PHP2620: Statistical Methods in Bioinformatics

http://www.stat.brown.edu/ZWU/teaching/php2620.aspx

Jan 23, 2014

Some slides taken/modified from Dr. Hao Wu's lecture

What is Bioinformatics

- The NIH Biomedical Information Science and Technology Initiative definition:

  Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

The beauty and challenge of high throughput

- Assaying thousands to millions of targets simultaneously
- Hypothesis generating vs Hypothesis driven
- Curse of dimensionality, “large p small N”
  - The problem of small N: where law of large numbers, central limit theorem, asymptotics no longer apply
- Computational burden/real time analysis
- Multiple testing and false positives

Course structure

- Statistical Methods for Genomic Data
  - Microarray and Sequencing
  - Differential expression (binding, methylation)
  - Clustering and classification
  - Genome wide association
  - Gene set analysis
  - Biological Sequence analysis

If time allows

- Other technologies: HTS, qRT-PCR

Review/Intro to molecular biology

Keep in mind these questions:

What do we want to make inference of?

What do we want to measure?

How do we measure?

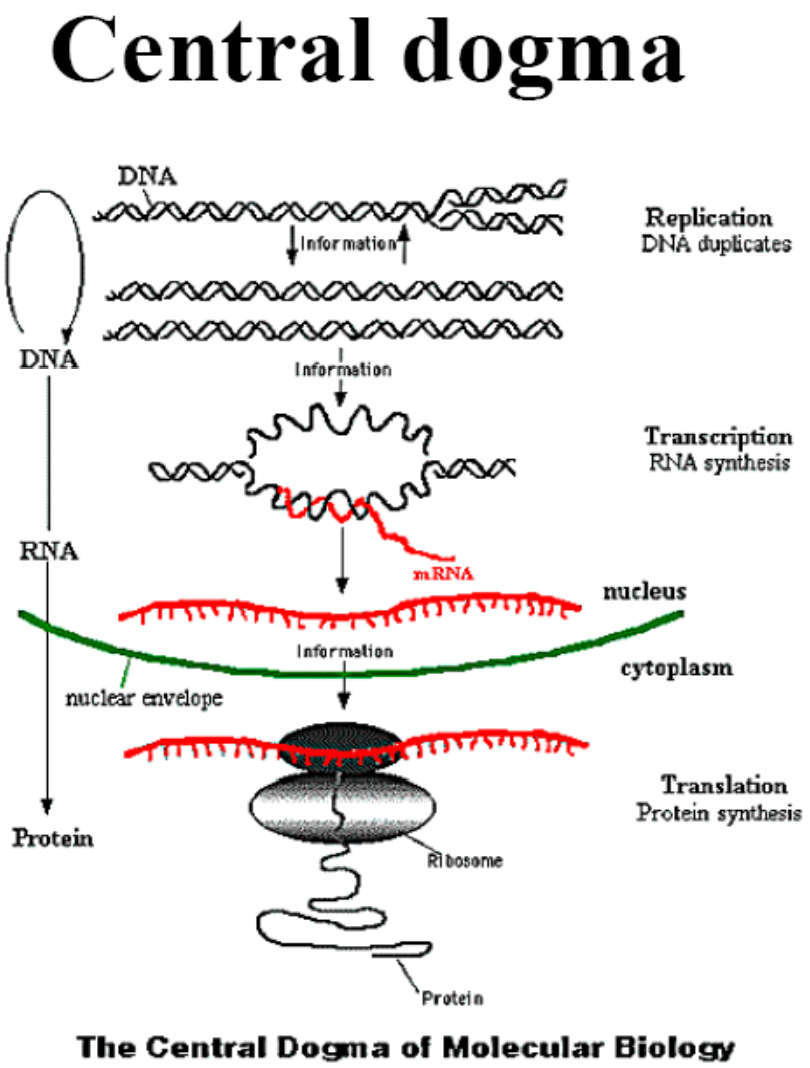
The flow of genetic information

DNA: Source code

compilation

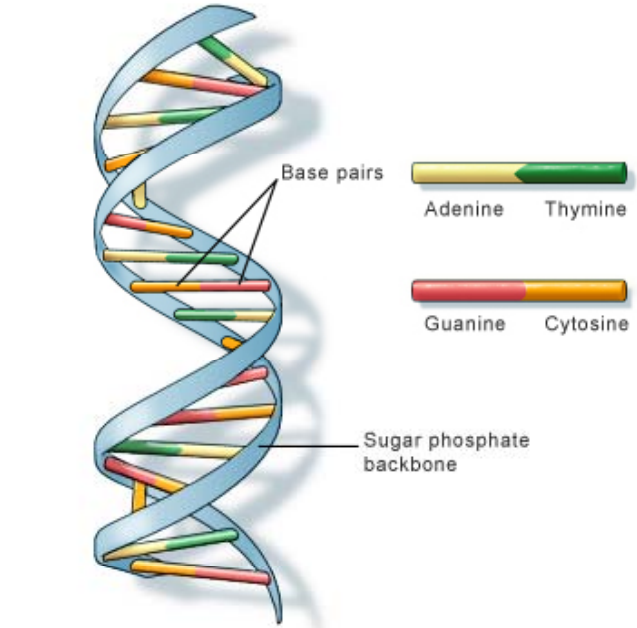
mRNA: Executable file

Protein: output/product

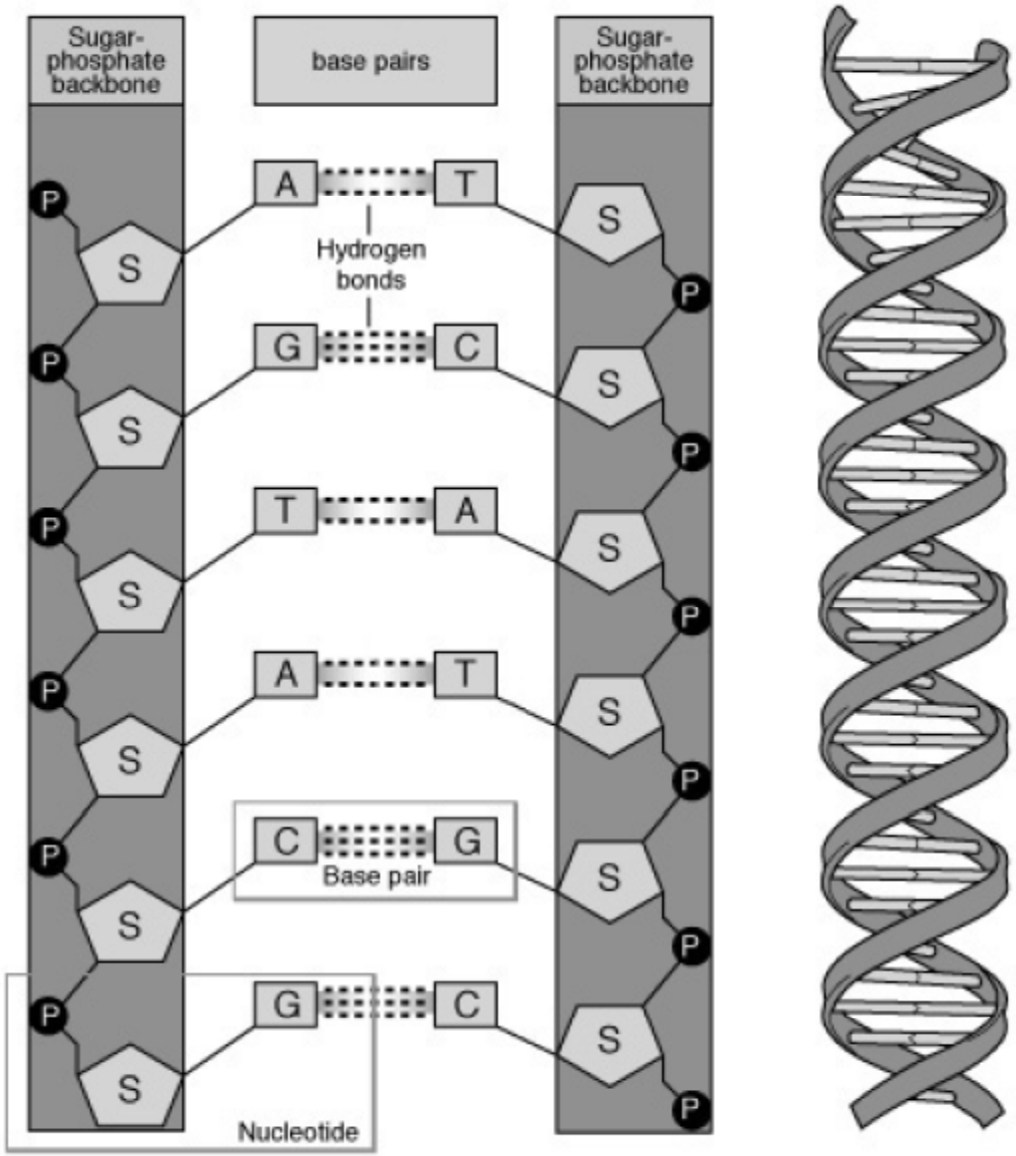


DNA (DeoxyriboNucleic Acid)

- Nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms and some viruses.
- DNA consists of two long polymers (strands) of simple units (bases) called nucleotides: A, C, G, T. So a DNA sequence can be thought as a long string of four letters.
- Base pairing: **A-T, C-G.**
- Two strands entwine in the shape of double helix.

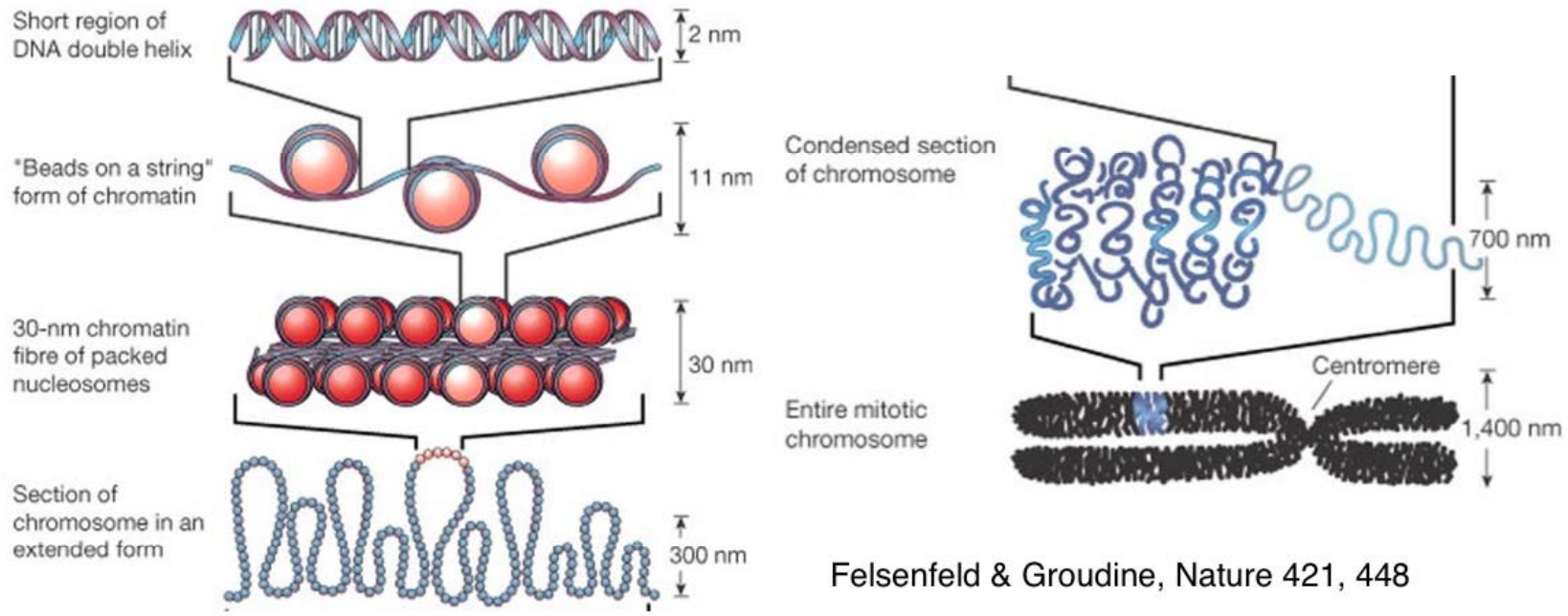


DNA


Sugar-phosphate backbone
base pairs
Sugar-phosphate backbone
A
T
Hydrogen bonds
G
C
T
A
A
T
C
G
Base pair
G
C
P
S
Nucleotide

- DNA
  - Genome: the entire collection of genetic information of an organism, in the form of DNA (with the exception of RNA viruses)
  - All cells in an organism contain the same genome (with the exception of some cells in the immune system) and the genome is usually fixed (with the exception of mutations)
  - Double stranded structure (like the positive/negative file of the same image, so each strand contains the same information)
  - 4 letter alphabet GCAT , G-C, A-T pairs

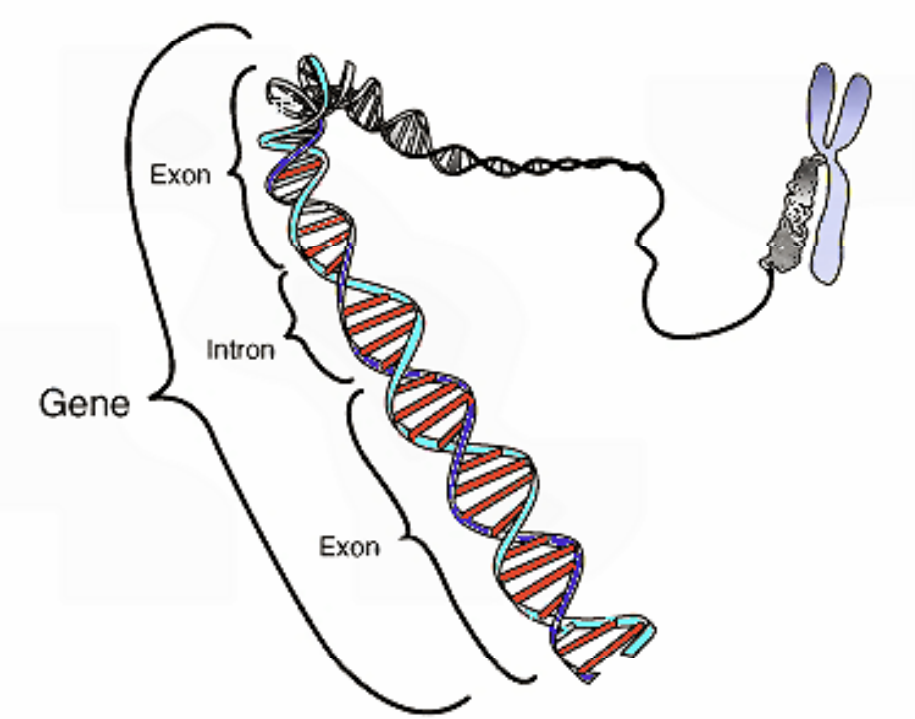
Chromosome: organized structure of DNA and proteins



- Ploidy: number of set of chromosomes in a cell.
  - monoploid, diploid or polyploid.
  - Human are diploid: cells have two copies of each chromosome, one from mother and one from father.

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.



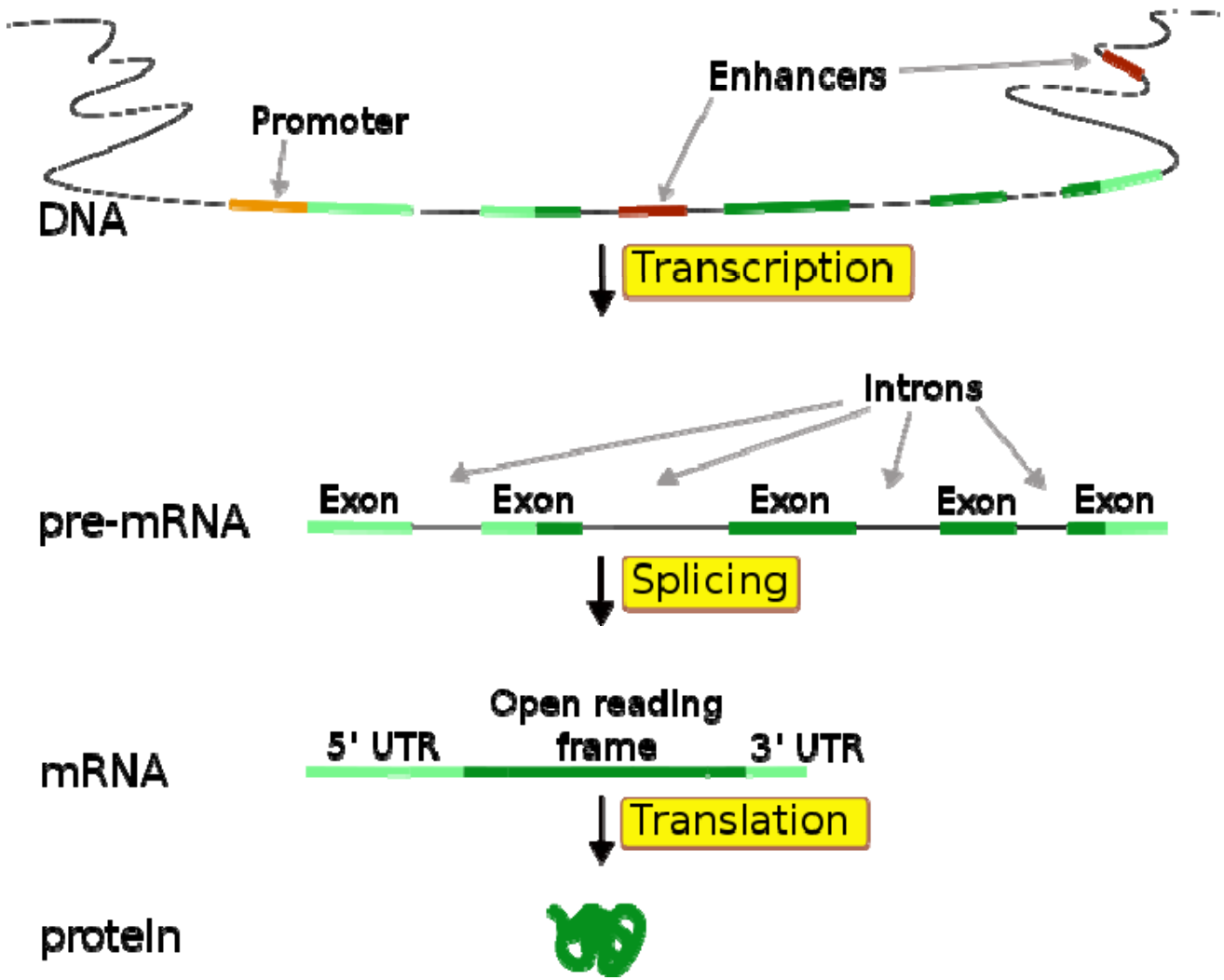
Allele

An *allele* is one variant of a given gene

Example: ABO blood type

- Three possible alleles: A, B, O
- 6 possible combinations of alleles
- 6 *genotypes* and 4 *phenotypes*
  - AA,AO
  - BB,BO
  - AB
  - OO

Gene structure and splicing



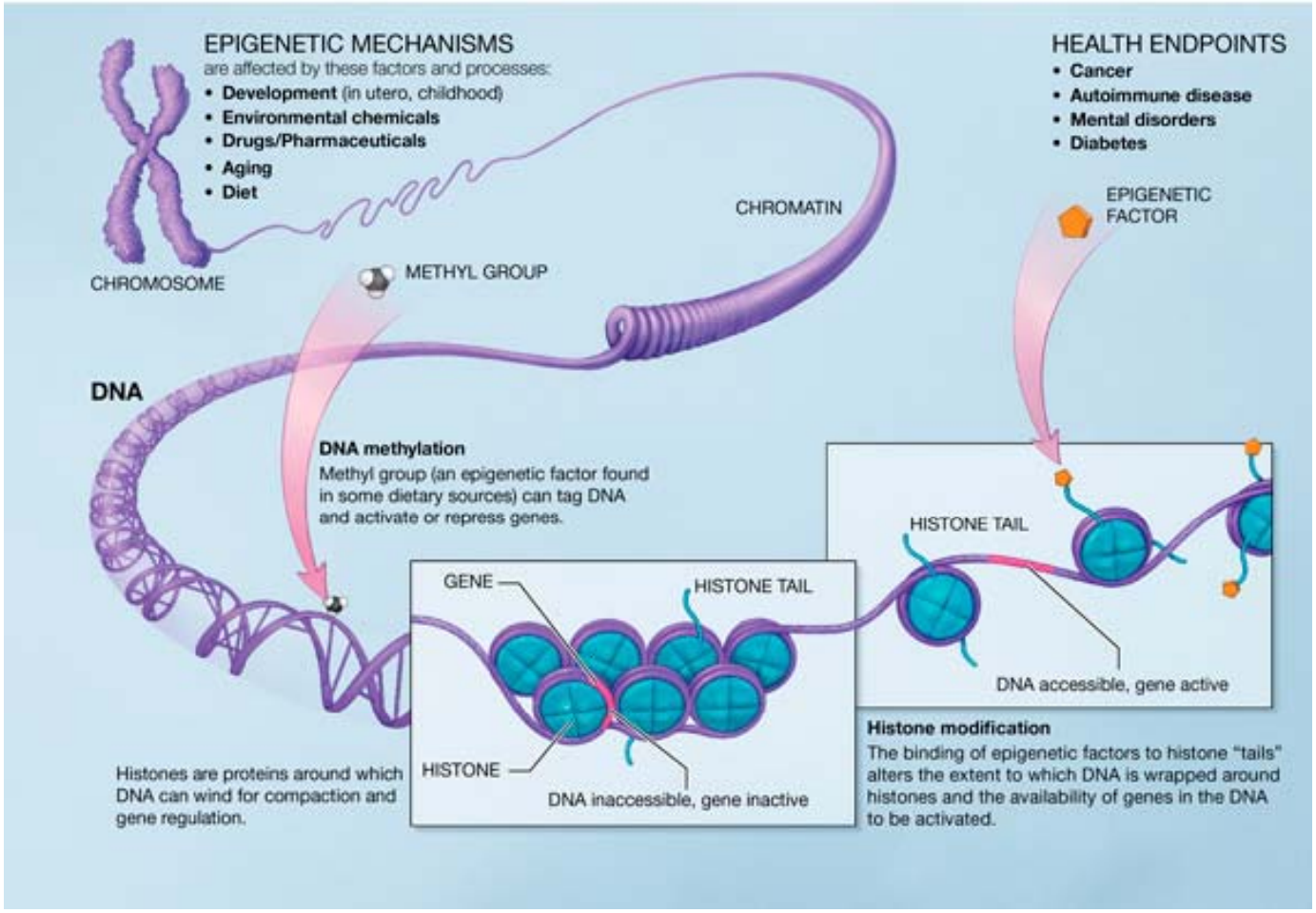
- In a nutshell (if you are not a biologist):
  - enhancer: a region for enhancing gene expression. Not necessarily closes to the gene. Each gene may have 0 or more than 1.
  - promoter: at the beginning of the gene, helps transcription. Each gene has 1.
  - exons: the coding sequence, will appear in the mature mRNA product (think about your actual code that will be run)
  - introns: the “spacer” between exons, will NOT be in the mRNA product (think about the comments after # in your code)
  - splicing: the process to remove introns and join exons.
  - alternative splicing: different splicing pattern for the same pre-mRNA. For example, mRNA could be from exons 1 and 2 or exons 1 and 3. Those are different “transcript” of the same gene.

RNA (Ribonucleic acid)

- Similar to DNA, but
  - RNA is *usually* single-stranded.
  - 4 letter alphabet GCAU, G-C, A-U pairs
  - The backbone is different from DNA.
- Many different types:
  - messenger RNA (mRNA): Coding RNA
  - Non-coding: ncRNA
    - tRNA (transfer) and rRNA (ribosomal)
    - Other Small RNAs: microRNA (miRNA), small interfering RNA (siRNA)…

Epigenetics

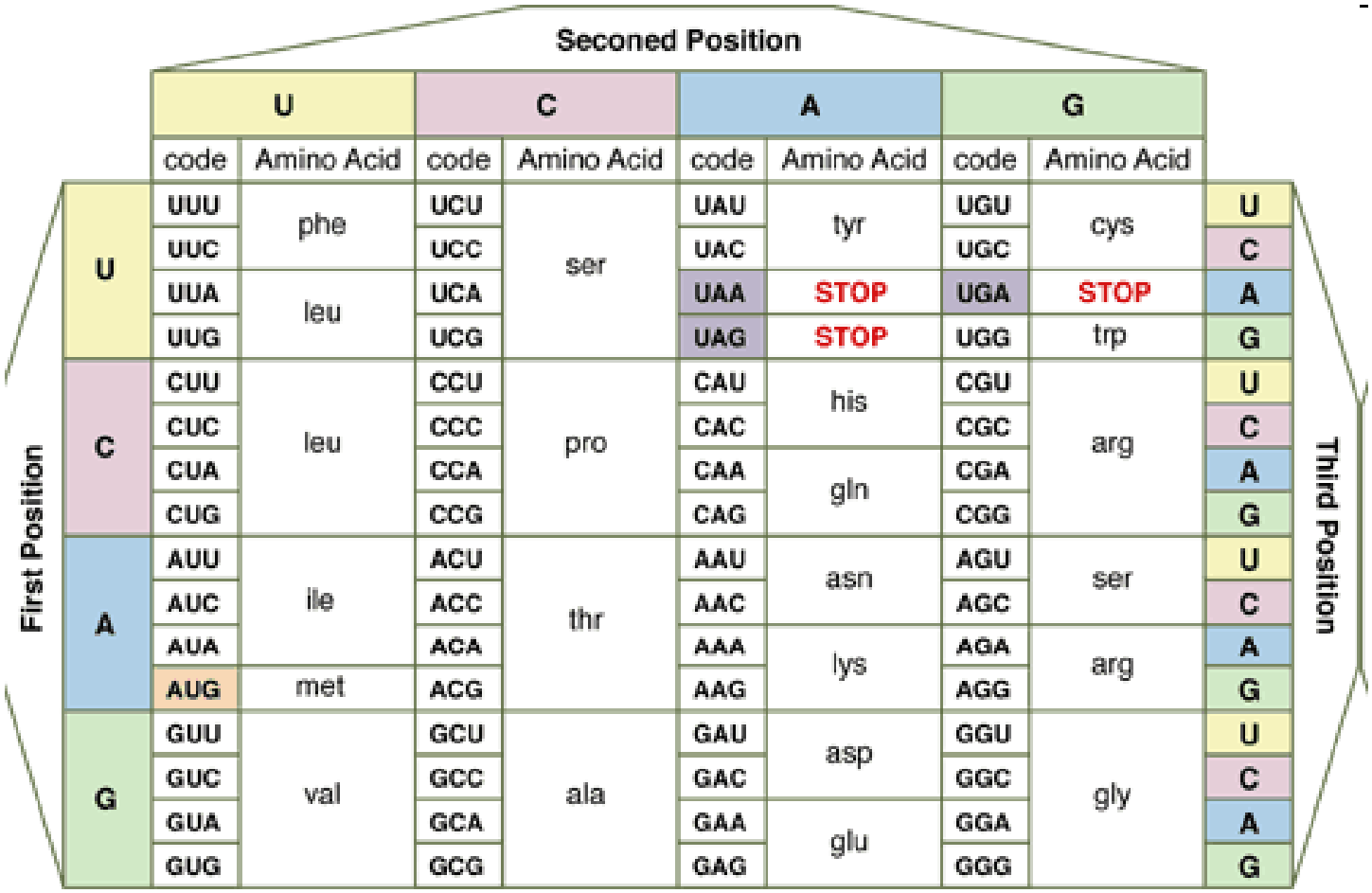
- **Non DNA** sequence related **heritable** mechanisms to control gene expressions. Examples: DNA methylation, histone modifications.



Protein

- The final product of gene expression process. The molecules that carry out most of the functions in a cell.
- 20-letter alphabet (amino acids).
- 3nt Codons mapping to AAs:
- 3D protein structure is often important for its function.

RNA codons to amino acids



First Position	U				C				A				G				Third Position
	code	Amino Acid			code	Amino Acid			code	Amino Acid			code	Amino Acid			
U	UUU	phe			UCU	ser			UAU	tyr			UGU	cys			U
	UUC				UCC				UAC				UGC				C
	UUA	leu			UCA				UAA	STOP			UGA	STOP			A
	UUG				UCG				UAG	STOP			UGG	trp			G
C	CUU	leu			CCU	pro			CAU	his			CGU	arg			U
	CUC				CCC				CAC				CGC				C
	CUA				CCA				CAA	gln			CGA				A
	CUG				CCG				CAG				CGG				G
A	AUU	ile			ACU	thr			AAU	asn			AGU	ser			U
	AUC				ACC				AAC				AGC				C
	AUA				ACA				AAA	lys			AGA	arg			A
	AUG	met			ACG				AAG				AGG				G
G	GUU	val			GCU	ala			GAU	asp			GGU	gly			U
	GUC				GCC				GAC				GGC				C
	GUA				GCA				GAA	glu			GGA				A
	GUG				GCG				GAG				GGG				G

Bioinformatics research topics

- What:
  - Sequencing a genome, comparative genomics
  - Identify genetic variations
  - Sequence a gene, studying different elements
  - Protein structure
- Who, when, where, how much
  - Genotypes and phenotypes
  - Copy numbers,
  - Gene expression in development, in different tissues, in response to disease or treatment
  - Protein levels in development, in different tissues, in response to disease or treatment
- How (regulation, interaction)
  - Gene x gene, protein x DNA, gene x environment ….

Some questions of interest about the genome

- What is the genome sequence: sequencing and assembly
- Can differences in our genomes explain the heterogeneity in a population?
  - Are the presence of certain mutations responsible for a disease?
  - Are certain polymorphisms of gene associated with a disease (disease markers)
- What do various parts of the genome do?
  - genes
  - Regulatory elements
  - Repeated sequences
- Evolution and comparison of genomes

Example questions about a transcriptome

- Which genes are expressed in specific tissues? What makes a neuron different from a muscle cell?
- When is a gene expressed? –genes involved in cell cycle, development, response to environmental change, signal transduction…
- What is going wrong?– what genes are associated with a disease? These may include the disease causing genes or genes that respond to a diseased state
- Are there patterns of gene expression associated with phenotype?

Some examples of basic techniques in molecular biology

- PCR: amplification
- Gel electrophoresis : separate molecules of different sizes
- Sequencing:
- Hybridization:
- Labeling: fluorescence or radioactivity
- Restriction enzyme digestion
- Protein binding DNA/RNA, immunoprecipitation

Where can noise arise in data acquisition?

Something to do before the first lab

- Install R
  - http://www.r-project.org/
  - From the menu, click “Download, packages- CRAN”
  - Select a mirror (at the bottom of the page, somewhere in USA would be a good choice.)
  - Download and Install R on your computer
  - It is possible to install R on your thumb drive if you want to use R on more than one computer, or in a computer lab that does not allow you to install software.
  - Try use R as a calculator by typing in the “R console” window
  - Download the Rintro files from the course website and try it out

A brief introduction to High-throughput experiments

High-throughput experiments

- Methods to conduct a large number of experiments simultaneously.
- Examples:
  - Microarrays.
  - Second generation sequencing.
  - Flow cytometry
  - …
- Pros: quick, cheap (per unit).
- cons: noisy, complicated data.

Microarray: multiplex lab-on-a-chip

- 2D array on a solid substrate that assays large amount of biological materials.
- Examples of microarrays:
  - DNA microarray:
    - Gene expression array.
    - SNP array.
    - Tiling arrays (ChIP-chip, array CGH).
    - Methylation array.
  - Protein microarray
  - Tissue arrays
  - Others …

DNA microarrays

- A collection of probes (short segments of DNAs).
  - Search phrases
- Detect and quantify target sequence (e.g., mRNA) by hybridization: sequence-specific interaction between two complementary strands of nucleic acid.
  - An exmple:

    `ATCGATTGAGCTCTAGCG`

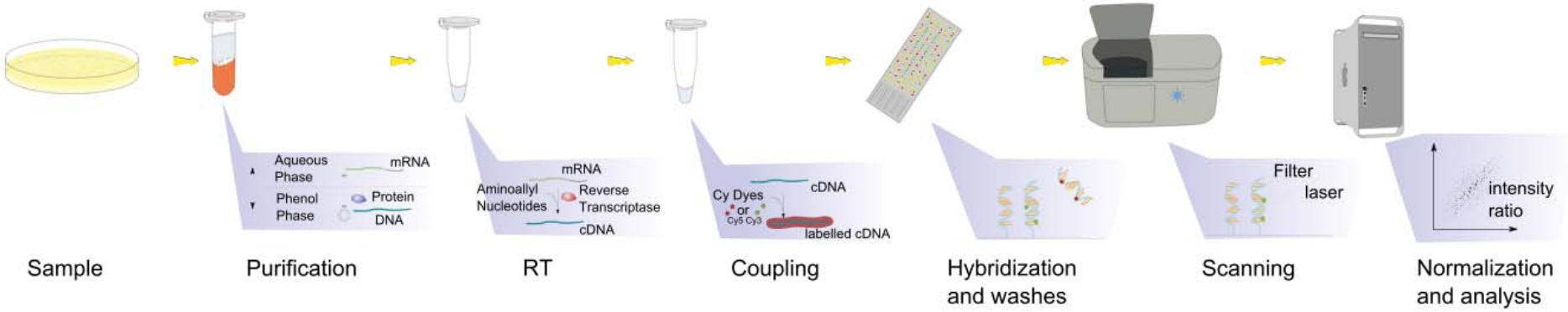
    `TAGCTAACTCGAGATCGC`

- Sort the labeled fragments in a sample (a complex collection of DNA fragments)
- the intensity on each probe represents the amount of matches each probe found (often with misspelling allowed)
-

Gene expression microarray

- Measure the gene expressions by the amount of mRNA.
- Each gene is targeted by many probes.
- Major manufactures:
  - Affymetrix
  - Illumina
  - Nimblegene

GE microarray procedures



- Data: a fluorescent intensity value (a non-negative floating-point number) for each probe.
- Goal:
  - find genes that are differentially expressed (produce different amount of mRNAs) among samples.
  - Identify genetic or epigenetic variations

Statistical challenges

- Data preprocessing
  - From raw measurements to the quantity of interest normalization/transformation/summarization.
- Statistical inferences: tests for DE (differentially expressed) genes.
- Pattern recognition, e.g., clusterings.
- Biological/clinical implications.

Tiling array

- DNA microarray with probes “tiling up” whole or specific regions of the genome.
- Used for genome-wide locational analysis: detect the locations of a specific modification of the genome.
- Examples:
  - ChIP-chip: Chromatin Immunoprecipitation (ChIP) followed by microarray (chip), for transcription factor binding sites (TFBS) detection.
  - arrayCGH: comparative genome hybridization, for copy number variation (CNV) detection.

Tiling array – ChIP-chip

- Chromatin Immunoprecipitation (ChIP) followed by microarray (chip).
- Goal: detect the locations of specific modifications of the genome, e.g., transcription factor binding sites (TFBS), histone modifications, etc.
- Also referred to as GWLA (genome-wide locational analysis).

DNA sequencing

- Technologies to determine the nucleotide bases from a DNA molecule.
- Traditional method: Sanger sequencing.
  - slow (low throughput) and expensive: took Human Genome Project (HGP) 13 years and $3 billion to sequence the entire human genome.
  - Relatively accurate.

Second generation sequencing

- Aka: high-throughput sequencing, next generation sequencing.
- Able to sequence large amount of short sequence reads in a short period:
  - high throughput: tens of millions sequences in a run.
  - Cheap: sequence entire human genome costs a few thousand dollars.
  - short read length: up to several hundred bps.

Applications

- DNA seq: sequence the genomic DNA in order to find variants or assemble reference genome.
- RNA seq: sequence the transcriptome (mRNA -> cDNA) in order to measure gene expressions or detect alternative splicing/gene fusion.
- MeDIP/ChIP-seq: detection protein-DNA binding or epegenitic modification sites. .

Available platforms

- Major player:
  - Illumina: HiSeq, IonTorrent
  - LifeTech SOLiD
  - Roche 454
- Others:
  - Complete Genomics
  - Pacific Bioscience

R and Bioconductor

R programing language

- Visit www.r-project.org to download/install R and reference manuals.
- I recommend using emacs with ESS (emacs speaks statistics), see http://www.biostat.wisc.edu/~kbroman/Rintro/ for details.
- Alternatively, there are easier environment such as Tinn-R and Revolution-R.

Bioconductor: a collection of R packages

- Provide many packages for genomics
- Installation: visit http://bioconductor.org
  - Basic installation: installing default (core) packages:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

  - Installing a specific package:

```
source("http://bioconductor.org/biocLite.R")
biocLite("limma")
```

Bioconductor installation

- Use biocLite.R script.
- Basic installation: installing default (core) packages:

```
source("http://bioconductor.org/biocLite.R")
biocLite("")
```

- Installing a specific package:

```
source("http://bioconductor.org/biocLite.R")
biocLite("limma")
```

Online resources: genome browser and public data repositories

UCSC Genome Browser

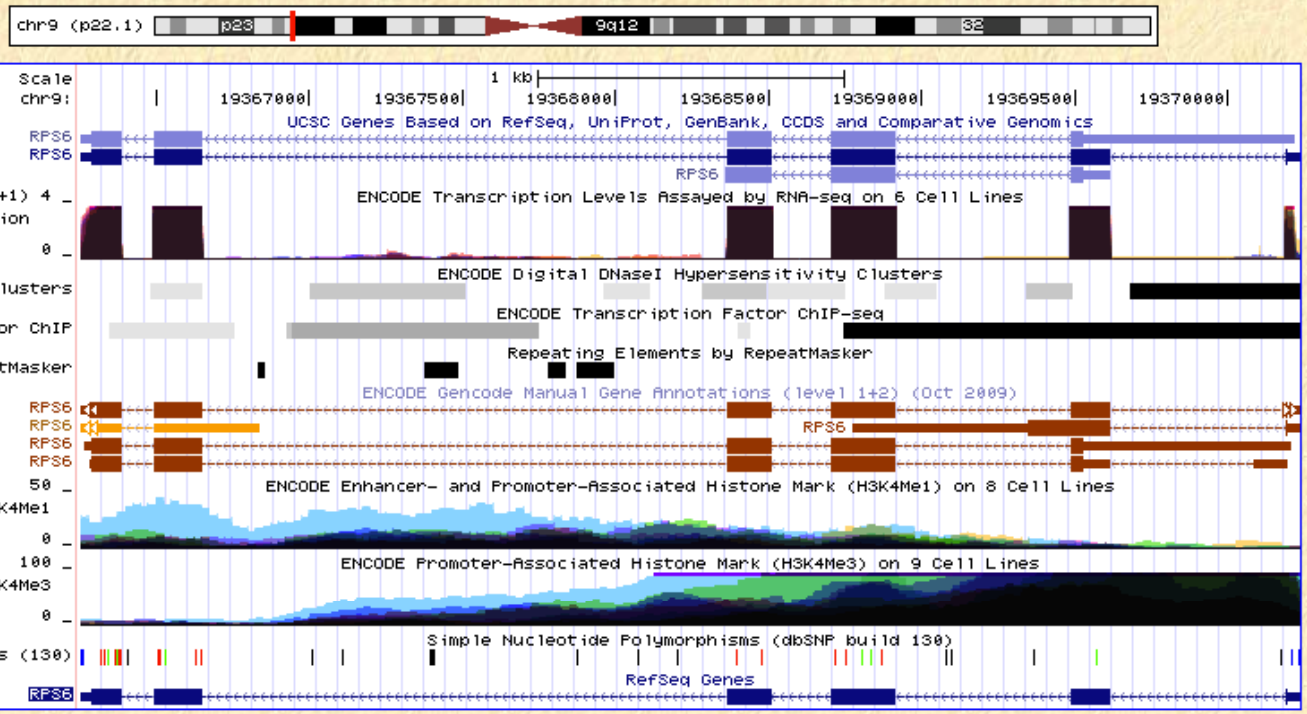
- Initially developed by Jim Kent on 2000 when he was a Ph.D. student in Biology.
- Host genomic annotation data for many species.
- The genome browser is a graphical viewer for visualizing genome annotations.
- Provide other tools for genomic data analysis and interfaces for querying the database.

Home | Genomes | Blat | Tables | Gene Sorter | PCR | DNA | Convert | Ensembl | NCBI | PDF/PS

UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr9:19,366,254-19,370,235 gene jump clear size 3,982 bp. configure



move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

move end < 2.0 >

Other genome browsers/databases

- General:
  - NCBI Map Viewer
  - Ensemble genome browser
- Other species specific genome browser
  - MGI: Mouse genome informatics
  - wormbase, Flybase, SGD (yeast), TAIR DB (arabidopsis), microbial genome database
- More or less the same, pick your favorite one.

Public high-throughput data repositories

- GEO: Gene expression omnibus.
  - Funded by NIG, part of NCBI.
  - Host array- and sequencing-based data.
- ArrayExpress: European version of GEO.
  - Better curated than GEO but has less data.
- SRA: sequence read archive.
  - Designed for hosting large scale high-throughput sequencing data, e.g., high speed file transfer.

Data are required to be deposited in one of the databases when paper is accepted!

Other public data resources

- TCGA (The Cancer Genome Atlas) data portal:
  - Host data generated by TCGA, a big consortium to study cancer genomics.
  - Huge collection of cancer related data: different types of genomic, genetic and clinical data for many different types of cancers.
- ENCODE (the **ENC**yclopedia **O**f **D**NA **E**lements) data coordination center:
  - Host data generated by ENCODE, a big consortium to study functional elements of human genome.
  - Rich collection of genomic and epigenomic data.
- Many others …

To do list after this class

1. Review slides.
2. Understand every highlighted term. Google and wikipedia are your friends for basic concepts.
3. Answer the exercise questions. Give a source for your answer.
4. Install R and Bioconductor on your computer.
5. Start to learn R by reading “R for beginners”: http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

Exercise

- How big is the human genome?
- How many genes are in the human genome?
- What genomes have been sequenced?
- What fraction of the human genome codes for protein?
- Which human gene is the largest?
- Which species has the largest genome?
- Which species has the genome most similar to human genome? How similar?

What do these terms refer to:

- Polymerase
- Transcription Factor
- Promoter
- cDNA
- Histone
- Homology and homologous genes
- Polymorphism
- Single nucleotide polymorphism (SNP)
- Copy number variation (CNV)

Exercise

- Suppose a genome's sequence was completely random, and any base (ATGC) is equally likely to appear at any position. IF we randomly select a 5-base window, what is the probability that you see
  - GGGGG
  - AAAAA
  - GAGAG
- How many examples of *GGGGG* do you expect to see in a region of size 1MB?
- Let X represent the number of G/C in a short window of 10. What random distribution would you use to model the random variable X?
- Suppose the genome sequence is random but the nucleotide frequency is not equal. Say the proportion of G is only 21%. How would you modify your model?

Exercise

- Consider a locus with two alleles, A (major allele) and a (minor allele).
- Suppose in the population the major allele frequency is 90% and the minor allele frequency is 10%. Under the assumption of random mating and no evolutionary influence (no mutation, selection etc), what are the proportion of the genotypes *AA,Aa* and *aa* in the population?
- In general, let *p* represent the Major Allele Frequency, what are the the proportion of the genotypes *AA,Aa* and *aa* in the population?