Identification and Annotation of Publicly Available Single Cell RNA-sequencing Data

Amrita Bhattacharya[1], Brad Farrell[1], Leonya Ivanov[1], Emmanuel Buys[2], Sondra Kopyscinski[1]

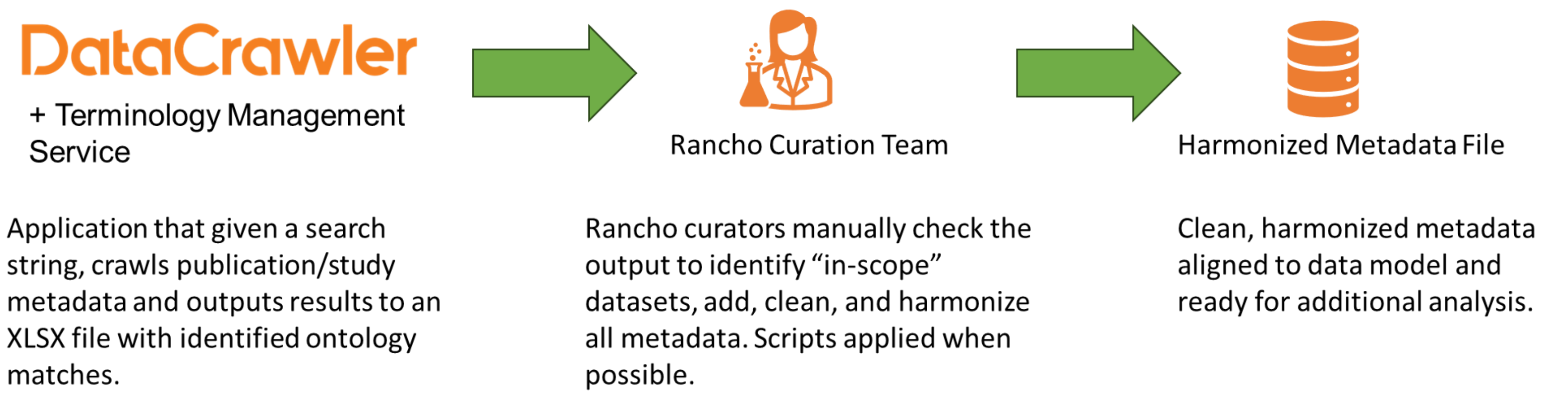
1 Rancho BioSciences, LLC, PO Box 7208, Rancho Santa Fe, CA 92067

2 Cellarity, 101 South St 6th floor, Somerville, MA 02143

Abstract

Single cell technology, specifically single cell RNA-sequencing (scRNA), has already begun transforming the drug discovery and development landscape across all therapeutic areas. However, finding and using publicly available data to test novel hypotheses and validate other findings remains a challenge. Rancho BioSciences has developed a **DataCrawler tool** that allows users to identify datasets available across a wide range of **publicly available** repositories, including NCBI and EMBL resources based on selected keywords. In addition, we use our **Terminology Management Service** to annotate tissue, treatment, and disease. During our collaboration with Cellarity, we enhanced our tool so that it crawls incrementally, allowing us to deliver only new datasets and added additional high-level annotations ("parent-level") in addition to specific disease and tissue annotations. Using the DataCrawler, we have identified **over 4,200 unique** human studies available on Sequence Read Archive (SRA), European Genome-phenome Archive (EGA), database of Genotypes and Phenotypes (dbGaP), and The Broad Institute's Single Cell Portal (SCP). Focusing only on studies with ex vivo clinical samples which are of the most interest to Cellarity's AI-based platform, we have further annotated and harmonized metadata for **over 1,800 studies** using Findable, Accessible, Interoperable, Reusable (FAIR) principles. These studies reveal an increase in the number of datasets with scRNA or single nuclei RNA-sequencing (snRNA) published each year, as well as an abundance of healthy or normal tissues that can be used for reference purposes. Various cancer types, including lymphomas, are the most common diseases profiled after healthy subjects. Easily accessibly tissue, such as blood and other bodily fluids, are the most profiled tissue. **The Rancho BioSciences' DataCrawler is an effective method to identify publicly available datasets and allow downstream users to glean insight from data quickly across various disciplines.**

Methods



Application that given a search string, crawls publication/study metadata and outputs results to an XLSX file with identified ontology matches.

Rancho curators manually check the output to identify "in-scope" datasets, add, clean, and harmonize all metadata. Scripts applied when possible.

Clean, harmonized metadata aligned to data model and ready for additional analysis.

Repositories on DataCrawler

PubMed	bioRxiv
GEO	AE
SRA	EGA
CT.gov	dbGaP
Figshare	SCP
Zendo	ProteomeXchange

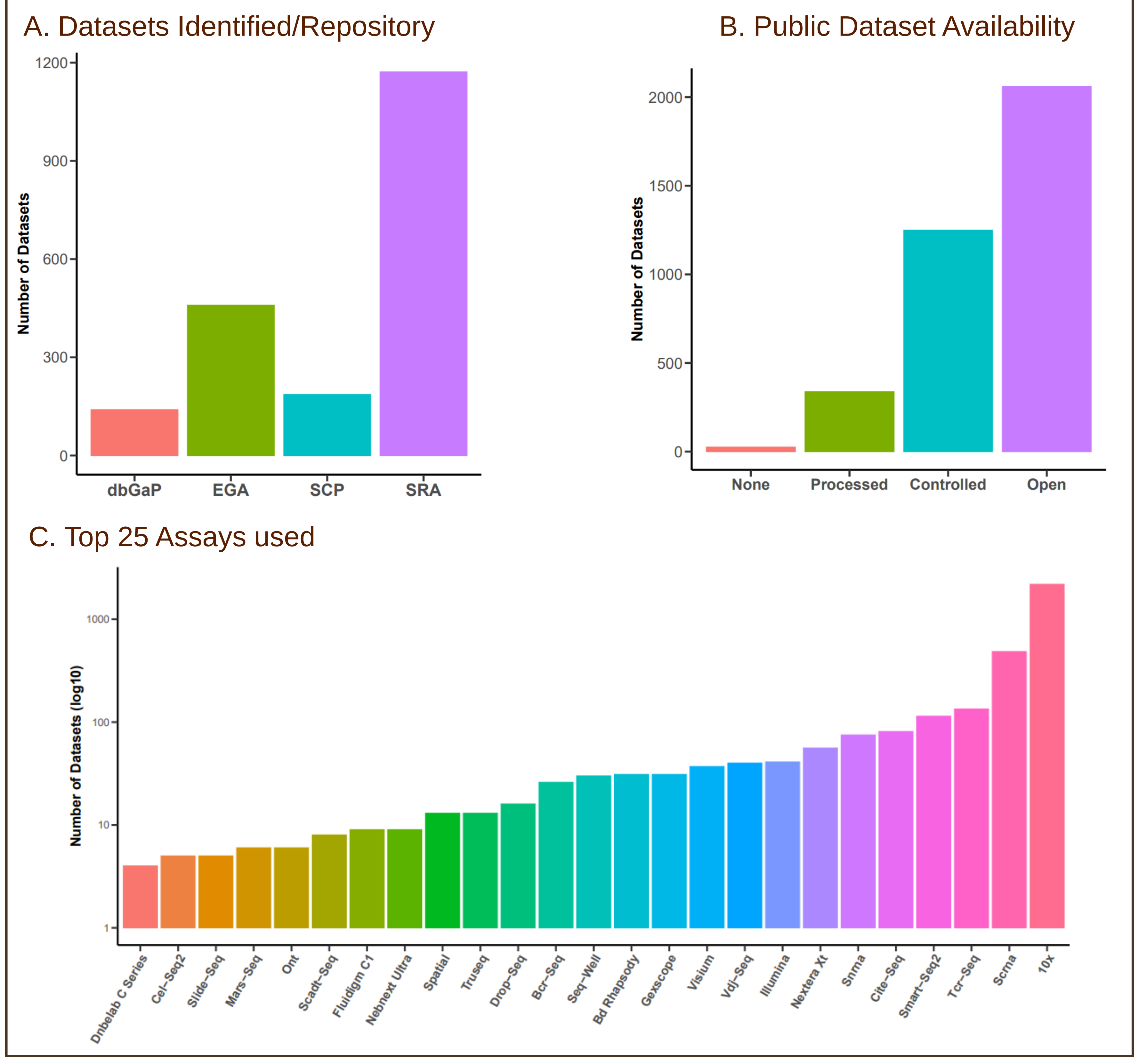
Example Keywords for DataCrawler

Assay	**scRNA-seq; snRNA-seq; methylation by array; proteomics**
Therapeutic area	Parkinson's Disease; breast cancer; oncology; dementia
Organism	Human; mouse; Arabidopsis thaliana; non-human primates

Keywords can be combined to give only the most relevant datasets, such as 'scRNA-seq *AND* human *AND* breast cancer'.

Results

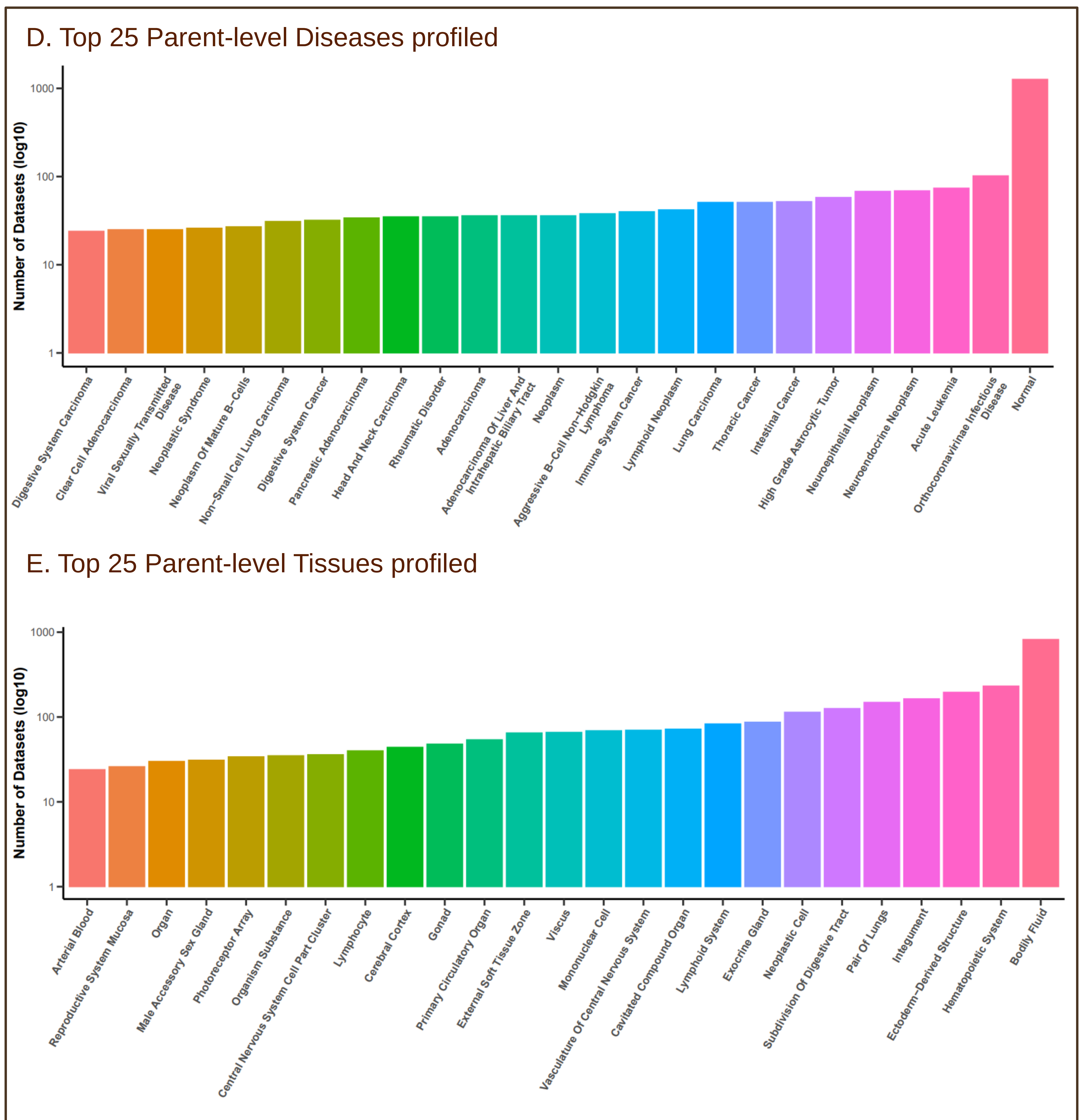
Figure 1


A. Datasets Identified/Repository

B. Public Dataset Availability

C. Top 25 Assays used

Results, Continued


D. Top 25 Parent-level Diseases profiled

E. Top 25 Parent-level Tissues profiled

Figure 1. High level summary of annotated single cell RNA-sequencing data
A. Number of datasets found in each repository B. Number of Datasets where expression data is publicly available ('Open') versus application-based ('Controlled') C. Top 25 assays found across all datasets D. Top 25 parent-level diseases found in all datasets E. Top 25 parent-level tissues found in all datasets

Figure 2

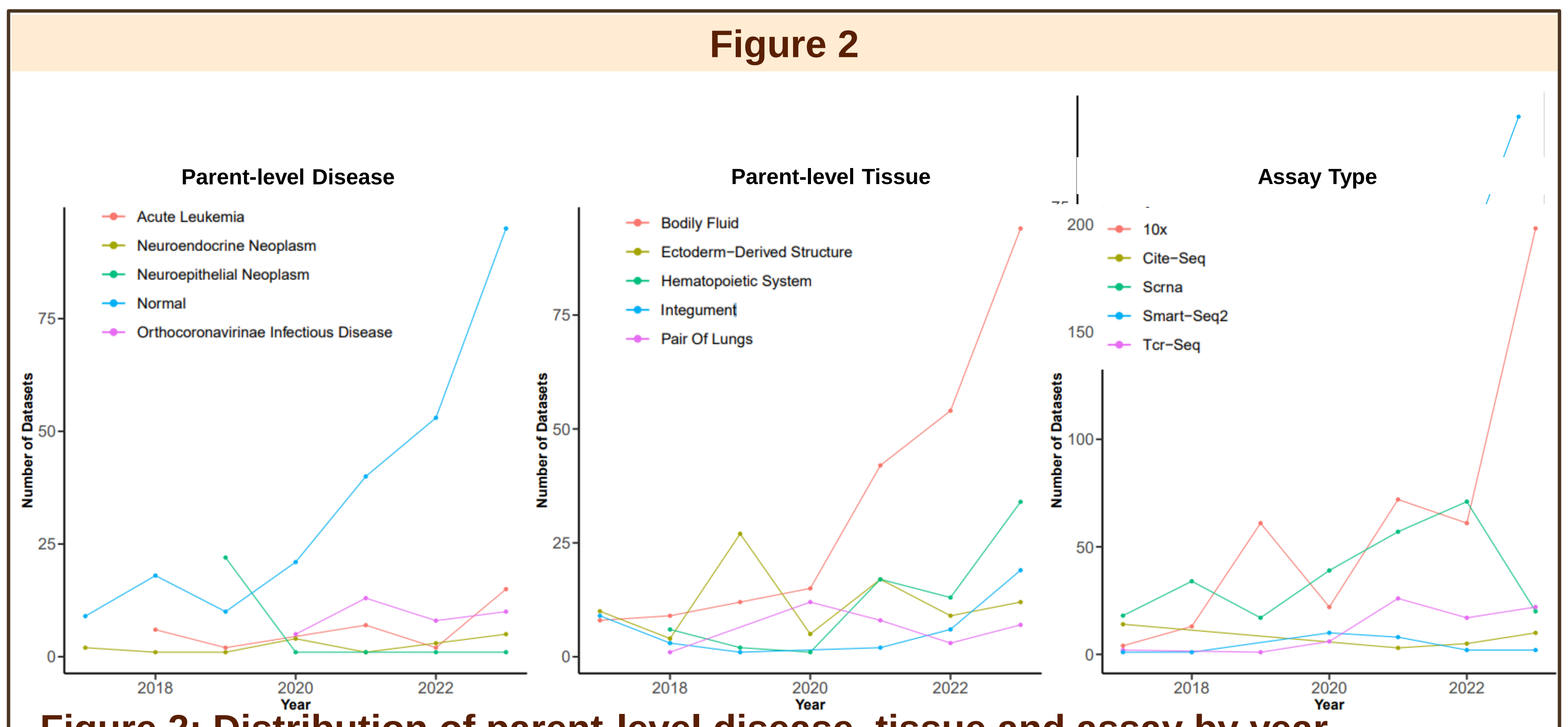


Figure 2: Distribution of parent-level disease, tissue and assay by year

A. Number of datasets for Top 5 parent-level diseases by publication year.
B. Number of datasets for top 5 parent-level tissues by publication year.
C. Number of datasets for top 5 assays or technologies used by publication year.

Conclusions

- More single cell-based datasets are being published each year
- Ideally, multiple repositories should be crawled to identify all possible data
- Healthy, or phenotypically normal, datasets are the most frequently found
- Easily accessible tissue, such as blood and other bodily fluids, are the most profiled tissue
- 10x Genomics-based technology is the most popular assay
- **Rancho BioSciences' DataCrawler is an effective method to identify publicly available datasets across repositories**
- **Visit Rancho BioSciences team at Booth # 525**

Acknowledgements

- A.B., B.F., L.I. and S.K. would like to thank the additional Rancho colleagues who have contributed to this project and poster including Cynthia J. Grondin, Dzmitry Fedarovich, Yuqing He, and Sarah Tahir.
- A.B., B.F., L.I. and S.K. would like to thank all our partners at Cellarity.