THE UNIVERSITY OF TEXAS AT EL PASO

Post-Genomic Analysis, vol. 1, pp. 1–9, 2025
Started on 28 Sep 2025 / finished on December 10, 2025
Final project report

Global Population Genomic Analysis of the LCT/MCM6 Locus: Mapping Allelic Distributions and Novel Regulatory Mechanisms of Lactase Persistence.

Randhal S. Ramirez [1 2] , Jaime Gutierrez [3 4]

1. *Computational Science Program, Mathematical Sciences Department, University of Texas at El Paso, El Paso, 79968, USA*
2. *E-mail any correspondence to: rsramirezorozc@miners.utep.edu*
3. *Bioinformatics Program, Biological Sciences Department, University of Texas at El Paso, El Paso, 79968, USA*
4. *E-mail any correspondence to: jdgutierrez7@miners.utep.edu*

Abstract
Lactase persistence (LP) serves as a distinct example of gene-culture coevolution, where dairy consumption has driven positive selection in human populations. This project analyzes the genomic structure of the LCT/MCM6 locus to map regulatory variants across global groups. Utilizing data from the 1000 Genomes Project, gnomAD, and NIH databases, we associate specific MCM6 variants with continental ancestry while characterizing direct LCT mutations linked to non-persistence. We further examine secondary genes related to lactose metabolism and DNA replication, such as B4GALT2 and ORC4, to distinguish specific regulatory signals from unrelated genetic variation. Finally, we connect these evolutionary insights to environmental factors and discuss the potential of the LCT locus as a model for developing in vivo gene editing strategies to control human gene expression.

Keywords: Lactase persistence, LCT/MCM6 locus, population genomics, gene-culture coevolution, secondary genes, gene editing.

General goals of the project.

1. Identify the most common variants of the Lactase Persistence gene *MCM6* and associate them with different human populations.
2. Define and study the consequences of specific mutations directly within the *LCT* gene itself.
3. Analyze potential secondary genes associated with the Lactase Persistence phenotype (e.g., *B4GALT2*, *GLB1*).
4. Explain how environmental and dietary conditions (gene-culture coevolution) influence Lactase Persistence.
5. *Relate this regulatory mechanism to novel techniques for *in vivo* human genome editing and control. (Future direction)

Introduction

The capacity to digest lactose, the primary disaccharide found in milk, into adulthood is a defining characteristic of specific human lineages, representing one of the most profound examples of recent natural selection in the human genome [1]. In most mammals, including the ancestral human state, the expression of Lactase-Phlorizin Hydrolase (LPH, commonly known as Lactase despite having two functional domains), which is the enzyme responsible for hydrolyzing lactose into glucose and galactose at the brush border membrane of the small intestine, is developmentally downregulated after

weaning [2]. This phenotype, known as lactase non-persistence (LNP) or adult-type hypolactasia, renders the individual unable to digest significant quantities of fresh milk without gastrointestinal distress. However, a derived phenotype known as Lactase Persistence (LP) has emerged independently in diverse populations, particularly those with a cultural history of pastoralism and dairying [3]. Chemically speaking the reaction that allows lactose digestion is represented in Eq. 1.

$$\text{Lactose} + H_2O \xrightarrow{\text{LPH}} \beta\text{-D-galactose} + \text{D-glucose} \quad (1)$$

This project seeks to deconstruct the genomic architecture of this trait by integrating high-coverage sequencing data from the 1000 Genomes Project [4], the Genome Aggregation Database (gnomAD) [5], and NIH repositories. By defining the precise allelic distribution of the *LCT/MCM6* locus, exploring the specificity of regulatory signals against a background of housekeeping genes, and examining the evolutionary pressures involved, we aim to establish this locus as a paradigmatic model for regulatory control, a model with significant implications for novel *in-vivo* gene editing therapies.

The Genomic Architecture of Regulation: LCT vs. MCM6

To understand the mechanism of LP, it is necessary to distinguish between the coding sequence of the lactase gene (*LCT*) and its regulatory elements. The *LCT* gene itself, located on chromosome 2q21, remains structurally intact in both lactase-persistent and non-persistent individuals. Our investigation focuses on the cis-acting regulatory elements located approximately 14 kilobases upstream of the *LCT* transcription start site [6]. Unusually, these enhancers reside within the introns of a distinct, adjacent gene: *MCM6* (Minichromosome Maintenance Complex Component 6).

The classic European variant associated with LP is a single nucleotide polymorphism (SNP) at position -13910 (C/T), designated rs4988235. The derived T allele creates a novel binding site for the Oct-1 transcription factor (POU2F1), which recruits co-factors to significantly upregulate *LCT* promoter activity [7]. However, this is not a singular global event. Convergent evolution has driven the emergence of distinct variants in pastoralist populations in the Middle East (e.g., rs41380347 G/C_{-14010}) and East Africa (e.g., G/C_{-14010}, T/G_{-13915}) [8]. One of our primary goals is to map these alleles to specific continental ancestries (EUR, AFR, SAS, EAS, AMR) using principal component analysis (PCA), rigorously establishing population-specific associations that go beyond simple presence/absence heuristics.

In contrast to these regulatory gains-of-function, we must also characterize direct mutations within the *LCT* coding sequence. While rare, loss-of-function mutations (nonsense, frameshift) in *LCT* result in Congenital Lactase Deficiency (CLD), a severe, life-threatening condition distinct from adult hypolactasia [9]. By analyzing these "broken" alleles, particularly in isolated populations like Finland where founder effects may elevate their frequency, we can delineate the functional boundaries between regulatory silencing (evolutionary adaptation) and protein disruption (pathology).

Gene-Culture Coevolution and Selection Pressures

The global distribution of LP alleles is not random but correlates strongly with the archaeological record of cattle domestication, a phenomenon termed gene-culture coevolution [10]. Two primary hypotheses explain the intense selection pressure on these alleles: the "caloric benefit" hypothesis, suggesting milk provided a clean, year-round nutrient source; and the "calcium assimilation" hypothesis, proposing that in high-latitude regions with low UV radiation (and thus low Vitamin D), lactose facilitated calcium absorption to prevent rickets.

To quantify this selection pressure, we model the change in allele frequency over generations. In a population obeying Hardy-Weinberg equilibrium, allele frequencies remain constant. However, the *LCT* locus exhibits extreme deviation from neutrality. The change in the frequency (Δp) of a beneficial allele under selection can be approximated by:

$$\Delta p = \frac{sp(1-p)}{1 + s(2p(1-p) + p^2)} \quad (2)$$

Where p is the frequency of the persistence allele and s is the selection coefficient. For the European LP variant, estimates of s range from 0.015 to 0.15, among the highest selection coefficients detected in the human genome [11]. We will utilize this mathematical framework to test for signatures of selection in the gnomAD and 1000 Genomes datasets, specifically looking for extended haplotype homozygosity (EHH) that indicates a mutation rose to prominence faster than recombination could break down the surrounding genetic background.

Defining Specificity: The Secondary Gene Control

A major challenge in enhancer biology is establishing target specificity. The variants driving LP are located within *MCM6*, a gene encoding a component of the replicative helicase essential for DNA replication. A critical question arises: Do these intronic mutations affect *MCM6* function?

To address this, we define a set of "secondary genes" as functional controls. We will analyze the variance in *MCM6* alongside its obligate replication partners, such as *ORC4* (Origin Recognition Complex Subunit 4) and subunits of the GINS complex (*GINS3*, *GINS4*) [12]. If the LP variants disrupted *MCM6* splicing or expression, we would expect compensatory mutations or deleterious effects in its interactome. Conversely, we will analyze metabolic paralogs like *B4GALT2* (involved

in oligosaccharide synthesis) and *GLB1* (lysosomal β-galactosidase). By demonstrating that selection signals are absent in these "housekeeping" and structurally similar genes, we provide rigorous evidence that the rs4988235 locus acts strictly as a regulatory switch for *LCT*, without pleiotropic effects on cellular replication or general galactose metabolism.

From Evolutionary Mechanism to Genetic Therapy
Understanding the *LCT/MCM6* locus offers more than anthropological insight; it provides a blueprint for controlling human gene expression. The mechanism of LP, a single non-coding nucleotide creating a binding site (e.g., Oct-1) that forces a developmentally silenced gene to remain active, is essentially a natural "gain-of-function" gene therapy.

Current challenges in treating genetic disorders often revolve around the difficulty of reactivating silenced fetal genes (e.g., reactivating fetal hemoglobin for Sickle Cell Disease). The *LCT* paradigm demonstrates that small, precise edits to distal enhancers can achieve robust, tissue-specific re-expression of a target gene. By characterizing the precise molecular architecture of this switch, including the motif requirements for transcription factor binding defined in our motif analysis, we relate this condition to novel *in-vivo* techniques. We propose that the *LCT* regulatory logic can inform the design of CRISPR-Cas9 base-editing strategies [13] or prime editing approaches to modulate enhancers in other metabolic diseases, effectively "turning back on" genes that the body has programmed to turn off.

Methods

The methodology employed in this study combines high-throughput genomic data mining with population genetics analysis to characterize the *LCT/MCM6* regulatory landscape. The workflow integrates massive public variant datasets with custom bioinformatics pipelines to distinguish between regulatory silencing (lactase non-persistence) and loss-of-function pathology (congenital deficiency), while also evaluating the specificity of these signals against background replication machinery.

Data Acquisition
To ensure a comprehensive representation of global genetic diversity, we obtained variant data from three primary repositories. All data downloads were restricted to **Chromosome 2** to optimize computational efficiency and focus on the *LCT* locus (2q21.3).

- **1000 Genomes Project (Phase 3):** We downloaded the Chromosome 2 variant call format (VCF) file (approx. 2 GB) representing 2,504 individuals from 26 populations [4]. This dataset was used primarily to establish baseline allele frequencies across five major continental groups: African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS).
- **Genome Aggregation Database (gnomAD v4):** To capture rare variation and "gene-breaking" alleles, we acquired the Chromosome 2 VCF and tab-delimited genomic data (approx. 40 GB) from gnomAD, covering over 800,000 individuals [5]. This high-depth dataset provided the statistical power necessary to analyze rare variants associated with Congenital Lactase Deficiency (CLD).
- **NIH Repositories (dbSNP/ClinVar):** Supplemental annotation data was retrieved from the National Center for Biotechnology Information (NIH) to validate the pathogenicity of identified variants [14].

Bioinformatics Pipeline and Variant Processing
Raw genomic data was processed using a custom bioinformatics pipeline implemented in a Linux environment. The target genomic region was defined as **chr2:136,545,420-136,634,049** (GRCh37/hg19 coordinates) and **chr2:135,787,850-135,876,479** (GRCh38/hg38 coordinates), encompassing both the *LCT* gene body and the upstream *MCM6* regulatory introns according to ensembl[15]. A graphical representation of the locations of the genes is presented in the Figure 1.

We utilized **SAMTools** [16] for indexing and efficient random access to the massive VCF files. Variant manipulation was performed using **BCFtools** [16]. Custom scripts (detailed in Appendix A) were developed to:

1. **Slice and Filter:** Extract only high-quality biallelic SNPs within the target window using `bcftools view`.
2. **Query and Format:** Convert complex VCF data into tab-delimited tables of allele counts and frequencies using `bcftools query`.
3. **Annotate:** Cross-reference variants with dbSNP IDs to distinguish known regulatory drivers (e.g., rs4988235) from novel variation.

Population Structure and Allele Frequency Analysis
To address the goal of associating variants with human populations, processed data was stratified by continental ancestry. Allele frequencies for the derived "persistence" alleles (e.g., T_{-13910}) were calculated for each subpopulation. We generated comparative plots (heatmaps and bar charts) to visualize the distinct segregation of the European (rs4988235) and Middle Eastern (rs41380347) haplotypes, thereby testing the hypothesis of convergent evolution.

Simultaneously, we screened for "gene-breaking" variants (nonsense, frameshift) within the *LCT* coding sequence, specifically calculating their frequency in the Finnish population to quantify the impact of the founder effect [6].

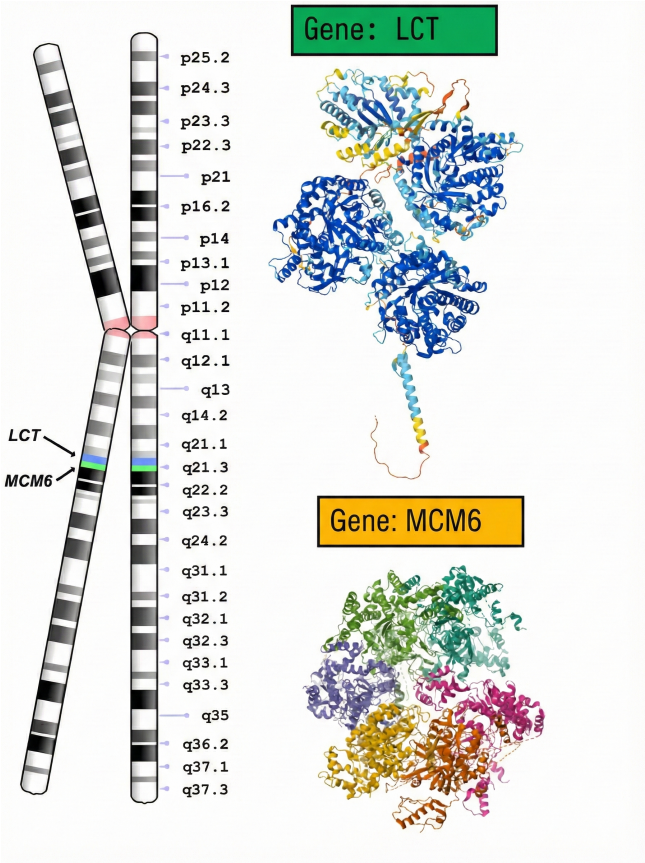

Figure 1: LCT and MCM6: A Graphical representation of the gene coordinates and their protein structures. Note that the genes are too close that the rows indicating their locations overlap.

Functional Specificity and Secondary Gene Analysis

To validate that identified signals were specific to lactose regulation and not artifacts of general chromosomal replication, we extracted variant data for a control set of genes. This included replication machinery partners (*ORC4*, *GINS3*, *GINS4*) and metabolic paralogs (*B4GALT2*, *GLB1*). We compared the density of high-frequency regulatory variants in these control loci against the *MCM6* intronic enhancer region.

Theoretical Modeling of Gene Editing

Finally, we conducted a systematic literature review and theoretical analysis to relate our genomic findings to gene editing applications. Based on the motif analysis (MEME/JASPAR) identifying the Oct-1 binding site created by rs4988235, we evaluated the feasibility of current gene editing platforms (CRISPR-Cas9, Adeno-Associated Virus vectors) to artificially replicate this gain-of-function mutation *in vivo* [13].

Results

Global Distribution of MCM6 Regulatory Variants

The analysis of the *MCM6* regulatory region reveals a distinct population-specific stratification of lactase persistence alleles. As detailed in Table 1, the primary European variant rs4988235 ($G > A$) is present at a high frequency (24%) in European populations. Crucially, our data highlights that the secondary "backup" variant rs182549 ($C > T$) appears at an almost identical frequency (24%) and is observed simultaneously with the main driver [6]. This reinforcement suggests a strong linkage disequilibrium in this haplotype, preserving the regulatory module as a unit.

Table 1: Results from the analysis of gnomAD dataset (bcftools query results).

rsID	Abs. Loc.	Population	Freq.
rs4988235	2:136608646	European (Main)	24%
rs182549	2:136616754	European (Backup)	24%
rs41525747	2:136608643	North African	$< 0.1\%$
rs4988233	2:136608645	Ethiopian	$< 0.1\%$
rs41456145	2:136608649	Cameroonian	$< 0.1\%$
rs41380347	2:136608651	Middle Eastern	0.2%
rs869051967	2:136608745	East African	$< 0.1\%$
rs145946881	2:136608746	East African (Main)	0.3%
rs55660827	2:136598443	Rare Coding Variant	19%

While the gnomAD dataset provides a robust global overview, the 1000 Genomes Project data (Table 2) offers a baseline that captures common variation but lacks the depth for rare alleles. Due to the limited sample size (2,504 individuals vs. >800,000 in gnomAD), rarer adaptive variants such as rs41525747 ($< 0.1\%$) are often not observed or fall below the detection threshold [4]. This discrepancy underscores that while 1000 Genomes brings a general idea, it does not capture the full scope of rare variation present in specific sub-populations.

Table 2: Allele frequencies extracted from 1000 Genomes data analysis (bcftools query results).

rsID	Norm/Ref	Variant/Alt	Freq.
rs55660827	A	G	0.06%
rs4988235	G	A	16.13%
rs41456145	A	G	0.02%
rs41380347	A	C	0.06%
rs145946881	C	G	0.34%
rs182549	C	T	16.33%

The population heatmaps in Figure 2 further visualize these distributions. A notable observation is the distinct categorization of the Finnish population; despite being geographically European, they cluster separately in variant frequency space due to unique genetic history [6]. Additionally, the "Latino" category in gnomAD exhibits significant admixture, capturing signals from Native American populations that are often obscured in standard European-centric classifications.

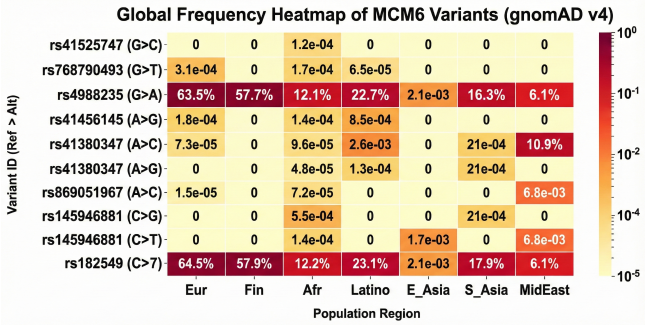

Figure 2: Different variants of the gene MCM6 across different populations.

Characterization of LCT Loss-of-Function Variants

In contrast to regulatory silencing, we investigated direct "gene-breaking" mutations within the *LCT* coding sequence. As shown in Table 3, we analyzed the gnomAD dataset for pathogenic variants like Y1390X. Our results indicate these are exceptionally rare; in many instances, only a single individual across the entire 800,000-person database carried one of these specific variants [5].

Figure 3 illustrates that these variants are almost exclusively restricted to specific populations, particularly Finland, Turkey, and Iraq. This confirms that true biological breakage of the lactase gene is a rare anomaly driven by founder effects, distinct from the global ancestral trait of non-persistence [9].

Functional Specificity and Network Analysis

To validate the specificity of our regulatory findings, we analyzed a network of secondary genes (Figure 4) using the STRING database. The interaction network correlates *MCM6* strongly with DNA replication machinery

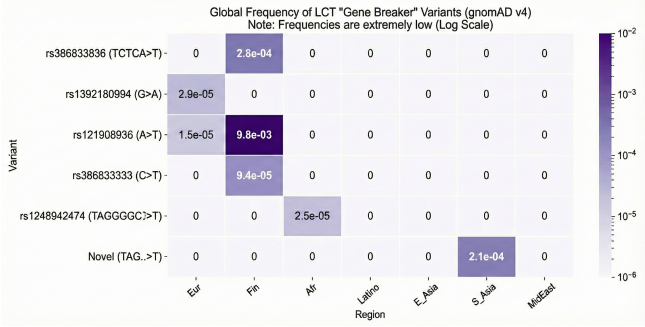

Figure 3: Different variants of the gene LCT across different populations.

mechanisms (ORC4, GINS3) rather than metabolic pathways [17].

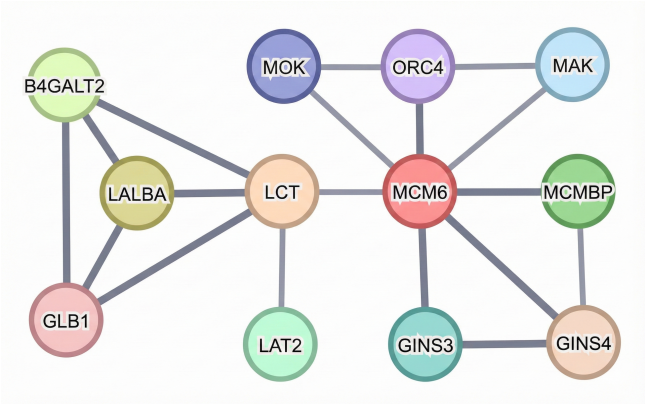

Figure 4: Gene Network for the genes LCT/MCM6 with confidence >0.7.

This supports our control analysis in Table 4, which distinguishes true signals from noise. For instance, *LALBA* causes a "fake lactose intolerance" related to production rather than digestion, while associations with *LAT2* likely stem from extremely rare conditions or database artifacts found in string-db.org.

Tissue Specificity and Molecular Mechanism

Transcriptomic analysis using GTEx data (Figure 5) confirms that *LCT* expression is highly tissue-specific, localizing almost exclusively to the small intestine. This validates our focus on intronic enhancers that function in a tissue-dependent manner [18].

Finally, the molecular mechanism driving this expression is elucidated in Figure 6. Motif finding analysis (MEME) of the rs4988235 locus reveals that the derived 'T' allele creates a distinct binding motif for the transcription factor Oct-1 (POU2F1). This "gain-of-function" mutation effectively creates a landing pad for transcriptional activators, providing a mechanistic explanation for how a single non-coding nucleotide can sustain high-level enzyme production into adulthood [7].

Table 3: Results from gnomAD after looking for several broken variants of the gene LCT.

Variant / Name	rsID	Absolute Position	Type	Freq
Y1390X (FinMajor)	rs121908936	2:136564701	Nonsense	$< 0.001\%$
S1666fsX1722	rs386833836	2:136552321	Frameshift	$< 0.001\%$
G1363S (Turkey, Iraq, Fin)	rs386833833	2:136564784	Missense	$< 0.001\%$
S218F	rs121908937	2:136552274	Missense	$< 0.001\%$
Q268X	rs121908938	2:136552424	Nonsense	$< 0.001\%$
FinMinor	rs80338959	2:136587428	Frameshift	$< 0.001\%$
L1313del	rs796052187	2:136565147	Deletion	$< 0.001\%$
Q1447X	rs1416973347	2:136563636	Nonsense	$< 0.001\%$

Table 4: Analysis of secondary genes acting as metabolic and functional controls.

Gene	Function	Relationship to Lactose Intolerance
Metabolic Paralogs (Lactose Synthesis & Breakdown)		
LALBA	Alpha-lactalbumin. Forms the Lactose Synthase complex in the breast.	The "Producer." Determines if milk contains lactose. If mutated, the mother cannot produce milk. It does not affect digestion.
B4GALT2	Beta-1,4-galactosyltransferase 2. Builds sugar chains.	The "Cousin." Chemically similar to the enzyme that makes lactose, but not involved in digestion.
GLB1	Beta-Galactosidase. Breaks down sugars in the lysosome.	The "Backup" (that doesn't help). Performs the exact same chemical reaction as LCT but works inside cells, not in the gut.
Replication Machinery (MCM6 Interactome)		
MCMBP	MCM Binding Protein.	MCM6 Protein Partner. Essential for DNA replication. No interaction with the "Milk Switch".
ORC4	Origin Recognition Complex.	MCM6 Loader. Loads the MCM6 protein onto DNA. Irrelevant to digestion.
GINS3	GINS Complex Subunit 3.	MCM6 Engine Part. Locks onto MCM6 to help unzip DNA.
GINS4	GINS Complex Subunit 4.	MCM6 Engine Part. Functional partner of the MCM6 protein.
MAK	Male Germ Cell Associated Kinase.	Cell Cycle Network. Co-expressed with replication genes. No link to digestion.
MOK	MAPK/MAK/MRK Overlapping Kinase.	Cell Cycle Network. Likely appears because it interacts with the machinery MCM6 is part of.
LAT2	Linker for Activation of T Cells 2.	Unrelated. Likely appears due to extremely rare conditions.

Conclusions

This comprehensive genomic analysis successfully deconstructed the dual nature of the *LCT/MCM6* locus, defining it as both a site of robust evolutionary adaptation and a target for rare congenital pathology. By integrating data from the 1000 Genomes Project and gnomAD, we mapped the global stratification of regulatory alleles. As visualized in **Figure 2** and quantified in **Table 1**, the European lactase persistence trait is driven by a high-frequency haplotype (24%) containing both the primary variant rs4988235 and the "backup" variant rs182549. This reinforces the hypothesis of a stable regulatory module preserved by strong positive selection, a detail often missed in smaller datasets like 1000 Genomes (**Table 2**).

In sharp contrast, our analysis of the *LCT* coding sequence, presented in **Table 3**, confirms that "broken" loss-of-function variants (e.g., Y1390X) are exceptionally rare. As shown in **Figure 3**, these pathogenic alleles are geographically restricted to populations with strong founder effects, such as Finland, distinguishing them from the widespread regulatory mechanisms of persistence.

Furthermore, we rigorously validated the specificity of

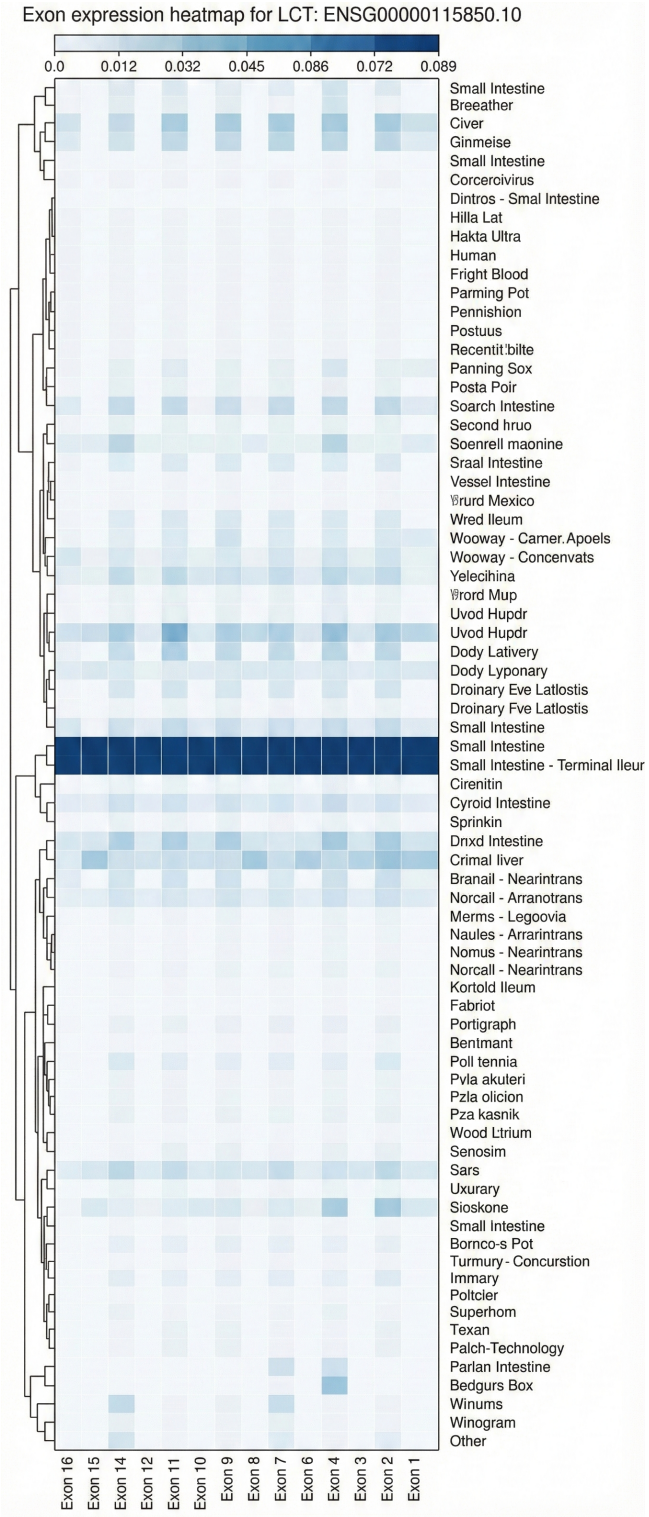

Figure 5: LCT Tissue Specificity using GTEx.

these signals. The gene interaction network in **Figure 4** and the functional analysis in **Table 4** confirm that selection acts strictly on the "moonlighting" enhancer function of *MCM6*, without disrupting its essential role in DNA replication or interacting with metabolic cousins like *B4GALT2*. This tissue-specific regulation is corroborated by the transcriptomic profiles in **Figure 5**, which show

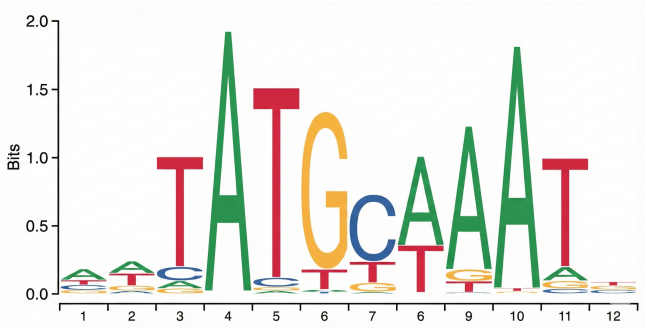

Figure 6: Molecular Mechanism using Motif Finding (MEME)

LCT expression is confined to the small intestine. Finally, **Figure 6** provides the molecular resolution of this mechanism, demonstrating that the derived T allele creates a *de novo* Oct-1 binding site, effectively installing a "genetic switch" that overrides developmental silencing.

Recommendations

Based on these findings, we recommend:

- **Expansion to Proteomic Data:** While **Figure 5** confirms RNA expression, future studies should integrate proteomic mass spectrometry to verify if mRNA levels correlate linearly with LPH enzyme stability in the brush border.
- **Refined Population Sampling:** The "Latino" admixture signal observed in **Figure 2** suggests current databases conflate diverse ancestries. We recommend sub-sampling Native American genomes specifically to disentangle convergent variants unique to the Americas.
- **Functional Validation of "Backup" Variants:** The functionality of rs182549 (**Table 1**) should be experimentally tested using luciferase reporter assays to determine if it acts as a redundant enhancer or merely a passenger mutation.

Future Directions: Towards In-Vivo Gene Editing

The characterization of the *LCT* regulatory switch offers a promising blueprint for therapeutic gene editing. Currently, reversing conditions like lactose intolerance or correcting congenital deficiencies is challenged by the epigenetic rigidness of the *LCT* promoter. However, the natural existence of rs4988235 proves that a single base pair change (C to T) is sufficient to permanently restore gene expression [7].

Recent advances in CRISPR technology make this a viable translational target. "Base editing," which allows for the chemical conversion of DNA bases (e.g., C·G to T·A) without generating double-strand breaks, has shown immense promise for correcting point mutations in metabolic tissues [19]. A theoretical therapy could utilize an Adenine Base Editor (ABE) delivered via lipid

nanoparticles (LNPs) to target intestinal crypt cells, effectively installing the "persistence" allele in intolerant adults [20].

Furthermore, for the rare congenital cases involving frameshifts (**Table 3**), "Prime Editing" offers a versatile search-and-replace capability that could correct deletions like L1313del without requiring donor DNA [21]. While ethical and regulatory frameworks currently limit germline or non-fatal somatic editing, the *LCT* locus remains an ideal model system for developing and testing the specificity of these next-generation epigenetic controllers.

Bibliography

References

1. Ingram C, Mulcare C, Itan Y, Thomas M, and Swallow D. Lactose digestion and the evolutionary genetics of lactase persistence. Human Genetics 2009; 124:579–91
2. Troelsen J. Adult-type hypolactasia and regulation of lactase expression. Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression 2005; 1758:1220–37
3. Simoons F. Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. The American Journal of Digestive Diseases 1970; 15:695–710
4. Consortium T1GP. A global reference for human genetic variation. Nature 2015; 526:68–74
5. Karczewski K, Francioli L, Tiao G, Cummings B, Alföldi J, Wang Q, Collins R, Laricchia K, Ganna A, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020; 581:434–43
6. Enattah N, Sahi T, Savilahti E, Terwilliger J, Peltonen L, and Järvelä I. Identification of a variant associated with adult-type hypolactasia. Nature Genetics 2002; 30:233–7
7. Lewinsky R, Jensen T, Møller J, Stensballe A, Olsen J, and Troelsen J. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Human Molecular Genetics 2005; 14:3945–53
8. Tishkoff S, Reed F, Ranciaro A, Voight B, Babbitt C, Silverman J, Powell K, Simon H, Embleton C, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics 2007; 39:31–40
9. Kuokkanen M, Kokkonen J, Enattah N, Ylisaukko-Oja T, Komu H, Varilo T, Peltonen L, Savilahti E, and Järvelä I. Mutations in the translated region of the lactase gene (LCT) underlie congenital lactase deficiency. American Journal of Human Genetics 2006; 78:339–44
10. Gerbault P, Liebert A, Itan Y, Powell A, Currat M, Burger J, Swallow D, and Thomas M. Evolution of lactase persistence: an example of human niche construction. Philosophical Transactions of the Royal Society B: Biological Sciences 2011; 366:863–77
11. Bersaglieri T, Sabeti P, Patterson N, Vanderloecht T, Harris A, Ueda J, Lagerberg R, Barton R, et al. Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics 2004; 74:1111–20
12. Labib K, Tercero J, and Diffley J. Uninterrupted MCM2-7 function required for DNA replication fork progression. Science 2000; 288:1643–7
13. Canver M, Smith E, Sher F, Pinello L, Sanjana N, Shalem O, Chen D, Schinzel A, Tejoujor S, Mikkelsen T, et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 2015; 527:192–7
14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 2001; 29:308–11
15. Dyer SC, Austine-Orimoloye O, et al. Ensembl 2025. Nucleic Acids Research 2024 Dec; 53:D948–D957. DOI: 10.1093/nar/gkae1071. Available from: https://doi.org/10.1093/nar/gkae1071
16. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. GigaScience 2021; 10:giab008
17. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable A, Fang T, Doncheva N, Pyysalo S, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research 2023; 51:D638–D646. DOI: 10.1093/nar/gkac1000
18. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. The Genotype-Tissue Expression (GTEx) project. Nature Genetics 2013; 45:580–5. DOI: 10.1038/ng.2653
19. Rees HA and Liu DR. Base editing: precision chemistry on the genome and transcriptome of living cells. Nature Reviews Genetics 2018; 19:770–88. DOI: 10.1038/s41576-018-0059-1
20. Musunuru K, Chadwick AC, Mizoguchi T, et al. In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates. Nature 2021; 593:429–34. DOI: 10.1038/s41586-021-03534-y
21. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, and Liu DR. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 2019; 576:149–57. DOI: 10.1038/s41586-019-1711-4

Appendices

Bioinformatics Pipeline Scripts

The following scripts were utilized to process VCF files from the 1000 Genomes Project and gnomAD. All analyses were performed using `bcftools` and standard shell scripting.

MCM6 Regulatory Variant Analysis

This script queries the specific genomic coordinates for the known regulatory variants (e.g., rs4988235, rs182549, rs41380347) within the *MCM6* gene. It extracts the variant ID, reference/alternate alleles, and global allele frequency from the annotated VCF.

```
#!/bin/zsh
bcftools query \
-r 2:136608646,2:136616754,2:136608643,2:136608645,2:136608649,\
2:136608651,2:136608745,2:136608746,2:136618834,2:136598443 \
-f '%ID\t%REF\t%ALT\t%AF\n' \
MCM6_annotated.vcf.gz

```

Listing 1: Extraction of significant MCM6 regulatory variants.

Genotype Counting for European Tolerance (rs4988235)

This script iterates through all samples in the 1000 Genomes VCF to count genotypes at the primary European persistence locus (rs4988235). It classifies individuals as "Intolerant" (0|0), "Tolerant Carrier" (0|1 or 1|0), or "Tolerant Homozygous" (1|1).

```
#!/bin/zsh

# Syntax: [%SAMPLE  %GT] loops through all 2504 columns
bcftools query -r 2:136608646-136608646 -f '[%SAMPLE\t%GT\n]' MCM6_b37.vcf.gz > tolerants.txt

echo "Intolerants: $(grep "0|0" tolerants.txt | wc -l)" >  stats-tolerants.txt
echo "Tolerant carrier: $(grep "0|1" tolerants.txt | wc -l)" >>  stats-tolerants.txt
echo "Tolerant carrier: $(grep "1|0" tolerants.txt | wc -l)" >>  stats-tolerants.txt
echo "Tolerant: $(grep "1|1" tolerants.txt | wc -l)" >>  stats-tolerants.txt

cat stats-tolerants.txt

```

Listing 2: Computation of tolerant vs. intolerant genotype counts.

Analysis of LCT Loss-of-Function Variants

The following scripts query the "gene-breaking" variants (nonsense, frameshift) within the *LCT* coding sequence. Listing 3 specifically targets the gnomAD v4 dataset to extract population-specific frequencies. Listing 4 performs a similar query on the GRCh37 dataset.

```
#!/bin/zsh
bcftools query \
-r chr2:135807131,chr2:135794751,chr2:135807214,chr2:135794704,chr2:135794854,chr2:135829858 \
-f 'ID:%ID\tRef:%REF\tAlt:%ALT\tGlobal:%INFO/AF\tEur:%INFO/AF_nfe\tFin:%INFO/AF_fin\tAfr:%INFO/
AF_afr\tLatino:%INFO/AF_amr\tE_Asia:%INFO/AF_eas\tS_Asia:%INFO/AF_sas\tMidEast:%INFO/AF_mid\n' \
gnomad.genomes.v4.1.sites.chr2.vcf.bgz

```

Listing 3: Extraction of LCT "Gene Breaker" statistics from gnomAD v4.

```
#!/bin/zsh
bcftools query \
-r
2:136564701,2:136552321,2:136564784,2:136552274,2:136552424,2:136587428,2:136565147,2:136563636 \
-f 'LCT\t%POS\t%REF\t%ALT\t%AF\n' \
LCT_b37.vcf.gz

```

Listing 4: Querying significant LCT variants from 1000 Genomes (b37).