Bio5488
Genomics
Spring, 2022

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

Location: Wherever your virtual background shows
Hopefully we’ll move to Holden in February
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
B

desired goal

A

B

the cloud

A

C

reality
A few TA–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 TA 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 (well, zoom)
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

4 credit
• 25% midterm
• 25% final
• 50% weekly assignments

3 credit
• 50% midterm
• 50% final

Audit/sit-in

What is the key to your success?
<table>
<thead>
<tr>
<th>Date</th>
<th>Day</th>
<th>Lecture/Lab</th>
<th>Lecturer</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/19/22</td>
<td>Wed</td>
<td>Genomics basics I</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>1/21/22</td>
<td>Fri</td>
<td>LAB 1: Introduction of statistical models and Python programming</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/24/22</td>
<td>Mon</td>
<td>Genomics basics II</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>1/26/22</td>
<td>Wed</td>
<td>Sequencing technology</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>1/28/22</td>
<td>Fri</td>
<td>LAB 2: Sequence technology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/31/22</td>
<td>Mon</td>
<td>Homology I</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>2/2/22</td>
<td>Wed</td>
<td>Homology II</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>2/4/22</td>
<td>Fri</td>
<td>LAB 3: Sequence comparison</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/7/22</td>
<td>Mon</td>
<td>Gene expression I</td>
<td>Lawson</td>
<td></td>
</tr>
<tr>
<td>2/9/22</td>
<td>Wed</td>
<td>Gene expression II</td>
<td>Lawson</td>
<td></td>
</tr>
<tr>
<td>2/11/22</td>
<td>Fri</td>
<td>LAB 4: Gene expression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/14/22</td>
<td>Mon</td>
<td>Epigenomics I</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>2/16/22</td>
<td>Wed</td>
<td>Epigenomics II</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>2/18/22</td>
<td>Fri</td>
<td>LAB 5: Epigenomics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/21/22</td>
<td>Mon</td>
<td>4D genome I</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>2/23/22</td>
<td>Wed</td>
<td>4D genome II</td>
<td>Wang</td>
<td></td>
</tr>
<tr>
<td>2/25/22</td>
<td>Fri</td>
<td>LAB 6: Long range interaction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/28/22</td>
<td>Mon</td>
<td>Single cell genomics I</td>
<td>Morris</td>
<td></td>
</tr>
<tr>
<td>3/2/22</td>
<td>Wed</td>
<td>Single cell genomics II</td>
<td>Morris</td>
<td></td>
</tr>
<tr>
<td>3/4/22</td>
<td>Fri</td>
<td>Lab 7: Single cell analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/7/22</td>
<td>Mon</td>
<td>Genetic variation I</td>
<td>Jin</td>
<td></td>
</tr>
<tr>
<td>3/9/22</td>
<td>Wed</td>
<td>Genetic variation II</td>
<td>Jin</td>
<td></td>
</tr>
<tr>
<td>3/11/22</td>
<td>Fri</td>
<td>MIDTERM EXAM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/14/22</td>
<td>Mon</td>
<td>Spring Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/16/22</td>
<td>Wed</td>
<td>Spring Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/18/22</td>
<td>Fri</td>
<td>Spring Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/21/22</td>
<td>Mon</td>
<td>Population genetics I</td>
<td>Jin</td>
<td></td>
</tr>
<tr>
<td>3/23/22</td>
<td>Wed</td>
<td>Population genetics II</td>
<td>Jin</td>
<td></td>
</tr>
<tr>
<td>3/25/22</td>
<td>Fri</td>
<td>LAB 9: Wright-Fisher Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/28/22</td>
<td>Mon</td>
<td>Metagenomics I</td>
<td>Baldrige</td>
<td></td>
</tr>
<tr>
<td>3/30/22</td>
<td>Wed</td>
<td>Metagenomics II</td>
<td>Baldrige</td>
<td></td>
</tr>
<tr>
<td>4/1/22</td>
<td>Fri</td>
<td>LAB 10: Metagenomics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/4/22</td>
<td>Mon</td>
<td>Functional assays I</td>
<td>White</td>
<td></td>
</tr>
<tr>
<td>4/6/22</td>
<td>Wed</td>
<td>Functional assays II</td>
<td>White</td>
<td></td>
</tr>
<tr>
<td>4/8/22</td>
<td>Fri</td>
<td>LAB 11: Identify Expression-Modulating Variants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/11/22</td>
<td>Mon</td>
<td>Variant effect analysis I</td>
<td>White</td>
<td></td>
</tr>
<tr>
<td>4/13/22</td>
<td>Wed</td>
<td>Variant effect analysis II</td>
<td>White</td>
<td></td>
</tr>
<tr>
<td>4/15/22</td>
<td>Fri</td>
<td>LAB 12: Synthetic gene assembly</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/18/22</td>
<td>Mon</td>
<td>Comparative genomics I</td>
<td>Buchser</td>
<td></td>
</tr>
<tr>
<td>4/20/22</td>
<td>Wed</td>
<td>Comparative genomics II</td>
<td>Buchser</td>
<td></td>
</tr>
<tr>
<td>4/22/22</td>
<td>Fri</td>
<td>LAB 13: Cis-regulatory evolution</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A little history
History of Bio5488

- **HGP**
- **1998** Bio5495 Eddy
- **2003** Bio5488 Cohen Mitra
- **2006** Bio5495 Brent
- **2012** Bio5488 Wang Conrad
- **2022** Bio5488 Wang

- **Computational Biology**
- **Microarray**
- **Systems/Synthetic Biology**
- **3rd-gen Sequencing**
- **Next-gen Sequencing**
- **ENCODE**
- **GWAS**
- **Roadmap** etc
### History of Genomics

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1865</td>
<td>Gregor Mendel: founding of genetics</td>
</tr>
<tr>
<td>1953</td>
<td>Watson and Crick: double helix model for DNA</td>
</tr>
<tr>
<td>1955</td>
<td>Sanger: first protein sequence, bovine insulin</td>
</tr>
<tr>
<td>1970</td>
<td>Needleman-Wunsch algorithm for sequence alignment</td>
</tr>
<tr>
<td>1977</td>
<td>Sanger: DNA sequencing</td>
</tr>
<tr>
<td>1978</td>
<td>The term “bioinformatics” appeared for the first time</td>
</tr>
<tr>
<td>1980</td>
<td>The first complete gene sequence (Bacteriophage FX174), 5386 bp</td>
</tr>
<tr>
<td>1981</td>
<td>Smith-Waterman algorithm for sequence alignment</td>
</tr>
<tr>
<td>1981</td>
<td>IBM: first Personal Computer</td>
</tr>
<tr>
<td>1983</td>
<td>Kary Mullis: PCR</td>
</tr>
<tr>
<td>1986</td>
<td>The term &quot;Genomics&quot; appeared for the first time: name of a journal</td>
</tr>
<tr>
<td>1986</td>
<td>The SWISS-PROT database is released for the first time</td>
</tr>
<tr>
<td>1987</td>
<td>Perl (Practical Extraction Report Language) is released by Larry Wall.</td>
</tr>
<tr>
<td>1990</td>
<td>BLAST is published</td>
</tr>
<tr>
<td>1995</td>
<td>The <em>Haemophilus influenzae</em> genome (1.8 Mb) is sequenced</td>
</tr>
<tr>
<td>1996</td>
<td>Affymetrix produces the first commercial DNA chips</td>
</tr>
<tr>
<td>2001</td>
<td>A draft of the human genome (3,000 Mbp) is published</td>
</tr>
</tbody>
</table>
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
- GTEx
- 4DN

- Comparative Genomics
- Evolution
- Systems/Synthetic Biology
- Human Pangenome
- Machine Learning
- Data mining
- Structural informatics
- Drug Design
- Statistical Modeling
- Database
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

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%

RPC1-11 65%

RPC1-5 3%

RPC1-13 2%

CTC 2%

CTD 2%

European 78%

African 2.4%

Latin American 1.3%

Asian 9%

Multiple 1.8%

East Asian 9%

Other Asian 2%

NR 6%
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
What makes us different?

Differences between individuals?

Differences between species?
How many genes do we have?

What we used to think

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 (1 pg = 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
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

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?
• 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