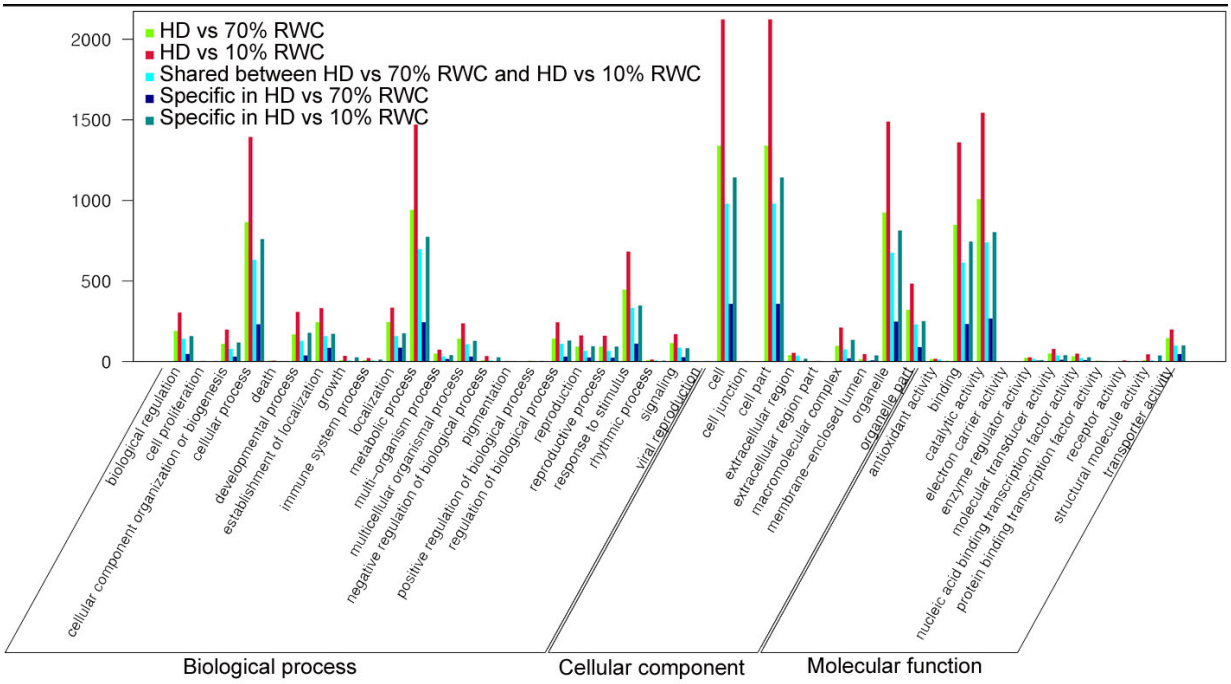

Fig. S10. GO enrichment of AS-DEGs.

Supplemental Tables

Table S1. Overview of the assembly input sequences.

Sequencing data	Library insert size (bp)	No. of lane	Raw data						Filtered data				
			Read length (bp)	Total reads (M)	Total bases (Gb)	Sequence depth (×)	Physical depth (×)		Read length (bp)	Total reads (M)	Total bases (Gb)	Sequence depth (×)	Physical depth (×)
Illumina reads	170	6	100PE	1280.47	128.05	75.72	128.72		95PE	1164.21	110.60	65.40	117.04
	200	2	100PE	381.04	38.10	22.53	45.07		90PE, 95PE	346.84	32.16	19.02	41.02
	250	7	150PE	1803.63	270.54	159.98	266.64		140PE	1446.94	202.57	119.79	213.91
	350	1	100PE	125.95	12.59	7.45	26.07		95PE	111.90	10.63	6.29	23.16
	500	2	100PE	285.60	28.56	16.89	84.44		95PE	242.1	23.00	13.60	71.58
	800	2	100PE, 90PE	307.47	29.00	17.15	145.45		95PE, 85PE	260.57	23.15	13.69	123.27
	2000	3	90PE	521.13	46.90	27.74	616.34		85PE	319.38	27.15	16.05	377.72
	5000	2	90PE	324.78	29.23	17.28	960.27		85PE	192.86	16.39	9.69	570.22
	10000	2	90PE, 49PE	267.80	17.91	10.59	1583.63		85PE, 44PE	138.54	7.56	4.47	819.26
	20000	2	90PE, 49PE	265.32	17.54	10.38	3137.87		85PE, 44PE	64.4	3.80	2.25	761.60
	40000	1	49PE	145.45	7.13	4.21	3440.35		44PE	39.45	1.74	1.03	933.04
	Total	30		5708.64	625.56	369.92	10434.84			4327.19	458.74	271.27	4051.83

Sequencing data	Library insert size (bp)	No. of run	Average read length (bp)	Total reads (M)	Total bases (Gb)	Sequence depth (×)		Average read length (bp)	Total reads (M)	Total bases (Gb)	Sequence depth (×)
454 sequences	600*	10	407.73	10.24	4.17	2.47		407.38	10.10	4.12	2.43
	1000**	17	557.40	20.98	11.69	6.91		556.61	20.67	11.51	6.81
	Total	27	508.32	31.21	15.87	9.38		507.63	30.78	15.62	9.24

* obtained from 454 GS FLX
** obtained from 454 GS FLX +

Table S2. Statistics of final genome assembly.

	Contig		Scaffold	
	Size (bp)	**Number**	**Size (bp)**	**Number**
N90	731	265300	857	108001
N80	1439	132776	6688	20586
N70	3355	69744	26097	8622
N60	6743	41691	66018	4899
N50	11187	26459	110988	3098
Longest	691061		1434191	
Total size	1328817553		1547684042	
Total number (>100bp)		659074		520969
Total number (>2kb)		99602		40367

Table S3. Estimation of B. hygrometrica based on K-mer statistics.

K-mer value	K-mer number	Per k-mer Depth (×)	Genome size (bp)	Used bases	Used reads	Depth (×)
17	62569613891	37	1691070645	75241940755	792020429	44.49

Table S4. Summary of assembly evaluation based on fosmid sequences.

fosmid ID	fosmid length	ratio	scaffold number	scaffold length	scaffold gap number	scaffold gap length
kjtajxa	32174	0.999814	125	9384964	53	53074
kjtavxa	42747	0.999977	86	1718604	13	12020
kjtawxa	31357	0.836974	45	1256271	16	13689
kjtaxa	40187	0.967726	72	1761535	17	20631
kjtaxxa	38621	0.994355	103	2503852	29	38726
kjtayxa	36123	0.598898	1	135609	2	1613
kjtazxa	35764	0.975422	84	2775006	31	28479
kjtbwxa	43556	0.993571	151	1038306	27	23114
kjtbxa	36126	0.981675	74	1281986	10	16347
kjtbxxa	34473	0.732631	1	289974	4	4649
kjtcbxa	34573	0.860874	7	128195	2	789
kjtakxa	32128	0.905627	7	1247759	1	899
kjtcxa	39779	0.976545	87	2230522	36	40620
kjtdxa	36122	0.999972	1	93582	6	7505
kjtexa	34429	0.985361	88	1476864	35	34630
kjtfxa	32014	0.999969	1	102894	7	4726
kjtgxa	34831	0.999943	1	261099	4	2284
kjthxa	35720	0.99734	40	757954	0	0
kjtjxa	33930	1	31	1600894	7	6399
kjtlxa	38266	1	46	5297711	11	9135
kjtmxa	33702	0.999496	150	8989752	63	54357
kjtnxa	34688	0.967078	107	1017393	13	12327
kjtamxa	38099	1	2	333043	0	0
kjtoxa	34052	0.975567	117	4782719	53	40630
kjtpxa	37172	0.999973	1	172196	6	2581
kjtqxa	36703	0.998175	118	6344951	56	71340
kjtrxa	27842	1	1	191194	7	4650
kjtsxa	32340	0.999969	1	64159	5	1444
kjttxa	35544	0.999887	163	1444057	16	9352

kjtvxa	36008	0.70451	111	4092878	43	52004
kjtanxa	33589	0.901307	98	5698818	52	39272
kjtapxa	33112	0.981789	97	2933243	17	13801
kjtaqxa	30853	0.958027	102	1328574	11	5978
kjtarxa	33493	0.99997	67	1472773	28	19323
kjtasxa	31141	0.99438	31	1140722	25	18833
kjtatxa	30352	0.996013	8	145037	0	0

Table S5. Summary of assembly assessment by ESTs.

Data set	Number	Total length (bp)	Covered by assembly (%)	With >90% sequence in one scaffold		With >90% sequence covered		With >50% sequence in one scaffold		With >50% sequence covered	
				number	Percent (%)	number	Percent (%)	number	Percent (%)	number	Percent (%)
All	2360	1107620	95.42	2195	93.01	2209	93.60	2232	94.58	2232	94.58
>200bp	2315	1100938	95.51	2157	93.17	2171	93.78	2191	94.64	2191	94.64
>500bp	874	464705	96.11	819	93.71	824	94.28	830	94.97	830	94.97

Table S6. Repetitive element annotation and statistics for the *B. hygrometrica* genome.

Type	Repeat Size(bp)	% of genome
TRF	62,678,253	4.05
RepeatMasker	288,898,449	18.67
RepeatProteinMask	420,952,717	27.20
De novo	1,154,894,710	74.62
Total	1,172,433,882	75.75

Table S7. Identification of non-coding RNA genes in *B. hygrometrica* genome.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		196	112.413	22,033	0.00142
tRNA		538	76.232	41,013	0.00264
rRNA	Total rRNA	1512	101.629	153,663	0.00988
	18S	191	322.597	61,616	0.00396
	28S	152	119.763	18,204	0.00117
	5.8S	50	131.84	6,592	0.00042
	5S	1119	60.099	67,251	0.00432
snRNA	Total snRNA	151	117.026	17,671	0.00114
	CD-box	82	93.817	7,693	0.00049
	HACA-box	12	141	1,692	0.00011
	splicing	57	145.368	8,286	0.00053

Table S8. General statistics of gene prediction and predicted protein-coding genes for *B. hygrometrica.*

	Gene set	Number	Average gene length (bp)	Total CDS length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	AUGUSTUS	154417	1876.02	126579084	820	4.39	187	311
	GENSCAN	152157	4980.07	142458258	936	5.09	184	988
Homolog	*A.thaliana*	23569	2092.08	22401237	950	3.82	249	405
	C.papaya	28770	1632.02	22664544	788	3.14	251	395
	C.sativus	33077	1644.17	27085536	819	2.97	275	418
	F.vesca	54980	1561.40	37403661	680	2.31	294	672
	G.max	29650	2917.42	29873874	1008	3.45	292	778
	V.vinifera	29023	2170.45	25434486	876	3.49	251	521
GLEAN		48915	2566.42	48228417	986	3.62	272	603
RNA-Seq based gene models		20087	2674.47	20899665	1040	4.48	232	469
Final set/BhV1.0		49374	2535.41	48253437	977	3.58	273	604

Table S9. Functional annotation of predicted genes for *B. hygrometrica*.

		Number	Percent (%)
Annotated	InterPro	18,618	37.71
	GO	14,176	28.71
	KEGG	12,159	24.63
	Swissprot	16,909	34.25
	TrEMBL	22,771	46.12
	Total Annotated	23,250	47.09
Unannotated		26,124	52.91
Total gene		49,374	

Table S10. Origin and evolution of ORFan genes.

	Method	Gene number	Percentage (%)
Total ORFan genes	-	14,391	100.00
Duplicated origin	Blastn in Boea gene set (e value < or =1e-4)	13,966	97.05
Frame-shift origin	Blastn to coding sequence of other speices (coding hits)	80	0.56
Denovo or a result of gene loss	Blastn to Intergenic or intron hits of other species	166	1.15
Unknown	No hits with any other species	179	1.24

Table S11. Statistics of ORFan genes in DEGs and all expressed genes during dehydration.

Treatment	Sample	No. of ORFan genes in DEGs	No. of DEGs	Perentage in DEGs (%)	No. of ORFan genes in each sample	No. of expressed genes in each sample	Perentage in expressed genes (%)	Sum
Dehydration	**HD**	122	8483	1.19	1755	20624	8.51	
	70% RWC	47	4599	1.02	1832	20381	8.99	128
	10% RWC	103	6794	1.60	2198	20977	10.48	

Table S12. Significantly enriched GO terms in DEGs.

Biologichal process		
GO term	**HD vs 70% RWC**	**HD vs 10% RWC**
GO:0008150:biological process	2562	3764
GO:0022610 : biological adhesion	0	2
GO:0065007 : biological regulation	294	449
GO:0001906 : cell killing	0	2
GO:0008283 : cell proliferation	5	9
GO:0071840 : cellular component organization or biogenesis	262	422
GO:0009987 : cellular process	1605	2435
GO:0016265 : death	17	22
GO:0032502 : developmental process	351	583
GO:0051234 : establishment of localization	420	568
GO:0040007 : growth	25	58
GO:0002376 : immune system process	27	43
GO:0051179 : localization	425	577
GO:0008152 : metabolic process	1722	2539
GO:0051704 : multi-organism process	99	133
GO:0032501 : multicellular organismal process	277	440
GO:0048519 : negative regulation of biological process	21	56
GO:0043473 : pigmentation	6	5
GO:0048518 : positive regulation of biological process	11	12
GO:0050789 : regulation of biological process	172	293
GO:0000003 : reproduction	171	287
GO:0022414 : reproductive process	166	278
GO:0050896 : response to stimulus	797	1147
GO:0007155 : cell adhesion	0	2
GO:0050789 : regulation of biological process	172	293
GO:0065008 : regulation of biological quality	116	166
GO:0065009 : regulation of molecular function	38	41
GO:0071554 : cell wall organization or biogenesis	54	65
GO:0044085 : cellular component biogenesis	59	101
GO:0016043 : cellular component organization	222	353
GO:0071841 : cellular component organization or biogenesis at cellular level	159	279
GO:0030029 : actin filament-based process	4	10
GO:0007154 : cell communication	33	46
GO:0007049 : cell cycle	34	56
GO:0022402 : cell cycle process	26	40
GO:0008219 : cell death	17	22
GO:0051301 : cell division	10	14
GO:0016049 : cell growth	22	43
GO:0048869 : cellular developmental process	91	125
GO:0019725 : cellular homeostasis	39	46
GO:0051641 : cellular localization	57	86
GO:0016044 : cellular membrane organization	15	20
GO:0044237 : cellular metabolic process	1234	1844
GO:0048610 : cellular process involved in reproduction	3	5

GO:0051716 : cellular response to stimulus	142	225
GO:0007059 : chromosome segregation	0	4
GO:0000910 : cytokinesis	6	6
GO:0032506 : cytokinetic process	2	2
GO:0016458 : gene silencing	11	34
GO:0010496 : intercellular transport	0	4
GO:0051651 : maintenance of location in cell	0	2
GO:0007017 : microtubule-based process	24	28
GO:0048523 : negative regulation of cellular process	10	23
GO:0048522 : positive regulation of cellular process	7	8
GO:0050794 : regulation of cellular process	136	221
GO:0032940 : secretion by cell	6	7
GO:0010118 : stomatal movement	15	18
GO:0006413 : translational initiation	0	2
GO:0008219 : cell death	17	22
GO:0007568 : aging	9	14
GO:0048532 : anatomical structure arrangement	7	8
GO:0048856 : anatomical structure development	207	346
GO:0048646 : anatomical structure formation involved in morphogenesis	17	25
GO:0009653 : anatomical structure morphogenesis	110	164
GO:0048869 : cellular developmental process	91	125
GO:0044111 : development involved in symbiotic interaction	2	2
GO:0048589 : developmental growth	13	22
GO:0021700 : developmental maturation	3	5
GO:0003006 : developmental process involved in reproduction	138	241
GO:0010073 : meristem maintenance	7	15
GO:0007275 : multicellular organismal development	240	396
GO:0051093 : negative regulation of developmental process	4	9
GO:0007389 : pattern specification process	28	50
GO:0009791 : post-embryonic development	100	167
GO:0050793 : regulation of developmental process	15	24
GO:0019827 : stem cell maintenance	2	6
GO:0051649 : establishment of localization in cell	346	82
GO:0051656 : establishment of organelle localization	5	3
GO:0045184 : establishment of protein localization	39	67
GO:0051236 : establishment of RNA localization	4	4
GO:0006810 : transport	384	519
GO:0016049 : cell growth	22	43
GO:0048589 : developmental growth	13	22
GO:0045926 : negative regulation of growth	0	3
GO:0045927 : positive regulation of growth	0	2
GO:0040008 : regulation of growth	3	0
GO:0002253 : activation of immune response	2	3
GO:0002252 : immune effector process	12	16
GO:0006955 : immune response	10	20
GO:0002684 : positive regulation of immune system process	2	3
GO:0002682 : regulation of immune system process	3	4

GO:0051641 : cellular localization	57	86
GO:0051234 : establishment of localization	420	568
GO:0033036 : macromolecule localization	19	40
GO:0051235 : maintenance of location	0	2
GO:0032879 : regulation of localization	10	14
GO:0009058 : biosynthetic process	406	591
GO:0009056 : catabolic process	162	228
GO:0070988 : demethylation	0	2
GO:0042445 : hormone metabolic process	15	20
GO:0043170 : macromolecule metabolic process	725	1223
GO:0032259 : methylation	8	19
GO:0009892 : negative regulation of metabolic process	11	37
GO:0006807 : nitrogen compound metabolic process	483	763
GO:0071704 : organic substance metabolic process	12	15
GO:0019637 : organophosphate metabolic process	40	38
GO:0055114 : oxidation-reduction process	41	63
GO:0042440 : pigment metabolic process	15	19
GO:0009893 : positive regulation of metabolic process	2	0
GO:0044238 : primary metabolic process	36	70
GO:0019222 : regulation of metabolic process	61	127
GO:0019748 : secondary metabolic process	72	99
GO:0044281 : small molecule metabolic process	398	584
GO:0009292 : genetic transfer	4	10
GO:0044419 : interspecies interaction between organisms	8	7
GO:0009856 : pollination	16	23
GO:0051707 : response to other organism	77	96
GO:0032504 : multicellular organism reproduction	9	15
GO:0007275 : multicellular organismal development	7	11
GO:0048609 : multicellular organismal reproductive process	9	15
GO:0043480 : pigment accumulation in tissues	6	5
GO:0051239 : regulation of multicellular organismal process	7	11
GO:0048316 : seed development	19	27
GO:0009845 : seed germination	10	10
GO:0009606 : tropism	34	32
GO:0043476 : pigment accumulation	6	5
GO:0032504 : multicellular organism reproduction	9	15
GO:0022414 : reproductive process	166	278
GO:0019953 : sexual reproduction	7	14
GO:0048610 : cellular process involved in reproduction	3	5
GO:0003006 : developmental process involved in reproduction	138	241
GO:0009566 : fertilization	3	8
GO:0022415 : viral reproductive process	4	2
GO:0007610 : behavior	2	2
GO:0051606 : detection of stimulus	8	19
GO:0048583 : regulation of response to stimulus	5	7
GO:0009628 : response to abiotic stimulus	293	407
GO:0009607 : response to biotic stimulus	90	110

GO:0042221 : response to chemical stimulus	358	548
GO:0009719 : response to endogenous stimulus	216	315
GO:0009605 : response to external stimulus	73	88
GO:0006950 : response to stress	383	539
GO:0007623 : circadian rhythm	3	7
GO:0007165 : signal transduction	61	80
GO:0009850 : auxin metabolic process	9	9
GO:0006081 : cellular aldehyde metabolic process	4	5
GO:0043449 : cellular alkene metabolic process	4	7
GO:0006725 : cellular aromatic compound metabolic process	87	104
GO:0044249 : cellular biosynthetic process	396	571
GO:0044262 : cellular carbohydrate metabolic process	126	164
GO:0044248 : cellular catabolic process	114	169
GO:0034754 : cellular hormone metabolic process	5	6
GO:0042180 : cellular ketone metabolic process	177	273
GO:0044255 : cellular lipid metabolic process	134	181
GO:0044260 : cellular macromolecule metabolic process	603	971
GO:0034641 : cellular nitrogen compound metabolic process	462	730
GO:0051186 : cofactor metabolic process	64	86
GO:0006091 : generation of precursor metabolites and energy	76	97
GO:0046483 : heterocycle metabolic process	148	206
GO:0010191 : mucilage metabolic process	5	6
GO:0031324 : negative regulation of cellular metabolic process	5	12
GO:0006730 : one-carbon metabolic process	40	73
GO:0006082 : organic acid metabolic process	170	264
GO:0006518 : peptide metabolic process	3	7
GO:0006793 : phosphorus metabolic process	99	139
GO:0015979 : photosynthesis	28	34
GO:0031323 : regulation of cellular metabolic process	42	81
GO:0006790 : sulfur compound metabolic process	17	34
GO:0009404 : toxin metabolic process	0	5
GO:0006805 : xenobiotic metabolic process	3	4
GO:0060255 : regulation of macromolecule metabolic process	38	96
GO:0019538 : protein metabolic process	406	636
GO:0005976 : polysaccharide metabolic process	57	86
GO:0010605 : negative regulation of macromolecule metabolic process	11	37
GO:0043412 : macromolecule modification	271	414
GO:0009057 : macromolecule catabolic process	38	70
GO:0009059 : macromolecule biosynthetic process	204	313
GO:0010467 : gene expression	222	440
GO:0044260 : cellular macromolecule metabolic process	603	971
GO:0044036 : cell wall macromolecule metabolic process	5	8
Cellular component		
GO term	**HD vs 70% RWC**	**HD vs 10% RWC**
GO:0005575 : cellular_component	2628	3922
GO:0005623 : cell	2595	3877
GO:0030054 : cell junction	7	11

GO:0044464 : cell part	2595	3877
GO:0031012 : extracellular matrix	3	5
GO:0005576 : extracellular region	94	116
GO:0032991 : macromolecular complex	188	387
GO:0016020 : membrane	921	1250
GO:0031974 : membrane-enclosed lumen	38	98
GO:0043226 : organelle	1721	2604
GO:0044422 : organelle part	524	785
GO:0005911 : cell-cell junction	7	10
GO:0000267 : cell fraction	17	15
GO:0071944 : cell periphery	181	244
GO:0042995 : cell projection	6	11
GO:0044463 : cell projection part	3	4
GO:0009986 : cell surface	0	2
GO:0012505 : endomembrane system	22	37
GO:0031975 : envelope	167	240
GO:0030312 : external encapsulating structure	150	198
GO:0044462 : external encapsulating structure part	3	4
GO:0005622 : intracellular	1852	2873
GO:0044424 : intracellular part	1845	2858
GO:0008287 : protein serine/threonine phosphatase complex	4	3
GO:0044420 : extracellular matrix part	2	2
GO:0044425 : membrane part	383	522
GO:0031090 : organelle membrane	183	257
GO:0019867 : outer membrane	9	8
GO:0034357 : photosynthetic membrane	39	46
GO:0005886 : plasma membrane	29	42
GO:0043233 : organelle lumen	38	98
GO:0043229 : intracellular organelle	1690	2566
GO:0043227 : membrane-bounded organelle	101	188
GO:0043228 : non-membrane-bounded organelle	101	188
GO:0044422 : organelle part	524	785
GO:0031982 : vesicle	243	285
GO:0044446 : intracellular organelle part	421	662
GO:0000313 : organellar ribosome	4	4
GO:0031410 : cytoplasmic vesicle	239	284
GO:0043231 : intracellular membrane-bounded organelle	1445	2244
GO:0043232 : intracellular non-membrane-bounded organelle	101	188
GO:0044446 : intracellular organelle part	421	662
GO:0005737 : cytoplasm	1017	1488
GO:0044444 : cytoplasmic part	1013	1484
GO:0031234 : extrinsic to internal side of plasma membrane	11	16
GO:0043229 : intracellular organelle	1690	2566
GO:0044446 : intracellular organelle part	421	662
GO:0019866 : organelle inner membrane	41	68
GO:0031968 : organelle outer membrane	9	8
GO:0030529 : ribonucleoprotein complex	42	145

GO:0009579 : thylakoid	133	166
GO:0044436 : thylakoid part	50	56
GO:0000151 : ubiquitin ligase complex	13	40
Molecular function		
GO term	**HD vs 70% RWC**	**HD vs 10% RWC**
GO: 0003674:Molecular function	2878	4261
GO:0016209:antioxidant activity	33	37
GO:0005488:binding	1670	2546
GO:0003824: catalytic activity	1998	2821
GO:0004601:peroxidase activity	6	6
GO:0043176 : amine binding	2	2
GO:0030246 : carbohydrate binding	17	22
GO:0031406 : carboxylic acid binding	64	14
GO:0003682 : chromatin binding	0	2
GO:0048037 : cofactor binding	92	132
GO:0008144 : drug binding	3	9
GO:0042562 : hormone binding	3	4
GO:0043167 : ion binding	544	726
GO:0008289 : lipid binding	17	25
GO:0051540 : metal cluster binding	26	33
GO:0003676 : nucleic acid binding	228	421
GO:0001871 : pattern binding	3	7
GO:0042277 : peptide binding	2	4
GO:0046906 : tetrapyrrole binding	8	11
GO:0009975 : cyclase activity	3	0
GO:0004133 : glycogen debranching enzyme activity	0	3
GO:0016787 : hydrolase activity	587	874
GO:0016853 : isomerase activity	56	81
GO:0016874 : ligase activity	82	118
GO:0016829 : lyase activity	85	130
GO:0016491 : oxidoreductase activity	438	554

Table S13. Summary of alternative splicing types in DEGs.

	Alternative splicing type	HD vs 70% RWC	HD vs 10% RWC	Shared between HD vs 70% RWC and HD vs 10% RWC	Specific in HD vs 70% RWC	Specific in HD vs 10% RWC
IA DEGs	Total	1166	1861	764	402	1133
	A3SS	892	1475	594	298	905
	A5SS	658	1189	425	233	766
	Retained Intron	559	795	364	195	453
	Skipped Exon	298	501	194	104	311
DA DEGs	Total	1186	2023	981	205	1006
	A3SS	799	1426	658	141	744
	A5SS	617	1036	494	123	540
	Retained Intron	448	790	376	72	392
	Skipped Exon	205	367	163	42	200

Table S14. TE statistics of de novo annotation or annotated in Repbase.

Type	**De novo TE prediction and type**		**Annotated Tes in Repbase**	
	Length (bp)	**% in genome**	**Length (bp)**	**% in genome**
DNA	5922377	0.38	3159308	0.20
LINE	2185943	0.14	463409	0.03
SINE	0	0.00	6666	0.00
LTR	1129691275	72.99	285448954	18.44
Unknown	16757152	1.08	4985	0.00
Total	1151489821	74.40	288898449	18.67

	The most abundant TE sub-type		
Sub-type	**Number**	**Repeat size (bp)**	**% in genome**
LTR/Gypsy	299063	148305429	9.58
LTR/Copia	274460	134369353	8.68
LTR/Caulimovirus	4305	2091553	0.14

Table S15. Synteny blocks within *B. hygrometrica* and between *B. hygrometrica, S. tuberosum* and *S. lycopersicum*.

Species	Number of synteny blocks	Average syntenic genes per block	Number of syntenic genes in all blocks	Mean syntanic block length (bp)
B. hygrometrica/B. hygrometrica	348	6.954	2,420	546,639.05
B. hygrometrica/S. tuberosum	560	9.9429	5,568	195,491.9268/450,308.1536*
B. hygrometrica/S. lycopersicum	3,951	7.5057	29,655	373,348.8185/51,003,550.8902*

Table S16. Percentage of syntenic length on each chromosome or scaffold in the genomes of *B. hygrometrica*, *S. tuberosum* and *S. lycopersicum*.

Species	Chromosome or scaffold ID	Syntonic length (bp)	Chromosome or scaffold length (bp)	Percentage on chromosome or scaffold (%)
S. tuberosum	PGSC0003DMB000000010	3261149	3949542	82.57
	PGSC0003DMB000000060	1500812	2190062	68.53
	PGSC0003DMB000000063	169326	2168858	7.81
	PGSC0003DMB000000083	1811076	1941144	93.30
	PGSC0003DMB000000099	1097190	1769961	61.99
	PGSC0003DMB000000204	862579	1149568	75.04
	PGSC0003DMB000000377	401477	587821	68.30
	PGSC0003DMB000000382	392433	573661	68.41
	PGSC0003DMB000000423	467795	510824	91.58
	PGSC0003DMB000000453	247347	466268	53.05
	PGSC0003DMB000000575	141040	297372	47.43
	PGSC0003DMB000000689	137488	199774	68.82
	chr03	15430799	51794595	29.79
	chr04	11747636	76018607	15.45
	chr06	13210781	59252670	22.30
	chr07	12691040	53013183	23.94
	chr08	11313235	42636723	26.53
	chr09	12903529	58431464	22.08
	chr1	14178088	86202534	16.45
	chr10	13218516	55065866	24.00
	chr11	11226839	45655905	24.59
	chr12	8803289	58334187	15.09
	chr2	10882160	47805827	22.76
	chr5	6373804	48683283	13.09
	whole genome (redundancies filtered)	152469428	727424546	20.96
	whole genome (redundancies included)	252172566	727424546	34.67
S. lypersicum	SL2.40ch01	18443934	90304244	20.42
	SL2.40ch02	19441429	49918294	38.95
	SL2.40ch03	24576974	64840714	37.9
	SL2.40ch04	15696590	64064312	24.5
	SL2.40ch05	10750913	65021438	16.53
	SL2.40ch06	17224323	46041636	37.41
	SL2.40ch07	19490458	65268621	29.86
	SL2.40ch08	20080235	63032657	31.86
	SL2.40ch09	25840413	67662091	38.19
	SL2.40ch10	14447081	64834305	22.28
	SL2.40ch11	13722333	53386025	25.7
	SL2.40ch12	12673966	65486253	19.35
	Whole genome (redundancies filtered)			
	Whole genome (redundancies included)			