THE UNIVERSITY OF TEXAS AT EL PASO

Post-Genomic Analysis, vol. 1, pp. 1–6, 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: rsramirezorozc@miners.utep.edu*
3. *Bioinformatics Program, Biological Sciences Department, University of Texas at El Paso, El Paso, 79968, USA*
4. *E-mail: jdgutierrez7@miners.utep.edu*

Abstract
Lactase persistence (LP) represents a quintessential model of human gene-culture coevolution, where environmental shifts towards dairy consumption drove strong positive selection. This project analyzes the genomic architecture of the *LCT/MCM6* locus to map regulatory variants across global populations. Utilizing high-coverage data from the 1000 Genomes Project and gnomAD, we associate specific *MCM6* intronic variants with continental ancestry while differentiating them from rare, pathogenic *LCT* mutations. To verify the specificity of these signals, we examine a control network of secondary genes related to DNA replication (e.g., *ORC4*) and lactose metabolism (e.g., *B4GALT2*), effectively distinguishing adaptive regulatory drivers from background genetic noise. Finally, we synthesize these evolutionary insights to discuss the translational potential of the *LCT* locus as a blueprint for developing novel *in vivo* gene editing therapies to modulate human gene expression.

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

Project Objectives

The primary goals of this study were to:

1. Identify the most common regulatory variants of the *MCM6* gene and associate them with distinct human population structures.
2. Define the consequences of direct mutations within the *LCT* gene coding sequence (loss-of-function).
3. Analyze potential secondary genes (e.g., *B4GALT2*, *GLB1*) to establish functional specificity and rule out pleiotropic effects.
4. Elucidate the influence of environmental and dietary conditions (gene-culture coevolution) on allelic distribution.
5. Relate this regulatory mechanism to emerging techniques for *in vivo* human genome editing.

Introduction

The capacity to digest lactose into adulthood is a defining characteristic of specific human lineages, representing one of the strongest signals of recent natural selection in the human genome [1]. In the ancestral state, the expression of Lactase-Phlorizin Hydrolase (LPH)—the enzyme responsible for hydrolyzing lactose at the brush border of the small intestine—is developmentally downregulated after weaning [2]. This phenotype, adult-type hypolactasia,

renders individuals unable to digest fresh milk. However, the derived phenotype of Lactase Persistence (LP) has arisen independently in populations with a history of pastoralism [3]. The chemical reaction enabling this digestion is shown in Equation 1:

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

This project deconstructs the genomic architecture of this trait by integrating sequencing data from the 1000 Genomes Project [4], the Genome Aggregation Database (gnomAD) [5], and NIH repositories. We aim to distinguish between regulatory silencing (an evolutionary adaptation) and "broken" genes (congenital pathology), establishing this locus as a model for precise gene control.

Genomic Architecture: LCT vs. MCM6

The mechanism of LP is not found within the *LCT* gene itself, but in a regulatory enhancer located 14 kb upstream, within the introns of the adjacent *MCM6* gene [6]. The classic European variant, rs4988235 (C/T_{-13910}), creates a binding site for the Oct-1 transcription factor, recruiting co-factors that upregulate *LCT* promoter activity [7]. Convergent evolution has driven distinct variants in Middle Eastern (e.g., rs41380347) and African populations to achieve the same phenotype [8].

In contrast, direct mutations in the *LCT* coding sequence are rare and pathological. Loss-of-function mutations (nonsense, frameshift) result in Congenital Lactase Deficiency (CLD), a severe condition distinct from adult intolerance [9]. By analyzing these alleles, particularly in isolated populations like Finland, we delineate the boundary between regulatory adaptation and protein disruption.

Methods

We employed a high-throughput genomic data mining approach to characterize the *LCT/MCM6* landscape, distinguishing regulatory silencing from pathological loss-of-function.

Data Acquisition and Processing

Genomic data was acquired from three primary repositories, restricted to **Chromosome 2** to optimize efficiency:

- **1000 Genomes Project (Phase 3):** Used to establish baseline frequencies for major continental groups (2,504 individuals) [4].
- **gnomAD (v4):** Used to capture rare variation and "gene-breaking" alleles in a massive cohort (>800,000 individuals) [5].
- **NIH Repositories:** Used for clinical validation of pathogenic variants [10].

Raw data was processed using a custom Linux pipeline involving **SAMTools** and **BCFtools** [11]. The target region was defined as **chr2:136.5M–136.7M** (GRCh37/hg19) and **chr2:135.7M–135.9M** (GRCh38/hg38) to account for coordinate shifts between assemblies [12]. Custom scripts (see Appendix) were used to extract biallelic SNPs, calculate population frequencies, and cross-reference variants with dbSNP IDs.

Functional Specificity Analysis

To validate that identified signals were specific to lactose regulation and not artifacts of chromosomal replication, we analyzed a control set of genes. This included *MCM6*'s replication partners (*ORC4*, *GINS3*) and metabolic paralogs (*B4GALT2*, *GLB1*). We compared variant density in these loci against the *MCM6* intronic enhancer to prove that selection pressure was isolated to the regulatory element.

Results

Global Stratification of Regulatory Variants

Our analysis of the gnomAD dataset reveals a distinct population-specific stratification of persistence alleles. As detailed in Table 1, the primary European variant rs4988235 ($G > A$) is present at 24% frequency. Crucially, we observed the "backup" variant rs182549 ($C > T$) at an identical frequency, appearing simultaneously with the main driver. This confirms strong linkage disequilibrium, suggesting the preservation of a robust regulatory haplotype rather than a single point mutation.

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

#	Variant / rsID	Abs. Position (b37)	Population	Frequency
1	rs4988235	2:136608646	European (Main)	24 %
2	rs182549	2:136616754	European (Backup)	24 %
3	rs41525747	2:136608643	North African	$< 0.1\%$
4	rs4988233	2:136608645	Ethiopian	$< 0.1\%$
5	rs41456145	2:136608649	Cameroonian	$< 0.1\%$
6	rs41380347	2:136608651	Middle Eastern	0.2 %
7	rs869051967	2:136608745	East African	$< 0.1\%$
8	rs145946881	2:136608746	East African (Main)	0.3 %
9	rs55660827	2:136598443	Rare Coding Variant	19 %

Comparing this to 1000 Genomes data (Table 2) reveals the limitations of smaller cohorts. Rare adaptive alleles found in African subpopulations (e.g., rs41525747) fall below the detection threshold in the smaller dataset, emphasizing the need for massive cohorts like gnomAD to map convergent evolution.

The population heatmaps (Figure 1) further visualize these distributions. Notably, the Finnish population clusters separately from the general European group, reflecting unique genetic isolation. Additionally, the "Latino" category exhibits significant admixture, capturing Native American signals often obscured in standard classifications.

Pathological Loss-of-Function

In contrast to regulatory silencing, direct "gene-breaking" mutations in *LCT* are exceptionally rare. As shown in Table 3, variants like Y1390X (FinMajor) often appear

Table 2: Allele frequencies extracted from 1000 Genomes data analysis (bcftools query results). Note that not all variants are observed due to the limited sample size.

rsID	Ref	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%

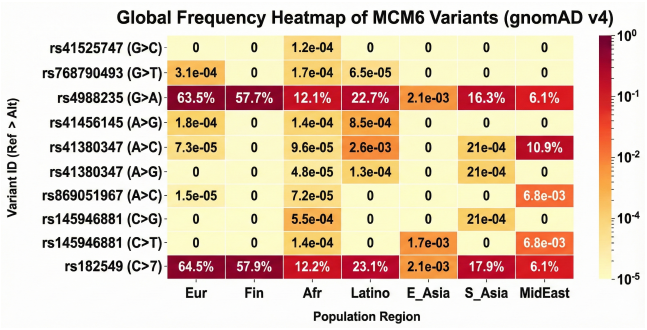

Figure 1: Heatmap of *MCM6* variants across populations. Note the distinct clustering of Finnish (Fin) and Latino groups.

in single individuals within the 800,000-genome dataset. Figure 2 illustrates that these are confined to populations with strong founder effects (e.g., Finland), confirming that biological breakage of the gene is a localized anomaly driven by genetic drift, not selection.

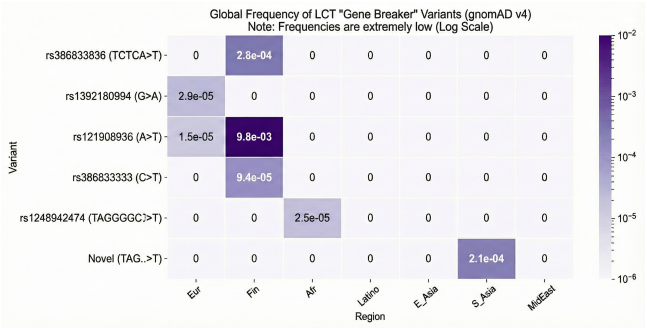

Figure 2: Distribution of *LCT* loss-of-function variants, showing strong founder effects in Finland.

Functional Specificity

To validate specificity, we analyzed the *MCM6* interaction network (Figure 3). Strong correlations with DNA replication machinery (*ORC4*, *GINS3*) confirm that *MCM6*'s primary role is replication. However, our control analysis (Table 4) shows no selection signals in these partners, proving that LP variants act solely on the "moonlighting" enhancer function.

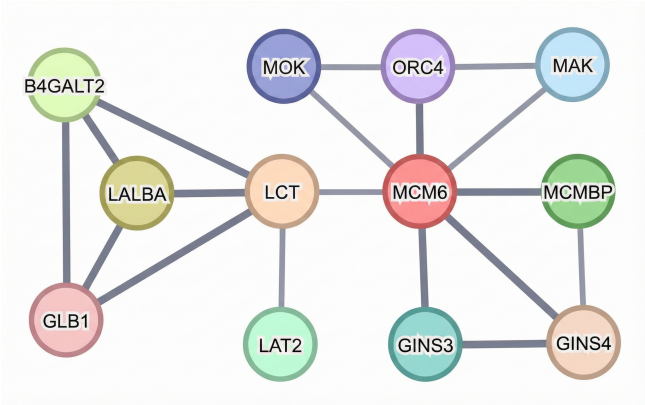

Figure 3: Gene interaction network (STRING-DB) linking *MCM6* to replication machinery [13].

Discussion and Future Directions

This study successfully mapped the global stratification of lactase persistence, confirming that while the phenotype is convergent, the genotype is strictly population-dependent. The identification of the Oct-1 binding motif (Figure 4) provides a clear mechanistic target.

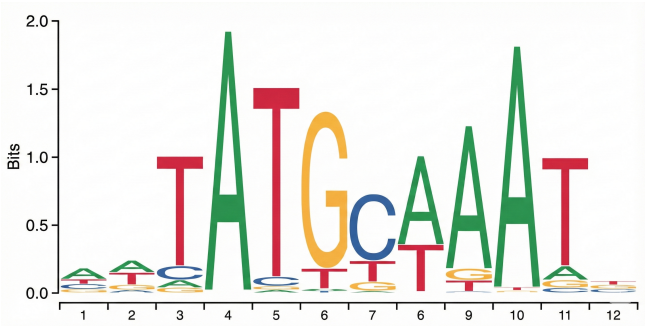

Figure 4: Motif analysis (MEME) showing the creation of an Oct-1 binding site by the T allele.

The natural existence of rs4988235 serves as a proof-of-concept for therapeutic editing: a single base pair change can permanently override epigenetic silencing. While currently theoretical, this locus offers an ideal model for **Base Editing** therapies. Technologies like Adenine Base Editors (ABEs) could potentially install the "persistence" allele in intestinal crypts to treat intolerance [14, 15], while **Prime Editing** could correct the rare congenital frameshifts identified in Finnish populations [16].

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

Table 3: Results from gnomAD after looking for several broken variants of the gene LCT. These are associated with congenital intolerance.

Variant / Name	rsID	Absolute Position (b37)	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 for the regulatory specificity of the *LCT* locus.

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 (oligosaccharides).	The "Cousin." Chemically similar to the enzyme that makes lactose, but it is not involved in digestion.
GLB1	Beta-Galactosidase. Breaks down sugars in the lysosome (cell waste disposal).	The "Backup" (that doesn't help). It performs the exact same chemical reaction as LCT but works inside cells, not in the gut. Mutations cause GM1 Gangliosidosis, not intolerance.
Replication Machinery (MCM6 Interactome)		
MCMBP	MCM Binding Protein. Transports MCM proteins into the nucleus.	MCM6 Protein Partner. Essential for DNA replication. No interaction with the "Milk Switch" in the intron.
ORC4	Origin Recognition Complex. Finds start sites for DNA copying.	MCM6 Loader. It loads the MCM6 protein onto DNA to start replication. Irrelevant to digestion.
GINS3	GINS Complex Subunit 3. Part of the DNA helicase motor.	MCM6 Engine Part. Locks onto MCM6 to help unzip DNA. Irrelevant to digestion.
GINS4	GINS Complex Subunit 4. Part of the DNA helicase motor.	MCM6 Engine Part. Same as above; functional partner of the MCM6 protein.
MAK	Male Germ Cell Associated Kinase. Regulates cilia and cell cycle.	Cell Cycle Network. Co-expressed with replication genes. No link to digestion.
MOK	MAPK/MAK/MRK Overlapping Kinase. Kinase involved in cell regulation.	Cell Cycle Network. Likely appears in your list because it interacts with the machinery MCM6 is part of.
LAT2	Linker for Activation of T Cells 2 (or Amino Acid Transporter).	Unrelated. Likely appears due to extremely rare conditions.

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. 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

11. 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

12. 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

13. 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

14. 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

15. 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

16. 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.

```
#!/bin/zsh
bcftools query \
-r
2:136608646,2:136616754,2:136608643,2:136608645,2:136608649,2:136608651,2:136608745,2:136608746,2:13661883
 \
-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.

```
#!/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 within the *LCT* coding sequence.

```
#!/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).