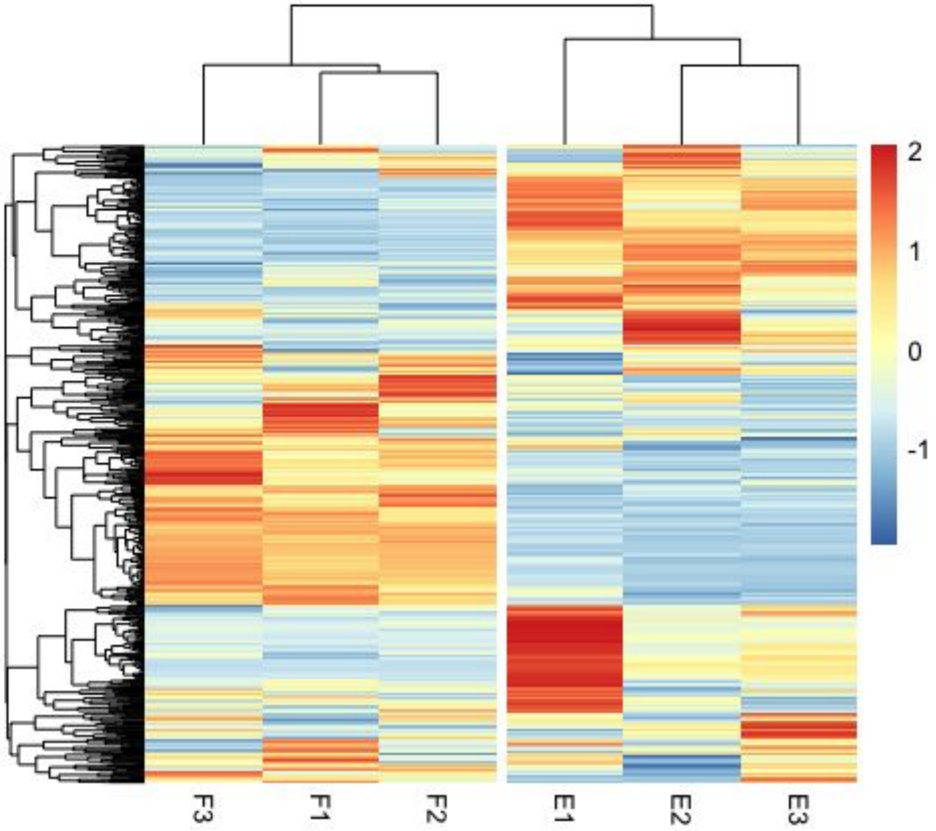
Introduction to Gene Expression

Part I

Anil Upreti

August 4,2021



Introduction

Course Objective

1. Give biologist the basic understanding of RNA sequencing
2. Give biologist the basic ability to analyze the RNA sequencing data (Focused on model organism mice)
3. Identification of the differentially expressed genes
4. Future directions for usage of RNAseq data analysis

Introduction

This course is …

1. Formal introduction to RNA sequencing analysis

This course is not …

1. Not a advance coding course
2. Not for downstream applications of RNA sequencing

What you do need to success ?

- Basic R, microsoft excel usage
- Basic understating of central dogma in Biology
- Practice, Practice, Practice

What you don’t need to succeed?

- Deep coding knowledge
- Deep R usage
- Deep understanding of sequencing methods

Gene Expression

Wednesday

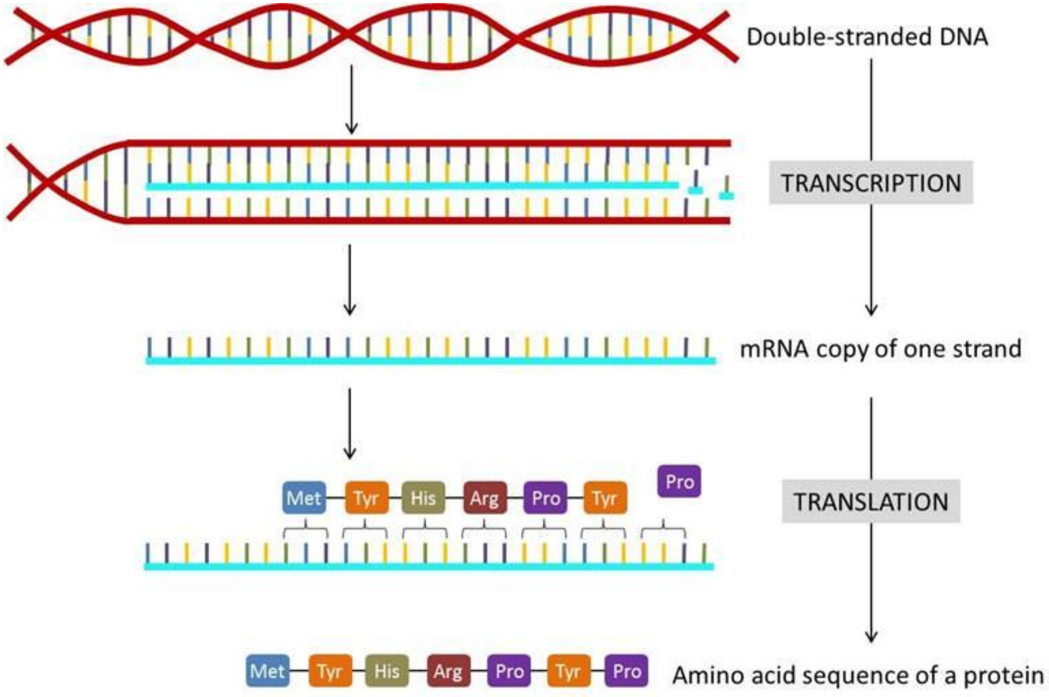
- Central dogma and sequencing
- Downloading RAW FASTA file from GEO
- Downloading reference transcriptome from GENECODE
- R basics (installation of DESeq2, pheatmap etc)

Thursday

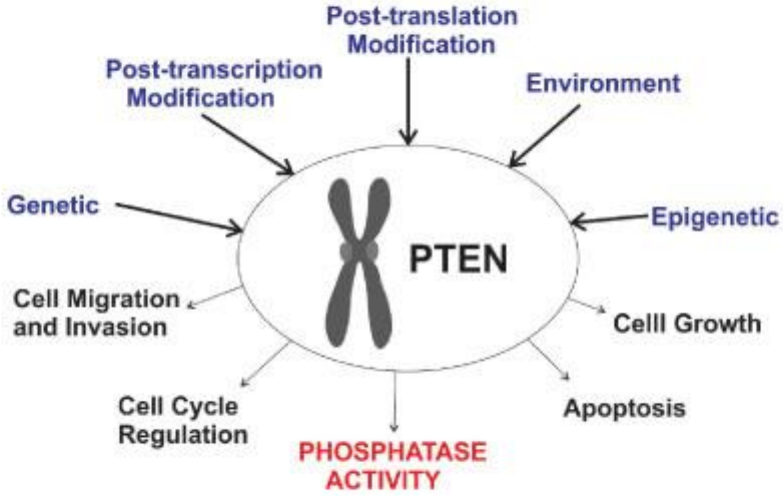
- Importing the count files in R
- Finding DEGs
- Finding gene of interest
- Data visualization
- Gene Ontology

Central dogma

Gene expression

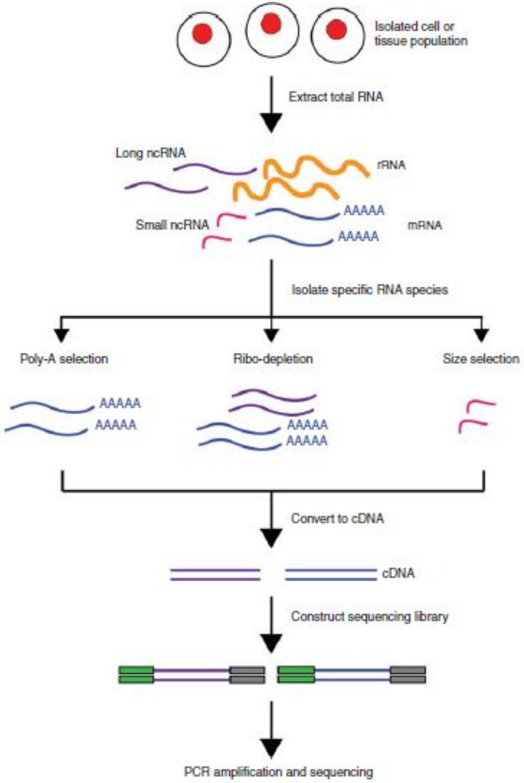


https://rwu.pressbooks.pub/bio103/chapter/the-central-dogma-genes-to-traits/



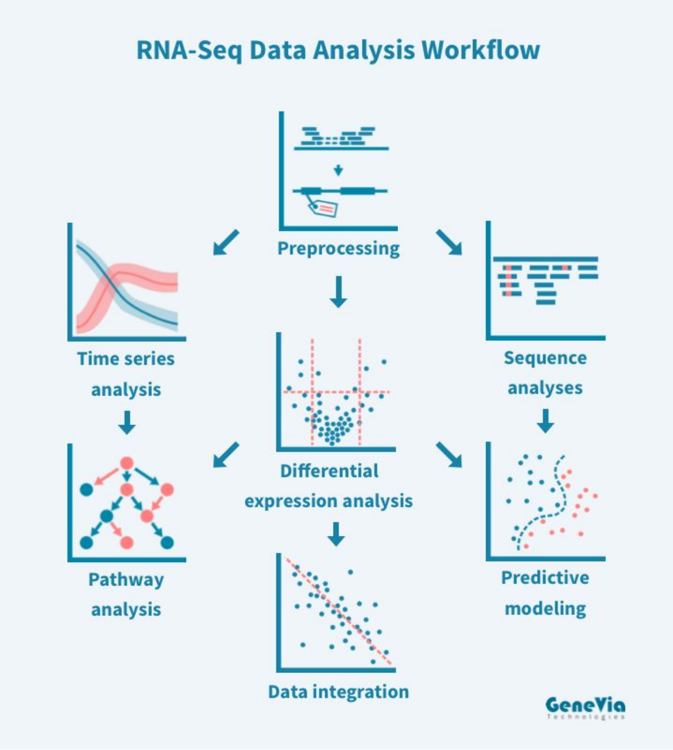
Silva KSF (2019) Genetic Counseling, Polymorphisms and Breast Cancer. J Fam Med Dis Prev 5:098. doi.org/10.23937/2469-5793/1510098

Overview of sequencing



https://www.rna-seqblog.com/introduction-to-rna-sequencing-and-analysis/

Analysis

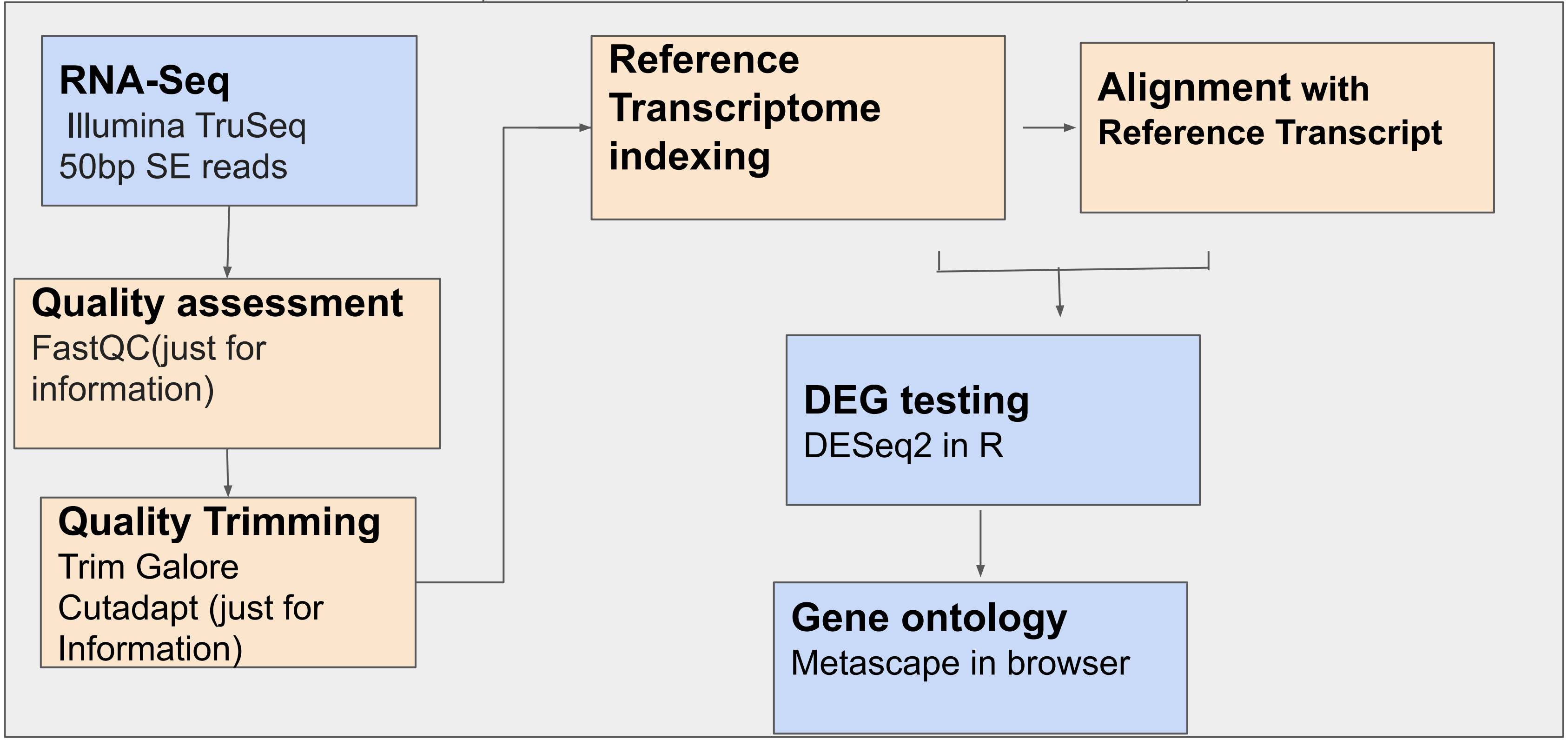


https://geneviatechnologies.com/bioinformatics-analyses/rna-seq-data-analysis/

Data processing pipeline

Genome: **GRCm38 from genecode**

Annotation: **Ensembl**



RNA-Seq
Illumina TruSeq
50bp SE reads

Quality assessment
FastQC(just for information)

Quality Trimming
Trim Galore
Cutadapt (just for Information)

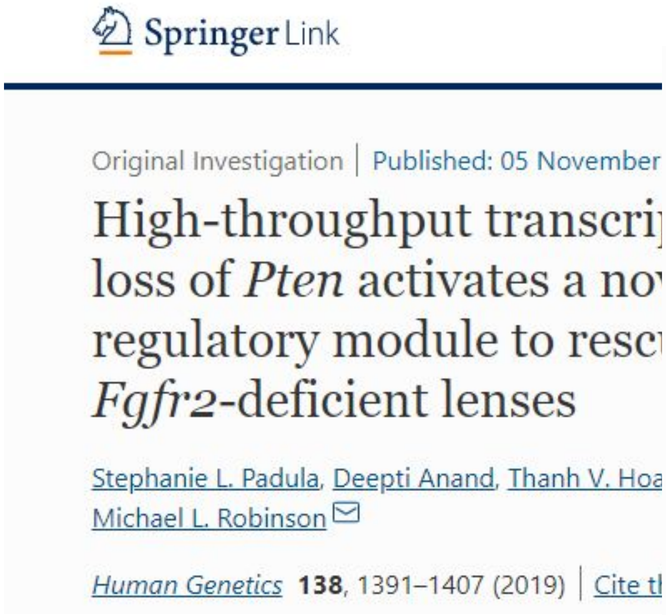
Reference Transcriptome indexing

Alignment with **Reference Transcript**

DEG testing
DESeq2 in R

Gene ontology
Metascape in browser

Finding the data of our interest...



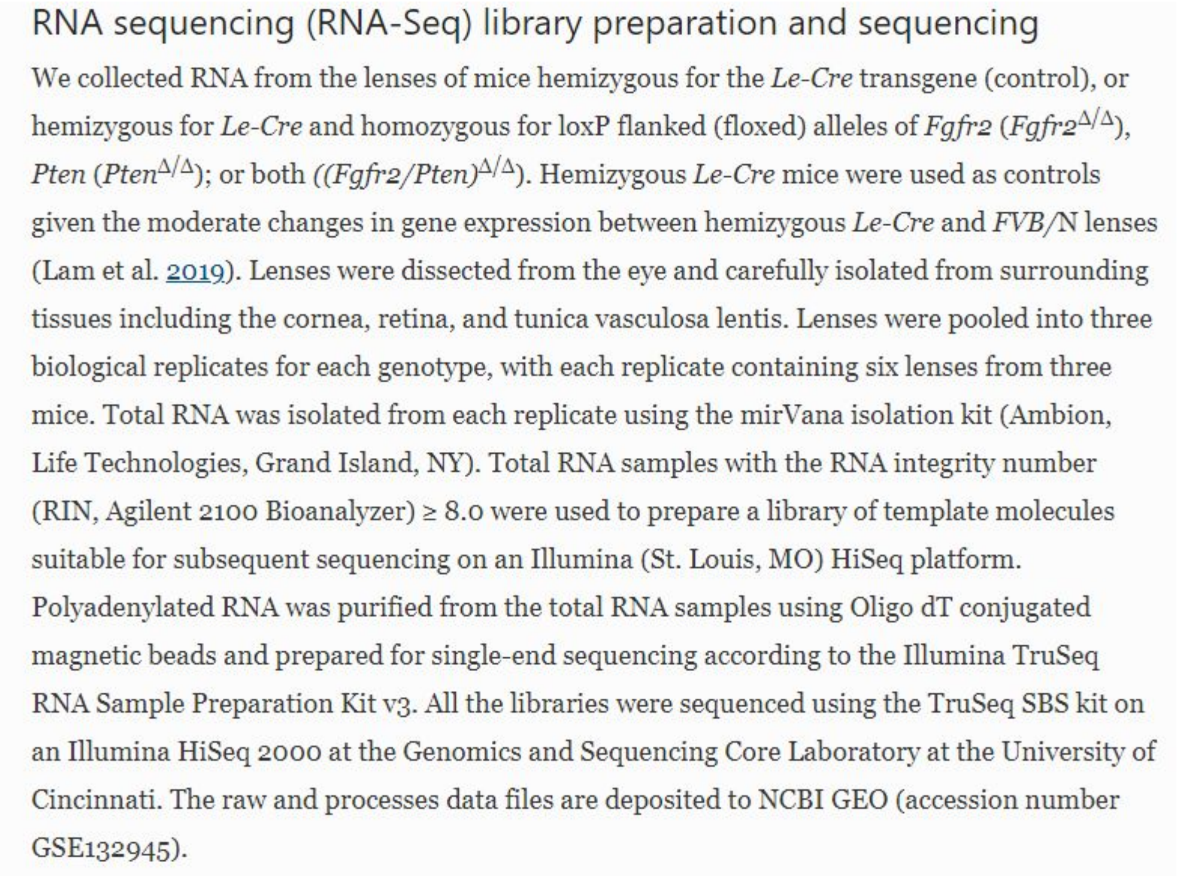
Springer Link

Original Investigation | Published: 05 November

High-throughput transcri loss of *Pten* activates a no regulatory module to resc *Fgfr2*-deficient lenses

Stephanie L. Padula, Deepti Anand, Thanh V. Hoa Michael L. Robinson ✉

Human Genetics **138**, 1391–1407 (2019) | Cite th

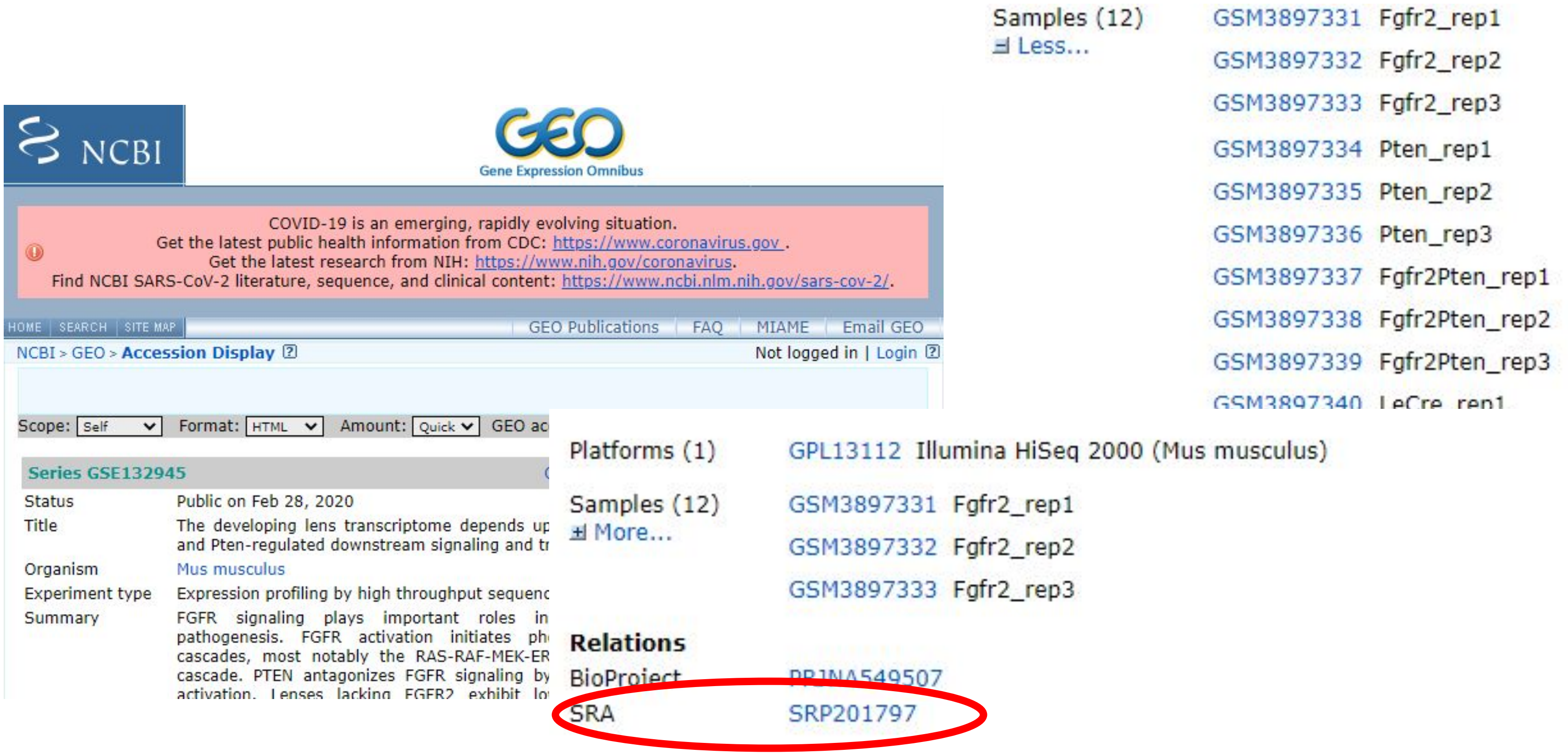


RNA sequencing (RNA-Seq) library preparation and sequencing

We collected RNA from the lenses of mice hemizygous for the *Le-Cre* transgene (control), or hemizygous for *Le-Cre* and homozygous for loxP flanked (floxed) alleles of *Fgfr2* ($Fgfr2^{\Delta/\Delta}$), *Pten* ($Pten^{\Delta/\Delta}$); or both (($Fgfr2/Pten)^{\Delta/\Delta}$). Hemizygous *Le-Cre* mice were used as controls given the moderate changes in gene expression between hemizygous *Le-Cre* and *FVB/N* lenses (Lam et al. 2019). Lenses were dissected from the eye and carefully isolated from surrounding tissues including the cornea, retina, and tunica vasculosa lentis. Lenses were pooled into three biological replicates for each genotype, with each replicate containing six lenses from three mice. Total RNA was isolated from each replicate using the mirVana isolation kit (Ambion, Life Technologies, Grand Island, NY). Total RNA samples with the RNA integrity number (RIN, Agilent 2100 Bioanalyzer) ≥ 8.0 were used to prepare a library of template molecules suitable for subsequent sequencing on an Illumina (St. Louis, MO) HiSeq platform. Polyadenylated RNA was purified from the total RNA samples using Oligo dT conjugated magnetic beads and prepared for single-end sequencing according to the Illumina TruSeq RNA Sample Preparation Kit v3. All the libraries were sequenced using the TruSeq SBS kit on an Illumina HiSeq 2000 at the Genomics and Sequencing Core Laboratory at the University of Cincinnati. The raw and processes data files are deposited to NCBI GEO (accession number GSE132945).

How to access the data?

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132945



COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

HOME | SEARCH | SITE MAP
GEO Publications | FAQ | MIAME | Email GEO
NCBI > GEO > Accession Display
Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO ac

Series GSE132945

Status	Public on Feb 28, 2020
Title	The developing lens transcriptome depends up and Pten-regulated downstream signaling and tr
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequenc
Summary	FGFR signaling plays important roles in pathogenesis. FGFR activation initiates ph cascades, most notably the RAS-RAF-MEK-ER cascade. PTEN antagonizes FGFR signaling by activation. Lenses lacking FGFR2 exhibit lo

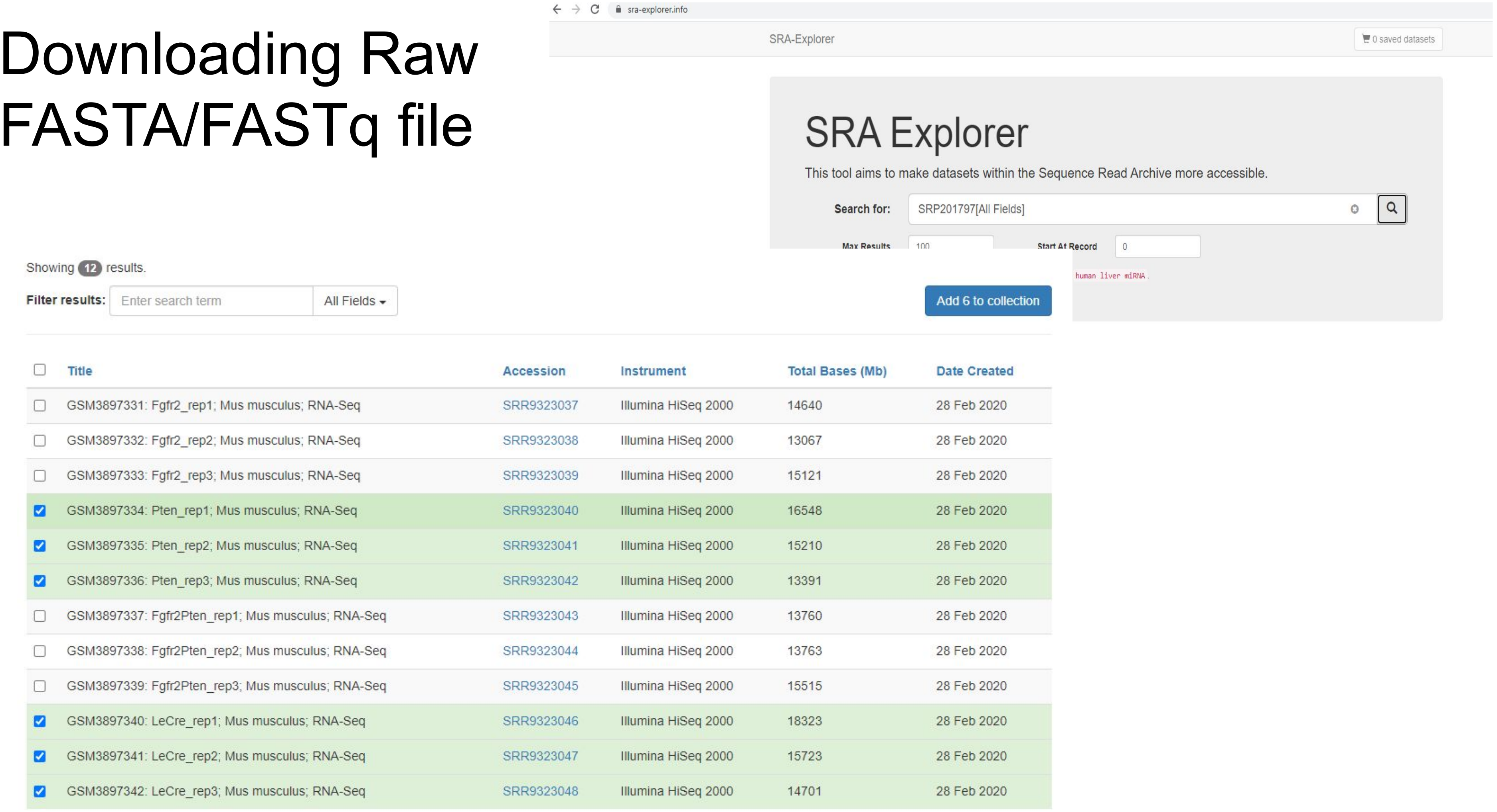
Platforms (1)	GPL13112	Illumina HiSeq 2000 (Mus musculus)
Samples (12) ⊞ More...	GSM3897331	Fgfr2_rep1
	GSM3897332	Fgfr2_rep2
	GSM3897333	Fgfr2_rep3

Relations

BioProject	PRJNA549507
SRA	SRP201797

Samples (12) ⊟ Less...	GSM3897331	Fgfr2_rep1
	GSM3897332	Fgfr2_rep2
	GSM3897333	Fgfr2_rep3
	GSM3897334	Pten_rep1
	GSM3897335	Pten_rep2
	GSM3897336	Pten_rep3
	GSM3897337	Fgfr2Pten_rep1
	GSM3897338	Fgfr2Pten_rep2
	GSM3897339	Fgfr2Pten_rep3
	GSM3897340	LeCre_rep1

Downloading Raw FASTA/FASTq file



SRA Explorer

This tool aims to make datasets within the Sequence Read Archive more accessible.

Search for: SRP201797[All Fields]

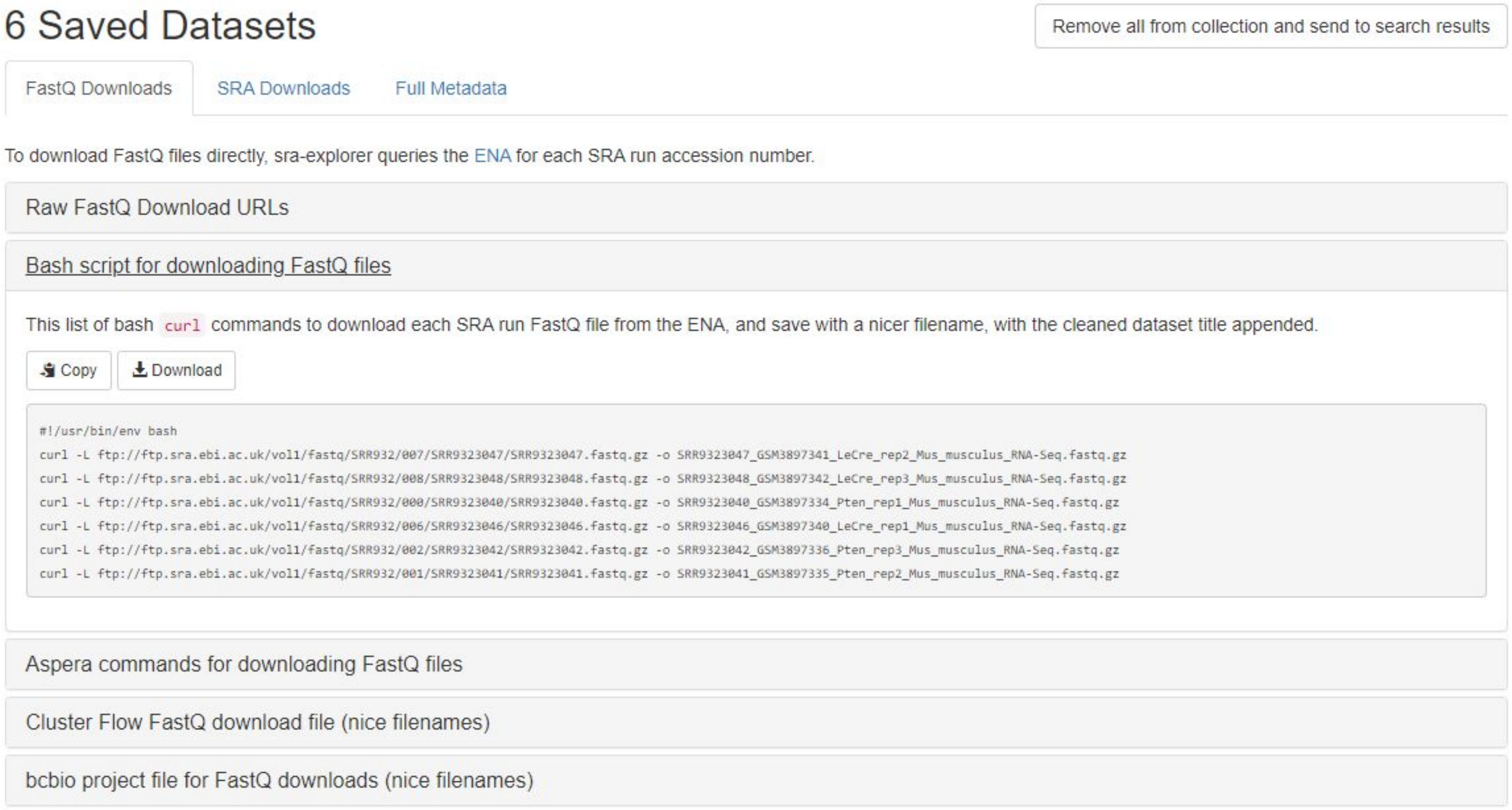
Max Results 100 **Start At Record** 0

human liver miRNA .

Showing 12 results.

Filter results: Enter search term | All Fields ▾

☐	Title	Accession	Instrument	Total Bases (Mb)	Date Created
☐	GSM3897331: Fgfr2_rep1; Mus musculus; RNA-Seq	SRR9323037	Illumina HiSeq 2000	14640	28 Feb 2020
☐	GSM3897332: Fgfr2_rep2; Mus musculus; RNA-Seq	SRR9323038	Illumina HiSeq 2000	13067	28 Feb 2020
☐	GSM3897333: Fgfr2_rep3; Mus musculus; RNA-Seq	SRR9323039	Illumina HiSeq 2000	15121	28 Feb 2020
☑	GSM3897334: Pten_rep1; Mus musculus; RNA-Seq	SRR9323040	Illumina HiSeq 2000	16548	28 Feb 2020
☑	GSM3897335: Pten_rep2; Mus musculus; RNA-Seq	SRR9323041	Illumina HiSeq 2000	15210	28 Feb 2020
☑	GSM3897336: Pten_rep3; Mus musculus; RNA-Seq	SRR9323042	Illumina HiSeq 2000	13391	28 Feb 2020
☐	GSM3897337: Fgfr2Pten_rep1; Mus musculus; RNA-Seq	SRR9323043	Illumina HiSeq 2000	13760	28 Feb 2020
☐	GSM3897338: Fgfr2Pten_rep2; Mus musculus; RNA-Seq	SRR9323044	Illumina HiSeq 2000	13763	28 Feb 2020
☐	GSM3897339: Fgfr2Pten_rep3; Mus musculus; RNA-Seq	SRR9323045	Illumina HiSeq 2000	15515	28 Feb 2020
☑	GSM3897340: LeCre_rep1; Mus musculus; RNA-Seq	SRR9323046	Illumina HiSeq 2000	18323	28 Feb 2020
☑	GSM3897341: LeCre_rep2; Mus musculus; RNA-Seq	SRR9323047	Illumina HiSeq 2000	15723	28 Feb 2020
☑	GSM3897342: LeCre_rep3; Mus musculus; RNA-Seq	SRR9323048	Illumina HiSeq 2000	14701	28 Feb 2020



6 Saved Datasets

To download FastQ files directly, sra-explorer queries the ENA for each SRA run accession number.

Raw FastQ Download URLs

Bash script for downloading FastQ files

This list of bash `curl` commands to download each SRA run FastQ file from the ENA, and save with a nicer filename, with the cleaned dataset title appended.

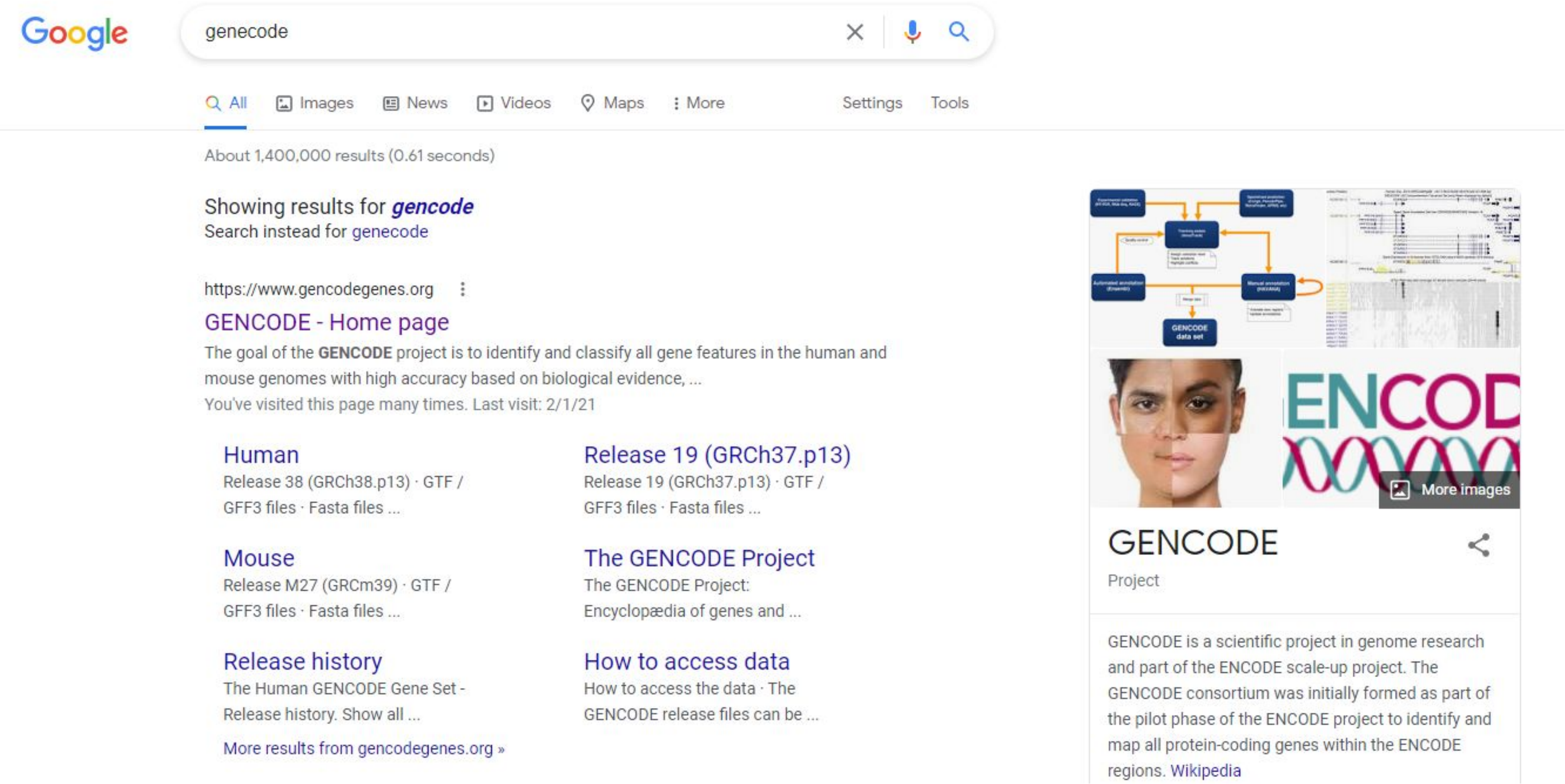
```
#!/usr/bin/env bash
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR932/007/SRR9323047/SRR9323047.fastq.gz -o SRR9323047_GSM3897341_LeCre_rep2_Mus_musculus_RNA-Seq.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR932/008/SRR9323048/SRR9323048.fastq.gz -o SRR9323048_GSM3897342_LeCre_rep3_Mus_musculus_RNA-Seq.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR932/000/SRR9323040/SRR9323040.fastq.gz -o SRR9323040_GSM3897334_Pten_rep1_Mus_musculus_RNA-Seq.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR932/006/SRR9323046/SRR9323046.fastq.gz -o SRR9323046_GSM3897340_LeCre_rep1_Mus_musculus_RNA-Seq.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR932/002/SRR9323042/SRR9323042.fastq.gz -o SRR9323042_GSM3897336_Pten_rep3_Mus_musculus_RNA-Seq.fastq.gz
curl -L ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR932/001/SRR9323041/SRR9323041.fastq.gz -o SRR9323041_GSM3897335_Pten_rep2_Mus_musculus_RNA-Seq.fastq.gz
```

Aspera commands for downloading FastQ files

Cluster Flow FastQ download file (nice filenames)

bcbio project file for FastQ downloads (nice filenames)

Finding reference genome/transcriptome?



About 1,400,000 results (0.61 seconds)

Showing results for ***gencode***
Search instead for genecode

https://www.gencodegenes.org

GENCODE - Home page

The goal of the **GENCODE** project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, ...
You've visited this page many times. Last visit: 2/1/21

Human
Release 38 (GRCh38.p13) · GTF / GFF3 files · Fasta files ...

Release 19 (GRCh37.p13)
Release 19 (GRCh37.p13) · GTF / GFF3 files · Fasta files ...

Mouse
Release M27 (GRCm39) · GTF / GFF3 files · Fasta files ...

The GENCODE Project
The GENCODE Project: Encyclopædia of genes and ...

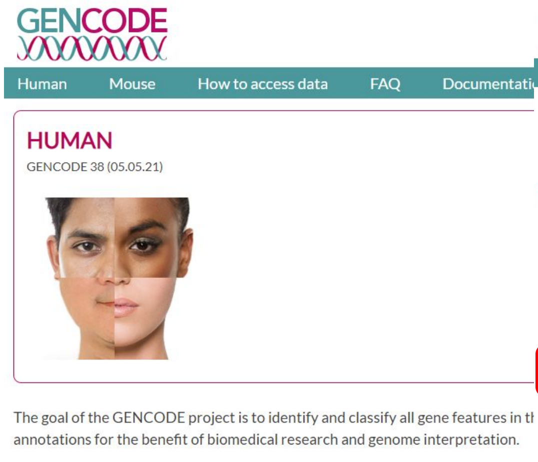
Release history
The Human GENCODE Gene Set - Release history. Show all ...

How to access data
How to access the data · The GENCODE release files can be ...

More results from gencodegenes.org »

GENCODE
Project

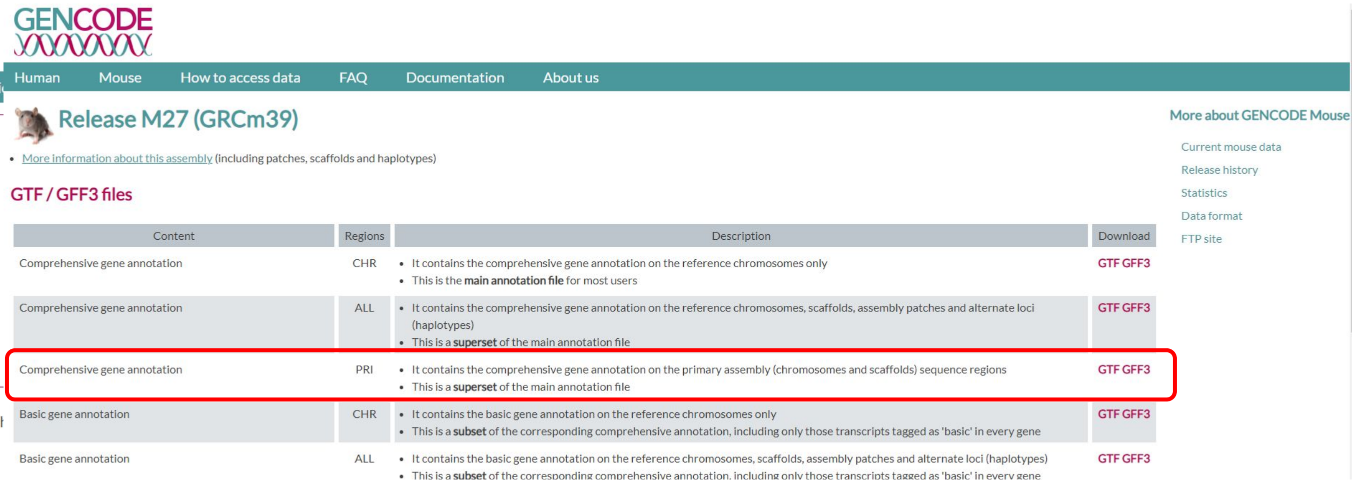
GENCODE is a scientific project in genome research and part of the ENCODE scale-up project. The GENCODE consortium was initially formed as part of the pilot phase of the ENCODE project to identify and map all protein-coding genes within the ENCODE regions. Wikipedia



HUMAN

GENCODE 38 (05.05.21)

The goal of the GENCODE project is to identify and classify all gene features in tl annotations for the benefit of biomedical research and genome interpretation.

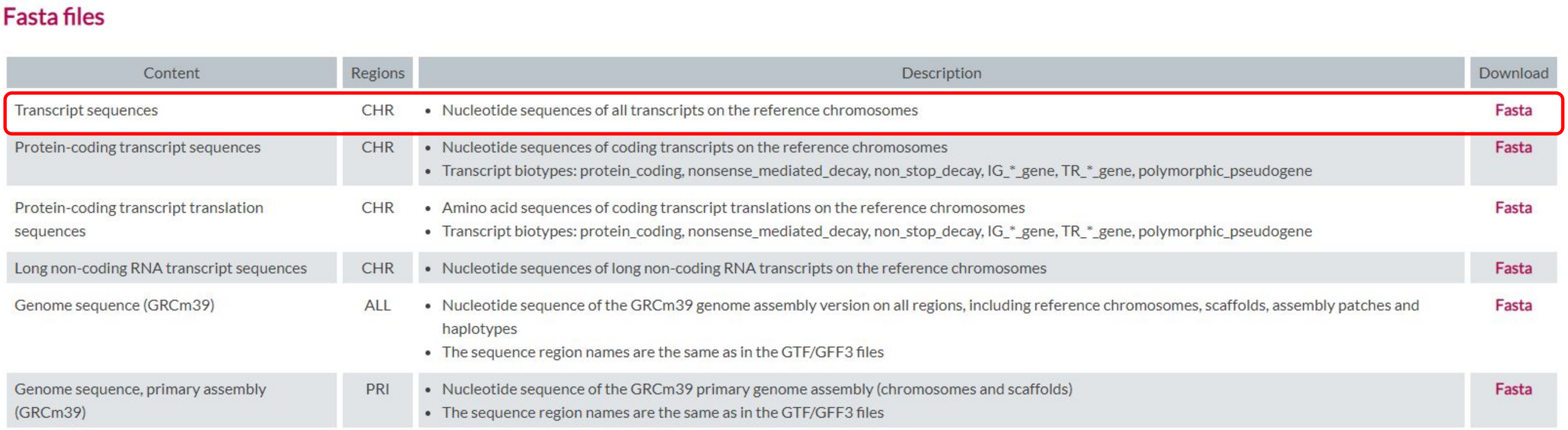


Release M27 (GRCm39)

- More information about this assembly (including patches, scaffolds and haplotypes)

GTF / GFF3 files

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only • This is the **main annotation file** for most users	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **superset** of the main annotation file	GTF GFF3
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions • This is a **superset** of the main annotation file	GTF GFF3
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Basic gene annotation	ALL	• It contains the basic gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3



Fasta files

Content	Regions	Description	Download
Transcript sequences	CHR	• Nucleotide sequences of all transcripts on the reference chromosomes	Fasta
Protein-coding transcript sequences	CHR	• Nucleotide sequences of coding transcripts on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene	Fasta
Protein-coding transcript translation sequences	CHR	• Amino acid sequences of coding transcript translations on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene	Fasta
Long non-coding RNA transcript sequences	CHR	• Nucleotide sequences of long non-coding RNA transcripts on the reference chromosomes	Fasta
Genome sequence (GRCm39)	ALL	• Nucleotide sequence of the GRCm39 genome assembly version on all regions, including reference chromosomes, scaffolds, assembly patches and haplotypes • The sequence region names are the same as in the GTF/GFF3 files	Fasta
Genome sequence, primary assembly (GRCm39)	PRI	• Nucleotide sequence of the GRCm39 primary genome assembly (chromosomes and scaffolds) • The sequence region names are the same as in the GTF/GFF3 files	Fasta

R packages needed

Tximport = Exporting the GFF file for processing by kallisto

GenomicFeatures = Basic package needed for lot of genomic analysis

AnnotationDbi = Basic package for all annotation packages

DESeq2 = DEG processing package

pheatmap = For making heatmaps

Tximport

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")


 BiocManager::install("tximport")
```

https://bioconductor.org/packages/release/bioc/html/tximport.html

GenomicFeatures

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("GenomicFeatures")
```

https://bioconductor.org/packages/release/bioc/html/GenomicFeatures.html

AnnotationDbi

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")


 BiocManager::install("AnnotationDbi")
```

https://bioconductor.org/packages/release/bioc/html/AnnotationDbi.html

DESeq2

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")


BiocManager::install("DESeq2")
```

https://bioconductor.org/packages/release/bioc/html/DESeq2.html

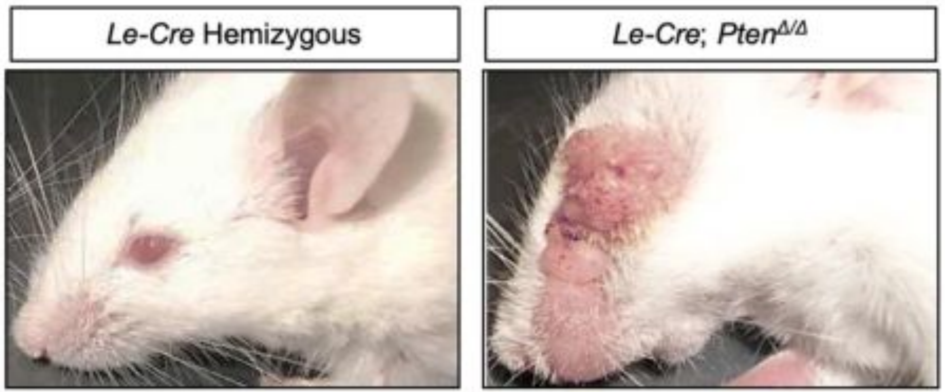
pheatmap

```
install.packages("pheatmap")
```

Understanding data..

What is the role of PTEN in lens development?

- Knock Out PTEN in mice Observe the phenotype
- Find the Differentially Expressed Genes (DEGs)
- Predict the cell signaling pathway leading to this effect
- Verify using wet lab analysis



PTEN present

PTEN deletion

DEGs when compared Le-Cre vs PTEN KO

Bonus

If you want to perform the RNA sequencing analysis on your own

I have provided step by step method for analysis ...

https://businesstech.co.za/news/business/104803/end-of-year-bonuses-may-bring-out-the-worst-in-people/

Anaconda install

Operating system: Windows 8 or newer, 64-bit macOS 10.13+, or Linux, including Ubuntu, RedHat, CentOS 6+, and others

- Minimum 5 GB disk space to download and install.

For Windows

https://docs.anaconda.com/anaconda/install/windows/

For macOS

https://docs.anaconda.com/anaconda/install/mac-os/

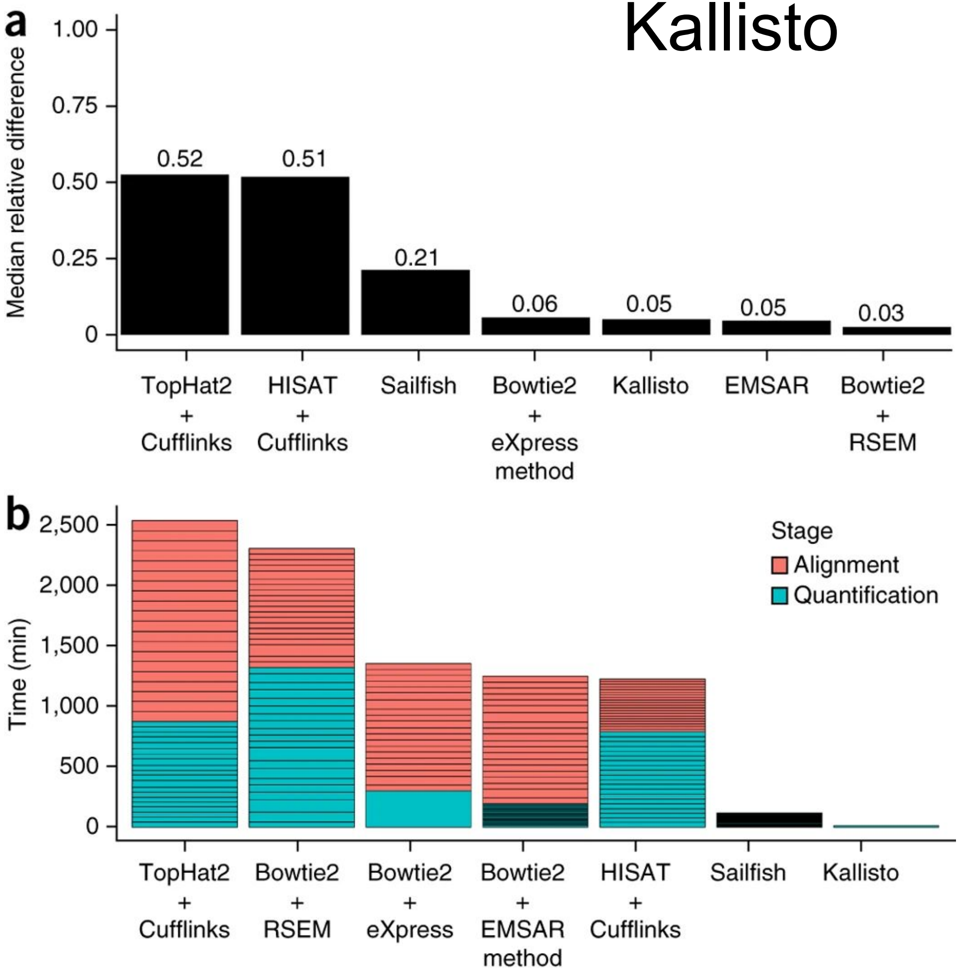
For Linux

https://docs.anaconda.com/anaconda/install/linux/

Anaconda

- Distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.),
- To simplify package management and deployment
- Runs in terminal /Command prompt

Kallisto



https://pachterlab.github.io/kallisto/

"RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases. We use kallisto to analyze 30 million unaligned paired-end RNA-seq reads in <10 min on a standard laptop computer. This removes a major computational bottleneck in RNA-seq analysis."

Nicolas L Bray, Harold Pimentel, Páll Melsted and Lior Pachter, Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology **34**, 525–527 (2016), doi:10.1038/nbt.3519

Creating your environment for RNA seq

```
Step 1:

Open Terminal / Command Prompt

Conda create --name RNA_seq

Step 2:

For windows : activate RNA_seq

For LINUX, macOS: source activate RNA_seq
```

FASTqc installation

```
Step 3:
conda install -c bioconda fastqc
Enter Y
```

Kallisto installation

```
Step 4:

conda install -c biobuilds kallisto

Enter Y
```

Checking all the packages and version installed

```
Step 5:

conda list
```

Exit environment

```
Windows: deactivate

macOS,LINUX : source deactivate
```

Quality control ….

Things needed

1. Le-cre control mice raw file (3)
2. Pten KO raw file (3)
3. Mouse reference transcriptome
4. Comprehensive gene annotation file(GTF)

Make a folder in Desktop named PTEN_KO_DEG_Analysis
Save all the files in this folder

Open terminal

```
For windows : activate RNA_seq
```

```
For LINUX, macOS: source activate RNA_seq
```

cd Desktop/PTEN_KO_DEG_Analysis

Quality of raw reads using FASTqc

Usage: fastqc file 1 file 2 file 3
Example: fastqc LeCre_rep1_Mus_musculus_RNA-Seq.fastq.gz

https://www.dreamstime.com/stock-illustration-quality-control-cartoon-inspector-checks-along-production-line-image91164438

Indexing the reference transcriptome

Kallisto indexing of mouse transcriptome

```
Builds a kallisto index

Usage: kallisto index [arguments] FASTA-files

Required argument:
-i, --index=STRING          Filename for the kallisto index to be constructed
```

```
kallisto index -i Mouse_transcriptome /home/Desktop/Pten/Mouse_transcript.fa.gz
```

Aligning the reads of our interest with reference transcriptome

Open terminal
Step 1: For windows : **activate RNA_seq**

For LINUX, macOS: **source activate RNA_seq**

Step 2: **cd Desktop/PTEN_KO_DEG_Analysis**
mkdir Kallisto_out

Step 3: **ls**
Check for the presence of all of the needed files

1. Le-cre control mice raw file (3)
2. Pten KO raw file (3)
3. Mouse reference transcriptome
4. Comprehensive gene annotation file(GTF)

Step 4: **kallisto quant -i index -o output --single -l 100 -s 2 file1.fastq.gz**

Example: kallisto quant -i ~/Desktop/PTEN_KO_DEG_Analysis/Mouse_transcriptome -o
~/Desktop/PTEN_KO_DEG_Analysis/Kallisto_out/L1 --single -l 100 -s 2 ~/Desktop/PTEN_KO_DEG_Analysi/L1.fq.gz

Making things easy

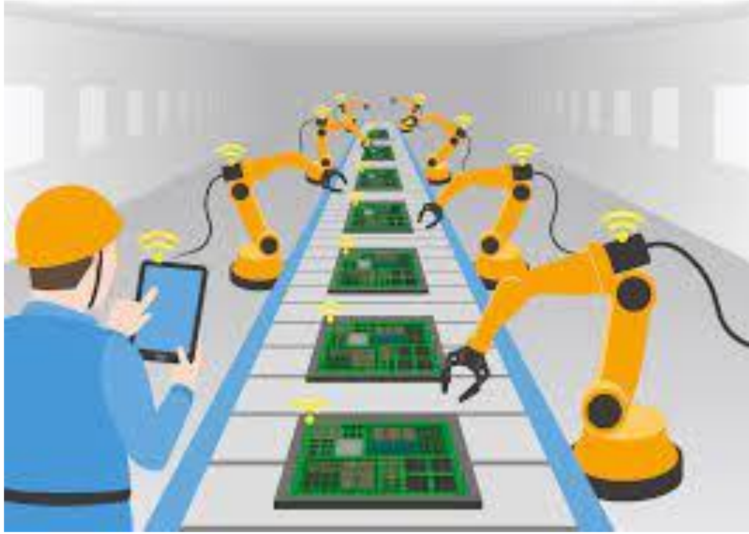
Second command will only run if the first ends with non zero

```
kallisto quant -i ~/Desktop/PTEN_KO_DEG_Analysis/Mouse_transcriptome -o
~/Desktop/PTEN_KO_DEG_Analysis/Kallisto_out/L1 --single -l 100 -s 2
~/Desktop/PTEN_KO_DEG_Analysi/L1.fq.gz &&
kallisto quant -i ~/Desktop/PTEN_KO_DEG_Analysis/Mouse_transcriptome -o
~/Desktop/PTEN_KO_DEG_Analysis/Kallisto_out/L2 --single -l 100 -s 2
~/Desktop/PTEN_KO_DEG_Analysi/L2.fq.gz
```

In the same way you can add all the remaining five sequence file

Output

- **abundances.h5** is a HDF5 binary file containing run info, abundance estimates, bootstrap estimates, and transcript length information length.
- **abundances.tsv** is a plaintext file of the abundance estimates. It does not contains bootstrap estimates. The first line contains a header for each column, including estimated counts, TPM, effective length.
- **run_info.json** is a json file containing information about the run



https://techcrunch.com/2016/04/21/the-automation-revolution-and-the-rise-of-the-creative-economy/