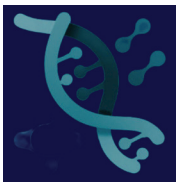

CASE STUDY

From Observational to Mechanistic Genomics: Establishing Causal Genomics as the Foundation for Scalable Variant Interpretation

Authors: Twist Bioscience and Codebreaker Labs

INTRODUCTION

Rapid advancements in next-generation sequencing (NGS) have produced a wealth of genomic data and exposed the staggering depth of human genetic variation. But in these depths, researchers have struggled to find meaningful insights. Codebreaker Labs, a pioneer of Causal Genomics, aims to improve this by solving a key challenge in genomic discovery—differentiating correlation and causation.

While NGS allows us to describe genetic variation in ever sharper detail, the functional consequences of that variation is often unknown. This is because modern genomics relies on brute-force, large-scale sequencing and association studies. With enough data, the effect of any given variant can be assumed by its association with a phenotype. Yet, such an approach is heavily dependent on variant frequency—the rarer a variant is, the harder it will be to gather enough data for statistically powered analyses. As rare variants make up a large portion of human genetic variation, association-based genomics does little more than gather sequencing data. The prolific absence of variant effect information in modern data sets greatly limits our ability to leverage genomics for drug development, diagnostics, clinical trial design, and AI modeling.[1]

“Once we know what a variant does, we can leverage that to more rapidly identify mechanisms underlying disease,” explained Ryan T. Gill, Co-Founder and CEO at Codebreaker. “This allows us to better predict targets, disease progress, drug response, and trial design among other valuable outcomes.”

To advance variant effect mapping, Codebreaker Labs has developed a platform for “Casual Genomics”, one that uses the tools of synthetic biology to prospectively map millions of variant effects at single-nucleotide resolution. Key to their success is the use of Twist’s DNA synthesis platform, which enables the efficient engineering of genomic variants on a large-scale. This efficiency, in turn, enables the team to reinvest in deep, single-cell phenotyping. Codebreaker’s approach promises to help the field of genomics become more mechanistic and transition its focus from correlation to causation on a large scale.

FROM OBSERVATIONAL TO CAUSAL

For the past two decades, genomics has largely been an exercise in observation, with researchers methodically recording genotypes and phenotypes in search of telling associations. Doing so often involves gathering multimodal data (such as genomic, transcriptomic, and clinical data) on a large scale. When paired with increasingly sophisticated computational tools, such an approach can be an effective way to discover clinically valuable biomarkers and shed light on potential gene functions, but it is also limited in some important ways:

1. Rare Variants Can Be Elusive: Though rare on an individual basis, as a collective group, rare variants are common.[2] The rarity of each variant means that they’re individually unlikely to be found in sufficient quantities within any given data set, thus frustrating attempts to draw statistically sound associations. Put simply, most studies lack sufficient power to draw conclusions about rare variants.

2. Incomplete and Biased Data Sets: Genetic datasets are prone to biases, often enriched for specific ancestral backgrounds which can limit representation of population-specific variants (many of which are rare). As a result, variants with potential clinical value are likely to be overlooked. Further, these data sets may contain incomplete or non-standardized phenotypic data that can confound or mislead study conclusions.

3. Correlation =/= Causation: Even when statistically strong associations can be found, the true mechanistic relationship between a given variant and its phenotype remains speculative until experimentally validated. Yet the scale of genomic observation has greatly outpaced validation studies. As a result, the genome is replete with variants of unknown significance.

These limitations mean that, despite our growing ability to observe the human genome, much of the data we gather lacks meaning. We may observe that variants are present, but we are unlikely to know if or why those variants influence human health and disease.

This is significant because data from genomic association studies is increasingly used to train AI and machine learning models. These models are expected to help researchers identify drug targets, interpret the significance of gene variants, and guide clinical trials. Without ground-truth data demonstrating variant effects, however, these models are built on assumptions. As such, the reliability of their predictions is greatly decreased.[1]

If we can produce ground-truthing data sets that match the scale of modern genomics—methodically demonstrating the downstream effect of genetic variants—we can improve AI training and extract more value from NGS' growing capabilities. Such is the goal of Codebreaker's Causal Genomics platform, which re-orients the data-generating process. Rather than passively analyzing what nature provides, they actively design and build the variants and measurements needed to understand human genetic variation.

"We're completely independent of variant frequency," explains Ryan Layer, Co-Founder and Chief Data Advisor at Codebreaker. "With our platform, we can just build and test whatever variants we want, and we can do it on a massive scale."

> **"Once we know what a variant does, we can leverage that to more rapidly identify mechanisms underlying disease. This allows us to better predict targets, disease progress, drug response, and trial design among other valuable outcomes."**
>
> RYAN T. GILL, CO-FOUNDER AND CEO
> CODEBREAKER

HOW CODEBREAKER'S CAUSAL GENOMICS PLATFORM WORKS

In essence, Causal Genomics is carried out by methodically testing the phenotypic effects for each of many thousands of individual variants in a high-throughput screening assay. The concept is born out of early work with CRISPR technology wherein researchers used Cas9 enzymes and a synthetic donor DNA template to introduce specific point mutations into a cell's genome (via homology direct repair).[3] By coding the template donor DNA in a barcoded DNA cassette, both editing and phenotyping of cells can be done in a pooled format, thereby enabling large-scale and high-throughput saturation variant screening.[4] Not only did these studies lay the methodological groundwork for Causal Genomics, but they illustrated three core principles that guide Codebreaker today:

- **Single-nucleotide perturbation is essential.** To truly interrogate the effect of an individual variant, researchers need the ability to manipulate the genome with single-nucleotide precision.
- **Endogenous context matters.** Editing endogenous loci, in particular if done in primary cells, preserves realistic regulatory and chromatin context, making the resulting variant-effect maps directly relevant to clinical and mechanistic questions. Causal Genomic screens should therefore seek to introduce single-nucleotide perturbations at endogenous loci through the use of base editors or, as described above, through the use of homology directed repair. Codebreaker Labs uses the latter method.
- **Synthetic DNA is the scaling engine.** Introducing genetic variants with precision requires the synthesis of many thousands of DNA fragments, each of which will serve as a template for homology directed repair. Synthesizing these fragments is thus a core enabling factor in Causal Genomics, as the quality and efficiency of DNA synthesis will determine the quality and scale of Causal Genomic screens. It is essential to find a reliable supplier of synthetic DNA.

Guided by these principles, Codebreaker Labs has developed a closed-loop, high-throughput Causal Genomics platform (**Figure 1**). The platform leverages Twist Bioscience's highly accurate, large-scale DNA synthesis platform to engineer a library of perturbed cells, each modified to contain one of the many thousands of genetic variants that are being studied. Single-cell assays are then used to generate tens of millions of phenotypic observations. This data provides a molecular-level view of how each variant alters cell function, in real time. As demonstrated in **Figure 1**, this model follows a design-build-test-learn format.

THE BENEFITS OF CAUSAL GENOMICS FOR RARE VARIANT STUDIES

Causal Genomics is particularly well suited to the study of rare variants. As described above, rare variants are collectively common. Yet, the rarity of any individual variant means that association studies will struggle to recruit enough carriers to draw statistically significant connections between the variant and a given phenotype. The rarer a variant is, the more people a study will need to recruit—and the more resources they will need to invest—in order to reach statistical significance.

Consider a variant with a minor allele frequency of 0.01%. If you need 1000 carriers to reach statistical significance, you would need to sequence roughly 5 million people. This illustrates how studies of rare variants can be a very costly endeavor, one that quickly becomes untenable.

> **"We're completely independent of variant frequency. With our platform, we can just build and test whatever variants we want, and we can do it on a massive scale."**
>
> RYAN LAYER, CO-FOUNDER, CHIEF DATA ADVISOR

In contrast the scale and cost of Causal Genomics is independent of variant frequency. Engineering and screening for an ultra-rare variant is as easy as engineering and screening for a common variant. Instead, the scale and cost of these studies is largely dictated by the efficiency of cell engineering and phenotyping. This means that the potential of Causal Genomics significantly benefits from advances in DNA synthesis which enable more efficient cell engineering.

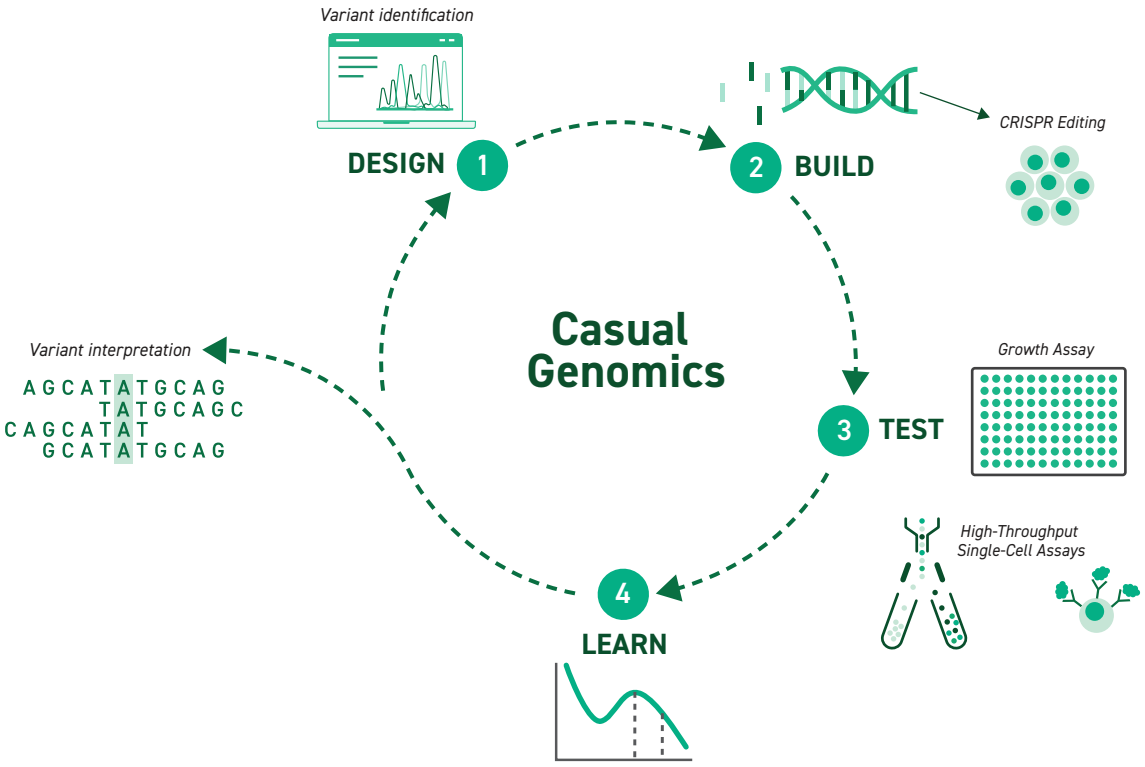

Figure 1: Codebreaker Labs' Design-Build-Test-Learn Cycle. 1 - Design: Variants of interest are identified using public databases and machine learning. Expression of affected genes is mapped across cell lines, and then mutagenesis libraries containing many thousands of variant donor DNA templates (and associated CRISPR guideRNAs) are designed in silico; **2 - Build**: Relevant experimental cell lines—such as primary human T cells—are selected for editing based on gene expression and biological context. Oligonucleotides encoding donor DNA and guide RNAs are synthesized on Twist Bioscience's synthetic DNA platform according to Codebreaker designs. Donor DNA is added into human cells via CRISPR-mediated homology directed repair; **3 - Test:** Engineered cells are phenotypically assayed using single-cell omics and growth assays, enabling the creation of experimentally validated maps that reveal variant effects; and **4 - Learn:** The resulting data is analyzed using machine learning and advanced AI techniques to infer functional consequences for variants by placing them in the context of biologically validated neighbors within a high-dimensional phenotypic space.

EFFICIENCY THROUGH DIRECT SYNTHESIS

In their effort to develop a high-throughput Causal Genomics platform, Codebreaker recognized the need for high quality synthetic DNA. Here, quality refers to the accuracy and uniformity of synthesized DNA.

Briefly, DNA synthesis is critical to Codebreaker's platform in two ways. First, synthetic oligonucleotides are used to build the donor template DNA that cells will use during homology directed repair. Each oligo will be precisely designed to contain the single nucleotide variant being studied. Second, synthetic DNA is used to build the guide RNAs (gRNA) that direct Cas nuclease activity, thus directing where the donor template DNA will be incorporated. In a Casual Genomic screen, this means designing tens of thousands of unique oligonucleotide sequences, each of which is valuable to the screen and each of which must be made precisely as it is designed.

This is why accuracy in DNA synthesis is essential. When synthesizing donor template DNA, errors will introduce new, unexpected genetic variants which may lead to erroneous phenotypic effects. Errors in gRNA synthesis can reduce editing efficiency and result in a large pool of unperturbed cells. At best, such errors are a waste of resources. At worst, they can mislead results interpretation and poison training data sets.

The uniformity of DNA synthesis is similarly important. A uniform oligonucleotide pool is one in which all oligo sequences are represented in equal proportions. This is difficult to achieve because some DNA sequences are easier to synthesize than others. A non-uniform oligo pool could result in under-representation of specific sequences. In the case of donor template DNA, non-uniformity can result in very little data being produced for some variants, reducing researchers ability to draw significant conclusions. Non-uniformity in gRNA sequencing can have a similar effect.

"Accurate and uniform oligo synthesis at scale is what unlocks our approach to Causal Genomics," Gill emphasized. "It also provides valuable differentiation from less accurate and uniform approaches such as base-editing."

Poor synthesis accuracy and uniformity can be compensated for by increasing experimental replicates, but doing so can be costly. For these reasons, Codebreaker set out to find a DNA synthesis provider who was capable of producing accurate and uniform synthesis on a large-scale. Their search ended with Twist Bioscience.

"We know from past experience that Twist can produce the high-quality oligos that this project demands," adds Tanya Warnecke, Co-Founder and CTO at Codebreaker. "Many of us have worked with Twist in our prior labs and it's left an impression. The quality they provide allows us to be efficient with our editing, which then opens the door to deeper phenotyping later in the pipeline."

Using Twist oligos, Codebreaker Labs is able to design large saturation screens wherein a single genetic loci is targeted for study. In such a screen, hundreds of donor oligos are designed to tile the space around the target locus, each encoding a unique SNV of interest. This allows for the methodical examination of every potential variant in a given locus. Donor oligos are produced as high-complexity oligo pools on Twist's silicon-based platform. Separately, Twist-synthesized gRNAs are combined with a Cas protein to form a ribonucleoprotein complex. This complex allows donor templates to be inserted into the target locus with extremely high efficiency, resulting in the precise introduction of designed variants.

The accuracy and uniformity of Twist's synthesis platform makes the build step of the design-build-test-learn cycle more efficient. Consequently, resources can be re-invested in the test stage, enabling the team to perform deep, single-cell phenotyping at scale. Already, the platform can achieve remarkable throughput and efficiency. In a 96-well plate, Codebreaker can generate >20,000 edited variants, enabling highly parallel mutagenesis or saturation editing campaigns.

"We know from past experience that Twist can produce the high-quality oligos that this project demands. Many of us have worked with Twist in our prior labs and it's left an impression. The quality they provide allows us to be efficient with our editing, which then opens the door to deeper phenotyping later in the pipeline."

TANYA WARNECKE, CO-FOUNDER, CTO

THE POTENTIAL VALUE OF CAUSAL GENOMICS

The most immediate application of scaled Causal Genomics is whole-genome sequencing interpretation in research and clinical settings. Saturation editing of BRCA1 and related genes has already shown that dense variant-effect maps can reclassify large numbers of variants of uncertain significance (VUS) and align strongly with established pathogenicity assessments.[4] Extending this gene-by-gene and panel-by-panel, Codebreaker can generate precomputed functional scores for high-priority clinical genes—hereditary cancer genes, cardiomyopathy genes, pharmacogenes—that clinical laboratories can consult alongside existing evidence. Instead of reporting "VUS" based solely on in silico prediction and limited observational data, laboratories can increasingly report "variant with measured effect in a disease-relevant assay", tightening interpretation and reducing ambiguity for patients and clinicians.

For mechanism-focused discovery biology, the same platform enables saturation and tiling experiments across regulatory elements, protein domains and pathways. Libraries that touch every base in an enhancer, every residue in a catalytic domain, or combinations across a signaling pathway can be built and edited at once, with phenotypes read out by fitness, reporter assays, or bulk and single-cell RNA-seq.[5, 6] These experiments provide detailed maps of which bases and motifs are necessary for function, how perturbations propagate through networks, and where biology is robust versus fragile. Such data feed directly into target identification, target validation and understanding of resistance mechanisms, and they are especially powerful when combined with high-content measurements such as single-cell transcriptional profiling.[5]

Causal Genomics data also provide a natural substrate for a range of modeling and AI efforts. Large observational models, including single-cell foundation models such as scGPT and sequence-based predictors, are powerful but limited by the quality and sparsity of their training labels.[7] Variant-effect maps derived from Causal Genomic screening offer high-quality, context-specific labels that can be used to train models from scratch for particular genes or pathways, to ground-truth and calibrate existing models, or to fine-tune broad foundation models.[6, 8] Codebreaker does not aim to compete in building all such models; instead, the intent is to serve as the causal data layer that enables model builders—academic groups, biotech companies, AI-native platforms—to train and validate their systems on mechanistic, experimentally grounded data.

Finally, these same variant-effect maps have clear implications for genotype-aware clinical trial design. If functional data indicate that certain variants strongly modify response or toxicity for a given mechanism, trials can be designed to enrich for carriers, exclude high-risk genotypes, or stratify randomization so that genotype is balanced across arms. Even in cases where no variant is used prospectively, post hoc analysis of trial outcomes using precomputed functional scores can help explain heterogeneity in response and guide the design of follow-on studies. In all cases, the key asset is the same: scalable, high resolution variant-effect data generated in the relevant biological context.

CONCLUSION

Observational studies have taken human genomics to a place where hundreds of thousands to millions of genomes can be read and linked with traits, and where increasingly sophisticated models can mine those datasets for associations. But reading genomes is not the same as understanding them, especially when clinical decisions, mechanistic insight and trial outcomes depend on the effects of specific variants in specific contexts.

Causal Genomics offers a path to that understanding at scale. Twist Bioscience's highly scalable and accurate oligonucleotide synthesis platform supports Codebreaker in transforming genomics into a proactive discipline, one where variant effects can be methodically precomputed across the genome.

REFERENCES

1. Goswami C, Chattopadhyay A, Chuang EY. Rare variants: data types and analysis strategies. Annals of Translational Medicine. 2021;9(12):961. doi:https://doi.org/10.21037/atm-21-1635

2. The 1000 Genomes Project Consortium. A Global Reference for Human Genetic Variation. Nature. 2015;526(7571):68-74. doi:https://doi.org/10.1038/nature15393

3. Garst AD, Bassalo MC, Pines G, et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nature Biotechnology. 2016;35(1):48-55. doi:https://doi.org/10.1038/nbt.3718

4. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature. 2014;513(7516):120-123. doi:https://doi.org/10.1038/nature13695

5. Kim HS, Grimes SM, Chen T, et al. Direct measurement of engineered cancer mutations and their transcriptional phenotypes in single cells. Nature Biotechnology. 2023;42(8):1254-1262. doi:https://doi.org/10.1038/s41587-023-01949-8

6. Stankey CT, Lee JC. The Role of ETS2 in Macrophage Inflammation. DNA and Cell Biology. Published online April 14, 2025. doi:https://doi.org/10.1089/dna.2025.0064

7. Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods. Published online February 26, 2024. doi:https://doi.org/10.1038/s41592-024-02201-0

8. McEwen AE, Tejura M, Fayer S, Starita LM, Fowler DM. Multiplexed assays of variant effect for clinical variant interpretation. Nature Reviews Genetics. Published online July 21, 2025. doi:https://doi.org/10.1038/s41576-025-00870-x