Glutenin Gene and Celiac Disease

Introduction:

Glutenin is a gluten protein that is a staple component in food worldwide across every culture. It acts as a key genetic determinant responsible for encoding proteins that contribute to the structural integrity and functionality of gluten. It's derived from wheat and related grains, which determines the quality of various wheat-based products [11]. It is an essential protein that is a key contributor to a piece of dough's viscoelastic properties which ultimately affect the outcomes of baked goods by influencing their texture, taste, water content, and structure. The significance of understanding glutenin lies in its implications for both agricultural practices and the food industry as it is incorporated into many meals, including most processed foods. Considering that the global demand for wheat-based products is increasing over time, the quality and composition of glutenin have become a center of focus [4,7]. As the population continues to grow, addressing the challenges and optimizing the properties of glutenin in wheat varieties becomes essential for ensuring the production of high-quality and nutritionally valuable food items. However, there is a growing need for innovating the molecule to change its functionality in the context of processed foods as many individuals suffer from a certain degree of gluten intolerance [4,7,11]. Celiac disease is a chronic autoimmune disorder that is triggered by the consumption of that affects 1% of the world population, inhibiting them from consuming gluten-containing products without consequences as demonstrated in Figure 1. The abnormal immune mistakes a specific component of gluten called the gliadin fraction as a harmful pathogen and activates an immune response against it. This immune response leads to inflammation and damage to the lining of the small intestine and over time, this damage can result in a range of gastrointestinal and systemic symptoms, including diarrhea, abdominal pain, malnutrition, and various other health problems. Additionally, a metanalysis study showcased that the proportion of individuals consuming gluten diets amongst those who are diagnosed, undiagnosed, and not diagnosed with Celiac disease increased from 5%, 51%, and 44% respectively in 2009-2010 to 16%,12%, and 72% in 2013-2014 as shown in Figure 2 [10]. By better understanding its molecular structure, we can create substitutes or treatments that do not detract from the regular Glutenin qualities while negativing undesirable consequences.

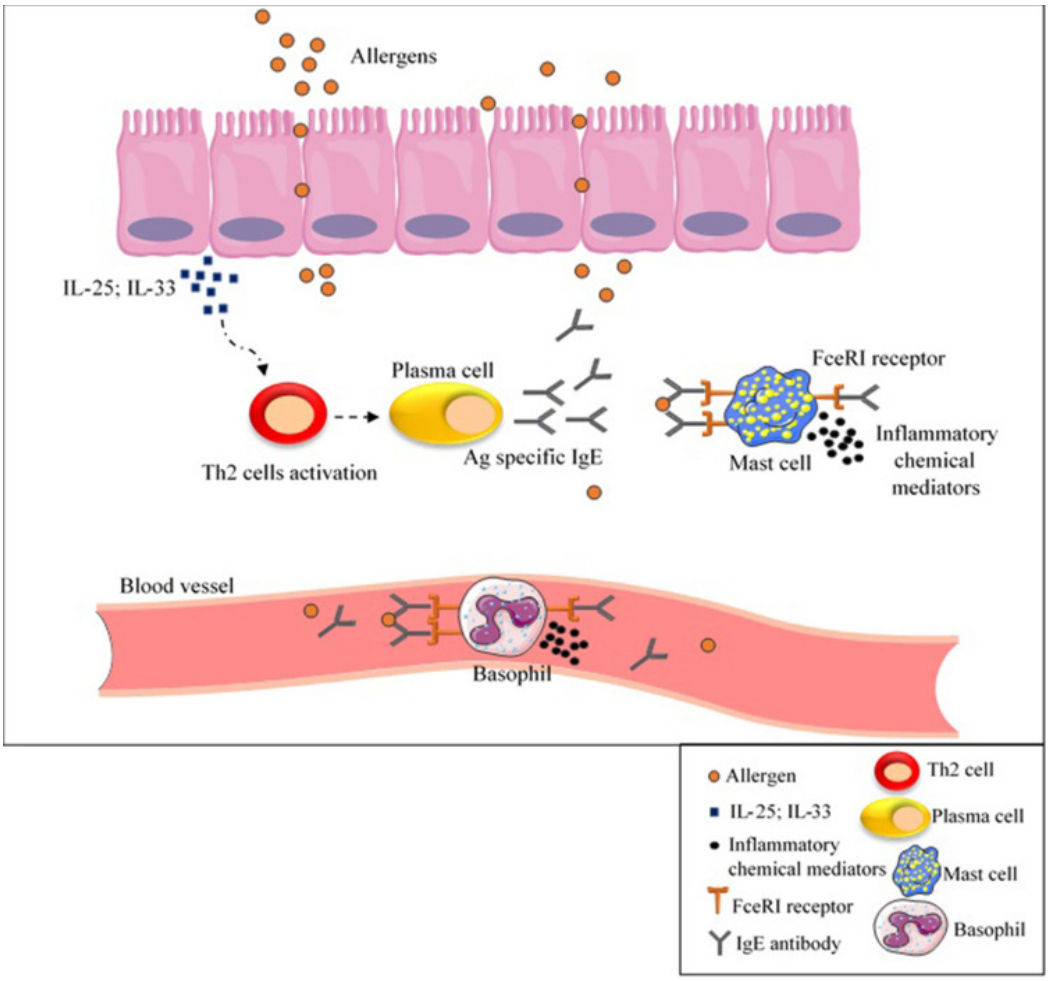


Figure 1: Cellular level depiction of the inflammatory auto immune response when gluten is consumed by a Gluten intolerant individual

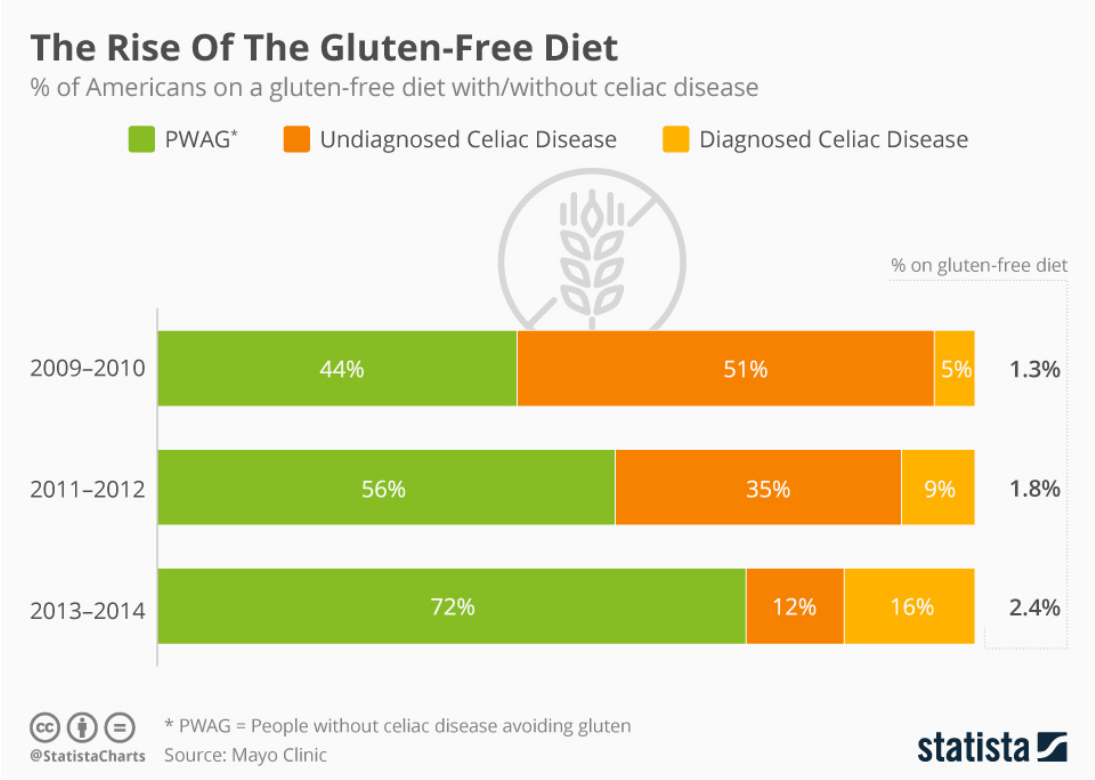


Figure 2: Incidences of Gluten-free diet across populations with, without, and undiagnosed with celiac disease from 2009-2014

I. Protein Sequence

Using the Uniprot website, https://www.uniprot.org, you can query for various relevant pieces of information for your target protein through various means including just typing the name of the compound such as Glutenin. You'll be redirected to an array of relevant proteins that you can choose from that give some high-level information such as the Protein name, gene name, and what organism it derives from shown in Figure 3. When selecting a specific protein, you can access various biological and chemically relevant information such as the function, amino acid size, and amino acid sequence etc as shown in Figures 4 and 5.

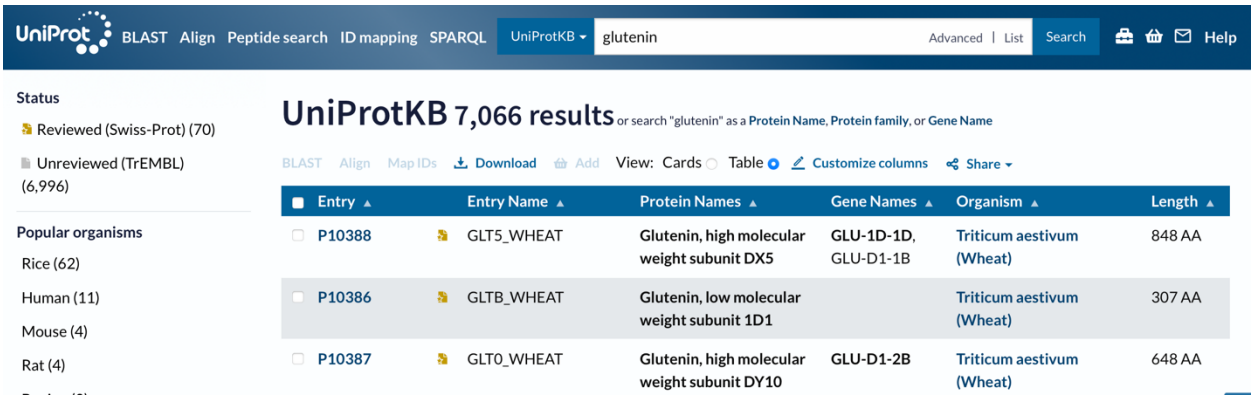


Figure 3: Uniprot search results for Glutenin

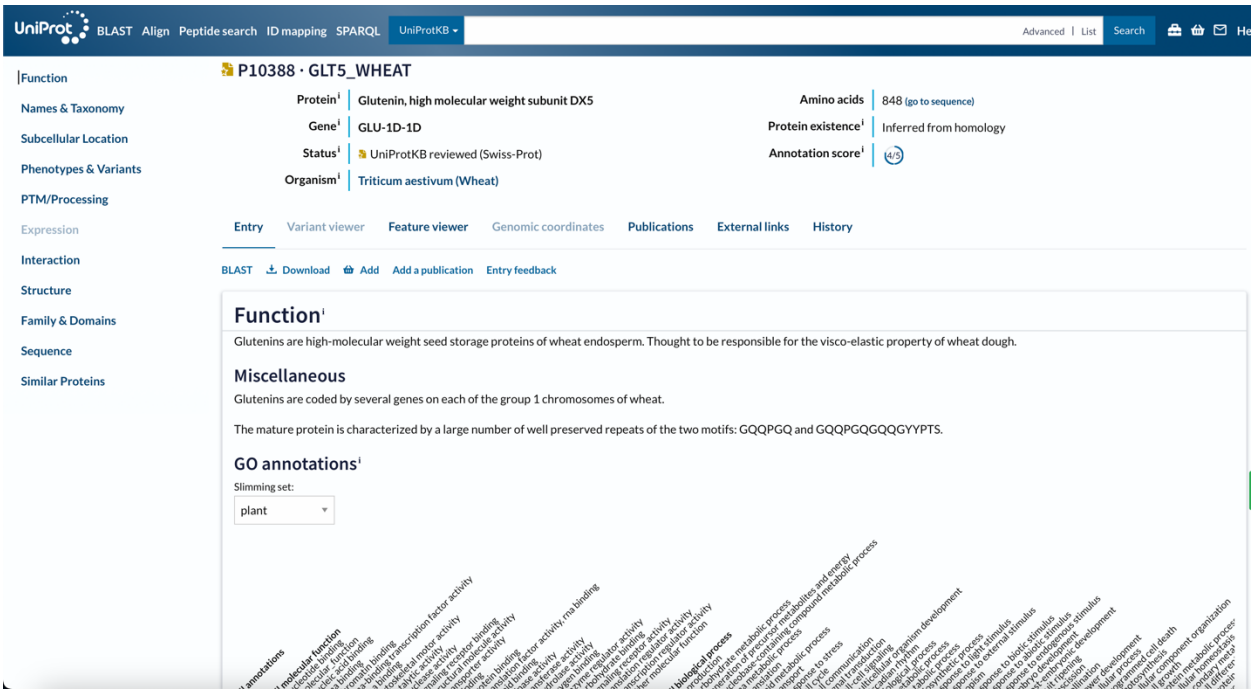
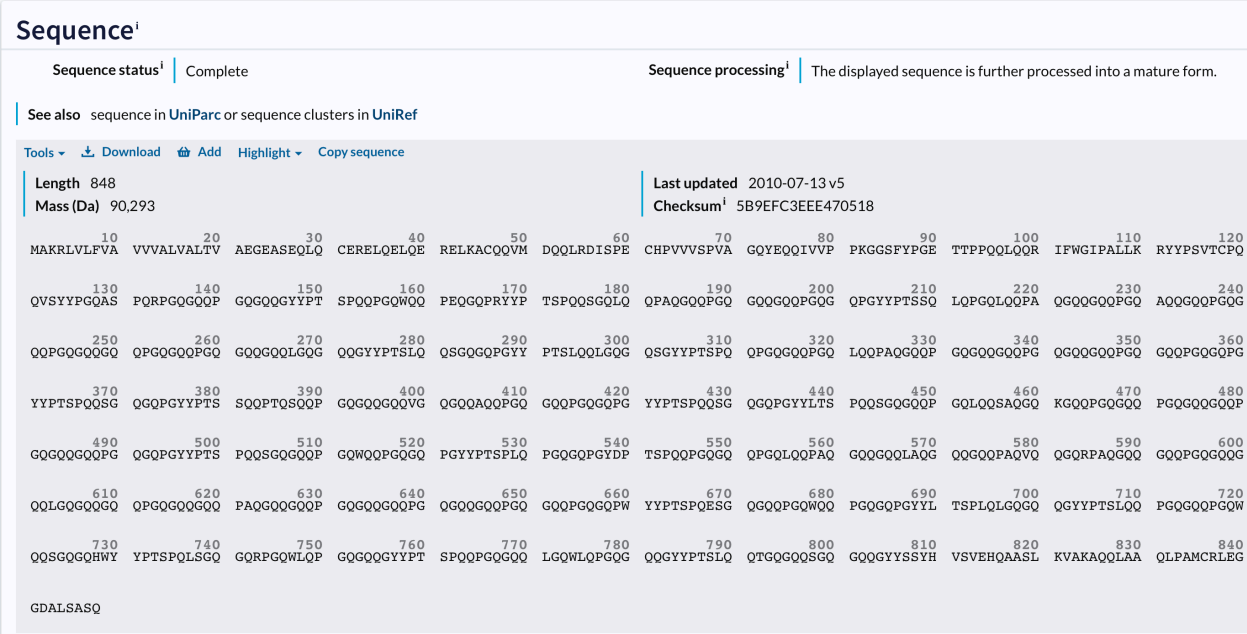


Figure 4: Uniprot molecular search for Glutenin P10388



Sequence[i]

Sequence status[i] | Complete

Sequence processing[i] | The displayed sequence is further processed into a mature form.

See also sequence in UniParc or sequence clusters in UniRef

Tools ▾ Download Add Highlight ▾ Copy sequence

Length 848
Mass (Da) 90,293

Last updated 2010-07-13 v5
Checksum[i] 5B9EFC3EEE470518

```
        10         20         30         40         50         60         70         80         90        100        110        120
MAKRLVLFVA VVVALVALTV AEGEASEQLQ CERELQELQE RELKACQQVM DQQLRDISPE CHPVVVSPVA GQYEQQIVVP PKGGSFYPGE TTPPQQLQQR IFWGIPALLK RYYPSVTCPQ

       130        140        150        160        170        180        190        200        210        220        230        240
QVSYYPGQAS PQRPGQGQQP GQGQQGYYPT SPQQPGQWQQ PEQGQPRYYP TSPQQSGQLQ QPAQGQQPGQ GQQGQQPGQG QPGYYPTSSQ LQPGQLQQPA QGQQGQQPGQ AQQGQQPGQG

       250        260        270        280        290        300        310        320        330        340        350        360
QQPGQGQQGQ QPGQGQQPGQ GQQGQQLGQG QQGYYPTSLQ QSGQGQPGYY PTSLQQLGQG QSGYYPTSPQ QPGQGQQPGQ LQQPAQGQQP GQGQQGQQPG QGQQGQQPGQ GQQPGQGQPG

       370        380        390        400        410        420        430        440        450        460        470        480
YYPTSPQQSG QGQPGYYPTS SQQPTQSQQP GQGQQGQQVG QGQQAQQPGQ GQQPGQGQPG YYPTSPQQSG QGQPGYYLTS PQQSGQGQQP GQLQQSAQGQ KGQQPGQGQQ PGQGQQGQQP

       490        500        510        520        530        540        550        560        570        580        590        600
GQGQQGQQPG QGQPGYYPTS PQQSGQGQQP GQWQQPGQGQ PGYYPTSPLQ PGQGQPGYDP TSPQQPGQGQ QPGQLQQPAQ GQQGQQLAQG QQGQQPAQVQ QGQRPAQGQQ GQQPGQGQQG

       610        620        630        640        650        660        670        680        690        700        710        720
QQLGQGQQGQ QPGQGQQGQQ PAQGQQGQQP GQGQQGQQPG QGQQGQQPGQ GQQPGQGQPW YYPTSPQESG QGQQPGQWQQ PGQGQPGYYL TSPLQLGQGQ QGYYPTSLQQ PGQGQQPGQW

       730        740        750        760        770        780        790        800        810        820        830        840
QQSGQGQHWY YPTSPQLSGQ GQRPGQWLQP GQGQQGYYPT SPQQPGQGQQ LGQWLQPGQG QQGYYPTSLQ QTGQGQQSGQ GQQGYYSSYH VSVEHQAASL KVAKAQQLAA QLPAMCRLEG

GDALSASQ
```

Figure 5: Uniprot amino acid sequence for Glutenin P10388

II. Protein Secondary and 3D Structures:

Deriving characteristics and functionality of a protein can be done using its amino acid to calculate relevant values such as the isoelectric point, which is the pH at which a protein has a net zero charge, as well as construct its higher-level structures such as its secondary and tertiary structure. We can utilize the calculator from the following website: http://isoelectric.org/calculate.php in order to calculate the isoelectric point for a given amino acid sequence which for Glutenin is 6.16 as shown in Figure 6. The figure showcases various methods for pI Estimation which specify the calculation method. It additionally showcases how the charge of the protein changes at different pH levels such as having a pH of 5.5, but when in a lysosomes environment, the charge is +2.5, indicating that the protein is positively charged at this pH. In order to construct the secondary structure, we can utilize the following website: https://www.compbio.dundee.ac.uk/jpred/, as shown in Figure 7. The majority of secondary structure is shown as coil and extracellular matrix.

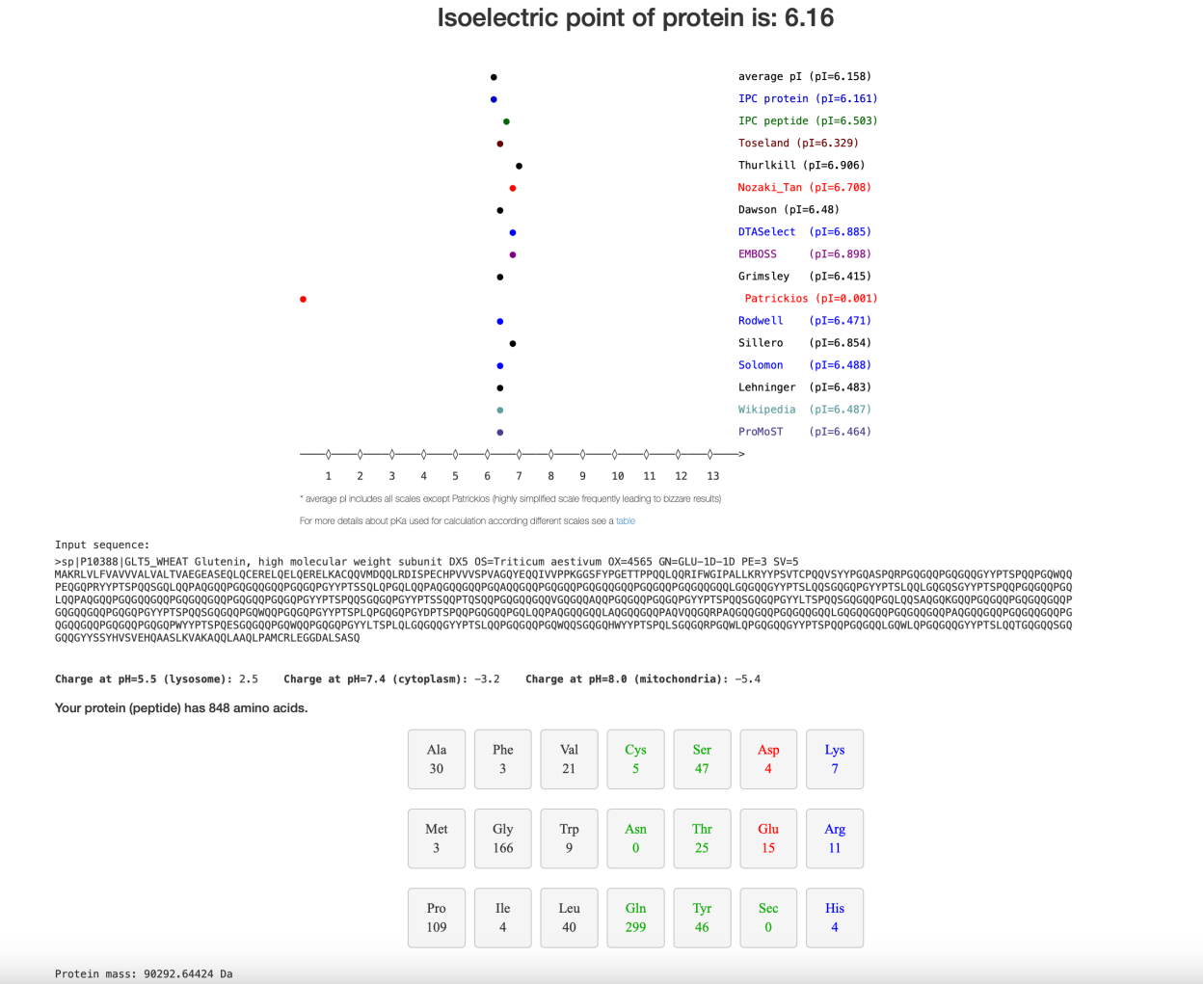


Figure 6: Isoelectric point of Glutenin

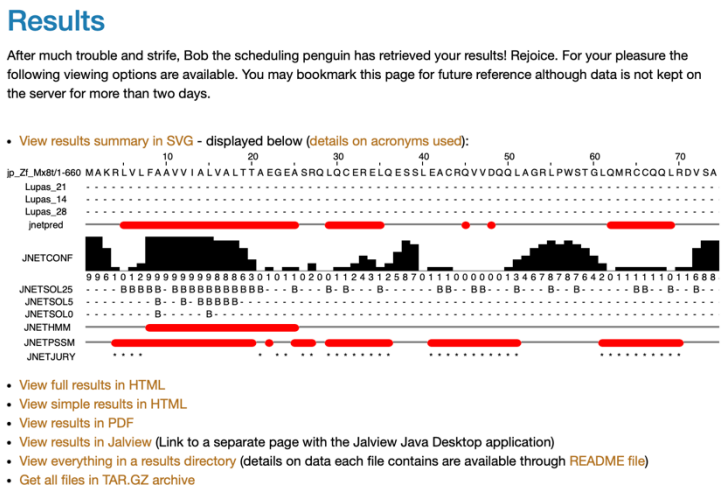


Figure 7: Secondary structure of Glutenin

The Protein 3D structure was developed through the use of the AlphaFold model which was created from the uniport website: https://www.uniprot.org which can be seen in Figure 8. The 3D shape, emphasizes an entirely coiled alpha helices. These helical structures are interspersed among the random coils and loops, which serve as a flexible connector between structured regions. There is a distinct multi-colored segment in the protein's center which may suggest that the region is of importance. In particular, it could be a binding site for Tissue Transglutaminase (tTG) to modify the gluten peptide to increase the affinity for HLA-DQ2 or HLA-DQ8 which are the human leukocyte antigen that can produce proinflammatory cytokines like interferon-gamma.

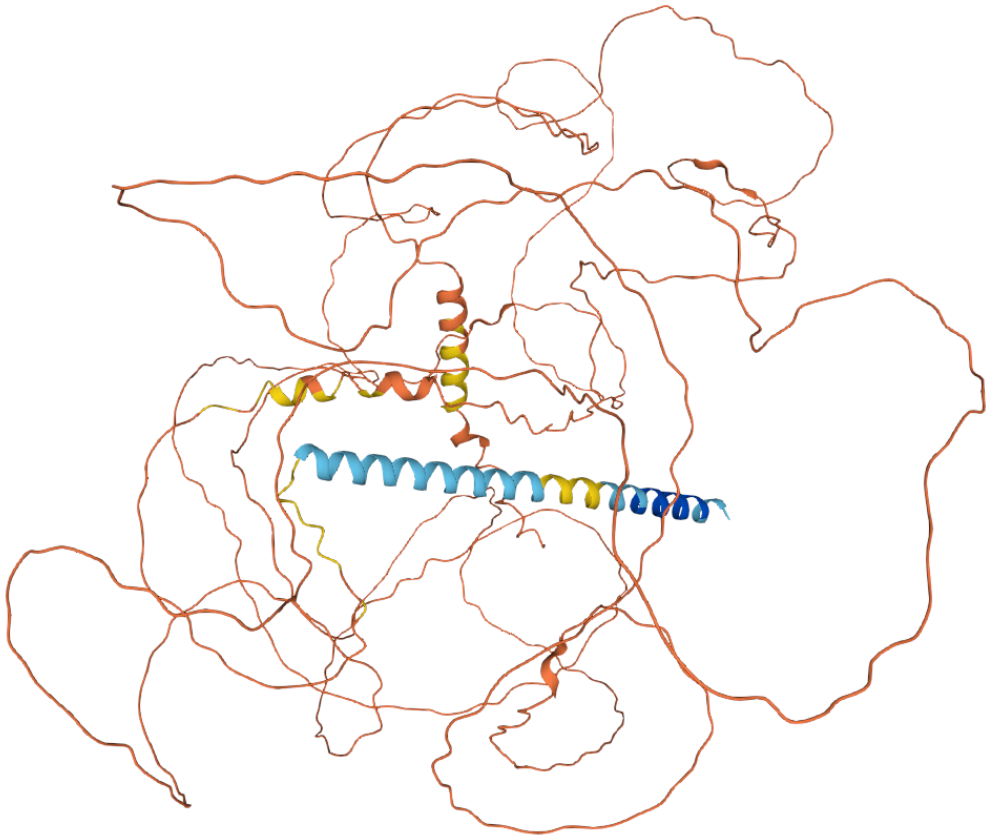


Figure 8: Tertiary structure of Glutenin

III. Multiple Protein Sequence Alignment

By utilizing the European Bioinformatics Institutes Clustal Omega tool (https://www.ebi.ac.uk/Tools/msa/clustalo/), several (~10) similar glutenin proteins are utilized in order to create a multiple protein sequence alignment to align all of these sequences and identify regions of similarity that could reveal functional, structural, or evolutionary relationships amongst the dataset. There are several conserved regions that have no change whatsoever in the specific region across all samples as well as some semi and non-conserved regions shown in Figures 9 and 10. We can additionally compare their evolutionary lineage by creating a phylogenetic tree shown in Figure 11 to showcase divergences that occurred over time and how those regions that mutated aren't critical to survival or are utilized to boost survival by allowing for a competitive advantage.

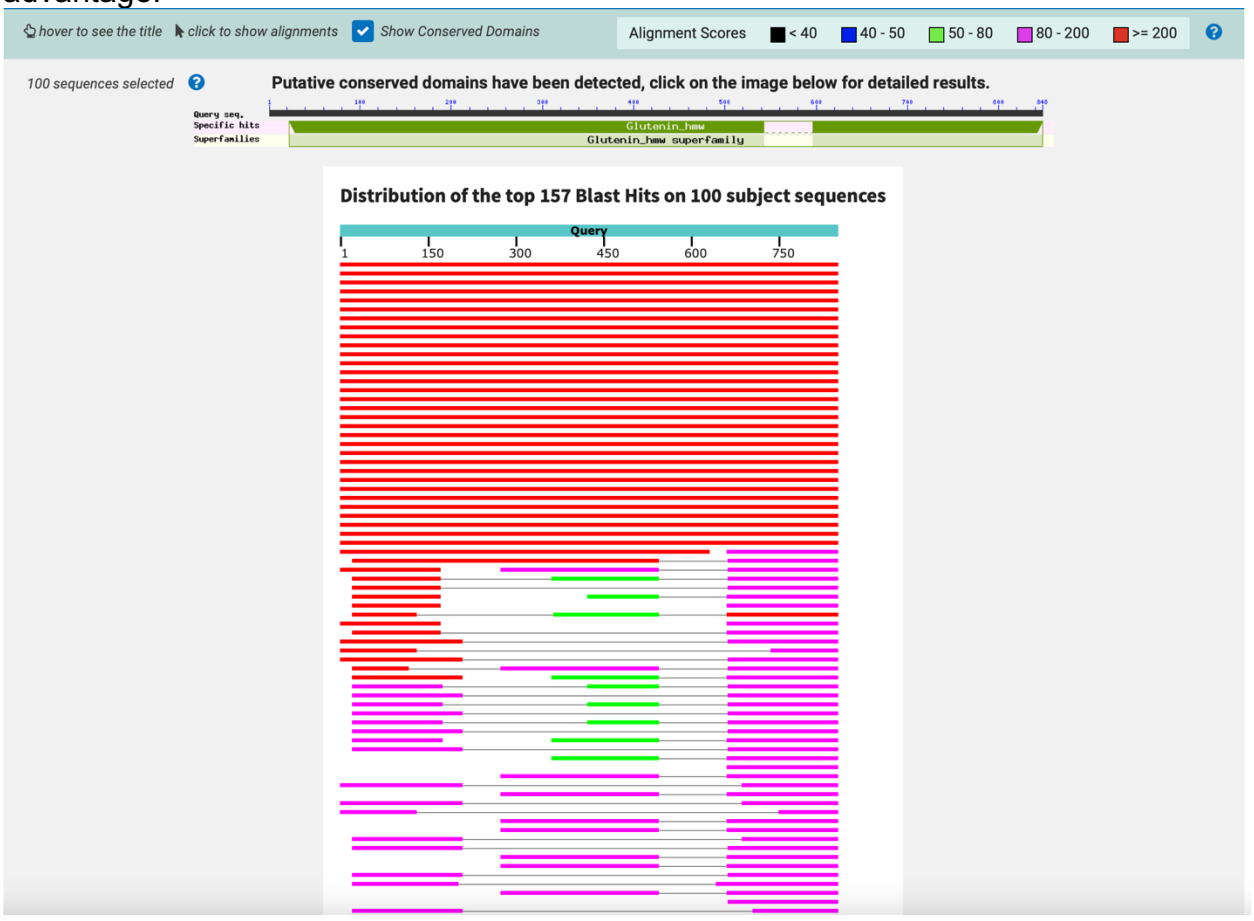


Figure 9: Conserved Domains in Glutenin Sequence Alignment

Figure 10: Multiple protein sequence alignment for 10 different Glutenin genes

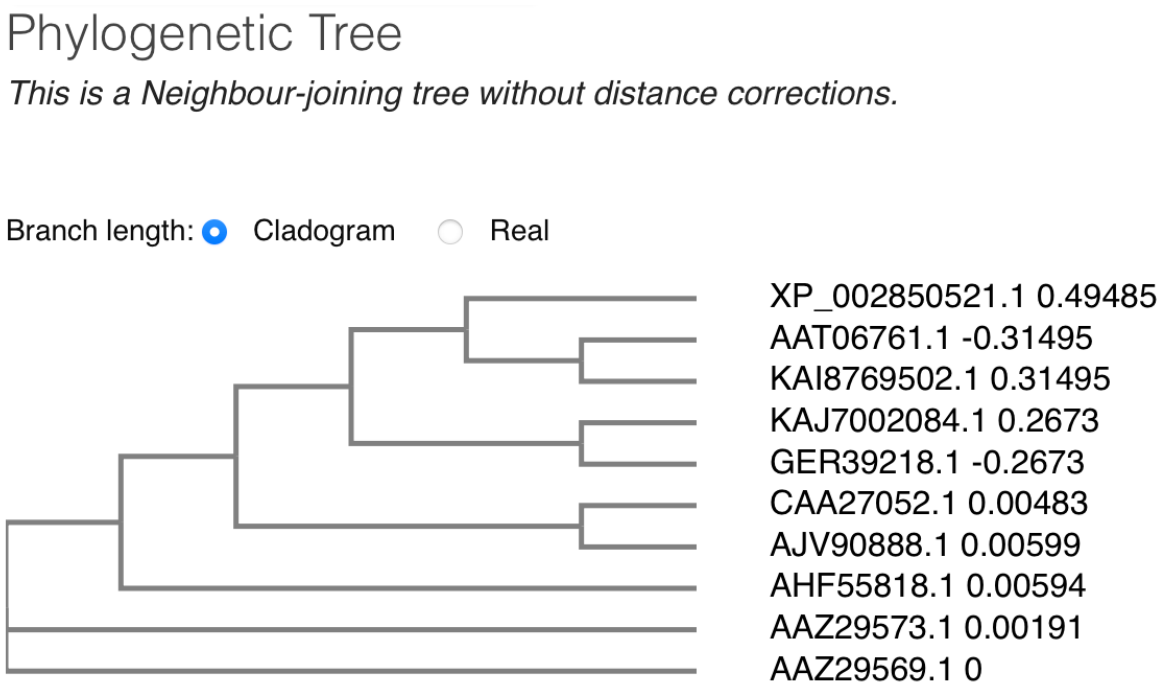


Figure 11: Phylogenetic tree to showcase evolutionary relationship of select Glutenin sequences

IV. Design and Discovery of Genetic Markets

a. Gene Querying

By utilizing the UCSC genome browser to search for the Glutenin gene, we can find the exact gene location on the chromosome which in this case is chr17:66798688-66799422, and can be visualized in Figure 12. The figure illustrates the position and distribution of Gliadin and Glutenin genes across several specific wheat chromosomes. These chromosomes are labeled as 1A, 1B, 1D, 6A, 6B, and 6D and the ω, γ, and α symbols represent different types of gliadins, with ω-gliadins located on Chromosome 1, γ-gliadins spread across Chromosomes 1A, 1B, and 1D, and α-gliadins found on Chromosome 6. Glutenin is categorized as High Molecular Weight (HMW) and Low

Molecular Weight (LMW), with HMW-glutenins present on Chromosome 1.

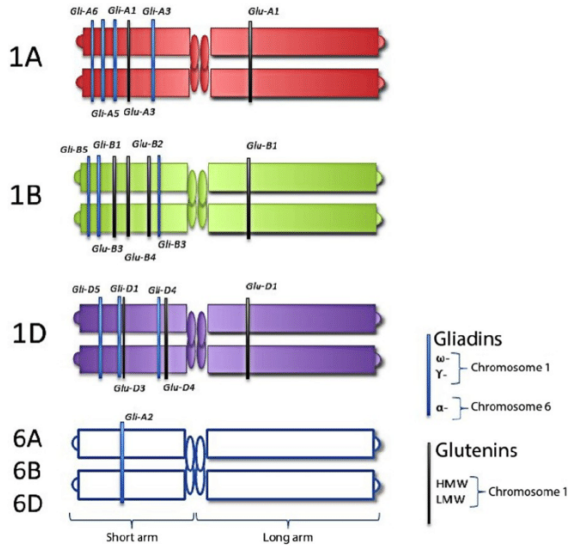


Figure 12: Labeled Chromosome which showcases positions of Gliadins and Glutenins

b. SNP searching

By utilizing the UCSC Genome Browser, Glutenin's SNPs can be identified by searching for it by name from the following link: https://genome.ucsc.edu/cgi-bin/hgGateway. The results can be showcased in Figure 13. We can infer a lot of additional information about this protein from this data including the genomic region it spans (about 7,184,400 bases to 7,184,500 bases, the several genetic variants with both ClinVar and dbSNP references, OMIM Allelic Variants and Phenotypes that display phenotypic information associated with certain alleles with distinction between pathogenic (red) and non-pathogenic (green) variants, the gene expression in 54 tissues from GTEx, as well as conservation across 100 vertebrates, including several mammalian species like humans, mice, and dogs. There is a wide variety of SNPs to choose from towards the bottom of the visualization that illuminates the specific mutation that occurred.

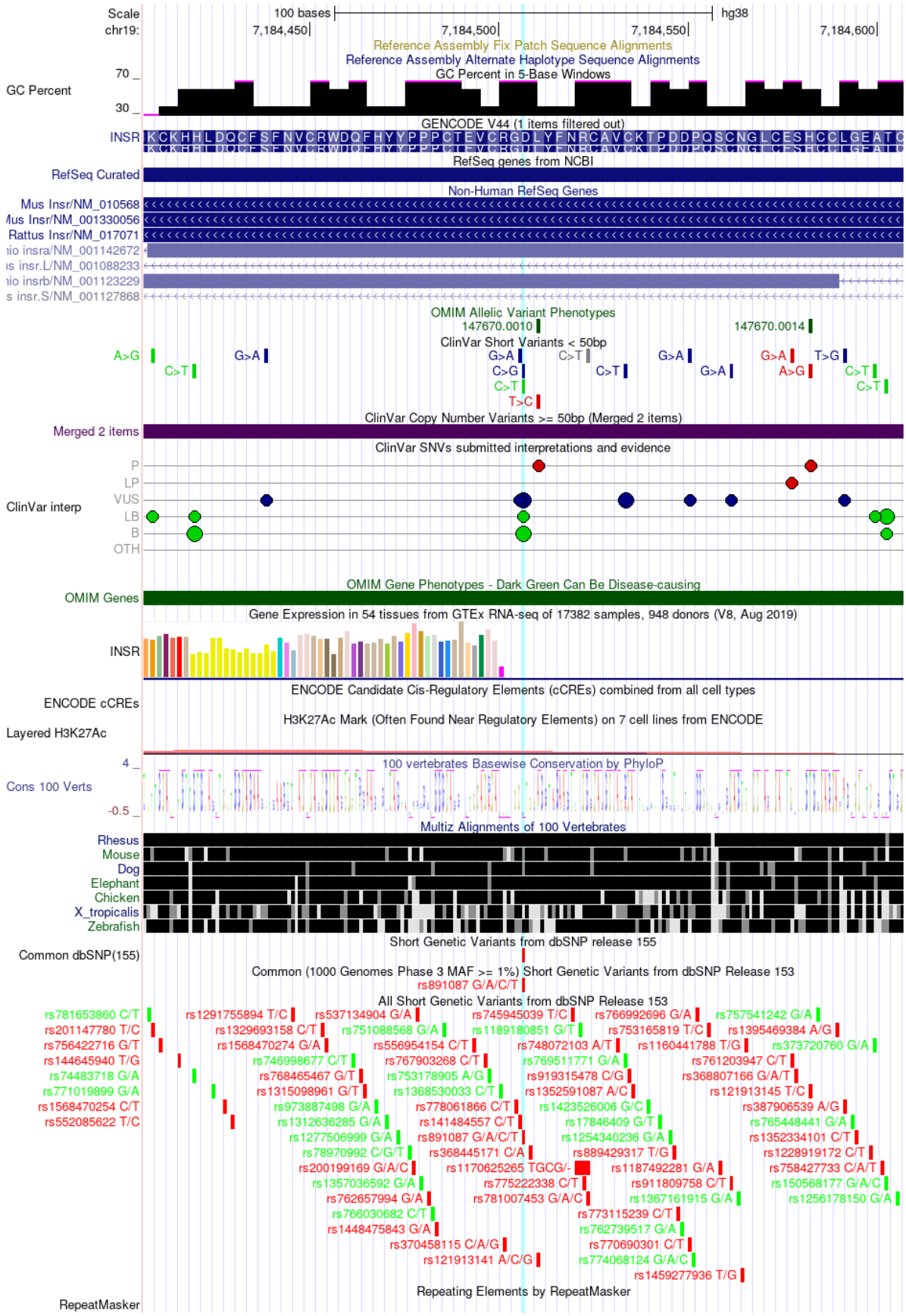

Scale
chr19:
100 bases
hg38
7,184,450
7,184,500
7,184,550
7,184,600
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GC Percent
GC Percent in 5-Base Windows
70
30
INSR
GENCODE V44 (1 items filtered out)
RefSeq Curated
RefSeq genes from NCBI
Non-Human RefSeq Genes
Mus Insr/NM_010568
Mus Insr/NM_001330056
Rattus Insr/NM_017071
Rattus insra/NM_001142672
Rattus insr.L/NM_001088233
Rattus insrb/NM_001123229
Rattus insr.S/NM_001127868
OMIM Allelic Variant Phenotypes
147670.0010
147670.0014
ClinVar Short Variants < 50bp
A>G
G>A
C>T
C>G
T>C
T>G
A>G
ClinVar Copy Number Variants >= 50bp (Merged 2 items)
Merged 2 items
ClinVar SNVs submitted interpretations and evidence
ClinVar interp
P
LP
VUS
LB
B
OTH
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing
OMIM Genes
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
INSR
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
100 vertebrates Basewise Conservation by PhyloP
4
-0.5
Cons 100 Verts
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Short Genetic Variants from dbSNP release 155
Common dbSNP(155)
Common (1000 Genomes Phase 3 MAF >= 1%) Short Genetic Variants from dbSNP Release 153
rs891087 G/A/C/T
All Short Genetic Variants from dbSNP Release 153
rs781653860 C/T
rs1291755894 T/C
rs537134904 G/A
rs745945039 T/C
rs766992696 G/A
rs757541242 G/A
rs201147780 T/C
rs1329693158 C/T
rs751088568 G/A
rs1189180851 G/T
rs753165819 T/C
rs1395469384 A/G
rs756422716 G/T
rs1568470274 G/A
rs556954154 C/T
rs748072103 A/T
rs1160441788 T/G
rs373720760 G/A
rs144645940 T/G
rs746998677 C/T
rs767903268 C/T
rs769511771 G/A
rs761203947 C/T
rs74483718 G/A
rs768465467 G/T
rs753178905 A/G
rs919315478 C/G
rs368807166 G/A/T
rs771019899 G/A
rs1315098961 G/T
rs1368530033 C/T
rs1352591087 A/C
rs121913145 T/C
rs1568470254 C/T
rs973887498 G/A
rs778061866 C/T
rs1423526006 G/C
rs387906539 A/G
rs552085622 T/C
rs1312636285 G/A
rs141484557 C/T
rs17846409 G/T
rs765448441 G/A
rs1277506999 G/A
rs891087 G/A/C/T
rs1254340236 G/A
rs1352334101 C/T
rs78970992 C/G/T
rs368445171 C/A
rs889429317 T/G
rs1228919172 C/T
rs200199169 G/A/C
rs1170625265 TGCG/-
rs1187492281 G/A
rs758427733 C/A/T
rs1357036592 G/A
rs775222338 C/T
rs911809758 C/T
rs150568177 G/A/C
rs762657994 G/A
rs781007453 G/A/C
rs1367161915 G/A
rs1256178150 G/A
rs766030682 C/T
rs773115239 C/T
rs1448475843 G/A
rs762739517 G/A
rs370458115 C/A/G
rs770690301 C/T
rs121913141 A/C/G
rs774068124 G/A/C
rs1459277936 T/G
Repeating Elements by RepeatMasker
RepeatMasker

Figure 13: UCSC Genome Browser generation for Glutenin to search for SNPs (Bottom of figure in green and red)

c. Primer Design

Primers can be designed by utilizing the previously discovered SNPs and utilizing the NCBI-Primer Blast functionality using a FASTA sequence and the following web link: https://www.ncbi.nlm.nih.gov/tools/primer-blast/. These primers are short and specific single-stranded DNA sequences that are a complement to the target DNA strand you're interested in to amplify specific DNA segments. By utilizing the SNP rs377244762, the following primer pairs were generated in Figures 14 and 15.

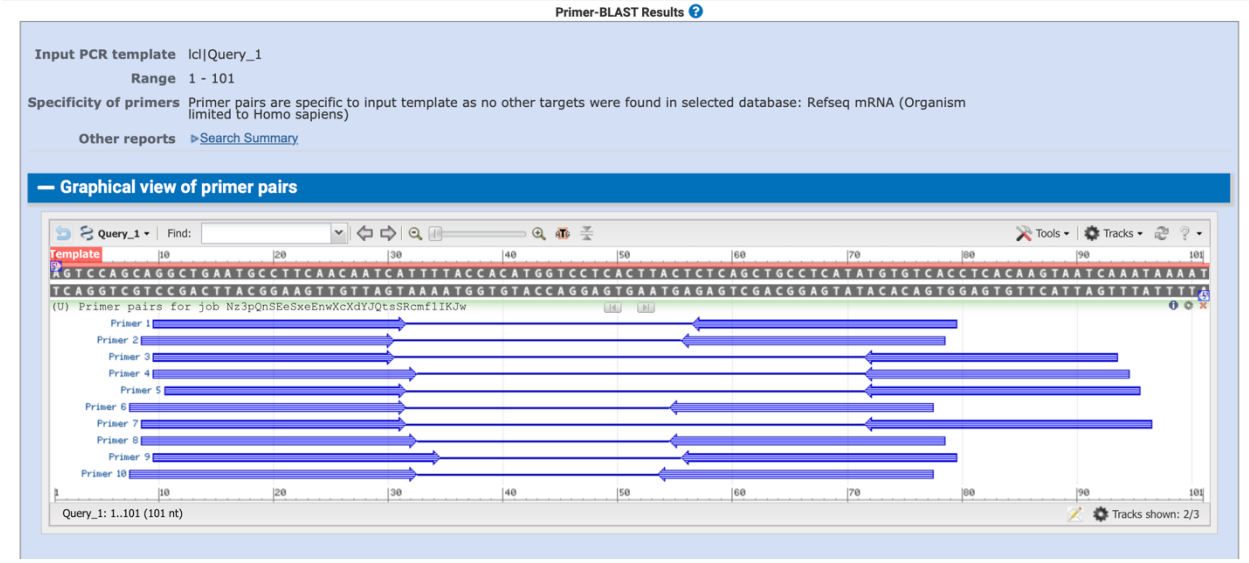
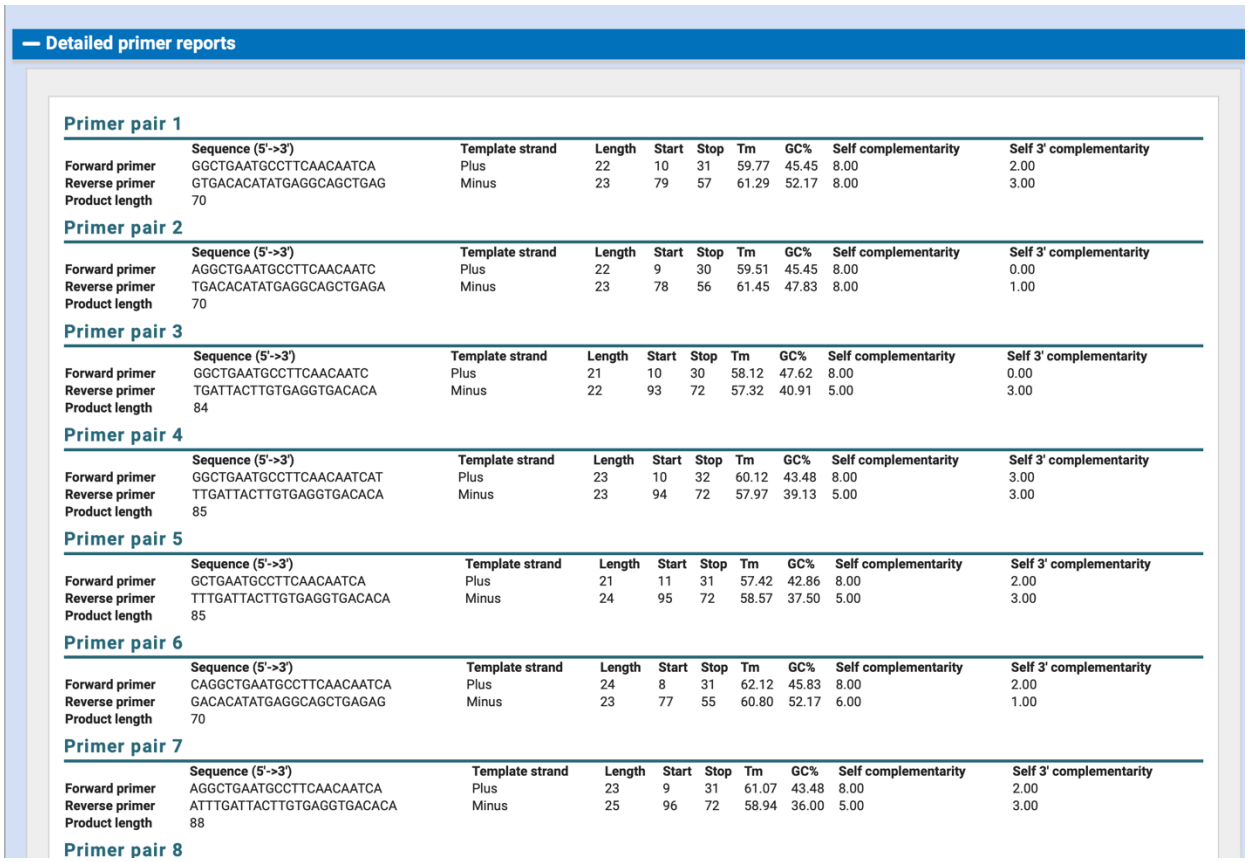


Figure 14: Primer pairs generated for SNP rs377244762



— Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GGCTGAATGCCTTCAACAATCA	Plus	22	10	31	59.77	45.45	8.00	2.00
Reverse primer	GTGACACATATGAGGCAGCTGAG	Minus	23	79	57	61.29	52.17	8.00	3.00
Product length	70								

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AGGCTGAATGCCTTCAACAATC	Plus	22	9	30	59.51	45.45	8.00	0.00
Reverse primer	TGACACATATGAGGCAGCTGAGA	Minus	23	78	56	61.45	47.83	8.00	1.00
Product length	70								

Primer pair 3

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GGCTGAATGCCTTCAACAATC	Plus	21	10	30	58.12	47.62	8.00	0.00
Reverse primer	TGATTACTTGTGAGGTGACACA	Minus	22	93	72	57.32	40.91	5.00	3.00
Product length	84								

Primer pair 4

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GGCTGAATGCCTTCAACAATCAT	Plus	23	10	32	60.12	43.48	8.00	3.00
Reverse primer	TTGATTACTTGTGAGGTGACACA	Minus	23	94	72	57.97	39.13	5.00	3.00
Product length	85								

Primer pair 5

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GCTGAATGCCTTCAACAATCA	Plus	21	11	31	57.42	42.86	8.00	2.00
Reverse primer	TTTGATTACTTGTGAGGTGACACA	Minus	24	95	72	58.57	37.50	5.00	3.00
Product length	85								

Primer pair 6

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CAGGCTGAATGCCTTCAACAATCA	Plus	24	8	31	62.12	45.83	8.00	2.00
Reverse primer	GACACATATGAGGCAGCTGAGAG	Minus	23	77	55	60.80	52.17	6.00	1.00
Product length	70								

Primer pair 7

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AGGCTGAATGCCTTCAACAATCA	Plus	23	9	31	61.07	43.48	8.00	2.00
Reverse primer	ATTTGATTACTTGTGAGGTGACACA	Minus	25	96	72	58.94	36.00	5.00	3.00
Product length	88								

Primer pair 8

Figure 15: Primer pairs generated for SNP rs377244762

V. Conclusion

In conclusion, the extensive research on glutenin and specifically its role in Celiac disease and other gluten-related disorders indicates that it is a growing dilemma for both medical science and the food industry. Current research suggests that Glutenin plays a critical role in triggering the autoimmune response in Celiac disease and it can be addressed in one of two fashions, therapeutic treatment to overcome the immune reaction or create a tolerable substitute for the protein [1,2,6]. We have a detailed mechanistic understanding of the biochemical pathways that are relevant to glutenin and enhance our understanding of its interaction with the human immune system [3,5,8]. This knowledge is not only vital for medical research but also for those who are severely considering the significant rise in gluten-free diets [10]. Additionally, genetic insights into glutenin are invaluable for developing gluten-free wheat varieties and understanding genetic predispositions to gluten-related disorders [12,13]. The future of glutenin research is being supported by advancements in bioinformatics tools[14]. Thus, the comprehensive study of glutenin's biochemical, genetic, and clinical aspects is crucial in addressing the challenges posed by gluten-related disorders and in driving innovations in food production [9].

References:

1. “What Is Celiac Disease?” *Celiac Disease Foundation*, celiac.org/about-celiac-disease/what-is-celiac-disease/. Accessed 4 Dec. 2023.

2. Sapone, A., Bai, J.C., Ciacci, C. *et al.* Spectrum of gluten-related disorders: consensus on new nomenclature and classification. *BMC Med* 10, 13 (2012). https://doi.org/10.1186/1741-7015-10-13

3. T. W. J. M. Van Herpen, M. Riley, C. Sparks, H. D. Jones, C. Gritsch, E. H. Dekking, R. J. Hamer, D. Bosch, E. M. J. Salentijn, M. J. M. Smulders, P. R. Shewry, L. J. W. J. Gilissen, Detailed Analysis of the Expression of an Alpha-gliadin Promoter and the Deposition of Alpha-gliadin Protein During Wheat Grain Development, *Annals of Botany*, Volume 102, Issue 3, September 2008, Pages 331–342, https://doi.org/10.1093/aob/mcn114

4. Sollid LM, Jabri B. Triggers and drivers of autoimmunity: lessons from coeliac disease. Nat Rev Immunol. 2013 Apr;13(4):294-302. doi: 10.1038/nri3407. Epub 2013 Mar 15. PMID: 23493116; PMCID: PMC3818716.

5. Wieser H. Chemistry of gluten proteins. Food Microbiol. 2007 Apr;24(2):115-9. doi: 10.1016/j.fm.2006.07.004. Epub 2006 Sep 7. PMID: 17008153.

6. Ludvigsson JF, Leffler DA, Bai JC, Biagi F, Fasano A, Green PH, Hadjivassiliou M, Kaukinen K, Kelly CP, Leonard JN, Lundin KE, Murray JA, Sanders DS, Walker MM, Zingone F, Ciacci C. The Oslo definitions for coeliac disease and related terms. Gut. 2013 Jan;62(1):43-52. doi: 10.1136/gutjnl-2011-301346. Epub 2012 Feb 16. PMID: 22345659; PMCID: PMC3440559.

7. Catassi C, Elli L, Bonaz B, Bouma G, Carroccio A, Castillejo G, Cellier C, Cristofori F, de Magistris L, Dolinsek J, Dieterich W, Francavilla R, Hadjivassiliou M, Holtmeier W, Körner U, Leffler DA, Lundin KE, Mazzarella G, Mulder CJ, Pellegrini N, Rostami K, Sanders D, Skodje GI, Schuppan D, Ullrich R, Volta U, Williams M, Zevallos VF, Zopf Y, Fasano A. Diagnosis of Non-Celiac Gluten Sensitivity (NCGS): The Salerno Experts' Criteria. Nutrients. 2015 Jun 18;7(6):4966-77. doi: 10.3390/nu7064966. PMID: 26096570; PMCID: PMC4488826.

8. Wieser, Herbert and Peter Koehler. “The Biochemical Basis of Celiac Disease.” *Cereal Chemistry*85 (2008): 1-13.

9. Belitz, H.D., Grosch, W. and Schieberle, P. (2009) Food Chemistry. 4th Edition, Springer-Verlag, Berlin, 1070 p.

10. McCarthy, Niall, and Felix Richter. “Infographic: The Rise of the Gluten-Free Diet.” *Statista Daily Data*, 18 Jan. 2017, www.statista.com/chart/7639/the-rise-of-

the-gluten-free-diet/.

11. Asri N, Rostami-Nejad M, Anderson RP, Rostami K. The Gluten Gene: Unlocking the Understanding of Gluten Sensitivity and Intolerance. Appl Clin Genet. 2021 Feb 11;14:37-50. doi: 10.2147/TACG.S276596. PMID: 33603437; PMCID: PMC7886246.

12. Li Y, Fu J, Shen Q, Yang D. High-Molecular-Weight Glutenin Subunits: Genetics, Structures, and Relation to End Use Qualities. Int J Mol Sci. 2020 Dec 26;22(1):184. doi: 10.3390/ijms22010184. PMID: 33375389; PMCID: PMC7795185.

13. Delorean, E., Gao, L., Lopez, J.F.C. *et al.* High molecular weight glutenin gene diversity in *Aegilops tauschii* demonstrates unique origin of superior wheat quality. *Commun Biol* 4, 1242 (2021). https://doi.org/10.1038/s42003-021-02563-7

14. Andermann T, Cano Á, Zizka A, Bacon C, Antonelli A. 2018. SECAPR—a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. *PeerJ* 6:e5175 https://doi.org/10.7717/peerj.5175

15. *Chromosomal Locaton of Glutenin and Gliadin Loci in Hexaploid Wheat ...*, www.researchgate.net/figure/Chromosomal-locaton-of-glutenin-and-gliadin-loci-in-hexaploid-wheat-The-high-molecular_fig1_273772041. Accessed 5 Dec. 2023.