Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 02

Pierre-Luc Germain

ETH Zürich

Plan for today

- Debriefing on the assignments

- The notion of gene
- Genome builds and transcriptome assemblies

- Practical:
  - *AnnotationHub* and *EnsDb* objects
  - *GenomicRanges* and their manipulations

Debriefing on the assignments

- Name the exercises **assignment.html** & **assignment.rmd**

- Do **not** change the folder structure. Always put it in the correct **week_xx** folder (e.g. week01)
  - If you don’t have the week folders already, just sync your repository with the parent one
  - Do not use subfolders (e.g. week01/out etc)

- Do not turn off messages and warnings in the chunk preamble, e.g. don’t do:
  - `` ```{r, warning=FALSE, message=FALSE} ``
  - or `knitr::opts_chunk$set(message = FALSE, warning = FALSE)`

- Many are registered to the course but didn’t give us their github username (nor have forked the repo)

A very brief history of genetics & genomics

… or why nobody knows how many genes we have

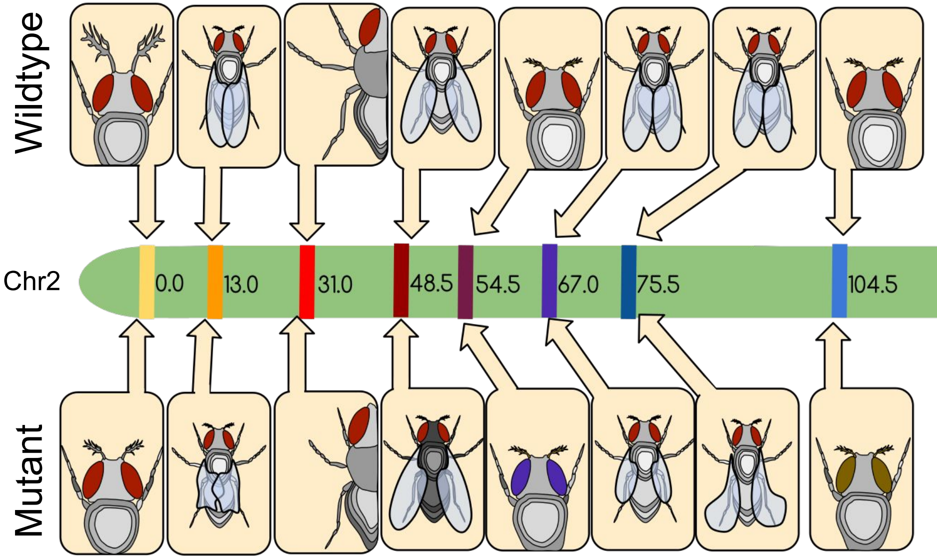
A very brief history of genetics & genomics

1900 - Rediscovery of Mendel’s work (1860s)
1903 - Chromosomes are hereditary units
1913 - Chromosomes are linear arrays of genes
1941 - Beadle & Tatum:
the one-gene-one-enzyme hypothesis
1944 - DNA is the genetic material
1951 - First protein sequences
1953 - DNA is a double helix
1961 - Jacob and Monod:
the *lac* operon
1977 - DNA sequencing
1977 - Eukaryotic genes are spliced
1995 - First bacterial genomes sequenced
2000 - Next Generation Sequencing (NGS)
2001 - Draft of the human genome
2003 - RNA-seq
2006 - ChIP-seq
2012 - ENCODE, ATAC-seq

A very brief history of genetics & genomics

1900 - Rediscovery of Mendel’s work (1860s)
1903 - Chromosomes are hereditary units
1913 - Chromosomes are linear arrays of genes
1941 - Beadle & Tatum:
the one-gene-one-enzyme hypothesis
1944 - DNA is the genetic material
1951 - First protein sequenced
1953 - DNA is a double helix
1961 - Jacob and Monod:
the *lac* operon
1977 - DNA sequencing
1977 - Eukaryotic genes are spliced
1995 - First bacterial genomes sequenced
2000 - Next Generation Sequencing (NGS)
2001 - Draft of the human genome
2003 - RNA-seq
2006 - ChIP-seq
2012 - ENCODE, ATAC-seq

Classical genetics: A “gene” is a unit of heredity

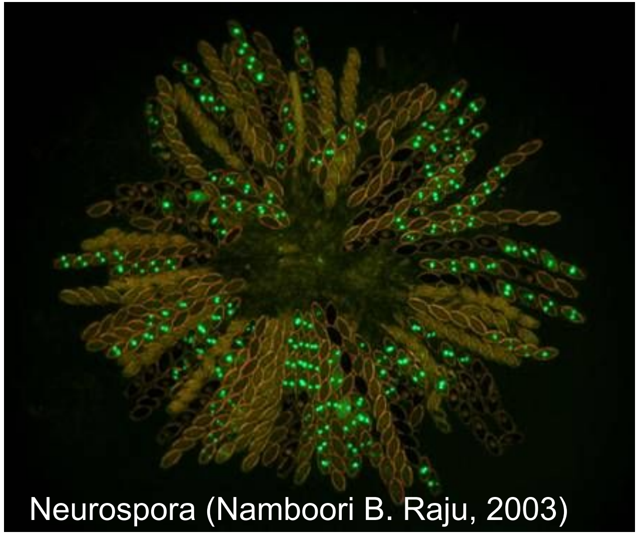


(Morgan's Drosophila genetic map, adapted from Twaanders17, CC BY-SA 4.0, via Wikimedia Commons)

A very brief history of genetics & genomics

1900 - Rediscovery of Mendel’s work (1860s)
1903 - Chromosomes are hereditary units
1913 - Chromosomes are linear arrays of genes
1941 - Beadle & Tatum:
the one-gene-one-enzyme hypothesis
1944 - DNA is the genetic material
1951 - First protein sequenced
1953 - DNA is a double helix
1961 - Jacob and Monod:
the *lac* operon
1977 - DNA sequencing
1977 - Eukaryotic genes are spliced
1995 - First bacterial genomes sequenced
2000 - Next Generation Sequencing (NGS)
2001 - Draft of the human genome
2003 - RNA-seq
2006 - ChIP-seq
2012 - ENCODE, ATAC-seq

Classical genetics:	A “gene” is a unit of heredity
	(+)
Early molecular genetics:	A “gene” is a part of DNA that encodes for a protein


Neurospora (Namboori B. Raju, 2003)

A very brief history of genetics & genomics

1900 - Rediscovery of Mendel's work (1860s)

1903 - Chromosomes are hereditary units

1913 - Chromosomes are linear arrays of genes

1941 - Beadle & Tatum:
the one-gene-one-enzyme hypothesis

1944 - DNA is the genetic material

1951 - First protein sequenced

1953 - DNA is a double helix

1961 - Jacob and Monod:
the *lac* operon

1977 - DNA sequencing

1977 - Eukaryotic genes are spliced

1995 - First bacterial genomes sequenced

2000 - Next Generation Sequencing (NGS)

2001 - Draft of the human genome

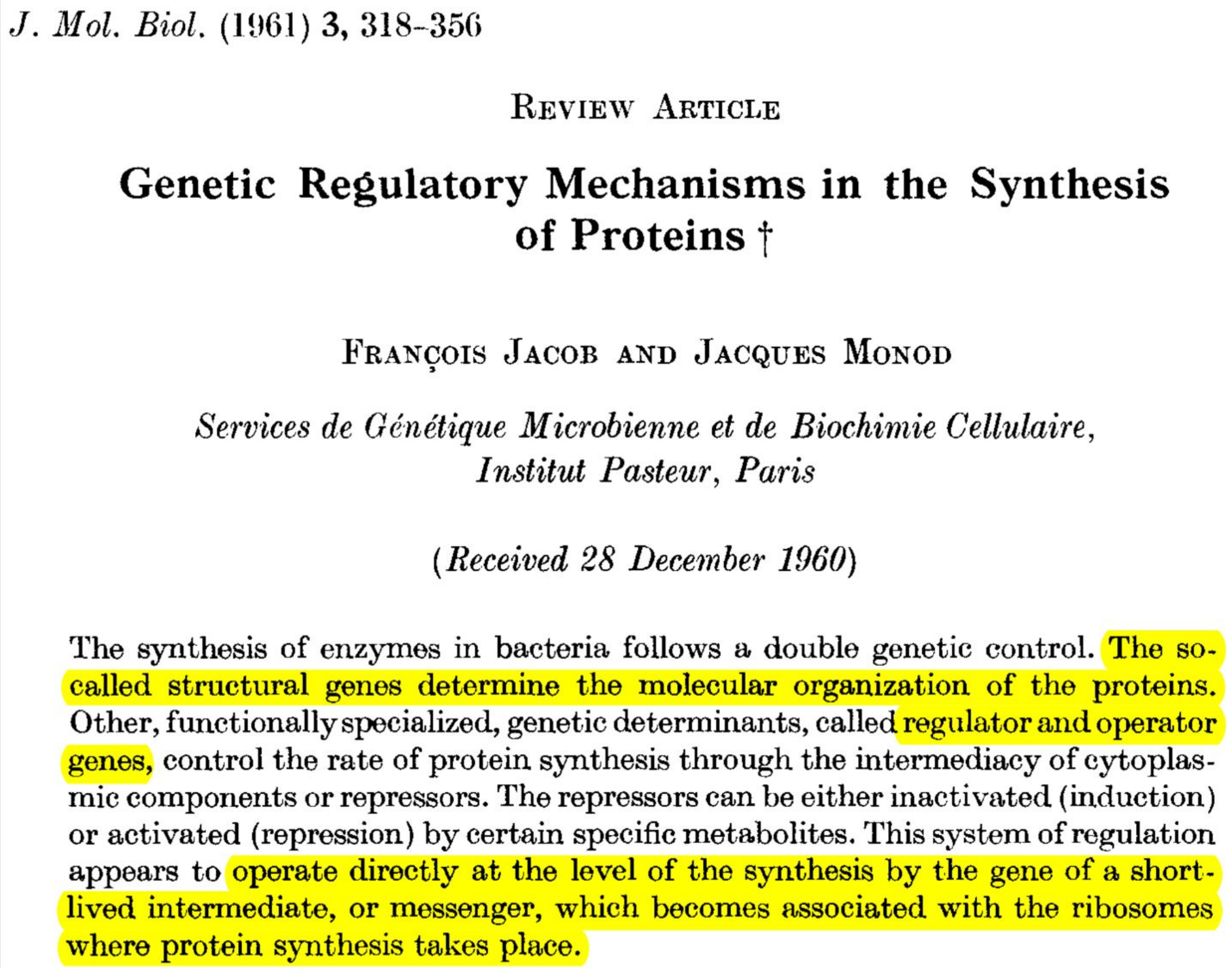
2003 - RNA-seq

2006 - ChIP-seq

2012 - ENCODE, ATAC-seq

Classical genetics:	A "gene" is a unit of heredity
Early molecular genetics:	A "gene" is a part of DNA that encodes for a protein

- "structural" genes encode for proteins
- "regulatory" genes → regulate the structural genes



J. Mol. Biol. (1961) **3**, 318–356

REVIEW ARTICLE

Genetic Regulatory Mechanisms in the Synthesis of Proteins †

FRANÇOIS JACOB AND JACQUES MONOD

Services de Génétique Microbienne et de Biochimie Cellulaire,
Institut Pasteur, Paris

(*Received 28 December 1960*)

The synthesis of enzymes in bacteria follows a double genetic control. The so-called structural genes determine the molecular organization of the proteins. Other, functionally specialized, genetic determinants, called regulator and operator genes, control the rate of protein synthesis through the intermediacy of cytoplasmic components or repressors. The repressors can be either inactivated (induction) or activated (repression) by certain specific metabolites. This system of regulation appears to operate directly at the level of the synthesis by the gene of a short-lived intermediate, or messenger, which becomes associated with the ribosomes where protein synthesis takes place.

A very brief history of genetics & genomics

1900 - Rediscovery of Mendel's work (1860s)
1903 - Chromosomes are hereditary units
1913 - Chromosomes are linear arrays of genes
1941 - Beadle & Tatum:
the one-gene-one-enzyme hypothesis
1944 - DNA is the genetic material
1951 - First protein sequenced
1953 - DNA is a double helix
1961 - Jacob and Monod:
the *lac* operon
1977 - DNA sequencing
1977 - Eukaryotic genes are spliced
1995 - First bacterial genomes sequenced
2000 - Next Generation Sequencing (NGS)
2001 - Draft of the human genome
2003 - RNA-seq
2006 - ChIP-seq
2012 - ENCODE, ATAC-seq

Classical genetics:	A "gene" is a unit of heredity
Early molecular genetics:	A "gene" is a part of DNA that encodes for a protein

- "structural" genes encode for proteins
- "regulatory" genes → regulate the structural genes

A very brief history of genetics & genomics

1900 - Rediscovery of Mendel's work (1860s)

1903 - Chromosomes are hereditary units

1913 - Chromosomes are linear arrays of genes

1941 - Beadle & Tatum:
the one-gene-one-enzyme hypothesis

1944 - DNA is the genetic material

1951 - First protein sequenced

1953 - DNA is a double helix

1961 - Jacob and Monod:
the *lac* operon

1977 - DNA sequencing

1977 - Eukaryotic genes are spliced

1995 - First bacterial genomes sequenced

2000 - Next Generation Sequencing (NGS)

2001 - Draft of the human genome

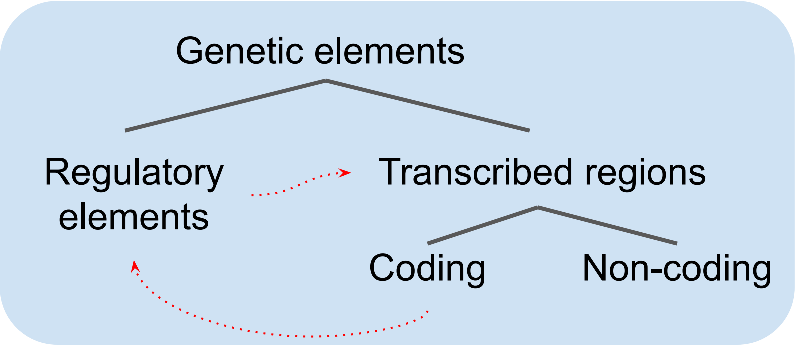
2003 - RNA-seq

2006 - ChIP-seq

2012 - ENCODE, ATAC-seq

Classical genetics:	A "gene" is a unit of heredity
Early molecular genetics:	A "gene" is a part of DNA that encodes for a protein

- "structural" genes encode for proteins
- "regulatory" genes → regulate the structural genes



Ignoble Nano-lectures:

- A 24sec technical summary
- A 7-words summary everyone can understand

So what’s a gene today?

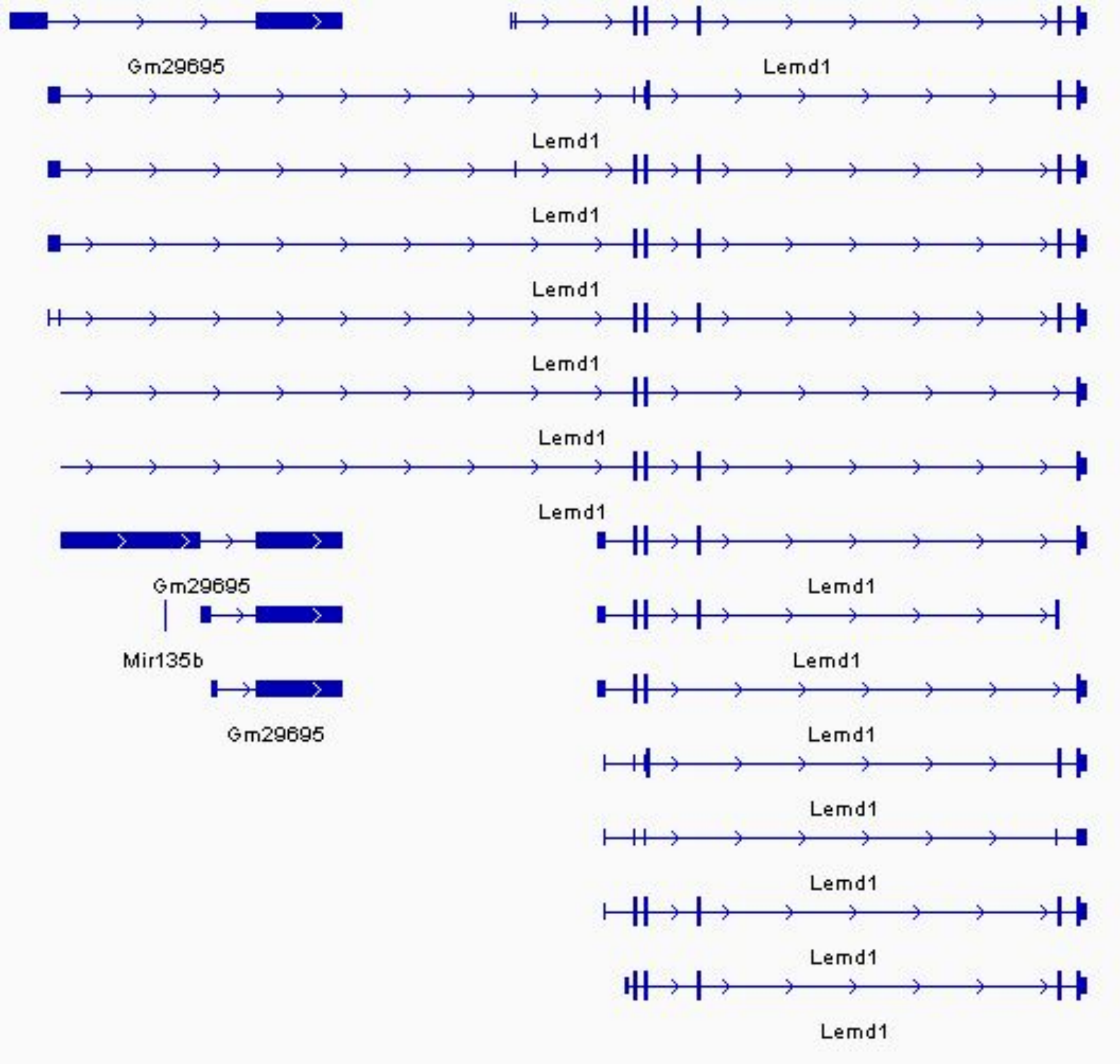
"The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products"

(Gerstein et al., 2007)

“On this view, genes represent a higher-order framework around which individual transcripts coalesce, creating a poly-functional entity that assumes different forms under different cellular states, guided by differential utilization of regulatory DNA."

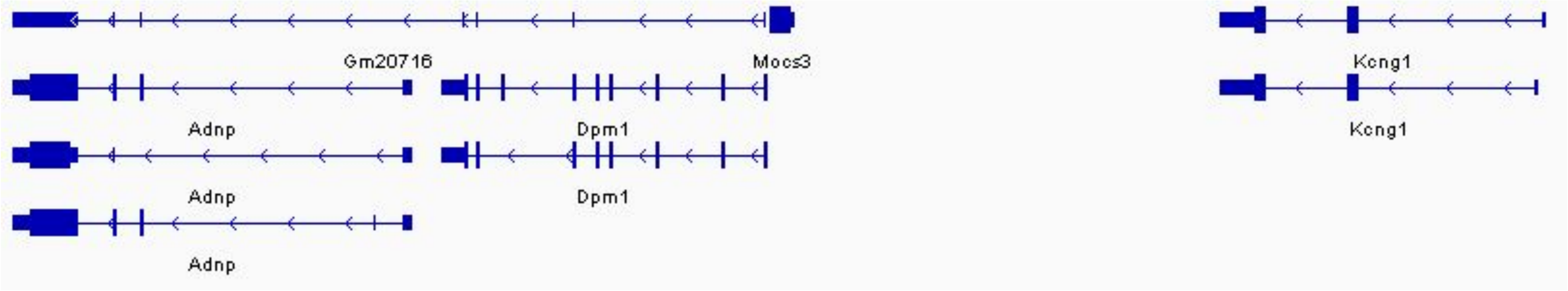
(Stamatoyannopoulos, 2012)

ENCODE



Transcripts tend to be grouped into genes depending on partially overlapping exonic sequences

But of course that’s not a universal rule…



Reference genomes and gene annotations

The **reference genome** refers to the **sequence** of the genome of a species.

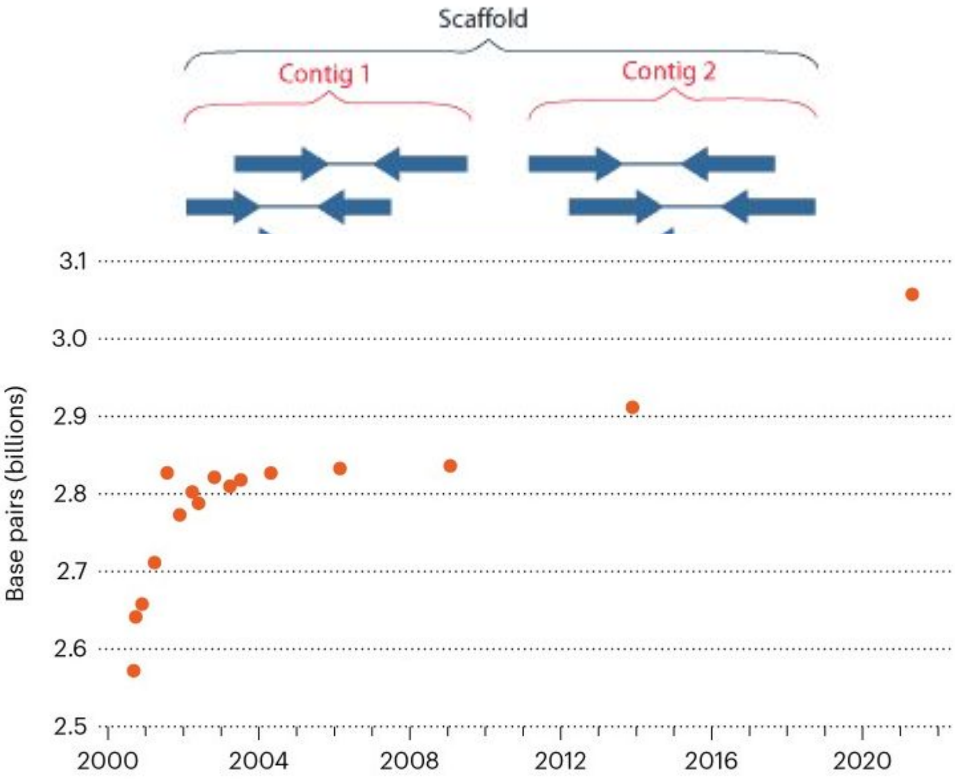
For mainstream organisms (e.g. human/mouse), there is one standardized genome (from the Genome Reference Consortium) which everyone uses, but different **build** versions of that genome, e.g.:

Human:

- GRCh38/hg38 (released in 2013)
- GRCh37/hg19 (released in 2009)
- ...

Mouse:

- GRCm39/mm39 (released in 2020)
- GRCm38/mm10 (released in 2011)
- GRCm37/mm9 (released in 2007)
- ...



(Reardon, Nature News 2021)

Reference genomes and gene annotations

The **reference genome** refers to the **sequence** of the genome of a species.

For mainstream organisms (e.g. human/mouse), there is one standardized genome (from the Genome Reference Consortium) which everyone uses, but different **build** versions of that genome, e.g.:

Human:

- GRCh38/hg38 (released in 2013)
- GRCh37/hg19 (released in 2009)
- ...

Mouse:

- GRCm39/mm39 (released in 2020)
- GRCm38/mm10 (released in 2011)
- GRCm37/mm9 (released in 2007)
- ...

Within each genome build, there are also **patches**:

GRCh38.p13 (released in 2019)
GRCh38.p12 (released in 2017)
…

Coordinates are stable *within* a build, but not *across* builds

Reference genomes and gene annotations

The **reference genome** refers to the **sequence** of the genome of a species.

For mainstream organisms (e.g. human/mouse), there is one standardized genome (from the Genome Reference Consortium) which everyone uses, but different **build** versions of that genome, e.g.:

Human:

- GRCh38/hg38 (released in 2013)
- GRCh37/hg19 (released in 2009)
- ...

Mouse:

- GRCm39/mm39 (released in 2020)
- GRCm38/mm10 (released in 2011)
- GRCm37/mm9 (released in 2007)
- ...

Annotations refer to the catalogues of regulatory elements, genes and transcripts associated to a genome.

There are two main sources for gene annotations (aka “gene builds”) :

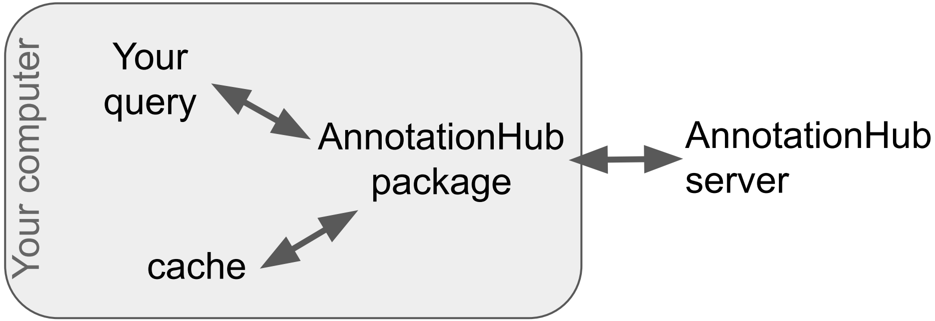
Ensembl	example transcript:
○ 104 (may 2021)	ENST000012343.2
○ 103 (nov 2020)	ENST000012343.1
○ 102 (april 2020)	ENST000012343.1
○ …	

(Also GENCODE → now pretty much equals Ensembl)

Accessing ensembl annotations

- http://ftp.ensembl.org/pub/
- contains various types of data, most importantly:
  - **.fasta** files: sequences
    - (e.g. DNA or cDNA)
  - **.gtf/.gff** files: gene models
    - (i.e. exon coordinates and inclusion into transcripts)
- Similar on GENCODE https://www.gencodegenes.org/

- **AnnotationHub**
- Standardized access to a large variety of annotations



Your computer

Your query

AnnotationHub package

cache

AnnotationHub server

- including genomes, gene annotations (e.g. EnsDb objects) and more

Further information

- See the documentation of the ensembldb package for manipulating `EnsDb` objects
    - (if you need to work from a gtf, see ensembldb::ensDbFromGtf)

- See the documentation of the GenomicRanges package for manipulating `GRanges` objects (and their derivatives)

Assignment for this week

- Using AnnotationHub, find and download the following annotations data:
  - The mouse (Mus Musculus) EnsDb object, version 102, genome build GRCm38
  - The mouse genome sequence (dna_sm) in TwoBit/2bit format for GRCm38
  - The drosophila melanogaster genome sequence (dna_sm) in TwoBit/2bit format for BDGP6

- Using the mouse EnsDb, find the following:
  - How many different *ensembl gene IDs* and *gene symbols* are there for protein-coding genes?
  - Plot the distribution of the (spliced) length of protein-coding transcripts
    - (tip: this will require you to extract exons of protein-coding transcripts from the database, and split them by transcript, before summing the width of the exons of each transcript)

Name your markdown file ‘`assignment.Rmd`’, render it, and put it (along with the produced html) in the `week02` folder of your repository, and push!