Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Only a third of Fusarium oxysporum formae speciales are monophyletic: a per-label and per-deposit audit of 363 public genomes

Beyza Akal
Ege University: Ege Universitesi
Yasin Kaymaz
yasin.kaymaz@ege.edu.tr

Ege University: Ege Universitesi https://orcid.org/0000-0002-9725-7536

Research Article

Keywords: Fusarium oxysporum, forma specialis, phylogenomics, genome data quality, effector, Panama disease

Posted Date: September 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10619539/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Only a third of *Fusarium oxysporum* formae speciales are monophyletic: a per-label and per-deposit audit of 363 public genomes

Beyza Akal[1], Yasin Kaymaz[1,*]

[1] Department of Bioengineering, Faculty of Engineering, Ege University, Izmir 35040, Türkiye

*Corresponding author: Yasin Kaymaz (yasin.kaymaz@ege.edu.tr)

ORCID: Beyza Akal https://orcid.org/0009-0008-2365-0504;

Yasin Kaymaz https://orcid.org/0000-0002-9725-7536

Acknowledgements We thank the Scientific and Technological Research Council of Türkiye (TÜBİTAK) for the graduate research stipend awarded to B.A.; Muhammed Bahattin Tanyolaç and Duygu Ateş for valuable discussions and for their expert guidance in the field; and the Graduate School of Natural and Applied Sciences of Ege University, and its Biotechnology graduate programme, for supporting this work.

Abstract

The *forma specialis* system classifies *Fusarium oxysporum* by host of isolation, and multilocus studies have long shown that it conflicts with phylogeny. That conflict has been described label by label but never measured across the whole system, and the public genome collection such measurements now depend on has not been audited deposit by deposit. We retrieved 816 *F. oxysporum* assemblies from NCBI, ran BUSCO uniformly, and inferred a maximum-likelihood phylogeny of 363 genomes from 1,455 strict-core loci (795,943 amino-acid positions). Of 36 formae speciales with at least two genomes, only 12 (33.3%) were monophyletic, every one of them with fewer than 12 representatives; among the ten best-sampled labels only *f. sp. lactucae* was, and the largest fragmented into seven to eleven independent lineages. Rarefaction, label permutation and deduplication left this unchanged: drawn down to three genomes, a large *forma specialis* had at most a 0.30 chance of appearing monophyletic. Twenty-eight genomes (7.7%) had a sister group carrying a different label. The panel was also redundant: 39 genomes (10.7%) carried no independent core-

genome information. Six of the 22 deposits labelled *f. sp. cubense* were Fol4287-derived, separating from the genuine ones with no overlap in whole-genome distance, assembly size or effector content; three are reported here for the first time. One further genome was not *F. oxysporum* at all. None of these would be removed by a completeness filter. We release a curated 324-genome panel and argue that *forma specialis* is an operational pathological category requiring independent evidence.

Keywords *Fusarium oxysporum* · *forma specialis* · phylogenomics · genome data quality · effector · Panama disease

Introduction

The *Fusarium oxysporum* species complex (FOSC) contains soil saprophytes, endophytes and the causal agents of vascular wilt on more than 100 plant hosts, making it one of the most economically damaging groups of soil-borne fungal pathogens (Dean et al., 2012; Gordon, 2017). Because members of the complex are morphologically indistinguishable and largely asexual, they have for a century been classified by host range: an isolate pathogenic on tomato is *f. sp. lycopersici*, one pathogenic on banana is *f. sp. cubense*, and so on (Michielse & Rep, 2009). More than 100 formae speciales have been described (Edel-Hermann & Lecomte, 2019), and because the category is defined by a pathogenicity test rather than by a character state, discriminating them molecularly has been a persistent practical problem (Kistler, 1997; Lievens et al., 2008). The concept itself, and the inconsistency with which it has been applied across the genus, has recently been reviewed in depth by Han et al. (2025).

The *forma specialis* system was never intended as a phylogenetic classification, and multilocus sequence data show that it is not one. The classic demonstration is *f. sp. cubense*, shown by translation elongation factor and mitochondrial genealogies to have at least five independent evolutionary origins (O'Donnell et al., 1998); comparable polyphyly has since been reported for many other labels (Baayen et al., 2000; van Dam et al., 2016). The genomic mechanism became clear with the reference genome of *F. oxysporum* f. sp. *lycopersici* strain 4287, in which roughly a quarter of the assembly consists of accessory, repeat-rich lineage-specific chromosomes that carry the effector genes required for host colonisation and can transfer horizontally between strains in culture (Ma et al., 2010). Host specificity

55 therefore resides in a mobile compartment while the core genome records vertical descent, so two
56 strains sharing a host may be phylogenetically remote and two nearly identical strains may differ in
57 host range (Ma et al., 2013; Rep & Kistler, 2010).

58 What has changed recently is scale. Genome-scale analyses of the complex now routinely
59 involve hundreds of assemblies: McTaggart et al. (2021) inferred phylogenetic species and detected
60 recombination from 40 single-copy orthologues across 410 genomes; Brenes Guallar et al. (2022)
61 clustered 537 public genomes on effector presence and absence alongside a core phylogeny; van
62 Westerhoven et al. (2024) inferred a maximum-likelihood backbone from 4,213 BUSCO loci across
63 618 quality-filtered genomes; and Lopez-Jimenez et al. (2026) have proposed a phylogenomic
64 framework from 4,286 orthologue coding sequences. Our panel is comparable to this body of work in
65 scale and smaller than several members of it, and we make no claim to the contrary.

66 What distinguishes the present study is its object rather than its size. Those studies infer FOSC
67 phylogeny in order to ask about recombination, effector repertoires or organellar evolution, and the
68 correspondence between *forma specialis* and phylogeny is noted in passing rather than measured.
69 Brenes Guallar et al. (2022) state that many formae speciales are polyphyletic and give clade counts
70 for two labels, and van Dam et al. (2018) report lineage counts for four; but no study has reported a
71 monophyly verdict and an independent-lineage count for *every* represented label under one stated
72 criterion, asked whether that result survives the fact that monophyly is arithmetically easier to satisfy
73 in a small group, or paired the measurement with a per-deposit audit of the public panel itself. Two
74 questions in particular have gone largely unasked. First, how much of the public panel is actually
75 independent? Sequencing projects commonly deposit many isolates from one outbreak, and re-
76 assemblies accumulate; redundancy inflates apparent support and biases any downstream summary.
77 Second, how reliable are the *forma specialis* labels attached to public deposits, given that they are
78 supplied by submitters and rarely re-verified?

79 Here we address these questions with a single uniform analysis. We retrieved every *F.*
80 *oxysporum* assembly available from NCBI, applied one BUSCO protocol across the panel, and
81 inferred a partitioned maximum-likelihood phylogeny from the strict core. We then use that tree to test

the monophyly of each *forma specialis* and count its independent lineages; to ask whether the answer survives rarefaction, label permutation and deduplication; to identify genomes whose nearest relatives carry a different label and resolve those conflicts where effector content can; and to establish what the failing deposits actually are.

We are explicit about what is and is not new. That formae speciales are polyphyletic is established (Baayen et al., 2000; O'Donnell et al., 1998; van Dam et al., 2016) and has been reported for individual labels on genome-scale data (Brenes Guallar et al., 2022; van Dam et al., 2018); that public genome collections require quality filtering is likewise known, and the most directly comparable study removed duplicate and artefactual FOSC deposits before its own analysis (Brenes Guallar et al., 2022). Our contribution is to make both systematic: a verdict and lineage count for every represented label under one criterion, tested against the null models that could explain it away, and an audit that names the deposits at fault and establishes what each of them is.

Materials and methods

Genome retrieval, metadata and panel selection

All *F. oxysporum* genome assemblies listed in NCBI at the time of analysis were retrieved (816 accessions), and the organism string as deposited was recorded. *Forma specialis* was extracted from that string by regular expression on the epithet following `f. sp.`, ignoring any trailing strain designation; 117 of the organism strings in the final panel carry a strain suffix, so matching on the full organism name would split single formae speciales into many apparent categories.

The phylogenetic panel comprises the 362 genomes carrying a *forma specialis* designation for which a BUSCO run completed, plus *Fusarium fujikuroi* IMI 58289 (GCF_900079805.1) as outgroup: 363 genomes. This is not a random sample of NCBI's holdings, because genomes without a label carry no information for the test at issue; the consequence is stated under Limitations.

BUSCO analysis and sequence quality control

BUSCO v6.0.0 (Tegenfeldt et al., 2025) was run in genome mode against the `hypocreales_odb10` lineage dataset (Manni et al., 2021; n = 4,494 markers), parallelised with GNU

Parallel (Tange, 2026). Across the 363 genomes of the final panel, completeness had a median of 99.7% (range 94.5–99.8%), with 322 genomes at or above 99.5%.

Single-copy BUSCO protein sequences were then filtered per sequence. A protein was rejected if it was empty, contained an internal stop codon, or contained more than 5% X residues; terminal stop characters were stripped. Rejections were logged per genome.

Core locus selection, alignment and phylogenetic inference

Across the 363 genomes, 4,493 distinct BUSCO loci were detected. A locus was retained for the primary analysis only if a quality-passing single-copy protein was present in all 363 genomes, which retained 1,455 loci (32.4%). Each locus was aligned with MAFFT v7.526 `--auto` (Katoh & Standley, 2013) and trimmed with trimAl v1.5.rev1 `-automated1` (Capella-Gutiérrez et al., 2009); trimmed alignments shorter than 50 aa were discarded. Accepted alignments were concatenated into an amino-acid supermatrix of 795,943 positions with one initial partition per locus. Because every locus is present in every genome by construction, the matrix has no locus-level missing data.

Inference used IQ-TREE v3.1.2 (Wong et al., 2026) with `-st AA -p <partitions> -m MFP+MERGE -rcluster 10 -B 1000 -alrt 1000 -bnni -T 32 -seed 1337`. ModelFinder (Kalyaanamoorthy et al., 2017) selected substitution models and merged partitions with similar evolutionary behaviour, reducing 1,455 initial partitions to 508. Branch support was assessed by ultrafast bootstrap (Hoang et al., 2018) and SH-aLRT (Guindon et al., 2010), 1,000 replicates each. Partition-model inference used the terrace-aware data structure of Chernomor et al. (2016). The tree was rooted post hoc on the outgroup.

Branch-length precision. The rooted tree is written with five decimal places, which rounds most terminal branches in this dataset to zero. All branch-length statistics were therefore computed on the full-precision unrooted treefile; only topology and support were taken from the rooted tree.

Locus-set sensitivity

Requiring 100% locus presence makes the matrix hostage to its worst genome. To quantify the cost, the locus-collection, alignment and trimming pipeline was repeated at a relaxed threshold of ≥

362/363 genomes, yielding 2,843 loci and 1,592,856 amino-acid positions. Both matrices are released so that the choice can be re-examined.

Monophyly, sister-group and rarefaction testing

For each *forma specialis* represented by ≥ 2 genomes we tested whether a node exists whose leaf set is exactly the set of genomes carrying that label. Where no such node exists, the number of independent lineages was counted as the number of *maximal* clades composed exclusively of members of that label, plus any members not contained in such a clade. A monophyletic *forma specialis* therefore has one lineage; a value of 8 means the label is spread over eight parts of the tree separated by genomes carrying other labels.

Separately, for every tip we examined its sister group. Where the focal genome carried a *forma specialis* designation, every member of its sister group carried one and the same designation, that designation differed from the focal genome's, and the subtending node had ultrafast bootstrap ≥ 95, the case was recorded as a sister-group label conflict. Requiring the sister group to be uniform in its label is what makes the conflict interpretable: a sister group carrying several different labels licenses no inference about either genome's assignment.

Monophyly is arithmetically easier to satisfy in a small group, so the association between sampling depth and fragmentation is confounded by construction. To separate the two we asked, for each *forma specialis* with ≥ 10 genomes, what the probability is that a subsample of k of its genomes would be recovered as monophyletic on the full 363-tip tree under the criterion of the preceding paragraph: the most recent common ancestor of the drawn genomes must subtend no genome carrying a different label. Genomes carrying the same label but not drawn are permitted inside the clade, since they cannot make the drawn set non-monophyletic. That condition is met exactly when every drawn genome falls inside a single exclusive clade of the label, so the probability is analytic, $P = \Sigma i\ \mathrm{C}(c_i_, k)/\mathrm{C}(n, k)$, where $c_i_$ are the sizes of the label's exclusive clades and n its total. We report these exact probabilities rather than Monte Carlo estimates; the identity was confirmed against exhaustive enumeration of every subset at $k = 2$ and $k = 3$, which reproduces it to machine precision for all six labels with more than twenty genomes. A random baseline was obtained by drawing k genomes from

the 362 ingroup genomes irrespective of label, 1,000 replicates per *k*. Two further controls were run on the unrarefied test: a fixed-margin label permutation, in which the 46 labels were reassigned at random over the 362 ingroup tips preserving the number of genomes per label, over 100,000 permutations; and a repeat of the whole analysis on the curated non-redundant panel.

Redundancy detection

IQ-TREE reports sequences identical across the whole alignment. These reports were resolved into groups by union–find, giving sets of genomes whose 795,943-position core sequences are indistinguishable. For the curated panel one representative per group was retained, the genome with the most quality-passing single-copy BUSCO proteins, ties broken by accession. The definition is deliberately conservative: genomes differing at a handful of core positions carry no independent signal either, but only exact matches are counted here.

SIX effector screen

All 363 assemblies were screened for the *Secreted In Xylem* family with tblastn (BLAST+ v2.16.0) using full-length SIX1–SIX15 reference proteins retrieved from GenBank, at $E < 10^{-5}$. References annotated by GenBank as partial were replaced with full-length equivalents, since a fixed query-coverage threshold against a truncated reference lowers the effective detection footprint.

A gene was scored present when the union of query positions covered by high-scoring pairs at ≥ 60% amino-acid identity against a single subject sequence spanned ≥ 50% of the reference protein length. The union rule is necessary rather than cosmetic: *SIX* genes contain introns, so tblastn against genomic DNA returns coding exons as separate and sometimes overlapping high-scoring pairs, and summing their lengths, the conventional shortcut, drives apparent coverage above 100% for ten of the fifteen genes in this panel and generates false positives. Requiring coverage to accumulate on one subject sequence prevents a gene from being assembled out of unlinked fragments of a poorly contiguous assembly.

The identity threshold was calibrated on a positive control rather than assumed. *SIX* alleles diverge substantially between formae speciales, so a threshold set too high scores true genes absent: at

≥ 70% identity the screen misses SIX1 and SIX6 in Fol4287 itself, even though both return near-full-length hits on chromosome 14 (query coverage 1.00 at 65.1% identity and 0.98 at 60.6% respectively). At ≥ 60% the screen recovers exactly the fourteen *SIX* genes known from Fol4287 and correctly scores only SIX4 absent, consistent with strain 4287 belonging to race 2, in which *SIX4*/*AVR1* is deleted. We therefore use 60%, and note that all prevalence figures below are lower bounds on the true prevalence of more divergent alleles.

Whole-genome distance

Because neither Mash, fastANI nor skani was available on the analysis machine, whole-genome distances were computed with a dependency-free MinHash implementation following Ondov et al. (2016): canonical k-mers (k = 21) were hashed with a 64-bit mixing function and the s = 20,000 smallest hashes retained as a sketch. The Jaccard index J of two sketches was estimated from the merged bottom sketch and the Mash distance taken as $D = -(1/k)\ \ln(2J/(1 + J))$. D saturates for divergent genomes, so values above roughly 0.2 are reported as "distant" rather than as a precise identity. All 363 × 363 pairwise distances were computed.

Reproducibility

Every table in this manuscript is regenerated from the primary outputs by a single script. Because no phylogenetics library was available on the analysis machine, tree parsing was implemented directly; as a correctness check, an independent second implementation re-derives the full monophyly table from the rooted treefile, returning the same 36 testable labels, the same 12 monophyletic verdicts, and the same lineage counts and clade-size vectors. Software versions are given in **Table S1**.

Results

From 816 public assemblies to a 1,455-locus strict core

The retrieval and filtering cascade is summarised in **Table S2**. Of 816 assemblies retrieved, 366 (44.9%) carried a *forma specialis* designation. Across the 363 genomes of the final panel, BUSCO completeness had a median of 99.7% and 322 genomes reached 99.5% or above: in terms of gene-space completeness the public FOSC panel is uniformly good.

Assembly quality is not. Assembly size ranges from 43.8 to 72.2 Mb (median 53.8), record count from 12 to 13,202 (median 2,225) and N50 from 47.8 to 5,280 kb (median 325 kb), the last distinctly bimodal with 103 genomes below 200 kb and 100 at or above 1 Mb. GC content by contrast is tightly unimodal (median 47.59%, interquartile range 47.45–47.75%), and contiguity is essentially uncorrelated with gene-space completeness (Spearman $\rho = +0.12$ against N50). Uniformly high BUSCO scores therefore conceal a panel whose assemblies differ by three orders of magnitude in contiguity (**Fig. S1**; per-genome values in **Table S3**).

Protein-level quality control told a different story. A mean of 4,445 single-copy BUSCO proteins were recovered per genome but only 4,392 passed sequence QC. The dominant cause was internal stop codons: 18,765 proteins were rejected on that ground, a mean of 51.7 per genome (range 43–83), so roughly 1.2% of the single-copy BUSCO proteins in a typical public assembly carry one; only 356 were rejected for excess ambiguity. The rate is broadly uniform across genomes rather than concentrated in a few poor assemblies, so it appears to be a systematic property of the *ab initio* gene prediction inside BUSCO rather than a marker of assembly quality; but any workflow taking BUSCO proteins at face value inherits it.

Of the 4,493 BUSCO loci detected, 1,455 (32.4%) were present as a quality-passing single-copy protein in every genome and formed the strict core (795,943 aligned positions). The resulting maximum-likelihood tree ($\log L = -3{,}993{,}698.70$; **Fig. 1**, with every tip labelled in **Fig. S2**) had a total length of 0.4497 substitutions per site, 20.2% of it on internal branches. Per-locus alignment statistics are in **Table S4**.

Only one third of formae speciales are monophyletic

Thirty-six formae speciales were represented by at least two genomes and could be tested (**Table 1**, **Fig. 2a**). Twelve (33.3%) were monophyletic.

Monophyly was strongly and inversely associated with sampling depth. Every monophyletic *forma specialis* had fewer than 12 representatives: *f. sp. lactucae* ($n = 11$), *f. sp. tulipae* and *f. sp. zingiberi* (5 each), *f. sp. passiflorae* and *f. sp. radicis-cucumerinum* (4 each), *f. sp. cepae*, *f. sp.*

medicaginis and *f. sp. narcissi* (3 each), and *f. sp. bulbigenum*, *f. sp. luffae*, *f. sp. momordicae* and *f. sp. raphani* (2 each). Among the ten formae speciales with ≥ 10 genomes, only *f. sp. lactucae* was monophyletic.

The larger the group, the more fragmented it was (**Table 1**): *f. sp. melonis* (53 genomes) into 8 independent lineages, *f. sp. vasinfectum* (41) into 9, *f. sp. fragariae* (37) into 11, *f. sp. niveum* (33) into 7, *f. sp. lycopersici* (31) into 8 and *f. sp. cubense* (22) into 8. Two distinct patterns are visible when the members of each label are traced along the tree (**Fig. 2b**). Some labels have a dominant clade with a scatter of outliers: *f. sp. lycopersici* has one clade of 21 genomes and seven other placements, and *f. sp. conglutinans* one clade of 9 and a single outlier. Others are split into large, comparably sized groups: *f. sp. vasinfectum* into clades of 19 and 14, *f. sp. niveum* into 17 and 9, *f. sp. spinaciae* into 6 and 5. The first pattern is consistent with a predominantly clonal *forma specialis* plus a few independent host jumps or mislabelled deposits; the second indicates multiple independent origins of the same host specificity, as documented for *f. sp. cubense* (O'Donnell et al., 1998). The recovery of 8 independent lineages for *f. sp. cubense* from 22 genomes is closely consistent with the ≥ 5 origins inferred from multilocus data and confirms that the earlier conclusion scales to genome-wide data.

That two thirds of labels fail monophyly should not be read as *forma specialis* carrying no phylogenetic information, and a permutation test makes the distinction explicit. Reassigning the 46 labels at random over the 362 ingroup tips while holding the number of genomes per label fixed, the expected number of monophyletic labels is 0.015; 98.6% of 100,000 permutations produced none at all, the maximum over all 100,000 was two, and none came close to twelve ($P < 10^{-5}$). Formae speciales are therefore far more phylogenetically clustered than chance; they are simply not, in most cases, *exactly* clades.

Polyphyly is not an artefact of sampling depth or redundancy

Every monophyletic label in **Table 1** has fewer than twelve genomes, and monophyly is easier to satisfy in a small group, so the association reported above is confounded by construction. Rarefaction separates the two (**Fig. 3a**). A subsample of size k is monophyletic exactly when all k drawn genomes fall inside one exclusive clade, so the probability has a closed form, $P = \Sigma i$ C($c_i_$,

k)/C(n, k) over the label's exclusive clade sizes $c_i_$; we therefore compute it rather than simulate it. Drawn down to k = 3, no large *forma specialis* exceeds a 30% chance of appearing monophyletic: *f. sp. lycopersici* 0.296, *f. sp. melonis* 0.194, *f. sp. niveum* 0.140, *f. sp. vasinfectum* 0.125, *f. sp. cubense* 0.110, *f. sp. fragariae* 0.015. By k = 7 every one is below 0.05 and by k = 15 all are below 0.001. Drawing k genomes at random from the 362 ingroup genomes irrespective of label never produced a monophyletic set in 1,000 replicates at any k.

Two internal controls establish that the test can return high values and is measuring what it claims to. *f. sp. lactucae*, the only well-sampled monophyletic label, has probability exactly 1 at every k, since all eleven of its genomes lie in a single exclusive clade. *f. sp. conglutinans*, which comprises one clade of nine genomes plus a single outlier, decays exactly as that structure dictates: the subsample is monophyletic precisely when the outlier is not drawn, giving $(10 - k)/10$, so 0.700, 0.500 and 0.300 at k = 3, 5 and 7. It is the only label in the panel whose curve is a straight line, and it is straight for that reason: a single contaminating genome makes the probability fall linearly, whereas a label broken into several comparable clades falls geometrically.

The result is also robust to the panel's redundancy. Repeating the whole analysis on the curated 324-genome panel returns the same 36 testable labels, the same 12 monophyletic (33.3%), the same ten well-sampled labels of which the same one is monophyletic, and the same number of independent lineages for every label but one. Only the sizes of the exclusive clades shrink, as they must when identical genomes are removed.

Finally, the strict-core locus definition is worth examining in its own right, because requiring 100% presence makes the matrix hostage to its worst genome. The locus presence spectrum is steep (**Fig. 4c**): 1,455 loci are present in all 363 genomes, but a further 1,388 are present in exactly 362, each missing from a single genome. Relaxing the requirement by one genome therefore almost doubles the matrix, from 1,455 to 2,843 loci and from 795,943 to 1,592,856 amino-acid positions; relaxing it to ≥ 361/363 gives 3,620 loci and to ≥ 360/363 gives 3,921 (**Table S5**). In a panel of this size the strict core discards roughly two thirds of the available phylogenetic signal in order to avoid any missing data, and the loci discarded are not systematically different, being simply the ones that

happened to fail in one arbitrary genome. We retain the strict core for the primary analysis so that no result can be attributed to missing-data artefacts, but investigators building similar frameworks should report the presence spectrum rather than only the final locus count, and should consider whether 100% presence buys enough to justify the cost.

Nearest relatives frequently carry a different label

A stronger test than monophyly asks what a genome's closest relative actually is. Twenty-eight genomes, 7.7% of the panel, had a sister group in which every member carried one and the same *forma specialis* label, different from the focal genome's, at ultrafast bootstrap ≥ 95 (**Table S6**; **Fig. S3**). Seven were reciprocal pairs, representing 14 tips and seven branches of the tree; the remaining 14 tips were single genomes sister to a larger, uniformly differently-labelled clade. The most striking cases involve a single genome sister to a large clade of another *forma specialis*: GCA_002234085.1 (*f. sp. melongenae*) is sister to 31 *f. sp. melonis* genomes at 99.9/97; GCA_016164145.2 (*f. sp. fragariae*) to 14 *f. sp. vasinfectum*, GCA_054130785.1 (*f. sp. phaseoli*) to 8 *f. sp. melonis*, GCA_001703035.1 (*f. sp. lycopersici*) to 4 *f. sp. radicis-cucumerinum* and GCA_009602625.1 (*f. sp. vasinfectum*) to 4 *f. sp. passiflorae*, the last four all at 100/100.

Each has two readings the core genome alone cannot distinguish: a genuine host jump within a clonal lineage, which is plausible because host specificity moves with accessory chromosomes independently of the core, or an error in the deposited metadata. Since effector presence/absence carries information the core genome does not (van Dam et al., 2016), we screened the whole panel for the *SIX* family and return to these genomes below.

One genome in ten carries no independent core-genome information

Fifty-nine of the 363 genomes fell into 20 groups whose 795,943-position core sequences are exactly identical (**Fig. 4d**; **Table S7**). Retaining one representative per group leaves 324 genomes, so 39 genomes (10.7% of the panel) are redundant at the core-genome level. Redundancy was concentrated in a few submissions: one group of 9 *f. sp. vasinfectum* genomes, one of 7 *f. sp. melonis*, one of 4, five groups of 3 and twelve of 2. Most are presumably multiple isolates from one outbreak,

or a clonal lineage sequenced repeatedly. That is legitimate practice, but it must be accounted for in any analysis treating genomes as independent observations.

Nineteen of the 20 groups were internally consistent in their label. The exception is notable: GCA_000149955.2, the *F. oxysporum* f. sp. *lycopersici* strain 4287 reference genome, has a core sequence identical to GCA_011316005.3 and GCA_016802225.1, both deposited as *F. oxysporum* f. sp. *cubense*. Because 795,943 amino-acid positions are involved, this is not a resolution artefact.

Six *f. sp. cubense* deposits carry the tomato-pathogen effector complement

Whole-genome comparison and effector profiling resolve that case decisively (**Table 2**). Both *f. sp. cubense* deposits have a whole-genome MinHash Jaccard index of exactly 1.000 against Fol4287; all three assemblies contain exactly 88 sequence records and differ in total length by 224 bp (61,471,697 bp against 61,471,473 bp), with the records simply renumbered; and all three carry an identical fourteen-gene *SIX* complement. The two deposits are Fol4287, not *f. sp. cubense*.

Extending the comparison to the whole label shows that they are not isolated cases. The 22 deposits carrying the *f. sp. cubense* label separate into two groups with no overlap on any of three independent axes (**Table 2**): MinHash distance to Fol4287 (six genomes at 0.000–0.016, sixteen at 0.023–0.028), assembly size (61.4–63.2 Mb against 44.9–51.1 Mb) and *SIX* gene count (13–14 against 0–6). The six also reproduce Fol4287's assembly architecture, with 85 or 88 sequence records and an N50 of 4.59–4.78 Mb, whereas the sixteen span 12 to 2,807 records. Because the three axes are computed from independent properties of the data and agree exactly on membership, any cut-off placed inside the gaps selects the same six genomes.

Three of the six are already known and must be credited. Brenes Guallar et al. (2022) describe GCA_016802225.1, together with GCA_016802195.1 and GCA_016802205.1, as *in silico* MISAG hybrids in which *f. sp. cubense* sequencing data were imputed onto the Fol4287 reference, and excluded all three from their 537-genome panel. Our screen recovers all three independently.

The other three are identified here. GCA_011316005.3 was published as an Indian *f. sp. cubense* race 1 (VCG 0124) genome whose accompanying announcement reports SIX1, SIX5, SIX9 and

SIX13 (Thangavelu et al., 2021); the deposited assembly carries nine effectors the announcement does not report and is whole-genome indistinguishable from Fol4287, so the assembly in the public record does not correspond to the strain its own genome announcement describes. GCA_052818615.1 matches Fol4287's record count (88), assembly length (61.46 against 61.47 Mb) and N50 to within 0.02%, at a MinHash distance of 0.004, closer to Fol4287 than either confirmed MISAG hybrid, while not being core-identical to it, the signature expected of a hybrid assembly built on the Fol4287 scaffold. GCA_014282265.3 shares the same 88-record architecture and thirteen of the fourteen Fol effectors but is 1.8 Mb larger, so we record it as Fol4287-derived without inferring how it was produced. Note that GCA_052818615.1 carries the second-deepest root-to-tip distance in the tree and is the sister of GCA_014770115.1, the genome shown below not to be *F. oxysporum*; the pairing of the two longest branches in the panel is the classic signature of long-branch attraction, so its position in the tree should not be over-interpreted. The evidence that it is Fol4287-derived does not rest on that position.

Six of the 22 genomes deposited as *f. sp. cubense* (27%) are therefore Fol4287-derived and carry no information about banana-pathogenic *F. oxysporum*. Removing the four whose provenance is documented leaves the headline result unchanged at 12 of 36 formae speciales monophyletic; only the internal structure of *f. sp. cubense* changes, from 22 genomes in 8 lineages with a largest exclusive clade of 11 to 18 genomes in 9 lineages with a largest clade of 7. Removing all six gives 16 genomes in 11 lineages, none larger than 2. That is, once the Fol4287 derivatives are excluded, the *f. sp. cubense* deposits in public sequence databases contain almost no repeated core-genome lineage at all.

One genome is not *Fusarium oxysporum*

The core genome of this complex is extremely shallow (**Fig. 4a**). The median terminal branch length was 2.26×10^{-6} substitutions per site; 31.4% of terminal branches were $\leq 10^{-6}$, 69.1% $\leq 10^{-5}$ and 81.5% $\leq 10^{-4}$ (**Table S8**). The single conspicuous exception was GCA_014770115.1 (*f. sp. capsici*), on a terminal branch of 0.2708, which is 60% of the entire tree length and 10^5-fold longer than the median.

That genome is not *F. oxysporum*. Four independent lines of evidence agree. Its whole-genome MinHash distance to Fol4287 is 0.2564 at a Jaccard index of 0.0023, so it shares essentially no 21-mers with the reference; it is further from *F. oxysporum* than the outgroup is, since *F. fujikuroi* IMI 58289 sits at 0.0845 (Jaccard 0.0926) from the same reference, placing it outside the *F. fujikuroi* species complex rather than merely outside the FOSC. Its GC content of 51.05% is 9.3 standard deviations above the panel mean and the highest in the set. It carries no detectable *SIX* gene. And NCBI's own assembly record flags it as atypical, with the warning "unverified source organism". Its BUSCO statistics are unremarkable, at 4,450 raw and 4,391 quality-passing single-copy proteins, which is precisely why no gene-space filter caught it: it is a good assembly of the wrong organism. We did not attempt to determine the species, and recommend that it be excluded from FOSC panels rather than flagged; we retain it here only so that the effect of its inclusion is documented, and removing it changes no conclusion.

Support across the tree was correspondingly uneven (**Fig. 4b**). Of 341 internal branches carrying support values, 190 (55.7%) reached ultrafast bootstrap ≥ 95 and 178 (52.2%) met the joint criterion SH-aLRT ≥ 80 and UFBoot ≥ 95. Roughly half the backbone is well supported and roughly a quarter effectively unresolved. This is the expected signature of a shallow, densely sampled complex, and a limit of the data rather than the method: with 795,943 positions already in the matrix, adding loci will not resolve branches whose true length is near zero.

SIX effector content is largely decoupled from core-genome descent

Screening all 363 genomes recovers the fourteen-gene reference profile of Fol4287 and shows the family to be unevenly distributed (**Fig. 3b**; per-effector counts and the reference proteins used in **Table S9**; per-genome calls in **Table S10**). SIX13 (56.2%), SIX11 (49.9%), SIX1 (48.2%), SIX6 (44.9%) and SIX8 (35.3%) are the common members; the remaining ten effectors are present in 9.6–27.5% of genomes. The per-genome distribution is strongly bimodal: 11 genomes (3.0%) carry none and 202 (55.6%) carry one to three, but a discrete cluster of 36 genomes (9.9%) carries twelve or more, separated from the rest by a gap at nine to eleven effectors that no genome in the panel occupies

(**Fig. 3c**). That cluster consists entirely of genomes labelled *f. sp. lycopersici* (30) and *f. sp. cubense* (6), the six discussed above.

Effector content and core divergence are close to orthogonal. Across all 65,341 ingroup pairs, the correlation between pairwise *SIX* Jaccard distance and pairwise core patristic distance is $r = 0.103$ ($P = 0.001$ by 999 permutations; permuted $r = -0.001 \pm 0.029$). The correlation is detectable only because the number of pairs is large: it accounts for about 1% of the variance. This is the two-compartment model of *F. oxysporum* genome evolution (Ma et al., 2010, 2013) stated as a measurement rather than an assumption.

Whether polyphyletic lineages share effectors is label-specific. Comparing mean pairwise *SIX* similarity between separate core-genome lineages of the same label against that within a lineage (**Fig. 3d**), *f. sp. mori* (1.00), *f. sp. conglutinans* (0.96), *f. sp. albedinis* (0.85), *f. sp. lycopersici* (0.84) and *f. sp. melonis* (0.80) retain most or all of their effector similarity across phylogenetically separate lineages, the pattern expected when host specificity is carried on a compartment that moves between distantly related genomes. *f. sp. vasinfectum* (0.09) and *f. sp. basilici* (0.00) show the opposite, their lineages sharing almost no effectors, as independent origins of host specificity or label error would produce. We therefore do not claim a single mechanism: both appear to operate, and which applies is a property of the individual *forma specialis*.

A fifteen-gene effector profile does not out-predict the core genome. In leave-one-out nearest-neighbour prediction of the deposited label, core patristic distance is correct for 0.864 of the 352 testable genomes and *SIX* profile distance for 0.699. (Ten genomes are not testable because their deposited label is carried by no other genome in the panel.) The gap is not an artefact of the panel's redundancy, which was the obvious confound: on the curated panel the figures are 0.856 and 0.684, and excluding core-identical neighbours from the candidate pool gives 0.866 and 0.696. This qualifies rather than contradicts the established result that effector profiles separate formae speciales better than *multilocus* phylogeny (van Dam et al., 2016): against a 795,943-position core-genome phylogeny, the fifteen *SIX* genes are simply too low-dimensional to compete. Effector-based assignment in the FOSC

uses hundreds of *mimp*-associated candidates (Brenes Guallar et al., 2022), not the *SIX* family alone, and our result should be read as a statement about *SIX* rather than about effectors in general.

Effector profiles nonetheless resolve most of the label conflicts. For each conflicting genome we asked whether its effector profile is closer to other genomes carrying its deposited label or to genomes carrying its sister group's label. Four of the 28 cannot be assessed this way because their deposited label is carried by no other genome, and four more carry no *SIX* gene at all. Of the remaining 20, seventeen are closer to the deposited label and three to the sister group. The clearest case is GCA_001703035.1, deposited as *f. sp. lycopersici* and sister to four *f. sp. radicis-cucumerinum* genomes at 100/100 support: it carries the complete fourteen-gene Fol complement, at mean *SIX* distance 0.06 to other *lycopersici* and 0.79 to *radicis-cucumerinum*. Its label is correct and its core-genome position genuinely lies among another *forma specialis*, a clean demonstration that the two axes are independent. On the evidence available, the majority of the sister-group conflicts we report are biology rather than metadata error.

A curated non-redundant panel

We provide the deduplicated panel as a reusable resource (**Table S11**): 324 of the 363 genomes, retaining for each core-identical group the genome with the most quality-passing single-copy BUSCO proteins. We recommend that this panel, rather than an unfiltered NCBI download, be used as the reference set for FOSC comparative work, and that the seven genomes resolved above be excluded: GCA_014770115.1, which is not *F. oxysporum*, and the six Fol4287-derived deposits carrying a *f. sp. cubense* label. Only GCA_014770115.1 is flagged by GC content, at 51.05% against a panel median of 47.59%; none of the seven is an outlier in contig count or N50, so none would be caught by a conventional contiguity filter. Assembly size does separate the six Fol4287 derivatives, at 61.4–63.2 Mb against a panel median of 53.8 Mb, but only once the *f. sp. cubense* label makes that size anomalous: a 61 Mb assembly is unremarkable for the panel as a whole. BUSCO completeness comes closest to a general filter and still fails: the four deposits of documented provenance sit at 98.4%, the 1.9th percentile, and the other two at 98.6% and 98.8%, all far above any threshold applied in practice,

and 7 of the 363 genomes score lower while being unremarkable in every other respect. This is the general lesson of the audit: the deposits that most need removing are the ones that look fine.

Discussion

Forma specialis is an operational category, not a phylogenetic one

That formae speciales are polyphyletic is not new, and it has been stated for individual labels on genome-scale data before (Brenes Guallar et al., 2022; van Dam et al., 2018). What our analysis adds is a verdict for every represented label under one criterion, and a test of whether that verdict means anything. The association with sampling depth needed testing rather than interpreting, because monophyly is easier to satisfy in a small group whatever the biology; rarefaction shows the fragmentation is real, since drawn down to the size of the small labels the large formae speciales are still almost never monophyletic while *f. sp. lactucae* remains monophyletic in every replicate. We would predict that *f. sp. lactucae* is the exception that will hold, and that *f. sp. cepae*, *f. sp. narcissi* and the other small groups will fragment as more genomes appear. That prediction rarefaction cannot itself test, since it can show that observed polyphyly is not an artefact of large samples but not what a small label would do if sampled more deeply.

Our effector data bear on *why* the two axes disagree without settling it. Effector content is almost independent of core-genome distance across the panel, which is what a horizontally mobile pathogenicity compartment predicts and is the mechanism usually invoked (Croll & McDonald, 2012; Ma et al., 2010). But the pattern across the polyphyletic labels is not uniform: some retain nearly all their effector similarity across separate core lineages and some retain almost none. A single explanation, transfer of a shared pathogenicity chromosome, does not fit both, and models of accessory-compartment evolution more elaborate than a strict two-speed genome are increasingly required to accommodate this kind of heterogeneity (Torres et al., 2020). The more defensible reading is that the *forma specialis* system aggregates outcomes of several distinct histories, and that treating "polyphyletic" as one phenomenon obscures this.

None of this means the system should be abandoned. It encodes something real and agronomically decisive, namely which crop an isolate can destroy, that no phylogenetic classification

captures, precisely because that trait sits on a horizontally mobile compartment. It fails only when used as a proxy for evolutionary relatedness, which is exactly how a *forma specialis* label in a genome database tends to be read. The constructive position, consistent with the recent reappraisal of the concept by Han et al. (2025), is that *forma specialis* is a pathological designation requiring pathogenicity or effector evidence, that phylogeny is a separate axis, and that public deposits should carry both.

Data quality in public FOSC genome space

Four quality findings emerge, none specific to *F. oxysporum* and all likely to apply to other rapidly growing pathogen genome collections.

First, redundancy: 10.7% of the panel is exactly identical to another member at 795,943 core positions, concentrated in a few large submissions. Analyses weighting genomes equally, such as allele-frequency summaries, population-structure inference and machine-learning training sets, will be biased toward whichever lineages were sequenced most heavily unless this is corrected.

Second, label conflict: 7.7% of genomes have a nearest relative bearing a different *forma specialis*. Some are real biology and some metadata error; the point is that a user cannot tell which from the database record alone.

Third, annotation noise: about 1.2% of the single-copy BUSCO proteins in a typical public assembly carry an internal stop codon. This is small, uniform and easy to filter, but invisible unless checked, and it propagates silently into any concatenation-based analysis that does not screen protein sequences.

Fourth, heterogeneous assembly quality behind homogeneous completeness scores. Gene-space completeness is nearly uniform (median 99.7%) while contiguity spans 12 to 13,202 records and is essentially uncorrelated with it, so BUSCO makes an attractive but misleading inclusion filter. Analyses sensitive to contiguity, such as synteny, accessory-chromosome or repeat content and structural variation, must filter on N50 or record count explicitly.

The *f. sp. cubense* case deserves emphasis because of what it costs. Banana Fusarium wilt is among the most consequential plant diseases now spreading, and *f. sp. cubense* is heavily studied for that reason. A panel assembled by metadata query returns 22 genomes under that label, of which six are Fol4287 derivatives carrying no information about banana-pathogenic *F. oxysporum*. Any study using those deposits to characterise *f. sp. cubense* effector repertoires, define diagnostic markers or train a host-range classifier would have drawn conclusions about a tomato pathogen, and the error is invisible to every filter normally applied, because apart from their size these are good assemblies.

Recommendations

For investigators building FOSC comparative datasets: deduplicate before analysis, using the curated panel provided here; report the locus presence spectrum rather than only the final core size; screen single-copy proteins for internal stops and ambiguity before concatenation; treat deposited *forma specialis* labels as hypotheses and confirm them by effector profiling where the label matters; and compute branch-length statistics from full-precision tree files, since rounded output silently converts short branches to zero.

For depositors: pairing a submission with the pathogenicity or effector evidence behind its *forma specialis* assignment, and flagging re-assemblies of existing strains, would remove most of the ambiguity documented here at negligible cost.

Limitations

The panel is the *forma specialis*-designated subset of available assemblies. This is appropriate for the question asked but means the tree is not a random sample of FOSC diversity, and unlabelled genomes, which include many endophytic and soil isolates, are absent; lineage-level conclusions should not be extrapolated to the complex as a whole. Labels are taken as deposited and were not verified against pathogenicity data: effector profiling adjudicates only 20 of the 28 sister-group conflicts, since four genomes carry no *SIX* gene and four have a deposited label carried by no other genome in the panel. A *SIX* profile is in any case evidence about effector content rather than about host range measured *in planta*, and where a label matters the standard remains a pathogenicity test. Only the core genome and the *SIX* family were analysed; the fifteen *SIX* genes are a small and biased

sample of the effectorome, and our finding that they under-predict *forma specialis* relative to the core genome should not be generalised to effectors as a class. The screen also detects only alleles at least 60% identical to a single reference protein over half its length, so every prevalence figure is a lower bound; the sensitivity of the result to that threshold is shown by the positive control, where raising it to 70% loses two of the fourteen genes Fol4287 is known to carry. Redundancy is defined conservatively as exact identity over the full supermatrix, and whole-genome MinHash shows the true figure is higher. Sampling is uneven across formae speciales, from 1 to 53 genomes; the rarefaction closes this concern in one direction only, since it asks what a large label would look like if sampled shallowly and cannot say whether the small monophyletic labels would survive deeper sampling. A single outgroup was used for rooting; the root position within the ingroup is not a focus of this study and should not be over-interpreted, and a second, more distant outgroup would test it directly. Finally, BUSCO-based sampling is biased toward conserved, slowly evolving loci by construction, which is desirable for deep relationships and contributes to the shallow terminal branches we report.

Conclusions

A uniform BUSCO-based analysis of 362 publicly available, *forma specialis*-designated *F. oxysporum* genomes and one outgroup shows that only a third of testable formae speciales are monophyletic, that among well-sampled labels the proportion falls to one in ten, and that the largest labels fragment into seven to eleven independent lineages. That result is not an artefact of the test: it survives rarefaction to the size of the small labels and deduplication, and the labels are nonetheless far more phylogenetically clustered than random reassignment would produce. Nearly 8% of genomes have a nearest relative carrying a different label, and for 17 of the 20 such cases that effector content can adjudicate, the complement supports the deposited label, making them biology rather than metadata error.

The audit of the panel itself returned findings of the same weight as the phylogenetics. Six of the 22 genomes deposited as *f. sp. cubense* are Fol4287-derived, three of them reported here for the first time and one contradicting its own published genome announcement, and one further genome is not *F. oxysporum* at all. None would be caught by an assembly-quality or completeness filter. Effector

content across the panel is almost orthogonal to core-genome divergence, confirming by measurement that the two compartments record different histories; but a fifteen-gene *SIX* profile predicts the deposited label less well than the core genome does, and effector sharing between separate lineages of one label varies from complete to absent, so no single mechanism accounts for *forma specialis* polyphyly.

We release the framework, per-genome quality metrics, effector profiles and a curated non-redundant panel, and conclude that *forma specialis* should be treated as an operational pathological category to be evidenced independently of phylogeny, and that a public genome panel assembled by metadata query should be treated as a hypothesis about its own contents until audited.

Statements and Declarations

Funding This study received no direct funding. B.A. was supported by a research stipend from the Scientific and Technological Research Council of Türkiye (TÜBİTAK), project number 224S905.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable. This study analysed only publicly available genome assemblies and involved no human participants, animals, or newly collected biological material.

Consent to participate Not applicable.

Consent for publication Not applicable.

Data availability All genome assemblies analysed are publicly available from NCBI; accessions and per-genome metadata are listed in **Table S3**. Analysis code, the derived data tables, the *SIX* effector presence/absence matrix, the 363 × 363 whole-genome MinHash distance matrix, the curated non-redundant panel and the scripts that regenerate every table and figure in this paper are available at https://github.com/BMGLab/fusarium-oxysporum-fsp-audit. The 795,943-position supermatrix, its partition scheme and both maximum-likelihood trees are archived at Zenodo (https://doi.org/10.5281/zenodo.21829290). No new sequence data were generated for this study.

Author contributions Conceptualization: Y.K.; methodology: Y.K. and B.A.; software: Y.K.; formal analysis: B.A.; investigation: B.A.; data curation: B.A.; visualization: B.A.; writing, original draft: B.A.; writing, review and editing: Y.K.; supervision: Y.K.; project administration: Y.K. Both authors read and approved the final manuscript.

Related work A companion study applying the framework built here to newly sequenced *F. oxysporum* isolates from tomato and pepper in Türkiye is in preparation and will be submitted separately. No newly sequenced genome appears in the present manuscript, and no result is shared between the two.

References

Baayen, R. P., O'Donnell, K., Bonants, P. J. M., Cigelnik, E., Kroon, L. P. N. M., Roebroeck, E. J. A., & Waalwijk, C. (2000). Gene genealogies and AFLP analyses in the *Fusarium oxysporum* complex identify monophyletic and nonmonophyletic formae speciales causing wilt and rot disease. *Phytopathology, 90*(8), 891–900. https://doi.org/10.1094/PHYTO.2000.90.8.891

Brenes Guallar, M. A., Fokkens, L., Rep, M., Berke, L., & van Dam, P. (2022). *Fusarium oxysporum* effector clustering version 2: An updated pipeline to infer host range. *Frontiers in Plant Science, 13*, 1012688. https://doi.org/10.3389/fpls.2022.1012688

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology, 65*(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Croll, D., & McDonald, B. A. (2012). The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathogens, 8*(4), e1002608. https://doi.org/10.1371/journal.ppat.1002608

Dean, R., Van Kan, J. A. L., Pretorius, Z. A., Hammond-Kosack, K. E., Di Pietro, A., Spanu, P. D., Rudd, J. J., Dickman, M., Kahmann, R., Ellis, J., & Foster, G. D. (2012). The Top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology, 13*(4), 414–430. https://doi.org/10.1111/j.1364-3703.2011.00783.x

Edel-Hermann, V., & Lecomte, C. (2019). Current status of *Fusarium oxysporum* formae speciales and races. *Phytopathology, 109*(4), 512–530. https://doi.org/10.1094/PHYTO-08-18-0320-RVW

Gordon, T. R. (2017). *Fusarium oxysporum* and the Fusarium wilt syndrome. *Annual Review of Phytopathology, 55*(1), 23–39. https://doi.org/10.1146/annurev-phyto-080615-095919

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology, 59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Han, S., Zhao, P., Wang, M., Jiang, X., Liu, F., Huang, J., & Cai, L. (2025). What is *Fusarium* forma specialis? *Mycology, 16*(4), 1522–1557. https://doi.org/10.1080/21501203.2025.2528352

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution, 35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods, 14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution, 30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kistler, H. C. (1997). Genetic diversity in the plant-pathogenic fungus *Fusarium oxysporum*. *Phytopathology, 87*(4), 474–479. https://doi.org/10.1094/PHYTO.1997.87.4.474

Lievens, B., Rep, M., & Thomma, B. P. H. J. (2008). Recent developments in the molecular discrimination of formae speciales of *Fusarium oxysporum*. *Pest Management Science, 64*(8), 781–788. https://doi.org/10.1002/ps.1564

Lopez-Jimenez, J., Daza, J. M., & Alzate, J. F. (2026). Towards a phylogenomic framework for the *Fusarium oxysporum* species complex. *International Journal of Molecular Sciences, 27*(14), 6255. https://doi.org/10.3390/ijms27146255

Ma, L.-J., van der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P. M., Kang, S., Shim, W.-B., Woloshuk, C., Xie, X., Xu, J.-R., Antoniw, J., Baker, S. E., Bluhm, B. H., … Rep, M. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature, 464*(7287), 367–373. https://doi.org/10.1038/nature08850

Ma, L.-J., Geiser, D. M., Proctor, R. H., Rooney, A. P., O'Donnell, K., Trail, F., Gardiner, D. M., Manners, J. M., & Kazan, K. (2013). *Fusarium* pathogenomics. *Annual Review of Microbiology, 67*(1), 399–416. https://doi.org/10.1146/annurev-micro-092412-155650

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution, 38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

McTaggart, A. R., James, T. Y., Shivas, R. G., Drenth, A., Wingfield, B. D., Summerell, B. A., & Duong, T. A. (2021). Population genomics reveals historical and ongoing recombination in the *Fusarium oxysporum* species complex. *Studies in Mycology, 99*, 100132. https://doi.org/10.1016/j.simyco.2021.100132

Michielse, C. B., & Rep, M. (2009). Pathogen profile update: *Fusarium oxysporum*. *Molecular Plant Pathology, 10*(3), 311–324. https://doi.org/10.1111/j.1364-3703.2009.00538.x

O'Donnell, K., Kistler, H. C., Cigelnik, E., & Ploetz, R. C. (1998). Multiple evolutionary origins of the fungus causing Panama disease of banana: Concordant evidence from nuclear and mitochondrial gene genealogies. *Proceedings of the National Academy of Sciences of the United States of America, 95*(5), 2044–2049. https://doi.org/10.1073/pnas.95.5.2044

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology, 17*(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Rep, M., & Kistler, H. C. (2010). The genomic organization of plant pathogenicity in *Fusarium* species. *Current Opinion in Plant Biology, 13*(4), 420–426. https://doi.org/10.1016/j.pbi.2010.04.004

Tange, O. (2026). *GNU Parallel 20260622* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.20959929

Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E. M., & Kriventseva, E. V. (2025). OrthoDB and BUSCO update: Annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research, 53*(D1), D516–D522. https://doi.org/10.1093/nar/gkae987

Thangavelu, R., Edwin Raj, E., Pushpakanth, P., Muthukathan, G., Murugan, L., & Uma, S. (2021). Draft genome resource of a novel virulent *Fusarium oxysporum* f. sp. *cubense* race 1 strain (VCG 0124) infecting Cavendish (AAA) group of banana in India. *Plant Disease, 105*(9), 2708–2710. https://doi.org/10.1094/PDIS-12-20-2737-A

Torres, D. E., Oggenfuss, U., Croll, D., & Seidl, M. F. (2020). Genome evolution in fungal plant pathogens: Looking beyond the two-speed genome model. *Fungal Biology Reviews, 34*(3), 136–143. https://doi.org/10.1016/j.fbr.2020.07.001

van Dam, P., Fokkens, L., Schmidt, S. M., Linmans, J. H. J., Kistler, H. C., Ma, L.-J., & Rep, M. (2016). Effector profiles distinguish formae speciales of *Fusarium oxysporum*. *Environmental Microbiology, 18*(11), 4087–4102. https://doi.org/10.1111/1462-2920.13445

van Dam, P., de Sain, M., ter Horst, A., van der Gragt, M., & Rep, M. (2018). Use of comparative genomics-based markers for discrimination of host specificity in *Fusarium oxysporum*. *Applied and Environmental Microbiology, 84*(1), e01868-17. https://doi.org/10.1128/AEM.01868-17

van Westerhoven, A. C., Dijkstra, J., Aznar Palop, J. L., Wissink, K., Bell, J., Kema, G. H. J., & Seidl, M. F. (2024). Frequent genetic exchanges revealed by a pan-mitogenome graph of a fungal plant pathogen. *mBio, 15*(12), e02758-24. https://doi.org/10.1128/mbio.02758-24

van Westerhoven, A. C., Fokkens, L., Wissink, K., Kema, G. H. J., Rep, M., & Seidl, M. F. (2025). Reference-free identification and pangenome analysis of accessory chromosomes in a major fungal plant pathogen. *NAR Genomics and Bioinformatics, 7*(2), lqaf034. https://doi.org/10.1093/nargab/lqaf034

Wong, T. K. F., Ly-Trong, N., Ren, H., Demotte, P., Baños, H., Roger, A. J., Susko, E., Bielow, C., De Maio, N., Goldman, N., Hahn, M. W., dos Reis, M., Vinh, L. S., Huttley, G., Lanfear, R., & Minh, B. Q. (2026). IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. *Molecular Biology and Evolution, 43*(5), msag117. https://doi.org/10.1093/molbev/msag117

Figure captions

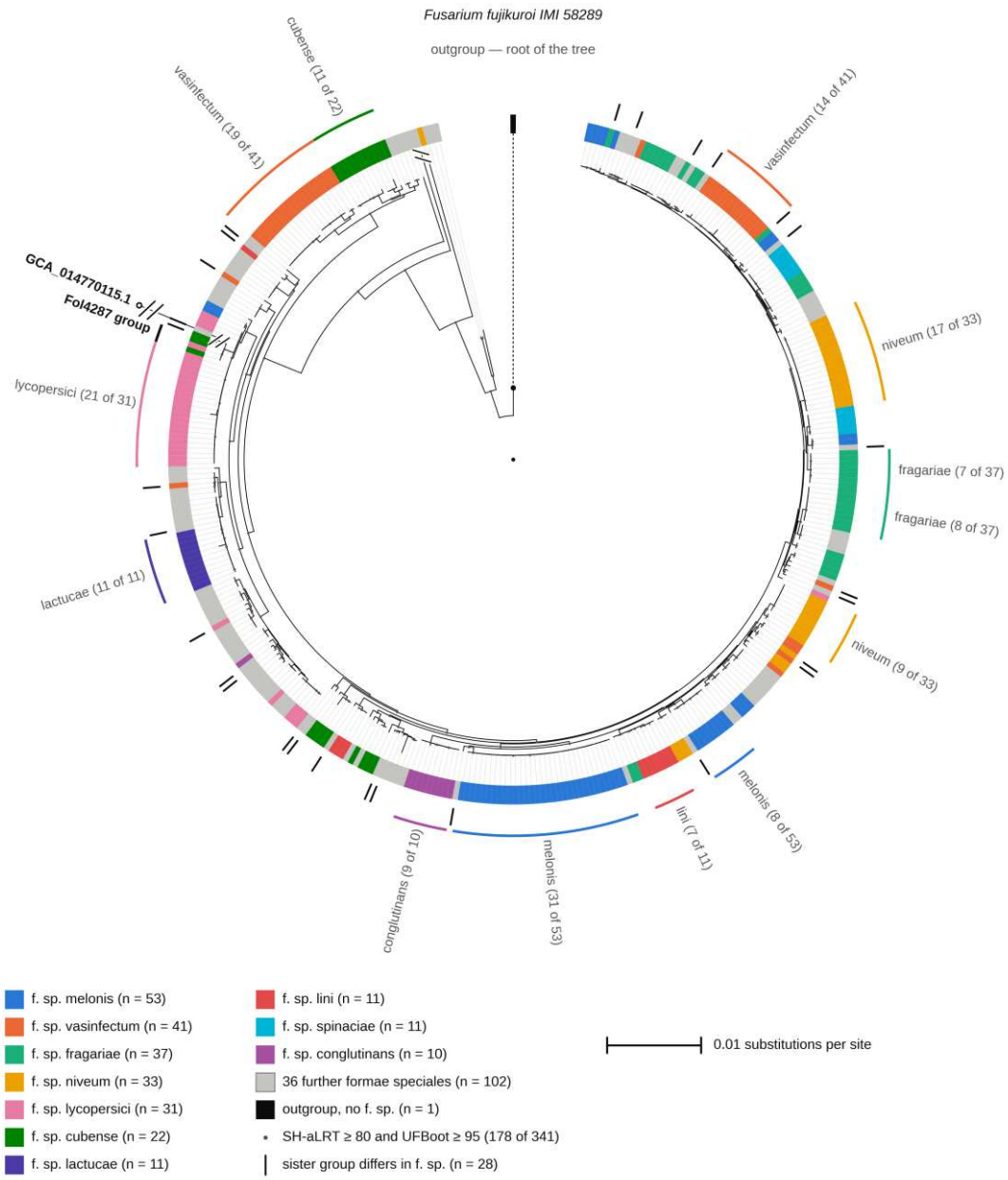


Fig. 1 Maximum-likelihood phylogeny of 363 *Fusarium oxysporum* species complex genomes, with *forma specialis* mapped onto the topology. The tree was inferred from 1,455 strict-core single-copy BUSCO proteins (795,943 amino-acid positions in 508 ModelFinder-merged partitions) and rooted on *Fusarium fujikuroi* IMI 58289, marked at the top of the fan. The outer ring gives the *forma specialis* of each genome as deposited in NCBI: the ten most heavily sampled are coloured and the remaining 36 are grey, with counts in the key. Coloured arcs bracket every maximal clade containing only one *forma specialis* and at least seven genomes, labelled with the epithet, the number of genomes in that clade and the total carrying the label; a label appearing in more than one bracket, or whose bracket accounts for only part of its genomes, is not monophyletic. Among the ten most heavily sampled labels only *f. sp. lactucae* (11 of 11) forms a single clade, whereas *f. sp. melonis*, *f. sp. vasinfectum*, *f. sp. fragariae*, *f. sp. niveum*, *f. sp. lycopersici* and *f. sp. cubense* occupy 8, 9, 11, 7, 8 and 8 lineages respectively. Grey dots mark the 178 of 341 supported internal branches reaching both SH-aLRT ≥ 80 and ultrafast bootstrap ≥ 95. Black ticks outside the ring mark the 28 genomes whose entire sister group carries a different *forma specialis*. The black bracket marks the tomato reference GCA_000149955.2 together with GCA_011316005.3 and GCA_016802225.1, both deposited as *f. sp. cubense*, whose 795,943-position core sequences are identical (Table S7). Radial distance is proportional to substitutions per site over the drawn range 0–0.02871, which contains 360 of the 363 tips. The three tips beyond that range are truncated at a double-slash

break: GCA_014770115.1 (root-to-tip 0.3135, terminal branch 0.2708, 60% of the total tree length), GCA_001888865.1 (0.0430) and GCA_052818615.1 (0.0429). The rooting step assigned the whole basal branch (0.00596 substitutions per site) to the ingroup side, so for display it is drawn split evenly across the root; no topology and no other branch length is altered. Tip labels cannot be resolved at this size and are given in full in Fig. S2

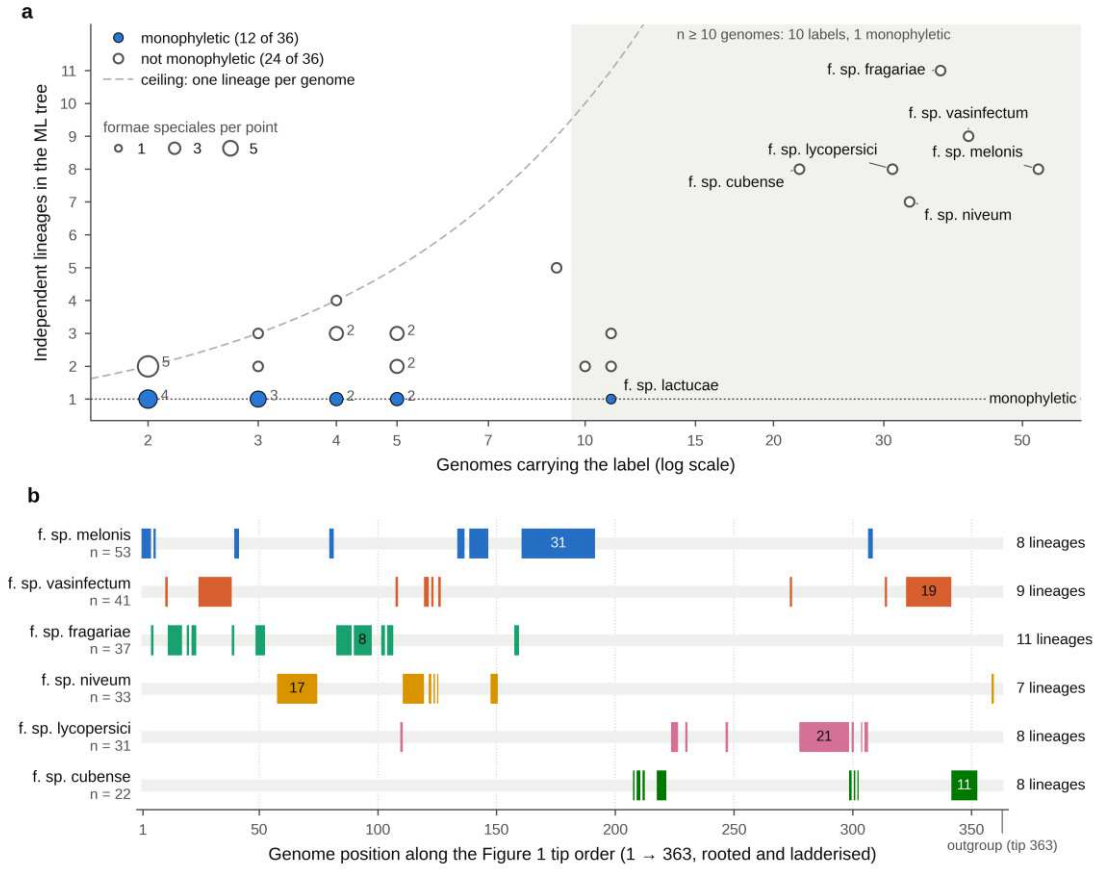


Fig. 2 Polyphyly of formae speciales. (**a**) Number of independent lineages against the number of genomes carrying the label, for the 36 formae speciales with at least two genomes (abscissa log-scaled). An independent lineage is a maximal clade composed exclusively of members of that label, so a monophyletic *forma specialis* has exactly one. Filled symbols, monophyletic (12 of 36); open symbols, not. Symbol area is proportional to the number of labels sharing a coordinate, printed beside every symbol representing more than one. The dashed line is the ceiling $y = n$. No trend line is fitted, because monophyly is more easily satisfied in a small group and part of the association between sampling depth and fragmentation is a property of the test rather than of biology. The shaded field contains the ten labels with $n \geq 10$, of which only *f. sp. lactucae* is monophyletic. (**b**) Where the members of the six largest formae speciales fall in the tip order of Fig. 1, unrolled onto a straight axis. Each bar is one independent lineage drawn across the tips it occupies, so a monophyletic label would appear as a single block; a bar one tip wide is a single genome placed among genomes carrying other labels. White hairlines separate lineages adjacent in tip order whose common ancestor also subtends genomes of other formae speciales. The number inside the widest bar of each row is that label's largest exclusive clade, and the total number of lineages is given at the right

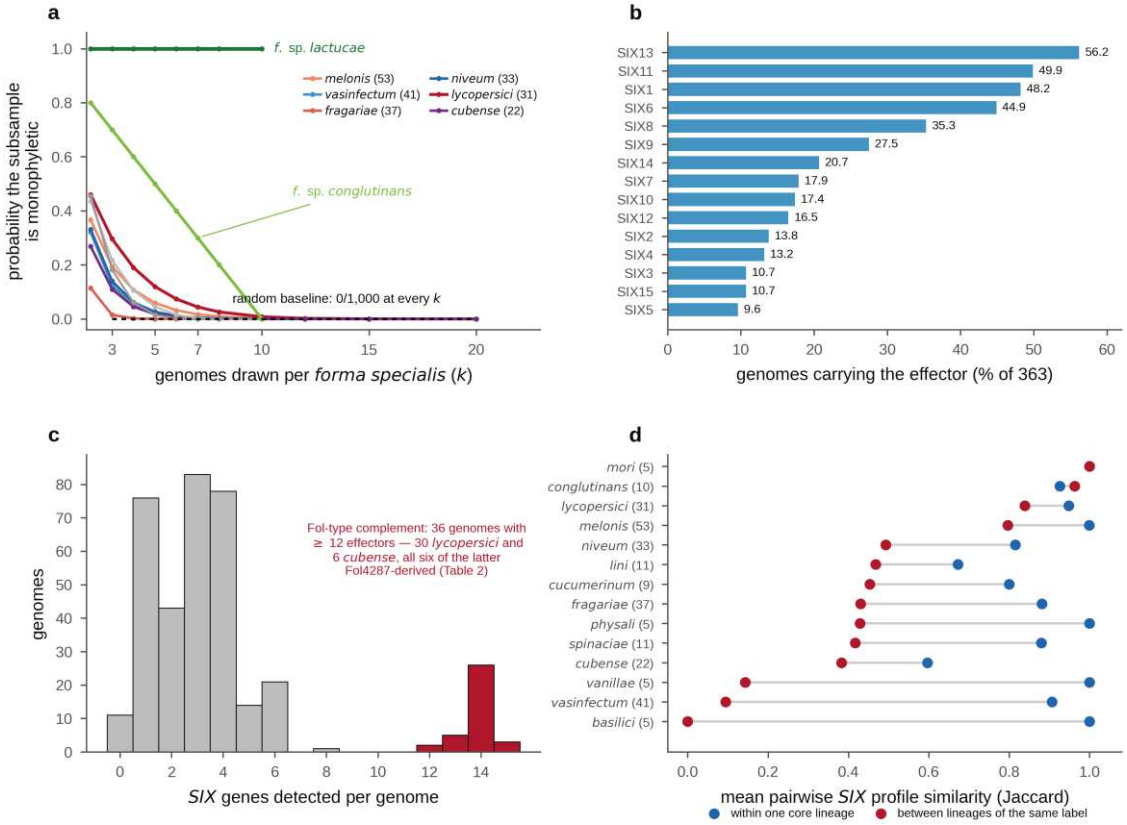


Fig. 3 Rarefaction of the monophyly test, and *SIX* effector content against core-genome descent. (**a**) Exact probability that a subsample of k genomes drawn from one *forma specialis* is recovered as monophyletic on the full 363-tip tree, meaning that the most recent common ancestor of the drawn genomes subtends no genome carrying a different label. The subsample satisfies that condition exactly when all k genomes fall inside one exclusive clade, so the probability is computed in closed form as $\Sigma i\ C(c_i_, k)/C(n, k)$ rather than simulated. Curves are drawn for the ten formae speciales with at least ten genomes and stop at each label's own size; the two annotated directly are the controls. The shape of each curve reports the label's internal structure. *f. sp. lactucae*, the only well-sampled monophyletic label, is flat at 1 because all eleven of its genomes lie in one clade. *f. sp. conglutinans* falls as a straight line because it is one clade of nine plus a single outlier: for $k \geq 2$ the subsample is monophyletic exactly when that one genome is not drawn, and $C(9, k)/C(10, k)$ reduces to $(10 - k)/10$, which is linear in k. The remaining eight labels are convex because their largest exclusive clade is a small fraction of the label, so the ratio falls off geometrically, roughly as $(c/n)^k$. The dashed line is the random baseline, k genomes drawn from the 362 ingroup genomes irrespective of label: 0 of 1,000 replicates at every k. (**b**) Prevalence of each *SIX* effector, as the percentage of the 363 genomes in which it was scored present. (**c**) Number of *SIX* genes per genome. The distribution is bimodal: 11 genomes carry none and 202 carry one to three, but a discrete group of 36 (red) carries twelve or more, comprising 30 genomes deposited as *f. sp. lycopersici* and 6 as *f. sp. cubense*. No genome carries nine, ten or eleven. (**d**) Mean pairwise *SIX* profile similarity (Jaccard) within a single core-genome lineage (blue) against that between separate lineages carrying the same label (red), for the fourteen labels with at least five genomes and more than one lineage, ordered by the between-lineage value; genome counts in parentheses. A short connector indicates a label whose phylogenetically separate lineages nonetheless share most of their effector complement; a long connector indicates lineages sharing almost none

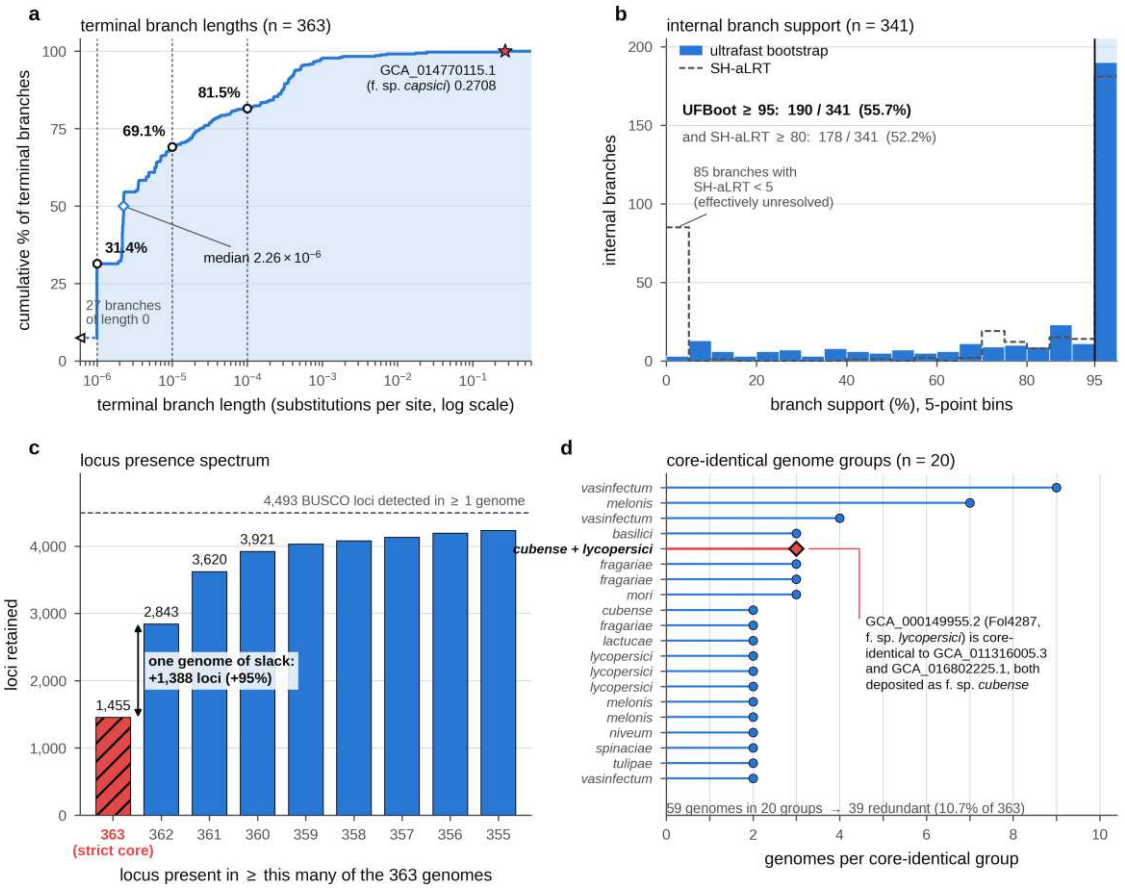


Fig. 4 Resolution and redundancy of the 363-genome public panel. (**a**) Empirical cumulative distribution of the 363 terminal branch lengths, on a logarithmic axis. Branch lengths were taken from the full-precision unrooted treefile, because the rooted treefile is written at five decimal places and rounds most terminal branches to zero. Dotted guides mark 10^{-6}, 10^{-5} and 10^{-4} substitutions per site; open circles and bold labels give the cumulative percentage at or below each. The open diamond marks the median. Twenty-seven branches are exactly zero and cannot be placed on a logarithmic axis; they form the first tread, drawn dotted against the left spine. The star at the right is GCA_014770115.1, whose terminal branch of 0.2708 is 10^5 times the median and the only conspicuous outlier. (**b**) Branch support across the 341 internal branches carrying support values, in 5-point bins. Filled bars, ultrafast bootstrap; dashed outline, SH-aLRT; the shaded region and vertical rule mark the UFBoot 95 threshold. 190 of 341 (55.7%) reach UFBoot ≥ 95 and 178 (52.2%) meet the joint criterion. (**c**) Locus presence spectrum: the number of BUSCO loci recovered as a quality-passing single-copy protein in at least the indicated number of genomes; the dashed line is the 4,493 loci detected in at least one. The strict core (red hatched bar) retains 1,455 loci, whereas allowing a locus to be missing from a single genome retains 2,843; the arrow marks this step. (**d**) Sizes of the 20 groups of genomes whose 795,943-position core sequences are identical, ordered by size and labelled with the *forma specialis* of the group. The groups contain 59 genomes, so 39 carry no independent core-genome information. Nineteen groups are internally label-consistent; the exception, in red, places Fol4287 with GCA_011316005.3 and GCA_016802225.1, both deposited as *f. sp. cubense*

Tables

Table 1 Monophyly of the 36 formae speciales represented by at least two genomes

forma specialis	Genomes	Monophyletic	Lineages	Sizes of exclusive clades
melonis	53	no	8	31;8;4;3;2;2;2;1
vasinfectum	41	no	9	19;14;2;1;1;1;1;1;1
fragariae	37	no	11	8;7;6;4;3;2;2;2;1;1;1

niveum	33	no	7	17;9;3;1;1;1;1
lycopersici	31	no	8	21;3;2;1;1;1;1;1
cubense	22	no	8	11;4;2;1;1;1;1;1
lactucae	11	yes	1	11
lini	11	no	3	7;3;1
spinaciae	11	no	2	6;5
conglutinans	10	no	2	9;1
cucumerinum	9	no	5	3;2;2;1;1
basilici	5	no	2	4;1
mori	5	no	2	4;1
physali	5	no	3	3;1;1
tulipae	5	yes	1	5
vanillae	5	no	3	3;1;1
zingiberi	5	yes	1	5
albedinis	4	no	3	2;1;1
lagenariae	4	no	3	2;1;1
nicotianae	4	no	4	1;1;1;1
passiflorae	4	yes	1	4
radicis-cucumerinum	4	yes	1	4
apii	3	no	2	2;1
cepae	3	yes	1	3
gladioli	3	no	3	1;1;1
medicaginis	3	yes	1	3
narcissi	3	yes	1	3
bulbigenum	2	yes	1	2

capsici	2	no	2	1;1
coriandrii	2	no	2	1;1
luffae	2	yes	1	2
melongenae	2	no	2	1;1
momordicae	2	yes	1	2
phaseoli	2	no	2	1;1
radicis-lycopersici	2	no	2	1;1
raphani	2	yes	1	2

[a] 12 of 36 formae speciales represented by >= 2 genomes are monophyletic (33.3%). The number of independent lineages is the number of maximal clades composed only of members of that forma specialis, plus any members not contained in such a clade, so a monophyletic forma specialis has exactly one. Tested on the rooted maximum-likelihood tree of 363 genomes. The ten formae speciales represented by a single genome cannot be tested and are listed in Table S3.

Table 2 Genomes deposited as *Fusarium oxysporum* f. sp. *cubense* that carry a Fol4287-type *SIX* effector complement

Accession	Deposited f. sp.	MinHash D	Size (Mb)	Records	SIX genes	Assessment
GCA_000149955.2	lycopersici	0.00000	61.47	88	14	reference
GCA_011316005.3	cubense	0.00000	61.47	88	14	Fol4287 re-deposition (core-identical)
GCA_016802225.1	cubense	0.00000	61.47	88	14	Fol4287 re-deposition (core-identical)
GCA_052818615.1	cubense	0.00419	61.46	88	14	Fol4287-derived, new here
GCA_016802195.1	cubense	0.01104	61.38	85	14	MISAG hybrid (Brenes Guallar et al., 2022)
GCA_016802205.1	cubense	0.01229	61.38	85	14	MISAG hybrid (Brenes Guallar et al., 2022)
GCA_014282265.3	cubense	0.01631	63.22	88	13	Fol4287-derived, new here

[a] Every genome deposited as f. sp. cubense that carries a Fol4287-type SIX complement (>= 10 of 15 effectors), with Fol4287 itself as the reference row. These six separate from the sixteen remaining cubense deposits on three independent axes with no overlap: MinHash distance to Fol4287 (0.000-0.016 against 0.023-0.028), assembly size (61.4-63.2 Mb against 44.9-51.1 Mb) and SIX gene count (13-14 against 0-6). Any cut-off inside those gaps selects the same six genomes. Whole-genome MinHash distance: k = 21, s = 20,000 canonical k-mers. Core-identity from IQ-TREE identical-sequence reporting over the 795,943-position supermatrix. Brenes Guallar et al. (2022) describe GCA_016802195.1, GCA_016802205.1 and GCA_016802225.1 as in silico MISAG hybrids in which f. sp. cubense reads were imputed onto the Fol4287 reference; the other three are identified here.

Electronic Supplementary Material

Online Resource 1 (PDF). Supplementary figures. **Fig. S1** assembly statistics of the 363-genome panel; **Fig. S2** the Fig. 1 phylogeny with every tip labelled, across six pages; **Fig. S3** the 28 sister-group *forma specialis* label conflicts.

Online Resource 2 (XLSX). Supplementary tables, one per sheet. **S1** software versions; **S2** dataset construction cascade; **S3** per-genome metadata, assembly statistics and BUSCO quality metrics for all 363 genomes; **S4** locus statistics for the strict-core alignments; **S5** locus presence spectrum; **S6** the 28 sister-group label conflicts; **S7** the 20 core-identical genome groups; **S8** terminal-branch-length and branch-support distributions; **S9** *SIX* effector prevalence and the reference proteins used; **S10** the per-genome *SIX* effector matrix; **S11** the curated 324-genome non-redundant panel.

Supplementary Files

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

- ESM1.pdf
- ESM1.pdf
- ESM1.pdf
- ESM2.xlsx
- ESM2.xlsx
- ESM2.xlsx