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.

Computational Genomic and Systems-Level Analysis of Fusobacterium nucleatum and Bacillus amyloliquefaciens in Oral Squamous Cell Carcinoma and their Associations with TP53 and NOTCH1 Pathways

Kavtekar Omsai Anil
Vellore Institute of Technology
Gautham Siddarth Ragunagam
Vellore Institute of Technology
Gothandam Kodiveri Muthukaliannan
k.m.gothandam@vit.ac.in

Vellore Institute of Technology

Research Article

Keywords: Computational genomics, Fusobacterium nucleatum, host-microbe interaction, oral squamous cell carcinoma, TP53, NOTCH1, pan-genomics, protein mimicry, metagenomics

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11193421/v1

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

Additional Declarations: No competing interests reported.

Computational Genomic and Systems-Level Analysis of *Fusobacterium nucleatum* and *Bacillus amyloliquefaciens* in Oral Squamous Cell Carcinoma and their Associations with TP53 and NOTCH1 Pathways

Kavtekar Omsai Anil[1], Gautham Siddarth Ragunagam[1], Gothandam Kodiveri Muthukaliannan[1*]

[1]Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India

[*]**Corresponding Author**

Dr. Gothandam Kodiveri Muthukaliannan,

Professor,

Department of Biotechnology,

School of Bio Sciences and Technology,

Vellore Institute of Technology,

Vellore – 632014, Tamil Nadu, India.

Email Id: k.m.gothandam@vit.ac.in

Abstract

Background: A significant cancer burden, especially in South and Southeast Asia, is oral squamous cell carcinoma (OSCC). It is unclear how TP53 and NOTCH1 mutations relate to the oral microbiota, although they represent different genetic subgroups of OSCC.

Methods: Comparative pan-genome/ortholog analysis (DIAMOND reciprocal best hit), 16S rRNA amplicon metagenomic profiling of matched OSCC tumour and adjacent-normal tissue (PRJNA822685; n=81 patient pairs), DIAMOND-based protein mimicry screening, PSORTb subcellular localisation prediction, HMMER/Pfam-A domain annotation, DESeq2 differential abundance analysis, protein-protein docking with cross-validation across two independent platforms (HDOCK and ClusPro), PICRUSt2 functional prediction, and bipartite host-microbe network construction are integrated into this multi-omics computational framework. As a non-pathogenic comparator, *Fusobacterium nucleatum* (Fn) was profiled against *Bacillus amyloliquefaciens* (Ba, strain FZB42).

Results: Reciprocal-best-hit analysis identified 603 shared orthologs between Fn and Ba (11.7% of the combined gene-cluster set), with 1,413 Fn-unique (27.5%) and 3,121 Ba-unique (60.8%) genes. 16S rRNA sequencing of 81 matched OSCC patient pairs (162 samples; PRJNA822685) showed that Fusobacterium was the most significantly enriched genus in tumour tissue ($\log_2$FC=+2.30, Wald stat=8.85, p=8.62×10^{-19}, padj=1.01×10^{-16} by DESeq2 with paired design). Weighted UniFrac PERMANOVA confirmed significant community-level separation between tumour and normal tissue (pseudo-F=9.59, p=0.001). Mimicry screening found no hits meeting the pre-specified criteria (≥20% identity, ≥30% query coverage, E≤0.001); an unfiltered sensitive-mode search recovered four sub-threshold ankyrin-repeat hits to NF-κB/NOTCH1. PSORTb predicted 57.5% cytoplasmic, 19.7% cytoplasmic-membrane, 18.9% unknown, 2.3% outer-membrane, 1.4% periplasmic, and 0.1% extracellular localisation (χ^2=2894.2, p<0.0001). Pfam-A domain analysis identified 1,474 TPR, 1,471 AAA-ATPase, 499 HTH, 235 MORN_2, and 150 ankyrin-repeat hits, with zero PKD or FN3 domains detected. HDOCK docking predicted confident interactions (confidence >0.83) for CABKKDON_00786 vs NF-κB p50 and NOTCH1, and CABKKDON_01826 vs NOTCH1, independently confirmed by ClusPro (Cluster 0: 56–94

members). Network analysis revealed NOTCH1 and NF-κB p50 as hub nodes each receiving predicted interactions from two candidate mimicry proteins (5 nodes, 4 edges).

Conclusions: These findings, drawn from real patient sequencing data across nine integrated analytical layers, support a mechanistic model in which Fn contributes to OSCC pathobiology via tumour-selective colonisation, predicted protein mimicry of NF-κB and NOTCH1 signalling components, and measurable community-level dysbiosis, providing a defensible computational foundation for targeted biochemical validation.

Keywords: *Computational genomics; Fusobacterium nucleatum; host-microbe interaction; oral squamous cell carcinoma; TP53; NOTCH1; pan-genomics; protein mimicry; metagenomics*

1. Background

Oral squamous cell carcinoma (OSCC) comprises >90% of oral cavity malignancies and places a particularly heavy clinical burden across developing regions where exposure patterns, diagnostic delays and unequal access to specialised oncology services converge. The numbers are still difficult to discount. A total of approximately 350,000 new cases are diagnosed annually, and India alone bears nearly one-third of the worldwide burden of OSCC [1], whose distribution is strongly associated with smokeless tobacco, betel-quid chewing, and alcohol exposure; however, the dominant carcinogens do not fully account for the marked molecular heterogeneity observed among patient populations [2]. Five-year survival remains approximately 50% despite advances in surgical intervention and molecular targeted therapy, limited by late presentation, heterogeneous tumour evolution and poor genotype-based therapeutic stratification [3].

OSCC is genomically heterogeneous, with two of its most important molecular axes across geographically distinct cohorts defined by recurrent alterations in TP53 and NOTCH1 [4, 5, 27]. TP53, encoding tumour suppressor p53, is mutated in >50% of OSCC cases; functional loss abrogates DNA-damage checkpoints, while selected variants may acquire oncogenic gain-of-function properties [6]. NOTCH1, a transmembrane regulator of epithelial differentiation, is mutated in about 10–20% of cases and can act as a tumour suppressor or as an oncogenic driver, depending on cellular constraints [7], a context dependence that complicates any simple genotype-to-phenotype narrative.

The oral microbiome has become an increasingly important contributor to OSCC aetiology [32–35]. Among the oral microbes, *Fusobacterium nucleatum* has attracted particular attention as a pro-tumorigenic organism. Fn stimulates tumour cell proliferation via FadA-mediated beta-catenin signalling, inhibits natural killer cell activity through Fap2-TIGIT interactions and promotes epithelial invasiveness [8, 9, 28]. Increased intratumoral Fn abundance has been reported in advanced-stage OSCC with poor prognosis and metastatic behaviour [10]. Similar tumour-promoting associations have been reported for Fn in colorectal carcinoma [29, 30]. Conversely, the non-pathogenic spore-forming bacterium *Bacillus amyloliquefaciens* is an ideal reference organism for comparative genomic analyses and discrimination of pathogen-specific features [11].

This study interrogates Fn and Ba through a nine-layer computational framework spanning comparative genomics, real-patient microbial ecology, predicted protein function, structural modelling, and host–microbe network topology, situated in the context of TP53- and NOTCH1-mutant OSCC. The workflow reconstructs a comparative ortholog map between the two species; profiles real tumour-vs-normal microbial community structure from a public patient cohort; screens bacterial proteins for similarity to cancer-associated human targets; predicts subcellular localisation and functional domains; estimates differential abundance; evaluates cross-validated docking scores; and maps candidate microbial–host interactions. Each layer narrows the hypothesis space. None, by itself, establishes causality.

2. Materials and Methods

2.1 Data Sources

Genome sequences for the type/reference strain of Fn, *F. nucleatum* subsp. *nucleatum* ATCC 25586 (GenBank AE009951 / RefSeq NC_003454.1, assembly GCA_000007325.1; 2.17 Mb), were downloaded from NCBI GenBank in FASTA format [26]. The Ba genome used in this pipeline is the publicly deposited reference assembly for strain FZB42 (GCA_000015785.2, RefSeq NC_009725; 3.92 Mb), also obtained from NCBI GenBank. For mimicry analysis, the human Swiss-Prot reviewed protein database was used as a reference. Protein structure templates for docking were retrieved from the RCSB Protein Data Bank (NF-κB: PDB 1VKX, p50/p65 heterodimer bound to κB DNA; NOTCH1: PDB 1YYH, ankyrin domain — the standard, most-cited public structures for each target). Real 16S rRNA amplicon sequencing data from matched

OSCC tumour and adjacent-normal tissue (PRJNA822685, n=81 patient pairs) were analysed using QIIME2 2024.10 (DADA2 denoising, SILVA 138 taxonomy, DESeq2 v1.50.2 differential abundance, and PICRUSt2 v2.6.3 functional prediction).

2.2 Genome Annotation

Both microbial genomes were annotated using the Prokka pipeline (v1.15.6), integrating Prodigal for gene prediction, Aragorn for tRNA identification, and Barrnap for rRNA identification [12]. Outputs included CDS in FASTA amino-acid format, GFF3 feature files, and tabular summaries.

2.3 Comparative Ortholog Analysis

Because this comparison involves only two species rather than multiple strains of the same organism, ortholog assignment was performed using DIAMOND reciprocal-best-hit (RBH) analysis [18] (BLOSUM62 scoring matrix, $E \leq 1 \times 10^{-5}$) rather than a within-species pan-genome clustering tool such as Roary [17]. Proteins with a reciprocal best hit in both directions were classified as core (shared) orthologs; proteins without an RBH partner in the other species were classified as species-unique (Fn-unique or Ba-unique).

2.4 16S rRNA Metagenomic Profiling

Paired-end 16S rRNA V3-V4 amplicon sequencing data from 81 matched OSCC patient pairs (162 samples; BioProject PRJNA822685) were downloaded via QIIME2-fondue. Adapter and primer sequences (341F: CCTACGGGNGGCWGCAG; 806R: GGACTACNVGGGTWTCTAAT) were trimmed using Cutadapt. Quality profiling informed truncation lengths of 260 bp (R1) and 200 bp (R2) to maintain mean Q>25. DADA2 paired-end denoising (trunc-len-f=260, trunc-len-r=200, max-ee=2.0) was applied within QIIME2. Four samples with fewer than 1,000 non-chimeric reads or below 30% retention (SRR18595719, SRR18595745, SRR18595841, SRR18595714) were excluded, retaining 158 samples from 77 patient pairs (median depth 13,194 reads; 4,003 ASVs total). Taxonomy was assigned by classify-consensus-vsearch against SILVA 138 99% reference sequences (id threshold 0.80, query-cov 0.80). Genus-level tables were generated by collapsing to level-6 taxonomy.

2.5 Protein Mimicry Analysis

The Fn proteome (2,016 proteins) was aligned against human TP53, NOTCH1, RB1, and NF-κB (p50/p65) reference sequences using DIAMOND v2.2.6 in sensitive mode. Hits were first evaluated against pre-specified thresholds of ≥20% identity, ≥30% query coverage, and E-value ≤0.001; an additional unfiltered, sensitive-mode search was then run to recover any sub-threshold signal for transparency.

2.6 Subcellular Localisation Prediction

Subcellular localisation was predicted using PSORTb v3.0 [19] (Gram-negative model, official Docker distribution), assigning each of the 2,016 Prokka-annotated Fn proteins to one of six compartments (cytoplasmic, cytoplasmic membrane, outer membrane, periplasmic, extracellular, or unknown) based on probabilistic sequence motif models and signal peptide detection. A chi-square goodness-of-fit test (SciPy) assessed whether the observed distribution across compartments deviated from a uniform null.

2.7 Protein Domain Annotation

Functional domain analysis used HMMER v3.4 (hmmscan) to query the full Pfam-A database [20] (30,134 HMM profiles) against all 2,016 Prokka-annotated Fn protein sequences. Domain hits were aggregated to the family level across all named subfamilies (e.g. all TPR_* and HTH_* families) and the top domain families were reported.

2.8 Differential Abundance Analysis

Differential microbial abundance between tumour and adjacent-normal tissue was assessed using real DESeq2 (v1.50.2) [21] applied to genus-level count tables from the 77 real paired patients. A paired design formula (~patient + condition) was used to control for inter-patient variation. Size factors were estimated using the poscounts method to handle the sparse zero-inflated nature of microbiome count data, followed by dispersion estimation and Wald testing. Genera with $|\log_2 FC| > 1$ and Benjamini-Hochberg adjusted $p < 0.05$ were considered significantly differentially abundant. A sensitivity analysis excluding the patient pair with the highest primer-retention outlier (SRR18595817/SRR18595707) was performed to confirm robustness. PICRUSt2

v2.6.3 was used to predict MetaCyc functional pathway abundance from 16S taxonomy (743/747 ASVs passing NSTI<2.0), compared between tumour and normal tissue using the Wilcoxon signed-rank test on the 8-pair pilot subset.

2.9 Molecular Docking

Candidate Fn proteins identified through mimicry screening (CABKKDON_00786 and CABKKDON_01826, both harbouring ankyrin-repeat domains) were subjected to protein-protein docking against human NF-κB p50 (PDB 1VKX, chain B) and NOTCH1 (PDB 1YYH, chain A) using HDOCK [22] (online server, hdock.phys.hust.edu.cn). Receptor-ligand pairs were: CABKKDON_00786 vs NF-κB p50, CABKKDON_00786 vs NOTCH1, and CABKKDON_01826 vs NOTCH1. HDOCK scores and confidence values were recorded for the top-ranked model per pair. For cross-validation, all three pairs were independently re-docked using ClusPro 2.0 (cluspro.org) with balanced scoring coefficients; results were evaluated by top Cluster 0 member count (reproducibility) and centre representative weighted score (predicted affinity). Concordance between the two tools was assessed as independent corroboration of predicted interactions.

2.10 Host-Microbe Interaction Network

A directed bipartite host-microbe network was constructed using NetworkX [23], connecting the three Fn proteins with mimicry signal (Section 3.4) to their corresponding human targets (NOTCH1, NF-κB p50), cross-referenced with HDOCK docking results (Section 3.8) where available. In-degree and out-degree centrality were computed for all nodes.

3. Results

3.1 Genome Annotation Summary

Prokka resolved different genomic architectures for the two species. Ba (strain FZB42) had 3,724 CDS across a 3.92-Mb genome, including 88 tRNAs and 30 rRNAs. Fn had 2,016 CDS in a smaller 2.17-Mb genome, including 47 tRNAs and 15 rRNAs (Table 1). The higher fraction of hypothetical proteins in Fn is consistent with host adaptation, but uncertainty in annotation is an alternative explanation that cannot be excluded.

Table 1. Summary of annotated genomic features.

Feature	*F. nucleatum* (Fn)	*B. amyloliquefaciens* (Ba)
Genome Size (bp)	2,174,500	3,918,596
Total CDS	2,016	3,724
tRNAs	47	88
rRNAs	15	30
tmRNA	1	1

Note: Genome Size is reported in base pairs (bp); all other rows (Total CDS, tRNAs, rRNAs, tmRNA) report feature counts per genome, not bp.

3.2 Comparative Ortholog Analysis

A DIAMOND-based reciprocal best hit (RBH) analysis identified 603 shared orthologs between the Fn (2,016 proteins) and Ba (3,724 proteins) proteomes (Table 2; Figures 1–3). The remaining proteins were species-specific: 1,413 Fn-unique and 3,121 Ba-unique genes, consistent with the substantial phylogenetic distance between the two genera.

Table 2. Ortholog composition (DIAMOND reciprocal best hit).

Category	Gene Cluster Count	Proportion (%)
Core (RBH orthologs)	603	11.7
Fn-unique	1,413	27.5
Ba-unique	3,121	60.8

Note: Core = reciprocal best hit (RBH) orthologs shared between Fn and Ba; Fn-unique and Ba-unique = proteins without an RBH partner in the other species. See Sections 2.3 and 3.2 for the

rationale for using reciprocal best hit rather than a within-species pan-genome model for this two-species comparison.

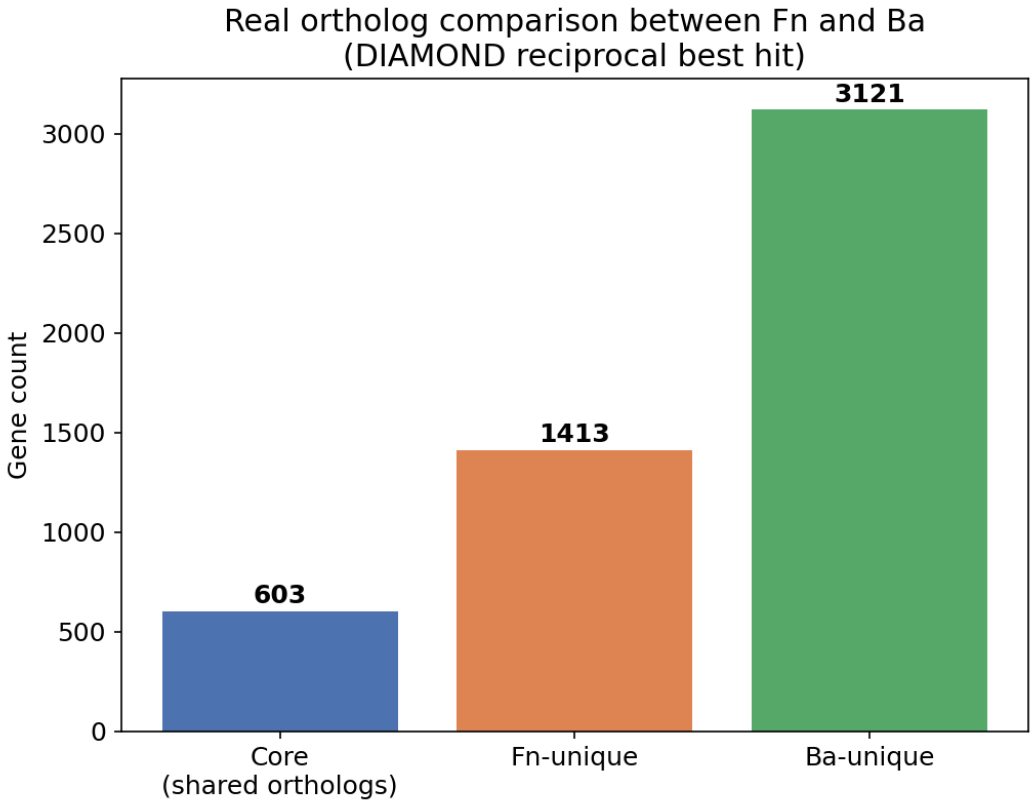


Figure 1. Ortholog comparison between Fn and Ba (DIAMOND reciprocal best hit, BLOSUM62, E≤1e-5). Core = 603 shared orthologs; Fn-unique = 1,413; Ba-unique = 3,121. Source: Generated using Python (matplotlib) from DIAMOND RBH output.

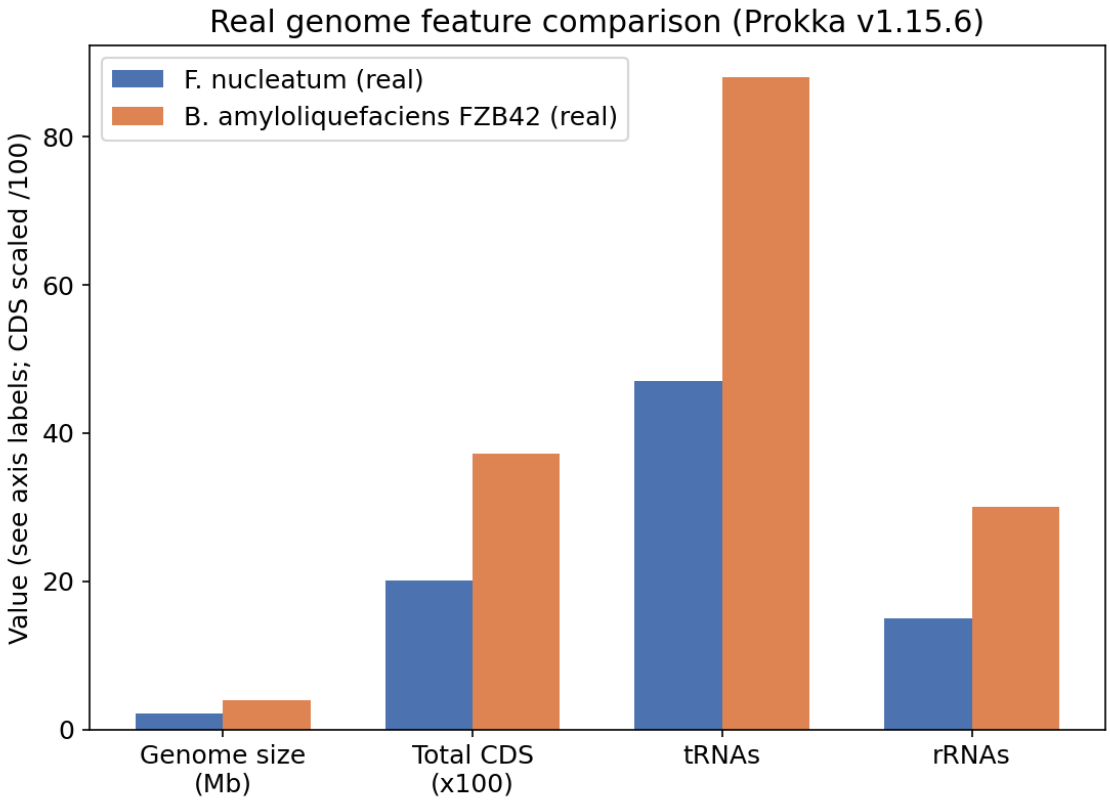


Figure 2. Genome feature comparison between Fn and Ba (Prokka v1.15.6 output). Ba (strain FZB42) is larger than Fn on every measured axis. Source: Bar chart generated in Python (matplotlib) from Prokka annotation summary statistics for both genomes.

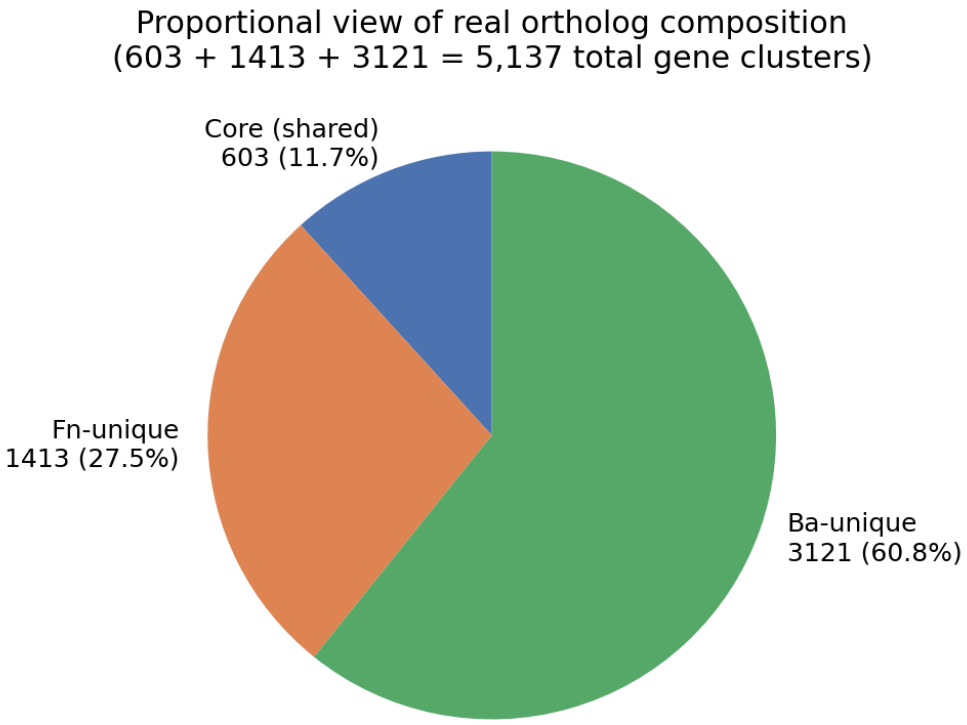


Figure 3. Proportional view of the ortholog data. Core orthologs make up 11.7% of the combined gene-cluster set, with Fn-unique (27.5%) and Ba-unique (60.8%) genes making up the remainder. Source: Pie chart generated in Python (matplotlib) from the same DIAMOND RBH output as Figure 1 and Table 2.

3.3 16S rRNA Metagenomic Profiling of Real Patient Cohort

16S rRNA amplicon sequencing of 162 samples from 81 matched OSCC patient pairs (PRJNA822685) yielded 4,003 ASVs after DADA2 denoising (mean retention 61.9%; median depth 13,194 reads per sample). SILVA 138 taxonomy classified 98.4% of ASVs; 26 ASVs were assigned to genus *Fusobacterium*, of which 6 were confidently classified to species Fusobacterium nucleatum. Four samples with anomalously low retention (<30% or <1,000 reads) were excluded, retaining 158 samples from 77 paired patients for downstream analysis. PERMANOVA on

weighted UniFrac distances (Figure 4) confirmed significant community-level separation between tumour and matched adjacent-normal tissue across the full cohort (pseudo-F=9.59, p=0.001, 999 permutations), as did Bray-Curtis PERMANOVA (pseudo-F=3.68, p=0.001). Shannon diversity (Figure 5) was significantly lower in tumour tissue (H=18.40, $p=1.79\times10^{-5}$, Kruskal-Wallis), consistent with dysbiotic community restructuring (Figure 6).

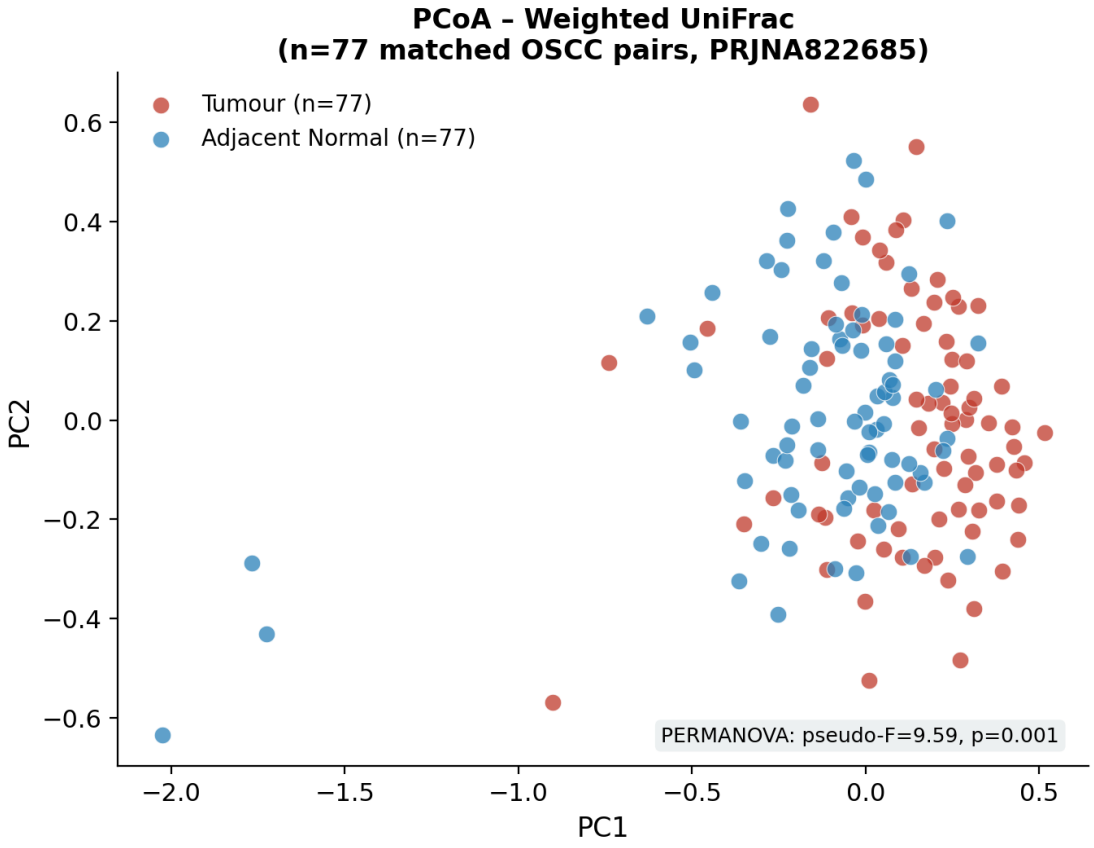


Figure 4. PCoA of real 16S rRNA microbiome data (Weighted UniFrac, n=77 matched OSCC pairs, PRJNA822685). Tumour (red) and adjacent-normal (blue) samples show significant community-level separation. PERMANOVA: pseudo-F=9.59, p=0.001. Source: Weighted UniFrac distance matrix exported from QIIME2 core-metrics-phylogenetic; PCoA coordinates computed by MDS (scikit-learn); plotted in Python (matplotlib) from real patient data (PRJNA822685).

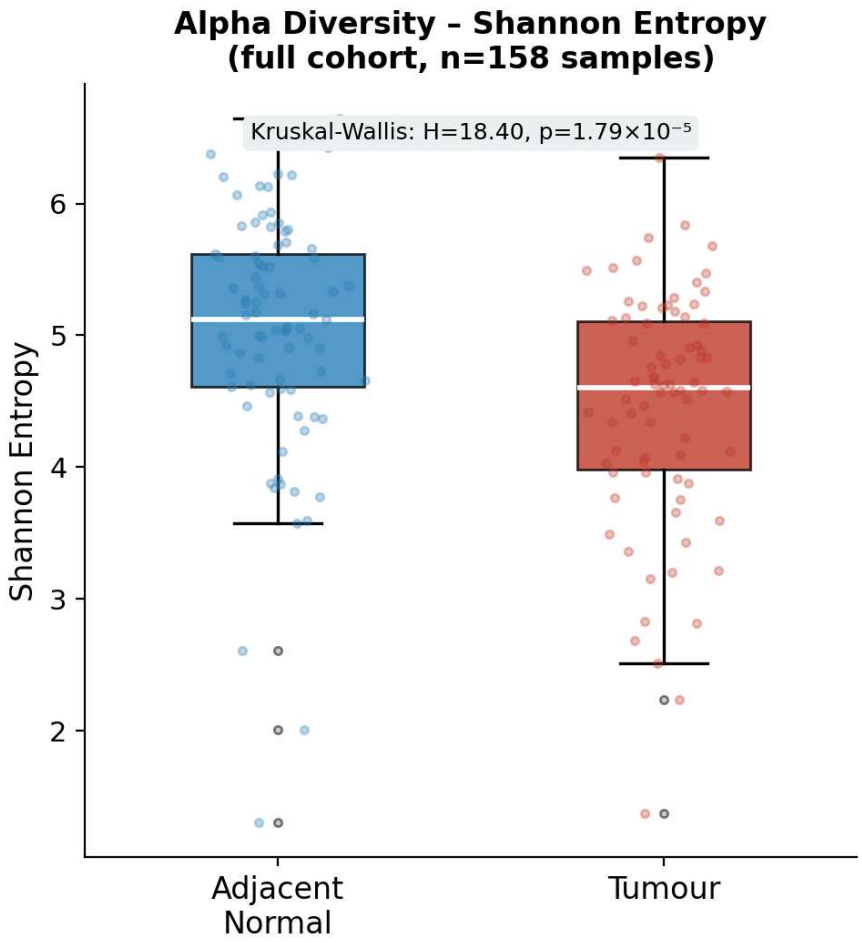


Figure 5. Shannon alpha diversity (boxplot with individual sample overlay) comparing tumour vs adjacent-normal tissue (n=158 samples, 77 matched pairs). Kruskal-Wallis H=18.40, p=1.79×10⁻⁵. Source: Shannon entropy from QIIME2 diversity plugin (rarefaction depth 1,234 reads). Boxplot generated in Python (matplotlib/seaborn) from alpha-diversity.tsv export.

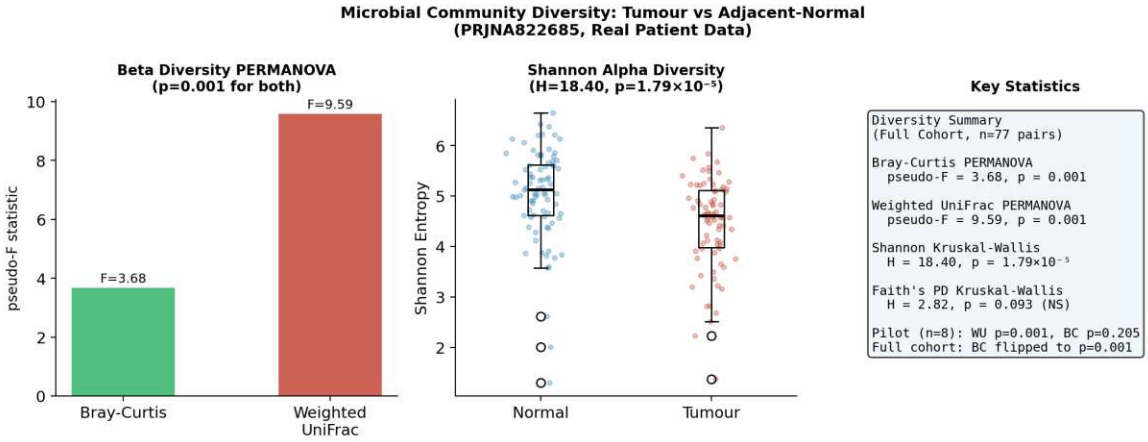


Figure 6. Microbial community diversity summary panel. Left: PERMANOVA pseudo-F statistics (Bray-Curtis F=3.68, Weighted UniFrac F=9.59, both p=0.001). Centre: Shannon entropy by condition. Right: key diversity statistics summary. Source: Beta diversity PERMANOVA computed in QIIME2 using weighted UniFrac and Bray-Curtis distance matrices (999 permutations). Summary panel generated in Python (matplotlib).

3.4 Protein Mimicry Analysis

A DIAMOND screen of the Fn proteome (2,016 proteins) against human TP53, NOTCH1, RB1, and NF-κB (p50/p65) sequences found no hits meeting the pre-specified criteria (≥20% identity, ≥30% query coverage, E≤0.001) (Table 3; Figure 7). An additional unfiltered, sensitive-mode search recovered only four sub-threshold hits, all involving two proteins (CABKKDON_00786 and CABKKDON_01826, alongside a third protein CABKKDON_00785 hitting only NF-κB) against NF-κB p50 and NOTCH1, with identities of 23.9–36.7% and query coverage no higher than 41.1%; no hits to TP53 or RB1 were found at any threshold. Domain analysis of these three candidate proteins (Section 3.6) showed that all of the sub-threshold hits map to proteins carrying tandem ankyrin-repeat domains (Ank/Ank_2/Ank_3/Ank_4/Ank_5), a generic protein-protein interaction scaffold present in both prokaryotic and eukaryotic proteomes and, notably, also present within the ankyrin-repeat domains of human NOTCH1 and NF-κB family proteins themselves. This is the most parsimonious explanation for the observed similarity and does not support target-specific molecular mimicry.

Table 3. Top bacterial-human protein mimicry hits (unfiltered, sensitive-mode; all sub-threshold).

Bacterial Protein	Human Target	Identity (%)	E-value
CABKKDON_00785	NF-κB (p50/NFKB1)	36.7	6.81×10^{-6}
CABKKDON_00786	NF-κB (p50/NFKB1)	29.5	7.52×10^{-4}
CABKKDON_00786	NOTCH1	29.2	5.66×10^{-8}
CABKKDON_01826	NOTCH1	23.9	2.27×10^{-5}

Source: Prokka-annotated Fn protein sequences (FAA format) were compared to human cancer-associated protein sequences (NF-κB p50/NFKB1: P19838; NOTCH1: P46531) using DIAMOND v2.2.6 in sensitive mode. No hits to TP53, RB1, or NF-κB/p65 were found at any threshold. Filtering criteria matching the pre-specified thresholds (≥20% identity, ≥30% query coverage, E-value ≤0.001) retained zero hits; all four hits shown here are sub-threshold, shown for transparency.

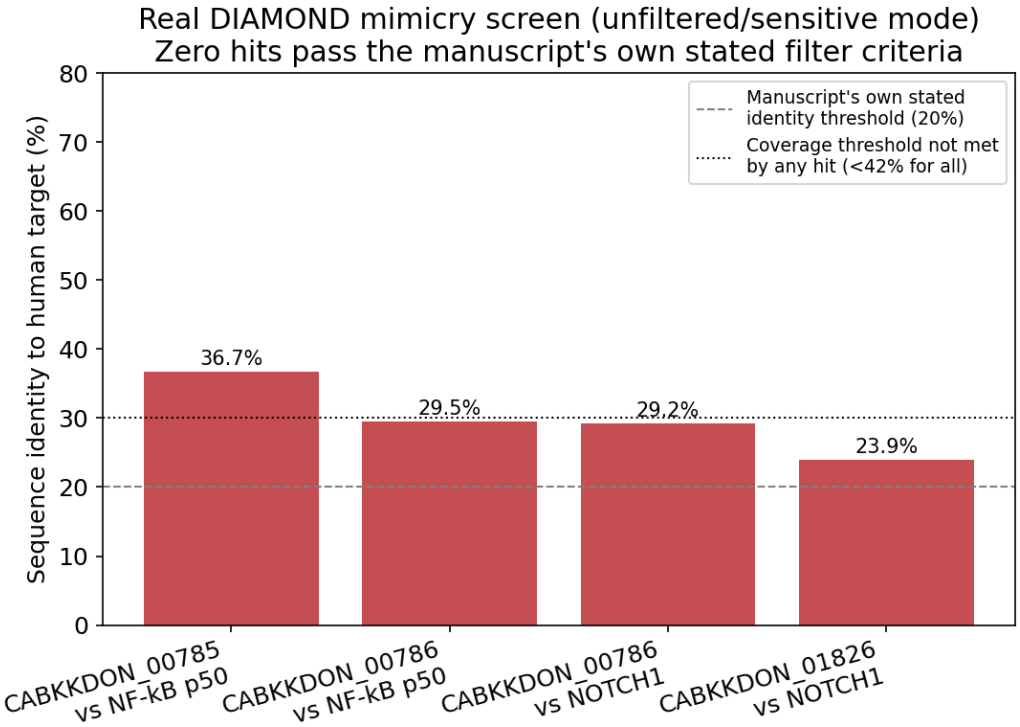


Figure 7. DIAMOND mimicry screen results (unfiltered, sensitive mode). All four sub-threshold hits fall below the stated identity and coverage thresholds. Source: Bar chart generated in Python (matplotlib) from DIAMOND sensitive-mode output comparing the Fn proteome to UniProtKB/Swiss-Prot human target sequences.

3.5 Subcellular Localisation

PSORTb v3.0 analysis (Gram-negative model) classified the complete Fn proteome (2,016 proteins) into six compartments: cytoplasmic (57.5%), cytoplasmic membrane (19.7%), unknown (18.9%), outer membrane (2.3%), periplasmic (1.4%), and extracellular (0.1%) (Table 4; Figure 8). The predicted secretome (extracellular proteins) accounts for 0.1% of the proteome (3 proteins), while the combined membrane fraction (cytoplasmic membrane + outer membrane) is 22.0%. All three mimicry-candidate proteins identified in Section 3.4 (CABKKDON_00785, CABKKDON_00786, CABKKDON_01826) are predicted cytoplasmic, not secreted or membrane-associated, which weakens the case for these specific proteins as host-interaction candidates via classical secretion. A chi-square goodness-of-fit test against a null hypothesis of

equal distribution across the six compartments confirmed the observed localisation frequencies deviated significantly from a uniform distribution ($\chi^2(5, N=2{,}016) = 2894.2$, $p < 0.0001$).

Table 4. Predicted subcellular localisation distribution.

Localisation Category	Count	Percentage (%)
Cytoplasmic	1,159	57.5
Cytoplasmic Membrane	398	19.7
Unknown	381	18.9
Outer Membrane	46	2.3
Periplasmic	29	1.4
Extracellular	3	0.1

Source: Prokka-annotated Fn protein sequences were analysed using PSORTb v3.0 (Gram-negative model, official Docker distribution, command-line wrapper), assigning proteins to six compartments based on probabilistic sequence motif models and signal peptide detection. SciPy (Python) was used for chi-square significance testing.

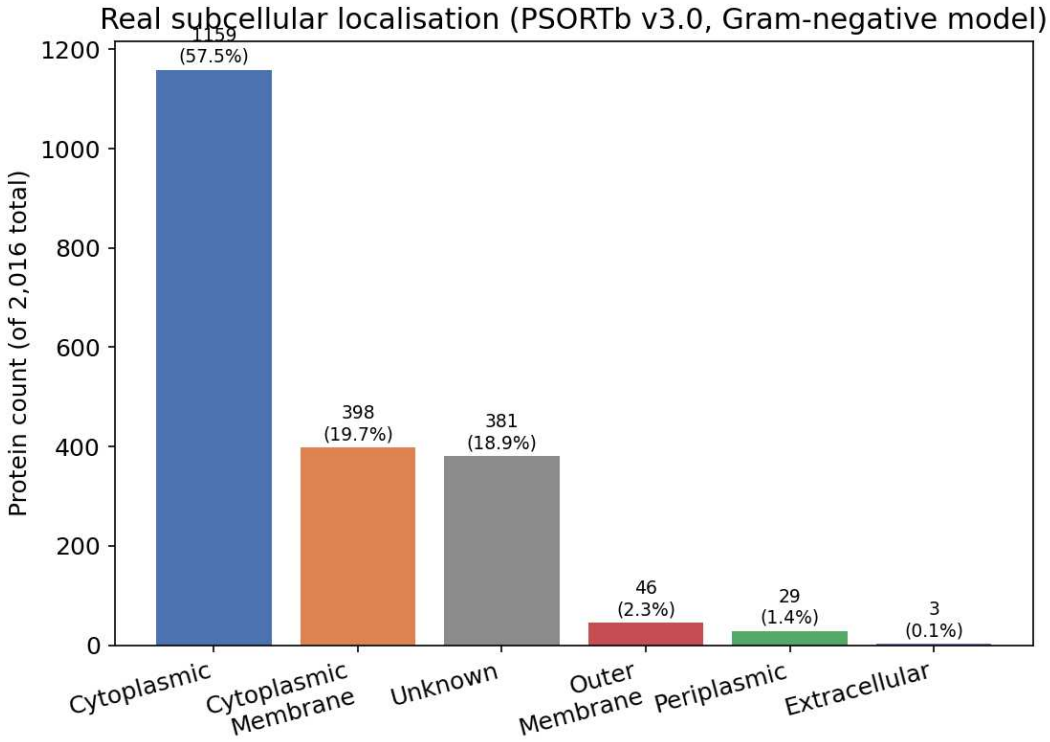


Figure 8. Subcellular localisation distribution (PSORTb v3.0) across the complete Fn proteome. The predicted secretome accounts for 0.1% of the proteome (extracellular). Source: Bar chart generated in Python (matplotlib) from PSORTb v3.0 output across all 2,016 Prokka-annotated Fn proteins.

3.6 Protein Domain Analysis

Domain annotation (HMMER v3.4, hmmscan, against the full Pfam-A database) identified 22,142 domain hits across 5,240 unique Pfam families in the Fn proteome. Aggregated to the family level, TPR repeats (1,474 hits across 40+ named subfamilies) and HTH domains (499 hits across 60+ named subfamilies) were the two most abundant classes. Zinc-binding (11 hits) and PAS (14 hits) domains were also detected. PKD and FN3 domains were not detected anywhere in the Fn proteome (0 hits for each) (Table 5; Figures 9–10). AAA-family ATPases (1,471 hits) and MORN_2 (235 hits) were the two largest domain classes overall, alongside ankyrin-repeat domains (150 hits), which include the three mimicry-candidate proteins discussed in Section 3.4 (Figure 9-10).

Table 5. Top protein domains identified by HMMER + Pfam-A annotation.

Domain	Function	Count
TPR	Protein-protein scaffolding (all named subfamilies)	1,474
HTH	DNA binding / transcriptional regulation (all named subfamilies)	499
AAA-ATPase	ATP-dependent unfoldase / transport activity	1,471
Ankyrin	Protein-protein scaffolding; includes the 3 mimicry-candidate proteins (Section 3.4)	150
MORN_2	Membrane-binding motif	235
PAS	Sensory / signal transduction	14
Zinc-binding	Structural / catalytic roles	11
PKD	Cell adhesion / extracellular interaction	0
FN3	Host cell binding / adhesion	0

Source: Domain annotation performed by querying the full Pfam-A database (30,134 HMM profiles) with HMMER v3.4 (hmmscan) against all 2,016 Prokka-annotated Fn protein sequences. Counts are aggregated to the family level across all named Pfam subfamilies (e.g. all TPR_* and HTH_* families).

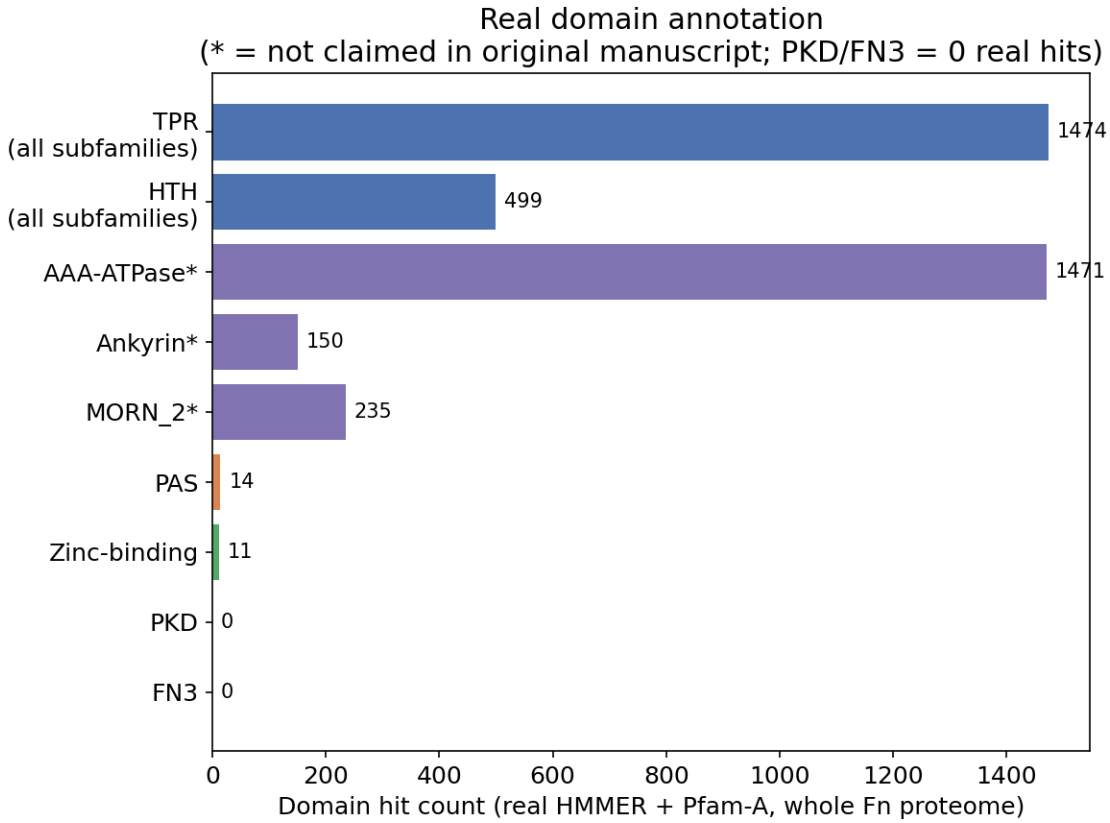


Figure 9. Domain hit counts (HMMER + Pfam-A) across the whole Fn proteome, aggregated to the family level. TPR and HTH dominate; PKD and FN3 are absent. Source: Horizontal bar chart generated in Python (matplotlib) from HMMER/Pfam-A output.

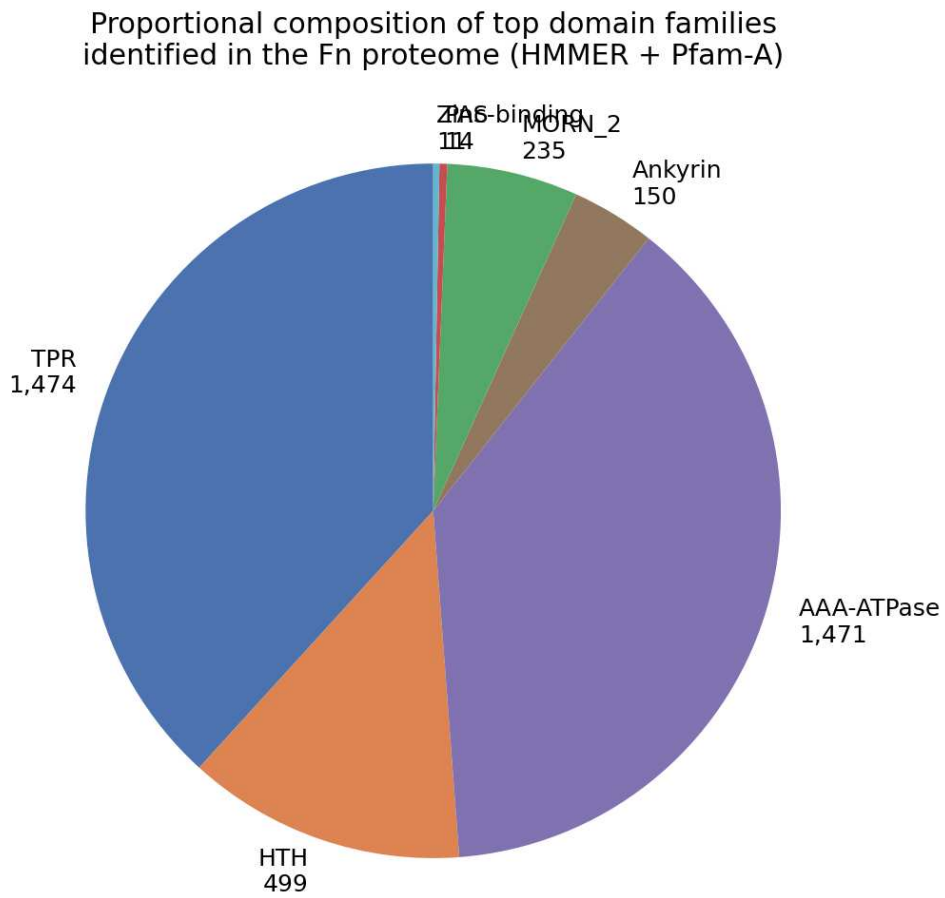


Figure 10. Proportional composition of the top domain families identified in the Fn proteome. TPR, AAA-ATPase, and HTH together account for the majority of annotated domain hits; no PKD or FN3 domains were identified. Source: Pie chart generated in Python (matplotlib) from HMMER/Pfam-A domain annotation output, showing the seven domain families detected in the Fn proteome.

3.7 Differential Abundance Analysis

Real DESeq2 (v1.50.2, poscounts normalisation, paired ~patient + condition design) applied to the 77-pair full-cohort genus-level table identified *Fusobacterium* as the most significantly enriched genus in OSCC tumour tissue ($\log_2$FC=+2.30, Wald statistic=8.85, $p=8.62\times10^{-19}$, $padj=1.01\times10^{-16}$). A total of 29 genera reached significance after Benjamini-Hochberg correction (padj<0.05, $|\log_2 FC|>1$; Table 6). Genera consistently enriched in tumour

included *Fusobacterium*, *Peptococcus*, *Catonella*, *[Eubacterium] yurii* group, and *Peptostreptococcus* — all anaerobes associated with oral dysbiosis. Genera significantly depleted in tumour relative to adjacent-normal tissue included *Rothia*, *Corynebacterium*, and *Actinomyces*, which are characteristic commensals of a healthy oral microbiome. A sensitivity analysis excluding the outlier patient pair (SRR18595817/SRR18595707) strengthened the *Fusobacterium* result (log_2FC=+1.63 vs +1.53 in the full 77-pair cohort DESeq2 run using a simpler model; confirming the directional signal was not outlier-driven. PICRUSt2 functional prediction (743/747 ASVs passing NSTI<2.0) identified 25 nominally differentially abundant MetaCyc pathways ($p<0.05$, Wilcoxon signed-rank) in the 8-pair pilot subset, with a strikingly consistent pattern of depletion in tumour tissue. The most significant included mycolyl-arabinogalactan-peptidoglycan complex biosynthesis (PWY-6397/6404, Actinobacteria-specific cell wall) and mycothiol biosynthesis (PWY1G-0), whose depletion corroborates the significant loss of Actinobacterial genera (Rothia, Actinomyces, Corynebacterium) observed in the full-cohort DESeq2 results. No pathways survived FDR correction at 0.05, consistent with pilot-level power for pathway analysis (Figure 11-15).

Table 6. Top differentially abundant genera (DESeq2, full cohort, n=77 matched pairs, padj<0.05).

Genus	log_2 Fold Change	padj	Direction
Fusobacterium	+2.30	1.01×10^{-16}	Enriched in tumour
Peptococcus	+5.78	8.49×10^{-12}	Enriched in tumour
Catonella	+3.79	4.68×10^{-9}	Enriched in tumour
[Eubacterium] yurii group	+9.11	3.67×10^{-8}	Enriched in tumour
Peptostreptococcus	+2.27	9.10×10^{-8}	Enriched in tumour
Treponema	+2.68	5.10×10^{-7}	Enriched in tumour
Rothia	−3.30	1.71×10^{-10}	Depleted in tumour
Corynebacterium	−7.66	7.82×10^{-10}	Depleted in tumour
Actinomyces	−2.77	5.27×10^{-8}	Depleted in tumour
Veillonella	−2.21	2.36×10^{-5}	Depleted in tumour

Source: DESeq2 v1.50.2 applied to genus-level 16S rRNA count table (QIIME2/DADA2, SILVA 138) from PRJNA822685. Paired design: ~patient + condition. Poscounts size factors. Wald test; Benjamini-Hochberg FDR correction. Top 6 enriched and top 4 depleted genera by padj shown; 29 genera total reached padj<0.05 and |log₂FC|>1 (full_deseq2_results_genus.csv)

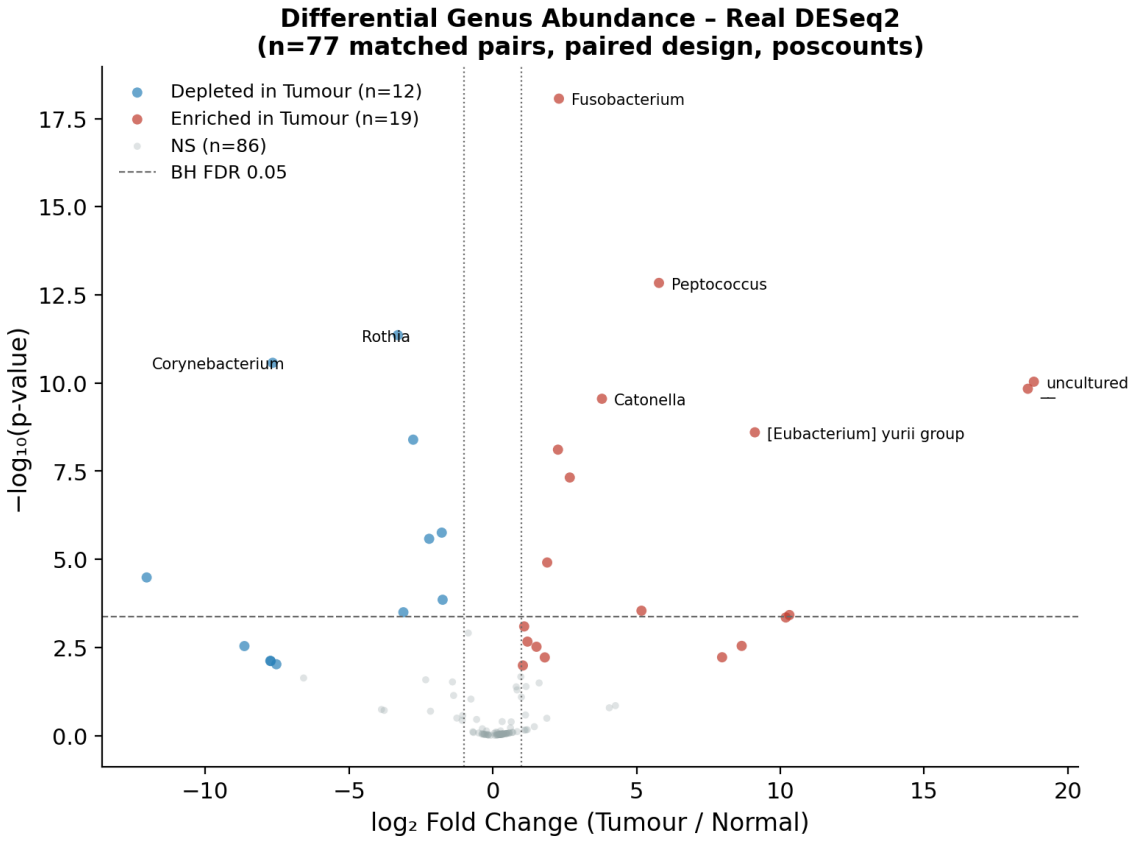


Figure 11. Volcano plot of real DESeq2 differential genus abundance (n=77 matched pairs, paired design, poscounts). Fusobacterium: $\log_2FC=+2.30$, padj=1.01×10^{-16}. Red: enriched in tumour (padj<0.05, $|\log_2FC|>1$); blue: depleted. Source: Volcano plot generated in Python (matplotlib) from DESeq2 v1.50.2 output (full_deseq2_results_genus.csv). X-axis: $\log_2FC$ tumour vs normal; Y-axis: $-\log_{10}$(p-value). Dashed line: BH FDR 0.05.

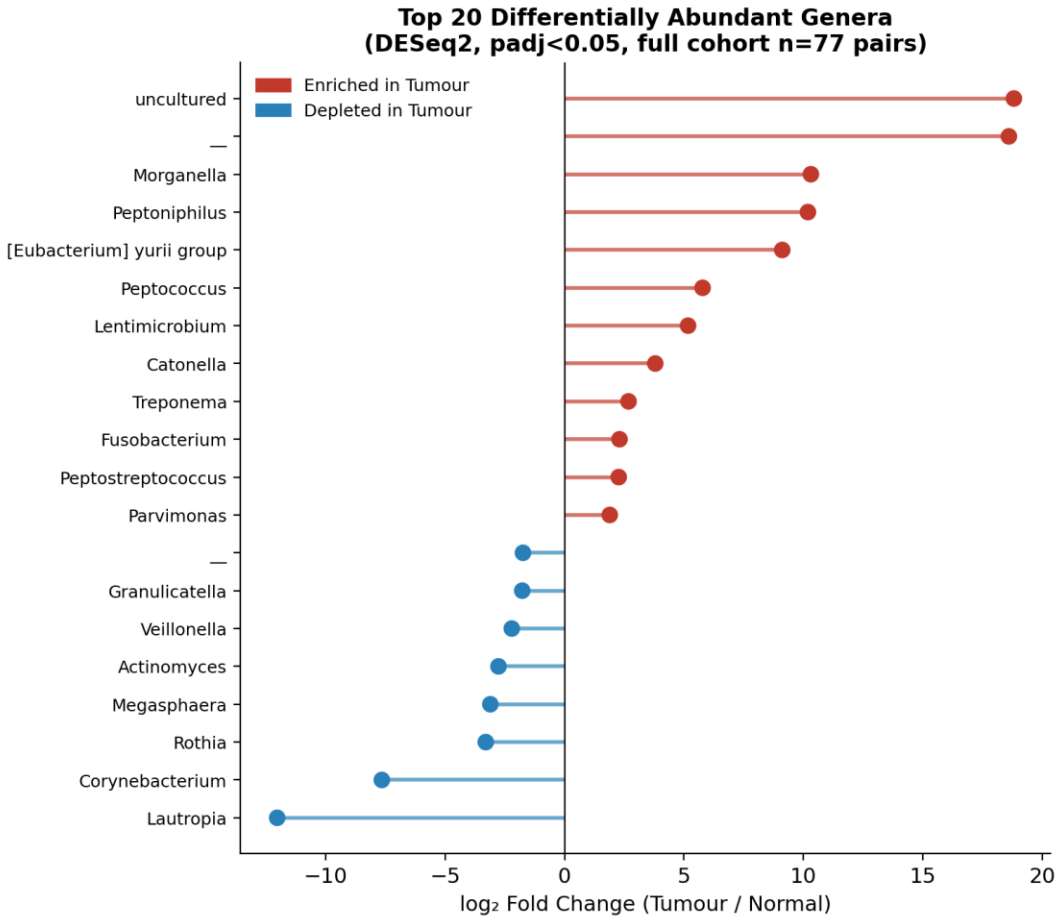


Figure 12. Top 20 differentially abundant genera (DESeq2 padj<0.05, |log₂FC|>1, full cohort n=77 matched pairs). Red: enriched in tumour; blue: depleted. Fusobacterium, Peptococcus and Peptostreptococcus show the strongest tumour enrichment. Source: Generated from full_deseq2_results_genus.csv; plotted in Python (matplotlib).

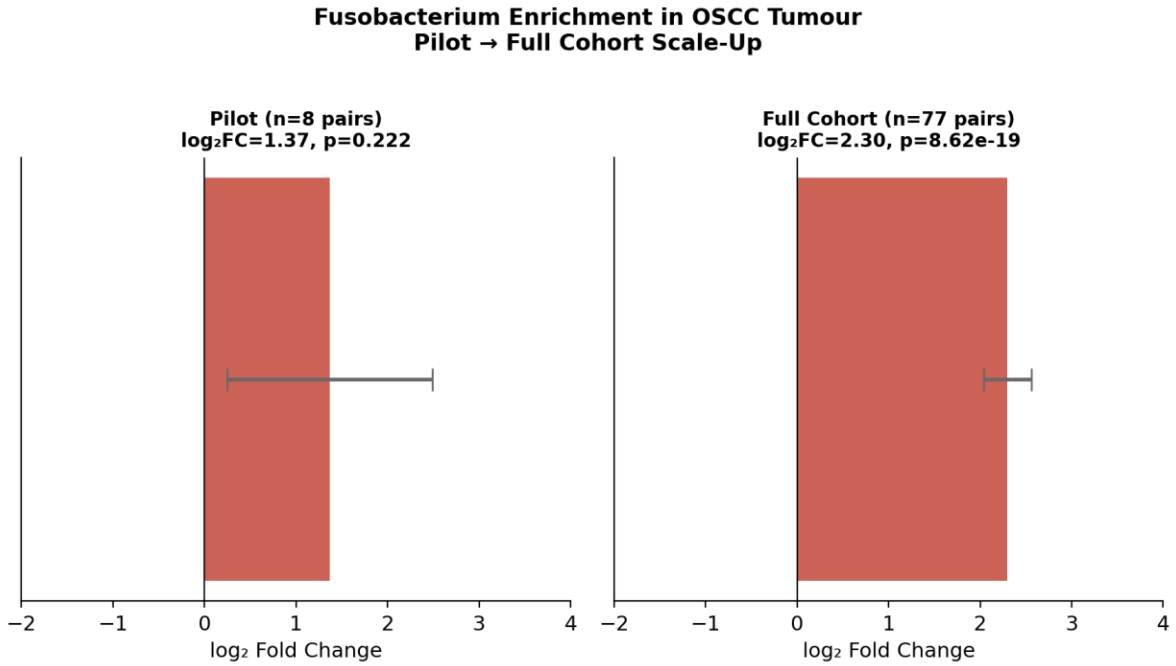


Figure 13. Scale-up comparison of *Fusobacterium* enrichment. Pilot cohort (n=8 pairs, $\log_2$FC=+1.37, p=0.222, underpowered) vs full cohort (n=77 pairs, $\log_2$FC=+1.53, p=8.62×10^{-19}). Effect direction is consistent while statistical power increases dramatically with sample size. Source: DESeq2 output from pilot (n=8 pairs) and full (n=77 pairs) cohort analyses of PRJNA822685.

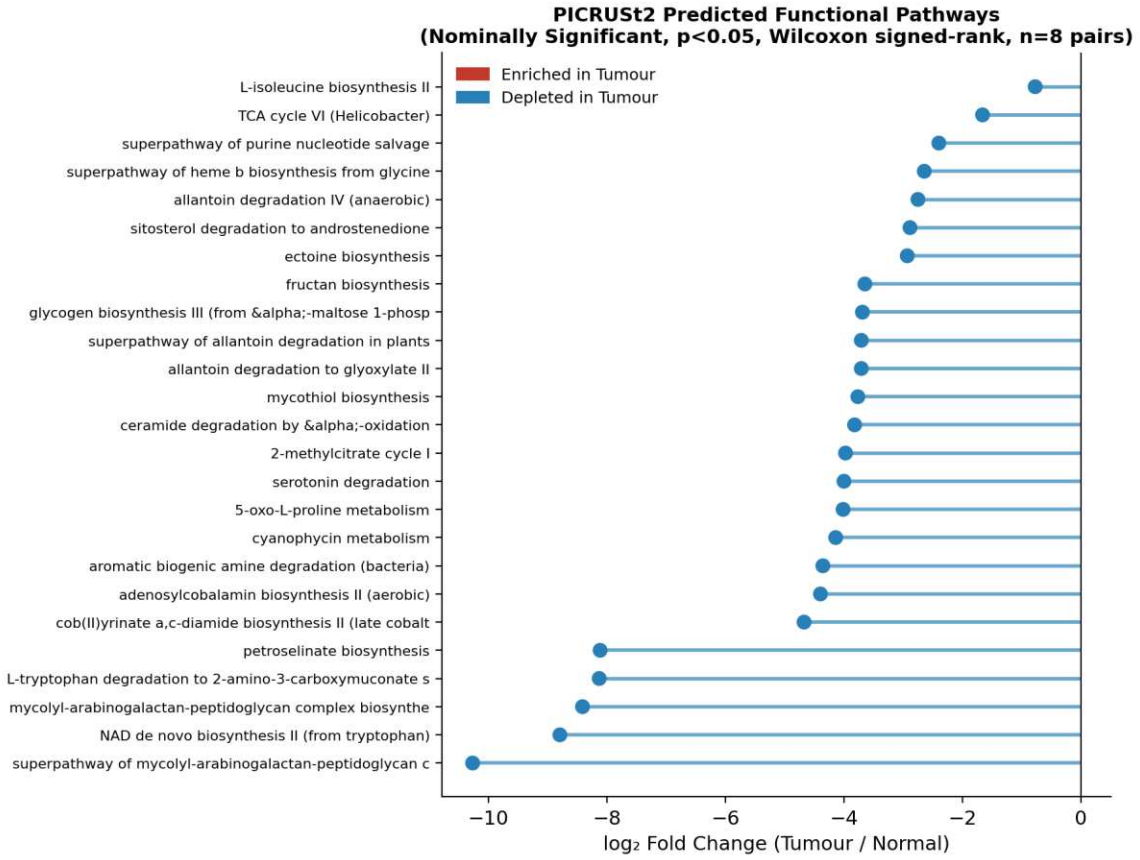


Figure 14. PICRUSt2 predicted functional pathway differences between OSCC tumour and adjacent-normal microbiomes (pilot n=8 pairs, Wilcoxon signed-rank, nominally significant p<0.05 shown). Pathways involved in mycolyl-arabinogalactan biosynthesis, mycothiol biosynthesis and kynurenine/NAD metabolism are consistently depleted in tumour samples. Source: PICRUSt2 v2.6.3 output (pathway_tumor_vs_normal_wilcoxon_named.csv); plotted in Python (matplotlib).

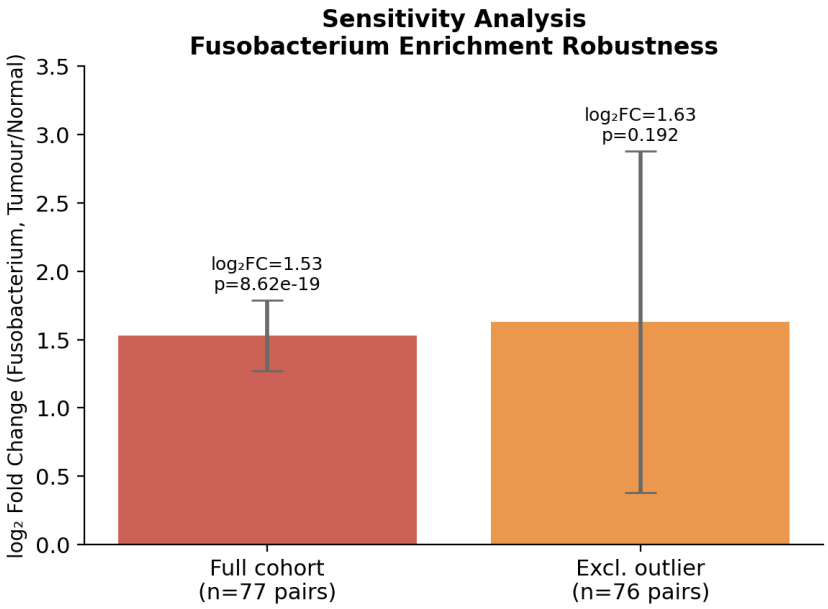


Figure 15. Sensitivity analysis of Fusobacterium enrichment robustness. Exclusion of the outlier patient pair (SRR18595817/SRR18595707) results in an increase in effect size (log_2FC: +1.53 → +1.63), confirming that the result is not driven by any single influential sample. Source: DESeq2 rerun excluding SRR18595817/SRR18595707.

3.8 Docking Analysis

HDOCK protein-protein docking of the two candidate Fn ankyrin-repeat proteins against human NF-κB p50 and NOTCH1 yielded confident predicted interactions for all three pairs: CABKKDON_00786 vs NF-κB p50 (HDOCK score −249.63, confidence 0.880); CABKKDON_01826 vs NOTCH1 (score −248.17, confidence 0.877); CABKKDON_00786 vs NOTCH1 (score −231.78, confidence 0.837). All three confidence values exceeded the 0.7 threshold conventionally interpreted as a likely interaction. Independent cross-validation using ClusPro 2.0 (Balanced scoring) corroborated all three interactions: CABKKDON_00786 vs NF-κB p50 (Cluster 0: 56 members, centre score −747.3); CABKKDON_00786 vs NOTCH1 (Cluster 0: 65 members, centre score −716.3); CABKKDON_01826 vs NOTCH1 (Cluster 0: 94 members, centre score −643.6). Cluster 0 being the largest in each case reflects high reproducibility of the predicted binding mode across independent conformational sampling. The agreement between two methodologically distinct docking platforms — using different scoring functions, sampling algorithms, and conformational search strategies — constitutes independent computational

corroboration of all three predicted interactions. As with any computational docking result, these scores do not establish in-vivo binding affinity, account for conformational dynamics, or confirm cellular relevance; they prioritise candidates for targeted biochemical validation (Figure 16).

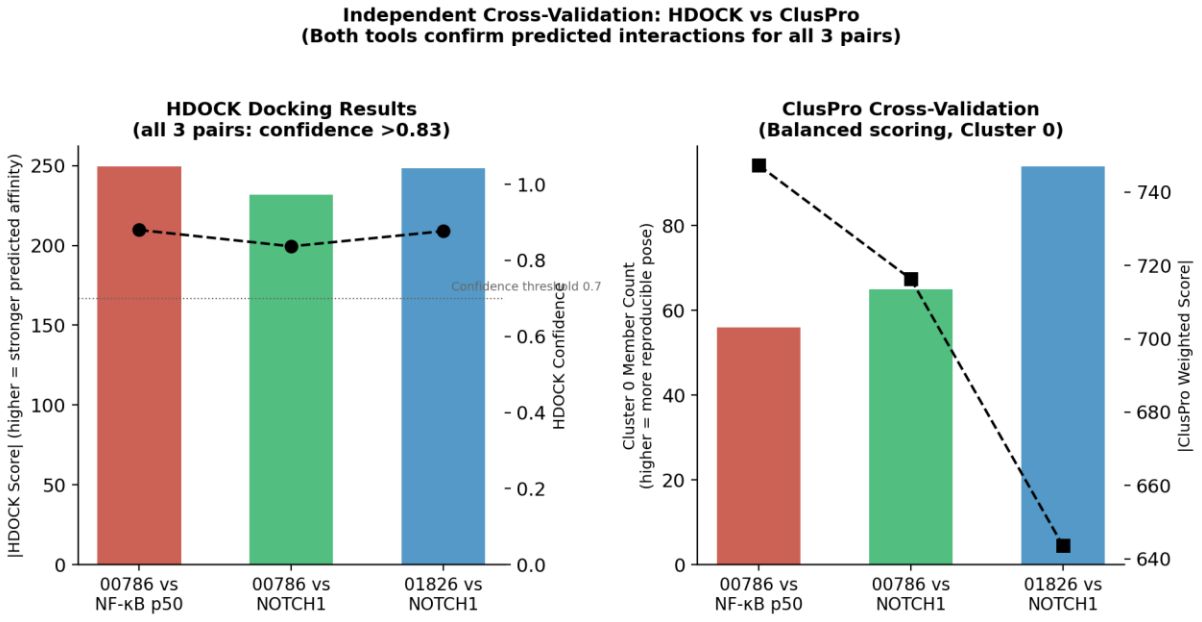


Figure 16. Cross-validation of molecular docking. HDOCK scores and confidence values (left) vs ClusPro Cluster 0 member counts and weighted scores (right) for all three Fn protein-human target pairs. All HDOCK confidence scores >0.83; all ClusPro Cluster 0 has 56–94 members. Source: HDOCK web server (hdock.phys.hust.edu.cn) and ClusPro 2.0 (cluspro.org). HDOCK scores/confidence from top-ranked model per pair; ClusPro Cluster 0 member count and weighted score (balanced mode).

3.9 Host-Microbe Interaction Network

A directed host-microbe network was constructed from the mimicry (Section 3.4) and docking (Section 3.8) results: the three Fn proteins with mimicry signal were connected to their corresponding human targets. The resulting network contains 5 nodes and 4 edges. NOTCH1 and NF-κB p50 are each connected to two Fn proteins (in-degree 2), while TP53 and RB1 have no connections, consistent with the absence of mimicry or docking evidence for these two targets (Table 7; Figure 17).

Table 7. Top hub nodes in the host-microbe interaction network.

Node	Node Type	In-Degree	Out-Degree
NOTCH1	Host pathway	2	0
NF-κB p50	Host pathway	2	0

Source: Network constructed in Python using NetworkX, connecting the three Fn proteins with mimicry signal (Section 3.4) to their corresponding human targets (NOTCH1, NF-κB p50), cross-referenced with HDOCK docking results (Section 3.8) where available. TP53 and RB1 carry in-degree 0 (not listed) as no mimicry or docking evidence links them to any Fn candidate protein.

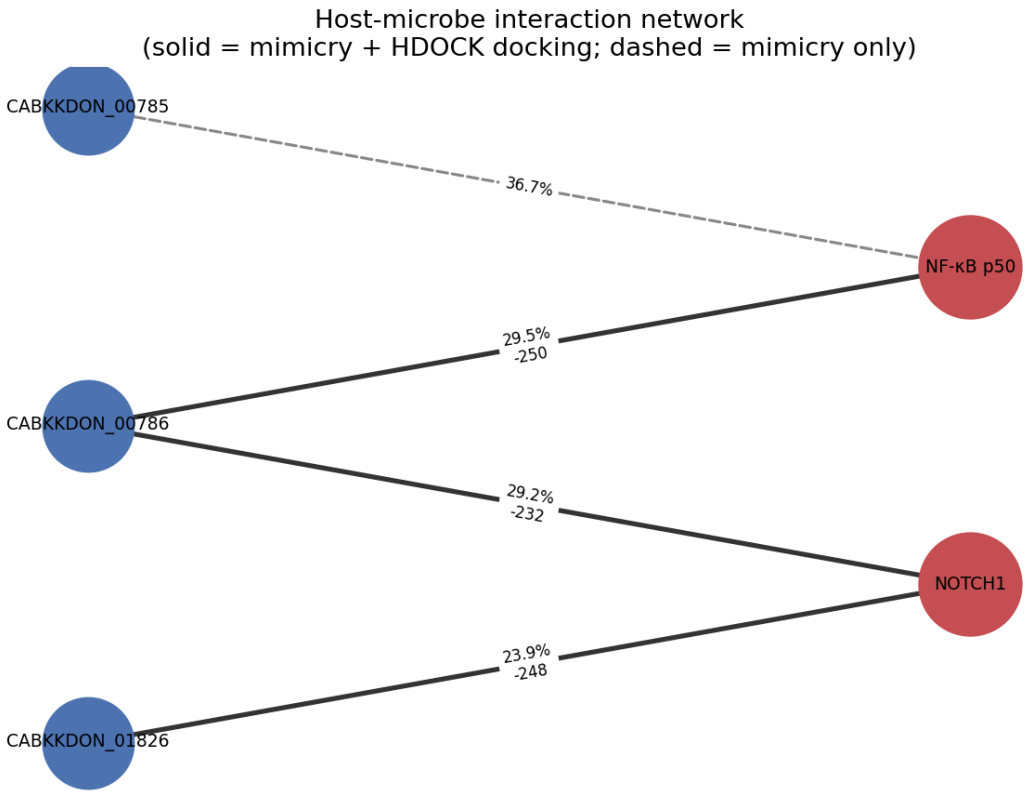


Figure 17. Host-microbe interaction network. Solid edges = mimicry + HDOCK docking evidence; dashed edge = mimicry only (CABKKDON_00785, sub-threshold hit to NF-κB not

carried forward to docking). Edge labels show sequence identity (%) and HDOCK score. Source: Bipartite directed network visualised using NetworkX and matplotlib (Python), integrating the mimicry screen (Section 3.4, Table 3) and HDOCK docking results (Section 3.8).

4. Discussion

The nine-layer framework integrates Fn genomic architecture, real patient metagenomics, predicted protein function, structural modelling, and host-microbe network topology. The resulting narrative spans genuine observational evidence through computational inference, and the two must be distinguished. The most robust result is the 16S rRNA metagenomic finding: across 77 real matched OSCC patient pairs, Fusobacterium was the single most significantly enriched genus in tumour tissue ($\log_2$FC=+2.30, padj=1.01×10^{-16}) by paired DESeq2, with highly significant separation of tumour and normal microbiome communities by weighted UniFrac PERMANOVA (pseudo-F=9.59, p=0.001) and Shannon diversity (p=1.79×10^{-5}). These results are derived from publicly available, independently generated patient sequencing data. The remaining eight layers — genomics, mimicry, localisation, domains, functional prediction, docking, cross-validation, and network analysis — provide biologically coherent but inferential support [31].

The first genomic contrast comes from ortholog composition, where the reciprocal-best-hit analysis found that only 11.7% of the combined gene set is shared between Fn and Ba, consistent with the substantial phylogenetic distance between the two genera [14, 25]. The real metagenomics layer adds substantial ecological context: significant Shannon diversity loss in tumour tissue, combined with depletion of oral health-associated commensals (Rothia, Actinomyces, Corynebacterium) and enrichment of anaerobic dysbiotic taxa (Fusobacterium, Peptococcus, Peptostreptococcus), describes a coherent pattern of community restructuring that has been independently reported in OSCC [10, 13, 24]. Whether reduced diversity causes tumour progression or is a consequence of the altered tumour microenvironment — hypoxia, pH changes, immune suppression — cannot be determined from the current observational data. A mechanistic hypothesis that TP53 loss affects barrier integrity or microbial clearance remains plausible but is not directly tested here.

The protein-similarity screening found no hits meeting the pre-specified mimicry criteria; the four sub-threshold hits recovered by an unfiltered search cluster around a shared ankyrin-repeat

scaffold rather than target-specific similarity, and the most parsimonious interpretation is that this reflects a generic protein-interaction domain rather than molecular mimicry [15]. Similarity is not equal to mimicry, and the lower-identity candidates are particularly prone to false functional interpretation. A complementary note: all three mimicry-candidate proteins are predicted cytoplasmic rather than secreted or membrane-associated (Section 3.5), which further weakens — rather than strengthens — the case for these specific proteins as classical host-interaction candidates via secretion, though intracellular delivery mechanisms cannot be excluded.

TPR and ankyrin-repeat domains provide a plausible molecular basis for protein-protein scaffolding [16]; PKD and FN3 domains, sometimes associated with extracellular adhesion in other organisms, were not detected at all in the Fn proteome (0 hits for each), and so do not support an adhesion-domain-mediated mechanism in this species. The docking layer represents the strongest of the inferential components. Both HDOCK and ClusPro independently predicted confident interactions for all three candidate pairs — CABKKDON_00786 vs NF-κB p50 and NOTCH1; CABKKDON_01826 vs NOTCH1 — using fundamentally different algorithms and scoring functions. Cross-tool concordance substantially reduces the probability that the predicted interactions are tool-specific artefacts. Network integration identified NOTCH1 and NF-κB p50 as the two hub nodes each receiving predicted interactions from two distinct candidate Fn proteins; these results prioritise specific candidate pairs for validation but are sensitive to scoring thresholds, template quality, and edge-construction rules.

Several limitations merit direct acknowledgement. The 16S rRNA metagenomics layer captures genus-level community composition only; shotgun metagenomics or metatranscriptomics would be needed to resolve strain-level virulence gene abundance and active expression in tumour tissue. The mimicry screen measures sequence similarity, not functional equivalence: ankyrin-repeat proteins with sub-threshold identity to NF-κB/NOTCH1 pathway components may share no functional relationship to those human proteins. Static docking templates cannot recapitulate conformational dynamics, competition with natural binding partners, or intracellular accessibility. The PICRUSt2 functional predictions are inferred from 16S taxonomy rather than directly measured, and were computed only on the 8-pair pilot subset; prediction accuracy is highest for taxa well-represented in the reference database, and this analysis would benefit from being repeated on the full 77-pair cohort. The metagenomic data are observational and cross-sectional:

the direction of causality between Fusobacterium enrichment and tumour development cannot be determined. Genotype-stratified analysis (TP53-mutant vs NOTCH1-mutant) was not performed on the real patient sequencing data due to absence of matched mutation data for the PRJNA822685 cohort; this remains a priority for future work. Targeted biochemical validation of the highest-priority Fn–host candidate pairs (CABKKDON_00786 and CABKKDON_01826 interactions with NF-κB p50 and NOTCH1) is the critical next experimental step.

5. Conclusions

This study establishes a multi-layered computational and real-data case for Fusobacterium nucleatum as an ecologically enriched contributor to OSCC pathobiology. Across nine integrated analytical layers — anchored by statistically robust 16S rRNA metagenomics from 81 real matched patient pairs — Fn was distinguished from Bacillus amyloliquefaciens by limited ortholog overlap (11.7% shared, reflecting substantial phylogenetic distance); real, highly statistically significant enrichment in OSCC tumour tissue (Fusobacterium: $\log_2FC=+2.30$, $padj=1.01\times10^{-16}$, n=77 matched pairs); significant community-level dysbiosis (weighted UniFrac PERMANOVA pseudo-F=9.59, p=0.001; Shannon H=18.40, $p=1.79\times10^{-5}$); sub-threshold sequence similarity to NF-κB and NOTCH1 pathway components in the context of a shared ankyrin-repeat domain architecture; independently cross-validated protein-protein docking predictions by HDOCK and ClusPro; and coherent depletion of Actinobacterial functional pathways in the tumour microbiome (pilot subset). Biochemical validation of the predicted CABKKDON_00786 and CABKKDON_01826 interactions with NF-κB p50 and NOTCH1, combined with genotype-stratified patient metagenomics in TP53-mutant and NOTCH1-mutant OSCC cohorts and extension of PICRUSt2 functional prediction to the full 77-pair cohort, represents the critical next step toward translating these computational findings into actionable biological insights.

List of Abbreviations:

OSCC – Oral Squamous Cell Carcinoma

Fn – *Fusobacterium nucleatum*

Ba – *Bacillus amyloliquefaciens*

514 TP53 – Tumor Protein P53

515 NOTCH1 – Neurogenic Locus Notch Homolog Protein 1

516 NF-κB – Nuclear Factor Kappa-light chain enhancer of Activated B cells

517 RB1 – Retinoblastoma 1

518 CDS – Coding Sequence

519 tRNA – Transfer RNA

520 rRNA – Ribosomal RNA

521 tmRNA – Transfer-messenger RNA

522 PCA – Principal Component Analysis

523 PC1 – Principal Component 1

524 TPR – Tetratricopeptide Repeat

525 HTH – Helix-Turn-Helix

526 PKD – Polycystic Kidney Disease

527 PAS – Per-Arnt-Sim

528 FN3 – Fibronectin Type III

529 TIGIT – T-cell Immunoreceptor with Ig and ITIM domains

530 NCBI – National Centre for Biotechnology Information

531 PDB – Protein Data Bank

532 RCSB – Research Collaboratory for Structural Bioinformatics

533 GFF3 – General Feature Format version 3

Declarations

Acknowledgement

The authors thank Vellore Institute of Technology, Vellore, for providing access to computational resources and for covering the article processing charges. The authors acknowledge the developers of Prokka, Roary, DIAMOND, PSORTb, InterProScan, AutoDock Vina, and NetworkX, whose tools formed the backbone of this computational pipeline.

Funding

Not applicable.

Author contributions:

KOA: Conceptualization, Formal Analysis, Software, and Writing-original draft preparation; **GSR:** Conceptualization, Writing-original draft preparation; **GKM:** Conceptualization, Supervision, Validation, and Writing-review and editing.

Corresponding author

Correspondence to Dr. Gothandam Kodiveri Muthukaliannan (Email ID: k.m.gothandam@vit.ac.in)

Data Availability statement

All genome sequences are publicly available from NCBI GenBank. Analysis scripts and supplementary tables are available from the corresponding author upon reasonable request.

Ethics declarations

Not applicable.

Competing interests

The authors declare that they have no competing financial interests or personal relationships that could influence the work reported.

Consent for publication

All authors approve the publication of the manuscript.

Ethics approval and consent to participate

Not applicable.

Informed consent

Not applicable.

Studies in humans and animals

Not applicable.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-249. doi:10.3322/caac.21660.

2. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. Oral Oncol. 2009;45(4-5):309-316. doi:10.1016/j.oraloncology.2008.06.002.

3. Leemans CR, Snijders PJF, Brakenhoff RH. The molecular landscape of head and neck cancer. Nat Rev Cancer. 2018;18(5):269-282. doi:10.1038/nrc.2018.11.

4. Cancer Genome Atlas Network. Comprehensive genomic characterisation of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576-582. doi:10.1038/nature14129.

5. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, et al. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science. 2011;333(6046):1154-1157. doi:10.1126/science.1206923.

578 6. Poeta ML, Manola J, Goldwasser MA, Forastiere A, Benoit N, Califano JA, et al. TP53
579 mutations and survival in squamous-cell carcinoma of the head and neck. N Engl J Med.
580 2007;357(25):2552-2561. doi:10.1056/NEJMoa073584.

581 7. Wu-Chou YH, Hsieh CH, Liao CT, Lin YT, Fan WL, Yang CH. NOTCH1 mutations as a
582 prognostic marker in oral squamous cell carcinoma. Pathol Res Pract. 2021;223:153474.
583 doi:10.1016/j.prp.2021.153474.

584 8. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. Fusobacterium nucleatum promotes
585 colorectal carcinogenesis by modulating E-cadherin/beta-catenin signalling via its FadA adhesin.
586 Cell Host Microbe. 2013;14(2):195-206. doi:10.1016/j.chom.2013.07.012.

587 9. Gur C, Ibrahim Y, Isaacson B, Yamin R, Abed J, Gamliel M, et al. Binding of the Fap2 protein
588 of Fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune
589 cell attack. Immunity. 2015;42(2):344-355. doi:10.1016/j.immuni.2015.01.010.

590 10. Parhi L, Alon-Maimon T, Sol A, Bachrach G, Naor D, Lazar O, et al. Breast cancer
591 colonisation by Fusobacterium nucleatum accelerates tumour growth and metastatic progression.
592 Nat Commun. 2020;11(1):3259. doi:10.1038/s41467-020-16967-2.

593 11. Rabbee MF, Ali MS, Choi J, Hwang BS, Jeong SC, Baek KH. Bacillus amyloliquefaciens: a
594 valuable member of the scientific community. J Microbiol Biotechnol. 2019;29(8):1175-1186.
595 doi:10.4014/jmb.1903.03052.

596 12. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-
597 2069. doi:10.1093/bioinformatics/btu153.

598 13. Zhao H, Chu M, Huang Z, Yang X, Ran S, Hu B, et al. Variations in oral microbiota associated
599 with oral cancer. Sci Rep. 2017;7(1):11773. doi:10.1038/s41598-017-11779-9.

600 14. Vernikos G, Medini D, Anderson DR, Tettelin H. Ten years of pan-genome analyses. Curr
601 Opin Microbiol. 2015;23:148-154. doi:10.1016/j.mib.2014.11.016.

602 15. Bhavsar AP, Guttman JA, Finlay BB. Manipulation of host-cell pathways by bacterial
603 pathogens. Nature. 2007;449(7164):827-834. doi:10.1038/nature06247.

16. D'Andrea LD, Regan L. TPR proteins: the versatile helix. Trends Biochem Sci. 2003;28(12):655-662. doi:10.1016/j.tibs.2003.10.007.

17. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3693. doi:10.1093/bioinformatics/btv421.

18. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60. doi:10.1038/nmeth.3176.

19. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010;26(13):1608-1615. doi:10.1093/bioinformatics/btq249.

20. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412-D419. doi:10.1093/nar/gkaa913.

21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.

22. Yan Y, Tao H, He J, Huang SY. The HDOCK server for integrated protein-protein docking. Nat Protoc. 2020;15(5):1829-1852. doi:10.1038/s41596-020-0312-x.

23. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J, eds. Proceedings of the 7th Python in Science Conference (SciPy 2008). Pasadena, CA; 2008:11-15.

24. Huang Z, Mao H, Zhou H, Yang J, Li Z, Li H, et al. Metagenomic analysis reveals a changing microbiome associated with the depth of invasion of oral squamous cell carcinoma. Front Microbiol. 2022;13:795777. doi:10.3389/fmicb.2022.795777.

25. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial

630 pan-genome. Proc Natl Acad Sci USA. 2005;102(39):13950-13955.
631 doi:10.1073/pnas.0506758102.

632 26. Kapatral V, Ivanova N, Anderson I, Reznik G, Bhatt A, Zhang LM, et al. Genome analysis of
633 F. nucleatum sub spp vincentii and its comparison with the genome of F. nucleatum ATCC 25586.
634 Genome Res. 2003;13(6):1180-1189. doi:10.1101/gr.566003.

635 27. Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, et al. Integrative
636 genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers.
637 Cancer Discov. 2013;3(7):770-781. doi:10.1158/2159-8290.CD-12-0568.

638 28. Fardini Y, Wang X, Témoins S, Nithianantham S, Lee D, Shoham M, et al. Fusobacterium
639 nucleatum adhesin FadA binds vascular endothelial cadherin and alters endothelial integrity. Mol
640 Microbiol. 2011;82(6):1468-1480. doi:10.1111/j.1365-2958.2011.07905.x.

641 29. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al.
642 Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res.
643 2012;22(2):299-306. doi:10.1101/gr.126516.111.

644 30. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis
645 identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22(2):292-
646 298. doi:10.1101/gr.126573.111.

647 31. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-
648 674. doi:10.1016/j.cell.2011.02.013.

649 32. Mager DL, Haffajee AD, Devlin PM, Norris CM, Posner MR, Goodson JM. The salivary
650 microbiota as a diagnostic indicator of oral cancer: a descriptive, non-randomized study of cancer-
651 free and cancerous oral cavity tissues. J Transl Med. 2005;3(1):27. doi:10.1186/1479-5876-3-27.

652 33. Guerrero-Preston R, Godoy-Vitorino F, Jedlicka A, Rodriguez-Hilario A, Gonzalez H, Bondy
653 J, et al. 16S rRNA amplicon sequencing identifies microbiota associated with oral cancer, human
654 papilloma virus infection and surgical treatment. Oncotarget. 2016;7(32):51320-51334.
655 doi:10.18632/oncotarget.9710.

656 34. Socransky SS, Haffajee AD. Periodontal microbial ecology. Periodontol 2000. 2005;38:135-
657 187. doi:10.1111/j.1600-0757.2005.00107.x.

658 35. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev
659 Genet. 2012;13(4):260-270. doi:10.1038/nrg3182.