RESEARCH PAPER

Computational Identification And Distribution Of Mirror Repeats In Transcripts Of Human PAH Gene

Laxita Singh[1], Namrata Dangi[2]*

[1]Ph.D. Scholar, Department of Zoology, Baba Mast Nath University, Asthal Bohar, Rohtak,124021, Haryana, India, Singhlaxita.21@gmail.com,

[2*]Assistant Professor, Baba Mastnath University, Asthal Bohar, Rohtak, 124021, Haryana, India, namrataphdzoology@gmail.com, Orcid id : 0000-0001-9029-6332

ABSTRACT

Repetitive DNA elements contribute significantly to genome organization and structural complexity. Among these, mirror repeats represent a distinct class of symmetric sequences capable of forming noncanonical DNA structures, yet their distribution in clinically important human genes remains poorly characterized. In this study, we performed a comprehensive computational analysis to identify and map mirror repeats within human phenylalanine hydroxylase (PAH) gene transcripts. Transcript sequences were retrieved and analyzed using a BLAST-based Fast Pairwise Complementary BLAST (FPCB) approach to enable sensitive detection of symmetric DNA motifs. Mirror repeats were systematically identified, classified as perfect or imperfect, and mapped across the coding sequence and exon regions.

Fragment-wise analysis of the PAH coding sequence identified a total of 86 mirror repeats, while exon-wise curation of the canonical transcript revealed 63 unique mirror repeats localized within coding exons. The detected mirror repeats exhibited a non-uniform distribution across coding fragments and exons. This study provides the first detailed descriptive catalog of mirror repeats in PAH gene transcripts and establishes a reproducible computational framework that can serve as a reference for future studies of symmetric DNA motifs.

KEYWORDS: *Mirror repeats, Symmetric DNA sequences, Phenylalanine hydroxylase (PAH) gene, PAH transcript variants, Coding sequence analysis, Exon-wise repeat distribution, Fast Pairwise Complementary BLAST (FPCB), Repetitive DNA elements, Computational genomics, In silico sequence analysis*

How to cite this article: *Singh L, Dangi N. Computational Identification And Distribution Of Mirror Repeats In Transcripts Of Human PAH Gene. Int J Drug Deliv Technol. 2026;16(15s): 375-383. DOI: 10.25258/ijddt.16.15s.44.*

INTRODUCTION

The human genome contains a substantial proportion of repetitive DNA sequences, which together form a major component of its overall sequence composition and structural organization (International Human Genome Sequencing Consortium, 2001; Ellegren, 2004; Treangen & Salzberg, 2012). These repetitive elements are interspersed throughout both coding and non-coding regions and exhibit considerable diversity in terms of sequence organization, evolutionary origin, and genomic distribution (Katti & Ranjekar, 2002; Li et al., 2003). Owing to their abundance and structural variability, repetitive DNA sequences **represent a major component of genome complexity and architectural heterogeneity**. The major categories of repetitive DNA present in the human genome are summarized in **Figure 1**, with the present work focusing specifically on **mirror repeat sequences**.

The increasing availability of high-quality genome assemblies, together with the development of computational approaches for sequence analysis, has facilitated the systematic detection and classification of repetitive DNA motifs at both genome-wide and locus-specific scales. Among the various classes of repetitive DNA, certain sequence motifs possess intrinsic symmetry or compositional features that allow them to deviate from the canonical B-form DNA structure. Such motifs are collectively referred to as **non-B DNA–forming sequences** and include a range of structurally distinct repeat types (Wells, 2007; Cer et al., 2011; Sharma, 2011; Bansal et al., 2022).

Mirror repeats constitute a distinct category of symmetric DNA sequences in which a nucleotide tract is repeated in reverse order on the same strand. Under specific physicochemical conditions, mirror repeats have been reported to participate in the formation of alternative DNA conformations, including triplex-associated structures (Frank-Kamenetskii & Mirkin, 1995; Mirkin, 2008). Large-scale computational surveys have indicated that mirror repeats are widely distributed across the human genome and are often generated through mechanisms involving tandem repeat expansion or local sequence duplication (McGinty et al., 2025).

The occurrence of non-B DNA–forming motifs, including mirror repeats, **has been reported in genomic contexts characterized by localized sequence variability across diverse biological systems** (Bacolla & Wells, 2004; Zhao et al., 2009; Duardo et al., 2023). However, the majority of existing investigations have primarily emphasized genome-wide assessments or comparative analyses across large genomic regions. In contrast, fewer studies have addressed the detailed organization of mirror repeats within individual human genes, particularly at the level of transcripts and individual exons. As a result, the fine-scale distribution patterns of mirror repeats within gene-specific contexts remain insufficiently documented for many clinically relevant loci (Yadav et al., 2022; McGinty et al., 2025).

*Author for Correspondence: Singhlaxita.21@gmail.com

The **phenylalanine hydroxylase (PAH) gene** encodes a critical enzyme involved in phenylalanine metabolism and is characterized by a well-defined exon–intron structure and multiple transcript variants (Goltsov et al., 1992; Woo et al., 1992; Lichter-Konecki et al., 1994; Flydal et al., 2013). Pathogenic alterations in this gene are known to cause **phenylketonuria**, a widely studied inherited metabolic disorder (Blau et al., 2010). Due to its extensive molecular annotation, established clinical relevance, and clearly defined transcript architecture, the PAH gene provides a suitable framework for detailed computational examination of sequence-level features (GeneReviews® Editors, 2025; MedlinePlus Genetics, 2023).

Despite comprehensive clinical and molecular characterization of the PAH gene, a systematic computational assessment of mirror repeat distribution across its transcript variants has not been comprehensively reported. The present study addresses this gap by identifying and cataloging mirror repeats within human PAH gene transcripts using a **BLAST-based FPCB computational strategy** (Bhardwaj et al., 2013). By presenting a detailed exon-wise and fragment-wise description of mirror repeat occurrences, this work establishes a **descriptive baseline** that may serve as a reference for subsequent comparative or functional analyses of repetitive DNA motifs in disease-associated genes.

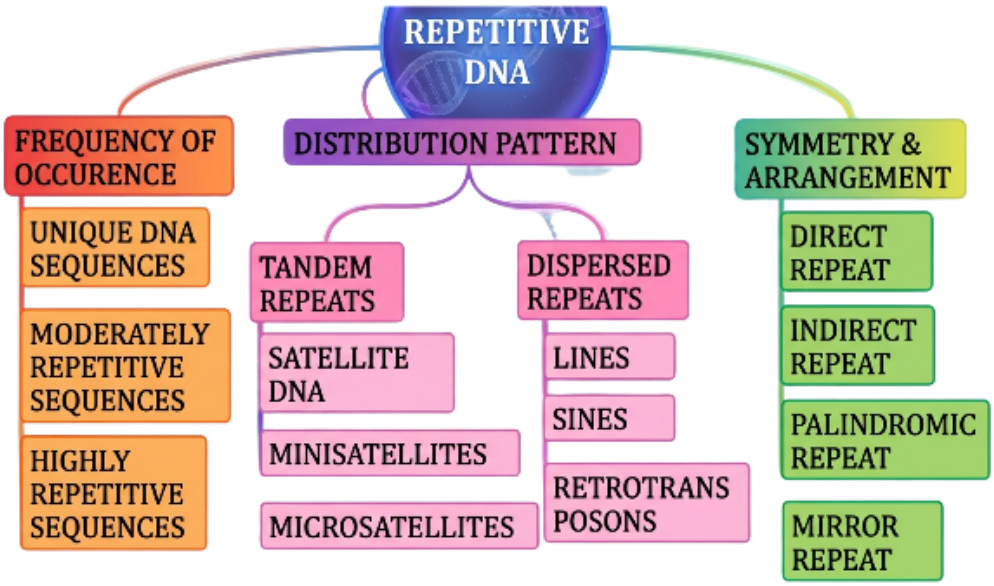


Figure1: Types of repetitive DNA Created in https://BioRender.com

Materials and Methods

Retrieval of PAH Transcript Sequences

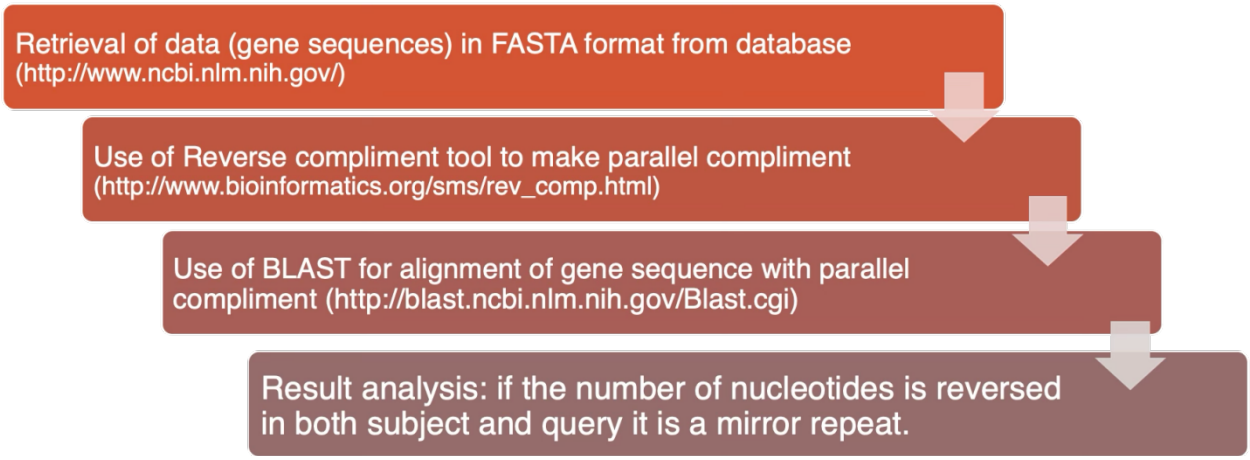
Nucleotide sequences corresponding to human phenylalanine hydroxylase (PAH) gene transcripts were retrieved from the National Center for Biotechnology Information (NCBI) database. The canonical transcript (NM_000277.3) and an alternative transcript variant (NM_001354304.2) were selected based on curated annotation, transcript completeness, and relevance to PAH gene expression (GeneReviews® Editors, 2025; MedlinePlus Genetics, 2023). All sequences were obtained in FASTA format and used exclusively for transcript-level computational analysis. Genomic DNA sequences, intronic regions, and upstream or downstream regulatory elements were not included in the present study.

Fragmentation Strategy for Mirror Repeat Detection

Mirror repeat identification was performed using a BLAST-based Fast Pairwise Complementary BLAST (FPCB) strategy, which has been described as an effective computational approach for detecting symmetric DNA motifs (Bhardwaj et al., 2013). Each PAH transcript sequence was fragmented into overlapping windows of 500 base pairs to ensure sensitive detection of mirror repeats across the coding sequence. Overlapping fragmentation was intentionally retained at this stage to maximize coverage and minimize the possibility of missing repeat motifs spanning fragment boundaries.

BLAST-Based Identification of Mirror Repeats

For each fragmented transcript sequence, a reverse complementary sequence was generated and aligned against the original fragment using the BLAST algorithm. BLAST was selected due to its high sensitivity, robustness, and widespread application in sequence similarity detection (Altschul et al., 1990; Altschul et al., 1997). All BLAST searches were performed using consistent parameters across analyses to maintain methodological uniformity. Alignments exhibiting sequence symmetry characteristic of mirror repeat patterns were extracted and subjected to further screening.



Curation and Exon-Wise Mapping of Mirror Repeats

Mirror repeats identified during fragment-wise analysis were initially retained to preserve detection sensitivity. For exon-wise analysis, detected mirror repeats were curated to remove redundant entries arising from overlapping coding sequence windows. Exon boundaries were determined using transcript annotation data obtained from NCBI. Curated mirror repeats were subsequently mapped to annotated exons of transcript variant 1, resulting in a set of unique exon-localized mirror repeats suitable for descriptive analysis. This curation approach ensured accurate representation of mirror repeat distribution at the exon level while avoiding inflation of repeat counts due to overlapping detections.

Classification of Mirror Repeats

Identified mirror repeats were classified as perfect or imperfect based on sequence symmetry. Perfect mirror repeats exhibited uninterrupted symmetry across the repeat region, whereas imperfect mirror repeats contained one or more mismatches within the symmetric motif. Classification was performed in accordance with established definitions used in previous studies of mirror repeats and non-B DNA-forming motifs (Frank-Kamenetskii & Mirkin, 1995; Mirkin, 2008; Bacolla & Wells, 2004).

Data Compilation and Visualization

All curated mirror repeats were compiled into tabular format, recording their sequence, positional coordinates, length, and classification. Fragment-wise and exon-wise mirror repeat counts were summarized using descriptive statistics. Visualization of mirror repeat distribution across the coding sequence and exon regions was performed using bar graphs and linear distribution plots. The analytical framework applied in this study is consistent with established computational methodologies for sequence pattern analysis and repeat detection (Gusfield, 1997; Pevzner, 2000). The present study was designed exclusively as a descriptive computational investigation, and no functional or experimental validation was performed.

RESULTS

1. Identification of Mirror Repeats in the PAH Gene

Computational screening of the human *phenylalanine hydroxylase* (*PAH*) gene transcript (NM_000277.3) using the BLAST-based FPCB strategy revealed a substantial abundance of mirror repeat (MR) sequences. Analysis of the complete coding sequence (CDS) spanning 3,759 bp identified a total of **86 mirror repeats**, comprising both **perfect mirror repeats** and **imperfect mirror repeats** containing mismatches or spacer nucleotides.

Mirror repeats were distributed throughout the length of the transcript, indicating a widespread but non-uniform presence across the gene. Most detected repeats were relatively short (10–20 bp); however, several longer imperfect mirror repeats were also observed, suggesting sequence heterogeneity within the CDS.

To facilitate systematic BLAST-based detection, the *PAH* CDS was divided into eight overlapping fragments of 500 bp each, with the final fragment comprising 259 bp. Mirror repeats were detected in all fragments, although their frequency varied considerably. Certain regions exhibited a higher density of mirror repeats, particularly those corresponding to longer exonic segments.

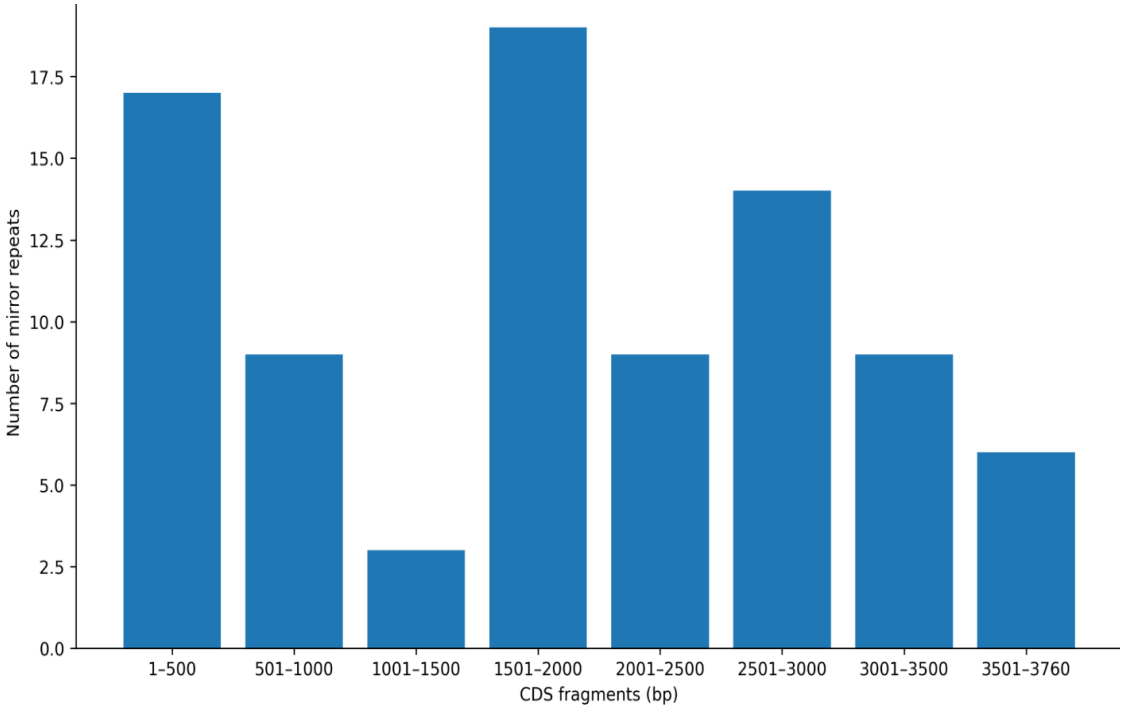


Figure 2: CDS-wise mirror repeat distribution in PAH transcript variant1 (n=86)

Fragment-Wise Detection of Mirror Repeats in the PAH CDS

We performed a fragment-wise analysis using sliding windows of 500 bp, yielding a comprehensive dataset of 315 BLAST hits that resolved into 86 distinct mirror repeats. The spatial distribution of these repeats across the CDS exhibited marked non-uniformity. Notably, the 1501–2000 bp fragment harbored the highest density of mirror repeats. In contrast, the 1001–1500 bp region displayed the lowest frequency. Across all fragments, perfect mirror repeats and those interrupted by short spacers predominated (collectively comprising 78.6% of detections), indicative of a pervasive bias toward highly conserved palindromic architectures within the CDS.

Table 1. Fragment-wise Distribution of Mirror Repeats in the *PAH* CDS (NM_000277.3)

CDS Region (bp)	CDS Length (bp)	Expected Threshold	No. of BLAST Hits	No. of Mirror Repeats	Perfect / Perfect with Spacer	Imperfect
1–500	500	20	29	17	16	1
501–1000	500	20	13	9	6	3
1001–1500	500	20	21	3	3	0
1501–2000	500	20	69	19	15	4
2001–2500	500	20	69	9	8	1
2501–3000	500	20	45	14	12	2
3001–3500	500	20	47	9	8	1

3501–3760	260	20	22	6	4	2
Total	3760	—	315	86	71	15

2. Exon-wise Distribution of Mirror Repeats in Transcript Variant 1

Exon-specific analysis of NM_000277.3 revealed that **63 mirror repeats** were localized within the 13 exonic regions. The distribution of mirror repeats varied considerably among exons. Longer exons generally harbored a greater number of repeats, whereas several shorter exons lacked detectable mirror repeats.

Exon 13, the longest exon, contained the highest number of mirror repeats (**36 repeats**), while exons 8, 9, and 10 showed no detectable mirror repeats. This uneven distribution highlights transcript-region–specific variation in mirror repeat occurrence.

Table 2. Exon-wise Distribution of Mirror Repeats in *PAH* Transcript Variant 1

Exons	Length of exon (bps)	Expected threshold	Number of hits	Number of mirror repeats
1. 1-174	174	100	8	4
2. 175-282	108	100	5	3
3. 283-466	184	100	12	8
4. 467-555	89	100	1	1
5. 556-623	68	100	1	1
6. 624-820	197	100	7	5
7. 821-956	136	100	5	3
8. 957-1026	70	100	0	0
9. 1027-1083	57	100	0	0
10. 1084-1179	96	100	0	0
11. 1180-1313	134	100	6	2
12. 1314-1429	116	100	2	0
13. 1430-3759	2330	100	401	36

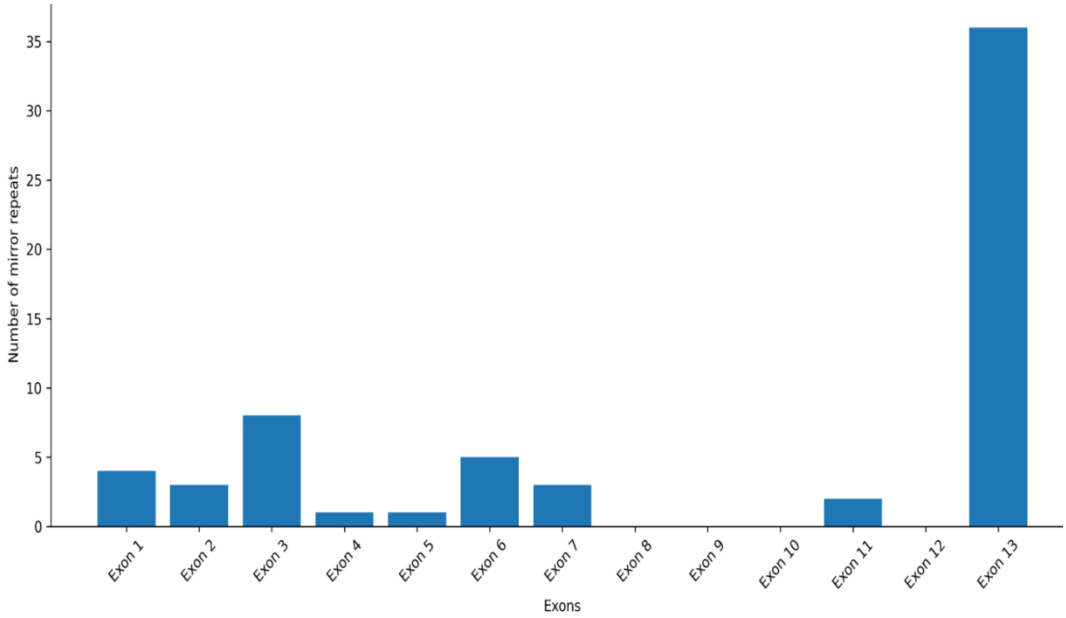


Figure 3: Exon wise mirror repeat frequency in PAH transcript variant 1 (n=63)

Comparative Interpretation of CDS-wise and Exon-wise Analyses

Fragment-wise analysis using overlapping CDS windows enabled sensitive detection of mirror repeats across the entire transcript, yielding 86 mirror repeats in total. However, exon-wise curation removed redundant detections and provided a refined set of **63 unique mirror repeats**, allowing accurate mapping of symmetric motifs to gene structure.

The non-uniform distribution observed at both CDS and exon levels indicates region-specific variability in mirror repeat occurrence. Certain CDS fragments and exons exhibit pronounced enrichment, whereas others are relatively devoid of such motifs. The exon-wise frequency of mirror repeats is illustrated in **Figure 3**, emphasizing the heterogeneous organization of mirror repeats along the *PAH* transcript.

3. Classification of Mirror Repeats

Following exon-wise curation, the identified mirror repeats in *PAH* transcript variant 1 (NM_000277.3) were classified based on **sequence symmetry** and **repeat length**. Mirror repeats were categorized into two major classes: **perfect mirror repeats (including repeats with spacers)**, which exhibit exact or near-exact symmetry, and **imperfect mirror repeats**, which contain mismatches or intervening spacer nucleotides. Each class was further subdivided into **short (≤7 bp)**, **medium (8–10 bp)**, and **long (>10 bp)** length categories.

Table 3. Classification of mirror repeats in PAH transcript variant 1 based on type and length

Mirror Repeat Type	**Length Category (bp)**	**No. of Mirror Repeats**	**Percentage (%)**
Perfect / Perfect with spacer	Short (≤7 bp)	28	44.4
	Medium (8–10 bp)	29	46.0
	Long (>10 bp)	14	22.2
Imperfect	Short (≤7 bp)	5	7.9
	Medium (8–10 bp)	6	9.5
	Long (>10 bp)	11	17.5
Total	—	**63**	**100**

DISCUSSION

Repetitive DNA elements represent a fundamental component of genome architecture, yet their organization within individual genes remains insufficiently characterized for many clinically relevant loci. In the present study, a systematic computational analysis was undertaken to identify and map mirror repeats within the human phenylalanine hydroxylase (PAH) gene transcript. Using a BLAST-based FPCB strategy, mirror repeats were detected, curated, and classified to generate a detailed descriptive profile of their distribution across the coding sequence and exon regions.

Analysis of overlapping 500 bp coding sequence fragments revealed the presence of 86 mirror repeats distributed non-uniformly along the PAH coding region. Fragment-wise analysis demonstrated that mirror repeats were not evenly represented across the coding sequence, indicating regional variation in repeat density. Such non-random distribution patterns are consistent with previous observations that repetitive DNA motifs often exhibit localized enrichment rather than uniform genomic dispersion. The fragment-wise approach employed in this study ensured sensitive detection of symmetric motifs while preserving positional resolution across the transcript.

Subsequent exon-wise curation resulted in the identification of 63 unique mirror repeats localized within annotated exons of transcript variant 1. The reduction in repeat count following exon-level mapping reflects the removal of redundant detections arising from overlapping coding sequence windows. Exon-wise analysis further revealed heterogeneity in mirror repeat abundance across individual exons, with longer exons contributing a greater number of detected repeats. This observation highlights exon length as an important factor influencing the absolute number of mirror repeats identified within coding regions, while also underscoring the importance of normalization and curation strategies in repeat analysis.

Classification of mirror repeats based on sequence symmetry identified both perfect and imperfect mirror repeats within the PAH transcript. The presence of imperfect mirror repeats, defined by minor mismatches within the symmetric region, reflects natural sequence variability and reinforces the need for inclusive detection thresholds in computational repeat identification. The classification framework applied in this study adheres to established definitions used in analyses of symmetric DNA motifs and provides a structured basis for describing mirror repeat diversity within a single gene context.

Importantly, the scope of this study was intentionally limited to descriptive characterization. No functional interpretations or biological consequences were inferred from the observed distribution patterns. Instead, the emphasis was placed on generating a rigorously curated, reproducible dataset that documents mirror repeat organization within the PAH gene. By maintaining a strictly descriptive focus, the present work establishes a reliable reference framework that can support future comparative, experimental, or functional investigations without overextending the interpretive scope of the current analysis.

CONCLUSION

This study presents the first comprehensive, transcript-level catalog of mirror repeats within the human PAH gene using a systematic computational approach. Through fragment-wise detection, exon-wise curation, and symmetry-based classification, mirror repeats were identified and mapped across the coding sequence of transcript variant 1 with clearly defined methodological thresholds.

The identification of both perfect and imperfect mirror repeats, along with their non-uniform distribution across coding fragments and exons, underscores the structural complexity of repetitive DNA organization within a single human gene. The analytical framework employed here demonstrates robustness, reproducibility, and adaptability for mirror repeat detection at the gene level. Overall, this work establishes a detailed descriptive baseline for mirror repeat organization in the PAH gene transcript. The curated dataset and methodological strategy presented here provide a valuable reference resource for future studies aimed at exploring the structural, comparative, or functional significance of mirror repeats in human genes.

References

1. McGinty, R., Lyskova, A., & Mirkin, S. M. (2025). The origin of mirror repeats in the human genome. *Nucleic Acids Research, 53*(12), gkaf619. https://doi.org/10.1093/nar/gkaf619
2. Bhardwaj, V., Gupta, S., Meena, S., & Sharma, K. (2013). FPCB: A simple and swift strategy for mirror repeat identification. *arXiv Preprint*. https://arxiv.org/abs/1312.3869
3. Yadav, S., Yadav, U., & Sharma, D. C. (2022). In silico evaluation of mirror repeats in HIV genome. *International Journal of Life Sciences and Pharma Research, 11*(5), 81–87.
4. Cer, R. Z., Donohue, D. E., Mudunuri, U. S., et al. (2011). Non-B DB: A database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Research, 39*(Database issue), D383–D391. https://doi.org/10.1093/nar/gkq954
5. Sharma, S. (2011). Non-B DNA secondary structures and their resolution by RecQ helicases. *Frontiers in Bioscience, 16*, 1060–1072. https://doi.org/10.2741/3725
6. Bansal, A., Kumar, A., Singh, A., & Kumar, R. (2022). Noncanonical DNA structures: Diversity and disease. *Frontiers in Genetics, 13*, 959258. https://doi.org/10.3389/fgene.2022.959258
7. Duardo, R. C., et al. (2023). Non-B DNA structures as a booster of genome instability. *DNA Repair*. Advance online publication. https://doi.org/10.1016/j.dnarep.2023.103586
8. Zhao, J., Bacolla, A., Wang, G., & Vasquez, K. M. (2009). Non-B DNA structure-induced genetic instability and evolution. *Molecular*

Carcinogenesis, 48(4), 273–285. https://doi.org/10.1002/mc.20467
9. GeneReviews® Editors. (2025). *Phenylalanine hydroxylase deficiency*. University of Washington, Seattle. https://www.ncbi.nlm.nih.gov/books/NBK1504/
10. MedlinePlus Genetics. (2023). *PAH gene: Phenylalanine hydroxylase*. U.S. National Library of Medicine. https://medlineplus.gov/genetics/gene/pah/
11. Flydal, M. I., Martinez, A., & Blau, N. (2013). Phenylalanine hydroxylase: Function, structure, and regulation. *IUBMB Life, 65*(3), 341–349. https://doi.org/10.1002/iub.1120
12. Goltsov, A. A., Eisensmith, R. C., & Woo, S. L.-C. (1992). A polymorphic STR system within the human phenylalanine hydroxylase gene. *Human Genetics, 89*, 347–350. https://doi.org/10.1007/BF00219105
13. Woo, S. L.-C., et al. (1992). Structural characterization of the 5′ regulatory region of the human phenylalanine hydroxylase gene. *Genomics, 12*(1), 84–92. https://doi.org/10.1016/0888-7543(92)90428-N
14. Lichter-Konecki, U., et al. (1994). Functional characterization of a unique liver gene promoter. *Journal of Biological Chemistry, 269*(12), 9137–9146.
15. Kim, Y. M., et al. (1999). Expression of phenylalanine hydroxylase in human tissues. *Journal of Biological Chemistry, 274*(2), 925–931. https://doi.org/10.1074/jbc.274.2.925
16. Blau, N., van Spronsen, F. J., & Levy, H. L. (2010). Phenylketonuria. *The Lancet, 376*(9750), 1417–1427. https://doi.org/10.1016/S0140-6736(10)60961-0
17. Wikipedia contributors. (2025). *Non-B DNA*. Wikipedia. https://en.wikipedia.org/wiki/Non-B_DNA
18. Wikipedia contributors. (2024). *Tandem repeat*. Wikipedia. https://en.wikipedia.org/wiki/Tandem_repeat
19. Wikipedia contributors. (2025). *Phenylalanine hydroxylase*. Wikipedia. https://en.wikipedia.org/wiki/Phenylalanine_hydroxylase
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
21. Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389
22. Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research, 27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573
23. Smit, A. F. A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0*. Institute for Systems Biology. https://www.repeatmasker.org
24. Frank-Kamenetskii, M. D., & Mirkin, S. M. (1995). Triplex DNA structures. *Annual Review of Biochemistry, 64*, 65–95. https://doi.org/10.1146/annurev.bi.64.070195.000433
25. Mirkin, S. M. (2008). Discovery of alternative DNA structures: A heroic decade (1979–1989). *Frontiers in Bioscience, 13*, 1064–1071. https://doi.org/10.2741/2756
26. Wells, R. D. (2007). Non-B DNA conformations, mutagenesis and disease. *Trends in Biochemical Sciences, 32*(6), 271–278. https://doi.org/10.1016/j.tibs.2007.04.003
27. Bacolla, A., & Wells, R. D. (2004). Non-B DNA conformations as determinants of mutagenesis and human disease. *Molecular Carcinogenesis, 40*(2), 95–102. https://doi.org/10.1002/mc.20019
28. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860–921. https://doi.org/10.1038/35057062
29. Brendel, V., Beckmann, J. S., & Trifonov, E. N. (1986). Linguistics of nucleotide sequences: Morphology and comparison of vocabulary. *Journal of Biomolecular Structure and Dynamics, 4*(1), 11–21. https://doi.org/10.1080/07391102.1986.10507469
30. Trifonov, E. N. (1989). The multiple codes of nucleotide sequences. *Bulletin of Mathematical Biology, 51*(4), 417–432. https://doi.org/10.1016/S0092-8240(89)80001-3
31. Katti, M. V., & Ranjekar, P. K. (2002). Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Science, 11*(3), 626–640. https://doi.org/10.1110/ps.37702
32. Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics, 5*(6), 435–445. https://doi.org/10.1038/nrg1348
33. Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics, 13*(1), 36–46. https://doi.org/10.1038/nrg3117
34. Li, W., Bernaola-Galván, P., Carpena, P., & Oliver, J. L. (2003). Isochores merit the prefix “iso”. *Computational Biology and Chemistry, 27*(1), 5–10. https://doi.org/10.1016/S1476-9271(02)00056-5
35. Sinden, R. R. (1994). *DNA structure and function*. Academic Press.
36. Rich, A., & Zhang, S. (2003). Z-DNA: The long road to biological function. *Nature Reviews Genetics, 4*(7), 566–572. https://doi.org/10.1038/nrg1116
37. Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: Computer science and*

computational biology. Cambridge University Press.
38. Pevzner, P. A. (2000). *Computational molecular biology: An algorithmic approach*. MIT Press.
39. Wootton, J. C., & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry, 17*(2), 149–163. https://doi.org/10.1016/0097-8485(93)85006-X
40. Karlin, S., & Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science, 257*(5066), 39–49. https://doi.org/10.1126/science.1621092
41. Smith, T. F., Waterman, M. S., & Burks, C. (1985). The statistical distribution of nucleic acid similarities. *Nucleic Acids Research, 13*(2), 645–656. https://doi.org/10.1093/nar/13.2.645