A Computational Approach to Discover Protein Formations of Fate: Investigating the Role of the HNF1A Mutation in MODY3

Sophia Wang and Jonah Bendell

Received October 24, 2025
Accepted November 22, 2025
Electronic access November 30, 2025

Genomic variations play a crucial role in determining disease susceptibility, and even small mutations can alter protein structure and cellular function. Unlike traditional experimental methods, this study uses an innovative computational approach to investigate genetic variants from human exome sequencing data, focusing on the HNF1A gene mutation associated with maturity-onset diabetes of the young type 3 (MODY3). Using data from the 1,000 Genomes Project (sample SRR701471), we aligned sequences with the Burrows-Wheeler Aligner (BWA), performed genotyping with SAMtools mpileup, and annotated variants using ANNOVAR, all on Midway3, a supercomputer at the University of Chicago. From over 112,000 high-quality variants, we identified 15 disease-annotated mutations, including a missense variant in HNF1A (serine to glycine). Protein modeling with AlphaFold and visualization in Visual Molecular Dynamics (VMD) revealed that this substitution causes a structural displacement of 18.34 Å between α-helices. This study demonstrates a computational workflow for variant identification and structural modeling, while highlighting the critical need for database validation. The HNF1A variant identified (rs1169305) was subsequently confirmed as a common benign polymorphism, underscoring that computational predictions must be validated against clinical databases before drawing pathogenicity conclusions. Although the computational confidence of this study is moderate, this approach highlights the power of genomic annotation and protein modeling in understanding disease mechanisms and guiding personalized treatment strategies for MODY3.

Keywords: computational genomics, HNF1A mutation, MODY3, exome sequencing, variant annotation, protein modeling, personalized medicine

Introduction

In cellular regulation, form is essential. Cellular regulation ensures that a cell maintains its proper form and function; shifts undetectable even by the world's most powerful microscopes in proteins, signaling pathways, or gene expression can disrupt this balance, ultimately leading to disease. Traditional biological experiments, such as gene cloning, PCR, Western blotting, and X-ray crystallography, are time-consuming, expensive, and limited in scale, making them unsuitable for analyzing the vast datasets generated by modern high-throughput sequencing. By integrating computer science, mathematics, and statistics, computational biology has become a powerful tool that has transformed our understanding of biology and disease. This study applies a computational approach to investigating a human's genomic information and their genetic diseases. Our approach exemplifies the revolutionary concept of personal genomics, which involves analyzing an individual's genetic makeup to gain a deeper understanding of their health and potential risks. We employed exome sequencing techniques[1], which provide confident data by targeting protein-coding regions. This technique uses new sequence technology that chooses DNA fragments complementary to known exonic sequences and focuses on the 1%–2% of the genome where 85% of disease-causing mutations occur. The sample analyzed (SRR701471) is from the 1,000 Genomes Project[2,3], a population genetics resource containing healthy individuals from diverse ethnic backgrounds. The provided patient exhibits 120,000 raw variations in their entire genome, and we chose to investigate one variation for in-depth analysis. This paper focuses on the mutation in the HNF1A gene[4], a transcription factor that is crucial for pancreatic β-cell function. Mutations in HNF1A are known to cause maturity-onset diabetes of the young type 3 (MODY3)[5,6], a disease sensitive to sulfonylureas. We investigated this gene through alignment, genotyping, and annotation to determine how the gene functions, to identify the mutation, and to understand its impact on individuals. It is essential to note that the presence of an HNF1A variant in population databases does not confirm disease expression. The individual sequenced (SRR701471) is from a healthy population cohort and may not exhibit MODY3 phenotype.

© The National High School Journal of Science 2025

Methodology

We analyzed genomic sequencing data from sample SRR701471, a healthy control individual from the 1,000 Genomes Project, on a high-performance computing cluster (Midway3)[7] hosted by the UChicago Research Computing Center (RCC). We applied the following steps in our research:

1. Obtain Data
   FASTQ files are typically used to store raw sequencing reads, and they contain both the nucleotide sequences and the quality scores that indicate the confidence of each base. The genome was delegated to us as two FASTQ[8] files ($SRR701471_1$.fastq and $SRR701471_2$.fastq). $SRR701471_1$.fastq contained the forward read, while $SRR701471_2$.fastq contained the reverse read.

2. DNA and RNA Sequences Alignment
   We utilized Burrows-Wheeler Aligner version 0.7.17-r1188 (BWA)[9], a software package for aligning short DNA and RNA sequences[10], to map sequencing reads to a reference genome database, such as the Human Genome Project reference. This software first creates an FM-index of the reference genome, which is based on the Burrows-Wheeler Transform, and then uses this index to map the reads quickly and efficiently. BWA has three main algorithms: BWA-backtrack, BWA-SW, and BWA-MEM. We used BWA-backtrack to identify the optimal placement of each read and to account for matches, mismatches, and gaps relative to the reference genome, reflecting the standard pipeline used for short-read data in the 1000 Genomes Project. The primary output of the BWA is a file in the Sequence Alignment/Map (SAM) format.

3. Variant Identification and Genotyping
   We employed mpileup version 1.6[11] for genotyping, which generates base-level quality scores for each nucleotide in the sequence. This tool assesses the sequence alignment data file, evaluates the sequencing coverage at every base position, and calculates Phred quality scores[12,13], which quantify the probability that a given base is incorrectly called. Higher Phred scores correspond to greater confidence in the accuracy of the base call. These scores assess whether a particular nucleotide is reliable for downstream analyses, such as variant identification and genotyping. The outputs are variant call format (VCF) files used in genomics to store genetic variation data.

4. Variant Quality Control and Filtering
   Following alignment with BWA-backtrack and variant calling with SAMtools mpileup, we applied quality filtering based on Phred scores. Variants with Phred quality scores >50 were retained, corresponding to $>99.999\%$ base-calling confidence because we chose to prioritize specificity (confidence) over sensitivity (variant yield), since 3D protein modeling of HNF1A required exceptionally accurate variant identification. This filtering reduced 156,432 raw variant calls to 112,076 high-confidence variants. Subsequent ANNOVAR annotation identified 14,719 variants in exonic regions. We did not apply any other filtering methods since the high Phred threshold already ensured exceptional base-calling quality, and visual inspection of coverage in IGV confirmed adequate depth across analyzed regions. Hardy-Weinberg equilibrium testing was not applicable as this analysis examined a single individual rather than a population.

5. Sequence Annotation
   Genetic variant annotation (ANNOVAR)[14], downloaded March 2023, was used to process the variant call format (VCF) files. ANNOVAR is an efficient software tool that functionally annotates genetic variants from significant amounts of sequencing data produced from high-throughput sequencing platforms. This tool converts the VCF file into ANNOVAR's input format and annotates the variants using multiple reference databases. Annotations include known variants, allele frequency information, and pathogenic mutations.

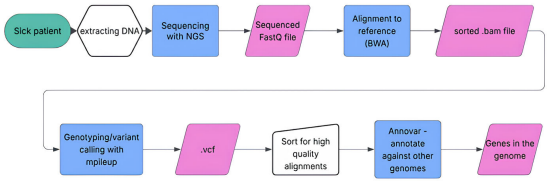

Fig. 1 Workflow overview

Figure 1 offers a concise overview of our workflow that we utilized to analyze this person's genome.

Results

From our alignment, we identified 112,076 high-quality variants (Phred Quality Score * $>$ 50) in our sequence. The genome contained 14,719 exonic variants and 174 indel variants, resulting from those high-quality mutations. Out of the

$*$ A Phred Quality Score is a measure of the accuracy of base calls made during DNA sequencing. A score of 50 or higher represents 99.999% accuracy or 1 base in 100,000 has an error during sequencing.

© The National High School Journal of Science 2025

exonic mutations, we identified 7,109 synonymous variants, 5,905 nonsynonymous variants, 47 frameshift variants, 72 insertion variants, 15 frameshift insertions, 57 non-frameshift insertions, 101 deletion variants, 32 frameshift deletions, and 69 non-frameshift deletions, displayed in Table 1 below: Each

Variant Type	Count	Description
Exonic Variants	14,719	Total variants in protein-coding regions
• Synonymous SNVs	7,109	Silent mutations
• Nonsynonymous SNVs	5,905	Missense mutations
• Stopgain SNVs	55	Nonsense mutations
• Stoploss SNVs	7	Loss of stop codon
• **Insertions (total)**	72	All insertion events
• Frameshift insertions	15	Indels not divisible by 3
• Non-frameshift insertions	57	In-frame insertions (multiples of 3)
• **Deletions (total)**	101	All deletion events
• Frameshift deletions	32	Indels not divisible by 3
• Non-frameshift deletions	69	In-frame deletions (multiples of 3)
• **All Frameshifts**	47	Combined frameshift events
• Unknown	1,250	Unknown/non-coding RNA
Intronic Variants	69,456	Non-coding regions
UTR/Regulatory/Intergenic	27,901	5' UTR, 3' UTR, upstream, downstream, intergenic, and other non-coding regulatory regions
Total High-Quality	112,076	All variants with Phred >50

Tab. 1 A list of the number of the different type of variants identified

high-quality variant had a quality score greater than 50, which ensures a low chance of mismatches. We then selected 15 specific variations in the genome to focus on, as presented in Table 2. These 15 variations were nonsynonymous SNVs or indels in exonic regions, documented in at least one disease, and noted as pathogenic in the annotation. From these 15 mutations, we then selected three particularly intriguing ones to research their implications and diseases further, using their Reference SNP† (rs) Report[15]. We ensured that each mutation was on varying chromosomes. Table 3 introduces our findings.

Figures 3 and 4 demonstrate that the number of genes and the length of a chromosome are directly proportional to the number of mutations, since more nucleotide bases have a possibility of changing or mutating. Linear regression analysis demonstrated that mutation count scales proportionally with chromosome length (slope (β_1) = 20,399 bp per mutation, 95% confidence interval(CI): 12,120–28,679, $R^2 = 0.56$, $p < 0.0001$). This indicates that mutations are distributed across the genome in proportion to chromosome size, with chromosome length explaining 56% of the variance in mutation counts (Figure 3). The number of mutations per chromosome also increased linearly with gene count ($\beta_1 = 4.84$ mutations per gene, 95% CI: 4.38–5.30, $R^2 = 0.96$, $p < 0.0001$), suggesting that genomic regions with higher gene density accumulate proportionally more variants (Figure 4). Figures 5, 6, 7, and 8 compare the number of exonic mutations versus intronic mutations for each chromosome and for each gene. Each chromosome presents more intronic variants than exonic variants, which is reasonable since intronic regions are more prevalent. The number of genes is also directly related to the number of intronic and exonic mutations, with a high coefficient of determination. Exonic mutations scaled strongly with the number of genes per chromosome ($\beta_1 = 0.49$ genes per exonic muta-

† A genetics database for human single nucleotide variations that tracks frequency across populations and effectn

Location of variant (Phred cut off = 50)	Type of variant	Implications
Chromosome 1 position 100206504	nonsynonymous SNV	Intermediate maple syrup urine disease type 2
Chromosome 1 position 114679616	nonsynonymous SNV	Muscle AMP deaminase deficiency
Chromosome 1 position 145927447	non-coding (affects expression)	Radial aplasia-thrombocytopenia syndrome
Chromosome 5 position 74685445	nonsynonymous SNV	Sandhoff disease, infantile type
Chromosome 5 position 177093242	nonsynonymous SNV	Cancer progression and tumor cell motility
Chromosome 7 position 117590400	nonsynonymous SNV	Congenital bilateral absence of the vas deference - Cystic fibrosis
Chromosome 9 position 133436862	nonsynonymous SNV	Upshaw-Schulman syndrome
Chromosome 11 position 18269312	nonsynonymous SNV	Serum amyloid A variant
Chromosome 11 position 48123823	nonsynonymous SNV	Carcinoma of colon
Chromosome 11 position 112086941	nonsynonymous SNV	Cowden syndrome 3
Chromosome 12 position 120999579	nonsynonymous SNV	Maturity-onset diabetes of the young, type 3
Chromosome 14 position 21321881	nonsynonymous SNV	Cone-rod dystrophy 13
Chromosome 16 position 27344882	nonsynonymous SNV	Atopy, resistance to acquired immunodeficiency syndrome low progression
Chromosome 16 position 30086309	synonymous SNV	Spondylocostal dysostosis 5
Chromosome 19 position 12899706	nonsynonymous SNV	Glutaric aciduria, type 1

Fig. 2 15 mutations present in the sequence and their implications.

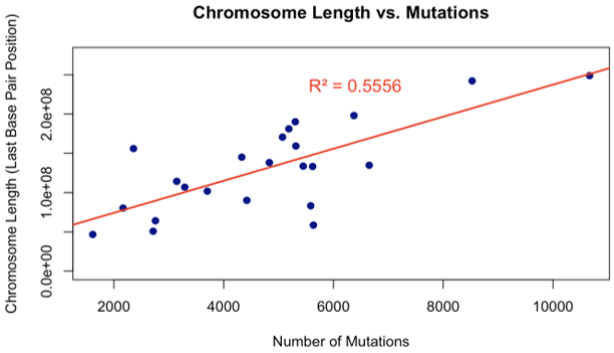

Fig. 3 Scatter plot comparing chromosome length vs. number of mutations on that chromosome

Gene and Location	Effect	Alleles/Frequency	Differences
Cowden Syndrome[16] Chromosome 11 Position 112086941 Gene: *SDHD*	A nonsynonymous, single-nucleotide mutation in the *SDHD* gene, a tumor suppressor that regulates cell division, leads to increased cell growth. It includes symptoms such as skin lesions, noncancerous growths, and an increased risk for cancer[17].	Latin Americans have the highest alternative gene (1.44%), while Africans have the lowest alternative gene (0%) . Globally: 0.9334%	Alleles = G > A / G > C Amino Acid = Glycine to Serine rsId = rs34677591
Maturity-onset Diabetes of the Young: Type 3 (MODY3) Chromosome 12 Position 120999579 Gene: *HNF1A*	A missense (nonsynonymous) single-nucleotide variant in the *HNF1A* gene. Although it is classified as benign, the mutation is contextually associated withMODY3 which causes reduced insulin production due to its inability to help the pancreas produce insulin, especially in young adults. Symptoms include thirst, urination, fatigue, blurred vision, and eventually, kidney disease/benign liver tumors.	Asians have the highest alternative gene (100%), while Africans have the lowest alternative gene (95.2%). Globally: 99.337% The high frequency suggests G is the common allele; A is rare.	Alleles = A > C / A > G / A > T Amino Acid = Serine to Glycine rsID = rs1169305
Maple Syrup Urine Disease Chromosome 1 Position 100206504 Gene: *DBT*	A nonsynonymous single-nucleotide mutation in the *DBT* gene, which produces an enzyme that helps break down certain proteins, causes a buildup of harmful substances in the blood and urine due to the body's inability to break down some amino acids. MSUD can cause symptoms such as poor feeding, vomiting, seizures, and even brain damage[18].	Europeans have the highest alternative gene (90.925%), while Africans have the lowest alternative gene (73.98%). Globally: 90.165%	Alleles = T > A / T > C Amino Acid = Serine to Glycine rsID = rs12021720

Tab 3. Three specific diseases, their mutation, and frequency

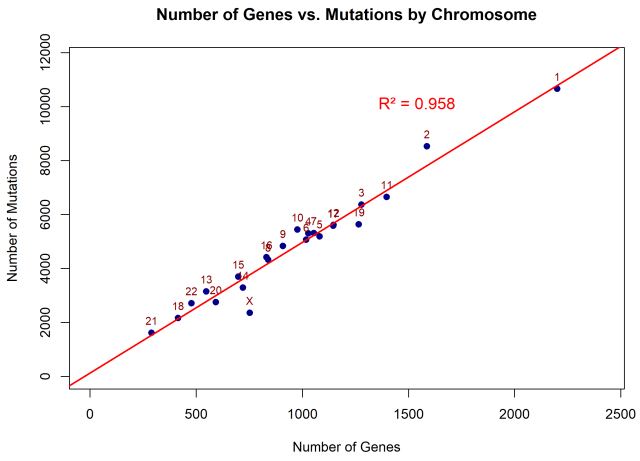

Fig. 4 Scatter plot comparing number of genes vs. number of mutations per chromosome.

tion, or equivalently, 2.04 exonic mutations per gene, 95% CI: 0.46–0.52, R^2 = 0.98, p < 0.0001). The extremely high R^2 value (0.98) indicates that gene count is an excellent predictor of exonic variant burden (Figure 5). Intronic mutations also correlated with gene count (β_1 = 0.19 genes per intronic mutation, or equivalently, 5.36 intronic mutations per gene, 95% CI: 0.17–0.21, R^2 = 0.95, p ¡ 0.0001). The higher ratio of intronic to exonic mutations reflects the larger size of intronic regions in the human genome (Figure 6). Figure 10 graphs the percent of exonic mutations per chromosome, showing no specific correlation between chromosome and exonic mutations. The percentage of variants classified as exonic showed no significant relationship with chromosome number (β_1= 0.15% per chromosome unit, 95% CI: -0.05 to 0.35, R^2 = 0.10, p = 0.13). This indicates that the proportion of exonic variants is relatively uniform across chromosomes (mean 13%), with no chromosome-specific enrichment or depletion. Figure 11 illustrates the distribution of various types of mutations. Figure 12 exhibits that, although most are synonymous, the number of nonsynonymous ones exceeds the average. Supposedly, "between one-quarter and one-third of point mutations in protein-coding DNA sequences are synonymous"[19]. Table 4 demonstrates a Chi-square test to test

© The National High School Journal of Science 2025

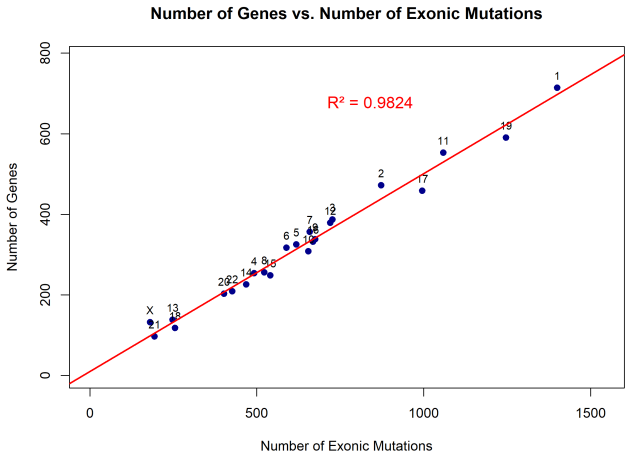

Fig. 5 Scatter plot comparing genes vs. exonic mutations

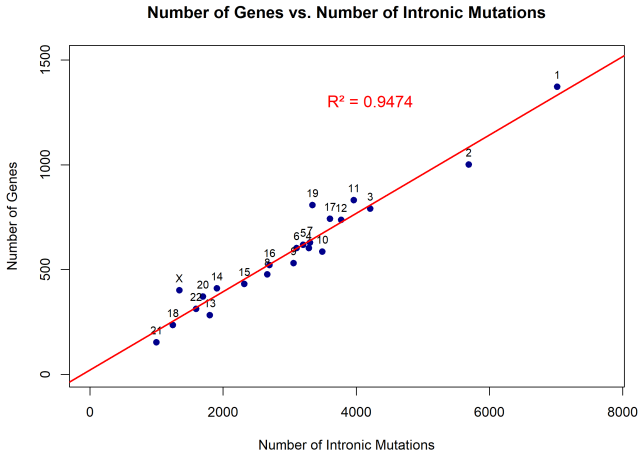

Fig. 6 Scatter plot comparing genes vs. intronic mutations.

whether the observed ratio is statistically significant.

$x^2 = 3,794$, degrees of freedom = 1, $\mathbf{p < 0.0001}$

The observed ratio (54.7% synonymous) significantly exceeds the expected 25-33% ($p < 0.0001$). Some possible explanations include that nonsynonymous mutations may be under-represented due to negative selection in the human population, our filtering criteria (Phred >50, depth filters) may have differentially affected the two categories, or the expected ratio from Shen et., al.[19] was derived from experimental mutation accumulation studies, which may not reflect standing variation in human populations subject to selection.

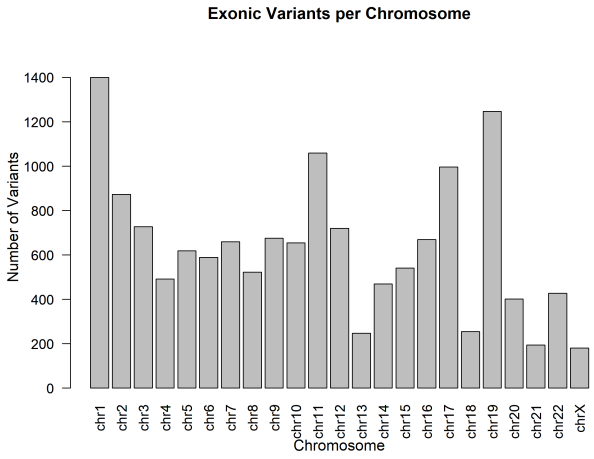

Fig. 7 Bar graph showing exonic variants per chromosome

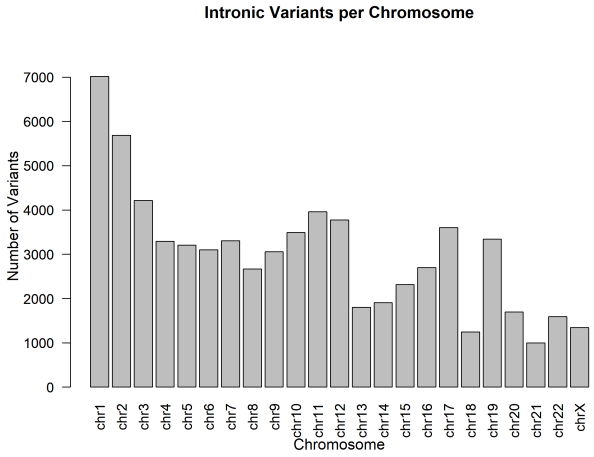

Fig. 8 Bar graph showing intronic variants per chromosome.

Category	Observed	Expected (30% syn)	$(O-E)^2/E$
Synonymous	7,109	3,912	2,656
Nonsynonymous	5,905	9,129	1,138

Fig. 9 Chi-square test for supposed synonymous mutations

Discussion

Among the various types of variations, one of the most interesting was a mutation in the HNF1A gene that is historically related to maturity-onset diabetes of youth, abbreviated as MODY3. The HNF1A variant rs1169305 (c.1720A>G) has been submitted to ClinVar in the context of Maturity-Onset Diabetes of the Young type 3 (MODY3), with historical annotations suggesting a missense change (p.Ser574Gly).

© The National High School Journal of Science 2025

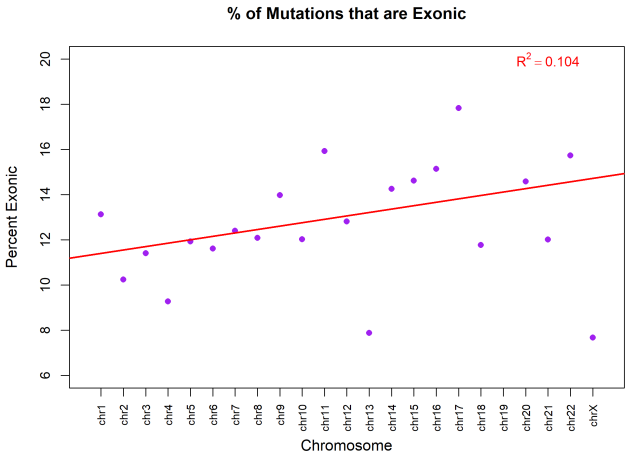

Fig. 10 Scatter plot showing the percentage of mutations per chromosome.

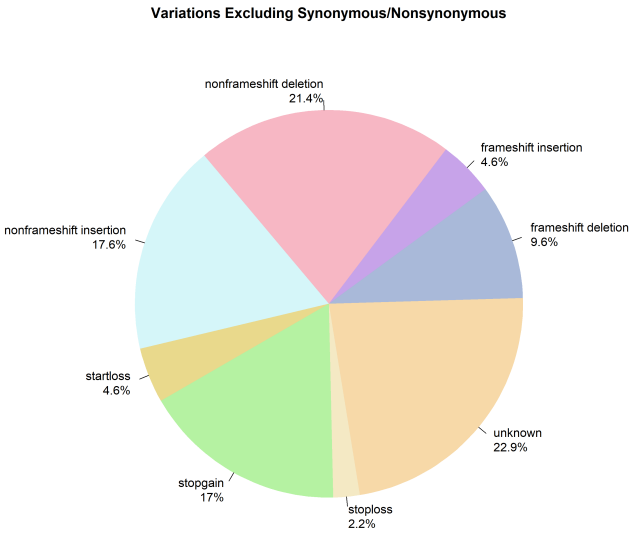

Fig. 11 Pie chart showing the percentages of all types of variants.

However, based on the ClinVar database, this variant is now classified as benign, and its high allele frequency in population databases further supports non-pathogenicity. The variant shows near-fixation in Asian populations (100%) and a global frequency of 99.3%, indicating that the G allele represents the ancestral, common variant. The reference genome's A allele at this position (0.663% globally) reflects the arbitrary nature of reference sequences, which are derived from a small number of individuals and do not necessarily represent the most common allele in human populations. This finding underscores the critical importance of validating computational predictions against population frequency databases and clinical variant repositories before making pathogenicity claims. Its high allele frequency in population databases further supports non-pathogenicity. MODY3 is a hereditary form of diabetes that stems from mutations in the HNF1A gene. While the HNF1A gene typically makes a protein (hepatocyte nuclear factor-1 alpha) that helps pancreatic cells detect sugar and release insulin, the HNF1A mutations produce faulty proteins that cannot properly activate genes necessary for glucose sensing and insulin production in pancreatic cells. As a result, these mutations inhibit the cells from detecting high blood sugar and releasing adequate insulin. The primary defect is early-onset diabetes due to progressive pancreatic failure. If such an individual's blood sugar is not well controlled, they can develop serious complications, including kidney damage, eye problems, nerve damage, and occasionally liver tumors.

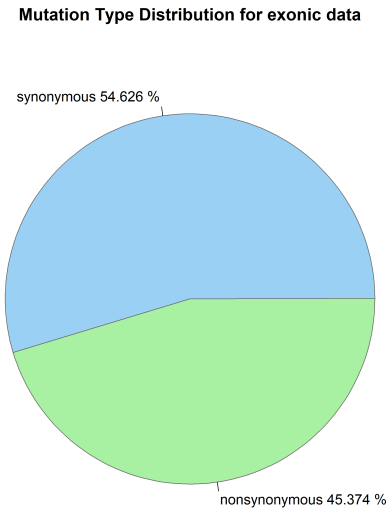

Fig. 12 Pie chart showing percentage of synonymous vs nonsynonymous mutations.

An example of a case study about MODY3 involved two different patients[20]:

- Patient 1: a 23-year-old man with a start codon mutation in the gene, which prevented him from producing the protein.
- Patient 2: a 19-year-old woman with a frameshift mutation in the gene; she produced a broken, shortened pro-

© The National High School Journal of Science 2025

tein that was nonfunctional.

Patient 1's diabetes was caught early on, and once doctors identified his genetic type, he responded well to diabetes pills and maintained good health with no complications. However, Patient 2 had suffered for 7 years without proper treatment and had already developed serious nerve and kidney damage before finally receiving insulin therapy. The difference in their quality of life underscores the importance of early genetic testing, especially for MODY3 patients. Unlike regular diabetes, this genetic variation responds much better to specific medications, but only under the circumstance that doctors exactly know what kind of disease they are treating.

Next, we modeled the gene using AlphaFold v2.3.2[21] and visualized it using VMD to obtain a better understanding of this mutation (Figure 13).

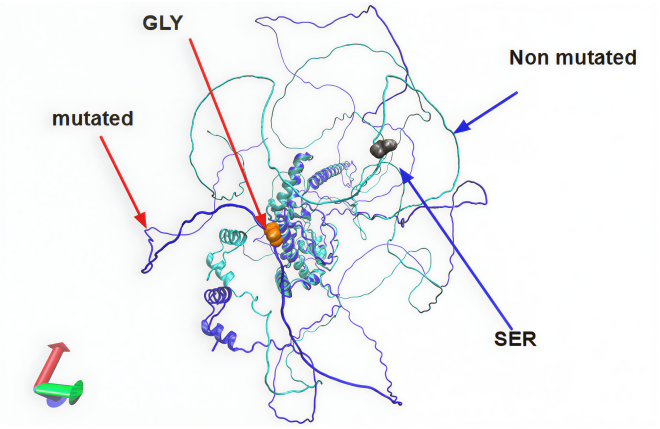

Fig. 13 Aligned proteins: mutated and non-mutated compared.

The dark blue model is the mutated model, and the light blue is the normal gene. Since the gene has not been experimentally analyzed yet, both models were created by AlphaFold as a hypothesis of its structure. Both models are aligned with each other. The orange is glycine, and the gray is serine. While the structures seem similar, the following key differences were identified:

1. The entire gene has drifted outward, revealing a weaker structure.
2. Shifting dynamics affect the interactions the protein would have otherwise.
3. The full shifting of the gene could indicate unstable structure and disrupted binding sites.

The variant identified in this study (p.Ser574Gly) is located at residue 574 within the C-terminal transactivation domain, approximately 300 residues downstream from the DNA-binding region. This domain is responsible for recruiting transcriptional co-activators rather than direct DNA contact. Figure 14 demonstrates the extent of this shift on the shape of the protein. The offset between these two α-helices is 18.34Å, which

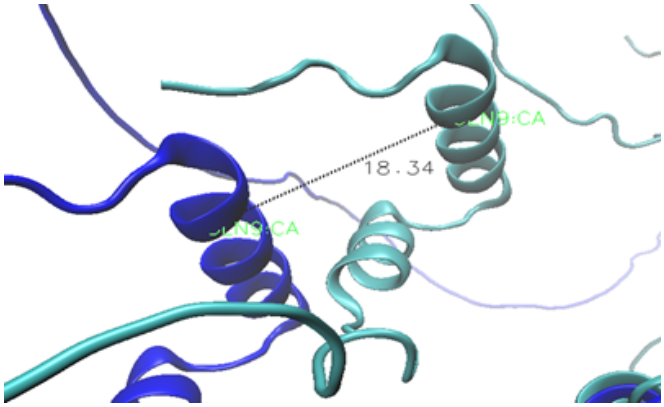

Fig. 14 Mutated and non-mutated proteins compared up close.

is significant in a molecule because of its small size (Figure 12). However, given the low model confidence and the variant's location outside the DNA-binding domain, we cannot conclude that this displacement directly affects DNA binding or causes disease, but it does change the structure of the protein. AlphaFold modeling of the wild-type and variant HNF-1α proteins yielded an overall confidence score (pLDDT) of approximately 53. While this represents low confidence overall by AlphaFold standards (scores >90 = high, 70-90 = good, 50-70 = low), it is important to note that confidence varies substantially across different protein regions. Critically, residue 574 is located within a predicted α-helix, a regular secondary structure element that AlphaFold models with greater reliability than disordered regions. We therefore present this structural analysis as a hypothesis-generating approach that could guide future experimental work, rather than definitive evidence of pathogenicity. The observed 18.34 Å displacement represents a testable prediction that requires experimental validation.

Limitations

There are few limitations of our study, ranging from ethical to computational. While the overall AlphaFold confidence score was moderate (53), the variant residue (574) is located within an α-helix, a structured region where AlphaFold predictions are more reliable (65-75). Nevertheless, the predicted structural displacement represents a computational hypothesis requiring experimental validation. The overall shape is present, but the turns still require investigation. In addition, this model is a snapshot of a dynamic structure that continuously changes,

which could amplify the differences between the two models. AlphaFold models are computational predictions, not experimentally determined structures.

Furthermore, sequencing errors during replication are a potential limitation of sequencing by next-generation sequencing (NGS)[22]. The Phred score, although used as a filter to include only mutations with a score larger than 50, still indicates a probability that the base calls are unreliable, which leads to false variant identification. Sequencing errors during replication could potentially misclassify benign variants as pathogenic or miss MODY3-causing mutations[23].

Personal genomics is also a significant ethical concern[24]. Genetic information that reveals a disposition to diabetes also affects the family of the patient since they share the same inheritance pattern; there are contradicting opinions on the publication of such data. Furthermore, misuse of data could lead to genetic discrimination and potentially endanger the patient[25].

Some downsides of genomic studies are that the data only presents the risk of having a disease without certainty, or data can have low coverage and still remain uncertain. A major caveat to studying human genomes is the legal complexities that accompany it. On one hand, personal genomics is the gateway to the center of a human's data and could be used for commercial purposes, but on the other hand, understanding this level of the disease would be game-changing in a clinical setting. For example, doctors could provide personalized healthcare, such prescribing sulfonylureas to support the protein in the action of releasing insulin instead of simply prescribing insulin[26].

Conclusion

Despite these limitations, our study provides valuable insights into the structural and functional impacts of the HNF1A mutation, demonstrating how computational modeling and genomic analysis can inform personalized treatment strategies. From the first base pair to 3 billion, a total of 120,000 variants is present within this sample human genome, 15 of which were associated with unique diseases. We narrowed these variants to research a mutation in the HNF1A gene, which is analyzed in relation to MODY3 as a proof-of-concept for structural modeling. This condition is sensitive to sulfonylureas and can be treated effectively if identified in a clinical setting. Our study explores new ground for this protein by visually examining its mutations, which allows the structural differences to be observed. Although computational confidence was moderate, this approach presents a proof-of-concept computational pipeline for genome-scale variant identification, annotation, and structural modeling for analyzing the source of a genetic disease.

Data Availability Statement

Raw sequencing data are publicly available from the 1,000 Genomes Project (SRR701471), or at https://www.ncbi.nlm.nih.gov/sra/SRR701471. Processed data (VCF files, annotated variants) and analysis scripts in R are available at `https://github.com/sophia0805/BIOS10007-Final`. The software versions used are as follows: BWA v0.7.17-r1188, SAMtools v1.6, ANNOVAR downloaded 03-2023, AlphaFold v2.3.2, and R v4.2.1. We used GRCh37/hg19 (hs37d5) as the human reference genome. All analyses were performed on Midway3 cluster (University of Chicago RCC). Both wild-type and mutant (p.Ser574Gly) HNF-1α protein structures were predicted using AlphaFold v2.3.2. Due to file size limitations, the original PDB model files are not permanently hosted, but can be fully reproduced using the AlphaFold pipeline (`https://github.com/deepmind/alphafold`) with the HNF1A protein sequence (UniProt: P20823, isoform 1) for wild-type, and the same sequence modified at position 574 (Ser to Gly) for the mutant model. The input sequences and AlphaFold parameters used are provided in the GitHub repository to ensure full reproducibility.

References

1 *Automated DNA Sequencing and Analysis*, ed. M. Adams, C. Fields and J. C. Ventor, Academic Press, London, San Diego, 1994.

2 The 1000 Genomes Project Consortium, *Nature*, 2015, **526**, 68–74.

3 O. Devuyst, *Peritoneal Dialysis International*, 2015, **35**, 676–677.

4 Y. Miyachi, T. Miyazawa and Y. Ogawa, *International Journal of Molecular Sciences*, 2022, **23**, 3222.

5 Q. Wen, Y. Li, H. Shah, J. Ma, Y. Lin, Y. Sun and T. Liu, *Open Medicine*, 2023, **18**, 20230705.

6 C. Bellanné-Chantelot, D. J. Lévy, C. Carette, C. Saint-Martin, J. P. Riveline, E. Langer, R. Valéro, J. F. Gautier, Y. Reznik, A. Sofa, A. Hartemann, S. Laboureau-Soares, M. Laloi-Michelin, P. Lecomte, I. Chaillous, D. Dubois-Laforgue, J. Timsit and French Monogenic Diabetes Study Group, *The Journal of Clinical Endocrinology and Metabolism*, 2011, **96**, E1346–E1351.

7 Research Computing Center, University of Chicago, *Midway3*, `https://rcc.uchicago.edu/midway3`, 2021, Accessed: 2025-11-25.

8 P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer and P. M. Rice, *Nucleic Acids Research*, 2010, **38**, 1767–1771.

9 H. Li and R. Durbin, *Bioinformatics*, 2009, **25**, 1754–1760.

10 K. M. Robinson, A. S. Hawkins, I. Santana-Cruz, R. S. Adkins, A. C. Shetty, S. Nagaraj, L. Sadzewicz, L. J. Tallon, D. A. Rasko, C. M. Fraser, A. Mahurkar, J. C. Silva and J. C. Dunning Hotopp, *Microbial Genomics*, 2017, **3**, e000122.

11 P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies and H. Li, *GigaScience*, 2021, **10**, giab008.

12 B. Ewing, L. Hillier, M. C. Wendl and P. Green, *Genome Research*, 1998, **8**, 175–185.

13 B. Ewing and P. Green, *Genome Research*, 1998, **8**, 186–194.

14 K. Wang, M. Li and H. Hakonarson, *Nucleic Acids Research*, 2010, **38**, e164.

© The National High School Journal of Science 2025

15 S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin, *Nucleic Acids Research*, 2001, **29**, 308–311.
16 M. Magaña, A. P. Landeta-Sa and Y. López-Flores, *American Journal of Dermatopathology*, 2022, **44**, e1–e6.
17 S. Kadiyala, Y. Khan, V. Miguel, M. N. Frone, F. Nwariaku, J. Rabaglia, S. Woodruff, E. E. King, S. S. Hathiramani, K. Pacak and H. K. Ghayee, *AACE Clinical Case Reports*, 2018, **4**, 186–190.
18 A. V. B. Margutti, W. A. Silva and D. F. Garcia, *Orphanet Journal of Rare Diseases*, 2020, **15**, 309.
19 X. Shen, S. Song, C. Li and J. Zhang, *Nature*, 2022, **606**, 725–731.
20 Y. Zhang, Y. Liu and Y. Wang, *Open Medicine*, 2023, **18**, 20230749.
21 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler and D. Hassabis, *Nature*, 2021, **596**, 583–589.
22 H. Satam, K. Joshi, U. Mangrolia, S. Aghoo, G. Zaidi, S. Rawool, R. P. Thakare, S. Banday, A. K. Mishra, G. Das and S. K. Malonia, *Biology*, 2023, **12**, 997.
23 M. Tosur and I. H. Philipson, *Journal of Diabetes Investigation*, 2022, **13**, 1465–1471.
24 J. Mathaiyan, A. Chandrasekaran and S. Davis, *Perspectives in Clinical Research*, 2013, **4**, 100–104.
25 K. B. Brothers and M. A. Rothstein, *Personalized Medicine*, 2015, **12**, 43–51.
26 K. I. Kendig, S. Baheti, M. A. Bockol, T. M. Drucker, S. N. Hart, J. R. Heldenbrand and V. Dinu, *Frontiers in Genetics*, 2019, **10**, 736.

© The National High School Journal of Science 2025