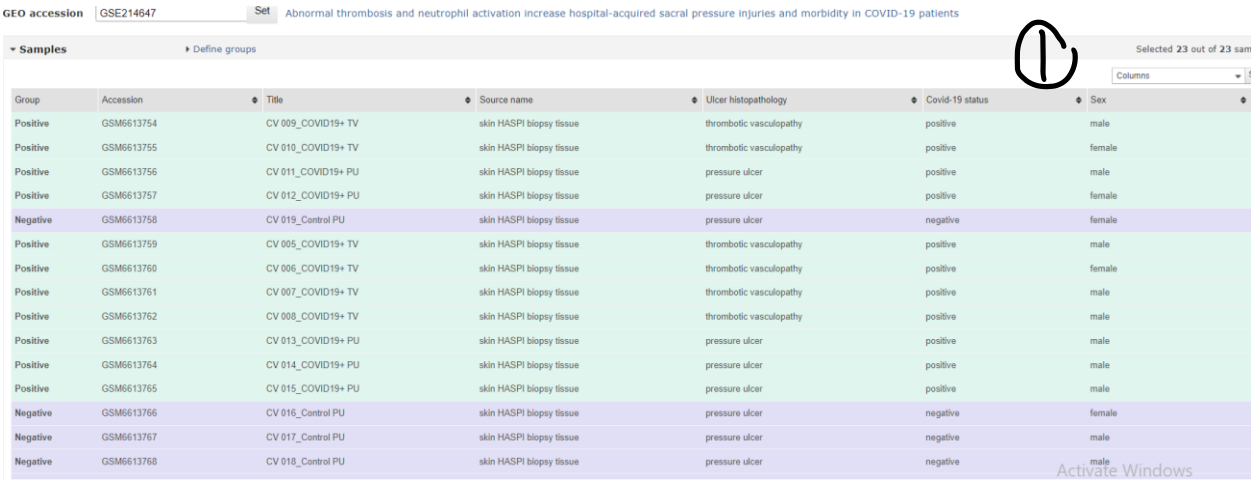
Question I: DGEA

The study is concerned with studying sacral pressure injuries in COVID-19-positive patients and COVID-19-negative patients to investigate how the viral infection affects sacral pressure injuries. Therefore, the main question of this workflow is to determine the differentially expressed genes across the two cohorts, with COVID-19 vs. without COVID-19.

Data Exploration (Figure below): There are 23 patients, 13 with covid and 10 without. The metadata table has 7 columns that describe the patients. I will not need all of them. Actually, I will only need the covid-19 status to define the two groups (Figure 1 below).

Groups (Cohorts): I am studying sacral pressure injuries in **COVID-19-positive vs. COVID-19-negative patients**. Therefore, the two cohorts I am studying can be chosen based on column number 7, which indicates **COVID-19 status** (Figure 1 below).



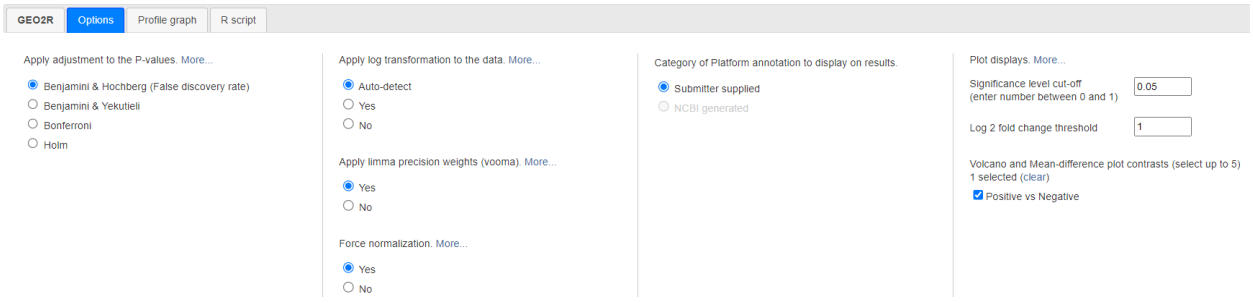
GEO accession GSE214647 Set Abnormal thrombosis and neutrophil activation increase hospital-acquired sacral pressure injuries and morbidity in COVID-19 patients

Samples Define groups Selected 23 out of 23 sam Columns

Group	Accession	Title	Source name	Ulcer histopathology	Covid-19 status	Sex
Positive	GSM6613754	CV 009_COVID19+ TV	skin HASPI biopsy tissue	thrombotic vasculopathy	positive	male
Positive	GSM6613755	CV 010_COVID19+ TV	skin HASPI biopsy tissue	thrombotic vasculopathy	positive	female
Positive	GSM6613756	CV 011_COVID19+ PU	skin HASPI biopsy tissue	pressure ulcer	positive	male
Positive	GSM6613757	CV 012_COVID19+ PU	skin HASPI biopsy tissue	pressure ulcer	positive	female
Negative	GSM6613758	CV 019_Control PU	skin HASPI biopsy tissue	pressure ulcer	negative	female
Positive	GSM6613759	CV 005_COVID19+ TV	skin HASPI biopsy tissue	thrombotic vasculopathy	positive	male
Positive	GSM6613760	CV 006_COVID19+ TV	skin HASPI biopsy tissue	thrombotic vasculopathy	positive	female
Positive	GSM6613761	CV 007_COVID19+ TV	skin HASPI biopsy tissue	thrombotic vasculopathy	positive	male
Positive	GSM6613762	CV 008_COVID19+ TV	skin HASPI biopsy tissue	thrombotic vasculopathy	positive	male
Positive	GSM6613763	CV 013_COVID19+ PU	skin HASPI biopsy tissue	pressure ulcer	positive	male
Positive	GSM6613764	CV 014_COVID19+ PU	skin HASPI biopsy tissue	pressure ulcer	positive	male
Positive	GSM6613765	CV 015_COVID19+ PU	skin HASPI biopsy tissue	pressure ulcer	positive	male
Negative	GSM6613766	CV 016_Control PU	skin HASPI biopsy tissue	pressure ulcer	negative	female
Negative	GSM6613767	CV 017_Control PU	skin HASPI biopsy tissue	pressure ulcer	negative	male
Negative	GSM6613768	CV 018_Control PU	skin HASPI biopsy tissue	pressure ulcer	negative	male

Activate Windows

Performing DGEA: After adjusting the parameters as follows (Figure 2), *Benjamini & Hochberg, Auto-detect log transformation, Apply Limma precision weights, Force normalization, Significance level of 0.05, Log2 fold change of 1, and checking the contrast for the two groups*, **69 differentially expressed genes were concluded**, of which **43 are under-expressed and 26 are over-expressed**. The reference is COVID-19-negative cohort. That is, when I say under-expressed, I mean it is under-expressed in the COVID-19-positive group in comparison to the reference COVID-19-negative group, and vice versa for over-expressed.



GEO2R Options Profile graph R script

Apply adjustment to the P-values. More...
- (•) Benjamini & Hochberg (False discovery rate)
- ( ) Benjamini & Yekutieli
- ( ) Bonferroni
- ( ) Holm

Apply log transformation to the data. More...
- (•) Auto-detect
- ( ) Yes
- ( ) No

Apply limma precision weights (vooma). More...
- (•) Yes
- ( ) No

Force normalization. More...
- (•) Yes
- ( ) No

Category of Platform annotation to display on results.
- (•) Submitter supplied
- ( ) NCBI generated

Plot displays. More...

Significance level cut-off (enter number between 0 and 1) 0.05

Log 2 fold change threshold 1

Volcano and Mean-difference plot contrasts (select up to 5) 1 selected (clear)
- [x] Positive vs Negative

Top 100 statistically differentially expressed genes: Since the question asks for **statistical significance**, I should sort according to the **p-adjusted** and locate the first 100 occurrences. The genes are illustrated in the figures below and the Excel sheet is also attached to the assignment.

Statistical Significance: Adjusted p-value cutoff of 0.05

Biological Significance: A log-fold change cutoff of 1

Visualization: Below are volcano plot, boxplots, and UMAP, all drawn by GEO2R.

Interpretation:

Volcano Plot: This gives a view on the genes that are both statistically and biologically differentially expressed since it plots **log2 fold change vs -log10 P-value**. To locate the truly differentially expressed genes, we should be **looking for the red (over-expressed) and blue balls (under-expressed), not the black**. From the figure only and visually, it seems that there are more downregulated genes than upregulated ones, and this is confirmed numerically below.

UMAP: Uniform Manifold Approximation and Projection is a **dimensionality reduction technique** useful for visualizing **how samples are related to each other**. It seems that the **COVID-19-negative patients are a little bit more clustered** and **similar** than the COVID-19-positive patients.

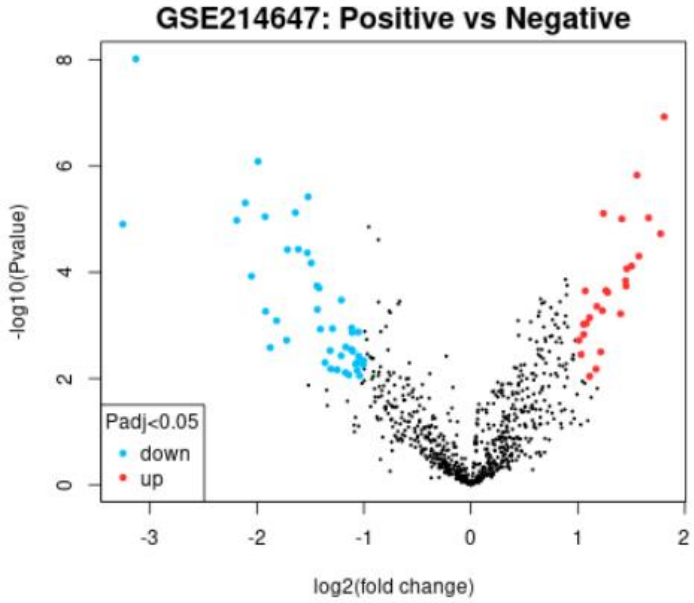
Boxplot: This shows the **distribution of values** (the word ‘values’ is not explained on the axis or on GEO2R official website) **per sample**. It is useful to investigate whether the samples are suitable for DGEA or not and if there are any outliers. From the figure, the data is **homogenous**, **no outliers**, and the **samples can be used for DGEA.** Before clicking ‘force normalization’, they were not homogenous or had similar distributions.

	A	B	C	D	E	F	G
1	ID	adj.P.Val	P.Value	t	B	logFC	G
2	CXCL13	7.67E-06	9.53E-09	-8.53713	10.21277	-3.13259	N
3	AIRE	4.71E-05	1.17E-07	7.40725	7.77781	1.80738	N
4	TMEM176A	0.00021825	8.14E-07	-6.58255	5.91835	-1.98985	N
5	PADI2	0.00029574	1.47E-06	6.33751	5.31887	1.55297	N
6	CIITA	0.00060905	3.79E-06	-5.95141	4.44194	-1.52183	N
7	HLA-DPB1	0.00066096	4.93E-06	-5.84466	4.17422	-2.10861	N
8	VCAM1	0.0007022	7.51E-06	-5.67585	3.80082	-1.64019	N
9	IL27	0.0007022	7.78E-06	5.66161	3.68402	1.23934	N
10	HLA-DRB3/4	0.0007022	8.95E-06	-5.60542	3.57908	-1.92335	N
11	PF4V1	0.0007022	9.4E-06	5.58561	3.5024	1.66156	N
12	HLA-DPA1	0.0007022	1.05E-05	-5.5423	3.47863	-2.18823	N
13	CAMP	0.0007022	9.82E-06	5.56822	3.46384	1.41124	N
14	HLA-DRA	0.0007666	1.24E-05	-5.4754	3.27362	-3.25447	N
15	ITGAL	0.00079731	1.39E-05	-5.43026	3.16764	-0.95461	N
16	TLR9	0.00100287	1.87E-05	5.31176	2.83453	1.77379	N
17	PTPN22	0.00122316	2.43E-05	-5.20757	2.6081	-0.86492	N
18	IL7R	0.00166948	3.69E-05	-5.04354	2.28306	-1.61318	N
19	HLA-DQB2	0.00166948	3.74E-05	-5.03825	2.26513	-1.71462	N
20	CPA3	0.00181299	4.28E-05	-4.98446	2.11473	-1.52903	N
21	IL1A	0.00199224	4.96E-05	4.92716	1.91849	1.57143	N
22	CD4	0.00255675	6.68E-05	-4.80988	1.72228	-1.49316	N
23	TG	0.00278541	7.62E-05	4.75796	1.49478	1.50252	N
24	XCL1/2	0.00299193	8.56E-05	4.71242	1.35533	1.45671	N
25	CD74	0.00396683	0.000118	-4.58502	1.09811	-2.05147	N
26	S1PR1	0.00435628	0.000135	4.53225	0.91292	0.88471	N
27	IFNA1	0.00445771	0.000144	4.50783	0.86619	1.44725	N
28	CDK8	0.00499765	0.00018	4.4201	0.79124	1.45097	N
29	KIT	0.00499765	0.000179	-4.4221	0.73882	-1.43821	N
30	FAM30A	0.00499765	0.000176	4.42904	0.59702	0.89892	N
31	JAK3	0.00529457	0.000198	-4.38413	0.66668	-1.41981	N
32	TREX1	0.00564171	0.000221	4.34021	0.57291	1.26469	N
33	CD8B	0.00564171	0.000225	4.33385	0.38966	1.06973	N
34	RBPJ	0.005814	0.000239	4.30994	0.41366	1.27979	N
35	ORAI1	0.0062933	0.000266	4.26707	0.43366	0.89372	N
36	MYD88	0.00729257	0.000317	4.19771	0.11883	0.65205	N
37	SPPL2B	0.00731417	0.000333	-4.17869	0.21618	-1.21078	N
38	AHR	0.00731417	0.00035	-4.15881	0.01623	-0.66606	N
39	POU2F2	0.00731417	0.000362	-4.14579	-0.00058	-0.86352	N
40	ENO1	0.00731417	0.000364	4.14398	-0.0293	0.68762	N
41	TBX21	0.00731417	0.000359	4.14961	-0.11601	0.8337	N
42	ANGPT1	0.00754023	0.000394	-4.11276	-0.11812	-0.67797	N
43	IL7	0.00754023	0.000386	4.12115	-0.269	0.6644	N

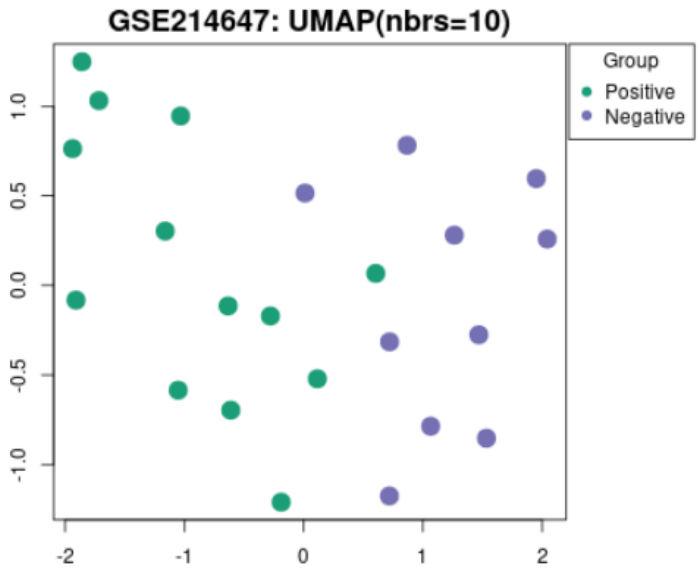
44	CCR1	0.00819756	0.000438	4.07052	-0.05653	1.17805	N
45	NUDT1	0.00841591	0.000461	4.05108	-0.10505	0.74837	N
46	CD45R0	0.0089785	0.000503	-4.01665	-0.19221	-1.43441	N
47	C3	0.00910205	0.000544	-3.98548	-0.28079	-1.9193	N
48	SIGLEC5	0.00910205	0.000525	3.99922	-0.28772	1.23009	N
49	TNFSF10	0.00910205	0.000535	-3.99155	-0.3785	-0.76349	N
50	CREB3L4	0.00910205	0.000555	3.9776	-0.64457	0.64013	N
51	MS4A2	0.00923048	0.000574	-3.96406	-0.48769	-0.75326	N
52	BCLAF1	0.0093213	0.000603	3.94467	-0.46468	1.39913	N
53	HSP90AB1	0.0093213	0.000601	3.94559	-0.50863	0.80283	N
54	TUBB	0.00932266	0.000615	3.93707	-0.46394	0.79849	N
55	MAPK7	0.00982378	0.000672	3.90165	-0.57376	0.63181	N
56	ITPR3	0.00982378	0.000663	3.907	-0.57395	0.88685	N
57	IRF7	0.01027391	0.000716	3.87674	-0.54888	1.10996	N
58	TNFRSF17	0.01102918	0.000782	3.84156	-0.89319	0.76361	N
59	IL6	0.01110165	0.000815	-3.82525	-0.66577	-1.81614	N
60	ARF1	0.01110165	0.000811	3.82713	-0.8015	0.44068	N
61	AP2A2	0.01150156	0.000858	3.8045	-0.81424	0.56005	N
62	RNF126	0.01184959	0.000899	3.78606	-0.76426	0.87481	N
63	LTA	0.0119135	0.000919	3.77744	-0.96954	1.07885	N
64	FLT1	0.01210521	0.000949	3.76471	-0.77196	1.0532	N
65	IRAK1	0.01216049	0.000968	3.75662	-0.92675	0.78891	N
66	PSTPIP1	0.01283421	0.00105	-3.72479	-0.9842	-0.79265	N
67	TOLLIP	0.01283421	0.00107	3.7178	-1.00788	0.90464	N
68	SDHA	0.01283421	0.00107	3.71684	-1.03262	0.68237	N
69	TRAF5	0.01306944	0.00112	-3.69784	-0.91693	-1.11413	N
70	MDM2	0.01306944	0.00112	3.69887	-1.00274	0.81555	N
71	CYBA	0.01316344	0.00115	-3.68922	-0.95961	-1.29371	N
72	TPSAB1/B2	0.01331476	0.00118	-3.67898	-1.14385	-1.40617	N
73	USP21	0.01438178	0.00131	-3.63723	-1.1295	-0.9915	N
74	SOCS3	0.01438178	0.00131	-3.637	-1.24378	-0.84175	N
75	HLA-DMA	0.01447512	0.00135	-3.62357	-1.15111	-1.10897	N
76	XBP1	0.01447512	0.00133	-3.62873	-1.25121	-1.05187	N
77	HCoV-229E_S	0.01560248	0.00147	3.58815	-1.32218	1.05442	
78	POS_D	0.01604637	0.00154	3.57162	-1.39575	0.44795	
79	UBE2S	0.01619143	0.00157	3.56281	-1.40823	0.75889	N
80	BCL2L1	0.01646179	0.00162	3.55102	-1.41397	0.74024	N
81	HMGB2	0.01732114	0.00172	3.52545	-1.45418	0.83001	N
82	RAG2	0.01739133	0.00175	3.51881	-1.70722	0.69396	N
83	IKBKB	0.01774329	0.00183	-3.50088	-1.37398	-0.97398	N
84	ATP6V0C	0.01774329	0.00183	3.50227	-1.57942	0.55378	N
85	CARD8	0.01794001	0.00192	3.48208	-1.40511	0.97458	N
86	CD209	0.01794001	0.00191	-3.48395	-1.46121	-1.72236	N

87	UBE2L3	0.01794001	0.00191	3.48474	-1.48036	1.01065
88	ATG9A	0.01836354	0.00199	3.46797	-1.56111	0.69773
89	IFNB1	0.01853927	0.00205	3.45491	-1.80885	0.79795
90	AICDA	0.01853927	0.00203	3.45941	-1.90523	0.6144
91	IL13	0.01887084	0.00211	3.44321	-1.8204	0.84695
92	UNC93B1	0.02250847	0.00255	-3.36711	-1.6858	-1.16925
93	CXCL12	0.022854	0.00262	-3.35645	-1.77632	-1.87591
94	LRRC41	0.02402805	0.00283	-3.32366	-1.87836	-1.11868
95	IL12RB1	0.02402805	0.00284	3.3229	-2.18379	0.66392
96	GZMM	0.02402805	0.00282	3.32572	-2.21021	0.62893
97	NPRL2	0.02431886	0.0029	-3.31371	-1.93059	-0.77429
98	BAD	0.02468394	0.00298	-3.30337	-1.84832	-1.3151
99	CSF1	0.02511048	0.00306	-3.29216	-1.86495	-1.10372
100	POLR3D	0.02518869	0.0031	3.28673	-1.84079	0.73425
101	ADORA2A	0.0253104	0.00315	3.28063	-2.04995	1.21341

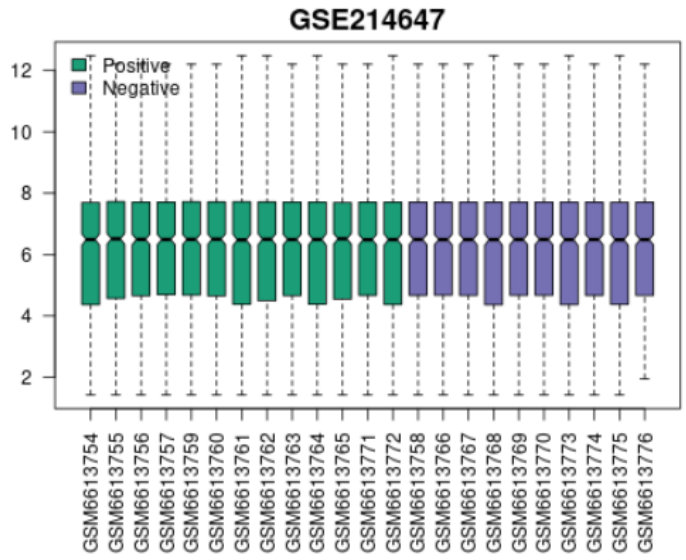
Bonus: Using **Reactome** online tool, I was able to investigate the pathways that these genes are involved within. They are mainly **immune and cytokine signaling**. That is, the upregulated DEGs perform various function that are related to immunity activation. The tool took an **input of a .txt file** called 'genes.txt' that has ONLY the 26 upregulated DEGs since I am interested in the differential phenomena from the perspective of the experimental, COVID-19 positive group.

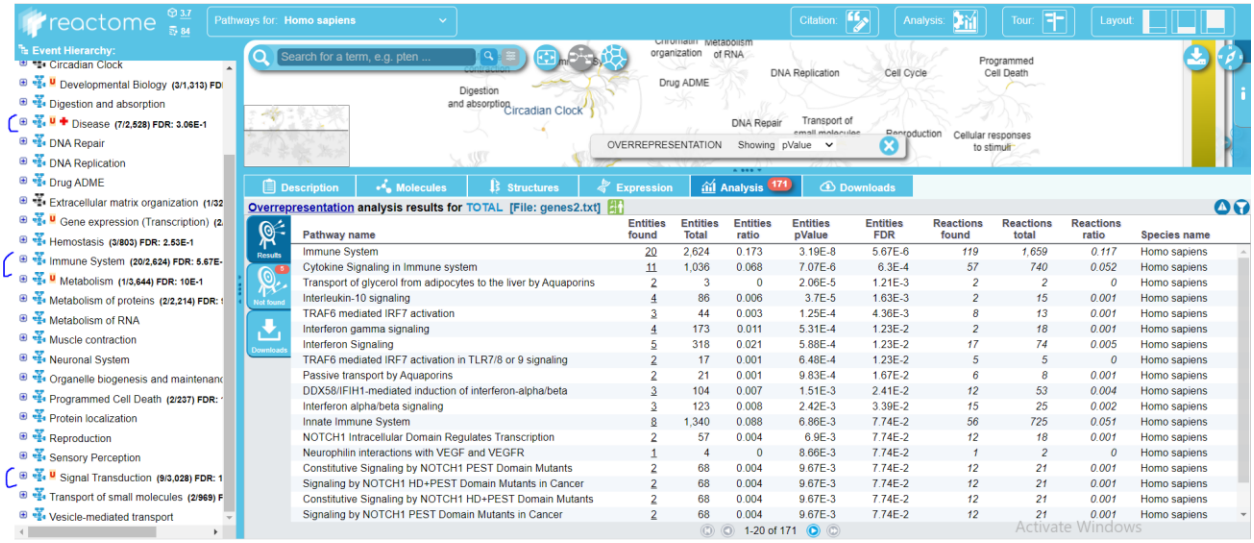


UMAP plot



Boxplot

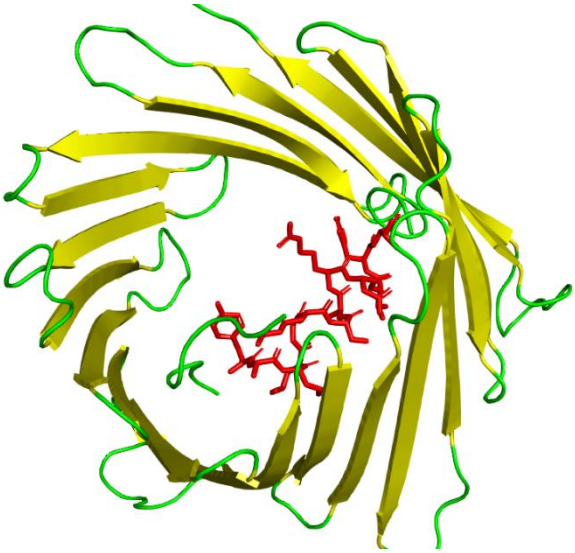


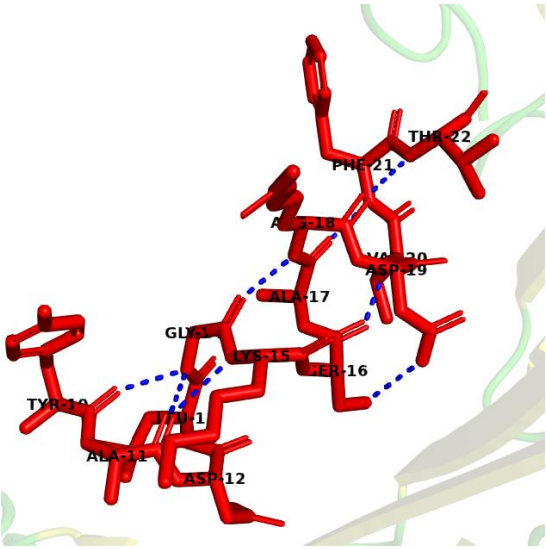


Conclusions: Based on the DGEA (statistical cutoff of adjusted p-value of 0.05 and biological cutoff of log-fold change of 1), it is concluded that there are 69 DEGs, 43 under-expressed and 26 over-expressed in the COVID-19-positive cohort in comparison to the COVID-19-negative cohort. Based on pathway analysis, these upregulated genes mostly participate in immune activation and cytokine signaling. This suggests that COVID-19 positive patients may be more affected by sacral pressure injuries since they devote most of their immune defenses towards viral elimination. However, the sample size of ONLY 23 patients is a major limitation of this analysis. Moreover, only 804 genes are studied; some important genes may not have been reported.

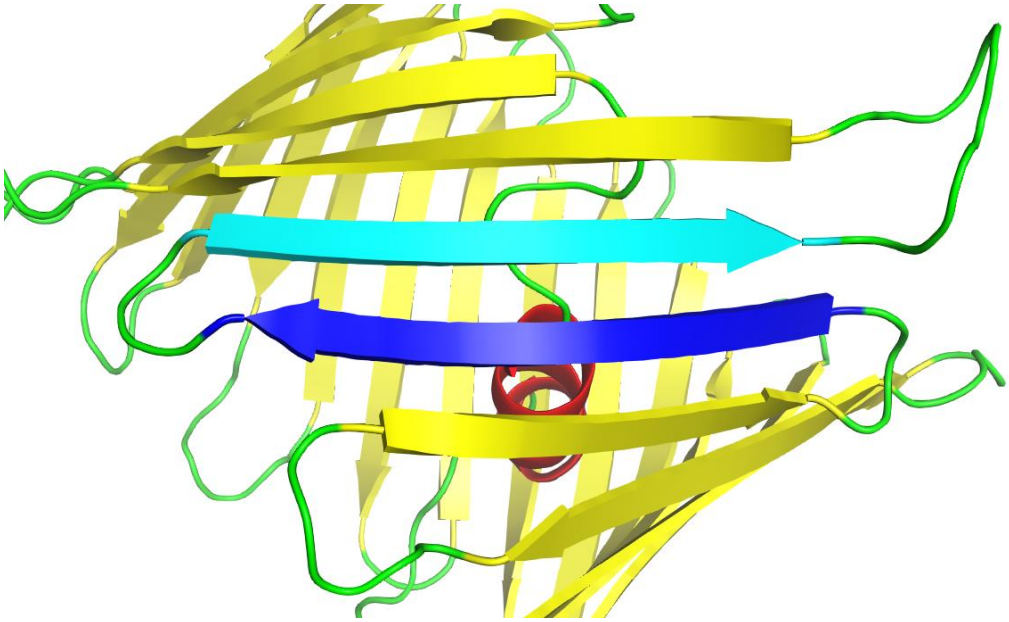
Question II: Protein Structure PyMol

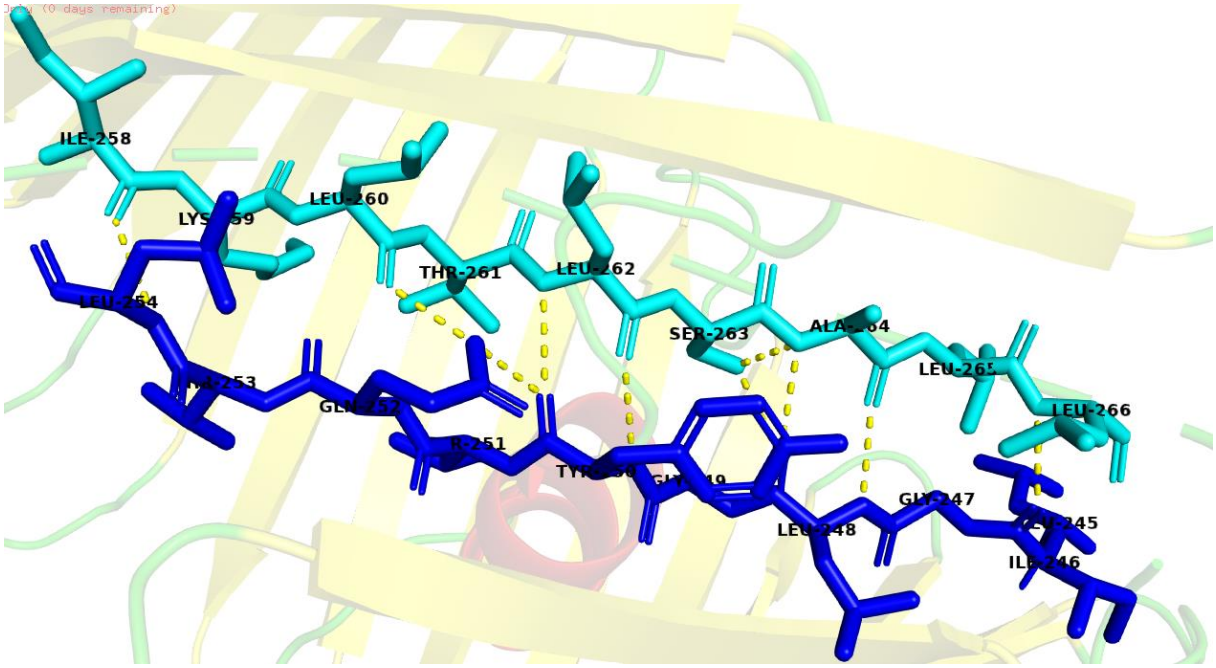
1. PDB: 2JK4. The helix is in red, shown as sticks. The hydrogen bonds within the helix backbone atoms are illustrated in blue.





2. PDB: 2JK4. The two antiparallel sheets are illustrated as cartoon in cyan and blue. The second figure illustrate the sheets in sticks representation and highlights the hydrogen bonds within the sheets' backbones in yellow.



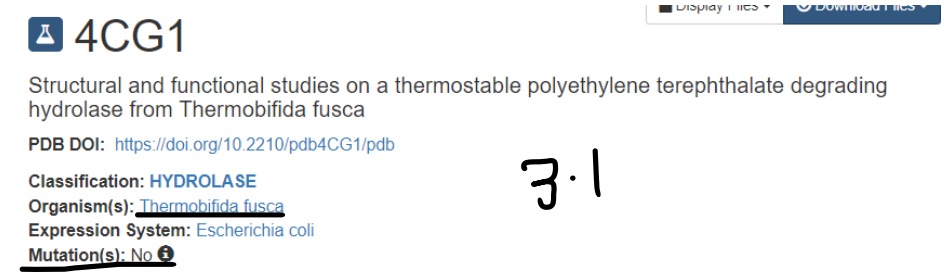


Question III: Homologous Proteins

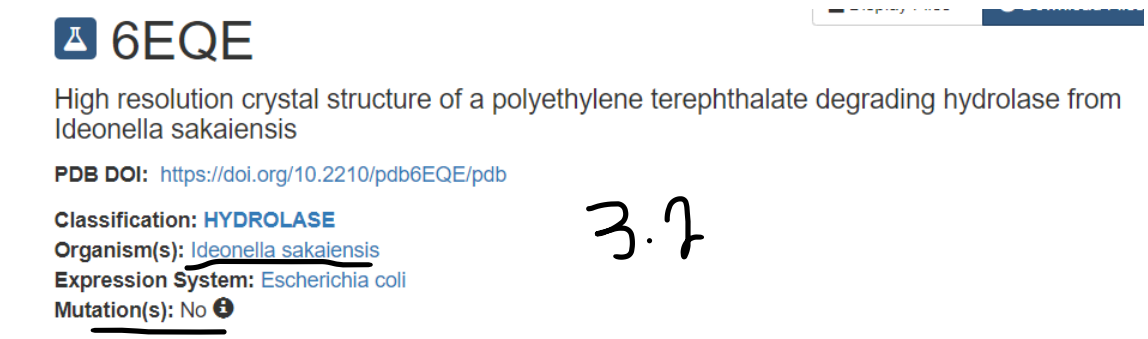
Part A)

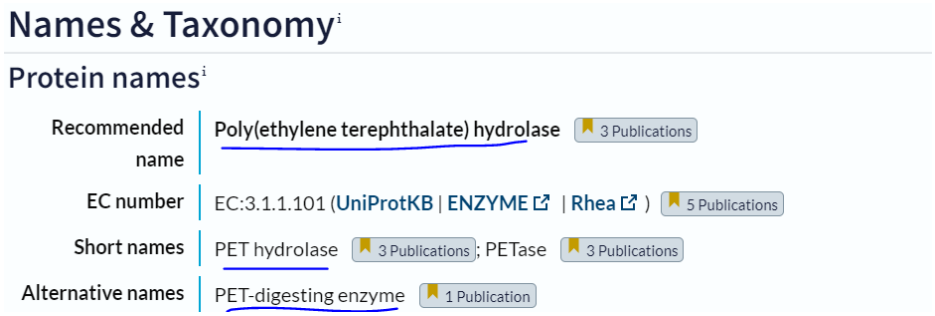
1.

- 4CG1 | Name: **Cutinase** (can also be called Polyethylene terephthalate (PET) hydrolase) | Organism: **Thermobifida fusca**



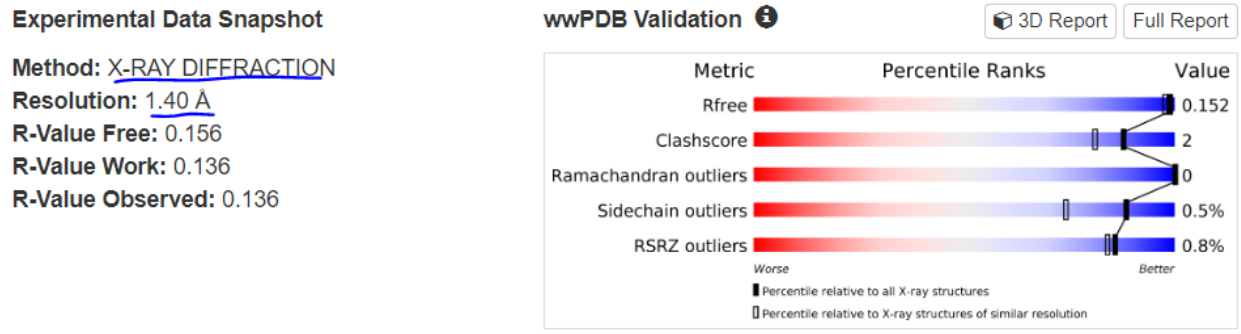
- 6EQE | Name: **Polyethylene terephthalate (PET) hydrolase** | Organism: **Ideonella sakaiensis**



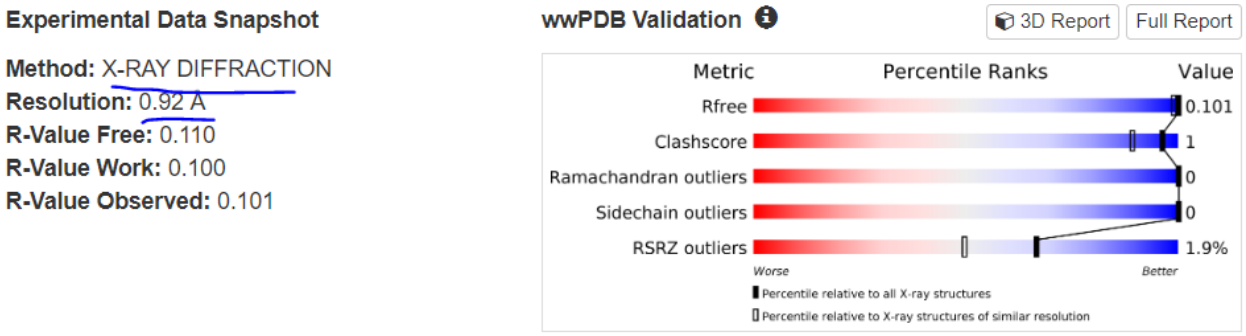


2.

- 4CG1 | Method: **X-RAY Diffraction** | Resolution: **1.40Å** | This is a **good resolution**.



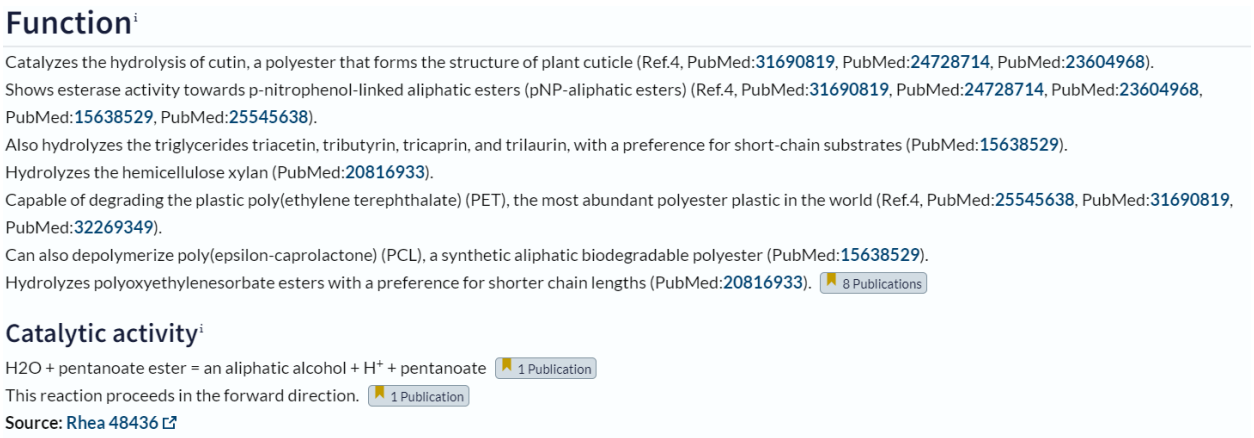
- 6EQE | Method: **X-RAY Diffraction** | Resolution: **0.92Å** | This is an **excellent resolution**.



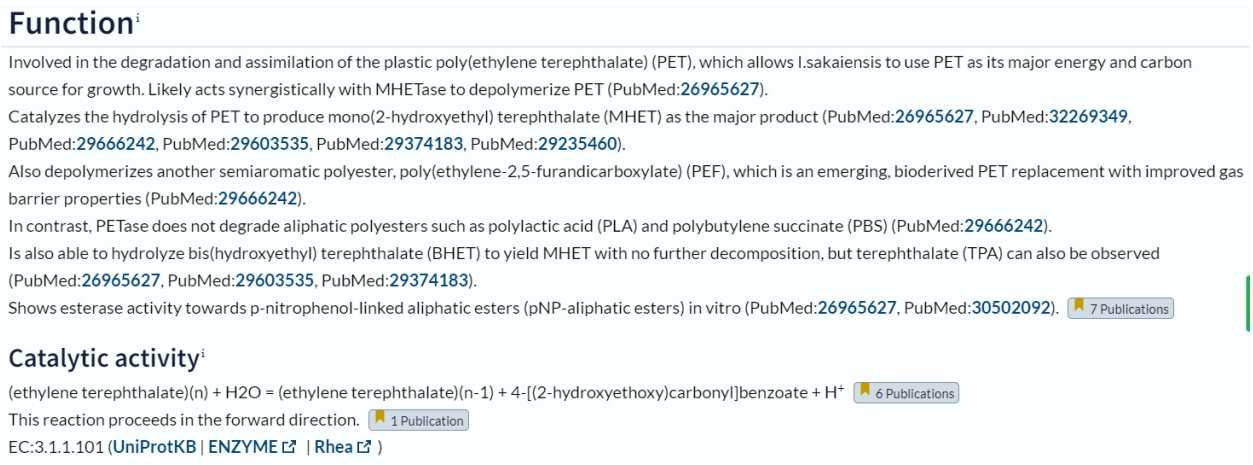
3. **None** of the structure is **mutated**, as per **figures 3.1 and 3.2 above**.

4.

- 4CG1 | Subcellular Location: **Secreted & on Periplasm** | Functions: **Hydrolysis of Cutin, Esterase activity towards p-nitrophenol-linked aliphatic esters, Hydrolysis of triglycerides, Hydrolysis of hemicellulose xylan, Degradation of PET**



- 6EQE | Subcellular Location: **Secreted** | Functions: **Degradation and assimilation of PET, Catalyze hydrolysis of PET, Does not degrade aliphatic polyesters.**



Function

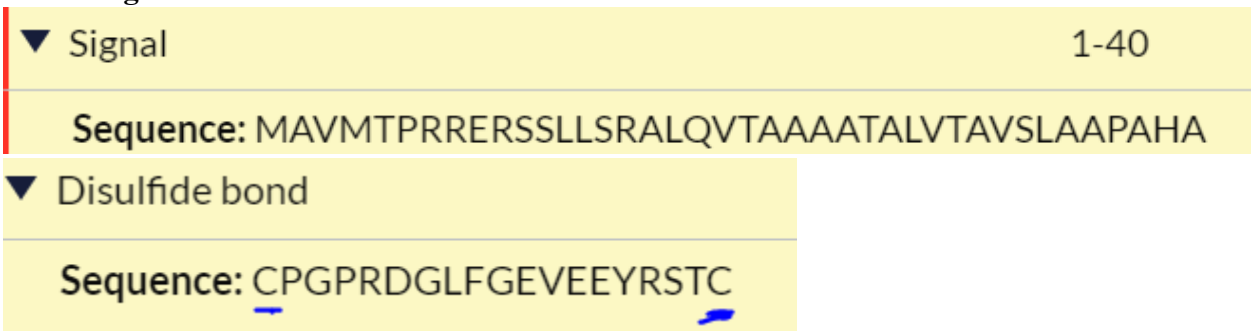
Involved in the degradation and assimilation of the plastic poly(ethylene terephthalate) (PET), which allows I.sakaiensis to use PET as its major energy and carbon source for growth. Likely acts synergistically with MHETase to depolymerize PET (PubMed:26965627).
Catalyzes the hydrolysis of PET to produce mono(2-hydroxyethyl) terephthalate (MHET) as the major product (PubMed:26965627, PubMed:32269349, PubMed:29666242, PubMed:29603535, PubMed:29374183, PubMed:29235460).
Also depolymerizes another semiaromatic polyester, poly(ethylene-2,5-furandicarboxylate) (PEF), which is an emerging, bioderived PET replacement with improved gas barrier properties (PubMed:29666242).
In contrast, PETase does not degrade aliphatic polyesters such as polylactic acid (PLA) and polybutylene succinate (PBS) (PubMed:29666242).
Is also able to hydrolyze bis(hydroxyethyl) terephthalate (BHET) to yield MHET with no further decomposition, but terephthalate (TPA) can also be observed (PubMed:26965627, PubMed:29603535, PubMed:29374183).
Shows esterase activity towards p-nitrophenol-linked aliphatic esters (pNP-aliphatic esters) in vitro (PubMed:26965627, PubMed:30502092). 7 Publications

Catalytic activity

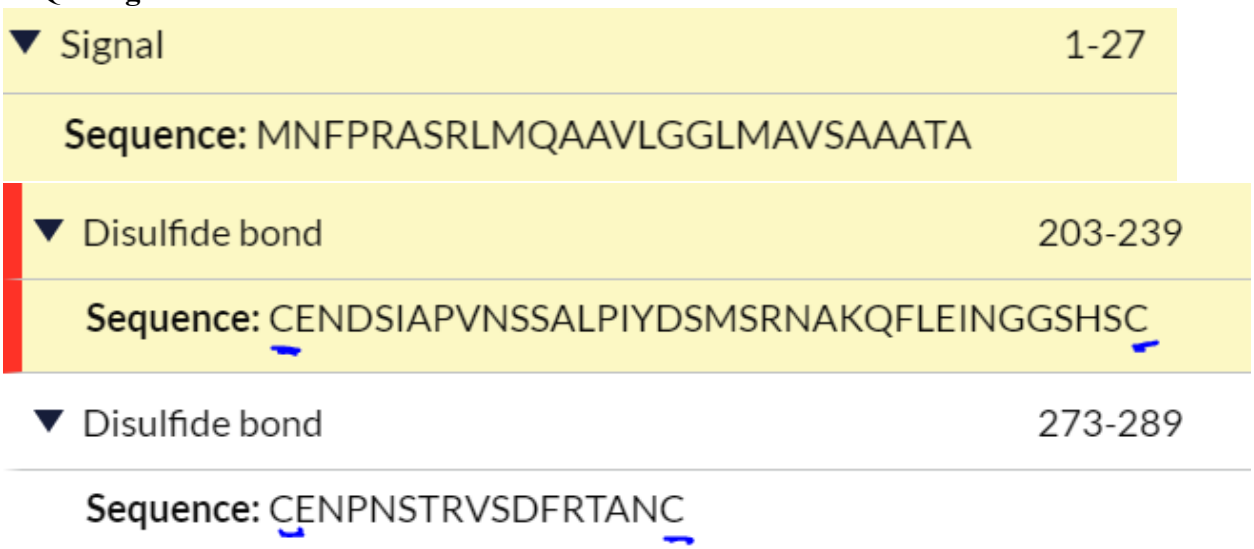
(ethylene terephthalate)(n) + H2O = (ethylene terephthalate)(n-1) + 4-[(2-hydroxyethoxy)carbonyl]benzoate + H^+ 6 Publications
This reaction proceeds in the forward direction. 1 Publication
EC:3.1.1.101 (UniProtKB | ENZYME | Rhea)

5.

- 4CG1: **Signal from 1-40 and disulfide bond from 281 to 299**



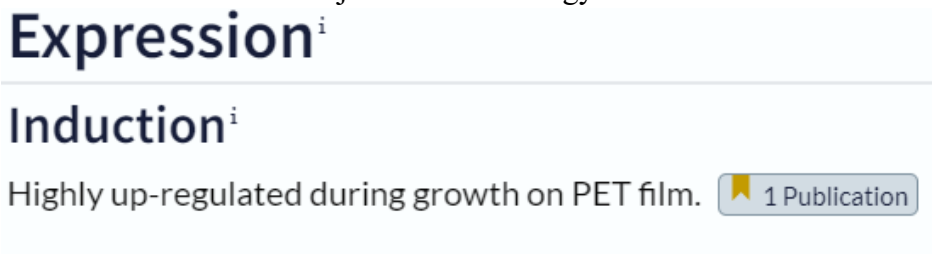
- 6EQE: **Signal from 1 to 27** and **disulfide bonds from 203 to 239 and from 273 to 289**



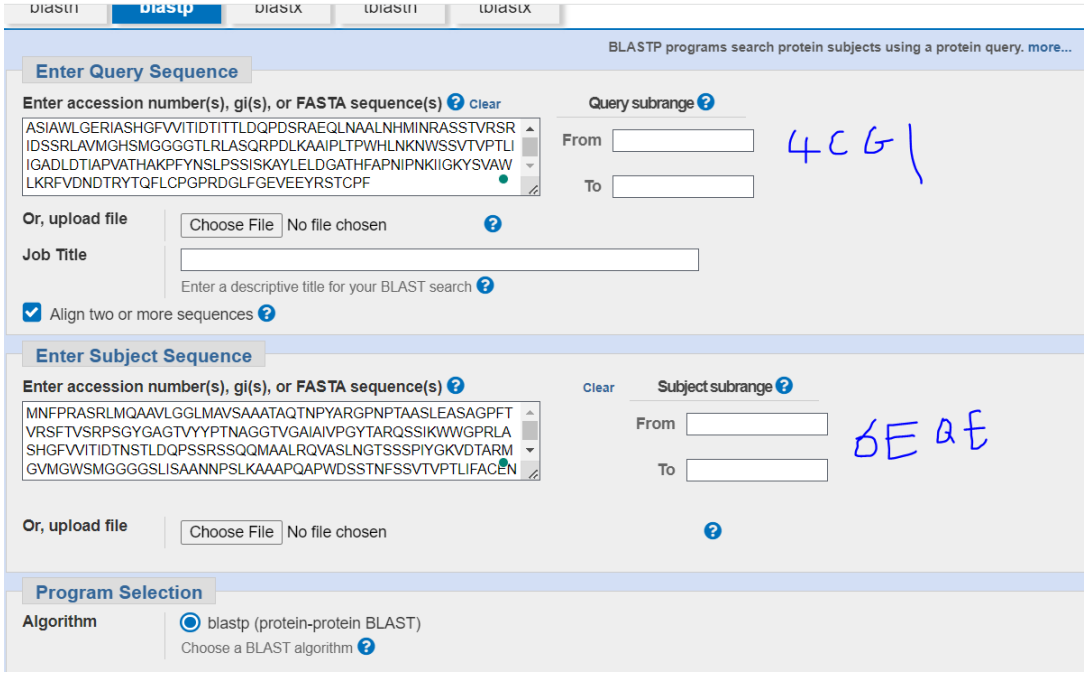
6.

- 4CGA: Not available on Uniprot, but we should expect behavior to the other protein since they are of the same family.

- 6EQE: Upregulated during growth on PET film, which makes sense this protein is a PETase and PET is the major source of energy and carbon to the bacterium.

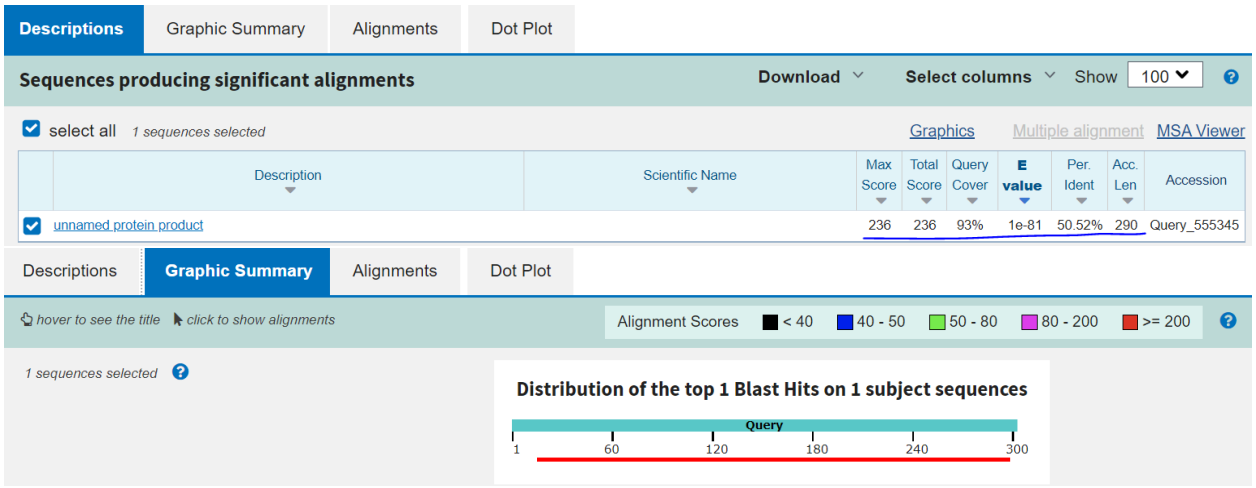


7. To study sequence homology, I will align the two proteins using BLASTP with all the default parameters. To conduct structural homology, I will align the two proteins on PyMol and investigate the RMSD.

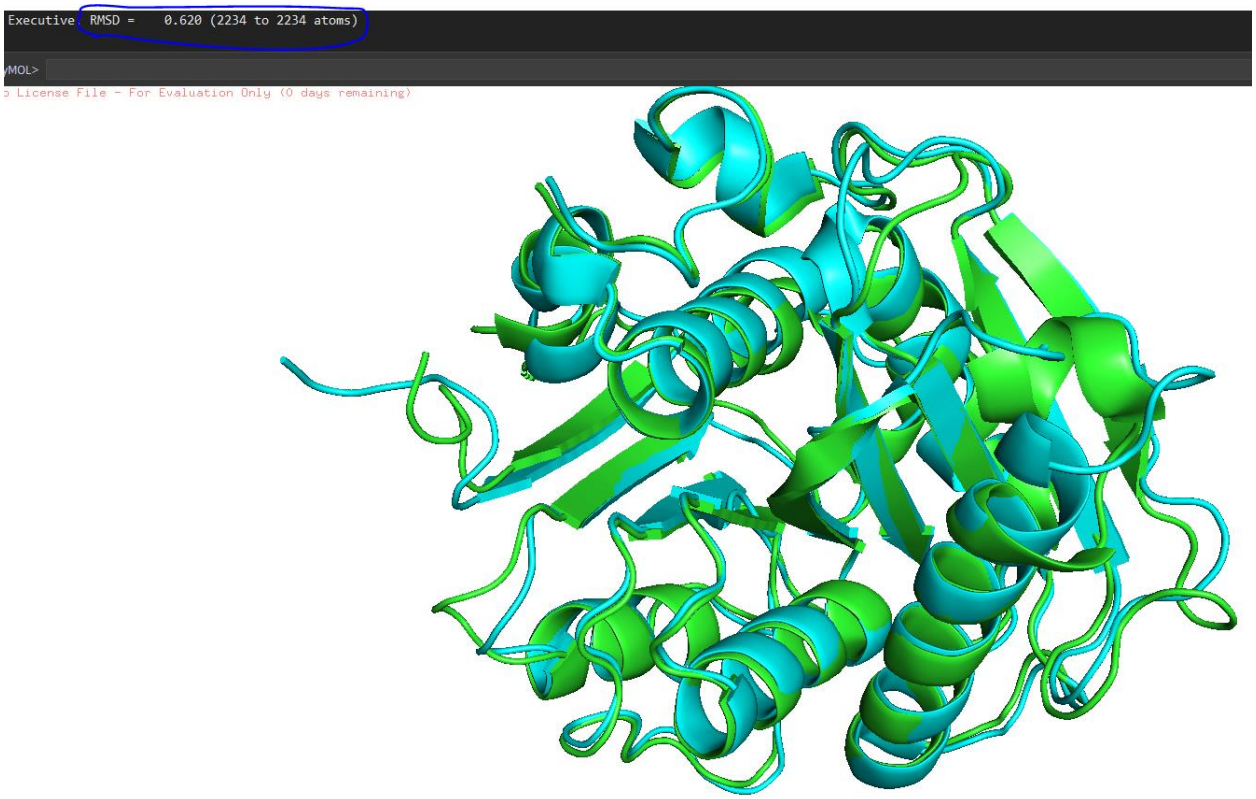


Based on BLAST results below, the sequences have 50% identity, max and total score of 236, query coverage of 93%, and E-value of $1e^{-81}$. Moreover, the graphical summary indicate high

alignment scores throughout the query sequence since all aligned regions are in red, indicating an alignment score >= 200, as per the figure below.



For the structure homology, I aligned the two sequences using the command “align 4CG1, 6EQE”. Based on the figure below, they are highly aligned. Green is 4CG1 and cyan is 6EQE. The RMSD is 0.620, supporting the conclusion of high alignment and structural homology.



Sequence Homology: Although the sequences have only 50% identity, they align throughout their sequences with high alignment scores as per the figures above, meaning they share high similarity. Moreover, homology is a hypothesis that is mostly based on similarity and

identity. Therefore, we cannot deduce homology directly, but we can say that it is highly likely the two sequences share sequence homology given their alignment results. Nevertheless, the reason for the low identity may be due to the fact that they are from different organisms.

Structural Homology: The sequences are highly homologous in structure since their RMSD alignment is significantly less than 2. This is in accordance with their high similarity and alignment results, and given the fact that they are both Hydrolases of the same family.

Overall, I would not say there is a discrepancy between sequence and structure homology. Even if they were not homologous in sequence, it is possible that two entirely different sequences would have the same protein structure and thus exhibit structural homology.

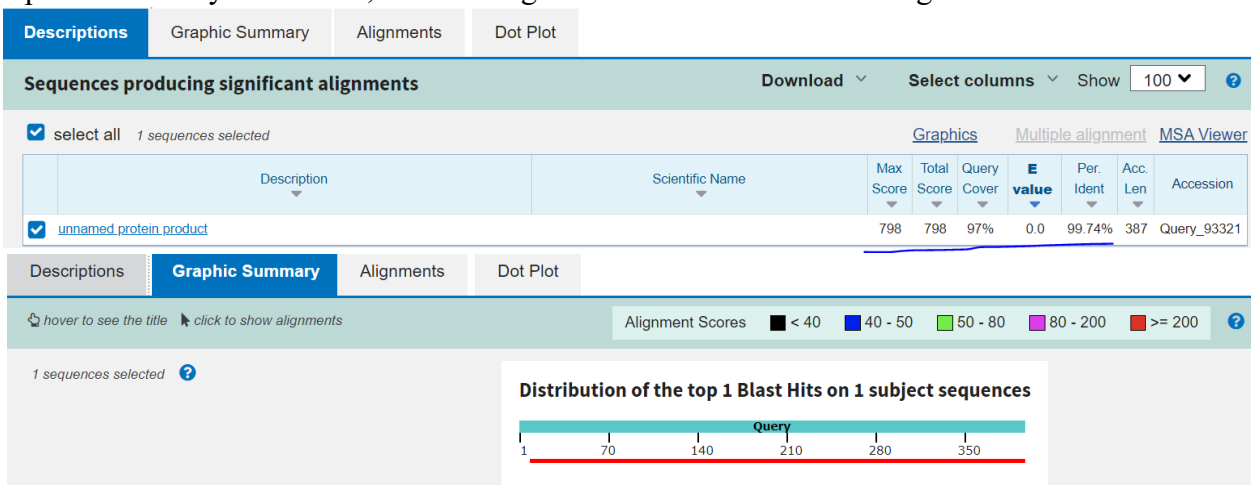
Part B)

The two proteins are definitely homologous in sequence and structure as indicated by BLASTP alignment and by PyMOL RMSD alignment. They also perform similar functions; that is, they are both dehydrogenases involved in fatty acid beta-oxidation.

First, on Uniprot, both PDB codes are given for the same protein entry, which means they are the same protein that was just experimentally determined more than time.

PDB	1EGC	X-ray	2.60 Å	A/B/C/D	26-421	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1EGD	X-ray	2.40 Å	A/B/C/D	26-421	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1EGE	X-ray	2.75 Å	A/B/C/D	26-421	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1T9G	X-ray	2.90 Å	A/B/C/D	26-421	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	2A1T	X-ray	2.80 Å	A/B/C/D	1-421	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	4P13	X-ray	1.73 Å	A/B/C/D	35-421	PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-P11310-F1	Predicted			1-421	AlphaFold

Second, they align almost perfectly, producing high alignment score throughout the sequence and a percent identity of 99.74%, with other good scores illustrated in the figures below.



Third, using “align 1EGE, 4P14” on PyMOL, the RMSD value is 0.492, which is very low, indicating almost identical structures. Pink is 4P13 and yellow is 1EGE.

Therefore, given the fact that they are almost identical in sequence and structure, and that they are both acyl-CoA dehydrogenases that perform the first step of mitochondrial fatty acid beta-oxidation, plus also having the same ligand which is Flavin-Adenine Dinucleotide (FAD), we can conclude that in general proteins that share high sequence and structural homology are expected to have the same function.

However, this can only be concluded if we do not have more information other than sequence and structure homology. That is, two proteins with high sequence and structural homology may behave differently if they are in different tissues or under different physiological conditions. Moreover, they can behave differently if they undergo different Post-Translational Modifications (PTMs) or bind different ligands at different binding sites. But in general, without lots of information, I agree with the hypothesis that says high sequence and structural homology likely indicates similar functionality, unless proven otherwise.