Unlocking Perturb-seq through Harmonization and Scalable Exploration

Rancho bio • sciences

Ben Ernest, Sondra Kopyscinski, Corey Oravetz, Dan Rozelle
Rancho BioSciences, PO Box 7208, Rancho Santa Fe, CA 92067

Background

Drug discovery remains one of the most formidable challenges in the preclinical pipeline, particularly the identification of novel therapeutic targets. Perturb-seq enables unprecedented insights into cellular function, but working with large, heterogeneous Perturb-seq datasets presents challenges in consistency and scale. Without harmonized data models and scalable access, interactive exploration and cross-study analysis are limited.

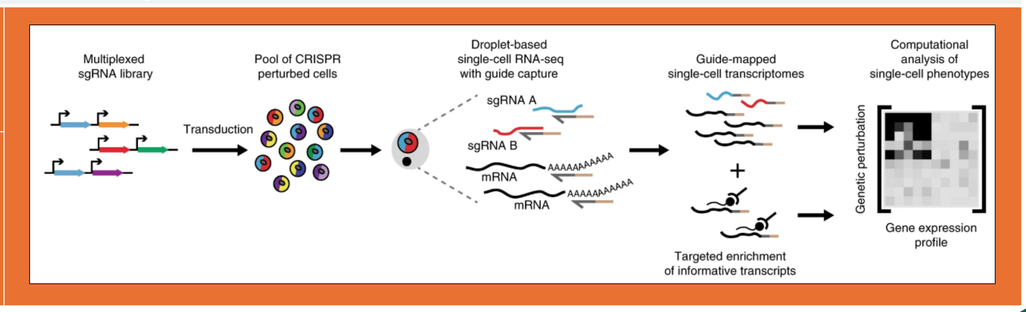
Rancho BioSciences combines high-quality, expert-driven data curation with scalable data infrastructure to unlock the full value of Perturb-seq data. Harmonized metadata ensures consistency across studies, while scalable storage and query systems enable efficient access to large, complex datasets. Interactive tools built on this foundation support rapid exploration, iteration, and interpretation. This integrated approach empowers clients to generate insights more quickly and confidently.

What is Perturb-seq Data?

Perturb-seq and related assays (e.g., CROP-seq, CRISP-seq) refer to the combination of genetic- (CRISPR, siRNA, shRNA) or chemical-based (small molecule) perturbations with gene expression sequencing at the individual cell level.

Why Use Perturb-seq Data?

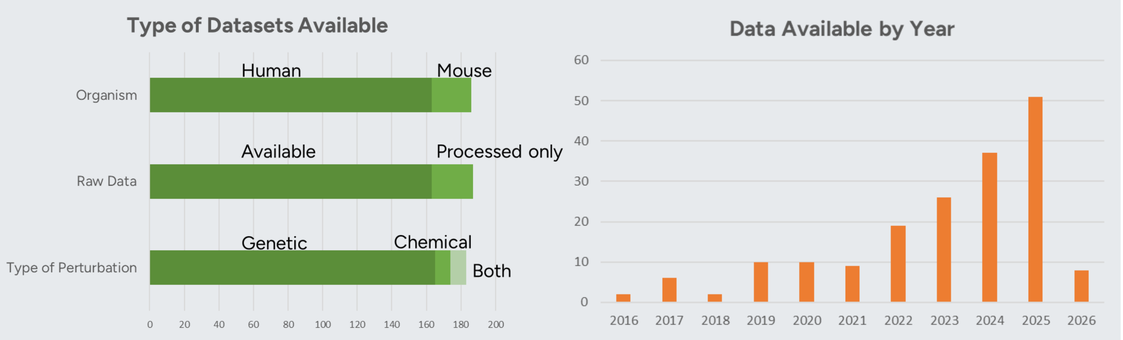
- Causal relationships: Perturb-Seq provides direct perturbation data (knockdowns, knockouts) that reveal cause-and-effect relationships between genes, rather than just correlational patterns in observational data
- Functional annotations: Each perturbation creates a controlled experiment, helping models learn gene function and regulatory networks
- Counterfactual learning: Models can learn "what happens if" scenarios, which is valuable for prediction and intervention design
- Richer training signal: The combination of perturbation identity + cellular response provides more structured supervision than unlabeled atlas data alone



Discovery through Quality

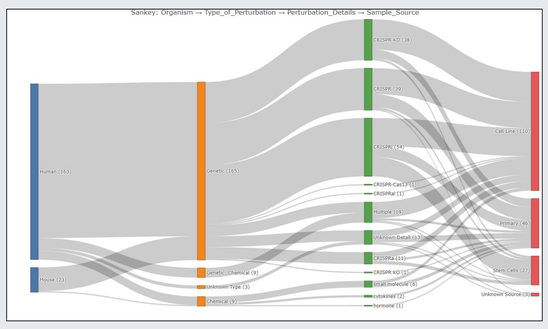
How much Perturb-seq Data is Available?

Perturb-seq is Publicly Available
Rancho has mined 6 different public repositories for available Perturb-seq datasets, including Gene Expression Omnibus, Sequence Read Archive, and FigShare.
As of Feb. 2026, there are over **180 publicly available studies**, including the Tahoe-100M dataset which includes > 100 million cells target with 1,100 small-molecules.



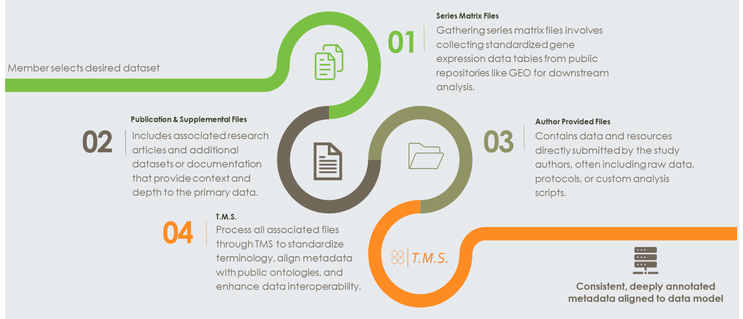
What of Type of Perturb-seq Data is Available?

There is a wide variety of Perturb-seq Data Available
Rancho has annotated study-level metadata for available datasets.



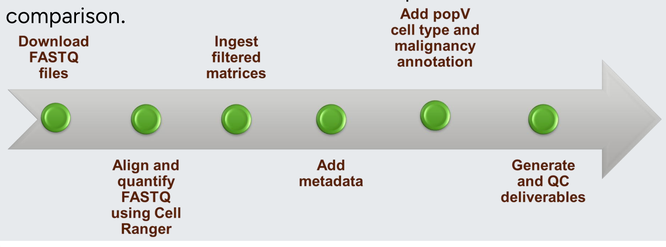
How do you get AI- and ML-ready Metadata?

Standardize Metadata
Rancho Data Integrity Specialists **harmonize** and **standardize** metadata from the repositories, publications, and other sources to an optimized data model and client-specified ontologies. **Rancho** uses our **Terminology Management Solution** to map values to ontologies in less time.



How do you get AI- and ML-ready Transcriptomic Data?

Bioinformatics Processing
Rancho Bioinformaticians download and process molecular data uniformly, better enabling cross-study comparison.



Deliverable	Format
Scanpy analysis object	h5ad
Seurat analysis object	RDS
README	txt
Metadata workbook	xlsx
QC directory	misc
manifest	json

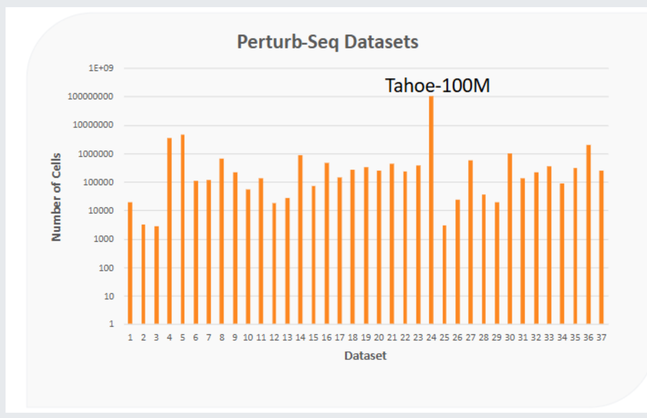
Scaling Up

Scalability Challenges from Perturb-seq

Perturb-seq Generates Large Datasets
Hundreds to thousands of cells per perturbation are required for statistical power, resulting in datasets ranging from hundreds of thousands to tens of millions of cells.

Large Data Strains Exploratory Analysis
Discoveries from Perturb-seq are made through exploration.

However, repeated filtering, querying, and visualization—central to exploratory analysis—are limited by standard in-memory approaches. **Rancho** has experience addressing these challenges, helping you derive insight faster and easier.

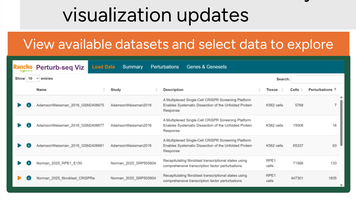


Perturb-seq Visualization App

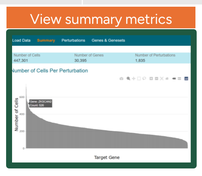
An R/Shiny application powered by Arrow, DuckDB, and Plotly enables scalable, interactive exploration of Perturb-seq data.

- Search and filter large DGE tables (50M+ rows)
- Explore cell and gene metadata
- Link table selections directly to visualization updates
- Dimensional reduction overlays for gene expression and metadata
- Interactive volcano plots and heatmaps
- Plotly-based hover, zoom, and styling controls

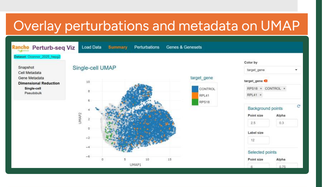
View available datasets and select data to explore



View summary metrics

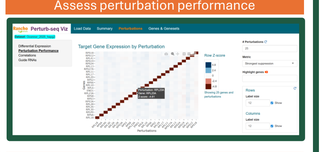


Overlay perturbations and metadata on UMAP

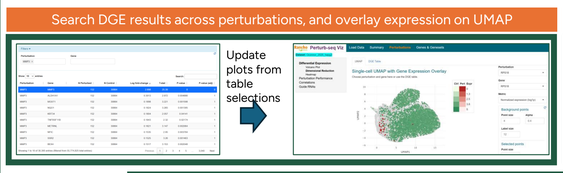


Visualize perturbation-driven expression via interactive volcano plots and heatmaps

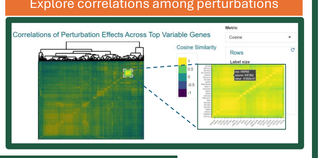
Assess perturbation performance



Search DGE results across perturbations, and overlay expression on UMAP



Explore correlations among perturbations



Perturb-seq Alliance

A pre-competitive, shared-cost community to identify, harmonize, and process publicly available Perturb-seq data

Perturb-seq data catalog
Generate updated list of available datasets with key metadata categories
Harmonized metadata
Consistent metadata terms between studies
Consistent Processing
Consistent gene model and transcriptomic pipeline
Visualization App
Interact with perturbation data

01 Harmonized data: unified model, aligned values
02 All data processed uniformly and with modern algorithms
03 Control over content AND portable formats
04 Keep your data forever

Rethinking Data Access for Scale

Shift from In-memory to On-demand Access
Large-scale datasets require moving beyond in-memory workflows to scalable, query-driven access patterns.

Leverage On-disk Data Systems
Multiple on-disk approaches enable scalable data access, each with tradeoffs in flexibility, performance, and complexity.

Simple and Scalable Data Architecture

Structured, Query-optimized Data
Rancho has transformed Perturb-seq datasets into structured, query-optimized formats to support efficient analysis and interactive exploration in a Shiny application.

Tools for Scalable Data Access
Rancho uses Arrow, DuckDB, and Polars are lightweight tools in R and Python that support lazy evaluation and query-driven workflows, avoiding full in-memory loading.

AnnData
- Processed using a standard Scanpy-based workflow
- Exported as collection of Parquet files (chunked as needed)

Parquet
- Columnar, compressed storage with explicit schema
- Single-file or partitioned directory structures
- Expression stored as tall tables, partitioned by gene for fast access

Analysis
- Querying and transformation using Arrow and DuckDB
- Data accessed on demand without full in-memory loading
- Enables efficient exploration and visualization in R/Shiny

Efficient Wrangling and Analysis

Familiar Workflows at Scale (Arrow)
Arrow enables dplyr-style querying with lazy execution on large, on-disk datasets.

Fast, SQL-based Querying (DuckDB)
Filtering, joining, and aggregation are performed directly on Parquet, efficiently handles very large tables (100M+ rows).

On-demand Data Access for Interactive Applications
Small data slices are retrieved in response to user input, enabling fast, scalable table rendering in Shiny.

Want to learn more? Visit us at Booth #405!