Benefits of a Statistics-first Approach to analyzing Nucleic Acid Sequence Data

Precision medicine, from diagnostics and therapeutics to genome and transcriptome variant calling, relies on increasing throughput precision of next generation sequencing. Yet, as the technology for sequencing has reduced the costs and increased the quantity of sequencing that can be done, the paradigm for analyzing sequence data remains highly similar to that developed decades ago when the human genome was first sequenced. In particular, these legacy methods require that the sample nucleic acid sequence be aligned against a reference genome to identify any sequence variant(s).

Why the legacy analytic paradigm of “assemble and align” is problematic

The use of the reference genome-based sequence assembly and alignment approach has multiple inherent limitations, including attenuated discovery power, and substantial resource requirements, with significant consequences in terms of speed and cost.

The reliance on reference genomes introduces several critical points of failure:

- **Signal Filtering (The Reference Paradox):** By aligning data to a pre-existing reference, researchers inadvertently filter out "non-reference" sequences. This eliminates rich biological signals and novel variants that may be of the greatest clinical interest.
- **Reference Incompleteness:** No reference genome is truly complete. In highly variable regions—such as the Major Histocompatibility Complex (MHC) or areas with large structural variants (SVs)—the lack of a matching reference leads to imprecise detection or total data loss.
- **The Diversity Gap:** Current assemblies often exclude sequence diversity from understudied populations and fail to account for the rapid evolution of viral and microbial genomes.
- **The Paralog Problem:** Approximately **54% of the human genome** consists of repetitive or paralogous regions. Mapping algorithms struggle to assign sequences to these positions, leading to misalignments or "dark regions" of the genome where data is simply discarded.

Eliminating Algorithmic Bias

Reference-based mapping introduces three major sources of systematic bias:

1. **Exclusion:** Filtering out sequences that don't map well.
2. **Misassignment:** Incorrectly assigning sequences to the wrong genomic origin.
3. **Heuristic Randomization:** Many common mappers use randomized heuristics, meaning they fail basic consistency tests (e.g., a read and its reverse complement may map differently, or permuted read orders may yield different variant calls).

Rosa's Statistics-First Approach

Rosa has developed a fundamentally different approach to address these limitations. It entails the use of statistical signal detection directly from raw sequencing data, without a requirement for a reference genome. It relies on a formalization of sequence variation, e.g., but not limited to k-mers (short, overlapping sequences of length *k*). This enables a direct analysis of sequence variation in raw sequence data assessing the full genetic diversity in the sample. Rosa has developed a set of proprietary statistical software tools to further analyze such sequence variants for various applications.

Analyzing raw sequence data directly allows for:

- **Unbiased Discovery:** Identifying myriad forms of variation (indels, fusions, SVs, CNVs, etc.) without prior knowledge of a reference.
- **Mathematical Rigor:** Analysis of k-mers provides **exact p-values** rather than the approximations common in permutation testing and machine learning heuristics.
- **Efficiency:** Use of k-mers allows analysis at **10x the speed** of legacy tools while requiring **2x less sequencing depth**

Emerging tools for measuring raw sequence diversity, such as k-mers, have the potential to increase signal detection rates to discover new biology missing from the reference genome as demonstrated, e.g., by the SPLASH suite of algorithms (SPLASH (5), SPLASH2 (6), sc-SPLASH(7)). These methods are an example of the power of statistics-first methods that operate on raw k-mers. For example, single cell SPLASH discovered a new family of secreted immune related genes missing from assemblies generated with the highest fidelity PacBio IsoSeq assembly algorithms. SPLASH2 and scSPLASH are also more than an order of magnitude more efficient than their reference-based analogues. Together, these findings demonstrate examples of the power of statistics-first methods in the realms of basic biology, and demonstrate their potential k-mer-based methods in biomedical applications.

Moreover, methods based on analysis of raw sequences, such as but not limited to k-mers, enable statistical tests that are rigorous and provide exact p values, versus approximate p values. An important but often unappreciated fact is that permutation testing, a mainstay of many machine learning and applied genomic statistical analyses, does not provide exact p values in general (8). Inaccurate p values translate into statistical inference that is also inaccurate.

This statistics- first analytical approach allows unbiased detection of all structural variation in nucleic acid sequence data and enables the Rosa proprietary platform to be highly accurate, faster, cheaper, requiring far less human intervention and be readily scalable. And, it utilizes

machine learning as an integral part of its tool development: the data output from the Rosa software platform is AI-ready.

Precise disease detection and monitoring

Alone, statistical detection of sequence variation is not enough to inform clinical or biomedical decision making. Rosa has developed proprietary tools that provide analysis and interpretation of the sequence variation identified, allowing sequence diversity to be linked to phenotype(s). Notably, the results attained by Rosa are orthogonal to results attained with assembly and alignment methods, as the Rosa results are based on sequence variation but not gene panels.

Sequence diversity can take many structural forms (indels, fusions, SV, CNV, MSI CIN, etc.) and vary from small, e.g., a single base change to large, e.g., chromosomal arm translocations. Many diseases are known to be caused by or correlated with certain genetic changes or mutations, and additional associations are being recognized as sequencing becomes more common. In cancer, genetic instability leading to sequence complexity takes many forms; genome instability is the hallmark of many cancers (9). Structural sequence variants have been associated with many diseases other than cancer, e.g., neurodegenerative diseases characterized by repeat expansions, e.g., ALS, Myotonic Dystrophy and many ataxias (10). Accurate and detection of the sequence variants associated with specific conditions enables earlier and more accurate diagnosis and leads to improved therapy.

Rosa has built and is building statistical tools that can be used for diagnostic or monitoring purposes, e.g. , to identify sequence variants for patient stratification, MCED and MRD, and for therapeutic purposes, e.g., target discovery and drug discovery and development.

1. https://academic.oup.com/bioinformatics/article/25/24/3207/235151
2. https://link.springer.com/article/10.1186/s13059-021-02558-x
3. https://www.nature.com/articles/s41592-022-01457-8
4. 4. Pelin Icer Baykal, Mike Simonov, Dhrithi Deshpande, Ful Belin Korukoglu, Jaden Moore, Karishma Chhugani, Cecilia Liu, Varuni Sarwal, Neha Rajkumar, Mohammed Alser, Niko Beerenwinkel, Serghei Mangul. Assessing genomic reproducibility of read alignment tools. bioRxiv 2025.05.08.652934; doi: https://doi.org/10.1101/2025.05.08.652934
5. K. Chuang, T. Baharav, G. Henderson, P. L. Wang, I. Zheludev and J. Salzman (2023). SPLASH: a statistical, reference-free genomic algorithm unifies biological discovery. (Cell) https://www.biorxiv.org/content/10.1101/2022.06.24.497555v4
6. M. Kokot, R. Dehghannasiri, T. Baharav, J. Salzman*, S. Deorowicz* (2024). SPLASH2 provides ultra-efficient, scalable, and unsupervised discovery on raw sequencing reads. Nature Biotechnology

7. Dehghannasiri, R., Kokot, M., Starr, A.L., Maziarz, J., Gordon, T., Tan, S.Y., Wang, P.L., Voskoboynik, A., Musser, J.M., Deorowicz, S. and Salzman, J., 2026. sc-SPLASH provides ultra-efficient reference-free discovery in barcoded single-cell sequencing. *(Nature Biotechnology, in press)*
8. *Romano, J.P., 1990. On the behavior of randomization tests without a group invariance assumption. Journal of the American Statistical Association, 85(411), pp.686-692.*
9. *Simovic-Lorenz, Milena, and Aurélie Ernst. "Chromothripsis in cancer." Nature Reviews Cancer 25.2 (2025): 79-92.*
10. Paulson, H., 2018. Repeat expansion diseases. *Handbook of clinical neurology*, *147*, pp.105-123.