The Plant Cell, Vol. 28: 1759–1768, August 2016, www.plantcell.org

COMMENTARY

Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences[OPEN]

Elisabeth Veeckman,[a,b] Tom Ruttink,[a,b] and Klaas Vandepoele[b,c,d,1]

[a]Institute for Agricultural and Fisheries Research, Plant Sciences Unit, Growth and Development, B-9090 Melle, Belgium
[c]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium
[d]Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium
[b]Bioinformatics Institute Ghent, Ghent University, B-9052 Ghent, Belgium

ORCID IDs: 0000-0003-0510-6317 (E.V.); 0000-0002-1012-9399 (T.R.); 0000-0003-4790-2725 (K.V.)

Genome sequencing is becoming cheaper and faster thanks to the introduction of next-generation sequencing techniques. Dozens of new plant genome sequences have been released in recent years, ranging from small to gigantic repeat-rich or polyploid genomes. Most genome projects have a dual purpose: delivering a contiguous, complete genome assembly and creating a full catalog of correctly predicted genes. Frequently, the completeness of a species' gene catalog is measured using a set of marker genes that are expected to be present. This expectation can be defined along an evolutionary gradient, ranging from highly conserved genes to species-specific genes. Large-scale population resequencing studies have revealed that gene space is fairly variable even between closely related individuals, which limits the definition of the expected gene space, and, consequently, the accuracy of estimates used to assess genome and gene space completeness. We argue that, based on the desired applications of a genome sequencing project, different completeness scores for the genome assembly and/or gene space should be determined. Using examples from several dicot and monocot genomes, we outline some pitfalls and recommendations regarding methods to estimate completeness during different steps of genome assembly and annotation.

INTRODUCTION

The ever-decreasing cost and the expanding capacity of genome sequencing using next-generation sequencing techniques has led to a remarkable increase in the number of available genome sequences. As of 2016, over 100 plant genomes have been sequenced, ranging from small (e.g., *Utricularia gibba*, 80 Mb) to huge, repeat-rich, or polyploid genomes (e.g., *Triticum aestivum*, 17 Gb), with many more expected in the years to come (Weigel and Mott, 2009; Chia et al., 2012; Michael and Jackson, 2013; Li et al., 2014). Ideally, a genome assembly represents a complete and contiguous genome sequence with a cumulative scaffold length equal to the haploid genome size (Figure 1A). In addition, a complete set of annotated genes offers a starting point for a detailed characterization of gene functions, biochemical and regulatory pathways, or quantitative trait loci. Genes are the nodes in a biological network, which offers valuable insights into protein complexes, regulatory interactions, and metabolic processes that determine the physiological and biochemical properties of a cell, an organ or an organism (Bassel et al., 2012).

Clearly, comparative genomics and evolutionary studies require complete genomes and gene sets. Well-assembled genome sequences are necessary to characterize different classes of repetitive elements to identify large-scale gene colinearity across related species and to reconstruct the organization and evolution of transposable elements (Bennetzen and Wang, 2014). Moreover, a complete gene catalog is required to test if the gain or loss of biochemical or signaling pathways in specific plant species can explain the structural and physiological adaptations required to survive in extreme environments. N50 is a commonly used contiguity measure denoting that 50% of the total assembly length is contained in scaffolds of length N50 or longer. Over the last 15 years, genome assemblies have displayed a large range of N50 values but indicate low contiguity even for relatively small genomes (Supplemental Figure 1), suggesting that fragmented draft genomes are generated for many plants. As this wealth of new plant genome sequences and gene catalogs expands and assembly strategies evolve, so too must the variety of methods used to measure their quality and completeness (Earl et al., 2011; Salzberg et al., 2012). As yet, no uniform metrics or standards are in place to estimate the completeness of a genome assembly or the annotated gene space, despite their importance for downstream analyses.

DEFINING THE EXPECTED GENOME SIZE AND GENE SPACE

A simple approach is common to all reported measures of completeness (Figures 1B and 1C). First, one measures the size of the assembled genome (i.e., total assembly length) or the gene space (i.e., the number of genes), in the following referred to as the "observed." Second, one selects a reference to define the expected genome size or gene space, here referred to as the "expected." To define the expected genome size, both physical measurements (e.g., nuclear weight) and computational methods

[1]Address correspondence to klaas.vandepoele@psb.vib-ugent.be.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Klaas Vandepoele (klaas.vandepoele@psb.vib-ugent.be).
[OPEN]Articles can be viewed without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.16.00349

COMMENTARY

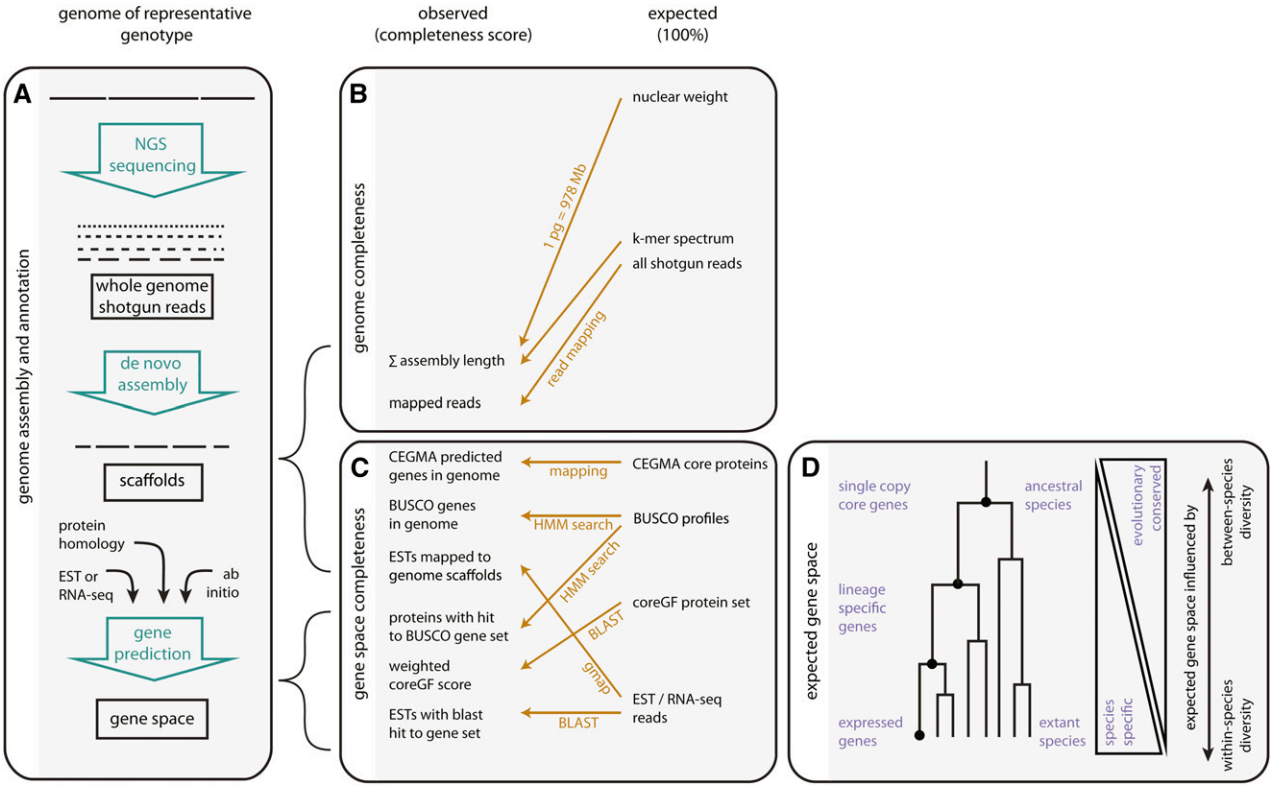


Figure 1. Framework for Genome Assembly and Gene Space Completeness Estimation.

(A) Workflow for genome assembly and annotation. A representative genotype is selected for sequencing and the whole-genome shotgun reads are assembled into incrementally longer contiguous scaffolds. In a final step, gene prediction provides the description of repetitive regions and the annotation of genes.

(B) and **(C)** Estimation of genome assembly and gene space completeness, respectively. Measures for the expected and observed size of the genome assembly and gene space are shown, connected by specific methods.

(D) The expected gene space can be estimated along an evolutionary scale, ranging from evolutionarily highly conserved to species-specific genes.

that analyze the sequence space (such as k-mer spectra) should be used. Furthermore, to define the expected gene space, one can rely on evolutionary conservation and use the gene space of related species as reference (interspecies comparisons). Alternatively, one can define a species-specific measure of the gene space by transcriptome or EST sequencing in the species itself (intraspecies comparisons; Figure 1D). Clearly, these methods rely on starkly contrasting assumptions, as further detailed below. All comparison methods (e.g., BLAST or read mapping) inherently assume directionality and set the external reference as the expected 100%. In all approaches, the observed measure is then expressed as a fraction of the expected and interpreted as completeness score for the genome assembly or gene space. Given the diversity of approaches, it is important to understand the underlying concepts to provide consistent and realistic measures of genome and gene space completeness.

The genome can be partitioned into two main fractions with contrasting characteristics in terms of assembly and annotation. The repetitive DNA fraction, mostly contained in heterochromatin, is difficult to assemble using short shotgun reads and is commonly collapsed or absent from draft genome assemblies. This partition generally contains transposable elements and relatively few coding genes. By contrast, the nonrepetitive sequence space, mostly contained in euchromatin, is relatively easy to assemble and is commonly assumed to represent the gene-rich partition. It is important to realize that methods to estimate genome or gene space completeness target these partitions of the genome differently, and although completeness scores may seem related, they should not be extrapolated between the two levels.

Here, we will outline the challenges of estimating the completeness of the genome assembly and annotated gene space. We first explain how the expected is defined for different measures of completeness and comment on the assumptions made by each method, including their strengths and weaknesses. Next, we will compare different measures of completeness in 12 recently published plant genomes and highlight several cases where dissimilar

completeness scores are the consequence of technical issues of assembly or annotation or due to strong gene function or expression biases in the expected gene space. Finally, we will provide some guidelines to determine more robust completeness scores and comment on the challenges facing future plant genome projects.

ESTIMATING THE COMPLETENESS OF A GENOME ASSEMBLY

The first step in a genome assembly workflow (Figure 1A) is selecting an individual that is representative of the species. For this individual, shotgun libraries are constructed with variable insert sizes, ranging from 100 bp to over 100 kb. Sequencing will yield reads of variable length, ranging from 100 bp to more than 10 kb, depending on the applied sequencing technology. These reads are then assembled into incrementally longer contiguous sequences in three steps. First, contigs are constructed through de novo assembly based on the overlap of short reads or de Bruijn k-mer graphs. In de Bruijn graphs, nodes represent k-mers and edges connect neighboring k-mers so that a traversal through this graph results in the reconstruction of a contiguous sequence. Second, the contigs are ordered into scaffolds using mate pairs, BAC-end sequences, or hybrid assembly with long sequencing reads. Finally, the scaffolds are ordered and anchored into pseudomolecules or linkage groups representing chromosomes. These chromosomal structures can further be validated and improved using optical mapping, cytogenetic mapping, Hi-C sequencing, genetic maps, population sequencing, or physical maps such as BAC minimal tiling paths (Mascher et al., 2013; Mendelowitz and Pop, 2014; Flot et al., 2015).

Two main factors affect the completeness and contiguity of the genome assembly: the level of heterozygosity and the length, abundance, and dispersal of duplicated regions or repetitive sequences (Wendel et al., 2016). Genome assembly algorithms attempt to reconstruct unique sequences in order to separate recently duplicated regions, closely related gene family members or highly conserved protein domains. As a result, allelic sequences in highly heterozygous species are often also reconstructed as independent sequences, thereby inflating the total assembly length and decreasing scaffold contiguity (e.g., *Malus domestica*; Velasco et al., 2010). Conversely, repeat regions are typically collapsed during assembly of short reads, thereby severely reducing the total assembled genome size and interrupting scaffold contiguity (e.g., *Lolium perenne*; Byrne et al., 2015). Highly polymorphic regions disturb sequence alignment during de novo assembly, lead to bubbles and branches in de Bruijn graphs, and cause breakpoints when de Bruijn graphs are resolved into contiguous sequences. Some of these issues may be overcome in the near future using third-generation long-read sequencing technologies.

The expected genome size of an organism can be measured using the physical properties of the nuclear genome: by reassociation kinetics of high molecular weight genomic DNA (C_ot assay), pulsed field gel electrophoresis, or, ideally, flow cytometry after DNA staining. These methods use standards of known molecular weight or reference species with a defined nuclear DNA mass (Zonneveld et al., 2005). The total assembled scaffold length (in Mb) can then be expressed as a fraction of the molecular weight of the nuclear DNA (in pg) using the standard average molecular weight of 1 pg per 978 Mb for the conversion. Strikingly, closely related species may display considerable variation in genome size, hence limiting the accuracy of interspecies comparative measures of completeness (Garcia et al., 2014). By contrast, flow cytometry-based measurements of genome size turn out to be fairly consistent across individuals within a species (Dolezel and Bartos, 2005), thus providing accurate estimates of the expected genome size within that species.

Alternatively, the expected genome size and repetitive sequence content can be estimated using computational methods, such as k-mer frequency spectra of the shotgun sequencing reads. A k-mer frequency spectrum shows the count distribution of all sequences of length k present in the read data. From the frequency plot, one can estimate the coverage depth, sequencing bias, data quality, problems in the assembly, and polymorphic rates (Liu et al., 2013). The genome size can be calculated by dividing the total number of k-mers in the read data by the peak value in the frequency plot (Supplemental Methods and Supplemental Figure 2). Furthermore, the percentage of the shotgun reads or BAC-end sequences that map onto the scaffolds yields a genome completeness score that indicates whether the shotgun read sequences have all been incorporated into the scaffolds. The read depth profile may further identify wrongly assembled, collapsed, or duplicated regions (Hunt et al., 2013; Rahman and Pachter, 2013). Conversely, one can control overassembly by analyzing whether all scaffolds are supported by read data. Just as the assembly algorithms are sensitive to genetic diversity and heterozygosity while searching for sequence overlap to build contiguous scaffolds, these assembly completeness methods rely on sequence identity for read mapping. Thus, completeness scores are inherently sensitive to mismatch stringency parameters in highly heterozygous genomes (Wendel et al., 2016).

ESTIMATING THE COMPLETENESS OF THE ANNOTATED GENE SPACE

In an ideal scenario, genome annotation describes repetitive regions and the complete set of protein-coding genes and various classes of noncoding RNAs with a correctly identified gene structure. Ab initio methods try to predict gene models by the detection of intrinsic signals such as codon composition or splice sites in the DNA sequence while extrinsic approaches make use of similarity to well-characterized proteins from related species or EST/RNA-seq transcript data of the species under investigation (Figure 1A). Unfortunately, most gene prediction methods suffer from false-positive and false-negative predictions as well as partially incorrect gene structures. Retraining gene prediction software to detect codon biases or specific splicing motifs is important both for obtaining high-quality gene models and for identifying species-specific genes lacking homologs in other plant families. Gene prediction benchmarks exist for different eukaryotic model

COMMENTARY

species and automated self-learning gene prediction approaches have been developed (Korf, 2004). However, in the absence of large species-specific transcript databases, generic gene prediction tools have been used for several plant genomes, compromising validation of the quality and completeness of the predicted gene catalog. Recently developed methods like MAKER-P and BRAKER1 offer a practical solution for some of these issues, provided that sufficient extrinsic information is available (Campbell et al., 2014; Hoff et al., 2016).

If the N50 is smaller than the average size of a gene, one can expect to annotate many partial gene models due to gene splitting, resulting in an overestimation of the number of genes in the genome. Clearly, such erroneous gene models will compromise the correct delineation of homologous gene families and orthologous genes, as well as the detection of protein domains. This obstructs the interpretation of gene family expansion or gene loss and any other downstream gene-based analysis, such as gene expression quantification through RNA-seq, annotation of ChIP-seq binding events, or gene network analysis.

DEFINING THE EXPECTED GENE SPACE ON A GLIDING EVOLUTIONARY SCALE

The expected gene space can be defined between two extremes on the evolutionary scale (Figure 1D). On one extreme, evolutionarily highly conserved reference gene sets are assumed to be present in the newly assembled genome. Thus, interspecies comparisons require the definition of the taxonomic range over which genes are expected to be conserved, relative to the species under investigation. The core eukaryotic genes mapping approach (CEGMA) has defined highly conserved eukaryotic genes, placing itself on a basal eukaryotic level in the tree of life (Parra et al., 2007, 2009). BUSCO, the successor of CEGMA, has defined sets of single-copy genes for various major clades, including plants (Simão et al., 2015). Finally, the PLAZA core gene families (coreGFs) are defined as highly conserved gene families in a majority of plants within predefined lineages (Van Bel et al., 2012). On the other extreme of the evolutionary scale, one can define all expressed genes in the sequenced individual as a reference for the expected gene space. In this case, transcript sequencing, optionally followed by de novo transcriptome assembly, and mapping provides empirical evidence to define the expected gene space in the organism under investigation. Below, we will further illustrate the underlying assumptions, strengths, and weaknesses for all four methods.

The CEGMA reference gene set comprises 458 genes that are highly conserved in six eukaryotic species (*Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*) and are assumed to be encoded in essentially all eukaryotic genomes. Notably, CEGMA was originally created to build a robust set of gene annotations to train gene prediction software in the absence of experimental transcriptome data, but it is not meant to provide a complete catalog of genes in a genome. Nevertheless, a subset of 248 single-copy core eukaryotic genes is frequently used to estimate genome completeness, where the CEGMA completeness score expresses the fraction of the 248 genes that can be accurately mapped onto the genome assembly (Figure 1C).

The CEGMA gene set dates back to the last common eukaryotic ancestor; thus, any extrapolation of the completeness score based on such a limited set of highly conserved proteins will fail to account for many genes unique to plant biology. In addition, as most plant genomes encode more than 20,000 genes, any bias present in such a small set of conserved core genes can lead to errors in the estimated completeness scores. We found that more than half of the 248 CEGMA genes from Arabidopsis are expressed across all the different conditions and organs contained in a nonredundant Arabidopsis expression atlas (Figure 2A). This reveals that many genes expressed in specific plant organs or developmental stages are missing. Gene Ontology enrichment further demonstrates the gene function bias in the 248 core eukaryotic genes: housekeeping functions (DNA metabolism, translation, cell cycle, and generation of precursor metabolites and energy) are overrepresented and the CEGMA set does not cover genes functioning in biological processes conserved in green plants, such as photosynthesis or development (Figure 2B).

BUSCO recently defined gene sets for six major phylogenetic clades to estimate completeness as well as the duplicated fraction of a genome sequence. Each gene is expected to be found single-copy in any newly sequenced genome when an appropriate clade is selected. The single-copy nature is used by BUSCO to estimate the level of redundancy in the genome assembly, but the frequency of small- and large-scale duplications, such as (paleo)polyploidy, in plants makes this feature less applicable. Whereas CEGMA works only on raw genome or transcript sequences and performs gene prediction prior to the completeness estimation, BUSCO can be applied to a genome sequence as well as to an annotated gene set (Supplemental Methods). The BUSCO plant profiles consist of 952 single-copy orthologs and analysis of the Arabidopsis best hits showed no expression bias and less gene function bias toward housekeeping genes compared with the CEGMA core genes (Figure 2).

The PLAZA coreGFs are a set of core gene families that are highly conserved in a majority of plant species within predefined evolutionary lineages. Three sets of coreGFs have been defined using the PLAZA 2.5 database: green plants (2928 coreGFs), rosids (6092 coreGFs), and monocots (7076 coreGFs), using a parsimony-based selection approach where complete conservation across all species is not required. This approach accounts for the observation that genes are indeed occasionally lost in some species and it tolerates potential annotation errors in a limited number of species. In contrast to CEGMA and BUSCO, coreGFs are not filtered for single-copy genes and can therefore better accommodate the frequent occurrence of whole-genome duplications in plants (Van de Peer et al., 2009). Consequently, the number of coreGF genes is 5 to 10 times higher compared with CEGMA or BUSCO gene sets. Similar to BUSCO and transcript mapping, coreGFs can be used to assess the completeness of an annotated gene set (for further details on

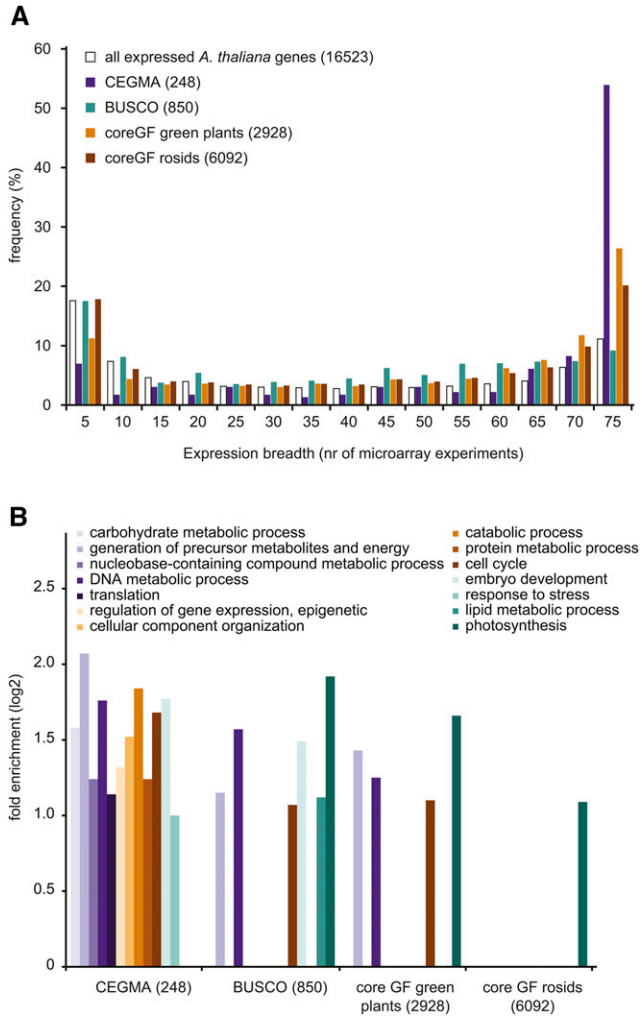


Figure 2. Expression and Gene Function Biases Associated with CEGMA, BUSCO, and coreGFs in Arabidopsis.

Expression and gene function biases were determined for the CEGMA set (248 single-copy core genes), the BUSCO profile best hits on the gene set of Arabidopsis (850), and the coreGFs for green plants (2928 gene families) and rosids (6092 gene families).
(A) Expression biases were determined by counting the number of microarray experiments in which a gene was expressed and compared with the expression breadth of the complete gene set of Arabidopsis.
(B) Gene function biases were estimated using Gene Ontology enrichment analysis of the PLAZA 3.0 Workbench. Gene Ontology slim Biological Process terms with at least 2-fold enrichment are shown ($P < 0.01$).

the calculation of the coreGF completeness scores, see Supplemental Methods). Expression breadth and gene function enrichment analysis reveals that the coreGF gene set is less biased toward ubiquitously expressed genes and does not strongly overrepresent specific gene functions (Figure 2). Furthermore, because coreGFs sample conserved gene families at different taxonomic levels within green plants, it offers a better representation of the gene function space of flowering plants compared with CEGMA and BUSCO.

By contrast, transcript mapping is a highly species-specific completeness assessment method that is independent of evolutionary conservation between species. This method uses large-scale EST or RNA-seq transcript sequencing to estimate how many of the transcribed genes are present in the gene space partition of the genome assembly of a given species (here referred to as "transcript mapping"). In this case, the expected gene space is defined as the total number of transcript sequences, either specifically generated to guide genome annotation of the sequenced genotype, or derived from public resources.

In an attempt to maximize the reference sequence data set, it is often tempting to use

COMMENTARY

transcripts from alternative genotypes within the species, or even from closely related species. However, when switching from organism-specific to intraspecies transcript mapping or to interspecies comparisons, one has to realize that the assumptions underlying the evolutionary conservation of gene sets may no longer hold and that increased sequence divergence gradually comes into play as well. Thus, all approaches are dependent on mapping stringency parameters, which should be adjusted to account for evolutionary divergence between species or genetic diversity within species. Although transcript mapping is highly dependent on the number of transcript sequences and the complexity captured by the different cDNA libraries, it is better able to capture fast-evolving and species-specific genes compared with evolutionary-based methods. A comparison of transcript mapping at the levels of the genome assembly and the annotated gene catalog indicates the completeness of the gene prediction. Depending on the library preparation method used, genes encoding different types of RNA (e.g., rRNAs, tRNAs, small nuclear RNA, long noncoding RNAs) can also be included.

INFLUENCE OF TRANSCRIPT MAPPING PARAMETERS ON GENE SPACE COMPLETENESS

In practice, de novo assembly often first leads to reconstruction of a partition of the genome that contains the euchromatic, gene rich, unique sequences in the genome. To evaluate the influence of transcript mapping data sets and parameters on imperfect genome assemblies when assessing gene space completeness, we simulated fragmented and incomplete genomes of Arabidopsis and rice (*Oryza sativa*). In short, we fragmented the genome into 10-kb sequences and randomly subsampled genomic fragments to simulate decreasing levels of completeness (50 to 100%). Random subsampling of a given fraction of the entire genome creates a reference that contains, proportionally, a "known" fraction of the gene space, independent of whether the repetitive DNA partition is included in the reference or not. We collected 1.5 and 1 M publicly available EST sequences for Arabidopsis and rice, respectively, and mapped them onto the partial reference assemblies. We then calculated mean and SD of the transcript mapping score across 100 replicate random subsamples (bins) with varying numbers of ESTs (range 100 to 300,000 ESTs) (Supplemental Methods). Finally, we compared the measured gene space completeness scores to the known fraction of the gene space to estimate the influence of EST mapping parameters (such as minimum percentage of coverage), and EST library size and complexity, because these typically vary across the reported completeness estimates. On average, the transcript mapping score is stable (SD < 1%) in bin sizes of at least 3000 ESTs, for both Arabidopsis and rice (Figure 3A). Transcript mapping estimates the completeness of the gene space at 61%, when only 50% of the Arabidopsis genome is used as reference, while for more complete genomes, the transcript mapping score converges to 97% (Figure 3A, upper panel). When partial EST mappings were filtered out (90% coverage filter), partial genomes are no longer overestimated, but more complete genomes seem incomplete (Figure 3A, lower panel). The latter might be related to the challenge of correctly aligning spliced transcript sequences to their corresponding genomic locus, comprising both exons and introns. These results show that it is important to consistently use and report the mapping parameters per comparison method. As stated above, it is important to note that transcript mapping scores should not be extrapolated to the completeness of the total genome assembly, but only apply to the gene space partition, even if the entire genome reference sequence is used for the EST mapping.

We also evaluated transcript mapping scores per library on various simulated genome incompleteness levels for Arabidopsis and rice to further define the relationship between transcript completeness score, actual completeness, and EST library size and complexity (Figure 3B). Both species display more variation in EST mapping score when smaller libraries are used to define the expected gene space, confirming the results from downsampling ESTs. If the libraries contain more than 10,000 ESTs, the EST mapping scores for Arabidopsis libraries converge to the same value as for subsampling bins of >10,000 ESTs. For rice, the convergence of EST mapping scores is not as clear. This indicates that the minimum library size needed for a reliable estimate depends on the species, perhaps as function of size and/or complexity of the genome. Several transcript libraries can be generated for a fraction of the cost of the entire genome sequencing project, which suffices to validate the gene space completeness test. Although RNA-seq is a valuable alternative to define the expected gene space, de novo assembly can lead to overestimation of the expected number of genes due to the construction of allelic transcripts or splice variants and fragmented transcripts and to underestimation due to the failure to reconstruct low-abundant transcripts (Honaas et al., 2016).

COMPARISON OF FOUR GENE SPACE COMPLETENESS METHODS

The completeness estimates of three methods based on evolutionary conserved gene sets (CEGMA, BUSCO, and coreGFs) and transcript mapping were compared (Figure 4) using 10 recently published plant genome data sets, including rosids and monocots (Supplemental Table 1). The two high-quality reference genomes of Arabidopsis and rice contain almost all of the CEGMA and coreGF core genes (completeness scores >99%; only 50 and 42 missing coreGFs for Arabidopsis and rice, respectively; Figure 4). In eight species, the CEGMA and BUSCO scores are higher than the coreGF score. Reporting only CEGMA or BUSCO scores generally leads to an overestimation of the gene space completeness. In some cases, the differences between the measured completeness scores are quite large. CEGMA scores are at least 5% higher than the coreGF score for more than half of the species, while for three species, this difference is even larger than 10%. These missing fractions in the expected gene space correspond to the projected absence of a few hundred to more than a thousand coreGF genes. The underlying reasons vary and can be illustrated in three specific cases. First, in *L. perenne*, the reported

COMMENTARY

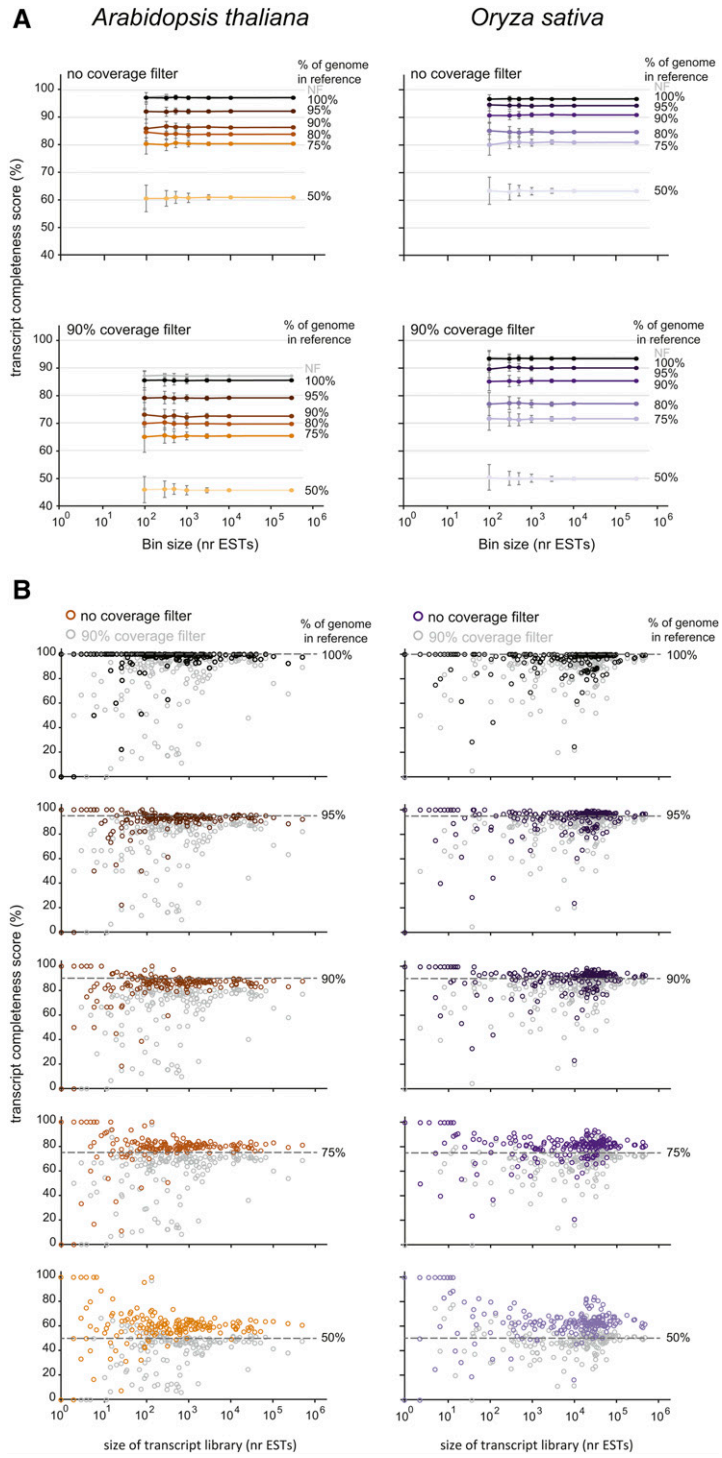


Figure 3. Evaluation of Transcript Completeness Scores.

To estimate the relationship between transcript completeness score, actual reference genome completeness, and EST library size and complexity, two approaches were compared using Arabidopsis or rice. For each species, the genome was cut into stretches of 10 kb, and fragments were randomly sampled to create partial genome references containing 50, 75, 80, 90, 95, and 100% of the original genome sequence. All publicly available EST sequences were mapped onto the respective partial genomes. In a first approach, all ESTs were pooled and random sampling for different EST bin sizes (range from 100 to 300,000) was performed 100 times. The mean and SD of the transcript completeness scores for each bin size and each partial genome is given in **(A)**. The lower graphs show mean transcript completeness scores and SD counting only mapped ESTs with a length coverage higher than 90%. **(B)** shows the transcript completeness score for each individual EST library (indicated by a circle) mapped onto the partial genomes. Completeness scores per library based on EST mappings with a length coverage higher than 90% are shown in gray in each graph.

CEGMA score of 96% indicates that the genome assembly is complete, yet 1709 coreGFs are missing from the predicted gene set. The score difference of BUSCO applied on the genome and gene set (97 and 90%, respectively) indicates a discrepancy in the gene space present in the genome assembly and annotated gene set. For this genome, the researchers generated a conservative, yet reliable, set of annotated genes by selecting only evidence-based gene models, i.e., supported by *Brachypodium distachyon* protein alignment and transcriptome assemblies (Byrne et al., 2015). The transcript mapping score of 96% on the genome assembly compared with the coreGF score of 76% on the predicted gene set corroborates that the gene space partition of the genome has been well assembled, but that gene prediction is incomplete. Indeed, mapping of Brachypodium proteins on the *L. perenne* genome assembly confirms that at least 924 of the 1709 missing coreGFs can be found using TBLASTN (E-value < 1e-10).

Second, we observed that *Phalaenopsis equestris* has a coreGF score of only 82%. It is important to note that the coreGFs are predefined at three evolutionary levels, rosids, monocots, and green plants. Monocot coreGFs were defined only using gene sets from the Poales, which are part of the commelinids. As *P. equestris* belongs to the Asparagales, a sister group to the commelinids,

COMMENTARY

the lower coreGF score could reflect potential gene loss in *P. equestris* and shows the importance of choosing an appropriate phylogenetic level at which an evolutionary conserved gene set is defined. A similar limitation exists for the BUSCO method applied to the genome assembly, as this approach uses an extrinsic gene prediction tool that was trained for maize (*Zea mays*), a member of the Poales. Therefore, low BUSCO scores on the genome could also be due to genes missed by the gene prediction step applied by BUSCO.

Third, although for most species the transcript mapping score lies within the same range as the CEGMA, BUSCO, and coreGF score, there are some exceptions. One example is *Cicer arietinum*, for which only 89% of the ESTs could be mapped on the genome sequence. More than half of the unmapped sequences are of non-plant origin, mostly from *Fusarium oxysporum*, illustrating how contaminations inflate the expected gene space and lead to an underestimation of the gene space completeness.

CONCLUSIONS AND GUIDELINES

Population resequencing studies in Arabidopsis, rice, potato (*Solanum tuberosum*), and maize have unveiled extensive genomic variation between individuals, including structural rearrangements, copy number variations, insertion-deletions, single nucleotide polymorphisms, and sequence repeats. This has led to the definition of "core" genome sequences (shared between all members of a species), "dispensable" genome sequences (present in only one or a few members), and "pan" genome sequences (the union, or full genome complement across all members). Hence, the variability of sequence conservation extends to the subspecies or individual organism level (Cao et al., 2011; Hirsch et al., 2014; Marroni et al., 2014). The dispensable genome contains genes with high biological relevance, illustrated by possible roles in adaptation to abiotic and biotic stresses (Hardigan et al., 2016), species diversification and development of novel gene functions (Wang et al., 2006), and agronomic and metabolic traits (Yao et al., 2015). This clearly limits the definition of the expected gene space and, consequently, the precision and accuracy of completeness estimates of both the genome and the gene space.

A complete genome assembly is essential for the study of chromosome structure and repeat content. Although a complete gene catalog is an important deliverable of a genome sequencing project, the genome assembly should not be restricted to the gene space partition, and alternative strategies of library preparation and assembly algorithms are needed to reconstruct the heterochromatic, repeat-rich sequence partition. Here, we discussed different measures to assess genome and gene space completeness and illustrated that large differences in completeness scores for the same genome can be found. Therefore, we advise assessing genome completeness both at the genome assembly and gene space level to reliably estimate the quality of all steps of assembly and annotation.

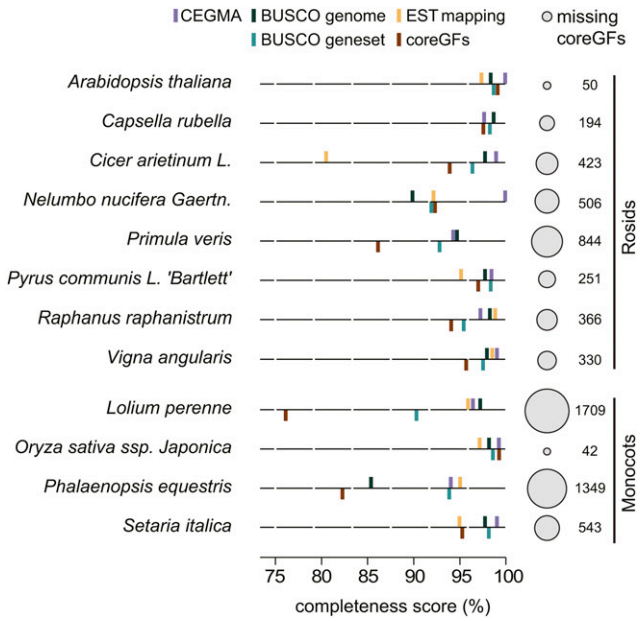


Figure 4. Comparison of CEGMA, BUSCO, CoreGF, and EST Mapping Completeness Scores for 12 Plant Genomes.

Twelve genomes within rosids and monocots were analyzed. Left: CEGMA, BUSCO, coreGF, and EST completeness scores per genome. The reported CEGMA score was obtained from the respective genome publications. We calculated the weighted coreGF score of the respective annotated gene sets using the rosid or monocot coreGFs according to lineage. The EST mapping completeness score is the percentage of publicly available EST sequences that could be mapped onto the genome. Right: Size of the circles and numbers indicate the number of missing coreGFs per genome.

COMMENTARY

Based on our observations, we suggest the following guidelines. (1) For genome assembly completeness, we suggest reporting the estimated genome size based on k-mer statistics of the raw sequence reads, together with the fraction of reads that map onto the assembled genome. In addition, a nuclear weight estimate should also be reported, obtained from an experimental method such as PFGE or flow cytometry using standardized references. Comparison of these measures highlights the fraction of the repeat DNA partition that was not assembled. (2) One should provide and compare gene space completeness score of methods based on evolutionary conservation and transcript mapping in order to limit the effect of erroneous assumptions underlying the expected gene space.

For interspecies comparisons, the core gene set used to model the expected number of genes ought to be defined at various levels of evolutionary conservation, but including a set as large as possible and without strong gene function or expression biases. Therefore, tools to define customized core gene sets should be developed so users can define the expected gene space at various phylogenetic levels, independent of the currently available predefined core sets.

For transcript mapping, preferably different cDNA libraries covering a range of organs and conditions should be included to secure a robust estimate of the expected number of genes. In the case of very low transcript mapping scores, one should check for contamination of the transcript data sets.

(3) The correct structural annotation of species-specific genes and fast-evolving genes poses big challenges for a full characterization of the gene space. Ideally, gene space completeness estimates should be applied to both the genome assembly and the annotated gene set, as large score differences can highlight loci in the genome assembly that were missed by the gene prediction. Identification of the missing coreGFs can be used for the targeted investigation of specific gene functions. Detection of genes that are missing only from the predicted gene space indicate that an optimization of the gene prediction algorithms is needed, since these tools frequently suffer from the lack of proper training in a newly sequenced organism. The absence of specific genes in the genome, and not just the assembly, should be independently confirmed using, for example, de novo assembled transcripts (Olsen et al., 2016) or hybridization-based molecular techniques.

(4) To perform cross-species gene and genome comparisons, one should work only with genome assemblies that have good contiguity. Highly fragmented genomes with low N50 values (for example, genomes where most contigs only contain one or a few genes) not only limit the detection of synteny or gene colinearity within and between species, but also suffer from split and partial gene models. Comparative genome studies aiming to identify genomic adaptations required for growth in a specific environmental niche (e.g., loss or gain of genes or pathways) should not rely on gene space validations using evolutionary conserved reference sets because these are blind to lineage-specific genes. Transcript mapping is a better means to verify species-specific biology.

We believe these pointers will help the next generation of plant scientists to assess the quality of new genome sequences in a transparent and balanced manner and to formulate a standard for delivering better plant genome sequences, which are the templates for new biological discoveries.

Supplemental Data

Supplemental Figure 1. N50 values for plant genomes published over the last 15 years.

Supplemental Figure 2. Theoretical example of a k-mer frequency spectrum.

Supplemental Table 1. Data sets used to evaluate genome assembly and gene space completeness measures.

Supplemental Methods. Genome size estimation using k-mer frequency spectra and implementation of gene space completeness measures

ACKNOWLEDGMENTS

We thank Ronnie de Jonge and Michiel Van Bel for critically reading the manuscript, Shu-Min Kao and Lieven Sterck for their help with the BUSCO analyses, and the editor and reviewers for helpful suggestions to improve the manuscript. This work was supported by the Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks" Project (no. 01MR0410W) of Ghent University.

AUTHOR CONTRIBUTIONS

E.V. retrieved and analyzed the data sets. E.V., T.R., and K.V. contributed to writing the article. K.V. coordinated the project.

Received May 3, 2016; revised July 13, 2016; accepted August 9, 2016; published August 10, 2016.

REFERENCES

Bassel, G.W., Gaudinier, A., Brady, S.M., Hennig, L., Rhee, S.Y., and De Smet, I. (2012). Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks. Plant Cell **24:** 3859–3875.

Bennetzen, J.L., and Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu. Rev. Plant Biol. **65:** 505–530.

Byrne, S.L., et al. (2015). A synteny-based draft genome sequence of the forage grass *Lolium perenne*. Plant J. **84:** 816–826.

Campbell, M.S., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. **164:** 513–524.

Cao, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. **43:** 956–963.

Chia, J.M., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. **44:** 803–807.

Dolezel, J., and Bartos, J. (2005). Plant DNA flow cytometry and estimation of nuclear genome size. Ann. Bot. (Lond.) **95:** 99–110.

Earl, D., et al. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. **21:** 2224–2241.

Flot, J.F., Marie-Nelly, H., and Koszul, R. (2015). Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. FEBS Lett. **589:** 2966–2974.

Garcia, S., et al. (2014). Recent updates and developments to plant genome size databases. Nucleic Acids Res. **42:** D1159–D1166.

Hardigan, M.A., et al. (2016). Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. Plant Cell **28:** 388–405.

COMMENTARY

Hirsch, C.N., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. Plant Cell **26:** 121–135.

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

Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y., Edger, P.P., Altman, N.S., Pires, J.C., Leebens-Mack, J.H., and dePamphilis, C.W. (2016). Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. PLoS One **11:** e0146062.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T.D. (2013). REAPR: a universal tool for genome assembly evaluation. Genome Biol. **14:** R47.

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

Li, J.Y., Wang, J., and Zeigler, R.S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience **3:** 8.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv preprint arXiv:1308.2012.

Marroni, F., Pinosio, S., and Morgante, M. (2014). Structural variation and genome complexity: is dispensable really dispensable? Curr. Opin. Plant Biol. **18:** 31–36.

Mascher, M., et al. (2013). Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J. **76:** 718–727.

Mendelowitz, L., and Pop, M. (2014). Computational methods for optical mapping. Gigascience **3:** 33.

Michael, T.P., and Jackson, S. (2013). The First 50 Plant Genomes. Plant Genome **6:** doi/10.3835/plantgenome2013.03.0001in.

Olsen, J.L., et al. (2016). The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. Nature **530:** 331–335.

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

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. Nucleic Acids Res. **37:** 289–297.

Rahman, A., and Pachter, L. (2013). CGAL: computing genome assembly likelihoods. Genome Biol. **14:** R8.

Salzberg, S.L., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. **22:** 557–567.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31:** 3210–3212.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., and Vandepoele, K. (2012). Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol. **158:** 590–600.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: a tale of duplications. Trends Plant Sci. **14:** 680–688.

Velasco, R., et al. (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat. Genet. **42:** 833–839.

Wang, W., et al. (2006). High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell **18:** 1791–1802.

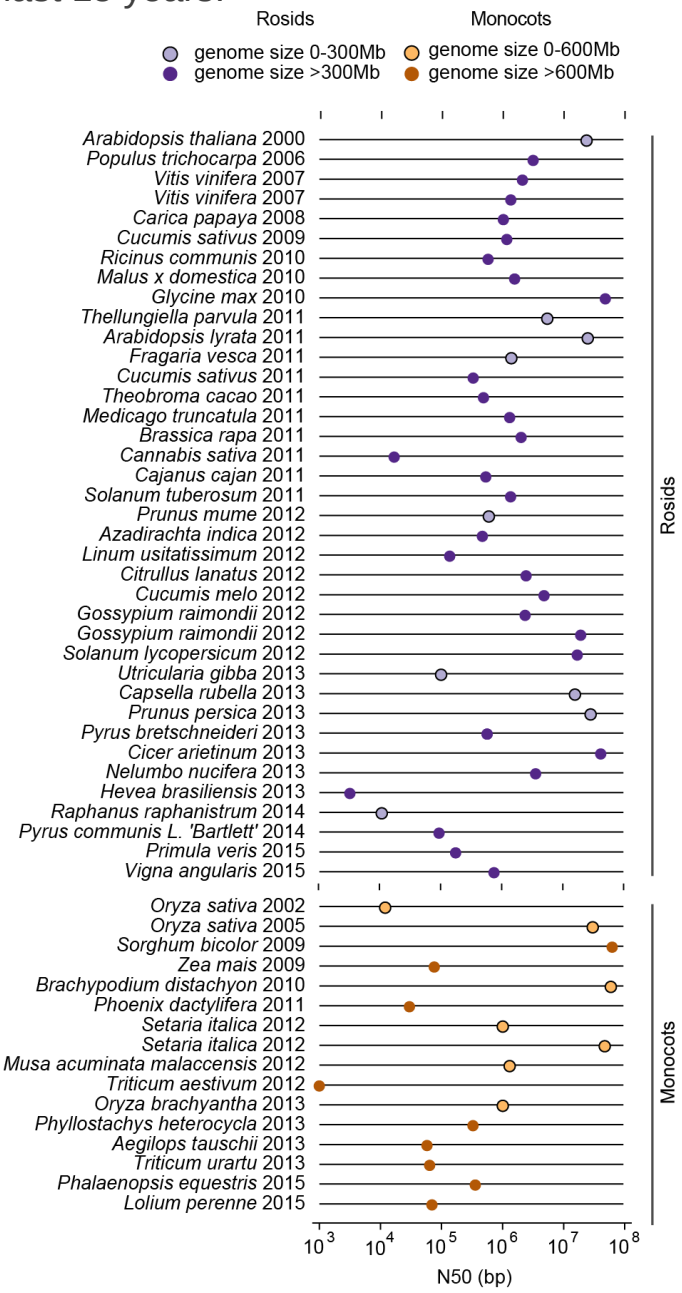
Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. **10:** 107.

Wendel, J.F., Jackson, S.A., Meyers, B.C., and Wing, R.A. (2016). Evolution of plant genome architecture. Genome Biol. **17:** 37.

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X., and Xie, W. (2015). Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol. **16:** 187.

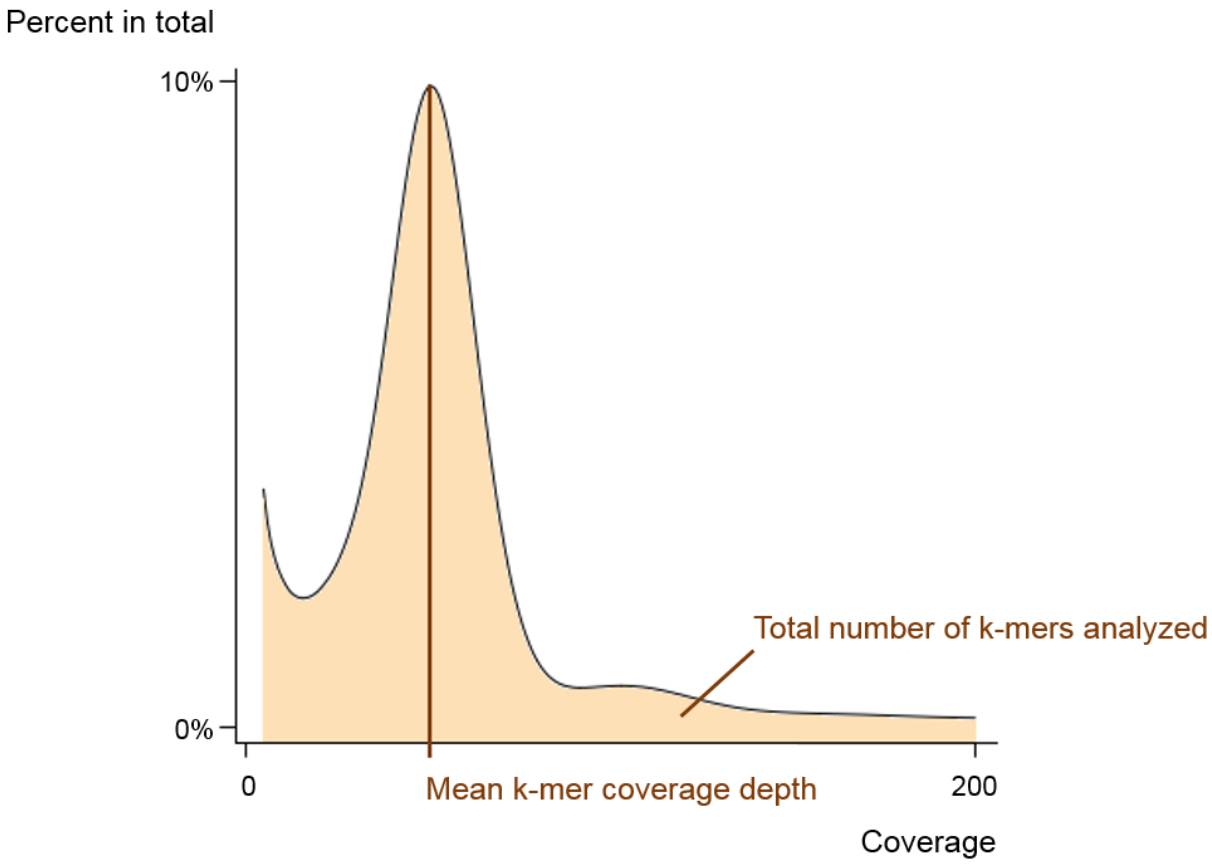
Zonneveld, B.J.M., Leitch, I.J., and Bennett, M.D. (2005). First nuclear DNA amounts in more than 300 angiosperms. Ann. Bot. (Lond.) **96:** 229–244.

Supplemental Figure 1. N50 values for plant genomes published over the last 15 years.



The N50 values of the first 50 published plant genomes were collected from Michael and Jackson (2013), complemented with the ten species used in the comparison of measures for genome and gene space completeness. The species are ordered according to their lineage (Rosids, purple; Monocots, orange) and publication date.

Supplemental Figure 2. Theoretical example of a k-mer frequency spectrum.



A k-mer frequency spectrum shows the percentage of k-mers analyzed classified by each coverage depth. The area under the curve reflects the total number of k-mers analyzed, while the peak value depicts the mean k-mer coverage depth. The genome size can be estimated as their respective ratio. An additional peak positioned left from the mean k-mer coverage depth would correspond to k-mers that are associated with SNPs or other polymorphisms. This allows to estimate the polymorphic rate in highly polymorphic genomes. Sometimes the quality of the sequencing data is poor. This would lead to a shift towards lower k-mer frequencies indicating the number of read errors. K-mers that are localized in repeat regions will not appear uniquely in the genome. Their frequencies will reflect their copy number in the genome, leading to a bump at higher coverage. Adapted from http://koke.asrc.kanazawa-u.ac.jp/HOWTO/kmer-genomesize.html

Supplemental Table 1. Datasets used to evaluate genome assembly and gene space completeness measures.

Species	Taxonomic clade	Size (Mb)	# sequences	Scaffold N50 (kb)	# ESTs	Ref.
Arabidopsis thaliana	Rosids	125	7	23,460	1.529.700	(Parra et al., 2007)
Capsella rubella	Rosids	219		15,100	NA	(Haudry et al., 2013); Slotte et al. (2013)
Cicer arietinum L.	Rosids	738	181.462	39,990	44.618	(Parween et al., 2015)
Nelumbo nucifera Gaertn.	Rosids	929	3334	3400	2207	(Ming et al., 2013)
Primula veris	Rosids	302		164	NA	(Nowak et al., 2015)
Pyrus communis L. 'Bartlett'	Rosids	265	142.083	27,400	450	(Chagné et al., 2014)
Raphanus raphanistrum	Rosids	254	68.331	10	81.524	(Moghe et al., 2014)
Vigna angularis	Rosids	443	3387	703	11.199	(Kang et al., 2015)
Lolium perenne	Monocots	1128	48.415	70	19.774	(Byrne et al., 2015)
Oryza sativa	Monocots	389	16	29,895	987.327	(Parra et al., 2007)
Setaria italica	Monocots	510	37.854	47,600	66.027	(Zhang et al., 2012)
Phalaenopsis equestris	Monocots	1086	236.185	359	5604	(Cai et al., 2015)

* CEGMA score reported in Figure 4 was obtained from this reference.

In total, twelve species including rosids and monocots were used to compare gene space completeness measures. Based on an initial list of 18 studies that used CEGMA to assess the completeness of a sequencing project within flowering plants, assembled sequence information could be retrieved for ten species. These datasets covered seven rosid species (*C. rubbella, C. arietinum L., N. nucifera Gaertn., P. veris, P. communics L.* 'Bartlett'*, R. raphanistrum, V. angularis*) and three monocots (*L. perenne*, *P. equestris*, *S. italica*). *A. thaliana* and *O. sativa* were also included as the oldest, high-quality reference genomes, which were sequenced using a gold-standard BAC-clone based approach and are thoroughly expert curated.

Supplemental Methods

Genome size estimation using k-mer frequency spectra

The genome size can be estimated by counting the k-mer frequencies. Several algorithms are available that count the number of occurrences of each substring of length k in raw sequencing data. The results are summarized in a histogram, leading to an empirical distribution of the DNA k-mers. Several models for the distribution of k-mers have been proposed that try to estimate genome characteristics more accurately in highly repetitive or heterozygous genomes (Liu et al., 2013).

The genome size can be calculated using the information one gets from a k-mer frequency spectrum (Supplemental Figure 2). First, the total number of k-mers analyzed should be determined. This number is equal to the area under the frequency curve, and can be calculated as the total number of reads multiplied by the number of k-mers that can be found in each read:

$$total\ number\ of\ kmers = total\ number\ of\ reads \times (read\ length - k + 1)$$

Next, the depth of coverage should be determined. For non-repetitive regions of the genome, the histogram should be normally distributed around a single peak. The peak value is the mean k-mer coverage depth in the sequencing data.

Finally, the genome size can be calculated as follows:

$$genome\ size = \frac{total\ number\ of\ kmers}{kmer\ coverage\ depth}$$

CEGMA

The CEGMA completeness score reports the number of conserved eukaryotic genes that could be found in the genome assembly using an accurate mapping protocol (Figure 1C). Partial and complete CEGMA scores refer to the presence of a gene fragment or a complete copy, respectively. For the ten species included in the comparison, the complete CEGMA score was extracted from the corresponding genome paper. The CEGMA score of *A. thaliana* is equal to 100%, as this species was one of the six eukaryotic species used to define the CEGMA core gene set.

BUSCO

Whereas CEGMA only works on raw genome or transcript sequences and performs gene prediction prior to the completeness estimation, BUSCO can be applied on a genome sequence as well as on an annotated gene set. For all species included in the comparison, the BUSCO plant profiles (only available upon request) were used to calculate genome and gene set completeness scores. For dicot and monocot species, Augustus was trained with *A. thaliana* or *Zea mays*, respectively. The completeness score was calculated as the percentage of complete and partially recovered BUSCO groups.

Transcript mapping score

For twelve species, EST sequences were obtained from the NCBI Nucleotide EST database (downloaded on October 12, 2015). The EST sequences were mapped to their respective reference genome using GMAP with default parameters (Wu and Watanabe, 2005). We collected all EST sequences that are publicly available for *A. thaliana* (186 libraries, ranging from 1 to 541,852 ESTs per library, 1,529,700 ESTs in total) and *O. sativa* (220 libraries, ranging from 1 to 53,637 ESTs per library, 987,327 ESTs in total). For *A. thaliana* and *O. sativa*, all ESTs were also mapped on simulated incomplete genomes. To simulate genome fragmentation, the genome was cut into pieces of 10kb and incomplete genomes were constructed by randomly selecting 50%, 75%, 80%, 90%, 95%, and 100% of these fragments. Next, we randomly sampled ESTs, to construct bin sizes containing 100 up to 300,000 ESTs, and for each bin we estimated which fraction was mapped onto the genome. Optionally, an extra filtering step was applied retaining

only EST mappings with >90% coverage. For each bin size, the mean and standard deviation of the transcript mapping score was calculated over 100 random replicates per bin size. In a second approach, each EST was assigned to its original EST library, and the transcript mapping score was calculated per EST library.

CoreGF completeness score

Three sets of coreGFs have been defined: green plants, based on conserved genes in 25 species including angiosperms, mosses and green algae (2928 coreGFs); rosids, based on conservation in 12 species (6092 coreGFs); and monocots, based on conservation in 5 species (7076 coreGFs). A BLAST-based sequence similarity search is applied per set of transcript sequences or predicted proteins to detect the presence of a coreGF, using one representative protein per coreGF. The representation across all individual coreGFs is summarized in a global weighted coreGF score, where large gene families get a smaller weight than single-copy families, as the former have a higher probability to be detected. All three coreGF sets, a preformatted BLAST database and a python3 script to calculate the coreGF score are available via ftp://ftp.psb.ugent.be/pub/plaza/plaza_public_02_5/coreGF/. More details can be found in the enclosed README file.

Expression and gene function bias of CEGMA and coreGFs

Gene function bias was determined through Gene Onotology enrichment analysis using the PLAZA 3.0 Dicots Workbench using GO source ‘primary’ (Proost et al., 2015) for the CEGMA core genes, BUSCO groups and the coreGFs of green plants and rosids. As there is no information available on which *A. thaliana* genes are present in the BUSCO plant profiles, the best hits of the *A. thaliana* gene set were used. The expression bias was assessed through *A. thaliana* gene expression analysis using the Compendium2 from the CORNET database (De Bodt et al., 2010). Highly similar experiments were removed by clustering the experiments using a 0.95 Pearson Correlation Coefficient threshold and taking into account the sample descriptions available in Gene Expression Omnibus. This resulted in an expression atlas of 75 experiments. Expression bias was determined for all expressed *A. thaliana* genes, the CEGMA core genes, the *A. thaliana*

BUSCO best hits (n=850 because some genes are not present on the ATH1 microarray) and the coreGFs of green plants and rosids. For each gene in these gene sets, we counted the number of experiments in which the gene is expressed (expression value > 2^7.5) and summarized the values in an expression breadth histogram.

Supplemental references

Byrne, S.L., Nagy, I., Pfeifer, M., Armstead, I., Swain, S., Studer, B., Mayer, K., Campbell, J.D., Czaban, A., Hentrup, S., Panitz, F., Bendixen, C., Hedegaard, J., Caccamo, M., and Asp, T. (2015). A synteny-based draft genome sequence of the forage grass Lolium perenne. Plant J **84,** 816-826.

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.C., Liu, K.W., Chen, L.J., He, Y., Xu, Q., Bian, C., Zheng, Z., Sun, F., Liu, W., Hsiao, Y.Y., Pan, Z.J., Hsu, C.C., Yang, Y.P., Hsu, Y.C., Chuang, Y.C., Dievart, A., Dufayard, J.F., Xu, X., Wang, J.Y., Wang, J., Xiao, X.J., Zhao, X.M., Du, R., Zhang, G.Q., Wang, M., Su, Y.Y., Xie, G.C., Liu, G.H., Li, L.Q., Huang, L.Q., Luo, Y.B., Chen, H.H., Van de Peer, Y., and Liu, Z.J. (2015). The genome sequence of the orchid Phalaenopsis equestris. Nature genetics **47,** 65-72.

Chagné, D., Crowhurst, R.N., Pindo, M., Thrimawithana, A., Deng, C., Ireland, H., Fiers, M., Dzierzon, H., Cestaro, A., and Fontana, P. (2014). The draft genome sequence of European pear (Pyrus communis L.'Bartlett'). PloS one **9,** e92644.

De Bodt, S., Carvajal, D., Hollunder, J., Van den Cruyce, J., Movahedi, S., and Inze, D. (2010). CORNET: A User-Friendly Tool for Data Mining and Integration. Plant physiology **152,** 1167-1179.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., Dewar, K., Stinchcombe, J.R., Schoen, D.J., Wang, X.W., Schmutz, J., Town, C.D., Edger, P.P., Pires, J.C., Schumaker, K.S., Jarvis, D.E., Mandakova, T., Lysak, M.A., van den Bergh, E., Schranz, M.E., Harrison, P.M., Moses, A.M., Bureau, T.E., Wright, S.I., and Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature genetics **45,** 891-U228.

Kang, Y.J., Satyawan, D., Shim, S., Lee, T., Lee, J., Hwang, W.J., Kim, S.K., Lestari, P., Laosatit, K., and Kim, K.H. (2015). Draft genome sequence of adzuki bean, Vigna angularis. Scientific reports **5**.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv preprint arXiv:1308.2012.

Ming, R., VanBuren, R., Liu, Y., Yang, M., Han, Y., Li, L.-T., Zhang, Q., Kim, M.-J., Schatz, M.C., and Campbell, M. (2013). Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). Genome biology **14,** R41.

Moghe, G.D., Hufnagel, D.E., Tang, H., Xiao, Y., Dworkin, I., Town, C.D., Conner, J.K., and Shiu, S.-H. (2014). Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish Raphanus raphanistrum and three other Brassicaceae species. The Plant Cell Online **26,** 1925-1937.

Nowak, M.D., Russo, G., Schlapbach, R., Huu, C.N., Lenhard, M., and Conti, E. (2015). The draft genome of Primula veris yields insights into the molecular basis of heterostyly. Genome biology **16**.

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

Parween, S., Nawaz, K., Roy, R., Pole, A.K., Suresh, B.V., Misra, G., Jain, M., Yadav, G., Parida, S.K., and Tyagi, A.K. (2015). An advanced draft genome assembly of a desi type chickpea (Cicer arietinum L.). Scientific reports **5**.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inze, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. Nucleic acids research **43,** D974-981.

Slotte, T., Hazzouri, K.M., Ågren, J.A., Koenig, D., Maumus, F., Guo, Y.-L., Steige, K., Platts, A.E., Escobar, J.S., and Newman, L.K. (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nature genetics **45,** 831-835.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21,** 1859-1875.

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., and Wang, W. (2012). Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. Nature biotechnology **30,** 549-554.