Bio5488 Genomics
Spring, 2026

http://genetics.wustl.edu/bio5488/

https://github.com/jinlab-washu/Bio5488-Genomics

Jan 12 & 14, 2026

Lectures: Mon, Wed 10:00-11:30 AM

Labs: Fri 10:00-11:30 AM

Location: Holden *In person~~~*

Washington University in St. Louis
School of Medicine

Objectives

- Provide a strong foundation in the theory and fundamental analysis techniques used in genomics research.

- Equip you with essential bioinformatics skills to effectively access, navigate, and utilize databases containing sequence data, expression data, and other genome-wide information.

A few MTE administrivia...

- If you didn't receive an email from genomics.bio5488@gmail.com last week, please email genomics.bio5488@gmail.com or talk to an MTE after class

- If you're taking the lab:
  - Read assignment 1
  - Attempt to install the required software
  - Bring your laptop to class on Friday (*in person!!!*)

Course website and communications

- http://www.genetics.wustl.edu/bio5488/
- https://github.com/jinlab-washu/Bio5488-Genomics
- Linux Primer
- Python Primer
- Lecture Notes
- Schedule
- Weekly Assignments and Answers
- Weekly Readings
- Post questions on Piazza: http://piazza.com/wustl/spring2026/biol5488/info

Grading

4 credit

- 25% midterm
- 25% final
- 50% weekly assignments

3 credit

- 50% midterm
- 50% final

Audit/sit-in

- There will be 13 assignments. (https://github.com/jinlab-washu/Bio5488-Genomics)
- New assignments will be released on Wednesdays, with links added to the syllabus on the course website.
- Assignments are due the following Friday by 11:59 PM. Check the syllabus for exact due dates.
- Assignments must be turned in on time. Late submissions (even 12:00 AM on Saturday) will receive a score of **0**, unless you have an approved extension from the MTEs.
- The two lowest-scoring labs will be dropped.
- To request an extension, email **genomics.bio5488@gmail.com** at least **1 day before the assignment deadline**.

Expectations

- Ensure you have completed the prerequisites. If not, reviewing the slides from Bio 5075: Introduction to Coding and Statistical Thinking in Genetics and Genomics will be helpful. https://sites.wustl.edu/bio5075/
- Attend every class unless not possible.
- Read all assigned articles whenever possible to enhance your understanding.
- Participate actively in class. If you're confused, don't hesitate to ask questions—questions are valuable and welcome!
- Put in the effort — it will pay off.

Pre-course work for students with no prior experience

1. UC Davis — MCB 192 (Online) (Gerald Quon): https://www.youtube.com/@quonbio/courses

2. Cold Spring Harbor Laboratory — Computational Genomics Course (David Hawkins, Danny Miller, Lauren Mills): https://millerlaboratory.com/cshl

3. University of Utah — Applied Computational Genomics (Aaron Quinlan): https://github.com/quinlan-lab/applied-computational-genomics

4. Coursera — Introduction to Genomics Technologies (Steven Salzberg, Jeff Leek): https://www.coursera.org/learn/introduction-genomics#modules

Instructors and MTEs

Si Jia Chen

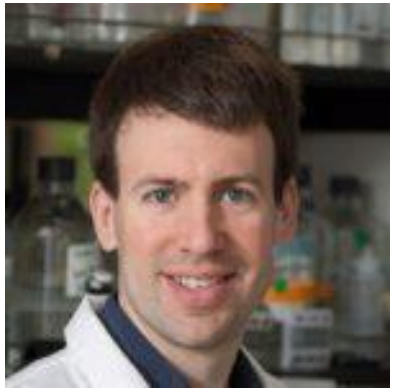
Yuchen Cheng

Ting Kuan Chu

Benjamin Van Court

Shantal Garcia

Nina Tekkey



Introduce yourself!

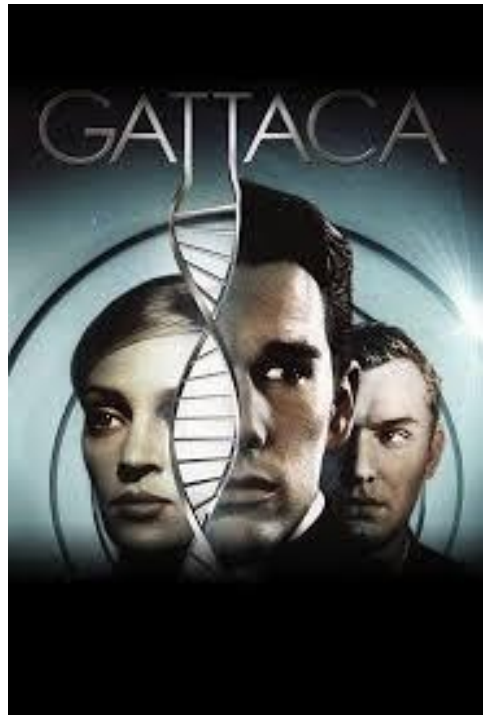
- What is your name, and what do you prefer to be called?
- Where are you from (or where did you grow up)?
- Which lab are you currently rotating, or what is your thesis lab?
- What's a favorite movie (or show) you've watched recently?
- What do you like to do outside the lab (hobbies, sports, creative stuff)?
- What are your research interests?
- What's a research question you'd love to answer someday (even if it feels ambitious)?
- What's been the biggest surprise for you so far in grad school?
- What's a "win" you're hoping to have by the end of this course?

Why should we study genomics?

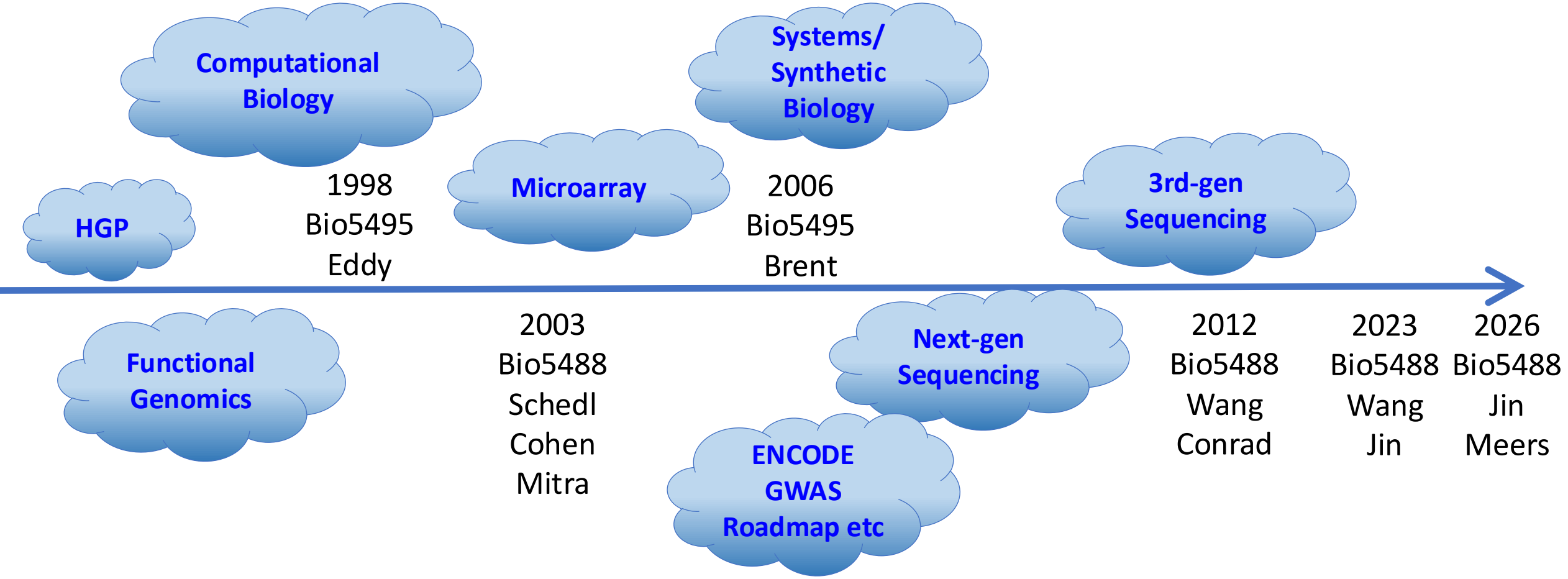
More than ever, biology is an inherently computational science. Therefore, to be a well-rounded scientist, you need to understand and take charge of every aspect of your research. In a nutshell, this means being able to run your experiment, analyze your data, and carefully think through any potential issues and caveats. Just like in any other field, you need to establish what results you expect, design strong positive and negative controls, and use the right tools to measure significance. This requires that you have the right tools to do so.

When it comes to analyzing your data, the process starts with questioning everything. You'll need to scrutinize your data thoroughly—essentially distrusting it—until you reach the point where you can confidently trust it. This course is here to teach you exactly how to do that.

A brief and incomplete history of genomics



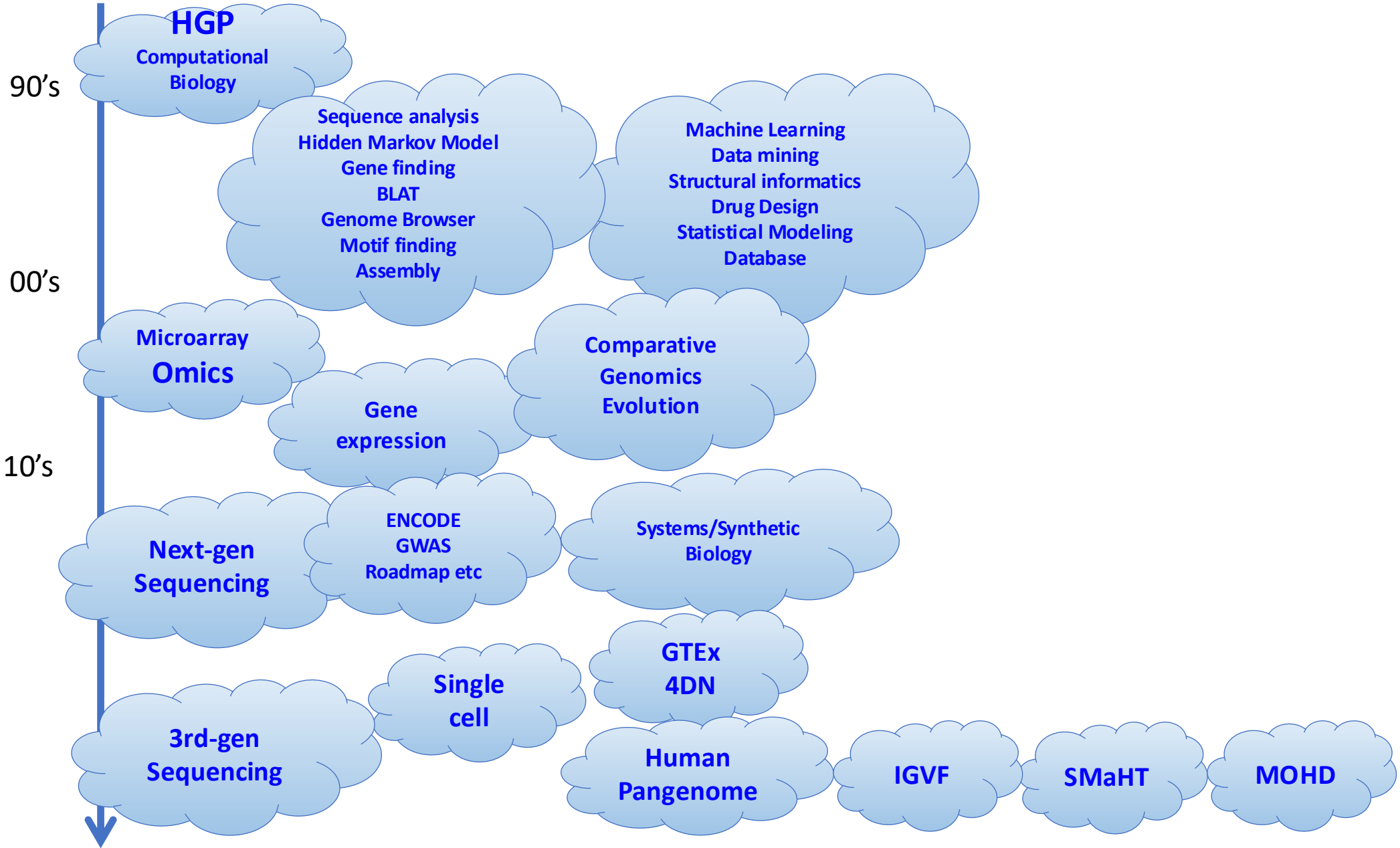
History of Bio5488



History of genomics

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 released for the first time
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



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.**

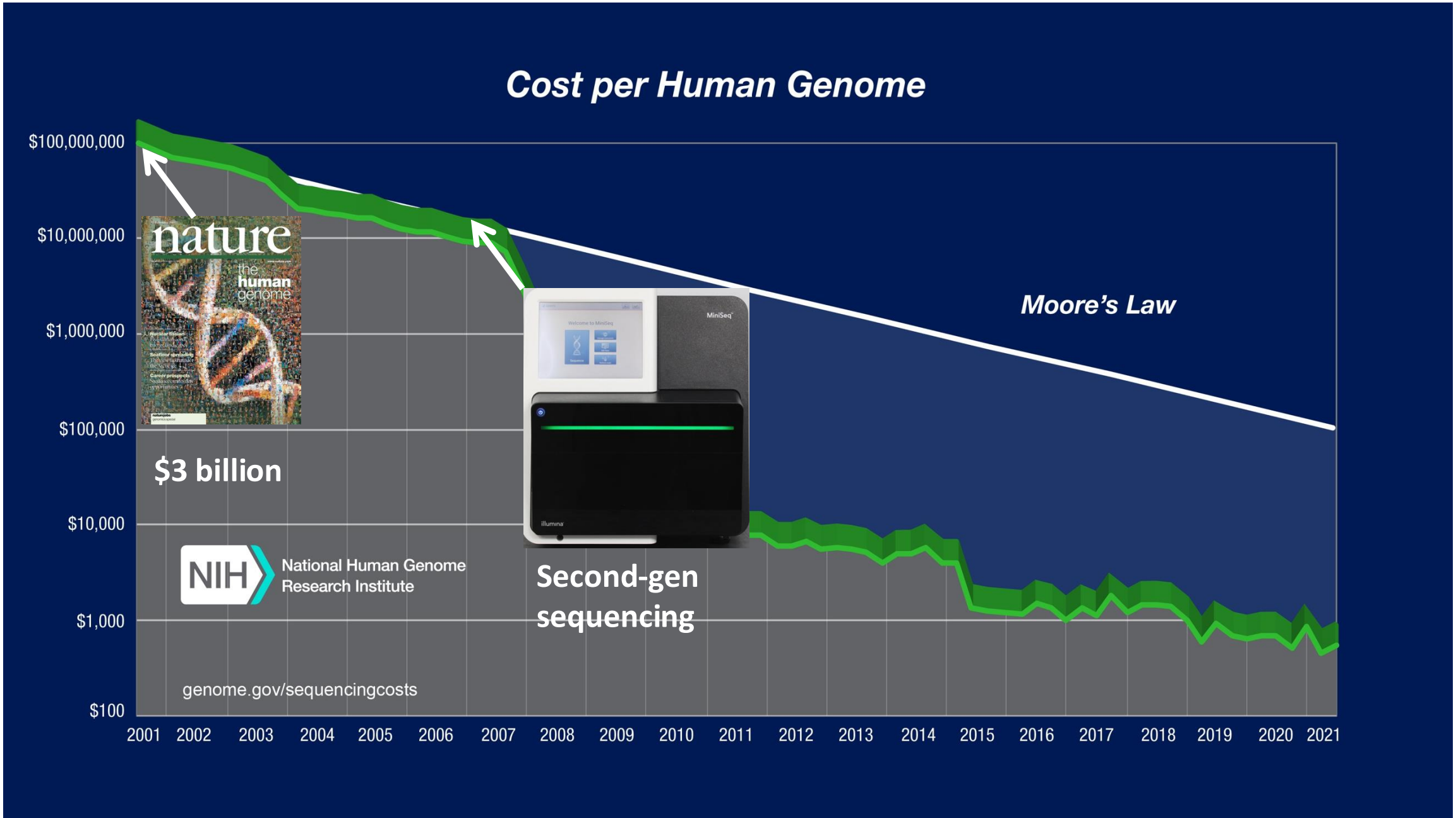
- What is genomics?
  - **The study of all of a person's genes (the genome), including interactions of those genes with each other and with the person's environment.**
  - **Biology in the big data era.**

The simple principles of genomics

The simple principles of genomics

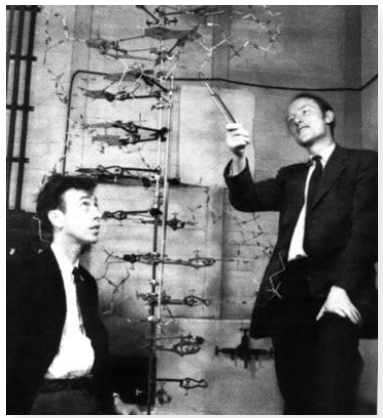
- **Characterize the genome**
  - How big
  - How many genes
  - How are they organized
- **Annotate the genome**
  - What, where, and how
- **Modern genomics: “ChIPer” vs “Mapper” vs ”CRISPRer”**
  - Direct measurement
  - Inference
  - Comparison
  - Evolution
- **From genome to molecular mechanisms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

Genome sequencing costs have dramatically declined



The Human genome: the "blueprint" of our body

Rosalind Franklin



James Watson
Francis Crick

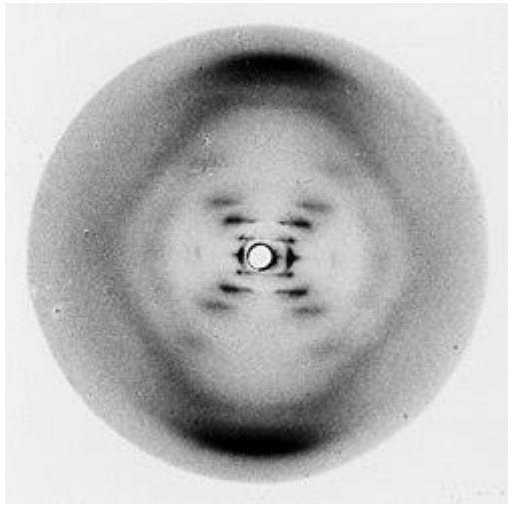
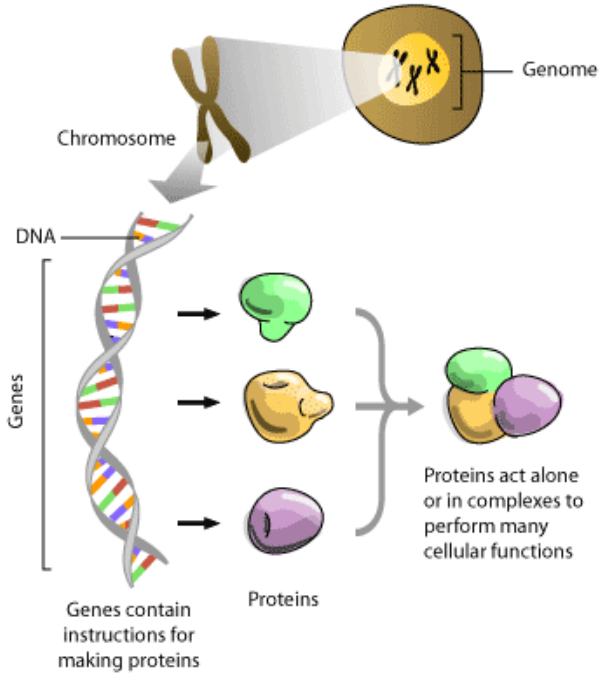
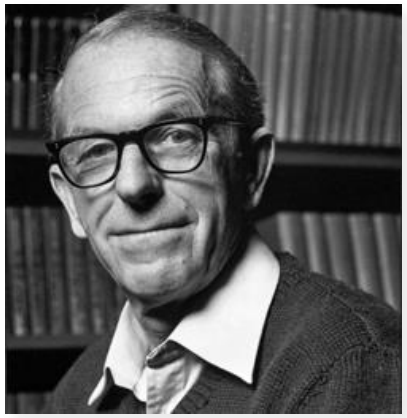


Photo 51



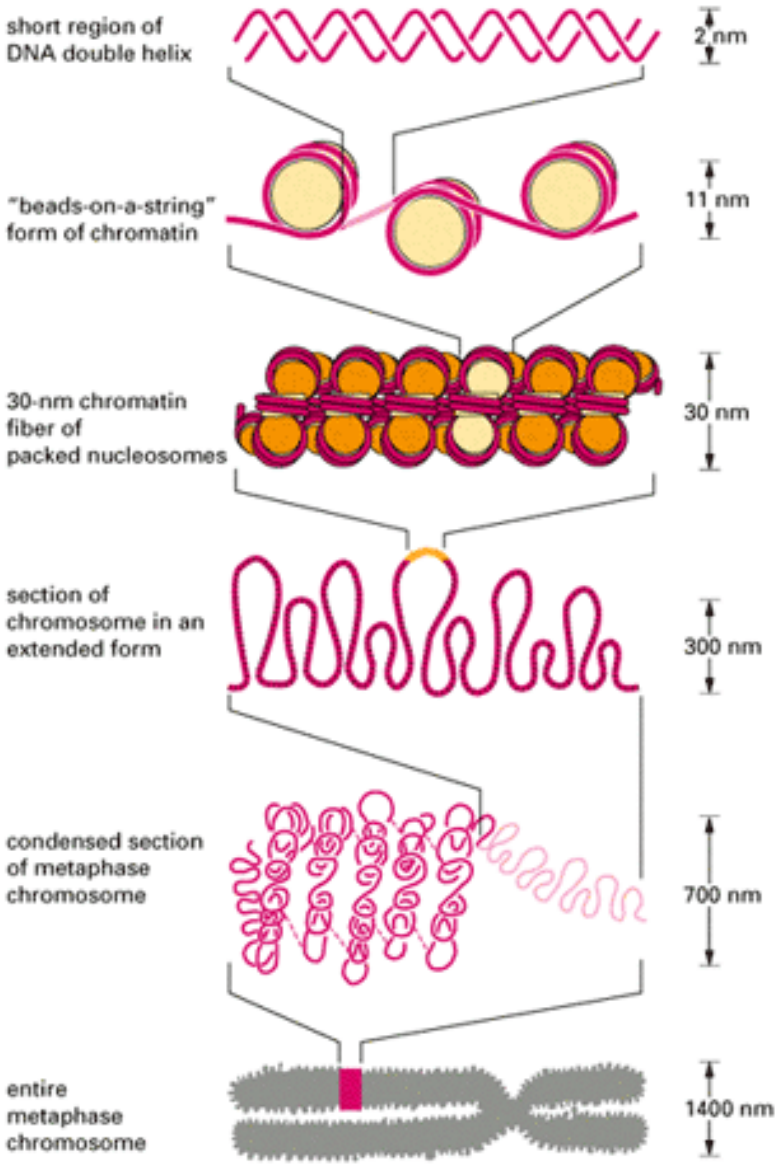


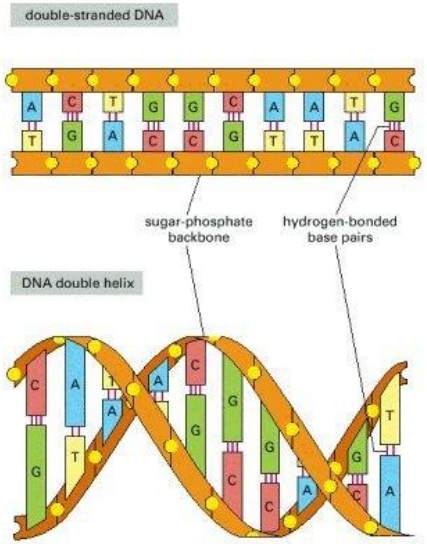
Fred Sanger

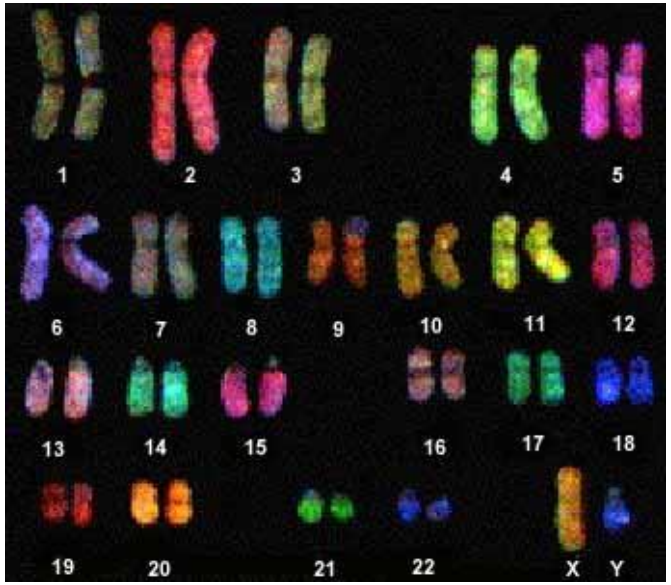
GTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCT
CCGAACAATAAAGATTCTACAATACTAGCTTTTATGGTTATG
AAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAA
ATTAACGAATCAAATTAACAACCATAGGATGATAATGCGATT
AGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG
ATGATTTTTGATCTATTAACAGATATATAAATGGAAAAGCTG
CATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTA
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}

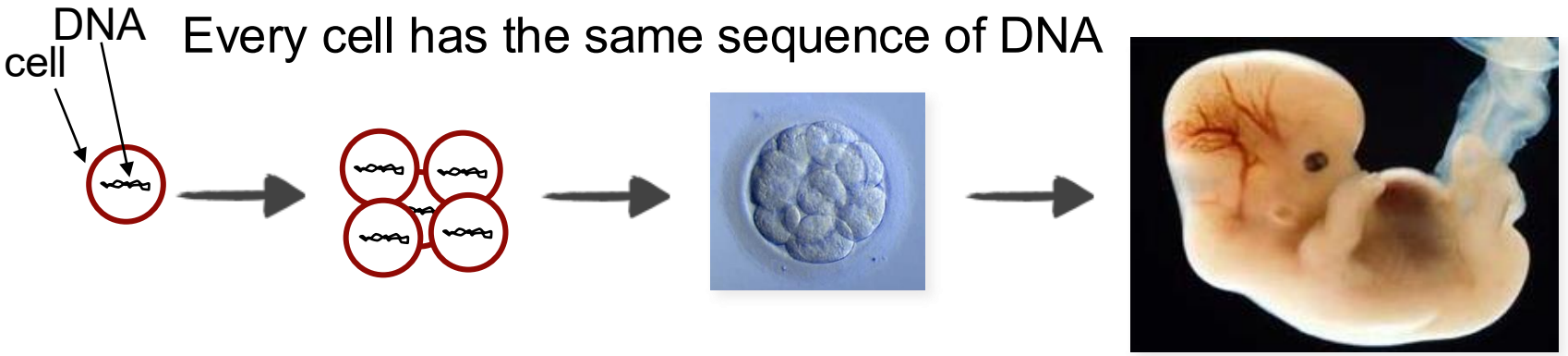
DNA, Chromosome, and Genome



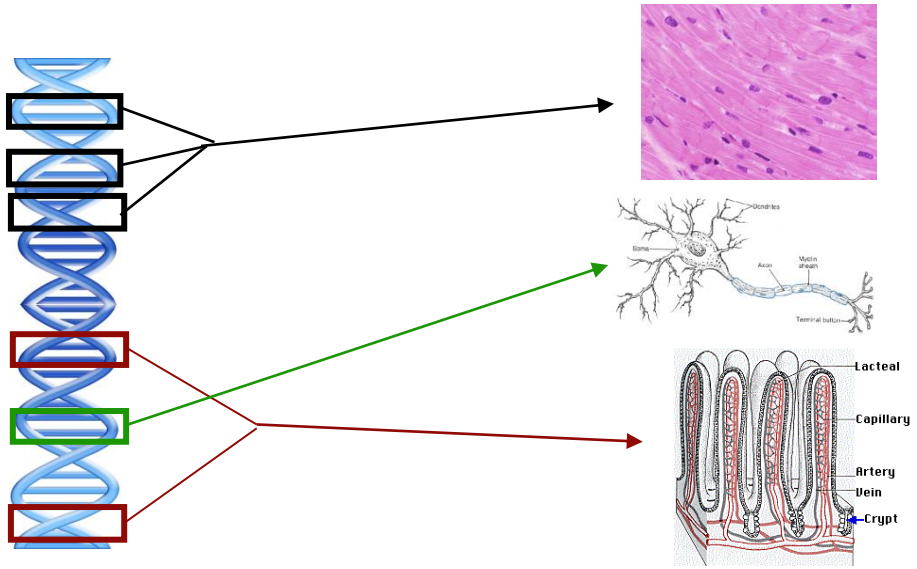




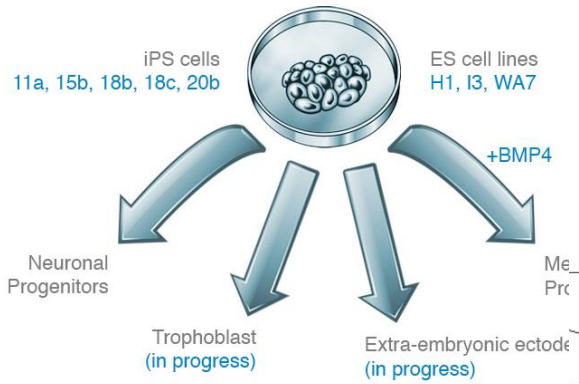
Building an organism



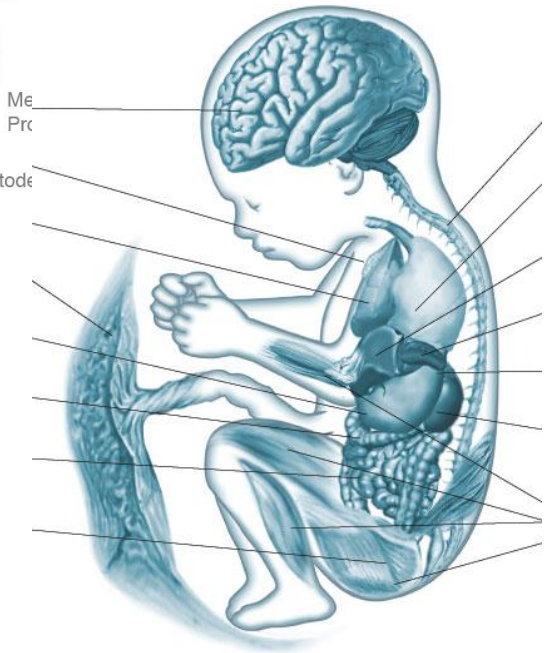
Subsets of the DNA sequence determine the identity and function of different cells



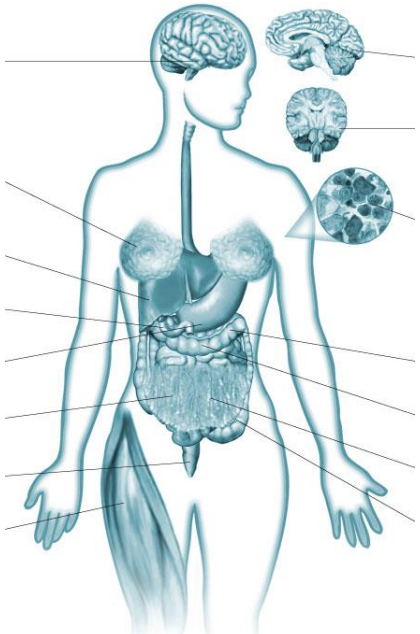
One genome, thousands of epigenomes



Embryonic stem cells



Fetal tissues



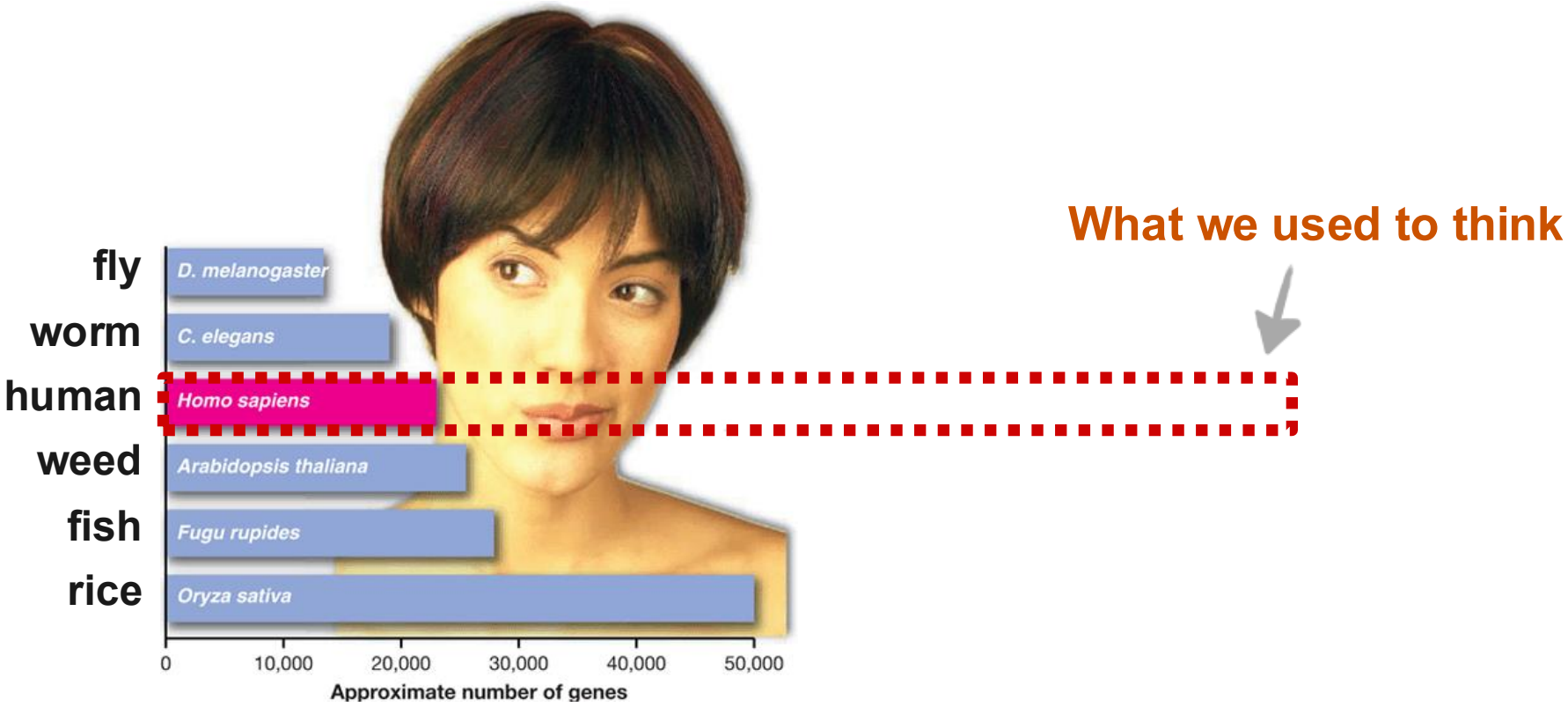
Adult cells and tissues

What makes us different?

Differences between individuals?

Differences between species?

How many genes do we have?

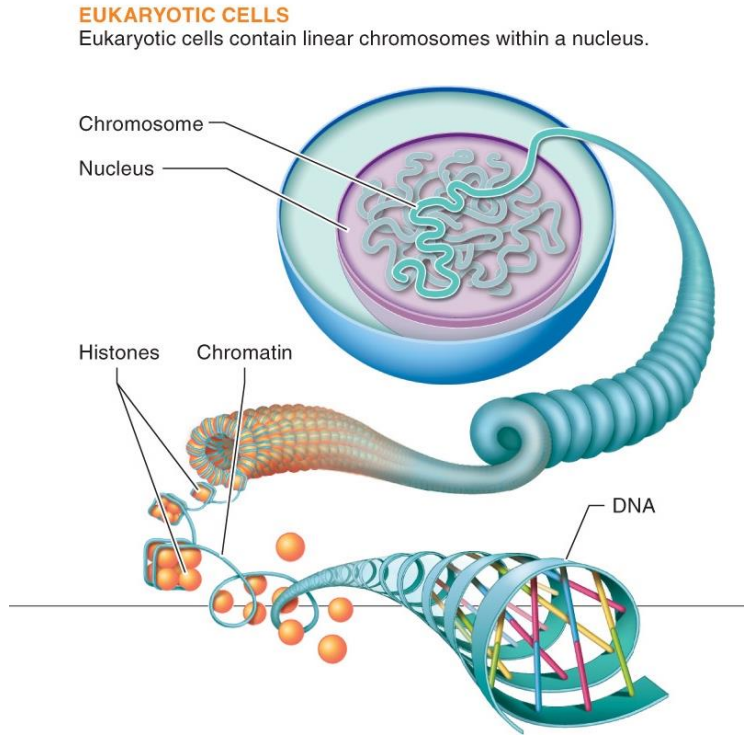


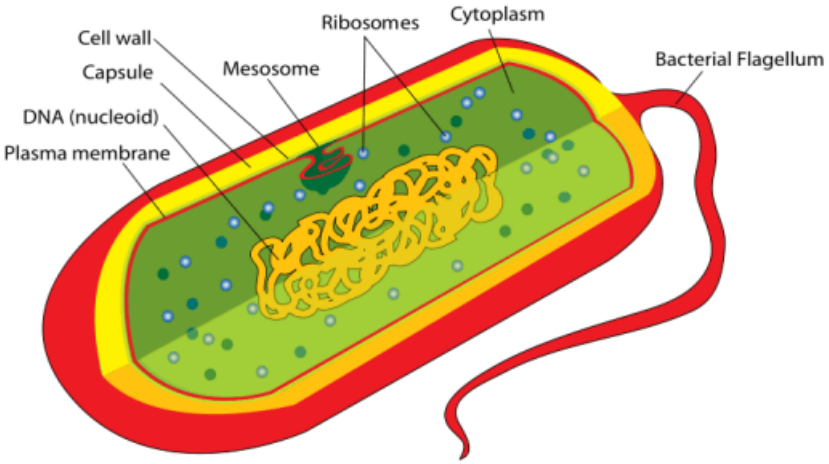
Science 2005

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

Prokaryotic versus Eukaryotic genome

	Prokaryotes	Eukaryotes
Size	- Small (1 – 2 μm)	- Large (5 – 100 μm)
Nucleus	- Absent	- Present and bounded by nuclear envelope
Content	- Gene dense - No centromere (no mitosis) - Circular DNA does not need telomeres - Usually lacks intros	- Mostly non-coding DNA - Usually contains introns and intergenic regions - Telomeres and centromere are present
Chromatin	- No histones - Supercoiling	- Histone bound (genome regulation)





https://digfir-published.macmillanusa.com/phelan3e/phelan3e_ch06_3.html

https://www.differencebetween.com/difference-between-prokaryotic-and-vs-eukaryotic-genome/

Complexity, Genome Size and the C-value Paradox

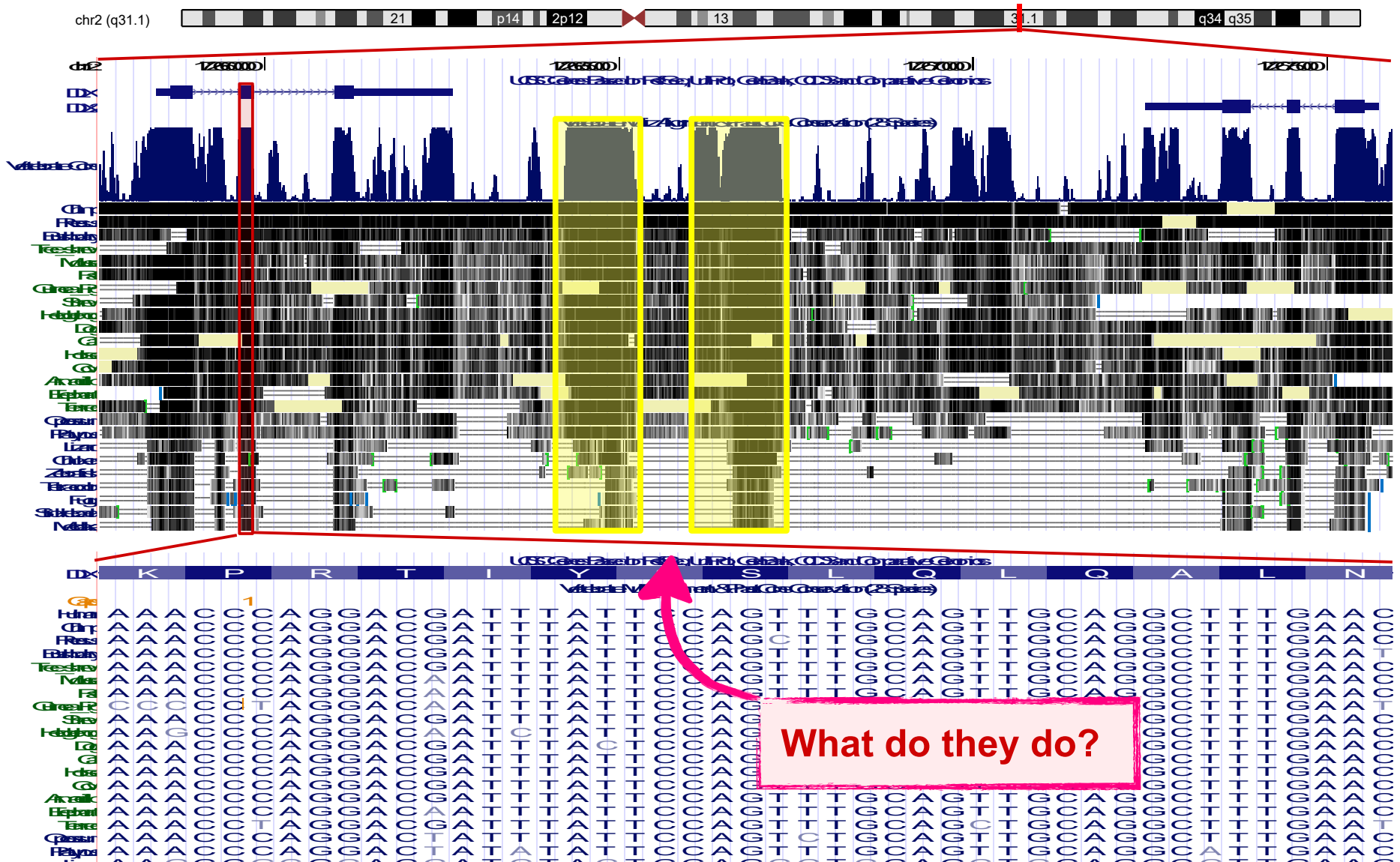
Organism	Genome Size (MB)
Amoeba	670,000
Fern	160,000
Salamander	81,300
Onion	18,000
Paramecium	8,600
Toad	6,900
Barley	5,000
Chimp	3,600
Gorilla	3,500
Human	**3,500**
Mouse	3,400
Dog	3,300
Pig	3,100
Rat	3,000
Boa Constrictor	2,100
Zebrafish	1,900
Chicken	1,200
Fruit fly	180
C. elegans	100
Plasmodium 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

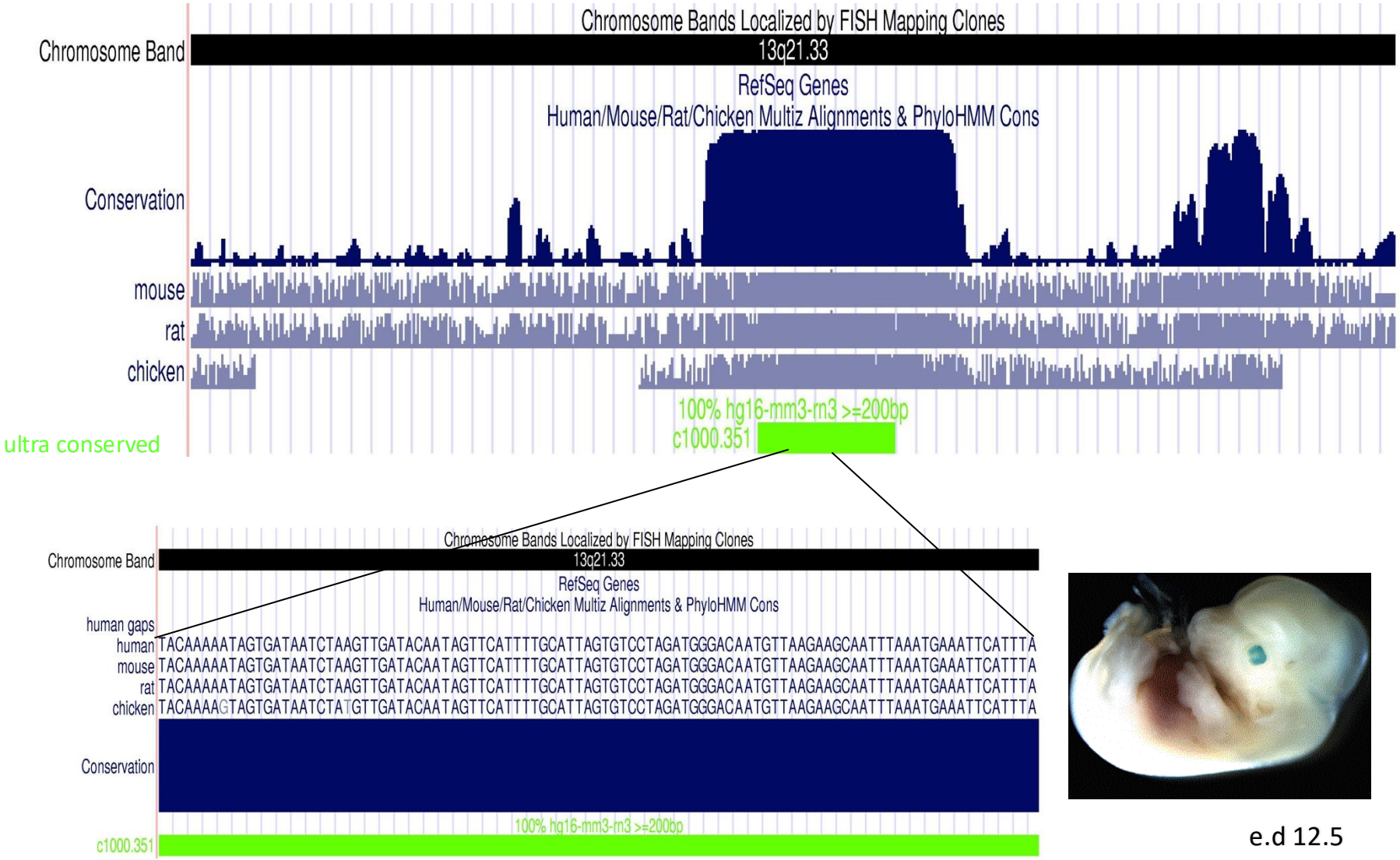
C-value: the amount of DNA contained within a haploid nucleus (e.g., a gamete) or one half the amount in a diploid somatic cell of a eukaryotic organism, expressed in picograms (1pg = 10^{-12} g).

Most functional information is non-coding

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



Ultra conserved elements



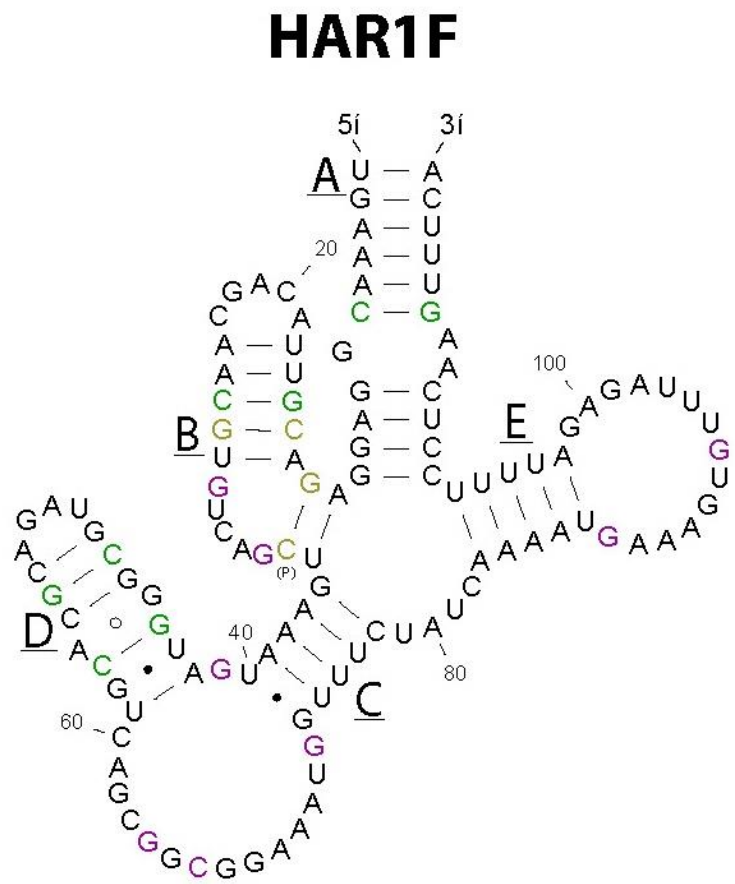
e.d 12.5

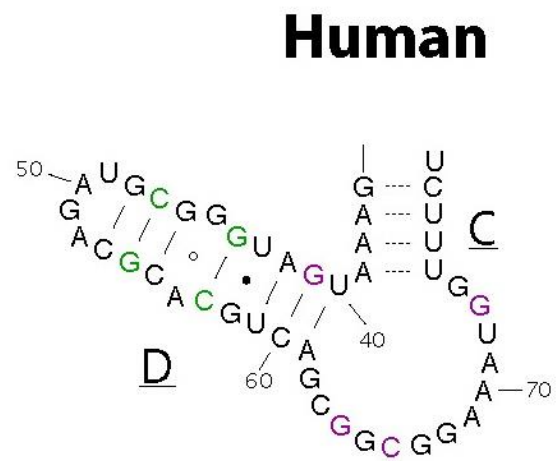
HARs: Human accelerated regions

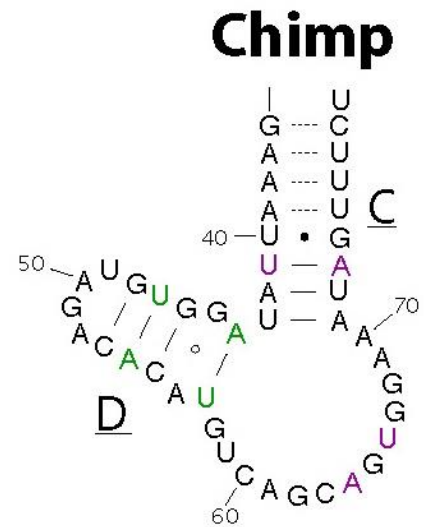
```
position          20        30        40        50
human     AGACGTTACAGCAACGTGTCAGCTGAAATGATGGCGTAGACGCACGT
chimpanzee AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
gorilla   AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
orangutan AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
macaque   AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
mouse     AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
dog       AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
cow       AGAAATTACAGCAATTCATCAGCTGAAATTATAGGTGTAGACACATGT
platypus  ATAAATTACAGCAATTTATCAAATGAAATTATAGGTGTAGACACATGT
opossum   AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
chicken   AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
```

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

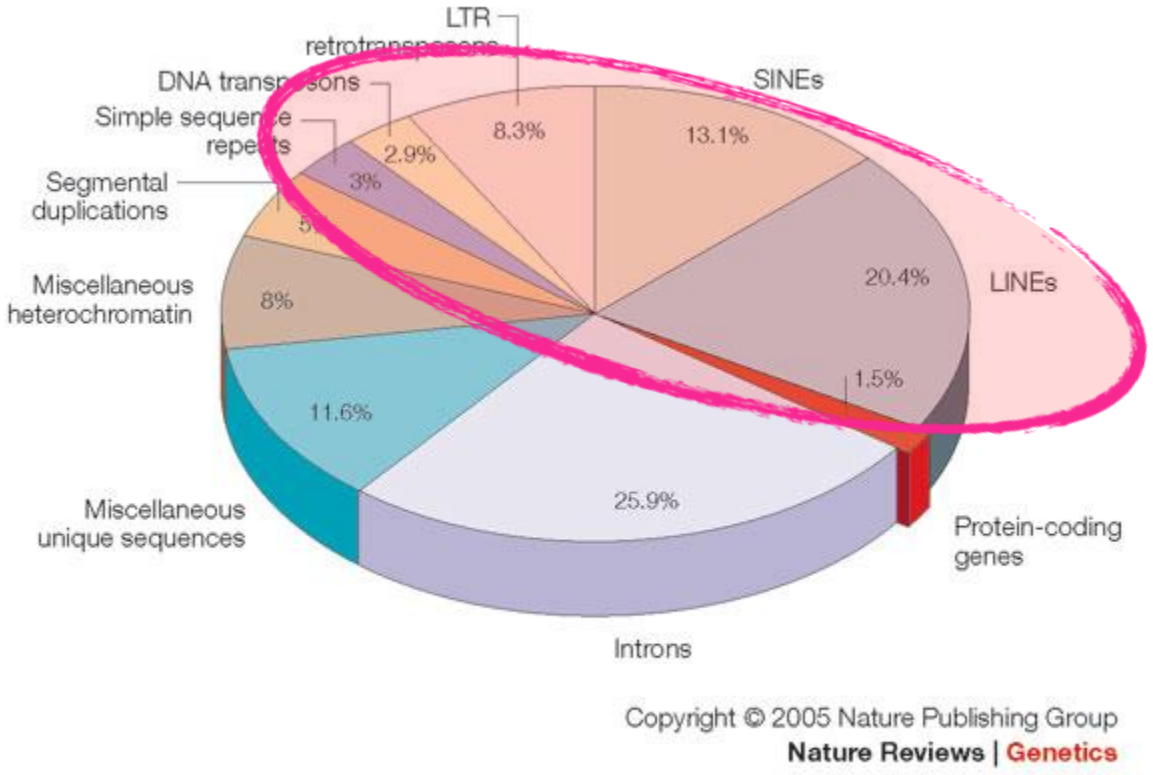
Human HAR1F differs from the ancestral RNA structure







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

Focus areas of genomics

The Future of Genomics: 10 Bold Predictions
https://youtu.be/5kAL11m_fwM

En Route to a "2020 Vision for Genomics"

1991-1995

1993-1998

1998-2003

2003-2010

2011-Present


Understanding Our Genetic Inheritance

The U.S. Human Genome Project:

The First Five Years FY 1991-1995

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
National Institutes of Health

U.S. DEPARTMENT OF ENERGY
Office of Energy Research
Office of Health and Environmental Research

POLICY FORUM

A New Five-Year Plan for the U.S. Human Genome Project

Francis Collins and David Galas*

SPECIAL SECTION GENOME

feature

A vision for the future of genomics research

A blueprint for

PERSPECTIVE

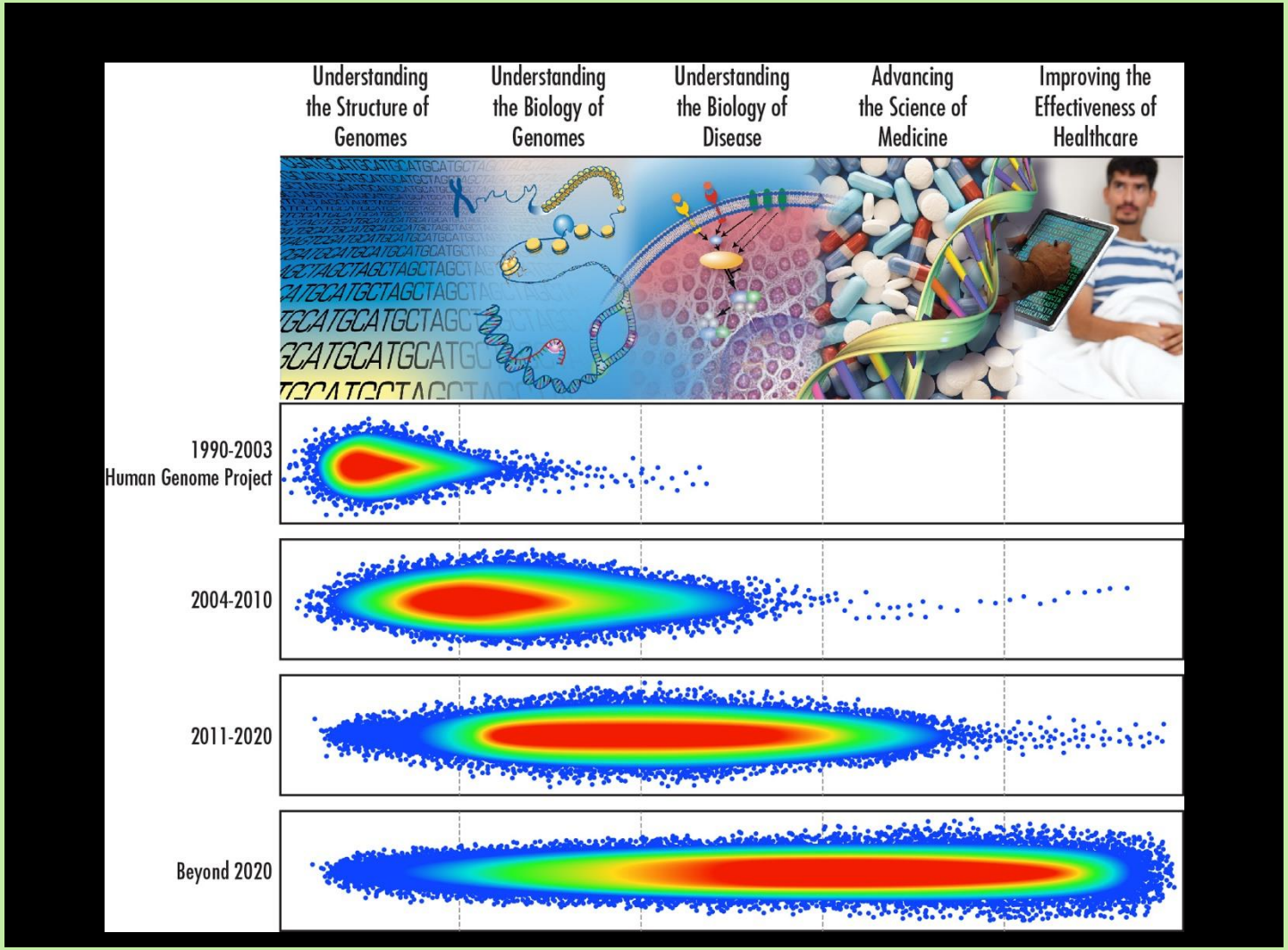
doi:10.1038/nature09764

Charting a course for genomic medicine from base pairs to bedside

Eric D. Green[1], Mark S. Guyer[1] & National Human Genome Research Institute*

There has been much progress in genomics in the ten years since a draft sequence of the human genome was published. Opportunities for understanding health and disease are now unprecedented, as advances in genomics are harnessed to obtain robust foundational knowledge about the structure and function of the human genome and about the genetic contributions to human health and disease. Here we articulate a 2011 vision for the future of genomics research and describe the path towards an era of genomic medicine.

2011-Present
En Route to Genomic Medicine



NHGRI 2020 Strategic Planning

Strategic vision document

Perspective

Strategic vision for improving human health at The Forefront of Genomics

https://doi.org/10.1038/s41586-020-2817-4

Received: 30 June 2020

Accepted: 4 September 2020

Published online: 28 October 2020

Check for updates

Eric D. Green[1]✉, Chris Gunter[1], Leslie G. Biesecker[1], Valentina Di Francesco[1], Carla L. Easter[1], Elise A. Feingold[1], Adam L. Felsenfeld[1], David J. Kaufman[1], Elaine A. Ostrander[1], William J. Pavan[1], Adam M. Phillippy[1], Anastasia L. Wise[1], Jyoti Gupta Dayal[1], Britny J. Kish[1], Allison Mandich[1], Christopher R. Wellington[1], Kris A. Wetterstrand[1], Sarah A. Bates[1], Darryl Leja[1], Susan Vasquez[1], William A. Gahl[1], Bettie J. Graham[1], Daniel L. Kastner[1], Paul Liu[1], Laura Lyman Rodriguez[1], Benjamin D. Solomon[1], Vence L. Bonham[1], Lawrence C. Brody[1], Carolyn M. Hutter[1] & Teri A. Manolio[1]

Starting with the launch of the Human Genome Project three decades ago, and continuing after its completion in 2003, genomics has progressively come to have a central and catalytic role in basic and translational research. In addition, studies increasingly demonstrate how genomic information can be effectively used in clinical care. In the future, the anticipated advances in technology development, biological insights, and clinical applications (among others) will lead to more widespread integration of genomics into almost all areas of biomedical research, the adoption of genomics into mainstream medical and public-health practices, and an increasing relevance of genomics for everyday life. On behalf of the research community, the National Human Genome Research Institute recently completed a multi-year process of strategic engagement to identify future research priorities and opportunities in

https://www.nature.com/articles/s41586-020-2817-4

Focus areas in genomics

- Basic Genomics & Genomic Technologies
- Genomics of Disease
- Genomic Data Science
- Genomics in Medicine & Health
- Society, Education, & Engagement

Basic genomics & genomic technologies

- Develop approaches for routine end-to-end sequencing of the human genome
- Improve incorporation of multi-omic data into research projects
- Advance the use of model organisms for validating genome function
- Develop technologies for ‘rewriting’ genomes using synthetic biology
- How can we better predict phenotypic consequence of genomic variants, moving from single variants to multiple variants?
- How can we routinely annotate genome and epigenome data?
- What is the most efficient way to put genes, regulatory elements, and associated genomic variants into pathways?
- New areas of genomic technology development are needed

Genomics of disease

- Improve understanding of gene-environment interactions
- Establish better ways to connect genomic structural variants to human disease
- Advance ability to incorporate phenotypic data into genomic studies of human disease
- Increase ancestral diversity in studies examining the genomics of disease
- What steps are needed to create high-quality, well-phenotyped, ancestrally diverse datasets?
- How do we improve understanding of how pathways and regulatory networks influence disease?
- How can a comprehensive understanding of the genomic architecture of inherited disease be achieved?
- What non-genomic data types are important for understanding the connection between genomic variants and disease risk?

Genomic data science

- Make genomic data accessible and shareable
- Find an appropriate balance between access to genomic data, information security, and the privacy of individuals
- Encourage inter-agency, international, and industry collaborations
- Develop standard formats and guidelines for genomic data
- What are the open computational problems in genomics?
- How can we promote genomic data sharing in an era of democratized genome sequencing?
- How can we integrate genomic data science into clinical care?
- What are barriers to ensuring integrity, security, and confidentiality of genomic data?
- How can we promote data science expertise in genomics?

Genomics in medicine & health

- Improve the integration of genomic information into routine medical practice
- Build better knowledgebases for predictive genomic medicine
- Perform rigorous evaluations of genomic diagnostic and therapeutic strategies
- Ensure that genomic health information has utility for all
- How best to reimagine and standardize sampling, consenting, and return of results to allow routine genome sequencing?
- What is needed for the iterative use of genomic information as a lifetime healthcare resource?
- What knowledgebases are needed to link functional data about genomic variants to medical relevance?
- What are the most effective ways to ensure that the benefits of genomic medicine are shared by all?

Society, education & engagement

- Identify barriers to ensuring equitable access to genomic medicine
- Develop genomic technologies in concert with community needs and preferences
- Empower informed decision-making about an individual's genomic information
- Provide appropriate training opportunities for scientists and clinicians (especially early in their careers)
- How best to engage stakeholders to promote individuals' informed use of genomic and healthcare data?
- What is needed to help people make well-informed decisions about the use of their genomic information?
- What strategies are needed to create a diverse workforce in genomics?
- How best to assess progress in getting scientific and public understanding of the interplay of genomic, environmental, and contextual influences on health?

Thinking Quantitatively

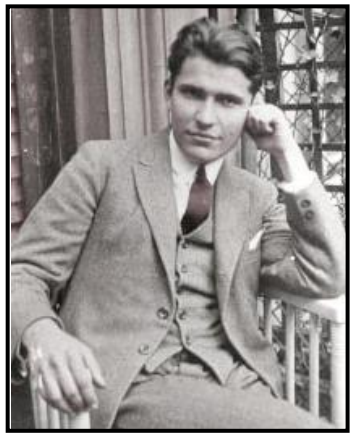
Biology is a Quantitative Science!!!

Gregor Mendel

1823-1884

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

Erwin Chargaff



1929-1992

Thinking Quantitatively

- Space
  - Be comprehensive
  - E.g., for rolling a die, S = {1,2,3,4,5,6}
- Signal to Noise Ratio
  - Sensitivity, specificity, dynamic range
  - What is my background control?
- Probability
- Distributions
  - Normal/Gaussian, Poisson, Binomial, Negative binomial, Multinomial, Extreme value, Hypergeometric, etc.
  - Discrete vs continuous
- The *P* value
- Bayes' rule
- Don't forget genetics!!!

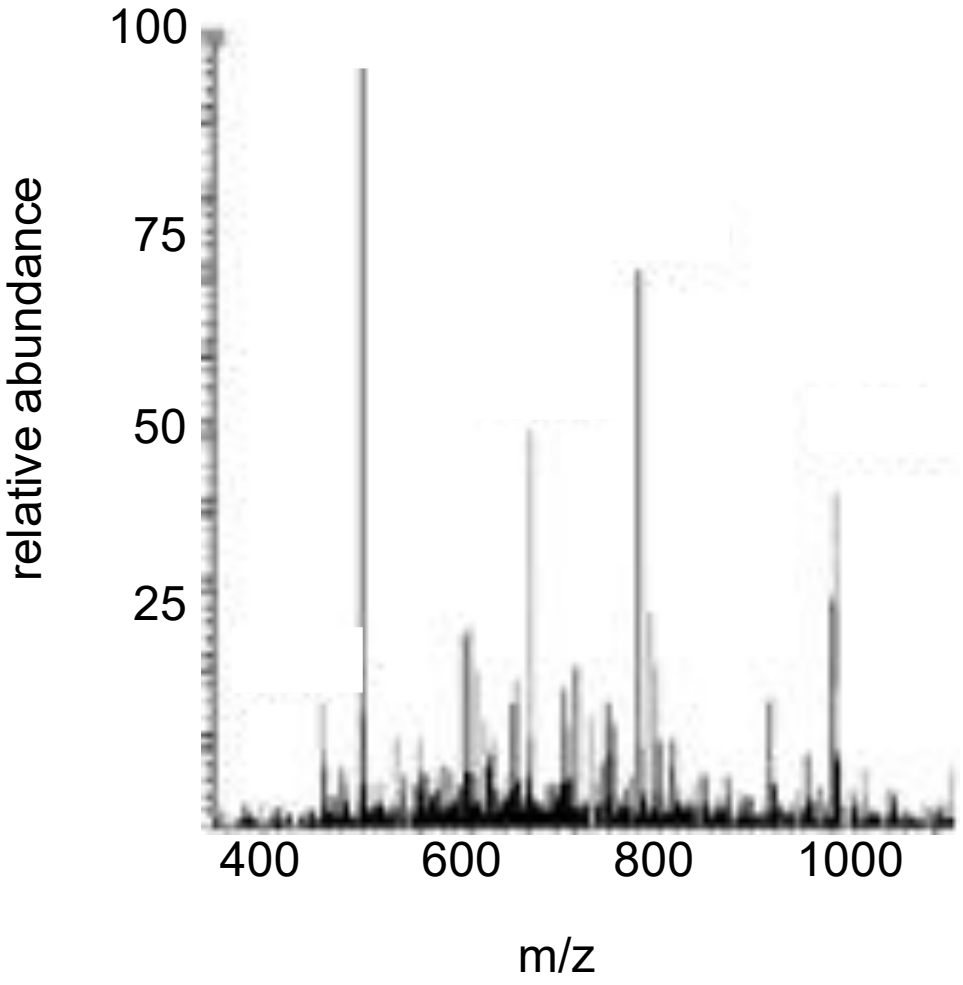
Bio 5075
Introduction to Coding and Statistical Thinking in Genetics and Genomics
White & Turner

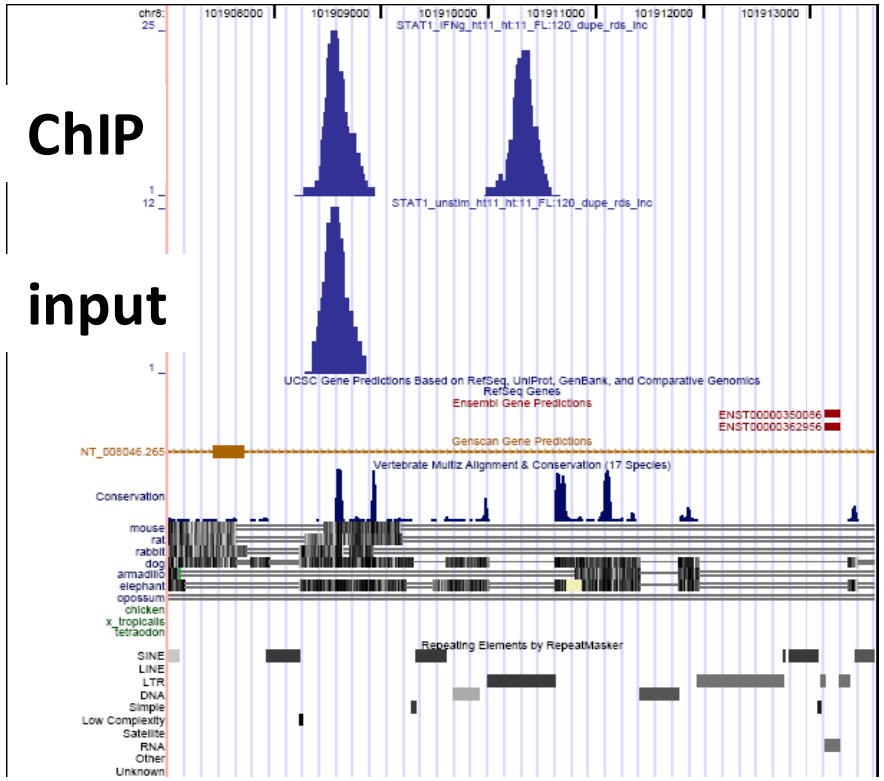
Simple principle:
what is your expectation?
what is your observation?

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)

Signal to Noise





Different sources of noise

Sensitivity and Specificity

- Sensitivity measures the proportion of positive test results out of all truly positive samples.
- Specificity measures the proportion of negative test results out of all truly negative samples.
- False results are also known as testing errors. The consequences of a testing error—a false positive or a false negative—are not equivalent

Example: Sensitivity and Specificity

	Disease	No Disease
Test Positive	True Positive (TP)	False Positive (FP)
Test Negative	False Negative (FN)	True Negative (TN)

Sensitivity = (TP) / (TP + FN)

Specificity = (TN) / (TN + FP)

	SARS-Cov-2 Positive	SARS-Cov-2 Negative
Test Positive	95	7
Test Negative	5	93

Sensitivity = 95 / (95 + 5) = 95%

Specificity = (93) / (93 + 7) = 93%

Probability Definitions and Notation

- Sample Space (Ω) = collection of possible outcomes of an experiment
  Example: die roll Ω = {1, 2, 3, 4, 5 ,6}
- Event (E) = subset of Ω
  Example: die roll is even E = {2, 4, 6}
- $\emptyset$ = null event or empty set
- A $\subset$ B = set A is a subset of set B, but not equal to set B
  Example: A = {1, 3}; B = {1, 3, 9}
- A $\cap$ B = both A and B occur
  Example: A $\cap$ B = {1, 3}
- A $\cup$ B = either A or B occurs
  Example: A $\cup$ B = {1, 3, 9}
- A $\cap$ B = $\emptyset$ => A and B are mutually exclusive

Probability Measure

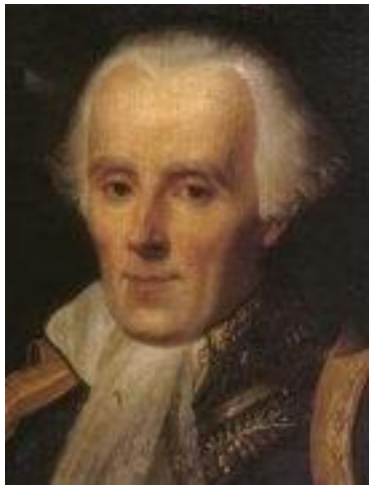
A probability measure ***P*** is a real valued function defined on the collection of possible events, satisfying the following conditions :

- For an event E $\subset \Omega$, $0 \leq P(E) \leq 1$
- $P(\Omega) = 1$
- If E1 and E2 are mutually exclusive events $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Slide credit: Brian Caffo

Probability

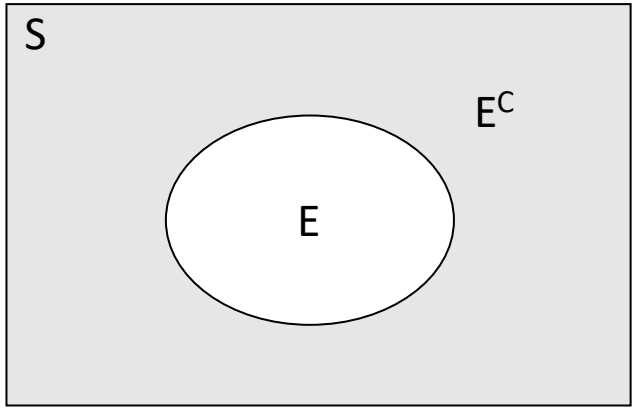
The numerical descriptions of the chance, or how likely an event is to occur. The probability of an event is a number between 0 and 1.

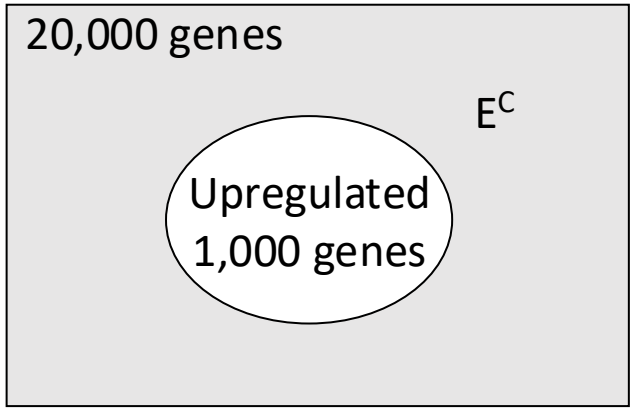


"The most important questions of life are …really only problems of probability."

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

$P(E)=E/S$
$P(E^c)=1-P(E)$





P(upregulated)
= 1,000/20,000
= 0.05

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?

$P(SP) = P(S)P(P) = 0.0688*0.0484 = 0.00329$

$P(SP^c) = 1 - P(SP) = 0.9966$

$P(\text{no SP anywhere in a 25 mer}) = P(SP^c)^{24} = 0.92$

$P(\text{at least one SP in a 25 mer}) = 1- P(SP^c)^{24} = 0.08$

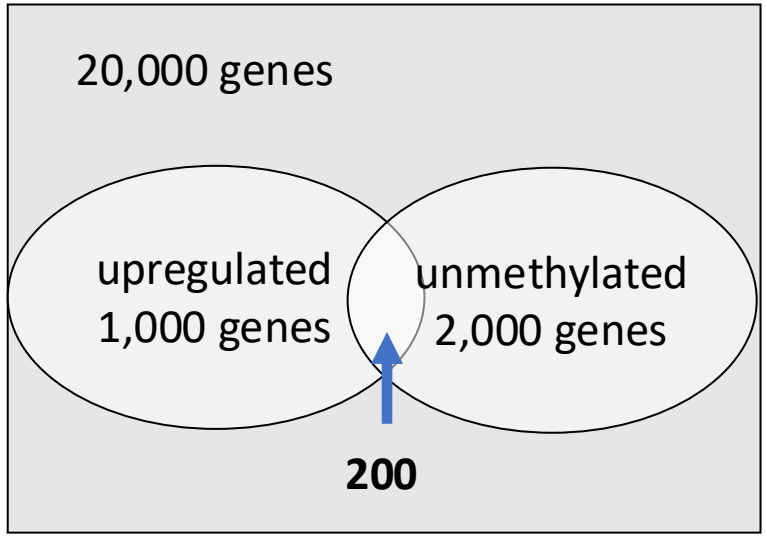
What is the probability that the last residue of a protein is either K or R?

$P(K)+P(R) = .0593 + .0532 = .11$

Conditional Probability

The probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) has already occurred.

P(A|B)



	Upregulated	Not upregulated	Row total
Unmethylated	**200**	**1800**	**2000**
Not unmethylated	**800**	**17200**	**18000**
Column total	**1000**	**19000**	**20000**

P(Upregulated) = 0.05
P(Unmethylated) = 0.1
P(Up and Un) = 0.01
(wait, why not 0.005??)

P(Up | Un) = 200/2000 = 0.1
P(Un | Up) = 200/1000 = 0.2

P(Up | Un) * P(Un) = 0.1*0.1 = 0.01
P(Un | Up) * P(Up) = 0.2*0.05 = 0.01

Derivative of Bayes' Rule

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

* $P(A \text{ and } B) = P(A \cap B)$

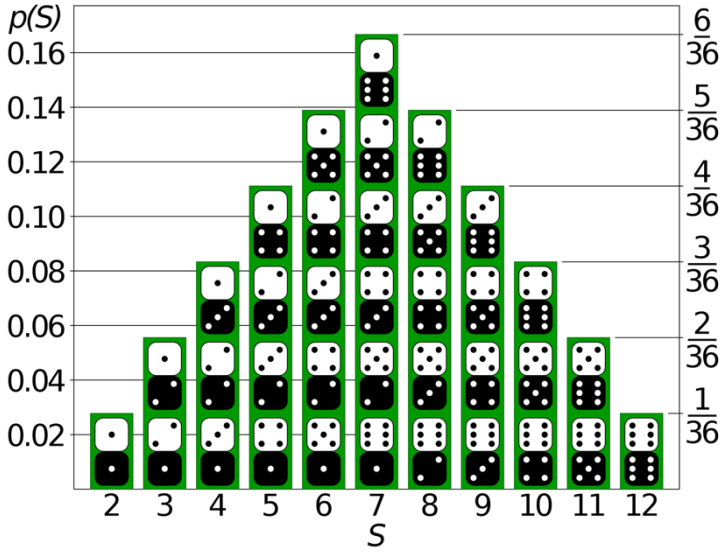
$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A) \times P(A)}{P(B)}$$

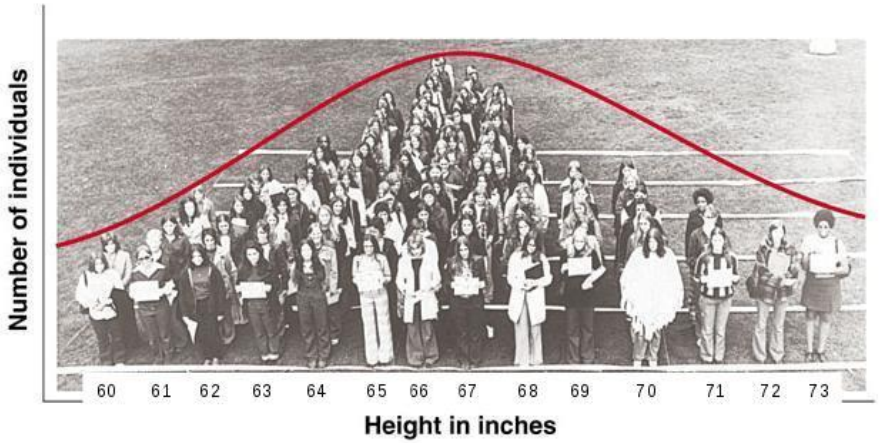
$$P(B|A) = \frac{P(B \text{ and } A)}{P(A)} = \frac{P(A \text{ and } B)}{P(A)}$$

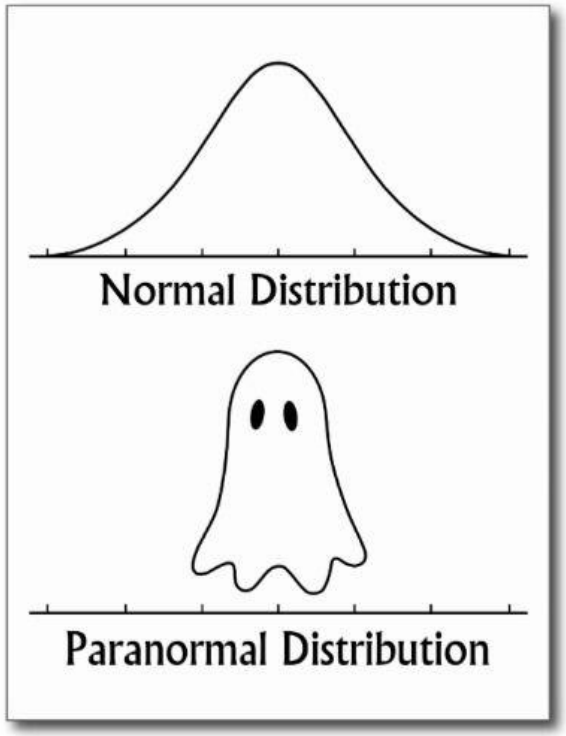
$$P(B|A) \times P(A) = P(A \text{ and } B)$$

Probability Distribution

The mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).







Random Variables

- Random variable: numerical outcome of an experiment
- Random variables could be discrete or continuous
- Discrete random variables can take on only a countable number of possible values. $P(X = k)$
- Continuous random variables can take on an infinite number of values $P(X \subseteq A)$

Slide credit: Brian Caffo

Example: Random Variables

- Discrete random variables:
  - The number of defective items in a batch of products
  - The number of customers arriving at a store in an hour

- Continuous random variables:
  - The weight of a newborn baby
  - The BMI of a subject for ten years after a baseline measurement

Probability Mass Function (PMF)

A PMF is a function over the sample space of a **discrete random variable *X*** which gives the probability that X is equal to a certain value.

Let X be a discrete random variable on a sample space *S*. Then the probability mass function f(x) is defined as

$$f(x) = P(X = x)$$

To be a valid PMF, the following conditions must hold:

- $f(x) \geq 0 \; for \; all \; x \in S$
- $\sum_{x \in S} f(x) = 1$

x	0	1	2	3	4	6
f(x)	0.05	0.1	0.2	0.4	0.2	0.05

https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/distribution-functions/probability-mass-function.html

Probability Density Function (PDF)

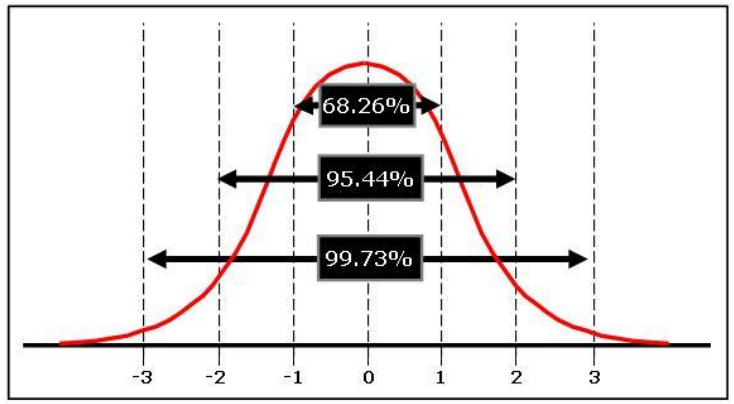
A PDF is a function associated with a **continuous random variable**

Area under PDFs correspond to probabilities for that random variable

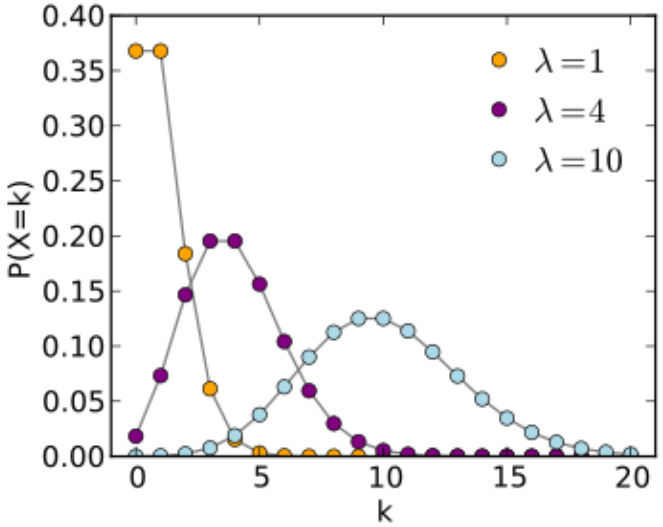
To be a valid PMF, the following conditions must hold:

- $f(x) \geq 0 \; for \; all \; x$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

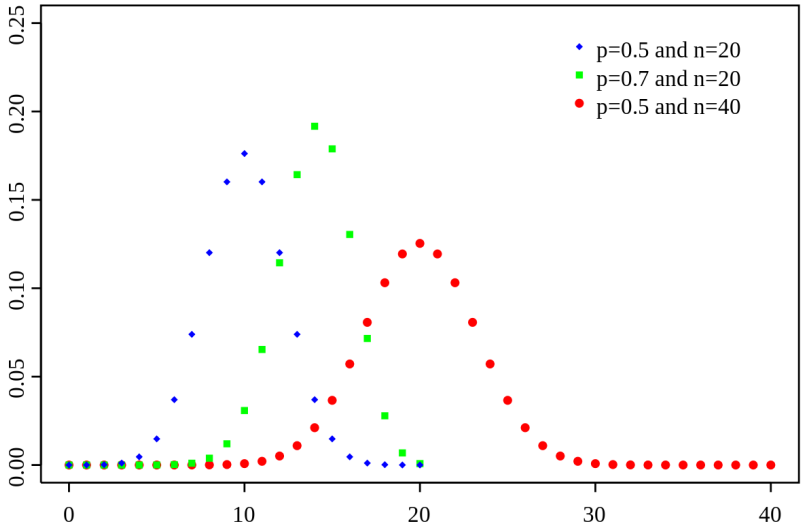
Slide credit: Brian Caffo



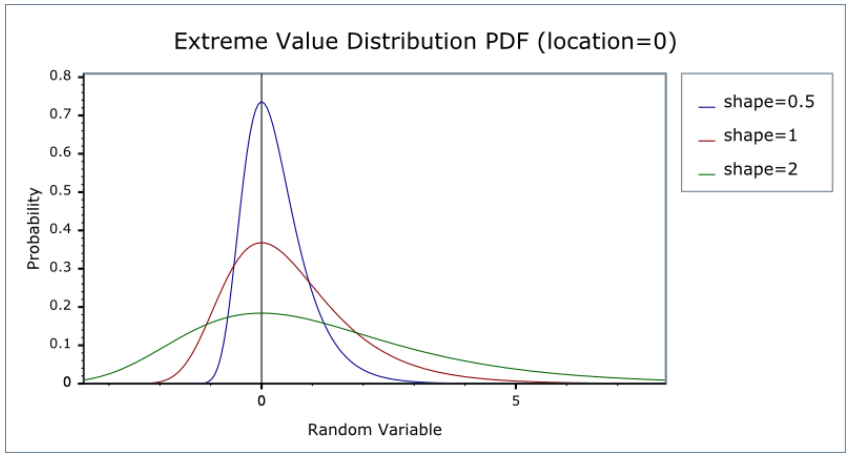
Normal (Gaussian)



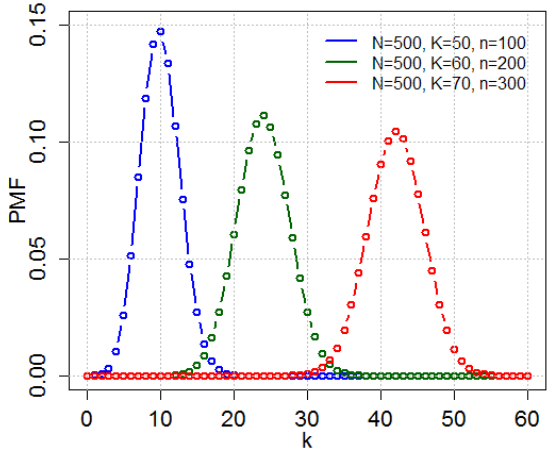
Poisson



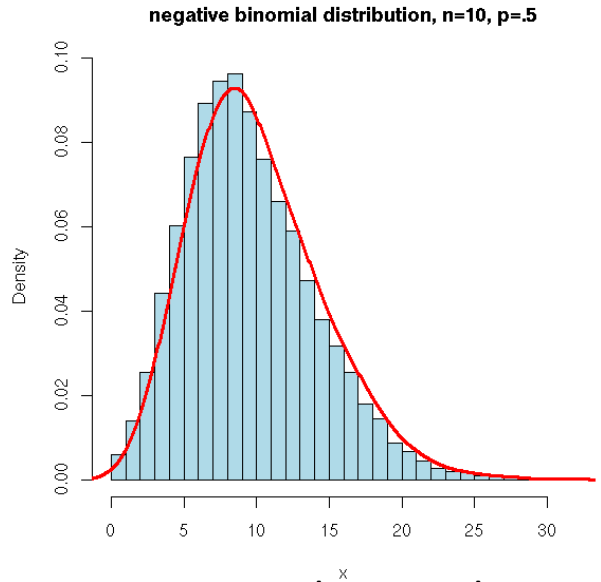
Binomial



Extreme value



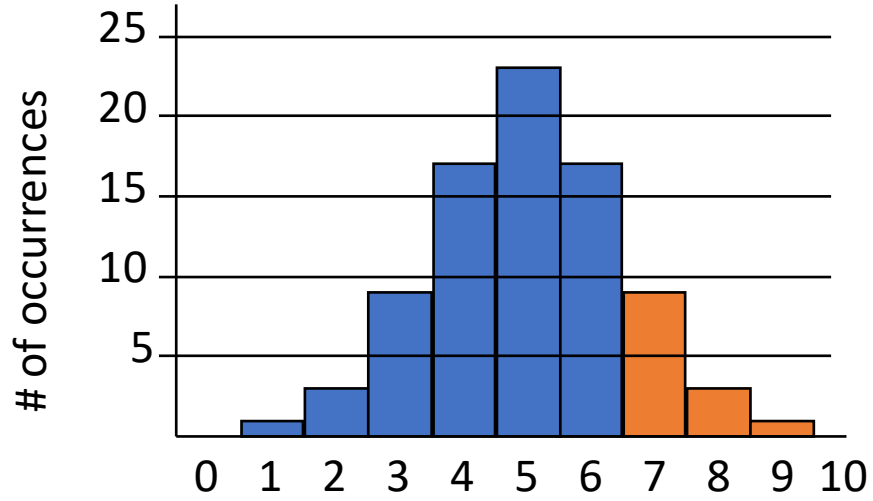
Hypergeometric



Negative binomial

The p-value

P-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.



What is the chance of getting seven or better?

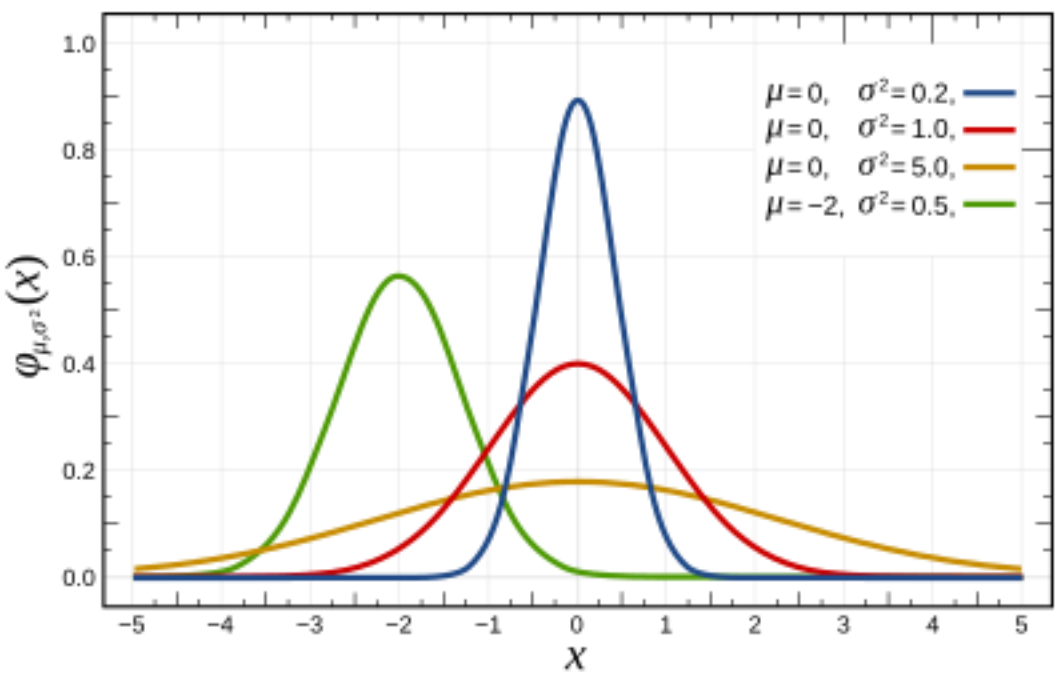
$$P = \frac{\#\,of\ trials\ \ that\ \ were\ \ seven\ \ or\ \ better}{total\ \#\ of\ trials}$$

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

$$P = 0.16$$

What is the probability of getting 5? What is the p-value of getting 5?

Normal (Gaussian) Distribution



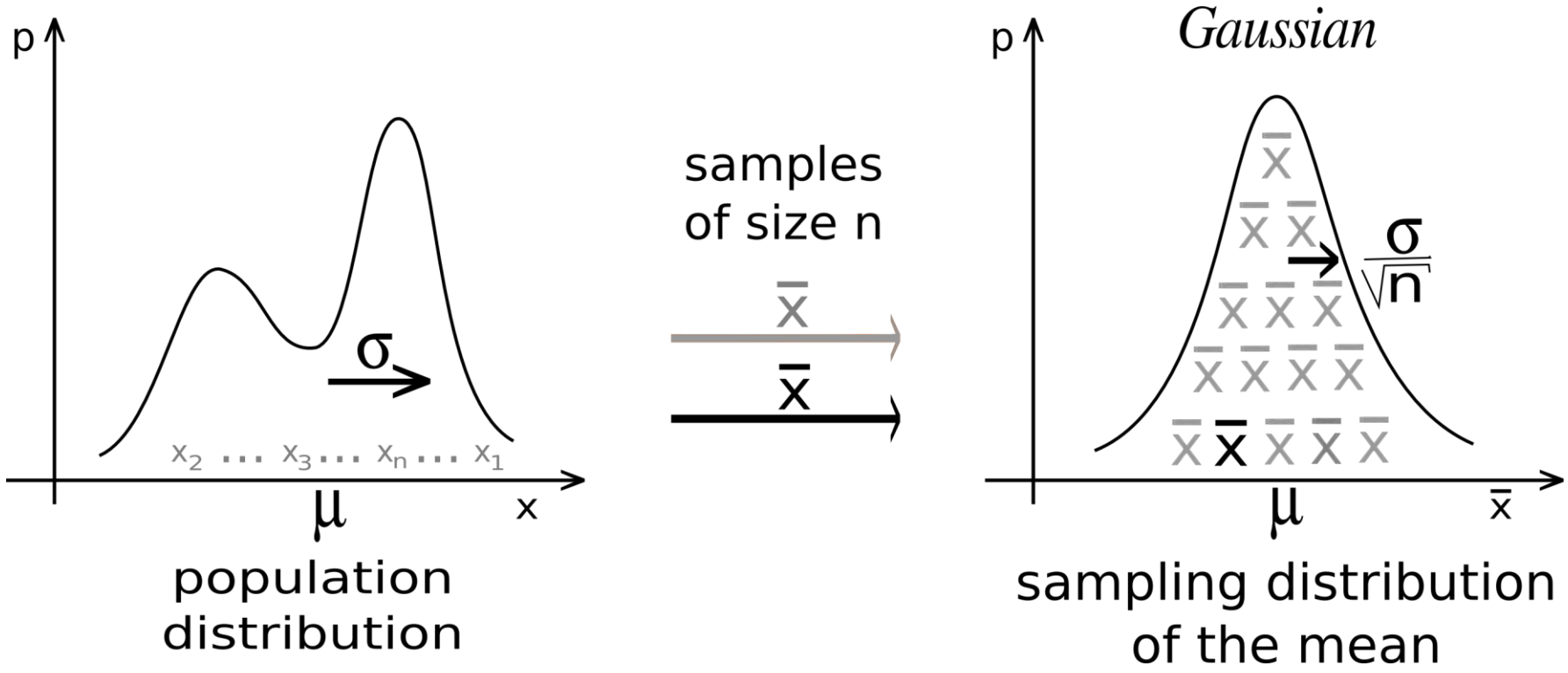
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

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

https://en.wikipedia.org/wiki/Normal_distribution

Central Limit Theorem

Under appropriate conditions, the distribution of a normalized version of the sample mean converges to a standard normal distribution. This holds even if the original variables themselves are not normally distributed.

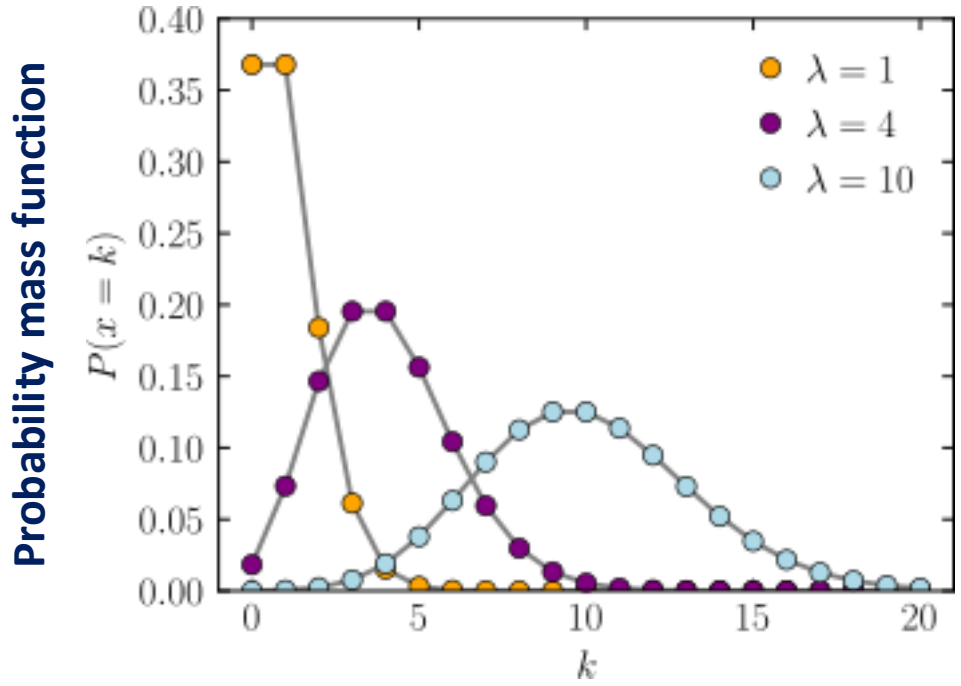


https://en.wikipedia.org/wiki/Central_limit_theorem

Terrific reference:
https://www.youtube.com/watch?v=YAlJCEDH2uY

Poisson Distribution

A discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event. In Poisson distribution, variance and mean are equal.



$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

https://en.wikipedia.org/wiki/Poisson_distribution

Example: Poisson Distribution

If we sequenced the genome to 10X coverage, what is the probability for a specific base to have been covered 5 times?

$P(5, 10) = 10^5 e^{-10}/5! = 0.038$ (3.8% bases are covered 5 times)

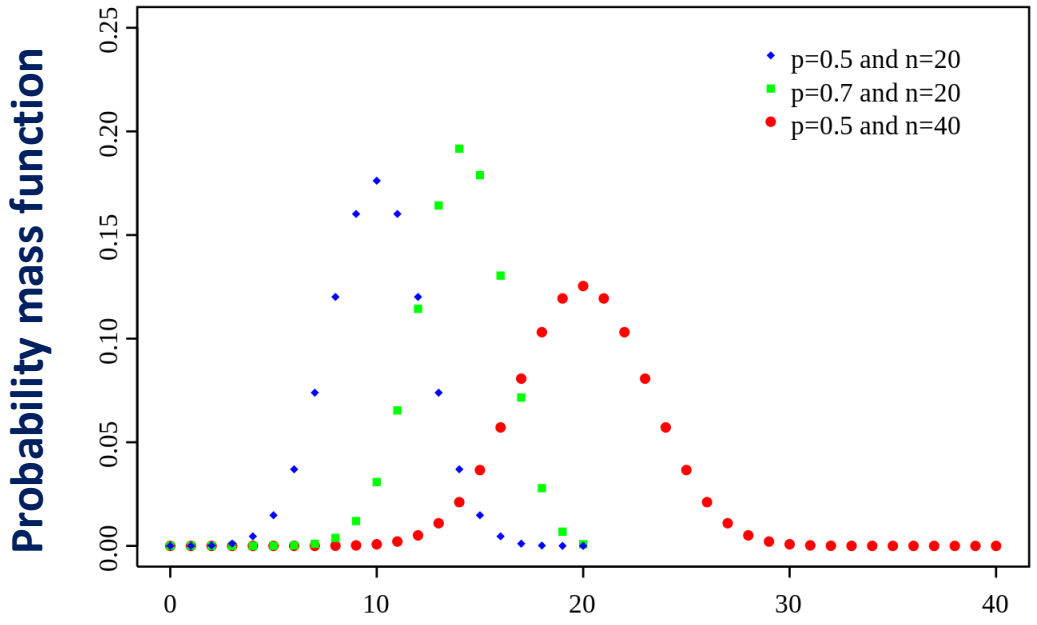
If we sequenced the genome to 10X coverage, what is the probability for a specific base to have been covered at least 5 times?

$P(5,10)+P(6,10)+\ldots+P(max,10) =$

$1-[P(0,10)+P(1,10)+P(2,10)+P(3,10)+P(4,10)] = 0.97$ (97% bases)

Binomial Distribution

A discrete probability distribution to describe the number of successes in a sequence of n independent experiments. The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N.



$$f(k, n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

https://en.wikipedia.org/wiki/Binomial_distribution

Example: Binomial Distribution

A CpG site is 50% methylated. Using bisulfite sequencing, I collected 4 reads showing methylated CpG, and 3 reads showing unmethylated CpG. What's the probability of seeing this?

$$P(4, 7, 0.5) = \binom{7}{4} 0.5^4 (1 - 0.5)^{7-4} = 0.273$$

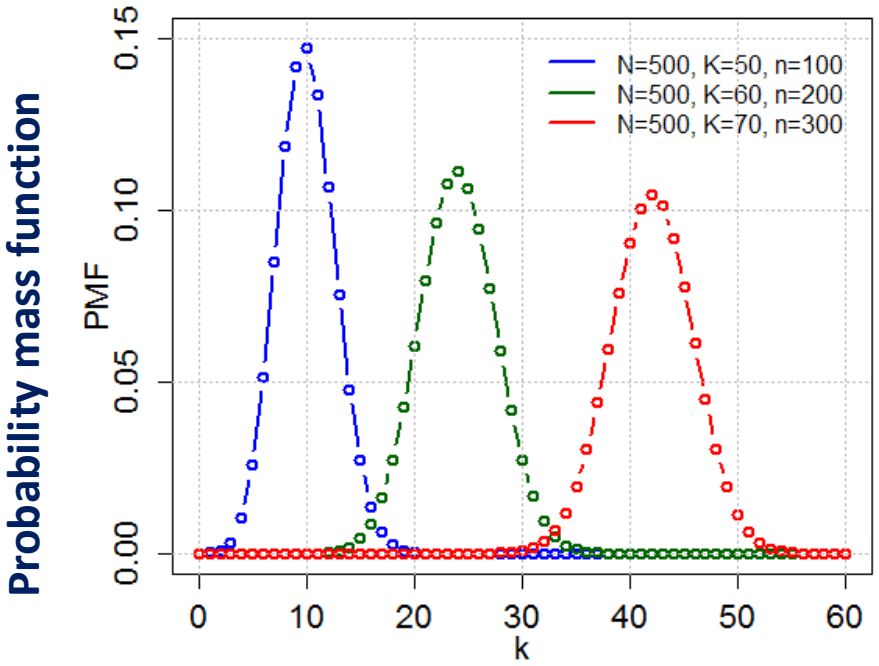
What is the p-value of seeing this?

Given the confidence interval we want (say, we want to be 95% confident about the estimated DNA methylation level), how does required sequencing coverage change as a function of DNA methylation level, using bisulfite sequencing?

Hypergeometric Distribution

A discrete probability distribution that describes the probability of k successes (random draws with a specified feature) in n draws, without replacement, from a finite population of size N containing exactly K objects with that feature. Each draw is either a success or a failure.

This is similar to the binomial distribution, with one key difference: *sampling is without replacement*. Consequently, the probability of success changes with each trial, unlike in the binomial distribution where the probabilities of success and failure remain constant.



$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where

- N is the population size,
- K is the number of success states in the population,
- n is the number of draws (i.e. quantity drawn in each trial),
- k is the number of observed successes,
- $\binom{a}{b}$ is a binomial coefficient.

https://en.wikipedia.org/wiki/Hypergeometric_distribution

Example: Hypergeometric Distribution

What's the probability of drawing all 4 kings in a random sample of 20 cards from a standard deck of cards?

N = 52: Number of cards in the deck
n = 20: Number of cards we sample
K = 4: Number of kings in the deck
k = 4: Number of kings we want (successes)

$$P(X = 4) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\binom{4}{4}\binom{52-4}{20-4}}{\binom{52}{20}} \approx 0.0179$$