Charlotte Whicher
charlotte@repositive.io

An overview of (human) genomic data sources
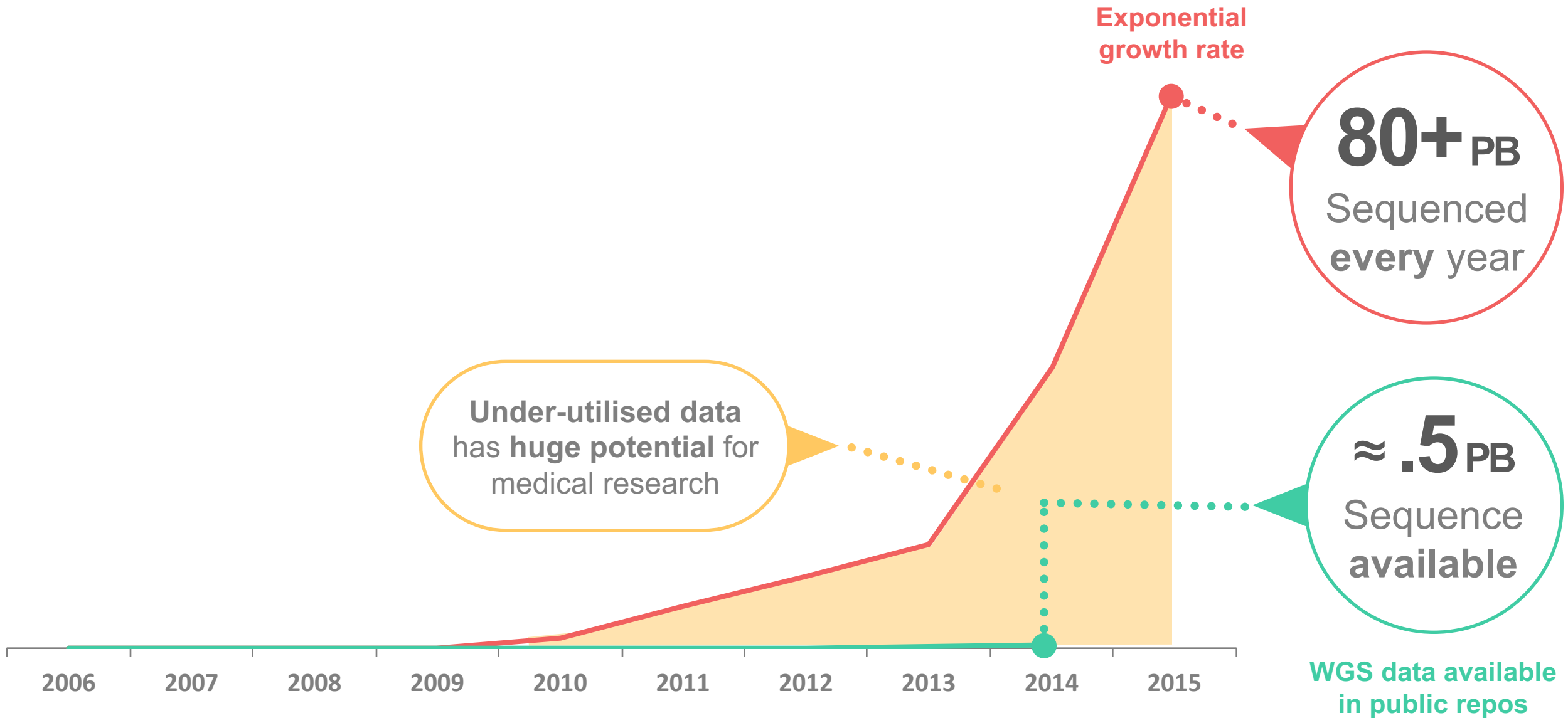
What is genomic data

Genomics is a discipline in genetics.

Genomes are the complete set of DNA within a single cell of an organism.

Genomics applies array, next generation sequencing methods, and bioinformatics to sequence, assemble, and analyse the function and structure of genomes.
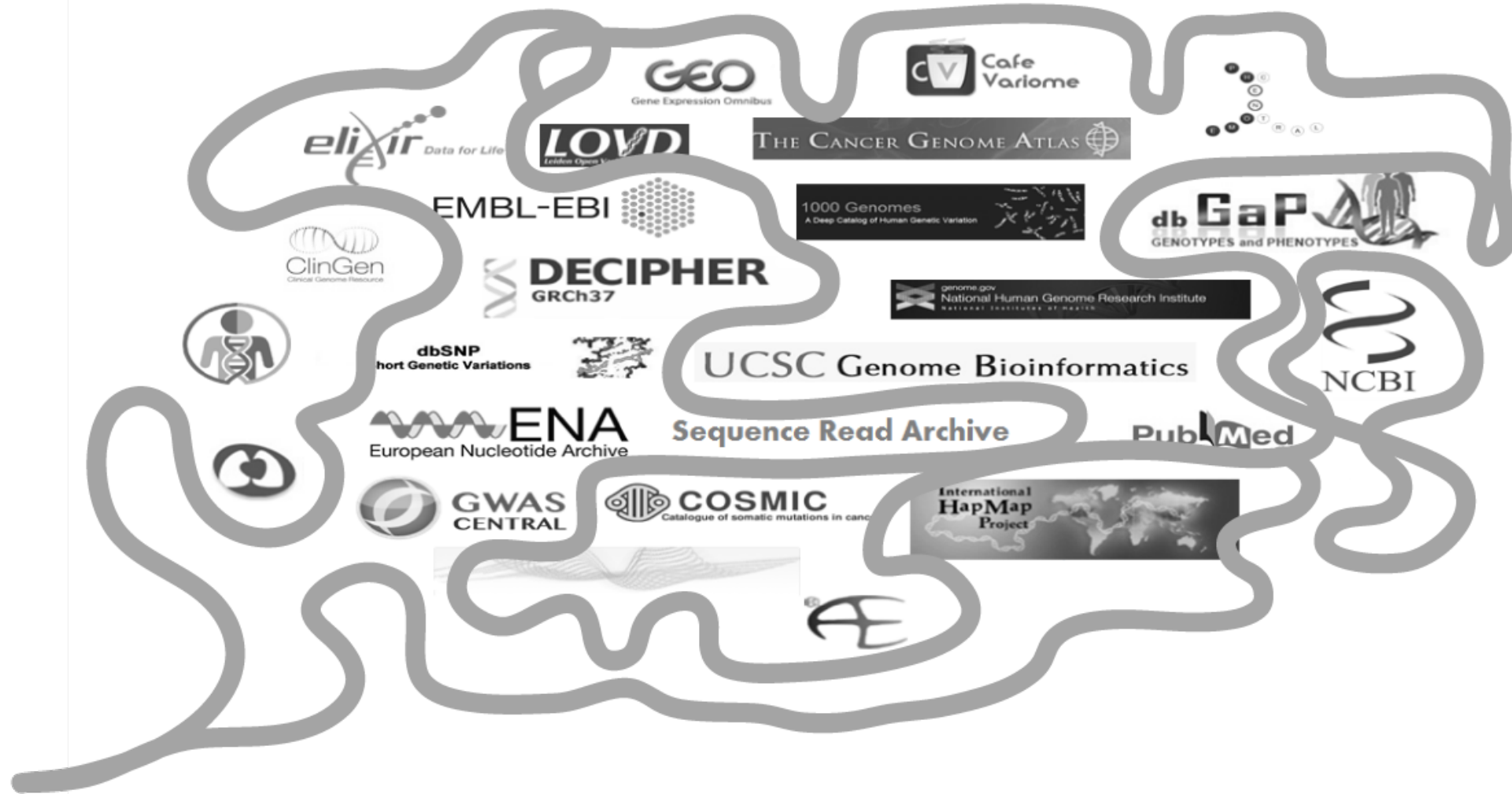
Large amounts of Data, but not accessible

Exponential growth rate
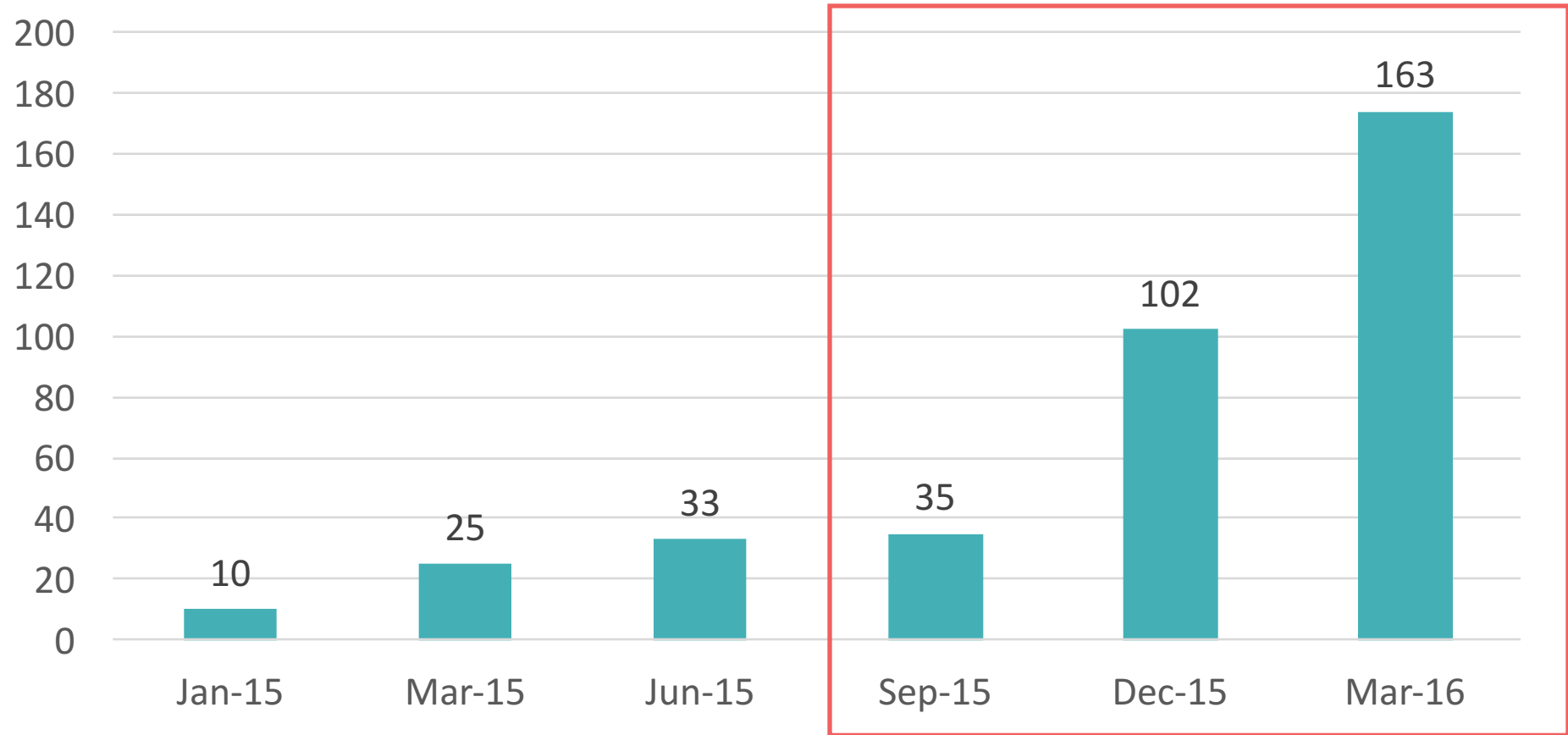
80+ PB
Sequenced **every** year

Under-utilised data has **huge potential** for medical research

≈ **.5** PB
Sequence **available**

2006   2007   2008   2009   2010   2011   2012   2013   2014   2015

WGS data available in public repos

Data is fragmented

It may be confusing

Hundreds of data sources

…but they aren't easy to find!

…at least 6 months work.

Machines & Data sources

5.600 62

947 50

660 25

26 0

68 0

88 3

23 *International*

Machines
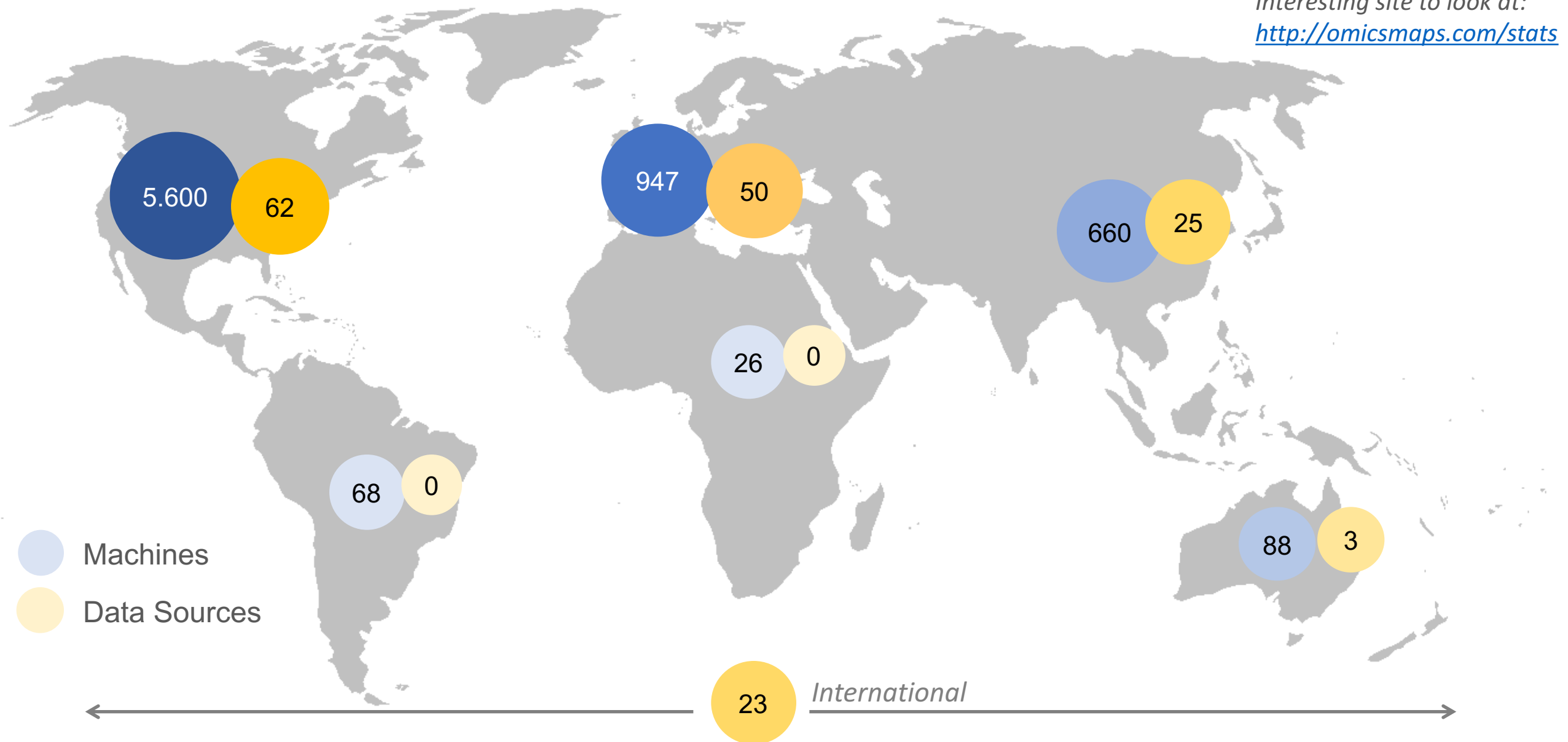
Data Sources

Main Data Sources (Repository) funders

EBI total 104 services, 19 repositories http://www.ebi.ac.uk/services/all

NCBI total 67 databases http://www.ncbi.nlm.nih.gov/guide/all/#databases

Data accessibility

Has both Open and Restricted access data within one repository.

Both 17%

Restricted 38%

Need to apply for access to the data.

■ Restricted

■ Unclear

■ Open

■ Both

Open 43%

Can download the data straight away or after logging in.

Unclear 2%
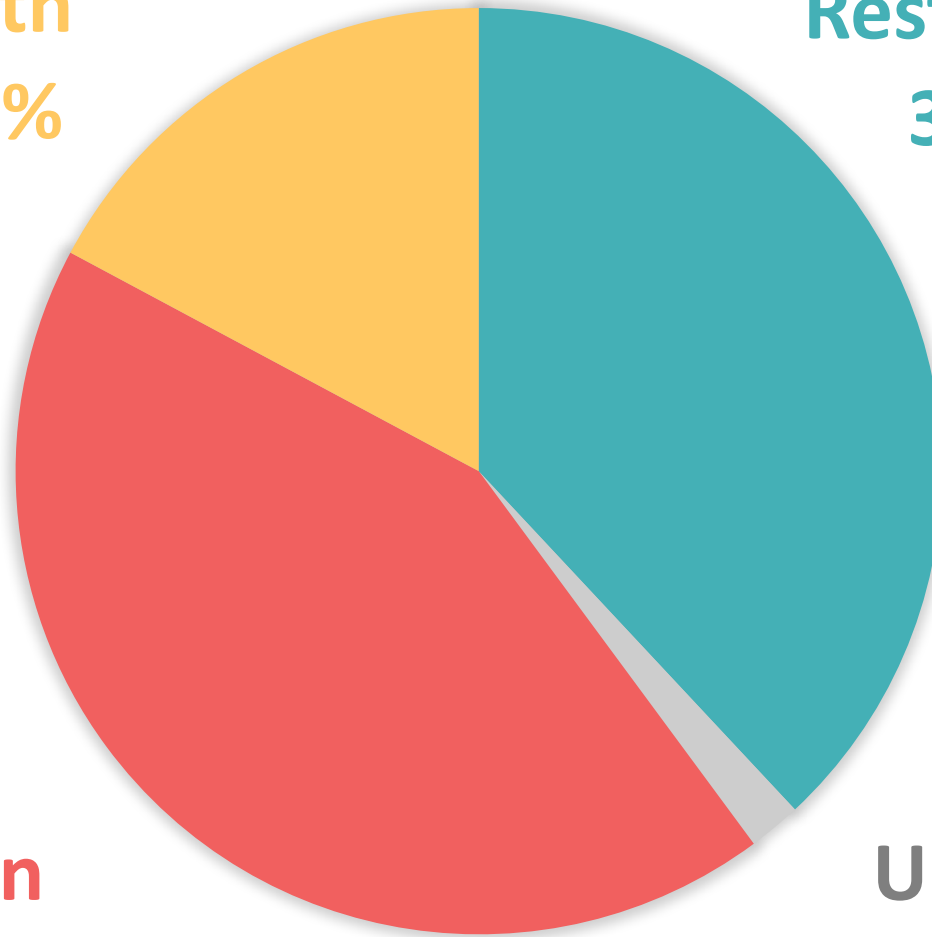
Sequenced ethnicities

Data source content

Dedicated to…
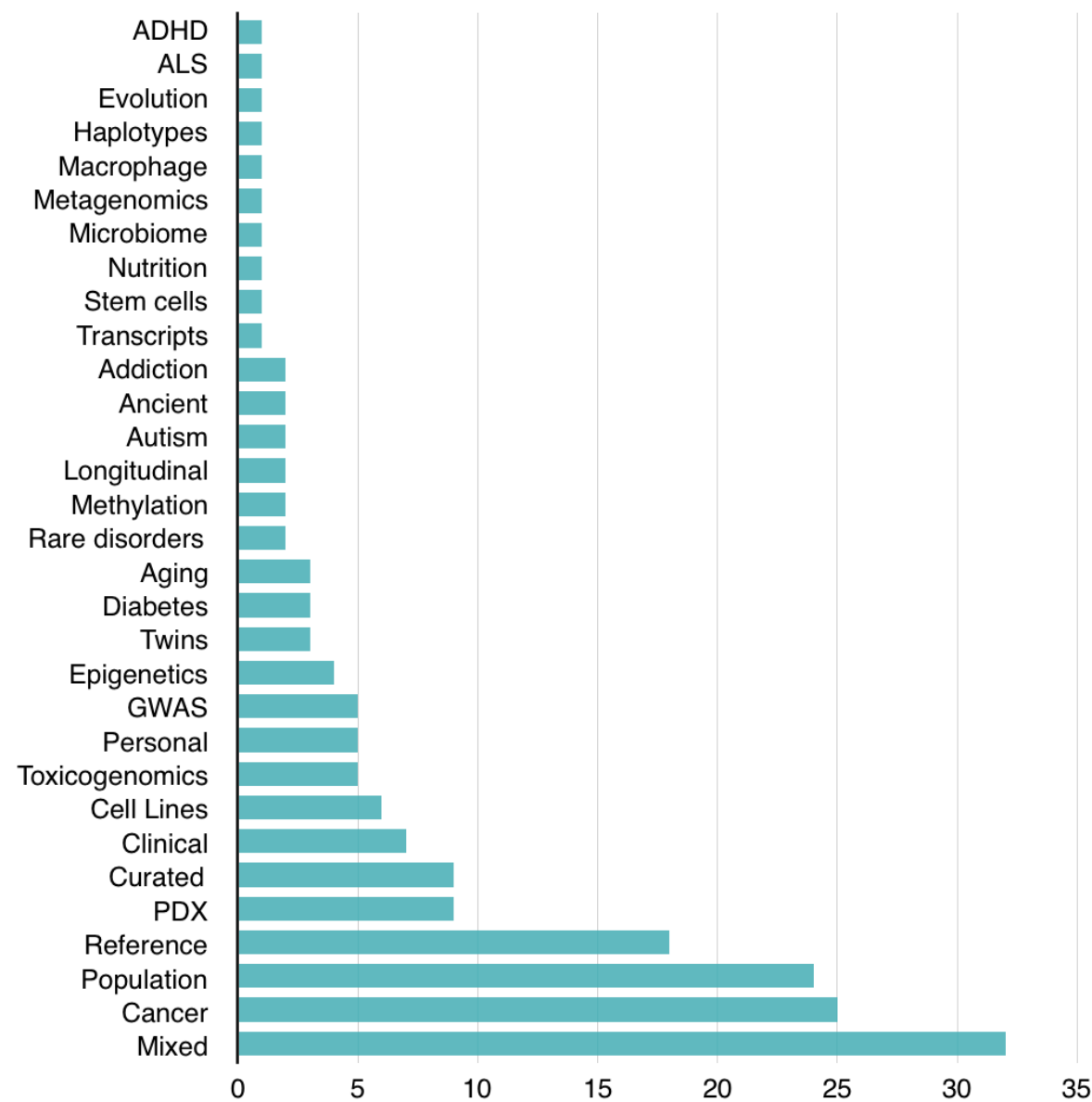
Assay Types
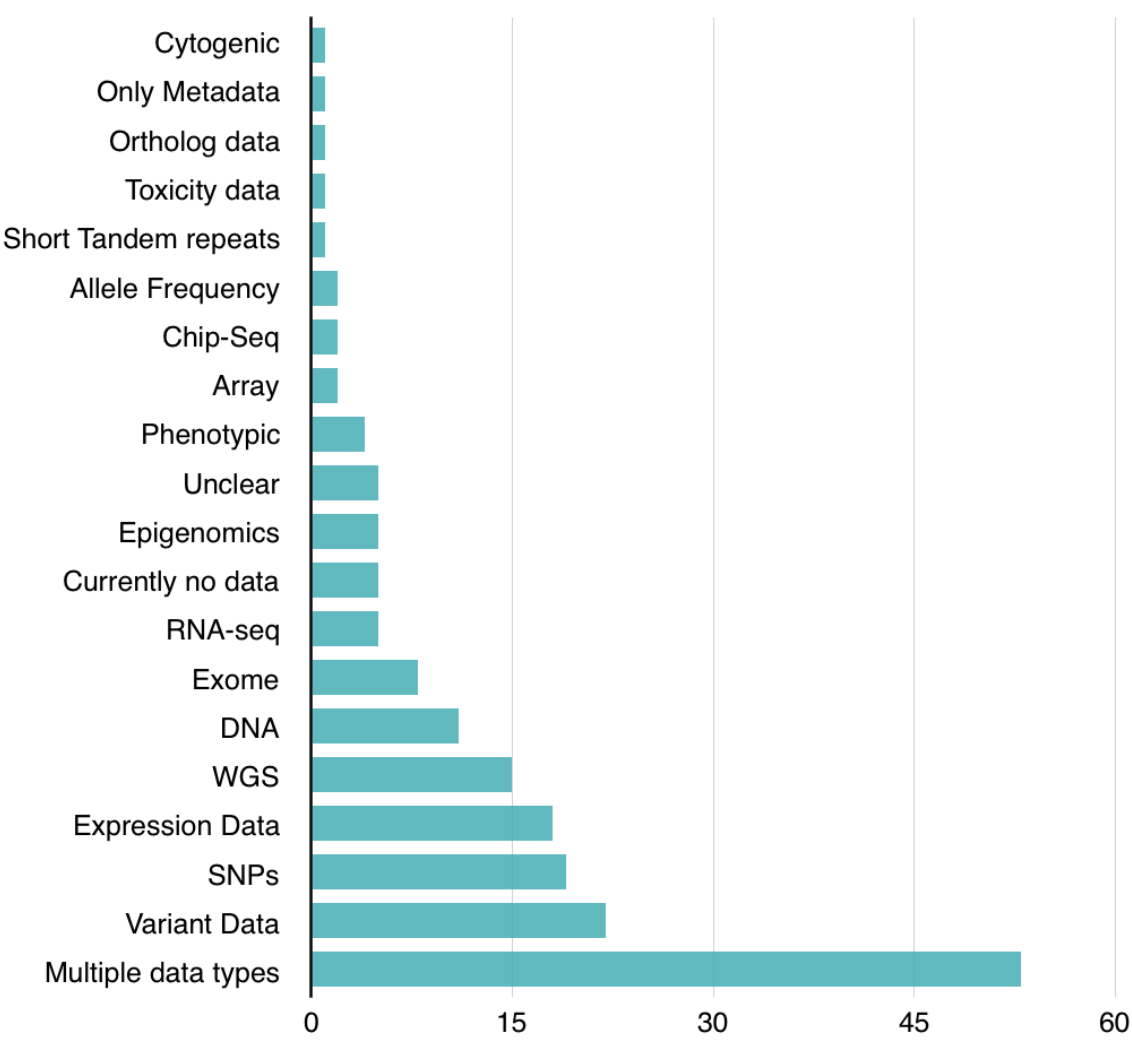
Number of samples in Data sources

Top 5:
GEO (1.8M)
PMI Cohort Program (1M)
Auria Biopankki (1M)
EGA (~0.6M)
SRA (~0.5M)

Sample # (Log10)
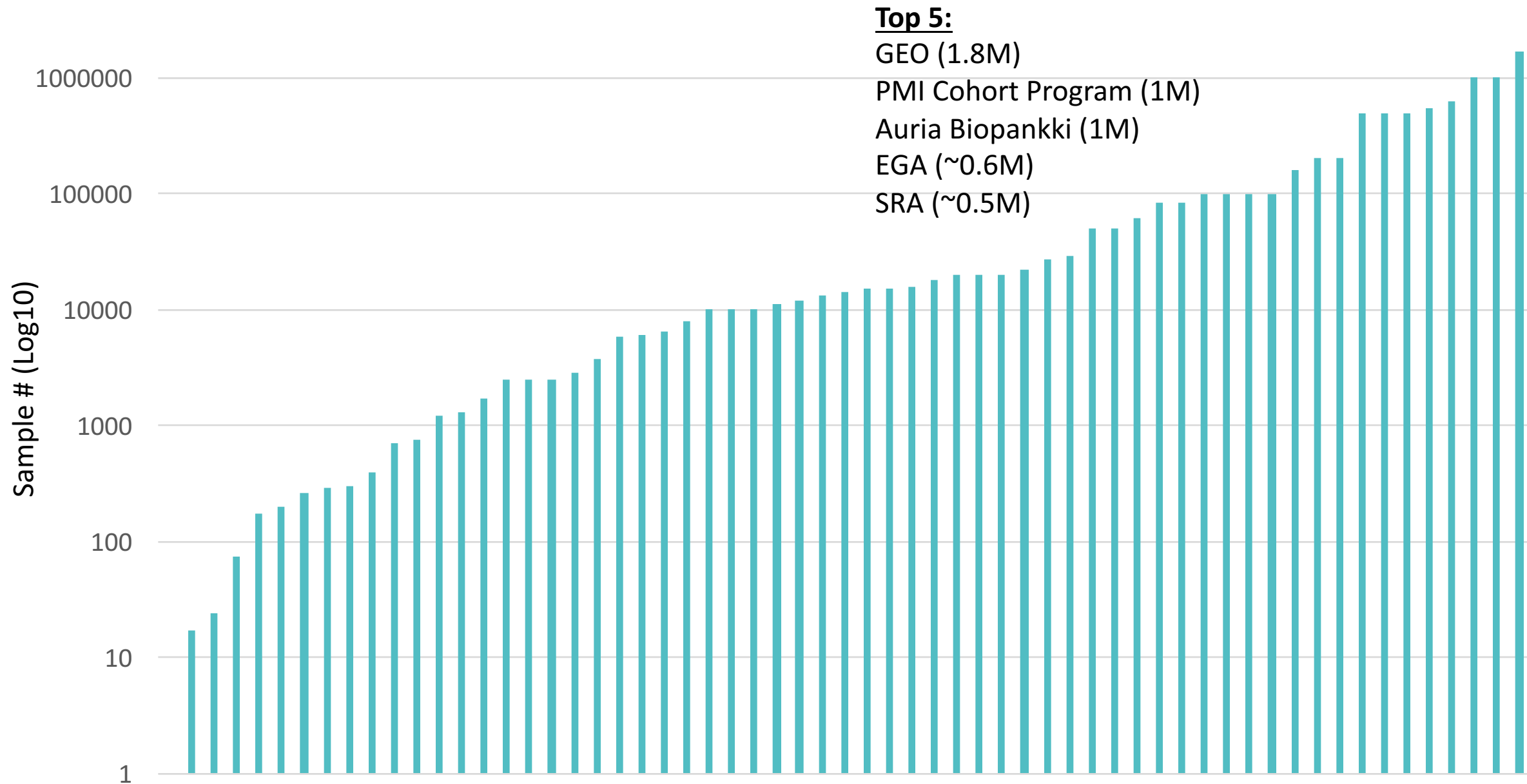
Online Data source 'types'

University – Affiliated to a university. Often only members of that university can upload/download to/from it.

Biobank – many have sequence data of their biological samples.

Catalogue – doesn't have raw data but lists studies/datasets.

Company – For profit organisation. Listing data is not their main purpose.

Repository – Can download from, has data from multiple institutions. Often can also upload your own data there.

Initiative/Consortium – Has a specific purpose/aim. Often focussed on a question or disease.

Access to NCBI data

Access to NCBI data

Access to NCBI data

Access to NCBI data

Access to NCBI data

Access to NCBI data

At Repositive our vision is

1-click to human genomic data access

*to make **finding data** as **easy** as **finding** a **book** on Amazon, booking a **hotel** on Expedia!*

A platform to make data accessible for research

The Repositive platform includes

Search, see **related** results

Discover and access

CWRU Circulating miRNAs in Breast Cancer
from GIGADB
Study investigating the role of miRNAs in breast cancer detection miRNA extracted from plasma on 20 breast cancer patients, 20 cont

public   85   3

Simpler workflow for data access
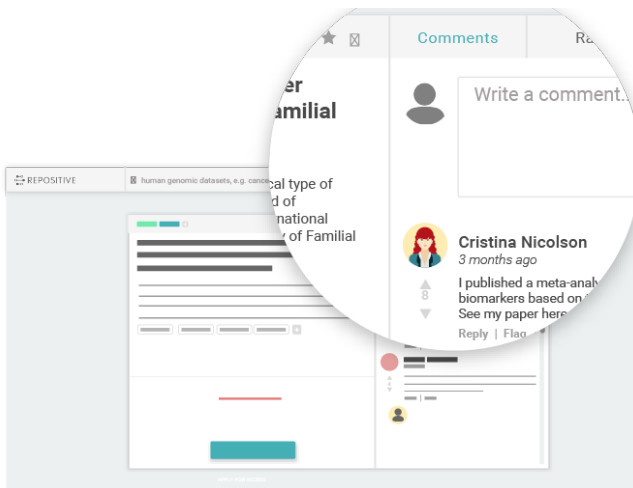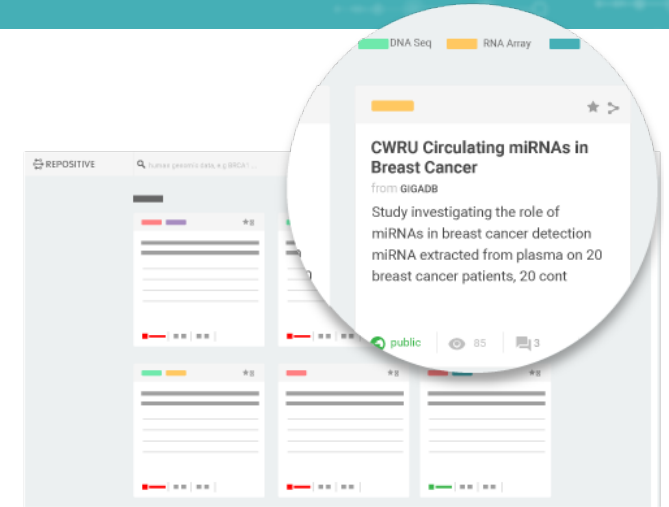
Co-annotate data & community **feedback**

Cristina Nicolson
3 months ago
I published a meta-anal... biomarkers based on... See my paper here...
Reply | Flag

Find **colleagues** & their **data interests**

Dan Fair
Bioinformatics scientist
58

7 Requests   12 Comr

Elena Watson
University of Cambridge