Lecture 10: Introduction to Genomics

Announcements

- Upcoming deadlines:
  - A2 due Tue Nov 1
  - A3 also released Tue, due Tue Nov 15
  - Midterm: In class, Mon Nov 7
    - 80 minutes
    - 1 page 8.5” x 11’’ of notes allowed (back and front)
    - No calculators allowed or needed
    - Covers material through “Genomics: Introduction”
    - Practice midterm released on Ed

Some biology basics: starting from DNA

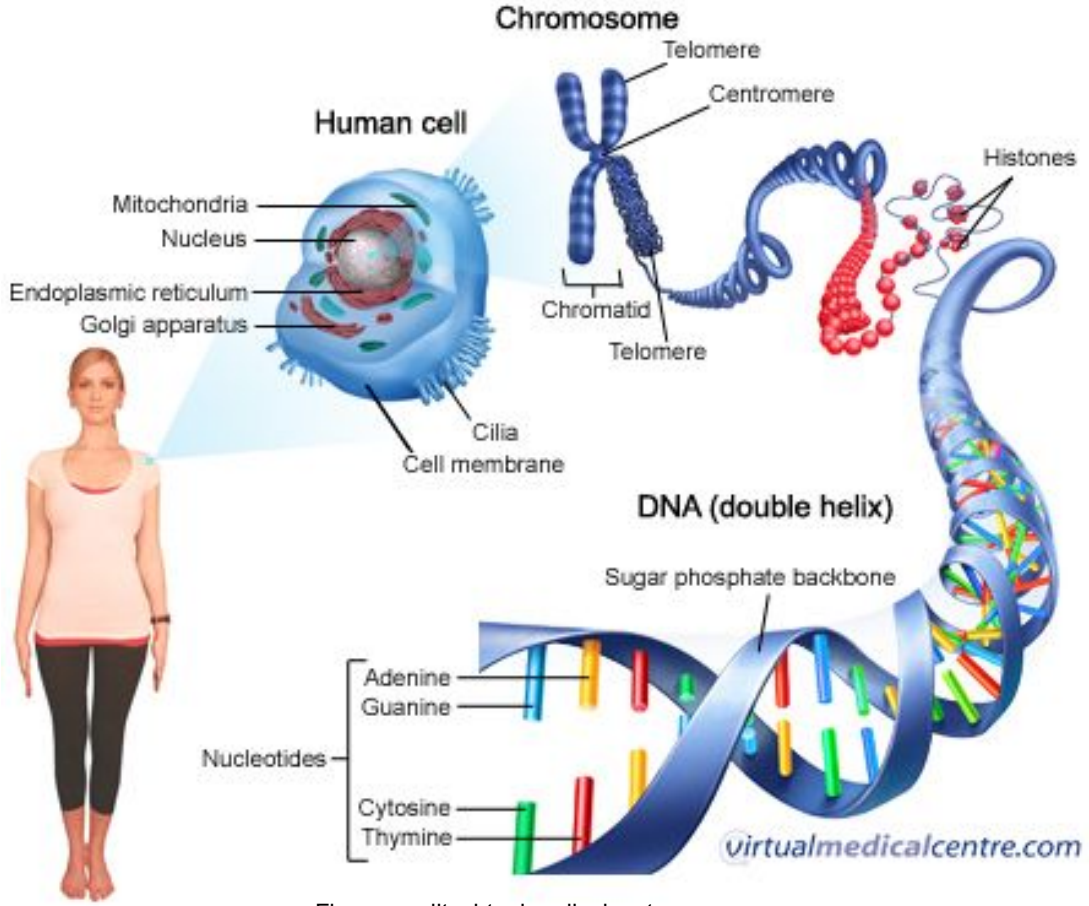


Figure credit: virtualmedicalcentre.com

Some biology basics: starting from DNA

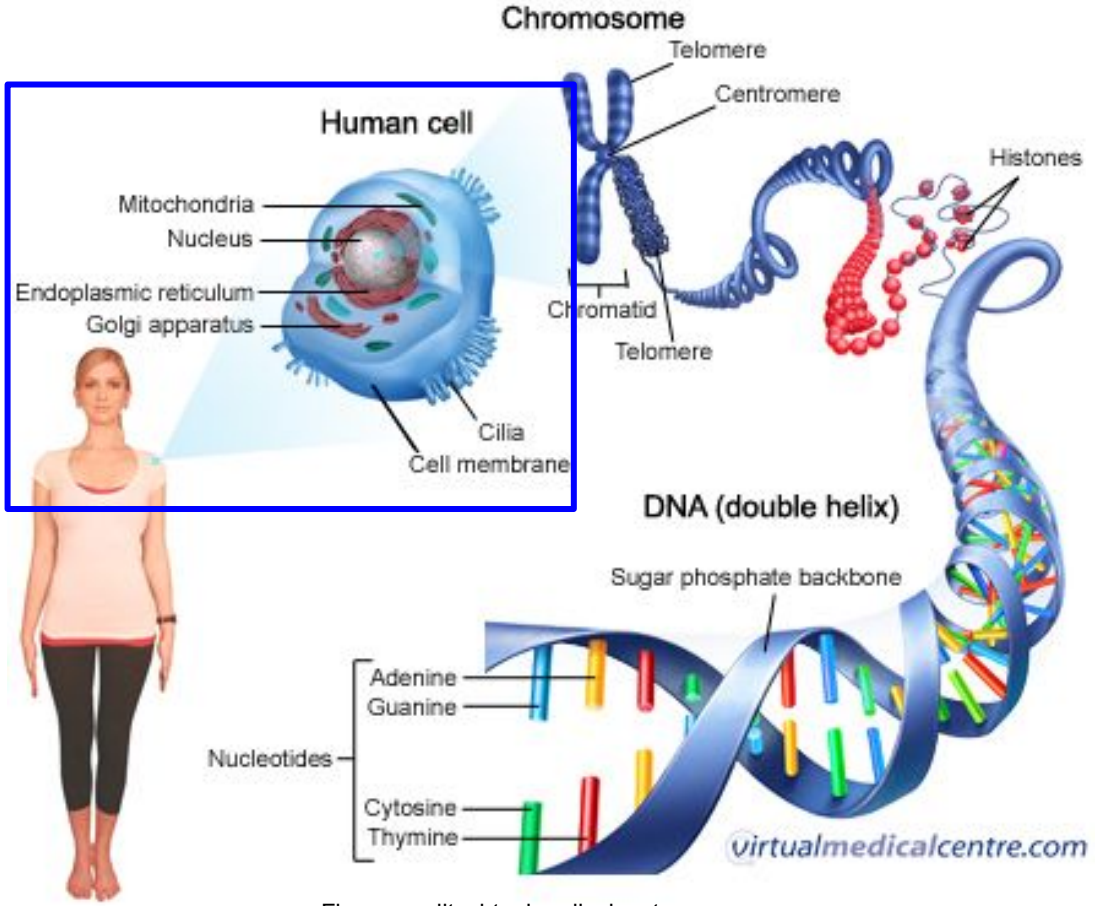


Figure credit: virtualmedicalcentre.com

Some biology basics: starting from DNA

~ 37 trillion cells in the human body

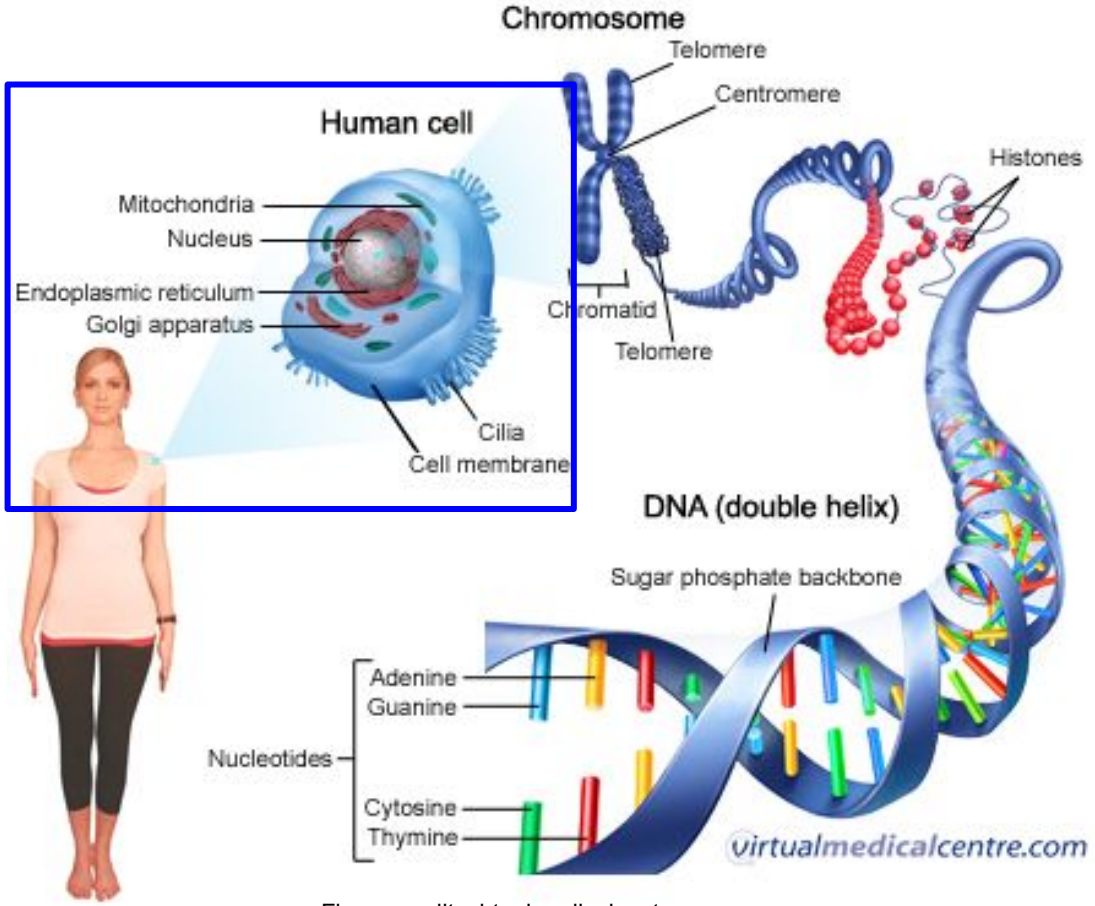


Figure credit: virtualmedicalcentre.com

Some biology basics: starting from DNA

Nucleus: “brain of the cell”. Contains genetic material in the form of DNA.

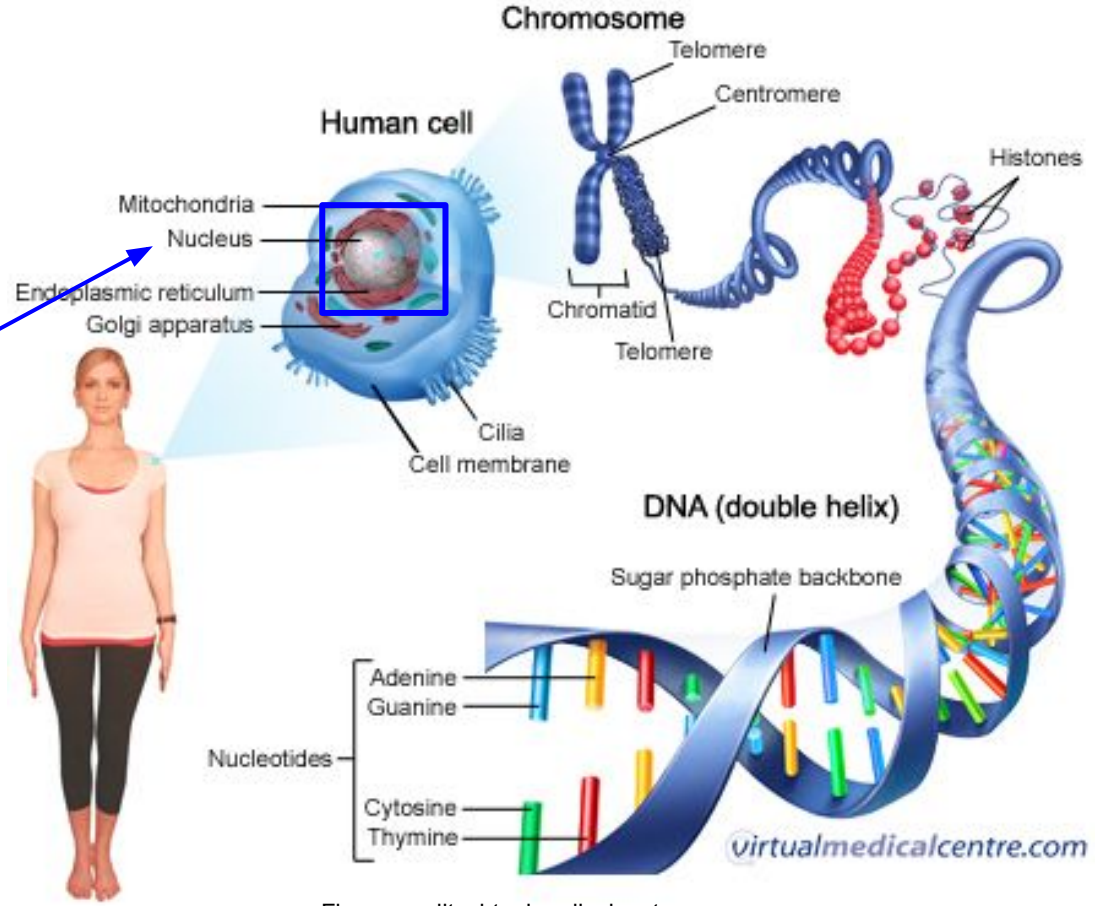


Figure credit: virtualmedicalcentre.com

Some biology basics: starting from DNA

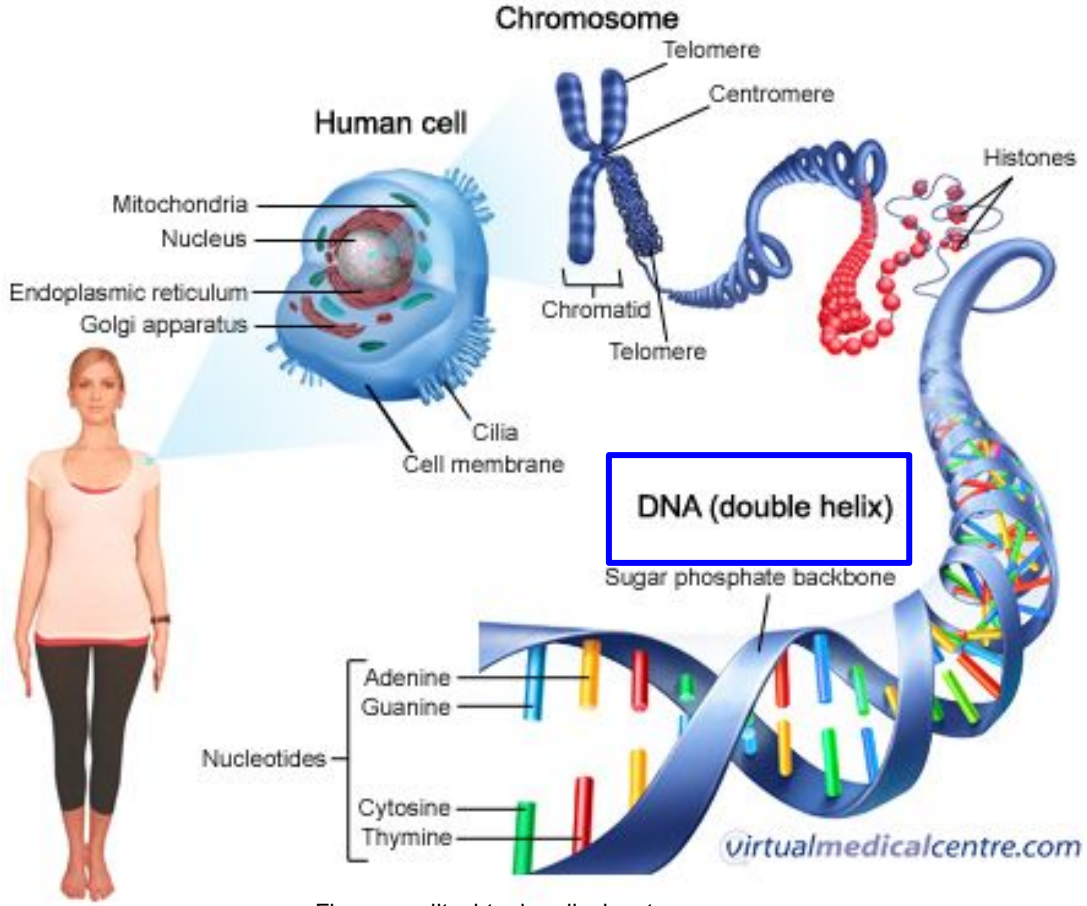


Figure credit: virtualmedicalcentre.com

Some biology basics: starting from DNA

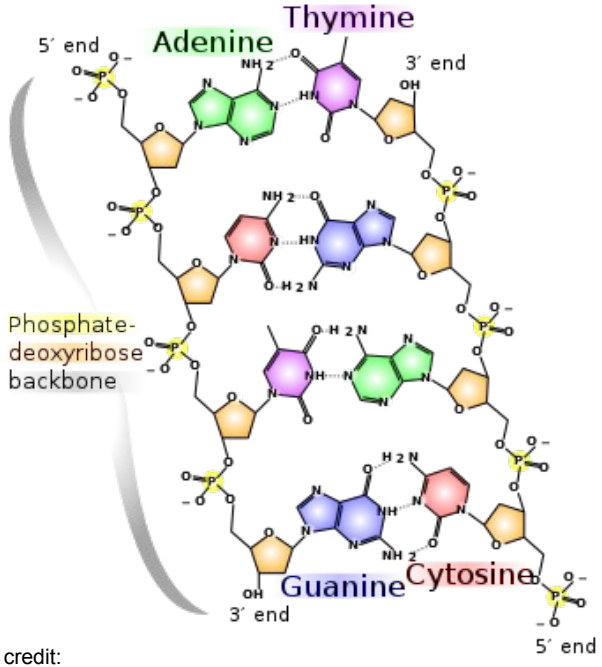


Figure credit: https://en.wikipedia.org/wiki/Nucleobase#/media/File:DNA_chemical_structure.svg

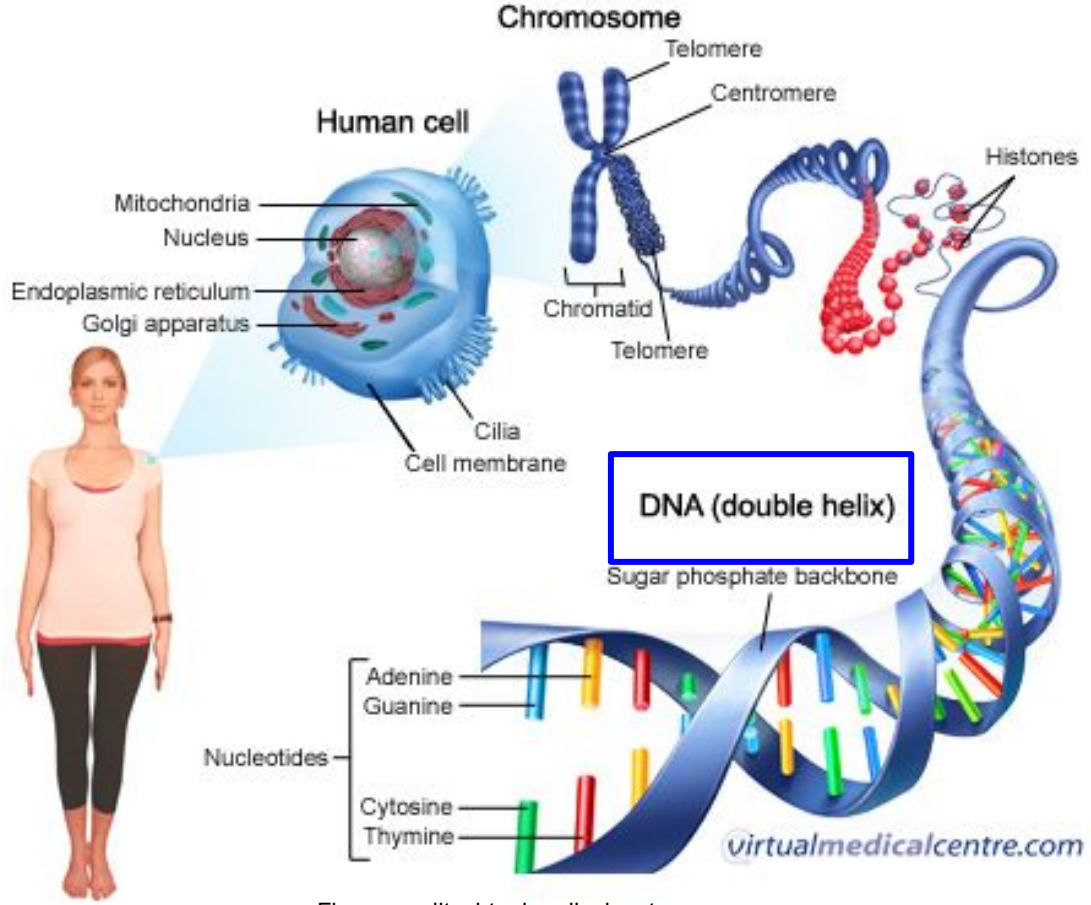


Figure credit: virtualmedicalcentre.com

Chromosomes and genes

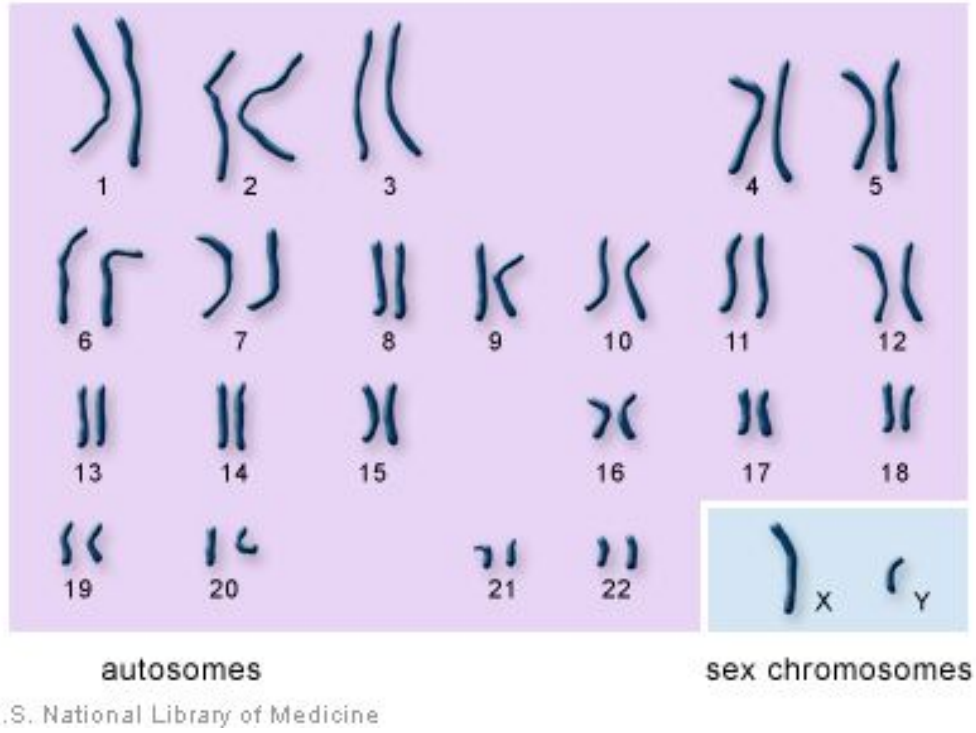


Figure credit: https://ghr.nlm.nih.gov/primer/illustrations/chromosomes.jpg

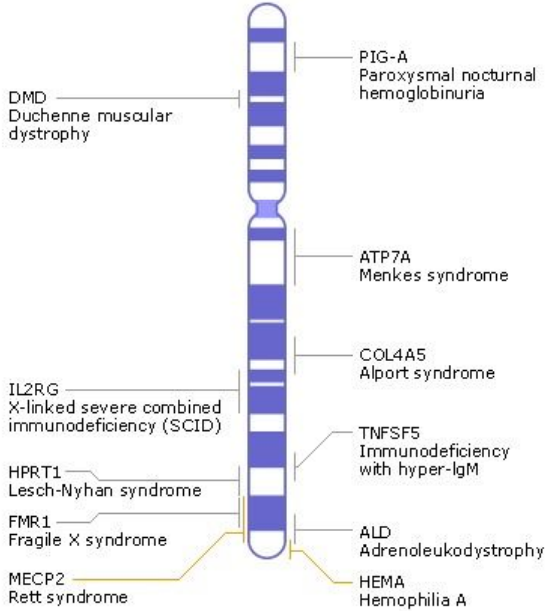


Figure credit: https://www.ncbi.nlm.nih.gov/books/NBK22266/bin/a01chr.jpg

Chromosomes and genes

23 pairs of chromosomes (22 autosomes + sex chromosomes)

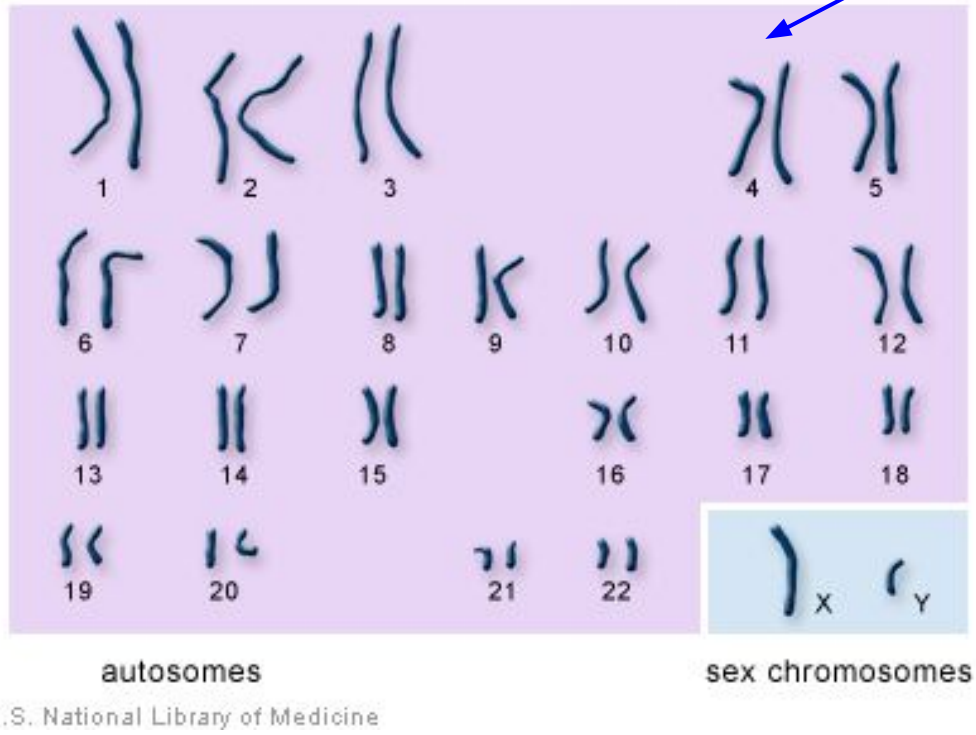


Figure credit: https://ghr.nlm.nih.gov/primer/illustrations/chromosomes.jpg

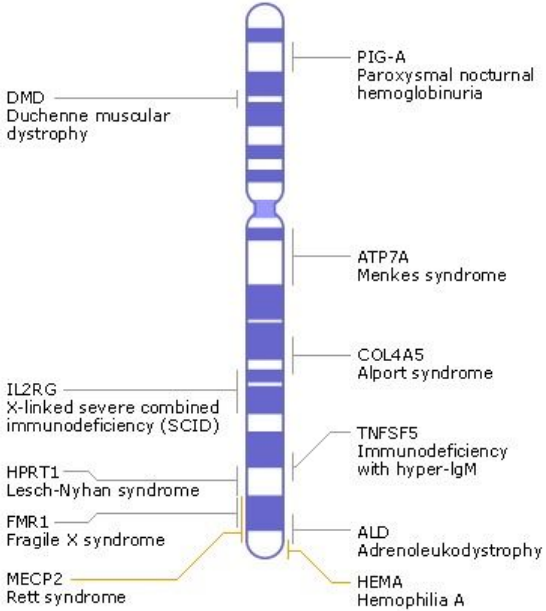


Figure credit: https://www.ncbi.nlm.nih.gov/books/NBK22266/bin/a01chr.jpg

Chromosomes and genes

Genes: segments of DNA within chromosomes

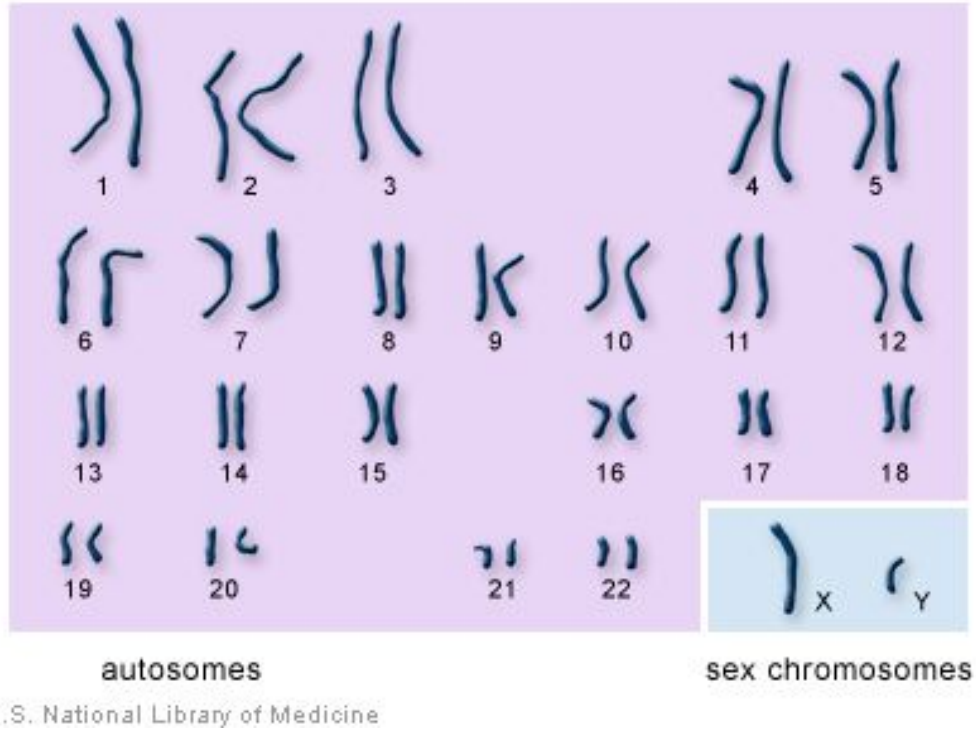


Figure credit: https://ghr.nlm.nih.gov/primer/illustrations/chromosomes.jpg

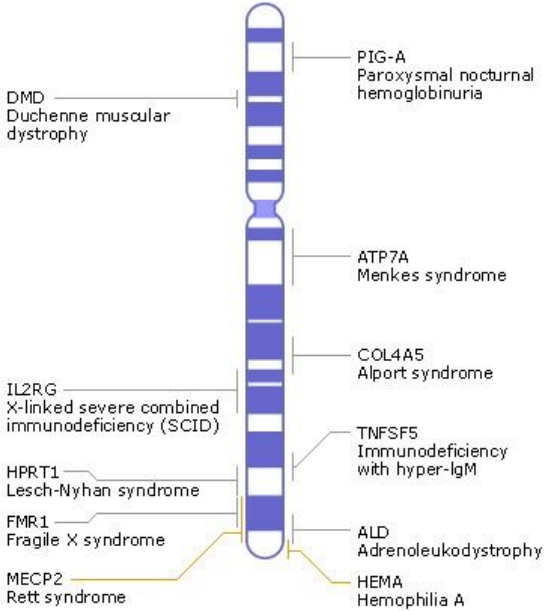


Figure credit: https://www.ncbi.nlm.nih.gov/books/NBK22266/bin/a01chr.jpg

Chromosomes and genes

Genes: segments of DNA within chromosomes

Genes provide code for proteins

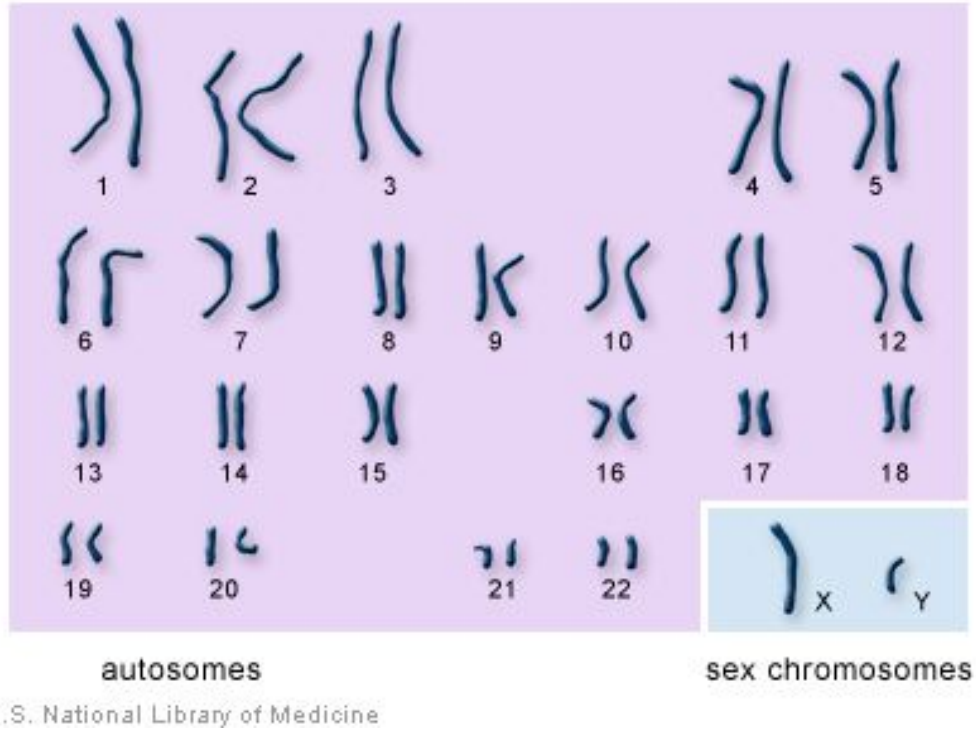


Figure credit: https://ghr.nlm.nih.gov/primer/illustrations/chromosomes.jpg

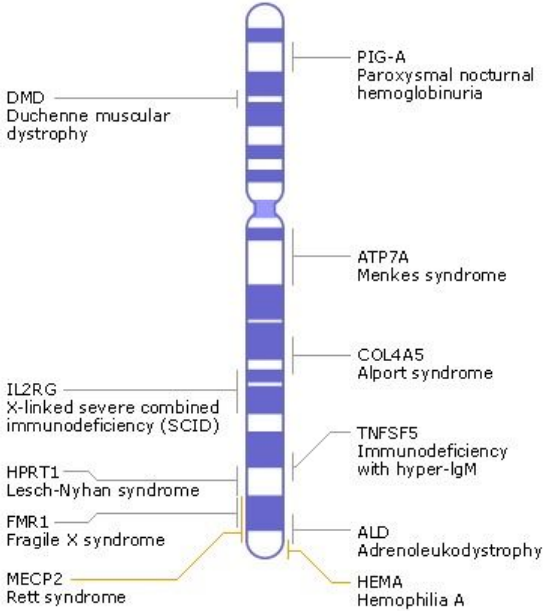


Figure credit: https://www.ncbi.nlm.nih.gov/books/NBK22266/bin/a01chr.jpg

Chromosomes and genes

Genes: segments of DNA within chromosomes

Genes provide code for proteins
But 99% of genes are “non-coding!”

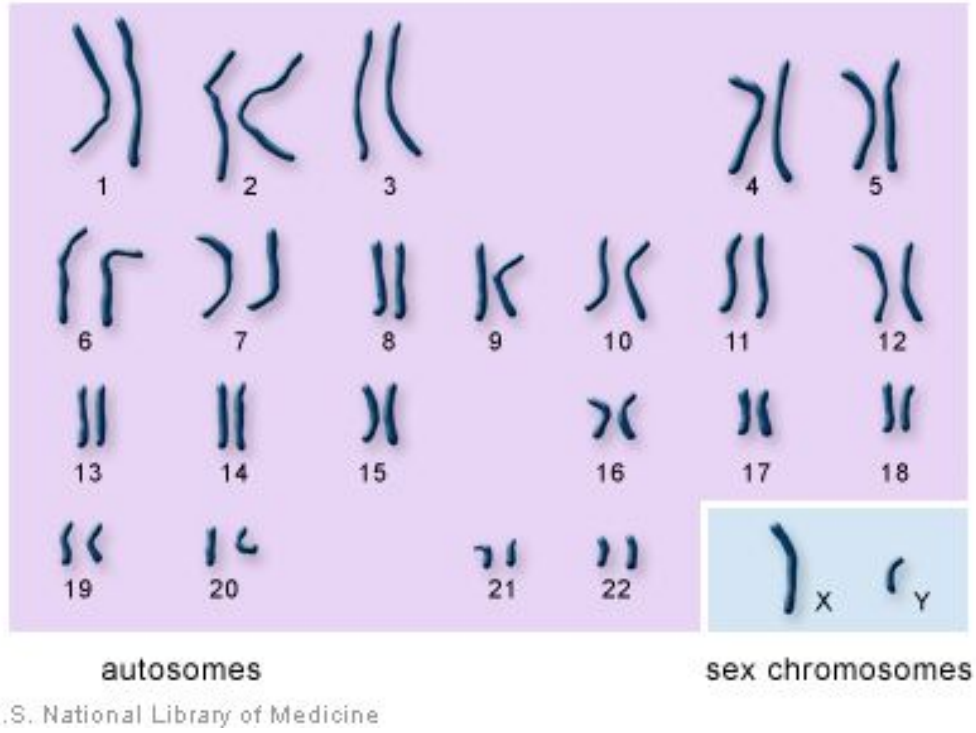


Figure credit: https://ghr.nlm.nih.gov/primer/illustrations/chromosomes.jpg

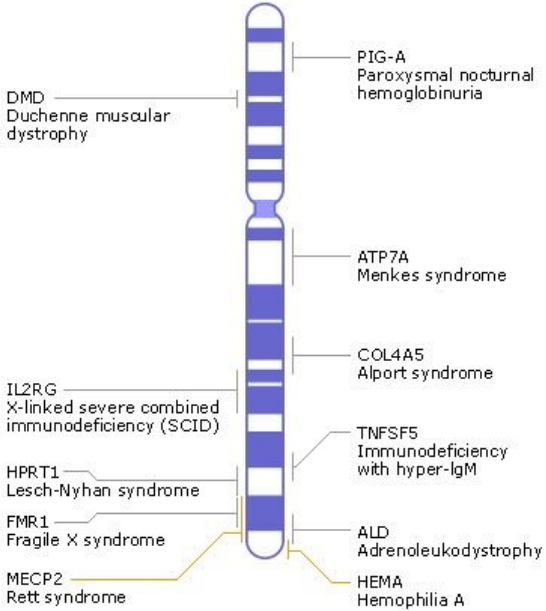
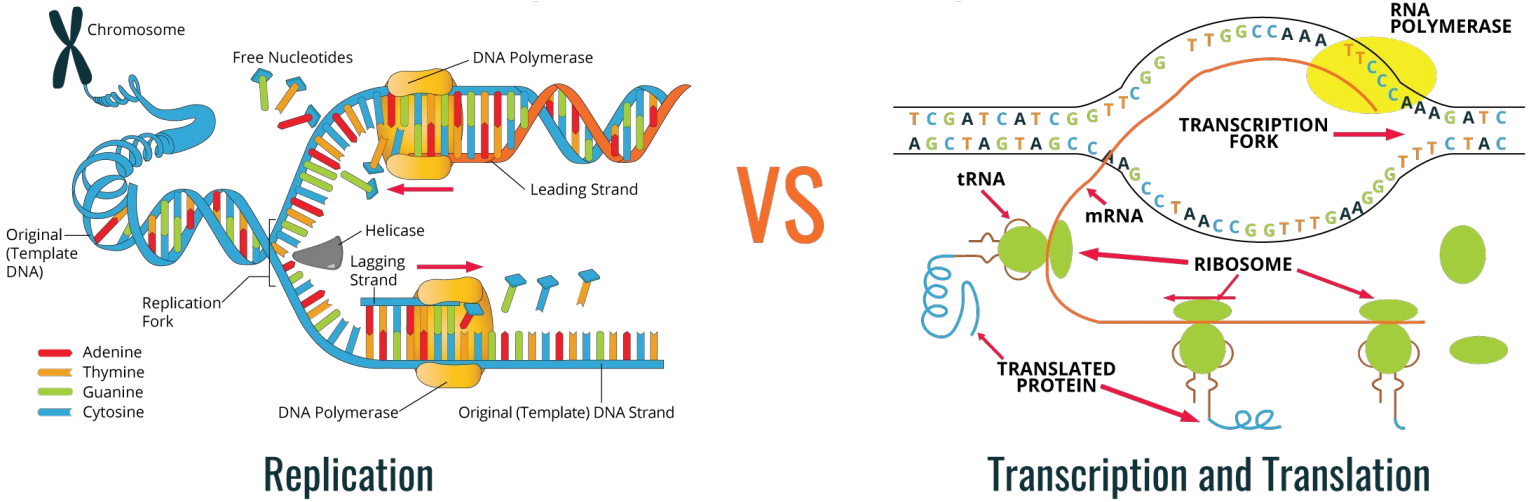


Figure credit: https://www.ncbi.nlm.nih.gov/books/NBK22266/bin/a01chr.jpg

DNA replication and transcription

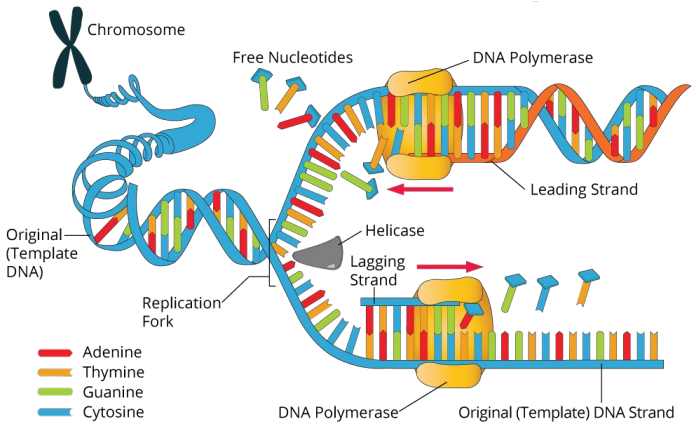


Replication

Transcription and Translation

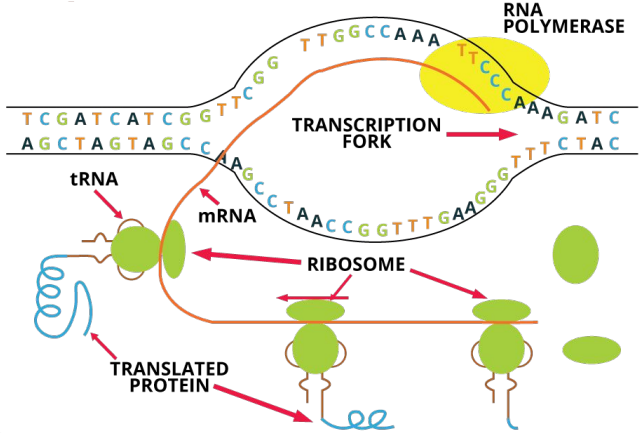
Figure credit:
https://www.bosterbio.com/media/images/MB_Replication_and_Transcription.png

DNA replication and transcription



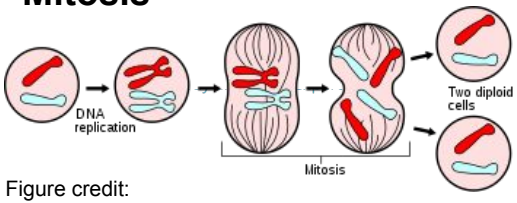
Replication

VS



Transcription and Translation

Mitosis



Meiosis

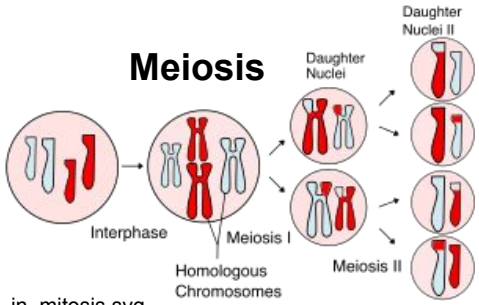
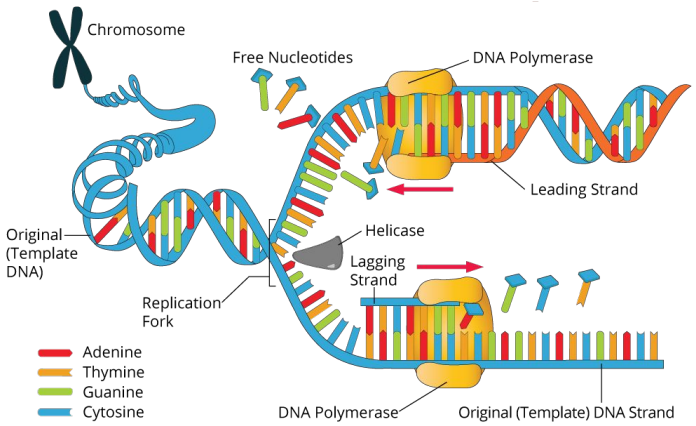


Figure credit:
https://en.wikipedia.org/wiki/Mitosis#/media/File:Major_events_in_mitosis.svg
https://en.wikipedia.org/wiki/Meiosis#/media/File:Meiosis_Overview_new.svg

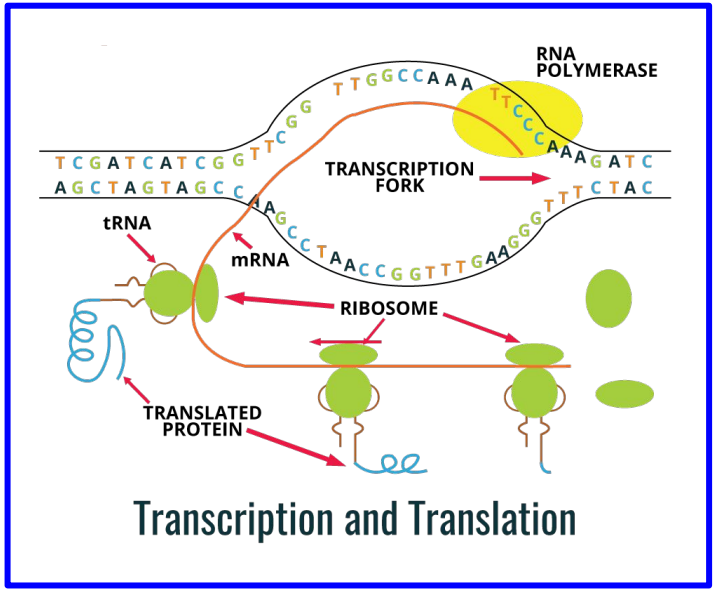
Figure credit:
https://www.bosterbio.com/media/images/MB_Replication_and_Transcription.png

DNA replication and transcription



Replication

VS



Transcription and Translation

Figure credit:
https://www.bosterbio.com/media/images/MB_Replication_and_Transcription.png

Transcription and translation

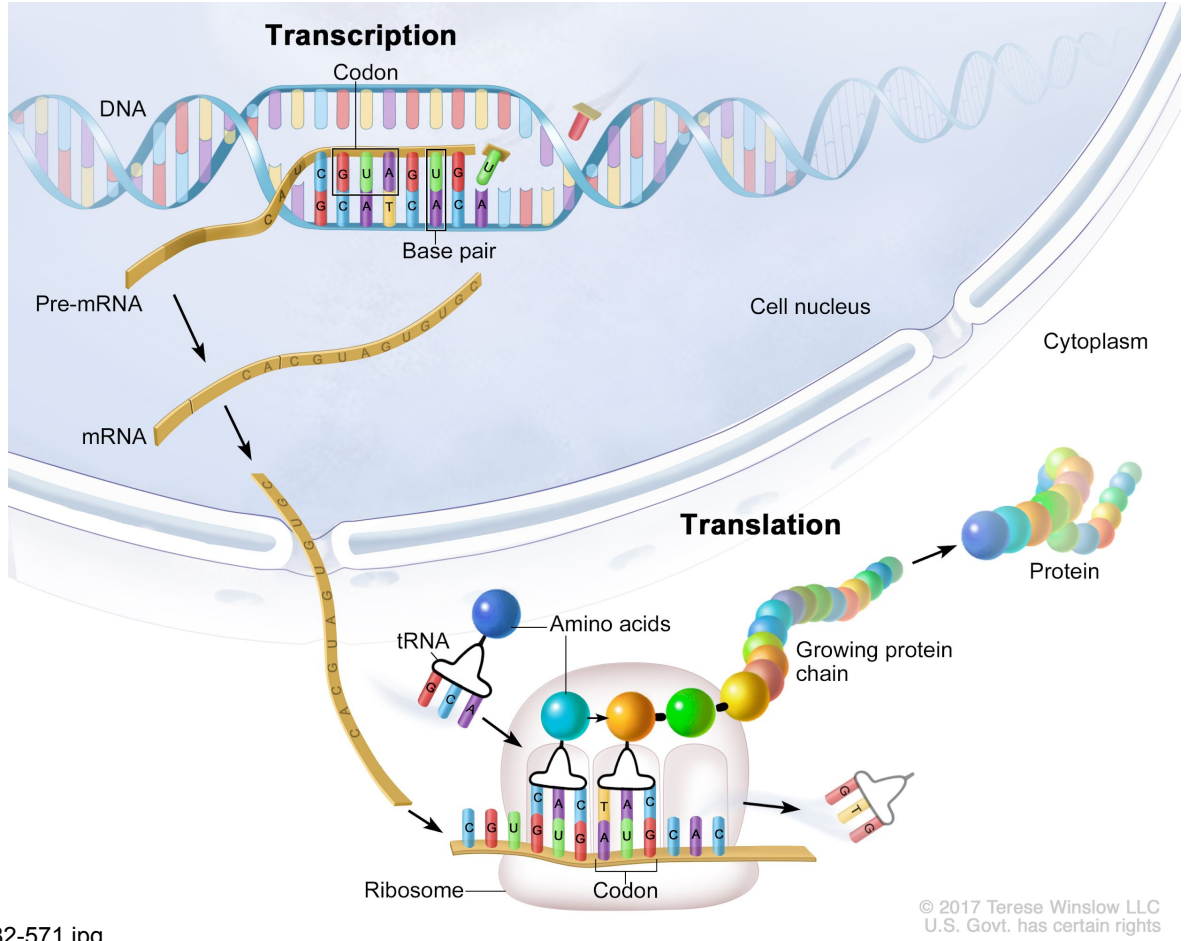


Figure credit: https://www.cancer.gov/images/cdr/live/CDR761782-571.jpg

Transcription and translation

Transcription: DNA -> RNA

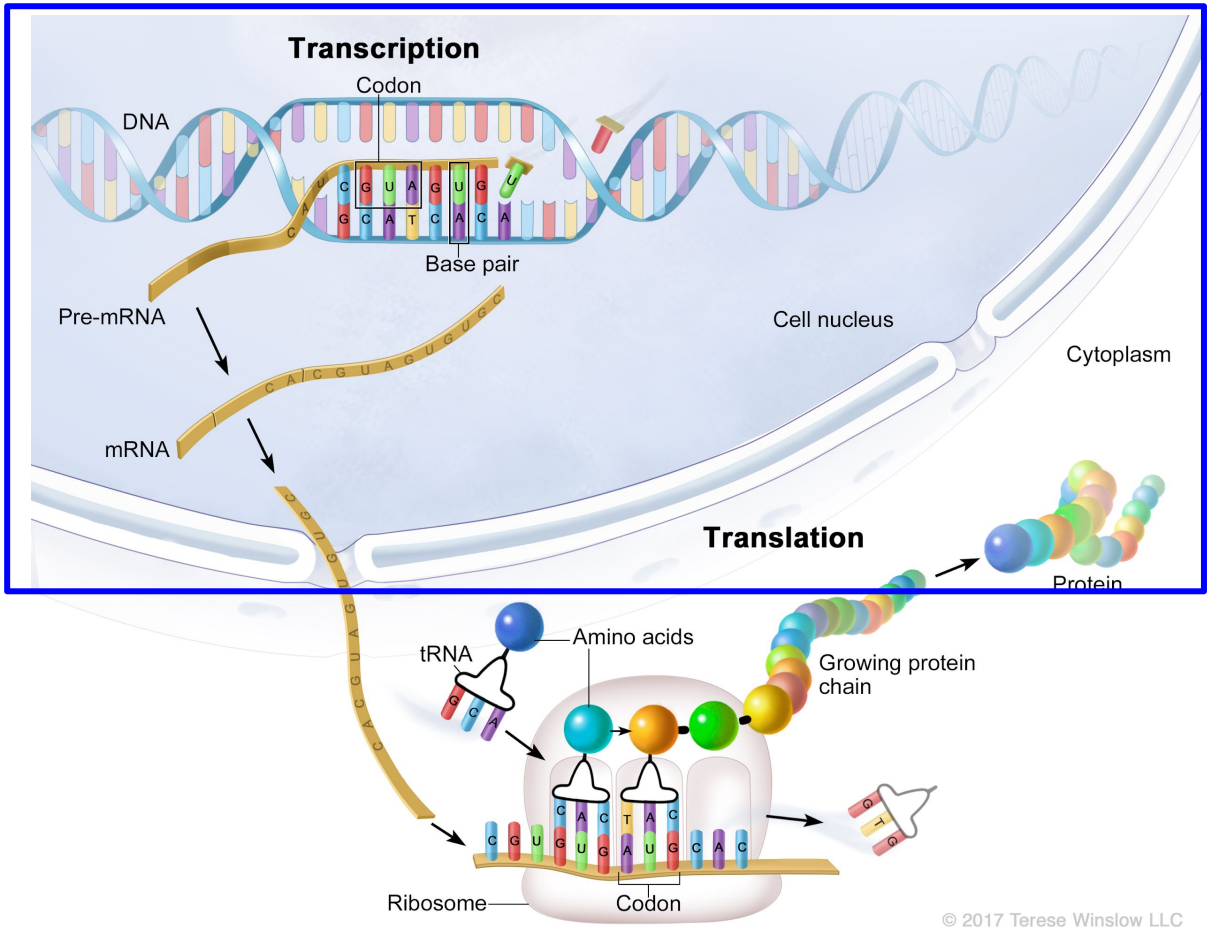


Figure credit: https://www.cancer.gov/images/cdr/live/CDR761782-571.jpg

Transcription and translation

Transcription: DNA -> RNA

Translation: RNA -> Protein

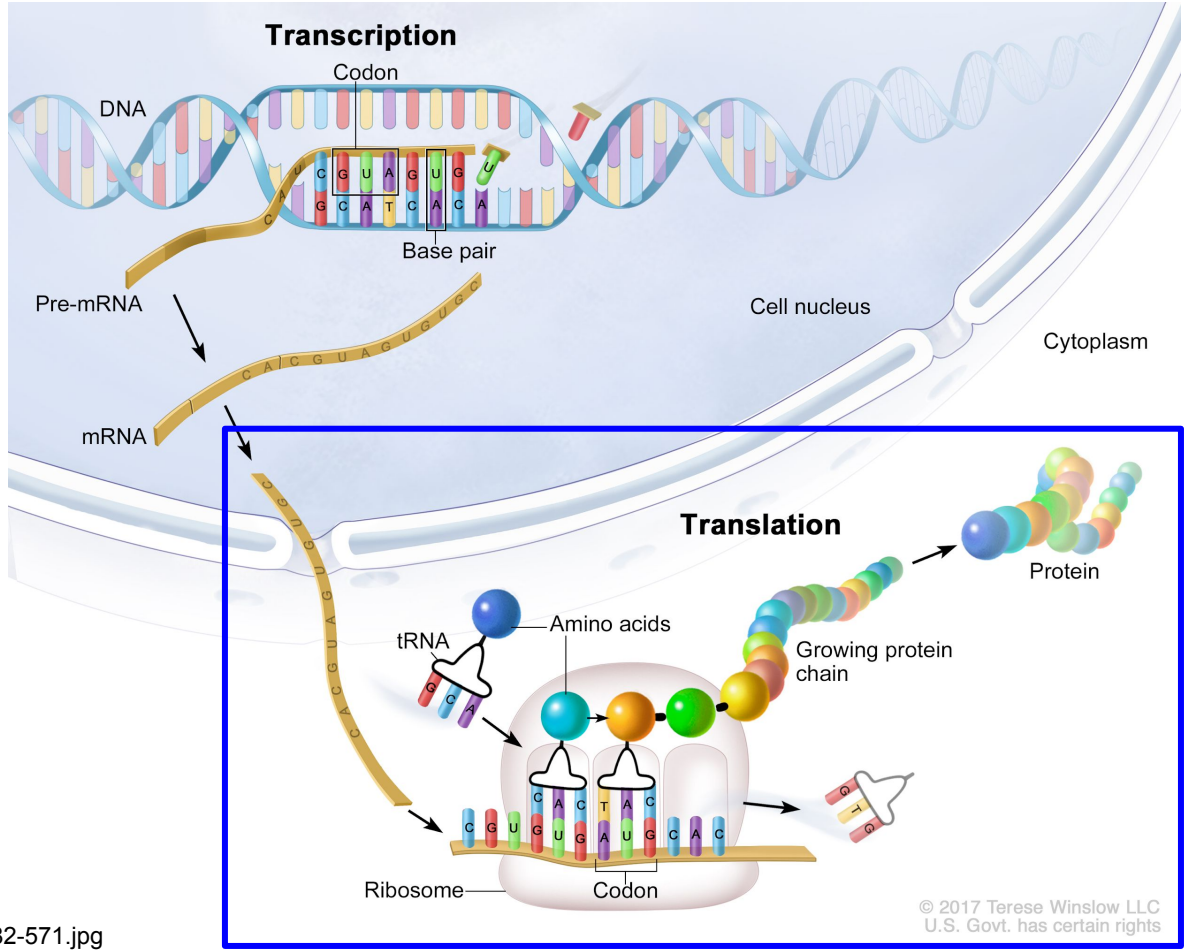
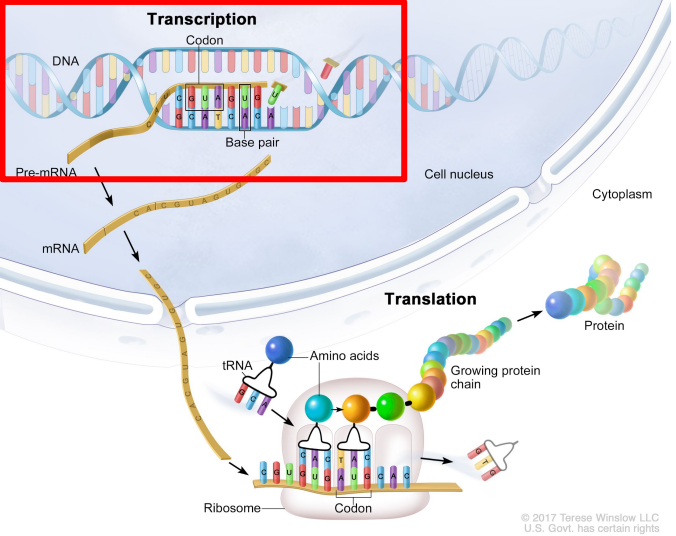


Figure credit: https://www.cancer.gov/images/cdr/live/CDR761782-571.jpg

DNA -> Pre-mRNA



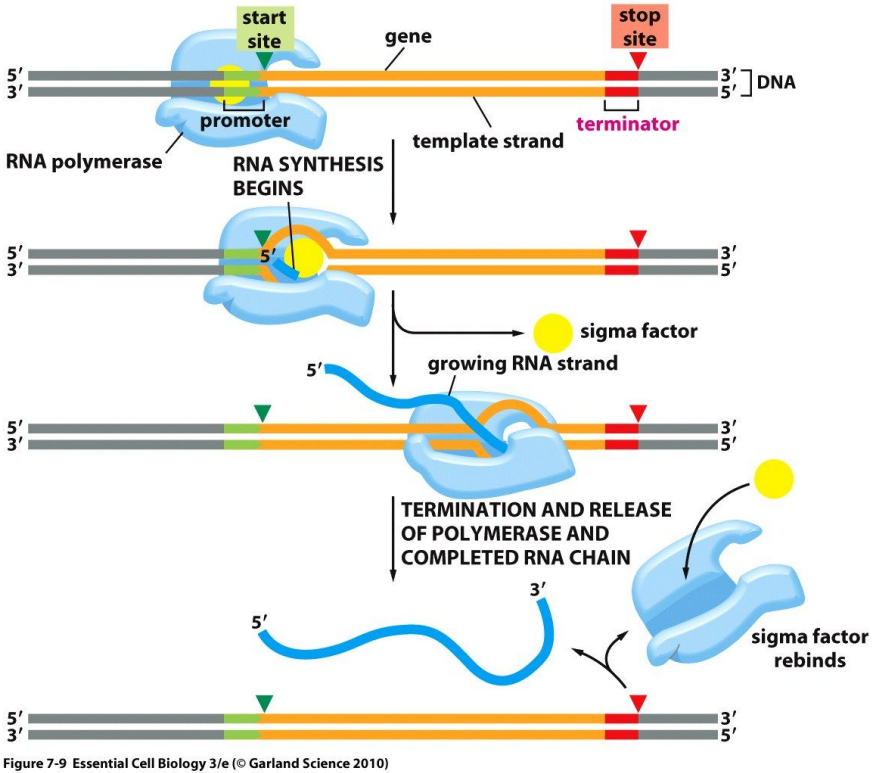
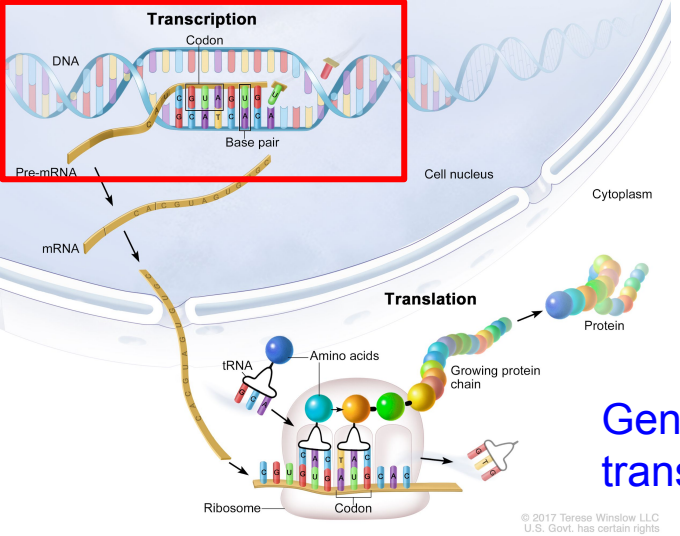


Figure 7-9 Essential Cell Biology 3/e (© Garland Science 2010)

Figure credit:
http://u18439936.onlinehome-server.com/craig.milgrim/Bio230/Outline/ECBFigures_Tables/Chapter_7/FigureJPGs/figure_07_09.jpg

DNA -> Pre-mRNA



Gene to transcribe

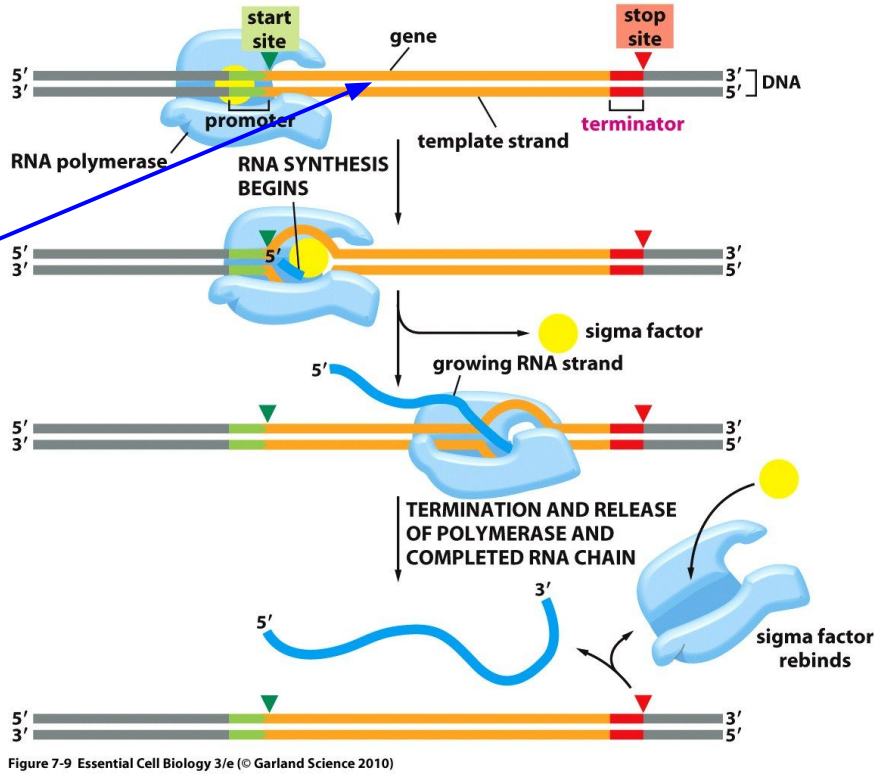
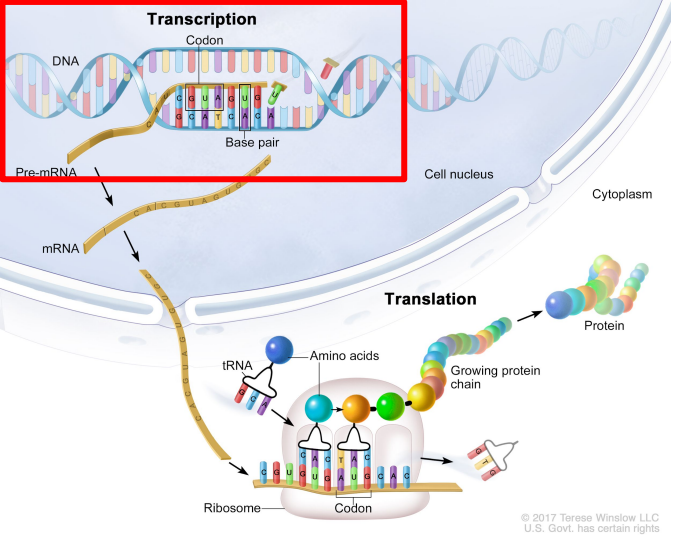


Figure 7-9 Essential Cell Biology 3/e (© Garland Science 2010)

Figure credit:
http://u18439936.onlinehome-server.com/craig.milgrim/Bio230/Outline/ECBFigures_Tables/Chapter_7/FigureJPGs/figure_07_09.jpg

DNA -> Pre-mRNA



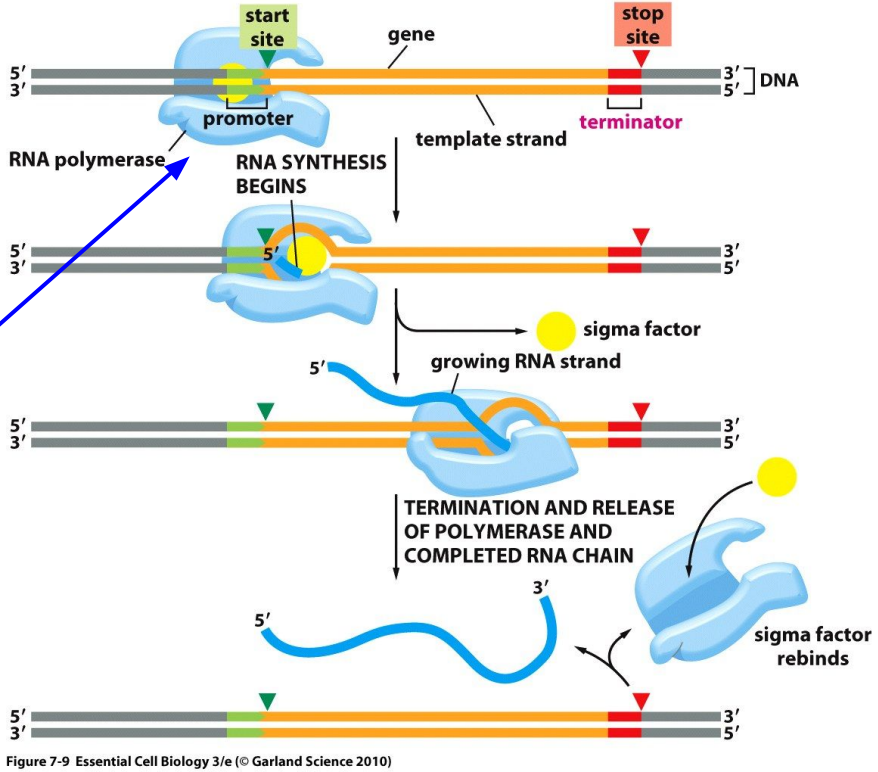
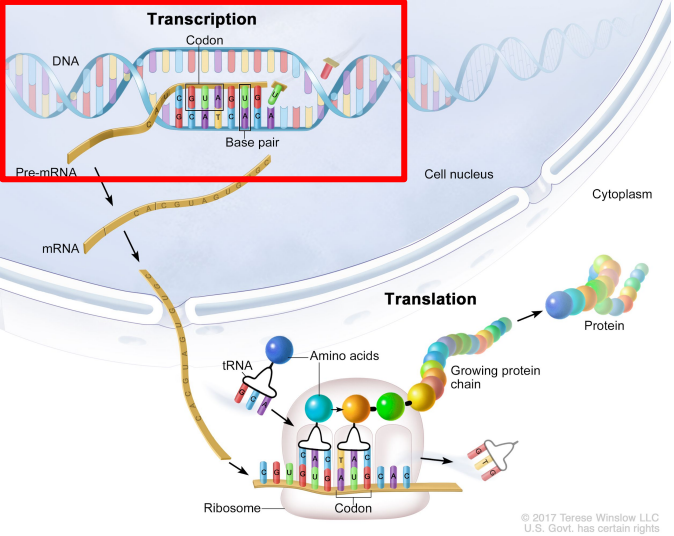


Figure 7-9 Essential Cell Biology 3/e (© Garland Science 2010)

RNA polymerase: enyzme that binds to promoter region and uses DNA template to synthesize complementary RNA

Figure credit:
http://u18439936.onlinehome-server.com/craig.milgrim/Bio230/Outline/ECBFigures_Tables/Chapter_7/FigureJPGs/figure_07_09.jpg

DNA -> Pre-mRNA



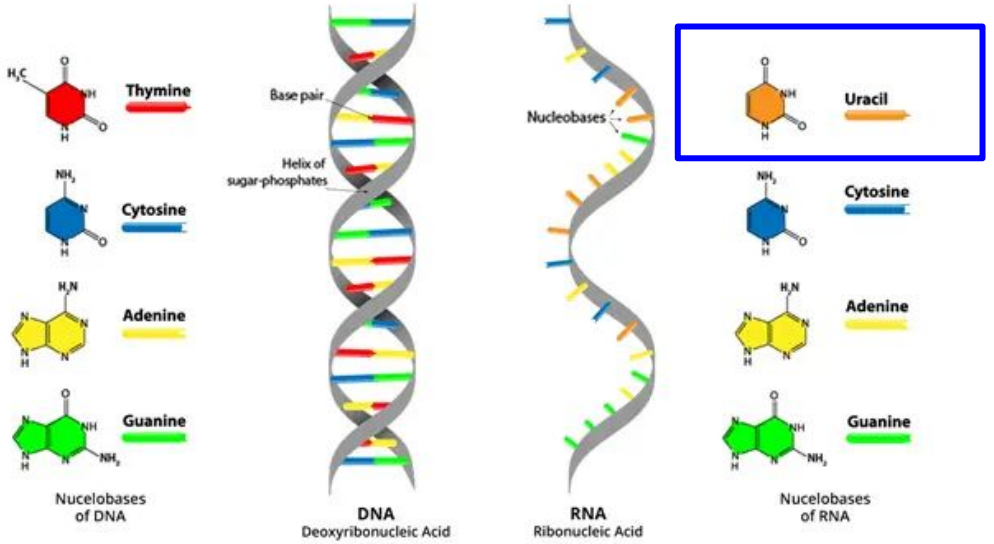
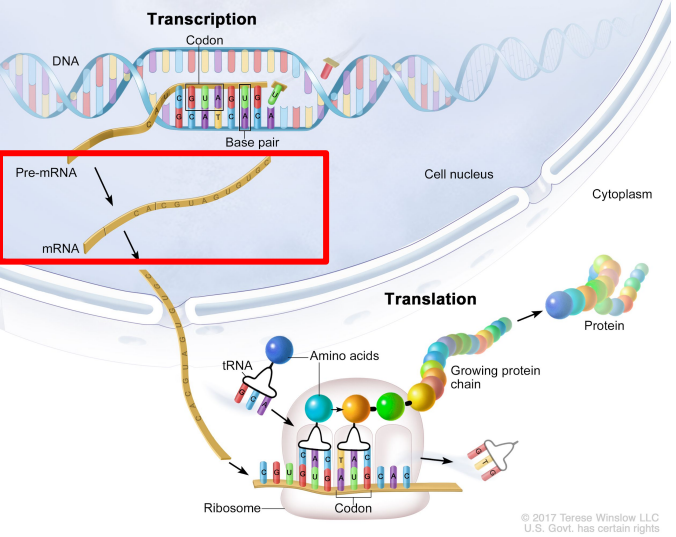


Figure credit:
https://cdn.technologynetworks.com/tn/images/thumbs/webp/640_360/what-are-the-key-differences-between-dna-and-rna-296719.webp?v=9503516

Pre-mRNA -> mRNA



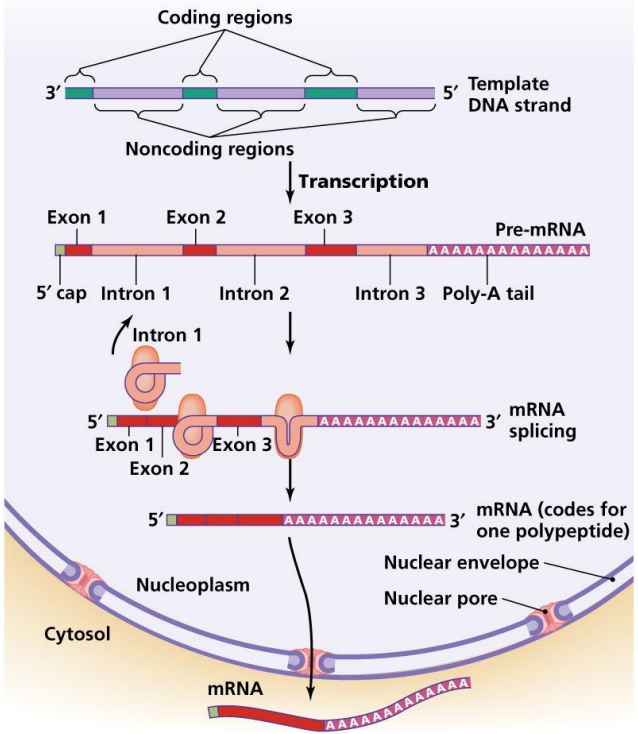
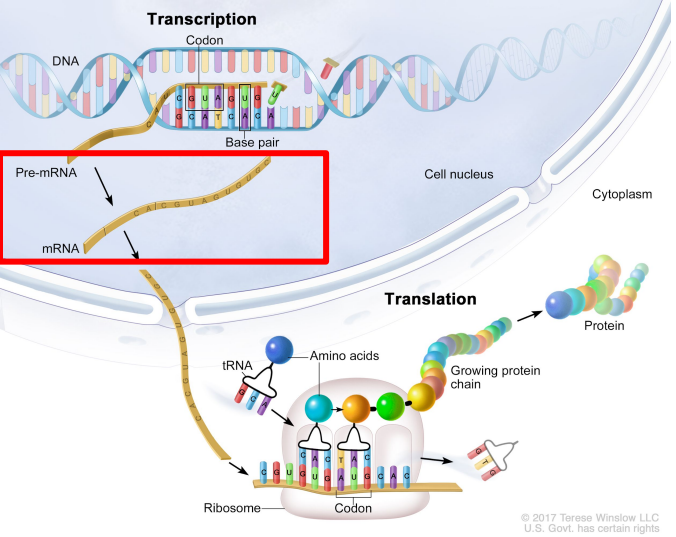


Figure credit: http://academic.pgcc.edu/~kroberts/Lecture/Chapter%207/transcription.html

Pre-mRNA -> mRNA



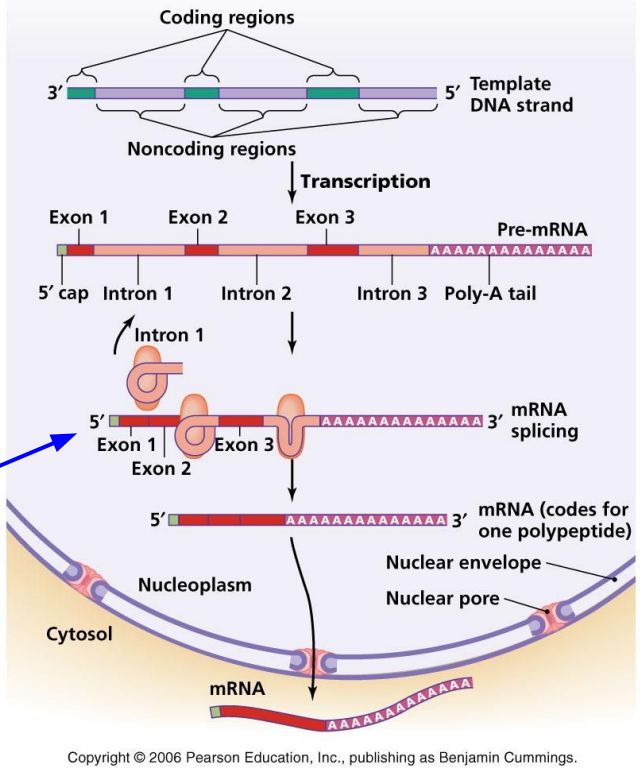
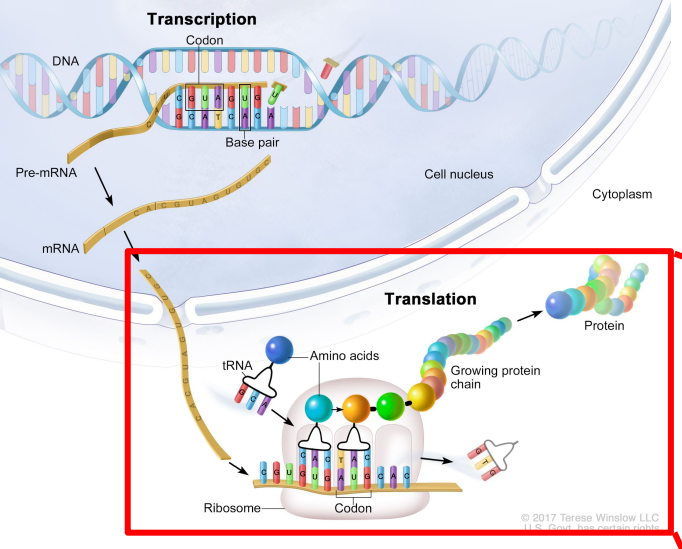


Figure credit: http://academic.pgcc.edu/~kroberts/Lecture/Chapter%207/transcription.html

mRNA splicing: remove introns (non-coding regions), splice together exons (coding regions)

mRNA -> Proteins



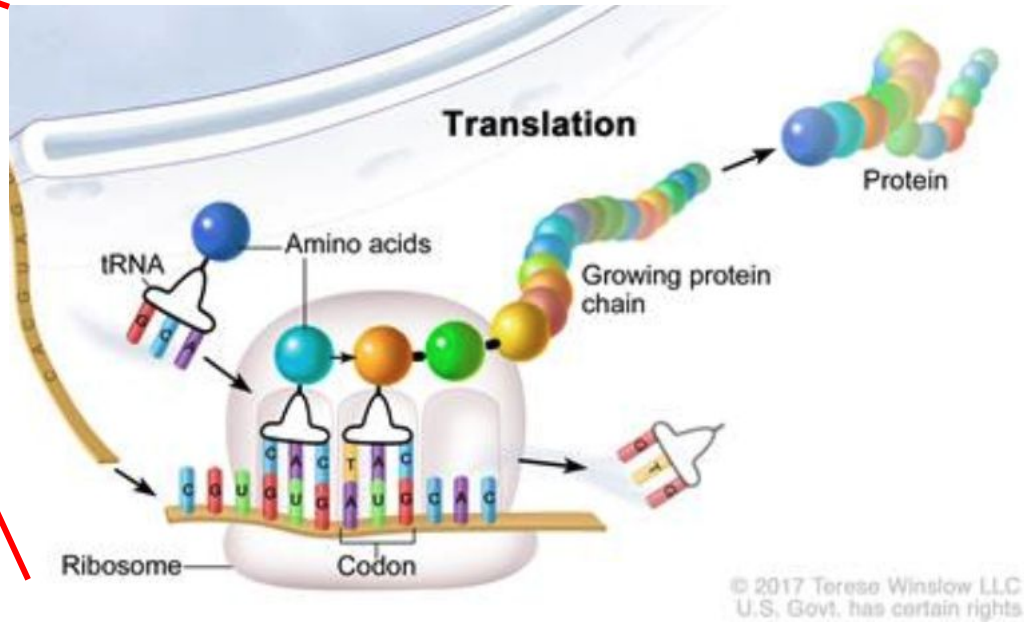
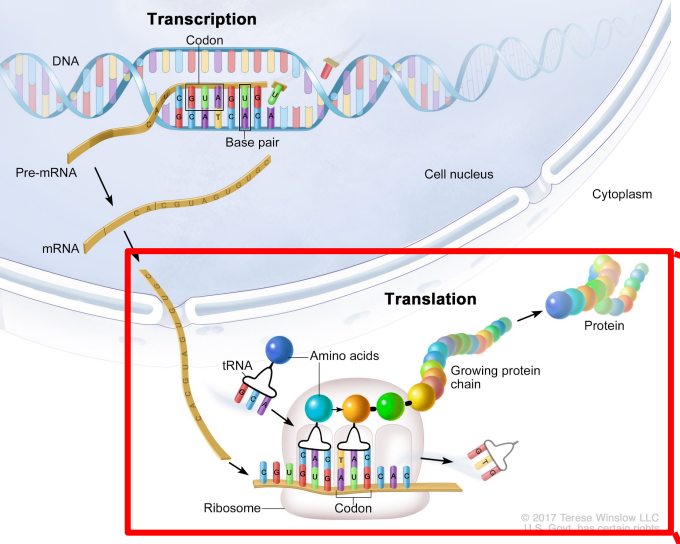


Figure credit: https://www.cancer.gov/images/cdr/live/CDR761782-571.jpg

mRNA -> Proteins



Ribosome: cell organelle that synthesizes proteins

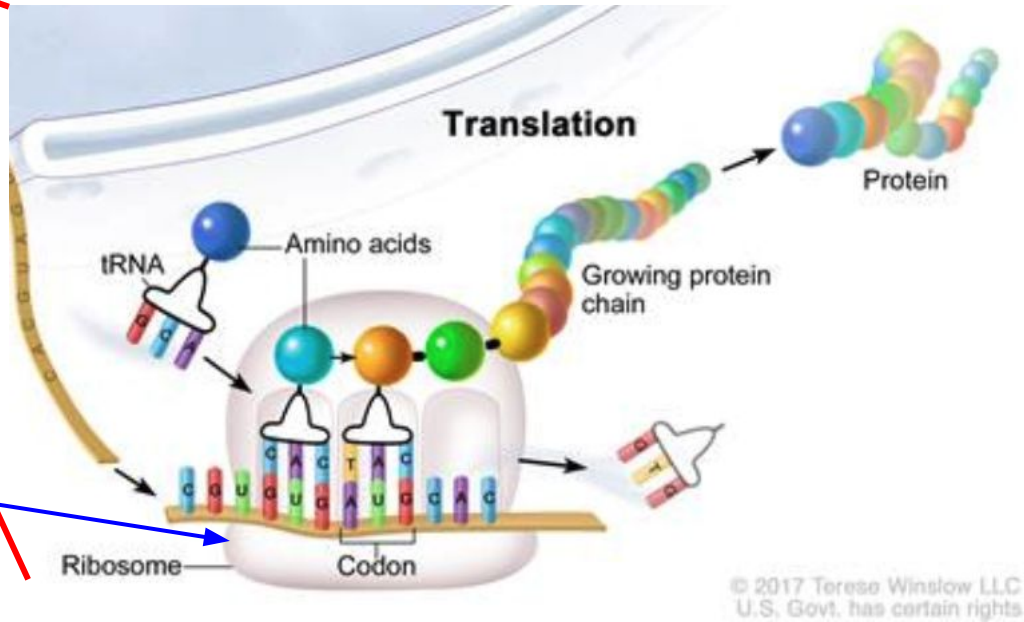
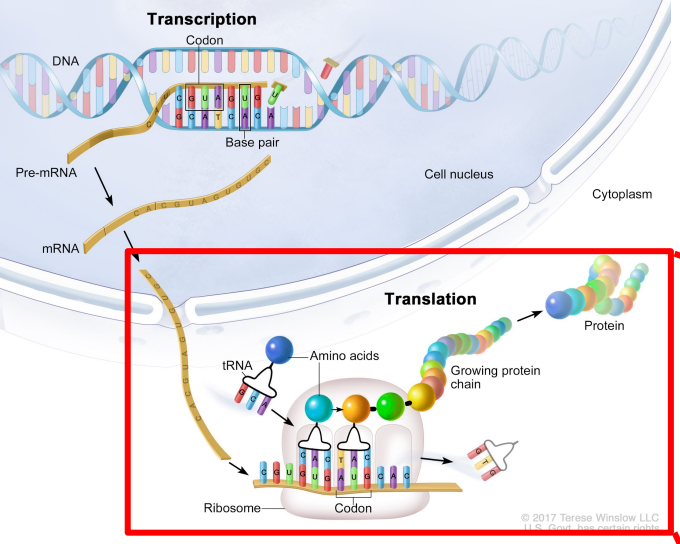


Figure credit: https://www.cancer.gov/images/cdr/live/CDR761782-571.jpg

mRNA -> Proteins



tRNA: molecule carrying amino acids corresponding to each 3-nucleotide codon

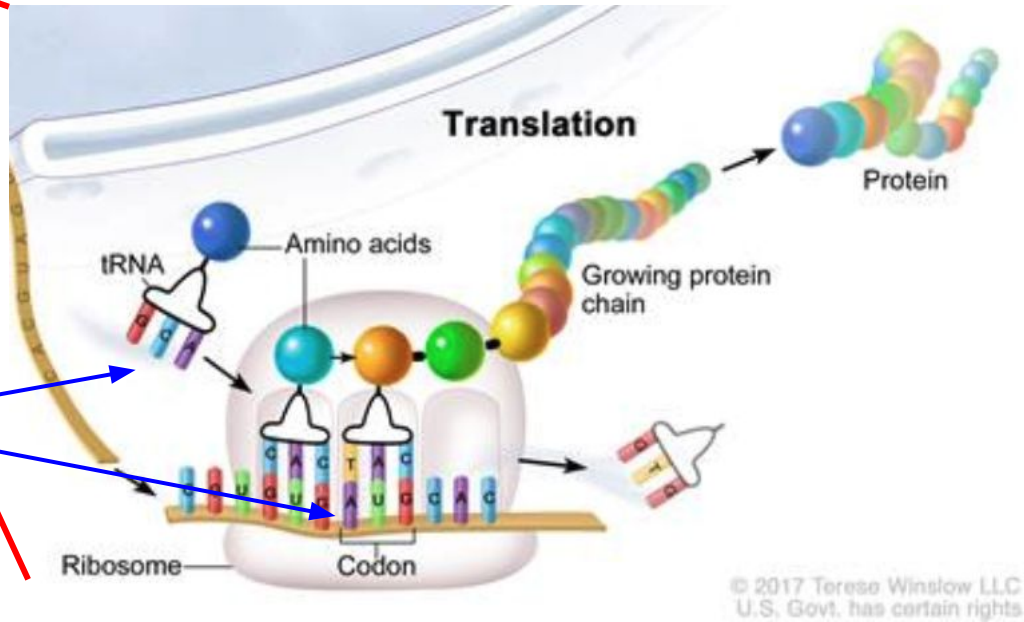
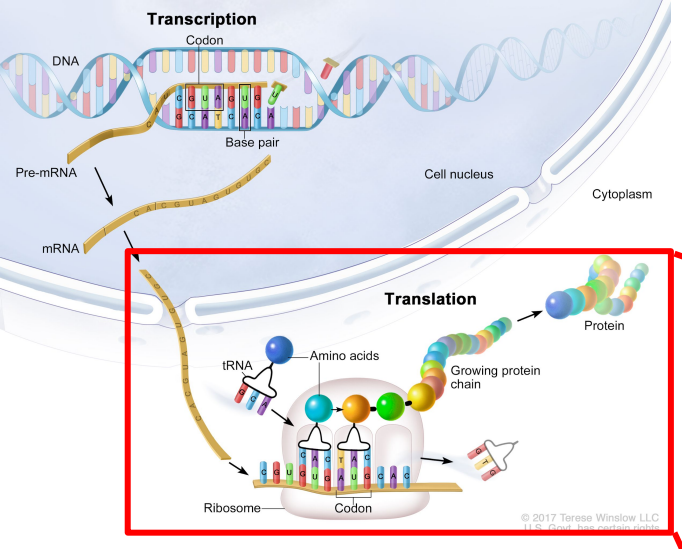


Figure credit: https://www.cancer.gov/images/cdr/live/CDR761782-571.jpg

mRNA -> Proteins



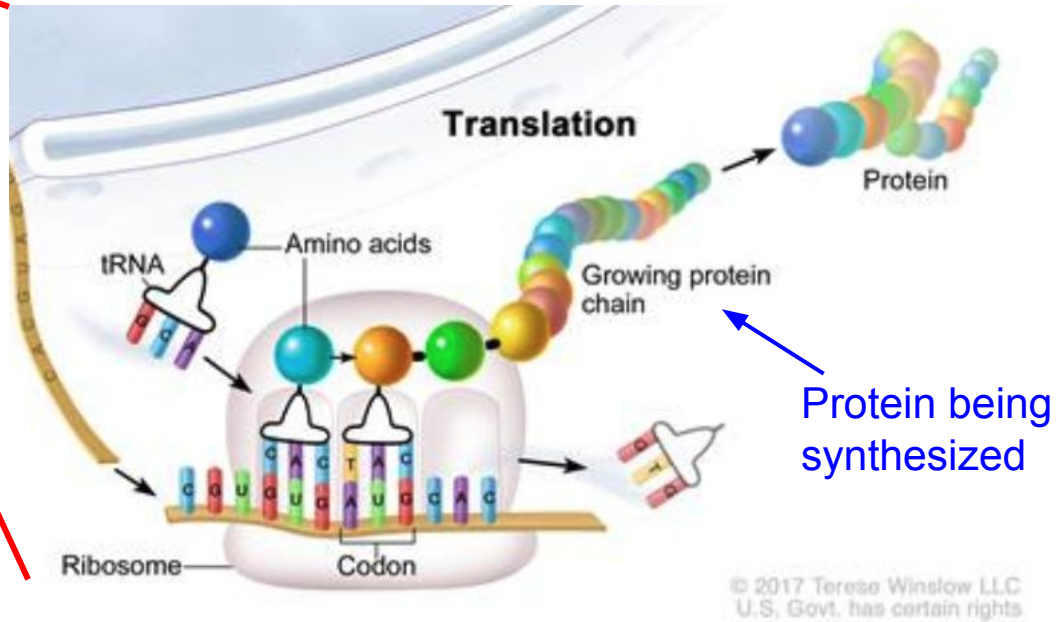
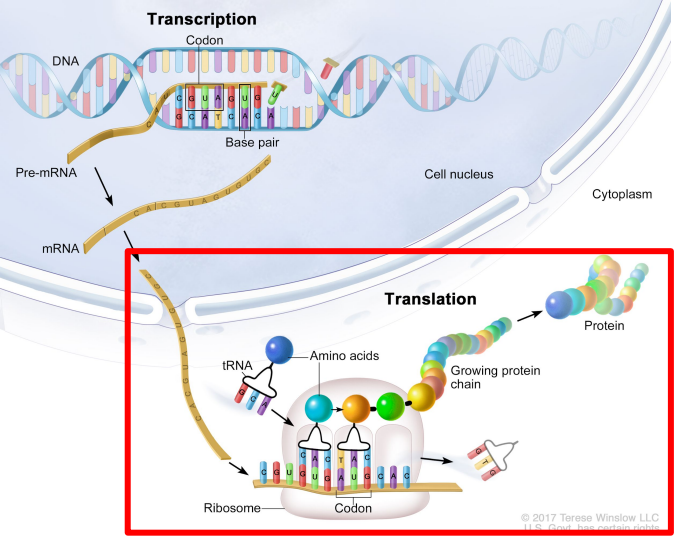


Figure credit: https://www.cancer.gov/images/cdr/live/CDR761782-571.jpg

mRNA -> Proteins



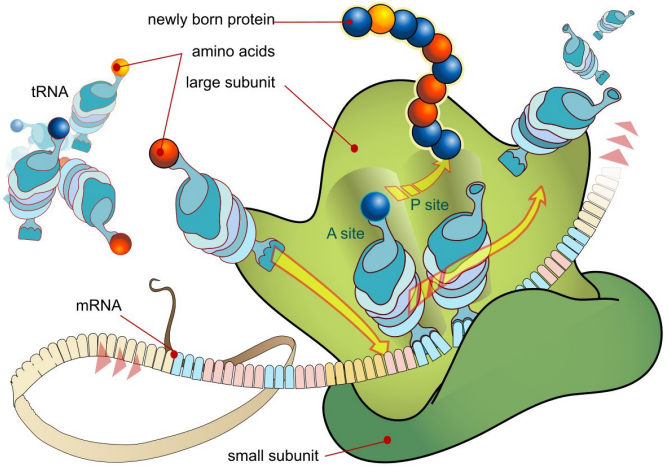
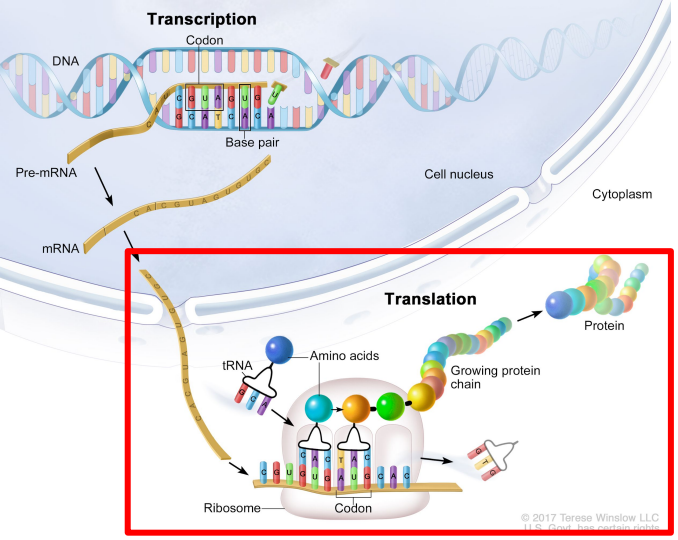
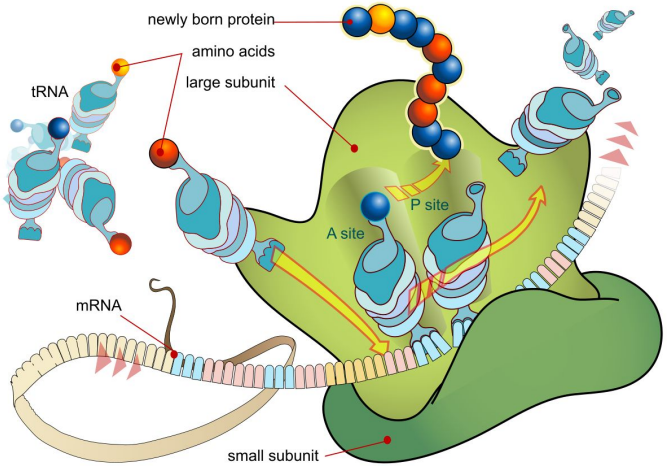


Figure credit:
https://en.wikipedia.org/wiki/Translation_(biology)#/media/File:Ribosome_mRNA_translation_en.svg
https://philschatz.com/biology-concepts-book/resources/Figure_09_04_02.jpg

mRNA -> Proteins





Codon -> amino acid mapping →

Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
A	AUU Ile, AUC Ile, AUA Ile, AUG Met	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn, AAA Lys, AAG Lys	AGU Ser, AGC Ser, AGA Arg, AGG Arg	U C A G
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp, GAA Glu, GAG Glu	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G

Figure credit:
https://en.wikipedia.org/wiki/Translation_(biology)#/media/File:Ribosome_mRNA_translation_en.svg
https://philschatz.com/biology-concepts-book/resources/Figure_09_04_02.jpg

Epigenomics

Study of processes that regulate how and when genes are turned on and off (“gene expression”)

Epigenomics

Study of processes that regulate how and when genes are turned on and off (“gene expression”)

- E.g. **transcription factors**: proteins that bind to the promoter and other noncoding regions, can enhance or repress transcription

Epigenomics

Study of processes that regulate how and when genes are turned on and off (“gene expression”)

- E.g. **transcription factors**: proteins that bind to the promoter and other noncoding regions, can enhance or repress transcription

- E.g. **DNA methylation**: addition of large methyl group to promoter region makes it difficult for proteins to bind -> represses transcription

Epigenomics

Study of processes that regulate how and when genes are turned on and off (“gene expression”)

- E.g. **transcription factors**: proteins that bind to the promoter and other noncoding regions, can enhance or repress transcription
- E.g. **DNA methylation**: addition of large methyl group to promoter region makes it difficult for proteins to bind -> represses transcription
- E.g. **Histone modification**: addition or removal of acetyl groups affects charge interaction to relax or tighten chromatin structure (easier for proteins to bind)

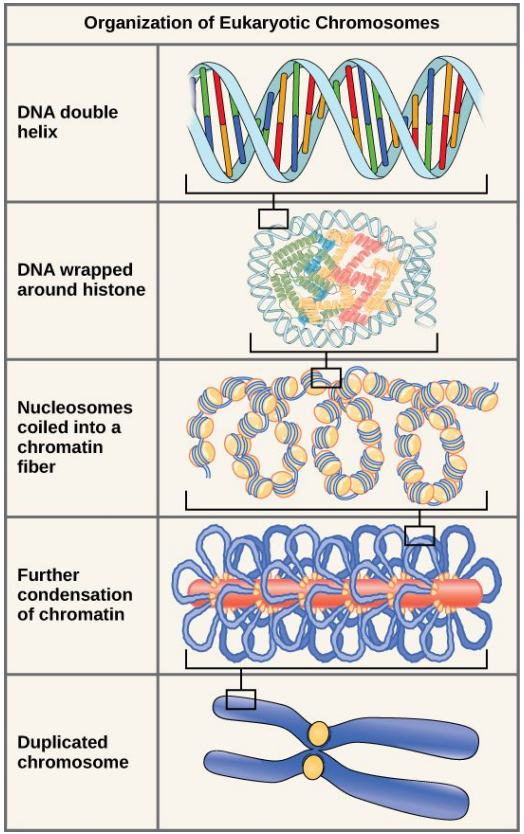


Figure credit:
https://philschatz.com/biology-concepts-book/resources/Figure_09_01_06.jpg

Transcriptomics

- Study of the transcriptome (the RNA of a cell)
- One reason of interest: Harder to measure proteins (the functional molecules!), but we can sequence RNA as a (highly imperfect) proxy for proteins to quantify cell state

Transcriptomics

- Study of the transcriptome (the RNA of a cell)
- One reason of interest: Harder to measure proteins (the functional molecules!), but we can sequence RNA as a (highly imperfect) proxy for proteins to quantify cell state

Proteomics

- Study of the proteins in a cell

Data: genomic sequencing

Produces readout of DNA template strands

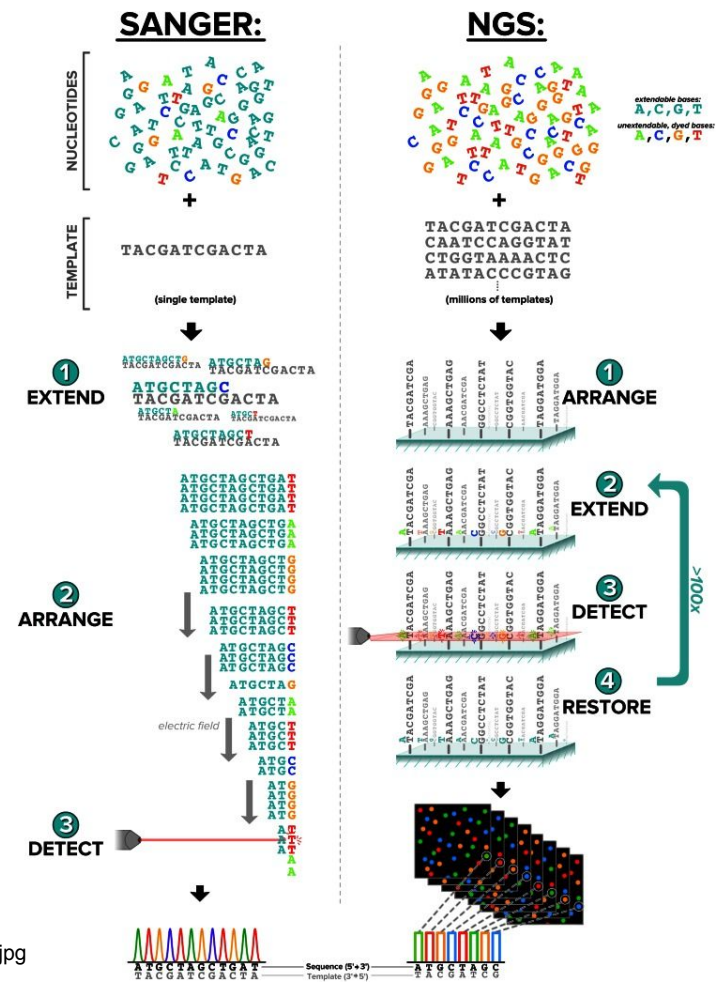


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

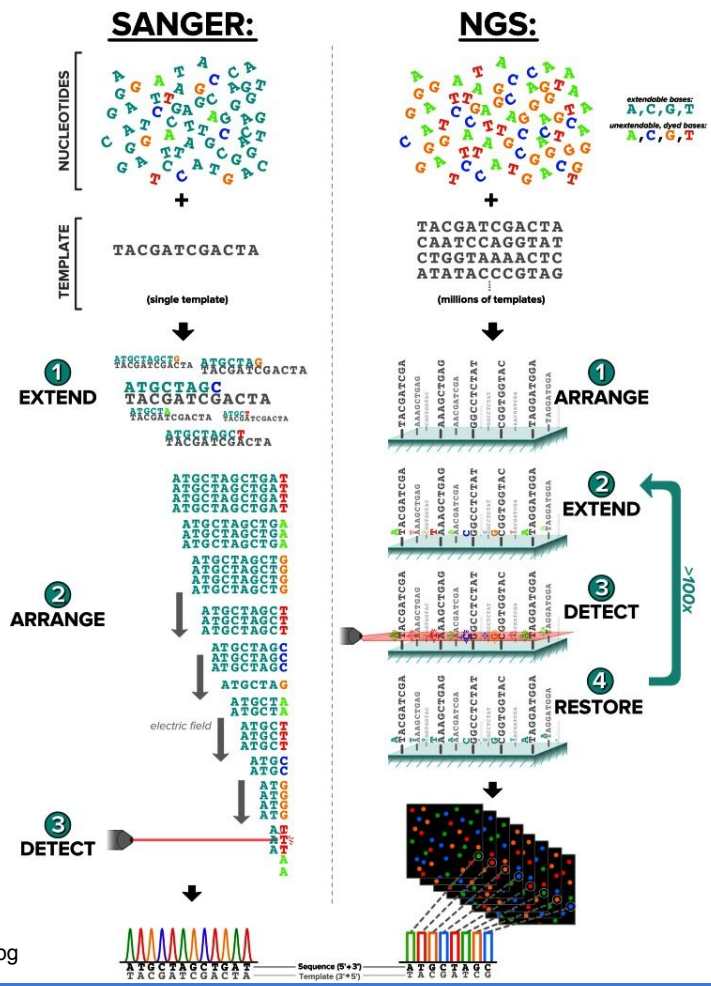


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Add some special (and fluorescently labeled) nucleotides that cause a chain being synthesized to terminate

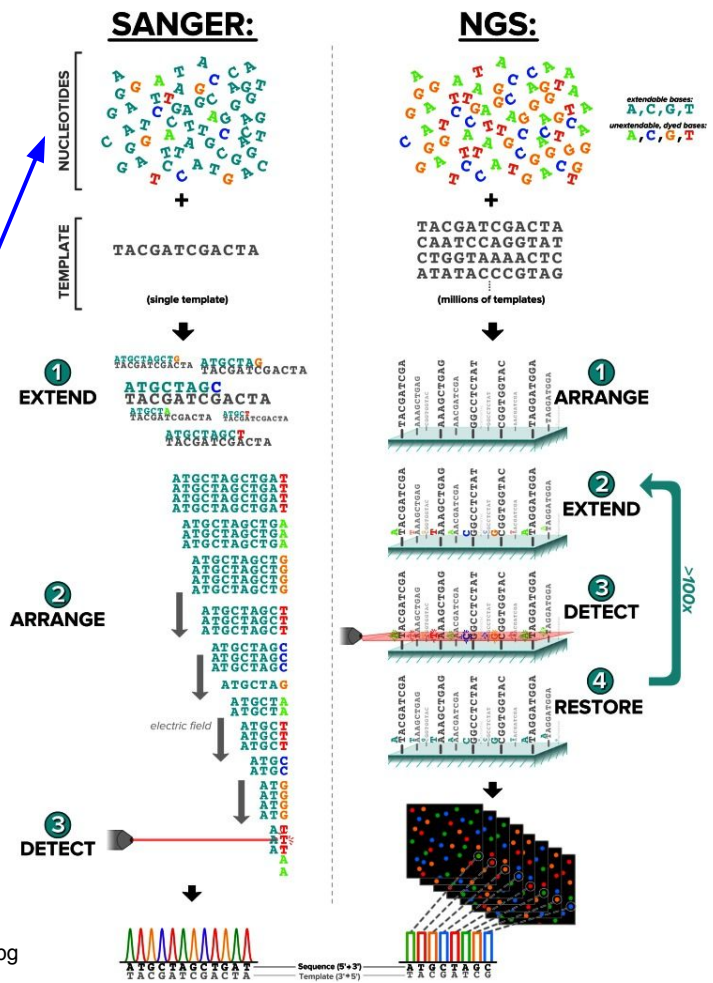


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Random interaction of nucleotides with template strand lead to chains of different early-terminated lengths

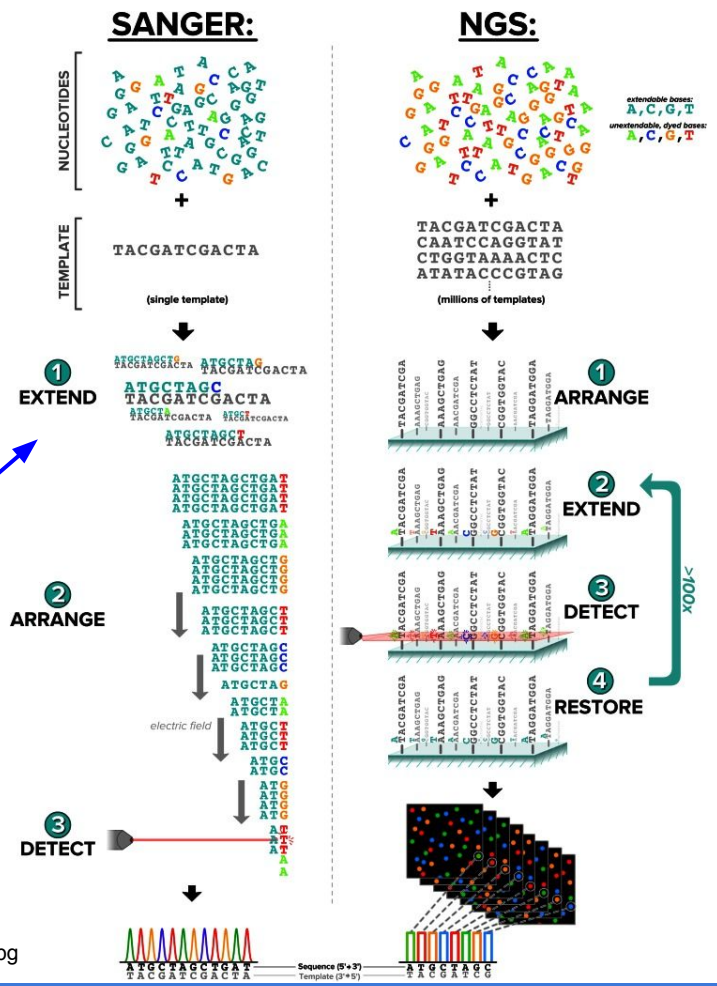


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Sorting by length (e.g. electrophoresis) gives sequence readout

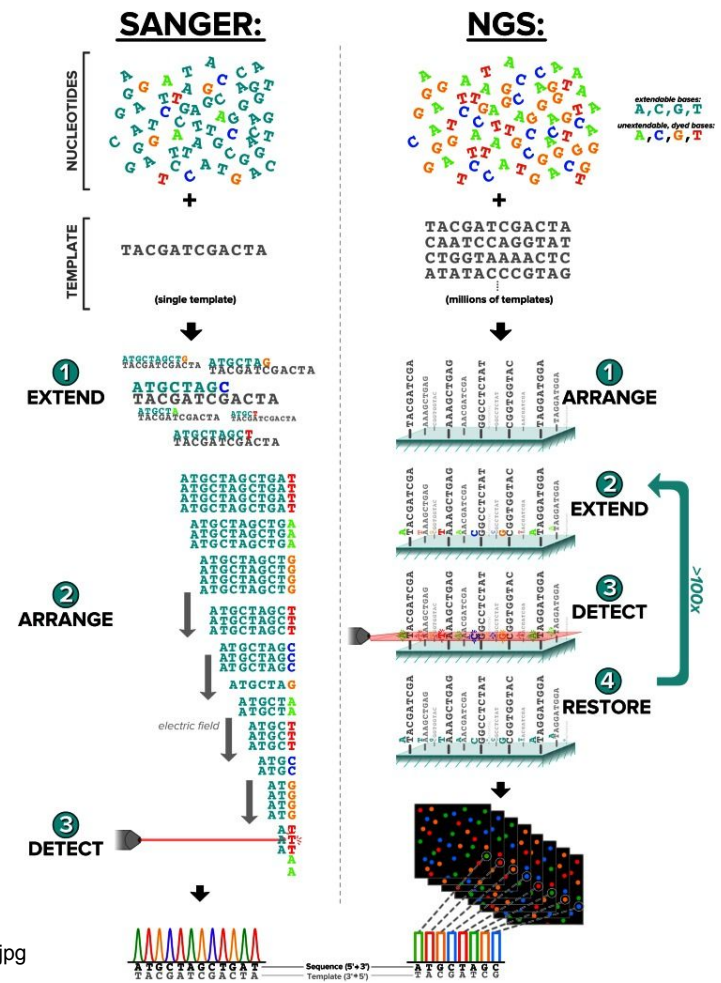


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Next-generation sequencing (NGS): Used since 2000s, based on massively parallelized sequencing of short sequences

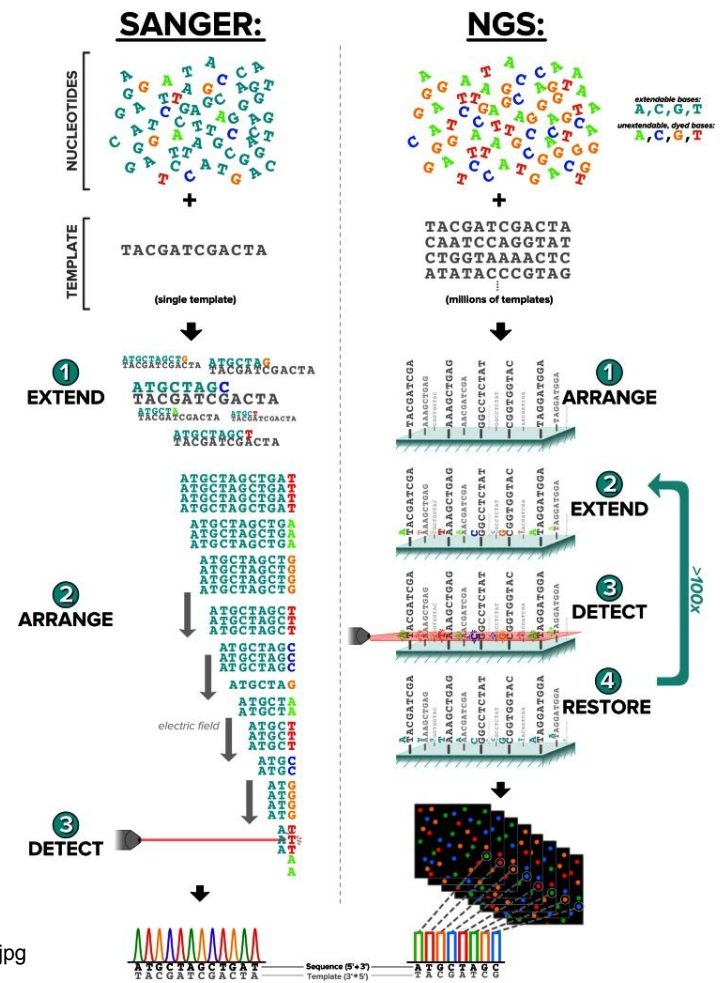


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Next-generation sequencing (NGS): Used since 2000s, based on massively parallelized sequencing of short sequences

Arrange many short templates on an array

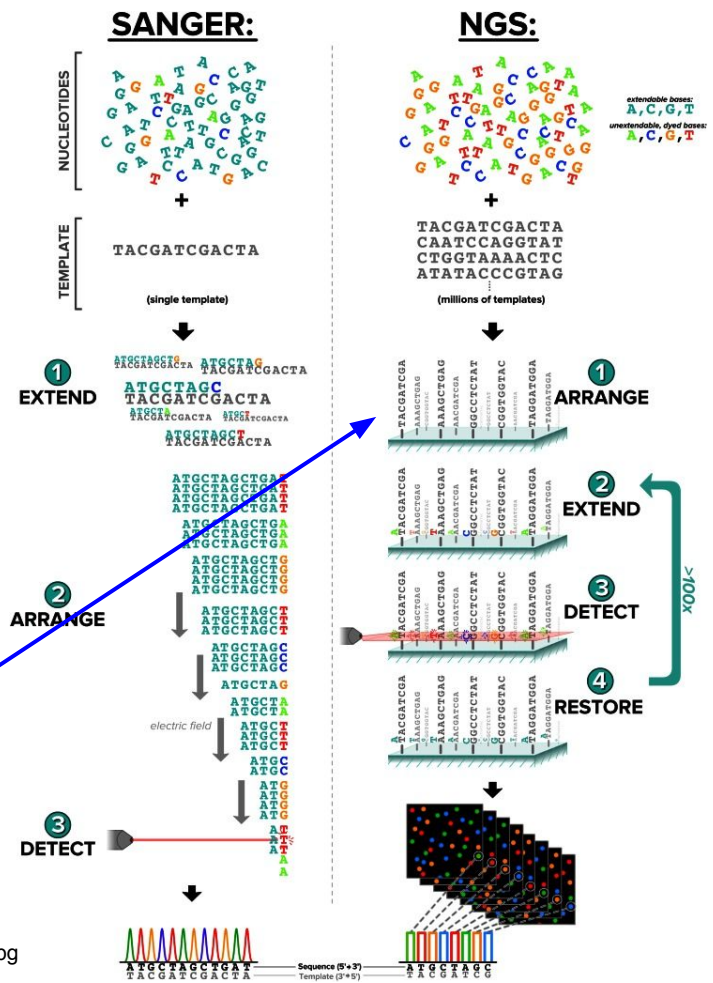


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Next-generation sequencing (NGS): Used since 2000s, based on massively parallelized sequencing of short sequences

Now all added nucleotides are chain-terminating

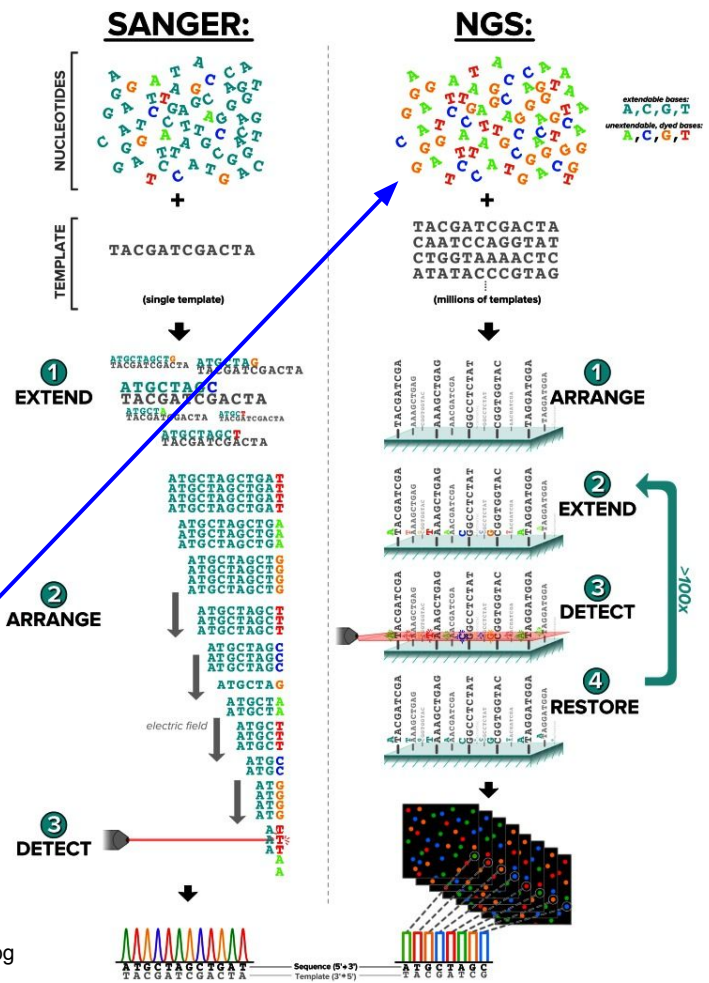


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Next-generation sequencing (NGS): Used since 2000s, based on massively parallelized sequencing of short sequences

All templates get next sequence element attached (and terminated), then read

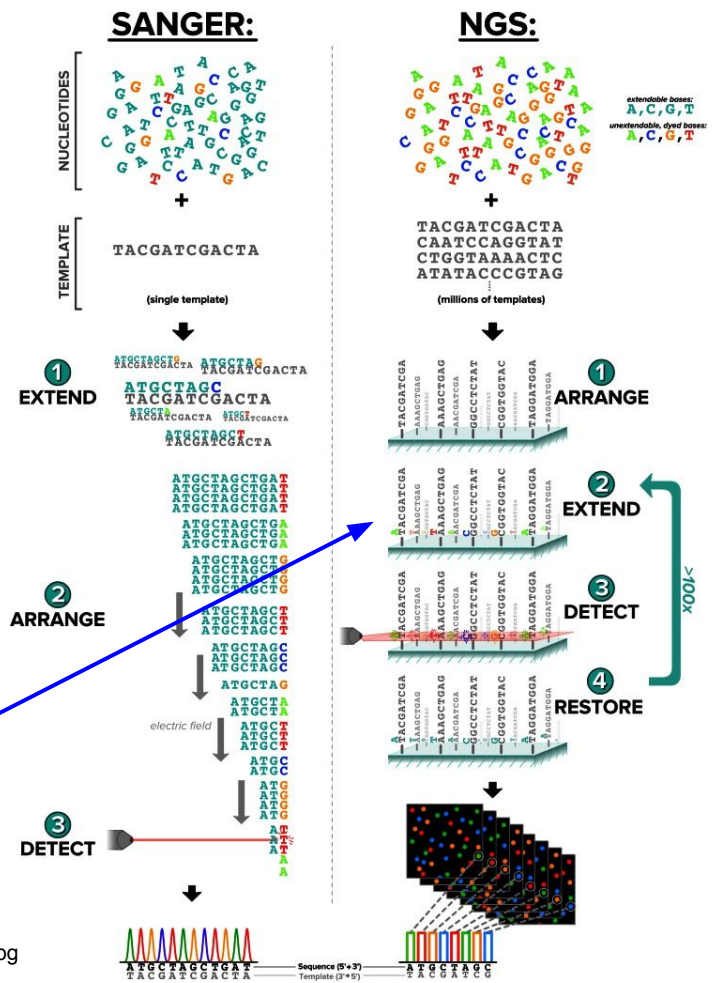


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Next-generation sequencing (NGS): Used since 2000s, based on massively parallelized sequencing of short sequences

Apply process to “restore” the chain-terminating nucleotides to be normal, then repeat to extend synthesizing sequence by one more nucleotide

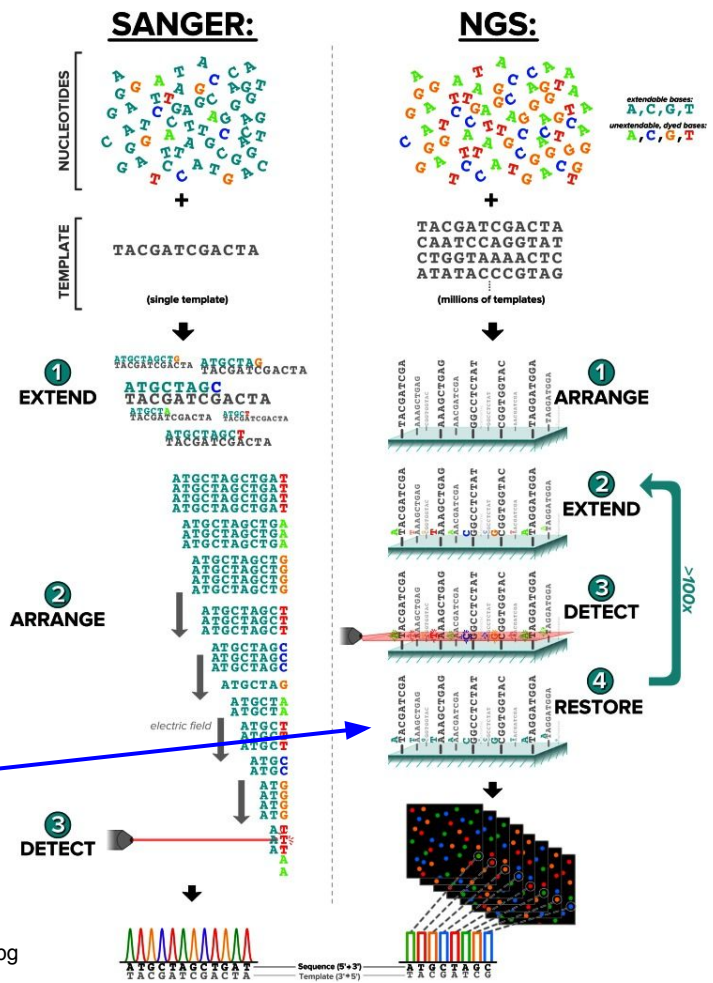


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: genomic sequencing

Produces readout of DNA template strands

Sanger sequencing: Invented in 1977, based on “chain termination”

Next-generation sequencing (NGS): Used since 2000s, based on massively parallelized sequencing of short sequences

Set of read-out images at every step gives sequences of all template strands. Then analyze data to reconstruct longer sequences.

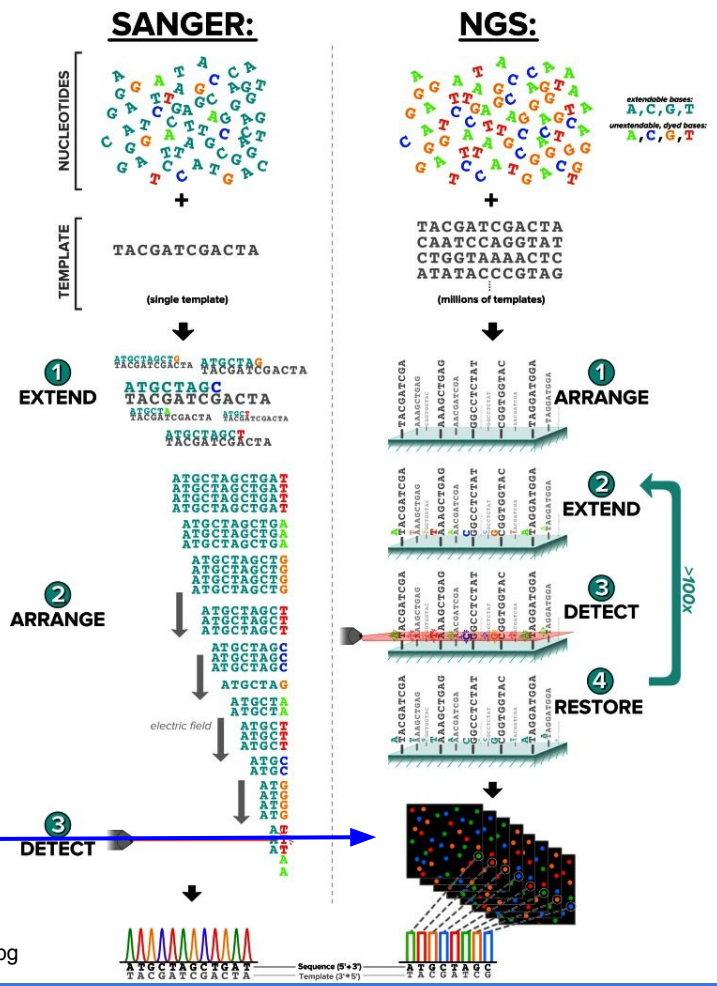


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig1_HTML.jpg

Data: DNA microarray

Produces relative expression of genes in normal vs disease tissue samples

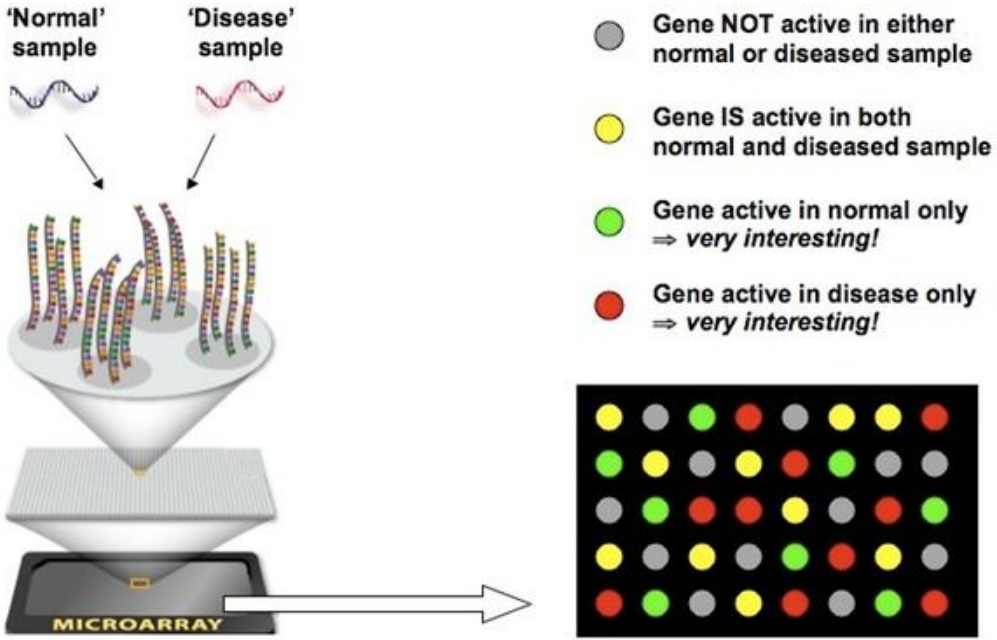


Figure credit: http://www.vce.bioninja.com.au/_Media/microarray_med.jpeg

Data: DNA microarray

Produces relative expression of genes in normal vs disease tissue samples

Isolate mRNA ("expressed genes") from tissue samples and synthesize complementary DNA (cDNA).

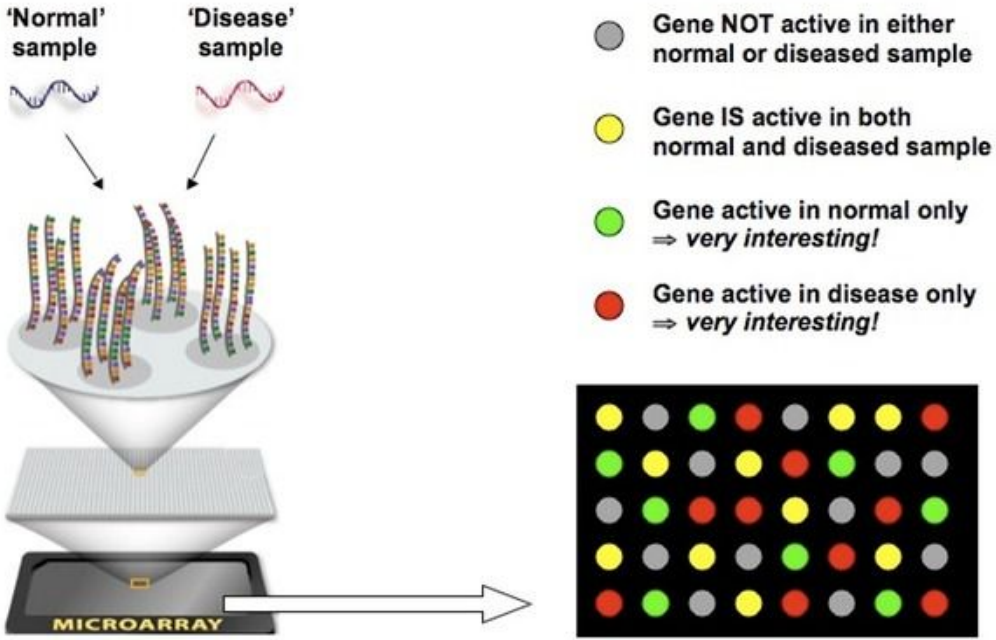


Figure credit: http://www.vce.bioninja.com.au/_Media/microarray_med.jpeg

Data: DNA microarray

Produces relative expression of genes in normal vs disease tissue samples

Isolate mRNA ("expressed genes") from tissue samples and synthesize complementary DNA (cDNA).

Use fluorescent tags to label cDNA from normal tissue green, and from disease tissue red

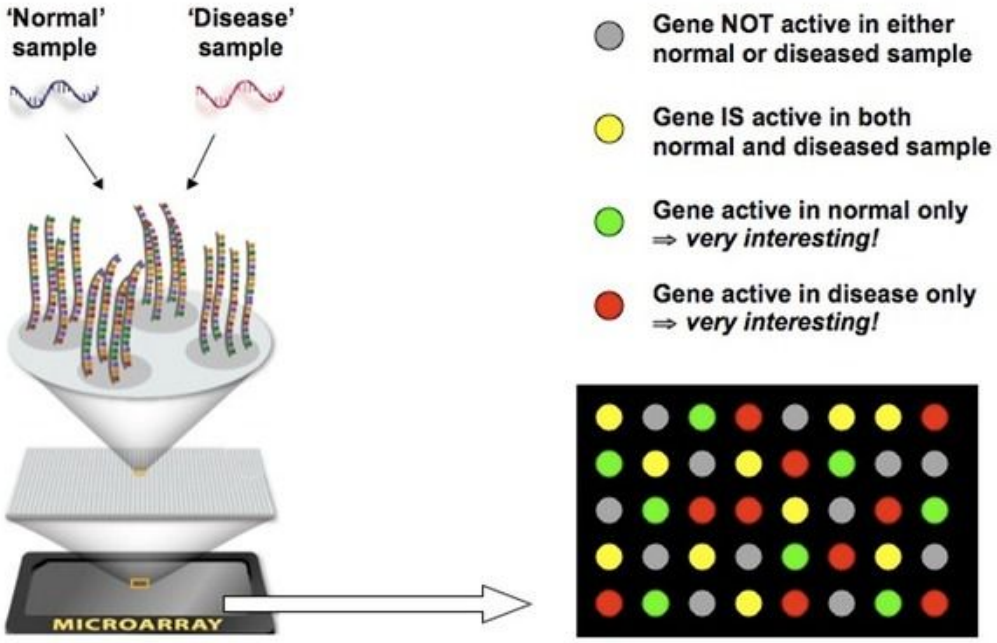
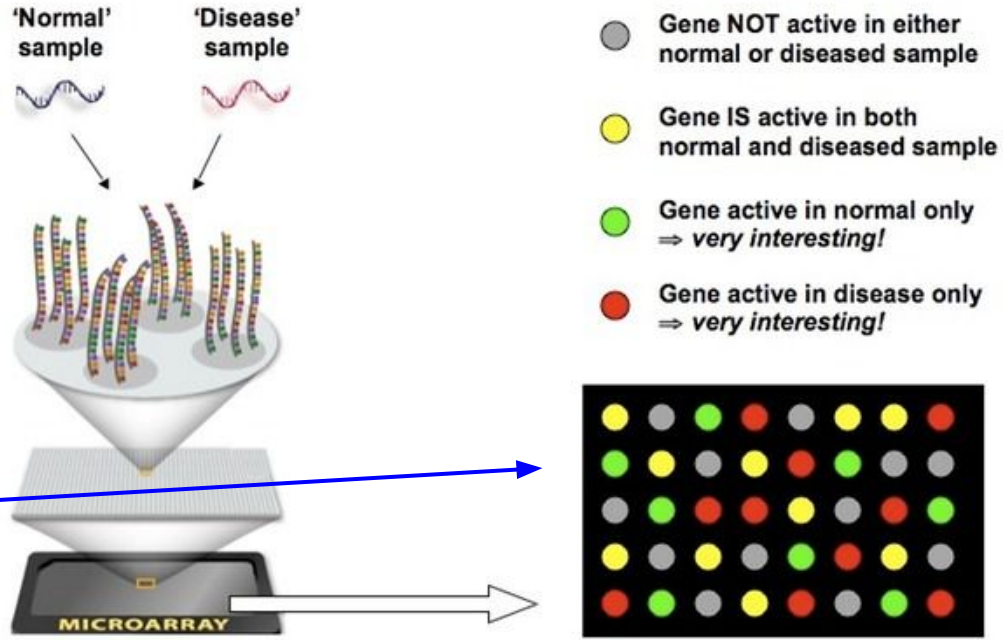


Figure credit: http://www.vce.bioninja.com.au/_Media/microarray_med.jpeg

Data: DNA microarray

Produces relative expression of genes in normal vs disease tissue samples



Each spot of DNA microarray contains single-stranded DNA corresponding to a gene

Figure credit: http://www.vce.bioninja.com.au/_Media/microarray_med.jpeg

Data: DNA microarray

Produces relative expression of genes in normal vs disease tissue samples

cDNA will bind to the corresponding DNA strands on microarray. Color indicates ratio of cDNA (relative gene expression) in the normal vs disease tissue

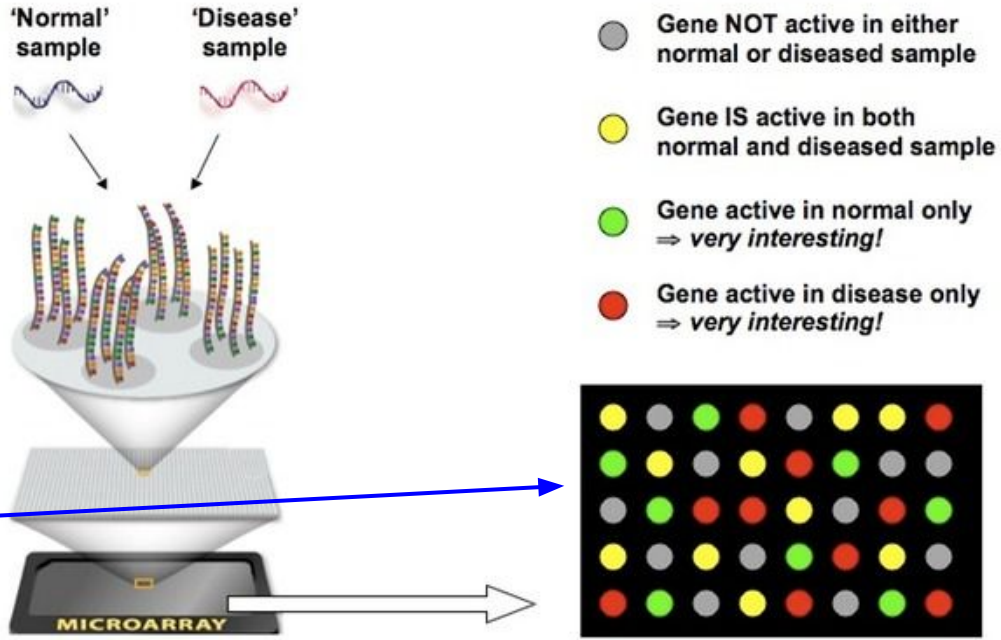


Figure credit: http://www.vce.bioninja.com.au/_Media/microarray_med.jpeg

Data: RNA-seq

Produces readout of mRNA content in a tissue sample

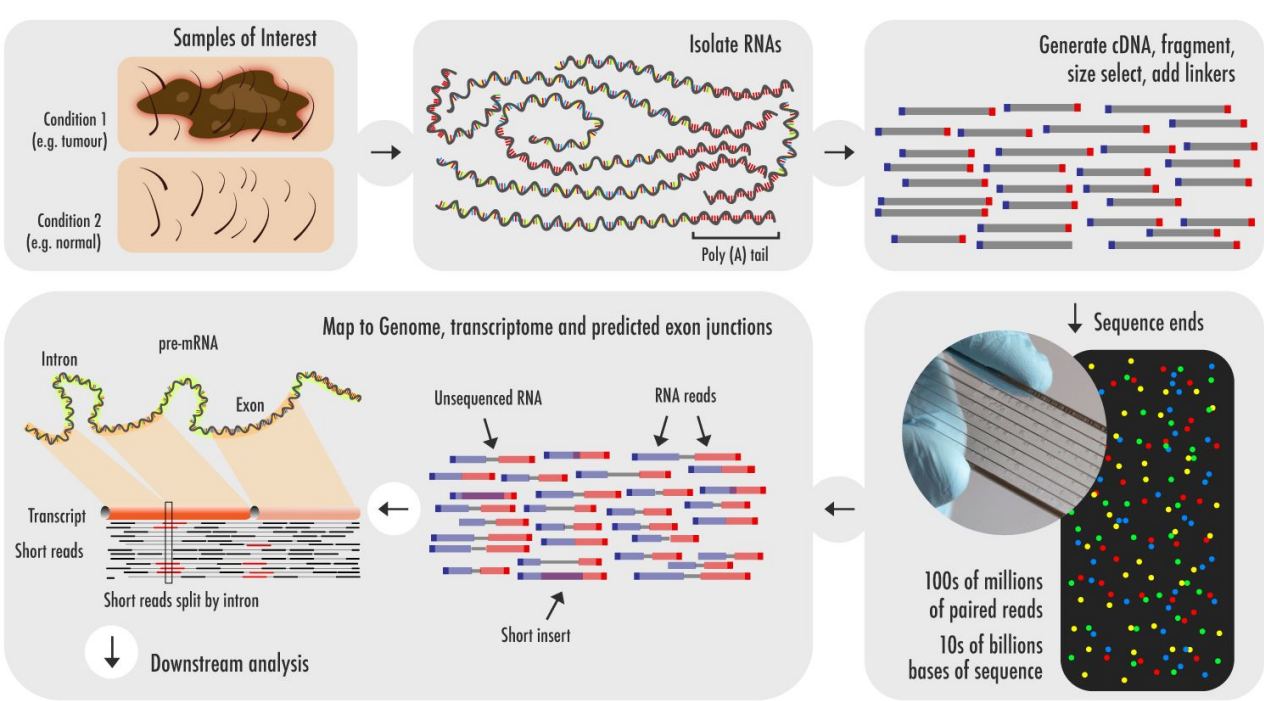


Figure credit: https://cdn.technologynetworks.com/tn/images/body/dnasequencinga1529596208892.png

Data: RNA-seq

Produces readout of mRNA content in a tissue sample

Isolate RNA and generate cDNA

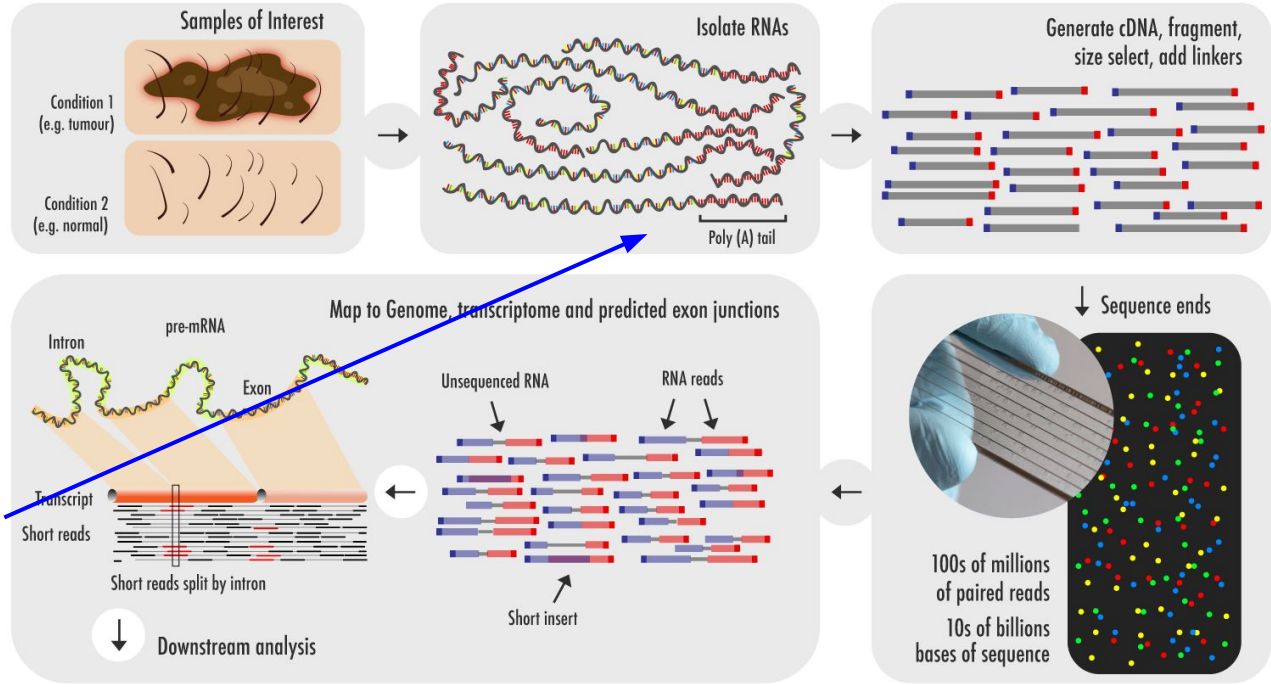
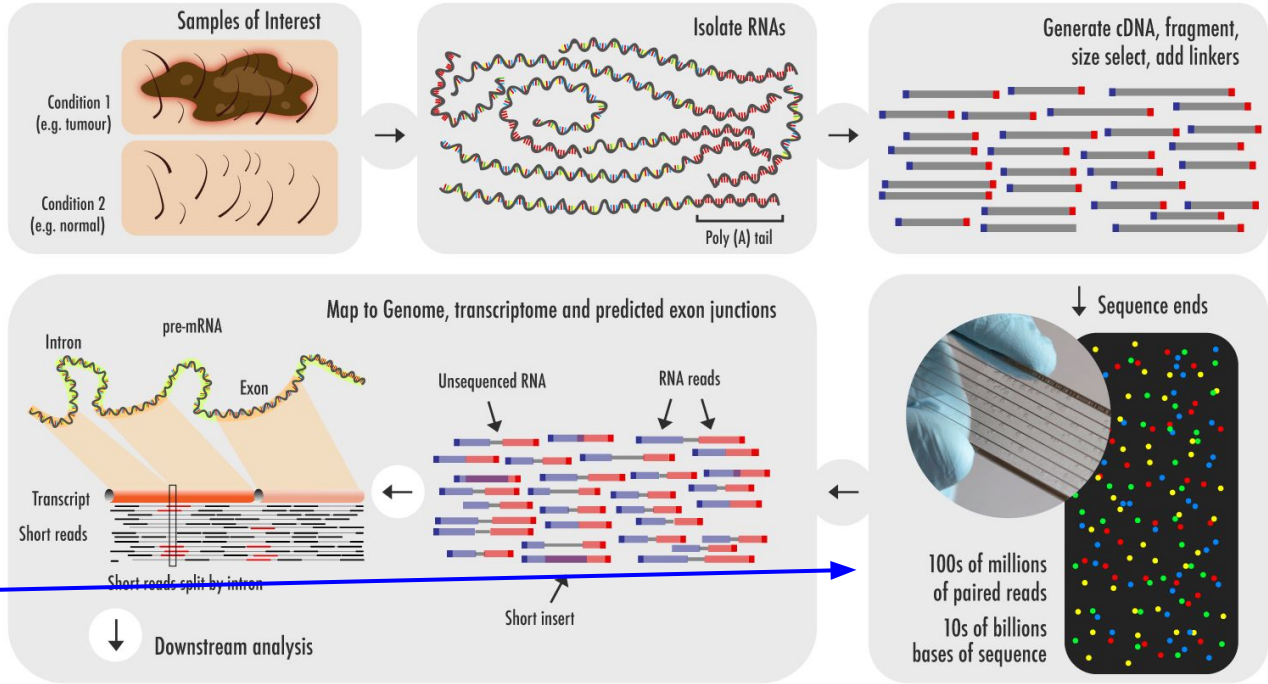


Figure credit: https://cdn.technologynetworks.com/tn/images/body/dnasequencinga1529596208892.png

Data: RNA-seq

Produces readout of mRNA content in a tissue sample



Use NGS to sequence cDNA

Figure credit: https://cdn.technologynetworks.com/tn/images/body/dnasequencinga1529596208892.png

Data: RNA-seq

Produces readout of mRNA content in a tissue sample

Map back to reference genome for analysis

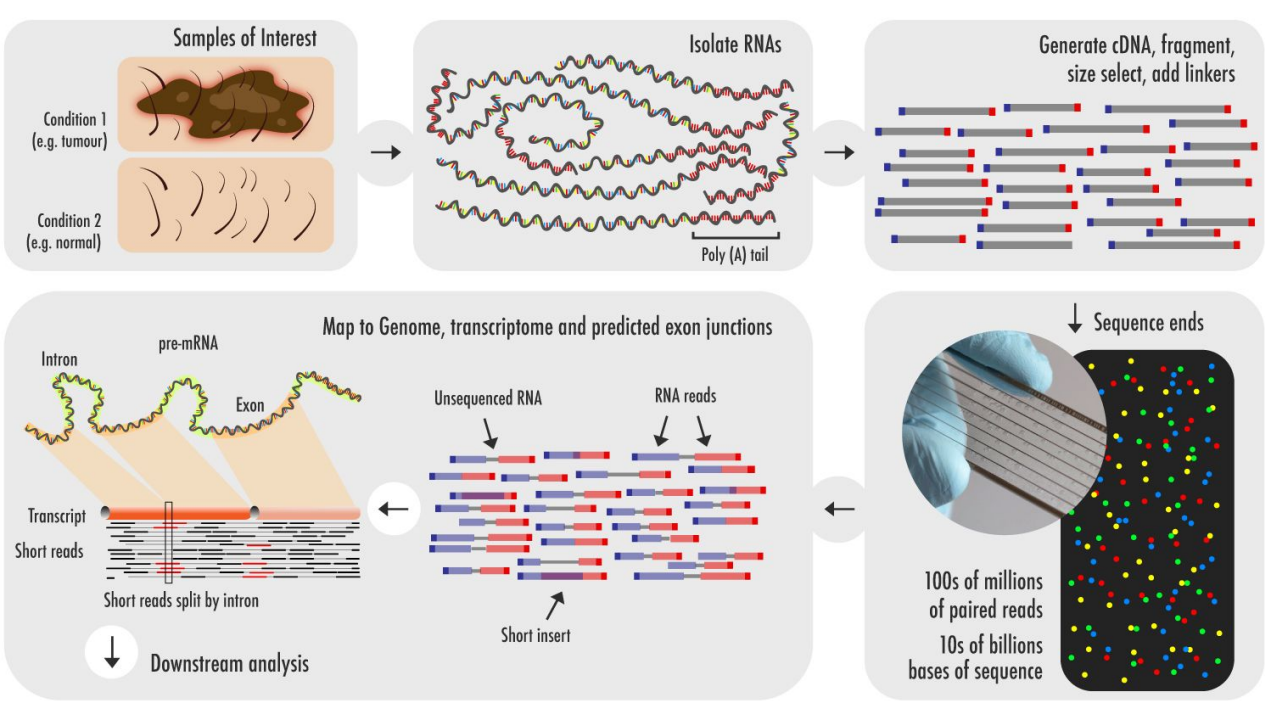
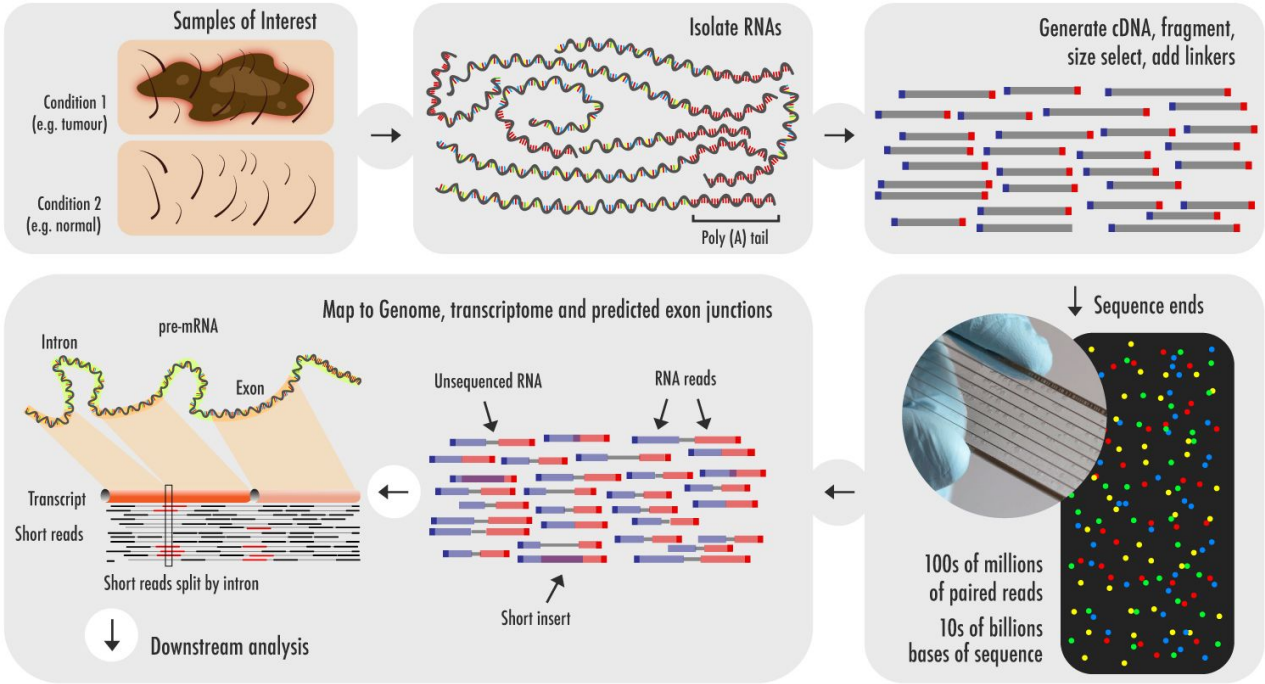


Figure credit: https://cdn.technologynetworks.com/tn/images/body/dnasequencinga1529596208892.png

Data: RNA-seq

Produces readout of mRNA content in a tissue sample



Map back to reference genome for analysis

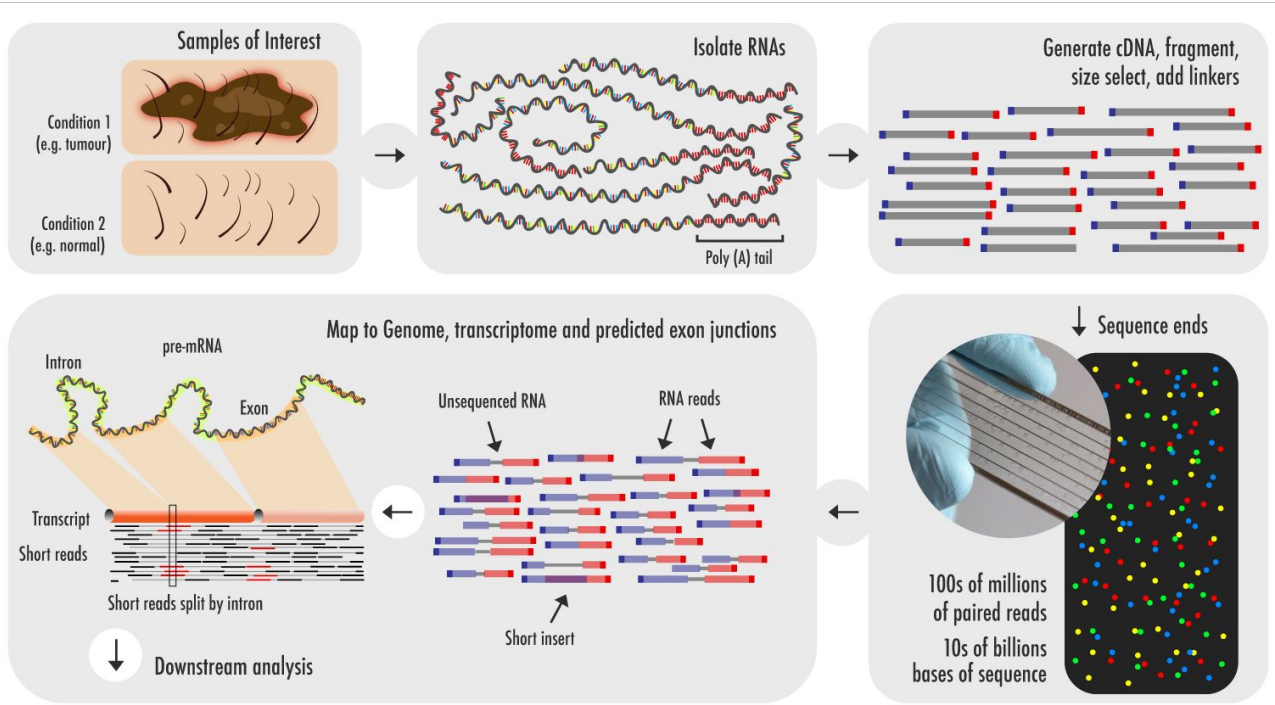
Now standard approach for transcriptomics study

Figure credit: https://cdn.technologynetworks.com/tn/images/body/dnasequencinga1529596208892.png

Data: RNA-seq

More recently in 2010s, single-cell RNA-seq!

Produces readout of mRNA content in a tissue sample



Map back to reference genome for analysis

Now standard approach for transcriptomics study

Figure credit: https://cdn.technologynetworks.com/tn/images/body/dnasequencinga1529596208892.png

Data: ChIP-seq

Produces reads of DNA sequences where a protein binds

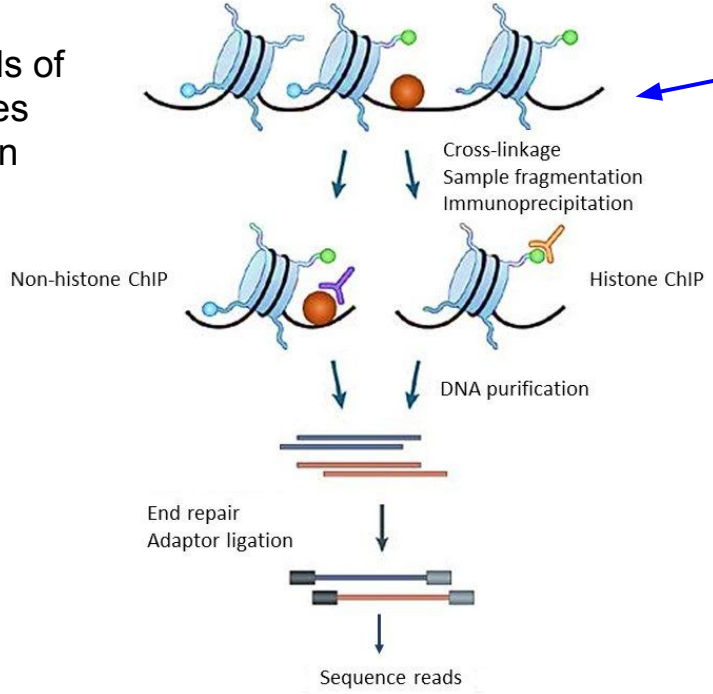


Figure credit: https://www.france-genomique.org/wp-content/uploads/2019/08/CHIP-selon-Park-1-e1566900408602.jpg

Use formaledehyde treatment to cross-link (fix) proteins to their bound DNA

Data: ChIP-seq

Produces reads of DNA sequences where a protein binds

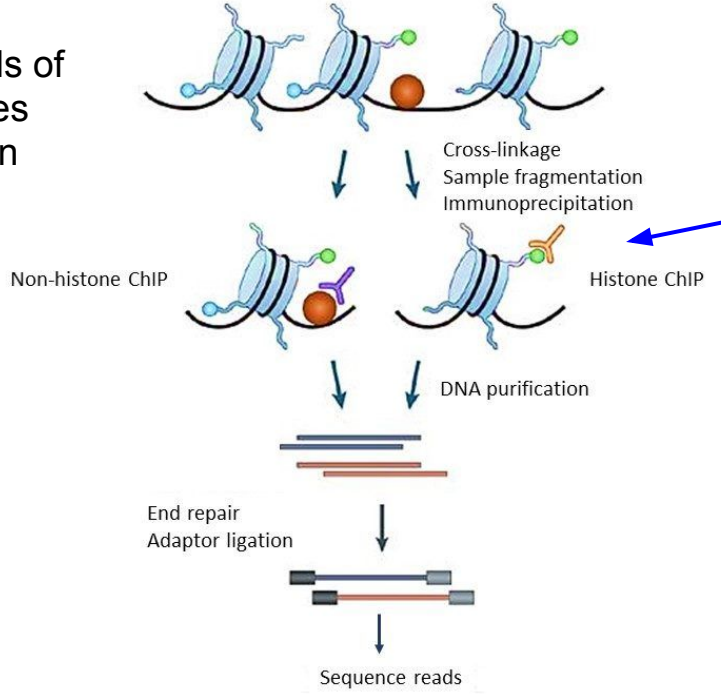
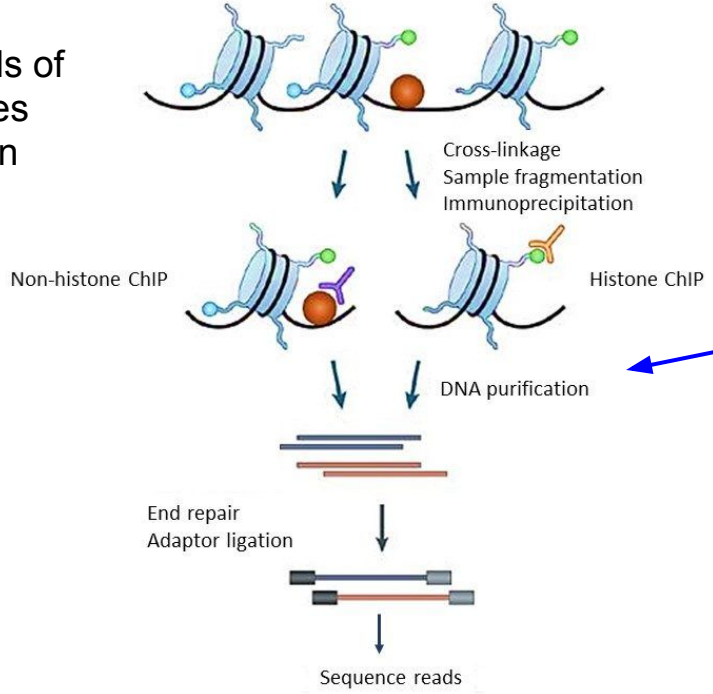


Figure credit: https://www.france-genomique.org/wp-content/uploads/2019/08/CHIP-selon-Park-1-e1566900408602.jpg

Disintegrate non-bound DNA -> what is left is DNA segments bound to protein

Data: ChIP-seq

Produces reads of DNA sequences where a protein binds



Treat sample to remove proteins

Figure credit: https://www.france-genomique.org/wp-content/uploads/2019/08/CHIP-selon-Park-1-e1566900408602.jpg

Data: ChIP-seq

Produces reads of DNA sequences where a protein binds

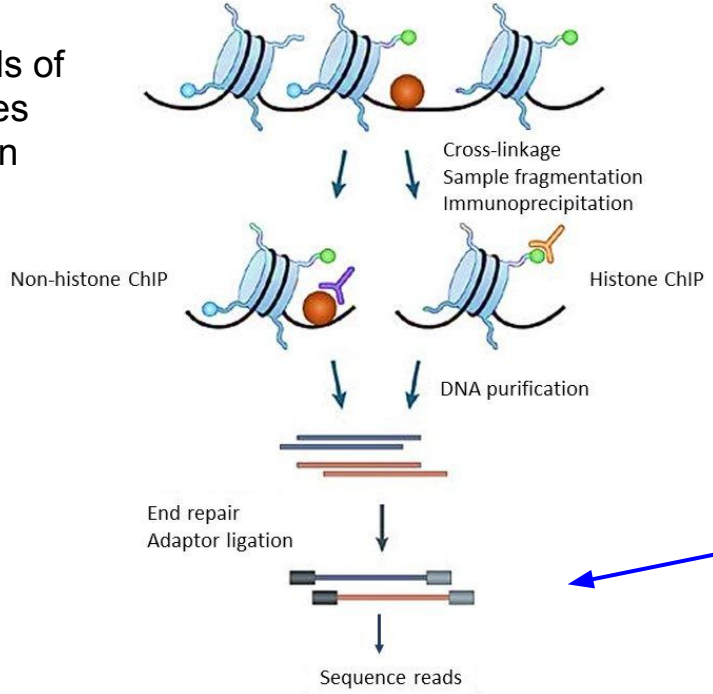


Figure credit: https://www.france-genomique.org/wp-content/uploads/2019/08/CHIP-selon-Park-1-e1566900408602.jpg

Use NGS to read-out remaining DNA sequences

Data: ChIP-seq

Produces reads of DNA sequences where a protein binds

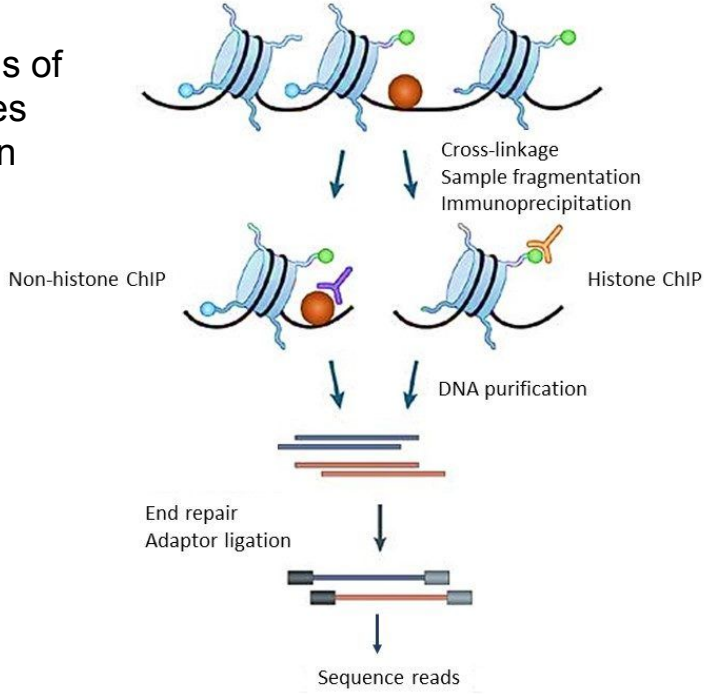


Figure credit:
https://www.france-genomique.org/wp-content/uploads/2019/08/CHIP-selon-Park-1-e1566900408602.jpg

Visualize distribution of locations on DNA where protein binds

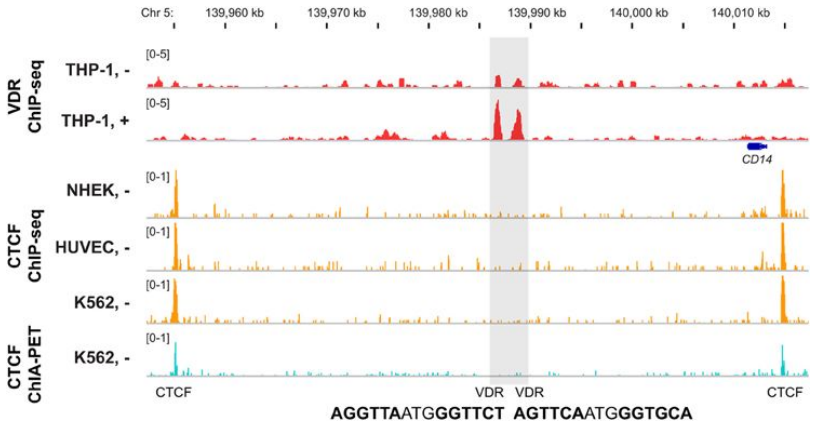


Figure credit:
https://www.researchgate.net/publication/262150050/figure/fig2/AS:272566950559751@1441996433141/Chromatin-domain-containing-VDR-binding-sites-The-IGV-browser-was-used-to-display-the.png

ENCODE: identifying and analyzing all functional elements in the human genome

- Launched by US National Human Genome Research Institute in 2003
- Contributions from worldwide consortium of research groups

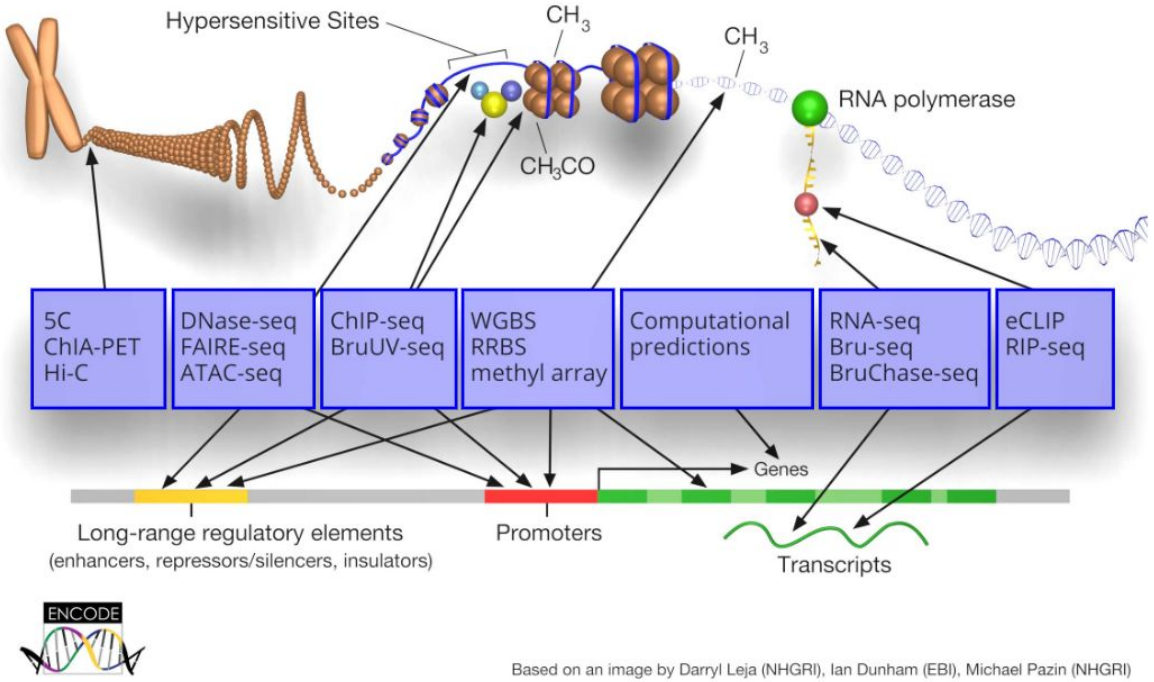
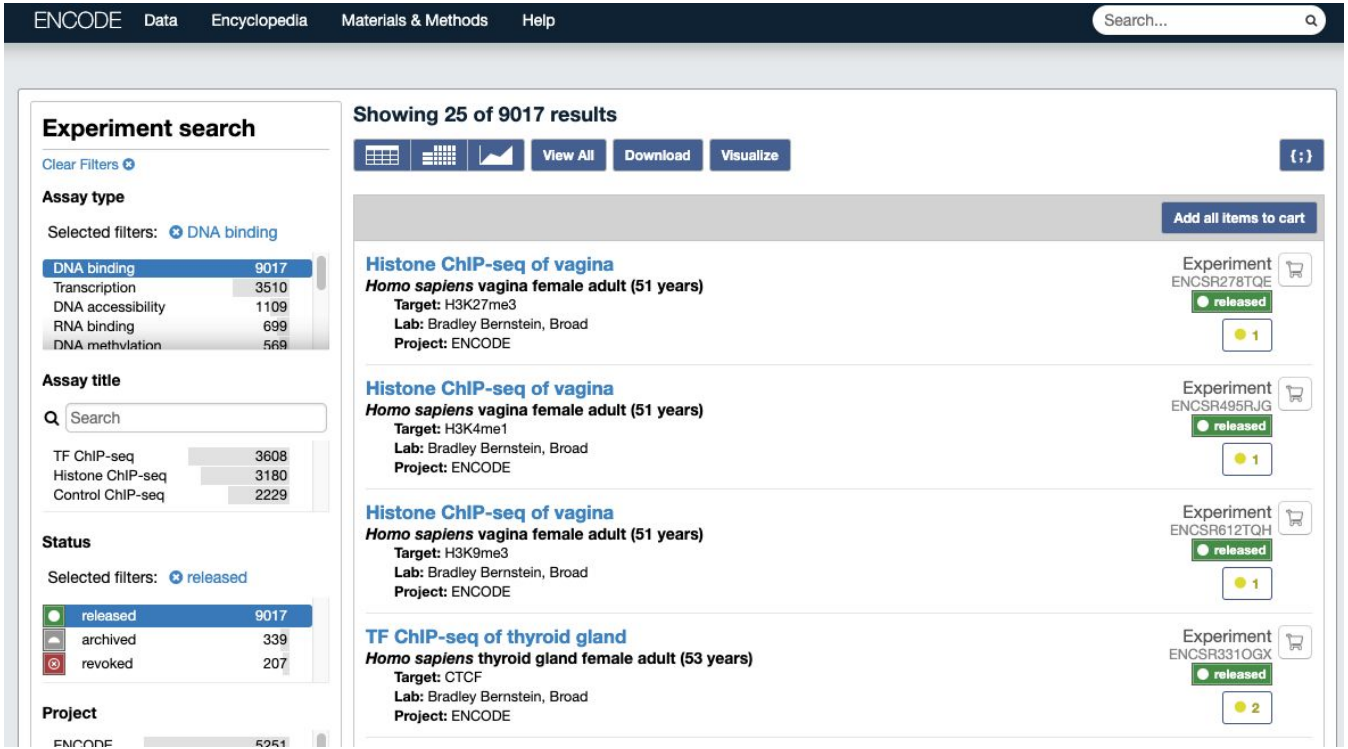
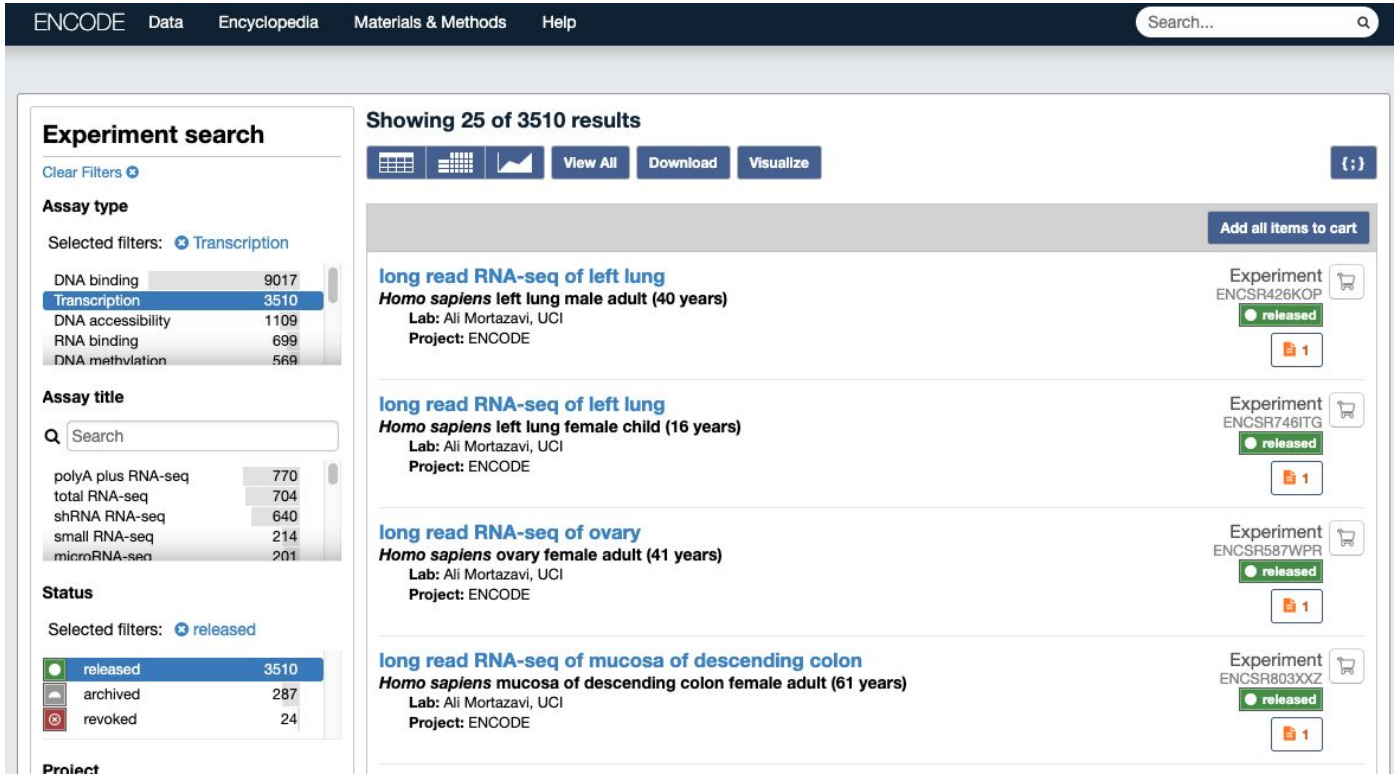


Figure credit: https://www.encodeproject.org/

ENCODE data


ENCODE
Data
Encyclopedia
Materials & Methods
Help
Search...
Experiment search
Clear Filters
Assay type
Selected filters: DNA binding
DNA binding 9017
Transcription 3510
DNA accessibility 1109
RNA binding 699
DNA methylation 569
Assay title
Search
TF ChIP-seq 3608
Histone ChIP-seq 3180
Control ChIP-seq 2229
Status
Selected filters: released
released 9017
archived 339
revoked 207
Project
ENCODE 5251
Showing 25 of 9017 results
View All
Download
Visualize
{;}
Add all items to cart
Histone ChIP-seq of vagina
Homo sapiens vagina female adult (51 years)
Target: H3K27me3
Lab: Bradley Bernstein, Broad
Project: ENCODE
Experiment
ENCSR278TQE
released
1
Histone ChIP-seq of vagina
Homo sapiens vagina female adult (51 years)
Target: H3K4me1
Lab: Bradley Bernstein, Broad
Project: ENCODE
Experiment
ENCSR495RJG
released
1
Histone ChIP-seq of vagina
Homo sapiens vagina female adult (51 years)
Target: H3K9me3
Lab: Bradley Bernstein, Broad
Project: ENCODE
Experiment
ENCSR612TQH
released
1
TF ChIP-seq of thyroid gland
Homo sapiens thyroid gland female adult (53 years)
Target: CTCF
Lab: Bradley Bernstein, Broad
Project: ENCODE
Experiment
ENCSR331OGX
released
2

ENCODE data


ENCODE
Data
Encyclopedia
Materials & Methods
Help
Search...
Experiment search
Clear Filters
Assay type
Selected filters: Transcription
DNA binding 9017
Transcription 3510
DNA accessibility 1109
RNA binding 699
DNA methylation 569
Assay title
Search
polyA plus RNA-seq 770
total RNA-seq 704
shRNA RNA-seq 640
small RNA-seq 214
microRNA-seq 201
Status
Selected filters: released
released 3510
archived 287
revoked 24
Project
Showing 25 of 3510 results
View All
Download
Visualize
{;}
Add all items to cart
long read RNA-seq of left lung
Homo sapiens left lung male adult (40 years)
Lab: Ali Mortazavi, UCI
Project: ENCODE
Experiment
ENCSR426KOP
released
1
long read RNA-seq of left lung
Homo sapiens left lung female child (16 years)
Lab: Ali Mortazavi, UCI
Project: ENCODE
Experiment
ENCSR746ITG
released
1
long read RNA-seq of ovary
Homo sapiens ovary female adult (41 years)
Lab: Ali Mortazavi, UCI
Project: ENCODE
Experiment
ENCSR587WPR
released
1
long read RNA-seq of mucosa of descending colon
Homo sapiens mucosa of descending colon female adult (61 years)
Lab: Ali Mortazavi, UCI
Project: ENCODE
Experiment
ENCSR803XXZ
released
1

ENCODE data

Common Cell Types: Tier 1 and Tier 2

Cell, tissue or DNA sample: Cell line or tissue used as the source of experimental material.

cell[1]	Tier[2]	Description[3]	Lineage[4]	Tissue[5]	Karyotype	Sex	Documents	Vendor ID
GM12878	1	B-lymphocyte, lymphoblastoid, International HapMap Project - CEPH/Utah - European Caucasion, Epstein-Barr Virus	mesoderm	blood	normal	F	ENCODE	Coriell GM12878
H1-hESC	1	embryonic stem cells	inner cell mass	embryonic stem cell	normal	M	ENCODE	WiCell Research Institute WA01
K562	1	leukemia, "The continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC	mesoderm	blood	cancer	F	ENCODE	ATCC CCL-243

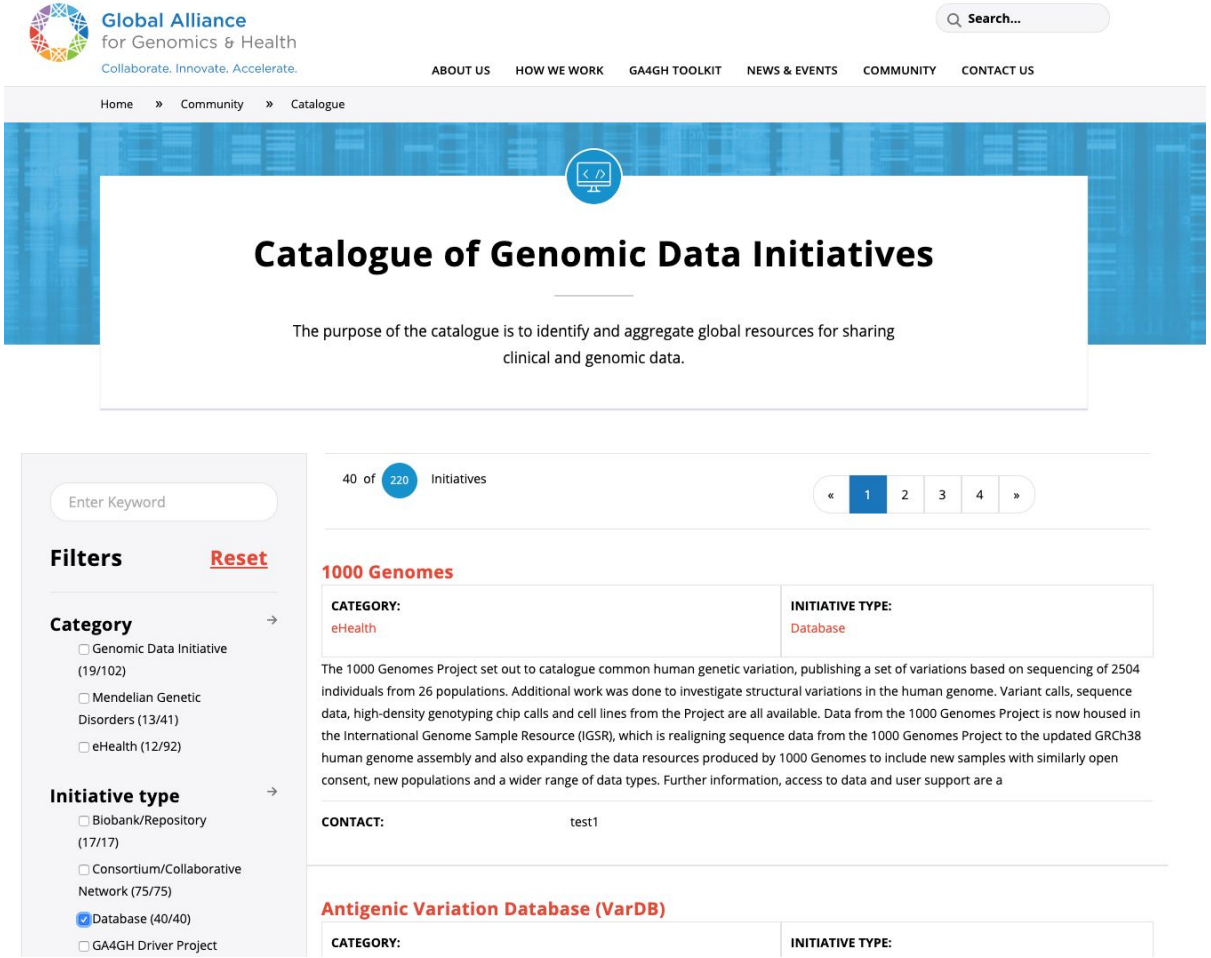
Total = 3

Cell, tissue or DNA sample: Cell line or tissue used as the source of experimental material.

cell[1]	Tier[2]	Description[3]	Lineage[4]	Tissue[5]	Karyotype	Sex	Documents	Vendor ID
A549	2	epithelial cell line derived from a lung carcinoma tissue. (PMID: 175022), "This line was initiated in 1972 by D.J. Giard, et al. through explant culture of lung carcinomatous tissue from a 58-year-old caucasian male." - ATCC, newly promoted to tier 2: not in 2011 analysis	endoderm	epithelium	cancer	M	Myers Crawford Stam	ATCC CCL-185

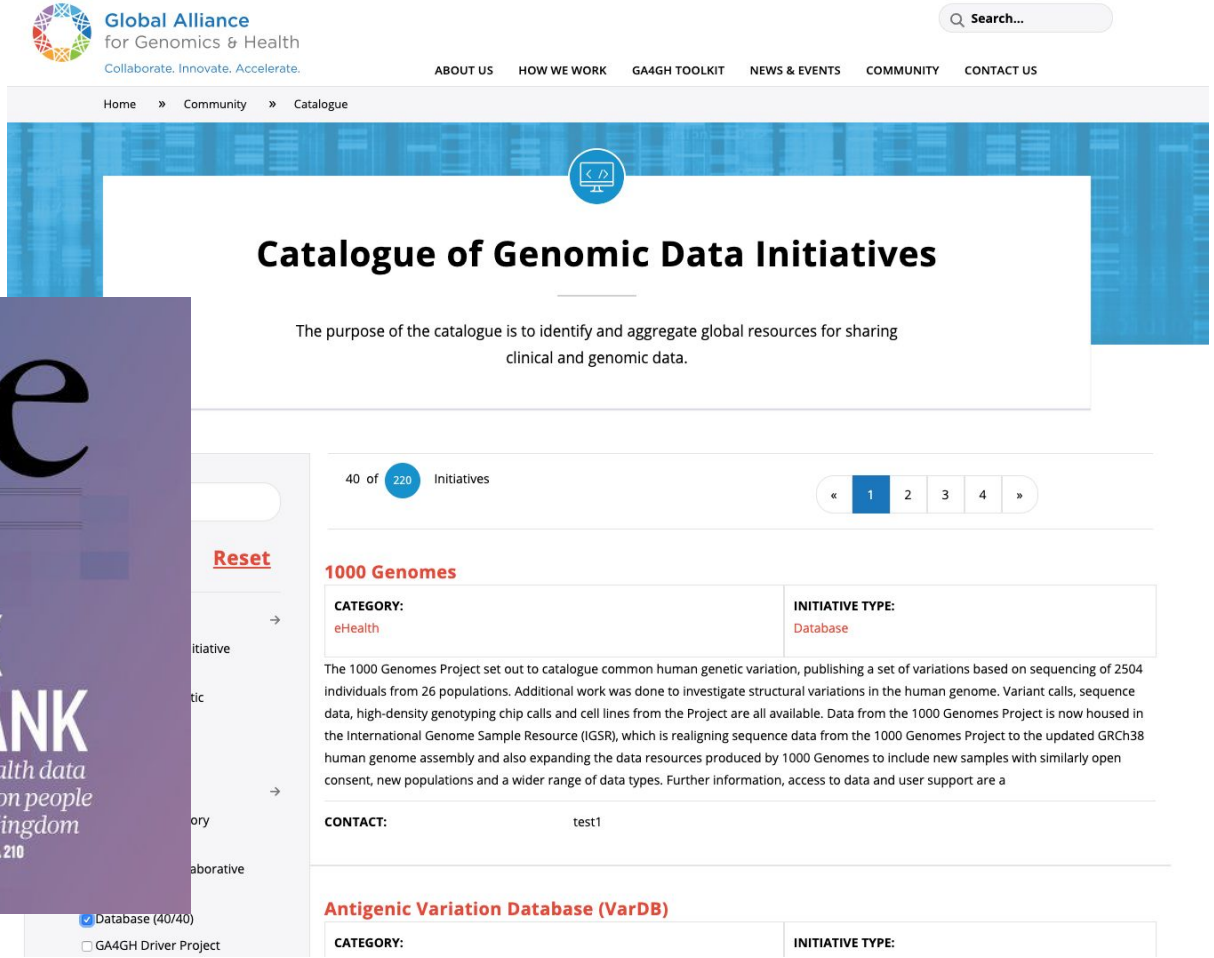
Other datasets

https://www.ga4gh.org/community/catalogue



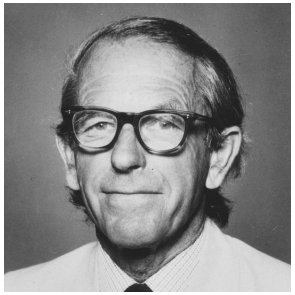
Other datasets

https://www.ga4gh.org/community/catalogue

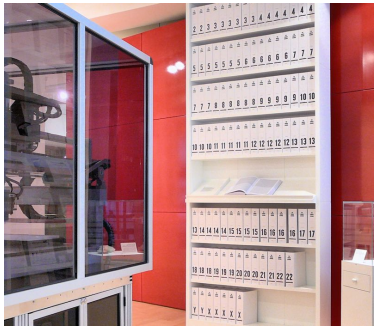


Genomics data

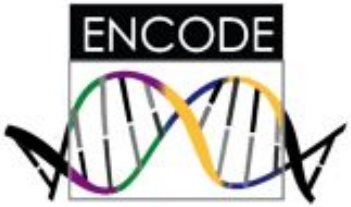
1953 - Watson and Crick discover double helix structures of DNA



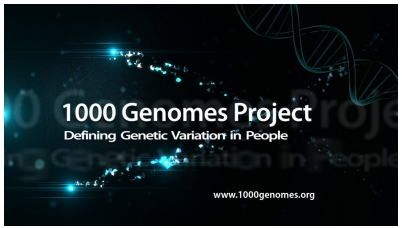
1977 - Fred Sanger sequences first full genome of a virus



1990 - 2003: Human Genome Project sequences full human genome



2003: ENCODE project launched to identify and characterize genes in human genome

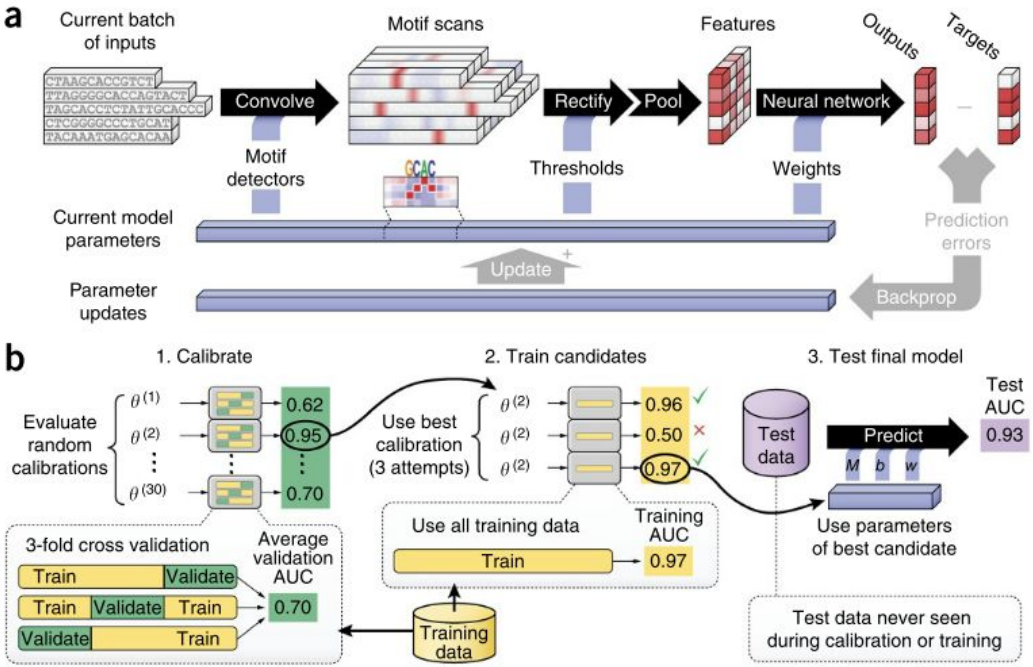


2008 - 2015: 1000 Genomes Project International effort to study human genetic variation

2006 - present: UK Biobank Project Genetic data and intended 30 years of health follow-up for 500k individuals in the UK

DeepBind

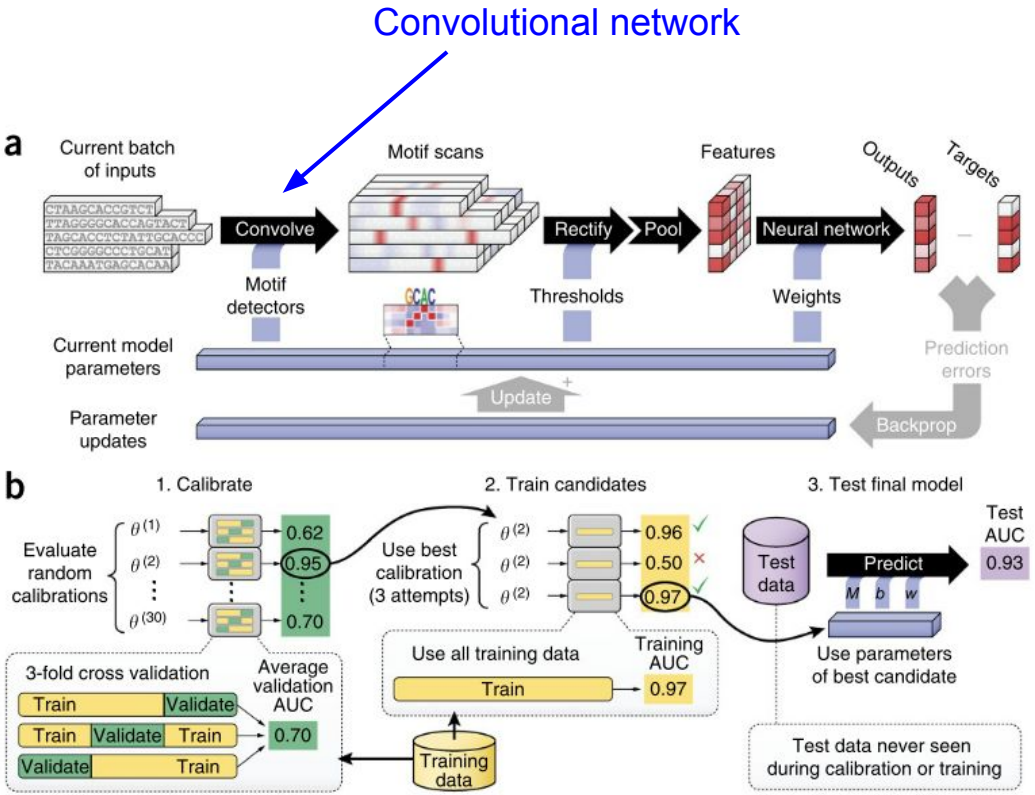
Input: DNA sequence
Output: Score of whether a particular protein will bind to the sequence or not



Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015.

DeepBind

Input: DNA sequence
Output: Score of whether a particular protein will bind to the sequence or not

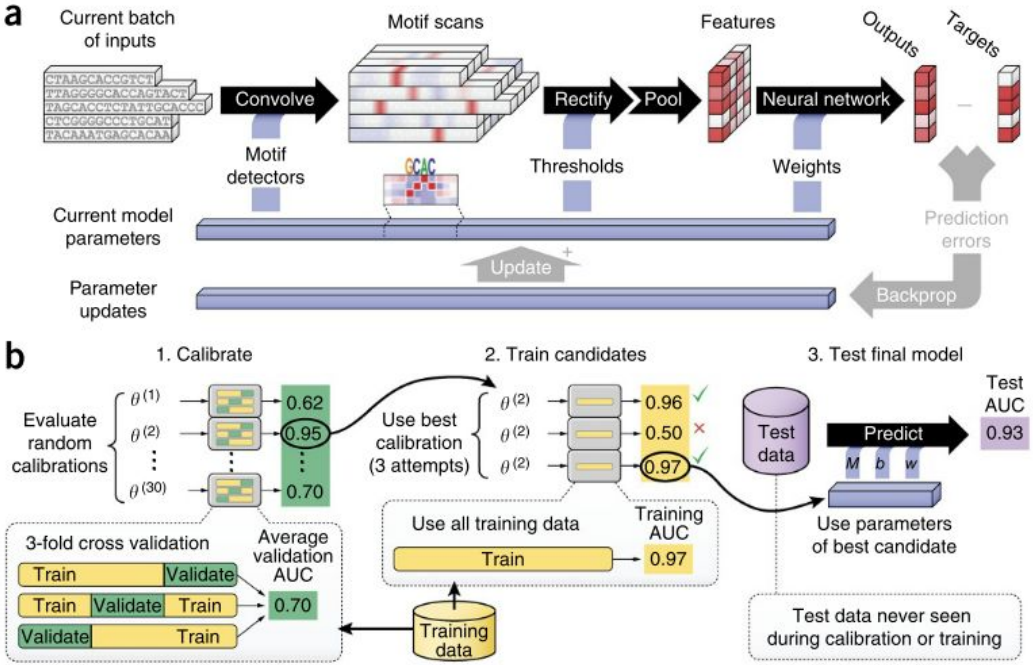


Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015.

DeepBind

Input: DNA sequence
Output: Score of whether a particular protein will bind to the sequence or not

- Processing to handle different sources of experimental (training) data and input / output data formats
- Trained on 12 TB of sequence data; learned 927 DeepBind models representing 538 transcription factor (TF) proteins and 194 RNA-binding proteins (RBPs)



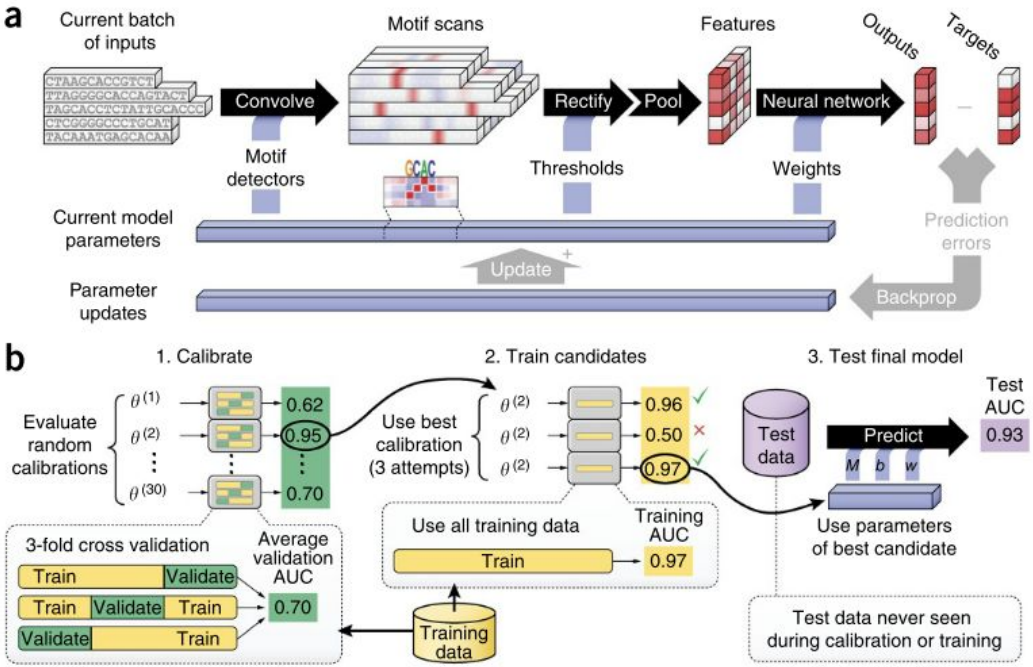
Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015.

DeepBind

Outperformed prior methods on the DREAM5 TF-DNA Motif Recognition Challenge

Input: DNA sequence
Output: Score of whether a particular protein will bind to the sequence or not

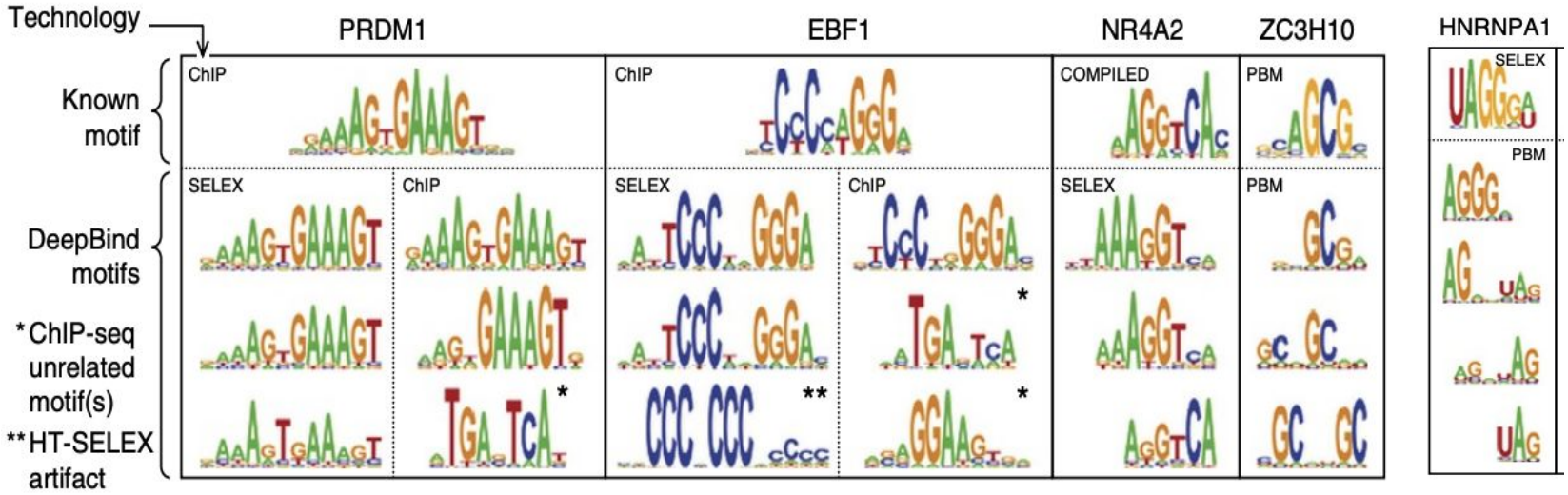
- Processing to handle different sources of experimental (training) data and input / output data formats
- Trained on 12 TB of sequence data; learned 927 DeepBind models representing 538 transcription factor (TF) proteins and 194 RNA-binding proteins (RBPs)



Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015.

DeepBind

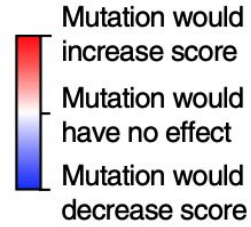
Learned DeepBind motifs

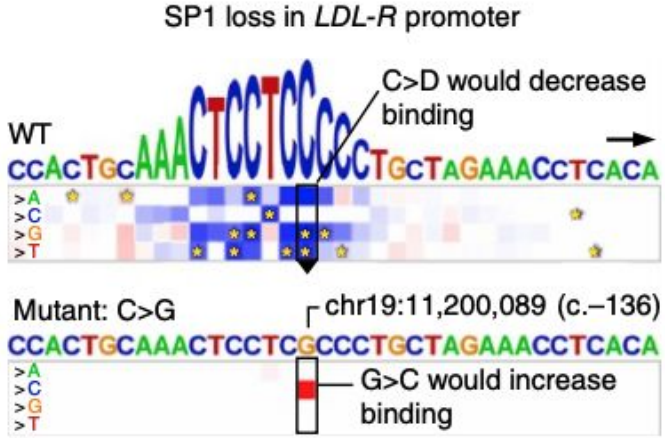


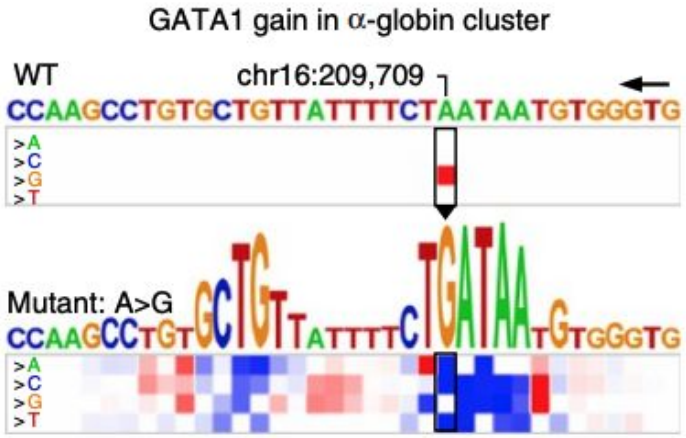
Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015.

DeepBind

Predicted effect of sequence mutations



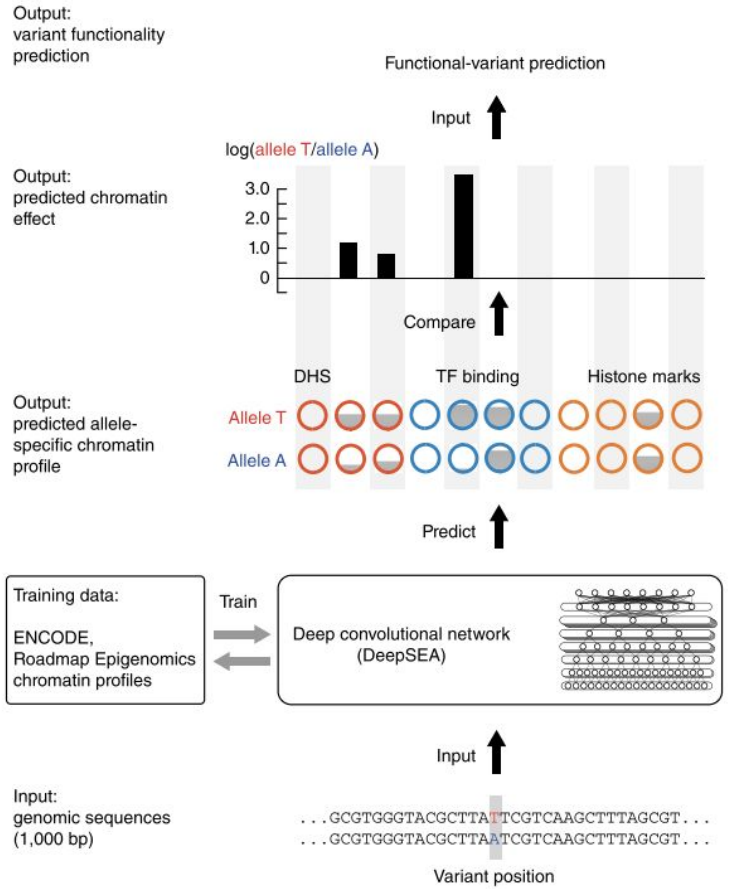




Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015.

DeepSea

Predict chromatin effects of (non-coding) sequence alterations with single-nucleotide sensitivity (SNPs: single nucleotide polymorphism)



Zhou and Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 2015.

DeepSea

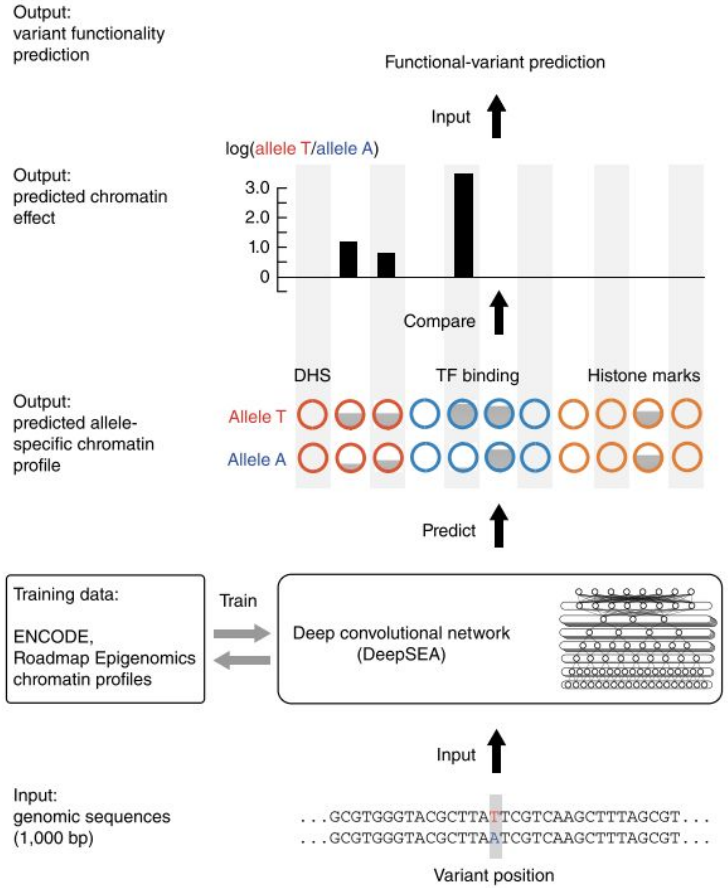
Predict chromatin effects of (non-coding) sequence alterations with single-nucleotide sensitivity (SNPs: single nucleotide polymorphism)

Input: DNA sequence pair with SNP

Output: Predicted chromatin effects (919 total)

- 690 transcription factor profiles
- 125 DNase I hypersensitive sites (DHS) profiles (looser chromatin structure, easier protein binding)
- 104 histone-mark profiles (histone modifications)

Multi-task training!



Zhou and Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 2015.

DeepSea

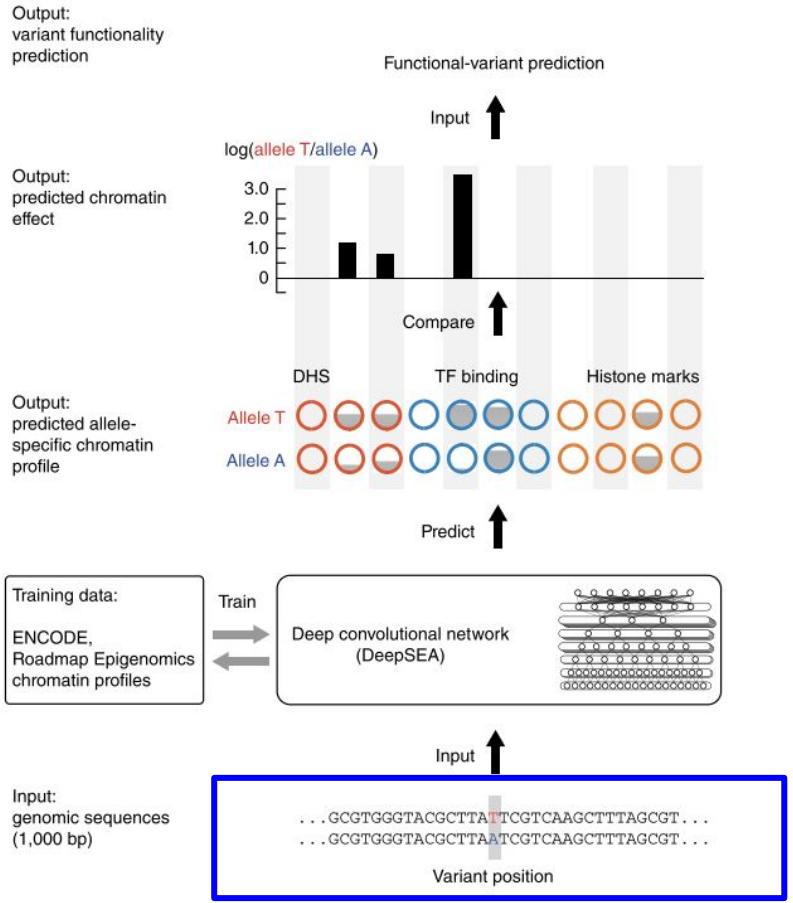
Predict chromatin effects of (non-coding) sequence alterations with single-nucleotide sensitivity (SNPs: single nucleotide polymorphism)

Input: DNA sequence pair with SNP

Output: Predicted chromatin effects (919 total)

- 690 transcription factor profiles
- 125 DNase I hypersensitive sites (DHS) profiles (looser chromatin structure, easier protein binding)
- 104 histone-mark profiles (histone modifications)

Multi-task training!



Zhou and Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 2015.

DeepSea

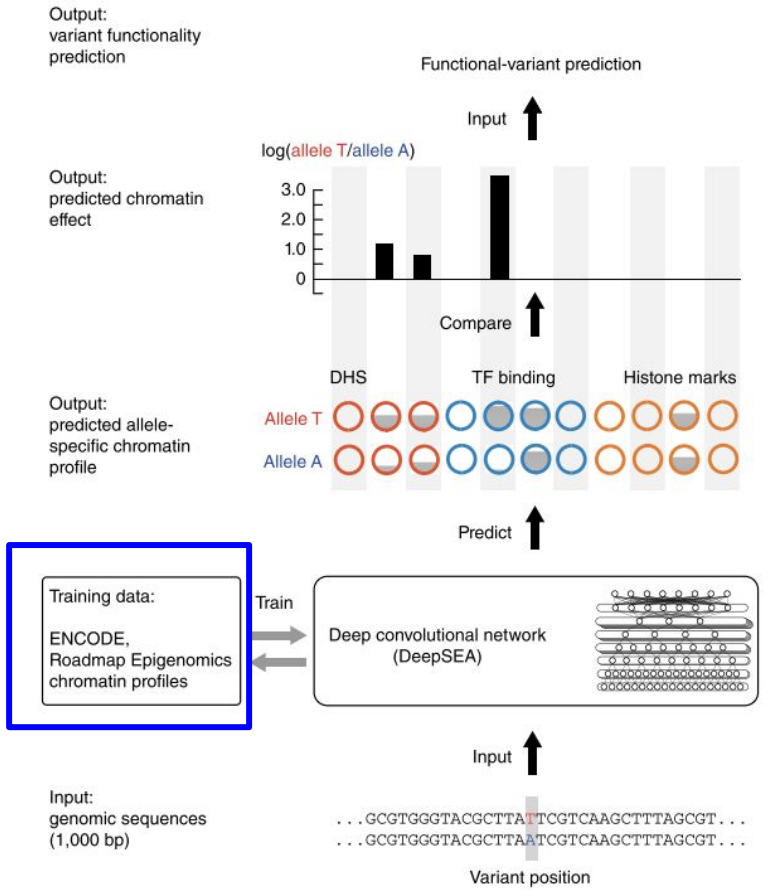
Predict chromatin effects of (non-coding) sequence alterations with single-nucleotide sensitivity (SNPs: single nucleotide polymorphism)

Input: DNA sequence pair with SNP

Output: Predicted chromatin effects (919 total)

- 690 transcription factor profiles
- 125 DNase I hypersensitive sites (DHS) profiles (looser chromatin structure, easier protein binding)
- 104 histone-mark profiles (histone modifications)

Multi-task training!



Zhou and Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 2015.

DeepSea

Predict chromatin effects of (non-coding) sequence alterations with single-nucleotide sensitivity (SNPs: single nucleotide polymorphism)

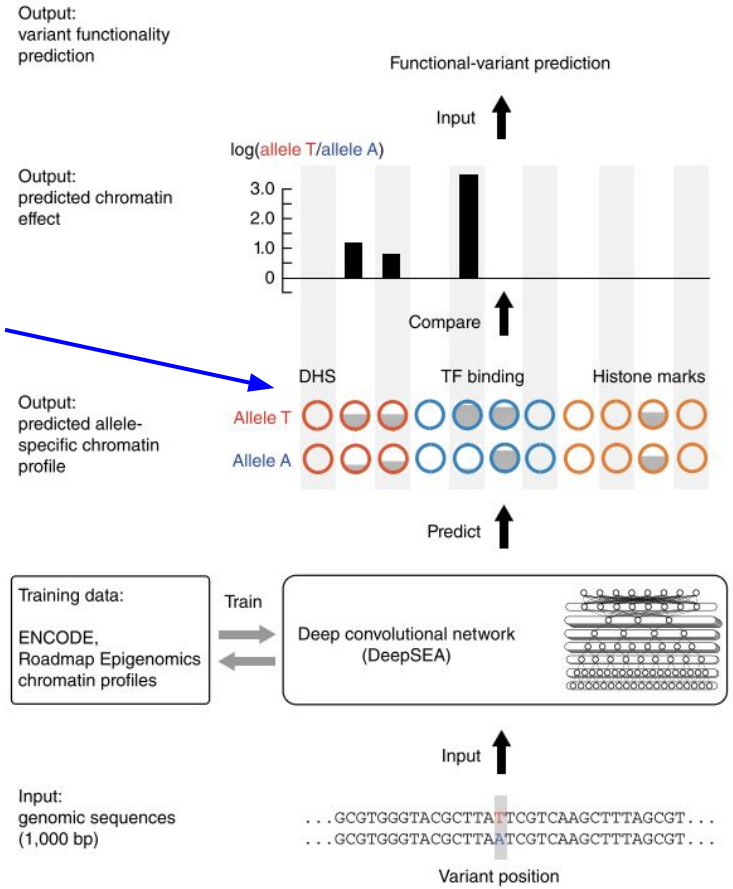
Multi-task prediction of 919 chromatin profiles, for each allele (variant)

Input: DNA sequence pair with SNP

Output: Predicted chromatin effects (919 total)

- 690 transcription factor profiles
- 125 DNase I hypersensitive sites (DHS) profiles (looser chromatin structure, easier protein binding)
- 104 histone-mark profiles (histone modifications)

Multi-task training!



Zhou and Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 2015.

DeepSea

Predict chromatin effects of (non-coding) sequence alterations with single-nucleotide sensitivity (SNPs: single nucleotide polymorphism)

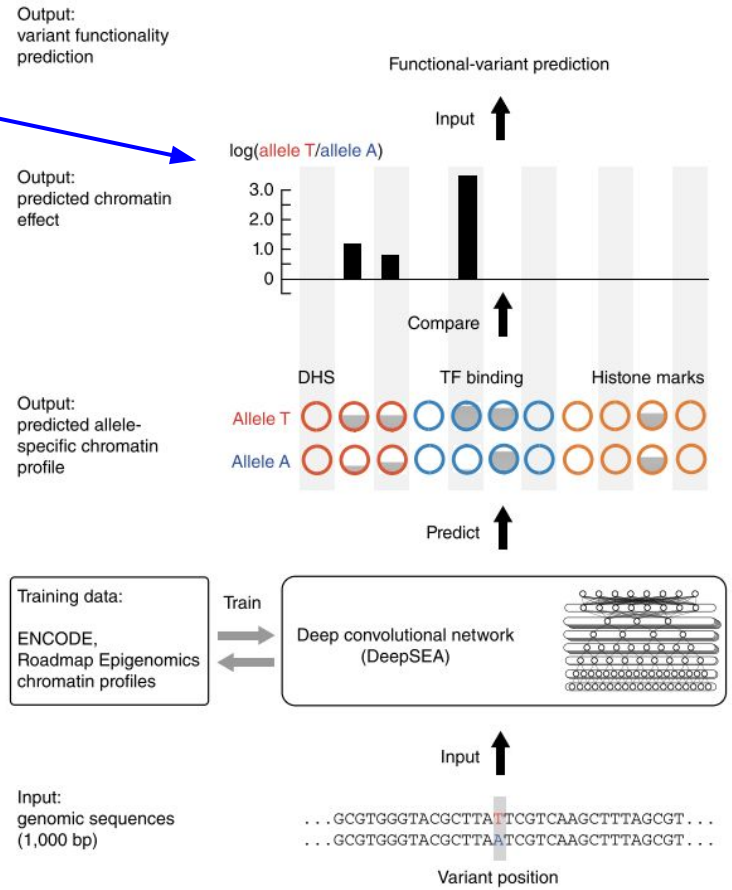
Input: DNA sequence pair with SNP

Output: Predicted chromatin effects (919 total)

- 690 transcription factor profiles
- 125 DNase I hypersensitive sites (DHS) profiles (looser chromatin structure, easier protein binding)
- 104 histone-mark profiles (histone modifications)

Multi-task training!

Interested in relative effect



Zhou and Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 2015.

DeepSea

Model Architecture:

1. Convolution layer (320 kernels. Window size: 8. Step size: 1.)
2. Pooling layer (Window size: 4. Step size: 4.)
3. Convolution layer (480 kernels. Window size: 8. Step size: 1.)
4. Pooling layer (Window size: 4. Step size: 4.)
5. Convolution layer (960 kernels. Window size: 8. Step size: 1.)
6. Fully connected layer (925 neurons)
7. Sigmoid output layer

Zhou and Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 2015.

DeepVariant

Variant calling: identifying variants from reference genome (SNPs, small indels, etc.)

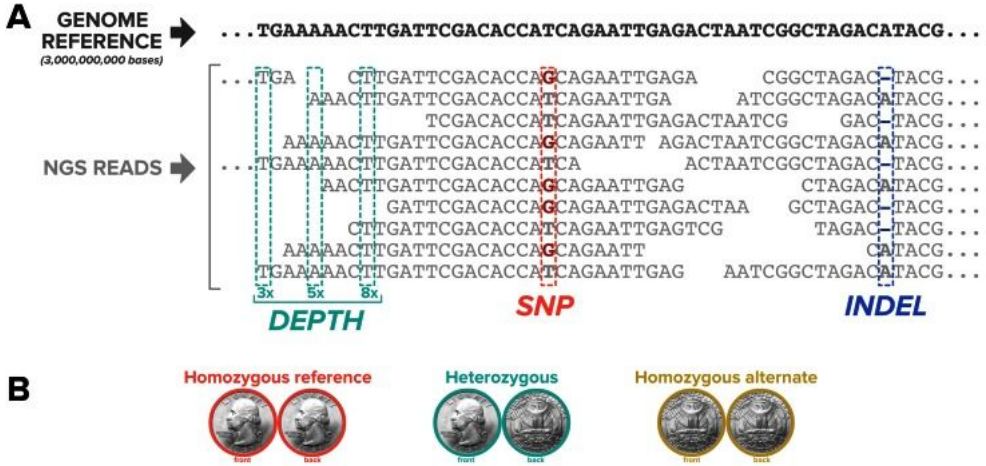


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig2_HTML.jpg

Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

DeepVariant

Variant calling: identifying variants from reference genome (SNPs, small indels, etc.)

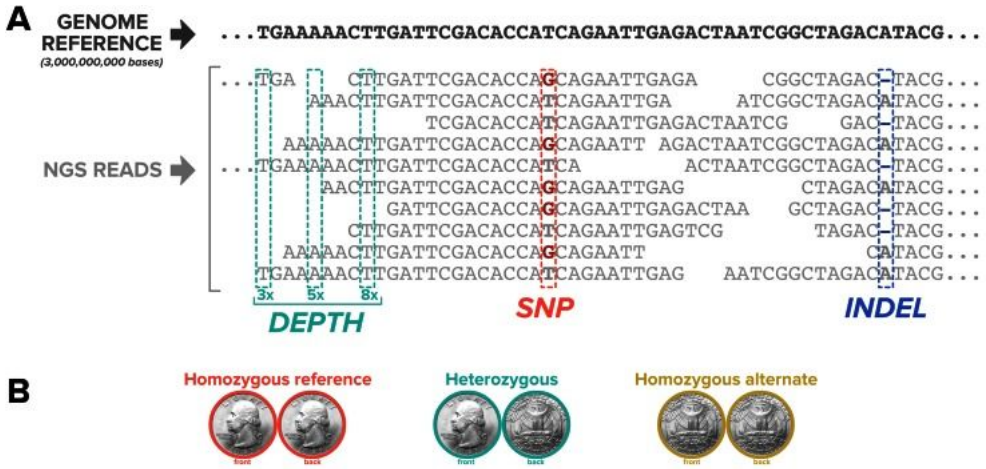
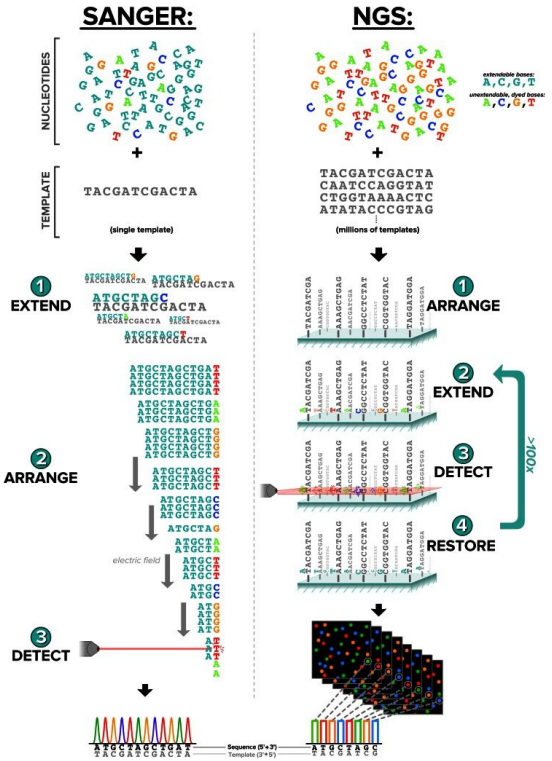


Figure credit: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4633438/bin/40142_2015_76_Fig2_HTML.jpg

Challenge with short, errorful sequence reads from NGS!

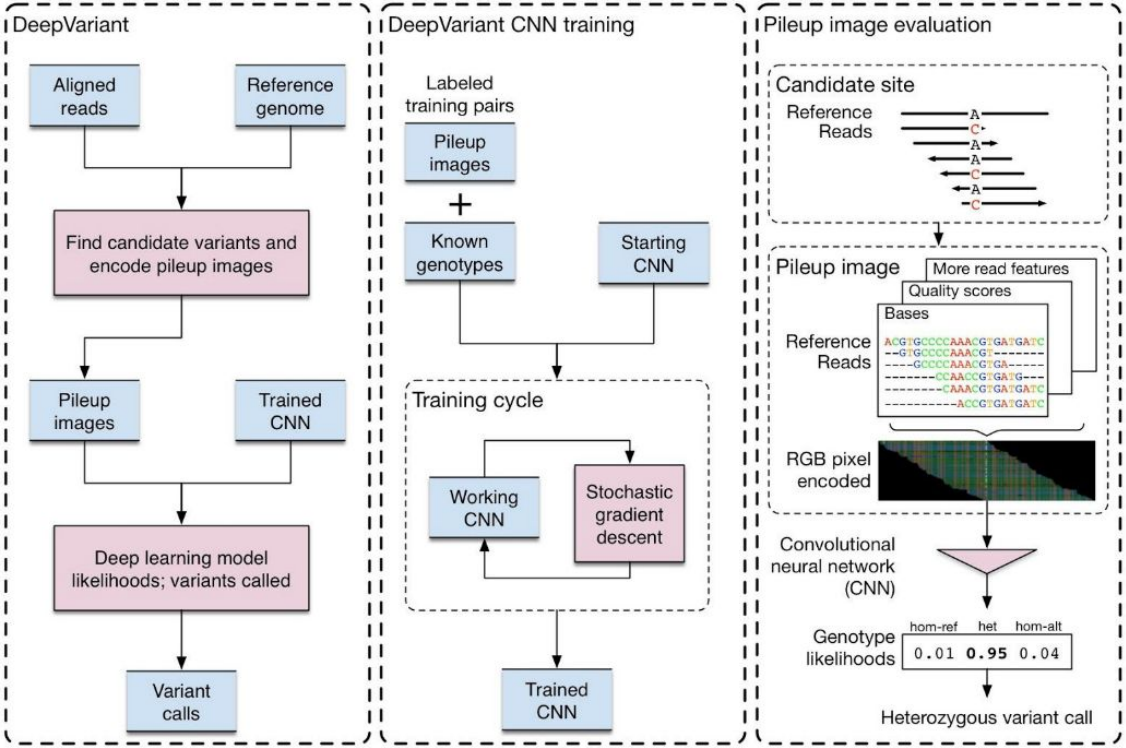


Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

DeepVariant

Input: “Pileup images” of reference sequence + NGS reads, + other features

Output: Categorical prediction of variant type (hom-ref, het, hom-alt), or no variant

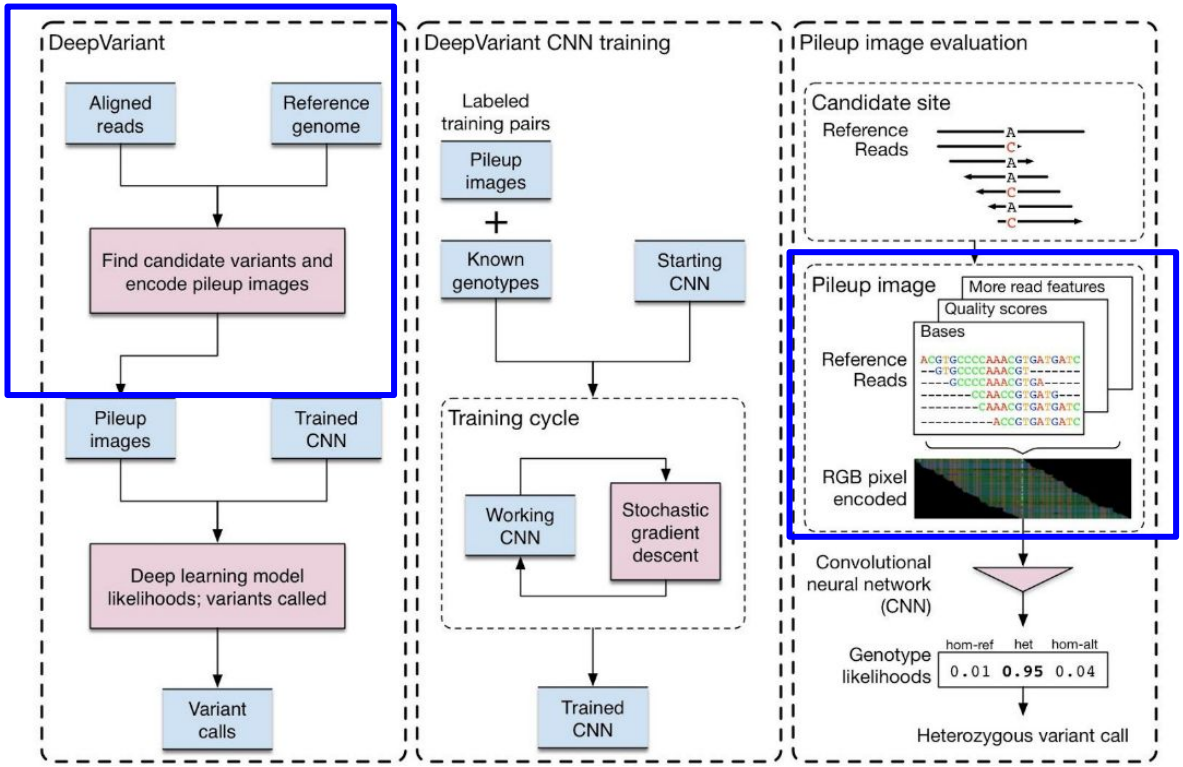


Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

DeepVariant

Input: “Pileup images” of reference sequence + NGS reads, + other features

Output: Categorical prediction of variant type (hom-ref, het, hom-alt), or no variant

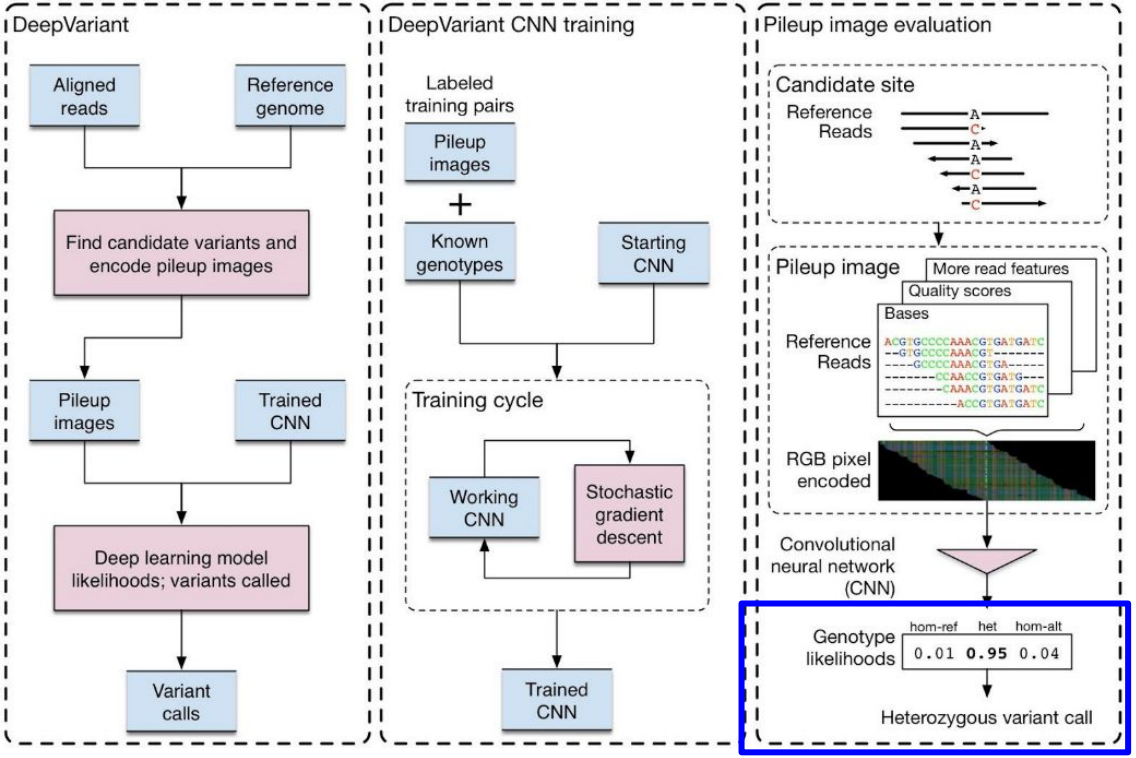


Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

DeepVariant

Input: “Pileup images” of reference sequence + NGS reads, + other features

Output: Categorical prediction of variant type (hom-ref, het, hom-alt), or no variant



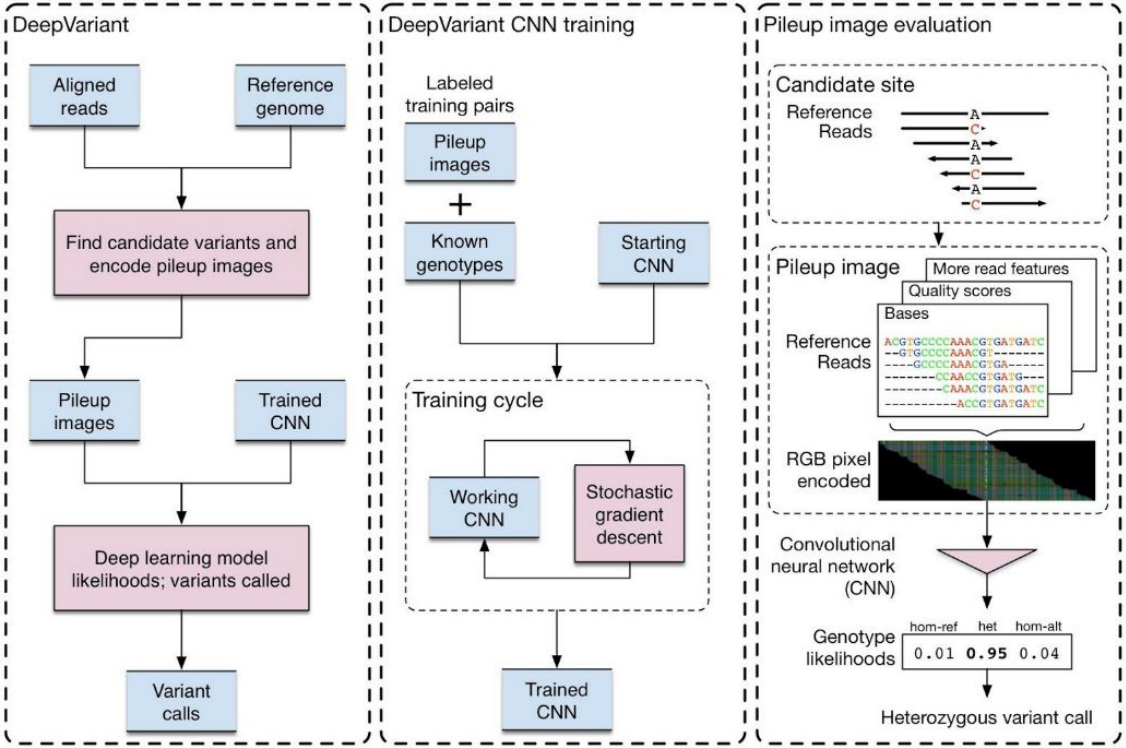
Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

DeepVariant

Input: “Pileup images” of reference sequence + NGS reads, + other features

Output: Categorical prediction of variant type (hom-ref, het, hom-alt), or no variant

Used an Inception v3 CNN



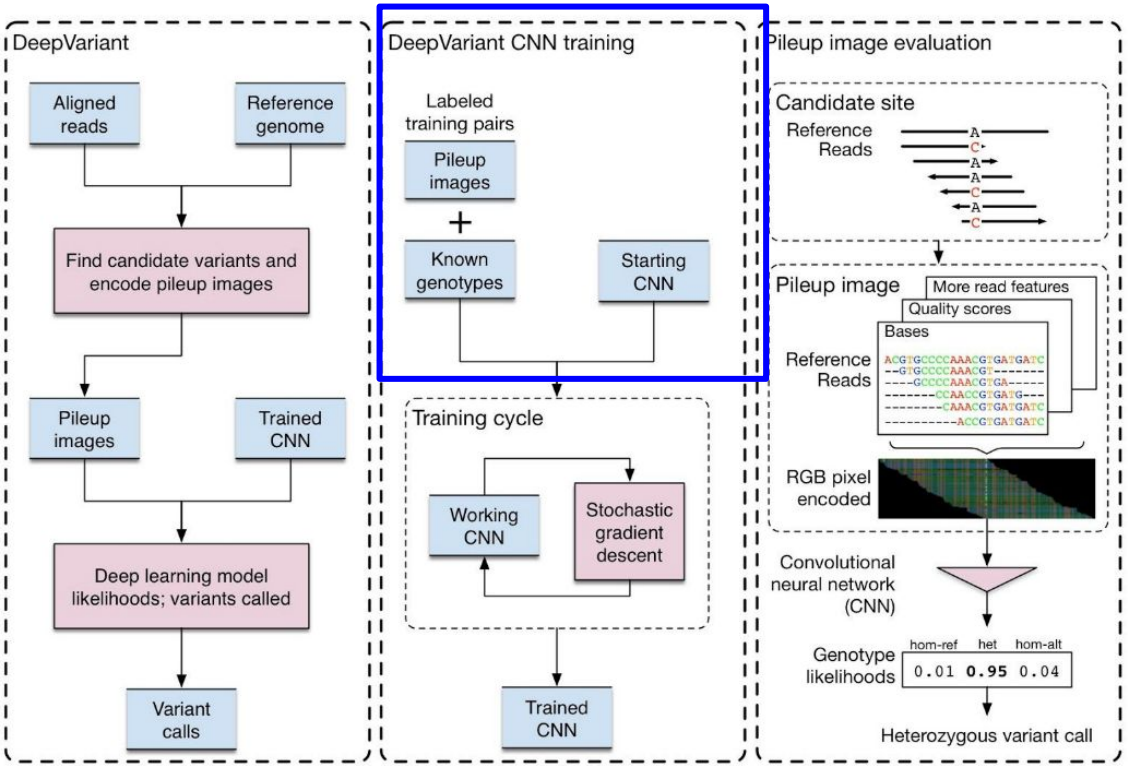
Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

DeepVariant

Input: “Pileup images” of reference sequence + NGS reads, + other features

Output: Categorical prediction of variant type (hom-ref, het, hom-alt), or no variant

Used an Inception v3 CNN



Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

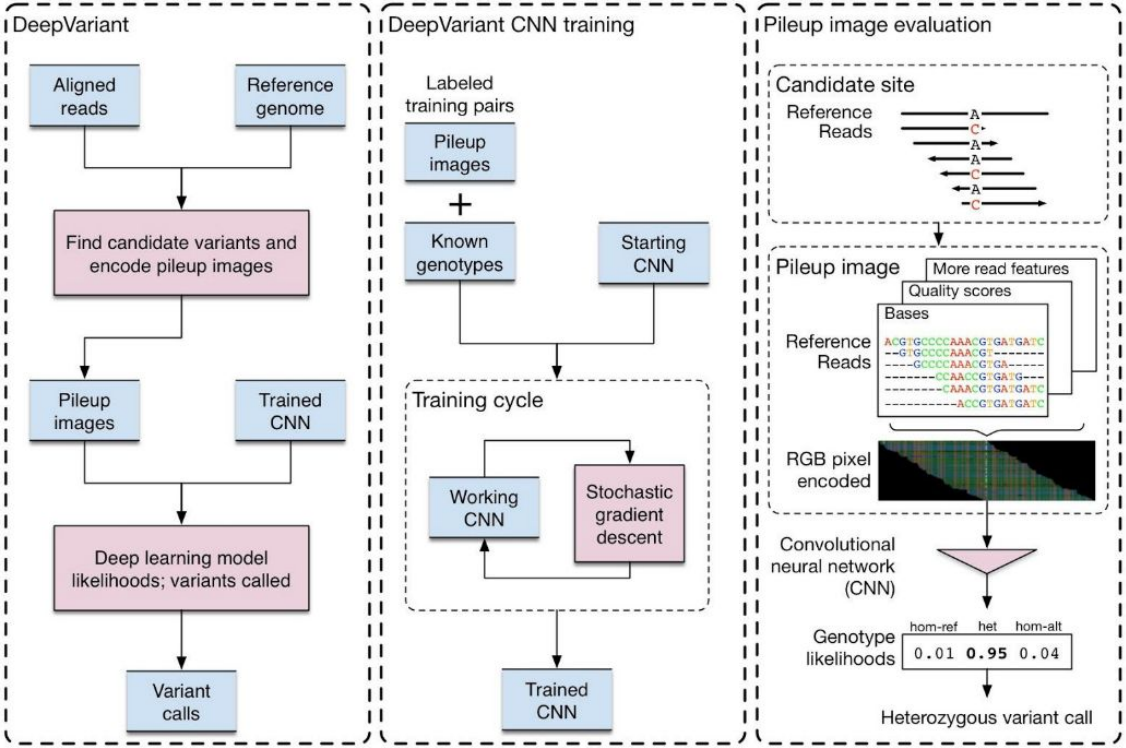
DeepVariant

Input: “Pileup images” of reference sequence + NGS reads, + other features

Output: Categorical prediction of variant type (hom-ref, het, hom-alt), or no variant

Used an Inception v3 CNN

Won highest performance for SNPs in the 2016 FDA variant calling Truth Challenge



Poplin et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 2018.

Remember: ChIP-seq

Produces reads of DNA sequences where a protein binds

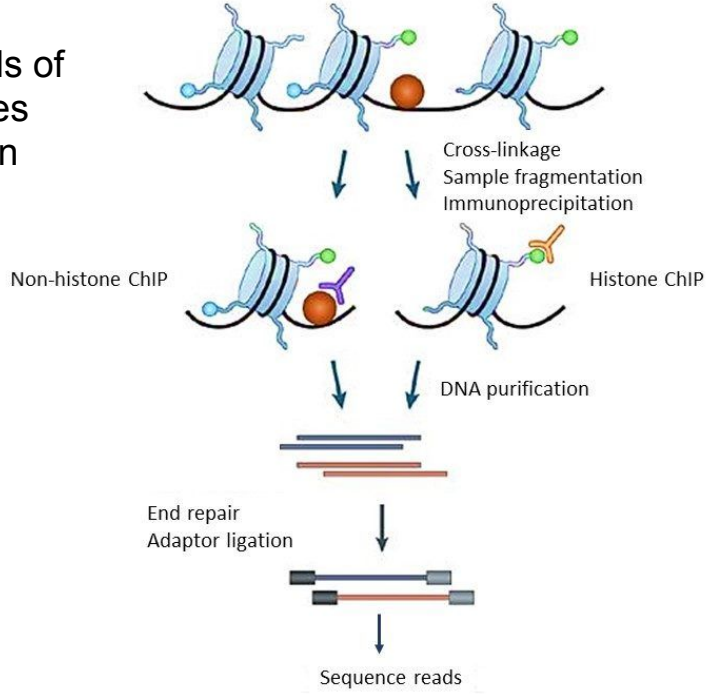


Figure credit:
https://www.france-genomique.org/wp-content/uploads/2019/08/CHIP-selon-Park-1-e1566900408602.jpg

Visualize distribution of locations on DNA where protein binds

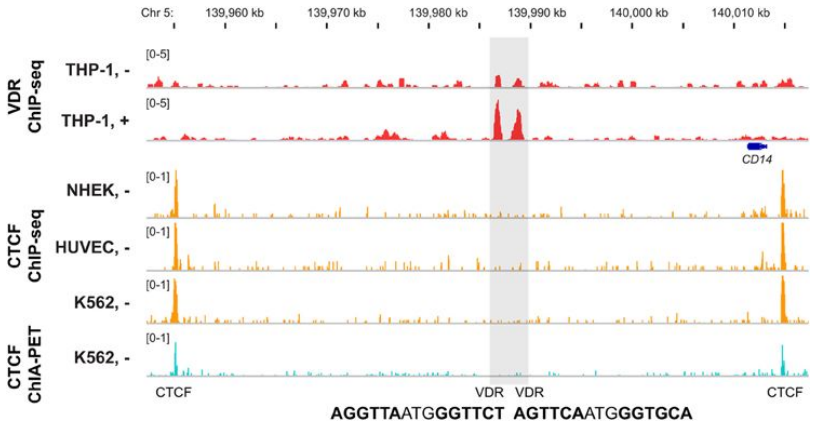
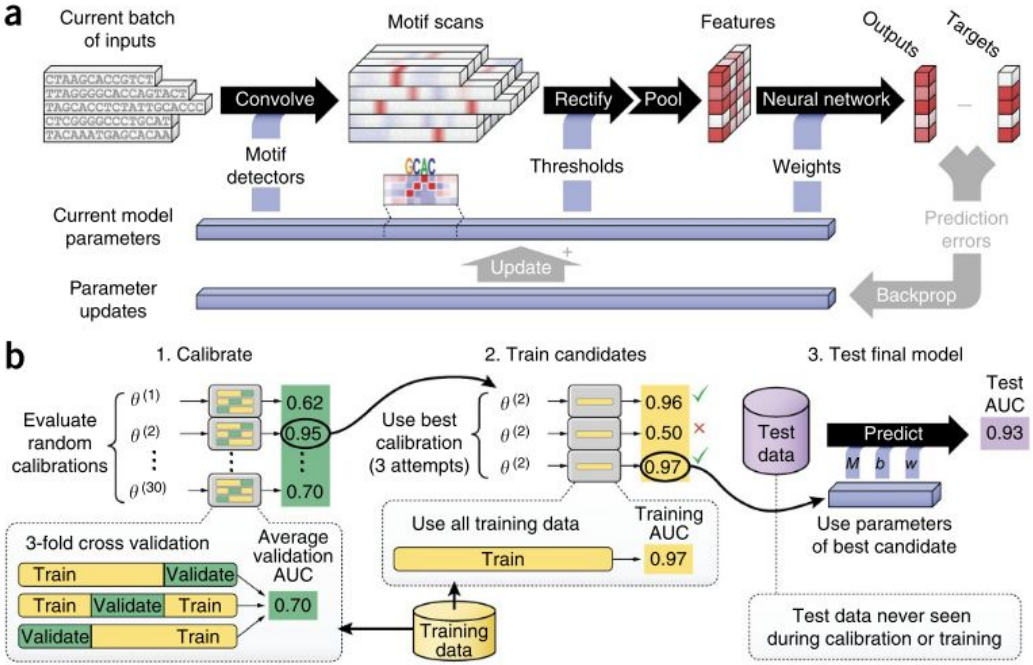


Figure credit:
https://www.researchgate.net/publication/262150050/figure/fig2/AS:272566950559751@1441996433141/Chromatin-domain-containing-VDR-binding-sites-The-IGV-browser-was-used-to-display-the.png

Remember: DeepBind

Input: DNA sequence
Output: Score of whether a particular protein will bind to the sequence or not

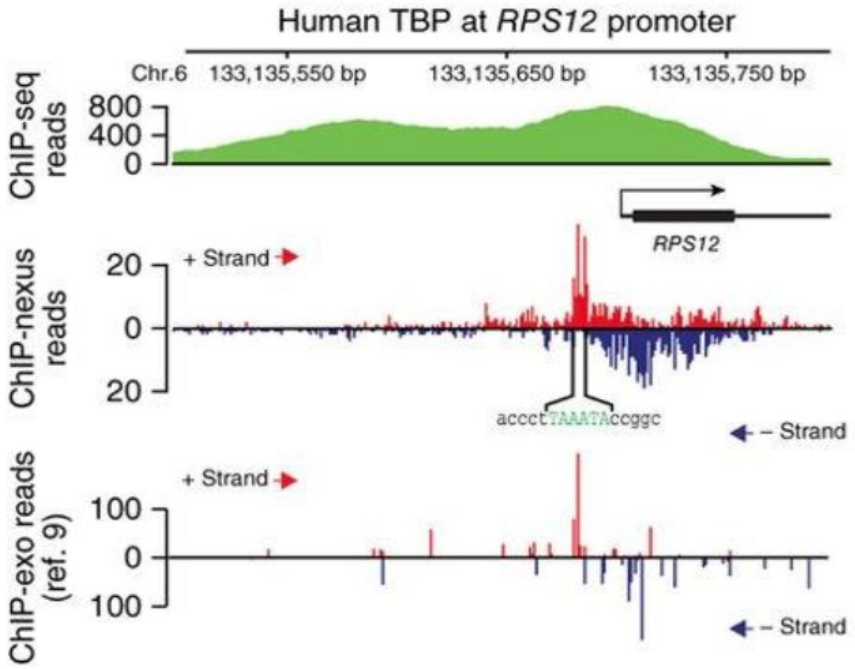
- Processing to handle different sources of experimental (training) data and input / output data formats
- Trained on 12 TB of sequence data; learned 927 DeepBind models representing 538 transcription factor (TF) proteins and 194 RNA-binding proteins (RBPs)



Alipanahi et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 2015.

More recently: ChIP-nexus vs. ChIP-seq

ChIP-nexus: newer technology that enables improved and higher-resolution data about transcription factor binding footprints on DNA (at individual base-pair resolution)

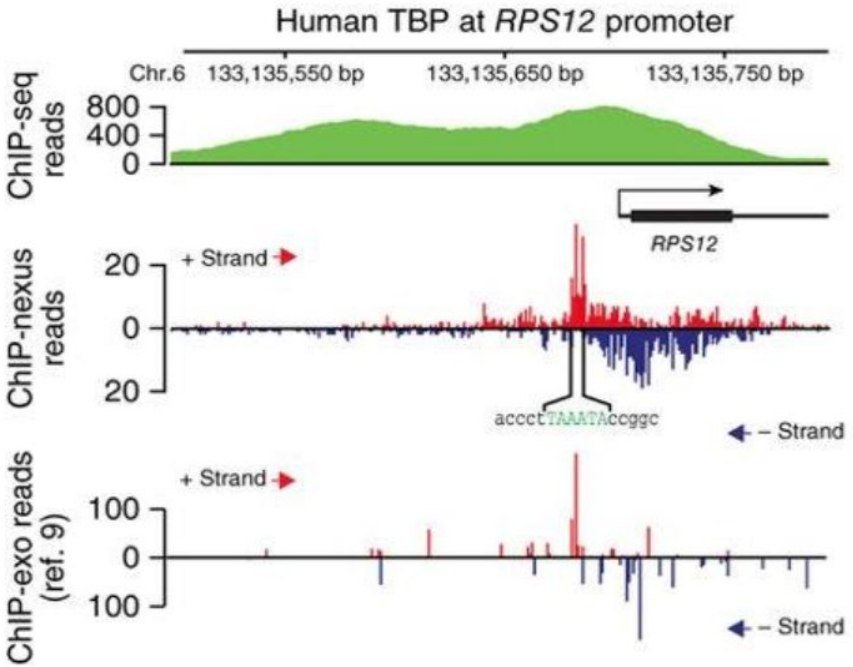


He et al. ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints. Nature Biotechnology, 2015.

More recently: ChIP-nexus vs. ChIP-seq

ChIP-nexus: newer technology that enables improved and higher-resolution data about transcription factor binding footprints on DNA (at individual base-pair resolution)

ChiP-seq

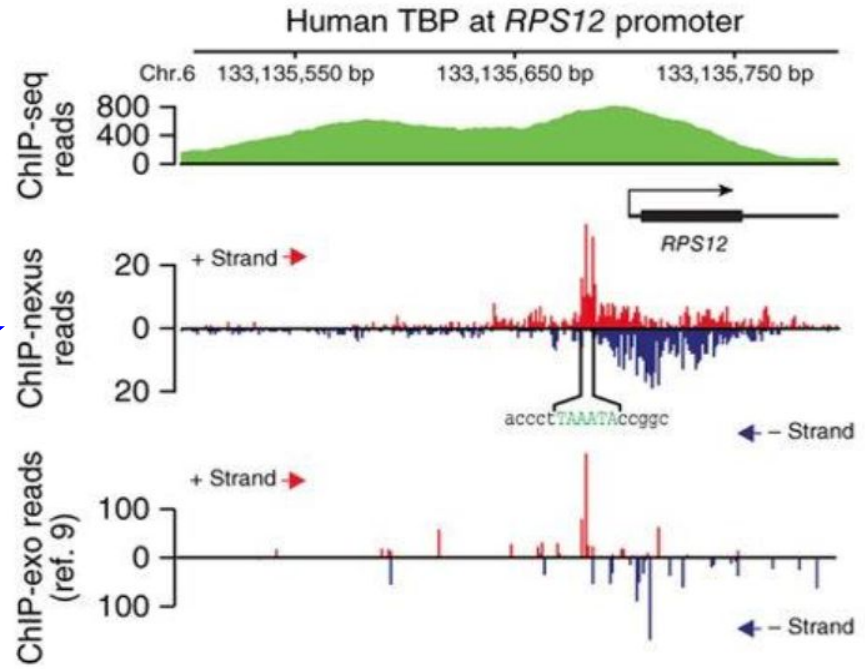


He et al. ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints. Nature Biotechnology, 2015.

More recently: ChIP-nexus vs. ChIP-seq

ChIP-nexus: newer technology that enables improved and higher-resolution data about transcription factor (TF) binding footprints on DNA (at individual base-pair resolution)

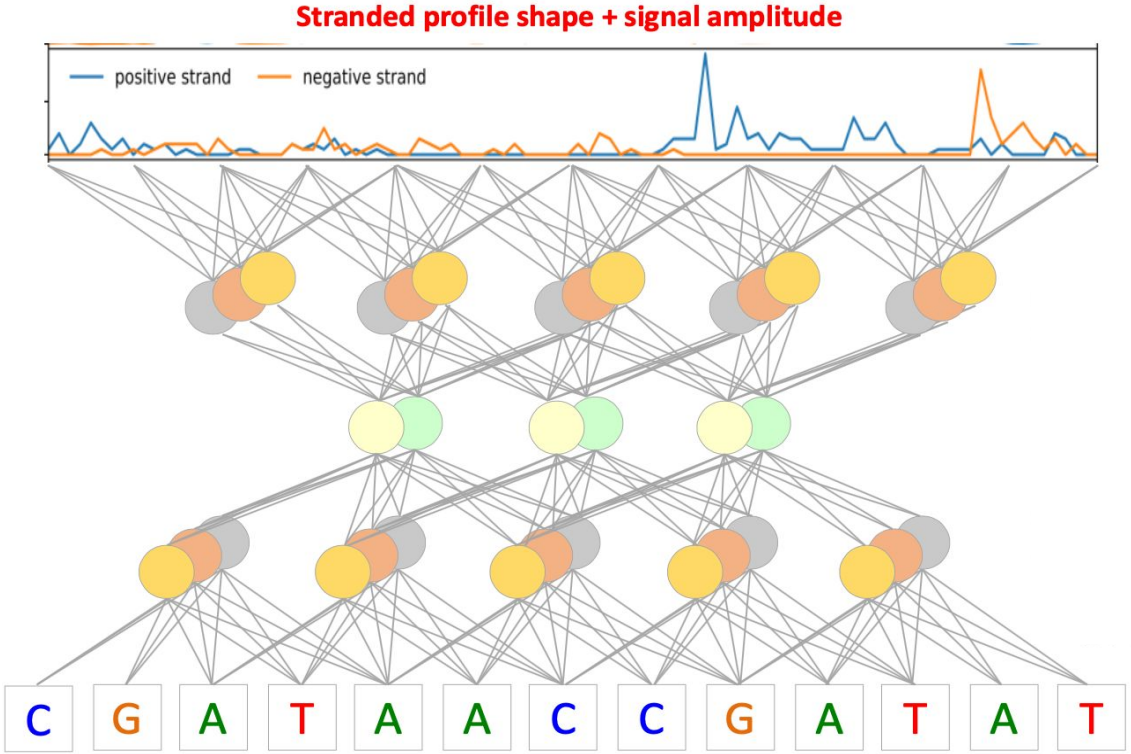
ChiP-nexus



He et al. ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints. Nature Biotechnology, 2015.

BPNet: DNA sequence to base-pair resolution profile regression

- Deep learning-based model based on ChiP-nexus data, that predicts TF binding profile at high, individual base-pair resolution

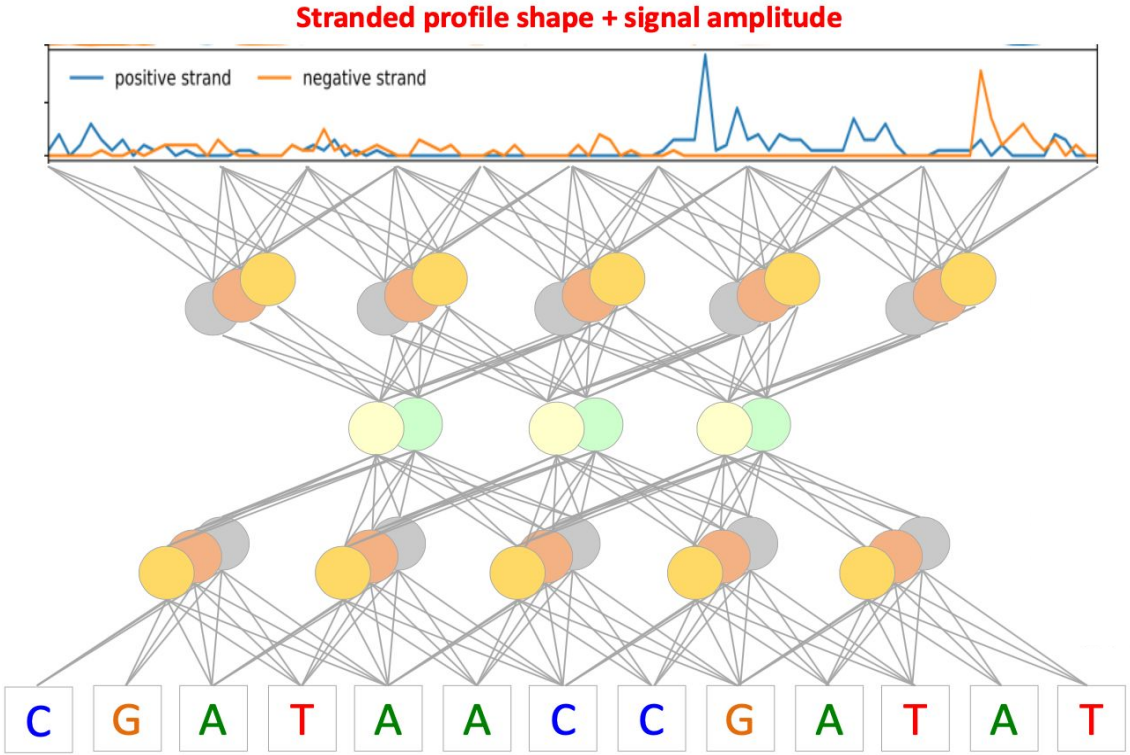


Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

Slide Credit: Anshul Kundaje

BPNet: DNA sequence to base-pair resolution profile regression

- Deep learning-based model based on ChiP-nexus data, that predicts TF binding profile at high, individual base-pair resolution
- Uses 1-D, **dilated** convolutional layers for greater increase of receptive field (extent of input used to produce a neuron output), instead of pooling layers -> maintains base-pair resolution

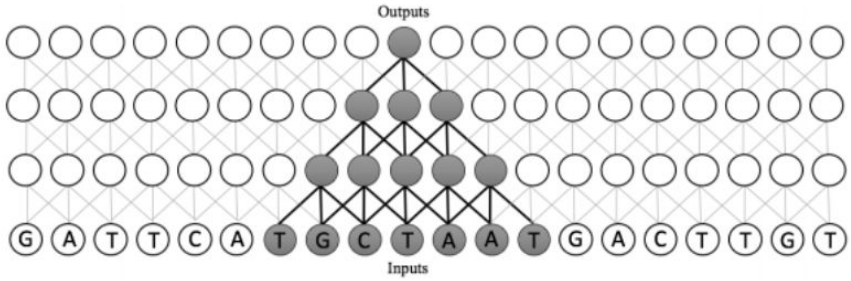


Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

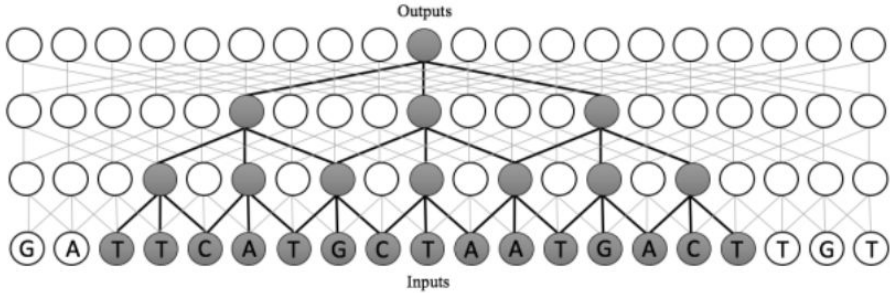
Slide Credit: Anshul Kundaje

Dilated convolutions instead of convolutions

- Greater increase of receptive field vs. standard convolution, for the same # of layers (avoids requiring many layers to increase receptive field which is more difficult to train)
- Pooling layers can also increase receptive field, but reduce resolution (whereas dilated convolutions can maintain high resolution)
- BPNet also includes residual connections (remember ResNets!) to improve ease of optimization for more effective training



(a) Convolution



(c) Dilated Convolution

Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.
Figure credit: Gupta et al. Dilated Convolutions for Modeling Long-Distance Genomic Dependencies, 2017.

Slide Credit: Anshul Kundaje

BPNet: Profile regression loss

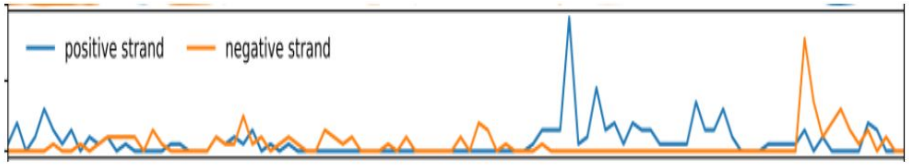
- Two-part loss function for optimizing prediction of the binding profile across the input sequence
  - MSE loss for log (total number of counts across the entire 1kb input sequence)
  - Multinomial loss for the likelihood of the observed count distribution over the sequence, compared to the predicted probabilities

Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

Slide Credit: Anshul Kundaje

BPNet: Profile regression loss

- Two-part loss function for optimizing prediction of the binding profile across the input sequence
  - MSE loss for log (total number of counts across the entire 1kb input sequence)
  - Multinomial loss for the likelihood of the observed count distribution over the sequence, compared to the predicted probabilities



Stranded profile shape + signal amplitude

$$Loss = -\log p_{mult.}(\mathbf{k}^{obs} \mid \mathbf{p}^{pred}, n^{obs}) + \lambda(\log(1+n^{obs}) - \log(1+n^{pred}))^2$$

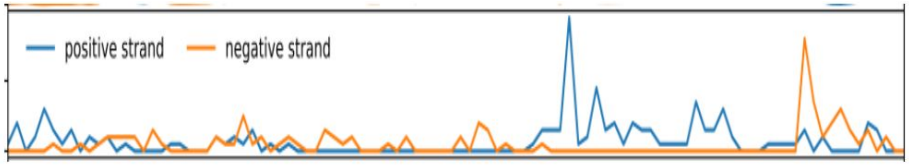
k^{obs}: vector of observed reads counts at each position
p^{pred}: learned multinomial prob. at each position
n^{obs}: total number of read counts across entire 1 kb

Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

Slide Credit: Anshul Kundaje

BPNet: Profile regression loss

- Two-part loss function for optimizing prediction of the binding profile across the input sequence
  - MSE loss for log (total number of counts across the entire 1kb input sequence)
  - Multinomial loss for the likelihood of the observed count distribution over the sequence, compared to the predicted probabilities



Stranded profile shape + signal amplitude

$$Loss = -\log p_{mult.}(\mathbf{k}^{obs} \mid \mathbf{p}^{pred}, n^{obs}) + \lambda(\log(1+n^{obs}) - \log(1+n^{pred}))^2$$

k^{obs}: vector of observed reads counts at each position
p^{pred}: learned multinomial prob. at each position
n^{obs}: total number of read counts across entire 1 kb

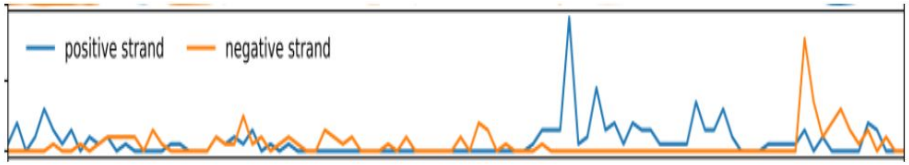
MSE loss

Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

Slide Credit: Anshul Kundaje

BPNet: Profile regression loss

- Two-part loss function for optimizing prediction of the binding profile across the input sequence
  - MSE loss for log (total number of counts across the entire 1kb input sequence)
  - Multinomial loss for the likelihood of the observed count distribution over the sequence, compared to the predicted probabilities



Stranded profile shape + signal amplitude

$$Loss = -\log p_{mult.}(\mathbf{k}^{obs} \mid \mathbf{p}^{pred}, n^{obs}) + \lambda(\log(1+n^{obs}) - \log(1+n^{pred}))^2$$

k^{obs}: vector of observed reads counts at each position
p^{pred}: learned multinomial prob. at each position
n^{obs}: total number of read counts across entire 1 kb

Multinomial loss

Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

Slide Credit: Anshul Kundaje

Multinomial loss component

$$Loss = -\log p_{mult.}(\mathbf{k}^{obs} \mid \mathbf{p}^{pred}, n^{obs}) + \lambda(\log(1+n^{obs}) - \log(1+n^{pred}))^2$$

k^{obs}: vector of observed reads counts at each position
p^{pred}: learned multinomial prob. at each position
n^{obs}: total number of read counts across entire 1 kb

Multinomial loss

Multinomial probability distribution

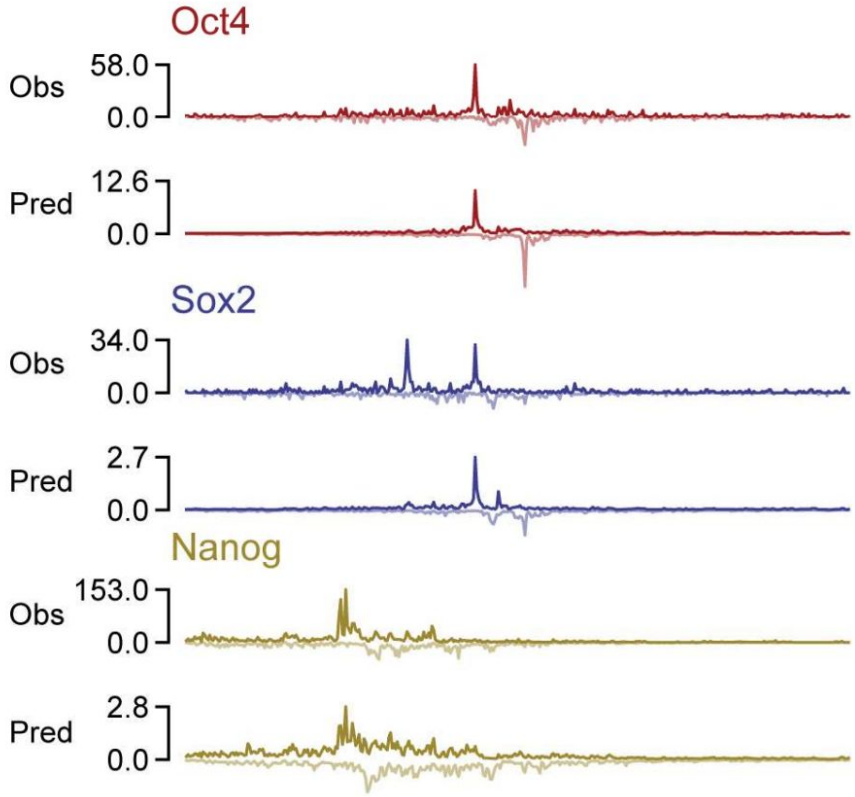
Suppose one does an experiment of extracting n^{obs} balls of 1000 different colors from a bag. Denote as p_i the probability that a given extraction will be in color *i*. Let k_i be the number of balls extracted of color i. The probability of this multinomial distribution is

$$p_{mult}([k_1, k_2 \ldots k_{1000}] \mid [p_1, p_2, \ldots, p_{1000}], n^{obs}) = \frac{n^{obs}!}{k_1!\, k_2! \ldots k_{1000}!} p_1^{k_1} p_2^{k_2} \ldots p_{1000}^{k_{1000}}$$

Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

Slide Credit: Anshul Kundaje

BPNet predicted TF profiles



Avsec et al. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code, 2019.

Slide Credit: Anshul Kundaje

More examples of deep learning in genomics

Epigenomics:

- Predicting methylation states, gene expression from histone modifications, etc.

Transcriptomics:

- Predicting phenotypes from transcriptome, identifying genes associated with transcriptomic data, etc.

Proteomics:

- Predicting secondary structure of proteins, protein-protein interactions, etc.

Summary

Today we covered:

- Biology basics for genomics
- Epigenomics, transcriptomics, proteomics
- Genomics data
- Examples of deep learning for genomics