Received: 18 April 2023 | Accepted: 13 October 2023

DOI: 10.1111/2041-210X.14243

Methods in Ecology and Evolution

APPLICATION

Assessing the quality of comparative genomics data and results with the *cogeqc* R/Bioconductor package

Fabricio Almeida-Silva[1,2] | Yves Van de Peer[1,2,3,4]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[2]VIB Center for Plant Systems Biology, VIB, Ghent, Belgium

[3]Department of Biochemistry, Genetics and Microbiology, Centre for Microbial Ecology and Genomics, University of Pretoria, Pretoria, South Africa

[4]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, China

Correspondence
Fabricio Almeida-Silva
Email: fabricio.almeidasilva@psb.vib-ugent.be

Funding information
H2020 European Research Council, Grant/Award Number: 833522; Universiteit Gent, Grant/Award Number: BOF.MET.2021.0005.01

Handling Editor: Francisco Balao

Abstract

1. Comparative genomics has become an indispensable part of modern biology due to the advancements in high-throughput sequencing technologies and the accumulation of genomic data in public databases. However, the quality of genomic data and the choice of parameters used in software tools used for comparative genomics can greatly impact the accuracy of results.
2. Here, we present *cogeqc*, an R/Bioconductor package that provides researchers with a toolkit to assess genome assembly and annotation quality, orthogroup inference, and synteny detection. The package offers context-guided assessments of assembly and annotation statistics by comparing observed statistics to those of closely-related species on NCBI. To assess orthogroup inference, *cogeqc* calculates a protein domain-aware orthogroup score that aims at maximising the number of shared protein domains within the same orthogroup. The assessment of synteny detection consists in representing anchor gene pairs as a synteny network and analysing its graph properties, such as clustering coefficient, node count, and scale-free topology fit.
3. The application of cogeqc to real datasets allowed for an evaluation of multiple parameter combinations for orthogroup inference and synteny detection, providing researchers in need for comparative genomics with guidelines to aid in the selection of the most appropriate tools and parameters for their specific data.
4. We demonstrate that the default parameters in orthogroup identification and synteny detection tools are not always the most suitable, highlighting the importance of performing assessments for each dataset. The assessment metrics provided by cogeqc will help researchers generate more accurate and reliable results.

KEYWORDS
evolutionary genomics, gene family, genome assembly, quality control

1 | INTRODUCTION

The availability of genomic data in public databases has increased exponentially due to advancements in sequencing technologies. However, a significant portion of this data is of poor quality, leading to incorrect or unreliable results in common analyses such as genome-wide synteny analysis and gene orthology detection (Feron & Waterhouse, 2022; Liu et al., 2018; Marks et al., 2021; Wang & Wang, 2022). Likewise, the choice of parameters in comparative genomics software can significantly impact the results obtained, as default parameters are typically optimised for particular (usually gold standard) datasets (Buchfink et al., 2021; Emms & Kelly, 2019). Therefore, careful data validation and quality control is crucial to ensure the accuracy and reliability of comparative genomics results.

Here, we present *cogeqc*, an R/Bioconductor package that can be used as a toolkit for assessing genome assembly and annotation statistics, orthogroup inference and synteny detection. The package offers context-guided assessments of assembly and annotation statistics by comparing observed values to those of closely related species on the National Center for Biotechnology Information (NCBI), while gene space completeness can be assessed with best universal single-copy orthologs (BUSCOs). The orthogroup inference assessment uses a protein domain-aware orthogroup score to maximise the number of shared protein domains within the same orthogroup. Finally, the assessment of synteny detection relies on representing anchor pairs as a synteny network and analysing its graph properties. The application of *cogeqc* to real datasets allowed for an evaluation of multiple parameter combinations for orthogroup inference and synteny detection, providing researchers with guidelines to aid in the selection of the most appropriate parameters for their specific data.

2 | IMPLEMENTATION

cogeqc is part of the Bioconductor ecosystem and, as such, can be easily integrated with other Bioconductor packages. Input data types are either base R or core Bioconductor classes (e.g. *DNAStringSet* and *AAStringSet* objects for DNA and protein sequences, respectively). For integration with external software tools (i.e. *BUSCO* (Simão et al., 2015) and OrthoFinder (Emms & Kelly, 2019)), we provide users with functions to read and parse their output for downstream analyses in *cogeqc*.

2.1 | Assessing genome assembly and annotation statistics

We propose a context-guided assessment of assembly and annotation statistics that consists in comparing observed values for common metrics (e.g. genome size, contiguity measures, number of genes, etc.) with those of closely related species on the National Center for Biotechnology Information (NCBI). For a particular taxon, the function *get_genome_stats()* extracts summary assembly and annotation statistics for all genomes on NCBI via the Datasets REST API (https://www.ncbi.nlm.nih.gov/datasets/) and returns a data frame with information on 35 variables, such as assembly level, scaffold and contig contiguity measures, number of coding and non-coding genes, and submitter data. In addition to the NCBI-provided statistics, the output data frame includes a variable 'CC ratio' representing the ratio of the number of contigs to the number of chromosome pairs, which has recently been proposed by (Wang & Wang, 2022) as a contiguity measure that compensates for the flaws of N50/L50 and allows cross-species comparisons.

Additionally, users can create a data frame containing assembly and annotation statistics for their own genome projects and pass it to the *compare_genome_stats()* function along with the output of *get_genome_stats()*. This function will add the user-provided statistics to a distribution of reference statistics from NCBI quality-checked genomes and report the percentile and rank of observed values in the distribution. Of note, statistics for NCBI genomes can also be obtained with the *getAssemblyStats()* function of the biomartr package (Drost & Paszkowski, 2017), but which is dramatically slower and limited to a single query species. Observed statistics can also be visually compared with reference statistics in publication-ready plots created by the function *plot_genome_stats()* (Figure 1a). Such context-guided assessments are particularly useful in cases when assembly statistics seem problematic (e.g. genome size is too large, number of genes is too small), but which are in fact due to genomic features of particular taxa, such as smaller genomes for parasitic (Xu et al., 2021), carnivorous (Palfalvi et al., 2020) and aquatic plants (An et al., 2019), and larger genomes for species with higher transposable element contents (Michael, 2014).

2.2 | Assessing gene space completeness

To assess gene space completeness, *cogeqc* relies on the identification of best universal single-copy orthologues (BUSCOs) (Simão et al., 2015). The function *run_busco()* is a wrapper that takes sequences as input (as FASTA files or as *AA/DNA/RNAStringSet* objects), runs BUSCO from the R session, and returns a data frame with the frequency of complete (duplicate and single copy), fragmented and missing BUSCOs. Users can also run BUSCO through the command line and read its output as a data frame using the function *read_busco()*. Finally, the function *plot_busco()* can be used to create publication-ready summary plots for both single-genome and batch modes (Figure 1b,c).

2.3 | Assessing orthogroup inference

We developed a protein domain-aware orthogroup assessment score that aims to maximise the number of shared protein domains within the same orthogroup while minimising the number of different orthogroups containing the same protein domain. The rationale

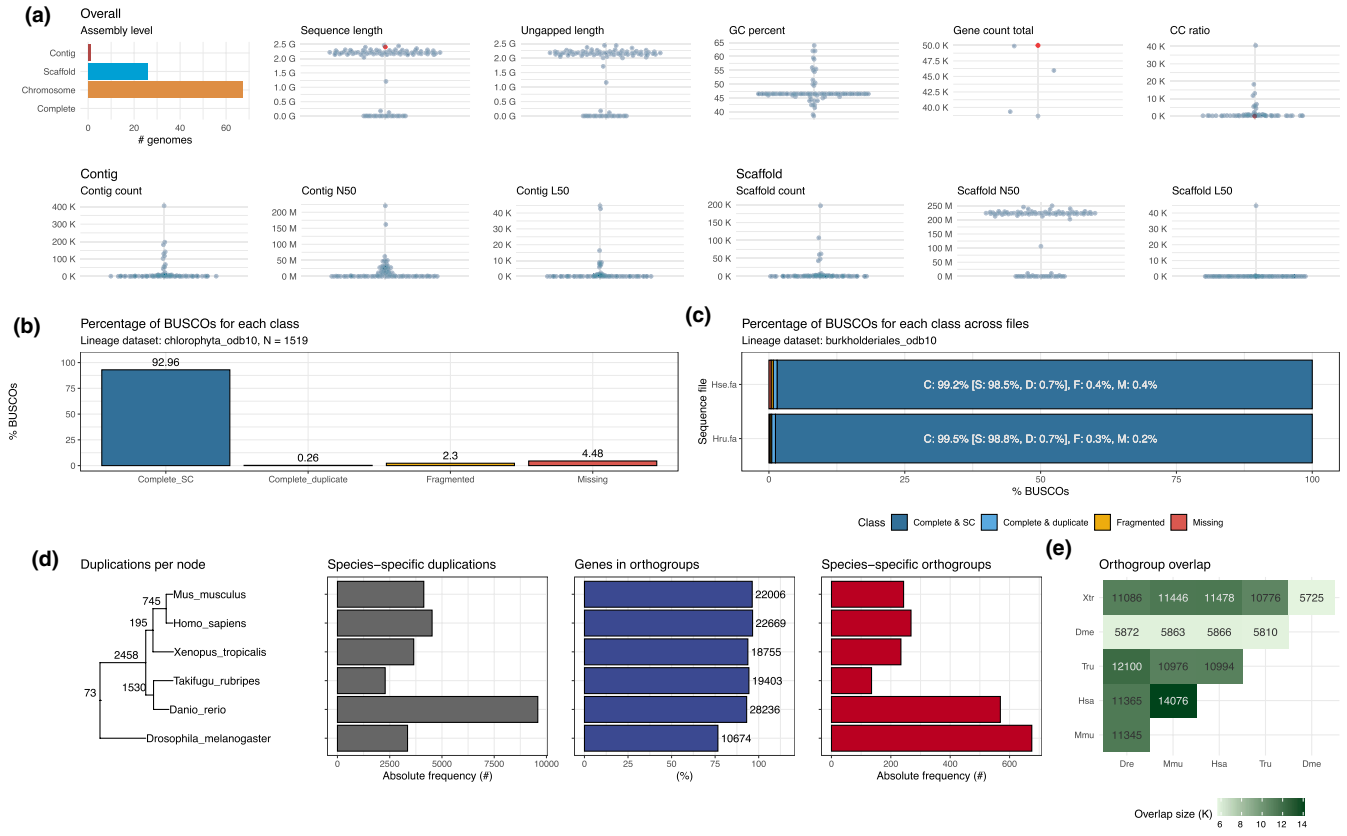

FIGURE 1 Summary of publication-ready plots that can be created with graphical functions in *cogeqc*. (a) Summary assembly and annotation statistics for all *Zea mays* genomes on the NCBI obtained with the function *plot_genome_stats()*. Red data points in the 'Sequence length' and 'Gene count total' panels represent simulated observed values passed by the user. (b) BUSCO summary statistics for a single genome obtained with the function *plot_busco()*. The data used in this figure are a BUSCO output for the *Ostreococcus tauri* genome. (c) BUSCO summary statistics for multiple genomes obtained with the function *plot_busco()*. The data used in this figure are a BUSCO output in batch mode for the bacteria *Herbaspirillum seropedicae* and *Herbaspirillum rubrisubalbicans*. (d) Summary comparative genomics statistics from OrthoFinder obtained with the function *plot_orthofinder_stats()*. (e) Heatmap of orthogroup overlap for pairwise species comparisons as obtained with the function *plot_og_overlap()*. All data used to create the figures are distributed with the package as example datasets.

for such approach is that genes that share the same protein domain are expected to have evolved from a common ancestor, so they should be assigned to the same orthogroup. Formally, orthogroup scores are calculated as:

$$Score_{OG} = Homogeneity - Dispersal.$$

The *homogeneity* term is the mean Sorensen-Dice index (Dice, 1945; Sorenson, 1948) for all pairwise combinations of genes in an orthogroup. The Sorensen-Dice index measures how similar two genes are in terms of the protein domains they have, and it ranges from 0 to 1, with 0 meaning that a gene pair does not share any protein domain, and 1 meaning that it shares all protein domains. Formally,

$$Homogeneity = \frac{1}{N_{pairs}} \sum_{i=1}^{N_{pairs}} SDI_i,$$

$$SDI(A,B) = \frac{2|A \cap B|}{|A| + |B|},$$

where A and B are the set of protein domains associated with genes A and B. Hence, an orthogroup with score 1 would have all genes with the exact same protein domains, while an orthogroup with score 0 would have a different protein domain for each gene. As individual genes in a gene family can lose domains and gain new ones, orthogroup scores can take any value from 0 to 1.

The *dispersal* term aims to correct for overclustering (i.e. orthogroup assignments that break 'true' gene families into an artificially large number of smaller subfamilies), and it describes the relative frequency of dispersed domains (i.e. the same domain in two or more orthogroups). This term penalises orthogroup assignments with the same protein domains in multiple orthogroups. We acknowledge that the presence of the same protein domain in multiple orthogroups can occur due to convergent evolution, but since convergent evolution of protein domains is rare (Gough, 2005), we assume that such patterns indicate overclustering of gene families.

Orthogroups inferred with OrthoFinder (Emms & Kelly, 2019) can be read with the function *read_orthogroups()*, and mean and median scores can be calculated with the function *assess_orthogroups()*. To ensure a higher accuracy in orthogroup assignments, we recommend running OrthoFinder with different

combinations of parameters and comparing the distributions of orthogroup scores in each run to select the best. Alternatively, if reference and reliable orthogroup assignments exist, users can compare their predicted orthogroups with reference orthogroups by using the function *compare_orthogroups()*, which can show the percentage of reference orthogroups that are preserved in predicted orthogroups.

Finally, comparative genomics statistics obtained with OrthoFinder can be read as a list of data frames with the function *read_orthofinder_stats()* and visualised with graphical functions that create publication-ready plots summarising statistics, such as *plot_orthofinder_stats()*, *plot_og_overlap()* and *plot_og_sizes()* (Figure 1d,e).

2.4 | Assessing synteny detection

We propose a network-based assessment of synteny (or collinearity, used here as synonyms) detection that consists in representing synteny relationships as a graph (i.e. a synteny network) and analysing topological properties of the graph to assess its quality. To infer synteny networks for mammalian and angiosperm genomes, (Zhao & Schranz, 2019) have run a synteny detection algorithm with multiple combinations of parameters and selected the best combination based on the clustering coefficient and number of nodes of each network. Ideally, a synteny network should have a large number of nodes (i.e. anchor pairs, duplicated genes retained from a large-scale duplication event) and a high clustering coefficient.

However, there is often a trade-off between the number of nodes and the clustering coefficient, with larger networks being more sparse, and smaller networks being more densely connected. To account for this trade-off, we use the product of the clustering coefficient and the number of nodes to assess networks. Additionally, as synteny networks and biological networks in general tend to be scale-free (i.e. the degree distribution follows a power-law distribution) (Barabási, 2009; Barabasi & Oltvai, 2004; Ravasz et al., 2002; Venancio et al., 2009; Zhao & Schranz, 2019), we added a term to the network score formula that considers how well the network fits a scale-free topology. Formally, the score of a synteny network is calculated as

$$Score = CNF,$$

where C is the network's clustering coefficient, N is the number of nodes, and F is the coefficient of determination (R^2) for the scale-free topology fit.

Synteny networks can be inferred with the R package syntenet (Almeida-Silva et al., 2023), and their scores can be calculated with the function *assess_synnet()*, which returns a data frame with the network's score and the observed values for each term of the formula above (C, N, and F). If users have multiple networks stored in a list, the function *assess_synnet_list()* can calculate scores for multiple networks at once.

3 | RESULTS AND DISCUSSION

3.1 | Assessing the completeness of Chlorophyta genomes

To demonstrate the usage of the functions to assess gene space completeness, we obtained genome sequences for all Chlorophyta genomes on Pico-PLAZA 3.0 ($N=16$) (Van Bel et al., 2022) and calculated their BUSCO scores (Text S1). All genomes were stored in the same directory and the function *run_busco()* was used to run BUSCO in batch mode using the lineage dataset *chlorophyta_odb10*. BUSCO scores were visualised with the function *plot_busco()* (Figure 2). We observed that Chlorophyta genomes on Pico-PLAZA are highly complete, with >90% complete BUSCOs (Figure 2). However, an exception is the algae *Helicosporidium* sp. (Trebouxiophyceae), with only 65.3% complete BUSCOs.

3.2 | Orthogroup assignments in different public databases perform equally well

We used the protein domain-aware orthogroup assessment implemented in *cogeqc* to assess orthogroup assignments in public databases, namely PLAZA Dicots 5.0 (Van Bel et al., 2022), OrthoDB (Kuznetsov et al., 2023), eggNOG (Hernández-Plaza et al., 2023), and HOGENOM (Penel et al., 2009). The rationale for this approach is that domain homogeneity for reference-quality orthogroups should be as high as possible, as indicated by our benchmark with the OrthoBench dataset (Text S2). As the species composition varies across databases, we used *Arabidopsis thaliana* as a representative species to assess orthogroups. For each database, orthogroups were filtered to include only *Arabidopsis* genes, and scores were calculated with the function *calculate_H()* using InterPro domain annotation obtained from PLAZA Dicots 5.0 (Text S3).

We observed that eggNOG orthogroups have lower scores than orthogroups from all other databases (Mann–Whitney U test, $p<0.01$). HOGENOM orthogroup scores are higher than OrthoDB scores, but lower than PLAZA. Finally, PLAZA orthogroup scores are higher than all other databases (Text S3, section 3). However, although differences were significant (Mann–Whitney U test, $p<0.01$), Wilcoxon effect sizes (r) were small, with the difference between eggNOG and HOGENOM being the only one with $r>0.1$ (Figure 3a). The small effect sizes suggest that the observed differences could be due to large sample sizes, as small p-values can be obtained if the sample size is large enough, even when differences are negligible (Sullivan & Feinn, 2012). Thus, despite some small differences, we conclude that all databases perform equally well in their orthogroup assignments.

3.3 | Assessing orthogroup inference under multiple combinations of OrthoFinder parameters

To infer orthogroups, OrthoFinder relies on similarity searches with DIAMOND (Buchfink et al., 2021), followed by normalisation of bit

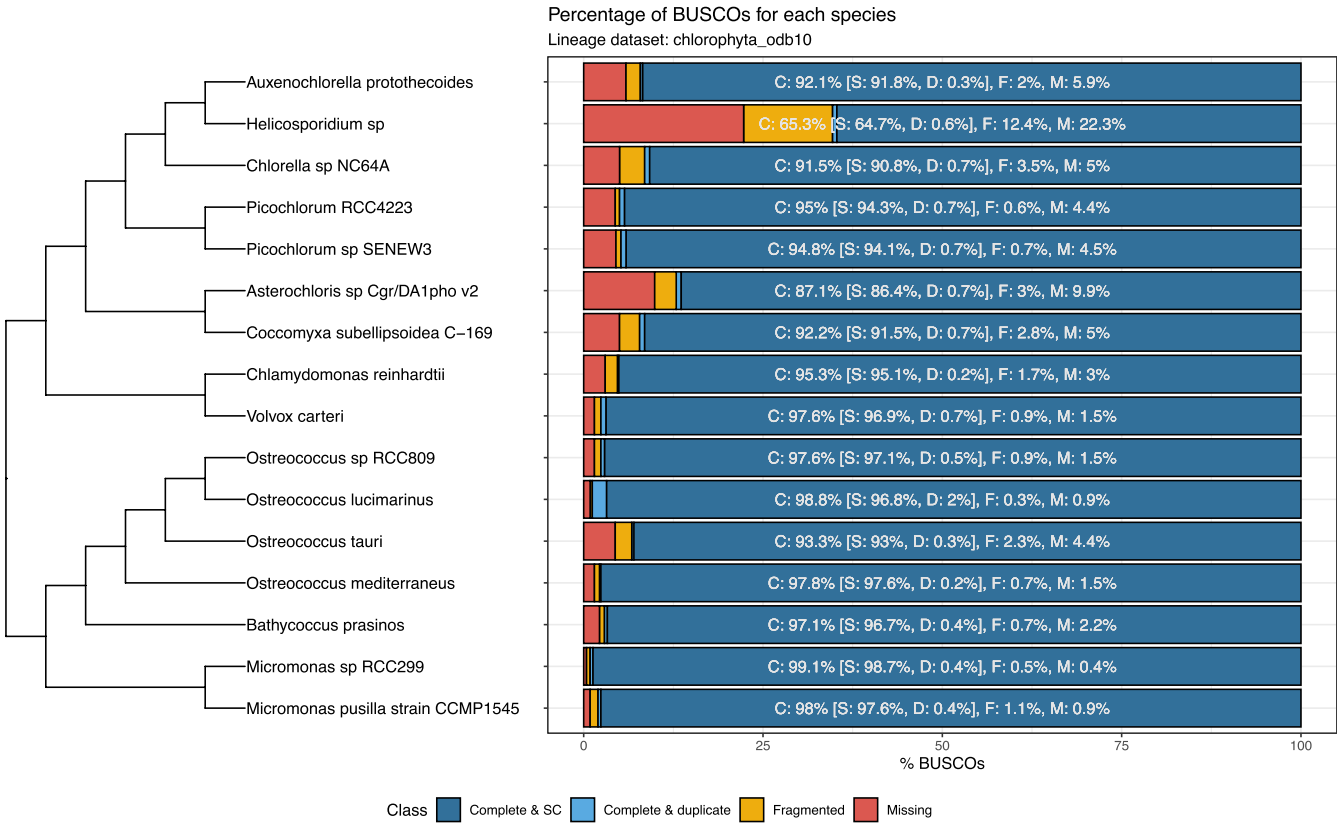

FIGURE 2 Species tree and BUSCO scores for Chlorophyta genomes in Pico-PLAZA 3.0. Scores were calculated using the function *run_busco()* to run BUSCO in batch mode using the chlorophyta_odb10 as lineage dataset. Except for *Helicosporidium* sp., all species have highly complete genomes, as demonstrated by their percentages of complete BUSCOs. Figures were generated by combining the output of the functions *plot_species_tree()* and *plot_busco()*, and code to recreate it is available in Text S1.

scores by gene length, and graph-based clustering using Markov clustering (MCL). By default, OrthoFinder runs DIAMOND in default mode, which is faster, but less accurate than its ultra-sensitive mode. To cluster genes into orthogroups, an MCL inflation parameter of 1.5 is used by default, with lower values resulting in a smaller number of larger clusters (i.e. low granularity), and higher values resulting in a greater number of smaller clusters (i.e. high granularity). To investigate the effect of different parameter combinations on orthogroup homogeneity, we downloaded the proteomes of 25 Brassicaceae species from PLAZA 5.0, Phytozome v13, BRAD and CoGe (Cheng et al., 2011; Goodstein et al., 2012; Lyons et al., 2008; Van Bel et al., 2022), and ran OrthoFinder with eight combinations of parameters by changing the DIAMOND mode (standard vs. ultra-sensitive mode) and the MCL inflation (*mcl* = 1, 1.5, 2, and 3). Orthogroup scores for each OrthoFinder run were obtained with the function *assess_orthogroups()* (Text S4).

A global comparison of the distributions of orthogroup scores shows that using an *mcl* = 1 dramatically reduces homogeneity scores as compared to every other *mcl* value (Mann–Whitney *U* test, $p < 0.01$; Figure 3b). Orthogroup scores for the default OrthoFinder mode (standard DIAMOND, *mcl* = 1.5) are much larger than runs with *mcl* = *1*, both with standard and ultra-sensitive DIAMOND mode (Mann–Whitney *U* test, $p < 0.001$, effect size >0.3; Figure 3b). To test for a possible bias resulting from the species choice, we inspected the distributions of orthogroup scores by OrthoFinder mode for each species separately. We observed that the species choice does not affect scores, as revealed by similar distribution shapes for all species (Figure 3c).

Furthermore, we analysed the effects of changing arguments for each parameter separately (i.e. same DIAMOND mode with different *mcl* values, and vice versa) to understand their individual relevance to orthogroup scores. We observed that increasing *mcl* values leads to significantly higher orthogroup scores, with scores following the order $3 > 2 > 1.5 > 1$, but the difference is only large between *mcl* values of 1 and other values (Mann–Whitney *U* test, $p < 0.001$, $r > 0.3$; Figure 3d). Wilcoxon effect sizes for the comparisons between *mcl* values of 1.5 and above are small ($r < 0.1$; tables 3 and 4 in Text S4), suggesting that significant differences could be due to large sample sizes ($N > 16{,}000$). Likewise, comparisons of orthogroup scores for OrthoFinder runs with different DIAMOND modes revealed significant but negligible differences ($p < 0.05$, $r < 0.04$; Figure 3e; table 5 in Text S4), indicating that changing DIAMOND modes has little impact in orthogroup scores and, hence, should not be a concern in orthogroup detection. This has been suggested by DIAMOND developers in their original manuscript (Buchfink et al., 2021), but we now confirm it with supporting data. Thus, we recommend running OrthoFinder with the standard DIAMOND mode, because it offers a 100-fold increase in speed compared to the ultra-sensitive

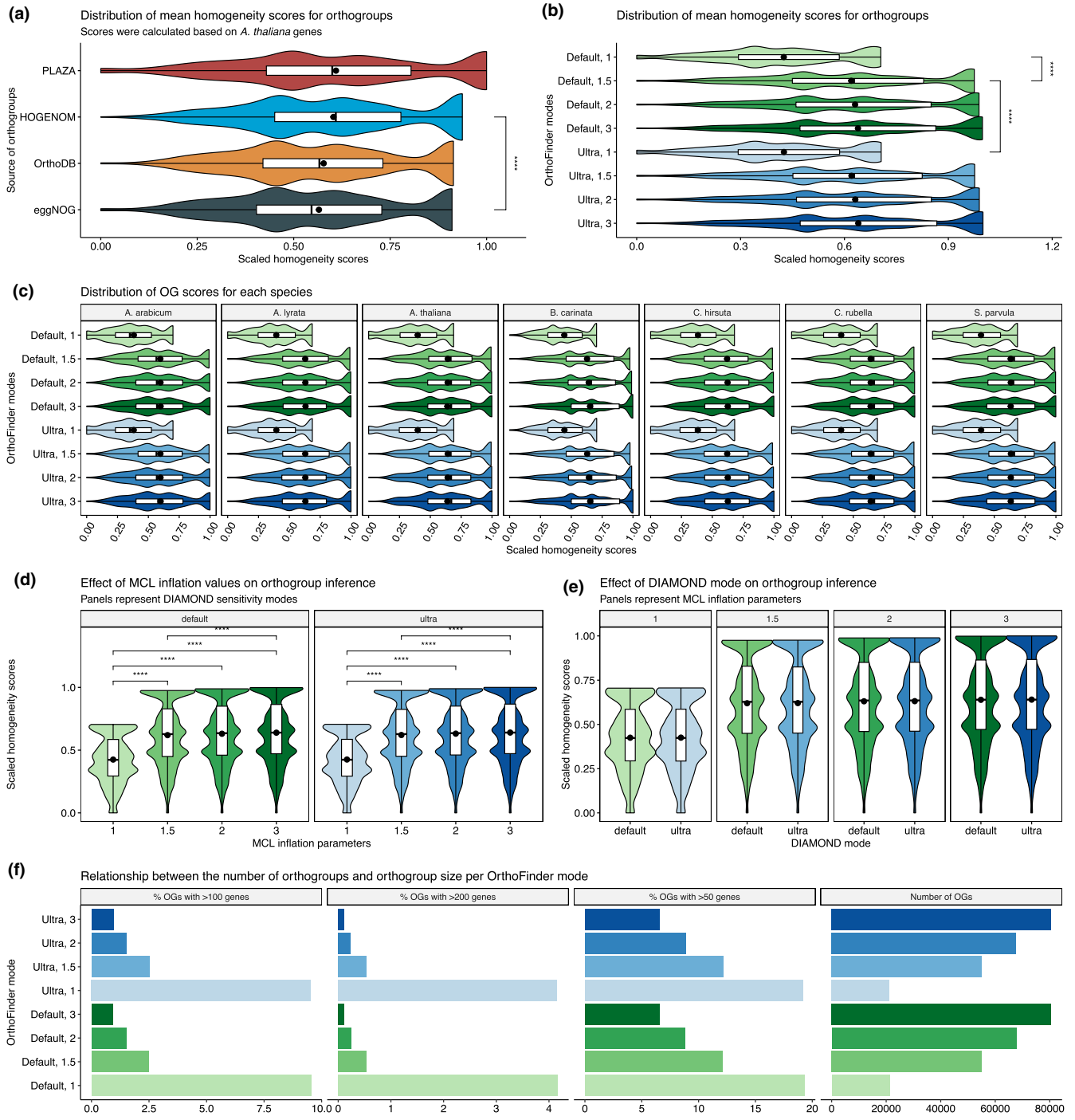

FIGURE 3 Assessment of orthogroup inference in public databases and for a Brassicaceae dataset. (a) Distribution of orthogroup scores for each database. Comparisons with Wilcoxon effect sizes ≥0.1 are highlighted, with asterisks representing significance levels. (b) Distribution of mean orthogroup scores for each OrthoFinder run. Comparisons between the default OrthoFinder mode (standard DIAMOND, *mcl* = 1.5) and other runs with Wilcoxon effect sizes ≥0.1 are highlighted. (c) Distribution of orthogroups scores obtained when considering each species individually. Distributions have the same shape regardless of the species choice. (d) Comparison of the distributions of orthogroup scores for OrthoFinder runs with the same DIAMOND mode, but different *mcl* values. Significant differences ($p < 0.05$, Mann–Whitney *U* test) are highlighted. (e) Comparison of the distributions of orthogroup scores for OrthoFinder runs with the same *mcl* value, but different DIAMOND modes. (f) Bar plots displaying the relationship between orthogroup count and size for each OrthoFinder run. Green and blue bars/distributions represent OrthoFinder runs with the default and ultra-sensitive DIAMOND modes, respectively. $^{*}p \leq 0.05$. $^{**}p \leq 0.01$, $^{***}p \leq 0.001$. $^{****}p \leq 0.0001$.

mode (Buchfink et al., 2021). To validate that orthogroup scores are corrected for overclustering, we ran OrthoFinder with *mcl* = 5 and confirmed that scores are penalised if the granularity is too high (Text S5).

Gene family evolution models often rely on phylogenetic birth-and-death processes, such as CAFE 5 (Mendes et al., 2020), Count (Csuros, 2010), and DeadBird (Zwaenepoel & Van de Peer, 2020). However, large variances in gene copy numbers can result in uninformative parameter estimates, and a common rule of thumb consists in removing orthogroups with ≥100 genes (Mendes et al., 2020). When comparing OrthoFinder runs with *mcl* values of 3 and 1.5, we observed a 2.7-fold increase in the percentage of orthogroups with ≥100 genes for the latter (Figure 3f). Hence, to reduce the number of discarded orthogroups, we recommend using OrthoFinder with *mcl* values of 3. However, since our dataset comprises a single plant family, we advise users to test different parameter combinations for more diverse (e.g. different families and orders) or more restricted (e.g. genus- or population-level) datasets.

3.4 | (Re)assessing the effectiveness of OrthoFinder's bit score normalisation

Bit scores are heavily influenced by gene length, which makes them a biased measure of sequence similarity (Emms & Kelly, 2015). To address this issue, OrthoFinder has introduced a normalisation technique that accounts for gene length, removing the correlation between gene length and bit scores (Emms & Kelly, 2015). To verify the effectiveness of OrthoFinder's normalisation, we performed a Spearman correlation test between orthogroups scores and median gene length and observed a significant but weak negative correlation ($\rho = -0.19$, $p < 0.001$; Figure 4a). Furthermore, we observed that the number of domains in a protein is moderately correlated with its length ($\rho = 0.416$, $p < 0.001$; Text S4), indicating that the number of domains can be a confounder. To address that, we calculated partial Spearman's correlation coefficients between orthogroup scores and median gene length while controlling for the number of domains, and we found a spurious correlation between these variables (Figure 4a; Text S4). Collectively, our findings shows that OrthoFinder's bit score normalisation is effective in reducing biases resulting from gene length.

3.5 | Functional analyses unveil biological processes associated with rapidly and slowly evolving gene families

We observed that all distributions of orthogroup scores produced by different OrthoFinder parameters have a similar shape. Using the default OrthoFinder mode, we were able to divide the distribution

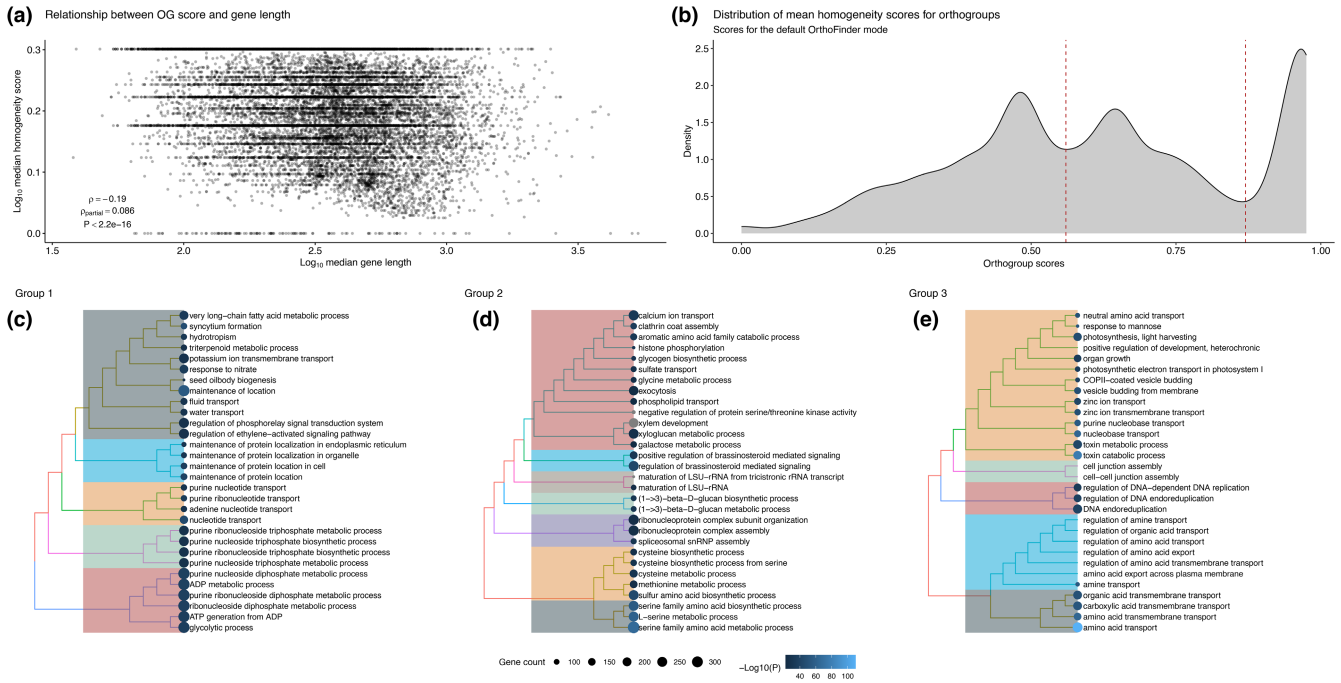

FIGURE 4 Validation of OrthoFinder's bit score normalisation and functional analyses of orthogroups clusters. (a) Relationship between orthogroup scores and orthogroup median length. Spearman's correlation coefficient (ρ), partial Spearman's correlation coefficient ($\rho_{partial}$), and *p*-value are indicated in the bottom left part of the plotting area. (b) Distribution of orthogroup scores for an OrthoFinder run with default parameters. Peaks were used to split the distribution in three clusters, with boundaries indicated by dashed red lines. (c) Tree plot of enriched functional terms for each orthogroup cluster. Enrichment analyses were performed with clusterProfiler (Wu et al., 2021) and visualised with enrichplot. Terms are grouped in a tree-like structure based on semantic similarities calculated with the function *pairwise_termsim()* from the enrichplot package.

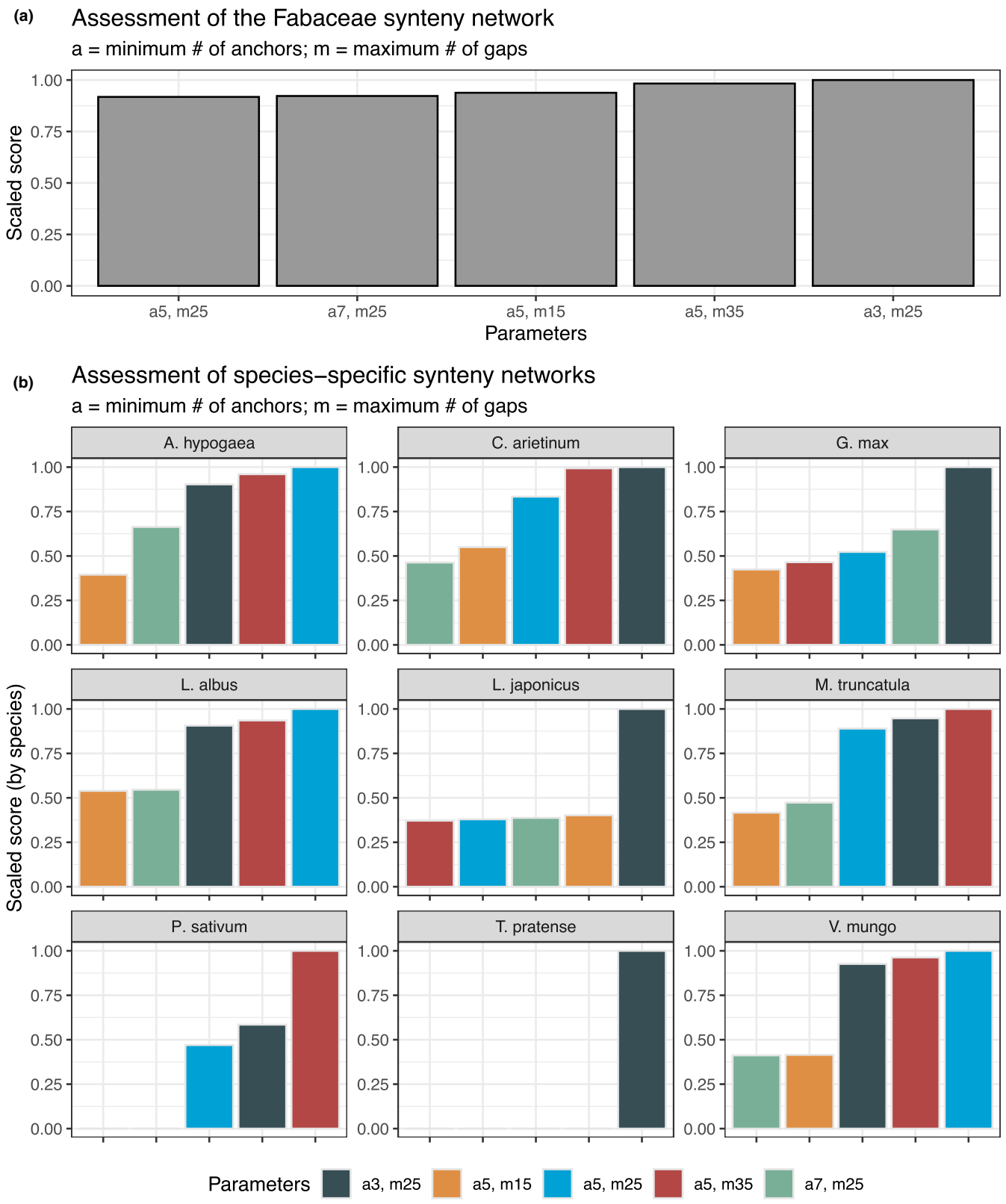

FIGURE 5 Assessment of synteny detection in Fabaceae species with different combinations of parameters. (a) Network scores for a synteny network containing all Fabaceae species. (b) Network scores for species-specific synteny networks. Networks with zero scores are due to clustering coefficients of 0. Colours represent different combinations of parameters.

of orthogroup scores into three clusters based on peaks (Figure 4b). Cluster 1 includes orthogroups with low scores, indicating faster evolutionary rates due to gain and loss of protein domains. Cluster 3 includes orthogroups with high scores, suggesting slower evolutionary rates due to shared protein domains by most or all members. Cluster 2 includes orthogroups with intermediate scores, indicating neither fast nor slow evolutionary rates. To investigate the functional profiles of each cluster, we conducted enrichment analyses of GO terms, MapMan bins, and InterPro domains using all genes in orthogroups as background.

For cluster 1, we found an enrichment of genes associated with ATP production, water and K^+ transport, seed oil body biogenesis, and response to nitrate and ethylene (Figure 4c). Orthogroups in cluster 2 were enriched in genes associated with sulfur amino acid metabolism, spliceosome biogenesis, β-1,3-glucan biosynthesis, response to brassinosteroids, xylem development, exocytosis, and calcium and sulphate transport. Finally, orthogroups in cluster 3 were enriched in genes involved in photosynthesis, zinc and amino acid transport, DNA replication, endocytosis, cell–cell junction assembly, and toxin catabolism. Our results are in line with previous studies, indicating that rapidly evolving families are associated with environmental response while slowly evolving families are associated with more basic cellular processes (Ngou et al., 2022; Wang et al., 2018).

3.6 | Graph-based assessment of synteny detection with different combinations of parameters

To detect synteny between genomic regions, one must explicitly define the minimum number of genes required to call a syntenic block or segment, and the maximum number of allowed gaps between genes. Here, we used the R package syntenet (Almeida-Silva et al., 2023) to infer synteny networks among Fabaceae genomes on PLAZA 5.0 (Van Bel et al., 2022) with 5 combinations of parameters: *a3m25*, *a5m15*, *a5m25*, *a5m35*, and *a7m25*, where 'a' stands for the minimum number of anchors to call synteny, and 'm' stands for the maximum number of gaps between genes (Text S6). We inferred species-specific (i.e. intragenomic) networks and a full network (i.e. all pairwise species comparisons).

For the full network, scaled network scores were very similar, but the parameter combination *a3m25* resulted in the best synteny network, with the largest number of nodes, scale-free topology fit, and overall score (Figure 5a). Interestingly, the network obtained with the default parameter combination of the MCScanX algorithm, *a5m25*, had the lowest score. When analysing each species' network separately, we observed that the best parameter combination depends on the species (Figure 5b), highlighting the importance of performing assessments for each dataset. The combinations *a7m25* and *a5m15* are typically the worst, leading to zero scores in some species due to clustering coefficients of zero, while *a3m25*, *a5m25*, and *a5m35* lead to the best scores in 45%, 33%, and 22% of the species, respectively (Figure 5b). Finally, we observed that the parameter combination that leads to the highest score in most species-specific networks (*a3m25*) is also the best combination for the full network (with all species) (Figure 5b). Although it suggests that this combination tends to be the best in most cases, we advise users to test multiple combinations whenever possible.

4 | CONCLUSION

cogeqc is an R package that can be used to assess the quality of genome assembly and annotation in a phylogenetic context, and to assess the quality of orthogroup inference and synteny detection. The package was designed to be user-friendly, easily integrate with other packages and software tools, and provide summary results as publication-ready plots. Applications to real datasets demonstrated how the package can be used to select optimal parameters in orthogroup inference and synteny analyses, providing users with general guidelines.

AUTHOR CONTRIBUTIONS

Fabricio Almeida-Silva and Yves Van de Peer conceived the ideas and designed methodology. Fabricio Almeida-Silva collected the data. Fabricio Almeida-Silva and Yves Van de Peer analysed the data. Fabricio Almeida-Silva and Yves Van de Peer led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Yves Van de Peer acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (no. 833522). Yves Van de Peer and Fabricio Almeida-Silva acknowledge funding from Ghent University (Methusalem funding, BOF.MET.2021.0005.01).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14243.

DATA AVAILABILITY STATEMENT

To ensure full reproducibility, all code and data used in this manuscript are available in a GitHub repository at https://github.com/almeidasilvaf/cogeqc_paper, and the code used in benchmarks are available in Supplementary Texts. A Zenodo archive of all data and code is also available (Almeida-Silva, 2023).

ORCID

Fabricio Almeida-Silva https://orcid.org/0000-0002-5314-2964
Yves Van de Peer https://orcid.org/0000-0003-4327-3730

REFERENCES

Almeida-Silva, F. (2023). almeidasilvaf/cogeqc_paper: Published paper (v1.0.0). *Zenodo*, https://doi.org/10.5281/zenodo.10013723

Almeida-Silva, F., Zhao, T., Ullrich, K. K., Schranz, M. E., & Van de Peer, Y. (2023). syntenet: An R/Bioconductor package for the inference and analysis of synteny networks. *Bioinformatics*, *39*(1), btac806.

An, D., Zhou, Y., Li, C., Xiao, Q., Wang, T., Zhang, Y., Wu, Y., Li, Y., Chao, D.-Y., & Messing, J. (2019). Plant evolution and environmental adaptation unveiled by long-read whole-genome sequencing of Spirodela. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(38), 18893–18899.

Barabási, A.-L. (2009). Scale-free networks: A decade and beyond. *Science*, *325*(5939), 412–413.

Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113.

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368.

Cheng, F., Liu, S., Wu, J., Fang, L., Sun, S., Liu, B., Li, P., Hua, W., & Wang, X. (2011). BRAD, the genetics and genomics database for Brassica plants. *BMC Plant Biology*, *11*, 1–6.

Csuros, M. (2010). Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, *26*(15), 1910–1912.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Drost, H.-G., & Paszkowski, J. (2017). Biomartr: Genomic data retrieval with R. *Bioinformatics*, *33*(8), 1216–1217.

Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 1–14.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 1–14.

Feron, R., & Waterhouse, R. M. (2022). Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes. *GigaScience*, *11*, giac006. https://doi.org/10.1093/gigascience/giac006

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., & Putnam, N. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186.

Gough, J. (2005). Convergent evolution of domain architectures (is rare). *Bioinformatics*, *21*(8), 1464–1471.

Hernández-Plaza, A., Szklarczyk, D., Botas, J., Cantalapiedra, C. P., Giner-Lamia, J., Mende, D. R., Kirsch, R., Rattei, T., Letunic, I., & Jensen, L. J. (2023). eggNOG 6.0: Enabling comparative genomics across 12 535 organisms. *Nucleic Acids Research*, *51*(D1), D389–D394.

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., & Zdobnov, E. M. (2023). OrthoDB v11: Annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, *51*(D1), D445–D451.

Liu, D., Hunt, M., & Tsai, I. J. (2018). Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinformatics*, *19*(1), 1–13.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., & Lisch, D. (2008). Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiology*, *148*(4), 1772–1781.

Marks, R. A., Hotaling, S., Frandsen, P. B., & VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nature Plants*, *7*(12), 1571–1578.

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*, *36*(22–23), 5516–5518.

Michael, T. P. (2014). Plant genome size variation: Bloating and purging DNA. *Briefings in Functional Genomics*, *13*(4), 308–317.

Ngou, B. P. M., Heal, R., Wyler, M., Schmid, M. W., & Jones, J. D. (2022). Concerted expansion and contraction of immune receptor gene repertoires in plant genomes. *Nature Plants*, *8*(10), 1146–1152.

Palfalvi, G., Hackl, T., Terhoeven, N., Shibata, T. F., Nishiyama, T., Ankenbrand, M., Becker, D., Förster, F., Freund, M., & Iosip, A. (2020). Genomes of the venus flytrap and close relatives unveil the roots of plant carnivory. *Current Biology*, *30*(12), 2312–2320.

Penel, S., Arigon, A.-M., Dufayard, J.-F., Sertier, A.-S., Daubin, V., Duret, L., Gouy, M., & Perrière, G. (2009). Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, *10*(6), 1–13.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, *297*(5586), 1551–1555.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Sorenson, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analysis of vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskabs Skrifter*, *5*, 1–5.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—Or why the P value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282.

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., & Vandepoele, K. (2022). PLAZA 5.0: Extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Research*, *50*(D1), 1468–1474.

Venancio, T. M., Balaji, S., Iyer, L. M., & Aravind, L. (2009). Reconstructing the ubiquitin network-cross-talk with other systems and identification of novel functions. *Genome Biology*, *10*, 1–18.

Wang, P., Moore, B. M., Panchy, N. L., Meng, F., Lehti-Shiu, M. D., & Shiu, S.-H. (2018). Factors influencing gene family size variation among related species in a plant family, *Solanaceae*. *Genome Biology and Evolution*, *10*(10), 2596–2613.

Wang, P., & Wang, F. (2022). A proposed metric set for evaluation of genome assembly quality. *Trends in Genetics*, *39*, 175–186.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., & Zhan, L. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovations*, *2*(3), 100141.

Xu, Y., Lei, Y., Su, Z., Zhao, M., Zhang, J., Shen, G., Wang, L., Li, J., Qi, J., & Wu, J. (2021). A chromosome-scale Gastrodia elata genome and large-scale comparative genomic analysis indicate convergent evolution by gene loss in mycoheterotrophic and parasitic plants. *The Plant Journal*, *108*(6), 1609–1623.

Zhao, T., & Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(6), 2165–2174. https://doi.org/10.1073/pnas.1801757116

Zwaenepoel, A., & Van de Peer, Y. (2020). Model-based detection of whole-genome duplications in a phylogeny. *Molecular Biology and Evolution*, *37*(9), 2734–2746.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Text S1. Assessing the completeness of Chlorophyta genomes.
Text S2. Assessing Orthobench orthogroups.
Text S3. Assessing orthogroup inference in public databases.
Text S4. Assessing orthogroup inference for Brassicaceae genomes.
Text S5. On overclustering correction.
Text S6. Assessing synteny detection in Fabaceae.

How to cite this article: Almeida-Silva, F., & Van de Peer, Y. (2023). Assessing the quality of comparative genomics data and results with the *cogeqc* R/Bioconductor package. *Methods in Ecology and Evolution*, *14*, 2942–2952. https://doi.org/10.1111/2041-210X.14243