Bio5488
Genomics
Spring, 2017

Lectures: Mon, Wed 10:00-11:30 am
Lab: Fri 10:00-11:30 am

4th floor classroom
4515 MSRB

Outline of the day

- Outline of the course
- What is genomics?
- A little history
- The simple principles of genomics
- Being quantitative
- From a student to an investigator

A few TA administrivia...

- If you didn't receive an email from bio5488wustl@gmail.com this week, please email bio5488wustl@gmail.com or talk to a TA after class

- If you're taking the lab:
  - Please fill out the anonymous survey
  - Read assignment 1
  - Attempt to install the required software
  - Bring your laptop to class on Friday

Course Web Site

- http://www.genetics.wustl.edu/bio5488/
- Linux Primer
- Python Primer
- Lecture notes
- Schedule
- Weekly Assignments and Answers
- Weekly Readings

Grading

4 credit

- ¼ midterm
- ¼ final
- ½ weekly assignments

3 credit

- ½ midterm
- ½ final

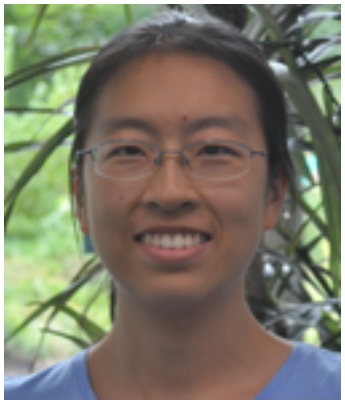
What is the key to your success?

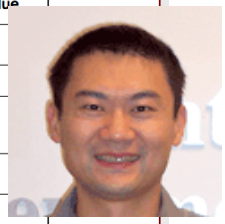
Washington University
Genetics – Bio5488

HOME | SCHEDULE 2017

Schedule 2017

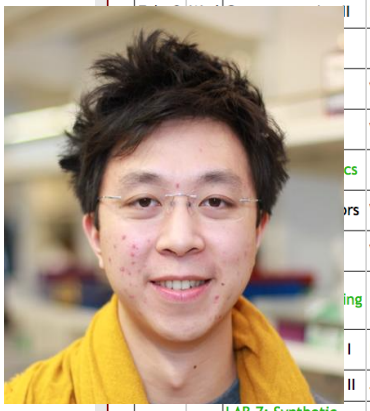
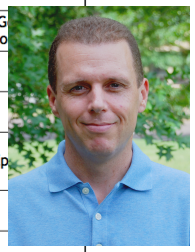
Date	Day	Lecture/Lab	Lecturer	Notes (if available)	Assignment due
		Martin Luther King			
			Wang/Conrad		
			TAs		
			Wang		
			Wang		
					LAB 1: Introduction to Python Programming
			Mitra		
			Mitra		
					LAB 2: Compa
			Lawson		
			Lawson		
					LAB 3: Sequen
			Wang		
			Wang		
					LAB 4: G Expressio
			Wang		
			Wang		
					LAB 5: Ep
			Dantas		
			Jim		
Mar. 3	Fri	LAB 7: Synthetic Gene Assembly Lab 8: Metagenomics			LAB 6: Motif Finding
Mar. 6	Mon	Metagenomics I	Dantas		
Mar. 8	Wed	Metagenomics II	Dantas		
Mar. 10	Fri	MIDTERM EXAM			
Mar. 13 -17	Mon	SPRING BREAK			
Mar. 20	Mon	Genetic variation I	Hall		
Mar. 22	Wed	Genetic variation II	Hall		LAB 7: Synthetic Gene Assembly Lab 8: Metagenomics

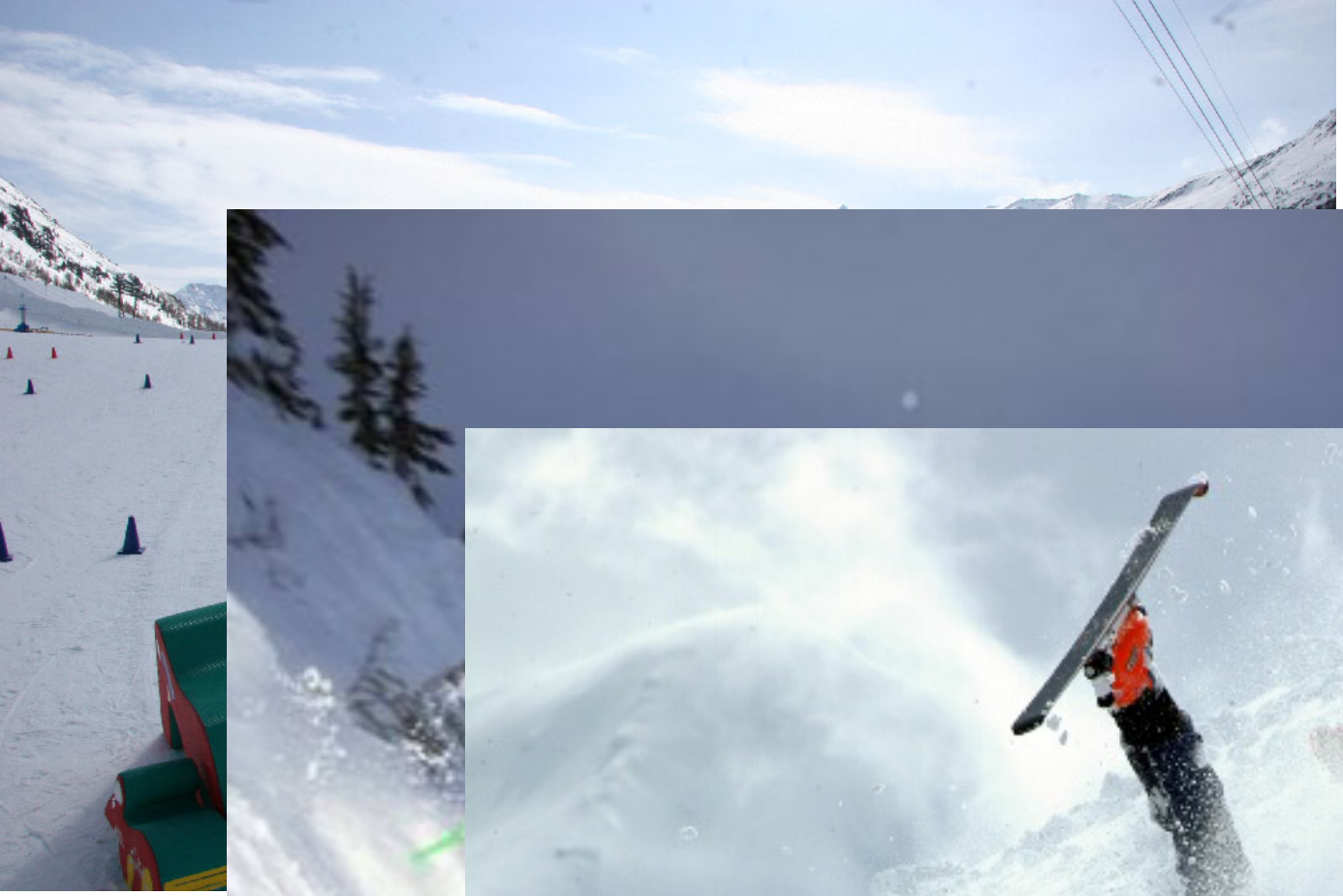




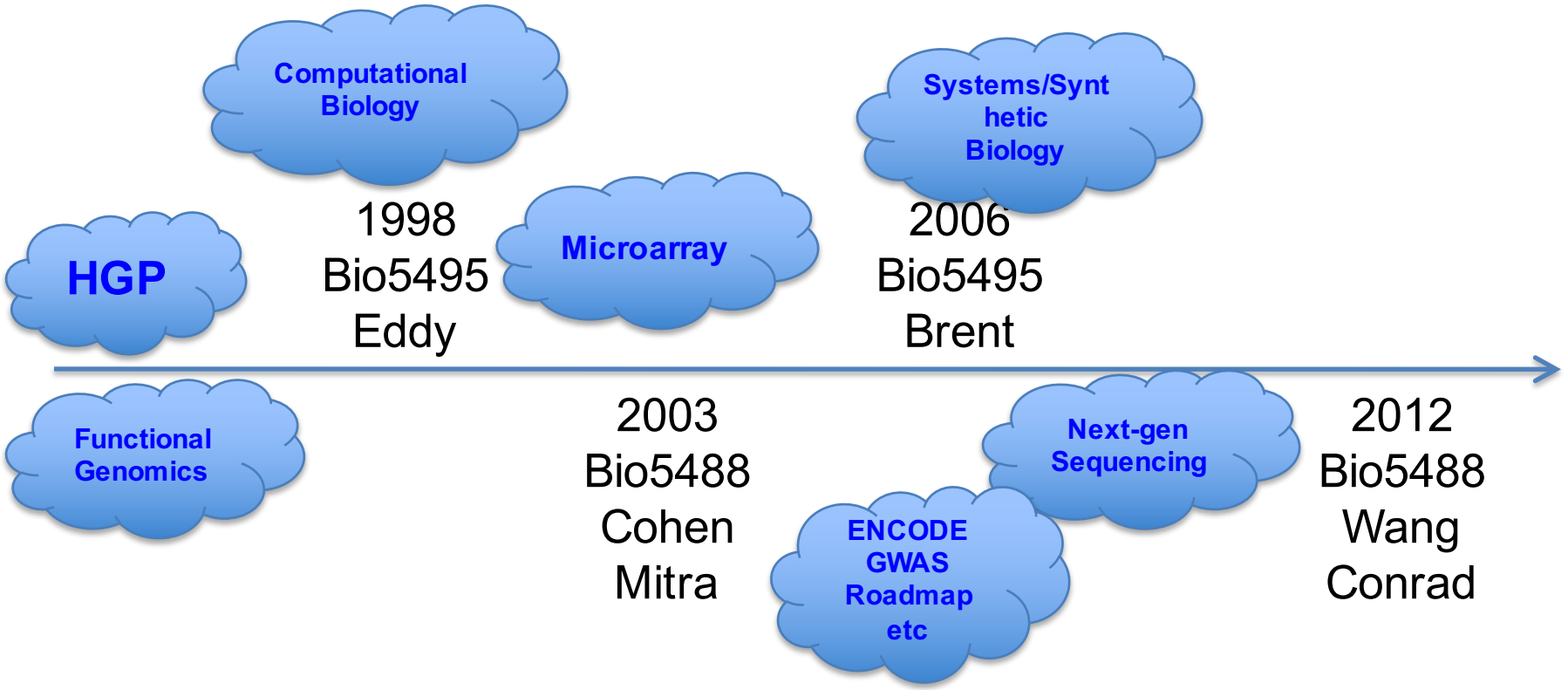






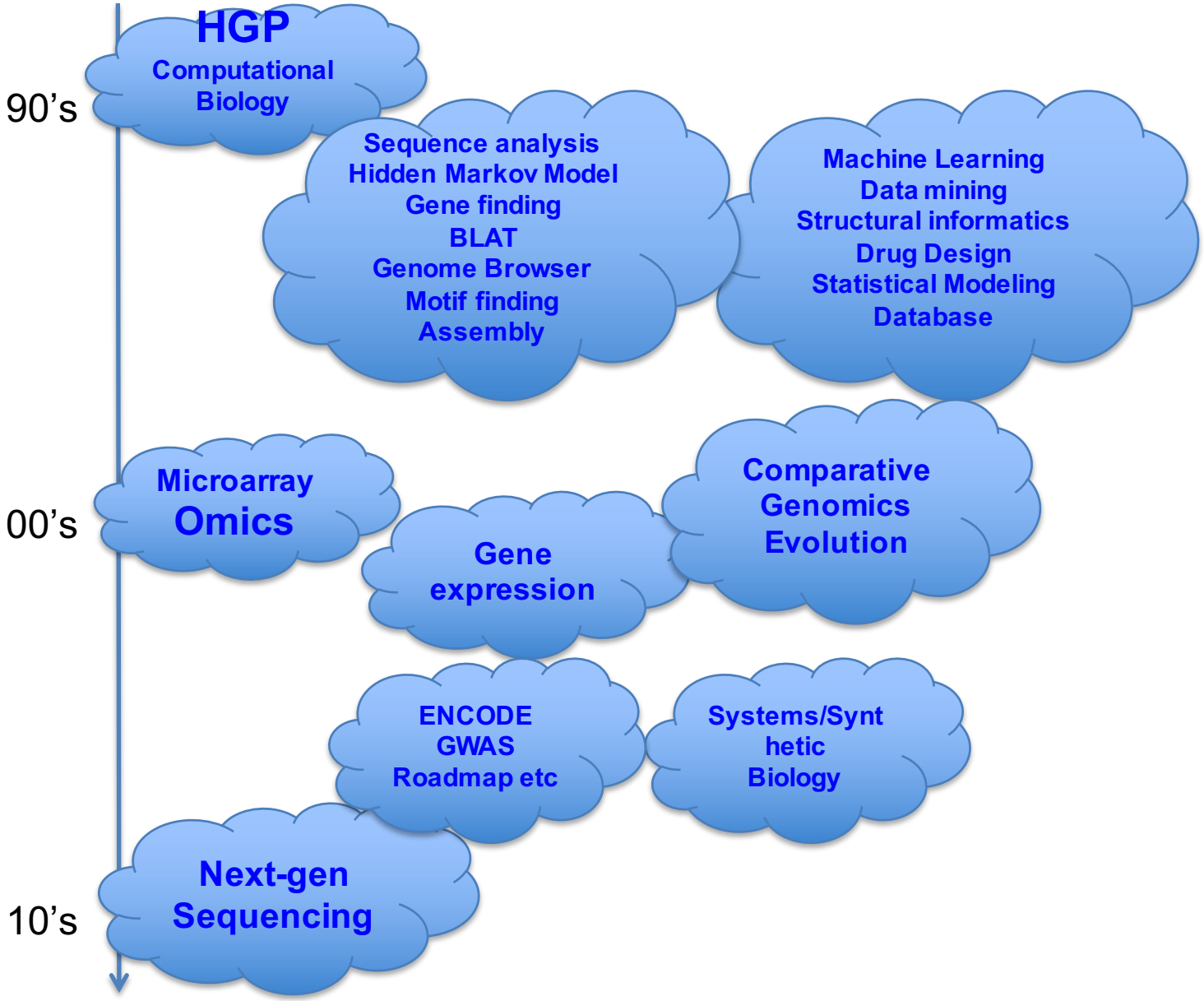
History of Bio5488



History of Genomics and Epigenomics

1865	Gregor Mendel: founding of genetics
1953	Watson and Crick: double helix model for DNA
1955	Sanger: first protein sequence, bovine insulin
1970	Needleman-Wunsch algorithm for sequence alignment
1977	Sanger: DNA sequencing
1978	The term “bioinformatics” appeared for the first time
1980	The first complete gene sequence (Bacteriophage FX174), 5386 bp
1981	Smith-Waterman algorithm for sequence alignment
1981	IBM: first Personal Computer
1983	Kary Mullis: PCR
1986	The term "Genomics" appeared for the first time: name of a journal
1986	The SWISS-PROT database is
1987	Perl (Practical Extraction Report Language) is released by Larry Wall.
1990	BLAST is published
1995	The *Haemophilus influenzea* genome (1.8 Mb) is sequenced
1996	Affymetrix produces the first commercial DNA chips
2001	A draft of the human genome (3,000 Mbp) is published

History of Genomics and Epigenomics



Genome, genetics, and genomics

- What is a genome?
  - **The genetic material of an organism.**
  - **A genome contains genes, regulatory elements, and other mysterious stuff.**

- What is genetics?
  - **The study of genes and their roles in inheritance.**

- What is genomics
  - **The study of all of a person's genes (the genome), including interactions of those genes with each other and with the person's environment.**
  - **Biology in big data era.**

The simple principles of genomics

- **Characterize the genome**
  - How big
  - How many genes
  - How are they organized
- **Annotate the genome**
  - What, where, and how
- **Modern genomics: “ChIPer” vs “Mapper”**
  - Direct measurement
  - Inference
  - Comparison
  - Evolution
- **From genome to molecular mechanisms to diseases**
  - Genomes/epigenomes of diseased cells
- **What do you want to learn from this class?**
  - Being quantitative
  - Concept/philosophy
  - Biology/technology/informatics
  - Problem solving skills
  - Do not forget genetics!!!

Motivation slides

Genomes sequenced

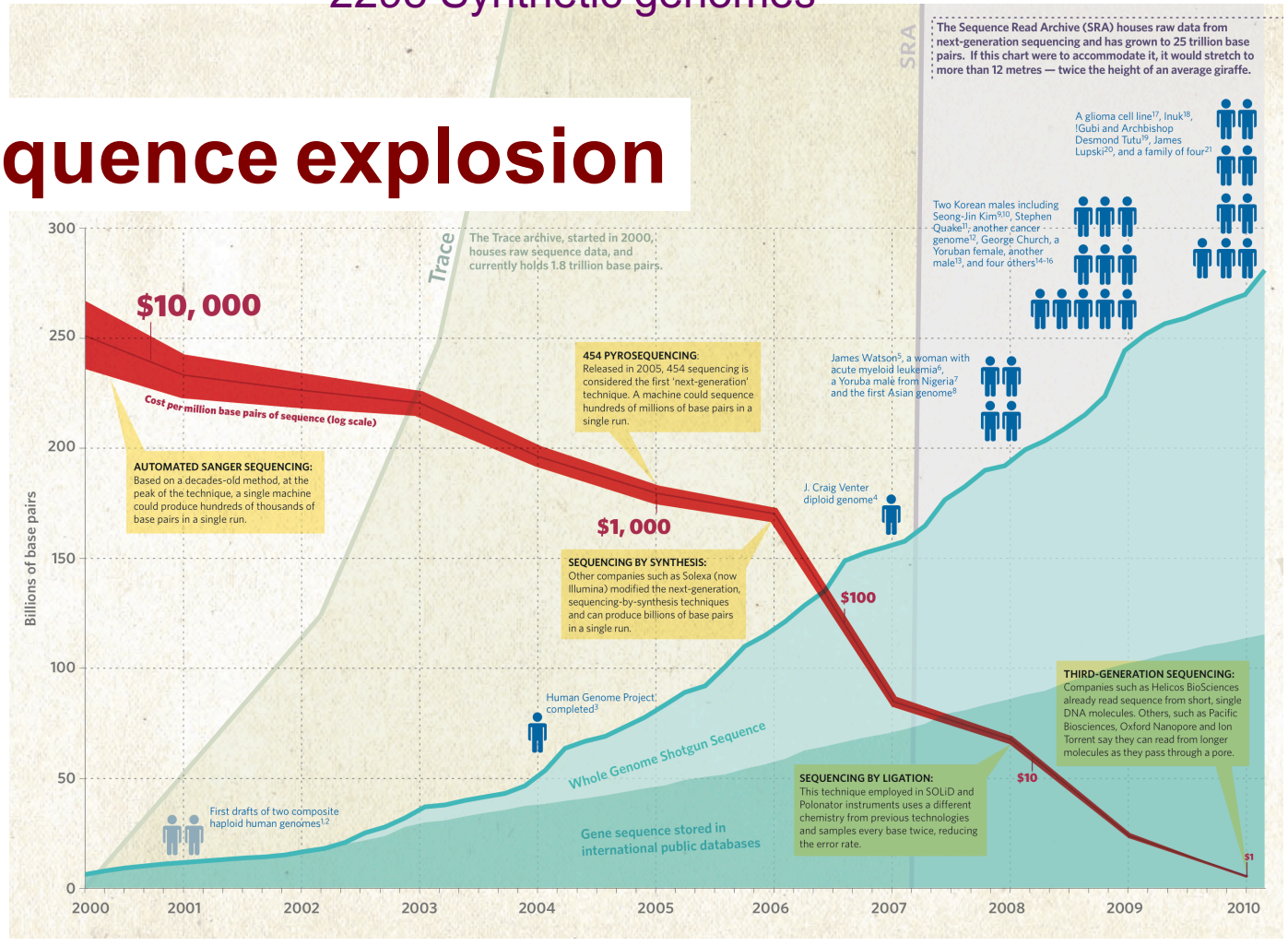
http://www.genomesonline.org/, (as of January 2017)

241,446 Bacteria, 2,142 Archaea, 14,737 Eukarya, 6,615 viruses

18385 metagenome samples

2298 Synthetic genomes

The sequence explosion



The Human Genome Project

The human genome completed! – 2001

June 26, 2000
President Clinton, with Craig Venter and Francis Collins, announces completion of "the first survey of the entire human genome."

February 15, 2001

The human genome completed, again! – 2010

April 1, 2010

10 years after draft sequence: What have we learned?

- Genome sequencing
- Functional elements
- Evolution of genome
- Basis of diseases
- Human history
- Computational biology

"If I was a senior in college or a first-year graduate student trying to figure out what area to work in, I would be a computational biologist."

-- During AAAS Meeting Jan 2010

COLLINS: .... Computational biologists are having a really good time and it's going to get better.

ROSE: Their day is coming?

COLLINS: Their day is here, but it's going to be even more here in a few years.

COLLINS: They're going to be the breakthrough artists ....

-- Mar 15, 2010, Charlie Rose Show

Francis Collins
NIH Director

Program Initiatives

$200 million

I. Facilitating Broad Use of Biomedical Big Data

- New Policies to Encourage Data & Software Sharing
- Catalog of Research Datasets to Facilitate Data Location & Citation
- Frameworks for the development of community-based standards
- Enabling Research Use of Clinical Data

II. Developing and Disseminating Analysis Methods and Software

- Software to Meet Needs of the Biomedical Research Community, both analytic software and management/processing software
- The creation of a Catalog of NIH-funded Software
- Facilitating Data Analysis: Access to Large-scale Computing

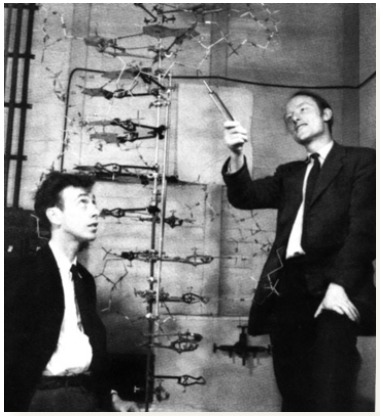
III. Enhancing Training for Biomedical Big Data

- Increase the Number of Computationally Skilled Biomedical Trainees
- Strengthen the Quantitative Skills of All Biomedical Researchers
- Enhance NIH Review and Program Oversight

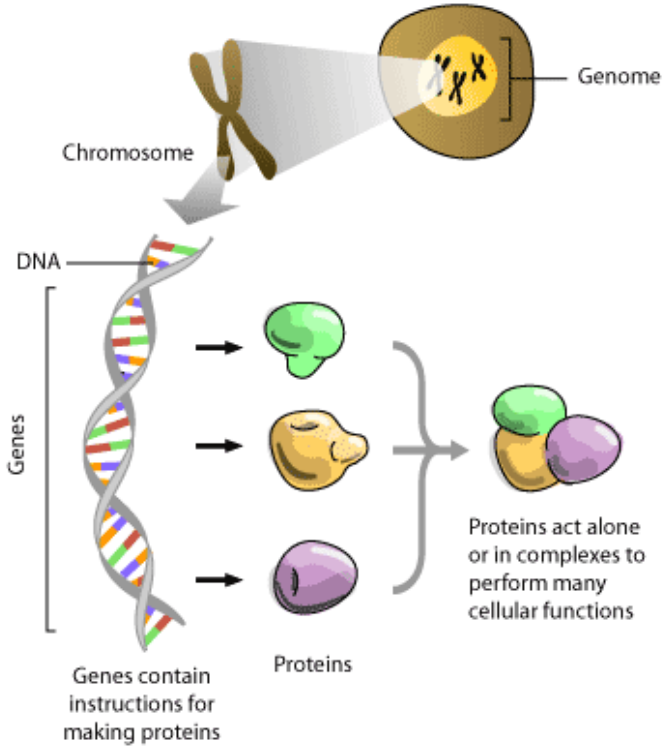
IV. Establishing Centers of Excellence for Biomedical Big Data

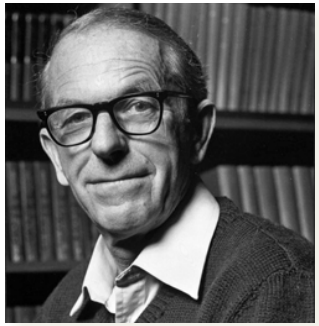
Advance the science of Big Data in the context of biomedical and behavioral research, and to create innovative new approaches, methods, software, and tools.

The Human genome: the "blueprint" of our body



James Watson
Francis Crick



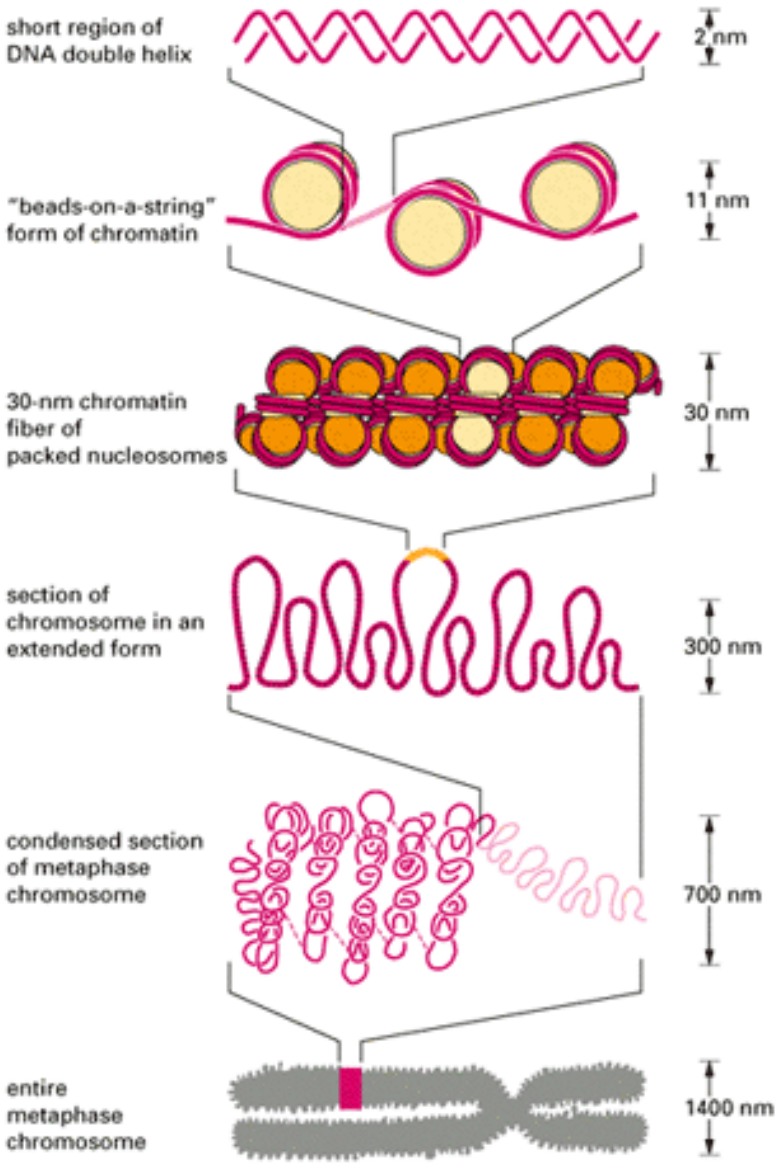


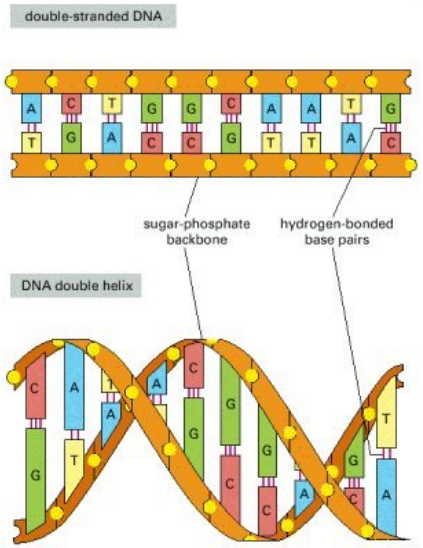
Fred Sanger

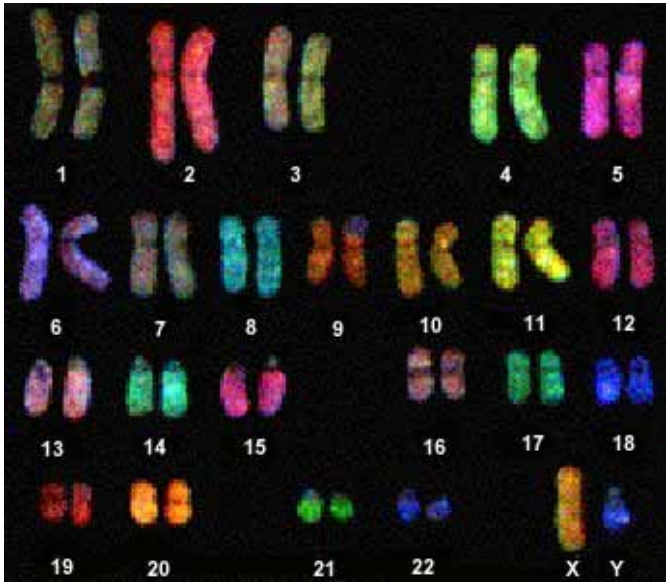
GTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCT
CCGAACAATAAAGATTCTACAATACTAGCTTTTATGGTTATG
AAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAA
ATTAACGAATCAAATTAACAACCATAGGATGATAATGCGATT
AGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG
ATGATTTTTGATCTATTAACAGATATATAAATGGAAAAGCTG
CATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTA
TTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAATT

- 10^{13} different cells in an adult human
- The cell is the basic unit of life
- DNA = linear molecule inside the cell that carries instructions needed throughout the cell's life ~ long string(s) over a small alphabet
- Alphabet of four (nucleotides/bases) {A,C,G,T}

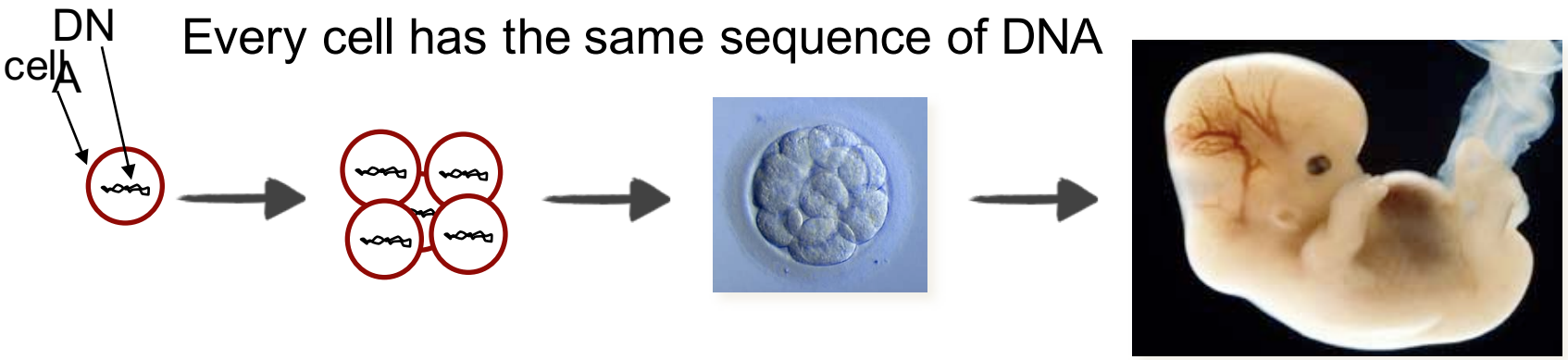
DNA, Chromosome, and Genome



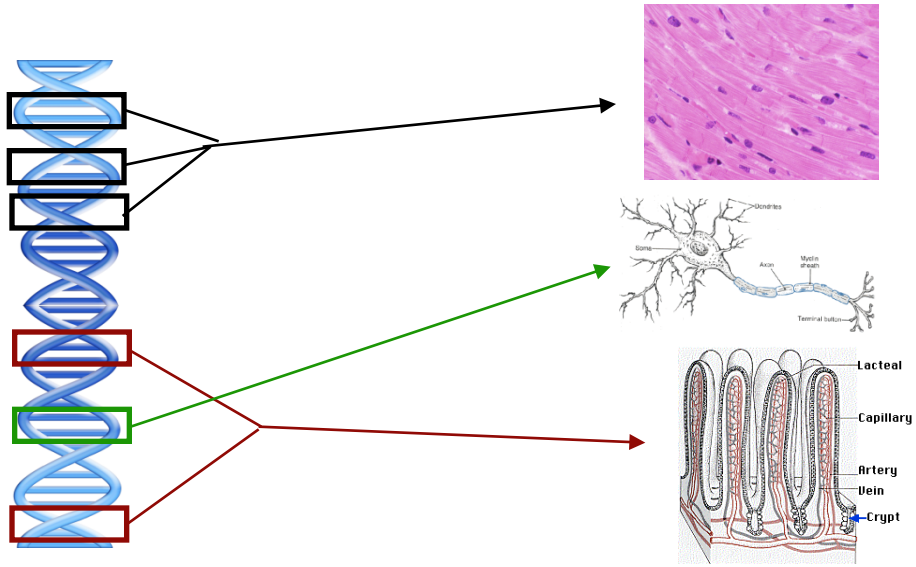




Building an Organism



Subsets of the DNA sequence determine the identity and function of different cells

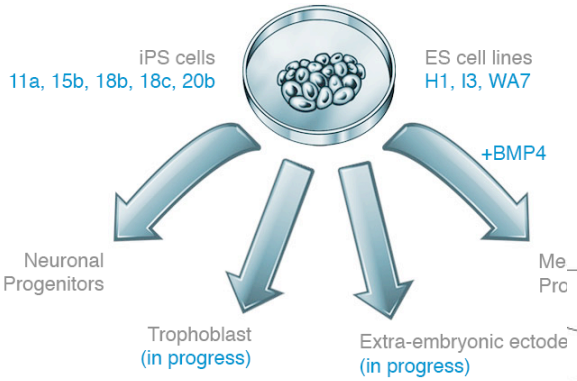


What makes us different?

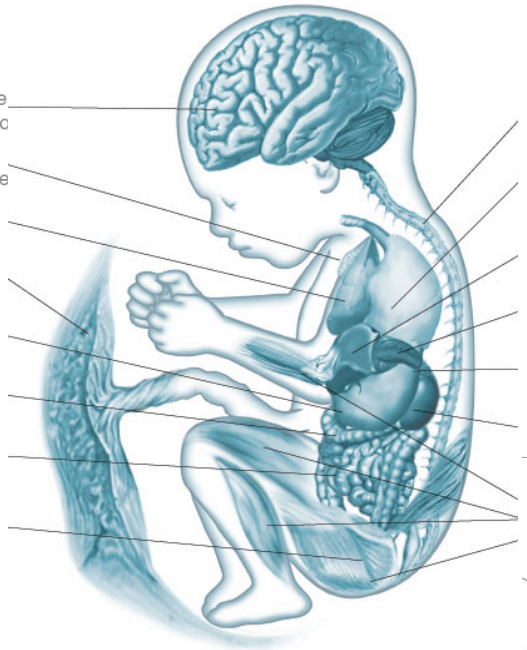
Differences between individuals?

Differences between species?

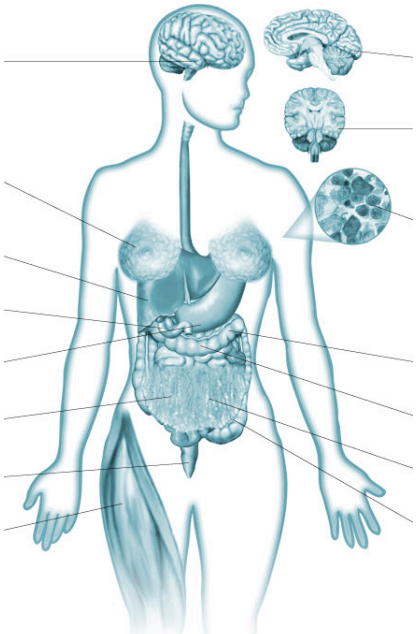
One genome, thousands of epigenomes



Embryonic stem cells



Fetal tissues



Adult cells and tissues

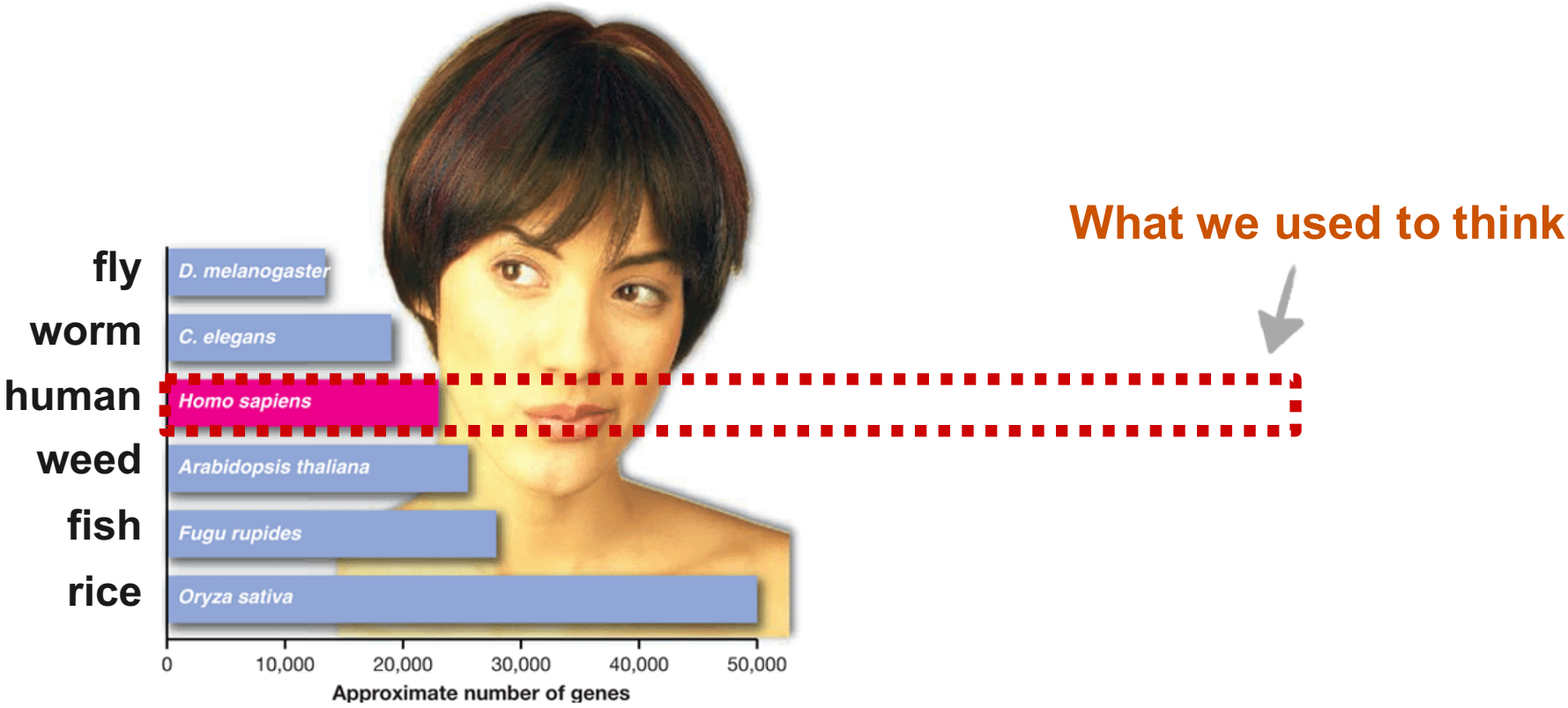
Understanding the genome

- View from 2000
  - Protein-coding genes 35,000 - 120,000
  - Regulatory sequence Less than protein-coding information
  - Transposons Junk DNA

All these are WRONG!

- We now know ...

How many genes do we have?



Science 2005

Gene numbers do not correlate with organism complexity. Many gene families are surprisingly old.

Complexity, Genome Size and the C-value Paradox

Organism	Genome Size (MB)
Amoeba	670,000
Fern	160,000
Salamander	81,300
Onion	18,000
Paramecium	8,600
Toad	6,900
Barley	5,000
Chimp	3,600
Gorilla	3,500
Human	**3,500**
Mouse	3,400
Dog	3,300
Pig	3,100
Rat	3,000
Boa Constrictor	2,100
Zebrafish	1,900
Chicken	1,200
Fruit fly	180
C. elegans	100
Plasmodium falciparum	25
Yeast, Fission	14
Yeast, Baker's	12
Escherichia coli	4.6
Bacillus subtilis	4.2
H. influenzae	1.8
Mycoplasma genitalium	0.60

www.genomesize.com

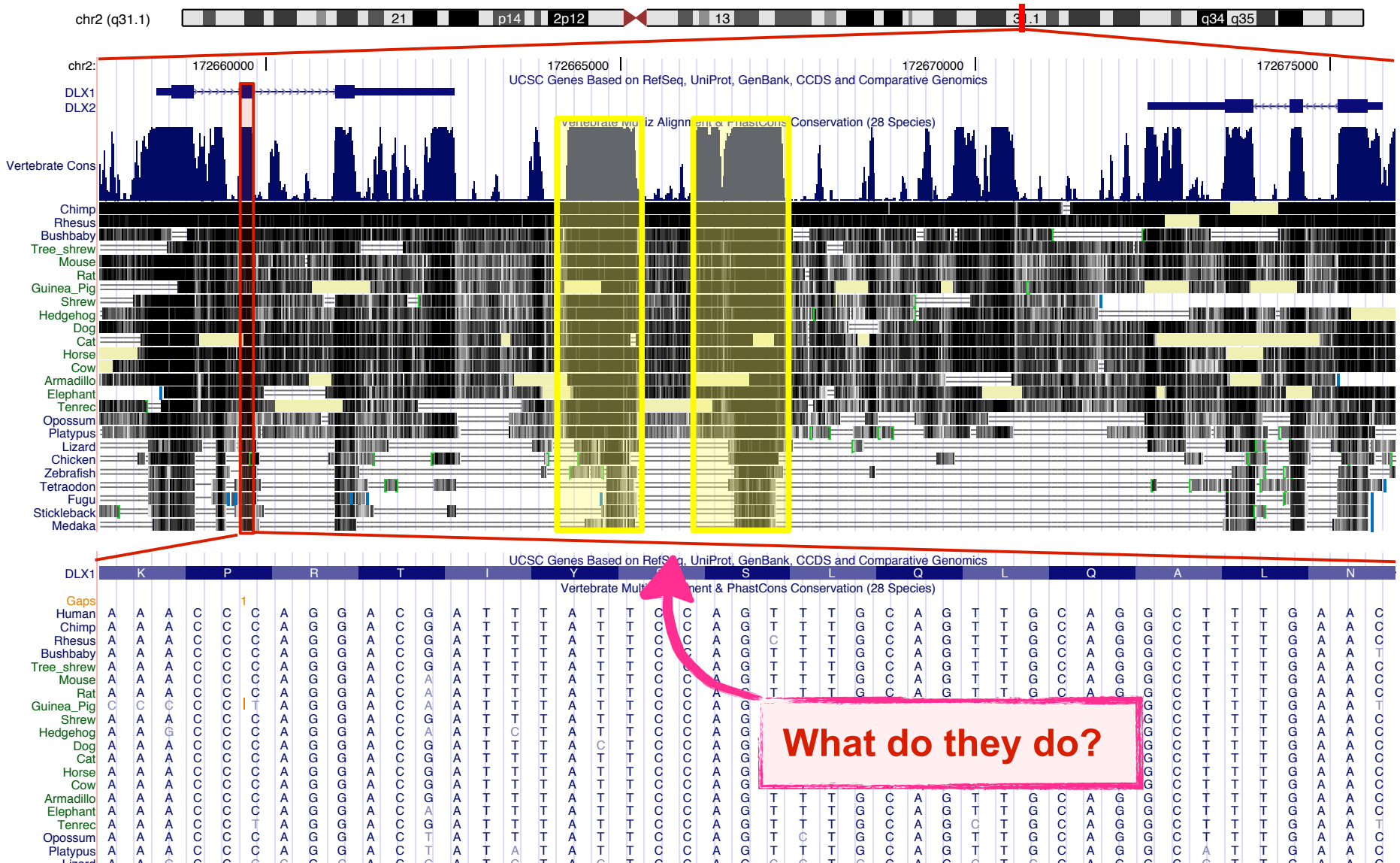
C-value: the amount of DNA contained within a haploid nucleus (e.g. a gamete) or one half the amount in a diploid somatic cell of a eukaryotic organism, expressed in picograms (1pg = 10^{-12} g).

Complexity and Organism Specific Genes

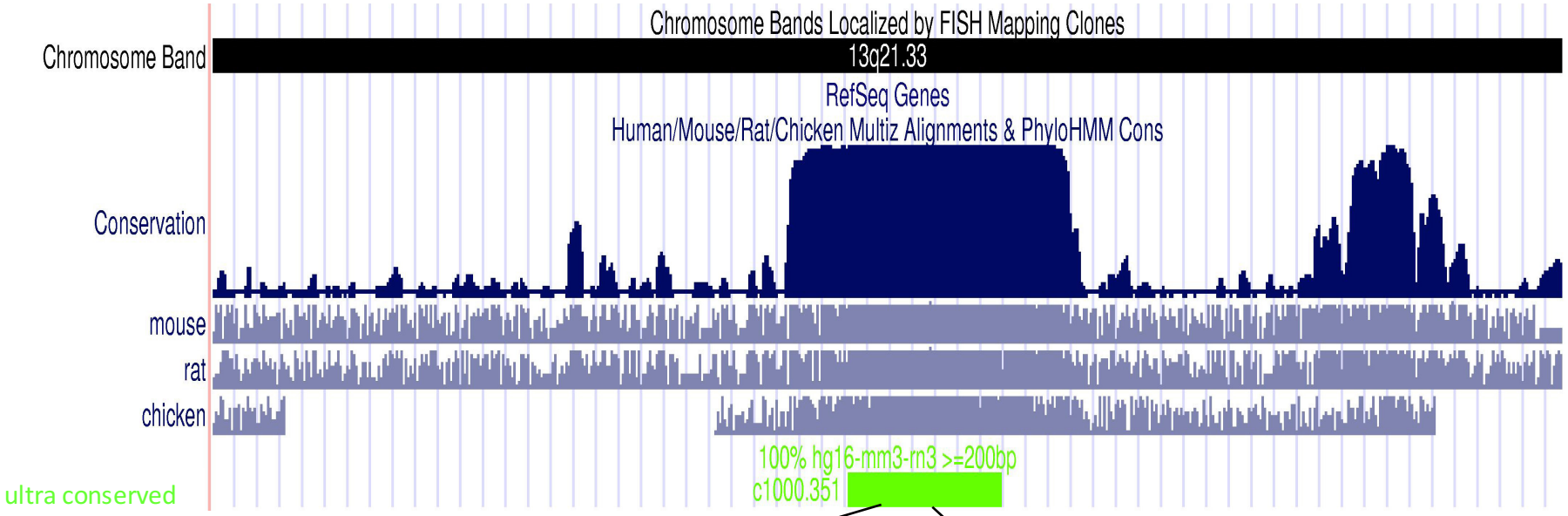
- Only 14 out of 731 genes on mouse chromosome 16 have no human homolog (Celera)
- Many genes specific to the mouse are olfactory receptors, also some differences in immunity and reproduction

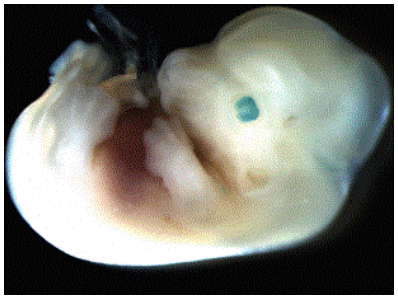
Most functional information is non-coding

- 5% highly conserved, but only 1.5% encodes proteins



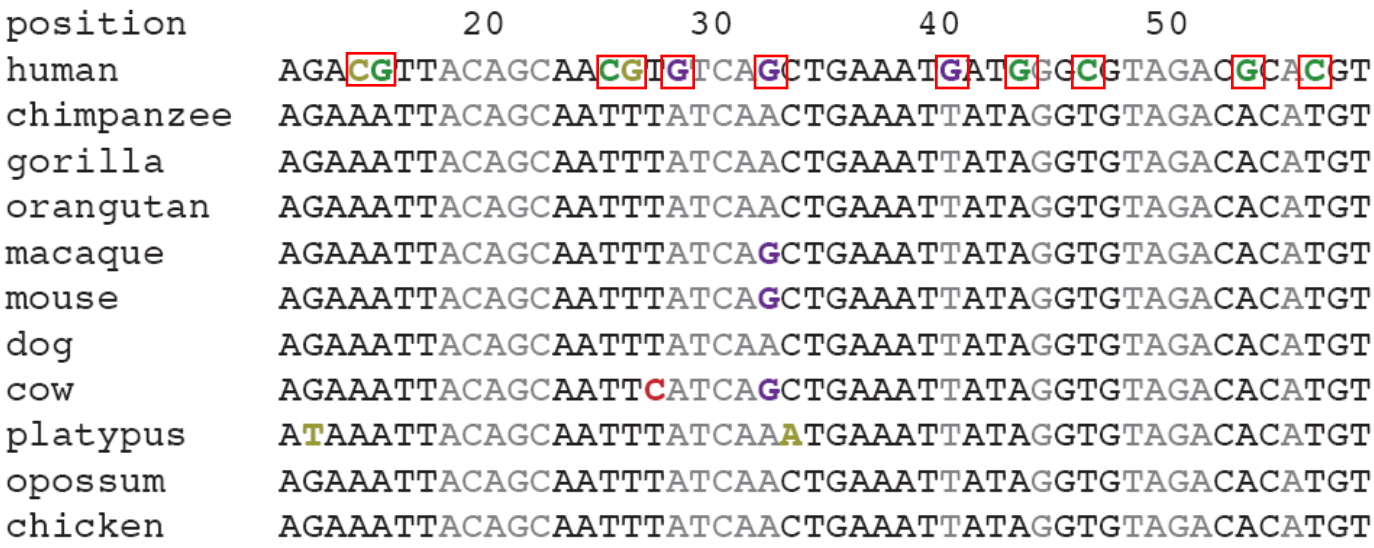
Ultra conserved elements





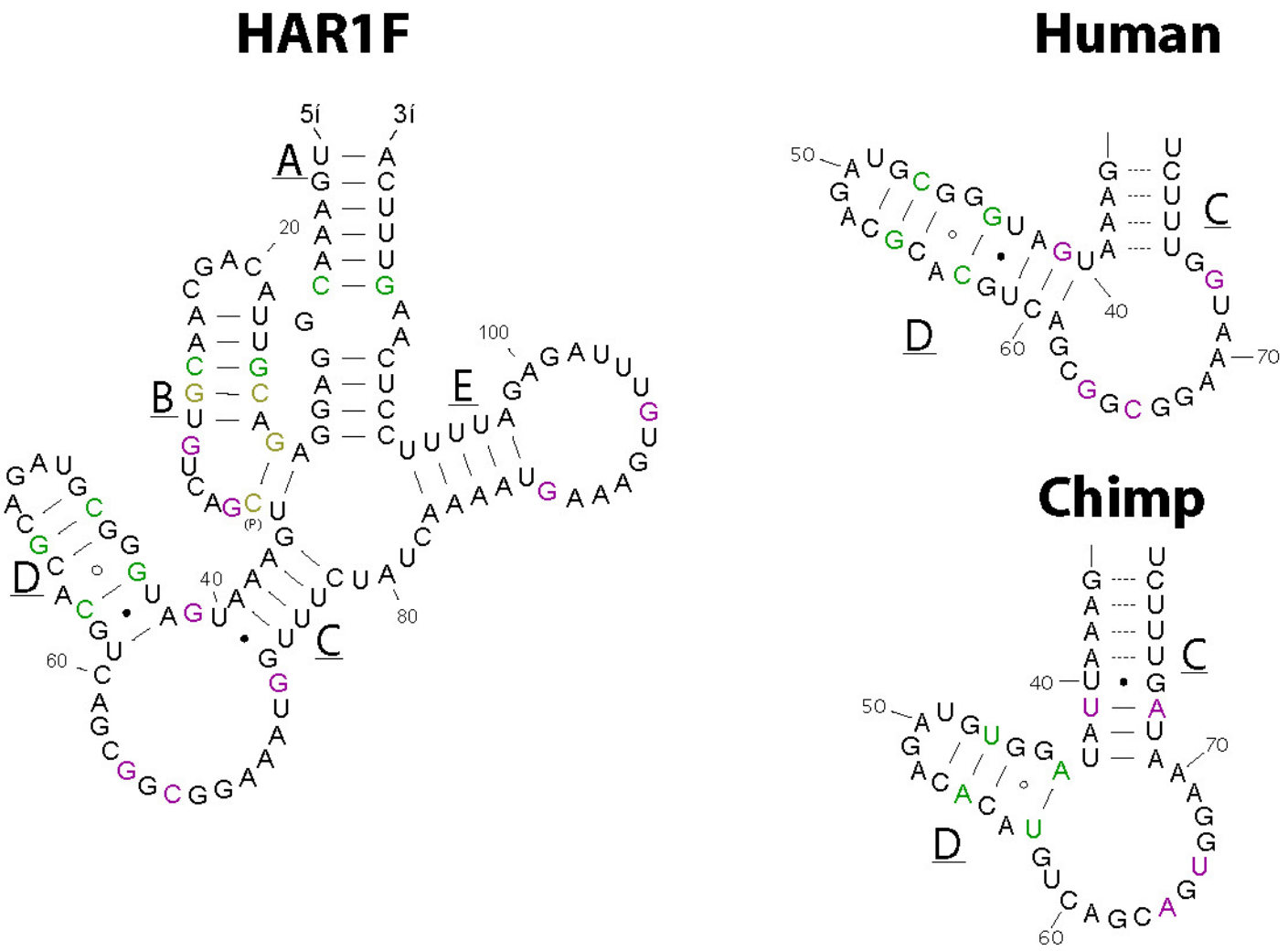
e.d 12.5

HARs: Human accelerated regions

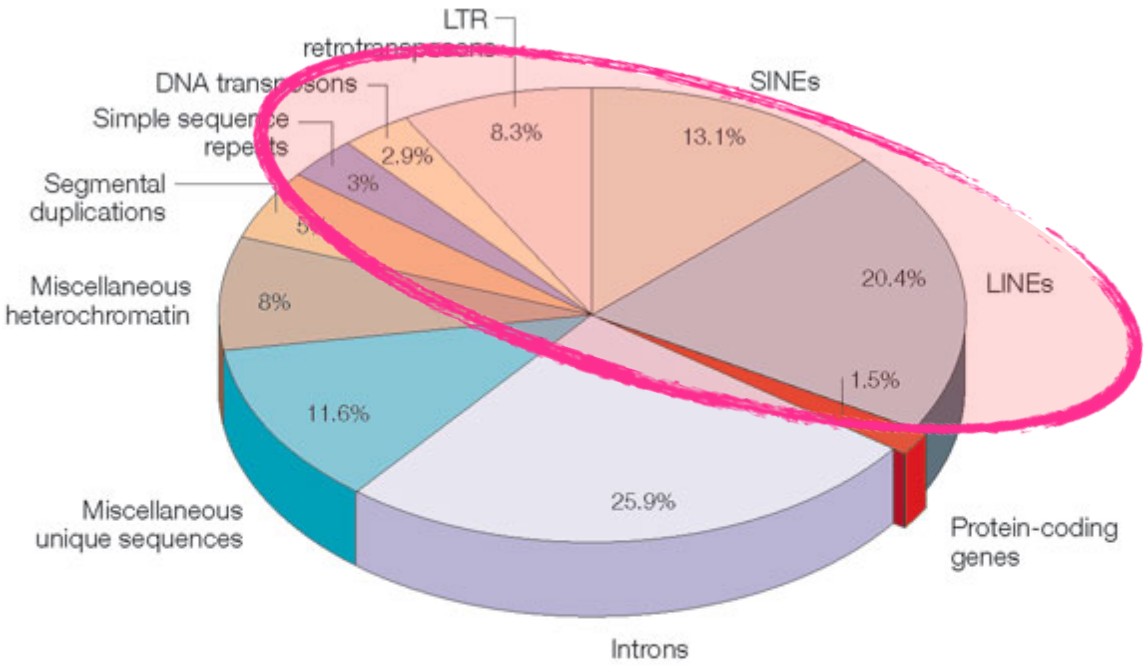


- 118 bp segment with 18 changes between the human and chimp sequences
- Expect less than 1

Human HAR1F differs from the ancestral RNA structure



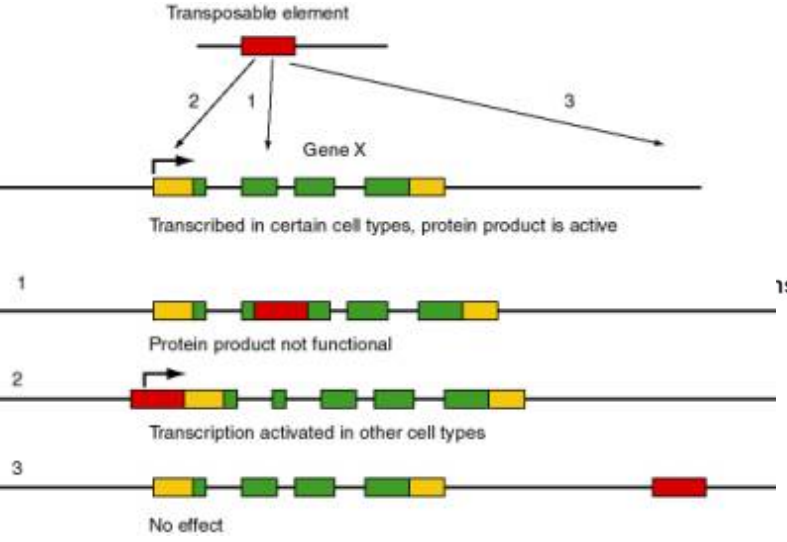
Main components in the Human genome



Barbara McClintock

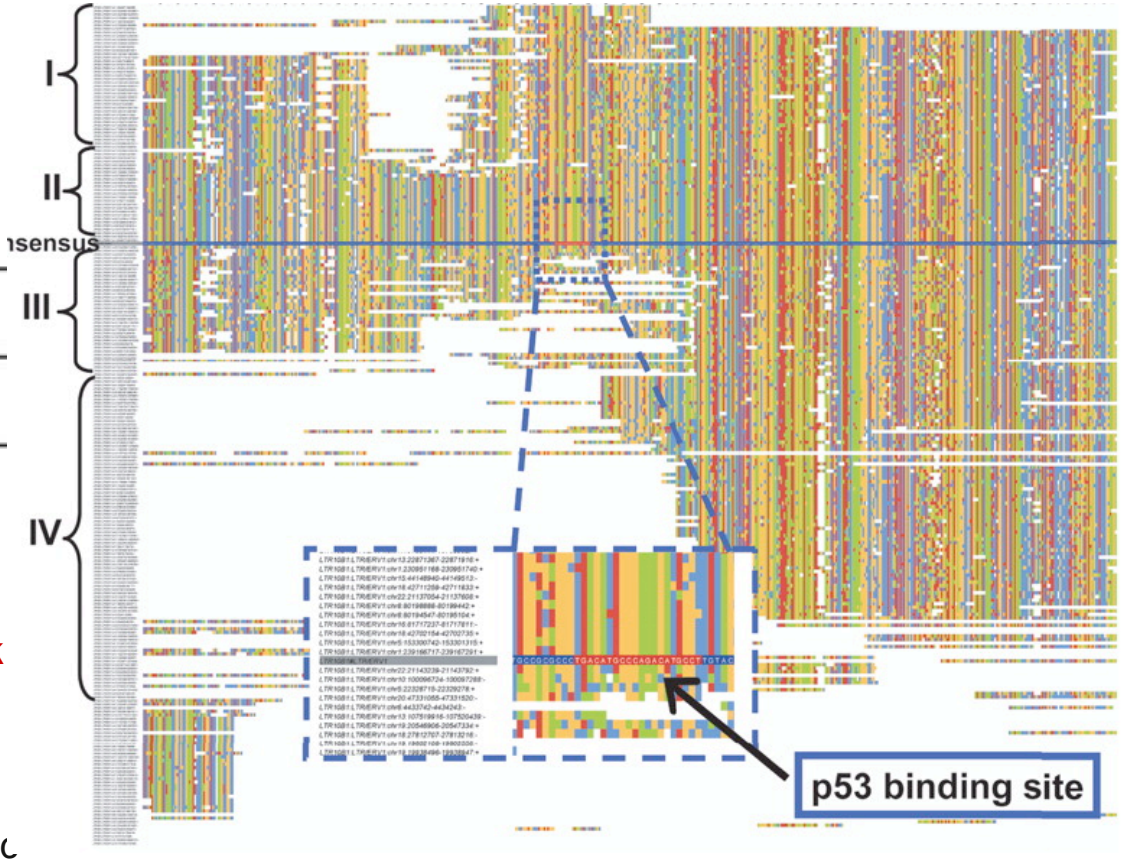
Only 1.5% of the human genome are protein-coding regions
Transposable elements make up almost half of the human genome

Transposable Elements (TEs)



TEs can shape transcriptional network

The LTR10 and MER61 families are particularly enriched for copies with a p53 site. These ERV families are primate-specific and transposed actively near the time when the New World and Old World monkey lineages split.

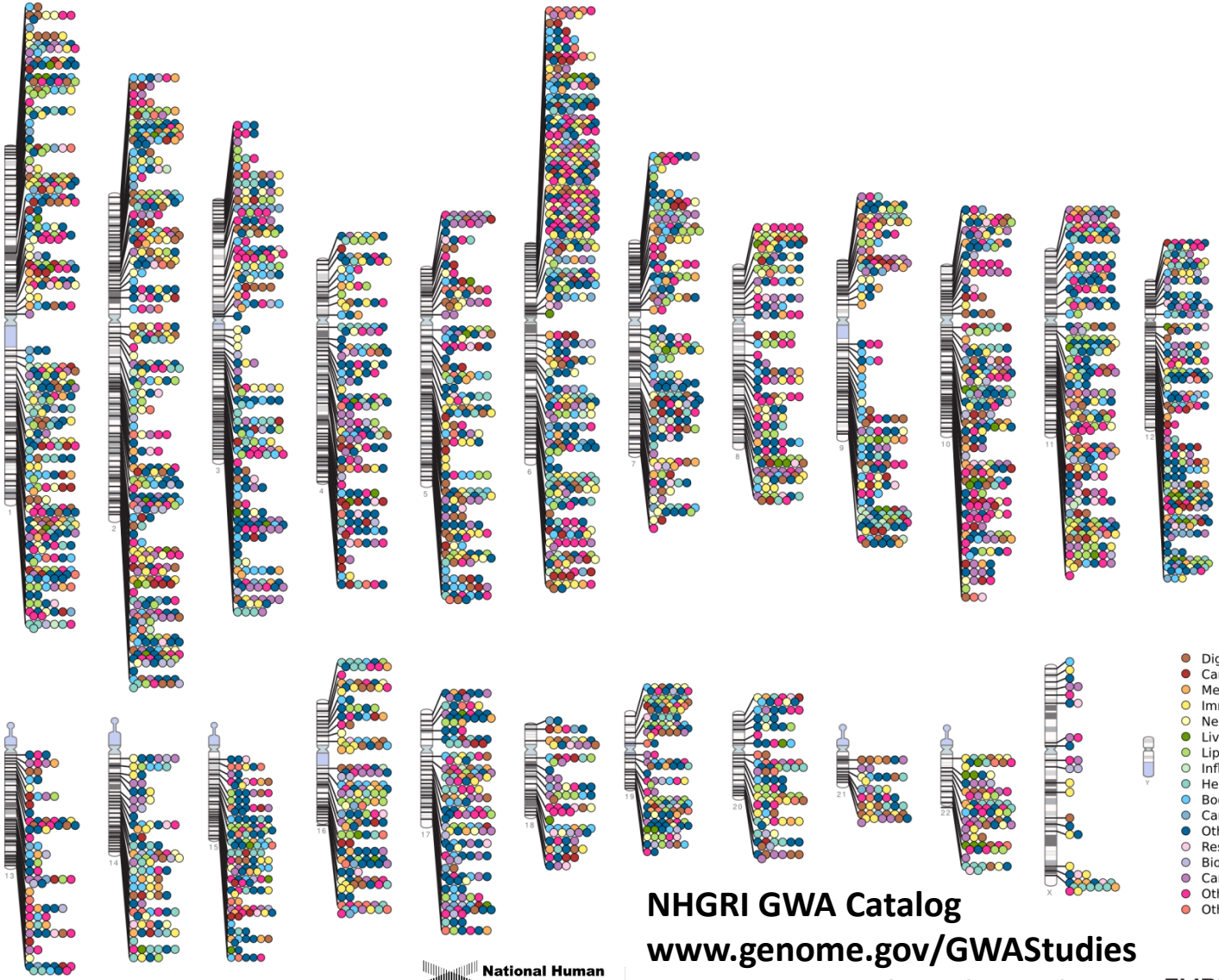


Evolutionary pattern of LTR10B1 genomic copies

Wang et al. PNAS 2007

Understanding Human Disease

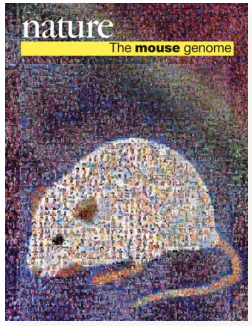
Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI



GENOME 10K©

Proposed in Nov 2009

ROADMAP epigenomics PROJECT

Epigenome

ENCODE Project

HapMap

1000 Genomes

THE CANCER GENOME ATLAS

TCGA

4D Nucleome

Thinking Quantitatively

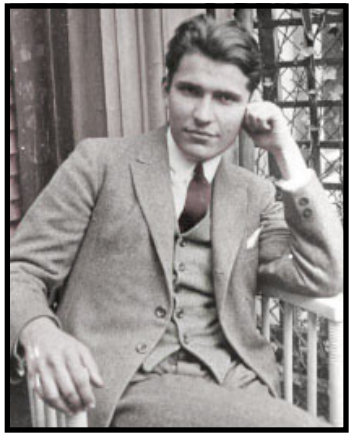
Biology Is A Quantitative Science!!!

Gregor Mendel

1823-1884

1) Mendel's Laws
2) Chargaff's Rules

Erwin Chargaff



1929-1992

Thinking Quantitatively

- Space
  - Be comprehensive
- Signal to Noise Ratio
  - Sensitivity, specificity, dynamic range
  - What is my background control?
- Distributions
  - Normal, Gaussian, Poisson, negative binomial, extreme value, hypergeometric, etc.
  - Discrete vs continuous
- The *P* value
- Statistics, Probability, Computation, and Informatics
- Don't forget genetics!!!

Simple principle:
what is your expectation?
what is your observation?

Spaces: Be comprehensive

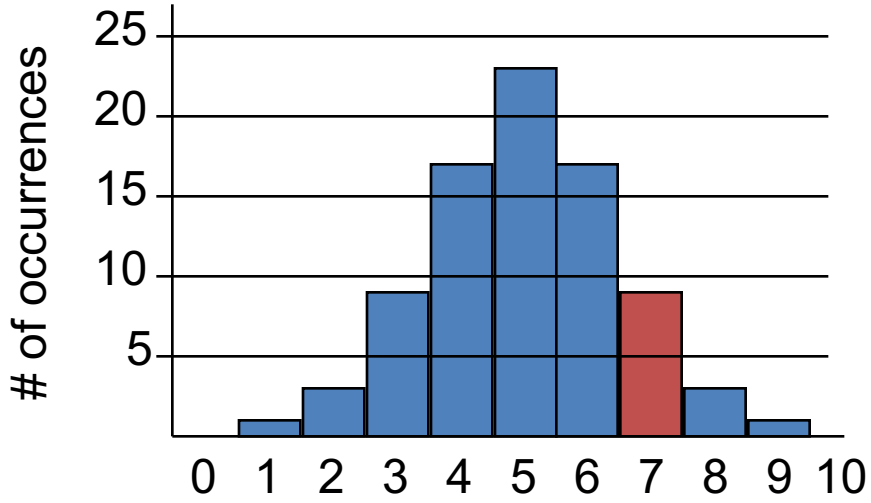
- Conditions: spatial, temporal, treatment – think about controlling for multiple variables
- Think globally – interaction between local features and global features (Placenta histone example)
- Be comprehensive about what assumptions are made – some we know, some we don't (genome assembly example)

Sensitivity, Specificity, and Dynamic Range

- **Sensitivity**
  - What is the smallest signal that can reliably be detected (signal to noise)?
  - True positive rate
- **Specificity**
  - How well can we discriminate between similar signals?
  - True negative rate
- **Dynamic Range**
  - What is the linear range of detection?
  - What is the range of natural variation?

The *P* value

*for discreet variables

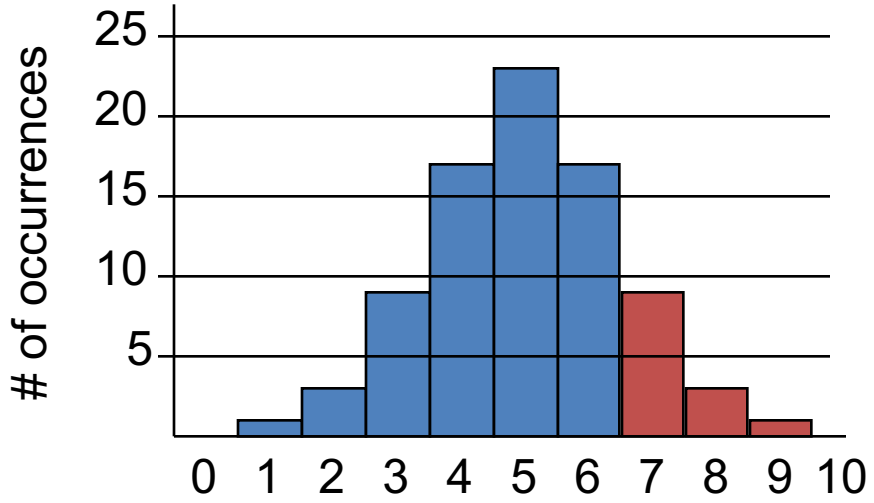


What is the chance of getting exactly seven?

$$P = \frac{\#\,of\ trials\ that\ were\ seven}{total\ \#\,of\ trials}$$

$$P = \frac{9}{1+3+9+17+23+17+9+3+1}$$

$$P = 0.10$$



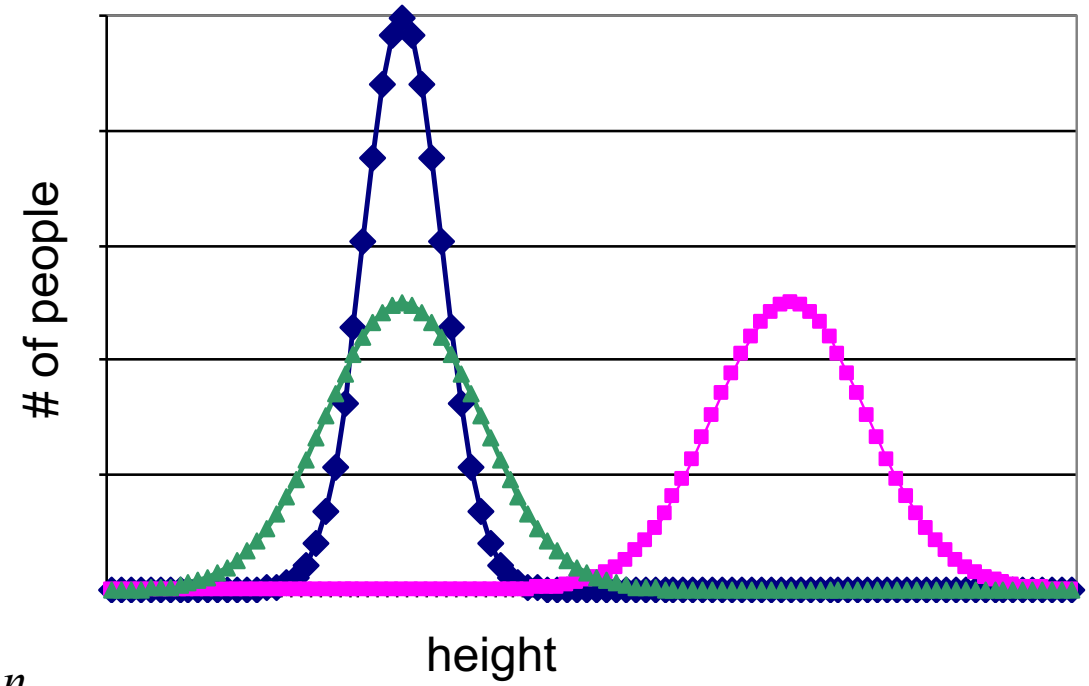
What is the chance of getting seven or better?

$$P = \frac{\#\,of\ trials\ that\ were\ seven\ or\ better}{total\ \#\ of\ trials}$$

$$P = \frac{9+3+1}{1+3+9+17+23+17+9+3+1}$$

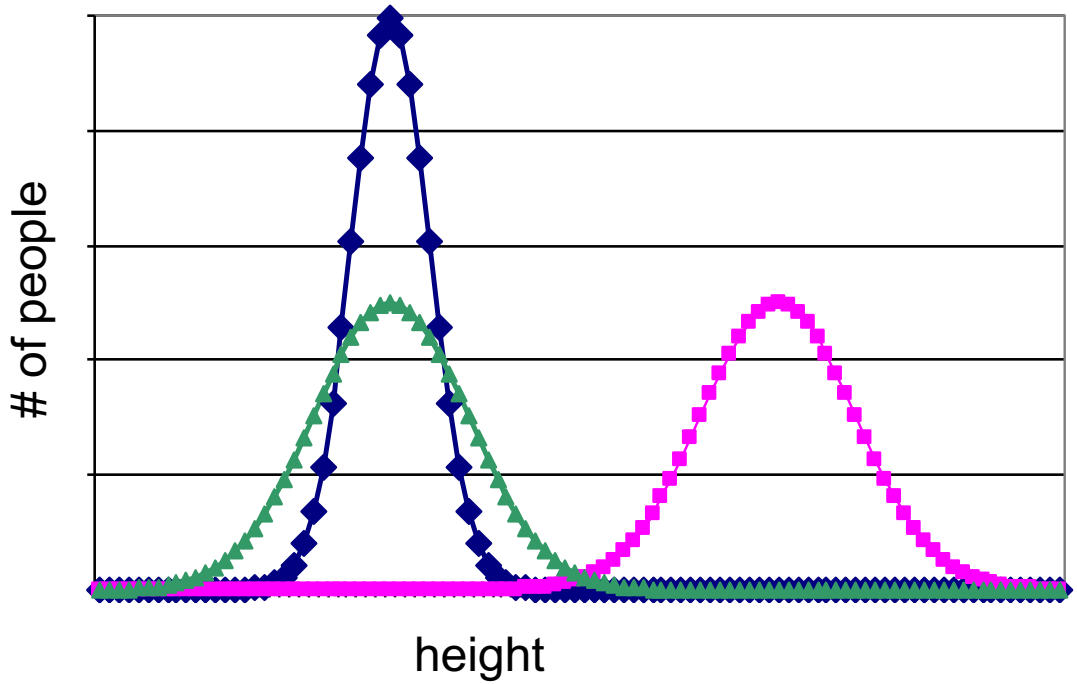
$$P = 0.16$$

Gaussian (Normal) Distributions I



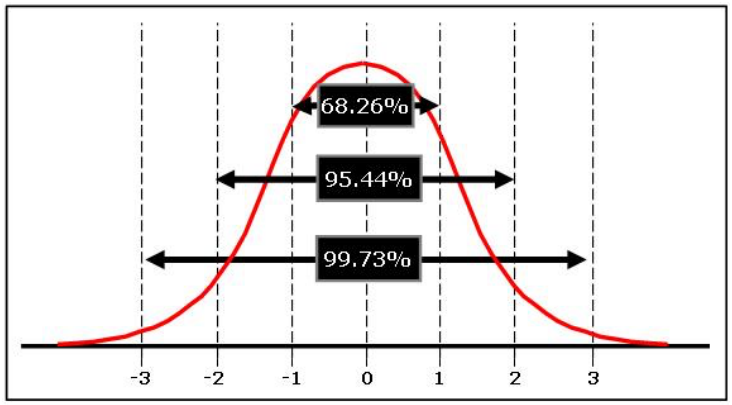
- Mean ($\bar{x}$) = $\frac{1}{n}\sum_{i=1}^{n} x_i$
- Standard Deviation ($\bar{s}$) = $\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Gaussian (Normal) Distributions II

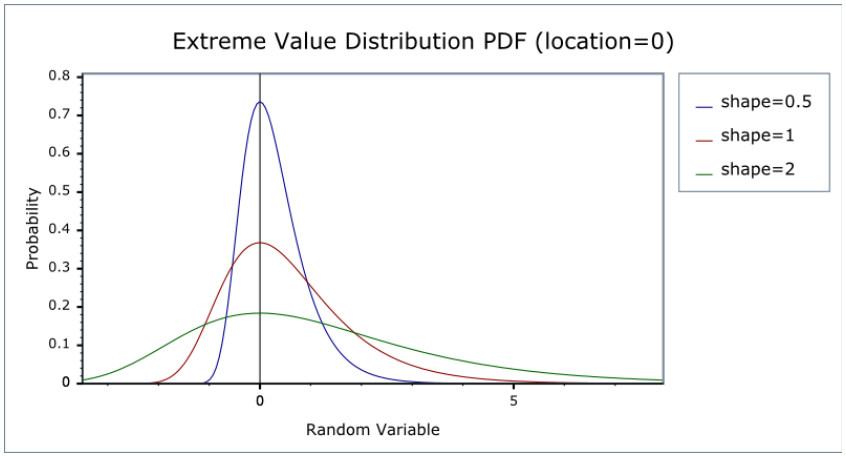


- Mean ($\bar{x}$ or μ)
- Standard Deviation (s or σ)
- Compare means (t-tests)
- Compare standard deviations (f-tests)
- Calculate *P*-values for particular results (z-tests)

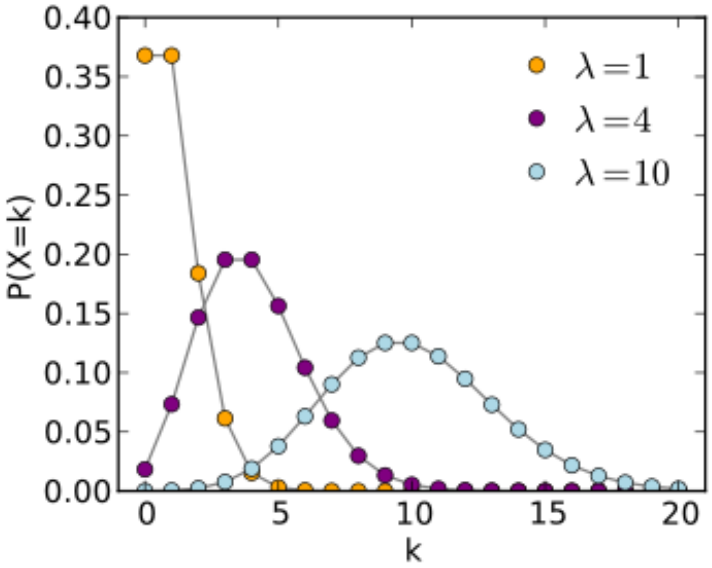
Distributions



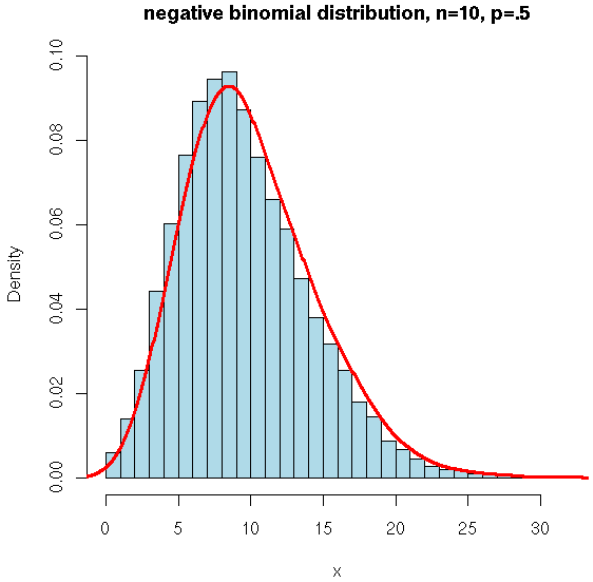
Normal (Gaussian)



Extreme value



Poisson

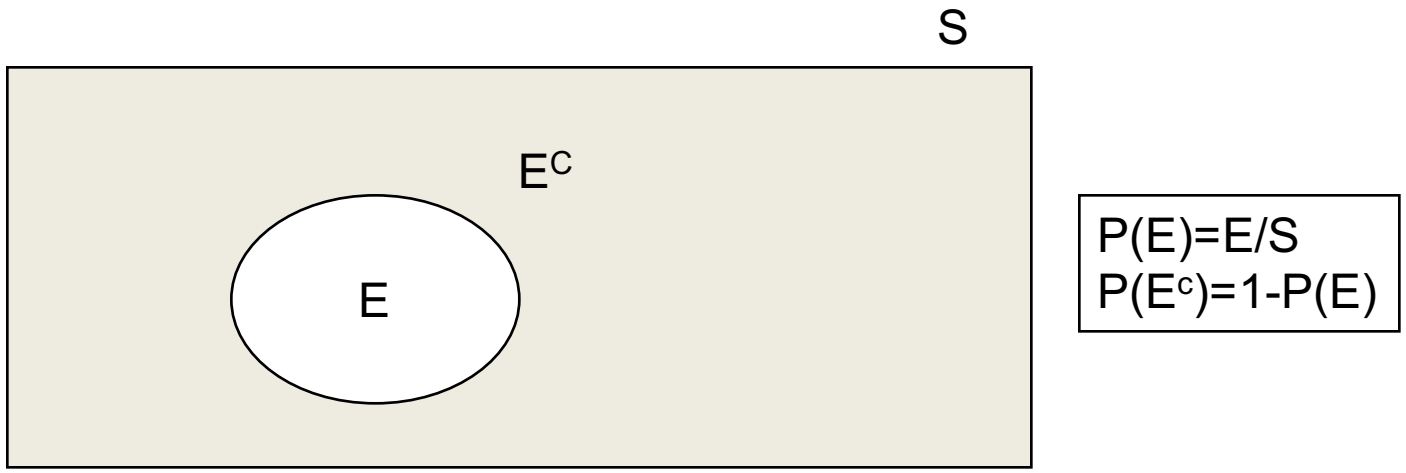


Negative binomial

"The most important questions of life are …really only problems of probability."

-Pierre Simon, Marquis de Laplace (1749-1827)

Probability: Computing the chance of a particular outcome of an experiment



Independent Events/Mutually Exclusive Events

What is the probability of A and B both occurring?

$P(AB) = P(A)*P(B)$ (Independent)

$P(A|B) = P(A)$ (Independent)

$P(A|B)*P(B) = P(B|A)*P(A)$ (Bayes rule)

$P(AB) = P(A|B) = 0$ (Mutually Exclusive)

What is the probability of A or B occurring (independent)?

$P(A) * P(B) + P(A) * 1\text{-}P(B) + 1\text{-}P(A) * P(B)$

both + A only + B only

What is the probability of A or B occurring (mutually exclusive)?

$P(A) + P(B)$

Example: Probability

Amino acid percentages of Swissprot

Ala	(A)	7.81	Gln	(Q)	3.94	Leu	(L)	9.62	Ser	(S)	6.88
Arg	(R)	5.32	Glu	(E)	6.60	Lys	(K)	5.93	Thr	(T)	5.45
Asn	(N)	4.20	Gly	(G)	6.93	Met	(M)	2.37	Trp	(W)	1.15
Asp	(D)	5.30	His	(H)	2.28	Phe	(F)	4.01	Tyr	(Y)	3.07
Cys	(C)	1.56	Ile	(I)	5.91	Pro	(P)	4.84	Val	(V)	6.71

What is the probability that a peptide of length 25 contains at least one SP motif?

What is the probability that the last residue of a protein is either K or R?

Example:
Counting permutations

Permutations:

Question: How many different 5-mer sequences can I make using each of the amino acids S, T, A, G, P once and only once?

Answer:

Example:
Counting combinations

Combinations:

Question: How different 5-mer sequences can I make using three Ser's and two Pro's

Answer:

Example: hypergeometric distribution

A particular cluster has 25 coexpressed genes in it. 15 of these genes are annotated as being involved in rRNA transcription.

Is 15/25 significant?

Example:
Sensitivity and specificity

Disease prevalence: 0.1%

Test sensitivity: 99%

Question: is the test worth taking?

Answer: