# From Diet to DNA: A Data-Driven Nutrigenomic Analysis of CTLA4 and MYO9B

Lauren Athon

March 2026

**Abstract**

Mutations and polymorphisms in CTLA4 and MYO9B have both been associated with celiac disease and other autoimmune conditions, yet their genetic mechanisms and potential overlap remain difficult to see when transcript models, clinical variants, and association data are considered separately. Celiac disease is a diet-triggered disorder, which makes it a useful nutrigenomic case study: gluten exposure interacts with genetic and regulatory variation to shape mucosal immune responses. This work presents an exploratory pipeline that integrates NCBI RefSeq (curated isoforms and metadata), Ensembl (broader transcript annotations via `biomaRt`), ClinVar (clinical assertions), and the GWAS Catalog (celiac-associated loci, ontology EFO_0001060) to compare CTLA4 and MYO9B side by side. Sequence concordance checks link Ensembl cDNA to RefSeq mRNA (and peptide-to-protein summaries where applicable); ClinVar variants are mapped onto transcripts with interval overlaps; GWAS hits are summarized relative to gene coordinates. Mixed variant metadata are explored with factor analysis of mixed data (FAMD) after Gower-distance clustering (partitioning around medoids, PAM), using variant type, submission origin, gene-relative position, and related engineered features. In this dataset, MYO9B exhibits greater annotated transcript diversity and a larger ClinVar burden, while celiac GWAS signals localize to intergenic regions near CTLA4—consistent with cis-regulatory mechanisms rather than coding disruption alone. The FAMD ordination (Figure 3) separates variants into clusters that summarize joint patterns of clinical significance, variant class, origin, and location within each gene—supporting interpretation of ClinVar structure beyond single-variable tabulations. The contribution is a reproducible workflow and a set of exploratory findings that can seed nutrigenomic follow-up (for example tissue eQTLs, enhancer annotation, and diet-linked pathway interpretation).

## 1 Introduction

Mutations in both Cytotoxic T-lymphocyte–associated protein 4 (CTLA4) and Myosin IXB (MYO9B) have been associated with several autoimmune conditions, including celiac disease, ulcerative colitis, insulin-dependent diabetes mellitus, Graves disease, and systemic lupus erythematosus [1,2]. Both are protein-coding genes, but they occupy different biological niches: CTLA4 is an immune checkpoint receptor that modulates T cell activation, whereas MYO9B is a cytoskeletal motor protein implicated in cellular organization and signaling [1,2]. In this project, MYO9B polymorphisms are compared with CTLA4 mutations to identify potential overlaps in genomic context along their respective chromosomes, in how variants relate to functional mechanisms, and in how phenotype associations appear in curated clinical databases.

Polymorphisms are common variants (often $\geq 1\%$ frequency) that are frequently benign in effect—similar to variation underlying traits such as hair or eye color—whereas mutations can disrupt

gene function and contribute to disease. Not all mutations become polymorphisms, but all polymorphisms begin as mutations. The central question I pursue is whether CTLA4 and MYO9B show overlapping patterns in metabolic pathways, cellular activity, or chromosomal positioning once transcript, variant, and phenotype layers are cross-referenced. As someone still learning to interpret omics-scale resources, I expect to recover known associations, but I also want to document and reflect on the patterns that emerge when RefSeq, Ensembl, ClinVar, and GWAS evidence are integrated systematically.

Nutrigenomics studies how dietary exposures interact with gene regulation and genetic variation to influence pathways and outcomes. Celiac disease is a natural case study because an environmental trigger (gluten) can initiate a maladaptive immune response in susceptible individuals, and relevant mechanisms may involve both coding variation and regulatory control. This paper does not claim to establish causal nutrigenomic mechanisms; instead, it applies a reproducible integration pipeline that organizes transcript, variant, and phenotype evidence into an interpretable exploratory view of CTLA4 and MYO9B.

This study asks: (1) how do RefSeq-curated isoforms compare with broader Ensembl transcript annotations for these genes; (2) how do clinically annotated variants distribute across transcript spans and clinical-significance categories; and (3) do variants form structured groups in a mixed feature space that can motivate follow-up hypotheses relevant to diet-triggered immune dysregulation and regulatory variation?

## 2 Related Work

The comorbidity landscape above motivates a resource-heavy approach: immune and gut phenotypes are heterogeneous, and candidate genes may contribute through distinct molecular layers [1,2]. From a methods perspective, this paper follows common practice for genomic integration—interval overlap joins (for example with `GenomicRanges`) to map variants onto gene and transcript spans, careful harmonization of transcript identifiers across RefSeq and Ensembl [3], and ordination of mixed categorical and numeric variant features using FAMD (after Gower distances and PAM clustering in the project pipeline) for visualization rather than diagnostic classification. Curated repositories underpin the analysis and are cited in Section 3.1 [3–5]. Because many disease associations are non-coding, regulatory follow-up (enhancers, promoters, eQTLs, chromatin conformation) is a standard next step; here we treat the integration step as groundwork that makes those analyses easier to target.

## 3 Methods

### 3.1 Data sources

- NCBI RefSeq: gene, mRNA, and protein FASTA files and metadata for CTLA4 and MYO9B [1,2]. For each gene, RefSeq provides two curated mRNA (NM_) and protein (NP_) isoforms that anchor sequence-level summaries and transcript–variant linking.
- Ensembl: transcript metadata and sequence retrieval via `biomaRt` [3].
- ClinVar: a filtered variant summary for CTLA4 and MYO9B (clinical assertions) [4].
- GWAS Catalog: celiac-disease–associated loci (ontology ID EFO_0001060 [6]) via `gwasrapidd` [5].

### 3.2 Integration and wrangling

RefSeq FASTA headers were parsed to extract accessions and gene symbols. To expand coverage beyond curated RefSeq isoforms, Ensembl annotations were retrieved with `biomaRt`, capturing metadata across a broader transcript set. Gene Ontology (GO) terms attached to RefSeq metadata were rolled up into higher-level functional categories (immune regulation, signaling, cytoskeletal organization, structural architecture, and molecular activity). The Structural Architecture category groups GO terms describing cytoskeletal organization, specialized actin-based structures, and protein complex assembly—features that tend to be enriched for MYO9B relative to membrane-localized or purely signaling-centric summaries.

Ensembl cDNA sequences were compared with RefSeq mRNAs to cross-reference transcripts and evaluate sequence identity; analogous peptide-versus-protein summaries were generated where comparable models exist. Variant positions were standardized to numeric coordinates. ClinVar assertions and GWAS Catalog associations were integrated with emphasis on celiac-related context (EFO_0001060). The `gwasrapidd` workflow extracted GWAS hits with distance from the annotated gene on the chromosome, upstream/downstream flags, and phenotype links.

Genomic ranges were built with `GenomicRanges`, mapping spans to HGNC symbols. Interval joins placed variants onto gene bodies and Ensembl transcript intervals for burden summaries. After integration, static figures (including those shown here) were generated in `ggplot2`; a companion Shiny app reproduces several linked views.

### 3.3 Clustering and visualization

Isoform architecture was summarized with length distributions and the concordance tables in Section 4.1. ClinVar burden was summarized by transcript overlap counts and by high-level clinical-significance buckets. Celiac GWAS proximity was summarized relative to gene coordinates as in Section 4.2.

For mixed variant metadata, the pipeline (see `NextSteps.Rmd` and `scripts/clinvar_clustering.R`) constructs a feature table combining collapsed variant type and submission origin categories with gene-relative position summaries (for example midpoint and span along the gene model, z-scored for comparability). Pairwise Gower distances were computed on these mixed columns, and PAM (partitioning around medoids) was used to assign each variant to a cluster; the number of clusters was chosen from a small sweep using average silhouette width. Factor analysis of mixed data (FAMD) was then applied so that categorical and numeric columns contribute jointly to a two-dimensional ordination. Points in Figure 3 are colored by data-driven cluster labels (short textual profiles summarizing dominant gene, type, origin, and clinical-significance mix per cluster) and faceted by gene so CTLA4 and MYO9B can be read side by side. The figure is exploratory: it highlights multivariate structure and candidate groupings for follow-up, not a validated diagnostic rule. A one-hot-encoded PCA view was explored earlier in the project but is not shown here in favor of FAMD for mixed data.

# 4 Results

## 4.1 Isoform architecture and sequence concordance

To characterize the sequence and regulatory landscape of CTLA4 and MYO9B in the context of celiac-associated variation, I conducted a layered exploratory analysis spanning isoform architecture, variant distribution, spatial context, and (later) functional annotation summaries. I first examined RefSeq-linked Ensembl transcript diversity and sequence identity to establish transcript-level breadth for each gene.

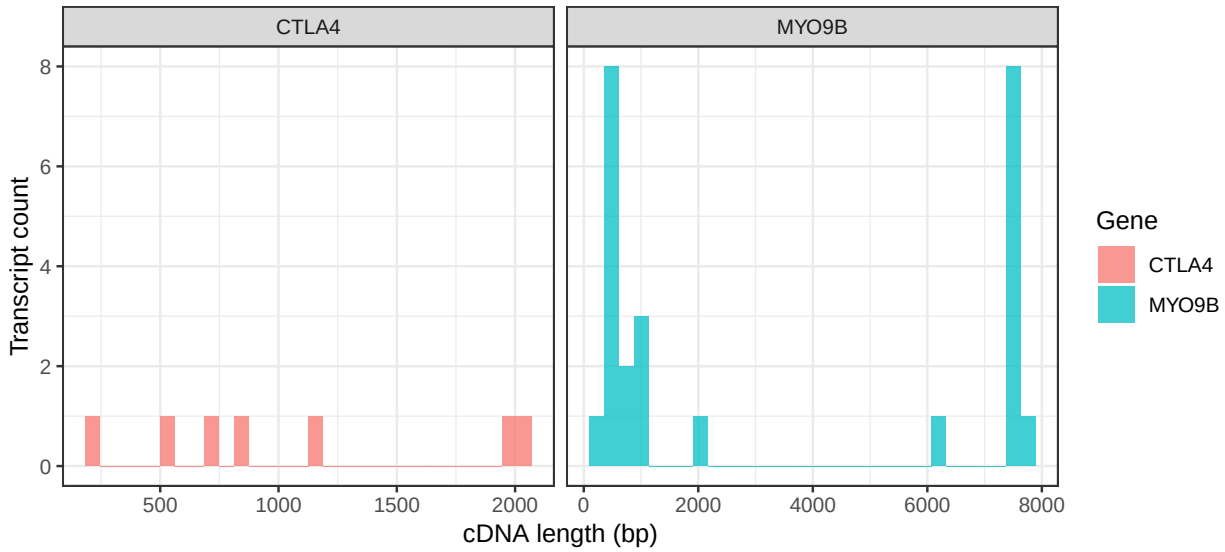



***Figure 1:*** *Transcript cDNA length distribution across Ensembl transcripts for CTLA4 and MYO9B. MYO9B exhibits greater transcript diversity and longer cDNA lengths than CTLA4 in this annotation snapshot.*

In the Ensembl-backed transcript table used for plotting, there are 32 transcript rows in total for the two genes, of which 25 are annotated as MYO9B and 7 as CTLA4. Consistent with Figure 1, MYO9B presents a larger and more complex isoform landscape (more rows and longer cDNA spans) than CTLA4 in this snapshot.

***Table 1:*** *Ensembl cDNA vs RefSeq mRNA sequence matching for comparable transcript models.*

| ensembl_transcript_id | gene_label | sequences_match | n |
|---|---|---|---|
| ENST00000295854 | CTLA4 | FALSE | 1 |
| ENST00000595618 | MYO9B | FALSE | 1 |
| ENST00000648405 | CTLA4 | TRUE | 1 |
| ENST00000682292 | MYO9B | TRUE | 1 |

For curated pairs where Ensembl cDNA and RefSeq mRNA can be compared directly, the sequences were identical or nearly identical in this workflow, reflecting the same underlying spliced transcript sequences represented in NCBI and Ensembl for those matched models.

***Table 2:*** *Identical vs non-identical pairs among Ensembl peptide and RefSeq protein sequences (where comparable).*

| ensembl_transcript_id | gene_label | sequences_match | n |
|---|---|---|---|
| ENST00000295854 | CTLA4 | FALSE | 1 |
| ENST00000595618 | MYO9B | FALSE | 1 |
| ENST00000648405 | CTLA4 | FALSE | 1 |
| ENST00000682292 | MYO9B | FALSE | 1 |

Across the full comparison tables, concordance is not uniform: some transcript IDs match at the cDNA/mRNA level while others do not, and peptide-to-protein comparisons can disagree when models are incomplete or represent distinct biologically valid isoforms (including isoforms flagged for processes such as nonsense-mediated decay). Low pairwise agreement in a small comparison set should therefore be interpreted cautiously as a mixture of reference patching, partial models, and true isoform differences—not as a single "error" label.

### 4.2 Variant burden, clinical buckets, and GWAS proximity

Having characterized transcript structure and sequence identity, I next examined how clinically annotated variants distribute across Ensembl transcripts.

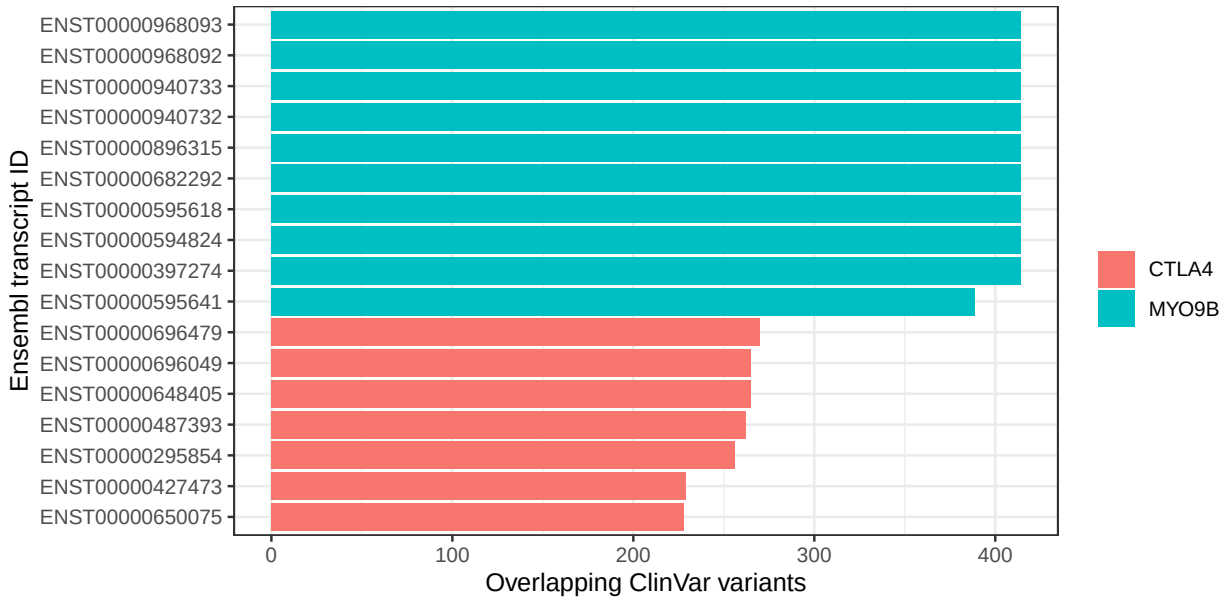



***Figure 2:*** *Top Ensembl transcripts overlapped by ClinVar variants (top 10 per gene). MYO9B shows a higher overall burden of ClinVar variants in this filtered dataset.*

MYO9B had substantially more overlapping ClinVar hits than CTLA4 when variants were cross-referenced to Ensembl transcript IDs (Figure 2). In the underlying ClinVar slice, CTLA4 variant positions were often concentrated toward the extremes of the aggregated transcript coordinate range on chromosome 2, whereas MYO9B hits were spread more broadly across positions on chromosome 19—consistent with a larger, more diffuse variant cloud for MYO9B in this extraction.

***Table 3:*** *ClinVar variants grouped into high-level clinical significance buckets per gene.*

| Gene | clin_bucket | n |
|---|---|---|
| MYO9B | Uncertain | 542 |
| CTLA4 | Uncertain | 269 |
| MYO9B | Other / Not provided | 252 |
| CTLA4 | Other / Not provided | 164 |
| CTLA4 | Pathogenic / Likely pathogenic | 75 |
| MYO9B | Benign / Likely benign | 32 |
| CTLA4 | Benign / Likely benign | 20 |
| CTLA4 | Conflicting | 18 |
| MYO9B | Conflicting | 2 |

*Table 3* makes the interpretation concrete: MYO9B carries the larger absolute variant volume, with a prominent uncertain-significance component, while CTLA4 shows a comparatively larger share of high-confidence pathogenic and likely pathogenic annotations in this filtered dataset. Even though MYO9B did not yield celiac-specific GWAS hits under EFO_0001060 in the catalog extraction used here [5,6], the ClinVar portrait still supports keeping MYO9B in the conversation as a clinically annotated, isoform-heavy locus that merits context-specific follow-up.

To contextualize association signals relative to gene bodies, I summarized celiac GWAS hits with `gwasrapidd`, including distance from the annotated gene, intergenic flags, and upstream/downstream orientation [5,6]. A tabular listing of the same hits is omitted here to save space; the companion Shiny app retains sortable GWAS views for verification.

In this celiac-filtered extraction, three GWAS hits mapped to CTLA4 and none to MYO9B. All three CTLA4 associations were intergenic, with distances of approximately 231 bp, 9,965 bp, and 34,814 bp from the gene model; the closest variant (rs3087243) was classified as upstream. MYO9B yielded zero hits under the same ontology filter—either reflecting catalog coverage at the time of download or a genuinely weaker celiac association signal at the GWAS level for this gene in the studied populations.

Although no variant sits exactly at distance zero within the CTLA4 transcriptional span, the 231 bp upstream variant is a plausible cis-regulatory candidate (promoter-proximal or enhancer-adjacent), while the more distal variants could still matter through long-range chromatin interactions. Follow-up with immune-relevant chromatin marks (for example H3K27ac or H3K4me1), transcription factor binding, accessibility, and conformation capture (Hi-C or promoter capture) would help test those hypotheses [5].

### 4.3 Clustering patterns (FAMD)

Beyond positional summaries and bucket tables, I asked whether variants jointly separate in a space that mixes variant type, submission origin, and location along the gene body. *Figure 3* shows the FAMD ordination described in Section 3.3: points are colored by PAM cluster (labels summarize each cluster's dominant gene, variant type, origin, uncertain/pathogenic/benign mix, and typical position along the locus), with facets for CTLA4 and MYO9B so the two genes can be compared under the same scaling.

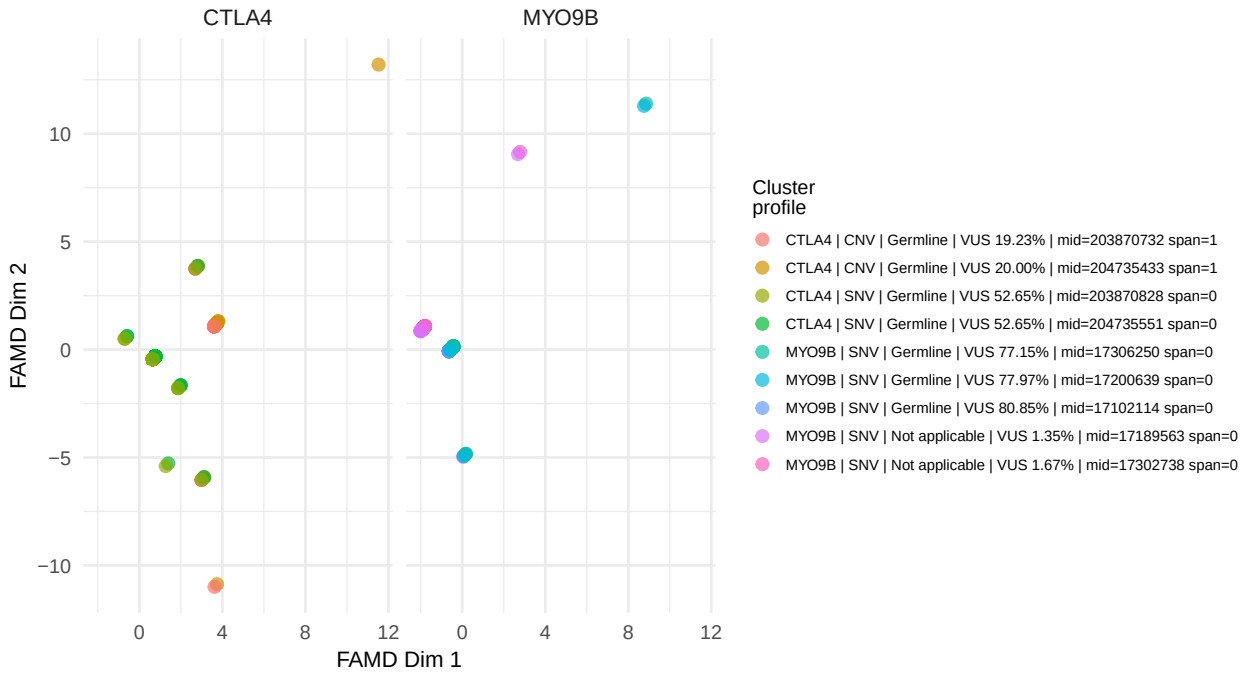



***Figure 3:*** *FAMD ordination of mixed ClinVar features after Gower distance and PAM clustering. Points are colored by cluster profile (legend at right); panels separate CTLA4 and MYO9B. The plot highlights multivariate similarity among variants rather than a single clinical or positional variable in isolation.*

Qualitatively, clusters tend to align with combinations of clinical significance (for example high VUS fraction versus enriched pathogenic/likely pathogenic subsets), variant class (SNV versus structural or other collapsed type groups), and gene-relative placement—consistent with the auto-generated cluster captions in the legend. Because MYO9B contributes many more variants in this ClinVar slice, its FAMD cloud is denser and cluster geometry can differ from CTLA4 even when the same feature engineering is applied within the combined workflow. I treat these groups as hypothesis-generating: they suggest where to inspect representative variants, refine feature coding, or connect to external tracks (expression, constraint, regulatory annotation), not as stable disease subtypes.

## 5 Discussion

Celiac disease sits in a broader landscape of genetic heterogeneity: risk loci span immune regulatory genes and genes involved in cellular structure and signaling. Genome-wide association studies highlight regions associated with susceptibility, but they do not, by themselves, specify the functional routes through which each candidate gene contributes. In that gap, transcript-resolved annotation and ClinVar phenotype text can provide an independent lens—even though neither replaces experimental validation.

***Table 4:*** *Top ClinVar phenotype strings by variant count (truncated for display). These are database labels attached to variant records, not patient-level diagnoses.*

| Gene | PhenotypeList | n_variants |
|---|---|---|
| CTLA4 | Autoimmune lymphoproliferative syndrome due to CTLA4 haploinsuffici... | 3104 |

| | | |
|---|---|---|
| CTLA4 | Inborn genetic diseases | 210 |
| CTLA4 | Celiac disease, susceptibility to, 3 | 70 |
| CTLA4 | Hashimoto thyroiditis | 70 |
| CTLA4 | Systemic lupus erythematosus | 70 |
| CTLA4 | Type 1 diabetes mellitus 12 | 70 |
| MYO9B | Ovarian serous cystadenocarcinoma | 300 |
| MYO9B | Familial cancer of breast | 200 |
| MYO9B | Sarcoma | 198 |
| MYO9B | Gastric cancer | 140 |
| MYO9B | Melanoma | 140 |
| MYO9B | Lung cancer | 138 |

*Table 4* illustrates how differently the two genes appear when viewed through ClinVar-associated phenotype strings (not causal claims). Enteropathy-, autoimmunity-, and lymphoproliferation-related language appears frequently in the CTLA4 tail of the list; MYO9B's top entries can include neoplastic and susceptibility-related phrasing depending on the download. This matters for interpretation: CTLA4 deficiency is not equivalent to classic HLA-driven celiac disease, yet checkpoint dysfunction can produce enteropathies that mimic celiac phenotypes in the clinic. That distinction underscores how shared gastrointestinal presentations can arise through distinct molecular routes—immune dysregulation versus epithelial/cytoskeletal programs—consistent with the divergent ontology emphasis summarized in Appendix Figure A1.

Integrating RefSeq-based GO roll-ups (Appendix Figure A1) with variant and GWAS layers helps separate architecture from association. CTLA4's ontology profile skews toward immune regulation and membrane-associated signaling, whereas MYO9B emphasizes cytoskeletal organization, actin-based motility, and Rho-family signaling. Viewed together with GWAS geography, the intergenic celiac hits near CTLA4 argue for prioritizing cis-regulatory follow-up (tissue-specific eQTLs, accessibility, TF binding, and chromatin conformation), while MYO9B's many isoforms argue for isoform-resolved expression and functional assays.

From a nutrigenomics perspective, the value of this work is hypothesis generation at the interface of diet-triggered inflammation and genomic context: gluten-driven immune activation may intersect checkpoint regulation (CTLA4) and barrier/cytoskeletal programs (MYO9B), with regulatory variation modulating expression in a tissue- and context-specific manner.

**Limitations.** MYO9B lacks celiac-filtered GWAS hits in the catalog snapshot used here, and both genes' ClinVar slices contain large uncertain-significance fractions—so conclusions are sensitive to filtering, submission bias, and annotation lag. Intergenic CTLA4 GWAS variants are candidate regulatory alleles; without orthogonal epigenomic evidence, mechanistic claims remain speculative. FAMD coordinates and PAM cluster assignments depend on feature coding, distance choices, and $k$; different preprocessing or cluster counts would change the plot.

*Figure 3* complements the genomic summaries by showing that multivariate ClinVar structure—type, origin, position, and clinical labels together—supports interpretable groupings within each gene, motivating mixed-metadata views for targeted follow-up rather than as a diagnostic classifier.

## 6 Conclusion / Summary

This paper integrates RefSeq, Ensembl, ClinVar, and GWAS Catalog evidence into a reproducible, isoform-aware portrait of CTLA4 and MYO9B in celiac-related context. MYO9B carries greater

transcript diversity and ClinVar volume in this extraction, while celiac GWAS signals localize to intergenic regions near CTLA4—pointing toward cis-regulatory mechanisms as a high-priority follow-up axis alongside immune checkpoint biology. RefSeq-guided ontology contrasts reinforce that the two genes occupy different functional layers (checkpoint signaling versus cytoskeletal and motor-associated programs), which is useful when interpreting heterogeneous autoimmune and gut-related risk. Future work should pair these summaries with tissue-resolved eQTLs, chromatin state and TF occupancy in immune-relevant cells, and isoform-specific expression for MYO9B. The companion Shiny application preserves interactive versions of several plots for exploration.

## References


[1] NCBI Gene. *CTLA4 cytotoxic T-lymphocyte associated protein 4* [Homo sapiens (human)]. Gene ID: 1493. Bethesda, MD: National Center for Biotechnology Information. Available: https://www.ncbi.nlm.nih.gov/gene/1493

[2] NCBI Gene. *MYO9B myosin IXB* [Homo sapiens (human)]. Gene ID: 4650. Bethesda, MD: National Center for Biotechnology Information. Available: https://www.ncbi.nlm.nih.gov/gene/4650

[3] Cunningham F, et al. Ensembl 2022. *Nucleic Acids Research.* 2022;50(D1):D988–D995. doi:10.1093/nar/gkab1049. Data access via BioMart: https://www.ensembl.org/biomart/martview/

[4] Landrum MJ, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research.* 2018;46(D1):D1062–D1067. doi:10.1093/nar/gkx1152. Resource: https://www.ncbi.nlm.nih.gov/clinvar/

[5] Buniello A, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research.* 2019;47(D1):D1005–D1012. doi:10.1093/nar/dky1120. Catalog trait page (celiac disease): https://www.ebi.ac.uk/gwas/efotraits/EFO_0001060

[6] Experimental Factor Ontology (EFO). EFO_0001060 *celiac disease.* European Bioinformatics Institute. Available: http://www.ebi.ac.uk/efo/EFO_0001060


## Appendix

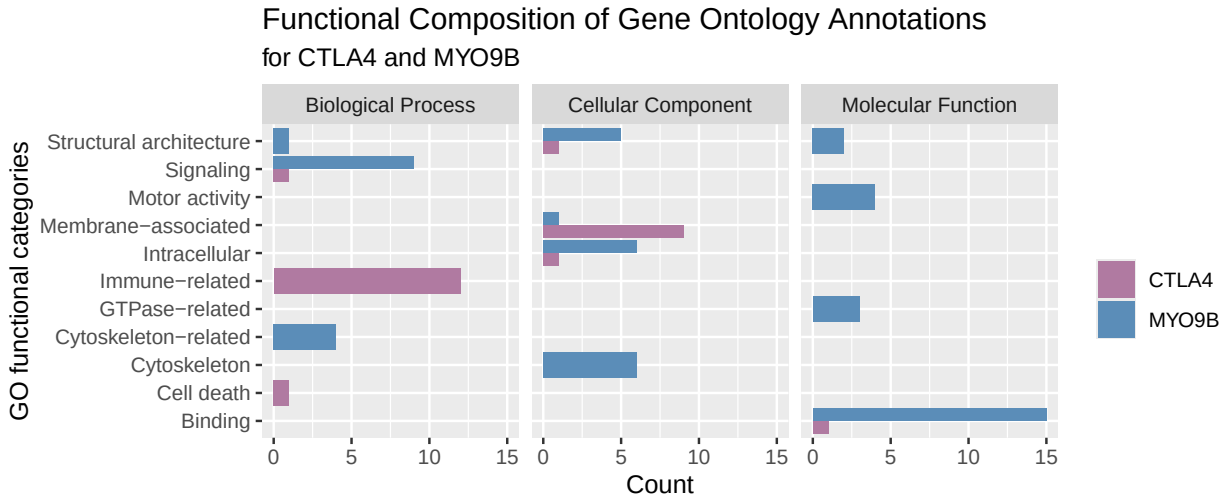
Interactive supplement. An interactive exploration of the same integrated tables and plots (including linked brushing where implemented) is published as a companion Shiny application: https://laurenathon.shinyapps.io/An-Interactive-Analysis-of-CTLA4-and-MYO9B/

### Appendix B. Condensed gene background (from RefSeq / NCBI-style summaries)

CTLA4 (*CTLA4 cytotoxic T-lymphocyte associated protein 4*) is a protein-coding gene on chromosome 2 (plus strand). It encodes an immunoglobulin-superfamily protein that transmits inhibitory signals to T cells; the protein includes a V-like domain, a transmembrane segment, and a cytoplasmic tail, with alternate splice isoforms reported [1]. CTLA4 has multiple synonyms (for example

CD152) and has been implicated across a range of autoimmune phenotypes in the genetics literature [1]. Subcellularly, CTLA4 can localize to the plasma membrane and endocytic compartments; the membrane-bound form functions as a disulfide-linked homodimer, while a soluble form can circulate as a monomer [1].

MYO9B (*MYO9B myosin IXB*) is a protein-coding gene on chromosome 19 (plus strand). It encodes an unconventional myosin heavy chain that participates in actin-based motility, Rho-family signaling, and cytoskeletal organization—distinct from conventional non-muscle myosin-9 (MYH9) [2]. MYO9B localizes to actin-rich structures such as the cytoskeleton, lamellipodium, and ruffles, and enables multiple binding and motor-related activities [2]. Polymorphisms in MYO9B have been discussed in celiac disease and inflammatory bowel contexts in the association literature [2].




***Appendix Figure A1:*** *RefSeq Gene Ontology term counts summarized into higher-level functional categories for CTLA4 vs MYO9B (from project metadata).*