

Statistics-First Genomic Analysis: Expanding Detection beyond the Limits of Reference Alignment in Precision Medicine

Executive Summary

Precision medicine relies on the paradigm that genotype determines phenotype. By identifying specific genetic variations, or expressed RNA variants, clinicians and researchers aim to diagnose disease, predict outcomes, and tailor therapeutics to individual patients. However, current genomic workflows depend heavily on alignment to a reference genome—a paradigm that inherently filters, biases, and drops critical sequence data, and is premised on the faulty premise that reference genomes exist. Tumors are highly heterogeneous, and many neurodegenerative, as well as other diseases are now known to have heterogeneous DNA, e.g. through repeat expansions.

This white paper introduces a fundamentally different approach based on a **statistics-first analysis of raw nucleic acid sequence data**. By operating directly on raw reads without initial reference mapping, ROSA provides a **panoptic** view of the genome and transcriptome, achieves **integrated detection** of genotype-to-phenotype associations, and delivers unprecedented computational **efficiency and statistical rigor**.

The Paradigm Shift: From Reference Alignment to Direct Statistics

Traditional Reference-Aligned Pipeline:

Raw Reads -> Read Alignment (Reference Bias / Data Loss and filters) -> Variant Calling -> Association Analysis

ROSA Statistics-First Pipeline:

Raw Reads (DNA/RNA) -> Direct Statistical Analysis -> Panoptic Interpretation, including Deep Learning-> Contextualization

The Bottleneck in Current Precision Medicine

Precision medicine requires an expansive view of genetic variation. Current standard analytical pipelines follow a rigid process: raw sequencing reads are mapped against a linear reference genome (such as T2T or a pan-genome), after which algorithms look for differences (variants) between the sample and these references. While this approach enabled the first era of genomics, it suffers from intrinsic limitations, even if patched with a pan-genome approach:

- **Reference Bias:** Sequences that deviate significantly from the reference—such as highly rearranged or mutated regions, novel insertions, highly polymorphic regions, or foreign genetic material—fail to align and are discarded. Splice and other RNA variants are difficult to detect; detection of methylated or edited bases require specialized aligners.

- **The "Invisible" Variant Problem:** A genetic variant that cannot be aligned cannot be detected, and a variant that is not detected cannot be appreciated. Current pipelines discard millions of valid sequencing reads per sample because they do not match the reference framework. It can be difficult to distinguish real biological variation and technical noise— as both cause alignments that deviate from the reference.
- **Sequencing Error Misinterpretation:** Standard mappers often confuse technical sequencing noise with real biological variation, or conversely, drop true low-frequency variants to suppress noise.
- **Loss of Context in Complex Phenotypes:** Complex traits and disease states are rarely driven by a single isolated single-nucleotide variant (SNV). They are the culmination of multiple interacting genetic elements across both coding and non-coding regions.

Rosa Advantages

1. Panoptic

The ROSA platform operates panoptically—providing a comprehensive view across the entirety of a sample's nucleic acid content. By removing reference mapping from the initial phase of discovery, ROSA evaluates **100% of raw read sequence data**:

- **Coding Sequence:** Traditional exon variation.
- **Non-Coding Regions & Regulatory Elements:** Enhancers, promoters, intrinsic repeat expansions, and non-coding RNAs that modulate expression.
- **Flanking Sequences & Structural Context:** Regional architectures that influence chromatin accessibility and expression dynamics.
- **Highly Polymorphic Loci:** In hypervariable regions such as HLA, KIR and other regions, reference alignment methods often struggle or fail completely to map them accurately. ROSA can detect these complex alleles without requiring specialized, locus-specific aligners.
- **Discovery:** Linkage of sequence variants to functional phenotypes offers a new systematic approach to functionalizing sequences including those that are currently considered "unaligned" or difficult or impossible to interpret. These could be new drug targets, new keys for therapeutic stratification or subject stratification. Moreover, all disease is polygenic.. Variation in disease manifestation and spectrum of severity are driven by coordinated networks of genetic elements in RNA and DNA. ROSA can link these variations to functional phenotypes.

2. Eliminate Blind Spots

Sequence elements previously dismissed as "unmapped noise" are retained and evaluated statistically, unlocking hidden drivers of disease. Many of the regions that are difficult to map consist of repeated sequences (about 50% of the entire genome) or highly polymorphic regions (e.g., HLA).

By being reference-free, the platform is unbiased across human ancestries, avoiding the legacy focus on European populations and the complexity and incompleteness of pan-genome mapping.

3. Statistical Rigor: Direct Linkage from Raw Sequence to Phenotype

ROSA uses mathematical representations of raw sequence space to measure pattern abundance directly across phenotypically characterized cohorts to identify those sequence patterns that statistically associate with the phenotype of interest.

- **Statistical Rigor:** By applying rigorous multiple testing corrections and non-parametric statistical models directly to sequence distributions, ROSA uncovers underreported and underappreciated genetic variations with high statistical confidence. This includes building high dimensional machine and deep learning models to link sequence to phenotype.
- **Unbiased Marker Identification:** Unbiased discovery reveals non-obvious epistatic interactions and non-coding drivers that hypothesis-driven or exon-focused pipelines .
- **Reducing Wet-Lab Discovery:** Traditional target discovery requires years of candidate gene selection, cell line engineering, and biomarker validation. ROSA performs high-powered association testing across vast raw datasets computationally, linking novel sequence motifs to clinical outcomes in minutes to hours. As AI generates more compounds, efficiency *in silico* screening has increased importance.

4. Efficiency and use in machine and deep learning

Unified DNA and RNA Insights. Because ROSA operates directly on raw nucleic acids, it applies the same statistical core to both genomic DNA and transcriptomic RNA sequence data. This allows unified multi-omics co-analysis—correlating genomic variant abundance directly with transcript expression and splicing events in a single statistical pipeline. This bypasses concerns about overtraining and data-snooping.

Deep and Machine Learning. ROSA is integrated into a deep learning model that parses terabytes to petabytes of data. Its smaller proof of principle prototypes perform blind "cross-cohort predictions," analyzing multivariate covariation in raw sequence data to predict disease phenotypes in independent studies. This methodology has demonstrated strong predictive accuracy:

- **Melanoma:** Predicting patient response to anti-PD-1 (nivolumab) across datasets (e.g., GSE28943) achieved a cross-study AUC >.76.
- **Parkinson's Disease:** Initial applications across two studies yielded accuracy >.69, which improved to >.87 using Rosa's latest deep learning model.

Scale. ROSA provides a scalable, efficient framework designed for modern petabyte-scale genomic repositories.

Efficiency. Rosa's platform operates at 10x the speed of legacy tools while requiring folds fewer— sometimes 10x less sequencing depth, drastically reducing computational costs.

In addition, The ROSA approach offers key advantages over traditional alignment pipelines across multiple architectural dimensions.

- By utilizing 100% of raw read data rather than dropping 5–20% or more as unmapped or filtered reads, ROSA eliminates the high reference dependencies that make traditional pipelines vulnerable to reference update shifts, using mapping solely for downstream annotation.
- Mathematically, ROSA simplifies computational overhead by replacing heavy CPU and RAM mapping ($\mathcal{O}(N^2)$ per reference base) with streamlined pattern counts and matrix operations.
- Additionally, ROSA provides a unified mathematical framework for DNA and RNA alongside native detection of variable alleles such as HLA and TCR/BCR unlike traditional pipelines that process DNA and RNA in isolated silos and rely on specialized target-enrichment pipelines for hypervariable loci.

Rosa's P2G (Phenotype-to-Genotype) technology has demonstrated power in identifying both known and novel disease-causing variants across diverse pathologies. Some example key achievements are highlighted for its completely unsupervised rediscovery of the following disease alleles:

- **Clinvar Rediscovery:** P2G successfully rediscovered the Pi*Z allele (rs28929474) in the *SERPINA1* gene, the common disease-causing variant for Alpha-1 Antitrypsin Deficiency (AATD), establishing that causal mutations can be identified ab initio from RNA-seq data.
- **Polymorphic Loci Analysis:** The platform excels in highly variable regions where mapping typically fails. It identified tuberculosis susceptibility loci (rs1042140 in HLA-DPB1) and HLA-A variants (rs1061235, rs1061156) associated with multiple autoimmune conditions, including rheumatoid arthritis and systemic lupus erythematosus.
- **Phenotype Prediction:** In immune checkpoint inhibitor-induced liver injury (ChILI), P2G identified *FAM210B* as a top predictor, consistent with established causal links to drug-induced liver injury risk. Furthermore, in drug-resistant epilepsy, it identified variants in *NAPB* with a prediction AUC >0.95, linking it to abnormal neurotransmitter release and seizures.

Broad Applications across Life Sciences

The ability to perform reference-free, statistics-first sequence analysis has potential application for any genetic analyses, e.g., for human health diagnostics, pharmacogenetics, and/or therapeutics. Additional potential applications are for diagnosis and treatment of veterinary diseases, agriculture (breeding), microbiome analyses and environmental assessments.

Conclusion

The future of precision medicine depends on seeing the genome as it truly is—not as it conforms to a pre-defined reference. By substituting initial reference alignment with statistical rigor applied directly to raw nucleic acid data, ROSA eliminates reference bias, recovers missing genetic signal, and establishes an efficient, panoptic paradigm for biomarker and target discovery across all domains of biological science.