massivebioinformatics.com info@massivebioinformatics.com

MASSIVE BIOINFORMATICS

FUNGAL GENOME ANALYSIS REPORT

by Massive Bioinformatics

Version: 1.0.0

Sample Name: S01

EXECUTIVE SUMMARY

Sample **S01** was sequenced on Oxford Nanopore Technology, yielding **160,959** reads. After quality filtering, **155,691** reads (96.7%) passed quality thresholds with a mean read length of **7,594 bp**.

De novo assembly produced **7 contigs** with a total genome size of **12,360,088 bp** and an N50 of **2,313,679 bp** (mean coverage: 89.7x).

BUSCO completeness assessment against the ascomycota_odb12 lineage showed **53.1%** complete orthologs (51.9% single-copy, 1.2% duplicated), with 5.3% fragmented and 41.6% missing.

Taxonomic identification classified this isolate as **Candidozyma auris** with 99.47% ANI.

Clade classification identified this isolate as **Clade I** (QC: PASS).

Reference-based variant calling identified **78,294** variants, of which **394** were high-impact and **10550** were moderate-impact.

ChroQueTas antifungal resistance screening detected **9 resistance-associated mutation(s)** in target genes with associated drug susceptibility predictions.

Key Metrics

Total Reads	**160,959**
Reads After Filtering	**155,691**
Species Identification	**Candidozyma auris**
Average Nucleotide Identity	**99.47%**
Genome Size	**12,360,088 bp**
Number of Contigs	**7**
N50	**2,313,679 bp**
BUSCO Completeness	**53.1%**
Clade Classification	**Clade I**
Resistance Variants	**5 variants**

TERMINOLOGY

Antimicrobial Resistance (AMR)

The ability of a microorganism to resist the effects of antimicrobial agents that would normally inhibit or kill it.

Antifungals

Pharmaceutical agents used to treat fungal infections by targeting essential fungal cellular processes such as ergosterol biosynthesis, cell wall synthesis, or nucleic acid metabolism.

Antifungal Resistance (AFR)

The acquired or intrinsic ability of a fungal pathogen to survive exposure to antifungal drugs at concentrations that would normally be lethal or inhibitory.

AFR Profiling

The systematic assessment of a fungal isolate's resistance or susceptibility to a panel of antifungal agents, typically through genomic analysis of known resistance-associated mutations.

Confidence Score

A numerical value assigned to resistance-associated mutations indicating the strength of evidence. A positive score denotes mutations reported to confer resistance; a negative score relates to mutations reported in susceptible strains.

MIC (Minimum Inhibitory Concentration)

The lowest concentration of an antimicrobial agent that prevents visible growth of a microorganism under defined conditions. Lower MIC values indicate greater susceptibility; higher values indicate resistance.

Drug Susceptibility

The degree to which a microorganism is affected by an antimicrobial agent, determined by MIC testing or genomic prediction of resistance markers.

Fungicide Resistance

The ability of a fungal organism to tolerate exposure to a fungicidal compound, often caused by mutations in drug target genes (e.g., ERG11/Cyp51 for azoles, FKS1 for echinocandins).

Clades

Distinct genetic lineages within a species that share a common ancestor. In Candidozyma auris, six clades (I–VI) have been identified, each associated with specific geographic regions and resistance profiles.

1. METHOD

1.1 Wet Lab Protocol

DNA Quantification: Genomic double-stranded DNA (dsDNA) concentration of each isolated bacterial sample was quantified using the Qubit™ dsDNA HS Assay Kit on a Qubit™ 4 Fluorometer, following the manufacturer's instructions.

End-Prep Reaction: Library preparation was performed following the "Ligation sequencing gDNA – Native Barcoding Kit 96 V14 (SQK-NBD114.96)" protocol. For each sample, 400 ng of DNA was transferred to a 0.2 mL PCR tube and brought to a final volume of 11 µL with nuclease-free water. End-prep reagents were then added: 1 µL Diluted DCS, 0.875 µL NEBNext FFPE DNA Repair Buffer, 0.5 µL NEBNext FFPE DNA Repair Mix, 0.875 µL Ultra II End-prep Reaction Buffer, and 0.75 µL Ultra II End-prep Enzyme Mix. The reaction was incubated in a thermal cycler at 20°C for 5 minutes followed by 65°C for 5 minutes.

Native Barcode Ligation: Following end-prep, native barcode ligation was performed by combining 0.75 µL of the end-prepped DNA with 2 µL nuclease-free water, 1.25 µL Native Barcode (NB01–96), and 5 µL Blunt/TA Ligase Master Mix. The mixture was briefly vortexed and incubated at room temperature for 20 minutes. Reactions were terminated using EDTA. Barcoded samples were pooled and purified using 0.4x AMPure XP (AXP) beads. The bead mixture was incubated at 37°C for 10 minutes, placed on a magnetic stand, and the supernatant was removed. Two washes with 125 µL Short Fragment Buffer (SFB) were performed before eluting DNA with 37 µL nuclease-free water.

Adapter Ligation: Adapter ligation was performed by combining 35 µL pooled barcoded DNA with 5 µL Native Adapter (NA), 50 µL 2x Rapid Buffer, and 10 µL T4 DNA ligase (Rapid). The mixture was incubated at 30°C for 10 minutes. A final AXP bead purification was performed: 40 µL AXP beads were added, incubated at 37°C for 10 minutes, washed with 125 µL SFB, and the adapted library was eluted in 25 µL Elution Buffer (EB). The final library concentration was measured by fluorometric quantification.

Sequencing: The sequencing library was prepared by combining 1170 µL Flow Cell Flush (FCF) with 30 µL Flow Cell Tether (FCT) as the priming mixture. The flow cell (R10.4.1 chemistry) was primed by drawing 200 µL air from the inlet port and adding 500 µL priming mixture. After a 5-minute incubation, 100 fmol of the adapted library was combined with 100 µL Sequencing Buffer (SB) and 68 µL Loading Beads (LIB), and 200 µL was loaded onto the flow cell via the inlet port. Sequencing was performed on a PromethION 2 Solo or GridION (Oxford Nanopore Technologies) and controlled using the MinKNOW software. Base-calling and demultiplexing were performed using Dorado with the super-accuracy model.

1.2 Bioinformatics Analysis

Raw sequencing data in POD5 format was base-called and demultiplexed using Dorado (Oxford Nanopore Technologies), producing FASTQ files for downstream analysis.

Quality Control: Long-read quality filtering was performed using Fastplong. Reads were trimmed (15 bp from the 5' end), and filtered based on minimum mean quality score (Q15), minimum read length (150 bp), and maximum percentage of unqualified bases (30%). Quality metrics including read length distributions, GC content, and per-base quality profiles were assessed before and after filtering.

De Novo Assembly: Genome assembly was performed using Flye, a long-read de novo assembler optimized for Oxford Nanopore data, with an estimated genome size of 12m and target assembly coverage of 50x. The raw assembly was post-processed to remove repeat-flagged contigs based on Flye's internal repeat classification, and remaining contigs were sorted by length in descending order. Filtered reads were re-aligned to the cleaned assembly using minimap2 (preset: map-ont) to generate a coordinate-sorted BAM file. The consensus sequence was then improved by neural-network-based polishing with Medaka (model: r1041_e82_400bps_sup_v5.0.0), which performs inference on signal-level features to correct systematic basecalling errors.

Quality Assessment: Assembly completeness was evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs) in genome mode against the auto lineage dataset, reporting the percentage of complete (single-copy and duplicated), fragmented, and missing orthologs. Additionally, EukCC2 was used to provide an independent estimate of eukaryotic genome completeness and contamination based on conserved marker gene sets. Taxonomic identification was performed using sourmash MinHash k-mer signatures (k=51, scaled=10,000) against the NCBI eukaryotic reference database, with species-level classification determined by the lowest common ancestor (LCA) approach using curated taxonomy lineages. Average nucleotide identity (ANI) was computed from the sourmash containment score to quantify similarity to the closest reference genome.

Reference-Based Variant Calling: Based on the sourmash species identification, a matching reference genome was automatically selected from the species-to-reference mapping table and downloaded from NCBI. Reads were aligned to the reference genome using minimap2 (preset: map-ont) and the alignment was polished with Medaka to improve consensus accuracy. Variant calling was performed using Clair3 in haploid-sensitive mode with the Oxford Nanopore SUP model (r1041_e82_400bps_sup_v5.0.0). Raw variants were filtered to retain high-confidence calls. Variants were functionally annotated using SnpEff with a custom-built database derived from the reference genome annotation (GFF3), classifying each variant by predicted impact (HIGH, MODERATE, LOW, MODIFIER) and effect type (missense, frameshift, stop gained/lost, synonymous, etc.).

Antifungal Resistance Screening: A targeted resistance gene filter was applied to extract protein-coding mutations (missense, frameshift, stop gained, stop lost) in 203 antifungal resistance-associated genes spanning azole targets (ERG11/CYP51, ERG3, ERG6), echinocandin targets (FKS1, FKS2), pyrimidine analogue resistance (FCY1, FCY2, FUR1), efflux pumps (CDR1, CDR2, MDR1), and transcriptional regulators (TAC1, MRR1, PDR1, UPC2). Detected mutations were cross-referenced against the FungAMR database (35,792 curated entries covering 95 fungal species and 246 genes) and MARDy database (76,098 entries across 120 species and 216 genes) to identify known resistance-associated variants with confidence scores and MIC values. In parallel, mutation-based antifungal resistance phenotype prediction was performed using ChroQueTas, which screens the polished assembly against the FungAMR protein database (105 resistance genes across 57 fungal species) using BLAST alignment with minimum identity of 90% and minimum coverage of 75%. For each gene, amino acid substitutions were identified relative to the reference protein sequences and cross-referenced with curated resistance mutations. ChroQueTas reports drug susceptibility changes for azoles, echinocandins, 5-flucytosine, and polyenes, with associated MIC breakpoints and confidence scores ranging from -3 (sensitivity-associated) to +3 (resistance-associated).

Virulence Factor Screening: The polished assembly was screened against PHI-base v4.19 (~10,000 curated pathogen-host interaction protein entries) using diamond blastx with a pre-filter of 40% identity and 70% subject coverage. Hits were then filtered to >= 80% protein identity and constrained to the sample's sourmash-identified species; if no same-species hits remained, results fell back to same-genus hits (Candida and Candidozyma are treated as the same genus to accommodate the 2024 reclassification of the C. auris clade). Each retained hit reports the PHI-base entry, gene name, host pathogen species, percent identity, e-value, and associated phenotype (e.g. unaffected pathogenicity, reduced virulence, loss of pathogenicity).

2. QC STATISTICS

Summary

	Before Filtering	After Filtering
Total Reads	160959	155691
Read Mean Length	7591	7594
GC Content	44.22%	44.27%

Filtering Result

	Count	%
Passed Filter Reads	155,691	96.73%
Low Quality Reads	5,248	3.26%
Too Many N Reads	0	0.00%
Too Short Reads	20	0.01%
Too Long Reads	0	0.00%

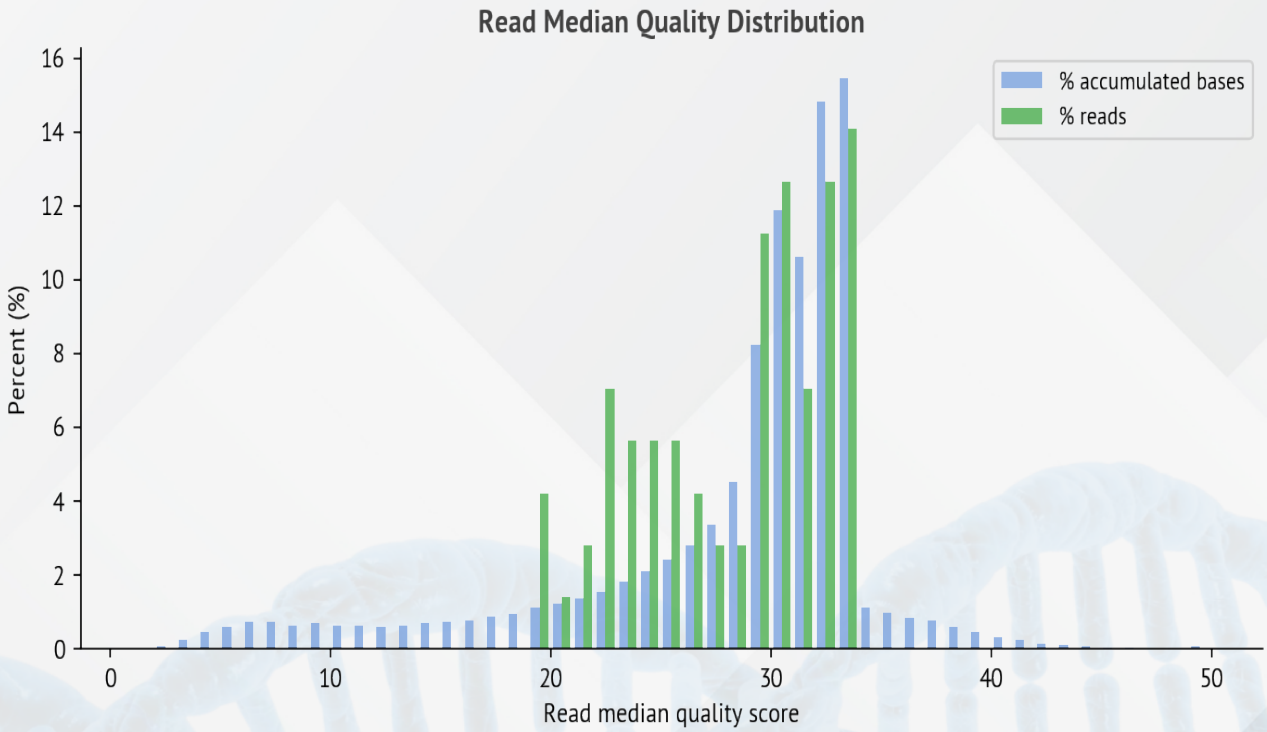


Figure 1. Per-Read Quality Distribution

3. ASSEMBLY STATISTICS

Assembly Summary

Total Contigs	7
Total Assembly Length	12,360,088 bp
Largest Contig	4,173,455 bp
N50	2,313,679 bp
Mean Coverage	89.7x

Contig Details

Contig	Length (bp)	Coverage	Circ.
contig_1	4,173,455	91x	N
contig_2	2,313,679	89x	N
contig_3	1,695,641	92x	N
contig_4	1,441,034	90x	N
contig_5	1,002,150	89x	N
contig_6	954,134	88x	N
contig_7	779,995	89x	N

4. GENOME COMPLETENESS

BUSCO (Benchmarking Universal Single-Copy Orthologs) assesses genome assembly completeness by searching for conserved genes expected to be present in a given lineage. A high percentage of complete BUSCOs indicates a high-quality, near-complete assembly.

4.1 EukCC2 Genome Quality Assessment

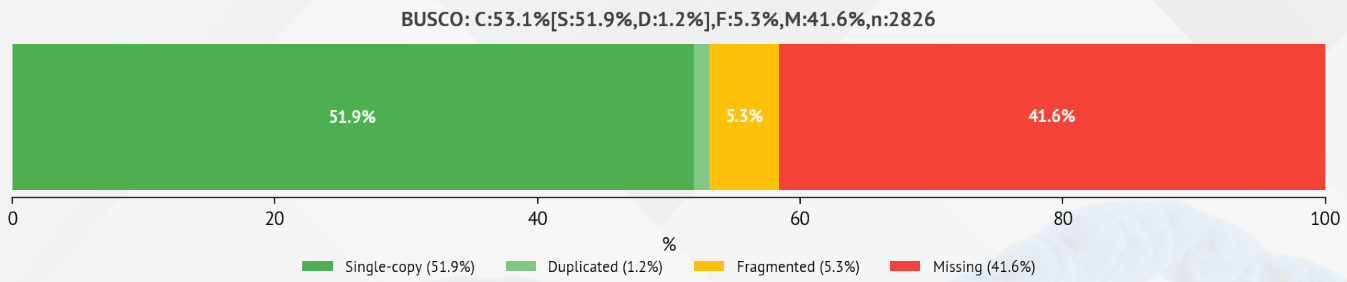
Completeness	100.0%
Contamination	0.0%

4.2 BUSCO Gene Space Completeness

C:53.1%[S:51.9%,D:1.2%],F:5.3%,M:41.6%,n:2826

Lineage: ascomycota_odb12

	Count	%
Complete BUSCOs	1500	53.1%
Single-copy	1466	51.9%
Duplicated	34	1.2%
Fragmented	149	5.3%
Missing	1177	41.6%
Total BUSCO Groups	2826	



5. TAXONOMIC IDENTIFICATION

Taxonomic identification was performed using sourmash, which compares MinHash signatures of the assembled genome against a reference database. The table below shows the classified lineage at each taxonomic rank with the fraction of the genome matched and the Average Nucleotide Identity (ANI) to the closest reference.

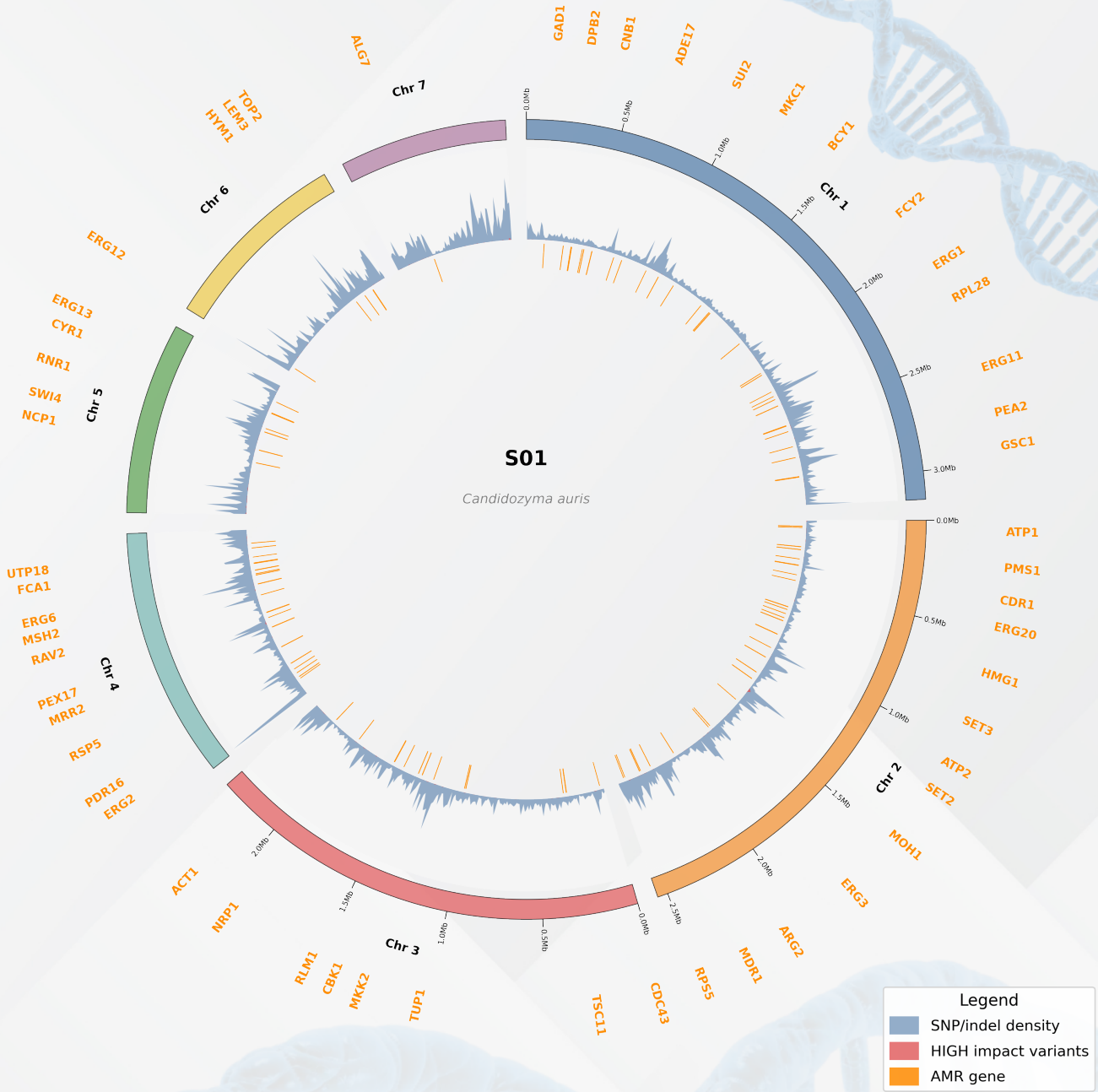
Rank	Lineage	Fraction	ANI
Superkingdom	Eukaryota	0.7633	99.47%
Phylum	Ascomycota	0.7633	99.47%
Class	Pichiomycetes	0.7633	99.47%
Order	Serinales	0.7633	99.47%
Family	Metschnikowiaceae	0.7633	99.47%
Genus	Candidozyma	0.7633	99.47%
Species	Candidozyma auris	0.7633	99.47%

5.1 Clade Classification

Clade classification for Candidozyma auris was performed using auriclass. C. auris is classified into six clades (I-VI), each with distinct geographic distributions and antifungal resistance profiles.

Clade	Clade I
QC Decision	PASS

6. GENOME OVERVIEW



7.1 Reference vs. Assembly Synteny

The dot plot below visualises the structural correspondence between the auto-selected reference genome (X-axis) and the polished de novo assembly (Y-axis), restricted to the ten longest contigs on each axis (sorted by length, descending). Each red segment is one **nucmer** alignment between a reference contig and an assembly contig. A clean diagonal indicates high collinearity between assembly and reference; off-diagonal blocks suggest rearrangements, and gaps indicate missing or extra sequence.

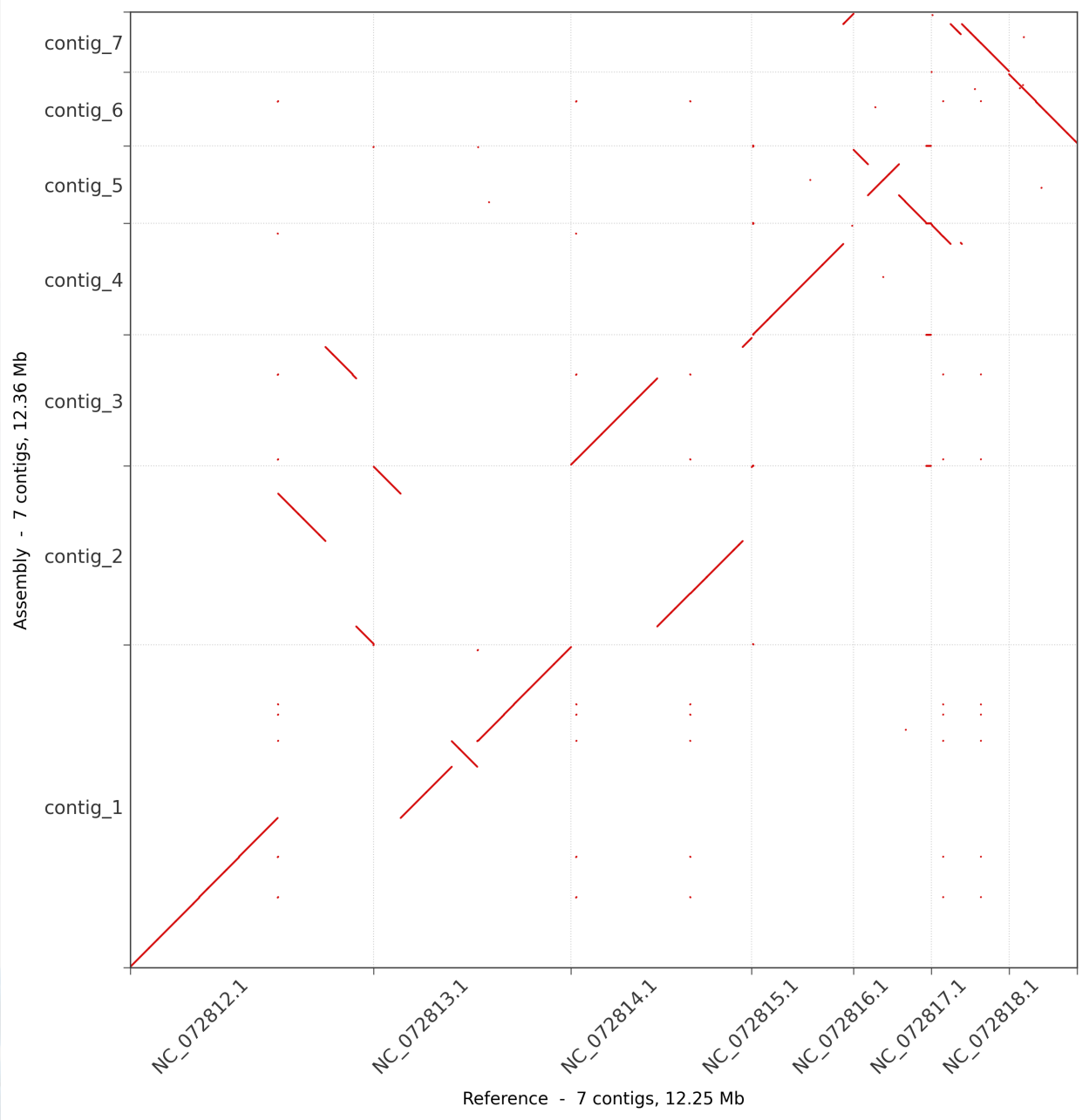


Figure 2. Reference vs. Assembly Synteny

8. ANTIFUNGAL RESISTANCE PROFILE

Protein-coding variants detected in 203 antifungal resistance genes were cross-referenced against two curated databases: **FungAMR** (35,792 entries from 501 studies, covering 208 drugs, 246 genes, 95 fungal species; Landry Lab, integrated into CARD) and **MARDy** (76,098 entries across 120 species, 216 genes, 230 antifungal compounds; University of Exeter). Mutations found in either database are shown below grouped by drug class.

The **Score** column shows the FungAMR confidence score: **positive values** indicate evidence of resistance, **negative values** indicate the mutation was found in susceptible strains. Higher absolute values (e.g., 8) indicate stronger evidence. **MIC** shows the reported Minimum Inhibitory Concentration.

8.1 Database-Validated AMR Mutations

Azoles

Gene	Mutation	Drug	Score	MIC (ug/mL)	FungAMR	MARDy
ERG11	Y132F	Fluconazole	8.0	>32	√	√
ERG11	Y132F	Itraconazole	8.0	16	√	√
ERG11	Y132F	Posaconazole	8.0	8	√	√
ERG11	Y132F	Voriconazole	8.0	8	√	√

Echinocandins

Gene	Mutation	Drug	Score	MIC (ug/mL)	FungAMR	MARDy
ERG11	Y132F	Anidulafungin	-8.0	0.12	√	√
ERG11	Y132F	Caspofungin	-8.0	0.25	√	√
ERG11	Y132F	Micafungin	-8.0	0.12	√	√

Polyenes

Gene	Mutation	Drug	Score	MIC (ug/mL)	FungAMR	MARDy
ERG11	Y132F	Amphotericin B	8.0	2	√	√

Pyrimidines

Gene	Mutation	Drug	Score	MIC (ug/mL)	FungAMR	MARDy
ERG11	Y132F	5-fluorocytosine	-8.0	0.12	√	√

8.2 ChroQueTas Phenotype Prediction

ChroQueTas (Chromosome Query Tool for antifungal susceptibility) is a tool that screens fungal genomes for known resistance-associated mutations by aligning the assembled genome against the **FungAMR** protein database using BLAST. Unlike our SnpEff-based variant calling which compares reads to a reference, ChroQueTas works directly on the polished assembly and compares translated protein sequences to curated resistance proteins.

The table below shows detected mutations with their associated drug predictions. The **MIC (sample/BP)** column shows the sample's MIC value and the confidence breakpoint score. A **negative breakpoint** indicates the mutation was reported in susceptible strains and is unlikely to confer resistance. A **positive breakpoint** indicates the mutation is associated with clinical resistance. **NA** indicates no MIC data was available for that drug-sample combination.

ChroQueTas Phenotype Prediction

Protein	Mutation	Drug Class	Drug	MIC (sample/BP)
ERG11	Y132F	Pyrimidines	5-fluorocytosine	NA/-8
ERG11	Y132F	Polyenes	Amphotericin B	8/-8
ERG11	Y132F	Echinocandins	Anidulafungin	NA/-8
ERG11	Y132F	Echinocandins	Caspofungin	8/-8
ERG11	Y132F	Azoles	Fluconazole	2/-8
ERG11	Y132F	Azoles	Itraconazole	8/-8
ERG11	Y132F	Echinocandins	Micafungin	NA/-8
ERG11	Y132F	Azoles	Posaconazole	8/-8
ERG11	Y132F	Azoles	Voriconazole	2/-8

Confidence Score Reference

1	Mutation created in same susceptible species background; resistance confirmed by measurements.
2	Mutation created in same gene but expressed in another species; resistance confirmed experimentally.
3	Effect quantified by bulk competition assays such as Deep Mutational Scanning.
4	Mutation identified by experimental evolution; high frequency in independent replicates suggests resistance.
5	Deletion of the gene causes resistance.
6	Overexpression or duplication of the gene causes resistance.
7	Significant association between mutation and resistance in a population (GWAS, QTL, or classical genetics).
8	Mutation identified in a resistant natural/clinical isolate without further validation.
NA	Absence of a positive confidence score report.
Negative (-)	Mutation was found NOT to cause resistance using the approach of the corresponding positive score.

9. VIRULENCE FACTORS

The table below lists genes in this assembly that match experimentally characterised virulence factors in the scientific literature. Each row corresponds to a high-identity homolog of a gene that has been functionally tested - typically via knockout, complementation, or overexpression studies - in this species (or a closely related one) and shown to influence the pathogen's ability to cause disease, evade host immunity, tolerate host conditions, or persist within tissues.

The **Phenotype** column records what happened to the source organism's virulence when the gene was perturbed in those studies - common outcomes are loss of pathogenicity, reduced virulence, increased virulence, and unaffected pathogenicity. These labels describe the experimental observation in the source species; they are not a direct prediction for this isolate. The hits are useful for narrowing in on candidate effectors, host-adaptation genes, and stress-response loci relevant to clinical behaviour, but they are starting points for follow-up rather than diagnostic markers.

Virulence Genes (PHI-base)

Gene	Source species	Phenotype	% id	Len	E-value	PHI-base
ERG25	Candida auris	reduced virulence	100.0	351	1.15e-233	PHI:125486
PDE2	Candida auris	reduced virulence	99.5	579	0.0	PHI:124327
GPA2	Candida auris	unaffected pathogenicity	99.1	451	1.22e-289	PHI:124324
RAS1	Candida auris	unaffected pathogenicity	98.8	243	2.66e-149	PHI:124323
PDE1	Candida auris	unaffected pathogenicity / reduced...	98.3	351	1.54e-225	PHI:124326

10. VARIANT SUMMARY

Variants were called using Clair3 against the auto-selected reference genome and annotated with SnpEff for functional effect prediction. The table below summarizes variants by impact category.

Variant Impact Summary

	Count
Total Variants	78294
HIGH Impact	394
MODERATE Impact	10550
LOW Impact	18657
MODIFIER	48693

10.1 AMR Resistance Gene Variants

Gene	Mutation	Effect	Impact
BUL1	I696fs	frameshift variant	HIGH
FLO8	S696fs	frameshift variant	HIGH
FLO8	E690fs	frameshift variant	HIGH
NRP1	T427fs	frameshift variant	HIGH
TUP1	Q157fs	frameshift variant	HIGH

REFERENCES & TOOLS

Version: 1.0.0

The following bioinformatics tools were used in this analysis workflow.

Tool	Version	Description
Fastplong	0.4.1	Long-read quality control
Flye	2.9.5	De novo genome assembler
minimap2	2.28	Sequence alignment
samtools	1.21	BAM/SAM processing
Medaka	2.0.1	Assembly polishing
BUSCO	6.0.0	Gene space completeness assessment
EukCC2	2.1	Eukaryotic genome completeness + contamination
sourmash	4.8.11	Taxonomic identification (MinHash)
Clair3	1.0.10	Variant calling (ONT)
SnpEff	5.4	Variant functional annotation
mosdepth	0.3.8	Read depth coverage profiling
Sniffles2	2.2	Structural variant calling
ChroQueTas	1.0	Antifungal resistance phenotype prediction
auriclass	0.7.0	C. auris clade classification
diamond	2.1	Protein homology search (virulence screen)

Databases

Database	Version	Description
FungAMR	2025-04-07	35,792 entries, 95 species, 246 genes
MARDy	2.0	76,098 entries, 120 species, 216 genes
sourmash	2025.01	NCBI eukaryotes k=51 signatures
BUSCO	eukaryota_odb12	Eukaryotic ortholog lineage dataset
EukCC2	1.2	Eukaryotic genome marker database
AuriClass	0.7.0	C. auris clade reference database
PHI-base	v4.19	10,023 pathogen-host interaction protein entries

MASSIVE
BIOINFORMATICS

Contact

massivebioinformatics.com
info@massivebioinformatics.com
+90 216 474 60 01

THANK YOU FOR CHOOSING US!

Website