

Rosa Bio: Novel Algorithms for interpretable, ultra-efficient precision medicine

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 throughput of data acquisition, the paradigm for analyzing sequence data remains highly similar to that developed decades ago when the human genome was first sequenced.

I: Categorical problems with current sequencing

- 1) The instrumentation used to sequence nucleic acids has advanced dramatically, allowing much more sequencing and lower cost, however the analytical methods used to analyze sequences - assembly and alignment. - originated in the human genome project 25 years ago: **These approaches are partially randomized and include heuristics introducing biases and blind spots, and noise and imprecision in downstream assays.** In summary,
 - a) they are computationally intensive and
 - b) they are lossy
 - c) they are expensive in human and compute time, and slow
 - d) they miss important variation at the sequence level- such as in HLA where there is high polymorphic diversity and in plants and microbes
 - e) in RNA, splice variants, microexons and other sequence variants are missed– including mobile genetic elements
 - f) in single cell RNA seq, the low depth of the sample means that alignment errors confound discrimination between a true variant and a mapping or sequence artifact
 - g) gene quantification is based on alignment, and suffers from batch effects and is generally non-reproducible across labs (Leek et al). Furthermore, different isoforms of the same gene can have opposite functions.
 - h) in DNA, recurrent mutations in spliceosomal RNAs have been shown to be driver mutations; and any such variant transcribed to RNA
 - i) Examples: rRNA, centromeres, telomeres, sines
 - i) for detecting modifications such as methylation, masking of the genome reduces its complexity and means more signals are missed and attenuated due to lower complexity – e.g. more ambiguity in the masked genome – in addition to computational cost
- 2) Long read technology has value for some, but not all, areas of biology and medicine. For example, liquid biopsies are emerging powerful approaches to monitor and diagnose diseases– but the limited physical size of nucleic acid in the blood, urine or ct fluid means that longer reads are not a panacea for precision medicine. **What is needed is a better way to extract information from short reads from circulating nucleic acids.**
- 3) Deep learning or ML models, including those trained on multimodal data such as protein, RNA and DNA (and beyond) are **currently limited because they are trained on ted signals extracted from current assembly and alignment algorithms.**

II: Rosa has developed highly efficient algorithms that overcome these limitations

While many specific applications of clinical genomics allow different human interpretations, there is a unified mathematical formulation of a general problem in diagnostics and therapeutics: namely, **statistically identifying a genomic signal that predicts a phenotype**. Phenotype, broadly defined, can be a set of variant calls in a genome (RNA or DNA) or a biomarker (of any sort, e.g., biochemical, or a clinical response).

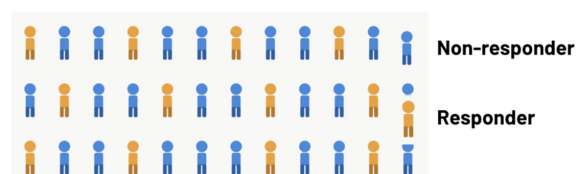
Rosa has used this idea to develop a **unified set of algorithms for sequence analysis that work on RNA, DNA, or any converted nucleic acid, and includes single cell and bulk measurements. It uses novel statistics-first approaches to analyzing all sequencing modalities and analytes.** This overcomes bespoke software design for different modalities and analytes. It is a solution consistent with the “the bitter lesson” taught by Sutton: “The bitter lesson is the observation in artificial intelligence that, in the long run, general approaches that scale with available computational power tend to outperform ones based on domain-specific understanding because they are better at taking advantage of the falling cost of computation over time.”.

P2G is Rosa's first generation of machine learning algorithms that bypass bespoke customized workflows. A single algorithm was designed and run blindly on hundreds of publicly available studies– phenotype includes drug response, disease status and disease subtype among others. In each case, features chosen by the model represent genomic features that predict the phenotype.

Rosa Core Algorithm: Phenotype 2 Genotype (P2G)

01

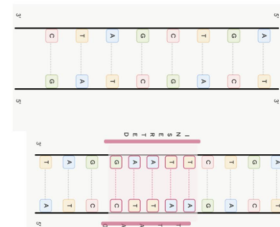
Today: Heterogeneous Patient Population



02

P2G: Unlocking Mechanism & Response

Unlocks underlying genetic cause, clinical response, and targetable biological mechanisms.



Advantages of P2G

Patient Stratification

Ab initio prediction enabling precise, robust patient stratification cohorts.

Signal Deconvolution

Interpretable attribution of both genetic variation and expression levels directly to clinical phenotype.

Practical Key Advantages

1. It is statistically efficient and fast
 - a. Easy to reanalyze legacy data and streamline pipeline design moving forward
2. Unbiased across ancestries
3. Integrates biological signals including gene expression, variants– including structural variants and highly polymorphic loci such as HLA or centromeres
4. Implicitly controls for batch effects
5. Single, panoptic streamlined algorithms unify: patient stratification, MRD, drug response prediction, disease status across multiple diseases, all with interpretable feature attribution
6. In scRNA-seq cell type, gene expression and morphology prediction including variant calls at the same resolution as bulk samples
7. Variant calling in RNA and DNA in tumor, including in single cell

Rosa's algorithms can rapidly parse terabytes or petabytes of data to extract signals that train a new generation of deep learning.

One aspect of the Rosa approach is that it provides interpretable summaries and performs completely blind cross-cohort drug and disease stratification. **A key property of Rosa's approach is that its features are interpretable and biologically general: including combinations of gene expression variant and splicing, copy number, structural variation and point mutations, mobile element insertions or repeat expansions, and beyond.**

Today, each of these biological features is tested separately by custom bespoke algorithm detection which loses unbiased statistical discovery power. Submodules of Rosa's algorithms can be used to restrict attention to one class of features (e.g. to detect methylation).

III. Examples of concomitant improvements in prediction and feature discovery

P2G is Rosa's first generation of machine learning algorithms that bypass bespoke customized workflows. A single algorithm was designed and run blindly on hundreds of publicly available studies— phenotype includes drug response, disease status and disease subtype among others. In each case, features chosen by the model represent genomic features that predict the phenotype.

A series of increasing complexity in testing and model development illustrates the power of Rosa's P2G technology in identifying reproducibly predictive features. This **allows the algorithm and model to learn phenotypes from sequences and feeds a virtuous cycle of development wherein future predictions improve**. Examples of phenotypic predictions are shown in Table 1(attached).

Illustrative examples of the application of P2G

ICI-induced Liver Damage. Checkpoint inhibitor-induced liver injury (ChILI) study of toxicity (phenotype) under combination immunotherapy (Anti-CTLA-4 + Anti-PD-1).data from GSE287540 was analyzed. Peripheral blood mononuclear cells (PBMCs) were sampled before and after immune checkpoint inhibitor (ICI) treatment in patients who either developed ChILI or remained free of liver injury. No predictive metrics were provided by the authors. One of the top predictors identified by P2G is FAM210B. This finding is consistent with a recent study based on Mendelian randomization establishing a causal link between FAM210B expression levels and drug-induced liver injury (DILI) risk.

Stratification of Drug-resistant Epilepsy. Drug resistant epilepsy (DRE) is heterogeneous, with about 30% of patients not responding to anti-seizure medications. Individual epilepsy types such as TLE-HS, FCD, and TSC each have characteristic transcriptional changes. P2G was used to provide insights into subtype stratification in epilepsy with data from GSE256068. <https://www.nature.com/articles/s41467-024-46592-2> . As one example, P2G identified variants in NAPB as a significant subtype prediction AUCs (>.95). NAPB has been shown to have the following functional properties: tied to epilepsy [<https://pubmed.ncbi.nlm.nih.gov/26235277/>]:

- synaptic vesicles cannot be efficiently recycled,
- neurotransmitter release becomes abnormal,
- neuronal networks become hyper excitable,
- seizures develop.

HLA variation. Notably, P2G is able to concurrently identify predictors that include HLA allele

variants represent a combination of known and novel biomarkers - candidates for monitoring and therapeutic stratification. **The concurrent identification of sequence variants of different types enables both stratification and the identification of gene interactions which is difficult to achieve using other approaches.**

III. Globally, P2G rediscovers known Clinvar variants across blind prediction tasks on a subset of the short read archive.

Genome wide association studies have been the historical method used to map genetic variation to disease phenotypes. However, statistical methods should be able to find such associations in expressed RNA from publicly available RNA sequencing data. Moreover, more efficient statistical methods should have power to identify such variants without the large cohorts required in GWAS. We used P2G to test this across multiple diseases.

Rosa variant calling within P2G maps and interprets features including assignment of sequence variants to known SNPs in dbSNP and Clinvar. Multiple examples of known true positive associations between SNPs and disease prediction were found in the features P2G identified in RNA-seq. These studies only **require a much lower sample size** (in the 100s) than those used for GWAS which are multiple orders of magnitude higher— which is possible due to higher statistical power.

As an illustration, AATD is Alpha-1 Antitrypsin Deficiency — a genetic disorder caused by mutations in the SERPINA1 gene, which encodes alpha-1 antitrypsin (AAT), a protein made mainly in the liver that normally protects lung tissue from damage by the enzyme neutrophil elastase. The most common disease-causing variant is the Pi*Z allele (rs28929474, a Glu342Lys substitution). People homozygous for this allele (PiZZ genotype) have severely reduced circulating AAT levels. P2G found a SNP in SERPINA1—rs28929474—in GSE109515 data was a true positive — the study's groups (PiZZ vs PiMZ) are defined by this genotype, establishing that causal mutations can be identified by P2G.

Other examples of *ab initio* rediscoveries by P2G are summarized below.

Gene/Variant	Study (GSE)	Disease Context	Finding
SERPINA1 — rs28929474 (Pi*Z allele, Glu342Lys)	GSE109515	Alpha-1 Antitrypsin Deficiency (AATD); PiZZ vs PiMZ	True positive — the study's genotype groups are defined by this variant. PiZZ (homozygous) causes severely reduced circulating AAT, which normally protects lung tissue from neutrophil elastase damage.
TTN — chr2:178,550,169 G>A (p.R18140*), chr2:178,579,800 G>T (p.S10049*), chr2:178,547,234 C>A (p.E19014*), chr2:178,536,357 C>A (p.E21047*)	GSE165303	Dilated cardiomyopathy vs. non-failing heart	All four are premature stop (truncating) variants classified Pathogenic/Likely_pathogenic, recapitulating the known truncating-variant mechanism for TTN-related DCM.
APOE — rs429358	GSE214846 (HCC) and GSE98894 (GEP-NET)	Hepatocellular carcinoma; neuroendocrine tumor	Coding SNP that, with rs7412, defines APOE ε2/ε3/ε4 haplotypes (known for Alzheimer's/ cardiovascular risk); also a literature-documented HCC susceptibility variant. A 2026 HCC GWAS review groups APOE with PNPLA3, TM6SF2, TERT, HSD17B13, HFE, and MTARC1 as aetiology-specific risk genes for non-viral (metabolic/alcohol-related) HCC, via lipid metabolism/hepatic steatosis.
SCGB1A1/CC16 — rs3741240 (A38G)	GSE201955 (asthma vs. control) and GSE309698 (pediatric severe exacerbation)	Asthma	Functionally validated variant that measurably lowers Clara Cell 16 protein/mRNA; repeated candidate-gene associations with asthma susceptibility and severity. Annotated "Established_risk_allele association."
HLA-DPB1 — rs1042140	GSE328391 (TB airway immunity) and GSE91061 (melanoma anti-PD-1)	Tuberculosis; melanoma immunotherapy response	Well-established, genome-wide-significant TB susceptibility locus (first identified in East/Southeast Asian GWAS); also detected in a melanoma anti-PD-1 response study.

HLA variants. Many disease-associated SNPs (as measured by GWAS) identified by P2G are in the Major Histocompatibility Locus (HLA). This illustrates P2G's defining power to identify variants in highly polymorphic loci where mapping strategies struggle

Examples include rs1042140 (Human Leukocyte Antigen-DPB1) in GSE328391 — a tuberculosis airway-immunity study. Human Leukocyte Antigen-DPB1 is a well-established, genome-wide-significant tuberculosis susceptibility locus first found in East/Southeast Asian genome-wide association studies (Png et al, Mahasirimongkol et al). The SNPs rs1061235 and rs1061156 in Human Leukocyte Antigen-A were P2G-chosen predictors in several autoimmune diseases including rheumatoid arthritis, systemic lupus erythematosus, giant cell arteritis, type 1 diabetes, and juvenile idiopathic arthritis/inflammatory bowel disease (McCormack et al, Carmona et al). Human Leukocyte Antigen-A associations are documented in several of these (giant cell arteritis especially, via Human Leukocyte Antigen-A*68).

Many other predictors were also identified, including, splice variants, sequences with multiple SNPs, and unmapped reads, among others. This implies that P2G **can improve the ability to perform patient stratification and disease monitoring by detection of multiple variants that affect phenotypic diversity**. The ability of P2G establishes its ability to identify true positive signals suggests the ability to detect other true variant signals that have not been identified

Section IV. Rosa machine and learning models were trained to enable blind cross-cohort prediction.

Rosa has built its algorithms stepwise, into a deep learning model. In brief, sequences identified in training cohorts are modeled — in terms of their multivariate covariation — i.e. how they are co-related. **The training data is used to develop a function on parsed, raw, sequence data.**

This function predicts disease phenotype. The concept behind this prediction is analogous to a polygenic risk score, though has much broader clinical application as well as distinct statistical and biological properties.. We call a model prediction on data from a study never seen in training a “**cross-cohort prediction**”. Below, we report a handful of examples of a growing number of cross-cohort predictions.

For example, a model trained on patient response in melanoma to nivolumab, an anti-PD-1, dataset (GSE28943) from a clinical trial run on samples from patients, including but not limited to melanoma, treated with anti-PD-1 (Baruch et al), cross study AUC >.76. The same methodology was applied to two studies of Parkinson’s disease. Cross study prediction accuracy >.69. The blind cross study prediction accuracy increased to >.87 using the first deep learning model produced with Rosa’s methodology.

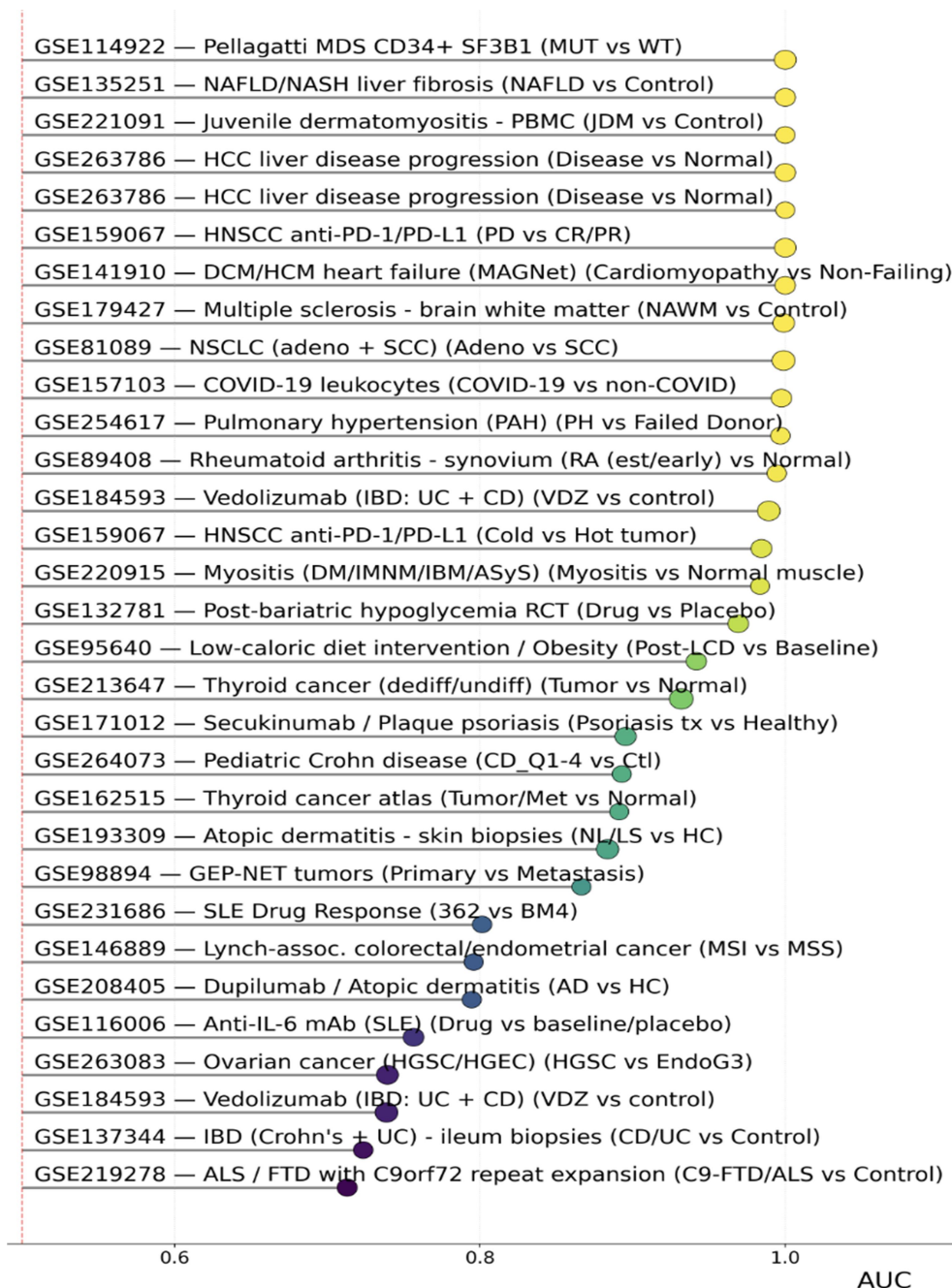
In addition to the value of the prediction itself, the features used for prediction give new insights into sequence data and functions of sequence variants. **Sequence variants and how they are co-varying are identified using a transformer architecture that feeds back into sequence parsing approaches that improve the deep learning model.** In addition, Rosa uses these covarying features to build more powerful and efficient training.

V: Conclusion and further steps

Rosa believes that its tools can enhance precision medicine in multiple ways. The ability to differentiate disease subpopulations can allow: 1) improved detection and monitoring of disease states, including cancer progression, 2) the ability to predict patients that will be harmed by (adverse effects) benefit from (responders) and/or not benefit from (non-responders) specific therapeutics, enhancing both drug safety and enabling more efficient drug development, and 3) allowing the identification and prediction of the effects of variant sequences for target discovery and drug development; 4) enabling new drugs to be developed, including cancer vaccines. HLA-restricted therapies (e.g. CAR-T, RNA drugs, anti-infectives and other agents).

Table 1

Data was randomly sampled from the short read archive (SRA). Each study was split into blind data never seen during training. Example AUCs with drug class defining color and study size reflected in dot size. All AUCs are reported on blind data never seen in model training.



References

- Baruch, Eran N., et al. "Cancer-Induced Nerve Injury Promotes Resistance to Anti-PD-1 Therapy." *Nature*, vol. 646, no. 8084, 2025, pp. 462–473. <https://doi.org/10.1038/s41586-025-09370-8>.
- Candelaria, P. V., Backer, V., Laing, I. A., et al. "Association between Asthma-Related Phenotypes and the CC16 A38G Polymorphism in an Unselected Population of Young Adult Danes." *Immunogenetics*, vol. 57, nos. 1–2, 2005, pp. 25–32. <https://doi.org/10.1007/s00251-005-0778-2>.
- Carmona, F. D., Mackie, S. L., Martin, J. E., et al. "A Large-Scale Genetic Analysis Reveals a Strong Contribution of the HLA Class II Region to Giant Cell Arteritis Susceptibility." *American Journal of Human Genetics*, vol. 96, no. 4, 2015, pp. 565–580. <https://doi.org/10.1016/j.ajhg.2015.02.009>.
- de Bakker, P. I., et al. [Title not provided]. 2006. PMID: 16998491.
- Ghouse, Jonas, et al. "Genome-Wide Meta-Analysis Identifies Nine Loci to Be Associated with Higher Risk of Hepatocellular Carcinoma Development." *JHEP Reports*, 2025, article 101485.
- Herman, D. S., Lam, L., Taylor, M. R., et al. "Truncations of Titin Causing Dilated Cardiomyopathy." *New England Journal of Medicine*, vol. 366, no. 7, 2012, pp. 619–628. <https://doi.org/10.1056/NEJMoa1110186>.
- Ku, M. S., Sun, H. L., Lu, K. H., Sheu, J. N., Lee, H. S., Yang, S. F., and Lue, K. H. "The CC16 A38G Polymorphism Is Associated with the Development of Asthma in Children with Allergic Rhinitis." *Clinical & Experimental Allergy*, vol. 41, no. 6, 2011, pp. 794–800. <https://doi.org/10.1111/j.1365-2222.2010.03679.x>.
- Leek, Jeffrey T. "SVASEq: Removing Batch Effects and Other Unwanted Noise from Sequencing Data." *Nucleic Acids Research*, vol. 42, no. 21, 2014, article e161.
- Mahasirimongkol, S., Yanai, H., Mushiroda, T., et al. "Genome-Wide Association Studies of Tuberculosis in Asians Identify Distinct At-Risk Locus for Young Tuberculosis." *Journal of Human Genetics*, vol. 57, 2012, pp. 363–367.
- McCormack, M., Alfirevic, A., Bourgeois, S., et al. "HLA-A*3101 and Carbamazepine-Induced Hypersensitivity Reactions in Europeans." *New England Journal of Medicine*, vol. 364, 2011, pp. 1134–1143.
- Pang, E., Alisjahbana, B., Sahiratmadja, E., et al. "A Genome-Wide Association Study of Pulmonary Tuberculosis Susceptibility in Indonesians." *BMC Medical Genetics*, vol. 13, 2012, article 5.
- Roberts, A. M., et al. "Integrated Allelic, Transcriptional, and Phenomic Dissection of the Cardiac Effects of Titin Truncations in Health and Disease." *Science Translational Medicine*, vol. 7, 2015, article 270ra6. <https://doi.org/10.1126/scitranslmed.3010134>.
- Sutton, Richard. "The Bitter Lesson." *Incomplete Ideas*, 13 Jan. 2019, <http://www.incompleteideas.net/Incldeas/BitterLesson.html>.