



From 34,599 genes to reproducible genomic signals

A custom survival-modeling pipeline tested genomic associations across **230,000 seeded model fits**—then measured which signals persisted.

Built for reproducible computational infrastructure.

34,599

genes screened

1,047

TCGA-BRCA patients

$\sim 10^{601}$

possible feature subsets

230,000

seeded model fits

23

terms shared across all runs

The challenge

Breast-cancer RNA-seq data create an **extreme high-dimensional** ($p \gg n$) **problem**: tens of thousands of molecular features, comparatively few patients, clinical covariates, and possible gene-by-treatment interactions. A single penalized modeling run can return an appealing answer that changes when the random seed changes. Exhaustive subset search is impossible; a one-seed result is not enough.

The pipeline

Midnight Mechanism built a parallelized, reproducible Cox/LASSO workflow that:

- joined expression and clinical data and enforced consistent preprocessing;
- screened **34,599 genes** while adjusting for age, race/ethnicity, treatment, and cancer stage;
- modeled both gene main effects and gene-by-treatment interactions;
- generated **230,000 seeded model fits** (2,000 across 115 outer runs); and
- aggregated selection frequency into an annotated consensus output for downstream investigation.

What the repetition revealed

Across 115 pipeline runs, 2,094 distinct terms were selected at least once. Only **23 were shared across all 115 runs**. Among treatment-interaction terms, the two most persistent appeared in 106 of 115 runs (92.2%) and 98 of 115 runs (85.2%). The pipeline converted seed sensitivity from a hidden modeling risk into a measurable property of every candidate signal.



THE RESULT

A rerunnable, interaction-aware pipeline that separates persistent genomic associations from unstable selections—and produces an auditable shortlist for biological review and external validation.

Why it matters for oncology teams

The deliverable was not a one-off notebook or a claim of clinically validated biomarkers. It was a **defensible computational system**: custom analysis logic, repeatable execution, explicit stability measurement, and interpretable outputs. The same approach can be adapted to proprietary cohorts, alternative endpoints, additional omics layers, and client-specific translational questions.



Have an oncology analysis that has outgrown ad hoc scripts?

Custom computational pipelines for translational oncology.



midnightmechanism.xyz

Reproducible. Scalable. Built around your question.