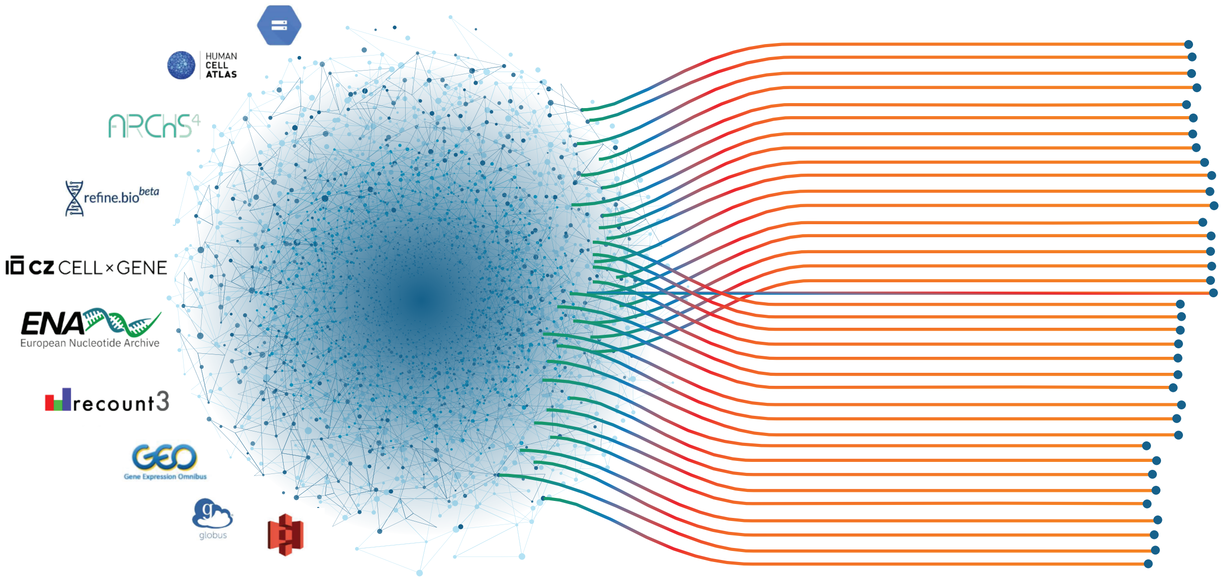
strand

# Data Curation, Integration And Harmonization



# #1 Data Ingestion, Harmonization, And Curation For Multiomics Datasets With Integrated Ontologies

## The Problem

In multi-omics research, achieving valuable outcomes requires systematic management and curation of diverse datasets while adhering to FAIR data principles.
The Needs include:

- Onboarding public datasets (data wrangling and curation)
- Curating external datasets to specified templates
- Curation of internal datasets (e.g., Benchling notebooks)
- Defining and updating ontology terms
- Managing data ingress and egress
- Onboarding restricted datasets

## Our Process

- Data sourcing: Processing of data from the comprehensive multi-omics data sources
- Controlled vocabularies
- Metadata Schema: Creation of a schema tailored to specific data needs for effective organization and retrieval.
- Standardization: Enhancing/customizing ontology dictionaries with standardized terms for better data organization and interoperability
- Clear Definitions: Providing clear definitions to reduce ambiguity and improve understanding.
- Training: Training users on controlled vocabularies to enhance data management and usage.

| | | |
|---|---|---|
| 1 | Large public databases | GEO Gene Expression Omnibus<br>ENA European Nucleotide Archive<br>NIH SRA |
| 2 | Published articles & associated supplementary material | PubMed |
| 3 | Data repositories on cloud, data transfer tools | amazon web services<br>S3<br>Google Cloud<br>globus |
| 4 | Custom databases | CZ CELL×GENE<br>HUMAN CELL ATLAS<br>ARCHS4 <py><br>recount3<br>refine.bio |
| 5 | In-house generated data | Lab<br>Collaborators<br>CROs |

| Description | Convention / Ontology | Example |
|---|---|---|
| A developmental stage is spatio-temporal region encompassing some part of the life cycle of an organism, e.g. blastula stage. This is specific to when the specimen was isolated from the organism. | An entry from OBO Human Developmental Stages (obo:hsapdv) or OBO Mouse Developmental Stages (obo:mmusdv) depending the species. | Carnegie stage 23, 9th week post-fertilization human stage, adolescent stage, third decade human stage, 50-year-old human stage. |
| A disease is the outward maniestation of one or more disorders. List the diseases (if known) which impact the source organism from which samples are derived. | Must be a disease name given in Mondo Disease Ontology (Mondo). Enter 'normal' if no known disease. | Alzheimer's disease |
| Must be a cell line given in the Cell Line Ontology (CLO). The name should be used and not the identifier. | Uberon, BTO. | U-2 OS cell |

## Sample metadata schema fields, descriptions and reference ontologies with example terms

| Metadata Fields | Revised Schema Field title |
|---|---|
| donor accession0 | rganism_accession |
| donor name | Organism_name |
| donor description0 | rganism_description |
| biological replicate0 | rganism_biological_replicate |
| donor genotype | Organism_genotype |
| donor species | Organism_species |
| donor biological sex0 | rganism_biological_sex |
| donor age | Organism_age |
| donor age units | Organism_age_unit |
| donor developmental stage | Organism_developmental_stage |
| donor disease0 | rganism_disease |
| donor disease model | Organism_disease_model |
| **Specimen** | |
| specimen accession | BiologicalSpecimen_accession |
| specimen name | BiologicalSpecimen_name |
| specimen type | BiologicalSpecimen_specimen_type |
| specimen tissue | BiologicalSpecimen_tissue |
| cell category | CellLine_cell_category |
| cell line | CellLine_cell_line |
| cell type | CellLine_cell_type |
| cell biological replicate | CellLine_biological_replicate |
| cell technical replicate | CellLine_technical_replicate |
| **Process** | |
| cell passage | CellLineProcessing_passage |
| cell population doubling | CellLineProcessing_population_doubling |

## Result / Impact

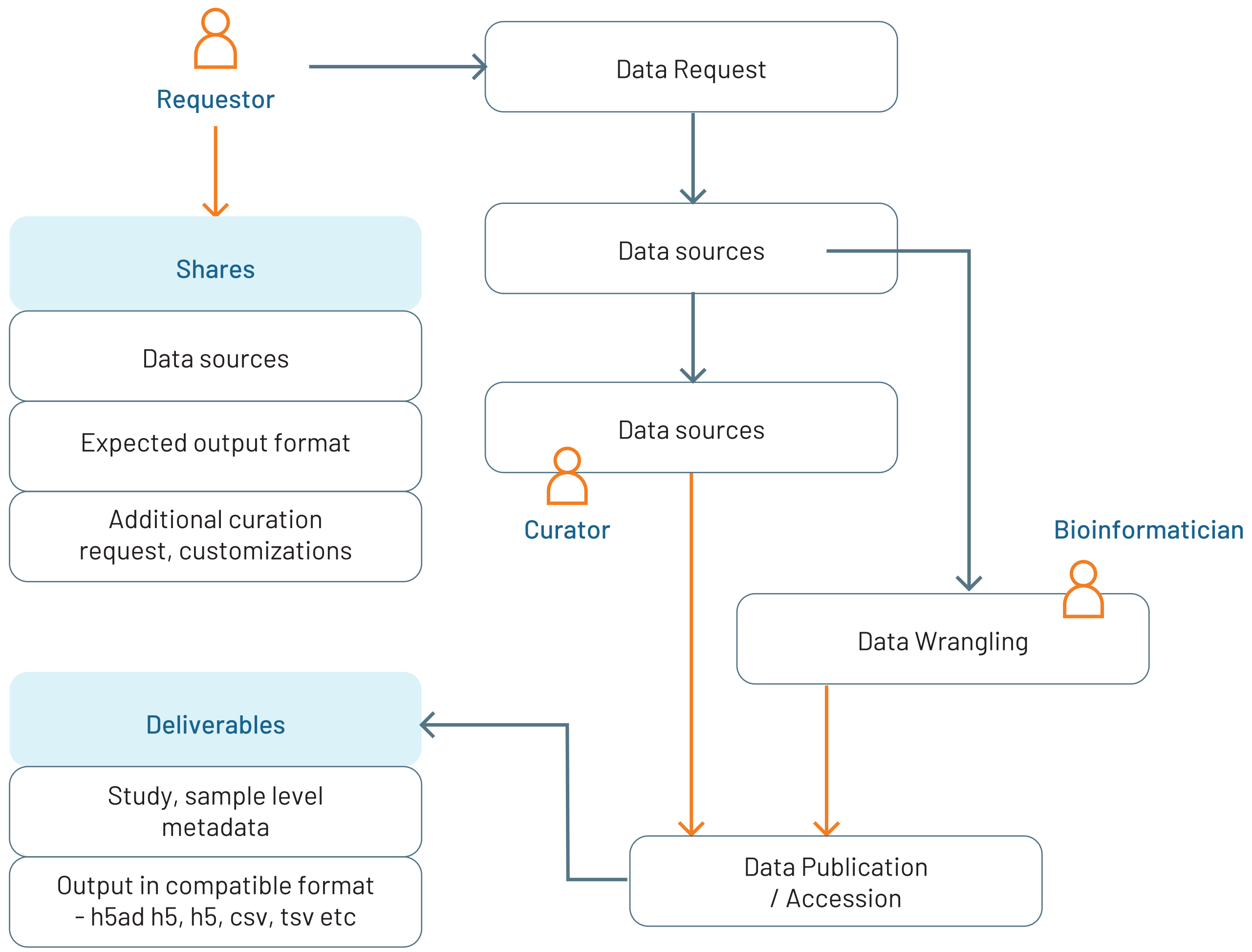
- Harmonized over 40 diverse datasets (both external and internal) for the customer

**We are developing LLM models to automate ontology mappings and are currently working on a proof of concept.**

## Data Onboarding

| | | |
|---|---|---|
| 1 | Data consolidation | Integrating data from various sources into a unified format for analysis |
| 2 | Data Upload | Uploading the final format to cloud or designated locations with appropriate access controls |
| 3 | Process Documentation | Documenting the entire process with Readmes and notes |

## Task Management


Requestor
Data Request
Shares
Data sources
Expected output format
Additional curation request, customizations
Data sources
Data sources
Curator
Bioinformatician
Data Wrangling
Deliverables
Study, sample level metadata
Output in compatible format - h5ad h5, h5, csv, tsv etc
Data Publication / Accession

## Achievements in Data harmonization and integration

1. **Metadata harmonization** — Organized data, enabling efficient downstream ML workflows
2. **Comprehensive** — Integrated large datasets across various disease conditions
3. **Integrated Atlas and Model** — Created a time-series model for studying cell-cell interactions and understanding cell types
4. **Improved Turnaround Time** — Reduced data ingestion time into the data lake, accelerating ML processes

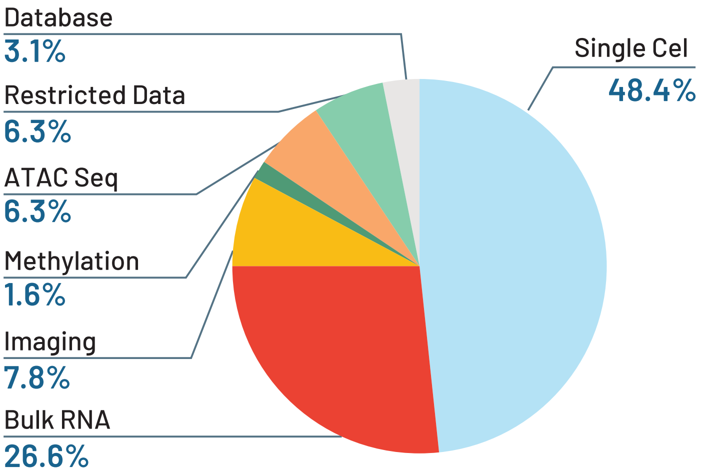
   ¥ 5-10 Datasets: 5 days turnaround
   ¥ 50+ Datasets: 2-3 weeks turnaround

## Harmonized Data enriched Collaborations

1. **Enhanced Collaboration** — Fostered cross-domain innovation and collaboration among data users
2. **Seamless Integration** — Achieved smooth integration with other datasets and systems
3. **Standards Compliance** — Adhered to global interoperabilitystandards set by international organizations/ industry consortia
4. **Reliable Versioning** — The version control implemented helped to track data changes, ensuring transparency and reproducibility

| | |
|---|---|
| Number of datasets filtered | 21,689 |
| Number of shortlisted studies | 7,168 |
| Number of samples | 63,533 |

Example of data volume managed for the customer as of August 2024 for an ongoing project



# #2 Scaling Omics Ingestion & Querying In Data Lakes To Improve Curation Efficiency

*LLM Solutions to FAIRify Omics & Metadata*

## The Problem

Biotech scale: 100s-1000s of datasets per day from public and internal sources
Pharma scale: Legacy data (for ex: RNA-Seq) at large scale, others small scale
Challenges:

- deposit all common data (preclinical pathology, biomarker, in vivo imaging) in one data lake
- multiple omics + non-omics modalities
- multiple ontologies and the need for a central ontology
- the central data lake has to serve each customer its own endpoint while controlling access to sensitive data

## Our Process

### Data Sources & Ontology

- Talk to stakeholders re: data sources + ontologies in use
- Arrive at consensus ontologies such as Uberon for anatomy
- Deposit central ontologies in data lake

### AI-enabled ingestion

- Curate all existing data to centralized schema + ontologies
- Use for ex. LinkML for ontology validation
- Use LLMs to assist curation of specific fields

### Data Lake Architecture + Governance

- Architect data lake to support common queries
- Use AWS dBs s.a. RDS, Dynamo, Redshift or Neptune depending on use case
- Parametrize w.r.t cost vs long term need

### Visualization + ML environments + AI tools

- Add downstream environments s.a. Sagemaker, Quicksight, and Shiny
- Write APIs to must-use platforms s.a. Omero
- Use AWS or equivalent for data governance
- Write custom AI tools

## Result / Impact

- We manually curated the 10x library prep field from free text for 3k datasets
- We adopted GPT powered approach for ontology control of the 10x library prep field
- The GPT-4 approach performed at an accuracy of 95% and was robust to new data
- Overall ingestion TAT improvement = 2.5-3x over purely manual curation

GPT 4 Turbo

**We improved ingestion TAT by 3x with a GPT-4 powered approach.**

# #3 Curation Of A Reference Knowledgebase To Derive Compound Effects

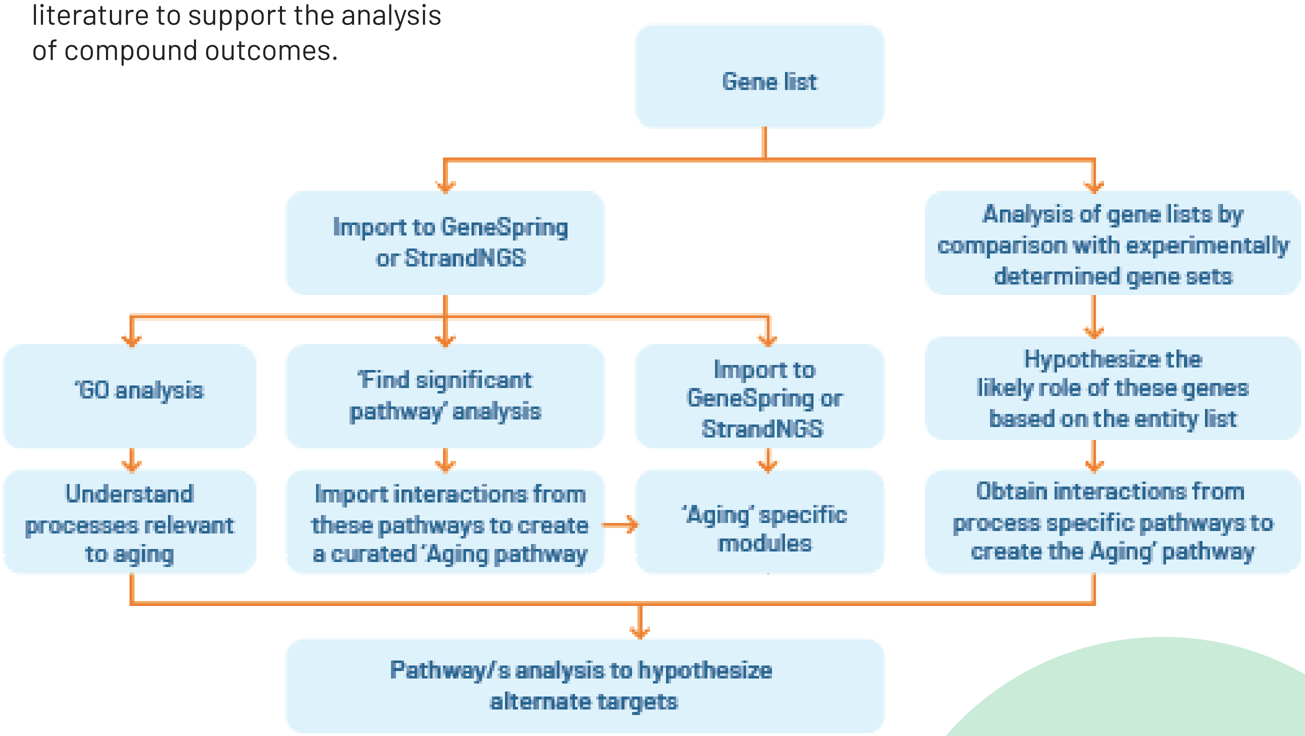
## The Problem

The Need: Build a comprehensive knowledgebase for compound outcome analysis

- Create a resource to evaluate the molecular, functional, and process-level outcomes of compounds of interest.
- Incorporate entities from various databases to build a comprehensive knowledgebase.
- Include additional entities from published studies to enrich the knowledgebase.
- Collect evidence statements from literature to support the analysis of compound outcomes.

## Our Solution

- Strand's NLP tool provided an exhaustive entity (gene) list with supporting statements, directions of interactions and interaction types.
- Curation scientists verified each NLP derived supporting statement and reviewed literature to include any missing entity/ interaction with supporting statements.
- A process from gene list to pathways was followed as depicted, for each of the client's functional/toxicology areas of interest.
- Overlaying the client's compound-treated gene expression data on the curated pathways, we discovered a gene expression signature in skin cells that mapped to the cholesterol synthesis pathway. This finding was particularly exciting for the client team, as it directly supported their ongoing research on a moisturizer-inducing compound.



Gene list

Import to GeneSpring or StrandNGS

Analysis of gene lists by comparison with experimentally determined gene sets

'GO analysis

'Find significant pathway' analysis

Import to GeneSpring or StrandNGS

Hypothesize the likely role of these genes based on the entity list

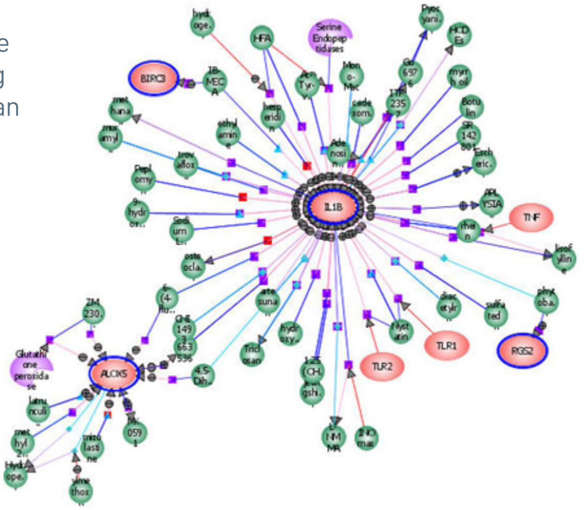
Understand processes relevant to aging

Import interactions from these pathways to create a curated 'Aging pathway

'Aging' specific modules

Obtain interactions from process specific pathways to create the Aging' pathway

Pathway/s analysis to hypothesize alternate targets



The process from gene lists to pathways using the aging pathway as an example

**This knowledgebase and pathways generated has been successfully used by our client in defining the biological function and toxicology effects of their compounds.**

# #4 Data Curation And Organization

## The Problem

- >=1000 public GEO partially curated datasets which lack to consistency required for in
- The Need: to develop a pipeline for processing external data that can also be extended to include internal data

## Our Process

- A streamlined process ensured that the large datasets were handled in a timely and accurate manner

**Data from User**

- Raw data - fastq
- Processed data - counts, metadata
- Analysis - txt, png, R codes etc

**Data Curator**

- Proposed the data hierarchy
- Validated the controlled vocabulary
- Improved controlled vocabulary by adding new terms
- Created annotations by reading source publications

**Data Manager**

- Transferred the data files to the proposed location
- Captured Information with respect to:
  - ○ Study
  - ○ Experiments within each study

**Bioinformatician**

- Created reference dataset with standardized sample names, metadata columns, cleanup of annotations (cell type, samples, etc)
- Validated the files created
- Transferred to staging location

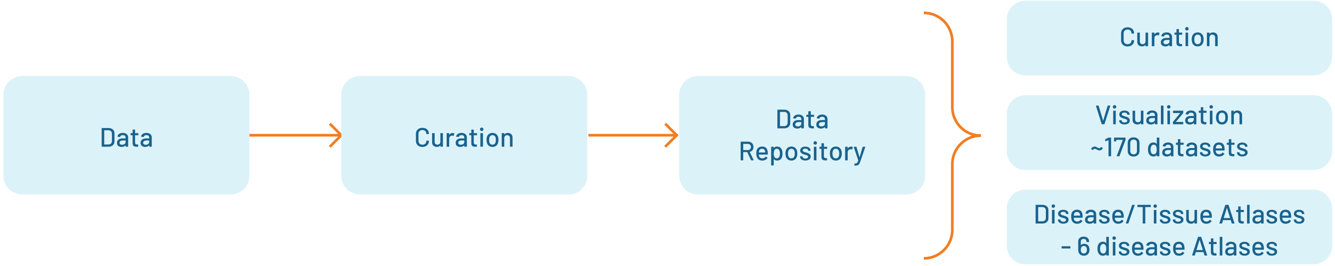
**Data Engineer**

- Pushed the reference dataset created to required databases.
- Upload to Client's platform
- Rechecked the data hierarchy

**We implemented a unified data platform for researchers across the company, supporting seamless integration of new data and metadata while enabling the incorporation of datasets from emerging experimental methodologies**

## Result / Impact

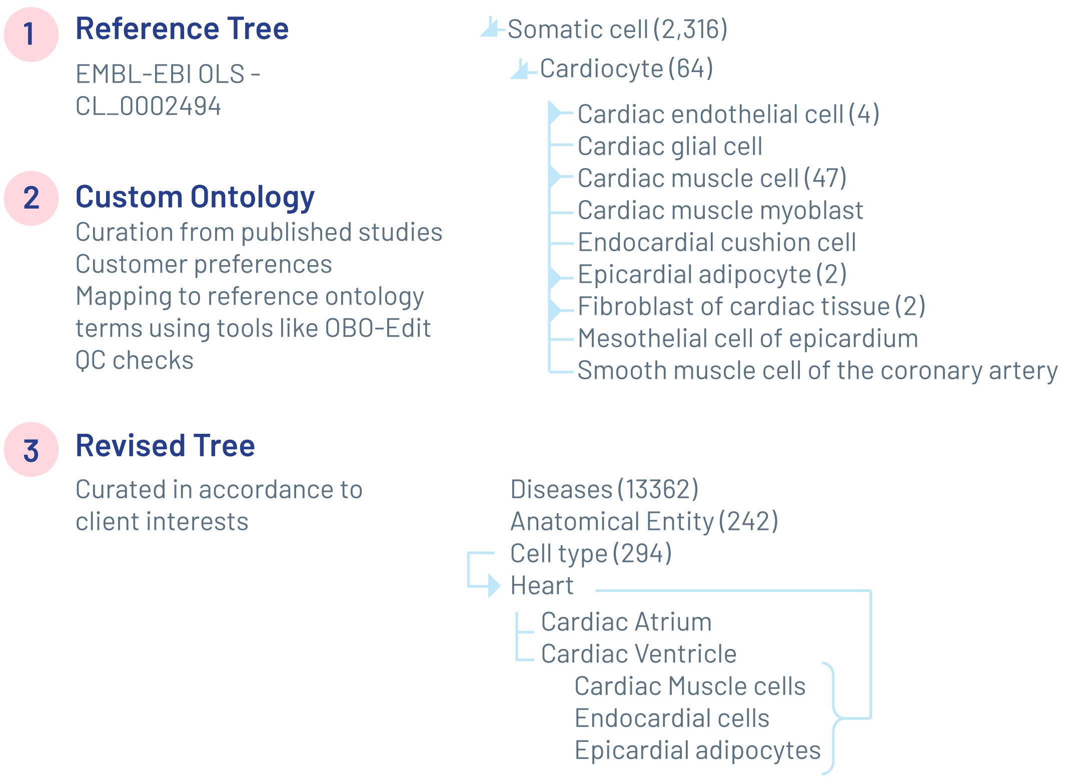
- A clean and well-organized dataset with a streamlined hierarchical structure that allows for easy tracking
- Organized a centralized data repository harboring 1500+ datasets
- Generated single unified datasets for analysis, across multiple datasets corresponding to each disease/area of interest
- Provided cleaned/organized data spanning diverse fields, such as:
- Cell level - disease, cell type, anatomy
- Study level - platform, sample matrix, processing
- Supported data integration/harmonization in collaboration with the client team



Data → Curation → Data Repository →

- Curation
- Visualization ~170 datasets
- Disease/Tissue Atlases - 6 disease Atlases

Cell type ontology custom curation for the client - an example



1 **Reference Tree**

EMBL-EBI OLS - CL_0002494

Somatic cell (2,316)
  Cardiocyte (64)
    Cardiac endothelial cell (4)
    Cardiac glial cell
    Cardiac muscle cell (47)
    Cardiac muscle myoblast
    Endocardial cushion cell
    Epicardial adipocyte (2)
    Fibroblast of cardiac tissue (2)
    Mesothelial cell of epicardium
    Smooth muscle cell of the coronary artery

2 **Custom Ontology**

Curation from published studies
Customer preferences
Mapping to reference ontology terms using tools like OBO-Edit
QC checks

3 **Revised Tree**

Curated in accordance to client interests

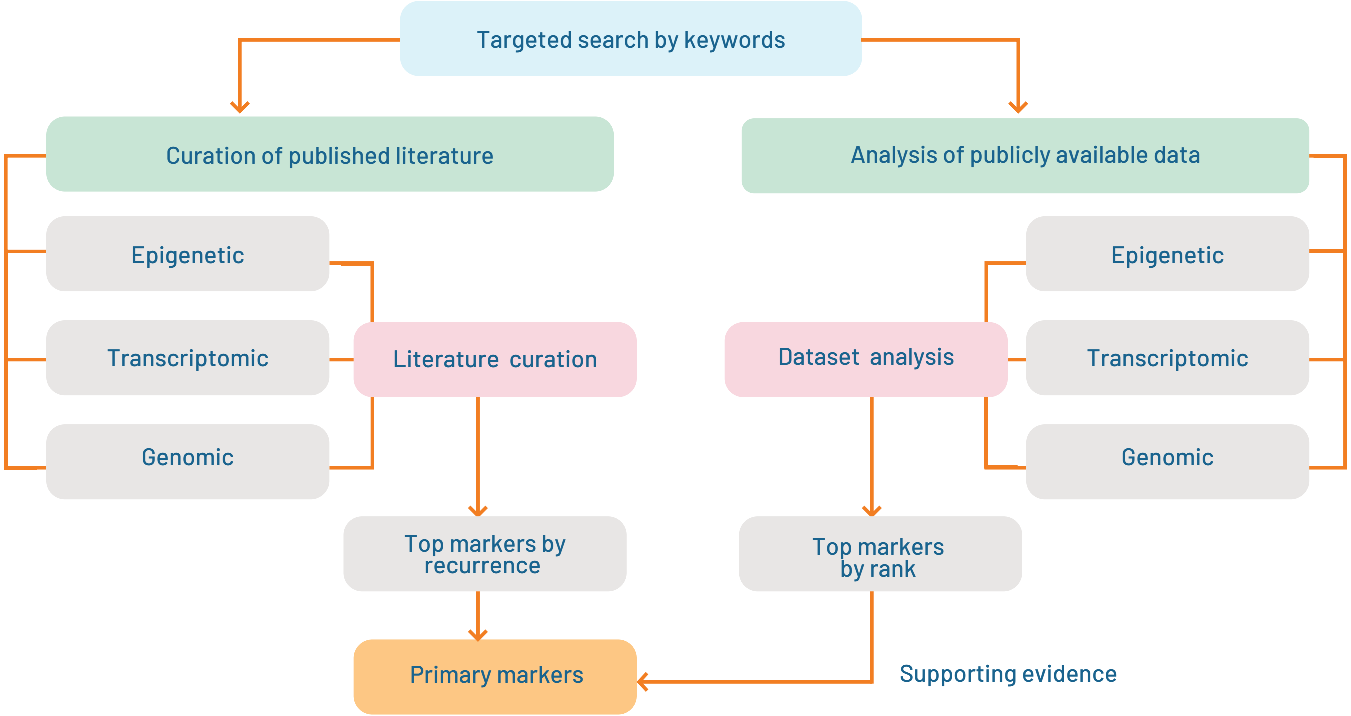
Diseases (13362)
Anatomical Entity (242)
Cell type (294)
Heart
  Cardiac Atrium
  Cardiac Ventricle
    Cardiac Muscle cells
    Endocardial cells
    Epicardial adipocytes

# #5 Potential Biomarkers For Early Detection Of Rheumatoid Arthritis

## The Problem

The Need: To identify potential non-invasive biomarkers for the early detection of RA within the 'window of opportunity'.

## Our Process



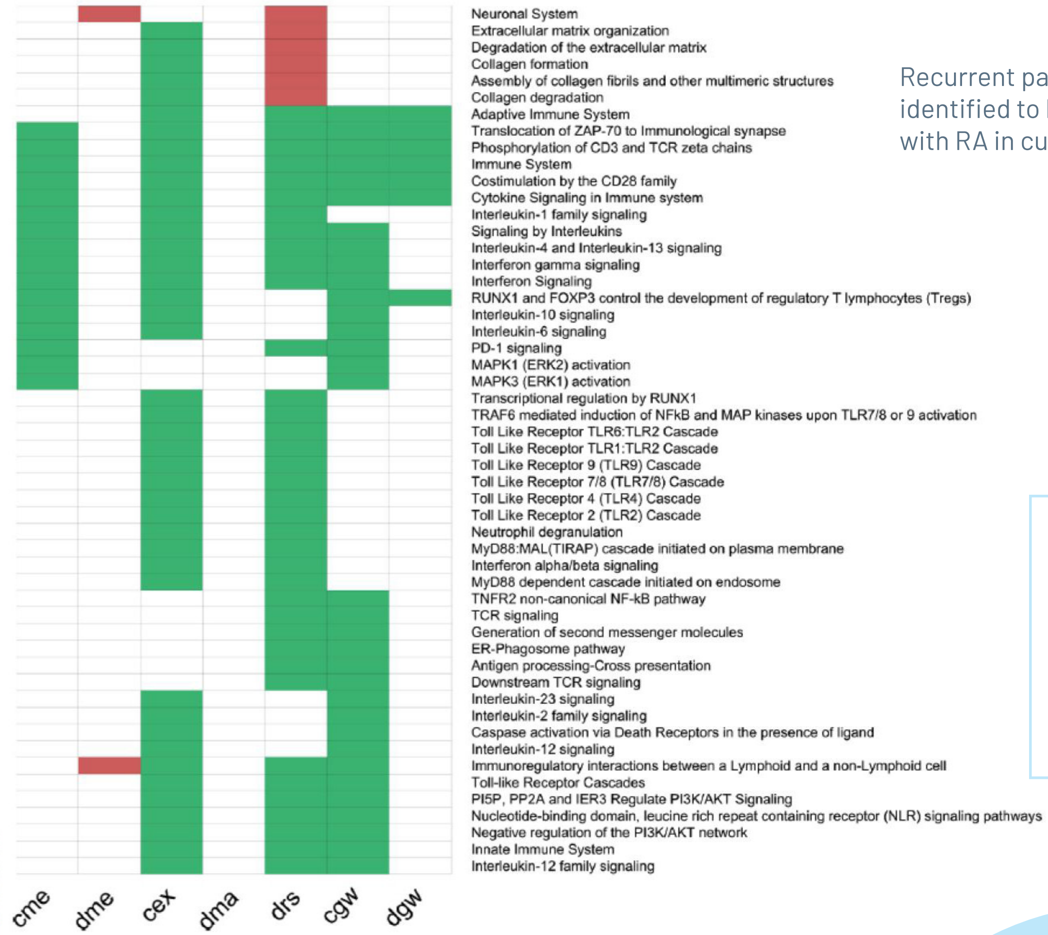
## Result / Impact

| Methylation markers | |
|---|---|
| 44 studies | : 135 genes |
| 3 datasets | : 2024 genes |
| Common trend | **: 12 genes** |

| Transcriptomic Markers | |
|---|---|
| 51 studies | : 301 genes |
| 14 datasets | : 2312 genes |
| Common trend | **: 44 genes** |

| Genomic Markers | |
|---|---|
| 44 studies | : 190 genes |
| 3 datasets | : 37 genes |
| Common trend | **: 21 genes** |

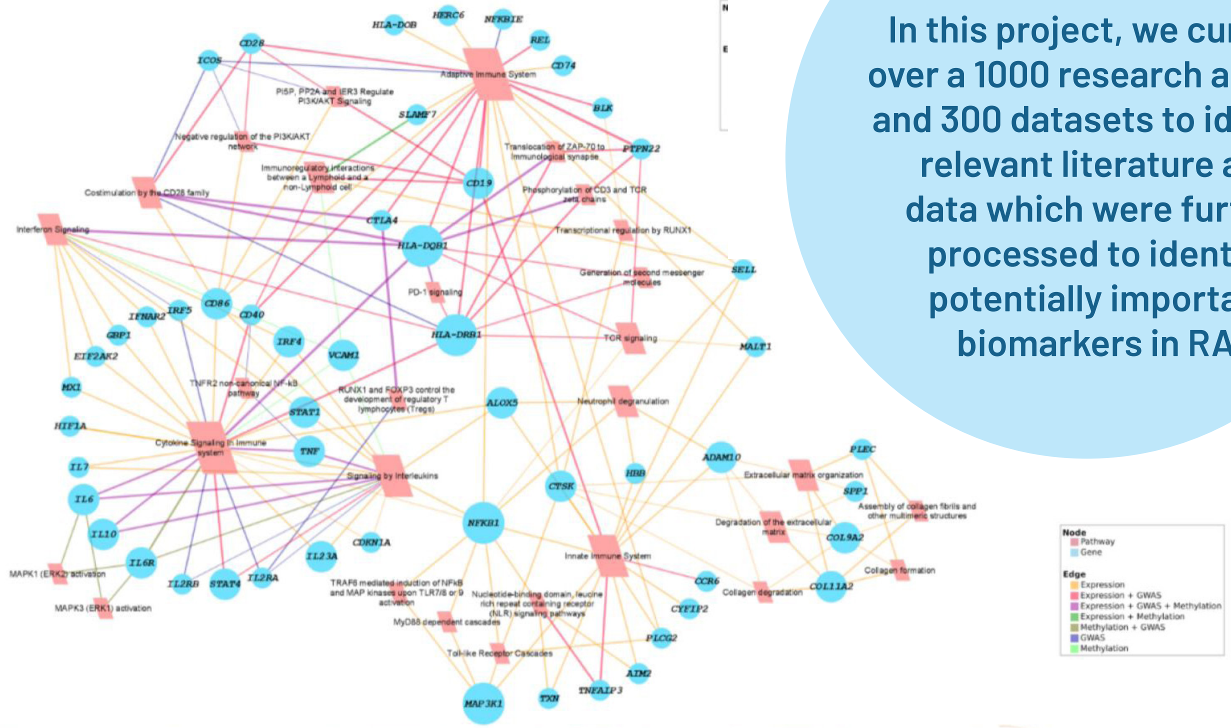
- Our analysis refined a list of top 30 markers from the initial pool of 20,000 genes, based on the recurrence across datasets analysis and literature curation.
- A comparison of the individual gene lists derived from data analysis and curation against all coding genes, followed by gene set enrichment analysis on methylation, microarray, and RNAseq datasets, revealed enriched pathways. Of these 52 pathways were identified as recurrent in 2 or more gene lists.
- We selected 53 genes from the enriched pathways (previous slide) which are potentially important biomarkers in RA with diverse mechanisms of action while sharing similar or common pathways.
- The 53 genes identified from the curation/ analysis exercise (above) when visually mapped to the 52 recurrent pathways , highlighted significant gene-pathway associations and shared pathways.
- Several top candidates linked to significantly enriched pathways and potential biomarkers for early RA, including key immune system targets.



Recurrent pathways across genes identified to be significantly associated with RA in curation and analysis efforts

cme: curation Methylation
cex: curation Expression
cgw: curation GWAS
dme: dataset Methylation
dma: dataset Microarray
drs: dataset RNAseq
dgw: dataset GWAS

Pathway to gene association generated from recurrent enriched pathways and genes across curation and analysis efforts


HLA-DOB
HERC6
NFKBIE
REL
CD74
CD28
ICOS
Adaptive Immune System
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling
SLAMF7
BLK
Negative regulation of the PI3K/AKT network
Translocation of ZAP-70 to Immunological synapse
PTPN22
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
CD19
Phosphorylation of CD3 and TCR zeta chains
Costimulation by the CD28 family
CTLA4
Interferon Signaling
Transcriptional regulation by RUNX1
HLA-DQB1
Generation of second messenger molecules
SELL
PD-1 signaling
CD86
IRF5
IFNAR2
CD40
GBP1
IRF4
HLA-DRB1
TCR signaling
MALT1
EIF2AK2
VCAM1
MX1
TNFR2 non-canonical NF-kB pathway
RUNX1 and FOXP3 control the development of regulatory T lymphocytes (Tregs)
ALOX5
Neutrophil degranulation
STAT1
HIF1A
Cytokine Signaling in Immune system
TNF
ADAM10
PLEC
IL7
Signaling by Interleukins
HBB
Extracellular matrix organization
CTSK
SPP1
Assembly of collagen fibrils and other multimeric structures
IL6
NFKB1
Degradation of the extracellular matrix
IL10
CDKN1A
COL9A2
IL6R
IL23A
Innate Immune System
Collagen formation
MAPK1 (ERK2) activation
IL2RB
STAT4
IL2RA
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways
CCR6
Collagen degradation
COL11A2
MAPK3 (ERK1) activation
MyD88 dependent cascades
CYFIP2
PLCG2
Toll-like Receptor Cascades
AIM2
TNFAIP3
MAP3K1
TXN
Node
Pathway
Gene
Edge
Expression
Expression + GWAS
Expression + GWAS + Methylation
Expression + Methylation
Methylation + GWAS
GWAS
Methylation

**In this project, we curated over a 1000 research articles and 300 datasets to identify relevant literature and data which were further processed to identify potentially important biomarkers in RA.**

# Omics CRO

## Curation

15 years of experience curating variants, genes, pathways and diseases for clinical reporting and pharma/biotech custom solutions

**~50**
**Molecular Biologists**

## Bioinformatics and Software

22 years of experience providing bioinformatics solutions to global instrument, diagnostic and pharma companies

**~220**
**SW Engineers, Bioinformaticians**

## Omics Assays

11 years of experience with sequencing-based diagnostics across oncology and genetics, at our CAP lab in India

**~90**
**Lab Scientists, Clin. Res. Scientists**

**24+** YEARS OF EXPERIENCE

**80,000+** Genetic Tests Reported

**500+** Projects Executed for Genomics Majors Globally

Presence in **20+** Countries

strand

7th Floor, MSR North Tower, #144, Outer Ring Road, Nagavara, Bengaluru - 560045

☎ +91 9980448044 hello@strandls.com strandls.com