MACHINE LEARNING FOR COMPUTATIONAL BIOLOGY
GENOMES, CIRCUITRY, DISEASE, THERAPEUTICS
6.8700 / 6.8701 / HST.507[J]

NEWLY REVAMPPED FOR FALL 2023!
NOW INCLUDES DEEP LEARNING & DRUG DISCOVERY

Epigenomics
Disease Circuitry
Protein Folding
Immunology
Gene Networks
Single-cell Analyses
Ligand Binding
Imaging Methods
Covers the foundations and frontiers of computational biology, combining theory with practice. Principles of machine learning and algorithm design, influential problems and techniques, analysis of large-scale biological datasets, applications to human disease and drug discovery. Genomes, sequence analysis, regulatory motifs, epigenomics; gene expression, networks, RNA; evolution, comparative genomics, phylogenetics; genetic variation, disease circuitry; therapeutics, drug discovery, protein folding and ligand binding. Combines classical algorithmic techniques such as Gibbs sampling, expectation maximization, hidden Markov models, Bayesian inference, with deep learning methods, including sequence and graph based models, geometric neural networks, and deep generative models.

Final Project can lead to thesis / paper
In-class quiz, no final exam
Satisfies TQE in AI & EC elective in Theory
First class: Thu Sept 7 at 1pm in 32-144

Lectures: TR1-2:30pm, Room 32-144
Prereqs: Algo (6.006), Prob (6.041), Bio (7.01)
Units: 12 (4-0-8)
Contact: MLCB@mit.edu
Goals for today: Course Introduction

1. Course overview:
   – Staff, students, responses to student survey
   – Foundations, frontiers, textbook, homework, quiz
   – Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   – What makes our field unique

3. Overview of course modules
   – Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   – Central Dogma of Molecular Biology
   – DNA, Epigenomics, RNA, Protein, Networks
   – Human genetics, Drug Discovery
I. Administrivia

Introduction to the course and its goals
Course organization and content
Homework and Quiz
Term Project
Introductions

• Lecturer: Manolis Kellis
  – MIT CSAIL, CompBio, Broad, Disease mechanism, Epigenomics, Cancer, Brain, Gene Regulation, Evolution, Single-cell genomics

• Lecturer: Regina Barzilay
  – MIT CSAIL, Jameel Clinic, AI, ML, NLP, Oncology, Chemistry, GNNs, Trusted AI, NLP,

• TA: Jeremy Wohlwend
  – MIT CSAIL, LLMs, Graph Neural Networks, Antibody Design, CNNs, Deep Learning, Imaging, 3D microscopy, Knowledge

• TA: Na Sun
  – MIT CSAIL, Computational Biology, Broad Institute, Neuroscience, Immunology, Disease Mechanism, Single-Cell

• TA: Benjamin James
  – MIT CSAIL, Computational Biology, Broad Institute, Epigenomics, Regulatory Circuitry, Single-Cell, Addiction, Neuroscience
Course Information

• Lectures
  – TR 1pm – 2:30, Room 32-144

• Recitations/Mentoring/OfficeHours:
  – On Friday at 3pm in 32-144
  – Recitations at MIT

• Course Website
  – https://canvas.mit.edu/courses/22618
  – or simply: http://compbio.mit.edu/MLCB (redirects to canvas)
  – All handouts, lectures, notes, etc will be posted here.

• Course calendar:
  – On Google, add public calendar: “MLCB23 Lectures”
Goals for the term

• Introduction to computational biology
  – Fundamental problems in computational biology
  – Algorithmic/machine learning techniques for data analysis
  – Research directions for active participation in the field
  – Understanding how methods work

• Ability to tackle research
  – Problem set questions: algorithmic rigorous thinking
  – Programming assignments:
    → hands-on experience w/ real datasets
  – Final project experience:
    → propose and carry out independent original research
    → present findings in conference format (written, oral)
Course content
Computation & Biology | Foundations & Frontiers

- **Duality #1 (x-axis):** Computation and Biology
  - **Important, relevant, current biology:**
    → Important biological problems
  - **Fundamental computer science:**
    → General techniques, principles

- **Duality #2 (y-axis):** Foundations and Frontiers
  - **Foundations:**
    - well-defined problems, general methodologies
    - ‘The classics’ of the field
  - **Frontiers:**
    - in-depth look at complex, current problems, open questions
    - combine techniques learned
    - opens to projects, research directions
Course organized around bio/comp modules

• Each module corresponds to an active area of research
  – Comparative genomics: Align/model genomes, DP, HMMs
  – Genes and Transcripts: RNA-seq, clustering, structure
  – Regulation: Epigenomics, TFs, Motifs, Network inference
  – Variation: Genetics, Human history, heritability, eQTLs
  – Evolution: Phylogeny, evolutionary sigs, WGD, assembly
  – Frontiers: Personal/Disease, 3D genomes, Pharma, Synth

• Each module contains: foundations
  – Dynamic programming, string matching, hashing, HMMs, EM, Gibbs Sampling, Clustering, Classification, Feature selection, SVMs, CRFs, Context-Free Grammars, phylogenetics, gene / species trees, evolutionary models, GWAS, disease mapping

• Each module contains: frontiers
  – Evolutionary signatures, Transcript analysis, lincRNAs, Network inference and analysis, Epigenomics, Recent human selection and ancestry, chromatin regulation, Missing heritability, 3D
# Course at a Glance

<table>
<thead>
<tr>
<th>Homeworks</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>No HW</td>
<td></td>
<td>1</td>
<td>Thu-Sep-07</td>
<td>L1</td>
<td>Course Overview, Machine Learning, Deep Learning, Inference, Genome, Proteins, Chemistry, Imaging</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Tue-Sep-12</td>
<td>L2</td>
<td>Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-Discr Learning</td>
</tr>
<tr>
<td>HW1 out Tue</td>
<td></td>
<td>2</td>
<td>Thu-Sep-14</td>
<td>L3</td>
<td>Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures</td>
</tr>
<tr>
<td>HW1 due Mon</td>
<td></td>
<td>2</td>
<td>Thu-Sep-14</td>
<td>L3</td>
<td>Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Tue-Sep-19</td>
<td>L4</td>
<td>Regulatory Genomics, Motifs, Information, ChIP, Gibbs Sampling, EM, CNNs for Genome Parsing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Thu-Sep-21</td>
<td>L5</td>
<td>Epigenomics: Signal Modeling, Peak calling, Chromatin states, 3D structure, Hi-C, Genome Topology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Tue-Sep-26</td>
<td>L6</td>
<td>Modeling Small Molecules using GNNs, Property Predictions, Characterizing Synergistic Effects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Thu-Sep-28</td>
<td>L7</td>
<td>Generative Models for Lead Optimization of Small Molecules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Thu-Oct-03</td>
<td>L8</td>
<td>Modeling Proteins using sequence and 3D Models; Protein Folding</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Thu-Oct-05</td>
<td>L9</td>
<td>Generative Models for Binding, Docking and Design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Tue-Oct-10</td>
<td>-</td>
<td>Holiday Columbus Day / Indigenous People Day</td>
</tr>
<tr>
<td>HW2 out Tue</td>
<td></td>
<td>6</td>
<td>Thu-Oct-10</td>
<td>L10</td>
<td>Uncertainty Estimation; Handling Bias in the Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Thu-Oct-17</td>
<td>L11</td>
<td>Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Thu-Oct-19</td>
<td>L12</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWA, multi-modal QTLs</td>
</tr>
<tr>
<td>HW3 due Mon</td>
<td></td>
<td>7</td>
<td>Thu-Oct-24</td>
<td>L13</td>
<td>Single-cell genomics, sc-multi-omics, non-linear embeddings, spatial transcriptomics, next-gen technologies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Thu-Oct-26</td>
<td>L14</td>
<td>Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimensionality Reduction, TF-enhancer-gene circuitry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Thu-Oct-26</td>
<td>L14</td>
<td>Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimensionality Reduction, TF-enhancer-gene circuitry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Tue-Oct-31</td>
<td>L15</td>
<td>Electronic Health Records, AI, UKBioBank, Medical Genomics, Population-Scale Cohorts, Multi-Ancestry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Thu-Nov-02</td>
<td>L16</td>
<td>Imaging methods for biological applications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Tue-Nov-07</td>
<td>L17</td>
<td>Comparative genomics, Alignment, Conservation, Evolutionary signatures, PhyloCSF, RNA structure, MorfELS,2conf</td>
</tr>
<tr>
<td>HW4 out Tue</td>
<td></td>
<td>10</td>
<td>Thu-Nov-09</td>
<td>L18</td>
<td>Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, IncRNAs, RNA modifications, miA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Thu-Nov-09</td>
<td>L18</td>
<td>Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, IncRNAs, RNA modifications, miA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Tue-Nov-14</td>
<td>L19</td>
<td>Introduction to Drug Discovery, Case Study: Herceptin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu-Nov-16</td>
<td>L20</td>
<td>In-Class Quiz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Tue-Nov-21</td>
<td>L21</td>
<td>Drug Localization &amp; Delivery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Thu-Nov-23</td>
<td>-</td>
<td>Thanksgiving Holiday</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Tue-Nov-28</td>
<td>L22</td>
<td>Genome Engineering, CRISPR/Cas9, Massively Parallel Reporter Assays, Multiplexing, RNA therapeutics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Tue-Nov-28</td>
<td>L22</td>
<td>Genome Engineering, CRISPR/Cas9, Massively Parallel Reporter Assays, Multiplexing, RNA therapeutics</td>
</tr>
<tr>
<td>HW4 due Mon</td>
<td></td>
<td>13</td>
<td>Thu-Nov-30</td>
<td>L23</td>
<td>Computational Approaches for Metabolomics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Tue-Dec-05</td>
<td>L24</td>
<td>Immunology and Vaccine Design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Tue-Dec-05</td>
<td>L24</td>
<td>Immunology and Vaccine Design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Tue-Dec-12</td>
<td>L25</td>
<td>Computational Biology at Novartis</td>
</tr>
<tr>
<td>Final Projects</td>
<td></td>
<td>15</td>
<td>Tue-Dec-12</td>
<td>L26</td>
<td>Project Presentations (6-8 mins per team). Write-ups due Mon, Slides Tue at 10am, Presentations at 1pm</td>
</tr>
</tbody>
</table>
Textbook / class notes / resources
(Optional) Books for the Course

- Durbin, Eddy, Krogh, Mitchison
- Jones, Pevzner
- Duda, Hart, Stork

Availability: BU Coop, MIT Coop, amazon.com (~$40-60)

All three books on reserve at the MIT and BU Engineering libraries
Book for the Course

Computational Biology:
Genomes, Networks, Evolution

MIT Course 6.047/6.878

Manolis Kellis & all of you!

... being compiled this year
by students like you!
... actually, including you!

Availability: Current version online on Stellar, for registered students only

Link to compiled scribe notes from 2014: http://tiny.cc/6047bookF14
Lectures and Scribing

• Each lecture will have a dedicated scribe who will take notes on the lecture
  – Please sign up to scribe for lecture on the sheet being passed around

• Build on notes from previous years
  – Available on course website

• Final draft of scribe notes due 6 days after lecture
  – Your grade depends on the improvement from previous year and completeness

• Some lectures need more work: multiple scribes

• Some tasks are better-suited to you than just scribing
  – E.g. figures, references, layout, macros, let us know!
Video recordings from previous years

Panopto (Fall 2021)
https://mit.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=7c716154-6516-4a49-9a81-adad0135dcb8

http://compbio.mit.edu/videos.html#MachineLearningCourse_F20

• Lecture 1 - Introduction (1h26)
• Lecture 2 - Dynamic Programming (1h22)
• Lecture 3 - Local alignment Hashing BLAST alignmentScores (1h27)
• Lecture 4 - HMMs 1 (1h21)
• Lecture 5 - HMMs 2 (1h21)
• Lecture 6 - Expression analysis Clustering Classification (1h23)
• Lecture 7 - RNA folding, RNA world, RNA structures (1h26)
• Lecture 8 - Epigenomics 1 (1h24)
• Lecture 9 - Epigenomics 2 and 3D genome (1h27)
• Lecture 10 - Regulatory Genomics and Motifs (1h22)
• Lecture 11 - Networks (1h27)
• Lecture 12 - Deep Learning (1h24)
• Lecture 13 - Population Genetics (1h28)
• Lecture 14 - GWAS and Disease Dissection (1h31)
• Lecture 15 - eQTLs expression Quantitative Trait Loci (1h27)
• Lecture16 - Systems Genetics and Heritability (1h25)
• Lecture17 - Comparative Genomics (1h25)
• Lecture 18 - Genome Evolution (1h26)
• Lecture 19 - Phylogenetics (1h29)
• Lecture 20 - Phylogenomics (1h28)
• Lecture 21 - Cancer Genomics (1h29)
• Lecture 01 - Introduction (1h15)
• Lecture 02 - Dynamic Programming (1h19)
• Lecture 03 - Hashing BLAST Database Search (1h23)
• Lecture 04 - HMMs Hidden Markov Models I (1h17)
• Lecture 05 - HMMs Hidden Markov Models II (1h15)
• Lecture 06 - Expression Analysis Clustering Classification (1h17)
• Lecture 07 - RNA world, RNA-seq, RNA folding (1h16)
• Lecture 08 - Epigenomics I (1h11)
• Lecture 09 - Epigenomics II (1h22)
• Lecture 10 - Regulatory Genomics (1h18)
• Lecture 11 - Network inference and analysis (1h16)
• Lecture 12 - Deep Learning (1h30)
• Lecture 13 - Population Genetics (1h18)
• Lecture 14 - GWAS (1h17)
• Lecture 15 - eQTLs Mediation (1h18)
• Lecture 16 - Systems Genetics (1h14)
• Lecture 17 - Comparative Genomics (1h22)
• Lecture 18 - Genome Evolution (1h21)
• Lecture 19 - Phylogenetics (1h20)
• Lecture 20 - Phylogenomics (1h22)
• Lecture 21 - Single-cell genomics (1h25)
• Lecture 22 - Cancer Genomics (1h26)
• Lecture 23 - Multi-Phenotype analyses (1h20)
• Lecture 24 - Genome Engineering (1h18)
• Lecture 25 - How to Present - Papers, Figures, Presentations (1h20)
• Lecture 11 1/2 - 6047 Buzzword Recitation (52 mins)
Fall 2020, 2019, 2018: YouTube, and ease of use anywhere

YouTube Playlist: (Fall 2021)
Fall 2020: https://www.youtube.com/playlist?list=PLypiXJdtIca6dEYlNoZJwBaz_CdsaoKJ
Fall 2019: https://www.youtube.com/playlist?list=PLypiXJdtIca6U5uQOCHjp9Op3gpa177fK
Fall 2018: https://www.youtube.com/playlist?list=PLypiXJdtIca6GBQwDTo4bIEDV8F4RCaAg
Fall 2021, 2022: Panopto, and awesome search capabilities

Panopto (Fall 2021)
https://mit.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=7c716154-6516-4a49-9a81-adad0135dcb8
Panopto (Fall 2022)
https://mit.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=176f8b23-0433-403d-8c26-af090151a28d

Speaker video

Search function!

Shared screen

Automatic transcript

2X speed

Automatic chapters (from slide headers)

Slide Navigation

Panopto:
- Automatic chapters (from slide headers)
- Search function!
- Shared screen
- Automatic transcript
- 2X speed
- Speaker video

Panopto links:
- Fall 2021: https://mit.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=7c716154-6516-4a49-9a81-adad0135dcb8
- Fall 2022: https://mit.hosted.panopto.com/Panopto/Pages/Sessions/List.aspx?folderID=176f8b23-0433-403d-8c26-af090151a28d
Lecture feedback (for any/every lecture!): https://forms.gle/zdD833bPFF596qQE8

1. Your interest in the overall topic: 1-5
2. The material actually presented 1-5
3. Quality of presentation
   – Quality of slides 1-5
   – Clarify of explanations 1-5
   – Usefulness of lecture notes 1-5
   – Were questions adequately answered 1-5
4. Pace:
   – Difficulty of the material: too easy - just right - too hard
   – Amount of material covered: too little - just right - too much
   – Pace of the lecture: too slow - just right - too fast
5. Comprehension (for each topic)
   – <20%, 20-40%, 40-60%, 60-80%, >80%
Homeworks and quiz
Details on Problem sets

• Each problem emphasizes one lecture (or two)
  – Practical problem: gain experience in techniques, write code, download datasets, carry out analysis, interpret your results, learn about behavior of problem/method

• Due Mondays at 11:59pm
  – Late policy: we are flexible, give or take a few hours
  – If more than a few hours, need prior arrangements, extensions typically not granted, except special circ.

• Submit all homeworks online from canvas page
  – No solutions distributed. If you’ve solved them, you know what you needed to learn/discover/achieve.
Details on the in-class quiz

• It’s not a midterm, and it’s not a final exam
  – It’s a quiz, friendly, fun, interesting, cute, fuzzy

• Demonstrate mastery of the material in 4 modules
  – Understand key points emphasized in lecture
  – Understand subtleties revealed in the psets
  – Ability to apply new skills to solve practical problems

• Types of questions
  – Knowledge questions: T/F justify, multiple choice
  – Deeper understanding questions: short answers
  – Practical problems: work through simple algorithm
  – Design problem(s): new/modified algorithm, need both knowledge and new idea, argue correctness
Final Project
Final Project: Original Research in Comp Bio

• A major aspect of the course is preparing you for original research in computational biology.
  – Framing a biological problem computationally
  – Gathering relevant literature and datasets
  – Solving it using new algorithms, machine learning
  – Interpreting the results biologically

• Also ability to present your ideas and research
  – Crafting a research proposal (fellowships/grants)
  – Working in teams of complementary skill sets
  – Review peer proposals, find flaws, suggest improvements
  – Receiving feedback and revising your proposal
  – Writing up your results in a scientific paper format
  – Presenting a research talk to a scientific audience

• Term project experience mirrors this process
Project Milestones

How much guidance would you like for each part of the term project

- Round 0: Self-introduction (due Week 2 Friday)
- Round 1: Literature search and paper description (due Week 4 Friday)
- Round 2: Team formation, project proposal, feasibility (due Week 7 Friday)
- Round 3: Office Hours, Update, Feedback (Meet Week 9 Friday)
- Round 4: Midcourse report (due Week 13, Friday)
- Round 5: Final report+slides (due Week 15, Friday)
It’s a team project

• Please make an effort to meet your peers!
• Form teams early with complementary expertise
# Final Project at a Glance

<table>
<thead>
<tr>
<th>Homeworks</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>No HW</td>
<td></td>
<td>1</td>
<td>Thu-Sep-07</td>
<td>L1</td>
<td>Course Overview, Machine Learning, Deep Learning, Inference, Genome, Proteins, Chemistry, Imaging</td>
</tr>
<tr>
<td>HW1 out Tue</td>
<td></td>
<td>2</td>
<td>Tue-Sep-12</td>
<td>L2</td>
<td>Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-Discr Learning</td>
</tr>
<tr>
<td>HW1 due Mon</td>
<td></td>
<td>2</td>
<td>Thu-Sep-14</td>
<td>L3</td>
<td>Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures</td>
</tr>
<tr>
<td>HW1 due Mon</td>
<td></td>
<td>3</td>
<td>Tue-Sep-19</td>
<td>L4</td>
<td>Regulatory Genomics, Motifs, Information, ChiP, Gibbs Sampling, EM, CNNs for Genome Parsing</td>
</tr>
<tr>
<td>HW1 due Mon</td>
<td></td>
<td>3</td>
<td>Thu-Sep-21</td>
<td>L5</td>
<td>Epigenomics: Signal Modeling, Peak calling, Chromatin states, 3D structure, Hi-C, Genome Topology</td>
</tr>
<tr>
<td>HW2 out Tue</td>
<td></td>
<td>4</td>
<td>Tue-Sep-26</td>
<td>L6</td>
<td>Modeling Small Molecules using GNNs, Property Predictions, Characterizing Synergistic Effects</td>
</tr>
<tr>
<td>HW2 due Mon</td>
<td></td>
<td>4</td>
<td>Thu-Sep-28</td>
<td>L7</td>
<td>Generative Models for Lead Optimization of Small Molecules</td>
</tr>
<tr>
<td>HW2 due Mon</td>
<td></td>
<td>5</td>
<td>Thu-Oct-03</td>
<td>L8</td>
<td>Modeling Proteins using sequence and 3D Models, Protein Folding</td>
</tr>
<tr>
<td>HW2 due Mon</td>
<td></td>
<td>5</td>
<td>Thu-Oct-05</td>
<td>L9</td>
<td>Generative Models for Binding, Docking and Design</td>
</tr>
<tr>
<td>HW2 due Mon</td>
<td></td>
<td>6</td>
<td>Thu-Oct-10</td>
<td>L10</td>
<td>Holiday Columbus Day / Indigenous People Day</td>
</tr>
</tbody>
</table>

**Modeling Biology II: Small Molecules & Proteins**

<table>
<thead>
<tr>
<th>Homeworks</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW3 out Tue</td>
<td></td>
<td>7</td>
<td>Thu-Oct-17</td>
<td>L11</td>
<td>Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function</td>
</tr>
<tr>
<td>HW3 due Mon</td>
<td></td>
<td>7</td>
<td>Thu-Oct-19</td>
<td>L12</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWASt, multi-modal QTLs</td>
</tr>
<tr>
<td>HW3 due Mon</td>
<td></td>
<td>8</td>
<td>Thu-Oct-24</td>
<td>L13</td>
<td>Single-cell genomics, omics, non-linear embeddings, spatial transcriptomics, next-gen technologies</td>
</tr>
<tr>
<td>HW3 due Mon</td>
<td></td>
<td>8</td>
<td>Thu-Oct-26</td>
<td>L14</td>
<td>Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimensionality Reduction, TF-enhancer-gene circuitry</td>
</tr>
</tbody>
</table>

**Drug Discovery I: Identifying Targets**

<table>
<thead>
<tr>
<th>Homeworks</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW4 out Tue</td>
<td></td>
<td>9</td>
<td>Tue-Oct-31</td>
<td>L15</td>
<td>Electronic Health Records, AIOUs, UKBioBank, Medical Genomics, Population-Scale Cohorts, Multi-Ancestry</td>
</tr>
<tr>
<td>HW4 due Mon</td>
<td></td>
<td>9</td>
<td>Thu-Nov-02</td>
<td>L16</td>
<td>Imaging methods for biological applications</td>
</tr>
<tr>
<td>HW4 due Mon</td>
<td></td>
<td>10</td>
<td>Tue-Nov-07</td>
<td>L17</td>
<td>Comparative genomics, Alignment, Conservation, Evolutionary signatures, PhyloCSF, RNA structure, MorfBLS2conf</td>
</tr>
<tr>
<td>HW4 due Mon</td>
<td></td>
<td>10</td>
<td>Thu-Nov-09</td>
<td>L18</td>
<td>Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, IncRNAs, RNA modifications, miRNA</td>
</tr>
</tbody>
</table>

**Modeling Biology III: Beyond Genomes: EHRs, Imaging, Evolution**

<table>
<thead>
<tr>
<th>Homeworks</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>11</td>
<td>Tue-Nov-14</td>
<td>L19</td>
<td>Introduction to Drug Discovery, Case Study: Herceptin</td>
</tr>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>11</td>
<td>Thu-Nov-16</td>
<td>L20</td>
<td>In-Class Quiz</td>
</tr>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>12</td>
<td>Thu-Nov-21</td>
<td>L21</td>
<td>Drug Localization &amp; Delivery</td>
</tr>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>12</td>
<td>Thu-Nov-23</td>
<td>L22</td>
<td>Thanksgiving Holiday</td>
</tr>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>13</td>
<td>Tue-Nov-28</td>
<td>L23</td>
<td>Genome Engineering, CRISPR/Cas9, Massively Parallel Reporter Assays, Multiplexing, RNA therapeutics</td>
</tr>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>13</td>
<td>Thu-Nov-30</td>
<td>L24</td>
<td>Computational Approaches for Metabolomics</td>
</tr>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>14</td>
<td>Tue-Dec-05</td>
<td>L25</td>
<td>Immunology and Vaccine Design</td>
</tr>
<tr>
<td>HW5 out Tue</td>
<td></td>
<td>14</td>
<td>Thu-Dec-07</td>
<td>L26</td>
<td>Computational Biology at Novartis</td>
</tr>
</tbody>
</table>

**Final Projects**

<table>
<thead>
<tr>
<th>Homeworks</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>No HW</td>
<td></td>
<td>15</td>
<td>Tue-Dec-12</td>
<td>L26</td>
<td>Project Presentations (6-8 mins per team). Write-ups due Mon, Slides Tue at 10am, Presentations at 1pm</td>
</tr>
</tbody>
</table>
Details on the final project

• Milestones ensure sufficient planning / feedback
  – Set-up: find project matching your skills and interests
  – Team: common interests and complementary skills
  – Inspiration: last year’s projects, and recent papers
  – Proposal: establish milestones, deliverables, expectations
  – Midcourse: see endpoint, outline report, methods, figures

• Periodic mentoring sessions
  – Senior students and postdocs can serve as your mentors
  – Group discussions to share ideas, guidance, feedback
  – Peer-review: think critically about peer proposals, receive feedback/suggestions, respond to critiques, adjust course

• Real-world experience, condensed in a single term
  – Grant/fellowships proposals, peer review, yearly reports, budget time/effort, collaboration, paper writing, give talk
Comm Lab: Help communicating your research!

A free resource for peer feedback from trained EECS grad students and postdocs.

Why people come to CommLab:

- Resume / CV: 63 appointments
- Graduate school appl. (incl. startup plans, RQE): 43 appointments
- Faculty package: 38 appointments
- Other report or essay: 35 appointments
- Oral presentation: 34 appointments
- Fellowship / scholarship appl.: 33 appointments
- Thesis: 33 appointments
- Manuscript: 31 appointments
- Poster / visual: 29 appointments
- Thesis proposal: 27 appointments
- Abstract: 14 appointments
- Grant: 8 appointments
- Lab report: 3 appointments

Total: 400 appointments

Number of appointments

"Very, very valuable. Thank you!"
—Elena Glassman, EECS PhD alumna

"I strongly encourage students to schedule a session; it’s a very impressive resource."
—Dirk Englund, professor

“"The experience and coaching helped me apply successfully for an important fellowship this year.”
—Joel Jean, EECS grad
Finding a research mentor / research advisor

- Chance to meet faculty at MIT/Broad/Harvard:
  - Through guest lectures and mentoring
  - Topics and papers covered in the lectures
  - Experts on: (1) human comparative genomics, (2) lincRNAs, (3) metabolic modeling, (4) disease mapping, selection, evolution and ecology (following four modules)

- Chance to meet senior students and postdocs:
  - On: coding genes, ncRNAs, regulatory motifs, networks, epigenomics, phylogenomics (again on each module)
  - Mentorship sessions with entire MIT CompBio group

- Your own personal research experience:
  - collaborators, datasets
  - learn active research directions, frontiers
  - living, breathing changing field
Putting it all together
Course Activities: Mens et Manus

- **Learning** (25 lectures * 1.5 hours)
- **Mentoring** (4-7 meetings * 1 hour)
- **4 problem sets: [30% of your grade]**
  - Out on Tuesdays, due Mondays in 2-3 weeks.
  - Each problem set: 7-10%, covers 3-4 lectures, contains 3-4 problems.
  - Algorithmic problems and programming assignments
- **Final project [40% of your grade]**
  - Introduction to research in computational biology (full term!)
  - Includes peer-reviewed NIH-style proposal and much feedback
- **Quiz [25% of your grade]**
  - In-class quiz. No final exam.
- **Office hours/recitations/lectures participation: 5% grade**
- **Collaboration policy [humans and AI]**
  - Collaboration allowed, but you must:
    - Work independently on each problem before discussing it
    - Write solutions on your own
    - Acknowledge sources and collaborators. No outsourcing.
  - **ChatGPT / LLM policy**
    - Acknowledge the way you would for a collaboration partner
    - Be transparent, save your chats, possibly submit w/ homework
Goals for today: Course Introduction

1. Course overview:
   – Staff, students, responses to student survey
   – Foundations, frontiers, textbook, homework, quiz
   – Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Learning more about you (first-day survey results)

3. Why Computational Biology:
   – What makes our field unique

3. Overview of course modules
   – Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   – Central Dogma of Molecular Biology
   – DNA, Epigenomics, RNA, Protein, Networks
   – Human genetics, Drug Discovery
Students taking this class Fall 2023
- Please make an effort to meet your peers!
- Form teams early with complementary expertise
What year are you?

- U1: 0 (0%)
- U2: 0 (0%)
- U3: 6 (10.7%)
- U4: 14 (25%)
- MEng: 7 (12.5%)
- G1: 17 (30.4%)
- G2: 5 (8.9%)
- G3: 4 (7.1%)
- G4+: 2 (3.6%)
- MS Student at Harvard: 1 (1.8%)
- Y2 Master of Biomedical Informatics: 1 (1.8%)

What major are you?

- 6-7: 12 (21.4%)
- 6: 12 (21.4%)
- 7: 1 (1.8%)
- 18C: 2 (3.6%)
- 20: 0 (0%)
- CSBi: 1 (1.8%)
- HST: 1 (1.8%)
- MAS: 1 (1.8%)
- HMS: 18 (32.1%)
- Harvard: 7 (12.5%)
- 10: 1 (1.8%)
- 6-4: 1 (1.8%)
- CBQQG: 1 (1.8%)
- 5: 1 (1.8%)
- 6-9: 1 (1.8%)
More than half the students have no research experience in computational biology.
Score your primary drivers/reasons for taking this class

- For the Machine Learning / Algorithms / Computational Methods
- For the Biological / Genomics / Health Applications
- For the Research Project Experience
- To obtain a new position in Machine Learning / Computational Biology
- Because my current/upcoming position requires the knowledge
- Just browsing/exploring a new field
- To fulfill a degree requirement
Course Goals

Understanding of Computational Biology and Applications:
- Acquire the fundamentals of computational biology and explore research questions in computational genomics.
- Gain experience with fundamental computational methods.
- Gain a comprehensive survey of computation and biology.
- Strengthen foundational computational biology knowledge and learn about lesser-known areas like drug discovery.
- Learn new computational methods and better understand current methods.
- Learn tools used in analysis for biology and the theory behind them.
- Gain skills in computational biology work and its applications.
- Understand computational methods behind tools like alignment algorithms or phylogeny programs.
- Gain exposure to computational biology, integrating biology and programming knowledge.
- Learn various techniques in computational biology.

Machine Learning (ML) in Biology:
- Learn about machine learning's application in different subfields of biology.
- Learn about geometric deep learning and other machine learning techniques for computational biology.
- Gain familiarity with ML algorithms and practical applications in genomics and computational biology.
- Deepen understanding of machine learning applications in the biology/healthcare space.
- Explore ML's application in a field of interest.
- Learn the application of ML in genomics.
- Understand advantages and disadvantages of specific ML algorithms and how to implement them.
- Strengthen ML skills, especially in relation to healthcare.
- Understand the workings of machine learning and its applications in specific fields.

Address Gaps & Expand Current Knowledge:
- Fill in knowledge gaps from previous genomics coursework/background.
- Gain a structured introduction to Computational Biology.
- Get a deeper understanding of underlying algorithms to inform research.
- Improve general knowledge in ML and Computational Biology.
- Understand better how genomics/bioinformatics data is analyzed.

Career & Research Intentions:
- Pivot into a more active role in the machine learning for computational biology community.
- Become a potent researcher in the drug discovery field.
- Explore computational biology for potential career opportunities.
- Integrate knowledge to explore new research areas.
- Hope to do related work in graduate school.
- Learn techniques catered to specific data for future research applications.

Curiosity & Exploration:
- Explore the reality of computational biology after having romanticized the field.
- Explore a new field and apply previous ML/stats experience.
- Explore computational biology as a potential career interest.
- Understand the mix between computational biology and machine learning.
- Explore how computational biology can be used in the future.
- Understand the applications of computational biology in medicine and other fields.

Recommendations & Course Requirements:
- Taking the course due to recommendations from lab members or peers.
- Satisfy the CSB comp bio course requirement or other academic requirements.

Specific Topics of Interest:
- Understand algorithms and principles of comparative genomics and evolution.
- Learn more about genomic evolution and inference algorithms.
- Gain more research project experience in computational biology.
- Enhance technical abilities and understand diverse applications of computational biology.
- Improve understanding of computational methods behind tools and medical applications of computational biology.
1. **Academic and Research Pathways:**
   - Apply for a PhD to deepen understanding, especially in fields such as Cancer genomics, quantitative biology, computational biology, infectious disease research, structural biology, and industrial engineering in health systems.
   - MD-PhD student aiming to become a cardiologist and lead research on the 3D genome and its implications for disease.
   - Interest in researching the 3D genome, transcription, and structural biology of viral proteins with goals like studying viral mechanisms, identifying drug targets, predicting mutational escape, and analyzing epidemiological dynamics.
   - Aspiration to eventually become a professor, especially in areas like structural biology.
   - Pursue graduate school after taking a break for research, especially to explore computational and experimental biology for environmental problems caused by climate change.

2. **Medical and Healthcare Field:**
   - Envision working in Health Information Exchanges (HIEs) and population health.
   - Become a physician scientist or work in biotech, especially using generative models and ML for precision psychiatry techniques.
   - MD-PhD goal to investigate the 3D genome's role in disease and develop potential therapies.
   - Possibly pursue a PhD in industrial engineering in health systems, focusing on applying ML methodologies to healthcare problems.

3. **Computational and Machine Learning Integration:**
   - Machine learning applications, especially in areas like EEG, and its use in understanding emerging pandemics.
   - Utilizing machine learning for tissue design, especially integrating synthetic and organic materials.
   - Desire to explore computational biology roles after working in software development.

4. **Industry and Biotech:**
   - Working in biotech, specifically utilizing generative models and machine learning for precision psychiatry.
   - Spend the next 3-5 years in a software-related role, potentially branching into computational biology.
   - Work on tissue design with a focus on the integration of synthetic and organic materials.

5. **Uncertain or Exploratory:**
   - Open to exploring various fields during a master's degree before committing to a PhD subject.
   - Some students expressed uncertainty or are still figuring out their paths.
   - Potential interest in pursuing bioinformatics or a computational component in Biology or Biomedical Engineering (BME).
Highly Specialized and In-depth Research with Publications:
1. Statistical genetics with a publication in a well-known journal.
2. Extensive machine learning for protein engineering, multiple projects including Bayesian statistics and ensemble models.
3. Evolutionary studies with a focus on new gene origins and competitive genomics.
4. Comprehensive work on evolution of mammalian gene-regulatory elements with a linked publication.
6. Genomic epidemiology with a focus on transmission inference.
7. Research focused on structural biology, especially the proteins of LINE-1 retrotransposon.
8. Clonality inference from mutation calling.

Computational, Machine Learning, and AI Applications in Biology:
1. Machine learning applied to biomedical image processing.
5. AI with a focus on precision medicine and clinical decision-making.
6. Computational epigenomics and analysis of cancer datasets.
7. Machine learning applied to soil carbon prediction.

Laboratory-based Research with Some Computational Elements:
1. Blood disorder research in the Li Lab.
2. Structural biology focusing on CryoEM techniques.
3. Solid tumor immunology with sequencing tools.
4. Work on stem cells and blastocyst organoids systems.
5. Research at the intersection of immunology, nanotechnology, and drug delivery.
6. Laboratory research on ovarian cancer with some programming elements.

Other Fields and Varied Research:
1. Research on computerized decision support in population health.
2. Epilepsy research.
4. Medical image classification and data mining.
5. Worked on Technoeconomic Analysis of green fuel options.
6. High-pressure, high-temperature chemical changes of organic matter.
7. Students with a more general or broad interest, e.g., "I don't really know I like lots of stuff."

Limited or No Research Exposure:
1. Some students have explicitly mentioned having no or limited research experience.
**Computational Biology Research**

**High Exposure/Expertise in Computational Biology:**
1. Research on horizontal gene transfer between prokaryotic and arthropod genomes using molecular evolutionary methods.
2. Molecular subtyping in medulloblastoma.
4. Using Broad PRISM datasets for drug activity prediction.
5. Work in the Sabeti Lab at Broad Institute analyzing genomic sequencing data.
8. Research in the Burge Lab studying RNA-binding proteins and 5’ splicing site evolution.
9. Work in the Kellis Lab analyzing genomic data and functional annotations.
10. Research on spatial sequencing analysis in PDAC and CRC.
11. Clonality inference from mutation calling.
12. Research in the Park Lab at HMS analyzing DNA samples from liquid biopsies for cancer detection.
13. Using autoencoders and gene expression data to predict transcription factor activities.
16. Project in the epigenomics space predicting enhancer or promoter activity in different types of cancer.
17. Using REDItools for RNA-seq on response to drugs in ovarian cancer cell lines.
18. UROPing in Berger Lab using protein structure-prediction tools.

**Moderate Exposure:**
1. Coursework focused on computational biology.
2. Will be starting a research rotation in Heng Li's lab.
3. UROP lab, using programming to identify transcription factors for blood cell development.
5. Lieberman Lab research on human microbiome and phage defense elements.
6. Microbiome research.
7. Work with Alex Lenail on stochastic models of cell-fate decision making.
8. Side project on developing an ML algorithm for predicting translational efficiency in toxoplasma.

**Minimal to No Research Experience in Computational Biology:**
1. Interested in pivoting to computational epigenomics.
2. Looking for labs applying machine learning to clinical problems.
3. Considering a project in the field as part of the Master of Biomedical Informatics capstone project.
4. Between labs/projects due to rotation structure.
5. Several students explicitly mentioned having no current research in computational biology.
ML Background

Highly Experienced:
- Have a bachelor's degree with an additional major in statistics.
- TA'd for graduate-level machine learning courses.
- Extensive research experiences spanning a variety of machine learning techniques and applications, including AI in finance, science, graph neural networks, reinforcement learning, and generative models.
- Specific advanced courses at renowned universities (e.g., Probability and Inference at Harvard).
- Deep learning courses with reputed professors and focused machine learning research with faculty members.
- Multiple courses in machine learning, probability, and statistics combined with research experiences, especially as UROPs.
- Honors theses centered on machine learning topics.
- In-depth application of machine learning in specialized sectors, such as medical imaging or bioinformatics.

Moderately Experienced:
- Taken a handful of machine learning, probability, and statistics courses.
- Some practical experience, either through class projects, independent studies, or minor research involvements.
- Engaged in more applied research rather than the development of new algorithms or data science techniques.
- Experience implementing and integrating ML models into applications, with less emphasis on the development side.
- Some research experience in labs or through undergraduate research opportunities.
- Training and experience with popular machine learning frameworks like PyTorch and TensorFlow.
- Minor or side coursework in machine learning or allied areas.

Beginner to Intermediate:
- Exposure to foundational machine learning courses.
- Worked on a few ML projects or have had some research exposure.
- Experience primarily from coursework with little to no practical or research exposure.
- Taken introductory classes in machine learning and associated fields like probability and statistics.
- Self-taught or have taken only a few formal courses.
- Exposure to machine learning concepts during internships or brief research opportunities.

Limited Experience:
- Minimal courses related to machine learning, often just introductory or foundational.
- Self-taught knowledge without formal training.
- Exposure to machine learning concepts but lacking in-depth training or experience.
- Experience is mainly from courses that touch upon machine learning but are not dedicated to it.

Minimal to No Experience:
- Students who have stated they have no formal training or very minimal exposure.
- Limited to foundational courses in probability or statistics without any specific machine learning training.
- Indicated by "N/A" or statements suggesting a lack of familiarity with machine learning.
Biology/genomics Background

Highly Experienced:
- Majors in biology, molecular biology, computational biology, or similar.
- PhD level qualifications in fields such as systems biology.
- Extensive undergraduate coursework covering a wide array of biology and genomics topics.
- Several years of research experiences in genomics and biology in academic labs.
- Participated in or won international competitions related to synthetic biology, such as iGEM.
- In-depth coursework from renowned institutions (e.g., MIT or Harvard) paired with relevant research experiences.

Moderately Experienced:
- Undergraduate majors or minors in biology, computational biology, or allied fields.
- Taken several dedicated courses in genetics, molecular biology, genomics, and similar.
- Engaged in research internships, undergraduate research opportunities (UROPs), or independent studies focused on biology or genomics.
- Practical experience from working in labs or on projects, but might lack extensive theoretical background.
- Completion of core biology sequences during undergraduate studies.

Beginner to Intermediate:
- Basic to intermediate coursework in biology or genomics.
- Some research experiences, either through class projects or minor research involvements.
- Self-taught knowledge paired with some formal courses.
- Engagement in more generalized biology courses with occasional focus on genomics.
- Completed introductory or foundational courses in biology at reputable institutions.

Limited Experience:
- Limited to foundational courses in biology without any specific advanced coursework.
- Some minor research or internship experiences, often without a deep dive into genomics.
- Minimal courses related to biology, often just introductory or foundational.
- Self-taught knowledge without much formal training or practical experience.

Minimal to No Experience:
- Students who have stated they have no training or very minimal exposure.
- Exposure to only a few or one foundational biology course without any deeper exploration.
- Indicated by "None" or statements suggesting a lack of familiarity with biology or genomics.

In summary, the range of biology/genomics backgrounds spans from students with extensive academic training and practical experience to those with minimal or no exposure. As with the machine learning backgrounds, this diversity can be harnessed in a classroom setting to promote collaborative learning and peer-to-peer knowledge sharing.
Algorithms Background

1. Advanced Level:
   - Students who have taken advanced graduate-level algorithm courses that are seminar-based and project-oriented.
   - Those who have had significant exposure through multiple classes like 6.006, 6.046, and even more advanced coursework.
   - Students with competitive programming experience, indicating a deeper practical understanding of algorithms.
   - Some have taken specialized courses in bioinformatics algorithms from renowned institutions.

2. Intermediate Level:
   - Students who have taken standard or intermediate courses in algorithms at renowned institutions like MIT and Harvard.
   - Those who have pursued topics like Introduction to Algorithms, Design and Analysis of Algorithms, and bioinformatics-focused algorithms.
   - Students with exposure to algorithm classes through their undergraduate study.
   - Individuals with experience from internships or research projects involving the analysis and implementation of algorithms.

3. Basic Level:
   - Students who have been exposed to foundational algorithm courses, like MIT's 6.006.
   - Those who have covered algorithms as a part of their CS major but haven't extensively applied them post their course.
   - Students who have undertaken proof-based and code-based algorithm courses in their undergraduate program.
   - Some have a basic understanding of algorithms in specialized areas, such as biomedical image analysis.

4. Minimal Exposure:
   - Students with limited experience through introductory classes in college.
   - Those who have only engaged with algorithm-related topics in a specific context, such as linear regressions or image analysis, without a broader study.
   - Students who have some foundational knowledge, like understanding search algorithms or dynamic programming.

5. No Formal Training:
   - A few students have no formal training in algorithms. Some might be familiar with algorithmic concepts through independent readings or projects but have never taken a formal course.
Goals for today: Course Introduction

1. Course overview:
   - Staff, students, responses to student survey
   - Foundations, frontiers, textbook, homework, quiz
   - Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   - What makes our field unique

3. Overview of course modules
   - Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   - Central Dogma of Molecular Biology
   - DNA, Epigenomics, RNA, Protein, Networks
   - Human genetics, Drug Discovery
Why Computational Biology?
Why Computational Biology: Last year’s answers

- Lots of data (* lots of data)
- There are rules
- Pattern finding
- It’s all about data
- Ability to visualize
- Simulations, temporal relationships
- Guess + verify (generate hypotheses for testing)
- Propose mechanisms / theory to explain observations
- Networks / combinations of variables
- Efficiency (reduce experimental space to cover)
- Informatics infrastructure (ability to combine datasets)
- Correlations, higher-order relationships
- Cycle from hypothesis generation to testing condensed
- Life itself is digital. Understand cellular instruction set
Why **Computational Biology**: Live in Zoom Chat F20

- Data-rich in a historically data-poor domain (Matthew West)
- Potential to do whatever you want without waiting for experiments (Stuti Khandwala)
- DNA is a massive dataset (Pablo X Villalobos)
- More efficient and in depth way to explore biology (Lilly K Edwards)
- There're tons of biological datasets waiting to be analyzed (Hieu Q Dinh)
- Because you can use other people’s datasets and then get good research done on a budget (Ari)
- Might be the biggest frontier of computing today (Erez Kaminski)
- More and more sequencing data are coming out (Evelyn Tong)
- New technologies - lots of data - (Manu Ponnapati)
- Biology benefits from approximation (Thomas Xiong)
- The need to integrate multi-omics data to gain more insights (Kathleen Sucipto)
- It's interesting and new (Daniel R Gutierrez)
- Can use expertise from other engineering fields to impact health (Swathi Manda)
- Complex patterns in biological data (Farhan Khodaei)
- Impact real human lives, important applications (Lucy Zhang)
- Answers questions not easily solvable by traditional experimental biology (Andrew D Hennes)
- Expands our horizons in asking biological questions (Dylan McCormick)
- Computational biology and simulations can help deconvolve results from experiments (Raina Thomas)
Genes Encode proteins

Regulatory motifs
Control gene expression

Genes

Regulatory motifs

Encode proteins

Control gene expression
Extracting signal from noise
Goal: A systems-level understanding of genomes and gene regulation:

- **The genome**: Map reads, align genes/genomes, assembly strategies
- **The genes**: Protein-coding exons, introns, non-coding RNA, RNA folding
- **The control regions**: Promoters, enhancers, insulators, chromatin states
- **The actual words**: Regulatory motifs, high-resolution accessibility maps
- **The regulators**: Transcription factors, chromatin modifiers, nucleosomes
- **The dynamics**: Changing maps between cell types, across development
- **The networks**: regulator → enhancer → target, ChIP-seq, correlated activity
- **The grammars**: TF/motif/mark combinations, predictive models
- **Human variation**: Human diversity, population genomics, linkage maps
- **Evolution**: Phylogenetics, phylogenomics, coalescent, human ancestry
- **GWAS/QTLs**: Genome variation ↔ organismal/molecular phenotypes
- **Disease**: Personal (epi)genomics, pharmacogenomics, synthetic biology
Goals for today: Course Introduction

1. Course overview:
   - Staff, students, responses to student survey
   - Foundations, frontiers, textbook, homework, quiz
   - Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   - What makes our field unique

3. Overview of course modules
   - Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   - Central Dogma of Molecular Biology
   - DNA, Epigenomics, RNA, Protein, Networks
   - Human genetics, Drug Discovery
# Course at a Glance

<table>
<thead>
<tr>
<th>Homeworks</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>Thu-Sep-07</td>
<td>L1</td>
<td>Course Overview, Machine Learning, Deep Learning, Inference, Genome, Proteins, Chemistry, Imaging</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Tue-Sep-12</td>
<td>L2</td>
<td>Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-Discr Learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Thu-Sep-14</td>
<td>L3</td>
<td>Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures</td>
</tr>
<tr>
<td>HW1 out Tue</td>
<td></td>
<td>2</td>
<td>Thu-Sep-14</td>
<td>L3</td>
<td>Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-Discr Learning</td>
</tr>
<tr>
<td>0=Self Introductions</td>
<td></td>
<td>2</td>
<td>Thu-Sep-14</td>
<td>L3</td>
<td>Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Tue-Sep-19</td>
<td>L4</td>
<td>Regulatory Genomics, Motifs, Information, ChIP, Gibbs Sampling, EM, CNNs for Genome Parsing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Thu-Sep-21</td>
<td>L5</td>
<td>Epigenomics: Signal Modeling, Peak calling, Chromatin states, 3D structure, Hi-C, Genome Topology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Tue-Sep-26</td>
<td>L6</td>
<td>Modeling Small Molecules using GNNs, Property Predictions; Characterizing Synergistic Effects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Thu-Sep-28</td>
<td>L7</td>
<td>Generative Models for Lead Optimization of Small Molecules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Thu-Sep-28</td>
<td>L7</td>
<td>Generative Models for Lead Optimization of Small Molecules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Tue-Oct-03</td>
<td>L8</td>
<td>Modeling Proteins using sequence and 3D Models; Protein Folding</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Thu-Oct-05</td>
<td>L9</td>
<td>Generative Models for Binding, Docking and Design</td>
</tr>
<tr>
<td>HW2 out Tue</td>
<td></td>
<td>5</td>
<td>Tue-Oct-10</td>
<td>-</td>
<td>Holiday Columbus Day / Indigenous People Day</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Thu-Oct-12</td>
<td>L10</td>
<td>Uncertainty Estimation; Handling Bias in the Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Thu-Oct-12</td>
<td>L10</td>
<td>Uncertainty Estimation; Handling Bias in the Data</td>
</tr>
<tr>
<td>HW3 out Tue</td>
<td></td>
<td>6</td>
<td>Thu-Oct-17</td>
<td>L11</td>
<td>Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function</td>
</tr>
<tr>
<td>2=Project proposal+feasibility</td>
<td></td>
<td>6</td>
<td>Thu-Oct-17</td>
<td>L11</td>
<td>Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Tue-Oct-24</td>
<td>L12</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS, multi-modal QTLs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Tue-Oct-26</td>
<td>L13</td>
<td>Single-cell genomics, sc-guMIcS, non-linear embeddings, spatial transcriptomics, next-gen technologies</td>
</tr>
<tr>
<td>HW3 due Mon</td>
<td></td>
<td>8</td>
<td>Tue-Oct-26</td>
<td>L14</td>
<td>Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimensionality Reduction, TF-enhancer-gene circuitry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Tue-Oct-31</td>
<td>L15</td>
<td>Electronic Health Records, AI/IOUs, UKBiobank, Medical Genomics, Population-Scale Cohorts, Multi-Ancestry</td>
</tr>
<tr>
<td>3=Offhrs, Update, Feedback</td>
<td></td>
<td>9</td>
<td>Thu-Nov-02</td>
<td>L16</td>
<td>Imaging methods for biological applications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Tue-Nov-07</td>
<td>L17</td>
<td>Comparative genomics, Alignment, Conservation, Evolutionary signatures, PhyloCSF, RNA structure, MorfBLS2conf</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Thu-Nov-09</td>
<td>L18</td>
<td>Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, IncRNAs, RNA modifications, miRNA</td>
</tr>
<tr>
<td>HW4 out Tue</td>
<td></td>
<td>10</td>
<td>Tue-Nov-14</td>
<td>L19</td>
<td>Introduction to Drug Discovery, Case Study: Herceptin</td>
</tr>
<tr>
<td>4=Midcourse report</td>
<td></td>
<td>10</td>
<td>Thu-Nov-16</td>
<td>L20</td>
<td>In-Class Quiz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu-Nov-21</td>
<td>L21</td>
<td>Drug Localization &amp; Delivery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Thu-Nov-23</td>
<td>-</td>
<td>Thanksgiving Holiday</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Tue-Nov-28</td>
<td>L22</td>
<td>Genome Engineering, CRISPR/Cas9, Massively Parallel Reporter Assays, Multiplexing, RNA therapeutics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Thu-Nov-30</td>
<td>L23</td>
<td>Computational Approaches for Metabolomics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Tue-Dec-05</td>
<td>L24</td>
<td>Immunology and Vaccine Design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Thu-Dec-07</td>
<td>L25</td>
<td>Computational Biology at Novartis</td>
</tr>
<tr>
<td>No HW</td>
<td></td>
<td>15</td>
<td>Tue-Dec-12</td>
<td>L26</td>
<td>Project Presentations (6-8 mins per team), Write-ups due Mon, Slides Tue at 10am, Presentations at 1pm</td>
</tr>
</tbody>
</table>

## Project planning

- HW1 out Tue
- HW2 out Tue
- HW3 out Tue
- HW4 out Tue

## Project execution

- HW1 due Mon
- HW2 due Mon
- HW3 due Mon
- HW4 due Mon

## Final Projects

15 Tue-Dec-12 L26 Project Presentations (6-8 mins per team), Write-ups due Mon, Slides Tue at 10am, Presentations at 1pm
Challenges in Computational Biology

1. Gene Finding
2. Sequence alignment
3. Database lookup
4. Genome Assembly
5. Regulatory motif discovery
6. Comparative Genomics
7. Evolutionary Theory
8. Gene expression analysis
9. Cluster discovery
10. Gibbs sampling
11. Protein network analysis
12. Metabolic modelling
13. Emerging network properties
### Aligning and Modeling Genomes

**Foundations vs. frontiers**
- **Foundations**: Classical computational methods / biological topics
- **Frontiers**: Latest developments, open questions, research areas
- **Duality for each**: basic problems / fundamental techniques

**Sequence alignment:**
- Local/global alignment: infer nucleotide-level evolutionary events
- Database search: scan for regions that may have common ancestry

**Hidden Markov Models**
- Hidden Markov Models (HMMs): Central tool in CS
- Decoding, evaluation, parsing, likelihood, scoring
Dynamic Programming Algorithms: Align, HMMs

- Sequence alignment
- Hidden Markov Models

- DP: Core computational technique
  - Pervasive in computer science, and computational biology
  - Fully explore exponential search spaces in poly time!
  - Greedy algorithms will not work, back-tracking, saving soln
  - Special requirements: Optimal substructure
  - Found in: alignment, HMMs, phylogeny, genetics, pop gen…
Gene expression analysis and transcripts

- **Computational foundations:**
  - Unsupervised Learning: Expectation Maximization
  - Supervised learning: generative/discriminative models
  - Read mapping, significance testing, splice graphs

- **Biological frontiers:**
  - PS2: Modeling conservation, GC content, CpG islands
  - L6/L7: Genome annotation and parsing
  - L8: Gene expression analysis: cluster genes/conditions
  - L9: Regulatory motif discovery: EM, gibbs sampling, info
Natural 1\textsuperscript{st} step: group similar rows/columns

**Clustering**

- Similar cell types
- Similarly-behaving groups of genes

Reveal common gene behaviors

Reveal common ‘conditions’

*Armstrong, Nature Gen 2002*

*Alizadeh, Nature 2000*
If labels are known: find more of same type

Classification

→ Classify diseases
→ Classify genes in different pathways

Find features that distinguish known classes
Find additional members of existing gene classes
Predict function of uncharacterized genes
Epigenomics and gene regulation

• **Computational Foundations**
  – Hidden Markov Models (HMMs): Central tool in CS
  – Decoding, evaluation, parsing, likelihood, scoring
  – Unsupervised Learning: Expectation Maximization
  – Supervised learning: generative/discriminative models

• **Biological frontiers:**
  – PS2: Modeling conservation, GC content, CpG islands
  – L6/L7: Genome annotation and parsing
  – L8: Gene expression analysis: cluster genes/conditions
  – L9: Regulatory motif discovery: EM, gibbs sampling, info
Motifs summarize TF sequence specificity

- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
  - Independence
  - Fixed spacing
Starting positions ⇔ Motif matrix

- given **aligned** sequences  ⇨ easy to compute profile matrix

- easy to find starting position probabilities

---

Key idea: Iterative procedure for estimating both, given uncertainty
(learning problem with hidden variables: the starting positions)
Multivariate HMM for Chromatin States

Enhancer | Transcription Start Site | Transcribed Region | DNA

Observed chromatin marks. Called based on a poisson distribution

1: K4me1
2: K4me3
3: K4me3
4: K4me3
5: K4me1
6: K36me3

High Probability Chromatin Marks in State

1: 0.8 K4me1 0.8 K27ac
2: 0.9 K4me3 0.8 K4me1
3: 0.9 K4me3
4: 0.7 K4me1
5: 
6: 0.9 K36me3

Most likely Hidden State

1 → 2 → 3 → 4 → 6 → 6 → 6 → 6 → 5 → 5 → 5

200bp intervals

All probabilities are learned from the data

Ernst and Kellis
Nature Biotech 2010
Phylogenetics / Phylogenomics

- Phylogenetics: Evolutionary models, Tree building, Phylo inference
- Phylogenomics: gene/species trees, reconciliation, coalescent, pops

Population genomics:

- Learning population history from genetic data (David Reich)
- Statistical genetics: disease mapping in populations (Mark Daly)
- Measuring natural selection in human populations (Pardis Sabeti)
- The missing heritability in genome-wide associations (Yaniv Erlich)

And we’re done! Last pset Nov 21\textsuperscript{st}, In-class quiz on Nov 22\textsuperscript{nd}

- No lab 4! Then entire focus shifts to projects, Thanksgiving, Frontiers
Characterizing sub-threshold variants in heart arrhythmia

Focus on sub-threshold variants (e.g. rs1743292 P=10^{-4.2})

Trait: QRS/QT interval

(1) Large cohorts, (2) many known hits
(3) well-characterized tissue drivers
Alignment: all species/genes share common ancestry

Slide credit: Serafim Batzoglou
Extinctions part of life
Phylogenetics

General Problem:
Infer complete ancestry of a set of ‘objects’ based on knowledge of their ‘traits’

‘Objects’ can be: Species, Genes, Cell types, Diseases, Cancers, Languages, Faiths, Cars, Architectural Styles

‘Traits’ can be: Morphological, molecular, gene expression, TF binding, motifs, words…

Historical record varies: Fossils, imprints, timing of geological events, ‘living fossils’, sequencing of extinct species, paintings, stories.

Today: Phylogenies using only extant species data ➔ gene trees (paralog / ortholog / homolog trees)
Challenges in Computational Biology

1. Gene Finding
2. Sequence alignment
3. Database lookup
4. Genome Assembly
5. Regulatory motif discovery
6. Comparative Genomics
7. Evolutionary Theory
8. Gene expression analysis
9. Cluster discovery
10. Gibbs sampling
11. Protein network analysis
12. Metabolic modelling
13. Emerging network properties
Structure of genetic code $\Leftrightarrow$ evolutionary signatures

- Substitutions that preserve AA properties tolerated in coding exons
- Leads to specific evolutionary signatures associated with protein-coding genes
- The code itself could be rediscovered simply based on observed substitution patterns

These specify different rates of codon substitution, which in turn lead to different probabilities of any given alignment:

$$\Pr(\text{Leaves}; Q_C, t) = \frac{1}{10^{117}}$$

$$\Pr(\text{Leaves}; Q_N, t) = \frac{1}{10^{152}}$$

$$\Pr(\text{Leaves}; Q_C, t) = \frac{1}{10^{275}}$$

$$\Pr(\text{Leaves}; Q_N, t) = \frac{1}{10^{254}}$$
Distance matrix ↔ Phylogenetic tree

<table>
<thead>
<tr>
<th></th>
<th>Hum</th>
<th>Mou</th>
<th>Rat</th>
<th>Dog</th>
<th>Cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Mouse</td>
<td>h.y.m</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Rat</td>
<td>h.y.r</td>
<td>m.r</td>
<td>0</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Dog</td>
<td>h.z.x.d</td>
<td>m.y.z.x.d</td>
<td>r.y.z.x.d</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Cat</td>
<td>h.z.x.c</td>
<td>m.y.z.x.c</td>
<td>r.y.z.x.c</td>
<td>d.c</td>
<td>0</td>
</tr>
</tbody>
</table>

**Goal:**
Minimize discrepancy between observed distances and tree-based distances

Map distances $D_{ij}$ to a tree

Tree implies a distance matrix $M_{ij}$

$$\min \sum_{ij} (D_{ij} - M_{ij})^2$$
‘Peeling’ algorithm for \( P(D|B,T) \) term

1. Assume sites \( j \) evolve independently.
   - Treat each column of the alignment in isolation

2. Assume branch independence, conditioned on parent
   - Expand total joint probability into prod of \( P(x_i|x_{\text{parent}}, t_i) \)
   - Only \( P(x_{2n-1}) \) remains, root prior, background nucl. freq.

3. We know how to compute \( P(x_i|x_{\text{parent}(i)}, t_i) \) for fixed pair
   - Defined by our sequence model (JC, K2P, HKY, etc)
   - Easily calculate for any given assignment of internal nodes

4. As internal node values are not known \( \rightarrow \) **marginalize**
   - Sum over all possible values of all internal/root nodes
   - Let \( x_{n+1}, \ldots, x_{2n-1} \) represent seqs of \( n-1 \) internal nodes
Two types of gene-tree species-tree reconciliation

- **Coalescent models of alleles in populations**
  - Deal with 1-to-1 orthologs
  - Estimate divergence times, pop sizes, etc
  - Models move backward in time
  - Cannot cope with duplication and loss

- **DL models of genes in species**
  - Deal with paralogous families
  - Estimate birth death rates
  - Models move forward in time
  - Cannot cope with incomplete lineage sorting

Gene tree

- A
- B
- C

Species tree

- A
- B
- C

Coalescence

Duplication & Loss

→ DLCoal
Evidence of Neanderthal → Human gene flow

Neand.-African tMRCA
~ 825 kya

Average human tMRCA
~ 500 kya

Population split
~ 350 kya

Human-human divergence is
AVERAGE

Human-human divergence is
HIGH
## Course at a Glance

<table>
<thead>
<tr>
<th>Homework</th>
<th>Project / Mentoring (Fri)</th>
<th>Week</th>
<th>Date</th>
<th>Lec</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>No HW</td>
<td></td>
<td>1</td>
<td>Thu-Sep-07</td>
<td>L1</td>
<td>Course Overview, Machine Learning, Deep Learning, Inference, Genome, Proteins, Chemistry, Imaging</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Modeling Biology I: DNA / RNA</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Tue-Sep-12</td>
<td>L2</td>
<td>Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-Discr Learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Thu-Sep-14</td>
<td>L3</td>
<td>Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Tue-Sep-19</td>
<td>L4</td>
<td>Regulatory Genomics, Motifs, Information, Chip, Gibbs Sampling, EM, CNNs for Genome Parsing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Thu-Sep-21</td>
<td>L5</td>
<td>Epigenomics: Signal Modeling, Peak calling, Chromatin states, 3D structure, Hi-C, Genome Topology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Thu-Oct-06</td>
<td>L6</td>
<td>Modeling Small Molecules using GNNs; Property Predictions; Characterizing Synergistic Effects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Thu-Oct-08</td>
<td>L7</td>
<td>Generative Models for Lead Optimization of Small Molecules</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Tue-Oct-03</td>
<td>L8</td>
<td>Modeling Proteins using sequence and 3D Models; Protein Folding</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Thu-Oct-05</td>
<td>L9</td>
<td>Generative Models for Binding, Docking and Design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Tue-Oct-10</td>
<td>-</td>
<td>Holiday Columbus Day / Indigenous People Day</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Thu-Oct-12</td>
<td>L10</td>
<td>Uncertainty Estimation; Handling Bias in the Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Thu-Oct-17</td>
<td>L11</td>
<td>Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Thu-Oct-19</td>
<td>L12</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMFWAS, multi-modal QTLs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Tue-Oct-24</td>
<td>L13</td>
<td>Single-cell genomics, sc-multomics, non-linear embeddings, spatial transcriptomics, next-gen technologies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Thu-Oct-26</td>
<td>L14</td>
<td>Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimensionality Reduction, TF-enhancer-gene circuitry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Tue-Oct-31</td>
<td>L15</td>
<td>Electronic Health Records, AI/IOUs, UKBioBank, Medical Genomics, Population-Scale Cohorts, Multi-Ancestry</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Thu-Nov-02</td>
<td>L16</td>
<td>Imaging methods for biological applications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Tue-Nov-07</td>
<td>L17</td>
<td>Comparative genomics, Alignment, Conservation, Evolutionary signatures, PhyloCSF, RNA structure, MorfEls2conf</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Thu-Nov-09</td>
<td>L18</td>
<td>Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, IncRNAs, RNA modifications, miA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu-Nov-14</td>
<td>L19</td>
<td>Introduction to Drug Discovery, Case Study: Herceptin</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu-Nov-16</td>
<td>L20</td>
<td>In-Class Quiz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Thu-Nov-21</td>
<td>L21</td>
<td>Drug Localization &amp; Delivery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>Thu-Nov-23</td>
<td>-</td>
<td>Thanksgiving Holiday</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Tue-Nov-28</td>
<td>L22</td>
<td>Genome Engineering, CRISPR/Cas9, Massively Parallel Reporter Assays, Multiplexing, RNA therapeutics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Thu-Nov-30</td>
<td>L23</td>
<td>Computational Approaches for Metabolomics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Tue-Dec-05</td>
<td>L24</td>
<td>Immunology and Vaccine Design</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Thu-Dec-07</td>
<td>L25</td>
<td>Computational Biology at Novartis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Tue-Dec-12</td>
<td>L26</td>
<td>Project Presentations (6-8 mins per team). Write-ups due Mon, Slides Tue at 10am, Presentations at 1pm</td>
</tr>
</tbody>
</table>

### Final Projects

- **WriteUp, Presentations**
- **Project Presentations (6-8 mins per team). Write-ups due Mon, Slides Tue at 10am, Presentations at 1pm**
Goals for today: Course Introduction

1. Course overview:
   – Staff, students, responses to student survey
   – Foundations, frontiers, textbook, homework, quiz
   – Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   – What makes our field unique

3. Overview of course modules
   – Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   – Central Dogma of Molecular Biology
   – DNA, Epigenomics, RNA, Protein, Networks
   – Human genetics, Drug Discovery
Biology primer

Quick introduction to molecular biology and information transfer within the cell
“Central dogma” of Molecular Biology

DNA makes RNA

RNA makes Protein
DNA: The double helix

- The most noble molecule of our time
DNA: the molecule of heredity

• Self-complementarity sets molecular basis of heredity
  – Knowing one strand, creates a template for the other
  – “It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.” Watson & Crick, 1953
DNA: chemical details

- Bases hidden on the inside
- Phosphate backbone outside

- Weak hydrogen bonds hold the two strands together
- This allows low-energy opening and re-closing of two strands

- Anti-parallel strands
- Extension 5’ → 3’ tri-phosphate coming from newly added nucleotide

The only pairings are:
- A with T
- C with G
DNA: the four bases

### The Nucleotides of DNA

<table>
<thead>
<tr>
<th></th>
<th>Adenine</th>
<th>Guanosine</th>
<th>Thymine</th>
<th>Cytosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purine</td>
<td></td>
<td>Purine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td></td>
<td></td>
<td>Pyrimidine</td>
<td>Pyrimidine</td>
</tr>
<tr>
<td>Strong</td>
<td></td>
<td></td>
<td>Weak</td>
<td></td>
</tr>
<tr>
<td>Amino</td>
<td></td>
<td></td>
<td></td>
<td>Amino</td>
</tr>
<tr>
<td>Keto</td>
<td></td>
<td></td>
<td></td>
<td>Keto</td>
</tr>
</tbody>
</table>
Goals for today: Course Introduction

1. Course overview:
   – Staff, students, responses to student survey
   – Foundations, frontiers, textbook, homework, quiz
   – Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   – What makes our field unique

3. Overview of course modules
   – Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   – Central Dogma of Molecular Biology
   – DNA, Epigenomics, RNA, Protein, Networks
   – Human genetics, Drug Discovery
“Central dogma” of Molecular Biology

DNA makes RNA makes Protein

Epigenomics
Chromosomes inside the cell

- Prokaryote cell
- Eukaryote cell
DNA packaging

• Why packaging
  – DNA is very long
  – Cell is very small

• Compression
  – Chromosome is 50,000 times shorter than extended DNA

• Using the DNA
  – Before a piece of DNA is used for anything, this compact structure must open locally

• Now emerging:
  – Role of accessibility
  – State in chromatin itself
  – Role of 3D interactions
Diverse epigenetic modifications

**Epigenetic Mechanisms**
- Development (in utero, childhood)
- Environmental chemicals
- Drugs/Pharmaceuticals
- Aging
- Diet

**DNA Methylation**
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

**Histone Modification**
The binding of epigenetic factors to histone “tails” alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

**Health Endpoints**
- Cancer
- Autoimmune disease
- Mental disorders
- Diabetes

Image source: http://nihroadmap.nih.gov/epigenomics/
Diversity of epigenetic modifications

- 100+ different histone modifications
  - Histone protein $\rightarrow$ H3/H4/H2A/H2B
  - AA residue $\rightarrow$ Lysine4(K4)/K36…
  - Chemical modification $\rightarrow$ Met/Pho/Ubi
  - Number $\rightarrow$ Me-Me-Me(me3)
  - Shorthand: H3K4me3, H2BK5ac

- In addition:
  - DNA modifications
  - Methyl-C in CpG / Methyl-Adenosine
  - Nucleosome positioning
  - DNA accessibility

- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

DNA wrapped around histone proteins
Epigenomics Roadmap across 100+ tissues/cell types

Diverse tissues and cells:
1. Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood...)
2. Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood...)
3. ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench, trophobl)

Diverse epigenomic assays:
1. Histone modifications
   - H3K4me3, H3K4me1
   - H3K36me3
   - H3K27me3, H3K9me3
   - H3K27/9ac, +20 more
2. Open chromatin:
   - DNase
3. DNA methylation:
   - WGBS, RRBS, MRE/MeDIP
4. Gene expression
   - RNA-seq, Exon Arrays

Art: Rae Senarighi, Richard Sandstrom
Deep sampling of 9 reference epigenomes (e.g. IMR90)

Chromatin state+RNA+DNAse+28 histone marks+WGBS+Hi-C
Diverse chromatin signatures encode epigenomic state

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl

1. Chromatin fiber
2. DNA
3. Nucleus
4. Nucleosome

- H3K4me3
- H3K4me1
- H3K27ac
- H3K36me3
- H4K20me3
- H3K79me3
- H3K27me3
- H3K9me3
- H3K9ac
- H3K18ac
Chromatin state annotations across 127 epigenomes

Reveal epigenomic variability: enh/prom/tx/repr/het

Anshul Kundaje
Goals for today: Course Introduction

1. Course overview:
   – Staff, students, responses to student survey
   – Foundations, frontiers, textbook, homework, quiz
   – Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   – What makes our field unique

3. Overview of course modules
   – Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   – Central Dogma of Molecular Biology
   – DNA, Epigenomics, RNA, Protein, Networks
   – Human genetics, Drug Discovery
“Central dogma” of Molecular Biology
Genes control the making of cell parts

- The gene is a fundamental unit of inheritance
  - Each DNA molecule $\leftrightarrow$ 10,000+ genes
  - 1 gene $\leftrightarrow$ 1 functional element (one “part” of cell machinery)
  - Every time a “part” is made, the corresponding gene is:
    - Copied into mRNA, transported, used as blueprint to make protein

- RNA is a temporary copy
  - The medium for transporting genetic information from the DNA information repository to the protein-making machinery is an RNA molecule
  - The more parts are needed, the more copies are made
  - Each mRNA only lasts a limited time before degradation
mRNA: The messenger

- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar

<table>
<thead>
<tr>
<th>DNA</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription</td>
<td>Translation</td>
</tr>
</tbody>
</table>

- Compatible base-pairing in hybrid

| uracil (RNA) | thymine (DNA) |
In Eukaryotes, not every part of a gene is coding
- Functional exons interrupted by non-translated introns
- During pre-mRNA maturation, introns are spliced out
- In humans, primary transcript can be $10^6$ bp long

From pre-mRNA to mRNA: Splicing

- Alternative splicing can yield different exon subsets for the same gene, and hence different protein products
RNA can be functional

- Single Strand allows complex structure
  - Self-complementary regions form helical stems
  - Three-dimensional structure allows functionality of RNA

- Four types of RNA
  - mRNA: messenger of genetic information
  - tRNA: codon-to-amino acid specificity
  - rRNA: core of the ribosome
  - snRNA: splicing reactions

- To be continued…
  - We’ll learn more in a dedicated lecture on RNA world
  - Once upon a time, before DNA and protein, RNA did all
RNA structure: 2\textsuperscript{nd}ary and 3\textsuperscript{rd}ary
Splicing machinery made of RNA

5' splice site  
U1 snRNP  
exon 1  
intron  
exon 2  
3' splice site

U4/U6 snRNP  
U5 snRNP

U4/U6 snRNP  
U5 snRNP

LARIAT FORMATION AND 5' SPlice SITE CLEAVAGE

3' SPLICE SITE CLEAVAGE AND JOINING OF TWO EXON SEQUENCES

excised intron sequence in the form of a lariat (will be degraded in nucleus)

portion of a primary transcript

portion of mRNA
Goals for today: Course Introduction

1. Course overview:
   - Staff, students, responses to student survey
   - Foundations, frontiers, textbook, homework, quiz
   - Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   - What makes our field unique

3. Overview of course modules
   - Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   - Central Dogma of Molecular Biology
   - DNA, Epigenomics, RNA, Protein, Networks
   - Human genetics, Drug Discovery
“Central dogma” of Molecular Biology

DNA \rightarrow \text{makes} \rightarrow RNA \rightarrow \text{makes} \rightarrow \text{Protein}
Proteins carry out the cell’s chemistry

- More complex polymer
  - Nucleic Acids have 4 building blocks
  - Proteins have 20. Greater versatility
  - Each amino acid has specific properties
- Sequence → Structure → Function
  - The amino acid sequence determines the three-dimensional fold of protein
  - The protein’s function largely depends on the features of the 3D structure
- Proteins play diverse roles
  - Catalysis, binding, cell structure, signaling, transport, metabolism
Protein structure

**Helix-turn-helix**

Common motif for DNA-binding proteins that often play a regulatory role as mRNA level transcription factors.

**Beta-barrel**

Some antiparallel b-sheet domains are better described as b-barrels rather than b-sandwiches, for example streptavidin and porin. Note that some structures are intermediate between the extreme barrel and sandwich arrangements.

**Alpha-beta horseshoe**

This placental ribonuclease inhibitor is a cytosolic protein that binds extremely strongly to any ribonuclease that may leak into the cytosol. 17-stranded parallel b-sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks as though it should. The strands are only very slightly slanted, being nearly parallel to the central `axis'.

![Diagram of Helix-turn-helix](image1)

![Diagram of Beta-barrel](image2)

![Diagram of Alpha-beta horseshoe](image3)
Protein building blocks

- **Amino Acids**

  - Amino acids with hydrophobic side groups:
    - Valine (Val)
    - Leucine (Leu)
    - Isoleucine (Ile)
    - Methionine (Met)
    - Phenylalanine (Phe)

  - Amino acids with hydrophilic side groups:
    - Asparagine (Asn)
    - Glutamic acid (Glu)
    - Glutamine (Gln)
    - Histidine (His)
    - Lysine (Lys)
    - Arginine (Arg)

  - Amino acids that are in between:
    - Glycine (Gly)
    - Alanine (Ala)
    - Serine (Ser)
    - Threonine (Thr)
    - Tyrosine (Tyr)
    - Tryptophan (Trp)

- Example of amino acid sequence and chemical structures.
From RNA to protein: Translation

- Ribosome
- tRNA
The Genetic Code

Use evolutionary and compositional properties to computationally discover protein-coding genes.
Summary: The Central Dogma

DNA makes RNA makes Protein

Inheritance

Messages

Reactions

protein
Goals for today: Course Introduction

1. Course overview:
   – Staff, students, responses to student survey
   – Foundations, frontiers, textbook, homework, quiz
   – Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   – What makes our field unique

3. Overview of course modules
   – Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   – Central Dogma of Molecular Biology
   – DNA, Epigenomics, RNA, Protein, Networks
   – Human genetics, Drug Discovery
Cellular dynamics and regulation
How cells move through this Central Dogma

Gene regulation

DNA makes RNA

RNA makes Protein
Animal/Human gene regulation: One genome ⇔ Many cell types

ACCAGTTACGACGGTCA
GGTACTGATACCCCCAA
ACGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCCTACAAAC
GTTTACAGGGTTACCGT
TGGGATTTGAAAAAAG
TTTGAGTTGGTTTTTTC
ACGTTAGAACCCTACCGT
TACCAGTA

Image Source: wikipedia
Eukaryotic Gene Regulation
Diverse roles for regulatory non-coding RNAs

• **Small RNA pathways (18-21 nt)**
  – microRNAs:
    • Repress genes by targeting their 3’ UTRs by complementarity
    • Double-stranded RNA is then recognized and degraded
    • Recently found to also target promoter regions in rare cases
  – piwiRNAs
    • Target and repress transposable elements in germline
  – snoRNAs
  – 21U-RNAs

• **Long non-coding RNAs (1000s nt, many exons)**
  – Scaffolds for protein/TF binding
  – Scaffolds for 3D structure of RNA
Regulation of Gene Expression

- Upstream of genes are **promoter** regions
- Contain promoter sequences or **motifs**
- **Transcription factors** (TFs) bind to motifs
- TFs recruit **RNA polymerase**
- Gene transcription

Examples:

- ATATAA
- CTGATAAGCAG
- GTGATACA
- AGCGAGACG
- AANAATAAA
- TAAATTAA
- GAACAAGAG
- AAATAAAA
Predicted motif drivers of enhancer modules

- Activator and repressor motifs consistent with tissues

Pouya Kheradpour
Network components reveal functional modules

- Feed-forward loops in developmental patterning
- Cooperation of master reg. & downstream reg.

Zeitlinger et al, Genes & Development 2007
Systematic motif dissection in 2000 enhancers: 5 activators and 2 repressors in 2 cell lines

Kheradpour et al. Genome Research 2013

54000+ measurements (x2 cells, 2x repl)

Active in HepG2 cells
- HNF1
- HNF4
- FOXA
- GATA
- NRF2
- ZFP161
- GFI1

Active in K562 cells
- HepG2 enhancers
- GFI1
- GATA
- NRF2
- K562 enhancers

Manipulations to the wildtype sequence
- Scramble
- Removal
- Max 1-bp decrease
- Least 1-bp change
- Max 1-bp increase
- Random 1-bp change (x2)

For ~10% of tested sequences

Add unique 10 nt tag for each candidate enhancer sequence (x10)

Sequences from other selected motif matches

Synthesize and construct plasmid pool

Total of ~55,000 distinct plasmids

Transfect K562 and HepG2 cells

Count plasmid tags (~30M reads each)

Count mRNA tags from each
Emerging properties of regulatory networks

- Hierarchical levels of regulatory control
  - Small number of backward-pointing edges
- Specific / distinct feedback by microRNAs at each level
  - Two classes of TFs: miRNA regulators and miR-regulated
From Systems Biology to Synthetic Biology

Jim Collins

- Components with known properties
- Assemble based on engineering goals / principles
- Implement within engineered cells and organisms
- Study behavior & adjust as needed

Jay Keasling
Over-express a single microRNA leads to new wing

- Discovery of sense/anti-sense miRNAs
- Regulatory switch selects between two developmental programs
- By over-expressing one strand (miRNAas), the balance is tilted
- Wing program launched vs. haltere

Stark et al, Genes & Development 2007
Goals for today: Course Introduction

1. Course overview:
   – Staff, students, responses to student survey
   – Foundations, frontiers, textbook, homework, quiz
   – Final project: teams, mentorship, challenge, relevance, originality, achievement, presentation

2. Why Computational Biology;
   – What makes our field unique

3. Overview of course modules
   – Genomes, Expression, Epigenomics, Networks, Genetics, Evolution, Frontiers

4. Biology primer (in the context of this course)
   – Central Dogma of Molecular Biology
   – DNA, Epigenomics, RNA, Protein, Networks
   – Human genetics, Drug Discovery
Brief intro to Human Genetics
The role of genetic alterations

DNA makes RNA

RNA makes Protein
Brief intro to human genetics

• Human genome: 3.2B letters, 2 copies, 23 chromosomes, 20k genes, ~3M common SNPs, ~500k haplotype blocks

Published Genome-Wide Associations through 09/2011
1,617 published GWA at p≤5X10⁻⁸ for 249 traits

NHGRI GWA Catalog
www.genome.gov/GWASTudies
The power and challenge of disease-association studies

- Large associated blocks with many variants: Fine-mapping challenge
- No information on cell type/mechanism, most variants non-coding
  ➜ Epigenomic annotations help find relevant cell types / nucleotides

Slide credit: Luke Ward, Mark Daly
The power of GWAS: reveal new disease genes

rs11209026 A G
Cases 22 976
Controls 68 932
Chi-sq = 24.5, p=7.3 x 10^-7

IL23R cytokine receptor on a subset of effector T-cells
More than 100 distinct regions of the genome associated to schizophrenia!!!
Interpreting non-coding variants

- Disease-associated SNPs enriched for enhancers in relevant cell types
- E.g. lupus SNP in GM enhancer disrupts Ets1 predicted activator
Mechanistic predictions for top disease-associated SNPs

**Lupus erythematosus in GM lymphoblastoid**
- **Disrupt activator Ets-1 motif**
  - **Loss of GM-specific activation**
  - **Loss of enhancer function**
  - **Loss of HLA-DRB1 expression**

**Erythrocyte phenotypes in K562 leukemia cells**
- **Creation of repressor Gfi1 motif**
  - **Gain K562-specific repression**
  - **Loss of enhancer function**
  - **Loss of CCDC162 expression**
Characterizing sub-threshold variants in heart arrhythmia

**Trait: QRS/QT interval**

(1) Large cohorts, (2) many known hits
(3) well-characterized tissue drivers

*Focus on sub-threshold variants (e.g. rs1743292 P=10^{-4.2})*
GWAS hits in enhancers of relevant cell types

<table>
<thead>
<tr>
<th>Trait</th>
<th>Abbrev</th>
<th>log</th>
<th>Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>ESC</td>
<td>4.7</td>
<td>0</td>
</tr>
<tr>
<td>Chronic lymphocytic leukemia</td>
<td>ESC</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Type 1 diabetes autoantibodies</td>
<td>Treg</td>
<td>4.8</td>
<td>0</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>Treg</td>
<td>4.1</td>
<td>0</td>
</tr>
<tr>
<td>Platelet counts</td>
<td>Th. pl.</td>
<td>4.6</td>
<td>0</td>
</tr>
<tr>
<td>Self-reported allergy</td>
<td>Th. pl.</td>
<td>4.9</td>
<td>0</td>
</tr>
<tr>
<td>Graves disease</td>
<td>Th. pl.</td>
<td>4.3</td>
<td>0</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>Th. pl.</td>
<td>4.7</td>
<td>0</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>Th. pl.</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Collage disease</td>
<td>Th. pl.</td>
<td>4.6</td>
<td>0</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>Th. pl.</td>
<td>5.5</td>
<td>0</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>Th. pl.</td>
<td>5.8</td>
<td>0</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>Th. pl.</td>
<td>5.4</td>
<td>0</td>
</tr>
<tr>
<td>Primary biliary cirrhosis</td>
<td>Th. pl.</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>Red blood cell traits</td>
<td>Th. pl.</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>Platelet counts</td>
<td>Th. pl.</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>Mean platelet volume</td>
<td>Th. pl.</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>Mean platelet volume</td>
<td>Th. pl.</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Th. pl.</td>
<td>5.6</td>
<td>0</td>
</tr>
<tr>
<td>Multiple myeloma</td>
<td>Th. pl.</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>Th. pl.</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>Th. pl.</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Acute respiratory distress</td>
<td>Th. pl.</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Liver glycemic levels (g-glut 1x)</td>
<td>Th. pl.</td>
<td>4.5</td>
<td>0</td>
</tr>
<tr>
<td>Amel. response to chem. (neutrophil)</td>
<td>Th. pl.</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>Th. pl.</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>High-like growth factors</td>
<td>Th. pl.</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Lactate dehydrogenase trans</td>
<td>Th. pl.</td>
<td>4.2</td>
<td>0</td>
</tr>
<tr>
<td>Cholesterol, total</td>
<td>Liver</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Cholesterol, total</td>
<td>Liver</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>LDL cholesterol</td>
<td>Liver</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>Liver</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>Metabolite levels</td>
<td>Liver</td>
<td>3.9</td>
<td>0</td>
</tr>
<tr>
<td>Platelet counts</td>
<td>Leuk</td>
<td>4.6</td>
<td>0</td>
</tr>
<tr>
<td>Primary biliary cirrhosis</td>
<td>Lymph</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>Leukocyte count per volume</td>
<td>Mnc. pl.</td>
<td>4.7</td>
<td>0</td>
</tr>
<tr>
<td>Inflammatory bowel disease</td>
<td>Mnc. pl.</td>
<td>4.7</td>
<td>0</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Mnc. pl.</td>
<td>4.7</td>
<td>0</td>
</tr>
<tr>
<td>Alzheimer's disease (late onset)</td>
<td>Bone</td>
<td>4.5</td>
<td>0</td>
</tr>
<tr>
<td>Fetal alcohol syndrome</td>
<td>Bone</td>
<td>4.5</td>
<td>0</td>
</tr>
</tbody>
</table>
Linking traits to their relevant cell/tissue types

- Liver
  - Cholesterol
  - Lipid metabolism phenotypes
  - Mean platelet volume
  - Mesenchymal
  - Liver
- Digestive
  - Adipocyte
  - Adipose Placenta
  - Fetal Muscle Trunk
  - Fetal Intestine
- Brain
  - Prefrontal Cortex
  - Cingulate Gyrus
- Heart
  - Left Ventricle
  - Mesenchymal
- B cells
  - CD4
  - CD45
  - CD19
  - Chronic lymphocytic leukemia
- T cells
  - CD4
  - CD8
  - Celiac and Rheumatoid arthritis
  - Self-reported allergy
- ES
  - Fibroblast
  - Keratinocyte
  - ES cell
- Platelet counts
- Alzheimer's (late onset)
- PR heart repolarization interval

- Type 1 diabetes autoantibodies
- Celiac
- Type 1 diabetes
- Crohn's
- Urate levels
- LDL cholesterol
- HDL cholesterol
- Liver enzyme levels (g-glut. transferase)
Methylation differences a causal component of AD

Methylation probes altered in AD are enriched in AD-associated SNPs

\[ \text{G} \rightarrow \text{M} \rightarrow \text{D} \]

\[ \text{G} \rightarrow \text{M} \leftarrow \text{D} \]

\[ \text{G} \rightarrow \text{D} \]

\[ \text{M} \leftarrow \]

Set-wise causality testing

AD predictive power reduced after removing meQTL effect
Uncovering the molecular basis of top obesity gene

Thermogenic stimuli (e.g. cold) → ARID5B KD (obesity) → C-to-T motif rescue (anti-obesity phenotypes)

Lean

ARID5B OE (anti-obesity) → IRX3, IRX5 knock-down★ (anti-obesity phenotypes)

ARID5B KD (obesity) → C-to-T motif rescue (anti-obesity phenotypes)

Obese

Browning mitochondrial thermogenesis

IRX3, IRX5 overexpression (pro-obesity phenotypes)

ARID5B OE (anti-obesity) → IRX3, IRX5 overexpression (pro-obesity phenotypes)

IRX3, IRX5 knock-down★ (anti-obesity phenotypes)

ARID5B KD (obesity) → T-to-C motif disruption (pro-obesity phenotypes)

UCP1, PGC1α, PRDM16

Lipid storage → White adipocytes
Model: beige $\Leftrightarrow$ white adipocyte development

Shift therapeutic focus from brain to adipocytes