High-throughput sequencing data and the impact of plant gene annotation quality

Vaattovaara, Aleksia Fanni Maria

2019-02-01

Vaattovaara , A F M , Leppälä , J M , Salojärvi , J T & Wrzaczek , M A 2019 , '
High-throughput sequencing data and the impact of plant gene annotation quality ' , Journal
of Experimental Botany , vol. 70 , no. 4 , ery43 , pp. 1069-1076 . https://doi.org/10.1093/jxb/ery434

http://hdl.handle.net/10138/300421
https://doi.org/10.1093/jxb/ery434

Journal of Experimental Botany, Vol. 70, No. 4 pp. 1069–1076, 2019
doi:10.1093/jxb/ery434 Advance Access Publication 24 December 2018
This paper is available online free of all access charges (see https://academic.oup.com/jxb/pages/openaccess for further details)

OPINION PAPER

High-throughput sequencing data and the impact of plant gene annotation quality

Aleksia Vaattovaara[1,†], Johanna Leppälä[2,†], Jarkko Salojärvi[1,3,*] and Michael Wrzaczek[1,*]

[1] Organismal and Evolutionary Biology Research Programme, Viikki Plant Science Centre, VIPS, Faculty of Biological and Environmental Sciences, University of Helsinki, Viikinkaari 1 (POB65), FI-00014 Helsinki, Finland
[2] Department of Ecology and Environmental Science, Umeå University, Linnaeus väg 6, SE-90187 Umeå, Sweden
[3] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore

[†] These authors contributed equally to this manuscript.

* Correspondence: michael.wrzaczek@helsinki.fi or jarkko@ntu.edu.sg

Received 16 August 2018; Editorial decision 28 November 2018; Accepted 28 November 2018

Abstract

The use of draft genomes of different species and re-sequencing of accessions and populations are now common tools for plant biology research. The *de novo* assembled draft genomes make it possible to identify pivotal divergence points in the plant lineage and provide an opportunity to investigate the genomic basis and timing of biological innovations by inferring orthologs between species. Furthermore, re-sequencing facilitates the mapping and subsequent molecular characterization of causative loci for traits, such as those for plant stress tolerance and development. In both cases high-quality gene annotation—the identification of protein-coding regions, gene promoters, and 5′- and 3′-untranslated regions—is critical for investigation of gene function. Annotations are constantly improving but automated gene annotations still require manual curation and experimental validation. This is particularly important for genes with large introns, genes located in regions rich with transposable elements or repeats, large gene families, and segmentally duplicated genes. In this opinion paper, we highlight the impact of annotation quality on evolutionary analyses, genome-wide association studies, and the identification of orthologous genes in plants. Furthermore, we predict that incorporating accurate information from manual curation into databases will dramatically improve the performance of automated gene predictors.

Keywords: Gene families, genome annotation, GWAS, high-throughput sequencing, phylogeny, translational research.

Introduction

The ongoing development of next-generation sequencing techniques has led to a remarkable decrease in the cost of genome sequencing. This is reflected in the increasing number of genome assemblies from all domains of life. For plants, more than 230 angiosperm genomes are currently available (Chen *et al.*, 2018). The advances in sequencing technology have also rapidly improved the quality, throughput, and the length of the reads. The result has been a dramatic increase in the amount and quality of information available for biological research, which often relies on gene model annotations representing exon–intron structures (including alternatively spliced isoforms), regulatory elements [e.g. promoter elements, enhancers, as well as 5′- and 3′-untranslated regions (UTRs)], and locations of transposable elements (TEs) and repeat sequences. Therefore, high-quality genome annotations are essential for analyses relying on genomic data. Since a large proportion of the genome is constituted of repetitive sequences and transposable elements varying, e.g., from 15% of the *Arabidopsis*

thaliana genome (de la Chaux *et al.*, 2012) up to 85% in maize (Schnable *et al.*, 2009) and bread wheat (Wicker *et al.*, 2018), genome assembly and annotation can be challenging. Errors in gene annotation have a strong impact on the results obtained, especially in phylogenomic analyses or in the functional interpretation of single-nucleotide polymorphisms detected in genome-wide association studies. These effects are more pronounced among tandemly duplicated genes and large gene families. However, the effects of erroneous gene annotations are still overlooked in many studies.

The importance of high-quality genome data

After the completion of a draft genome sequence, a standard approach is to first identify TEs and other repetitive DNA sequences and then to mask these parts to facilitate the prediction of protein-coding genes from the rest of the genome. A majority of TEs are located in heterochromatic regions but they can also be found in close proximity of genes. For example, in maize 33% of genes have TE insertions in introns (Schnable *et al.*, 2009) and 15% of Arabidopsis genes have TE insertions located within 500 bp from the 3′ or 5′ end of the coding region (Hollister *et al.*, 2011). This may hamper gene annotation, as parts of the genes (in this case intron sequences) could be masked due to repeat annotation. Alternatively, repeat masking that is too cautious can result in false positive gene predictions and produce inflated numbers of predicted genes. These in turn can cause problems in estimating the size of gene families, since TEs often contain fragments of functional genes. Therefore, it is important to develop ways to measure the accuracy of TE identification and annotation, a subject that has recently received attention (Hoen *et al.*, 2015). Gene annotation is typically initiated by so-called *de novo* gene prediction software such as Glimmer (Delcher *et al.*, 1999), SNAP (Korf, 2004), Augustus (Stanke *et al.*, 2008), EuGene (Foissac *et al.*, 2008), Genemark (Ter-Hovhannisyan *et al.*, 2008), or BRAKER (Hoff *et al.*, 2016). The programs learn a statistical model that predicts gene models from genome sequences. The model parameters are trained using evidence from RNA-sequencing data, expressed sequence tags (ESTs), annotated gene models in related species, or by using predictor parameters optimized for a model species. These automated predictions are then typically combined with evidence from RNA-sequencing and known gene models from other sources, for example from related species, using combiner software packages such as EVidenceModeler (Haas *et al.*, 2008), JIGSAW (Allen and Salzberg, 2005), or MAKER (Cantarel *et al.*, 2008). Proteogenomics have also been utilized to improve gene annotations in plants (Castellana *et al.*, 2008; Chapman and Bellgard, 2017). Several recent genome papers have paid specific attention to gene annotation quality and its improvement, for example in the genomes of kiwifruit (Pilkington *et al.*, 2018), silver birch (Salojärvi *et al.*, 2017), wheat (International Wheat Genome Sequencing Consortium, 2018), and melon (Ruggieri *et al.*, 2018). The kiwifruit and the silver birch publications in particular have highlighted the importance of manual curation of automatically predicted gene models.

The quality of a gene annotation is typically assessed through the presence and correctness of well-known single-copy genes (a set of 1440 in plants) called 'Benchmarking Universal Single-Copy Orthologs' (BUSCO; Simão *et al.*, 2015) or using a smaller set of genes in the 'Core Eukaryotic Genes Mapping Approach' (CEGMA) (Parra *et al.*, 2007). The BUSCO scores for well-annotated genomes vary between 95–97%, with 3–5% of missing or fragmented gene annotations (Raymond *et al.*, 2018; Springer *et al.*, 2018). However, conserved single-copy genes do not necessarily provide a suitable indicator for annotation quality for genes that have duplicates, such as members of gene families. This is particularly relevant for plant genomes where over 80% of genes belong to gene families (Guo, 2013). To evaluate gene family annotations, 'core gene families' (coreGFs) (Li *et al.*, 2016; Veeckman *et al.*, 2016) have been proposed in order to provide a measure for the presence of conserved gene families in a genome. Plants have experienced several whole-genome duplications that, together with local tandem duplications, have contributed to expansions of gene families. Due to their high sequence similarity, recently expanded gene families are highly susceptible to annotation problems and tools such as *DuplicationDetector* (Djedatin *et al.*, 2017) and NLR-Parser (Steuernagel *et al.*, 2015) have been developed to detect and correct problems following initial annotation.

Only a few comparisons of the quality of plant genome assemblies and annotations are currently available. Shangguan *et al.* (2013) assessed the quality of 32 plant genomes by mapping ESTs to genome sequences and found that, at the time of their study, the quality of many plant genomes was lower than had previously been assumed, whilst Veeckman *et al.* (2016) used ESTs, CEGMA genes, BUSCO genes, as well as coreGFs to evaluate the annotation quality of 12 plant genomes. In both studies, well-annotated plant genomes such as the model plant species *Arabidopsis thaliana* and rice (*Oryza sativa*) received high scores for genome and annotation quality. Arabidopsis and rice gene annotations have been heavily curated as a result of continuous input from the scientific community, resulting in constantly improving gene annotations. This shows the importance of constant reannotation and correction of gene models, but importantly also serves as a reminder to be aware of potential annotation errors when analysing the genomes of non-model species. However, even well-curated genomes can still contain annotation errors. For example, the gene models for the receptor-like protein kinases (RLKs) *AtCrk16* and *AtCrk17* (Vaattovaara *et al.*, 2018, Preprint) were found to contain annotation errors in the version TAIR10 (TAIR, 2010) of the *Arabidopsis thaliana* genome (the ectodomain region of *Crk16* was annotated as part of *Crk17* and the gene model for *Crk16* was truncated to contain only the kinase domain). This has been corrected in Araport11 (Cheng *et al.*, 2017), but the old versions are still listed as splice variants. Similarly, the gene model for the protein kinase *AtHt1* was only partially predicted (missing 45 AA from the N terminus) in the TAIR10 annotation for *Arabidopsis thaliana* (Hõrak *et al.*, 2016). More drastic gene annotation errors were identified during re-sequencing

of resistance genes from the tomato genome when 317 previously unannotated NB-LRR genes were revealed (Jupe *et al.*, 2013).

The issues leading to incorrect annotations of gene models can be diverse. The causative error types behind incorrect gene model predictions can be categorized into sequencing errors, assembly errors, or difficulties with annotation. Sequencing errors typically occur in regions with low sequencing coverage where errors in individual reads can introduce stop codons or frame shifts, resulting in erroneous gene models. The end result of a genome assembly is typically a large number of scaffolds, assemblies where contiguous genomic sequences (typically referred to as contigs) are linked by gaps filled by ambiguous character (Ns). Assembly errors, meaning the erroneous linking of contigs, may lead to truncated or fused gene models or frame shifts. Gaps in assemblies can also result in partial gene models if they overlap with a gene. Similarly, genes may be located at the edges of scaffolds, causing partial or missing annotations. Similar to assembly errors, annotation errors can lead to erroneous gene models, but there can also be problems in gene structure, such as missing or extra exons. A single gene may be split into several genes, or several genes could be fused into a single-gene model, or the splice sites can be misplaced. In case of gene family members organized in tandem repeats, automated gene predictors can in some cases predict a fused-gene model by combining exons in consecutive genes. Naturally, genes can also be entirely missed by annotation algorithms. Even in 'simpler' prokaryotic genomes, where genes generally do not contain introns, unannotated genes pose a problem (Warren *et al.*, 2010).

Information from transcriptome sequencing yields high-quality evidence for genes with high transcript abundance and is therefore an invaluable source for gene annotation. In particular, long introns can cause problems in the annotation of the gene models, as gene prediction programmes may split a single gene into truncated partial-gene models. Evidence from transcript data can be helpful in many such cases. Notably, plant genes can contain very long introns, such as *OsMADS50* with the first intron being 27.6 kb long (Tadege *et al.*, 2003); however, this is very rare. In Arabidopsis less than 1% of introns are longer than 1kb (Chang *et al.*, 2017), whereas the same figure in Norway spruce is 24% (Nystedt *et al.*, 2013). There are limitations to the information that can be obtained from transcripts as it is challenging to obtain a comprehensive set of transcripts for all genes in a genome. Typically, only 60–70% of the genes encoded in a genome are expressed in the sampled material (Salojärvi *et al.*, 2017). A large number of transcripts are differentially regulated in response to circadian rhythms, developmental cues, environmental signals, or stress conditions. In multicellular organisms, genes may also have expression patterns that are highly cell- or tissue-specific, resulting in low abundance of corresponding mRNAs in some tissues. In addition to the biological challenges, *de novo* assembled transcriptome data can contain assembly errors. A reference-guided transcriptome assembly, on the other hand, can suffer from assembly errors in a genome that, when used to support gene prediction, can lead to errors in gene models. Although combiner software can search for stop codons to identify the approximate end of a gene (Haas *et al.*, 2008), partial transcripts can lead to the annotation of truncated gene models in the absence of other evidence. For members of large gene families with high sequence similarities it can be difficult to distinguish splice variants and recently diverged gene models, especially if the transcriptome data is sequenced from a different individual. This problem can be expected to become more prominent in the future as more genomes from (auto)polyploid plants become available. Finally, a complementary source of gene annotation, use of predicted gene models and proteins from other available genomes, can possibly lead to inherited erroneous annotations, as has been observed in case of functional annotations (Gilks *et al.*, 2002). Further support for the correctness of gene models can come from the domain composition of the encoded protein. Identification of partial Pfam domains (https://pfam.xfam.org/) has been found to be an excellent indicator of possible annotation errors in gene models (Triant and Pearson, 2015). Therefore, while partial protein domains can arise through incomplete duplications, the identification of a truncated domain warrants additional validation of the gene annotation.

Correct gene annotation with a high degree of completeness is essential for the functional annotation of the genes (Jones *et al.*, 2007; Schnoes *et al.*, 2009), such as the gene ontology assignment and identification of conserved protein domains, and for all subsequent analyses utilizing this information. Variation in the quality of genome and gene annotations can especially cause problems in comparative and evolutionary analyses. Thus, it is necessary to manually validate the annotation quality of different genomes and data sets extracted from available genomes.

The importance of accurate phylogenies for translational research

While small annotation errors in gene models do not always drastically alter the results of phylogenetic analyses, partial or absent gene models can result in false tree topologies that hamper the interpretation of gene relationships, especially for translational research. Model organisms such as *Arabidopsis thaliana* are a common choice for investigation of gene or protein function (Davis, 2004). Model organisms may have little commercial or agricultural relevance but have relatively small genomes, a short generation time, and can be propagated easily in the laboratory. However, research on a model organism is often carried out to improve the traits of crop species. Orthologous genes by definition have similar functions in different species, thereby allowing transfer of the functional information. With single-copy genes and small gene families the inference of orthologous gene pairs between species is usually simple, but the situation can be more complicated in the case of large gene families. Genome duplications and tandem duplications result in paralogous genes, which can hamper the identification of orthologs that have the same function (Box 1).

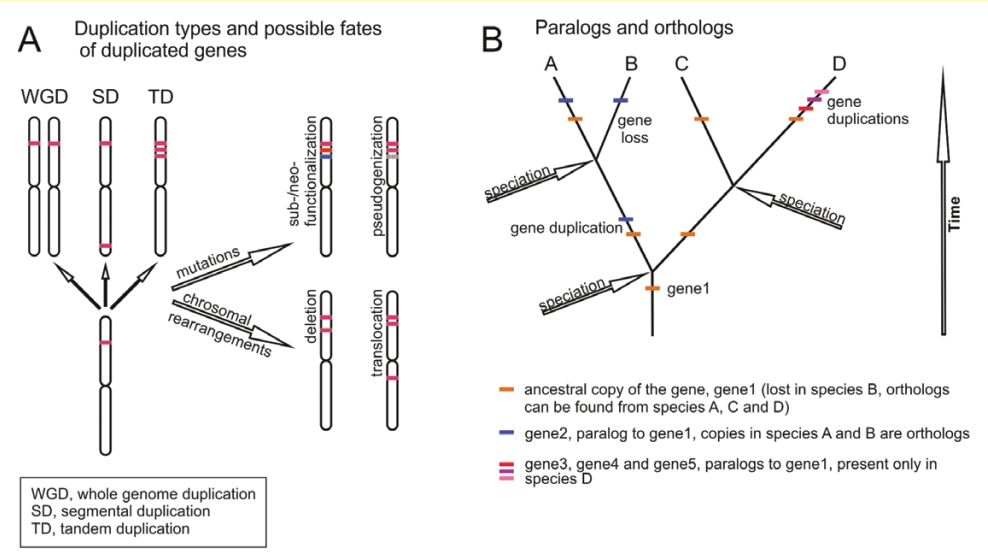
A high-quality phylogenetic tree containing curated gene models from gene family members is an efficient way to investigate relationships among genes from different species (Box

Box 1. Gene families

Genes in plant genomes are rarely pure single-copy genes with one-to-one orthologs in different species, but instead belong to gene families, a group of genes with common ancestry. An analysis of eight plant and algal genomes found that 86.4% of genes belonged to gene families (Guo, 2013). Gene families can be part of larger superfamilies, for example the plant receptor-like kinase gene family is part of the larger superfamily of protein kinases. The size of gene families can vary drastically, from small families with only few members to very large families with more than thousand. Notably, the number of members in gene families can vary considerably between different species.

Gene families evolve through duplications, pseudogenization, and gene-loss events (see part A of the figure). Duplicated genes can result from whole-genome duplication (WGD) or triplication (WGT) events, from tandem duplication events, from segmental duplication within or between chromosomes, or from duplications mediated by transposable elements (Panchy *et al.*, 2016). A general hypothesis is that following duplication, the duplicates are under decreased selection pressure, which allows mutations to accumulate. This can lead to distribution of ancestral functions between duplicated genes, sub-functionalization, or acquisition of novel functions compared to the ancestral gene (neofunctionalization; Conant and Wolfe, 2008). Genes can be also turned into pseudogenes as a result of the slow accumulation of deleterious mutations, or they can be lost during chromosomal rearrangements. Under strict selection, new duplicates can be removed from the genome to keep the gene as a single-copy gene.

Gene families constitute a striking challenge for translational research. In large gene families the recognition of orthologs (genes separated by speciation) from paralogs (genes that emerge as the result of duplications within species) between distant species can be difficult or even impossible (see part B of the figure). Tandem duplications in particular can lead to large lineage-specific expansions and thus to lineage-specific genes without orthologous genes outside of that specific taxon. The most common problem in translational studies is the assumption that genes having the highest sequence similarity between species are orthologs. This can be erroneous for large gene families, which may have gained and lost genes at different rates in different lineages. In addition, model species have lineage- or species-specific genes and they may have lost certain genes or even gene families. For these reasons, transferring information from large gene families of model species to crops remains a considerable challenge for translational research.



2). However, for large gene families in particular, the high sequence similarity of genes from different species is not necessarily sufficient evidence for orthologs or paralogs, as the similarity-based search does not account for lineage-specific gene duplications and losses. Within gene families, some members may have evolved similarly to conserved single-copy genes, thus facilitating the recognition of orthologs. However, other members may have experienced lineage-specific expansions, thus making the inference of orthologs challenging or even impossible. Synteny, the conservation of the local ordering of a set of genes, provides further evidence of orthology. Finally, transcriptional evidence can be used to determine whether the expression of putative orthologs is conserved, since the altered regulation of gene expression, for example due to changes in the promoter region, is not visible in phylogenetic trees based on the coding region.

Genome-wide association studies benefit from high-quality annotations

Identifying causal variants for quantitative traits in model plants and crops is essential for improving agronomically important traits and fundamental for finding adaptive variants in evolutionary biology. High-throughput genome sequencing enables this by providing information on single-nucleotide polymorphisms (SNPs) that can be analysed in a genome-wide association study (GWAS). GWAS utilizes naturally occurring phenotypic variation within populations or species and identifies statistically associated genotypic variation (for a recent review in plants see Ogura and Busch, 2015).

Interpretation of results from GWAS analyses can be challenging. The genetic architecture underlying many common traits is frequently complex (that is, polygenic) and therefore the sizes of the effects of the associated SNPs that have been identified by the GWAS are often small (Ingvarsson and Street, 2011; Visscher *et al.*, 2017), although the proportion of the variation in the trait explained by significant SNPs is generally found to be larger in plants (Huang *et al.*, 2010; Li *et al.*, 2010; Ingvarsson and Street, 2011). Thus, it can be challenging to distinguish true 'small effect' associations from artefacts. In addition, several studies have shown that the underlying genetics for adaptive traits can cause significant associations to be thousands of base pairs away from the causative locus, due to multiple alleles being present at a locus and recent positive selection (causing positively associated SNPs to be spread over longer regions due to linkage disequilibrium; e.g. Atwell *et al.*, 2010; Kerdaffrec *et al.*, 2016). Thorough fine-mapping of quantitative trait loci is highly time-consuming, and so high-quality annotation information is essential for determining the location of candidate causative SNPs for validation: SNPs may be located either in protein-coding regions where they can result in synonymous or non-synonymous substitutions, or alternatively in introns, UTRs, or intergenic regions where they possibly have a lower impact on function. Therefore, high-quality structural and functional gene annotations are essential for predicting the likely candidate genes or genomic regions and nucleotide variant(s) that underlie an investigated trait. GWASs frequently concentrate on SNPs located in protein-coding genes. Non-synonymous SNPs can be identified from high-quality annotation data, and prior information, including gene expression data and computational methods, can be used to predict the effect on protein function (Tang and Thomas, 2016). For example, the weeping phenotype of the silver birch garden cultivar 'Youngii' was predicted to result from a premature stop codon detected in the *LAZY* gene, known to result in a similar relaxed phenotype in maize, rice, Arabidopsis, and peach (Salojärvi *et al.*, 2017). However, SNP variants outside coding regions can also be functional, for example by affecting regulation of gene expression or transcript stability. In addition, transposable elements can affect the expression of nearby genes, resulting in changes in the plant phenotype (reviewed by Cui and Cao, 2014). It is therefore crucial to improve the annotation and functional understanding of the non-coding parts of genomes, which typically comprise the majority of the genome. Recent studies have identified long non-coding RNAs (Liu *et al.*, 2015), open chromatin regions (Rodgers-Melnick *et al.*, 2016), small open reading frames (ORFs; Hellens *et al.*, 2016), epigenomes (Kawakatsu *et al.*, 2016), and transcription factor binding sites (cistromes; O'Malley *et al.*, 2016). It would be valuable to obtain such information from many different tissue and cell types. Increased knowledge of the functional roles of non-coding regions will greatly improve identification of putative causal SNPs from GWASs. In human genetics, tools to predict the functions of non-coding SNPs have been developed (reviewed by Nishizaki and Boyle, 2017) and hopefully similar tools will soon become available for plant model species, and subsequently more widely for non-model species.

Similar to phylogenetic approaches, GWASs are also strongly affected by the quality of genome assemblies. When high-throughput sequencing is used for genotyping, the SNPs are typically identified against a reference genome. A genome sequence that is absent from the reference but present in some individuals is usually excluded in standard analyses. If the size of the excluded insert is large and it has a causative variant, the remaining associated SNPs may be too far away and thus the size of their effect may be too low to be detected. Functional interpretation is therefore not possible because the variants, which potentially have a causative effect, are not included in the analysis.

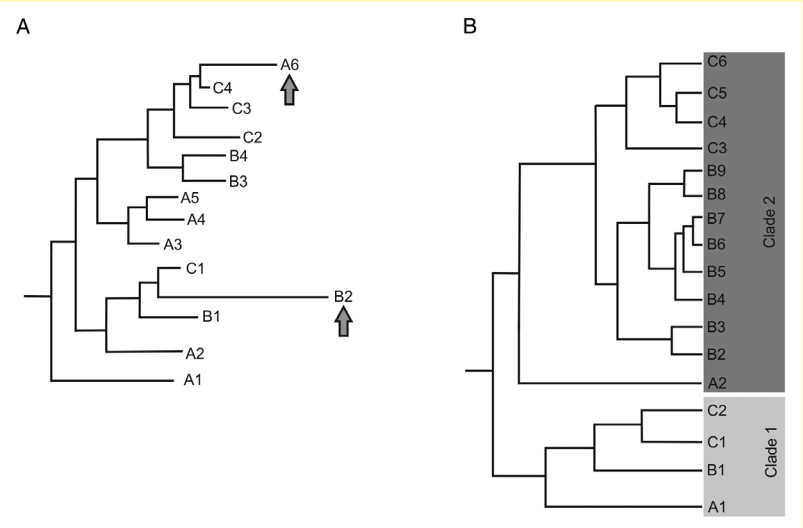
Using additional resources such as gene-expression and protein-interaction networks together with high-quality genome annotation will in the future result in improved understanding of GWAS results, and ultimately in improved understanding of the genetic architecture of different traits. Careful reannotation of regions containing significant SNPs is strongly advisable to verify their location in coding or non-coding regions. In the future, GWASs will benefit from integration of improved, curated annotations of both protein coding and non-coding regions into the available reference genome information.

Box 2. Phylogenies

Phylogenetic trees represent the evolutionary distances between given data samples. In phylogenomics, phylogenies are usually either species trees that represent the relationships between species, or gene (or protein) trees that represent evolutionary relationships between genes and proteins. The distances in the tree are estimated based on the similarities and differences between data samples; for genes and proteins the data are derived from sequence alignments. Nucleotide sequence alignments are informative for closely related sequences, while amino acid alignments can be more practical for more divergent data sets. The construction of phylogenetic trees can be carried out by different methods based on distance matrices, parsimony, maximum-likelihood, or Bayesian methods (Yang and Rannala, 2012). With large data sets, the estimation of the phylogenetic tree is a so-called NP-hard (non-deterministic polynomial-time) problem where there exists a single, correct answer, but it is exponentially hard to identify, and therefore the result represents the optimum from among the trees explored by the search algorithm. The reliability of the splits represented by the nodes of the tree is commonly evaluated by bootstrapping (Felsenstein, 1985; Holder and Lewis, 2003).

Annotation errors can cause severe problems for the inference of phylogenetic trees. Short stretches of missing or extra sequences may only affect the branch lengths in the tree, but missing or additional exons can lead to long branches or even differences in the branching order in a phylogeny. Missing gene models on the other hand can lead to false estimations of the relationships between genes. In part (A) of the figure, which represents members of a gene family from three different species indicated by the letters A–C, two examples of possible annotation errors in the phylogenetic tree are presented. The branch leading to gene B2 is long compared to the branches leading to the other genes in the phylogeny. This indicates either diversifying evolution or that there are problems with the annotation. In the case of gene A6, an annotation error is even more likely, as the gene is grouped with the genes from species C.

The relationships between genes from different species can be defined from comprehensive phylogenies. This is useful for identifying orthologs and paralogs in gene families, for example for translational research where information on gene or protein function in a model species is used to improve crop performance. The recognition of lineage-specific gene expansions and contractions from the phylogeny is important for correct interpretation of the relationships of gene family members between species. In part (B) of the figure, which again represents members of a gene family from three different species, the ancestral gene has duplicated prior to the separation of the species, giving rise to two related clades. In both clades a single gene from species A is placed closest to the root of the tree. In Clade 1, gene B1 is orthologous to A1, while in the lineage leading to species C, gene duplication has taken place, making ortholog identification more difficult. In Clade 2, both species B and C have undergone several lineage-specific gene duplication events. For species B, the inference of orthologous genes between species is not possible based only on the phylogeny. For species C, the gene C3 could possibly be the ortholog for A2 as they are the most similar genes between these species, but sub-functionalization between these genes is also possible.



Conclusions

Over the past decade, high-throughput sequencing has improved significantly and genome data for different plant species, accessions, and populations are now widely available. Reference genomes can be used as a tool for evolutionary analyses, for the mapping and characterization of agriculturally important traits, and for translational research. However, despite the increasing number and quality of available genomes, the quality of their assembly and annotation is variable, and in some cases this can represent a serious problem for detailed analyses.

Overall, as sequencing technology develops, assemblies become more contiguous and contain fewer assembly errors. At the same time, gene annotation software is developing rapidly and is already able to overcome commonly observed annotation problems, whilst RNAseq is able to provide reliable evidence for gene model structures. The development of combiner software, which is able to make composite gene predictions based on several data sources, has been a major recent advance in gene annotation, and our view is that it should be the main focus when aiming to further improve annotation quality, perhaps by including more diverse information sources such as phylogenomics, synteny, and information on tandem expansions.

For the time being, careful validation and *de novo* annotation of gene models is particularly important not only for members of large gene families, but also for genes where transcriptional evidence is difficult to obtain, such as those transcribed in response to specific environmental stimuli or that have high tissue specificity. Manual curation of annotations allows the identification of orthologous genes with higher confidence for translational research, and also strengthens the GWAS approach for the identification of causative loci for traits of high importance to agriculture.

In spite of the increasing quality of gene and genome annotations, we want to emphasize that the careful reannotation of genomic regions of interest remains an important tool for analyses of gene families, for translational approaches, and also for GWASs and mutant screens. The procedure should consist of reannotation of TEs, ORFs, and exon–intron structures, as well as promoter and UTR regions. Encoded proteins should be checked for the completeness of Pfam domains and for whether they correspond to the typical domain architecture of related proteins. Making these curated annotations available in genome databases will dramatically increase the quality of genomes and gene annotations. More importantly, more accurate information in databases makes it possible to improve automated gene predictors, which in turn will reduce the effort required for manual reannotation of genomic features.

Acknowledgements

This work was supported by the Doctoral Programme in Plant Sciences (DPPS) of the University of Helsinki and the Finnish Cultural Foundation (to AV) and the Academy of Finland (grant numbers 275632, 283139, and 312498, to MW). AV, JS, and MW are members of the Centre of Excellence in the Molecular Biology of Primary Producers (2014–2019) funded by the Academy of Finland (grant numbers 271832 and 307335).

References

Allen JE, Salzberg SL. 2005. JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics **21,** 3596–3603.

Atwell S, Huang YS, Vilhjálmsson BJ, *et al*. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature **465,** 627–631.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research **18,** 188–196.

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP. 2008. Discovery and revision of Arabidopsis genes by proteogenomics. Proceedings of the National Academy of Sciences, USA **105,** 21034–21038.

Chang N, Sun Q, Hu J, An C, Gao AH. 2017. Large introns of 5 to 10 kilo base pairs can be spliced out in Arabidopsis. Genes **8,** 200.

Chapman B, Bellgard M. 2017. Plant proteogenomics: improvements to the grapevine genome annotation. Proteomics **17,** 1700197.

Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, Lin Z, Tang H, Zhang L. 2018. The sequenced angiosperm genomes and genome databases. Frontiers in Plant Science **9,** 418.

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal **89,** 789–804.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nature Reviews Genetics **9,** 938–950.

Cui X, Cao X. 2014. Epigenetic regulation and functional exaptation of transposable elements in higher plants. Current Opinion in Plant Biology **21,** 83–88.

Davis RH. 2004. The age of model organisms. Nature Reviews Genetics **5,** 69–76.

de la Chaux N, Tsuchimatsu T, Shimizu KK, Wagner A. 2012. The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. Mobile DNA **3,** 2.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Research **27,** 4636–4641.

Djedatin G, Monat C, Engelen S, Sabot F. 2017. *DuplicationDetector*, a light weight tool for duplication detection using NGS data. Current Plant Biology **9–10,** 23–28.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39,** 783–791.

Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, Van de Peer Y, Rouze P, Schiex T. 2008. Genome annotation in plants and fungi: EuGene as a model platform. Current Bioinformatics **3,** 87–97.

Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA. 2002. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics **18,** 1641–1649.

Guo YL. 2013. Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. The Plant Journal **73,** 941–951.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biology **9,** R7.

Hellens RP, Brown CM, Chisnall MAW, Waterhouse PM, Macknight RC. 2016. The emerging world of small ORFs. Trends in Plant Science **21,** 317–328.

Hoen DR, Hickey G, Bourque G, *et al*. 2015. A call for benchmarking transposable element annotation methods. Mobile DNA **6,** 13.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics **32,** 767–769.

Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews Genetics **4,** 275–284.

Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 2011. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proceedings of the National Academy of Sciences, USA **108,** 2322–2327.

Hõrak H, Sierla M, Tõldsepp K, *et al*. 2016. A dominant mutation in the HT1 kinase uncovers roles of MAP kinases and GHR1 in CO_2-induced stomatal closure. The Plant Cell **28,** 2493–2509.

Huang X, Wei X, Sang T, *et al*. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics **42,** 961–967.

Ingvarsson PK, Street NR. 2011. Association genetics of complex traits in plants. New Phytologist **189,** 909–922.

International Wheat Genome Sequencing Consortium. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science **361,** eaar7191.

Jones CE, Brown AL, Baumann U. 2007. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics **8,** 170.

Jupe F, Witek K, Verweij W, *et al*. 2013. Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. The Plant Journal **76,** 530–544.

Kawakatsu T, Huang SC, Jupe F, *et al*. 2016. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell **166,** 492–505.

Kerdaffrec E, Filiault DL, Korte A, Sasaki E, Nizhynska V, Seren U, Nordborg M. 2016. Multiple alleles at a single locus control seed dormancy in Swedish *Arabidopsis*. eLife **5,** e22502.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics **5,** 59.

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO. 2010. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences, USA **107,** 21199–21204.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. The Plant Cell **28,** 326–344.

Liu J, Wang H, Chua NH. 2015. Long noncoding RNA transcriptome of plants. Plant Biotechnology Journal **13,** 319–328.

Nishizaki SS, Boyle AP. 2017. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. Trends in Genetics **33,** 34–45.

Nystedt B, Street NR, Wetterbom A, *et al*. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature **497,** 579–584.

O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and epicistrome features shape the regulatory DNA landscape. Cell **165,** 1280–1292.

Ogura T, Busch W. 2015. From phenotypes to causal sequences: using genome wide association studies to dissect the sequence basis for variation of plant development. Current Opinion in Plant Biology **23,** 98–108.

Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of gene duplication in plants. Plant Physiology **171,** 2294–2316.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23,** 1061–1067.

Pilkington SM, Crowhurst R, Hilario E, *et al*. 2018. A manually annotated *Actinidia chinensis* var. *chinensis* (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. BMC Genomics **19,** 257.

Raymond O, Gouzy J, Just J, *et al*. 2018. The *Rosa* genome provides new insights into the domestication of modern roses. Nature Genetics **50,** 772–777.

Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. 2016. Open chromatin reveals the functional maize genome. Proceedings of the National Academy of Sciences, USA **113,** E3177–3184.

Ruggieri V, Alexiou KG, Morata J, *et al*. 2018. An improved assembly and annotation of the melon (*Cucumis melo* L.) reference genome. Scientific Reports **8,** 8088.

Salojärvi J, Smolander OP, Nieminen K, *et al*. 2017. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nature Genetics **49,** 904–912.

Schnable PS, Ware D, Fulton RS, *et al*. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science **326,** 1112–1115.

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Computational Biology **5,** e1000605.

Shangguan L, Han J, Kayesh E, Sun X, Zhang C, Pervaiz T, Wen X, Fang J. 2013. Evaluation of genome sequencing quality in selected plant species using expressed sequence tags. PLoS ONE **8,** e69890.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31,** 3210–3212.

Springer NM, Anderson SN, Andorf CM, *et al*. 2018. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nature Genetics **50,** 1282–1288.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics **24,** 637–644.

Steuernagel B, Jupe F, Witek K, Jones JD, Wulff BB. 2015. NLR-parser: rapid annotation of plant NLR complements. Bioinformatics **31,** 1665–1667.

Tadege M, Sheldon CC, Helliwell CA, Upadhyaya NM, Dennis ES, Peacock WJ. 2003. Reciprocal control of flowering time by OsSOC1 in transgenic Arabidopsis and by FLC in transgenic rice. Plant Biotechnology Journal **1,** 361–369.

TAIR (The Arabidopsis Information Resource). 2010. https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/README_TAIR10.txt at www.arabidopsis.org.

Tang H, Thomas PD. 2016. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics **203,** 635–647.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Research **18,** 1979–1990.

Triant DA, Pearson WR. 2015. Most partial domains in proteins are alignment and annotation artifacts. Genome Biology **16,** 99.

Vaattovaara A, Brandt B, Rajaraman S, *et al*. 2018. Mechanistic insights into the evolution of DUF26-containing proteins in land plants. BioRxiv, 493502. [Preprint].

Veeckman E, Ruttink T, Vandepoele K. 2016. Are we there yet? Reliably estimating the completeness of plant genome sequences. The Plant cell **28,** 1759–1768.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. American Journal of Human Genetics **101,** 5–22.

Warren AS, Archuleta J, Feng WC, Setubal JC. 2010. Missing genes in the annotation of prokaryotic genomes. BMC Bioinformatics **11,** 131.

Wicker T, Gundlach H, Spannagl M, *et al*. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biology **19,** 103.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics **13,** 303–314.