

# Data Repositories

The Rockefeller University

Bioinformatics Resource Centre

# Getting hold of HTS data

- From public repositories
- From collaborators
- By sequencing some of your own material!

# Public Repositories

- Several public sources of HTS data exist.
- First concentrating on those acting as repositories.
  - GEO (Gene Expression Omnibus)
  - ENA (European Nucleotide Database)
  - SRA (Short Read Archive)

GEO (<https://www.ncbi.nlm.nih.gov/geo/>)

- GEO holds different types of biological datasets.
- Very popular for submission of data accompanying publication.
- Captures **metadata**, **processed files** and **raw data**.
- GEO was **not built for HTS data**

# GEO - Quick Tour

# SRA ([www.ncbi.nlm.nih.gov/sra](http://www.ncbi.nlm.nih.gov/sra))

- NCBI's HTS specific repository.
- Sequencing specific metadata.
- Stores **Raw data** (in **SRA format**)
- SRA format - requires SRA Toolkit
- Lost then regained funding?

# SRA - Quick Tour

ENA (<https://www.ebi.ac.uk/ena>)

- ENA acts as a european HTS repository.
- Mirrors much of SRA.
- Stores Raw data
- No SRA formats
  - fastq by default.



# ENA - Quick Tour

# Other Repositories

- Many repositories contain processed or unprocessed data.
- These typically are the result of a consortium's data release policies.
- Good example is Encode site.  
(<https://www.encodeproject.org/>)
- UCSC has many useful links to genomics data in various formats.  
(<http://hgdownload.soe.ucsc.edu/downloads.html>)

# Other Repositories

- Other specialist repositories exist.
- **ReCount** database provides standardised counts for user analysis.
- Other databases like Immgen/Bodymap provide RNAseq for specific cells/tissues.

← → ↻ bowtie-bio.sourceforge.net/recount/

Please note that to use the ExpressionSets below, you will need to install **Bioconductor** and run the command `library(Biobase)`

» The Datasets

Study	PMID	Species	Number of biological replicates	Number of uniquely aligned reads	ExpressionSet	Count table	Phenotype table	Notes
bodymap	22496456	human	19	2,197,622,796	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	Illumina Human BodyMap 2.0 -- tissue comparison
cheung	20856902	human	41	834,584,950	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	HapMap - CEU
core	19056941	human	2	8,670,342	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	lung fibroblasts
gillad	20009012	human	6	41,356,738	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	liver; males and females
maq	20167110	human	14 (technical)** 2 (biological)	71,970,164	<a href="#">original pooled</a>	<a href="#">original pooled</a>	<a href="#">original pooled</a>	experiment: MAQC-2
montgomery	20220756	human	60	*886,468,054	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	HapMap - CEU
pickrell	20220758	human	69	*886,468,054	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	HapMap - YRI
sultan	18599741	human	4	6,573,643	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	cell type comparison
wang	18978772	human	22	223,929,919	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	tissue comparison
katz.mouse	21057496	mouse	4	14,368,471	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	control vs. CUG-BP1 knockdown myoblasts
mortazavi	18516045	mouse	3	61,732,881	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	tissue comparison
trapnell	20436464	mouse	4	111,376,152	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	time course
yang	20363980	mouse	1	27,883,862	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	hybrid cell line, X always inactive
bottomly	21455293	mouse	21	343,445,340	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	2 inbred mouse strains
nagalakshmi	18451266	yeast	4	7,688,602	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	priming technique comparison
hammer	20452967	rat	8	158,178,477	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	experimental vs. control at 2 time points
modencodeworm	19181841	worm	46	1,451,119,823	<a href="#">link</a>	<a href="#">link</a>	<a href="#">link</a>	developmental time course
modencodfly	21179090	fly	147 (technical)** 30 (biological)	2,278,788,557	<a href="#">original pooled</a>	<a href="#">original pooled</a>	<a href="#">original pooled</a>	developmental time course

\*Montgomery and Pickrell read counts are for both datasets combined.  
 \*\*These studies originally contained tables with unpooled technical replicates. The unpooled tables are available under the "original" links, while tables with pooled technical replicates are available under the "pooled" links.

# Reference data

- Reference Genome available from many locations.
- Different assemblies
  - Major Revisions - Change locations
  - Minor Revisions - Update annotation
- Genome sequence stored as FASTA.
- Gene build as GFF3 or GTF.
- **IGenomes** contains full annotation files for many genomes.

