Pucker *et al. BMC Res Notes (2017) 10:667*
https://doi.org/10.1186/s13104-017-2985-y

BMC Research Notes

RESEARCH NOTE

Open Access

Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence

Boas Pucker, Daniela Holtgräwe and Bernd Weisshaar*

Abstract

Objective: The *Arabidopsis thaliana* Niederzenz-1 genome sequence was recently published with an ab initio gene prediction. In depth analysis of the predicted gene set revealed some errors involving genes with non-canonical splice sites in their introns. Since non-canonical splice sites are difficult to predict ab initio, we checked for options to improve the annotation by transferring annotation information from the recently released Columbia-0 reference genome sequence annotation Araport11.

Results: Incorporation of hints generated from Araport11 enabled the precise prediction of non-canonical splice sites. Manual inspection of RNA-Seq read mapping and RT-PCR were applied to validate the structural annotations of non-canonical splice sites. Predictions of untranslated regions were also updated by harnessing the potential of Araport11's information, which was generated by using high coverage RNA-Seq data. The improved gene set of the Nd-1 genome assembly (GeneSet_Nd-1_v1.1) was evaluated via comparison to the initial gene prediction (GeneSet_Nd-1_v1.0) as well as against Araport11 for the Col-0 reference genome sequence. GeneSet_Nd-1_v1.1 contains previously missed non-canonical splice sites in 1256 genes. Reciprocal best hits for 24,527 (89.4%) of all nuclear Col-0 genes against the GeneSet_Nd-1_v1.1 indicate a high gene prediction quality.

Keywords: Genome annotation, Splicing, Araport11, Gene prediction hints, Reciprocal best hit

Introduction

Eukaryotic genes are transcribed as a primary transcript that is subsequently converted to a mature mRNA through several processing steps including splicing. During splicing, introns [1–3] are removed from the primary transcript while exons are retained. The process is catalyzed by a RNA protein complex called a spliceosome, which exists in several variants. Based on the spliceosome variant that acts on a given intron, eukaryotic introns are classified as U2-type introns [4] that appear very frequently, or rare U12-type introns [5], respectively [6]. The highly conserved sequences at the termini of introns are not sufficient to distinguish between both types, since the U12-spliceosome can remove AT-AC introns, some other non-canonical intron variants, as well as some introns of the canonical GT-AG type [6–9]. Canonical GT-AG and non-canonical intron variants including AT-AC introns can coexist within the same gene, potentially with an effect on gene expression due to the slow removal of U12-type introns [10]. Several extremely rare terminal intron sequences were discovered and often discussed as potential artifacts, e.g. introns with GT-GG or TT-AG termini [11–14]. Further details regarding exceptional splicing events have recently been reviewed [15, 16].

Splicing processes were investigated intensively in the plant model system *Arabidopsis thaliana* [17–22], resulting in very well annotated splice sites throughout the reference genome sequence [23]. Despite attempts

*Correspondence: bernd.weisshaar@uni-bielefeld.de
Faculty of Biology & Center for Biotechnology, Bielefeld University, Bielefeld, Germany

to annotate non-canonical splice sites automatically [24, 25], ab initio gene prediction without experimental support from e.g. RNA-Seq data ("external hints") does not support the detection and annotation of non-canonical splice sites on genome sequence assemblies at a satisfying level [26–28]. By generating high quality gene prediction hints based on the recently released Araport11 annotation of the Col-0 sequence [29, 30], we improved the gene set generated by ab initio gene prediction based on the *A. thaliana* Niederzenz-1 (Nd-1) sequence [31].

To correlate and compare gene structures from related genomes, the first step is to define "orthologous" gene couples. Such couples can efficiently be determined by evaluating reciprocal best BLAST hits (RBHs) [32–35]. Each RBH couple consists of two genes, one from each of the two genome sequences (or genomes) to compare, which display the highest scoring hit in the other data set in a reciprocal manner [36]. RBH couples are the basis for gene-centric comparative genomics [32–35] and can also be used for synteny analysis or as guidance in a genome assembly [31].

Main text

Methods

Analysis of candidate genes

In total, 45 randomly selected Col-0 genes with non-canonical splice sites were manually inspected in a RNA-Seq read mapping produced with STAR [37] based on Araport11 data sets (listed in [30]). Reads were required to map with at least 90% of their length and 95% similarity. The number of selected cases was a compromise between the required accuracy of the results and a manageable amount for individual manual inspection. Corresponding loci in the Nd-1 sequence were identified via tblastn [38]. Gene structures around non-canonical splice sites in the Nd-1 assembly sequence [31] were annotated manually for further investigation.

Primer combinations for RT-PCR included one primer bridging an exon–exon junction with 100–500 nt distance to the other primer (Table 1). Oligonucleotides were purchased from Metabion (http://www.metabion.com/). Total RNA was isolated as described before [39]. DNAse I (M0303L, New England Biolabs) digestion was performed according to the suppliers' protocol. cDNA synthesis was carried out using 1 µg of total RNA and ProtoScript II Reverse Transcriptase (M0368L, New England Biolabs) based on the suppliers' protocol. Q5 High-Fidelity DNA polymerase (M0491L, New England Biolabs) was employed according to the suppliers' recommendations (including PCR cycling conditions) for generation of amplicons. The size of the amplicons was checked by agarose gel electrophoresis. Samples were purified for sequencing by ExoSAP-IT (78201.1.ML ThermoFisher Scientific) treatment as described [40]. Sanger sequencing on ABI3730XL was applied to reveal the entire sequences as described [41]. Finally, the correct annotation of the non-canonical splice sites in the candidate genes was inspected via sequence alignments generated with MAFFT [42].

Hint-based gene prediction

All representative transcript sequences of protein coding genes in the Col-0 nucleome within the Araport11 annotation, as well as the first transcripts of At4g01800 and At3g10350, were mapped to the Nd-1 genome sequence via BLAT [43]. Perl scripts provided in the AUGUSTUS package filterPSL.pl and blat2hints.pl (http://bioinf.uni-greifswald.de/augustus/binaries/scripts/) were used to convert the BLAT output into valid hints. AUGUSTUS 3.2.1 [44, 45] was run on the Nd-1 genome sequence incorporating these hints.

Table 1 The oligonucleotides listed were applied in RT-PCRs to validate non-canonical splice sites selected candidate genes in Nd-1

Name	Gene	Sequence	Length	Orientation	Recommended annealing temperature [°C]
S015	At1g79350 (*FGT1*)	GCTTCCCTGGAGTGCTGATCG	21	Forward	61
S016	At1g79350 (*FGT1*)	TCGGGTTCATCAATCGAGCATCC	23	Reverse	61
S017	At1g79350 (*FGT1*)	AAGAACAGGTAGTTTCTCCTGCTCC	25	Reverse	60
S003	At4g01800 (*AGY1*)	ACTGGTGAAGGGAAAACGCTTG	22	Forward	59
S004	At4g01800 (*AGY1*)	AATGTATATCCCGCTCAAAGGCTG	24	Reverse	59
S005	At4g01800 (*AGY1*)	TCTTCTGCTTTTCATCAACAGTGTAATG	28	Reverse	58
S018	At4g27500 (*PPI1*)	AGCCGCAGAAGGAAGAAAAGC	21	Forward	59
S019	At4g27500 (*PPI1*)	ACGCGATGAGACGAATTCCGAG	22	Forward	61
S020	At4g27500 (*PPI1*)	CTCTTGGGATCGTTTCTGGTCC	22	Reverse	59

Comparison of gene predictions

Calculation of gene prediction statistics as well as comparison to the Col-0 annotation via identification of RBHs was carried out by custom Python scripts as previously described [31]. ParsEval [46] was applied to compare the GeneSet_Nd1_v1.0 and GeneSet_Nd1_v1.1 in more detail.

Results and discussion

When analyzing the protein coding genes predicted in the recently released *A. thaliana* Nd-1 genome sequence [31], we observed complete absence of introns with non-canonical splice sites in the initially predicted gene set (GeneSet_Nd-1_v1.0). The structural annotation was performed ab initio using AUGUSTUS 3.2. By comparing the GeneSet_Nd-1_v1.0 with the Araport11 gene set for the Col-0 reference genome sequence [23, 29, 30], we identified several loci with gene structures showing mis-annotated introns or even a lack of gene prediction for the Nd-1 case. For the present study, we focused on protein encoding genes in the nuclear genome sequence since this gene set was previously predicted ab inito. The annotation update provided here will further support *A. thaliana* pan-genomic research by redefining the gene set for the accession Nd-1. Moreover, researchers interested in single genes and their Nd-1 alleles will be able to access a high quality annotation for comparison to Araport11 for the Col-0 reference sequence.

In total, the Araport11 gene set contains 1267 genes which display non-canonical splice sites to generate the respective representative transcript. This 'representative transcript' has been defined as the transcript isoform containing the longest protein coding sequence (CDS) [30]. We established a set of well investigated genes consisting of At1g79350 (*FGT1*) [47–49], At4g01800 (*AGY1*) [47, 50–52] and At4g27500 (*PPI1*) [53–57] as examples for genes containing confirmed introns with non-canonical slice sites in their main transcript isoform. Despite high sequence conservation between Col-0 and Nd-1, the gene structures predicted at these loci in GeneSet_Nd-1_v1.0 did not match the Araport11 annotation [29, 30], indicating that *bona fide* genes were missed by ab initio annotation of the Nd-1 genome sequence because they contain introns with non-canonical splice sites (Fig. 1).

When analyzing the Araport11 data set of Col-0 protein coding nuclear genes, which is based on very high coverage RNA-Seq information, we identified 39 different pairs of splice donor and splice acceptor sites (i.e. intron types) that need removal in order to generate the representative transcript isoforms. In total, the Araport11 structural annotation dataset contains 119,097 splice site pairs (introns) in nuclear protein coding genes that are spliced out of the primary transcript to produce the representative transcript. Of these, 117,732 (98.9%) were canonical GT-AG splice site pairs, while 1196 (1.0%) were GC-AG pairs and 81 (0.1%) were AT-AC pairs. In addition, diverse and less frequent splice site pairs sum up to 88 (0.1%) cases. These less frequent splice site pairs occur with very low frequencies and case numbers between one and nine.

When considering all transcript isoforms of all genes annotated in Araport11, 125 different splice site pairs are annotated. Obviously, non-protein coding genes contribute a huge proportion to splice site variation. Despite the very high quality of the *A. thaliana* Col-0 reference sequence, sequencing errors or collapsed gene sequences [58] could explain at least a fraction of the very rare splice site pairs [11].

Representative structures of protein encoding genes from Araport11 were used to produce gene prediction hints for the Nd-1 genome sequence (see "Methods"). This information transfer was done to harness the improvement potential of 1267 annotated protein

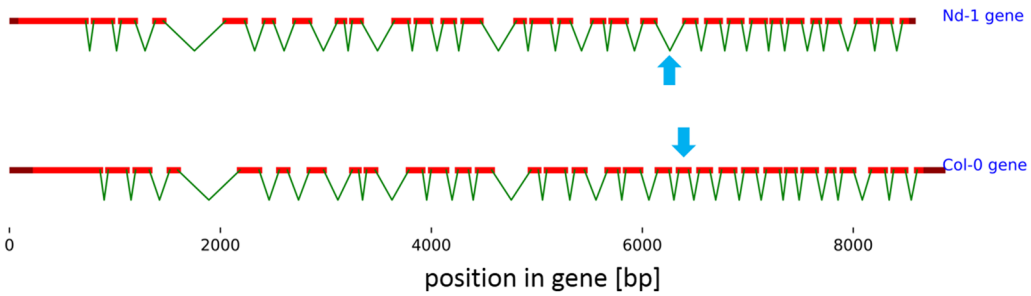

Fig. 1 Representative gene structure of missed non-canonical splice sites in ab initio gene prediction on the Nd-1 genome sequence. Gene structures of *At1g79350*.1 and the corresponding reciprocal best BLAST hit (RBH) of the ab initio gene prediction in Nd-1 (GeneSet_Nd-1_v1.0) are displayed. The non-canonical splice sites were missed leading to a difference at exon 20 (blue arrows). Despite this deviation, the structure of $At1g79350^{Nd-1}$ was predicted very well by AUGUSTUS [44, 45]

encoding genes in the Col-0 reference sequence containing various non-canonical splice sites in their representative transcript. Gene prediction on the Nd-1 genome sequence using these hints revealed 30,834 genes (GeneSet_Nd-1_v1.1, Additional file 1) exceeding the number of predicted genes in the GeneSet_Nd-1_v1.0 by 2164. Detailed comparison revealed a match of 91.2% in respect to predicted CDS features and a match of 50.2% concerning UTR features, respectively. Vast changes in the UTR prediction could be explained by the incorporated hints, since the ab initio prediction of these regions is error-prone. A slight reduction in the average CDS length from 1086 bp (median) in the GeneSet_Nd-1_v1.0 compared to an average length of 1041 bp (median) in the GeneSet_Nd-1_v1.1 was observed. There are 135,356 introns with 30 different pairs of donor and acceptor splice sites in the GeneSet_Nd-1_v1.1 (Additional file 2), supporting the assumption that some minor splice sites in the Araport11 annotation might be due to sequencing errors [11]. Splice site pairs were distinguished into 134,004 (99.0%) GT-AG splice site pairs, 1080 (0.8%) GC-AG splice site pairs, 66 (0.05%) AT-AC splice site pairs and 206 (0.15%) diverse and less frequent splice site pairs. In total, 1256 genes within the GeneSet_Nd-1_v1.1 contain introns with non-canonical splice sites. Their average transcript length is 2003 bp (median) consisting on average of ten protein encoding exons. Compared to the average number of four annotated exons in all genes of GeneSet_Nd-1_v1.1, we see a clear accumulation of non-canonical splice sites in exon-rich transcripts. This overrepresentation of exon-rich transcripts among the non-canonically spliced transcripts is supported by the Araport11 annotation where the average exon number of protein encoding transcripts with non-canonical splice sites is also ten. Manual inspection identified At4g01800 and At3g10350 as genes where the representative transcript in Araport11 does not require processing of non-canonical splice site pair, but another strongly expressed isoform does. Therefore, we expect the number of genes with non-canonical splice sites in Col-0 to be slightly higher than 1267 as deduced from the representative transcript data set.

Reciprocal best BLAST hit (RBH)-based comparison of the new GeneSet_Nd1_v1.1 and the Araport11 annotation revealed 24,527 gene couples (Additional file 3). The number of RBHs within the hint-based GeneSet_Nd1_v1.1 is strongly increased compared to the ab initio predicted GeneSet_Nd1_v1.0. We expect a further increase in prediction accuracy if the underlying sequence would be available with enhanced continuity, as for example possible if generated by SMRT sequencing, and if incorporation of additional hints from RNA-Seq data would be possible. High sensitivity mapping of Col-0 exon sequences to the Nd-1 genome sequence might discover small matches leading to further prediction improvements. Gene duplications are a special challenge in this process, because exon sequences might map to only one copy in the Nd-1 genome sequence. This might explain a part of the observed difference between the Col-0 annotation and the Nd-1 gene prediction concerning the number of transcripts with non-canonical splice sites.

Non-canonical splice sites in the reciprocal best hits (RBHs) of the three candidate genes *FGT1*, *AGY1* and *PPI1* in the GeneSet_Nd1_v1.1 were confirmed by Sanger sequencing of amplicons generated from cDNA. *FGT1* contained 31 exons and displayed a GC-CT splice site pair in intron 20 (Fig. 2). *AGY1* contained 20 exons and displayed a GA-AG splice site pair in intron 4. *PPI1* contained 7 exons and displayed a GA-AG splice site pair in intron 6.

Limitations

Allowing an increased number of alternative splicing possibilities deviating from the GT-AG rule would render ab initio prediction of gene structures almost impossible. Since the number of non-canonical splice sites is low, the ratio of false positive predictions would strongly increase. Incorporation of evidence from RNA-Seq experiments or high quality annotations of related genome sequences

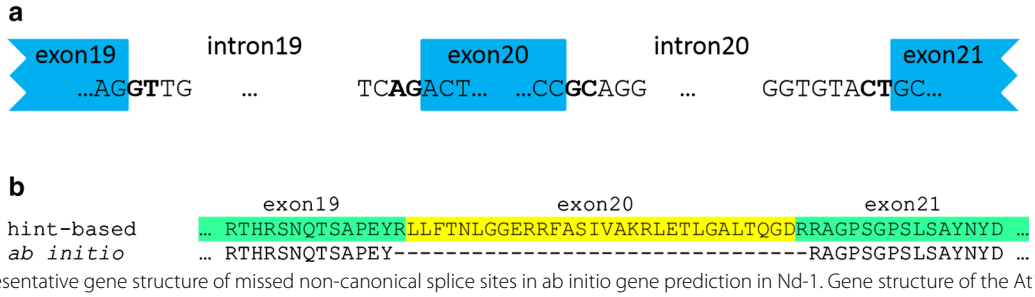

Fig. 2 Representative gene structure of missed non-canonical splice sites in ab initio gene prediction in Nd-1. Gene structure of the At1g79350 RBH in the hint-based gene prediction (GeneSet_Nd-1_v1.1) on the Nd-1 genome sequence is displayed (**a**). The non-canonical splice sites were missed in the ab initio gene prediction leading to a skipping of exon 20 (highlighted in yellow) (**b**)

into a gene prediction process with AUGUSTUS [44, 45] or a combination of AUGUSTUS and GeneMark [59] within BRAKER1 [60] is most probably the best way to achieve high quality gene predictions. Annotating new genome sequences via transfer of annotations from model species and adding additional expression data derived hints was successfully carried out several times before and has recovered many non-canonical splice sites [61–65]. Other promising approaches are completely based on homology to predict gene structures [66]. Nevertheless, the accurate prediction of non-canonical splice sites remains a challenge. Anyway, it will be a general contribution to accuracy to pay attention to non-canonical splice sites when applying ab initio gene prediction.

Additional files

Additional file 1. GeneSet_Nd1_v1.1. Gene prediction of the Nd-1 genome sequence containing genes with non-canonical splice sites.

Additional file 2. Non-canonical splice sites in Nd-1. All occurrences of the different splice site pairs within the first transcript of predicted Nd-1 genes in GeneSet_Nd1_v.1.1 are listed.

Additional file 3. Reciprocal Best Hits. Gene couples with reciprocal best hits between the Araport11 annotation of Col-0 and the GeneSet_Nd1_v1.1 are listed.

Authors' contributions

BP, DH and BW conceived and designed research. BP conducted experiments. BP, DH and BW interpreted the data. BP, DH and BW wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Katharina Kemmet for her help with the validation of non-canonical splice sites in Nd-1 genes. In addition, the authors wish to thank the members of the Genome Research Team at Bielefeld University as well as the Bioinformatics Resource Facility for their excellent assistance and support.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All data generated during this study are included in this published article and its additional files.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

We acknowledge the financial support of the German Research Foundation (DFG) and the Open Access Publication Fund of Bielefeld University for the article processing charge.

Publisher's Note

Received: 25 May 2017 Accepted: 23 November 2017
Published online: 04 December 2017

References

1. Gilbert W. Why genes in pieces? Nature. 1978;271(5645):501.
2. Kinniburgh AJ, Mertz JE, Ross J. The precursor of mouse beta-globin messenger RNA contains two intervening RNA sequences. Cell. 1978;14(3):681–93.
3. Breathnach R, Chambon P. Organization and expression of eukaryotic split genes coding for proteins. Ann Rev Biochem. 1981;50:349–83.
4. Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc Natl Acad Sci USA. 1978;75(10):4853–7.
5. Jackson IJ. A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. 1991;19(14):3795–8.
6. Dietrich RC, Incorvaia R, Padgett RA. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol Cell. 1997;1(1):151–60.
7. Hall SL, Padgett RA. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science. 1996;271(5256):1716–8.
8. Tarn WY, Steitz JA. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell. 1996;84(5):801–11.
9. Tarn WY, Steitz JA. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns. Science. 1996;273(5283):1824–32.
10. Patel AA, McCarthy M, Steitz JA. The splicing of U12-type introns can be a rate-limiting step in gene expression. EMBO J. 2002;21(14):3804–15.
11. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 2000;28(21):4364–75.
12. Dietrich RC, Peris MJ, Seyboldt AS, Padgett RA. Role of the 3′ splice site in U12-dependent intron splicing. Mol Cell Biol. 2001;21(6):1942–52.
13. Abril JF, Castelo R, Guigó R. Comparison of splice sites in mammals and chicken. Genome Res. 2005;15(1):111–9.
14. Niu X, Luo D, Gao S, Ren G, Chang L, Zhou Y, Luo X, Li Y, Hou P, Tang W, et al. A conserved unusual posttranscriptional processing mediated by short, direct repeated (SDR) sequences in plants. J Genet Genom. 2010;37(1):85–99.
15. Sharp PA, Burge CB. Classification of introns: U2-type or U12-type. Cell. 1997;91(7):875–9.
16. Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. Nat Rev Genet. 2016;17(7):407–21.
17. Shukla GC, Padgett RA. Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants. RNA. 1999;5(4):525–38.
18. Wu Q, Krainer AR. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. Mol Cell Biol. 1999;19(5):3225–36.
19. Zhu W, Schlueter SD, Brendel V. Refined annotation of the Arabidopsis genome by complete expressed sequence tag mapping. Plant Physiol. 2003;132(2):469–84.
20. Zhu W, Brendel V. Identification, characterization and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. Nucleic Acids Res. 2003;31(15):4561–72.
21. Lewandowska D, Simpson CG, Clark GP, Jennings NS, Barciszewska-Pacak M, Lin CF, Makalowski W, Brown JW, Jarmolowski A. Determinants of plant U12-dependent intron splicing efficiency. Plant Cell. 2004;16(5):1340–52.
22. Szcześniak MW, Kabza M, Pokrzywa R, Gudyś A, Makałowska I. ERISdb: a database of plant splice sites and splicing signals. Plant Cell Physiol. 2013;54(2):e10.
23. Initiative The Arabidopsis Genome. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408(6814):796–815.
24. Brendel V, Xing L, Zhu W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics. 2004;20(7):1157–69.
25. Sparks ME, Brendel V. Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. Bioinformatics. 2005;21(3):iii20–30.
26. Brent MR, Guigó R. Recent advances in gene structure prediction. Curr Opin Struct Biol. 2004;14(3):264–72.
27. Goel N, Singh S, Aseri TC. A comparative analysis of soft computing techniques for gene prediction. Anal Biochem. 2013;438(1):14–21.

28. Huang Y, Chen SY, Deng F. Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction. Comput Struct Biotechnol J. 2016;14:298–303.
29. Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, Rosen BD, Cheng CY, Moreira W, Mock SA, et al. Araport: the Arabidopsis information portal. Nucleic Acids Res. 2015;43(Database issue):D1003–9.
30. Cheng CY, Krishnakumar V, Chan A, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 2017;89:789–804. https://doi.org/10.1111/tpj.13415
31. Pucker B, Holtgräwe D, Rosleff Sörensen T, Stracke R, Viehöver P, Weisshaar B. A de novo genome sequence assembly of the *Arabidopsis thaliana* accession Niederzenz-1 Displays presence/absence variation and strong synteny. PLoS ONE. 2016;11(10):e0164321.
32. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
33. Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 2008;24(3):319–24.
34. Ward N, Moreno-Hagelsieb G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS ONE. 2014;9(7):e101850.
35. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
36. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278(5338):631–7.
37. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
38. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
39. Stracke R, Holtgräwe D, Schneider J, Pucker B, Rosleff Sörensen T, Weisshaar B. Genome-wide identification and characterisation of R2R3-MYB genes in sugar beet (*Beta vulgaris*). BMC Plant Biol. 2014;14:249.
40. Stracke R, Huep G, Weisshaar B. Use of mutants from T-DNA insertion populations generated by high-throughput screening. In: Meksem K, Kahl G, editors. The handbook of plant mutation screening. Weinheim: Wiley-VCH; 2010. p. 31–54.
41. Stracke R, Ishihara H, Huep G, Barsch A, Mehrtens F, Niehaus K, Weisshaar B. Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the *Arabidopsis thaliana* seedling. Plant J. 2007;50(4):660–77.
42. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
43. Kent WJ. BLAT–the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
44. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19(Suppl 2):ii215–25.
45. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27(6):757–63.
46. Standage DS, Brendel VP. ParsEval: parallel comparison and analysis of gene structure annotations. BMC Bioinform. 2012;13:187.
47. Dal Bosco C, Lezhneva L, Biehl A, Leister D, Strotmann H, Wanner G, Meurer J. Inactivation of the chloroplast ATP synthase gamma subunit results in high non-photochemical fluorescence quenching and altered nuclear gene expression in *Arabidopsis thaliana*. J Biol Chem. 2004;279(2):1060–9.
48. Wang Y, Zhang WZ, Song LF, Zou JJ, Su Z, Wu WH. Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis. Plant Physiol. 2008;148(3):1201–11.
49. Brzezinka K, Altmann S, Czesnick H, Nicolas P, Gorka M, Benke E, Kabelitz T, Jähne F, Graf A, Kappel C, et al. Arabidopsis FORGETTER1 mediates stress-induced chromatin memory through nucleosome remodeling. Elife. 2016;5:e17061.
50. Ascencio-Ibáñez JT, Sozzani R, Lee TJ, Chu TM, Wolfinger RD, Cella R, Hanley-Bowdoin L. Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. Plant Physiol. 2008;148:1.
51. Liu D, Gong Q, Ma Y, Li P, Li J, Yang S, Yuan L, Yu Y, Pan D, Xu F, et al. cpSecA, a thylakoid protein translocase subunit, is essential for photosynthetic development in Arabidopsis. J Exp Bot. 2010;61(6):1655–69.
52. Skalitzky CA, Martin JR, Harwood JH, Beirne JJ, Adamczyk BJ, Heck GR, Cline K, Fernandez DE. Plastids contain a second sec translocase system with essential functions. Plant Physiol. 2011;155(1):354–69.
53. Morandini P, Valera M, Albumi C, Bonza MC, Giacometti S, Ravera G, Murgia I, Soave C, De Michelis MI. A novel interaction partner for the C-terminus of *Arabidopsis thaliana* plasma membrane H+ -ATPase (AHA1 isoform): site and mechanism of action on H+ -ATPase activity differ from those of 14-3-3 proteins. Plant J. 2002;31(4):487–97.
54. Viotti C, Luoni L, Morandini P, De Michelis M. Characterization of the interaction between the plasma membrane H-ATPase of *Arabidopsis thaliana* and a novel interactor (PPI1). FEBS J. 2005;272(22):5864–71.
55. Anzi C, Pelucchi P, Vazzola V, Murgia I, Gomarasca S, Piccoli MB, Morandini P. The proton pump interactor (Ppi) gene family of *Arabidopsis thaliana*: expression pattern of Ppi1 and characterisation of knockout mutants for Ppi1 and 2. Plant Biol. 2008;10(2):237–49.
56. Bonza MC, Fusca T, Homann U, Thiel G, De Michelis MI. Intracellular localisation of PPI1 (proton pump interactor, isoform 1), a regulatory protein of the plasma membrane H(+)-ATPase of *Arabidopsis thaliana*. Plant Biol. 2009;11(6):869–77.
57. Thieme CJ, Rojas-Triana M, Stecyk E, Schudoma C, Zhang W, Yang L, Miñambres M, Walther D, Schulze WX, Paz-Ares J, et al. Endogenous Arabidopsis messenger RNAs transported to distant tissues. Nat Plants. 2015;1(4):15025.
58. Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, Synek L. Dissecting a hidden gene duplication: the *Arabidopsis thaliana* SEC10 locus. PLoS ONE. 2014;9(4):e94077.
59. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;2014(15):e119.
60. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32(5):767–9.
61. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449(7161):463–7.
62. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43(5):476–81.
63. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al. The genome of the mesopolyploid crop species Brassica rapa. Nat Genet. 2011;43(10):1035–9.
64. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, Zhao M, Ma J, Yu J, Huang S, et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun. 2013;5:3930.
65. Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F, Tafer H, Rupp O, Sorensen TR, Stracke R, Reinhardt R, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014;505(7484):546–9.
66. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;44(9):e89.