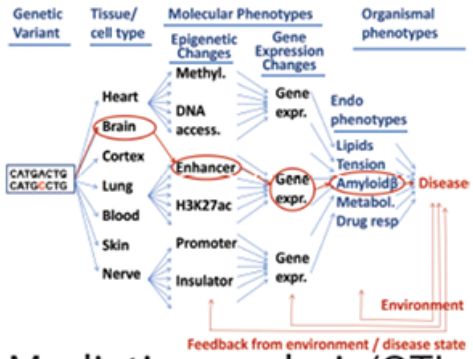
6.047 (U, AUS) - 6.878 (G, TQE) - HST.507

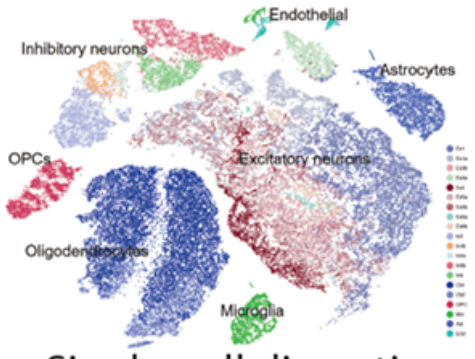
Fall 2019

Machine Learning in Genomics

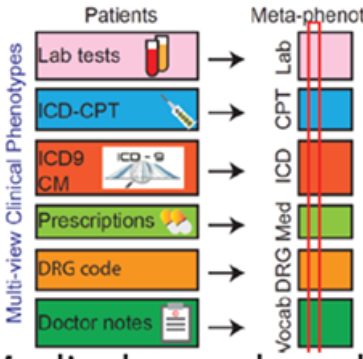
Dissecting the circuitry of human disease



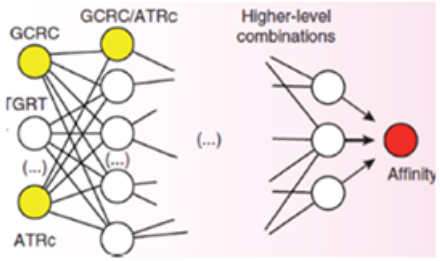
Mediation analysis/QTLs



Single-cell dissection

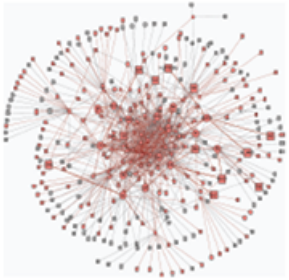


Medical record models

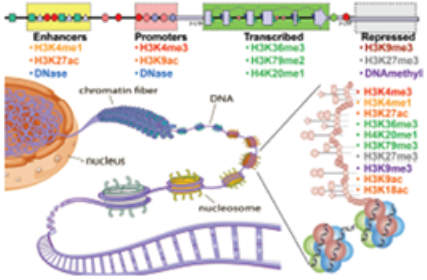


Deep learning

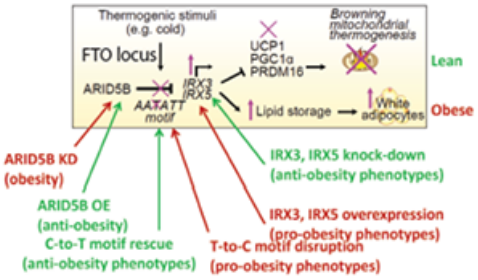
DNA motifs



Gene networks



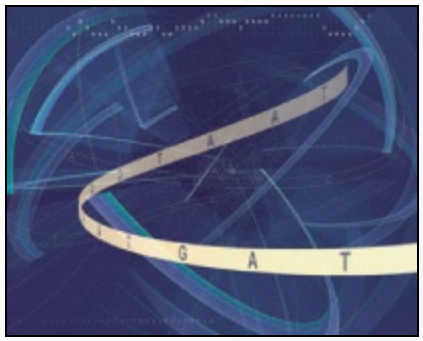
Epigenomics



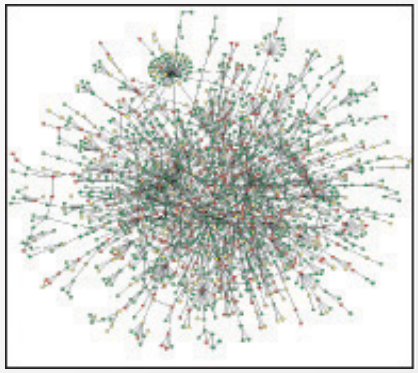
Manipulate disease circuitry

Multi-phenotype

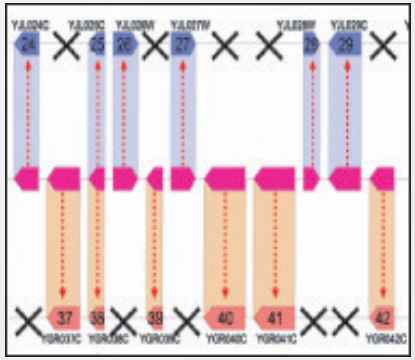
Prof. Manolis Kellis - MIT / CSAIL / Broad Institute



Rapid database search



Protein interaction network



Genome duplication

Machine Learning in Genomics

MIT 6.047 / 6.878
HSPH IMI.231
HST.507

Prof. Manolis Kellis

TA: Samuel Kim

I. Administrivia

Introduction to the course and its goals

Course organization and content

Homework and Quiz

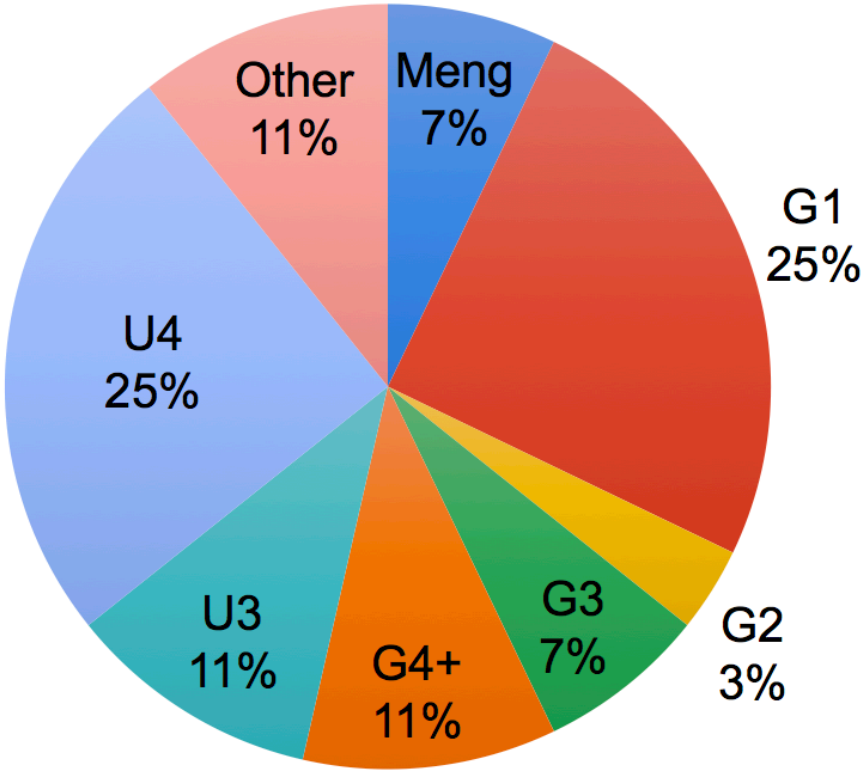
Term Project

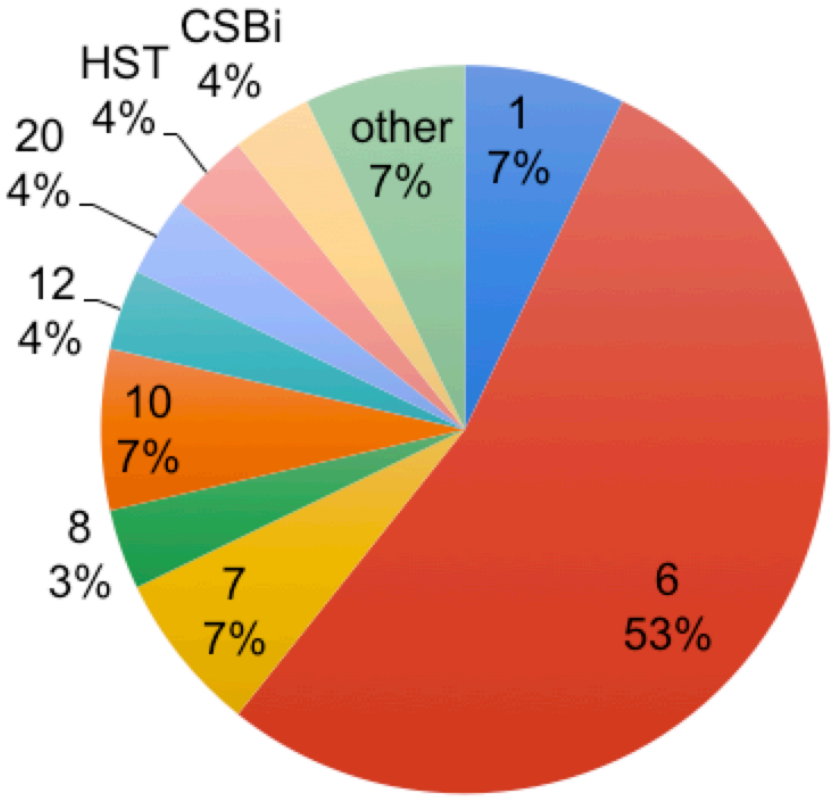
Introductions

- **Lecturer**
  - Manolis Kellis
    (MIT CSAIL, Computational Biology, Broad Institute)
  - My own research:
    Comparative genomics, Gene Regulation, Evolution, Epigenomics, Phylogenomics, Disease genomics, etc

- **TA**
  - Samuel Kim
  - 4th year PhD Student. Statistical genetics, Heritability, Gene network
  - Integrated circuits, synthetic biology background

Students (from first-day survey)





Are you currently doing research in computational biology?

- **No 46% (13/28). Yes 54% (15/28)**

- Advisor to sv.ai and hackathon participant
- Alexander Lab at WHOI: developing metatranscriptomic pipelines to analyze ocean biological datasets of eukaryotic phytoplankton. I use trimmomatic, salmon, fastqc, Trinity, etc.
- Am a student in the Sculpting Evolution group in the Media Lab. None of my current projects have a large computational biology component, but I would like to expand that aspect of the project.
- Cordero Lab at CEE, microbial community assembly, now 16S rRNA data but look forward to metagenomics
- Ecological systems biology in Gore Lab

- I am analyzing ~1 million single nuclei profiles with spatial indexes from the human and mouse cerebellum in a first author role. We aim to pair the well-characterized cell type morphologies in the cerebellum with their transcriptomes and to elucidate molecular spatial compartmentalization within and across cell types. I began this project as an RA in the Macosko lab at the Broad Institute

- I currently work in the Bueno laboratory at Brigham and Women's hospital. To date the lab has collected extensive genomic and transcriptomic data in lung cancer and mesothelioma. I'm personally driving a project examining the histological heterogeneity of mesothelioma using single cell (Seq-Well). In addition, we are using single cell transcriptomics (10X) to understand lymph node and tumor microenvironment in the context of treated and untreated lung cancer. I'm the nested computational biologist in the lab, and so asked to assist in all computational biology questions that come up (although my expertise is primarily in transcriptomics). As such, I feel this course could help provide a stronger foundation in genomics (variant calling, population structure, heritability, etc) that I never developed in my PhD, which would allow me to better explore the field on my own.

- I will be starting out at the Gifford Lab this semester. The project is not clearly defined yet, but it will pertain to antibody design.

- I will be working in computational biology. My project is not yet well-defined but I will begin by working with single cell RNA seq data
- I work with Professor Alexander Gusev at DFCI. My current project includes using a deep neural network (DNN) based classifier to predict primary sites of cancers of unknown primary. Primary datasets we're using are Dana-Farber Cancer Institute Profile data and AACR Project GENIE data.

- I would like to use more computational biology in my research in David Sontag's machine learning for healthcare lab. I'm currently comfortable using gene expression data in exploratory and predictive ML, but would like to broaden the biological/genetic data types I can work with. Specifically interested in using this data for precision medicine research.

- I'm working in the Collins lab to develop a diagnostic tool for Inflammatory bowel disease using the human gut microbiome

- Regev Lab: Single cell RNA seq analysis +/- exp. conditions

- Yes, my lab is a genetics lab, and we've been looking at the effect of enhancer elements on gene expression and disease.
- Yes. Working in White lab analyzing large datasets of peptide data form the EGFR pathway to understand the effect of point mutations and to try to generate a peptide-level model of species interactions.

Why are you taking this class? What do you hope to learn from it?

- A treatment for p1RCC. There has been no increase in OS for the last 14 years for this disease.
- To become more rigid in my understanding of the **underlying algorithms** that I use in my research.
- I hope to **learn computational biology methods** that I can use in my research.
- I wish to **advance my knowledge** and skills in computational biology, particularly about genomics/metagenomics.
- I'm interested in quantitative model and experiment of gene expression, regulatory network. I'm also interested in processing the large gene seq data.
- I've implemented many of the course topics in my projects without formal training. I'd like to "back-learn" the fundamentals so that my technical skills can be generalized to other -omics projects.
- Solidify and **reaffirm well-hashed concepts** and expand knowledge of machine learning; especially in genomics (as opposed to transcriptomics).
- To **learn about the state of the art** in computational biology, and the term project sounds like a useful experience.
- I want to understand the many problems and tools available in computational biology, as well as learn the important statistical considerations and assumptions when working with biological/genomic data
- I'd like to know more about a computational genomics and **how a ML can be effectively used in genomics.**
- I am interested in the intersection of machine learning and biological sciences, specifically for healthcare purposes. I hope to learn **how to use ML to solve problems in the space of healthcare**.
- Genomics is a skill set I'd like to pick up, given the increase in NGS etc.; also I'm mostly project driven, and would like to create a publication-worthy project at the end of the term (at least conference proceedings)
- I hope to learn new ways to apply computer science to biology. I think computational biology is a field I want to be a part of and dive into this fall.
- I want a stronger background in **practical applications of computational techniques** to biology. I would like to learn about clear examples of computational techniques actually being useful for improving human health and where I can fit in to help us live longer and healthier lives.
- To gain exposure to the **application of computer science in biology**
- how to build versatile evolutionary algorithms for program synthesis architectures that meta-reason.
- I love biology and really enjoyed 006 last semester, and I've always been interested in ML, so in general I'd like to learn about various computation bio ideas and become better versed in the vocabulary
- I want to learn how to apply computer science skills in my future career, likely in the biotech field.
- I would like to **practice using applications of machine learning.**
- Learn the **fundamentals about computational biology** to mesh biology with my computer science skills.
- Mostly to build on a previous knowledge base.
- I would like to learn what kind of scientific problems in biology you can address with computational methods.
- Integrative and comprehensive genomic and biology tools for understanding the problems in evolutionary and computational biology.
- Because my current position requires the knowledge
- I would like to know the algorithmic (and **mathematical) basis** for a lot of the computational biology work, because I would like to pursue a PhD in the field.
- I hope to the learn the general procedure for applying computational methods to biological problems. I have some foundation in both areas, but I am looking forward to a class that bridges these two fields.

Interest in specific topics (Phew)

	Don't know	Lowest	Low	Medium	High	Highest
Dynamic programming/Alignment	0	1	4	13	6	4
HMMs/Gene Finding	4	0	1	9	10	4
Gene expression analysis	1	2	2	7	10	6
RNA biology	2	0	1	10	11	4
Epigenomics	1	0	1	6	11	9
3D genome	3	1	2	14	5	3
Motif Discovery	6	3	2	3	7	7
Networks	1	1	3	4	11	8
Deep Learning	0	1	3	3	11	10
Population genetics	0	1	5	6	6	10
Disease associations	0	0	4	9	8	7
Quantitative Traits eQTLs	9	0	3	7	4	5
Linear Mixed Models / Heritability	4	0	4	6	7	7
Comparative Genomics	2	1	3	5	6	10
Phylogenetics/Phylogenomics	6	1	5	3	5	7
Single-Cell Biology	0	4	0	7	5	12
Electronic Health Records	0	12	2	8	4	2
Cancer Genomics	0	1	2	6	5	13

Other topics of interest

- single cell sequence
- Protein structure prediction, Personalized Medicine
- Just more about clustering and visualization

- How to create a pipeline between patients with rare diseases (who are willing to open their data) to your class.

- cellular automatons and their ability to model systems in nature

- Bayesian models and their applications in biology. Discussions surrounding clever ways to understand cancer with coupled transcriptomics and genomics data are also of great interest to me.
- any topic related to metagenomics
- Analysis of repetitive DNA

Course Information

- Lectures
  - TR 1pm – 2:30, Room 32-141
- Recitations:
  - On Friday at 3pm in 4-237
  - Recitations at MIT
- TA office hour:
  - Survey shows R works better than T
  - Tentatively R: 2:30-3:30 (after class; location: TBD)
- Course Website
  - http://stellar.mit.edu/S/course/6/fa19/6.047/
  - or simply: compbio.mit.edu/6.047 (redirects to stellar)
  - All handouts, lectures, notes, etc will be posted here.
- Course calendar:
  - On Google, add public calendar: “6.047 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

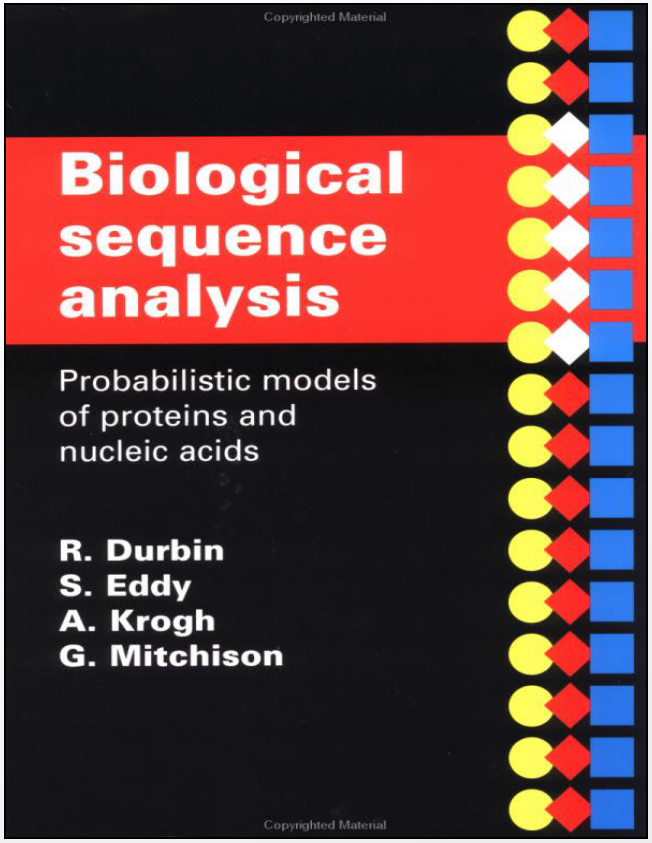
Course organized around bio/comp modules

- Each module corresponds to an active area of research
  - **1: Comparative genomics: Align/model genomes, DP, HMMs**
  - **2: Genes and Transcripts: RNA-seq, clustering, structure**
  - **3: Regulation: Epigenomics, TFs, Motifs, Network inference**
  - **4: Variation: Genetics, Human history, heritability, eQTLs**
  - **5: Evolution: Phylogeny, evolutionary sigs, WGD, assembly**
  - **6: Frontiers: Personal/Disease, 3D genomes, Pharma, Synth**
- For each module: First half ⇔ the 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
- For each module: Second half ⇔ the frontiers
  - Evolutionary signatures, Transcript analysis, lincRNAs, Network inference and analysis, Epigenomics, Recent human selection and ancestry, chromatin regulation, Missing heritability, 3D

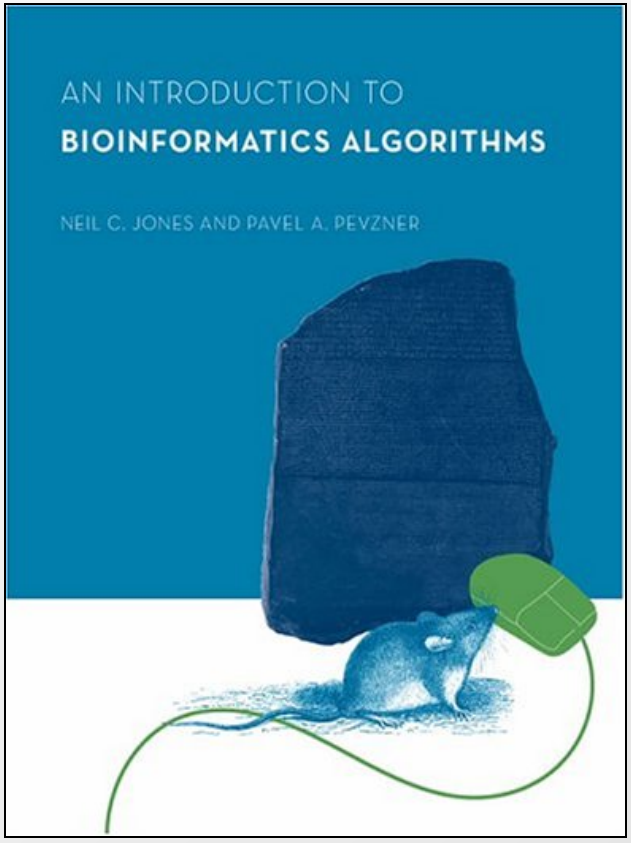
Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Mon 9/23**	PS1 out on:L1-L5	1	Thu, Sep 5	Introduction		L1	Algorithms, Machine Learning, Networks, Course Overview	1
	due Mon 9/23		Fri, Sep 6			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 10	Module I:	Foundations	L2	Dynamic Programming, Reusing computation, Iterative Functions, Exponential / Poly	2,3
			Thu, Sep 12			L3	Database search, Rapid string matching, Hashing	3
			Fri, Sep 13			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 17		Frontiers	L4	HMMs1: Evaluation, Parsing, posterior decoding, learning, HMM architectures	7,8
			Thu, Sep 19			L5	HMMs2: Applications, architectures, memory, gene finding, chromatin states	7,8
			Fri, Sep 20			No Classes - Student Holiday		
Find prev project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Mon 10/7**	PS2 out on:L6-R4	4	Tue, Sep 24	Module II:	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 26			L7	RNA structure and function. RNA world, RNA-seq, transcript structure, RNA folding	14,15
			Fri, Sep 27			R3	Recitation 3: Supervised Learning and Random Forest Classification	
			Fri, Sep 27	cts, self introductions, mentor intro, example projects, teamwork 32D-507				
		5	Tue, Oct 1		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 3			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due Mon 10/7		Fri, Oct 4			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
			Fri, Oct 4	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Thu 10/17. Presented on Fri 10/18**	PS3 out on:L10-R6	6	Tue, Oct 8	Module III:	Foundations	L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Thu, Oct 10			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Fri, Oct 11			R5	Recitation 5: Communication Lab	
		7	Tue, Oct 15			No Classes - Columbus Day Holiday		
			Thu, Oct 17		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due Mon 10/21		Fri, Oct 18			R6	Recitation 6: Motif Discovery, WEEDER, In vitro Motif Discovery - PBMs, Selex	
			Fri, Oct 18	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507				
Evaluate/discuss three peer proposals, NIH review format. **Reviews back Mon 10/28**	PS4 out on:L13-R8	8	Tue, Oct 22	Module IV:	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 24			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 25			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 25	Panel Review: Discuss Peer Projects. Feedback sent out from group reviews. 32D-463 (Star).				
		9	Tue, Oct 29		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/7**	due Mon 11/4		Thu, Oct 31			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
			Fri, Nov 1			R8	Recitation 8: Phylogenetic distance metrics, Coalescent Process	
	PS5 out on:L17-R10	10	Tue, Nov 5	Module V:	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 7			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 8			No Recitation, Veterans Day		
Continue making subst. progress on proposed milstones.Write outline of final report. **Midcourse report due Mon 11/25**	due Fri 11/15	11	Tue, Nov 12		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
			Thu, Nov 14			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
			Fri, Nov 15			R9	Recitation 9: Quiz Review	
	No more psets! (work on your final project)	12	Tue, Nov 19	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 21			L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Fri, Nov 22			R10	Recitation 10: Project Feedback, results, interpretation, directions	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/8**		13	Tue, Nov 26	Module VI:	Frontiers	L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Thu, Nov 28			No lecture, thanksgiving break - Thu Nov 28, 2019		
			Fri, Nov 29			No recitation, thanksgiving break		
		14	Tue, Dec 3			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 5			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 6			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Presentations on Tue 12/10**		15	Tue, Dec 10			L25	Final Presentations - Part I (1pm). 32-141 (Classroom)	
			Tue, Dec 10			L25	Final Presentations - Part I (2:30pm). 32D-463 (Star)	

Textbook / class notes / resources

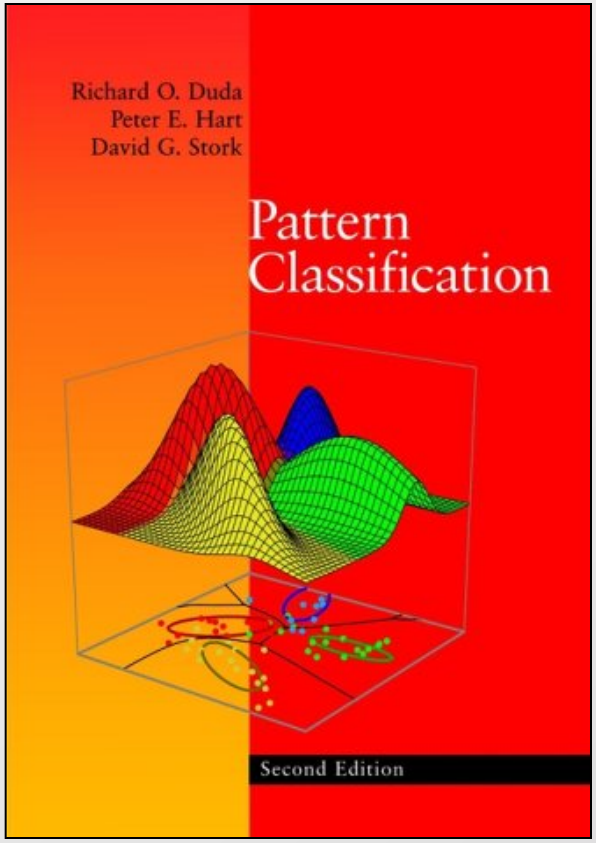
(Optional) Books for the Course



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!

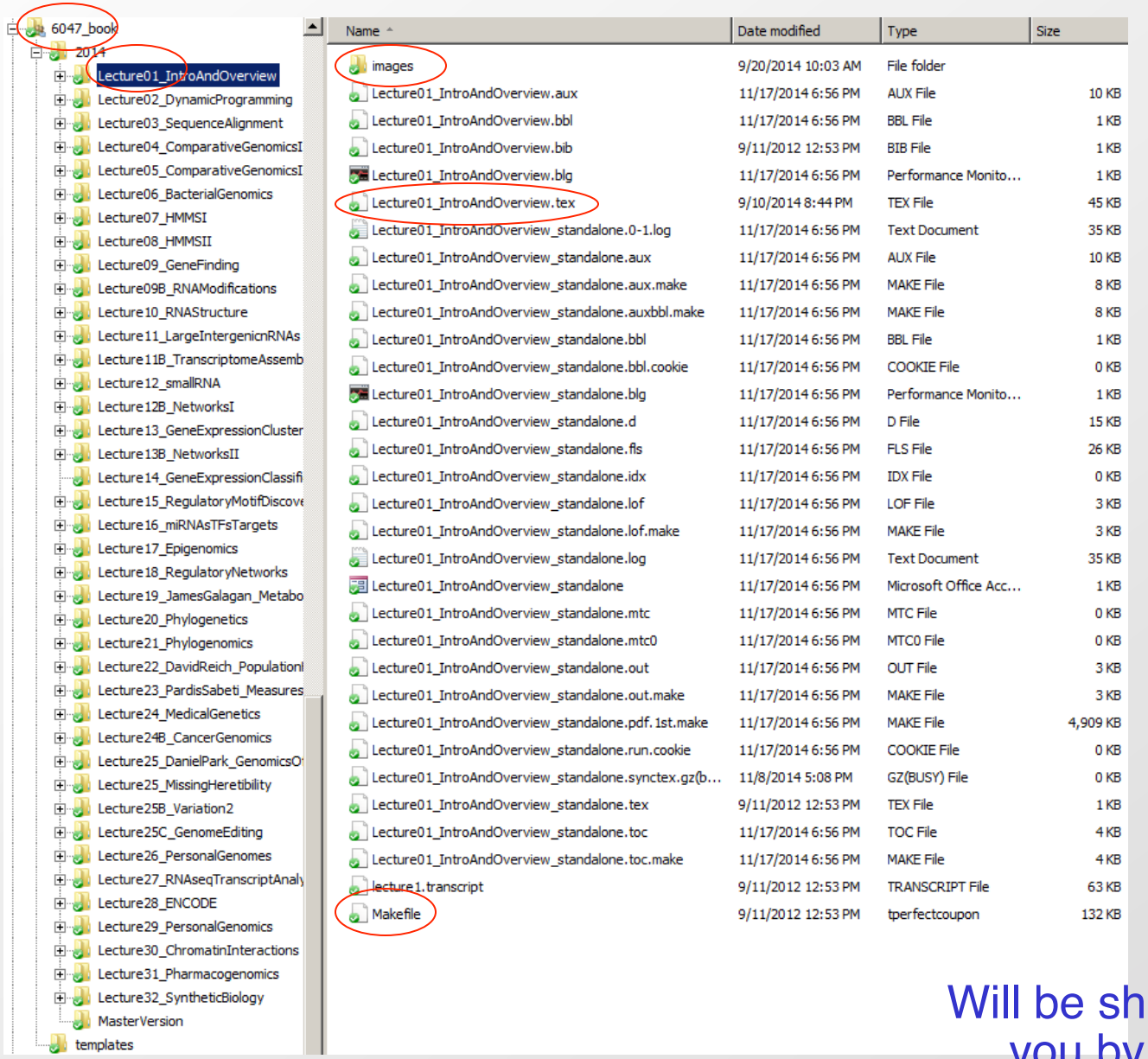
Availability: Current version online on Stellar, for registered students only

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!

Scribing details – DropBox



Will be shared with you by the TA

Sign up here if you haven't already

Lecture	Date	Topic	Existing chapters
L1	Thu, Sep 5	Algorithms, Machine Learning, Networks, Course Overview	1
L2	Tue, Sep 10	Dynamic Programming, Reusing computation, Iterative Functions, Exponential / Poly	2,3
L3	Thu, Sep 12	Database search, Rapid string matching, Hashing	3
L4	Tue, Sep 17	HMMs1: Evaluation, Parsing, posterior decoding, learning, HMM architectures	7,8
L5	Thu, Sep 19	HMMs2: Applications, architectures, memory, gene finding, chromatin states	7,8
L6	Tue, Sep 24	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
L7	Thu, Sep 26	RNA structure and function. RNA world, RNA-seq, transcript structure, RNA folding	14,15
L8	Tue, Oct 1	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
L9	Thu, Oct 3	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
L10	Tue, Oct 8	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
L11	Thu, Oct 10	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
L12	Thu, Oct 17	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
L13	Tue, Oct 22	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
L14	Thu, Oct 24	Disease Association Mapping, GWAS, organismal phenotypes	31
L15	Tue, Oct 29	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
L16	Thu, Oct 31	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
L17	Tue, Nov 5	Comparative genomics and evolutionary signatures	4
L18	Thu, Nov 7	Genome Scale Evolution, Genome Duplication	4,5.7
L19	Tue, Nov 12	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
L20	Thu, Nov 14	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
L21	Thu, Nov 21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
L22	Tue, Nov 26	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
L23	Tue, Dec 3	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
L24	Thu, Dec 5	Genome Engineering with CRISPR/Cas9 and related technologies	36

- https://tinyurl.com/compbioscribe

	Slides	Audio	Notes	Video1	Video2		
Module I: Comparative Genomics						Foundations	**Lecture 1 - Intro and Overview: Genomes and** Administrivia, Genomes, Information flow, Systems
							Lecture 2 - Dynamic Programming / Sequence Alignment Dynamic Programming, Sequence Alignment
							Lecture 3 - Hashing, Database Search, BLAST algorithm Sequence alignment II, review, local vs. global alignment, semi-numerical string matching, BLAST algorithm, probabilistic interpretation of score matrices (addendum - Linear-time deterministic string matching)
						Frontiers	**Lecture 4 - Comparative Genomics I - Evolutionary Signatures1** Evolutionary signatures of protein-coding genes
							Lecture 5 - Comparative Genomics II - Evolutionary Signatures2 Evolutionary signatures for diverse classes of functional elements
							Lecture 5 - Comparative Genomics III - Evolution Mechanisms of evolutionary change, Genome Duplication
Module II: Coding and Non-coding Genes						Foundations	**Lecture 6 - Hidden Markov Models I - Generation, Evaluation, Parsing** Intro to HMMs
							Lecture 7 - Hidden Markov Models II: Posterior Decoding, Learning Increasing state space, Posterior decoding, Supervised/Unsupervised Learning
						Frontiers	**Lecture 8 - Gene Identification: Gene structure, Semi-Markov, CRFs** Capturing gene structure, Semi-Markov models, Conditional Random Fields, Emerging lines of evidence
							Lecture 9 - RNA structure RNA world, folding algorithms, DP nussinov, energy models, probabilistic models, genomics of ncRNAs
Module III: Networks and Gene Regulation						Foundations	**Lecture 10A - Expression Clustering** Module III intro, Gene regulation, Microarrays, Expression Clustering, K-means, Fuzzy K-means, Expectation Maximization, Hierarchical Clustering, Hypergeometric
							Lecture 10B - Classification Clustering reprise, Bayesian Classification, Naive Bayes, Support Vector Machines
							Lecture 11 - Regulatory Motif Discovery TF binding, EM, EM extensions, Gibbs Sampling, Information Content, DNA/protein motifs
						Frontiers	**Lecture 12 - Regulatory Genomics** De novo motif discovery using comparative genomics, target prediction and motif instance identification, microRNA hairpin prediction, mature microRNA prediction
							Lecture 13 - Regulatory Networks Network structure, network inference, network-based prediction
							Lecture 14 - Epigenomics and chromatin states Using combinations of chromatin marks to interpret the human genome
Module IV: Evolution						Foundations	**Lecture 15 - Phylogenetics, Evolutionary Models, Tree Building** Introduction to phylogenetics, models of evolution, and tree building algorithms
							Lecture 16 - Phylogenomics Studying phylogenetics at the genome level, gene/species tree reconciliation, coalescence
							Lecture 17 - Population genomics Statistical genetics and human disease mapping
						Frontiers	**Lecture 18 - Population genetics and recent selection**
							Lecture 19 - Population history Population genomics and recent human history
Frontiers						Guest Lectures	**Lecture 20 - Metabolic modeling** Systems biology for modeling metabolism and regulation
							Lecture 21 - Bacterial Genomics and Microbiomics Systems biology for modeling metabolism and regulation
							Lecture 22 - Large intergenic non-coding RNAs Genome regulation by large intergenic non-coding RNAs

Will be posted on Stellar after each lecture

Lecture feedback: https://goo.gl/rV5XJi

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
  - Theoretical problem: pen/paper, explore algorithmic / statistical / machine learning aspect in detail/depth. (Typically additional advanced problem for 6.878)
- 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 stellar 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

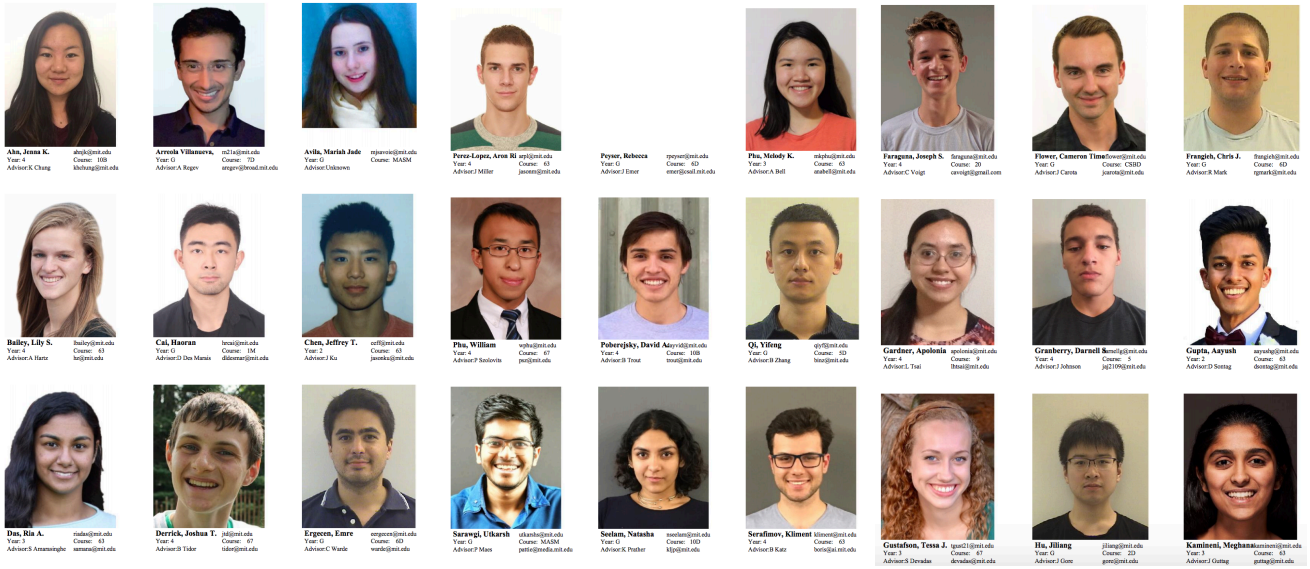
Final Project

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 imprvmts
  - 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

It's a team project

- Please make an effort to meet your peers!
- Form teams early with complementary expertise



Final Project at a Glance

	Project	Psets
Project planning	Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Mon 9/23**	PS1 out on:L1-L5 **due Mon 9/23**
	Find prev project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Mon 10/7**	PS2 out on:L6-R4 **due Mon 10/7**
	Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Thu 10/17. Presented on Fri 10/18**	PS3 out on:L10-R6 **due Mon 10/21**
	Evaluate/discuss three peer proposals, NIH review format. **Reviews back Mon 10/28**	PS4 out on:L13-R8 **due Mon 11/4**
	Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/7**	PS5 out on:L17-R10 **due Fri 11/15**
Project execution	Continue making subst. progress on proposed milstones.Write outline of final report. **Midcourse report due Mon 11/25**	No more psets! (work on your final project)
	Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/8**	
	Conference format slide pres. **Presentations on Tue 12/10**	

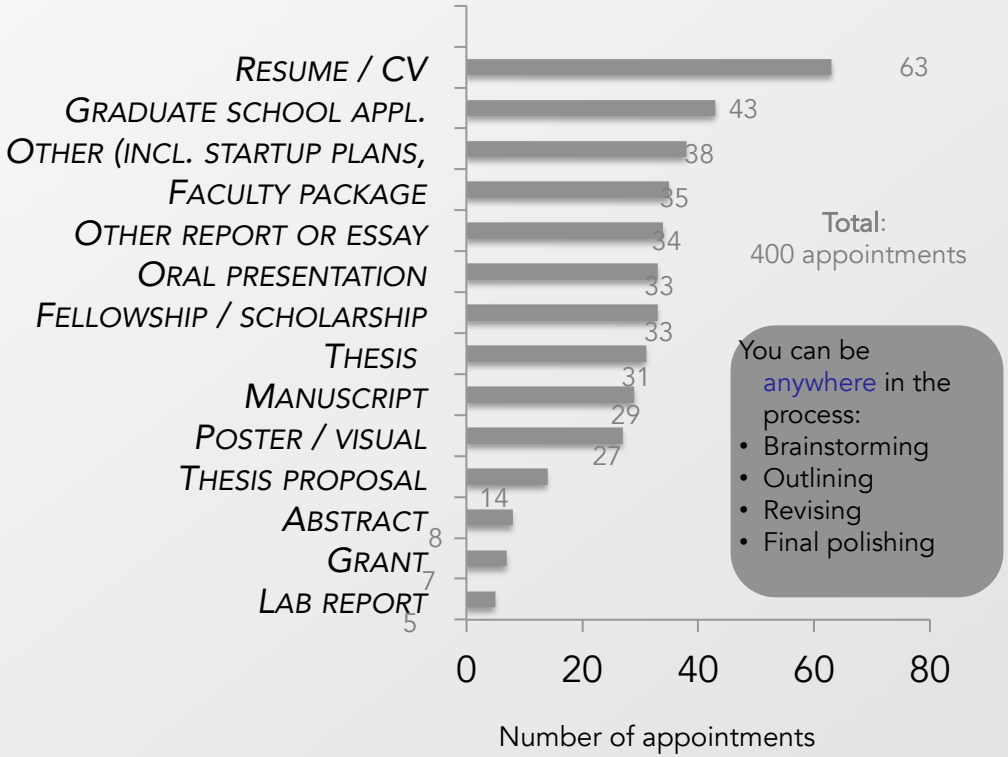
Details on the final project

- 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:



"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 Grading

- Grading:

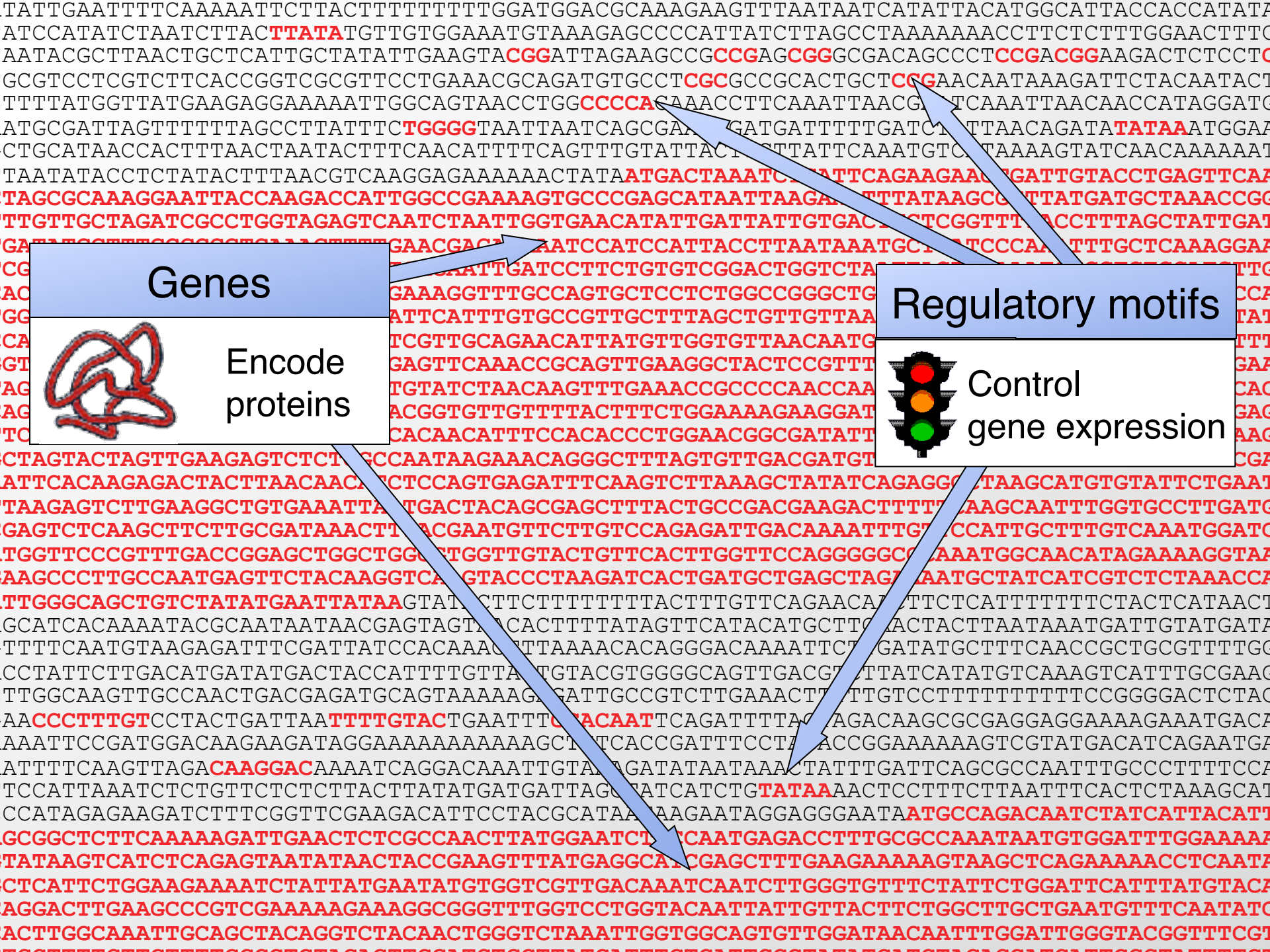
Problem sets 30%	Final Project 40%	Midterm 20%	Scrib10%

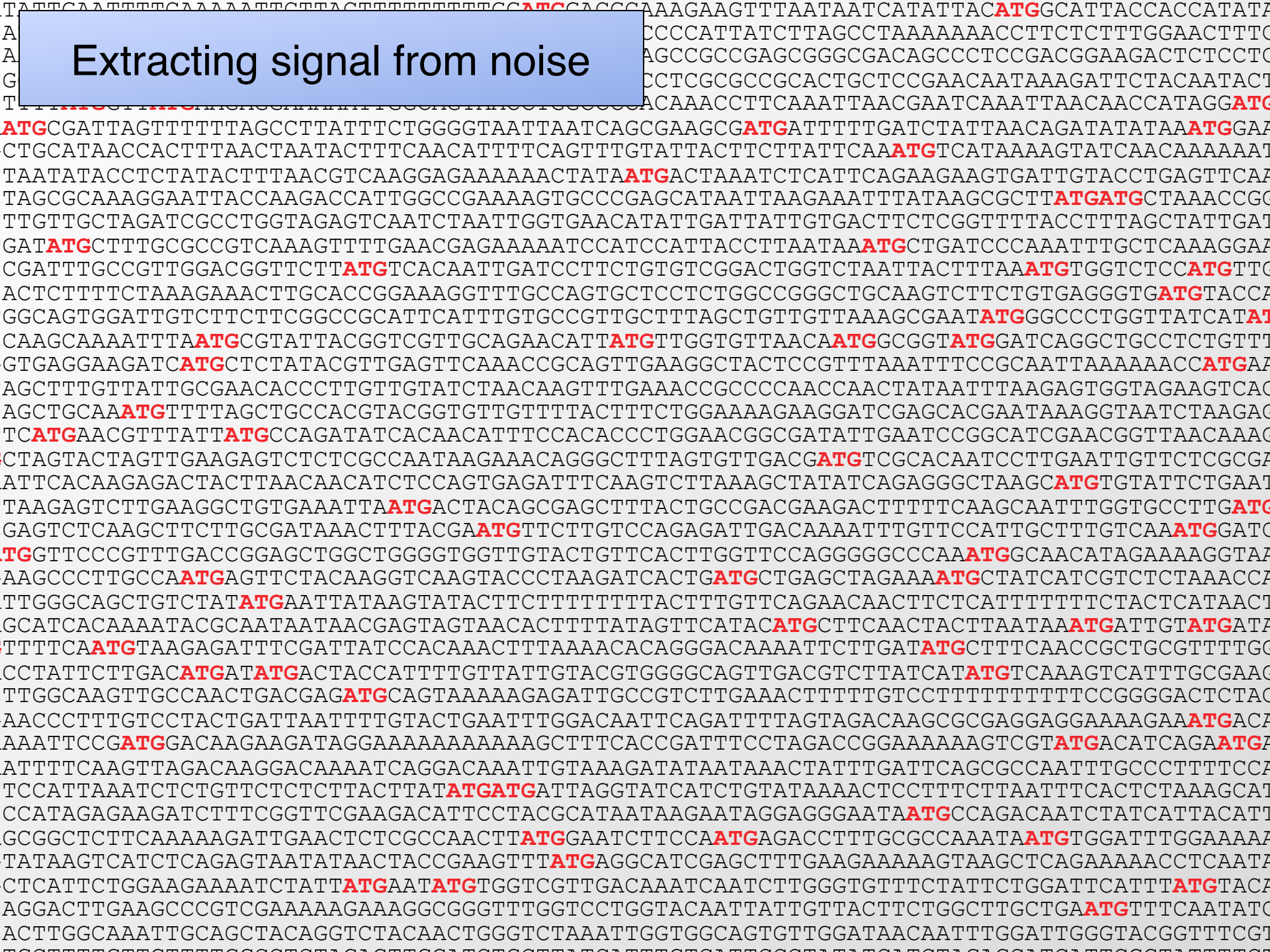
- 4 problem sets:
  - Each problem set: 7-10%, covers 3-4 lectures, contains 3-4 problems.
  - Algorithmic problems and programming assignments (PS1 out now)
  - Graduate version includes additional problem on current research
- Final project
  - Introduction to research in computational biology (7 weeks!)
  - Includes peer-reviewed NIH-style proposal and much feedback
- Quiz
  - In-class quiz (Tue Nov 22). No final exam.
- Collaboration policy
  - Collaboration allowed, but you must:
    - Work independently on each problem before discussing it
    - Write solutions on your own
    - Acknowledge sources and collaborators. No outsourcing.

Why Computational Biology ?

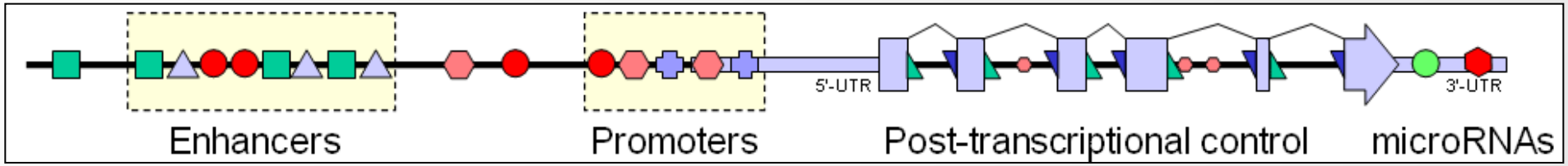
Why Computational Biology: Last year's answers

- 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


Genes
Encode proteins
Regulatory motifs
Control gene expression


Extracting signal from noise

The components of genomes and gene regulation



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

Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Tue 9/26**	PS1 out on:L1-L5	1	Thu, Sep 7	Introduction		L1	Intro: Biology, Algorithms, Machine Learning, Course Overview	1
			Fri, Sep 8			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 12	Module I: Aligning and Modeling Genomes	Foundations	L2	Alignment I: Dynamic Programming, Global and local alignment	2,3
			Thu, Sep 14			L3	Alignment II: Database search, Rapid string matching, BLAST, BLOSUM	3
			Fri, Sep 15			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 19		Frontiers	L4	Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms	7,8
			Thu, Sep 21			L5	Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch	8
	due Tue 9/26		Fri, Sep 22				No classes - student holiday	
			Mon, Sep 25	Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507				
Identify previous project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Tue 10/3**	PS2 out on:L6-R4	4	Tue, Sep 26	Module II: Gene Expression and Epigenomics	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 28			L7	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr	14,15
			Fri, Sep 29			R3	Recitation 3: Affinity Propagation Clustering and Random Forest Classification	
		5	Tue, Oct 3		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 4			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due Tue 10/10		Fri, Oct 6			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Tue 10/19. Presented on Fri 10/20**			Fri, Oct 6	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
	PS3 out on:L10-R6	6	Tue, Oct 10	Module III: Regulatory Genomics and Networks	Foundations		No Classes - Columbus Day Holiday	
			Thu, Oct 12			L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Fri, Oct 13			R5	Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex	
		7	Tue, Oct 17			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Thu, Oct 19		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due Tue 10/24		Fri, Oct 20			R6	Recitation 6: Networks review, Recommendation systems, EHR, PheWAS	
			Fri, Oct 20	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507 at 4-5pm				
Evaluate/discuss three peer proposals, NIH review format. **Review Panels Fri 10/27** **Reviews back Tue 10/31**	PS4 out on:L13-R8	8	Tue, Oct 24	Module IV: Population Genetics and Disease Genomics	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 26			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 27			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 27	Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507				
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/9**		9	Tue, Oct 31		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
			Thu, Nov 2			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
	due Tue 11/7		Fri, Nov 3			R8	Recitation 8: Rare Variants, ExAC	
	PS5 out on:L17-R10	10	Tue, Nov 7	Module V: Comparative Genomics and Evolution	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 8			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 10				No Recitation, Veterans Day	
Continue making substantial progress on proposed milstones.Write outline of final report. **Midcourse report due Wed 11/22.**		11	Tue, Nov 14		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
			Thu, Nov 16			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
	due Fri 11/17		Fri, Nov 17			R9	Recitation 9: Phylogenetic distance metrics, Coalescent Process	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/10**	No more psets! (work on your final project)	12	Tue, Nov 21	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 23				No lecture, thanksgiving break - Thu Nov 26, 2015	
			Fri, Nov 24				No recitation, thanksgiving break	
		13	Tue, Nov 28	Module VI: Current Research Directions	Frontiers	L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Thu, Nov 30			L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Fri, Dec 1			R10	Recitation 10: Project Feedback, results, interpretation, directions	
		14	Tue, Dec 5			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 7			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 8			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Talks on Tue 12/12**		15	Tue, Dec 12			L25	Final Presentations - Part I (11am). 32-G8 reading room	
			Tue, Dec 12			L25	Final Presentations - Part I (1pm). 32-141	

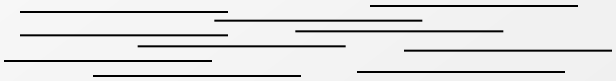
* readings refer to chapters in compiled 2016 scribe notes, available in the materials folder on Stellar

** recitation topics will be adjusted to respond to lecture and student needs

Overview of the 5 modules

Challenges in Computational Biology

4 Genome Assembly



5 Regulatory motif discovery

1 Gene Finding

DNA

2 Sequence alignment

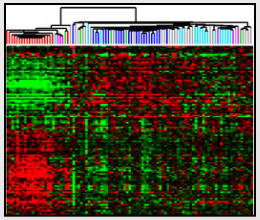
6 Comparative Genomics

7 Evolutionary Theory

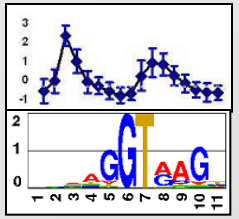
3 Database lookup

8 Gene expression analysis

RNA transcript

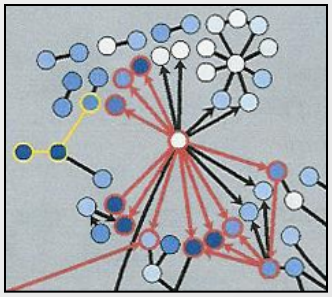


9 Cluster discovery



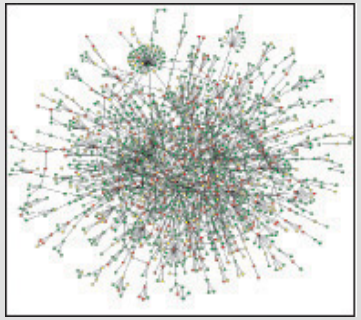
10 Gibbs sampling

11 Protein network analysis



12 Metabolic modelling

13 Emerging network properties

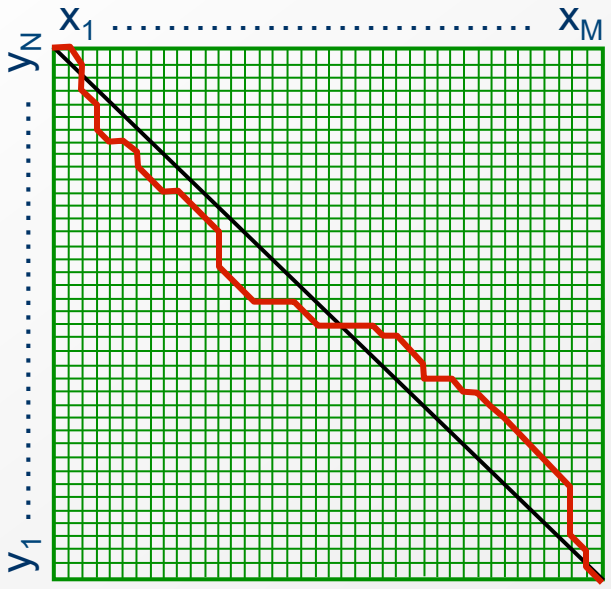


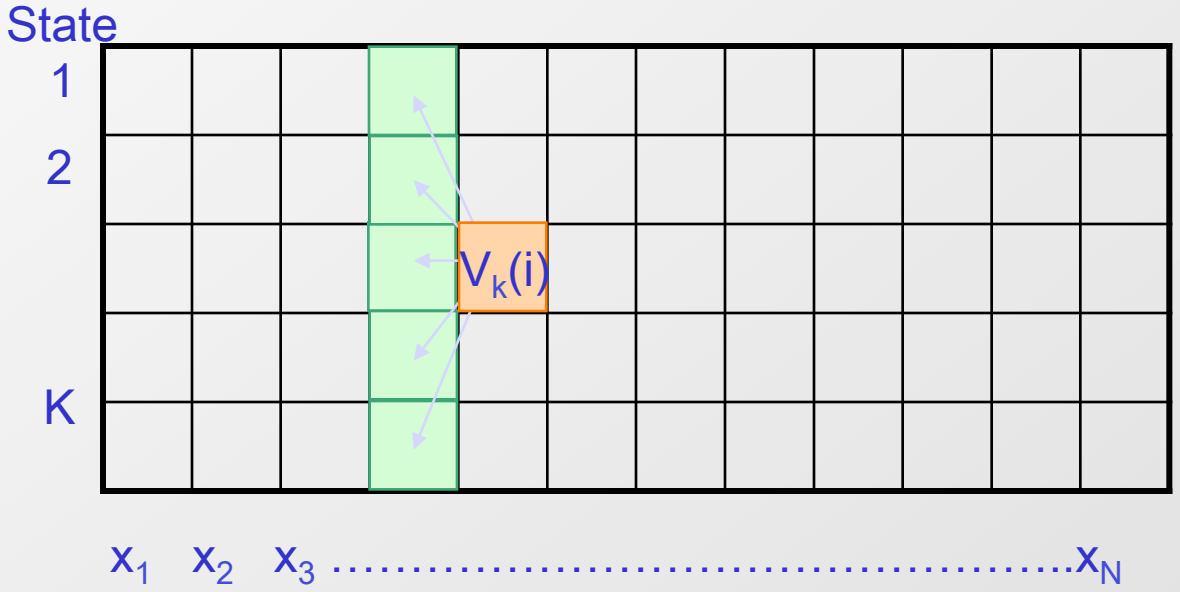
Module 1: Aligning and Modeling Genomes

Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Tue 9/26**	PS1 out on:L1-L5 **due Tue 9/26**	1	Thu, Sep 7	Introduction		L1	Intro: Biology, Algorithms, Machine Learning, Course Overview	1
			Fri, Sep 8			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 12	Module I: Aligning and Modeling Genomes	Foundations	L2	Alignment I: Dynamic Programming, Global and local alignment	2,3
			Thu, Sep 14			L3	Alignment II: Database search, Rapid string matching, BLAST, BLOSUM	3
			Fri, Sep 15			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 19		Frontiers	L4	Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms	7,8
			Thu, Sep 21			L5	Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch	8
			Fri, Sep 22				No classes - student holiday	
			Mon, Sep 25	Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507				

- 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…

Module II: Gene expression analysis and transcripts

Identify previous project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Tue 10/3**	PS2 out on:L6-R4	4	Tue, Sep 26	Module II: Gene Expression and Epigenomics	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 28			L7	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr	14,15
			Fri, Sep 29			R3	Recitation 3: Affinity Propagation Clustering and Random Forest Classification	
		5	Tue, Oct 3		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 4			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
Form teams of two, specify project goals, division of work, milestones.	due Tue 10/10		Fri, Oct 6			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
			Fri, Oct 6		Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507			

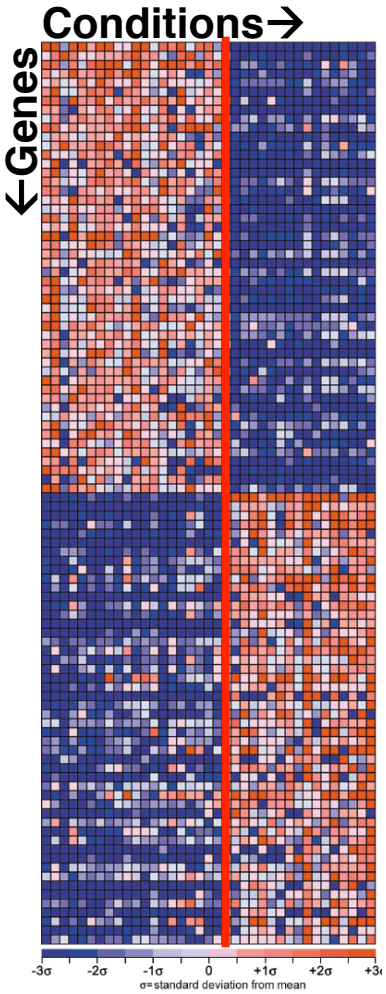
- Computational foundations:
  - Unsupervised Learning: Expectation Maximization
  - Supervised learning: generative/discriminative models
  - Read mapping, significance testing, splice graphs
- Biological frontiers:
  - PS2: Modeling conservation, GC content, CpG islands
  - L6/L7: Genome annotation and parsing
  - L8: Gene expression analysis: cluster genes/conditions
  - L9: Regulatory motif discovery: EM, gibbs sampling, info

Natural 1st step: group similar rows/columns

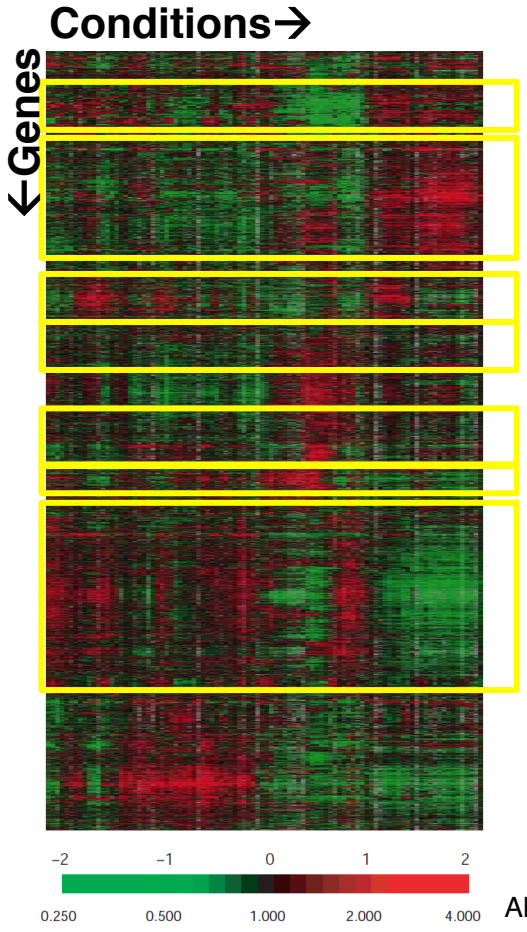
Clustering

➔ Similar cell types

➔ Similarly-behaving groups of genes



Armstrong, Nature Gen 2002



Alizadeh, Nature 2000

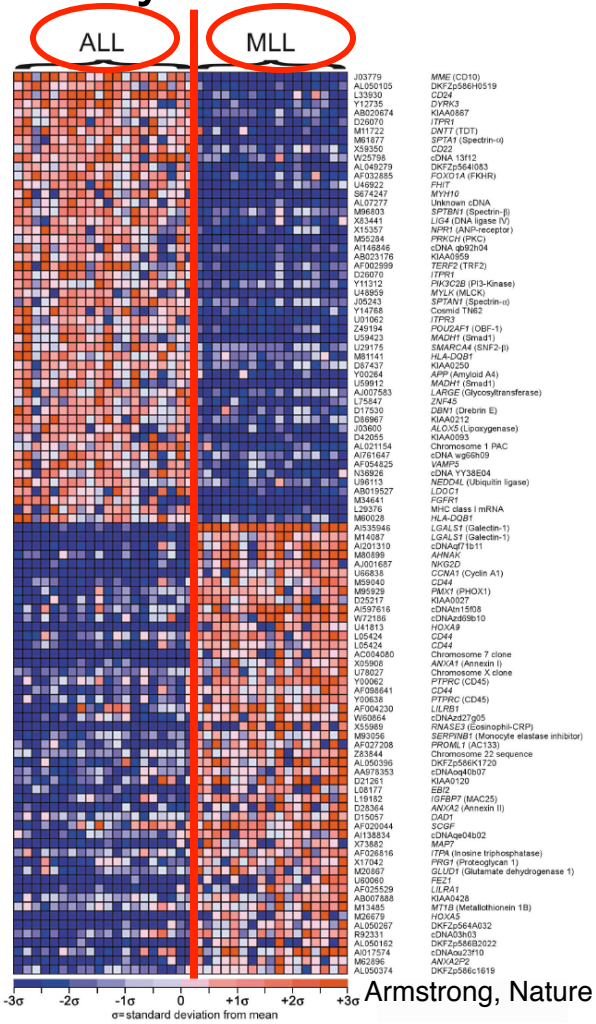
Reveal common 'conditions'

Reveal common gene behaviors

If labels are known: find more of same type

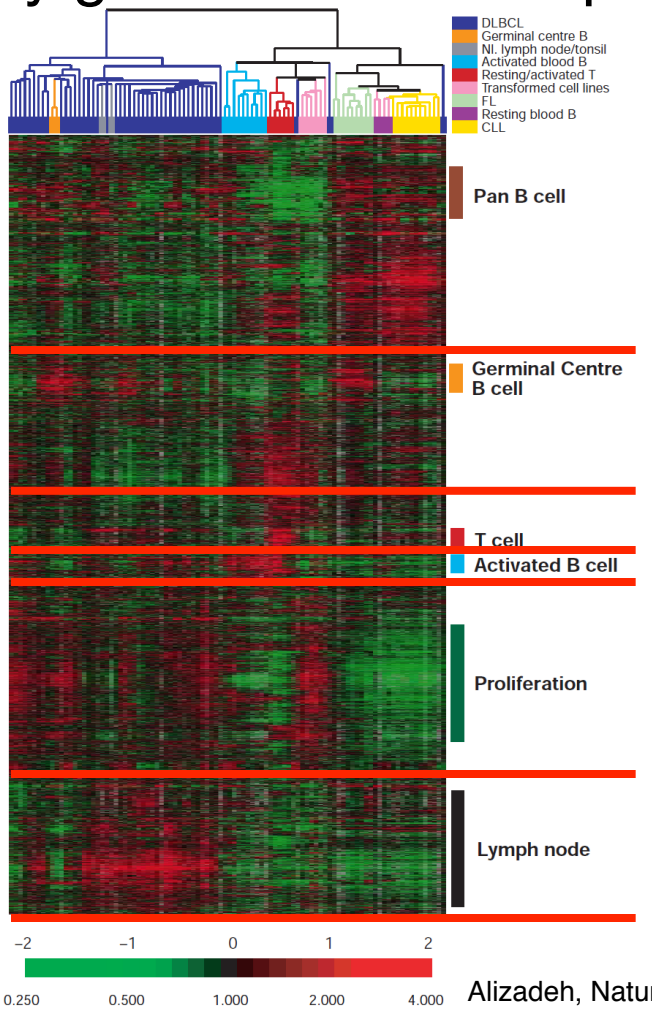
Classification

➔ Classify diseases



Armstrong, Nature Gen 2002

➔ Classify genes in different pathways



Alizadeh, Nature 2000

Find features that distinguish known classes

Find additional members of existing gene classes
Predict function of uncharacterized genes

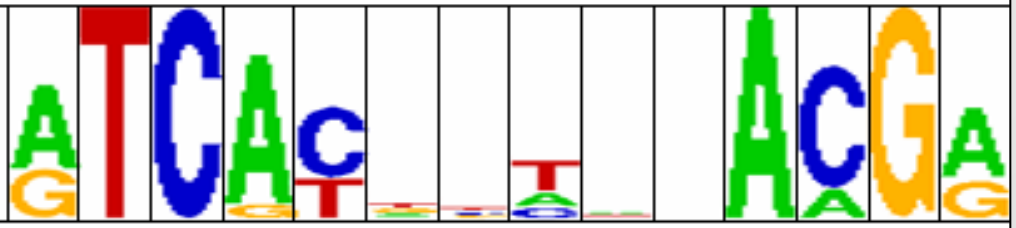
Module III: Epigenomics and gene regulation

datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Tue 10/19. Presented on Fri 10/20**	PS3 out on:L10-R6 **due Tue 10/24**	6	Tue, Oct 10	Module III: Regulatory Genomics and Networks	Foundations		No Classes - Columbus Day Holiday	
			Thu, Oct 12			L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Fri, Oct 13			R5	Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex	
		7	Tue, Oct 17			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Thu, Oct 19		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
			Fri, Oct 20			R6	Recitation 6: Networks review, Recommendation systems, EHR, PheWAS	
			Fri, Oct 20		Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507 at 4-5pm			

- 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

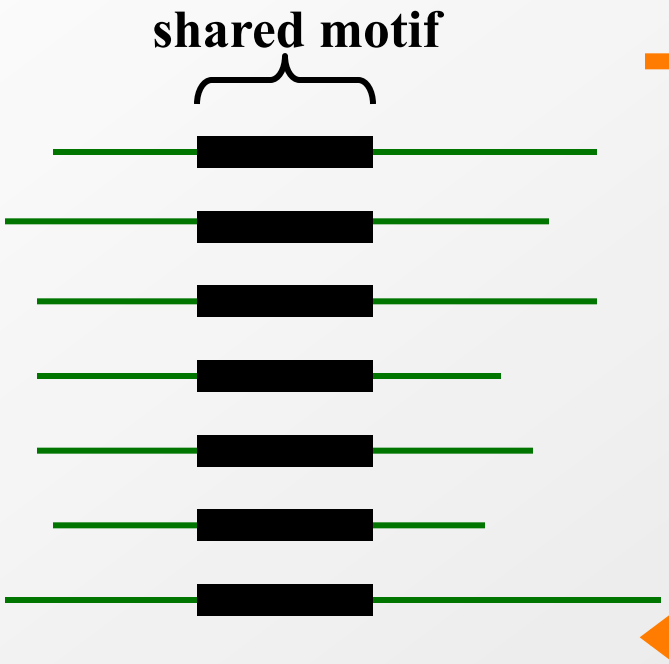
Target genes bound by ABF1 regulator		Coordinates		Genome sequence at bound site
ACS1	acetyl CoA synthetase	-491	-479	|ATCATTCTGGACG|
ACS1	acetyl CoA synthetase	-433	-421	|ATCATCTCGGACG|
ACS1	acetyl CoA synthetase	-311	-299	|ATCATTTGCCACG|
CHA1	catabolic L-serine dehydratase	-280	-254	A|ATCACCGCGAACG|GA
ENO2	Enolase	-470	-461	ggcgttat|GTCACTAACGACG|tgcacca
HMR	silencer	-256	-283	ATCAATAC|ATCATAAAATACG|AACGATC
LPD1	lipoamide dehydrogenase	-288	-300	gat|ATCAAAATTAACG|tag
LPD1	lipoamide dehydrogenase	-301	-313	gat|ATCACCGTTGACG|tca
PGK	phosphoglycerate kinase	-523	-496	CAAACAA|ATCACGAGCGACG|GTAATTTC
RPC160	RNA pol III/C 160 kDa subunit	-385	-349	|ATCACTATATACG|TGAA
RPC40	RNA pol III/C 40 kDa subunit	-137	-116	|GTCACTATAAACG|
rpL2	ribosomal protein L2	-185	-167	TAAT|aTCAcgtcACACG|AC
SPR3	CDC3/10/11/12 family homolog	-315	-303	|ATCACTAAATACG|
YPT1	TUB2	-193	-172	CCTAG|GTCACTGTACACG|TATA

Position		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Position Weight Matrix (PWM)	A	**56**	4	4	**81**	4	23	15	27	31	31	**89**	23	4	**58**
	G	**32**	4	4	12	4	31	23	4	19	23	4	4	**89**	**35**
	C	4	4	**89**	4	**58**	12	23	19	19	23	4	**69**	4	4
	T	4	**89**	4	4	**35**	35	39	50	31	23	4	4	4	4
Motif Logo															
Consensus		**R**	**T**	**C**	**A**	**Y**	N	N	H	N	N	**A**	**C**	**G**	**R**

- 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



sequence positions

	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

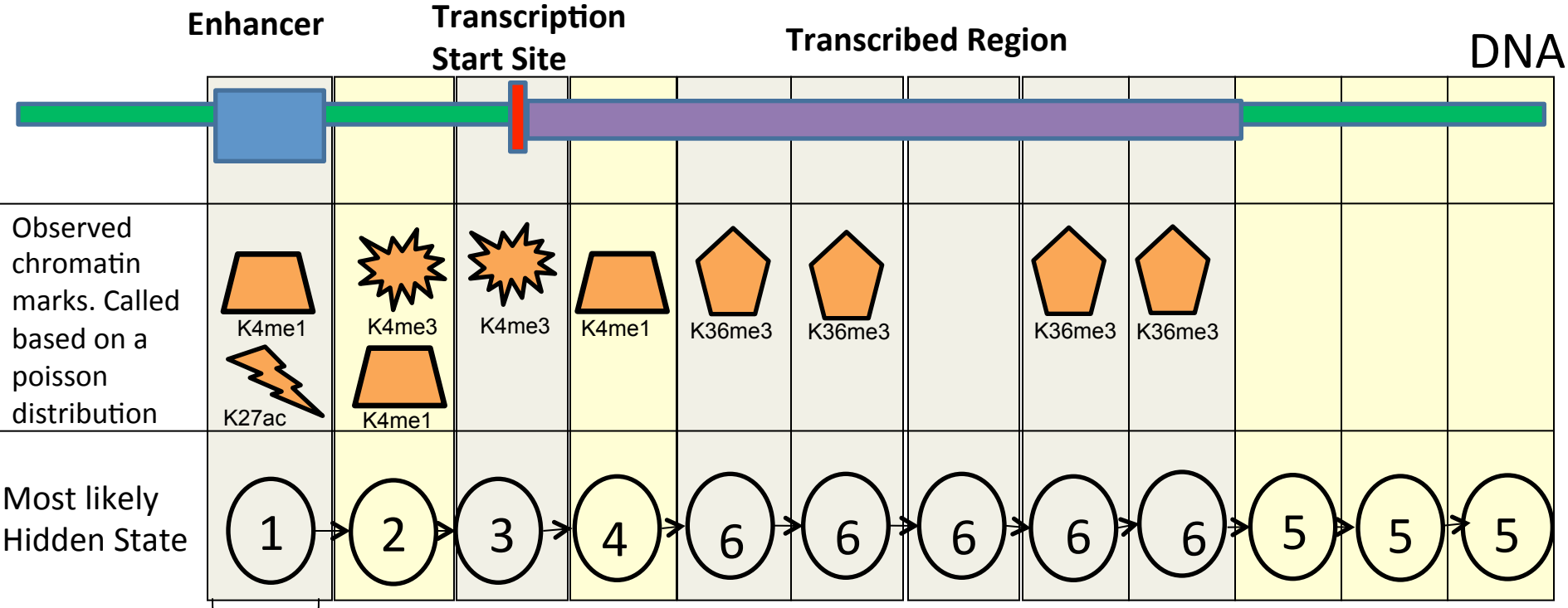
given profile matrix

- easy to find starting position probabilities

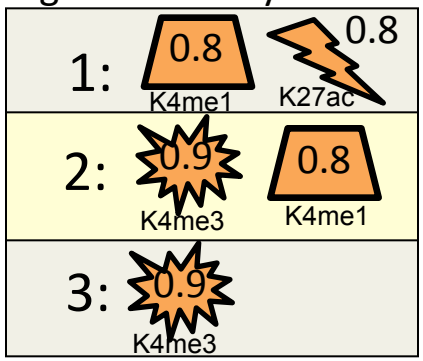
Key idea: Iterative procedure for estimating both, given uncertainty

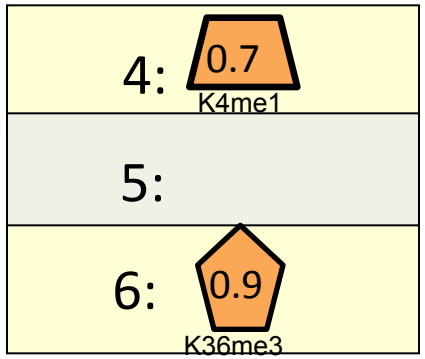
(learning problem with hidden variables: the starting positions)

Multivariate HMM for Chromatin States

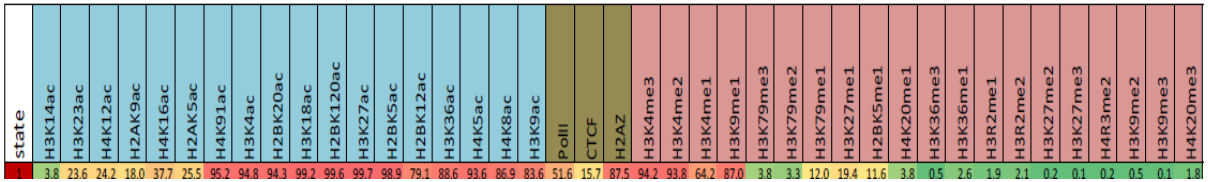


High Probability Chromatin Marks in State





All probabilities are learned from the data



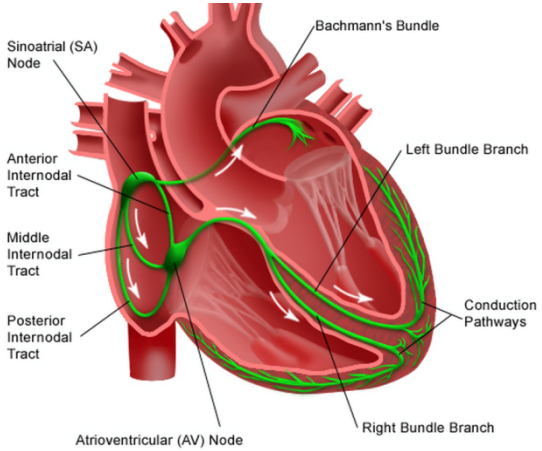
Ernst and Kellis
Nature Biotech 2010

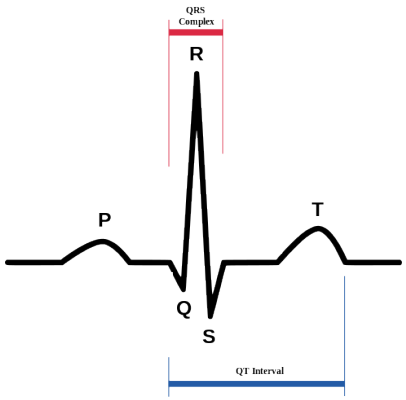
Modules IV and V: Evolution/phylogeny/populations

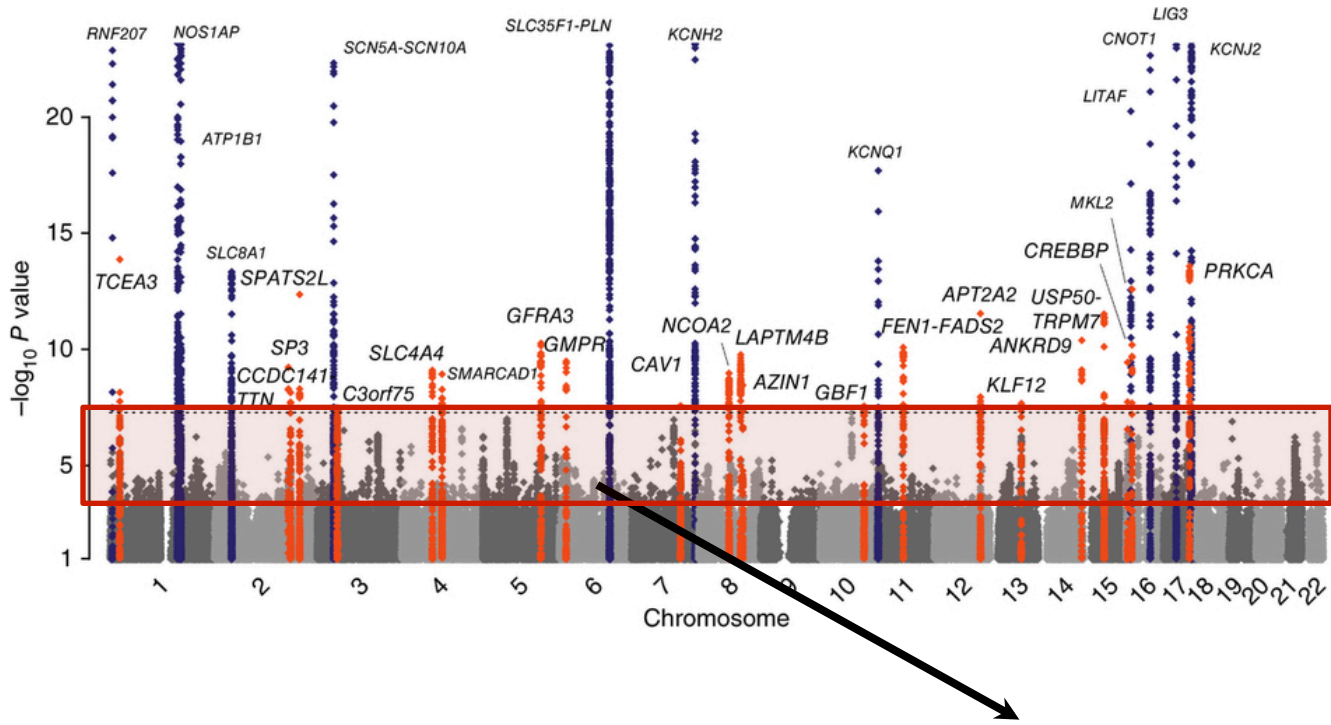
Evaluate/discuss three peer proposals, NIH review format. **Review Panels Fri 10/27** **Reviews back Tue 10/31**	PS4 out on:L13-R8	8	Tue, Oct 24	Module IV: Population Genetics and Disease Genomics	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 26			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 27			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 27		Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507			
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/9**	due Tue 11/7	9	Tue, Oct 31		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
			Thu, Nov 2			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
			Fri, Nov 3			R8	Recitation 8: Rare Variants, ExAC	
	PS5 out on:L17-R10	10	Tue, Nov 7	Module V: Comparative Genomics and Evolution	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 8			L18	Genome Scale Evolution, Genome Duplication	4,5.7
Continue making substantial progress on proposed milstones.Write outline of final report. **Midcourse report due Wed 11/22.**			Fri, Nov 10				No Recitation, Veterans Day	
	due Fri 11/17	11	Tue, Nov 14		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
			Thu, Nov 16			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
			Fri, Nov 17			R9	Recitation 9: Phylogenetic distance metrics, Coalescent Process	

- 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 21st, In-class quiz on Nov 22nd
  - No lab 4! Then entire focus shifts to projects, Thanksgiving, Frontiers

Characterizing sub-threshold variants in heart arrhythmia







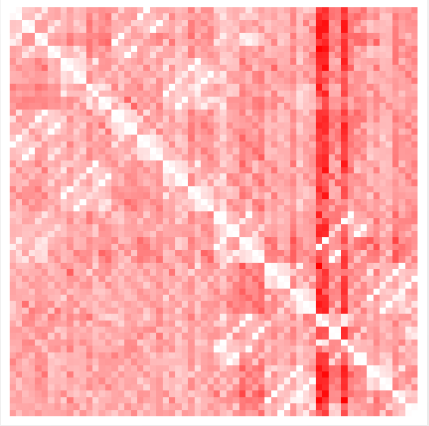
Focus on sub-threshold variants (e.g. rs1743292 $P=10^{-4.2}$)

Trait: QRS/QT interval

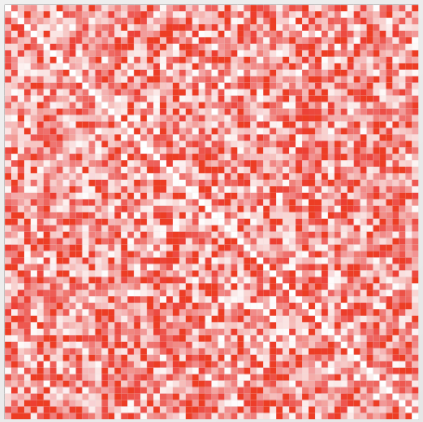
(1) Large cohorts, (2) many known hits
(3) well-characterized tissue drivers

Structure of genetic code ⇔ 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

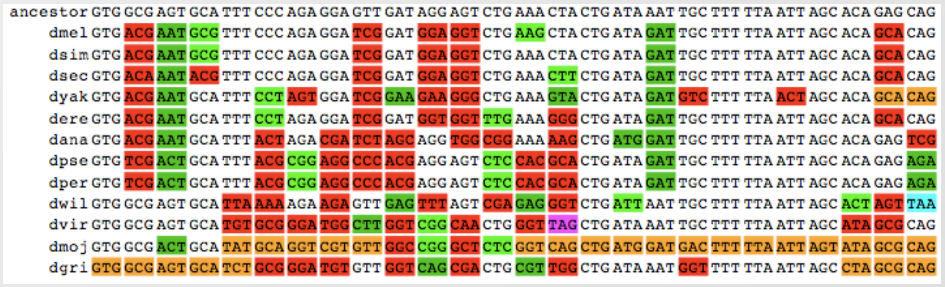


$\mathbf{Q}_C$ estimated from known coding regions



$\mathbf{Q}_N$ estimated from non-coding regions

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



$$\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t}) = \frac{1}{10^{117}}$$

$$\Pr(\text{Leaves}; \mathbf{Q}_N, \underline{t}) = \frac{1}{10^{152}}$$

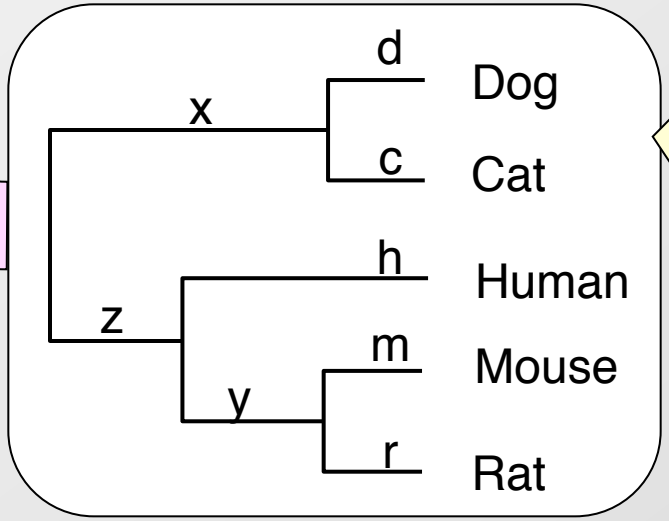
$$\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t}) = \frac{1}{10^{275}}$$

$$\Pr(\text{Leaves}; \mathbf{Q}_N, \underline{t}) = \frac{1}{10^{254}}$$

Distance matrix ⇔ Phylogenetic tree

	Hum	Mou	Rat	Dog	Cat
Human	0	4	5	7	6
Mouse	h.y.m	0	3	8	5
Rat	h.y.r	m.r	0	9	7
Dog	h.z.x.d	m.y.z.x.d	r.y.z.x.d	0	2
Cat	h.z.x.c	m.y.z.x.c	r.y.z.x.c	d.c	0

Tree implies a distance matrix M_{ij}

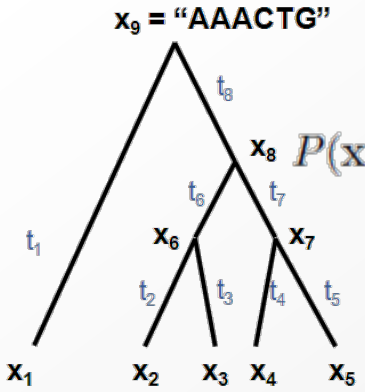


Map distances D_{ij} to a tree

$\min \sum_{ij} (D_{ij}-M_{ij})^2$

Goal:

Minimize discrepancy between observed distances and tree-based distances

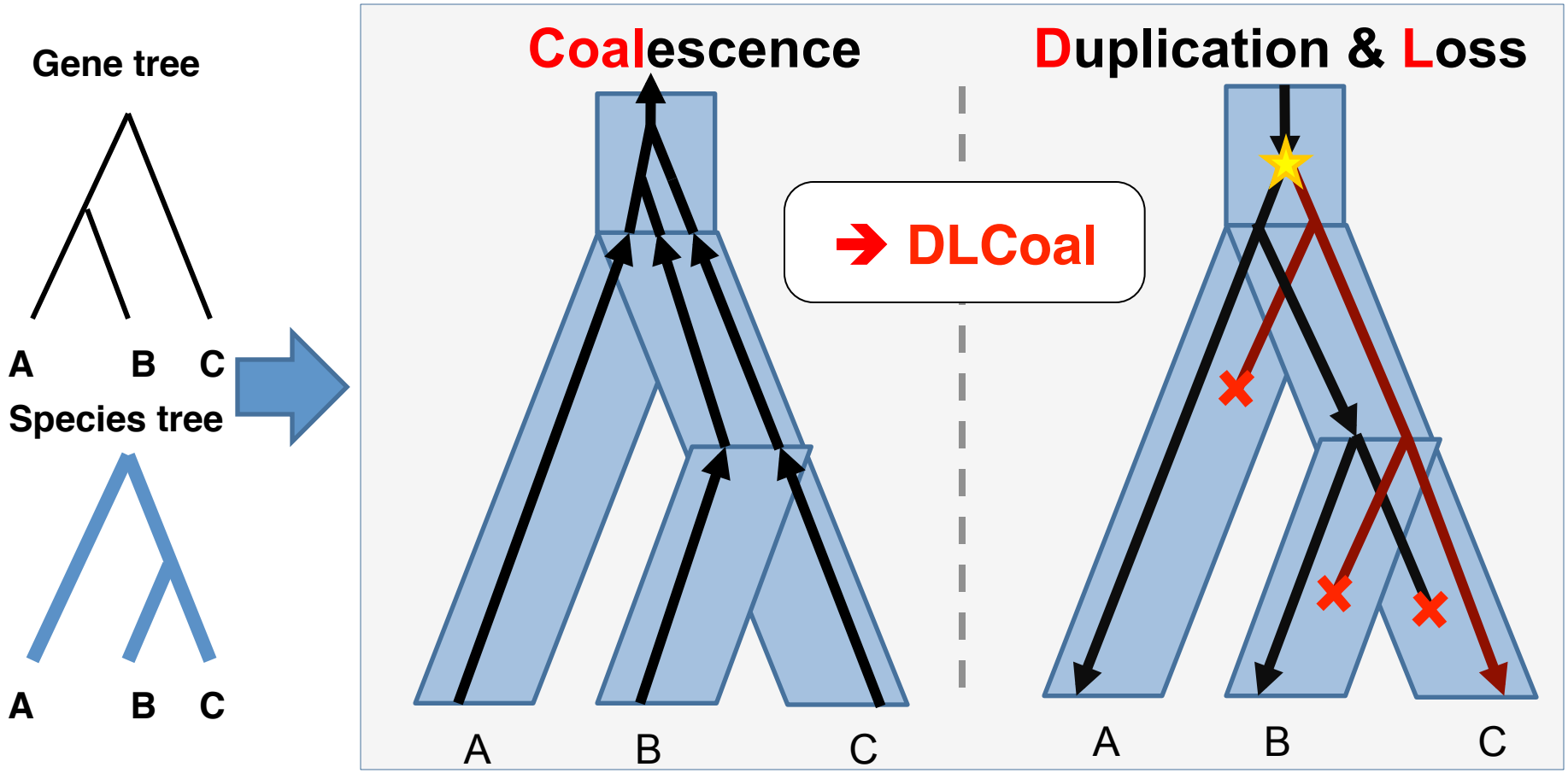


'Peeling' algorithm for P(D|B,T) term

$$P(x_1,\ldots,x_{2n-1}|T,t) = P(x_1|x_2,\ldots,x_{2n-1},T,t)P(x_2|x_3,\ldots,x_{2n-1},T,t)\ldots P(x_{2n-1}|T,t)$$
$$= P(x_1|x_{parent(1)},t_1)P(x_2|x_{parent(2)},t_2)\ldots P(x_{2n-1})$$
$$= P(x_{2n-1})\prod_{i=1}^{2n-2} P(x_i|x_{parent(i)},t_i)$$

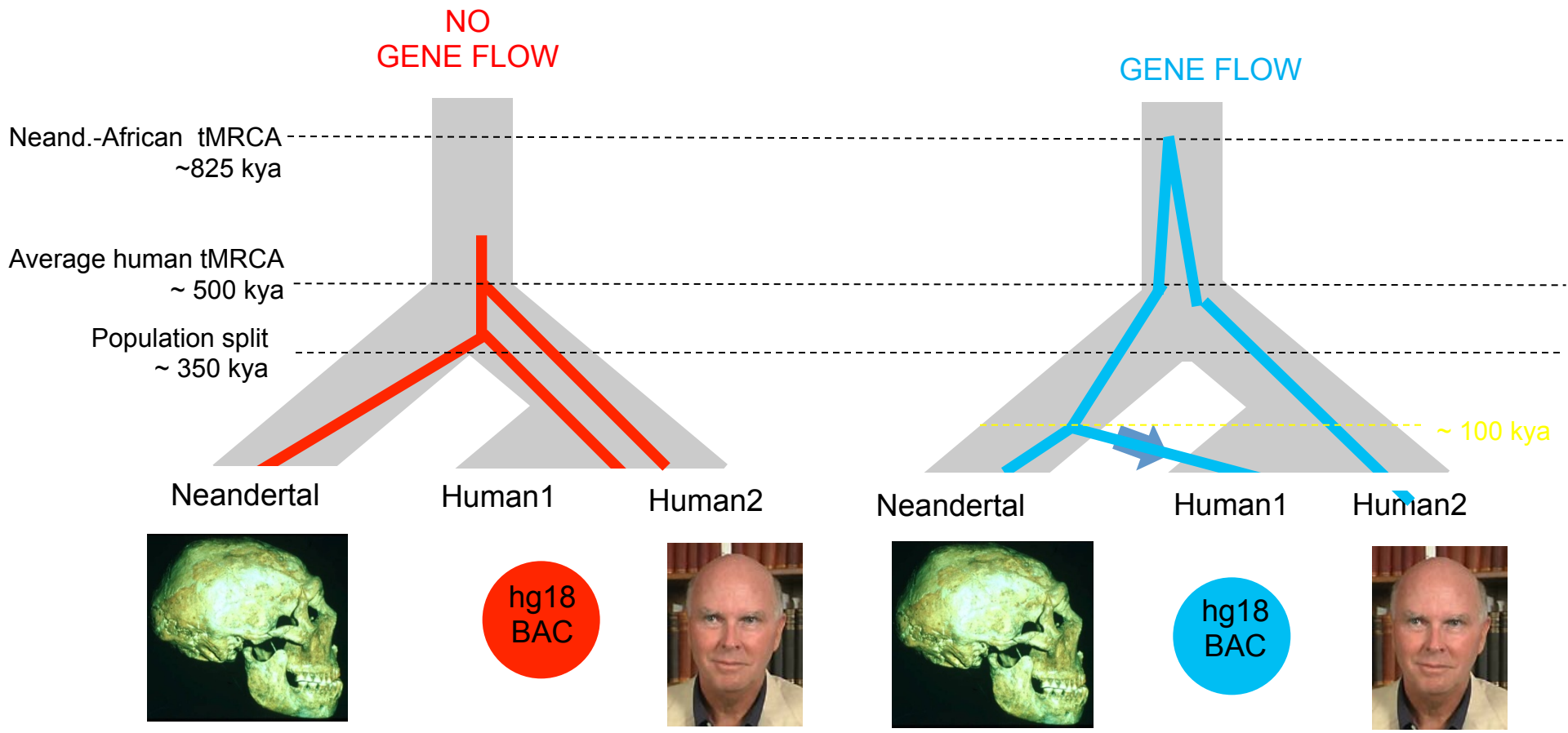
1. Assume **sites j evolve independently**.
   - ➔ Treat each column of the alignment in isolation
2. Assume **branch independence**, conditioned on parent
   - ➔ Expand total joint probability into prod of $P(x_i|x_{parent},t_i)$
   - ➔ Only $P(x_{2n-1})$ remains, root prior, background nucl. freq.
3. We know how to compute **$P(x_i|x_{parent(i)},t_i)$** 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 → **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

Evidence of Neanderthal→Human gene flow



Human-human divergence is

AVERAGE

Human-human divergence is

HIGH

Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Tue 9/26**	PS1 out on:L1-L5	1	Thu, Sep 7	Introduction		L1	Intro: Biology, Algorithms, Machine Learning, Course Overview	1
			Fri, Sep 8			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 12	Module I: Aligning and Modeling Genomes	Foundations	L2	Alignment I: Dynamic Programming, Global and local alignment	2,3
			Thu, Sep 14			L3	Alignment II: Database search, Rapid string matching, BLAST, BLOSUM	3
			Fri, Sep 15			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 19		Frontiers	L4	Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms	7,8
			Thu, Sep 21			L5	Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch	8
	due Tue 9/26		Fri, Sep 22				No classes - student holiday	
			Mon, Sep 25	Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507				
Identify previous project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Tue 10/3**	PS2 out on:L6-R4	4	Tue, Sep 26	Module II: Gene Expression and Epigenomics	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 28			L7	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr	14,15
			Fri, Sep 29			R3	Recitation 3: Affinity Propagation Clustering and Random Forest Classification	
		5	Tue, Oct 3		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 4			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due Tue 10/10		Fri, Oct 6			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Tue 10/19. Presented on Fri 10/20**			Fri, Oct 6	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
	PS3 out on:L10-R6	6	Tue, Oct 10	Module III: Regulatory Genomics and Networks	Foundations		No Classes - Columbus Day Holiday	
			Thu, Oct 12			L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Fri, Oct 13			R5	Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex	
		7	Tue, Oct 17			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Thu, Oct 19		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due Tue 10/24		Fri, Oct 20			R6	Recitation 6: Networks review, Recommendation systems, EHR, PheWAS	
			Fri, Oct 20	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507 at 4-5pm				
Evaluate/discuss three peer proposals, NIH review format. **Review Panels Fri 10/27** **Reviews back Tue 10/31**	PS4 out on:L13-R8	8	Tue, Oct 24	Module IV: Population Genetics and Disease Genomics	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 26			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 27			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 27	Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507				
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/9**		9	Tue, Oct 31		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
			Thu, Nov 2			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
	due Tue 11/7		Fri, Nov 3			R8	Recitation 8: Rare Variants, ExAC	
	PS5 out on:L17-R10	10	Tue, Nov 7	Module V: Comparative Genomics and Evolution	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 8			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 10				No Recitation, Veterans Day	
Continue making substantial progress on proposed milstones.Write outline of final report. **Midcourse report due Wed 11/22.**		11	Tue, Nov 14		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
			Thu, Nov 16			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
	due Fri 11/17		Fri, Nov 17			R9	Recitation 9: Phylogenetic distance metrics, Coalescent Process	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/10**	No more psets! (work on your final project)	12	Tue, Nov 21	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 23				No lecture, thanksgiving break - Thu Nov 26, 2015	
			Fri, Nov 24				No recitation, thanksgiving break	
		13	Tue, Nov 28	Module VI: Current Research Directions	Frontiers	L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Thu, Nov 30			L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Fri, Dec 1			R10	Recitation 10: Project Feedback, results, interpretation, directions	
		14	Tue, Dec 5			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 7			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 8			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Talks on Tue 12/12**		15	Tue, Dec 12			L25	Final Presentations - Part I (11am). 32-G8 reading room	
			Tue, Dec 12			L25	Final Presentations - Part I (1pm). 32-141	

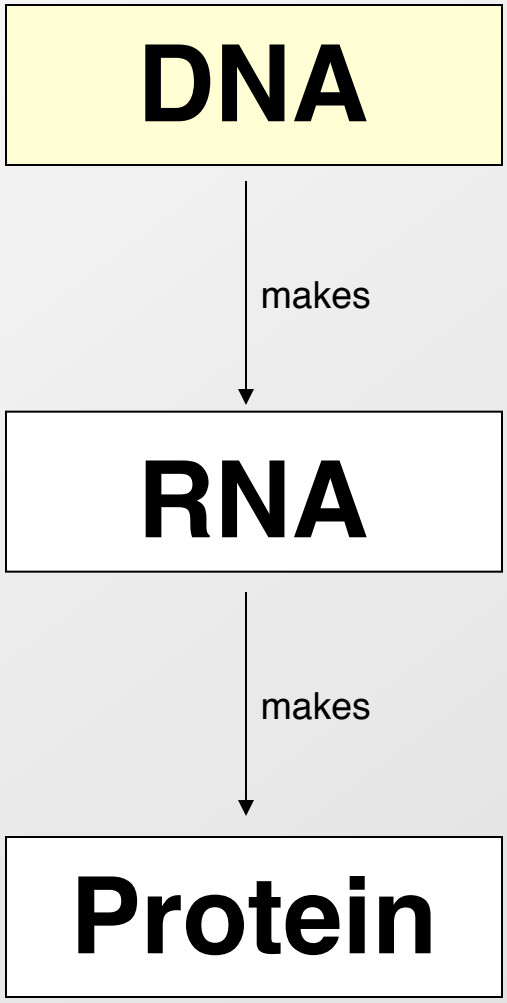
* readings refer to chapters in compiled 2016 scribe notes, available in the materials folder on Stellar

** recitation topics will be adjusted to respond to lecture and student needs

Biology primer

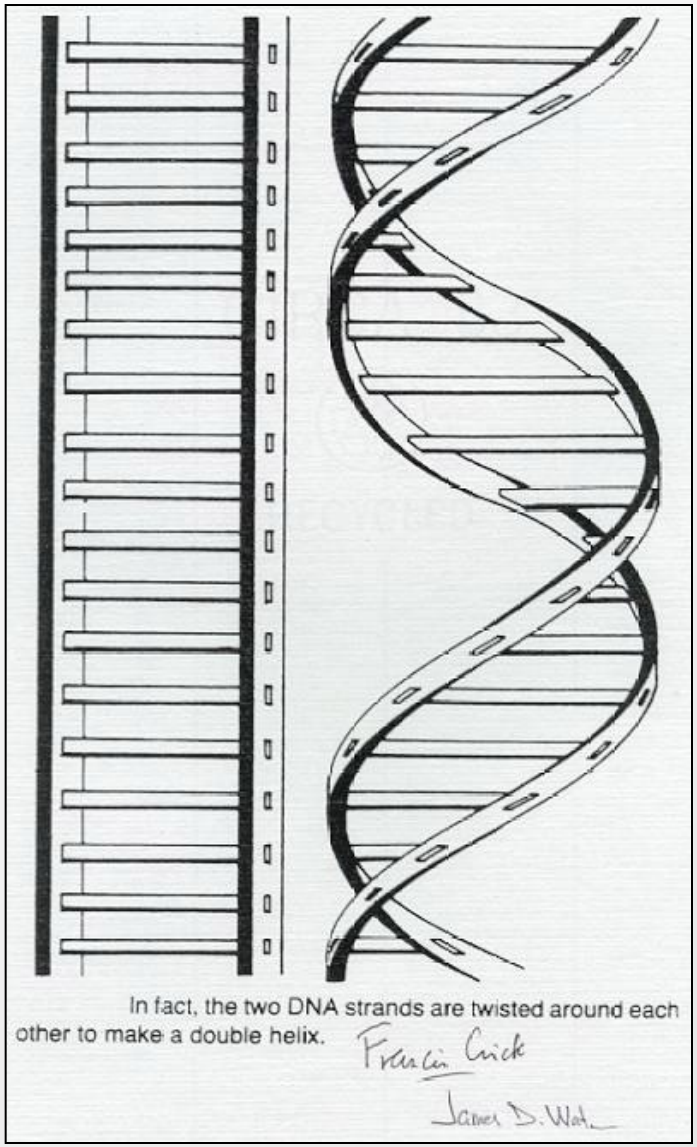
Quick introduction to molecular biology
and information transfer within the cell

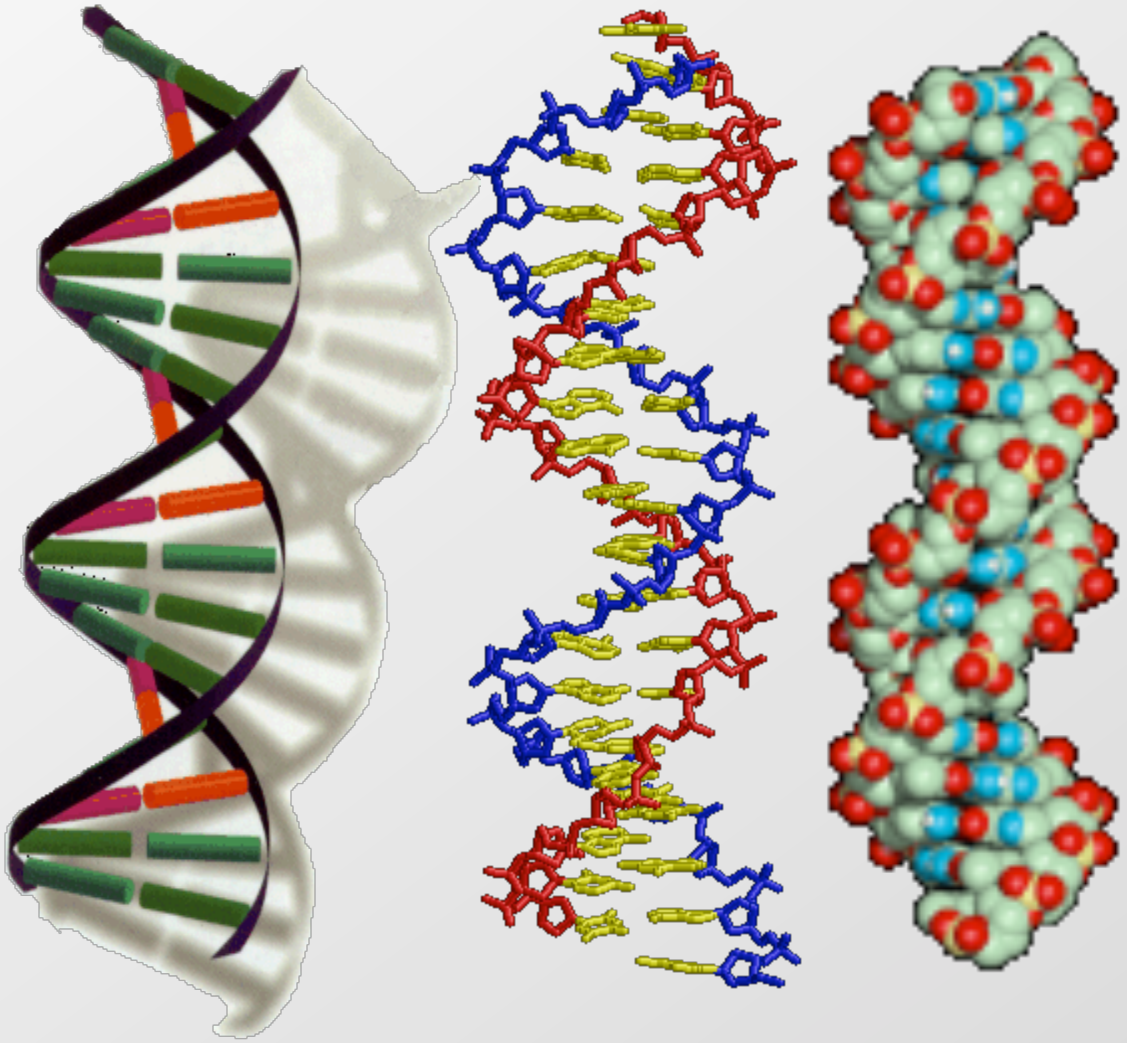
“Central dogma” of Molecular Biology



DNA: The double helix

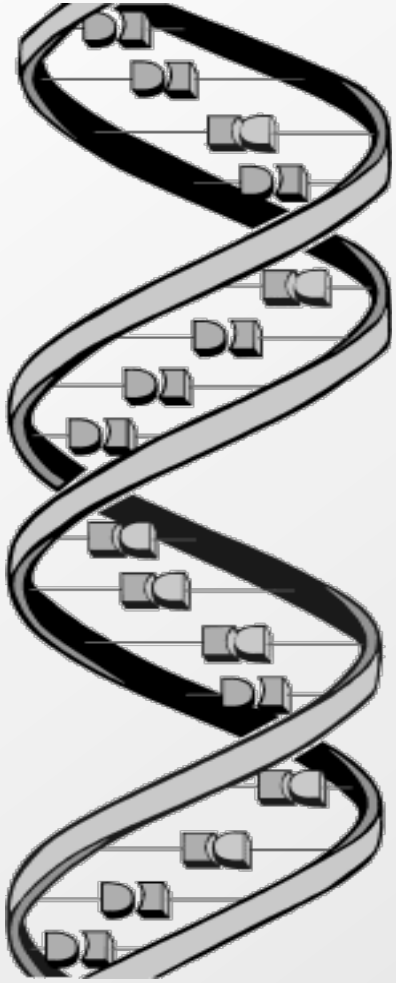
- The most noble molecule of our time

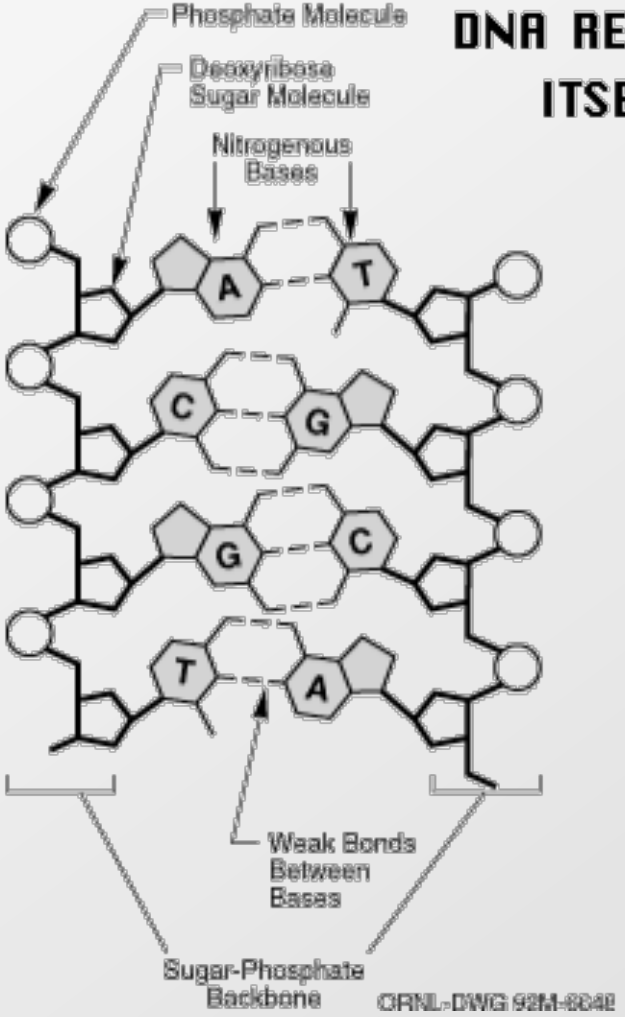


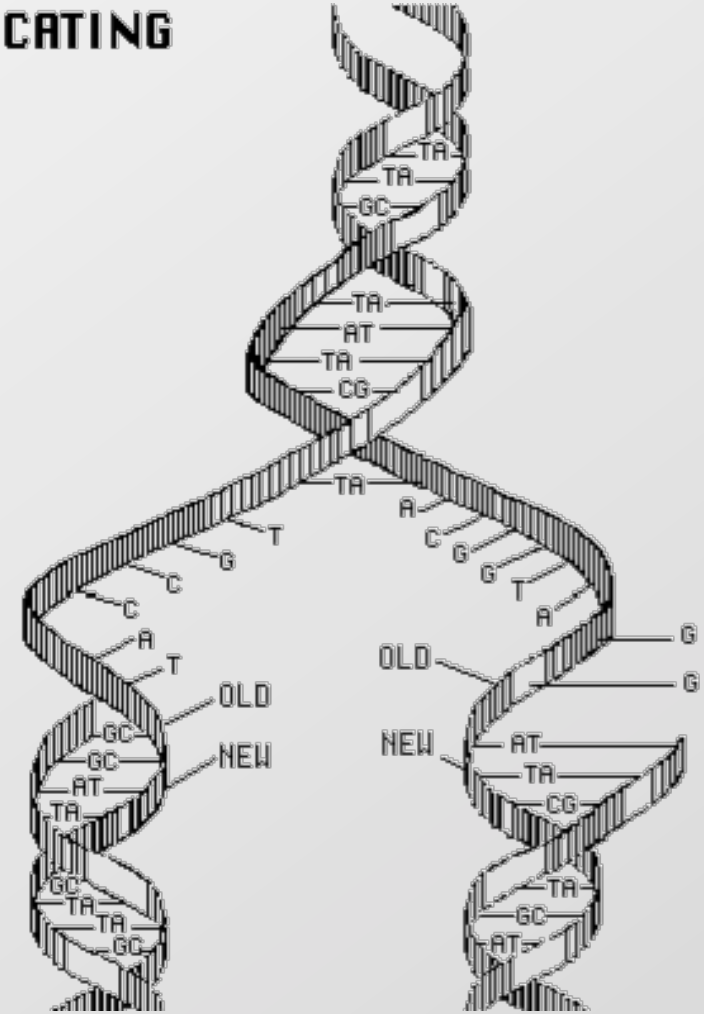


DNA: the molecule of heredity

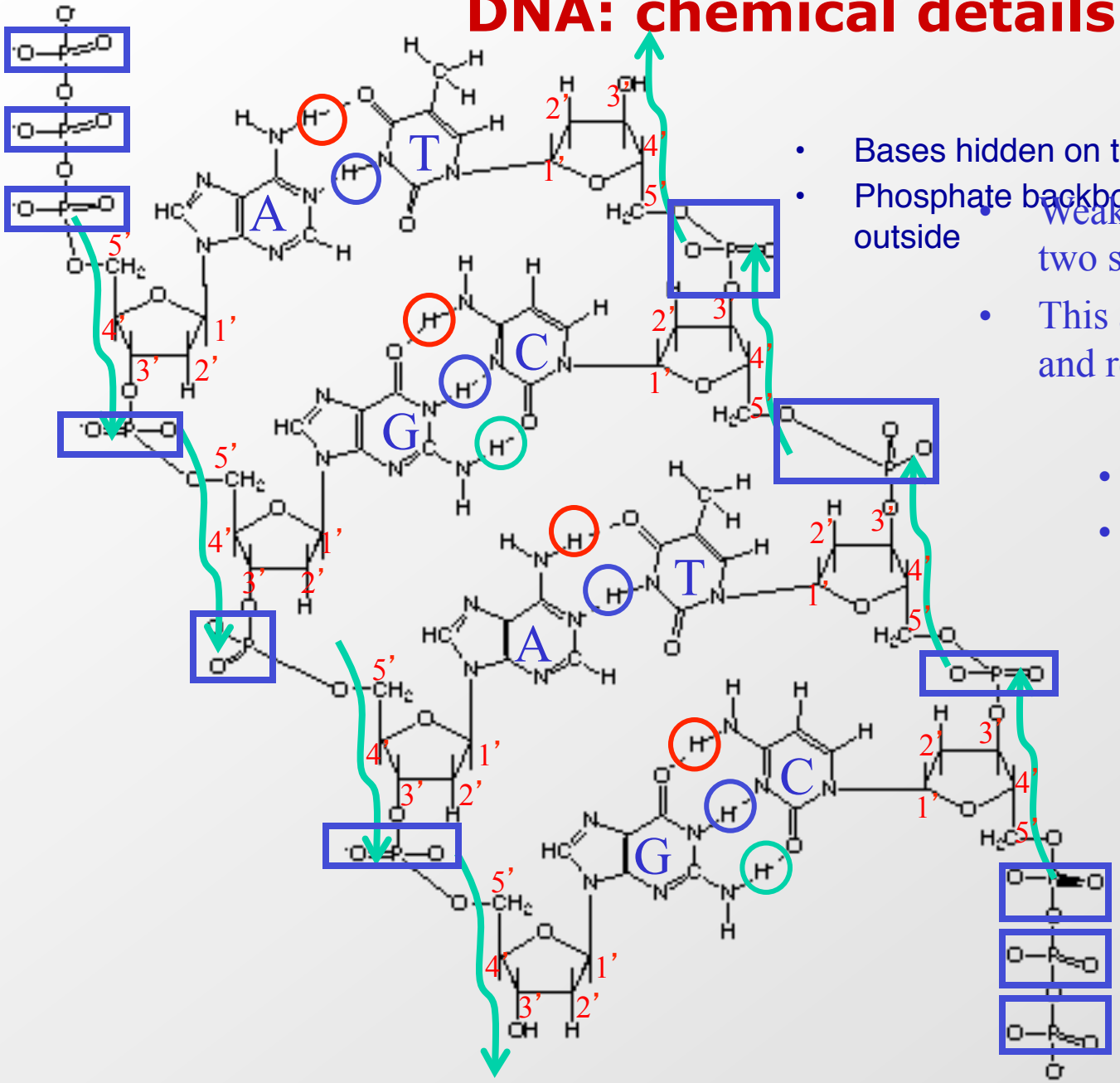
- Self-complementarity sets molecular basis of heredity
  - Knowing one strand, creates a template for the other
  - “It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.” Watson & Crick, 1953







DNA: chemical details



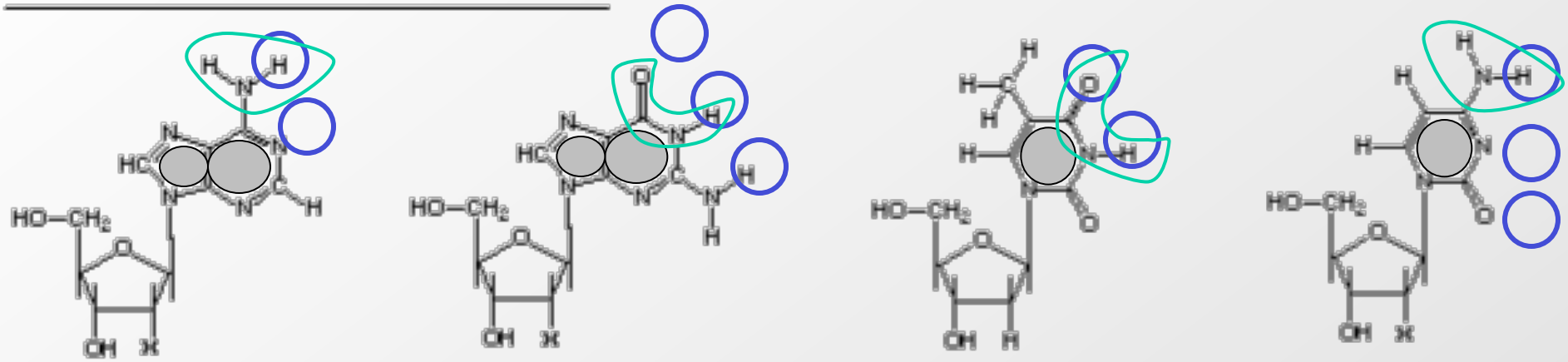
- Bases hidden on the inside
- Phosphate backbone outside
- Weak hydrogen bonds hold the two strands together
- This allows low-energy opening and re-closing of two strands
- Anti-parallel strands
- Extension 5’→3’ tri-phosphate coming from newly added nucleotide

The only parings are:

- A with T
- C with G

DNA: the four bases

The Nucleotides of DNA

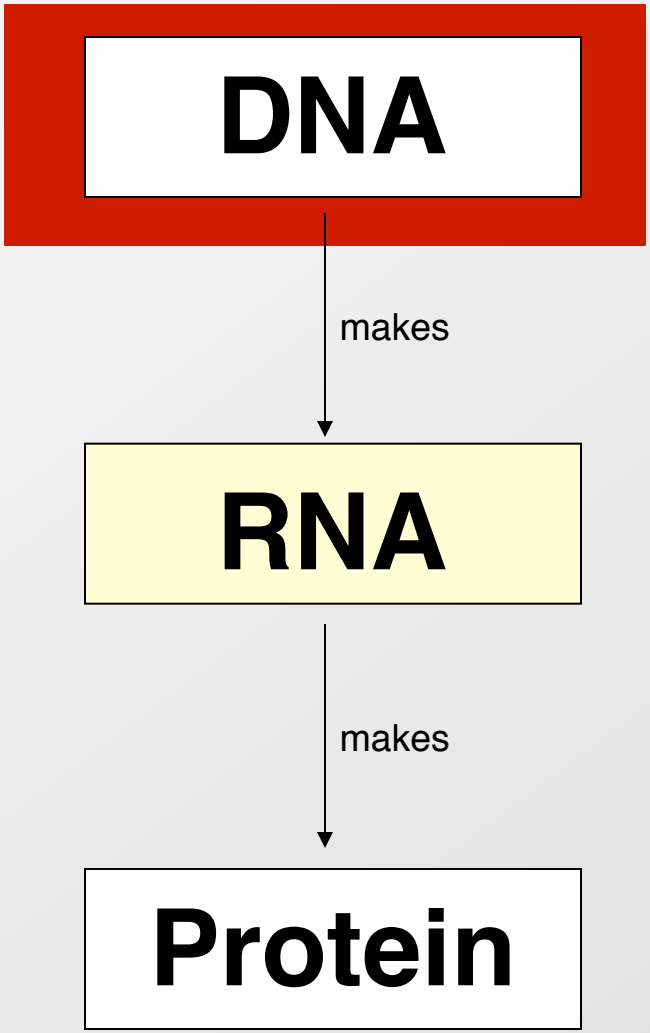


Adenine | Guanosine | Thymine | Cytosine

Purine	Purine		
		Pyrimidine	Pyrimidine
Weak		Weak	
	Strong		Strong
Amino			Amino
	Keto	Keto	

Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Mon 9/23**	PS1 out on:L1-L5	1	Thu, Sep 5	Introduction		L1	Algorithms, Machine Learning, Networks, Course Overview	1
			Fri, Sep 6			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 10	Module I:	Foundations	L2	Dynamic Programming, Reusing computation, Iterative Functions, Exponential / Poly	2,3
			Thu, Sep 12			L3	Database search, Rapid string matching, Hashing	3
			Fri, Sep 13			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 17		Frontiers	L4	HMMs1: Evaluation, Parsing, posterior decoding, learning, HMM architectures	7,8
	due		Thu, Sep 19			L5	HMMs2: Applications, architectures, memory, gene finding, chromatin states	7,8
	Mon 9/23		Fri, Sep 20				No Classes - Student Holiday	
Find prev project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Mon 10/7**	PS2 out on:L6-R4	4	Tue, Sep 24	Module II:	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 26			L7	RNA structure and function. RNA world, RNA-seq, transcript structure, RNA folding	14,15
			Fri, Sep 27			R3	Recitation 3: Supervised Learning and Random Forest Classification	
			Fri, Sep 27	cts, self introductions, mentor intro, example projects, teamwork 32D-507				
		5	Tue, Oct 1		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 3			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due		Fri, Oct 4			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
	Mon 10/7		Fri, Oct 4	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Thu 10/17. Presented on Fri 10/18**	PS3 out on:L10-R6	6	Tue, Oct 8	Module III:	Foundations	L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Thu, Oct 10			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Fri, Oct 11			R5	Recitation 5: Communication Lab	
		7	Tue, Oct 15				No Classes - Columbus Day Holiday	
			Thu, Oct 17		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due		Fri, Oct 18			R6	Recitation 6: Motif Discovery, WEEDER, In vitro Motif Discovery - PBMs, Selex	
	Mon 10/21		Fri, Oct 18	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507				
Evaluate/discuss three peer proposals, NIH review format. **Reviews back Mon 10/28**	PS4 out on:L13-R8	8	Tue, Oct 22	Module IV:	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 24			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 25			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 25	Panel Review: Discuss Peer Projects. Feedback sent out from group reviews. 32D-463 (Star).				
		9	Tue, Oct 29		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/7**	**due**		Thu, Oct 31			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
	Mon 11/4		Fri, Nov 1			R8	Recitation 8: Phylogenetic distance metrics, Coalescent Process	
	PS5 out on:L17-R10	10	Tue, Nov 5	Module V:	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 7			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 8				No Recitation, Veterans Day	
Continue making subst. progress on proposed milstones.Write outline of final report. **Midcourse report due Mon 11/25**		11	Tue, Nov 12		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
	due		Thu, Nov 14			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
	Fri 11/15		Fri, Nov 15			R9	Recitation 9: Quiz Review	
	No more psets! (work on your final project)	12	Tue, Nov 19	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 21			L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Fri, Nov 22			R10	Recitation 10: Project Feedback, results, interpretation, directions	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/8**		13	Tue, Nov 26	Module VI:	Frontiers	L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Thu, Nov 28				No lecture, thanksgiving break - Thu Nov 28, 2019	
			Fri, Nov 29				No recitation, thanksgiving break	
		14	Tue, Dec 3			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 5			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 6			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Presentations on Tue 12/10**		15	Tue, Dec 10			L25	Final Presentations - Part I (1pm). 32-141 (Classroom)	
			Tue, Dec 10			L25	Final Presentations - Part I (2:30pm). 32D-463 (Star)	
				* readings refer to chapters in compiled 2018 scribe notes, available in the materials folder on Stellar				
				** recitation topics will be adjusted to respond to lecture and student needs				

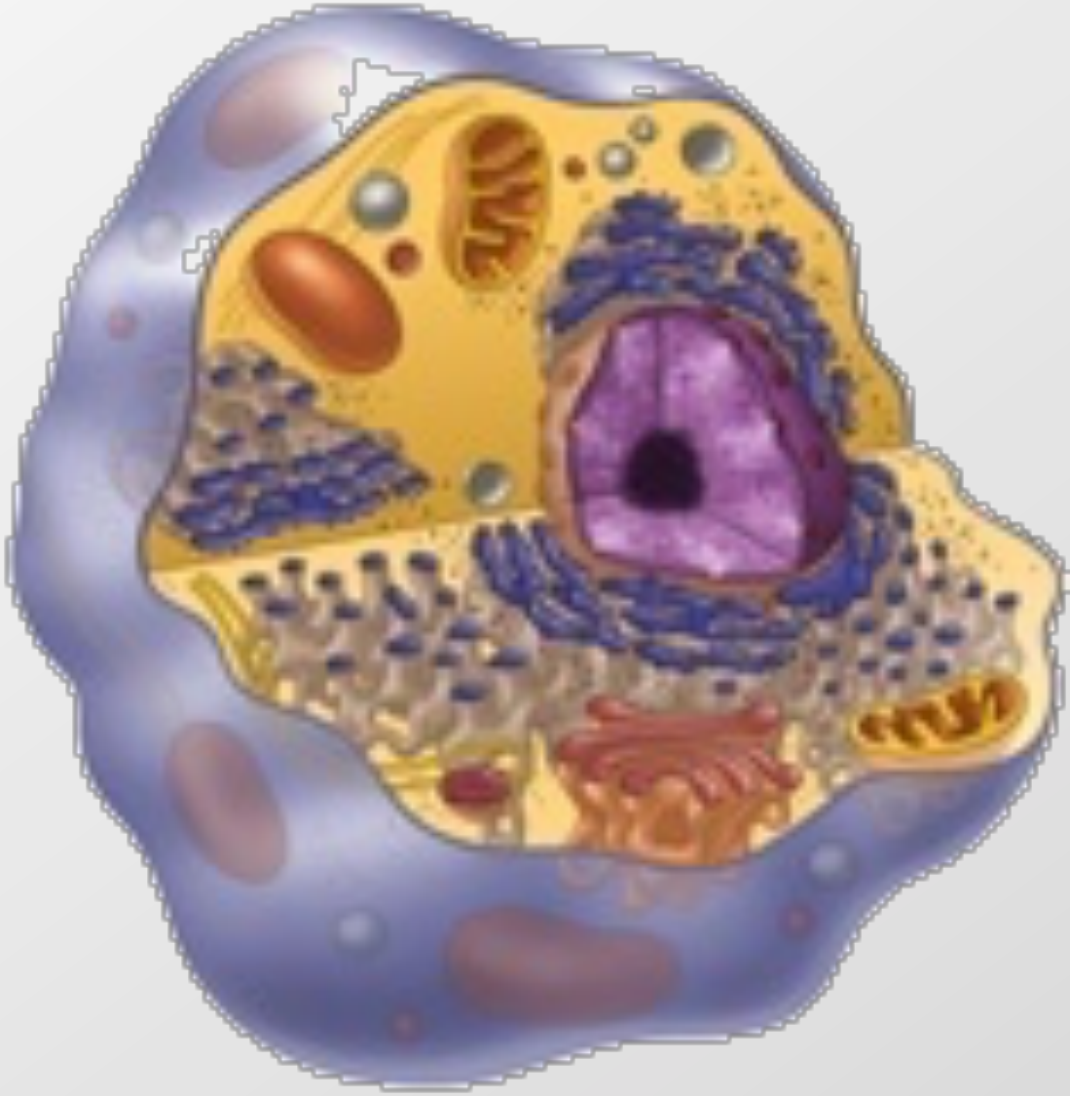
“Central dogma” of Molecular Biology



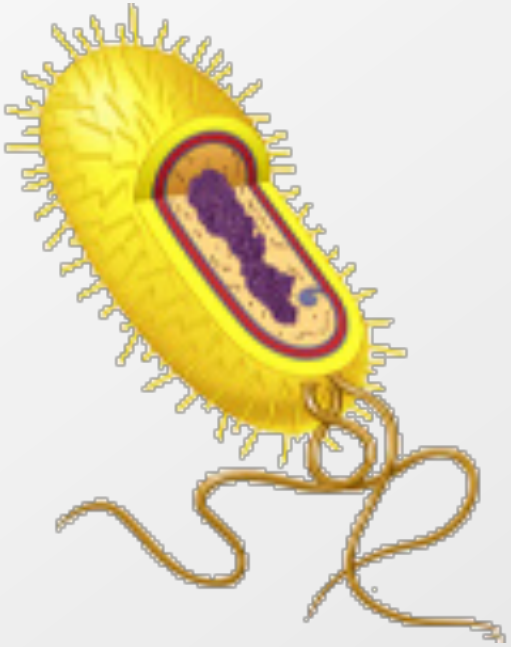
Epigenomics

Chromosomes inside the cell

- Eukaryote cell

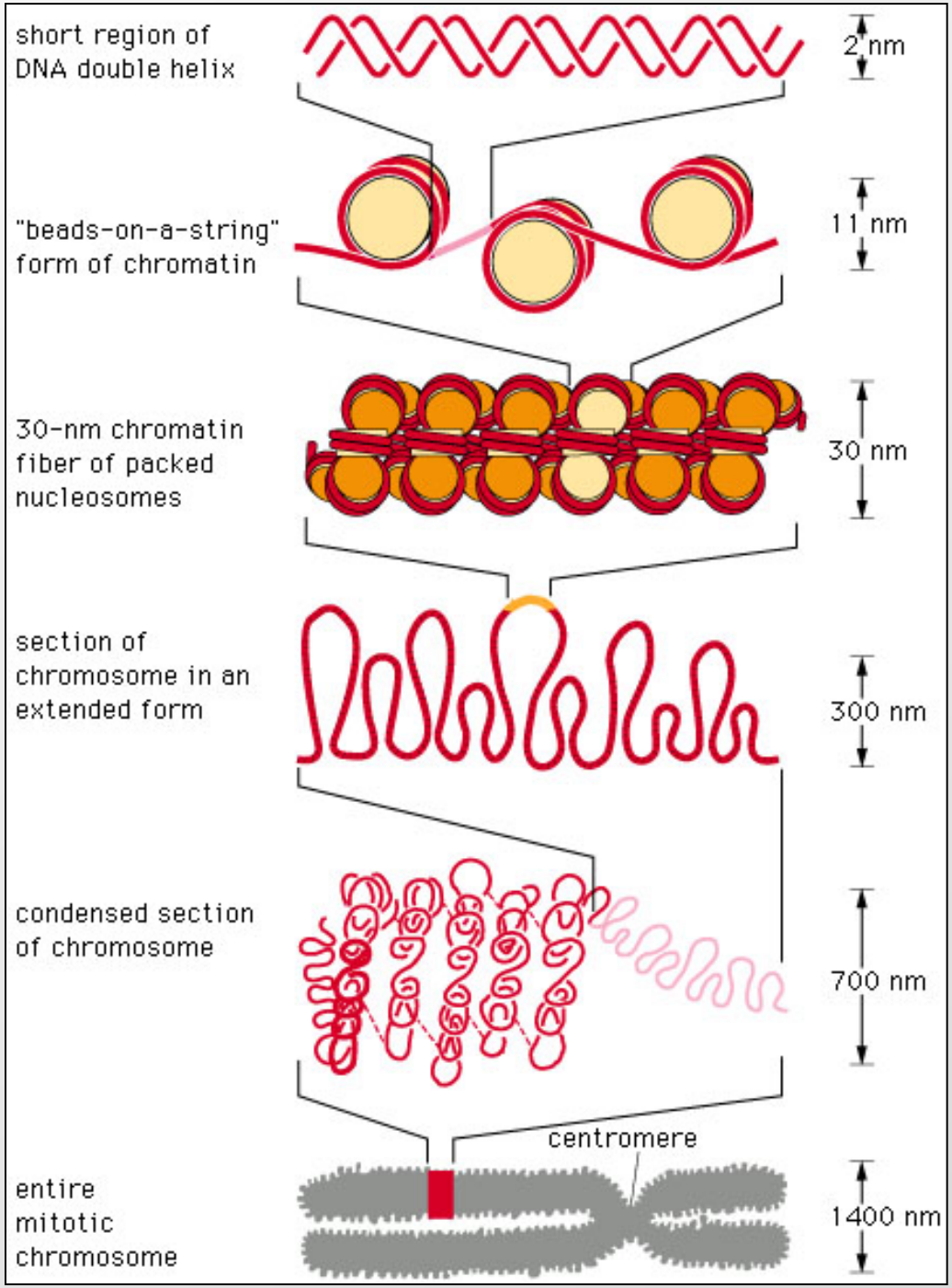


- Prokaryote cell



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

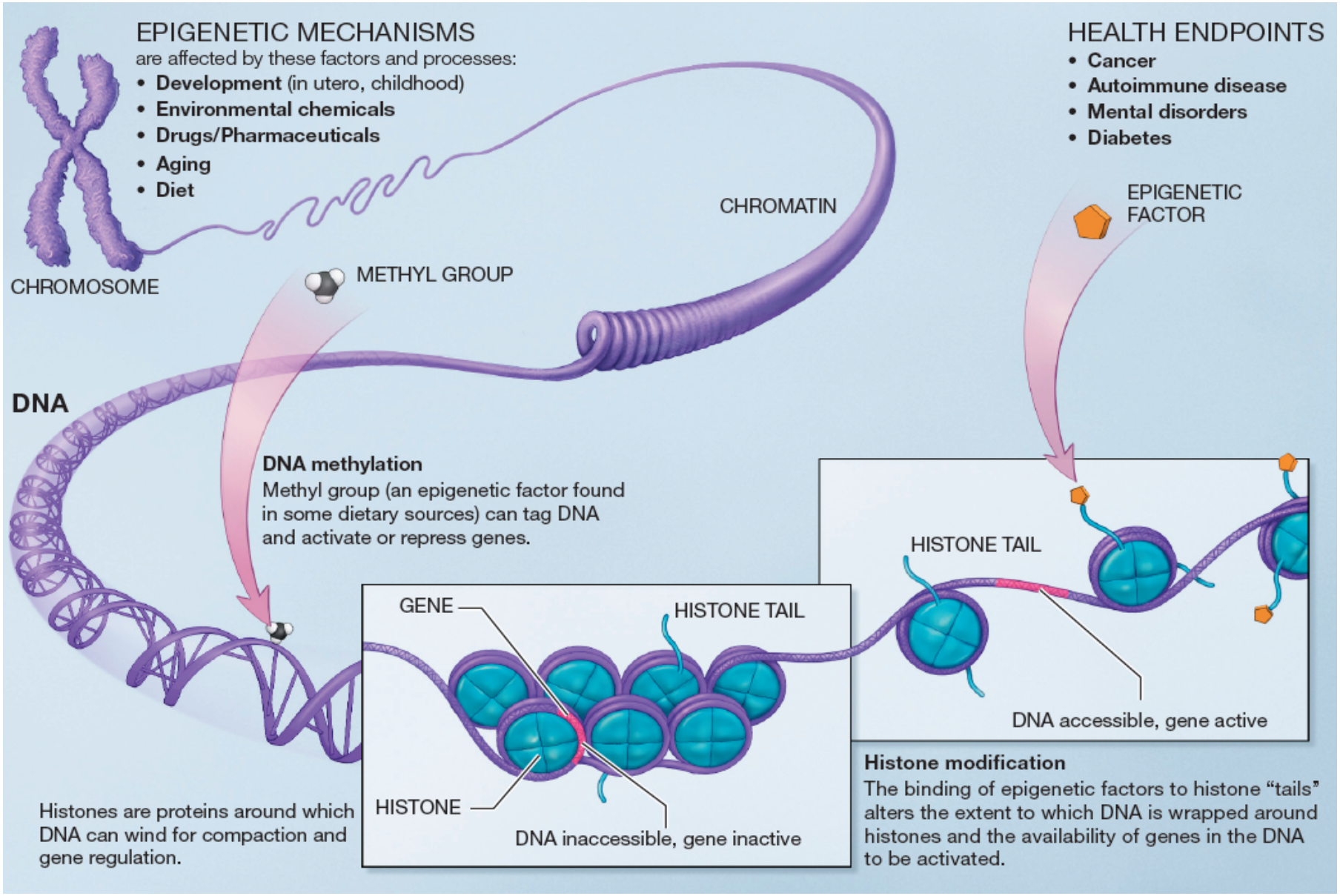
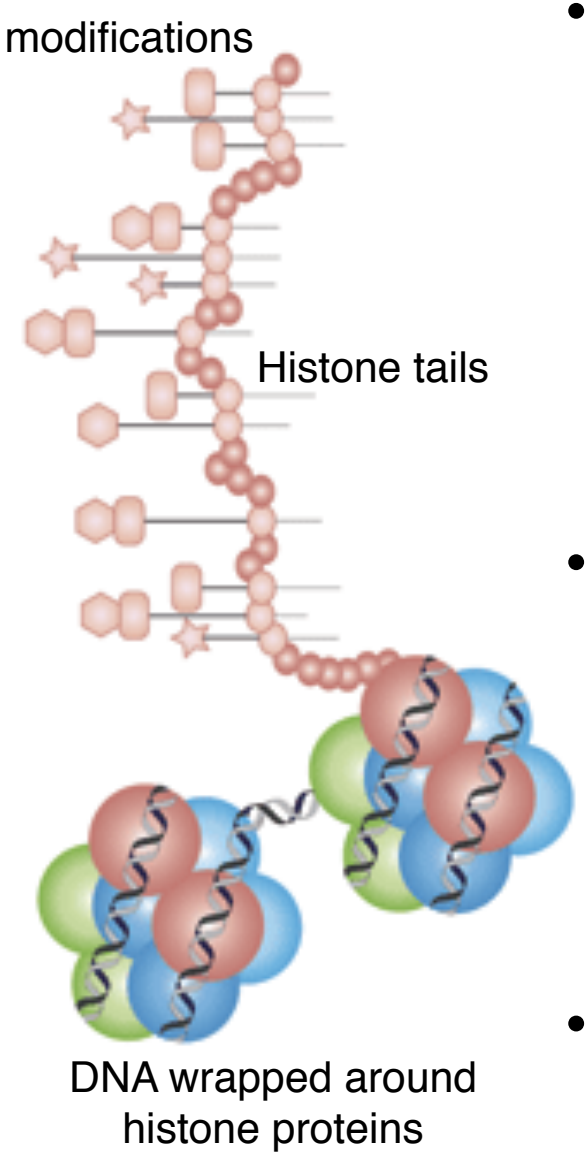


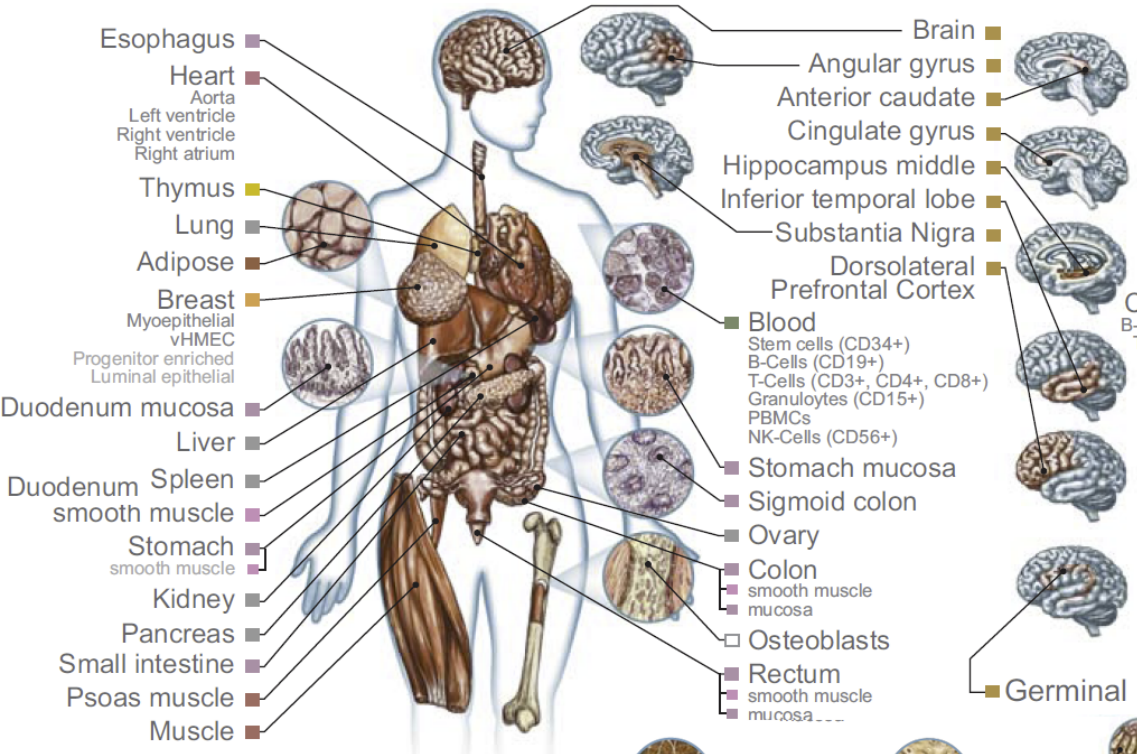
Image source: http://nihroadmap.nih.gov/epigenomics/

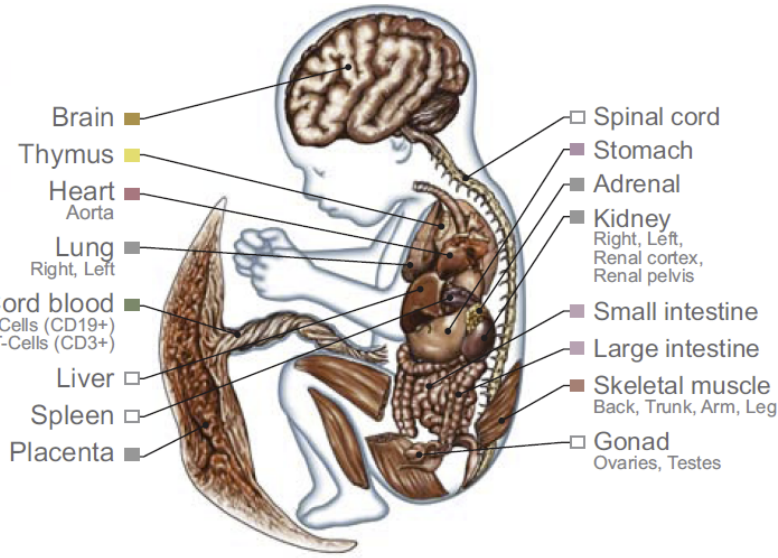
Diversity of epigenetic modifications



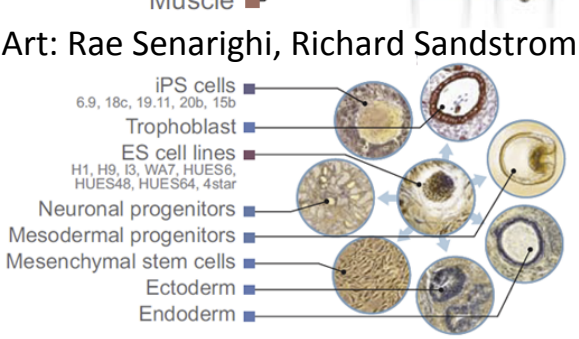
- 100+ different histone modifications
  - Histone protein → H3/H4/H2A/H2B
  - AA residue → Lysine4(K4)/K36…
  - Chemical modification → Met/Pho/Ubi
  - Number → Me-Me-Me(me3)
  - Shorthand: H3K4me3, H2BK5ac
- In addition:
  - DNA modifications
  - Methyl-C in CpG / Methyl-Adenosine
  - Nucleosome positioning
  - DNA accessibility
- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

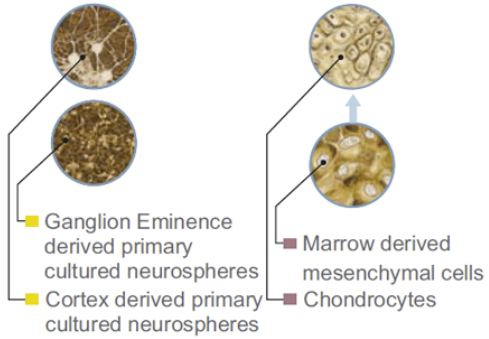
Epigenomics Roadmap across 100+ tissues/cell types

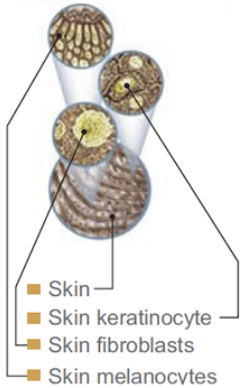




Art: Rae Senarighi, Richard Sandstrom







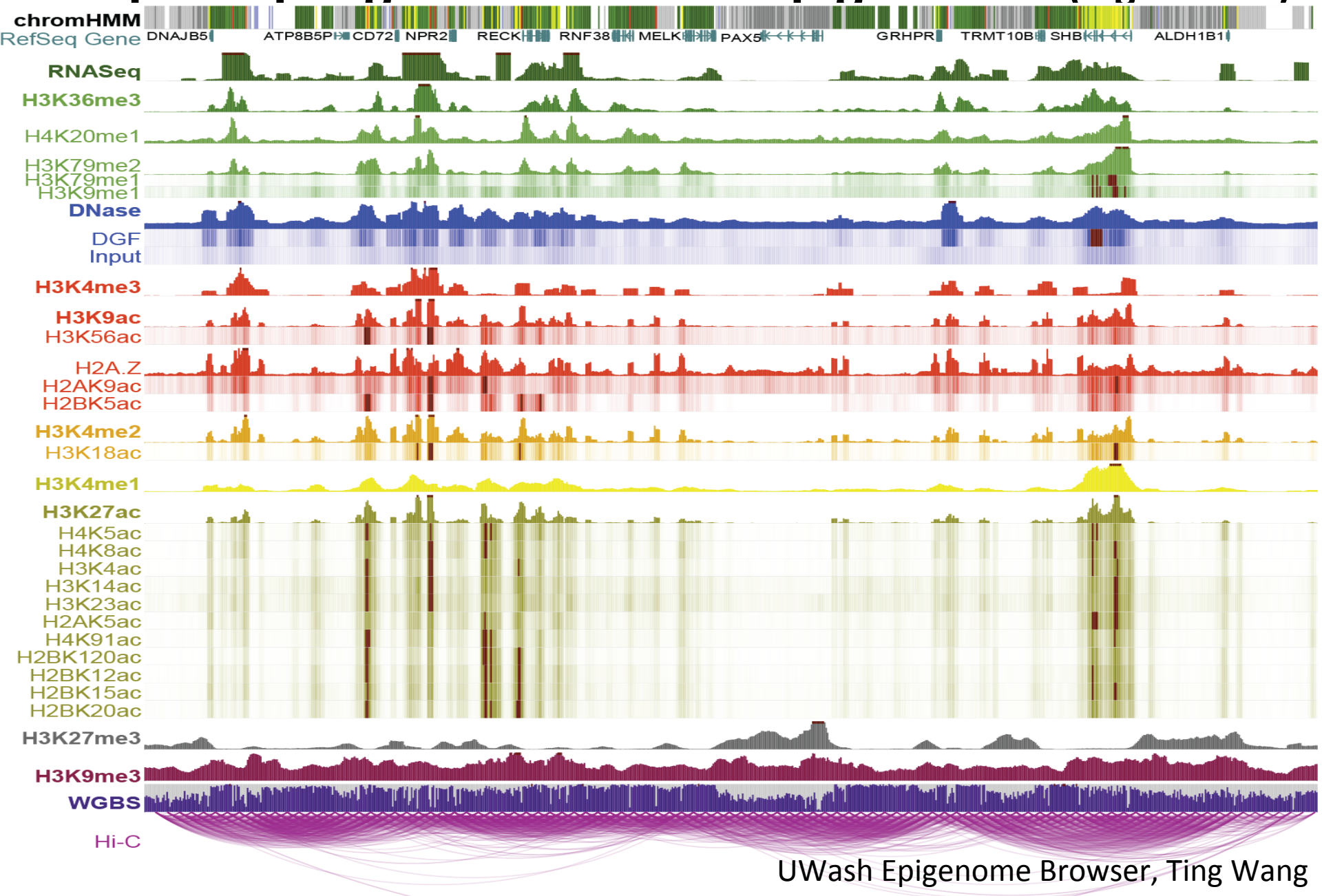
Diverse epigenomic assays:

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**

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)

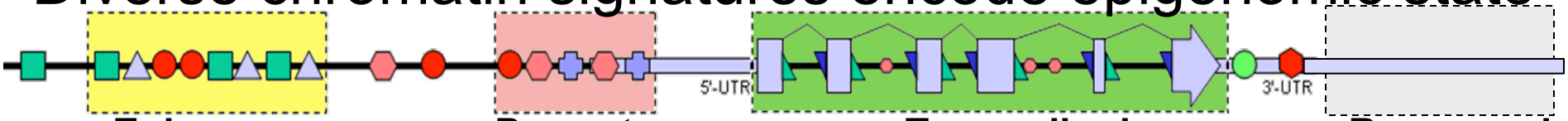
Deep sampling of 9 reference epigenomes (e.g. IMR90)



UWash Epigenome Browser, Ting Wang

Chromatin state+RNA+DNAse+28 histone marks+WGBS+Hi-

Diverse chromatin signatures encode epigenomic state

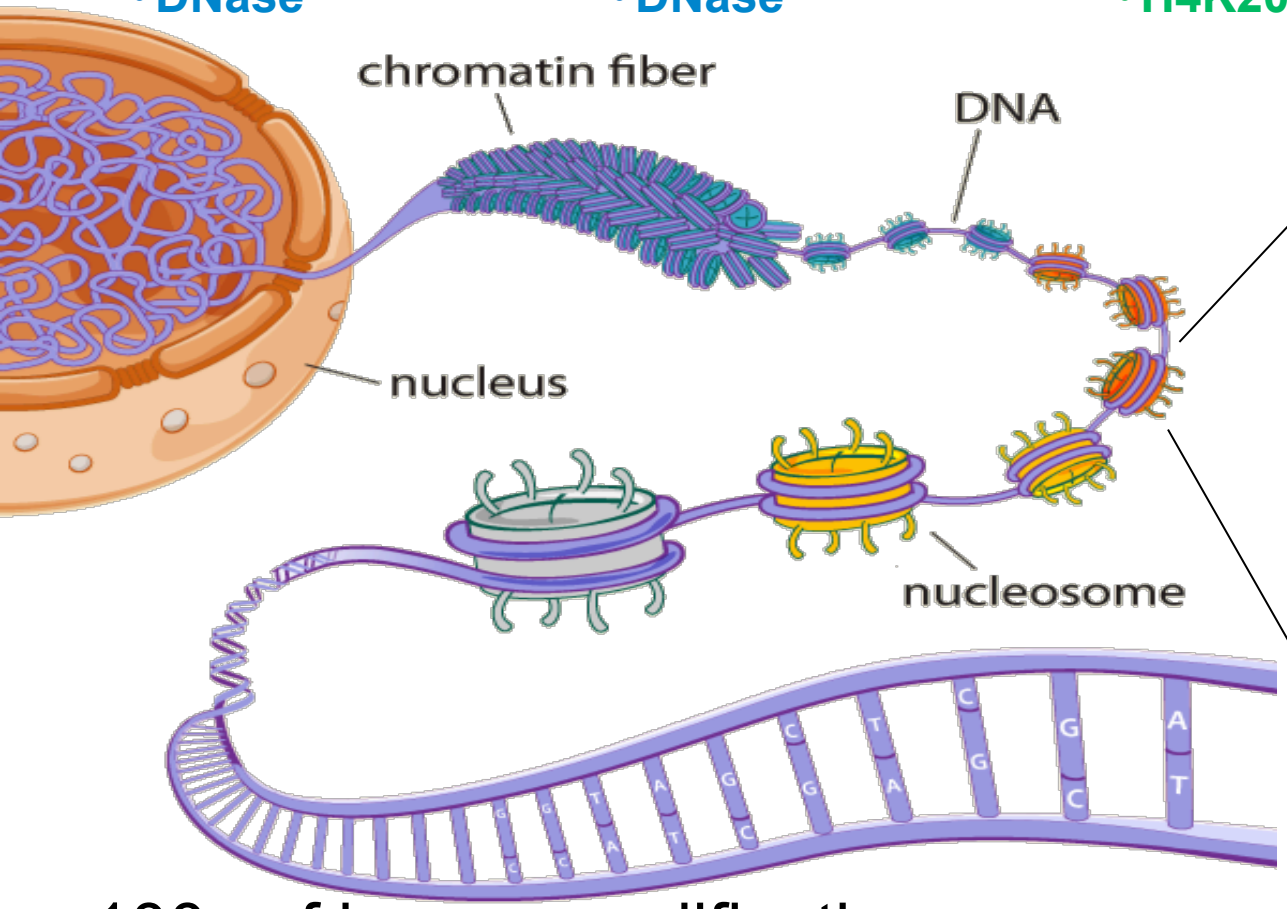


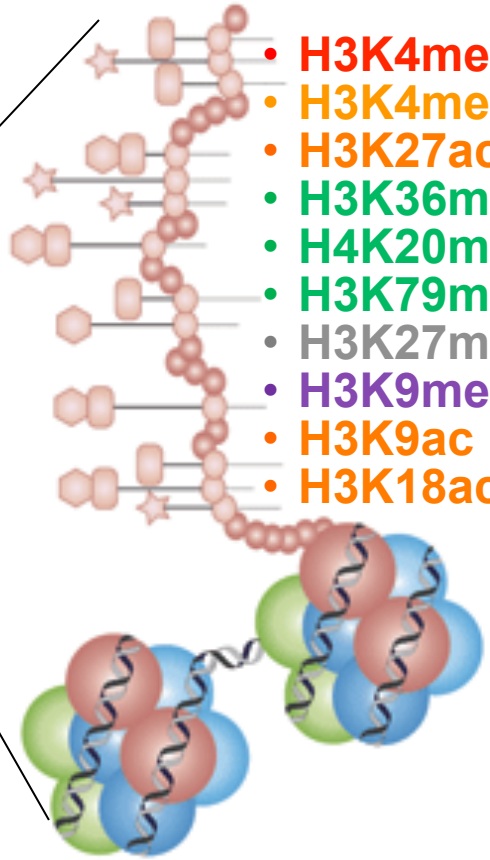
Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl





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

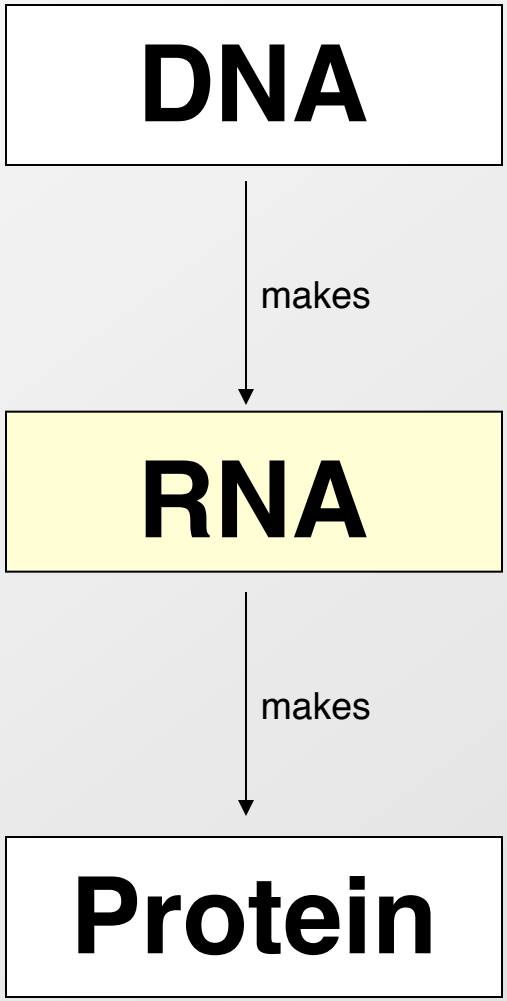
Chromatin state annotations across 127 epigenomes

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

Anshul Kundaje

Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Mon 9/23**	PS1 out on:L1-L5	1	Thu, Sep 5	Introduction		L1	Algorithms, Machine Learning, Networks, Course Overview	1
			Fri, Sep 6			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 10	Module I:	Foundations	L2	Dynamic Programming, Reusing computation, Iterative Functions, Exponential / Poly	2,3
			Thu, Sep 12			L3	Database search, Rapid string matching, Hashing	3
			Fri, Sep 13			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 17		Frontiers	L4	HMMs1: Evaluation, Parsing, posterior decoding, learning, HMM architectures	7,8
	due		Thu, Sep 19			L5	HMMs2: Applications, architectures, memory, gene finding, chromatin states	7,8
	Mon 9/23		Fri, Sep 20				No Classes - Student Holiday	
Find prev project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Mon 10/7**	PS2 out on:L6-R4	4	Tue, Sep 24	Module II:	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 26			L7	RNA structure and function. RNA world, RNA-seq, transcript structure, RNA folding	14,15
			Fri, Sep 27			R3	Recitation 3: Supervised Learning and Random Forest Classification	
			Fri, Sep 27	cts, self introductions, mentor intro, example projects, teamwork 32D-507				
		5	Tue, Oct 1		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 3			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due		Fri, Oct 4			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
	Mon 10/7		Fri, Oct 4	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Thu 10/17. Presented on Fri 10/18**	PS3 out on:L10-R6	6	Tue, Oct 8	Module III:	Foundations	L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Thu, Oct 10			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Fri, Oct 11			R5	Recitation 5: Communication Lab	
		7	Tue, Oct 15				No Classes - Columbus Day Holiday	
			Thu, Oct 17		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due		Fri, Oct 18			R6	Recitation 6: Motif Discovery, WEEDER, In vitro Motif Discovery - PBMs, Selex	
	Mon 10/21		Fri, Oct 18	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507				
Evaluate/discuss three peer proposals, NIH review format. **Reviews back Mon 10/28**	PS4 out on:L13-R8	8	Tue, Oct 22	Module IV:	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 24			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 25			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 25	Panel Review: Discuss Peer Projects. Feedback sent out from group reviews. 32D-463 (Star).				
		9	Tue, Oct 29		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/7**	due		Thu, Oct 31			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
	Mon 11/4		Fri, Nov 1			R8	Recitation 8: Phylogenetic distance metrics, Coalescent Process	
	PS5 out on:L17-R10	10	Tue, Nov 5	Module V:	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 7			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 8				No Recitation, Veterans Day	
Continue making subst. progress on proposed milstones.Write outline of final report. **Midcourse report due Mon 11/25**	due	11	Tue, Nov 12		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
	Fri 11/15		Thu, Nov 14			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
			Fri, Nov 15			R9	Recitation 9: Quiz Review	
	No more psets! (work on your final project)	12	Tue, Nov 19	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 21			L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Fri, Nov 22			R10	Recitation 10: Project Feedback, results, interpretation, directions	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/8**		13	Tue, Nov 26	Module VI:	Frontiers	L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Thu, Nov 28				No lecture, thanksgiving break - Thu Nov 28, 2019	
			Fri, Nov 29				No recitation, thanksgiving break	
		14	Tue, Dec 3			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 5			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 6			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Presentations on Tue 12/10**		15	Tue, Dec 10			L25	Final Presentations - Part I (1pm). 32-141 (Classroom)	
			Tue, Dec 10			L25	Final Presentations - Part I (2:30pm). 32D-463 (Star)	

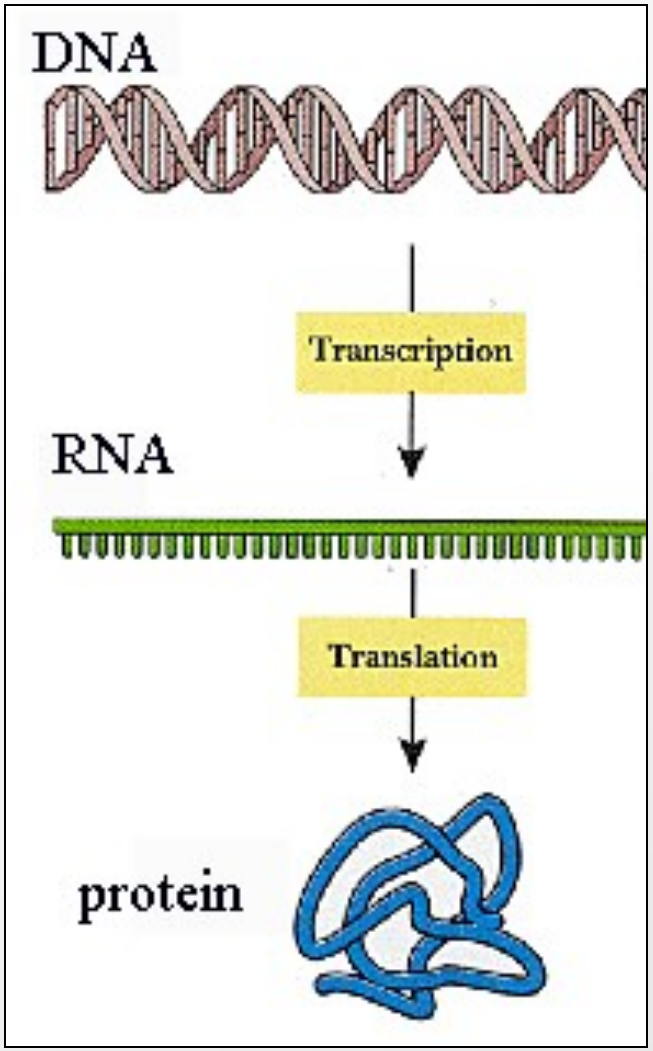
“Central dogma” of Molecular Biology



Genes control the making of cell parts

- The gene is a fundamental unit of inheritance
  - Each DNA molecule ⇔ 10,000+ genes
  - 1 gene ⇔ 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

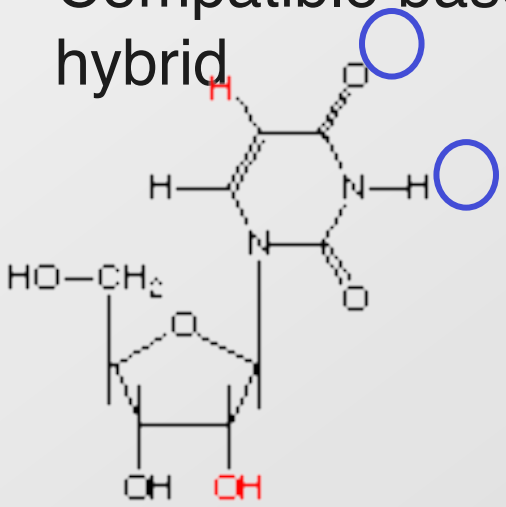
mRNA: The messenger



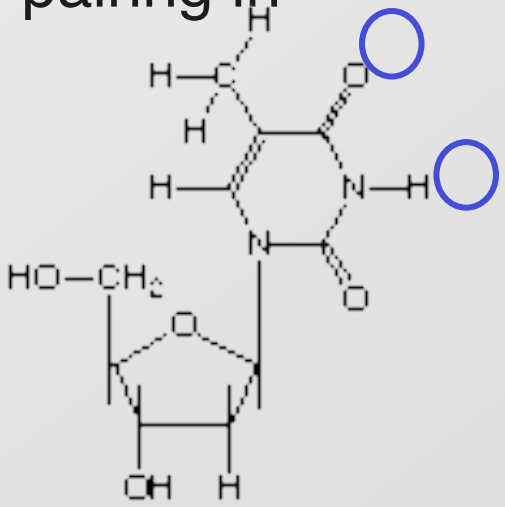
- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar

```
A T T A C G G T A C C G T
| | | | | | | | | | | |
U A A U G C C A U G G C A
```

  - Compatible base-pairing in hybrid

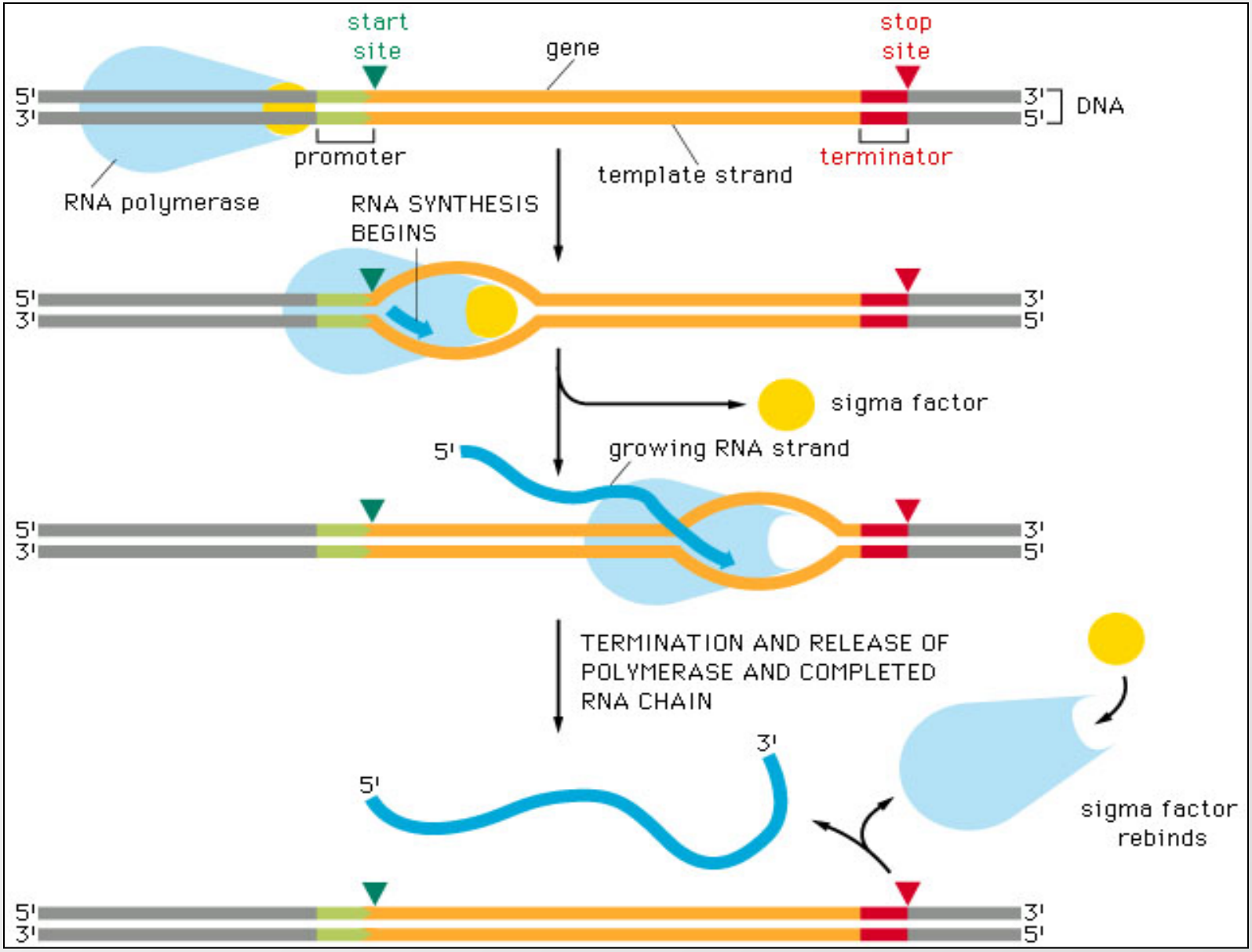


uracil (RNA)



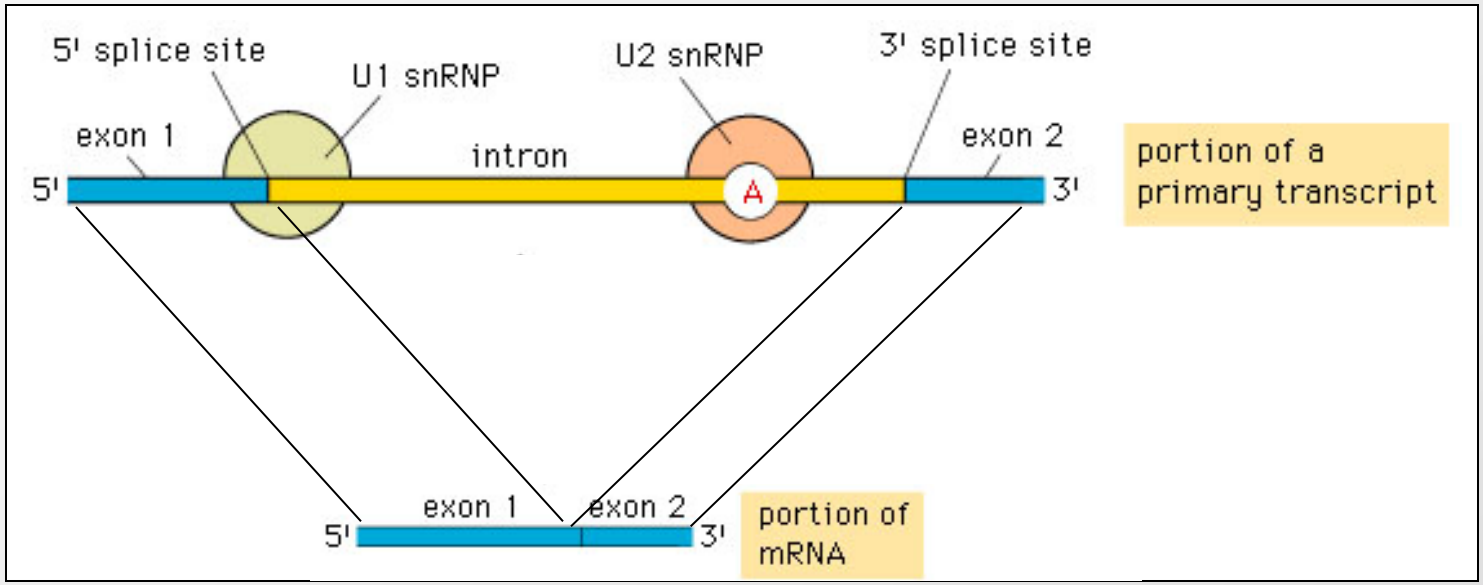
thymine (DNA)

From DNA to RNA: Transcription



From pre-mRNA to mRNA: Splicing

- 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

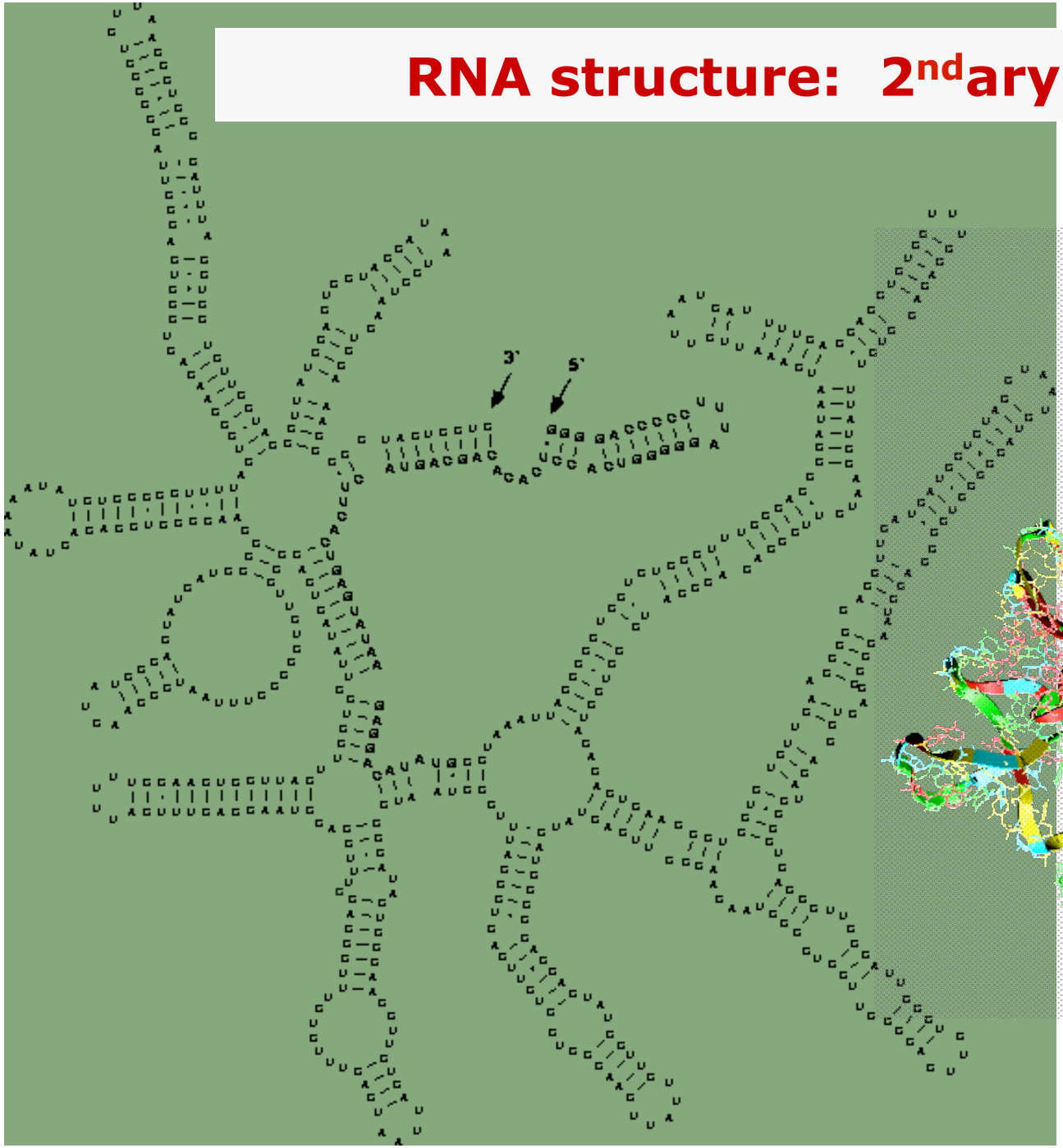


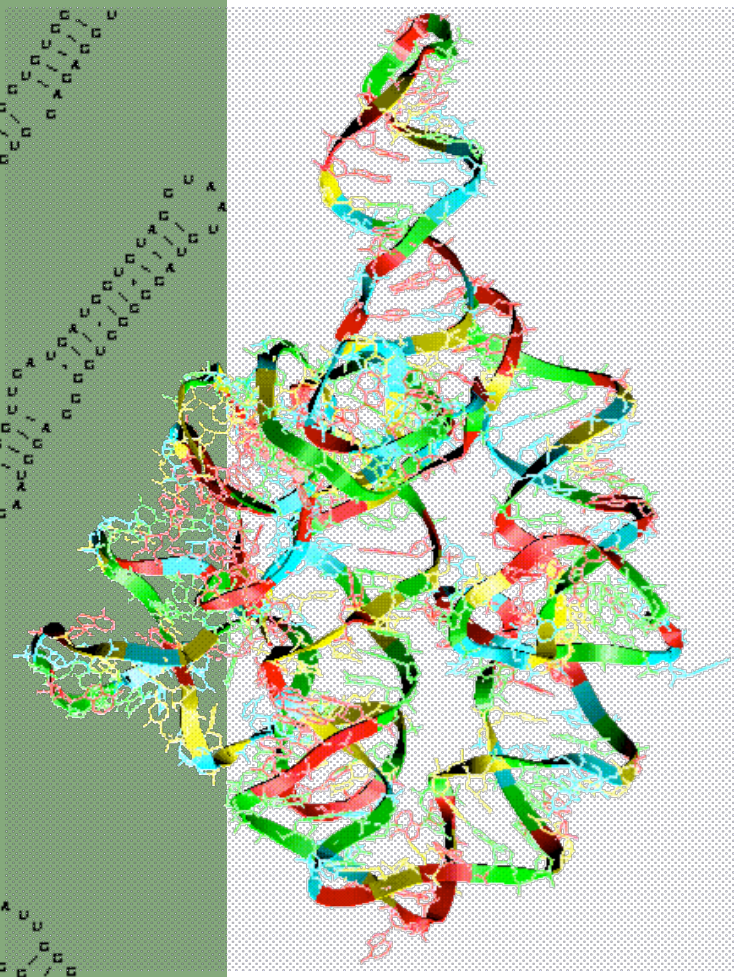
  - Alternative splicing can yield different exon subsets for the same gene, and hence different protein products

RNA can be functional

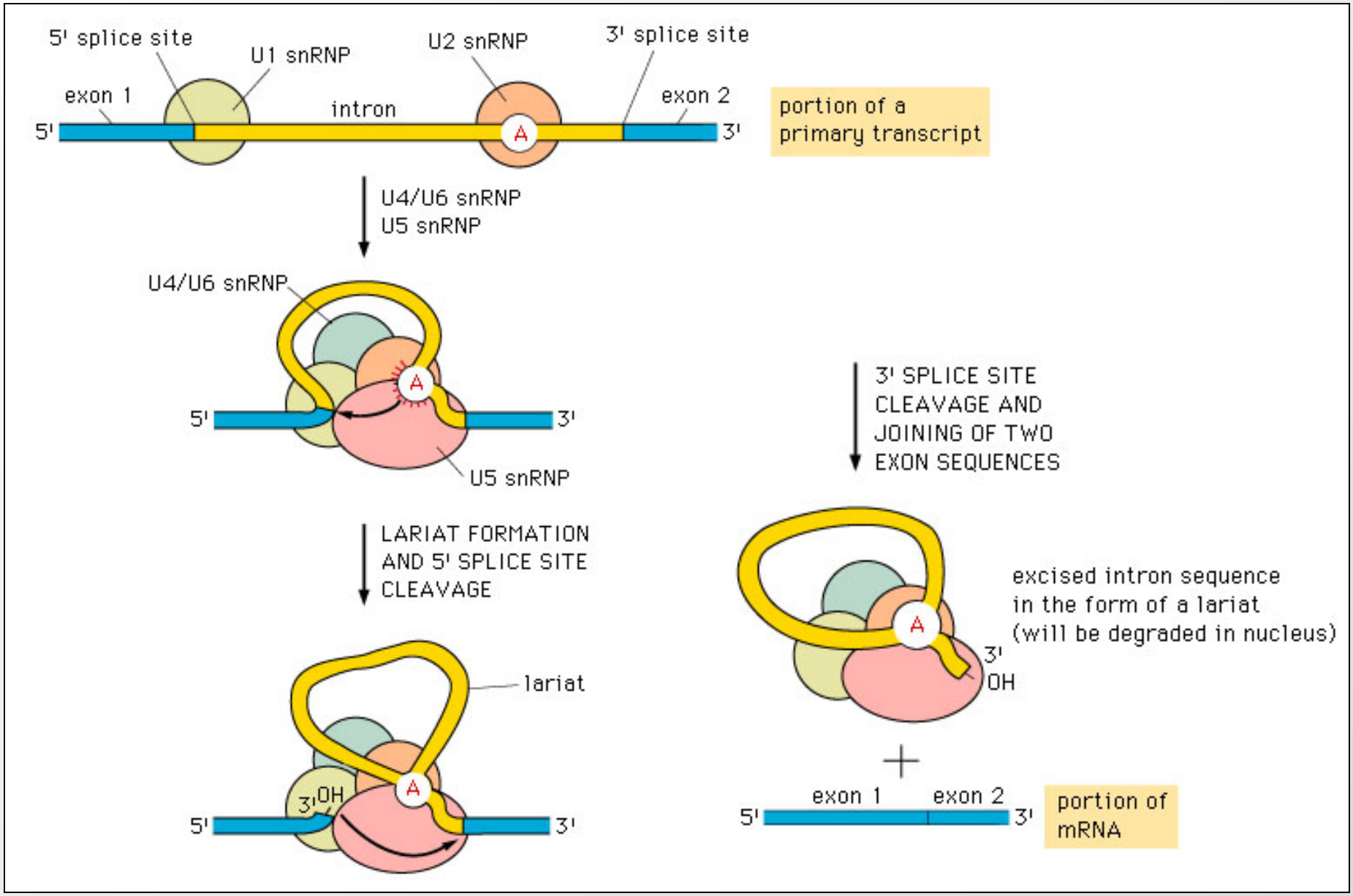
- Single Strand allows complex structure
  - Self-complementary regions form helical stems
  - Three-dimensional structure allows functionality of RNA
- Four types of RNA
  - mRNA: messenger of genetic information
  - tRNA: codon-to-amino acid specificity
  - rRNA: core of the ribosome
  - snRNA: splicing 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: 2ndary and 3rdary

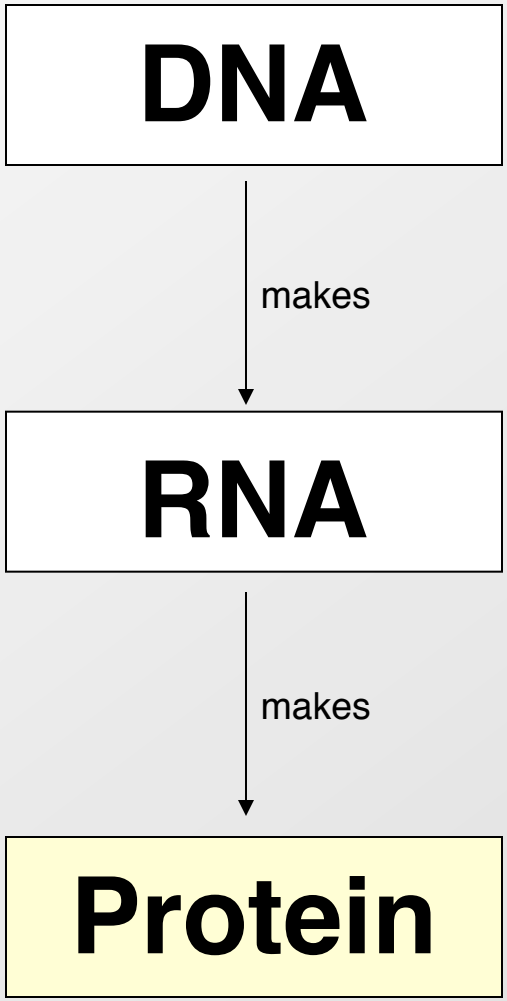




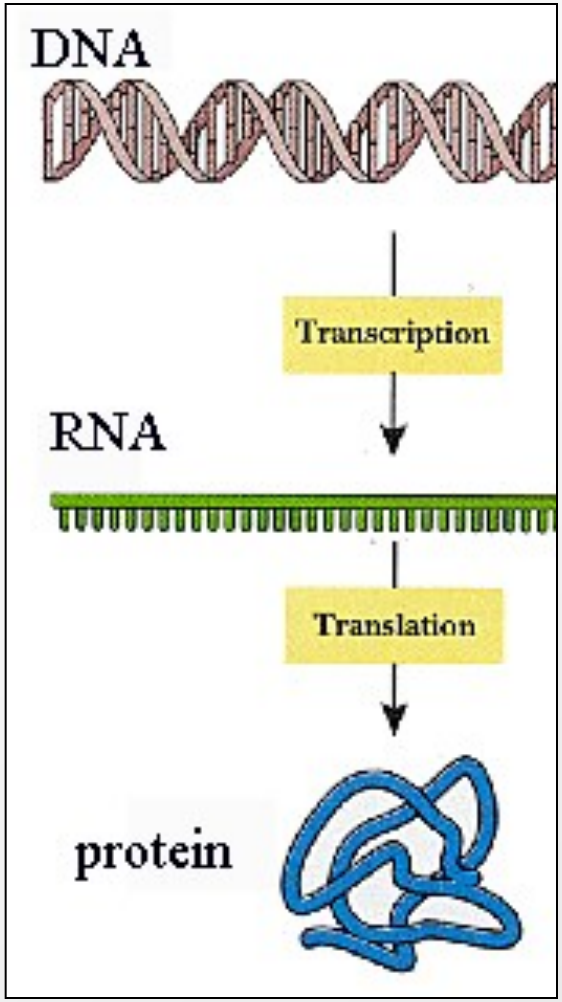
Splicing machinery made of RNA



“Central dogma” of Molecular Biology

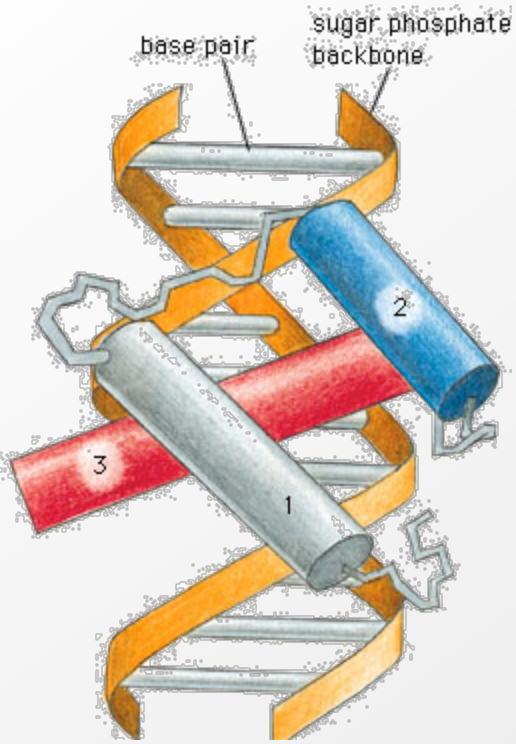


Proteins carry out the cell's chemistry



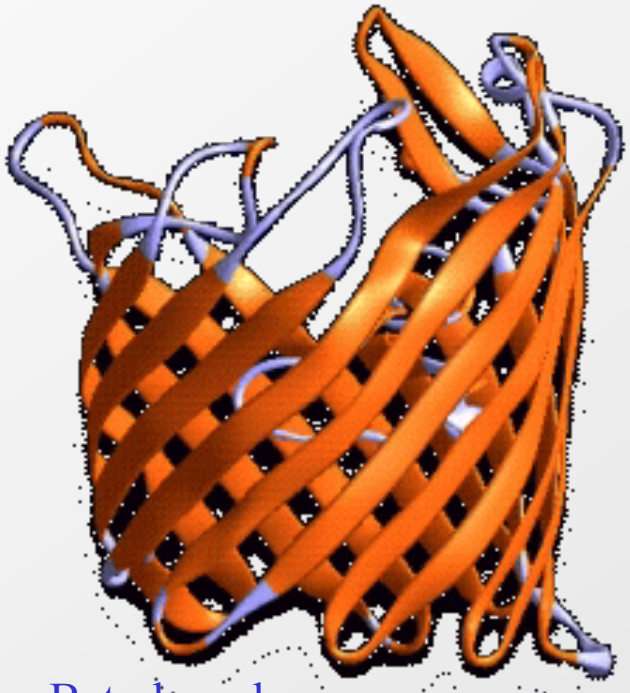
- More complex polymer
  - Nucleic Acids have 4 building blocks
  - Proteins have 20. Greater versatility
  - Each amino acid has specific properties
- Sequence → Structure → Function
  - The amino acid sequence determines the three-dimensional fold of protein
  - The protein's function largely depends 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 streptavadin and porin. Note that some structures are

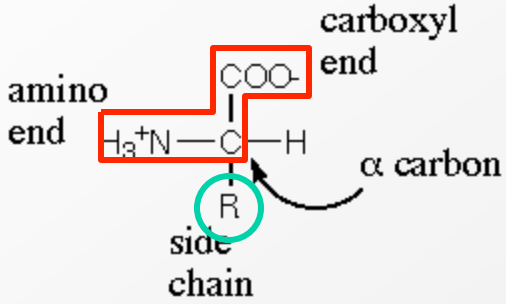
intermediate between the extreme barrel and sandwich arrangements.

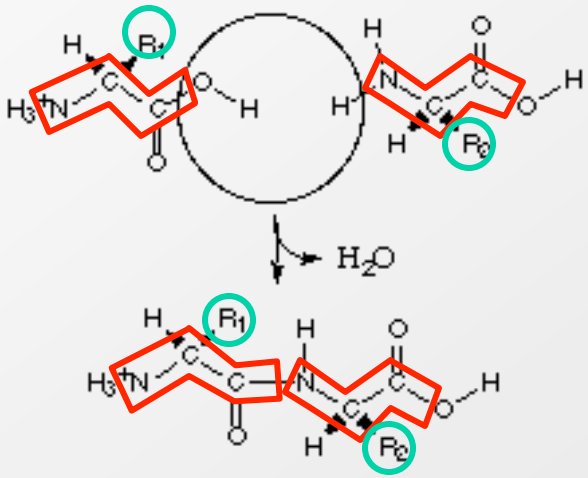
Alpha-beta horseshoe

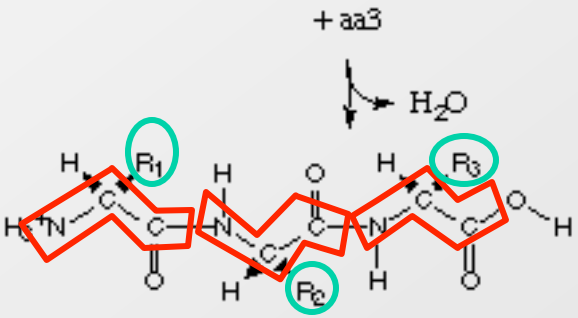
this placental ribonuclease inhibitor is a cytosolic protein that binds extremely strongly to any ribonuclease that may leak into the cytosol. 17-stranded parallel b sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks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

Protein building blocks

- Amino Acids

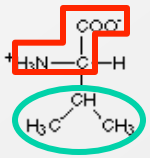


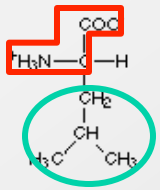


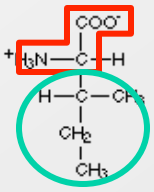


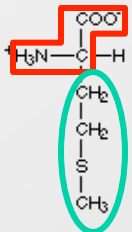
etc...

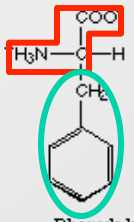
Amino acids with hydrophobic side groups


Valine (val)

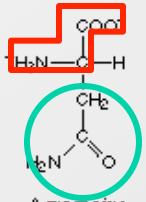

Leucine (leu)

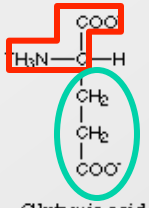

Isoleucine (ile)

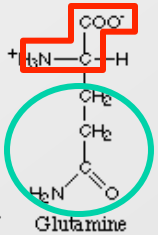

Methionine (met)

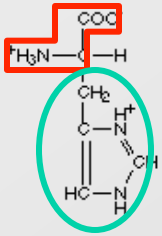

Phenylalanine (phe)

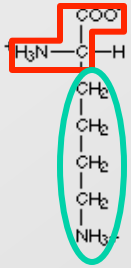
Amino acids with hydrophilic side groups

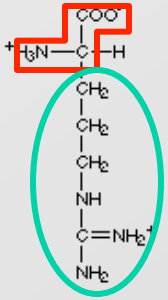

Asparagine (asn)

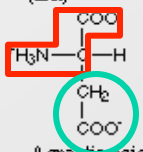

Glutamic acid (glu)


Glutamine (gln)

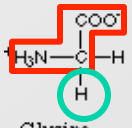

Histidine (his)

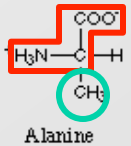

Lysine (lys)

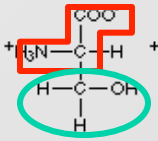

Arginine (arg)

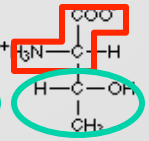

Aspartic acid (asp)

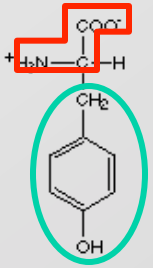
Amino acids that are in between

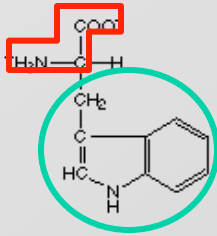

Glycine (gly)

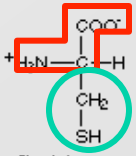

Alanine (ala)

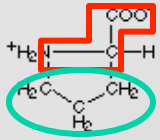

Serine (ser)


Threonine (thr)


Tyrosine (tyr)

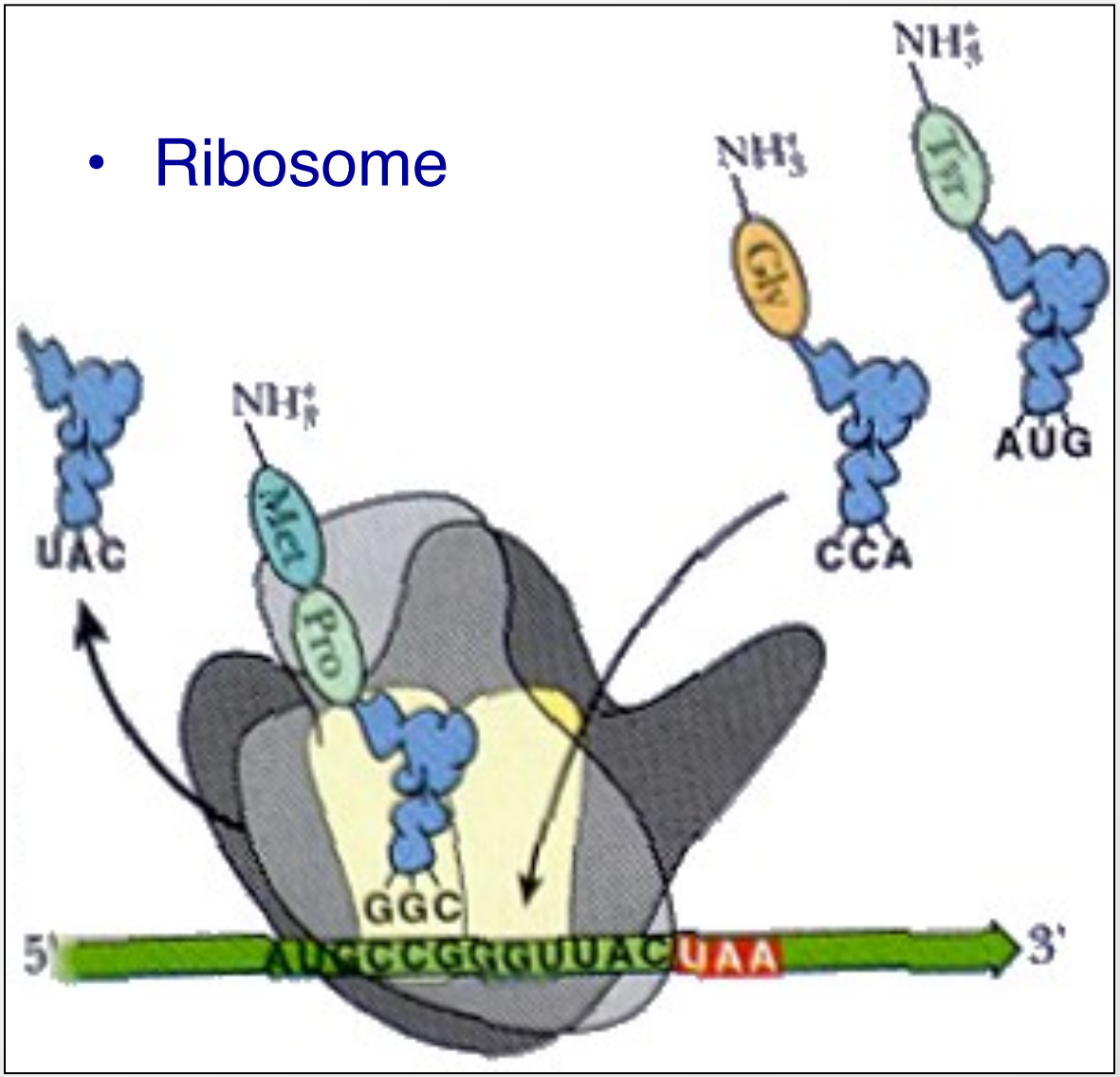

Tryptophan (trp)


Cysteine (cys)

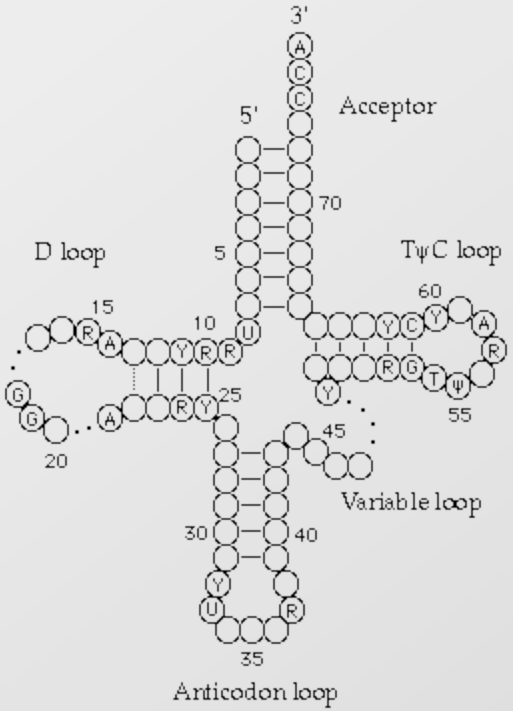

Proline (pro)

From RNA to protein: Translation

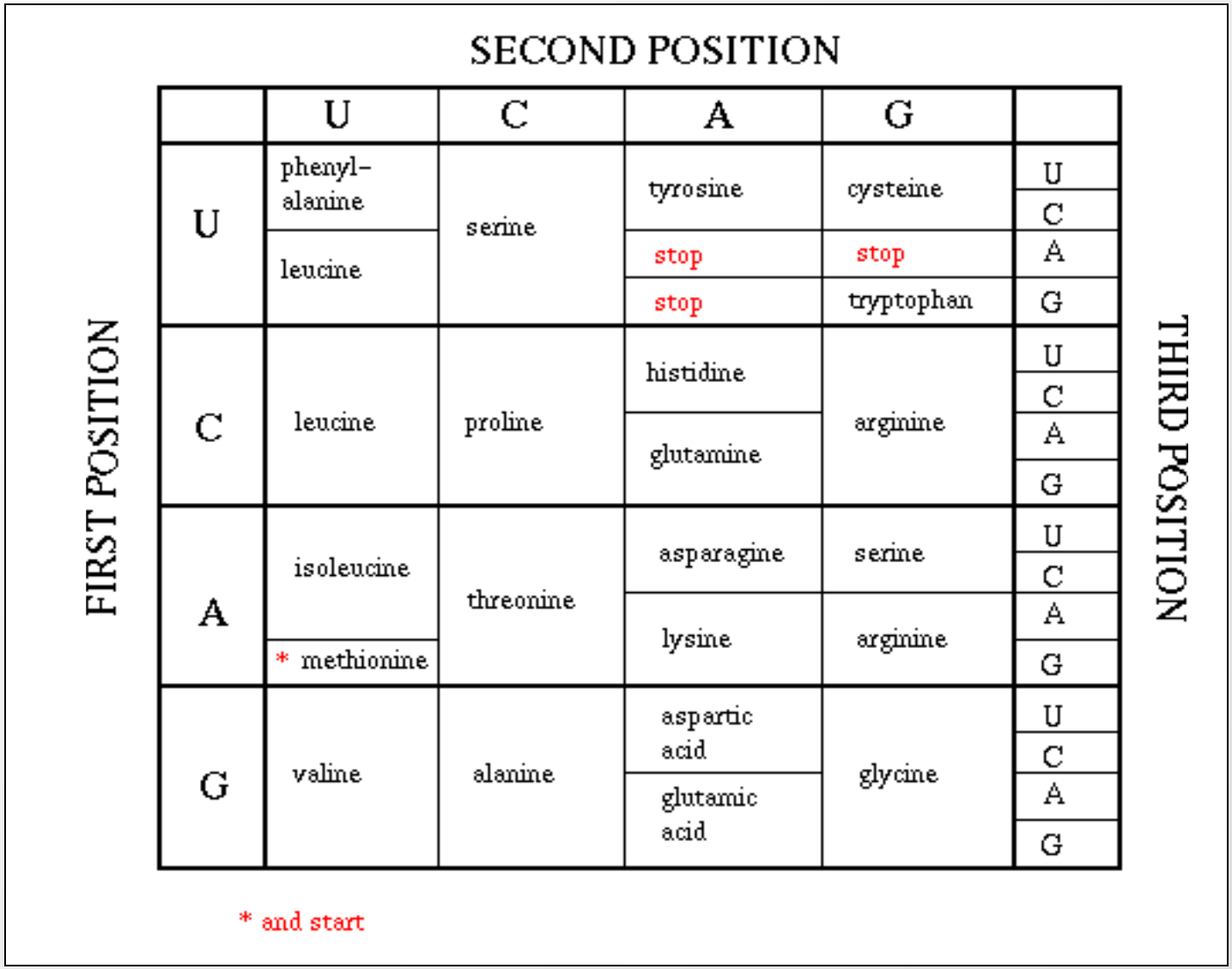
- Ribosome



- tRNA



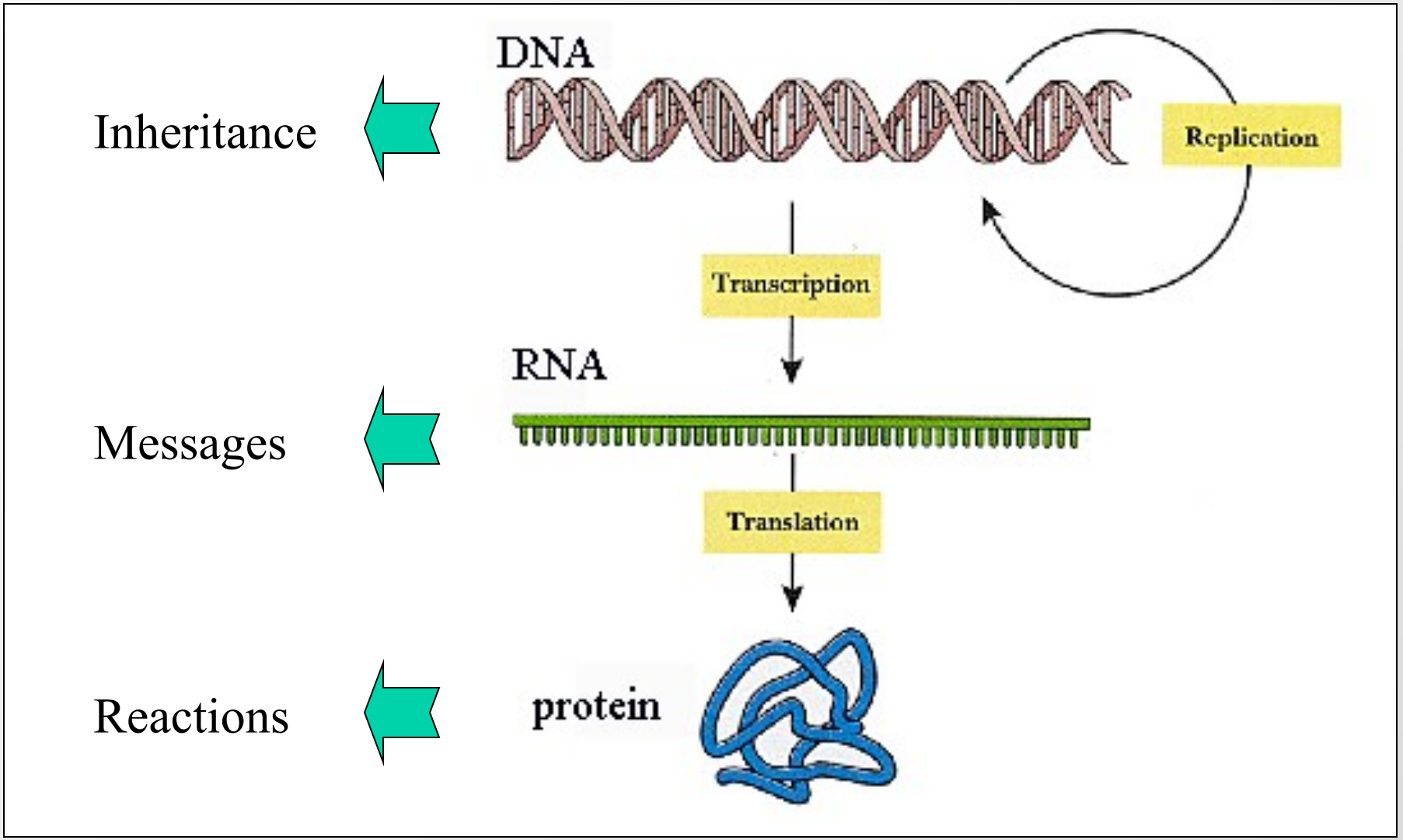
The Genetic Code



➔ Use evolutionary and compositional properties to computationally discover protein-coding genes

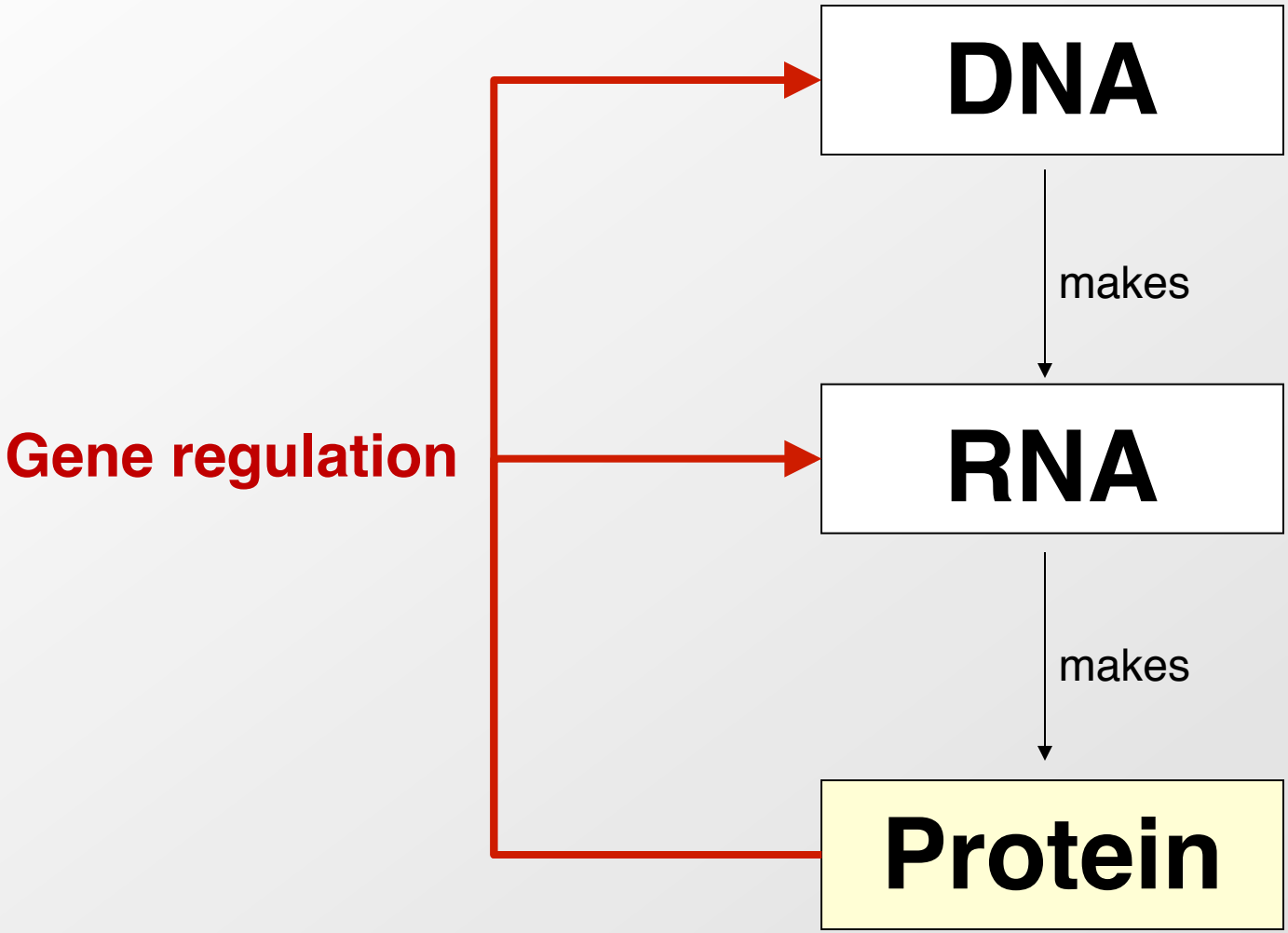
Summary: The Central Dogma

DNA makes RNA makes Protein



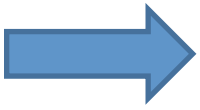
Cellular dynamics and regulation

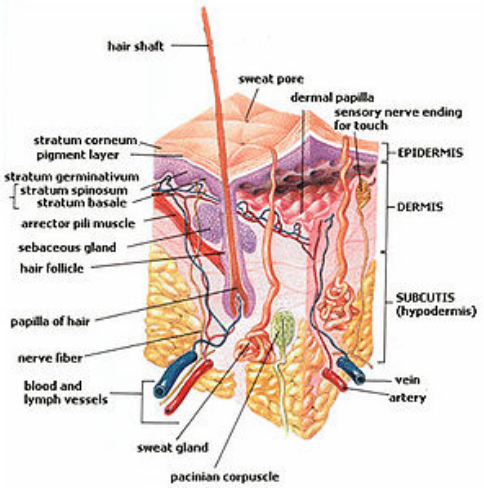
How cells move through this Central Dogma

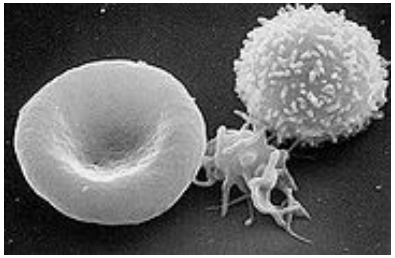


Animal/Human gene regulation: One genome ⇔ Many cell types

ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA







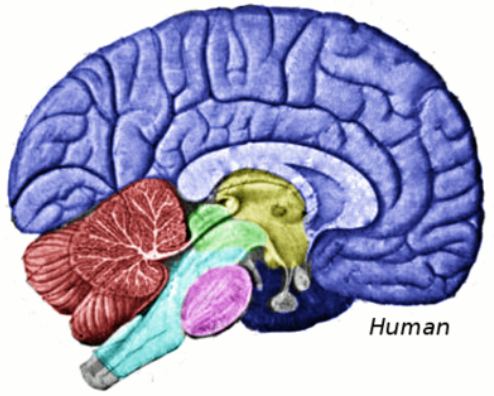
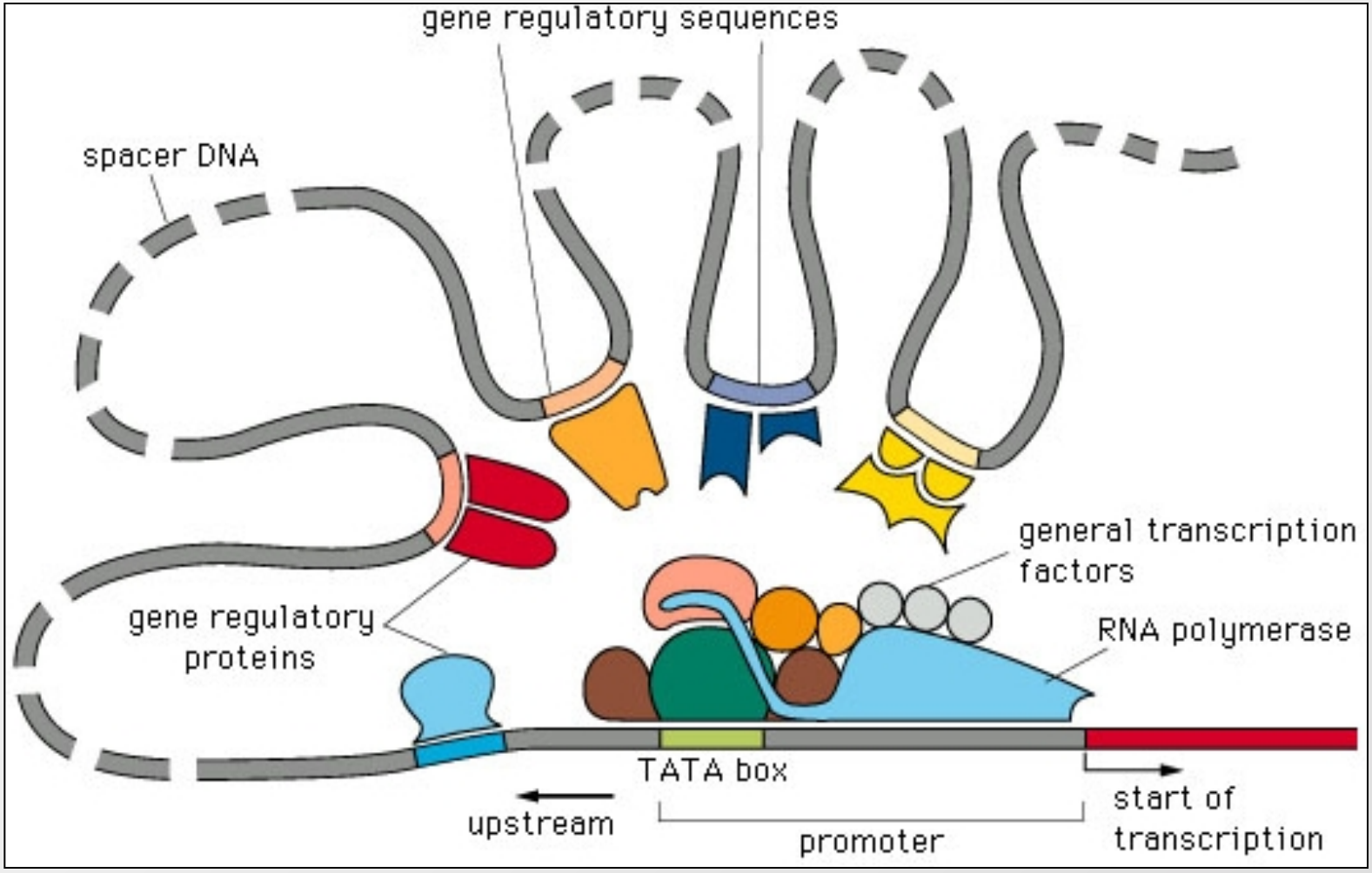


Image Source wikipedia

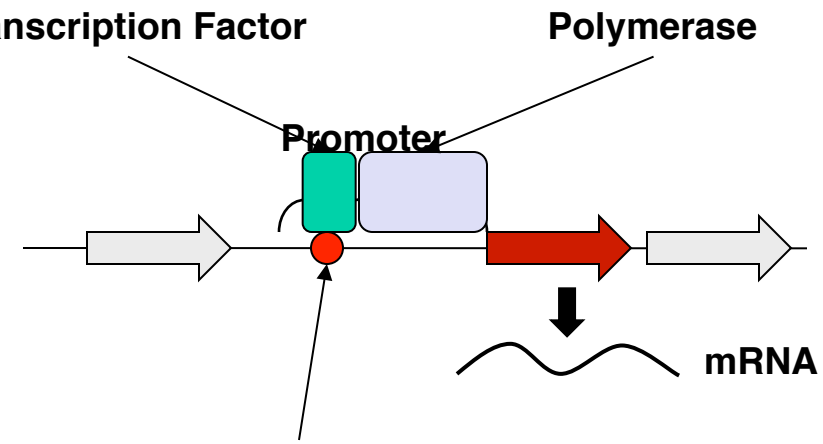
Eukaryotic Gene Regulation



Diverse roles for regulatory non-coding RNAs

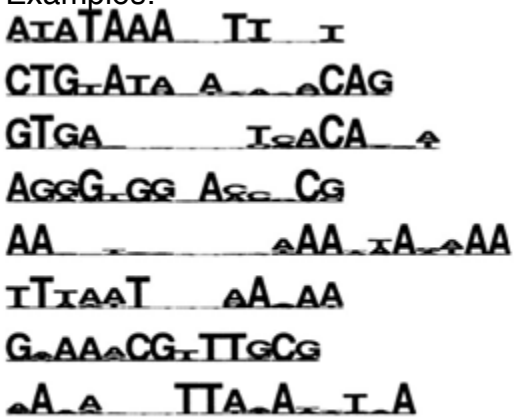
- **Small RNA pathways (18-21 nt)**
  - microRNAs:
    - Repress genes by targeting their 3’ UTRs by complementarity
    - Double-stranded RNA is then recognized and degraded
    - Recently found to also target promoter regions in rare cases
  - piwiRNAs
    - Target and repress transposable elements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



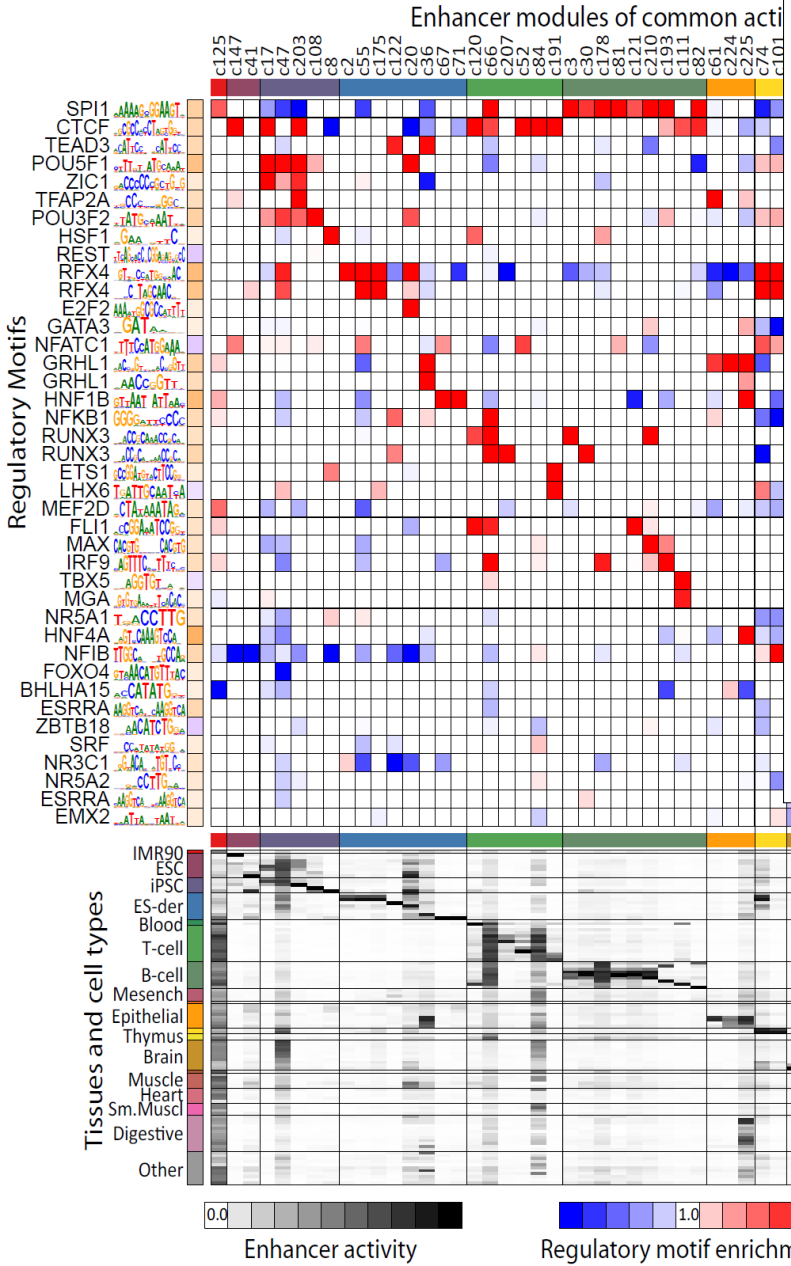
Transcription Factor Binding Site

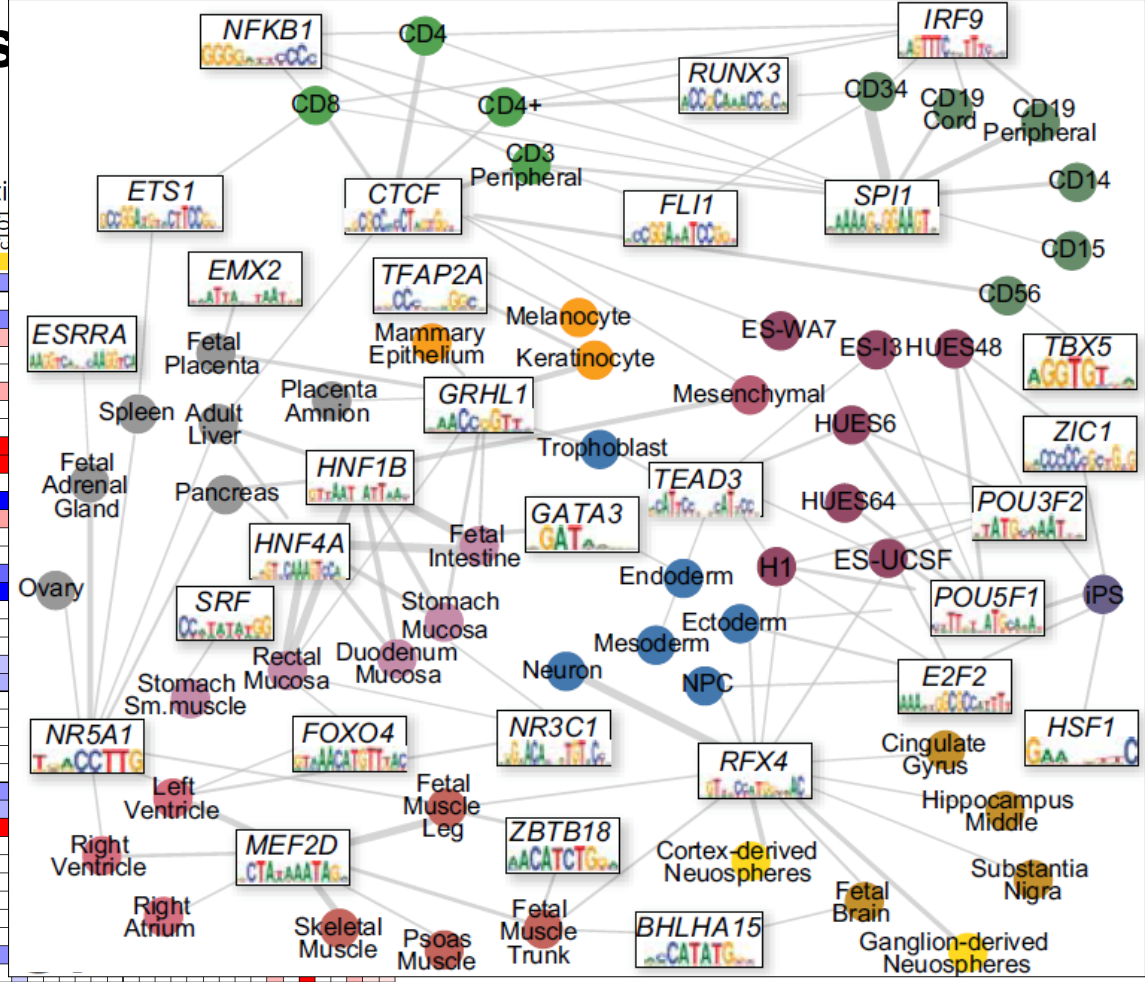
Examples:



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

Predicted motif drivers of enhancer modules

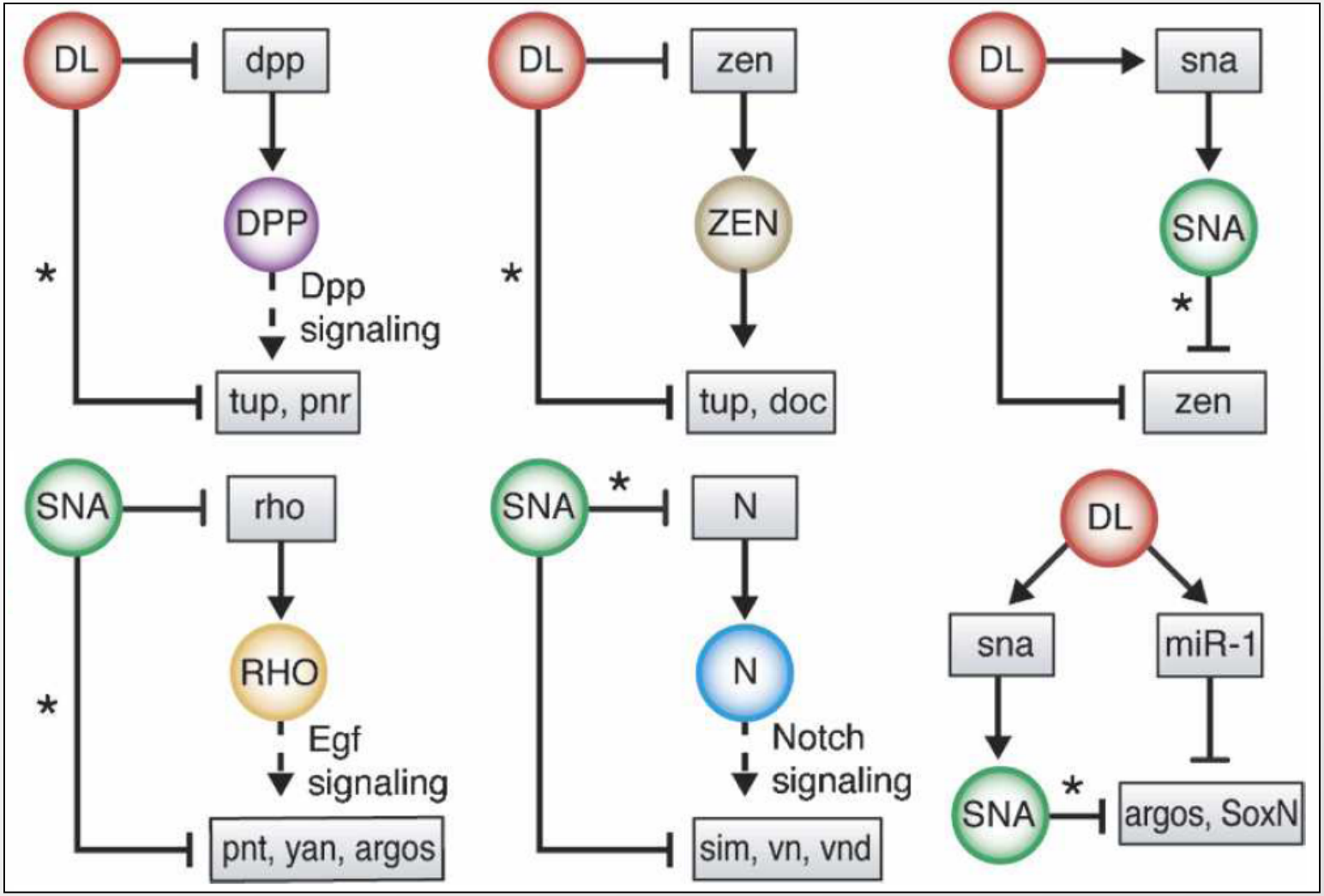




- **Activator and repressor motifs consistent with tissues**

Pouya Kheradpour

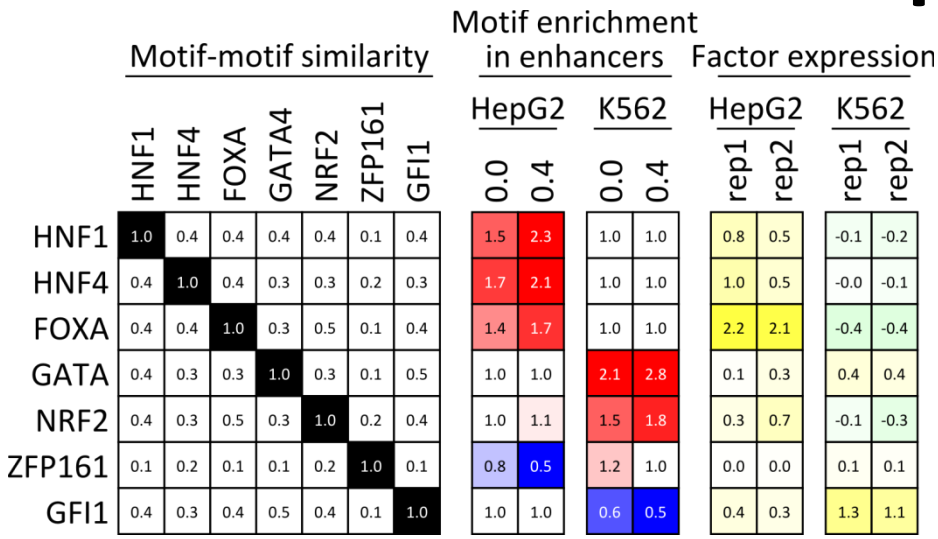
Network components reveal functional modules

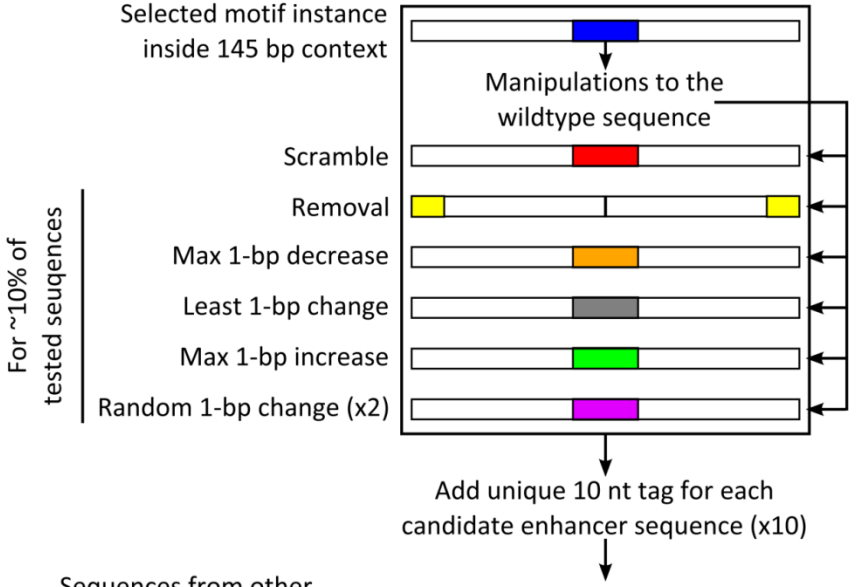


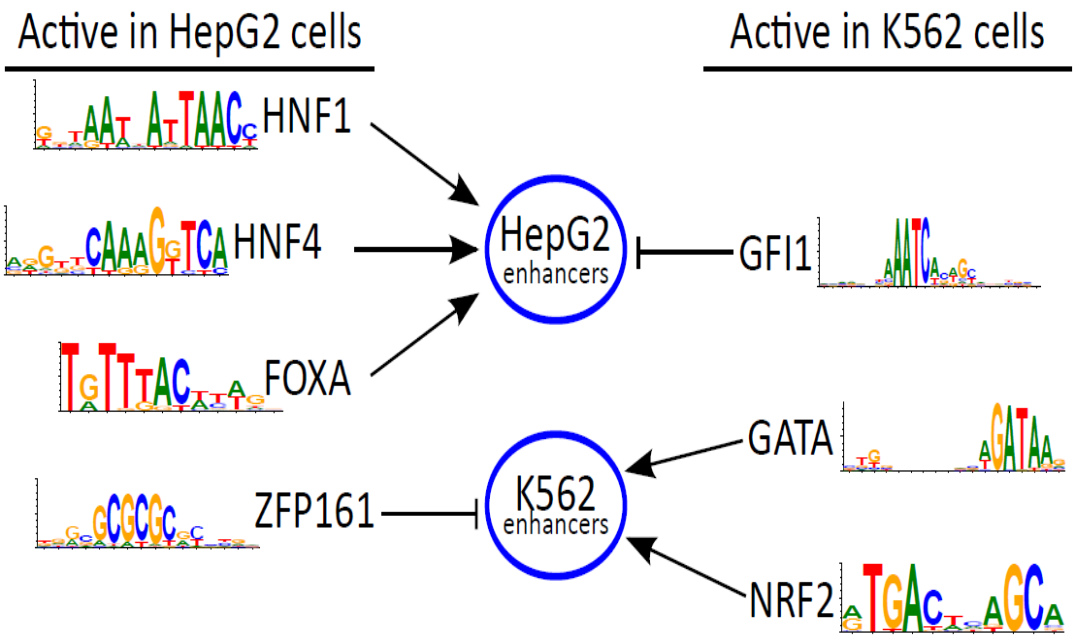
- Feed-forward loops in developmental patterning
- Cooperation of master reg. & downstream reg.

Zeitlinger *et al*, Genes & Development 2007

Systematic motif dissection in 2000 enhancers: 5 activators and 2 repressors in 2 cell lines



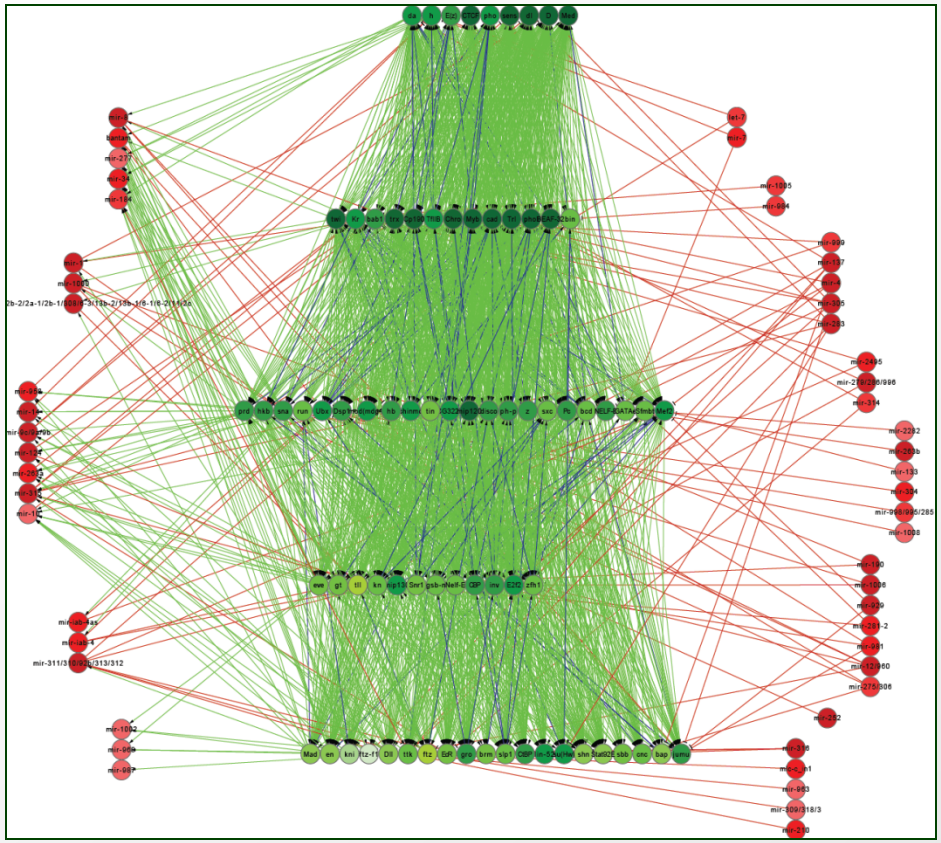


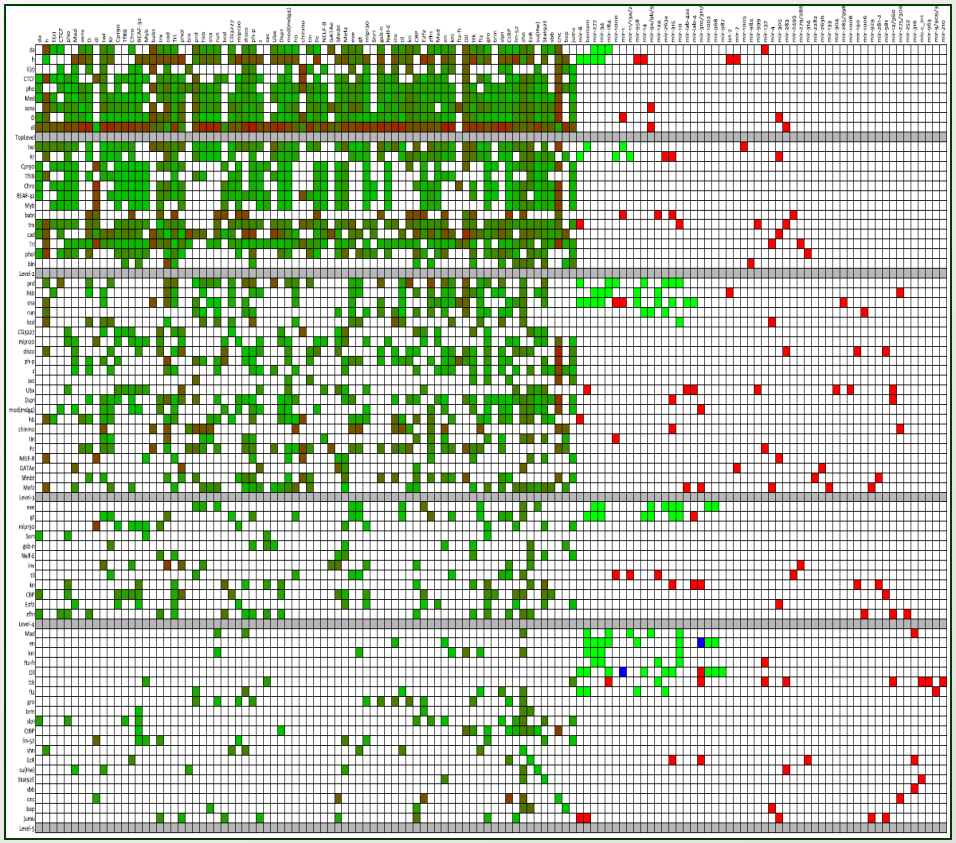


54000+ measurements (x2 cells, 2x repl)

Kheradpour *et al* Genome Research 2013

Emerging properties of regulatory networks

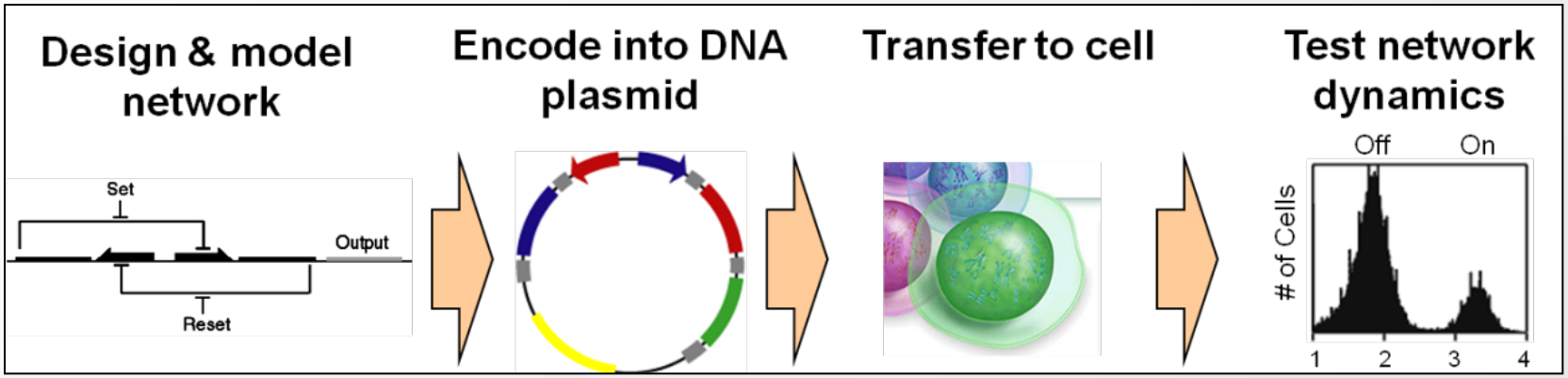




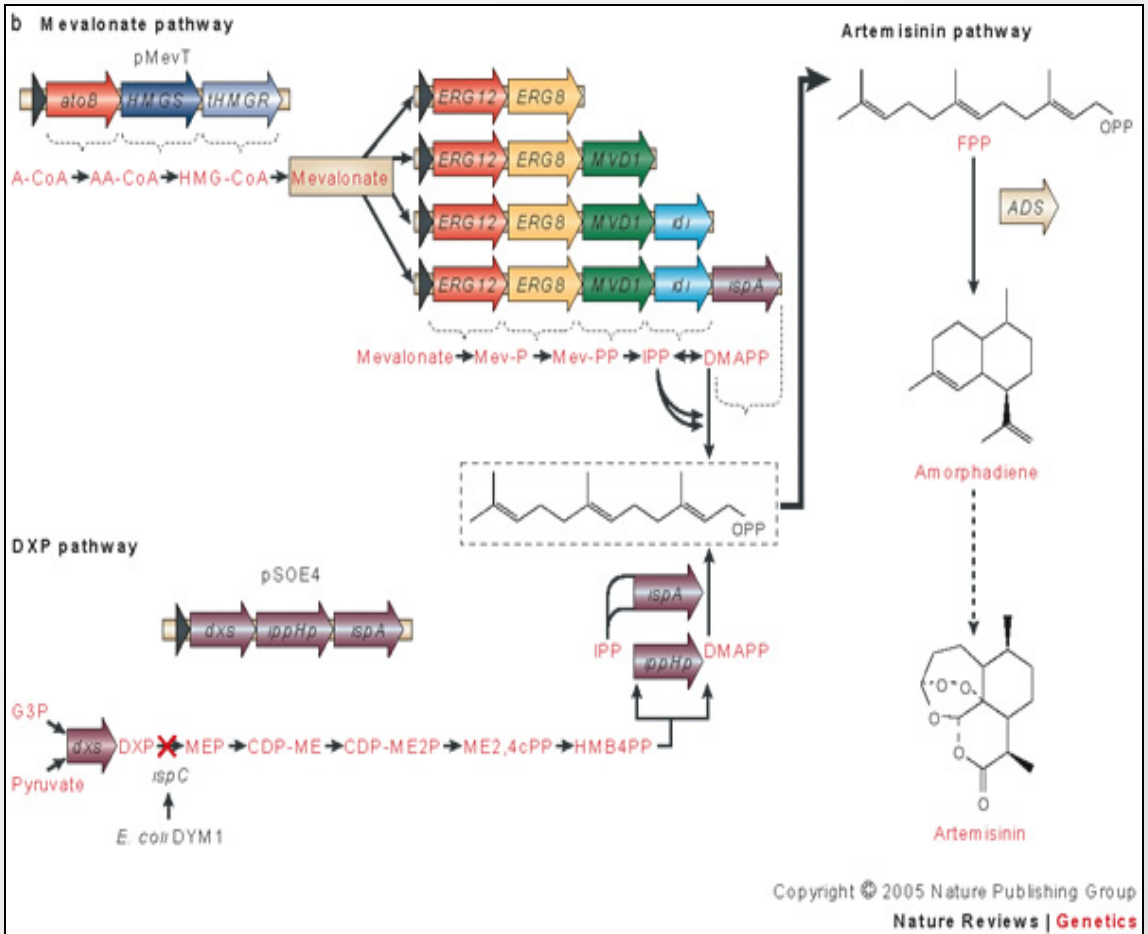
- Hierarchical levels of regulatory control
  - Small number of backward-pointing edges
- Specific / distinct feedback by microRNAs at each level
  - Two classes of TFs: miRNA regulators and miR-regulated

From Systems Biology to Synthetic Biology

Synthetic Regulatory Networks



Synthetic Metabolic Pathways

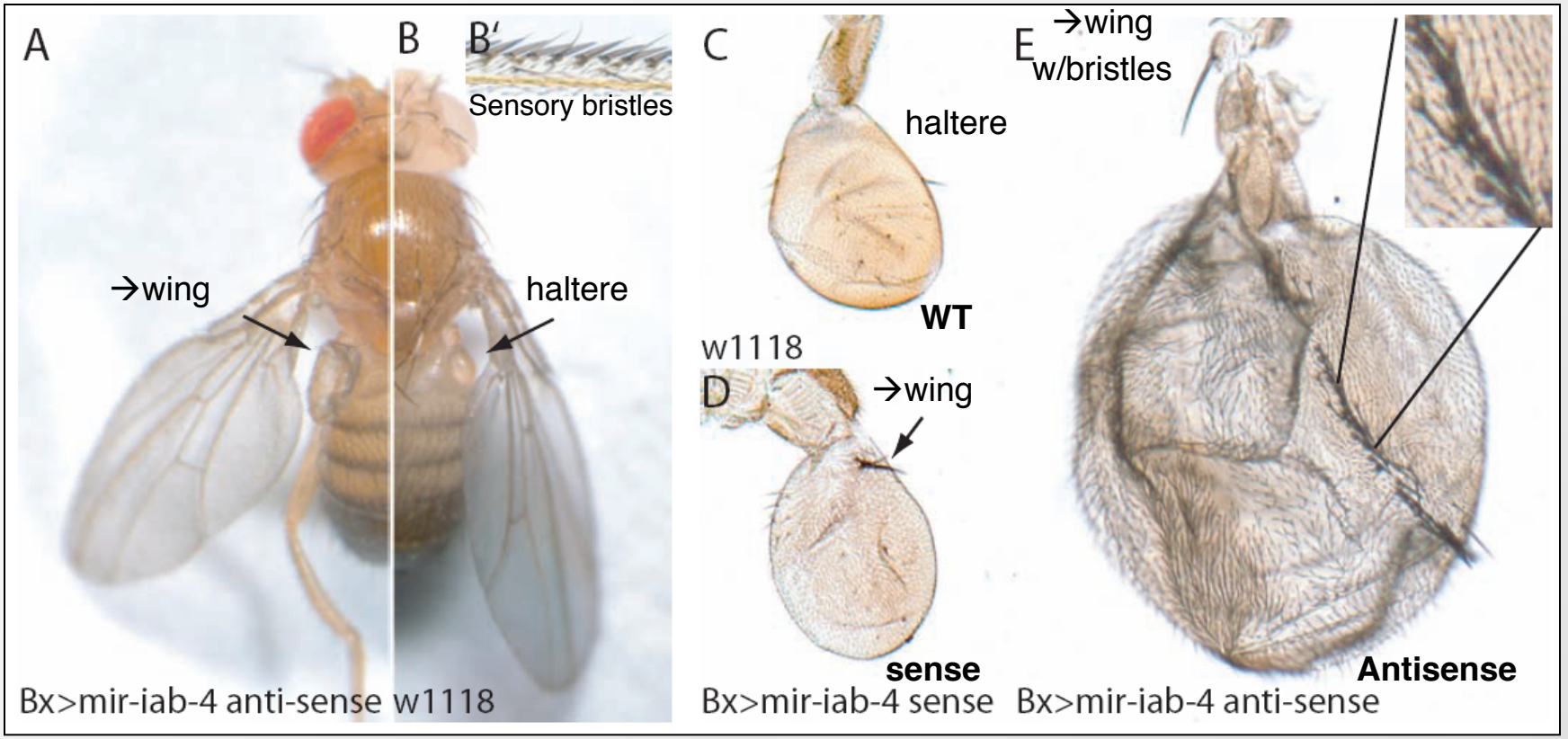


Jim Collins

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

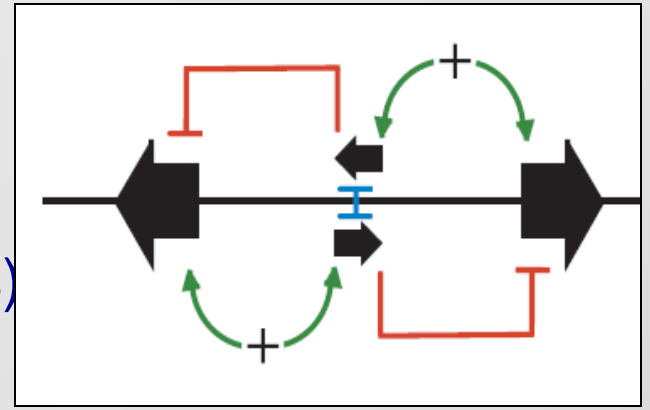
Jay Keasling

Over-express a single microRNA leads to new wing



Note: C,D,E same magnification

- 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

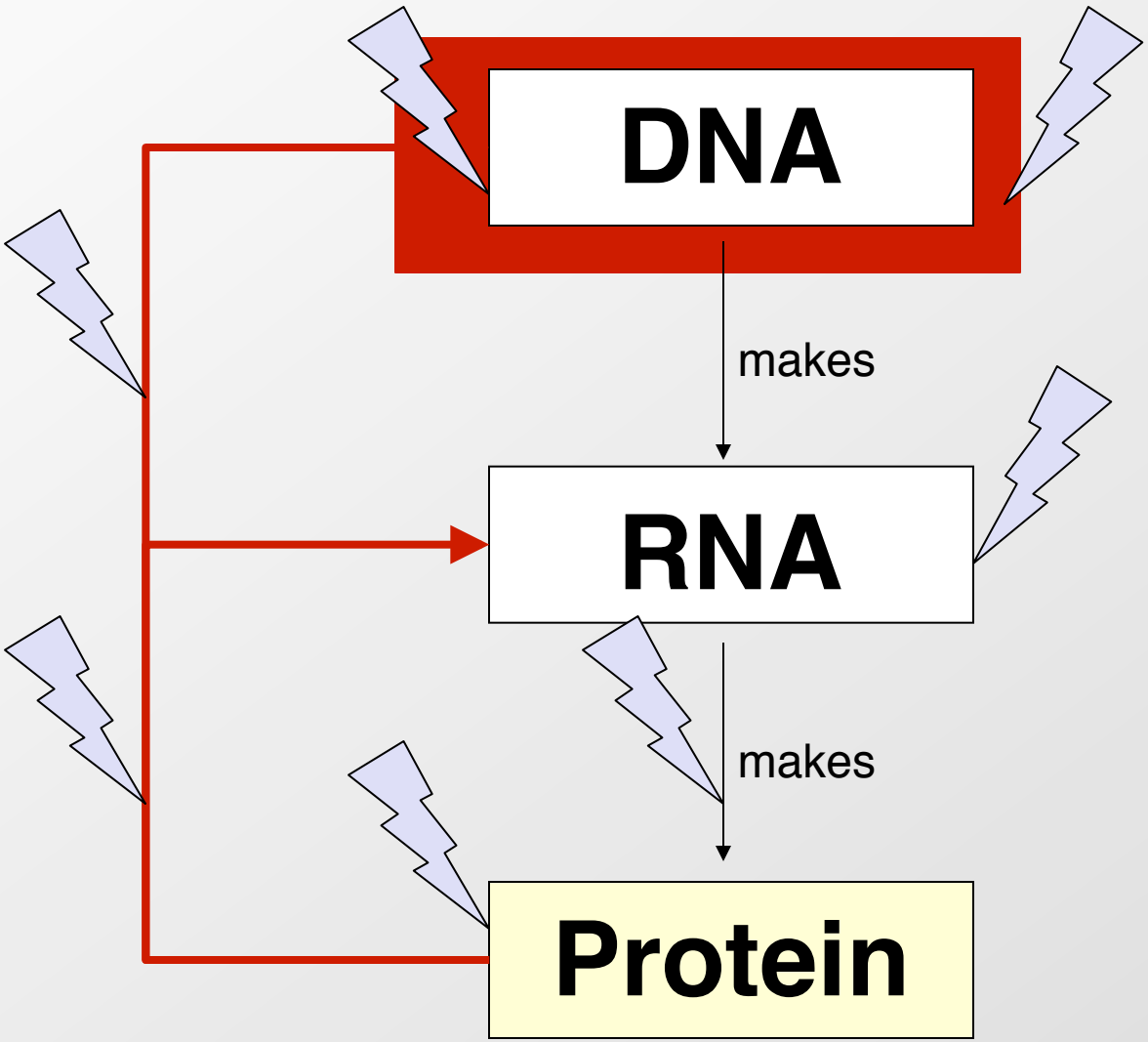


Stark *et al*, Genes&Development 2007

Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Mon 9/23**	PS1 out on:L1-L5	1	Thu, Sep 5	Introduction		L1	Algorithms, Machine Learning, Networks, Course Overview	1
			Fri, Sep 6			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 10	Module I:	Foundations	L2	Dynamic Programming, Reusing computation, Iterative Functions, Exponential / Poly	2,3
			Thu, Sep 12			L3	Database search, Rapid string matching, Hashing	3
			Fri, Sep 13			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 17		Frontiers	L4	HMMs1: Evaluation, Parsing, posterior decoding, learning, HMM architectures	7,8
	due Mon 9/23		Thu, Sep 19			L5	HMMs2: Applications, architectures, memory, gene finding, chromatin states	7,8
			Fri, Sep 20			No Classes - Student Holiday		
Find prev project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Mon 10/7**	PS2 out on:L6-R4	4	Tue, Sep 24	Module II:	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 26			L7	RNA structure and function. RNA world, RNA-seq, transcript structure, RNA folding	14,15
			Fri, Sep 27			R3	Recitation 3: Supervised Learning and Random Forest Classification	
			Fri, Sep 27	cts, self introductions, mentor intro, example projects, teamwork 32D-507				
		5	Tue, Oct 1		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 3			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due Mon 10/7		Fri, Oct 4			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
			Fri, Oct 4	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Thu 10/17. Presented on Fri 10/18**	PS3 out on:L10-R6	6	Tue, Oct 8	Module III:	Foundations	L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Thu, Oct 10			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Fri, Oct 11			R5	Recitation 5: Communication Lab	
		7	Tue, Oct 15			No Classes - Columbus Day Holiday		
			Thu, Oct 17		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due Mon 10/21		Fri, Oct 18			R6	Recitation 6: Motif Discovery, WEEDER, In vitro Motif Discovery - PBMs, Selex	
			Fri, Oct 18	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507				
Evaluate/discuss three peer proposals, NIH review format. **Reviews back Mon 10/28**	PS4 out on:L13-R8	8	Tue, Oct 22	Module IV:	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 24			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 25			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 25	Panel Review: Discuss Peer Projects. Feedback sent out from group reviews. 32D-463 (Star).				
		9	Tue, Oct 29		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/7**	due Mon 11/4		Thu, Oct 31			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
			Fri, Nov 1			R8	Recitation 8: Phylogenetic distance metrics, Coalescent Process	
	PS5 out on:L17-R10	10	Tue, Nov 5	Module V:	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 7			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 8			No Recitation, Veterans Day		
Continue making subst. progress on proposed milstones.Write outline of final report. **Midcourse report due Mon 11/25**	due Fri 11/15	11	Tue, Nov 12		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
			Thu, Nov 14			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
			Fri, Nov 15			R9	Recitation 9: Quiz Review	
	No more psets! (work on your final project)	12	Tue, Nov 19	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 21			L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Fri, Nov 22			R10	Recitation 10: Project Feedback, results, interpretation, directions	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/8**		13	Tue, Nov 26	Module VI:	Frontiers	L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Thu, Nov 28			No lecture, thanksgiving break - Thu Nov 28, 2019		
			Fri, Nov 29			No recitation, thanksgiving break		
		14	Tue, Dec 3			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 5			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 6			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Presentations on Tue 12/10**		15	Tue, Dec 10			L25	Final Presentations - Part I (1pm). 32-141 (Classroom)	
			Tue, Dec 10			L25	Final Presentations - Part I (2:30pm). 32D-463 (Star)	

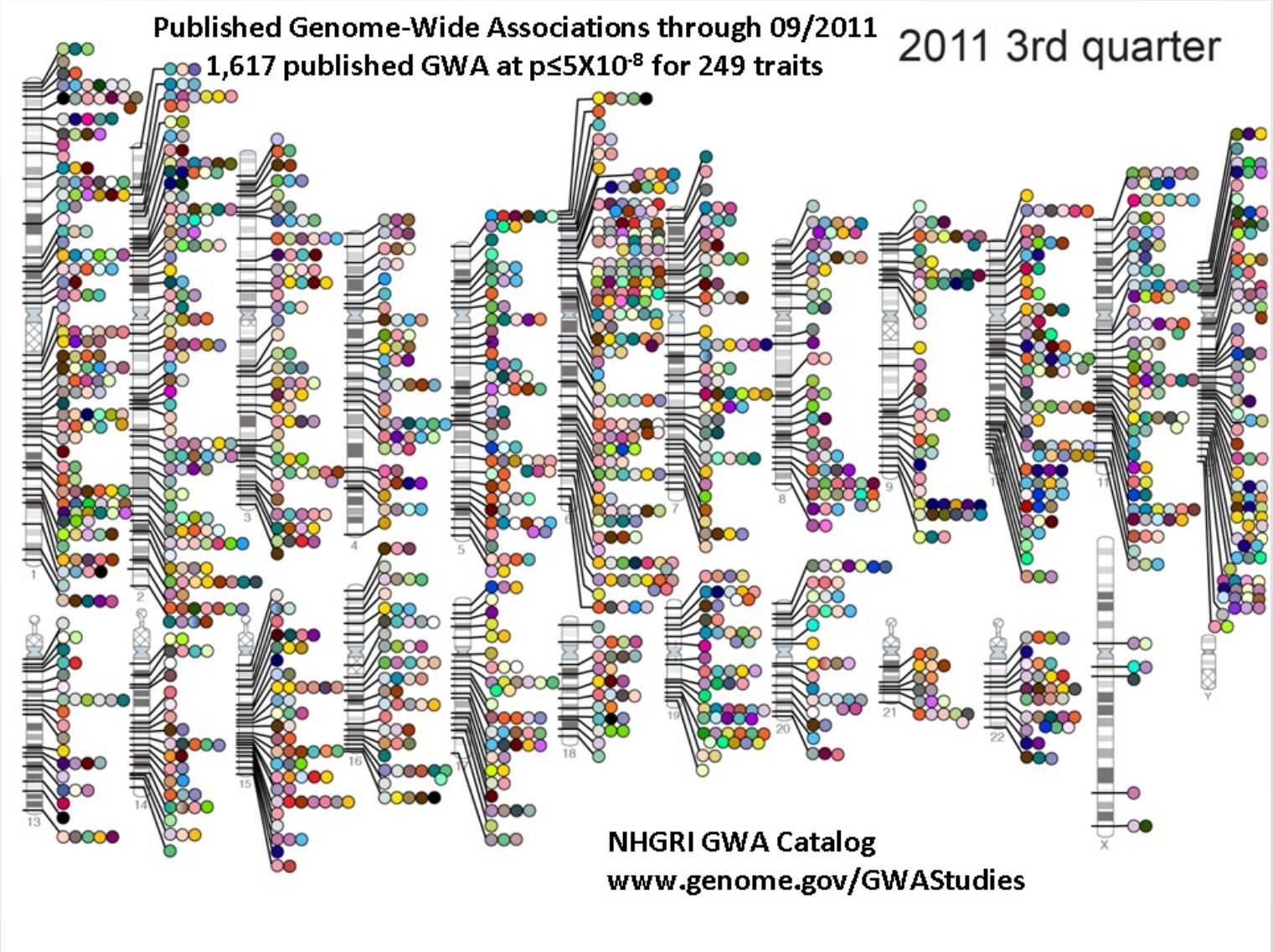
Brief intro to Human Genetics

The role of genetic alterations

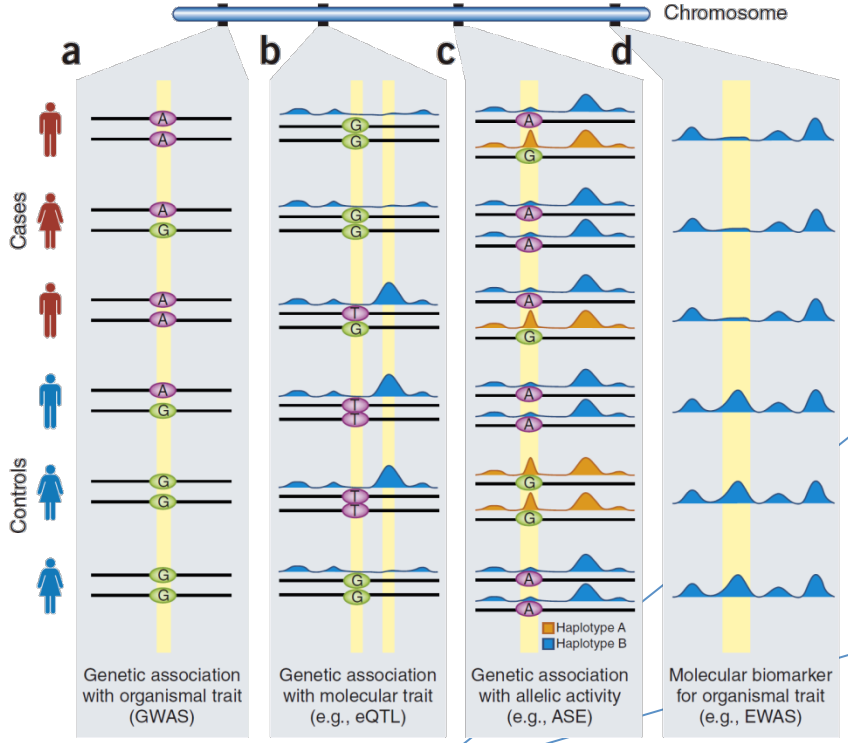


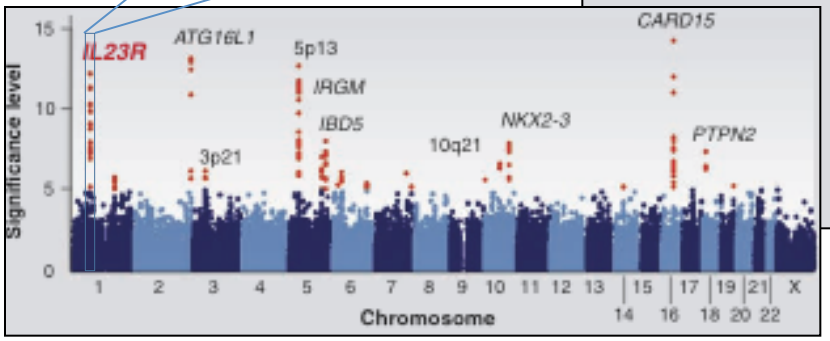
Brief intro to human genetics

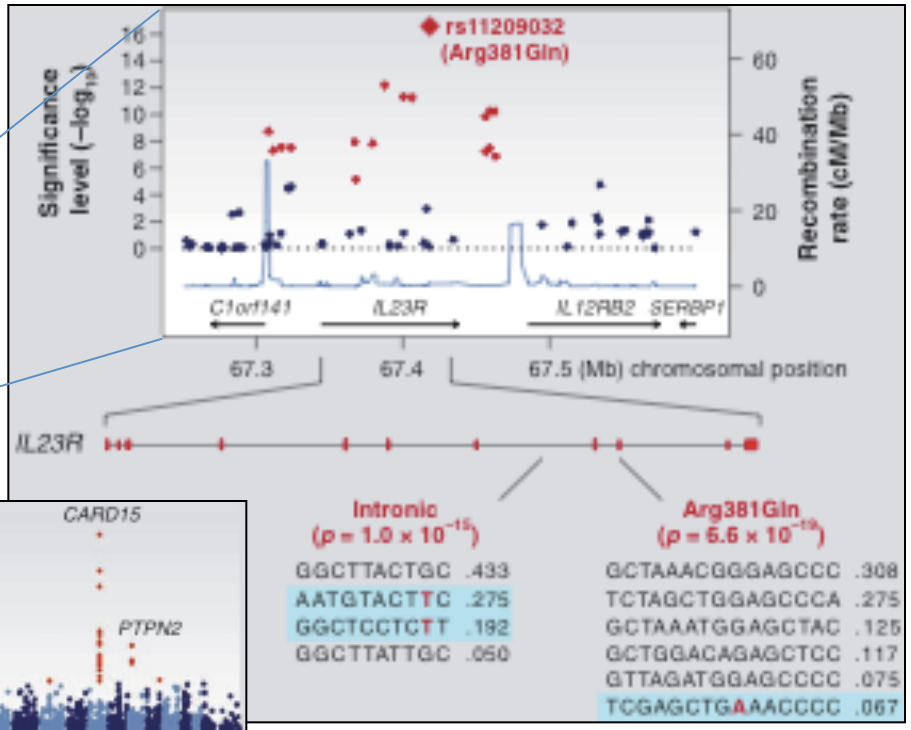
- **Human genome**: 3.2B letters, 2 copies, 23 chromosomes, 20k genes, ~3M common SNPs, ~500k haplotype blocks



The power and challenge of disease-association studies





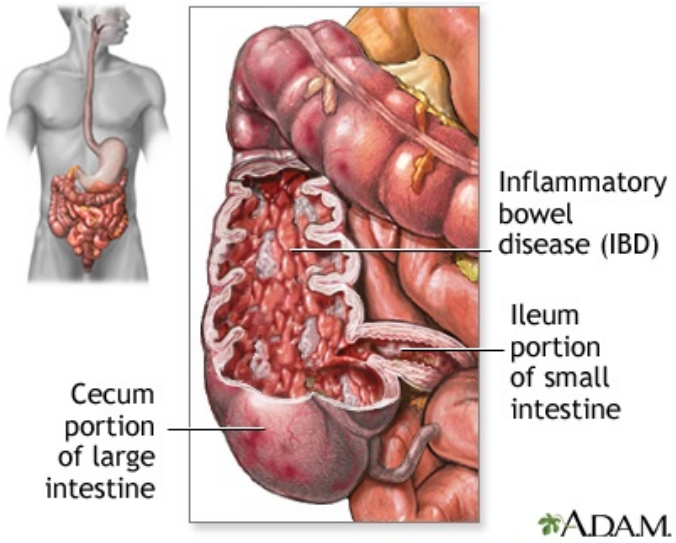


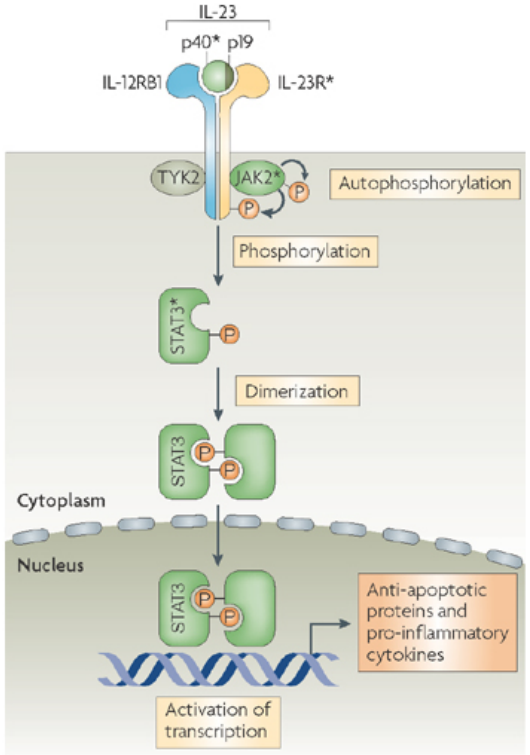
Slide credit: Luke Ward, Mark Daly

- **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**

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

Genomewide association in schizophrenia with 40,000 cases

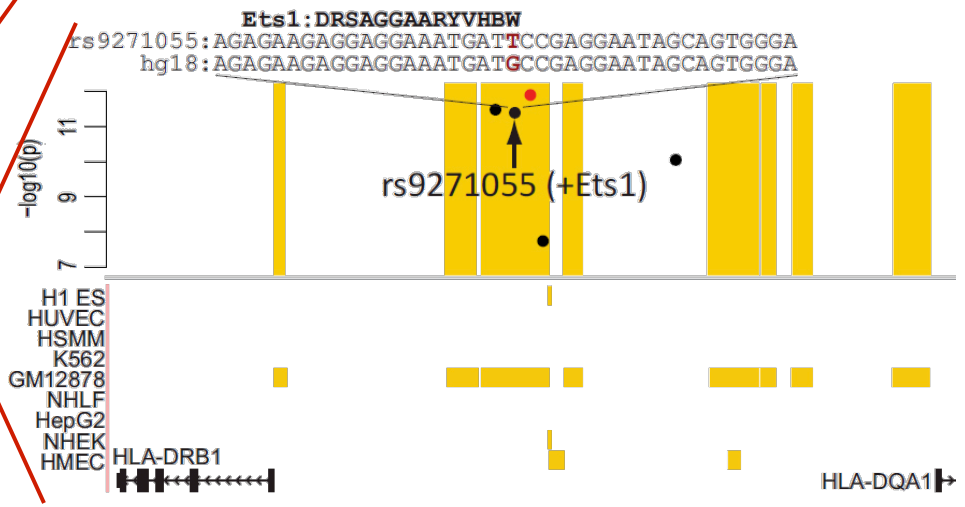
Stephan Ripke

More than 100 distinct regions of the genome associated to schizophrenia!!!

Interpreting non-coding variants

Phenotype	Top Cell Type	Total #SNPs from Study	#SNPs in enh. States 4 and 5	p-value	FDR	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
Erythrocyte phenotypes (Ref. 38)	K562	35	9	$<10^{-7}$	0.02	9	17	4	0	0	1	2	1	1
Blood lipids (Ref. 39)	HepG2	101	13	$<10^{-7}$	0.02	3	5	0	11	2	3	3	4	3
Rheumatoid arthritis (Ref. 40)	GM12878	29	7	2.0×10^{-7}	0.03	0	0	15	0	2	0	0	2	3
Primary billary cirrhosis (Ref. 41)	GM12878	6	4	6.0×10^{-7}	0.03	0	11	41	0	0	0	0	8	8
Systemic lupus erythromatosus (Ref. 42)	GM12878	18	6	9.0×10^{-7}	0.03	0	4	21	0	5	8	0	3	5
Lipoprotein cholesterol/triglycerides (Ref. 43)	HepG2	18	5	1.2×10^{-6}	0.03	17	8	0	24	3	6	4	3	3
Hematological traits (Ref. 44)	K562	39	7	1.7×10^{-6}	0.03	0	12	10	2	1	0	0	1	0
Hematological parameters (Ref. 45)	K562	28	6	2.2×10^{-6}	0.03	0	15	7	0	5	7	7	3	2
Colorectal cancer (Ref. 46)	HepG2	4	3	3.8×10^{-6}	0.03	0	0	0	66	0	12	0	12	12
Blood pressure (Ref. 47)	K562	9	4	5.0×10^{-6}	0.04	0	30	14	0	10	6	7	5	11

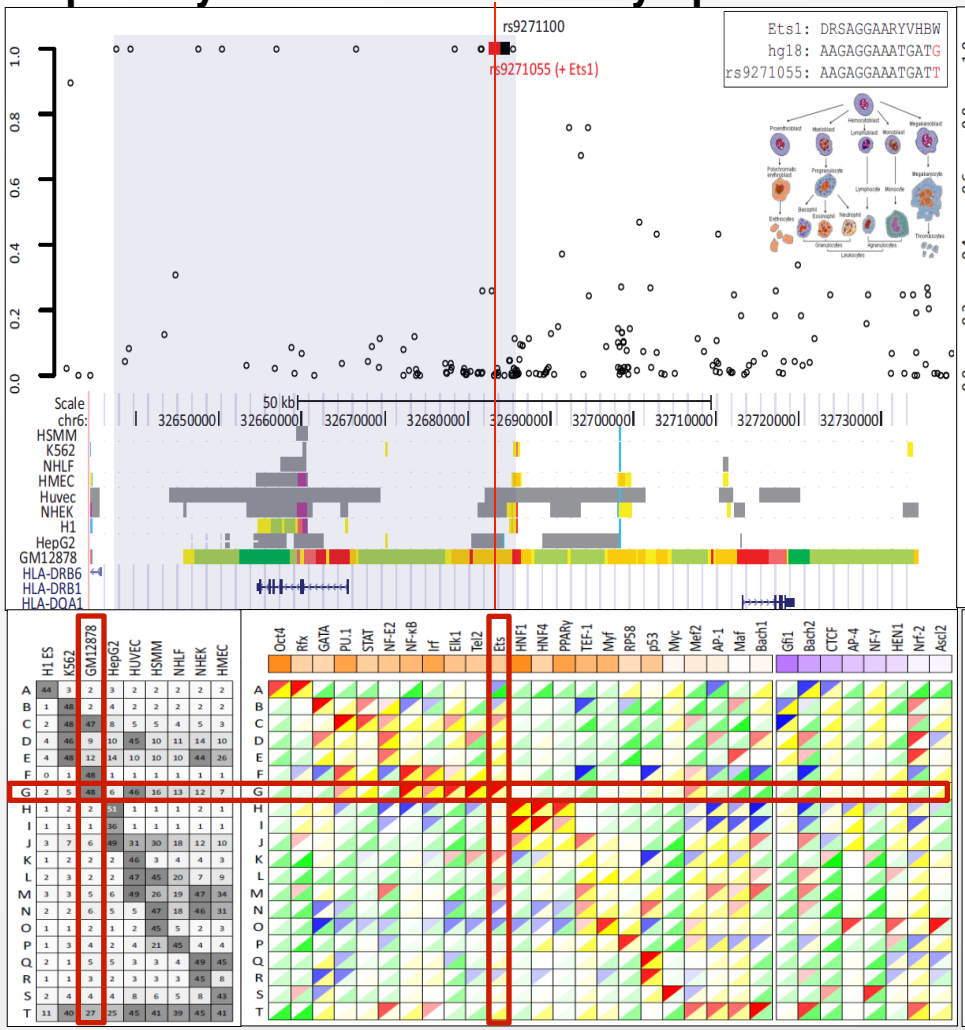
SNP	H1 ES	K562	GM	HepG2	Huvec	HSMM	NHLF	NHEK	HMEC	Chrom. Band	Gene	Link Sco	Distanc
rs13385731										2p22			
rs10036748										5q33			
rs1385374										12q24	MGC16384	-	1
rs2230926										6q23	TNFAIP3	3.7	7
rs4728142										7q32	IRF5	-	4
rs9271100										6p21	HLA-DRB1	4.5	19
rs4917014										7p12	IKZF1	2.2	38
rs7812879										8p23	BLK	2.9	11
rs2205960										1q25			
rs548234										6q21			



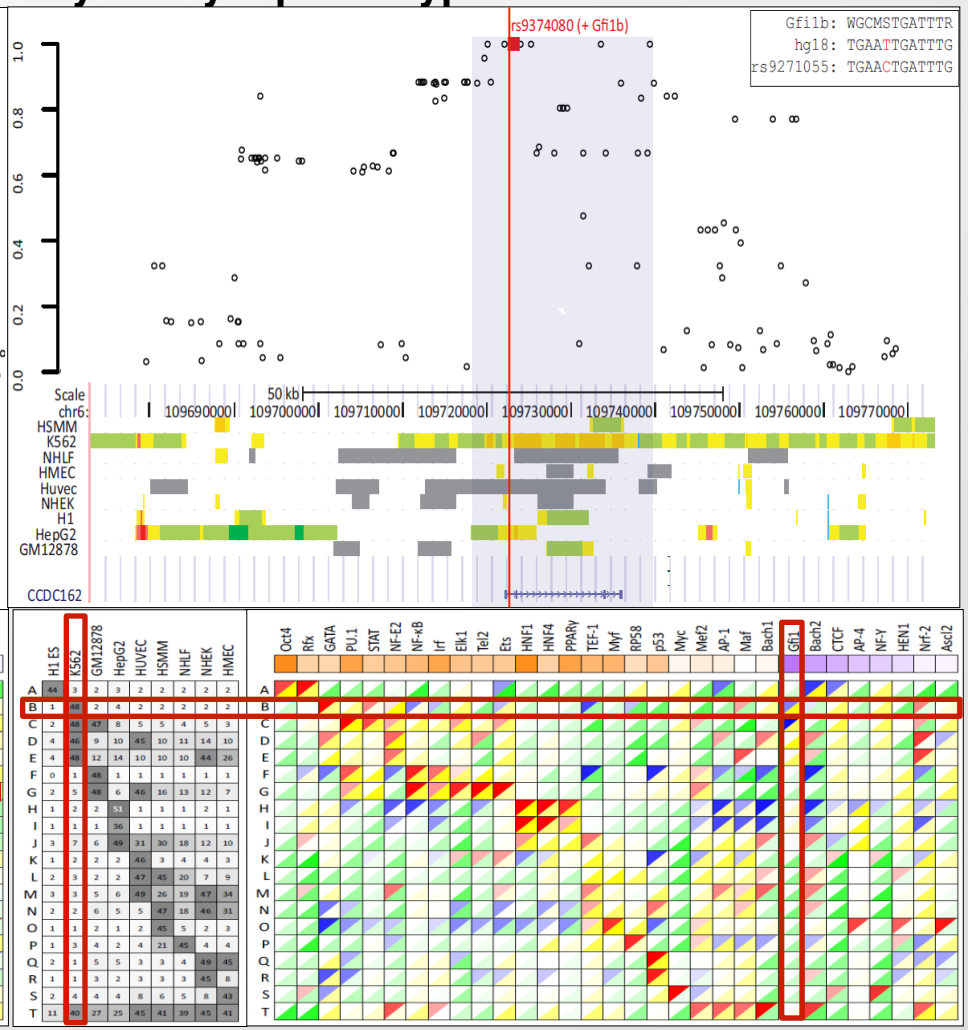
- **Disease-associated SNPs enriched for enhancers in relevant cell types**
- **E.g. lupus SNP in GM enhancer disrupts Ets1 predicted activator**

Mechanistic predictions for top disease-associated SNPs

Lupus erythromatosus in GM lymphoblastoid



Erythrocyte phenotypes in K562 leukemia cells



Disrupt activator Ets-1 motif

➔ **Loss of GM-specific activation**

➔ **Loss of enhancer function**

➔ **Loss of HLA-DRB1 expression**

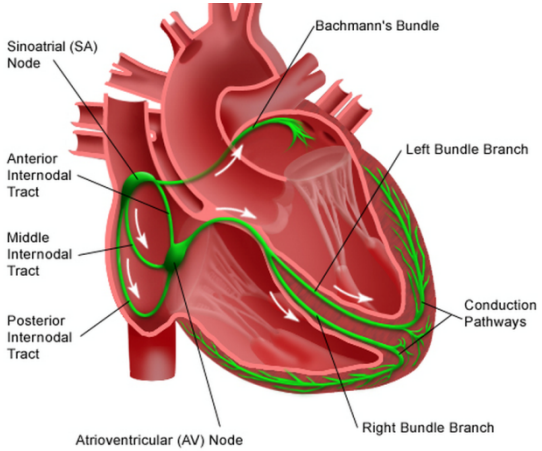
Creation of repressor Gfi1 motif

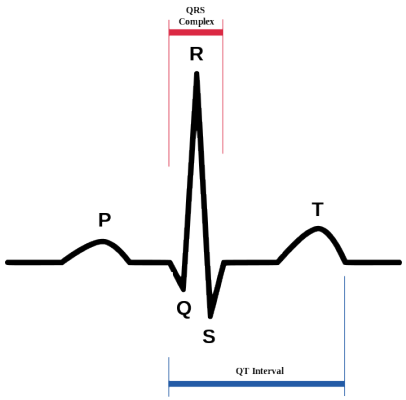
➔ **Gain K562-specific repression**

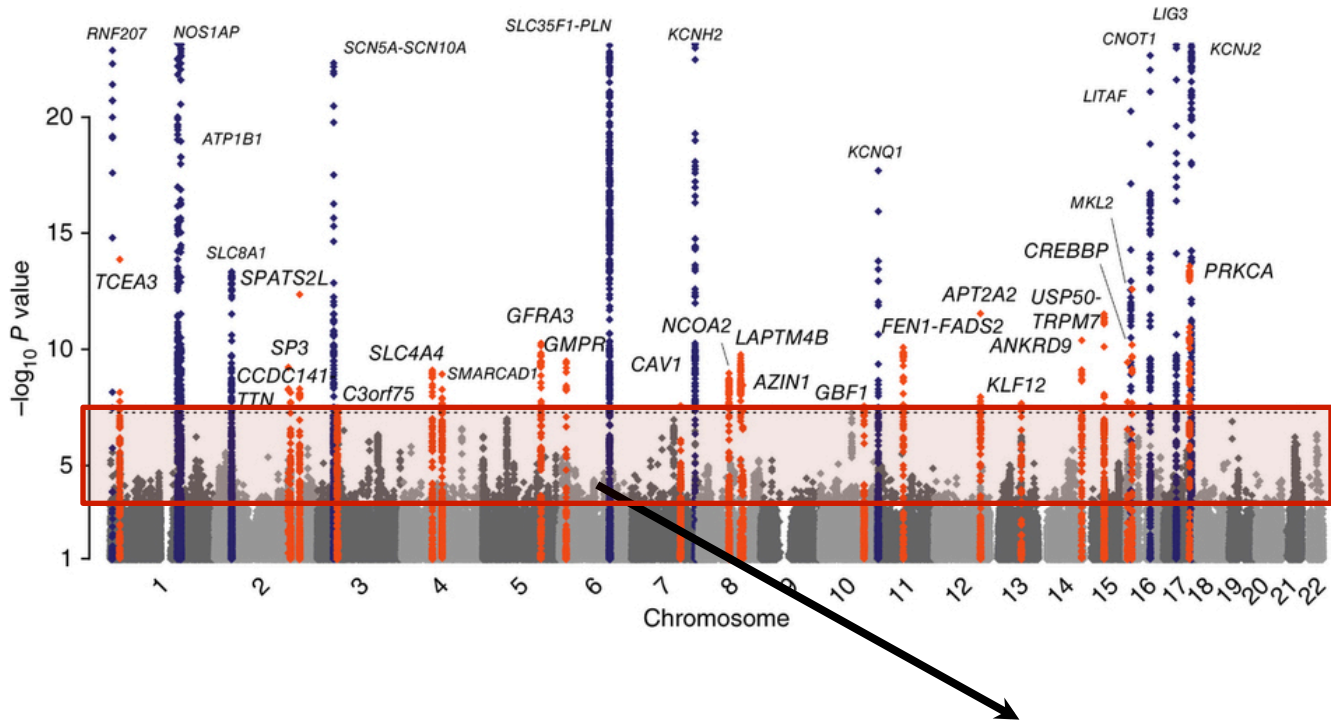
➔ **Loss of enhancer function**

➔ **Loss of CCDC162 expression**

Characterizing sub-threshold variants in heart arrhythmia





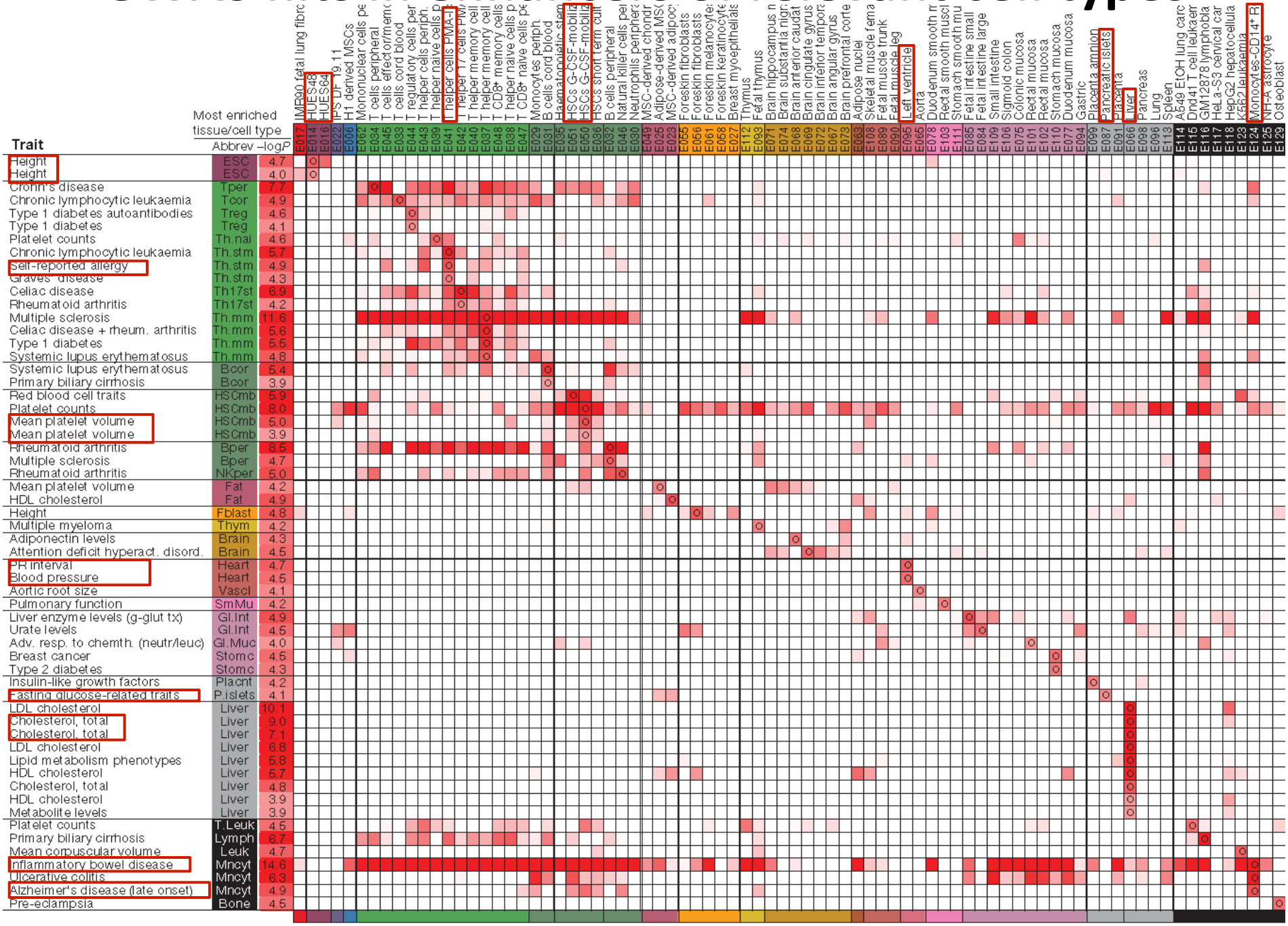


Focus on sub-threshold variants (e.g. rs1743292 $P=10^{-4.2}$)

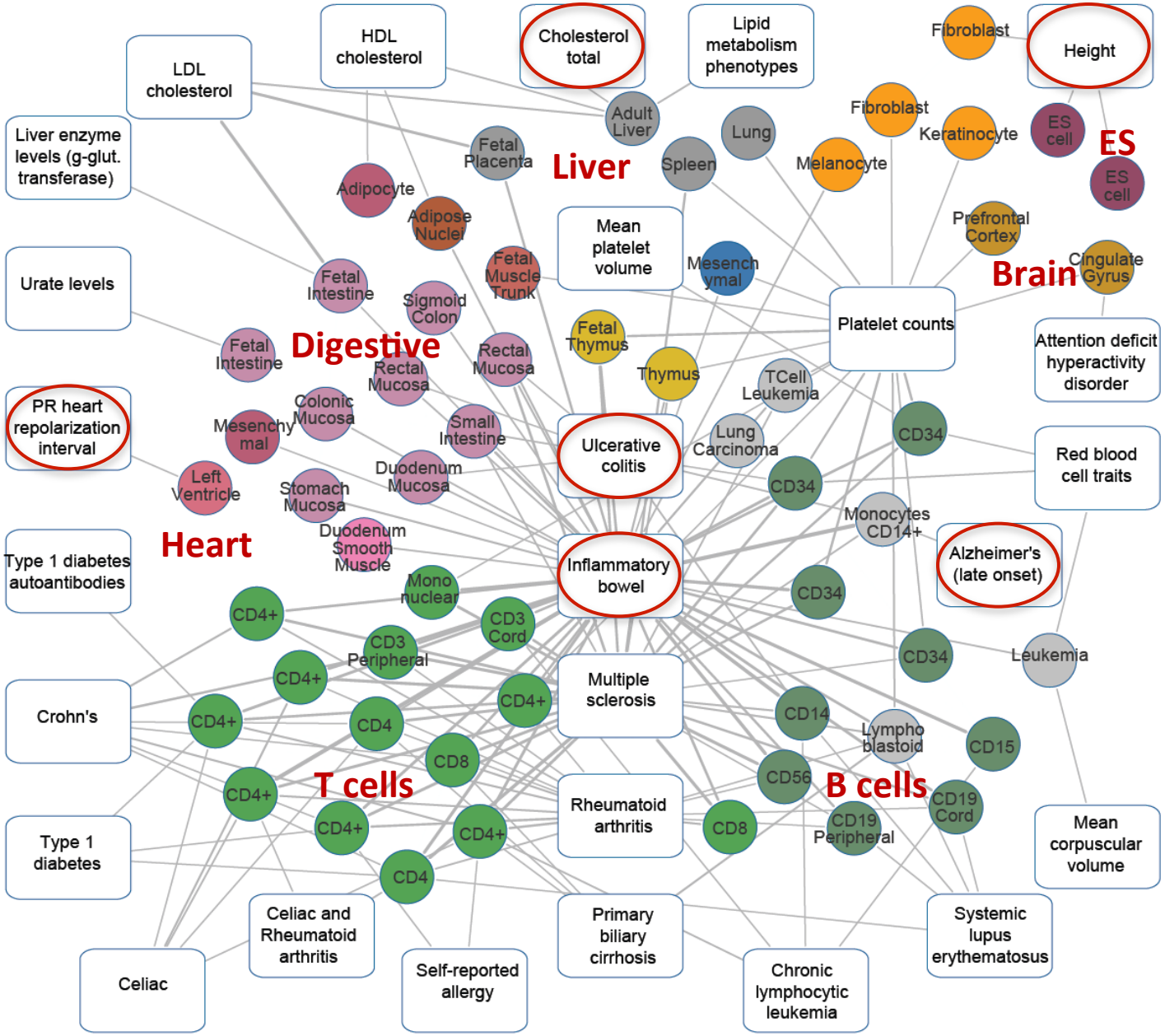
Trait: QRS/QT interval

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

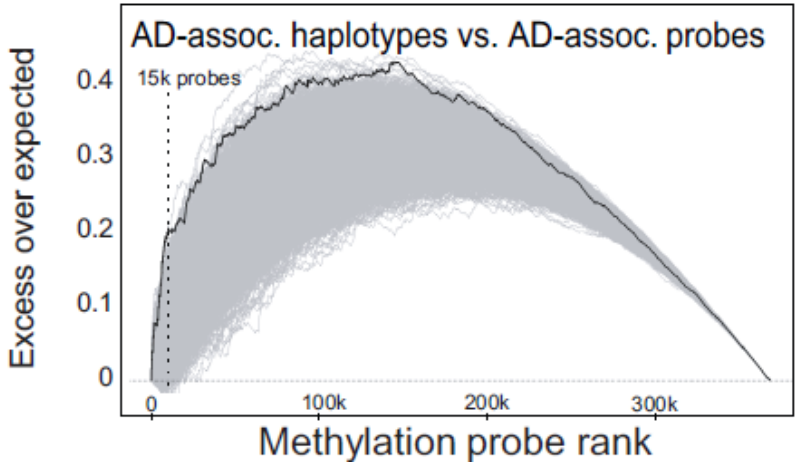
GWAS hits in enhancers of relevant cell types



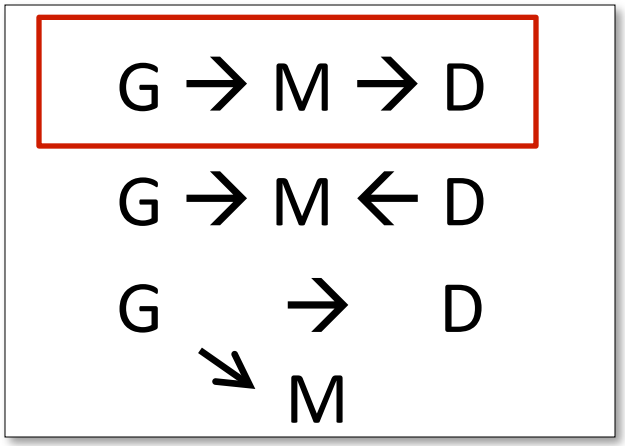
Linking traits to their relevant cell/tissue types



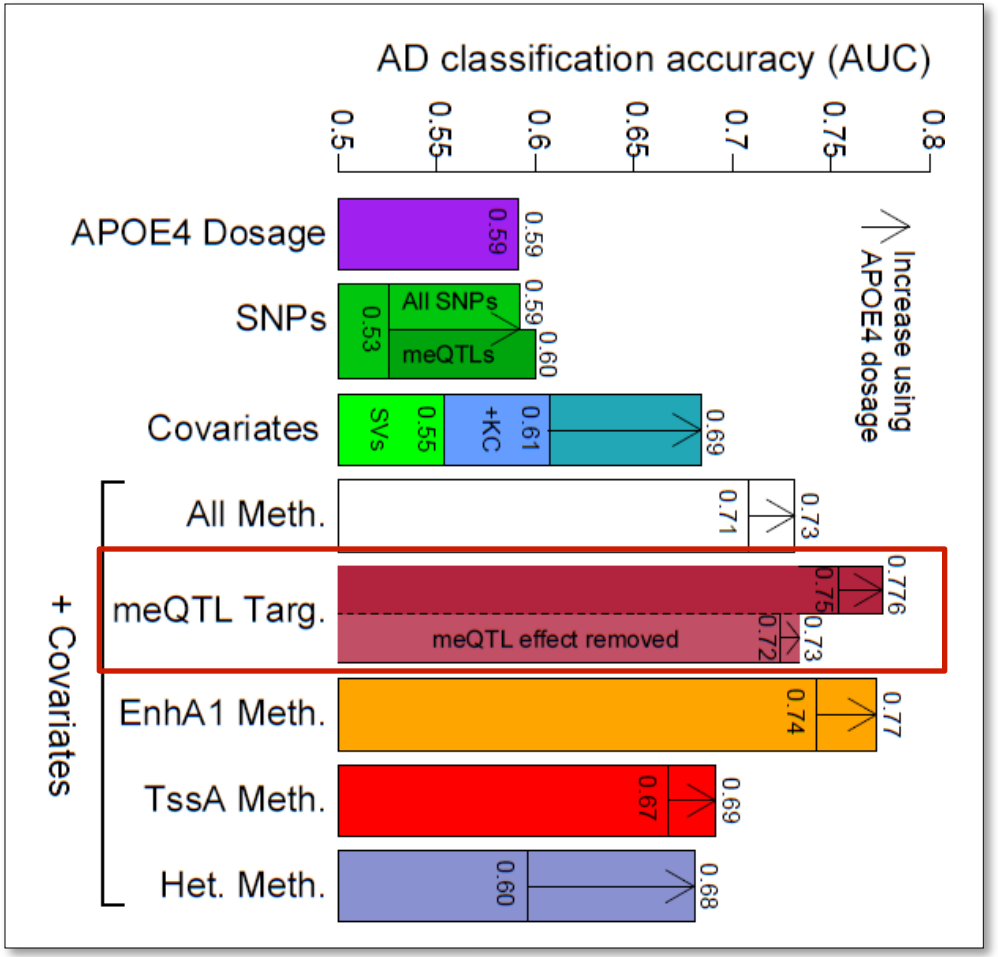
Methylation differences a causal component of AD



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

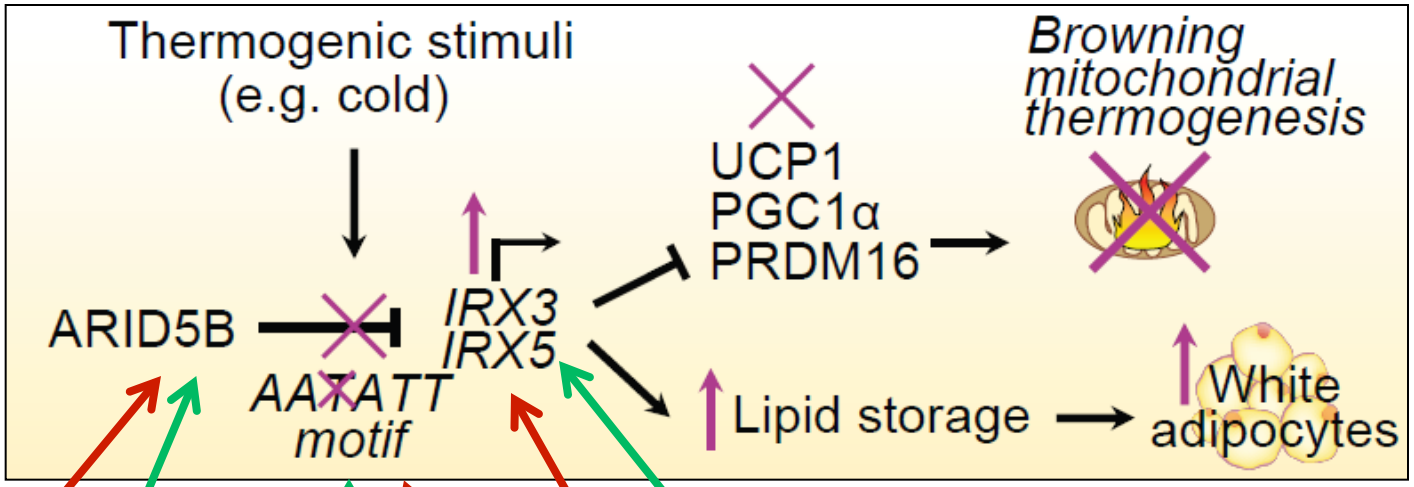


Set-wise causality testing



AD predictive power reduced after removing meQTL effect

Uncovering the molecular basis of top obesity gene



ARID5B KD (obesity)

ARID5B OE (anti-obesity)

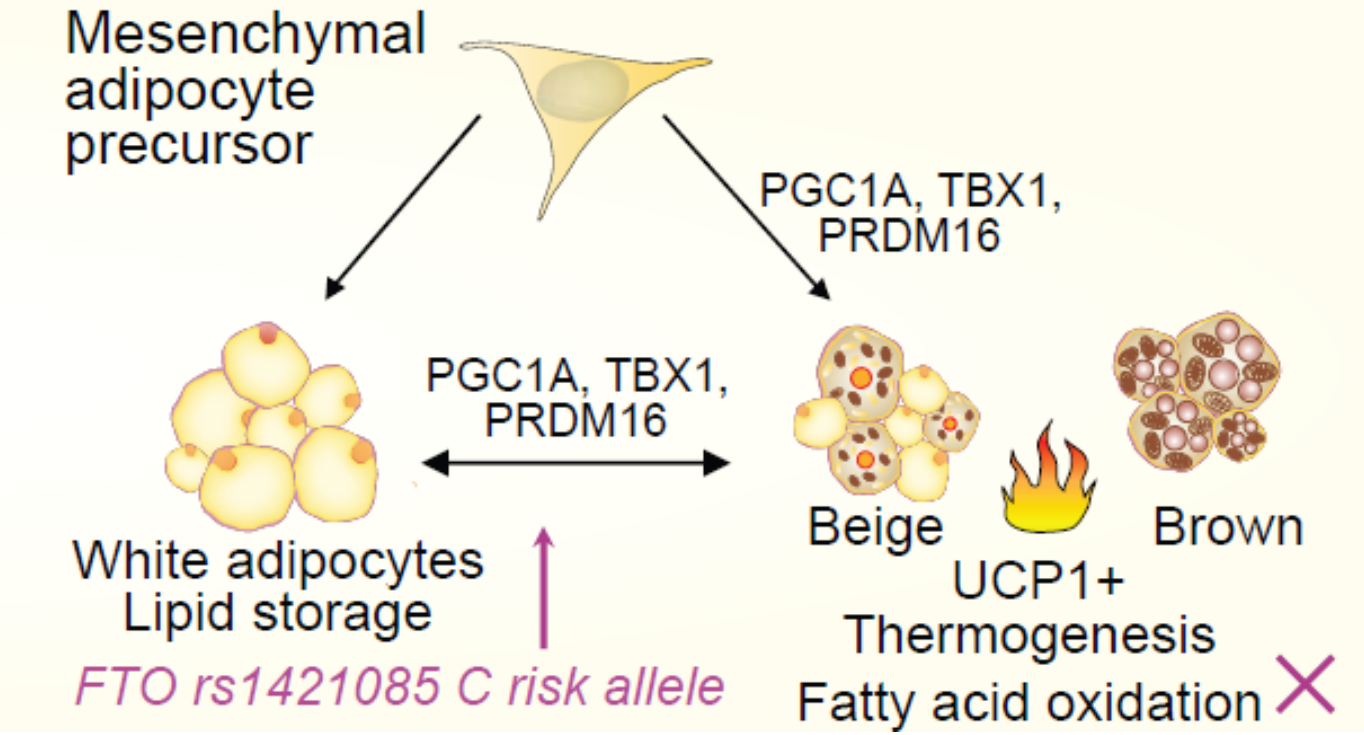
IRX3, IRX5 knock-down ★ (anti-obesity phenotypes)

IRX3, IRX5 overexpression (pro-obesity phenotypes)

★ C-to-T motif rescue (anti-obesity phenotypes)

T-to-C motif disruption (pro-obesity phenotypes)

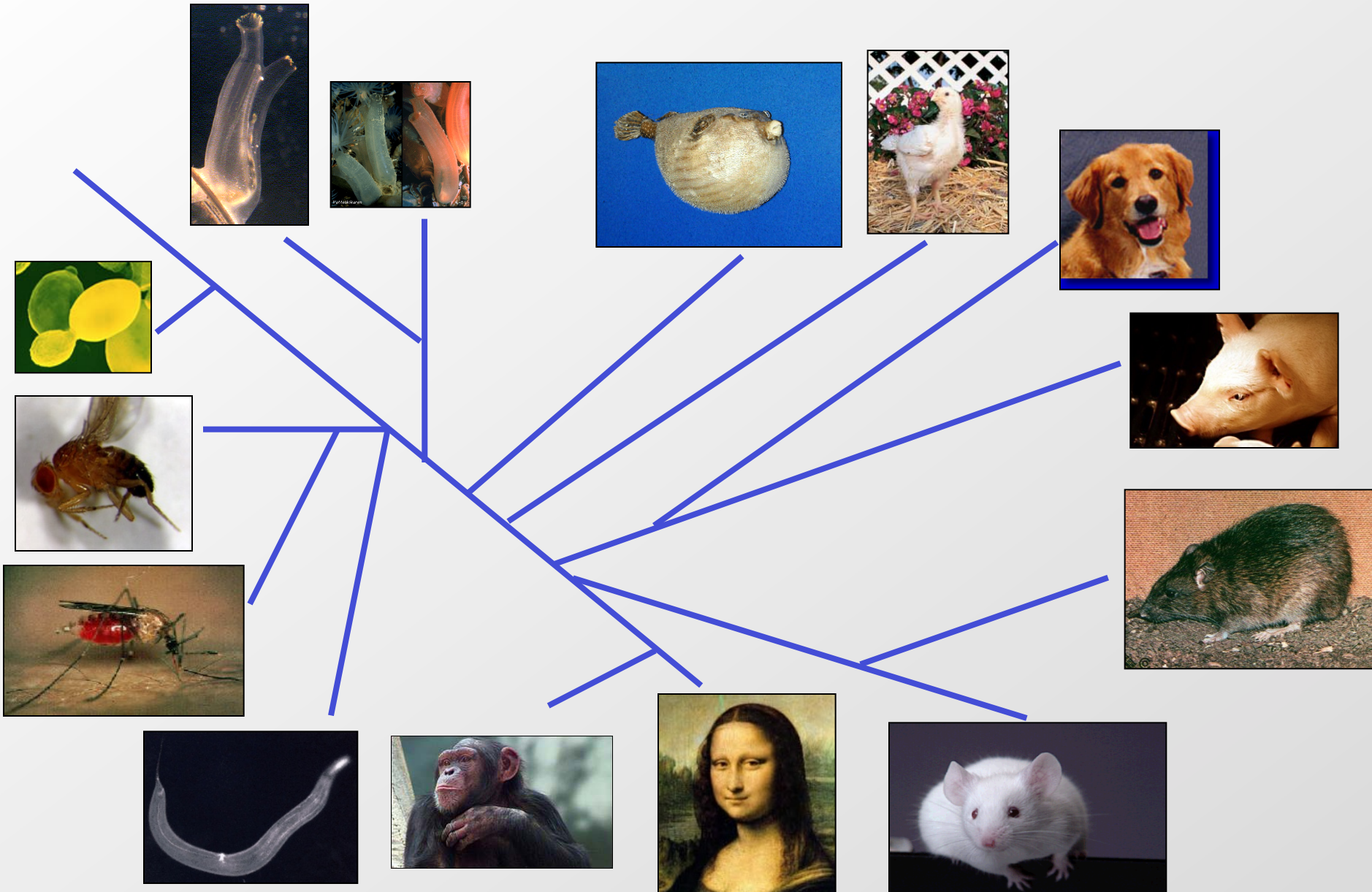
Model: beige ⇔ white adipocyte development



Shift therapeutic focus from brain to adipocytes

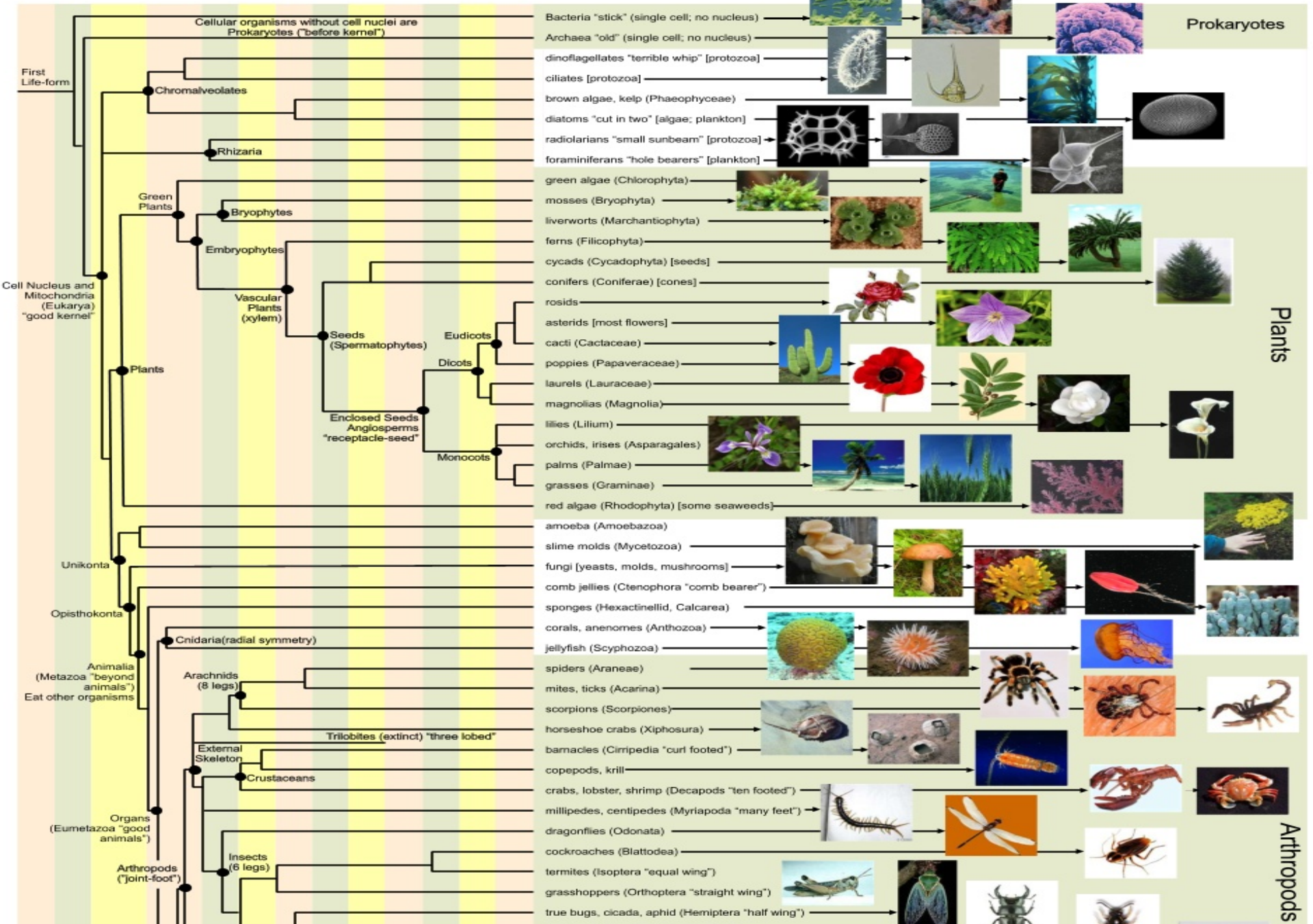
Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Mon 9/23**	PS1 out on:L1-L5	1	Thu, Sep 5	Introduction		L1	Algorithms, Machine Learning, Networks, Course Overview	1
			Fri, Sep 6			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 10	Module I:	Foundations	L2	Dynamic Programming, Reusing computation, Iterative Functions, Exponential / Poly	2,3
			Thu, Sep 12			L3	Database search, Rapid string matching, Hashing	3
			Fri, Sep 13			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 17		Frontiers	L4	HMMs1: Evaluation, Parsing, posterior decoding, learning, HMM architectures	7,8
	due		Thu, Sep 19			L5	HMMs2: Applications, architectures, memory, gene finding, chromatin states	7,8
	Mon 9/23		Fri, Sep 20				No Classes - Student Holiday	
Find prev project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Mon 10/7**	PS2 out on:L6-R4	4	Tue, Sep 24	Module II:	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 26			L7	RNA structure and function. RNA world, RNA-seq, transcript structure, RNA folding	14,15
			Fri, Sep 27			R3	Recitation 3: Supervised Learning and Random Forest Classification	
			Fri, Sep 27	cts, self introductions, mentor intro, example projects, teamwork 32D-507				
		5	Tue, Oct 1		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 3			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due		Fri, Oct 4			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
	Mon 10/7		Fri, Oct 4	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Thu 10/17. Presented on Fri 10/18**	PS3 out on:L10-R6	6	Tue, Oct 8	Module III:	Foundations	L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Thu, Oct 10			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Fri, Oct 11			R5	Recitation 5: Communication Lab	
		7	Tue, Oct 15				No Classes - Columbus Day Holiday	
			Thu, Oct 17		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due		Fri, Oct 18			R6	Recitation 6: Motif Discovery, WEEDER, In vitro Motif Discovery - PBMs, Selex	
	Mon 10/21		Fri, Oct 18	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507				
Evaluate/discuss three peer proposals, NIH review format. **Reviews back Mon 10/28**	PS4 out on:L13-R8	8	Tue, Oct 22	Module IV:	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 24			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 25			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 25	Panel Review: Discuss Peer Projects. Feedback sent out from group reviews. 32D-463 (Star).				
		9	Tue, Oct 29		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/7**	due		Thu, Oct 31			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
	Mon 11/4		Fri, Nov 1			R8	Recitation 8: Phylogenetic distance metrics, Coalescent Process	
	PS5 out on:L17-R10	10	Tue, Nov 5	Module V:	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 7			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 8				No Recitation, Veterans Day	
Continue making subst. progress on proposed milstones.Write outline of final report. **Midcourse report due Mon 11/25**		11	Tue, Nov 12		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
	due		Thu, Nov 14			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
	Fri 11/15		Fri, Nov 15			R9	Recitation 9: Quiz Review	
	No more psets! (work on your final project)	12	Tue, Nov 19	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 21			L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Fri, Nov 22			R10	Recitation 10: Project Feedback, results, interpretation, directions	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/8**		13	Tue, Nov 26	Module VI:	Frontiers	L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Thu, Nov 28				No lecture, thanksgiving break - Thu Nov 28, 2019	
			Fri, Nov 29				No recitation, thanksgiving break	
		14	Tue, Dec 3			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 5			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 6			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Presentations on Tue 12/10**		15	Tue, Dec 10			L25	Final Presentations - Part I (1pm). 32-141 (Classroom)	
			Tue, Dec 10			L25	Final Presentations - Part I (2:30pm). 32D-463 (Star)	

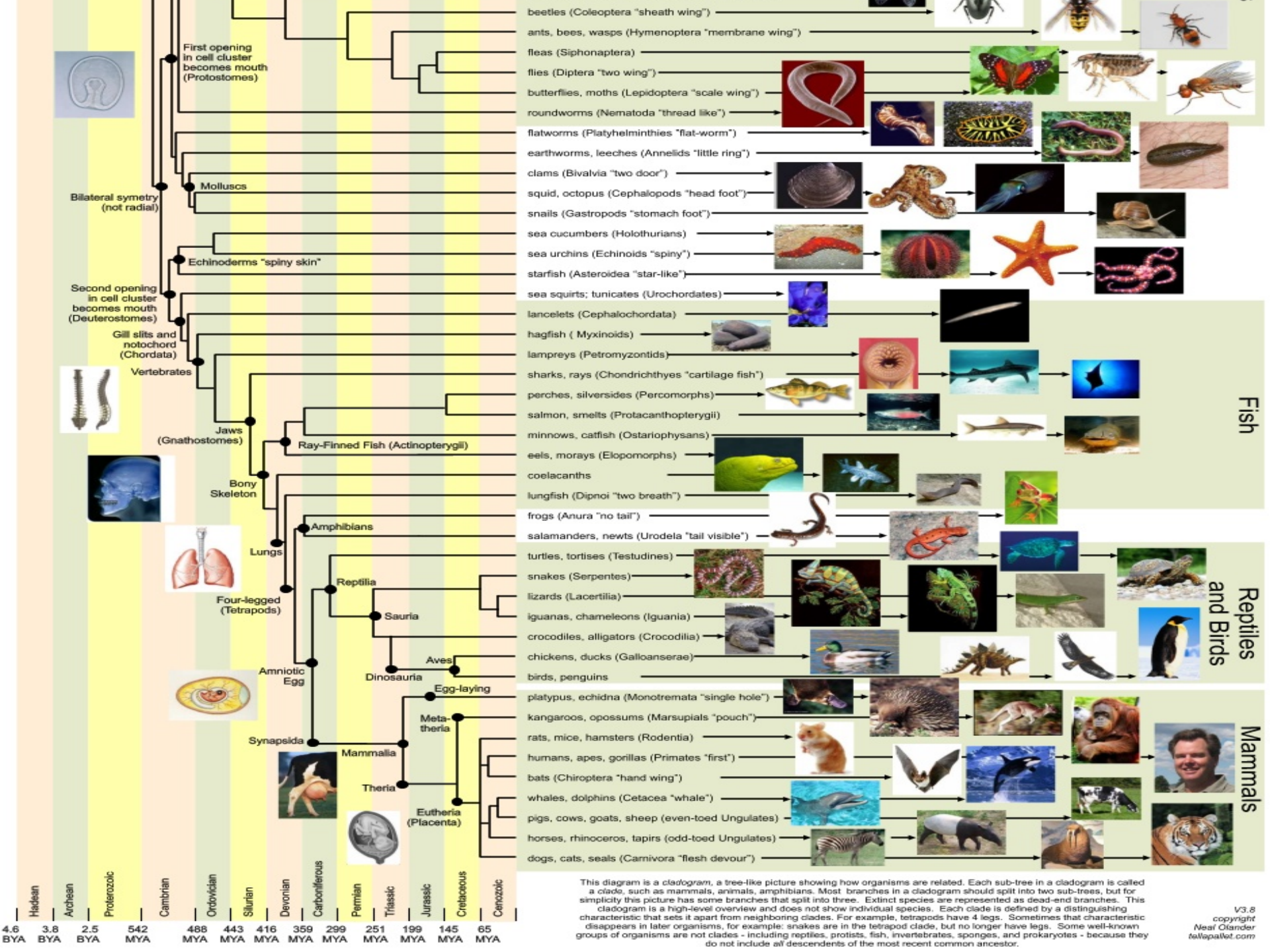
Alignment: all species/genes share common ancestry



Slide credit: Serafim Batzoglou

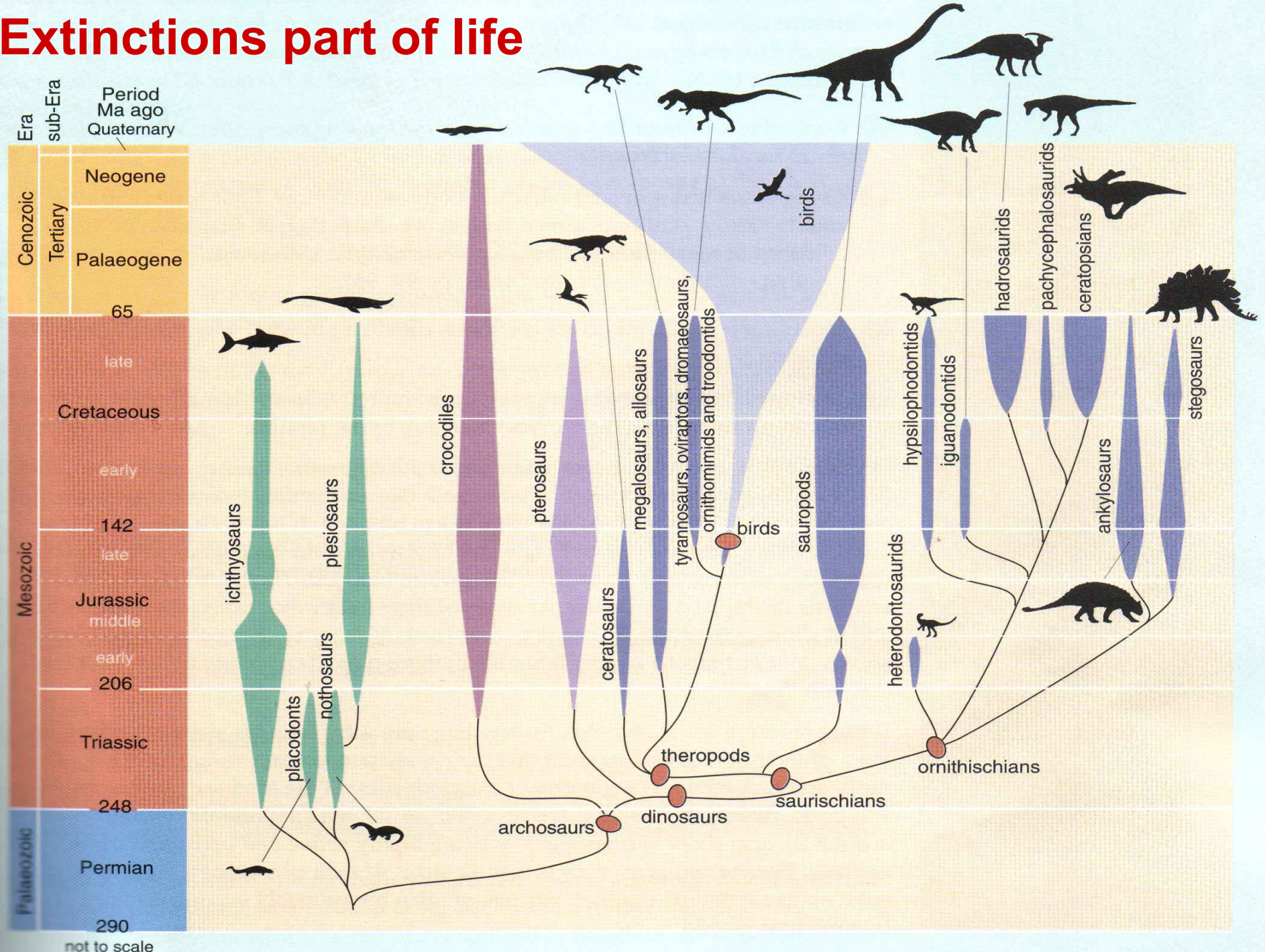
Tree of Life


Cellular organisms without cell nuclei are Prokaryotes ("before kernel")
First Life-form
Chromalveolates
Rhizaria
Green Plants
Bryophytes
Embryophytes
Vascular Plants (xylem)
Seeds (Spermatophytes)
Eudicots
Dicots
Enclosed Seeds Angiosperms "receptacle-seed"
Monocots
Cell Nucleus and Mitochondria (Eukarya) "good kernel"
Plants
Unikonta
Opisthokonta
Cnidaria(radial symmetry)
Animalia (Metazoa "beyond animals") Eat other organisms
Arachnids (8 legs)
Trilobites (extinct) "three lobed"
External Skeleton
Crustaceans
Organs (Eumetazoa "good animals")
Insects (6 legs)
Arthropods ("joint-foot")
Bacteria "stick" (single cell; no nucleus)
Archaea "old" (single cell; no nucleus)
Prokaryotes
dinoflagellates "terrible whip" [protozoa]
ciliates [protozoa]
brown algae, kelp (Phaeophyceae)
diatoms "cut in two" [algae; plankton]
radiolarians "small sunbeam" [protozoa]
foraminiferans "hole bearers" [plankton]
green algae (Chlorophyta)
mosses (Bryophyta)
liverworts (Marchantiophyta)
ferns (Filicophyta)
cycads (Cycadophyta) [seeds]
conifers (Coniferae) [cones]
rosids
asterids [most flowers]
cacti (Cactaceae)
poppies (Papaveraceae)
laurels (Lauraceae)
magnolias (Magnolia)
lilies (Lilium)
orchids, irises (Asparagales)
palms (Palmae)
grasses (Graminae)
red algae (Rhodophyta) [some seaweeds]
Plants
amoeba (Amoebazoa)
slime molds (Mycetozoa)
fungi [yeasts, molds, mushrooms]
comb jellies (Ctenophora "comb bearer")
sponges (Hexactinellid, Calcarea)
corals, anenomes (Anthozoa)
jellyfish (Scyphozoa)
spiders (Araneae)
mites, ticks (Acarina)
scorpions (Scorpiones)
horseshoe crabs (Xiphosura)
barnacles (Cirripedia "curl footed")
copepods, krill
crabs, lobster, shrimp (Decapods "ten footed")
millipedes, centipedes (Myriapoda "many feet")
dragonflies (Odonata)
cockroaches (Blattodea)
termites (Isoptera "equal wing")
grasshoppers (Orthoptera "straight wing")
true bugs, cicada, aphid (Hemiptera "half wing")
Arthropods



This diagram is a *cladogram*, a tree-like picture showing how organisms are related. Each sub-tree in a cladogram is called a *clade*, such as mammals, animals, amphibians. Most branches in a cladogram should split into two sub-trees, but for simplicity this picture has some branches that split into three. Extinct species are represented as dead-end branches. This cladogram is a high-level overview and does not show individual species. Each clade is defined by a distinguishing characteristic that sets it apart from neighboring clades. For example, tetrapods have 4 legs. Sometimes that characteristic disappears in later organisms, for example: snakes are in the tetrapod clade, but no longer have legs. Some well-known groups of organisms are not clades - including reptiles, protists, fish, invertebrates, sponges, and prokaryotes - because they do not include all descendents of the most recent common ancestor.

Extinctions part of life



Phylogenetics

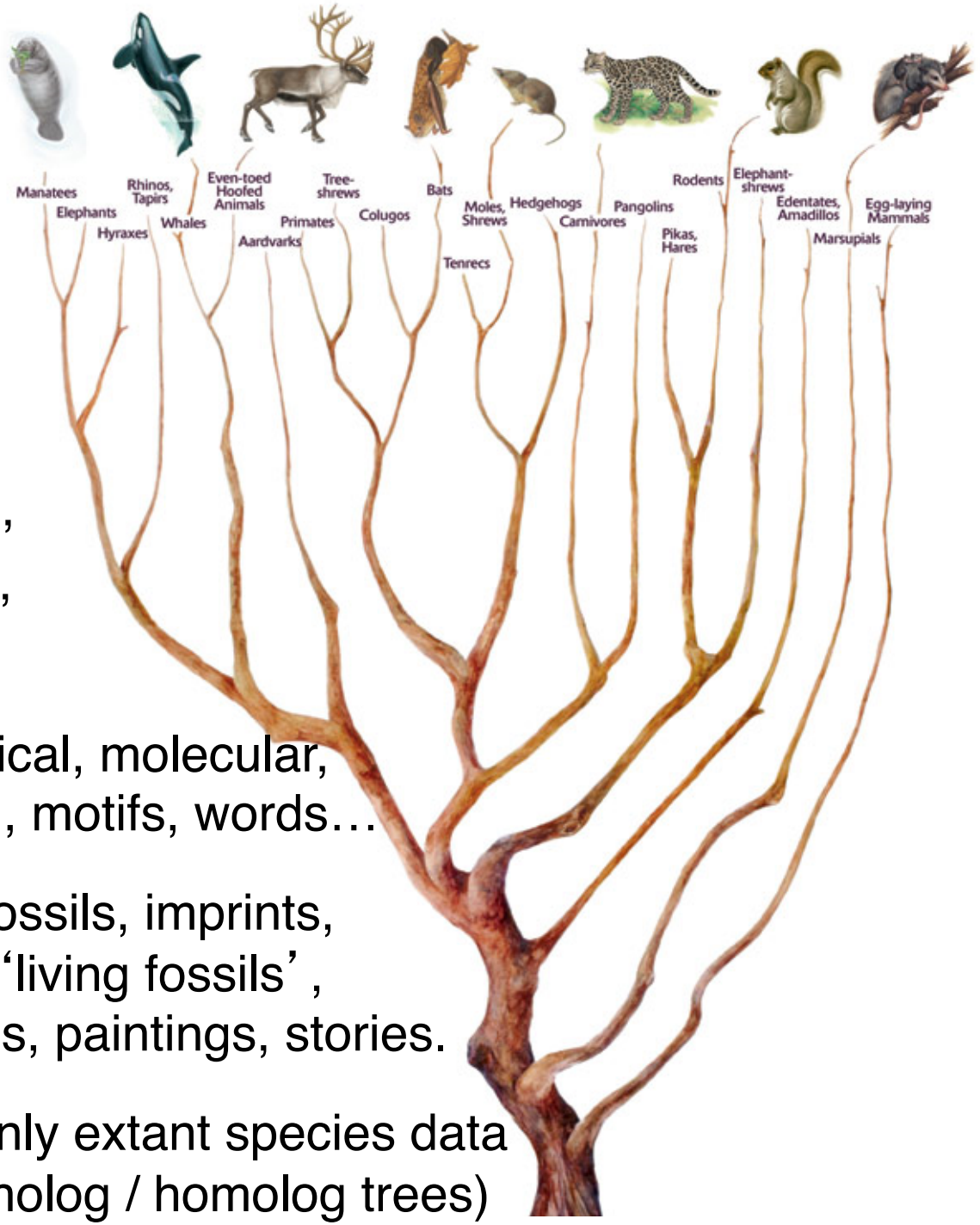
General Problem:
Infer complete ancestry of a set of '**objects**' based on knowledge of their '**traits**'

'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…

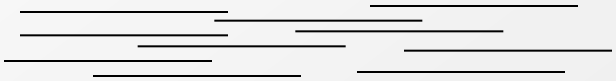
Historical record varies: Fossils, imprints, timing of geological events, 'living fossils', sequencing of extinct species, paintings, stories.

Today: Phylogenies using only extant species data
➔ **gene trees** (paralog / ortholog / homolog trees)



Challenges in Computational Biology

4 Genome Assembly



5 Regulatory motif discovery

1 Gene Finding

DNA

2 Sequence alignment

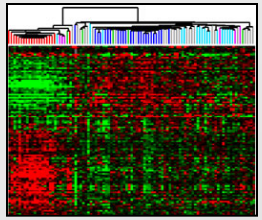
6 Comparative Genomics

7 Evolutionary Theory

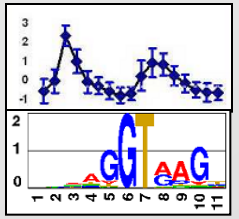
3 Database lookup

8 Gene expression analysis

RNA transcript

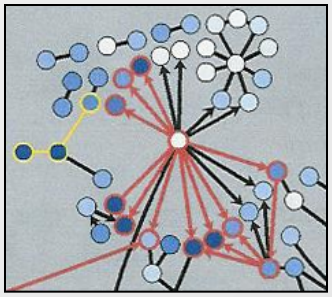


9 Cluster discovery



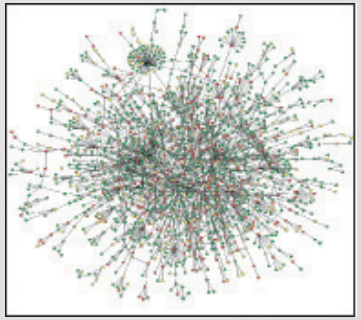
10 Gibbs sampling

11 Protein network analysis



12 Metabolic modelling

13 Emerging network properties



Project	Psets	Week	Date	Topic		Lec	Topic	Read*
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Mon 9/23**	PS1 out on:L1-L5	1	Thu, Sep 5	Introduction		L1	Algorithms, Machine Learning, Networks, Course Overview	1
			Fri, Sep 6			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 10	Module I:	Foundations	L2	Dynamic Programming, Reusing computation, Iterative Functions, Exponential / Poly	2,3
			Thu, Sep 12			L3	Database search, Rapid string matching, Hashing	3
			Fri, Sep 13			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 17		Frontiers	L4	HMMs1: Evaluation, Parsing, posterior decoding, learning, HMM architectures	7,8
	due		Thu, Sep 19			L5	HMMs2: Applications, architectures, memory, gene finding, chromatin states	7,8
	Mon 9/23		Fri, Sep 20				No Classes - Student Holiday	
Find prev project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Project area/team due Mon 10/7**	PS2 out on:L6-R4	4	Tue, Sep 24	Module II:	Foundations	L6	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Thu, Sep 26			L7	RNA structure and function. RNA world, RNA-seq, transcript structure, RNA folding	14,15
			Fri, Sep 27			R3	Recitation 3: Supervised Learning and Random Forest Classification	
			Fri, Sep 27	cts, self introductions, mentor intro, example projects, teamwork 32D-507				
		5	Tue, Oct 1		Frontiers	L8	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 3			L9	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	22
	due		Fri, Oct 4			R4	Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute	
	Mon 10/7		Fri, Oct 4	Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507				
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Project proposal due Thu 10/17. Presented on Fri 10/18**	PS3 out on:L10-R6	6	Tue, Oct 8	Module III:	Foundations	L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Thu, Oct 10			L11	Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels	20,21
			Fri, Oct 11			R5	Recitation 5: Communication Lab	
		7	Tue, Oct 15				No Classes - Columbus Day Holiday	
			Thu, Oct 17		Frontiers	L12	Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder	20.7
	due		Fri, Oct 18			R6	Recitation 6: Motif Discovery, WEEDER, In vitro Motif Discovery - PBMs, Selex	
	Mon 10/21		Fri, Oct 18	Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507				
Evaluate/discuss three peer proposals, NIH review format. **Reviews back Mon 10/28**	PS4 out on:L13-R8	8	Tue, Oct 22	Module IV:	Foundations	L13	Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq	30
			Thu, Oct 24			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 25			R7	Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation	
			Fri, Oct 25	Panel Review: Discuss Peer Projects. Feedback sent out from group reviews. 32D-463 (Star).				
		9	Tue, Oct 29		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS	32
Address peer evaluations, revise aims, scope, list of final deliverables / goals. **Response due Thu 11/7**	due		Thu, Oct 31			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	31
	Mon 11/4		Fri, Nov 1			R8	Recitation 8: Phylogenetic distance metrics, Coalescent Process	
	PS5 out on:L17-R10	10	Tue, Nov 5	Module V:	Foundations	L17	Comparative genomics and evolutionary signatures	4
			Thu, Nov 7			L18	Genome Scale Evolution, Genome Duplication	4,5.7
			Fri, Nov 8				No Recitation, Veterans Day	
Continue making subst. progress on proposed milstones.Write outline of final report. **Midcourse report due Mon 11/25**	due	11	Tue, Nov 12		Frontiers	L19	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
			Thu, Nov 14			L20	Phylogenomics: Gene/species trees, reconciliation, coalescent, ARGs	28
	Fri 11/15		Fri, Nov 15			R9	Recitation 9: Quiz Review	
	No more psets! (work on your final project)	12	Tue, Nov 19	Quiz	Foundations	Quiz	In Class Quiz (the only quiz - the class has no final exam) - covers L1-L20,R1-R9	
			Thu, Nov 21			L21	Single-cell genomics: technology, analysis, microfluidics, applications, insights	37
			Fri, Nov 22			R10	Recitation 10: Project Feedback, results, interpretation, directions	
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Written report due Sun 12/8**		13	Tue, Nov 26	Module VI:	Frontiers	L22	Mining human phenotypes, PheWAS, UK Biobank, meta-phenotypes+imputation	34
			Thu, Nov 28				No lecture, thanksgiving break - Thu Nov 28, 2019	
			Fri, Nov 29				No recitation, thanksgiving break	
		14	Tue, Dec 3			L23	Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface	35
			Thu, Dec 5			L24	Genome Engineering with CRISPR/Cas9 and related technologies	36
			Fri, Dec 6			R11	Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills	
Conference format slide pres. **Presentations on Tue 12/10**		15	Tue, Dec 10			L25	Final Presentations - Part I (1pm). 32-141 (Classroom)	
			Tue, Dec 10			L25	Final Presentations - Part I (2:30pm). 32D-463 (Star)	

Please provide feedback:
https://goo.gl/rV5XJi