PARAGON

(Pan-species Analysis of RNA Architecture & Genomic Organization Networks)

Revealing and Understanding the Non-coding Genome

Ananth Kumar, PhD

Exploring the Genome's Uncharted Territory

Over 98 % of the human genome does not code for proteins (Poller W et al, 2023). This vast molecular landscape, often called genomic dark matter, contains the regulatory instructions that control when, where, and how much of each gene is expressed in our cells. While genomic dark matter includes various regulatory elements, much of it is transcribed into non-coding RNAs (ncRNAs), which act as molecular switches and coordinators of gene activity. For example, microRNAs (miRNAs) can silence specific genes, long non-coding RNAs (lncRNAs) control which genes are accessible for activation, and small nuclear RNAs (snRNAs) help process genetic instructions into functional forms. Despite their fundamental cellular importance, ncRNAs have been largely underexplored. Landmark initiatives like the Human Genome Project and recent breakthroughs such as AlphaFold, have advanced our understanding of how DNA and proteins drive cell functioning and disease pathology. Consequently, molecular biology remains fundamentally protein-centric, and modern medicine reflects this bias: the **majority of FDA approved drugs target proteins** (The Human Protein Atlas). The products of the non-coding genome represent an unexplored universe of therapeutic targets, yet have been long dismissed as "junk" (Warner K et al, 2018). The ENCODE project (a groundbreaking global initiative curating data on transcription, chromatin structure, histone marks, DNA modifications, and protein binding) revealed that the vast **majority of the non-coding transcriptome is biochemically active**, recognizing ncRNAs as essential cellular operators (The ENCODE Project). Despite this, only a few dozen lncRNAs in humans and plants have been rigorously characterized, out of tens of thousands that exist (Mattick JS, 2023 & Palos K et al, 2023).

NcRNAs encompass diverse molecular classes that are technically challenging to study. They exist at low abundance, display heterogeneous expression, adopt dynamic structures, and often have redundant functions. (Ponting CP and Haerty W, 2022 & Mattick JS, 2023). Conflicting results and divergent interpretations pervade the field, making it difficult for clinicians, drug developers, and academics to draw meaningful conclusions about ncRNA roles in health and disease. **These barriers have stifled progress and deterred new entrants to the field.**

Current efforts to map ncRNAs are led primarily by the ENCODE project. ENCODE faces incomplete tissue coverage, reliance on lab models, low dataset resolution, and challenges in data integration. Since the COVID-19 mRNA vaccine boom, RNA programs like the RNome project and Singapore's NIBRA have emerged (The International Human RNome Project Consortium, 2025 & https://nirba.sg/). But none specifically target the non-coding transcriptome. Academic labs produce isolated, high quality studies on individual ncRNAs, but lack coordination, scalable tools, standardization, and systematic coverage.

To bridge the above gap, we need a focused research platform for ncRNAs, one that will **create a level playing field for everybody, accelerate ncRNA discovery across species**, ultimately enabling new avenues for understanding, targeting, and treating disease. I propose the creation of PARAGON, an integrated platform that systematically catalogs the structural and sequence profiles, regulatory mechanisms, and interactions of distinct ncRNA subgroups.

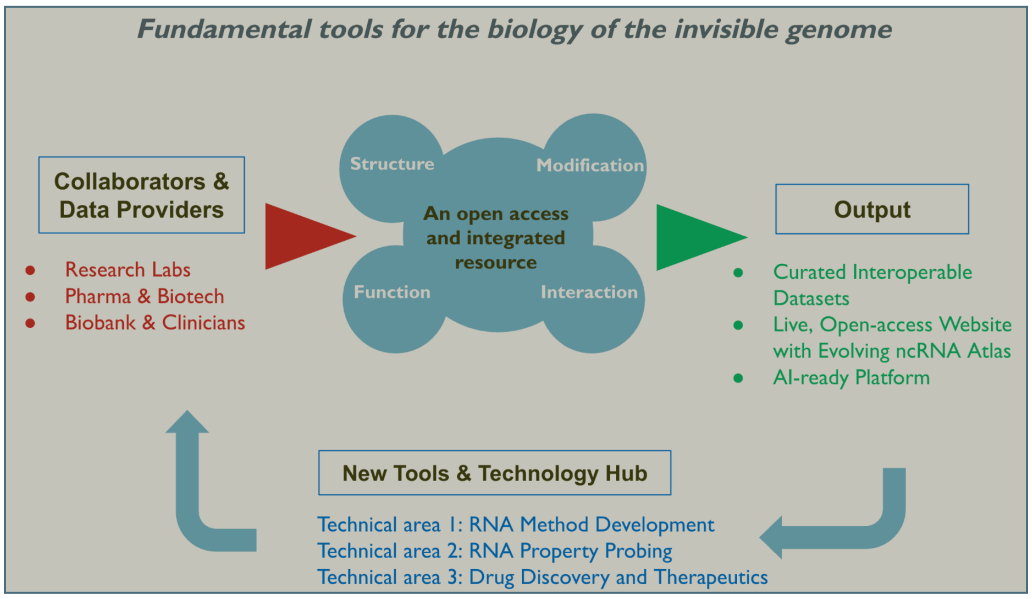

Image: Overview of PARAGON

PARAGON: From Dark Matter to Discovery

PARAGON will provide a comprehensive platform for ncRNAs, designed to catalog, map and analyze them via open, scalable, and integrated tools. Advances in long-read sequencing, RNA multi-omics, transformer models, and the unmet medical need for identifying new druggable targets makes this the optimal time to launch PARAGON. The program is built on four pillars: RNA structural biology, RNA modification mapping, RNA-protein interaction mapping, and RNA functional genomics. Together, these will

generate a rich, multi dimensional view of ncRNAs. A central data science team will ensure curated, interoperable datasets and develop machine learning platforms for actionable insights. A live, open-access website will showcase an evolving ncRNA atlas with visualization tools linking ncRNAs to tissues, functions, diseases, and molecular interactions. The platform will remain human-centric while accommodating cross-species data, enabling discovery of conserved elements and species-specific functions. Rigorous data governance protocols will safeguard privacy and ensure data integrity.

Five Years to Transform RNA Biology

PARAGON is a five-year, £25 million program spanning three technical areas and structured in four stages. It includes a six-month launch phase focused on engaging end users, refining market needs, establishing governance, securing IP agreements and hiring key personnel. The year-long second stage will be dedicated to building the experimental platform, integrating data pipelines, and delivering a v1.0 data platform. The 18-month scale-up stage will include transitioning to high-throughput experimental pipelines, expanding the web platform, releasing early RNA language models, and conducting user review and soliciting feedback. The two-year translational phase will establish PARAGON as a central ncRNA data hub, enabling cross-species analyses, initiating preclinical and IP activities, and securing long-term funding.

What Success Looks Like

PARAGON's success would **transform genomic 'dark matter' into a mapped, accessible atlas with powerful analytical and technical tools**. By decoding RNA's molecular grammar, machine learning models trained on PARAGON data will uncover hidden patterns in the transcriptome, revealing new therapeutic target classes such as pathogenic protein-RNA interfaces and splice junctions. Clinically, ncRNA-based biomarkers could enable non-invasive early detection of pancreatic and lung cancers, where survival depends on catching disease at its earliest stages. PARAGON data will catalyze for-profit ventures, strengthening the RNA therapeutics sector projected to reach $25 billion by 2034. By 2035, this comprehensive understanding of RNA biology could enable wearable biosensors that continuously monitor circulating ncRNA biomarkers for real-time health management, and mRNA vaccine synthesizers in local clinics capable of rapid immunization against emerging pathogens, fundamentally shifting medicine from reactive treatment to proactive prevention. Beyond health, PARAGON's pan-species framework will reveal how organisms adapt to environmental changes by fine-tuning their epitranscriptome to maintain ecosystem resilience, with implications for climate change and agriculture.

References

1. Poller W, Sahoo S, Hajjar R, Landmesser U, Krichevsky AM. Exploration of the Noncoding Genome for Human-Specific Therapeutic Targets-Recent Insights at Molecular and Cellular Level. Cells. 2023 Nov 20;12(22):2660.

2. The Human Protein Atlas The human proteome in druggable

3. Warner, K., Hajdin, C. & Weeks, K. Principles for targeting RNA with drug-like small molecules. Nat Rev Drug Discov 17, 547–558 (2018).

4. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004 Oct 22;306(5696):636-40.

5. Mattick JS. RNA out of the mist. Trends Genet. 2023 Mar;39(3):187-207. doi: 10.1016/j.tig.2022.11.001. Epub 2022 Dec 15. PMID: 36528415.

6. Palos K, Yu L, Railey CE, Nelson Dittrich AC, Nelson ADL. Linking discoveries, mechanisms, and technologies to develop a clearer perspective on plant long noncoding RNAs. Plant Cell. 2023:35(6):1762–1786.

7. Ponting CP, Haerty W. Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. Annu Rev Genomics Hum Genet. 2022 Aug 31;23:153-172.

8. The International Human RNome Project Consortium. Unlocking the regulatory code of RNA: launching the Human RNome Project. Genome Biol 26, 367 (2025).

9. NIRBA