Online Resource 1

Supplementary Figures S1-S7

ARTICLE TITLE

Lineage predicts SIX effector repertoire but not *Avr2* race genotype in *Fusarium oxysporum* isolates from Türkiye

JOURNAL

World Journal of Microbiology and Biotechnology

AUTHORS

Beyza Akal, Duygu Ateş, Muhammed Bahattin Tanyolaç, Yasin Kaymaz

AFFILIATION

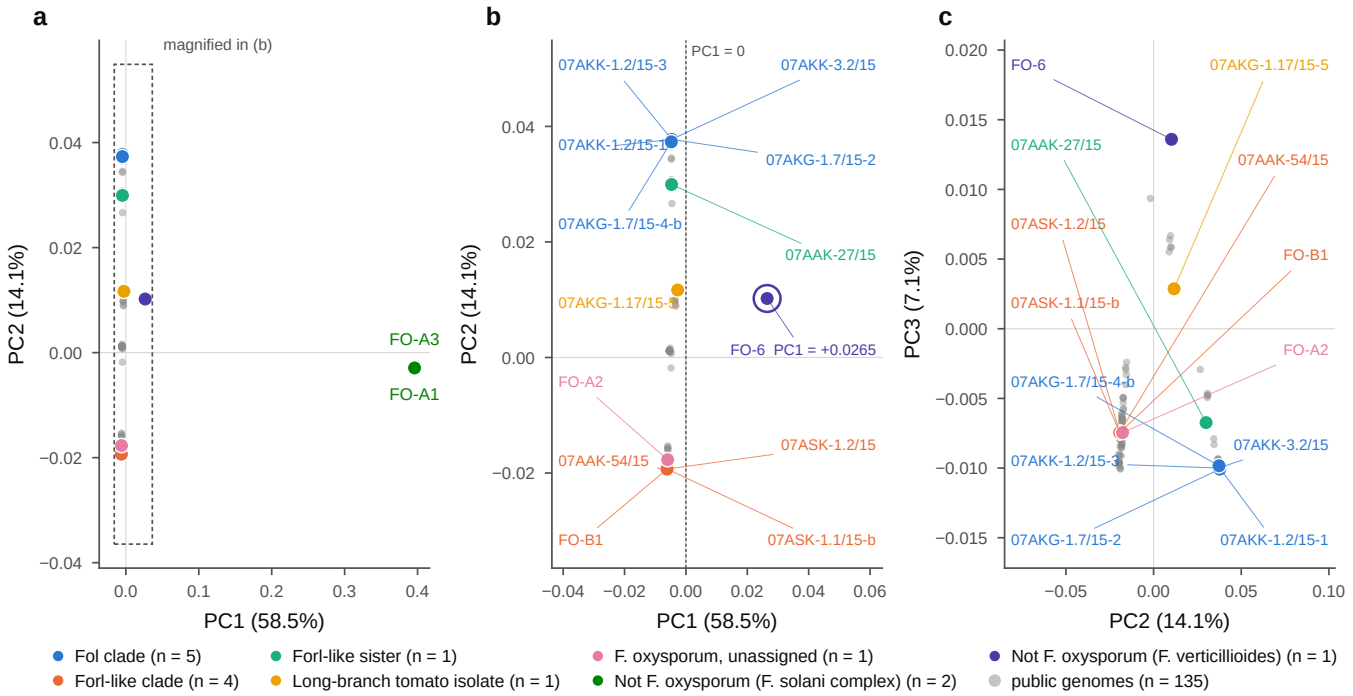
Bioengineering Department, Faculty of Engineering, Ege University, Izmir 35040, Türkiye (all authors)

CORRESPONDING AUTHOR

Yasin Kaymaz · Bioengineering Department, Faculty of Engineering, Ege University, Izmir 35040, Türkiye
yasin.kaymaz@ege.edu.tr

CONTENTS OF THIS FILE

Online Resource 2 (ESM_2.xlsx) contains Supplementary Tables S1-S24.



n = 150 genomes (135 distinct public + 15 from Türkiye), the same panel as Figure 2. Colour is the lineage recovered in Figure 2, not the laboratory pathogenicity label.
FO-6 is marked in (b): PC1 = +0.0265, opposite in sign to all 135 public genomes (-0.0070 to -0.0033) and to the
12 confirmed F. oxysporum isolates from Türkiye (-0.0062 to -0.0027); its reassignment to F. verticillioides is given in Section 3.3.

Fig. S1 Principal coordinates analysis of the Fol-cluster supermatrix PCoA of amino-acid identity distances computed from the same 4,464-locus supermatrix as Fig. 2 (n = 150 genomes: the 135 distinct public genomes plus the 15 isolates from Türkiye). PC1 explains 58.5%, PC2 14.1% and PC3 7.1% of the variance. Points are coloured by the lineage recovered in Fig. 2, never by the laboratory pathogenicity label, which is confounded with host of isolation (Section 4.5). (**a**) PC1 versus PC2 over the full range: PC1 is almost entirely the FO-A3/FO-A1 axis, so the remaining 148 genomes collapse onto one line; the separation is a result and is therefore drawn to scale. The dashed rectangle marks the region magnified in (b). (**b**) PC1 versus PC2 over the dense region, with the 13 remaining isolates from Türkiye labelled. FO-6 is resolved here at PC1 = +0.0265, on the opposite side of zero from all 135 public genomes (−0.0070 to −0.0033) and from the 12 confirmed *F. oxysporum* isolates from Türkiye (−0.0062 to −0.0027); it is marked and its species reassignment is given in Section 3.3. (**c**) PC2 versus PC3 for the same 13 isolates, where the structure among the *F. oxysporum* isolates resolves. The coordinate file holds 150 of the 152 tree tips; the two absentees are exactly the superseded duplicate accessions GCA_000733055.2 and GCA_001703025.1, so the PCoA was already de-duplicated and no re-run was required

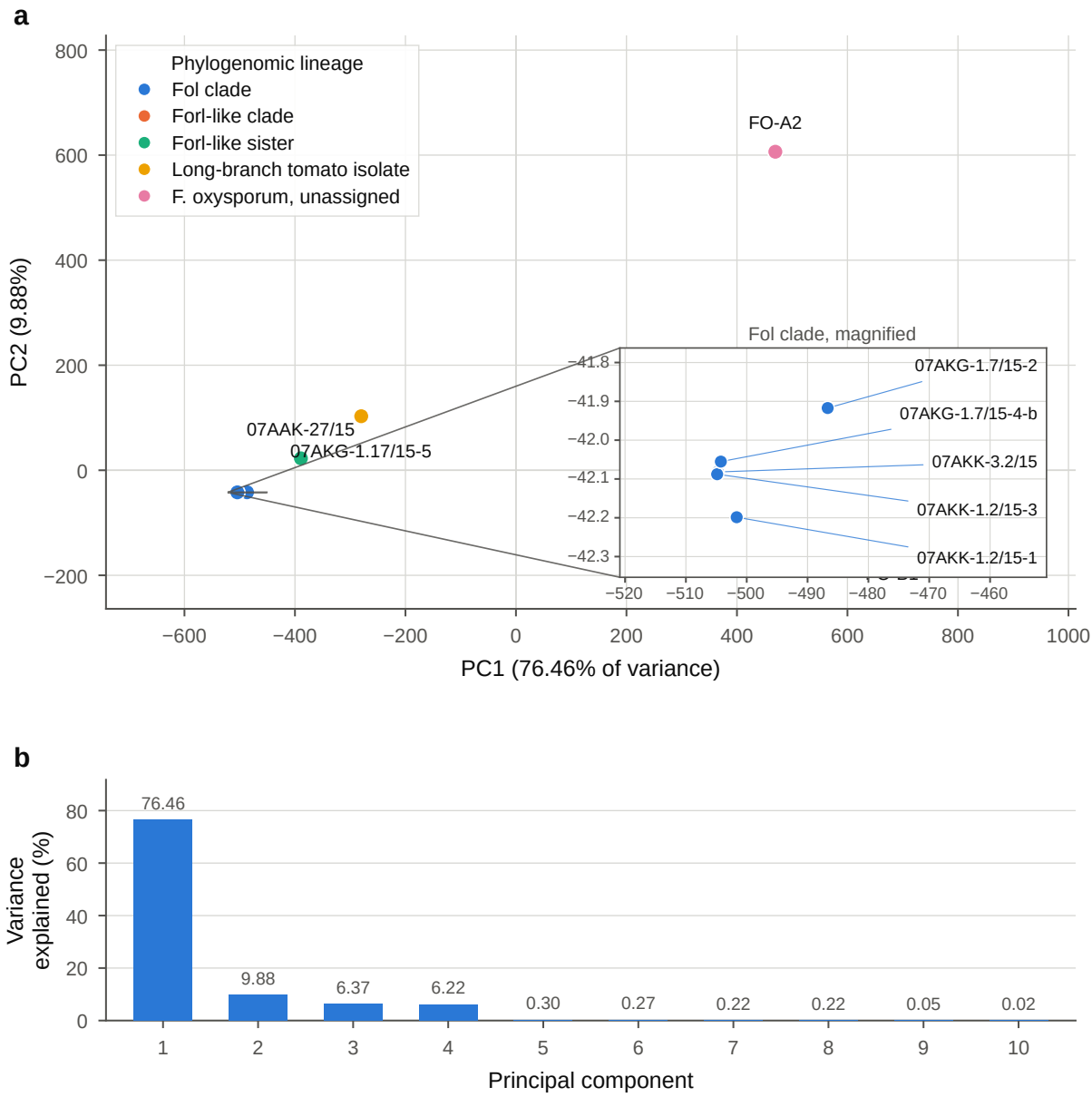
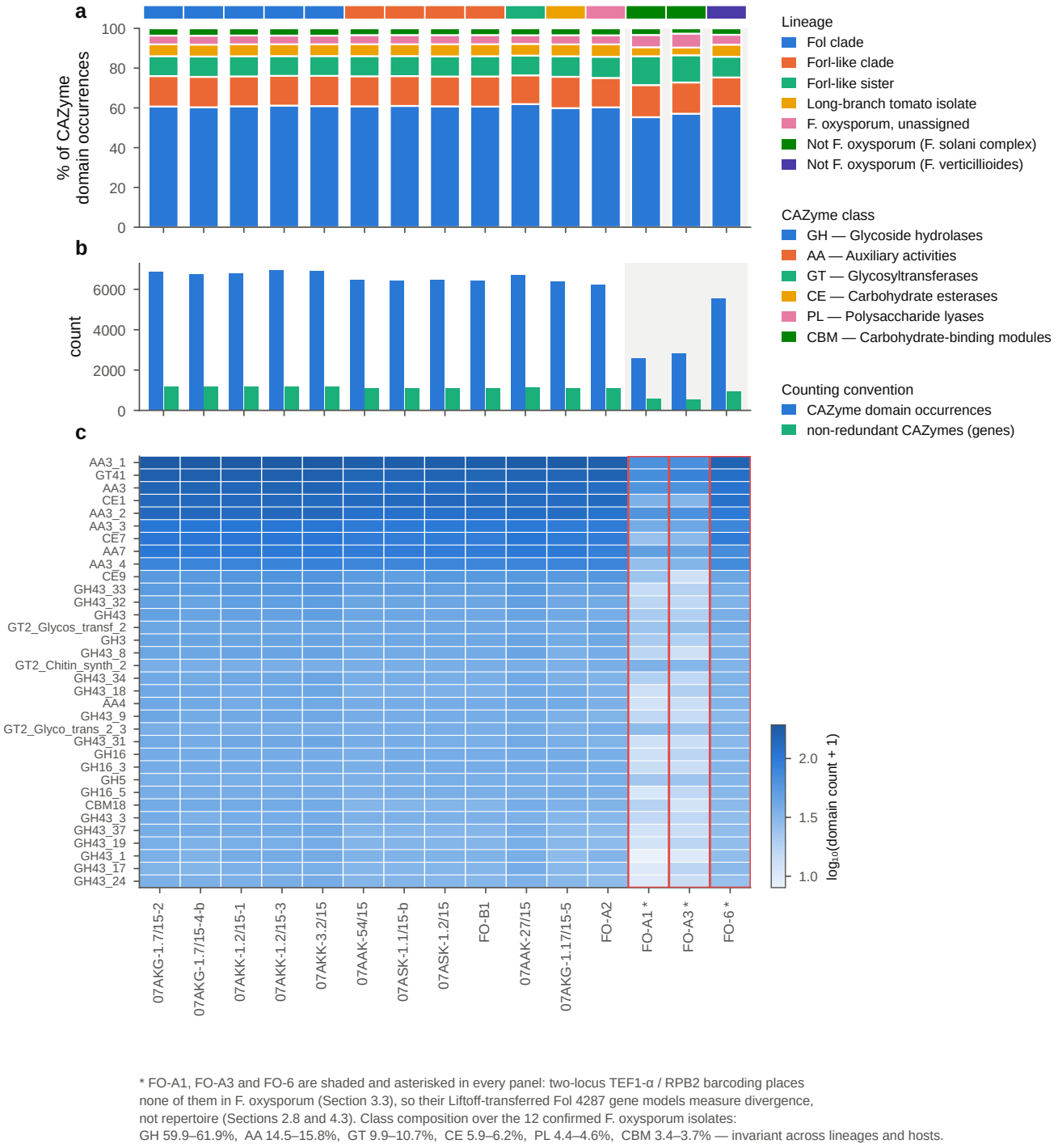


Fig. S2 Genotype principal component analysis of the 12 deeply sequenced isolates PCA of the filtered joint-call genotype matrix. Of 13,913,835 records in the callable-region VCF, 731,390 sites were polymorphic among the 12 isolates (700,918 single-nucleotide variants and 30,472 indels) and these are the sites the PCA uses; monomorphic records contribute an all-zero column after per-variant centring and are dropped without any effect on the eigenvalues or eigenvectors. This is not the same quantity as the 702,353 non-reference single-nucleotide variants reported in Section 3.8, which counts every filtered record including those at which all 12 isolates share the same non-reference genotype. Missing genotypes are row-mean imputed and variants are centred before singular-value decomposition, exactly as in `variant_calling/pca.py`. (**a**) PC1 (76.46% of variance) versus PC2 (9.88%), coloured by phylogenomic lineage, with a magnified inset over the five f. sp. *lycopersici*-clade isolates, which are mutually near-identical at the core-genome level and would otherwise overplot as a single mark. (**b**) Scree plot of the variance explained by each principal component. The three isolates excluded for low mapped depth against Fol4287 are FO-A3 (6.3×), FO-A1 (6.2×) and FO-6 (10.4×): the same three isolates that the two-locus barcoding of Section 2.4 subsequently placed outside *F. oxysporum*, and, at true sequencing depths of 72.5×, 71.7× and 114.8×, not low-coverage libraries (Section 3.1; Online Resource 2, Table S19). The depth threshold of ≥ 49× and the species assignment select identical sets of twelve isolates, so this figure required no re-run after the reassignment and the 12 isolates shown are exactly the 12 confirmed *F. oxysporum* isolates analysed throughout Sections 3.4–3.8



* FO-A1, FO-A3 and FO-6 are shaded and asterisked in every panel: two-locus TEF1-α / RPB2 barcoding places none of them in F. oxysporum (Section 3.3), so their Liftoff-transferred Fol 4287 gene models measure divergence, not repertoire (Sections 2.8 and 4.3). Class composition over the 12 confirmed F. oxysporum isolates: GH 59.9–61.9%, AA 14.5–15.8%, GT 9.9–10.7%, CE 5.9–6.2%, PL 4.4–4.6%, CBM 3.4–3.7% — invariant across lineages and hosts.

Fig. S3 Carbohydrate-active enzyme (CAZyme) repertoire of the 15 isolates from Türkiye dbCAN v13 `hmmscan` annotation (`--domE 1e-3`) of the predicted proteomes; isolates are ordered by lineage and then by strain code. (**a**) Class composition as a percentage of CAZyme domain occurrences, in the six CAZy classes. (**b**) Totals under both counting conventions, plotted side by side because they differ by roughly six-fold and are easily conflated: domain occurrences (2,587–6,942 across the 15 isolates) and non-redundant CAZymes, that is, proteins carrying at least one CAZy domain (524–1,186). Class sums in (a) and (b) reconcile with Online Resource 2, Table S32 for every isolate. (**c**) Heat map of the 34 most abundant of the 636 dbCAN families detected, on a $\log_{10}(\text{count} + 1)$ scale. Panel (c) is built from the per-sample dbCAN family tables. FO-A3, FO-A1 and FO-6 are shaded and asterisked in every panel: none is *F. oxysporum* (Section 3.3), so their counts are Liftoff annotation-transfer artefacts and not biology (Sections 2.8 and 4.3). The class-composition percentages quoted in Section 3.6 are computed over the 12 confirmed *F. oxysporum* isolates only

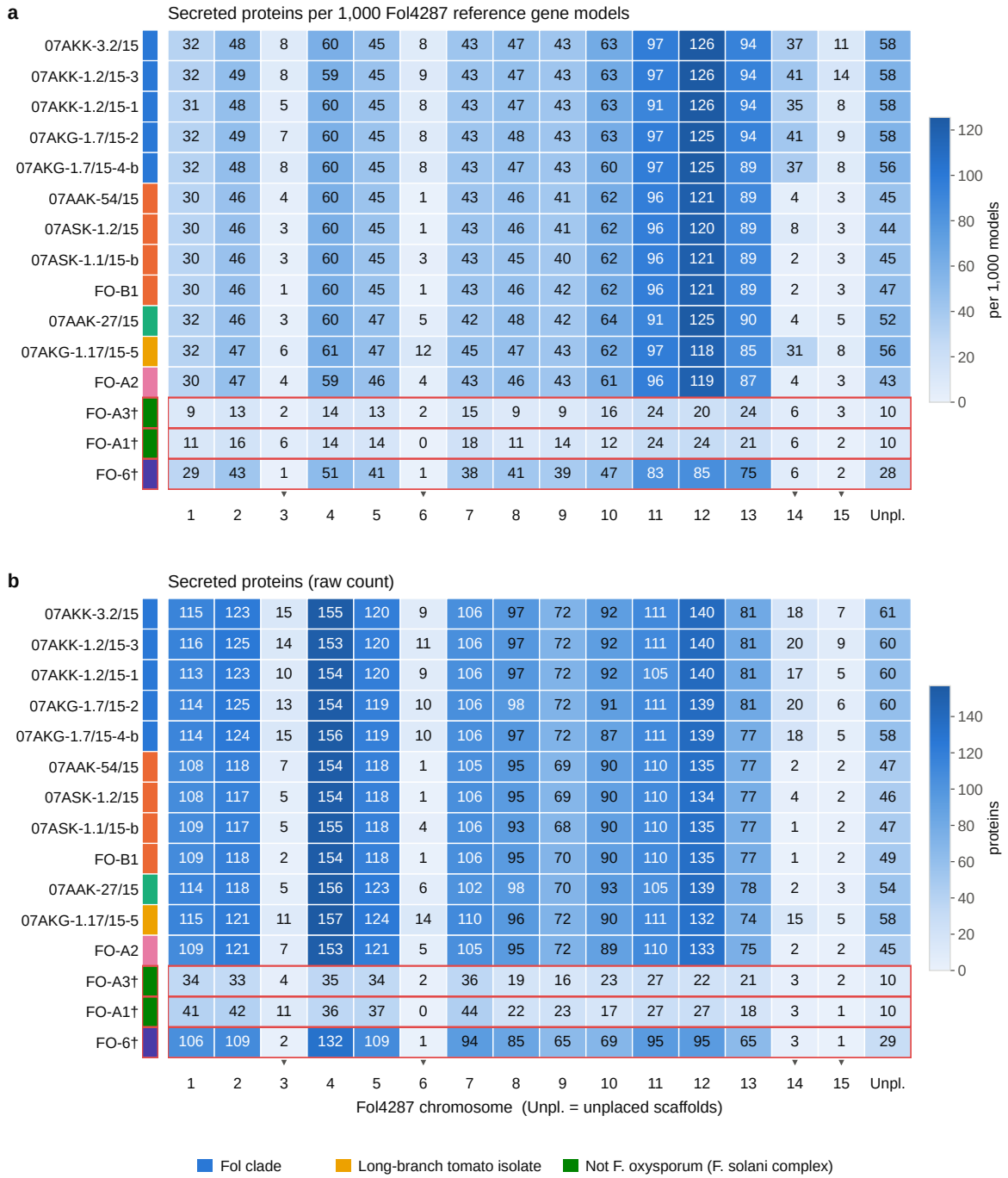


Fig. S4 Chromosomal distribution of secreted proteins Each secreted protein is assigned to the Fol4287 sequence carrying its orthologue; this is a reference coordinate, identical in provenance for all 15 isolates, and not a de novo assembly position. (**a**) Secreted proteins per 1,000 Fol4287 reference gene models on the same sequence. Normalisation is necessary because raw counts are confounded on both axes; a column is tall largely because that chromosome carries more reference gene models. (**b**) The same matrix as raw counts, for transparency. Accessory chromosomes 3, 6, 14 and 15 are marked. The "Unpl." column reports secreted proteins whose Fol4287 orthologue lies on an unplaced scaffold (`NW_*`): 694 of the 17,083 secreted-protein records across the 15 isolates, 76 distinct proteins, 10–61 records per isolate. With that column included, all 15 row sums reconcile exactly with the per-isolate secretome sizes in Online Resource 2, Table S30. FO-A3, FO-A1 and FO-6 are outlined in red and daggered; none is *F. oxysporum* (Section 3.3), so their secretome sizes are Liftoff annotation-transfer artefacts (Sections 2.8 and 4.3). All 15 isolates are retained in this figure so that the artefact is visible; the 17,083 records it totals are the 15-isolate total, of which 15,343 belong to the 12 confirmed isolates

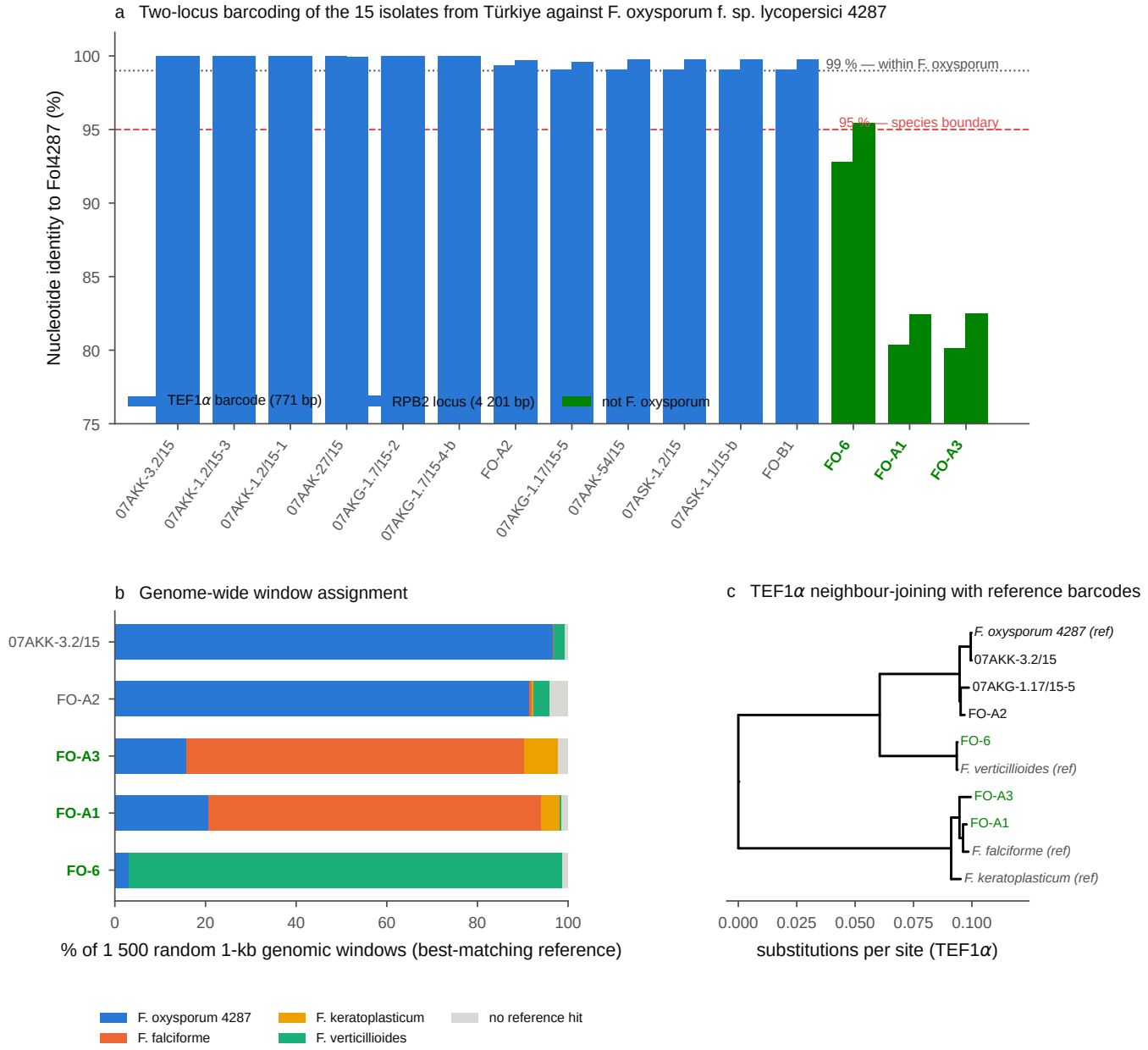


Fig. S5 Two-locus species assignment of the 15 isolates from Türkiye (**a**) Nucleotide identity of each isolate to *F. oxysporum* f. sp. *lycopersici* 4287 at the 771 bp *TEF1-α* barcode region and at the 4,201 bp *RPB2* locus, computed over alignment columns at which neither sequence carries a gap or an N. The three isolates reassigned outside *F. oxysporum* are coloured and bold-labelled. Reference lines mark 99% and 95% identity. (**b**) Genome-wide assignment of randomly sampled 1-kb windows per isolate (a sampling target of 1,500 windows, 1,465–1,497 of which were recovered per isolate) to the best-matching of the four references, shown for a representative confirmed isolate (07AKK-3.2/15), for FO-A2, and for the three reassigned isolates; windows with no reference hit are shown in grey. (**c**) Neighbour-joining tree of the *TEF1-α* barcode region for a subset of isolates together with the four reference barcodes, showing that FO-A3 and FO-A1 group with the *Fusarium solani* complex references and FO-6 with *F. verticillioides* 7600

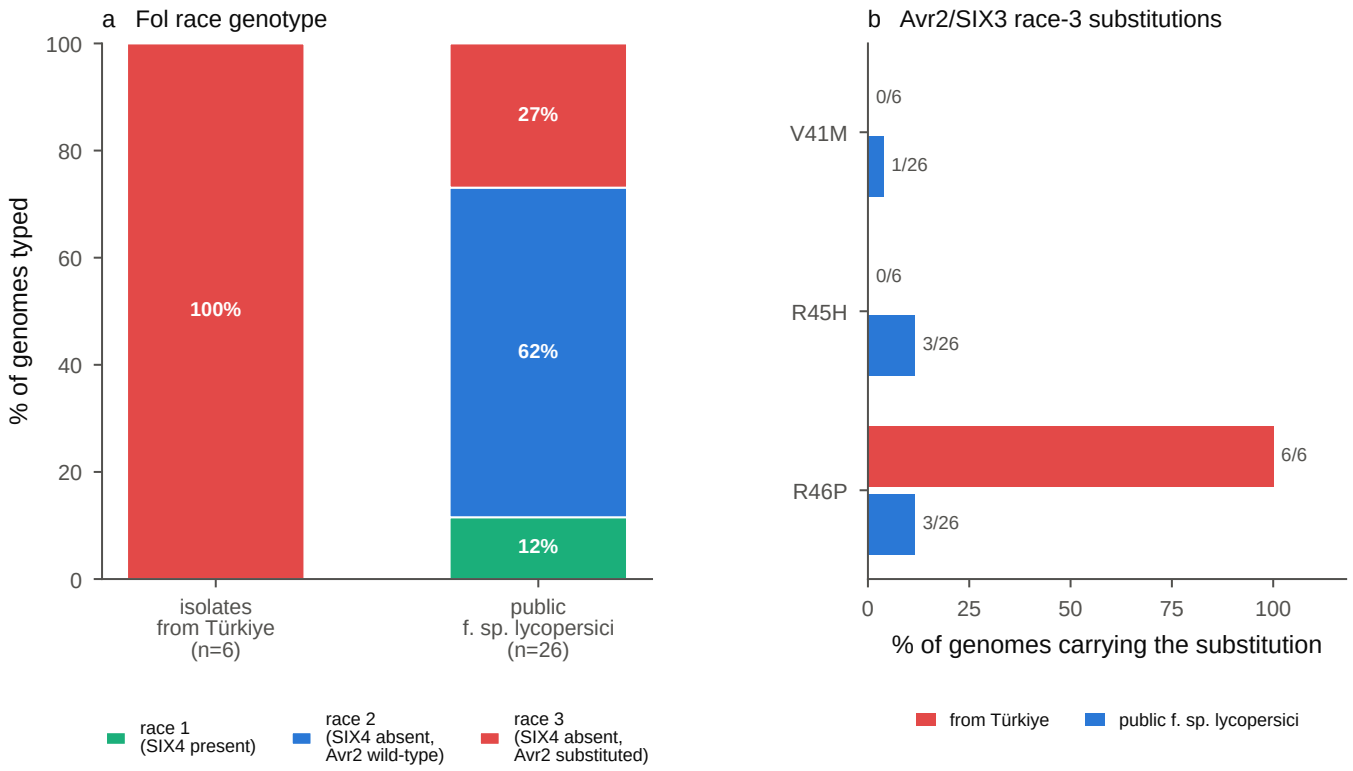


Fig. S6 *Avr2*/*SIX3* race genotype of the isolates from Türkiye against the public f sp. *lycopersici* panel. (**a**) Race genotype composition, as a percentage of genomes typed, for the six *SIX3*-positive isolates from Türkiye and for the 26 public genomes registered as f. sp. *lycopersici* from which an *Avr2* sequence could be recovered. Race 1 is defined by *SIX4* presence; race 2 by *SIX4* absence with a wild-type *Avr2*; race 3 by *SIX4* absence with any of the substitutions V41M, R45H or R46P. (**b**) Frequency of each of the three race-3 substitutions in the two groups, with counts over denominators printed beside each bar. All six isolates from Türkiye carry R46P; the public panel carries R45H in 3, R46P in 3 and V41M in 1 of 26. Because five of the six isolates from Türkiye are near-clones, this panel is a description of allele carriage and not a test of enrichment; the conservative Fisher test that collapses the near-clones to a single observation is not significant (one-sided p = 0.095; Online Resource 2, Table S23)

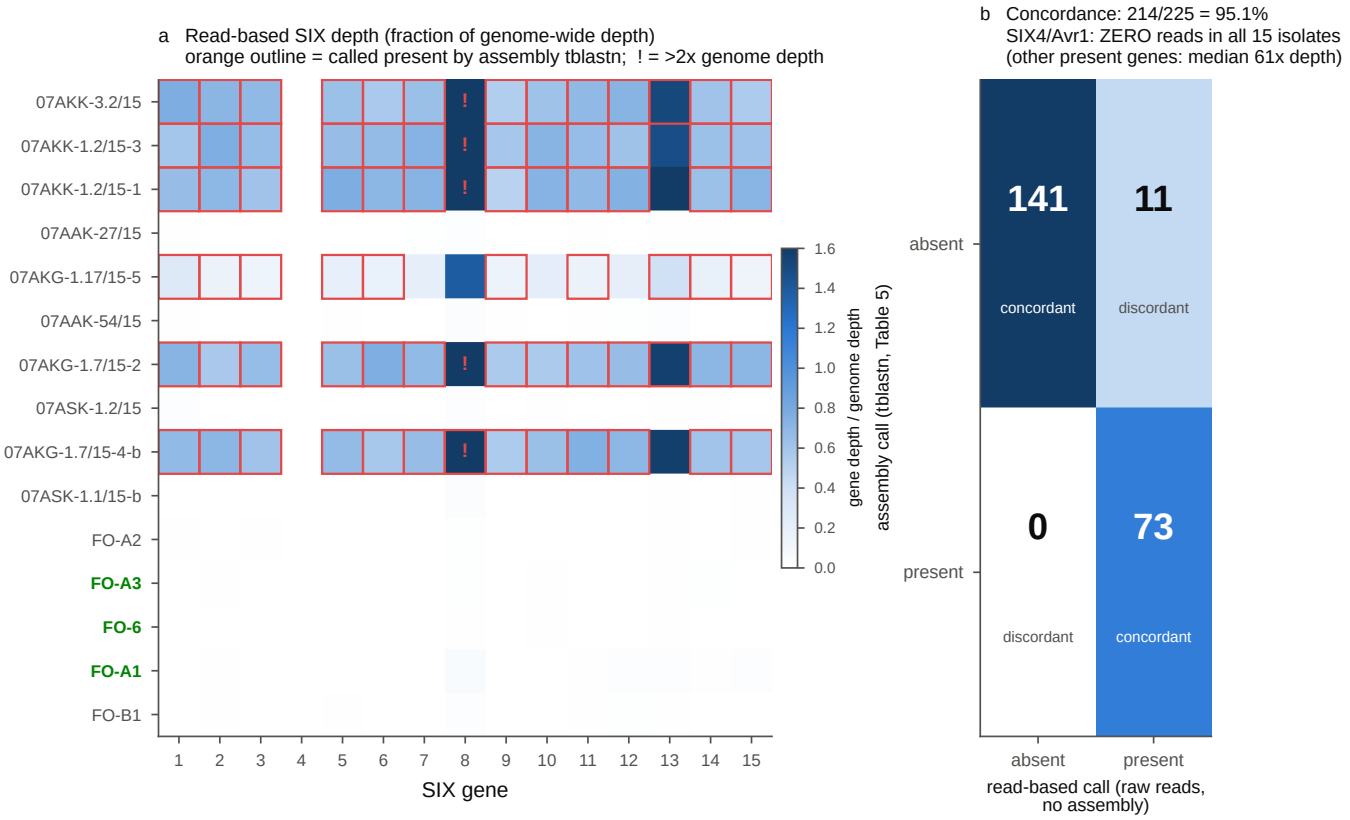


Fig. S7 *Secreted In Xylem* (SIX) presence/absence called from raw reads, with no assembly in the procedure (**a**) Heat map of read-based depth at each of the 15 SIX loci in each of the 15 isolates, expressed as a fraction of that isolate's genome-wide sequencing depth so that reference-recruitment artefacts are visible rather than hidden. Rows for the three isolates reassigned outside *F. oxysporum* (Section 3.3) are coloured and bold. An orange outline marks the cells called present by the assembly-based tblastn screen of Table 2, so that agreement and disagreement between the two methods can be read directly off the panel. An exclamation mark marks cells at more than twice the genome-wide depth: these are reference-recruitment artefacts of a 15-locus reference, not high copy number, and are not reported as presence (Section 2.9; Online Resource 2, Table S21). (**b**) Confusion matrix of the two call sets over all 225 isolate × gene comparisons: 141 absent by both methods, 73 present by both, 11 present from reads only and none present from the assembly only, giving 214/225 = 95.1% concordance. The ten calls that formerly ran assembly-present/reads-absent are the ten withdrawn by the contig criterion of Section 2.9, so the residual disagreement runs in one direction only. *SIX4/Avr1* recruited zero reads in all 15 isolates, against a median depth of 61× for the SIX genes scored present. The reference used here was built exclusively from public genomes (Online Resource 2, Table S22), so this test is independent of the assemblies from Türkiye