Bio5488
Genomics
Spring, 2018

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

4th floor classroom
4515 Couch Medical Research Building
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:
  – Read assignment 1
  – Attempt to install the required software
  – Bring your laptop to class on Friday
Workshop this afternoon

• The WashU Epigenome Browser

• A practice run for a workshop @ Keystone Symposium on DNA and RNA methylation (Vancouver, 1/22/18)

• Instructors:
  – Renee Sears
  – Josh Jang

• 1-3pm, 4001B
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?
<table>
<thead>
<tr>
<th>Date</th>
<th>Day</th>
<th>Lecture/Lab</th>
<th>Lecturer</th>
<th>Notes</th>
<th>Assignment due</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/18</td>
<td>Mon</td>
<td>Martin Luther King holiday</td>
<td>Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/19</td>
<td>Wed</td>
<td>Course Introduction</td>
<td>Conrad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/21</td>
<td>Fri</td>
<td>LAB 1: Introduction of statistical models and Python programming</td>
<td>Taka</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/26</td>
<td>Mon</td>
<td>Sequencing technology I</td>
<td>Nitros</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/28</td>
<td>Wed</td>
<td>Sequencing technology II</td>
<td>Nitros</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/30</td>
<td>Fri</td>
<td>LAB 2: Sequence technology</td>
<td>Nitros</td>
<td></td>
<td>LAB 1: Introduction of statistical models and Python programming</td>
</tr>
<tr>
<td>2/6</td>
<td>Mon</td>
<td>Homology I</td>
<td>Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/8</td>
<td>Wed</td>
<td>Homology II</td>
<td>Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/10</td>
<td>Fri</td>
<td>LAB 3: Sequence comparison</td>
<td>Lawson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/15</td>
<td>Mon</td>
<td>Gene expression I</td>
<td>Lawson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/17</td>
<td>Wed</td>
<td>Gene expression II</td>
<td>Lawson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/19</td>
<td>Fri</td>
<td>LAB 4: Gene expression</td>
<td>Lawson</td>
<td></td>
<td>LAB 2: Sequencing technology</td>
</tr>
<tr>
<td>2/26</td>
<td>Mon</td>
<td>Epigenomics I</td>
<td>Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2/28</td>
<td>Wed</td>
<td>Epigenomics II</td>
<td>Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/1</td>
<td>Fri</td>
<td>LAB 5: Epigenomics</td>
<td>Wang</td>
<td></td>
<td>LAB 4: Gene expression</td>
</tr>
<tr>
<td>3/6</td>
<td>Mon</td>
<td>DNA binding factors</td>
<td>Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/8</td>
<td>Wed</td>
<td>3D genome</td>
<td>Wang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/10</td>
<td>Fri</td>
<td>LAB 6: motif finding</td>
<td>Wang</td>
<td></td>
<td>LAB 5: Epigenomics</td>
</tr>
<tr>
<td>3/13</td>
<td>Mon</td>
<td>Single cell genomics I</td>
<td>Conrad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/15</td>
<td>Wed</td>
<td>Single cell genomics II</td>
<td>Conrad</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/17</td>
<td>Fri</td>
<td>Lab 7: Single cell analysis</td>
<td>Conrad</td>
<td></td>
<td>LAB 6: motif finding</td>
</tr>
<tr>
<td>3/20</td>
<td>Mon</td>
<td>Metagenomics I</td>
<td>Dantas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/22</td>
<td>Wed</td>
<td>Metagenomics II</td>
<td>Dantas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/23</td>
<td>Fri</td>
<td>MIDTERM EXAM</td>
<td>Lab 8: Metagenomics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/26</td>
<td>Mon</td>
<td>Spring break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/28</td>
<td>Wed</td>
<td>Spring break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3/30</td>
<td>Fri</td>
<td>Spring break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/2</td>
<td>Mon</td>
<td>Genetic variation I</td>
<td>Hall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/4</td>
<td>Wed</td>
<td>Genetic variation II</td>
<td>Hall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/10</td>
<td>Mon</td>
<td>Genetic variation II</td>
<td>Hall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4/12</td>
<td>Wed</td>
<td>LAB 9: Genetic variation</td>
<td></td>
<td></td>
<td>LAB 7: Single cell analysis</td>
</tr>
<tr>
<td>4/13</td>
<td>Fri</td>
<td>Lab 8: Metagenomics</td>
<td></td>
<td></td>
<td>LAB 8: Metagenomics</td>
</tr>
<tr>
<td>4/19</td>
<td>Mon</td>
<td>Population genetics I</td>
<td>Conrad</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A little history
History of Bio5488

- HGP
  - Computational Biology
  - 1998 Bio5495 Eddy

- Functional Genomics

- Microarray
  - 2003 Bio5488 Cohen Mitra

- Systems/Synthetic Biology
  - 2006 Bio5495 Brent

- Next-gen Sequencing
  - 2012 Bio5488 Wang Conrad

- ENCODE
  - GWAS Roadmap etc
The Origin of "Genomics": 1987

EDITORIAL
A New Discipline, A New Name, A New Journal

Genomics (1987)

"For the newly developing discipline of [genome] mapping/sequencing (including the analysis of the information), we have adopted the term GENOMICS..."
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 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 and Epigenomics

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

- **00's**
  - Microarray Omics
  - Gene expression
  - Comparative Genomics
  - Evolution
  - ENCODE
  - GWAS
  - Roadmap etc
  - Machine Learning
  - Data mining
  - Structural informatics
  - Drug Design
  - Statistical Modeling
  - Database

- **10's**
  - Next-gen Sequencing
  - Systems/Synthetic Biology
  - Single cell
  - GTEx
  - 4DN

- **GTEx**
  - 4DN
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”
  – 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
Three Decades of Genomics

- Human Genome Sequenced for First Time by the Human Genome Project
- Cost of Sequencing a Human Genome Reduced Nearly ~1 Million-Fold
- Many Tens of Thousands of Human Genomes Sequenced
- Profound Advances in Understanding How the Human Genome Functions
- Significant Advances in Unraveling the Genomic Bases of Human Disease
- Vivid Examples of Genomic Medicine in Action Now Emerging
‘Hot Areas’ in Genomic Medicine

- Cancer Genomics
- Pharmacogenomics
- Rare Genetic Disease Diagnostics
- Genomics of Pregnancy
- Clinical Genomics Information Systems
The sequence explosion

The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

- A llama cell line\(^\text{5,6}\), rull\(^\text{6}\), Tollan and Archibird\(^\text{6}\), Desmond Tutu\(^\text{7}\), James Lupa\(^\text{7}\), and a family of four\(^\text{7}\),
- Two Korean males including Soey\(^\text{8}\), Je Kim\(^\text{8}\), Stephen Quale\(^\text{9}\), another cancer genome\(^\text{8}\), George Church, a Yoruban female, another male\(^\text{9}\), and four others\(^\text{9}\),
- James Wang\(^\text{9}\), a woman with acute myeloid leukaemia\(^\text{9}\), a robot from Jig/></span>

The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

- Human Genome Project completed\(^\text{10}\)
- Whole Genome Shotgun Sequence
- Gene sequence stored in international public databases
- SEQUENCING BY LIGATION
  - This technique employed in SOLiD and 454 platforms allows sequence from short single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and IonTorrent say they can read from longer molecules as they pass through a pore.

**454 PYROSEQUENCING:** Released in 2005, 454 sequencing is considered the first next-generation technique. A machine could sequence hundreds of millions of base pairs in a single run.

**SEQUENCING BY SYNTHESIS:** Other companies such as Solexa (now Illumina) modified the next-generation sequencing by-synthesis techniques and can produce billions of base pairs in a single run.

**AUTOMATED SANGER SEQUENCING:** Based on a decade-old method, at the peak of the techniques, a single machine could produce hundreds of thousands of base pairs in a single run.

**Cost per-million base pairs of sequence (log scale):**

- $10,000
- $1,000
- $100
- $10

**Genomes sequenced**

http://www.genomesonline.org/, (as of January 2018)

- 256,794 Bacteria,
- 2,439 Archaea,
- 21,646 Eukarya,
- 27,920 Metagenome samples
- 8,876 Viruses
- 2,298 Synthetic genomes
“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
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\}

GTCGCGTTTCCCCTGAAAACGCAGATGTGCCTCGCCTCGCCGCACGTGCT
CCGAACAATAAGATTCTACAATACTAGCTTTTTATGTTATG
AAGAGGAAAAATTGGCAGTAACCTGGCCCCAACAAACCTTCACA
ATTAACGAATCAAATTTAACAACCATAGGATGATAATCGGATT
AGTATTATTTTGATCTATAACAGATATATAAATGAAAAGCTG
CATAACCACTTTTAACTAATACCTTCCATTTTACGTGTTGTA
TTACTTCTTATTCAAATGTCTATAAAAGTATCAACAAAAATT
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.
What makes us different?

Differences between individuals?

Differences between species?
One genome, thousands of epigenomes

Embryonic stem cells

iPS cells 11a, 15b, 18b, 18c, 20b

ES cell lines H1, I3, WA7

+ BMP4

Neuronal Progenitors

Trophoblast (in progress)

Extra-embryonic ectoderm (in progress)

Fetal tissues

Adult cells and tissues
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.

Science 2005
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

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% hg16-mm3-rm3 >=200bp
c1000.351

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

Only **1.5%** of the human genome are **protein-coding regions**

**Transposable elements** make up **almost half** of the human genome

Barbara McClintock
Transposable Elements (TEs)

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
Thinking Quantitatively
Biology Is A Quantitative Science!!!

1) Mendel’s Laws
2) Chargaff’s Rules
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

• Probability

• Bayes rule

• 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?

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

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

$$P = 0.10$$

What is the chance of getting seven or better?

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

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

$$P = 0.16$$
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

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

\[ P(E^c) = 1 - P(E) \]
Independent Events/Mutually Exclusive Events

What is the probability of A and B both occurring?
\[ P(AB) = P(A) \times P(B) \]  (Independent)
\[ P(A|B) = P(A) \]  (Independent)
\[ P(A|B) \times P(B) = P(B|A) \times P(A) \]  (Bayes rule)
\[ P(AB) = P(A|B) = 0 \]  (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)?
\[ P(A) + P(B) \]
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?

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: