## 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)

- 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

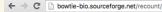
# Other Repositories

- Many repositories contain processed or unprocessed data.
- These typically are the result or 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.



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	link	link	link	Illumina Human BodyMap 2.0 tissue comparison
cheung	20856902	human	41	834,584,950	link	link	link	HapMap - CEU
core	19056941	human	2	8,670,342	link	link	link	lung fibroblasts
gilad	20009012	human	6	41,356,738	link	link	link	liver; males and femlae
maqc	20167110	human	14 (technical)** 2 (biological)	71,970,164	original pooled	original pooled	original pooled	experiment: MAQC-2
montgomery	20220756	human	60	*886,468,054	link	link	link	HapMap - CEU
pickrell	20220758	human	69	*886,468,054	link	link	link	HapMap - YRI
sultan	18599741	human	4	6,573,643	link	link	link	cell type comparison
wang	18978772	human	22	223,929,919	link	link	link	tissue comparison
katz.mouse	21057496	mouse	4	14,368,471	link	link	link	control vs. CUG-BP1 knockdown myoblasts
mortazavi	18516045	mouse	3	61,732,881	link	link	link	tissue comparison
trapnell	20436464	mouse	4	111,376,152	link	link	link	time course
yang	20363980	mouse	1	27,883,862	link	link	link	hybrid cell line, X alwa
bottomly	21455293	mouse	21	343,445,340	link	link	link	2 inbred mouse strains
nagalakshmi	18451266	yeast	4	7,688,602	link	link	link	priming technique comparison
hammer	20452967	rat	8	158,178,477	link	link	link	experimental vs. contra at 2 time points
modencodeworm	19181841	worm	46	1,451,119,823	link	link	link	developmental time course
modencodefly	21179090	fly	147 (technical)** 30 (biological)	2,278,788,557	original pooled	original pooled	original pooled	developmental time course

<sup>\*</sup>Montgomery and Pickrell read counts are for both datasets combined.

<sup>\*\*</sup>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 Revisisons 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.