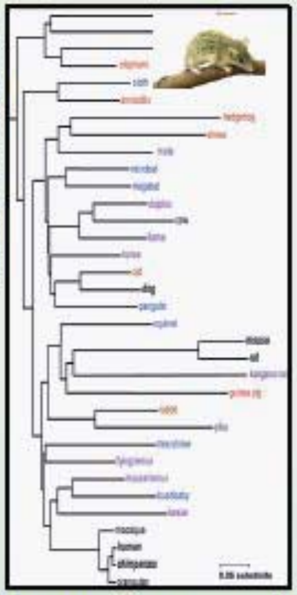
6.047 (U) - 6.878 (G, TQE) - HST.507

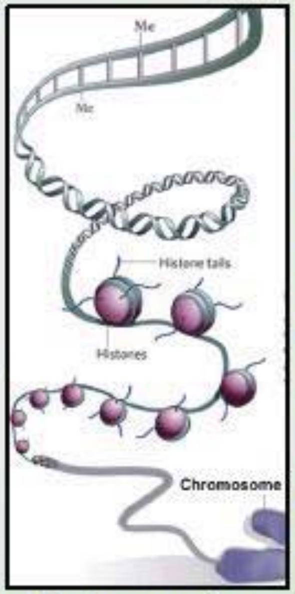
Fall 2015

Computational Biology

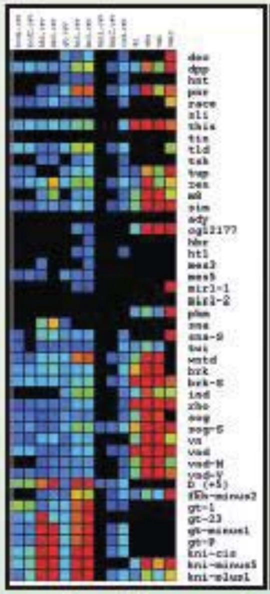
Genomes - Networks - Evolution

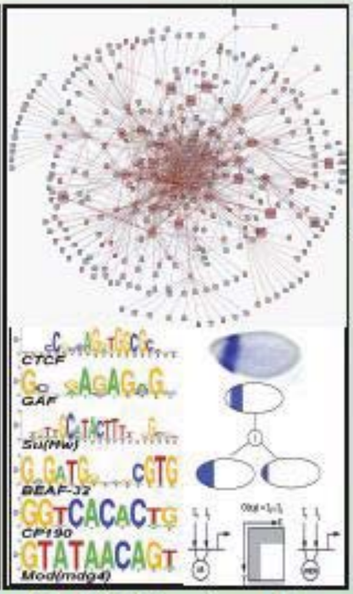
Learn about:


Comparative genomics

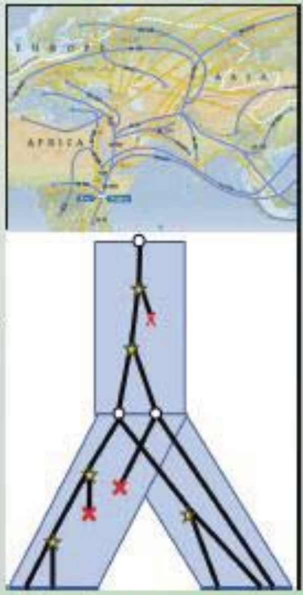


Epigenomics

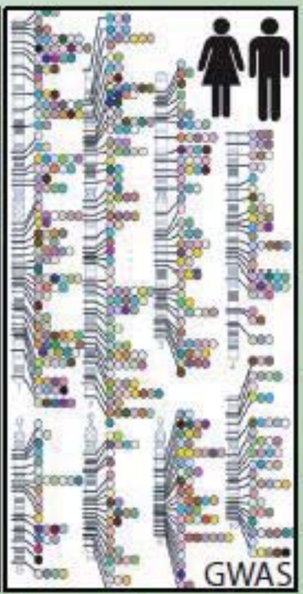

Functional genomics



Motifs & networks



Phylogenomics



Personal genomics

... and much more

Prof. Manolis Kellis - MIT / CSAIL / Broad Institute

Covers the algorithmic and machine learning foundations of computational biology combining theory with practice. We cover both foundational topics in computational biology, and current research frontiers. We study fundamental techniques, recent advances in the field, and work directly with current large-scale biological datasets.

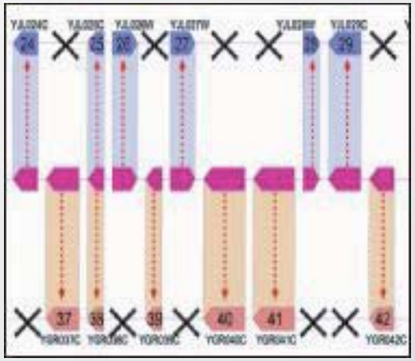
Computational Biology:
Genomes, Networks, Evolution

MIT 6.047 / 6.878
HSPH IMI.231
HST.507

Prof. Manolis Kellis

Rapid database search

Protein interaction network



Genome duplication

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, etc

Course Information

- Lectures
  - TR 1pm – 2:30
- Recitations:
  - On Friday at 3pm
  - Recitations at MIT (HST/HSPH students can join)
  - All handouts, lectures, notes, etc will be posted here.

- Course calendar:
  - On Google, add public calendar: “6.047 Lectures” and “6.047 due dates”

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)

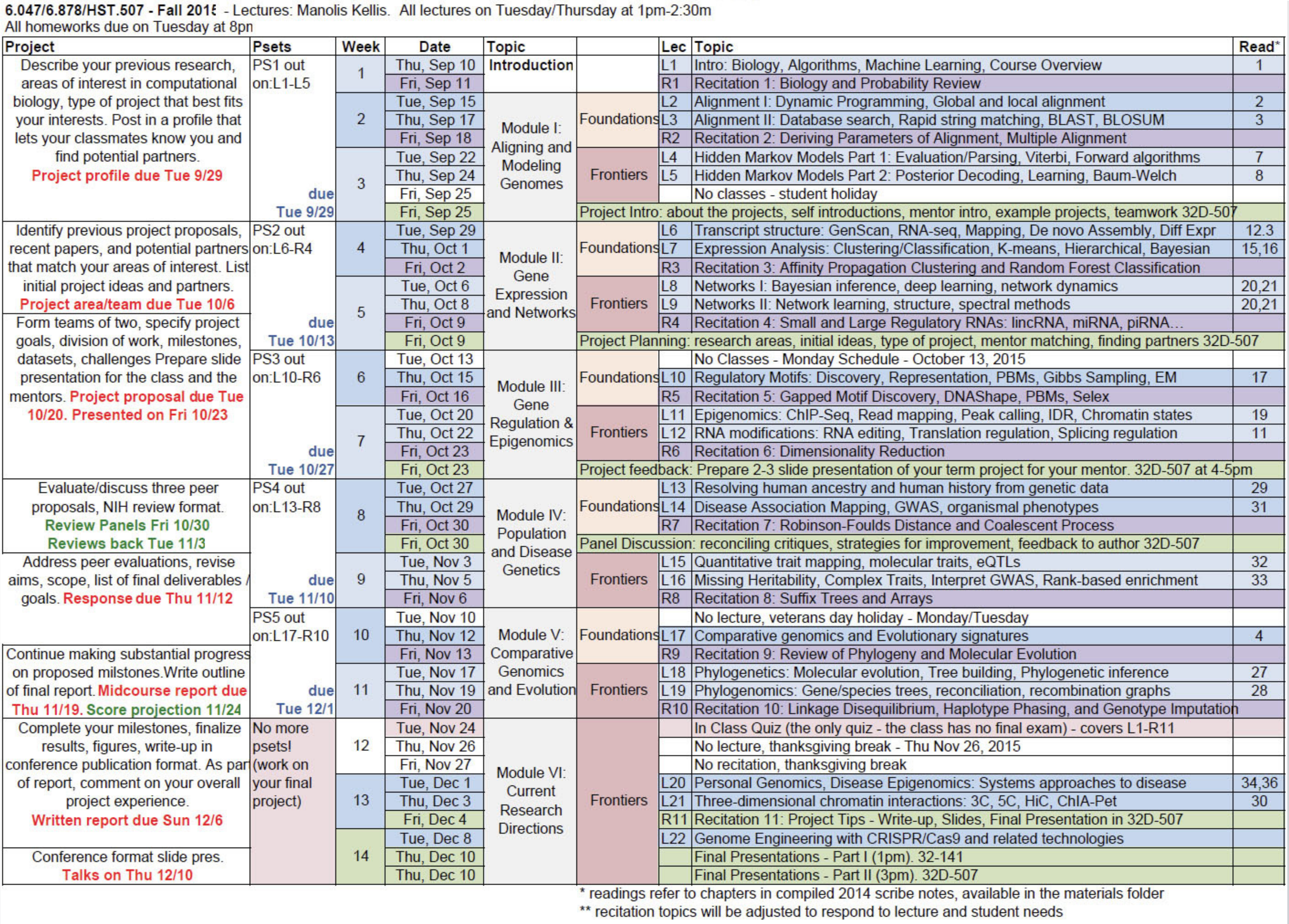
Course content

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



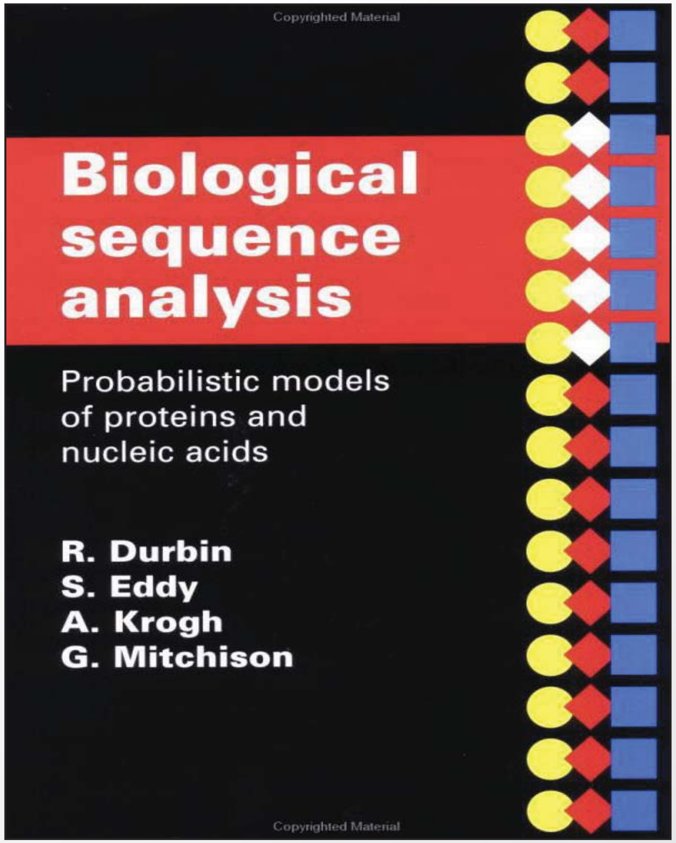
6.047/6.878/HST.507 - Fall 2015 - Lectures: Manolis Kellis. All lectures on Tuesday/Thursday at 1pm-2:30m
All homeworks due on Tuesday at 8pm

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 9/29**	PS1 out on:L1-L5	1	Thu, Sep 10	**Introduction**		L1	Intro: Biology, Algorithms, Machine Learning, Course Overview	1
			Fri, Sep 11			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 15	Module I: Aligning and Modeling Genomes	Foundations	L2	Alignment I: Dynamic Programming, Global and local alignment	2
			Thu, Sep 17			L3	Alignment II: Database search, Rapid string matching, BLAST, BLOSUM	3
			Fri, Sep 18			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 22		Frontiers	L4	Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms	7
			Thu, Sep 24			L5	Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch	8
	due		Fri, Sep 25				No classes - student holiday	
	Tue 9/29		Fri, 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 10/6**	PS2 out on:L6-R4	4	Tue, Sep 29	Module II: Gene Expression and Networks	Foundations	L6	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr	12.3
			Thu, Oct 1			L7	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Fri, Oct 2			R3	Recitation 3: Affinity Propagation Clustering and Random Forest Classification	
		5	Tue, Oct 6		Frontiers	L8	Networks I: Bayesian inference, deep learning, network dynamics	20,21
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 10/20. Presented on Fri 10/23**			Thu, Oct 8			L9	Networks II: Network learning, structure, spectral methods	20,21
	due		Fri, Oct 9			R4	Recitation 4: Small and Large Regulatory RNAs: lincRNA, miRNA, piRNA…	
	Tue 10/13		Fri, Oct 9		Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507			
	PS3 out on:L10-R6	6	Tue, Oct 13	Module III: Gene Regulation & Epigenomics	Foundations		No Classes - Monday Schedule - October 13, 2015	
			Thu, Oct 15			L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Fri, Oct 16			R5	Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex	
		7	Tue, Oct 20		Frontiers	L11	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 22			L12	RNA modifications: RNA editing, Translation regulation, Splicing regulation	11
	due		Fri, Oct 23			R6	Recitation 6: Dimensionality Reduction	
	Tue 10/27		Fri, Oct 23		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 10/30** **Reviews back Tue 11/3**	PS4 out on:L13-R8	8	Tue, Oct 27	Module IV: Population and Disease Genetics	Foundations	L13	Resolving human ancestry and human history from genetic data	29
			Thu, Oct 29			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 30			R7	Recitation 7: Robinson-Foulds Distance and Coalescent Process	
			Fri, Oct 30		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 11/12**		9	Tue, Nov 3		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs	32
	due		Thu, Nov 5			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	33
	Tue 11/10		Fri, Nov 6			R8	Recitation 8: Suffix Trees and Arrays	
	PS5 out on:L17-R10	10	Tue, Nov 10	Module V: Comparative Genomics and Evolution	Foundations		No lecture, veterans day holiday - Monday/Tuesday	
			Thu, Nov 12			L17	Comparative genomics and Evolutionary signatures	4
Continue making substantial progress on proposed milstones.Write outline of final report. **Midcourse report due Thu 11/19. Score projection 11/24**			Fri, Nov 13			R9	Recitation 9: Review of Phylogeny and Molecular Evolution	
		11	Tue, Nov 17		Frontiers	L18	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
	due		Thu, Nov 19			L19	Phylogenomics: Gene/species trees, reconciliation, recombination graphs	28
	Tue 12/1		Fri, Nov 20			R10	Recitation 10: Linkage Disequilibrium, Haplotype Phasing, and Genotype Imputation	
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 12/6**	No more psets! (work on your final project)	12	Tue, Nov 24	Module VI: Current Research Directions	Frontiers		In Class Quiz (the only quiz - the class has no final exam) - covers L1-R11	
			Thu, Nov 26				No lecture, thanksgiving break - Thu Nov 26, 2015	
			Fri, Nov 27				No recitation, thanksgiving break	
		13	Tue, Dec 1			L20	Personal Genomics, Disease Epigenomics: Systems approaches to disease	34,36
			Thu, Dec 3			L21	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	30
			Fri, Dec 4			R11	Recitation 11: Project Tips - Write-up, Slides, Final Presentation in 32D-507	
		14	Tue, Dec 8			L22	Genome Engineering with CRISPR/Cas9 and related technologies	
Conference format slide pres. **Talks on Thu 12/10**			Thu, Dec 10				Final Presentations - Part I (1pm). 32-141	
			Thu, Dec 10				Final Presentations - Part II (3pm). 32D-507	

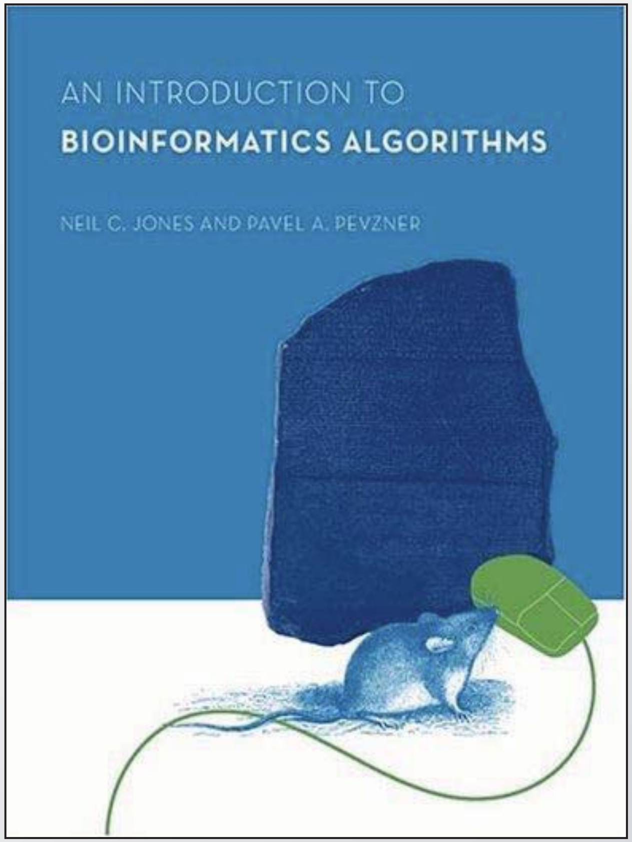
* readings refer to chapters in compiled 2014 scribe notes, available in the materials folder
** recitation topics will be adjusted to respond to lecture and student needs

Textbook / class notes / resources

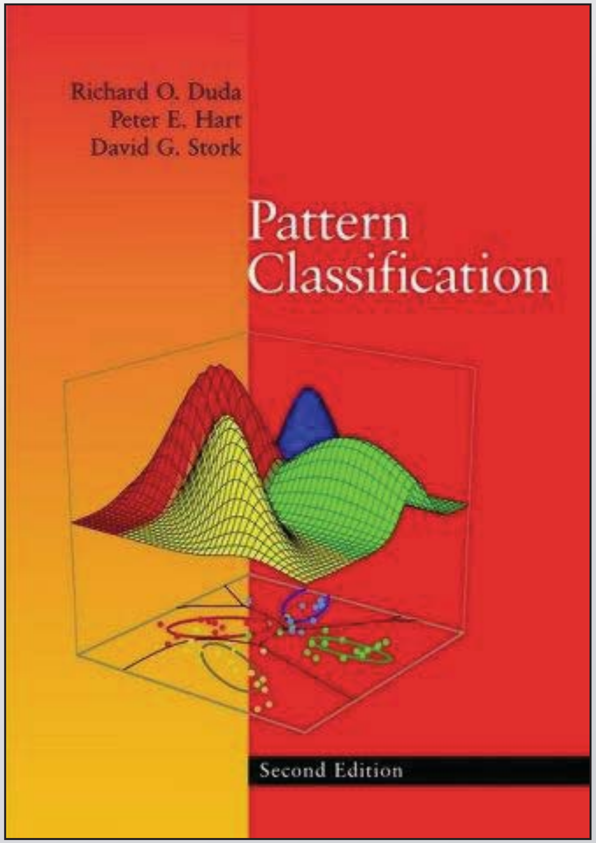
(Optional) Books for the Course


Copyrighted Material
Biological sequence analysis
Probabilistic models of proteins and nucleic acids
R. Durbin
S. Eddy
A. Krogh
G. Mitchison
Copyrighted Material

Durbin, Eddy, Krogh, Mitchison



Jones, Pevzner



Duda, Hart, Stork

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

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

New this year!! 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!

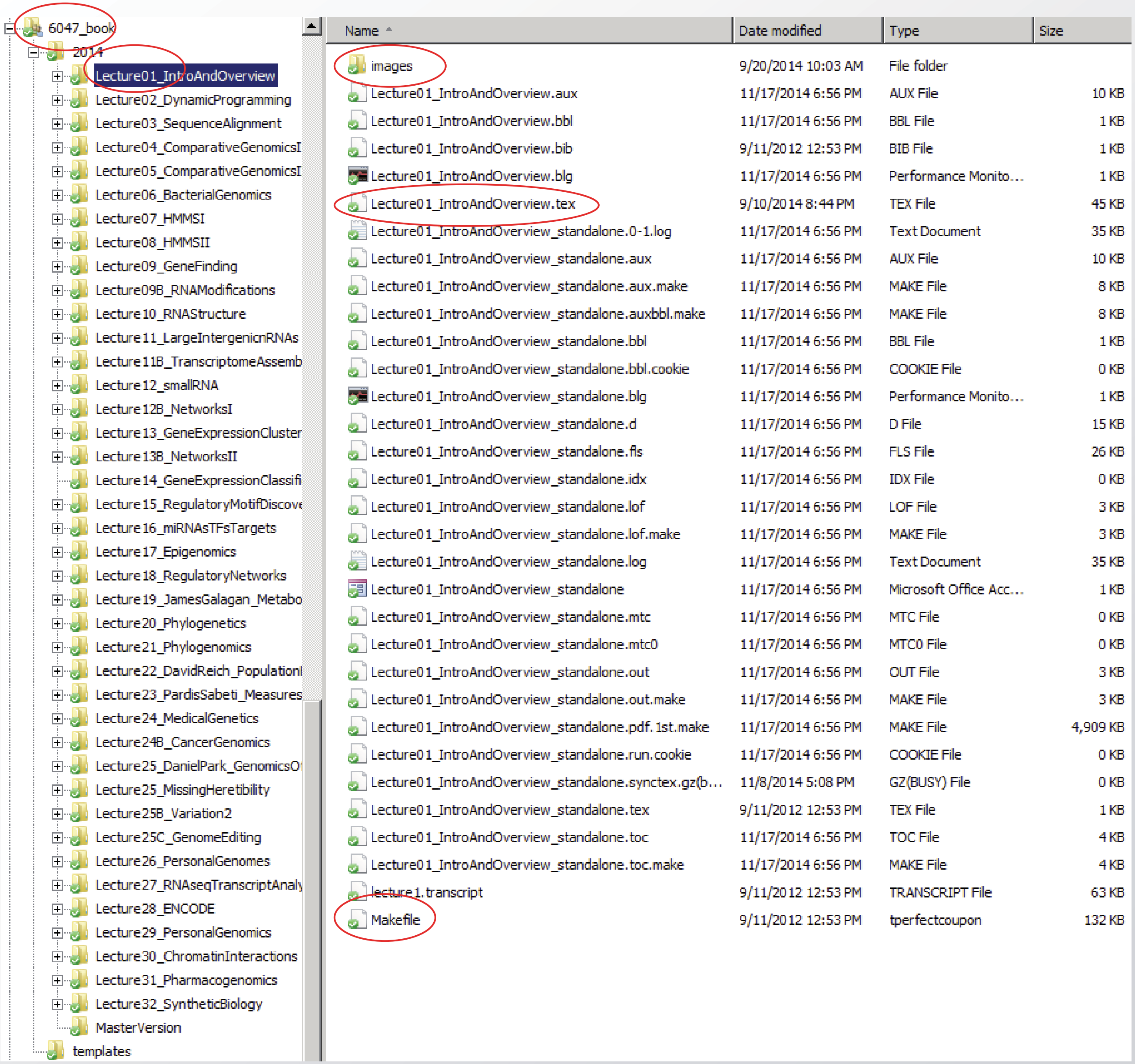
… actually, including you!

Availability: Online PDF

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
- Complete draft of scribe notes: before prev. lecture
  - Unless it’s not there from previous year (this is rare)
- 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!

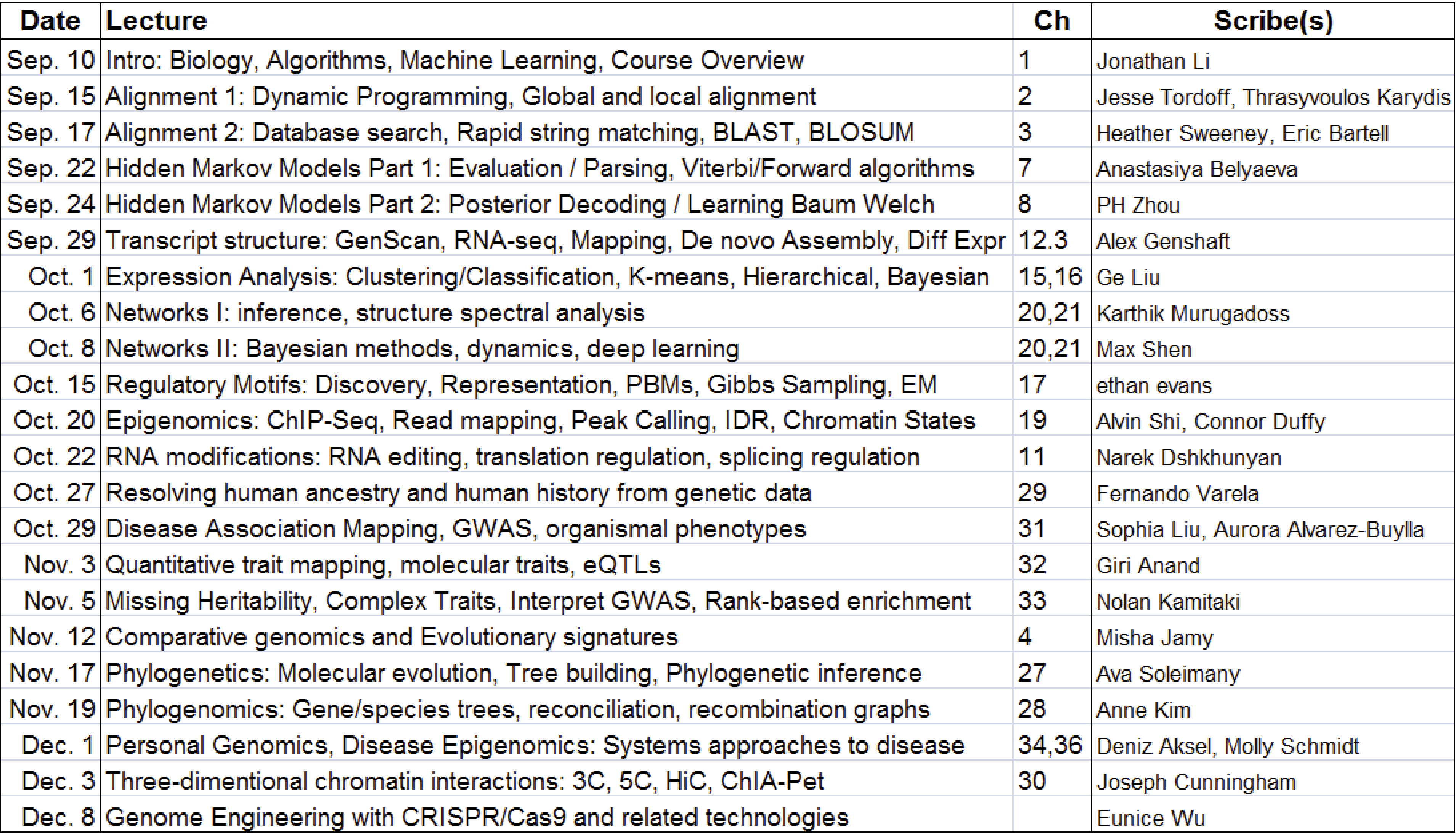
Scribing details – DropBox 6047_book LaTex



- 6047_book
  - 2014
    - Lecture01_IntroAndOverview
    - Lecture02_DynamicProgramming
    - Lecture03_SequenceAlignment
    - Lecture04_ComparativeGenomicsI
    - Lecture05_ComparativeGenomicsI
    - Lecture06_BacterialGenomics
    - Lecture07_HMMSI
    - Lecture08_HMMSII
    - Lecture09_GeneFinding
    - Lecture09B_RNAModifications
    - Lecture10_RNAStructure
    - Lecture11_LargeIntergenicnRNAs
    - Lecture11B_TranscriptomeAssemb
    - Lecture12_smallRNA
    - Lecture12B_NetworksI
    - Lecture13_GeneExpressionCluster
    - Lecture13B_NetworksII
    - Lecture14_GeneExpressionClassifi
    - Lecture15_RegulatoryMotifDiscove
    - Lecture16_miRNAsTFsTargets
    - Lecture17_Epigenomics
    - Lecture18_RegulatoryNetworks
    - Lecture19_JamesGalagan_Metabo
    - Lecture20_Phylogenetics
    - Lecture21_Phylogenomics
    - Lecture22_DavidReich_PopulationI
    - Lecture23_PardisSabeti_Measures
    - Lecture24_MedicalGenetics
    - Lecture24B_CancerGenomics
    - Lecture25_DanielPark_GenomicsO
    - Lecture25_MissingHeretibility
    - Lecture25B_Variation2
    - Lecture25C_GenomeEditing
    - Lecture26_PersonalGenomes
    - Lecture27_RNAseqTranscriptAnaly
    - Lecture28_ENCODE
    - Lecture29_PersonalGenomics
    - Lecture30_ChromatinInteractions
    - Lecture31_Pharmacogenomics
    - Lecture32_SyntheticBiology
    - MasterVersion
  - templates

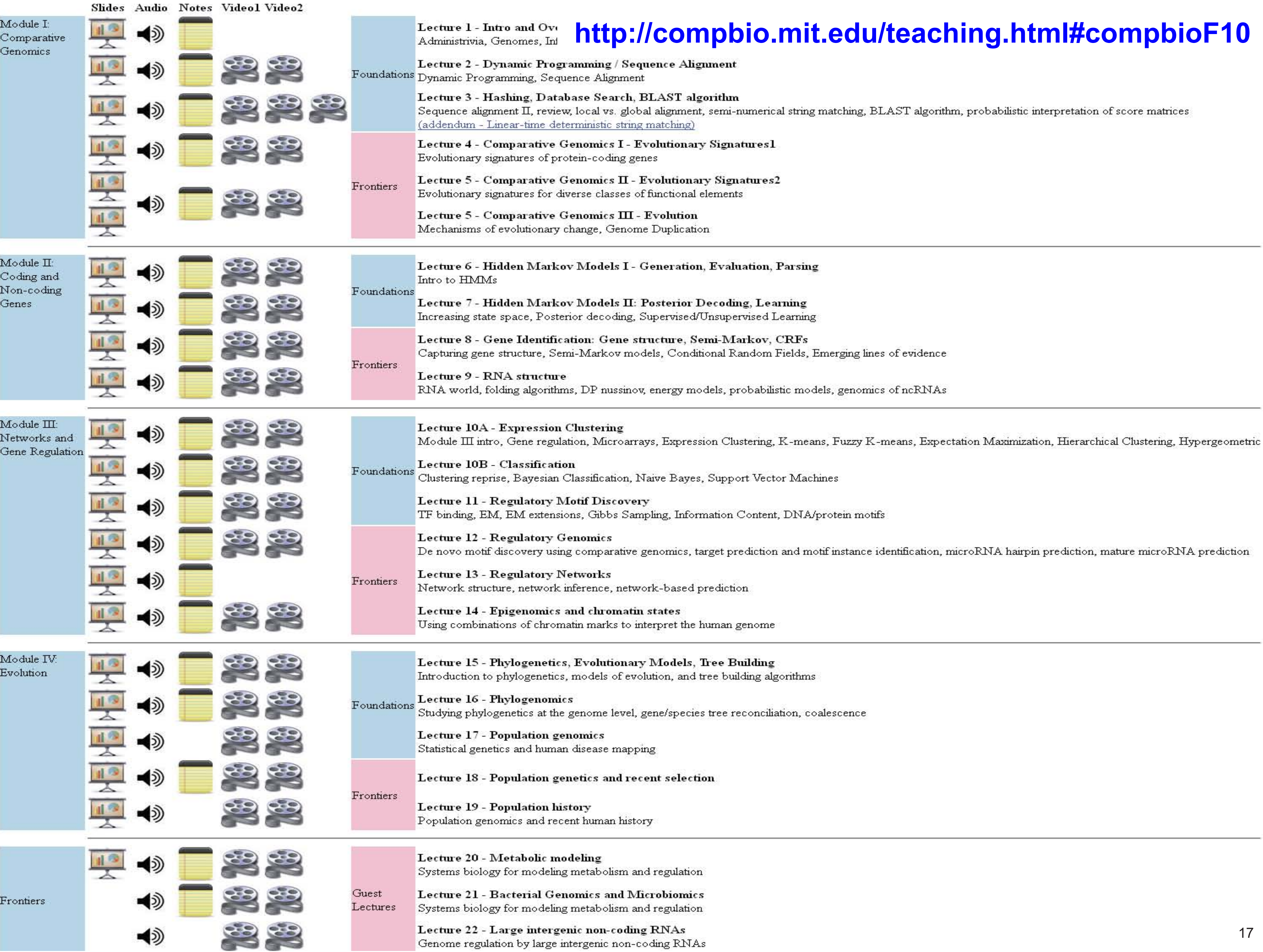
Name	Date modified	Type	Size
images	9/20/2014 10:03 AM	File folder	
Lecture01_IntroAndOverview.aux	11/17/2014 6:56 PM	AUX File	10 KB
Lecture01_IntroAndOverview.bbl	11/17/2014 6:56 PM	BBL File	1 KB
Lecture01_IntroAndOverview.bib	9/11/2012 12:53 PM	BIB File	1 KB
Lecture01_IntroAndOverview.blg	11/17/2014 6:56 PM	Performance Monito...	1 KB
Lecture01_IntroAndOverview.tex	9/10/2014 8:44 PM	TEX File	45 KB
Lecture01_IntroAndOverview_standalone.0-1.log	11/17/2014 6:56 PM	Text Document	35 KB
Lecture01_IntroAndOverview_standalone.aux	11/17/2014 6:56 PM	AUX File	10 KB
Lecture01_IntroAndOverview_standalone.aux.make	11/17/2014 6:56 PM	MAKE File	8 KB
Lecture01_IntroAndOverview_standalone.auxbbl.make	11/17/2014 6:56 PM	MAKE File	8 KB
Lecture01_IntroAndOverview_standalone.bbl	11/17/2014 6:56 PM	BBL File	1 KB
Lecture01_IntroAndOverview_standalone.bbl.cookie	11/17/2014 6:56 PM	COOKIE File	0 KB
Lecture01_IntroAndOverview_standalone.blg	11/17/2014 6:56 PM	Performance Monito...	1 KB
Lecture01_IntroAndOverview_standalone.d	11/17/2014 6:56 PM	D File	15 KB
Lecture01_IntroAndOverview_standalone.fls	11/17/2014 6:56 PM	FLS File	26 KB
Lecture01_IntroAndOverview_standalone.idx	11/17/2014 6:56 PM	IDX File	0 KB
Lecture01_IntroAndOverview_standalone.lof	11/17/2014 6:56 PM	LOF File	3 KB
Lecture01_IntroAndOverview_standalone.lof.make	11/17/2014 6:56 PM	MAKE File	3 KB
Lecture01_IntroAndOverview_standalone.log	11/17/2014 6:56 PM	Text Document	35 KB
Lecture01_IntroAndOverview_standalone	11/17/2014 6:56 PM	Microsoft Office Acc...	1 KB
Lecture01_IntroAndOverview_standalone.mtc	11/17/2014 6:56 PM	MTC File	0 KB
Lecture01_IntroAndOverview_standalone.mtc0	11/17/2014 6:56 PM	MTC0 File	0 KB
Lecture01_IntroAndOverview_standalone.out	11/17/2014 6:56 PM	OUT File	3 KB
Lecture01_IntroAndOverview_standalone.out.make	11/17/2014 6:56 PM	MAKE File	3 KB
Lecture01_IntroAndOverview_standalone.pdf.1st.make	11/17/2014 6:56 PM	MAKE File	4,909 KB
Lecture01_IntroAndOverview_standalone.run.cookie	11/17/2014 6:56 PM	COOKIE File	0 KB
Lecture01_IntroAndOverview_standalone.synctex.gz(b...	11/8/2014 5:08 PM	GZ(BUSY) File	0 KB
Lecture01_IntroAndOverview_standalone.tex	9/11/2012 12:53 PM	TEX File	1 KB
Lecture01_IntroAndOverview_standalone.toc	11/17/2014 6:56 PM	TOC File	4 KB
Lecture01_IntroAndOverview_standalone.toc.make	11/17/2014 6:56 PM	MAKE File	4 KB
lecture1.transcript	9/11/2012 12:53 PM	TRANSCRIPT File	63 KB
Makefile	9/11/2012 12:53 PM	tperfectcoupon	132 KB

Sign up here if you haven’t already



Date	Lecture	Ch	Scribe(s)
Sep. 10	Intro: Biology, Algorithms, Machine Learning, Course Overview	1	Jonathan Li
Sep. 15	Alignment 1: Dynamic Programming, Global and local alignment	2	Jesse Tordoff, Thrasyvoulos Karydis
Sep. 17	Alignment 2: Database search, Rapid string matching, BLAST, BLOSUM	3	Heather Sweeney, Eric Bartell
Sep. 22	Hidden Markov Models Part 1: Evaluation / Parsing, Viterbi/Forward algorithms	7	Anastasiya Belyaeva
Sep. 24	Hidden Markov Models Part 2: Posterior Decoding / Learning Baum Welch	8	PH Zhou
Sep. 29	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr	12.3	Alex Genshaft
Oct. 1	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16	Ge Liu
Oct. 6	Networks I: inference, structure spectral analysis	20,21	Karthik Murugadoss
Oct. 8	Networks II: Bayesian methods, dynamics, deep learning	20,21	Max Shen
Oct. 15	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17	ethan evans
Oct. 20	Epigenomics: ChIP-Seq, Read mapping, Peak Calling, IDR, Chromatin States	19	Alvin Shi, Connor Duffy
Oct. 22	RNA modifications: RNA editing, translation regulation, splicing regulation	11	Narek Dshkhunyan
Oct. 27	Resolving human ancestry and human history from genetic data	29	Fernando Varela
Oct. 29	Disease Association Mapping, GWAS, organismal phenotypes	31	Sophia Liu, Aurora Alvarez-Buylla
Nov. 3	Quantitative trait mapping, molecular traits, eQTLs	32	Giri Anand
Nov. 5	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	33	Nolan Kamitaki
Nov. 12	Comparative genomics and Evolutionary signatures	4	Misha Jamy
Nov. 17	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27	Ava Soleimany
Nov. 19	Phylogenomics: Gene/species trees, reconciliation, recombination graphs	28	Anne Kim
Dec. 1	Personal Genomics, Disease Epigenomics: Systems approaches to disease	34,36	Deniz Aksel, Molly Schmidt
Dec. 3	Three-dimentional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	30	Joseph Cunningham
Dec. 8	Genome Engineering with CRISPR/Cas9 and related technologies		Eunice Wu

http://compbio.mit.edu/teaching.html#compbioF10



	Slides	Audio	Notes	Video1	Video2			
Module I: Comparative Genomics	✓	✓	✓				Foundations	**Lecture 1 - Intro and Ov** Administrivia, Genomes, Inf
	✓	✓	✓	✓	✓			**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

Lecture feedback:

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 Tuesdays at 8pm
  - 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
  - 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 planning

Project execution

★Assignment due
★Feedback rec'd

Project	Psets	Week	Date	Topic	
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** with PS1	PS1 out on:L1-L5 due Mon 9/23	1	Thu, Sep 04	Introduction	
			Fri, Sep 05		
		2	Tue, Sep 10	Module I: Aligning and Modeling Genomes	Foundations
			Thu, Sep 12		
			Fri, Sep 13		
		3	Tue, Sep 17		Frontiers
			Thu, Sep 19		
			Fri, Sep 20		
			Fri, Sep 20		Project Intro: ★★
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 Mon 10/7** with PS2	PS2 out on:L6-R5 due Mon 10/7	4	Tue, Sep 24	Module II: Genes and Transcripts	Foundations
			Thu, Sep 26		
			Fri, Sep 27		
		5	Tue, Oct 01		Frontiers
			Thu, Oct 03		
			Fri, Oct 04		
			Fri, Oct 04		Project Planni★
Form teams of two, specify project goals, division of work, milestones, datasets, challenges Prepare slide presentation for the class and the mentors. **Presented on 10/18.** Formal Project Proposal, in form of NIH proposal.	PS3 out on:L10-R8 due Mon 10/21	6	Tue, Oct 08	Module III: Regulation, Epigenomics, Networks	Foundations ★
			Thu, Oct 10		
			Fri, Oct 11		
		7	Tue, Oct 15		Frontiers
			Thu, Oct 17		
			Fri, Oct 18		
			Thu, Oct 17		Project feedba★
Evaluate/discuss three peer proposals, NIH review format. **Reviews / Panel Discussion Mon 10/26. Written reviews due**	PS4 out on:L15-R10 due Mon 11/04	8	Tue, Oct 22	Module IV: Evolution and Phylogenetics	Foundations
			Thu, Oct 24		
			Fri, Oct 25		
			Fri, Oct 26		Panel Discuss★
Address peer evaluations, revise aims, scope, list of final deliverables / goals. Revised		9	Tue, Oct 29		Frontiers
			Thu, Oct 31		
			Fri, Nov 01		
Continue making substantial progress on proposed milstones.Write outline of final report. **Midcourse progress report Due on Mon 11/18. Project final score projection from course staff by Friday**	PS5 out on:L15-R10 due Mon 11/18	10	Tue, Nov 05	Module V: Population Genetics and Demography	Foundations
			Thu, Nov 07		
			Fri, Nov 08		
		11	Tue, Nov 12		Frontiers
			Thu, Nov 14		
			Fri, Nov 15		
			Fri, Nov 15		Progress feed★
Complete your milestones, finalize results, figures, write-up in conference publication format. As part of report, comment on your overall project experience. **Final written report Due 12/7**	No more psets! (work on your final project)	12	Tue, Nov 19	Quiz	Quiz
			Thu, Nov 21	Module VI: Current Research Directions	Frontiers ★
			Fri, Nov 22		
		13	Tue, Nov 26		
			Thu, Nov 28		
			Fri, Nov 29		
		14	Mon, Dec 02		One-on-one m★
			Tue, Dec 03		Frontiers ★★
			Thu, Dec 05		
			Fri, Dec 06	Final Presentations	
Conference format slide presentation. **Talks on 12/10**		15	Tue, Dec 10		
			Tue, Dec 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

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%**

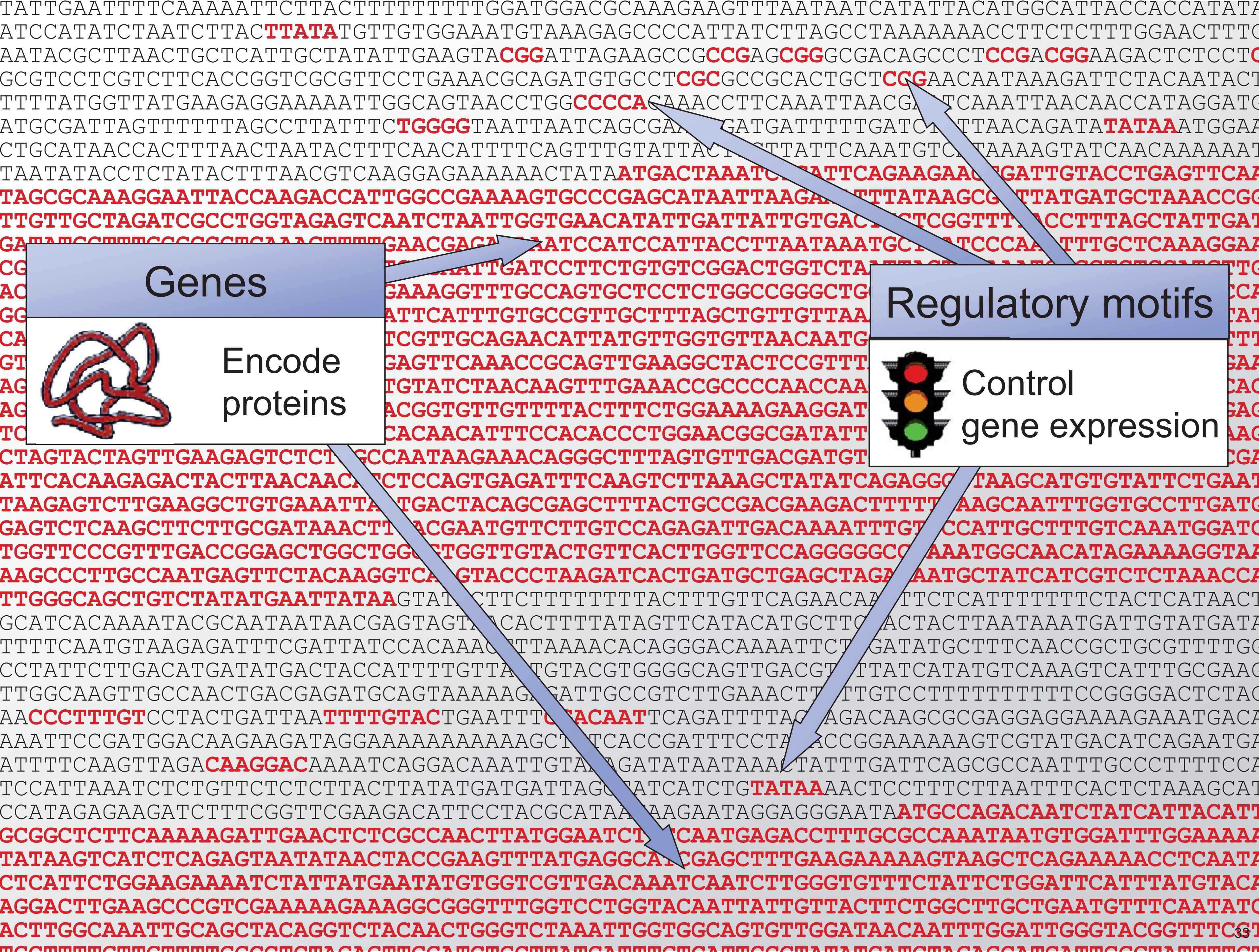
- 4 problem sets:
  - Each problem set: 7-10%, covers 3-4 lectures, contains 3-4 problems.
  - Algorithmic problems and programming assignments (PS1 out now)
  - Graduate version includes additional problem on current research
- Final project
  - Introduction to research in computational biology (7 weeks!)
  - Includes peer-reviewed NIH-style proposal and much feedback
- Quiz
  - In-class quiz (Tue Nov 15). 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

```
TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAG
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATG
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGG
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAG
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTTCCGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATT
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCG
```



Genes

Encode proteins

Regulatory motifs

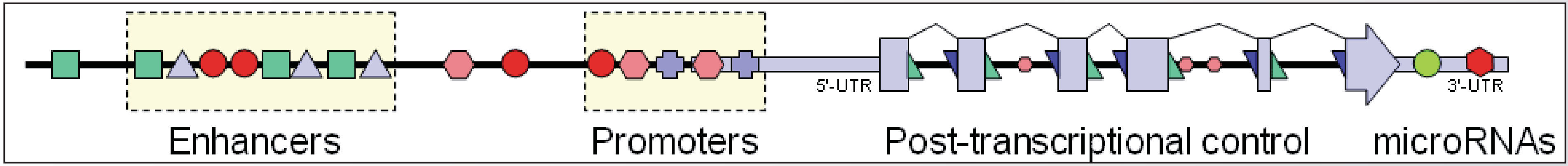
Control gene expression

TATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
ATCCATATCTAATCTTAC**TTATA**TGTTGTGGAAATGTAAAGAGCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
AATACGCTTAACTGCTCATTGCTATATTGAAGTA**CGG**ATTAGAAGCCG**CCG**AG**CGG**GCGACAGCCCT**CCG**A**CGG**AAGACTCTCCTC**C**
GCGTCCTCGTCTTCACCGGTCGCGTTCCTGAAACGCAGATGTGCCT**CGC**GCCGCACTGCT**CCG**AACAATAAAGATTCTACAATACT
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGG**CCCCA**CAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTC**TGGGG**TAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATA**TATAA**ATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATA**ATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA**
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAC
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAG
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATG
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGC
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAC
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTTCCGGGGACTCTAC
AA**CCCTTTGT**CCTACTGATTAA**TTTTGTAC**TGAATTT**GGACAAT**TCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGA**CAAGGAC**AAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTG**TATAA**AACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATA**ATGCCAGACAATCTATCATTACATT**
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATAT
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTC

Extracting signal from noise

```
TATTGAATTTTCAAAAATTCTTACTTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATA
A                              CCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTC
A                              AGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
G                              CCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAATACT
T                              ACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATG
ATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAACAGATATATAAATGGAA
CTGCATAACCACTTTAACTAATACTTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGAT
GATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTG
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGCCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTCATTTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTTGGTGTTAACAATGGCGGTATGGATCAGGCTGCCTCTGTTT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAAAAAACCATGAA
AGCTTTGTTATTGCGAACACCCTTGTTGTATCTAACAAGTTTGAAACCGCCCCAACCAACTATAATTTAAGAGTGGTAGAAGTCAC
AGCTGCAAATGTTTTAGCTGCCACGTACGGTGTTGTTTTACTTTCTGGAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAC
TCATGAACGTTTATTATGCCAGATATCACAACATTTCCACACCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAAAC
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTCACAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCTTGATG
GAGTCTCAAGCTTCTTGCGATAAACTTTACGAATGTTCTTGTCCAGAGATTGACAAAATTTGTTCCATTGCTTTGTCAAATGGATG
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTATAAGTATACTTCTTTTTTTTTACTTTGTTCAGAACAACTTCTCATTTTTTTTCTACTCATAACT
GCATCACAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTCATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTCGATTATCCACAAACTTTAAAACACAGGGACAAAATTCTTGATATGCTTTCAACCGCTGCGTTTTGC
CCTATTCTTGACATGATATGACTACCATTTTGTTATTGTACGTGGGGCAGTTGACGTCTTATCATATGTCAAAGTCATTTGCGAAC
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCCTTTTTTTTTTCCGGGACTCTAC
AACCCTTTGTCCTACTGATTAATTTTGTACTGAATTTGGACAATTCAGATTTTAGTAGACAAGCGCGAGGAGGAAAAGAAATGACA
AAATTCCGATGGACAAGAAGATAGGAAAAAAAAAAAAGCTTTCACCGATTTCCTAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAATCAGGACAAATTGTAAAGATATAATAAACTATTTGATTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAAATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCAT
CCATAGAGAAGATCTTTCGGTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGAATAATGCCAGACAATCTATCATTACATT
GCGGCTCTTCAAAAAGATTGAACTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTTGGAAAAA
TATAAGTCATCTCAGAGTAATATAACTACCGAAGTTTATGAGGCATCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCTGGAAGAAAATCTATTATGAATATGTGGTCGTTGACAAATCAATCTTGGGTGTTTCTATTCTGGATTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCCTGGTACAATTATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAATTGCAGCTACAGGTCTACAACTGGGTCTAAATTGGTGGCAGTGTTGGATAACAATTTGGATTGGGTACGGTTTCGT
```

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

6.047/6.878/HST.507 - Fall 2015 - Lectures: Manolis Kellis. All lectures on Tuesday/Thursday at 1pm-2:30m
All homeworks due on Tuesday at 8pm

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 9/29**	PS1 out on:L1-L5	1	Thu, Sep 10	**Introduction**		L1	Intro: Biology, Algorithms, Machine Learning, Course Overview	1
			Fri, Sep 11			R1	Recitation 1: Biology and Probability Review	
		2	Tue, Sep 15	Module I: Aligning and Modeling Genomes	Foundations	L2	Alignment I: Dynamic Programming, Global and local alignment	2
			Thu, Sep 17			L3	Alignment II: Database search, Rapid string matching, BLAST, BLOSUM	3
			Fri, Sep 18			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
		3	Tue, Sep 22		Frontiers	L4	Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms	7
			Thu, Sep 24			L5	Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch	8
	due Tue 9/29		Fri, Sep 25				No classes - student holiday	
			Fri, 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 10/6**	PS2 out on:L6-R4	4	Tue, Sep 29	Module II: Gene Expression and Networks	Foundations	L6	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr	12.3
			Thu, Oct 1			L7	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
			Fri, Oct 2			R3	Recitation 3: Affinity Propagation Clustering and Random Forest Classification	
		5	Tue, Oct 6		Frontiers	L8	Networks I: Bayesian inference, deep learning, network dynamics	20,21
			Thu, Oct 8			L9	Networks II: Network learning, structure, spectral methods	20,21
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 10/20. Presented on Fri 10/23**	due Tue 10/13		Fri, Oct 9			R4	Recitation 4: Small and Large Regulatory RNAs: lincRNA, miRNA, piRNA…	
			Fri, Oct 9		Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D-507			
	PS3 out on:L10-R6	6	Tue, Oct 13	Module III: Gene Regulation & Epigenomics	Foundations		No Classes - Monday Schedule - October 13, 2015	
			Thu, Oct 15			L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
			Fri, Oct 16			R5	Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex	
		7	Tue, Oct 20		Frontiers	L11	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
			Thu, Oct 22			L12	RNA modifications: RNA editing, Translation regulation, Splicing regulation	11
	due Tue 10/27		Fri, Oct 23			R6	Recitation 6: Dimensionality Reduction	
			Fri, Oct 23		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 10/30** **Reviews back Tue 11/3**	PS4 out on:L13-R8	8	Tue, Oct 27	Module IV: Population and Disease Genetics	Foundations	L13	Resolving human ancestry and human history from genetic data	29
			Thu, Oct 29			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
			Fri, Oct 30			R7	Recitation 7: Robinson-Foulds Distance and Coalescent Process	
			Fri, Oct 30		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 11/12**		9	Tue, Nov 3		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs	32
	due Tue 11/10		Thu, Nov 5			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	33
			Fri, Nov 6			R8	Recitation 8: Suffix Trees and Arrays	
	PS5 out on:L17-R10	10	Tue, Nov 10	Module V: Comparative Genomics and Evolution	Foundations		No lecture, veterans day holiday - Monday/Tuesday	
			Thu, Nov 12			L17	Comparative genomics and Evolutionary signatures	4
Continue making substantial progress on proposed milstones.Write outline of final report. **Midcourse report due Thu 11/19. Score projection 11/24**			Fri, Nov 13			R9	Recitation 9: Review of Phylogeny and Molecular Evolution	
		11	Tue, Nov 17		Frontiers	L18	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
	due Tue 12/1		Thu, Nov 19			L19	Phylogenomics: Gene/species trees, reconciliation, recombination graphs	28
			Fri, Nov 20			R10	Recitation 10: Linkage Disequilibrium, Haplotype Phasing, and Genotype Imputation	
Complete your milestones, finalize results, figures, write-up in conference publication format. As par of report, comment on your overall project experience. **Written report due Sun 12/6**	No more psets! (work on your final project)	12	Tue, Nov 24	Module VI: Current Research Directions	Frontiers		In Class Quiz (the only quiz - the class has no final exam) - covers L1-R11	
			Thu, Nov 26				No lecture, thanksgiving break - Thu Nov 26, 2015	
			Fri, Nov 27				No recitation, thanksgiving break	
		13	Tue, Dec 1			L20	Personal Genomics, Disease Epigenomics: Systems approaches to disease	34,36
			Thu, Dec 3			L21	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet	30
			Fri, Dec 4			R11	Recitation 11: Project Tips - Write-up, Slides, Final Presentation in 32D-507	
		14	Tue, Dec 8			L22	Genome Engineering with CRISPR/Cas9 and related technologies	
Conference format slide pres. **Talks on Thu 12/10**			Thu, Dec 10				Final Presentations - Part I (1pm). 32-141	
			Thu, Dec 10				Final Presentations - Part II (3pm). 32D-507	

* readings refer to chapters in compiled 2014 scribe notes, available in the materials folder

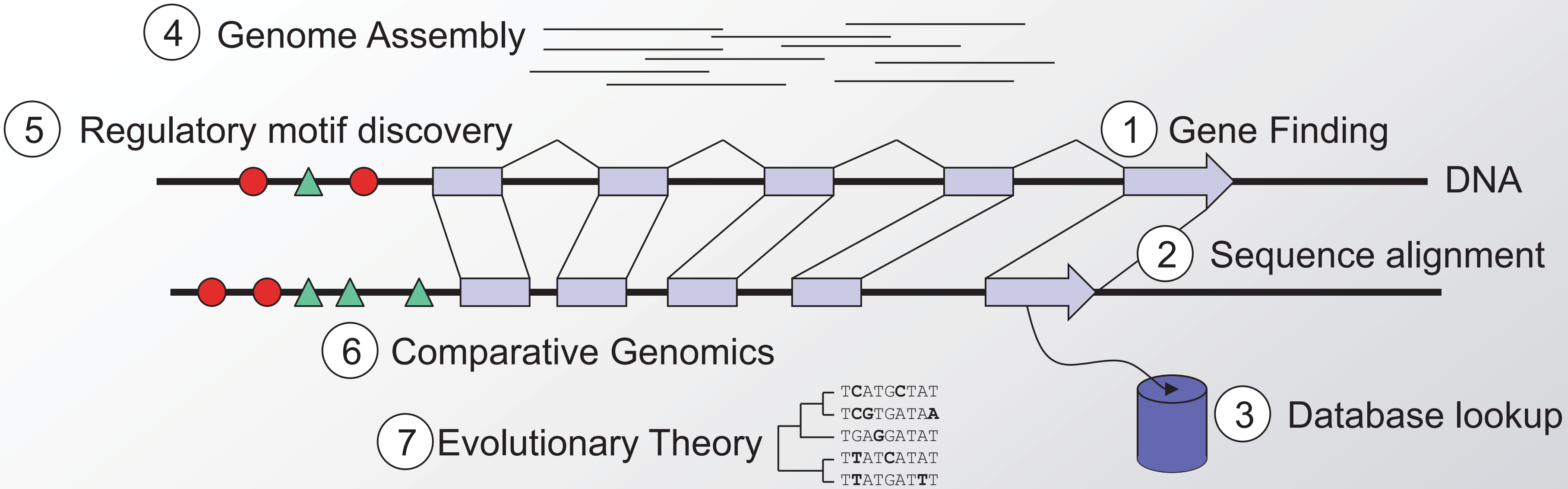
** recitation topics will be adjusted to respond to lecture and student needs

Coupling each topic with foundational CS tools

Lect	Fundamental bio problem	Foundational comp. tool
1	Introduction	
2	Sequence alignment	Dynamic programming
3	Database search	Hashing
4,5	Modeling biological signals	HMMs/Modeling/Learning/EM
6,7	Transcriptome analysis	Clustering / EM
8,9	Regulatory networks	Graph algorithms, spectral analysis
10	Regulatory motifs	Information/Gibbs Sampling/EM
11	Epigenomics	Classification / Modeling
13-16	Population Genetics	Statistical modeling and inferrence
18-19	Gene trees and species trees	Phylogenetics/Bayesian inference

Overview of the 5 modules

Challenges in Computational Biology



4 Genome Assembly

5 Regulatory motif discovery

1 Gene Finding

DNA

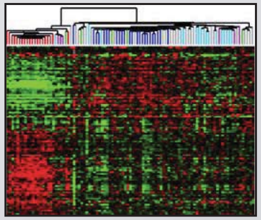
2 Sequence alignment

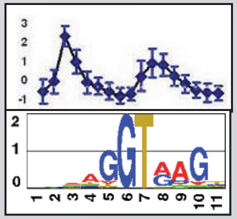
6 Comparative Genomics

3 Database lookup

7 Evolutionary Theory

8 Gene expression analysis

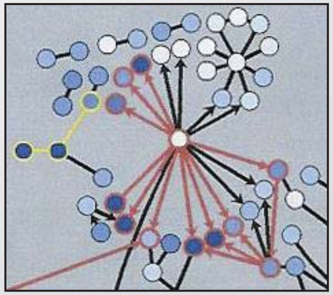




9 Cluster discovery

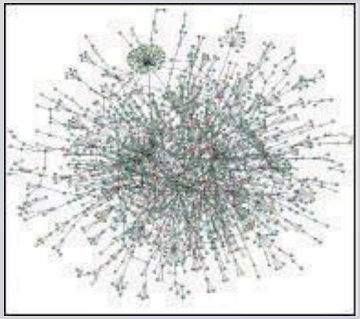
10 Gibbs sampling

11 Protein network analysis

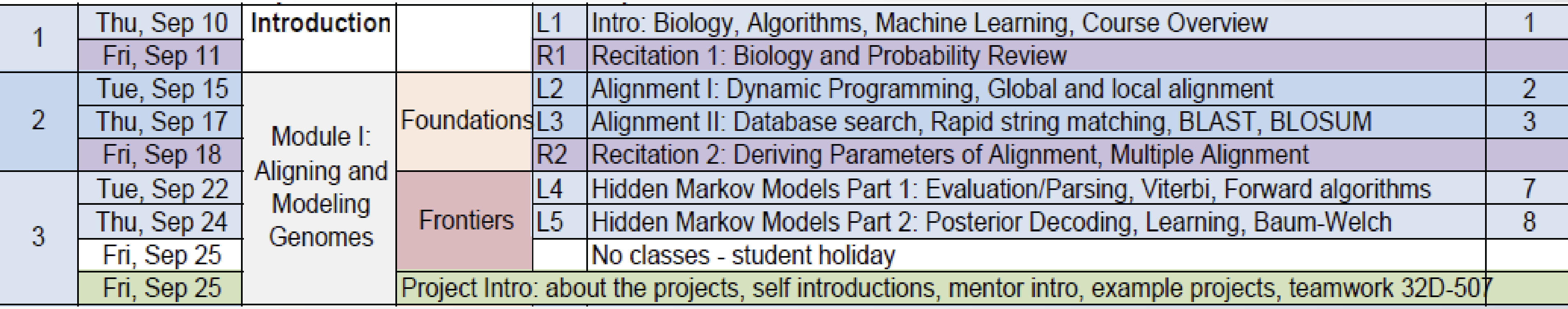


12 Metabolic modelling

13 Emerging network properties



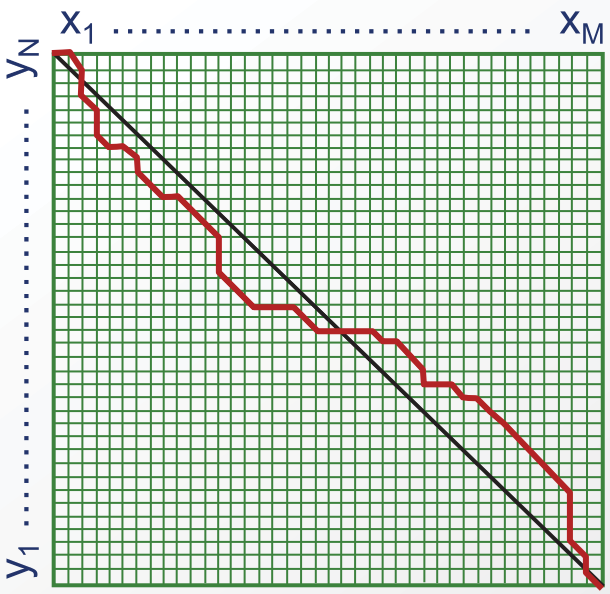
Module 1: Aligning and Modeling Genomes

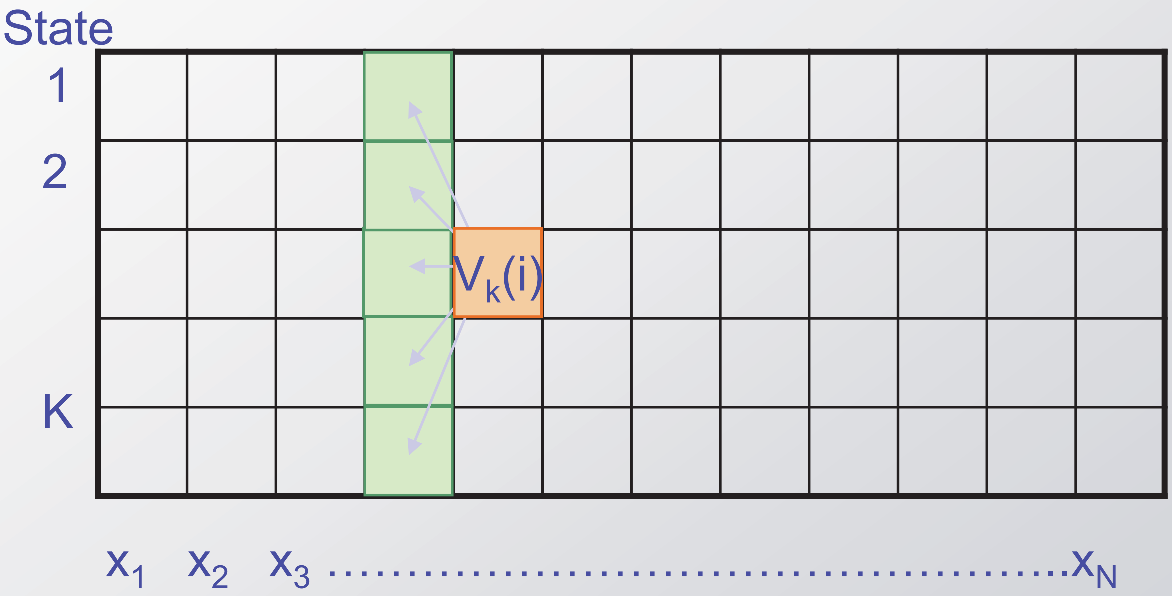


1	Thu, Sep 10	Introduction		L1	Intro: Biology, Algorithms, Machine Learning, Course Overview	1
	Fri, Sep 11			R1	Recitation 1: Biology and Probability Review	
2	Tue, Sep 15	Module I: Aligning and Modeling Genomes	Foundations	L2	Alignment I: Dynamic Programming, Global and local alignment	2
	Thu, Sep 17			L3	Alignment II: Database search, Rapid string matching, BLAST, BLOSUM	3
	Fri, Sep 18			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment	
3	Tue, Sep 22		Frontiers	L4	Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms	7
	Thu, Sep 24			L5	Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch	8
	Fri, Sep 25				No classes - student holiday	
	Fri, 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

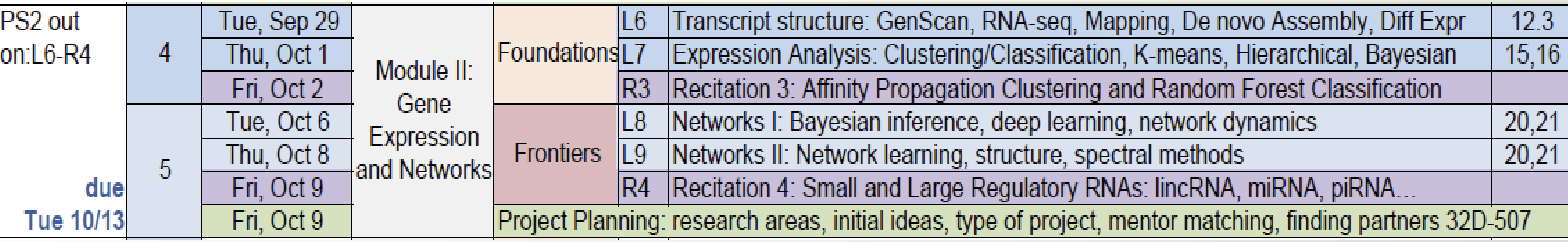
Dynamic Programming Algorithms: Align, HMMs




State
1
2
K
$V_k(i)$
x_1 x_2 x_3 .... x_N

- 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



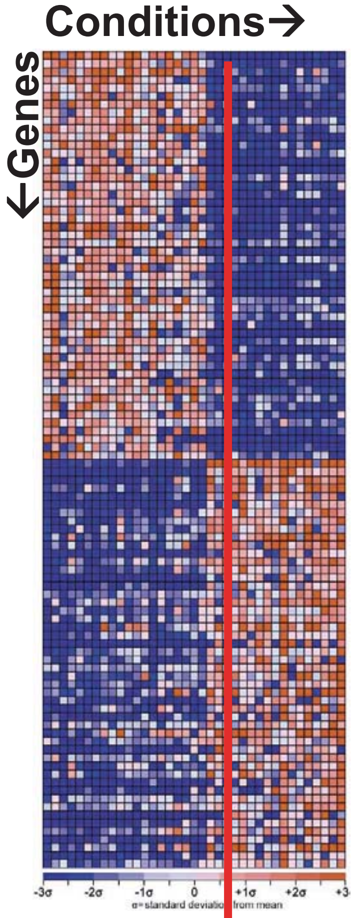
PS2 out	4	Tue, Sep 29	Module II: Gene Expression and Networks	Foundations	L6	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr	12.3
on:L6-R4		Thu, Oct 1			L7	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian	15,16
		Fri, Oct 2			R3	Recitation 3: Affinity Propagation Clustering and Random Forest Classification	
	5	Tue, Oct 6		Frontiers	L8	Networks I: Bayesian inference, deep learning, network dynamics	20,21
		Thu, Oct 8			L9	Networks II: Network learning, structure, spectral methods	20,21
due		Fri, Oct 9			R4	Recitation 4: Small and Large Regulatory RNAs: lincRNA, miRNA, piRNA...	
Tue 10/13		Fri, Oct 9		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

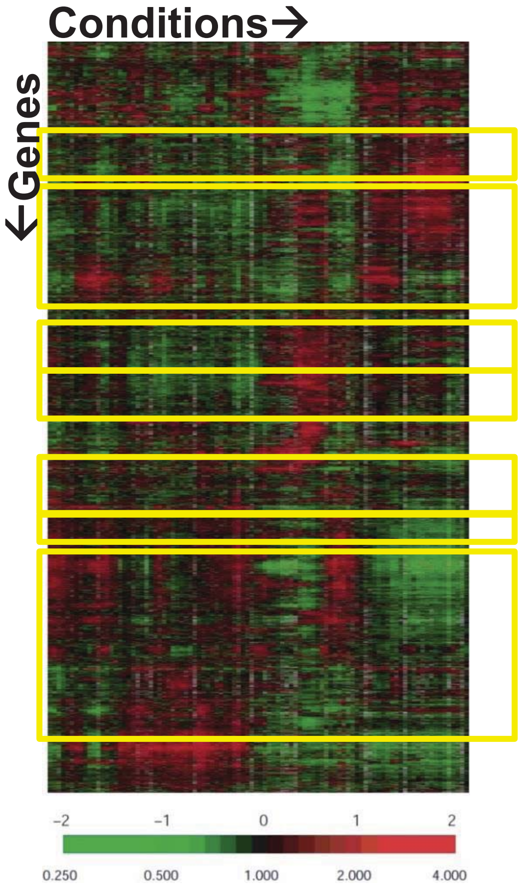
➔ Similar cell types



Armstrong, Nature Gen 2002

Reveal common ‘conditions’

➔ Similarly-behaving groups of genes



Alizadeh, Nature 2000

Reveal common gene behaviors

Source: Armstrong, Scott A. et al. "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia." Nature Genetics 30, no. 1 (2002): 41-47.

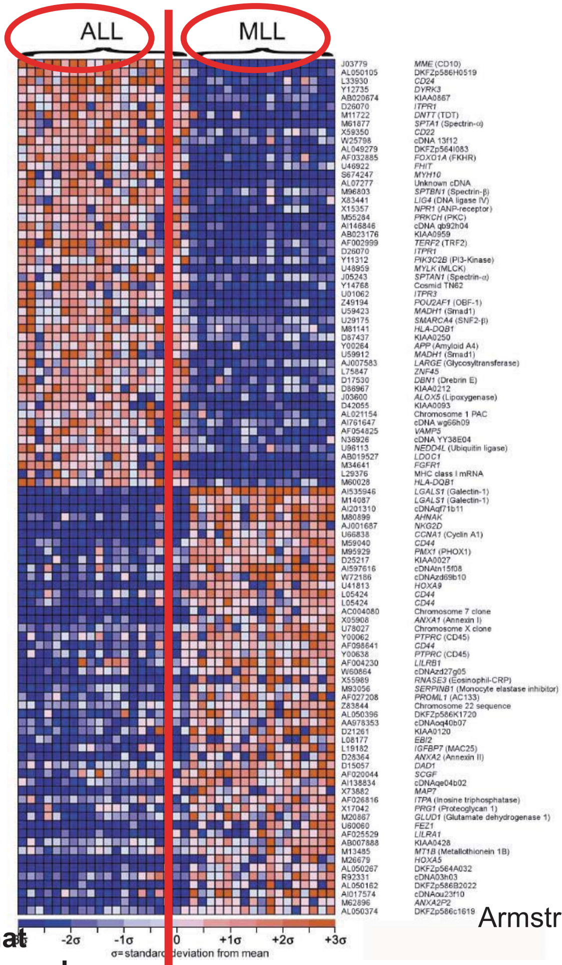
Source: Alizadeh, Ash A., Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S.Lossos, Andreas Rosenwald, Jennifer C. Boldrick et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature 403, no. 6769 (2000): 503-511.

If labels are known: find more of same type

Classification

➔ Classify diseases

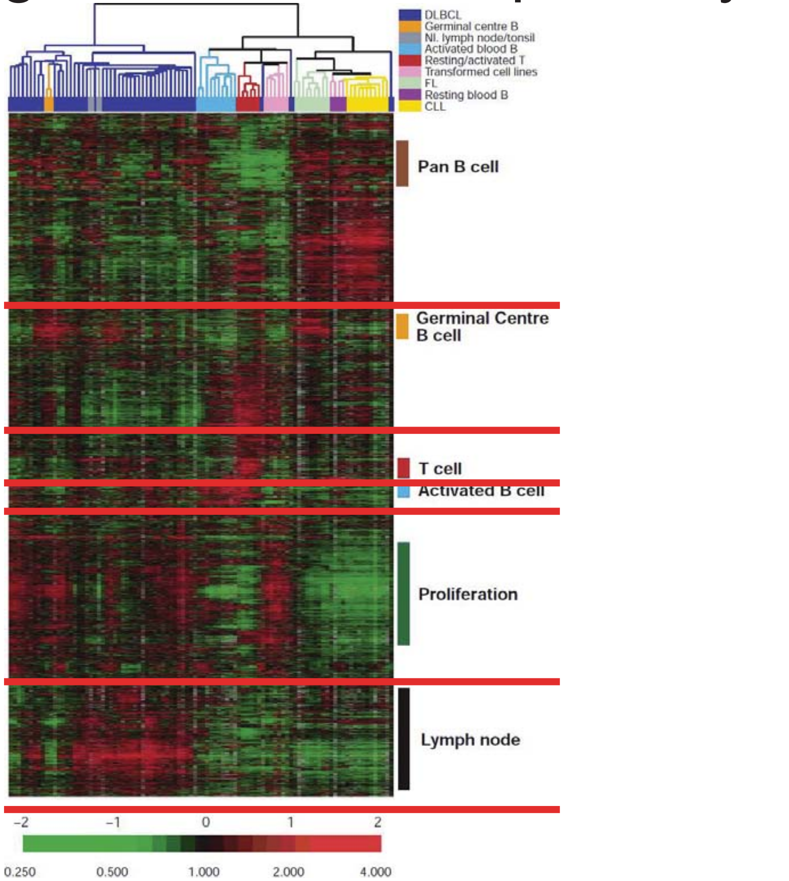
➔ Classify genes in different pathways



Armstrong, Nature Gen 2002

Find features that distinguish known classes

Source: Armstrong, Scott A. et al. "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia." Nature Genetics 30, no. 1 (2002): 41-47.



Alizadeh, Nature 2000

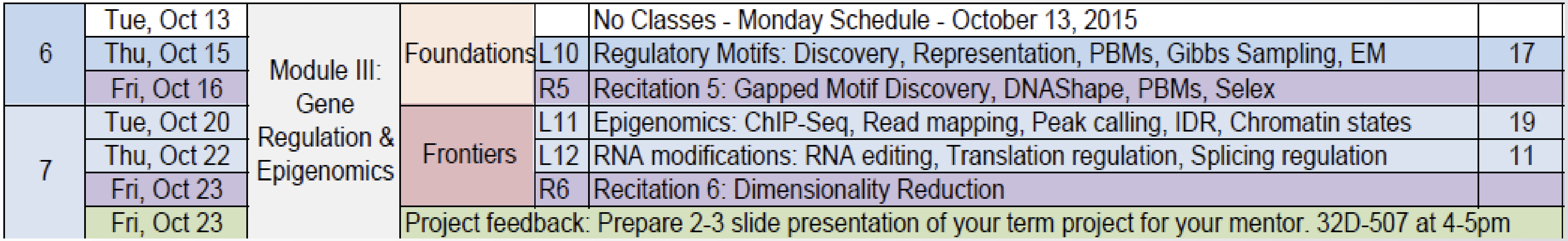
Find additional members of existing gene classes

Predict function of uncharacterized genes

Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Alizadeh, Ash A., Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S.Lossos, Andreas Rosenwald, Jennifer C. Boldrick et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature 403, no. 6769 (2000): 503-511.

Module III: Epigenomics and gene regulation

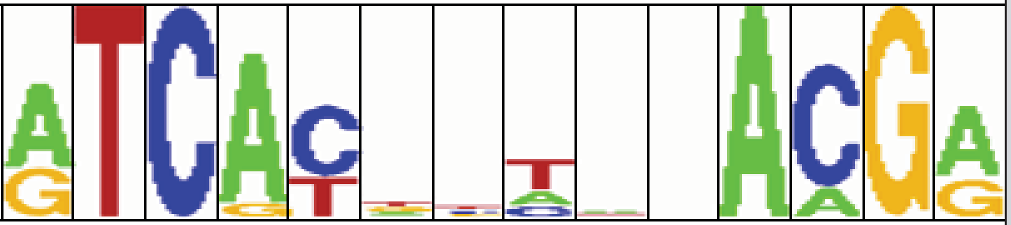


6	Tue, Oct 13	Module III: Gene Regulation & Epigenomics	Foundations		No Classes - Monday Schedule - October 13, 2015	
	Thu, Oct 15			L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM	17
	Fri, Oct 16			R5	Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex	
7	Tue, Oct 20		Frontiers	L11	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states	19
	Thu, Oct 22			L12	RNA modifications: RNA editing, Translation regulation, Splicing regulation	11
	Fri, Oct 23			R6	Recitation 6: Dimensionality Reduction	
	Fri, Oct 23		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

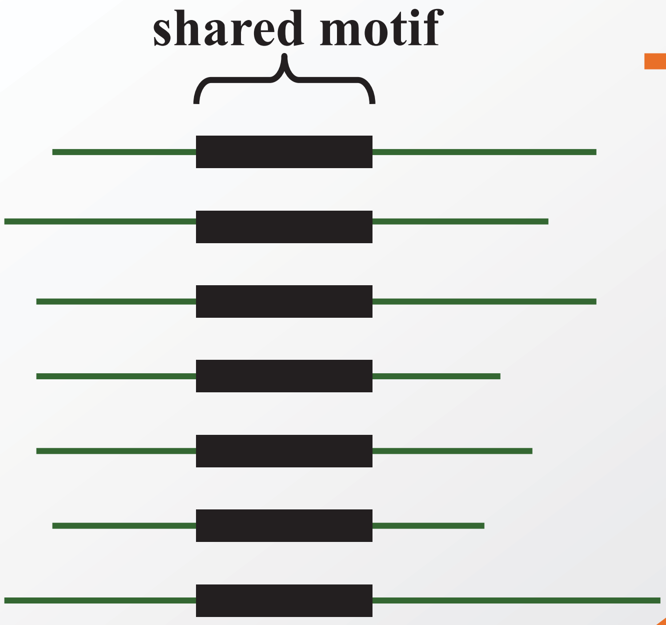
Target genes bound by ABF1 regulator		Coordinates		Genome sequence at bound site
ACS1	acetyl CoA synthetase	-491	-479	|ATCATTCTGGACG|
ACS1	acetyl CoA synthetase	-433	-421	|ATCATCTCGGACG|
ACS1	acetyl CoA synthetase	-311	-299	|ATCATTTGCCACG|
CHA1	catabolic L-serine dehydratase	-280	-254	A|ATCACCGCGAACG|GA
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

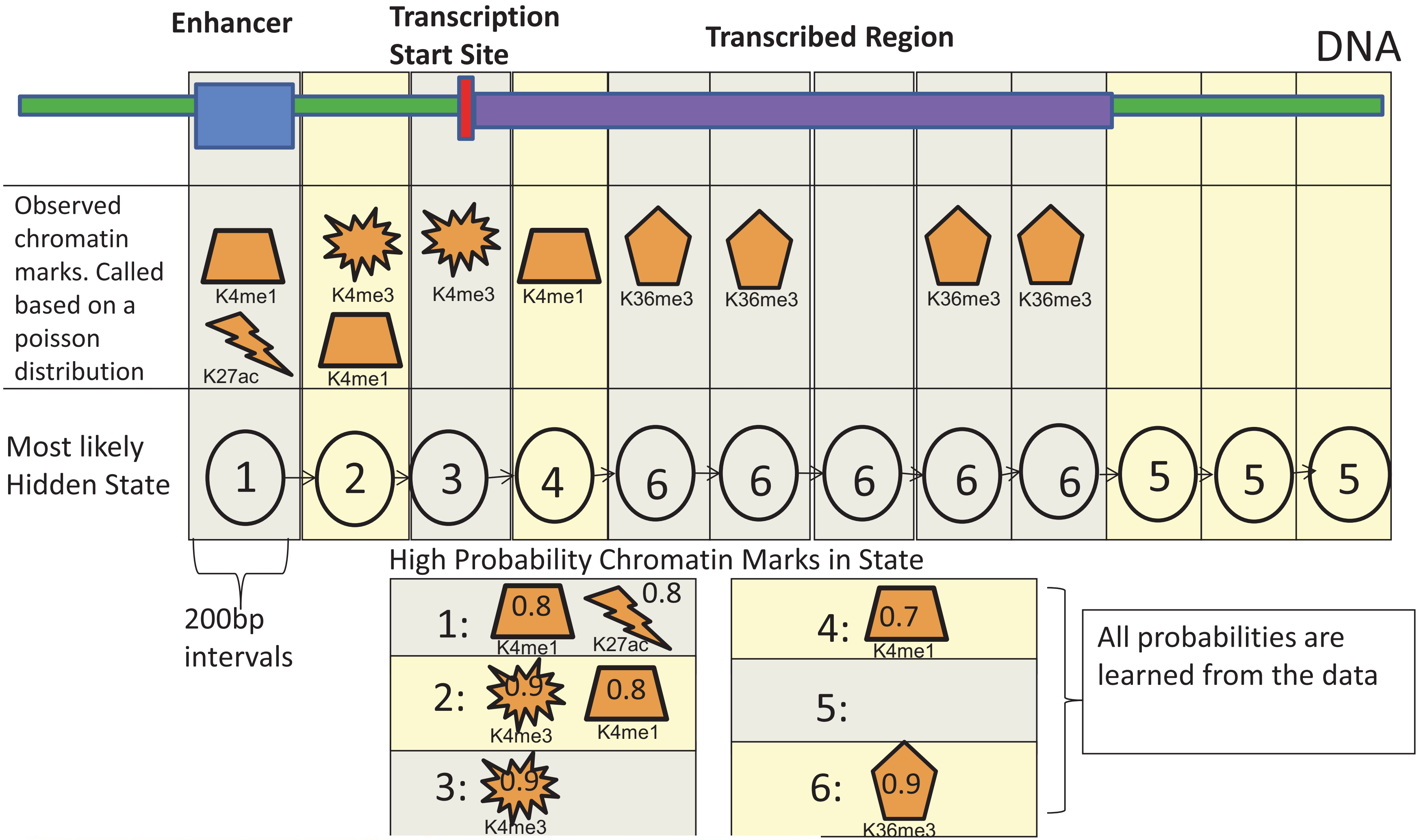
← **given profile matrix**

- **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



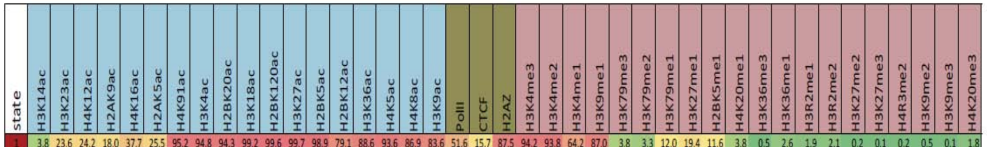
Observed chromatin marks. Called based on a poisson distribution

Most likely Hidden State

200bp intervals

High Probability Chromatin Marks in State

All probabilities are learned from the data



state	H3K14ac	H3K23ac	H4K12ac	H2AK9ac	H4K16ac	H2AK5ac	H4K91ac	H3K4ac	H2BK20ac	H3K18ac	H2BK120ac	H3K27ac	H2BK5ac	H2BK12ac	H3K36ac	H4K5ac	H4K8ac	H3K9ac	PolII	CTCF	H2AZ
1	3.8	23.6	24.2	18.0	37.7	25.5	95.2	94.8	94.3	99.2	99.6	99.7	98.9	79.1	88.6	93.6	86.9	83.6	51.6	15.7	87.5

state	H3K4me3	H3K4me2	H3K4me1	H3K9me1	H3K79me3	H3K79me2	H3K79me1	H3K27me1	H2BK5me1	H4K20me1	H3K36me3	H3K36me1	H3R2me1	H3R2me2	H3K27me2	H3K27me3	H4R3me2	H3K9me2	H3K9me3	H4K20me3
1	94.2	93.8	64.2	87.0	3.8	3.3	12.0	19.4	11.6	3.8	0.5	2.6	1.9	2.1	0.2	0.1	0.2	0.5	0.1	1.8

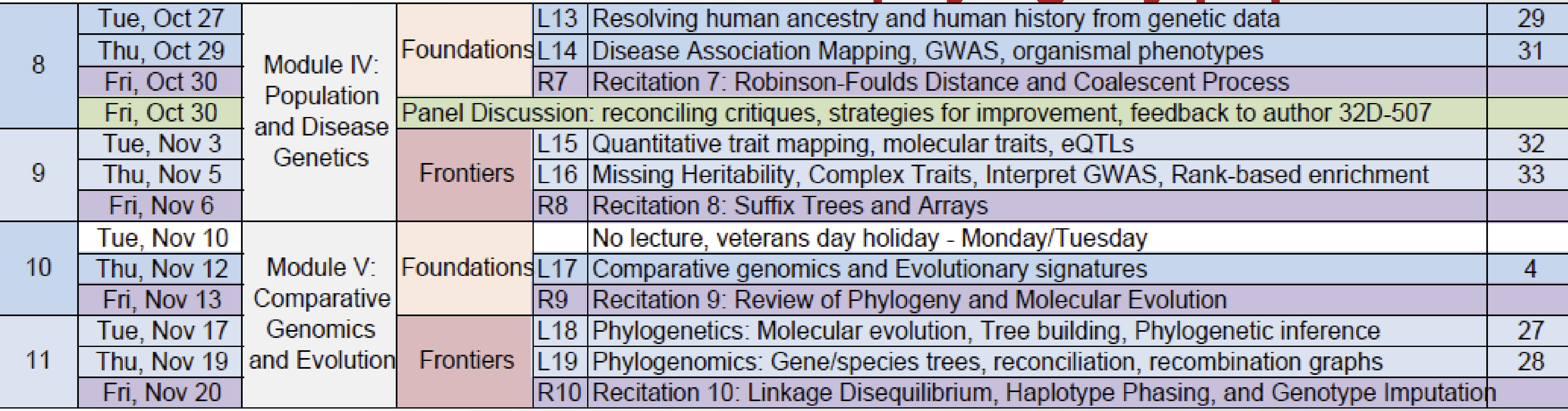
Ernst and Kellis

Nature Biotech 2010

Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

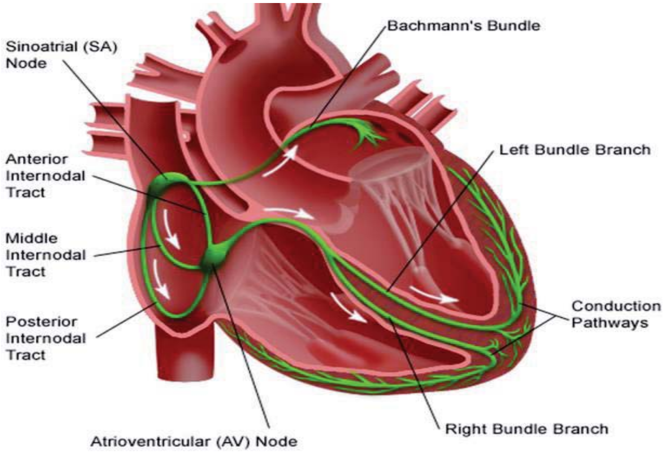
Modules IV and V: Evolution/phylogeny/populations

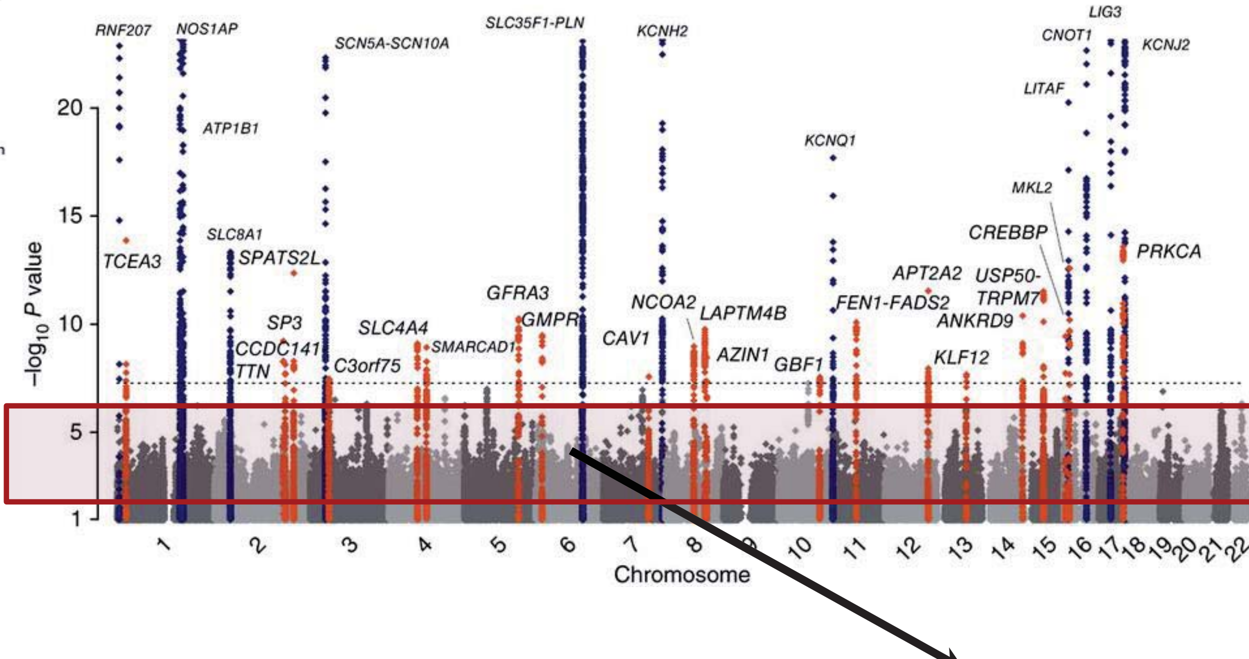


8	Tue, Oct 27	Module IV: Population and Disease Genetics	Foundations	L13	Resolving human ancestry and human history from genetic data	29
	Thu, Oct 29			L14	Disease Association Mapping, GWAS, organismal phenotypes	31
	Fri, Oct 30			R7	Recitation 7: Robinson-Foulds Distance and Coalescent Process	
	Fri, Oct 30		Panel Discussion: reconciling critiques, strategies for improvement, feedback to author 32D-507			
9	Tue, Nov 3		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs	32
	Thu, Nov 5			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment	33
	Fri, Nov 6			R8	Recitation 8: Suffix Trees and Arrays	
10	Tue, Nov 10	Module V: Comparative Genomics and Evolution	Foundations		No lecture, veterans day holiday - Monday/Tuesday	
	Thu, Nov 12			L17	Comparative genomics and Evolutionary signatures	4
	Fri, Nov 13			R9	Recitation 9: Review of Phylogeny and Molecular Evolution	
11	Tue, Nov 17		Frontiers	L18	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference	27
	Thu, Nov 19			L19	Phylogenomics: Gene/species trees, reconciliation, recombination graphs	28
	Fri, Nov 20			R10	Recitation 10: Linkage Disequilibrium, Haplotype Phasing, and Genotype Imputation	

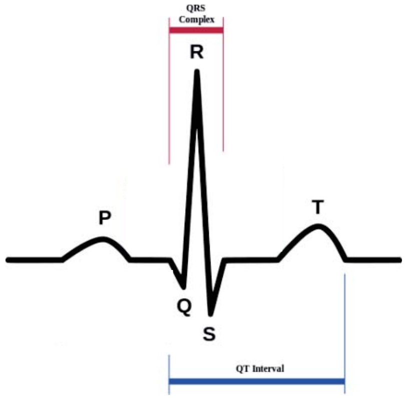
- Phylogenetics / Phylogenomics
  - Phylogenetics: Evolutionary models, Tree building, Phylo inference
  - Phylogenomics: gene/species trees, reconciliation, coalescent, pops
- Population genomics:
  - Learning population history from genetic data (David Reich)
  - Statistical genetics: disease mapping 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





Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Arking, Dan E. et al. "Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization." Nature Genetics 46, no. 8 (2014): 826-836.

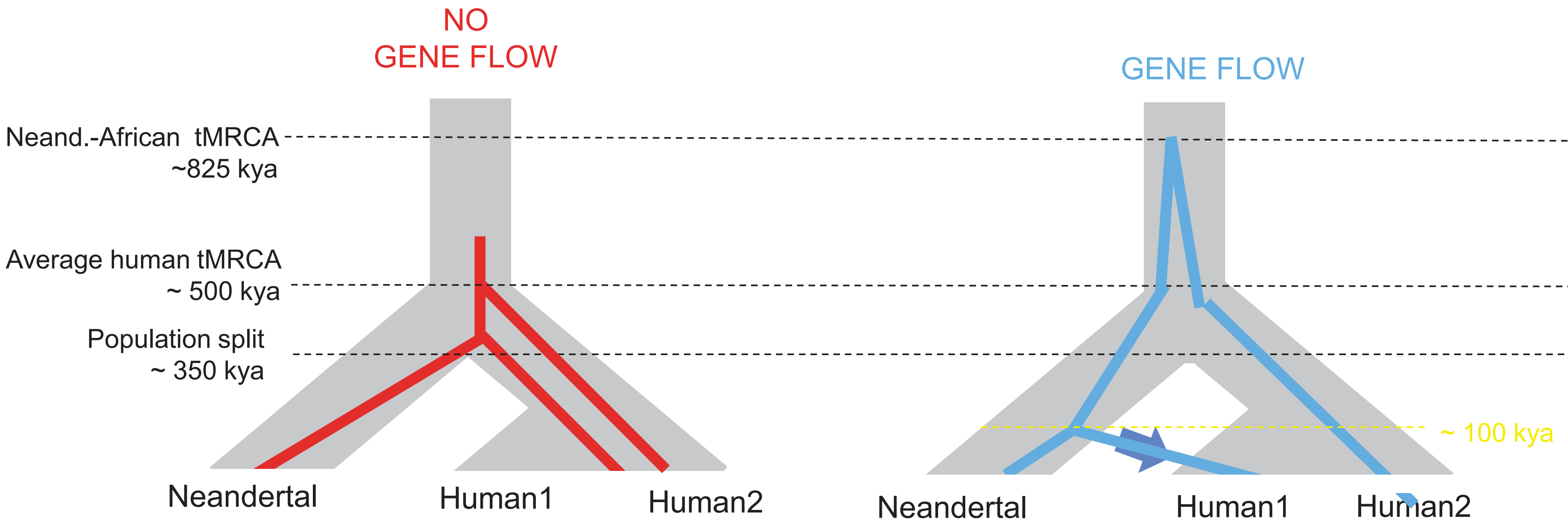


Focus on sub-threshold variants (e.g. rs1743292 P=$10^{-4.2}$)

Trait: QRS/QT interval

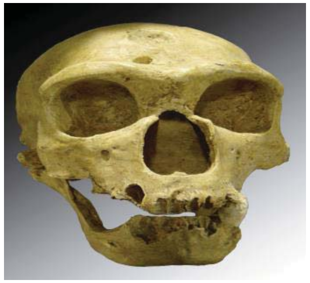
(1) Large cohorts, (2) many known hits
(3) well-characterized tissue drivers

Evidence of Neanderthal→Human gene flow



Courtesy of Luna04 on wikipedia.
License: CC BY.

hg18
BAC



Courtesy of Luna04 on wikipedia.
License: CC BY.

hg18
BAC

Human-human divergence is

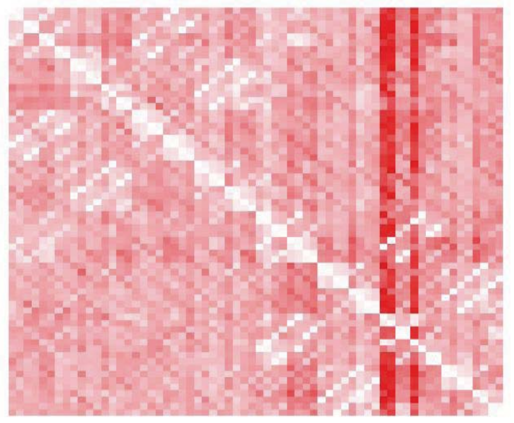
AVERAGE

Human-human divergence is

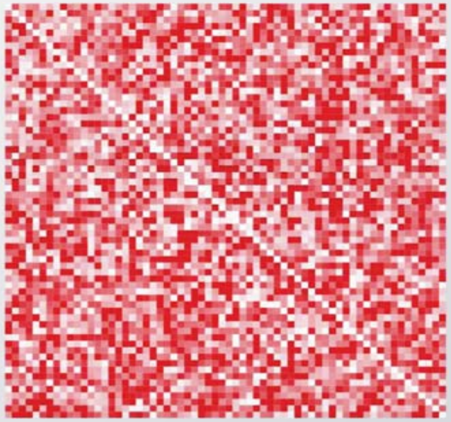
HIGH

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

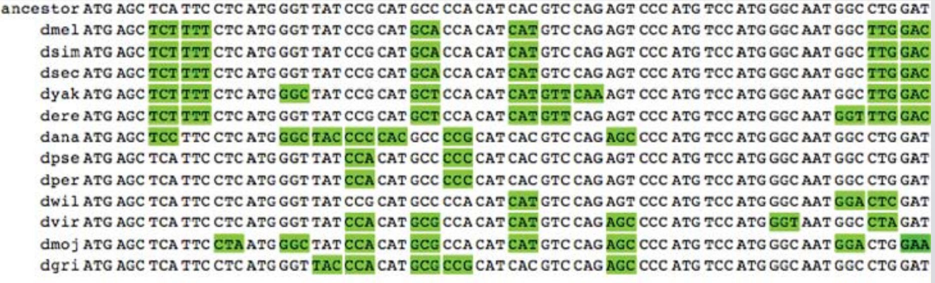


$\boldsymbol{Q}_C$ estimated from known coding regions

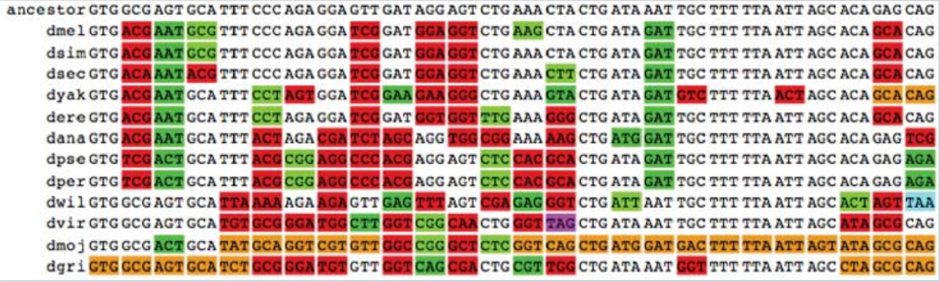


$\boldsymbol{Q}_N$ estimated from non-coding regions

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



```
ancestor ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dmel ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsim ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dsec ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCA CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dyak ATG AGC TCT TTT CTC ATG GGC TAT CCG CAT GCT CCA CAT CAT GTT CAA AGT CCC ATG TCC ATG GGC AAT GGC TTG GAC
    dere ATG AGC TCT TTT CTC ATG GGT TAT CCG CAT GCT CCA CAT CAT GTT CAG AGT CCC ATG TCC ATG GGC AAT GGT TTG GAC
    dana ATG AGC TCC TTC CTC ATG GGC TAC CCC CAC GCC CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dpse ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dper ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCC CCC CAT CAC GTC CAG AGT CCC ATG TCC ATG GGC AAT GGC CTG GAT
    dwil ATG AGC TCA TTC CTC ATG GGT TAT CCG CAT GCC CCA CAT CAT GTC CAG AGT CCC ATG TCC ATG GGC AAT GGA CTC GAT
    dvir ATG AGC TCA TTC CTC ATG GGT TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGT AAT GGC CTA GAT
    dmoj ATG AGC TCA TTC CTA ATG GGC TAT CCA CAT GCG CCA CAT CAT GTC CAG AGC CCC ATG TCC ATG GGC AAT GGA CTG GAA
    dgri ATG AGC TCA TTC CTC ATG GGT TAC CCA CAT GCG CCG CAT CAC GTC CAG AGC CCC ATG TCC ATG GGC AAT GGC CTG GAT
```



```
ancestor GTG GCG AGT GCA TTT CCC AGA GGA GTT GAT AGG AGT CTG AAA CTA CTG ATA AAT TGC TTT TTA ATT AGC ACA GAG CAG
    dmel GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAG CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsim GTG ACG AAT GCG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTA CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dsec GTG ACA AAT ACG TTT CCC AGA GGA TCG GAT GGA GGT CTG AAA CTT CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dyak GTG ACG AAT GCA TTT CCT AGT GGA TCG GAA GAA GGG CTG AAA GTA CTG ATA GAT GTC TTT TTA ACT AGC ACA GCA CAG
    dere GTG ACG AAT GCA TTT CCT AGA GGA TCG GAT GGT GGT TTG AAA GGG CTG ATA GAT TGC TTT TTA ATT AGC ACA GCA CAG
    dana GTG ACG AAT GCA TTT ACT AGA CGA TCT AGC AGG TGG CGG AAA AAG CTG ATG GAT TGC TTT TTA ATT AGC ACA GAG TCG
    dpse GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dper GTG TCG ACT GCA TTT ACG CGG AGG CCC ACG AGG AGT CTC CAC GCA CTG ATA GAT TGC TTT TTA ATT AGC ACA GAG AGA
    dwil GTG GCG AGT GCA TTA AAA AGA AGA GTT GAG TTT AGT CGA GAG GGT CTG ATT AAT TGC TTT TTA ATT AGC ACT AGT TAA
    dvir GTG GCG AGT GCA TGT GCG GGA TGG CTT GGT CGG CAA CTG GGT TAG CTG ATA AAT TGC TTT TTA ATT AGC ATA GCG CAG
    dmoj GTG GCG ACT GCA TAT GCA GGT CGT GTT GGC CGG GCT CTC GGT CAG CTG ATG GAT GAC TTT TTA ATT AGT ATA GCG CAG
    dgri GTG GCG AGT GCA TCT GCG GGA TGT GTT GGT CAG CGA CTG CGT TGG CTG ATA AAT GGT TTT TTA ATT AGC CTA GCG CAG
```

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

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

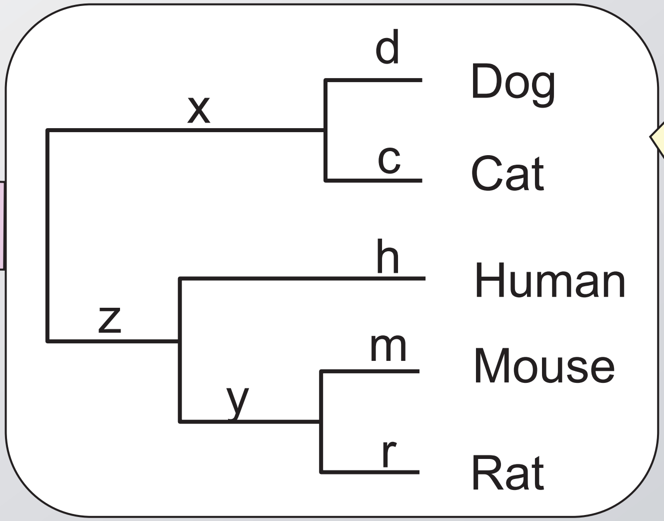
$$\Pr(\text{Leaves}; \mathbf{Q}_C, \underline{t}) = \frac{1}{10^{275}}$$

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

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}



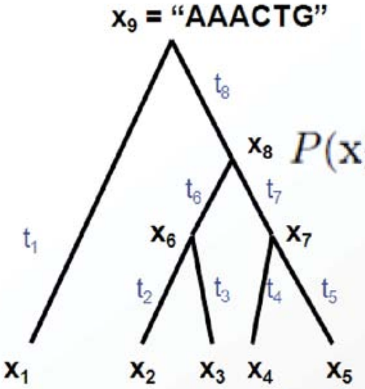
Map distances D_{ij} to a tree

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

Goal:

Minimize discrepancy between observed distances and tree-based distances

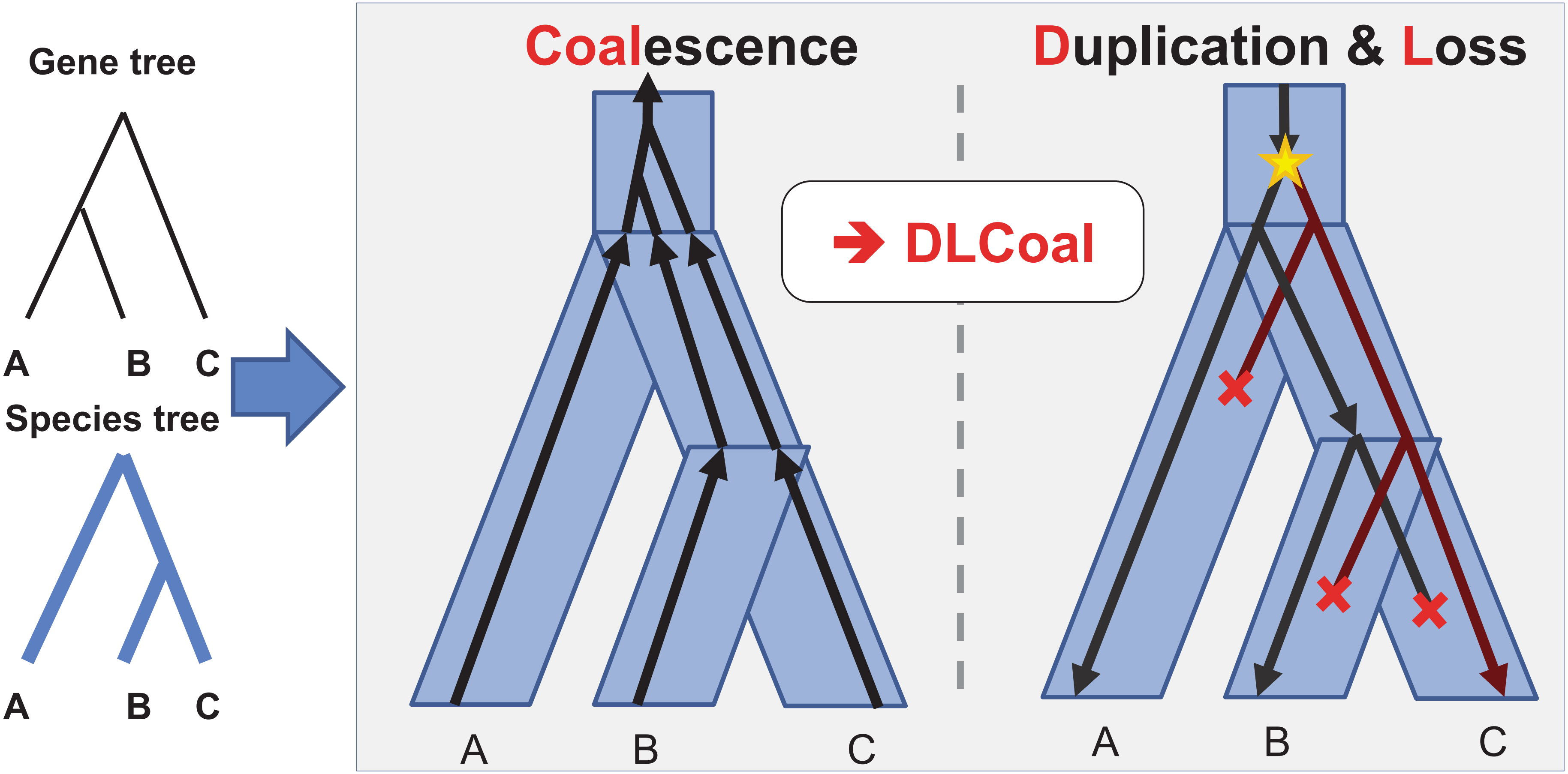
‘Peeling’ algorithm for P(D|B,T) term



$$
\begin{aligned}
P(\mathrm{x}_1,\ldots,\mathrm{x}_{2n-1}|T,\mathbf{t}) &= P(\mathrm{x}_1|\mathrm{x}_2,\ldots,\mathrm{x}_{2n-1},T,\mathbf{t})P(\mathrm{x}_2|\mathrm{x}_3,\ldots,\mathrm{x}_{2n-1},T,\mathbf{t})\ldots P(\mathrm{x}_{2n-1}|T,\mathbf{t}) \\
&= P(\mathrm{x}_1|\mathrm{x}_{\mathrm{parent}(1)},t_1)P(\mathrm{x}_2|\mathrm{x}_{\mathrm{parent}(2)},t_2)\ldots P(\mathrm{x}_{2n-1}) \\
&= P(\mathrm{x}_{2n-1})\prod_{i=1}^{2n-2} P(\mathrm{x}_i|\mathrm{x}_{\mathrm{parent}(i)},t_i)
\end{aligned}
$$

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 $\mathbf{P(x_i|x_{parent(i)},t_i)}$ 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

6.047/6.878/HST.507 - Fall 201 - Lectures: Manolis Kellis. All lectures on Tuesday/Thursday at 1pm-2:30pm
All homeworks due on Tuesday at 8pn

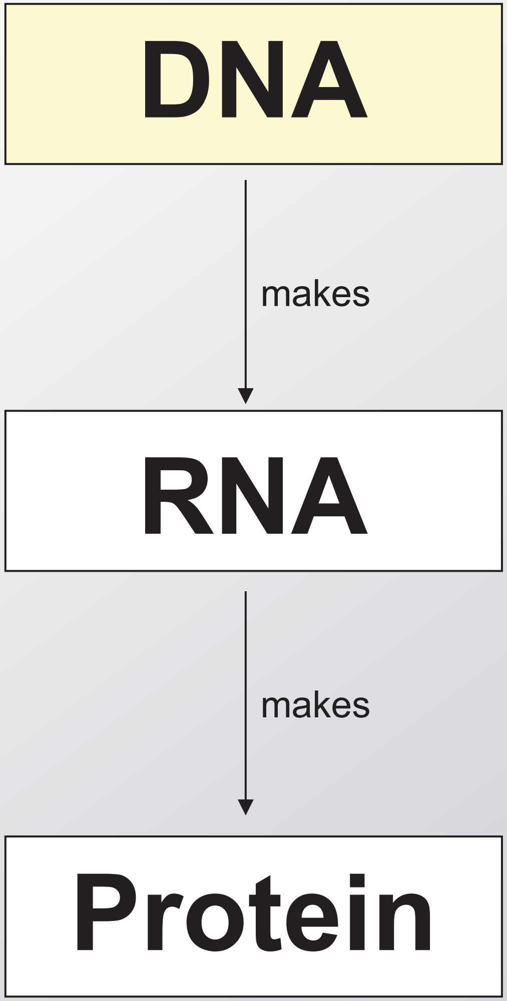
Project	Psets	Week	Date	Topic		Lec	Topic
Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. **Project profile due Tue 9/29**	PS1 out on:L1-L5	1	Thu, Sep 10	**Introduction**		L1	Intro: Biology, Algorithms, Machine Learning, Course Overview
			Fri, Sep 11			R1	Recitation 1: Biology and Probability Review
		2	Tue, Sep 15	Module I: Aligning and Modeling Genomes	Foundations	L2	Alignment I: Dynamic Programming, Global and local alignment
			Thu, Sep 17			L3	Alignment II: Database search, Rapid string matching, BLAST, BLOSUM
			Fri, Sep 18			R2	Recitation 2: Deriving Parameters of Alignment, Multiple Alignment
		3	Tue, Sep 22		Frontiers	L4	Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms
			Thu, Sep 24			L5	Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch
	due		Fri, Sep 25				No classes - student holiday
	Tue 9/29		Fri, Sep 25		Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-50		
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 10/6**	PS2 out on:L6-R4	4	Tue, Sep 29	Module II: Gene Expression and Networks	Foundations	L6	Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr
			Thu, Oct 1			L7	Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian
			Fri, Oct 2			R3	Recitation 3: Affinity Propagation Clustering and Random Forest Classification
		5	Tue, Oct 6		Frontiers	L8	Networks I: Bayesian inference, deep learning, network dynamics
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 10/20. Presented on Fri 10/23**			Thu, Oct 8			L9	Networks II: Network learning, structure, spectral methods
	due		Fri, Oct 9			R4	Recitation 4: Small and Large Regulatory RNAs: lincRNA, miRNA, piRNA…
	Tue 10/13		Fri, Oct 9		Project Planning: research areas, initial ideas, type of project, mentor matching, finding partners 32D		
	PS3 out on:L10-R6	6	Tue, Oct 13	Module III: Gene Regulation & Epigenomics	Foundations		No Classes - Monday Schedule - October 13, 2015
			Thu, Oct 15			L10	Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM
			Fri, Oct 16			R5	Recitation 5: Gapped Motif Discovery, DNAShape, PBMs, Selex
		7	Tue, Oct 20		Frontiers	L11	Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states
			Thu, Oct 22			L12	RNA modifications: RNA editing, Translation regulation, Splicing regulation
	due		Fri, Oct 23			R6	Recitation 6: Dimensionality Reduction
	Tue 10/27		Fri, Oct 23		Project feedback: Prepare 2-3 slide presentation of your term project for your mentor. 32D-507 at 4-5		
Evaluate/discuss three peer proposals, NIH review format. **Review Panels Fri 10/30** **Reviews back Tue 11/3**	PS4 out on:L13-R8	8	Tue, Oct 27	Module IV: Population and Disease Genetics	Foundations	L13	Resolving human ancestry and human history from genetic data
			Thu, Oct 29			L14	Disease Association Mapping, GWAS, organismal phenotypes
			Fri, Oct 30			R7	Recitation 7: Robinson-Foulds Distance and Coalescent Process
			Fri, Oct 30		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 11/12**		9	Tue, Nov 3		Frontiers	L15	Quantitative trait mapping, molecular traits, eQTLs
	due		Thu, Nov 5			L16	Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment
	Tue 11/10		Fri, Nov 6			R8	Recitation 8: Suffix Trees and Arrays
	PS5 out on:L17-R10	10	Tue, Nov 10	Module V: Comparative Genomics and Evolution	Foundations		No lecture, veterans day holiday - Monday/Tuesday
			Thu, Nov 12			L17	Comparative genomics and Evolutionary signatures
Continue making substantial progress on proposed milstones.Write outline of final report. **Midcourse report due Thu 11/19. Score projection 11/24**			Fri, Nov 13			R9	Recitation 9: Review of Phylogeny and Molecular Evolution
		11	Tue, Nov 17		Frontiers	L18	Phylogenetics: Molecular evolution, Tree building, Phylogenetic inference
	due		Thu, Nov 19			L19	Phylogenomics: Gene/species trees, reconciliation, recombination graphs
	Tue 12/1		Fri, Nov 20			R10	Recitation 10: Linkage Disequilibrium, Haplotype Phasing, and Genotype Imputatio
Complete your milestones, finalize results, figures, write-up in conference publication format. As par of report, comment on your overall project experience. **Written report due Sun 12/6**	No more psets! (work on your final project)	12	Tue, Nov 24	Module VI: Current Research Directions	Frontiers		In Class Quiz (the only quiz - the class has no final exam) - covers L1-R11
			Thu, Nov 26				No lecture, thanksgiving break - Thu Nov 26, 2015
			Fri, Nov 27				No recitation, thanksgiving break
		13	Tue, Dec 1			L20	Personal Genomics, Disease Epigenomics: Systems approaches to disease
			Thu, Dec 3			L21	Three-dimensional chromatin interactions: 3C, 5C, HiC, ChIA-Pet
			Fri, Dec 4			R11	Recitation 11: Project Tips - Write-up, Slides, Final Presentation in 32D-507
		14	Tue, Dec 8			L22	Genome Engineering with CRISPR/Cas9 and related technologies
Conference format slide pres. **Talks on Thu 12/10**			Thu, Dec 10				Final Presentations - Part I (1pm). 32-141
			Thu, Dec 10				Final Presentations - Part II (3pm). 32D-507

* readings refer to chapters in compiled 2014 scribe notes, available in the materials folder
** 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

makes

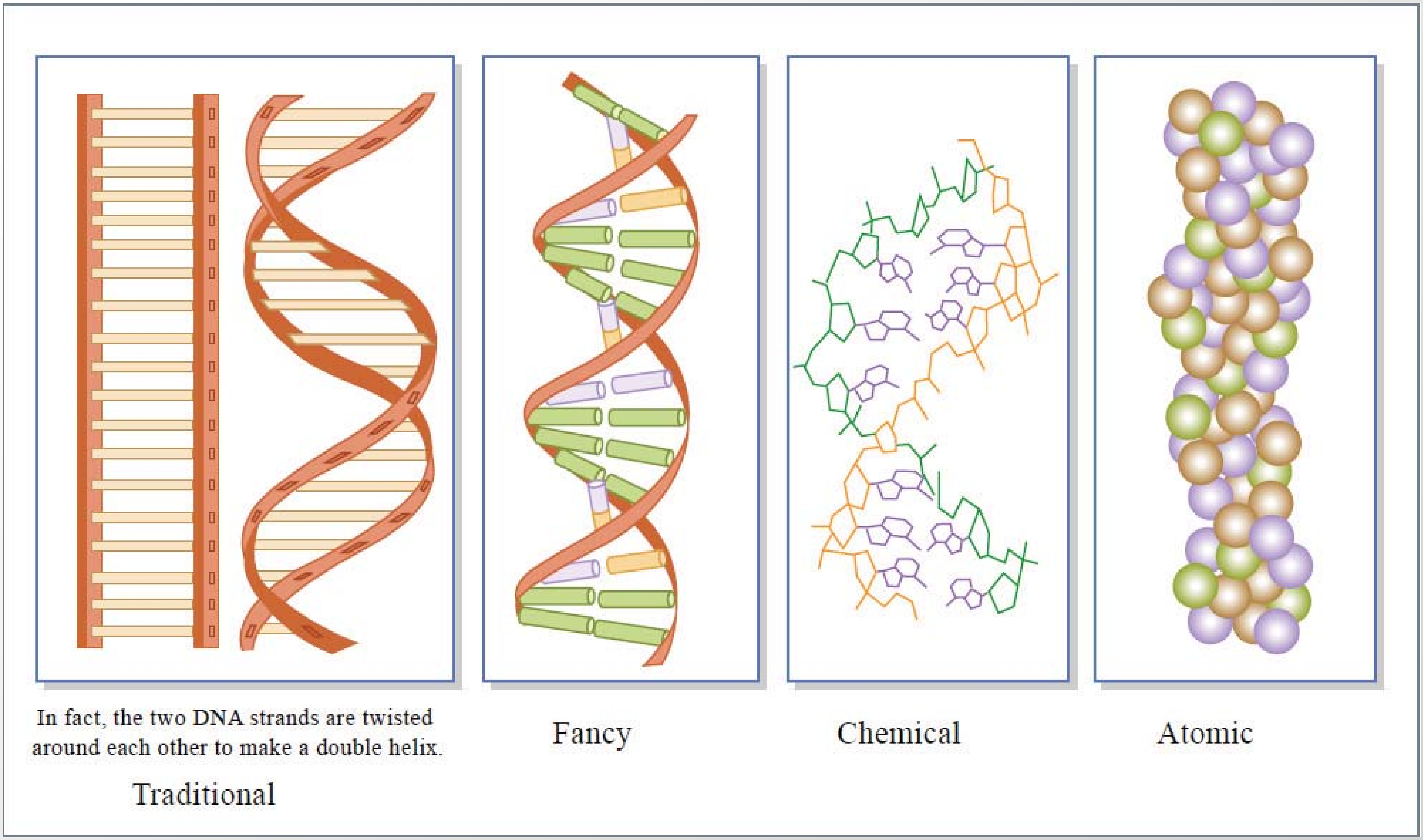
RNA

makes

Protein

DNA: The double helix

- The most noble molecule of our time



In fact, the two DNA strands are twisted around each other to make a double helix.

Traditional

Fancy

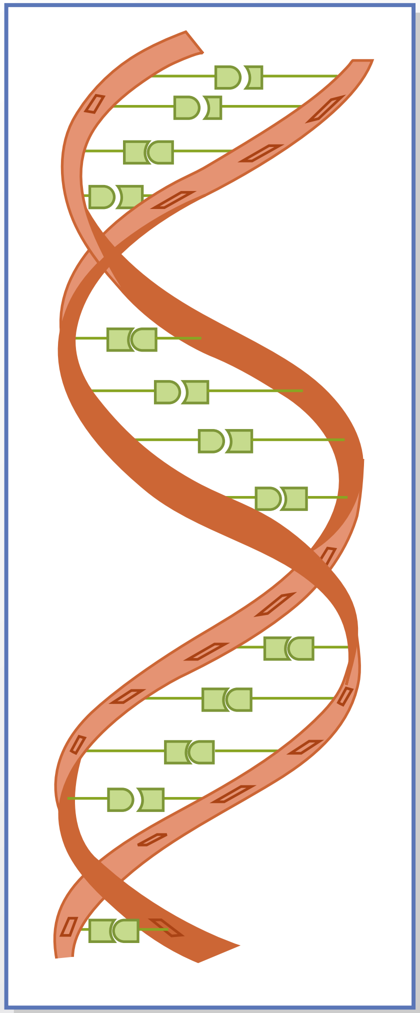
Chemical

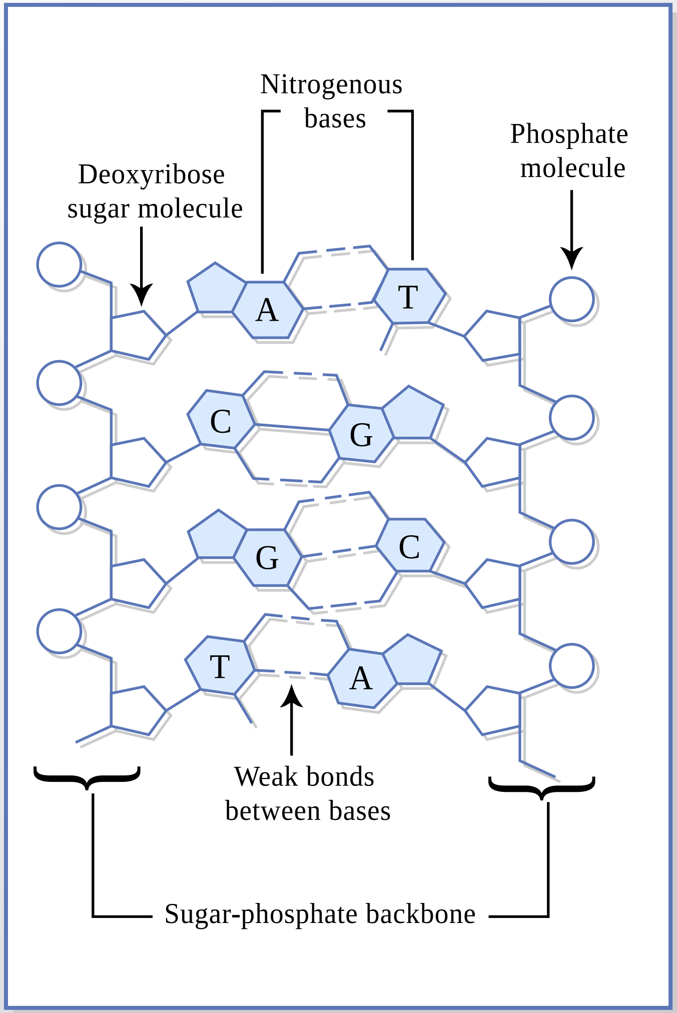
Atomic

Image by MIT OpenCourseWare.

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





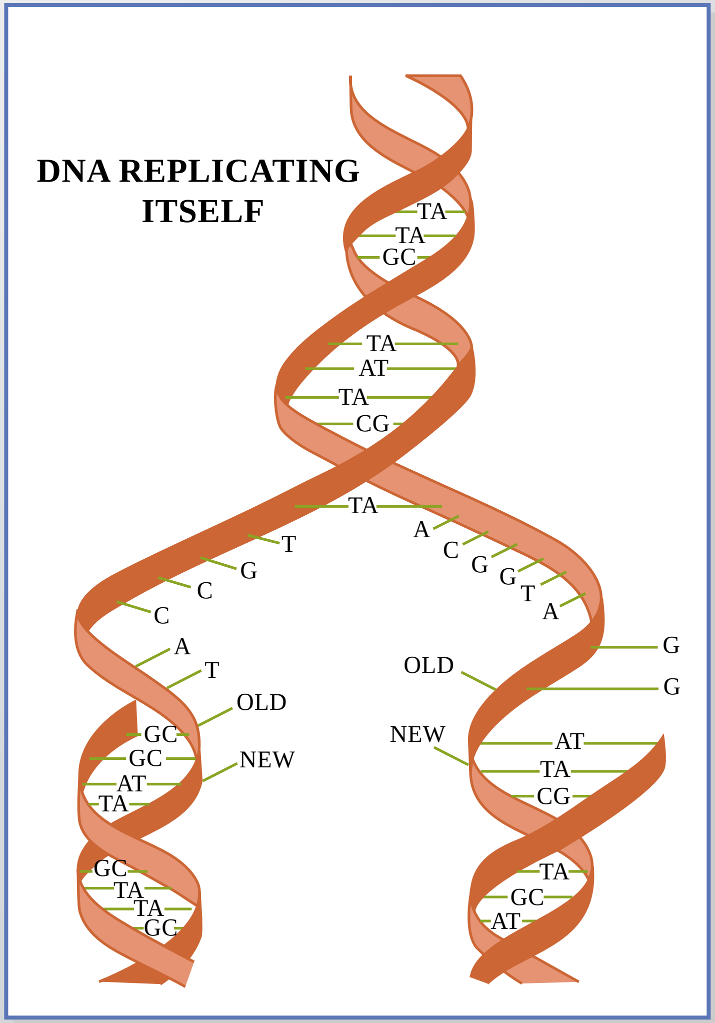
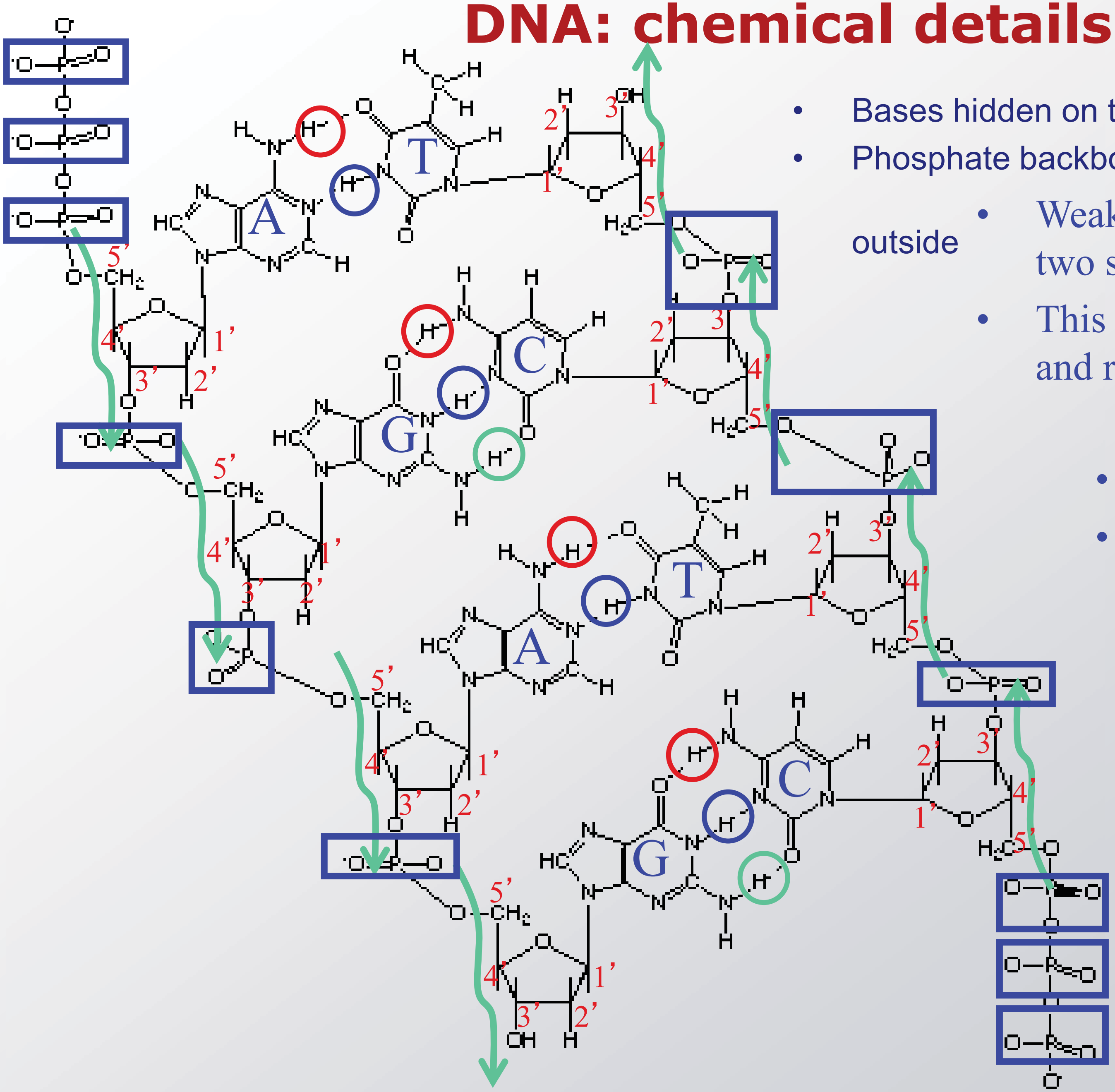


Image by MIT OpenCourseWare.

DNA: chemical details



- Bases hidden on the inside
- Phosphate backbone outside

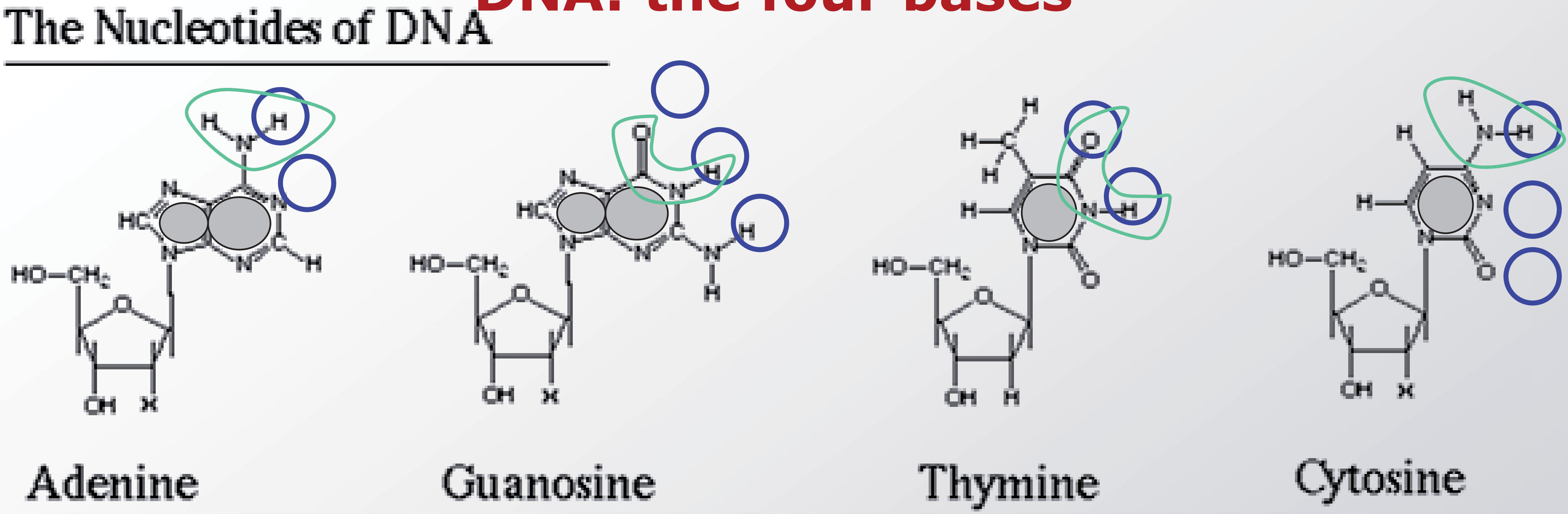
- Weak hydrogen bonds hold the two strands together
- This allows low-energy opening and re-closing of two strands

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

The only parings are:

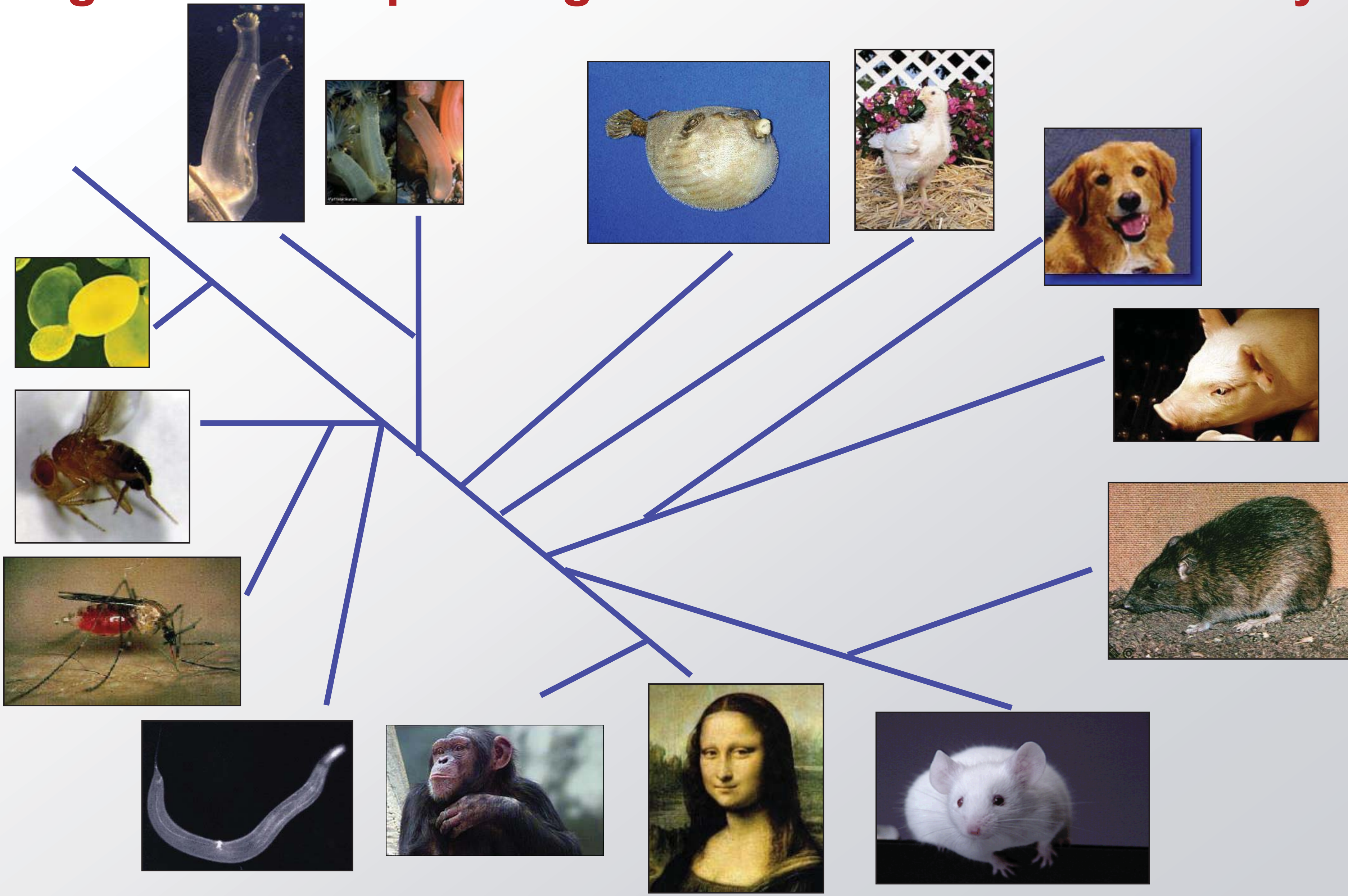
- A with T
- C with G

DNA: the four bases



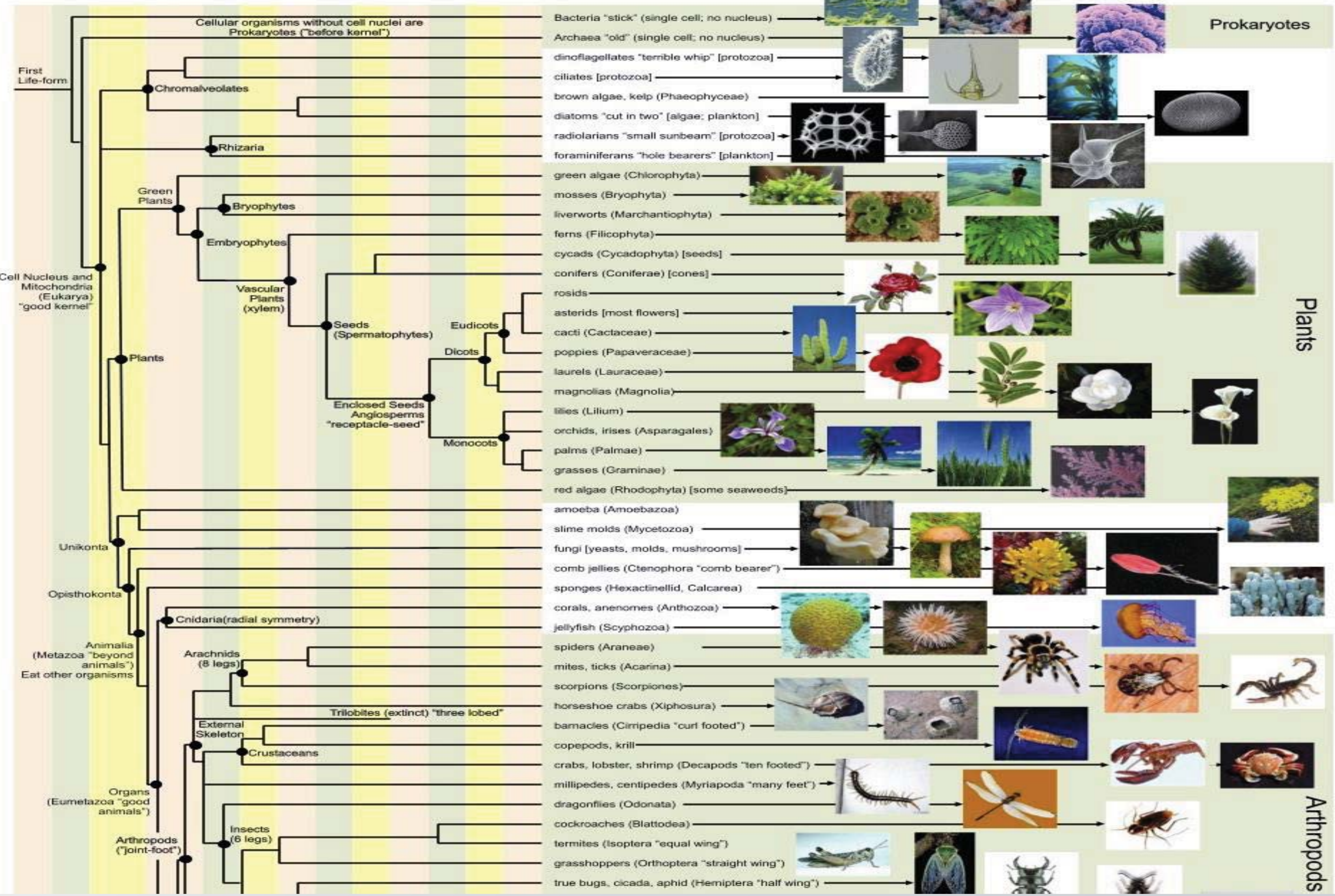
Purine	Purine		
		Pyrimidine	Pyrimidine
Weak		Weak	
	Strong		Strong
Amino			Amino
	Keto	Keto	

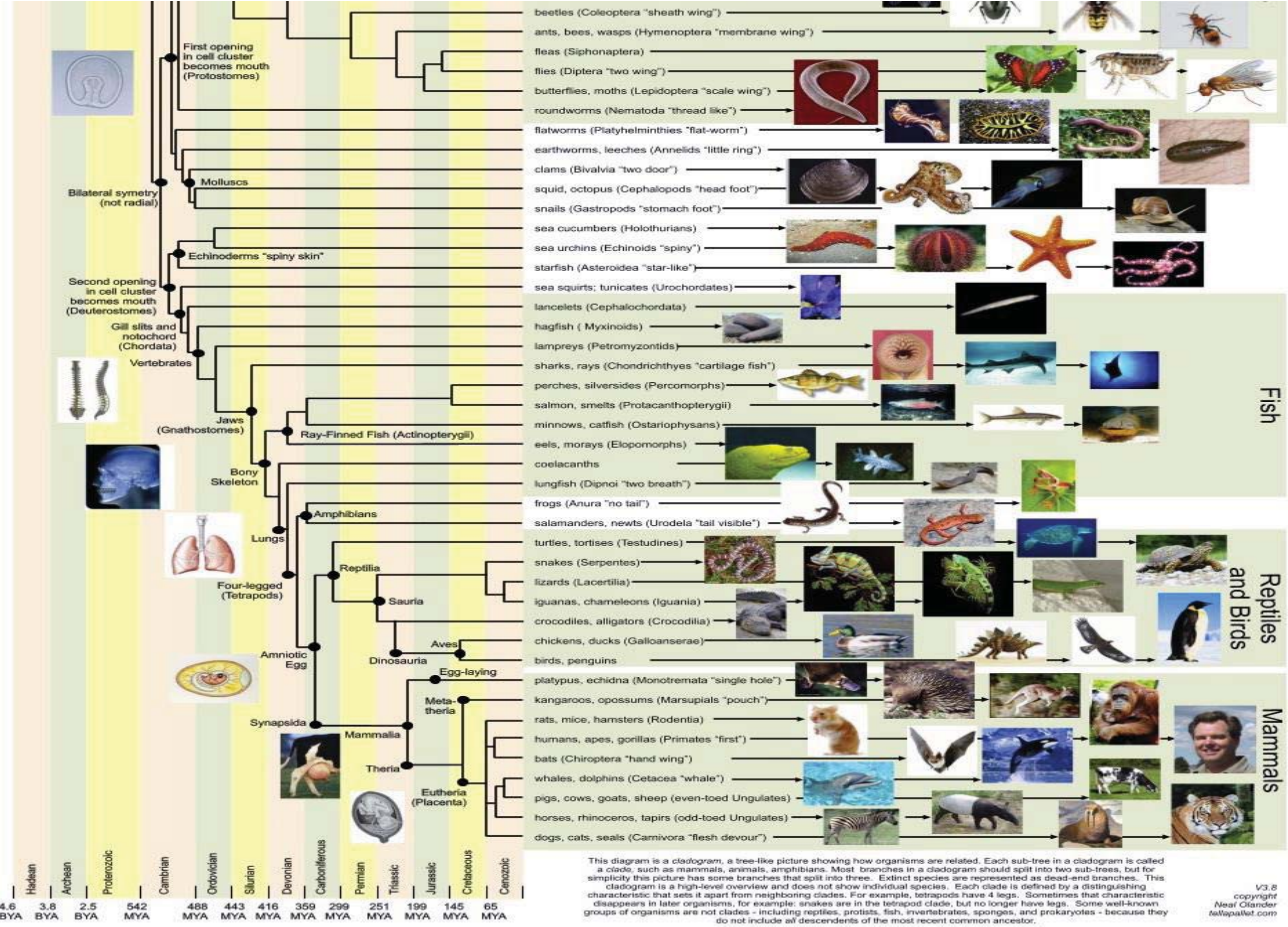
Alignment: all species/genes share common ancestry



Slide credit: Serafim Batzoglou

Tree of Life


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


beetles (Coleoptera "sheath wing")
ants, bees, wasps (Hymenoptera "membrane wing")
fleas (Siphonaptera)
flies (Diptera "two wing")
butterflies, moths (Lepidoptera "scale wing")
roundworms (Nematoda "thread like")
flatworms (Platyhelminthies "flat-worm")
earthworms, leeches (Annelids "little ring")
clams (Bivalvia "two door")
squid, octopus (Cephalopods "head foot")
snails (Gastropods "stomach foot")
sea cucumbers (Holothurians)
sea urchins (Echinoids "spiny")
starfish (Asteroidea "star-like")
sea squirts; tunicates (Urochordates)
lancelets (Cephalochordata)
hagfish (Myxinoids)
lampreys (Petromyzontids)
sharks, rays (Chondrichthyes "cartilage fish")
perches, silversides (Percomorphs)
salmon, smelts (Protacanthopterygii)
minnows, catfish (Ostariophysans)
eels, morays (Elopomorphs)
coelacanths
lungfish (Dipnoi "two breath")
frogs (Anura "no tail")
salamanders, newts (Urodela "tail visible")
turtles, tortises (Testudines)
snakes (Serpentes)
lizards (Lacertilia)
iguanas, chameleons (Iguania)
crocodiles, alligators (Crocodilia)
chickens, ducks (Galloanserae)
birds, penguins
platypus, echidna (Monotremata "single hole")
kangaroos, opossums (Marsupials "pouch")
rats, mice, hamsters (Rodentia)
humans, apes, gorillas (Primates "first")
bats (Chiroptera "hand wing")
whales, dolphins (Cetacea "whale")
pigs, cows, goats, sheep (even-toed Ungulates)
horses, rhinoceros, tapirs (odd-toed Ungulates)
dogs, cats, seals (Carnivora "flesh devour")
Fish
Reptiles and Birds
Mammals
First opening in cell cluster becomes mouth (Protostomes)
Molluscs
Bilateral symetry (not radial)
Echinoderms "spiny skin"
Second opening in cell cluster becomes mouth (Deuterostomes)
Gill slits and notochord (Chordata)
Vertebrates
Jaws (Gnathostomes)
Ray-Finned Fish (Actinopterygii)
Bony Skeleton
Amphibians
Lungs
Reptilia
Sauria
Four-legged (Tetrapods)
Aves
Amniotic Egg
Dinosauria
Egg-laying
Meta-theria
Synapsida
Mammalia
Theria
Eutheria (Placenta)
Hadean
Archean
Proterozoic
Cambrian
Ordovician
Silurian
Devonian
Carboniferous
Permian
Triassic
Jurassic
Cretaceous
Cenozoic
4.6 BYA
3.8 BYA
2.5 BYA
542 MYA
488 MYA
443 MYA
416 MYA
359 MYA
299 MYA
251 MYA
199 MYA
145 MYA
65 MYA

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.

V3.8
copyright
Neal Olander
tellapallet.com

Extinctions part of life

Phylogenetic tree showing archosaurs, dinosaurs, birds, etc. through geologic time removed due to copyright restrictions.

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

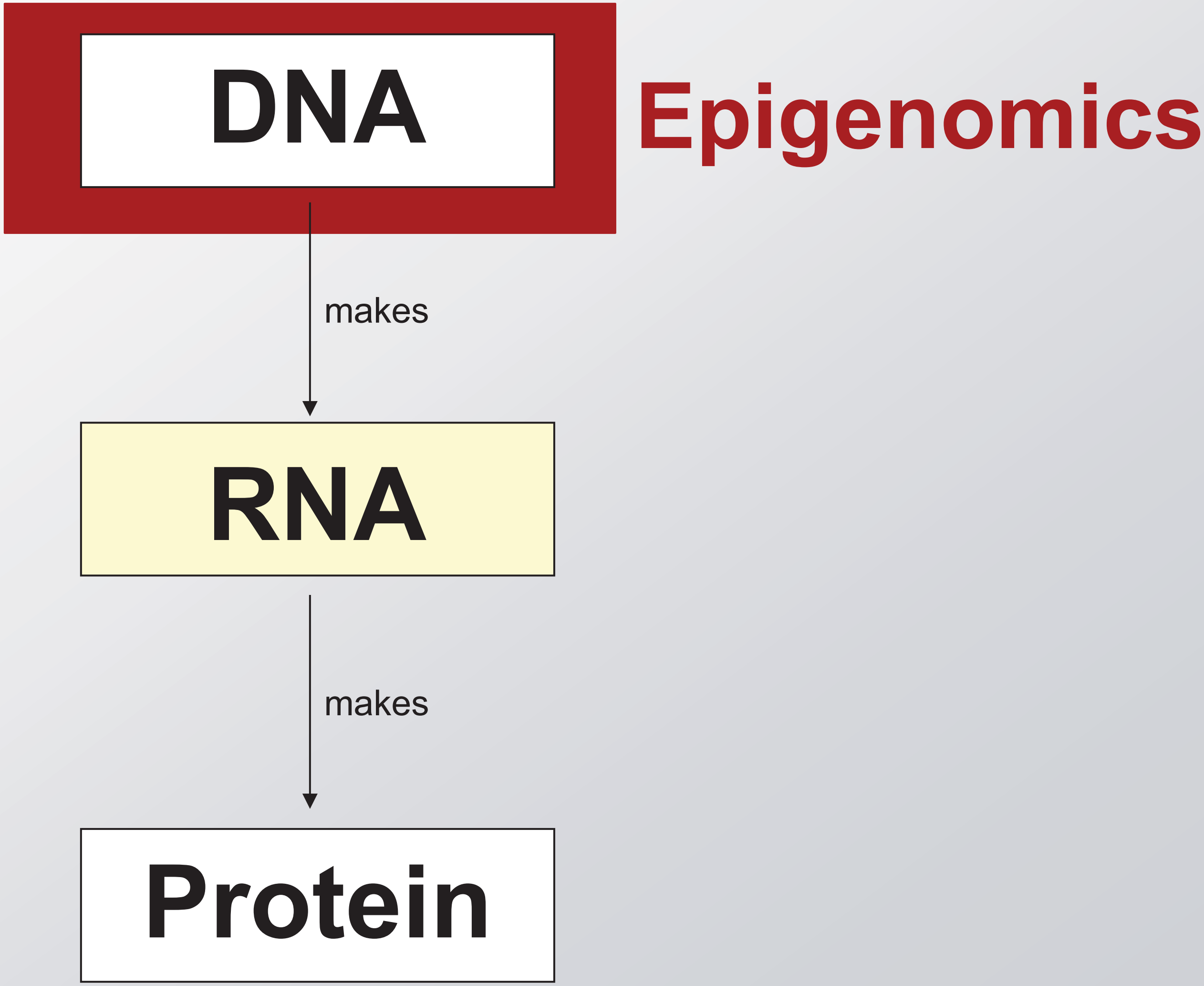
Mammal family tree removed due to copyright restrictions.

‘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)

“Central dogma” of Molecular Biology



DNA

Epigenomics

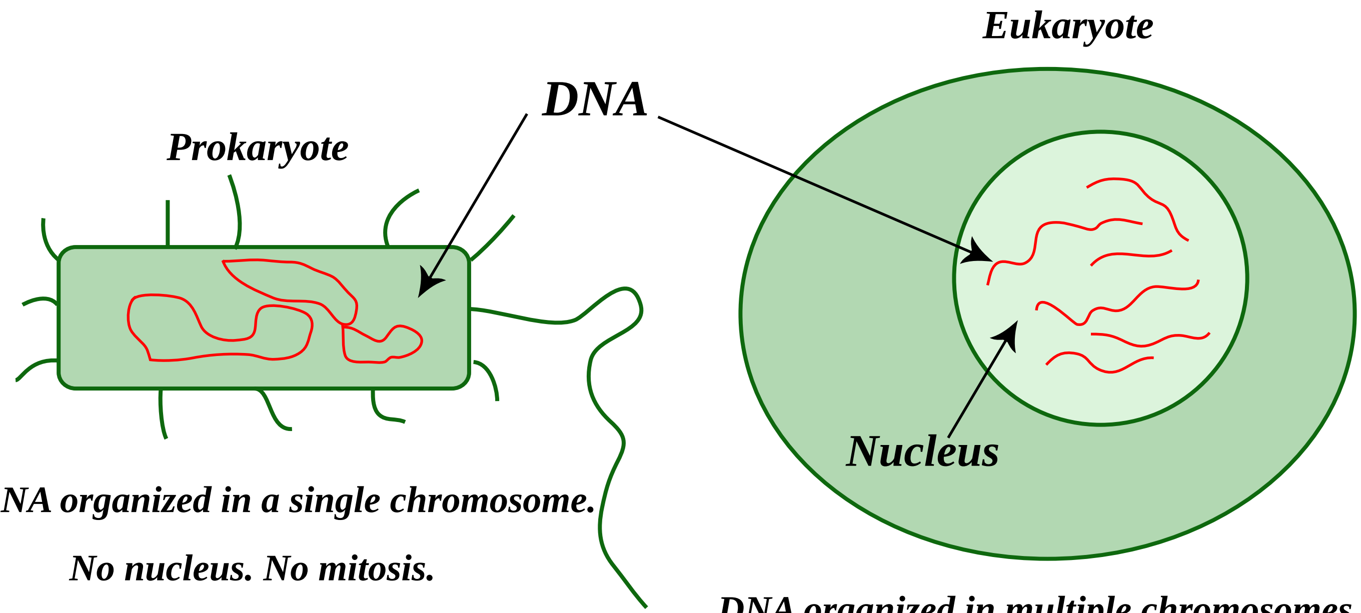
makes

RNA

makes

Protein

Chromosomes inside the cell



DNA organized in a single chromosome.

No nucleus. No mitosis.

DNA organized in multiple chromosomes inside a nucleus.

Mitotic division

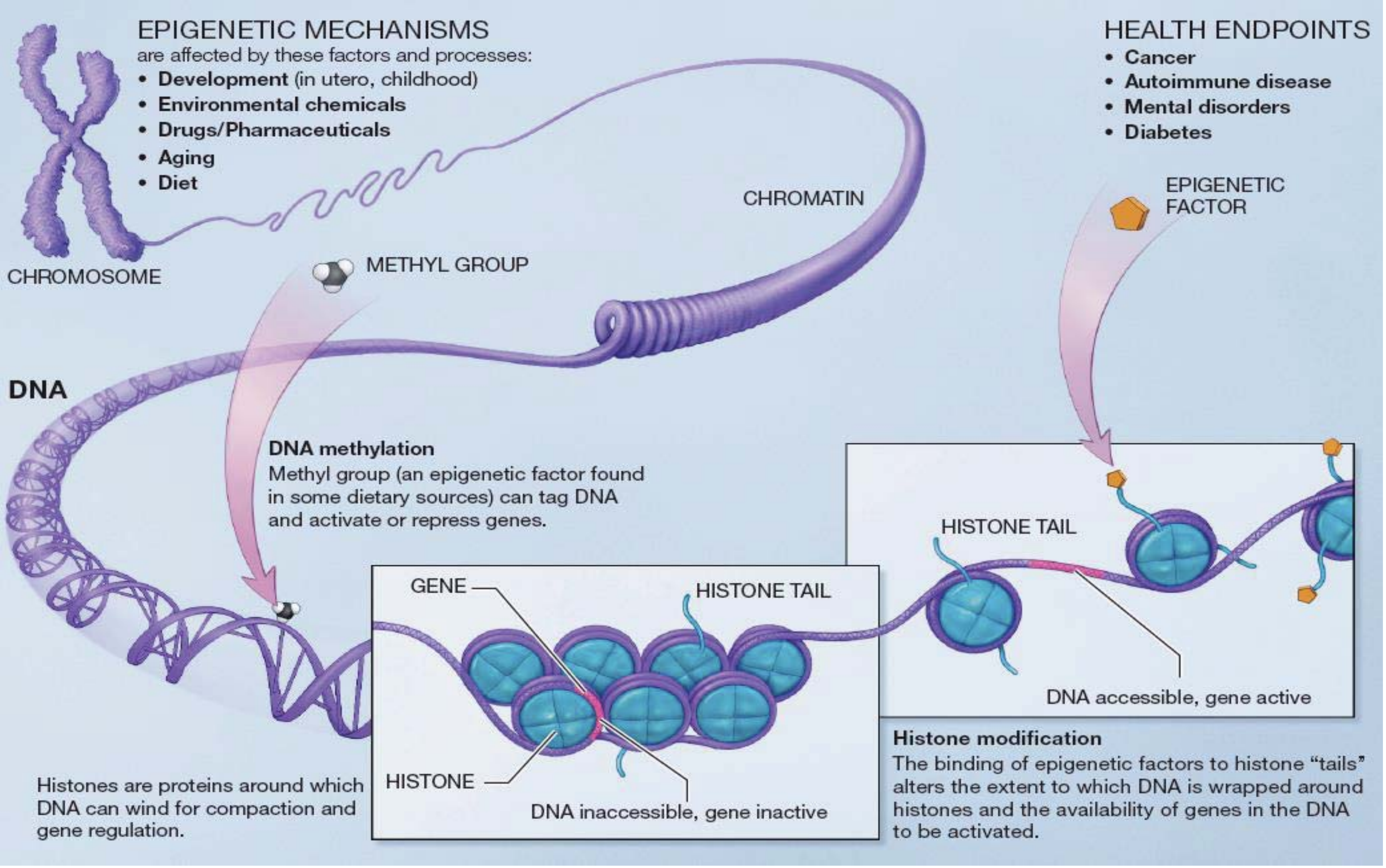
Figures by MIT OpenCourseWare.

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

Image removed due to copyright restrictions.
Please see: Figure 8-10 from Alberts, Bruce, and Martin Raff. Essential Cell Biology. New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

Diverse epigenetic modifications



EPIGENETIC MECHANISMS
are affected by these factors and processes:

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

HEALTH ENDPOINTS

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

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

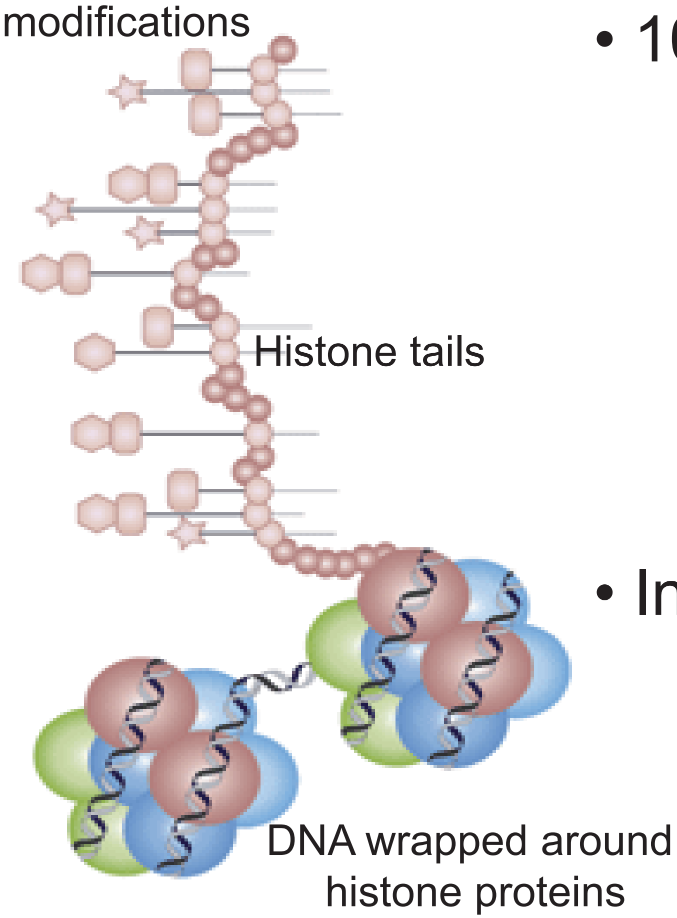
Histones are proteins around which DNA can wind for compaction and gene regulation.

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

Courtesy of the National Institutes of Health; in the public domain.

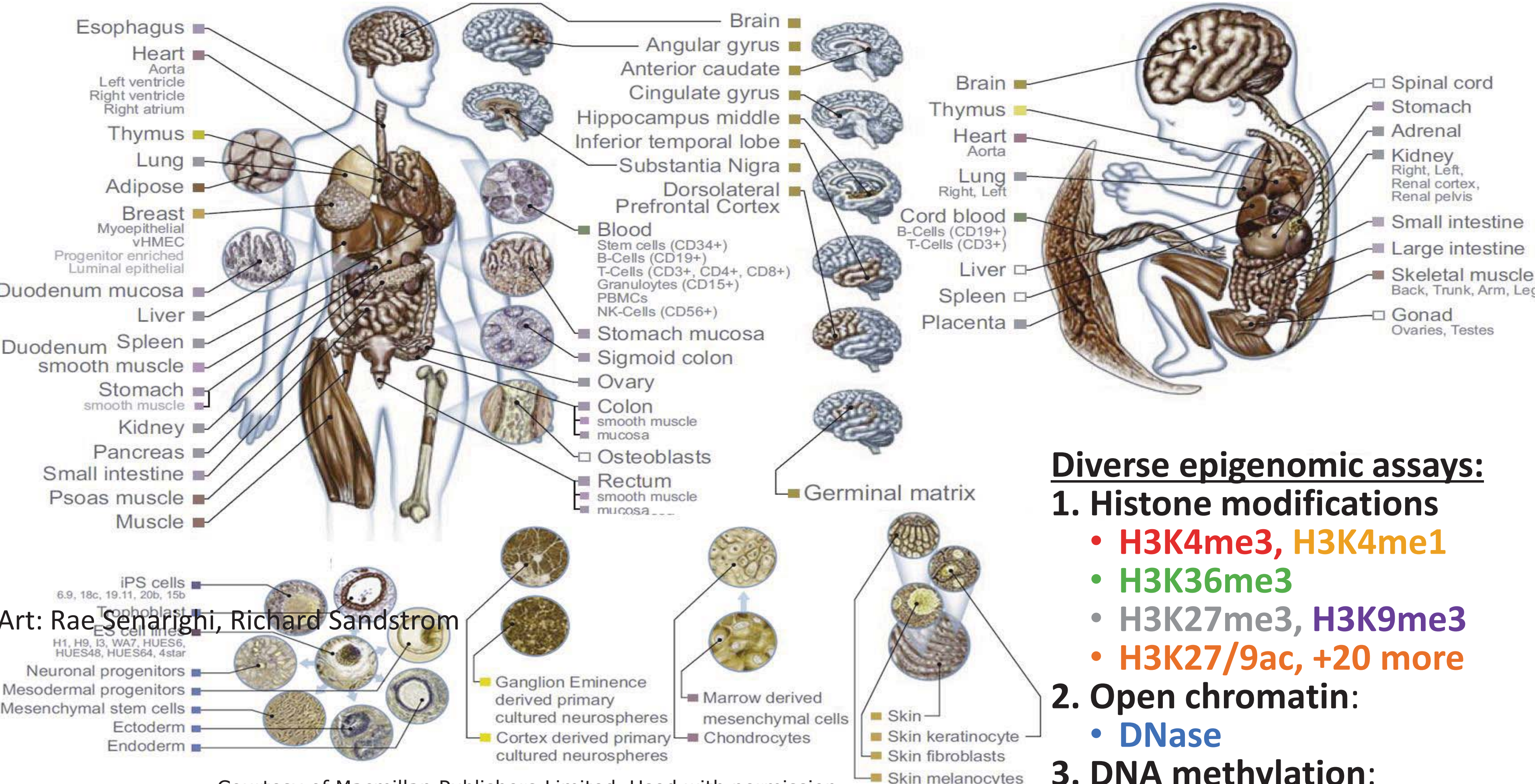
Image source: http://nihroadmap.nih.gov/epigenomics/

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

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

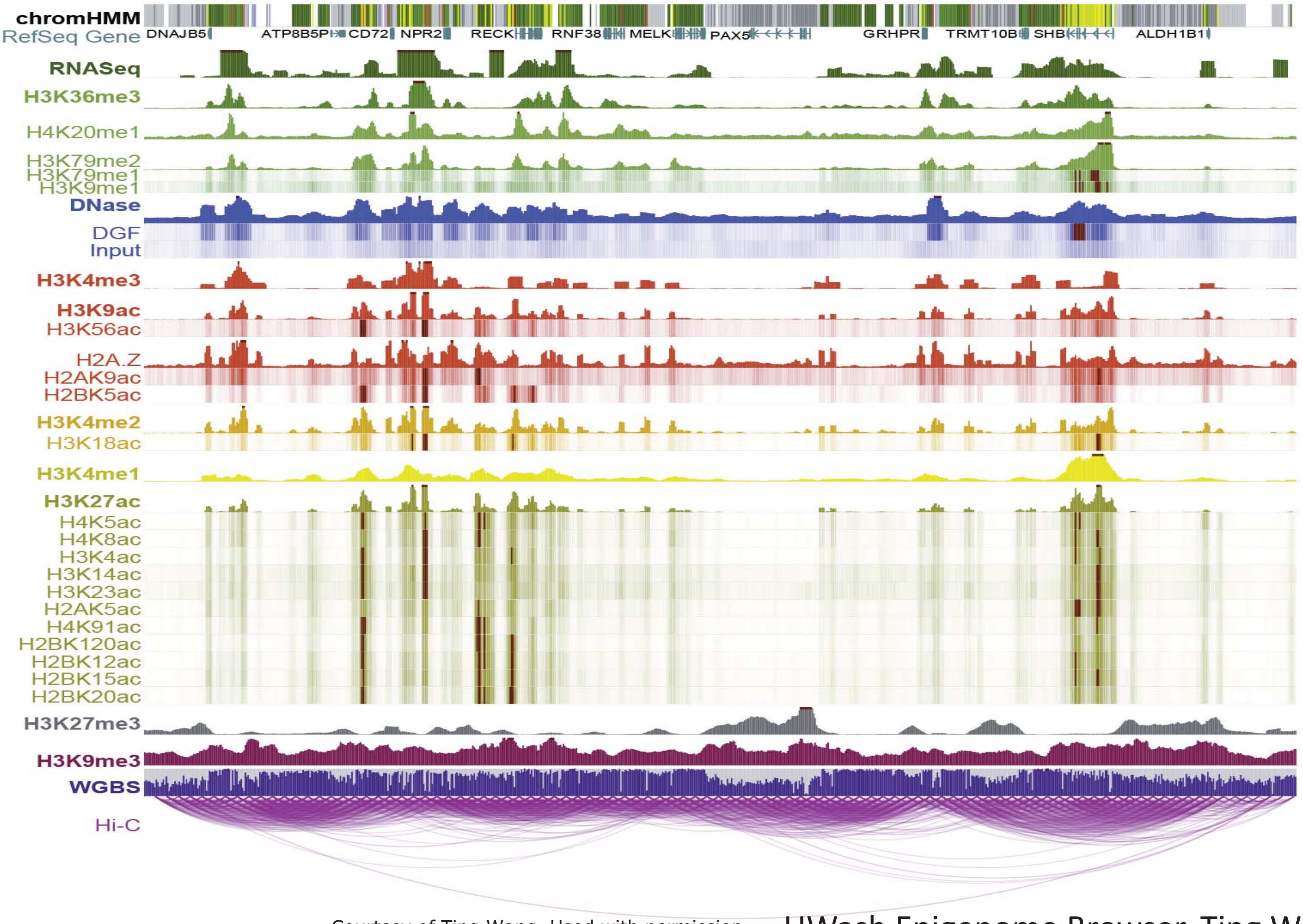
Diverse epigenomic assays:

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)

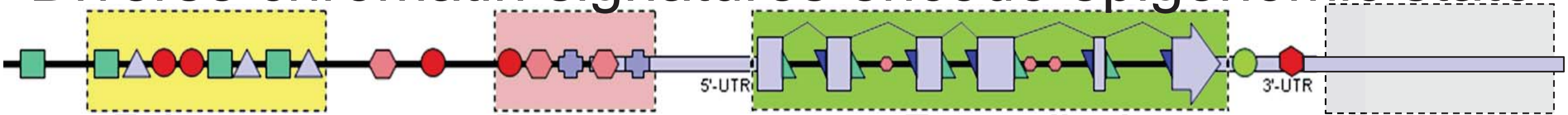


Courtesy of Ting Wang. Used with permission.

UWash Epigenome Browser, Ting Wang

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

Diverse chromatin signatures encode epigenomic state

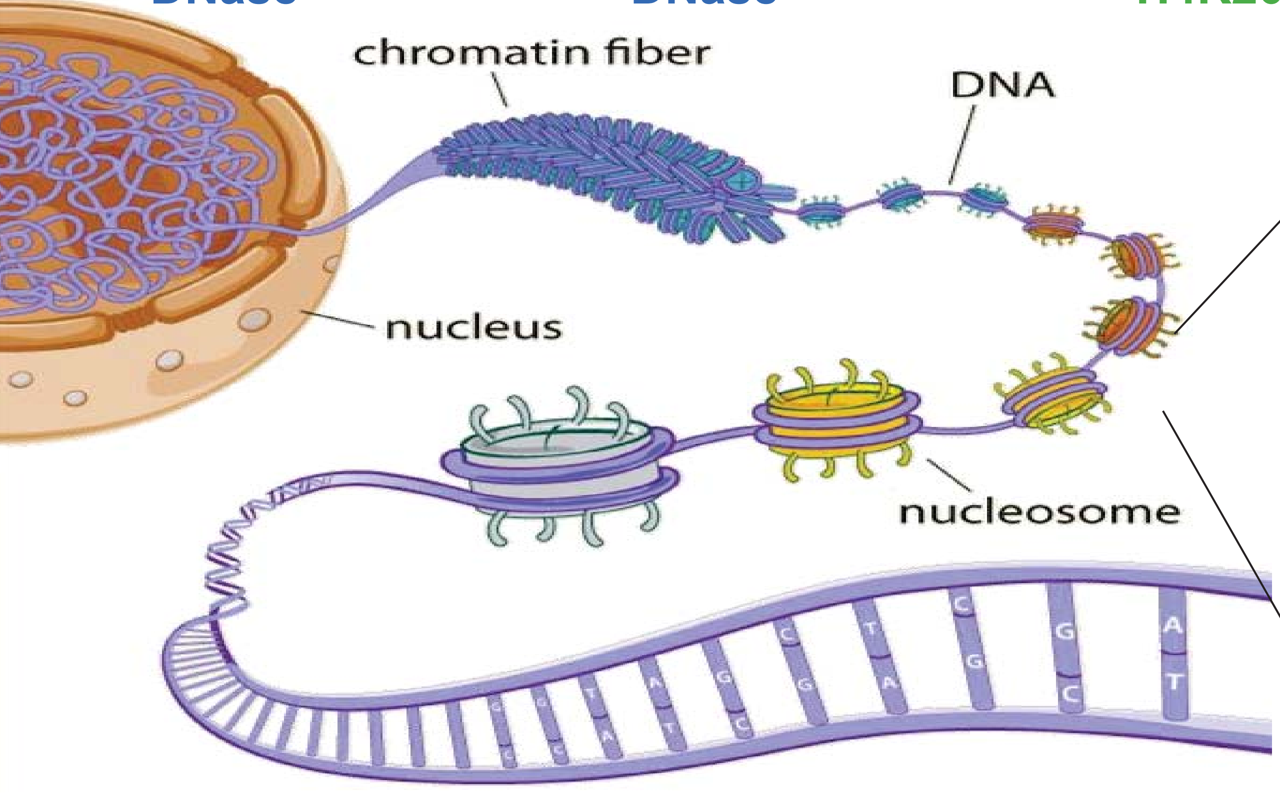


Enhancers
- H3K4me1
- H3K27ac
- DNase

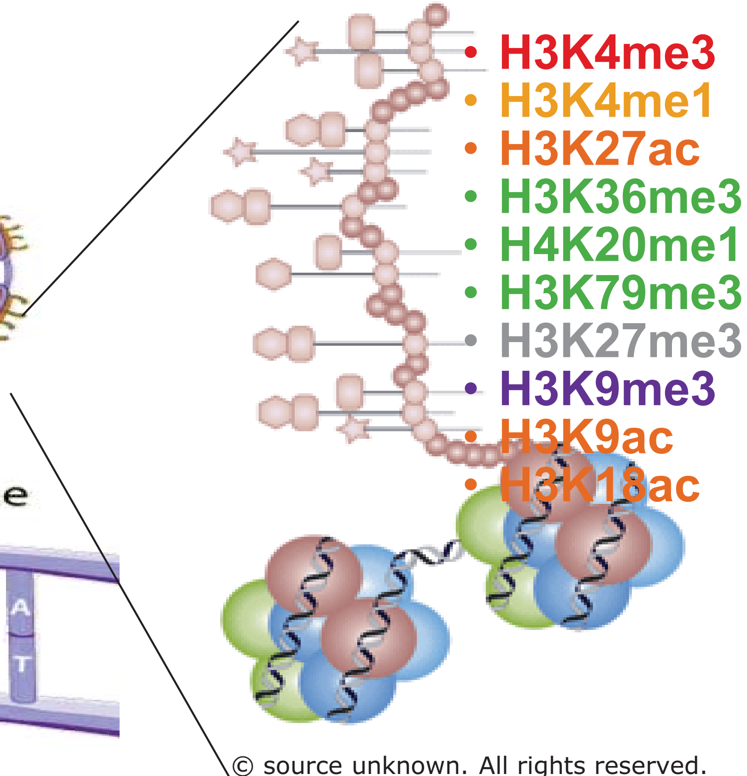
Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl



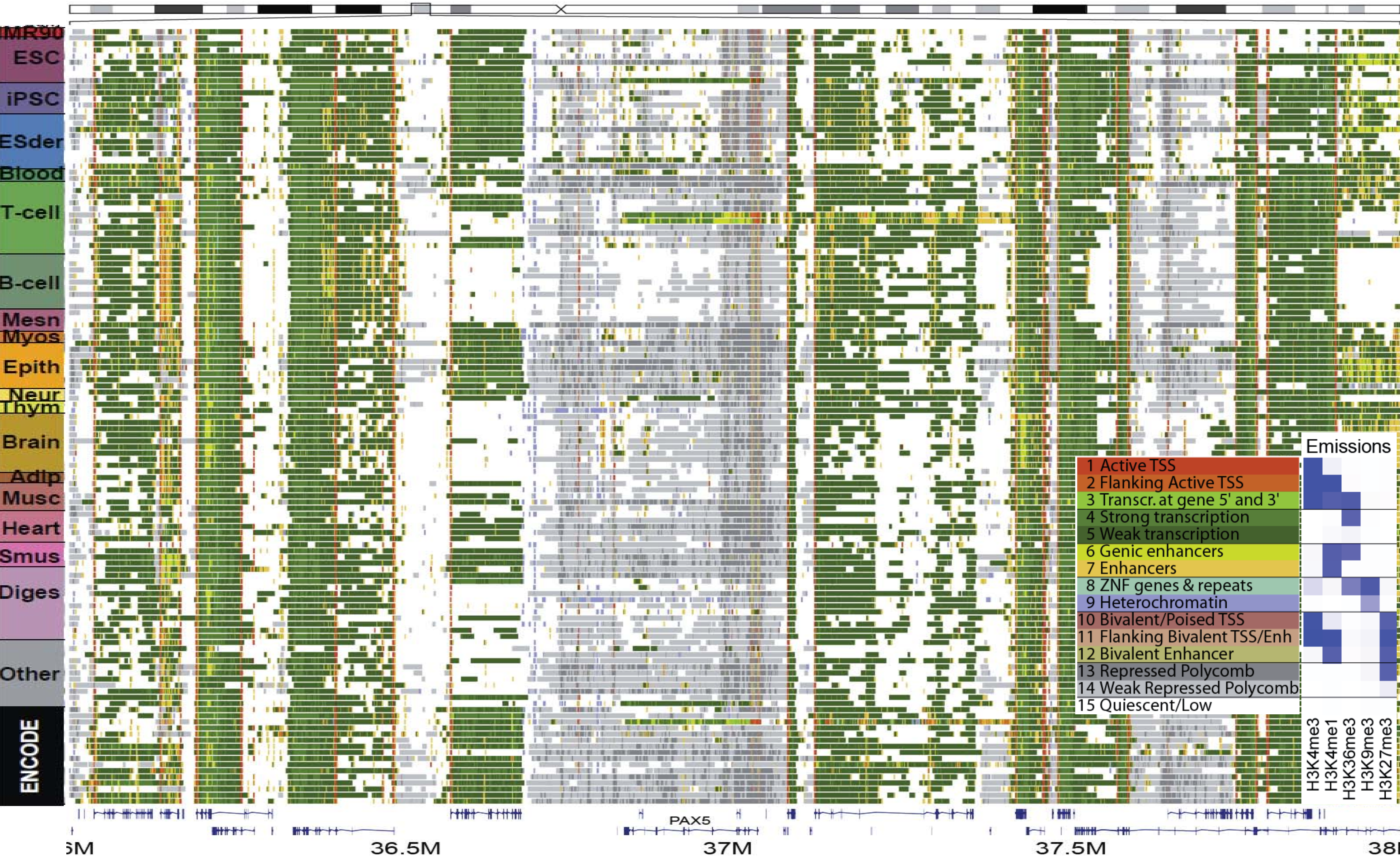
Courtesy of Broad Communications. Used with permission.



- H3K4me3
- H3K4me1
- H3K27ac
- H3K36me3
- H4K20me1
- H3K79me3
- H3K27me3
- H3K9me3
- H3K9ac
- H3K18ac

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

Chromatin state annotations across 127 epigenomes

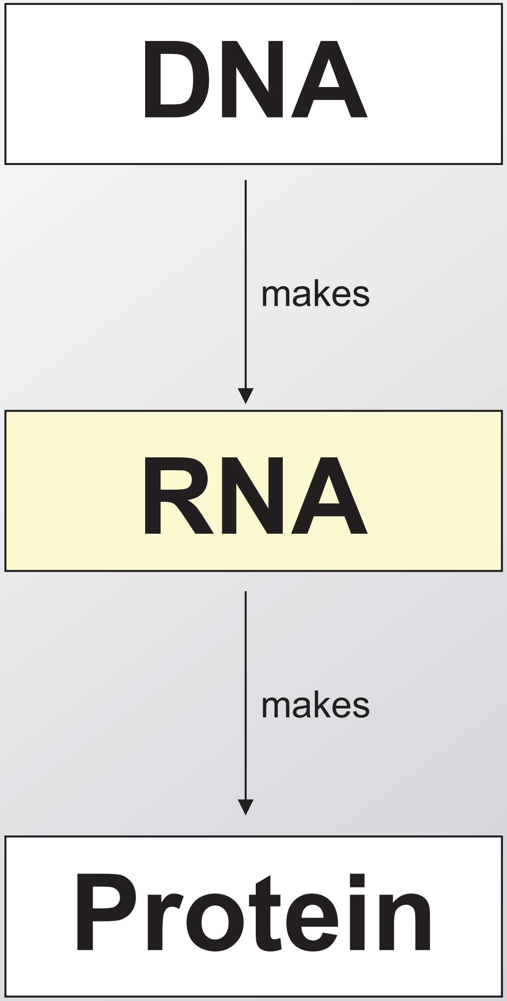


Courtesy of Anshul Kundaje. Used with permission.

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

Anshul Kundaje

“Central dogma” of Molecular Biology



DNA

makes

RNA

makes

Protein

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

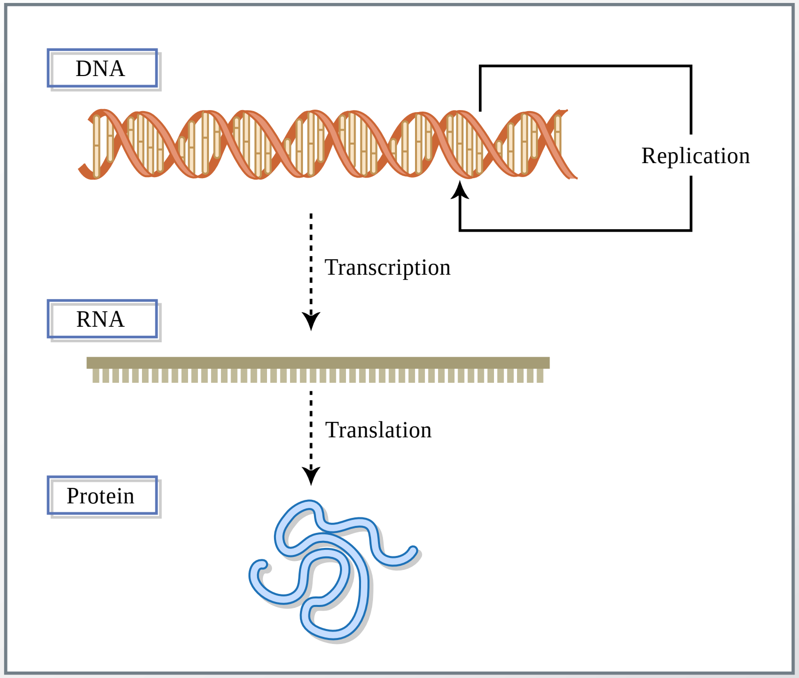


Image by MIT OpenCourseWare.

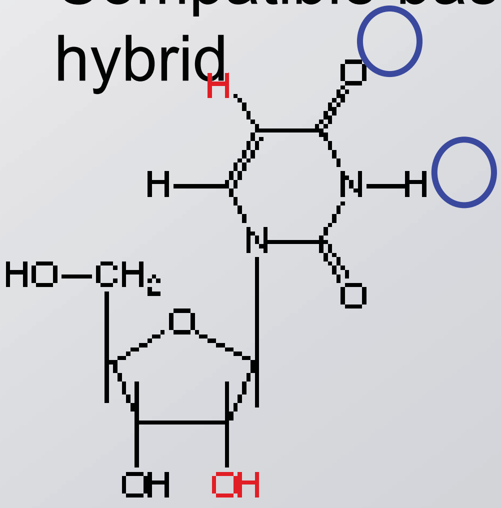
- Information changes medium
  - single strand vs. double strand
  - ribose vs. deoxyribose sugar

A T T A C G G T A C C G T

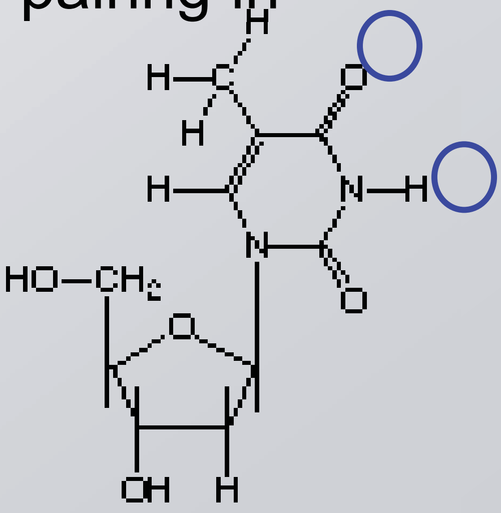
|| || || || ||| ||| ||| || || ||| ||| ||| ||

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

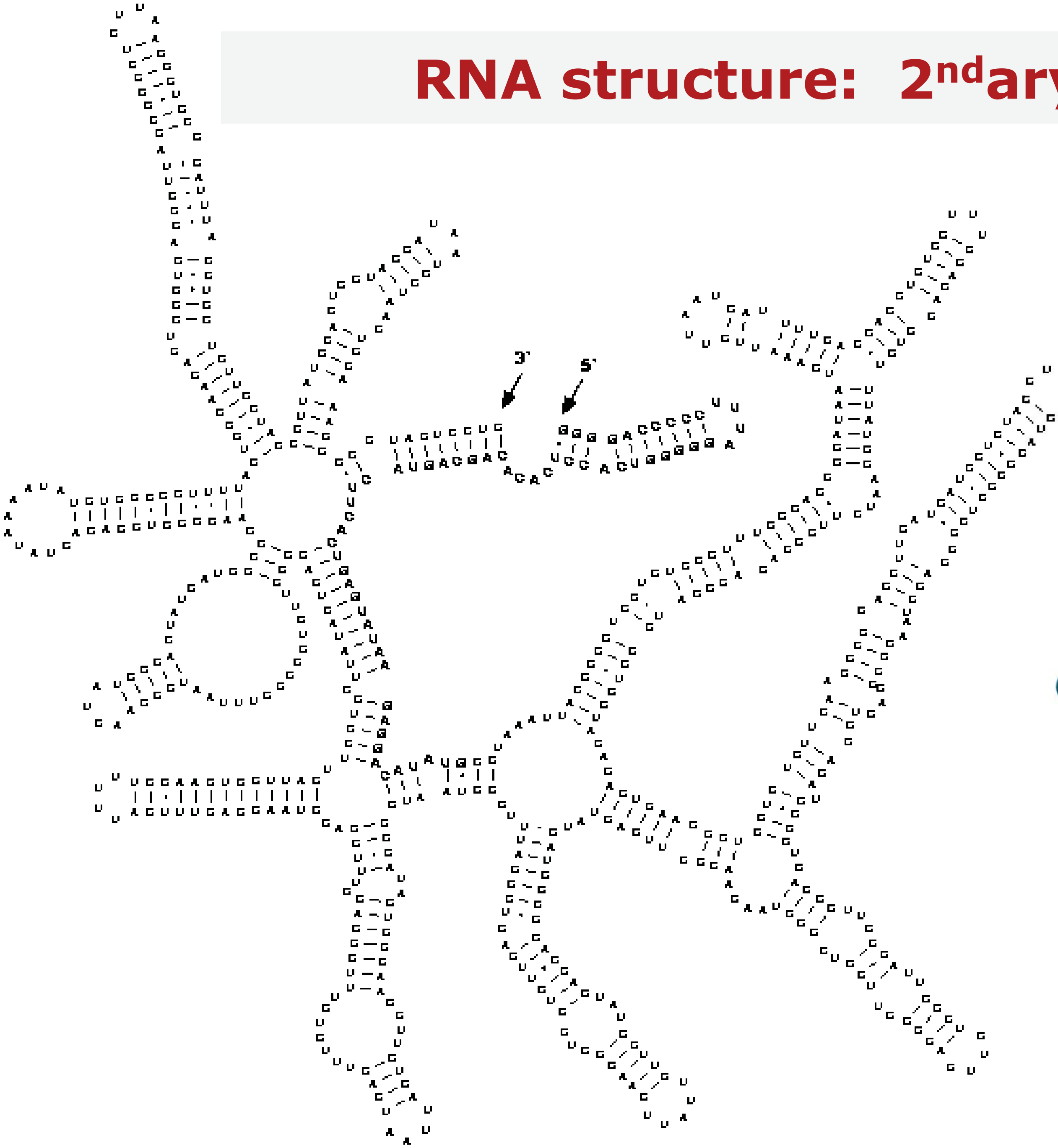
Image removed due to copyright restrictions.
Please see: Figure 7-16 from Alberts, Bruce, and Martin Raff. Essential Cell Biology.
New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

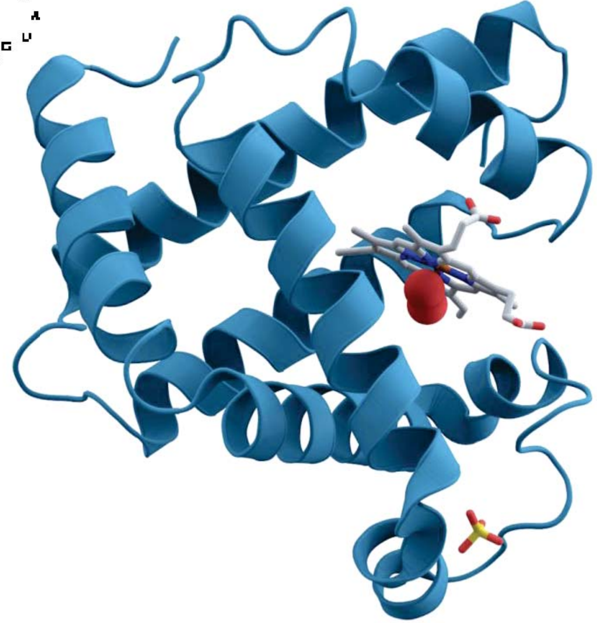
- Alternative splicing can yield different exon subsets for the same gene, and hence different protein products

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



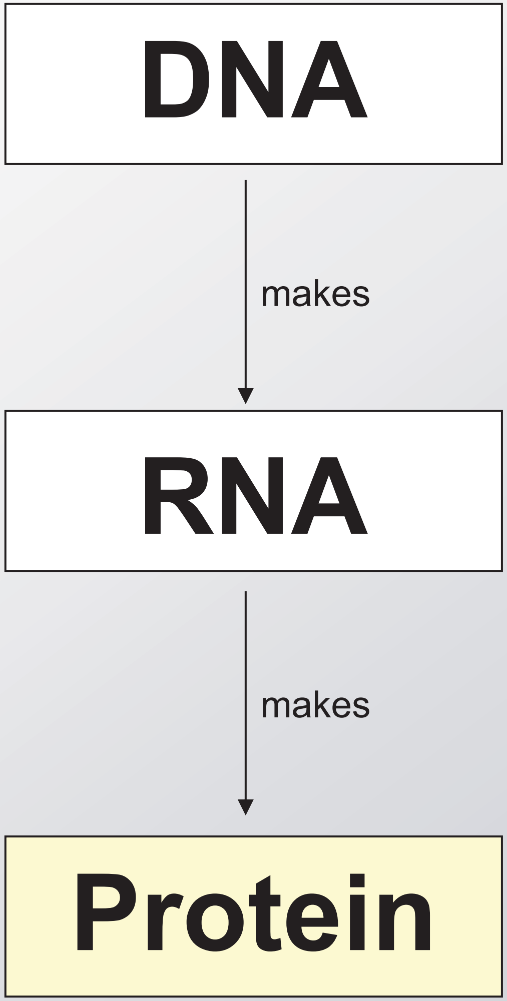


Courtesy of SStructView

Splicing machinery made of RNA

Image removed due to copyright restrictions.
Please see: Figure 7-16 from Alberts, Bruce, and Martin Raff. Essential Cell Biology.
New York, NY: Garland Publishing Inc., 1997. ISBN: 0815320450.

“Central dogma” of Molecular Biology



DNA

makes

RNA

makes

Protein

Proteins carry out the cell's chemistry

- More complex polymer
  - Nucleic Acids have 4 building blocks
  - Proteins have 20. Greater versatility
  - Each amino acid has specific properties

Sequence → Structure → Function

  - The amino acid sequence determines the three-dimensional fold of protein
  - The protein's function largely depends on the features of the 3D structure

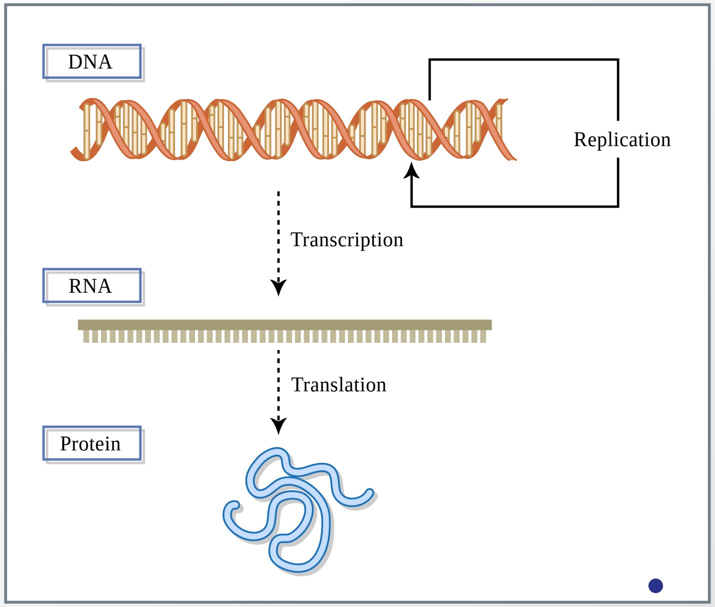


Image by MIT OpenCourseWare.

- Proteins play diverse roles
  - Catalysis, binding, cell structure, signaling, transport, metabolism

Protein structure

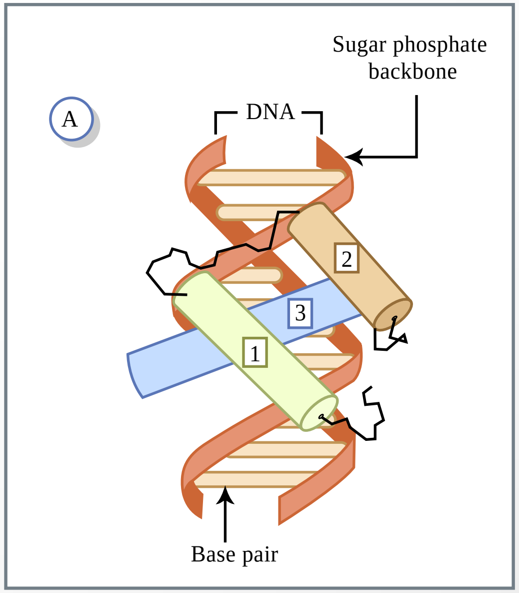


Image by MIT OpenCourseWare.

Helix-turn-helix

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

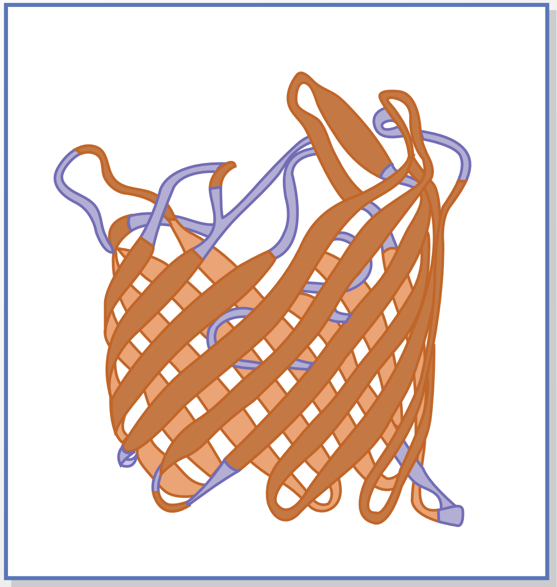


Image by MIT OpenCourseWare.

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.

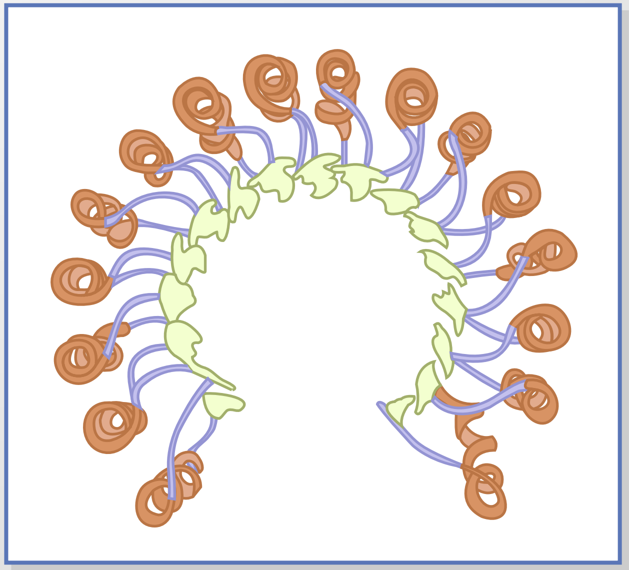


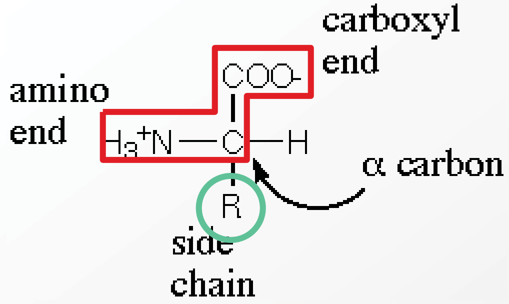
Image by MIT OpenCourseWare.

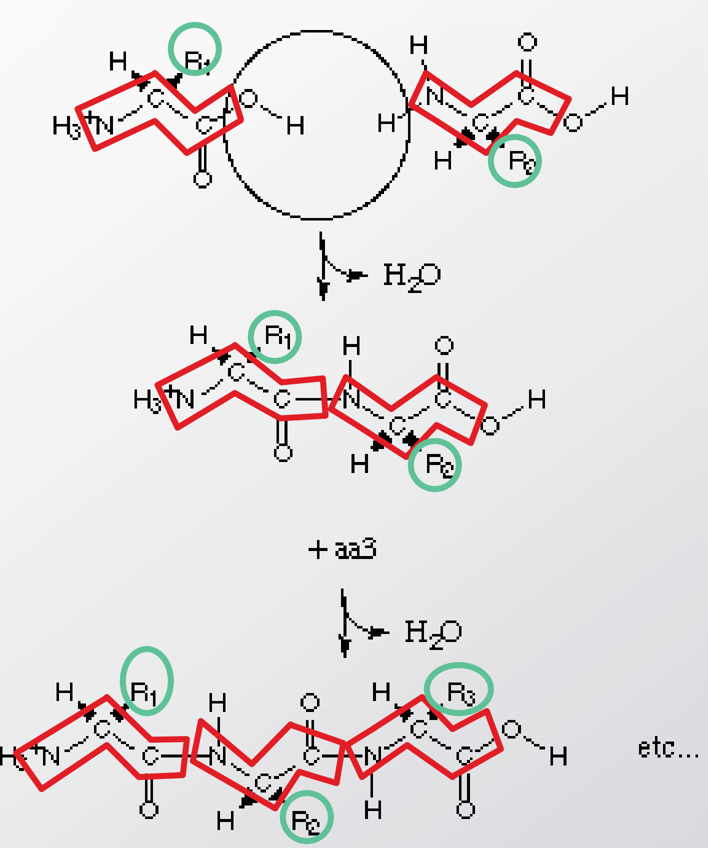
Alpha-beta horseshoe

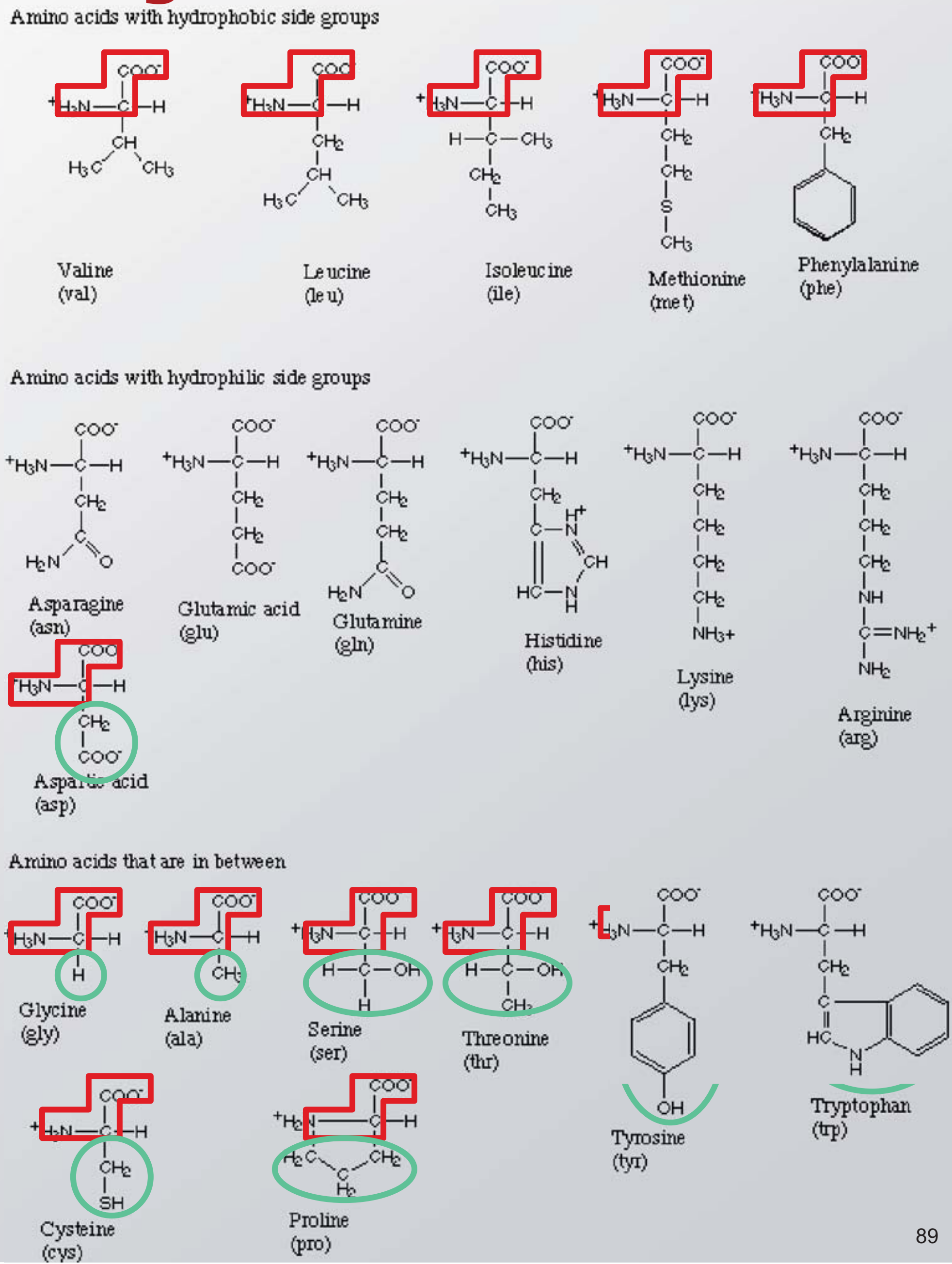
this placental ribonuclease inhibitor is a cytosolic protein that binds extremely strongly to any ribonuclease that may leak into the cytosol. 17-stranded parallel b sheet curved into an open horseshoe shape, with 16 a-helices packed against the outer surface. It doesn't form a barrel although it looks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

Protein building blocks

- Amino Acids







From RNA to protein: Translation

- Ribosome

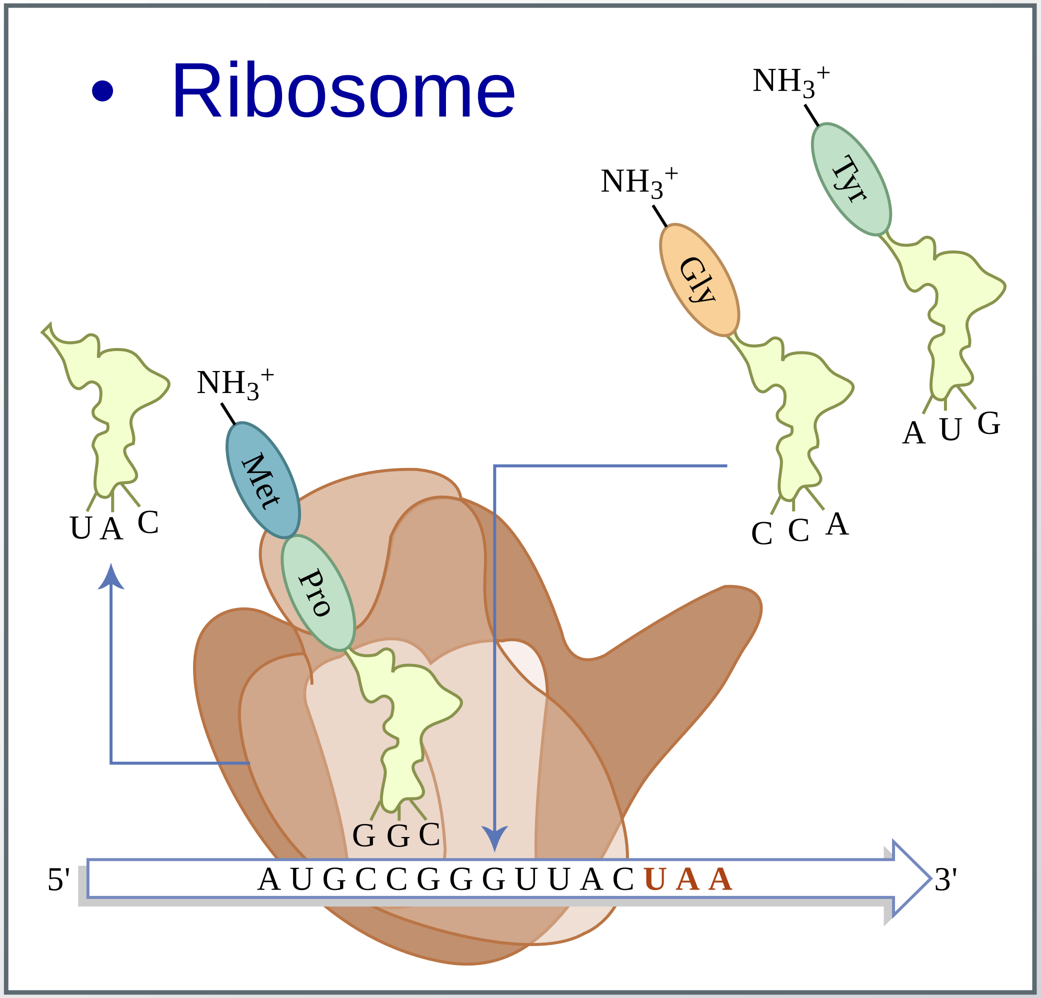
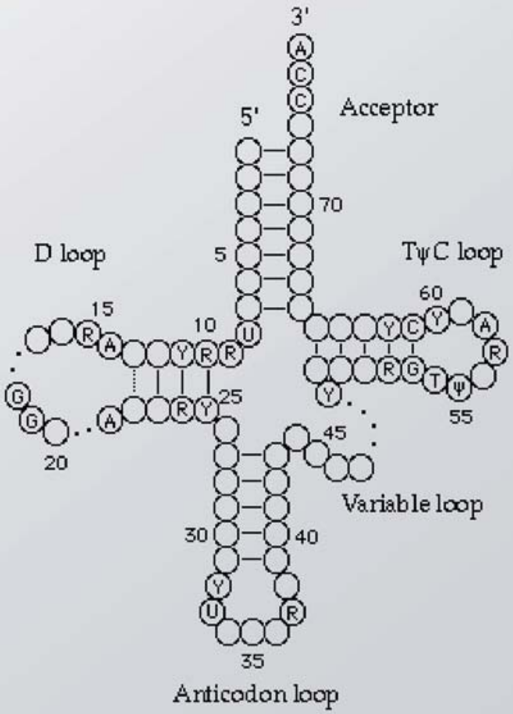
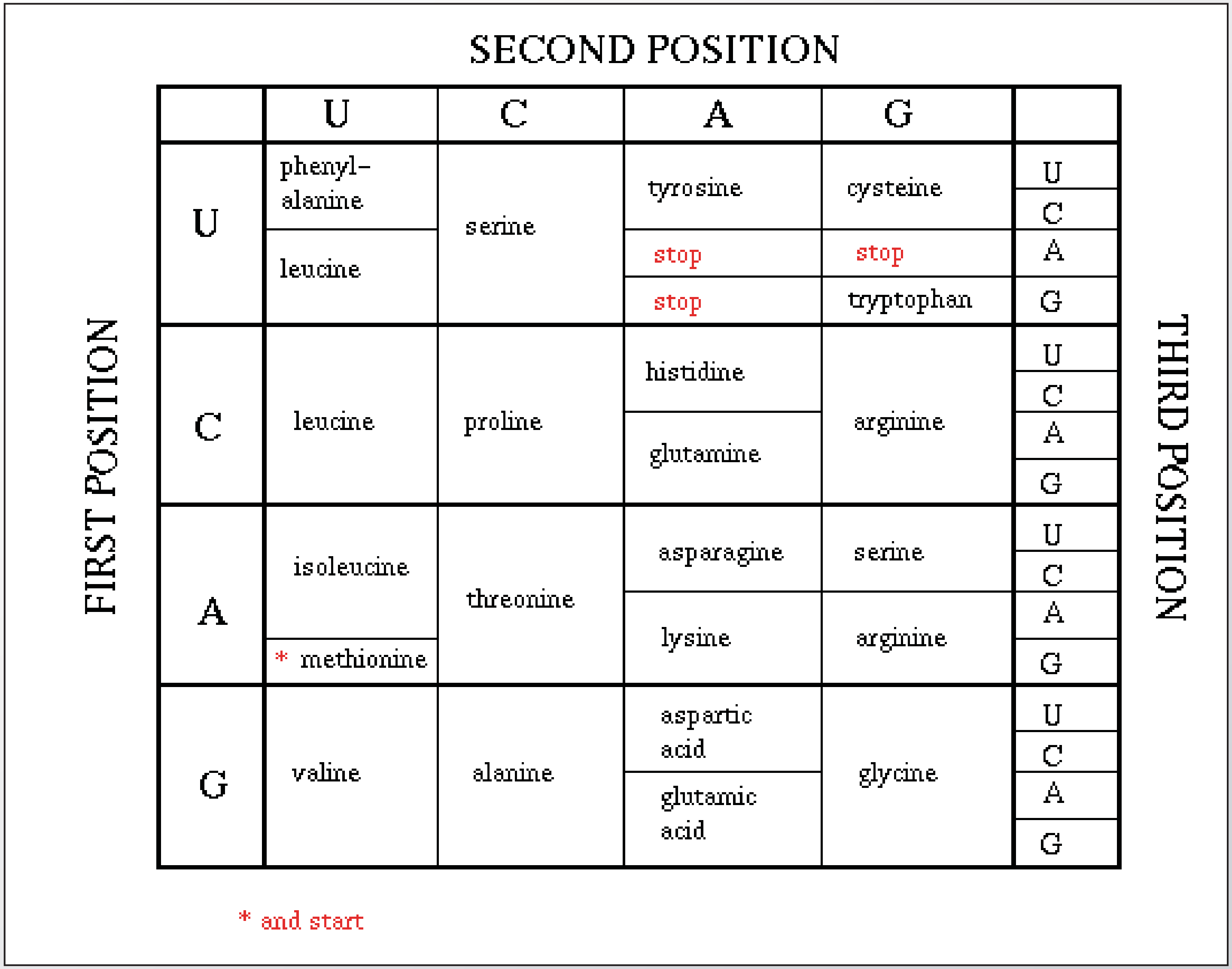


Image by MIT OpenCourseWare.

- tRNA



The Genetic Code



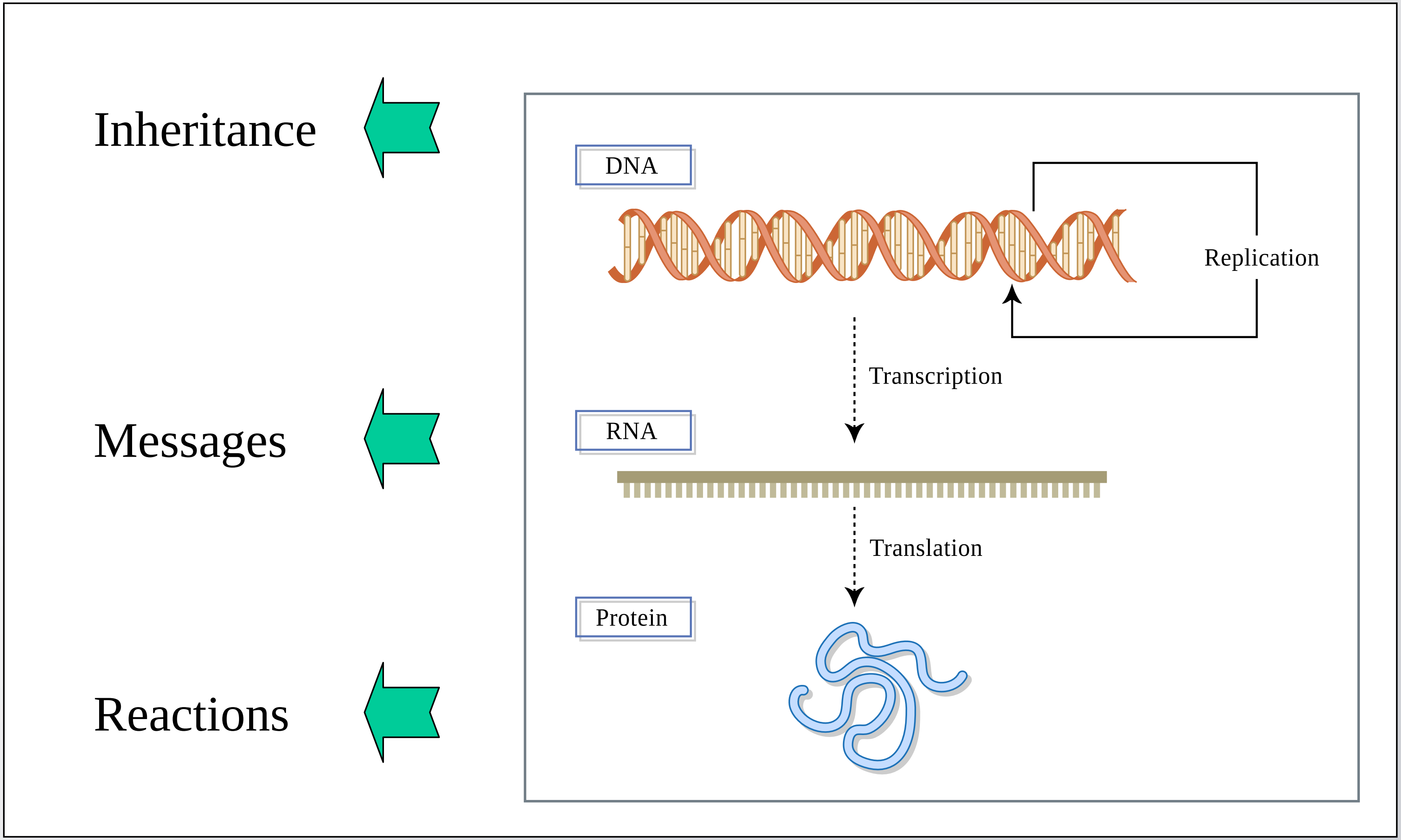
FIRST POSITION	SECOND POSITION				THIRD POSITION
	U	C	A	G	
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
			lysine	arginine	A
	* methionine				G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

* and start

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

Summary: The Central Dogma

DNA makes RNA makes Protein



Inheritance

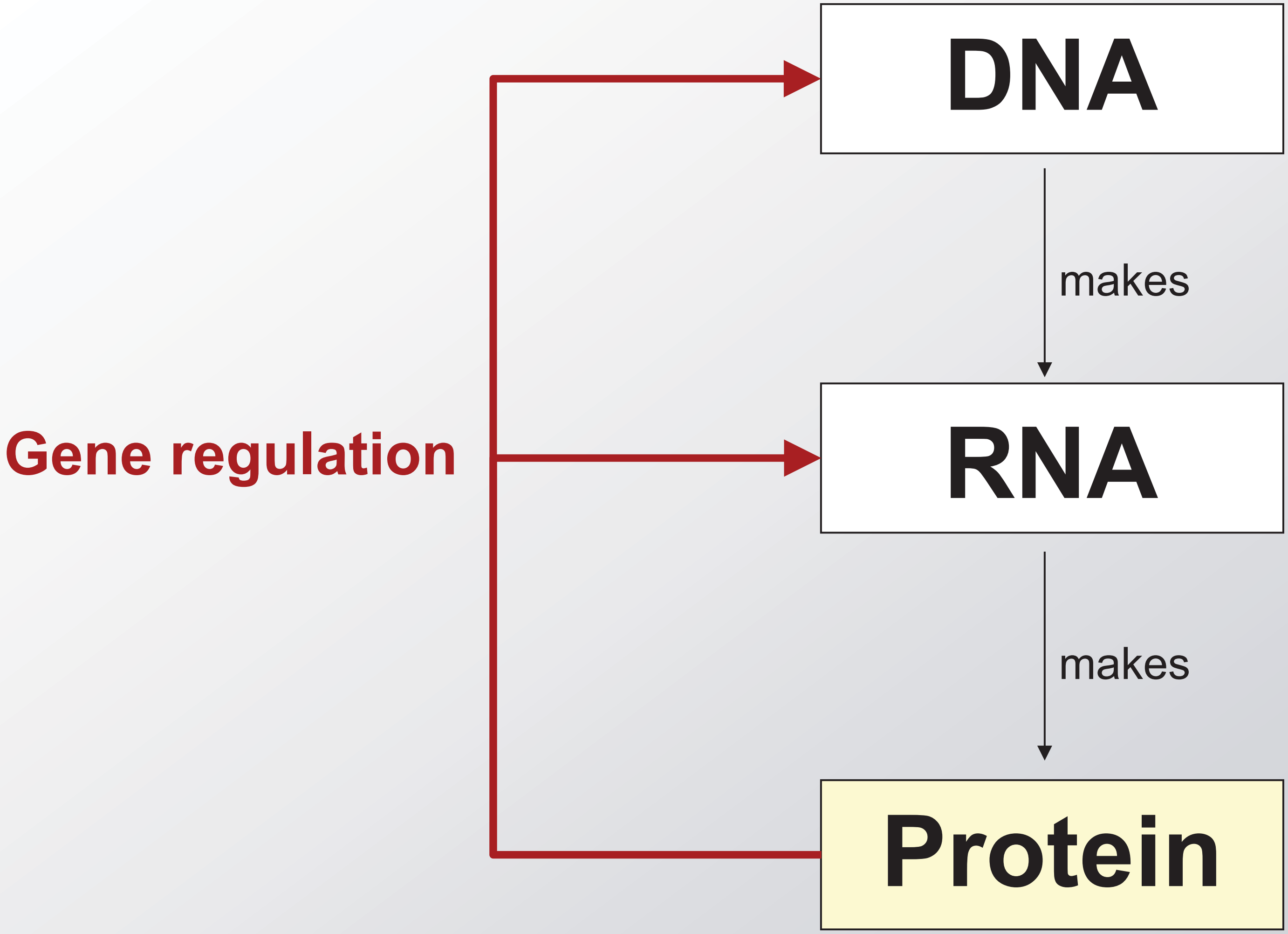
Messages

Reactions

Image by MIT OpenCourseWare.

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

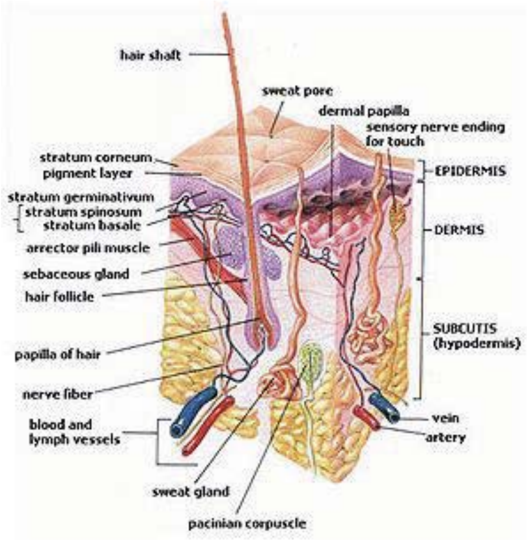


Image in the public domain.

Images of a heart, red blood cell, and a brain removed due to copyright restrictions.

Image Source wikipedia

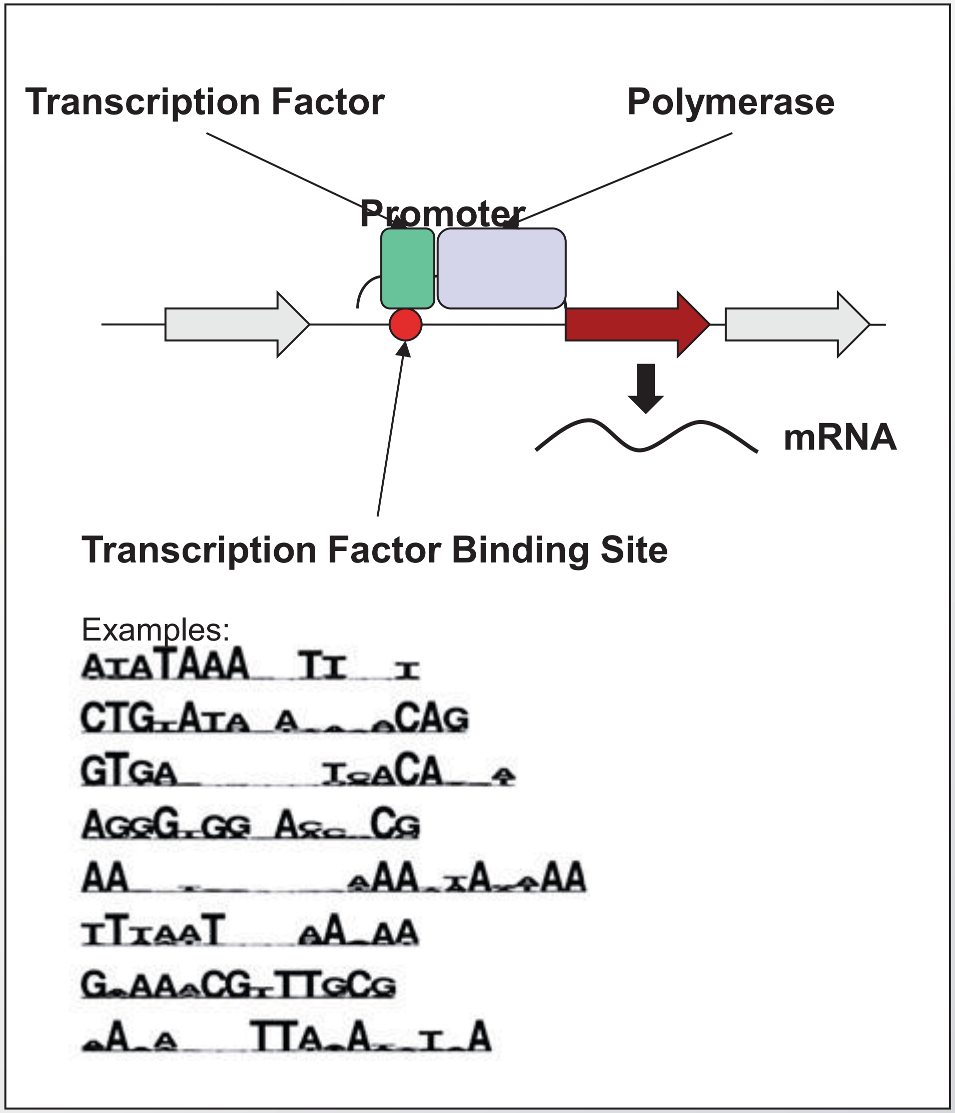
Eukaryotic Gene Regulation

Cartoon depicting eukaryotic gene regulation removed due to copyright restrictions.

Diverse roles for regulatory non-coding RNAs

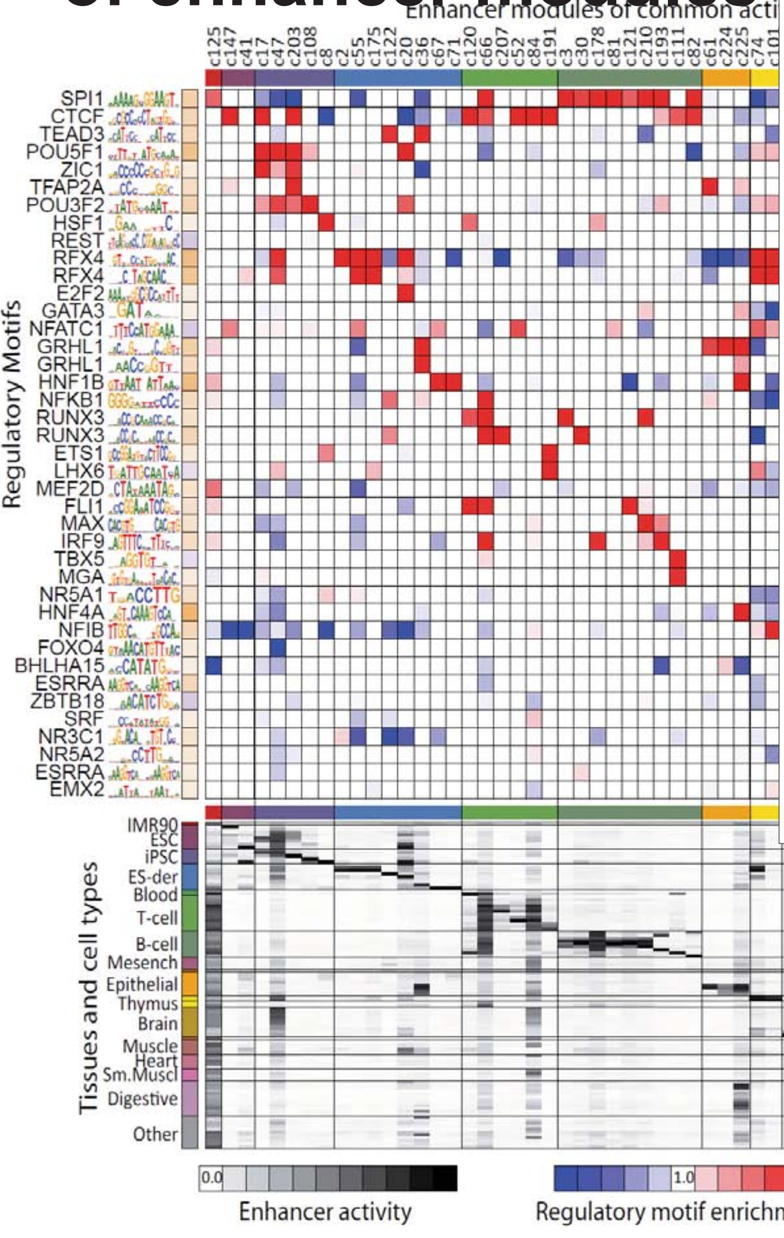
- **Small RNA pathways (18-21 nt)**
  - microRNAs:
    - Repress genes by targeting their 3' UTRs by complementarity
    - Double-stranded RNA is then recognized and degraded
    - Recently found to also target promoter regions in rare cases
  - piwiRNAs
    - Target and repress transposable elements in germline
  - snoRNAs
  - 21U-RNAs
- **Long non-coding RNAs (1000s nt, many exons)**
  - Scaffolds for protein/TF binding
  - Scaffolds for 3D structure of RNA

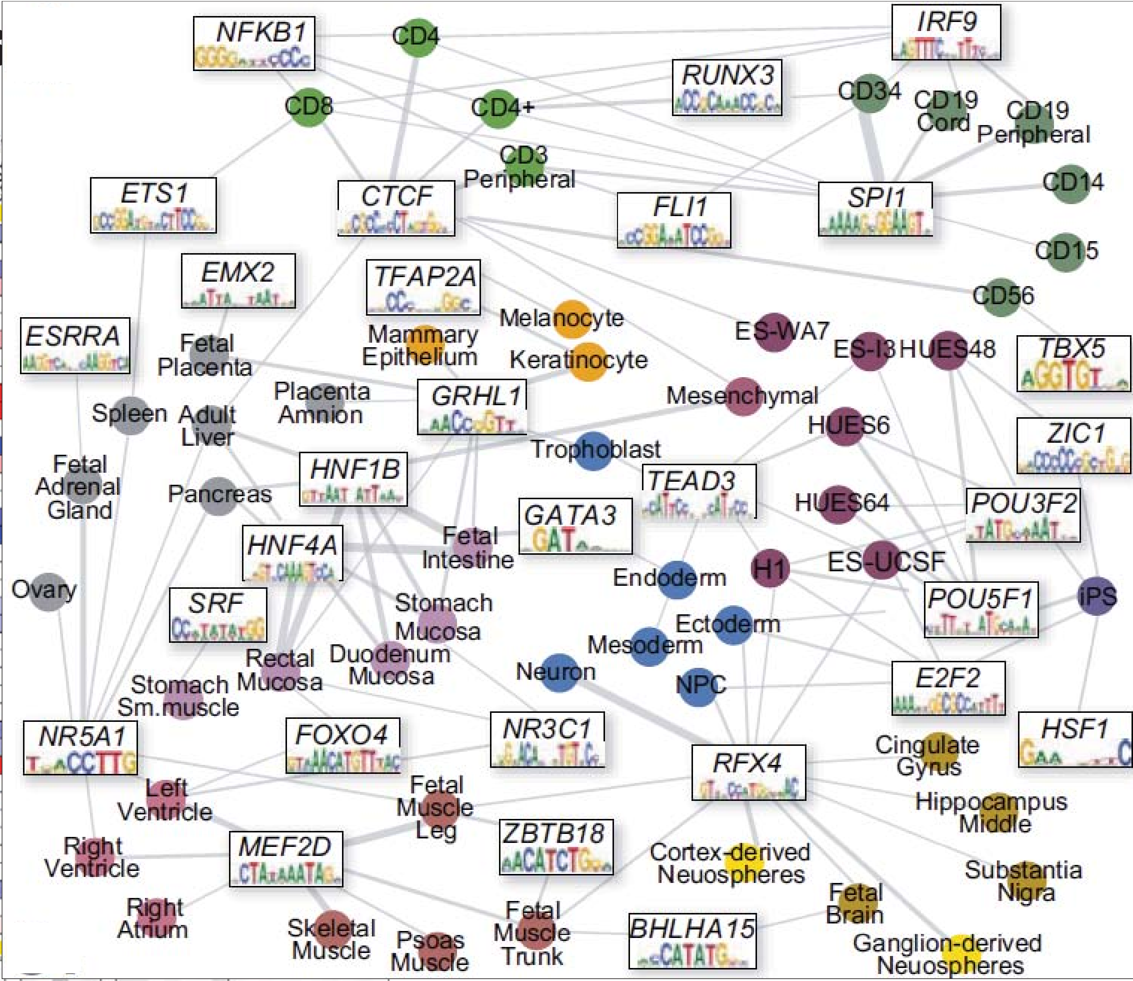
Regulation of Gene Expression



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

Predicted motif drivers of enhancer modules



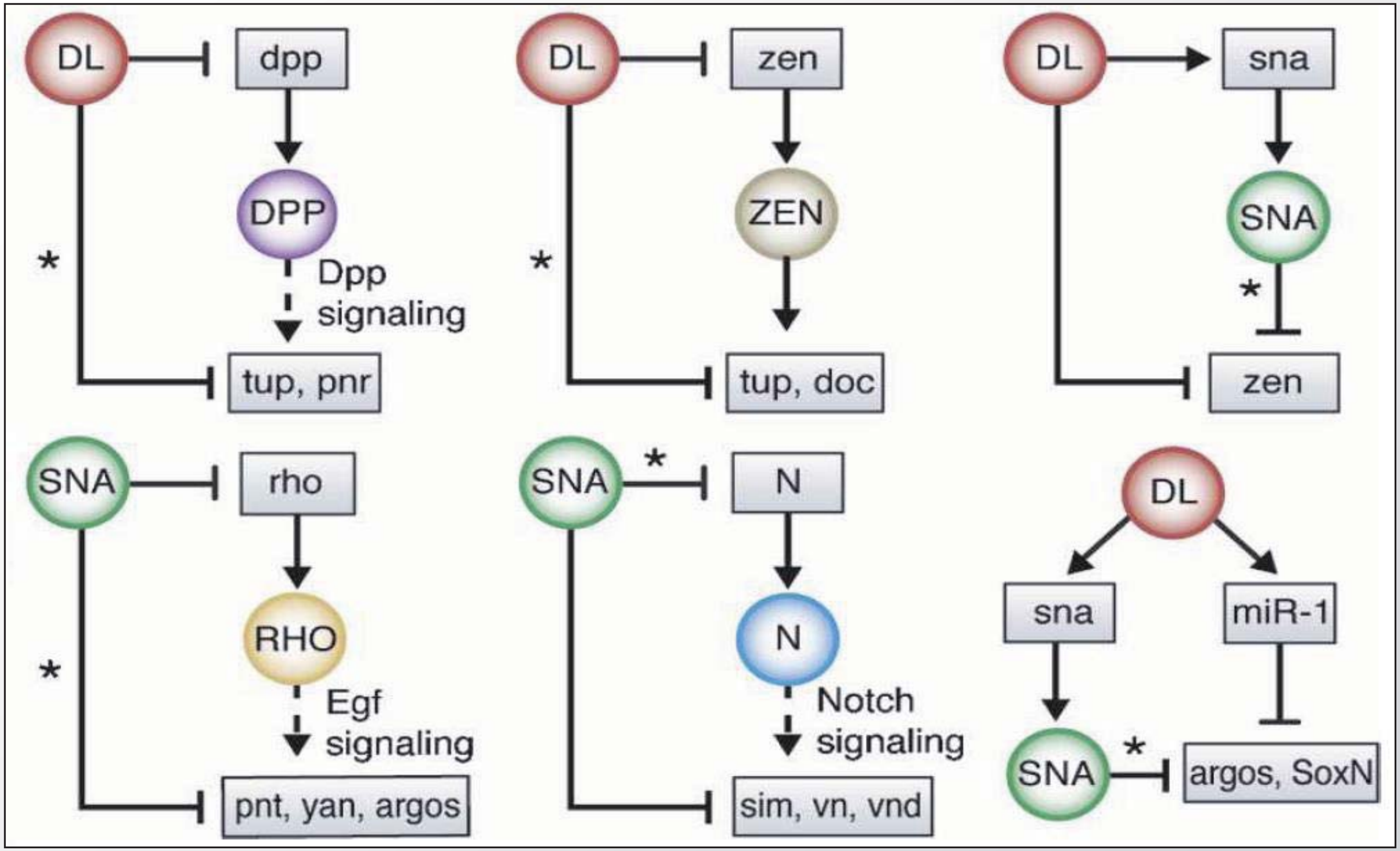


- **Activator and repressor motifs consistent with tissues**

Courtesy of Macmillan Publishers Limited. Used with permission.

Pouya Kheradpour

Network components reveal functional modules



Source: Zeitlinger, Julia et al. "Whole-genome ChIP–chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo." Genes & Development 21, no. 4 (2007): 385-390.

- 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

Figure 1: selection of activator and repressor motifs removed due to copyright restrictions. Source: Kheradpour, Pouya et al. "Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay." Genome Research 23, no. 5 (2013): 800-811.

54000+ measurements (x2 cells, 2x repl)

Kheradpour *et al* Genome Research 2013

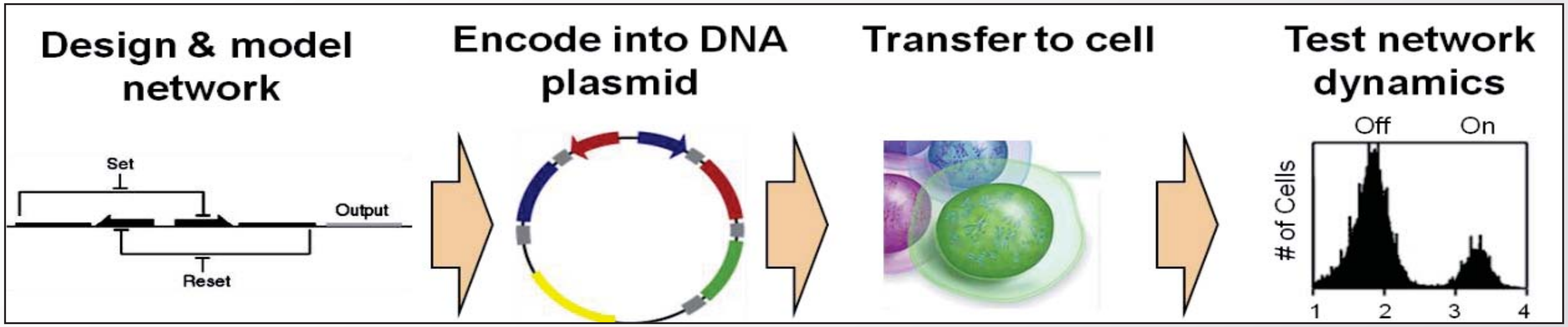
Emerging properties of regulatory networks

Figures removed due to copyright restrictions.

- 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

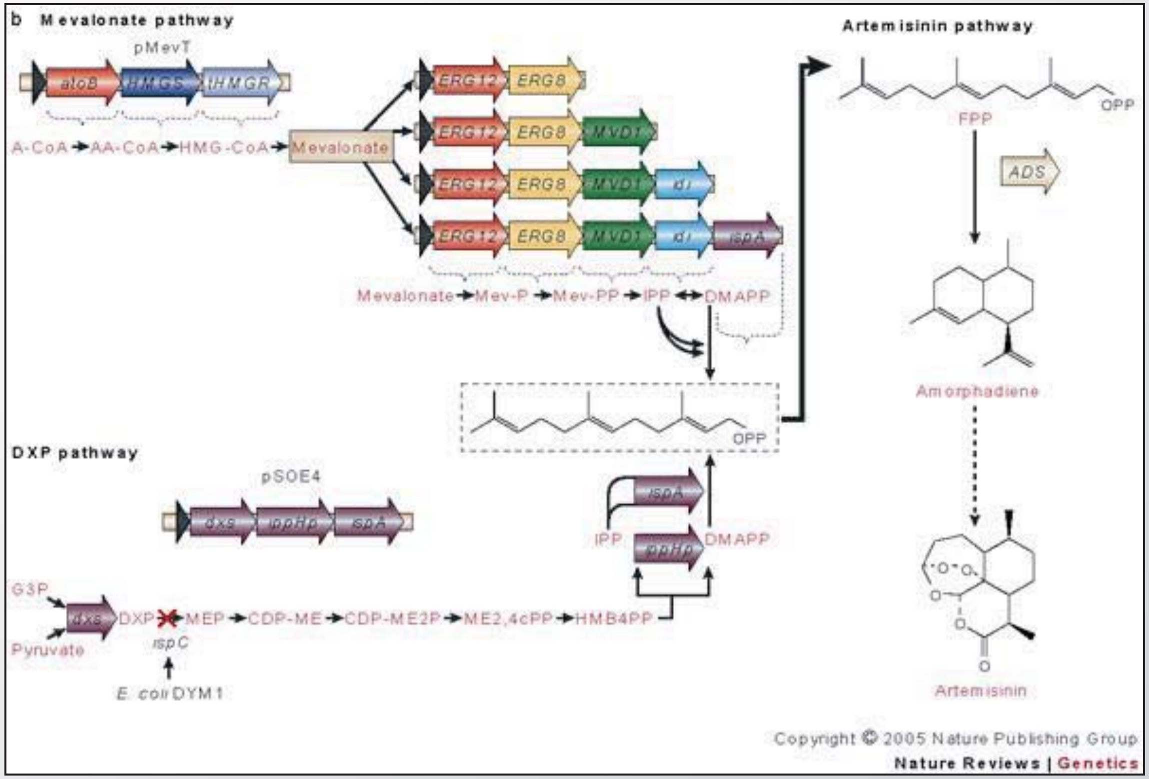
From Systems Biology to Synthetic Biology

Synthetic Regulatory Networks



Design & model network → Encode into DNA plasmid → Transfer to cell → Test network dynamics

Synthetic Metabolic Pathways



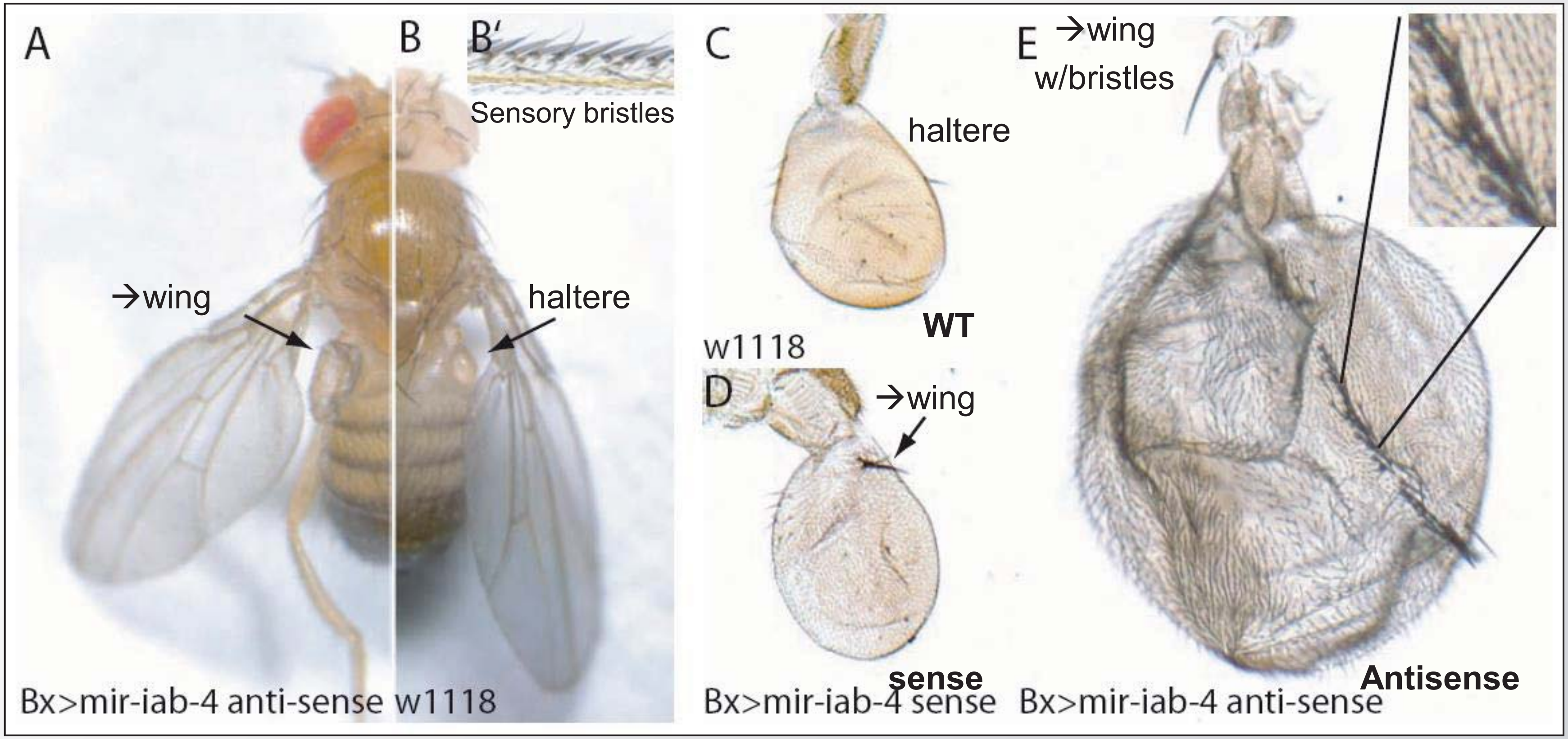
Courtesy of Macmillan Publishers Limited. Used with permission. Source: Benner, Steven A. and A. Michael Sismour. "Synthetic biology." Nature Reviews Genetics 6, no. 7 (2005): 533-543.

Jim Collins

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

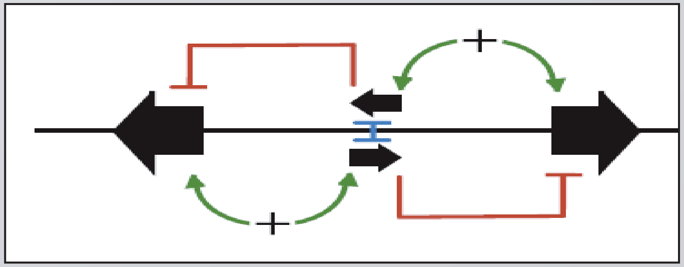
Jay Keasling

Over-express a single microRNA leads to new wing



Note: C,D,E same magnification

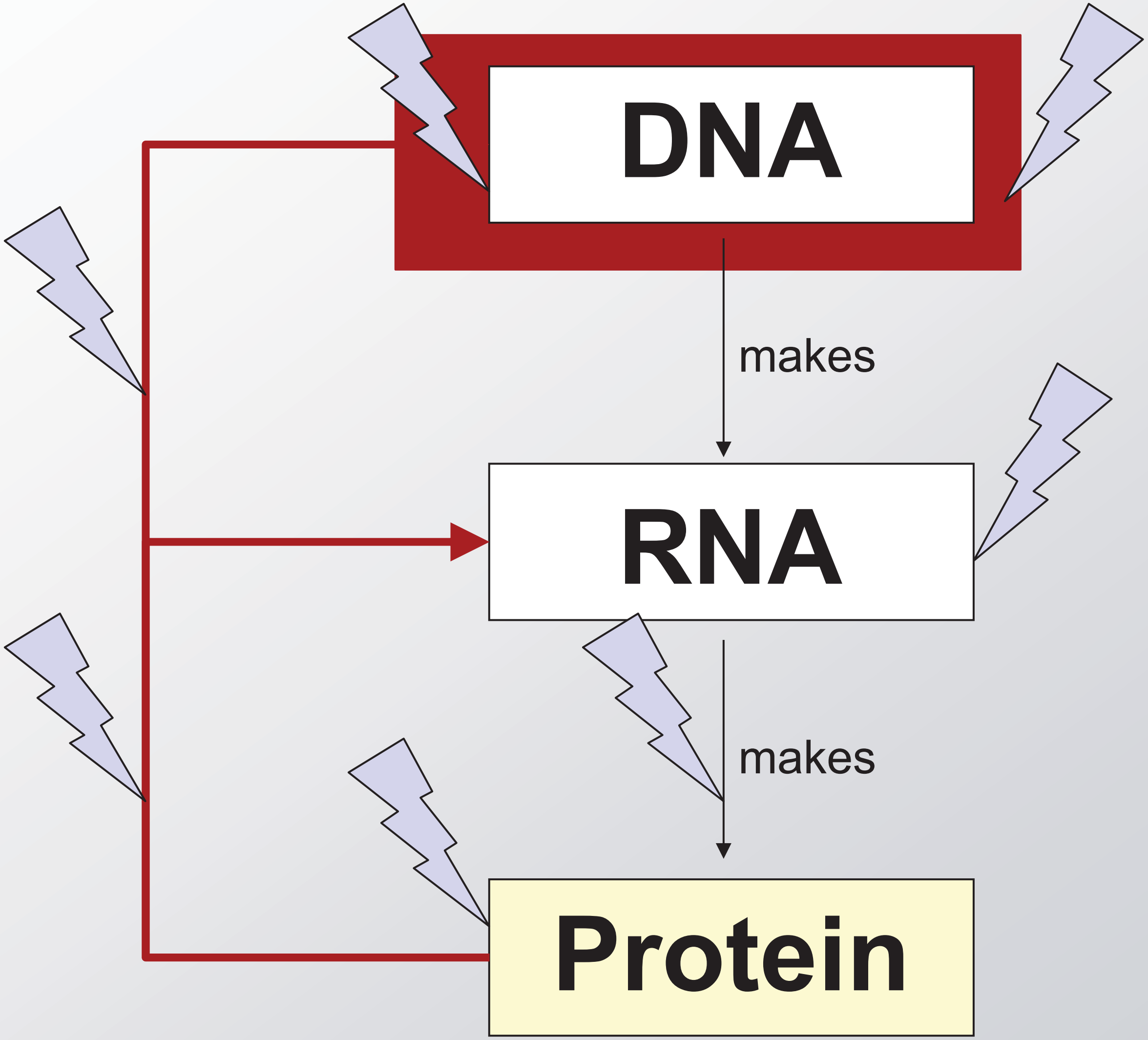
- Discovery of sense/ant_i-sense miRNAs
- Regulatory switch selects between two developmental programs
- By over-expressing one strand (miRNAas) the balance is tilted
- Wing program launched vs. haltere



Stark *et al*, Genes&Development 2007

Brief intro to Human Genetics

The role of genetic alterations



DNA

makes

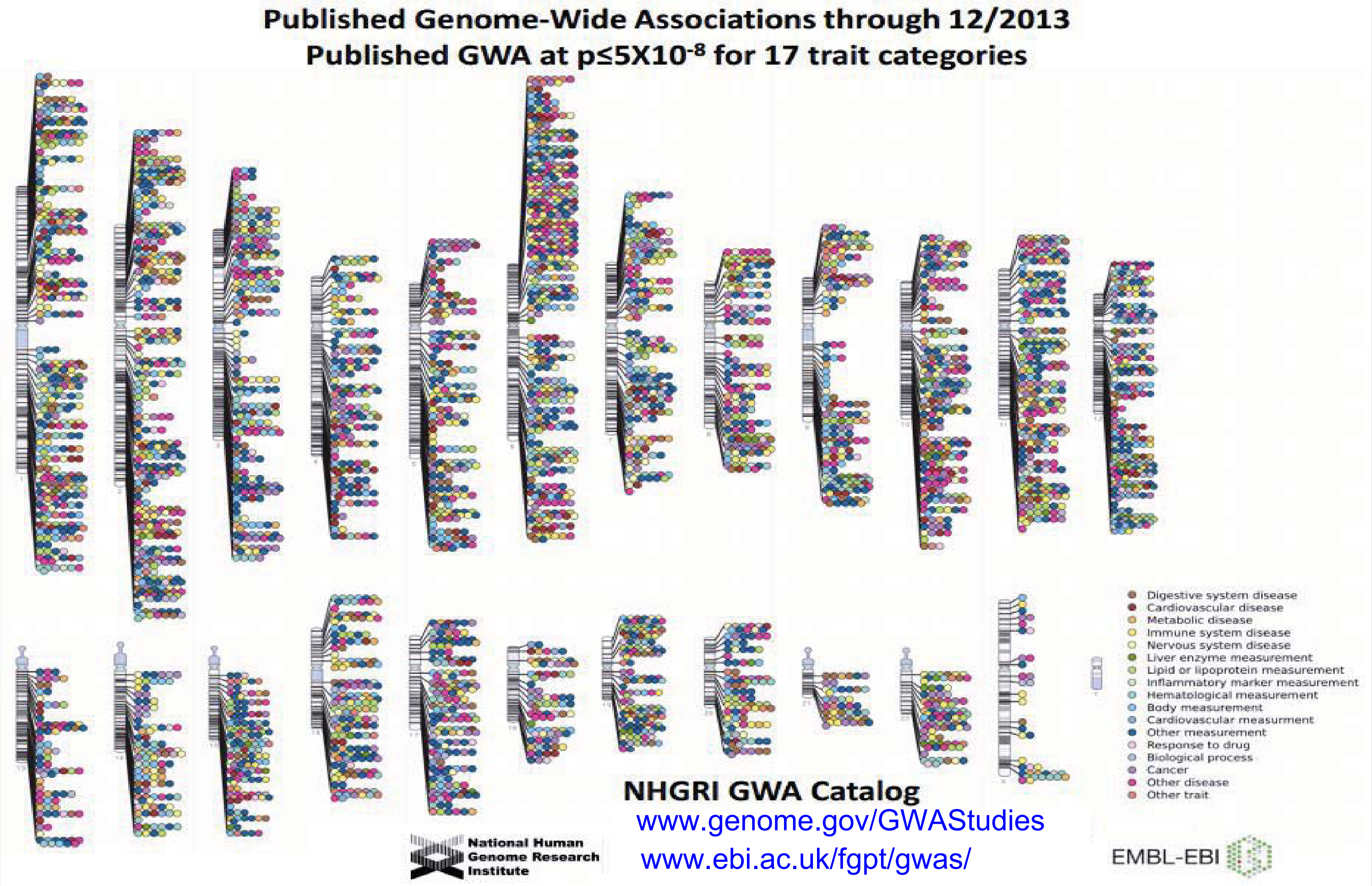
RNA

makes

Protein

Brief intro to human genetics

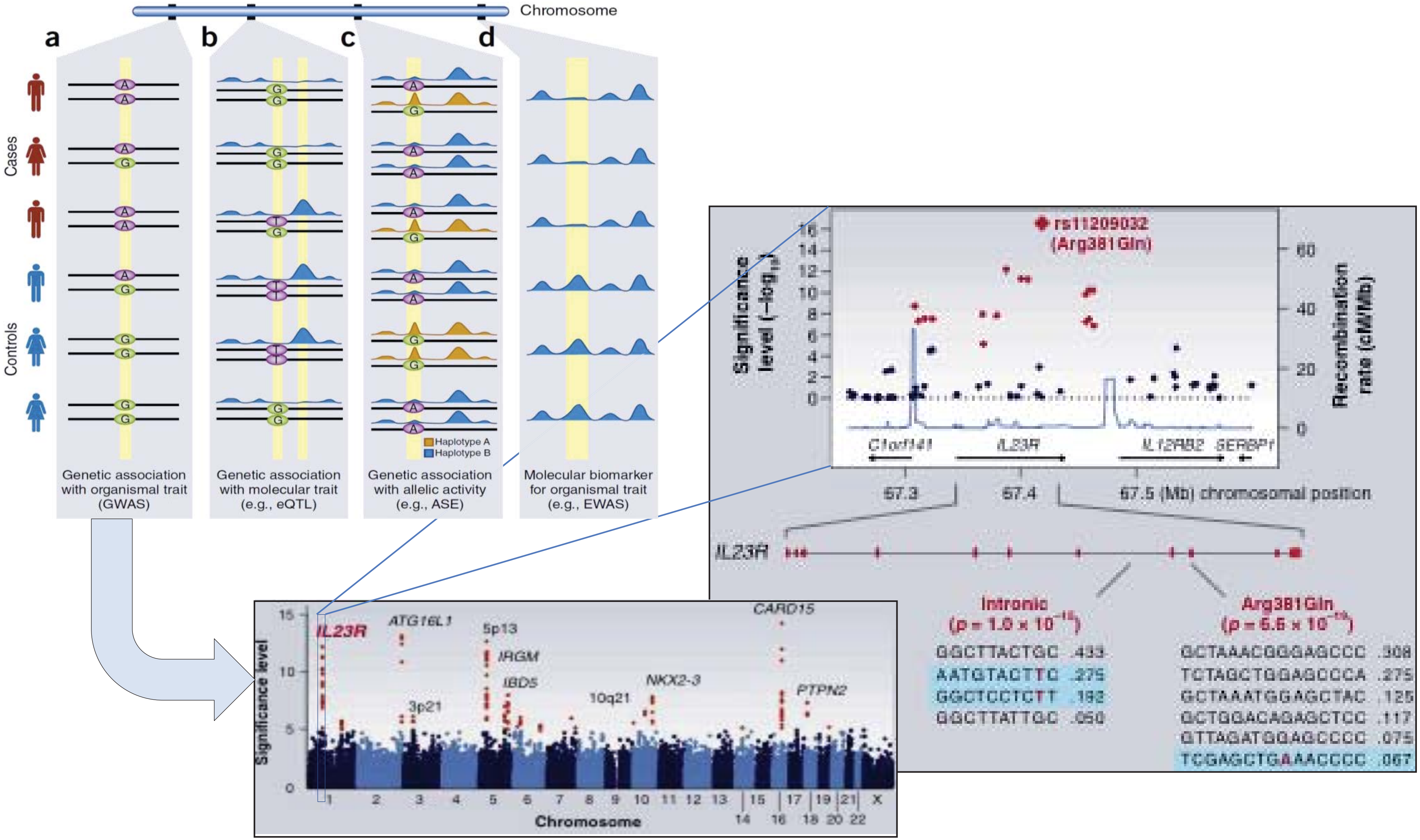
- **Human genome**: 3.2B letters, 2 copies, 23 chromosomes, 20k genes, ~3M common SNPs, ~500k haplotype blocks



www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

Figure in the public domain. Created by Darryl Leja and Teri Manolio, NHGRI; Tony Burdett, Dani Welter, and Helen Parkinson, EBI.

The power and challenge of disease-association studies

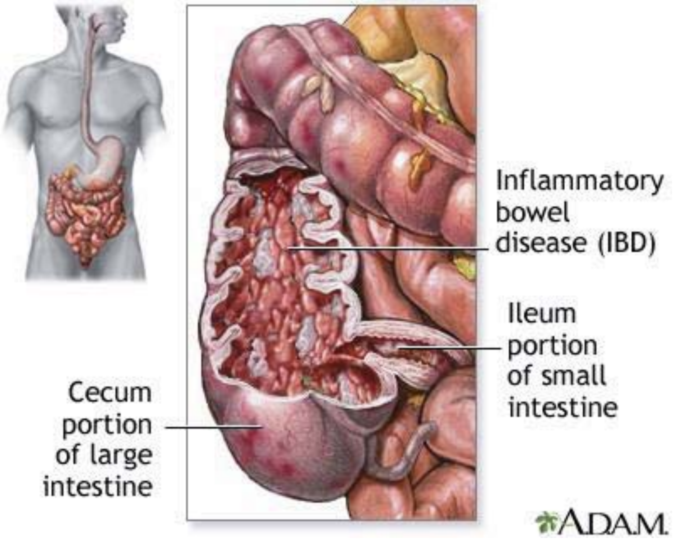


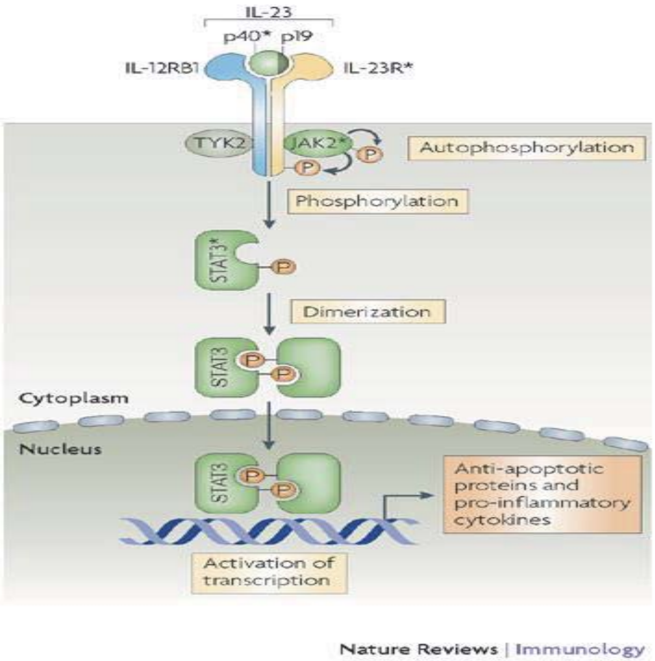
Slide credit: Luke Ward, Mark Daly

- **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


Inflammatory bowel disease (IBD)
Ileum portion of small intestine
Cecum portion of large intestine
ADAM



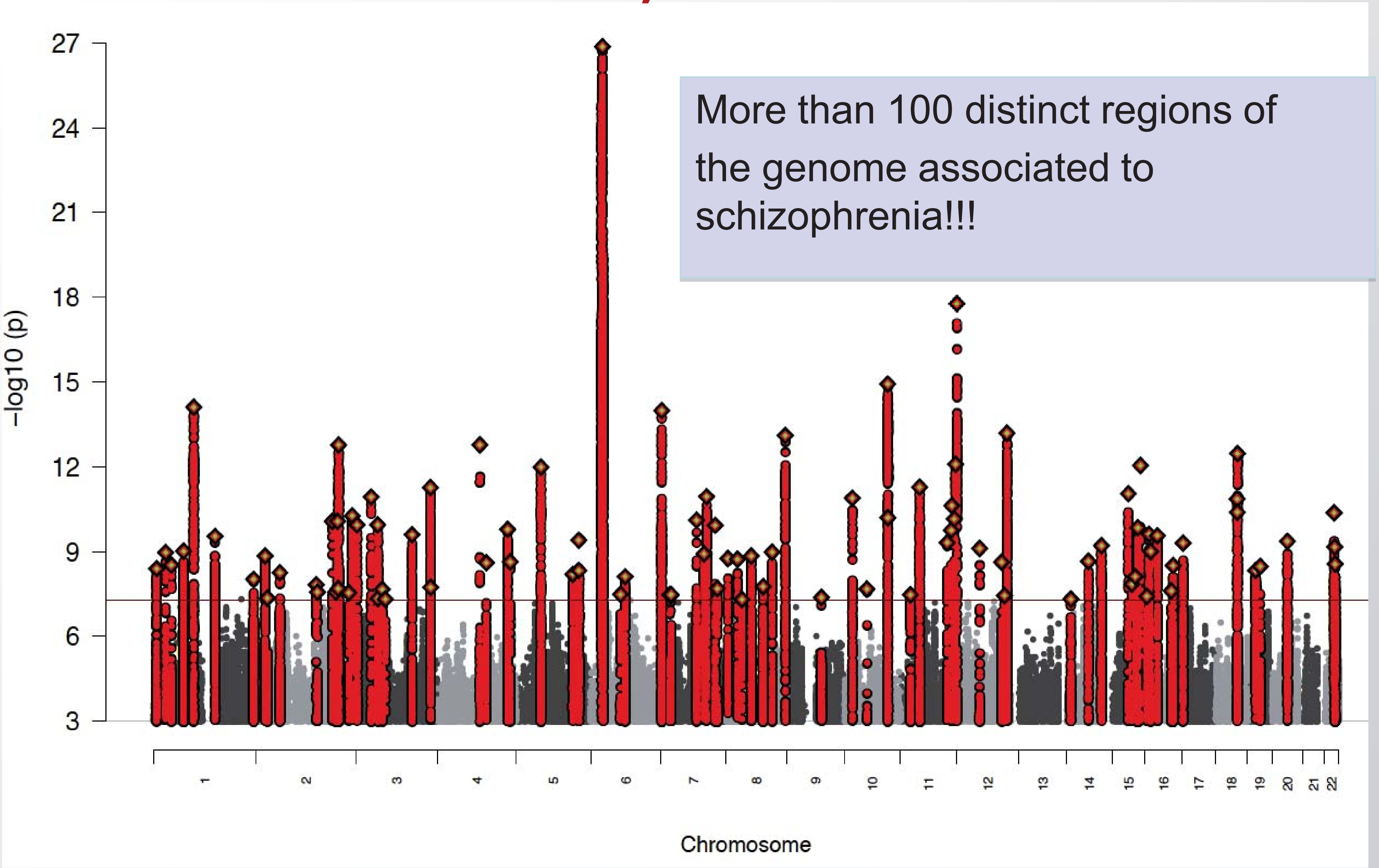
Courtesy of Macmillan Publishers Limited. Used with permission. Source: Cho, Judy H. "The genetics and immunopathogenesis of inflammatory bowel disease." Nature Reviews Immunology 8, no. 6 (2008): 458-466.

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



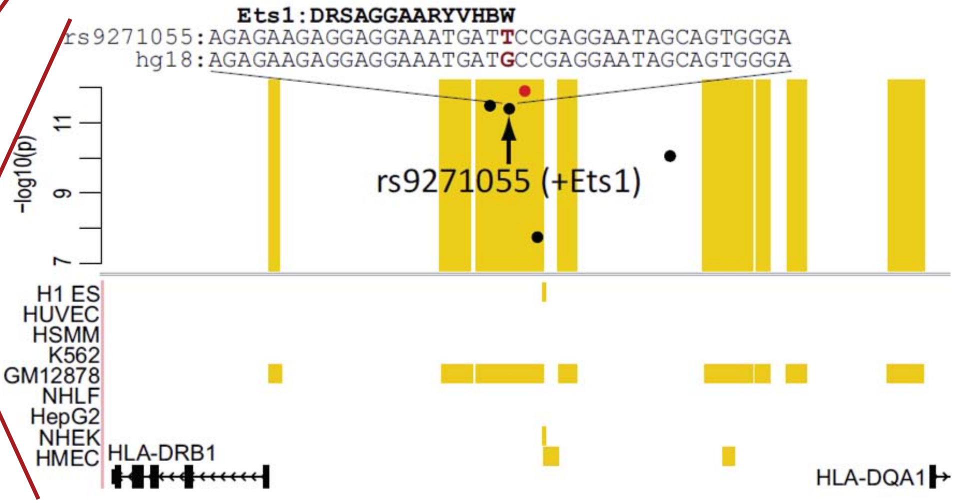
More than 100 distinct regions of the genome associated to schizophrenia!!!

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Ripke, Stephan et al. "Biological insights from 108 schizophrenia-associated genetic loci." Nature 511, no. 7510 (2014): 421.

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 Sc	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

Figures removed due to copyright restrictions.

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

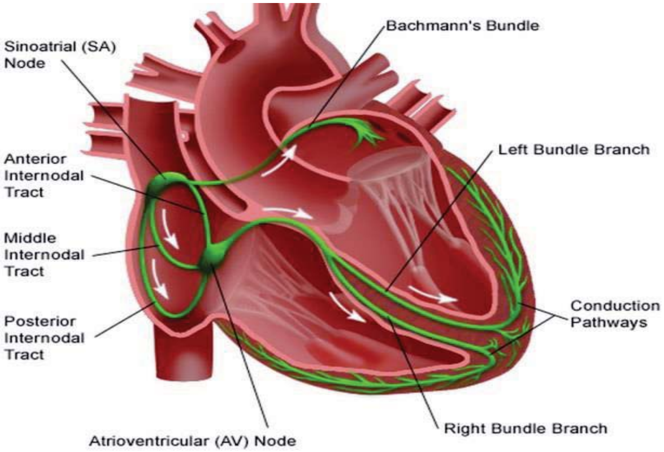
Chromatin state annotations across 127 epigenomes

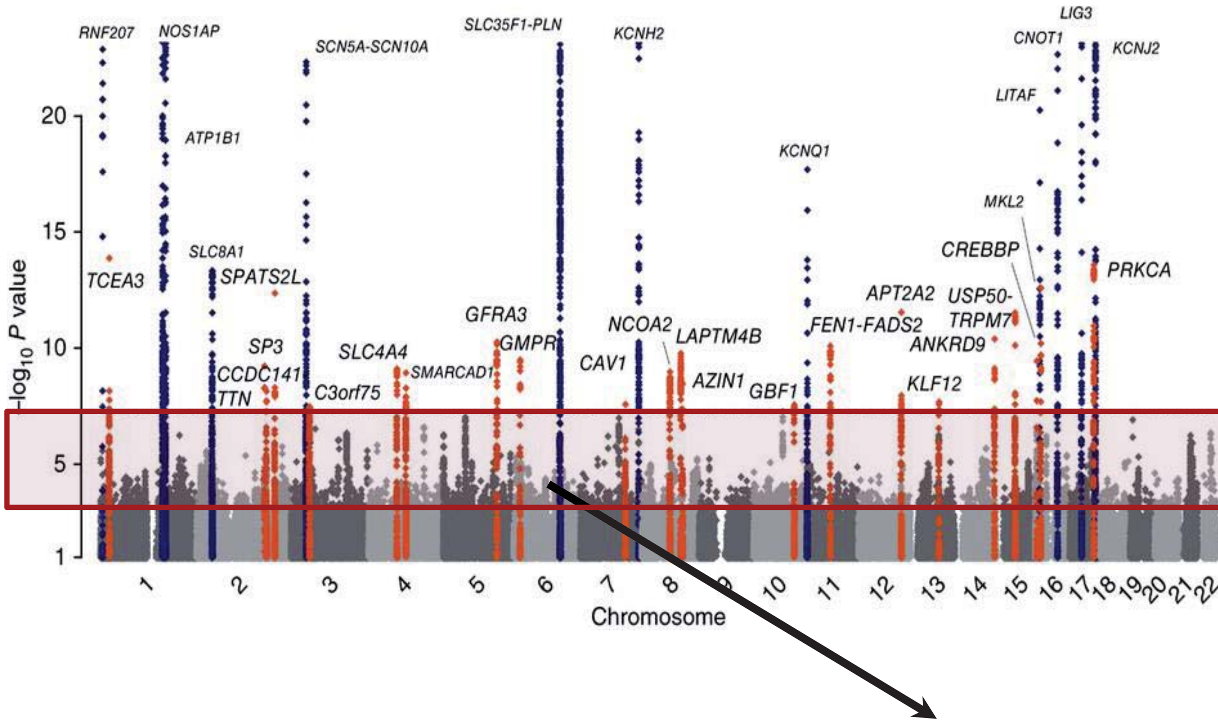
Figures removed due to copyright restrictions.

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

Anshul Kundaje

Characterizing sub-threshold variants in heart arrhythmia

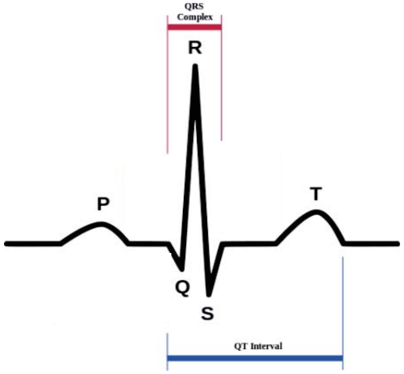




Focus on sub-threshold variants

(e.g. rs1743292 P=$10^{-4.2}$)

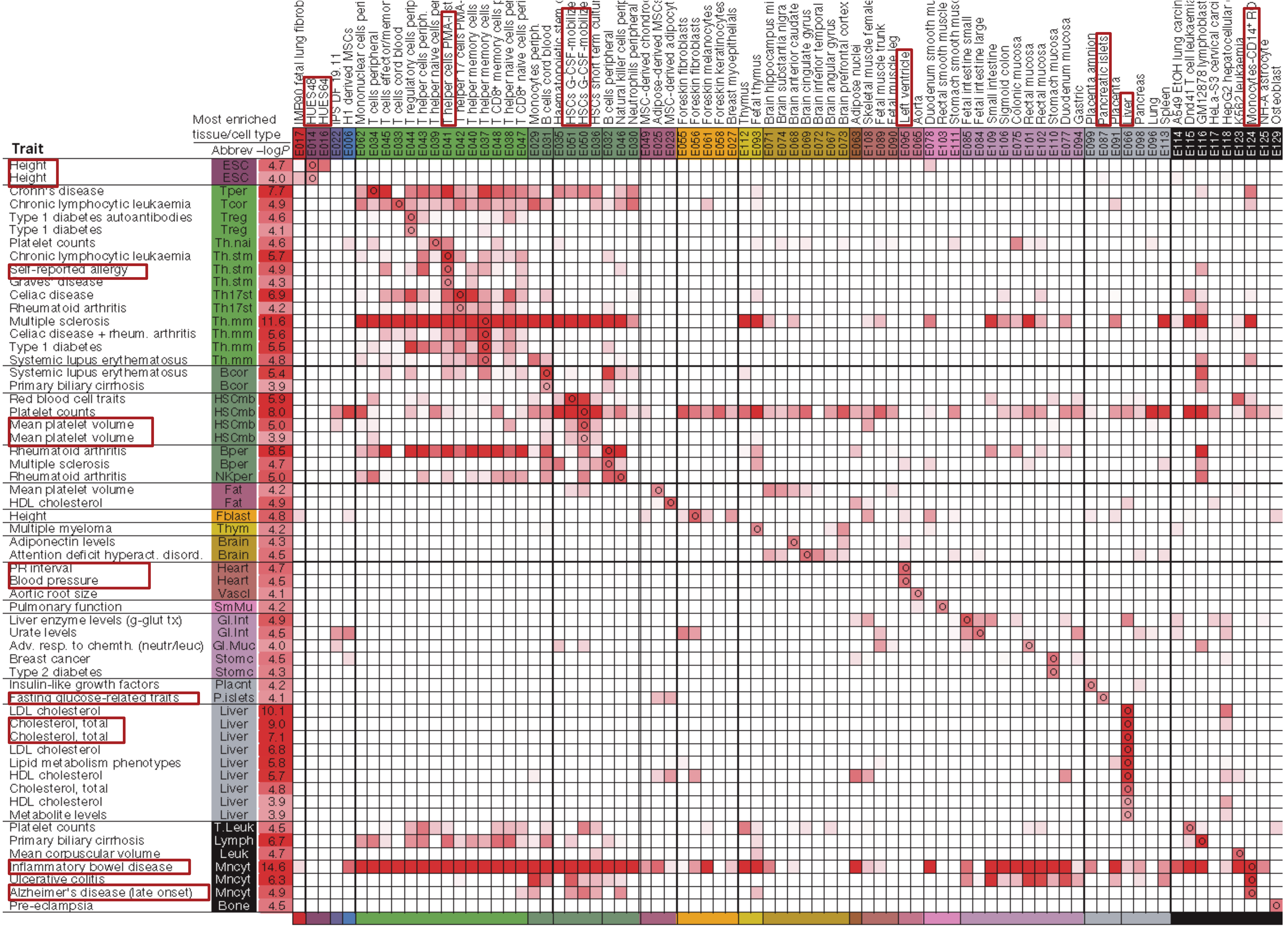
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Arking, Dan E. et al. "Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization." Nature Genetics 46, no. 8 (2014): 826-836.



Trait: QRS/QT interval

(1) Large cohorts, (2) many known hits

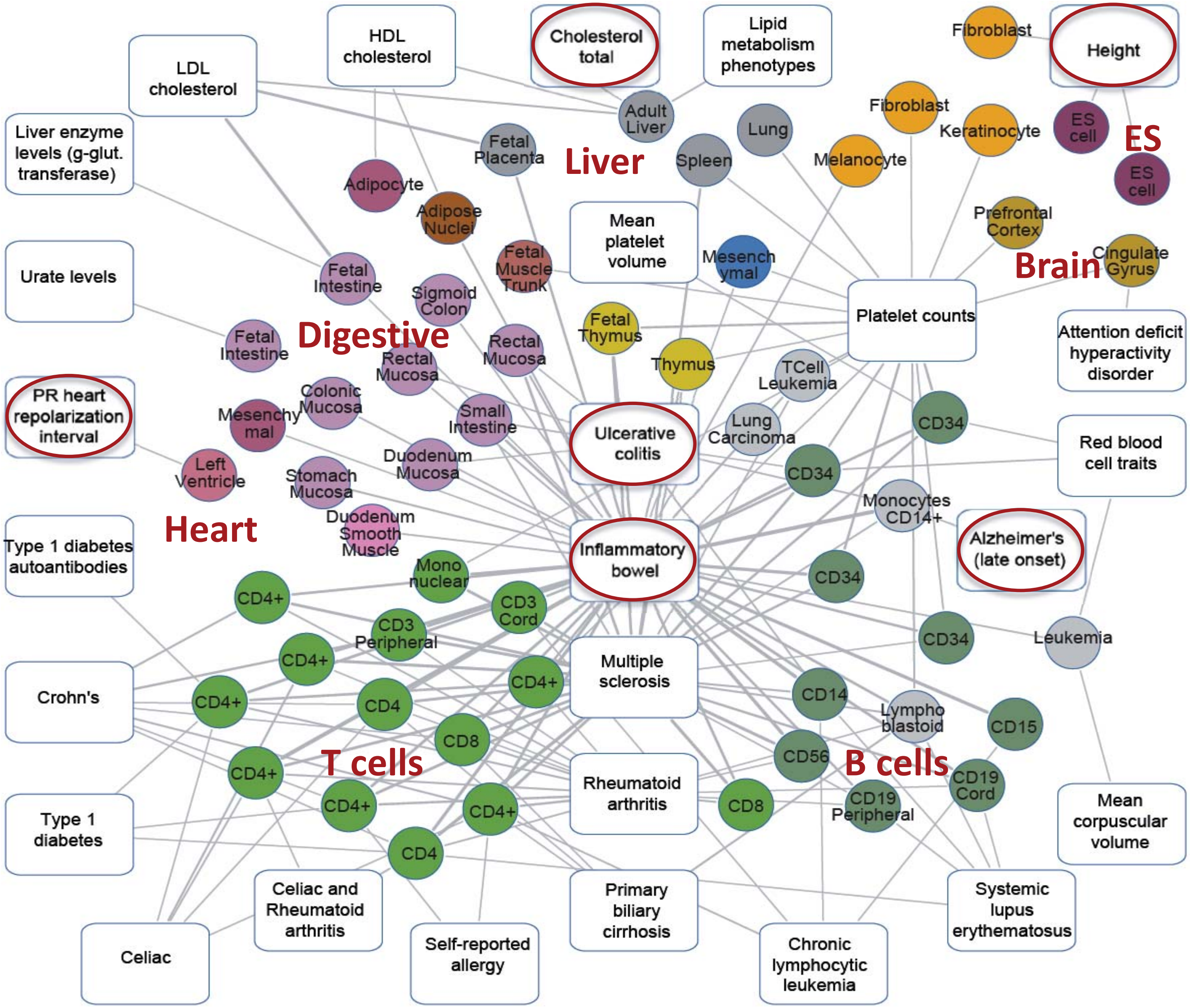
(3) well-characterized tissue drivers



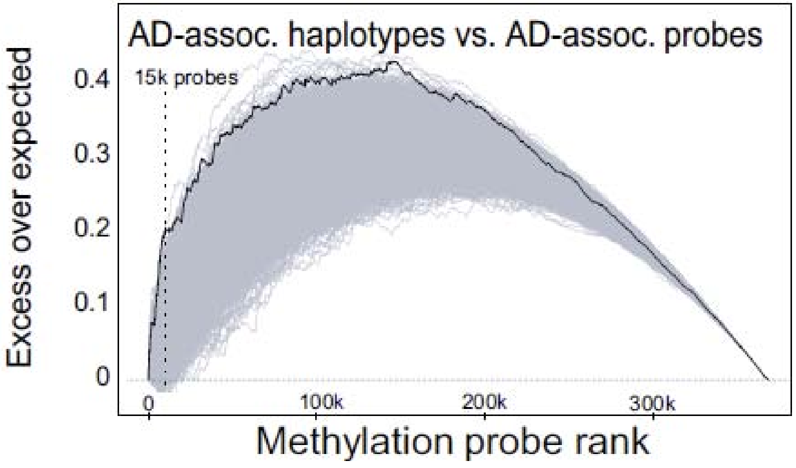
Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

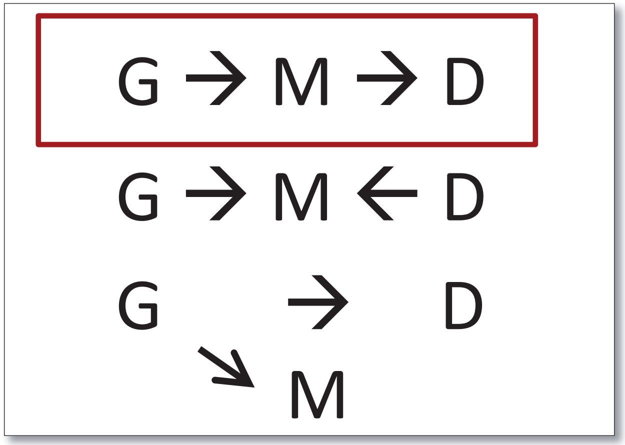
Linking traits to their relevant cell/tissue types



Methylation differences a causal component of AD



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

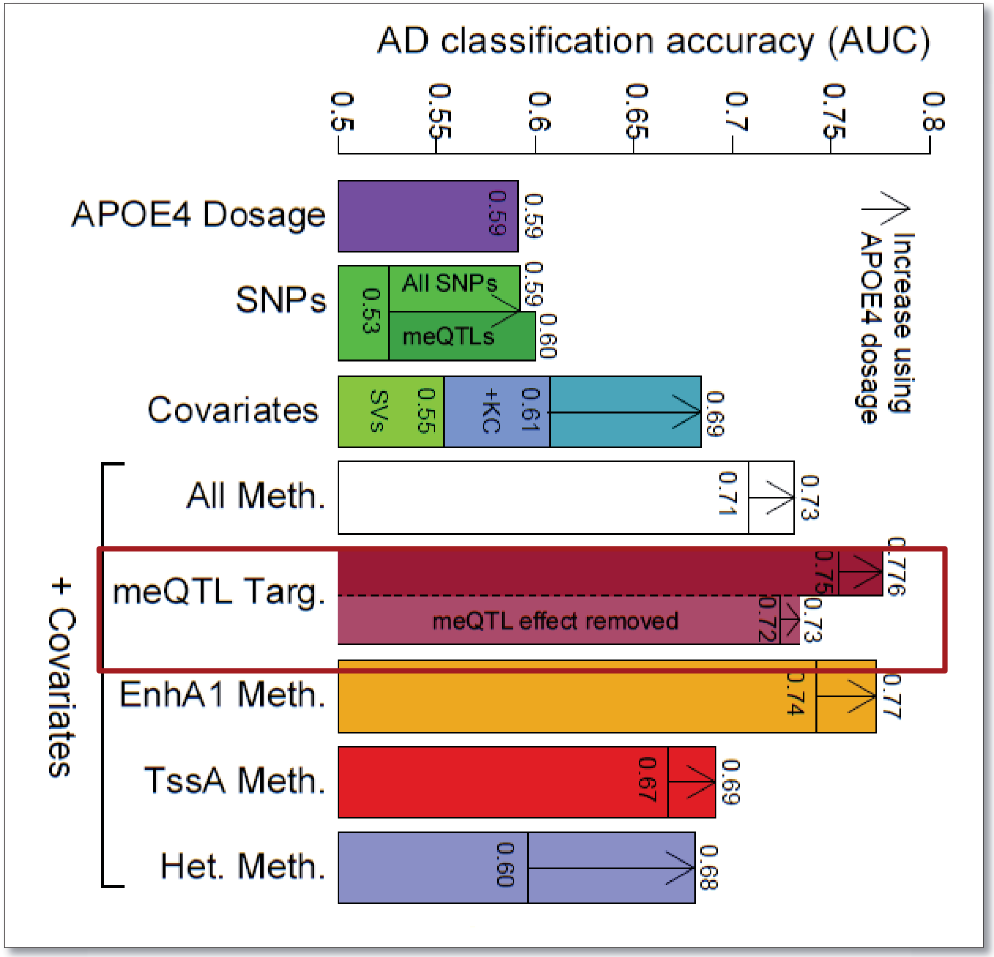


G → M → D

G → M ← D

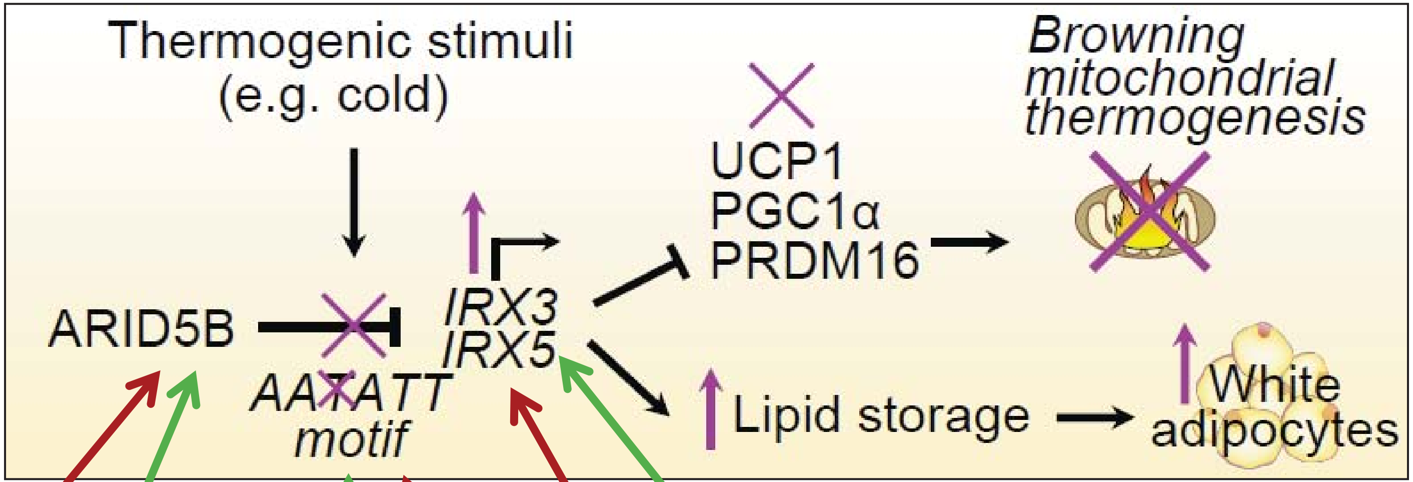
G → D
↘ M

Set-wise causality testing



AD predictive power reduced after removing meQTL effect

Uncovering the molecular basis of top obesity gene



Lean

Obese

ARID5B KD
(obesity)

ARID5B OE
(anti-obesity)

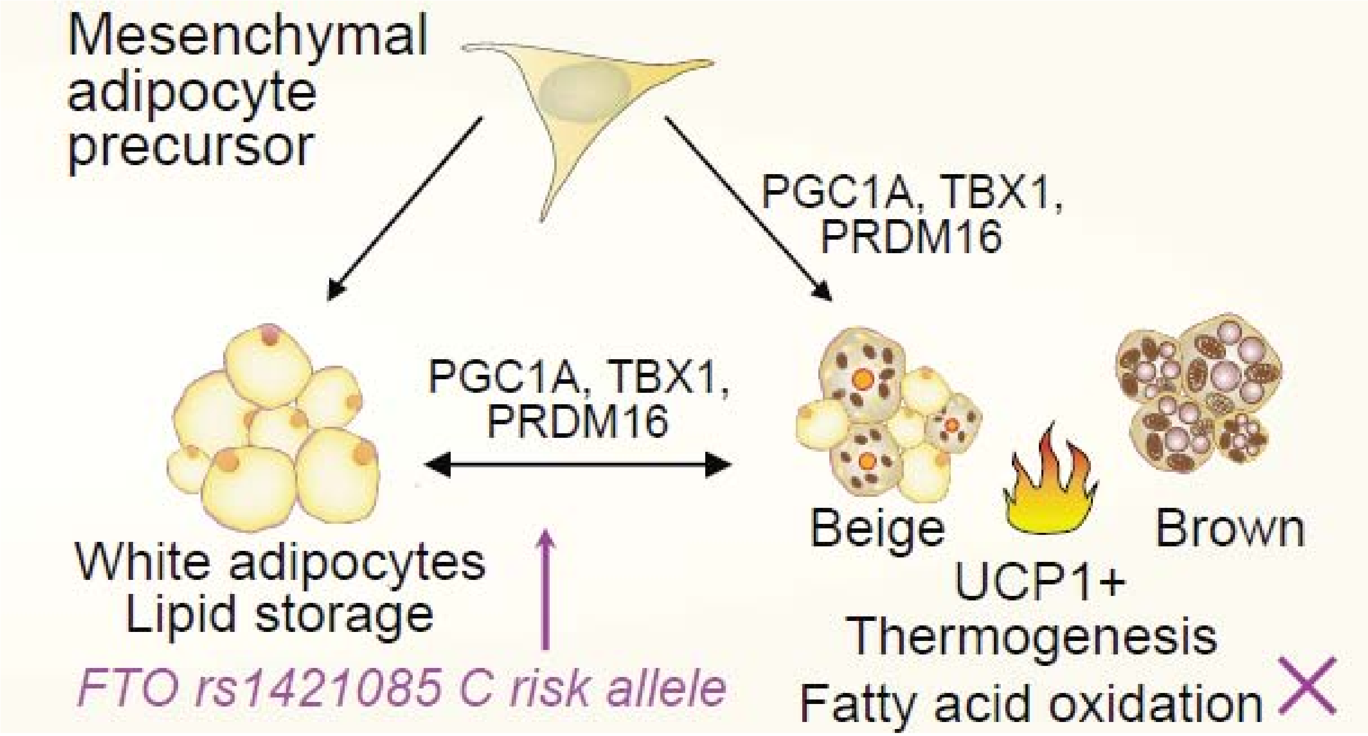
★ C-to-T motif rescue
(anti-obesity phenotypes)

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

IRX3, IRX5 overexpression
(pro-obesity phenotypes)

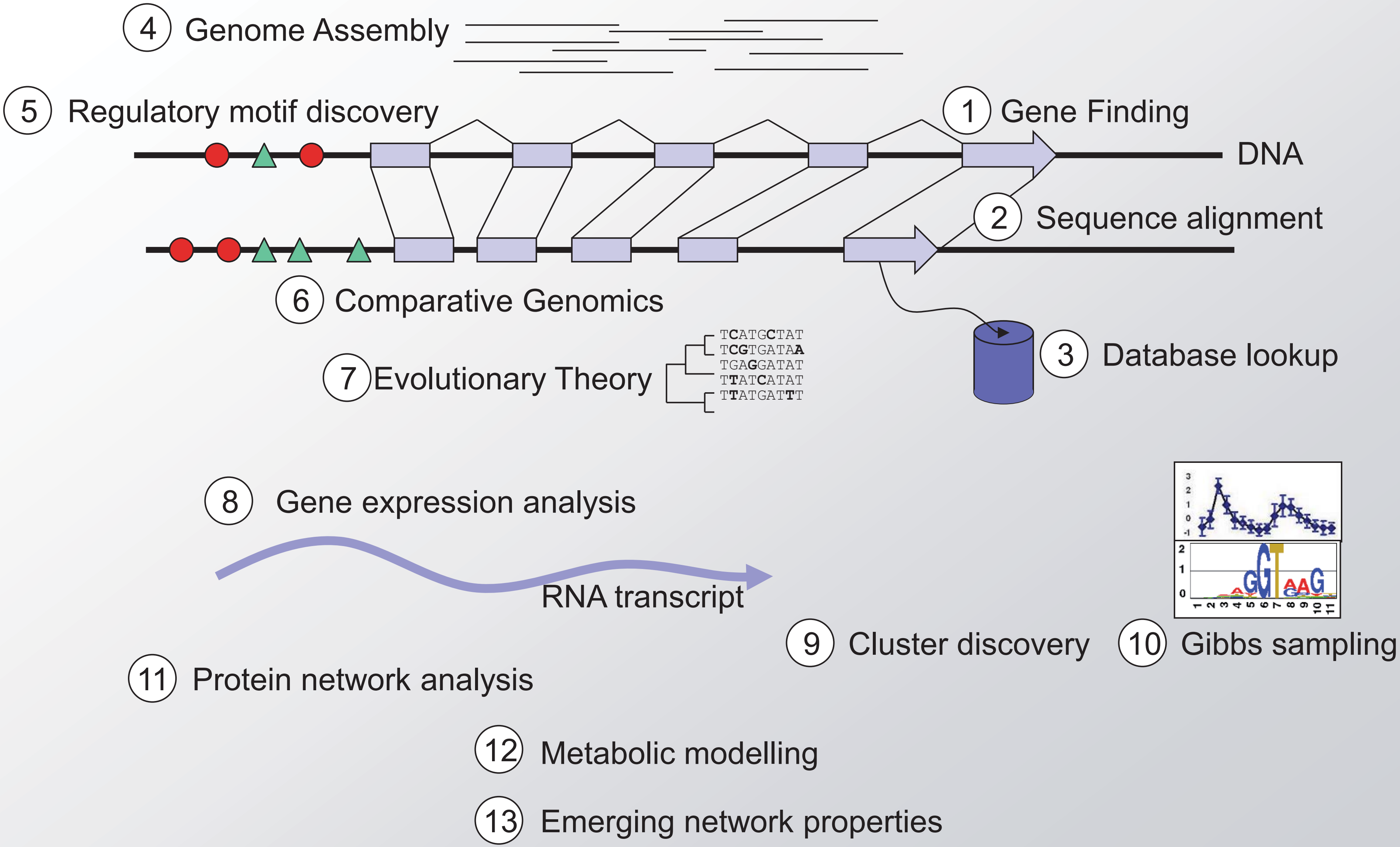
IRX3, IRX5 knock-down ★
(anti-obesity phenotypes)

Model: beige ⬄ white adipocyte development



Shift therapeutic focus from brain to adipocytes

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