Welcome to

Foundations of Computational and Systems Biology

7.36/20.390/6.802 (undergrad version)

and

7.91/20.490/6.874/HST.506 (grad version)

taught by

Profs Chris Burge, Ernest Fraenkel, David Gifford

and TAs

with guest lectures by

George Church (Harvard), Doug Lauffenburger, Ron Weiss

with video recording by MIT's OCW

7.91/20.490 and 6.874/HST.506
are graduate-level survey courses in computational biology

target audience: graduate students with solid biology background and comfort with quantitative approaches

7.36/20.390 and 6.802
are upper-level undergraduate survey courses in computational biology

target audience: upper-level undergraduates with solid biology background and comfort with quantitative approaches

goal:
to develop understanding of foundational methods in computational biology, to enable students to contextualize and understand the basis of a good portion of the research literature in this growing field, *and gain exposure to research in this field**.

**graduate versions only*

This course is **not**

a systems biology class

- but topics important for analyzing complex systems will be presented

a synthetic biology class

- but there will be a guest lecture related to synthetic biology

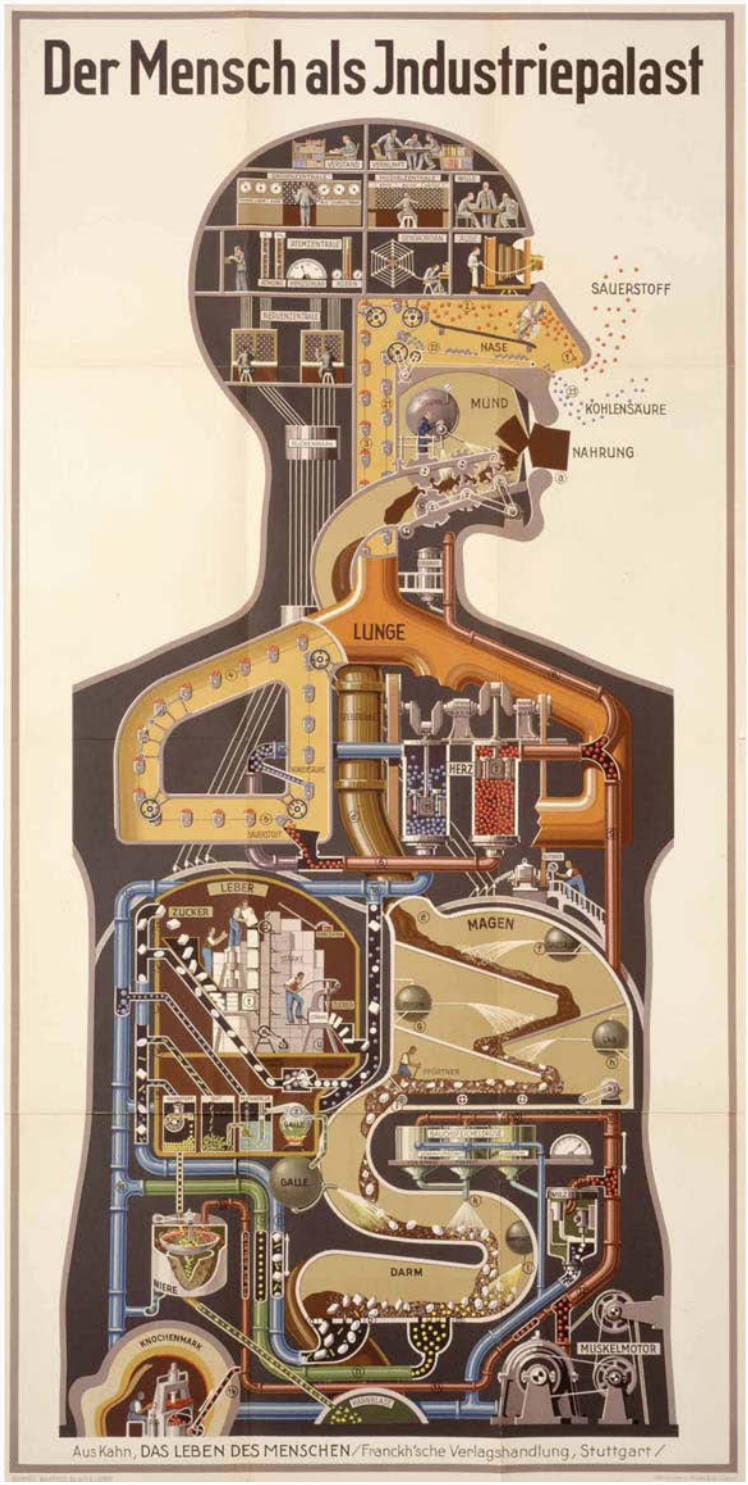
an algorithms class*

- we do not assume prior expertise in designing or analyzing algorithms
- the essential ideas behind a variety of algorithms *will* be discussed, but we will not usually address details of implementation
- however, you will have an opportunity to implement at least one bioinformatics algorithm as part of homeworks

*Except 6.874, which includes additional AI content

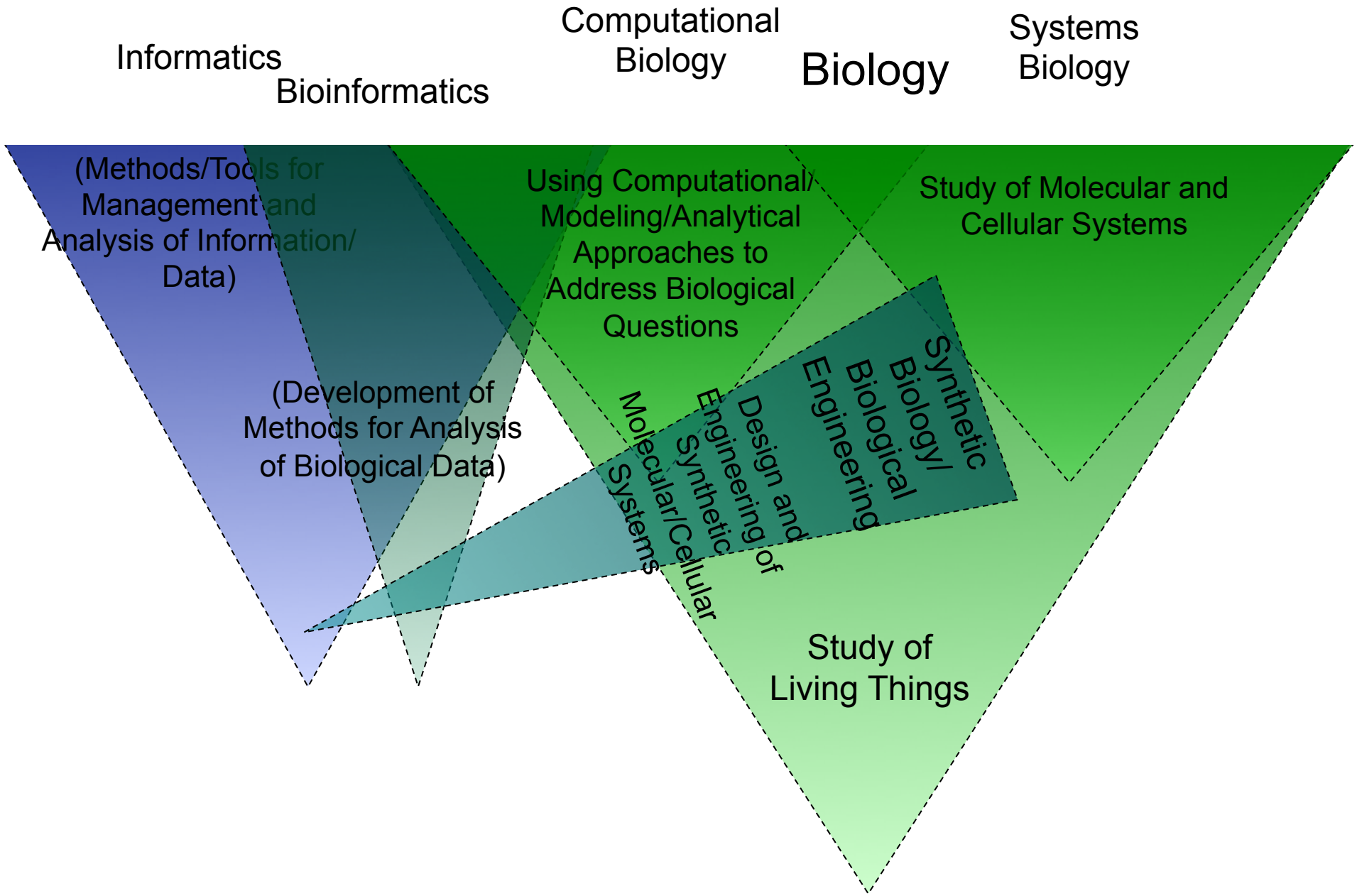
Plan for Today

- History of Computational and Systems Biology
- Review of Course
  - Mechanics
  - Organization
  - Content



Brief, Anecdotal History of Computational and Systems Biology

Overlapping Fields



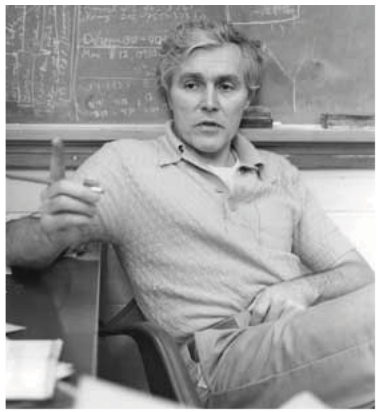
The 1970s and Earlier - Sequence Databases, Similarity Matrices and Molecular Evolution

Margaret Dayhoff

How do protein sequences evolve?

How should similarity between two proteins be scored to most accurately detect homology?

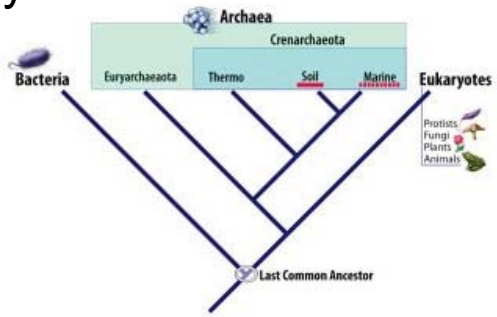
- First protein sequence databases / protein family classification
- PAM matrices for protein sequence comparisons (still used!)

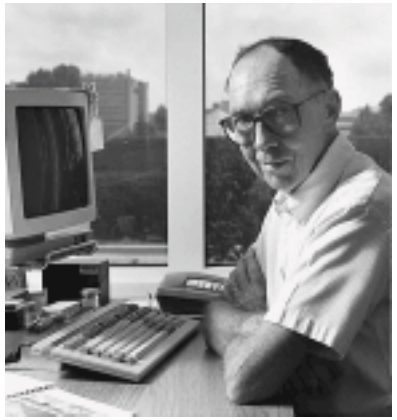


Carl Woese

What can molecular sequences tell us about organismal evolution?

- Molecular classification of life
- Molecular clocks
- Use of ribosomal RNA to infer phylogeny
- Discovery of third 'domain' of life - Archaea





Russ Doolittle

The 1980s: Sequence Alignment/Search

Which specific residues/positions in a pair of proteins are homologous?

- Smith-Waterman alignment algorithm

Photographs of scientists removed due to copyright restrictions.

What RNA secondary structure has minimum folding free energy?

- Nussinov algorithm
- Zuker algorithm

How to rapidly and reliably find homologs to a query sequence in a sequence database?

- FastA and BLAST algorithms and associated statistics

Photographs of scientists removed due to copyright restrictions.

Al Gore Learns to Search PubMed

NCBI Director David Lipman (far left) coaches Vice President Gore (seated) as he searches PubMed.
NIH Director Harold Varmus (center) and NLM Director Donald Lindberg (far right) look on.

Photograph by the National Center for Biotechnology Information; in the public domain.

The ‘90s: HMMs, Ab Initio Protein Structure Prediction, Genomics, Comparative Genomics

How to identify domains in a protein?

How to identify genes in a genome?

Hidden Markov Models as a framework for such problems

How to study gene expression globally, infer gene function from expression?

- Microarrays and clustering

How to predict protein function by comparing genomes?

- gene fusions, phylogenetic profiling, etc.

How to predict protein structure directly from primary sequence?

- Rosetta algorithm

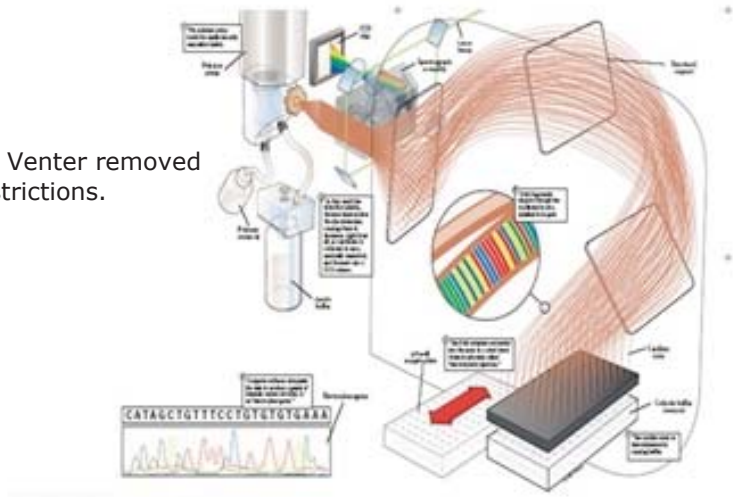
The 2000s Part 1:

The human genome is sequenced, assembled, annotated

genomics becomes fashionable

Photograph of genome-project pioneers removed due to copyright restrictions. See the photograph on nature.com.

Photograph of Craig Venter removed due to copyright restrictions.



Photographs of Jim Kent removed due to copyright restrictions.

Photographs of Ewan Birney removed due to copyright restrictions.

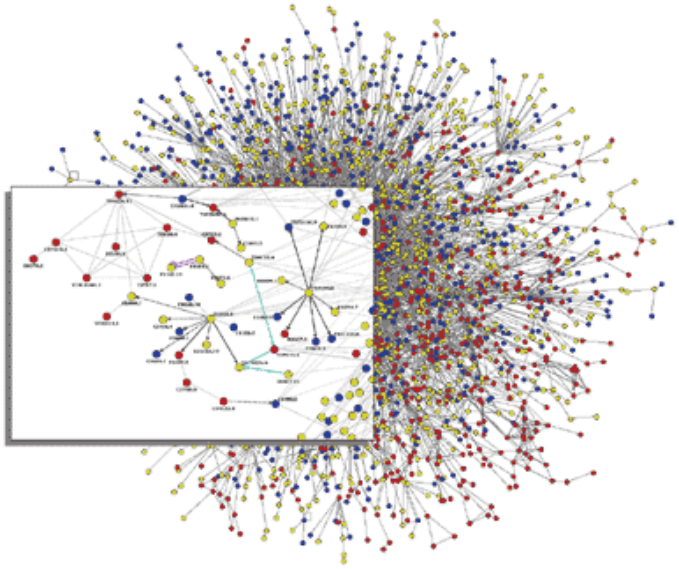
The 2000s Part 2:
Biological Experiments Become High-Throughput, Computational Biology Becomes more Biological

Massively parallel data collection - transcriptomics, proteomics, interactomics, metagenomics

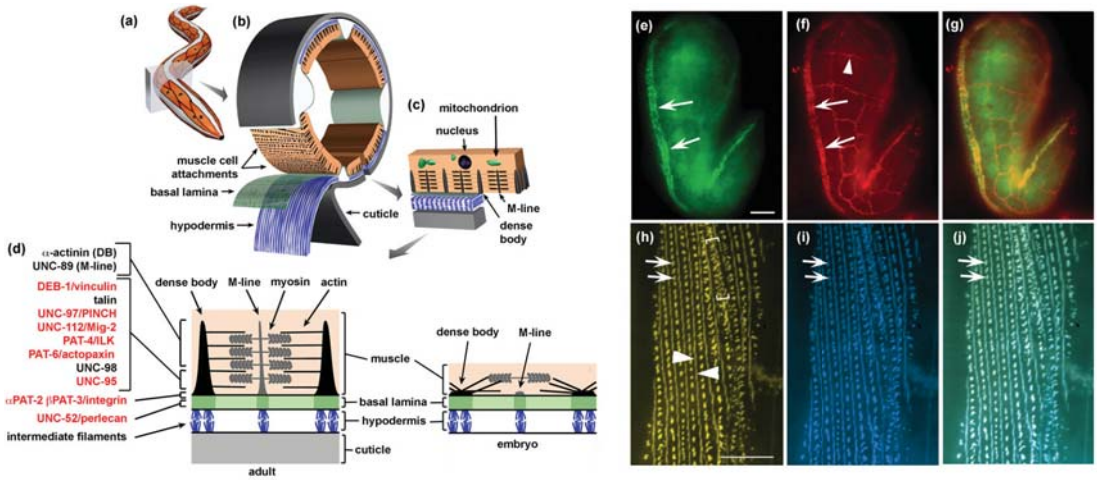
Using sequence and array data to address fundamental questions about transcription, splicing, microRNAs, translation, epigenetics, protein structure/function, development, evolution, disease, etc.

Integrated computational/experimental approaches

Rise of bioimage informatics

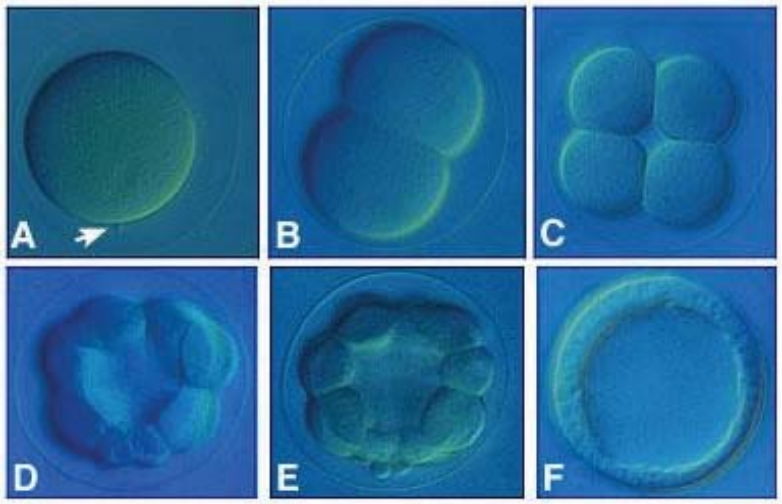


Courtesy of Marc Vidal. Used with permission.



Courtesy of Donald G. Moerman and Benjamin D. Williams. License: CC-BY.
Source: Moerman, D. G. and Williams, B. D. "Sarcomere Assembly in C. Elegans Muscle" (January 16, 2006), WormBook, ed. The C. elegans Research Community, WormBook.

Photograph of Eric Davidson removed due to copyright restrictions.



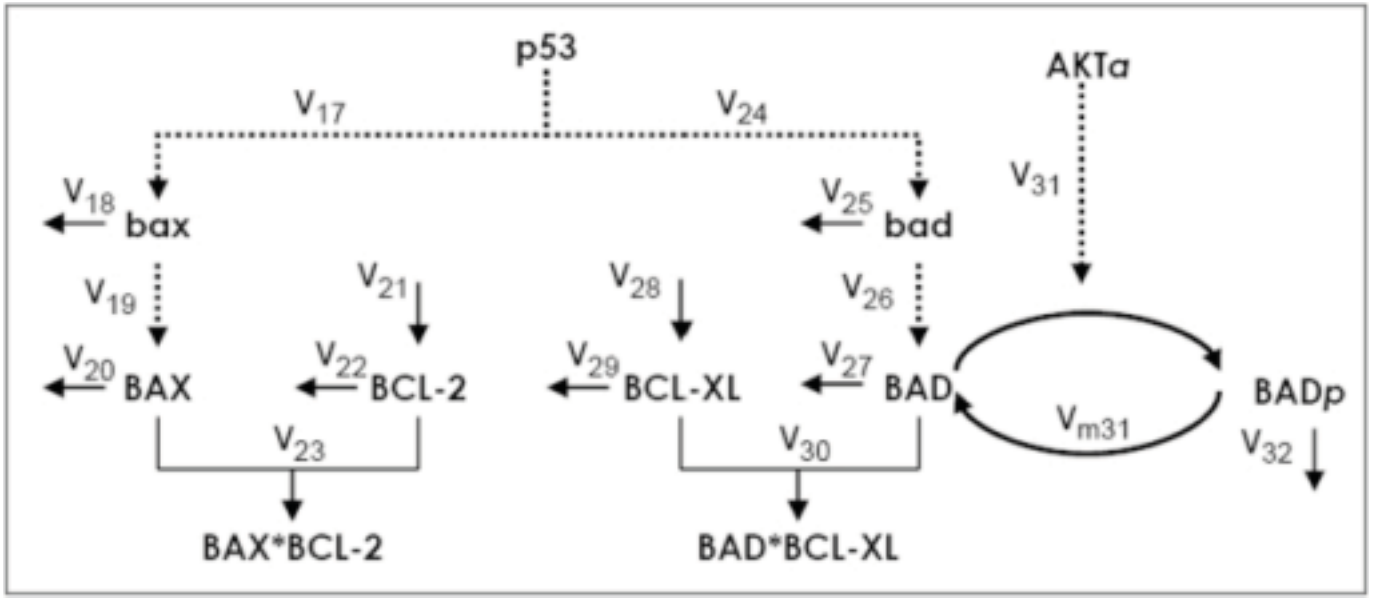
Courtesy of Charles Ettensohn. Used with permission.

Computational model of the gene regulatory netowrk controlling sea-urchin embryonic development removed due to copyright restrictions. See the image here.

The 2000s Part 3

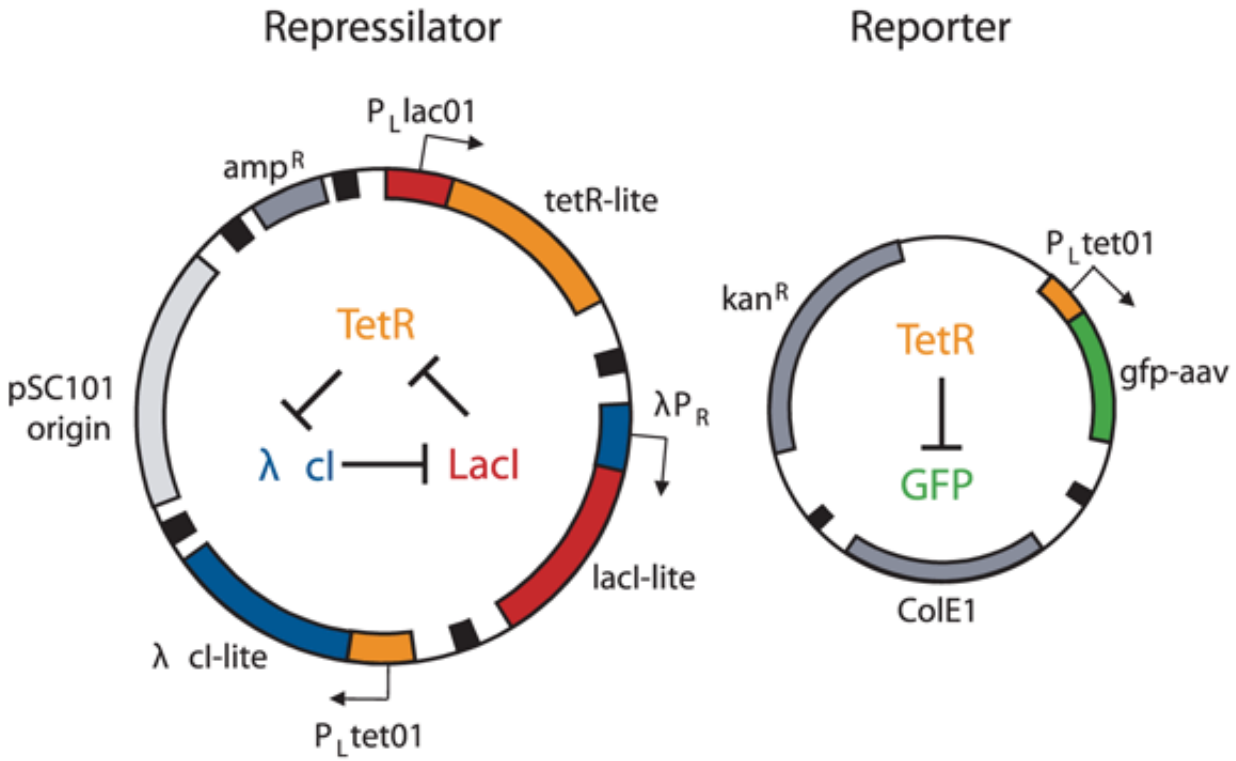
Systems Biology

Models of gene and protein networks in development, disease, etc.



The 2000s Part 4:
Synthetic Biology & Biological Engineering

Design of regulatory networks using biological components



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Elowitz, Michael B., and Stanislas Leibler. "A Synthetic Oscillatory Network of Transcriptional Regulators." *Nature* 403, no. 6767 (2000): 33S-8.

$$\frac{d[A]}{dt} = \frac{V_a}{1+\frac{[B]}{K_{iB}}} - k_a[A]$$

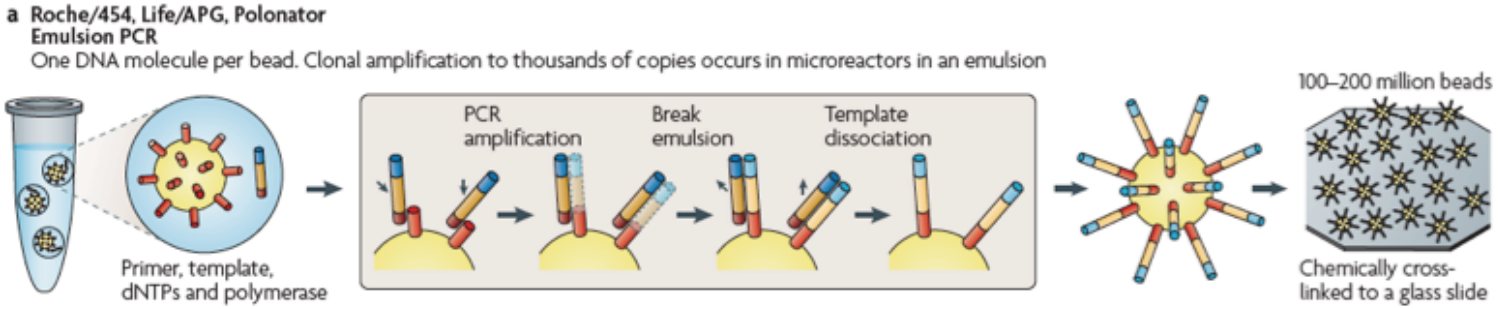
$$\frac{d[B]}{dt} = \frac{V_b}{1+\frac{[C]}{K_{iC}}} - k_b[B]$$

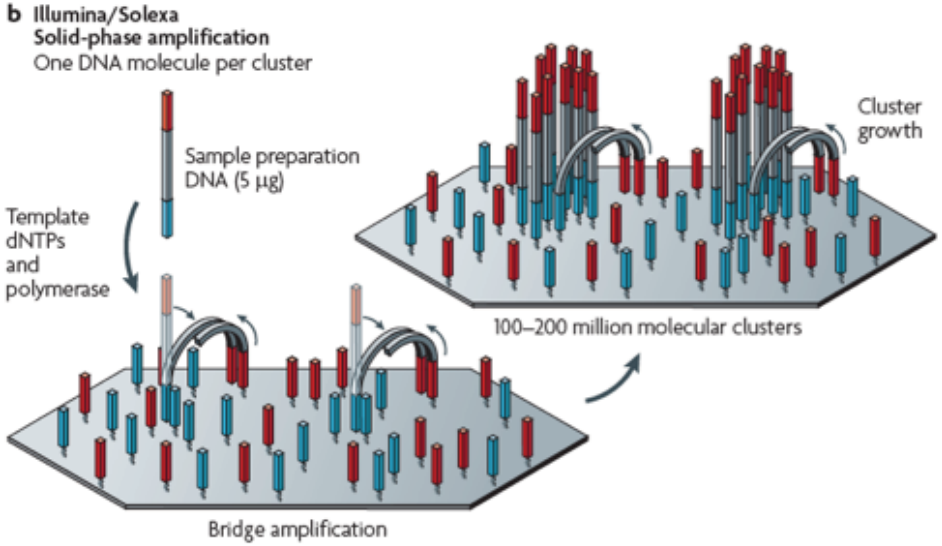
$$\frac{d[C]}{dt} = \frac{V_c}{1+\frac{[E]}{K_{iE}}} - k_c[C]$$

$$\frac{d[D]}{dt} = \frac{V_d}{1+\frac{K_{aB}}{[B]}} - k_d[D]$$

$$\frac{d[E]}{dt} = \frac{V_e}{\left(1+\frac{K_{aD}}{[D]}\right)\left(1+\frac{[C]}{K_{iC'}}\right)} - k_e[E]$$

Late 2000s / Early 2010s





Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Metzker, Michael L. "Sequencing Technologies-the Next Generation. " *Nature Reviews Genetics* 11, no. 1 (2009): 31-46.

Metzker NRG 2010

Next-gen sequencing finds applications across biology

Genome sequencing

Transcriptome sequencing

Protein-DNA intrxns (ChIP-seq)

Protein-RNA intrxns (CLIP-seq)

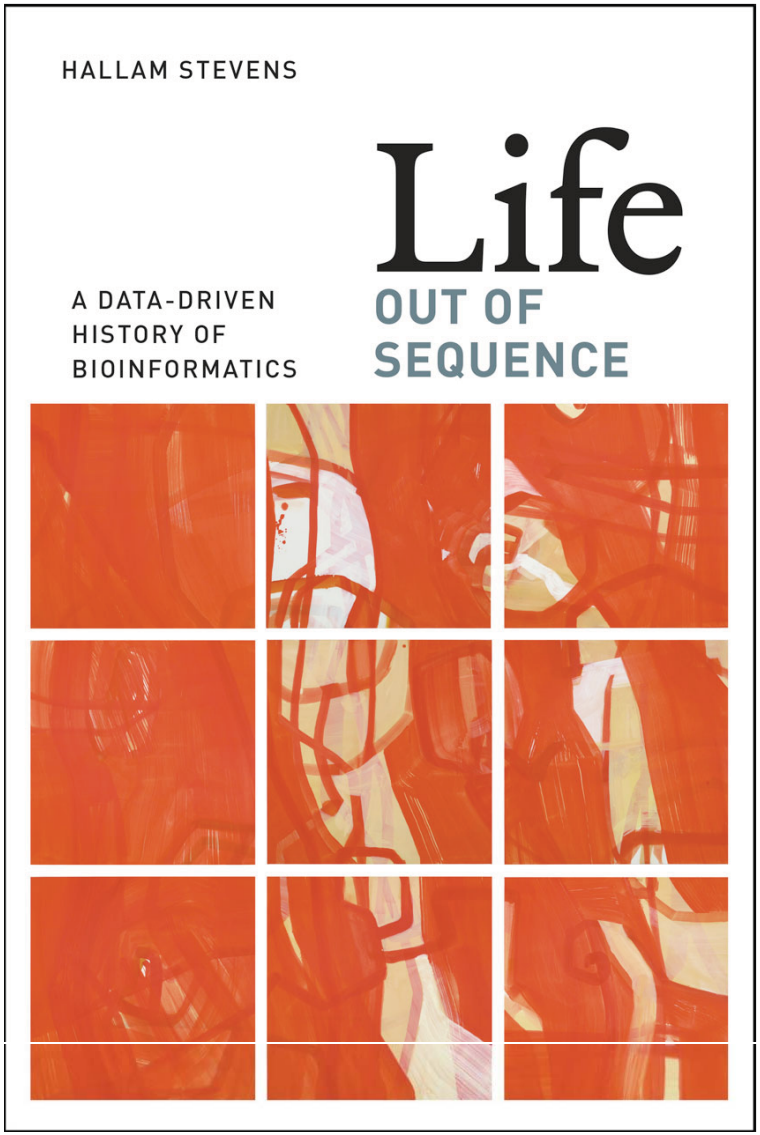
Translatome

Methylome

Open chromatome

Photograph of Barbara Wold removed due to copyright restrictions.

For those who would like a proper history of the field



Photograph of Hallam Stevens removed due to copyright restrictions.

Source: Stevens, Hallam." Life Out of Sequence: A Data-Driven History of Bioinformatics ." *University of Chicago Press*, 2013.

Online Access

All course materials, including copies of lecture slides, will be distributed via course website:

Auditors/Listeners

The class may be audited only by permission of one of the instructors. Please meet us after class and tell us who you are and why you want to audit. Students are strongly encouraged to take the course for credit.

A look at the syllabus

Topics:

Genomic Analysis I (CB)

Genomic Analysis II – 2nd Gen Sequencing (DG)

Modeling Biological Function (CB)

Proteomics (EF)

Regulatory Networks (EF, DG + DL)

Computational Genetics (DG)

Guest lectures: Doug Lauffenburger (signaling networks), Ron Weiss (synth bio),

George Church (genetics/genomics)

Topics will include a discussion of motivating questions, experimental methods and the interplay between experiment and computation

Motivating Questions

What instructions are encoded in our (and other) genomes?

How are chromosomes organized?

What genes are present?

What regulatory circuitry is encoded?

Can the transcriptome be predicted from the genome?

Can the proteome be predicted from the transcriptome?

Can protein function be predicted from sequence?

Can evolutionary history be reconstructed from sequence?

Motivating Questions

What would you need to measure if you wanted to discover

- the causes of disease?
- the mechanisms of existing drugs?
- the metabolic pathways in a micro-organism?

What kind of modeling would help you use these data to

- design new therapies?
- re-engineer organisms for new purposes?

What can we currently measure?

What does each type of data mean individually?

How do we integrate the data to understand the system?

Course Schedule, Part I

Assignments	Wk	Date	Lecturer	Topic Area	Lec	Topic
ALL DUE @ NOON	1	Tue, Feb 04	all		L1	Course Introduction: History of Computational Biology, Overview of Course, Course Policies and Mechanics, DNA Sequencing Technologies
		Thu, Feb 06	CB	Genomic Analysis	L2	Local Alignment (BLAST) and Statistics
		Fri, Feb 07			R1	Recitation 1
Interests Due	2	Tue, Feb 11	CB		L3	Global Alignment of Protein Sequences (NW, SW, PAM, BLOSUM)
		Thu, Feb 13	CB		L4	Comparative Genomic Analysis of Gene Regulation
		Fri, Feb 14			R2	Recitation 2
	3	Tue, Feb 18	no class			No Class - President's Day
PS1 Due		Thu, Feb 20	DG	Genomic Analysis - Next Gen Sequencing	L5	Library complexity and BWT
		Fri, Feb 21			R3	Recitation 3
Teams Due	4	Tue, Feb 25	DG		L6	Genome assembly
		Thu, Feb 27	DG		L7	ChIP-Seq analysis (DNA-protein interactions)
		Fri, Feb 28			R4	Recitation 4
	5	Tue, Mar 04	DG		L8	RNA-seq analysis (expression, isoforms)
		Thu, Mar 06	CB	Modeling biological function	L9	Modeling & Discovery of Sequence Motifs (Gibbs sampler, alternatives)
Aims Due		Fri, Mar 07			R5	Recitation 5
	6	Tue, Mar 11	CB		L10	Markov & Hidden Markov Models of Genomic and Protein Features
Ps2 Due		Thu, Mar 13	CB		L11	RNA Secondary Structure - Biological Functions & Prediction
		Fri, Mar 14			R6	Recitation 6
	7	Tue, Mar 18	EXAM			

Course Schedule, Part II

Assignments	Wk	Date	Lecturer	Topic Area	Lec	Topic
		Thu, Mar 20	EF	Proteomics	L12	Introduction to Protein Structure; Structure Comparison & Classification
Research Strategy		Fri, Mar 21			R7	Recitation 7
	8	Tue, Mar 25	no class			No Class - Spring Break
		Thu, Mar 27	no class			
	9	Tue, Apr 01	EF		L13	Predicting protein structure
PS3 Due		Thu, Apr 03	EF		L14	Predicting protein interactions
		Fri, Apr 04			R8	Recitation 8
	10	Tue, Apr 08	EF	Regulatory Networks	L15	Gene Regulatory Networks
		Thu, Apr 10	EF		L16	Protein Interaction Networks
		Fri, Apr 11			R9	Recitation 9
	11	Tue, Apr 15	EF		L17	Computable Network Models
PS4 Due		Thu, Apr 17	DG		L18	Chromatin and DNase-seq analysis
		Fri, Apr 18			R10	Recitation 10
Written Report	12	Tue, Apr 22	no class			No class - Patriots Day
		Thu, Apr 24	DG	Computational Genetics	L19	eQTL discovery and analysis
		Fri, Apr 25			R11	Recitation 11
	13	Tue, Apr 29	DG		L20	Human genetics, SNPs, and GWAS
PS5 Due		Thu, May 01	Guest		L21	
		Fri, May 02			R12	Recitation 12
	14	Tue, May 06	EXAM			
		Thu, May 08	Guest		L22	
		Fri, May 09			R13	Recitation 13
	15	Tue, May 13				Presentations
		Thu, May 15				Presentations

Is this the right course for me?

Other alternatives are available, some more specialized:

7.57 Quantitative Biology for Graduate Students (only for bio grad students)
7.81 Systems Biology (Gore)
6.581/20.482 Foundations of Algorithms and Computational Techniques in Systems Biology (Tidor, White)
6.047/6.878 Computational Biology: Genomes, Networks, Evolution (Kellis)
6.502/6.582/HST.949 Molecular Simulations (Stultz)
6.877/HST.949 Computational Evolutionary Biology (Berwick)
18.417 Introduction to Computational Molecular Biology (Waldispuhl)
18.418 Topics in Computational Molecular Biology (Berger)
10.555J Bioinformatics: Principles, Methods and Applications (Stephanopoulos, Rigoutsos)

Text Book

The following text is recommended (not required) for this course is available through Amazon and at the COOP and will be on reserve at Hayden library

Understanding Bioinformatics, Zvelebil & Baum (Garland Science)

This text contains helpful background information for some of the lectures and relevant chapters or sections will be mentioned from time to time. However, we have also selected the following texts as particularly useful in selected areas, if you are looking for further information. …

Basic Probability and Statistics

A primer covering basic concepts in probability and statistics that are useful for this class is available at the class web site. The workbook is designed for students to complete at their own pace, and/or to use as a reference. It includes explanatory material and many examples drawn from biology. Students who have less background in these areas or need a refresher are strongly encouraged to read the primer and do the examples. Your TAs will be able to answer questions about this material throughout the semester.

Homework

Five written or computer-based homework assignments will be posted on the course web site. These are designed to promote deeper understanding of the principles and algorithms discussed in class and to provide hands-on experience with bioinformatics tools.

Your score for the homework portion of the course will be based on a maximum of 100 points, but the total points available on the homeworks will be 120. This means that you can miss one homework (or a portion of one homework) and still do fairly well on this component if you have done well on the other homeworks. For example, a student who obtained perfect marks on 4 of 5 homeworks, each valued at 24 points would get 96 points for the homework component of the course, almost as good as a student who completed all 5 homeworks, earning 90% of points on each, since 0.9 x 120 = 108, which would earn the maximum score of 100. Because of this, no make-up assignments will be offered. Of course, it is still to your advantage to do all five homeworks, as this will help you to learn the material in more depth, help prepare you for exams, etc. Please note that the point values of individual homeworks may vary somewhat from the 24 point average value, depending on their length and level of difficulty.

The dates that the assignments are due are included in the syllabus below. Homework must be turned in to the appropriate box outside of the Biology education office (68-120), or submitted on-line, by the assigned time to be eligible for full credit.

Late assignments

Assignments are due at noon on the dates indicated on the syllabus. Assignments received electronically, or in the appropriate box outside the Biology Education Office (68-120), within 24h of the time they are due will be eligible for 50% credit. If necessary, you may turn in your written portion and your programming portion separately. For example, if you turn in your written portion on time and your programming by the late due date, your written work will be eligible for full points but the programming will be eligible for only 50% of points. You may not further sub-divide your submissions. Because answer keys will be posted, no homework will be accepted after the extended deadline.

Collaboration on Problem Sets

The goal of the problem sets is to reinforce the material and sometimes to explore a topic in greater depth. You may talk with other students about the problems and work on them together. However, you should write up your own solutions. Copying someone else's solutions will not improve your understanding of the material and is not acceptable. Duplicate or nearly identical homeworks from different students will receive a score of zero. This has happened. We notice. Don't let it happen to you!

Collaboration on Programming in Problem Sets

You must write your own code on problem sets. You may discuss the programming problems with other students. The following two simple rules should make it clear what is not permitted:
do not copy or reuse code from any source (except the sample code provided)
do not share your code with anyone else in the class.

Intellectual Honesty

We hope and trust that academic misconduct will not occur during this course. We nevertheless want to emphasize that we will be rigorous in our enforcement of Institute rules. It is the policy of the Biology Department to keep a record of all cases of academic misconduct and to forward cases to the Dean of Undergraduate and Student Affairs.

Recitations

Three recitation sections will be offered each week. Times will be determined on the first day of class. These will be led by the TAs and will provide students an opportunity to ask questions about material presented in lecture, the readings or the homeworks. Recitations are required for 6.874, optional for other versions. However, if you are having difficulty in the class, you are *strongly* encouraged to attend regularly.

Python Instruction

The homework assignments will include problems that involve writing programs in the scripting language Python. Python is widely used for bioinformatics and computational biology. Programming will not be taught in lecture, but because some students may have little or no programming experience, hands-on tutorials in Python will be offered by the teaching assistants during the first and second weeks of classes. Also, you may ask questions related to programming assignments at weekly recitations. Be sure to attend recitation if you are struggling with the programming assignments.

There will be an “Intro to Python” session, targeting those with little or no previous programming experience.

Please bring a laptop if you have one.

- Notes outlining the materials covered in these sessions, as well as short exercises designed to help you get up to speed (NOT for credit), are posted on course website.
- You are encouraged to look at these materials, especially "Starting Python Programming” before attending the introductory sessions.

Project Component (7.91/20.490/6.874/HST.506 only)

Students registered for one of the graduate versions of this course will also complete a computational biology research project during the semester.

There are 6 student assignments related to project component:
1) Submit background/interests for posting to course website
2) Choose Teams and submit Project Title and 1 Paragraph Summary
3) Submit Specific Aims (1 page)
4) Submit Research Strategy (2 pages)
5) Submit Final Written report (~5 pages)
6) Oral presentations
Due dates for these assignments are listed on the course syllabus.

This Project component of this course is designed to give you practice in applying computational methods to contemporary problems in biology. Students design and carry out projects working in a group (maximum: 5 students, unless approved by instructor) or by themselves. We suggest that you choose people to work with who have skills complementary to yours.

Course	Project	AI problems
7.36/20.390/6.802	NO	NO
7.91/20.490/HST.506	YES	NO
6.874	YES	YES

Exams/Grading/Honesty

Exams

There will be two 80-minute exams (non cumulative). Exam dates are noted on the calendar below. Students are expected to take the exams at the scheduled times. There is no final exam.

Grading

Undergraduate versions of course (6.802, 7.36, 20.390):

Homework (out of max 100 points):	36%
Exams	62%
Peer Review	2%

Graduate Bio/BE/HST versions of course (7.91, 20.490, HST.506):

Homework (out of max 100 points):	30%
Exams	48%
Project	20%
Peer Review	2%

Graduate EECS version of course (6.874):

Homework (out of max 100 points):	25%
Exams	48%
Project	20%
Extra AI-related problems	5%
Peer Review	2%

An additional 1% extra credit may be awarded for exceptional class participation.

Topic 1 - Announcements

PSet

• PSet1 will be posted tonight. **Due: Thurs Feb 20 @ noon**
involves basic molecular biology, probability/statistics

• Pset2 will be posted soon. Look at the programming problem to give you a feeling for the level of programming that will be needed.

Probability/Stats

• For review of basic probability/statistics concepts related to the course, see

Statistics Primer (posted)

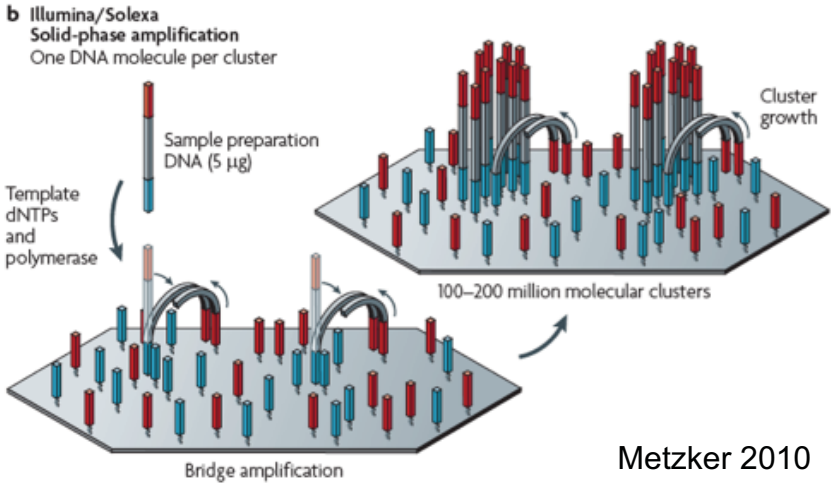
Sequencing Technologies

• Read Metzker review (posted)

Reading on Sequence Alignment / Statistics

• Over the next several days, read Chapters 4, 5 of Z&B for background on sequence alignment

Genomic Analysis I

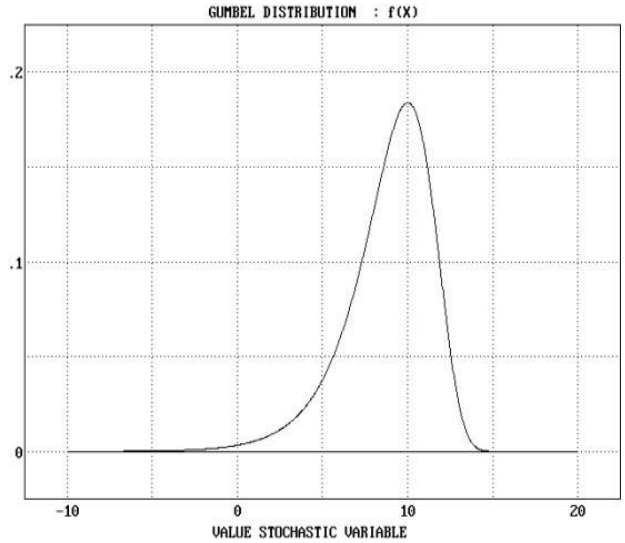


Metzker 2010

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Metzker, Michael L. "Sequencing Technologies-the Next Generation." *Nature Reviews Genetics* 11, no. 1 (2009): 31-46.

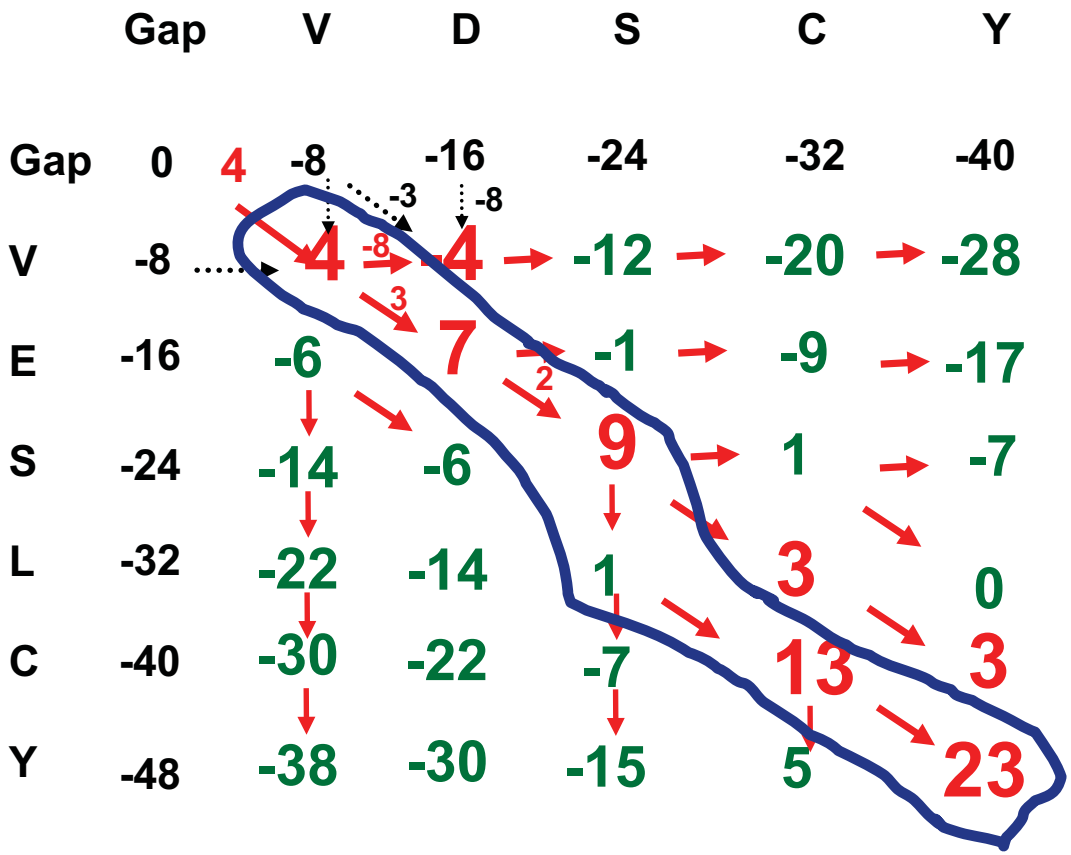
Sequencing Technologies (L1/L2)

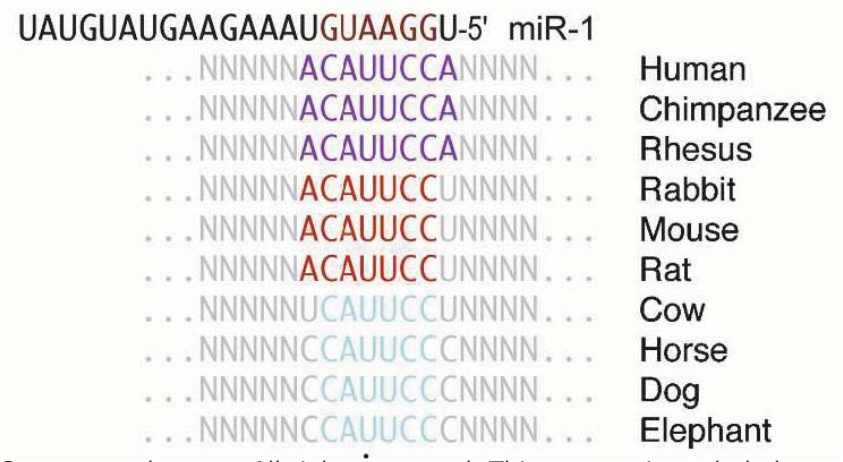
Sequence search/statistics with BLAST (L2)
(local ungapped sequence alignment)

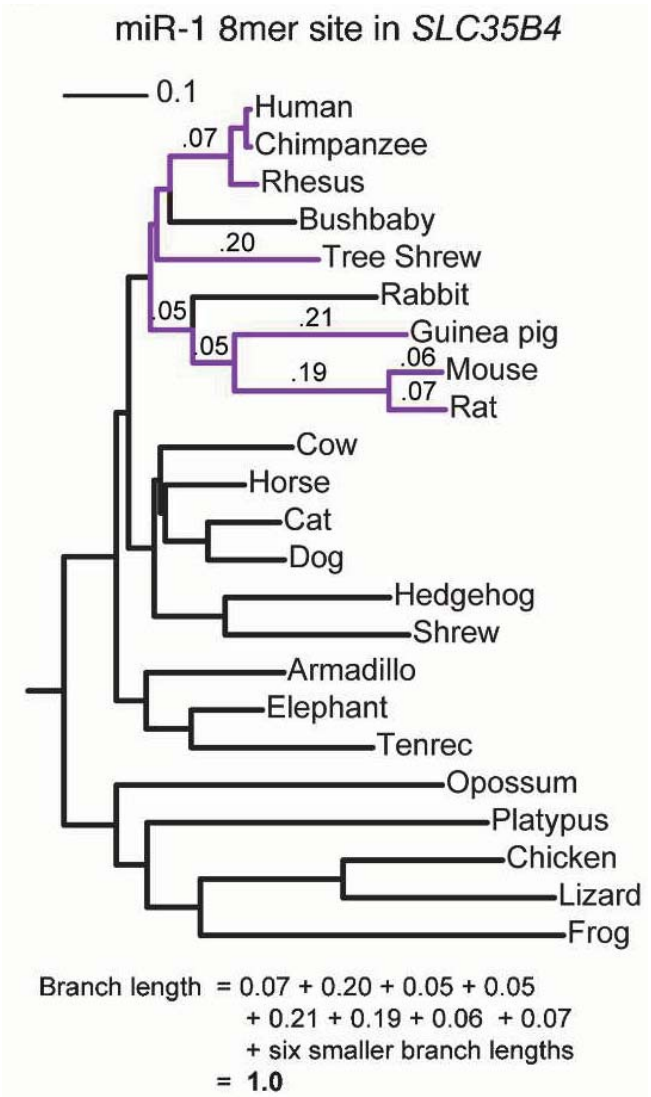


$$P(S > x) = 1 - \exp[-KMN\, e^{-\lambda x}]$$

Global Sequence Alignment (L3)





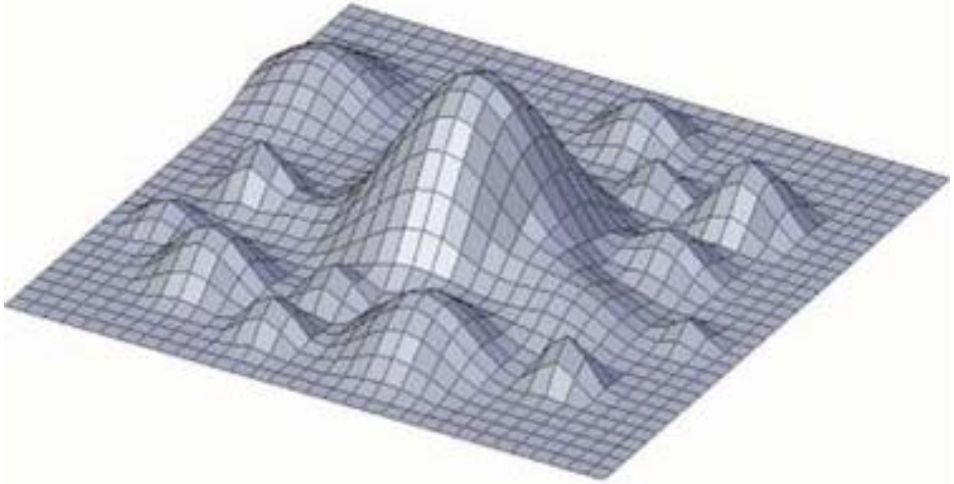


Source: Figure 1B of Friedman, Robin C., Kyle Kai-How Farh, et al. "Most Mammalian mRNAs are Conserved Targets of MicroRNAs." *Genome Research* 19, no. 1 (2009): 92-10S.

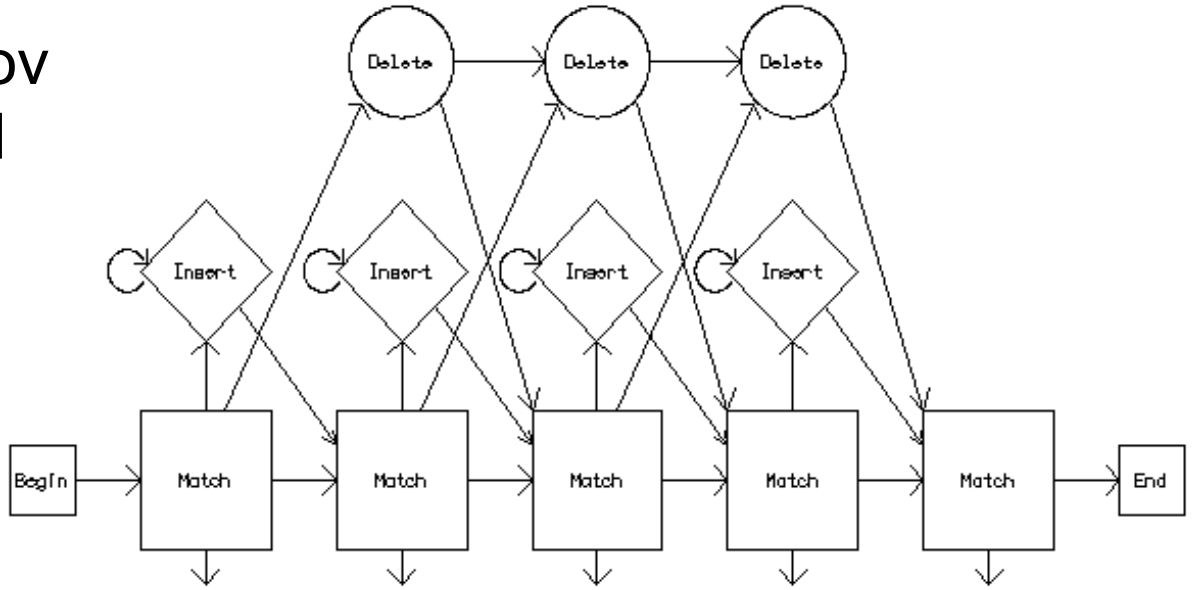
Comparative Genomic Analysis of Gene Regulation (L4)

Modeling Biological Function

Modeling & Discovery of Sequence Motifs (L9)

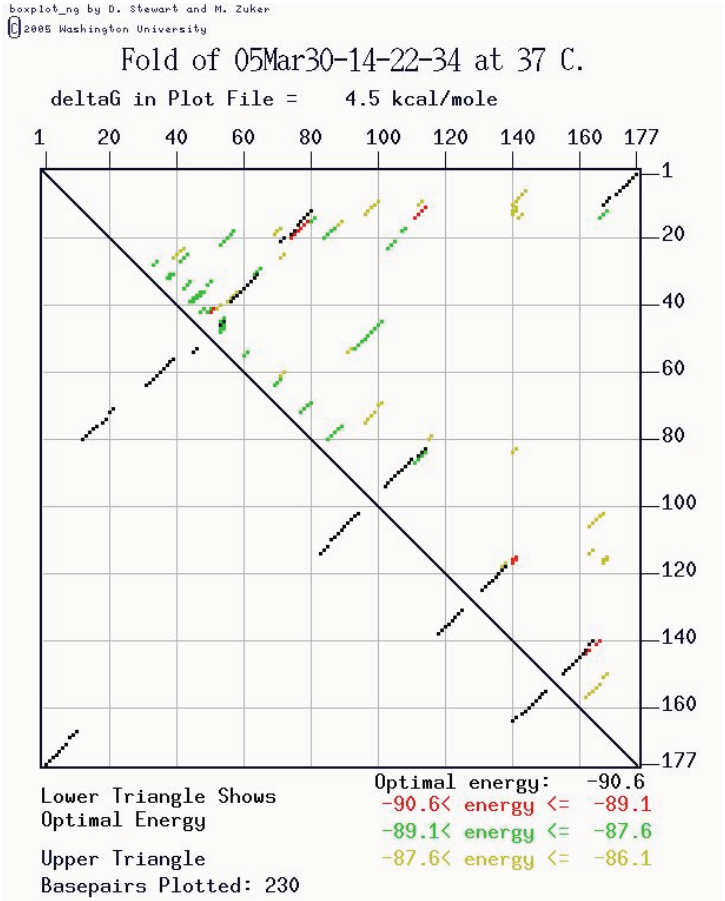


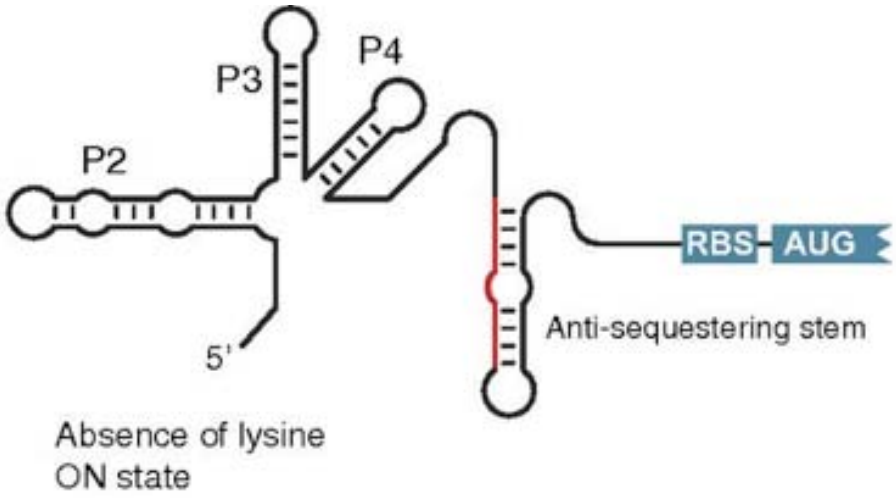
Markov & Hidden Markov Models of Genomic and Protein Features (L10)

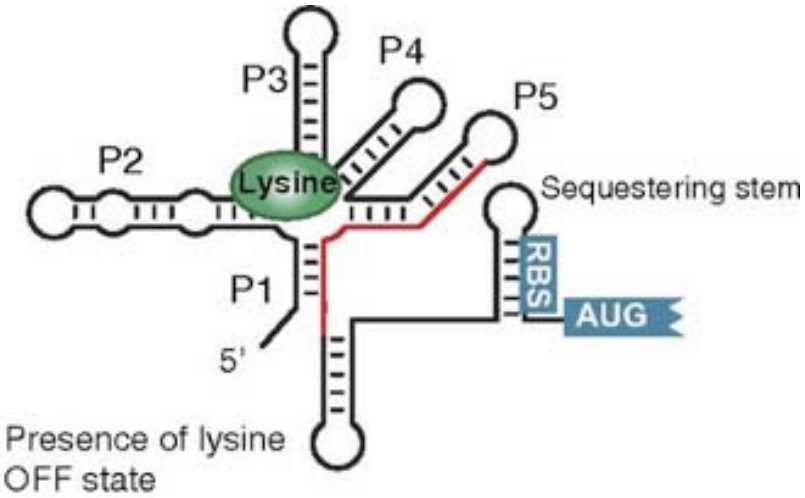


Source: Hashad, Attalah, Khalaed Kamal, et al. "Improving Virus C type 4 Interferon using Bioinformatics Techniques." (PDF) *Journal of Emerging Trends in Computing and Information Sciences,* no. 6 (2012): 88S-9S.

RNA Secondary Structure - Biological Functions & Prediction (L11)



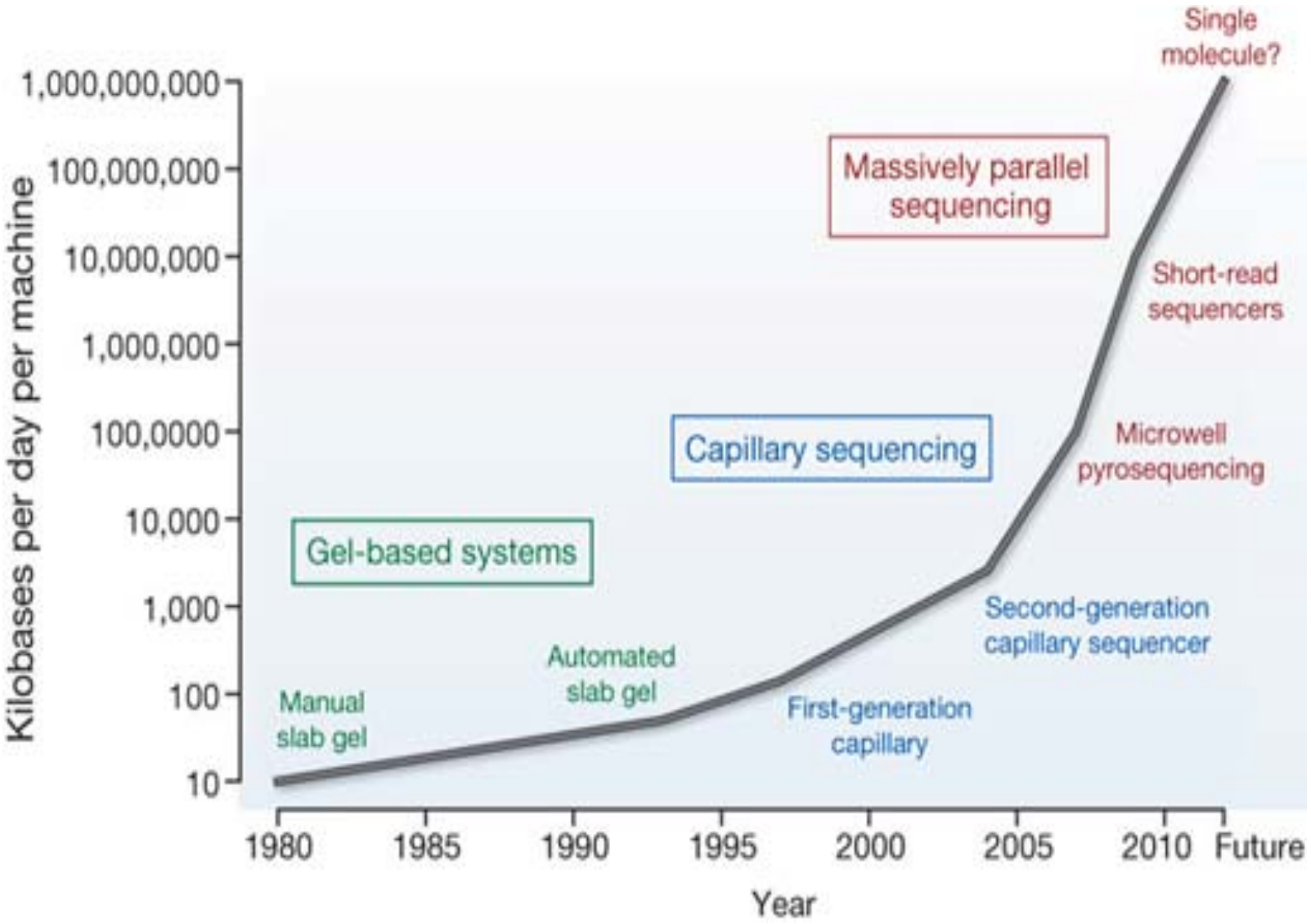




Genomic Analysis Module

Next Generation Sequencing

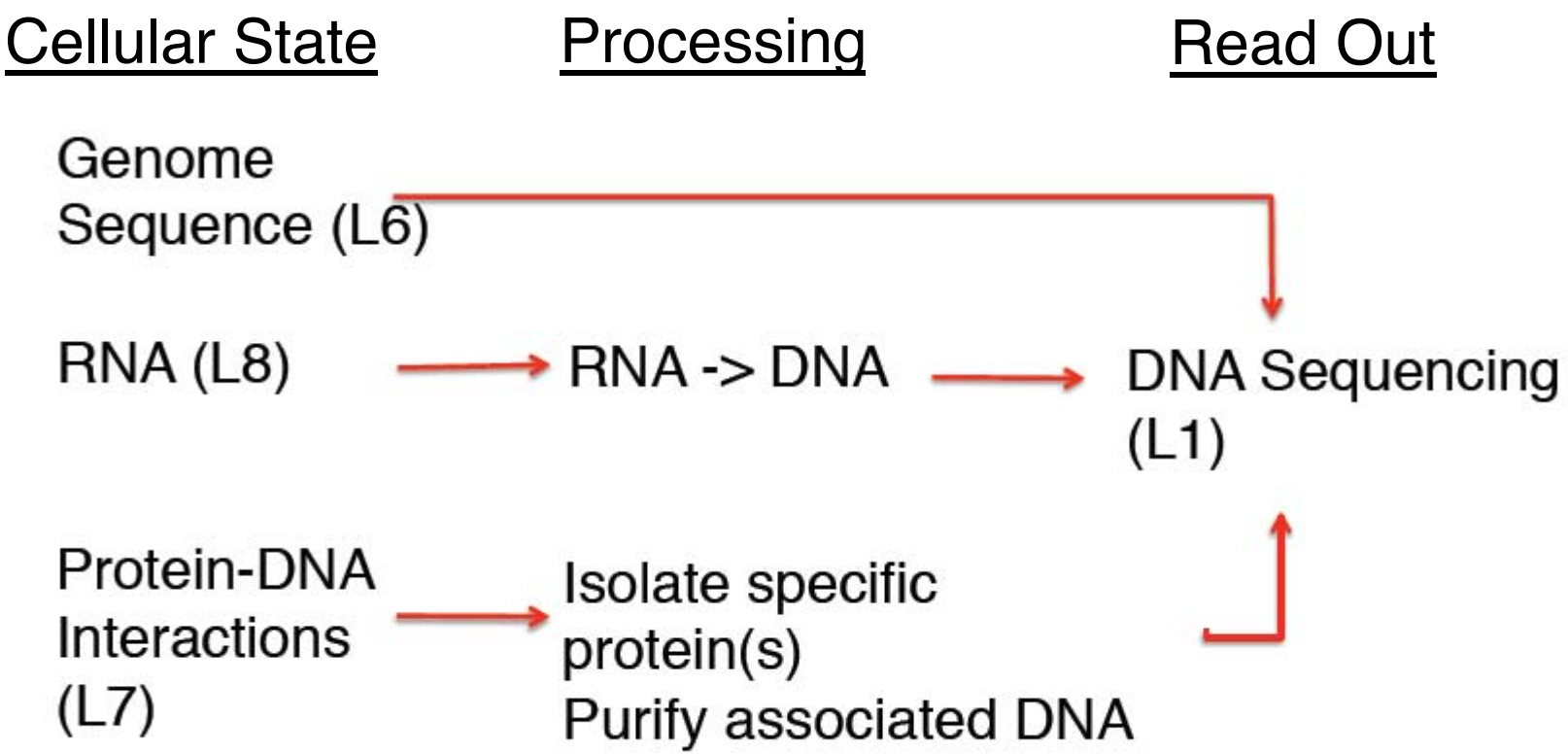
DNA Sequencing Technology is improving more than exponentially



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Stratton, Michael R., Peter J. Campbell, et al. "The Cancer Genome." *Nature* 4S8, no. 7239 (2009): 719-24.

http://genome.wellcome.ac.uk/doc_WTX059576.html

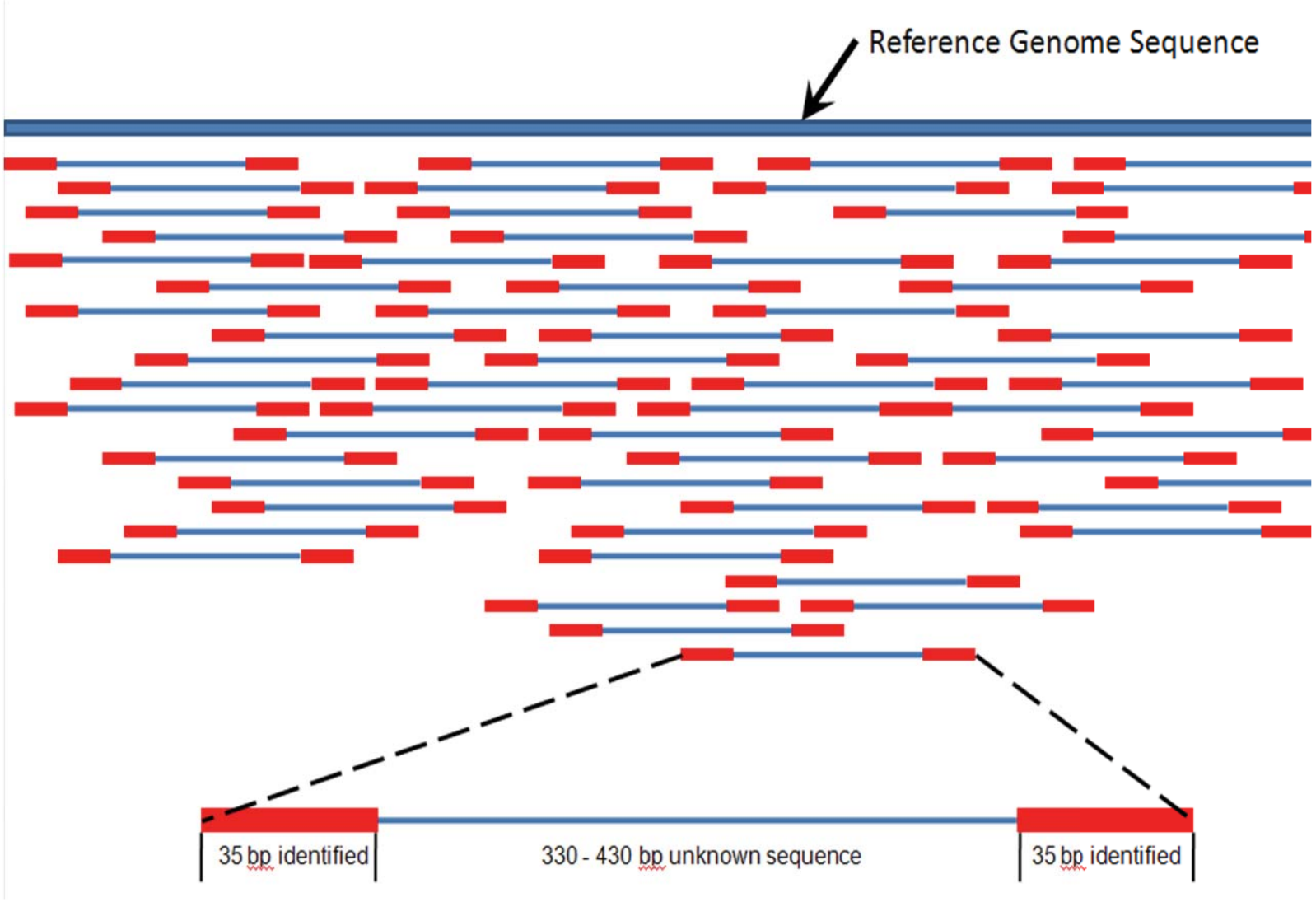
Idea – Use DNA sequencing to measure diverse biological state information



Genomic Analysis Module
Next Generation Sequencing

- L5 – How to index and process millions of DNA sequence reads using the Burrows-Wheeler Transform (BWT) to build genome indexes
- L6 – How to assemble a reference genome sequence from 10^8 short sequence reads
- L7 – How to discovery where regulatory proteins occupy the genome (ChIP-seq analysis)
- L8 – How to measure RNA expression and isoforms using high-throughput DNA sequencing (RNA-seq analysis)

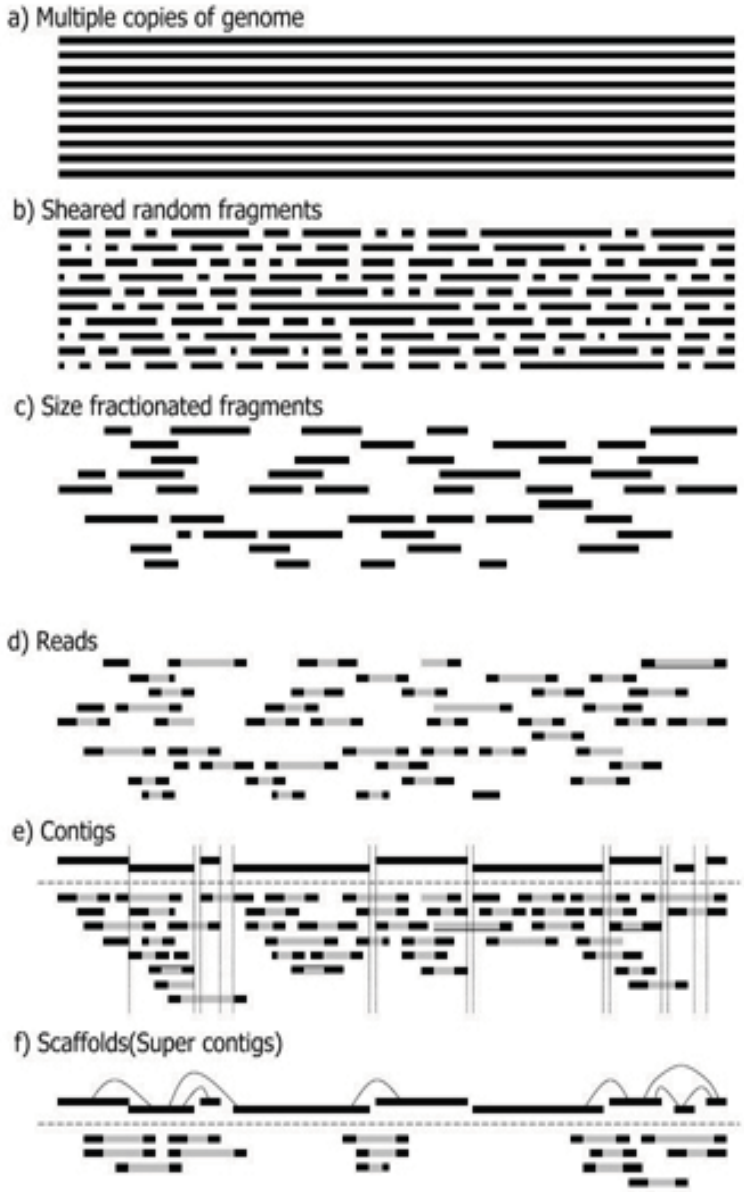
Reads aligned to a reference genome for interpretation (L5)



Courtesy of Suspencewl on wikipedia. Image is in the public domain.

http://en.wikipedia.org/wiki/Genomics

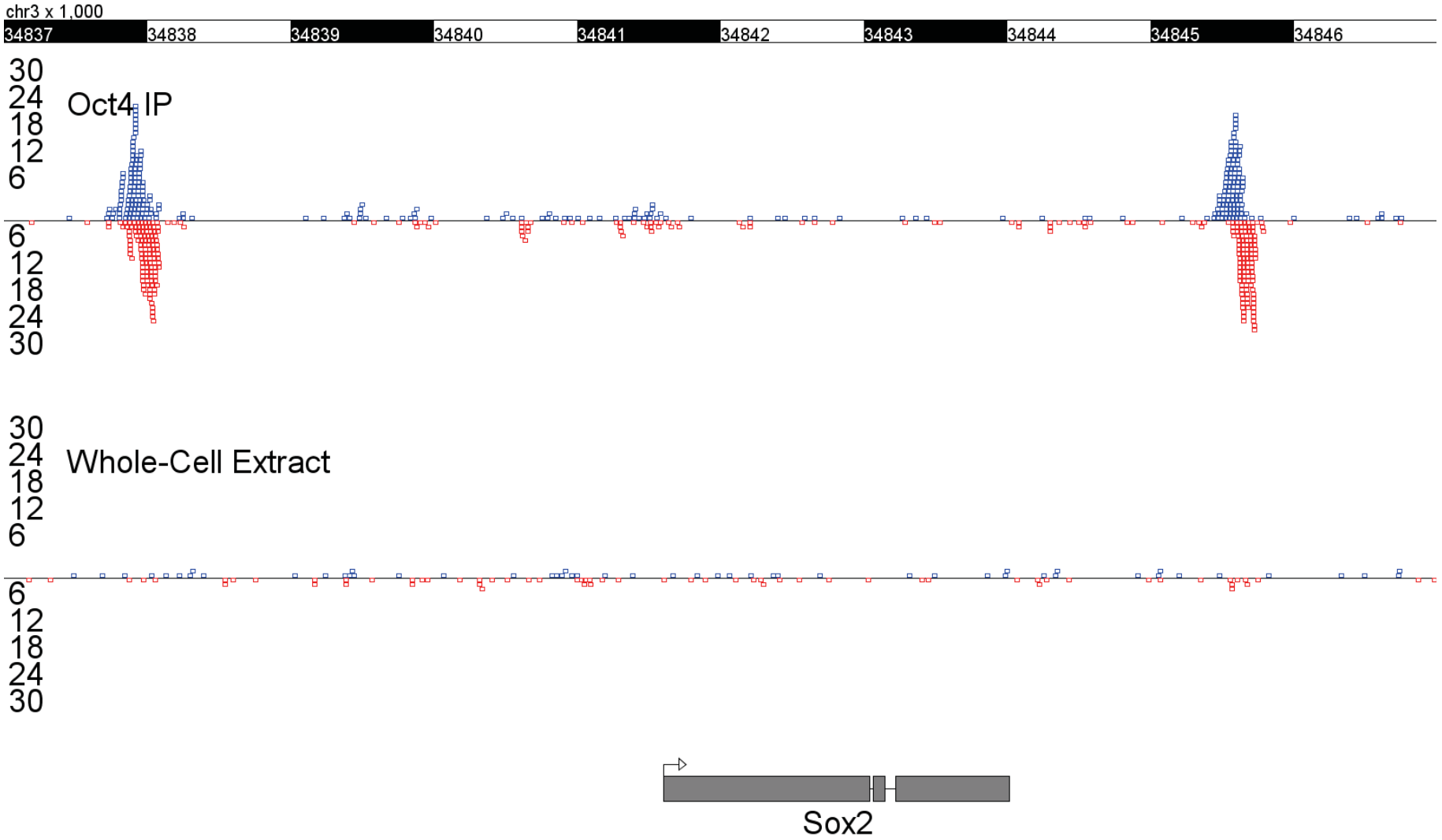
Reference genomes are assembled from millions of short reads (L6)



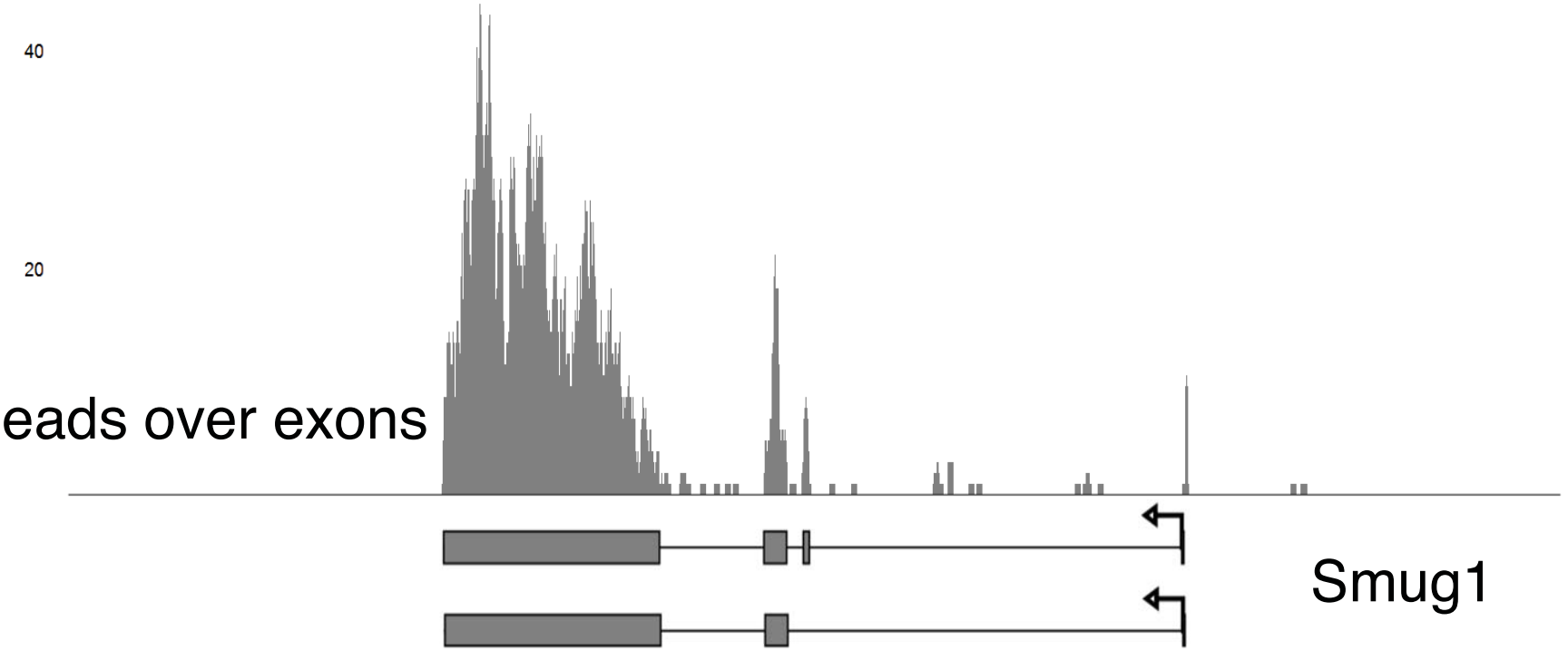
Courtesy of Shinichi Morishita. Used with permission.

http://www.k.u-tokyo.ac.jp/pros-e/person/shinichi morishita/shinichi morishita.htm

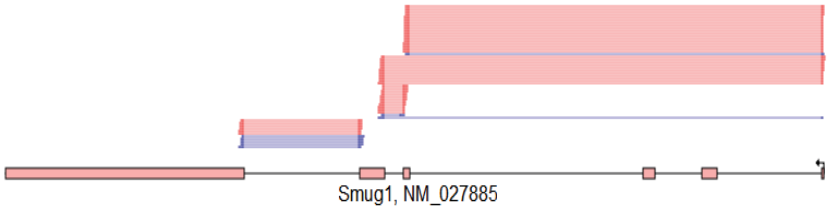
ChIP-seq reveals where key genomic regulators bind to the genome (L7)



RNA-seq reveals both RNA expression levels and isoforms (L8)



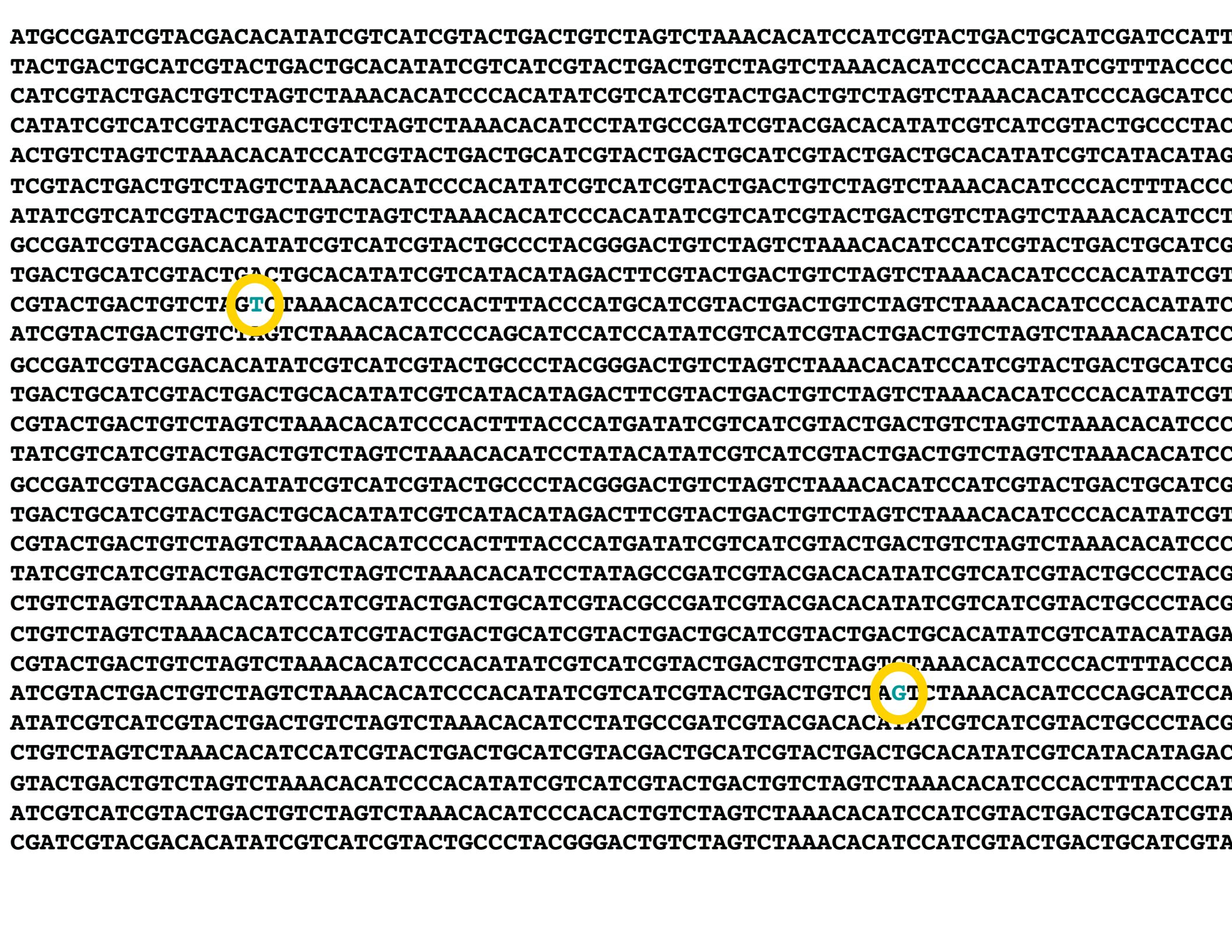
Junction reads (split between exons)



Computational Genetics Module

Understanding genome function in health and disease

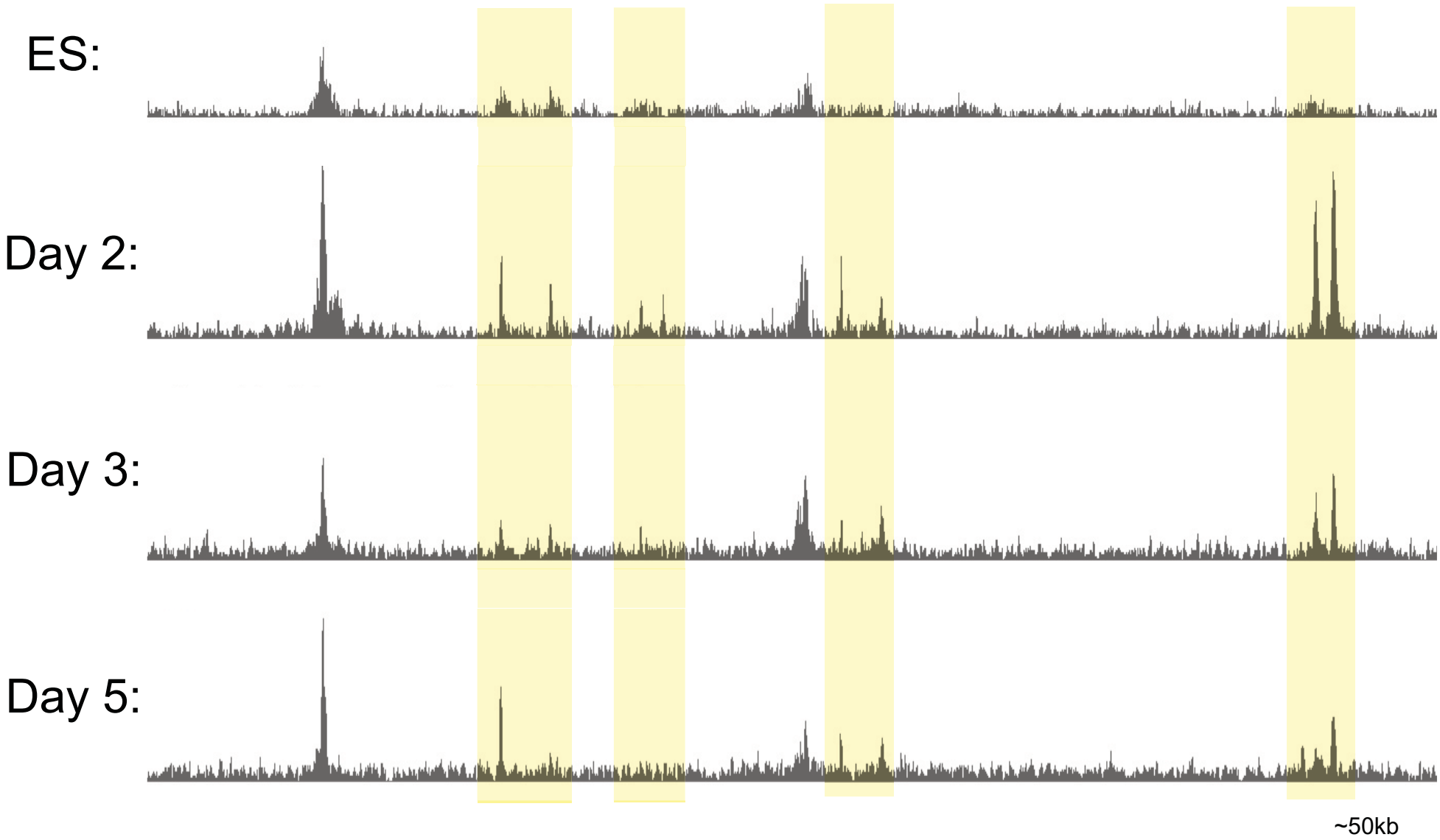
ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGATCCATT
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTTTACCCC
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCC
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCTAC
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATACATAG
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCC
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCT
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCG
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATC
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCG
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCG
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCATCGTACTGCCCTACG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCATCGTACTGCCCTACG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATACATAGA
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCA
ATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCA
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCTACG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATATCGTCATACATAGAC
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCAT
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTA
CGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTA


ATGCCGATCGTACGACACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGATCCATT
TACTGACTGCATCGTACTGACTGCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTTTACCCC
CATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCC
CATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCTAC
ACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATACATAG
TCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCC
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCT
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCG
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGCATCGTACTGACTGTCTAGTCTAAACACATCCCACATATC
ATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCATCCATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCG
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCC
GCCGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCG
TGACTGCATCGTACTGACTGCACATATCGTCATACATAGACTTCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGT
CGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCATGATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCC
TATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATAGCCGATCGTACGACACATATCGTCATCGTACTGCCCTACG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGCCGATCGTACGACACATATCGTCATCGTACTGCCCTACG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACTGACTGCATCGTACTGACTGCACATATCGTCATACATAGA
CGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCA
ATCGTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCAGCATCCA
ATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCTATGCCGATCGTACGACACATATCGTCATCGTACTGCCCTACG
CTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTACGACTGCATCGTACTGACTGCACATATCGTCATACATAGAC
GTACTGACTGTCTAGTCTAAACACATCCCACATATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACTTTACCCAT
ATCGTCATCGTACTGACTGTCTAGTCTAAACACATCCCACACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTA
CGATCGTACGACACATATCGTCATCGTACTGCCCTACGGGACTGTCTAGTCTAAACACATCCATCGTACTGACTGCATCGTA

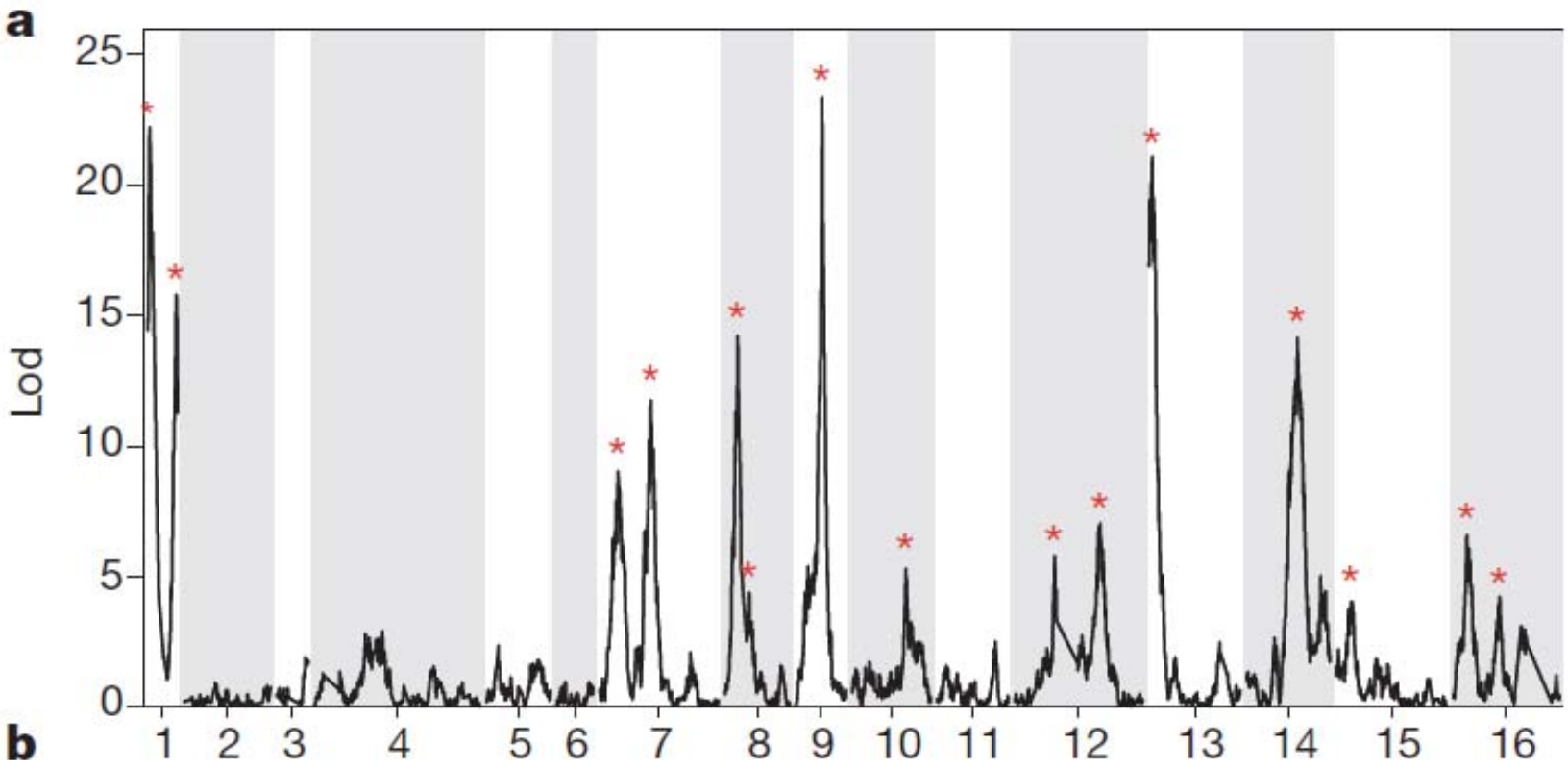
Computational Genetics Module

- L18 – How to use epigenetic and chromatin information to reveal functional genomic elements
- L18 – How to discover Quantitative Trait Loci (QTLs) that can predict quantitative traits (such as cellular growth rate)
- L19 – How to perform Genome Wide Association Studies (GWAS) that can discover human variants that predict an increased risk for specific diseases

Chromatin accessibility changes can reveal genome functional elements (L18)



Quantitative Trait Loci (QTLs) can predict phenotype (L19)



NATURE | VOL 494 | 14 FEBRUARY 2013

Finding the sources of missing heritability in a yeast cross

Joshua S. Bloom[1,2], Ian M. Ehrenreich[1,3], Wesley T. Loo[1,2], Thúy-Lan Võ Lite[1,2] & Leonid Kruglyak[1,4,5]

Source: Bloom, Joshua S., Ian M. Ehrenreich, et al. "Finding the Sources of Missing Heritability in a Yeast Cross." *Nature* 494, no. 7436 (2013): 234-37.

GWAS analysis can identify human variants associated with disease (L20)

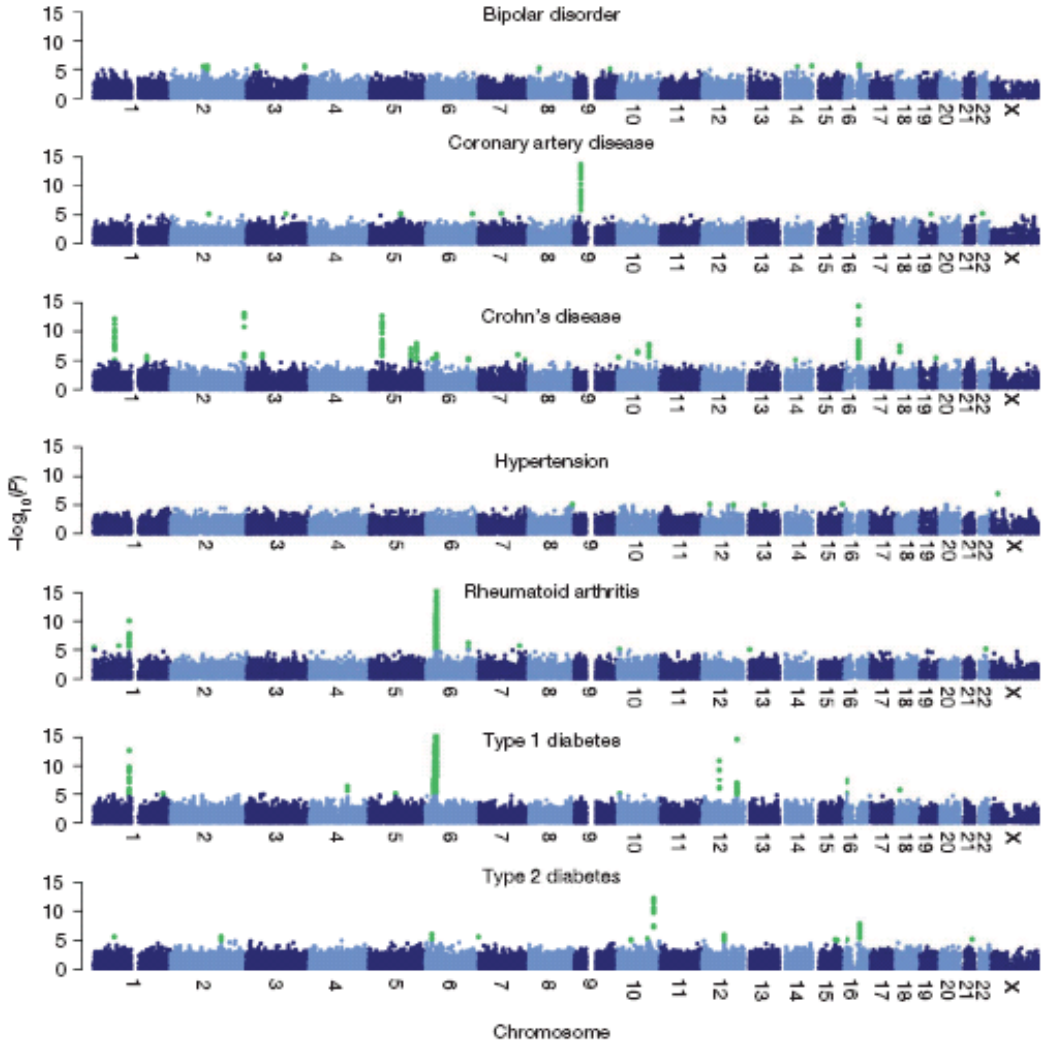


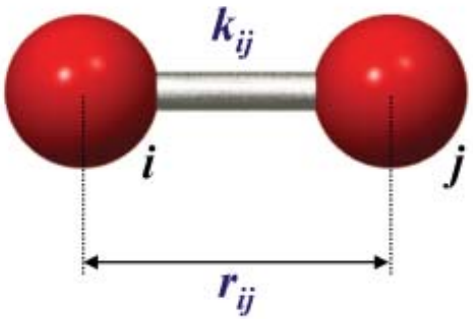
Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases $-\log_{10}$ of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with P values $<1\times10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P\text{ value})=15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.

Burton, Paul R., David G. Clayton, et al. "Genome-wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature* 447, no. 714S (2007): 661-78.

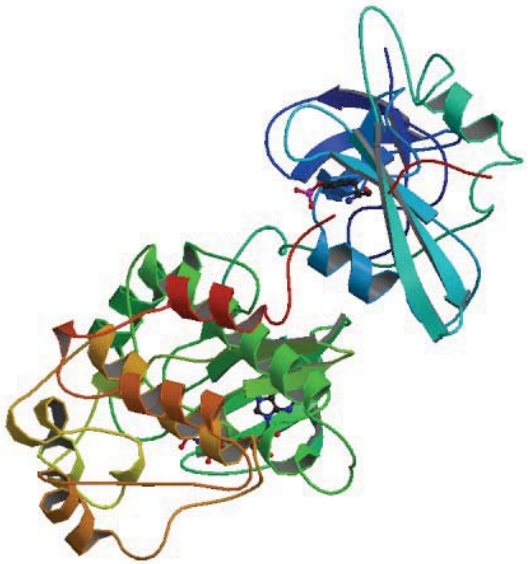
 Nature Vol 447l7 June 2007l doi:10.1038/nature05911

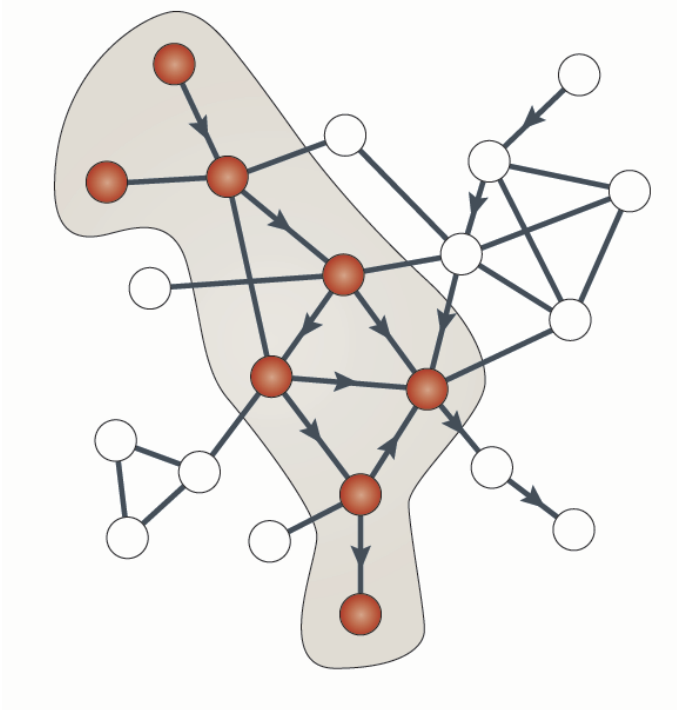
- L12 - Introduction to Protein Structure; Structure Comparison & Classification
- L13 - Predicting protein structure
- L14 - Predicting protein interactions
- L15 - Gene Regulatory Networks
- L16 - Protein Interaction Networks
- L17 - Computable Network Models

Modeling Scales



$$U_{bond} = \sum_{bonds} K_b(b - b^0)^2,$$

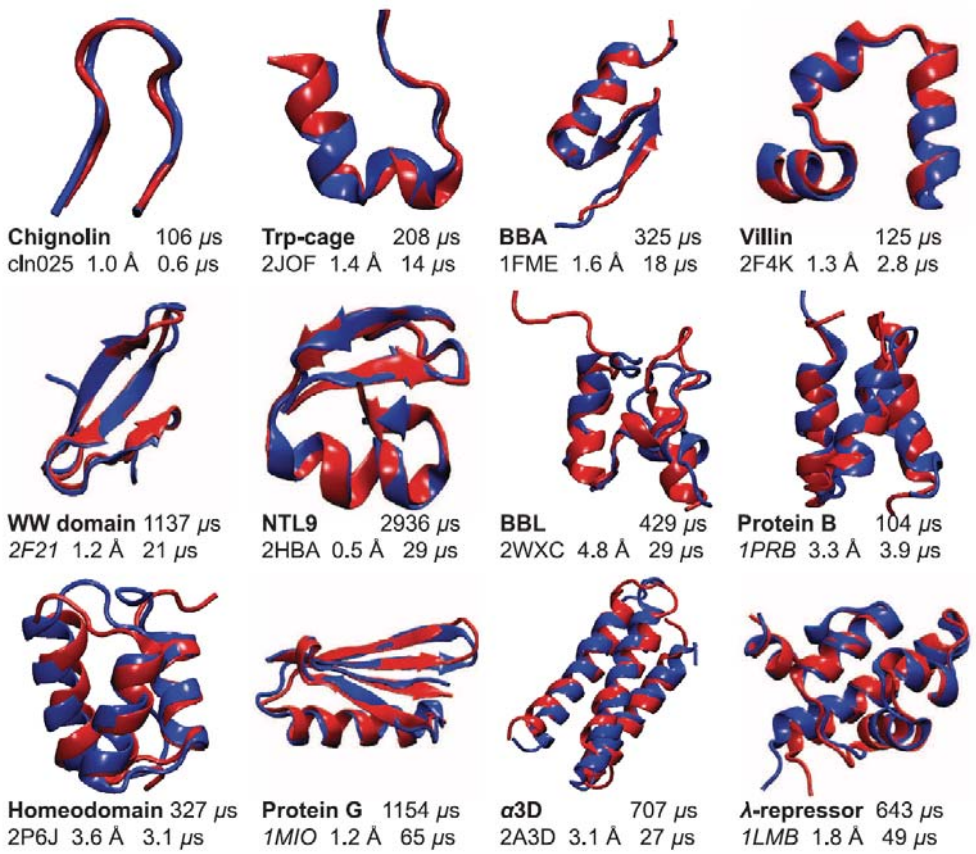




Atom **Protein** **Network**

Source: Barabasi, Albert-László, Natali Gulbahce, et al."Network Medicine: A Network-based Approach to Human Disease." *Nature Reviews Genetics* 12, no. 1 (2011): 56-68.

Predicting Protein Structure (L13)



Source: Lindorff-Larsen, Kresten, Stefano Piana, et al. "How Fast-folding Proteins Fold." *Science* 334, no. 60SS (2011): S17-20.

Predicting Protein Structure
Man vs. Machine (L13)

The New York Times

In a Video Game, Tackling the Complexities of Protein Folding

By JOHN MARKOFF
Published: August 9, 2010

Gamers 1, computer 0.

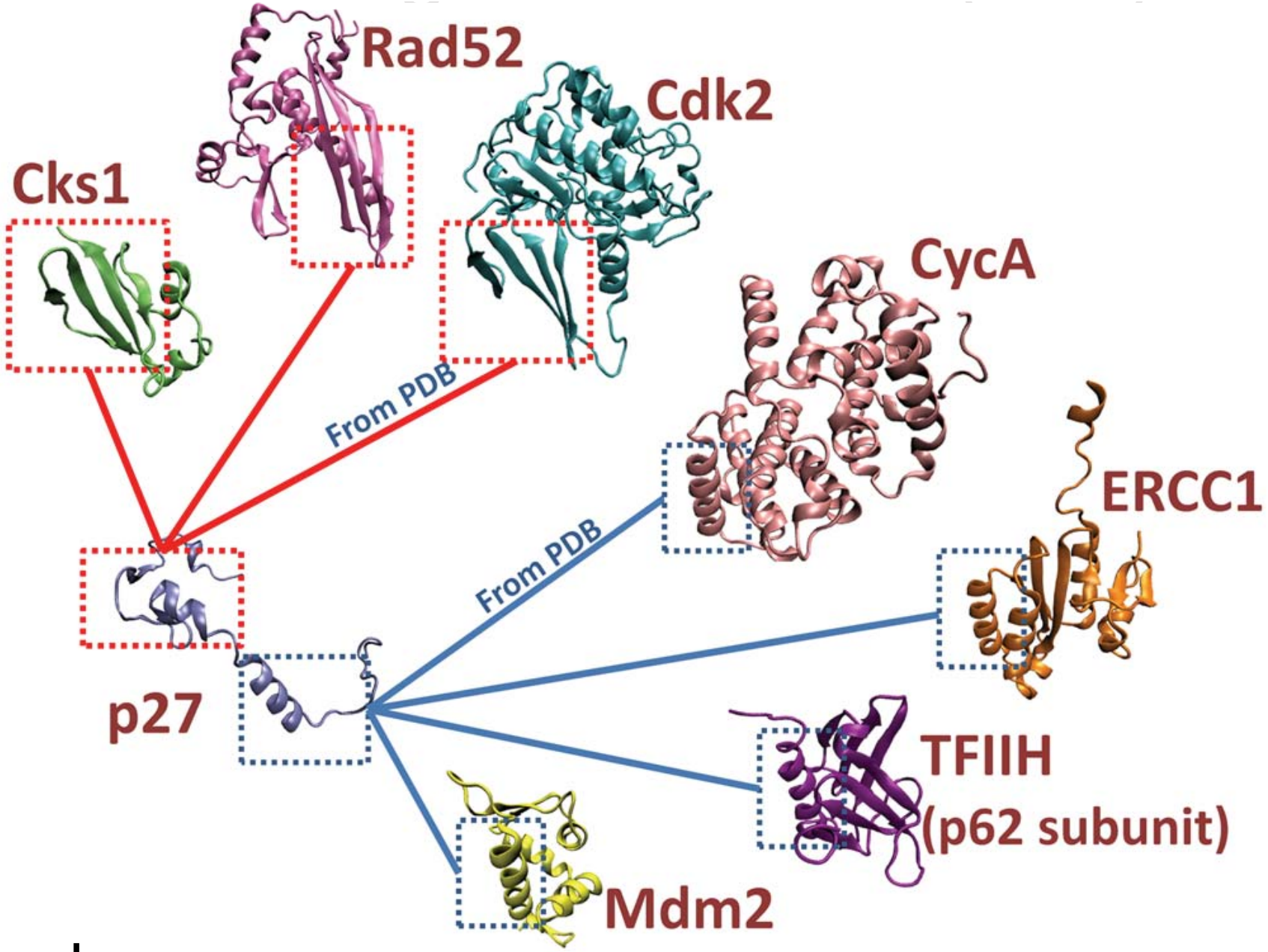
Enlarge This Image

In a match that pitted video game players against the best known computer program designed for the task, the gamers outperformed the software in figuring out how 10 proteins fold into their three-dimensional configurations.

Source: Markoff, John. " In a Video Game, Tackling the Complexities of Protein Folding." *The New York Times*, 2010.

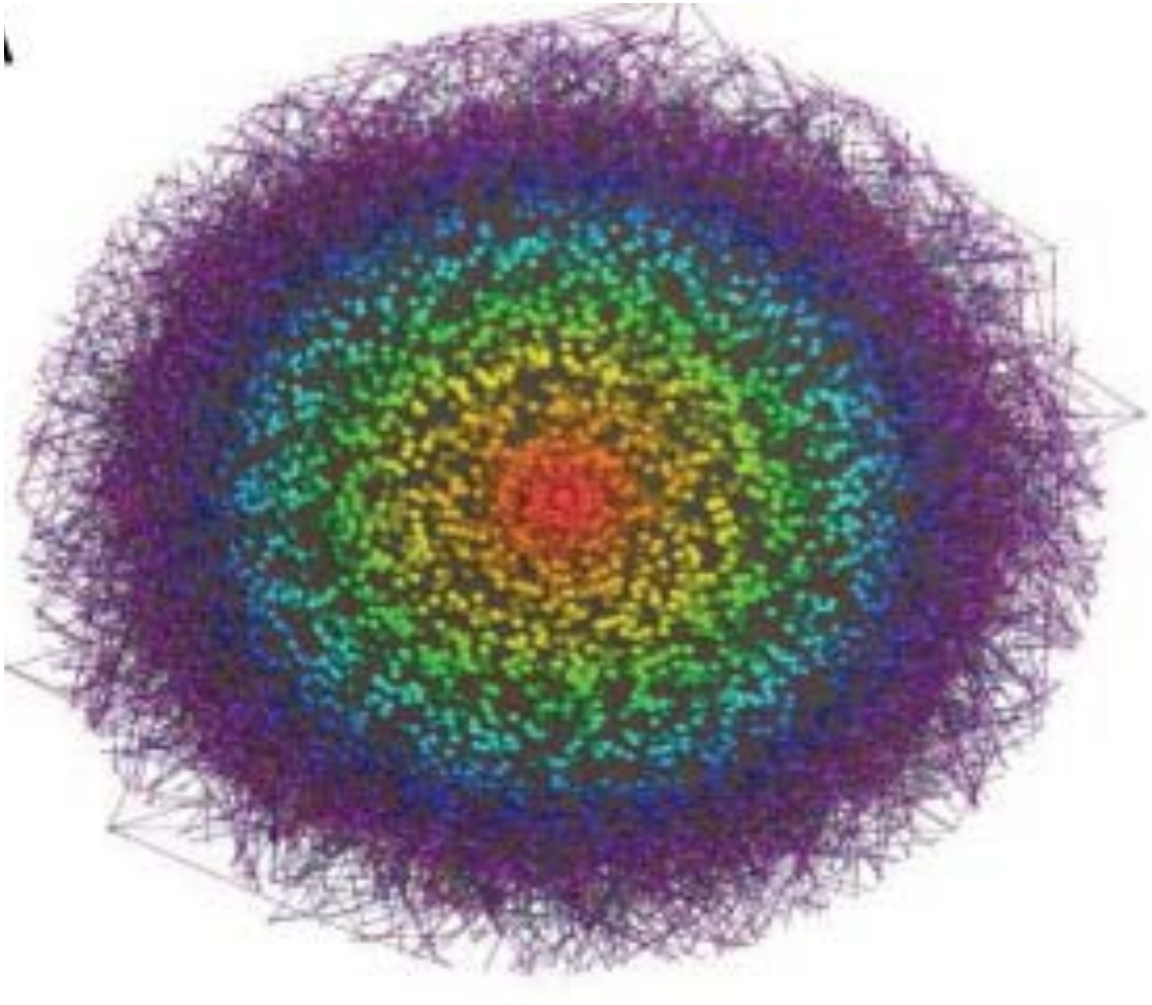
Source: Shaw, David E., Martin M. Deneroff, et al. "Anton, A Special-purpose Machine for Molecular Dynamics Simulation." Communications of the ACM S1, no. 7 (2008): 91-7.

Predicting Interactions (L14)



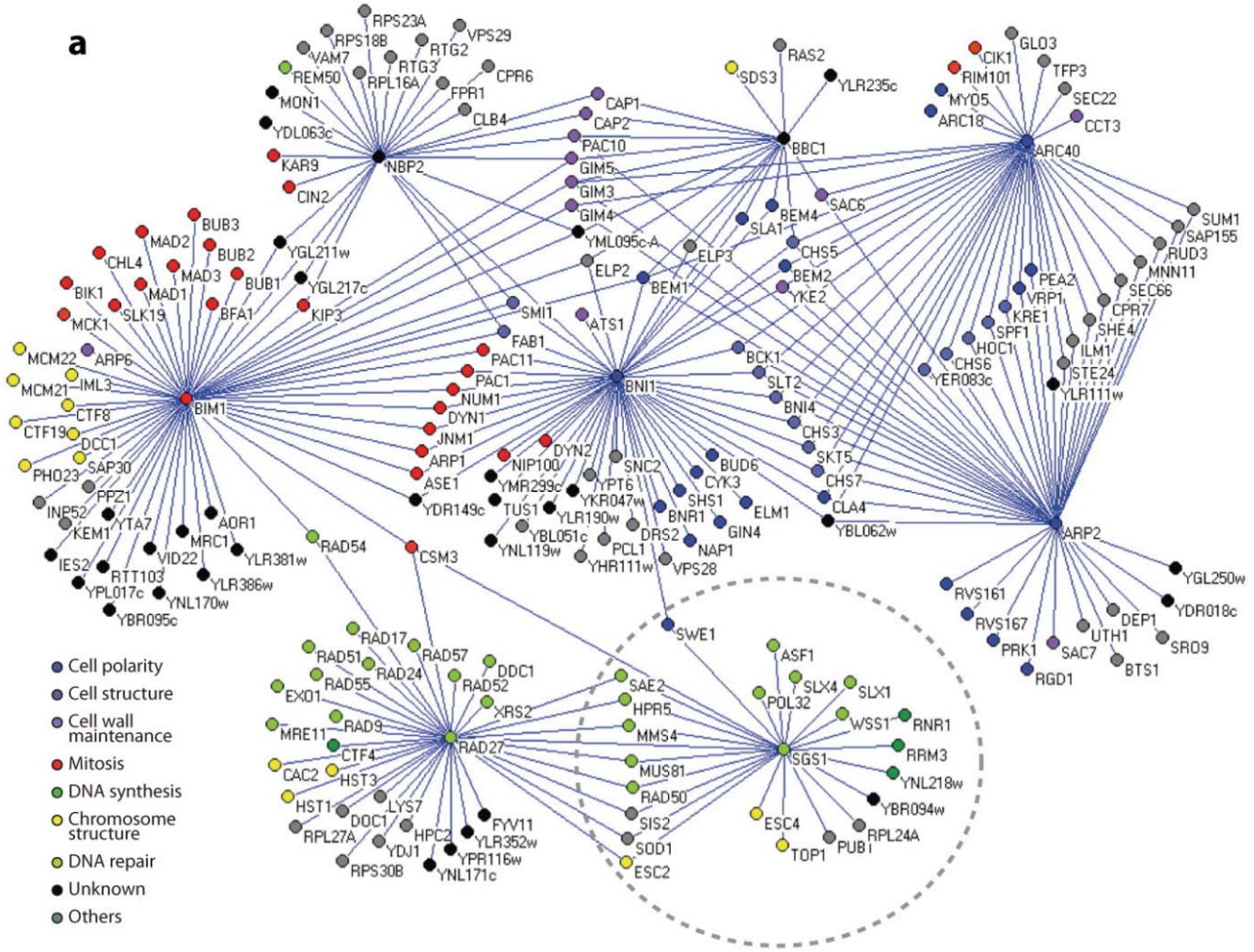
N. Tuncbag

Gene Regulatory Networks (L15)



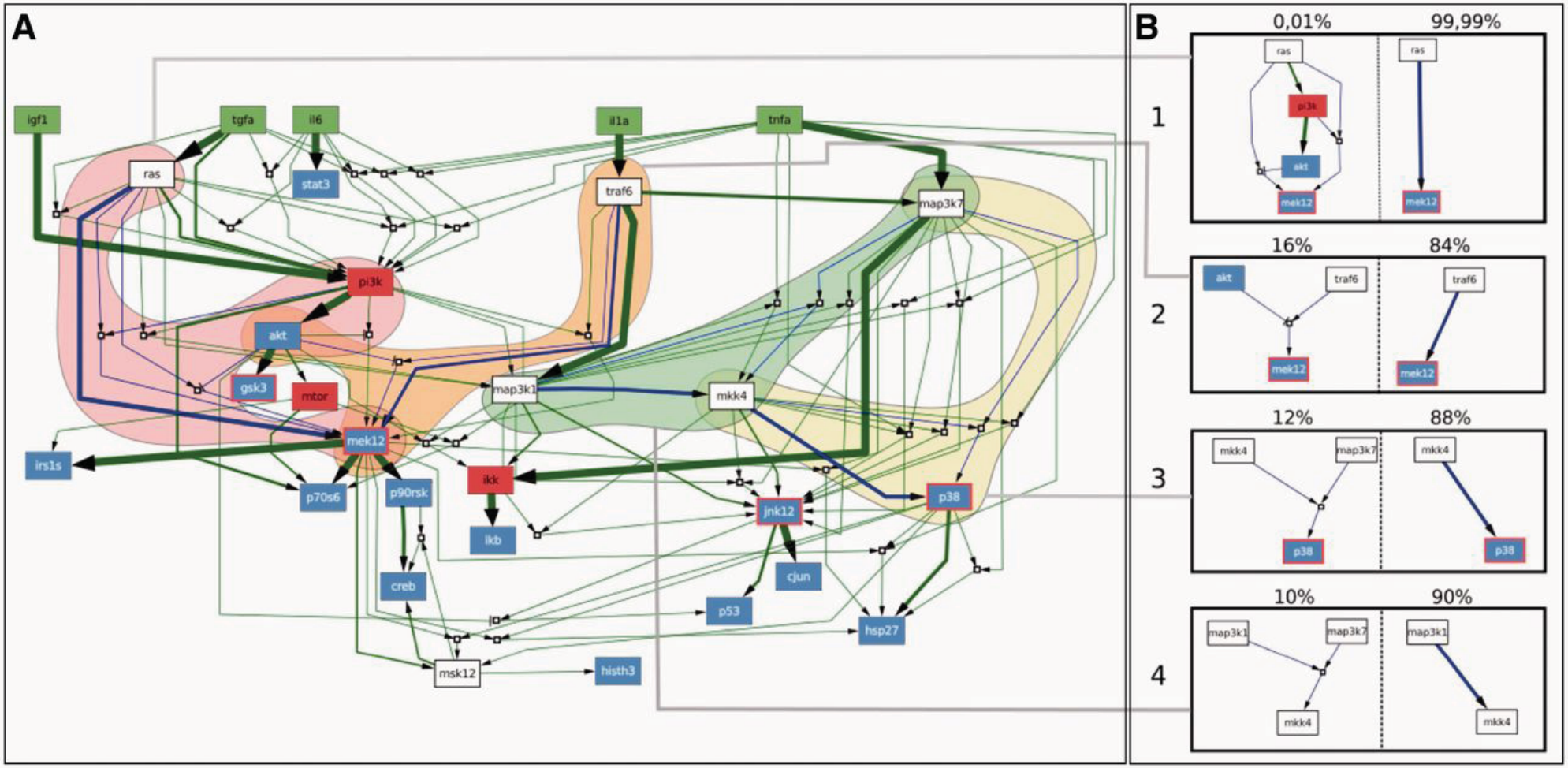
Source: Sumazin, Pavel, Xuerui Yang, et al. "An Extensive MicroRNA-mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma." *Cell* 147, no. 2 (2011): 370-81.

Interaction Networks (L16)



Dixon *et* al. (2009) Annual Review of Genetics Vol. 43: 601-625 (doi:10.1146/annurev.genet.39.073003.114751)

Computable Models (L17)



Source: Guziolowski, Carito, Santiago Videla, et al. "Exhaustively Characterizing Feasible Logic Models of a Signaling Network using Answer Set Programming." *Bioinformatics* 30, no. 13 (2014): 1942-2.

Guziolowski C et al. Bioinformatics 2013;29:2320-2326

Course	Project	AI problems
7.36/20.390/6.802	NO	NO
7.91/20.490/HST.506	YES	NO
6.874	YES	YES

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014