Outline

• Outline of the course
• What is genomics?
• A little history
• The simple principles of genomics
  • Technologies, bioinformatics, big data
  • How to solve a problem using genomics?
• Being quantitative
  • Probability, biostatistics, machine learning
• From a student to an investigator
• Focus areas of genomics in the near future
A few MTE administrivia...

• If you didn’t receive an email from genomics.bio5488@gmail.com this week, please email genomics.bio5488@gmail.com or talk to a 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 Web Site

- [http://www.genetics.wustl.edu/bio5488/](http://www.genetics.wustl.edu/bio5488/)
  - Linux Primer
  - Python Primer
  - Lecture Notes
  - Schedule
  - Weekly Assignments and Answers
  - Weekly Readings

- Canvas
- Piazza
## Grading

<table>
<thead>
<tr>
<th>4 credit</th>
<th>3 credit</th>
<th>Audit/sit-in</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 25% midterm</td>
<td>• 25% final</td>
<td>• 50% midterm</td>
</tr>
<tr>
<td>• 25% final</td>
<td>• 50% final</td>
<td>• 50% final</td>
</tr>
</tbody>
</table>

**What is the key to your success?**
# Genetics – BIO5488

## Syllabus 2024

### Weekly Schedule

<table>
<thead>
<tr>
<th>Day</th>
<th>Lecture/Lab</th>
<th>Lecturer</th>
<th>Assignment due</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/22/24</td>
<td>Mon</td>
<td>Genomics basics II</td>
<td>Assignment 1 2023 Lab 1 Slides</td>
</tr>
</tbody>
</table>

Please check Canvas and Piazza if you have any questions on class, assignments, exams etc.

### Course Information

- **Professor:** Wang
- **TAs:** Jin, Lawon, White, Buchser, TAs, TAs

### Materials

- Reading:
  - Jindal et al., 2022
  - Astley et al., 1996
- Slides:
  - Lecture Slides
  - Supplemental

### Assignments

- Assignment 1: 2023 Lab 1 Slides
- Assignment 2
- Assignment 3
- Assignment 4
- Assignment 5
- Assignment 6
- Assignment 7
- Assignment 8
- Assignment 9
- Assignment 10
- Assignment 11
- Assignment 12
- Assignment 13

### Exams

- Midterm Exam
- Final Exam

### Additional Resources

- Websites:
  - Washington University School of Medicine in St. Louis

---

**Note:** The schedule and assignments are subject to change.
A little history
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  Needleman-Wunsch algorithm for sequence alignment
1978  Sanger: DNA sequencing
1977  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 influenzae 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

90's
- HGP
  - Computational Biology
  - Sequence analysis
  - Hidden Markov Model
  - Gene finding
  - BLAT
  - Genome Browser
  - Motif finding
  - Assembly

00's
- Microarray Omics
  - Gene expression

10's
- Next-gen Sequencing
  - ENCODE
  - GWAS
  - Roadmap etc
- 3rd-gen Sequencing
- Single cell

10's
- Comparative Genomics
  - Evolution

- Systems/Synthetic Biology

- GTEx
- 4DN
- Human Pangenome

- Machine Learning
  - Data mining
  - Structural informatics
  - Drug Design
  - Statistical Modeling
  - Database

- 90's
- 00's
- 10's
- 3rd-gen Sequencing
- Human Pangenome
- GTEx
- 4DN
- Single cell
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
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!!!
Motivation slides
First Draft of the Human Genome Sequence Released

1000 Genome Project

The Human Heredity and Health in Africa (H3Africa)

Electronic Medical Records and Genomics (eMERGE)

Simons Diversity Study

Population Architecture using Genomics and Epidemiology (PAGE)

Implementing Genomics in Practice (IGNITE)

International HapMap Project

Human Genome Project Completed

Applicants Accepted: UK Biobank

Trans-Omics for Precision Medicine (TOPMed) Program

GWAS Catalog

Global Alliance for Genomics and Health (GA4GH)

Launch USA All of Us Biobank

Launch of the Human Pangenome Reference Consortium

European 57%

African 37%

East Asian 6%

Unknown 13%

European 78%

African 2.4%

Latin American 1.3%

Multiple 1.8%

East Asian 9%

Other Asian 2%

NR 6%

European 57%

African 37%

East Asian 6%

Unknown 13%

European 78%

African 2.4%

Latin American 1.3%

Multiple 1.8%

East Asian 9%

Other Asian 2%

NR 6%

Lorem ipsum

NR 6%

East Asian (9%)

Other Asian (2%)

Latin American 1.3%

African 2.4%

Multiple 1.8%

European 78%

African 2.4%

Latin American 1.3%

Multiple 1.8%
Cost per Human Genome

Moore’s Law

NIH National Human Genome Research Institute
genome.gov/sequencingcosts
The Human genome: the “blueprint” of our body

$10^{13}$ 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

Every cell has the same sequence of DNA.

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?

Gene numbers do not correlate with organism complexity. Many gene families are surprisingly old.
## Complexity, Genome Size and the C-value Paradox

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

**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

What do they do?
Ultra conserved elements

Chromosome Bands Localized by FISH Mapping Clones

13q21.33

RefSeq Genes

Human/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons

Conservation

mouse
rat
chicken

ultra conserved

100% hg19-mm3-m3 =>2000bp
c1000.351

Chromosome Bands Localized by FISH Mapping Clones

13q21.33

RefSeq Genes

Human/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons

Conservation

c1000.351

100% hg19-mm3-m3 =>2000bp

e.d 12.5
**HARs: Human accelerated regions**

<table>
<thead>
<tr>
<th>position</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>AGAC</td>
<td>TTAC</td>
<td>AGCA</td>
<td>ACCTG</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>gorilla</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>orangutan</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>macaque</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>mouse</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>dog</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>cow</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>platypus</td>
<td>ATAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>opossum</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
<tr>
<td>chicken</td>
<td>AGAA</td>
<td>ATTAC</td>
<td>AGCA</td>
<td>ATGGT</td>
</tr>
</tbody>
</table>

- 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

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”
2011-Present
En Route to Genomic Medicine
Strategic Vision for improving human health at The Forefront of Genomics

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...
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!!!

1) Mendel’s Laws
2) Chargaff’s Rules

Gregor Mendel
1823-1884

Erwin Chargaff
1929-1992
Thinking Quantitatively

• Space
  • Be comprehensive
• 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
Fundamentals of Biostatistics
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, 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?
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)

\[ P(E) = \frac{E}{S} \]

\[ P(E^c) = 1 - P(E) \]

\[ P(\text{upregulated}) = \frac{1,000}{20,000} = 0.05 \]
Example: Probability

Amino acid percentages of Swissprot

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

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

\[
P(\text{SP}) = P(S)P(P) = 0.0688 \times 0.0484 = 0.00329
\]
\[
P(\text{SP}^c) = 1 - P(\text{SP}) = 0.9966
\]
\[
P(\text{no SP anywhere in a 25 mer}) = P(\text{SP}^c)^{24} = 0.92
\]
\[
P(\text{at least one SP in a 25 mer}) = 1 - P(\text{SP}^c)^{24} = 0.08
\]

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

\[
P(K) + P(R) = 0.0593 + 0.0532 = 0.11
\]
Conditional Probability

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

<table>
<thead>
<tr>
<th></th>
<th>Upregulated</th>
<th>Not upregulated</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmethylated</td>
<td>200</td>
<td>1800</td>
<td>2000</td>
</tr>
<tr>
<td>Not unmethylated</td>
<td>800</td>
<td>17200</td>
<td>18000</td>
</tr>
<tr>
<td>Column total</td>
<td>1000</td>
<td>19000</td>
<td>20000</td>
</tr>
</tbody>
</table>

\[ P(\text{Upregulated}) = 0.05 \]
\[ P(\text{Unmethylated}) = 0.1 \]
\[ P(\text{Up and Un}) = 0.01 \]

\[ P(\text{Up} | \text{Un}) = \frac{200}{2000} = 0.1 \]
\[ P(\text{Un} | \text{Up}) = \frac{200}{1000} = 0.2 \]

\[ P(\text{Up} | \text{Un}) \times P(\text{Un}) = 0.1 \times 0.1 = 0.01 \]
\[ P(\text{Un} | \text{Up}) \times P(\text{Up}) = 0.2 \times 0.05 = 0.01 \]
Independent Events/Mutually Exclusive Events

What is the probability of A and B both occurring?
\[
P(AB) = P(A) \times P(B) \quad \text{(Independent)}
\]
\[
P(A|B) = P(A) \quad \text{(Independent)}
\]
\[
P(A|B) \times P(B) = P(B|A) \times P(A) \quad \text{(Bayes rule)}
\]
\[
P(AB) = P(A|B) = 0 \quad \text{(Mutually Exclusive)}
\]

What is the probability of A or B occurring (independent)?
\[
P(A) \times P(B) + P(A) \times 1 - P(B) + 1 - P(A) \times P(B)
\]
both + A only + B only

What is the probability of A or B occurring (mutually exclusive)?
\[
P(A) + P(B)
\]
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).
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?

\[ P = \frac{\text{# of trials that were seven or better}}{\text{total # of trials}} \]

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

\[ P = 0.16 \]

What is the probability of getting 5? What is the p-value of getting 5?
Gaussian (Normal) Distribution
(The Central Limit Theorem)

- Mean ($x$ or $\mu$)
- Standard Deviation ($s$ or $\sigma$)
- Compare means ($t$-tests)
- Compare standard deviations ($f$-tests)
- Calculate $P$-values for particular results ($z$-tests)

Expression level of a single gene

$$\text{Mean (x)} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

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

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

\[ f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \]
**Poisson Distribution**

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

\[
P(5, 10) = \frac{10^5 e^{-10}}{5!} = 0.038 \text{ (3.8% bases are covered 5 times)}
\]

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

\[
P(5,10)+P(6,10)+\ldots+P(\text{max,10}) = 1-[P(0,10)+P(1,10)+P(2,10)+P(3,10)+P(4,10)] = 0.97 \text{ (97% bases)}
\]
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 \( N \).

\[
f(k, n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}
\]
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?
A discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws, without replacement, from a finite population of size N that contains exactly K objects with that feature, wherein each draw is either a success or a failure.

\[
P(X \geq s) = \sum_{i=s}^{n} \binom{m}{i} \binom{N-m}{n-i} \cdot \binom{N}{n}
\]

where \( \binom{a}{b} = \frac{a!}{b!(a-b)!} \)
Hypergeometric Probability Distribution

\[ P(X \geq s) = \sum_{i=s}^{n} \binom{m}{i} \binom{N-m}{n-i} \binom{N}{n} \]

where

\[ \binom{a}{b} = \frac{a!}{b!(a-b)!} \]

\[ P(\text{intersection} \geq 200) = 10^{-23} \]
Example: Sensitivity and specificity

Natural immune to covid-19: 1%
Test sensitivity/specificity: 99%
Question: is the test worth taking?
Answer:

\[ P(\text{Immune} \mid \text{positive test}) \]

\[
P(\text{Imm}) = 1\%, \ P(\text{notImm}) = 99%
\]
\[
\text{SEN} = (+\mid\text{Imm}) = 0.99
\]
\[
\text{SPE} = (-\mid\text{notImm}) = 0.99
\]

True Negative = \( P(\text{notImm} ^ {-}) = P(\text{notImm}) \cdot P(-\mid\text{notImmune}) = 0.99 \cdot 0.99 \)

True Positive = \( P(\text{Imm} ^ {+}) = P(\text{Imm}) \cdot P(+\mid\text{Imm}) = 0.01 \cdot 0.99 \)

False Positive = \( P(\text{notImm} ^ {+}) = P(\text{notImm}) \cdot P(+\mid\text{notImm}) = 0.99 \cdot 0.01 \)

False Negative = \( P(\text{Imm} ^ {-}) = P(\text{Imm}) \cdot P(-\mid\text{Imm}) = 0.01 \cdot 0.01 \)

\[
P(+) = P(+ ^ {\text{Imm}}) + P(+ ^ {\text{notImm}}) = 0.01 \cdot 0.99 + 0.99 + 0.01
\]

\[
P(\text{Imm}\mid+) = P(\text{Imm} ^ {+})/P(+) = 0.01 \cdot 0.99 / 0.02 \cdot 0.99 = 50%
\]