RNA-seq data analysis workshop for biologists, part I

from raw data to read counts

Reference

Genome

Database of all the DNA
of the organism

new transcripts
gene isoforms

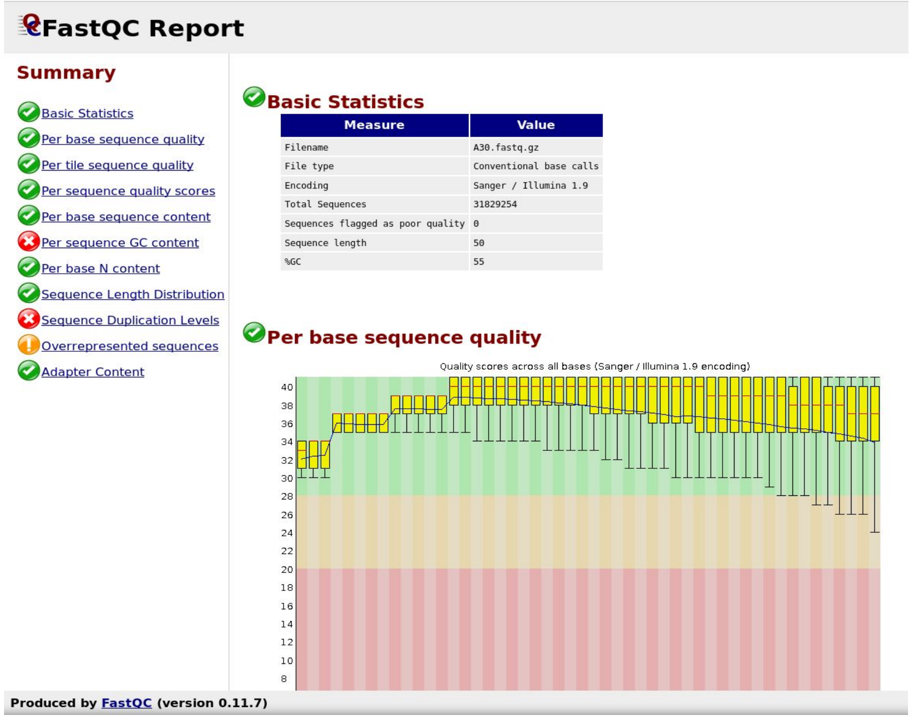
Transcriptome

Database of all known
transcripts for the organism

more accurate quantification

1. **Genome or transcriptome?**
2. Where can I find the reference?

What are the steps of an RNA-Seq analysis?



FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content

✓ Basic Statistics

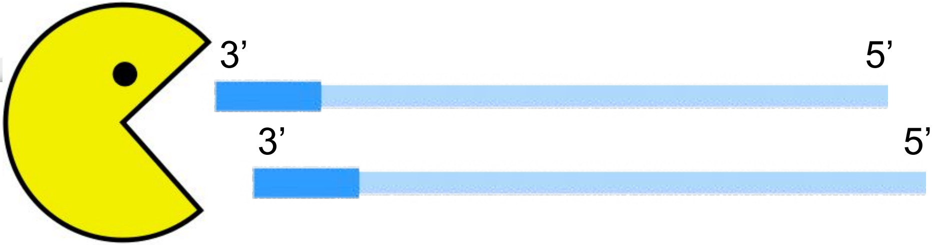
Measure	Value
Filename	A30.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	31829254
Sequences flagged as poor quality	0
Sequence length	50
%GC	55

✓ Per base sequence quality

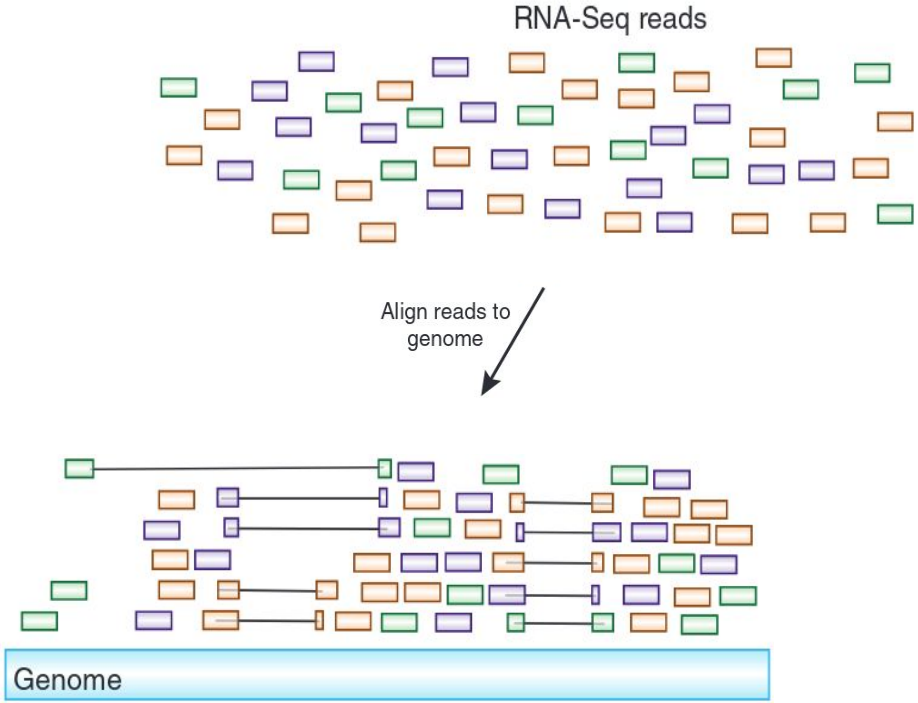
Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by FastQC (version 0.11.7)

1. **Quality control**
2. Alignment or mapping
3. Count reads
4. Differential expression



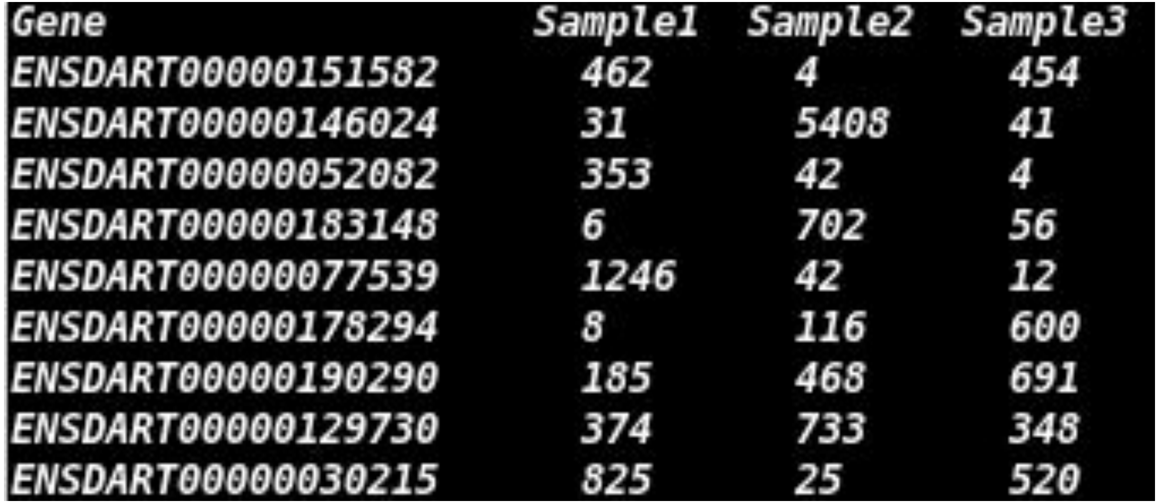
What are the steps of an RNA-Seq analysis?



B. Haas, M. Zody. Advancing RNA-Seq analysis. Nature biotechnology. 28:5. 2010.

1. Quality control
2. **Alignment or mapping**
3. Count reads
4. Differential expression

What are the steps of an RNA-Seq analysis?

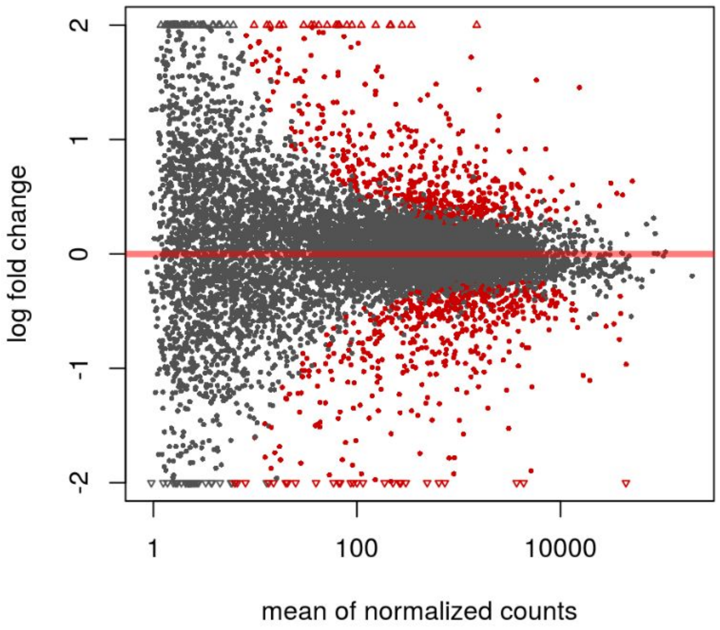


```
Gene                 Sample1  Sample2  Sample3
ENSDART00000151582    462      4        454
ENSDART00000146024    31       5408     41
ENSDART00000052082    353      42       4
ENSDART00000183148    6        702      56
ENSDART00000077539    1246     42       12
ENSDART00000178294    8        116      600
ENSDART00000190290    185      468      691
ENSDART00000129730    374      733      348
ENSDART00000030215    825      25       520
```

1. Quality control
2. Alignment or mapping
3. **Reads counting**
4. Differential expression

What are the steps of an RNA-Seq analysis?

Will be covered in second part of the workshops



1. Quality control
2. Alignment or mapping
3. Count reads
4. **Differential expression**

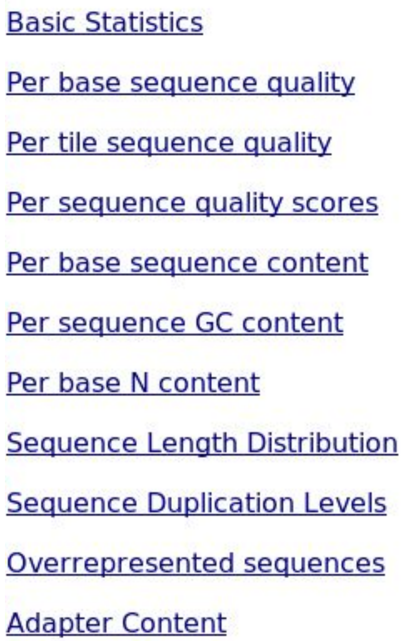
Quality control

Tools:

FastQC

MultiQC

What statistics we are interested in?



Basic Statistics

Per base sequence quality

Per tile sequence quality

Per sequence quality scores

Per base sequence content

Per sequence GC content

Per base N content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences

Adapter Content

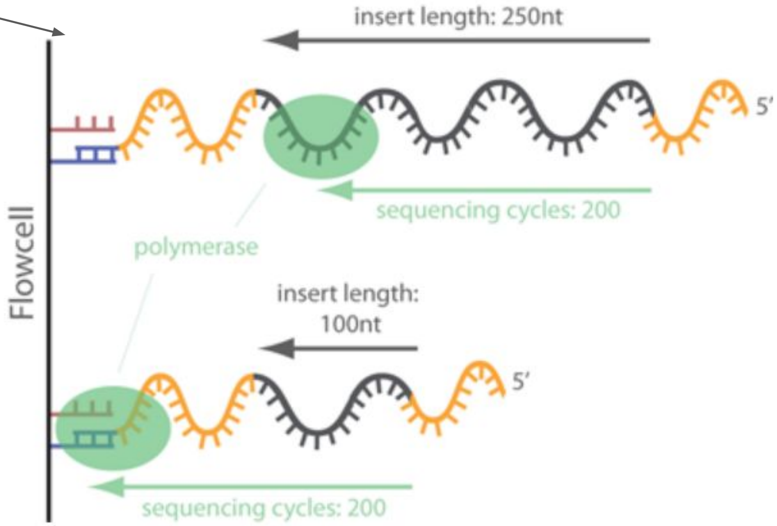
1. **Quality visualization**
2. Reads filtering / trimming

Quality control

1. Quality visualization
2. **Reads filtering / trimming**

Why we filter or trimm our precious data:

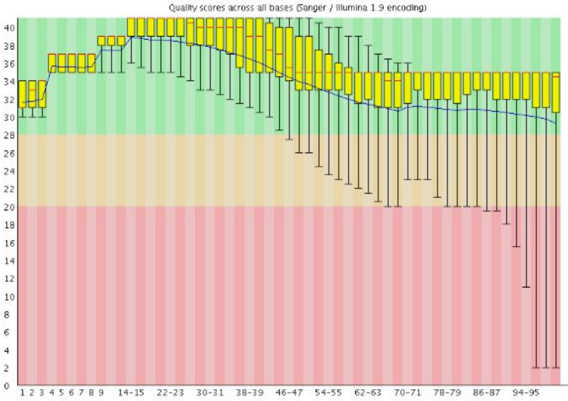
- adapters
- low read quality
- rRNA
- mtDNA

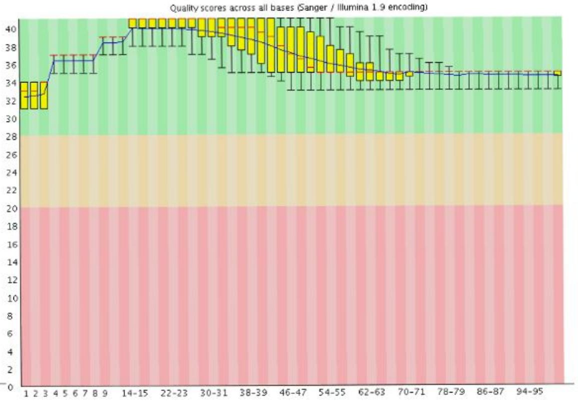


http://www.ecseq.com/support/ngs/trimming-adapter-sequences-is-it-necessary

Quality control

Quality trimming





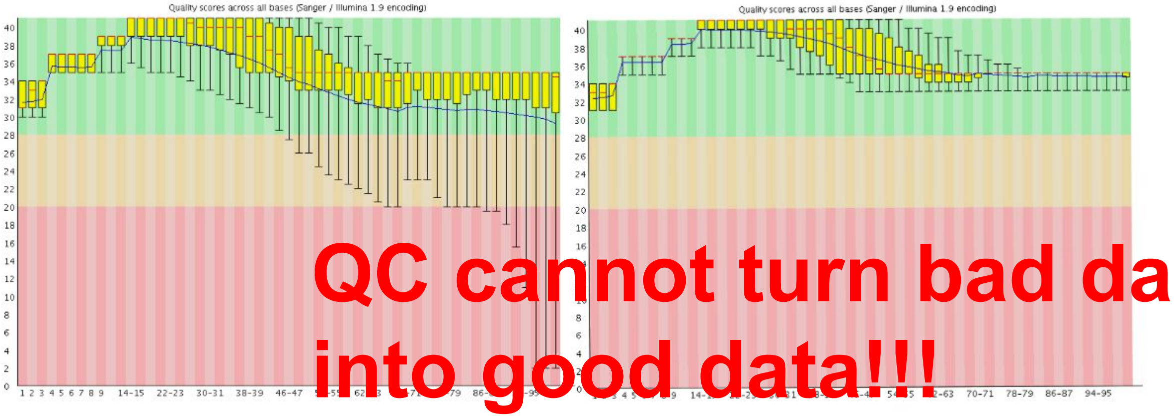
1. Quality visualization
2. **Reads filtering / trimming**

Adapter trimming

Quality control

Quality trimming

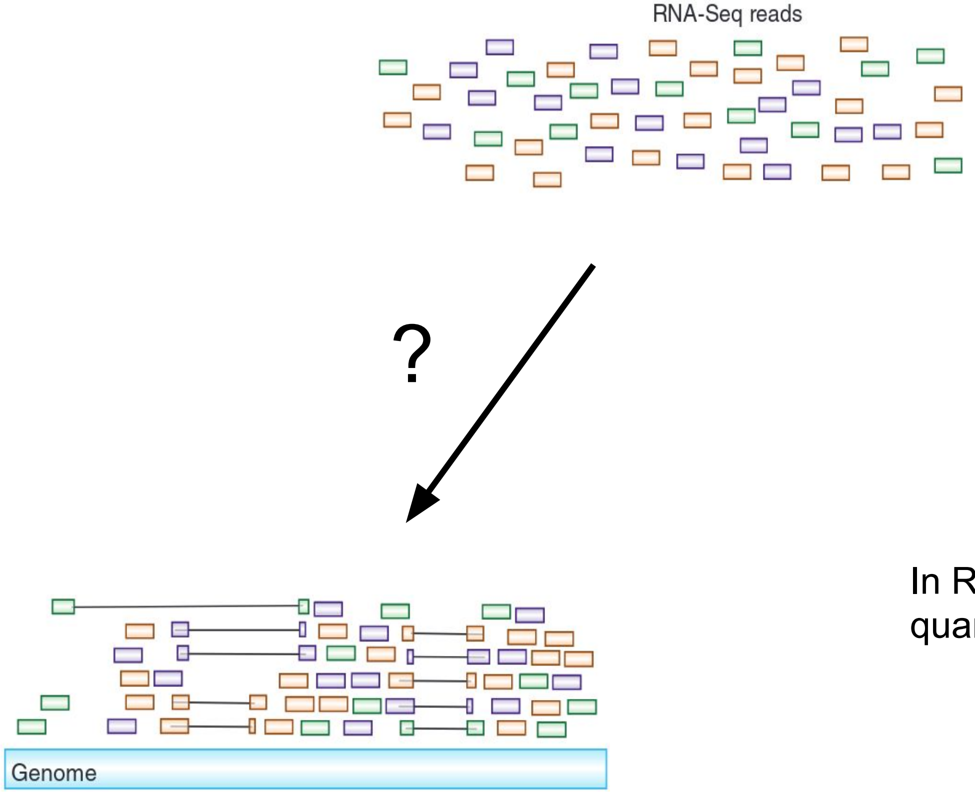
1. Quality visualization
2. **Reads filtering / trimming**



QC cannot turn bad data into good data!!!

Adapter trimming

Alignments or mapping



1. **How does it work?**
2. What are the options?

In RNA-seq we are interested in quantification

Alignments or mapping

Alignment methods

- STAR
- HiSat2
- BWA
- BBMap
- Subjunc

Mapping methods

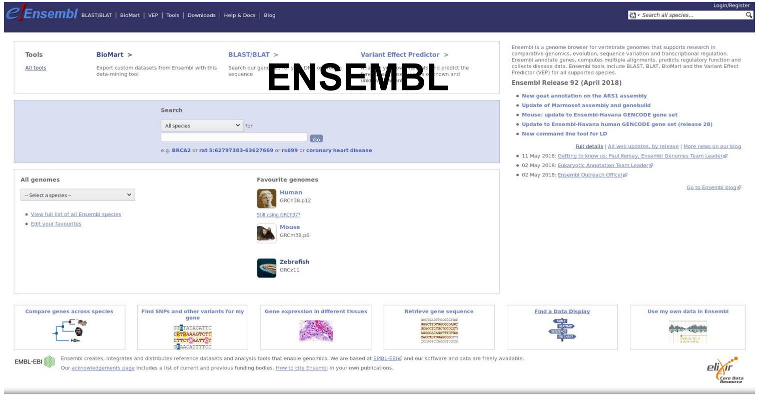
- Salmon
- Kalisto

1. How does it work?
2. **What are the options?**

Reference

1. Genome or transcriptome?
2. **Where can I find the reference?**

ENSEMBL



e!Ensembl BLAST/BLAT | BioMart | VEP | Tools | Downloads | Help & Docs | Blog

Login/Register

Search all species...

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes

-- Select a species --

- View full list of all Ensembl species
- Edit your favourites

Favourite genomes

Human
GRCh38.p12

Still using GRCh37?

Mouse
GRCm38.p6

Zebrafish
GRCz11

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 92 (April 2018)

- New goat annotation on the ARS1 assembly
- Update of Marmoset assembly and genebuild
- Mouse: update to Ensembl-Havana GENCODE gene set
- Update to Ensembl-Havana human GENCODE gene set (release 28)
- New command line tool for LD

Full details | All web updates, by release | More news on our blog

- 11 May 2018: Getting to know us: Paul Kersey, Ensembl Genomes Team Leader
- 02 May 2018: Eukaryotic Annotation Team Leader
- 02 May 2018: Ensembl Outreach Officer

Go to Ensembl blog

Compare genes across species | Find SNPs and other variants for my gene | Gene expression in different tissues | Retrieve gene sequence | Find a Data Display | Use my own data in Ensembl

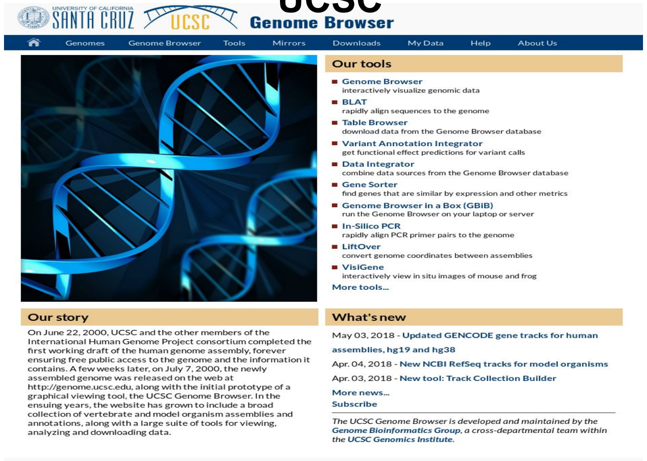
EMBL-EBI

Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available.

Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

elixir Core Data Resource

UCSC



UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC Genome Browser

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Our tools

- Genome Browser
  interactively visualize genomic data
- BLAT
  rapidly align sequences to the genome
- Table Browser
  download data from the Genome Browser database
- Variant Annotation Integrator
  get functional effect predictions for variant calls
- Data Integrator
  combine data sources from the Genome Browser database
- Gene Sorter
  find genes that are similar by expression and other metrics
- Genome Browser in a Box (GBIB)
  run the Genome Browser on your laptop or server
- In-Silico PCR
  rapidly align PCR primer pairs to the genome
- LiftOver
  convert genome coordinates between assemblies
- VisiGene
  interactively view in situ images of mouse and frog

More tools...

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains. A few weeks later, on July 7, 2000, the newly assembled genome was released on the web at http://genome.ucsc.edu, along with the initial prototype of a graphical viewing tool, the UCSC Genome Browser. In the ensuing years, the website has grown to include a broad collection of vertebrate and model organism assemblies and annotations, along with a large suite of tools for viewing, analyzing and downloading data.

What's new

May 03, 2018 - Updated GENCODE gene tracks for human assemblies, hg19 and hg38

Apr. 04, 2018 - New NCBI RefSeq tracks for model organisms

Apr. 03, 2018 - New tool: Track Collection Builder

More news...

Subscribe

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the UCSC Genomics Institute.