

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

Audrey Lapendry

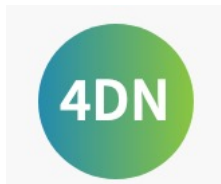
Club Bioinfo – Thursday 05 May 2022



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

Sequences databases:





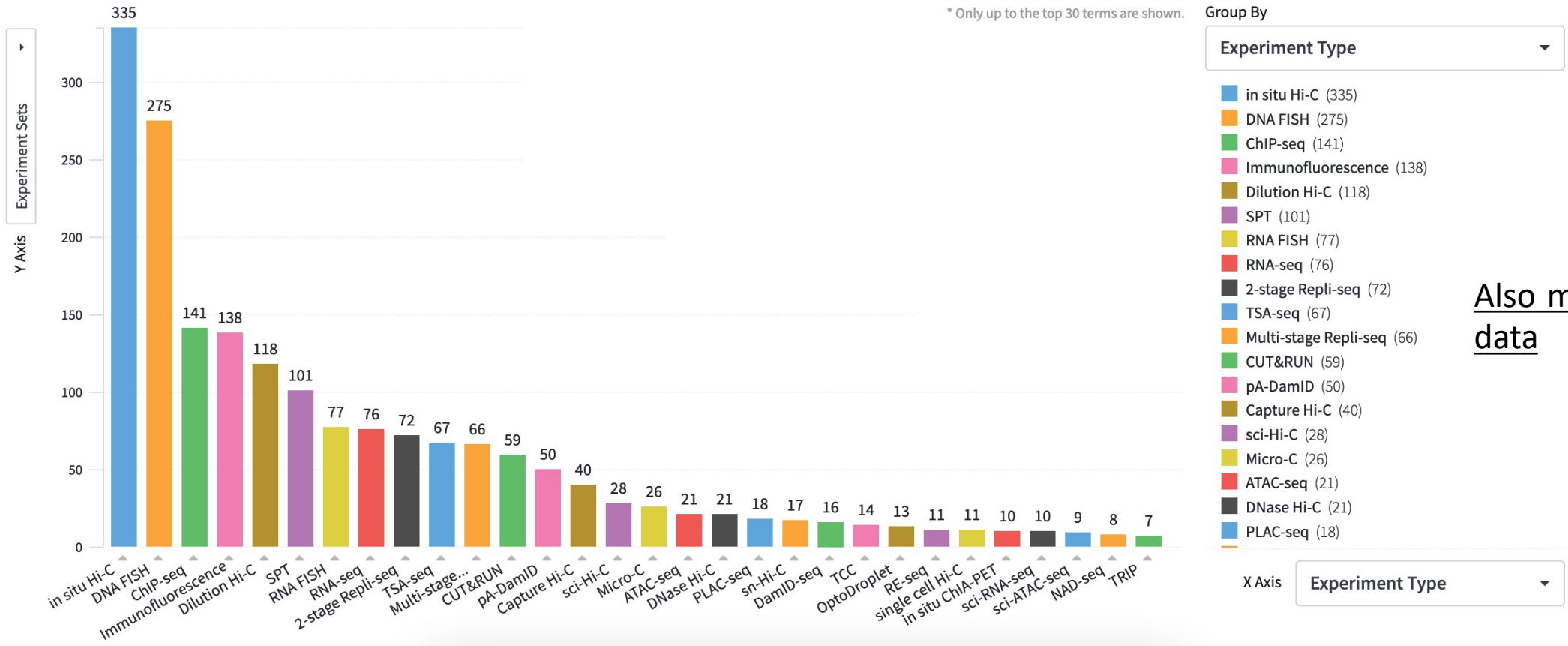
4DN Data Portal



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

Objective: “understand the principles behind the 3D organization of the nucleus and the role of nuclear organization that plays in gene expression and cellular function”

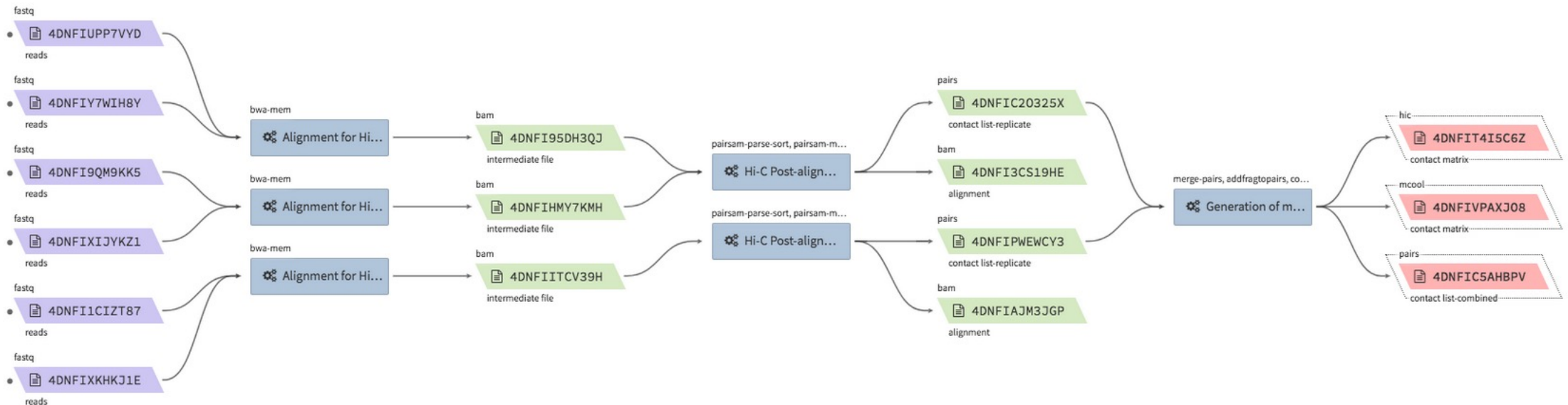
Funding: National Institutes of Health

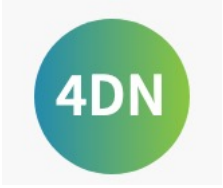




All the data present in the portal has been analyzed in an automated way, with the same procedure by technology

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:

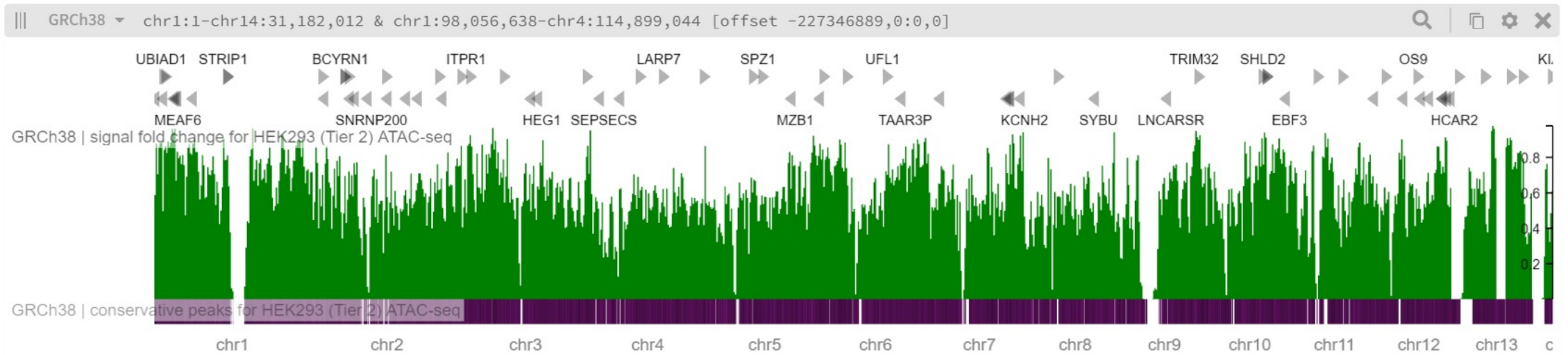
Raw datas are also availables.

File Type
contact list-combined (pairs)
contact matrix (hic)
contact matrix (mcool)
boundaries (bed)
insulation score-diamond (bw)
compartments (bw)
alignments (bam)
contact list-replicate (pairs)
alignments (bam)
contact list-replicate (pairs)

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



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

Objective: "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

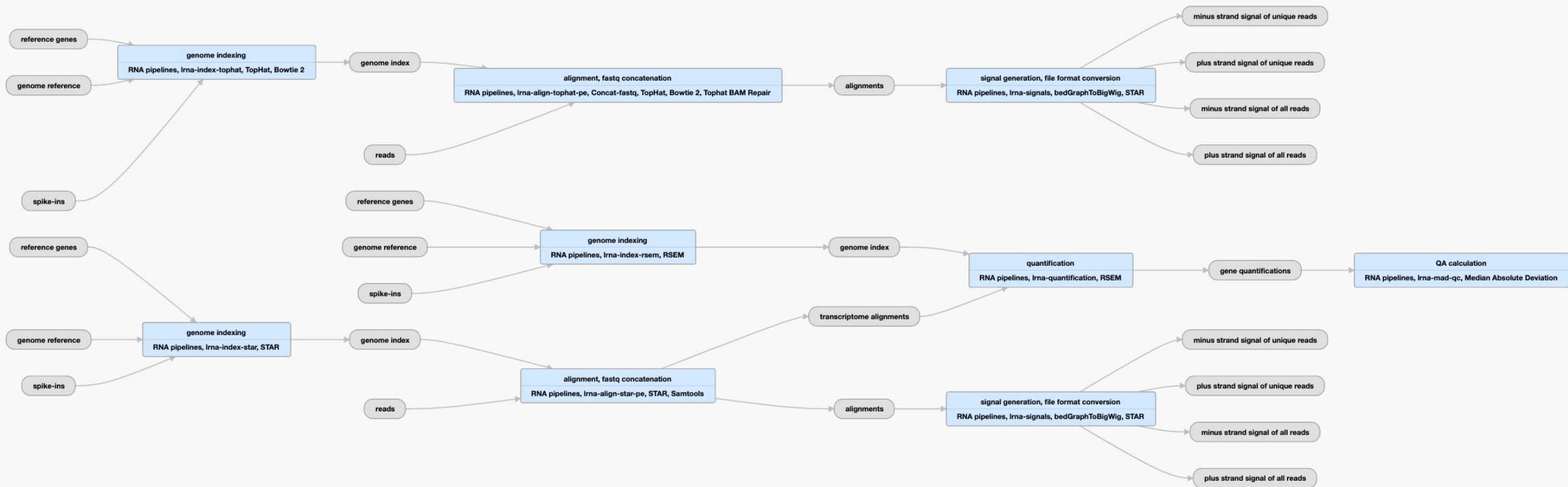
ENCODE: Encyclopedia of DNA Elements

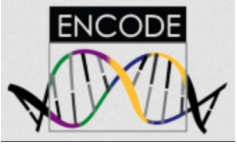
TF ChIP-seq	4075	DNAme array	251	microRNA counts	114
Histone ChIP-seq	3482	eCLIP	225	Repli-seq	104
DNase-seq	1574	small RNA-seq	212	RNA Bind-n-Seq	102
Mint-ChIP-seq	943	WGBS	211	RRBS	96
total RNA-seq	781	long read RNA-seq	197	Hi-C	72
polyA plus RNA-seq	741	RAMPAGE	155	Repli-chip	59
ATAC-seq	415	ChIA-PET	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
					<u>Etc.</u>



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.

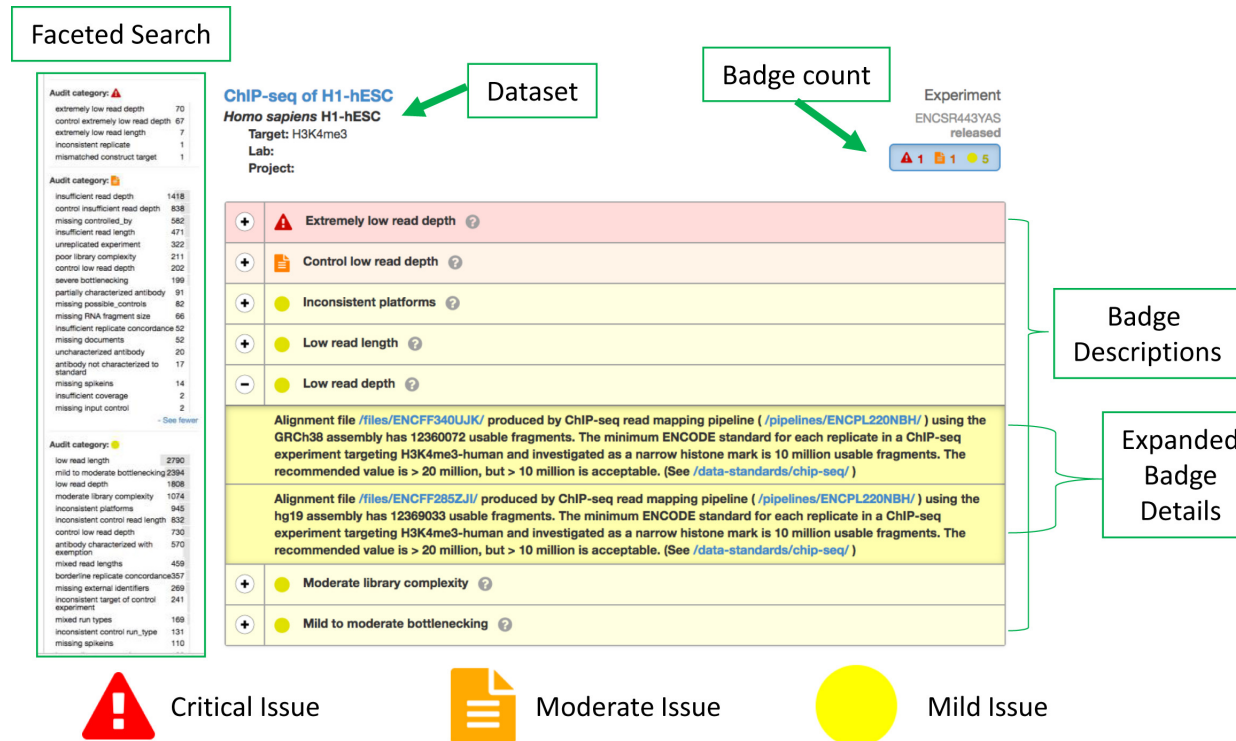
bigWig	plus strand signal of unique reads
tsv	gene quantifications
bigWig	minus strand signal of unique reads
bam	alignments
tsv	transcript quantifications
bigWig	plus strand signal of all reads
tsv	transcript quantifications
bigWig	minus strand signal of all reads
bam	transcriptome alignments

Steps to download files:

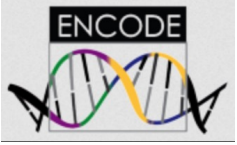
- Select the datasets of interest (and the file type wanted)
- Then, It automatically creates a file with the datasets. The curl command for the download is also given.
- Use the API REST of ENCODE to get the metadata



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



(Davis and al. 2017)



ENCODE integrates a data visualization tool for some types of data





GEO



GEO = public data repository

Objective: “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

Etc.



Essential nomenclature of GEO

- GEO = Gene Expression Omnibus, a public data repository
- GSE = identifier associated to a dataset, often corresponding to the data produced during a publication
- GSM = experiences that are part of a GSE
- SRA = high throughput sequencing data storage, format: .sra = compressed format of FASTQ files and SRX 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 extracted using NucleoSpin® RNA ((Macherey-Nagel)

3 µg 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

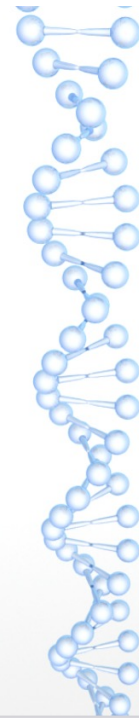
Genome_build: hg19

Supplementary_files_format_and_content: CTSS

Supplementary_files_format_and_content: TPM expression table text file

Club bioinfo GEO Deposit

By Jean-Baptiste Claude






GEO Deposit



Club Bioinfo
07 Novembre 2019
Jean-Baptiste Claude

1

Comparisons of 4DN, ENCODE & GEO

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

Alternatives: European Nucleotide Archive or ArrayExpress from European Bioinformatics Institute and DNA Database of Japan