Michael Hart
Bioinformatics
Final Project Report

Investigating a *C. auris* protein that could contribute to surface adhesion in clinical settings

C. auris has recently emerged as a lethal threat in clinical settings due to its multidrug resistance. This danger is compounded by strong adhesive properties that make it difficult to remove from medical equipment (Kean et al. 2018). The proteins that give this adhesiveness have not yet been fully investigated, limiting the range of possible countermeasures (Muñoz et al. 2018; Kean et al. 2018).

My research aim was to investigate the uncharacterized *C. auris* protein CJI97_000055 to determine if it is an adhesin that contributes to *C. auris*'s strong adhesive properties. An InterProScan search revealed that CJI97_000055 contains a large segment belonging to the glycophosphatidylinositol (GPI)-anchored superfamily, which has been implicated as a unique contributor to C. auris's tenacious stickiness to medical devices (Kean et al. 2018; Muñoz et al. 2018). The GPI anchor is a modification added to the C-terminus of proteins after translation that allows the modified protein to be anchored in the cell membrane's outer leaflet, in some cases contributing to cell-cell contact and adhesion (Paulick and Bertozzi, 2008). Several of the analyses I undertook will be detailed in this paper, including a multiple sequence alignment of homologs, phylogenetic trees, 3D protein structure prediction, and motif discovery.

I found a good amount of evidence supporting the hypothesis that the protein is a contributor to *C. auris* adhesion. According to an InterPro search of CJI97_000055, the protein contains one domain belonging to the GPI-anchored superfamily (implicated in surface adhesion) and another domain belonging to the KRE9 superfamily (involved in cell wall synthesis). My multiple sequence alignment shows relatively high conservation in both of these domains. Phylogenetic trees built from the multiple sequence alignment suggest that five different strains of *C. auris* carry identical homologs of CJI97_000055, suggesting the importance of this protein for *C. auris*. Several other analyses add more weight to the theory that CJI97_000055 is an important adhesin, but definite conclusions are hampered by a lack of structural characterization of the protein and limited experimental evidence from gene knockout experiments.

Gene Structure and Function:

Multiple Sequence Alignment

Running BLAST searches of CJI97_000055's protein sequence revealed the homologs listed in Table 1, which I put into a multiple sequence alignment using MUSCLE in the program MEGA-X. Figure 1 is the multiple sequence alignment built in Jalview with the Zappo color configuration (which colors by amino acid properties) and secondary structures predicted by JPred. Signal peptide and superfamily annotation was added based on the InterProScan results for CJI97_000055.

Species_Strain	Protein Name	RefSeq Accession Numbe	Sequence Length (AA)	% Coverage	% Identity
Candida auris_B11221	CJI97_000055	XP_028892459	272	(reference)	(reference)
Candida_auris_B11221	CJI97_000056	XM_029032145.1	266	92%	50%
Candida_auris_B11220	CJI96_0001486	QEO20671.1	272	100%	100%
Candida_auris_B8441	B9J08_000049	PIS58603.1	272	100%	100%
Candida_auris_JCM_15448	CAJCM15448_1085	GBL48811.1	272	100%	100%
Candida_auris_6684	QG37_05757	KND97379.2	272	100%	100%
Candida_duobushaemulonis_B09383	CXQ87_002839	XP_025335632.1	270	100%	72.06%
Candida_duobushaemulonis_B09383	CXQ87_002838	XP_025335631.1	269	99%	45.93%
Candida_pseudohaemulonis_B12108	C7M61_001738	XP_024715017.1	270	100%	70.96%
Candida_pseudohaemulonis_B12108	C7M61_001737	XP_024715016.1	269	92%	44.44%
Candida_haemulonis_B11899	CXQ85_000062	XP_025342037.1	272	100%	70.22%
Candida_haemulonis_B11899	CXQ85_000063	XP_025342038.1	253	93%	48.03%
Candida lusitaniae	ATCC 42720	XP_002618139.1	267	98%	55.76%
Lodderomyces_elongisporus	NRRL YB-4239	XM_001526065.1	268	97%	51.67%
Candida_tropicalis	MYA-3404	XP_002548030.1	271	98%	49.45%
Candida_orthopsilosis	Co 90-125	XP_003868628.2	196	66%	49.18%
Candida_guilliermondii	PGUG_01687	XP_001486016.1	274	100%	48.58%
Candida_albicans	SC5314	XP_723077.1	271	99%	46.72%
Candida parapsilosis	CPAR2_404060	NA	269	99%	45.99%
Candida_dubliniensis	CD36	XP_002419330.1	271	99%	45.99%
Debaryomyces_hansenii_CBS767	CBS767	XP_457401.2	270	100%	45.85%
Debaryomyces_hansenii_CBS767	CBS767	XP_457402.2	270	100%	44.29%
Candida krusei	C5L36_0D02170	XP_029322951.1	277	95%	40.66%
Saccharomyces_cerevisiae_S288C	S288C	NP_012361.1	276	96%	38.49%
Saccharomyces_cerevisiae_S288C	S288C	NP_010234.1	268	90%	35.32%
Candida_glabrata_CBS_138	CAGL0C00363g	XP_445197.1	276	91%	37.22%
Candida_glabrata_CBS_138	CAGL0H07997g	XP_447146.1	265	91%	37.22%

Table 1

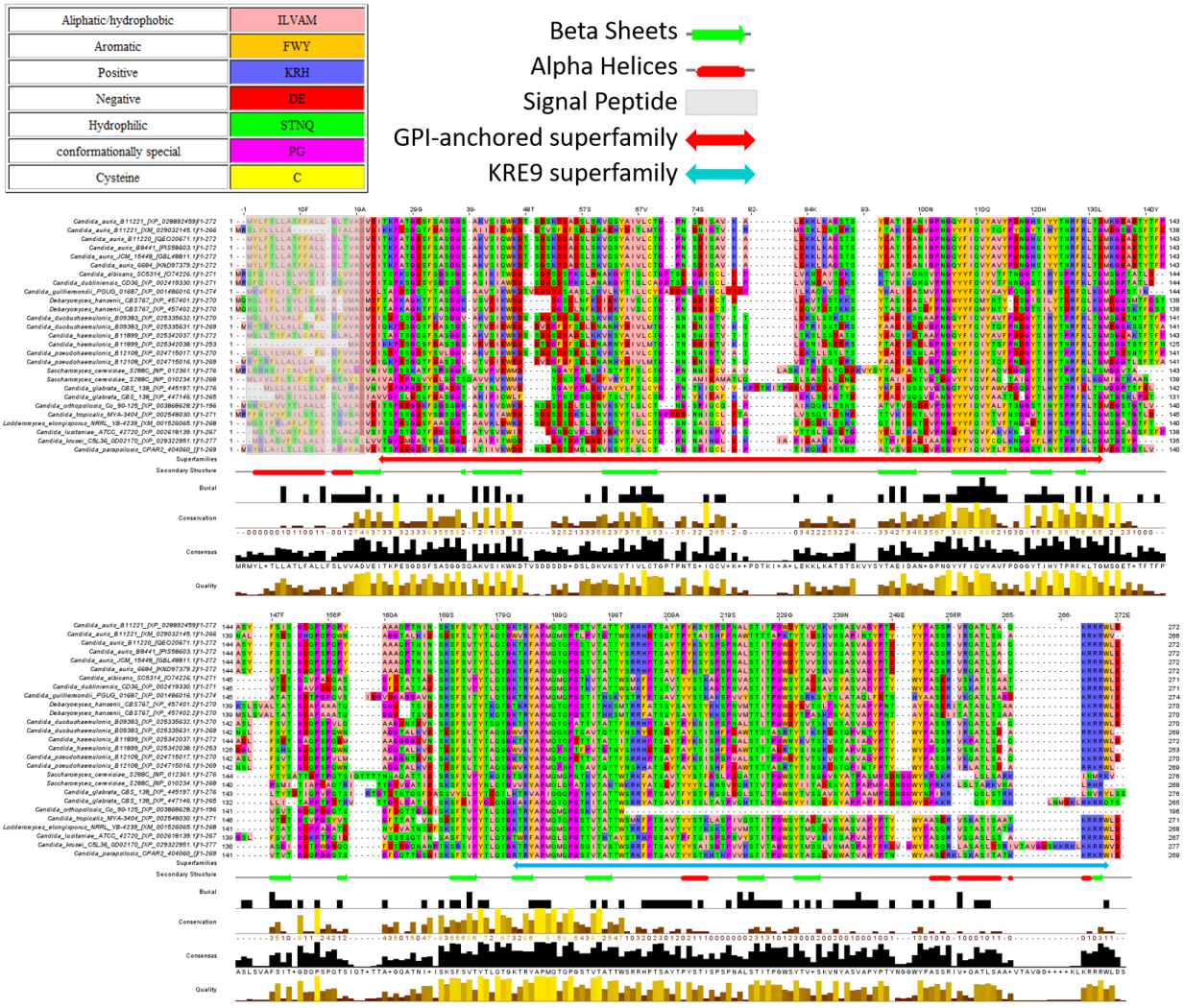

Figure 1: Multiple Sequence Alignment

The multiple sequence alignment shows good conservation for most of the non-gap regions of the GPI-anchored domain, with some amino acids retaining 100% conservation for many residues (ex. the proline at 26 bp, the tryptophan at 45 bp, and the leucine at 73 bp). For the KRE9 domain the conservation quickly drops halfway through. The signal peptide region at the beginning is mostly hydrophobic and the KRE9 domain is largely hydrophilic (which suggests that this region is not entirely buried in the cell wall), but the GPI-anchored domain contains a broad mix of amino acid properties. There are about 12 predicted beta sheets and about four predicted alpha helices. The GPI-anchored domain is mostly beta sheets, and the KRE9 domain is a mix of beta sheets and alpha helices.

Gene Structure and Function:

Protein Structure Prediction

CJI97_000055 is an uncharacterized protein and did not return any good template matches from BLAST Protein Database or Swiss Model searches, so I used the protein threading method of the program I-TASSER to generate predicted models of the protein's structure.

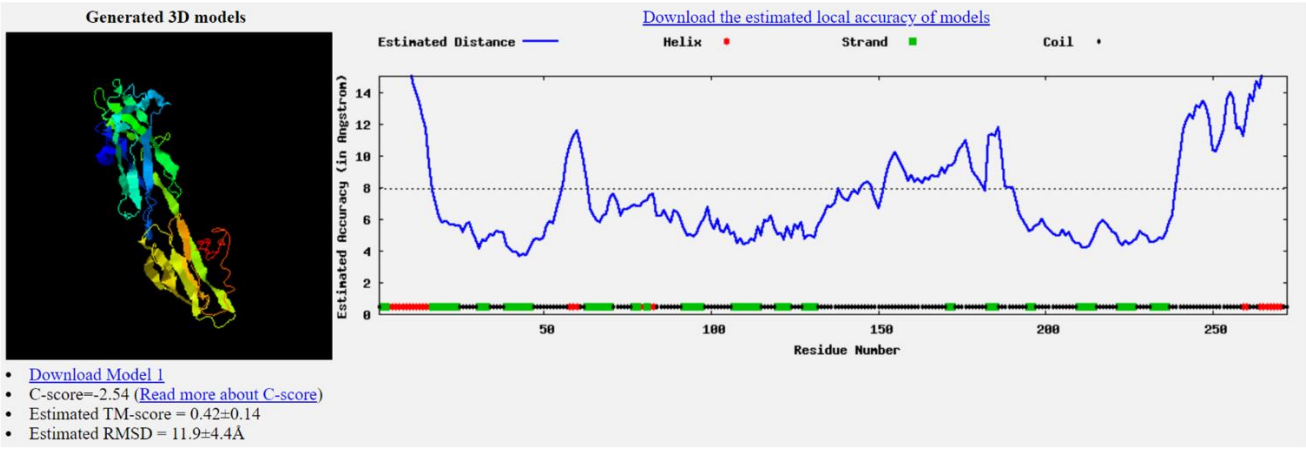

Figure 2: First of Top 5 final models predicted by I-TASSER
C-score > -1.5 indicates a model of correct global topology; these C-scores are below -1.5, so the models are low-confidence

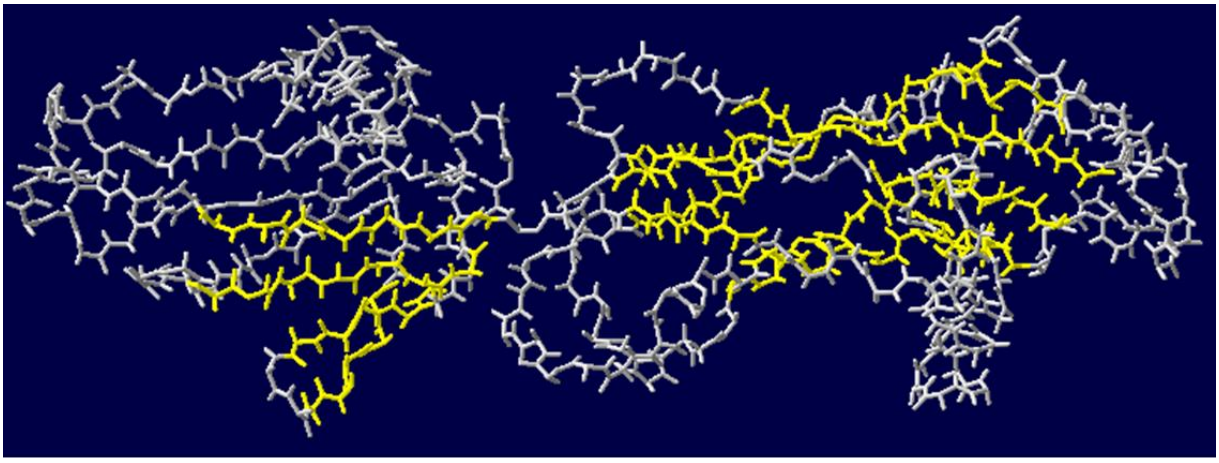

Figure 3: PDB file of predicted structure for CJI97_000055 in DeepView
Yellow = beta sheets

Figure 2 is the first of the "Top 5" final models predicted by I-TASSER. The chart to the right of the protein shows the estimated accuracy (in angstroms) of each region of the secondary structure, revealing higher predicted precision for most of the beta sheets but lower predicted precision for the alpha helices. Figure 3 is the image produced by SwissProt DeepView when given the PDB file of the first of the "Top 5" final models predicted by I-TASSER. I colored the secondary structures, revealing eight beta sheets (yellow) clustered in fours on each side of the

protein. I was initially surprised that the predicted alpha helices at the beginning and end of the protein were not depicted in the PDB DeepView file; this may be because those regions have lower conservation and confidence in their structure prediction.

I-TASSER also provides a report of gene ontology terms predicted to be associated with CJI97_000055. The gene ontology terms predicted from the top three templates were “cell wall”, “binding”, “pathogenesis”, and “cholesterol binding”. The gene ontology terms associated with the consensus prediction of all of the templates included “sterol binding”, “multi-organism process”, and ”cell wall”. Some of these gene ontology terms (“binding”, “cell wall”) make sense for CJI97_000055 because it is known to contain a GPI-anchored domain involved in adhesion and the KRE9 domain involved in cell wall synthesis, but it is difficult to determine how the other terms may or may not relate to surface adhesion in *C. auris*.

Regarding all of the results from I-TASSER, the confidence-related scores (Z-score, C-score, etc.) were all relatively poor, and either barely met a confidence threshold or fell below it. For example, for Figure 2 the C-Score is measured between 0-1 and my C-scores are all 0.1 or lower, so the results are extremely low-confidence. This indicates that the protein CJI97_000055 is a hard target to model and that the I-TASSER results should be cited with caution.

Gene & Genome Evolution:

Phylogenetic Analysis

I undertook a phylogenetic analysis to gather evidence regarding when and in what species CJI97_000055 expanded compared to closely related Candida species. Figure 4 and Figure 5 are maximum parsimony and maximum likelihood trees, respectively. They were generated from a multiple sequence alignment of CJI97_000055 homologs in the MEGA-X program. Both my maximum parsimony tree and maximum likelihood trees match fairly well with Fig. 3 of Muñoz et al. 2018 (which is a maximum likelihood tree) when rooted on the branch containing *C. krusei*, *C. glabrata*, and *S. cerevisiae*, which is assumed to be the outgroup. One striking feature of both trees is that protein sequences from five strains of *C. auris* (B11221, B11220, B8441, JCM_15448, 6884) are grouped in one clade with a bootstrap value of 100; referencing the multiple sequence alignment in Figure 1 also reveals that there are zero amino

acid substitutions in these five sequences. This suggests that the protein is critical for functionality in the *C. auris* life cycle and cannot tolerate amino acid substitutions, thus preventing any fixation of new mutations.

My phylogenetic trees both suggest that the CJI97_000055 protein underwent duplication events followed by speciation events in strains of the species *C. auris*, *C. duobushaemulonis*, *C. haemulonis*, *S. cerevisiae*, and *C. glabrata*. The most noticeable difference between the trees is that the maximum likelihood tree has S. cerevisiae and C. glabrata display a duplication followed by two speciation events, whereas the maximum parsimony tree shows a speciation followed by one duplication. The maximum likelihood tree has higher bootstrap values for these branchings and requires fewer evolutionary events, so according to the parsimony principle the maximum likelihood tree is in fact more likely to be correct.

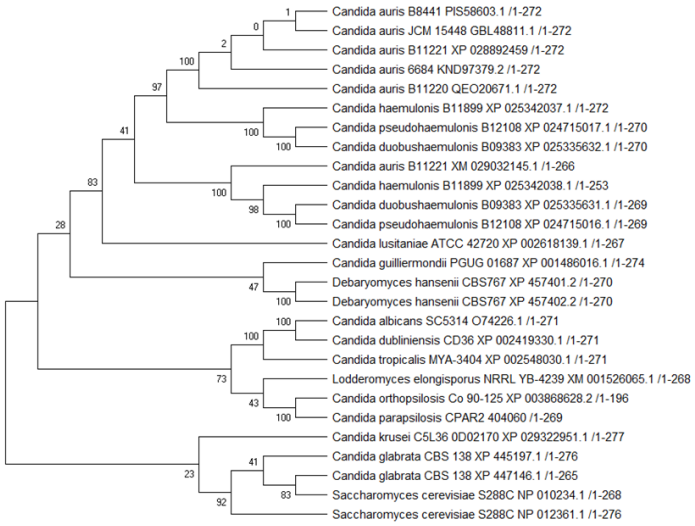

FIGURE 4: Maximum Parsimony Tree

Maximum Parsimony analysis of taxa

The evolutionary history was inferred using the Maximum Parsimony method. Tree #1 out of 5 most parsimonious trees (length = 1540) is shown. The consistency index is (0.636856), the retention index is (0.693188), and the composite index is 0.451923 (0.441461) for all sites and parsimony-informative sites (in parentheses). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches [1]. The MP tree was obtained using the Subtree-Pruning-Regrafting (SPR) algorithm (pg. 126 in ref. [2]) with search level 1 in which the initial trees were obtained by the random addition of sequences (10 replicates). This analysis involved 27 amino acid sequences. There were a total of 323 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [3].

1. Felsenstein J. (**1985**). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-791.

2. Nei M. and Kumar S. (**2000**). *Molecular Evolution and Phylogenetics.* Oxford University Press, New York.

3. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* **35**:1547-1549.

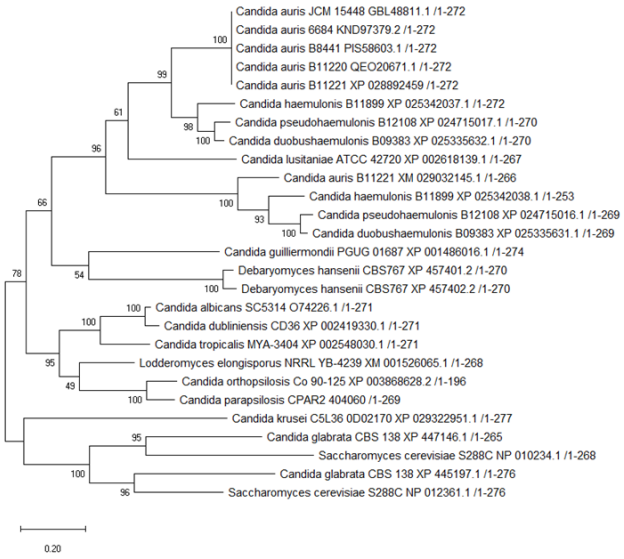

FIGURE 5: Maximum Likelihood Tree

Evolutionary analysis by Maximum Likelihood method
The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model [1]. The tree with the highest log likelihood (-8507.65) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 27 amino acid sequences. There were a total of 323 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [2].

1. Jones D.T., Taylor W.R., and Thornton J.M. (**1992**). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**: 275-282.

2. Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (**2018**). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution* **35**:1547-1549.

Global Analyses & Models:

Motif Discovery

Figure 6 shows the motif discovery output from the MEME program. When searching for a total of 100 motifs between 6 and 50 amino acids long, MEME returned many motifs that are conserved in every sequence from the multiple sequence alignment. The most highly conserved motifs appear in the ~20 to ~130 range and ~180 to ~280 range, which overlaps with the predicted location of the GPI-anchored superfamily domain and KRE9 superfamily domain. Serine and threonin (hydrophobic) appear the most in the MSA, with frequences of 0.119 and 0.125, respectively. The largest motif (1) appears in the KRE9 domain.

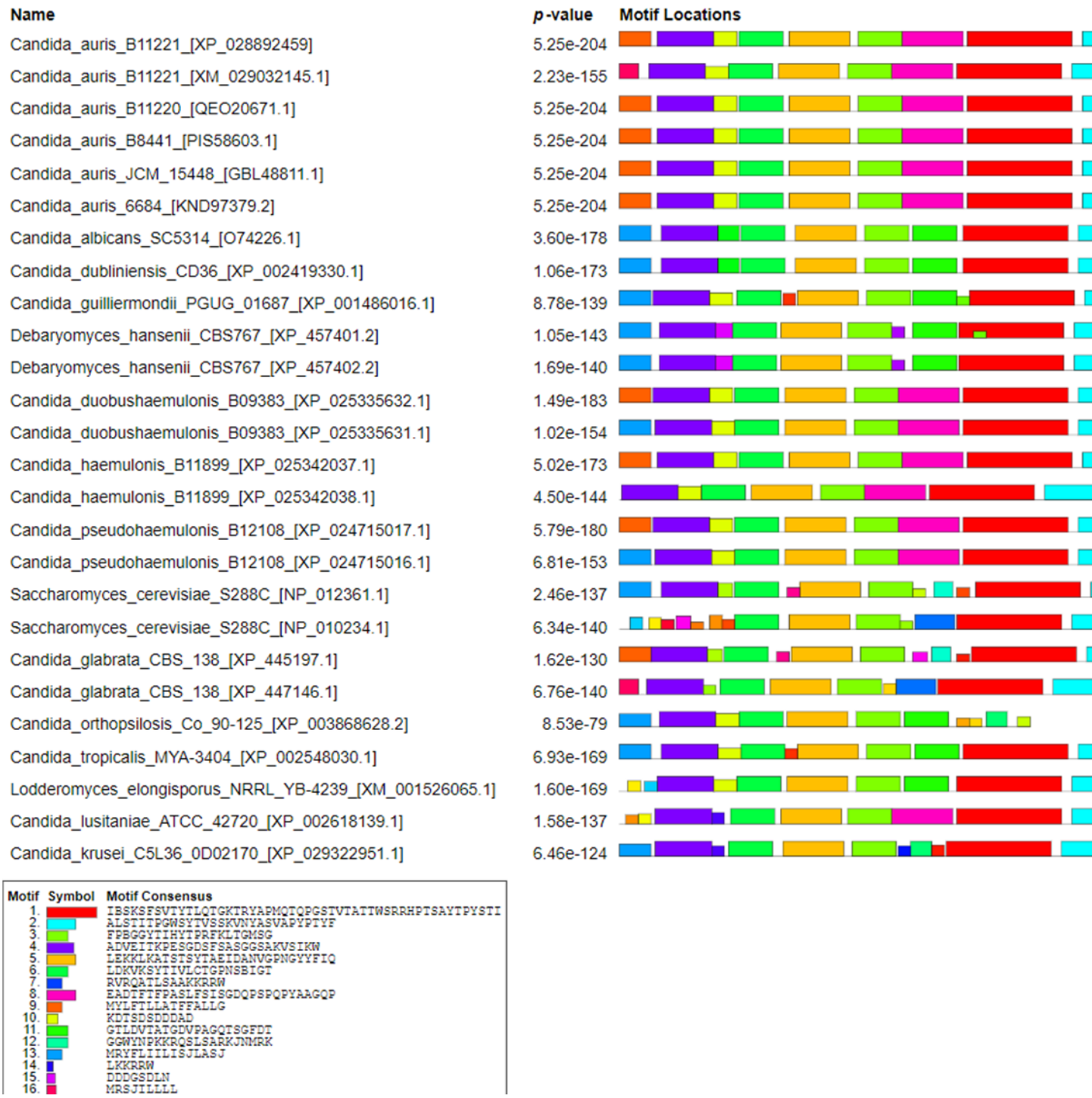
Name
p-value
Motif Locations
Candida_auris_B11221_[XP_028892459] 5.25e-204
Candida_auris_B11221_[XM_029032145.1] 2.23e-155
Candida_auris_B11220_[QEO20671.1] 5.25e-204
Candida_auris_B8441_[PIS58603.1] 5.25e-204
Candida_auris_JCM_15448_[GBL48811.1] 5.25e-204
Candida_auris_6684_[KND97379.2] 5.25e-204
Candida_albicans_SC5314_[O74226.1] 3.60e-178
Candida_dubliniensis_CD36_[XP_002419330.1] 1.06e-173
Candida_guilliermondii_PGUG_01687_[XP_001486016.1] 8.78e-139
Debaryomyces_hansenii_CBS767_[XP_457401.2] 1.05e-143
Debaryomyces_hansenii_CBS767_[XP_457402.2] 1.69e-140
Candida_duobushaemulonis_B09383_[XP_025335632.1] 1.49e-183
Candida_duobushaemulonis_B09383_[XP_025335631.1] 1.02e-154
Candida_haemulonis_B11899_[XP_025342037.1] 5.02e-173
Candida_haemulonis_B11899_[XP_025342038.1] 4.50e-144
Candida_pseudohaemulonis_B12108_[XP_024715017.1] 5.79e-180
Candida_pseudohaemulonis_B12108_[XP_024715016.1] 6.81e-153
Saccharomyces_cerevisiae_S288C_[NP_012361.1] 2.46e-137
Saccharomyces_cerevisiae_S288C_[NP_010234.1] 6.34e-140
Candida_glabrata_CBS_138_[XP_445197.1] 1.62e-130
Candida_glabrata_CBS_138_[XP_447146.1] 6.76e-140
Candida_orthopsilosis_Co_90-125_[XP_003868628.2] 8.53e-79
Candida_tropicalis_MYA-3404_[XP_002548030.1] 6.93e-169
Lodderomyces_elongisporus_NRRL_YB-4239_[XM_001526065.1] 1.60e-169
Candida_lusitaniae_ATCC_42720_[XP_002618139.1] 1.58e-137
Candida_krusei_C5L36_0D02170_[XP_029322951.1] 6.46e-124
Motif Symbol Motif Consensus
1. IBSKSFSVTYTLQTGKTRYAPMQTQPGSTVTATTWSRRHPTSAYTPYSTI
2. ALSTITPGWSYTVSSKVNYASVAPYPTYF
3. FPBGGYTIHYTPRFKLTGMSG
4. ADVEITKPESGDSFSASGGSAKVSIKW
5. LEKKLKATSTSYTAEIDANVGPNGYYFIQ
6. LDKVKSYTIVLCTGPNSBIGI
7. RVRQATLSAAKKRRW
8. EADTFTFPASLFSISGDQPSPQPYAAGQP
9. MYLFTLLATFFALLG
10. KDTSDSDDDAD
11. GTLDVTATGDVPAGQTSGFDT
12. GGWYNPKKRQSLSARKJNMRK
13. MRYFLIILISJLASJ
14. LKKRRW
15. DDDGSDLN
16. MRSJILLLL

Figure 6

Global Analyses & Models:

Cartoon Summary Model

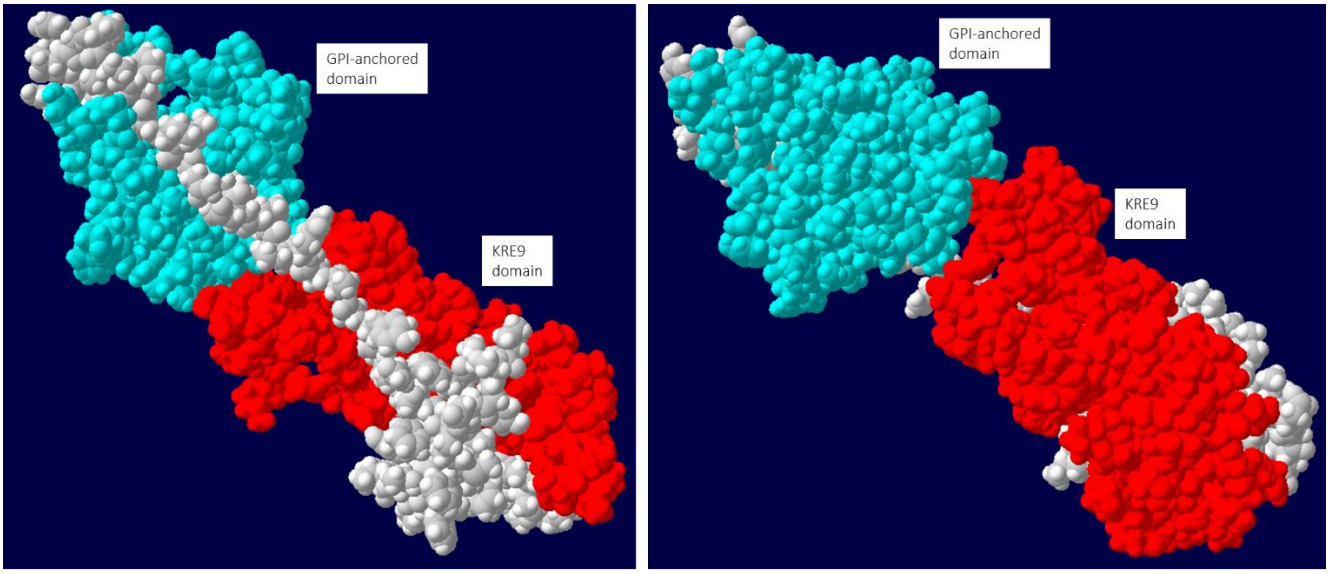

Figure 7: Front and Back Views of I-TASSER predicted protein model of *C. auris* protein CJI97_000055
Red=GPI-anchored superfamily domain; Teal=KRE9 superfamily domain

Experiments with *C. auris* have shown that adhesin-related GPI-anchored cell wall genes experience upregulation throughout the formation of biofilm, suggesting that GPI-anchored domains play an important role within cellular adhesion (Kean et al. 2018). Figure 7 is an I-TASSER-predicted PDB file in DeepView with the GPI-anchored domain colored teal and the KRE9 domain colored red, shown from two different angles. Although my protein is not confirmed to be an adhesin, the cartoon summary model in Figure 8 (adapted from an original figure presented in Kean et al.) indicates that CJI97_000055 upregulation should be seen at the intermediate (12 hour) timepoint during *C. auris* adhesion.

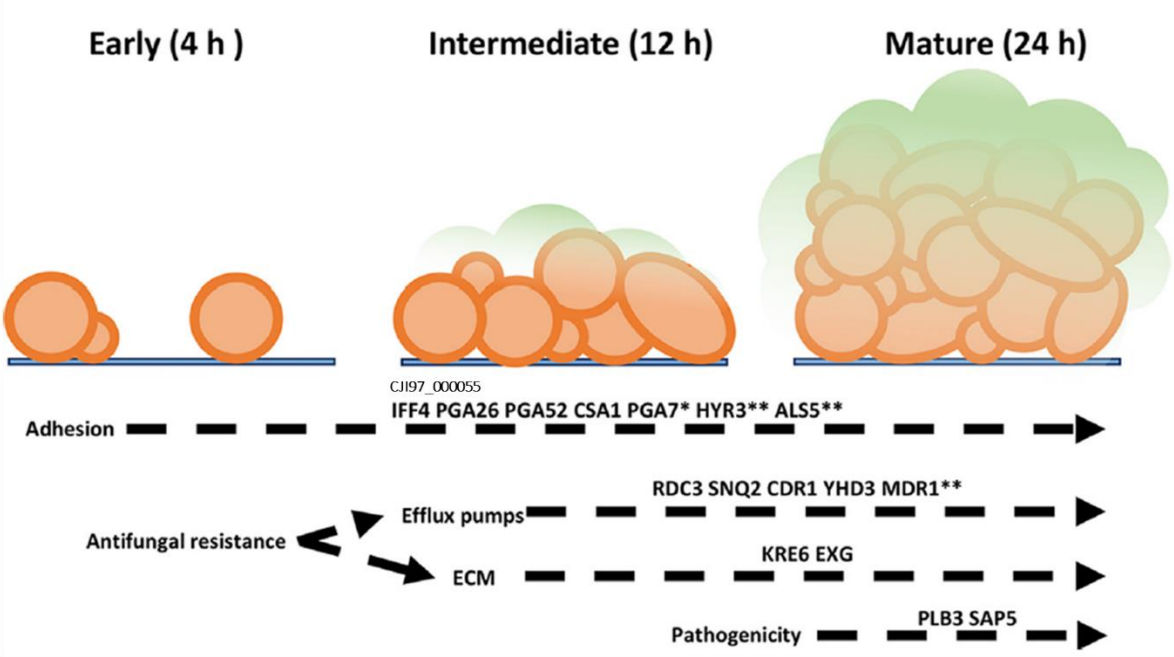

Figure 8: Possible Path of CJI97_000055 Upregulation for Biofilm Formation (biofilm formation image from Kean et al.)

Discussion

CJI97_000055 probably contributes to *C. auris*'s strong adhesive properties. The strongest evidence to support this is the fact that multiple databases and algorithms recognized CJI97_000055 as containing a GPI-anchored domain, and proteins that have this domain have been shown to be upregulated during *C. auris* biofilm formation (Kean et al. 2018). Additionally, FungalRV's "Adhesin Predictor" program scored the CJI97_000055 amino acid sequence as 1.1209, which is above the threshold for being classified as an adhesin. The I-TASSER results are less convincing because although the CJI97_000055 protein model and its predicted templates sequences were associated with gene ontology terms such as "cell wall" and "binding" that could be related to biofilm adhesion, the relatively low confidence-related scores associated with the I-TASSER models indicate that these results should be cited with caution and highlight the need for actual structural characterization of the CJI97_000055 protein to confirm the results.

There are several wet lab experiments that could be performed to determine if CJI97_000055 is actually a *C. auris* adhesin. It has been observed that in *C. auris* some GPI-linked cell wall proteins are upregulated in early biofilm formation which highlights their importance in the first stage of adherence (Kean et al. 2018). If CJI97_000055 is indeed an adhesin, it should

also be upregulated during biofilm formation as demonstrated in figure 8. This could be tested by a replication of the Kean et al. experimental procedure that also measures levels CJI97_000055 translation. Additionally, simple gene knockout experiments could be performed on CJI97_000055 to see if the absence of the protein leads to less effective biofilm formation and surface adhesion. Since there are many proteins involved in GPI-anchored adhesion, knocking out different amounts of the proteins simultaneously could help determine which proteins play the most critical role in surface adhesion.

References

Kean R., Delaney C., Sherry L., Borman A., Johnson EM., Richardson MD., Rautemaa-Richardson R., Williams C., Ramage G. Transcriptome assembly and profiling of Candida auris reveals novel insights into biofilm-mediated resistance. mSphere 3:e00334-18. (2018).

Muñoz J.F., Gade L., Chow N.A. et al. Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species. Nat Commun 9, 5346 (2018) doi:10.1038/s41467-018-07779-6.

Paulick M. G., Bertozzi C. R. The glycosylphosphatidylinositol anchor: a complex membrane-anchoring structure for proteins. Biochemistry. 2008;47(27):6991–7000. doi: 10.1021/bi8006324.

Bailey T. and Elkan C. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers". Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California. (1994).