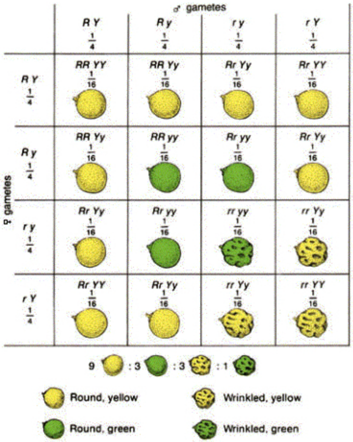
STAT588/BIOL588: Genomic Data Science
Lecture 3: Review Basic Terminology of Genetics

Guest Lecture: Dr. Shannon Davis
Department of Biological Sciences

Objectives of Lecture 3

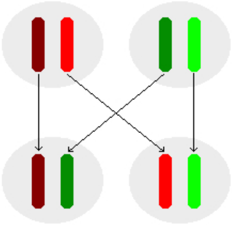
- Biology in a nutshell
  - Central dogma of molecular biology
  - Chromosomes, genes, DNA, RNA, and proteins
  - Gene expression
  - Genetic variation
  - Mutations
- Technologies for Genome Analysis

Mendelian Genetics (1866)



♀ gametes \ ♂ gametes	R Y $\frac{1}{4}$	R y $\frac{1}{4}$	r y $\frac{1}{4}$	r Y $\frac{1}{4}$
R Y $\frac{1}{4}$	RR YY $\frac{1}{16}$	RR Yy $\frac{1}{16}$	Rr Yy $\frac{1}{16}$	Rr YY $\frac{1}{16}$
R y $\frac{1}{4}$	RR Yy $\frac{1}{16}$	RR yy $\frac{1}{16}$	Rr yy $\frac{1}{16}$	Rr Yy $\frac{1}{16}$
r y $\frac{1}{4}$	Rr Yy $\frac{1}{16}$	Rr yy $\frac{1}{16}$	rr yy $\frac{1}{16}$	rr Yy $\frac{1}{16}$
r Y $\frac{1}{4}$	Rr YY $\frac{1}{16}$	Rr Yy $\frac{1}{16}$	rr Yy $\frac{1}{16}$	rr YY $\frac{1}{16}$

9 : 3 : 3 : 1
Round, yellow
Round, green
Wrinkled, yellow
Wrinkled, green



Segregation of alleles in the production of sex cells

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

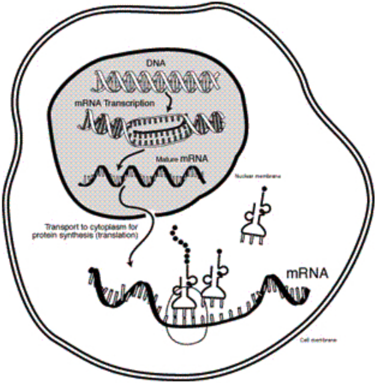
Mendelian Genetics Translates to Modern Genetics

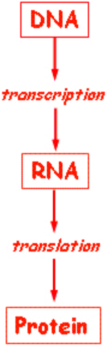
- A parent contributes only a single chromosome within a pair to the offspring.
- A fixed location on a chromosome pair is called a locus, and only those loci coding (for proteins or functional RNA) are typically called genes.
- An allele is the state or type of genetic info at a locus on a single chromosome. 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.

Central Dogma of Biology: Classic View

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




DNA

transcription

RNA

translation

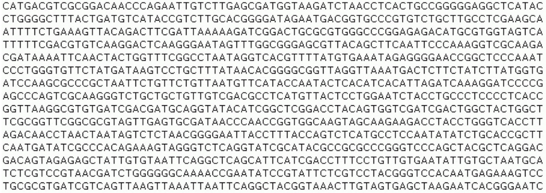
Protein

DNA (DeoxyriboNucleic Acid)

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

DNA sequence

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

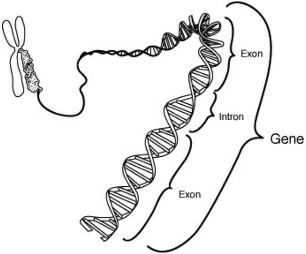


```
CATGACGTCGCGGACAACCCAGAATTGTCTTGAGCGATGGTAAGATCTAACCTCACTGCCGGGGGAGGCTCATAC
CTGGGGCTTTACTGATGTCATACCGTCTTGCACGGGGATAGAATGACGGTGCCCGTGTCTGCTTGCCTCGAAGCA
ATTTTCTGAAAGTTACAGACTTCGATTAAAAAGATCGGACTGCGCGTGGGCCCGGAGAGACATGCGTGGTAGTCA
TTTTTCGACGTGTCAAGGACTCAAGGGAATAGTTTGGCGGGAGCGTTACAGCTTCAATTCCCAAAGGTCGCAAGA
CGATAAAATTCAACTACTGGTTTCGGCCTAATAGGTCACGTTTTATGTGAAATAGAGGGGAACCGGCTCCCAAAT
CCCTGGGTGTTCTATGATAAGTCCTGCTTTATAACACGGGGCGGTTAGGTTAAATGACTCTTCTATCTTATGGTG
ATCCAAGCGCCCGCTAATTCTGTTCTGTTAATGTTCATACCAATACTCACATCACATTAGATCAAAGGATCCCCG
AGCCCAGTCGCAAGGGTCTGCTGCTGTTGTCGACGCCTCATGTTACTCCTGGAATCTACCTGCCCTCCCCTCACC
GGTTAAGGCGTGTGATCGACGATGCAGGTATACATCGGCTCGGACCTACAGTGGTCGATCGACTGGCTACTGGCT
TCGCGGTTCGGCGCGTAGTTGAGTGCGATAACCCAACCGGTGGCAAGTAGCAAGAAGACCTACCTGGGTCACCTT
AGACAACCTAACTAATAGTCTCTAACGGGGAATTACCTTTACCAGTCTCATGCCTCCAATATATCTGCACCGCTT
CAATGATATCGCCCACAGAAAGTAGGGTCTCAGGTATCGCATACGCCGCGCCCGGGTCCCAGCTACGCTCAGGAC
GACAGTAGAGAGCTATTGTGTAATTCAGGCTCAGCATTCATCGACCTTTCCTGTTGTGAATATTGTGCTAATGCA
TCTCGTCCGTAACGATCTGGGGGGCAAAACCGAATATCCGTATTCTCGTCCTACGGGTCCACAATGAGAAAGTCC
TGCGCGTGATCGTCAGTTAAGTTAAATTAATTCAGGCTACGGTAAACTTGTAGTGAGCTAAGAATCACGGGAATC
```

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

Gene

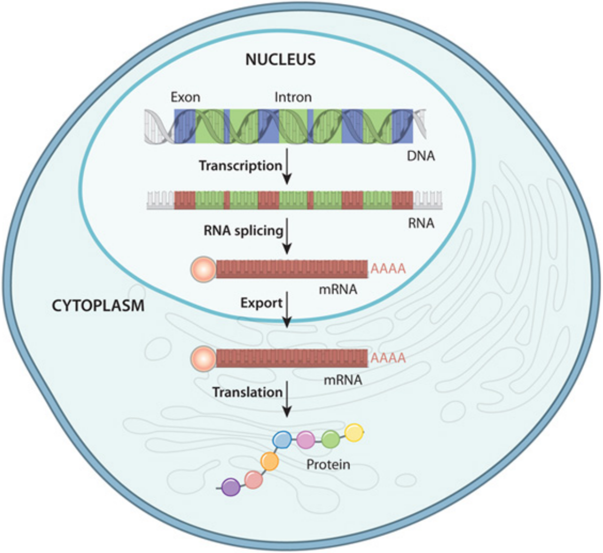
- A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions.
- Or simply, a piece of “useful” DNA sequence.



In a nutshell (for statistician)

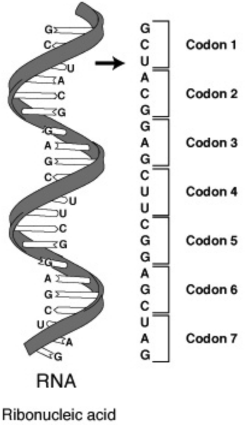
- **enhancer**: a region for enhancing gene expression. Not necessarily closes to the gene.
- **promoter**: at the beginning of the gene, helps transcription.
- **exons**: the "useful" part of the gene, will appear in the mRNA product.
- **introns**: the "spacer" between exons, will NOT be in the mRNA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

Gene structure and splicing



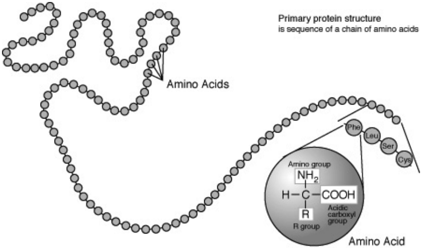
RNA (RiboNucleic Acid)

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



Protein

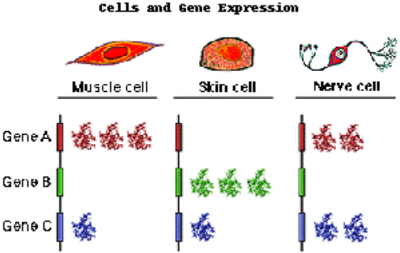
- The final project of gene expression process, workhorses in the cells.
- A chain of amino acid.
- Every 3 nucleotide is translated into one amino acid during translation.
- There are 20 types of amino acids, so a protein can be thought as a string from a 20 character alphabet.
- 3D protein structure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



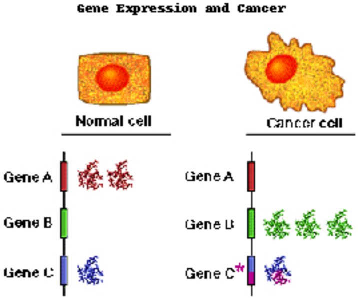
Gene Expression

Gene expression is a term that is used to describe the entire process of translation and transcription of a gene. 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

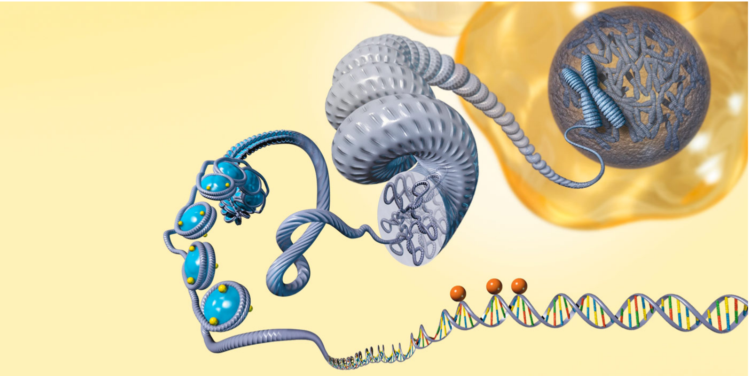


gene expression in the same tissue, but different points in time

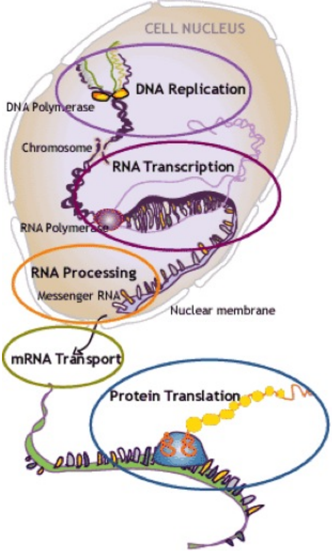


Epigenetics

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



Putting it all together



source:

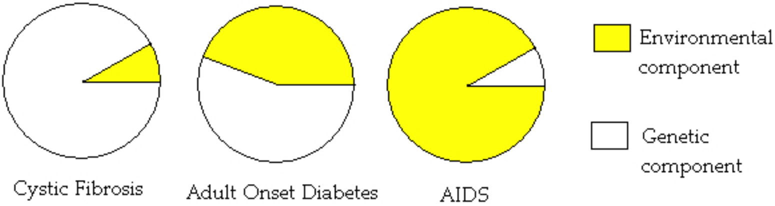
http://www.nobelprize.org/educational/medicine/dna/index.html

- DNA sequence:
  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 of Variation

Environment Vs. Gene

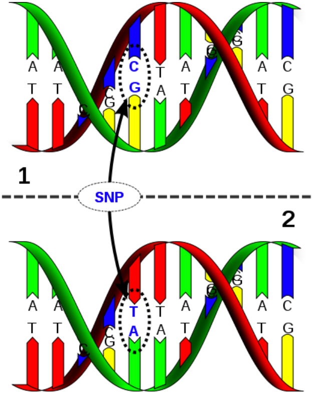
Any two individuals are 99.9% identical in their DNA



Genetic Variations (Polymorphisms)

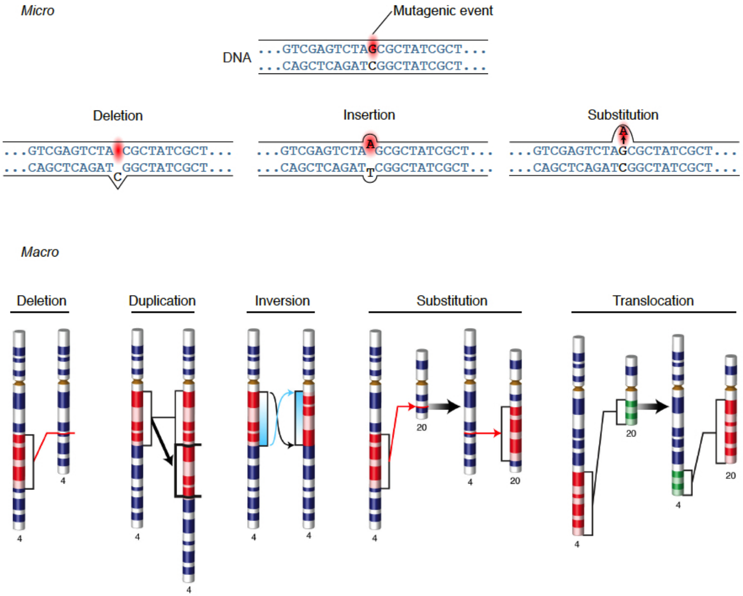
That 0.1 % is very important in defining our differences

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

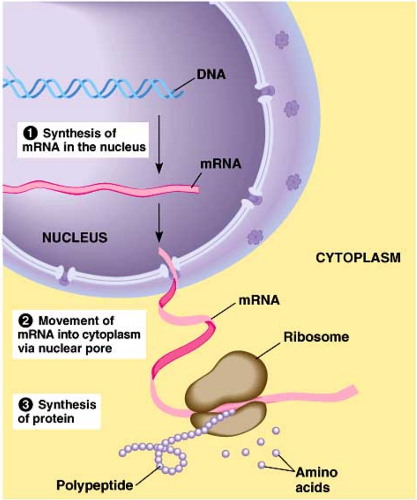


source: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp

Mutations

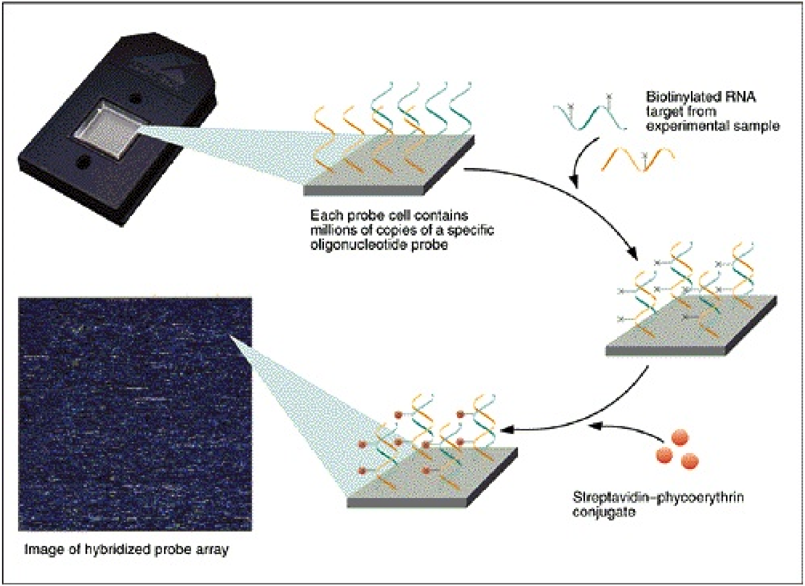

Micro
Mutagenic event
DNA
...GTCGAGTCTAGCGCTATCGCT...
...CAGCTCAGATCGGCTATCGCT...
Deletion
...GTCGAGTCTA CGCTATCGCT...
...CAGCTCAGAT GGCTATCGCT...
C
Insertion
A
...GTCGAGTCTA GCGCTATCGCT...
...CAGCTCAGAT CGGCTATCGCT...
T
Substitution
A
...GTCGAGTCTAGCGCTATCGCT...
...CAGCTCAGATCGGCTATCGCT...
Macro
Deletion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

Genome Analysis Technologies

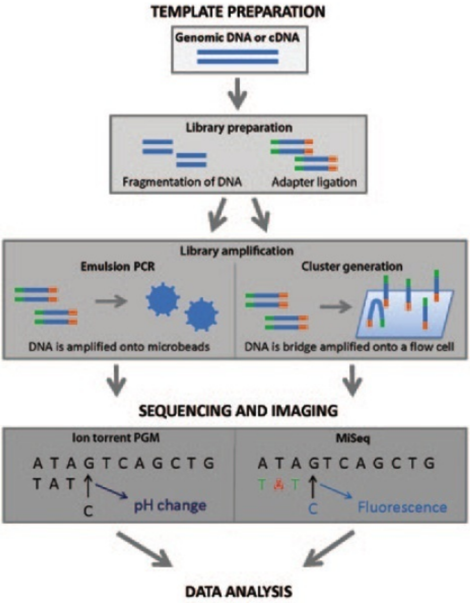

DNA
1 Synthesis of mRNA in the nucleus
mRNA
NUCLEUS
CYTOPLASM
mRNA
2 Movement of mRNA into cytoplasm via nuclear pore
Ribosome
3 Synthesis of protein
Amino acids
Polypeptide

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

General Steps in Obtaining Gene Expression Data


Biotinylated RNA target from experimental sample
Each probe cell contains millions of copies of a specific oligonucleotide probe
Streptavidin-phycoerythrin conjugate
Image of hybridized probe array

General Steps in Next-Generation Sequencing



Next

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