

DOWNLOADING THE PDC GENES FROM NCBI

- Setting Email for Entrez:
 - The `Entrez.email` variable is set to provide the email address to NCBI's Entrez utilities. This is a requirement by NCBI to identify users and prevent abuse of their services.
- Function `download_gene_sequence`:
 - This function downloads the gene sequence from NCBI using specified parameters.
 - Parameters:
 - `accession_number`: Accession number of the gene sequence in the NCBI database.
 - `start`: Start position of the gene sequence.
 - `end`: End position of the gene sequence.
 - `pdc_gene_number`: Identifier for the PDC (Pyruvate decarboxylase) gene.
 - `reverse_complement` (optional): Boolean flag indicating whether to retrieve the reverse complement of the sequence.
 - Steps:
 - Print Start Message: Displays a message indicating the start of the download process for the specified gene sequence.
 - Fetch Sequence: Uses NCBI's Entrez utility (`efetch`) to fetch the gene sequence from the nucleotide database.
 - Read Sequence: Reads the fetched sequence using Biopython's `SeqIO module`.
 - Close Handle: Closes the handle to the sequence data retrieved from NCBI.
 - Reverse Complement: Optionally applies reverse complement to the sequence if specified.
 - Write to File: Writes the downloaded gene sequence to a FASTA file named according to the PDC gene number.
 - Print Success Message: Displays a message indicating successful download of the gene sequence.
 - Returns: The downloaded sequence as a `SeqRecord` object.
- Downloading PDC1 Gene Sequence:
 - Downloads the PDC1 gene sequence from NCBI.
 - Parameters:
 - `accession_number_pdc1`: Accession number of the PDC1 gene sequence.
 - `start_pdc1`: Start position of the PDC1 gene sequence.
 - `end_pdc1`: End position of the PDC1 gene sequence.
 - Flags reverse complement to be applied.
- Downloading PDC5 Gene Sequence:
 - Downloads the PDC5 gene sequence from NCBI.
 - Parameters:

- `accession_number_pdc5`: Accession number of the PDC5 gene sequence.
 - `start_pdc5`: Start position of the PDC5 gene sequence.
 - `end_pdc5`: End position of the PDC5 gene sequence.
- Downloading PDC6 Gene Sequence:
 - Downloads the PDC6 gene sequence from NCBI.
 - Parameters:
 - `accession_number_pdc6`: Accession number of the PDC6 gene sequence.
 - `start_pdc6`: Start position of the PDC6 gene sequence.
 - `end_pdc6`: End position of the PDC6 gene sequence.
 - Flags reverse complement to be applied.

Applying Blast

Blast through code.

- I have provided the code to blast the PDC's with *Saccharomyces cerevisiae* genome
- Importing Necessary Modules:
 - The code imports required modules from Biopython, namely `NCBIWWW` and `NCBIXML`, for performing BLAST searches and parsing BLAST results, respectively.
- Defining `perform_blast()` Function:
 - This function performs a BLAST search against the NCBI BLAST database using the provided query sequence.
 - It returns a handle to the BLAST search results.
- Defining `parse_blast_results()` Function:
 - This function parses the BLAST search results to identify SNPs (Single Nucleotide Polymorphisms).
 - It iterates through each BLAST record, alignment, and High Scoring Pair (HSP) to identify SNPs.
 - Identified SNPs are stored in a set.
- Iterating Over Files and Performing BLAST Searches:
 - The code iterates over a list of file indices representing different genes.(ie PDC1, DC2, PDC3)

- For each gene, it prepares a query sequence by extracting the sequence from the corresponding FASTA file.
 - It then performs a BLAST search using the `perform_blast()` function and parses the results using the `parse_blast_results()` function.
- Additional Notes:
 - Biopython's functionality is utilised for BLAST searches and result parsing.
 - The code specifically searches for SNPs in the *Saccharomyces cerevisiae* genome using the provided query sequences.
 - SNPs are identified by comparing the query and subject sequences from BLAST alignments.
- Summary:
 - This script automates the process of searching for SNPs in the *Saccharomyces cerevisiae* genome using BLAST.
 - It defines functions for performing BLAST searches and parsing results, iterates over a list of genes, and extracts sequences from FASTA files to perform BLAST searches.
 - Finally, it prints the identified SNPs
- OUTPUT ->


```
-----
For PDC 1
****Alignment****
Sequence: gi|1586061137|gb|CP036478.1| Saccharomyces cerevisiae strain ySR128 chromosome XII, complete sequence
Length: 1076801
E-value: 0.0
ATGCTGAAATTACTTGGTAAATTTGTTGAAAGATTAAAGCAAGTCACGTTAACACCGTTTCGGTTG...
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...|
ATGCTGAAATTACTTGGTAAATTTGTTGAAAGATTAAAGCAAGTCACGTTAACACCGTTTCGGTTG...

****Alignment****
Sequence: gi|1586061137|gb|CP036478.1| Saccharomyces cerevisiae strain ySR128 chromosome XII, complete sequence
Length: 1076801
E-value: 0.0
ATGCTGAAATTACTTGGTAAATTTGTTGAAAGATTAAAGCAAGTCACGTTAACACCGTTTCGGTTG...
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...|
ATGCTGAAATAACCTTAGGTAAATTTGAAAGATTGAGCCAAGTCACGTTAACACCGTCTTCGGTTG...

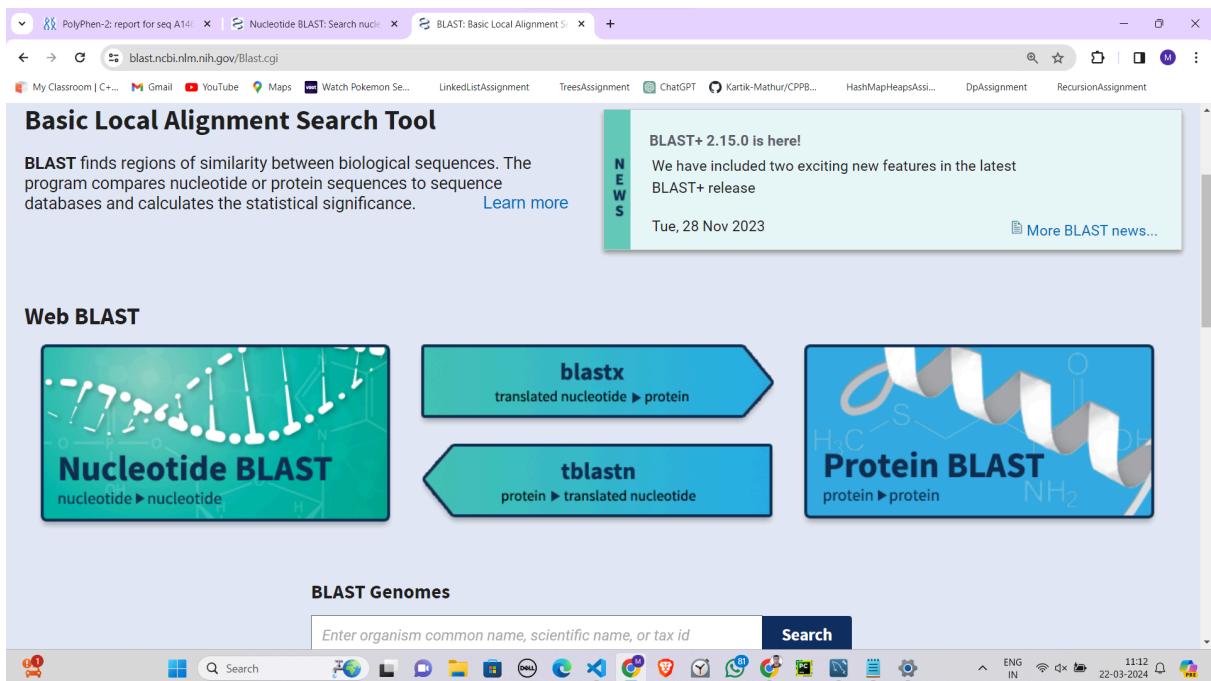
****Alignment****
Sequence: gi|1389182097|gb|CP029160.1| Saccharomyces cerevisiae strain SY14 chromosome I, complete sequence
Length: 11848804
E-value: 0.0
ATGCTGAAATTACTTGGTAAATTTGTTGAAAGATTAAAGCAAGTCACGTTAACACCGTTTCGGTTG...
|||||||||||||||||||||||||||||||||||||||||||||||||||||||...|
ATGCTGAAATTACTTGGTAAATTTGTTGAAAGATTAAAGCAAGTCACGTTAACACCGTTTCGGTTG...

Number of snps -> 783
Some of the snps ..... -> [12, 15, 16, 18, 21, 27, 30, 33, 42, 44, 45, 51, 54, 55, 56]
```

- NOTE- *The above ss is only for pdc 1. Also the number of snp indicates number of positions and does not include number of combinations of mutations on a single position. Therefore my number of snps might be less.*
- NOTE: The number of snp is indicated for separate PDC strains
 - First printing the PDC gene number
 - Then printing the subject sequence information
 - Then the alignment
 - At the end of each pdc I am also printing the total number of SNP found and printing first 10 or 15 positions.

BLAST through UI

- First visit the official blast website ([link for blast website](#))



- Then Click on the Nucleotide Blast (nucleotide -> nucleotide) button
- After this step the following page will open

BLAST® » blastn suite

Standard Nucleotide BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

Or, upload file No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

- Copy paste the fasta file of the pdc gene 1 in the QUERY Sequence Box shown in the above web page. Also enter the organism name ie Saccharomyces cerevisiae across which you need to perform blast of the query. After filling this information simply click on the blast button.

Sequences producing significant alignments		Download		Select columns		Show 100	<input type="button" value="Feedback"/>	
		GenBank	Graphics	Distance tree of results	MSA Viewer			
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain ySR128 chromosome XII, complete sequence	Saccharomyces cer...	3125	5087	100%	0.0	100.00%	1076801	CP036478.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain SY14 chromosome I, complete sequence	Saccharomyces cer...	3125	5903	100%	0.0	100.00%	11848804	CP029160.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain BY4742 chromosome XII, complete sequence	Saccharomyces cer...	3125	5087	100%	0.0	100.00%	1104511	CP026300.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain DBVPG6765 chromosome XII sequence	Saccharomyces cer...	3125	5093	100%	0.0	100.00%	1022186	CP020168.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain S288c chromosome XII sequence	Saccharomyces cer...	3125	5087	100%	0.0	100.00%	1075542	CP020134.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain HB_S_GIMBLETTROAD_14 chromosome XII sequence	Saccharomyces cer...	3125	5093	100%	0.0	100.00%	1077529	CP008264.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain HB_C_TUKITUKI1_16 chromosome XII sequence	Saccharomyces cer...	3125	5093	100%	0.0	100.00%	1077438	CP008400.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain HB_C_TUKITUKI2_10 chromosome XII sequence	Saccharomyces cer...	3125	5093	100%	0.0	100.00%	1077494	CP008383.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain WI_S_JASA_5 chromosome XII sequence	Saccharomyces cer...	3125	5093	100%	0.0	100.00%	1076962	CP008349.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain T52 chromosome XII sequence	Saccharomyces cer...	3125	5089	100%	0.0	100.00%	1077515	CP008502.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain HB_C_KOROKIPO_3 chromosome XII sequence	Saccharomyces cer...	3125	5093	100%	0.0	100.00%	1077202	CP008468.1
<input checked="" type="checkbox"/> Saccharomyces cerevisiae strain HB_C_OMARUNUI_14 chromosome XII sequence	Saccharomyces cer...	3125	5095	100%	0.0	100.00%	1077859	CP008451.1

- The above page will appear on the screen as a result.
- Note that there is a column of percent identity as highlighted in the above image. As we are finding SNP (Single Nucleotide Polymorphism) which means that only one nucleotide should change in the alignment. Implies that Percent Identity = $\{len(query)-1/len(query)\}*100$

- So as the len of pdc gene 1 is 1692 so Percent Identity = $(1691/1692) * 100$ ie approx 99.94%
 - In the more general sense, it is not necessary for only one nucleotide to change in a sequence. Check the references section for more details.

Blast search results for PB Assignment-2									
	Query		Database		Statistics		Summary		Details
	Query ID	Length	Database ID	Length	Score	E-value	Bit score	Expect	Link
1	Saccharomyces cerevisiae YPH499 DNA, chromosome 12, complete sequence	5087	Saccharomyces cerevisiae	5087	100%	0.0	100.0%	1220291	AP028843.1
2	Saccharomyces cerevisiae strain 12651cf5 macromosome XII	5087	Saccharomyces cerevisiae	5087	100%	0.0	100.0%	1078171	CP027005.1
3	Saccharomyces cerevisiae strain 12651 chromosome XII	5083	Saccharomyces cerevisiae	5093	100%	0.0	100.0%	787108	CP027010.1
4	Saccharomyces cerevisiae strain P19001 chromosome XII	5093	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1051807	CP027151.1
5	Saccharomyces cerevisiae strain DBY105cf5 macromosome XII	5093	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1042978	CP030324.1
6	Saccharomyces cerevisiae strain 1-174 chromosome XII	5087	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1020547	CP031721.1
7	Saccharomyces cerevisiae strain YLM878 chromosome XII	5083	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1008608	CP033881.1
8	Saccharomyces cerevisiae strain YLM875 chromosome XII	5083	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1020394	CP033881.1
9	Saccharomyces cerevisiae strain 1-1528 chromosome XII	5093	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1019782	CP036321.1
10	Saccharomyces cerevisiae strain YLM891 chromosome XII	5093	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1075117	CP035541.1
11	Saccharomyces cerevisiae S288C chromosome XII	5125	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1042878	CP029481.1
12	Saccharomyces cerevisiae YJM981 chromosome XII genomic sequence	5093	Saccharomyces cerevisiae YJM...	5125	100%	0.0	100.0%	1893211	CP004671.1
13	Saccharomyces cerevisiae LW25911 chromosome 12	5087	Saccharomyces cerevisiae	5125	100%	0.0	100.0%	1029517	CP029532.1
14	TPA: Saccharomyces cerevisiae S288C chromosome XII, complete sequence	5125	Saccharomyces cerevisiae S288C	5125	100%	0.0	100.0%	1078177	CP009845.1
15	Saccharomyces cerevisiae S288C endopeptidase leucylcarboxypeptidase I (PDC1), partial	5125	Saccharomyces cerevisiae S288C	5125	100%	0.0	100.0%	1892	MM_001181931.1
16	S.cerevisiae chromosome XII reading frame ORF R1044c	3125	Saccharomyces cerevisiae	3125	100%	0.0	100.0%	3027	CT2216.1
17	Saccharomyces cerevisiae strain SGD1 chromosome XII	4989	Saccharomyces cerevisiae	3121	100%	0.0	99.94%	1065425	CP066483.1
18	Saccharomyces cerevisiae strain CDR02R of H-thiaminease XII sequence	5082	Saccharomyces cerevisiae	3121	100%	0.0	99.94%	1077502	CP008001.1
19	Saccharomyces cerevisiae strain LA207 chromosome XII	5084	Saccharomyces cerevisiae	3121	100%	0.0	99.94%	1059687	CP011973.1
20	Saccharomyces cerevisiae strain YPS128 chromosome XII sequence	5137	Saccharomyces cerevisiae	3120	100%	0.0	99.94%	1027157	CP022018.1
21	Saccharomyces cerevisiae strain YPS128 chromosome XII	5137	Saccharomyces cerevisiae	3120	100%	0.0	99.94%	1043065	CP009730.1
22	Saccharomyces cerevisiae strain BC1167 chromosome XII	5137	Saccharomyces cerevisiae	3120	100%	0.0	99.94%	1027684	CP009744.1
23	Saccharomyces cerevisiae strain YPS606 chromosome XII	5137	Saccharomyces cerevisiae	3120	100%	0.0	99.94%	1028814	CP009681.1
24	Saccharomyces cerevisiae strain UWY05357_767.3 chromosome XII	5143	Saccharomyces cerevisiae	3120	100%	0.0	99.94%	1027091	CP009361.1
25	Saccharomyces cerevisiae strain WYOF557_2421 chromosome XII	5137	Saccharomyces cerevisiae	3120	100%	0.0	99.94%	1065541	CP009300.1
26	Saccharomyces cerevisiae strain HLL1 genome assembly, chromosome 12	4637	Saccharomyces cerevisiae	3120	100%	0.0	99.94%	1031746	LBB135112



<input checked="" type="checkbox"/>	TPA: <i>Saccharomyces cerevisiae</i> S288C chromosome XII, complete sequence	<i>Saccharomyces cer...</i>	3125	5087	100%	0.0	100.00%	1078177	BK006945.2	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae S288C indolepyruvate decarboxylase 1 (PDC1), partial m...	<i>Saccharomyces cer...</i>	3125	3125	100%	0.0	100.00%	1692	NM_001181931.1	
<input checked="" type="checkbox"/>	S.cerevisiae chromosome XII reading frame ORF YLR044c	<i>Saccharomyces cer...</i>	3125	3125	100%	0.0	100.00%	3023	Z73216.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain SK23 chromosome XII	<i>Saccharomyces cer...</i>	3121	4969	100%	0.0	99.94%	1065425	CP046463.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain CDRDR_sf_H chromosome XII sequence	<i>Saccharomyces cer...</i>	3121	5082	100%	0.0	99.94%	1077582	CP008009.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain LAN210 chromosome XII	<i>Saccharomyces cer...</i>	3121	5084	100%	0.0	99.94%	1059687	CP011679.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain YPS128 chromosome XII sequence	<i>Saccharomyces cer...</i>	3120	5137	100%	0.0	99.94%	1027157	CP020219.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain YPS128 chromosome XII	<i>Saccharomyces cer...</i>	3120	5137	100%	0.0	99.94%	1043035	CP093792.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain BC187 chromosome XII	<i>Saccharomyces cer...</i>	3120	5137	100%	0.0	99.94%	1027669	CP093744.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain YPS606 chromosome XII	<i>Saccharomyces cer...</i>	3120	5137	100%	0.0	99.94%	1028614	CP093648.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain UWOPS3_787_3 chromosome XII	<i>Saccharomyces cer...</i>	3120	5143	100%	0.0	99.94%	1027081	CP093616.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain UWOPS7_2421 chromosome XII	<i>Saccharomyces cer...</i>	3120	5137	100%	0.0	99.94%	1060541	CP093600.1	
<input checked="" type="checkbox"/>	Saccharomyces cerevisiae strain HLJ1 genome assembly_chromosome: 12	<i>Saccharomyces cer...</i>	3120	4637	100%	0.0	99.94%	1031746	LR813511.2	



- Click on alignment section set alignment view as dot plot.(See below)

Job Title PDC1 YLR044C SGDID:S000004034
RID ZRWJ3K09013 Search expires on 03-22 16:39 pm Download All
Program BLASTN Citation
Database nt See details
Query ID lclQuery_460151
Description PDC1 YLR044C SGDID:S000004034
Molecule type dna
Query Length 1692
Other reports Distance tree of results MSA viewer

Filter Results

Organism only top 20 will appear exclude
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity E value Query Coverage

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Alignment view Pairwise Pairwise
3 sequences selected
Query-anchored with dots for identities
Query-anchored with letters for identities
Flat query-anchored with dots for identities
Saccharomyces cerevisiae strain YPS128 chromosome XII sequence
Sequence ID: CP020219.1 Length: 1027157 Number of Matches: 2

Range 1: 227563 to 229254 GenBank Graphics ▾ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
3120 bits(1689)	0.0	1691/1692(99%)	0/1692(0%)	Plus/Minus

Query 1 ATGTCGAAATTACTTTGGTAATATTGTGCAAGGATAAAGCAAGTCACCGTAAAC 60
Sbjct 229254 ATGTCGAAATTACTTTGGTAATATTGTGCAAGGATAAAGCAAGTCACCGTAAAC 229195

Query 61 ACCGTTTCGGTTGCCCAAGGTACCTCAACTGTCTTGTGGACAAAGATCTACGAAGT 120
Sbjct 229194 ACCGTTTCGGTTGCCCAAGGTACCTCAACTGTCTTGTGGACAAAGATCTACGAAGT 229135

Query 121 GAAGGTATGAGATGGTGTGAAACGCCAAATTGAGCTTGCTACGGCCCTGATGGT 180

Feedback

- On choosing **Saccharomyces cerevisiae** strain **YPS128** chromosome **XII** sequence I get an alignment overview with the SNP position 1521

Range 2: 405513 to 407204 GenBank Graphics ▾ Next Match ▲ Previous Match ▲ First Match

Query	Subject	Sequence	Start	End
1081	Sbjct 228174	GCTTCTACCCCCATTGAAGCAAGAACGGATGGTGGAAACCAATTGGTAACCTCTTGCAAGAA	228175	1140
1141	Sbjct 228114	GGTGATGTTGCATTGCTGAAACCGGTACCTCCGCTTCGGTATCAACCAAACCACTTTC	228115	1200
1201	Sbjct 228054	CCAAACACACCTACGGTATCTCAAGCTTATGGGGTCCATTGGTTACCACTGGT	227995	1260
1261	Sbjct 227994	GCTACCTGGGTGCTGCTTCGCTGCTGAAGAAATTGATCCAAAGAGAGGTTATCTTA	227875	1320
1321	Sbjct 227934	TTCATTGGTACGGTTCTTGCATTGACTGTTCAAAGAAATCTCACCATGATCAGATGG	227875	1380
1381	Sbjct 227874	GGCTTGAAGCCATACTGTTGCTCTGAACAAACGATGGTACCACTTGAAGGGTTGATT	227815	1440
1441	Sbjct 227814	CACGGTCAAAGGCTAATACAACGAAATTCAAGGGGGACCACCTACCTTGGCCA	227755	1500
1501	Sbjct 227754	ACTTCGGTGCTAAGGACTATGAAACCCACAGAGTCGCTACCCGGTGAATGGGAAAG	227695	1560
1561	Sbjct 227694	TTGACCCAAGACAAGTCTTCAACGACACTCTAAGATGATTGAAATCATGGT	227635	1620
1621	Sbjct 227634	CCAGTCTTGGATGCTCCACAAAATTGGTTAACAAAGCTAAGTTGACTGCTGCTACCAAC	227575	1680
1681	Sbjct 227574	GCTAAGCAATAA 1692	227563	

Feedback

Range 2: 393754 to 395445 GenBank Graphics

[▼ Next Match](#) [▲ Previous Match](#) [▲ First Match](#)

Score 1967 bits(1065)	Expect 0.0	Identities 1492/1701(88%)	Gaps 18/1701(1%)	Strand Plus/Plus
Query 1 Sbjct 393754	ATGTCTGAAATTACTTGGTAAATATTGTCGAAAGATTAAAGCAAGTCACGTTAAC	60		
A.C.A.....A.T.....G.GC.....TG....	393813		
Query 61 Sbjct 393814	ACCGTTTCGGTTGCCAGGTGACTTCACCTGTC-CTTGGACAAGATCTACGAAGT	119		
C.....T.....T...-.....T..C.T.T....	393872		
Query 120 Sbjct 393873	TGAAGGTATGAGATGGCTGGTAACGCCAACGAATTGAACGCTGCTTACGCCGTGATGG	179		
	CA.....T.....T.....C.T.T.....	393932		
Query 180 Sbjct 393933	TTACGCTCGTATCAAGGGTATGCTGTATCATCACCACTCGGTGCGGTGAATTGTC	239		
C.....T.T.....T.....T.....	393992		
Query 240 Sbjct 393993	TGCTTGAACCGTATTGCCGTTCTACGCTGAACACGTCGGTGTGACCGTTGG	299		
T.....T.....	394052		
Query 300 Sbjct 394053	TGTCCCATCCATCTCGCTCAAGCTAACGAAATTGTTGTCACCAACACCTGGTAACGG	359		
	...T.....T.....T.....T.....	394112		
Query 360 Sbjct 394113	TGACTTCACTGTTTCCACAGAACATGCTGCCAACATTCTGAAACCACGTGCTATGATCAC	419		
C.....	394172		
Query 420 Sbjct 394173	TGACATTGCTACCGCCCCAGCTGAAATTGACAGATGTATCAGAACCACTACGTCACCCA	479		
	...T.....A.....T.....C.....ACT.....	394232		
Query 480 Sbjct 394233	AAGACCAAGTCTACTTAGGTTGCCAGCTAACCTGGTCGACTGAAACGTCAGCTAACATT	539		
G.....T.....C.....C.....	394292		
Query 540 Sbjct 394293	GTTGCAAACCTCAAATTGACATGTCTTGAAAGCCAACGATGCTGAATCCGAAAAGGAAGT	599		
	A.....G.....T.....C.....G.T.....GCT.....	394352		
Query 600 Sbjct 394353	CATT-GACACCATCTTGGCTTGGCAAGGATGCTAACGAAACCCAGTTATCTGGCTGATG	658		
	TG.....A.....TG.TG.T.AA.....A.....	394411		
Query 659 Sbjct 394412	CTTGTGTTCCAGACACGACGTCAGGCTGAAACTAAGAAGTTGATTGACTTGAATCAAT	718		
GC.....T.....T.....T.....G.....	394471		
Query 719 Sbjct 394472	TCCCAGCTTCGTCACCCAAATGGTAAGGGTCCATTGACGAACAAACCCAAGATAACG	778		
T.A.....G.T.....	394531		
Query 779 Sbjct 394532	GTGGTGTTCAGCTGGTACCTGTCCAAGCCAGAAGTTAAGGAAGCCGTTGAATCTGCTG	838		
T.....T.GA.....A.G.T.A.....	394591		

- For strain **Saccharomyces cerevisiae strain UWOPS87_2421**
chromosome XII SNP is 1521

PolyPhen-2: report for seq A14 | NCBI BlastNC_001144.5:23230 | +

blast.ncbi.nlm.nih.gov/Blast.cgi?2211401230

My Classroom | C... Gmail YouTube Maps Watch Pokemon Se... LinkedListAssignment TreesAssignment ChatGPT Kartik-Mathur/CP9... HashMapHeapsAss... DpAssignment RecursionAssignment

Sbjct 226534

Query 1441 CACGGTCCAAGGCTAACACGAAATTCAAGGTTGGACCAC

Sbjct 226474

Query 1501 ACTTCGGTGCTAAGGACTATGAAACCCACAGAGTCGCTACCAACG

Sbjct 226414

Query 1561 TTGACCAAGACAAGTCTTCAACGACAACCTAACAGTCAGAACATGA

Sbjct 226354

Query 1621 CCAGTCTCGATGCTCCACAAACTGGTTGAACAAGCTAACGTTGA

Sbjct 226294

Query 1681 GCTAACGAAATAA 1692 226223

Range 2: 404202 to 405893 GenBank Graphics

▼ Next Match ▲ Previous Match ▲ First Match

Score 2017 bits(1092) Expect 0.0 Identities 1501/1701(88%) Gaps 18/1701(1%) Strand Plus/Plus

Query 1 ATGTCTGAAATTACTTGGTAAATATTGTCGAAAGATTAAAGC.

Feedback

- For strain **Saccharomyces cerevisiae strain SK23 chromosome XII**
SNP is 592

Score	Expect	Identities	Gaps	Strand
3121 bits(1690)	0.0	1691/1692(99%)	0/1692(0%)	Plus/Minus
Query 1 Sbjct 226706		ATGTCGAAATTACTTTGGTAAATTTGTCGAAGATTAAGCAAGCTAACGTTAAC	68	226647
Query 61 Sbjct 226586		ACCGTTTCGGTTGCCAGGTGACTTCACCTTGCTTGGACAAGATCTACGAAGTT	128	226587
Query 121 Sbjct 226586		GAAGGTATGAGATGGGCTGGTAAAGCCACGAAATTGAAAGCTGCTTACGCCGCTGATGGT	180	226527
Query 181 Sbjct 226526		TACGCTCGTATCAAGGGTATGCTTGTATCATCACCACTTCGGTGCCTGGAATTGCT	240	226467
Query 241 Sbjct 226466		GCTTGAAACGGTATTGCCGGTCTTACGCTGAACACGTGGTTTGCACTGTTGGT	300	226407
Query 301 Sbjct 226406		GTCCCCATCCATCTGCTCAAGCTAAGCAATTGTTGCAACCACCTTGGTAACGGT	360	226347
Query 361 Sbjct 226346		GACTTCACTGTTTCCACAGAATGCTGCCAACATTCTGAACACACTGCTATGATCACT	420	226287
Query 421 Sbjct 226286		GACATTGCTACCCGCCAGCTGAATAAGCAGATGTAATCAGAACACCTTAGTCAACCAA	480	226227
Query 481 Sbjct 226226		AGACCAAGCTACTTAGGTTGCCAGCTAACTTGGTCGACTTGAAACGTCCCAGCTAAGTTG	540	226167
Query 541 Sbjct 226166		TTGCAAACTCCAATTGACATGCTTGTGAAGCCAACGATGCTGAATCCGAAAAGGAAGTC	600	226107
Query 601 Sbjct 22601		ATTGACCCATCTGGCTTGGCAAGGATGCTAAGAACCCAGTTCTGGCTGATGCT	660	22601

- Similarly following the same steps for PDC 5 we get SNP at
 - 879 for **Saccharomyces cerevisiae strain WI_S_OAKURA_4 chromosome XII sequence**
 - 1214 for **Saccharomyces cerevisiae strain WA_C_CODDINGTON_2 chromosome XII, partial sequence**
 - And many more
- Similarly following the same steps for PDC 6 we get SNP at
 - 1142 for **Saccharomyces cerevisiae strain L_1528 chromosome VII**
 - 1437 for **Saccharomyces cerevisiae YJM1356 chromosome VII genomic sequence**
 - And many more

EXPASY TOOL FOR TRANSLATION

- Use this link to open the page ([expassy link](#))

The screenshot shows the Expasy Translate tool interface. At the top, there are tabs for Home, Programmatic Access, and Contact. Below that is the title "Translate tool". A text area titled "DNA or RNA sequence" contains the nucleotide sequence: ATGCTTTCCATTCAAGGAAAGATAATATTGTC... Below this is a "Genetic codes" dropdown set to "Standard". To the right, there are "Output format" options: "Verbose: Met, Stop, spaces between residues" (unchecked), "Compact: M, -, no spaces" (checked), "Includes nucleotide sequence" (unchecked), and "Includes nucleotide sequence, no spaces" (unchecked). Under "DNA strands", both "forward" and "reverse" are checked. At the bottom are "reset" and "TRANSLATE!" buttons.

- Copy paste the nucleotide sequence in the box and press on translate button.
- The output will be the following where there would be multiple frames shown in red and we would have to chose

The screenshot shows the results of the translation. It displays two frames: "5'3' Frame 1" and "5'3' Frame 2". The sequence for Frame 1 is: MLSIQQRYNICLMAERHPKWTQLELAKWAYETFQLPKIPSQGTISRLLARKSTYMNCKEHEKDANRLRKPNNLLVRKILQEWSQSL... The sequence for Frame 2 is: CFPFSKDIIFV-WRRGTQSGRNLNWQNGLMRRSSCQKFHKPAQYRVWCQGNQLI-IVKS... Both sequences are partially redacted with a red box.

BIOPYTHON CODE FOR TRANSLATION

Reading the Fasta file and extracting the sequence from it

```
▶ Click here to ask Blackbox to help you code faster
def extractSeq(file_name):
    for seq_record in list(SeqIO.parse(file_name, "fasta"))[:5]:
```



```
        pdc_gene_seq = str(seq_record.seq)

    return pdc_gene_seq
```



```
□ Click here to ask Blackbox to help you code faster
def translate_sequence(sequence):
    seq = Seq(sequence)
    protein_sequence = seq.translate()
    return str(protein_sequence)

sequ = extractSeq("PDC5.fasta")
sequStrain = extractSeq("strain.fasta")
print(translate_sequence(sequ))
print(translate_sequence(sequStrain))
```

- Function `extractSeq(file_name)`:
 - Purpose: This function extracts DNA sequences from a FASTA file.
 - Parameters:
 - `file_name` (string): The name of the FASTA file containing DNA sequences.
 - Returns:
 - A string representing the DNA sequence extracted from the FASTA file.
 - Description:
 - Iterates over the first five sequence records in the FASTA file specified by `file_name`.
 - Converts the DNA sequence of each record to a string.
 - Returns the DNA sequence of the first record found.
- Function `translate_sequence(sequence)`:
 - Purpose: This function translates a DNA sequence into a protein sequence.

- Parameters:
 - `sequence` (`string`): The DNA sequence to be translated.
- Returns:
 - A string representing the translated protein sequence.
- Description:
 - Converts the input DNA sequence into a `Seq` object.
 - Utilizes Biopython's `translate()` function to translate the DNA sequence into a protein sequence.
 - Converts the translated protein sequence into a string and returns it.
- Main Code:
 - Calls `extractSeq("PDC5.fasta")` to extract a DNA sequence from the "PDC5.fasta" file.
 - Calls `translate_sequence()` to translate the extracted DNA sequence into a protein sequence.
 - Prints the translated protein sequence to the console.
- Overall Description:
 - The code snippet first extracts a DNA sequence from a FASTA file using the `extractSeq()` function.
 - It then translates the extracted DNA sequence into a protein sequence using the `translate_sequence()` function.
 - Finally, it prints the translated protein sequence to the console.
 - This code can be used to quickly extract and translate DNA sequences for further analysis in protein-related studies.

POLYPHEN PREDICTION TOOL

- Go to the PolyPhen-2 web server: [PolyPhen-2](#).
- Paste or upload the amino acid sequence of the protein you want to analyse.
- Ensure that your input sequence is in FASTA format and contains only the amino acid sequence (without headers or additional information).
- Choose the type of mutation you want to predict and its position. (Ex G -> A at 143 position)

Not secure genetics.bwh.harvard.edu/pph2/

PolyPhen-2 prediction of functional effects of human nsSNPs

Home About Help Downloads Batch query WHESS.db

PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Please, use the form below to submit your query.

21-Jun-2021: Server has been migrated to new hardware. Note, all queries were terminated and user sessions data discarded in the process, hence you will need to resubmit your query if affected. We apologize for the inconvenience caused.

Query Data

Protein or SNP Identifier	
Protein sequence in FASTA format	
Position	
Substitution	AA1 A R N D C E Q G H I L K M F P S T W Y V AA2 A R N D C E Q G H I L K M F P S T W Y V
Query description	

Submit Query | Clear | Check Status | Display advanced query options

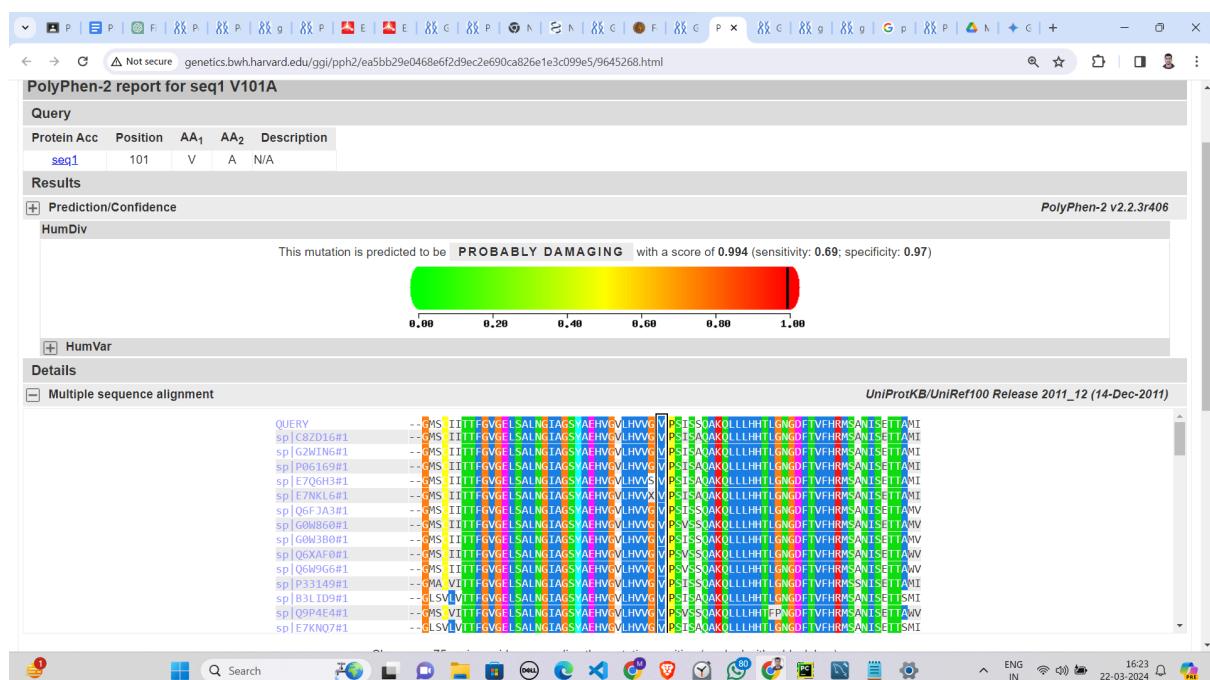
Software & web support: van adzhubey Web design & development: biobyte solutions



PREDICTIONS

1. Result 1

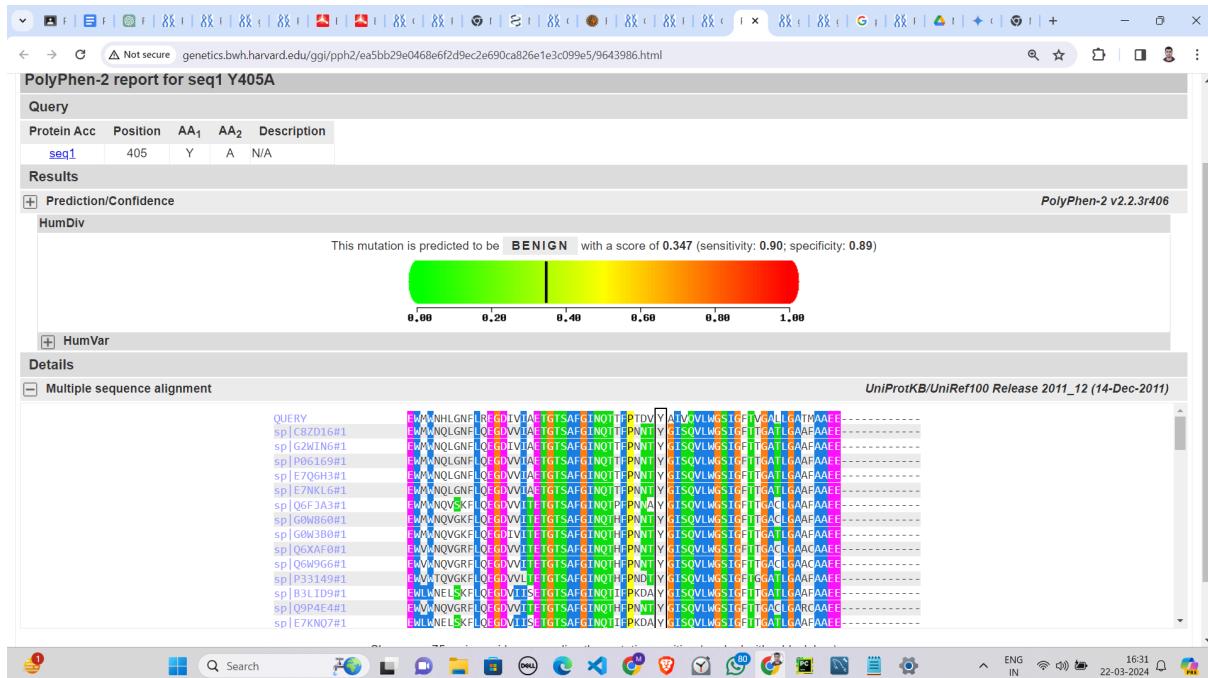
- a. **STRAIN USED:** Pyruvate decarboxylase 6 (PDC6) - *Saccharomyces cerevisiae*
- b. **POSITION:** 101
- c. **MUTATION:** Valine -> Alanine (V -> A)
- d. **PREDICTION:** From the report, the mutation seq1 V101A is labelled as "PROBABLY DAMAGING" with a score of 0.994 (sensitivity: 0.96, specificity: 0.97). This score is quite high, suggesting a high confidence in the prediction. The sensitivity and specificity values also indicate a high true positive rate and a low false positive rate, respectively.



2. Result 2

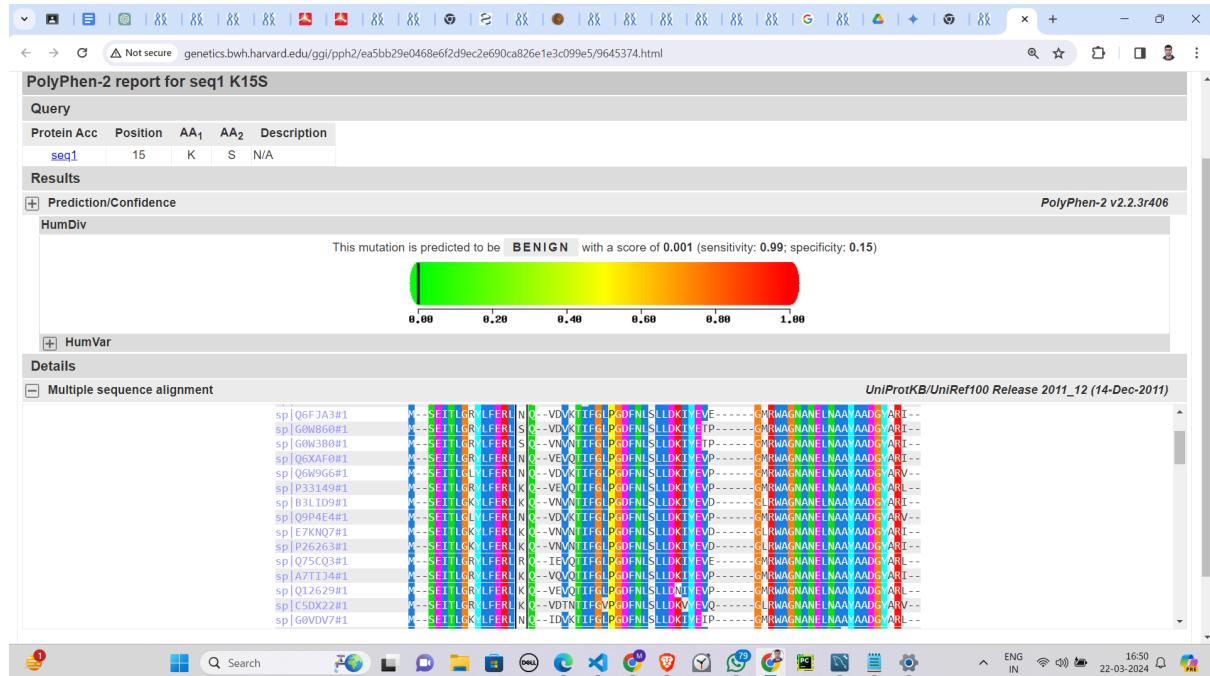
- a. **STRAIN USED:** Pyruvate decarboxylase 5 (PDC5) - *Saccharomyces cerevisiae*
- b. **POSITION:** 406
- c. **MUTATION:** Tyrosine → Alanine (Y → A)
- d. **PREDICTION:** This report indicates that a specific mutation (seq1 Y405A) is predicted to be "BENIGN" with a score of 0.347. The sensitivity is 0.90 and specificity is 0.89, suggesting high accuracy in the prediction model. The confidence in this prediction is visualised on a gradient scale from green (benign) to red (probably damaging), with this particular mutation falling into the green zone.

Based on this information, one can predict that this particular protein mutation (Y405A) would not have damaging effects or contribute to disease development due to its benign nature as indicated by the low score of 0.347 on PolyPhen-2's scale.



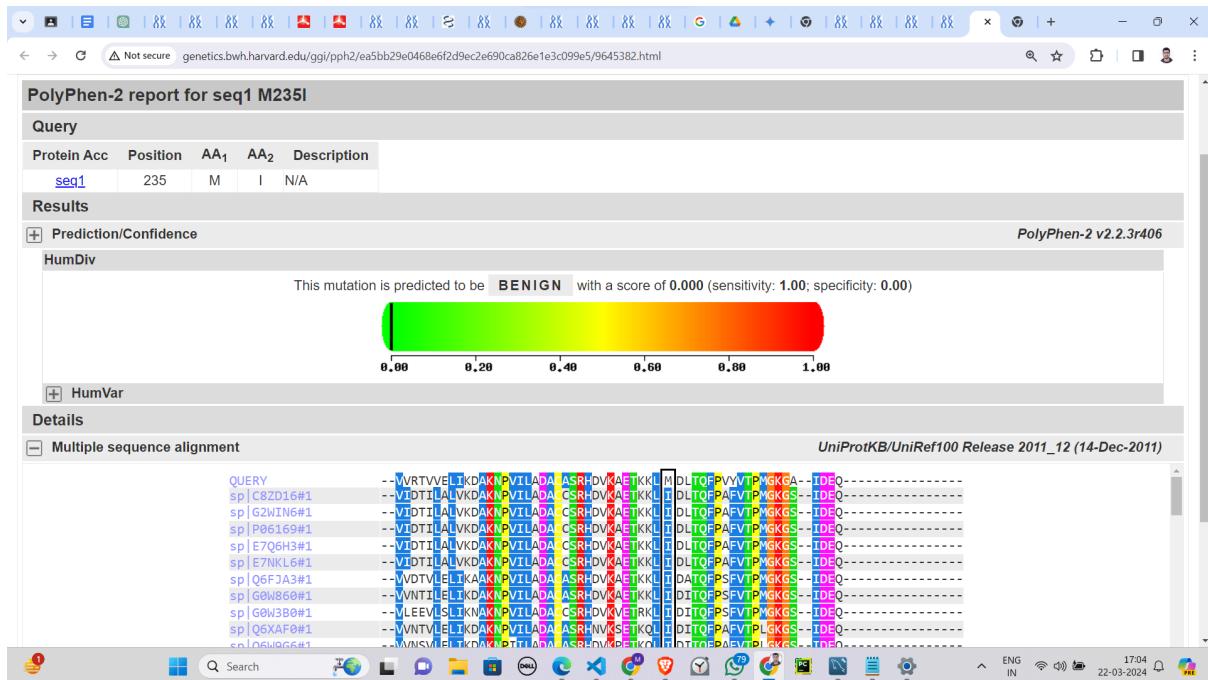
3. Result 3

- a. **STRAIN USED:** Pyruvate decarboxylase 1 (PDC1) - *Saccharomyces cerevisiae*
- b. **POSITION:** 15
- c. **MUTATION:** Lysine → Serine (K → S)
- d. **PREDICTION:** The report you've shared indicates that a specific mutation (seq1 K15S) is predicted to be "BENIGN" with a low score of 0.001 and sensitivity and specificity as 0.99 and 0.15 respectively. In the context of PolyPhen-2, a "benign" prediction suggests that the mutation is not expected to have a significant impact on the protein's function. The score, which ranges from 0 (benign) to 1 (probably damaging), provides a measure of confidence in this prediction. A score of 0.001 is very close to 0, indicating a high confidence that the mutation is benign



4. Result 4

- a. **STRAIN USED:** Pyruvate decarboxylase 5 (PDC5) - *Saccharomyces cerevisiae*
- b. **POSITION:** 235
- c. **MUTATION:** Methionine-> Isoleucine (M -> I)
- d. **PREDICTION:** This report indicates that a specific mutation (seq1 M235I) is predicted to be "BENIGN" with a score of 0.000. The sensitivity is 1 and specificity is 0. As it is labelled as "benign" and the score is 0.000, it suggests that this mutation might not have significant implications on the protein's function.



CITATIONS

1. Effects of pyruvate decarboxylase (pdc 1, pdc 5) gene knockout on the production of metabolites in two haploid *Saccharomyces cerevisiae* strains
Zhang W; Kang J; Wang C; Ping W; Ge J;
<https://pubmed.ncbi.nlm.nih.gov/33881948/>
2. Development of Single Nucleotide Polymorphism (SNP)-Based Triplex PCR Marker for Serotype-specific *ⁱEscherichia coli* Detection
Rahman et al.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8874422/>

3. PolyPhen-2 Tutorials

The Jackson Laboratory is an independent, nonprofit organization focusing on mammalian genetics research to advance human health. Our mission is to discover precise genomic solutions for disease and empower the global biomedical community in the shared quest to improve human health.

https://www.jax.org/-/media/jaxweb/files/education-and-learning/tutorials/polyphen2_tutorials_clickable.pdf

4. ExPASy: The proteomics server for in-depth protein knowledge and analysis

Gasteiger et al.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC168970/>

