Innovative Solutions for OMICs Metadata Harmonization from Multiple Sources

June 23rd, 2021

Peter Woollard, PhD, Computational Biologist / Manager, GSK
Bing Zhou, PhD, Senior Scientist, Rancho Biosciences
Andy Hope, PhD, Business Development, Rancho Biosciences

WEBINAR IS BEING RECORDED.
ALL PARTICIPANTS WILL BE ON MUTE.

Speakers

Bing Zhou, Ph.D.
Senior Scientist, Curation & QC Manager
bing.zhou@RanchoBiosciences.com

Peter Woollard
Business Metadata Manager
peter.m.woollard@gsk.com

Andy Hope, Ph.D.
Business Development
andy.hope@ranchobiosciences.com

Agenda

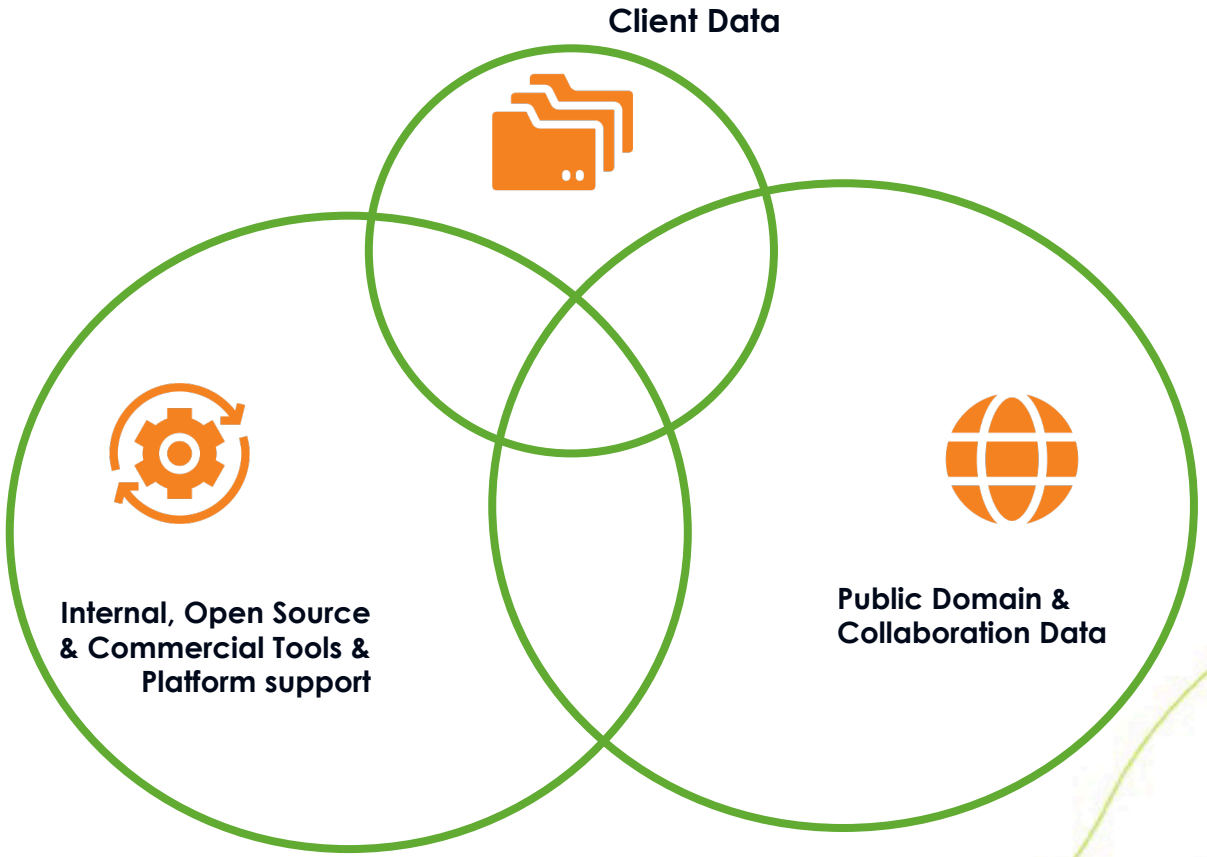
- Background and Introduction
  - Study Background
  - Objectives and Related Data Collection
- Metadata Harmonization: Innovative Solutions
  - Overview of Data Harmonization Workflow
  - Project Deliverables and Reusable Resources
- Significance and Application
- Pre-curated, harmonized datasets now available
- Discussion
- Questions & Answers

Why? Judicious Use of Public Data Can Become a Competitive Advantage

Availability of ever-increasing amounts public datasets can **serve to augment** the private experiments of pharma companies **many fold**.

Difficult and time consuming for pharma companies to exploit this valuable source of research data.

Effort spent by Scientists:
80% wrangling/curating/integrating versus **20% on analysis**



Rancho biosciences

Background and Introduction

- Study Background
- Objectives and Related Data Collection

Public Data Sources, Challenges

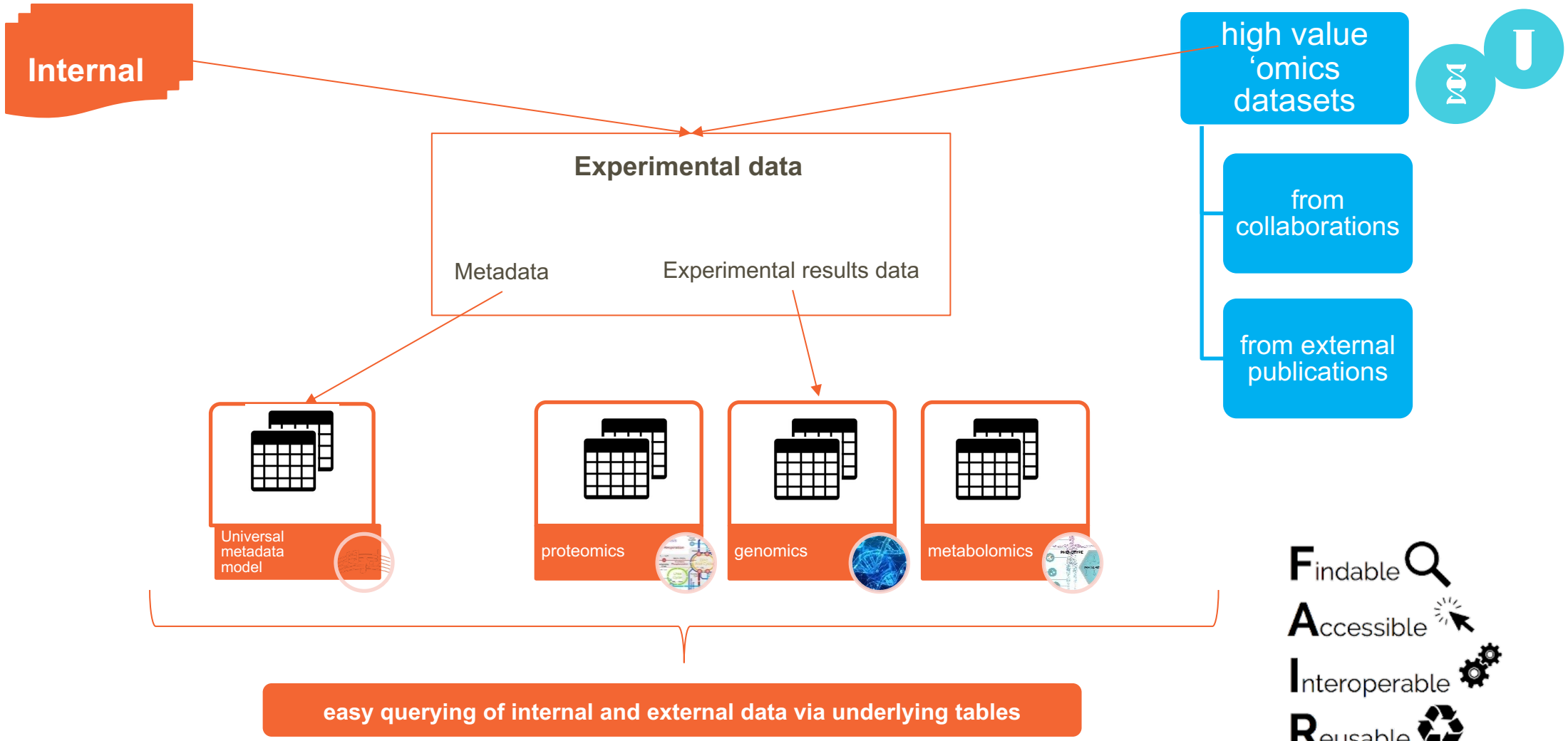
- What types of metadata to collect?
- What reference source to use?
- Why repeatedly enter the same metadata into non-aligned tools?

Metadata is inconsistent

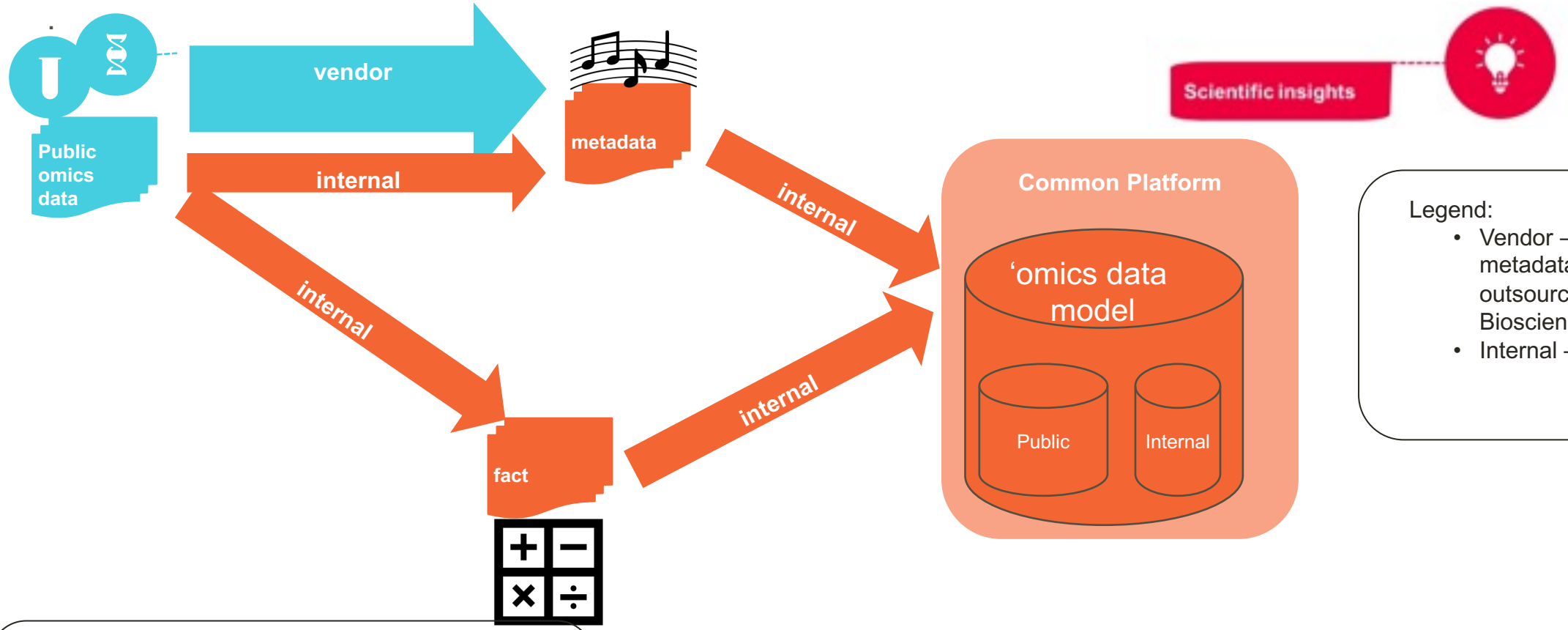
- Metadata is sparse
- Wasting time finding/preparing data
- Want to use specific ontologies
- Differing accuracy/recall thresholds

Enabling Better and Faster Scientific Decisions

– By creating an integrated, interoperable ‘omics’ data environment



Harmonising Metadata from External Sources

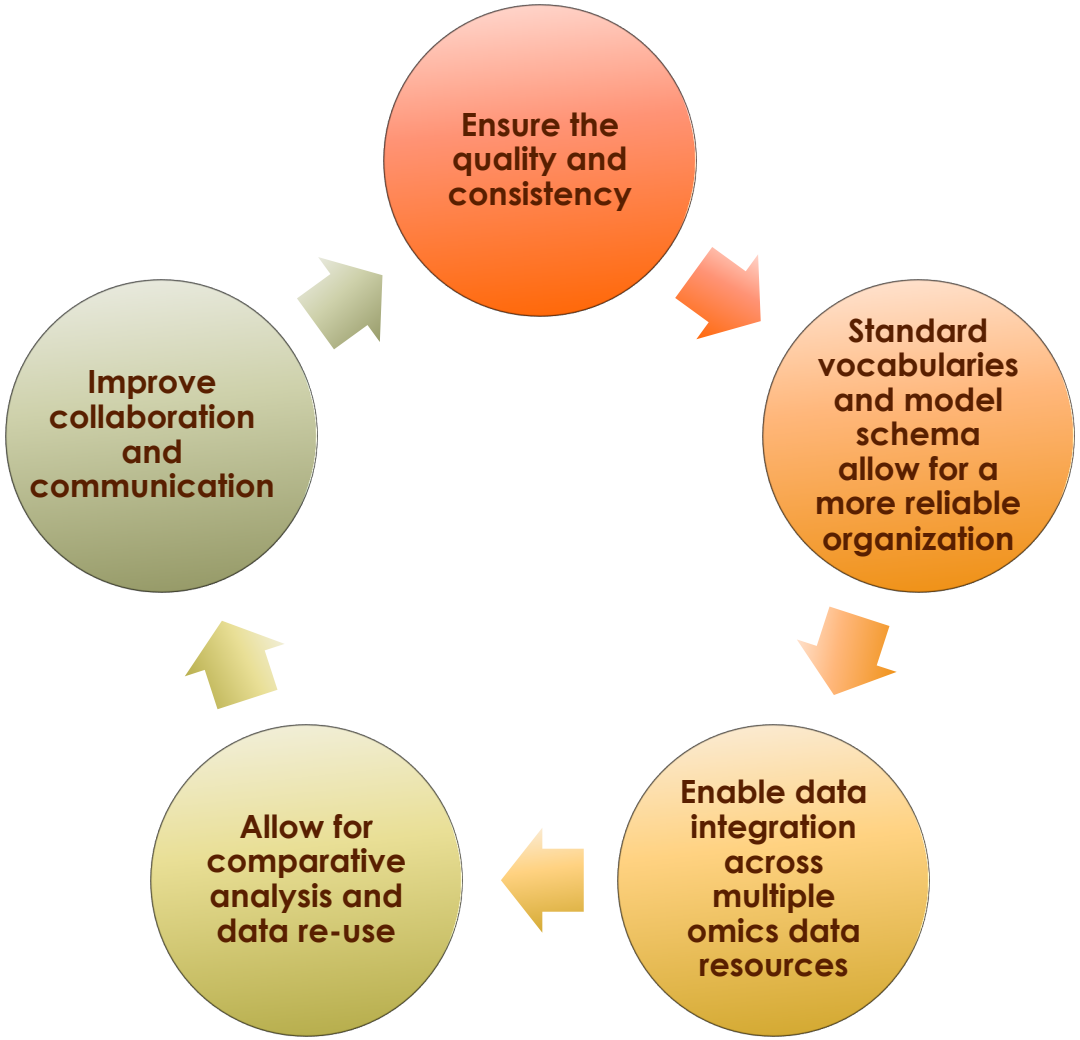


Legend:
- Vendor – much of the metadata curation is outsourced to Rancho Biosciences
- Internal – GSK FTE focused

Standards
- Universal data model
  - Core metadata entities to collect
  - Entity definitions
- Preferred:
  - ontologies
  - entity naming (=column naming)

Main metadata Curation:
- Core metadata entities identified in source
- Values mapped to preferred ontology terms
- Additional useful entities and values collected

Importance of Metadata Harmonization



FAIR principles

Objectives of OMICs Metadata Harmonization Initiative

- Facilitates data harmonization for pooling data from disparate public sources, that cover a range of omics data types to be integrated with internally generated datasets at GSK.
- To maintain a standard, extensible multi-domain data model (version 1 created by Rancho early in 2019).
- To reduce the time and effort investigators are spending harmonizing the data and to support use and re-use of the data assets more easily.

Landscape of Data Collection

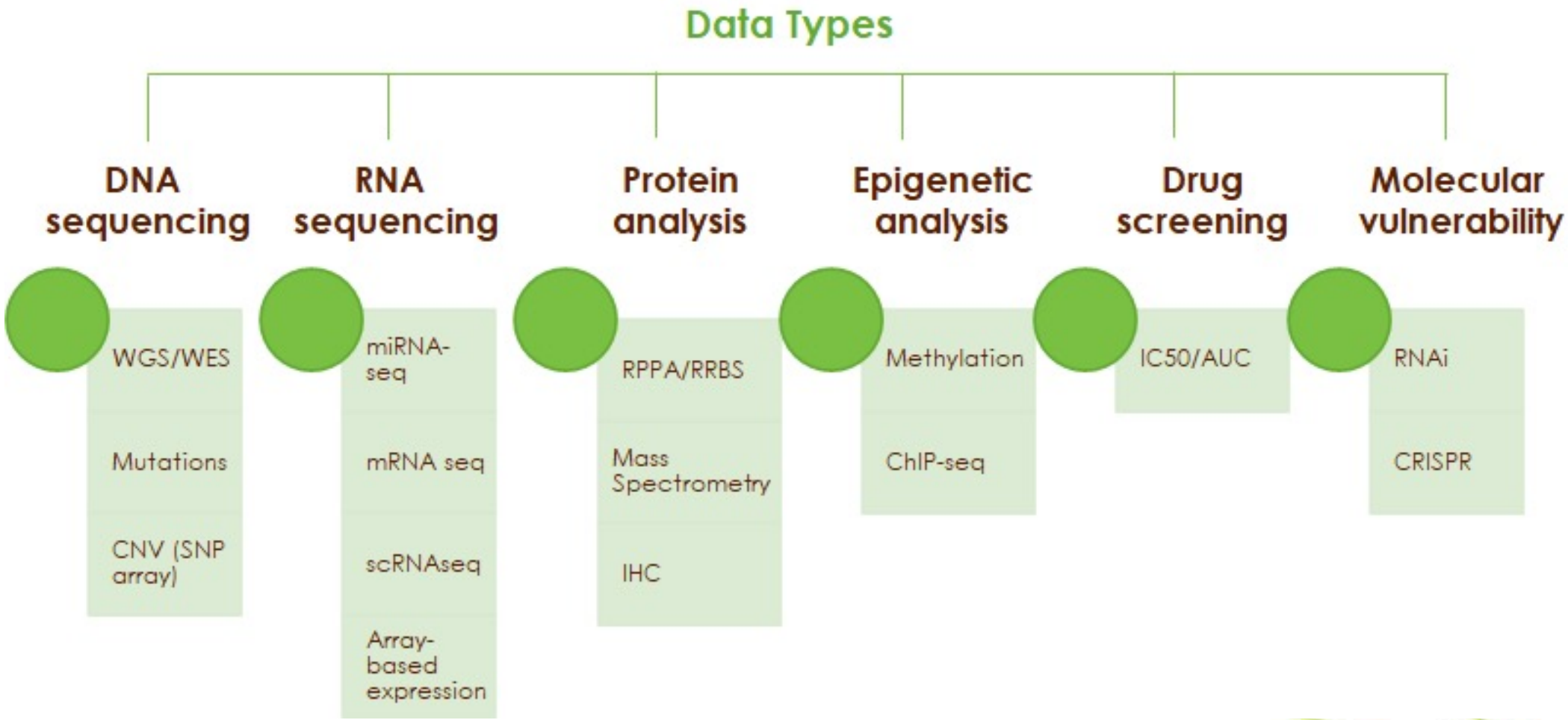
Considerations for metadata complexity assessment

- Are the raw metadata files already specified and easy to be downloaded?
- How many raw metadata files are involved? Is data integration needed?
- How many data types are involved?
- How many metadata variables need to be aligned to the data model?
- Are the raw data in a standard or well-organized format? What does the data structure look like?
- Is a metadata pre-cleaning step required prior to model alignment?

Number	Name	Complexity
1	Blueprint	Med
2	CCLE	Low
3	Depmap	Low
4	ENCODE	High
5	GEO (324 I/O studies)*	High
6	GTEx	Med
7	L1000/LINCS	High
8	Roadmap	Low
9	TCGA	Med
10	Achilles	Med
11	Cancerrxgene	Low
12	CGCI	High
13	HCA	High
14	HPA	High
15	PRISM	Low
16	DICE	Med
17	Cytoimmunogenomics	Med
18	SRA(2)/Array Express (3)	Med

* Immunology/Oncology

Summary of Data Types



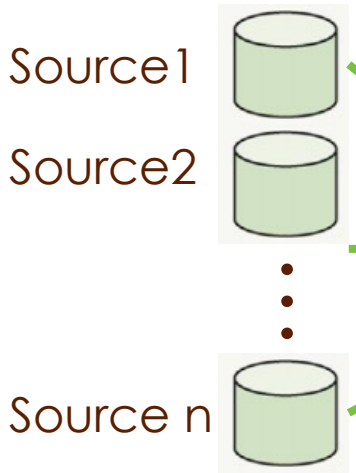
Rancho biosciences

Metadata Harmonization: Innovative Solutions

- Overview of Data Harmonization Workflow
- Project Deliverables and Reusable Resources

An Overview of Data Harmonization Process

Aggregation across multiple data sources



Uniform metadata harmonization

1. Source data nominating and assessment
2. Alignment with respect to a common data model
3. Clean and quality control accepted data with domain-relevant standard terms
4. Delivery harmonized metadata in a uniform format

Reproducible data processing

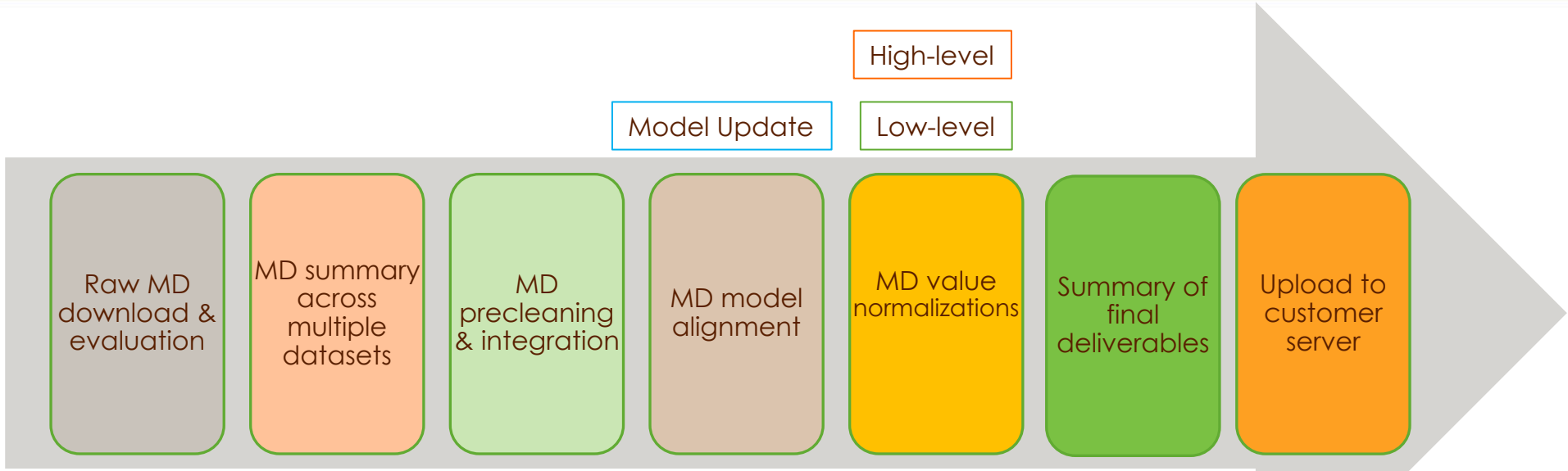
Process with common bioinformatic pipelines

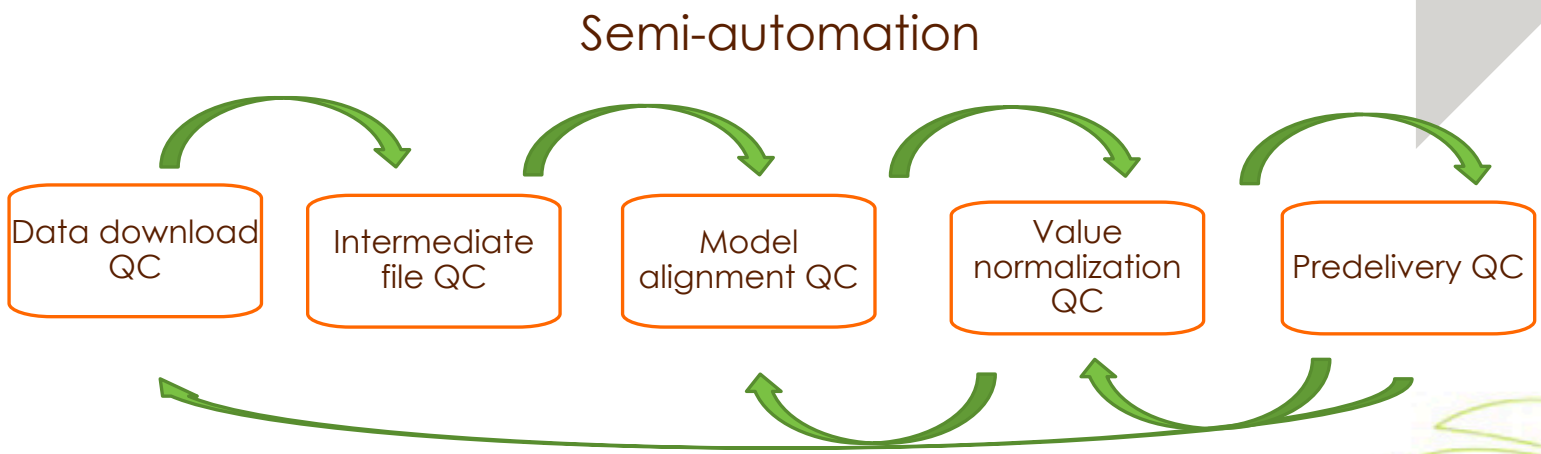
Post-process and share

1. Quality control, i.e., filtering of poor-quality data, removing outliers
2. Integrative visualization and analysis platform

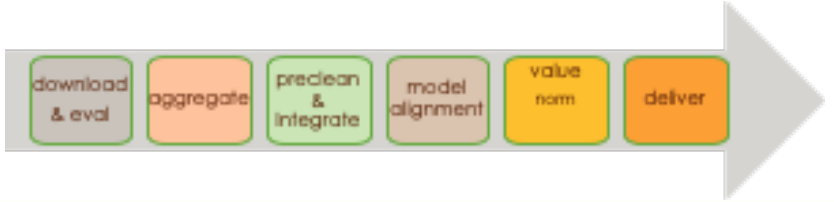
Optimization Feedback

Metadata (MD) Harmonization Workflow





Metadata Summary Examples



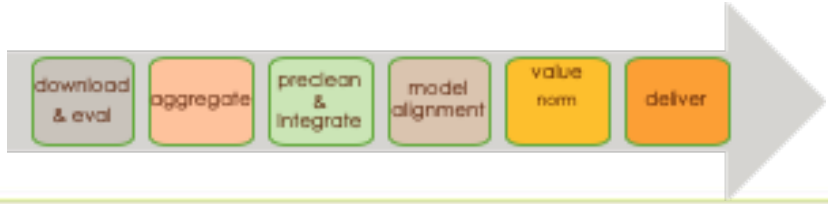
Attribute_per_row
(R script Summary)
1 file

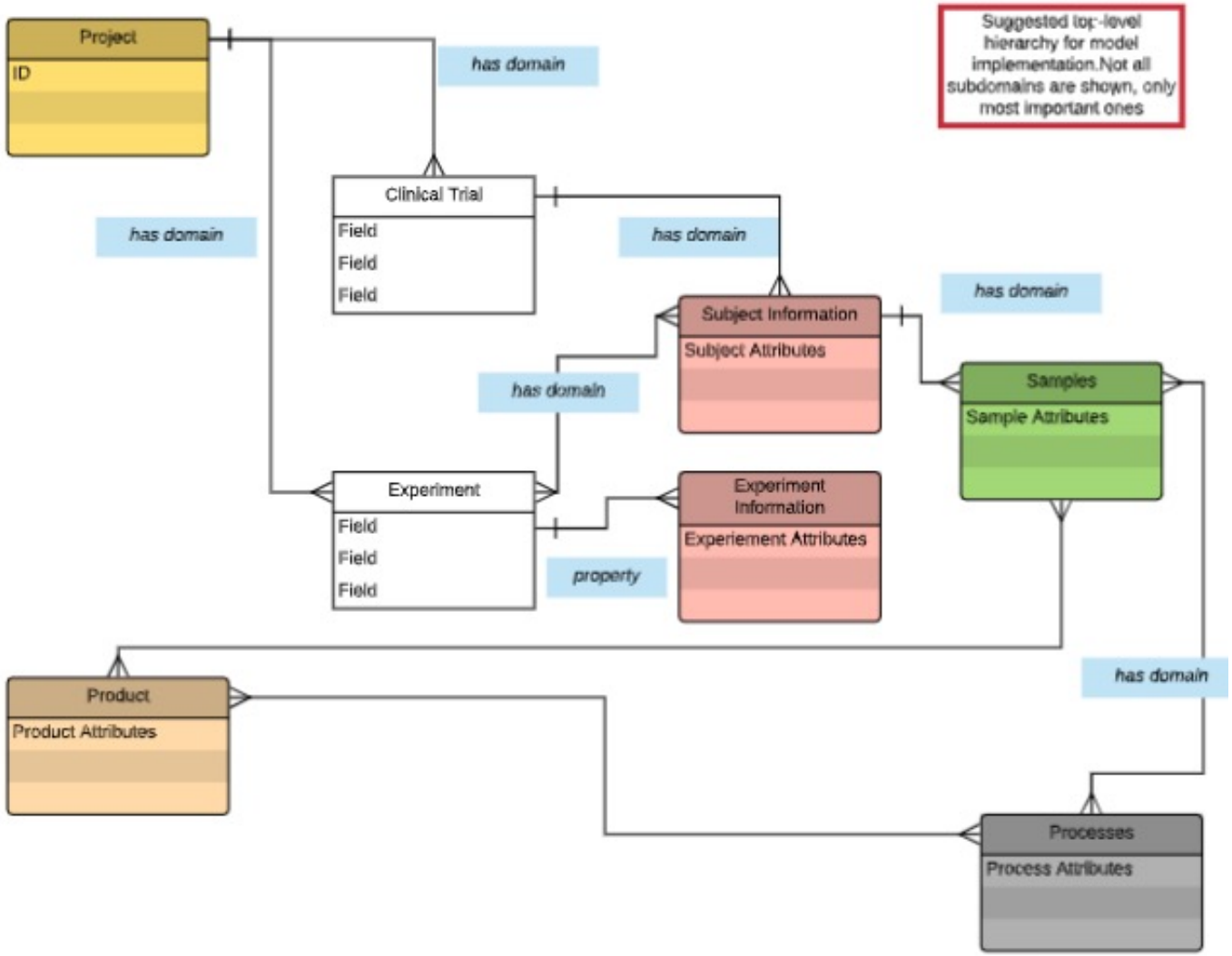
	A	B	C	D	E	F	G
	UniqueColumnName (CN)	CN Project Source	CN File Source	CN Count	Unique Values (UV)	UV Type	UV Counts
	Activation and Stimulation	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	lipopolysaccharide | dimethyl	logical;logical;character	16
	Address	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	415 Main St, Cambridge, MA 0	character;logical;logical;cha	84
	Age	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	13 | 63 | 58 | 43 | 47 | 54 | 62	logical;double;logical;doub	8152
	Age at Diagnosis	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	12 | 7 | 11 | 5 | 6 | 8 | 9 | 4 | 1	logical;logical;double;integ	90
	Age units	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	years | year | weeks | days	logical;character;character;	4
	Alcohol Use History	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	Drinker;1 unit/day;50 years | N	logical;character;logical;cha	14
	Biological replication	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	1 | 2 | 1;2 | 3 | 4 | 1;3 | 2;3 | 1	logical;character;logical;do	42
	Biological Substance Nam	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	none | granulocyte-macropha	logical;character;logical;cha	392
0	Cell Type	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	epithelial cell | lymphoblast | l	logical;character;logical;cha	294
1	Chemical Substance Nam	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	vorinostat | resiquimod | MC2	logical;character;logical;cha	24529
2	Country	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	United States of America | Uni	logical;character	10
3	Department affiliation	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	National Cancer Institute, Nati	logical;character;logical;cha	36
4	Disease	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	renal cell carcinoma | striated	character;logical;logical;cha	539
5	Disease Stage	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	primary neoplasm | metastasi	logical;character;character;	35
6	DNA type	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	genomic DNA | chromatin | DI	logical;character;character;	4
7	Dose	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	100 | 0 | 500;5;50 | 2 | 10 | 0.3	logical;character;integer;ch	5416
8	Dose units	Achilles;Blueprint;Cancerrxgene;CCLE;CGCI;Depma	Achilles_2.0_Me	15	ng/ml | ug/ml | nM | % | mM	logical;character;character;	48

Value_per_row
(Python script Summary)
15+1 files

Field Name	Unique Value	Project Name	Source File
Disease	Acute Biphenotypic Leukemia	Depmap	Depmap_Metadata.txt
Disease	Acute Eosinophilic Leukemia	Depmap	Depmap_Metadata.txt
Disease	Acute Myeloid Leukemia with Maturation	Depmap	Depmap_Metadata.txt
Disease	Acute Myelomonocytic Leukemia	Depmap	Depmap_Metadata.txt
Disease	Adult Acute Myeloid Leukemia	Cancerrxgene	Cancerrxgene_Metadata.txt
Disease	Adult Acute Myeloid Leukemia with Maturation	Achilles	Achilles_v2.4.3_Metadata.txt
Disease	Allergic Rhinitis	GEO	GEO_GSE75011_Metadata.txt
Disease	Ampulla of Vater Adenocarcinoma	Depmap	Depmap_Metadata.txt
Disease	Anaplastic Oligodendroglioma	Depmap	Depmap_Metadata.txt
Disease	Aneurysm	GEO	GEO_GSE106520_Metadata.txt
Disease	Autoimmune Disease	GEO	GEO_GSE119918_Metadata.txt
Disease	B Acute Lymphoblastic Leukemia	Depmap	Depmap_Metadata.txt
Disease	B-cell childhood acute lymphoblastic leukemia	Cancerrxgene	Cancerrxgene_Metadata.txt
Disease	B-cell lymphoma	Depmap; CCLE; Cancerrxgene	Depmap_Metadata.txt; CCLE_Metadata.txt; Cancerrxge
Disease	B-cell lymphoma;non-Hodgkin lymphoma	LINCS	LINCS_GSE92742_Metadata.txt
Disease	B-cell non-hodgkin lymphoma	CCLE	CCLE_Metadata.txt
Disease	Benign Prostatic Hyperplasia	Depmap; Cancerrxgene	Depmap_Metadata.txt; Cancerrxgene_Metadata.txt

Data Model Overview

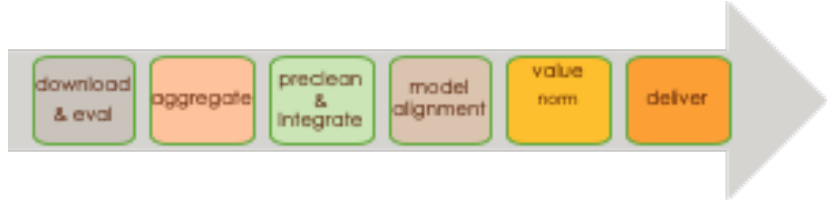




Combination of complex structures across multiple domains:

- Project
- Sample
- Process
- Product

Model Alignment and Precleaning



Normalized metadata **aligned to data model**

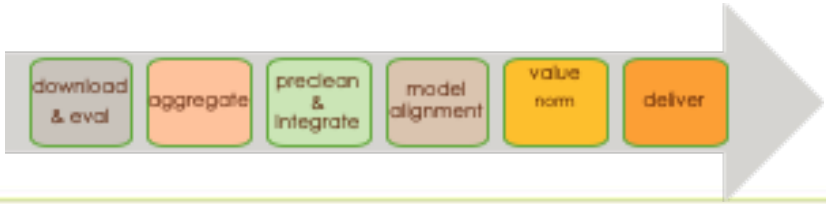
Annotations from data source

Project Name	UniqueColumnName (CN)	Model_Field Name (MFN)	Model_Code	KeyMetadata (y/n)	Unique Values (UV)	Source Description
GTEx	SMGEBTCHT	Technology type	1.1.1.8	y	Standard Exome Sequencing v	Type of genotype o
GTEx	SMGEBTCH	Experiment ID	1.1.4.1.3	y	LCSET-4574;GTEx_OM25_Dec_	Genotype or Expres
GTEx	SEX	Sex	1.1.7.3.1	y	Female;Male	The Donor's Identif
GTEx	AGE	Age	1.1.7.3.4	y	20 to 29;30 to 39;40 to 49;50 t	Elapsed time since
GTEx	SUBJID	Subject ID	1.1.7.7	y	GTEX-1117F;GTEX-111CU;GTE	Subject ID, GTEx Pu
GTEx	SAMPID	Sample ID	2.4.3.1.1	y	GTEX-1117F-0003-SM-58Q7G;	Sample ID, GTEx Pu
GTEx	SMUBRID	Name of cell type	2.5.1.1.14	y	fibroblast derived cell line;K562;lymphoblast	
GTEx	SMUBRID	Tissue name	2.5.1.1.5	y	0013756;0009834;0002190;00	Uberon ID, anatom
GTEx	/	Age units	1.1.7.3.5	y	years	Split from "Age"
GTEx	/	Disease	2.5.1.1.8	y	Healthy	Rancho added meta
GTEx	/	Sample Type	2.5	y	Solid Tissue Normal	Rancho added meta
GTEx	/	Project ID	TBD	y	GTEx	GTEx_Analysis_v8_
GTEx	/	Organism name	1.1.7.6.1	y	Homo sapiens	Rancho added meta
GTEx	/	Taxonomy ID	1.1.7.6.2	y	NCBITAX:9606	Rancho added meta
GTEx	SMGEBTCHD	Start date	1.1.4.1.14		01/15/2014;01/28/2014;09/2(	Date of genotype o

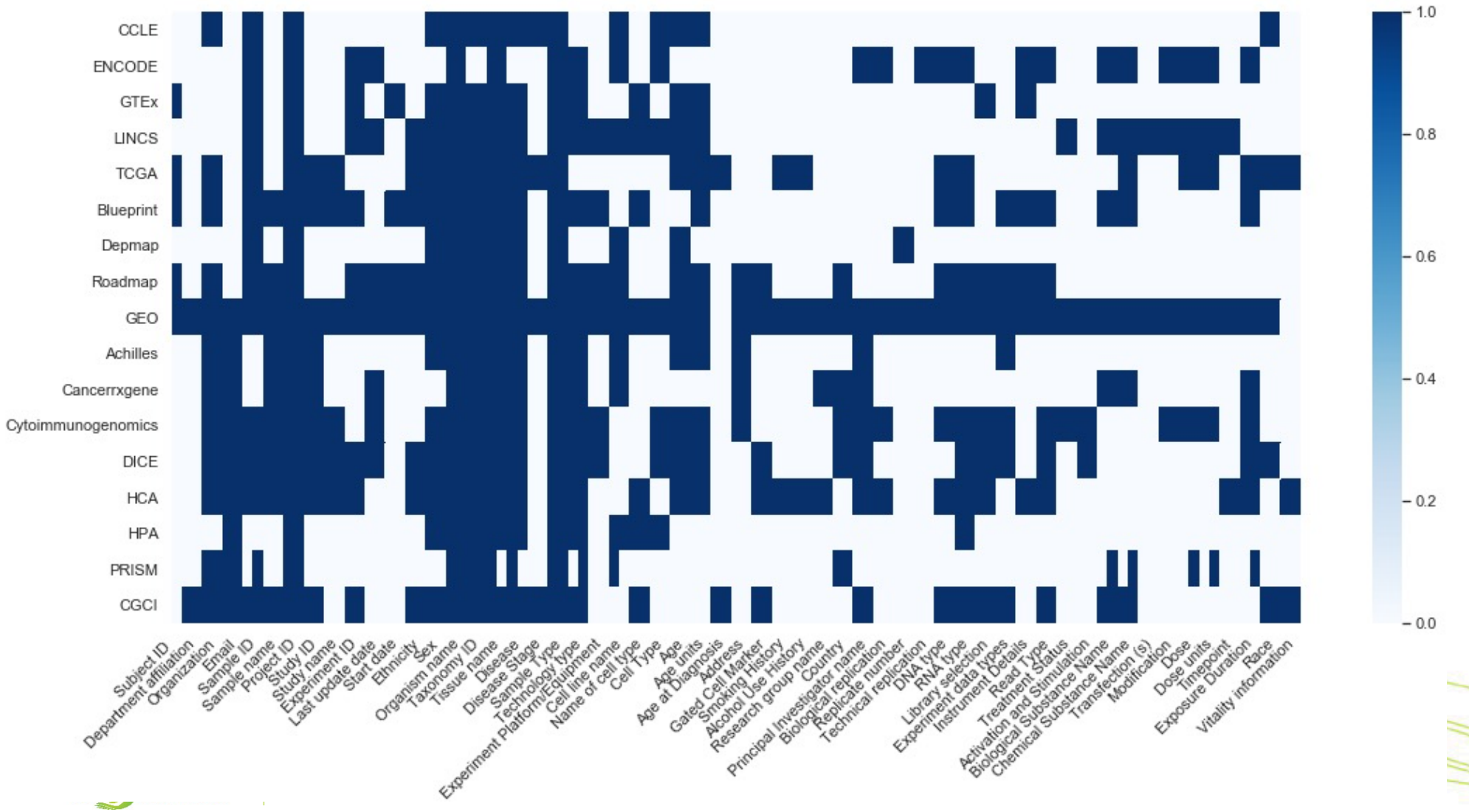
Additional metadata added by Rancho

TBD: **Updated attributes** to data model

Model Alignment Overview



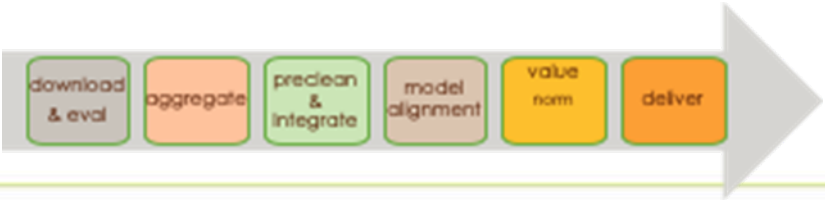
57 Main Metadata Attributes Across Multiple Datasets



1=Data Available

0=No Data

Value Harmonization Overview



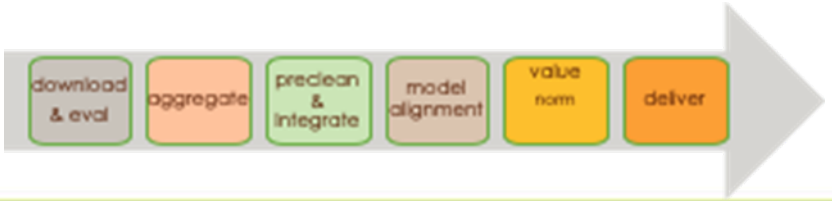
Step I: Metadata summary based on raw/pre-cleaned metadata

Step II: Map to Preferred Ontologies (Low-level)

Step III: Map to Preferred ontologies (High-level) for selected entities (Disease, Tissue, Cell Type)

- ✓ Combine **robust Fuzzy matching tool** with multiple rounds of manual review
- ✓ Domain-relevant standard terms, **linking terminology sources and codes**
- ✓ Map to standard terms (e.g. , DOID, UBERON, CL) at **different levels of granularity** to facilitate classification

Quality Assurance Procedures



Quality Assurance Check List

It is important to define and enforce quality assurance standards before, during, and after the collection and extraction of data. Rigorous QCs are required for each workflow across the data management life cycle to ensure the quality of deliverables from Rancho BioSciences. Here are some general QC considerations:

☐ QC Source input files

- Check whether received data (if source data is provided by customers) are consistent with your expectation and that it fits the proposed workflow
- Check file number and size are consistent with data source, which should be properly documented in data download and storage summary (data downloading source, data version, file types, downloading date, CRC32/MD5 or other checksums, etc.)

☐ Check column headers

- Spelling errors (turn on spell checker)
- Maintained original column headers or aligned to standard Data Model fields (if desired)
- Check for duplicated or empty column headers

☐ Check rows

- Sample number QC
- Empty or duplicated rows

☐ Check values

- Check whether unique values line up in proper column names (unclear annotations need to be investigated)
- Spelling errors
- Check for data inconsistency
  - Capitalization differences between values
  - Round-up of numerical values
  - Representation of "null" values (NA, Null, "/", dot, blank cells, etc.)
  - Check for inconsistency of unique value pairs for related metadata entities (e.g., subject level information: "Subject ID vs Age"; Controlled vocabularies vs corresponding code/link)
  - Check for inconsistency of "raw vs. cleaned" unique value pairs (e.g., Spot checking of 10~20% of the records curated)
  - Check for inconsistent sample or subject identifiers
- Check for categorical data ending with numbers or links ending with numbers (possibly caused by dragging down values)
- Check for values not mapped to standard terminology (if desired ontologies are defined)
- Check for missing values due to omission or inadequate data extraction
- Check for special extended ASCII characters (e.g., use "[^\x00-\x7F]+" regex to search for special characters in Notepad++)
- Check for extra spaces and hidden symbols (e.g., look for Leading/ Trailing spaces or White spaces in between characters by enabling symbol visualization with a text editor, such as Notepad++)
- Statistical analysis of quantitative data (script based or manual QC)
  - Look for impossible or anomalous values (e.g., human "Age" >200 years)
  - Look for outliers using normal probability plots, regression plots, or scatter plots
  - Check means, medians, percentiles, % missing data

☐ Check file type/extension

☐ Spot checking and change logs

- Keep logs for all change decisions, alterations, and benchmark testing
- Make sure spot checking (10~25% data) compliance with QC specifications

Project Deliverables & FTP Data Transfer

- **ReadMe**

Raw data download and storage summary, Summary of deliverables

- **Harmonized Metadata**
- **Metadata_Summary_By_Script**

  Summary of metadata by Python or R

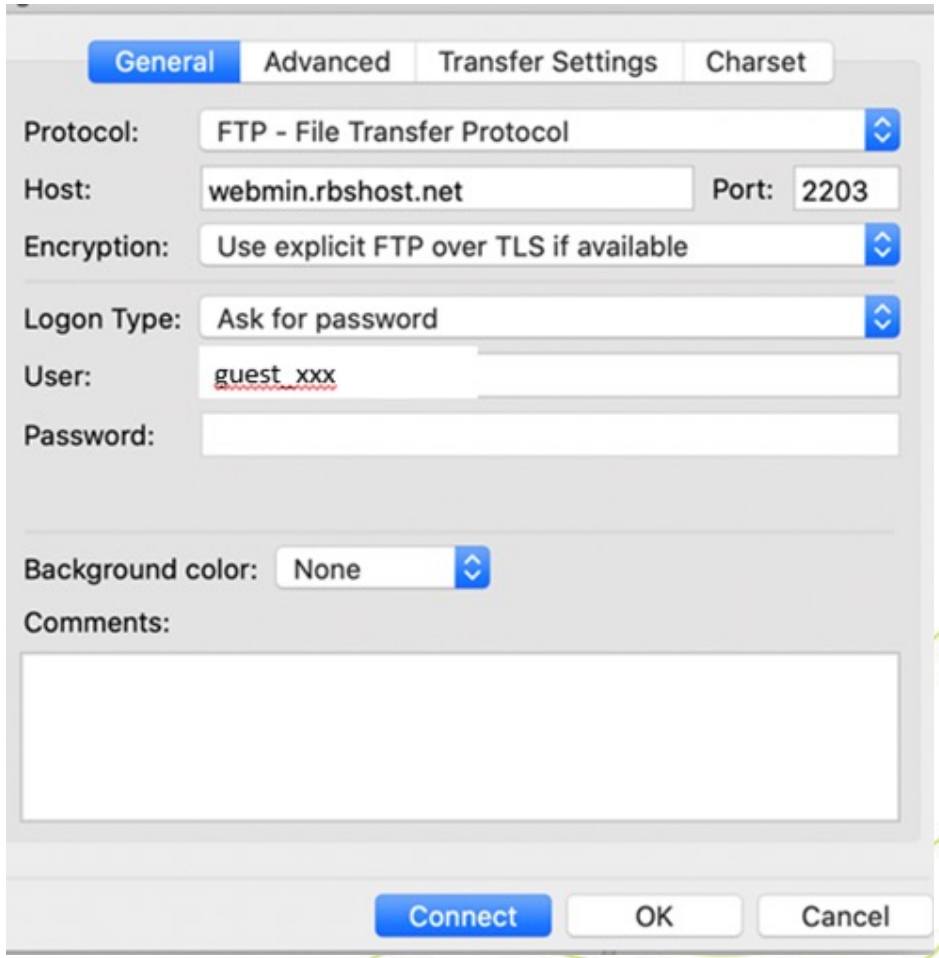
- **Metadata_Model_Alignment_Mapping**

  metadata unique columns/values, model fields/code, metadata annotations

- **Metadata_Value_Normalization_Revision_Mapping**

  metadata values mappings to preferred ontologies and additional high-level mappings

- **Metadata_Quality_Control_Report**
- **Updated Data Model**



Example of Metadata Summary Report

"ReadMe" Folder

Figure 4: Summary of subject specific "Module entity" examples

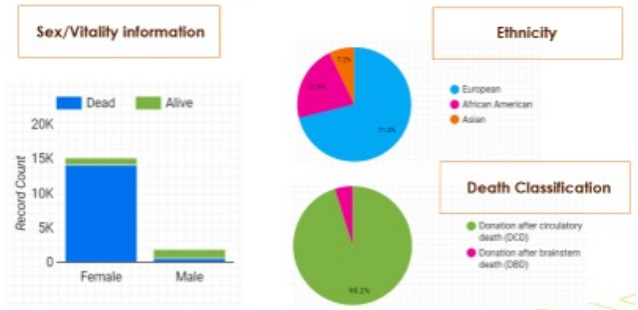


Figure 4: Summary of subject specific "Module entity" metadata fields, including "Sex" (Bar graph broken down by "Vitality information"), "Ethnicity" (top right) and "Death Classification" (bottom right) for donors included in all three HCA studies.

Figure 5: Summary of HCA "Process type" and "Core" metadata entity examples

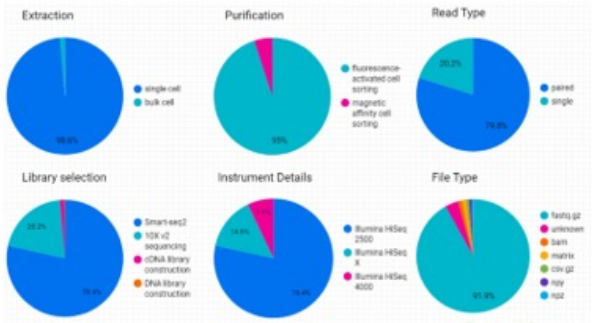


Figure 5: Pie chart summary of HCA "Process type" metadata entities, including "Extraction" method, "Purification" method, "Read Type" (these are the top left, middle, and right graphs) and "Library selection", "Instrument Details" (these are the bottom left and middle graphs), and "Core" type entity "File Type" (bottom right), for the three HCA studies.

Summary of Deliverables

This summary describes Batch 2 deliverable to GSK team on Data Harmonization Project that covers HCA dataset. Below is the brief description of each file/folder in the deliverable package.

1. "ReadMe" Folder

Files in this folder describe deliverables to GSK team on Data Harmonization Project for the HCA project.

1) Summary of Deliverables for GSK Batch 2 HCA project GSK_Metadata_Harmonization_Batch 2_Report-HumanCellAtlas.2020.04.10.docx

2) Summary of downloaded metadata files on Rancho internal storage server GSK_Batch2_Download_And_Storage_summary.xlsx

2. "Metadata" Folder

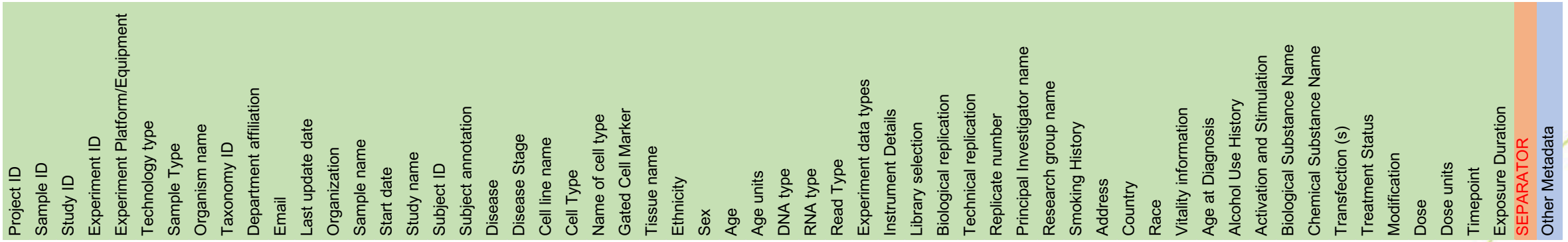
This folder contains normalized and aligned final metadata files for each project:

- HCA_Metadata.txt

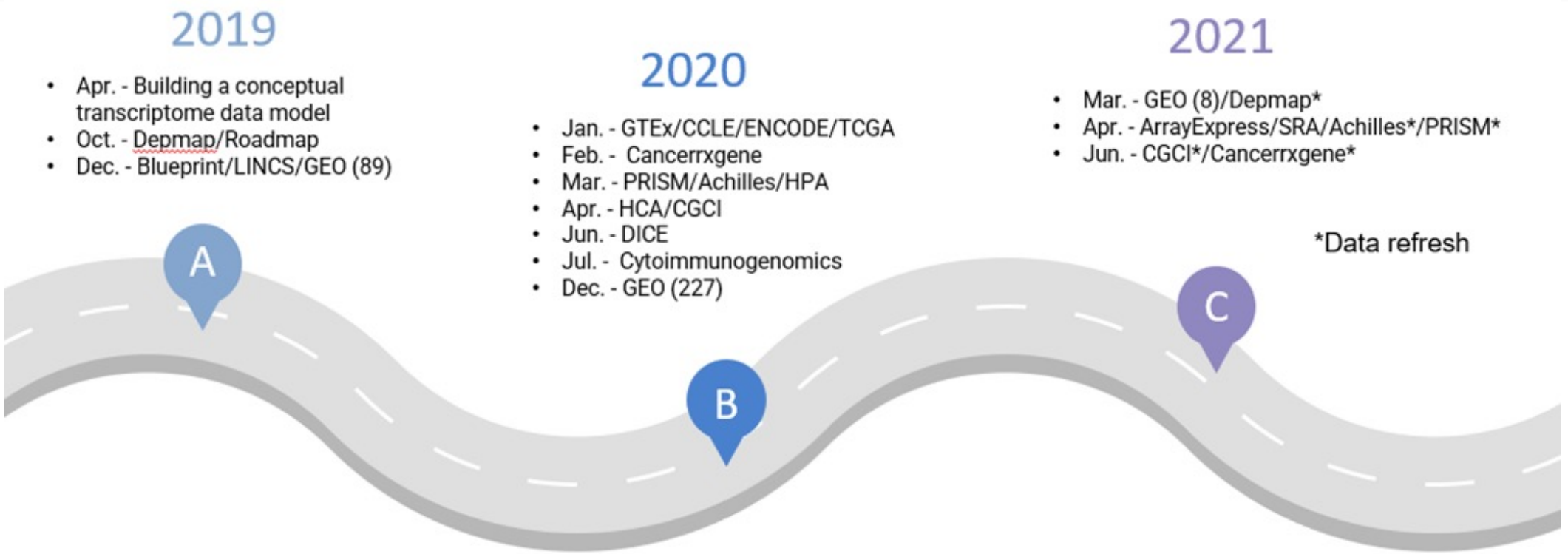
3. "Metadata_Model_Alignment_Mapping" Folder

Harmonized Metadata Layout

57 Main Metadata Entities Aligned to Data Model and Harmonized



Metadata Harmonization: Roadmap



- Developed a standard, extensible data model to accommodate the data & metadata from various public data sources
- Collection of 18 datasets harmonized using standard ontologies
- Data refresh when there is a new data release for on-going project

Rancho biosciences

Reusable Resources and Lessons Learned

Reusable Templates

Metadata Download and Storage Summary	Model Alignment Mapping file	Value Normalization Mapping file	Metadata Quality Control Template
• Project name • File directory • Download source, date, version • File type • File name • File description • File size	• Project Name • Unique Column Name (CN) • Model Field Name (MFN) • Model Code • Unique Values (UV) • Source Description	• Project ID • Raw values • Standard values • Ontology Link • Ontology Name • Curator Comments	• QC check list • QC Findings • QC Results

Valuable Ontology Mapping Resource

Semi-automatic curation has been an essential component of this effort to produce the **"Gold Standard"** in standard term mappings

- Standard format **"Cell line name"** mappings (**~4,000**)
- Low-level standard term mappings for **"Disease"** (**~1,800**), **"Tissue name" (~1,400), and "Cell Type" (~700)**
- **"Chemical Substance Name"** mappings (**~20,000** from LINCS, **>6,000** from other datasets), **"Biological Substance Name"** mappings **(~400**)

Lessons Learned

1. A successful data alignment and normalization requires **teamwork – active collaboration and frequent communication**
2. **Document** early and often, good **design and specifications at early phase** to avoid unnecessary iterations
3. Developing a conceptual **data model** early in the process is a key underpinning to everything else
4. Don’t underestimate the amount of data **quality checking** that is required at all levels
5. The data normalization/documentation/implementation is a **living process**

Rancho biosciences

Significance and Application

Metadata Summary Across All Datasets

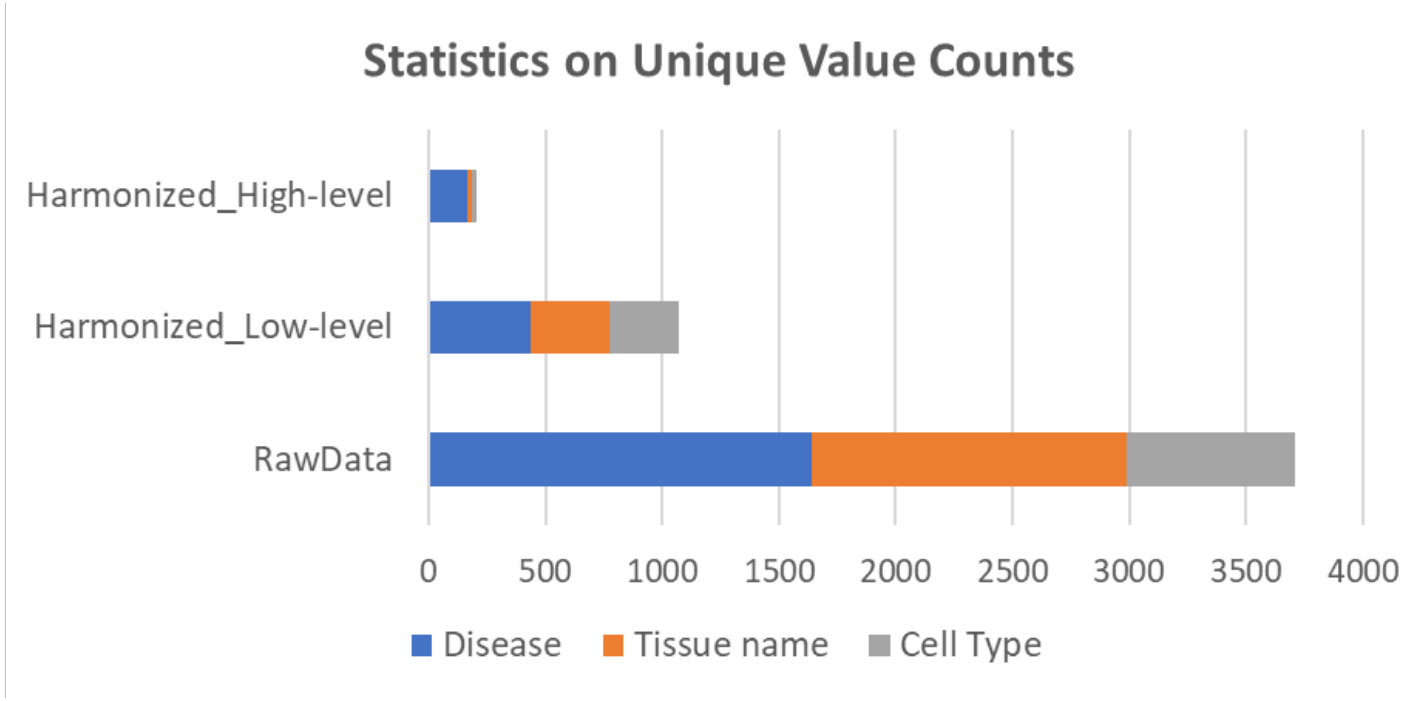
Project Name	Source File	Source File Count	Rows of Data
SRA/ArrayExpress	SRA_SRP078466_Metac	5	68
Cytoimmunogenomics	Cytoimmunogenomics_	1	805
Achilles	Achilles_2.0_Metadata.t	6	1,225
CCLE	CCLE_Metadata.txt	1	1,461
Dice	DICE_Metadata.txt	1	1,544
Depmap	Depmap_Metadata.txt;I	2	1,823
Roadmap	Roadmap_GSE16256_M	9	3,256
CGCI	CGCI_Metadata_BLGSP_	11	10,017
GEO	GEO_GSE100376_Meta	324	14,895
HCA	HCA_Metadata.txt	1	16,936
GTEx	GTEx_Metadata.txt	1	22,951
Blueprint	Blueprint_Metadata.txt	1	29,678
PRISM	PRISM_PrimaryScreen_(	4	54,829
ENCODE	ENCODE_Metadata.txt	1	398,316
TCGA	TCGA_biospecimen_Me	2	432,576
Cancerrxgene	Cancerrxgene_Metadata	2	446,146
LINCS	LINCS_GSE70138_Meta	3	1,668,313
HPA	HPA_Metadata_normal	21	95,575,200
		Total Count	98,680,039

Project Name	"Subject ID" Count
SRA/ArrayExpres	19
HCA	30
Cytoimmunogen	33
DICE	91
Roadmap	240
HPA	399
CGCI	616
Blueprint	636
GTEx	980
GEO	1790
TCGA	11,364
Total Subject	16198

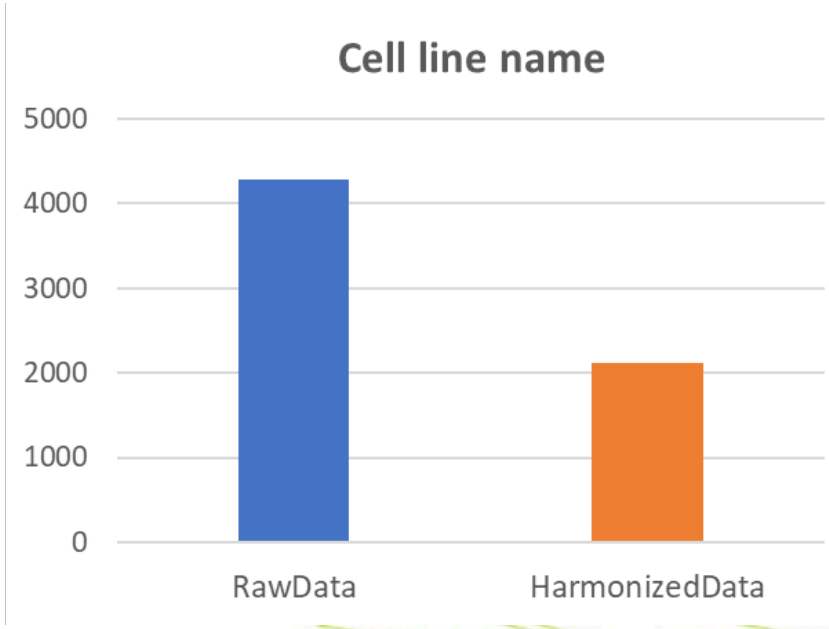
Total Rows of data: 98,680,039

Total Subject: 16,198

Metadata Summary Across All Datasets



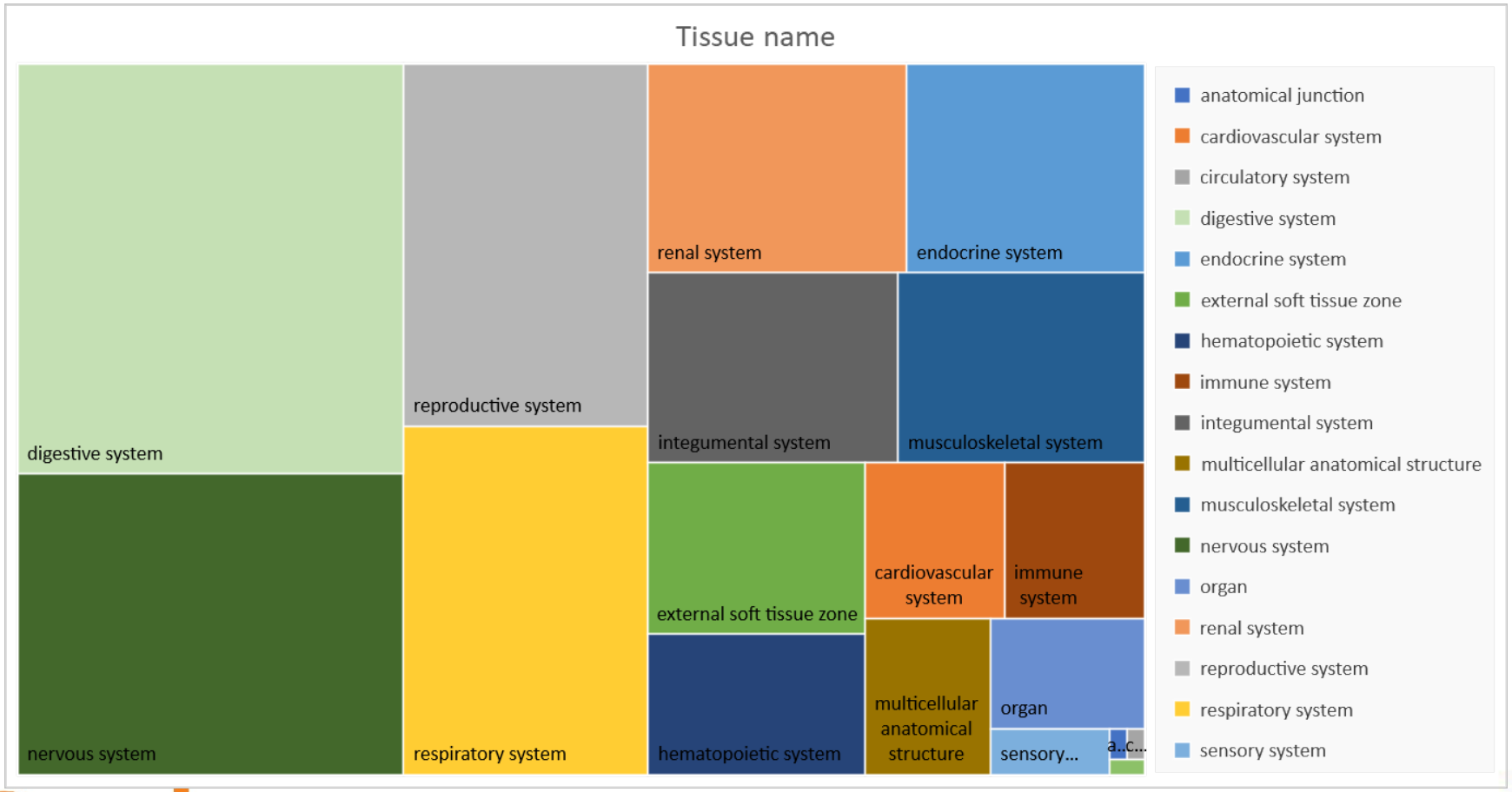
- ~60 metadata entities
- ~130 Technology types
- 7 Organisms
- 27 Countries

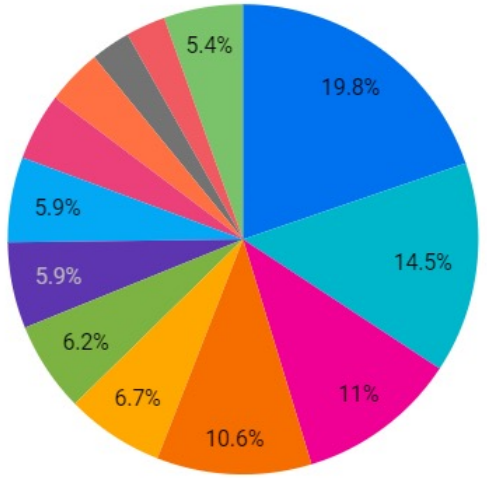


Metadata Summary Across All Datasets

Classification of "Tissue name" Across All Datasets

>300 Low-level Tissue terms are mapped to 18 high-level terms (UBERON)



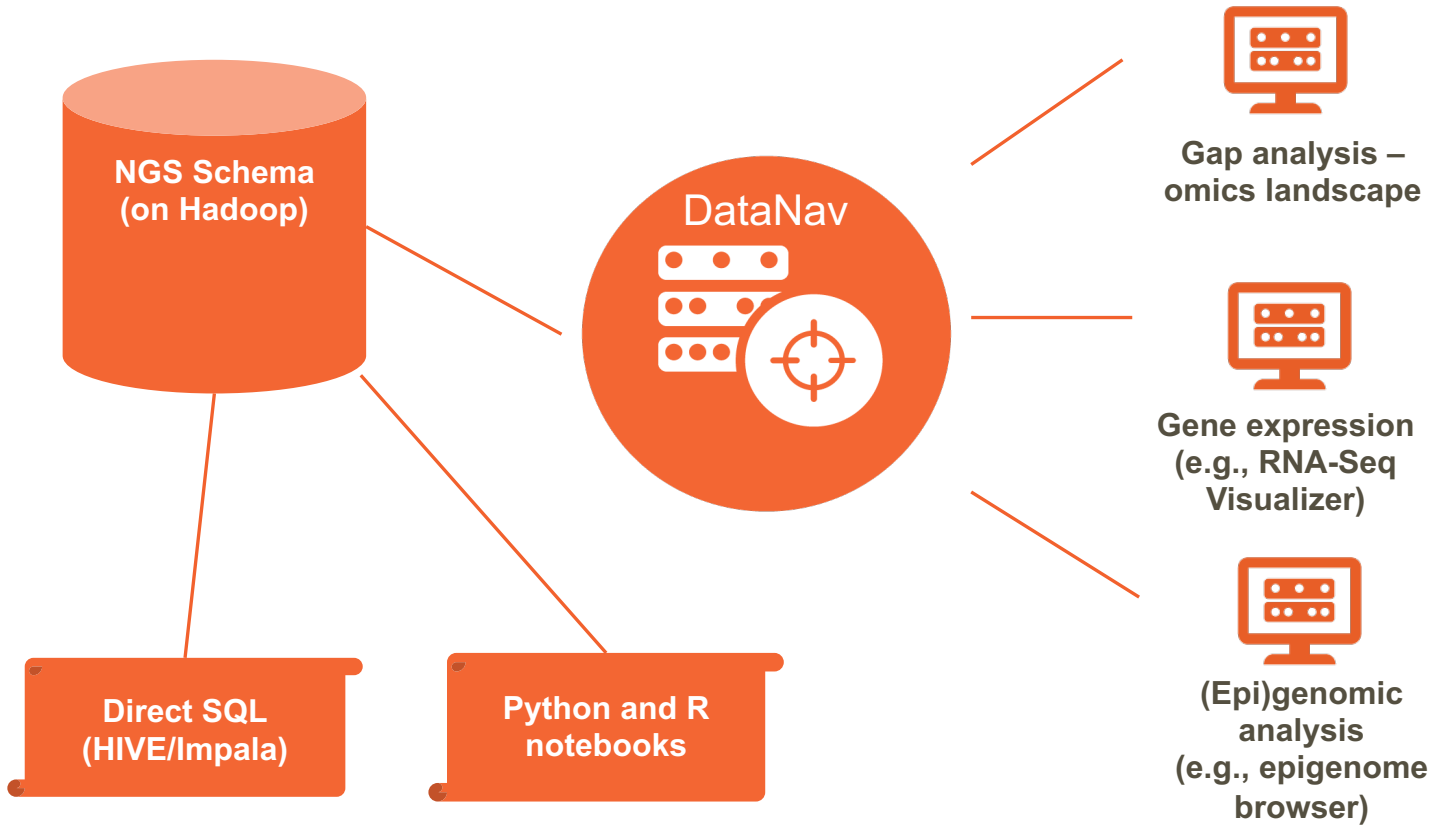


- digestive system|UBERON:0001007
- nervous system|UBERON:0001016
- reproductive system|UBERON:0000990
- respiratory system|UBERON:0001004
- renal system|UBERON:0001008
- endocrine system|UBERON:0000949
- integumental system|UBERON:0002416
- musculoskeletal system|UBERON:0002204
- external soft tissue zone|UBERON:0034929
- hematopoietic system|UBERON:0002390
- cardiovascular system|UBERON:0004535
- immune system|UBERON:0002405
- others

Project Significance to GSK

- This effort, focused on Oncology & Immunology, has become a key part of GSK's R&D ecosystem
- For first time allowing integration of internal and external data together and going from raw to controlled data.
  - Integrated and harmonized data advances our collective understanding of the captured observations and measurements which brings new insights from advanced analytics to create impact.
- The OMICs metadata harmonization initiative has helped to drive internal discovery decisions and efforts

Harmonising Metadata from External Sources



Rancho biosciences

High Quality, Curated Datasets Now Available

Pre-curated Datasets – One Time Fee

Number	Name	1st Delivered	Update	Fee
1	Depmap	4Q 19	1Q 21	$35K
2	Roadmap	4Q 19		$20K
3	Blueprint	4Q 19		$30K
4	L1000/LINCS	4Q 19		$35K
5	GEO (316 I/O* datasets)	4Q 19 (87) 4Q 20 (229)		$30K $90K
6	GTEx v8	1Q 20		$30K
7	CCLE	1Q 20		$20K
8	TCGA	1Q 20		$30K
9	ENCODE	1Q 20		$35K
10	Cancerrxgene	1Q 20	2Q 21	$30K
11	PRISM	1Q 20	1Q 21	$40K
12	Achilles	1Q 20	1Q 21	$50K
13	Human Protein Atlas	1Q 20		$35K
14	Human Cell Atlas	2Q 20		$35K
15	CGCI	2Q 20	2Q 21	$65K
16	DICE	2Q 20		$30K
17	Cytoimmunogenomics	3Q 20		$30K
18	SRA(2)/ArrayExpress (3)	1Q 21		$25K

Welcome suggestions of other datasets where a cost-sharing model can be leveraged, including in other therapeutic areas.

* Immunology/Oncology

Next steps

- All datasets curated in this effort are now available for purchase and delivered “as-is”
  - Any work to align with client’s specific infrastructure will be scoped as a separate effort, in addition to the pre-curated fee.
- Maintenance?
  - For a fee of $5K, client can subscribe to an alert service for 12 months.
  - Once a quarter, Rancho will provide a roster indicating where there have been updates to all data sources.
  - If there are updates in datasets, then a “refresh project” for those interested in receiving updated versions can be discussed under a cost share model.

Questions and Discussions

Thank you!

Rancho Bio Sciences

16955 Via Del Campo Suite 200
San Diego, CA 92127

Phone: (760) 717-7881

services@ranchobiosciences.com

Phone: (760) 717-7881

services@ranchobiosciences.com

Additional Questions 1

- Q: Was single cell data included?
  - We curated a couple of single-cell datasets, such as included in “Cytoimmunogenomics” and “HCA” sources. Some metadata attributes that are essential for single-cell studies, such as “Number of cells”, “Cell Sorter”, “Gated Cell Marker”, “Library selection”, “Technology type”, “Read Type”, “Read Strand”, “Experiment Platform” and “Instrument Details” are captured if related information is available.
  - Metadata is annotated at the sample level. Predicted cell types are down to the downstream users.
- Q: Is the data model proprietary? Are there plans to publish this, or are you licensing the model to potential other customers?
  - The metadata delivery data models are no longer proprietary. The conceptual data model is proprietary.
  - Rancho can help clients develop data models tailored to their needs.
- Q: Can you speak to automated approaches to labeling or filter data of questionable or unverified accuracy?
  - Using Python or R-based scripts, we generate a metadata summary for controlled metadata entities. This provides information on unique value of each entity. The script will also check whether all values for controlled columns exist in our value normalization mapping files. All standard terms in value normalization files are validated using v-lookups and scripts against standard ontology term/code pairs maintained in Rancho’s ontology database. Using scripta, we also extract unique value pairs from related metadata entities. These automated approaches help us to identify data that are questionable or inconsistent. Meanwhile, statistical analysis is applied to quantitative data, to identify “impossible” values, missing values, outliers, etc.

Additional Questions 2

- Q: dbGaP has the metadata template for NGS data submission to SRA. Does your universal metadata model harmonize and integrate all NCBI's data elements for NGS data?
  - The current effort has focused on open-access datasets. Controlled studies that are deposited in dbGaP are not included in this initiative. There are a couple of SRA studies in our 2021 delivery, where almost all metadata entities (sample level and project level) pulled from SRA have been successfully aligned to the current data model.
- Q: What were some of the key points you used within GSK and “metadata-positive” folks to inspire other areas of the organization to curate their metadata?
  - A mix of carrot and stick!
  - If users use the internal requests submission tools and fill out the minimum metadata, their sequencing and mass spec. etc. requests and data processing is much smoother for them. The tools have mandatory and expected fields, with controlled vocabulary lookup/pull-downs. Also, many metadata values are automatically added by context. GSK’s goal is for the data to be already annotated with metadata rather than getting people to annotate after the event and to automatically pass metadata along with the data where appropriate.
  - For others with different data types from different processes, that ultimately want to put theirs into the same end schema, they need to provide minimal metadata using appropriate vocabulary terms. etc. Their data can then be pipelined and made visible in the same tools as with the other data.
  - From the “carrot” side of things, if the key end scientific users get the metadata they need to answer questions, they will continue to commission more wet experiments.
  - There have also been several talks within the R&D org. about the power of metadata.