BMC Bioinformatics

SOFTWARE

Open Access

CompàreGenome: a command-line tool for genomic diversity estimation in prokaryotes and eukaryotes

Gabriele Moro[1*], Rossano Atzeni[2*], Ali Al-Subhi[3] and Maria Giovanna Marche[1]

*Correspondence:
gmoro@bioecopest.com

[1] Technology Park of Sardinia, Bioecopest Srl, SP 55 Km 8.400, Tramariglio, Alghero, SS, Italy
[2] Center for Advanced Studies, Research and Development in Sardinia, CRS4 Srl, Pula, CA, Italy
[3] Plant Sciences Department, College of Agricultural and Marine Sciences, Sultanate Qaboos University, Al-Khoud, Sultanate of Oman

Abstract

Background: The increasing availability of sequenced genomes has enabled comparative analyses of various organisms. Numerous tools and online platforms have been developed for this purpose, facilitating the identification of unique features within selected organisms. However, choosing the most appropriate tools can be unclear during the initial stages of analysis, often requiring multiple attempts to match the specific characteristics of the data. Here, we introduce CompàreGenome, a command-line tool specifically designed for genomic diversity estimation analyses. Suitable for both prokaryotes and eukaryotes, this tool is particularly valuable in the early stages of studies when little information is available about the genetic differences or similarities among compared organisms.

Results: In all the tests conducted, CompàreGenome successfully identified specific genetic features of the selected organisms, detected the most conserved genes, pinpointed highly divergent ones, and functionally annotated these genes. This provided insights into biological processes, molecular functions, and cellular components associated with each gene. The tool also distinguished organisms at the strain level and quantified genetic distances using three distinct analytical methods.

Conclusion: CompàreGenome empowers users to explore genomic differences among organisms, translating technical outputs from various tools into actionable insights for biologists. While primarily tested on small microbial genomes, the tool has potential applications for larger genomes. CompàreGenome is implemented in Bash, R, and Python and is freely available under an LGPL-2.1 license.

Keywords: Comparative genomics, Bioinformatics tool, Functional annotation, Microorganisms

Background

Comparative genomic analyses enable the comparison of different organisms using whole-genome sequences. These analyses address a wide range of biological questions, driving the development of numerous command-line tools and online platforms [1]. Among these tools, many focus on identifying genetic variations, such as single nucleotide polymorphisms (SNPs) and insertions/deletions (e.g., VarScan, GATK, VarDict)

[2–4], using genetic variability to estimate similarity and evolutionary relationships between genomic regions (e.g., HSDecipher, eggNOG) [5, 6]. Tools like ACT and Artemis [7–9] visualize BLAST-based genome comparisons effectively for a few genomes, while Mauve [10] provides detailed insights into conserved genomic regions, aiding evolutionary studies. BRIG [11], available both as a command-line tool and online platform, offers circular visualization of prokaryotic genomes, providing an immediate overview of differences across species.

Using such tools, scientists can compare genomes to identify organism-specific features and highlight relevant differences. For example, Li et al. identified the genetic basis of virulence in *Beauveria bassiana*, an entomopathogenic fungus with strains of varying pathogenicity [12]. Similarly, Ann et al. revealed the strain-specific anti-inflammatory effects of *Limosilactobacillus fermentum* SMFM2017-NK2 [13]. These studies relied on multiple genomic comparison tools, underscoring the challenges of selecting and combining appropriate tools, particularly at the initial stages of analysis.

To address this challenge, we developed CompàreGenome, a command-line tool designed for comparing whole genomes directly. Its primary objective is to estimate genomic diversity among organisms through gene-to-gene comparisons, including identifying and annotating conserved and divergent genes, analysing their biological impact, and quantifying genetic distances. By distinguishing organisms at the species level and beyond, CompàreGenome supports intra- and interspecific comparative studies and aids in identifying strains of the same species or subgroups within populations. It is also suitable for evolutionary studies, providing preliminary results that can be further explored using more specialized tools. Designed for Unix-like systems, CompàreGenome leverages existing tools for both installation and analysis.

Implementation

Pipeline structure and output description

CompàreGenome requires a reference genome and at least two query genomes as input. The reference genome, provided in GENBANK format, serves as the basis for identifying homologous gene sequences in the query genomes. Selection of the reference genome is critical: for intraspecific comparisons, it should belong to the same species, while for interspecific analyses, the most taxonomically related species is recommended. Repeating the analysis with different reference genomes and merging the results is also a good practice. Query genomes, in FASTA format, should be assembled to at least at contig level, with overly short contigs or scaffolds removed.

Typically, contigs or scaffolds shorter than 500 bp are excluded, but other metrics like N50 may guide this decision [14, 15]. The CompàreGenome pipeline uses external tools for sequence alignment, genetic distance quantification, and graphical result visualization (Table 1). External tool dependencies are managed via Conda environments, ensuring reproducibility by isolating tool versions and dependencies, ensuring a consistent and isolated execution environment, preventing dependency conflicts and facilitating reproducibility. Starting from reference gene sequences and query genome assemblies, CompàreGenome identifies homologous genes and performs pairwise comparisons using BLASTN [16]. Each alignment generates a similarity score (scale: 0–100), estimating genetic similarity among gene sequences. These scores group genes

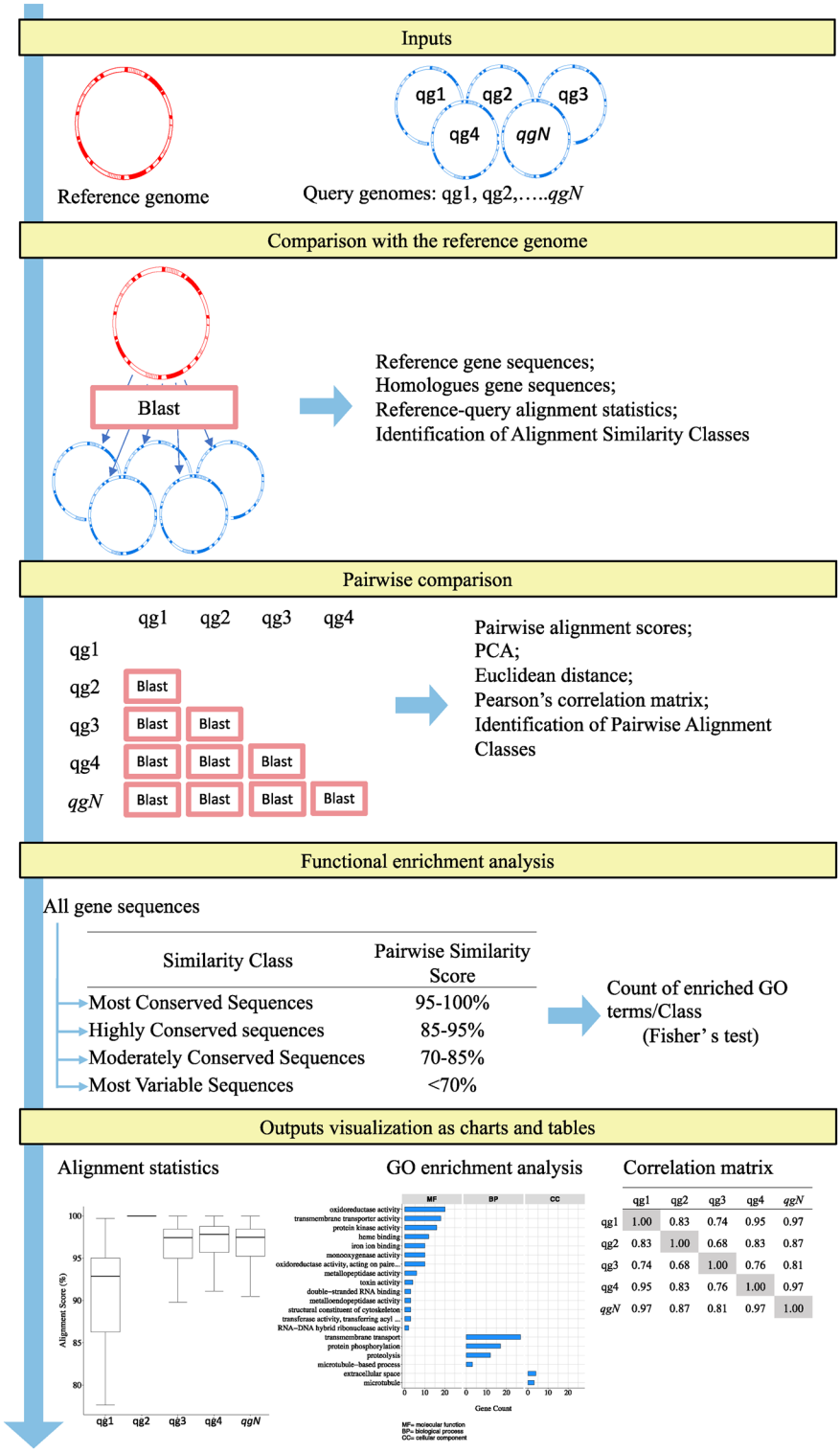
Table 1 External tools utilized by CompàreGenome

Tool	Version	Application
Anaconda	4.14.0 [35]	Tool's installation
Biopython	1.82 [19]	Processing input fasta and gbk files
BLAST +	2.15.0 [16]	Gene sequence alignment
R packages		
factoextra	1.0.7 [22]	PCA analysis
ggtree	3.10.0 [23]	Visualization of Euclidean distance analysis
agricolae	1.3_7 [36]	General file editing
SeqinR	4.2_36 [20]	Fasta files editing
gplots	3.1.3 [21]	Data visualization
phangorn	2.11.1 [24]	Visualization of Euclidean distance analysis
ggrepel	0.9.4 [37]	Visualization of PCA analysis data
cowplot	1.1.2 [25]	Visualization of PCA analysis data
GO.db	3.12.1 [26]	GO enrichment analysis

into four similarity classes (95–100%, 85–95%, 70–85%, and < 70%) to distinguish highly conserved, moderately conserved, and highly variable genes. A Gene Ontology (GO)-based enrichment analysis [17, 18] is performed, providing functional annotation and revealing the biological implications of genetic variability. Similarity scores also quantify genetic distances between query genomes, enabling users to characterize relationships among the organisms.

The pipeline follows these steps (Fig. 1):

1. Reference gene sequence retrieval. Using the Biopython [19] package, the pipeline extracts nucleotide sequences, sequence lengths, product names, and associated GO terms (if available) for each gene in the reference genome. Genes without annotations are labelled as "Unclassified."
2. Homologous gene identification. The extracted reference sequences are used to identify homologous sequences within query genomes through BLAST + blastn [16]. Alignments are scored using the Reference Similarity Score (RSS), calculated from Identity (percentage of matching base pairs) and Coverage (percentage of overlap with the reference sequence).
3. Reference genome comparisons. RSS values are used to group reference sequences into four Reference Similarity Classes (RSCs): 95–100%; 85–95%; 70–85%; < 70%. Data distribution and RSC values are evaluated with statistical summaries (e.g., medians), leveraging SeqinR [20] and gplots [21] for analysis.
4. Pairwise comparisons between query genomes. Homologous sequences undergo alignment across all query genomes. Alignment quality is summarized as a Pairwise Similarity Score (PSS), calculated as the average and standard deviation of alignment scores (Identity and Coverage). These PSS values quantify genetic distances through methods like PCA and Euclidean distance. This step employs R packages: factoextra [22], ggtree [23], phangorn [24], and cowplot [25].
5. GO-based enrichment analysis. Based on PSS values, gene sequences are grouped into four Pairwise Similarity Classes (PSC), named as Most Conserved Sequences, Highly Conserved sequences, Moderately Conserved Sequences and Most Vari-



	qg1	qg2	qg3	qg4	qgN
qg1	1.00	0.83	0.74	0.95	0.97
qg2	0.83	1.00	0.68	0.83	0.87
qg3	0.74	0.68	1.00	0.76	0.81
qg4	0.95	0.83	0.76	1.00	0.97
qgN	0.97	0.87	0.81	0.97	1.00

Fig. 1 Pipeline structure overview. The pipeline compares one reference genome and 2 to N query genomes, starting with reference vs. query and query vs. query comparisons. Following this step, Gene Ontology (GO) enrichment analysis is performed on the defined similarity classes. The results are presented as tables and visualisations, offering a comprehensive comparison

able Sequences, including sequences with PSS equal to 95–100%, 85–95%, 70–85% and < 70%, respectively. GO enrichment analysis is performed for each PSC, exploring functional annotations for Molecular Function, Biological Process, and Cellular Component categories using GO.db [26].

Results

Case study: comparison of different strains of *Beauveria bassiana*

CompàreGenome was tested on several species, both prokaryotes and eukaryotes, and the results confirmed its sensitivity and accuracy in performing genetic comparisons. In this study, we present a validation test on the entomopathogenic fungus *Beauveria bassiana*. CompàreGenome was used to estimate genetic differences among several known strains of this species, as well as a related species, *Beauveria brongniartii* (Table 2).

Using publicly available whole-genome sequences and newly sequenced genomes, CompàreGenome effectively distinguished between the different strains of *B. bassiana* and *B. brongniartii* (Fig. 2). Moreover, it successfully identified the genes responsible for these differences.

All of *B. bassiana* showed slight but measurable differences from the reference genome (Fig. 2a). Approximately 80% of the gene sequences were highly conserved within the *B. bassiana* genome, with similarity scores ranging from 95 to 100%. The remaining gene sequences exhibited varying degrees of similarity, predominantly falling into the 85–95% and < 70% groups, with only a small proportion included in the 70–85% group.

Principal Component Analysis (PCA) and Euclidean distance analysis revealed that the genomes of the *B. bassiana* strains were highly correlated, as expected for closely related organisms (Fig. 2c-d). Furthermore, consistently with the source material, CompàreGenome identified Bb_ATCC74040s1, Bb_ATCC74040s2, and Bb_ARSEF3097 [27] as belonging to the same strain, demonstrating the tool's sensitivity and accuracy (Supplementary Tables S5 and S6). This is notable since all three genome sequences were derived from the same strain (*B. bassiana* ATCC74040) but were sequenced at different times using different methods (see Methods for Bb_ATCC74040s1 and Bb_ATCC74040s2, and reference [27] for Bb_ARSEF3097).

Despite the differences in sequencing methods and times, CompàreGenome successfully identified the samples Bb_ATCC74040s1, Bb_ATCC74040s2, and

Table 2 List and references of the genome assemblies analysed in the case study

Label	Species	Accession number
Bbrongniartii	*B. brongniartii*	PRJNA879330
Bb_ARSEF2860	*B. bassiana*	PRJNA38719
Bb_ARSEF8028	*B. bassiana*	PRJNA260878
Bb_D1_5	*B. bassiana*	PRJNA178080
Bb_JEF_007	*B. bassiana*	PRJNA352877
Bb_ARSEF3097	*B. bassiana*	PRJNA624104
Bb_ATCC74040s1	*B. bassiana*	PRJNA1203612
Bb_ATCC74040s2	*B. bassiana*	PRJNA1203920
Reference genome	*B. bassiana*	PRJNA38719

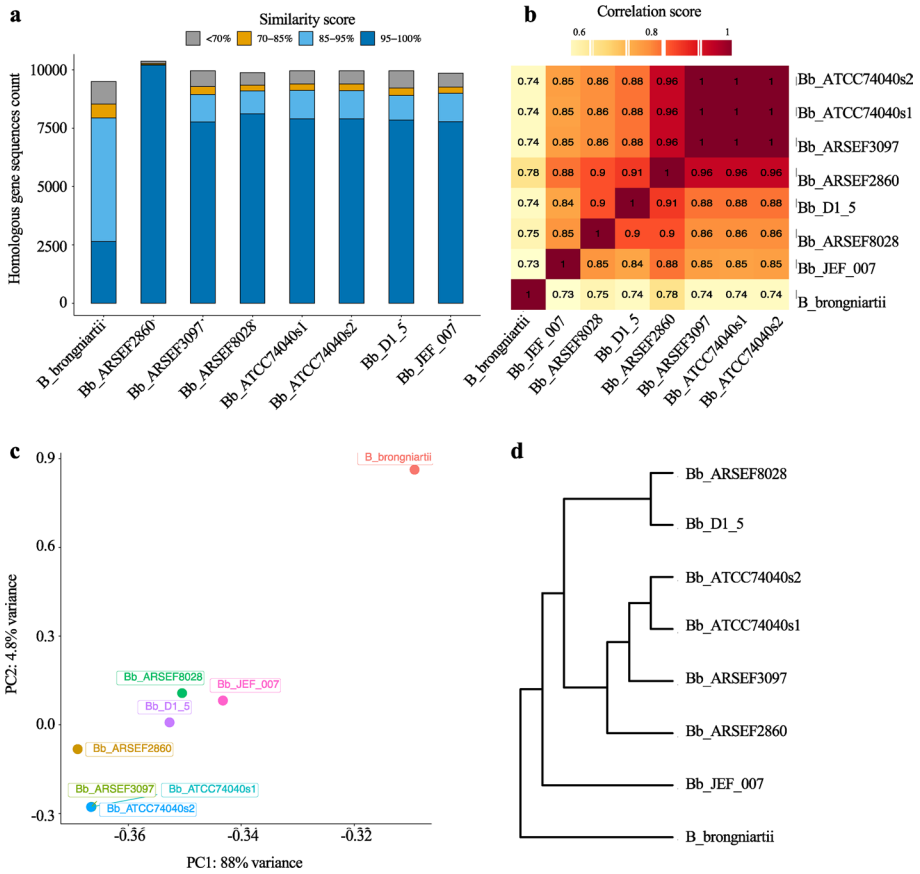


Fig. 2 Comparative analysis on *Beauveria* spp. Distribution of homologous gene sequences within 4 levels of similarity score, resulting by comparison of the 8 query genomes with the reference genome (**a**); Pearson's correlation matrix based on the mean gene similarity scores (**b**); Principal component analysis based on gene pairwise similarity scores (**c**); tree representing Euclidean distance calculated on gene pairwise similarity scores (**d**)

Bb_ARSEF3097 as identical strains, grouping them consistently across all analyses. These samples were assigned to the same position in the PCA plot (overlapping dots in Fig. 2c), gave a perfect correlation score of 1 (Fig. 2b), were placed on the same branch in the Euclidean distance tree (Fig. 2d), and reported a sequence similarity greater than 98% in 10,317 of the 10,364 *B. bassiana* genes (Supplementary Table S6).

Although all the strains analysed belong to the same species, the correlation levels between ARSEF2860, ARSEF8028, D1_5, and JEF_007 were lower compared to the Bb_ATCC74040 samples, with correlation scores ranging from 0.84 to 1. The highest correlation was observed among the three ATCC74040 samples (Bb_ATCC74040s1, Bb_ATCC74040s2, and Bb_ARSEF3097), while the lowest correlation was associated with the Bb_JEF_007 strain (Fig. 2b). In contrast, the species *B. brongniartii* was consistently identified as significantly different from *B. bassiana* strains across all analyses. Only around 25% of the gene sequences were highly conserved between *B. bassiana* and *B. brongniartii*, while over 50% exhibited similarity scores in the 85–95% range (Fig. 2a). PCA, Euclidean distance, and Pearson correlation analyses all indicated that *B. brongniartii* formed a distinct group, separate from the *B. bassiana* strains.

Additionally, CompàreGenome identified over- and under-represented GO terms, providing information on the biological relevance of genes within each similarity class (Table 3). Among the 75 enriched GO categories, the majority (34 terms) were associated with the "Most Variable" class, followed by "Highly Conserved" (21 terms) and "Moderately Conserved" (20 terms), while no enriched terms were found for the "Most Conserved" class. Notably, the enriched categories "toxin activity" and "mycotoxin biosynthetic process" deserve special attention, as they may help explain variations in insecticidal activity reported within *B. bassiana* strains [38, 39].

Discussion

Comparative genomics is a fundamental area of biological research, enabling the identification of relevant genes and uncovering evolutionary relationships among species [28]. Various tools facilitate direct whole-genome comparisons, offering detailed insights into the species under investigation.

CompàreGenome is a command-line tool tailored for comparative genomic analysis, especially beneficial during the initial stages. It requires only basic proficiency in bash scripting and bioinformatics. Results are presented as intuitive tables and plots, allowing for straightforward interpretation. Unlike other tools, CompàreGenome translates sequence alignment data into genomic distances and biologically meaningful insights through Gene Ontology (GO) annotation. Genes are categorized into four similarity classes based on similarity scores (95–100%, 85–95%, 70–85%, and < 70%), helping users identify genes of interest tailored to the study's focus. For instance, the 95–100% class is crucial for intra-species analyses in prokaryotes, while the 70–85% and < 70% classes are more relevant for inter-species comparisons [29, 30]. In contrast, orthologous sequences in the human genome generally exhibit minimum similarity levels of 70–80% [31].

CompàreGenome outputs include analysis metrics and robust statistical summaries of alignment scores, aiding in the identification of genes with potential biological relevance. It is versatile, applicable to prokaryotic and eukaryotic genomes, to identify genetic similarities, differences, and associated biological effects.

The tool, however, is primarily gene-focused. In prokaryotes, where genic regions dominate the genome, CompàreGenome delivers comprehensive comparisons. For eukaryotes, where genic regions can constitute less than 50% of the genome, regulatory elements and non-coding regions might not be included in the analysis.

CompàreGenome is efficient and scalable. On a dual core i5 processor with 8 GB of RAM, installation takes approximately 30 min, while analysing three fungal genomes (~ 35 Mb) requires 1 h and 18 min. On an 8-core i7 processor with 40 GB of RAM, installation takes 14 min, and analysis time drops to 46 min (Supplementary Table S3). Genome size significantly impacts runtime. For instance, analysing three *Arabidopsis thaliana* genomes (~ 135 Mb) takes 126 min, while analysing three *Bacillus cereus* genomes (~ 6.3 Mb) takes only 9 min. Processing nine genomes increases runtime slightly, to 141 min for *A. thaliana* and 10 min for *B. cereus* (Supplementary Table S4).

Although the tool provides substantial information, further improvements are needed to optimize performance and address genome features excluded in the current version, such as regulatory and non-coding regions (Supplementary Table S1).

Table 3 GO enrichment analysis results

Similarity Class	Pairwise Similarity Score	Enriched GO category (Top 10 most significant)	Observed gene count	Expected gene count	Log2 fold change	Fisher's P value
Most Conserved Sequences (n = 0)	95—100%	–	–	–	–	–
Highly Conserved sequences (*n* = 21)	85% to < 95%	Protein-arginine deiminase activity	7	1.04	2.753	4.64E-04
		Mycotoxin biosynthetic process	10	2.45	2.027	6.71E-04
		Toxin activity	8	1.98	2.013	2.43E-03
		Serine-type peptidase activity	14	5.38	1.380	2.52E-03
		NADP binding	9	3.11	1.531	8.26E-03
		Cholestenol delta-isomerase activity	3	0.28	3.405	1.04E-02
		Sterol metabolic process	3	0.28	3.405	1.04E-02
		N,N-dimethylaniline monooxygenase activity	6	1.60	1.903	1.13E-02
		Proteolysis	28	16.61	0.753	1.15E-02
		Nitrogen compound metabolic process	7	2.17	1.689	1.20E-02
Moderately Conserved Sequences (*n* = 20)	70% to < 85%	Protein phosphorylation	17	4.97	1.773	1.83E-05
		Protein kinase activity	16	5.03	1.668	7.34E-05
		Heme binding	12	3.32	1.856	2.03E-04
		Monooxygenase activity	10	2.47	2.016	3.10E-04
		Oxidoreductase activity, acting on paired				
		Donors, with incorporation or reduction of	10	2.59	1.948	4.39E-04
		Molecular oxygen				
		Extracellular space	4	0.42	3.245	1.55E-03
		Double-stranded RNA binding	3	0.18	4.052	1.80E-03
		Metallopeptidase activity	6	1.24	2.279	2.24E-03
		Iron ion binding	10	3.47	1.528	3.35E-03
		Structural constituent of cytoskeleton	3	0.24	3.637	3.39E-03

Table 3 (continued)

Similarity Class	Pairwise Similarity Score	Enriched GO category (Top 10 most significant)	Observed gene count	Expected gene count	Log2 fold change	Fisher's P value
Most Variable Sequences ($n=34$)	< 70%	Protein dimerization activity	20	1.76	3.510	1.07E-13
		Serine-type endopeptidase activity	17	3.13	2.444	9.14E-08
		Proteolysis	22	7.53	1.546	1.57E-05
		Nucleoside metabolic process	9	1.28	2.809	2.00E-05

All the gene sequences were first grouped into 4 similarity classes according to the sequence similarity within the query genomes. Enrichment was calculated by comparison of the expected vs. observed gene count for each GO term ($P < 0.05$, Fisher's test). Shown the top 10 most enriched categories (full list available in the supplementary file)

Conclusion

CompàreGenome is a versatile command-line pipeline implemented in bash, R, and Python, designed for comparative genomic analysis of both prokaryotes and eukaryotes. Originally developed for bacterial strain comparisons, it is suitable for various microbial species and can handle larger genomes with sufficient computational resources or by analysing chromosomes separately.

The pipeline requires basic bash scripting knowledge and effectively translates gene sequence alignment data into actionable biological insights, making it a valuable tool for researchers in life sciences.

Methods

B. bassiana isolation and DNA extraction

The samples Bb_ATCC74040s1 and Bb_ATCC74040s2 were isolated from the commercial product Naturalis (CBC [Europe] Srl) and cultured on potato dextrose agar at 28 °C. Genomic DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen) following the manufacturer's instructions and further purified with ethanol precipitation.

Genome sequencing and assembly

The Bb_ATCC74040s1 and Bb_ATCC74040s2 genomes were sequenced in April 2022 and April 2023, respectively, using Illumina NovaSeq 6000 (2×150 bp paired-end mode) at Eurofins Genomics. Raw reads were evaluated with FastQC v0.12.0 [32], trimmed using Trim Galore v0.6.5, and assembled de novo with SPAdes v3.15.5 at the scaffold level. Assembly quality was assessed with QUAST v5.0.2 [34].

Analysis with CompàreGenome

The analysis included eight genome assemblies in FASTA format: six publicly available assemblies (*B. bassiana* ARSEF8028, D1_5, JEF_007, ARSEF3097, ARSEF2860, and *B. brongniartii*), and two newly sequenced *B. bassiana* assemblies (Bb_ATCC74040s1

and Bb_ATCC74040s2). ARSEF8028, provided in GenBank format, served as the reference genome.

Default pipeline settings were used. Similarity scores were calculated using the formula:

$$Similarity\ score = \begin{cases} Coverage \times Identity(\%) & if\ \ Coverage \leq 100 \\ Coverage - \big(Coverage - 100\big) \times Identity(\%) & if\ \ Coverage > 100 \end{cases}$$

Similarity scores informed genetic distance calculations via PCA, Euclidean distance, and correlation matrix analyses. GO enrichment analysis was performed for genes within each similarity class.

Availability and requirements

Project name: CompàreGenome.

Project home page: https://bioecopest.com/comparegenome

Operating system(s): MacOS, Linux.

Programming language: Bash, R, Python.

Other requirements: Anaconda 4.14.0 or higher.

Licence: LGPL-2.1

Any restrictions to use by non-academics: None.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06036-0.

Additional file 1.

Acknowledgements
The authors would like to acknowledge Mulugeta Nega and Thomas Hamm from the University of Tübingen, Germany, for critical reading of the manuscript and testing the tool. And also Adi Kliot and Preetom Regon, from the Volcani Institute, State of Israel, and Gianfranco Pischedda from the University of Sassari, Italy, for testing the tool.

Author contributions
G.M. designed the pipeline structure, developed the core scripts and wrote the manuscript. R.A. and G.M. engineered the scripts and tested the tool. M.G.M. conducted the pre-sequencing laboratory work. G.M., R.A., M.G.M. and A.A. reviewed and edited the manuscript. G.M. and R.A. equally contributed to this paper.

Funding
The project was fully funded by Bioecopest Srl.

Availability of data and materials
The full code of CompàreGenome is available in the project webpage and as GitHub repository. https://bioecopest.com/comparegenome https://github.com/gmoro-bioecopest/CompareGenome The genome assemblies of the strain isolated and sequenced in this study have been deposited on the NCBI website with accession numbers PRJNA1203612 and PRJNA1203920.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 June 2024 Accepted: 3 January 2025
Published online: 13 January 2025

References

1. Yu J, Blom J, Glaeser SP, Jaenicke S, Juhre T, Rupp O, Schwengers O, Spänig S, Goesmann A. A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies. J Biotechnol. 2017;261:2–9.
2. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
3. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 1st ed. O'Reilly Media; 2020.
4. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44(11):e108–e108.
5. Zhang X, Hu Y, Cheng Z, Archibald JM. HSDecipher: a pipeline for comparative genomic analysis of highly similar duplicate genes in eukaryotic genomes. STAR Protoc. 2023;4(1):102014.
6. Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR, Kirsch R, Rattei T, Letunic I, Jensen LJ, Bork P, von Mering C, Huerta-Cepas J. eggNOG 6.0: enabling comparative genomics across 12535 organisms. Nucleic Acids Res. 2023;51(D1):D389–94.
7. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J. ACT: the Artemis comparison tool. Bioinformatics. 2005;21(16):3422–3.
8. Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream M-A. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics. 2008;24(23):2672–6.
9. Waseem M, Basheer A, Anwer F, Shahid F, Zaheer T, Ali A. Genomics, metagenomics, and pan-genomics approaches in COVID-19. Omics approaches and technologies in COVID-19. Elsevier; 2023.
10. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14(7):1394–403.
11. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011;12(1):402.
12. Li JX, Fernandez KX, Ritland C, Jancsik S, Engelhardt DB, Coombe L, Warren RL, van Belkum MJ, Carroll AL, Vederas JC, Bohlmann J, Birol I. Genomic virulence features of Beauveria bassiana as a biocontrol agent for the mountain pine beetle population. BMC Genomics. 2023. https://doi.org/10.1186/s12864-023-09473-4.
13. Ann S, Choi Y, Yoon Y. Comparative Genomic Analysis and Physiological Properties of Limosilactobacillus fermentum SMFM2017-NK2 with Ability to Inflammatory Bowel Disease. Microorganisms. 2023;11(3):547.
14. Li Y, Hu Y, Bolund L, Wang J. State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics. 2010;4:271.
15. Narzisi G, Mishra B. Comparing de novo genome assembly: the long and short of it. PLoS ONE. 2011. https://doi.org/10.1371/journal.pone.0019175.
16. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: Architecture and applications. BMC Bioinfor. 2009;10:1–9.
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
18. Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, Hartline E, Fey P, Thomas PD, Albou L-P, Ebert D, Kesling MJ, Mi H, Muruganujan A, Huang X, Mushayahama T, LaBonte SA, Siegele DA, Antonazzo G, Attrill H, Brown NH, Garapati P, Marygold SJ, Trovisco V, dos Santos G, Falls K, Tabone C, Zhou P, Goodman JL, Strelets VB, Thurmond J, Garmiri P, Ishtiaq R, Rodríguez-López M, Acencio ML, Kuiper M, Lægreid A, Logie C, Lovering RC, Kramarz B, Saverimuttu SCC, Pinheiro SM, Gunn H, Su R, Thurlow KE, Chibucos M, Giglio M, Nadendla S, Munro J, Jackson R, Duesbury MJ, Del-Toro N, Meldal BHM, Paneerselvam K, Perfetto L, Porras P, Orchard S, Shrivastava A, Chang H-Y, Finn RD, Mitchell AL, Rawlings ND, Richardson L, Sangrador-Vegas A, Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov DM, Harris MA, Oliver SG, Rutherford K, Wood V, Hayles J, Bähler J, Bolton ER, De Pons JL, Dwinell MR, Hayman GT, Kaldunski ML, Kwitek AE, Laulederkind SJF, Plasterer C, Tutaj MA, Vedi M, Wang S-J, D'Eustachio P, Matthews L, Balhoff JP, Aleksander SA, Alexander MJ, Cherry JM, Engel SR, et al. The gene ontology resource: enriching a gold mine. Nucleic Acids Res. 2021;49(D1):D325–34.
19. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3.
20. Charif D, Lobry JR. SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Structural approaches to sequence evolution: Molecules, networks, populations. Berlin, Heidelberg; 2007.
21. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B. gplots: Various R Programming Tools for Plotting Data. 2022. https://CRAN.R-project.org/package=gplots
22. Alboukadel K, Fabian M. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. 2020. https://CRAN.R-project.org/package=factoextra
23. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.
24. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3.
25. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. 2020. https://CRAN.R-project.org/package=cowplot
26. Carlson M. GO.db: A set of annotation maps describing the entire Gene Ontology. 2020. https://bioconductor.org/packages/release/data/annotation/html/GO.db.html

27. Atzeni R, Moro G, Marche MG, Uva P, Ruiu L. Genome sequence of beauveria bassiana strain ATCC 74040, a widely employed insect pathogen. Microbiol Resour Announc. 2020. https://doi.org/10.1128/MRA.00446-20.
28. Touchman J. Comparative genomics. Nature Edu Knowl. 2010;3(10):13.
29. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):1–8.
30. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci. 2009;106(45):19126–31.
31. Rosenberg MS. Evolutionary distance estimation and fidelity of pair wise sequence alignment. BMC Bioinformatics. 2005;6(102):1–9.
32. Babraham bioinformatics-FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc
33. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19(5):455–77.
34. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
35. Anaconda Software Distribution. Computer software. https://docs.anaconda.com/
36. De Mendiburu F. agricolae: Statistical Procedures for Agricultural Research. 2021. https://CRAN.R-project.org/package=agricolae
37. Slowikowski K. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. 2021. https://CRAN.R-project.org/package=ggrepel
38. Wang H, Peng H, Li W, Cheng P, Gong M. The toxins of Beauveria bassiana and the strategies to improve their virulence to insects. Front Microbiol. 2021. https://doi.org/10.3389/fmicb.2021.705343.

Publisher's Note