Advanced Bioinformatics

Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2019

Colin Dewey
colin.dewey@wisc.edu
www.biostat.wisc.edu/bmi776/

Agenda Today

- Introductions
- Course information
- Overview of topics

Course Web Site

- www.biostat.wisc.edu/bmi776/
- Syllabus and policies
- Readings
- Tentative schedule
- Lecture slides (draft posted before lecture)
- Announcements
- Homework
- Project information
- Link to Piazza discussion board

Your Instructor: Colin Dewey

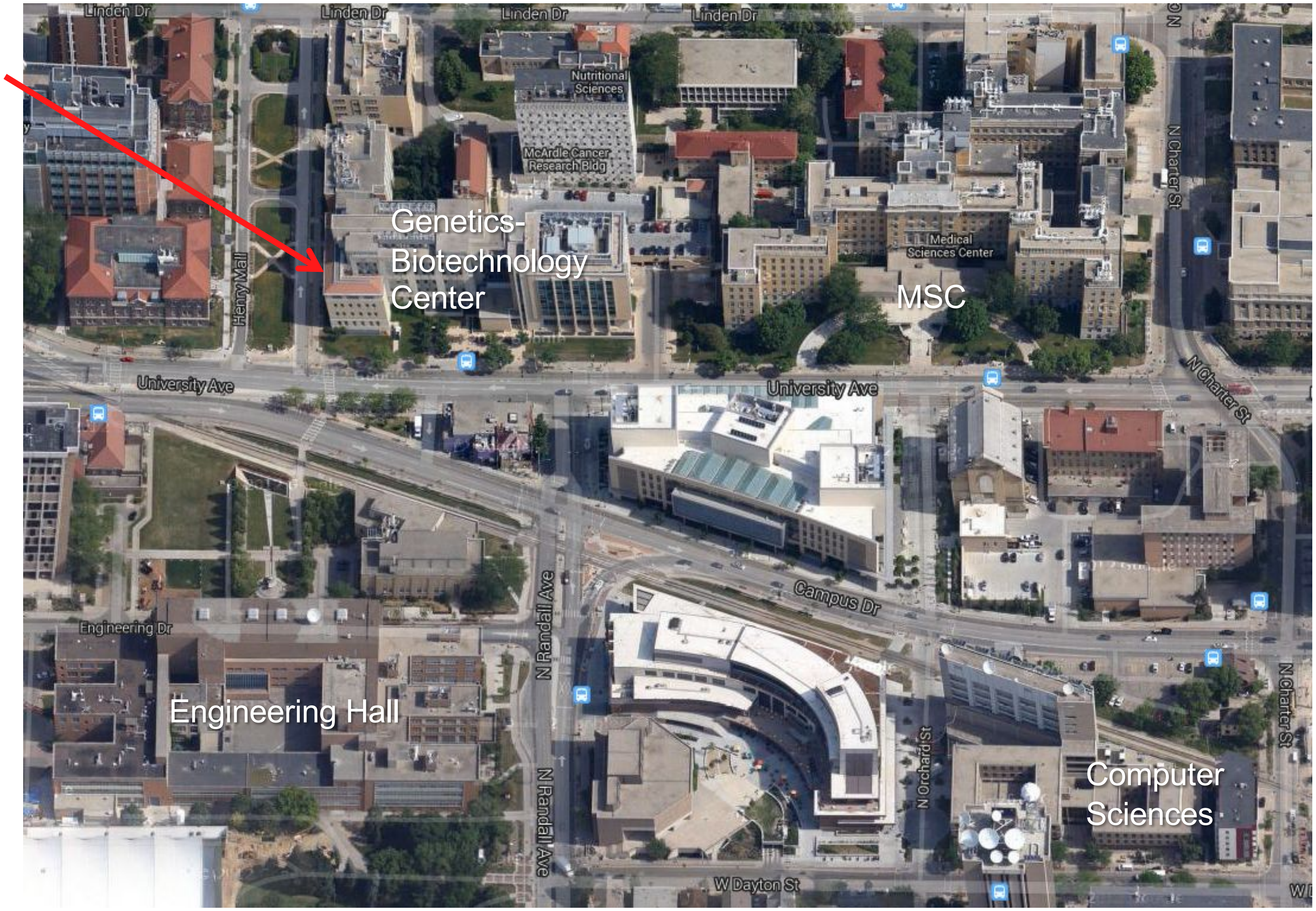
- email: colin.dewey@wisc.edu
- website: www.biostat.wisc.edu/~cdewey/
- office: 2128 Genetics-Biotechnology Center

- Professor in the department of Biostatistics & Medical Informatics with an affiliate appointment in Computer Sciences

- research interests: probabilistic modeling, biological sequence evolution, analysis of “next-generation” sequencing data (RNA-seq in particular), whole-genome alignment

Finding My Office:
2128 Genetics-Biotechnology Center

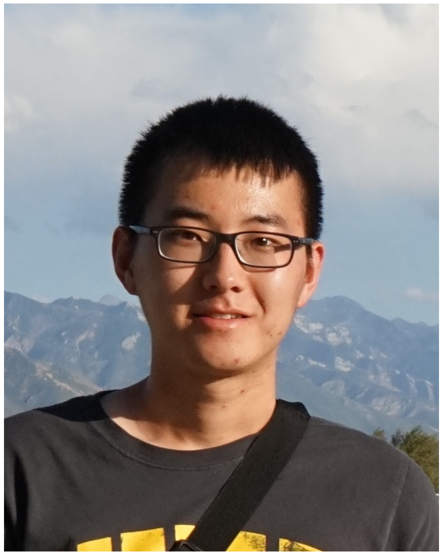
my office



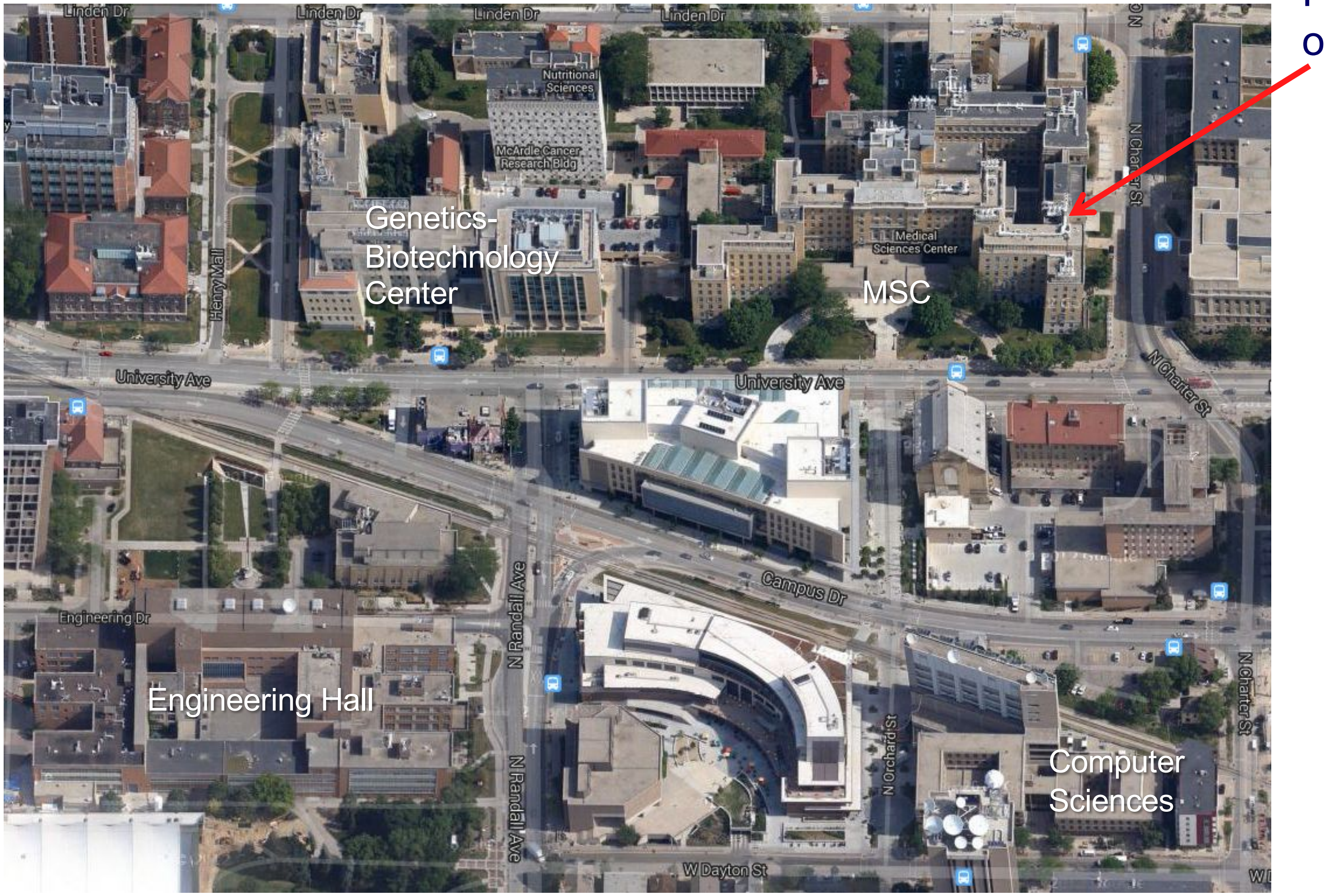
- slightly confusing building(s)
- best bet: use Henry Mall main entrance

Course TA

- Fangzhou Mu
  - fmu2@wisc.edu
  - MSC 6729
  - Graduate student
    - Pharmacy & CS



Finding Fangzhou's Office: MSC 6729



Fangzhou's office

- **very** confusing building
- best bet: use **420 North Charter St entrance**

Office Hours

- To be announced
- Will begin next week
- Doodle poll to determine a good office hour schedule for TA and me
  - Please fill out poll to increase the likelihood that our office hours will work for you!
- You are encouraged to visit our office hours!

You

- So that we can all get to know each other better, please tell us your
  - name
  - major or graduate program
  - research interests and/or topics you’re especially interested in learning about
  - favorite programming language

Course Requirements

- 4 or 5 homework assignments: ~40%
  - Written exercises
  - Programming (Python)
  - Computational experiments (e.g. measure the effect of varying parameter *x* in algorithm *y*)
  - Five late days permitted

- Project: ~25%
- Midterm: ~15%
- Final exam: ~15%
- Class participation: ~5%

Exams

- Midterm: Tuesday, March 12, in class
- Final: Sunday May 5, 12:25-2:25 PM

- Let me know *immediately* if you have a conflict with either of these exam times

Computing Resources for the Class

- Linux servers in Dept. of Biostatistics & Medical Informatics
  - No “lab”, must log in remotely (use WiscVPN)
  - Will create accounts for everyone on course roster
  - Two machines

    mi1.biostat.wisc.edu

    mi2.biostat.wisc.edu
  - HW0 tests your access to these machines
  - Homework must be able to run on these machines

- CS department usually offers Unix orientation sessions at beginning of semester

Programming Assignments

- All programming assignments require Python
  - Project can be in any language

- Have a Python 3 environment on biostat servers
  - Permitted packages on course website
  - Can request others

- HW0 will be Python introduction

- Use Piazza for Python discussion
  - If you know Python, please help answer questions

Project

- Design and implement a new computational method for a task in molecular biology
- Improve an existing method
- Perform an evaluation of several existing methods
- Run on real biological data
- Suggestions will be provided
- Not simply your existing research
- Can email me now to discuss ideas

Participation

- Do the assigned readings before class
- Show up to class
- No one will have the perfect background
  - Ask questions about computational or biological concepts
- Correct me when I am wrong
  - Seriously, it will happen
- Piazza discussion board
  - Questions and answers

Piazza Discussion Board

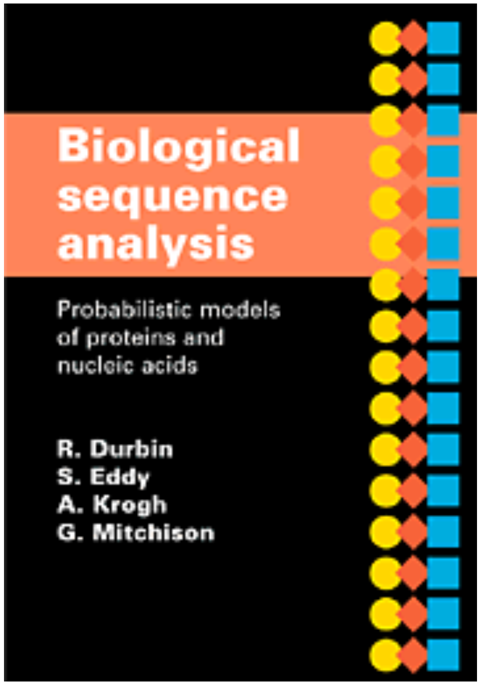
- Instead of a mailing list
- http://piazza.com/wisc/spring2019/bmics776/home
- Post your questions to Piazza instead of emailing the instructor or TA
  - Unless it is a private issue or project-related
- Answer your classmates' questions
- Announcements will also be posted to Piazza
- Supplementary material for lecture topics

Course Readings

- Mostly articles from the primary literature
- Must be using a campus IP address to download some of the articles (can use WiscVPN from off campus)

Recommended textbook

- *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Cambridge University Press, 1998.



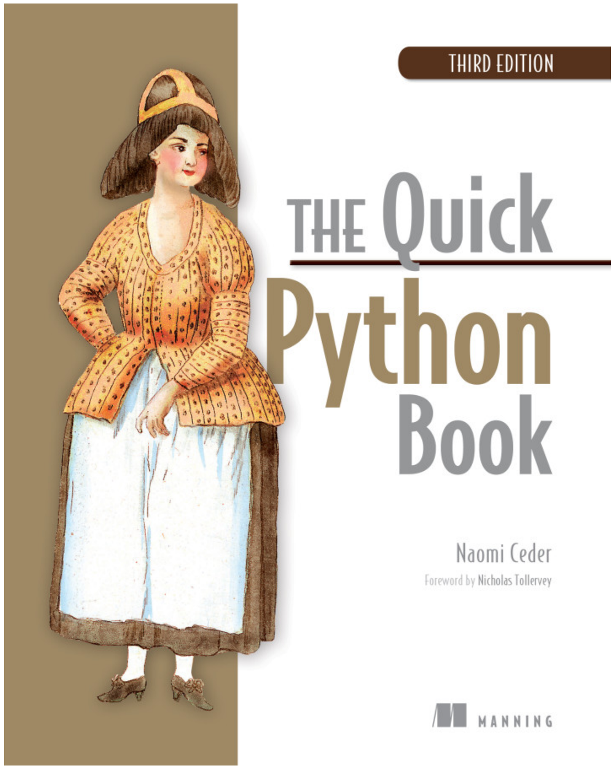
Biological sequence analysis

Probabilistic models of proteins and nucleic acids

R. Durbin
S. Eddy
A. Krogh
G. Mitchison

Python references

- https://docs.python.org
- If you want a book:
  - Python 3 for programmers
- Many other good books and online resources



https://www.manning.com/books/the-quick-python-book-third-edition

Prerequisites

- BMI/CS 576 or equivalent
- Knowledge of basic biology and methods from that course will be assumed
- May want to go over the material on the 576 website to refresh
- http://www.biostat.wisc.edu/bmi576/

What you should get out of this course

- An understanding of some of the major problems in computational molecular biology
- Familiarity with the algorithms and statistical techniques for addressing these problems
- How to think about different data types
- At the end you should be able to
    - Read the bioinformatics literature
    - Apply the methods you have learned to other problems both within and outside of bioinformatics
    - Write a short bioinformatics research paper

Major Topics to be Covered (the algorithms perspective)

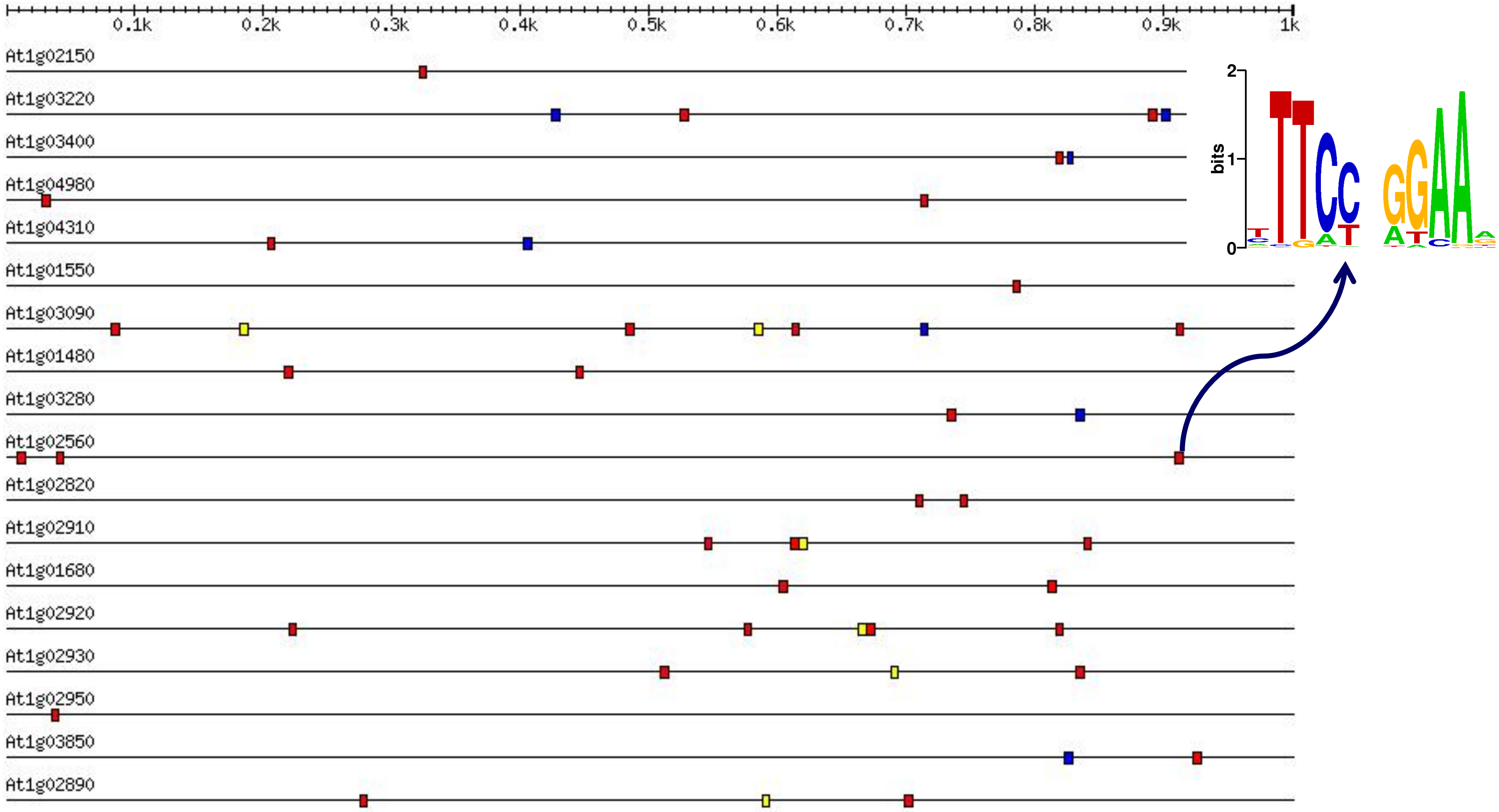
- Expectation Maximization
- Gibbs sampling
- Mutual information
- Network flow algorithms
- Stochastic context free grammars
- Multiple hypothesis testing correction
- Convolutional neural networks
- Linear programming
- Tries and suffix trees
- Markov random fields

Major Topics to be Covered (the task perspective)

- Modeling of motifs and *cis*-regulatory modules
- Identification of transcription factor binding sites
- Transcriptome quantification
- Transcriptome assembly
- RNA sequence and structure modeling
- Regulatory information in epigenomic data
- Genotype analysis and association studies
- Mass spectrometry peptide and protein identification
- Pathways in cellular networks
- Large-scale sequence alignment

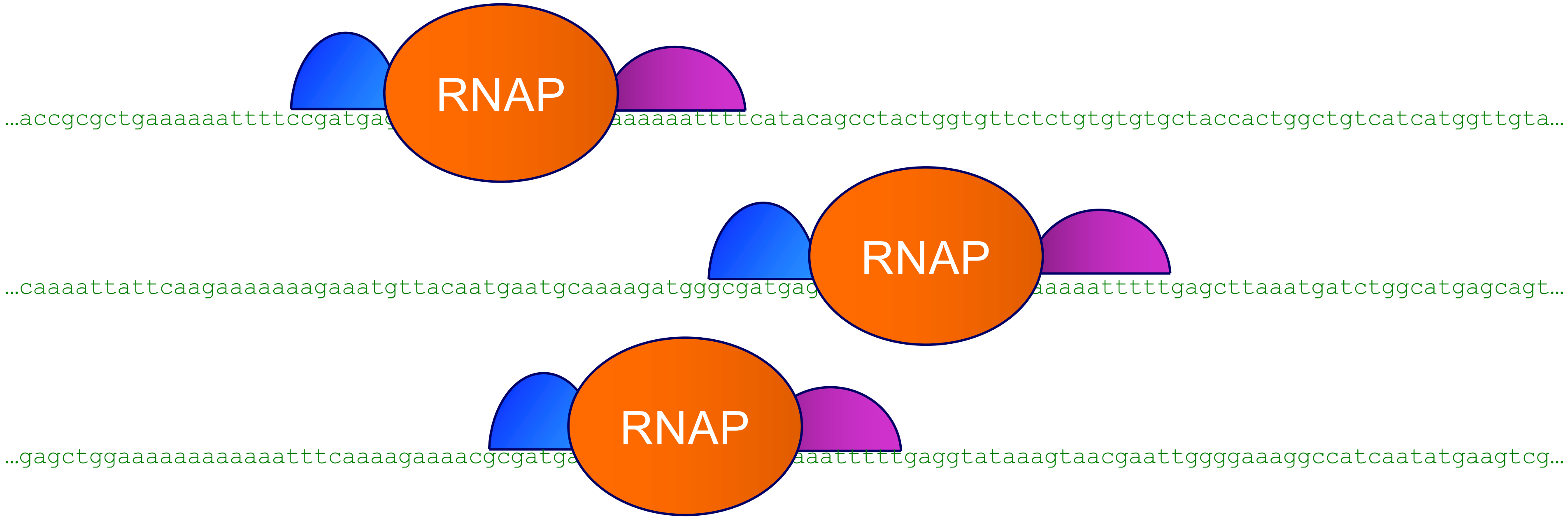
Motif Modeling

What sequence motif do these promoter regions have in common?



cis-Regulatory Modules

What configuration of sequence motifs do these promoter regions have in common?



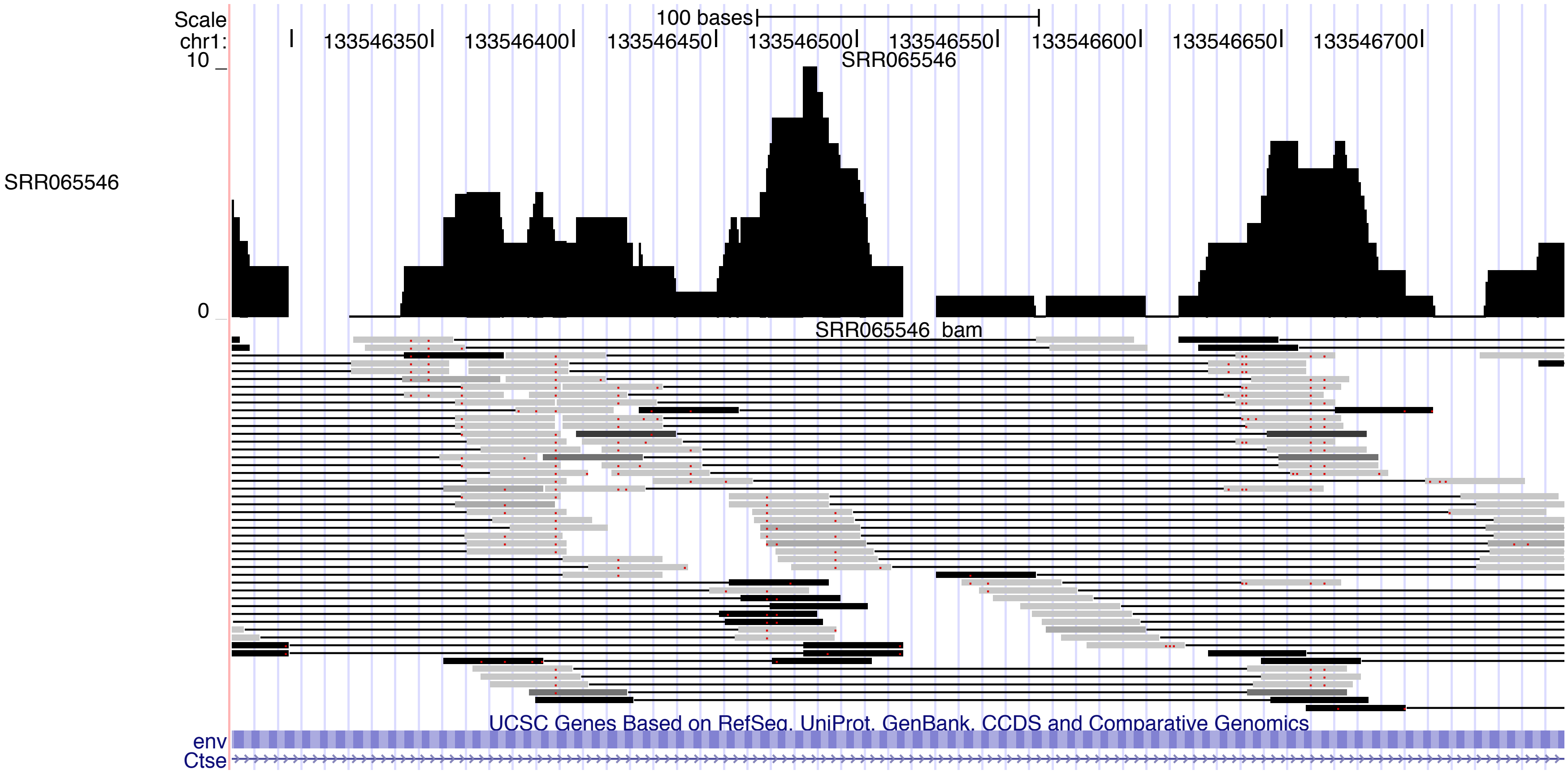
…accgcgctgaaaaaattttccgatgag aaaaaattttcatacagcctactggtgttctctgtgtgtgctaccactggctgtcatcatggttgta…

…caaaattattcaagaaaaaaagaaatgttacaatgaatgcaaaagatgggcgatgag aaaaattttttgagcttaaatgatctggcatgagcagt…

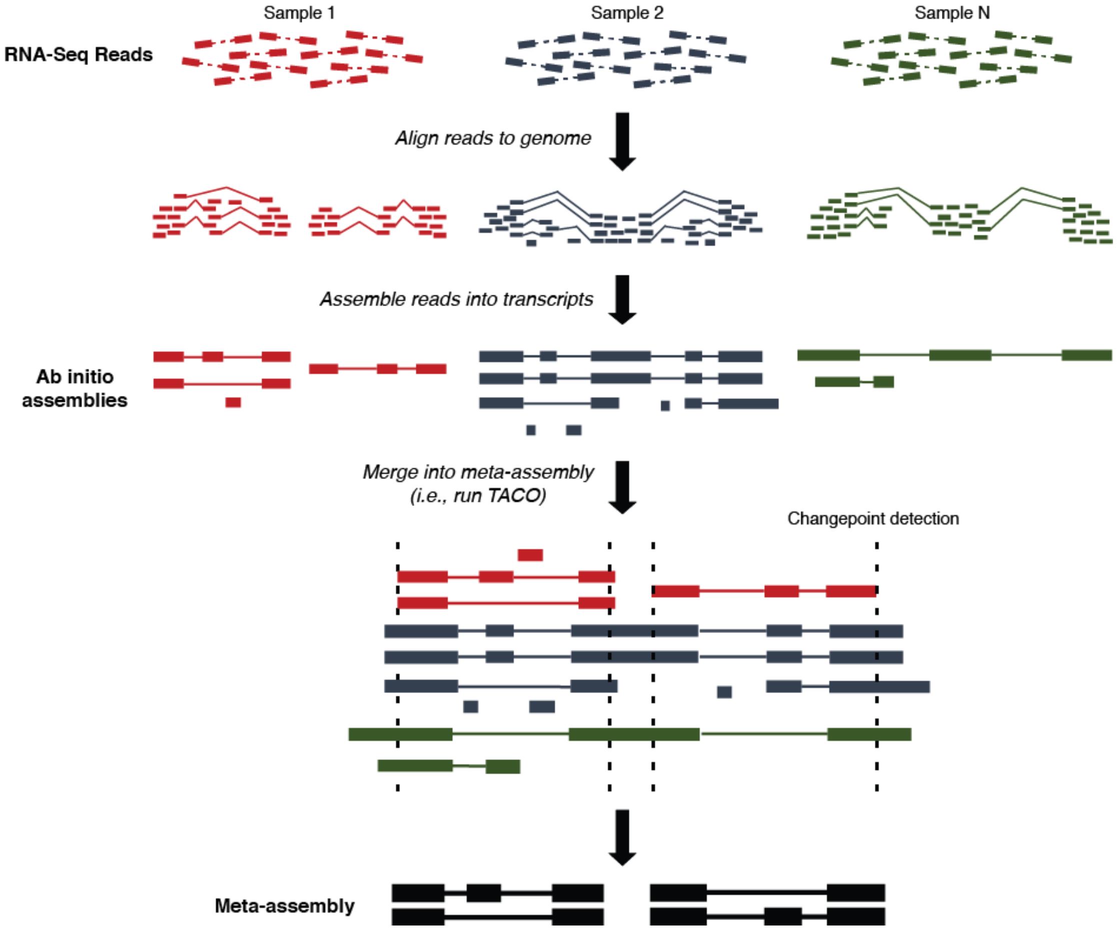
…gagctggaaaaaaaaaaaatttcaaaagaaaacgcgatga aaatttttgaggtataaagtaacgaattggggaaaggccatcaatatgaagtcg…

Transcriptome Analysis with RNA-Seq

What genes are expressed and at what levels?



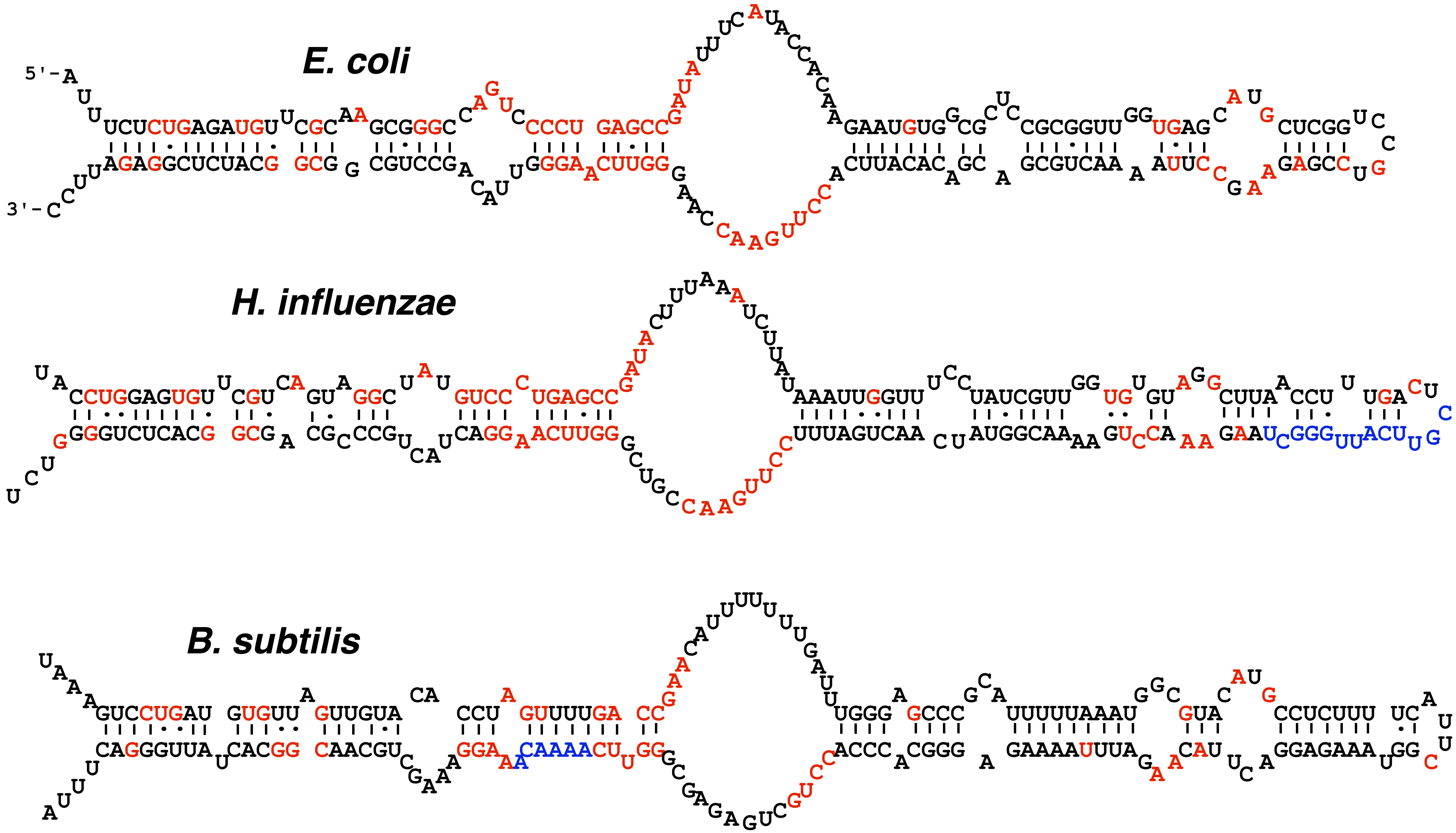
Transcriptome assembly



https://tacorna.github.io/

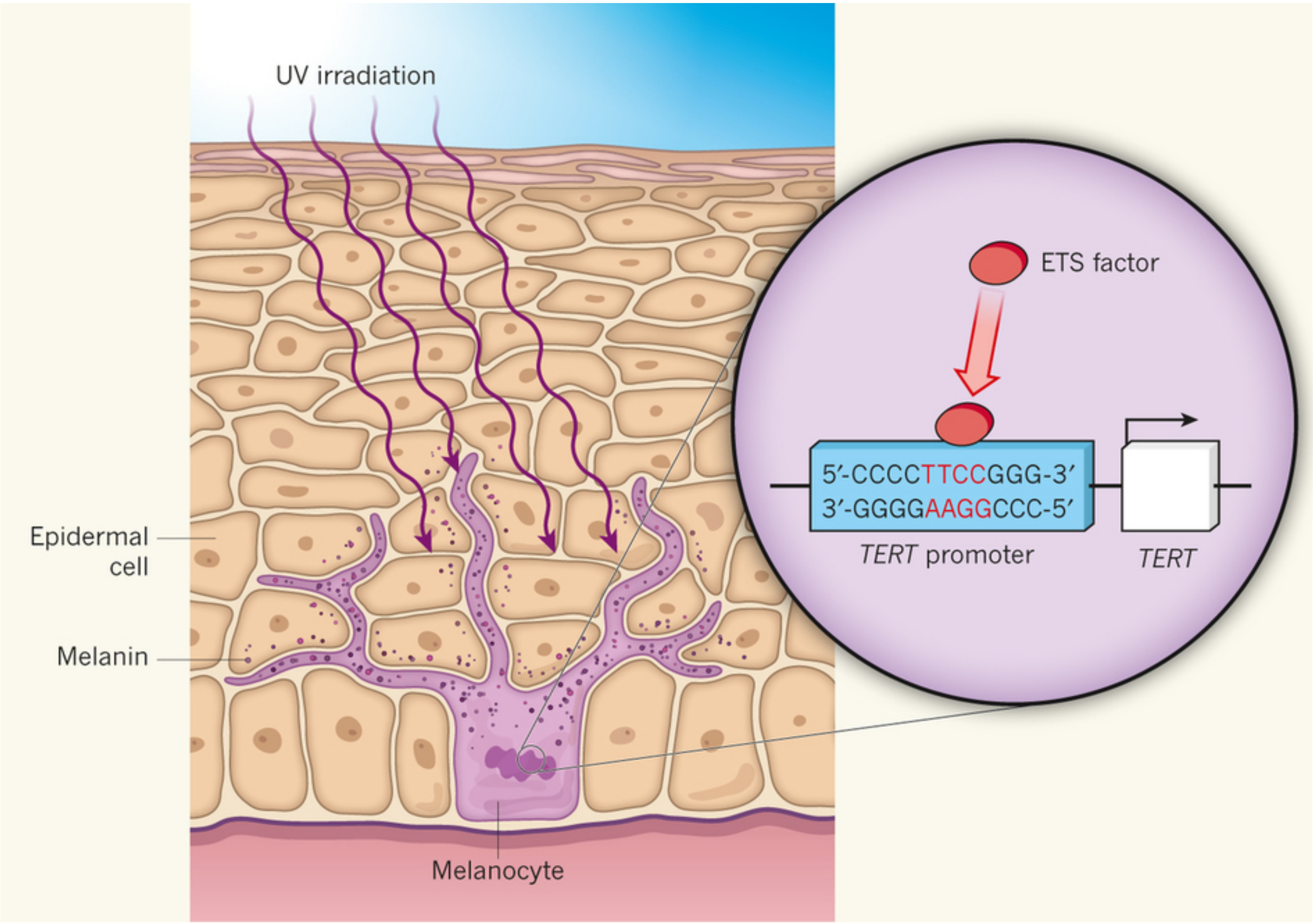
RNA Sequence and Structure Modeling

How can we identify sequences that encode this RNA structure?



Noncoding Genetic Variants

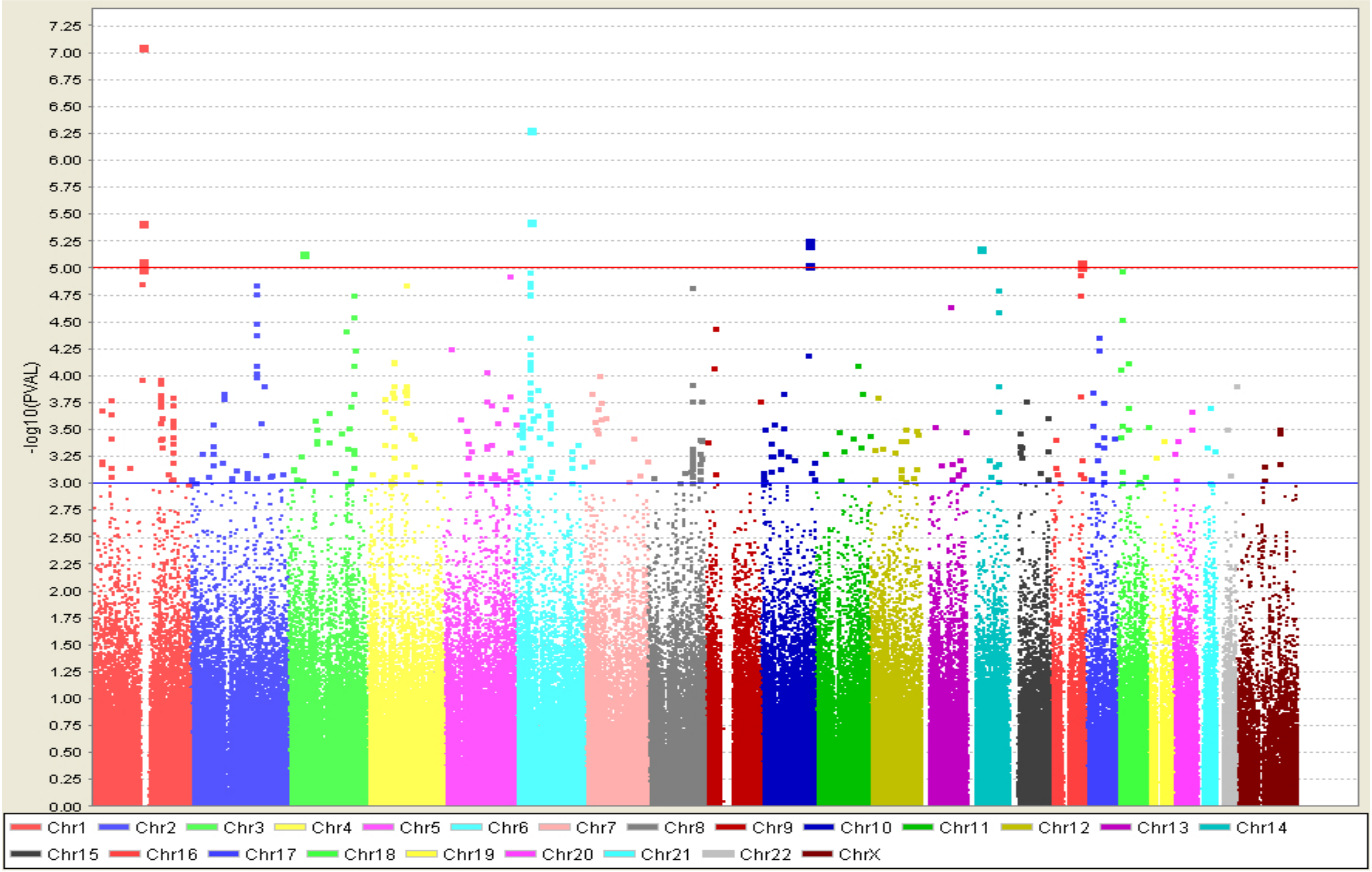
How do genetic variants outside protein coding regions impact phenotypes?



Patton and Harrington, *Nature*, 2013

Genome-wide Association Studies

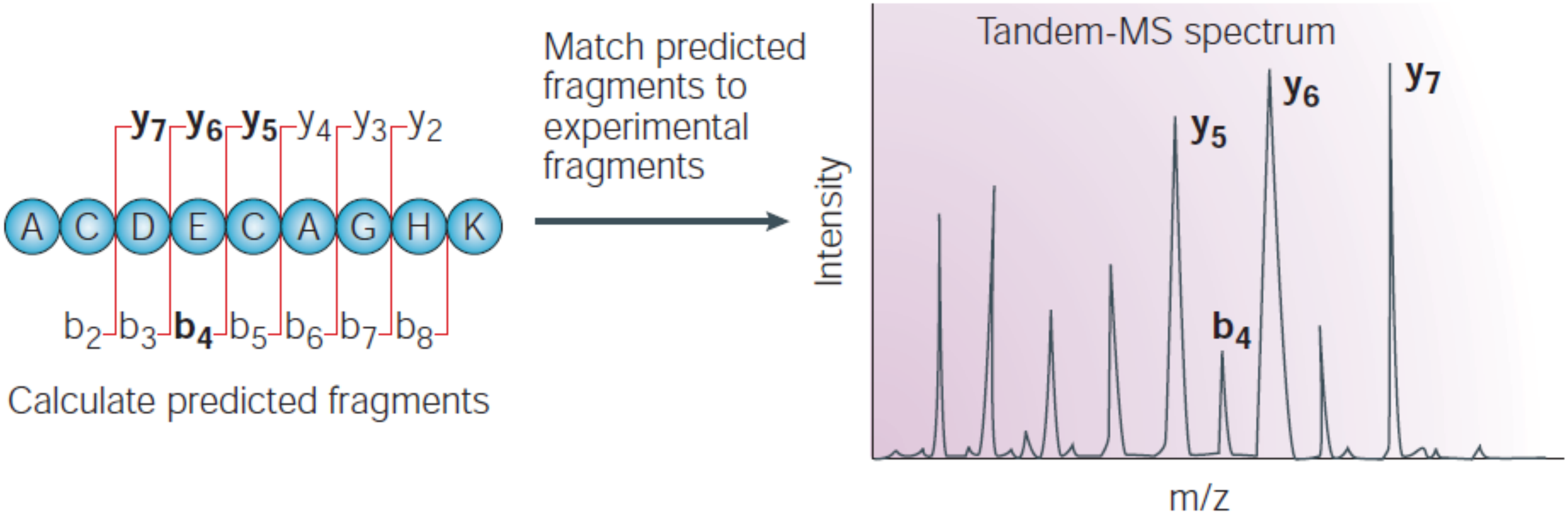
Which genes are involved in diabetes?



Type 2 diabetes association P values by chromosome (386,731 markers). The x-axis is the genomic position by chromosome 1-22 and X (by color), and the y-axis is the negative base 10 logarithm of the P value.

Proteomic Analysis with Mass Spectrometry

What proteins are expressed and at what levels?



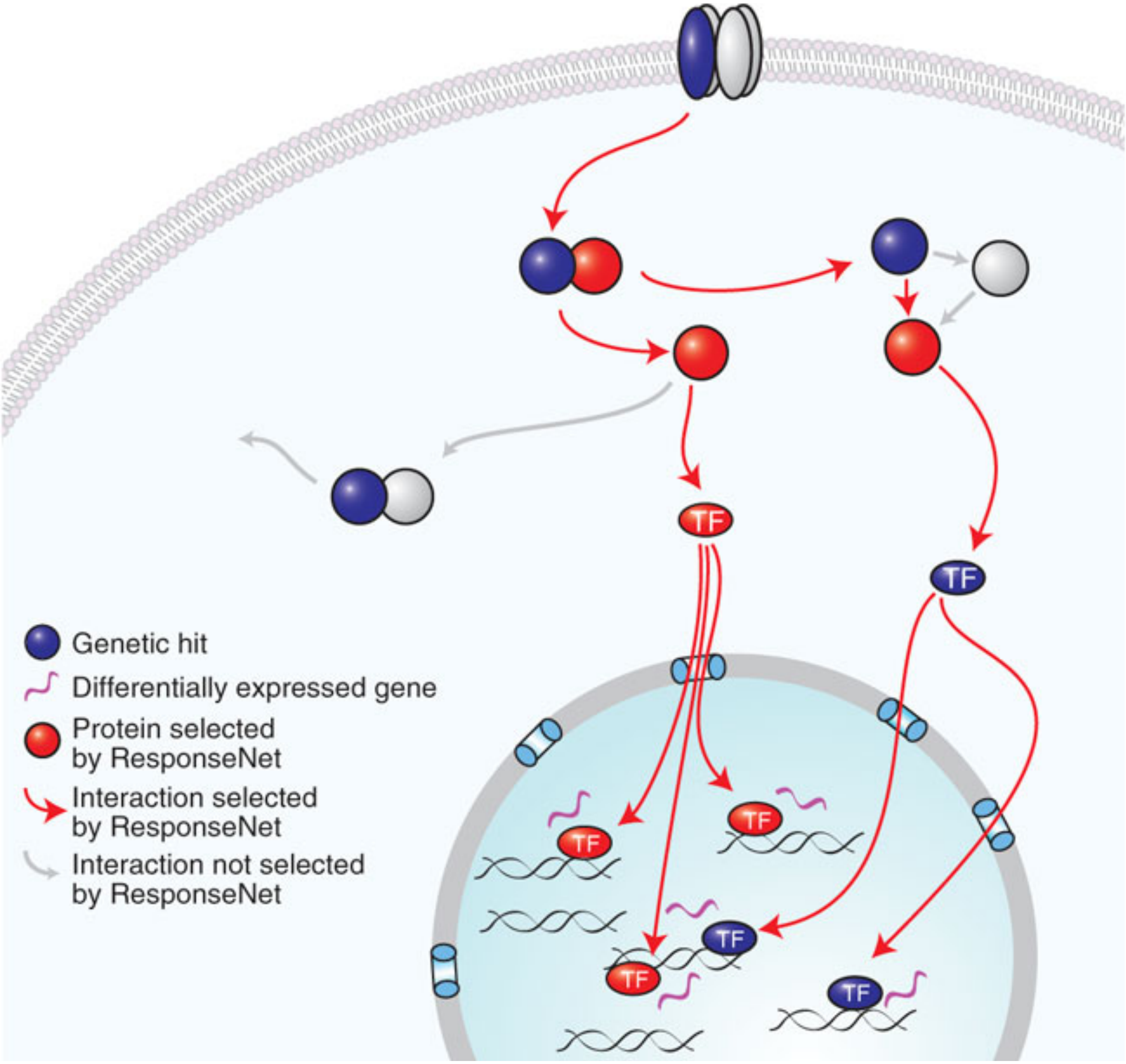
Calculate predicted fragments

Match predicted fragments to experimental fragments

Steen and Mann, *Nature Reviews Molecular Cell Biology*, 2004

Identifying Signaling Pathways

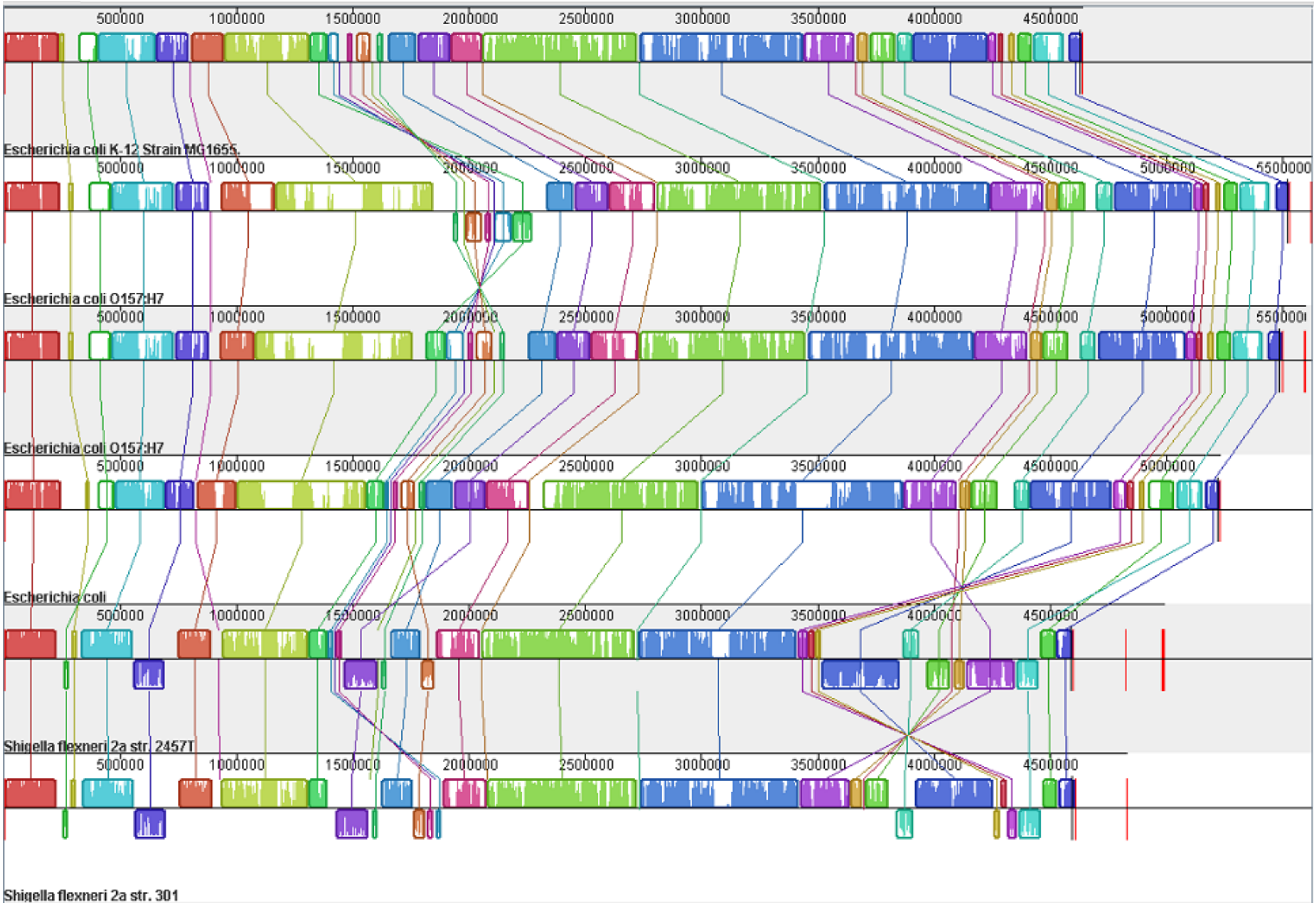
How do proteins coordinate to transmit information?


TF
TF
TF
TF
TF
TF
TF
Genetic hit
Differentially expressed gene
Protein selected by ResponseNet
Interaction selected by ResponseNet
Interaction not selected by ResponseNet

Yeger-Lotem et al., *Nature Genetics,* 2009

Large Scale Sequence Alignment

What is the best alignment of these 6 genomes?



Other Topics

- Many topics we aren’t covering
  - Protein structure prediction
  - Protein function annotation
  - Metagenomics
  - Metabolomics
  - Graph genomes
  - Single-cell sequencing
  - Text mining
  - Others?

Reading Groups

- Computational Systems Biology Reading Group
  - http://lists.discovery.wisc.edu/mailman/listinfo/compsysbiojc
- AI Reading Group
  - http://lists.cs.wisc.edu/mailman/listinfo/airg
- ComBEE Python Study Group
  - https://combee-uw-madison.github.io/studyGroup/

- Many relevant seminars on campus