Reference-based QUantification Of gene Dispensability (QUOD)

Katharina Frey[1,2], Bernd Weisshaar[1,*] and Boas Pucker[1,3]

[1]Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University, 33615 Bielefeld, Germany

[2]Graduate School DILS, Bielefeld Institute for Bioinformatics Infrastructure (BIBI), Bielefeld University, 33615 Bielefeld, Germany

[3]Evolution and Diversity, Department of Plant Sciences, University of Cambridge, Cambridge, UK

*Correspondence: weisshaa@cebitec.uni-bielefeld.de

Abstract

Dispensability of genes in a phylogenetic lineage, e.g. a species, genus, or higher-level clade, is gaining relevance as most genome sequencing projects move to a pangenome level. Most analyses classify genes as core genes, which are present in (almost) all investigated individual genomes, and dispensable genes, which only occur in a single or a few investigated genomes. The binary classification as ‘core’ or ‘dispensable’ is often based on arbitrary cutoffs of presence/absence in the analysed genomes. Even when extended to ‘conditionally dispensable’, this concept still requires the assignment of genes to distinct groups.

Here, we present a new method which overcomes this distinct classification by quantifying gene dispensability and present a dedicated tool for reference-based QUantification Of gene Dispensability (QUOD). As a proof of concept, sequence data of 966 *Arabidopsis thaliana* accessions were processed to calculate a gene-specific dispensability score based on normalised coverage in read mappings. We validated this score by comparison of highly conserved Benchmarking Universal Single Copy Orthologs (BUSCOs) to all other genes. The average scores of BUSCOs were significantly higher than the scores of non-BUSCOs. Scores of genes involved in specialised metabolism were generally lower than scores of genes belonging to primary metabolism, indicating higher dispensability of genes involved in secondary metabolic processes. Analysis of variation demonstrated lower variation values between replicates of a single accession than between iteratively, randomly selected accessions from the whole dataset.

Instead of classifying a gene as core or dispensable, QUOD assigns a dispensability score to each gene. Hence, QUOD facilitates the identification of candidate dispensable genes which often underlie lineage-specific adaptation to varying environmental conditions.

Introduction

Genetic variation is not restricted to single nucleotide polymorphisms or small insertions and deletions but extends also to (large) structural variations. These structural variations include copy number variations (CNVs) and presence/absence variations (PAVs), which can cause substantial variation of the gene content among individual genomes [1,2]. The comparative analysis of multiple genomes of the same phylogenetic clade allows the identification of PAVs underlying phenotypic traits. In the case of crop species, the identification of PAVs underlying specific agronomic traits which only occur in a single or a few species is feasible [3–5]. As more highly contiguous genome sequences become available, pangenomes are suitable to describe and investigate the gene set diversity of a biological clade, e.g. species, genus or higher [6,7].

Genes of a pangenome are thought to be divided into a core and a dispensable gene set, the latter is also often referred to as ‘accessory’ in the literature. Core genes occur in all or almost all investigated genomes, whereas dispensable genes only occur in a single or a few genomes [8]. Sometimes, a third category of ‘conditionally dispensable’ genes is invoked [9]. However, this distinct classification relies on one or multiple arbitrary cutoffs. Some studies consider genes as ‘core’ if these genes occur in at least 90 % of the investigated genomes [10]; in other studies, only genes which are found in all genomes are part of the core genome [11]. In addition, dependency groups might influence the dispensability of certain genes. The possibility that two genes might be ‘replaced’ by a specific number of other genes has to be considered. Some genes, of e.g. a gene family, might be required in a specific proportion and therefore are only conditionally dispensable [9]. Further, assemblies of genomes or transcriptomes might be incomplete leading to artificially missing genes [12]. One way to circumvent this is to rely on a high-quality reference genome sequence, thus avoiding additional assemblies which are potential sources of errors.

Here, we present QUOD - a bioinformatic tool to quantify gene dispensability. An *A. thaliana* dataset of about 1,000 accessions was used to calculate a per gene dispensability score derived from the coverage of all genes in the given genomes. This score was validated by comparison of scores of BUSCOs and scores of secondary metabolite genes. Our tool is easy to use for all kinds of plant species and datasets of different levels of phylogenetic resolution. QUOD extends the distinct

classification of genes as ‘core’ and ‘dispensable’ based on an arbitrary threshold to a continuous dispensability score.

Methods

Selection and preprocessing of datasets

Genomic reads (FASTQ format) of the investigated genomes were retrieved from the Sequence Read Archive (SRA) [13] via fastq-dump. BWA-MEM [14] was applied to map all genomic paired-end Illumina reads to the corresponding reference genome sequence. For *A. thaliana*, all available 1,135 datasets [15] (File S1) were subjected to a mapping against the AthNd-1_v2c genome sequence [16]. The resulting BAM files of these mappings were subjected to QUOD.

Calculation of gene dispensability scores – QUOD

QUOD calculates a reference-based gene dispensability score for each annotated gene based on a supplied mapping file (BAM) and annotation of the reference sequence (GFF) (https://github.com/KatharinaFrey/QUOD). The tool is written in Python3 and consists of six different modules. During the first part of the analysis, the read coverage per position (1) as well as the read coverage per gene (2) are calculated. In the next step, genomes with an average coverage below a given cutoff (default=10) are discarded and excluded from further analyses (3). Finally, an input matrix is constructed (4) and a dispensability score is determined for each gene (5). Optionally, the results can be visualized as a colored histogram and a violin plot (6).

The dispensability score (ds(g)) is calculated as follows:

$$\text{dispensability score (gene g)} = \frac{\sum\left(\frac{\text{average coverage of gene g}}{\text{mean coverage over all genes in genome x}}\right)}{\text{total number of genomes}}$$

Identification of plastid sequences

Genes of the *A. thaliana* test set with high similarity to plastid sequences were flagged via BLASTp [17] of the encoded peptides against all organelle peptide sequences obtained from the National Center for Biotechnology Information (NCBI). As a control, the sequences were also searched against themselves. Peptide sequences of Nd-1 with a score ratio ≥ 0.8 were considered plastid-like sequences when comparing BLAST hits against self-hits [16].

Score comparison between contrasting gene sets

Genes annotated in AthNd-1_v2c were classified with BUSCO v3 [18] running in protein mode on the encoded peptide sequences using 'brassicales odb10' as reference [19]. BUSCOs include single-copy genes and universal genes which are present in > 90% of all species in the reference dataset and are used to measure the completeness of assemblies and annotations [18]. The scores (≤ 2) of BUSCO and non-BUSCO genes were compared using matplotlib [20] for visualization (boxplot) and a Mann–Whitney U test implemented in the Python package SciPy [21] for determination of the significance.

Primary and secondary metabolite genes were identified using the KEGG Orthology for *A. thaliana* (ath00001.keg) downloaded from the KEGG database [22] and a list of primary and secondary metabolic pathways obtained from Mukherjee *et al.* [23]. In total, 1,887 genes of Nd-1 were assigned to the primary metabolism, whereas 440 genes were assigned to the secondary metabolism. The average scores of these groups of genes were compared statistically by a Mann-Whitney U test within a custom Python script using the package SciPy [21] and a boxplot was created using matplotlib.

A list of Nd-1 transposable element (TE) genes, which are Nd-1 gene structures overlapping with sequences annotated as TEs, was obtained from Pucker *et al.* [16]. First, the score distribution (ds > 2 is set to ds = 2) of TE and non-TE genes was determined using a Mann–Whitney U test implemented in the Python package SciPy [21]. Next, the minimal distance of each gene to its closest TE gene was calculated after extracting the gene positions from the Nd-1 annotation file. Pearson's correlation coefficient between per gene distance to TEs and the gene dispensability score (ds > 2 is set to ds = 2) was determined.

Correlation of gene length and exon number with the dispensability score

Length and number of exons per gene were extracted from the Nd-1 annotation file. The Pearson's correlation coefficients of gene length and exon number, respectively, with the gene dispensability score (ds > 2 is set to ds = 2) were calculated via SciPy [21] for the whole dataset as well as for three large *A. thaliana* gene families (TAPscan [24]), namely MYBs, AP2/EREBP and WRKYs .

Variation between replicates

A total of 14 genomic datasets of the *A. thaliana* accession Col-0 were received from the SRA (File S2) to assess the technical variation between replicates of the same accession. Each dataset was mapped to the TAIR10 reference genome sequence using BWA-MEM. The mappings were then subjected to QUOD expecting a dispensability score close to one for each gene. As the distributions are different (Kolmogorov-Smirnov test, $p \approx 3e\text{-}27$) and the sample size (n) is high, the Levene's test was selected to test for equal variances, regarding the gene dispensability scores. The test was applied for (1) the dataset including replicates only and (2) iteratively (100x), randomly chosen subsets (n=14) of the dataset including all available *A. thaliana* accessions.

Functional annotation

All genes of the *A. thaliana* Nd-1 genome sequence were functionally annotated via reciprocal best blast hits (RBHs) and best BLAST hits against Araport11 [16].

Data Availability

The tool QUOD (QUOD.py) can be downloaded from GitHub (https://github.com/KatharinaFrey/QUOD).

Results

In this study, a bioinformatic tool was developed to calculate a gene-specific dispensability score based on the normalised coverage in a read mapping. QUOD allows the quantification of dispensability by calculation of a score for each gene (Figure 1). The binary classification of gene dispensability can be compared to the original method of mRNA detection by endpoint RT-PCR providing only qualitative results [25–27] which was replaced by quantitative analyses like RNA-Seq.

The gene dispensability score would initially be dependent on the sequencing depth per genome. By division of the average coverage of gene g by the mean coverage over all genes in genome x, the score is normalised for differences in the sequencing depth of the investigated genomes. A low value indicates that a gene is likely to be missing in some genomes and therefore more likely dispensable than a gene with a higher dispensability score. Due to this quantification approach, this method is not based on an arbitrary cutoff to determine the core genome and the dispensable genome of any given pangenome dataset. An example: Using a cutoff of ‘gene n occurs in at least 90 % of all genomes’ to be considered a ‘core’ gene (dark blue), genes 1,2,4,5 and 6 (dark grey) would be considered ‘dispensable’ (Figure 1). However, considering the coverage (right panel), is e.g. gene 1 truly dispensable? QUOD does not rely on any thresholds for the classification of genes into ‘core’ and ‘dispensable’, but provides a score based on the normalized coverage in a read mapping. The genes could theoretically be ranked as well using the percentage of presence/absence of a gene in the investigated genomes. However, this approach would still rely on a threshold, like e.g. the number of mapped reads for a gene to be present in a genome, to determine the presence/absence of a specific gene.

Figure 1: Illustration of the QUOD method using an artificial dataset. On the left side, genes are classified as ‘core’ (dark blue) or ‘dispensable’ (dark grey) according to a

cutoff. On the right side, gene dispensability is quantified according to a dispensability score based on the normalised coverage in a read mapping (a-f: investigated genomes). Coloring of genes (right side) indicates different dispensability scores. Extremely rare genes can be easily detected using QUOD.

As a proof of concept, *A. thaliana* sequencing datasets of 1,135 accessions were downloaded and mapped to the *A. thaliana* Nd-1 genome sequence. All genomes with less than 10-fold read coverage were discarded. The remaining sequencing datasets covering 966 accessions were analysed with QUOD to calculate a dispensability score for each gene (Figure 2). Genes with low dispensability scores, colored in pink, are considered to be likely dispensable, whereas genes with dispensability scores close to one (dark purple/dark blue) are most likely considered core genes. The origin of high scoring, bright blue colored genes can be explained by the genome wide distribution of the dispensability scores (see below).

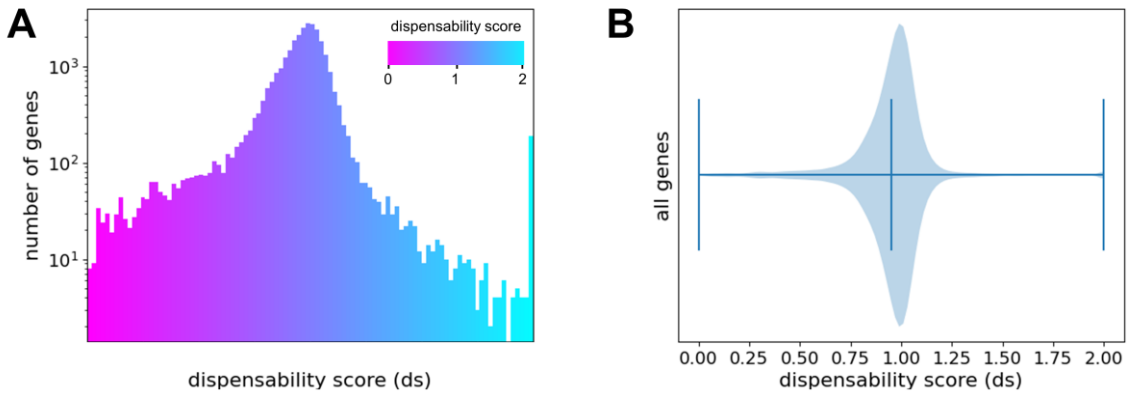

Figure 2: Distribution of the gene dispensability scores for the *A. thaliana* datasets. A) Histogram colored according to the dispensability score. The x-axis represents the dispensability score, whereas the y-axis shows the number of genes in each bin in logarithmic scale. B) Violin plot representing the dispensability score (y-axis) of all genes (x-axis).

Further validation of the reliability of the gene dispensability quantification was achieved by comparison of BUSCOs and non-BUSCOs (Figure 3A) as well as primary and secondary metabolite genes (Figure 3B). BUSCO genes show significantly higher scores than non-BUSCO genes (U test, $p \approx 1e\text{-}122$). Secondary metabolite genes show significantly lower scores than primary metabolite genes (U test, $p \approx 6e\text{-}11$).

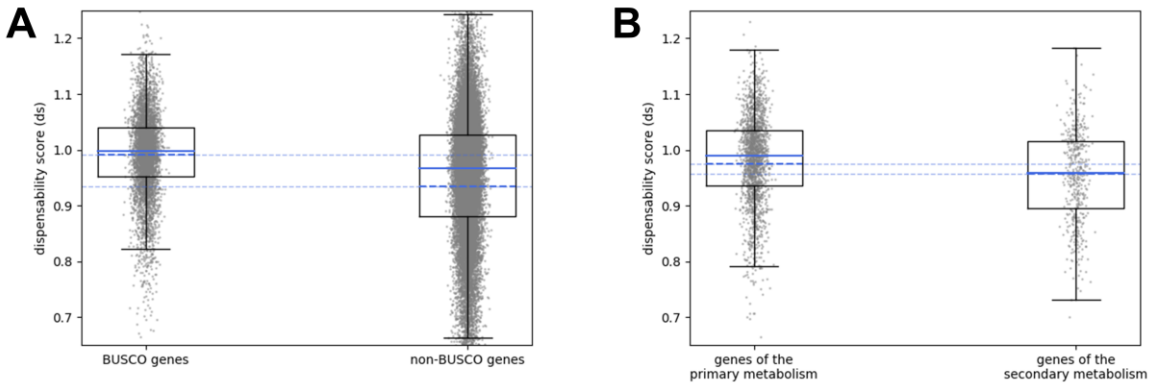

Figure 3: Box plot of the gene dispensability scores of (A) BUSCO and non-BUSCO genes and (B) primary and secondary metabolite genes. The mean is represented by the dashed, blue line, the other blue line represents the median of the scores. A) BUSCO genes (n = 4586, mean ds ≈ 0.991) show significantly higher scores than non-BUSCO genes (n = 25354, mean ds ≈ 0.934) (U test, p≈1e-122). B) Secondary metabolite genes (n = 440, mean ds ≈ 0.957) show significantly lower scores than primary metabolite genes (n = 1,887, mean ds ≈ 0.975) (U test, p≈6e-11).

Functional annotation of BUSCO outliers, which include genes with dispensability scores below 0.75 or above 1.25, revealed, amongst others, several repeat proteins, transmembrane proteins, a ‘stress induced protein’, and multiple hypothetical proteins (File S3).

Next, the genome-wide distribution of genes with specific gene dispensability scores was investigated in *A. thaliana*. A high dynamic across centromeric regions is visible based on the colored heatmap (Figure 4). High and low scoring genes cluster in centromeric and telomeric regions.

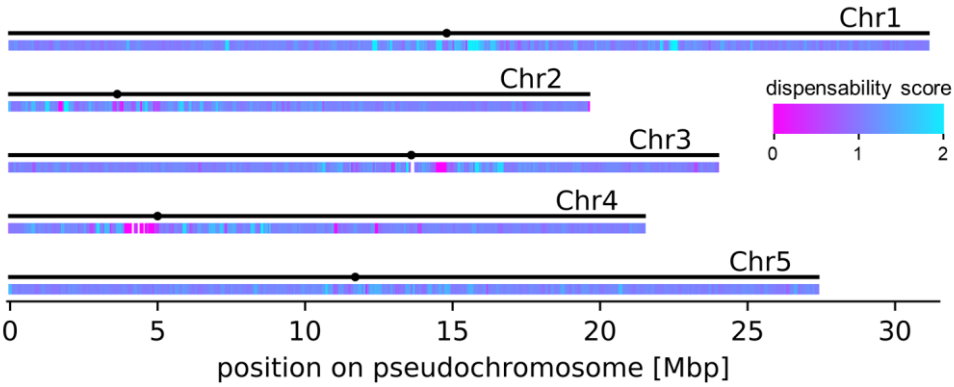

Figure 4: Genome-wide distribution of genes with different dispensability scores in *A. thaliana* Nd-1. The colored heatmap shows the respective gene dispensability scores. There are high (blue) and low (pink) scoring genes clustered in repetitive regions, including centromeric and telomeric areas. The x-axis represents the size (in Mbp) of each pseudochromosome in the assembly. The black dots represent the position of the centromeres of the five chromosomes in the AthNd1_v2c assembly [16].

As high and low scoring genes cluster in repetitive regions (mainly centromeres), the score distribution of TE genes was investigated (File S4). Scores of TE genes are evenly distributed across all dispensability scores. In total, the mean score of TE genes (mean ds ≈ 0.919) is significantly lower when compared to non-TE genes (mean ds ≈ 0.953) (U test, p ≈ 6e-08), which are more frequent across scores close to one. Moreover, the minimal distance of each gene to its closest TE gene and the dispensability scores revealed no correlation (File S4).

To test the hypothesis whether genes with lower dispensability scores/more likely dispensable genes are shorter and whether introns accumulate in core genes, the correlation of the gene dispensability score with gene length and exon number, respectively, were determined for the whole dataset and for three selected gene families separately. However, no clear trend was detectable (File S5).

Further, the variation between replicates of the same accession (Col-0) was determined (File S6). The variation of the gene dispensability score distribution of the replicate dataset (one accession) ($\sigma^2 \approx 0.0226$) is significantly lower than the variation between all iteratively, randomly selected subsets of *A. thaliana* accessions ($\sigma^2 \approx 0.0392$) (Levene's test, p ≈ 4e-19).

The function of the 100 genes with the lowest gene dispensability scores was examined in detail for the *A. thaliana* (File S7) dataset. Fourteen genes of this *A.*

thaliana test set are annotated as “disease resistance proteins”, whereas seven genes are annotated as transposons/transposases. Four genes are described as hypothetical proteins and 24 genes have no functional annotation.

Discussion

QUOD was developed for the quantification of gene dispensability in plant pangenome datasets, but the application is not limited to the plant kingdom. However, an accurate determination of gene dispensability scores free of systematic biases might rely on a uniform selection of genomes from the respective taxonomic group and on uniform read coverage of genes. In addition, non-random fragmentation of DNA prior to sequencing [28] may cause biases. The variation among replicates of the same accession (Col-0; $\sigma^2 \approx 0.0226$) might be attributed to technical biases, e.g. during sequencing library preparation.

We validated the reliability of the gene dispensability score by showing that BUSCO genes are more conserved and can be more likely considered core genes than non-BUSCOs as BUSCOs are associated with significantly higher scores (Figure 3A). In addition, secondary metabolite genes are not highly conserved and often taxon-specific [29] and therefore more likely dispensable as shown by the significantly lower dispensability scores in comparison to genes of primary metabolic processes (Figure 3B). Based on the distribution of the scores in the boxplot (Figure 3), the difference between BUSCOs and non-BUSCOs appears small, even though the difference is significant (U test, $p \approx 1e\text{-}122$). The expected difference in the dispensability scores of BUSCOs and non-BUSCOs might not be exceptionally high as multiple-copy genes are not included in the BUSCO gene set [18]. Further, functional annotation of BUSCO outliers revealed several repeat proteins and transmembrane proteins. Repeat proteins might lead to read mapping errors and consequently artificial variations in coverage and dispensability scores. Transmembrane proteins could be involved in biotic stress response and therefore might not be essential for some accessions [30]. This could explain the absence in some genomes resulting in low dispensability scores of these genes. Therefore, many important, high-scoring genes might lie outside of the BUSCO sampling space. The difference between scores of

primary and secondary metabolite genes might be higher as well, as strongly variable secondary metabolite genes might not be detected in our analyses due to the RBH-based annotation.

The genome-wide distribution of all gene dispensability scores of the *A. thaliana* genomes reveals the origin of exceptionally high dispensability scores (Figure 4). High scoring genes, which are colored in light blue (Figure 4), might be TEs and duplicated genes associated with collapsed sequences in the assembly. An accurate determination of the dispensability scores of these genes might be possible using ideal genome sequences without any collapsed regions as well as with error-free mapping. However, high scoring, light blue colored genes could still be useful to determine amplified TEs and duplicated genes. High and low scoring genes cluster in repetitive regions, like e.g. centromeres or telomeres, due to variations in the recombination rate [31] and active TEs in these regions. It was previously proposed, that likely dispensable genes are located closer to TEs which are important factors in pangenome evolution [32]. In our study, TE genes are widely distributed across all dispensability scores as TEs can occur in a high copy number in genomes leading to high scores and can as well be dispensable. Other studies detected a high number of TEs in the dispensable genome [33], however, only certain TE families might be truly dispensable. A high-quality annotation of transposons and a following exclusion of these genes from the analysis or improved read mapping to the consensus sequence might improve the results. Moreover, heterochromatin or genome-purging mechanisms [34] could influence the gene dispensability scores in these regions. Additionally some of the high scoring genes were flagged as plastid-like sequences as original sequencing data from plants contain a high amount reads originating from plastid sequences [35,36]. Biases due to this plastid read contamination inflate the coverage of sequences with high similarity to plastid sequences, resulting in an exceptionally high gene dispensability score. Therefore, especially genes with dispensability scores above two (ds > 2) must be investigated carefully and are therefore excluded from the analyses. Very similar sequences, e.g. members of a gene family or close paralogs, might cause read mapping errors confounding biases in the dispensability scores of these genes. Even though it was previously proposed that dispensable genes are located close to TEs [32], this result was not observed in our study. One limitation is the accurate assignment of reads to repetitive sections of the reference sequence during the read mapping [12]. Further, only a fraction of

transposons might be correctly assembled and annotated due to several computational challenges in highly repetitive and peri-centromeric regions [37]. Therefore, a different strategy might be needed to accurately quantify dispensability of TEs.

Functional annotation of the 100 most likely dispensable genes revealed a high number of uncharacterised proteins, disease resistance proteins as well as transposons and transposases in the *A. thaliana* genomes. TEs were detected in other studies as contributors to large structural variations between species and individuals and substantial part of the dispensable genome [33]. Previous pangenome analyses also revealed that the dispensable genome comprises functions like e.g. ‘defense response’, ‘diseases resistance’, ‘flowering time’ and ‘adaptation to biotic and abiotic stress’ [10,32,38]. Therefore, in depth investigation of genes with low dispensability scores can result in the identification and characterisation of phenotypic variation [39] and important agronomic traits [38]. We envision several applications for the gene dispensability score generated by QUOD: (1) more accurate prediction if a PAV is associated with a specific trait, (2) development of dependency gene networks, and (3) improved modeling of the evolutionary value of genes.

Conclusion

QUOD (reference-based QUantification Of gene Dispensability) overcomes the problem of labeling genes as ‘core’ or ‘dispensable’ through implementation of a quantification approach. Instead of a distinct classification, QUOD provides a ranking of all genes based on assigned gene-specific dispensability scores and therefore does not rely on any thresholds.

Author Contributions

KF, BW and BP designed the study, performed the experiments, analysed the data, and wrote the manuscript. All authors read and approved the final version of this manuscript.

Funding

This research received no external funding.

Acknowledgments

We thank members of Genetics and Genomics of Plants for discussion of preliminary results. We are very grateful to Janik Sielemann and Nathanael Walker-Hale for helpful comments on the manuscript. We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University. We thank the CeBiTec Bioinformatic Resource Facility team for great technical support.

Conflicts of Interest

The authors declare no conflict of interest.

Supplementary Files

File S1: SRA IDs of datasets downloaded to conduct the QUOD analysis of the *A. thaliana* genomes.

File S2: SRA IDs of datasets downloaded to conduct the analysis of replicates (Col-0).

File S3: Functional annotation of BUSCO outliers with a dispensability score smaller than 0.75 or greater than 1.25.

File S4: Distribution of scores of TE genes and non-TE genes and correlation of the distance to the closest TE gene with the gene dispensability score of the *A. thaliana* genomes.

File S5: Correlation of gene length and exon number with the dispensability scores of the *A. thaliana* genomes.

File S6: Analysis of variance of the gene dispensability score calculated for replicates of the *A. thaliana* Col-0 accession and iteratively, randomly chosen subsets of the whole dataset.

File S7: Functional annotation of the 100 most likely dispensable genes of the *A. thaliana* genomes.

References

1. Springer, N. M. *et al.* Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet* **5**, e1000734 (2009).
2. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nat Genet* **39**, S7–S15 (2007).
3. Tao, Y., Zhao, X., Mace, E., Henry, R. & Jordan, D. Exploring and Exploiting Pan-genomics for Crop Improvement. *Molecular Plant* **12**, 156–169 (2019).
4. Lu, F. *et al.* High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat Commun* **6**, 6914 (2015).
5. Swanson-Wagner, R. A. *et al.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research* **20**, 1689–1699 (2010).
6. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ‘pan-genome’. *Proceedings of the National Academy of Sciences* **102**, 13950–13955 (2005).
7. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Current Opinion in Microbiology* **23**, 148–154 (2015).
8. Golicz, A. A., Batley, J. & Edwards, D. Towards plant pangenomics. *Plant Biotechnol J* **14**, 1099–1105 (2016).
9. Marroni, F., Pinosio, S. & Morgante, M. Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology* **18**, 31–36 (2014).
10. Zhao, Q. *et al.* Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet* **50**, 278–284 (2018).

11. Li, Y. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**, 1045–1052 (2014).

12. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat Methods* **8**, 61–65 (2011).

13. Leinonen, Rasko and Sugawara, Hideaki and Shumway, Martin and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Research* **39**, D19–D21 (2010).

14. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

15. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).

16. Pucker, B. *et al.* A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. *PLoS ONE* **14**, e0216233 (2019).

17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

18. Simão, Felipe A and Waterhouse, Robert M and Ioannidis, Panagiotis and Kriventseva, Evgenia V and Zdobnov, Evgeny M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

19. Zdobnov, E. M. *et al.* OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* **45**, D744–D749 (2017).

20. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

21. Jones, E., Oliphant, T., Peterson, P. & others. SciPy: Open source scientific tools for Python. (2001).

22. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).

23. Mukherjee, D., Mukherjee, A. & Ghosh, T. C. Evolutionary Rate Heterogeneity of Primary and Secondary Metabolic Pathway Genes in *Arabidopsis thaliana*. *Genome Biol Evol* **8**, 17–28 (2016).

24. Wilhelmsson, P. K. I., Mühlich, C., Ullrich, K. K. & Rensing, S. A. Comprehensive Genome-Wide Classification Reveals That Many Plant-Specific Transcription Factors Evolved in Streptophyte Algae. *Genome Biology and Evolution* **9**, 3384–3397 (2017).

25. Wang, A. M., Doyle, M. V. & Mark, D. F. Quantitation of mRNA by the polymerase chain reaction. *Proceedings of the National Academy of Sciences* **86**, 9717–9721 (1989).

26. Gilliland, G., Perrin, S., Blanchard, K. & Bunn, H. F. Analysis of cytokine mRNA and DNA: detection and quantitation by competitive polymerase chain reaction. *Proceedings of the National Academy of Sciences* **87**, 2725–2729 (1990).

27. Chiang, P. W. *et al.* Use of a fluorescent-PCR reaction to detect genomic sequence copy number and transcriptional abundance. *Genome Research* **6**, 1013–1026 (1996).

28. Poptsova, M. S. *et al.* Non-random DNA fragmentation in next-generation sequencing. *Sci Rep* **4**, 4532 (2015).

29. Peregrín-Alvarez, J. M., Sanford, C. & Parkinson, J. The conservation and evolutionary modularity of metabolism. *Genome Biol* **10**, R63 (2009).

30. Dodds, P. N. & Rathjen, J. P. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nat Rev Genet* **11**, 539–548 (2010).

31. Nachman, M. Variation in recombination rate across the genome: evidence and implications. *Current Opinion in Genetics & Development* **12**, 657–663 (2002).

32. Gordon, S. P. *et al.* Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat Commun* **8**, 2184 (2017).

33. Morgante, M., Depaoli, E. & Radovic, S. Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology* **10**, 149–155 (2007).

34. Lee, S.-I. & Kim, N.-S. Transposable Elements and Genome Size Variations in Plants. *Genomics Inform* **12**, 87 (2014).

35. Yu, J. *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**, 79–92 (2002).

36. Tang, J. *et al.* A Comparison of Rice Chloroplast Genomes. *Plant Physiol.* **135**, 412–420 (2004).

37. Platt, R. N., Blanco-Berdugo, L. & Ray, D. A. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol Evol* **8**, 403–410 (2016).

38. Golicz, A. A. *et al.* The pangenome of an agronomically important crop plant Brassica oleracea. *Nat Commun* **7**, 13390 (2016).

39. Hirsch, C. N. *et al.* Insights into the Maize Pan-Genome and Pan-Transcriptome. *Plant Cell* **26**, 121–135 (2014).