Identifying Causal Genetic Variants Through Single Cell Sequencing

Introduction

Today, human genomics can sequence millions of genomes, profile billions of single cells, and train models that link DNA sequences to chromatin states and gene expression.

These massive datasets are used to connect genetic variants to biological functions, and to interpret how changes in the genome influence health and disease, guiding functional interpretation and the prioritization of therapeutic targets.

But the hardest clinical and discovery decisions still hinge on rare and private variants.

When a genetic variant is rare, the sample cohort must be sufficiently large to identify enough individuals carrying the variant to understand its impact. Therefore, the rarer the mutation, the larger the study cohort must be. For instance, a variant with a population frequency of 0.01% would require a study of 1 million individuals (the 1/f limitation). And if it is a private variant (present only in one individual or family line), it is virtually impossible to detect.

No matter how sophisticated, computational models trained solely on observational data, whether simple risk scores or complex deep learning systems, inevitably inherit these constraints.

Causal Genomics addresses this gap by shifting the focus from correlation to causation. Rather than asking which variants are statistically associated with a trait, it focuses on identifying which genetic variants directly cause a particular trait or disease. This field has been essential for uncovering causal genetic factors that informed mechanism-based drug development, rather than relying on gene associations where the variations may be correlated but not actually causal, leading to misleading targets and failed therapeutics.

By directly testing and validating causal relationships between genetic variants and molecular or cellular phenotypes, Causal Genomics provides a more direct path to discovering rare or private variants that would otherwise be missed by observational studies alone.

To enable this discovery, the field leverages synthetic biology and multiplexed assays of variant effect (MAVEs) to design, build, test, and learn from libraries of precisely defined variants in relevant cellular contexts. Foundational MAVE studies have shown that single–nucleotide–resolution maps of variant effects are possible when variants are intentionally created through genome engineering, rather than relying on chance sampling of natural variation.

Causal Genomics, however, can only scale effectively when every stage of the Design–Build–Test–Learn (DBTL), an iterative framework for experimental optimization, cycle grows together.

The 1/f limitation of observational genomics and the impedance mismatch between a fast Design/Build stage and a slow Test are two sides of the same problem.

This application note describes how Codebreaker and Parse Biosciences address both by combining Codebreaker's Causal Genomics DBTL platform, and Parse's Evercode™ single cell RNA-seq. Together, they form an impedance-matched, exponentially scaling DBTL loop that feeds modern biological models.

DBTL and Impedance Matching

In synthetic biology, scientists often use a structured process called the Design–Build–Test–Learn (DBTL) cycle to engineer biological systems.

The DBTL loop is a standard pattern in engineering, software, and synthetic biology:

researchers design a system, build it, test it, and use what they learn to improve the design in the next cycle.

In Causal Genomics DBTL is used to understand how genetic variations cause phenotypic changes. Design selects what to test: which variants, loci, and the experimental set up to read out the perturbation effects.

Design selects what to test: which variants, loci, and the experimental set up to read out the perturbation effects.

Build creates the genetic elements via synthesis and editing: DNA synthesis, CRISPR perturbations, base editing.

Test runs the experiment to collect the data to measure the effects on the phenotypes such as editing outcomes on gene expression or regulatory mechanisms.

Learn analyzes the results, builds predictive models, and uses those insights to inform the next round of design.

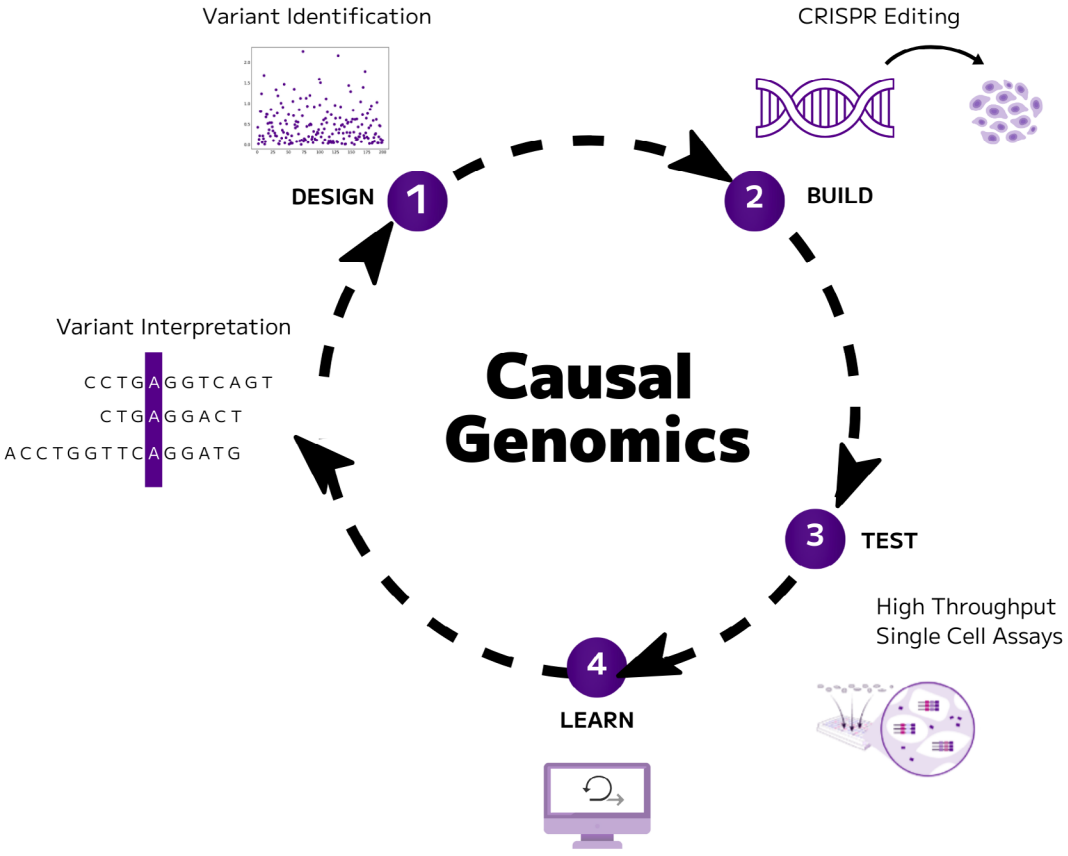

Figure 1: The DBTL loop is a standard pattern in engineering, software, and synthetic biology: researchers design a system, build it, test it, and use what they learn to improve the design in the next cycle.

A simple principle governs the overall speed of this loop: practical throughput is set by the slowest stage, and no matter how much researchers improve Design, Build, or Learn, the total cycle time won't improve unless the slowest step is sped up.

In practice, the overall throughput is bottlenecked by the stage with the lowest capacity among Design (D), Build (B), Test (T), and Learn (L). Thus, even if one can design and build a million variants per cycle, an assay capable of testing only one hundred thousand constrains the usable throughput to that number.

Over the past decade, Design and Build have grown exponentially. Array-based synthesis, pooled cloning, and barcoded perturbations have enabled researchers to increase library size and complexity while still maintaining low incremental costs.

On the other hand, Test has not kept this pace, and its growth has been linear. ELISAs and small multiplex immunoassays scale linearly with plates and robots. FACS can process a large number of events, but running multiple different conditions remains tied to instrument time. High-content imaging is information-rich but limited to the low throughput of the microscope.

If Test capacity increases only linearly while Design and Build scale exponentially, Test ultimately becomes the limiting factor, causing the entire DBTL pipeline to behave as a linear system.

We call the requirement that D, B, T, and L improve synergistically impedance matching. For Causal Genomics to operate as a truly scalable platform, Test must scale to match Design and Build while delivering broadly informative phenotypes.

The Test Layer We Need: Exponenial and Generalizable

\Therefore, the Test assay at the end of a Causal Genomics platform must meet two core conditions.

First, its output must scale at the same rate as Design and Build to meet the impedance-matching throughput requirement. In practical terms, as library design and synthesis, sequencing, and automation improve, the assay should naturally support more variants or cells growing from tens of thousands to millions per run without redesigning the assay.

Second, each measurement must capture enough biological information to answer many different questions.

A useful way to think about this in Causal Biology is:

> Useful information over time ≈ (number of cell variants measured) × (how well it reflects the true phenotype).

Which means that to get useful information over time, both the number of cells or variants measured and the amount of biological information delivered by each measurement must both be high.

To answer questions about mechanisms of action, toxicity, biomarkers, drug resistance and more, and to train biological analysis models, the assay must provide informationally broad phenotypes, not just one piece of information.

Many common assays fail as they do not meet both these criteria.

ELISA and small multiplex panels offer reasonable sample throughput but deliver only a handful of analytes per sample, which is ideal for targeted validation but narrow for discovery.

FACS, on the other hand, can collect measurements from millions of individual cells in one experiment. However, a FACS experiment records only a small number of parameters per cell, typically constrained by the available fluorescence channels and scattering measurements. In addition, FACS is limited in how many experimental

conditions can be explored, since each condition must be run separately on the instrument.

Cell painting generates highly detailed morphological fingerprints by imaging cells across multiple fluorescent stains, capturing thousands of features that describe cell shape, structure, and organization, but the number of cells or conditions that can be profiled in one single experiment is limited by imaging hardware and downstream analysis.

At the core of the DBTL loop, there must be a Test assay whose throughput matches Design and Build. A test where the number of cells or variants that can be measured per unit time increases as sequencing and computation ability improve, and where each measurement provides a rich, high-dimensional phenotype interpretable across many biological questions. In other words, Test should provide a molecular snapshot of many aspects of the cell (whole transcriptome profile, pathways activities, cell state, identity, response to stimuli), all at once.

Single cell RNA sequencing, especially when implemented through split–pool combinatorial indexing, meets these requirements: it scales the DBTL cycle by barcoding millions of cells and by sequencing and analyzing the resulting data in a single experiment.

The Codebreaker and Parse Platforms

Codebreaker's platform designs variant libraries at single nucleotide resolution for genes and regions where observational associations obtained through GWAS and eQTL studies are still unclear or underpowered, or where the standard computational and statistical models don't reliably predict a function.

The platform builds thousands to millions of variant libraries using array-based oligo synthesis, introduces these libraries into cells via pooled cloning or genome editing, and applies pooled selection assays. The effects of each variant introduced are then quantified by next-generation sequencing, enabling high-throughput tracking of individual variants.

This provides an exponential Design/Build substrate: each generation of synthesis and automation supports larger, more complex libraries at lower effective cost.

Parse's Evercode™ chemistry adds the missing Test layer.

The Evercode workflow uses combinatorial barcoding to uniquely label cells or nuclei resulting in sequencing ready libraries.

At the core of the technology are cell/nuclei fixation and a plate-based split-pool barcoding assay.

The Evercode™ split-pool barcoding process labels individual cells or nuclei using a combinatorial indexing strategy, eliminating the need for physical isolation of single cells.

Cells or nuclei are first fixed and permeabilized. Fixation preserves RNA integrity and enables samples to be stored for months prior to barcoding and library preparation, making the technology well suited for large or multi-site studies. Permeabilization turns each cell or nucleus into a self-contained reaction chamber for downstream molecular reactions. As a result, the workflow does not require microfluidic instrumentation and can be performed in any laboratory using standard equipment, enabling straightforward scale up and automation.

Barcoding begins with the first split-pool round, in which cells or nuclei are distributed across the wells of a 96-well plate containing well-specific barcodes. In this initial reaction, barcodes are ligated to mRNA molecules. The samples are then pooled, mixed, and redistributed into a second plate, where a second barcode is added. This split-pool process is repeated for a total of four rounds, with the final round attaching sequencing primers.

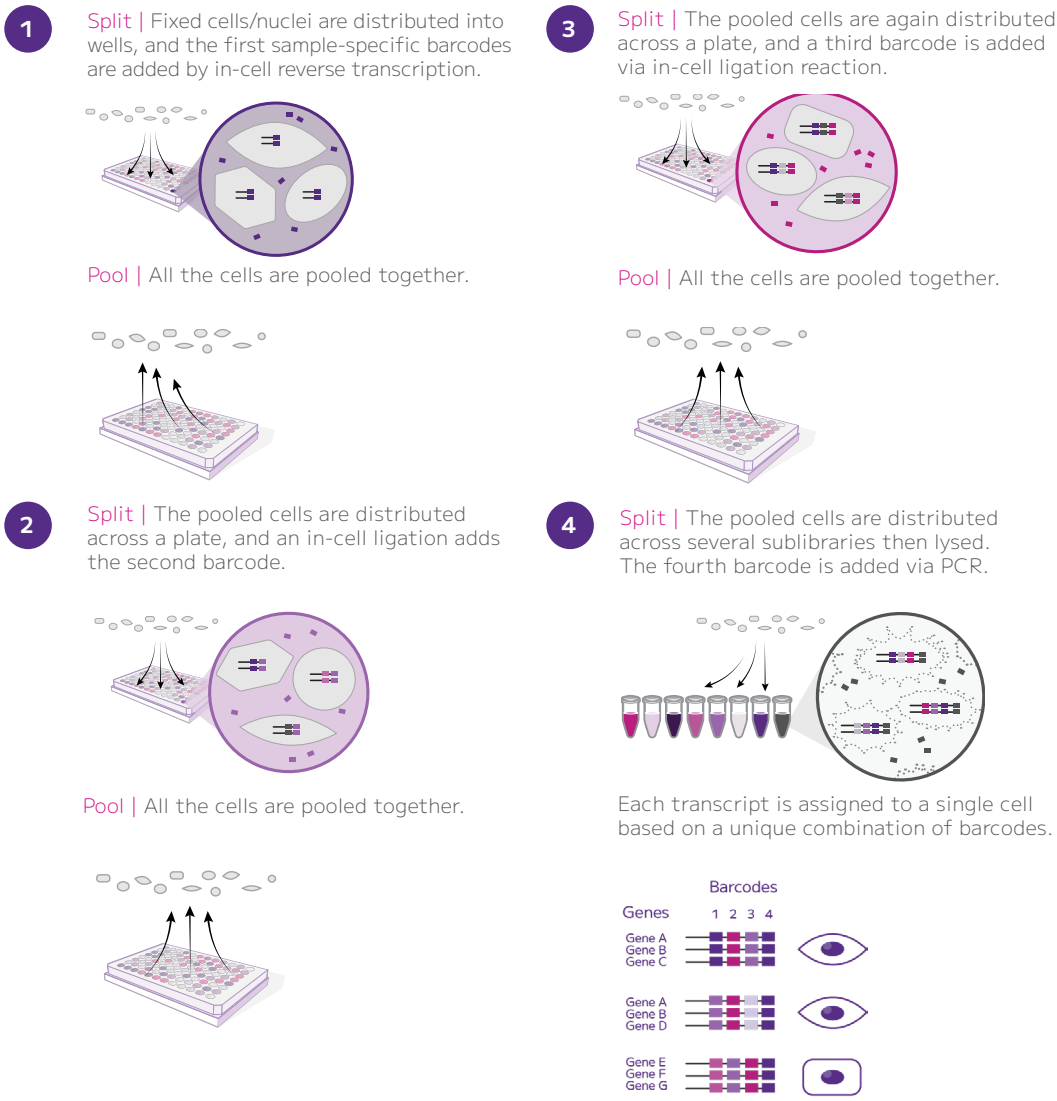

Figure 2: The Parse Biosciences Combinatorial Barcoding workflow. Cells are first fixed and permeabilized, turning them into their own reaction vessels. The split-pool barcoding process then labels cells with an exponentially large number of barcode combinations making it possible to easily scale beyond other technologies.

Because cells are randomly redistributed between plates after each round, the number of possible barcode combinations is extremely large. This makes it highly unlikely that two cells acquire the same combination of four barcodes. The resulting unique barcode combination serves as a cell-specific identifier, allowing transcripts from the same cell to be grouped computationally after sequencing.

Following barcoding, libraries are sequenced and reads are decoded based on their barcode combinations to generate single cell RNA expression profiles.

Evercode WT kits are available in multiple sizes, supporting experiments up to 5 million cells or nuclei. The kits enable processing of up to 384 samples in parallel using a plate-based, instrument-free workflow that runs on standard laboratory equipment and is easy to automate. Experimental capacity can be expanded by increasing the number of barcodes per round or the number of rounds, leveraging higher sequencing output, or running multiple plates in parallel.

Moreover, the innovative Parse Biosciences Gigalab expands automation to support larger-scale barcoding, starting at 10 million cells and beyond. This enables higher throughput, great consistency, and faster turnaround for population-scale studies.

The output is a whole transcriptome view of cell state, capturing expression of thousands of genes per cell. These data reveal cell identity, activation status, pathway usage, and stress responses. A

single dataset can be reused for many downstream applications, including variant-effect mapping, mechanism-of-action studies, toxicity and off-target profiling, biomarker discovery, and training or fine-tuning data-centric biological models.

In the combined Codebreaker × Parse platform, Codebreaker designs and builds libraries using exponential technologies, while Evercode single cell RNA-seq makes the test phase exponential and high-dimensional.

Finally, the analytics platforms from Codebreaker and Parse enable the learn phase, incorporating these data into increasingly powerful models.

Choosing a Back-end Test Assay

The same impedance and information criteria can be used to compare Evercode single-cell RNA-seq with common alternatives.

Assay Type	Throughput and scaling	Phenotype generizeability	AI Utiilty
Evercode scRNA-seq	10^4–10^7 cells per run; scales with barcodes & reads	Whole-transcriptome; broad use	High
sc multi-omics	10^4–10^5 cells; similar scaling	Adds regulatory or protein layers	High (spec.)
sc proteomics	10^3–10^6 cells; instrument-limited	Dozens–thousands of proteins	Med-high
sc chromatin / methylation	10^4–10^5 cells; indexing-based	Regulatory and lineage information	High
Cell painting / HCI	10^4–10^5 conditions; microscope-limited	Morphology-centric	Medium
FACS / flow cytometry	10^6–10^8 events; conditions time-limited	Tens of markers; good for gating	Low
ELISA / multiplex bulk	10^2–10^4 samples; plate-limited	Few analytes; narrow but precise	Low

Taken together, the DBTL and 1/f frameworks make Evercode-based single cell RNA-seq a natural default for the Test layer. It matches the scale of Design and Build through barcoding, sequencing depth, and automation, while delivering high-dimensional, generalizable phenotypes that can directly power biological models.

Conclusion

The Codebreaker and Parse platform establishes a new operating model for Causal Genomics. A model designed to study biology at full resolution, where individual-specific variants are observable, experiments are large enough to reflect real complexity, and computational learning scales with the data to reveal biological signals.

By integrating design, experimentation, and analysis in an impedance-matched DBTL loop, this approach moves genomics beyond association and toward a truly causal, model-driven understanding of biology. The result is more reliable target discovery, better therapeutic decisions, and a faster path from genome to mechanism to medicine.

About Codebreaker Labs

Codebreaker Labs is pioneering Causal Genomics by empirically mapping the functional effects of genetic variants at scale, moving beyond correlation to causation in genomics. Using high-throughput synthetic biology and single cell phenotyping, the company builds large experimentally validated datasets of variant effects to accelerate biological insight and AI-driven interpretation of genomic variation. For more information, visit codebreakerlabs.io or contact partners@codebreakerlabs.io.

About Parse Biosciences

Parse Biosciences is a global life sciences company whose mission is to accelerate progress in human health and scientific research. Empowering researchers to perform single cell sequencing with unprecedented scale and ease, its pioneering approach has enabled groundbreaking discoveries in cancer treatment, tissue repair, stem cell therapy, kidney and liver disease, brain development, and the immune system. For more information, visit parsebiosciences.com.