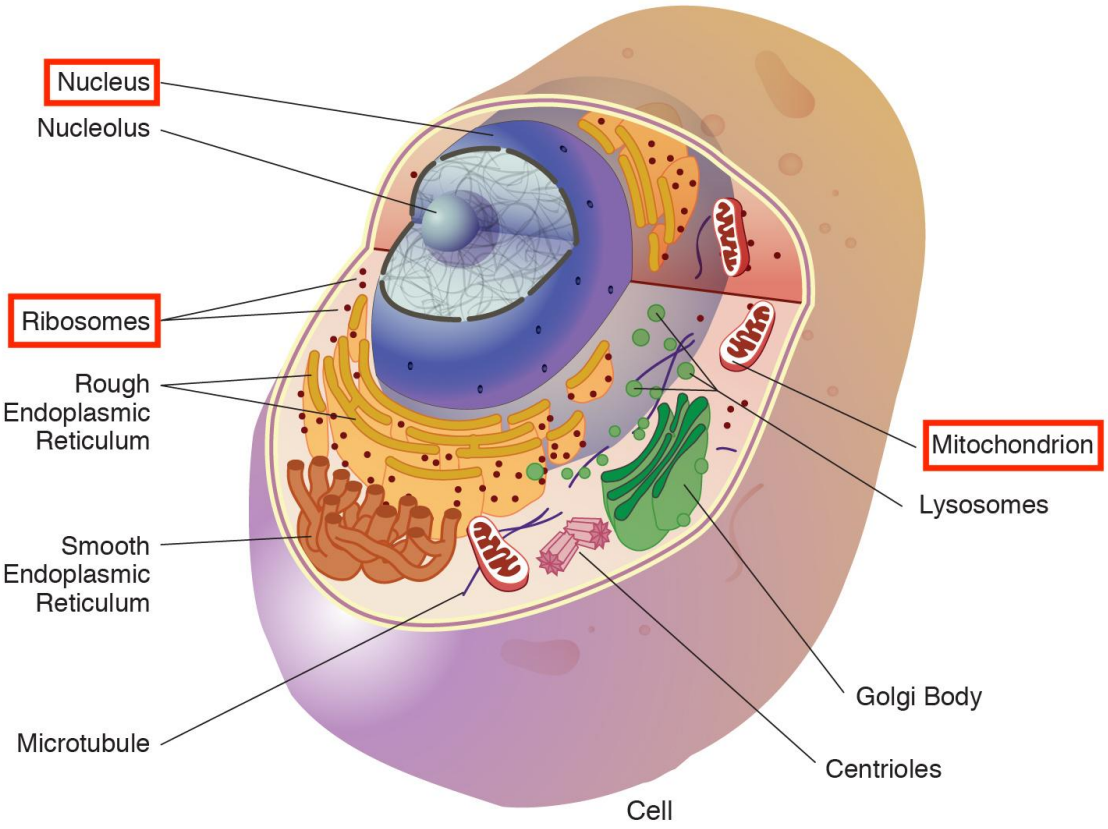
Key Questions to Consider

- How is information about the function of our cells stored in DNA?
- How do variations in DNA affect our health?
- How are DNA variants measured?
- Once we have identified variants, what evidence is used to analyze the effect of a variant?

Genetics 101

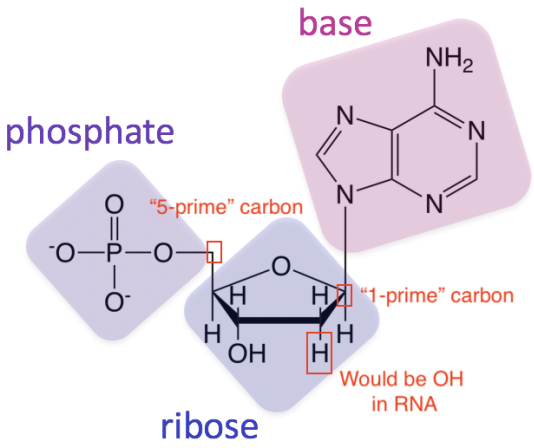
The Cell

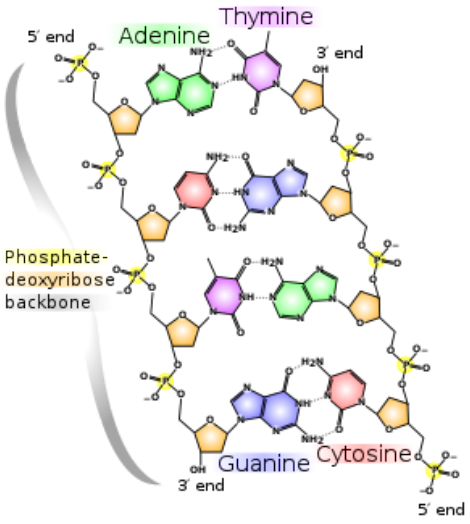
- Most of the genome lives in the **nucleus** (nuclear DNA/nuclear genome)
- A small number of genes live in the **mitochondria** (mtDNA, mitochondrial genome, etc)
- Proteins are initially assembled by **ribosomes**



Nucleic Acids

- Deoxyribonucleic acid (DNA) and RNA (differ by one oxygen atom)
- In these organic chemistry diagrams, each black vertex is a carbon atom, and all other atoms are labeled with their chemical symbol.
- On the left, a DNA monomer. On the right, 2 DNA polymers.





DNA

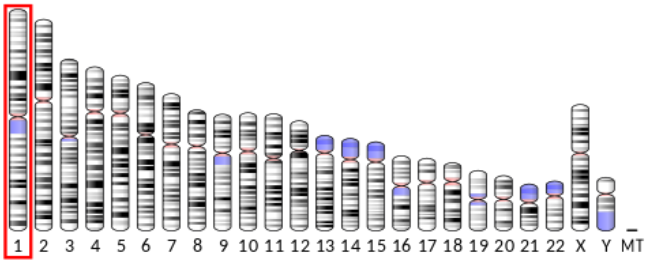
- DNA is a double-stranded polymer of **nucleotides**.
- DNA is directional:
    - At the **5-prime (5')** end, there is a phosphate group sticking out. At the **3-prime (3')** end, there is a hydroxyl group sticking out.
    - By convention, we write single-stranded DNA sequences **starting at the 5' end**
- There are 4 nucleotides, distinguished by their 'nucleobase' (or just **'base'**). Adenine, Cytosine, Guanine, Thymine.
    - A binds with T. C binds with G
- In total, the human genome contains ~3e9 base pairs (bp).

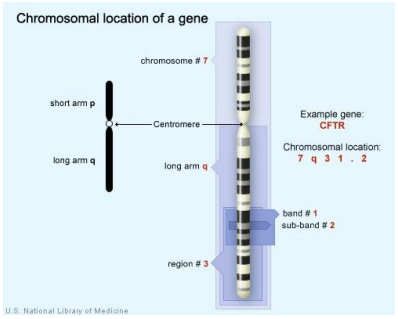
RNA

- Another polymer of nucleic acids.
- Can be structural/functional as well as information storage (e.g. ribosomes which translate messenger RNAs into proteins are themselves mostly built out of RNA)
- **Uracil** replaces **Thymine**

Chromosomes and Inheritance

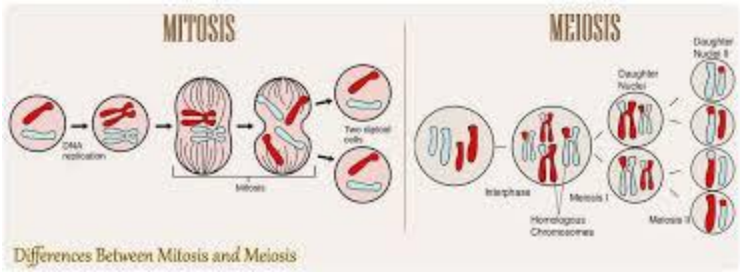
- DNA in humans (like most organisms with a large genome) is arranged into several contiguous pieces. These pieces spend most of their time wound tightly around **histone** proteins. These complexes of DNA and protein are called **chromosomes.**
- Healthy humans inherit **23** chromosomes from mom and 23 from dad. **(23+23 = 46 total chromosomes)**
- Each set of 23 consists of:
    - **Autosomal chromosomes** 1-22, **(autosomes)**. These are numbered by size, with **1** being largest.
    - A sex chromosome (X, or Y). **(sex chromosomes).** Females inherit an X chromosome from dad (genotype XX), males inherit a Y (genotype XY).
- Mitochondria are inherited from mom.
    - Mitochondria have their own tiny genome (16 kbp), and this includes ~37 genes important for mito function.
    - There are also "**nuclear mitochondrial genes**" (which affect mitochondrial function, but live on chromosomes in the nucleus). These encode the rest of the >1000 proteins required for mito function.
- Almost all cells in the body have a copy of all chromosomes, even though a particular cell type will use only a very small subset.
    - We will mention mechanisms for inactivating parts of chromosomes.
    - Red blood cells shed their nucleus and don't carry genetic material
- Near the middle of each chromosome is a **centromere**, an attachment point when the cell wants to drag the chromosomes around during division.
    - Relative to this centromere, the short arm of the chromosome is called **p** and the long arm is called **q.**
- Giemsa staining (a standard protocol) is used to look for large-scale abnormalities. This staining gives a characteristic **banded** appearance (some regions predictably absorb more stain and become darker).
    - This banding pattern can be used to describe the location of genes, for example CFTR (the gene whose mutation results in Cystic Fibrosis) is located on 7q31.2 (https://ghr.nlm.nih.gov/gene/CFTR#location)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT


Chromosomal location of a gene
chromosome # 7
short arm p
Centromere
long arm q
long arm q
band # 1
sub-band # 2
region # 3
Example gene:
CFTR
Chromosomal location:
7 q 3 1 . 2
U.S. National Library of Medicine

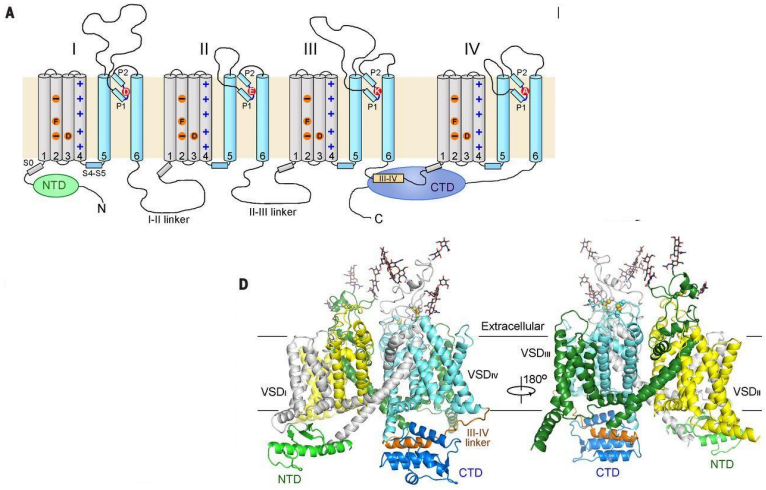
Germline and Genetic Diversity

- Change happens! (Mechanisms of DNA damage, repair and mutagenesis)
    - DNA replication is not perfect
    - Environmental processes can also cause changes in DNA sequence (e.g. UV radiation on your skin)
    - Some parts of the genome are designed to be mutated and recombined, e.g. T-cell receptor and B-cell receptor (to produce diverse and adaptive immune protection)
    - DNA repair mechanisms are constantly at work to fix small mistakes
- **Evolution** takes place (very slowly) when mutations occur naturally and happen to confer slightly improved fitness to the host organism
- Some regions of the genome are **highly conserved** between the modern human genome and our evolutionary ancestors - from this we can infer that changes in these regions are not compatible with healthy life.
- Short-term variability is also desirable.
    - The production of sex cells (aka **germ/germline** cells) is called **meiosis,** and includes a stage of **crossing-over** to produces diversity in the next generation by mix-and-match from existing sequences.
    - Contrast this with **mitosis**, whose goal is to produce 2 accurate copies from 1 parent cell when growing/replenishing an organ.



Proteins

- Proteins perform most of the functions of our cells; destroying toxins, metabolizing nutrients, catalyzing chemical reactions (**enzymes**)
- Proteins (aka **polypeptides**) are built from intricately folded polymers of **amino acids**.
    - The linear sequence of amino acids gives a protein its **primary structure.**
    - These amino acid chains form into simple motifs (alpha-helices and beta-pleated sheets), giving **secondary structure**.
    - These simple shapes assemble into an overall 3D **tertiary structure.**
    - Multiple subunits can be linked together, giving **quaternary structure.**
- Like with nucleic acids, amino acid polymers have a directionality.
    - The **5'** end of the DNA coding strand corresponds to the **N** (amino) terminus.
    - The **3'** end corresponds to the **C** (carboxy) terminus
- For example, consider the voltage-gated sodium channel (a vital protein in neurons):



- In humans, there are 20 main amino acids, each described in our DNA and RNA by 3-base sequences called **codons**
    - The variable side chain of the amino acid gives different properties
    - There is also a special **start** codon, and several **stop** codons, that help control transcription

RNA codon table [edit]

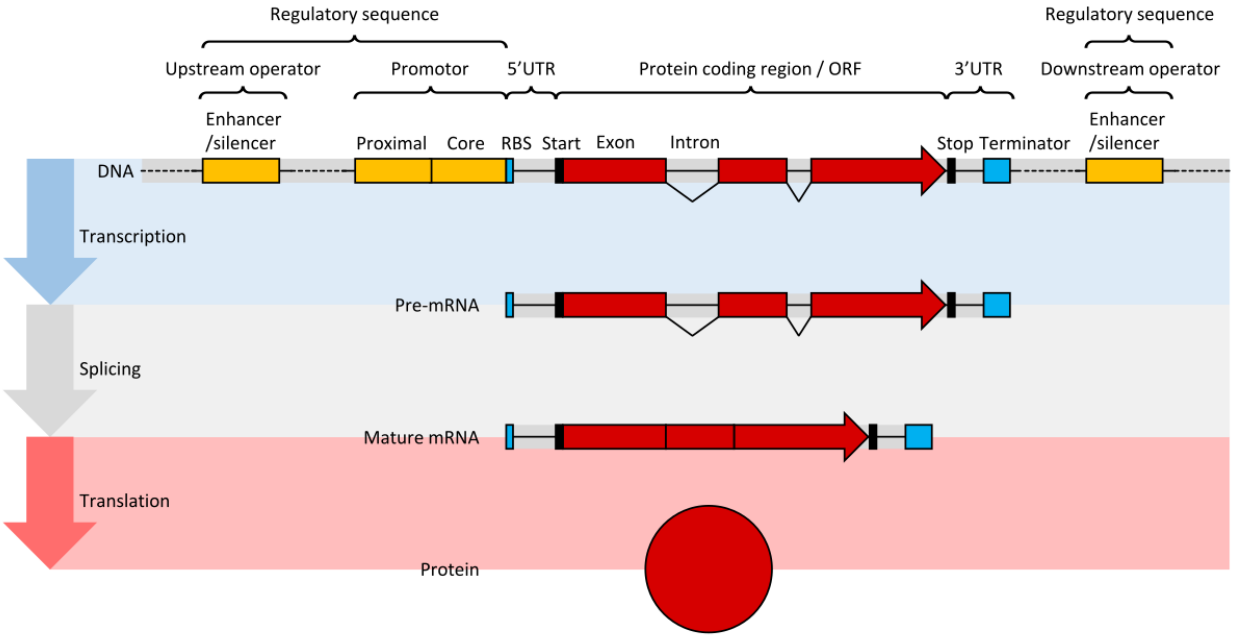
Amino-acid biochemical properties	Nonpolar	Polar	Basic	Acidic

Termination: stop codon

Standard genetic code

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (*Ochre*)[B]	UGA	Stop (*Opal*)[B]	A
	UUG[A]		UCG		UAG	Stop (*Amber*)[B]	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG[A]		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

The Central Dogma (DNA → RNA → Proteins)



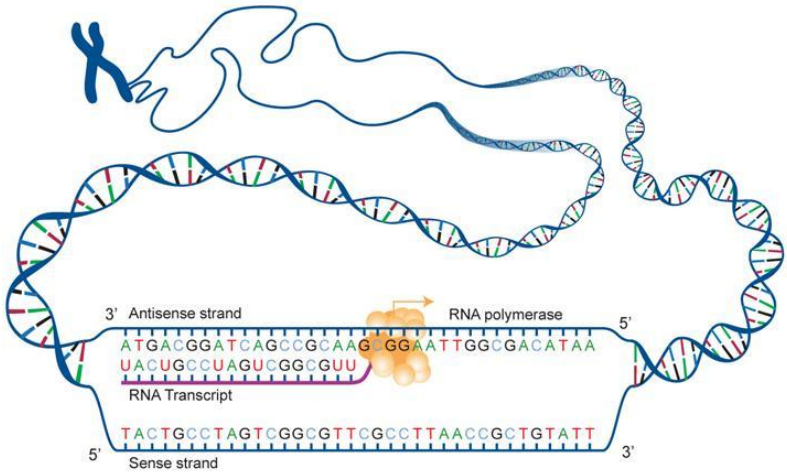
Genes

- The "central dogma" of biology says: DNA is **transcribed** into RNA, which is **spliced** to remove non-coding regions, and then **translated** into proteins.
- Each protein subunit is encoded by a single gene.
- Genes have a typical structure:
    - Upstream regulatory sequences, e.g. **transcription factor binding sites**
    - Promoter
    - 5' untranslated region (**UTR**)
    - An interspersed collection of coding regions (**exons**), and non-coding regions (**introns**)
    - 3' untranslated region (**UTR**)
    - Downstream regulatory sequences
- **"Coding"** refers to parts of the sequence that directly describe an amino acid. **"Non-coding"** does NOT mean unimportant!
- The collection of all exons is called the **exome.** This is about 1% of the total size of the **genome**

Transcription

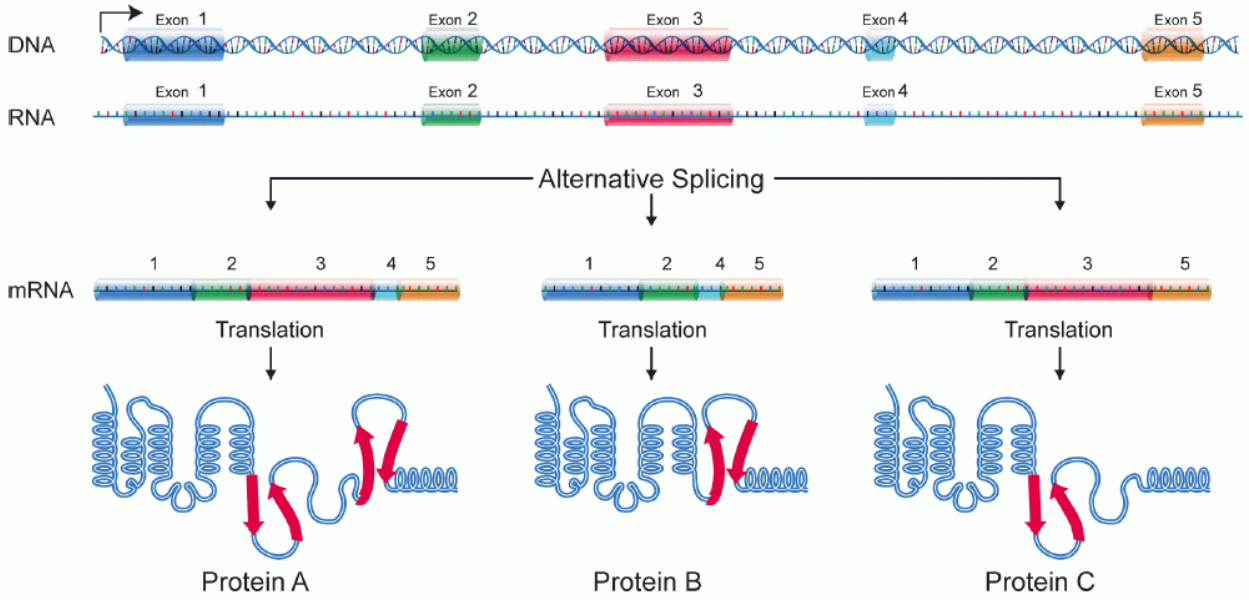
- Regulatory mechanisms (see Epigenetics) cause a certain gene to be accessible

- Transcription factors recruit the machinery for transcription (especially the **RNA polymerase)** to the beginning of a gene.
- The RNA polymerase copies the antisense strand of the DNA to produce a **precursor messenger RNA (mRNA)**



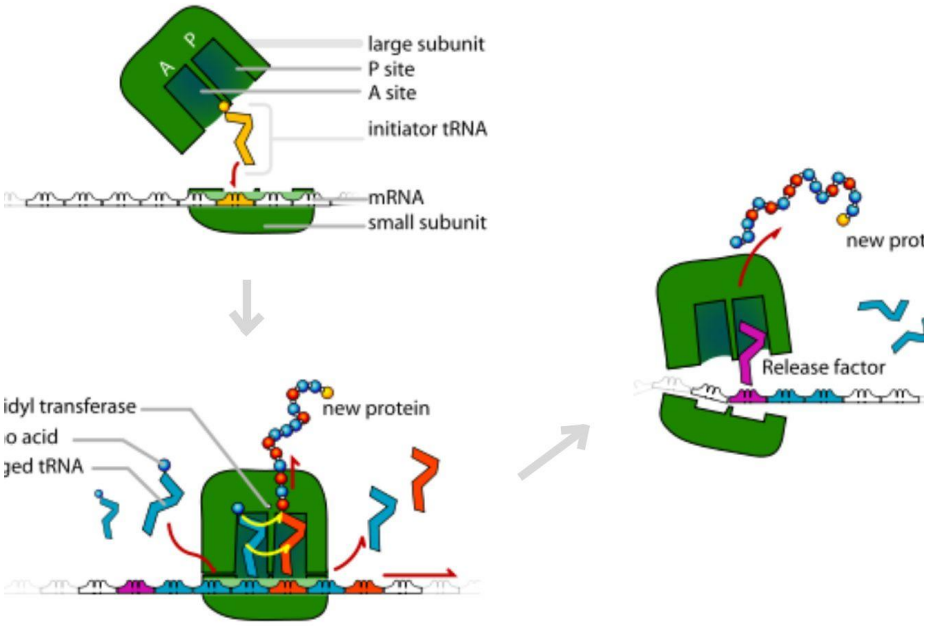
Splicing

- The precursor mRNA is processed into a **mature mRNA** by the addition of a 5' cap and a 3' tail, as well as by the selective **removal of introns**
- By choosing which introns to remove, a single gene can be processed into several **splice isoforms**



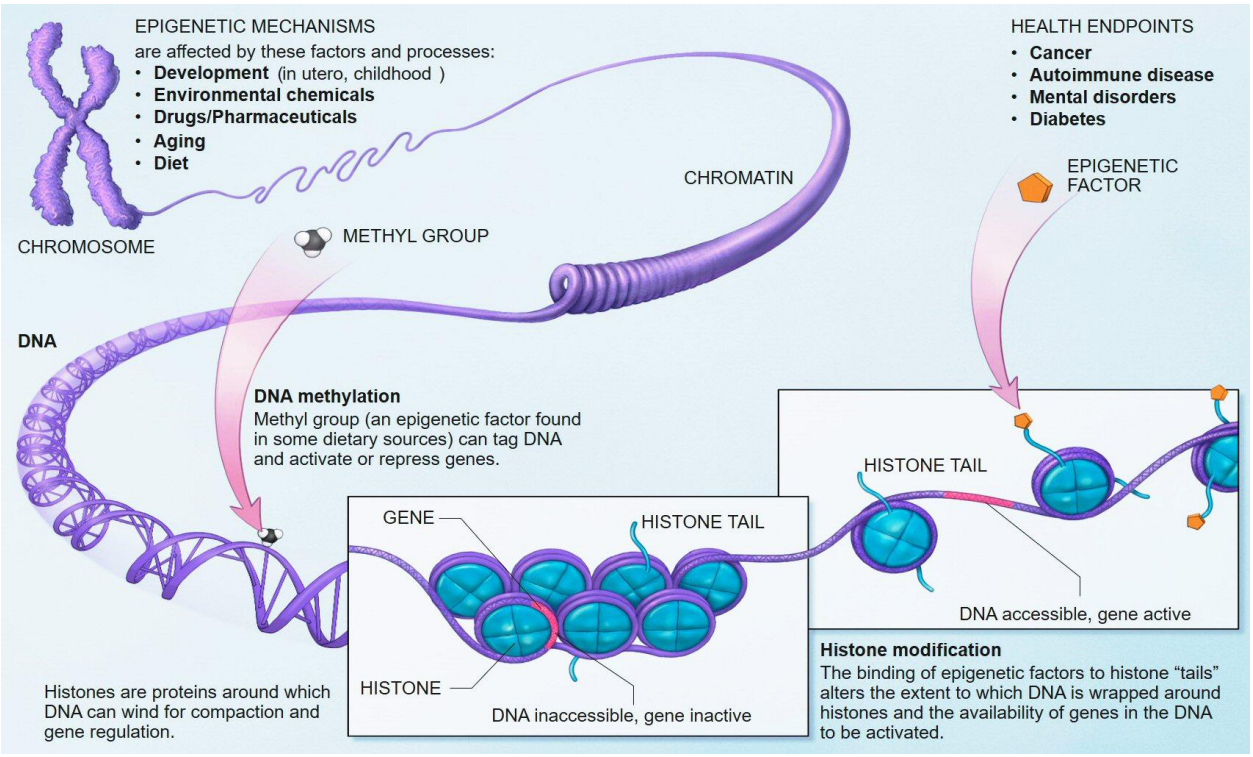
Translation

- The processed mRNA is brought to a **ribosome**
- The ribosome scans the mRNA, matching a single codon at a time, and assembling a linear polypeptide



Epigenetics

- If all cells carry the same genome, why don't they all behave the same way? Epigenetics!
- This broad set of regulatory mechanisms is less well understood, but includes
    - Histone acetylation, which helps open/close chromatin to select which genes are accessible for transcription (can be measured by **ChIP-seq**)
    - DNA methylation, (can be measured by **bisulfite sequencing**)
- We can identify **enhancers** and other regulatory sequences by catching two regions of DNA in close contact. This 3D structure can be measured by techniques like **chromosome conformation capture**



Mutations

Single Nucleotide Variants (SNV)

- SNV are sometimes called Single Nucleotide Polymorphisms (SNPs)
- Recall that some amino acids are encoded by multiple codons
- Changing a single base can result in:
    - The same amino acid (**synonymous** change)
    - A different amino acid (**missense** change)
    - A premature stop codon (**nonsense** change)
- Notice that there is a finite set of possible SNV: **(3) * (3e9) ~= 1e10**
    - Many papers that attempt to predict the impact of SNVs will provide a pre-calculated table for all possible SNV

Insertion / Deletion (indel)

- If we gain or lose a multiple of 3 bases, the **reading frame** of the overall sequence is preserved, albeit with a few extra or few missing amino acids.
    - This may still be severe, or may introduce a premature stop codon (**nonsense** change)
- If the number of gained/lost bases is NOT a multiple of 3, we have a **frameshift** mutation. Everything upstream is preserved, and everything downstream is changed. This is typically severe.

Bioinformatics 101

When looking for disease-related genetic variation, our typical goal is:

1. Identify a person, tissue, or cell that is affected by the disease
2. Measure the DNA (RNA, proteins, or metabolites present (genomics, transcriptomics, proteomics, metabolomics))
3. Identify changes relative to a healthy control
4. Try to understand how these changes cause the disease, and how they can be repaired or compensated

Here, we will describe genomics only, though all of these domains have analogous concepts.

Reference Genomes

- By comparing DNA sequences to a reference that we consider “normal”, we can decide when a mutation is “abnormal”
- Wikipedia description of the reference genome:

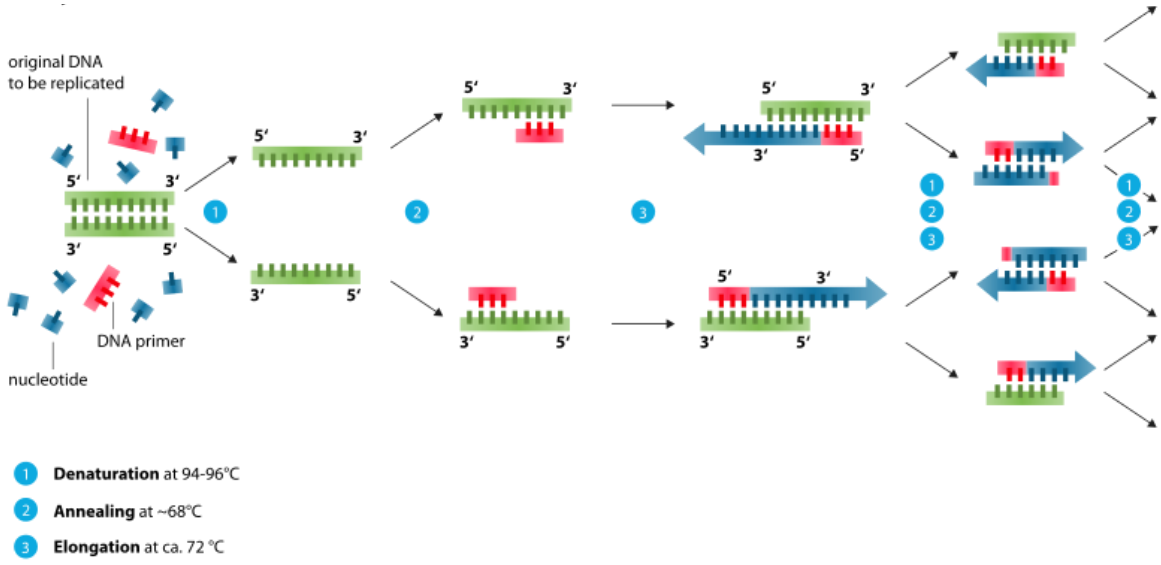
    - *The human reference genome is derived from **thirteen anonymous volunteers** from Buffalo, New York. Donors were recruited by advertisement in The Buffalo News, on Sunday, March 23, 1997.*
- As time passes, these samples are re-processed using the latest chemical and computational methods to produce better, more complete "**builds**"
- Latest: **GRCh38,** aka **hg38** (Genome Reference Consortium, human build 38)

Illumina Sequencing

NOTE - this is just one sequencing technology, but it's the current market leader.

Prerequisites: PCR

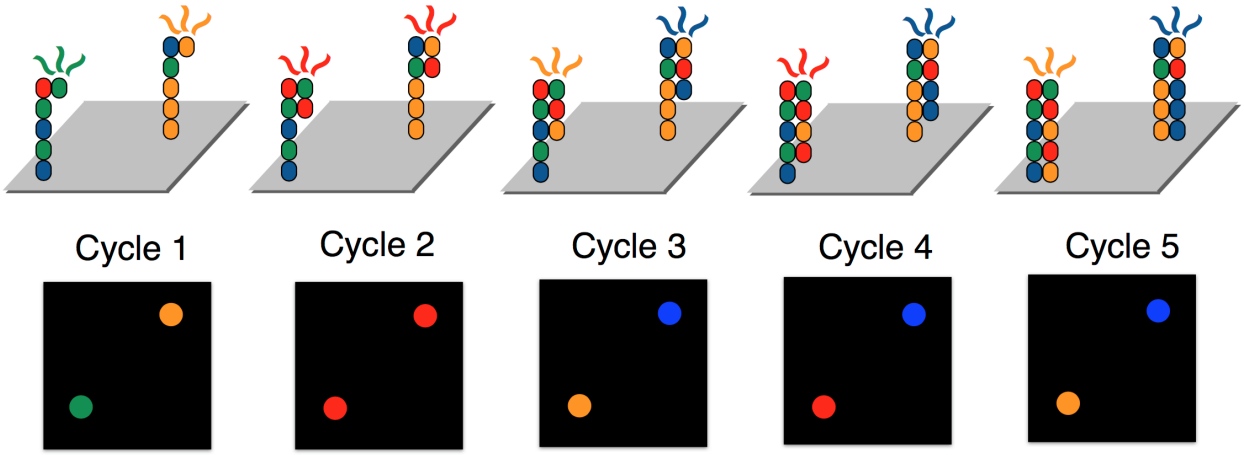
- Polymerase chain reaction (PCR) is a standard method for repeatedly duplicating a small sample of DNA to produce many identical copies



Prerequisites: Chain-terminating Sequencing-by-Synthesis

- Illumina uses "chain-terminating" nucleotides for their "sequencing-by-synthesis" procedure
    - We have a template single strand of DNA that we want to read
    - We have a collection of modified nucleotides
        - Each flavor is bound a unique color of fluorescent molecule (suppose A-Red, C-Green, T-Blue, G-Yellow)
        - They are "blocked" so that only a single nucleotide can be added at a time
    - We apply the special blocked nucleotides - only a single one type will match our template (e.g. A-Red)

- We illuminate the strand and record the color (e.g. Red) - this tells us which nucleotide got added (e.g. A), and therefore tells us what that position on our template was (e.g. T)
- We can cut off the colored fluorescent molecule, remove the “blocking” mechanism, and repeat for another cycle

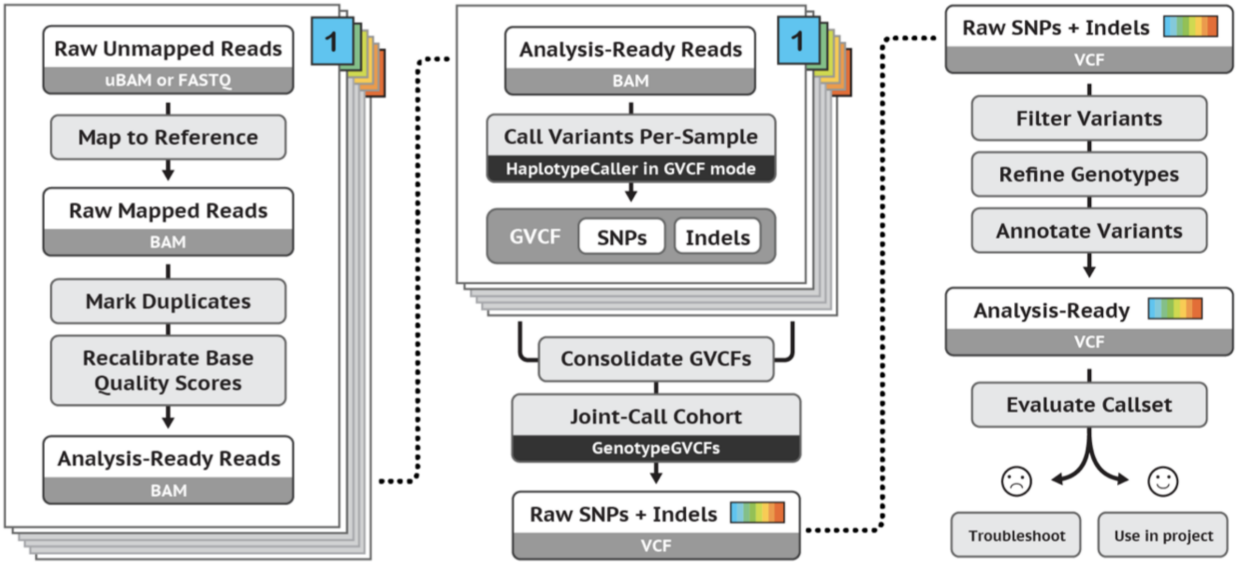


Illumina Process (https://www.youtube.com/watch?v=fCd6B5HRaZ8)

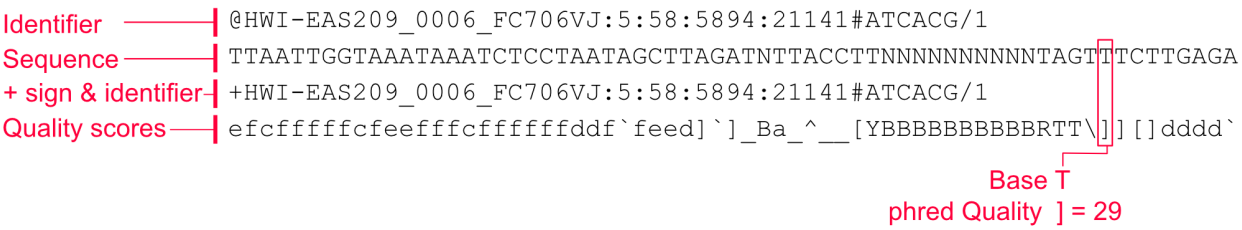
1. Collect sample (blood, saliva, tissue sample, etc) and extract nucleic acids
2. Amplify sample by PCR
3. Prepare “sample library”
    a. Randomly fragment, e.g. by ultrasound
    b. Tag all fragments from a single sample with a unique barcode sequence (so that later we can combine multiple samples to be processed together)
    c. Add sequencing primers (used during the actual sequencing-by-synthesis)
    d. Add a 5’ and a 3’ adapters (for bridge amplification later)
4. Bridge amplification
    a. The sequencing chip is pre-made with a sparse “lawn” of adapters. These match the adapters we added to our fragments
    b. We wash our sample on, and amplify each fragment to produce small isolated colonies of identical strands
    c. This is necessary because the brightness from a single fluorescent molecule is hard to detect; we need to measure several at once.
5. Sequencing by synthesis
6. Base scoring and calling (aka **primary analysis**)

Primary, Secondary, and Tertiary Analysis

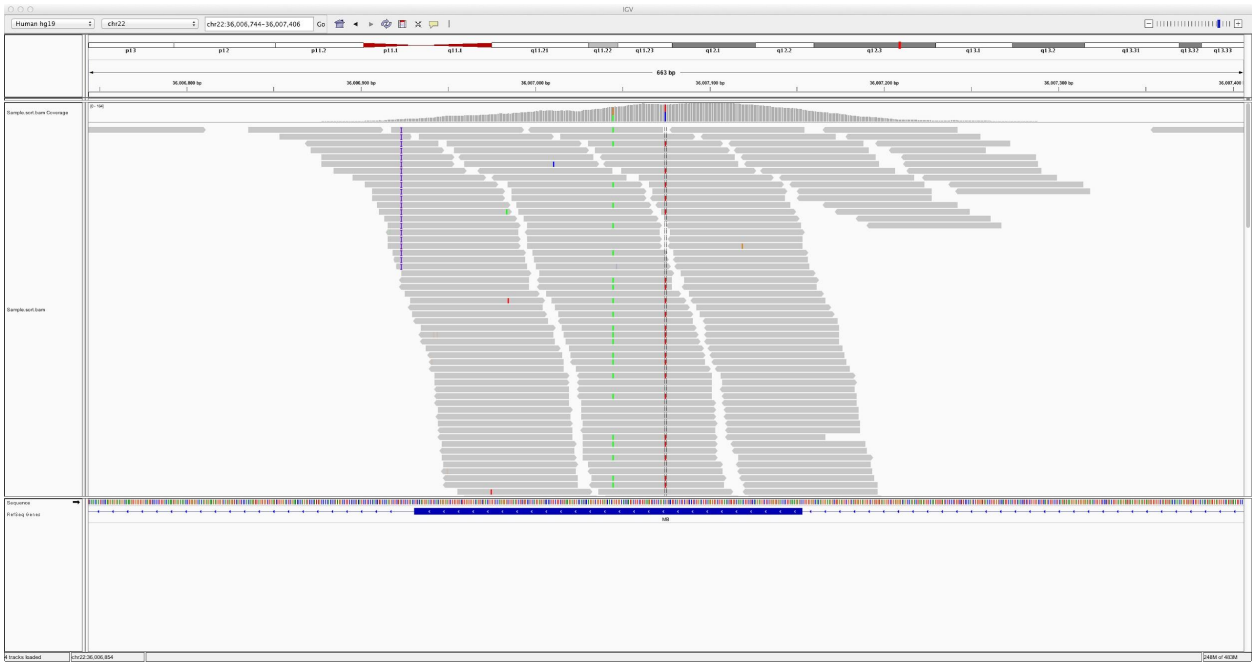
- See Best Practices Workflows – GATK
    - For example, for germline (inherited) SNV and small indel detection:



- Recall that **Primary analysis** was done on the sequencer to produce raw reads with a quality score for each base
    - Quality scores are given using the “phred” scale: q = -10 log_10 e, where e is the probability of error. E.g. 10 means 90% accuracy, 20 means 99% accuracy, etc.
    - **Raw read data** received from a sequencer is typically stored in **FASTQ** format



- **Secondary analysis** consists of:
    - **Aligning reads** to infer the contiguous sequence of the original DNA sample, and variant calling (identifying changes relative to a reference sequence). Aligned reads are typically stored in **SAM/BAM** format (BAM is just the binary version of a Sequence Alignment/Map format - https://samtools.github.io/hts-specs/SAMv1.pdf).
        - In the case of new sequences (e.g. bacterial sequencing) *de novo* alignment can be used, where we only rely on substring matching to each other to reconstruct the linear sequence
        - The **depth of coverage** achieved during sequencing refers to the average number of reads that cover each nucleotide



- **Calling variants**
  - Based on the reads covering a position, and the quality of bases in each read, we can determine whether we believe there was a change relative to the reference sequence
  - Variants are typically stored in a **VCF** (Variant Call Format) file (https://en.wikipedia.org/wiki/Variant_Call_Format)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

- **Tertiary Analysis** consists of assembling evidence for each variant to determine which variants may be disease-related, and which ones may be related to a particular condition of interest

Evidence of Disease-Relevance

See, for example, the ACMG Standards and Guidelines for the Interpretation of Sequence Variants

- Databases describing diseases and disease-causing variants in genes
    - ClinVar
    - HGMD (**http://www.hgmd.cf.ac.uk/ac/index.php - also has a nice table describing different variant types)**
    - OMIM
- Rates of mutation at different positions of the genome in healthy individuals
    - gnomAD collects “minor allele frequencies” (the rate of mutation) across the genome
- Known gene structure (locations of exons, etc)
    - Ensembl’s Variant Effect Predictor
- Algorithms for predicting pathogenicity (known regulatory and structural motifs, sequence context, evolutionary conservation, etc)
    - SIFT
    - PolyPhen
    - CADD
    - DANN
    - Many, many others
    - Key pitfall: some models just collect/are trained on the predictions of other models
- Protein modeling
    - We can try to predict the effect of the variant on the 3D structure of the protein. Note that protein folding is HARD!
- Using gene expression data to link the gene of interest to the tissue of interest
    - GTEx
- Experimental models
    - Using techniques like CRISPR, it is fast becoming feasible to create cell-line or animal models with targeted mutations, to directly measure the effect of a specific variant on a control genome.
- Comparing phenotype and genotype inheritance patterns
    - If a variant is truly disease-causing, then all family members or subjects in a study who carry the variant should also exhibit the disease. (In the simplified model of a monogenic, fully penetrant disease…)

Related Reading

Deep Learning in Variant Analysis

Here are a few papers whose abstract looks interesting

- “A universal SNP and small-indel variant caller using deep neural networks” - https://www.nature.com/articles/nbt.4235.pdf
- “A multi-task convolutional deep neural network for variant calling in single molecule sequencing” - https://www.nature.com/articles/s41467-019-09025-z
- “Deep Learning for Genomics: A Concise Overview” - https://arxiv.org/pdf/1802.00810.pdf
- “ Deep learning in bioinformatics” - https://academic.oup.com/bib/article/18/5/851/2562808

Bioinformatics Overview

- “Hitchhiker’s Guide to NGS” https://www.goldenhelix.com/media/pdfs/whitepapers/Hitchhikers-Guide-to-NGS.pdf
- Illumina Sequencing https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf
- Niklas SM - Master’s thesis https://dash.harvard.edu/bitstream/handle/1/33789915/SMEDEMARK-MARGULIES-MASTEROFMEDICALSCIENCESTHESIS-2016.pdf