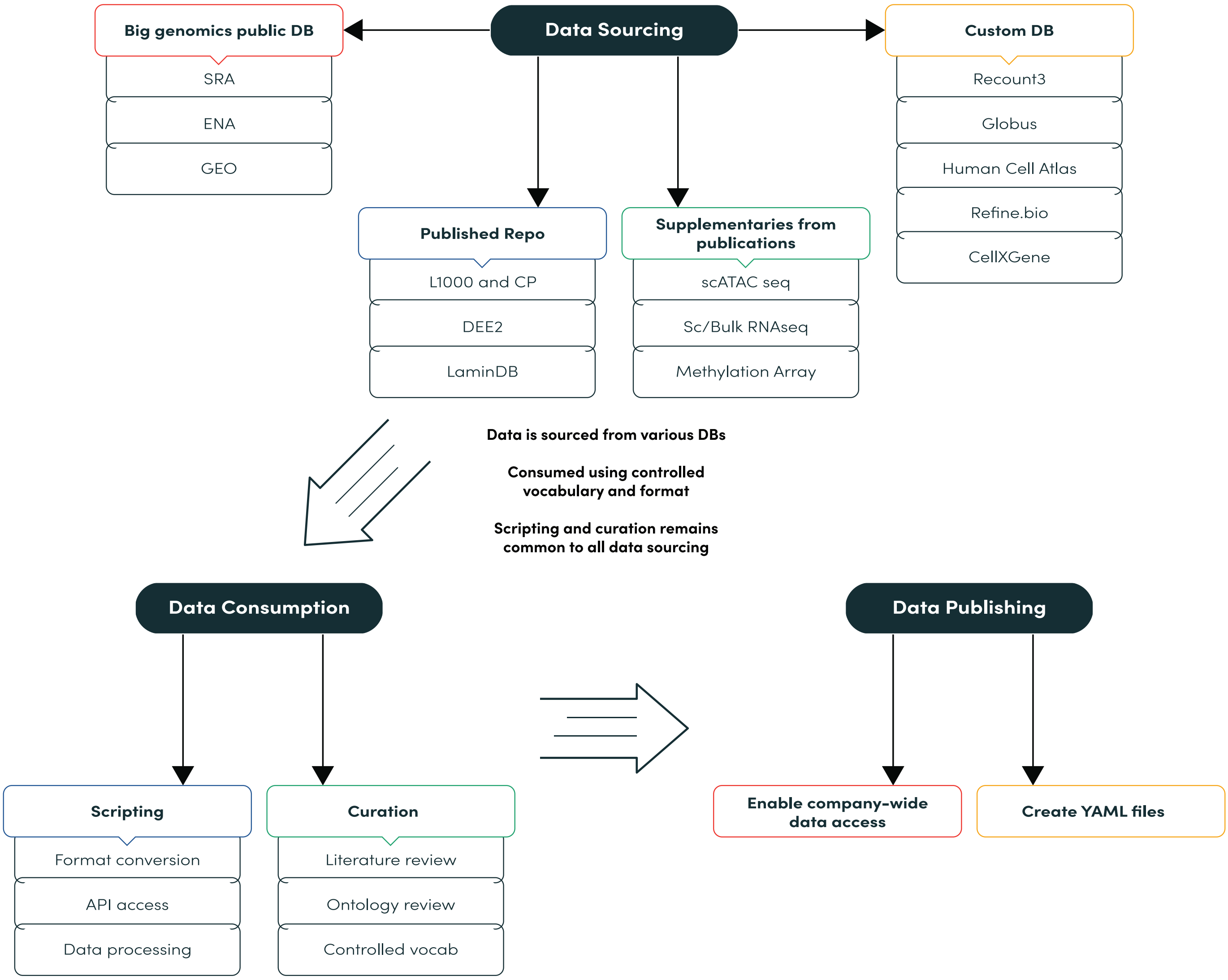
WE SOLVE PROBLEMS IN GENOMICS.

From Sourcing to Consumption:

Strand’s Data Harmonization Process

Strand acquires, processes, and curates diverse types of biomedical data and makes it analysis-ready



Data Sourcing

Big genomics public DB
- SRA
- ENA
- GEO

Custom DB
- Recount3
- Globus
- Human Cell Atlas
- Refine.bio
- CellXGene

Published Repo
- L1000 and CP
- DEE2
- LaminDB

Supplementaries from publications
- scATAC seq
- Sc/Bulk RNAseq
- Methylation Array

Data is sourced from various DBs

Consumed using controlled vocabulary and format

Scripting and curation remains common to all data sourcing

Data Consumption

Scripting
- Format conversion
- API access
- Data processing

Curation
- Literature review
- Ontology review
- Controlled vocab

Data Publishing
- **Enable company-wide data access**
- **Create YAML files**

Strand's Workflow

Our team of data stewards and curators has established a two step process for harmonizing data from various sources:

Data sourcing

Access data from four main types of repositories

- Big genomics public databases
- Published repositories
- Publication supplementaries
- Custom databases

Data consumption

- Convert acquired data to standard formats
- Develop custom scripts for format conversion, API access and data processing
- Curate data by performing thorough literature and ontology reviews using controlled vocabulary dictionaries

Data publishing

- Create YAML files post curation
- Enable access to this data across the client's organization

Client Collaboration

Operational Overview

- Strand is currently implementing this workflow for a notable California-based biotechnology company. Our data harmonization (DH) team collaborates with the client, onboarding required datasets mainly from public databases and occasionally from research publications.
- A streamlined process follows:
  - The data steward steps in to understand the native data format in the various sources and converts it to the client-preferred structure
  - The data curator reviews each dataset and manually populates missing or inconsistent data fields based on metadata schemata. Curation is performed at the study, sample and sometimes at the single-cell level based on client-specific controlled vocabulary ontologies
  - The data is then made available for release into the client's repository

Work management

- Data onboarding requests are processed through the JIRA ticketing system
- Large ticket items are subdivided and assigned to Strand's data stewards and curators
- The steward sources and restructures the dataset, while the curator populates both the common and specific data fields using a combination of scripting and manual entry

Client's Problem	Definition	Time Taken
Cell type	A cell type is a distinct morphological or functional form of the cell.	3 minutes
Title of the publication	A title is a textual entity that summarily describes some entity /the study.	1 minute
Study ID	A study identifier is an identifier that identifies some study.	Automated
Experiment	Study design in terms of control vs disease, treatment vs naive, moderate vs severe etc.	2 minutes
Disease	Association with a disease phenotype or a cell line used to study a disease.	2 minutes

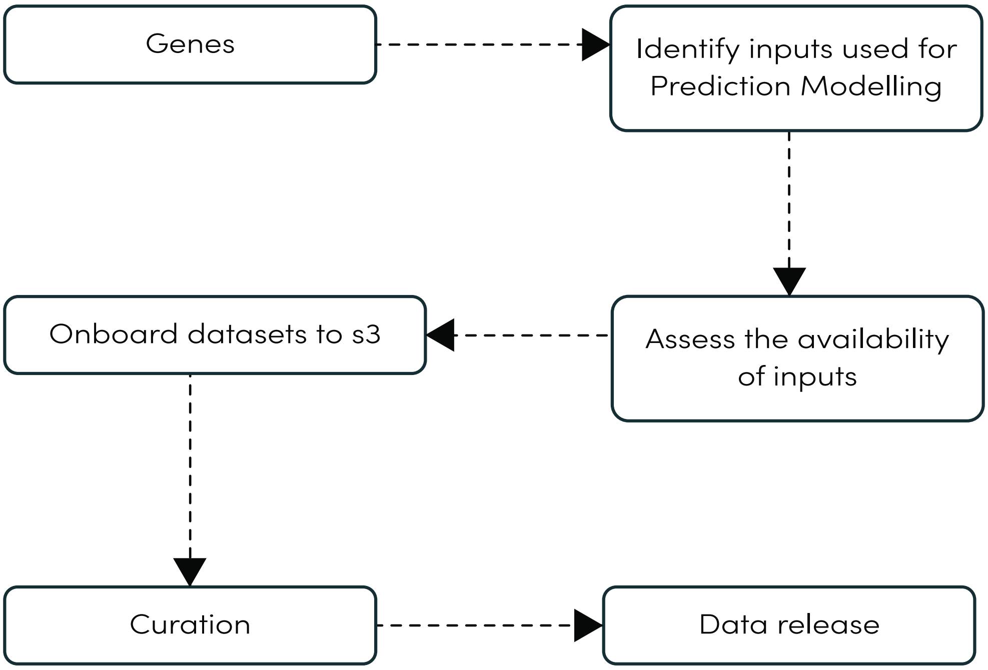
Representative examples of turnaround times per data fields

- Over the past 6 months, Strand has managed a substantial data volume, handling 5-15 requests monthly processing over 1 TB of metadata and onboarding more than 20 TB of raw sequencing data

Specific Highlights from the Collaboration

Cell-painting

- Cell painting is a high-throughput imaging assay capturing a wide array of cellular phenotypes
- Strand has recently ventured into harmonizing cell painting datasets
- Following the client's direction to replicate results from a publication that utilized cell painting and L-1000 assay datasets, our team is working on onboarding and processing the respective datasets
- We have established a workflow that involves onboarding datasets from Github repositories, identifying inputs used for prediction modeling, assessing input availability, onboarding the dataset to S3, curation and data release


Genes

Identify inputs used for Prediction Modelling

Assess the availability of inputs

Onboard datasets to s3

Curation

Data release

Agile Hyper Automation Progression

- In the past six months of engagement, our team has developed ways to automate a significant part of the workflow
- Despite the dynamic nature of the project, our analysis indicates that 50% of the incoming requests follow repetitive patterns, while the remaining 50% are novel requiring some process adaptation
- We have automated the workflow for the predictable 50% of requests, and for the unique client tickets, we have defined clear next steps
- This approach allows us to make quick decisions in response to a wide range of data ingestion requests