Hundred-million-scale gene-encoded natural diverse components repository

Wei Chen,[1,2] Chi Song,[1,2] Liang Leng,[1,2] and Shilin Chen[1,2,*]

[1]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China

[2]Institute of Herbgenomics, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China

*Correspondence: slchen@cdutcm.edu.cn (S.C.)

Received: April 22, 2025; Accepted: July 2, 2025; Published Online: July 7, 2025; https://doi.org/10.59717/j.xinn-life.2025.100146

Citation: Chen W., Song C., Leng L., et al. (2025). Hundred-million-scale gene-encoded natural diverse components repository. *The Innovation Life* **3**:100146.

On April 17th, 2025, the world's first Gene-encoded Natural Diverse Components Repository (GNDC) was officially launched at the 3rd "1K Herb Genomes Project" conference in Chengdu, Sichuan Province. The release marks a significant leap forward in digitizing and decoding the biochemical wealth of traditional medicine using genomics and artificial intelligence.

"GNDC achieves deep integration between genomic data and artificial intelligence, ushering in a new era of intelligent and precise natural product discovery," said Professor Shilin Chen, academician of the Chinese Academy of Engineering and the principal investigator of GNDC. "It offers a transformative foundation for future drug discovery and modernization of traditional Chinese medicine."

Available at https://cbcb.cdutcm.edu.cn/gndc/, GNDC was built upon nuclear and organelle genome data from 1,037 medicinal species recorded in eight authoritative pharmacopoeias (Brazilian, Egyptian, European, Indian, Japanese, Korean, the Pharmacopoeia of the People's Republic of China, and U.S. Pharmacopoeia's Herbal Medicines Compendium),[1] supplemented by multiple publicly available resources. GNDC currently catalogs over 234 million gene-encoded herbal components, encompassing 2.32 million secondary metabolites, 229 million peptides, 2.38 million small RNAs, and 0.26 million carbohydrate molecules (Figure 1). By combining genomic, compositional, and functional information, GNDC provides a level of chemical insight unmatched by existing natural product databases.

GNDC's development was enabled by a self-developed genome traversal algorithm in combination with multi-omics technologies-including transcriptomics, small RNA sequencing, and peptide mass spectrometry, enabling systematic mining and validation of natural diverse components.[2] Furthermore, GNDC innovatively integrates artificial intelligence algorithms such as natural language processing (NLP), generative adversarial networks (GANs), and graph convolutional networks (GCNs) to achieve automated identification, precise classification, and functional annotation of components.

Unlike conventional databases that sort natural components mainly by their chemical structures or source species, GNDC introduces a classification system rooted in the central dogma of molecular biology. Within this framework, natural components are categorized into direct gene-encoded components, such as small RNA and small peptides, and indirect gene-encoded components, including secondary metabolites and carbohydrates. These components are essential for an organism's ability to adapt, survive, and thrive in various environmental conditions.[3] They also play key roles in cellular communication, defense mechanisms, and metabolic regulation. From a medicinal standpoint, the vast diversity of these components presents an opportunity to discover novel bioactive compounds with therapeutic potential.[4] The greater the diversity in natural components, the more opportunities there are to identify new drugs and treatment strategies, often targeting previously unexplored biological pathways.[5] Therefore, the exploration of this diversity is central to advancing drug discovery.

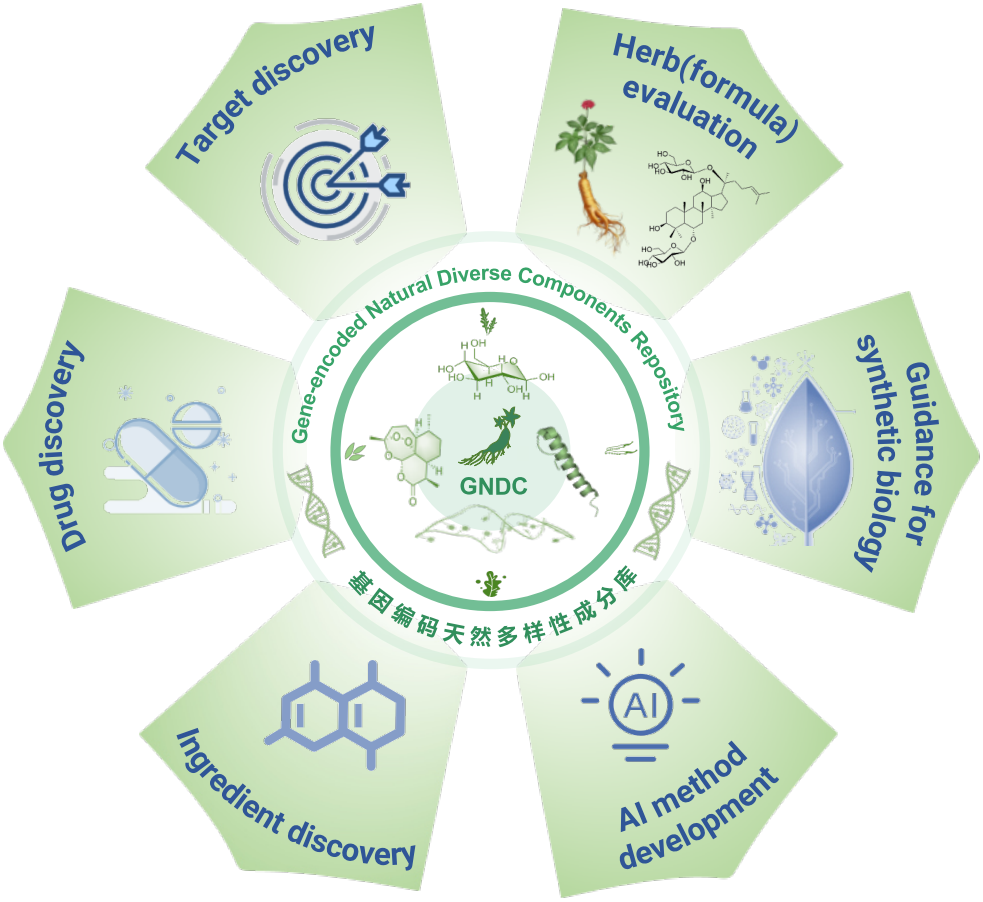

Figure 1. The prospective applications of GNDC could lead to disruptive innovations in traditional medicine research

The GNDC project represents both a technological breakthrough and a conceptual shift in natural product research. Traditionally reliant on labor-intensive extraction and chemical analysis, natural product discovery is now moving toward a computational-first approach. With GNDC, candidate compounds can be predicted from genomic data before laboratory validation, reversing the traditional workflow and accelerating the discovery process.

With over 234 million structurally annotated and functionally predicted entries, GNDC vastly expands the chemical search space available for pharmacological screening, offering a high-resolution alternative to traditional compound libraries. GNDC sets a new standard for digital infrastructure in traditional medicine research. By classifying compo

nents into direct and indirect gene-encoded ones, GNDC enables a more holistic, systems-level understanding of the complex mixtures of compounds in herbal medicines, which often act synergistically. Moreover, GNDC's comprehensive, structured, and interoperable design will facilitate data sharing, AI applications, and cross-disciplinary collaboration at an unprecedented scale. By bridging traditional ethnopharmacology with modern data science, GNDC not only supports the scientific modernization of herbal medicine, but also accelerates drug innovation in the era of big data.

GNDC not only serves as a comprehensive repository of existing genomic knowledge but also establishes a pioneering platform for advancing research into the diverse natural components encoded by genes. Each category of components within GNDC presents distinct challenges. While secondary metabolites have been extensively studied, a major unresolved issue remains in effectively translating genomic data into biosynthetic pathway inference. Similarly, carbohydrates pose difficulties due to their indirect genomic encoding, complicating efforts to interpret their biosynthesis from sequence data. For directly gene-encoded components, a critical bottleneck lies in the high-throughput experimental validation required to confirm the vast number of compounds predicted through genome mining. In future, efforts will be directed toward establishing dynamic update mechanisms that integrate emerging multi-omics data, enabling the platform to stay aligned with the latest scientific advancements.

REFERENCES

1. Liao B., Hu H., Xiao S., et al. (2022). Global Pharmacopoeia Genome Database is an integrated and mineable genomic database for traditional medicines derived from eight international pharmacopoeias. *Sci. China Life Sci.* **65**:809–817. DOI:10.1007/s11427-021-1968-7
2. Chen W., Yu Z. Y., Leng L., et al. (2025). Artificial Intelligence-Curated repository of gene-encoded natural diverse components from herbal medicines. *The Innovation* DOI: 10.1016/j.xinn.2025.101011.
3. Xu Z., Tian Y., Wang J., et al. (2024). Convergent evolution of berberine biosynthesis. *Sci Adv* **10**:eads3596. DOI:10.1126/sciadv.ads3596
4. Bi Z., Li H., Liang Y., et al. (2025). Emerging paradigms for target discovery of traditional medicines: A genome-wide pan-GPCR perspective. *The Innovation* **6**:100774. DOI:10.1016/j.xinn.2024.100774
5. Leng L., Xu Z., Hong B., et al. (2024). Cepharanthine analogs mining and genomes of Stephania accelerate anti-coronavirus drug discovery. *Nat. Commun.* **15**:1537. DOI:10.1038/s41467-024-45690-5

FUNDING AND ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation of Sichuan (No. 2024ZDZX0019). The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

W.C., L.L., C.S. and S.C. conceived, organized and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.