Data Alliances: A Collaborative Opportunity for FAIR and AI-ready Data

Rancho bio • sciences

Lavinia Woodward, Sondra Kopyscinski, Ben Ernest, Candace Ruff, Anne Cooley, Panos Agioutantis, Konstantin Bobkov, Marissa Hirst, Nicole Leyland, Viktoria Andreeva, John Garrison, Alena Fedarovich, Dan Rozelle

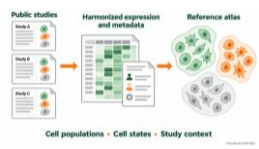
Rancho Biosciences, PO Box 7208, Rancho Santa Fe, CA 92067

Publicly available biomedical data are dispersed across repositories, publications and regulatory documents. Fragmented metadata and inconsistent processing can make them difficult to find, assess and use to answer scientific questions. Rancho's precompetitive Data Alliances enable members to share the preparation investment, creating shared, curated, harmonized and documented resources aligned with FAIR principles, for use and reuse across teams and programs.



From To Fragmented To Reusable Data For Research and AI

Single Cell data alliance



Which cell populations and states are associated with disease, and where is target RNA detected?

DATA RESOURCE

Harmonized public single-cell expression data, cell annotations and study context, with reusable reference atlases.

112 million reprocessed and annotated cells

1,154 analysis-ready datasets

14 cell-type atlases

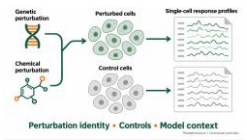
APPLICATIONS

Disease biology • Target RNA assessment • Reference annotation • AI model development and evaluation

CURRENT STATUS

Established resource; Phase 2 ongoing.

Perturb-Seq data alliance



How do cells respond to genetic or chemical perturbations?

DATA RESOURCE

Planned harmonized expression data linking perturbations to measured single-cell responses, with controls, experimental metadata and quality-control documentation. v

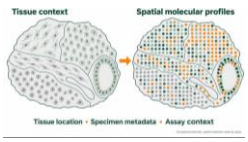
APPLICATIONS

Functional target assessment • Pathway analysis • Response heterogeneity • Predictive model development and evaluation

CURRENT STATUS

Early development; visualization app prototype; member input to guide further development.

Spatial data alliance



Where do disease-associated molecular signals occur within tissue?

DATA RESOURCE

Curated molecular profiles linked to tissue coordinates, specimen metadata and assay documentation.

33 delivered datasets

26 public studies

4 platform families: Visium, GeoMx, CosMx and MERFISH

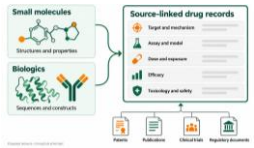
APPLICATIONS

Target RNA localization • Tumor microenvironment analysis • Spatial method benchmarking

CURRENT STATUS

Phase 1 delivered; seeking members for Phase 2.

Drug data alliance



Research question
How do molecular design and experimental context relate to drug performance?

DATA RESOURCE

Two proposed arms for small molecules and biologics, linking structures, sequences and constructs to pharmacology, efficacy, toxicology and safety data from patents, publications, clinical trials and regulatory documents.

APPLICATIONS

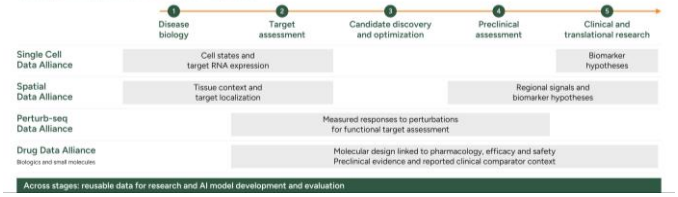
Drug comparison • Toxicology and safety assessment • Preclinical-to-clinical assessment • AI model development and evaluation

CURRENT STATUS

Scope in development for the small-molecule and biologics arms.

Where Alliance Data Can Support Drug Development

Selected potential applications across an iterative R&D process



Value Delivered by Phase 1 of the Single Cell Alliance

DELIVERED RESOURCE

99.4M cells in the shared resource

1,025 datasets

848 public studies

THE PREPARATION BEHIND THE RESOURCE

~90% of datasets were not fully author-annotated

Incomplete author annotation at source

7,389 → 901

Author cell-type labels mapped to common concepts

~16,960 SPECIALIST HOURS TO RECREATE THE RESOURCE

Estimate: 848 studies × ~20 hours of preparation per study

Benefits of Alliance Membership

Share the preparation investment
Spread data curation, harmonization and processing costs across participating organizations.

Access a broader resource
Benefit from studies selected for the wider alliance, including datasets prioritized by other members.

Reduce duplicated preparation
Use standardized processing, harmonized metadata and quality-control outputs to move more quickly to analysis.

Assess suitability for research and AI
Use documented biological context, annotations and provenance to evaluate datasets for your scientific question.

Shape future priorities
Help direct dataset selection and the resource's continued development.

Reuse across teams and programs
Apply the same documented resources to different research questions without rebuilding the underlying datasets.

Exchange knowledge and access expertise
Share methods and applications with peers, supported by Rancho's scientific, bioinformatics and data-engineering specialists.

A lasting data asset: Single Cell Phase 1. Members retain delivered data, metadata and documentation for continued internal use after membership ends.

99.4M cells

1,025 datasets

848 studies