My experience using public databases (4DN, ENCODE & GEO)

Audrey Lapendry

Club Bioinfo – Thursday 05 May 2022



Public databases = making data accessible to the scientific community

- Pros
- Consultation, recovery and exploitation of data (often) for free
- Cons
- Sometimes difficult to use it effectively
- Variable data quality













4DN Data Portal = platform to search, visualize and download nucleomics data

<u>Objective</u>: "understand the principles behind the 3D organization of the nucleus and the role of nuclear organization that plays in gene expression and cellular function" Funding: National Institutes of Health





All the data present in the portal has been analyzed in an automated way, with the same procedure by technology

Example of a workflow for Hi-C data analysis:





The datasets can be easily downloaded in various format

Example of files types for Hi-C data analysis:

Raw datas are also availables.



Steps to download files:

- Connect to 4DN with a Google or GitHub account (free)
- Select the datasets of interest (and the file type wanted)

- Then, It automatically creates a file with the datasets and metadata (descriptions of how the data were acquired). The curl command for the download is also given.



The data visualization tool HiGlass is integrate in the 4DN portal



Availables genomes: GRCh38 GRCm38 dm6



ENCODE = public research consortium that has produced a lot of data which have been made available

<u>Objective:</u> "build a comprehensive parts list of functional elements in the genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active"

Funding: National Human Genome Research Institute

ENCODE: Encyclopedia of DNA Elements

TF ChIP-seq	4075	DNAme array	251	microRNA counts	114	
Histone ChIP-seq	3482	eCLIP	225	Repli-seq	104	
DNase-seq	1574	small RNA-seq	212	RNA Bind-n-Seq	102	
Mint-ChIP-seq	943	WGBS	211	RRBS	96	
total RNA-seq	781	long read RNA-seq	197	Hi-C	72	<u>Etc.</u>
polyA plus RNA-seq	741	RAMPAGE	155	Repli-chip	59	
ATAC-seq	415	ChIA-PET	141	PAS-seq	40	
microRNA-seq	369	RNA microarray	128	BruChase-seq	32	
scRNA-seq	355	genotyping array	121	RNA-PET	31	
snATAC-seq	302	CAGE	117	polyA minus RNA-seq	31	

All the data present in ENCODE has been analyzed in an automated way, with the same procedure by technology

Example of a workflow for RNA-seq data analysis:





The datasets can be easily downloaded in various format

Example of file types for RNA-seq data analysis:

Raw datas are also availables.

	plus strand
bigWig	signal of
	unique reads
tev	gene
LSV	quantifications
	minus strand
bigWig	signal of
	unique reads
bam	alignments
tou	transcript
LSV	quantifications
	plus strand
bigWig	signal of all
	Teaus
tsv	transcript
1.34	quantifications
	minus strand
bigWig	signal of all
	reads
bam	transcriptome
	alignments

Steps to download files:

- Select the datasets of interest (and the file type wanted)
- Then, It automatically creates a file with the datasets. The curl command for the download is also given.
- Use the API REST of ENCODE to get the metadata

The datasets are automated audits according to the quality of the data and the completion of metadata

Audit category: A externely low read depth 70 corror externely low read depth 77 inconstraint replicate memory low read length 7 inconstraint replicate 1 Audit category: A	ChIP-seq of H1-hESC Homo sapiens H1-hESC Target: H3K4me3 Lab: Project: Badge count Experiment ENCSR443YAS released A 1 B 1 5
insufficient read depth 1418 control insufficient read depth 838 missing controlled by 582 insufficient read length 471 unreplicated experiment 322 poor library complexity 211	Extremely low read depth @
control low read depth 202 severe bottlenecking 199 partially characterized antibody 91	
missing possible_controls 82 missing RNA fragment size 66	Inconsistent platforms Badge
missing documents 52 uncharacterized antibody 20	Low read length @ Descriptions
antibody not characterized to 17 standard missing spikeins 14 insufficient coverane 2	Low read depth
maang input control 2 -See forwer Audit category. Not to waad length 2730 mid to moarthe bottneschart 2934 low waad depth 1608 modernike Insay complexity 1074 inconsister plantoma 945 inconsister plantoma 945	Alignment file /files/ENCFF340UJK/ produced by ChIP-seq read mapping pipeline (/pipelines/ENCPL220NBH/) using the GRCh38 assembly has 12360072 usable fragments. The minimum ENCODE standard for each replicate in a ChIP-seq experiment targeting H3K4me3-human and investigated as a narrow histone mark is 10 million usable fragments. The recommended value is > 20 million is accentrable. (See (data-standards/chip-seq/) Alignment file /files/ENCFF2852JI/ produced by ChIP-seq read mapping pipeline (/pipelines/ENCPL220NBH/) using the hg19 assembly has 12369033 usable fragments. The minimum ENCODE standard for each replicate in a ChIP-seq exempted targeting H3(fms), but > 10 million is accentrable. (See (data-standards/chip-seq/)
antibody characterized with 570 exemption	recommended value is > 20 million, but > 10 million is acceptable. (See /data-standards/chip-seq/)
mixed read rendmis 459 borderline replicate concordance357 missing external identifiers 269 inconsistent target of control 241	Moderate library complexity
experiment mixed run types 169 inconsistent control run_type 131 missing spikeins 110	Mild to moderate bottlenecking
Crit	ical Issue Mild Issue Mild Issue
	(Davis and al. 2017)



ENCODE integrates a data visualization tool for some types of data







GEO = public data repository

<u>Objective</u>: "provide a public archive to store massive volumes of published **high-throughput functional** genomic data generated by the international research community"

Funding: National Center for Biotechnology Information

GEO: Gene Expression Omnibus

Series type	Count		
Expression profiling by array	65,141	Genome binding/occupancy profiling by array	235
Expression profiling by genome tiling array	753	Genome binding/occupancy profiling by genome tiling array	2,373
Expression profiling by high throughput sequencing	63,048	Genome binding/occupancy profiling by high throughput sequencing	27,125
Expression profiling by SAGE	239	Genome binding/occupancy profiling by SNP array	18
Expression profiling by MPSS	20	Methylation profiling by array	1,315
Expression profiling by RT-PCR	874	Methylation profiling by genome tiling array	2,092
Expression profiling by SNP array	14	Methylation profiling by high throughput sequencing	4,116
Genome variation profiling by array	855	Methylation profiling by SNP array	17
Genome variation profiling by genome tiling array	1,540	Protein profiling by protein array	352
Genome variation profiling by high throughput sequencing	271	Protein profiling by Mass Spec	9
Genome variation profiling by SNP array	1,472	SNP genotyping by SNP array	854



Essential nomenclature of GEO

- <u>GEO</u> = Gene Expression Omnibus, a public data repository
- <u>GSE</u> = identifier associated to a dataset, often corresponding to the data produced during a publication
- <u>GSM</u> = experiences that are part of a GSE
- <u>SRA</u> = high throughput sequencing data storage, <u>format</u>: .sra = compressed format of FASTQ files and <u>SRX</u> identifier

Steps to download files:

- Raw data → Select the identifier of the datasets of interest and download it with the SRA toolkit, e.g. sratoolkit/bin/fastq-dump SRR260219 (it also convert in FASTQ)
- Analysed data → Select the datasets of interest, go directly on the GSE or GSM pages to obtain the curl command to run

Non-homogeneity of the data submitted, each team analyses data in a different way

Extracted molecule total RNA

Extraction protocol Total RNA was extracte using NucleoSpin® RNA ((Macherey-Nagel)

 $3 \mu g$ of total RNA from each sample were subjected to reverse transcription with random primers. The 5-end cap structure was biotinylated and captured with streptavidin-coated magnetic beads (Thermo Fisher). After ligation of 5' and 3' adaptors, second-strand cDNA was synthesized, followed by exonuclease I (New England BioLabs) digestion. The purified CAGE libraries were sequenced using single-end reads of 50 bp on the Illumina HiSeq 2500 (Illumina, USA).

Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina HiSeq 2500

Data processing rDNA and low quality read filtering: MOIRAI pipeline (Hasegawa et al. 2014) genome alignment: STAR(Ver 2.5.3a_modified) Count CAGE defined transcriptional start sites (CTSS): overlapping with FANTOM5 robust promoter set Normalization: Tag Per Million Genome_build: hg19 Supplementary_files_format_and_content: CTSS Supplementary files format and content: TPM expression table text file

Club bioinfo GEO Deposit

By Jean-Baptiste Claude



Comparisons of 4DN, ENCODE & GEO

	Data analysed in a homogeneous way	Facility to find data and metadata	Most represented type of data
4DN	Yes	Yes	Hi-C
	Yes	Yes	TF ChIP-seq
Gene Expression Omnibus	No	It depends of the project	RNA-seq and microarray

<u>Alternatives</u>: European Nucleotide Archive or ArrayExpress from European Bioinformatics Institute and DNA Database of Japan