BST227
Introduction to Statistical Genetics

Lecture 1:
Introduction and Overview of Genetic Disease

http://aryeelab.org/BST227

BST227 Housekeeping Details

Instructor:

Martin Aryee
aryee.martin@mgh.harvard.edu
MGH, Room 6016, 149 13th Street, Charlestown
http://aryee.mgh.harvard.edu/

Office Hours: Monday, 12:30PM - 1:30PM, Building 2/419

Teaching Assistant:

Caleb Lareau
caleblareau@g.harvard.edu

Office Hours: Monday, 6PM-7PM, FXB G11

Section: Wednesday, 5:30PM-7:30PM, FXB G13

Course Organization

Basic statistical prerequisites: Bayes rule, hypothesis testing, confidence intervals, estimation, chi-square tests for independence & goodness of fit, linear regression, logistic regression.

Course Materials: http://aryeelab.org/BST227

Required Readings:
The Fundamentals of Modern Statistical Genetics
by Nan Laird and Christoph Lange
(available from the COOP or online stores)

Reference books:
- Statistics in Human Genetics by Pak Sham
- Statistical Methods for Genetic Epidemiology by Duncan Thomas

Course Organization

Outcome Measures:

- Homework: 60%
- Project: 30%
- Class and Section Participation: 10%

Homework:
Homework assignments are due Monday via the Canvas online dropbox by end of day (midnight).

Students should feel free to discuss approaches to solving the problems in working through homework problems, but each student must turn in their own solution, written entirely in his or her own words and not copied from another source. In addition, you should not share your final solutions with another student.

Project:
The class will be divided into teams, each of which will complete a data analysis of GWAS project, write a report and present their results during the last week of class.

Lecture overview

23-Oct	Background
25-Oct	Mendel's Laws and genetic models for disease
30-Oct	Hardy Weinberg Equilibrium and Recurrence risk Ratios
1-Nov	An overview of linkage and association
6-Nov	Introduction to GWAS
8-Nov	Population Substructure
13-Nov	Analysis of rare variants and non-SNP variation
15-Nov	Variant calling from high-throughput sequencing data
20-Nov	Estimating Heritability from Family Data and from Genome Data

Identifying disease variants

Thanksgiving

27-Nov	Using the 3-dimensional organization of DNA to interpret variants
29-Nov	Epigenetic enrichment analysis
4-Dec	Genetic association studies in cancer
6-Dec	Manipulating DNA to validate GWAS hits
11-Dec	Epigenome-wide Association Studies (EWAS)
13-Dec	Project presentations

Interpreting disease variants

Today: Review of basic concepts

- Review of genetics & molecular genetics
- Types of genetic variants and data
- Types of genetic disease

- What is the human genome?
- How is it organized?

Human Genome

22 autosomes, 2 sex chromosomes

Diploid cells:

2 copies of 22 autosomes

1 pair of sex chromosomes (XX or XY)

= 46 chromosomes in total

Telomeres:

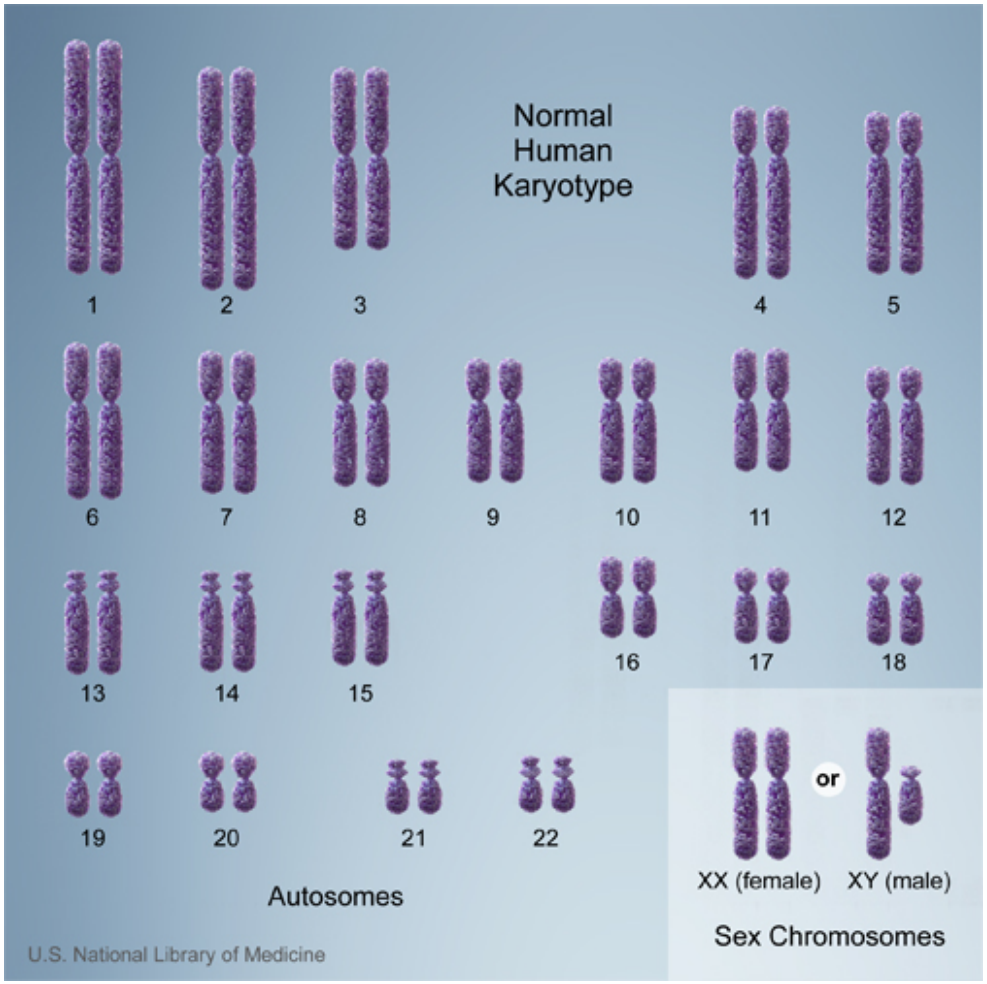
Ends of the chromosomes

Centromere:

the two copies join here during replication

Two arms:

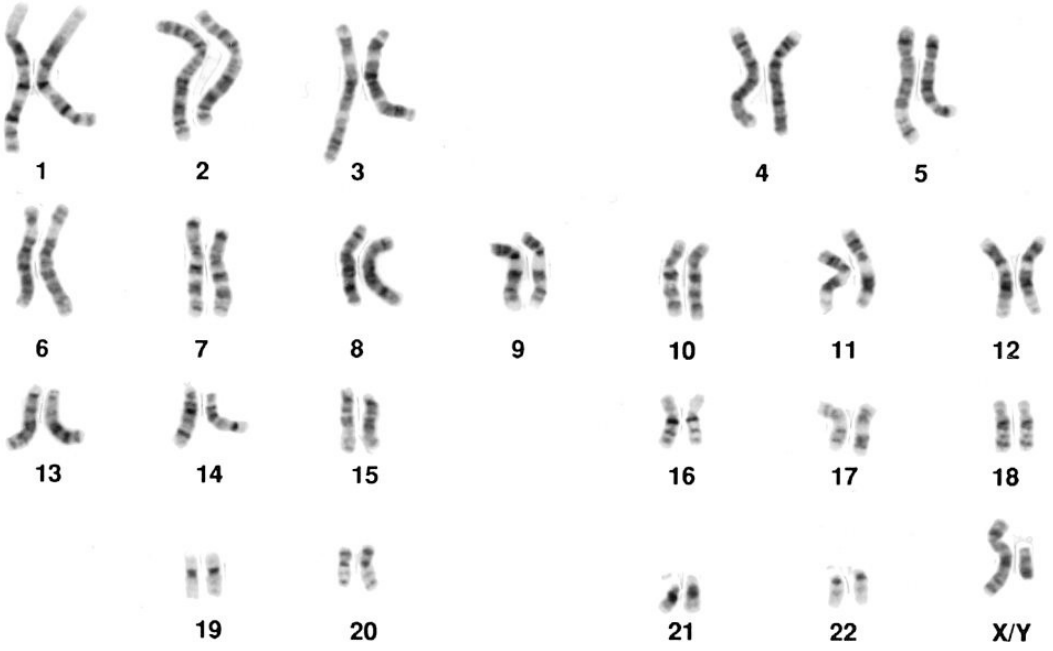
p - short (short, ‘petit’) and q (long)



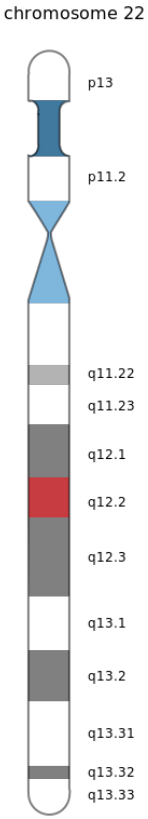
Human Genome

Locus:
Particular location in the genome

Band: segment within each arm



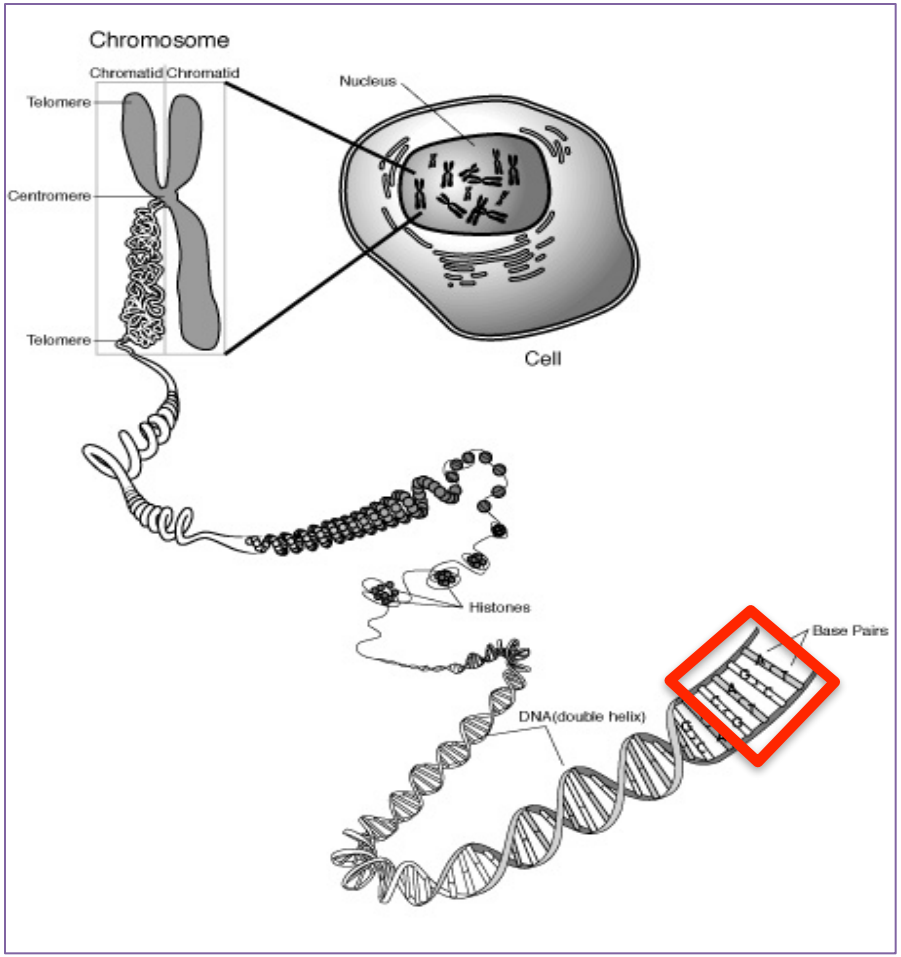
Normal Karyotype



What are chromosomes made of?

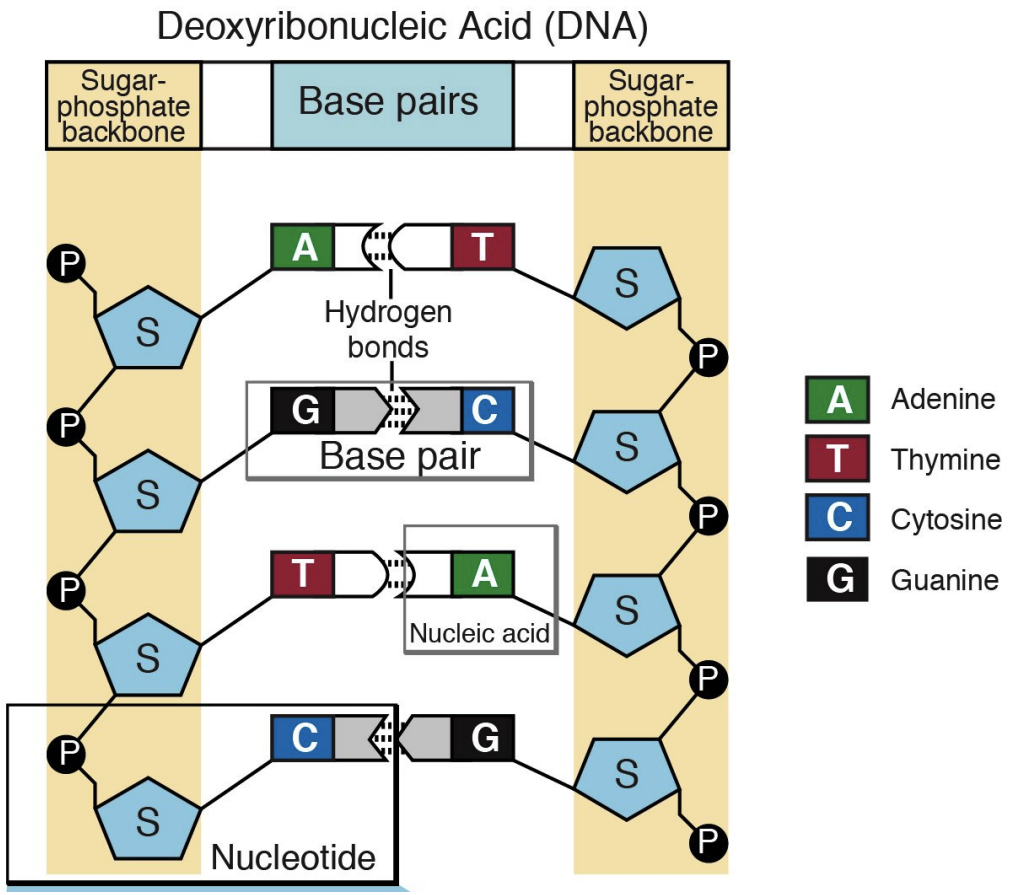
Each chromosome is made up of two strands of deoxyribonucleic acid (DNA) in a double helix arrangement

2 meters of DNA per cell!



Deoxyribonucleic Acid (DNA)

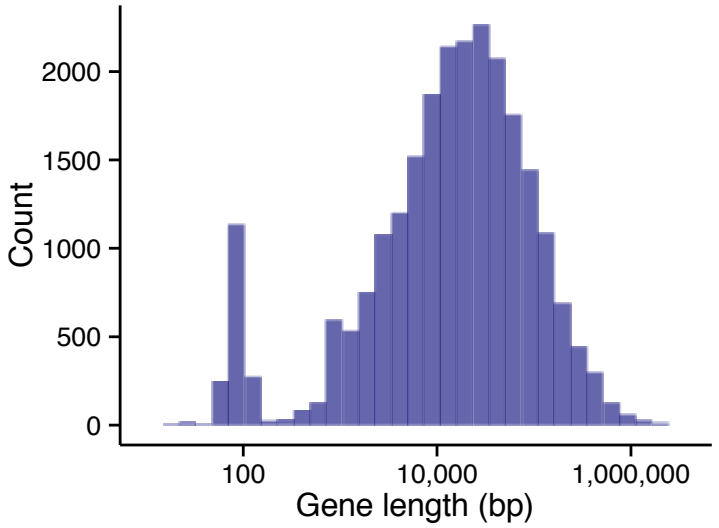
Each ‘ladder rung’ has 2 complementary base pairs



http://knowgenetics.org/nucleotides-and-bases/
NHGRI

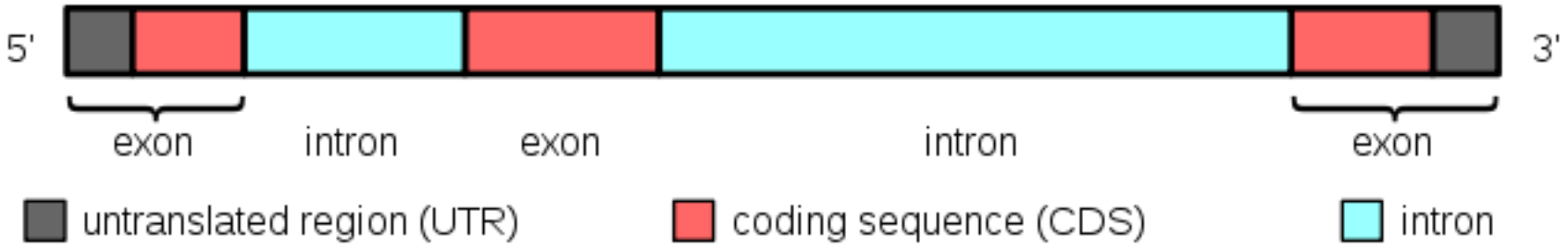
Genome, Genes and Genetic Loci

- The human **genome** is made up of 3 billion base-pairs.
- **Genes** are linear stretches of DNA. They are the main functional units of heredity.
- Genes code for a specific functional molecules (RNAs and proteins).
- The human genome contains about **20,000 protein coding genes** and thousands of non-protein coding genes
- A **genetic locus** can be a gene, a base pair, or any specific region of DNA.



Gene structure

- A gene is divided into alternating sequence blocks called **exons** and **introns.**
- **Exons** are relatively short and often code for amino acids
- On average there are 9 exons and 8 introns per human gene.

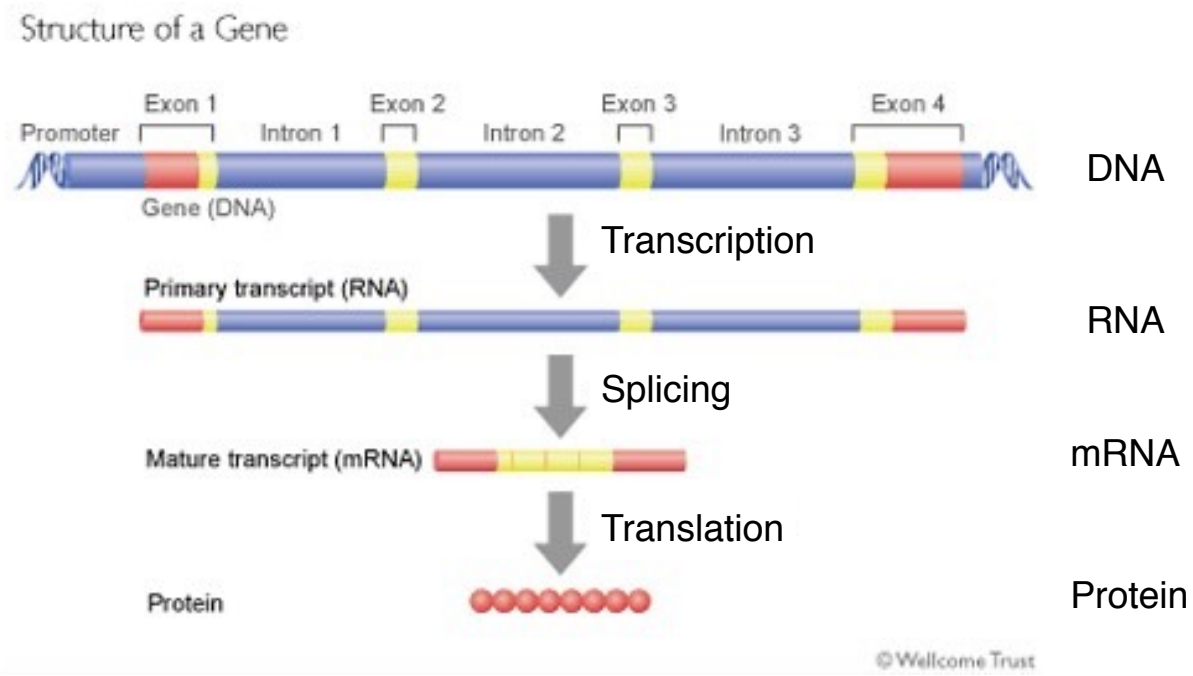


Genes and gene expression

- Genes are **expressed** to produce functional RNA and protein molecules
- The **central dogma of molecular biology** describes the two-step process, **transcription** and **translation**, by which the information in genes flows into proteins:

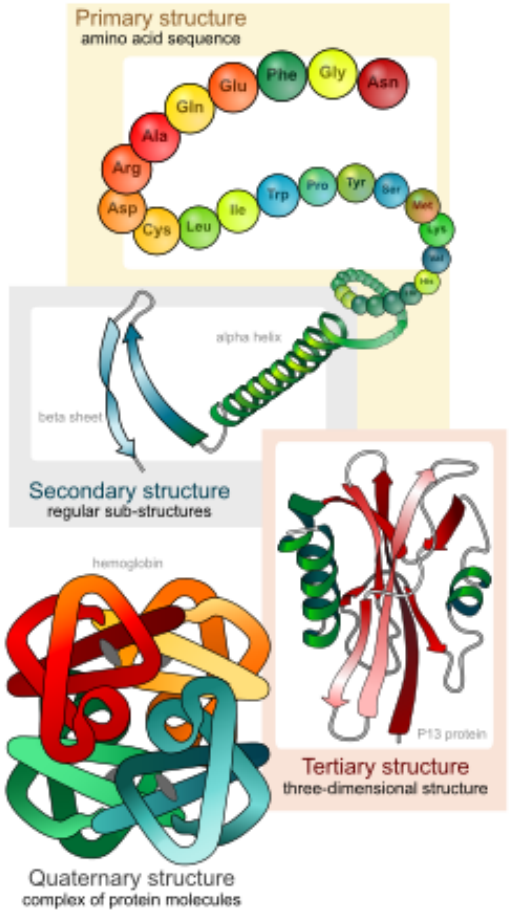
  DNA → RNA → protein

- In the human genome, coding exons make up 1%, introns make up 26%



Amino acids and proteins

Alanine	**Leucine**
Arginine	**Lysine**
Asparagine	**Methionine**
Aspartic acid	**Phenylalanine**
Cysteine	**Proline**
Glutamic acid	**Serine**
Glutamine	**Threonine**
Glycine	**Tryptophan**
Histidine	**Tyrosine**
Isoleucine	**Valine**

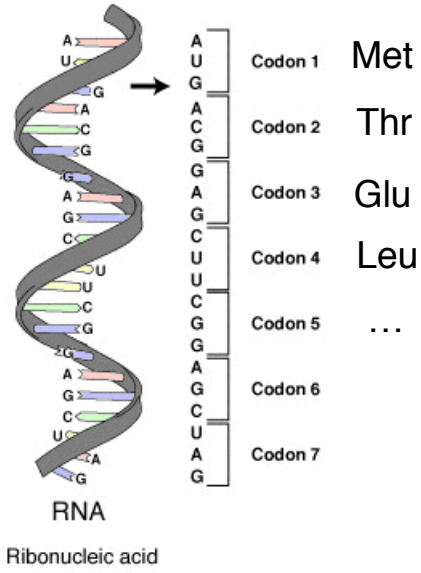


By LadyofHats [Public domain], via Wikimedia Commons

Proteins consist of chains of amino acid residues

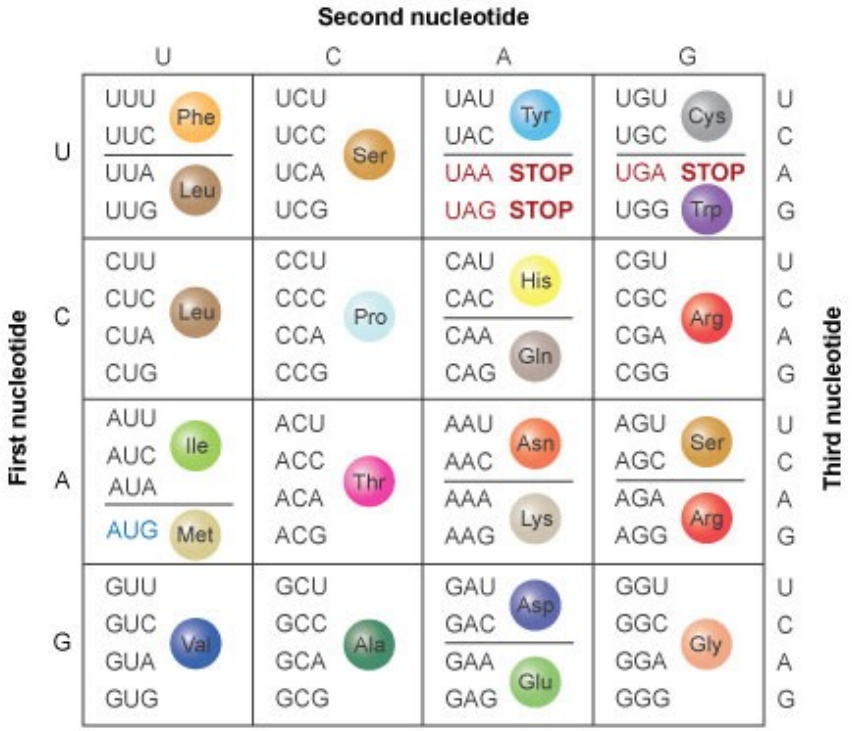
The Genetic Code

- DNA bases: A G C T
- RNA bases: A G C U
- A three-nucleotide **codon** in a nucleic acid sequence specifies a single amino acid



http://en.wikipedia.org/wiki/Genetic_code

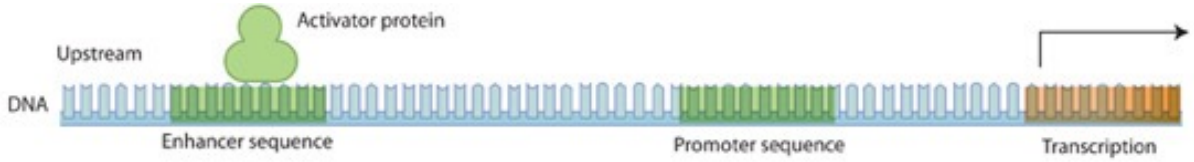
4 by 4 by 4

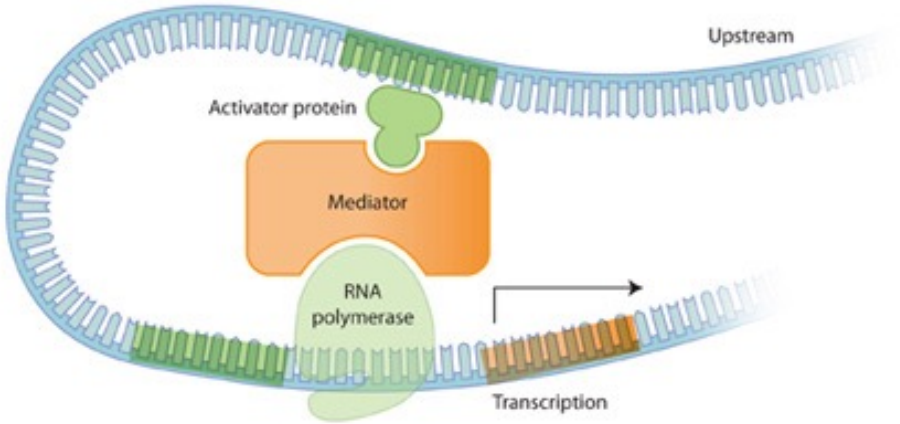


2014 Nature Education

Regulation of gene expression

- A **promoter** is a region of DNA adjacent to a gene that contains binding sites for proteins involved in transcription
- **Enhancers** are other regulatory DNA sequences involved in regulating gene expression. May be far (1Mb+) away from gene.





More on Nov 29

How much of the genome is functional?

80% ?

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

8% ?

OPEN ACCESS Freely available online

PLOS | GENETICS

8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage

Chris M. Rands[1], Stephen Meader[1], Chris P. Ponting[1]*, Gerton Lunter[2]*

1 MRC Functional Genomics Unit, Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford, United Kingdom, **2** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

Example: CCR5 gene

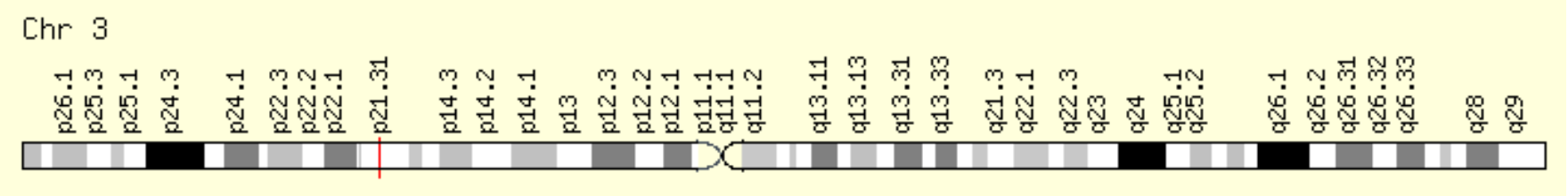
The CCR5 gene provides instructions for making a protein called Chemokine (C-C motif) Receptor 5.

Size: 6,065 bp

Cytogenetics is the study of chromosomal structure, location and function in cells. It includes the study of chromosome number and appearance (karyotyping), the physical location of genes on chromosomes, and chromosomal behaviour in processes such as cell division

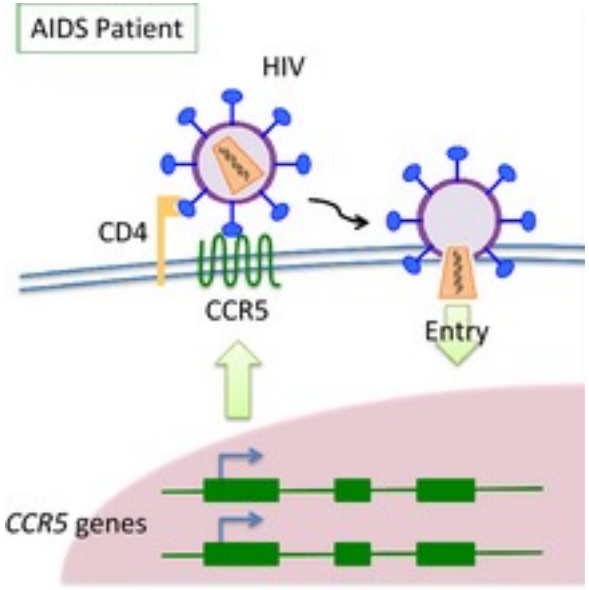
Cytogenetic location:

3p21.31 - short (p) arm of chromosome 3 at position 21.31



Molecular location:

Chromosome 3, base pairs 46,411,633 to 46,417,697



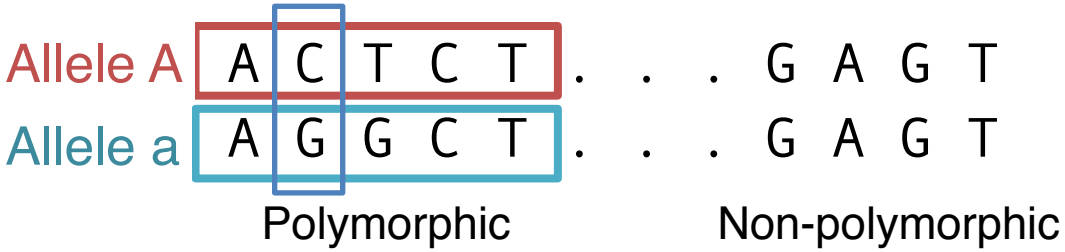
Dev Growth Differ. 2014 Jan;56(1):63-77. doi: 10.1111/dgd.12107. Epub 2013 Dec 11.

Genetic correction using engineered nucleases for gene therapy applications.

Li HL1, Nakano T, Hotta A.

More on Dec 6

Some definitions



- **Alleles or Variants**: Different forms (i.e. different DNA sequences) of the same gene or genetic locus. Many different types of variation.

- **Polymorphic:** Polymorphic loci have several different alleles. At other loci, there is no variation from person to person.

- **Genetic Marker:** Random variable providing information on DNA sequence at a particular locus; many different types of loci and types of information. Markers are inherently categorical. Often labelled A or a, A or B.

- **Disease Susceptibility Locus:** A genetic locus thought to be causal for the disease. Markers are always observable, DSL may not be.

From alleles to genotypes

- Chromosomes come in pairs in diploid organisms (e.g. humans)
- **Genotype:** Pair of alleles at a locus
- **Example:** two possible alleles A and a, what are possible genotypes?
- **Heterozygote:** genotype with different alleles (Aa)
- **Homozygote:** have the same alleles (AA, aa)

How Genetic Variants Arise: Mutations Create Variation

DNA undergoes frequent chemical change, especially when it is being **replicated**. Most of these changes are quickly repaired. Those that are not result in a 'mutation'.

Thus, mutation is a failure of DNA repair.

Mutations give rise to new DNA variants, they may cause disease, they can be protective against disease, or they may be harmless variations that can be used as genetic markers.

Mutations vs. variants

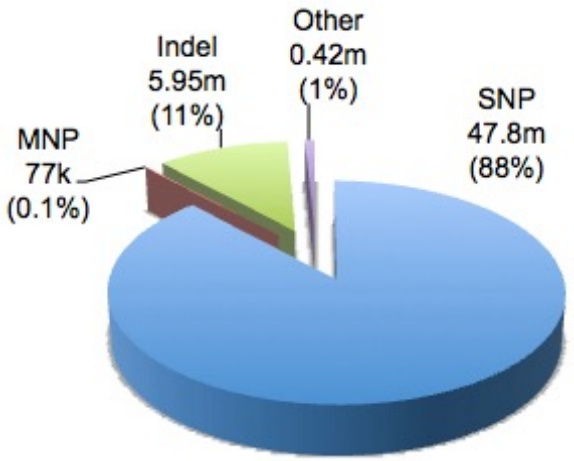
- We tend to use the term mutation in a restrictive way
- A new genetic variant is created by a mutation during replication or reproduction (an error in replication produces a new type of variant)
- Subsequently we refer to the mutation as a variant

Examples of genetic variants

- Single nucleotide polymorphisms (SNPs)
- Insertions/deletions
- Tandem repeats
- Structural variants:
  - Translocations
  - Inversions
  - Copy Number Variants (CNVs)

Examples of genetic variants:
Single nucleotide polymorphisms (SNPs)

- SNPs involve base substitutions, NOT change in # or rearrangement of base pairs.
- Most commonly, these variants are found in the DNA between genes and have no known function.
- About 11.5m known 'common' SNPs (>1% allele frequency)
- Useful genetic markers for Genome Wide Association Studies



dbSNP v135
Massgenomics.org

Single Nucleotide Polymorphism (SNP) example

Right: Two forms of the same chromosomal locus

Here only fifth base pair is polymorphic:

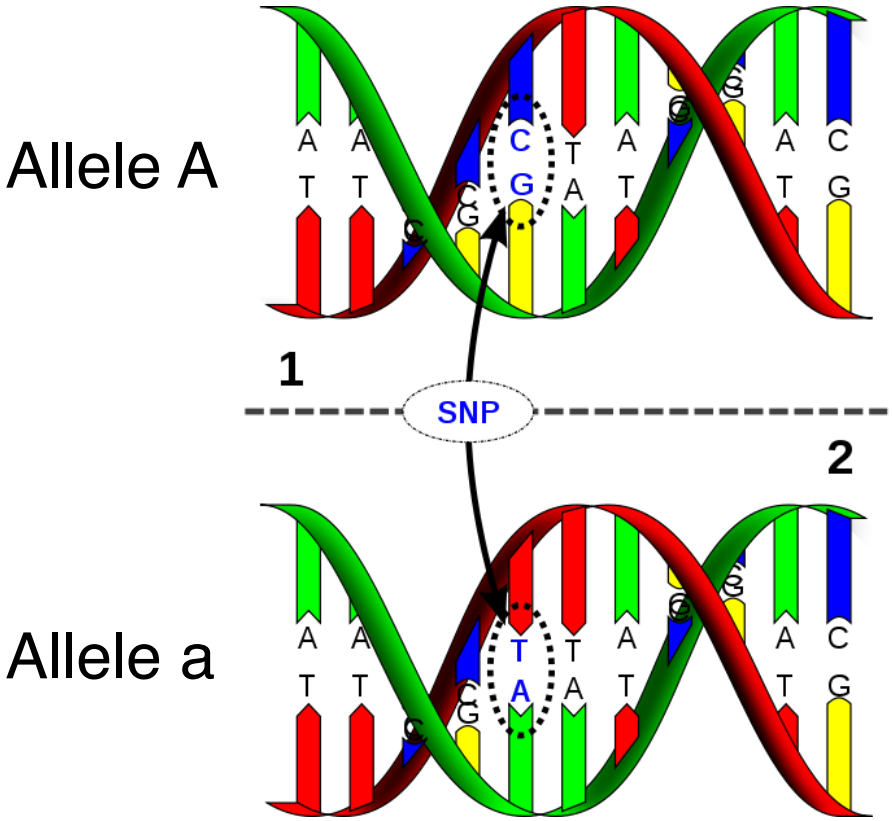
Either C or T alleles

Possible genotype labels at this SNP include:

[CC,CT,TT], or alternatively [AA,Aa,aa]

This person has genotype CT (a.k.a. Aa)

Genotypes could be named CT, AB, 12, wt…



http://isogg.org/wiki/Single-nucleotide_polymorphism

Examples of genetic variants: Indels

- Extra base pairs may be added (insertions) or removed (deletions) from DNA. Collectively, these mutations are called **indels**.

- Deletions of small numbers of base pairs are common

A C G A T C C A C T G

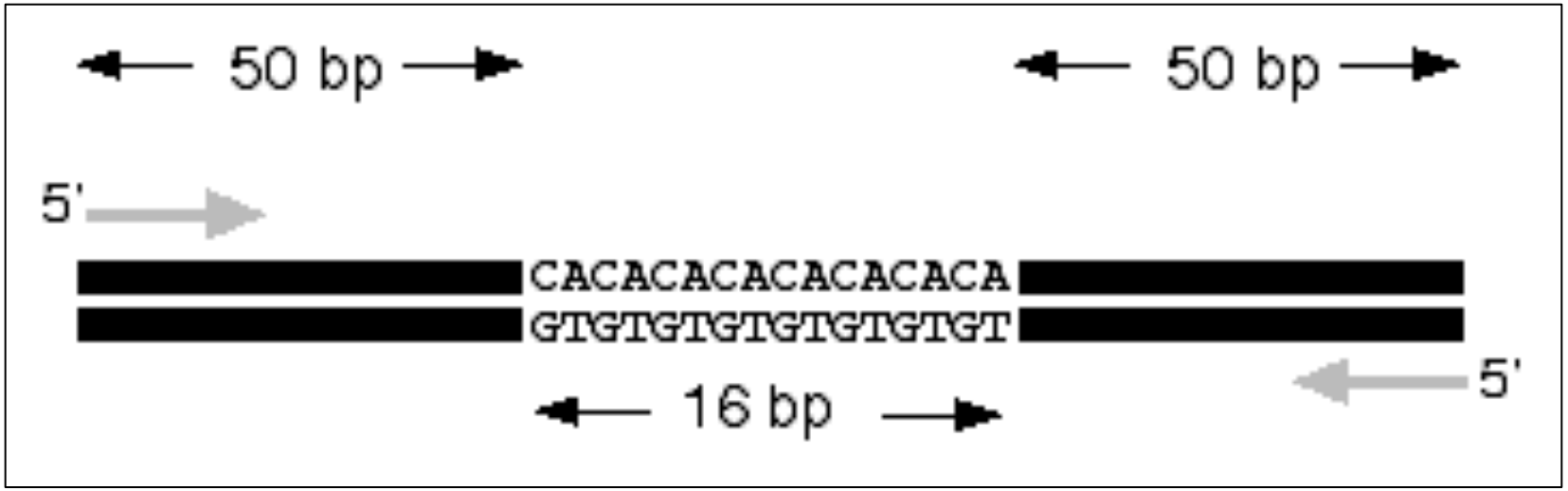
3bp deletion A C G C A C T G

2bp insertion A C G A C C T C C A C T G

Examples of genetic variants:
Short tandem repeats (Microsatellites)

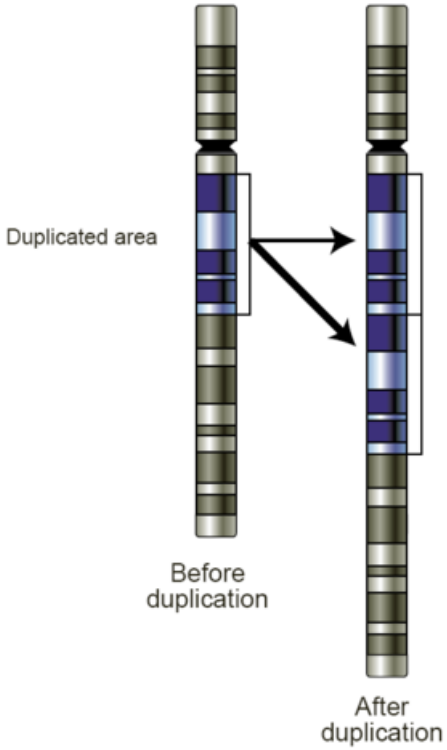
Microsatellites are repeated DNA sequences, where the repeating unit is 2-5bp long, e.g. CACACACA. Most often found in non-coding regions. The number of repeats can vary widely (3—30 or more). Number of genotypes can be very large. Often used for DNA fingerprinting.

Genotype data might be 13,22 or 3,30



Examples of genetic variants:
Copy Number Variants (CNVs)

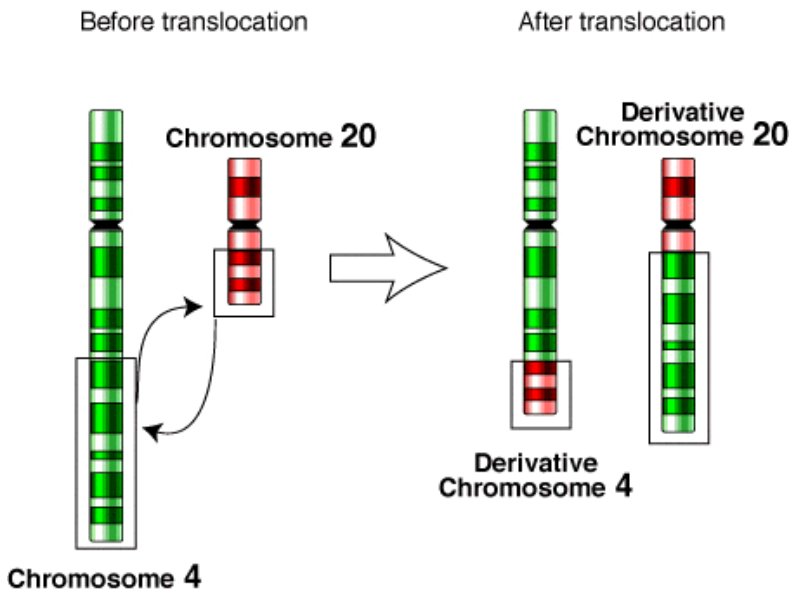
- CNVs involve deletions or duplications of an entire gene or region.
- So a person can have 0, 1, 2, 3, 4… copies of a gene



National Human Genome Research Institute

Other structural variants

- Translocations of one section of a chromosome to another non-homologous chromosome
- Inversions



National Human Genome Research Institute

Genetic Diseases

- An illness caused by abnormalities in the genome
- Types of Genetic Diseases:
  Mendelian vs Complex Disorders
- Relationship of Genetic Variants to Disease

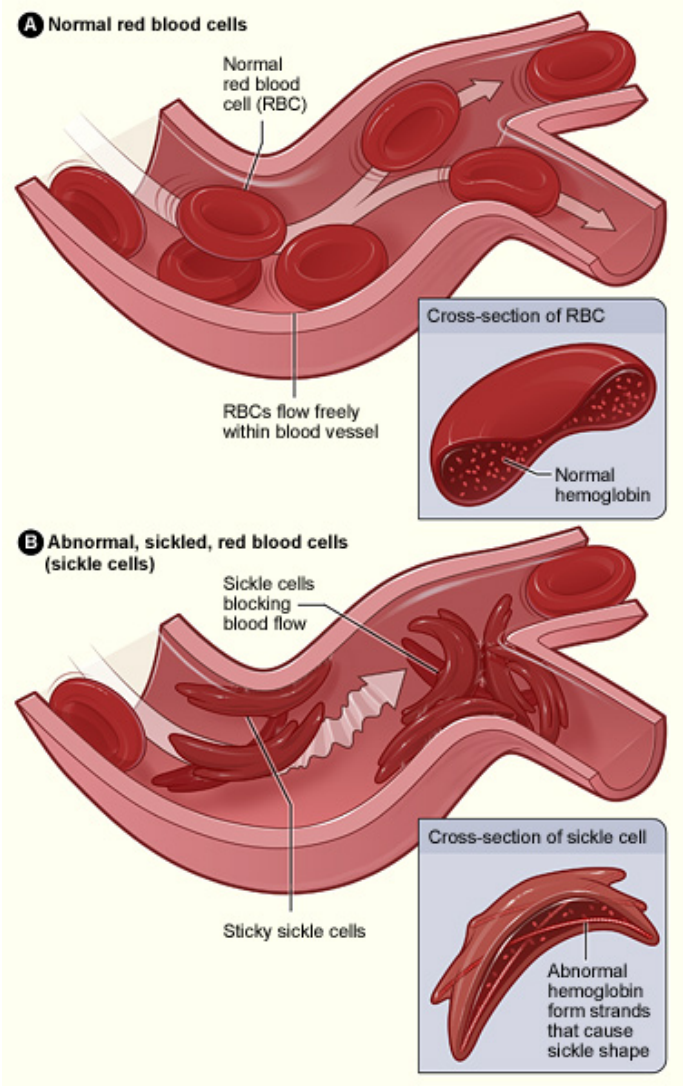
Characteristics of Mendelian Disorders

- Caused by a genetic variant in a single gene (contrast with polygenic disease model)
- Deterministic disease models: P(disease) either 0 or 1 depending on # of genetic variants
- No environmental causes, but can have gene-environment interactions

Sickle Cell Anemia—Classic Mendelian Disorder

First described genetic disorder in the medical literature

- Blood disorder caused by defective red blood cells
- Caused by a single SNP in the Hemoglobin gene
- Two copies of the variant cause Sickle Cell Anemia



NHLBI

A SNP in the Sickle Cell Gene

HBB Sequence in Normal Adult Hemoglobin (Hb A):

Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser
	3			6			9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

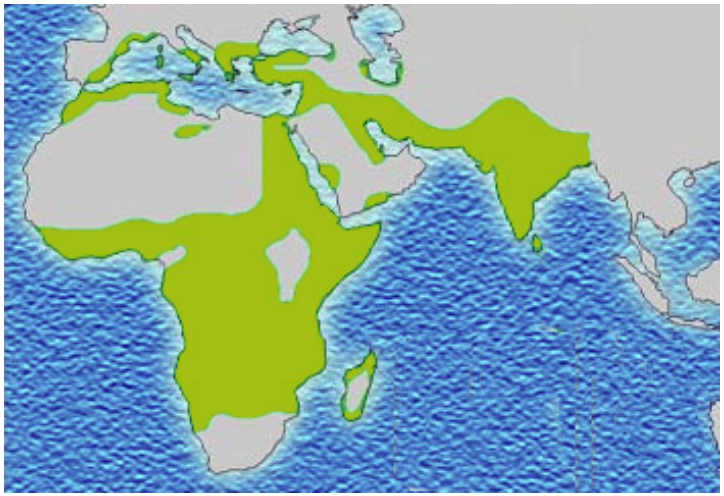
Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser
	3			6			9

Alleles often called A and S.
AA no disease, AS 'Sickle cell trait', SS have Sickle Cell Anemia

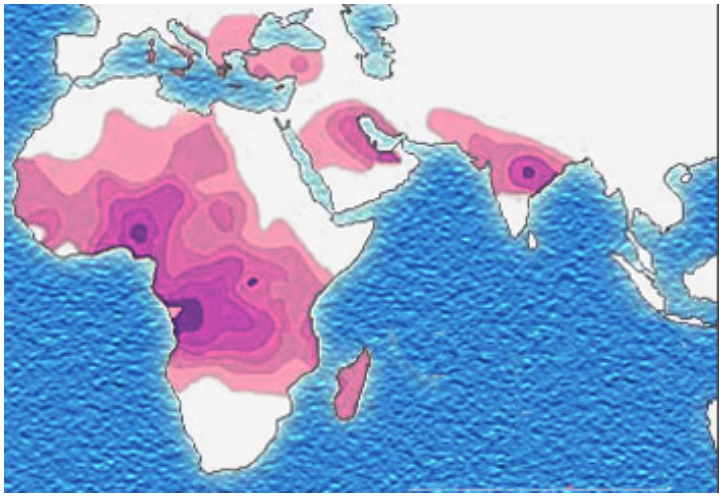
Sickle Cell Anemia—Selection Phenomenon

- One copy causes ‘sickling trait’ and protects against the malaria parasite falciparum
- Selection keeps the variant frequency common

Historic distribution of malaria



Distribution of sickle-cell trait

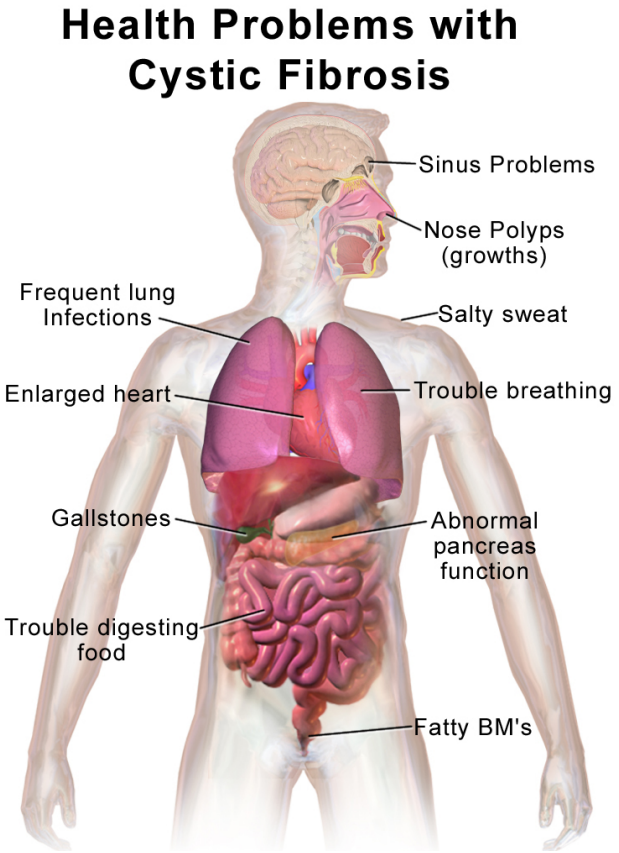


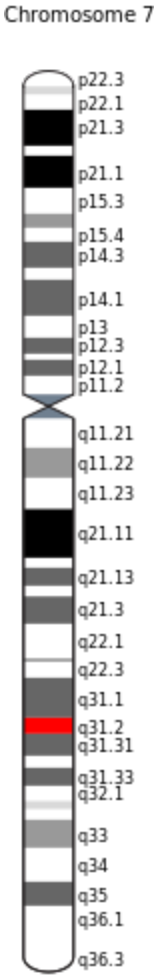
Other Mendelian Diseases: Cystic Fibrosis

- Caused by defects in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Over 1000 different variants cause cystic fibrosis.

not one SNP

- Unlike Sickle Cell disease, no single variant is responsible for all cases of cystic fibrosis. People with cystic fibrosis inherit two variant genes, but the variants need not be the same.





Microsatellite repeat disorder: Fragile X Syndrome

- A non-coding region in the Fragile X mental retardation (FMR1) gene on the human X chromosome contains a locus where the triplet CGG is normally repeated (CGGCGGCGGCGG, etc.) from 5 to 50 or even 100 times without causing a harmful outcome.
- Longer repeats tend to grow longer still from one generation to the next (to as many as 4000 repeats).
- Long (>200) repeats lead to gene silencing and a microscopically visible constriction in the X chromosome: fragile X.
- Males with fragile X show a number of harmful effects including mental retardation. Females who inherit a fragile are only mildly affected.

Complex Traits and Disorders

- A variety of environmental and genetic risk factors
- Many ‘susceptibility’ genes
- May have gene-gene and gene environment interactions as well
- Examples?

Complex trait examples

Gene symbol	Variant(s)	Environmental exposure	Outcome and nature of interaction
Genes for skin pigmentation (for example, *MC1R*)	Variants for fair skin colour	Sunlight or ultraviolet light B	Risk of skin cancer is higher in people with fair skin colour that are exposed to higher amounts of sunlight
CCR5	Δ-32 deletion	HIV	Carriers of the receptor deletion have lower rates of HIV infection and disease progression
MTHFR	Ala222Val polymorphism	Folic acid intake	Homozygotes for the low activity Ala222Val variant are at different risk of colorectal cancer and adenomas if nutritional folate status is low
NAT2	Rapid versus slow acetylator SNPs	Heterocyclic amines in cooked meat	Red meat intake is more strongly associated with colorectal cancer among rapid acetylators
F5	Leiden prothrombotic variant	Hormone replacement	Venous thromboembolism risk is increased in factor V Leiden carriers who take exogenous steroid hormones
UGT1A6	Slow-metabolism SNPs	Aspirin	Increased benefit of prophylactic aspirin use in carriers of the slow metabolism variants
APOE	*E4* allele	Cholesterol intake	Exaggerated changes in serum cholesterol in response to dietary cholesterol changes in *APOE4* carriers

Complex disorder example: Type 2 Diabetes

- Affects 7% of US population
- The disorder is related to problems with the hormone insulin, a regulator of glucose metabolism
- Risk affected by mutations in genes involved in insulin production, regulation
- Polymorphisms in >30 genes have been associated with disease risk, but explain only ~10% of heritability

How were/are disease genes found?

- **Segregation Analysis** (Pedigrees - no genetic data)

- **Linkage Analysis** (Families with disease and genetic data)

- **Association Analysis** (Cases and controls and/or families)

How was the Sickle cell anemia genetic variant located?

- Segregation Analysis: Pedigree analysis of large African families proved the traits were inherited and discovered the nature of the genetic model underlying disease (1920's)
- Variant suspected to be in hemoglobin gene because sickling was evident under microscope
- Defect in the hemoglobin gene located via laboratory studies

Next Time:

Mendel's Laws
Basic Probability Models for Disease