Quick intro to the MIT CompBio Group

Who we are

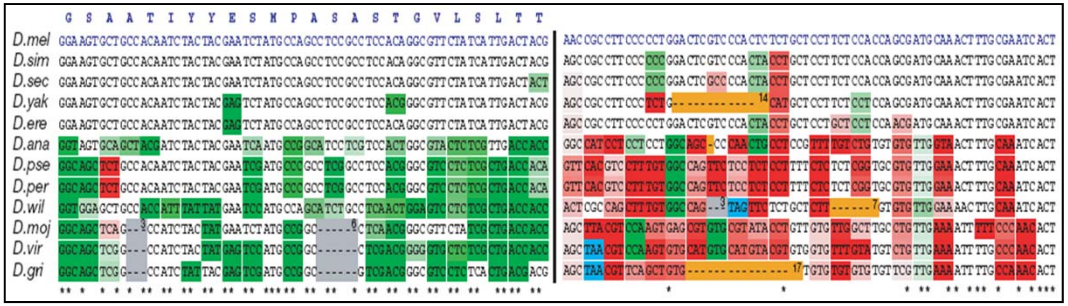
Stata3

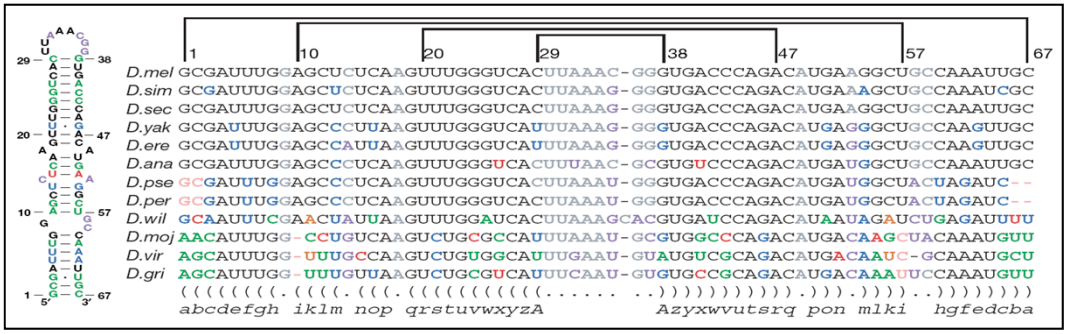
Stata4

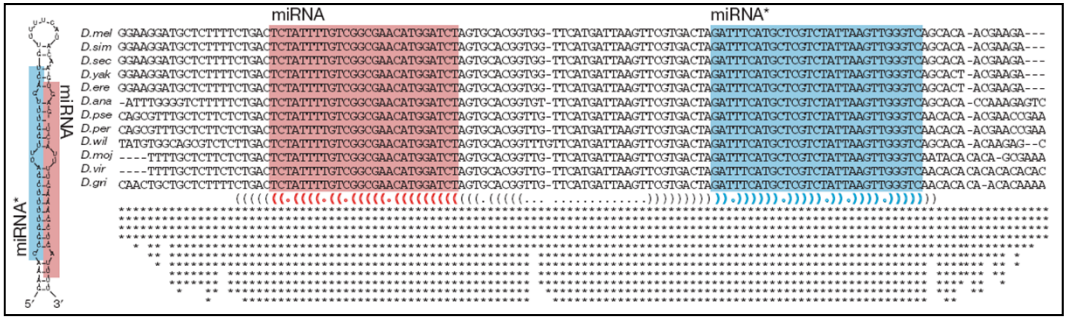
What we do: Research synopsis

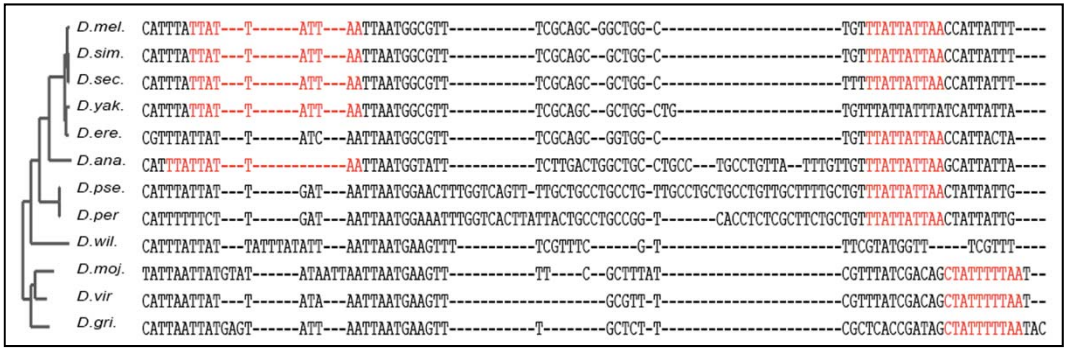
- **Why biology in a computer science group?**
- **Fundamental biological questions:**
  1. Interpreting the human genome.
  2. Revealing the logic of gene regulation.
  3. Principles of evolutionary change.
- **Algorithmic/machine learning methods:**
  - Comparative genomics: evolutionary signatures
  - Regulatory genomics: motifs, networks, models
  - Epigenomics: chromatin states, dynamics, disease
  - Phylogenomics: evolution at the genome scale
- **Defining characteristics of our group:**
  - Learn genomic rules, exploit nature of problems
  - Interdisciplinary collaborations, high biology impact

(1) Comparative genomics: evolutionary signatures









Protein-coding signatures

- 1000s new coding exons
- Translational readthrough
- Overlapping constraints

Non-coding RNA signatures

- Novel structural families
- Targeting, editing, stability
- Structures in coding exons

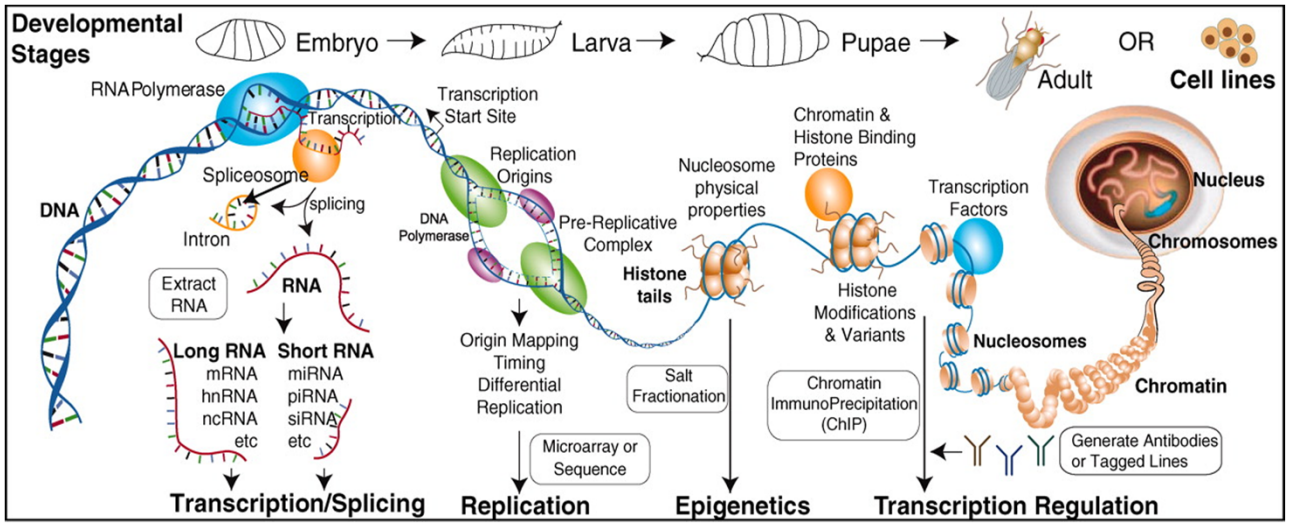
microRNA signatures:

- Novel/expanded miR families
- miR/miR* arm cooperation
- Sense/anti-sense switches

Regulatory motif signatures

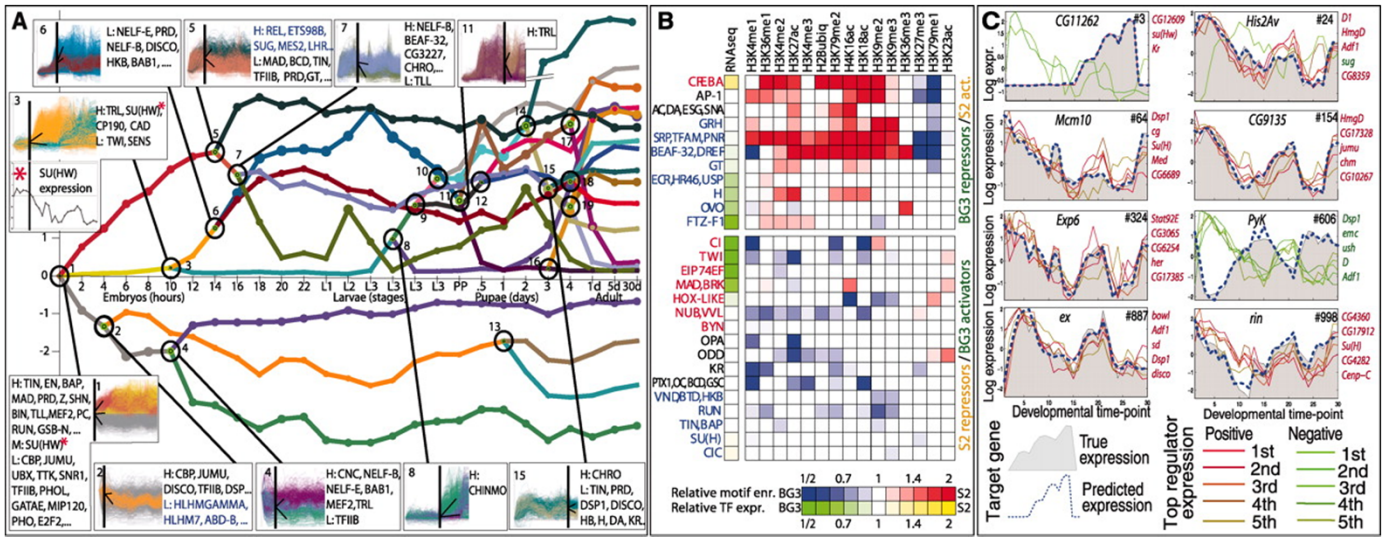
- Systematic motif discovery
- Regulatory motif instances
- TF/miRNA target networks
- Single binding-site resolution

(2) Regulatory genomics: circuits, predictive models



ENCODE/modENCODE

- 4-year effort, dozens of experimental labs
- Integrative analysis
- Systematic genome annotation
- Flagship NIH project



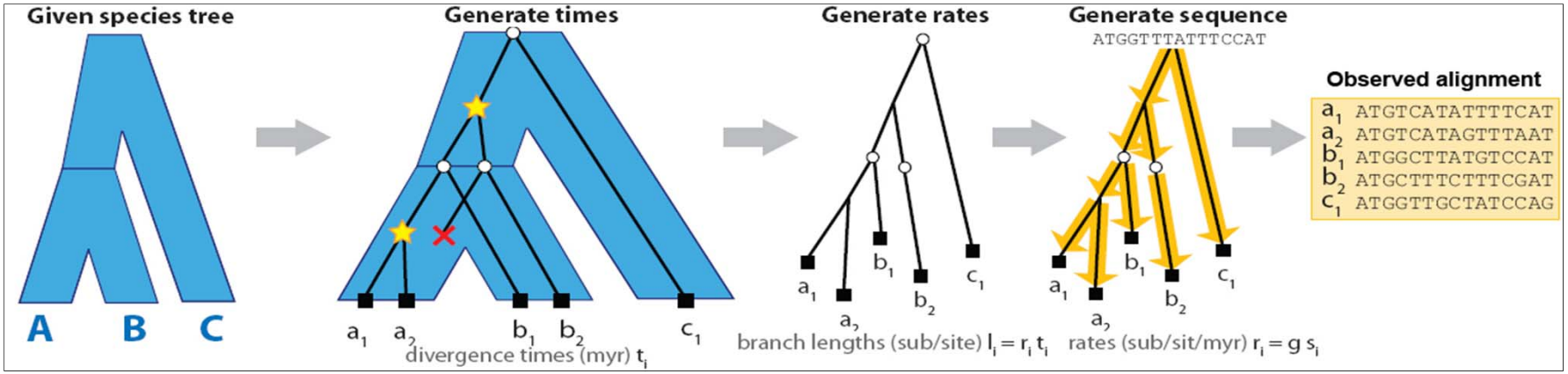
Predictive models of gene regulation

- Infer networks
- Predict function
- Predict regulators
- Predict gene expression

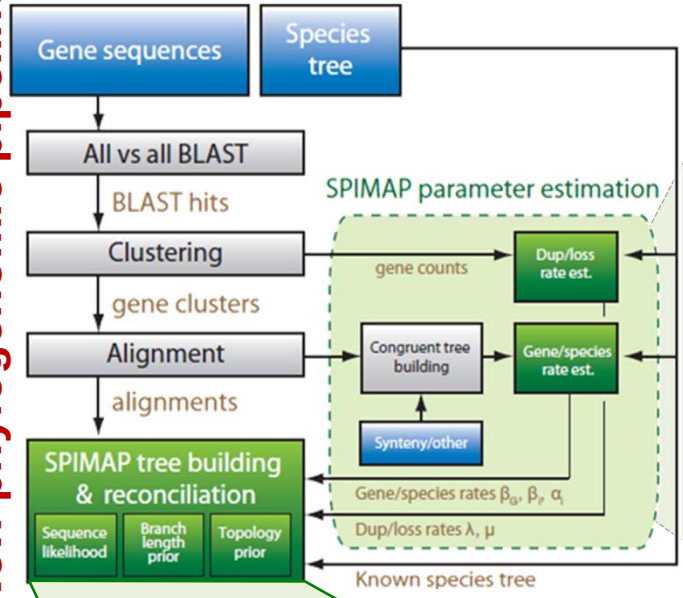
- **Initial annotation of the non-coding genome, from 20% to 70%**
- **Systems biology for an animal genome for the first time possible**
- **Students and postdocs are co-first authors, leadership roles**

(3) Phylogenomics: Bayesian gene-tree reconstruction

Generative model

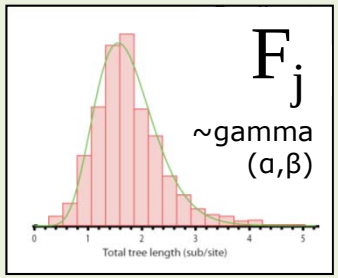


New phylogenomic pipeline



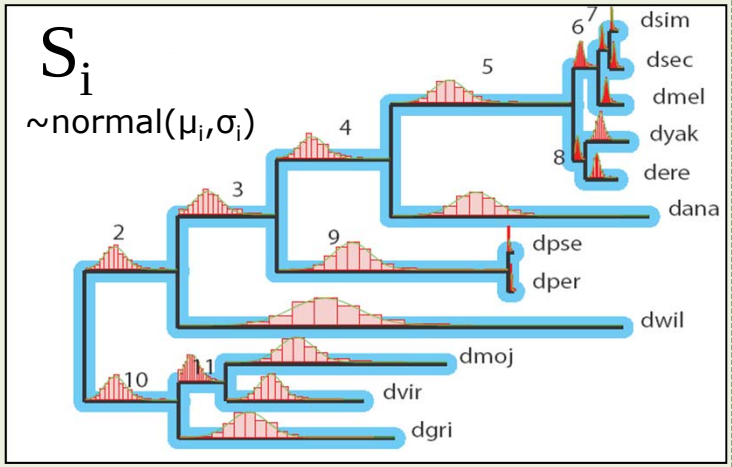
Two components of gene evolution

1. Family rate



Selective pressures on gene function

2. Species-specific rates



Population dynamics of the species

Bayesian formulation

Length l, Topology T, Reconciliation R

$$\hat{\mathbf{l}}, \hat{T}, \hat{R} = \underset{\mathbf{l},T,R}{\operatorname{argmax}} P(\mathbf{l}, T, R | \mathbf{D}, \theta) = \underset{\mathbf{l},T,R}{\operatorname{argmax}}$$

Alignment data **D**, species-level parameters θ

Sequence likelihood	Branch length prior	Topology prior
$P(\mathbf{D}\|\mathbf{l},T)$	$P(\mathbf{l}\|T,R,\theta)$	$P(T,R\|\theta)$
HKY model (traditional)	Learned F_j,S_i distributions	Birth-Death process

Jason Ernst

Discovery and characterization of chromatin states for systematic annotation of the human genome

Jason Ernst & Manolis Kellis

nature biotechnology

(4) Vignette on Epigenomics

Using chromatin information to understand human diseases

Pouya Kheradpour

Mapping and analysis of chromatin state dynamics in nine human cell types

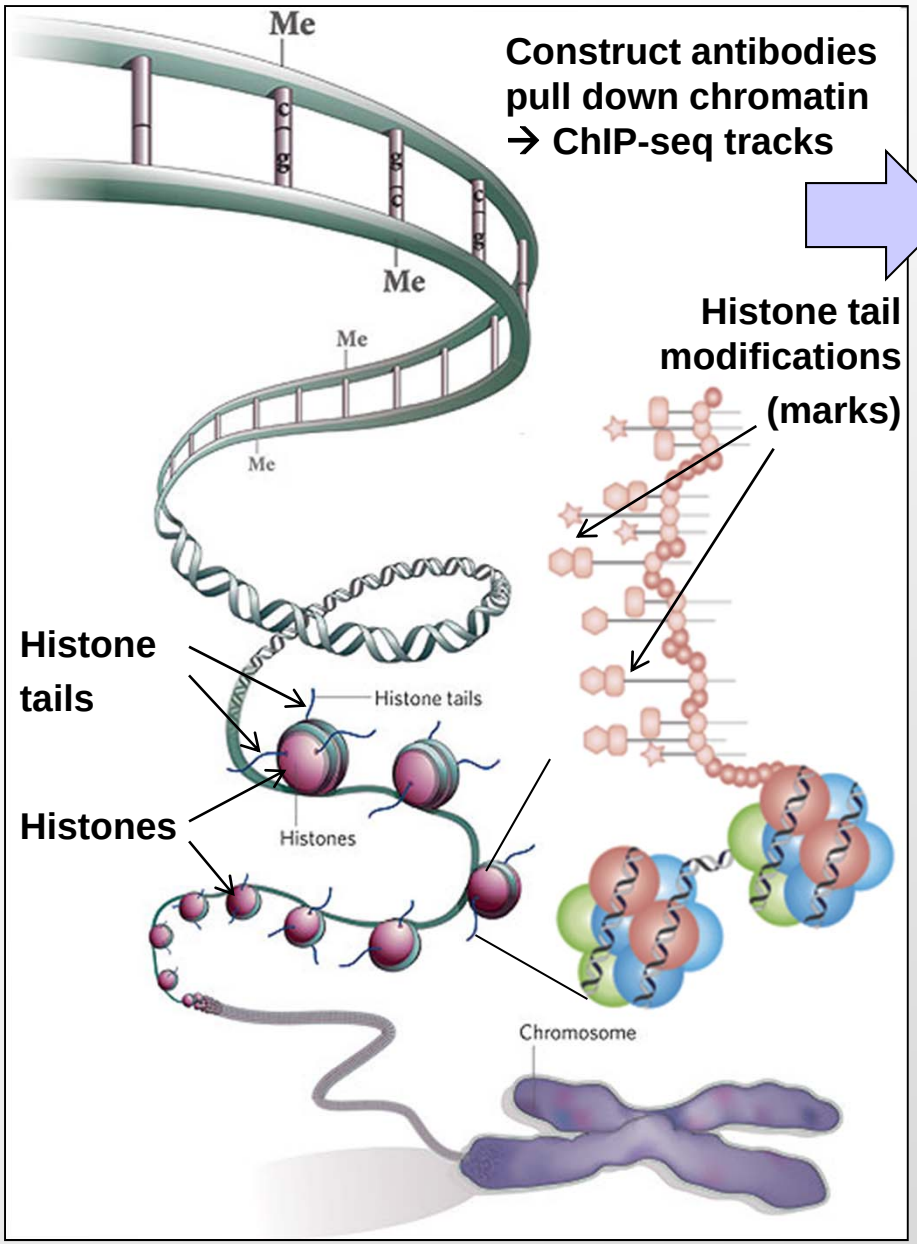
Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis & Bradley E. Bernstein

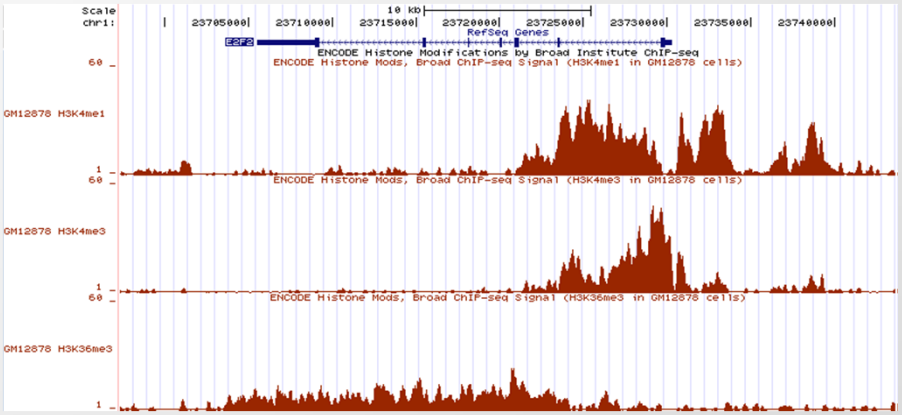
doi:10.1038/nature09906

Abstract | Full Text | PDF

nature

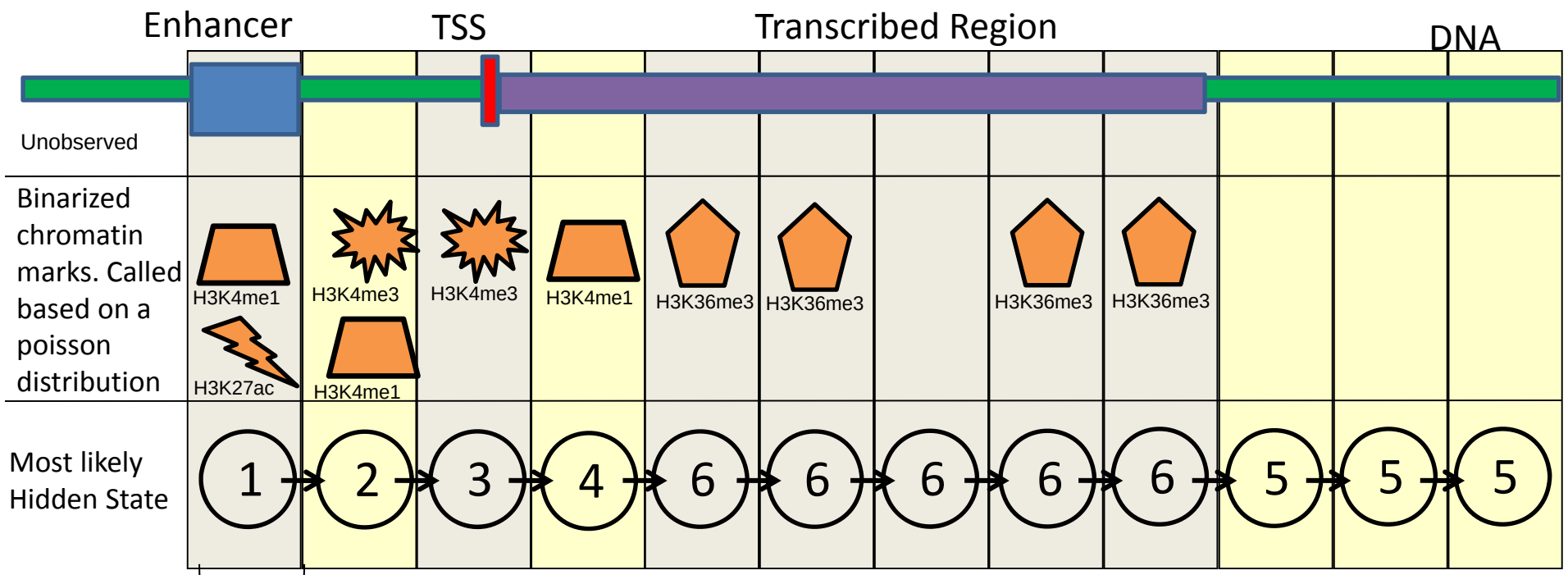
Challenge of data integration in many marks/cells





- **Dozens of chromatin tracks**
  - Understand their function
  - Reveal their combinations
  - Annotate systematically
- **Our approach: learn common chromatin states**
  - Explicitly model combinations
  - Unsupervised approach, probabilistic model

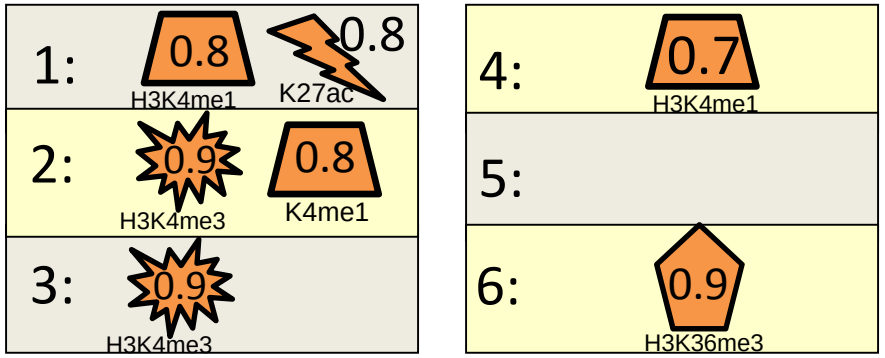
Our approach: Multivariate Hidden Markov Model (HMM)



200 base pair interval

Emission distribution is a product of independent Bernoulli random variables

High Probability Chromatin Marks in State

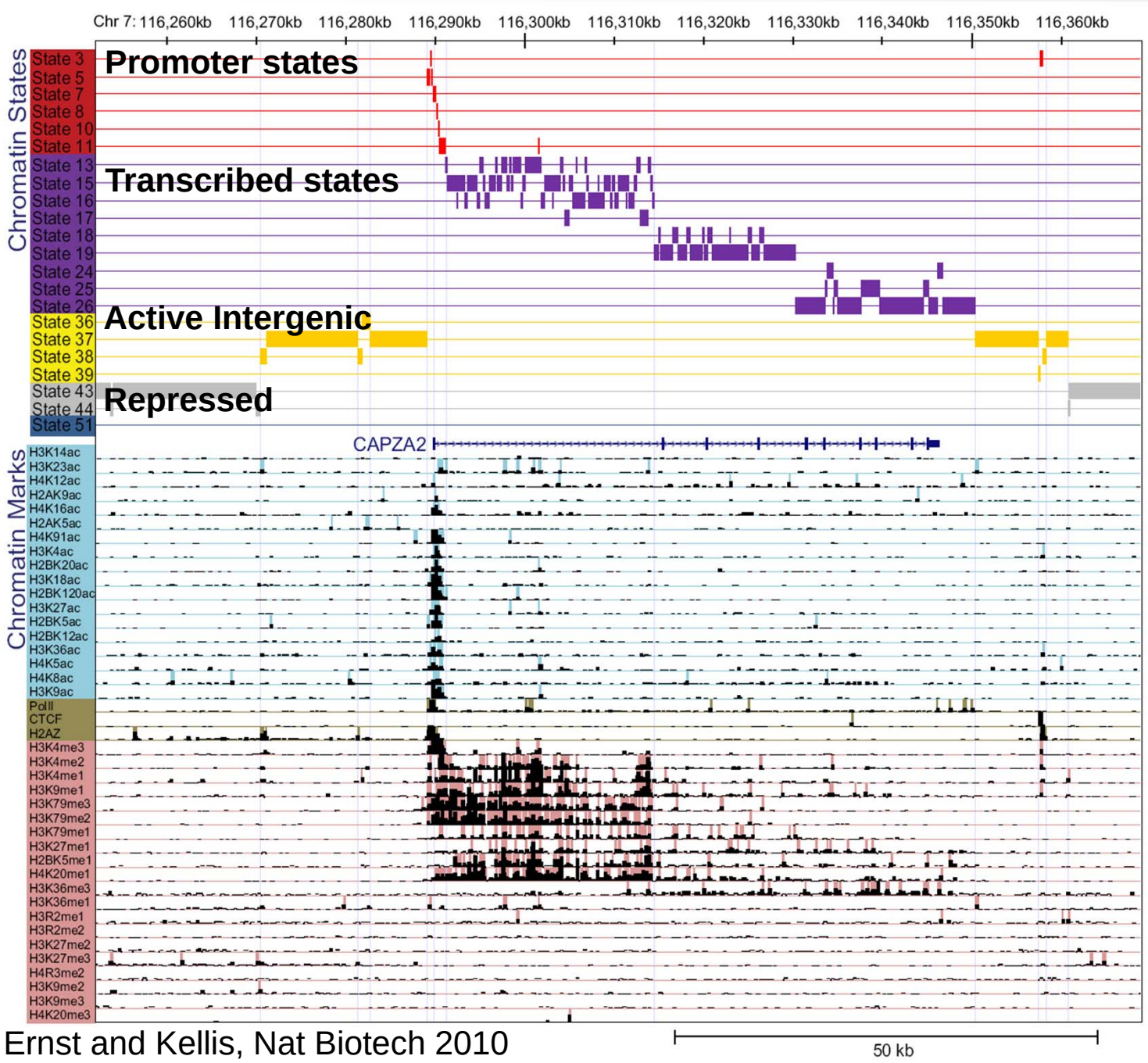


All probabilities are learned from the data

Binarization leads to explicit modeling of mark combinations and interpretable parameters

Ernst and Kellis, Nat Biotech 2010

From ‘chromatin marks’ to ‘chromatin states’



Ernst and Kellis, Nat Biotech 2010

- Learn *de novo* significant combinations of chromatin marks
- Reveal functional elements, even without looking at sequence
- Use for genome annotation
- Use for studying regulation dynamics in different cell types

ENCODE: Study nine marks in nine human cell lines

9 marks

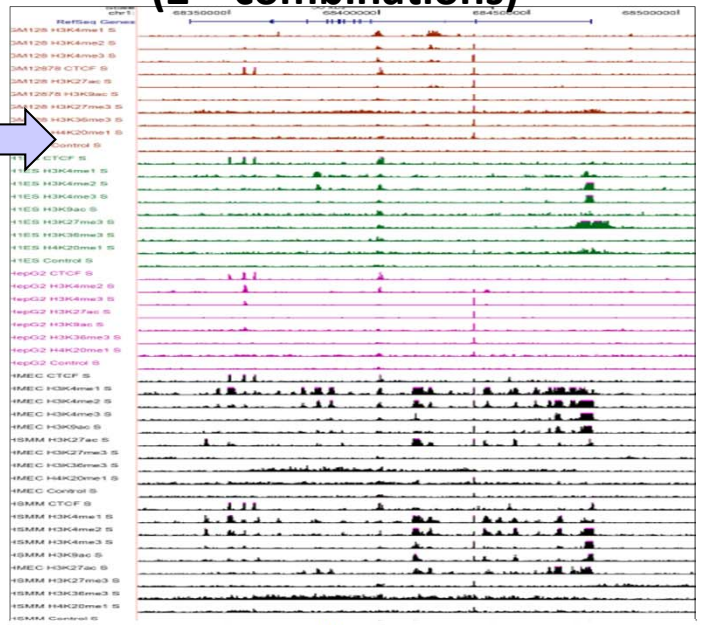
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

X

9 human cell types

HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

81 Chromatin Mark Tracks (2^{81} combinations)

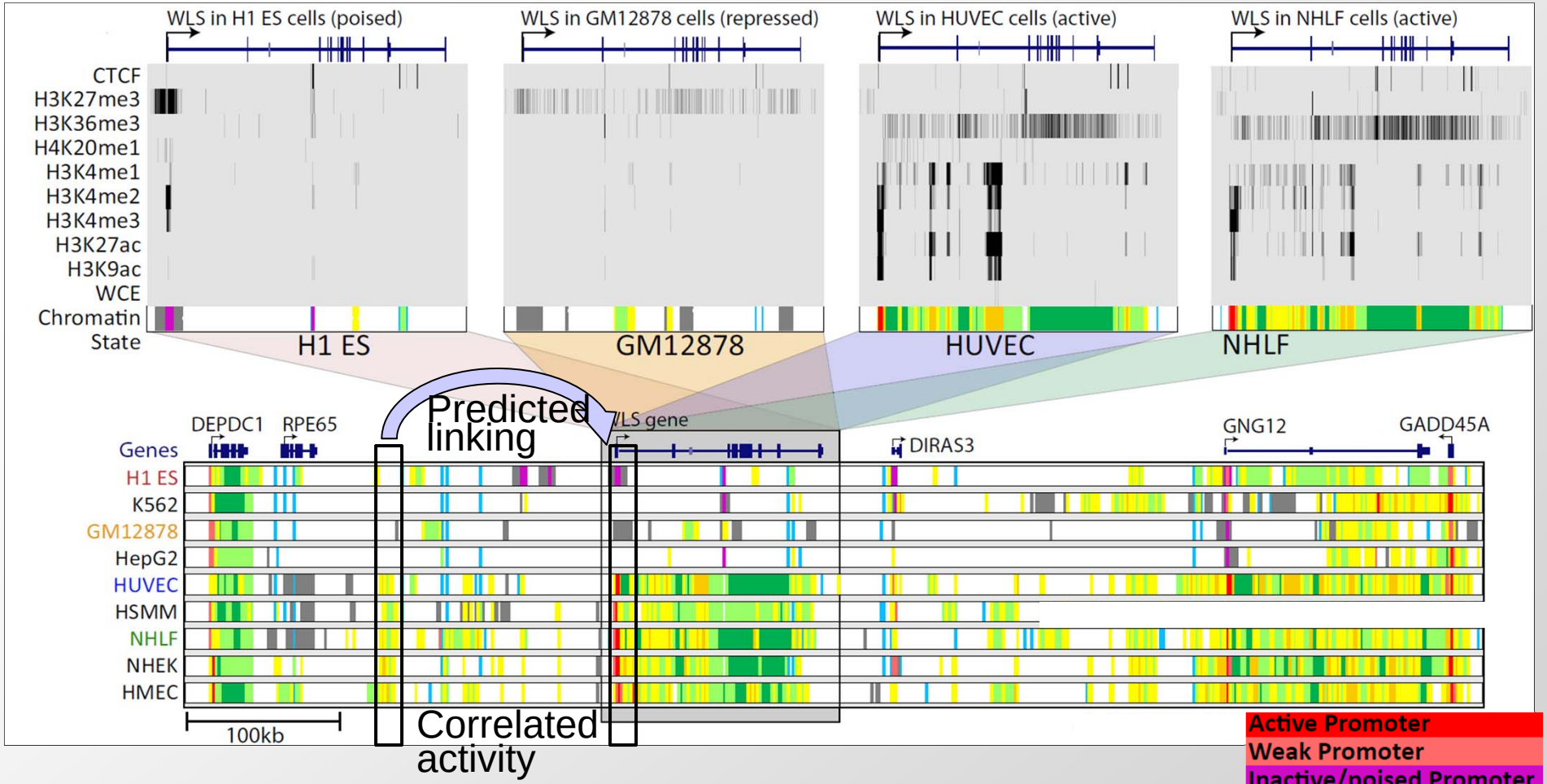


Brad Bernstein ENCODE Chromatin Group

b.

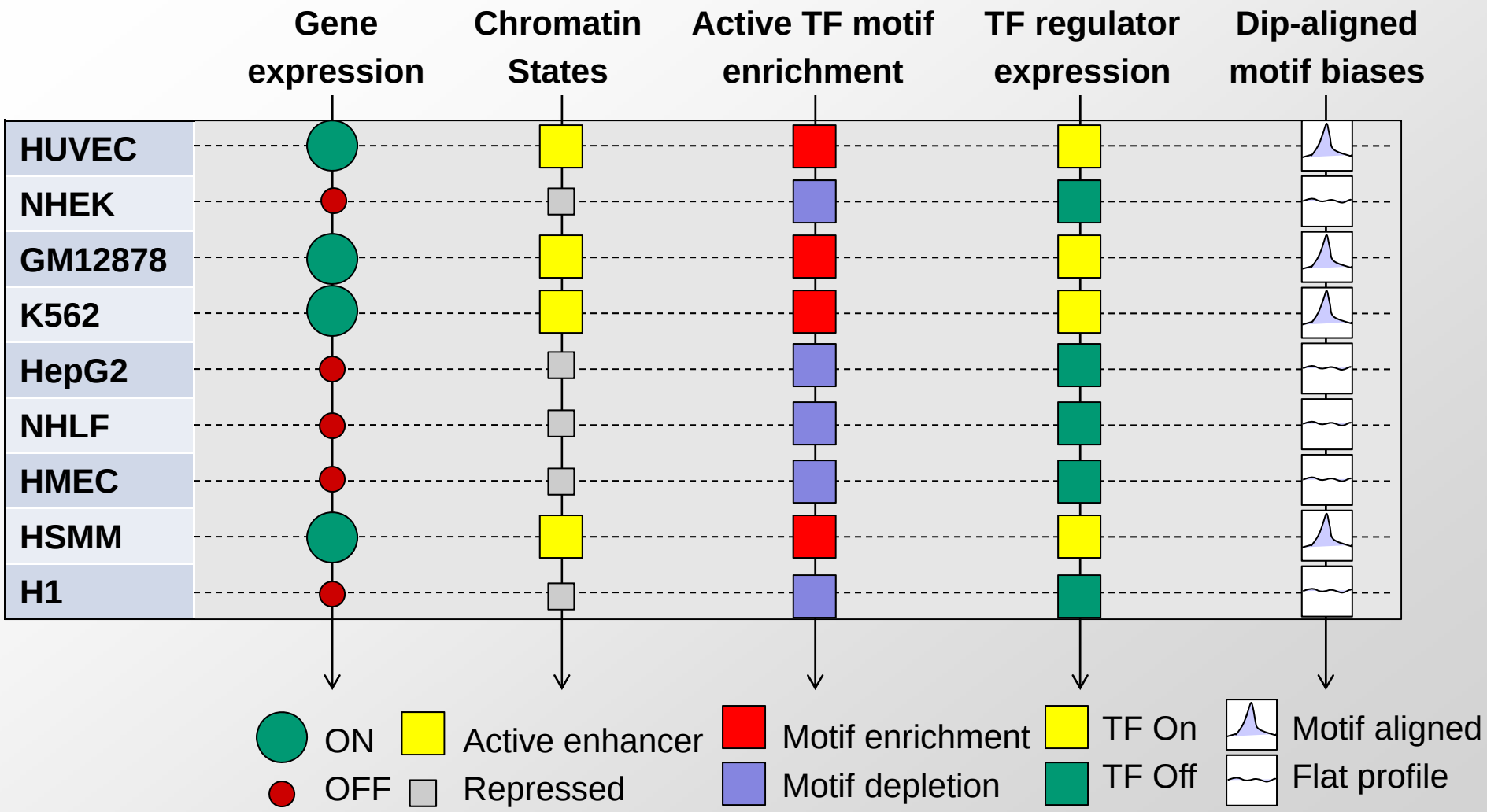
State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage: Median	Coverage: H1 ES	Coverage: GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
Chromatin States 1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active Promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak Promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised Promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb-repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV
	Chromatin Mark Observation Frequency (%)										(%)	(fold)		(kb)	(%)	Functional enrichments (fold)						

Chromatin states dynamics across nine cell types



- **Single annotation track for each cell type**
- **Summarize cell-type activity at a glance**
- **Can study 9-cell activity pattern across ↓**

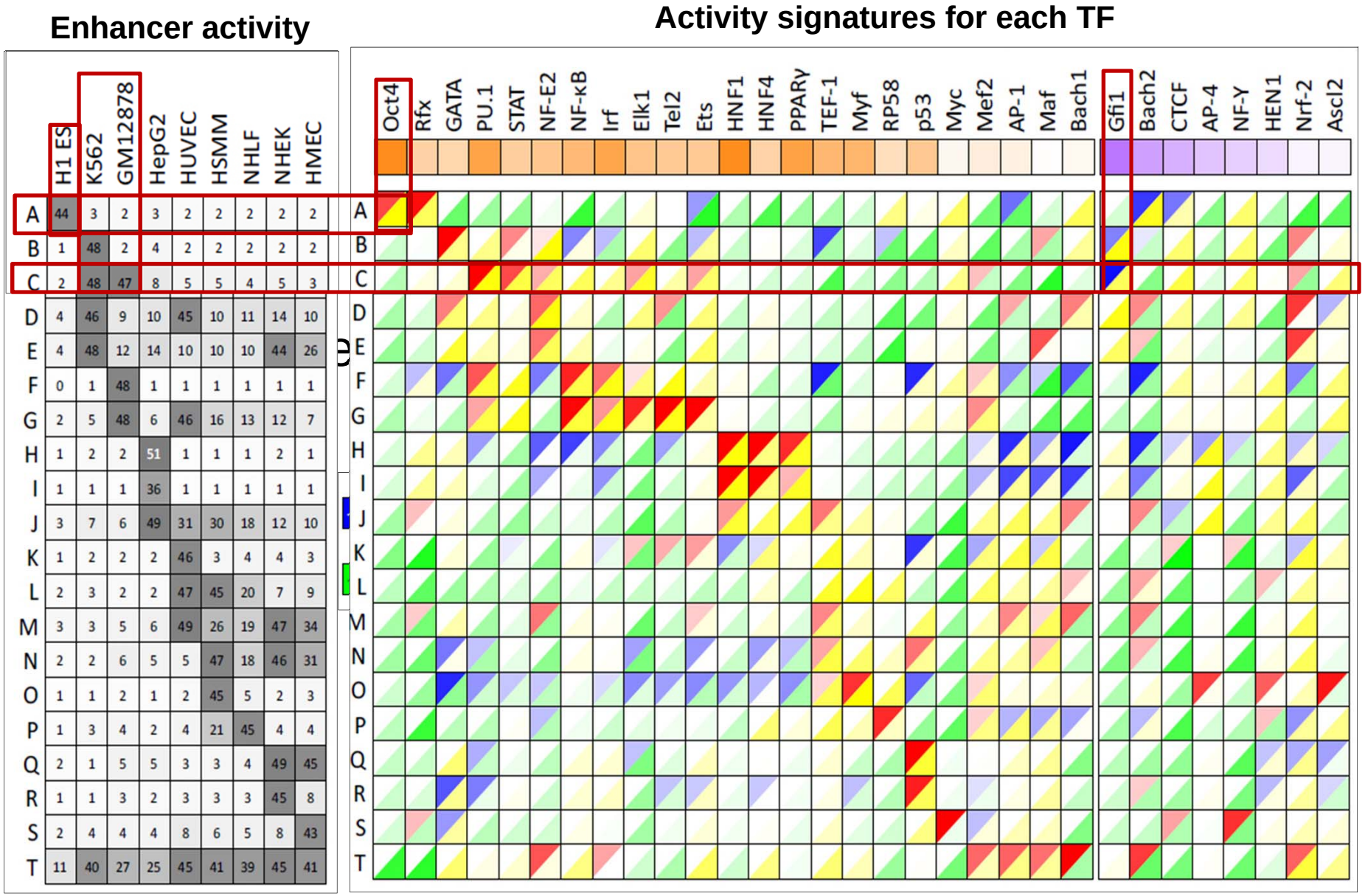
Multi-cell activity profiles and their correlations



Chromatin state & gene expression → link enhancers and target genes

TF motif enrichment & TF expression → reveal activators / repressors

Coordinated activity reveals activators/repressors

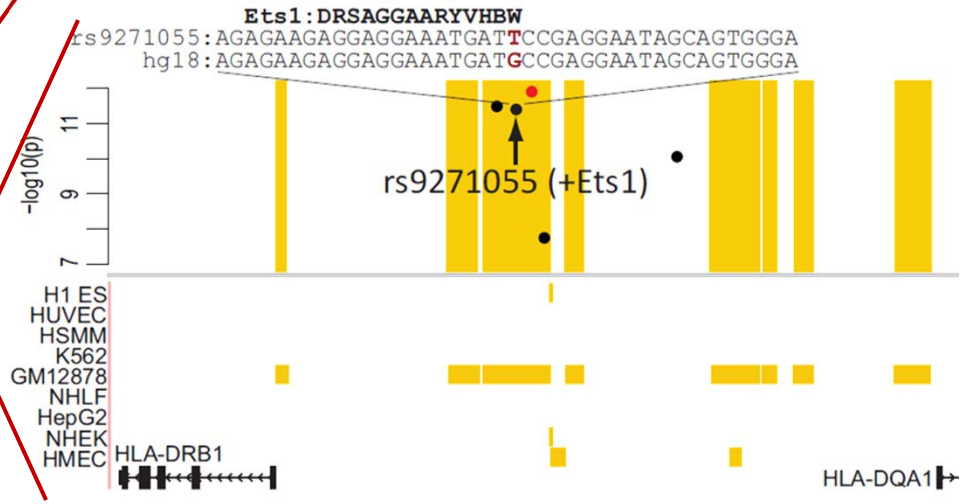


- Enhancer networks: Regulator → enhancer → target gene

Revisiting disease-associated 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**

Regulatory roles revealed for many studies

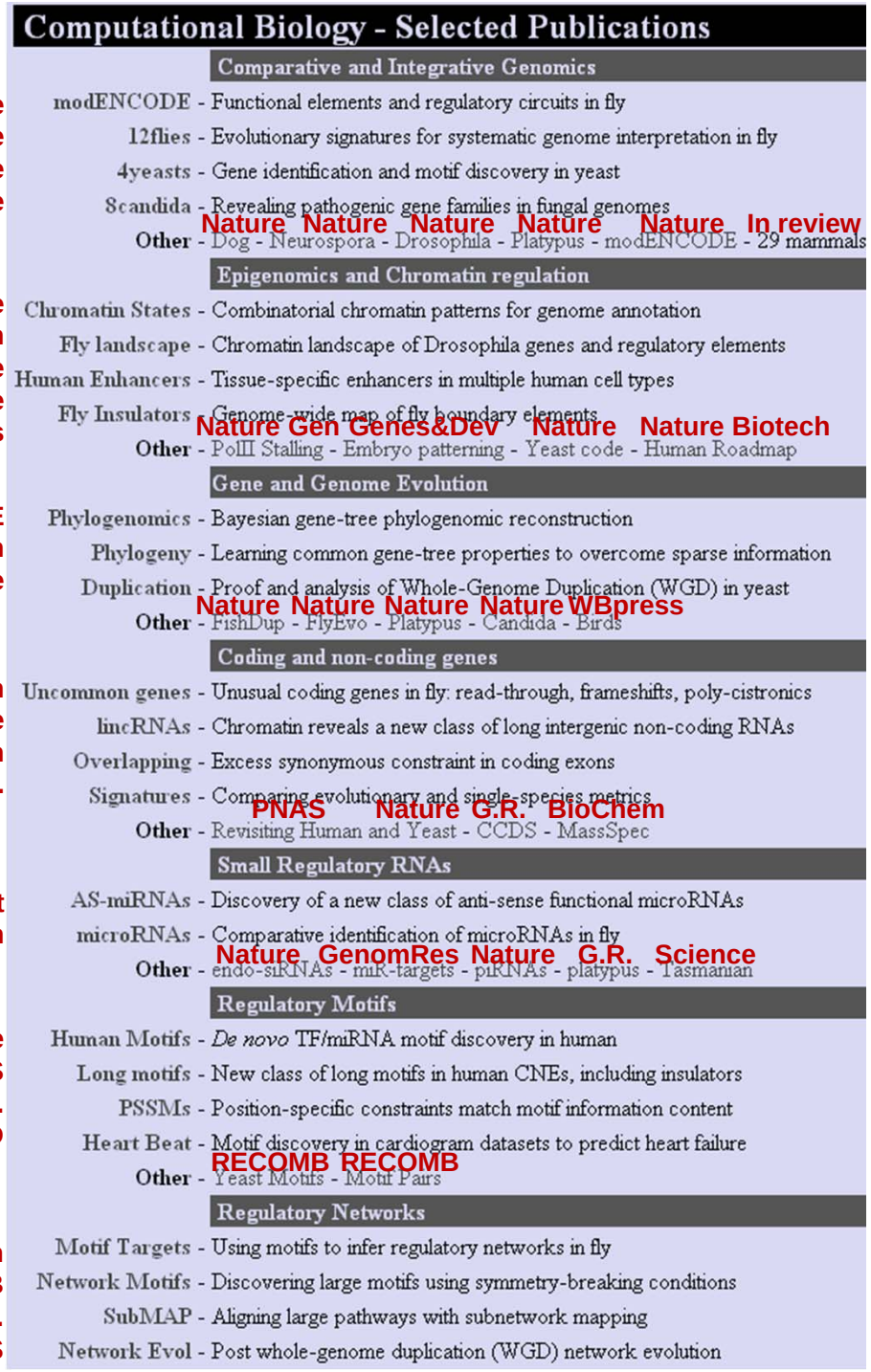
Title	Author/ Journal	Total #SNPs	Fold	Cell Type	# SNPs in enhancers	FDR
Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium.	Ganesh et al Nat Genet 2009	35	17	K562	9	0.02
Biological, clinical and population relevance of 95 loci for blood lipids	Teslovich et al Nature 2010	101	11	HepG2	13	0.02
Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci	Stahl et al Nat Genet 2010	29	15	GM12878	7	0.03
Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis	Liu et al Nat Genet 2010	6	41	GM12878	4	0.03
Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus.	Han et al Nat Genet 2009	18	21	GM12878	6	0.03
Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans.	Kathiresan et al Nat Genet 2008	18	24	HepG2	5	0.03
Genome-wide association study of hematological and biochemical traits in a Japanese population	Kamatani et al Nat Genet 2009	39	12	K562	7	0.03
A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium.	Soranzo et al Nat Genet 2009	28	15	K562	6	0.03
Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer.	Houlston et al Nat Genet 2008	4	66	HepG2	3	0.03
Genome-wide association study identifies eight loci associated with blood pressure.	Newton-Chen Nat Genet 2009	9	30	K562	4	0.04

We work hard…

We aim to further our understanding of the **human genome** by computational integration of large-scale functional and comparative genomics datasets.

- We use comparative genomics of multiple related species to recognize **evolutionary signatures** of protein-coding genes, RNA structures, microRNAs, regulatory motifs, and individual regulatory elements.
- We use combinations of epigenetic modifications to define **chromatin states** associated with distinct functions, including promoter, enhancer, transcribed, and repressed regions, each with distinct functional properties.
- We develop **phylogenomic** methods to study differences between species and to uncover evolutionary mechanisms for the emergence of new gene functions

Our methods have led to numerous new insights on diverse regulatory mechanisms, uncovered evolutionary principles, and provide mechanistic insights for previously uncharacterized disease-associated SNPs



But we also have fun

Kayaking

Whitewater Rafting

BBQs

Canadian Thanksgiving

Sailing

Want to get started?

6.047/6.878 – Computational Biology: Genomes, Networks, Evolution

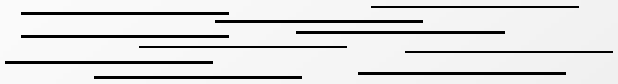
6.047 Computational Biology: Genomes, Networks, Evolution

(Subject meets with 6.878J, HST.507J)
Prereq: 6.006, 6.041, Biology (GIR); or permission of instructor
Units: 3-0-9
Lecture: *TR9.30-11* (4-159) **Recitation:** *F1* (36-156)

Covers the algorithmic and machine learning foundations of computational biology, combining theory with practice. Principles of algorithm design, influential problems and techniques, and analysis of large-scale biological datasets. Topics include (a) genomes: sequence analysis, gene finding, RNA folding, genome alignment and assembly, database search; (b) networks: gene expression analysis, regulatory motifs, biological network analysis; (c) evolution: comparative genomics, phylogenetics, genome duplication, genome rearrangements, evolutionary theory. These are coupled with fundamental algorithmic techniques including: dynamic programming, hashing, Gibbs sampling, expectation maximization, hidden Markov models, stochastic context-free grammars, graph clustering, dimensionality reduction, Bayesian networks.
M. Kellis

Challenges in computational biology

4 Assembly/BWT

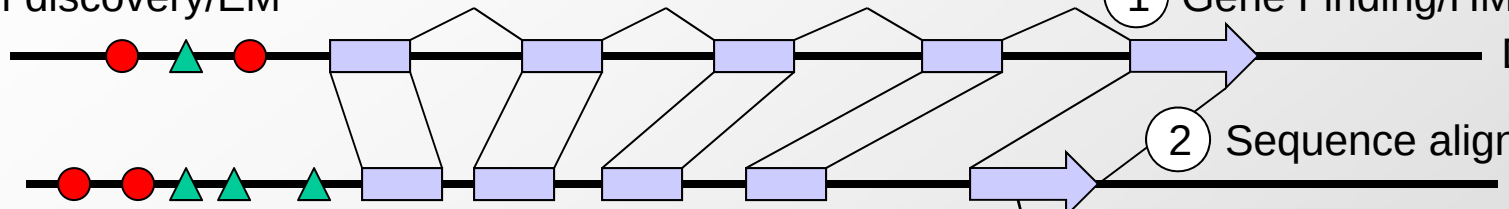


5 Motif discovery/EM

1 Gene Finding/HMMs

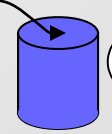
DNA

2 Sequence alignment/DP



6 Comparative Genomics/model fits

7 Evolutionary Theory and phylogenetics

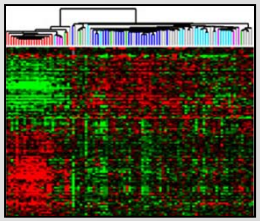


3 Rapid lookup

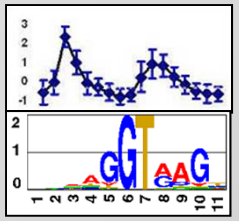
BLAST/Hashing

8 Gene expression/Bayesian classification

RNA transcript

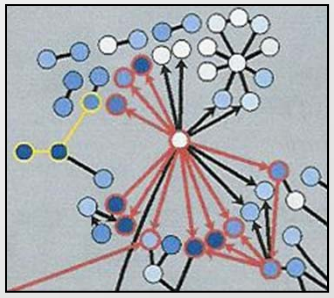


9 Cluster discovery unsupervised learning



10 Motif finding/ Gibbs sampling

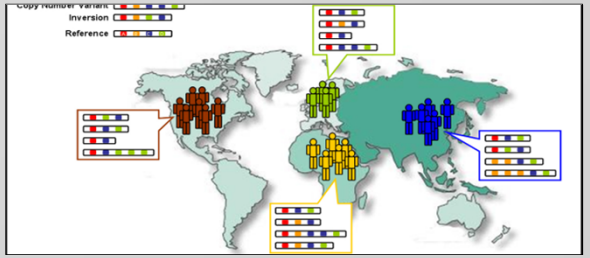
11 Network analysis/Graph theory



12 Metabolic modeling/systems

13 Predictive models/regression

14 Population Genomics/ coalescent theory/selection



4 modules: Genomes, Genes, Networks, Evolution

Psets	Week	Date	Topic	Category	Lec	Topic
PS1 out covers:L1-L5	1	Thu, Sep 09	Introduction		L1	Intro: Biology, Algorithms, Machine Learning
		Fri, Sep 10			R1	Recitation1: Probability, Statistics, Biology
	2	Tue, Sep 14	Module I: Comparative Genomics	Foundations	L2	Global/local alignment/DynProg
		Thu, Sep 16			L3	StringSearch/Blast/DB Search
		Fri, Sep 17			R2	Recitation 2 - Multiple, Progressive, Phylogenetic, Whole-genome alignment
	3	Tue, Sep 21		Frontiers	L4	Comparative genomics, Selection, Evolutionary signatures of coding genes
due Mon 9/27		Thu, Sep 23			L5	Evolutionary Signatures of RNA structures, miRNAs, motifs, nucleotides
		Fri, Sep 24			R3	Recitation 3 - Evolutionary signatures and measures of selection
PS2 out covers:L6-R5	4	Tue, Sep 28	Module II: Coding and non-coding genes	Foundations	L6	HMMs1 - Evaluation / Parsing
		Thu, Sep 30			L7	HMMs2 - PosteriorDecoding/Learning
		Fri, Oct 01			R4	Recitation4 - Posterior decoding review, Baum-Welch Learning
	5	Tue, Oct 05		Frontiers	L8	Gene finding in practice: CRFs, Discriminative training, GHMMs, RNAseq
due Wed 10/13		Thu, Oct 07			L9	Structural RNAs: Fold prediction and genome-wide annotation
		Fri, Oct 08			R5	Recitation 5 - SVMs and discriminative learning
PS3 out covers:L10-R8	6	Thu, Oct 14	Module III: Networks and Gene Regulation	Foundations	L10	Expression Analysis: Clustering, Classification, Feature Selection
		Fri, Oct 15			R6	Recitation 6 - Microarrays/RNAseq expression analysis
	7	Tue, Oct 19			L11	Regulatory Motif Discovery: Gibbs Sampling, Expectation Maximization
		Thu, Oct 21			L12	TF/miRNA target prediction and regulatory network inference
		Fri, Oct 22			R7	Recitation 7 - Entropy, Information, Background models
	8	Tue, Oct 26		Frontiers	L13	Regulatory Network Analysis, Bayesian Function Prediction
due Mon 11/1		Thu, Oct 28			L14	Epigenomics: Chromatin marks, chromatin states, and their dynamics
		Fri, Oct 29			R8	Recitation 8 - Regulatory genomics and epigenomics
PS4 out covers:L15-R10	9	Tue, Nov 02	Module IV: Evolution and Phylo-genomics	Foundations	L15	Phylogenetics: Molecular Evolution, Tree Building, Phylogenetic inference
		Thu, Nov 04			L16	Phylogenomics: gene/species trees,reconciliation,coalescent,populations
		Fri, Nov 05			R9	Recitation 9 - Gene Trees, Species Trees, Reconciliation
	10	Tue, Nov 09			L17	Population genomics: Statistical genetics and human disease mapping
		Fri, Nov 12			R10	Recitation 10 - Population genetics and genomics
	11	Tue, Nov 16		Frontiers	L18	Population genomics: Measuring natural selection in human populations
due Mon 11/22		Thu, Nov 18			L19	Population genomics: Learning population history from genetic data
		Fri, Nov 19			R10	Recitation 11 - Quiz Review
	12	Tue, Nov 23	In Class (Friendly) Quiz		L20	In Class Quiz (the only Quiz) - covers L1-R10

- Foundations and frontiers. Practical/algorithmic problems

Final project: mentoring, milestones

Project planning

Project execution

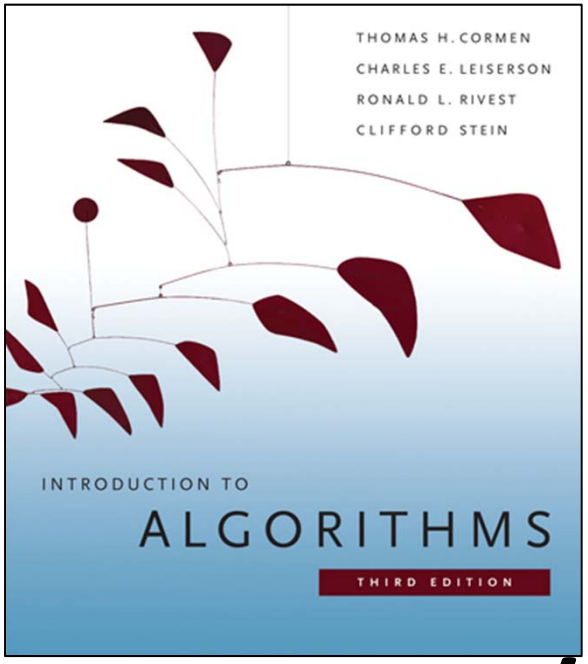
Week	Term Project	Psets
1	**I. Project Set-up.** Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post these in a profile that lets your classmates know you and find potential partners. **Due Mon 9/27** with PS1	PS1 out covers:L1-L5
2		
3		due Mon 9/27
4	**II.Brainstorming.** Identify previous project proposals, recent papers, and potential partners that match your areas of interest. List initial project ideas and partners. **Due Wed 10/13** with PS2	PS2 out covers:L6-R5
5		due Wed 10/13
6	**III. Proposal.**Form teams (2or1), specify project goals, division of work, milestones, datasets, challenges, aims, algorithms, in form of NIH proposal. **Due Mon 10/25.**	PS3 out covers:L10-R8
7		
8	**b.Review.**Evaluate/discuss 3 peer proposals, NIH format, provide suggestions/advice. **Due Mon 11/1** with PS3.	due Mon 11/1
9	**c.Response.** Bullet response to reviews, revise scope/aims as needed.	PS4 out covers:L15-R10
10	**IV.Midterm progress report.** Continue making progress on proposed milestones. Adjust aims/deliverables as needed. Write outline of final report, with draft methods/figs/tables **Due Mon 11/22** with PS4	
11		due Mon 11/22
12	**V. Final project report.** Finish methods, milestones, results, figures, write-up in conference paper format. In report, comment on lessons learned and overall project experience. **Due Mon 12/6.**	Complete Final Project
13		
14	**VI.Final class presentation** ~10min conference talk on your final project. **Thu 12/9**	

Week	Date	Lec
1	Thu, Sep 09	Module I: Comparative Genomics
	Fri, Sep 10	
2	Tue, Sep 14	
	Thu, Sep 16	
	Fri, Sep 17	
3	Tue, Sep 21	
	Thu, Sep 23	
	Fri, Sep 24	
	Fri, Sep 24	Project Mentoring
4	Tue, Sep 28	
	Thu, Sep 30	
	Fri, Oct 01	
5	Tue, Oct 05	Module II: Coding and non-coding genes
	Thu, Oct 07	
	Fri, Oct 08	
6	Tue, Oct 12	
	Thu, Oct 14	
	Fri, Oct 15	
	Fri, Oct 15	Project Mentoring
7	Tue, Oct 19	
	Thu, Oct 21	
	Fri, Oct 22	Module III: Networks and Gene Regulation
8	Tue, Oct 26	
	Thu, Oct 28	
	Fri, Oct 29	
	Fri, Oct 29	Project Mentoring
9	Tue, Nov 02	
	Thu, Nov 04	
	Fri, Nov 05	Module IV: Evolution and Phylo-genomics
10	Tue, Nov 09	
	Thu, Nov 11	
	Fri, Nov 12	
	Fri, Nov 12	Project Mentoring
11	Tue, Nov 16	
	Thu, Nov 18	
	Fri, Nov 19	
12	Tue, Nov 23	Final Project Mode
	Thu, Nov 25	
	Fri, Nov 26	
13	Tue, Nov 30	
	Thu, Dec 02	
	Fri, Dec 03	
	Fri, Dec 03	Project Mentoring
14	Tue, Dec 07	
	Thu, Dec 09	9:30am
	Thu, Dec 09	3pm

For more info: compbio.mit.edu (or talk to any of us…)

6.006

Introduction to Algorithms



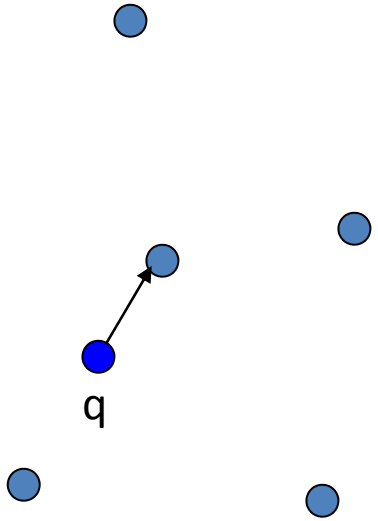
Lecture 25 2/3

(High-dimensional) geometry

Or: 6.850 “Geometric Computing” preview

Nearest Neighbor

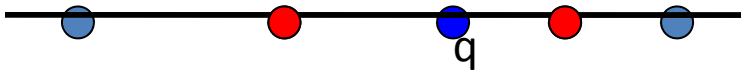
- Given: a set P of n points in R^d
- Nearest Neighbor: for any query q, returns a point $p \in P$ minimizing the Euclidean distance $||p-q||$



The case of d=1

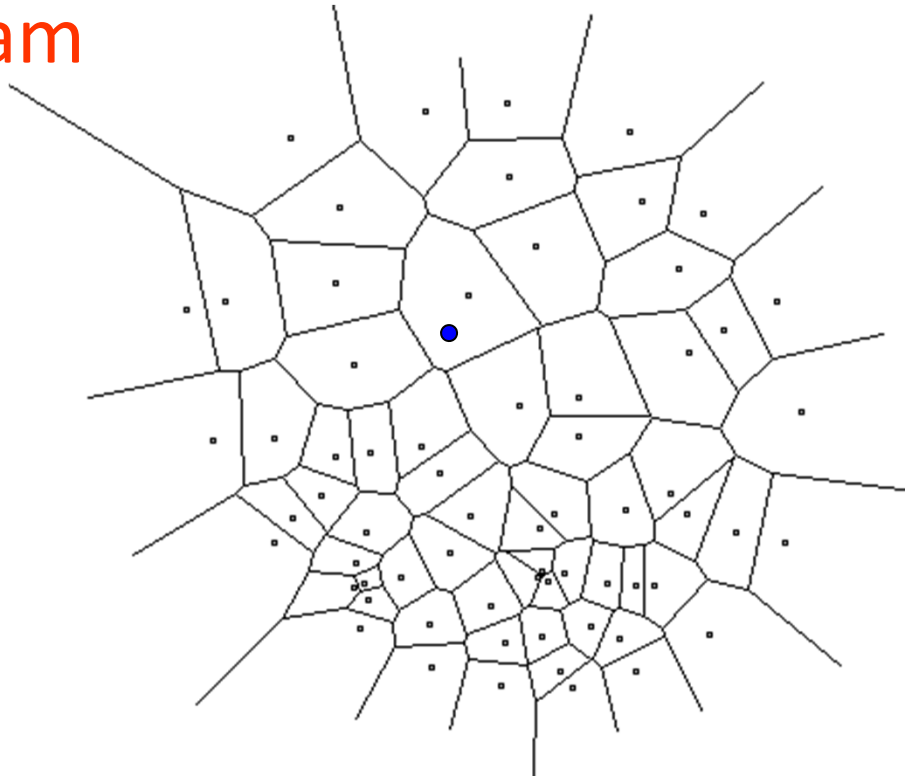
- Build BST
- Next-larger(q): finds the next element after element q
- Next-smaller(q): analogous
- The closer of the two is the nearest neighbor
- Performance:
  - Space: O(n)
  - Query time: O(log n)

  assuming balanced BST



The case of d=2

- Compute Voronoi diagram
- Given q, perform point location
- Performance:
  - Space: O(n)
  - Query time: O(log n)



The case of d>2

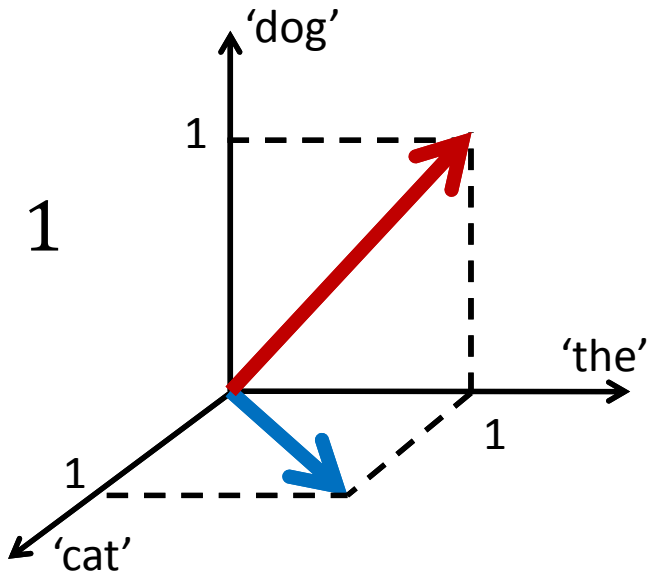
- Voronoi diagram has size $n^{O(d)}$

  "Curse of dimensionality"
- We can also perform a linear scan: $O(dn)$ time
- Both are pretty bad if d,n=few million
- Why would d be a few million ?

Example: Vector Space Model
[Salton, Wong, Yang 1975]

- Treat each document D as a vector of its words
  - One coordinate $D(w)$ for every possible word w
- Example:
  - D_1 = "the cat"
  - D_2 = "the dog"
- Similarity between vectors?
  - **Dot product:**

$$D_1 \cdot D_2 = 1$$



$$D_1 \cdot D_2 = \sum_w D_1(w) \cdot D_2(w)$$

http://portal.acm.org/citation.cfm?id=361220

Vector Space Model ctd
[Salton, Wong, Yang 1975]

- We have

$$||D_1 - D_2||^2 = ||D_1||^2 + ||D_2||^2 - 2\, D_1 \cdot D_2$$

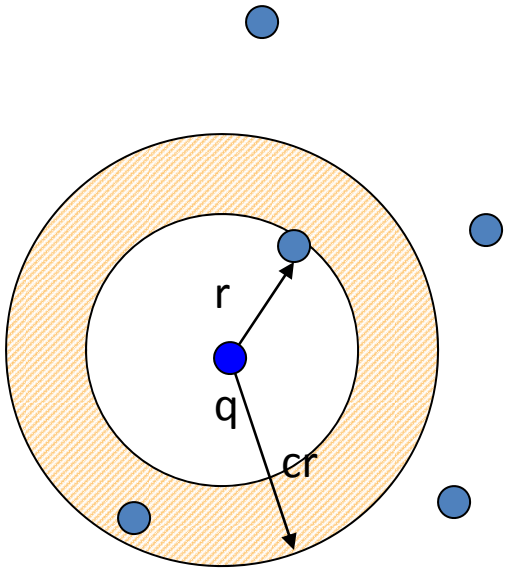
- If we normalize D_1, D_2 then

$$||D_1 - D_2||^2 = 2 - 2\, D_1 \cdot D_2$$

- Minimizing Euclidean distance = Maximizing Dot Product
- Many other applications: searching for similar bio sequences, similar images, etc

Approximate Near Neighbor

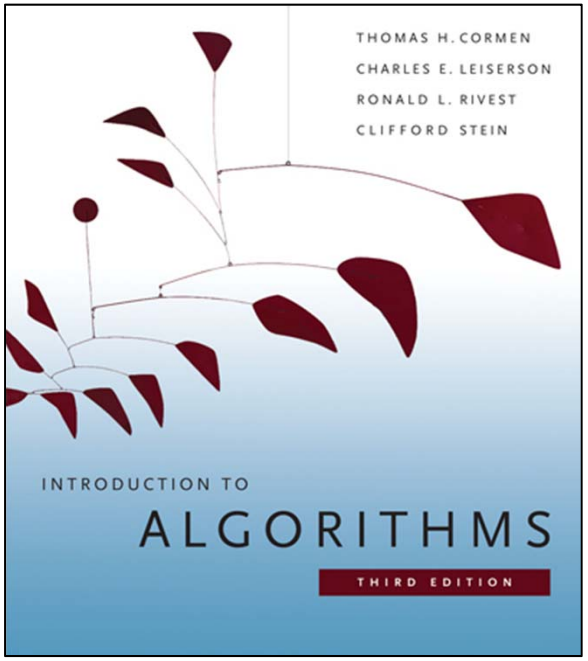
- c-Approximate Nearest Neighbor: build data structure which, for any query q
  - returns $p' \in P$, $||p-q|| \leq cr$,
  - where r is the distance to the nearest neighbor of q
- Can beat the curse
- Example algorithm:
  - Space: $O(dn+n^{1+1/c})$
  - Query time: $O(dn^{1/c})$



- 6.850 “Geometric Computing”
- Next Spring
- Coming to lecture room near you!

6.006

Introduction to Algorithms



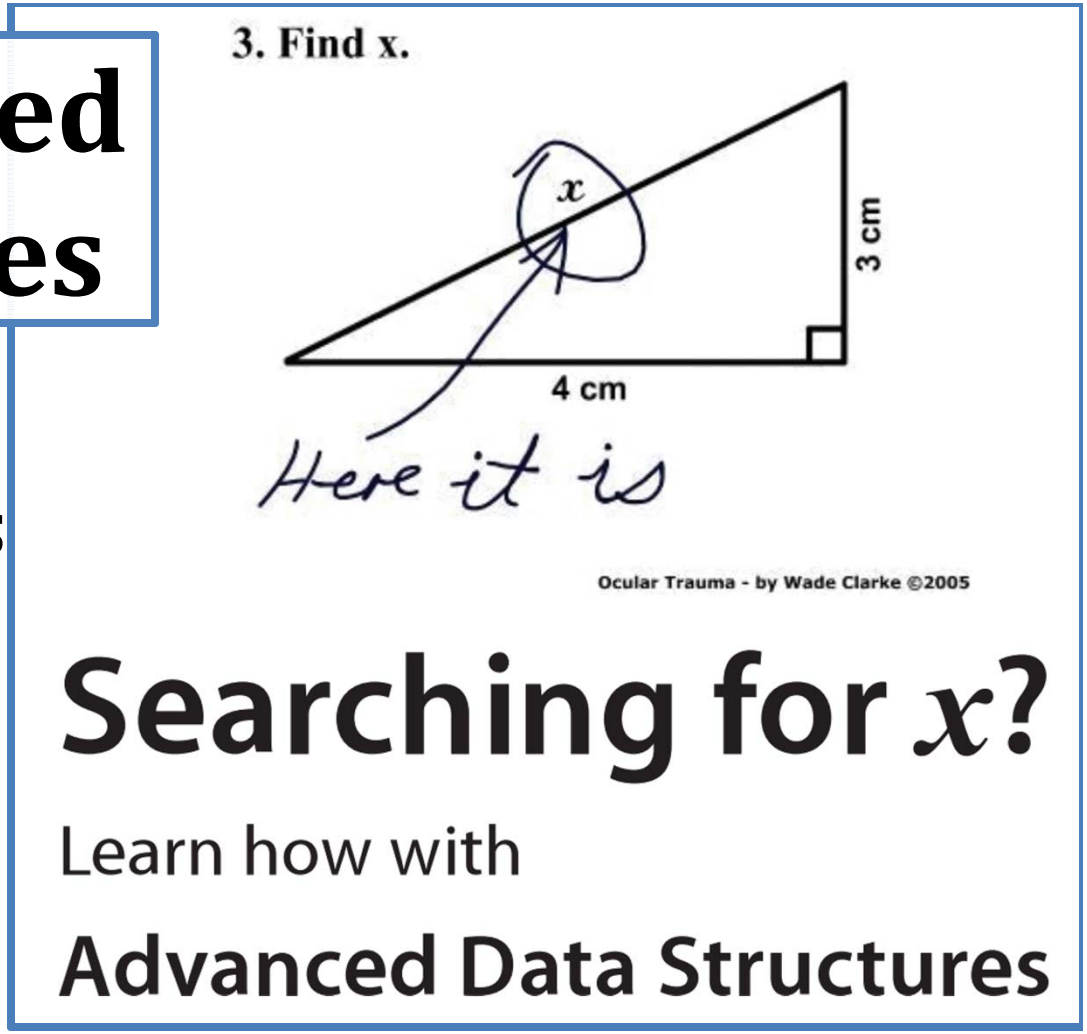
Lecture 26c: Beyond

Prof. Erik Demaine

Erik's Main Research Areas

- Data structures
- Graph algorithms
- Geometric folding algorithms
- Recreational algorithms

6.851: Advanced Data Structures



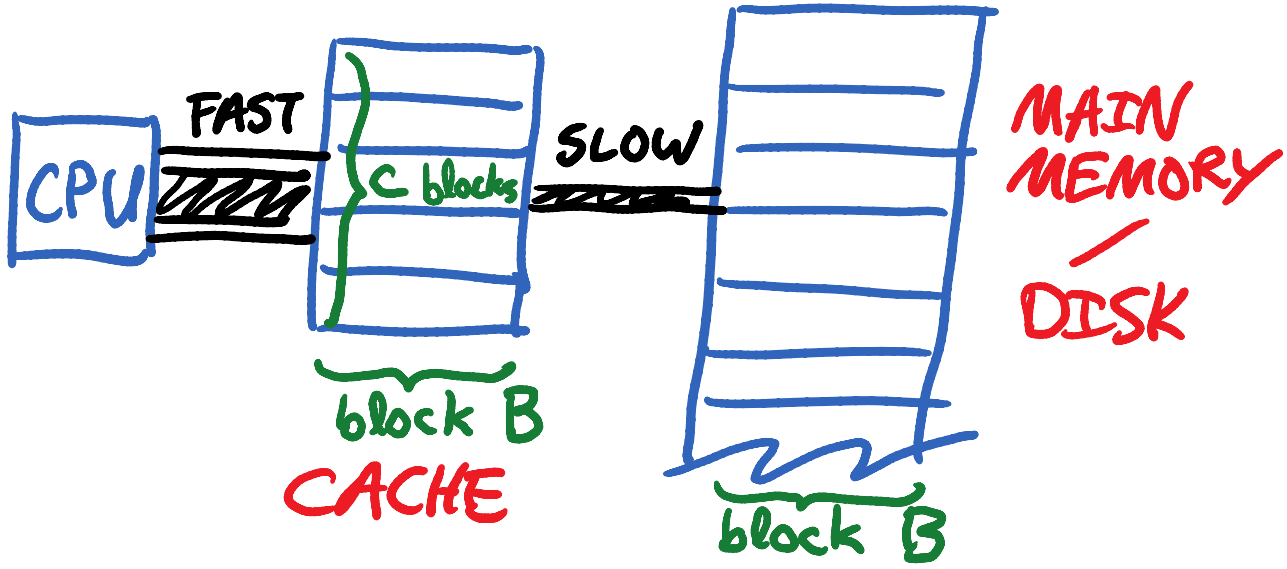
- Dynamic graphs
- Integer structures
- String structures
- Geometric DSs
- Reducing space
- Time travel
- Realistic models: cache, disk, GPUs, …
- Best possible binary search trees

Integer Data Structures

- Store n integers in range $\{0, 1, \ldots, u-1\}$ subject to insert, delete, successor, and predecessor in
  - $O(\lg \lg u)$ time *[van Emde Boas]*
  - $O\left(\frac{\lg n}{\lg \lg u}\right)$ time *[fusion trees]*
  - $O\left(\sqrt{\frac{\lg n}{\lg \lg n}}\right)$ time *[combination]*

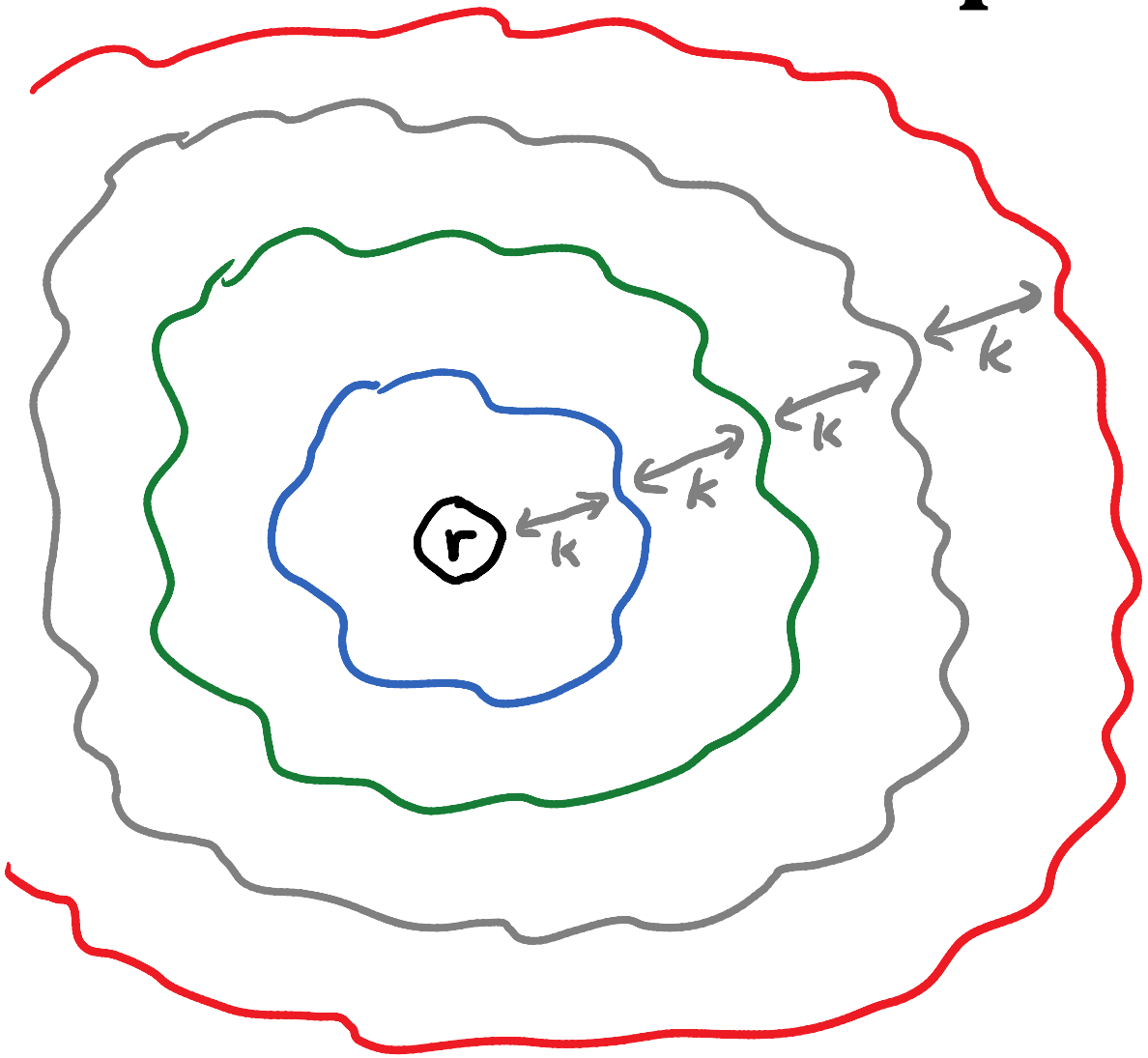
Cache-Efficient Data Structures

- Memory transfers happen in blocks:



- Searching takes $O(\log_B n)$ memory trasnfers
- Sorting takes $O\left(\frac{n}{B}\log_C \frac{n}{B}\right)$ memory transfers
- …even if you don't know B or C!

Approximation Algorithms in Planar Graphs (& more)



- Delete every kth BFS layer
- Rings are width-$O(k)$ "trees"
  - Most problems solved in $n^{O(k)}$
- Patch up at a factor $\approx 1 + 1/k$
- Approximate better as $k \to \infty$

Erik's Lecture 5 in 6....

courses.csail.mit.edu/6.849/fall10/lectures/L05.html

6.849: Geometric Folding Algorithms: Linkages, Origami, Polyhedra (Fall 2010)

Prof. Erik Demaine

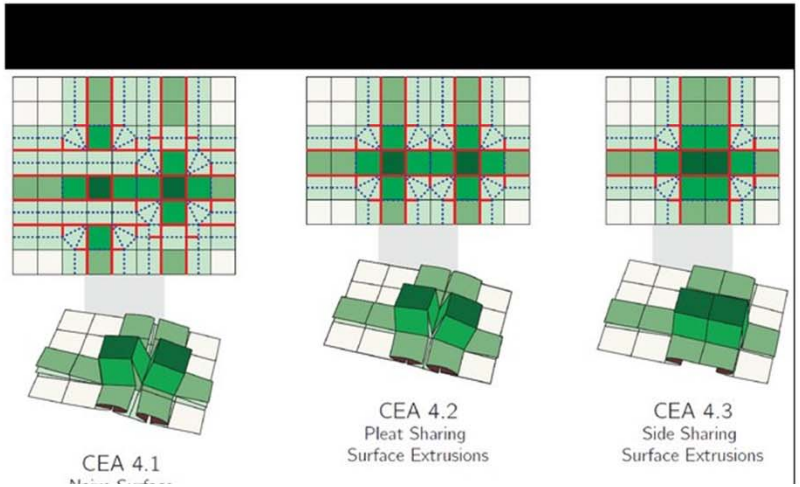
[Home] [Problem Sets] [Project] [Lectures] [Problem Session Notes]

Lecture 5 Video [previous]

[+] **Universal hinge patterns:** box pleating, polycubes; orthogonal maze folding. **NP-hardness:** introduction, reductions; simple foldability; crease pattern flat foldability; disk packing (for tree

Download Video: 360p

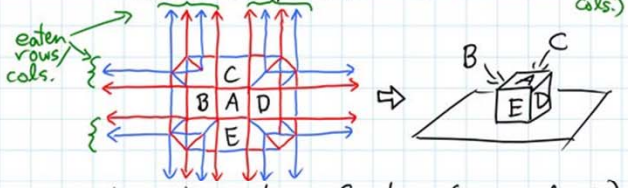
Slides, page 5/20 • [previous page] • [next page] • [PDF]



Handwritten notes, page 1/7 • [previous page] • [next page] • [PDF]

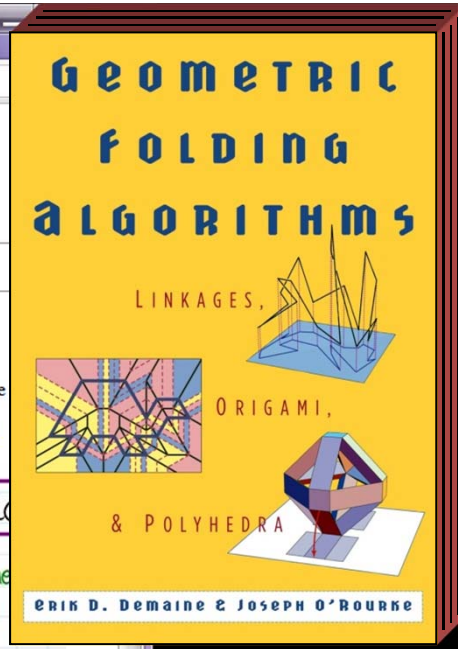
6.849 | Lecture 5 | Sept. 22, 201

Universal hinge patterns: (for origami transforme
[Benbernou, Demaine, Demaine, Oradya 2010]
- suppose crease pattern required to be subset of fixed "hinge pattern"
  (e.g. Origamizer uses completely different creases for every model)
- n×n box-pleat pattern can make any polycube of O(n) cubes, seamless:
  - cube gadget turns O(1) rows & columns into a cube sticking out of sheet ~ even if bumps elsewhere (not in eaten rows/cols.)



- to make a tree of cubes: (= any polycube)
  - make a leaf
  - conceptually remove it } "post order traversal"
  - repeat
- actually need to reserve space ahead of time for all the cube gadgets

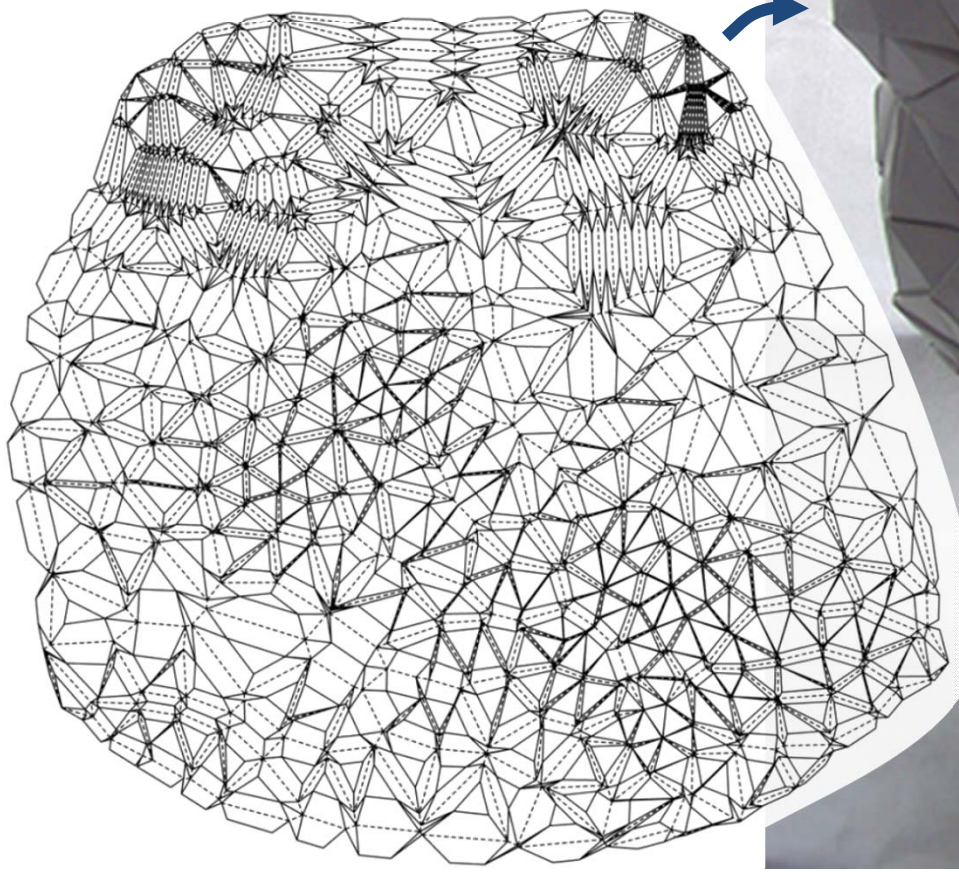
Handwritten notes, page 1/7 • [previous page] • [next page] • [PDF]

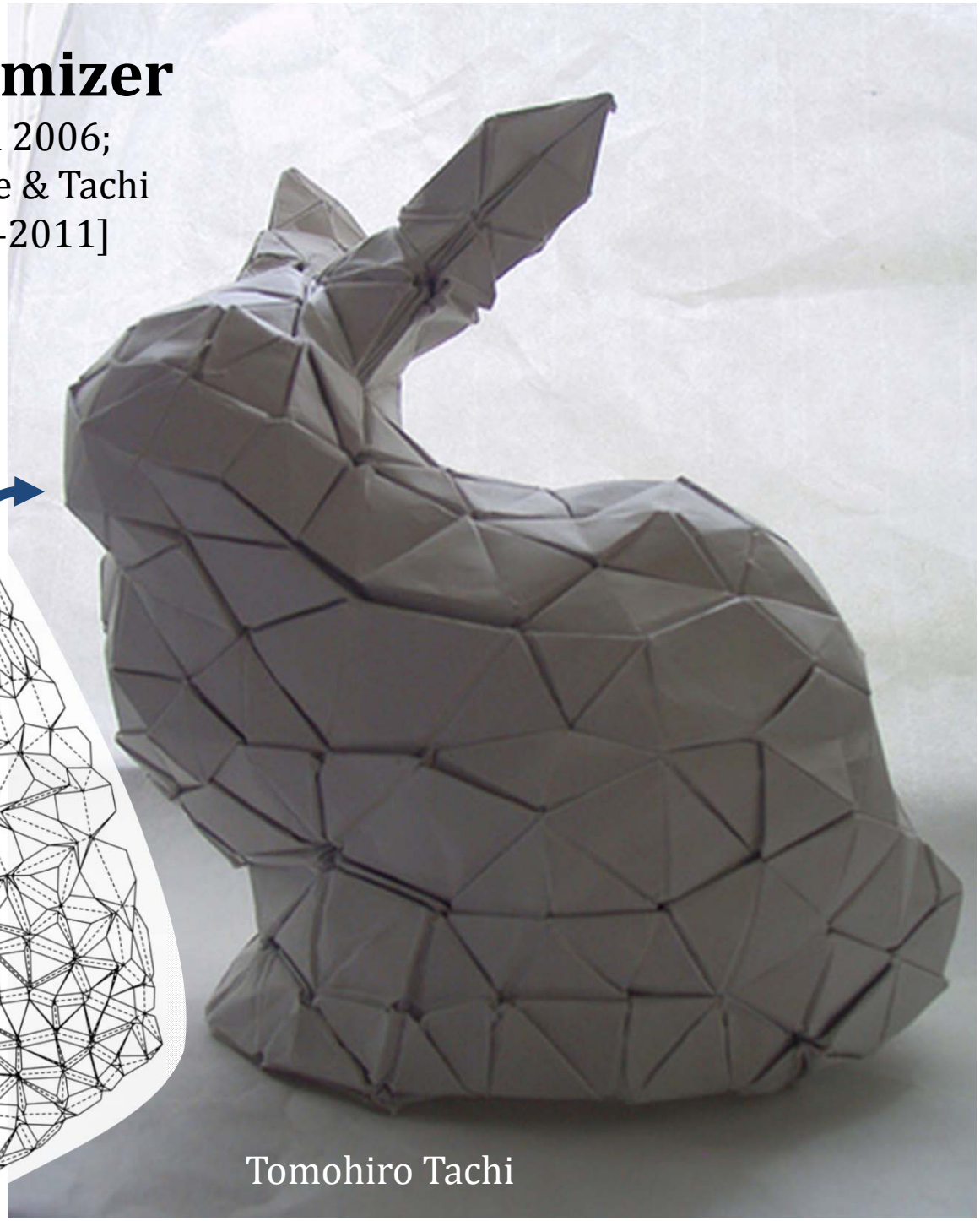



OFFICIAL SELECTION WISCONSIN FILM FESTIVAL 2009
OFFICIAL SELECTION ASHLAND 2009
OFFICIAL SELECTION BROOKLYN FILM FESTIVAL 2009
WINNER BEST DOCUMENTARY INDIE SPIRIT FILM FESTIVAL 2009
WINNER BEST DOCUMENTARY CINEMA SOCIETY OF SAN DIEGO 2009
OFFICIAL SELECTION TRANSILVANIA FILM FESTIVAL 2009
OFFICIAL SELECTION MAUI FILM FESTIVAL 2009
OFFICIAL SELECTION NEWPORT BEACH FILM FESTIVAL 2009
OFFICIAL SELECTION ASIAN FILM FESTIVAL 2009
OFFICIAL SELECTION CLEVELAND 2009
OFFICIAL SELECTION HAMPTONS 2008
WINNER AUDIENCE AWARD NEW HAMPSHIRE FILM FESTIVAL 2008
WINNER AUDIENCE AWARD RHODE ISLAND INTERNATIONAL FILM FESTIVAL 2008
OFFICIAL SELECTION SAVANNAH FILM FESTIVAL 2008
OFFICIAL SELECTION SANTE FE FILM FESTIVAL 2008
OFFICIAL SELECTION TEMECULA 2008
THE SCIENCE OF ART.
THE ART OF SCIENCE.
BETWEEN THE FOLDS
A NEW DOCUMENTARY FROM GREEN FUSE FILMS
[i]NDEPENDENT LENS
PBS

Origamizer

[Tachi 2006;
Demaine & Tachi
2009–2011]



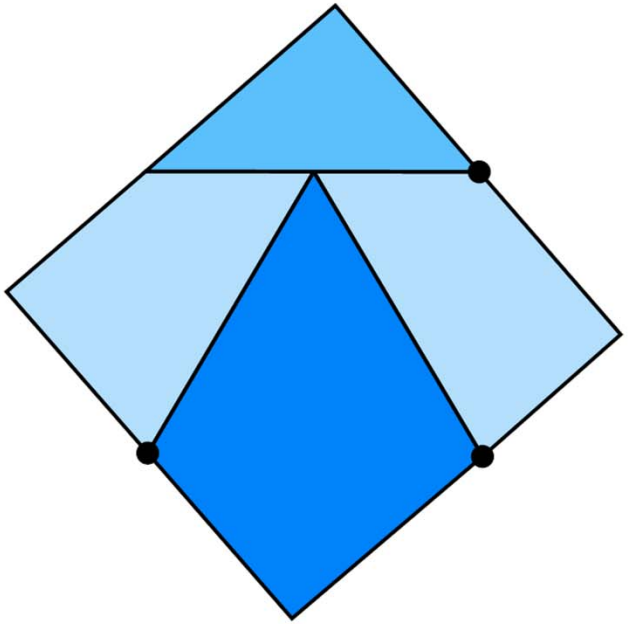


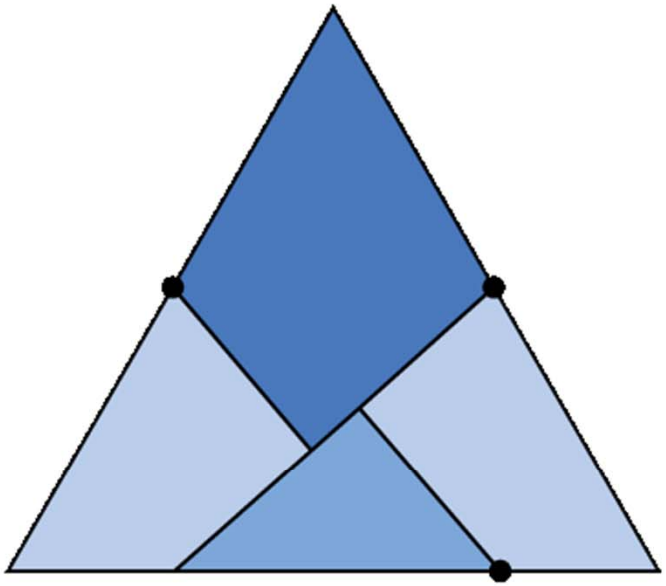
“Natural Cycles”

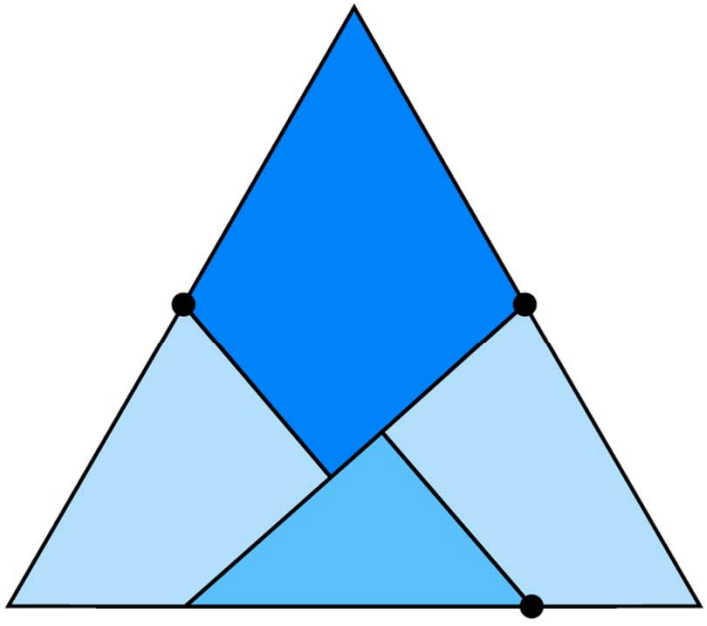
Erik & Martin Demaine

Renwick Gallery,
Smithsonian American
Art Museum, 2012

Hinged Dissection





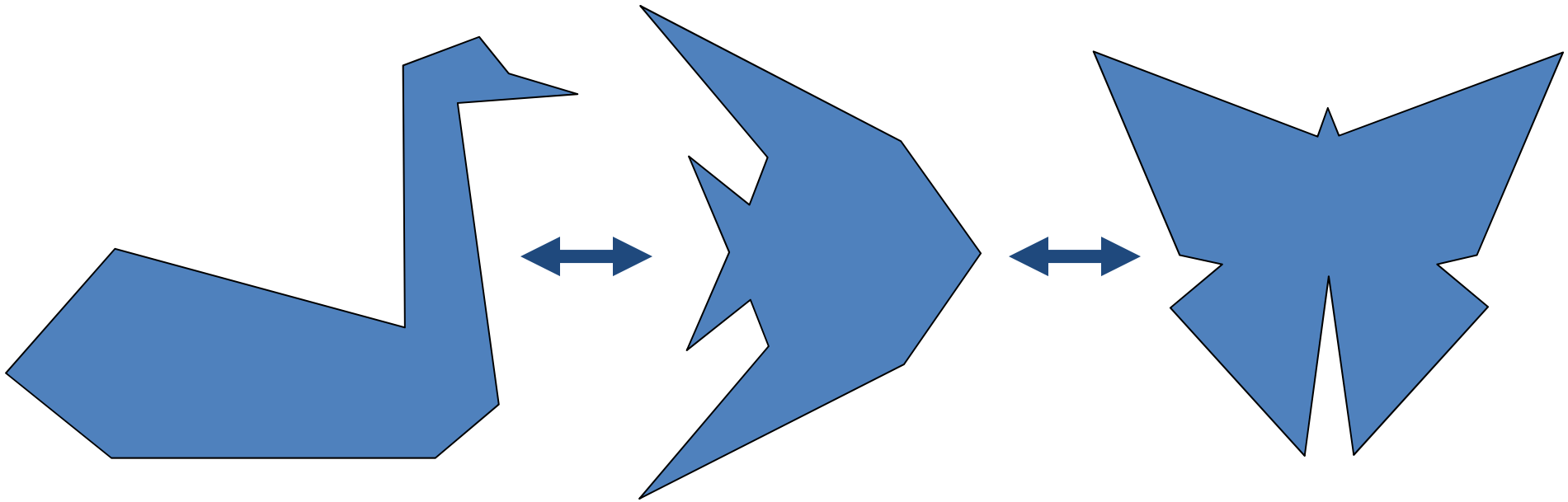


[Dudeney 1902]

Hinged Dissection Universality

[Abbott, Abel, Charlton, Demaine, Demaine, Kominers 2008]

- Theorem: For *any finite set* of polygons of equal area, there is a hinged dissection that can fold into any of the polygons, *continuously without self-intersection*



Complexity of Games & Puzzles

[Demaine, Hearn & many others]

0 players (simulation)

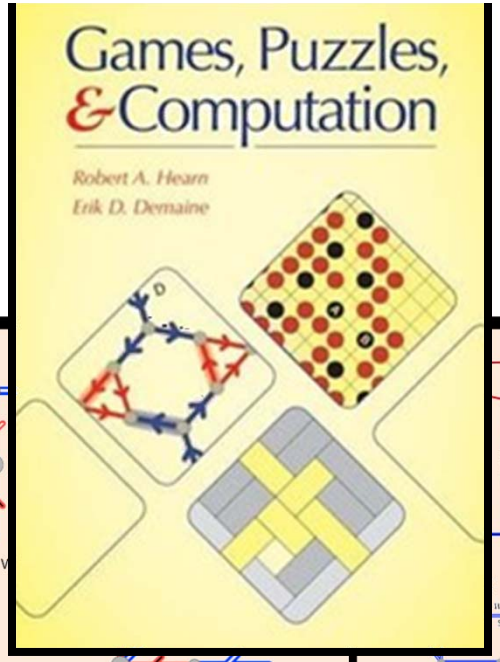
1 player (puzzle)

2 players (game)

team, imperfect info

Constraint Logic

[Hearn & Demaine 2009]



unbounded

bounded

PSPACE

PSPACE

EXPTIME

Undecidable

P

NP

PSPACE

NEXPTIME

0 players
(simulation)

1 player
(puzzle)

2 players
(game)

team,
imperfect info

Follow-On Algorithms Classes

- 6.046: Intermediate Algorithms
- 6.047: Computational Biology
- 6.854: Advanced Algorithms
- 6.849: Geometric Folding Algorithms
- 6.850: Geometric Computing
- 6.851: Advanced Data Structures
- 6.852: Distributed Algorithms
- 6.853: Algorithmic Game Theory
- 6.855: Network Optimization
- 6.856: Randomized Algorithms
- 6.857: Network and Computer Security

Follow-On Theory Classes

- 6.045: Automata, Computability, Complexity
- 6.840: Theory of Computing
- 6.841: Advanced Complexity Theory
- 6.842: Randomness & Computation
- 6.845: Quantum Complexity Theory