Application Area	Name	Primary Data Size	Source
Biobank	Broad-Pan UKBB QTL dataset (all by all)	40 TB	https://pan.ukbb.broadinstitute.org/docs/hail-format
	LD	40 TB	https://pan.ukbb.broadinstitute.org/docs/hail-format
	TCGA VCF's (+gVCF's)	10 TB	TCGA Data Portal + Client Pipeline
	PPP pQTLs	702 TB	UK Biobank
	deCODE Proteomics summary stats	5 TB	Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease.
	deCODE GWAS	48 TB	https://www.decode.com/ukbsummary/
	gtex QTL	460 GB	https://console.cloud.google.com/storage/browser/gtex-resources/GTEx_Analysis_v8_QTLs/GTEx_Analysis_v8_EUR_eQTL_all_associations
	ClinVar		
	dbNSFP		
SingleCell	Human cell atlas 40+ million cells	4 TB (expression data only)	https://data.humancellatlas.org/explore/projects
Proteomics	Pride	3 TB	

![paradigm4]

Platform Aspect	REVEAL Platform
Immediately available Population Scale Data	Immediately available solution to centralize all single cell data Ease of use with low code Data management layer with schema tailored for fast retrieval and compute based on metadata and scalability to 100s of TBs of data
Costs	Predictable costs Domain expert workflows Extensible: Build your own workflows in Python or R
Integration/ Interoperability	APIs to retrieve data in a format that can be integrated with other data and return data in whatever formats are required by downstream analysis: e.g. data frames in python or R; parquet files, bioinformatic-specific formats. Interoperable with flexFS and AI/ML packages
Governance	Integrated provenance, versioning, traceability, and logs SSO Granular, slice-level access control
Public Data Sets	Publicly available datasets available pre-loaded for simplified validation