Data Alliances: A collaborative opportunity for FAIR and AI-ready data

Rancho bio • sciences

Hillary Mosso, Sondra Kopyscinski, Ben Ernest, Candace Ruff, Anne Cooley, Panos Agioutantis, Konstantin Bobkov, Marissa Hirst, Nicole Leyland, Viktoria Andreeva, Dan Rozelle

Rancho Biosciences, PO Box 7208, Rancho Santa Fe, CA 92067

Problem statement

Public biomedical data is abundant but rarely analysis-ready. Fragmented formats, inconsistent annotations, and accessibility barriers prevent researchers from leveraging this resource at scale. Our Alliances provide a collaborative framework to harmonize data by domain and disease type, delivering FAIR, AI-ready datasets that accelerate research.

Disease Agnostic
- Access to high-quality data across diseases and cell types
- Customizable to your needs

Flexible and Compatible
- Interoperable, FAIR
- Use any tools, workflows and pipelines with the data
- Platform-agnostic

Standardized Data
- Consistent standards across datasets
- Ready for immediate analysis
- Processed, harmonized, AI/ML machine readable

Maximize ROI
- Minimal membership fees
- NO cost to maintain access
- Shared investment across members

01 Harmonized data: unified model, aligned values

02 All data processed uniformly and with modern algorithms

03 Control over content & portable formats

04 Keep your data forever

Single Cell data alliance

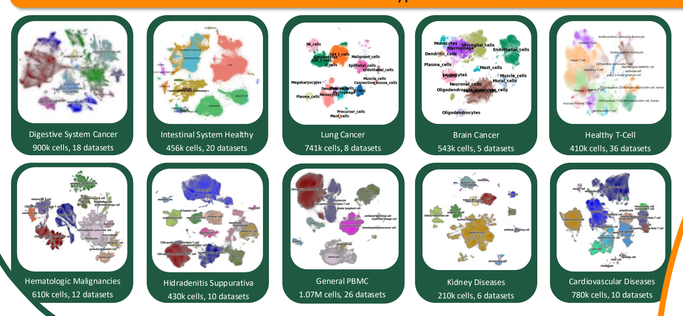
Description

Supporting a collaborative community of biomedical researchers with AI-ready **single-cell** data, governance, and inspiration
- Mining public data: 9,000+ public single cell datasets available, member request process
- Rich, harmonized metadata: comprehensive study, dataset, donor, sample, cell-level annotations, version-controlled ontologies
- Advanced processing pipeline: uniform mapping and quantification, standardized QC, multi-modal data support
- Flexible data delivery: h5ad, Seurat, TileDB, optimized for downstream analysis

Data

112 million Reprocessed, and annotated single-cells	37,851 Batch-corrected samples	17,854 Biological donors	1,154 Analysis-ready datasets for Seurat and Scanpy
100% Cell-types annotated	48% of datasets include author-provided annotations	48% Cells are from diseased samples	

14
Number of cell type atlases delivered



Oncology data alliance

Description

A collaborative initiative aimed at generating large-scale, disease-specific content across multiple datasets. The content will be integrated, re-analyzable, harmonized, and portable. The goal is to extract value from vast amounts of disparate public domain oncology data.
- Metadata is harmonized using an interoperable data model that is used among the alliances
- Standardized processing workflow customized for data type and flexible for various raw data types
- Delivered directly to client and platform independent

Data

Bulk RNA-seq

TCGA 33 cancer types	GTeX Healthy	GEO 8 cancer types

As **healthy** counterpart to TCGA

Next steps

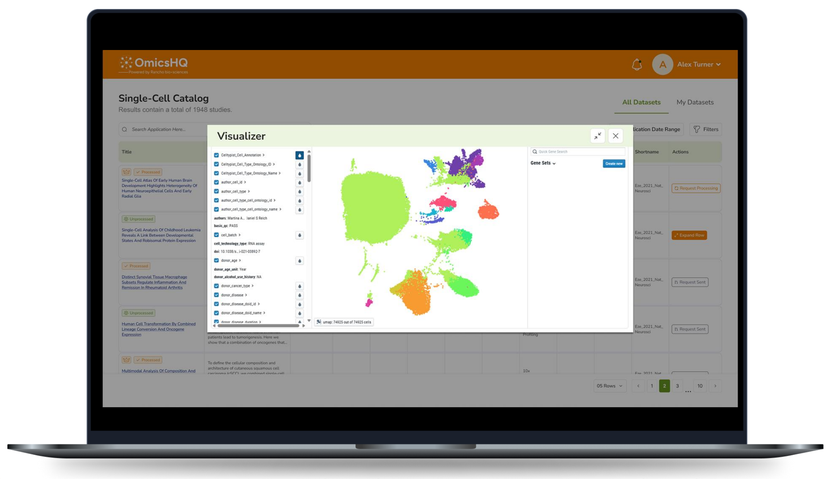
Flexibility to choose any **oncology** data of interest, including various cancer types and data types.

Transcriptomic · Gene variants · Phenotypical · Proteomics · Epigenetics

OmicsHQ

Description

Advance discovery through a comprehensive, high-quality, and standardized multi-omics data platform from public and proprietary sources, accessible to every scientist.



Catalog · Visualize · Request · Curation

Spatial data alliance

Description

To accelerate **spatial biology** research by creating standardized, interoperable, and AI-ready transcriptomics datasets, faster and more cost effectively than any single organization
- Aggregate and harmonize spatial transcriptomics datasets across platforms and tissue types
- Establish a coherent spatial data model based on Rancho's ontology and curation framework
- Enable cross-platform benchmarking and reproducibility
- Provide access to ready-to-use, quality-controlled datasets for analysis and AI applications

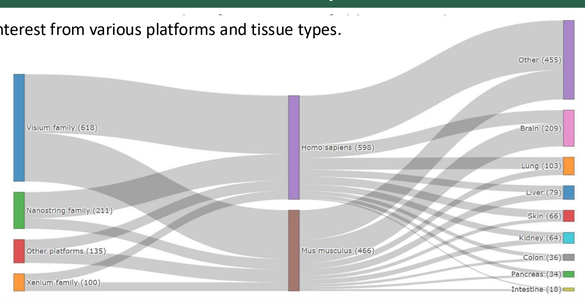
Data

33 Total Datasets (26 studies)	21 Human 12 Mouse	13 Diseased 10 both 8 Healthy	25 10x Visium v1/2** 6 GeoMx 1 CosMx 1 MerFish	17 Oncology 8 Healthy 4 Immunology 2 Liver Disease 2 NeuroDegen. 4 Brain · 6 Colon 3 Lung · 6 Liver 5 Kidney · 3 Skin

**cell deconvolution completed for all Visium v1/2 datasets

Next steps

Flexibility to choose ten datasets of interest from various platforms and tissue types.



Perturb-Seq data alliance

Description

Perturb-seq and related assays refer to the combination of genetic- (CRISPR, siRNA, shRNA) or chemical-based (small molecule) perturbations with gene expression sequencing e individual cell level
- 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 networksalone
- Counterfactual learning: Models can learn "what happens if" scenarios, which is valuable for prediction and intervention designatlas data alone
- Richer training signal: The combination of perturbation identity + cellular response provides more structured supervision than unlabeled atlas data alone

Next steps

As of Feb. 2026, there are over 180 publicly available studies to choose from, including:
- Tahoe-100M (> 100 million cells target with 1,100 small-molecules)