

Introduction to high-throughput experiments and data analysis

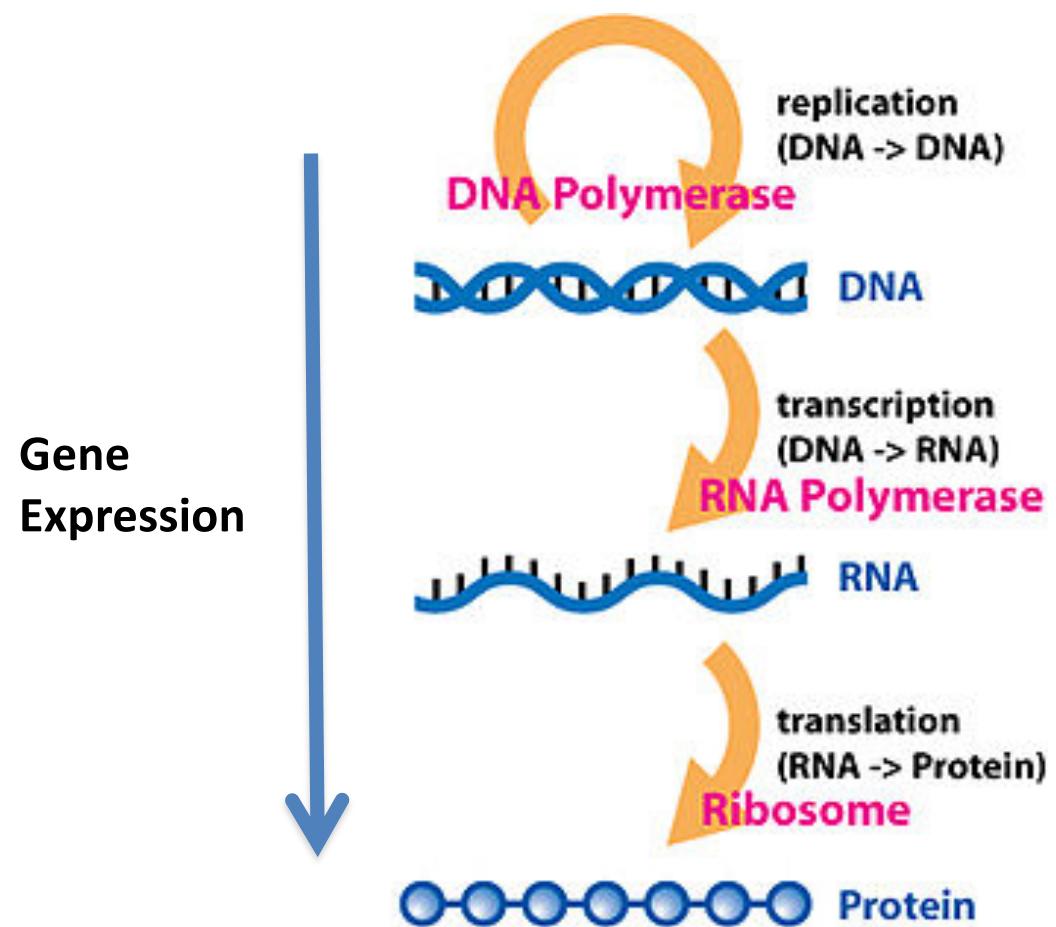
Outline

- Biology in a nutshell.
- High-throughput experiments:
 - microarrays.
 - Second generation sequencing.
- R and Bioconductor.
- Online resources: genome browser and public data repositories.

Biology in a nutshell

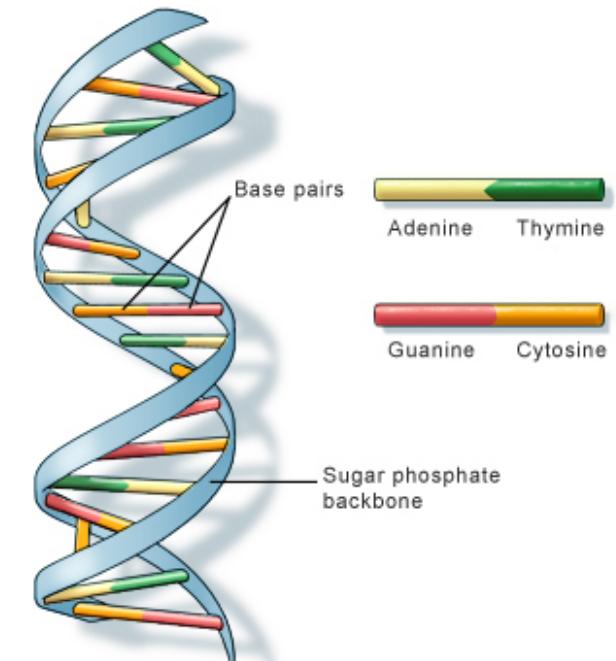
Central dogma of molecular biology

- By Francis Crick
(1970) *Nature*:
"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that information cannot be transferred back from protein to either protein or nucleic acid."



DNA (DeoxyriboNucleic Acid)

- A molecule contains the genetic instruction of all known living organisms and some viruses.
- Resides in the cell nucleus, where DNA is organized into long structures called **chromosomes**.
- Most DNA molecule consists of two long polymers (**strands**), where two strands entwine in the shape of a double helix.
- Each strand is a chain of simple units (**bases**) called **nucleotides**: A, C, G, T.
- The bases from two strands are complementary by **base pairing**: A-T, C-G.



U.S. National Library of Medicine

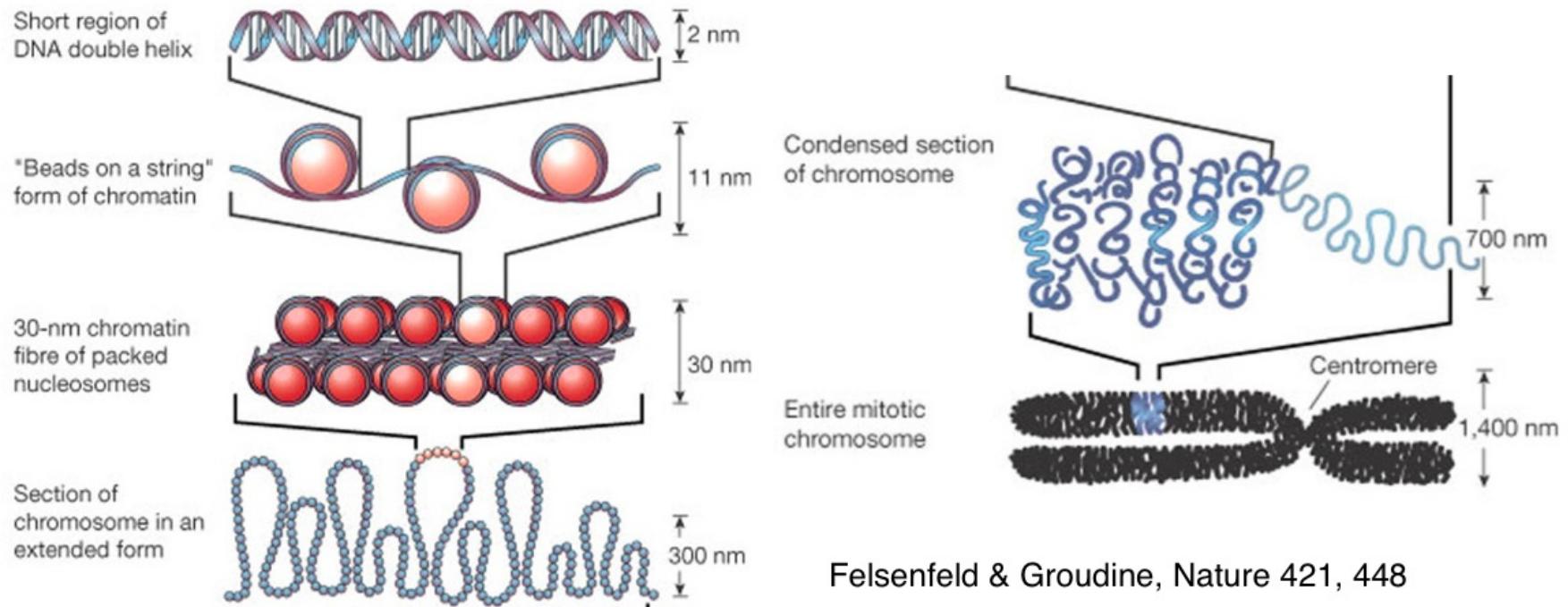
DNA sequence

- The order of occurrence of the bases in a DNA molecule is called the **sequence** of the DNA. The DNA sequence is usually stored in a big text file:

```
ACAGGTTTGTGGTACCGAGTTCTTCATGAGGGACCATCTATCACAAACAG  
AGAAAGCACTTGGATCCACCAGGGCTGCCAGGGGAAGCAGCATGGGAGC  
CTGAACCATGAAGCAGGAAGCACCTGTCTGTAGGGGGAAAGTGATGGAAGG  
ACATGGGCACAGAAGGGTAGGTTTGTCTGGAGGACACTGGGAGTG  
GCTCCTGGCATTGAAACAGGTGTAGAAGGATGTGGTGGACCTACAGA  
CAGACTGGAATCTAAGGGACACTTGAATCCCAGTGTGACCATGGTCTTA  
AGGACAGGTTGGggccaggcacagtggctcatgcctgtaatcccagcact
```

- Some interesting facts:
 - Total length of the human DNA is **3 billion bases**.
 - Difference in DNA sequence between two individuals is less than 1%.
 - Human and chimpanzee have 96% of the sequences identical. Human and mouse: 70%.

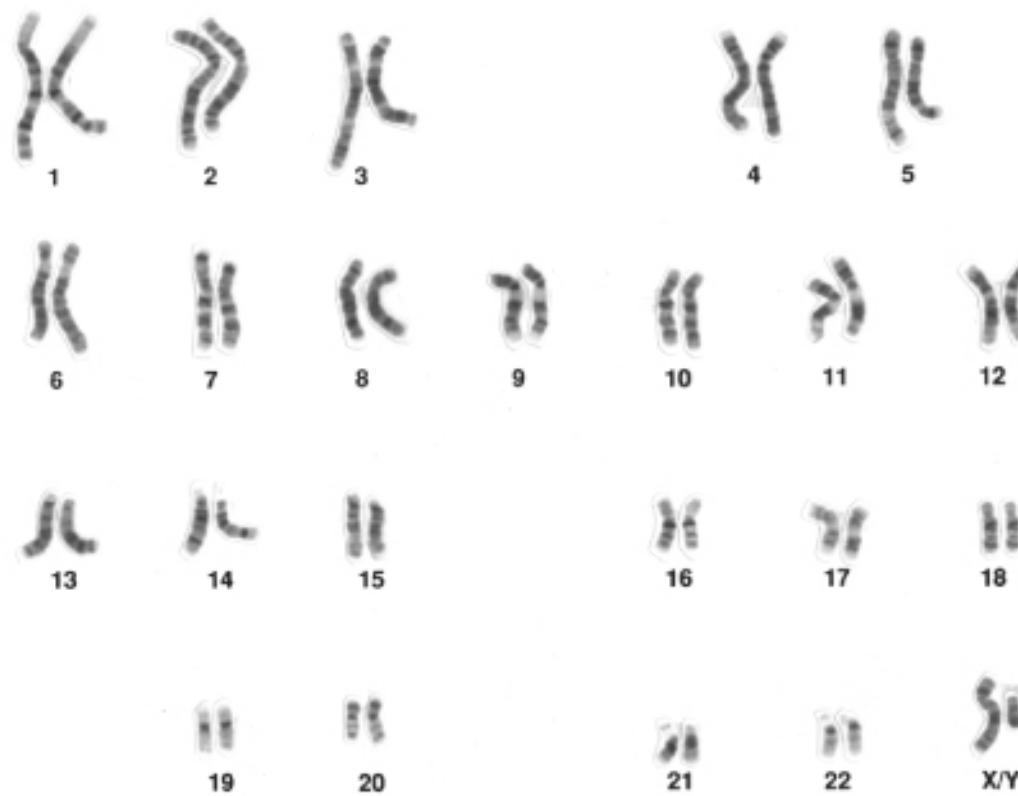
Chromosome: organized structure of DNA and proteins



- **Ploidy:** number of set of chromosomes in a cell.
 - monoploid, diploid or polyploid.
 - Human are diploid: cells have two copies of each chromosome, one from mother and one from father.

Genome

- All of the heritable biological information needed to build and maintain a living example of that organism.
- Or simply, the full set of chromosomes.

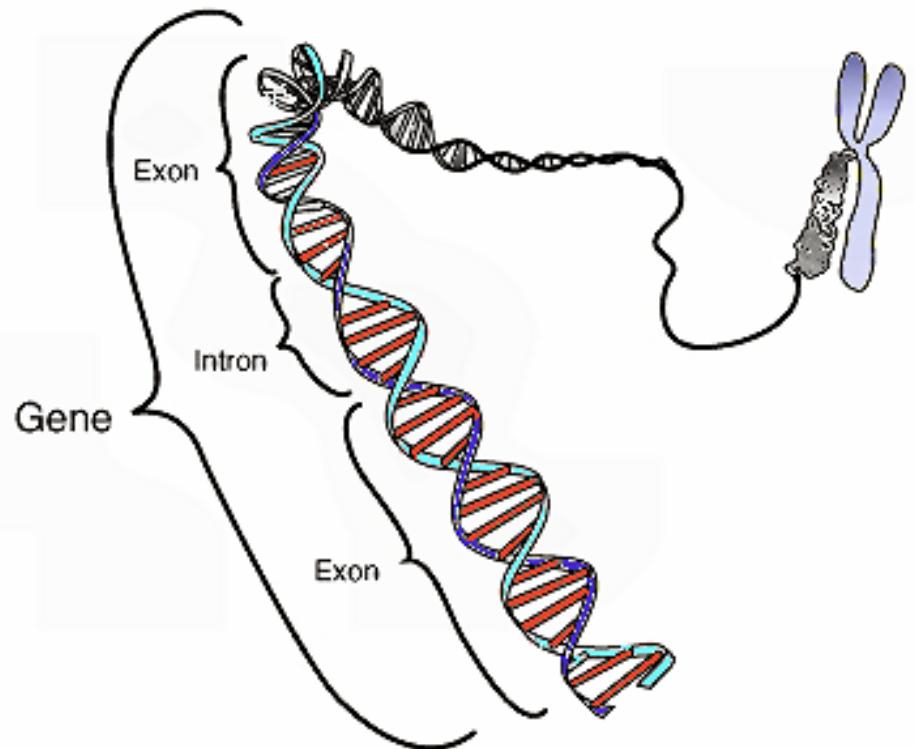


Genomes of different model organisms

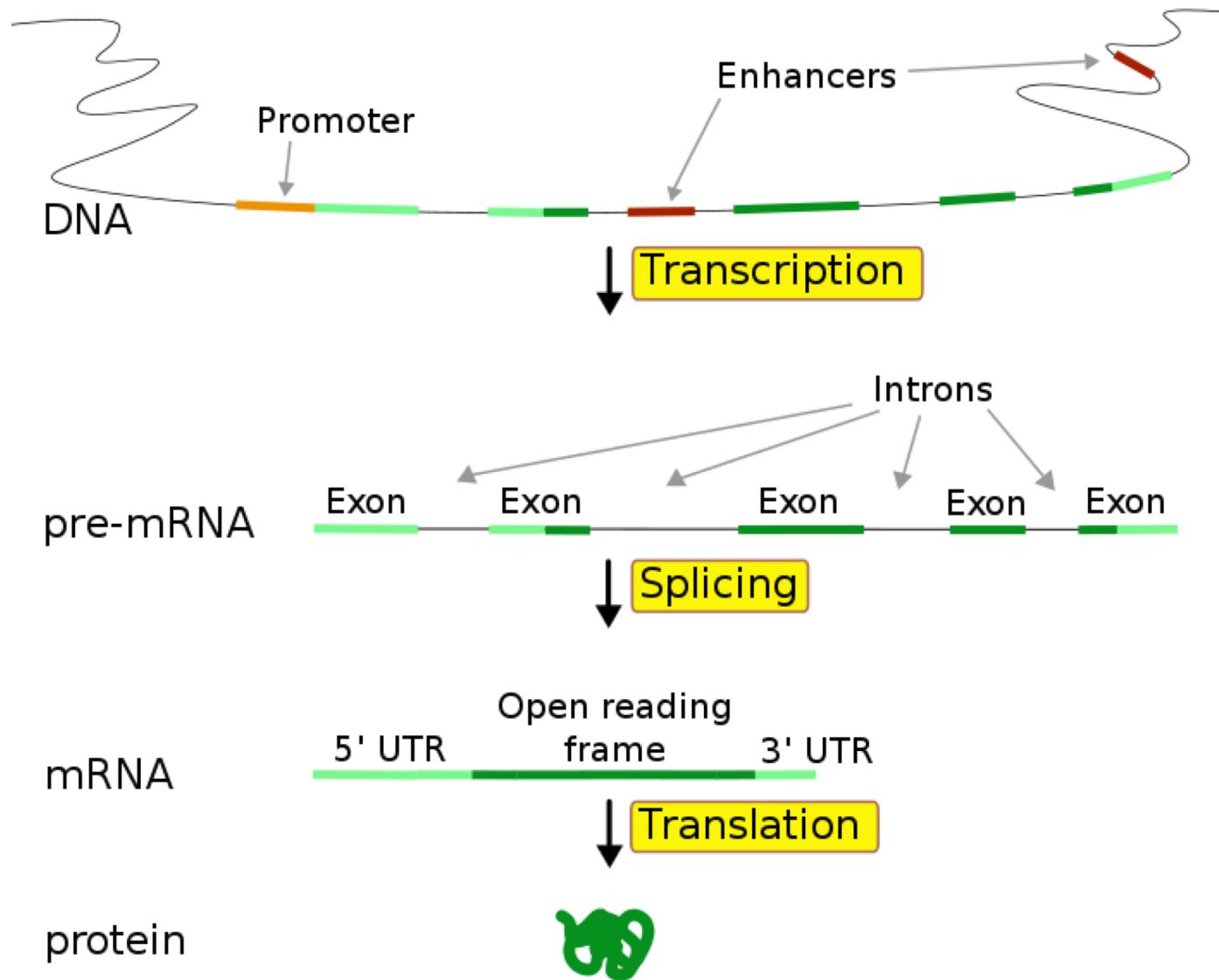
Organism	Genome size (bp)	# genes
E. coli	4.6M	4,300
S. cerevisiae (yeast)	12.5M	5,800
C. elegans (worm)	100M	20,000
A. thaliana (plant)	115M	28,000
D. melanogaster (fly)	123M	13,000
M. musculus (mouse)	3G	23,800
H. sapiens (human)	3.3G	25,000

Gene

- A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions.
- Or simply, a piece of “useful” DNA sequence.



Gene structure and splicing



- In a nutshell (for biostatisticians):
 - **enhancer**: a region for enhancing gene expression. Not necessarily close to the gene.
 - **promoter**: at the beginning of the gene, helps transcription.
 - **exons**: the “useful” part of the gene, will appear in the mRNA product.
 - **introns**: the “spacer” between exons, will NOT be in the mRNA product.
 - **splicing**: the process to remove introns and join exons.
 - **alternative splicing**: different splicing pattern for the same pre-mRNA. For example, mRNA could be from exons 1 and 2 or exons 1 and 3. Those are different “transcripts” of the same gene.

RNA (Ribonucleic acid)

- Similar to DNA, but
 - RNA is usually single-stranded.
 - The base U is used in place of T.
 - The backbone is different.
- Many different types: mRNA, tRNA, rRNA, miRNA, snoRNA, etc.

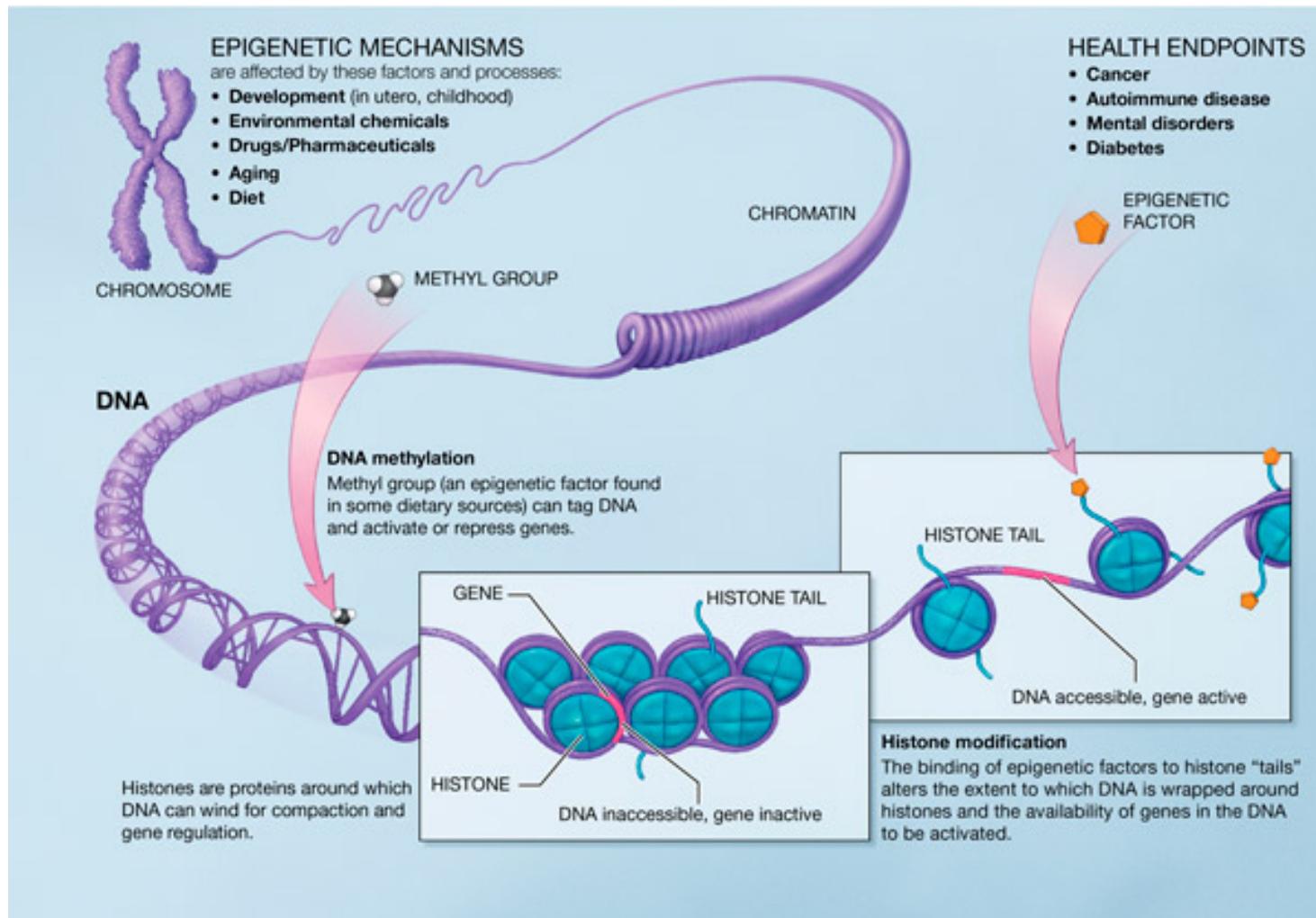
Protein

- The final product of gene expression process, workhorses in the cells.
- A chain of amino acid.
- Every 3 nucleotide is translated into one amino acid during translation.
- There are 20 types of amino acids, so a protein can be thought as a string from a 20-character alphabet.
- 3D protein structure is often important for its function.



Epigenetics

- Non-DNA sequence related, heritable mechanisms to control gene expressions. Examples: DNA methylation, histone modifications.



What is computational biology

- Use mathematical/statistical models to study biological mechanisms.
- Imagine biological system as a machine.
 - Bench biologists (“**web lab**”) perform experiments to collect data to measure the outputs of the machine.
 - Computational biologists (“**dry lab**”) make inferences about how the machine works based on data.

Examples of computational biology researches

- DNA sequence analysis:
 - sequence alignment and searching.
 - gene and motif finding.
 - evolution: phylogenetic trees.
- Transcriptional analysis:
 - compare gene expression by measuring mRNA quantity in different sample (expression microarrays, RNA-seq).
 - detect alternative splicing and gene fusion.
 - locational analysis: detect protein (transcription factor) binding or epigenetic modification.

- Epigenetics:
 - Detect, compare and characterize DNA methylation or histone modifications.
 - Epigenetic regulation of gene expression.
- Protein:
 - protein sequence alignment.
 - protein expressions (protein arrays).
- Joint analysis:
 - Jointly model multiple –omics data to decipher gene expression process or understand their relationships.
- Disease biomarker discovery.

A brief introduction to High-throughput experiments

High-throughput experiments

- Methods to conduct a large number of experiments simultaneously.
- Examples:
 - Microarrays.
 - Second generation sequencing.
 - Flow cytometry
 - ...
- Pros: quick, cheap.
- cons: lower accuracy, complicated data.

Microarray

- 2D array on a solid substrate that assays large amount of biological materials.
- Examples of microarrays:
 - DNA microarray:
 - Gene expression array.
 - SNP array.
 - Tiling arrays (ChIP-chip, array CGH).
 - Methylation array.
 - Protein microarray
 - Others ...

DNA microarrays

- A collection of many spots, each has a certain type of probes (short segments of DNAs).
- Detect and quantify target sequence (e.g., mRNA) by **hybridization**: sequence-specific interaction between two complementary strands of nucleic acid.
 - An example:

ATCGATTGAGCTCTAGCG

TAGCTAACTCGAGATCGC

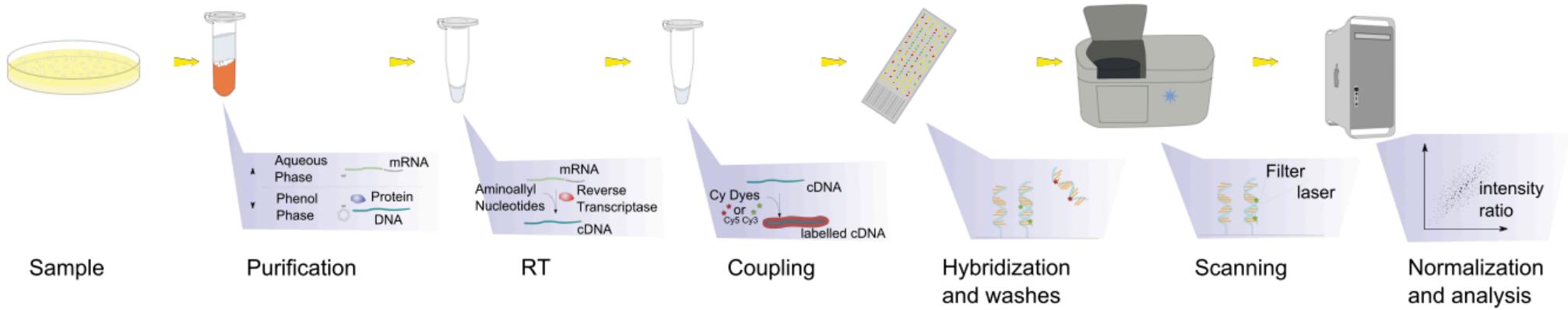
- DNA segments in the sample will stick to probes with complementary sequences.
- Each probe has a reading (intensity), which measures the **RELATIVE** amount of target sequence in the sample.

Gene expression microarray

- Measure the gene expressions by the amount of mRNA.
- Each gene is targeted by many probes.
- Major manufacturers:
 - Affymetrix
 - Illumina
 - Nimblegen (now acquired by Roche).



GE microarray procedures



- Data: a fluorescent intensity value (a non-negative floating-point number) for each probe.
- Goal: find genes that are differentially expressed (produce different amount of mRNAs) among samples.

Statistical challenges for expression arrays

- Data normalization, transformation, and summarization.
- Statistical inferences: tests for DE (differentially expressed) genes.
- Pattern recognition, e.g., clustering.
- Biological/clinical implications.

DNA sequencing

- Technologies to determine the nucleotide bases from a DNA molecule.
- Traditional method: Sanger sequencing.
 - slow (low throughput) and expensive: took Human Genome Project (HGP) 13 years and \$3 billion to sequence the entire human genome.
 - Relatively accurate.

Next-generation sequencing (NGS)

- Aka: high-throughput sequencing, second-generation sequencing.
- Able to sequence large amount of short sequence reads in a short period:
 - high throughput: hundreds of millions sequences in a run.
 - Cheap: sequence entire human genome costs a few thousand dollars.
 - short read length: up to several hundred bps.

NGS Applications

- **DNA-seq:** sequence the genomic DNA in order to find variants or assemble reference genome.
- **RNA-seq:** sequence the transcriptome (mRNA -> cDNA) in order to measure gene expressions or detect alternative splicing/gene fusion.
- **MeDIP/ChIP-seq:** detect protein-DNA binding or epigenetic modification sites.
- **BS-seq:** Single nucleotide resolution DNA methylation.
- Basically everything microarrays can do.

Single cell sequencing

- Traditional “bulk” sequencing: measures the average signal of many cells, which ignores the inter-cellular heterogeneities:
 - Difference among different cell types.
 - Biological variation among the same type of cell.
- Single cell sequencing:
 - Experiment is performed for each cell individually.
 - Provides more detailed, higher resolution information.

Available platforms

- Major player:
 - Illumina: Genome Analyzer, HiSeq, MiSeq
 - LifeTech: SOLiD, IonTorrent
- Others:
 - Oxford Nanopore
 - Pacific Bioscience

Statistical challenges for second generation sequencing data

- Sequence alignment.
- Data transformation and normalization.
- Goal specific:
 - RNA-seq: differential gene/isoform expression or splicing, new gene/exon discovery.
 - ChIP-seq: peak detection, differential peak,
 - BS-seq: differential methylation.

R and Bioconductor

R programming language

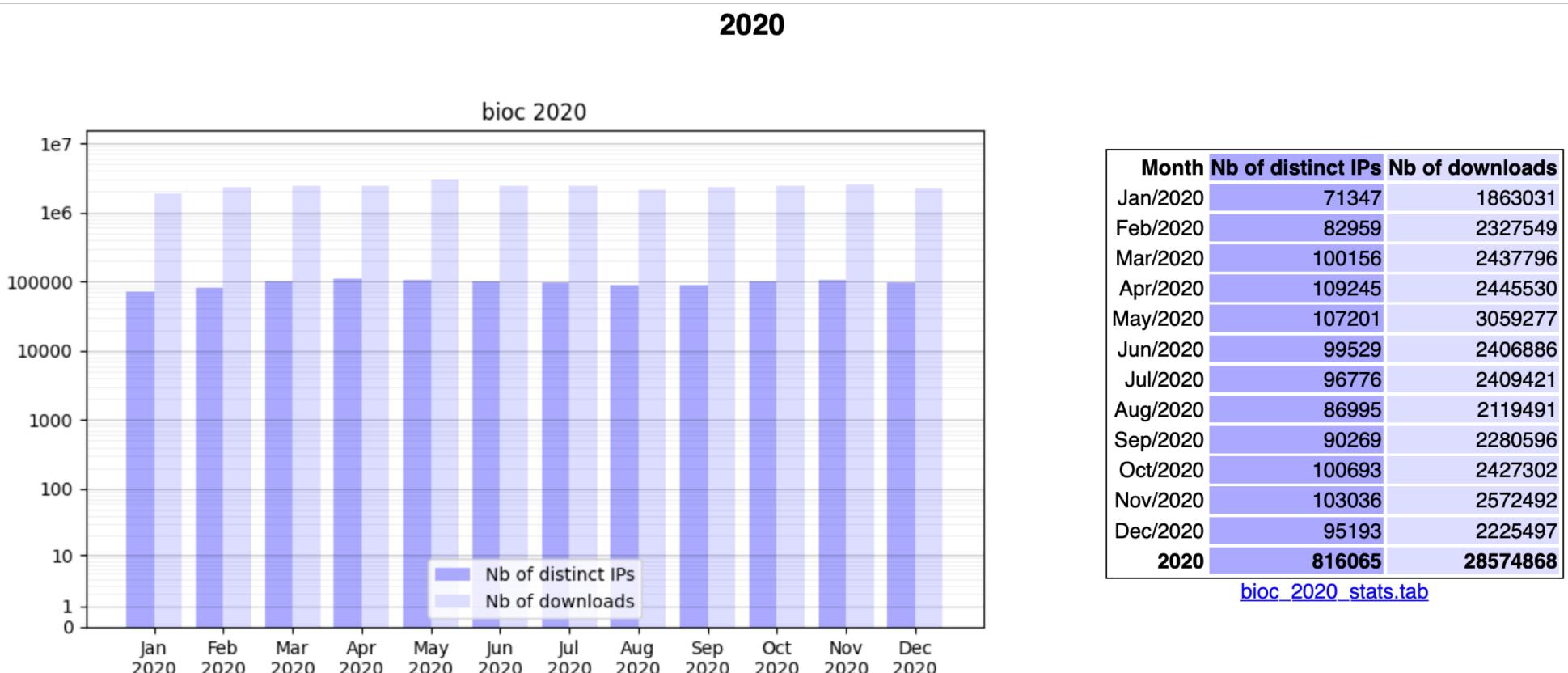
- THE programming language and environment for statisticians.
- Free and open source.
- Easy and intuitive.
- Contains a large collection of add-on “packages”.
- Provides extensive graphics capabilities and interfaces to lower level languages (C, Fortran, etc.)
- Evolve rapidly: several (7) years ahead of SAS.
- Relatively slow, but with easy interfaces with other languages.
- Visit www.r-project.org to download/install R and reference manuals.

R environment

- You can write programs in a text editor, and copy/paste into R console.
- IDEs (Integrated Development Environment) available (such as **Rstudio**), which are much more convenient.
- For geeks, I recommend using emacs with ESS. See <http://www.biostat.wisc.edu/~kbroman/Rintro/> for details.

Bioconductor: a collection of R packages

- Started by Rob Gentleman (Fred Hutch), with a few junior(at that time) faculty members.
- Becoming the *de facto* language for genomic data analysis.



Functionalities

- “*Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker images](#).*”
- Currently (Aug 2021) provides 3500+ packages for:
 - microarrays.
 - second generation sequencing.
 - other high-throughput assays.
 - annotation and data.
- Most of the packages are contributed.

Bioconductor installation

- Use `BiocManager::install()`.
- Basic installation: installing default (core) packages:

```
if (!requireNamespace("BiocManager"))
  install.packages("BiocManager")
BiocManager::install()
```

- Installing a specific package:
`BiocManager::install("limma")`
- Upgrading also use `BiocManager::install()`
 - It's a good habit to upgrade bioconductor periodically.

Online resources: genome browser and public data repositories

UCSC Genome Browser

- Initially developed by Jim Kent on 2000 when he was a Ph.D. student.
- Host genomic annotation data for many species.
- The genome browser is a graphical viewer for visualizing genome annotations.
- Provide other tools for genomic data analysis and interfaces for querying the database.

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr1:11,102,837-11,267,747 164,911 bp.

enter position, gene symbol, HGVS or search terms

go



Scale
chr1:

11,150,000

50 kb | Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments

11,200,000 | hg38
11,250,000

Alt Haplotypes

MTOR
MTOR

GENCODE v29 Comprehensive Transcript Set (only Basic displayed by default)

RNU6-537P

11,150,000 | Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments

ANGPTL7 RNU6-291P

MTOR
MTOR-AS1
ANGPTL7

NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Annotation Release 109 (2018-03-29)

RNU6-537P

MTOR

MTOR-AS1

UCSC annotations of RefSeq RNAs (NM_* and NR_*)

ANGPTL7 RNU6-291P

OMIM Alleles

OMIM Allelic Variants

MTOR

Gene Expression in 53 tissues from GTEx RNA-seq of 6555 samples (570 donors)

Layered H3K27Ac

MTOR-AS1
RNU6-537P

ANGPTL7

RNU6-291P

RPL39P6

H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

DNase Clusters

4.88

DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)

100 vertebrates Basewise Conservation by PhyloP

Cons 100 Vertebr

-4.5

Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey

Multiz Alignments of 100 Vertebrates

Common SNPs(151)

Simple Nucleotide Polymorphisms (dbSNP 151) Found in >= 1% of Samples

Repeating Elements by RepeatMasker

SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Other genome browsers/databases

- General:
 - NCBI Map Viewer
 - Ensemble genome browser
- Other species specific genome browser
 - MGI: Mouse genome informatics
 - wormbase, Flybase, SGD (yeast), TAIR DB (arabidopsis), microbial genome database
- More or less the same, pick your favorite one.

Public high-throughput data repositories

- **GEO:** Gene expression omnibus.
 - Host array- and sequencing-based data.
- **ArrayExpress:** European version of GEO.
 - Better curated than GEO but has less data.
- **SRA:** sequence read archive.
 - Designed for hosting large scale high-throughput sequencing data, e.g., high speed file transfer.

Data are required to be deposited in one of the databases when paper is accepted!

Other public data resources

- TCGA (The Cancer Genome Atlas) data portal (<https://portal.gdc.cancer.gov>):
 - Host data generated by TCGA, a big consortium to study cancer genomics.
 - Huge collection of cancer related data: different types of genomic, genetic and clinical data for many different types of cancers.
- ENCODE (the **EN**Cyclopedia **O**f **D**N**A** **E**lements) data coordination center (<https://www.encodeproject.org/about/data-access>):
 - Host data generated by ENCODE, a big consortium to study functional elements of human genome.
 - Rich collection of genomic and epigenomic data.
- Many others: GTEx, Roadmap Epigenomics, Human Cell Atlas, UK biobank, etc.

To do list after this class

1. Review slides.
2. Read wikipedia pages for DNA, gene, genome, DNA microarray and DNA sequencing.
3. Install R and Bioconductor on your computer.
4. Start to learn R by reading “R for beginners”:
http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf