

RESEARCH ARTICLE

# Improving the Annotation of *Arabidopsis lyrata* Using RNA-Seq Data

Vimal Rawat<sup>1</sup>, Ahmed Abdelsamad<sup>2,3</sup>, Björn Pietzenek<sup>2</sup>, Danelle K. Seymour<sup>4</sup>, Daniel Koenig<sup>4</sup>, Detlef Weigel<sup>4</sup>, Ales Pecinka<sup>2\*</sup>, Korbinian Schneeberger<sup>1\*</sup>

**1** Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829, Cologne, Germany, **2** Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829, Cologne, Germany, **3** Department of Genetics, Cairo University, 12613, Giza, Egypt, **4** Department for Molecular Biology, Max Planck Institute for Developmental Biology, Spemannstrasse 35–39, 72076, Tübingen, Germany

\* [schneeberger@mpipz.mpg.de](mailto:schneeberger@mpipz.mpg.de) (KS); [pecinka@mpipz.mpg.de](mailto:pecinka@mpipz.mpg.de) (AP)



**OPEN ACCESS**

**Citation:** Rawat V, Abdelsamad A, Pietzenek B, Seymour DK, Koenig D, Weigel D, et al. (2015) Improving the Annotation of *Arabidopsis lyrata* Using RNA-Seq Data. PLoS ONE 10(9): e0137391. doi:10.1371/journal.pone.0137391

**Editor:** Nicholas James Provart, University of Toronto, CANADA

**Received:** June 12, 2015

**Accepted:** August 17, 2015

**Published:** September 18, 2015

**Copyright:** © 2015 Rawat et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Short read data presented in this paper are available through NCBI Short Read Archive under accession number GSE69222. Short reads from the cold treatment are deposited at the European Nucleotide Archive under accession number PRJEB6701.

**Funding:** This work was funded by German Research Foundation (dfg.de) grant AP1829-2 to AP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Abstract

Gene model annotations are important community resources that ensure comparability and reproducibility of analyses and are typically the first step for functional annotation of genomic regions. Without up-to-date genome annotations, genome sequences cannot be used to maximum advantage. It is therefore essential to regularly update gene annotations by integrating the latest information to guarantee that reference annotations can remain a common basis for various types of analyses. Here, we report an improvement of the *Arabidopsis lyrata* gene annotation using extensive RNA-seq data. This new annotation consists of 31,132 protein coding gene models in addition to 2,089 genes with high similarity to transposable elements. Overall, ~87% of the gene models are corroborated by evidence of expression and 2,235 of these models feature multiple transcripts. Our updated gene annotation corrects hundreds of incorrectly split or merged gene models in the original annotation, and as a result the identification of alternative splicing events and differential isoform usage are vastly improved.

## Introduction

*Arabidopsis lyrata* is a predominantly self-incompatible, perennial plant species that diverged from a common ancestor with *A. thaliana* approximately 10 million years ago [1]. Despite its evolutionary closeness, its genome size is estimated to be between 230 to 245 Mb, or one and a half times as large as the *A. thaliana* genome [2,3]. Except for *A. thaliana*, *A. lyrata* is the only species within the family of Brassicaceae with a reference assembly exclusively based on high quality dideoxy sequencing. This 207 Mb *A. lyrata* reference assembly attributed the difference in genome size between the two species to the accumulation of many small deletions in the *A. thaliana* genome primarily in non-coding regions and transposable elements (TEs) [1]. In addition, *A. lyrata* has experienced recent genome expansion due to activity of TEs, in particular *Copia* long terminal repeat (LTR) retrotransposons [1,4,5], which is the basis for species-specific patterns in DNA methylation [6].

**Competing Interests:** The authors have declared that no competing interests exist.

As *A. lyrata* is the closest fully assembled relative of *A. thaliana*, it serves as an important out-group for evolutionary studies within *A. thaliana* [7–9]. Moreover, recent advances in sequencing technology have enabled the assembly of an increasing number of Brassicaceae genomes and their close relatives [4,5,10–19], which, together, are leveraged for comparative genomics in this family. Intra- as well as inter-species comparisons, however, heavily rely on the gene annotations of each species involved and high quality annotations even in the non-model species become essential.

Methods for gene model annotation profited considerably from the invention of high-throughput RNA sequencing (RNA-seq) [20,21]. Identification of genuine transcription start and termination sites as well as intron/exon borders is a non-trivial task when using only reference sequences and homology data. Now, information on spliced alignments from RNA-seq data can improve the identification of gene models [22,23] and also enable the annotation of variant isoforms [24]. In particular, the gene annotations of model species have been updated regularly despite only minor changes to the reference genome sequence [25].

The current gene annotation of the *A. lyrata* includes 32,670 genes and was generated using a combination of *ab initio* gene prediction, homology to known proteins, as well as gene sequences and expression data from related species [1]. Even though the gene models were analyzed for their expression support using RNA-seq data, gene prediction methods integrating RNA-seq alignment information had not been developed at the time the assembly had been generated. In a recent study, Haudry and colleagues supplemented the original annotation with additional putatively transcribed regions in order to study the conservation of non-coding sequences among related Brassicaceae species [14]. They integrated the results of additional *ab initio* gene predictions, RNA-seq data alignments and homology searches against the genes of *A. thaliana* in order to mask potentially un-annotated coding sequences and regions that recently lost coding potential due to mutations.

Building upon the major efforts of the first annotation of *A. lyrata* genome (version-1 from hereon) we have updated the gene models using diverse RNA-seq samples. Our annotation (version-2 from hereon) has changed the coordinates of 29,141 of the original 32,670 gene models, removed 1,286 and added 1,295 new models. This update corrected hundreds of gene models, which were wrongly merged or split in version-1, and also separated transposable element genes from other protein coding genes. Finally, we have analyzed the transcriptional response of *A. lyrata* to heat stress to show the improved utility of version-2 for the identification of differential isoform usage and pre-mRNA splicing.

## Results and Discussion

### Improving the *A. lyrata* gene annotation using transcriptional data

We sequenced the transcriptome of various *A. lyrata* aerial tissues, including whole rosettes, dissected shoot apices, complete inflorescences, as well as vegetative rosettes exposed to cold and heat stress (see [Materials and Methods](#)). In total, we generated over 290 million single-end, strand unspecific short reads using Illumina sequencing technology after poly-A purification ([Table 1](#)). Short reads were aligned to the *A. lyrata* reference assembly [1] using Bowtie v2.1.0 [26] and the splice junction mapper TopHat v2.0.9 [27] (see [Materials and Methods](#)). We could align 89% of all reads, out of which 85% aligned uniquely and were used for further analysis. The proportion of unaligned reads was comparable to the proportion of unaligned reads in similar experiments with *A. thaliana*, which presumably has one of the most complete reference genome sequences. Over 10% of the aligned reads matched to putative intergenic regions indicating that some gene models may have been missed in the original version of the *A. lyrata* gene annotation. Visual inspection of these intergenic alignments revealed the expected

**Table 1. Short read statistics (read numbers in millions).**

Tissue	Sample	Read length (bp)	Raw reads	Reads aligned	Reads aligned uniquely (full-length alignments)	Reads aligned uniquely (spliced alignments)
Rosette (WT)	Rep 1	96	16.0	11.9	8.3	3.0
Rosette (WT)	Rep 2	96	12.0	9.1	6.0	2.7
Rosette (Heat stressed)	Rep 1	96	17.6	13.1	9.7	3.4
Rosette (Heat stressed)	Rep 2	96	5.9	4.6	3.1	1.2
Rosette (Recovered)	Rep 1	96	15.6	12.6	5.6	4.1
Rosette (Recovered)	Rep 2	96	5.8	4.5	2.8	1.3
Shoot apical meristem (WT)	Rep 1	101	14.5	12.7	7.5	3.6
Rosette (WT)	Rep 1	101	19.6	17.5	9.3	5.6
Rosette (WT)	Rep 2	101	18.1	16.2	9.3	4.8
Inflorescence (WT)	Rep 1	75	32.0	30.2	12.9	9.3
Inflorescence (WT)	Rep 2	75	32.0	29.5	12.6	9.0
Rosette (WT and cold stressed)	Rep 1	75–100	102.7	96.7	59.7	24.6
			291.8	258.6	146.8	72.6

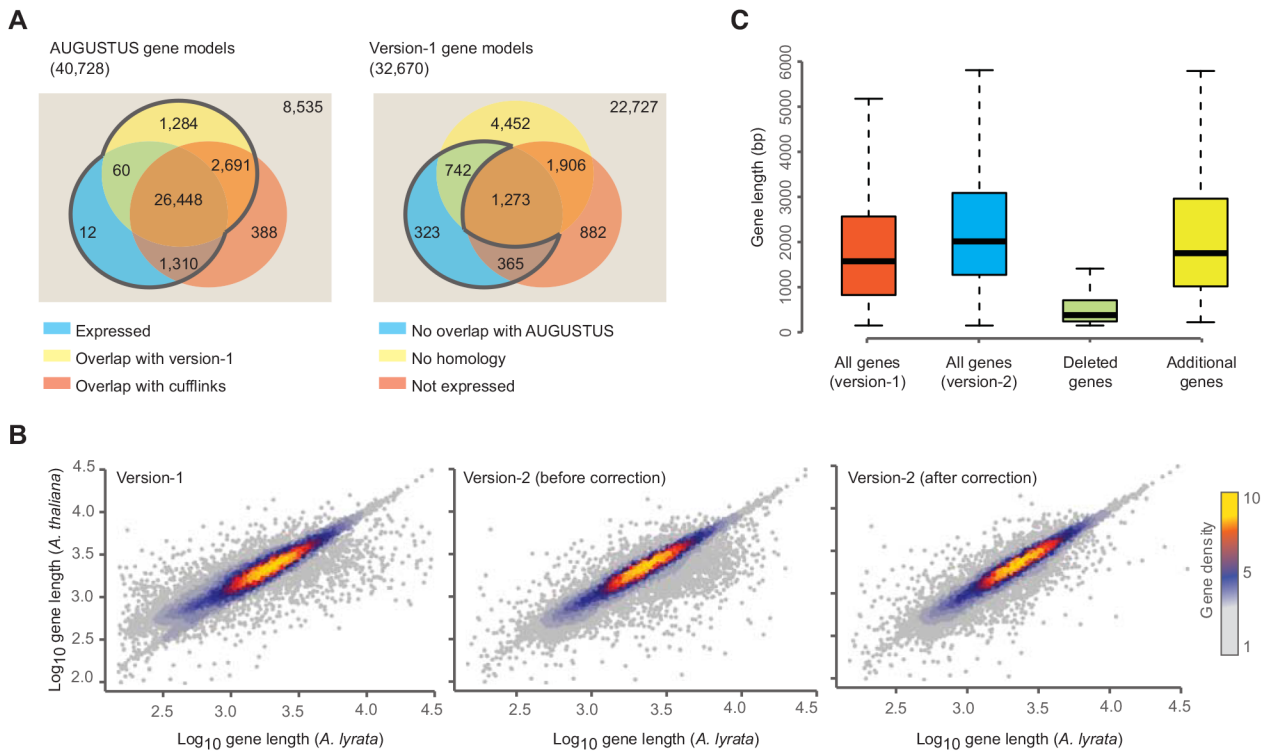
doi:10.1371/journal.pone.0137391.t001

patterns for spliced transcripts indicating instances of unidentified gene models and cases where transcription exceeded known gene boundaries (see Fig A in [S1 Dataset](#)).

New gene models were predicted from short read alignment data using Cufflinks 2.1.1 [22] independently for each tissue. In total, Cufflinks predicted 31,194 distinct gene models across all samples. An additional RNA-seq alignment-guided gene prediction using Augustus v.3.0.1 [28] identified 40,728 gene models, including 27,830 genes, which were supported by at least five RNA-seq reads. Moreover, 30,483 and 30,837 of Augustus predicted gene models overlapped with version-1 and Cufflinks predictions, respectively (see [Materials and Methods](#) and Fig B in [S1 Dataset](#)).

We combined 31,793 Augustus gene models with evidence of transcription or that overlapped with version-1 gene models to update the *A. lyrata* gene annotation (Fig 1A). To ensure that we were not excluding any true gene models in version-1, we included 1,430 version-1 gene models that were not overlapping with any of the new gene models, but showed either evidence of expression or featured an ortholog in at least one of the Brassicaceae species *A. thaliana* [29], *Capsella rubella* [4], *Brassica rapa* [10], *Schrenkiella parvula* [11] and *Arabis alpina* [5]. This increased the number of gene models to 33,223 (see [Materials and Methods](#)). To identify and to correct cases where incorrect gene models may have been introduced into the version-2 annotation, we utilized the very close phylogenetic relationship between *A. lyrata*, *A. thaliana* and *C. rubella*. We compared all gene models that were considerably different between version-1 and version-2 to *A. thaliana* and *C. rubella* orthologs (see [Materials and Methods](#)). If the length of the version-1 open reading frame was closer to that of the orthologs, we retained the version-1 gene model. This resulted in 548 version-2 gene models being replaced with 688 of the original version-1 gene models (Fig 1B). After additional removal of redundant gene models we obtained a final set of 33,221 non-redundant gene models.

Based on a recent annotation of *A. lyrata* TEs [14] and sequence similarity to TE genes of *A. thaliana* [25], we annotated 2,089 of the protein coding gene models as TE protein coding genes (see [Materials and Methods](#)). Without these, version-2 comprised of 31,132 gene models, which is ~13% more than in *A. thaliana* [25]. Although tRNA genes were described in the original analysis of the *A. lyrata* genome [1], version-1 lacks information regarding these loci. By



**Fig 1. Updating the gene model annotation of *A. lyrata*.** (A) Left, version-2 gene models predicted by Augustus [28]. Number of gene models overlapping with version-1 (yellow), genes predicted with Cufflinks (red), and genes with expression evidence (blue). Right, gene models of the version-1 annotation. Number of models without overlap to version-2 models (yellow), without orthologs in five other Brassicaceae (red), and without significant expression evidence (blue). (B) Correlation of the lengths of *A. lyrata* gene models with the length of their orthologous gene models in *A. thaliana*. Left, *A. lyrata* version-1 gene models. Correlations using version-1 gene models (left), version-2 gene models before (middle) and after (right) the homology-based correction of gene models. (C) Length distribution of gene models including genes that were removed or newly added in the version-2.

doi:10.1371/journal.pone.0137391.g001

rerunning tRNAScan [30], we identified 660 tRNA genes coding for all 20 amino acids. For completeness, we also incorporated 170 recently published miRNA genes into the new annotation file [31].

Altogether, we updated the coordinates of 29,141 of the original gene models, removed 1,286 entire (mostly short) gene models, and added 1,295 new models (Fig 1C). Only 2,243 remained unaltered (including 688 version-1 gene models re-introduced due to their superior similarity to orthologs). The new annotation accounted for 31,132 non-TE related gene models including 27,084 multi-exonic genes of which 2,236 featured at least one alternative isoform (Table 2).

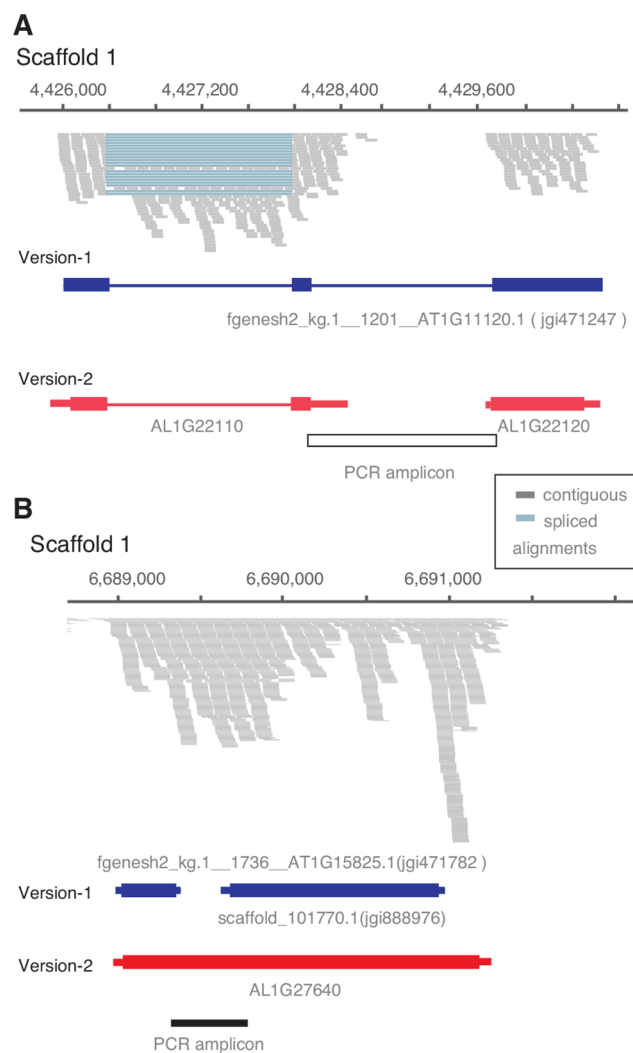
**Table 2. Comparison of version-1 and version-2 annotations.**

	# version-1	# version-2
Gene models	32,670	33,221
Predicted transcripts	32,670	35,805
Protein coding genes	32,670	31,132
TE coding genes	-	2,089
miRNA genes	-	170
tRNA genes	-	660
Featuring ortholog (in at least one Brassicaceae)	23,996	24,146

doi:10.1371/journal.pone.0137391.t002

### Validating differences in gene model structure

Even after the above-mentioned homology-based gene length adjustments, we found cases where the corresponding gene models from the two annotations varied drastically in length. This included instances where multiple version-1 gene models were fused to form a single gene model in version-2 or vice versa (Fig 2). In total, 161 version-1 genes were split (accounting for 530 genes in version-2) and 1,729 version-1 gene models were merged (accounting for 775 gene models in version-2). We randomly selected 14 version-1 gene models that had been split into multiple gene models in version-2, and another 14 gene models that had been merged in version-2, for PCR validation (see Fig C and D in S1 Dataset). One split case did not yield gDNA bands indicating a technical problem in primer design. For three merge cases we obtained cDNA bands of the expected size, but were not able to amplify genomic DNA for primer validation. This was most likely due to large gDNA amplicon size (2.4–5 kbp) and



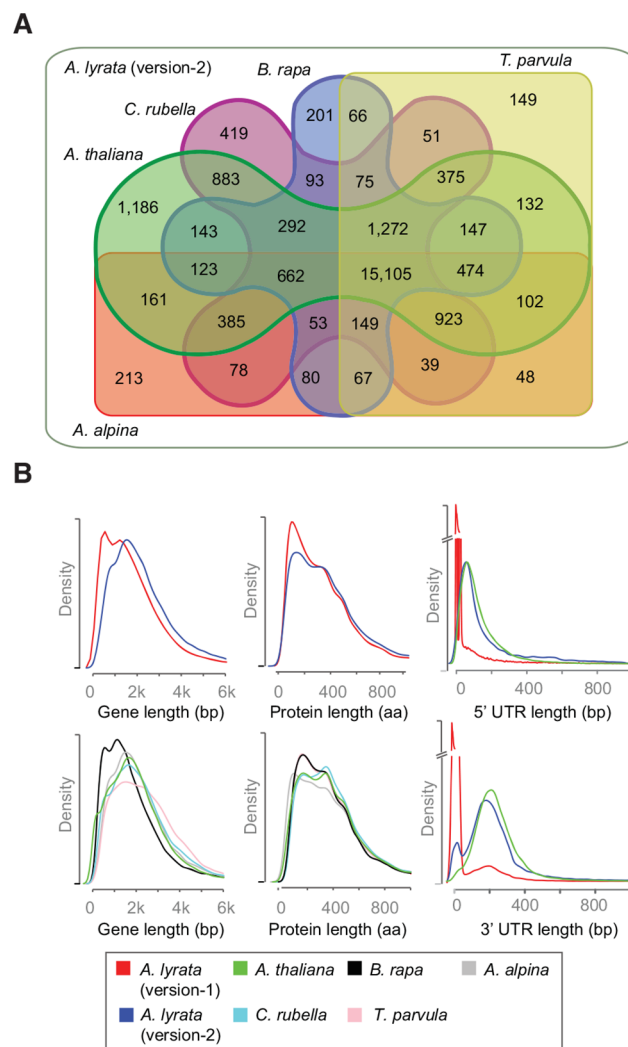
**Fig 2. Examples of version-1 gene models split and merged in *A. lyrata* gene annotation version-2. (A)** Example of a gene model that was split into two gene models in version-2. Reverse transcription-PCR could not confirm the connection of both. **(B)** Example of version-1 gene models that were merged during the annotation update. Reverse transcription-PCR confirmed presence of a transcript bridging the two version-1 genes.

doi:10.1371/journal.pone.0137391.g002

rendered the results of these cases inconclusive. For all 24 remaining cases, PCR results fully confirmed the annotation of the new gene models.

### *A. lyrata* version-2 annotation in contrast to other Brassicaceae annotations

For both *A. lyrata* annotations we predicted orthologous relationships between *A. lyrata* and five other Brassicaceae species (see [Materials and Methods](#)). Using version-2 gene models, 77.5% of genes had an ortholog in at least one species (24,146 out of 31,132) ([Fig 3A](#)), compared to 73% for version-1 (23,996 out of 32,670) (see Fig E in [S1 Dataset](#)). The number of genes with orthologs in all five Brassicaceae was also slightly higher for version-2 with 15,105 genes versus 14,850 genes with version-1.



**Fig 3. Comparing the *A. lyrata* gene annotation version-2 with the annotations of five other Brassicaceae.** (A) Orthologous gene models shared between *A. lyrata* (version-2), *A. thaliana* [29], *A. alpina* [5], *B. rapa* [10], *C. rubella* [4] and *S. parvula* [11]. (B) Gene, Protein and UTR length distributions of above-mentioned species including the new and old *A. lyrata* annotations. UTR distribution is only shown for *A. lyrata* and *A. thaliana* because of poor UTR annotation in some of the other species.

doi:10.1371/journal.pone.0137391.g003

The removal of many short gene models in version-2 changed the distribution of gene model lengths (Figs 1C and 3B). Version-1 has an excess of gene models shorter than 1 kb with a second mode around 1.5 kb, which describes a bimodal distribution that was only reflected by gene length distribution of *B. rapa*. In contrast version-2 had only a single mode around 1.7 kb, similar to the four other species. The length distribution of predicted protein sequences in version-1 had also been distinct from the other Brassicaceae species, and this discrepancy largely disappeared with version-2. A third factor that contributed to the length differences between the genes of version-1 and version-2 were differences in UTR annotations (Fig 3B). In version-1 33% of the genes were annotated without UTR information, however, in version-2 only 5% remained without 3' and 5' UTR annotation. The absolute and relative contributions of individual features are shown in Fig F in S1 Dataset. Though, absolute increase in genomic space for all gene features was observed but CDS and UTRs are benefited the most. We also observed little decrease in intronic genome space, which can be explained by introduction of splice variants previously missing from version-1 annotation.

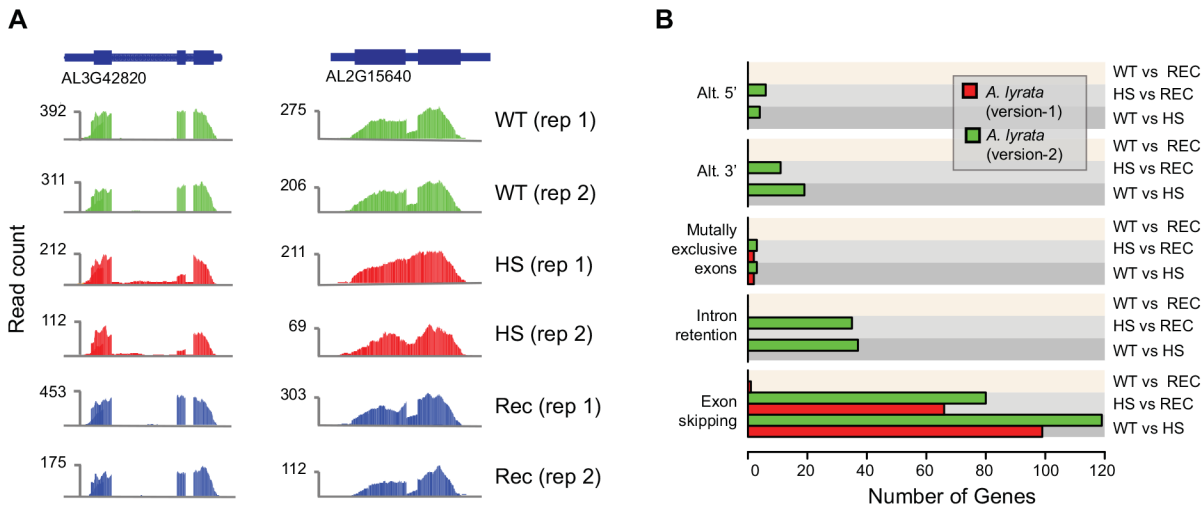
Whether the bimodal distribution in *B. rapa* reflects similar ambiguity in gene annotations, or mirrors particular characteristics of *B. rapa*, including its ancient genome triplication and subsequent fractionation, is not known.

## New annotation enabled improved identification of alternative splicing events

The availability of multiple isoforms from individual gene models in version-2 enables quantitative expression comparisons between annotated isoforms. We analyzed RNA-seq data from *A. lyrata* rosette tissues from untreated (WT), heat stressed (HS), and recovered (REC) samples in duplicate (see Materials and Methods). We first analyzed the data for differential gene expression using Cuffdiff v.2.0.2 [32]. WT and REC differed from HS at 3,114 and 2,962 genes, whereas only 106 genes differed between WT and REC. This indicates, as expected, a strong effect of heat stress on gene expression (see Materials and Methods). Cuffdiff was also used to estimate differential expression between isoforms. We identified differential isoform expression at 283, 15 and 119 genes when comparing WT with HS, WT with REC, and HS with REC, respectively. In contrast, as version-1 does not include different isoforms, which are prerequisite for isoform expression analysis as implemented in Cuffdiff, it was not possible run this analysis using version-1.

We investigated differential splicing using a second tool, MATS v3.0.8 [33], which does not rely on prior isoform annotations and only identifies differences in individual splicing events, but not between entire transcripts. With version-2, MATS identified 177, 0 and 130 differential splicing events distributed over 187 distinct gene models in the three comparisons (Fig 4; see Materials and Methods). MATS reported only 99, 1 and 67 events affecting 103 gene models using version-1. The overlap of different splicing events was very high (95 out of 103 (version-1) and 187 (version-2) gene models). Thus, almost all gene models with differential splicing events predicted based on version-1 were also predicted using version-2, however, the results based on version-2 revealed many more gene models. This was partially due to newly added genes (10 cases), but the most improvement came from the updates to exon-intron boundaries of existing gene models indicating that the new gene annotation improved the overall utility of this resource.

The isoform-dependent (Cuffdiff) and-independent (MATS) analyses identified only 37 common gene models. Even though Cuffdiff revealed fewer events as compared to the MATS analysis, it did identify 100 genes with differential isoform usage that were not included in the set of genes with multiple isoforms. This suggests that differential isoform expression analysis profits from prior isoform annotation, however, should not only rely on existing isoforms.



**Fig 4. Heat stress induces alternative splicing events.** (A) Examples of differentially expressed isoforms in response to heat stress in *A. lyrata*. AL3G42820 expresses a second isoform that lacks the middle exon in heat-treated samples (HS). Transcripts from wild-type (WT) and recovery (REC) samples contain all three exons. AL2G15640 retains an intron in response to heat stress (HS) while wild-type (WT) and recovery (REC) samples show partial intron splicing. (B) Number of differential splicing events, including alternative 5' and 3' splice sites, mutually exclusive exons, intron retention, and exon skipping events identified with MATs based on version-1 and version-2 annotations.

doi:10.1371/journal.pone.0137391.g004

### Availability of the annotation and gene naming conventions

The version-2 annotation can be found in [S2 Dataset](#). The gene identifiers have been updated following the annotation principles applied to *A. thaliana* [29]. Gene numbering follows the physical order of genes on chromosomes, where each gene is named “AL” followed by scaffold number, then a “G” (for the first 8 scaffolds corresponding to the eight chromosomes) or “U” (for unanchored scaffolds) and finally a unique number incremented by 10, to leave flexibility for genes that were missed in this annotation. Genes that were removed from version-1 can be found in [S3 Dataset](#). A mapping of the gene model identifiers of version-1 to version-2 can be found in [S4 Dataset](#).

### Conclusions

The updated annotation includes 31,132 gene models with 35,805 transcripts. We also reported 1,304 gene models that were erroneously split or merged in the previous annotation. Validation of these models strongly supported our updates highlighting the importance of employing species-specific RNA-seq data for annotating genomes.

We also provided a first annotation of alternative splicing events in *A. lyrata*. Using RNA-seq samples for a heat stress experiment we demonstrated the improved utility of the version-2 annotation for differential isoform expression studies. This revised genome annotation advances the reference sequence of *A. lyrata* as a community resource for comparative and functional studies.

### Material and Methods

#### Plant material

*Arabidopsis lyrata* subsp. *lyrata* MN47 plants were grown in soil under long day conditions (16 hours light, 21°C: 8 hours dark, 16°C). Vegetative rosettes and dissected shoot apices of three week old plants and entire inflorescences of flowering plants were harvested as mock treated



samples. For heat stress and recovery treatments we incubated three week old plants at 37°C for 6 hours or for an additional 48 hours at 21°C, respectively. Cold stressed samples were treated as described [6].

## Nucleic acids isolation and RNA-seq library preparation

DNA was isolated using Nucleon Phytopure kit (GE Healthcare). For total RNA isolation, samples were flash frozen in liquid nitrogen and used with Qiagen RNeasy<sup>®</sup> Plant Mini Kit, including an on-column DNase I digestion. Total RNA integrity was confirmed on the Agilent BioAnalyzer. Barcoded libraries were constructed using the Illumina TruSeq RNA kit with average of 1 µg of total RNA as starting material. The manufacturer's protocol was precisely followed with one exception in the cold-treated samples where 12 PCR cycles were used instead of the recommended 15. The library quality was monitored on a Bioanalyzer 2100 (Agilent) and the libraries were sequenced as 100-bp single end reads using Illumina sequencing.

## RNA-seq read mapping and gene predictions

RNA-seq data was mapped to the *A. lyrata* reference genome assembly [1] using Bowtie v1.0.0 [26] and TopHat v2.0.10 [27]. Cufflinks v2.0.2 [22] was used for *de novo* transcript identification in all tissues separately. Cuffmerge (from the Cufflinks suite) was used to merge transcript annotation files obtained for three tissues separately. In addition, all short reads were aligned to the reference assembly of *A. lyrata* using BLAT v.34 [34] to generate an evidence file for guided gene prediction using Augustus v3.0.0. *A. lyrata* specific configuration file was generated using the version-1 annotation. To estimate agreement between Augustus and version-1 gene models, gene models with  $\geq 30\%$  overlap (in respect to the shorter gene model) were considered. Gene models supported by five or more RNA-seq reads were considered as expressed irrespective of gene length.

To identify cases where wrong gene models were introduced in version-2, we first compared version-2 proteins (23,181 comparable proteins) with corresponding version-1 proteins. A total of 1,037 proteins were identified as outliers, where protein length difference was outside the range of  $\pm$  standard deviation of the distribution of length differences. For these cases version-1 and version-2 protein sequences were further compared against the proteins of their orthologs in *A. thaliana* [29] and *C. rubella* [4]. If both orthologs were more similar in length to the protein of version-1, the respective version-2 gene model was replaced with version-1.

## Ortholog identification

Orthologous gene identification for both version-1 and version-2 was done separately at protein level using reciprocal best hits using blastall v2.2.25 [35] and an e-value cutoff 0.001 among five Brassicaceae species.

## Identification of TE genes in version-2

Version-2 gene models harboring complete TEs [14] within their coding regions or were entirely spanned by a TE were annotated as "TE coding genes". In addition 3,909 *A. thaliana* TE genes [25] and TIGR Brassicaceae specific repeat database [36] were used to identify TE genes using blastn v2.2.25 [35].

## cDNA preparation and PCR

Plants were grown on soil under long day conditions until the five-leaf stage reached after approximately three weeks. cDNA samples were prepared from 1 µg total RNA of mock-

treated rosettes using RevertAid First Strand cDNA Synthesis Kit with oligo d(T) primers (Thermo Scientific). Reverse transcriptase minus samples were processed in the same way without enzyme addition. PCR reactions were done in an Eppendorf thermal cycler using a standard program and the products were visualized on agarose gels stained with ethidium bromide. The PCR primer sequences can be found in [S5 Dataset](#).

## Differential gene expression and alternative splicing

Cufflinks [22] was used to calculate differential gene expression level (FPKM) with p-value < 0.01 and log<sub>2</sub>-fold change difference of more than 2. MATS [33] was used to investigate differential splicing events with over 0.01% splicing difference at a p-value < 0.01 and a false discovery rate of less than 1%. To control for false positives, genes with 10,000 fold or more expression difference were excluded.

## Supporting Information

### S1 Dataset. Supplementary figures.

(DOCX)

### S2 Dataset. General feature formatted (GFF) file describing version-2 annotation.

(ZIP)

### S3 Dataset. GFF file describing genes that were removed from version-1.

(ZIP)

### S4 Dataset. Table describing the mapping of version-1 to version-2 gene models.

(XLSX)

### S5 Dataset. Primer information for gene model validation.

(XLSX)

## Acknowledgments

We thank Geo James Velikkakam for helpful discussion on gene annotation tools, J. de Meaux for seeds of *A. lyrata*, B. Eilts and R. Gentges for technical support, A. Platts, S. Wright and M. Blanchette for insights in their gene annotation and annotation of conserved non-coding sequences, and M. Koornneef for critical reading of the manuscript.

## Author Contributions

Conceived and designed the experiments: KS AP. Performed the experiments: AP AA. Analyzed the data: VR. Contributed reagents/materials/analysis tools: AA BP DKS DK DW. Wrote the paper: VR KS AA AP BP DKS DK DW.

## References

1. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* 2011, May; 43(5):476–81. doi: [10.1038/ng.807](#) PMID: [21478890](#)
2. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, et al. Evolution of genome size in brassicaceae. *Ann Bot* 2005; 95(1):229–35. PMID: [15596470](#)
3. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. The dynamic ups and downs of genome size evolution in brassicaceae. *Mol Biol Evol* 2009, Jan; 26(1):85–98. doi: [10.1093/molbev/msn223](#) PMID: [18842687](#)

4. Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, et al. The capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* 2013, Jul; 45(7):831–5. doi: [10.1038/ng.2669](https://doi.org/10.1038/ng.2669) PMID: [23749190](https://pubmed.ncbi.nlm.nih.gov/23749190/)
5. Willing E-M, Rawat V, Mandáková T, Maumus F, James GV, Nordström KJ, et al. Genome expansion of *arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants* 2015; 1(2).
6. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. Evolution of DNA methylation patterns in the brassicaceae is driven by differences in genome organization. *PLoS Genet* 2014, Nov; 10(11): e1004785. doi: [10.1371/journal.pgen.1004785](https://doi.org/10.1371/journal.pgen.1004785) PMID: [25393550](https://pubmed.ncbi.nlm.nih.gov/25393550/)
7. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, et al. Reference-guided assembly of four diverse *arabidopsis thaliana* genomes. *Proc Natl Acad Sci U S A* 2011, Jun 21; 108(25):10249–54. doi: [10.1073/pnas.1107739108](https://doi.org/10.1073/pnas.1107739108) PMID: [21646520](https://pubmed.ncbi.nlm.nih.gov/21646520/)
8. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple *arabidopsis thaliana* populations. *Nat Genet* 2011, Aug 28; 43(10):956–63. doi: [10.1038/ng.911](https://doi.org/10.1038/ng.911) PMID: [21874002](https://pubmed.ncbi.nlm.nih.gov/21874002/)
9. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in *arabidopsis thaliana* lines from sweden. *Nat Genet* 2013, Jun 23; 45(8):884–90. doi: [10.1038/ng.2678](https://doi.org/10.1038/ng.2678) PMID: [23793030](https://pubmed.ncbi.nlm.nih.gov/23793030/)
10. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, et al. The genome of the mesopolyploid crop species *brassica rapa*. *Nat Genet* 2011, Aug 28; 43(10):1035–9. doi: [10.1038/ng.919](https://doi.org/10.1038/ng.919) PMID: [21873998](https://pubmed.ncbi.nlm.nih.gov/21873998/)
11. Dassanayake M, Oh D-H, Haas JS, Hernandez A, Hong H, Ali S, et al. The genome of the extremophile crucifer *thellungiella parvula*. *Nat Genet* 2011, Aug 7; 43(9):913–8. doi: [10.1038/ng.889](https://doi.org/10.1038/ng.889) PMID: [21822265](https://pubmed.ncbi.nlm.nih.gov/21822265/)
12. Wu H-J, Zhang Z, Wang J-Y, Oh D-H, Dassanayake M, Liu B, et al. Insights into salt tolerance from the genome of the *thellungiella salsuginea*. *Proc Natl Acad Sci U S A* 2012, Jul 9; 109(30):12219–24. doi: [10.1073/pnas.1209954109](https://doi.org/10.1073/pnas.1209954109) PMID: [22778405](https://pubmed.ncbi.nlm.nih.gov/22778405/)
13. Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, et al. The reference genome of the halophytic plant *eutrema salsugineum*. *Front Plant Sci* 2013; 4:46. doi: [10.3389/fpls.2013.00046](https://doi.org/10.3389/fpls.2013.00046) PMID: [23518688](https://pubmed.ncbi.nlm.nih.gov/23518688/)
14. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* 2013, Aug; 45(8):891–8. doi: [10.1038/ng.2684](https://doi.org/10.1038/ng.2684) PMID: [23817568](https://pubmed.ncbi.nlm.nih.gov/23817568/)
15. Cheng S, van den Bergh E, Zeng P, Zhong X, Xu J, Liu X, et al. The *tarenaya hassleriana* genome provides insight into reproductive trait and genome evolution of crucifers. *Plant Cell* 2013, Aug; 25(8):2813–30. doi: [10.1105/tpc.113.113480](https://doi.org/10.1105/tpc.113.113480) PMID: [23983221](https://pubmed.ncbi.nlm.nih.gov/23983221/)
16. Oh D-H, Hong H, Lee SY, Yun D-J, Bohnert HJ, Dassanayake M. Genome structures and transcriptomes signify niche adaptation for the multiple-ion-tolerant extremophyte *schrenkiella parvula*. *Plant Physiol* 2014, Apr; 164(4):2123–38. doi: [10.1104/pp.113.233551](https://doi.org/10.1104/pp.113.233551) PMID: [24563282](https://pubmed.ncbi.nlm.nih.gov/24563282/)
17. Kitashiba H, Li F, Hirakawa H, Kawanabe T, Zou Z, Hasegawa Y, et al. Draft sequences of the radish (*raphanus sativus* L.) Genome. *DNA Res* 2014, May 16; 21(5):481–90. doi: [10.1093/dnares/dsu014](https://doi.org/10.1093/dnares/dsu014) PMID: [24848699](https://pubmed.ncbi.nlm.nih.gov/24848699/)
18. Lobréaux S, Manel S, Melodelima C. Development of an *arabis alpina* genomic contig sequence data set and application to single nucleotide polymorphisms discovery. *Mol Ecol Resour* 2014, Mar; 14(2):411–8. doi: [10.1111/1755-0998.12189](https://doi.org/10.1111/1755-0998.12189) PMID: [24128264](https://pubmed.ncbi.nlm.nih.gov/24128264/)
19. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, et al. The brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun* 2014; 5:3930. doi: [10.1038/ncomms4930](https://doi.org/10.1038/ncomms4930) PMID: [24852848](https://pubmed.ncbi.nlm.nih.gov/24852848/)
20. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in *arabidopsis*. *Cell* 2008, May 2; 133(3):523–36. doi: [10.1016/j.cell.2008.03.029](https://doi.org/10.1016/j.cell.2008.03.029) PMID: [18423832](https://pubmed.ncbi.nlm.nih.gov/18423832/)
21. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, Jan; 10(1):57–63. doi: [10.1038/nrg2484](https://doi.org/10.1038/nrg2484) PMID: [19015660](https://pubmed.ncbi.nlm.nih.gov/19015660/)
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010, May; 28(5):511–5. doi: [10.1038/nbt.1621](https://doi.org/10.1038/nbt.1621) PMID: [20436464](https://pubmed.ncbi.nlm.nih.gov/20436464/)
23. Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* 2011; 12:540. doi: [10.1186/1471-2164-12-540](https://doi.org/10.1186/1471-2164-12-540) PMID: [22047402](https://pubmed.ncbi.nlm.nih.gov/22047402/)

24. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res* 2010, Jan; 20(1):45–58. doi: [10.1101/gr.093302.109](https://doi.org/10.1101/gr.093302.109) PMID: [19858364](https://pubmed.ncbi.nlm.nih.gov/19858364/)
25. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The *Arabidopsis* information resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res* 2012, Jan; 40(Database issue):D1202–10. doi: [10.1093/nar/gkr1090](https://doi.org/10.1093/nar/gkr1090) PMID: [22140109](https://pubmed.ncbi.nlm.nih.gov/22140109/)
26. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012, Apr; 9(4):357–9. doi: [10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) PMID: [22388286](https://pubmed.ncbi.nlm.nih.gov/22388286/)
27. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with *RNA*-seq. *Bioinformatics* 2009, May 1; 25(9):1105–11. doi: [10.1093/bioinformatics/btp120](https://doi.org/10.1093/bioinformatics/btp120) PMID: [19289445](https://pubmed.ncbi.nlm.nih.gov/19289445/)
28. Stanke M, Waack S. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics* 2003, Oct; 19(Suppl 2):215–25.
29. The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000; 408(6814):796–815. PMID: [11130711](https://pubmed.ncbi.nlm.nih.gov/11130711/)
30. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997, Mar 1; 25(5):955–64. PMID: [9023104](https://pubmed.ncbi.nlm.nih.gov/9023104/)
31. Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, Laubinger S, et al. MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. *Plant Cell* 2010, Apr; 22(4):1074–89. doi: [10.1105/tpc.110.073999](https://doi.org/10.1105/tpc.110.073999) PMID: [20407027](https://pubmed.ncbi.nlm.nih.gov/20407027/)
32. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of *RNA*-seq experiments with Tophat and cufflinks. *Nat Protoc* 2012, Mar; 7(3):562–78. doi: [10.1038/nprot.2012.016](https://doi.org/10.1038/nprot.2012.016) PMID: [22383036](https://pubmed.ncbi.nlm.nih.gov/22383036/)
33. Shen S, Park JW, Huang J, Dittmar KA, Lu Z-X, Zhou Q, et al. MATS: A bayesian framework for flexible detection of differential alternative splicing from *RNA*-seq data. *Nucleic Acids Res* 2012, Apr; 40(8):e61. doi: [10.1093/nar/gkr1291](https://doi.org/10.1093/nar/gkr1291) PMID: [22266656](https://pubmed.ncbi.nlm.nih.gov/22266656/)
34. Kent WJ. BLAT—the blast-like alignment tool. *Genome Res* 2002; 12(4):656–64. PMID: [11932250](https://pubmed.ncbi.nlm.nih.gov/11932250/)
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990, Oct 5; 215(3):403–10. PMID: [2231712](https://pubmed.ncbi.nlm.nih.gov/2231712/)
36. Ouyang S, Buell CR. The TIGR plant repeat databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* 2004, Jan 1; 32(Database issue):D360–3. PMID: [14681434](https://pubmed.ncbi.nlm.nih.gov/14681434/)