Advanced Bioinformatics

Biostatistics & Medical Informatics 776
Computer Sciences 776
Spring 2022

Daifeng Wang

daifeng.wang@wisc.edu

www.biostat.wisc.edu/bmi776/

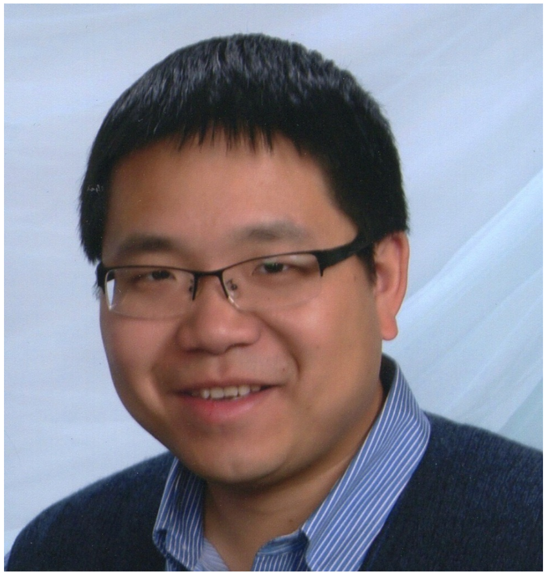
Agenda Today

- 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: Daifeng Wang



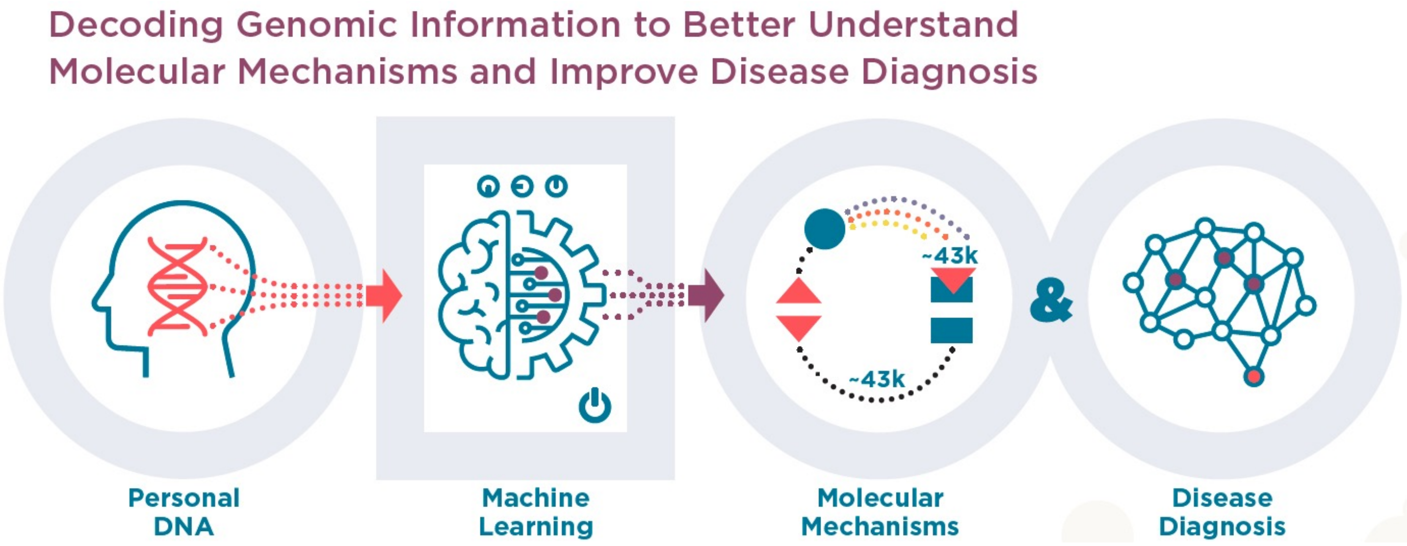
- Email: daifeng.wang@wisc.edu
- Website: https://daifengwanglab.org/
- Office: Waisman Center 517
- Class Location: 2534 Engineering Hall
- Class Times: Tue/Thu 1:00 – 2:15 pm CST
- Office for Office Hours: Medical Sciences Center 6740
- Office Hours: Tues 2:30-3:30pm, Thus 2:30-3:30pm CST

- Assistant Professor in the Department of Biostatistics & Medical Informatics and Investigator in Waisman Center

- Research interests: interpretable machine learning, network biology, functional genomics, comparative genomics, brain diseases, precision medicine

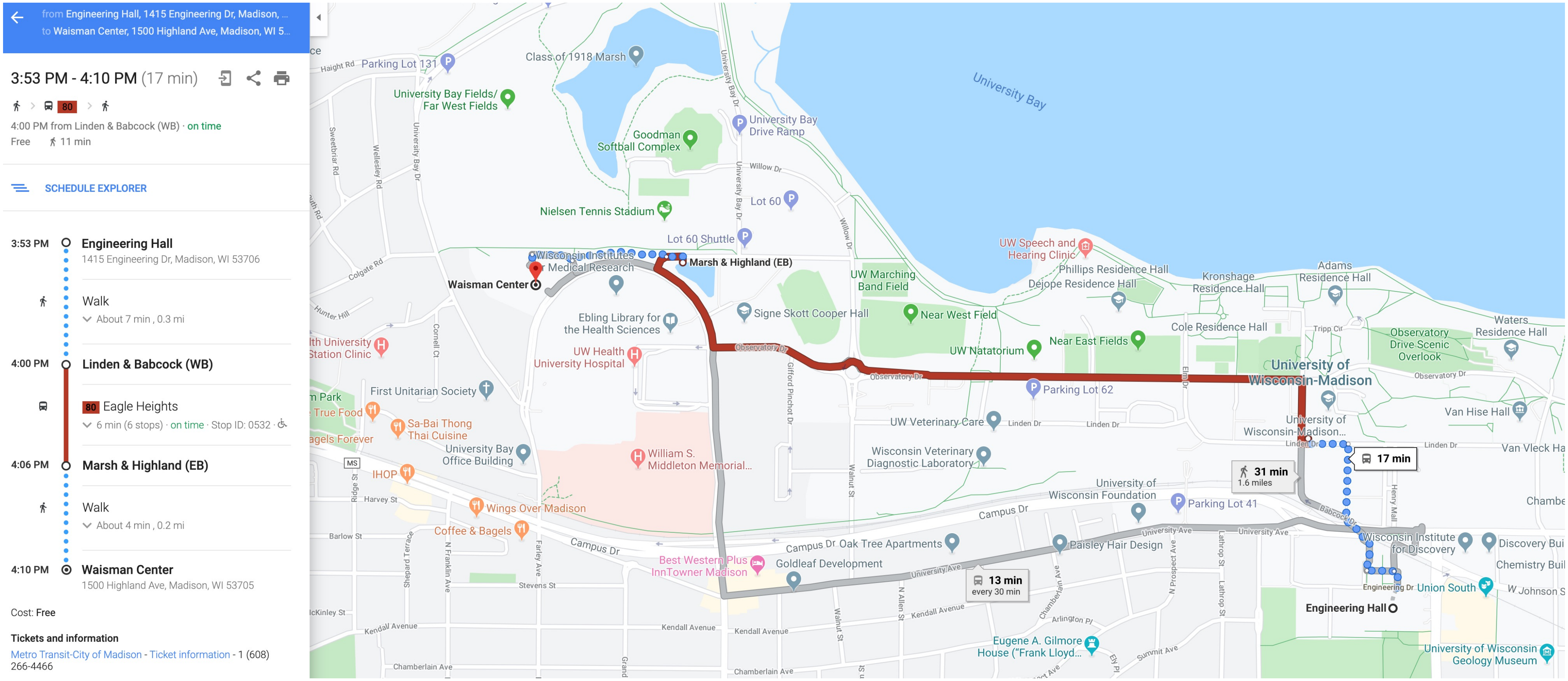
My research in Waisman Center

- Mission of Waisman Center
  - *Advance knowledge about human development, developmental disabilities, and neurodegenerative diseases*
- Goal of my research


Decoding Genomic Information to Better Understand Molecular Mechanisms and Improve Disease Diagnosis
~43k
~43k
&
Personal DNA
Machine Learning
Molecular Mechanisms
Disease Diagnosis

https://www.waisman.wisc.edu/2020/01/07/new-researcher-uses-machine-learning-to-decode-genomic-information/

Finding My Office:
517 Waisman Center



from Engineering Hall, 1415 Engineering Dr, Madison, ...
to Waisman Center, 1500 Highland Ave, Madison, WI 5...

3:53 PM - 4:10 PM (17 min)

80

4:00 PM from Linden & Babcock (WB) · on time

Free 11 min

SCHEDULE EXPLORER

3:53 PM **Engineering Hall**
1415 Engineering Dr, Madison, WI 53706

Walk
About 7 min , 0.3 mi

4:00 PM **Linden & Babcock (WB)**

80 Eagle Heights
6 min (6 stops) · on time · Stop ID: 0532

4:06 PM **Marsh & Highland (EB)**

Walk
About 4 min , 0.2 mi

4:10 PM **Waisman Center**
1500 Highland Ave, Madison, WI 53705

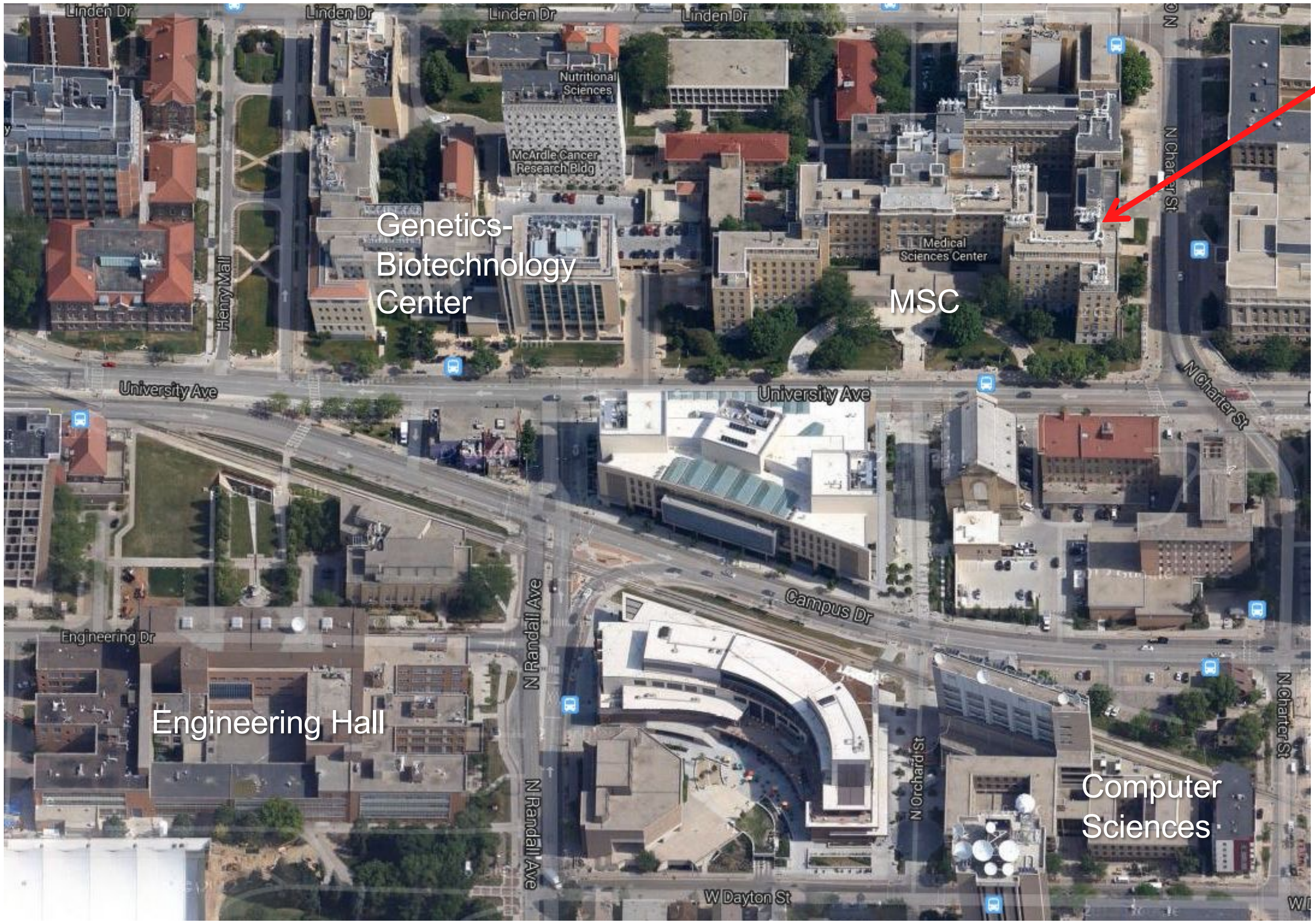
Cost: Free

Tickets and information

Metro Transit-City of Madison - Ticket information - 1 (608) 266-4466

- Far away, most west building
- Take No. 80 Bus or Bike/Walk for exercise

Finding the Office for in-person Office Hours: MSC 6740

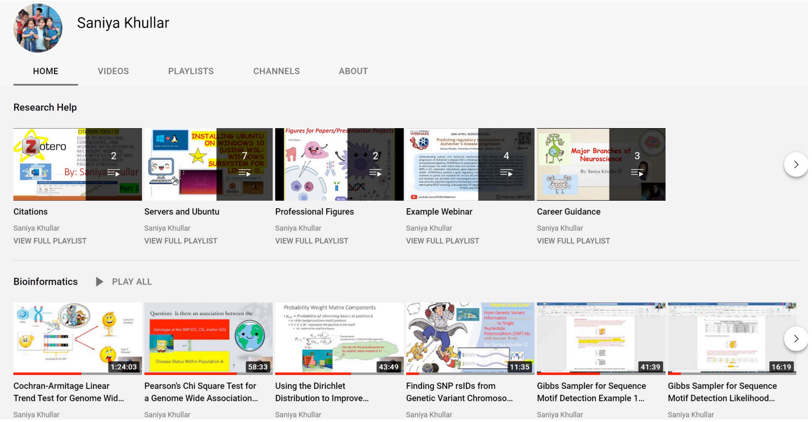


Office for Office Hours

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

Course TA:
Saniya Khullar

- Email: skhullar2@wisc.edu
- Skype: saniya0605
- Office Hours (1 virtual, 1 in-person Office Hour):
  - **Virtual:** Fridays 9 am to 10 am CST (https://zoom.us/j/2593679726)
  - **In-person:** Mondays 9 a.m. to 10 a.m. CST in Waisman 520
  - Available by appointment as well
- Ph.D. candidate
  - Biomedical Data Science
- Educational YouTube channel (with content related to Advanced Bioinformatics and Beyond): https://www.youtube.com/c/SaniyaKhullar



Saniya Khullar

Research Help

Citations	Servers and Ubuntu	Professional Figures	Example Webinar	Career Guidance
Saniya Khullar	Saniya Khullar	Saniya Khullar	Saniya Khullar	Saniya Khullar
2	7	2	4	3

Bioinformatics

Cochran-Armitage Linear Trend Test for Genome Wid…	Pearson's Chi Square Test for a Genome Wide Association…	Using the Dirichlet Distribution to Improve…	Finding SNP rsIDs from Genetic Variant Chromoso…	Gibbs Sampler for Sequence Motif Detection Example 1…	Gibbs Sampler for Sequence Motif Detection Likelihood…
Saniya Khullar	Saniya Khullar	Saniya Khullar	Saniya Khullar	Saniya Khullar	Saniya Khullar
1:24:03	58:33	43:49	11:35	41:39	16:19

Course TA:
Ting Jin

- Email: tjin27@wisc.edu
- Ph.D. student
  - Biomedical Data Science
- Office Hours (Virtual)
  - Friday: 10 – 11am (Zoom)
- Grading student assignments and providing feedback

Our Course TAs are here to help support you!
Please do reach out with any questions!

Office Hours

- Tue 2:30-3:30pm, Thu 2:30-3:30pm
  - MSC 6740
- Will begin next week
- Free to schedule an individual meeting
  - Waisman Center or Zoom
- You are encouraged to visit our office hours!

You

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

Course Requirements

- 4 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: Thursday, March 10, in class
- Final: Monday May 8, 10:05 AM – 12:05 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

- Resources:
  - TA Saniya prepared a video on working on a remote server (pushing, pulling, running files remotely to Biostat servers, WiscVPN): Video Link along with other helpful videos on Servers
  - 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

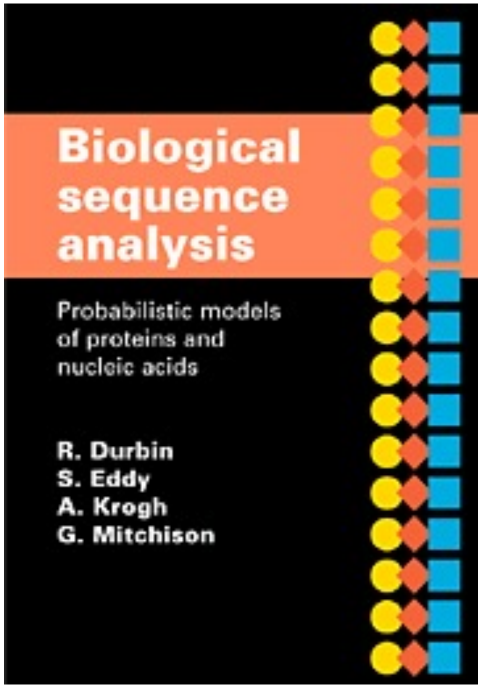
- Instead of a mailing list
- https://piazza.com/wisc/spring2022/bmics776
- Post your questions to Piazza instead of emailing the instructor or TA
  - Unless it is a private issue or project-related
- Answer your classmates' questions
- Announcements will also be posted to Piazza
- Supplementary material 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

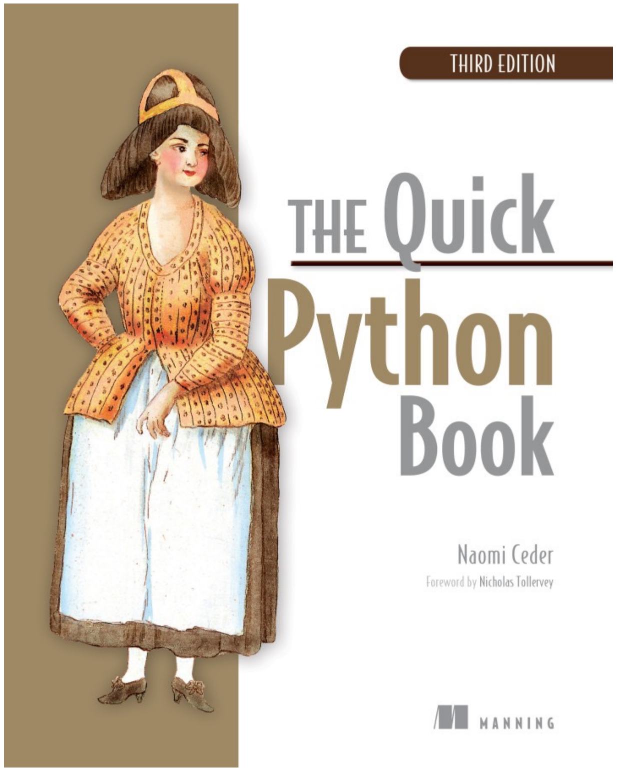
Recommended online reading

- Translational Bioinformatics
  - https://collections.plos.org/translational-bioinformatics

PLOS
COMPUTATIONAL BIOLOGY A Peer-Reviewed, Open Access Journal
OPEN ACCESS
Translational Bioinformatics
ploscollections.org/translationalbioinformatics

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 biology and bioinformatics
- Familiarity with the techniques for addressing these problems
  - Computational, statistical, machine learning
- 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)

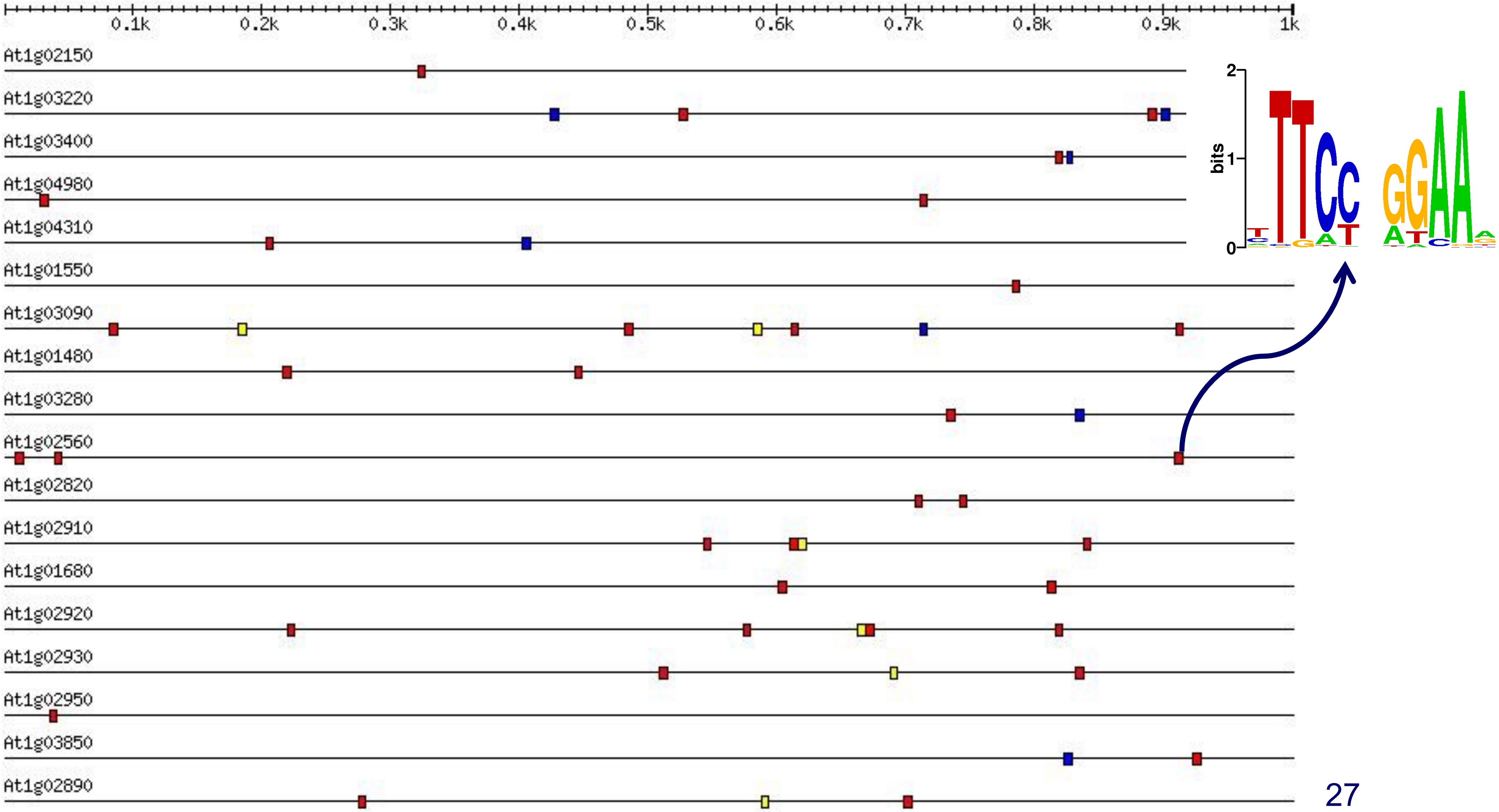
- Expectation Maximization
- Gibbs sampling
- Mutual information
- Network flow algorithms
- Multiple hypothesis testing correction
- Deep learning (e.g., Convolutional neural networks)
- Linear programming
- Clustering
- More machine learning approaches (e.g., manifold alignment)

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
- Regulatory information in epigenomic data
- Genotype analysis and association studies
- Quantitative Trait Locus (QTL) Analysis
- Pathways in cellular networks
- Single cell RNA-seq
- Gene regulatory network

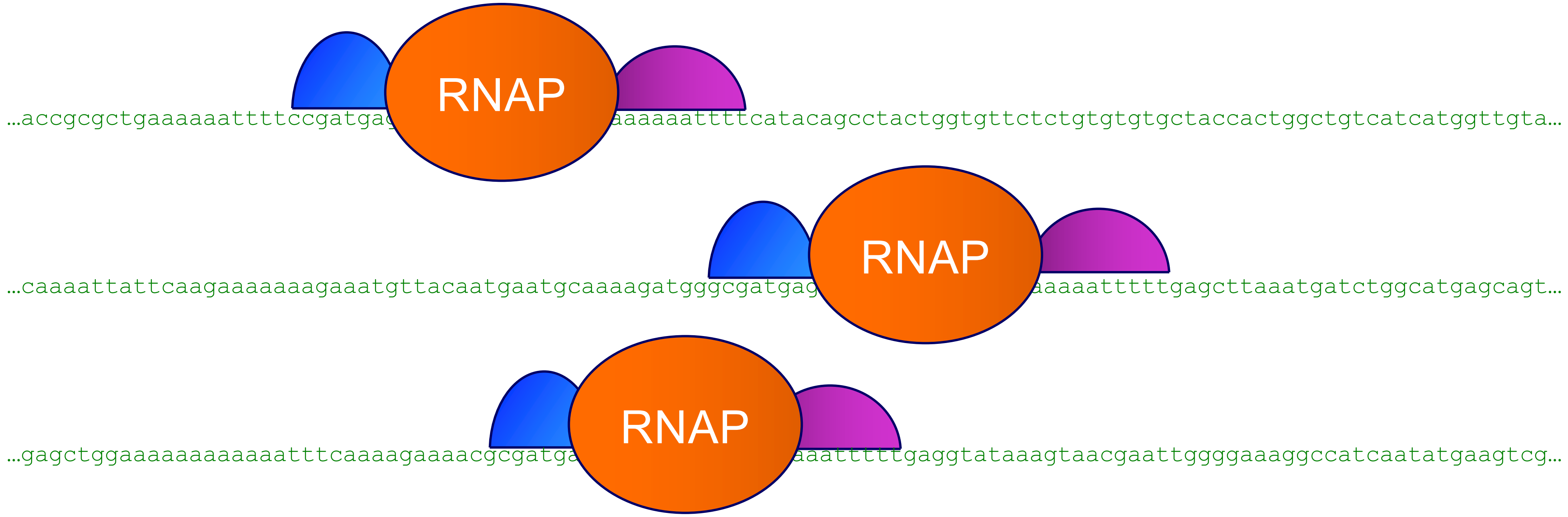
Motif Modeling

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?



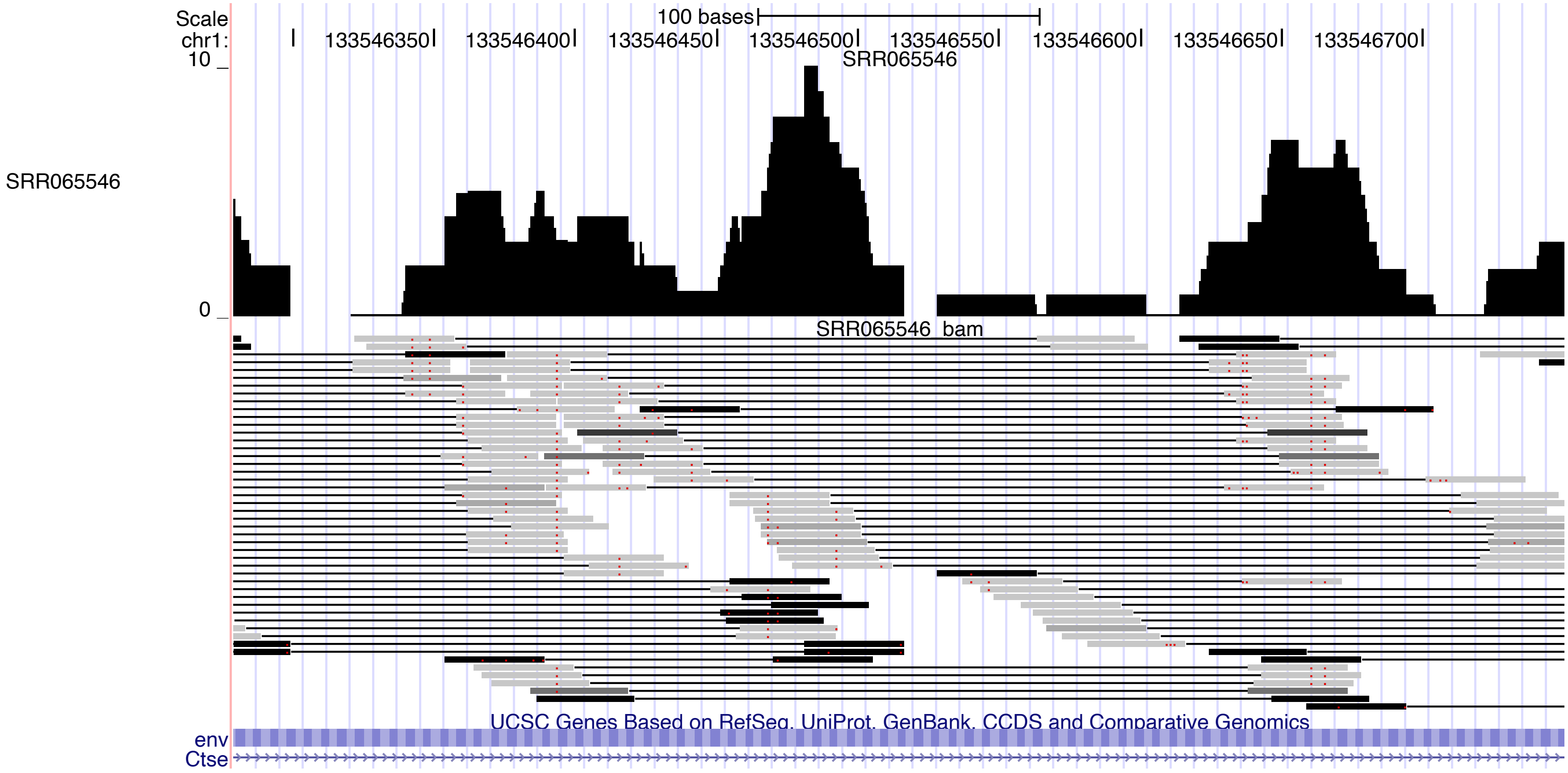
…accgcgctgaaaaaattttccgatgag aaaaaattttcatacagcctactggtgttctctgtgtgtgctaccactggctgtcatcatggttgta…

…caaaattattcaagaaaaaaagaaatgttacaatgaatgcaaaagatgggcgatgag aaaaatttttgagcttaaatgatctggcatgagcagt…

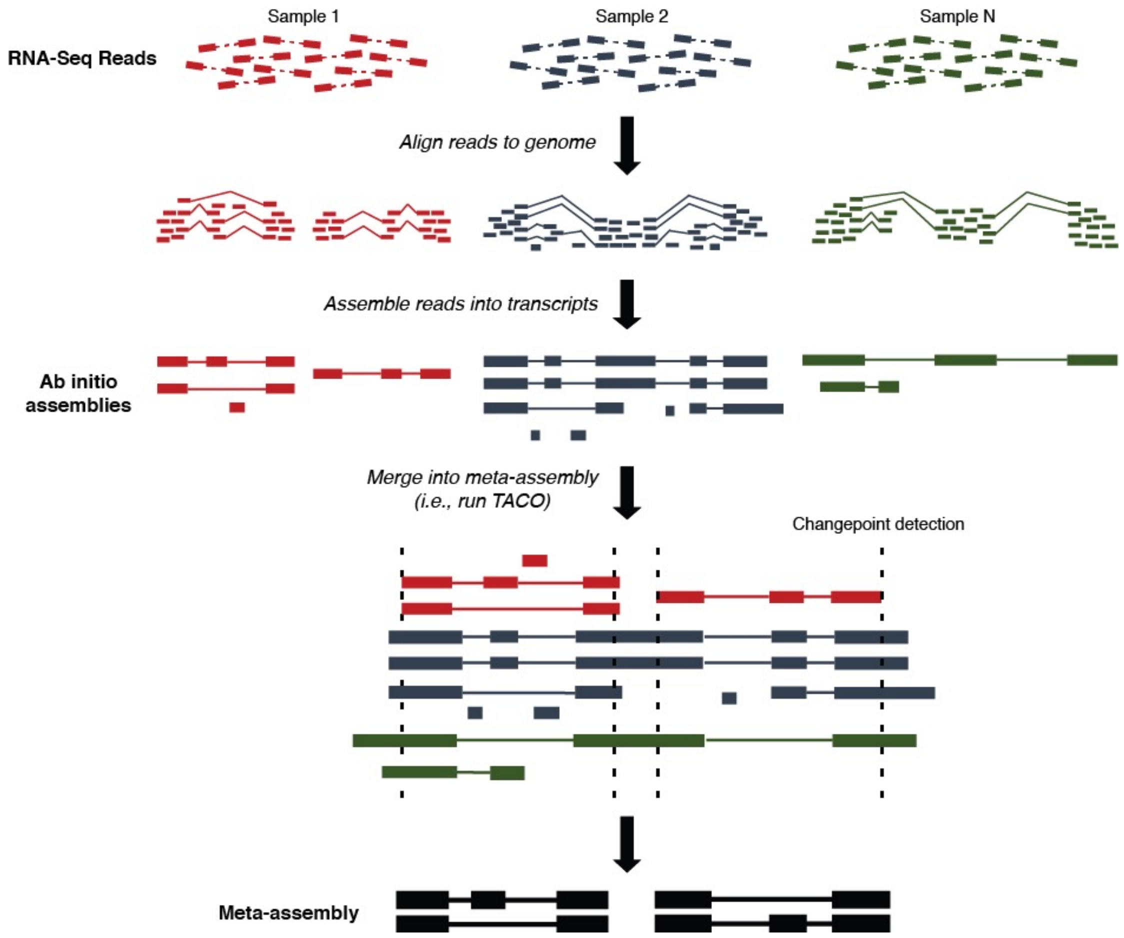
…gagctggaaaaaaaaaaaaatttcaaaagaaaacgcgatga aaatttttgaggtataaagtaacgaattggggaaaggccatcaatatgaagtcg…

Transcriptome Analysis with RNA-Seq

What genes are expressed and at what levels?



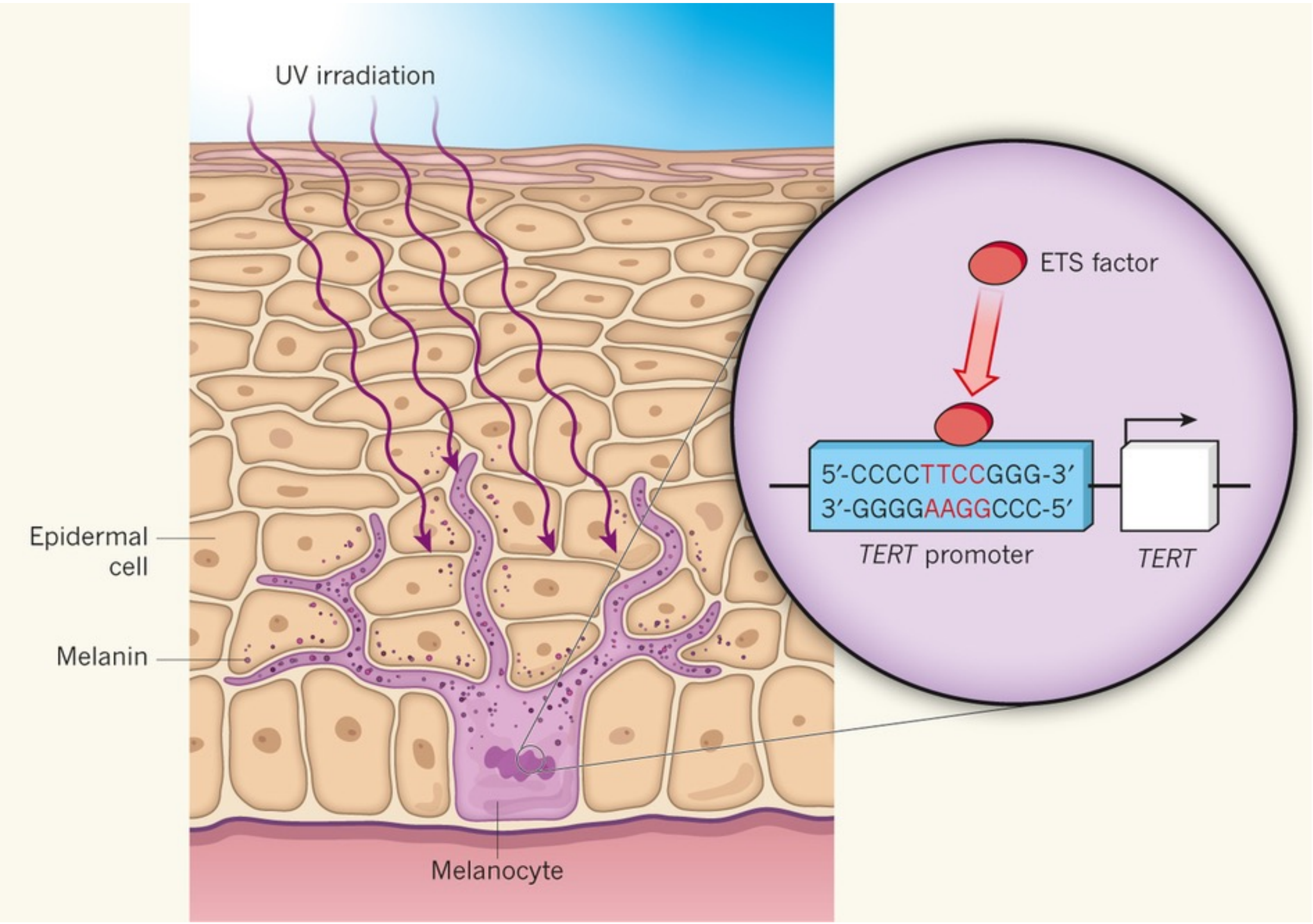
Transcriptome assembly



https://tacorna.github.io/

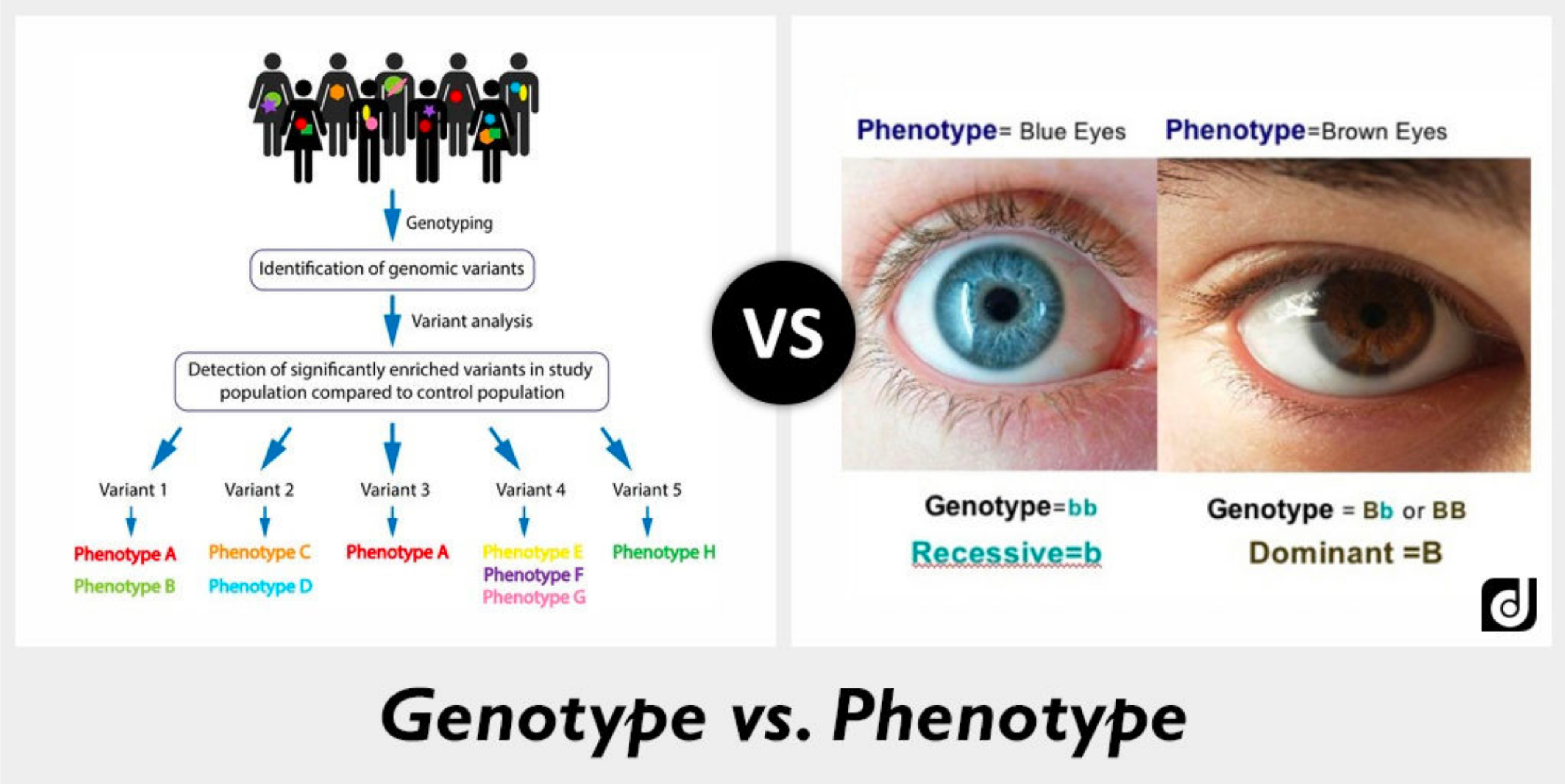
Noncoding Genetic Variants

How do genetic variants outside protein coding regions impact phenotypes?



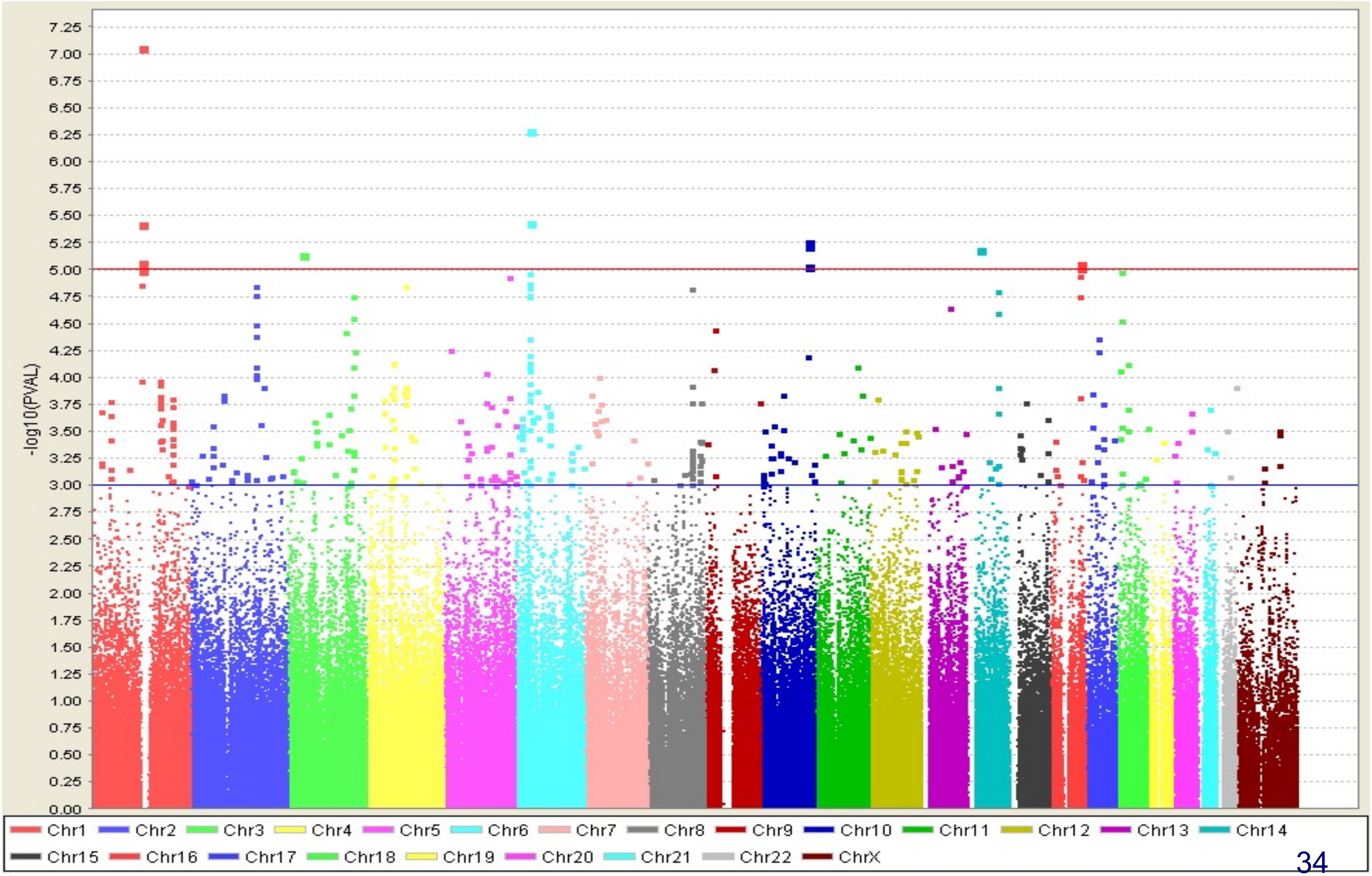
Patton and Harrington, *Nature*, 2013

Genotype to Phenotype



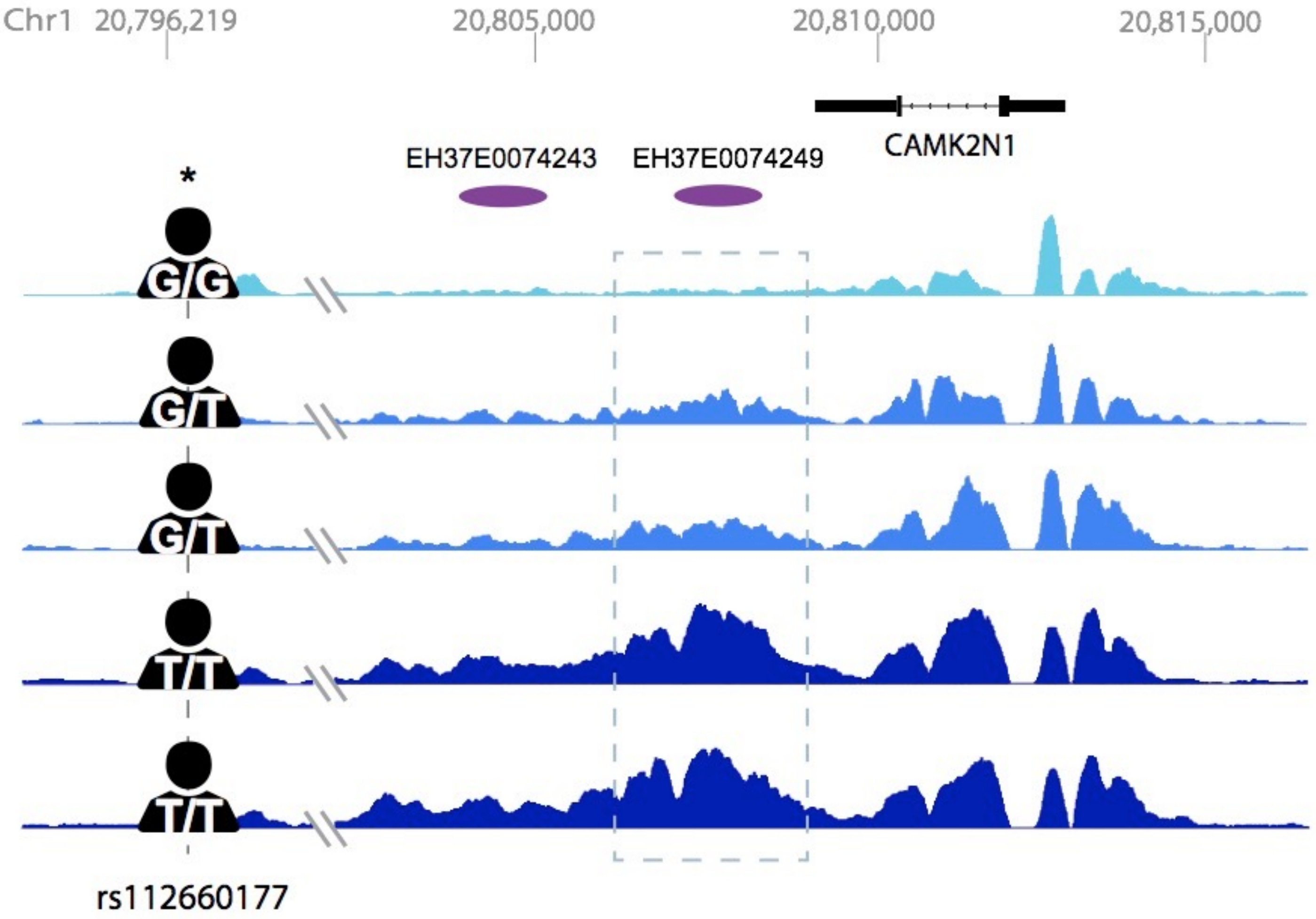
Genome-wide Association Studies

Which genes are involved in diabetes?



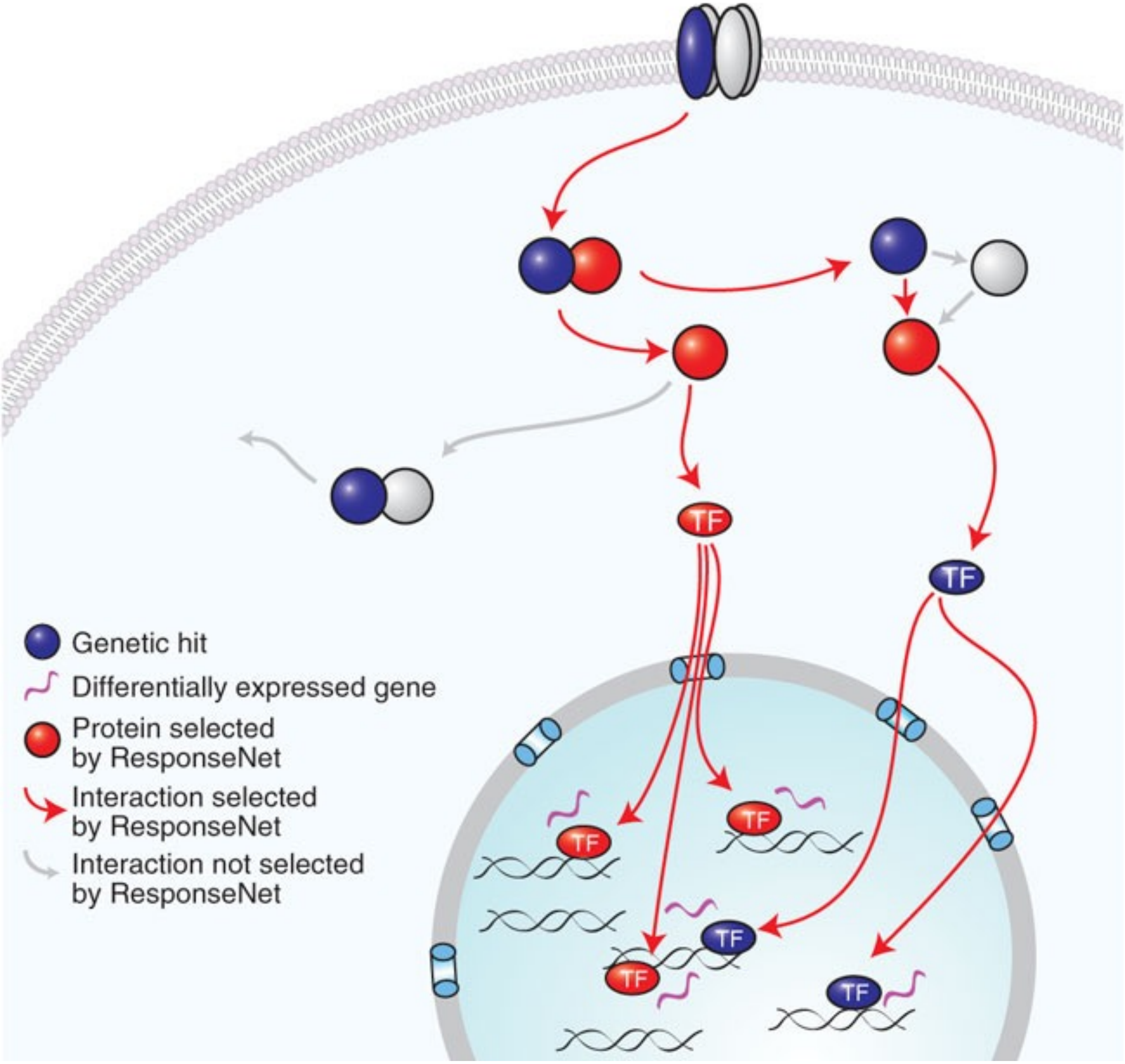
Type 2 diabetes association P values by chromosome (386,731 markers). The x-axis is the genomic position by chromosome 1-22 and X (by color), and the y-axis is the negative base 10 logarithm of the P value.

Quantitative Trait Locus (QTL) analysis



Identifying Signaling Pathways

How do proteins coordinate to transmit information?

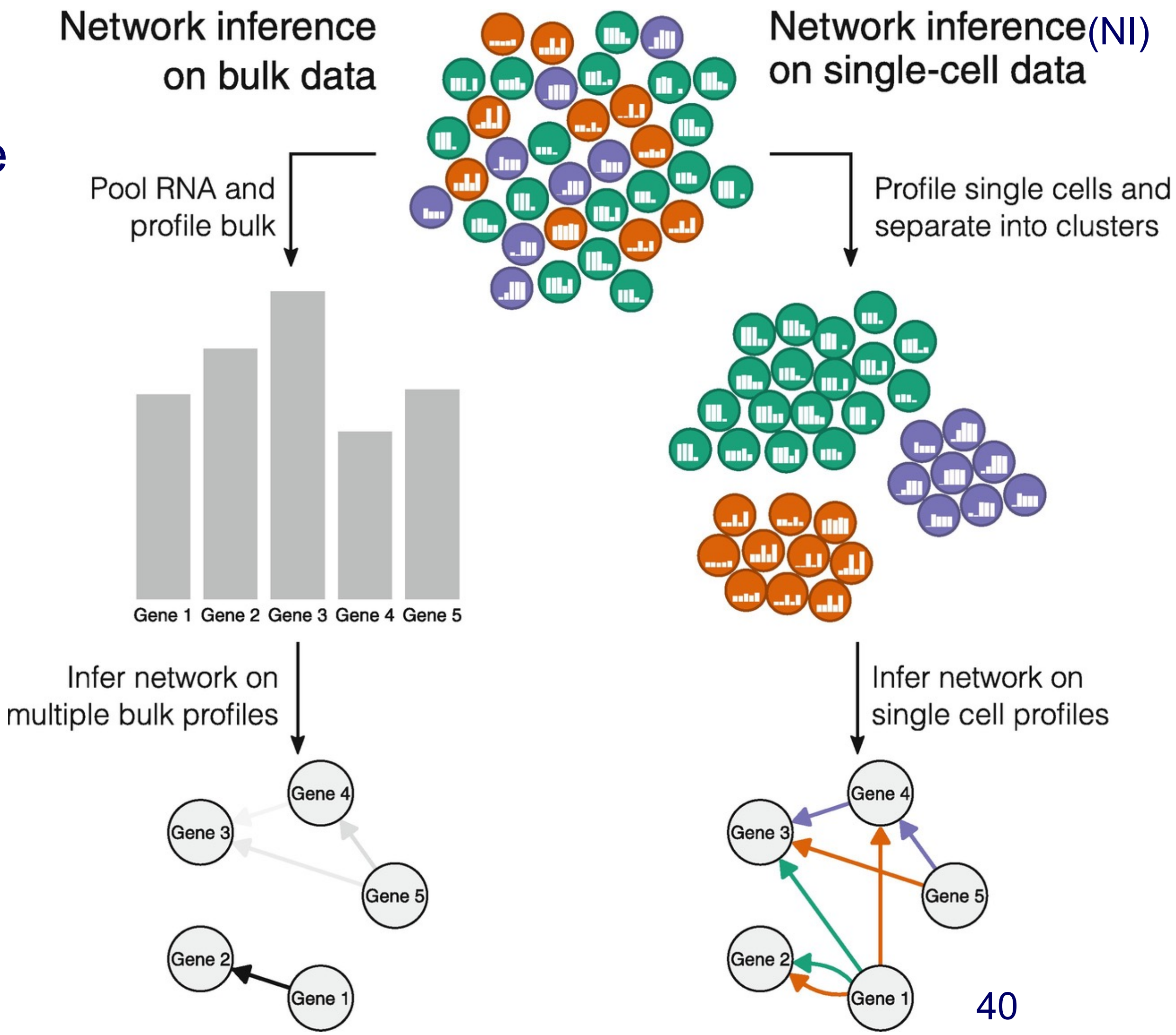

TF
TF
TF
TF
TF
TF
TF
Genetic hit
Differentially expressed gene
Protein selected by ResponseNet
Interaction selected by ResponseNet
Interaction not selected by ResponseNet

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

Cell-type gene regulatory networks

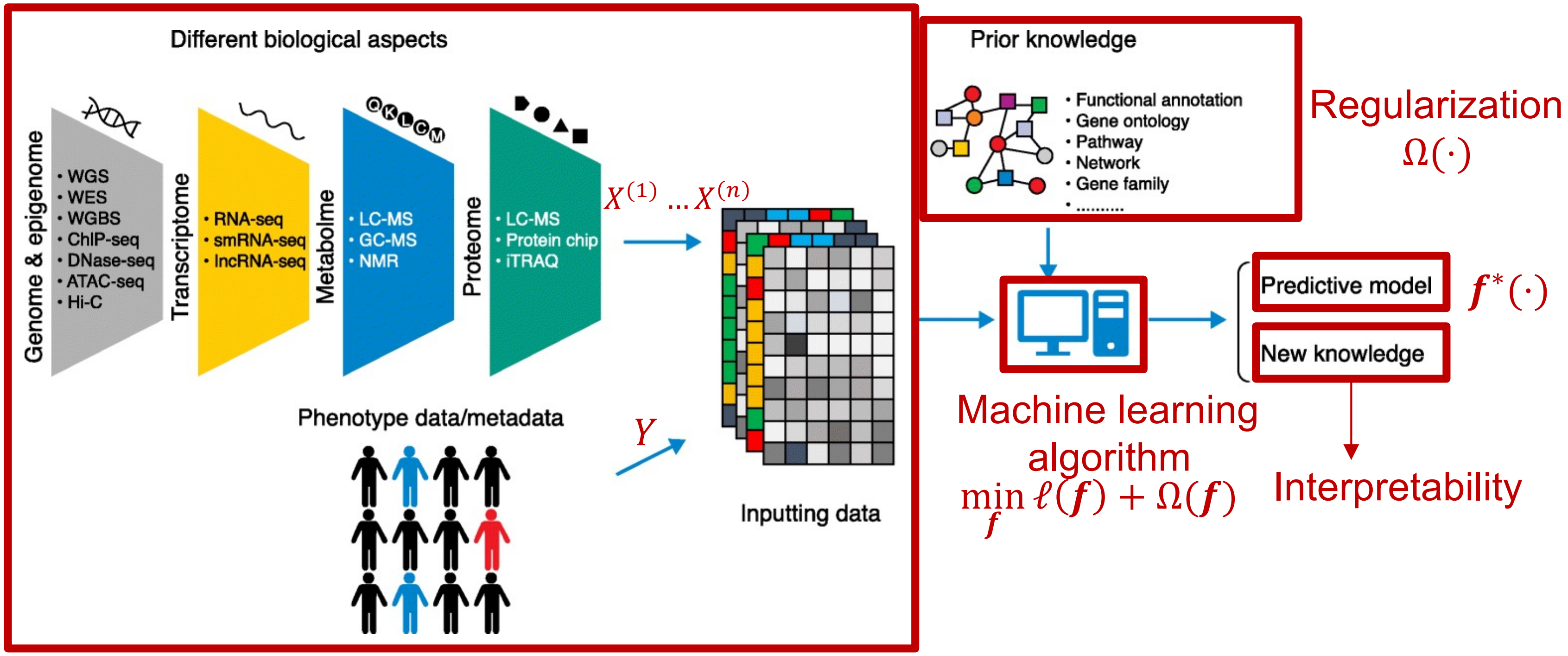
- Cell-type-specific GRNs would be key tools for the study of cellular heterogeneity

- Cell-type-specific GRNs will reveal key regulatory factors and circuits for specific cell types, facilitating mapping between disease-associated variants and affected cell types


Network inference on bulk data
Network inference(NI) on single-cell data
Pool RNA and profile bulk
Profile single cells and separate into clusters
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Infer network on multiple bulk profiles
Infer network on single cell profiles
Gene 4
Gene 3
Gene 5
Gene 2
Gene 1
Gene 4
Gene 3
Gene 5
Gene 2
Gene 1

Todorov H., Cannoodt R., Saelens W., Saeys Y. (2019) Network Inference from Single-Cell Transcriptomic Data. In: Sanguinetti G., Huynh-Thu V. (eds) Gene Regulatory Networks. Methods in Molecular Biology, vol 1883. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-8882-2_10

Machine Learning for Multiomics



$X^{(1)} \dots X^{(n)}$

Y

Regularization $\Omega(\cdot)$

Machine learning algorithm $\min_{\boldsymbol{f}} \ell(\boldsymbol{f}) + \Omega(\boldsymbol{f})$

$\boldsymbol{f}^*(\cdot)$

Interpretability

Xu. Genome Bio. 2019

Other Topics

- Many topics we aren’t covering
  - Protein structure prediction
  - Protein function annotation
  - Metagenomics
  - Metabolomics
  - Graph genomes
  - Mass Spectrometry
  - 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