Statistics for Human Genetics and Molecular Biology
Lecture 1: Review Basic Terminology of Genetics

Dr. Yen-Yi Ho (yho@umn.edu)

Sep 09, 2015

Logistics

Lectures	M W F
& Labs:	1:25 to 2:15
Office Hours :	Yen-Yi MW 2:30-3:30
	Cavan MW 2:30-3:30
	Zhiyuan (Jason) Xu Tue 3-4p in Mayo A446
Textbook:	Foulkes (2009): Applied Statistical Genetics with R
	Hahne, Huber, Gentleman, and Falcon (2008): Bioconductor Case Studies
	John Verzani's SimpleR notes
Website:	http://www.biostat.umn.edu/∼cavanr/pubh7445.html

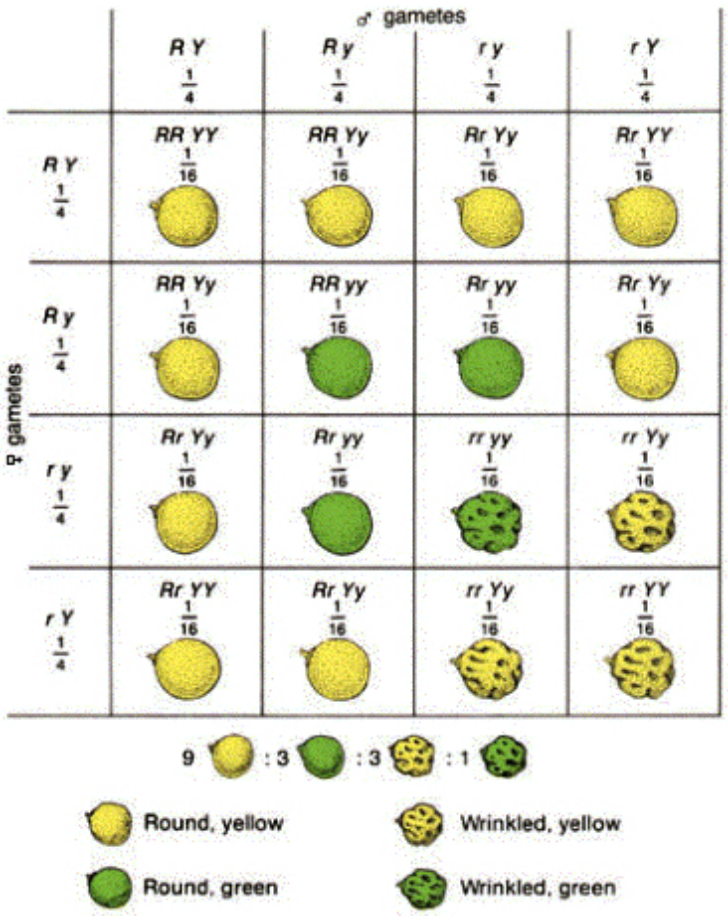
Goals for the Course

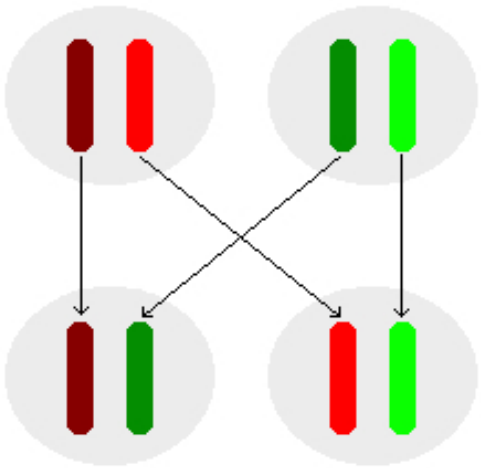
- Basic knowledge of R
- Basics of statistics for human genetics
- Basics of genetic data analyses using R/Bioconductor
- Interpreting results and simple diagnoses

Objectives of Lecture 1

- Review basic terminology of genetics
  - Central dogma of molecular biology
  - Chromosomes, genes, DNA, RNA, and proteins
  - Gene expression
  - Genetic variation
  - Mutations
- Technologies for Genome Analysis

Mendelian Genetics (1866)





Segregation of alleles in the production of sex cells

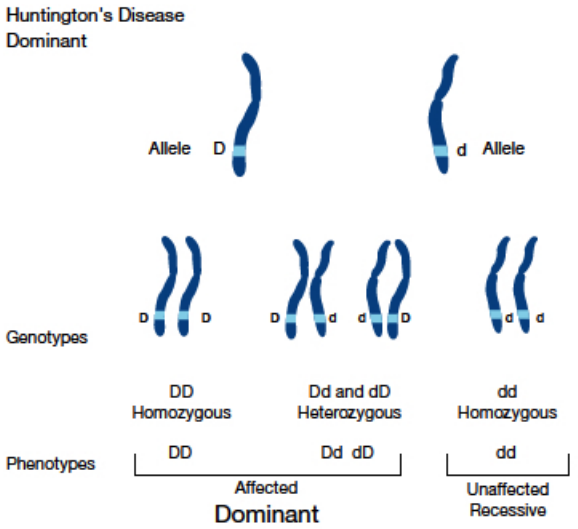
1. the principle of segregation
2. the principle of independent assortment

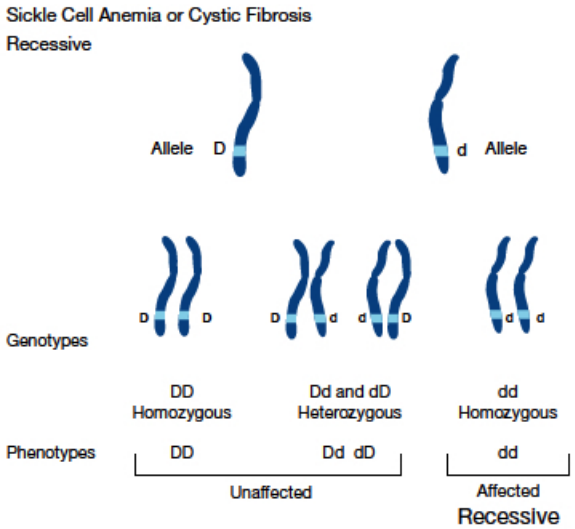
Mendelian Genetics Translates to Modern Genetics

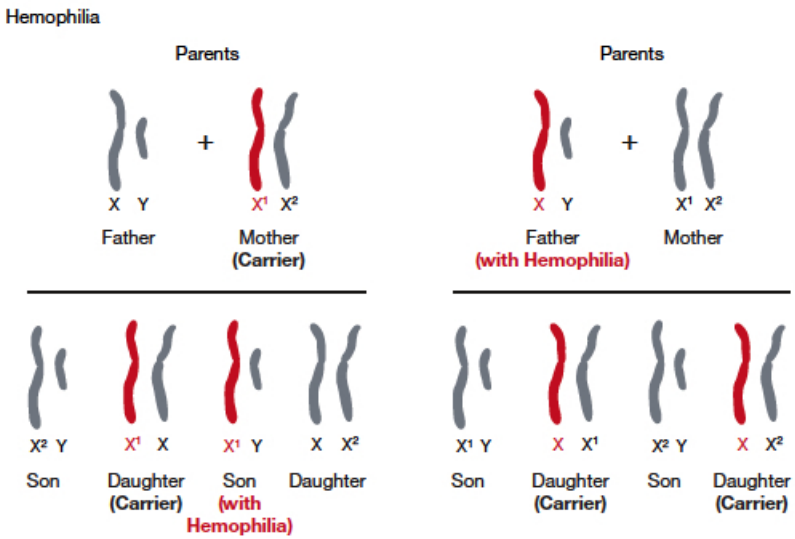
- A parent contributes only a single chromosome within a pair to the offspring.
- A fixed location on a chromosome pair is called a locus, and only those loci coding (for proteins or functional RNA) are typically called genes.
- An allele is the state or type of genetic info at a locus on a single chromosome. Thus there are two alleles at each locus in an individual (for autosomes, and for sex chromosomes in females).

- Example: A particular disease locus has two possible allele types in the population: d (the disease allele) and D (normal).
- Genotype: the joint (unordered) state of the two alleles. Could be dd, DD (called homozygous genotypes), or Dd (heterozygous genotype).
- Alleles that are common in the population are often called wild type while disease alleles are called mutant.
- Phenotype: an observed trait we care about, such as disease status, etc.

Mendelian Genetics Translates to Modern Genetics

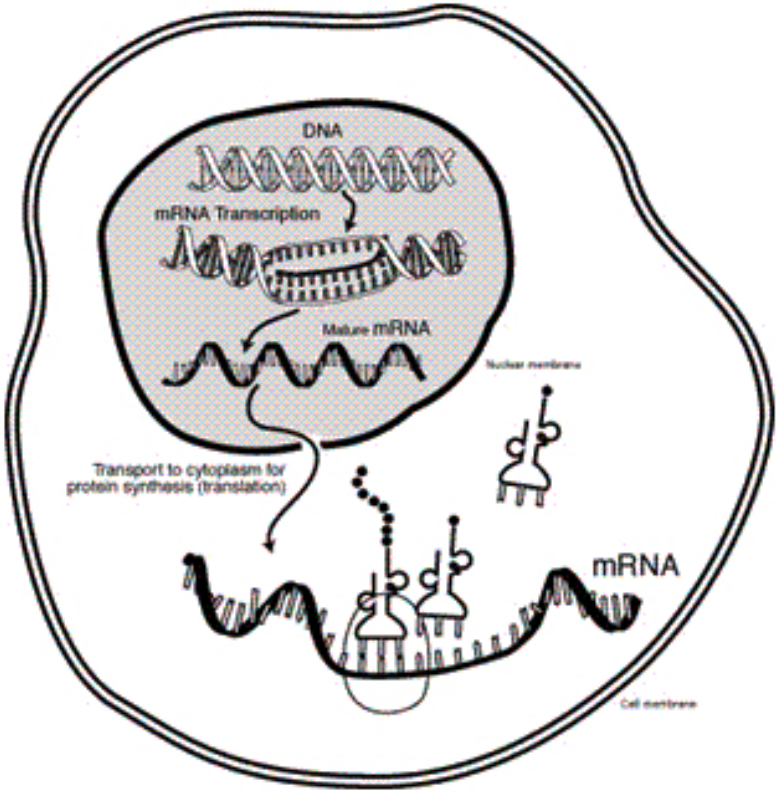


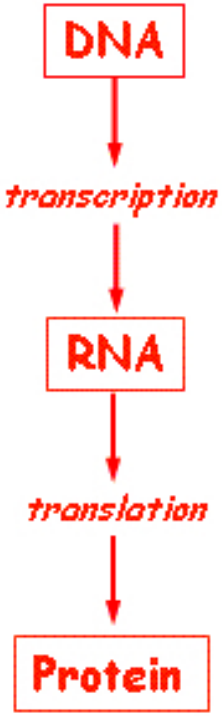




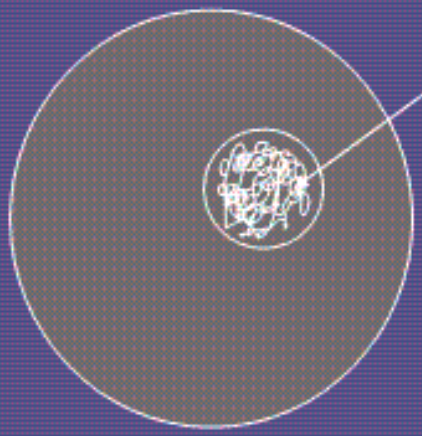
Adapted from NHGRI Talking Glossary

Central Dogma of Biology: Classic View

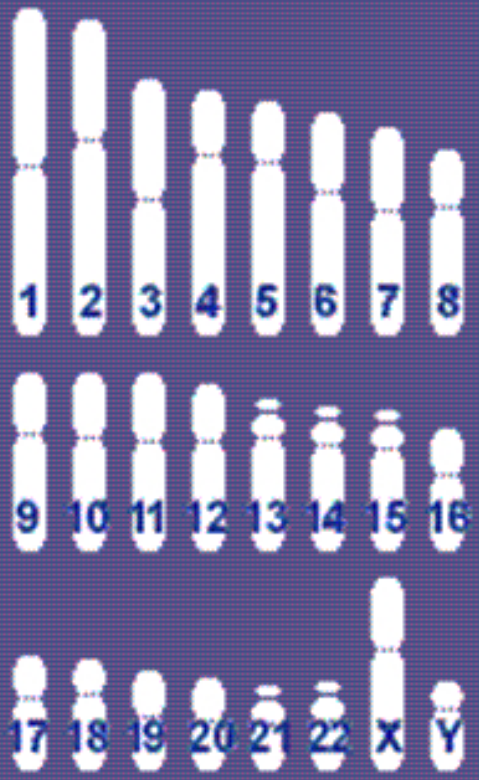




Example: Human genome



Nucleus containing DNA

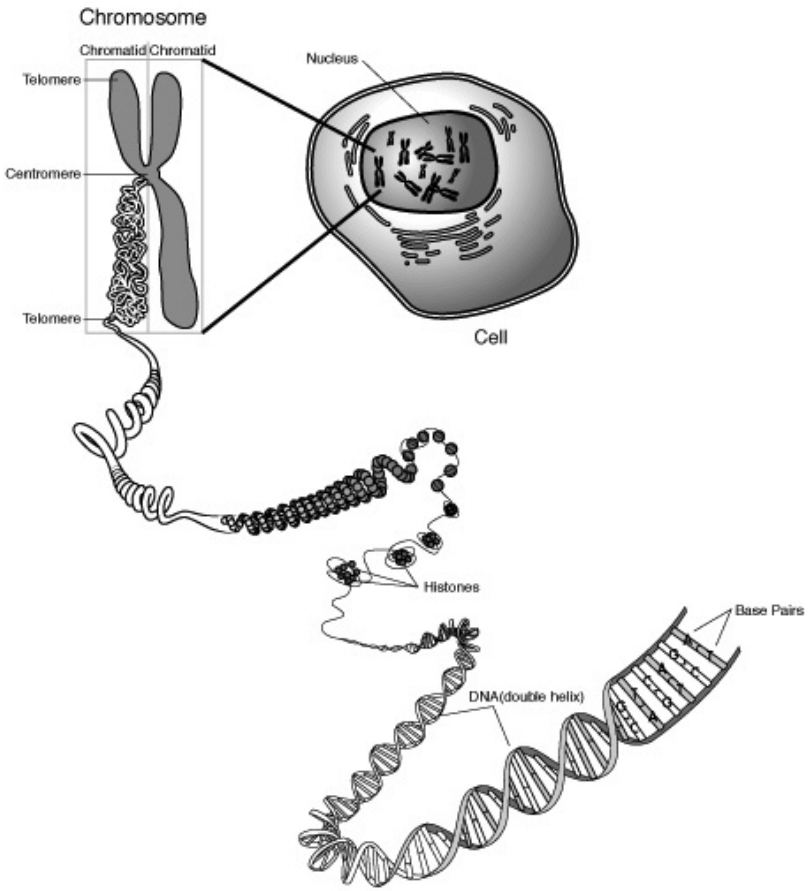


DNA is organized into chromosomes:
22 pairs of autosomes (1-22) and 1 pair of sex chromosomes (X,Y).

Genes, the functional units of heredity, are carried on chromosomes.

Plus the mitochondrial DNA

Base Pairs



IUPAC code	Base
a	adenine
c	cytosine
g	guanine
t (or u)	thymine (or uracil)
r	a/g
y	c/t
s	g/c
w	a/t
k	g/t
m	a/c
b	c/g/t
d	a/g/t
h	a/c/t
v	a/c/g
n	any base
./ -	gap

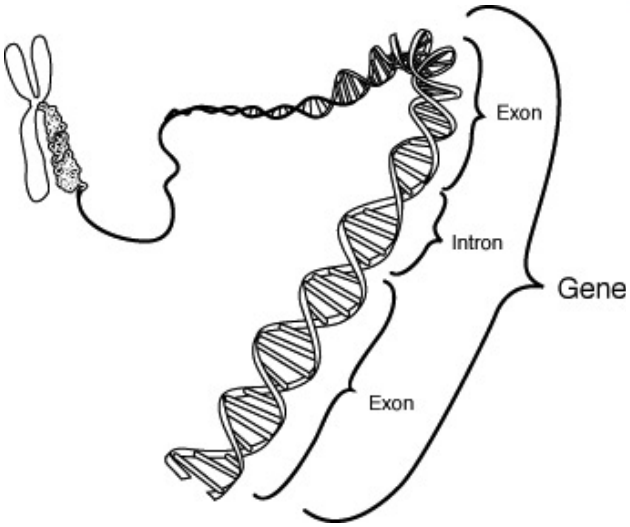
Humans have $\approx 3 \times 10^9$ base pairs in their nuclear genome.

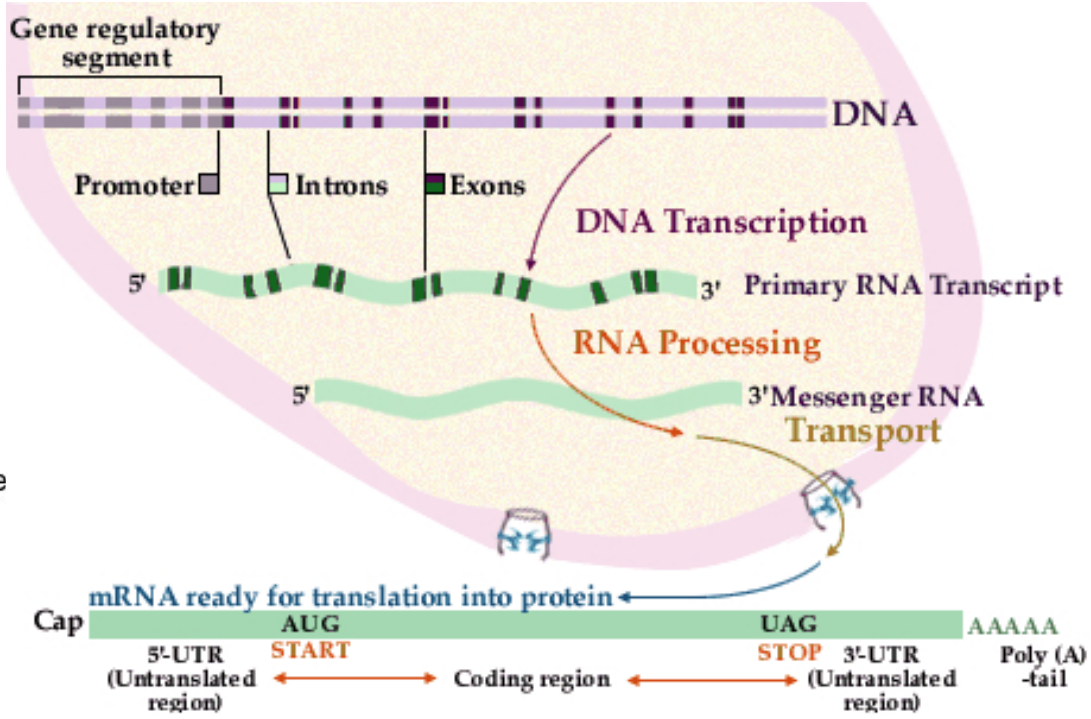
Gene

Gene: a functional and inheritable element in the genome, usually codes for a protein; human genome ≈20,000 genes.

The gene consists of three major structures:

- Regulatory segment
- Exons
- Introns

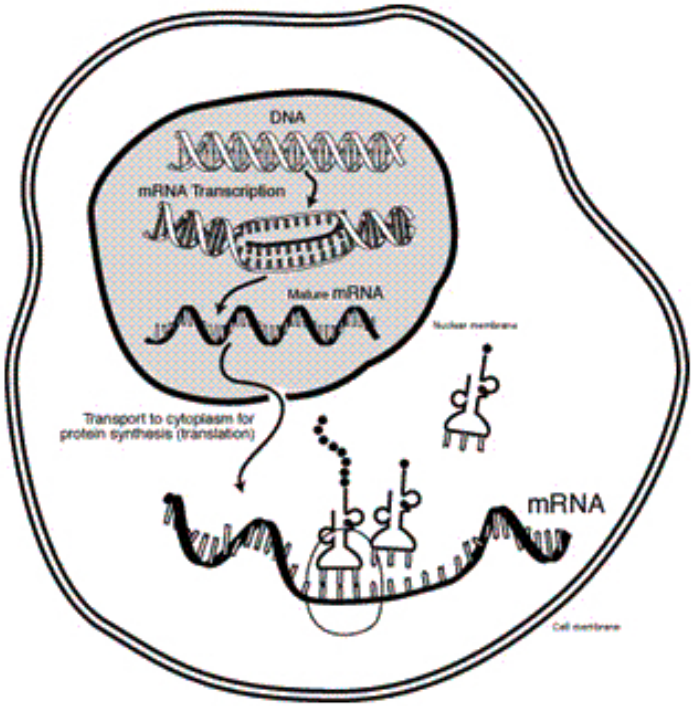


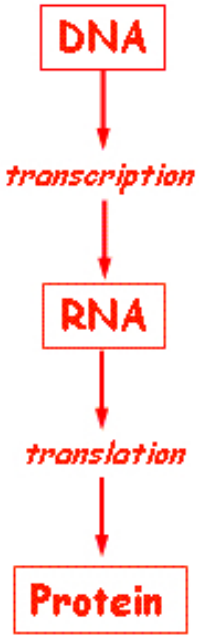


souce: http://www.nobelprize.org/educational/medicine/dna/a/replication/gene.html

Transcription

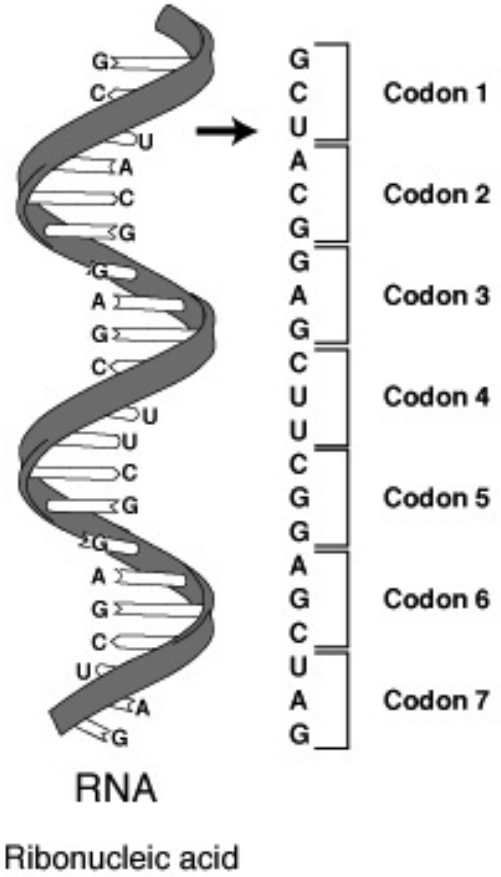
Transcription is the process of making RNA from DNA.





Translation

Translation is the process of translating the sequence of nucleotide bases in DNA/RNA into a sequence of amino acids in a protein.



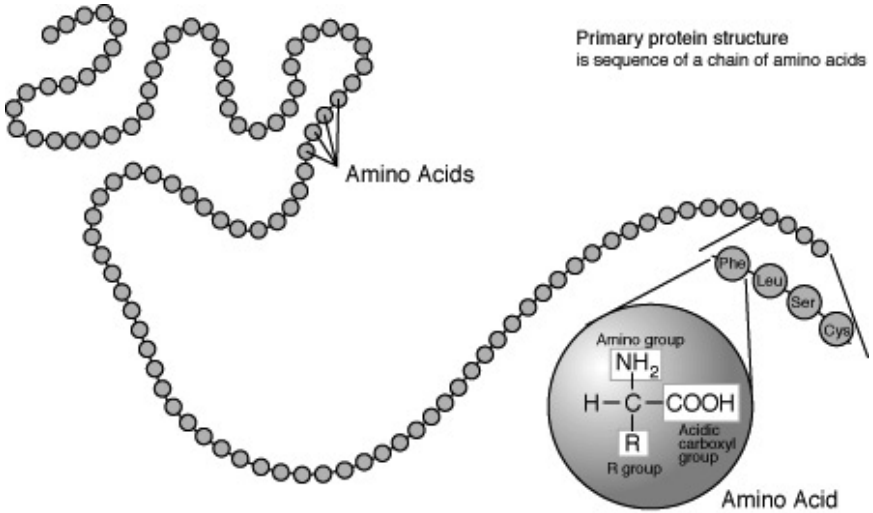
SECOND POSITION

FIRST POSITION

	U	C	A	G	
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
			lysine	arginine	A
	* methionine				G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

THIRD POSITION

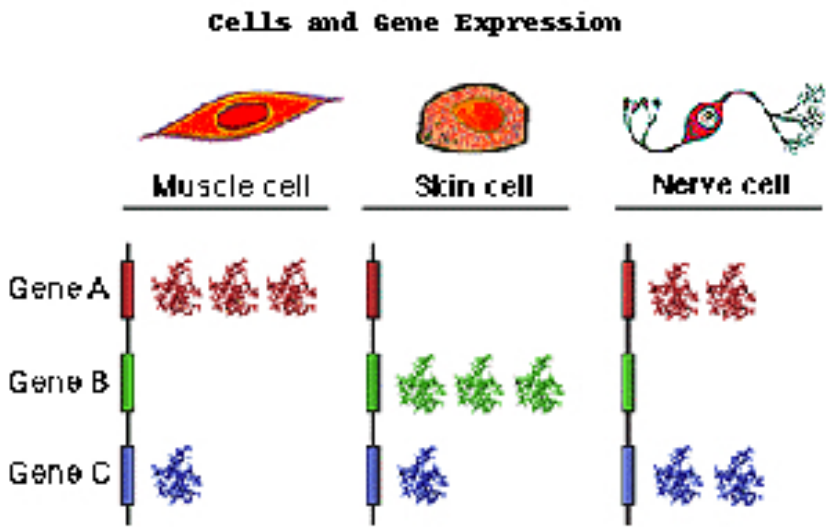
* and start



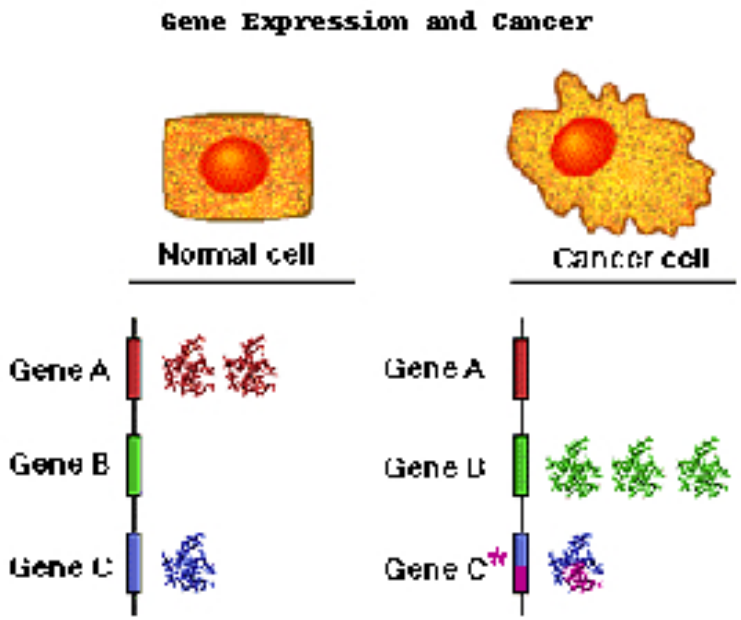
Gene Expression

Gene expression is a highly specific process. Only a small fraction of the genes are expressed, or turned "on," in any particular type of cell.

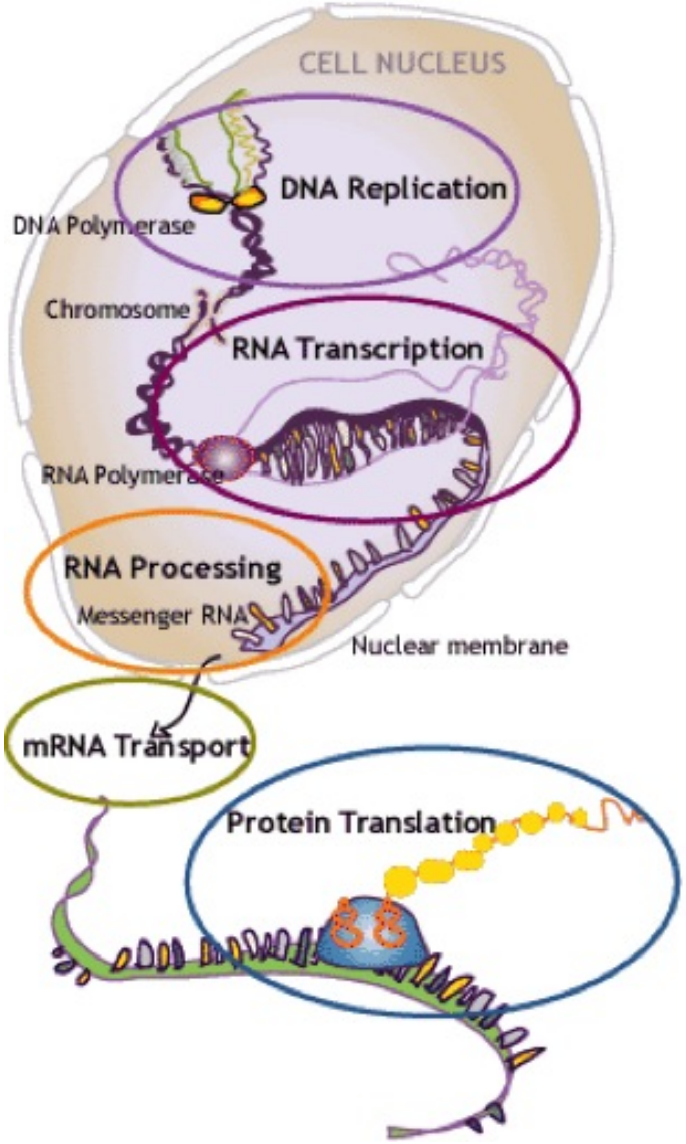
gene expression in different tissues



gene expression in the same tissue, but different points in time



Putting it all together



- DNA:
  Info on chromosome is static, and essentially the same across cells within the individual
- mRNA:
  Not as relevant as protein, but easier to quantify
- Protein:
  Difficult to quantify globally, though very relevant

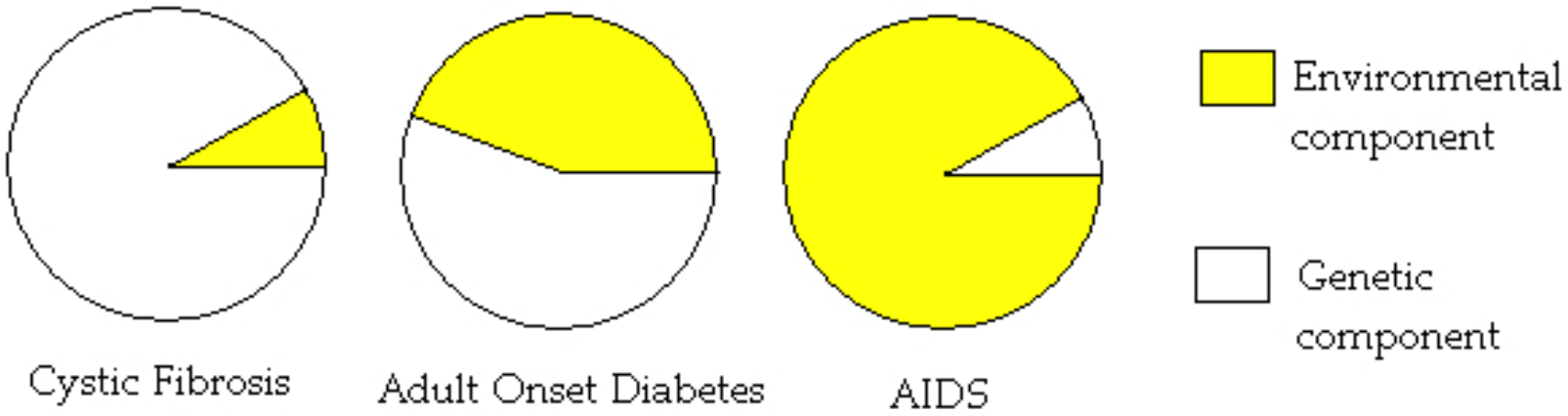
source:

http://www.nobelprize.org/educational/medicine/dna/index.html

Source of Variation

Environment Vs. Gene

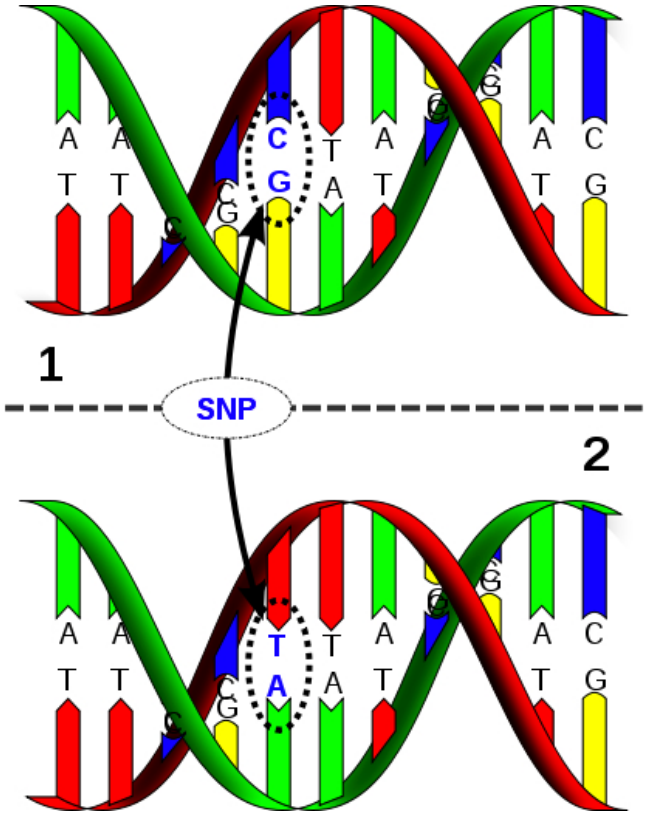
Any two individuals are 99.9% identical in their DNA



Genetic Variations (Polymorphisms)

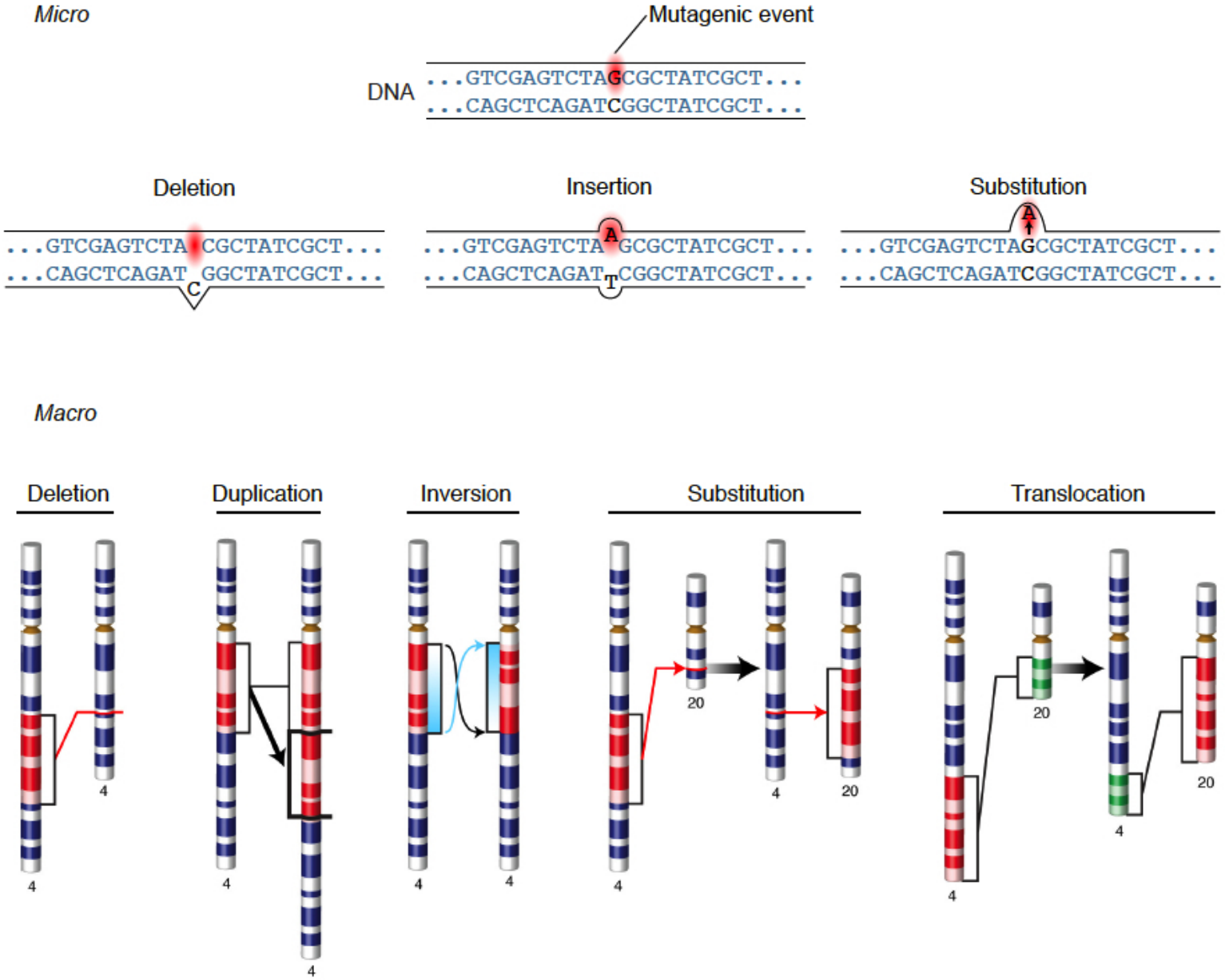
That 0.1 % is very important in defining our differences

- single nucleotide polymorphisms (SNPs, every 300 nucleotide on average)
- small-scale mutation, insertions, deletions
- copy number variations (AAGAAGAAGAAG)

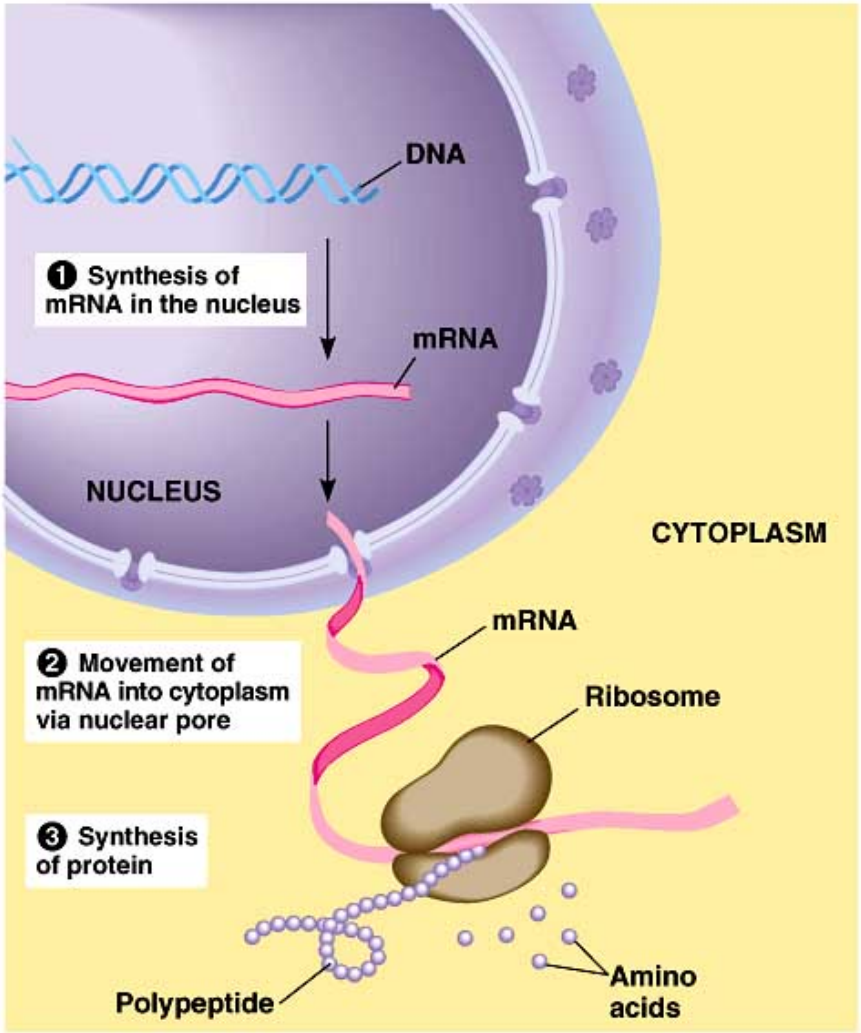


source: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp

Mutations

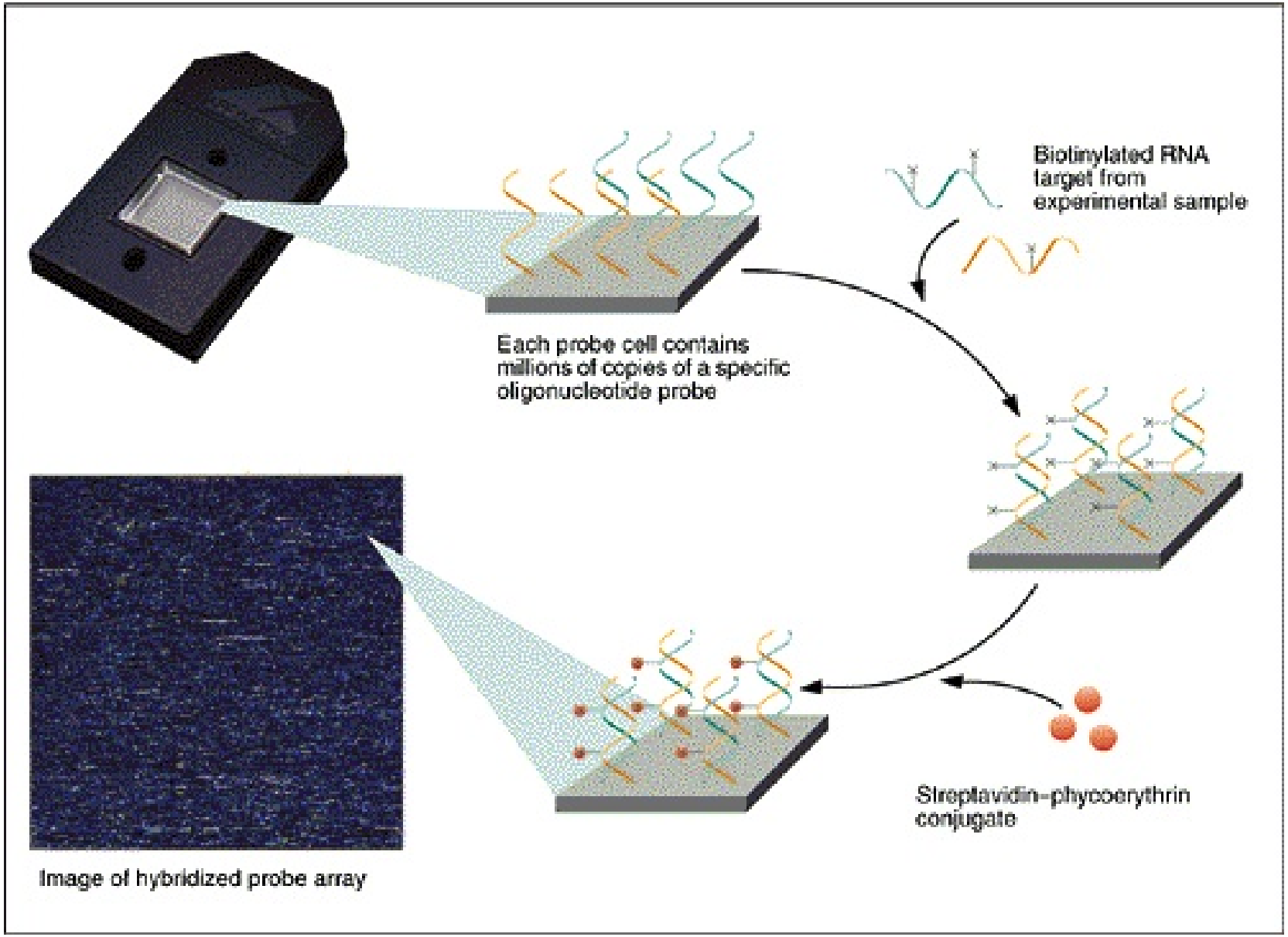

Micro
Mutagenic event
DNA
...GTCGAGTCTAGCGCTATCGCT...
...CAGCTCAGATCGGCTATCGCT...
Deletion
...GTCGAGTCTA CGCTATCGCT...
...CAGCTCAGAT GGCTATCGCT...
C
Insertion
A
...GTCGAGTCTA GCGCTATCGCT...
...CAGCTCAGAT CGGCTATCGCT...
T
Substitution
A
...GTCGAGTCTAGCGCTATCGCT...
...CAGCTCAGATCGGCTATCGCT...
Macro
Deletion
Duplication
Inversion
Substitution
Translocation
4
20

Genome Analysis Technologies

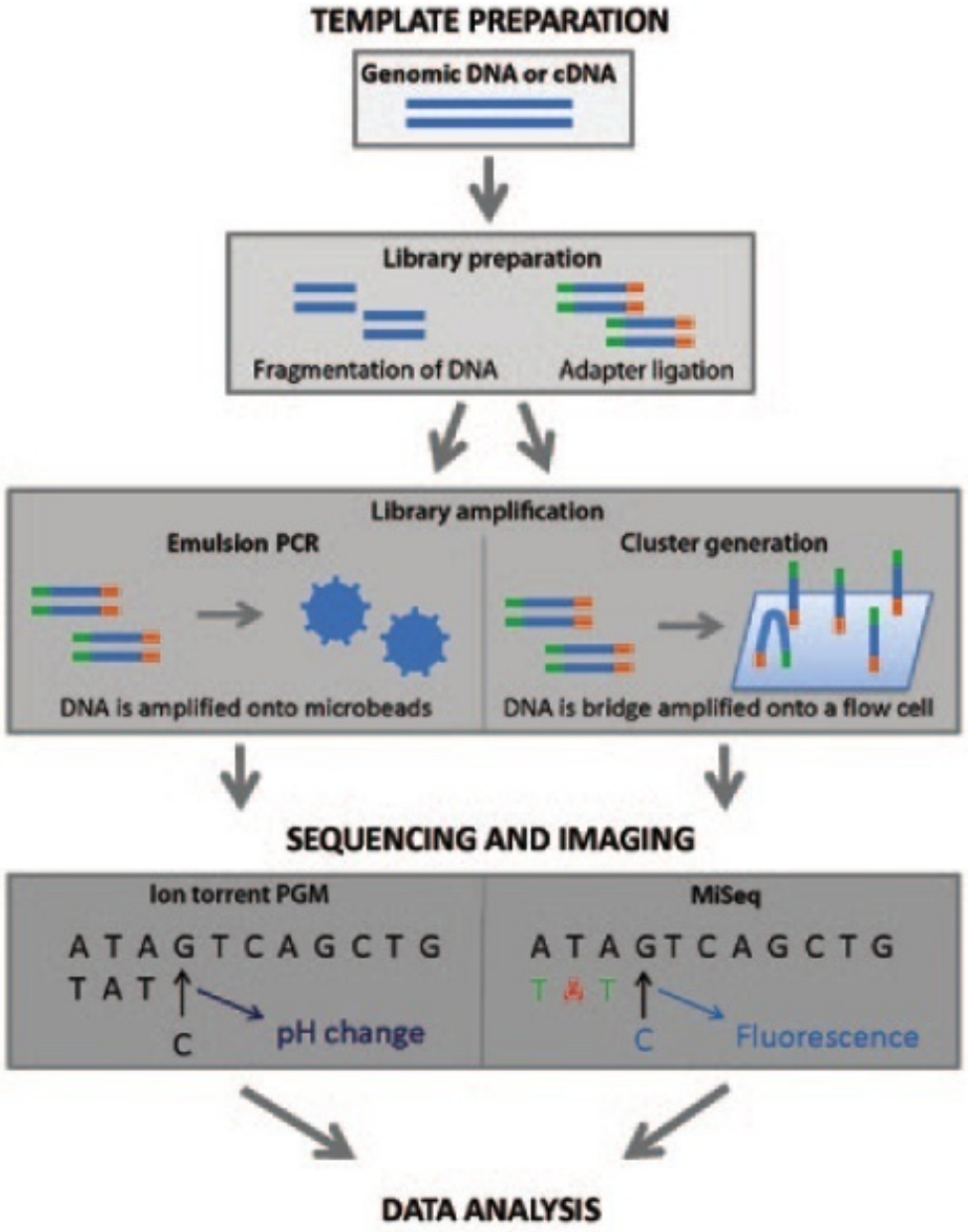


1. DNA
   - Microarrays: SNP, Copy number variation (CNV), Methylation
   - DNA sequencing: SNP, Insertion, Deletion, Mutation, CNV, Methylation
2. mRNA
   - Microarrays
   - RNA sequencing
3. Protein
   - 2-D electrophoresis
   - Maldi-Tof mass spec

General Steps in Obtaining Gene Expression Data



General Steps in Next-Generation Sequencing



Next Lecture

- Review basic terminology of population genetics
  - Crossing Over
  - DNA Recombination
  - Genetic Markers
  - Genetic Association Analysis
- Structures of Genetic Data