An evolutionary view of the Fusarium genome

Daniel A. Gomez-Chavarria
Centro Nacional de Secuenciación Genómica - CNSG, Universidad de Antioquia

Alvaro L. Rua-Giraldo
University of Antioquia

Juan F. Alzate
jfernando.alzate@udea.edu.co

University of Antioquia

Research Article

Keywords: Fusarium, phylogenomics, comparative genomics, evolution, genome evolution

Posted Date: October 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3457417/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Genomics on March 22nd, 2024. See the published version at https://doi.org/10.1186/s12864-024-10200-w.

Abstract

Fusarium is an Ascomycota with several relevant pathogenic species of plants and animals. Some phytopathogenic species have received special attention due to their negative economic impact on the agricultural industry around the world. Traditionally, identification and taxonomic analysis of *Fusarium* have relied on morphological and phenotypic features, including the hosts of the fungus, leading to taxonomic conflicts that have been solved using molecular systematic technologies. In this work, we applied a phylogenomic approach that allowed us to resolve the evolutionary history of the species complexes of the genus and present evidence that supports the *F. ventricosum* species complex as the most basal lineage of the genus. Additionally, we present evidence that proposes modifications to the previous hypothesis of the evolutionary history of the *F. staphyleae*, *F. newnesense*, *F. nisikadoi*, *F. oxysporum*, and *F. fujikuroi* species complexes. Evolutionary analysis showed that the genome GC content have a tendency to be lower in more modern lineages, below 49.3%, while genome size gain and losses are present during the evolution of the genus. Interestingly core genome duplication events have a positive correlation with the genome size. Evolutionary and genome conservation analysis supports the F3 hypothesis of *Fusarium* as a more compact and conserved group in terms of genome conservation. By contrast the most basal clades, outside the F3 hypothesis only share 8.8% of its genomic sequences with the F3 clade.

BACKGROUND

Fungi can behave as saprophytes, endophytes, and pathogens, but only a few of them represent a risk to other living beings (1). *Fusarium* is an Ascomycota that encompasses several relevant pathogenic species of plants and animals, including humans (2). Currently, some phytopathogenic species have taken special notoriety as responsible for economic losses valued at billions of dollars per year for the agricultural industry around the globe, due to their potential to generate devastating epidemics in almost any crop (cereals, vegetables, ornamental plants, fruits, flowers, etc.) (3, 4). Another of the characteristics for which *Fusarium* species stands out is its ability to produce mycotoxins, which contaminate agricultural products rendering them unsafe for human or animal consumption. Several of these mycotoxins have been associated with cancer and hormonal disorders in humans and farm animals (5). Moreover, some *Fusarium* species can be found causing disease in humans; ranging from onychomycosis, skin infections, and keratitis in immunocompetent individuals to invasive or disseminated infections, mainly in neutropenic and immunosuppressed patients (6). Also, some wall components and cell metabolites of this genus have been implicated in allergic processes in hypersensitive individuals (7). Nowadays, it is estimated that *Fusarium* species can comprise more than 400 phylogenetically distinct species, most of them discovered in the last 25 years (8).

In recent decades, the agricultural sector witnessed with fear, the re-emergence of *Fusarium*, with the appearance *of Fusarium oxysporum f.sp. cubense* tropical race 4 (Foc TR4). This pathogen started to affect the Cavendish banana crops of this cultivar around the 1970s and now is present in all continents where banana is grown causing millionaire losses. Cavendish cultivar was the solution to the appearance of the *F. o. fsp cubense* (Foc) race 1, the *Fusarium* species responsible for the well-known Panama disease, that was originally described in Australia, in 1874, and destroyed all monoculture of banana cultivars ‘Gros Michel’ around the globe by the begging of the XX century (9).

The destructive power of *Fusarium* is not restricted to crop devastation, some species produce secondary metabolites like mycotoxins, that can be toxic to humans and animals, including gibberellins and the mycotoxins fusarins, fumonisins, and trichothecenes (2, 5, 10, 11). These toxins are produced by the fungus on stored agricultural products, or even directly on the growing plant (12). In the period 1930–1940 in the Volga and Ural regions, the presence of fusariotoxins in winter cereals claimed the lives of tens of thousands of people (13).

Fusarium as a taxonomic group was first described in 1809 by Johann Heinrich Friedrich Link (Link, 1809). However, it went unnoticed until the publication "Die Fusariem" by Wollenweber and Reinking in 1935 where 65 species, 55 varieties, and 22 forms of *Fusarium* were described (14). Historically, other alternative generic names have been proposed for *Fusarium* isolates based on the morphology of sexual stages like *Gibberella*, *Nectria*, and *Neocosmospora*. Nonetheless, only the *Fusarium* genus name should be used to avoid misunderstandings (2). To date, more than 400 phylogenetically distinct species in 23 monophyletic species complexes are included in the genus *Fusarium* although not all of them have been formally designated (15).

Traditionally, *Fusarium* species classification relies heavily on morphological and phenotypical characteristics, which includes the affected host organism. The macro and microscopic structures analyzed are highly variable and similarities between closely related species set the origin for several of the taxonomic inconsistencies observed until the first decade of this century. With the scientific and technological developments at the end of the XX century, genetic and biochemical features were added to solve morphological disagreements. Nevertheless, is not until recent years, using modern phylogenetic methods, that evolutionary approaches started to clarify the chaotic taxonomic assignment of Fusaria (3, 11, 13, 16, 17). Several of these first phylogenetic works were based on multilocus sequence analysis of conserved genes, or partial genes like *cmdA*, *rpb1, rpb2*, *tef1*, and *tub2* (11, 16). One of the noteworthy successes reached with molecular phylogenetics is recognizing *Fusarium* as a monophyletic group and the setting of the initial hypothesis of evolutionary relationships within the genus (11).

Even using molecular systematics, many published phylogenies have been conflictive or debated, especially those that have proposed major nomenclatural changes. Sometimes it has been difficult to draw the limit between species and infraspecific lineages, and some results do not offer enough solidity to explain the evolutionary history of the genus, representing a problem for the academic community of *Fusarium* (3, 11).

With this work, we aimed to contribute to a deeper understanding of *Fusarium's* evolutionary relationships and various genomic features of *Fusarium* reference strains, including GC content, genome size, genome and core proteome conservation, and ancestral gene duplication events. Our results, derived from a phylogenomic approach utilizing 559 conserved proteins, provide evidence supporting the *F. ventricosum* species complex as the most basal lineage within the genus. Additionally, our study offers a novel perspective on the evolutionary history of the *F. sta*phyleae, *F. newnesense*, *F. nisikadoi*, *F. oxysporum*, and *F. fujikuroi* species complexes.

METHODS

Reference genomes

The NCBI-Datasets website was consulted for *Fusarium* genomes on 06/06/22. This result was filtered by the criterion "reference genomes RefSeq" from which a total of 224 genomes were obtained which made up the initial dataset (18). Subsequently, the assemblies were processed with the BUSCO v3 program (19) in order to obtain additional quality metrics of the *Fusarium* genome assemblies. In July of the same year, a new reference genome of the *ventricosum* species complex was added, *Fusarium robinianum* CBS430 (GCA_024115165). This genome was included in our dataset. Additionally, the reference genome that we originally downloaded on June/22 downloaded as *F. ventricosum* NRRL 25729 (GCA_013623725) was removed from the database and a new entry with the same genome was added as *F. robinianum* NRRL 25729 (GCA_013623725.1). We used this new RefSeq genome entry and the *ventricosum* species complex is represented by two *F. robinianum* genomes, strains NRRL 25729 and CBS430. The detailed list of the accession numbers of the genomes used in this study as well as the descriptive statistics of them can be found in Supplementary table 1.

A boxplot analysis was implemented, allowing us to know the dispersion ranges of genome metrics values of central tendency measures such as median and quartiles; all this to identify the outliers. The metrics analyzed included assembly length (Assembled genome size), assembly N50, Largest scaffold, scaffold count, BUSCO completeness, BUSCO single copy genes, BUSCO duplicated genes, BUSCO fragmented genes, and BUSCO missing genes.

We proceeded to apply filters in the initial dataset to filter out those genomes that showed low-quality values. For this, we set as threshold the upper quantile limit of scaffold count (≤ 4,225) and also the BUSCO completeness and BUSCO single copy genes ratios to ≥ 90%. The result of these filters was the removal of 24 low-quality genomes from the initial dataset:

GCA_001680625-*Fusarium*_azukicola, GCA_001680685-*Fusarium*_brasiliense, GCA_011036685-*Fusarium*_cf_nygamai, GCA_001680505-*Fusarium*_cuneirostrum, GCA_012978535-*Fusarium*_sp_NRRL_62957, GCA_013266265-*Fusarium*_sp_NRRL_66182, GCA_001680515-*Fusarium*_phaseoli, GCA_0131864252-*Fusarium*_sp_NRRL_62610, GCA_014764975-*Fusarium*_sp_DS_682, GCA_013186395-*Fusarium*_sp_NRRL_62944, GCA_013363185-*Fusarium*_secorum, GCA_006518225-*Fusarium*_neocosmosporiellum, GCA_013184375-*Fusarium*_newnesense, GCA_013618265-*Fusarium*_albidum, GCA_008711595-*Fusarium*_xyrophilum, GCA_013624395-*Fusarium*_sp_KOD_1611, GCA_013623595-*Fusarium*_nematophilum, GCA_013010345-*Fusarium*_sp_NRRL_22101, GCA_014824365-*Fusarium*_sp_NRRL_66894, GCA_014824405-*Fusarium*_sp_NRRL_66896, GCA_012978555-*Fusarium*_sp_NRRL_62941, GCA_013186415-

*Fusarium*_sp_NRRL_66088, GCA_013363215-*Fusarium*_acuminatum, GCA_0134167852-*Fusarium*_sp_BWC1.

Proteome annotation and filtering

Using the BUSCO (19) hypocreales_odb10 reference and AUGUSTUS genome annotator (with *Fusarium graminearum* reference training set) (20, 21) we identified a set of ancestral single-copy proteins of the *Fusarium* reference genomes (18). Every downloaded genome was annotated with the same strategy. The annotated proteins in the genome of *Fusarium oxysporum* GCF_000271745, identified by the program as single-copy proteins, were then compared using BLASTP (22) with a database of the ancestral conserved proteins of the BUSCO hypocreales_odb10 databank. Those proteins that were above the filtering criteria e-value = 0, %id = 80, and bit score > = 600, were kept as candidates for the phylogenomic analysis, 996 in total. Then, these 996 proteins were annotated using the EGGNOG-MAPPER web server (23). Those sequences whose assignment of eggNOG orthologous group did not coincide with Hypocreales, family Nectriaceae, or that presented a duplicated KEGG KO assignment code were eliminated. After this, a set of 559 proteins was kept as the selected markers for the phylogenomic analysis (Supplementary table 2).

Phylogenomic analysis

To construct the 559-protein alignment super-matrix a combination of BLASTP searches and individual MAFFT alignments(24) was performed. The 559 filtered/curated proteins (from single-copy conserved genes) of the reference strain *Fusarium oxysporum* GCF_000271745 were used as a query to catch the respective ortholog in each of the 225 proteomes using BLASTP (The same process was applied for the three *Trichoderma* outgroups). Each set of orthologous proteins was aligned individually using the MAFFT aligner. The 559 individual protein alignments were concatenated using the program catsequences (https://github.com/ChrisCreevey/catsequences). As an outgroup, we included 3 *Trichoderma* reference species: *Trichoderma oligosporum* (ASM1526638v1), *Trichoderma viride* (ASM789649v1), and *Trichoderma afroharzianum* (ASM2073690v1). One maximum likelihood phylogenomic tree was computed using IQ-TREE v2(25) performing 5000 UFB pseudoreplicates(26) and using the different partitions models(27) for each protein with the options -m MFP + MERGE and -rcluster 15. The partition scheme used can be found in the supplementary material. The Log-likelihood of consensus tree was − 7139388.639034. Furthermore, we computed gene concordance factors (gCF) and site concordance factors (sCF) using IQTREE v. 2.2.0.8, employing the "−gcf" and "−scf" options to assess the level of genealogical agreement. gCF is defined as the proportion of gene trees that include a specific branch considered "decisive" for each branch of a species tree, while sCF is defined as the percentage of decisive alignment sites supporting that particular branch.

Genome-to-genome comparisons and Average Amino Acid Identity (AAI) Analysis

Genome alignment and comparative analysis were conducted using the DNADIFF program from the MUMMER v3 software (REF). All *Fusarium* genomes were aligned with each other, and the fraction of aligned bases and average nucleotide identity were extracted from the '.report' file. Subsequently, a non-redundant table was created and imported into R for further analysis.

The Average Amino Acid Identity score was calculated using the EzAAI program (REF). To do this, single-copy proteomes annotated with BUSCO were used as input for the comparisons. An all-vs-all comparison of single-copy proteomes was performed, and a non-redundant summarized table containing the AAI score values and proteome coverage ratio for all comparisons was constructed and utilized for the statistical analysis.

Statistical Analysis and Graphics

Boxplots and quantile analysis were conducted using R and RStudio v.4.1.3 platforms, with the assistance of the ggplot2 library (28, 29). Graphics, including boxplots and scatter plots, were generated using the ggplot2 package. Phylogenetic trees were visually edited using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Correlation analysis between GC content and genome size, as well as the ratio of core genome duplicated genes and genome size, was performed in R using the cor.test function with the Spearman's test. Group comparisons were assessed using the Kruskal-Wallis Rank Sum Test with the R function kruskal.test

RESULTS

Evolutionary History of the Fusarium Species Complexes:

In this study, we incorporated a substantial number of reference species genomes obtained from the NCBI RefSeq database, encompassing all the *Fusarium* species complexes documented to date (Supplementary Table 1). Subsequently, we employed an extensive set of loci, which consisted of 559 conserved single-copy proteins, to establish a phylogenomic framework within the *Fusarium* genus for subsequent analyses.

The selection of these 559 proteins was based on three specific criteria: i) encoded by single-copy genes, ii) annotated as Nectriaceae proteins by the Eggnogmapper tool, and iii) showing no duplicated KO annotation terms. These criteria were adopted to minimize the inclusion of paralogous proteins, which could potentially introduce noise into the phylogenetic reconstruction. The list of the selected proteins and their annotations can be found in Supplementary Table 2.

Our reconstructed phylogenomic tree aligns with prior research, supporting *Fusarium* as a monophyletic group (100 UFB support). Furthermore, the tree illustrates all currently accepted species complexes as monophyletic groups with 100% UFB support (Fig. 1). Within the basal clade of our phylogenomic tree, the *F. ventricosum* complex takes the lead, followed, in order, by the *F. dimerum*, *F. albidum*, and *F. staphyleae* species complexes.

Subsequently, the tree delineates two major branches: one encompassing the *F. solani* and *F. decemcellulare* lineages and the other clade comprising the remaining species complexes. The most basal lineages within the latter branch are, in order, the *F. buxicola*, *F. buharicum*, *F. lateritium*, and *F. torreyae* species complexes. This final branch encompasses most of the described species complex groups to date. Within this branch, we identify two primary lineages: one encompassing the species complexes *tricintum*, *heterosporum*, *incarnatum*, *equiseti*, *chlamydosporum*, and *sambucinum*, and the other including the species complexes *concolor*, *babinda*, *burgessi*, *redolens*, *newnesense*, *nisikadoi*, *oxysporum*, *claminii*, and *fujikuroi*. Supplementary Fig. 1 presents the complete, uncollapsed phylogenomic tree, comprising 225 *Fusarium* genomes analyzed.

NCBI RefSeq database Fusarium genome quality analysis

The NCBI RefSeq database had, until June 2022, 224 *Fusarium* reference genomes, each representing a distinct species. In the subsequent month, an additional reference strain, CBS430 of *F. robinianum*, was introduced and included in our study. Additionally, we observed a name change for the *F. ventricosum* NRRL 25729 strain to *F. robinianum* NRRL 25729 in the NCBI datasets database.

We conducted an analysis of various genome quality parameters within the database, including assembly length (genome size), assembly N50, largest scaffold, and scaffold count. Additionally, we performed a BUSCO genome quality control assessment, evaluating metrics such as completeness, single copy genes, duplicated genes, fragmented genes, and missing genes.

As depicted in the boxplots in Fig. 2, genome statistics exhibited relatively compact value dispersions for assembly length and BUSCO genome completeness. However, the presence of outlier values and genomes of questionable quality became evident. For the other assessed metrics, such as N50 and scaffold count, we observed a wider dispersion of values, indicating a heterogeneous performance in genome sequencing experiments. Notably, there was an approximately one-order-of-magnitude difference between the lower and upper limits of the boxplots, suggesting varying levels of success in the genome assembly process. Similar trends were observed in the BUSCO genome quality metrics, with some genomes performing poorly.

To enhance the quality and reliability of subsequent analyses, we implemented genome quality thresholds: i) scaffold number (≤ 4,225); and ii) BUSCO metrics completeness and single copy genes (≥ 90%). Applying these filters led to the removal of 24 genomes from the dataset (For a detailed list of excluded genomes, please refer to the methods section). After refining the dataset, we reevaluated the general statistics of the *Fusarium* genomes, resulting in an improved value dispersion of the BUSCO metrics, as illustrated in Fig. 3.

Evolution of the Fusarium Genome: GC Ratio, Size, and Core Genome Duplications

We grouped the RefSeq genomes according to their respective species complexes and created box plots illustrating the variations of the genome GC content, assembled genome size, and the fraction of duplicated core genes within each group. To gain insights into the evolutionary trends of these genome features, we arranged the species complexes in the plots based on their positions in the phylogenomic tree (Fig. 4). The analysis revealed that variations in these genomic features depended on the studied group, and in most cases, genome disparities were narrower within each species complex compared to the entire genus. These findings were statistically significant for all three analyzed variables, with p-values below 2.2e-16 in all cases.

In the *Fusarium* genus, assuming the F1 hypothesis, there are striking variations in GC content, ranging from 43.4–55.2%. From an evolutionary perspective, it is interesting to note that the ancestral clades, basal to the F3 group, generally exhibit higher GC content, exceeding 51% (species complexes *ventricosum*, *albidum*, *staphyleae*, *solani*, *decemcellulare*, and *buxicola*), except for the *F. dimerum* species complex. The remaining species complexes within the F3 group display median GC content values below 49.3%.

Historically, the genus *Fusarium* has been described as having a broad range of genome sizes, ranging from 33 to 60 Mbp. However, our analysis reveals that genome sizes range from 32.05 to 65.63 Mb. Nevertheless, within each species complex, the variations are more discrete and show significant changes during evolution. The *F. redolens* species complex exhibits the largest genomes (median size 52.6 Mb), while the *F. staphyleae* complex shows the most reduced genome versions (median 33.25 Mb).

As a complementary analysis, we investigated whether there is a relationship between genome size and duplication events in the *Fusarium* core genome. To do so, we quantified and plotted the proportion of duplicated core genes within each genome. As shown in Fig. 4, panel C, duplication events of core genes vary depending on the species complex, with the *F. solani*, *F. decemcellulare*, and *F. redolens* complexes having the highest proportions at 0.6%. Additionally, the boxplot graph reveals a consistent trend between genome size and the proportion of core genes duplicated. To confirm this observation, we conducted a correlation analysis between genome size and duplicated genes. As seen in Fig. 5, panel A, there is a strong correlation between genome size and the proportion of duplicated core genome genes ($R = 0.71$, $p < 2.2e-16$). In contrast, GC content does not exhibit such a correlation ($R = 0.017$, $p = 0.81$) (Fig. 5, panel B).

Genome conservation in Fusarium

To gain a deeper understanding of genome conservation among *Fusarium* reference species, we employed a genome-to-genome alignment strategy using the MUMMER program. Our analysis encompassed the calculation of both the proportion of the aligned genomic blocks and the nucleotide identity within them. The results were visualized using boxplots, wherein the different species complexes were grouped, and their order was determined by the previous phylogenomic tree, providing an evolutionary perspective (Fig. 6). Median values of the proportion of genome-aligned bases within each group exhibited a range from 11.23–94.99%, with a decrease to 8.82% when comparing between species complexes. In contrast, the median values of nucleotide identity within each complex spanned from 84.71–99.92%, with a median of 84.50% for comparisons between species complexes. The overall median values for genome-aligned bases and nucleotide identity, when comparing genomes within each corresponding species complex, were 75% and 90%, respectively. These differences in the fraction of genome-aligned bases and the average nucleotide identity were found to be statistically significant, with a p-value < 2.2e-16.

In addition, we conducted a scatter plot analysis comparing the proportion of aligned genome versus nucleotide identity, both within and between species complexes. This analysis revealed a general trend where inter species-complexes exhibited lower proportions of genome aligned blocks, while intra species-complexes comparisons tended to have larger proportion (Fig. 7).

In addition, we sought to quantify the extent of conservation within the core proteome of the *Fusarium* genus. To achieve this, we annotated and compared all the conserved single-copy proteins using the BUSCO software. The average amino acid identity (AAI) of these conserved single-copy proteins for each genome and the overall coverage ratio of the single-copy proteome are visualized in the scatter plot presented in Fig. 9. Different colors are used to denote comparisons between and within species complexes. The analysis demonstrates that the single-copy core proteome of *Fusarium* is markedly well-conserved among the examined reference genomes, with at least 90% of it present in nearly all tested species and displaying an AAI value exceeding 70%. Notably, when comparing within or between species complexes, intra-lineage comparisons yielded higher AAI values, consistently above 88%, for nearly all proteomes, and coverage ratios exceeding 0.91.

A more detailed analysis, utilizing boxplots and differentiating the F3 lineage within the *Fusarium* genus, reveals that the basal species complexes in the F3 lineage tend to exhibit lower AAI and proteome coverage ratios compared to the F3 lineage as a whole. Specifically, the median AAI value for inter-complex comparisons was 86.32%, whereas intra-complex comparisons yielded a median AAI value of 96.32%. Interestingly, the core proteome coverage was quite similar in both comparison groups, with values of 0.98 and 0.99 for inter-species complexes and intra-species complexes comparisons, respectively (Fig. 10). These differences in the AAI score and core proteome coverage ratio were statistically significant, with both obtaining a p-value < 2.2e-16.

In order to provide a more comprehensive understanding of genome conservation within the *Fusarium* genus, especially in relation to the F3 hypothesis, we categorized the genomes into two primary groups: the F3 lineage and the basal taxa. We then conducted genome-to-genome alignment comparisons to assess their conservation patterns. As depicted in Fig. 11, Panel A, genomes within the F3 hypothesis clade demonstrated a notably higher level of conservation, with a median alignment success rate of 12.6%. In contrast, when comparing alignment coverage between the F3 lineage and the basal taxa, there was a significant decline to 3.19%. Within the basal taxa themselves, the median proportion of aligned genome reached 12%. Interestingly, the average nucleotide identity of the aligned genome blocks displayed similar values across all three tested groups, measuring at 84.8% for basal taxa, 84.6% for the F3 lineage, and 84.5% for inter-lineage comparisons (see Fig. 11, Panel B).

Furthermore, when examining the conservation of the core proteome, subtle yet significant differences emerged. The AAI value within the F3 group reached 87.5, whereas it dropped to 81.5 for comparisons between the F3 and basal genera species. In terms of core proteome coverage, the median value within the F3 group was 0.983, while it reduced to 0.968 when comparing the F3 group with basal genera species. Importantly, all comparisons among the three groups yielded statistically significant results, with p-values consistently below $p < 2.2e-16$ in all cases.

DISCUSSION

Taxonomy based on morphological or phenotypical characteristics of microscopic organisms has been exceptionally challenging since their discovery more than three centuries ago (30). Although classical taxonomic methods prompted progress for *Fusarium* studies, major concerns emerged with the advent of molecular systematics (16). Thankfully, the recent advances in NGS and bioinformatics allowed mycologists to start clarifying the complex relationships within this fungal taxon and to reveal novel species (2, 11, 17, 31). In the case of *Fusarium*, advances are remarkable, leading to the identification of approximately 400 species and a better understanding of the genus's evolutionary history (8). As a first step in this work, we assessed the quality of the genomes available in the NCBI RefSeq database for *Fusarium* species. While most of these genomes are reliable, some exhibited poor quality indicators. These low-quality genomes should be avoided as references for comparative genomic analysis, emphasizing the need for new, trustworthy genome projects for these species.

Our phylogenetic analysis, based on 559 conserved single-copy proteins, confirms that the currently recognized 28 species complexes of the *Fusarium* genus form well-supported monophyletic groups. Furthermore, 198 (91%) nodes in our tree have a support level of ≥ 95%, with nodes F1, F2, and F3 achieving 100% support (31).

The depicted phylogeny also aligns with more recent studies, supporting the concept of the 'broad' *Fusarium* clade, which encompasses 17 species complexes and 11 allied genera (32). The tree topology also gives support to the 'narrow' concept of the genus *Fusarium* sensu stricto, which includes only the 17 species complexes, often referred to as the F3 hypothesis (32–35). The branch lengths observed in the phylogenomic tree further indicate a wider evolutionary divergence between the 'allied' genera and the species at the F3 node.

Concordance factors (gcf and scf) were calculated to refine the accuracy of our conclusions regarding phylogenetic reconstructions. For instance, F1 (gcf 98.8%, scf 91.7%), F2 (gcf 55%, scf 46.9%), and F3 (gcf 83.4%, scf 71.6%) showed significant support. A gcf value above 50% suggests that more than half of the ortholog proteins used support the node positions. It's worth noting that even low support values do not necessarily indicate an indeterminate phylogeny; instead, they provide information about the degree of relationship or congruence, helping to elucidate the evolutionary history of the species (36).

We have identified some topological differences when comparing our phylogenomic results to previously published phylogenies, particularly at the basal nodes. One significant difference involves the positioning of the *F. ventricosum* species complex. In our phylogenetic analyses, the *F. ventricosum* complex emerges as the most ancestral taxon within *Fusarium*, which contrasts with the hypothesis presented by Geiser and O'Donnell (2, 11, 31, 37). In those phylogenies, the *F. ventricosum* and *F. dimerum* species complexes were grouped as an ancestral monophyletic clade, albeit with low support, such as 64% ML-BS, less than 50 maximum parsimony bootstrap (MP-BS), and Bayesian posterior probability (B-PP) of 1.0 (31). O'Donnell in 2013 reported ML-BS and BP-BS values below 70%, and Geyser (2021) noted BS values under 70 and BPP below 0.99, with a gene concordance factor (gcf) of 0, indicating minimal support for this node in their analyses. In all these cases, the authors acknowledged the necessity for additional studies to clarify the position of these complexes, given the low support values. Conversely, in other phylogenies presented by Lombard et al. (34), Gräfenhan et al. (33), Han et al. (32), and Chen Y.P et al. (38), the *F. ventricosum* complex is placed as ancestral to the *F. dimerum* complex, in concordance to our findings.

Regarding the basal clades of the genus, F1 hypothesis, we observed four basal complexes, each with strong support: i) ventricosum complex (allied genus *Rectifusarium*) node is well-supported with UFB 100%, gcf 98.8%, and scf 91.7%, ii) *F. dimerum* complex (allied genus *Bisifusarium*) node shows support with UFB 100%, gcf 32.9%, and scf 38.4%, iii) *F. albidum* complex (allied genus *Luteonectria*) node has support with UFB 100%, gcf 55%, and scf 46.9%, and iv) *F. staphyleae* complex (allied genus *Geejayessia*) node receives BS 100%, gcf 26.4%, and scf 35.9%. Within this node, we find the allied genus *Nothofusarium* forming a monophyletic clade with *Geejayessia*. This genus is found in the phylogenies presented by Han et. Al. (32) and Chen et. al. (38) with robust bootstrap support.

The position of the *F. staphyleae* species complex has been a topic of debate. O'Donnell et al (11) proposed it as the 4th most ancestral clade, but its position lacked support, warranting further analysis. In contrast, Geyser et al. 2021 (37), positioned the *F. staphyleae* complex within a clade containing the *F. solani*, *F. decemcellulare*, and *F. buxicola* species complexes. However, this placement also had relatively low support (81% ML-BS and 0.99 BPP). Our phylogenomic tree supports the hypothesis that *Setofusarium setosum* (allied genus *Setofusarium*) and *F. staphyleae* (allied genus *Albonectria*) represent ancestral lineages of the *F. solani* (allied genus *Neocosmopora*) and *F. buxicola* (allied genus *Cyanonectria*) species complexes.

In summary, our results support the hypothesis that the most ancestral clades within the *Fusarium* genus, in sequence, are *F. ventricosum*, *F. dimerum*, *F. albidum*, and *F. staphyleae* species complexes. They are followed by the *F. solani + decemcellulare* clade and then the *F. buxicola*, *F. buharicum*, *F. lateritium*, and *F. torreyae* complexes. Notably, these last four complexes share the same topology and receive 100% UFB support in both our phylogeny and the one proposed by Geiser et al. in 2021 (37). Subsequently, lineages beyond *F. buxicola* (allied genus *Cyanonectria*) form the F3 clade, *Fusarium* s. str., comprising 17 complexes of *Fusarium* species. This taxonomic proposal aligns with the phylogenies presented by Han et. al. (32) and Chen et. al(38).

Regarding the species complexes *F. sambucinum* (FSAMSC), *F. incarnatum-equiseti* (FIESC), *F. babinda* (FBSC), *F. concolor* (FCOSC), *F. burgessi* (FBurSC), and *F. redolens* (FRSC), no topological differences are observed concerning the phylogenies presented by Geiser (2021) (37), Chen (2023) (38), and S. L. Han (2023) (32). These nodes in our analysis received strong support with 100% UFB. While Geyser et al. reported poor support for the *F. concolor* clade, our results provide robust evidence for the monophyly of this species complex, with 100% UFB support.

We also found that the *F. newnesense* (FnewSC), *F. nisikadoi* (FNSC), *F. oxysporum* (FOSC), and *F. fujikuroi* (FFSC) species complexes form a monophyletic clade in agreement with Geiser et al. (2021) (37), Han et. al. (32), Chen et. al. (38), and Crous et. al. (35). However, our phylogeny suggests a different evolutionary history within this clade. In our analysis, *F. fujikuroi* emerges as the most ancestral lineage, followed by *F. oxysporum* and the *F. nisikadoi + newnesense* branches. Geiser et al. (2021) reported less consistent support for this clade, with some bootstrap values below 90% (37).

The evolutionary analysis of genome GC content suggests that ancestral *Fusarium* lineages, the ‘allied’ genera, have higher GC content ratios, exceeding 50%, while most modern clades in the F3 clade have shown a reductive trend in this index, with values dropping to 47–48% in most species complexes. Conversely, genome size did not exhibit a consistent vertical evolutionary trend. While ancestral lineages, such as *F. ventricosum* and *F. dimerum*, showed smaller genomes below 38 Mb, the remaining species complexes displayed significant gains and losses in genomic content, resulting in genome sizes fluctuating between 32 and 66 Mbp.

Notably, substantial genome gains occurred in both basal taxa and within the F3 hypothesis, particularly in the *F. albidum*, *F. solani*, *F. decemcellulare*, *F. oxysporum*, *F. newnesense*, *F. redolens*, and *F. burgessii* complexes. This phenomenon has been previously discussed in other studies where the gain and loss of accessory chromosomes drove changes in genome sizes (39). However, our analysis revealed that these genome gains strongly correlate with duplications of conserved genes within the core genome. A similar phenomenon of genome expansion associated with duplications of ancestral genes has been previously observed in Archaea (40).

As a result, genome size and GC content appear to be distinct characteristics within each species complex.

Genome conservation in *Fusarium*, as measured by the proportion of the genome aligned between different species complexes, appears to be relatively low, with only approximately 12% of the genome aligning within the species complexes of the basal clade (‘allied’ genera) or within the species complexes of the F3 clade (*Fusarium* sensu stricto). However, when comparing these two clades, the alignment proportion drops to nearly a quarter, indicating a more distant evolutionary relationship between these two groups.

Conversely, when examining the core proteome of the genus, a notably higher level of conservation becomes evident, with at least 90% of the proteins detected in almost all tested reference strains.

Furthermore, the average amino acid identity score (AAI) excedes 75%, with a median value of 86% when making comparisons between species complexes. One plausible interpretation of this phenomenon aligns with previous reports, suggesting that the basal groups of the F3 clade exhibit a significant evolutionary distance from the *Fusarium* sensu stricto group (F3). This provides support for arguments favoring the classification of the basal groups into different genera outside the genus *Fusarium* (32, 35, 41).

Another intriguing observation emerges from our scatter plot analysis, encompassing both genome and core proteome comparisons. Notably, several species exhibit similar divergence profiles in terms of genome and proteome conservation, regardless of their classification within the same species complexes. This discovery suggests that certain species, traditionally grouped within the same species complex, show evolutionary distances and molecular divergences comparable to those observed between species categorized in different complexes. From a genomic standpoint, this raises questions about potential challenges within the ongoing classification framework of the *Fusarium* genus.

CONCLUSIONS

Our study represents a significant step forward in understanding the taxonomy, evolution, and genome dynamics within the *Fusarium* genus. The advent of molecular systematics, coupled with recent advances in NGS and bioinformatics, has provided us with invaluable tools to tackle the intricate relationships within this fungal taxon. Furthermore, our results indicate a broader evolutionary divergence between the 'allied' genera (basal clades) and the species within the F3 clade.

Our evaluation of the quality of genomes in the NCBI RefSeq database for *Fusarium* species highlights the importance of reliable reference genomes for comparative genomic analyses. While most genomes are dependable, some exhibit poor quality indicators, emphasizing the need for trustworthy genome projects for these species.

Our findings also shed light on the evolution of genome GC content and genome size within *Fusarium* species complexes. Ancestral lineages and some modern clades exhibit distinct patterns in these genomic characteristics. Notably, genome expansions correlate strongly with duplications of ancestral conserved genes within the core genome.

While genome conservation within *Fusarium* species complexes appears relatively low at the genomic level, the core proteome exhibits a notably higher level of conservation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the data used in this study is publicly available in the NCBI DATASETS database (https://www.ncbi.nlm.nih.gov/datasets/taxonomy/5506/). The detailed list of accessions is described in the methods section.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by Centro Nacional de Secuenciación Genómica -CNSG, Universidad de Antioquia, Medellin, Colombia.

Authors' contributions

DAGC conducted the phylogenetic analysis, contributed to manuscript writing, and reviewed it. ALRG participated in data interpretation, contributed to manuscript writing, and reviewed it. JFA was responsible for the overall study design, conducted bioinformatic data processing, and was a major contributor in writing the manuscript. All authors have reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Knogge W. Fungal Infection of Plants. Plant Cell. 1996;8:1711–22.
2. Ma LJ, Geiser DM, Proctor RH, Rooney AP, O'Donnell K, Trail F et al. Fusarium Pathogenomics. http://dx.doi.org/101146/ annurev-micro-092412-155650 [Internet]. 2013 Sep 11 [cited 2022 Sep 27];67:399–416. Available from: https://www.annualreviews.org/doi/abs/10.1146/annurev-micro-092412-155650.
3. Aoki T, O'Donnell K, Geiser DM. Systematics of key phytopathogenic Fusarium species: current status and future challenges. J Gen Plant Pathol. 2014;80(3):189–201.
4. Arie T. Fusarium diseases of cultivated plants, control, diagnosis, and molecular and genetic studies. J Pestic Sci. 2019;44(4):275–81.
5. Woloshuk CP, Shim WB. Aflatoxins, fumonisins, and trichothecenes: a convergence of knowledge. FEMS Microbiol Rev [Internet]. 2013 Jan 1 [cited 2022 Sep 27];37(1):94–109. Available from:

https://academic.oup.com/femsre/article/37/1/94/558665.

6. Batista BG, de Chaves MA, Reginatto P, Saraiva OJ, Fuentefria AM. Human fusariosis: An emerging infection that is difficult to treat. Rev Soc Bras Med Trop [Internet]. 2020 Jun 1 [cited 2022 Sep 27];53:1–7. Available from: http://www.scielo.br/j/rsbmt/a/MkBhLpgzF4fd5cs3H6jMk3c/?lang=en.
7. Yeh CC, Tai HY, Chou H, Wu KG, Shen HD. Vacuolar Serine Protease Is a Major Allergen of *Fusarium proliferatum* and an IgE-Cross Reactive Pan-Fungal Allergen. Allergy Asthma Immunol Res. 2016;8(5):438.
8. O'Donnell K, Whitaker BK, Laraba I, Proctor RH, Brown DW, Broders K, et al. DNA Sequence-Based Identification of *Fusarium*: A Work in Progress. Plant Dis. 2022;106(6):1597–609.
9. Maymon M, Sela N, Shpatz U, Galpaz N, Freeman S. The origin and current situation of Fusarium oxysporum f. sp. cubense tropical race 4 in Israel and the Middle East. Sci Rep. 2020;10(1):1590.
10. Munkvold GP. Fusarium species and their associated mycotoxins. Methods in Molecular Biology [Internet]. 2017 [cited 2022 Sep 27];1542:51–106. Available from: https://link.springer.com/protocol/10.1007/978-1-4939-6707-0_4.
11. O'Donnell K, Rooney AP, Proctor RH, Brown DW, McCormick SP, Ward TJ, et al. Phylogenetic analyses of RPB1 and RPB2 support a middle Cretaceous origin for a clade comprising all agriculturally and medically important fusaria. Fungal Genet Biol. 2013;52:20–31.
12. Perincherry L, Lalak-Kańczugowska J, Stępień Ł. Fusarium-Produced Mycotoxins in Plant-Pathogen Interactions. Toxins (Basel). 2019;11(11):664.
13. Stakheev AA, Samokhvalova LV, Ryazantsev DYu, Zavriev SK, MOLECULAR GENETIC APPROACHES FOR INVESTIGATION OF TAXONOMY, AND SPECIFIC IDENTIFICATION OF TOXIN-PRODUCING Fusarium SPECIES. : ACHIEVEMENTS AND PROBLEMS (review). Sel'skokhozyaistvennaya Biologiya. 2016;51(3):275–84.
14. Wollenweber HW. and ROA. Die Fusarien: Ihre Beschreibung Schadwirkung Und Bekämpfung. P. Parey J.W. Edward; 1935.
15. Mirghasempour SA, Studholme DJ, Chen W, Zhu W, Mao B. Molecular and Pathogenic Characterization of Fusarium Species Associated with Corm Rot Disease in Saffron from China. J Fungi. 2022;8(5):515.
16. Lombard L, Sandoval-Denis M, Lamprecht SC, Crous PW. Epitypification of Fusarium oxysporum – clearing the taxonomic chaos. Persoonia - Molecular Phylogeny and Evolution of Fungi. 2019;43(1):1–47.
17. O'Donnell K, Ward TJ, Robert VARG, Crous PW, Geiser DM, Kang S. DNA sequence-based identification of Fusarium: Current status and future directions. Phytoparasitica. 2015;43(5):583–95.
18. O'leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, Mcveigh R et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res [Internet]. 2015 [cited 2020 Jul 14];44:733–45. Available from: http://www.ncbi.nlm.nih.gov/books/.

19. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
20. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Web Server issue):W435–9.
21. Hoff KJ, Stanke M. Predicting Genes in Single Genomes with AUGUSTUS. Curr Protoc Bioinformatics [Internet]. 2018 Nov 22 [cited 2020 Jul 14];65(1):e57. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.57.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
23. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol [Internet]. 2017 [cited 2020 Jul 14];34(8):2115–22. Available from: http://creativecommons.
24. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol [Internet]. 2013;30(4):772–80. https://doi.org/10.1093/molbev/mst010.
25. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44(W1):W232–5.
26. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. Mol Biol Evol. 2018;35(2):518–22.
27. Chernomor O, von Haeseler A, Minh BQ. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst Biol. 2016;65(6):997–1008.
28. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria. ; 2013. Available from: http://www.r-project.org/.
29. Team R, Development Core A, Language, and Environment for Statistical Computing. R Foundation for Statistical Computing [Internet]. 2018 [cited 2021 Aug 30];2:https://www.R-project.org. Available from: http://www.r-project.org/.
30. Hugenholtz P, Chuvochina M, Oren A, Parks DH, Soo RM. Prokaryotic taxonomy and nomenclature in the age of big sequence data. ISME J. 2021;15:1879–92.
31. Geiser DM, Aoki T, Bacon CW, Baker SE, Bhattacharyya MK, Brandt ME, et al. One Fungus, One Name: Defining the Genus *Fusarium* in a Scientifically Robust Way That Preserves Longstanding Use. Phytopathology. 2013;103(5):400–8.
32. Han SL, Wang MM, Ma ZY, Raza M, Zhao P, Liang JM, et al. *Fusarium* diversity associated with diseased cereals in China, with an updated phylogenomic assessment of the genus. Stud Mycol. 2023;104(1):87–148.
33. Gräfenhan T, Schroers HJ, Nirenberg HI, Seifert KA. An overview of the taxonomy, phylogeny, and typification of nectriaceous fungi in Cosmospora, Acremonium, Fusarium, Stilbella, and Volutella. Stud Mycol. 2011;68:79–113.

34. Lombard L, van der Merwe NA, Groenewald JZ, Crous PW. Generic concepts in *Nectriaceae*. Stud Mycol. 2015;80(1):189–245.

35. Crous PW, Lombard L, Sandoval-Denis M, Seifert KA, Schroers HJ, Chaverri P, et al. Fusarium: more than a node or a foot-shaped basal cell. Stud Mycol. 2021;98:100116.

36. Minh BQ, Hahn MW, Lanfear R. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. Mol Biol Evol. 2020;37(9):2727–33.

37. Geiser DM, Al-Hatmi AMS, Aoki T, Arie T, Balmas V, Barnes I, et al. Phylogenomic Analysis of a 55.1-kb 19-Gene Dataset Resolves a Monophyletic *Fusarium* that Includes the *Fusarium solani* Species Complex. Phytopathology. 2021;111(7):1064–79.

38. Chen Y, Su P, Hyde K, Maharachchikumbura S. Phylogenomics and diversification of Sordariomycetes. Mycosphere. 2023;14(1):414–51.

39. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature. 2010;464(7287):367–73.

40. Sheridan PO, Raguideau S, Quince C, Holden J, Zhang L, Gaze WH, et al. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat Commun. 2020;11(1):5494.

41. Hill R, Buggs RJA, Vu DT, Gaya E. Lifestyle Transitions in Fusarioid Fungi are Frequent and Lack Clear Genomic Signatures. Mol Biol Evol. 2022;39(4).

Figures

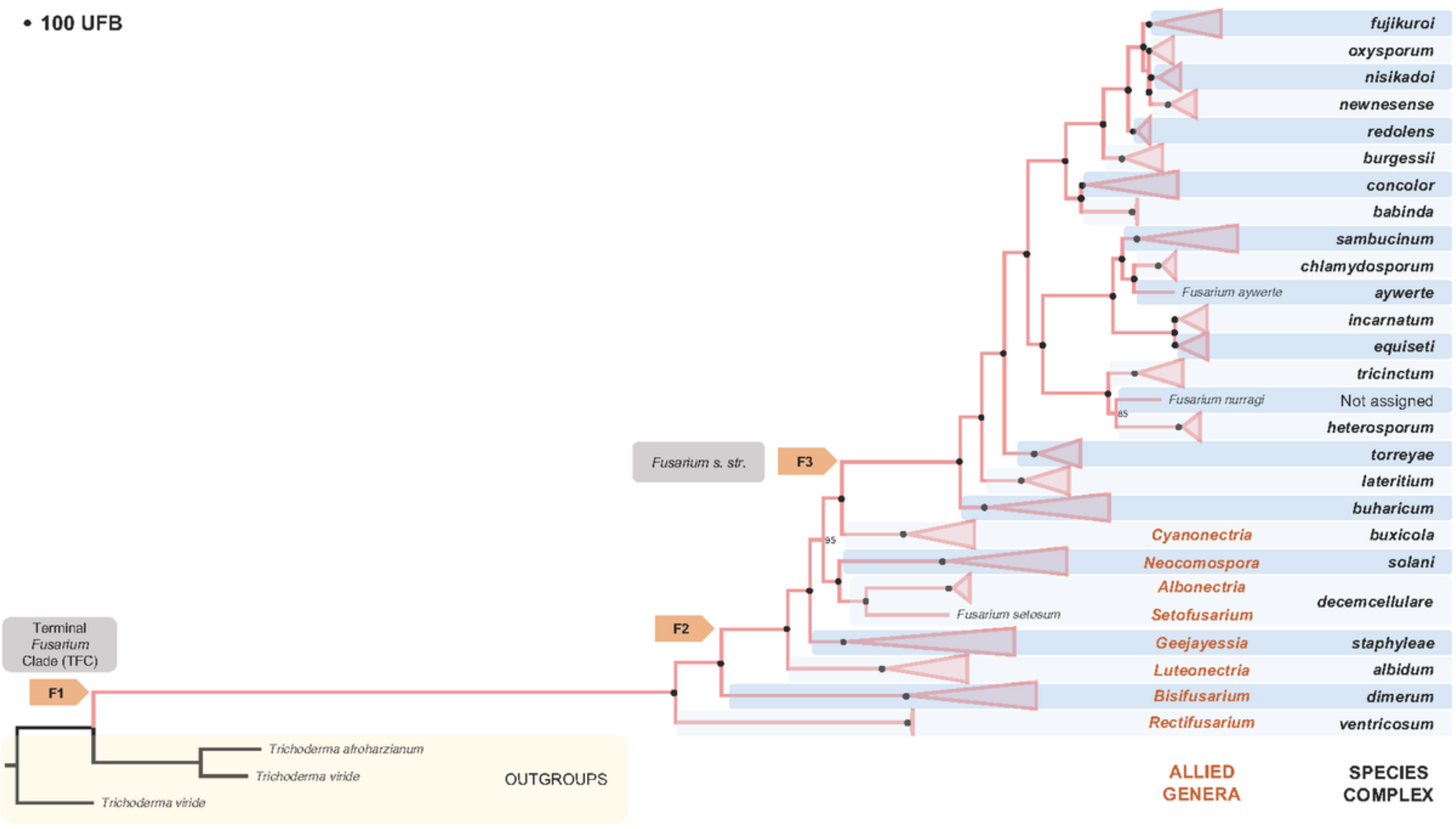

Figure 1

Maximum-likelihood phylogenomic tree based on 559 single-copy conserved proteins. The tree collapsed at the established *Fusarium* species complexes. The black circle (·) denotes 100% UFB support. The UFB value is indicated in the nodes with support below 100. Trichoderma was included as an outgroup.

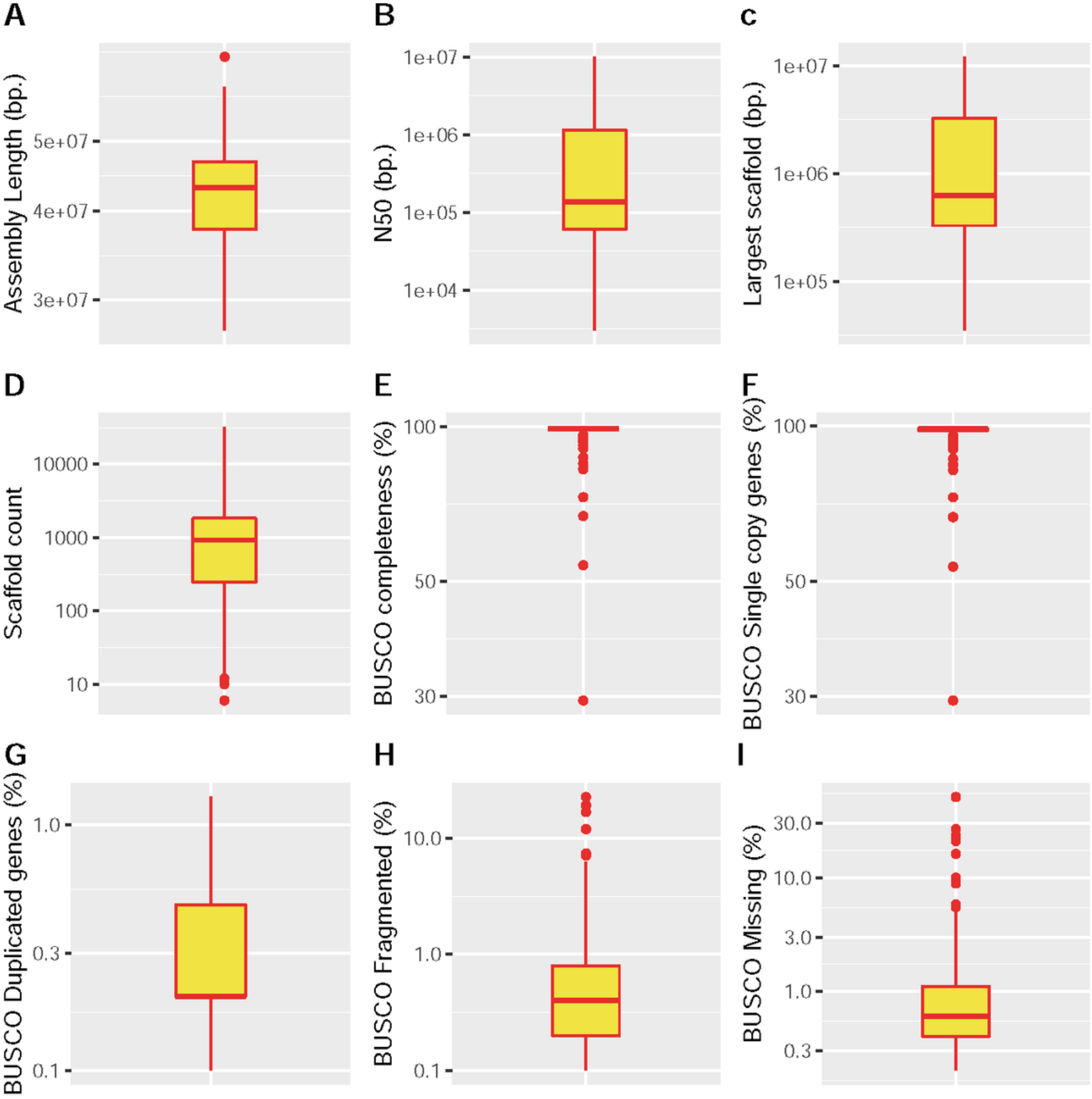

Figure 2

Box plot analysis of the *Fusarium* reference genome quality metrics of all 225 genome assemblies downloaded from the NCBI datasets database. The analyzed metrics include **A.** Assembly length, **B.** Assembly N50, **C.** Largest scaffold length in bp., **D.** Scaffold count, **E.** BUSCO genome completeness, **F.** BUSCO single copy genes detected, **G.** BUSCO duplicated genes, **H.** BUSCO fragmented genes, and **I.** BUSCO missing genes. Outliers are presented as black dots.

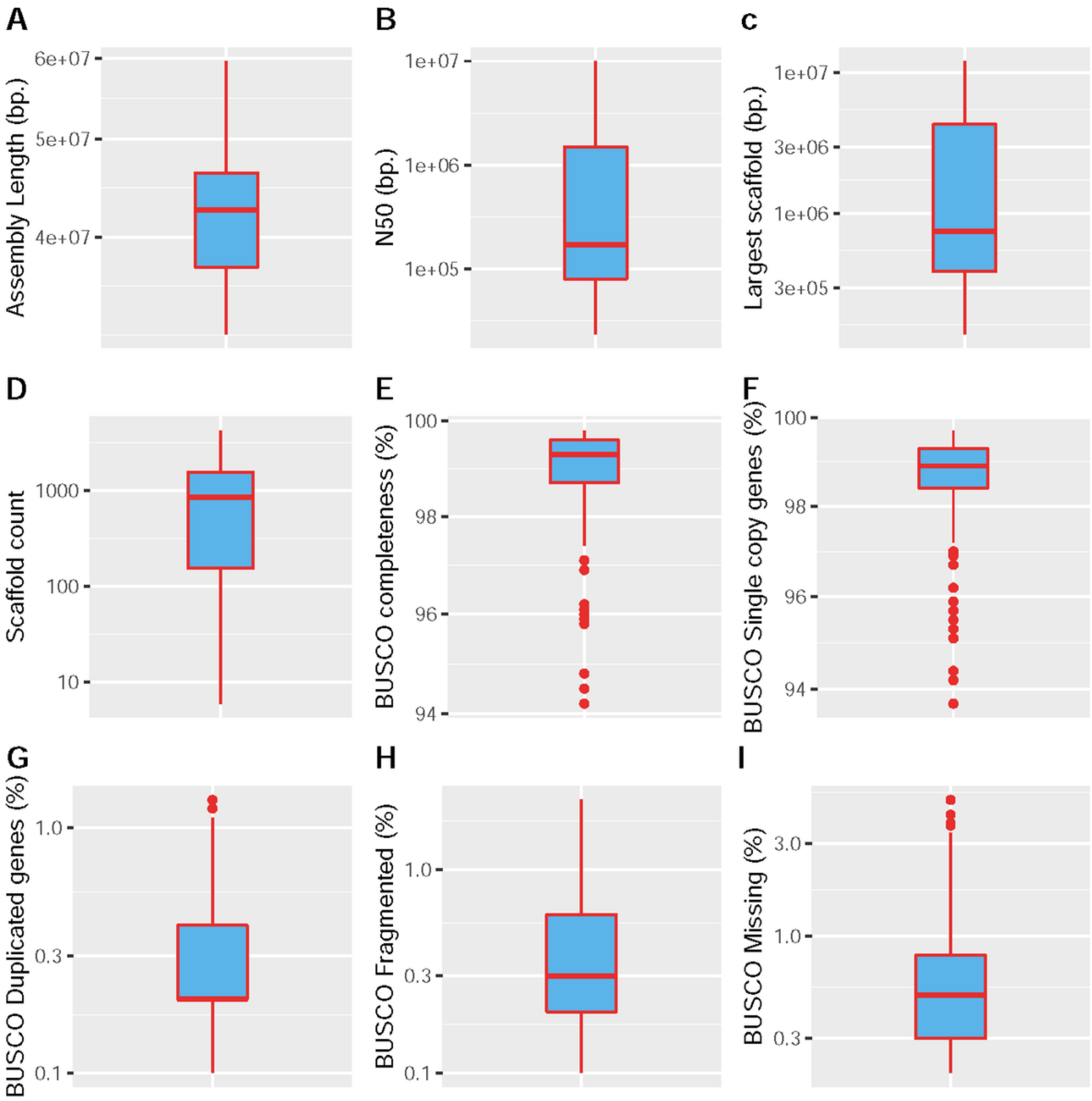

Figure 3

Box plot analysis of the *Fusarium* reference genome quality metrics after removing 24 low-quality genomes. **A.** Assembly length, **B.** Assembly N50, **C.**Largest scaffold length in bp., **D.** Scaffold count, **E.** BUSCO genome completeness, **F.** BUSCO single copy genes detected, **G.**BUSCO duplicated genes, **H.** BUSCO fragmented genes and **I.** BUSCO missing genes. Outliers are presented as black dots.

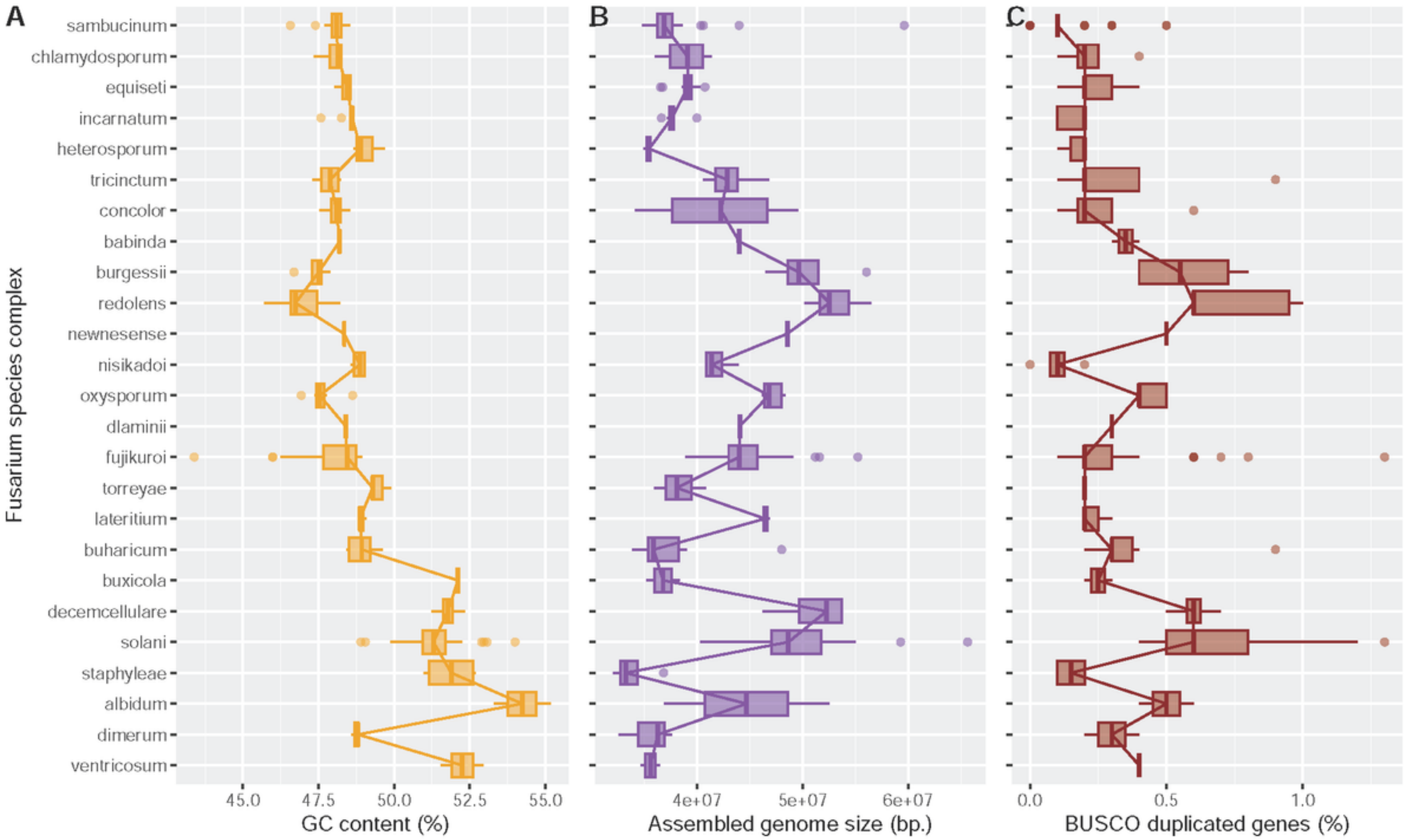

Figure 4

Evolution of GC Content, Genome Size, and Duplications of Core Genes in *Fusarium*. The species complexes are organized as presented in the collapsed tree, with the most ancestral clades positioned at the bottom of the graph. Lines connect the median values of the boxplots. Kruskal-Wallis rank sum test p< 2.2e-16. **A.**Box plots illustrating the evolution of genome GC content. **B.** Box plots depicting the evolution of genome size. **C.** Box plots displaying the evolution of the ratio of duplicated core genes.

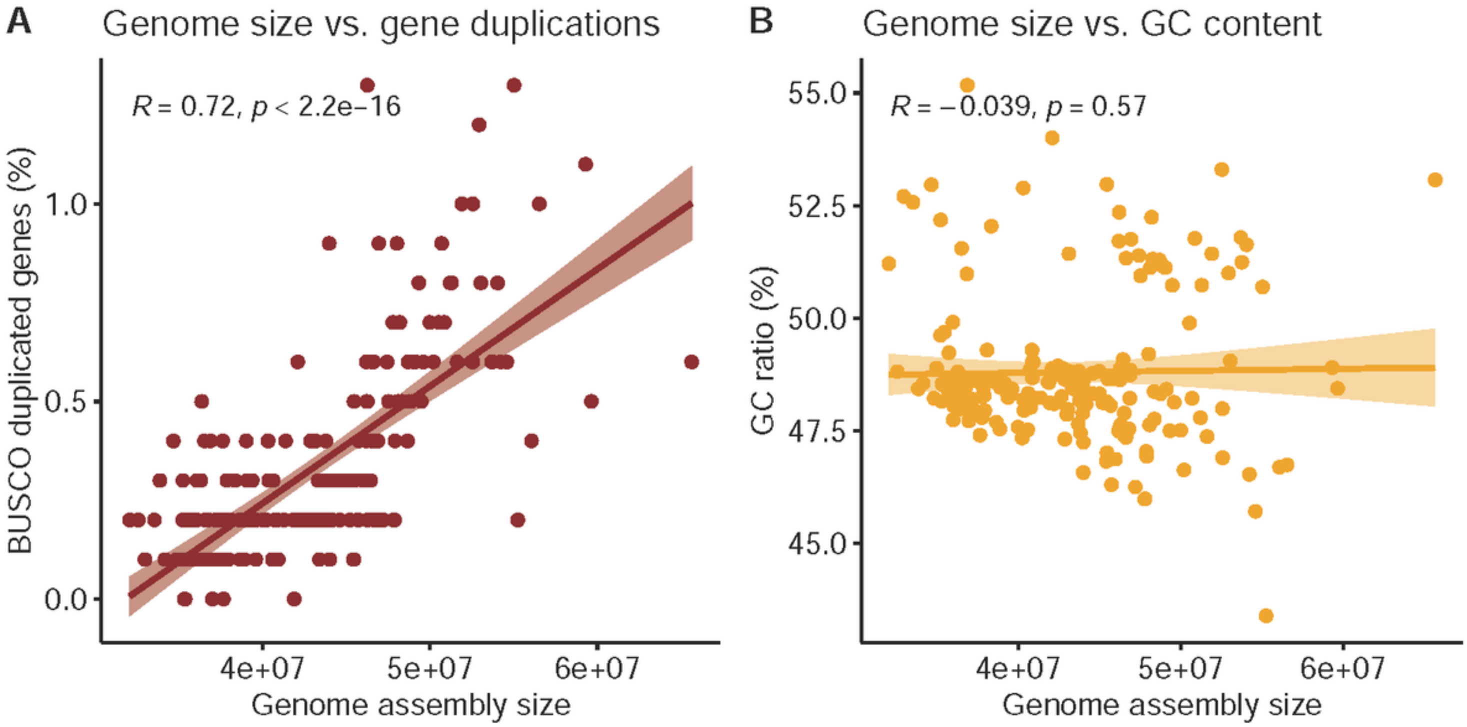

Figure 5

Correlation Analysis Between Genome Size and Duplicated Core Genes in *Fusarium*.

A. Spearman correlation analysis that illustrates the relationship between genome size (x-axis) and the ratio of duplicated core genes (y-axis) in *Fusarium*. Each data point represents a species (p-value < 2.2e-16, rho = 0.720358).

B. Spearman correlation analysis that illustrates the relationship between genome size (x-axis) and GC content (y-axis) in *Fusarium*. Each data point represents a species (p-value = 0.5707, rho = -0.03878989).

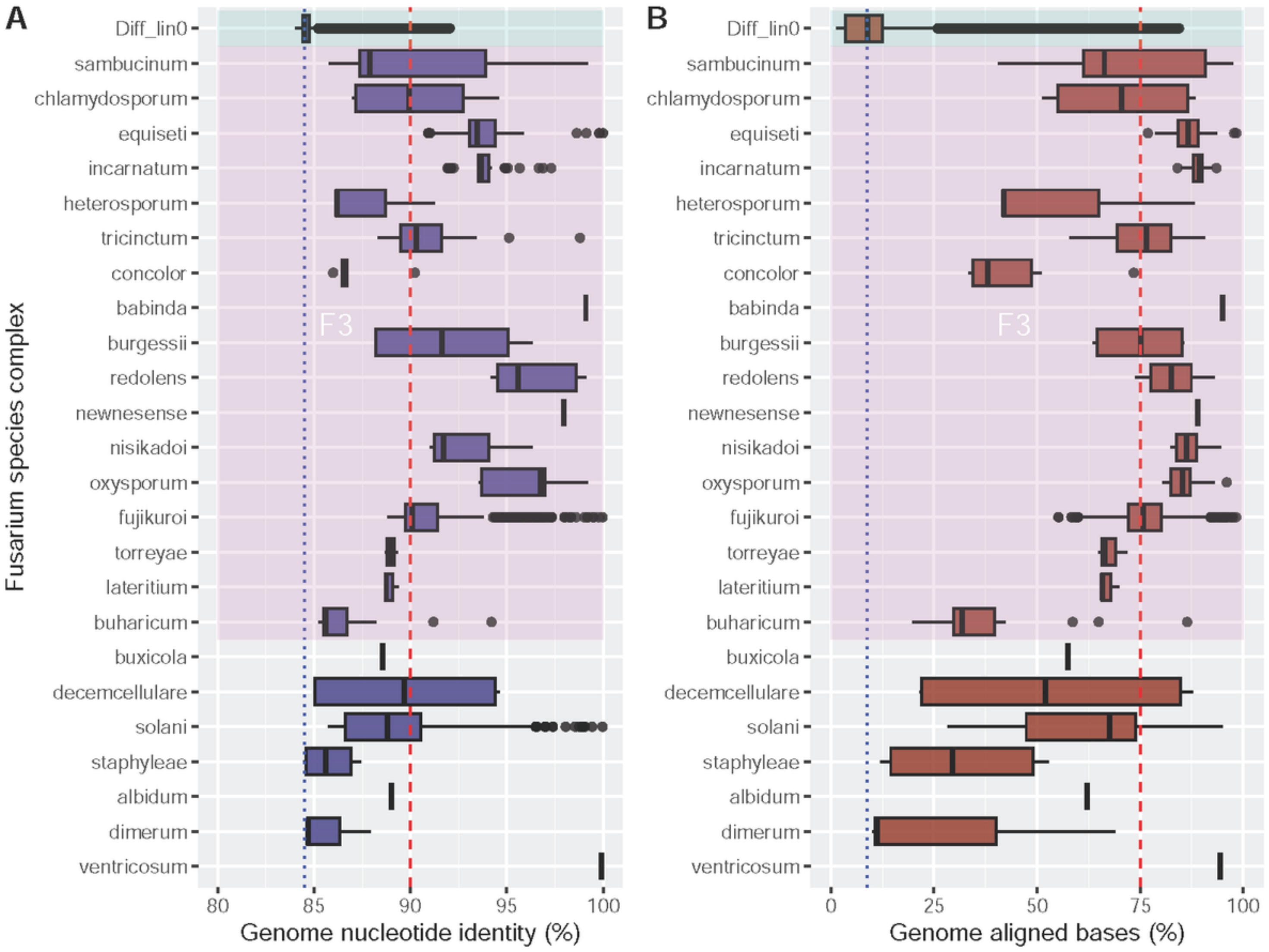

Figure 6

Genome Conservation Among *Fusarium* Species Complexes.

A. Box plots display variations in nucleotide identity across different genomes of the *Fusarium* species complexes. Comparisons between different species complexes are depicted in the 'Diff_lin' category. The species complexes are organized as presented in the collapsed tree, with the most ancestral clades at the bottom of the graph. The dashed red line represents the median value when comparing within the same species complexes, while the dotted blue line signifies the median value for comparisons between species complexes. Kruskal-Wallis rank sum test p< 2.2e-16. The pink rectangle the depicts the species complexes of the *Fusarium* senso stricto clade (F3)

B. Box plots display variations in genome-aligned bases across different genomes of the *Fusarium* species complexes. Comparisons between different species complexes are depicted in the 'Diff_lin' category. The species complexes are organized as presented in the collapsed tree, with the most ancestral clades at the bottom of the graph. The dashed red line represents the median value when comparing within the same species complexes, while the dotted blue line signifies the median value for comparisons between species complexes. Kruskal-Wallis rank sum test p< 2.2e-16. The pink rectangle the depicts the species complexes of the *Fusarium* senso stricto clade (F3)

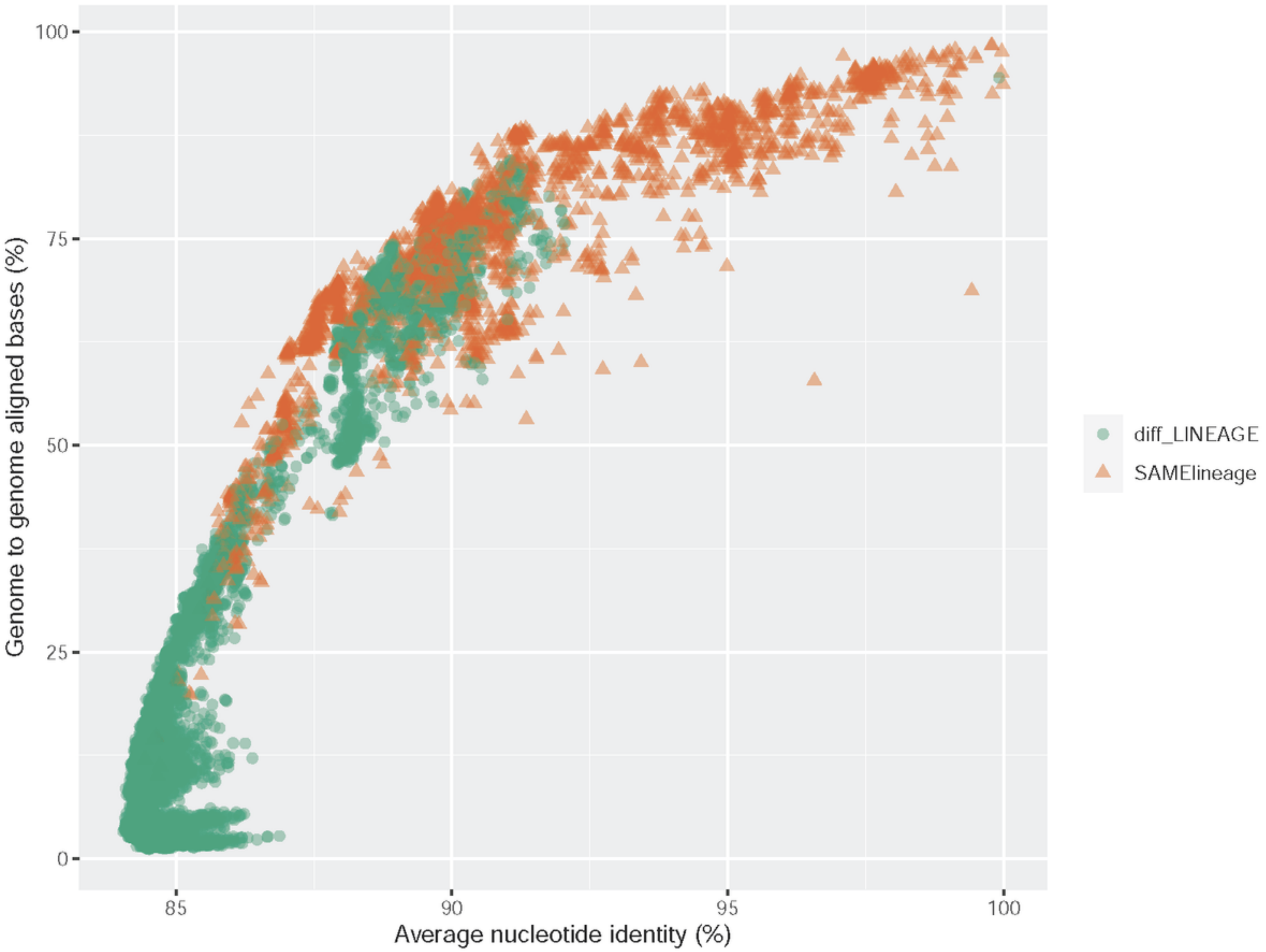

Figure 7

Scatter Plot Analysis of the Relationship Between Nucleotide Identity and Genome-to-Genome Aligned Bases in *Fusarium*.

This scatter plot illustrates the correlation between nucleotide identity and the proportion of genome-to-genome aligned bases across different genomes of the *Fusarium* reference species. Each data point represents a pairwise comparison between two genomes, with nucleotide identity on the x-axis and the proportion of aligned genome bases on the y-axis. The color and shape code distinguishes comparisons within the same species complexes (intra-species-complex, shown in orange triangles, SAMElineage) and comparisons between species complexes (inter-species-complex, shown in green circles, diff_LINEAGE). Kruskal-Wallis rank sum test p< 2.2e-16.

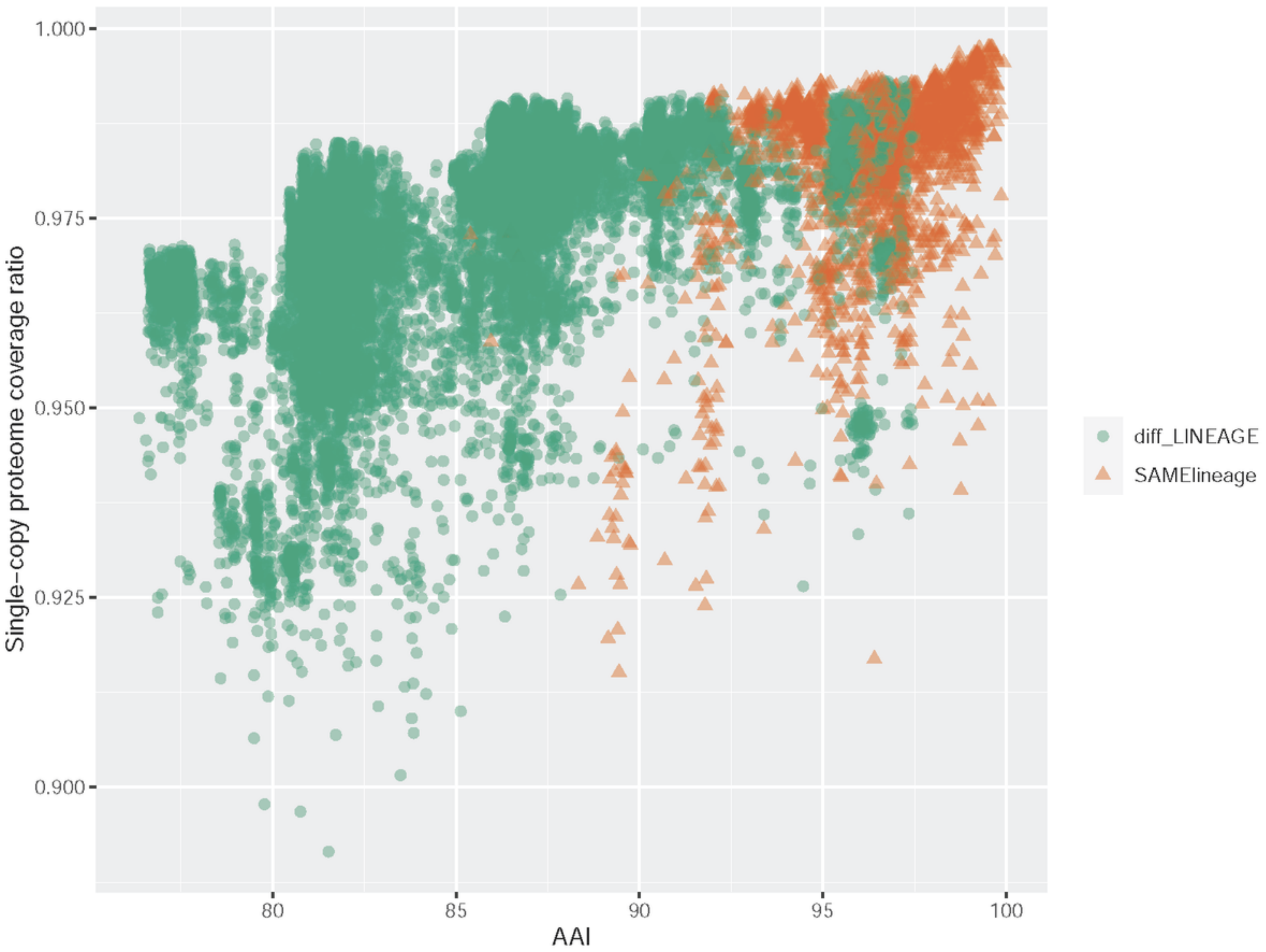

Figure 8

Scatter Plot Analysis of the Relationship Between AAI and Core Proteome Coverage in *Fusarium*.

This scatter plot illustrates the correlation between Average Amino Acid Identity (AAI) and the coverage of the core proteome across different proteomes of the *Fusarium*reference species. Each data point represents a pairwise comparison between two proteomes, with the AAI score on the x-axis and the ratio of the proteome coverage on the y-axis. Data points are color-coded and marked with different shapes to distinguish comparisons within the same species complexes (intra-species-complex, shown in orange triangles, SAMElineage) and comparisons between species complexes (inter-species-complex, shown in green circles diff_LINEAGE). The plot provides insights into the relationship between AAI and core proteome conservation within and between species complexes.

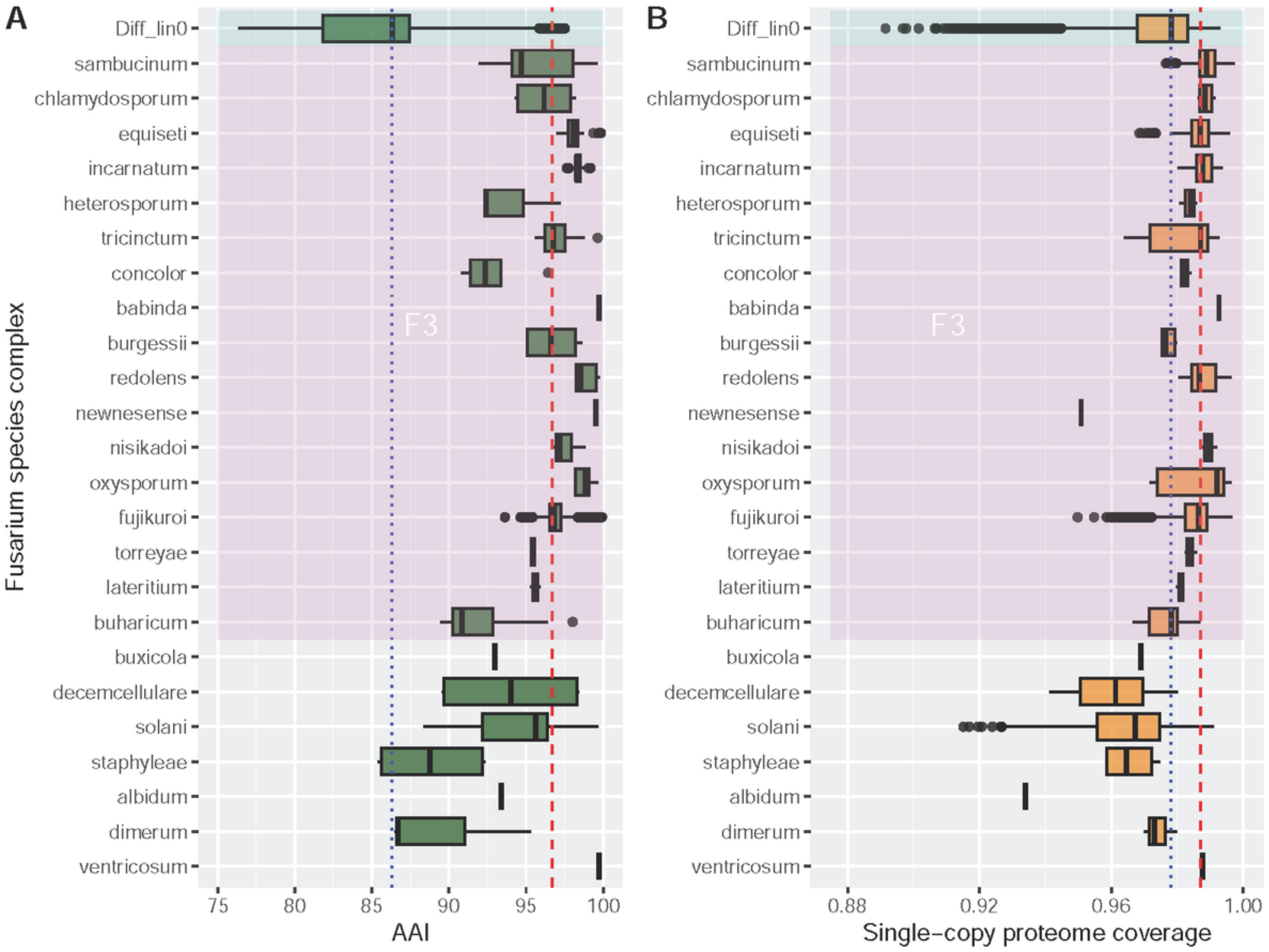

Figure 9

Core proteome Conservation Among *Fusarium* Species Complexes.

A. Box plots display variations in Average Amino Acid Identity (AAI) across different genomes of the *Fusarium* species complexes. AAI comparisons between different species complexes are depicted in the 'Diff_lin' category. The species complexes are organized as presented in the collapsed tree, with the most ancestral clades at the bottom of the graph. The dashed red line represents the median value when comparing within the same species complexes, while the dotted blue line signifies the median value for comparisons between species complexes. Kruskal-Wallis rank sum test p< 2.2e-16. The pink rectangle the depicts the species complexes of the *Fusarium*senso stricto clade (F3).

B. Box plots display variations in the coverage of the core proteome across different genomes of the *Fusarium* species complexes. AAI comparisons between different species complexes are depicted in the 'Diff_lin' category. The species complexes are organized as presented in the collapsed tree, with the most ancestral clades at the bottom of the graph. The dashed red line represents the median value when comparing within the same species complexes, while the dotted blue line signifies the median value for comparisons between species complexes. Kruskal-Wallis rank sum test p< 2.2e-16. The pink rectangle the depicts the species complexes of the *Fusarium*senso stricto clade (F3)

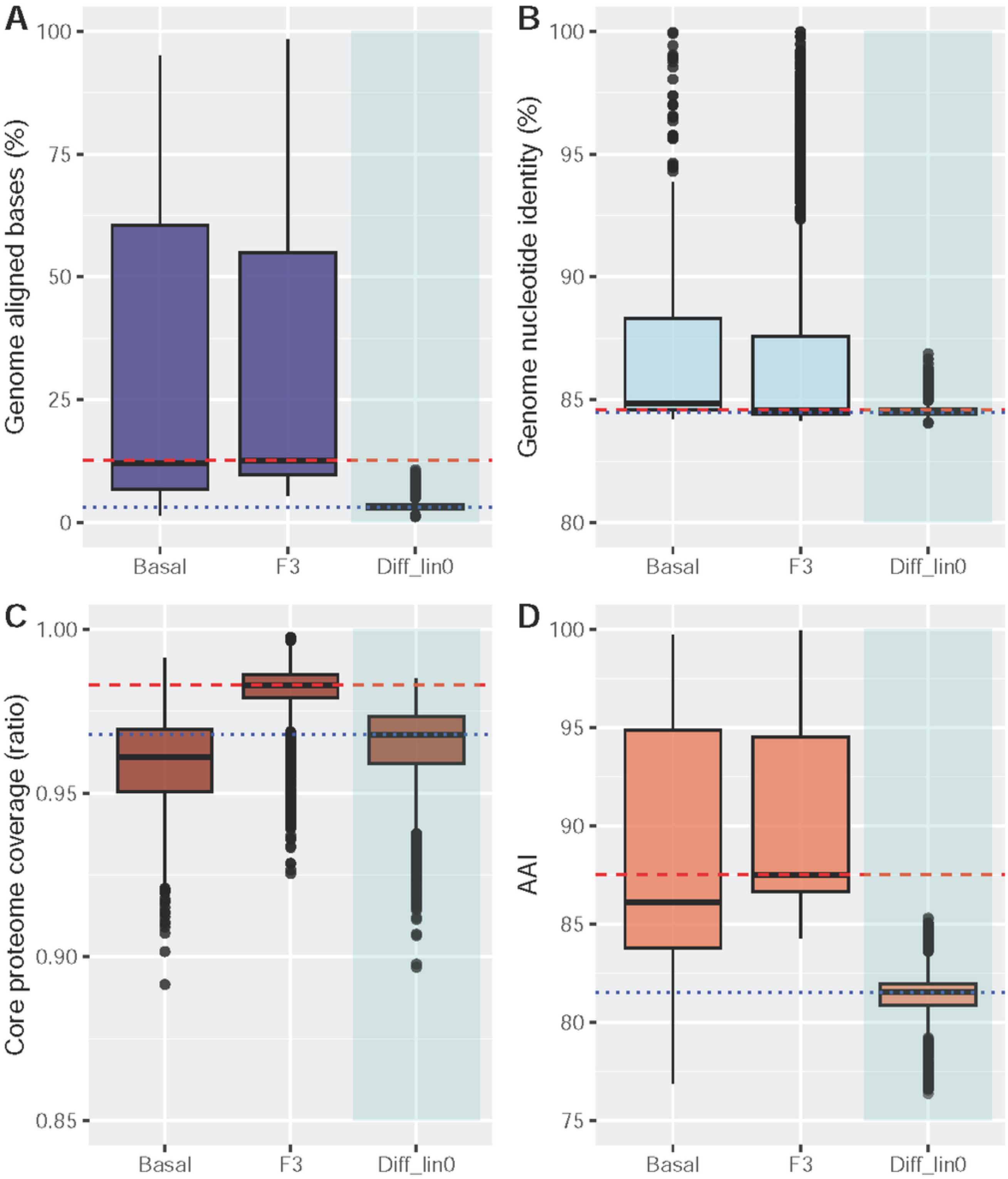

Figure 10

Genome and Core Proteome Conservation Among *Fusarium* Basal (Allied genera and the F3 (*Fusarium* sensu stricto) Species.

Box plots display variations in genome-aligned bases (**A.**) and nucleotide identity (**B.**) across different genomes of the *Fusarium* basal reference species ('allied' genera - **Basal**), the *Fusarium*sensu stricto - F3 reference species (F3), and between the basal and F3 reference species (Diff_lin). The dashed red line represents the median value when comparing within the F3 species, while the dotted blue line signifies the median value for comparisons between F3 and Basal reference species. Kruskal-Wallis rank sum test p< 2.2e-16.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplFig1Fusariumcompletetreerev.pdf
- SupplementaryTable1GenomeData.xlsx
- SupplementaryTable2559OrthologouesProteinsFusarium.xlsx