Gene space completeness in complex plant genomes

Michiel Van Bel[1,2], François Bucchini[1,2] and Klaas Vandepoele[1,2,3*]

[1] Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium

[3] Bioinformatics Institute Ghent, Ghent University, 9052 Ghent, Belgium

[*] To whom correspondence should be addressed. Tel: +32 9 3313822; Fax: +32 9 3313809; Email: klaas.vandepoele@psb.vib-ugent.be

Author contact information: email addresses

Michiel Van Bel Michiel.vanbel@psb.vib-ugent.be
François Bucchini Francois.bucchini@psb.vib-ugent.be
Klaas Vandepoele Klaas.vandepoele@psb.vib-ugent.be

Author information: ORCID

Michiel Van Bel https://orcid.org/0000-0002-1873-2563
François Bucchini https://orcid.org/0000-0002-5686-7960
Klaas Vandepoele https://orcid.org/0000-0003-4790-2725

Abstract

Genome annotations offer ample opportunities to study gene functions, biochemical and regulatory pathways, or quantitative trait loci in plants. Determining the quality and completeness of a genome annotation, and maintaining the balance between them, are major challenges, even for genomes of well-studied model organisms. In this review, we present a historical overview of the complexity in different plant genomes and discuss the hurdles and possible solutions in obtaining a complete and high-quality genome annotation. We illustrate there is no clear-cut answer to solve these challenges for different gene types, but provide tips on guiding the iterative process of generating a superior genome annotation, which is a moving target as our knowledge about plant genomics increases and additional data sources become available.

Keywords

Gene Annotation; Genome Complexity

Introduction

The increased availability of genome sequences and their associated gene products catalogs has created immense opportunities in virtually all fields of life sciences: from revealing the functional entities encoded in the genomes, over the analysis of pathways that have led to the current diversity of metabolites produced by different species, to determining the genomic loci associated with traits and susceptibility to disease. The enormous costs associated with genome sequencing of the first eukaryotic model species have been offset by the boost they have given to the development of new and groundbreaking DNA sequencing technologies, paving the way for studying the genomes of both model and non-model species.

The interest in plant genome sequences has mainly been driven by the wide variety of crop species that are globally cultivated. Food crops such as rice [1], maize [2], and wheat [3], are obviously the most important species to sequence in order to safeguard the access to produce, but are also among the more complex and difficult genomes to assemble. Other plant species have also been sequenced because of economic, cultural, and ecological reasons, such as poplar [4], tobacco plant [5] , rubber tree [6], and the parasite *Cuscuta* [7]. Specific characteristics of plants and their genomes make sequencing, assembling, and annotating their genomes more difficult compared to other eukaryotes. The plant cell wall, in combination with other factors such as secondary metabolites, causes the DNA extraction to be more challenging in plants than in animals [8]. Other characteristics are related to genome plasticity [9]: autopolyploidy and allopolyploidy events lead to complex, sometimes hybrid, genomes in which many genes are present in multiple copies [3], while transposon activity can increase the size of the genomes to unparalleled sizes by the insertion of regions containing highly repetitive DNA [10], exemplified by various gymnosperm sequencing projects [11,12].

Raw genome sequences form the basis for further domain-specific knowledge extraction: the various regions within the genome, such as gene loci, regulatory regions, and QTLs, need to be detected and annotated. These regions and associated products can subsequently be functionally annotated, resulting in an integrated and multi-layered network. In order to maximize extraction of useful knowledge from this biological network and prevent the inference of incorrect conclusions, it needs to be as comprehensive as possible. This in turn relies on the completeness of its constituent underlying data sets, which are interdependent themselves. As such, the following two key questions need to be answered: is the genome sequence complete, and is the set of products complete? The first question implies the full sequencing and assembly of difficult-to-sequence regions, such as centromeres, telomeres and other high-repeat sequences, as well as the detection and separation of allelic variations to deal with heterozygosity. This problem is partially solved: for most organisms the total assembled genome size has not changed much over the years (Supplementary Figure 1). Rather, optimization of the chromosome assembly and gap removal are major foci of current genome assembly projects [13,14]. The answer to the second question is multi-facetted: where initial gene annotation efforts were mostly concentrating on protein-coding genes, the focus has shifted to include the multitude of classes of RNA genes as well. Additionally, new transcriptome profiling initiatives cause the total number of genes to be in constant flux, even for major well-studied model species such as human [15].

In this review, we give an overview of the increasing levels of complexity identified in plant genomes, together with associated challenges related to completeness assessments of the different types of

gene products. Furthermore, we show that a continuous effort is required to extract high-quality gene information from newly and previously sequenced plant genomes, through the application of improved gene annotation methods and novel experimental approaches.

Next-generation sequencing increases our understanding of the complexity of plant genomes

Observation of gene annotations from a historical perspective provides key insights into the changing landscape of the annotations of protein-coding loci, RNA loci, and protein-coding isoforms for a set of model organisms (Figure 1). While the number of protein-coding loci is quasi stable once a baseline genome has been assembled (small increases and decreases notwithstanding), the number of annotated protein-coding isoforms has seen a sharp increase in recent years for many species. This observation holds true for plant species, as well as for most other model species added as comparison. Figure 1 also indicates that the initial gene annotations often did not include RNA genes, or were restricted to certain RNA-types when they did. Throughout the years however, the biological importance of RNA genes has become clearer leading to a large increase of annotated RNA genes, initially in animals but more recently also in the plant kingdom.

The time points that coincide with the increased number of annotated isoforms can easily be connected with the rise in popularity of RNA sequencing (RNA-seq): the detected transcripts from these RNA-seq experiments had to be made concordant with the known genomic loci, resulting in expanding the gene models with new isoforms. In order to fully capture isoforms, a multitude of RNA-seq experiments is required over a wide variety of conditions and developmental stages, in as many organs or cell types as possible [16]. Recent advances in reliable annotation of new isoforms have led to various studies presenting different sets of isoforms, which are currently not all integrated in a single reference genome annotation. For example, the recent Reference Transcript Dataset for Arabidopsis (AtRTD2) reported 82,190 non-redundant transcripts from 34,212 genes, many of which are not present in AraPort11 [16,17]. The larger average number of isoforms per locus in AtRTD2 compared to Araport11 (2.68 vs. 1.75, respectively) offers new opportunities to study the functional consequences of isoform switching in plants [18].

Spearheaded by large-scale RNA sequencing experiments, a strong increase in the number of annotated non-coding RNA genes can also be observed. Apart from well-studied types of RNAs such as rRNAs, tRNAs, miRNAs, or snoRNAs, thousands of long non-coding RNAs (lncRNAs) have recently been reported for several plant models including rice [19,20], maize [20,21], and Arabidopsis [22]. LncRNA genes are usually longer than 200 nucleotides and lack an open-reading frame coding for a protein. Many of these lncRNAs show highly tissue-specific expression and several mechanisms have been identified where, both in cis or in trans, they play a role in mediating (post-)transcriptional gene silencing and regulating the transcription machinery, histone modifications, and the RNA processing machinery [23]. For most plant species these novel RNA genes have not yet been unified in reference genome annotations, indicating that researchers currently need to combine multiple gene annotation

sources in order to get a complete view on the RNA gene space, for example when studying transcriptional responses.

Additionally, focusing on a single strain of an organism hides the true complexity of gene space variation: sequencing the genomes of a variety of accessions has been applied in different crop species in order to study gene diversity and evolutionary adaptations that can be used to improve breeding [24]. By using a *de novo* genome assembly approach, rather than mapping the data to the reference genome, new strain-specific variations can be discovered and catalogued, resulting in the pan-genome for a given species [25-27]. For example, a presence-absence variation analysis in 67 rice accessions provided evidence for 10,872 gene models that were absent in the *O. sativa* Nipponbare reference annotation [25], with 90% of those gene models being absent from the reference assembly. This indicates that the missing gene models are not absent due to problems with the gene annotation, but are rather true examples of pan-genomic gene expansion. A study using 54 *Brachypodium distachyon* lines revealed a similar picture where 7,135 out of 17,195 genes that were present in 3 to 52 of the lines, were not present in the reference gene annotation [26]. The resulting pan-genome can thus show significant gene variability within a single species. As many of the newly discovered *B. distachyon* pan-genes lack homologs in related cereals (i.e. species-specific) and are as such less evolutionarily constrained compared to core genes present in multiple species, methods estimating gene space completeness based on evolutionary conservation will not be strongly affected by the presence or absence of pan-genes [26]. Constructing the pan-genome of a species is only one of the possible ways to fully capture the intra-species complexity of an organism [28]. Recent research tries a different approach by merging all assemblies of a pan-genome into a single genome graph [29], which is the graph representation of small nucleotide polymorphisms (SNPs), small insertions and deletions, as well as longer stretches of allelic variation and strain variation.

Gene space completeness in plant genome sequencing projects

CEGMA (Core Eukaryotic Genes Mapping Approach) was one of the initial approaches towards defining a list of conserved genes that can be used to assess the completeness of a genome annotation [30]. Specifically, it relied on the evolutionary conservation of orthologous genes within eukaryotes to estimate the expected gene space and assessed what fraction of these orthologs was present in a given genome or gene set. The most popular successor, BUSCO (Benchmarking Universal Single-Copy Orthologs), uses an approach that allows for the natural variation of conserved genes within clades by having distinct sets of nearly-universal single-copy orthologs for a variety of eukaryotic clades [31]. BUSCO relies on the selection of predominantly single-copy genes for the creation of orthologous datasets, which poses major problems for flowering plant species that have undergone (ancient) whole-genome duplications or are allopolyploids [32]. As an alternative to single-copy orthologs, the PLAZA Core Gene Families ('coreGFs') are a set of gene families that are highly conserved in a majority of plant species and that are not filtered for single-copy genes [33]. CoreGFs consequently cover a larger number of genes to model the expected gene space and show less functional biases compared to single-copy genes in plants [13,34]. For instance, proteins that encode for key components of the eukaryotic cell machinery, such as transcriptions factors, ribosomal proteins, kinases, and histones, are absent from (quasi) single-copy gene sets, but are present in the CoreGFs (Supplementary Table 3). A perfect completeness score solely based on single-copy gene families does therefore not reflect a

biologically functional and complete genome. Both BUSCO and PLAZA coreGFs currently define different sets of conserved genes in order to model the expected gene space at different evolutionary scales within the green plant lineage.

An overview of gene space completeness estimates for eight plant species using BUSCO Embryophyta and the PLAZA 'green plants' coreGFs is shown in Figure 2 (Supplementary Methods 2). For most species, assembly and annotation updates steadily increase the completeness without expanding the gene space (Figure 1 and 2). Updated gene annotations mainly correspond to improved assemblies or the application of improved gene prediction tools. Applying more advanced gene prediction algorithms can boost the annotation quality only up to a certain degree. Additional improvements can be achieved by better utilizing existing data sources, such as protein homology information, or by integrating new data sources, such as new sequencing technologies. Examples of the latter include single-molecule long-read sequencing such as PacBio Iso-seq and Nanopore RNA-seq, which are used to identify novel genes and isoforms, as well as to correct known isoform structures [35,36]. Ribo-seq provides a means to determine whether predicted isoforms and detected transcripts are also translated into proteins, without requiring a full proteomics screening [37]. The wide diversity of these novel sequencing methods highlights how they all capture different data types and provide different evidences for the gene prediction algorithms to integrate and interpret, suggesting that their combination is necessary to fully capture the gene space [38].

In contrast to protein-coding genes, few methods are currently available to assess the completeness of the RNA gene space. Most genome annotation projects use consensus secondary structures and covariance models from the RFAM database to search for different types of known RNAs in newly sequenced genomes. Despite their importance for gene regulation, and in contrast to protein-coding genes, many of the recently identified plant lncRNAs lack strong conservation. Starting from a set of 5,362 unique Arabidopsis long intergenic non-coding RNAs (lincRNAs), sequence similarity searches in 10 Brassicaceae genomes revealed that for only 22% homologs could be identified tracing back to the most recent common ancestor [39]. Interestingly, conserved lincRNAs showed higher expression levels than non-conserved lincRNAs and within the conserved set, sequence homologs were detected in all studied Brassicaceae genomes for 93 *Arabidopsis thaliana* lincRNAs. In cereals, it was found that 20% of rice lincRNAs (2,281/11,229 lincRNAs) showed detectable sequence conservation to the maize genomic sequence, although only 5% of them (117/2,281) had sequence similarity to maize lincRNAs [20]. These results indicate that, within specific taxonomic clades, small sets of conserved lncRNAs exist, which could be used to define the expected gene space for lncRNAs. Apart from modeling the expected RNA gene space based on evolutionary conservation, transcript mapping based on species-specific RNA-seq datasets offers an alternative approach where one can determine what fraction of expressed transcript reads overlaps with annotated RNA gene models [13]. Additionally, reads mapping to specific genomic regions lacking annotated genes can potentially indicate new RNA loci absent from the available genome annotation. Adapted sequencing methodologies will be required to target RNA genes, as standard RNA sequencing protocols use library preparation methods that select for poly-A tails that are present in eukaryotic mRNAs but are not always present in RNA gene transcripts [40].

Whereas improved sequencing technologies expand our view of different gene types encoded by plant genomes, incorrect reference data used in gene annotation procedures can leave genuine genes to be left unannotated in the genome. For example, a recent study reported strong differences (between

158 and 463) in the number of resistance genes for four different Brassicaceae genome annotations [41]. Different approaches for repeat masking were identified to be, at least partially, responsible for this discrepancy. The Repbase database, which collects repeat sequences used for the masking of repetitive DNA, was contaminated with motifs from transposable elements fused with resistance genes. Calling repeats using *ab initio* programs may not circumvent this issue, as these tools may still find transposable elements fused with genes. Consequently, a careful analysis of repeat sequences used for masking is mandatory to avoid over-masking and underestimating the copy number for specific gene functions [10].

No gene annotation approach is flawless and although striving for perfection could be attempted, rarely do genome consortia have the diversity of skill, time and funding to achieve it in a timely fashion [42]. Therefore, a better approach might be to use additional solutions to identify and annotate missing conserved genes, once an initial annotation has been made available. Based on orthologous protein information, OrthoFiller offers the distinct advantage to iteratively and automatically add missing gene models and as such increase the gene space completeness [43].

Balancing gene space completeness and gene annotation quality

Merely attempting to increase the number of annotated genes per genome, without paying attention to the quality of the gene models, will quickly lead to a "race-to-the-bottom" where the quality of genome projects is evaluated using misguided criteria. The quality of gene models can still be evaluated however, using a combination of different metrics [13]. Complementary to experimentally determined full-length transcript sequences, comparative metrics are frequently used to determine the number of partial genes by exploiting gene homology information from related species. As an example, BUSCO reports the putative fragmentation of the gene models based on a gene length comparison of annotated genes against their orthologs [31].

The total gene space of model species changes over time, as seen in Figure 1. This does not, however, give an indication of how many partial genes are present in each annotation version, and whether these partial genes are corrected in later versions. A historical perspective on putative full-length and partial genes in *Zea mays* provides insight into the continuous efforts of the research community to produce high-quality gene models. By using the same methodology to determine gene fragmentation as BUSCO, but using a similarity search against SwissProt [44] as reference in order to evaluate all genes, Figure 3 shows a Sankey Diagram illustrating the flow of partial and full-length genes between consecutive maize annotation versions (Supplementary Methods 3). Each new version introduces new loci into the annotation, not all of them full-length: the fraction of partial new genes is even close to 50% with the transition from RefGenV3 to RefGenV4. The same transition does re-annotate a significant portion of known partial loci into full-length genes (~55%), while the remaining known partial genes are either kept the same (~25%) in RefGenV4 or get removed (~20%). Additionally, a smaller number of full-length genes from RefGenV3 are shortened and became partials in RefGenV4. This leads to the overall view where, despite clear efforts to add new gene models and improve existing gene models, both the number of full-length genes and the fraction of partial genes slightly increases between RefGenV3 and RefGenV4. This is in contrast to *Arabidopsis thaliana*, for which both the

number and status of partial genes stay relatively constant over time (Supplementary Figure 2). By using an expression compendium of 169 data sets generated using Curse [45], a significant difference in the overall fraction of expressed genes can be shown between the complete (93%) and partial (81%) genes in RefGenV3 (Supplementary Methods 4 and Supplementary Table 4 and 5). This difference is even more outspoken in RefGenV4 where the fraction of partial genes with expression drops to 55% while the fraction of expressed complete genes remains stable (91%). Gene expression can as such be used as a means to evaluate the biological relevance of partial genes.

To evaluate and improve gene models, and consequently correct partial genes, new software tools have been developed. PASA uses spliced alignments of RNA-seq data to infer gene structures [46], with EVidenceModeler being the toolchain build on top of PASA to further improve gene models using additional evidences [46]. Rather than solely relying on the intrinsic short-read nature of Illumina RNA-seq data, LOREAN can also make use of long-read cDNA sequences to create the gene models [38]. In contrast to these methods, OMGene uses an orthology approach to optimize gene structures without requiring experimental data [47]. Application of this method yielded hundreds of modified gene structures in different plant genomes, even for a species like Arabidopsis which has gone through multiple cycles of re-annotation. For the genomes of *C. papaya* and *T. cacao* more than 900 genes were modified, suggesting that protein-coding gene models in these less well-studied genomes can still be substantially refined [47].

Conclusion

The combined use of metrics that evaluate gene quality and gene space completeness is required to confidently establish the quality of genome annotation. Furthermore, a continuous effort is needed to integrate new gene models and isoforms in a unified manner into reference genome annotations for each plant species. Ideally, such reference genome annotations comply with the following rules:

i) Genome annotations must describe protein-coding genes, RNA genes, and optionally transcript isoforms in order to fully capture the different levels of gene complexity which can be extracted from experimental data such as RNA-seq.

ii) Genome annotations need to be evaluated using completeness scores for protein-coding and RNA genes. Completeness estimation methods modelling the expected gene space based on either evolutionary conservation or transcript mapping are complementary to assess if gene loci remained undetected. Missing genes may point to biological gene loss, imperfections in repeat masking or other gene prediction artefacts.

iii) After a first round of gene prediction and quality evaluation, additional tools to further optimize gene models or workflows to detect lncRNA genes should be used.

Even in the absence of an improved genome assembly, the application of new annotation pipelines, manual curation, and the integration of novel experimental data sources are all necessary to enhance the quality of gene annotations. Continuously iterating over these processes enables approaching the optimum genome annotation, a constantly moving target.

Acknowledgments

We thank Dries Vaneechoutte for help compiling the *Zea mays* expression compendium, and Heike Sprenger for comments on the manuscript. François Bucchini is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. H2020-MSCA-ITN-2015-675752.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**:793-800.
2. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al.: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**:1112-1115.
•• 3. International Wheat Genome Sequencing Consortium: **Shifting the limits in wheat research and breeding using a fully annotated reference genome**. *Science* 2018, **361**. This study describes an improved 14.5-Gb chromosome-scale assembly for the hexaploid wheat genome allowing to resolve the inherent complexity of gene families related to important agronomic traits.
4. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al.: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**:1596-1604.
5. Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV: **The tobacco genome sequence and its comparison with those of tomato and potato**. *Nat Commun* 2014, **5**:3833.
6. Tang C, Yang M, Fang Y, Luo Y, Gao S, Xiao X, An Z, Zhou B, Zhang B, Tan X, et al.: **The rubber tree genome reveals new insights into rubber production and species adaptation**. *Nat Plants* 2016, **2**:16073.
7. Vogel A, Schwacke R, Denton AK, Usadel B, Hollmann J, Fischer K, Bolger A, Schmidt MH, Bolger ME, Gundlach H, et al.: **Footprints of parasitism in the genome of the parasitic flowering plant Cuscuta campestris**. *Nat Commun* 2018, **9**:2515.
8. Abdel-Latif A, Osman G: **Comparison of three genomic DNA extraction methods to obtain high DNA quality from maize**. *Plant Methods* 2017, **13**:1.
9. Leitch AR, Leitch IJ: **Genomic plasticity and the diversity of polyploid plants**. *Science* 2008, **320**:481-483.
10. Goerner-Potvin P, Bourque G: **Computational tools to unmask transposable elements**. *Nat Rev Genet* 2018, **19**:688-704.
11. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al.: **The Norway spruce genome sequence and conifer genome evolution**. *Nature* 2013, **497**:579-584.
12. Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marcais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, et al.: **Sequencing and assembly of the 22-gb loblolly pine genome**. *Genetics* 2014, **196**:875-890.
• 13. Veeckman E, Ruttink T, Vandepoele K: **Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences**. *Plant Cell* 2016, **28**:1759-1768. A paper which describes, using examples from several dicot and monocot genomes, pitfalls and recommendations regarding methods to estimate completeness during different steps of genome assembly and annotation.
14. Schreiber M, Stein N, Mascher M: **Genomic approaches for studying crop evolution**. *Genome Biol* 2018, **19**:140.
15. Willyard C: **New human gene tally reignites debate**. *Nature* 2018, **558**:354-355.

• 16. Zhang R, Calixto CPG, Marquez Y, Venhuizen P, Tzioutziou NA, Guo W, Spensley M, Entizne JC, Lewandowska D, Ten Have S, et al.: **A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing**. *Nucleic Acids Res* 2017, **45**:5061-5073. This paper describes a comprehensive Reference Transcript Dataset for Arabidopsis (AtRTD2) offering improved transcript-level expression quantification in the presence of alterative splicing (as demonstrated in reference Vaneechoutte et al., 2017).
17. 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**:D1003-1009.
18. Vaneechoutte D, Estrada AR, Lin YC, Loraine AE, Vandepoele K: **Genome-wide characterization of differential transcript usage in Arabidopsis thaliana**. *Plant J* 2017, **92**:1218-1231.
19. Zhang YC, Liao JY, Li ZY, Yu Y, Zhang JP, Li QF, Qu LH, Shu WS, Chen YQ: **Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice**. *Genome Biol* 2014, **15**:512.
20. Wang H, Niu QW, Wu HW, Liu J, Ye J, Yu N, Chua NH: **Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits**. *Plant J* 2015, **84**:404-416.
21. Li L, Eichten SR, Shimizu R, Petsch K, Yeh CT, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE, et al.: **Genome-wide discovery and characterization of maize long non-coding RNAs**. *Genome Biol* 2014, **15**:R40.
22. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH: **Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis**. *Plant Cell* 2012, **24**:4333-4345.
23. Liu J, Wang H, Chua NH: **Long noncoding RNA transcriptome of plants**. *Plant Biotechnol J* 2015, **13**:319-328.
24. Farfan ID, De La Fuente GN, Murray SC, Isakeit T, Huang PC, Warburton M, Williams P, Windham GL, Kolomiets M: **Genome wide association study for drought, aflatoxin resistance, and important agronomic traits of maize hybrids in the sub-tropics**. *PLoS One* 2015, **10**:e0117737.
25. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, et al.: **Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice**. *Nat Genet* 2018, **50**:278-284.
•• 26. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, et al.: **Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure**. *Nat Commun* 2017, **8**:2184. An example of a grass pan-genomics project where whole-genome *de novo* assembly and annotation of 54 *Brachypodium distachyon* lines yields a pan-genome containing nearly twice the number of genes found in any individual genome.
27. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA, et al.: **The pangenome of an agronomically important crop plant Brassica oleracea**. *Nat Commun* 2016, **7**:13390.
28. Golicz AA, Batley J, Edwards D: **Towards plant pangenomics**. *Plant Biotechnol J* 2016, **14**:1099-1105.
29. Paten B, Novak AM, Eizenga JM, Garrison E: **Genome graphs and the evolution of genome inference**. *Genome Res* 2017, **27**:665-676.
30. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**:1061-1067.
31. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO applications from quality assessments to gene prediction and phylogenomics**. *Mol Biol Evol* 2017.
32. Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K: **The flowering world: a tale of duplications**. *Trends Plant Sci* 2009, **14**:680-688.

33. Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K: **Dissecting plant genomes with the PLAZA comparative genomics platform**. *Plant Physiol* 2012, **158**:590-600.
34. De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y: **Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants**. *Proc Natl Acad Sci U S A* 2013, **110**:2898-2903.
• 35. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D: **Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing**. *Nat Commun* 2016, **7**:11708. An application of long-read technologies to transcriptome sequencing in maize results in improved gene and isoform identification as well as transcript expression quantification.
36. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C: **Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells**. *Nat Commun* 2017, **8**:16027.
37. Calviello L, Ohler U: **Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome**. *Trends Genet* 2017, **33**:728-744.
38. Cook D, Espejo Valle-Inclan J, Pajoro A, Rovenich H, Thomma B, Faino L: **Long Read Annotation (LoReAn): automated eukaryotic genome annotation based on long-read cDNA sequencing**. *Plant Physiol* 2018.
•• 39. Nelson AD, Forsythe ES, Devisetty UK, Clausen DS, Haug-Batzell AK, Meldrum AM, Frank MR, Lyons E, Beilstein MA: **A Genomic Analysis of Factors Driving lincRNA Diversification: Lessons from Plants**. *G3 (Bethesda)* 2016, **6**:2881-2891. A study applying a comparative genomic and phylogenetic approach to uncover factors influencing lincRNA detection and conservation in the plant family Brassicaceae.
40. Wierzbicki AT, Haag JR, Pikaard CS: **Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes**. *Cell* 2008, **135**:635-648.
• 41. Bayer PE, Edwards D, Batley J: **Bias in resistance gene prediction due to repeat masking**. *Nat Plants* 2018, **4**:762-765. This paper reports that different Brassicaceae genome annotations exhibit strong differences in resistance gene content, which is caused by different approaches to perform repeat masking prior gene prediction.
42. Papanicolaou A: **The life cycle of a genome project: perspectives and guidelines inspired by insect genome projects**. *F1000Res* 2016, **5**:18.
43. Dunne MP, Kelly S: **OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations**. *BMC Genomics* 2017, **18**:390.
44. UniProt Consortium T: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2018, **46**:2699.
45. Vaneechoutte D, Vandepoele K: **Curse: Building expression atlases and co-expression networks from public RNA-Seq data**. *under review* 2018.
46. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments**. *Genome Biol* 2008, **9**:R7.
• 47. Dunne MP, Kelly S: **OMGene: mutual improvement of gene models through optimisation of evolutionary conservation**. *BMC Genomics* 2018, **19**:307. The OMGene method improves gene model accuracy in the absence of experimental data by optimizing the consistency of multiple sequence alignments of orthologous proteins from multiple species.

Figures

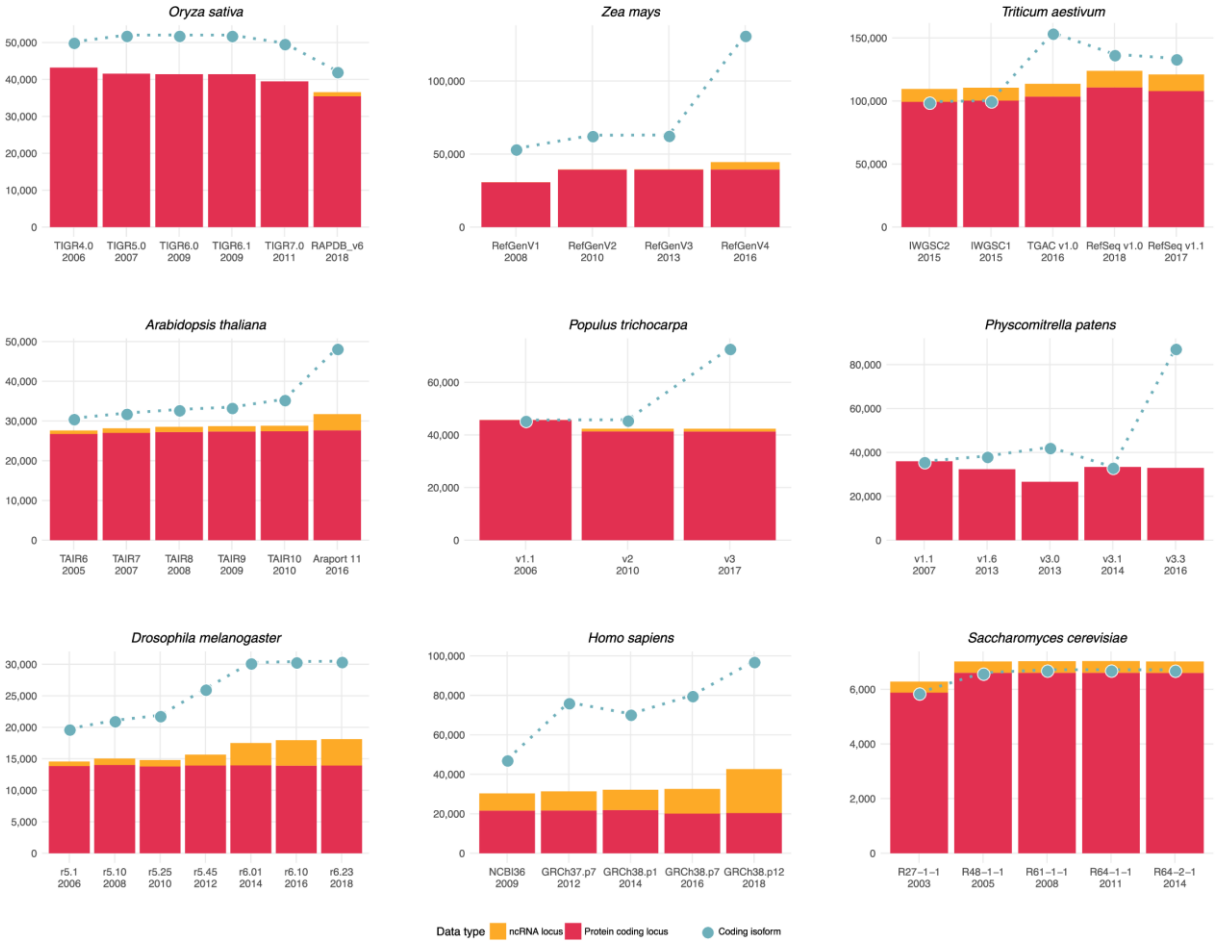


Figure 1. Historical overview of the number of annotated protein-coding loci, isoforms, and non-coding RNA loci in nine model organisms: *Oryza sativa*, *Zea mays*, *Triticum aestivum*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Physcomitrella patens*, *Homo sapiens*, *Drosophila melanogaster, Saccharomyces cerevisiae.* Data collected from various sources (see Supplemental Table 1 and 2).

Figure 2. Gene space completeness for various genome annotation versions from eight plants: *Oryza sativa*, *Zea mays*, *Triticum aestivum*, *Arabidopsis thaliana*, *Populus trichocarpa*, *Physcomitrella patens*, *Theobroma cacao*, and *Carica papaya*. Gene space completeness was determined using BUSCO Plants and PLAZA 'green plants' coreGFs.

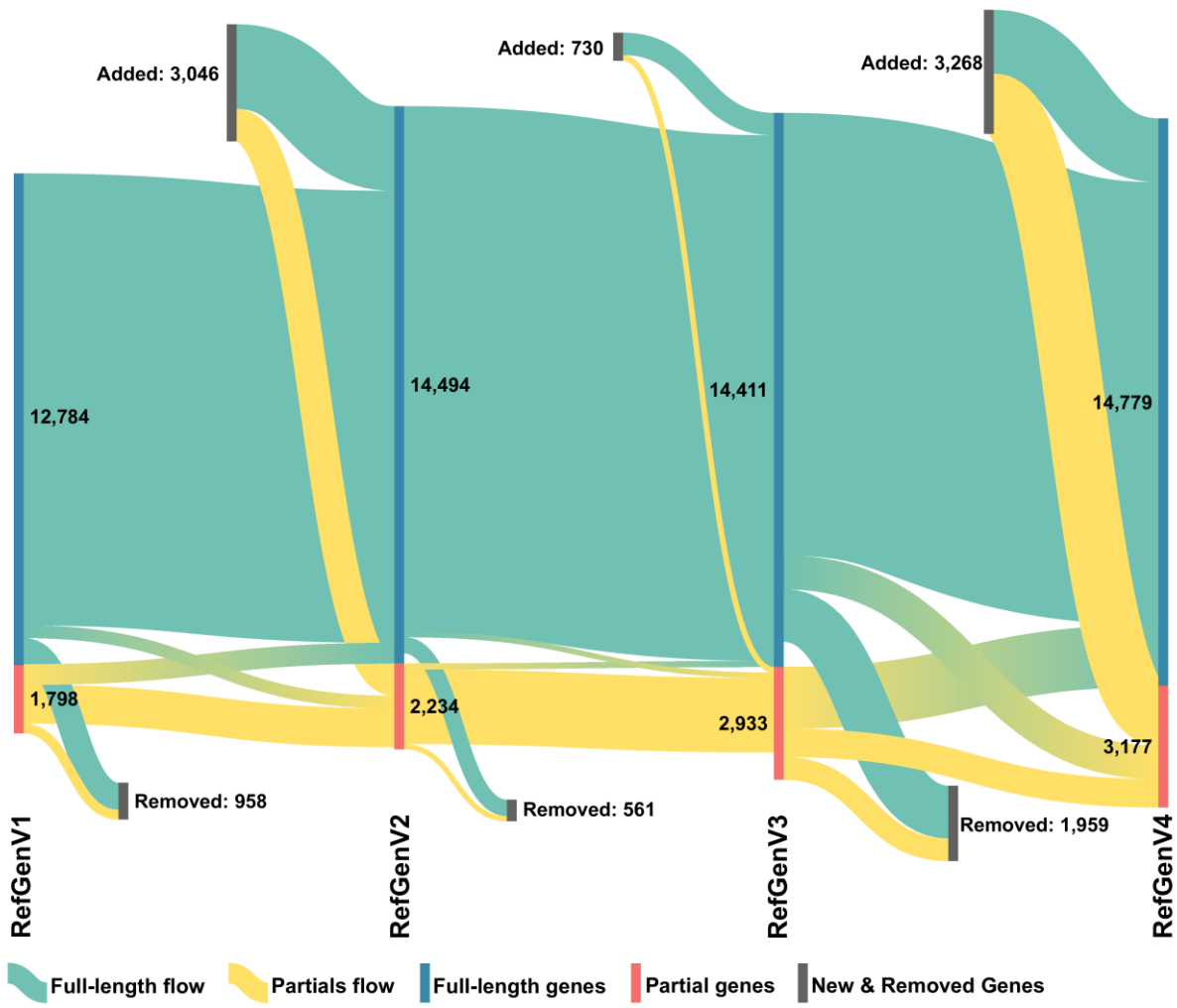


Figure 3. Historical overview over consecutive gene annotation versions of putative partial genes in the model organism *Zea mays*. The figure only contains genes for which at least 10 hits were present in the SwissProt reference proteome database, in order to prevent the incorrect inference of gene fragmentation.

Supplementary Data - Gene space completeness in complex plant genomes

Michiel Van Bel[1,2], François Bucchini[1,2] and Klaas Vandepoele[1,2,3*]

[1] Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium

[3] Bioinformatics Institute Ghent, Ghent University, 9052 Ghent, Belgium

[*] To whom correspondence should be addressed. Tel: +32 9 3313822; Fax: +32 9 3313809; Email: klaas.vandepoele@psb.vib-ugent.be

Supplementary Methods

Supplementary Method 1: Data processing Figure 1

We gathered annotation and assembly information for a set of model organisms (see Supplementary Table 1 for sources). These annotation data sets consisted of a variety of file types: protein FASTA files, ncRNA FASTA files, GFF3 files, text files, etc. Since not all genome projects provide their annotation data in the same manner (often due to the wide variety within the GFF3 comment field to indicate feature information), custom scripts and Linux commands were used for each organism and version. For example:

1) Using different methods to determine the number of isoforms and protein-coding loci in *Drosophila melanogaster*, based on different input files in order to assert the correctness of the results:

Determining the number of isoforms from the protein FASTA file:

```
zcat dmel-all-translation-r6.23.fasta.gz | grep '>' | grep 'type=protein' | wc -l
```

Determining the number of protein loci from the protein FASTA file:

```
zcat dmel-all-translation-r6.23.fasta.gz | grep '>' | grep 'type=protein' | cut -d ' ' -f 6 |
cut -d '=' -f 2 | cut -d ',' -f 1 | sort -u | wc -l
```

Determining the number of isoforms from the GFF3 file:

```
zcat dmel-all-r6.23.gff.gz | grep -P '\tmRNA\t' | wc -l
```

Determining the number of protein loci from the GFF3 file:

```
zcat dmel-all-r6.23.gff.gz | grep -P '\tgene\t' | grep UniProt | wc -l`
```

2) Using different methods to determine the number of isoforms and protein-coding loci in *Physcomitrella patens*, based on different input files in order to assert the correctness of the results:

Determining the number of isoforms from the protein FASTA file (v1.1):

```
zcat proteins.Phypa1_1.FilteredModels.fasta.gz | grep -c ">"
```

Determining the number of isoforms from the GFF3 file (v1.1):

```
zcat Phypa1_1.FilteredModels.gff.gz | grep -v "#" | grep -P '\tCDS\t' | cut -f 9 | cut -d ';'
-f 2 | cut -d ' ' -f 3 | sort -u  | wc -
```

Supplementary Method 2: Data processing Figure 2

Protein FASTA files corresponding to various genome annotation versions were gathered for eight plant species (Supplementary Table 1): *Oryza sativa*, *Zea mays*, *Triticum aestivum, Arabidopsis thaliana*, *Populus trichocarpa*, *Physcomitrella patens*, *Theobroma cacao*, and *Carica papaya*. Gene space completeness was then estimated using BUSCO [1] and PLAZA 'green plants' core gene families (coreGFs) [2]. BUSCO 3.0 was run in protein mode, using 1,440 *Embryophyta* profiles selected from OrthoDB v9. The set of PLAZA coreGFs used to perform completeness analyses consists of 2,928 'green plants' gene families that are highly conserved in 25 *Viridiplantae* species spanning angiosperms, mosses, and green algae, retrieved from PLAZA 2.5 (ftp://ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/coreGF/coreGF_plaza2.5_greenplants.txt).
CoreGF completeness analysis was performed as described in [3] but including a filtering of partial genes, using the following protocol: i) for a query proteome a protein similarity search against the PLAZA 2.5 protein database was run using DIAMOND 0.9.18 [4] in 'more sensitive' mode with a maximum e-value cut-off of 10^{-5} ii) for each query protein, self-hits were excluded and the gene

family associated with the top hit was retrieved. Only hits where the alignment length is larger than 50% of the median protein length of the gene family members were considered (filtering of partial genes), and iii) the core GF completeness score was calculated as the sum of the weights of represented coreGFs divided by the total weights of all 2,928 coreGFs.

Supplementary Method 3: Data processing Figure 3

The basis for this figure are the various *Zea mays* gene annotations, gathered as described in Supplementary Method 1. The next step was performing a protein sequence similarity search (e-value cutoff 1e-05) against SwissProt (release September 2018) [5] for the protein FASTA file of each *Zea mays* genome annotation version. These results were then post processed, and partial genes were delineated using a similar process as the one used in BUSCO to detect fragmented genes [1]: for each query protein, its length was compared against the lengths of the hit proteins. This comparison consists of checking whether the query protein length is smaller than the mean length of the hit proteins minus two times the standard deviation of the lengths of the hit protein. While in BUSCO this sequence length comparison is done against the length of the profile to which a protein is assigned, we needed a measurement to could deal with all sequences present in the proteome (as BUSCO only has a subset of the proteins integrated in the profiles representing conserved orthologous genes), hence the use of SwissProt. To avoid the under- or overestimation of the fraction of partial genes, we only took into account those proteins for which there were 10 or more BLASTP hits. Genes for which there were no hits or less than 10 hits were excluded from the analysis. Subsequently, the partials per *Zea mays* version were compared against the partials of the previous version using common gene identifiers information (i.e. RefGenV2 is compared to RefGenV1, etc.). We used the gene identifier conversion table provided at Gramene for the conversion between RefGenV3 and RefGenV4, due to the different gene identifier construction scheme. This way, a structured overview of the data flows between *Zea mays* versions was constructed. The Sankey diagram was constructed using the online software tool SankeyMatic (http://sankeymatic.com/build/), with post processing of the resulting SVG in InkScape.

Importantly, the indication that a gene is partial is only putative: it is possible that through evolution genes lose protein domains and remain functional, and the similarity search is also an imperfect measurement, both in terms of method and in terms of reference database. However, to compare different annotation versions we here assume that genes detected as putatively fragmented are a proxy for partial gene models.

The choice to use SwissProt as the reference gold standard database, and not any other databases such as PLAZA [6], was made because the SwissProt database only contains curated protein sequences. Although the SwissProt database does contain relatively few plant protein sequences (https://web.expasy.org/docs/relnotes/relstat.html), quality was valued over quantity in order correctly quantify the number of partial proteins.

Supplementary Methods 4: *Zea mays* expression analysis

Starting from publicly available RNA-Seq expression data, the Curse/Prose suite [7] was used to generate a maize B73 expression atlas covering 169 different samples comprising different organs

and stress conditions. After running Kallisto [8] using the RefGenV3 and RefGenV4 transcript files, genes were considered expressed if the TPM value in at least one of the 169 samples was higher or equal than two (TPM>=2). For each gene evaluated as being expressed, the number of experiments in which the TPM value is two or higher, was counted. For the various gene sets (e.g. all the complete genes in RefGenV3), these experiment counts were averaged for the expressed genes (Supplementary Table 3).

The TPM cutoff value was selected based on the paper 'Genome-Wide Analysis of Alternative Splicing in Zea mays: Landscape and Genetic Regulation' [9], which determines the optimal FPKM cutoff for isoform detection at FPKM=1.3. Based on this result, the TPM cutoff was set at 2.0 in order to be fully confident that the observed expression values of these (putative) partial genes are not due to spurious hits.

Supplementary Tables

Supplementary Table 1: Data sources Figure 1

Gathering the data and statistics for the various gene annotation versions of the model species was more complicated than anticipated: not only was finding the data non-trivial, it was also important to try and limit the annotations to a single provider, in order to reduce variations introduced by different gene annotation pipelines.

Species	Annotation Version	Year	Source
Arabidopsis thaliana	TAIR6	2005	ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR6_genome_release/README_TAIR6
Arabidopsis thaliana	TAIR7	2007	ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR7_genome_release/README
Arabidopsis thaliana	TAIR8	2008	ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release/README
Arabidopsis thaliana	TAIR9	2009	ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR9_genome_release/readme_TAIR9.txt
Arabidopsis thaliana	TAIR10	2010	ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/README_TAIR10.txt
Arabidopsis thaliana	Araport 11	2016	https://www.araport.org/data/araport11
Homo sapiens	NCBI36 (Ensembl 54)	2009	ftp://ftp.ensembl.org/pub/release-54/fasta/homo_sapiens/
Homo sapiens	GRCh37.p7 (Ensembl 67)	2012	ftp://ftp.ensembl.org/pub/release-67/fasta/homo_sapiens/
Homo sapiens	GRCh38.p1 (Ensembl v78)	2014	ftp://ftp.ensembl.org/pub/release-78/
Homo sapiens	GRCh38.p7 (Ensembl v87)	2016	ftp://ftp.ensembl.org/pub/release-87/gff3/homo_sapiens/
Homo sapiens	GRCh38.p12 (Ensembl v93)	2018	ftp://ftp.ensembl.org/pub/release-93/gff3/homo_sapiens/
Zea mays	release-4a.53 (RefGen_v1)	2008	PLAZA 2.0

Zea mays	release-5a (RefGen_v2)	2010	PLAZA 3.0
Zea mays	release-5b (RefGen_v3)	2013	PLAZA CNB 2.0
Zea mays	RefGen_v4	2016	PLAZA 4.0
Oryza sativa	TIGR4.0	2006	http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/
Oryza sativa	TIGR5.0	2007	http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/
Oryza sativa	TIGR6.0	2009	http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/
Oryza sativa	TIGR6.1	2009	http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/
Oryza sativa	TIGR7.0	2011	http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/
Oryza sativa	RAPDB v6	2018	https://rapdb.dna.affrc.go.jp/download/irgsp1.html
Drosophila melanogaster	r5.1	2006	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/
Drosophila melanogaster	r5.10	2008	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/
Drosophila melanogaster	r5.25	2010	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/
Drosophila melanogaster	r5.45	2012	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/
Drosophila melanogaster	r6.01	2014	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/
Drosophila melanogaster	r6.10	2016	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/
Drosophila melanogaster	r6.23	2018	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/
Saccharomyces cerevisiae	R27-1-1 (UCSC sacCer1)	2003	https://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/
Saccharomyces cerevisiae	R48-1-1	2005	https://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/

Saccharomyces cerevisiae	R61-1-1 (UCSC sacCer2)	2008	https://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/
Saccharomyces cerevisiae	R64-1-1 (UCSC sacCer3)	2011	https://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/
Saccharomyces cerevisiae	R64-2-1	2014	https://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/
Triticum aestivum	IWGSC2 (PlantEnsembl 25)	2015	ftp://ftp.ensemblgenomes.org/pub/release-25/plants/gff3/triticum_aestivum/
Triticum aestivum	IWGSC1+POPSEQ (PlantEnsembl 29)	2015	ftp://ftp.ensemblgenomes.org/pub/release-29/plants/gff3/triticum_aestivum/
Triticum aestivum	TGAC v1.0 (PlantEnsembl 32)	2016	ftp://ftp.ensemblgenomes.org/pub/release-32/plants/gff3/triticum_aestivum/
Triticum aestivum	RefSeq 1.0 (PlantEnsembl 40)	2018	ftp://ftp.ensemblgenomes.org/pub/release-40/plants/gff3/triticum_aestivum
Physcomitrella patens	v1.1	2007	https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Ppatens
Physcomitrella patens	v1.6	2013	https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Ppatens
Physcomitrella patens	v3.0	2013	https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Ppatens
Physcomitrella patens	v3.1	2014	https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Ppatens
Physcomitrella patens	v3.3	2016	https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Ppatens
Populus trichocarpa	V1.1	2006	PLAZA 1.0
Populus trichocarpa	V2.0	2010	PLAZA 3.0
Populus trichocarpa	V3.0	2017	PLAZA 4.0

Supplementary Table 2: Raw data Figure 1

Species	Version	Year	# ncRNA loci	#Protein coding loci	#Coding Isoforms
Arabidopsis thaliana	TAIR6	2005	838	26751	30690
Arabidopsis thaliana	TAIR7	2007	1123	27029	32007
Arabidopsis thaliana	TAIR8	2008	1288	27235	32916
Arabidopsis thaliana	TAIR9	2009	1312	27379	33501
Arabidopsis thaliana	TAIR10	2010	1359	27416	35485
Arabidopsis thaliana	Araport 11	2016	4063	27655	48359
Homo sapiens	NCBI36	2009	8663	21627	47458
Homo sapiens	GRCh37.p7	2012	9594	21715	76454
Homo sapiens	GRCh38.p1	2014	10295	21871	70608
Homo sapiens	GRCh38.p7	2016	12502	20090	80058
Homo sapiens	GRCh38.p12	2018	22210	20385	97334
Zea mays	RefGenV1	2008	80	30678	53764
Zea mays	RefGenV2	2010	38	39323	62929
Zea mays	RefGenV3	2013	156	39323	63074
Zea mays	RefGenV4	2016	4976	39498	131496
Oryza sativa	TIGR4.0	2006	0	43208	50159
Oryza sativa	TIGR5.0	2007	0	41561	52026
Oryza sativa	TIGR6.0	2009	0	41407	52031
Oryza sativa	TIGR6.1	2009	0	41386	52010
Oryza sativa	TIGR7.0	2011	0	39102	49119
Oryza sativa	TIGR7.0	2011	0	39460	49812
Oryza sativa	RAPDB_v6	2018	1080	35452	42229
Drosophila melanogaster	r5.1	2006	711	13854	19781
Drosophila melanogaster	r5.10	2008	1004	14023	21099
Drosophila melanogaster	r5.25	2010	1018	13781	21909
Drosophila melanogaster	r5.45	2012	1718	13927	26108
Drosophila melanogaster	r6.01	2014	3534	13953	30276
Drosophila melanogaster	r6.10	2016	4030	13907	30438
Drosophila melanogaster	r6.23	2018	4189	13931	30506
Saccharomyces cerevisiae	R27-1-1	2003	400	5878	5878
Saccharomyces cerevisiae	R48-1-1	2005	412	6604	6604
Saccharomyces cerevisiae	R61-1-1	2008	418	6607	6717
Saccharomyces cerevisiae	R64-1-1	2011	424	6607	6717
Saccharomyces cerevisiae	R64-2-1	2014	417	6600	6713
Triticum aestivum	IWGSC2	2015	10121	99354	99354
Triticum aestivum	IWGSC1+POPSEQ	2015	10121	100342	100344
Triticum aestivum	TGAC v1.0	2016	10076	103539	154102
Triticum aestivum	RefSeq v1.0	2018	13044	110790	137056
Triticum aestivum	RefSeq v1.1	2017	13044	107891	133744
Populus trichocarpa	v1.1	2006	5	45654	45654
Populus trichocarpa	v2	2010	967	41376	45778
Populus trichocarpa	v3	2017	1012	41335	73012
Physcomitrella patens	v1.1	2007	0	35938	35938
Physcomitrella patens	v1.6	2013	0	32273	38354
Physcomitrella patens	v3.0	2013	0	26610	42392
Physcomitrella patens	v3.1	2014	0	33362	33362
Physcomitrella patens	v3.3	2016	0	32926	87533

Supplementary Table 3: Functional biases CoreGF versus single-copy families

This table lists GO terms that are significantly (p-value<=0.05) enriched in CoreGFs (n=7398), but are not enriched in (quasi) single-copy gene families (n=498). The functional information and gene family data are taken from PLAZA Dicots 4.0 [6]. The GO terms shown in the table are at a minimum depth of 5 in the GO graph in order to only present the most informative terms.

GO Term	GO Aspect	Description
GO:0006612	BP	protein targeting to membrane
GO:0072330	BP	monocarboxylic acid biosynthetic process
GO:0006605	BP	protein targeting
GO:0000398	BP	mRNA splicing, via spliceosome
GO:0071396	BP	cellular response to lipid
GO:0000280	BP	nuclear division
GO:0042727	BP	flavin-containing compound biosynthetic process
GO:0072350	BP	tricarboxylic acid metabolic process
GO:0048278	BP	vesicle docking
GO:0072599	BP	establishment of protein localization to endoplasmic reticulum
GO:1901663	BP	quinone biosynthetic process
GO:0045047	BP	protein targeting to ER
GO:0006400	BP	tRNA modification
GO:0006401	BP	RNA catabolic process
GO:0009902	BP	chloroplast relocation
GO:0016570	BP	histone modification
GO:0016573	BP	histone acetylation
GO:0051603	BP	proteolysis involved in cellular protein catabolic process
GO:0006650	BP	glycerophospholipid metabolic process
GO:0006414	BP	translational elongation
GO:0006412	BP	translation
GO:0046394	BP	carboxylic acid biosynthetic process
GO:0046395	BP	carboxylic acid catabolic process
GO:0006767	BP	water-soluble vitamin metabolic process
GO:0072657	BP	protein localization to membrane
GO:0072655	BP	establishment of protein localization to mitochondrion
GO:0019941	BP	modification-dependent protein catabolic process
GO:0006633	BP	fatty acid biosynthetic process
GO:0006511	BP	ubiquitin-dependent protein catabolic process
GO:0006626	BP	protein targeting to mitochondrion
GO:0006505	BP	GPI anchor metabolic process
GO:0030833	BP	regulation of actin filament polymerization
GO:0006506	BP	GPI anchor biosynthetic process
GO:0006568	BP	tryptophan metabolic process
GO:0043623	BP	cellular protein complex assembly
GO:0002097	BP	tRNA wobble base modification
GO:0002098	BP	tRNA wobble uridine modification
GO:0043632	BP	modification-dependent macromolecule catabolic process
GO:0006576	BP	cellular biogenic amine metabolic process
GO:0051258	BP	protein polymerization
GO:0006206	BP	pyrimidine nucleobase metabolic process
GO:0006207	BP	'de novo' pyrimidine nucleobase biosynthetic process
GO:0019750	BP	chloroplast localization
GO:0016485	BP	protein processing
GO:0019752	BP	carboxylic acid metabolic process
GO:0043648	BP	dicarboxylic acid metabolic process
GO:0006661	BP	phosphatidylinositol biosynthetic process
GO:0006664	BP	glycolipid metabolic process
GO:0006544	BP	glycine metabolic process
GO:0006418	BP	tRNA aminoacylation for protein translation
GO:0006779	BP	porphyrin-containing compound biosynthetic process
GO:0016579	BP	protein deubiquitination
GO:0006553	BP	lysine metabolic process
GO:0044743	BP	protein transmembrane import into intracellular organelle
GO:0006547	BP	histidine metabolic process
GO:0019856	BP	pyrimidine nucleobase biosynthetic process
GO:0006366	BP	transcription from RNA polymerase II promoter
GO:0006367	BP	transcription initiation from RNA polymerase II promoter
GO:0006497	BP	protein lipidation
GO:0019674	BP	NAD metabolic process

GO:0043604	BP	amide biosynthetic process
GO:0006586	BP	indolalkylamine metabolic process
GO:0006352	BP	DNA-templated transcription, initiation
GO:0042886	BP	amide transport
GO:0006357	BP	regulation of transcription from RNA polymerase II promoter
GO:0048193	BP	Golgi vesicle transport
GO:0018205	BP	peptidyl-lysine modification
GO:0006289	BP	nucleotide-excision repair
GO:0046854	BP	phosphatidylinositol phosphorylation
GO:0042255	BP	ribosome assembly
GO:0019363	BP	pyridine nucleotide biosynthetic process
GO:0034622	BP	cellular macromolecular complex assembly
GO:0071806	BP	protein transmembrane transport
GO:0065002	BP	intracellular protein transmembrane transport
GO:0006298	BP	mismatch repair
GO:0019359	BP	nicotinamide nucleotide biosynthetic process
GO:0006260	BP	DNA replication
GO:0006261	BP	DNA-dependent DNA replication
GO:0042278	BP	purine nucleoside metabolic process
GO:0042158	BP	lipoprotein biosynthetic process
GO:0009658	BP	chloroplast organization
GO:0016071	BP	mRNA metabolic process
GO:0016073	BP	snRNA metabolic process
GO:0032787	BP	monocarboxylic acid metabolic process
GO:0008213	BP	protein alkylation
GO:0006397	BP	mRNA processing
GO:0016180	BP	snRNA processing
GO:0070972	BP	protein localization to endoplasmic reticulum
GO:0008380	BP	RNA splicing
GO:0006084	BP	acetyl-CoA metabolic process
GO:0006085	BP	acetyl-CoA biosynthetic process
GO:0009119	BP	ribonucleoside metabolic process
GO:0009117	BP	nucleotide metabolic process
GO:0018193	BP	peptidyl-amino acid modification
GO:0008154	BP	actin polymerization or depolymerization
GO:0009247	BP	glycolipid biosynthetic process
GO:0046834	BP	lipid phosphorylation
GO:0006072	BP	glycerol-3-phosphate metabolic process
GO:0030041	BP	actin filament polymerization
GO:0009108	BP	coenzyme biosynthetic process
GO:0008064	BP	regulation of actin polymerization or depolymerization
GO:0033014	BP	tetrapyrrole biosynthetic process
GO:0009165	BP	nucleotide biosynthetic process
GO:0000105	BP	histidine biosynthetic process
GO:0046474	BP	glycerophospholipid biosynthetic process
GO:0033365	BP	protein localization to organelle
GO:0015833	BP	peptide transport
GO:0070585	BP	protein localization to mitochondrion
GO:0072525	BP	pyridine-containing compound biosynthetic process
GO:0072528	BP	pyrimidine-containing compound biosynthetic process
GO:0046488	BP	phosphatidylinositol metabolic process
GO:0015994	BP	chlorophyll metabolic process
GO:0000377	BP	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO:0070647	BP	protein modification by small protein conjugation or removal
GO:0071616	BP	acyl-CoA biosynthetic process
GO:0009070	BP	serine family amino acid biosynthetic process
GO:0070646	BP	protein modification by small protein removal
GO:0009085	BP	lysine biosynthetic process
GO:0009089	BP	lysine biosynthetic process via diaminopimelate
GO:0000027	BP	ribosomal large subunit assembly
GO:0043039	BP	tRNA aminoacylation
GO:0000387	BP	spliceosomal snRNP assembly
GO:0043043	BP	peptide biosynthetic process
GO:0044257	BP	cellular protein catabolic process
GO:0009067	BP	aspartate family amino acid biosynthetic process
GO:0009066	BP	aspartate family amino acid metabolic process
GO:0009069	BP	serine family amino acid metabolic process
GO:0000375	BP	RNA splicing, via transesterification reactions
GO:1901607	BP	alpha-amino acid biosynthetic process
GO:0016591	CC	DNA-directed RNA polymerase II, holoenzyme
GO:0009941	CC	chloroplast envelope

GO:0022626	CC	cytosolic ribosome
GO:0000793	CC	condensed chromosome
GO:0032040	CC	small-subunit processome
GO:0016887	MF	ATPase activity
GO:0016407	MF	acetyltransferase activity
GO:0016410	MF	N-acyltransferase activity
GO:0070035	MF	purine NTP-dependent helicase activity
GO:0004519	MF	endonuclease activity
GO:0015405	MF	P-P-bond-hydrolysis-driven transmembrane transporter activity
GO:0000049	MF	tRNA binding
GO:0003678	MF	DNA helicase activity
GO:0004527	MF	exonuclease activity
GO:0002161	MF	aminoacyl-tRNA editing activity
GO:0004222	MF	metalloendopeptidase activity
GO:0030983	MF	mismatched DNA binding
GO:0019843	MF	rRNA binding
GO:0019829	MF	cation-transporting ATPase activity
GO:0005525	MF	GTP binding
GO:0005524	MF	ATP binding
GO:0016791	MF	phosphatase activity
GO:0016417	MF	S-acyltransferase activity
GO:0015035	MF	protein disulfide oxidoreductase activity
GO:0004386	MF	helicase activity
GO:0008757	MF	S-adenosylmethionine-dependent methyltransferase activity
GO:0004003	MF	ATP-dependent DNA helicase activity
GO:0016462	MF	pyrophosphatase activity
GO:0019201	MF	nucleotide kinase activity
GO:0019205	MF	nucleobase-containing compound kinase activity
GO:0042625	MF	ATPase coupled ion transmembrane transporter activity
GO:0008312	MF	7S RNA binding
GO:0042623	MF	ATPase activity, coupled
GO:0017136	MF	NAD-dependent histone deacetylase activity
GO:0017016	MF	Ras GTPase binding
GO:0019783	MF	ubiquitin-like protein-specific protease activity
GO:0008536	MF	Ran GTPase binding
GO:0008408	MF	3'-5' exonuclease activity
GO:0017111	MF	nucleoside-triphosphatase activity
GO:0030554	MF	adenyl nucleotide binding
GO:0019001	MF	guanyl nucleotide binding
GO:0032561	MF	guanyl ribonucleotide binding
GO:0032559	MF	adenyl ribonucleotide binding
GO:0034979	MF	NAD-dependent protein deacetylase activity
GO:0032550	MF	purine ribonucleoside binding
GO:0008026	MF	ATP-dependent helicase activity
GO:0003924	MF	GTPase activity
GO:0044769	MF	ATPase activity, coupled to transmembrane movement of ions, rotational mechanism
GO:0031267	MF	small GTPase binding
GO:0008135	MF	translation factor activity, RNA binding
GO:0004812	MF	aminoacyl-tRNA ligase activity
GO:0003729	MF	mRNA binding
GO:0031078	MF	histone deacetylase activity (H3-K14 specific)
GO:0032041	MF	NAD-dependent histone deacetylase activity (H3-K14 specific)
GO:0003951	MF	NAD+ kinase activity
GO:0003887	MF	DNA-directed DNA polymerase activity
GO:0003743	MF	translation initiation factor activity
GO:0008080	MF	N-acetyltransferase activity
GO:0004843	MF	thiol-dependent ubiquitin-specific protease activity

Supplementary Table 4: *Zea mays* expression information for RefGenV3 and RefGenV4 subsets shown in Figure 3.

Version	Partialness Type	Subset	#Genes	#Expressed Genes	%Expressed Genes	Average (#samples expressed)	%Average (#samples expressed)
RefGenV3	complete	all	14,411	13,370	92.78%	107	63.31%
		consistent	13,674	12,714	92.98%	107	63.31%
		new	583	504	86.45%	91	53.85%
		removed	1,370	1,142	83.36%	88	52.07%
	partial	all	2,226	1,795	80.64%	101	59.76%
		consistent	1,944	1,557	80.09%	99	58.58%
		new	147	106	72.11%	87	51.48%
		removed	589	436	74.02%	85	50.30%
RefGenV4	complete	all	14,779	13,389	90.59%	104	61.54%
		consistent	11,515	10,725	93.14%	105	62.13%
		new	1,676	1,177	70.23%	87	51.48%
	partial	all	3,177	1,757	55.30%	81	47.93%
		consistent	740	473	63.92%	87	51.48%
		new	1,592	625	39.26%	48	28.40%

Supplementary Table 5: *Zea mays* expression compendium.

#studyAccession	experimentAccession	sampleAccession	runAccessions	groupName	studyTitle
SRP109003	SRX2909945	SRS2276526	SRR5674388	10mm_ear_primordia	RNA-seq of maize ear primordia in response to drought
SRP109003	SRX2909946	SRS2276527	SRR5674387	10mm_ear_primordia	RNA-seq of maize ear primordia in response to drought
SRP109003	SRX2909947	SRS2276528	SRR5674386	10mm_ear_primordia	RNA-seq of maize ear primordia in response to drought
SRP109545	SRX2925340	SRS2290611	SRR5691823	1d_BGM_infested	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP109545	SRX2925341	SRS2290612	SRR5691824	1d_BGM_infested	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP109545	SRX2925324	SRS2290594	SRR5691807	1d_wounded	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP109545	SRX2925325	SRS2290596	SRR5691808	1d_wounded	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP029238	SRX339789	SRS473613	SRR957448	2-4mm_ear_primordioum_tip	Maize proteome and transcriptome atlas
SRP029238	SRX339790	SRS473612	SRR957449	2-4mm_ear_primordioum_tip	Maize proteome and transcriptome atlas
SRP029238	SRX339791	SRS473614	SRR957450	2-4mm_ear_primordioum_tip	Maize proteome and transcriptome atlas
SRP109545	SRX2925336	SRS2290608	SRR5691819	2g_BGM_infested	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP109545	SRX2925337	SRS2290607	SRR5691820	2g_BGM_infested	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP109545	SRX2925320	SRS2290590	SRR5691799	2h_wounded	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP109545	SRX2925321	SRS2290592	SRR5691801	2h_wounded	Maize transcriptional responses to specialist and generalist spider mite herbivores
SRP101911	SRX2641029	SRS2048553	SRR5344570,SRR5344571	3d_drought	Zea mays cultivar:B73 Raw sequence reads
SRP114563	SRX3054171	SRS2400434	SRR5888346	4d_heat_stress_leaf	Zea mays Transcriptome or Gene expression
SRP114563	SRX3054175	SRS2400430	SRR5888342	4d_heat_stress_leaf	Zea mays Transcriptome or Gene expression
SRP114563	SRX3054195	SRS2400411	SRR5888322	4d_heat_stress_leaf	Zea mays Transcriptome or Gene expression
SRP109003	SRX2909951	SRS2276531	SRR5674382	5mm_ear_primordia	RNA-seq of maize ear primordia in response to drought
SRP109003	SRX2909952	SRS2276533	SRR5674381	5mm_ear_primordia	RNA-seq of maize ear primordia in response to drought
SRP109003	SRX2909953	SRS2276534	SRR5674380	5mm_ear_primordia	RNA-seq of maize ear primordia in response to drought
SRP029238	SRX339786	SRS473609	SRR957445	6-8mm_ear_primordium_tip	Maize proteome and transcriptome atlas
SRP029238	SRX339787	SRS473610	SRR957446	6-8mm_ear_primordium_tip	Maize proteome and transcriptome atlas

SRP029238	SRX339788	SRS473611	SRR957447	6-8mm_ear_primordium_tip	Maize proteome and transcriptome atlas
ERP024506	ERX2154021	ERS1871579	ERR2096645,ERR2096644	bract	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154022	ERS1871580	ERR2096647,ERR2096646	bract	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154023	ERS1871581	ERR2096648,ERR2096649	bract	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
SRP095179	SRX2425991	SRS1862008	SRR5113836	F.graminearum_kernel	Transcriptome profiling of two inbred lines with distinct responses to Gibberella ear rot disease identified candidate genes for resistance in maize
SRP095179	SRX2425992	SRS1862009	SRR5113837	F.graminearum_kernel	Transcriptome profiling of two inbred lines with distinct responses to Gibberella ear rot disease identified candidate genes for resistance in maize
SRP097894	SRX3558102	SRS2831332	SRR6468136	heat_stress	Transcriptome analysis of B vitamin deficiency in plants
SRP106663	SRX2792109	SRS2173583	SRR6807863	heat_stress	Zea mays subsp. mays Raw sequence reads
SRP106663	SRX2792110	SRS2173584	SRR6807864	heat_stress	Zea mays subsp. mays Raw sequence reads
SRP106663	SRX2792111	SRS2173585	SRR6807865	heat_stress	Zea mays subsp. mays Raw sequence reads
SRP111315	SRX3045731	SRS2343174	SRR5878414	bract	Zea mays Mo17 Genome sequencing and assembly
SRP135909	SRX3804718	SRS3198041	SRR6849465	endosperm	Zea mays strain:B73 Transcriptome or Gene expression
SRP135909	SRX3804723	SRS3198045	SRR6849460	endosperm	Zea mays strain:B73 Transcriptome or Gene expression
SRP097894	SRX2520724	SRS1942333	SRR5206843	apical_meristem	Transcriptome analysis of B vitamin deficiency in plants
SRP097894	SRX2520725	SRS1942330	SRR5206844	apical_meristem	Transcriptome analysis of B vitamin deficiency in plants
SRP097894	SRX2520726	SRS1942332	SRR5206845	apical_meristem	Transcriptome analysis of B vitamin deficiency in plants
SRP106663	SRX2792105	SRS2173578	SRR6807855	cold_stress	Zea mays subsp. mays Raw sequence reads
SRP106663	SRX2792107	SRS2173580	SRR6807856	cold_stress	Zea mays subsp. mays Raw sequence reads
SRP106663	SRX2792108	SRS2173582	SRR6807858	cold_stress	Zea mays subsp. mays Raw sequence reads
SRP135909	SRX3804720	SRS3198042	SRR6849463	coleoptilar_nodes	Zea mays strain:B73 Transcriptome or Gene expression
SRP135909	SRX3804727	SRS3198040	SRR6849456	coleoptilar_nodes	Zea mays strain:B73 Transcriptome or Gene expression
SRP106663	SRX2792102	SRS2173575	SRR5903687	control_stress	Zea mays subsp. mays Raw sequence reads
SRP106663	SRX2792103	SRS2173576	SRR6807861	control_stress	Zea mays subsp. mays Raw sequence reads
SRP106663	SRX2792104	SRS2173577	SRR6807862	control_stress	Zea mays subsp. mays Raw sequence reads
SRP135909	SRX3804721	SRS3198043	SRR6849462	ears	Zea mays strain:B73 Transcriptome or Gene expression

SRP135909	SRX3804725	SRS3198039	SRR6849458	ears	Zea mays strain:B73 Transcriptome or Gene expression
SRP079373	SRX1973316	SRS1581562	SRR3948319	embryo	Transcriptome of rtcs and wild-type embryos
SRP079373	SRX1973317	SRS1581563	SRR3948320	embryo	Transcriptome of rtcs and wild-type embryos
SRP079373	SRX1973318	SRS1581564	SRR3948321	embryo	Transcriptome of rtcs and wild-type embryos
SRP029238	SRX339777	SRS473601	SRR957436	embryos	Maize proteome and transcriptome atlas
SRP029238	SRX339778	SRS473600	SRR957437	embryos	Maize proteome and transcriptome atlas
SRP029238	SRX339774	SRS473597	SRR957433	embryos_20DAP	Maize proteome and transcriptome atlas
SRP029238	SRX339775	SRS473598	SRR957434	embryos_20DAP	Maize proteome and transcriptome atlas
SRP029238	SRX339776	SRS473599	SRR957435	embryos_20DAP	Maize proteome and transcriptome atlas
SRP029238	SRX339779	SRS473602	SRR957438	embryos	Maize proteome and transcriptome atlas
SRP029238	SRX339756	SRS473579	SRR957415	endosperm_12DAP	Maize proteome and transcriptome atlas
SRP029238	SRX339757	SRS473580	SRR957416	endosperm_12DAP	Maize proteome and transcriptome atlas
SRP029238	SRX339758	SRS473581	SRR957417	endosperm_12DAP	Maize proteome and transcriptome atlas
SRP029238	SRX339762	SRS473585	SRR957421	endosperm_crown	Maize proteome and transcriptome atlas
SRP029238	SRX339763	SRS473586	SRR957422	endosperm_crown	Maize proteome and transcriptome atlas
SRP029238	SRX339764	SRS473587	SRR957423	endosperm_crown	Maize proteome and transcriptome atlas
SRP029238	SRX339815	SRS473638	SRR957474	female_spikelets	Maize proteome and transcriptome atlas
SRP029238	SRX339780	SRS473604	SRR957439	germinating_kernels	Maize proteome and transcriptome atlas
SRP029238	SRX339781	SRS473603	SRR957440	germinating_kernels	Maize proteome and transcriptome atlas
SRP029238	SRX339782	SRS473605	SRR957441	germinating_kernels	Maize proteome and transcriptome atlas
SRP029238	SRX339771	SRS473593	SRR957430	growth_zone	Maize proteome and transcriptome atlas
SRP029238	SRX339772	SRS473595	SRR957431	growth_zone	Maize proteome and transcriptome atlas
SRP029238	SRX339773	SRS473596	SRR957432	growth_zone	Maize proteome and transcriptome atlas
SRP098550	SRX2527292	SRS1948152	SRR5217088,SRR5217087	husk	Genome-wide gene expression profile in inner stem tissue of V2 seedlings and husks (maize B73) [RNA-Seq]
SRP098550	SRX2527293	SRS1948153	SRR5217090,SRR5217089	husk	Genome-wide gene expression profile in inner stem tissue of V2 seedlings and husks (maize B73) [RNA-Seq]
SRP098550	SRX2527294	SRS1948154	SRR5217091,SRR5217092	husk	Genome-wide gene expression profile in inner stem tissue of V2 seedlings and husks (maize B73) [RNA-Seq]
SRP137145	SRX3883792	SRS3123073	SRR6939388	immature_leaves	Zea mays cultivar:B73 and Teosinte Raw sequence reads
SRP137145	SRX3883796	SRS3123077	SRR6939384	immature_leaves	Zea mays cultivar:B73 and Teosinte Raw sequence reads

SRP098550	SRX2527286	SRS1948146	SRR5217080,SRR5217079	inner_stem	Genome-wide gene expression profile in inner stem tissue of V2 seedlings and husks (maize B73) [RNA-Seq]
SRP098550	SRX2527287	SRS1948147	SRR5217081	inner_stem	Genome-wide gene expression profile in inner stem tissue of V2 seedlings and husks (maize B73) [RNA-Seq]
SRP098550	SRX2527288	SRS1948148	SRR5217082	inner_stem	Genome-wide gene expression profile in inner stem tissue of V2 seedlings and husks (maize B73) [RNA-Seq]
SRP029238	SRX339816	SRS473639	SRR957475	internode_6-7	Maize proteome and transcriptome atlas
SRP029238	SRX339817	SRS473641	SRR957476	internode_6-7	Maize proteome and transcriptome atlas
SRP029238	SRX339818	SRS473640	SRR957477	internode_6-7	Maize proteome and transcriptome atlas
SRP029238	SRX339819	SRS473642	SRR957478	internode_7-8	Maize proteome and transcriptome atlas
SRP029238	SRX339820	SRS473643	SRR957479	internode_7-8	Maize proteome and transcriptome atlas
SRP029238	SRX339821	SRS473644	SRR957480	internode_7-8	Maize proteome and transcriptome atlas
SRP102375	SRX2664283	SRS2066044	SRR5368989	jasmonic_acid	Transcriptome profiling of maize responses to Ostrinia furnacalis and jasmonate jasmonic acid
SRP102375	SRX2664284	SRS2066045	SRR5368990	jasmonic_acid	Transcriptome profiling of maize responses to Ostrinia furnacalis and jasmonate jasmonic acid
SRP102375	SRX2664285	SRS2066046	SRR5368991	jasmonic_acid	Transcriptome profiling of maize responses to Ostrinia furnacalis and jasmonate jasmonic acid
SRP116320	SRX3141064	SRS2473827	SRR5985051	leaf_section_2	Developmental gradients of maize leaf
SRP116320	SRX3141035	SRS2473799	SRR5985080	leaf_section_4	Developmental gradients of maize leaf
SRP116320	SRX3141068	SRS2473830	SRR5985047	leaf_section_4	Developmental gradients of maize leaf
SRP116320	SRX3141040	SRS2473804	SRR5985075	leaf_section_6	Developmental gradients of maize leaf
SRP116320	SRX3141075	SRS2473837	SRR5985040	leaf_section_6	Developmental gradients of maize leaf
SRP116320	SRX3141038	SRS2473802	SRR5985077	leaf_section_8	Developmental gradients of maize leaf
SRP116320	SRX3141039	SRS2473803	SRR5985076	leaf_section_8	Developmental gradients of maize leaf
SRP050435	SRX793139	SRS776926	SRR1688292	low_phosphate_leaf	Zea mays Transcriptome or Gene expression
SRP050435	SRX793140	SRS776927	SRR1688293	low_phosphate_leaf	Zea mays Transcriptome or Gene expression
SRP050435	SRX793135	SRS776922	SRR1688288	low_phosphate_root	Zea mays Transcriptome or Gene expression
SRP050435	SRX793136	SRS776924	SRR1688289	low_phosphate_root	Zea mays Transcriptome or Gene expression
SRP029238	SRX339783	SRS473607	SRR957442	mature_leaf	Maize proteome and transcriptome atlas
SRP029238	SRX339784	SRS473606	SRR957443	mature_leaf	Maize proteome and transcriptome atlas
SRP029238	SRX339785	SRS473608	SRR957444	mature_leaf	Maize proteome and transcriptome atlas

SRP137145	SRX3883788	SRS3123069	SRR6939392	mature_leaves	Zea mays cultivar:B73 and Teosinte Raw sequence reads
SRP137145	SRX3883790	SRS3123071	SRR6939390	mature_leaves	Zea mays cultivar:B73 and Teosinte Raw sequence reads
SRP137145	SRX3883791	SRS3123072	SRR6939389	mature_leaves	Zea mays cultivar:B73 and Teosinte Raw sequence reads
SRP029238	SRX339807	SRS473630	SRR957466	mature_pollen	Maize proteome and transcriptome atlas
SRP029238	SRX339808	SRS473631	SRR957467	mature_pollen	Maize proteome and transcriptome atlas
SRP029238	SRX339809	SRS473632	SRR957468	mature_pollen	Maize proteome and transcriptome atlas
SRP114563	SRX3054173	SRS2400432	SRR5888344	non-heat_leaf	Zea mays Transcriptome or Gene expression
SRP114563	SRX3054177	SRS2400428	SRR5888340	non-heat_leaf	Zea mays Transcriptome or Gene expression
SRP114563	SRX3054181	SRS2400424	SRR5888336	non-heat_leaf	Zea mays Transcriptome or Gene expression
SRP097894	SRX3558055	SRS2831283	SRR6468183	ozone_stress	Transcriptome analysis of B vitamin deficiency in plants
SRP097894	SRX3558056	SRS2831284	SRR6468182	ozone_stress	Transcriptome analysis of B vitamin deficiency in plants
SRP097894	SRX3558057	SRS2831285	SRR6468181	ozone_stress	Transcriptome analysis of B vitamin deficiency in plants
ERP024506	ERX2154018	ERS1871576	ERR2096639,ERR2096638	pericarp	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154019	ERS1871577	ERR2096640,ERR2096641	pericarp	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154020	ERS1871578	ERR2096643,ERR2096642	pericarp	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
SRP135909	SRX3804715	SRS3198038	SRR6849468	pollen	Zea mays strain:B73 Transcriptome or Gene expression
SRP135909	SRX3804716	SRS3198037	SRR6849467	pollen	Zea mays strain:B73 Transcriptome or Gene expression
SRP029238	SRX339759	SRS473582	SRR957418	preicarp_and_aleurone	Maize proteome and transcriptome atlas
SRP029238	SRX339760	SRS473583	SRR957419	preicarp_and_aleurone	Maize proteome and transcriptome atlas
SRP029238	SRX339761	SRS473584	SRR957420	preicarp_and_aleurone	Maize proteome and transcriptome atlas
SRP029238	SRX339801	SRS473624	SRR957460	primary_root	Maize proteome and transcriptome atlas
SRP029238	SRX339802	SRS473625	SRR957461	primary_root	Maize proteome and transcriptome atlas
SRP029238	SRX339803	SRS473626	SRR957462	primary_root	Maize proteome and transcriptome atlas
SRP095130	SRX2423172	SRS1860237	SRR5110860	primary_root_air	Growth is required for perception of water availability to pattern plant root branches
SRP095130	SRX2423173	SRS1860238	SRR5110861	primary_root_air	Growth is required for perception of water availability to pattern plant root branches
SRP029238	SRX339794	SRS473617	SRR957453	root_cortex	Maize proteome and transcriptome atlas
SRP029238	SRX339797	SRS473620	SRR957456	root_cortex	Maize proteome and transcriptome atlas

SRP029238	SRX339793	SRS473616	SRR957452	root_elongation	Maize proteome and transcriptome atlas
SRP029238	SRX339796	SRS473619	SRR957455	root_elongation	Maize proteome and transcriptome atlas
SRP029238	SRX339799	SRS473622	SRR957458	root_elongation	Maize proteome and transcriptome atlas
SRP081501	SRX2545792	SRS1964668	SRR5238833	root_hair	Diversification of Root Hair Development Genes in Vascular Plants
SRP081501	SRX2545793	SRS1964669	SRR5238834	root_hair	Diversification of Root Hair Development Genes in Vascular Plants
SRP081501	SRX2545794	SRS1964671	SRR5238835	root_hair	Diversification of Root Hair Development Genes in Vascular Plants
SRP029238	SRX339792	SRS473615	SRR957451	root_maturation	Maize proteome and transcriptome atlas
SRP029238	SRX339795	SRS473618	SRR957454	root_maturation	Maize proteome and transcriptome atlas
SRP029238	SRX339798	SRS473621	SRR957457	root_maturation	Maize proteome and transcriptome atlas
SRP135909	SRX3804722	SRS3198044	SRR6849461	root_tip	Zea mays strain:B73 Transcriptome or Gene expression
SRP135909	SRX3804724	SRS3198046	SRR6849459	root_tip	Zea mays strain:B73 Transcriptome or Gene expression
SRP097894	SRX3558094	SRS2831321	SRR6468144	salt_stress	Transcriptome analysis of B vitamin deficiency in plants
SRP029238	SRX339804	SRS473627	SRR957463	secondary_root	Maize proteome and transcriptome atlas
SRP029238	SRX339805	SRS473628	SRR957464	secondary_root	Maize proteome and transcriptome atlas
SRP029238	SRX339806	SRS473629	SRR957465	secondary_root	Maize proteome and transcriptome atlas
ERP024506	ERX2154030	ERS1871588	ERR2096663,ERR2096662	seedling_leaf	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154032	ERS1871590	ERR2096666,ERR2096667	seedling_leaf	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
SRP118717	SRX3206840	SRS2532637	SRR6061627	seedling_root_tip	Zea mays B73, seedling root tip (0-5mm), RNA-seq, transcription data
SRP118717	SRX3206841	SRS2532637	SRR6061626	seedling_root_tip	Zea mays B73, seedling root tip (0-5mm), RNA-seq, transcription data
SRP118717	SRX3206842	SRS2532637	SRR6061625	seedling_root_tip	Zea mays B73, seedling root tip (0-5mm), RNA-seq, transcription data
ERP024506	ERX2154025	ERS1871583	ERR2096652,ERR2096653	seedling_shoot	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154026	ERS1871584	ERR2096655,ERR2096654	seedling_shoot	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154015	ERS1871573	ERR2096633,ERR2096632	silk	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154016	ERS1871574	ERR2096634,ERR2096635	silk	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154017	ERS1871575	ERR2096637,ERR2096636	silk	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
SRP029238	SRX339810	SRS473633	SRR957469	silks	Maize proteome and transcriptome atlas

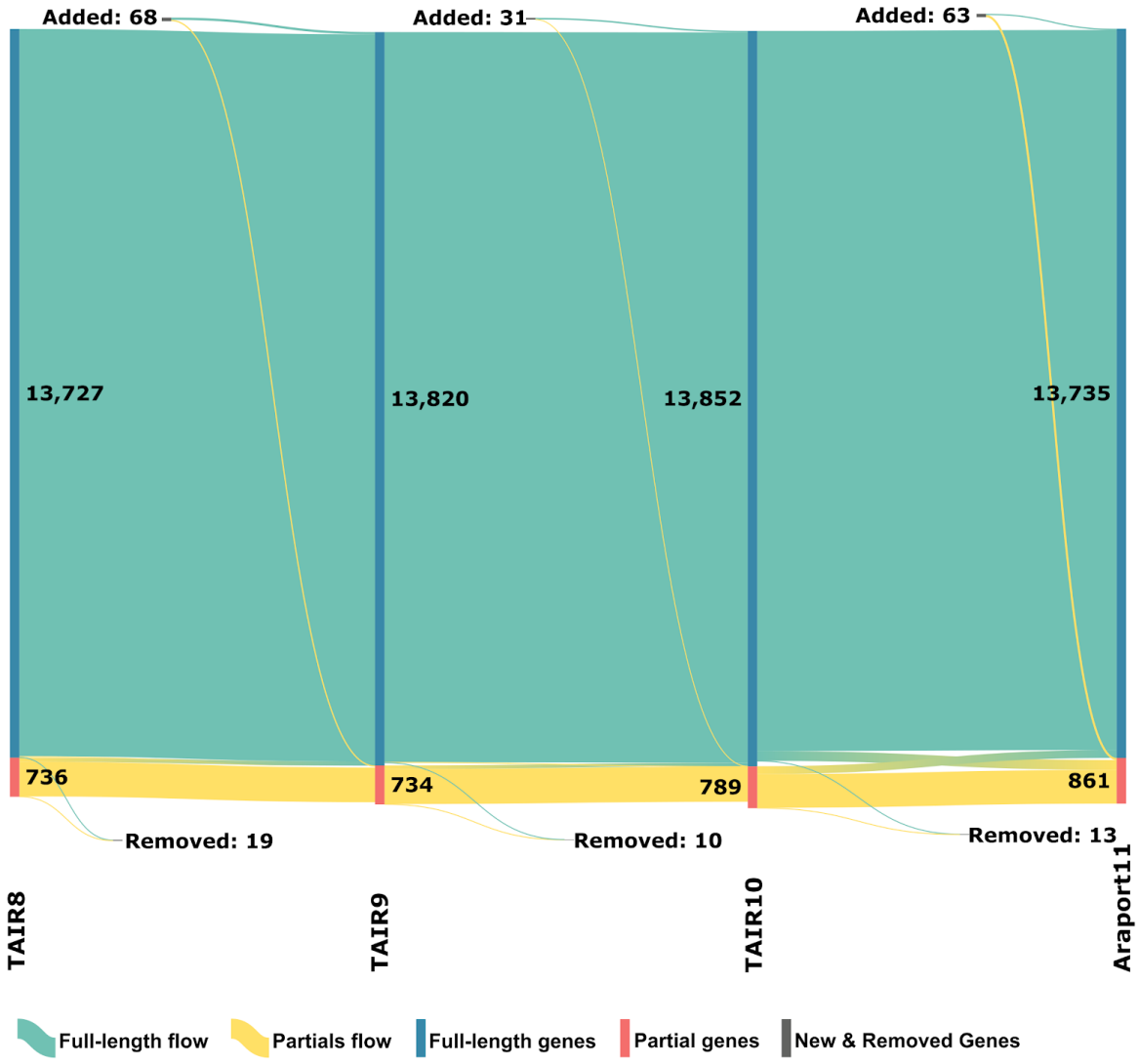
SRP029238	SRX339811	SRS473634	SRR957470	silks	Maize proteome and transcriptome atlas
SRP029238	SRX339768	SRS473592	SRR957427	stomatal_division_zone	Maize proteome and transcriptome atlas
SRP029238	SRX339769	SRS473591	SRR957428	stomatal_division_zone	Maize proteome and transcriptome atlas
SRP029238	SRX339770	SRS473594	SRR957429	stomatal_division_zone	Maize proteome and transcriptome atlas
SRP029238	SRX339765	SRS473588	SRR957424	symmetrical_division_zone	Maize proteome and transcriptome atlas
SRP029238	SRX339766	SRS473589	SRR957425	symmetrical_division_zone	Maize proteome and transcriptome atlas
SRP029238	SRX339767	SRS473590	SRR957426	symmetrical_division_zone	Maize proteome and transcriptome atlas
SRP115608	SRX3100534	SRS2436573	SRR5941976	tassel_control	tassel Transcriptome under water deficit
SRP115608	SRX3100535	SRS2436574	SRR5941975	tassel_drought	tassel Transcriptome under water deficit
SRP029238	SRX339822	SRS473645	SRR957481	veg-meristem	Maize proteome and transcriptome atlas
SRP029238	SRX339823	SRS473646	SRR957482	veg-meristem	Maize proteome and transcriptome atlas
ERP024506	ERX2154027	ERS1871585	ERR2096657,ERR2096656	whole_seedling	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154028	ERS1871586	ERR2096659,ERR2096658	whole_seedling	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing
ERP024506	ERX2154029	ERS1871587	ERR2096660,ERR2096661	whole_seedling	A Phylogenetically Based Comparative Transcriptional Landscape of Maize and Sorghum Obtained by Single-molecule Sequencing

Supplementary Figures

Supplementary Figure 1: Historical overview of genome size for reference organisms

Historical overview of genome sizes of different model organisms over time, in megabase (MB).

Supplementary Figure 2: Historical overview partial and full-length genes in *Arabidopsis thaliana*



Supplementary References

1. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO applications from quality assessments to gene prediction and phylogenomics**. *Mol Biol Evol* 2017.
2. Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K: **Dissecting plant genomes with the PLAZA comparative genomics platform**. *Plant Physiol* 2012, **158**:590-600.
3. Veeckman E, Ruttink T, Vandepoele K: **Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences**. *Plant Cell* 2016, **28**:1759-1768.
4. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods* 2015, **12**:59-60.
5. UniProt Consortium T: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2018, **46**:2699.
6. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K: **PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics**. *Nucleic Acids Res* 2018, **46**:D1190-D1196.
7. Vaneechoutte D, Vandepoele K: **Curse: Building expression atlases and co-expression networks from public RNA-Seq data**. *under review* 2018.
8. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification**. *Nat Biotechnol* 2016, **34**:525-527.
9. Thatcher SR, Zhou W, Leonard A, Wang BB, Beatty M, Zastrow-Hayes G, Zhao X, Baumgarten A, Li B: **Genome-wide analysis of alternative splicing in Zea mays: landscape and genetic regulation**. *Plant Cell* 2014, **26**:3472-3487.