Introduction to Bioinformatics: Past and Present

Juan Cui, Associate Professor

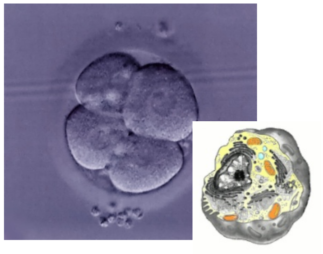
System Biology and Biomedical Informatics (SBBI) Laboratory

Computer Science and Engineering, UNL

Outline

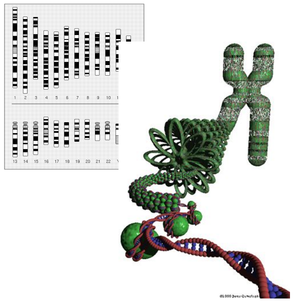
- Introduction to Bioinformatics
  - ❑ Historical milestones
- Omics and big data challenges
- Example projects
- Summary

The Basics



cel

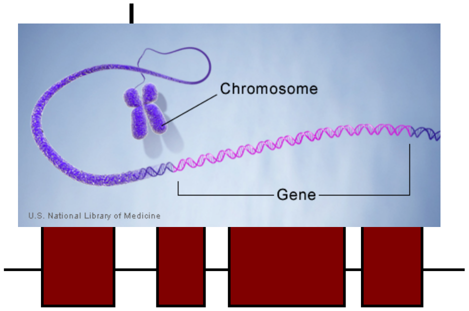
l



chromosome

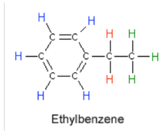
ccgtacgtacgtagagtgctagtctagtcgtagcgccgtagtcga
tcgtgtgggtagtagctgatatgatgcgaggtaggggataggata
gcaacagatgagcggatgctgagtgcagtggcatgcgatgtcg
atgatagcggtaggtagacttcgcgcataaagctgcgcgagatg
attgcaaagragttagatgagctgatgctagaggtcagtgactga
tgatcgatgcatgcatggatgatgcagctgatcgatgtagatgca
ataagtcgatgatcgatgatgatgctagatgatagctagatgtgat
cgatggtaggtaggatggtaggtaaattgatagatgctagatcgt
aggta…………………………………

genome: DNA sequence

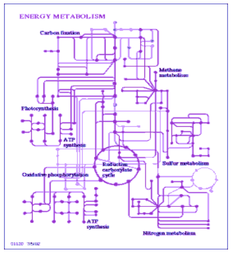


gene

protein



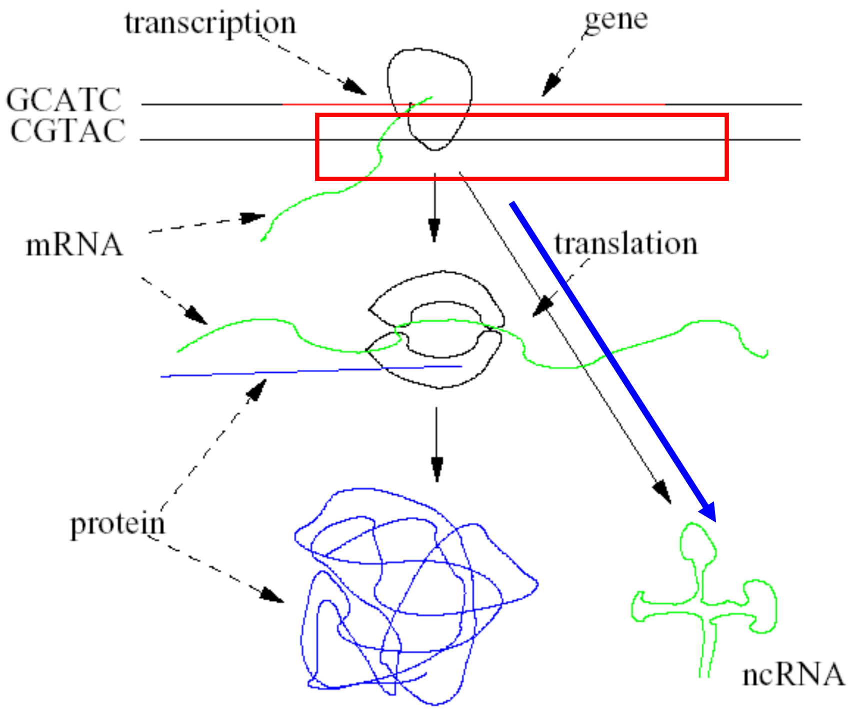
metabolite

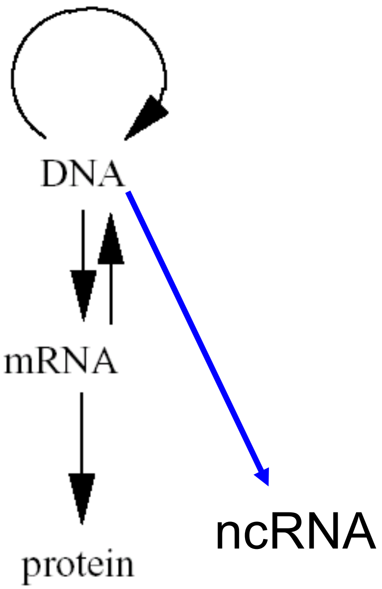


metabolic pathway

Cells are the smallest unit of life, the basic building blocks of all living things.
Chromosomes are threadlike structures of nucleic acids and protein found in the nucleus of most living cells, carrying genetic information in the form of genes.
Genes are regions of DNA that encodes functional RNAs or proteins and are the molecular units of heredity.
Proteins are large biomolecules consisting of one or more long chains of amino acid residues. They perform a vast array of functions.
Metabolites are substances formed in or necessary for metabolism.

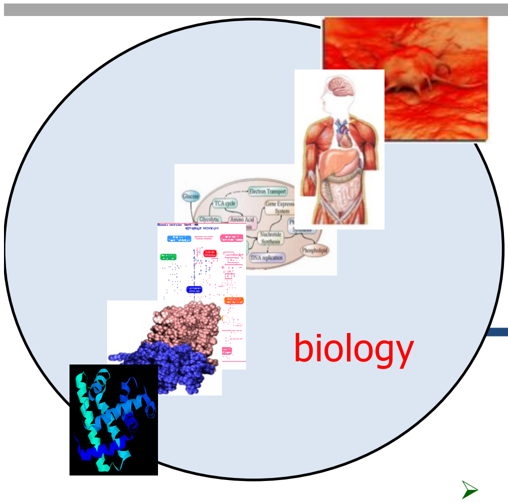
Central Dogma





ncRNA

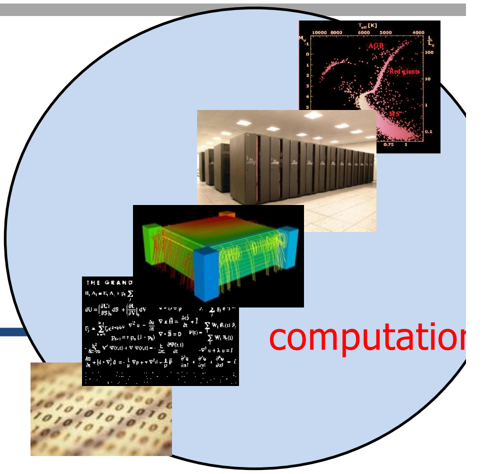
Bioinformatics (Computational Biology)



data management; data mining; modeling; prediction; theory formulation

bioinformatics

genes, proteins, protein complexes, pathways, cells, organisms, ecosystem



- This interdisciplinary science … is about *providing computational support to studies on linking the behavior of cells, organisms and populations to the information encoded in the genomes*.

an indispensable part of biological science

-- Temple Smith, *Current Topics in Computational Molecular Biology (2002)*

engineering aspect

scientific aspect

computer science, biology, medicine, statistics,

mathematics, physics, chemistry, engineering,…

Bioinformatics

- It is about developing and using computational techniques to
  - ❑ analyze and interpret biological data
  - ❑ predict structures and functions of biological entities
  - ❑ model the dynamic behavior of biological processes and systems
  - ❑ ……
- People have used mathematical or computing techniques to solve biological problems since early 1900’s
  - ❑ e.g., evolution and genetic analyses by Ronald Fisher, J.B.S. Haldane, S. Wright (founders of population genetics, study of the distributions and changes of allele frequency in a population)
- So what is new?

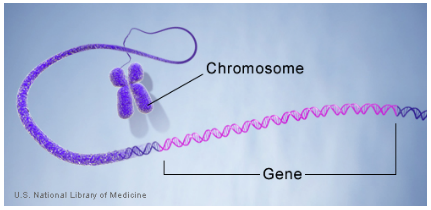
Allele: one of two or more alternative forms of a gene that arise by mutation and are found at the same place on a chromosome.

A Historical Perspective: gene discovery

- Realization of the existence of “gene” in our cells by Hermann Müller, a student of Thomas Hunt Morgan (1921)
  - ❑ The role the chromosome plays in heredity in drosophila; X-rays could induce mutations
  - ❑ *Both won Nobel Prize in Physiology or Medicine*

- Understanding of the physical natures of genes by
  - ❑ Fred Sanger (e.g., 1949), for “determination of base sequences in nucleic acids”
  - ❑ E. Chargraff (e.g., 1950), J. Kendrew (e.g., 1958)

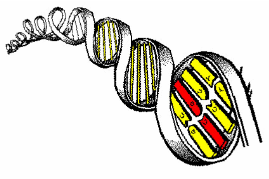
  in 40’ and 50’s



A **gene** is a locus (or region) of DNA that encodes a functional RNA or protein product, and is the molecular unit of heredity.

A Historical Perspective: DNA discovery

- Understanding of the double helical structure of DNA by James Watson and Frances Crick in 1953

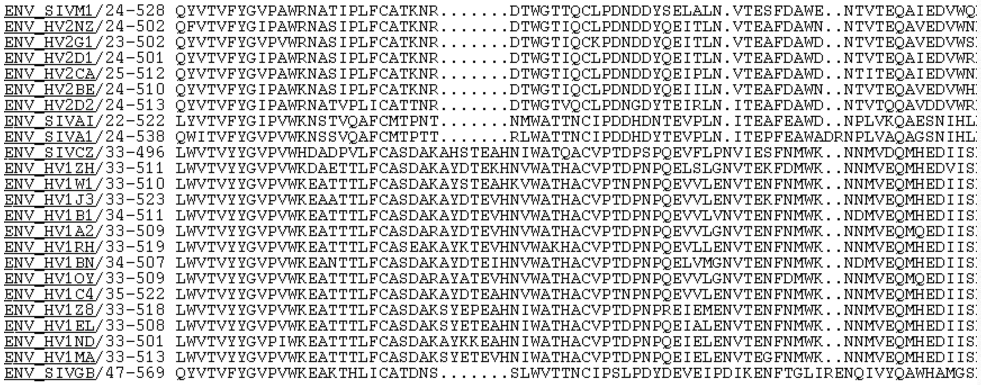


- Development of sequencing technology, first of proteins ar genomic DNA, based on the work of
  - ❑ Fred Sanger on sequencing of insulin (1956),
  - ❑ Walter Gilbert and Allan Maxam on sequencing of Lactose operator (1977)

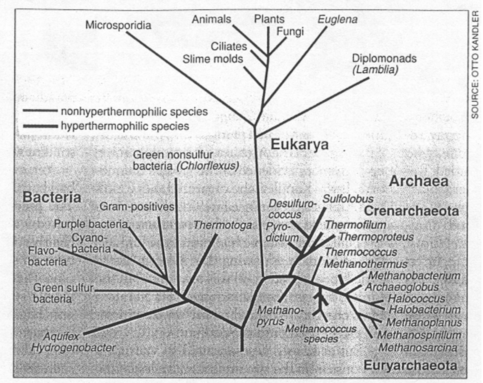
which demonstrated that the genetic sequence of a genome, including human's, is sequence-able!

A Historical Perspective: protein discovery

- Development of a *science* of analyzing protein and DNA sequences, particularly in
  - protein sequence analyses and evolution by Margaret Dayhoff (60’s)
  - phylogenetic analyses and comparative sequence analyses by W. Fitch and E. Margoliash (1967) and by R. Doolittle (1983)



```
ENV_SIVM1/24-528 QYVTVFYGVPAWRNATIPLFCATKNR.......DTWGTTQCLPDNDDYSELALN.VTESFDAWE..NTVTEQAIEDVWQ
ENV_HV2NZ/24-502 QFVTVFYGIPAWRNASIPLFCATKNR.......DTWGTIQCLPDNDDYQEITLN.VTEAFDAWN..NTVTEQAVEDVWN
ENV_HV2G1/23-502 QYVTVFYGVPVWRNASIPLFCATKNR.......DTWGTIQCKPDNDDYQEITLN.VTEAFDAWD..NTVTEQAVEDVWS
ENV_HV2D1/24-501 QYVTVFYGIPAWRNASIPLFCATKNR.......DTWGTIQCLPDNDDYQEITLN.VTEAFDAWD..NTVTEQAIEDVWR
ENV_HV2CA/25-512 QYVTVFYGVPAWKNASIPLFCATKNR.......DTWGTIQCLPDNDDYQEIPLN.VTEAFDAWD..NTITEQAIEDVWN
ENV_HV2BE/24-510 QYVTVFYGIPAWKNASIPLFCATKNR.......DTWGTIQCLPDNDDYQEIILN.VTEAFDAWN..NTVTEQAVEDVWH
ENV_HV2D2/24-513 QYVTVFYGIPAWRNATVPLICATTNR.......DTWGTVQCLPDNGDYTEIRLN.ITEAFDAWD..NTVTQQAVDDVWR
ENV_SIVAI/22-522 LYVTVFYGIPVWKNSTVQAFCMTPNT.......NMWATTNCIPDDHDNTEVPLN.ITEAFEAWD..NPLVKQAESNIHL
ENV_SIVA1/24-538 QWITVFYGVPVWKNSSVQAFCMTPTT.......RLWATTNCIPDDHDYTEVPLN.ITEPFEAWADRNPLVAQAGSNIHL
ENV_SIVCZ/33-496 LWVTVYYGVPVWHDADPVLFCASDAKAHSTEAHNIWATQACVPTDPSPQEVFLPNVIESFNMWK..NNMVDQMHEDIIS
ENV_HV1ZH/33-511 LWVTVYYGVPVWKDAETTLFCASDAKAYDTEKHNVWATHACVPTDPNPQELSLGNVTEKFDMWK..NNMVEQMHEDVIS
ENV_HV1W1/33-510 LWVTVYYGVPVWKEATTTLFCASDAKAYSTEAHKVWATHACVPTNPNPQEVVLENVTENFNMWK..NNMVEQMHEDIIS
ENV_HV1J3/33-523 LWVTVYYGVPVWKEAATTLFCASDAKAYDTEVHNVWATHACVPTDPNPQEVVLENVTEKFNMWK..NNMVEQMHEDIIS
ENV_HV1B1/34-511 LWVTVYYGVPVWKEATTTLFCASDAKAYDTEVHNVWATHACVPTDPNPQEVVLVNVTENFNMWK..NDMVEQMHEDIIS
ENV_HV1A2/33-509 LWVTVYYGVPVWKEATTTLFCASDARAYDTEVHNVWATHACVPTDPNPQEVVLGNVTENFNMWK..NNMVEQMQEDIIS
ENV_HV1RH/33-519 LWVTVYYGVPVWKEATTTLFCASEAKAYKTEVHNVWAKHACVPTDPNPQEVLLENVTENFNMWK..NNMVEQMHEDIIS
ENV_HV1BN/34-507 LWVTVYYGVPVWKEANTTLFCASDAKAYDTEIHNVWATHACVPTDPNPQELVMGNVTENFNMWK..NDMVEQMHEDIIS
ENV_HV1OY/33-509 LWVTVYYGVPVWKEATTTLFCASDARAYATEVHNVWATHACVPTDPNPQEVVLGNVTENFDMWK..NNMVEQMQEDIIS
ENV_HV1C4/35-522 LWVTVYYGVPVWKEATTTLFCASDAKAYDTEAHNVWATHACVPTNPNPQEVVLENVTENFNMWK..NNMVEQMHEDIIS
ENV_HV1Z8/33-518 LWVTVYYGVPVWKEATTTLFCASDAKSYEPEAHNIWATHACVPTDPNPREIEMENVTENFNMWK..NNMVEQMHEDIIS
ENV_HV1EL/33-508 LWVTVYYGVPVWKEATTTLFCASDAKSYETEAHNIWATHACVPTDPNPQEIALENVTENFNMWK..NNMVEQMHEDIIS
ENV_HV1ND/33-501 LWVTVYYGVPIWKEATTTLFCASDAKAYKKEAHNIWATHACVPTDPNPQEIELENVTENFNMWK..NNMVEQMHEDIIS
ENV_HV1MA/33-513 LWVTVYYGVPVWKEATTTLFCASDAKSYETEVHNIWATHACVPTDPNPQEIELENVTEGFNMWK..NNMVEQMHEDIIS
ENV_SIVGB/47-569 QYVTVFYGVPVWKEAKTHLICATDNS.......SLWVTTNCIPSLPDYDEVEIPDIKENFTGLIRENQIVYQAWHAMGS
```



SOURCE: OTTO KANDLER

Tree of life. The Woese family tree shows that most life is one-celled, and that the oldest cells were hyperthermophiles.

A Historical Perspective: computational methods

- Development of sequence comparison algorithms
  - ❑ Needleman and Wunsch (1970)
  - ❑ Smith and Waterman (1981)

- Organization of biological data into databases
  - ❑ Protein Data Bank (PDB, 1973) of protein structures
  - ❑ GENBANK (1982) of DNA sequences

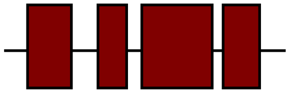
- Computational methods for gene finding in genomic sequences
  - ❑ Work by Borodovsky, Claverie, Uberbacher from mid-80’s to early 90’s

A Historical Perspective: High-throughput techniques

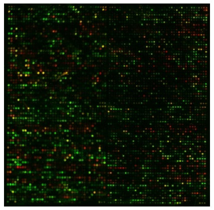
- Sequencing of Human and other genomes
  - □ (1986 – 2003)

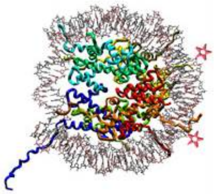
- Development of “high-throughput” measurement technologies
  - □ microarray chips for functional states of genes
  - □ two-hybrid systems for protein-protein interactions
  - □ structural genomics for structure determination
  - □ ………

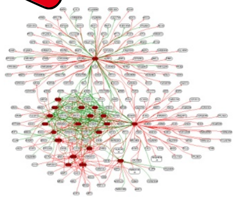
ccgtacgtacgtagagtgctagt
ctagtcgtagcgccgtagtcgatc
gtgtgggtagtagctgatatgatg
cgaggtaggggataggatagca
acagatgagcggatgctgagtg
cagtggcatgcgatgtgatagct
agatgtgatcgatggtaggtagg
atggtaggt



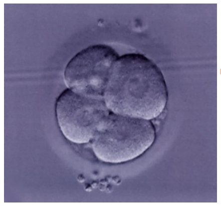
gene



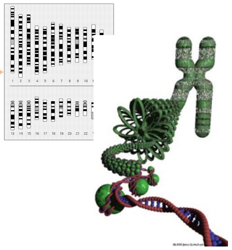




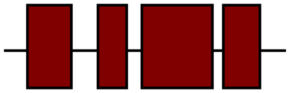
Omics Techniques



cell



chromosomes

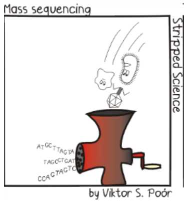


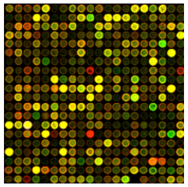
gen
e

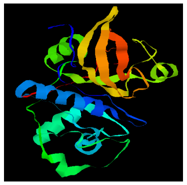
Genomics

(Transcriptomics)

atgcatgcatggatgatgcagctgatcgatgtag
atgcaataagtcgatgatcgatgatgatgctaga
tgatagctagatgtgatcgatggtaggtaggatg
gtaggtaaattgatagatgctagatcgtaggta
……

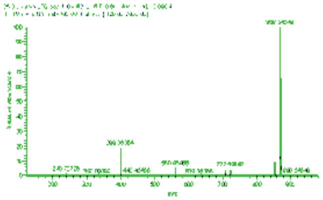


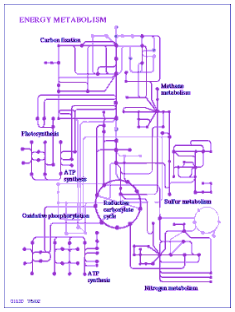




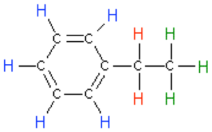
protein
s

Proteomics



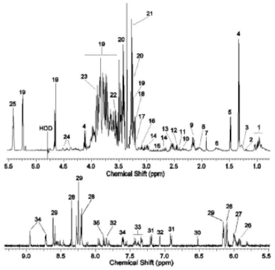


metabolic pathways



metabolite
s

Metabolomics



Molecular Biology is Becoming An Information Science!

- These "high-throughput" probing technologies and others are being used to generate enormous amounts of data about

  the *existence*,

  the *structure*,

  the *functional state,*

  the *relationship* of biological molecules and machineries …..

- … how to analyze and interpret these data ? …..

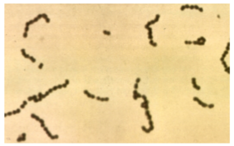
It is the amount & the type of biological data *about the cellular states, molecular structures and functions, generated by high-throughput technologies*, that have driven the rapid advancement of bioinformatics!

So what is new?

An Example of Computation for Biology

- *Lactococcus* is a premier model microorganism for a wide array of studies in molecular biology, *producing lactic acid through glucose fermentation*
  - ❑ **is nonpathogenic**

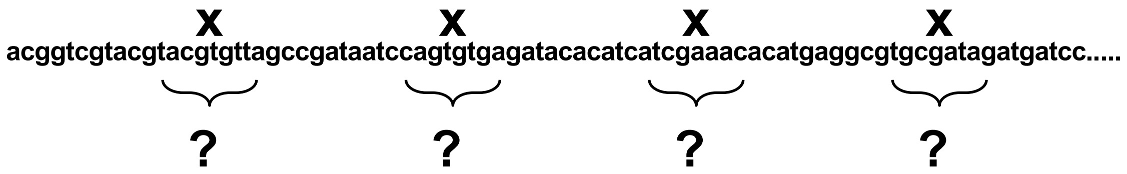
- *Streptococcus* is closely related to *Lactococcus* and could become pathogenic



- Question: What make one pathogenic and the other non-pathogenic?
  - ❑ specific genes?
  - ❑ unique pathways?
  - ❑ different regulatory mechanisms?
  - ❑ ……

An Example of Computation for Biology

- X years ago, …. to search for potential genes that possibly make the difference, researchers had to
  - ❑ remove various parts of DNA sequence,
  - ❑ then observe if they may have any relevance

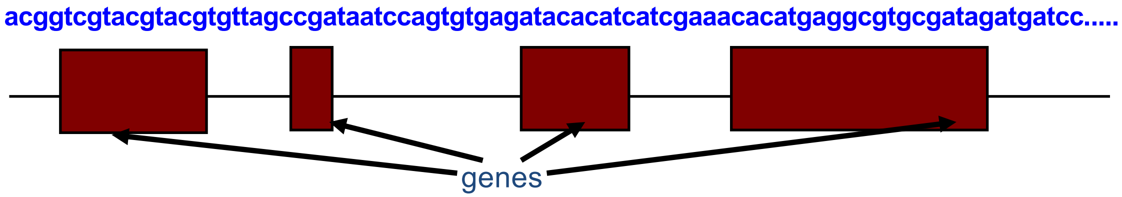


acggtcgtacgtacgtgttagccgataatccagtgtgagatacacatcatcgaaacacatgaggcgtgcgatagatgatcc.....

This could be a very lengthy process ……

An Example of Computation for Biology

- Since the Human genome project (1986), computational scientists have developed computer programs to locate genes in genomic DNA sequence
  - ❑ GRAIL, Gene-Scan, Glimmer, …….


acggtcgtacgtacgtgttagccgataatccagtgtgagatacacatcatcgaaacacatgaggcgtgcgatagatgatcc.....
genes

- With gene-prediction programs, researchers only need to knock-out regions predicted to be genes in their search for relevant genes

An Example of Computation for Biology

- Over the years, many genes have been thoroughly studied in different organisms, e.g., human, mouse, fly, …., rice, …
  - ❑ their biological functions have been identified and documented

- Computational scientists have developed computer programs to associate newly identified genes to genes with known functions!
  - ❑ Existing methods can associate > 60% of newly identified genes to genes with known functions

- Now, researchers only need to knock-out genes with possibly relevant functions in their search for understanding of a particular biological process ……

More Advanced Computation

- Computational programs have been springing out that can predict
  - □ if two proteins interact with each other
  - □ if a group of gene products work in the same pathway
  - □ functions of genes at a genome scale
  - □ ……..

These capabilities allow researchers to study complex biology problems like understanding the difference between *Lactococcus* and *Streptococcus* in a more efficient and systematic manner

Computation for Biology at Different Level

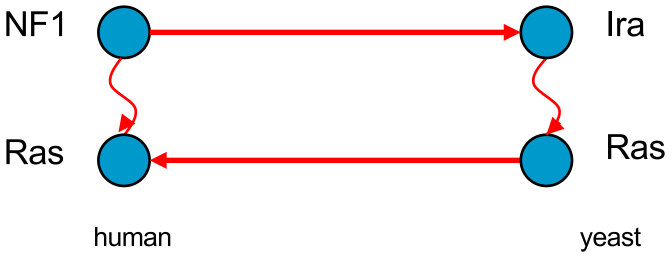
Biocomputing (bioinformatics, computational biology), in conjunction with large-scale bio-data, facilitates tackling large, complex biological problems at systems level

More examples……

Examples of “Computation for Biology”

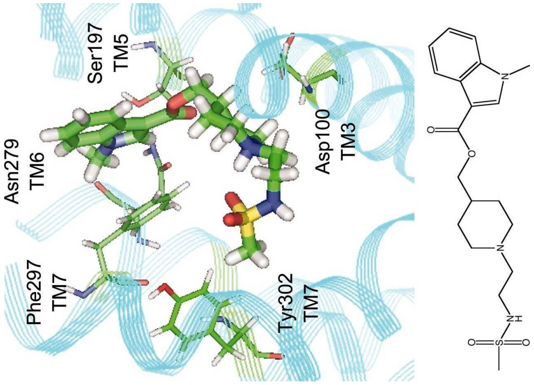
➢Suggesting functions of newly identified genes

- ❑ It was known that mutations of NF1 are associated with inherited disease *neurofibromatosis* 1; but little is known about the molecular basis of the disease
- ❑ Sequence search found that NF1 is homologous to a *yeast* protein called Ira, which is a GAP-type protein and known to regulate the function of a second type of protein called Ras
- ❑ Hypothesis: NF1 regulates Ras in human cell; follow-up experiments verified this.



Examples of “Computation for Biology”

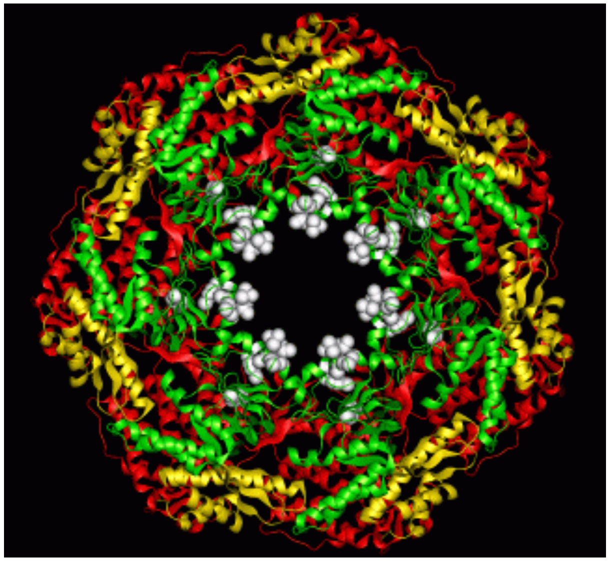
- Computer-assisted drug design
  - 3D structure models of G protein-coupled receptors were used to computationally screen 100,000+ compounds as possible drug targets and 100 were selected
  - Follow-up experiments confirmed a high hit rate of 12%-21%



OM Becker, et al, PNAS, 2004, 101:11304-11309

Examples of “Computation for Biology”

- Computational studies reveal the functional mechanism of GroEL heptamer (chaperonin, a protein complex that facilitates protein folding）



14-subunit double-toroid assembly

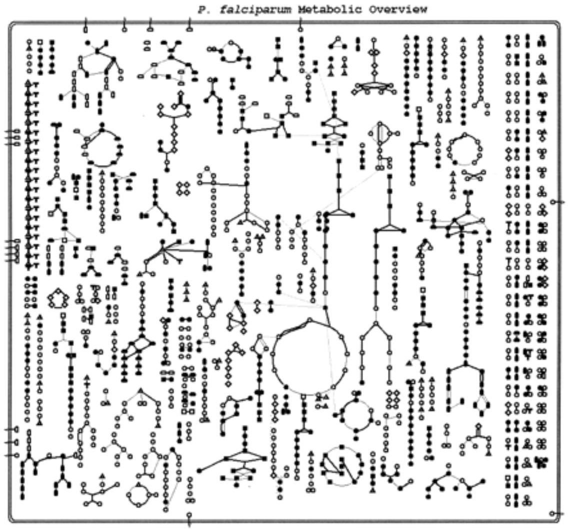
GroEL conformational changes using a targeted molecular dynamics (TMD) method

Courtesy of JP Ma’s lab

Examples of "Computation for Biology"

- Computational analysis of *Plasmodium falciparum* metabolism
  - Plasmodium causes human malaria

  - computational prediction of metabolic pathways of plasmodium
  - computational simulations have helped to identify 216 "chokepoints" in this pathway model
  - among all 24 previously suggested drug targets, 21 target at the "chokepoints"
  - among the three popular drugs for malaria, they all targeted at the "chokepoints"

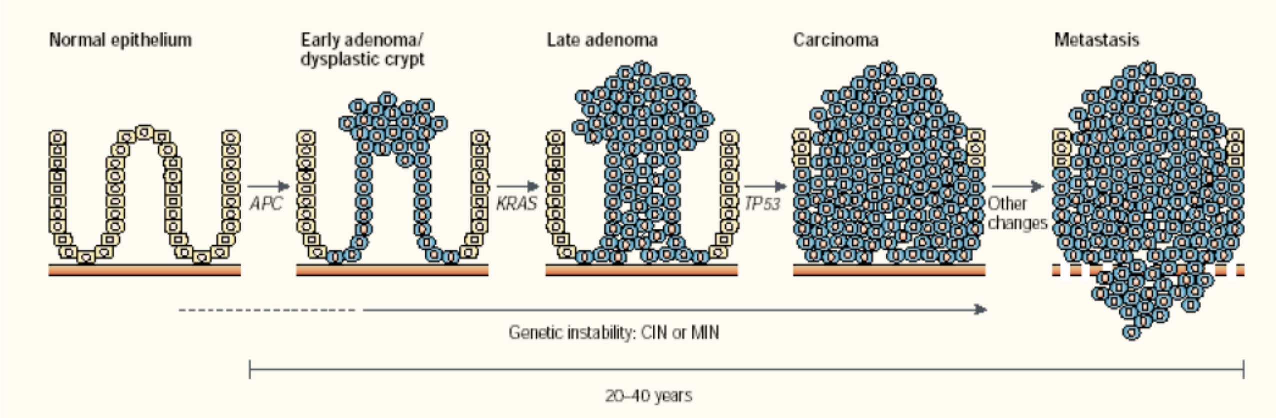


A "chokepoint reaction" is defined as a reaction that either consumes a unique substrate or produces a unique product in the PlasmoCyc metabolic network.

Yeh I. et. al. Genome Research 2004, 14:917-24

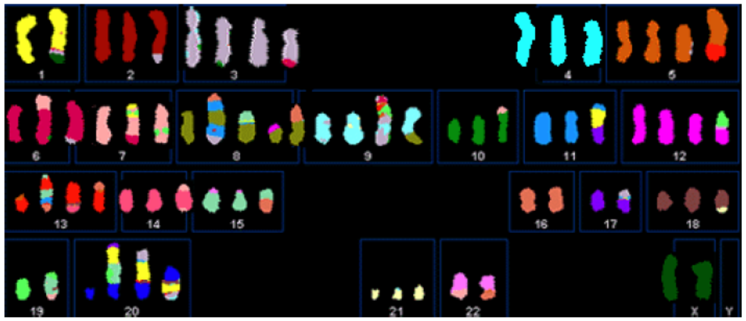
Examples of “Computation for Biology”

- Study cancer genome evolution



Rajagopalan, 2003

Cancer is a Genetic Disease



http://www.path.cam.ac.uk/

SKY Paint of MCF7 Breast Tumor Cell Line

Computation for Biology

Though computation may not solve a biological problem directly, it can help quickly narrow down the search space

Searching a needle in a haystack …

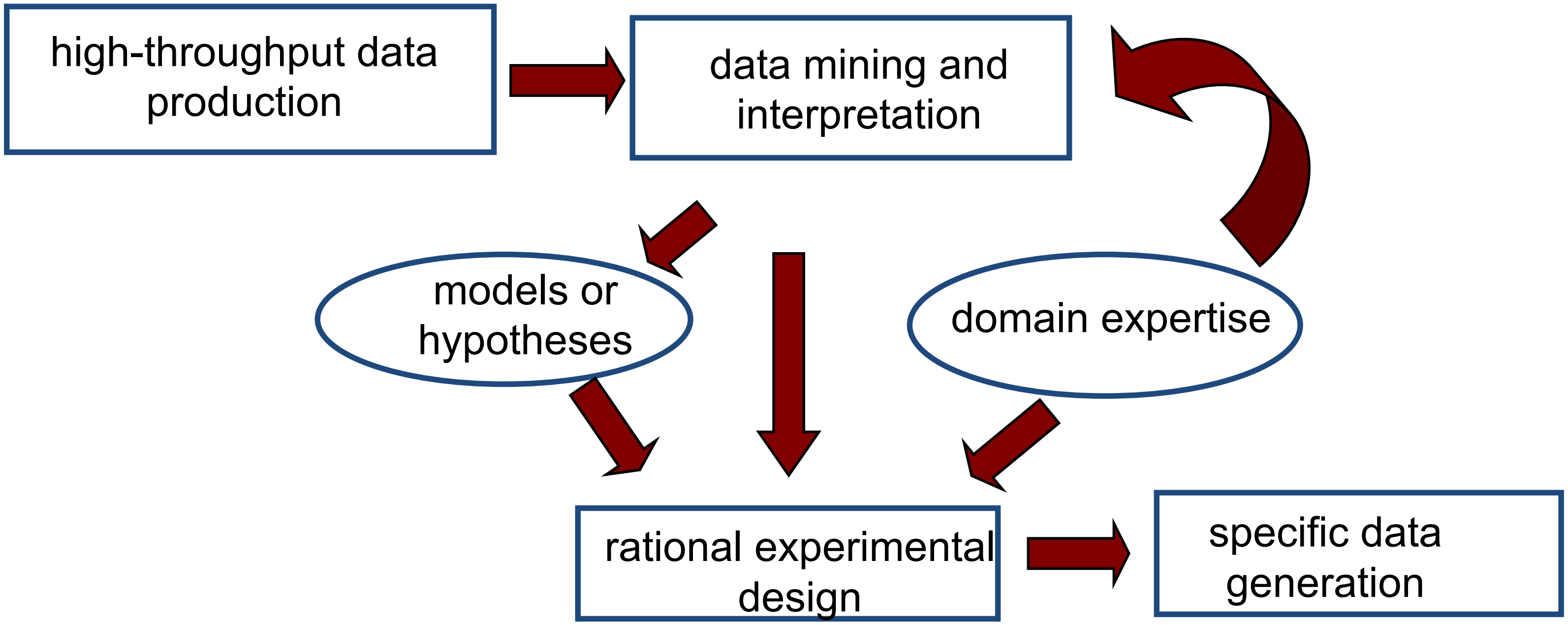
Change of Paradigm in Biology

- The human genome sequencing project has led to fundamental changes in how biological science is done!
  - ❑ It represents biology's first foray into 'big science' – *Science*, editorial, 2003

- The coordinated efforts in "high-throughput" production of biological data beyond sequences have fueled the rapid transition of biology from "*cottage industry science*" to "*big science*"
  - ❑ functional genomic data
  - ❑ structural genomic data
  - ❑ proteomic data
  - ❑ metabolomic data
  - ❑ ……

Change of Paradigm

Data driven discovery



high-throughput data production

data mining and interpretation

models or hypotheses

domain expertise

rational experimental design

specific data generation

Integrative Biology

Interdisciplinary approaches for molecular and cellular life sciences.

"Howdy, want to do biology together?"

"omic" data

Computation for Biology

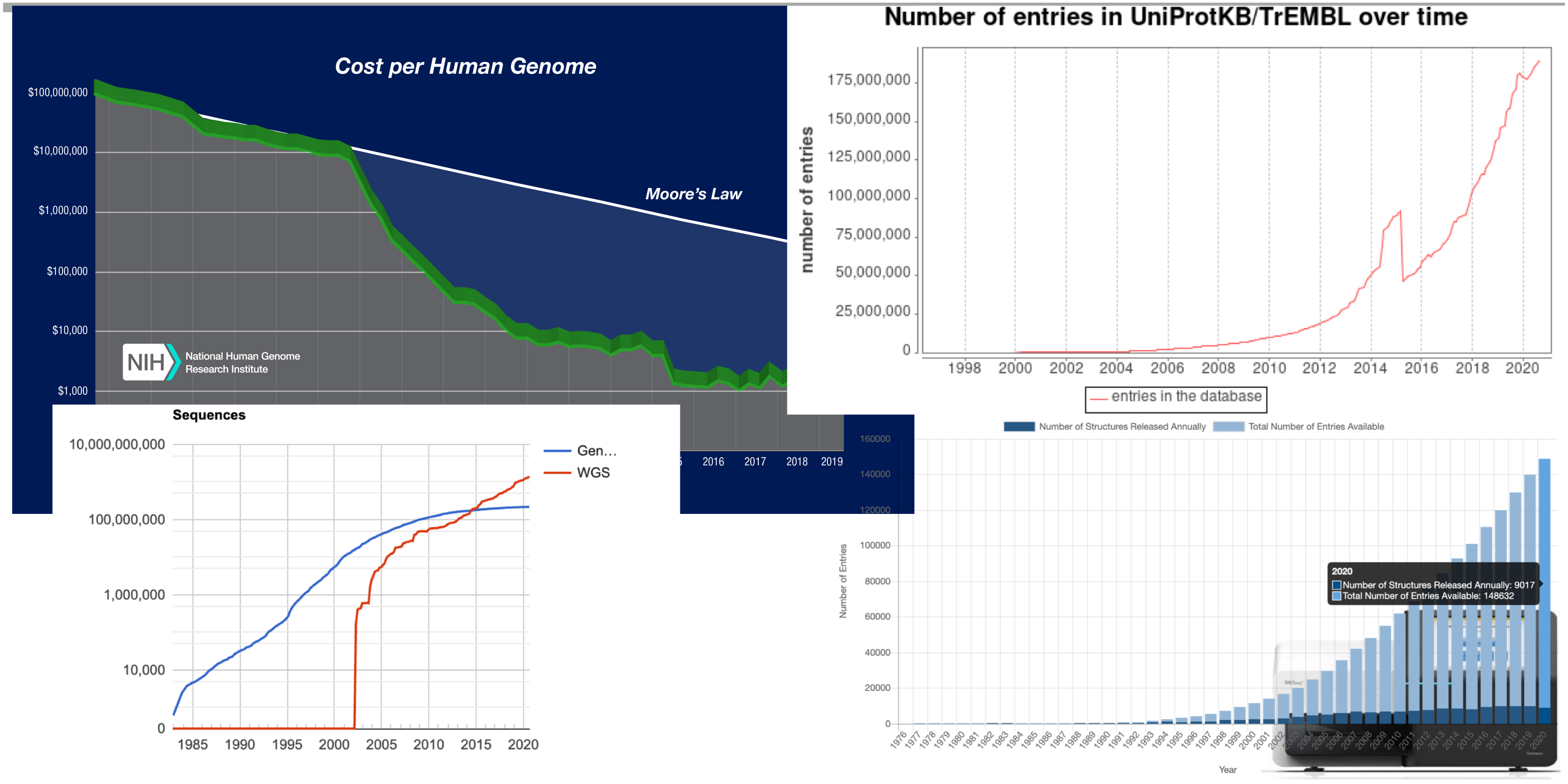
- There are increasingly more successful examples of employing computational techniques to study (or help to study) complex biological problems, in many fronts of biological research

- We begin to see computational techniques with *predictive capabilities* that can help to generate new hypotheses and guide experimental designs

Summary

- The driving force of bioinformatics is biological data production through “high-throughput” technologies
- Computation is becoming increasingly *indispensable* in biological research
- Combination of high-throughput data generation and computation allows scientists to look at more complex biological problems at systems level

Sequencing Is Becoming Much Faster and Cheaper



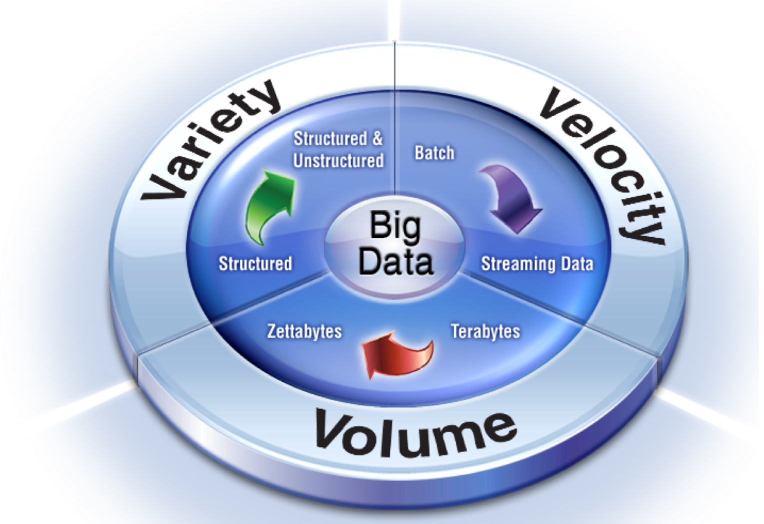
Big Data Challenges

Variety: Complexity of data in many different structures

Velocity: Streaming data and large volume data movement

Volume: Scale from Terabytes to Petabytes (1K TBs) to Zetabytes (1B TBs)

Too big,
too unstructured,
too many different sources

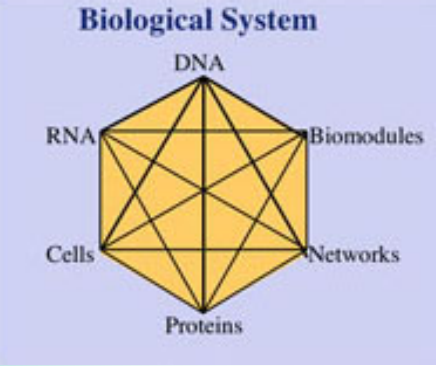


NSF: http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607

health-related data is expected to double every 73 days by 2020

Translational Research & Systems Biology

Genomics



Transcriptomics

Proteomics

Metabolomics

Epigenomics

single elements

Systems Biology

1990 1995 2000 2005 2010 2015 2020

Decoding genomes

integrating elements with a systems approach

systems approach for translational research
(computational systems biology)

Examples of Data-Driven Cancer Research

Genome Study

Which genetic mutations are associated with cancer formation and progression?

How cancer genomes evolve?

Cui J. et al, International J. Cancer 2014

Qin Ma et al. Nucleic Acids Research 2013

Biomarker Discovery

Can we find a gene or protein signature in cancer?

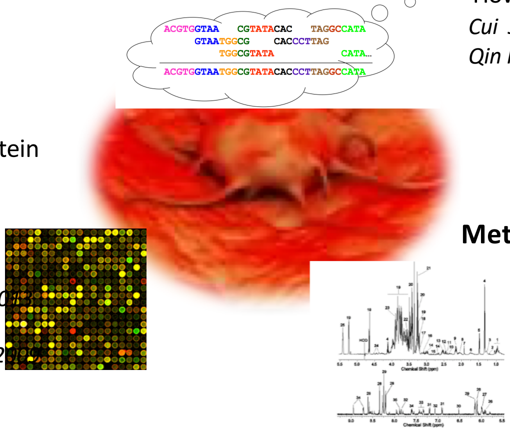
Cui J., et al, Nucleic Acids Res. 2010

Hong S., Cui J., et al, PLoS ONE, 2011

Dong X., et al, Diagnostic pathology , 2013

Cui J., et al, Bioinformatics, 2008

Q Liu, J Cui et. al, BMC bioinformatics, 2009


ACGTGGTAA CGTATACAC TAGGCCATA
GTAATGGCG CACCCTTAG
TGGCGTATA CATA...
ACGTGGTAATGGCGTATACACCCTTAGGCCATA

Metabolic Network

How ATP-production works in cancer?

Cui, et. al., J. Molecular Cell Biology, 2012

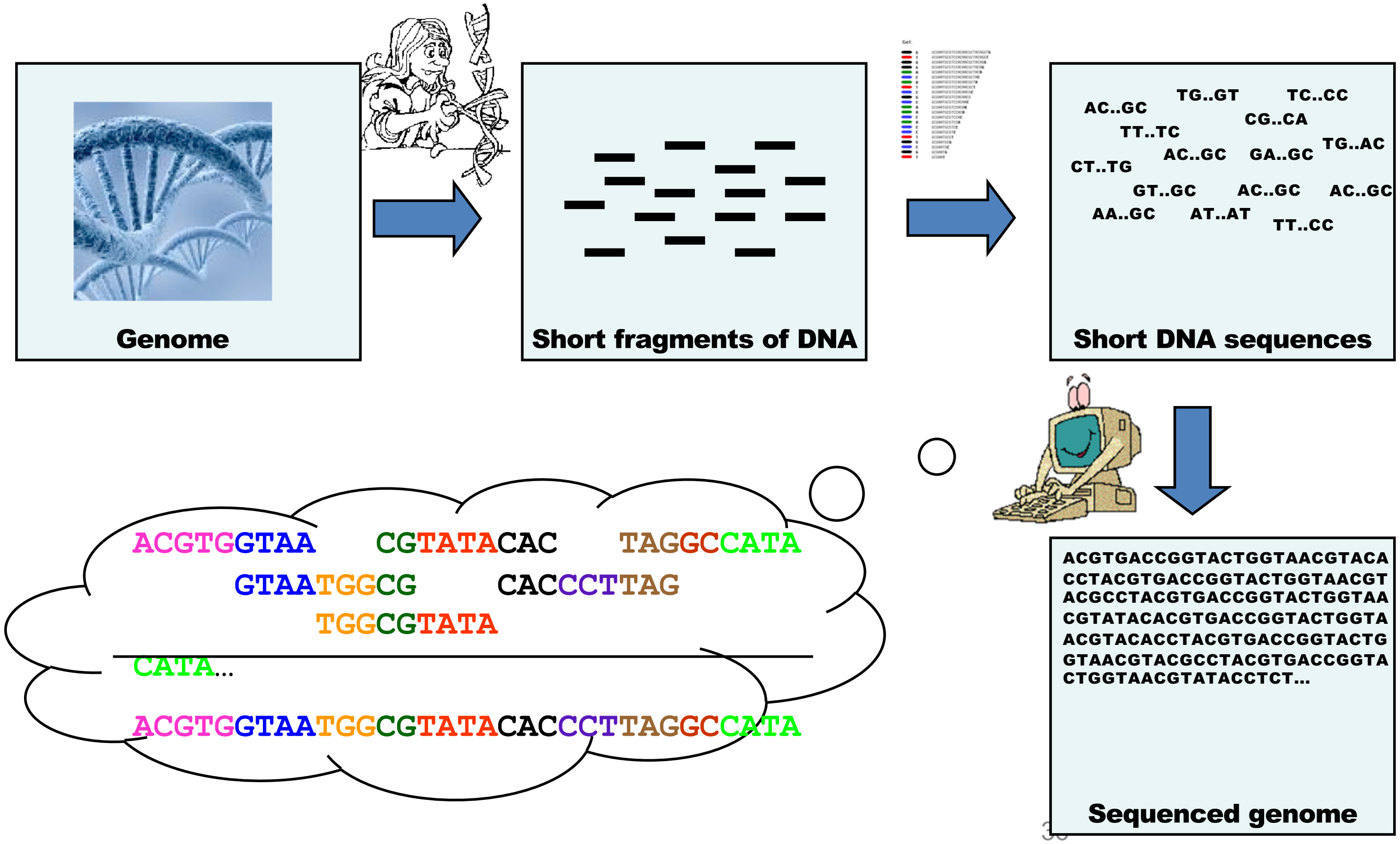
"Utilizing the molecular changes we observed in cancer to make discoveries towards understanding cancer behavior"

Example: Detection of Genetic Mutations in Cancer

Experiments:

- Whole genome sequencing of 5 pairs of gastric cancer and control genomes
- 60 × and ~360M reads/ sample

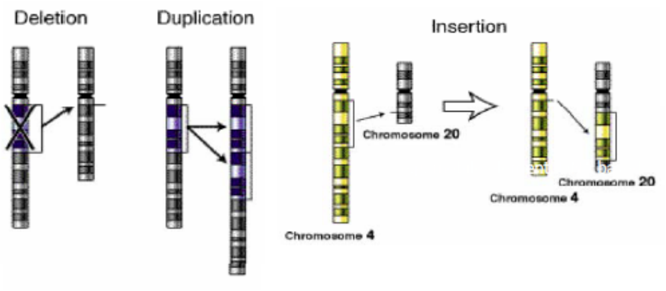
Genome Sequencing

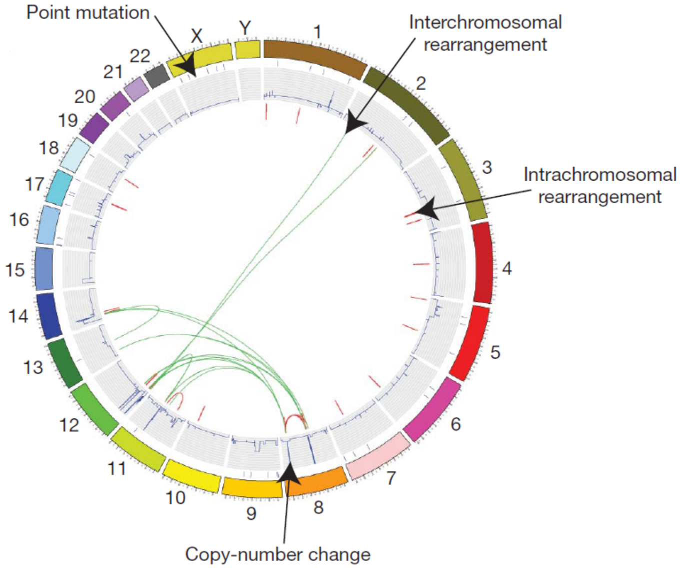


ACGTGACCGGTACTGGTAACGTACA
CCTACGTGACCGGTACTGGTAACGT
ACGCCTACGTGACCGGTACTGGTAA
CGTATACACGTGACCGGTACTGGTA
ACGTACACCTACGTGACCGGTACTG
GTAACGTACGCCTACGTGACCGGTA
CTGGTAACGTATACCTCT...

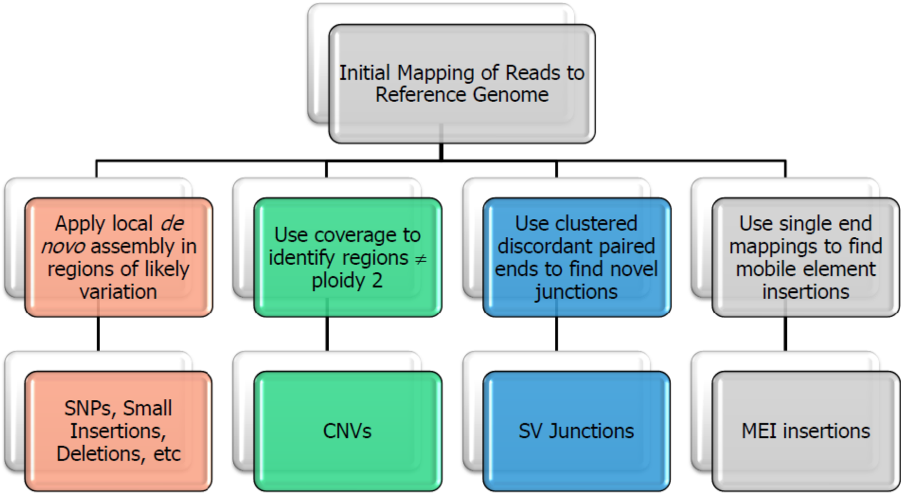
Different Types of Mutations

- Single Nucleotide Polymorphisms (SNP)
- Large structure variations
  - – Insertion/deletions
  - – Duplications
  - – Inversions
  - – Copy number variations
  - – Translocation
  - -…





Local *de novo* Assembly



Initial Mapping of Reads to Reference Genome

Apply local *de novo* assembly in regions of likely variation

Use coverage to identify regions ≠ ploidy 2

Use clustered discordant paired ends to find novel junctions

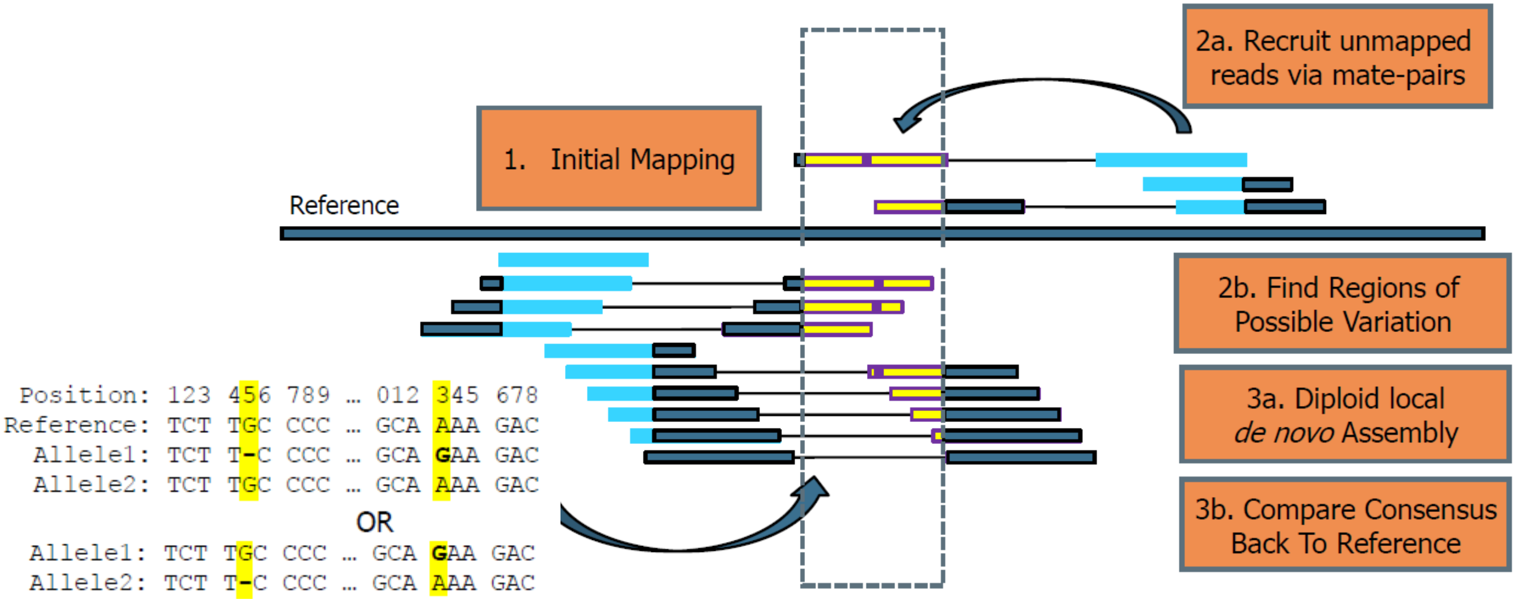
Use single end mappings to find mobile element insertions

SNPs, Small Insertions, Deletions, etc

CNVs

SV Junctions

MEI insertions



1. Initial Mapping

2a. Recruit unmapped reads via mate-pairs

2b. Find Regions of Possible Variation

3a. Diploid local *de novo* Assembly

3b. Compare Consensus Back To Reference

```
Position: 123 456 789 … 012 345 678
Reference: TCT TGC CCC … GCA AAA GAC
  Allele1: TCT T-C CCC … GCA GAA GAC
  Allele2: TCT TGC CCC … GCA AAA GAC
```

OR

```
  Allele1: TCT TGC CCC … GCA GAA GAC
  Allele2: TCT T-C CCC … GCA AAA GAC
```

Somatic mutations

A median of 9,700 somatic point mutations per tumor
315 genes with non-synonymous mutations

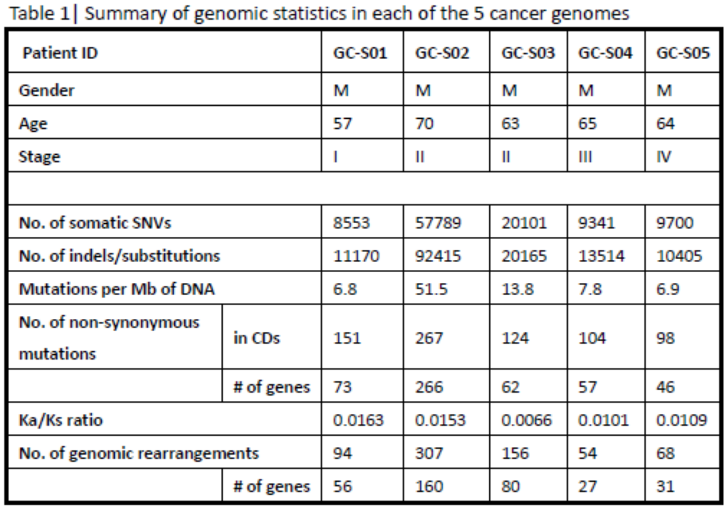
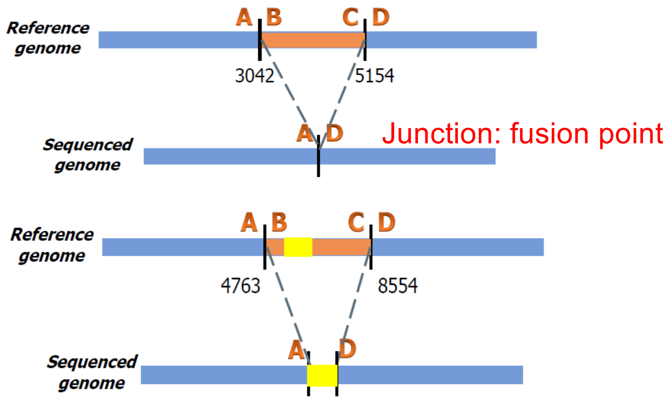


Table 1| Summary of genomic statistics in each of the 5 cancer genomes

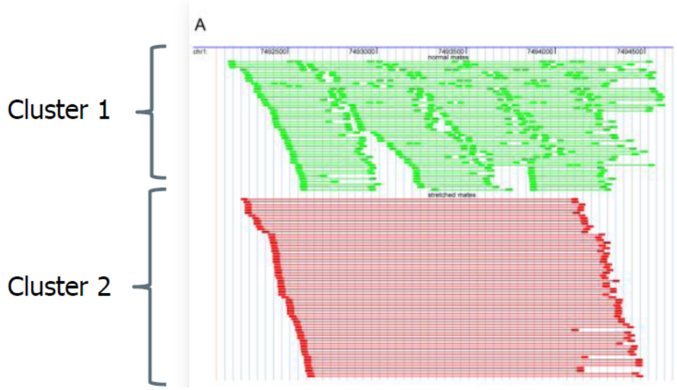
Patient ID		GC-S01	GC-S02	GC-S03	GC-S04	GC-S05
Gender		M	M	M	M	M
Age		57	70	63	65	64
Stage		I	II	II	III	IV
No. of somatic SNVs		8553	57789	20101	9341	9700
No. of indels/substitutions		11170	92415	20165	13514	10405
Mutations per Mb of DNA		6.8	51.5	13.8	7.8	6.9
No. of non-synonymous mutations	in CDs	151	267	124	104	98
	# of genes	73	266	62	57	46
Ka/Ks ratio		0.0163	0.0153	0.0066	0.0101	0.0109
No. of genomic rearrangements		94	307	156	54	68
	# of genes	56	160	80	27	31

Structure Variations

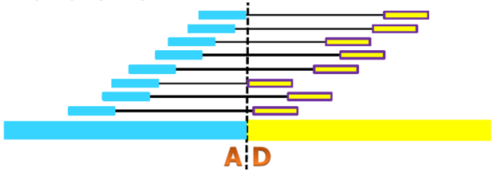


679 rearrangements were identified based 747 junctions;
Recurrent deletions in **tumor suppressors** (FHIT, DACH2 and WWOX) and **Oncogenes** (EGFR)

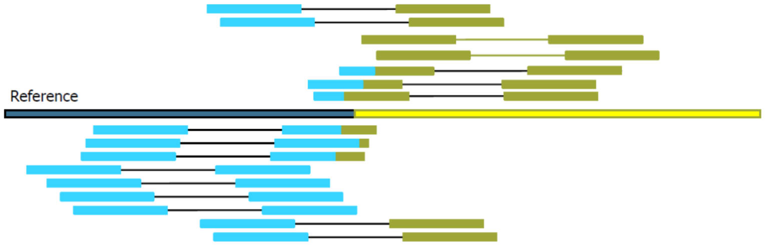
Clustering



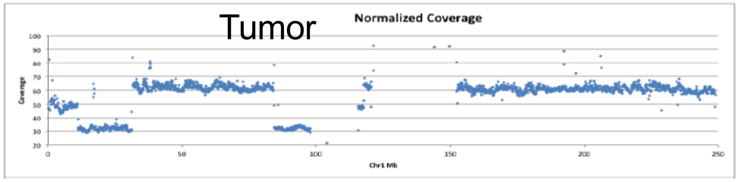
Find cluster

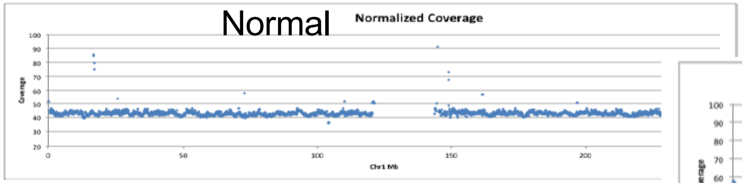


Junction assembly



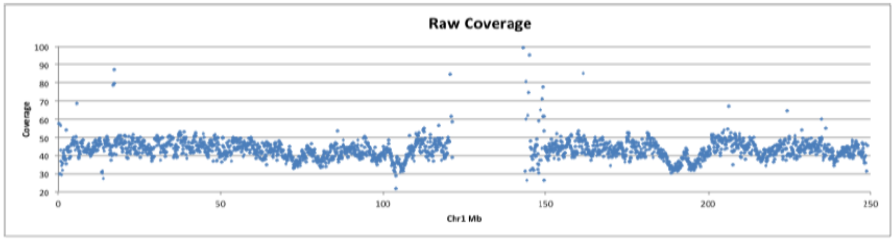
Copy Number Variations

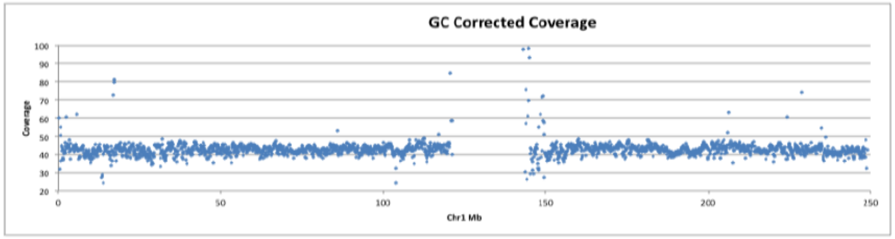


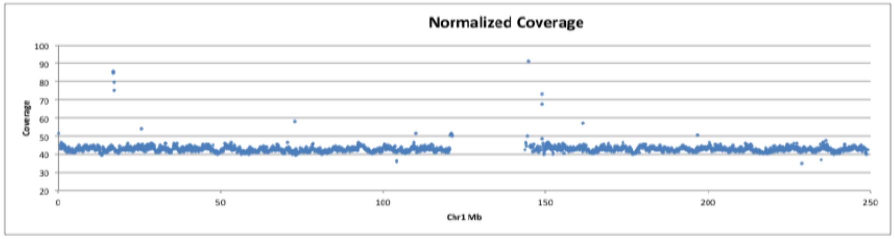


Deletions		Amplification			
KPNA6		RLF		IL11RA	
NKD1		ARAF		LRRIQ1	
C1QL2		C8orf34		MIR4469	
DAZ1		RFC3		MTF2	
DAZ2		ABCA8		MTMR1	
DAZ3		ALCAM		N4BP2L1	
DAZ4		ARID2		NLGN1	
DPF3		C8orf40		POLB	
ECHS1		CCDC102A		POTEA	
GOLGA4		CCL19		PRKG1	
HSD17B11		CCL21		RFC3	
MOB2		CCL27		RGPD1	
NDUFV3		CDH18		RGPD2	
NFIB		CHRNA6		RGS22	
NLGN4Y		CHRNB3		RIMS1	
PRRX2		CNTNAP2		RNF170	
SALL3		DKK4		SFI1	
UNC5A		DLG2		SGK196	
UTY		DNAJB5		SI	
		EHD2		SLAIN1	
		ERC2		SLC20A2	
		FAM205A		SLC9A7	
		FNTA		STK3	
		GALT		THAP1	
		GRIA4		UBQLN2	
		HACE1		UNC13B	
		HGSNAT		VDAC3	
		HOOK3		WFDC10B	
		IKBKB			

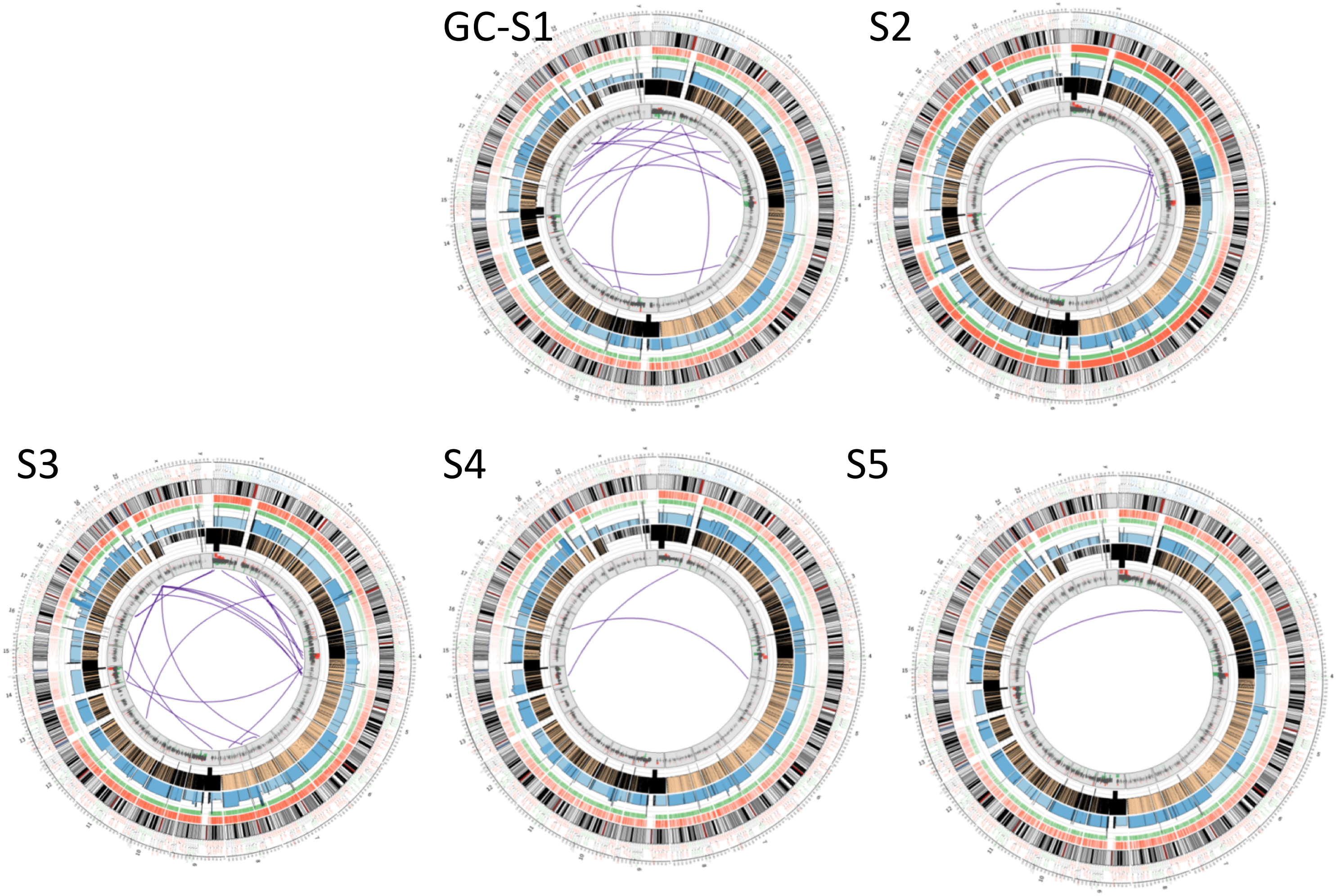
Preprocess







Mutations in Five Gastric Tumors



Current Research Focus at SBBI Lab

Integrated Machine Learning and Stochastic Modeling for Understanding Disease-related Cell Signaling and Regulation

Examples from SBBI Lab

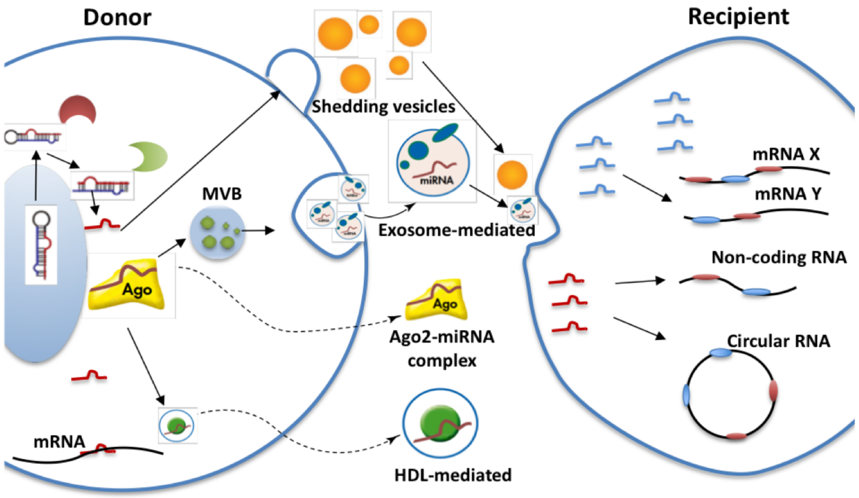
- Molecular characterization of human cell-cell communication via microRNA regulation
  - ❑ To uncover mechanisms underlying *RNA sorting and regulation*

- Elucidation of disease-associated gene regulation
  - ❑ To understand multifaceted gene regulation networks

- Analysis of energy metabolism in humans
  - ❑ To monitor *glucose and energy production through computational modeling*

MicroRNAs

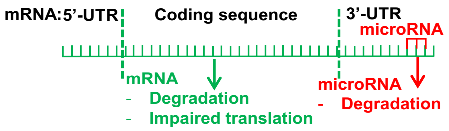
- Small non-coding RNAs, ~21nt long
- Important in post-transcriptional gene silencing in eukaryotes
- Functional study is largely based on the reliable identification of gene targets
  - Current predictions reply on <u>sequence and structural features</u>
  - <u>Cooperative and competitive binding</u> is not well-characterize



Schematic diagram of miRNA transfer between cells and competitive miRNA binding

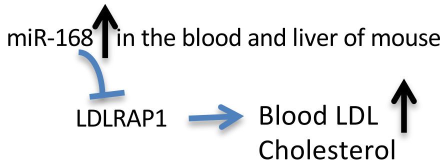
Static

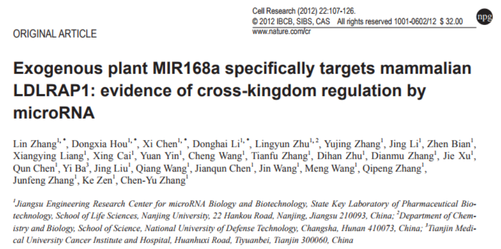
Dynamic



Dietary MicroRNAs

- Plant borne microRNAs
  - ❑ Rice: miR-168 (Zhang, *Cell Research*, 2012)





ORIGINAL ARTICLE

Cell Research (2012) 22:107-126.
© 2012 IBCB, SIBS, CAS All rights reserved 1001-0602/12 $ 32.00
www.nature.com/cr

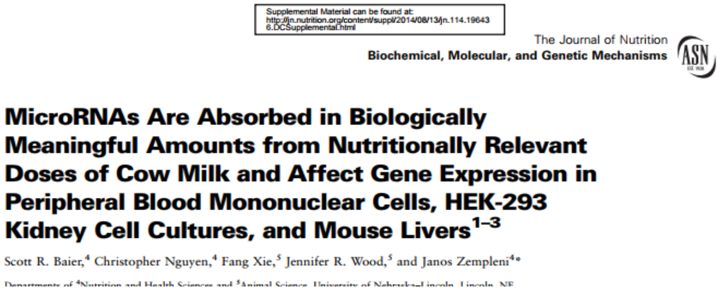
Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA

Lin Zhang[1,*], Dongxia Hou[1,*], Xi Chen[1,*], Donghai Li[1,*], Lingyun Zhu[1,2], Yujing Zhang[1], Jing Li[1], Zhen Bian[1], Xiangying Liang[1], Xing Cai[1], Yuan Yin[1], Cheng Wang[1], Tianfu Zhang[1], Dihan Zhu[1], Dianmu Zhang[1], Jie Xu[1], Qun Chen[1], Yi Ba[3], Jing Liu[1], Qiang Wang[1], Jianqun Chen[1], Jin Wang[1], Meng Wang[1], Qipeng Zhang[1], Junfeng Zhang[1], Ke Zen[1], Chen-Yu Zhang[1]

[1]Jiangsu Engineering Research Center for microRNA Biology and Biotechnology, State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, 22 Hankou Road, Nanjing, Jiangsu 210093, China; [2]Department of Chemistry and Biology, School of Science, National University of Defense Technology, Changsha, Hunan 410073, China; [3]Tianjin Medical University Cancer Institute and Hospital, Huanhuxi Road, Tiyuanbei, Tianjin 300060, China

  - ❑ Honeysuckle: miR-2911
    - ❖ 48 hours stay in mouse serum and urine (Yang, Cell Research, 2015)

- Animal borne microRNAs
  - ❑ Cow milk: miR-29b and -200c (Baier et. al. NJ, 2014)
  - ❑ Chicken egg: miR-181a/b (Howard, K. 2015) found in human plasma



Supplemental Material can be found at:
http://jn.nutrition.org/content/suppl/2014/08/13/jn.114.196436.DCSupplemental.html

The Journal of Nutrition
Biochemical, Molecular, and Genetic Mechanisms

MicroRNAs Are Absorbed in Biologically Meaningful Amounts from Nutritionally Relevant Doses of Cow Milk and Affect Gene Expression in Peripheral Blood Mononuclear Cells, HEK-293 Kidney Cell Cultures, and Mouse Livers[1–3]

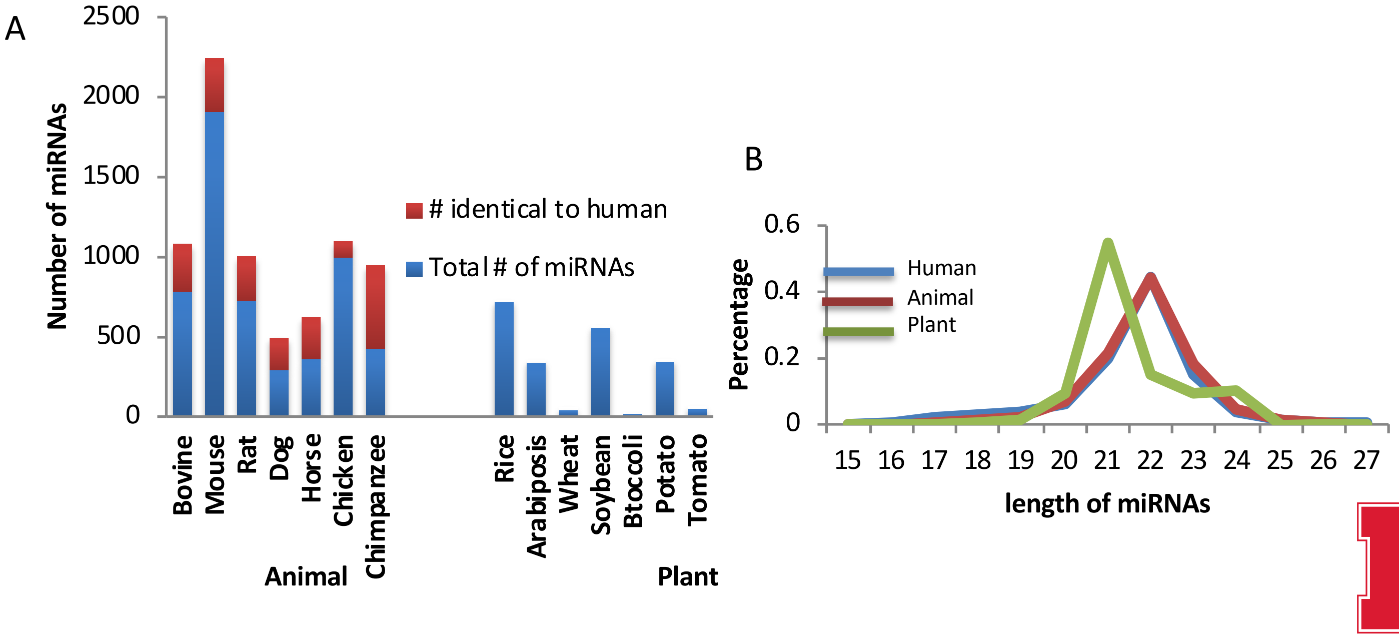
Scott R. Baier,[4] Christopher Nguyen,[4] Fang Xie,[5] Jennifer R. Wood,[5] and Janos Zempleni[4*]

Departments of [4]Nutrition and Health Sciences and [5]Animal Science, University of Nebraska–Lincoln, Lincoln, NE

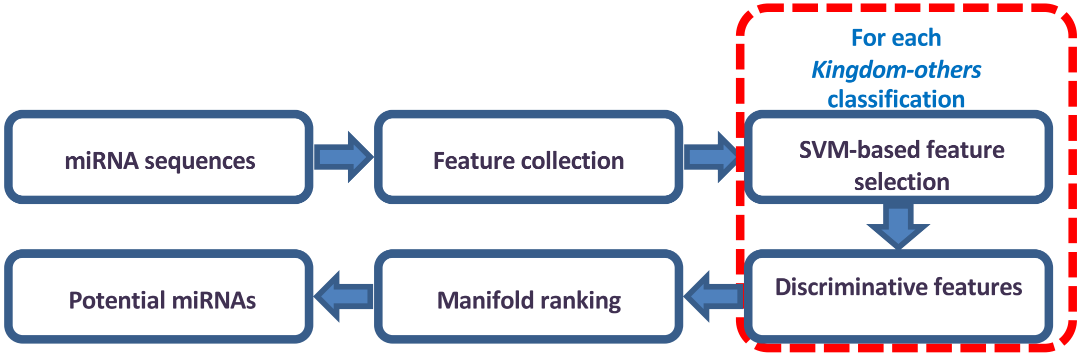
Molecular Features of MicroRNAs

➢ Comparative analysis to identify sequence features that contribute to microRNA secretion

- All mature microRNA sequences from 5 kingdoms (*miRBase)*
- Human circulating microRNAs
- Dietary microRNAs
- 1200+ features



Bioinformatics Workflow: Feature Selection and Prediction



For each *Kingdom-others* classification

miRNA sequences

Feature collection

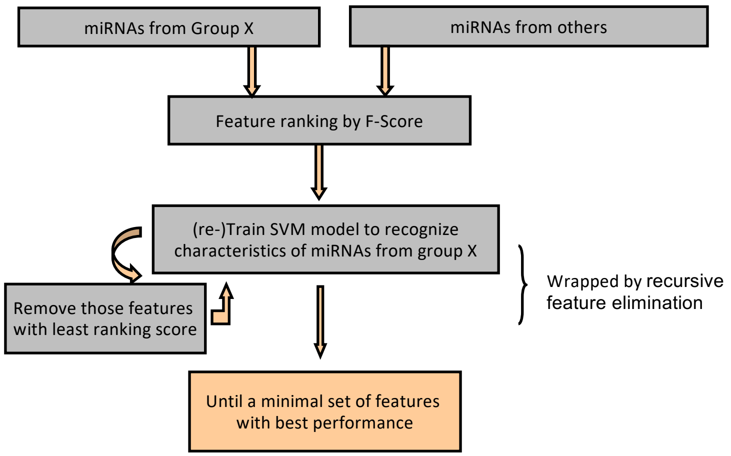
SVM-based feature selection

Discriminative features

Manifold ranking

Potential miRNAs

- Identify a set of most discriminative features



miRNAs from Group X

miRNAs from others

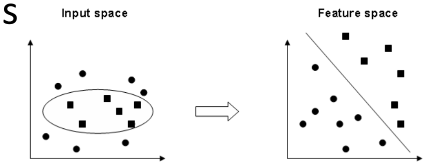
Feature ranking by F-Score

(re-)Train SVM model to recognize characteristics of miRNAs from group X

Remove those features with least ranking score

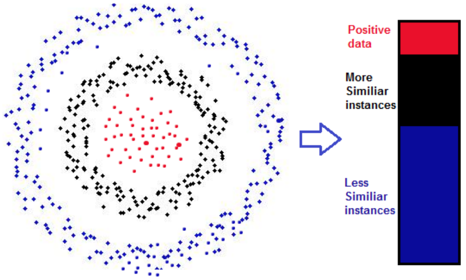
Wrapped by recursive feature elimination

Until a minimal set of features with best performance

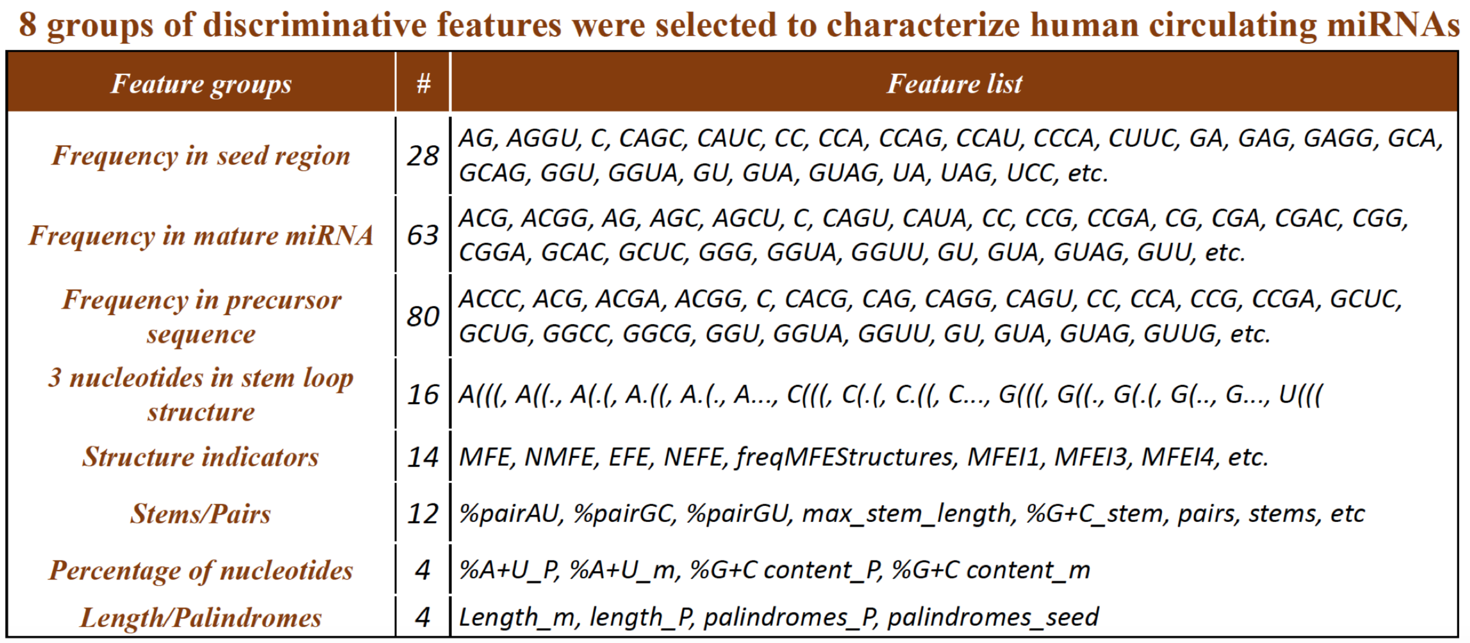


$$F(i) \equiv \frac{\left(\bar{\boldsymbol{x}}_i^{(+)} - \bar{\boldsymbol{x}}_i\right)^2 + \left(\bar{\boldsymbol{x}}_i^{(-)} - \bar{\boldsymbol{x}}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{\boldsymbol{x}}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{\boldsymbol{x}}_i^{(-)}\right)^2}$$

F-Score: Chen *et al.* Feature extraction, 2006.



Distinguishing Molecular Features



8 groups of discriminative features were selected to characterize human circulating miRNAs

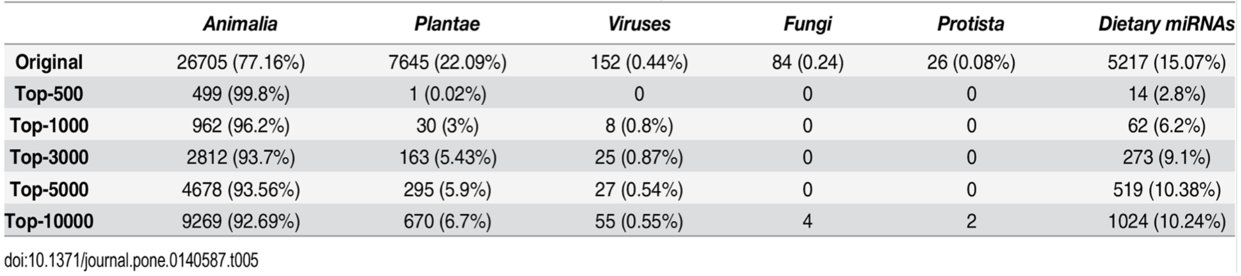
Feature groups	#	*Feature list*
Frequency in seed region	28	*AG, AGGU, C, CAGC, CAUC, CC, CCA, CCAG, CCAU, CCCA, CUUC, GA, GAG, GAGG, GCA, GCAG, GGU, GGUA, GU, GUA, GUAG, UA, UAG, UCC, etc.*
Frequency in mature miRNA	63	*ACG, ACGG, AG, AGC, AGCU, C, CAGU, CAUA, CC, CCG, CCGA, CG, CGA, CGAC, CGG, CGGA, GCAC, GCUC, GGG, GGUA, GGUU, GU, GUA, GUAG, GUU, etc.*
Frequency in precursor sequence	80	*ACCC, ACG, ACGA, ACGG, C, CACG, CAG, CAGG, CAGU, CC, CCA, CCG, CCGA, GCUC, GCUG, GGCC, GGCG, GGU, GGUA, GGUU, GU, GUA, GUAG, GUUG, etc.*
3 nucleotides in stem loop structure	16	*A(((, A((., A(.(, A.((, A.(., A..., C(((, C(.(, C.((, C..., G(((, G((., G(.(, G(.., G..., U(((*
Structure indicators	14	*MFE, NMFE, EFE, NEFE, freqMFEStructures, MFEI1, MFEI3, MFEI4, etc.*
Stems/Pairs	12	*%pairAU, %pairGC, %pairGU, max_stem_length, %G+C_stem, pairs, stems, etc*
Percentage of nucleotides	4	*%A+U_P, %A+U_m, %G+C content_P, %G+C content_m*
Length/Palindromes	4	*Length_m, length_P, palindromes_P, palindromes_seed*

Accuracy: 90.0% (Sensitivity: 84.7%; Specifity:95.4%)

Shu and Cui, PLoS ONE, 2015

Prediction of Circulating MicroRNAs

- Final ranking of all microRNAs based on transportability through *Manifold Ranking*



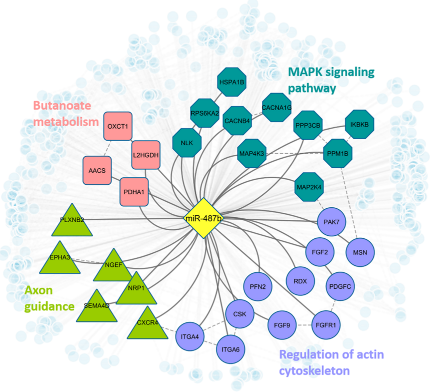
	Animalia	*Plantae*	*Viruses*	*Fungi*	*Protista*	*Dietary miRNAs*
Original	26705 (77.16%)	7645 (22.09%)	152 (0.44%)	84 (0.24)	26 (0.08%)	5217 (15.07%)
Top-500	499 (99.8%)	1 (0.02%)	0	0	0	14 (2.8%)
Top-1000	962 (96.2%)	30 (3%)	8 (0.8%)	0	0	62 (6.2%)
Top-3000	2812 (93.7%)	163 (5.43%)	25 (0.87%)	0	0	273 (9.1%)
Top-5000	4678 (93.56%)	295 (5.9%)	27 (0.54%)	0	0	519 (10.38%)
Top-10000	9269 (92.69%)	670 (6.7%)	55 (0.55%)	4	2	1024 (10.24%)

doi:10.1371/journal.pone.0140587.t005

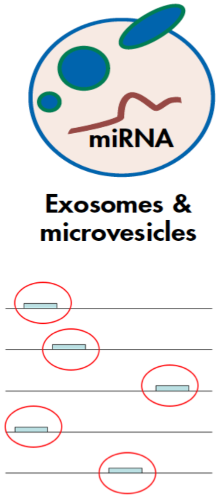
Shu *et al.* Plos One, 2015.

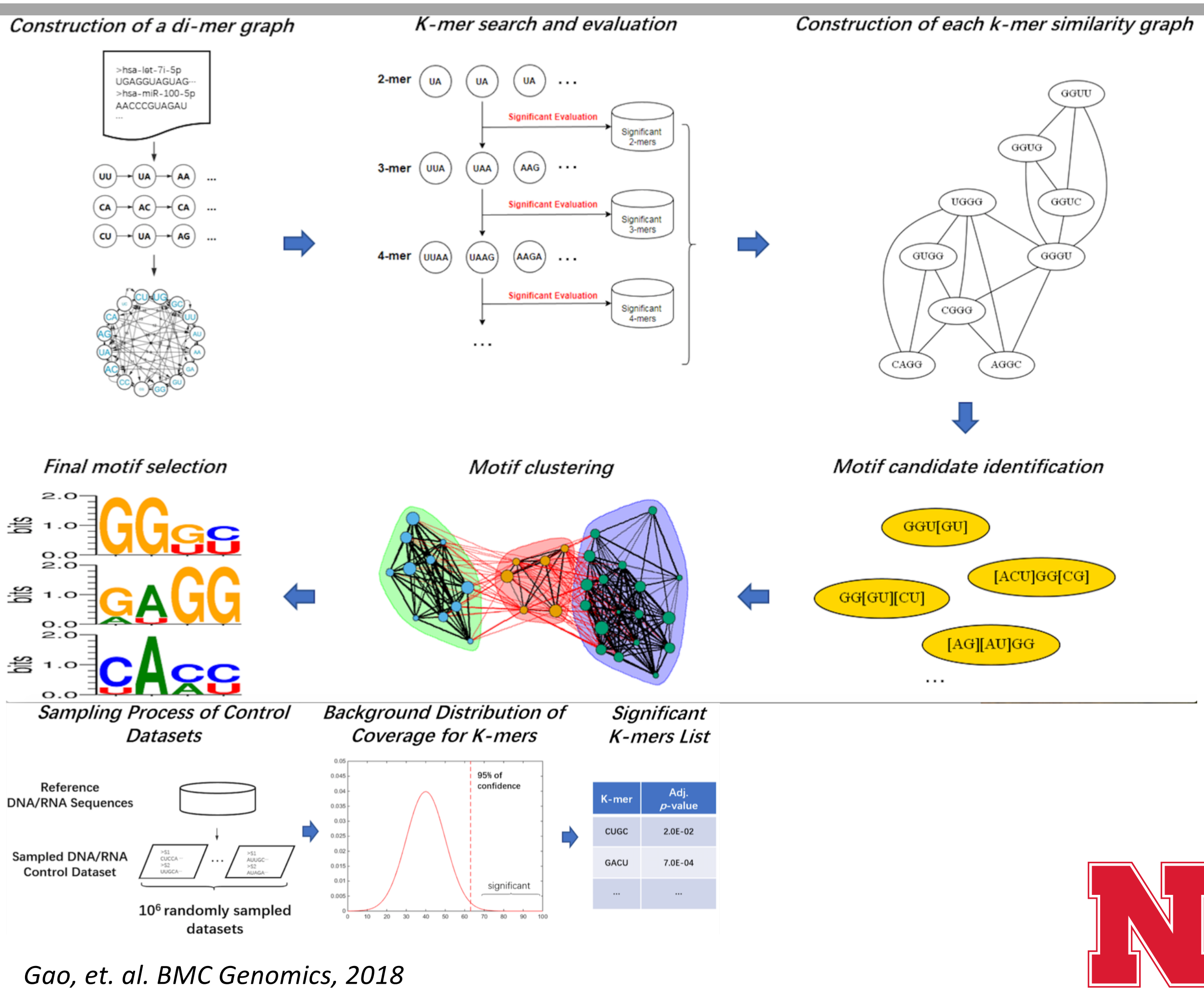
- 345 dietary miRNAs have been identified
  - 9 cow miRNAs have been validated in a cow milk feeding study.
  - Possible functional implication in humans, e.g., bta-miR-487b

- Viral microRNAs validated in pervious studies:
  - ❑ 9 of Epstein–Barr virus (EBV)
  - ❑ 14 of *Rhesus lymphocryptovirus* (rLCV)



Common Sequence Motifs



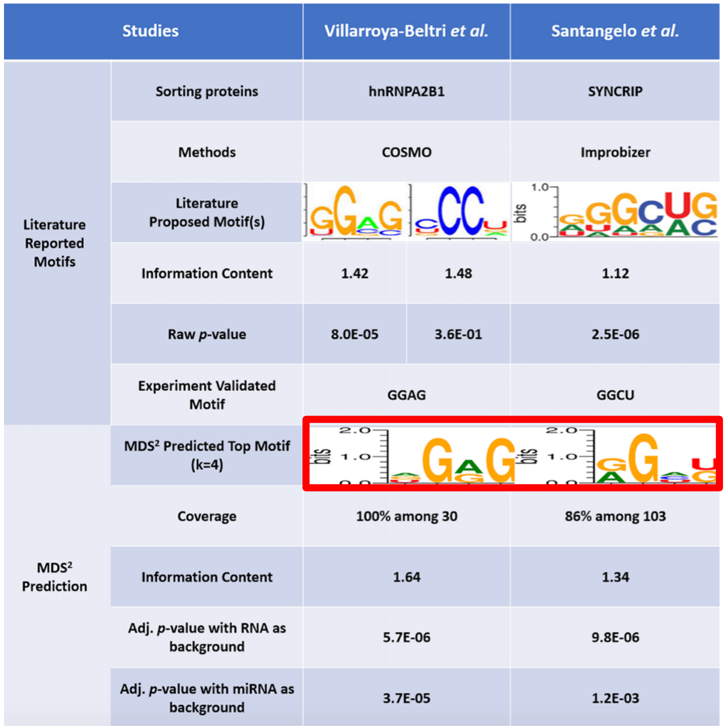


K-mer	Adj. *p*-value
CUGC	2.0E-02
GACU	7.0E-04
...	...

Gao, et. al. BMC Genomics, 2018

Motifs Enriched in Exosomal MicroRNAs

- Comparison with reported motifs



Studies		Villarroya-Beltri *et al.*		Santangelo *et al.*
Literature Reported Motifs	Sorting proteins	hnRNPA2B1		SYNCRIP
	Methods	COSMO		Improbizer
	Literature Proposed Motif(s)			
	Information Content	1.42	1.48	1.12
	Raw *p*-value	8.0E-05	3.6E-01	2.5E-06
	Experiment Validated Motif	GGAG		GGCU
MDS² Prediction	MDS² Predicted Top Motif (k=4)			
	Coverage	100% among 30		86% among 103
	Information Content	1.64		1.34
	Adj. *p*-value with RNA as background	5.7E-06		9.8E-06
	Adj. *p*-value with miRNA as background	3.7E-05		1.2E-03

Example:

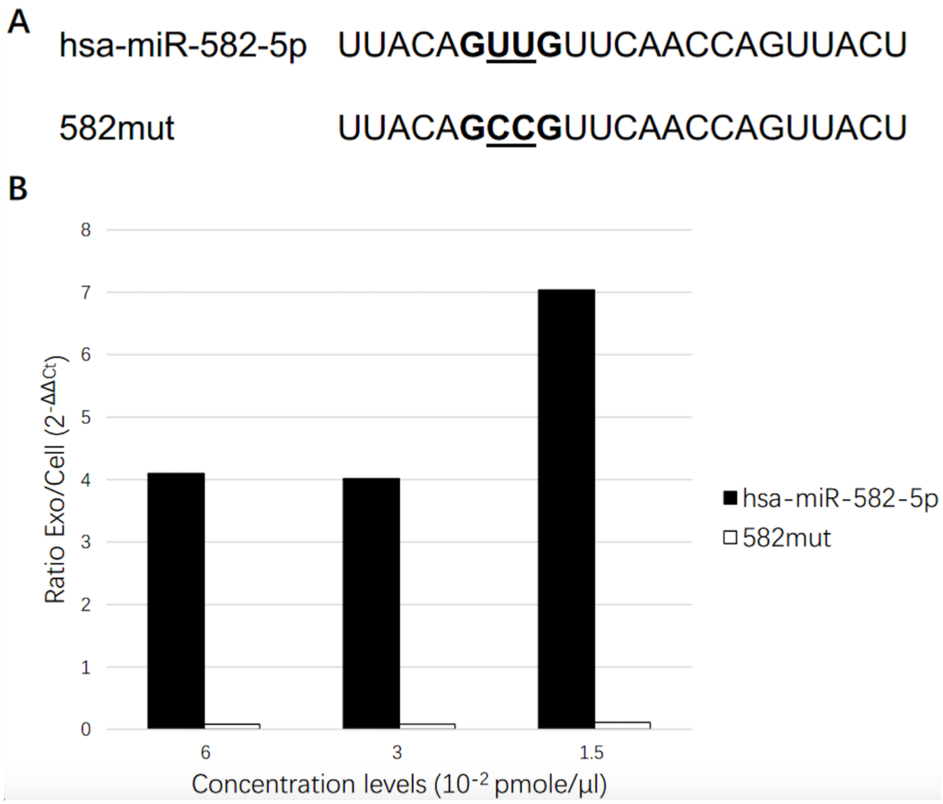
- Motif (**GGAG**) is enriched in exosomal microRNAs secreted from T-cells (Villarroya- Beltri et al., 2013)

- Analysis of exosomal microRNA sequences collected from different resources, >30 groups

Vesiclepedia EVpedia ExoCarta

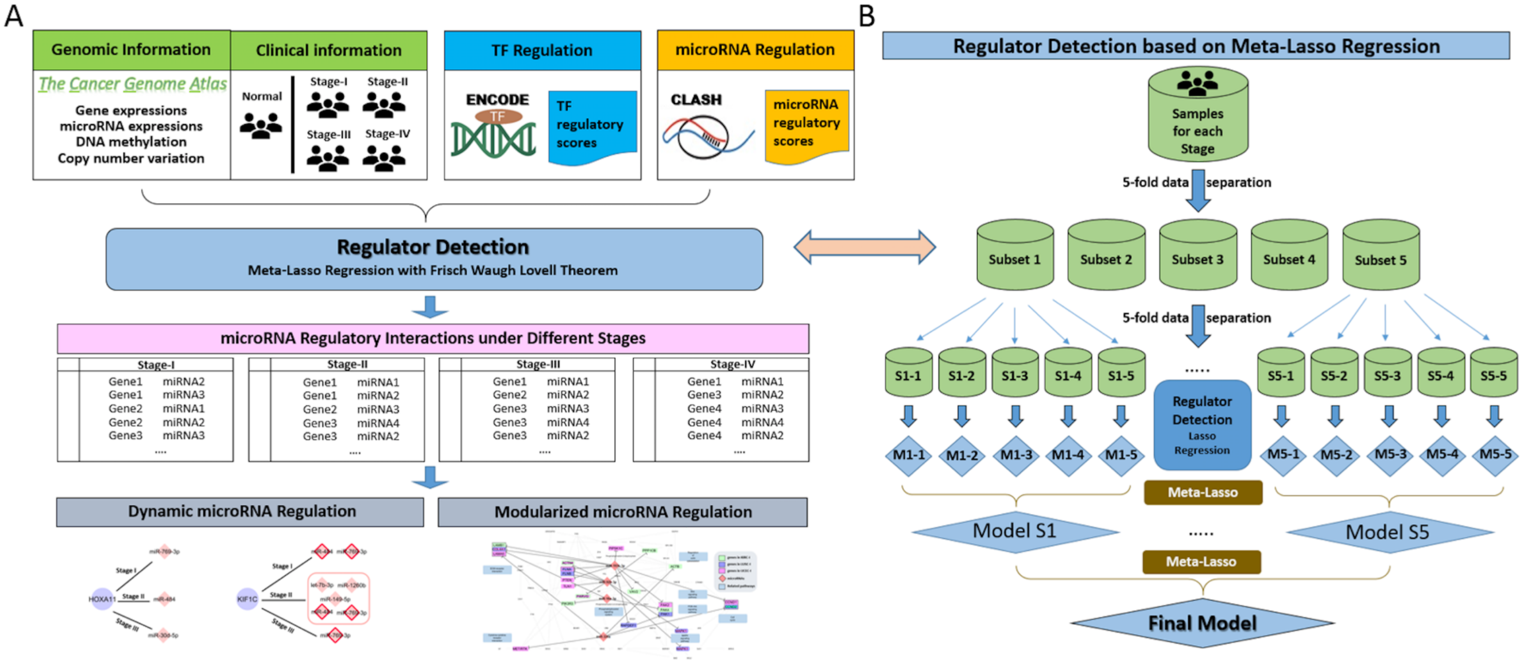
Experimental Validation

SW620 (colon cancer) cell transfection and RT-qPCR test



Gao, et. al. BMC Genomics, 2018

Modeling Dynamic MicroRNA Regulation



Stage-I	
Gene1	miRNA2
Gene1	miRNA3
Gene2	miRNA1
Gene2	miRNA2
Gene3	miRNA3
....	

Stage-II	
Gene1	miRNA1
Gene1	miRNA2
Gene2	miRNA3
Gene3	miRNA4
Gene3	miRNA2
....	

Stage-III	
Gene1	miRNA1
Gene2	miRNA2
Gene3	miRNA3
Gene3	miRNA4
Gene3	miRNA2
....	

Stage-IV	
Gene1	miRNA1
Gene3	miRNA2
Gene4	miRNA3
Gene4	miRNA4
Gene4	miRNA2
....	

(A) The identification pipeline of conditional miRNA regulatory interactions; (B) Meta-Lasso Regression utilized to detect the microRNA regulators of genes in each cancer stage.

$$\max_{\boldsymbol{\beta}_0, \boldsymbol{g}, \boldsymbol{\zeta}} \left\{ \sum_{m=1}^{M} \ell_m(\beta_{m0}, \boldsymbol{g}, \boldsymbol{\zeta}_m) - \sum_{j=1}^{p} |g_j| - \lambda \sum_{j=1}^{p} \sum_{m=1}^{M} |\zeta_{mj}| \right\}$$

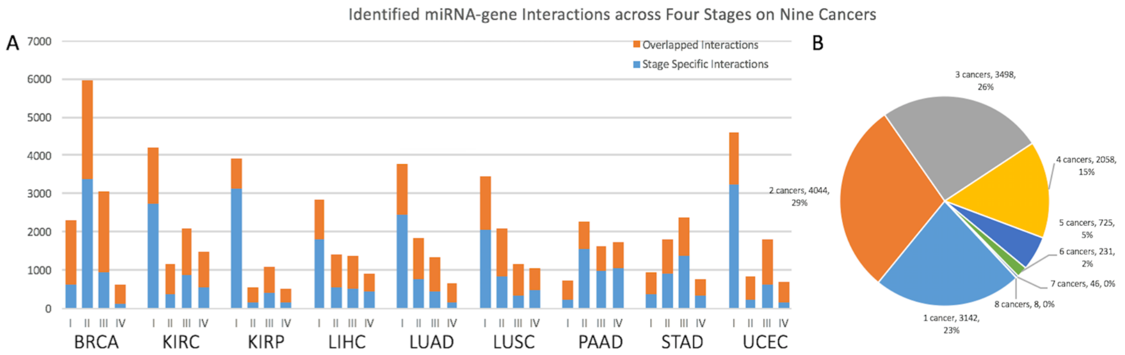
where $l_m\left(\beta_{m_0}, g, \zeta_m\right)$ is the log-likelihood function of the m-th dataset; M denotes the number of individual datasets; g_p is the effect of the p-th regulator (out of P regulators) at the overall condition; and ζ_{mp} is the effect of the p-th regulator at the m-th dataset (out of M datasets).

Shu, et. al., Scientific Reports, 2017

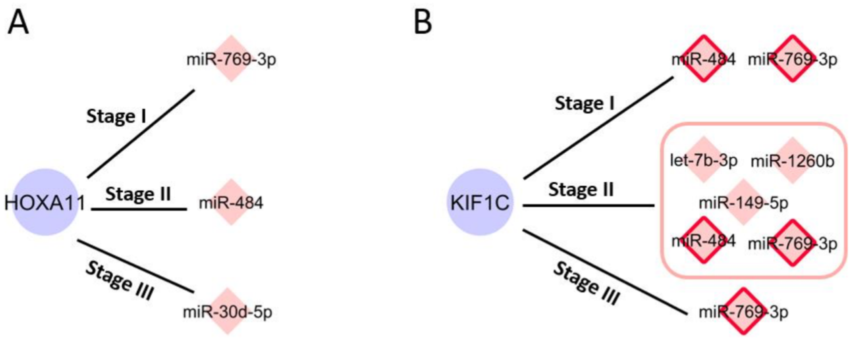
Biometrics. 2014 Dec;70(4):872-80

Conditional Gene Regulation in Human Cancers

- Overview of the miRNA-mRNA interactions identified in nine cancers



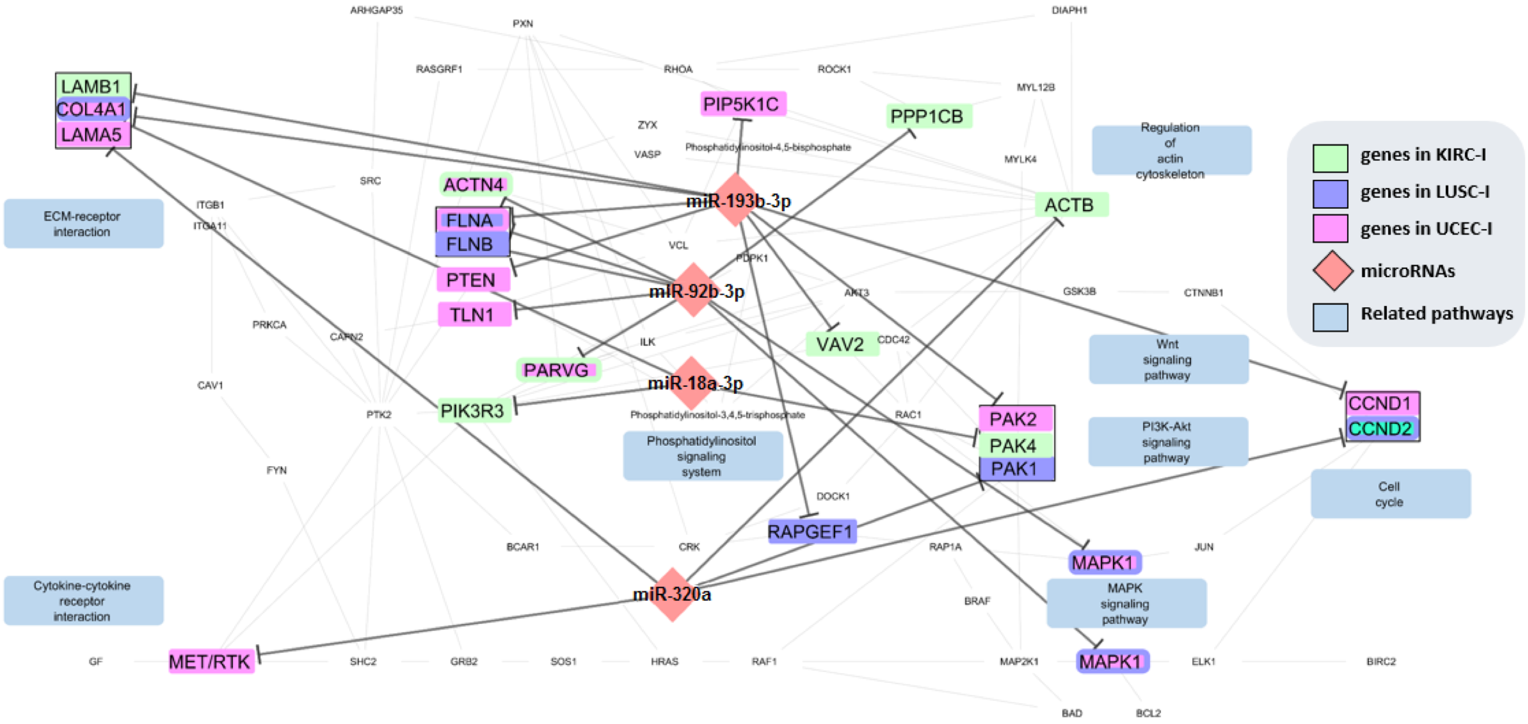
- Illustration of the dynamic miRNA-mediated gene regulation (e.g., kidney cancer)



Shu, et. al. Scientific Reports, 2017

Modularized MicroRNA Regulation in Cancers

Modularized microRNA Regulation in Focal Adhesion Pathway

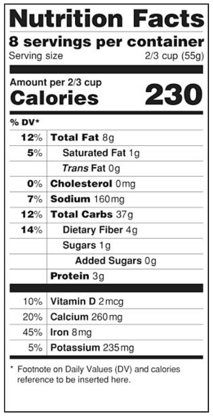


- 4,134 miRNA modules were identified
- 4–5 miRNAs consistently co-regulate the same set of pathways across multiple conditions.

Shu, et. al. Scientific Reports, 2017

Stochastic Modeling of Glucose and Energy Metabolism

Food Intake



Nutrition Facts

8 servings per container

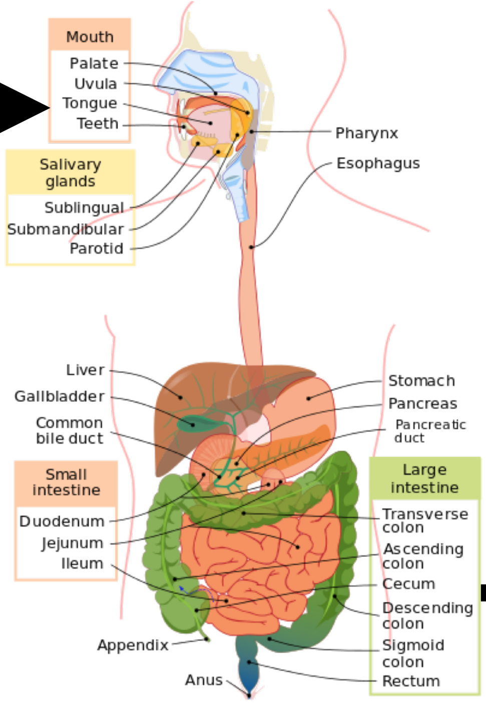
Serving size 2/3 cup (55g)

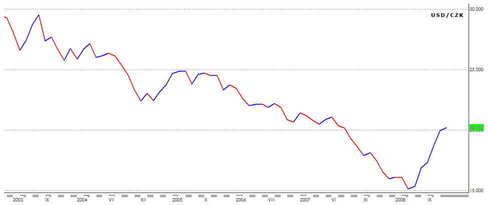
Amount per 2/3 cup

Calories 230

% DV*	
12%	**Total Fat** 8g
5%	Saturated Fat 1g
	Trans Fat 0g
0%	**Cholesterol** 0mg
7%	**Sodium** 160mg
12%	**Total Carbs** 37g
14%	Dietary Fiber 4g
	Sugars 1g
	Added Sugars 0g
	Protein 3g
10%	Vitamin D 2mcg
20%	Calcium 260mg
45%	Iron 8mg
5%	Potassium 235mg

* Footnote on Daily Values (DV) and calories reference to be inserted here.

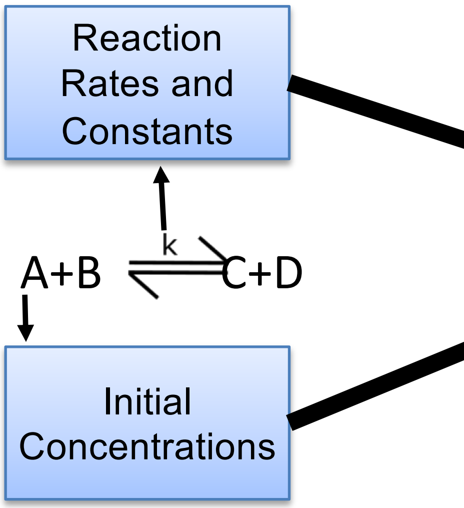


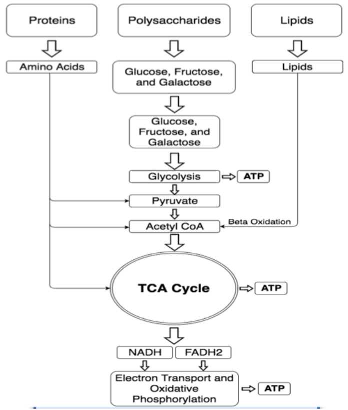


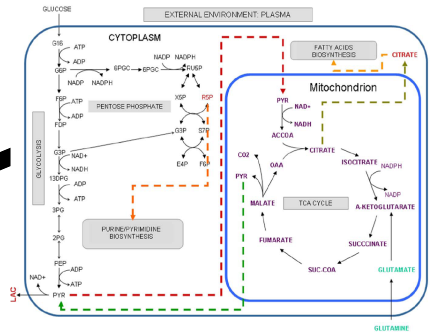
Cellular Glucose, ATP

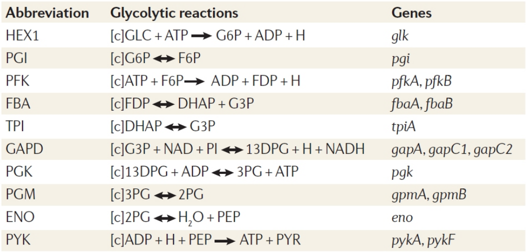
- As per the CDC, 9.4% of US population have diabetes and ~35% are obese.
- Diabetes is a disease where the blood glucose reaches abnormal levels.
- Insulin plays a key role in the regulation of glucose uptake from blood by cells.

The Model

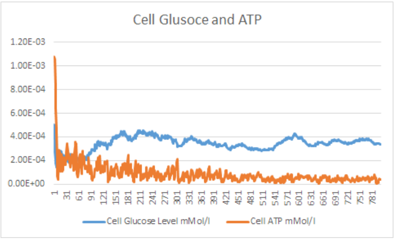




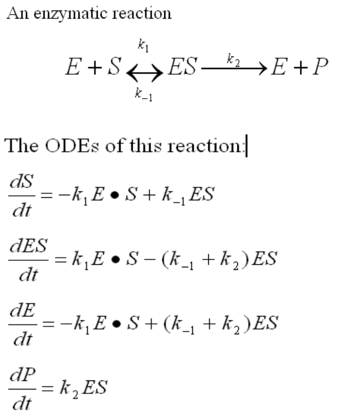




Abbreviation	Glycolytic reactions	Genes
HEX1	[c]GLC + ATP → G6P + ADP + H	*glk*
PGI	[c]G6P ↔ F6P	*pgi*
PFK	[c]ATP + F6P → ADP + FDP + H	*pfkA, pfkB*
FBA	[c]FDP ↔ DHAP + G3P	*fbaA, fbaB*
TPI	[c]DHAP ↔ G3P	*tpiA*
GAPD	[c]G3P + NAD + PI ↔ 13DPG + H + NADH	*gapA, gapC1, gapC2*
PGK	[c]13DPG + ADP ↔ 3PG + ATP	*pgk*
PGM	[c]3PG ↔ 2PG	*gpmA, gpmB*
ENO	[c]2PG ↔ H_2O + PEP	*eno*
PYK	[c]ADP + H + PEP → ATP + PYR	*pykA, pykF*



Step-Wise Concentrations



An enzymatic reaction

$$E + S \underset{k_{-1}}{\overset{k_1}{\longleftrightarrow}} ES \xrightarrow{k_2} E + P$$

The ODEs of this reaction:

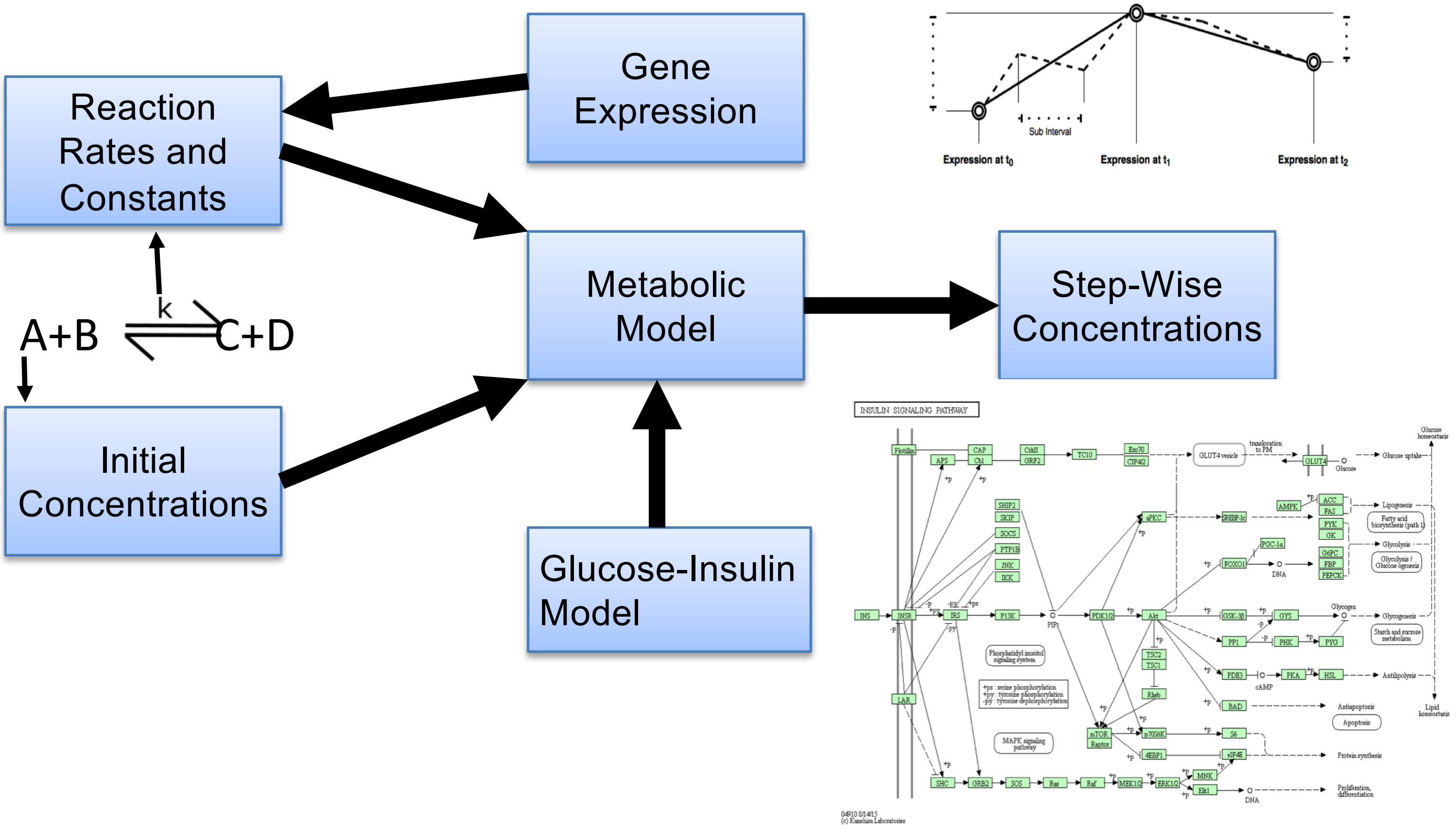
$$\frac{dS}{dt} = -k_1 E \bullet S + k_{-1} ES$$

$$\frac{dES}{dt} = k_1 E \bullet S - (k_{-1} + k_2) ES$$

$$\frac{dE}{dt} = -k_1 E \bullet S + (k_{-1} + k_2) ES$$

$$\frac{dP}{dt} = k_2 ES$$

The Model



Reaction Rates and Constants

Gene Expression

Metabolic Model

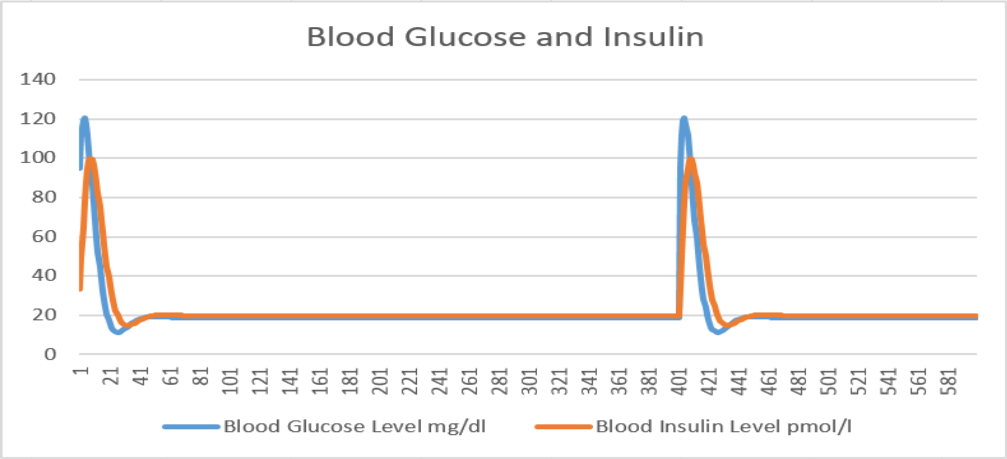
Step-Wise Concentrations

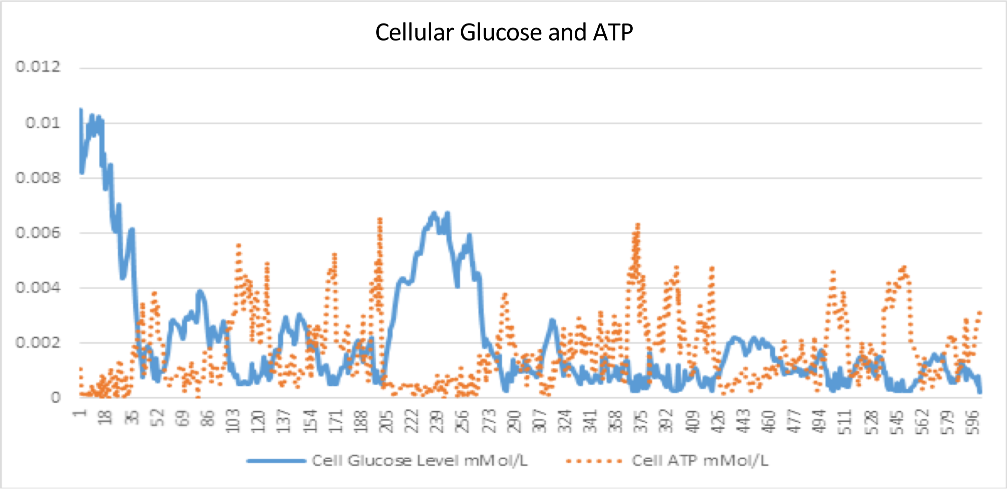
Initial Concentrations

Glucose-Insulin Model

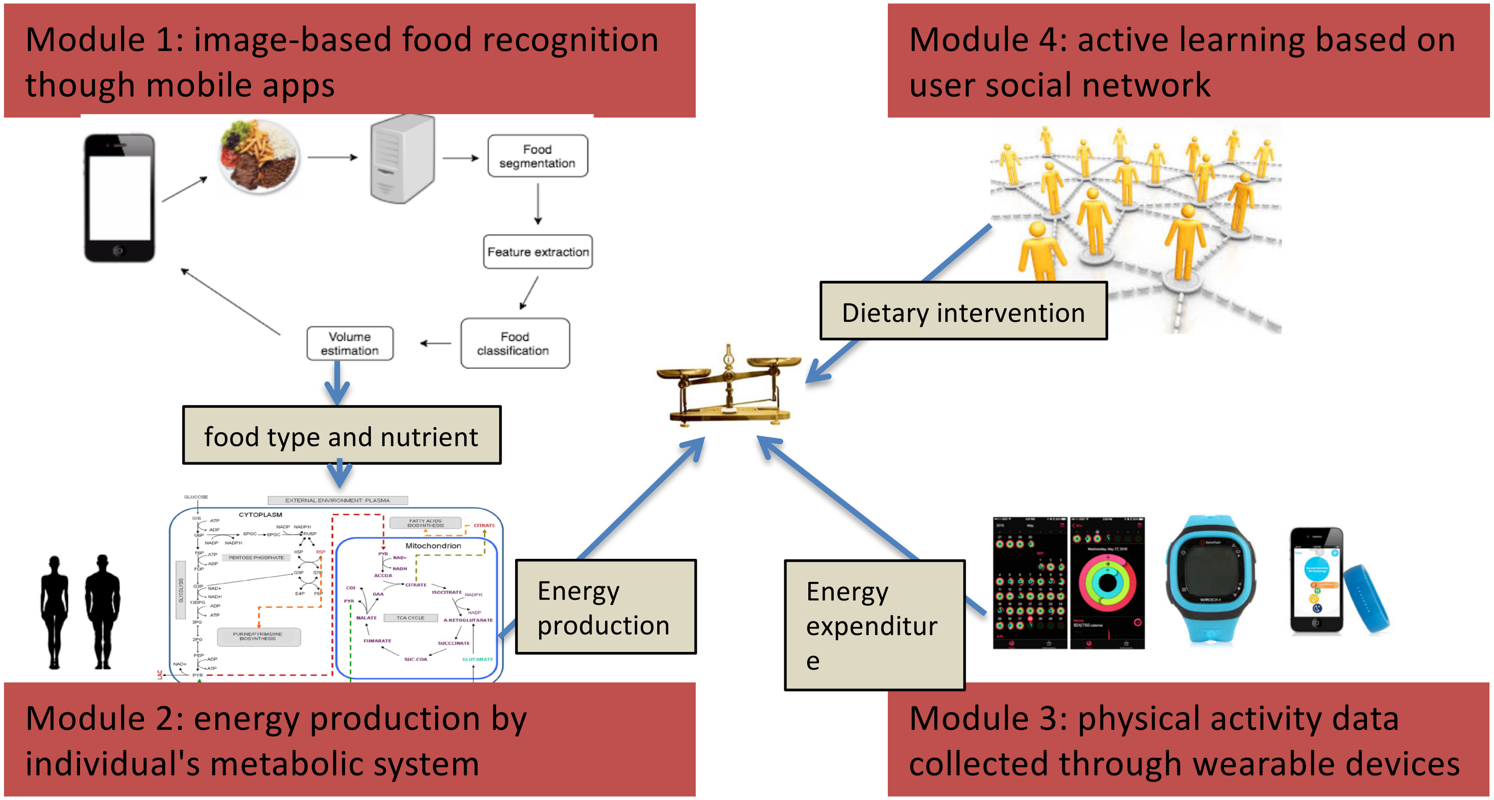
$A+B \overset{k}{\rightleftharpoons} C+D$

Results: Glucose, ATP, and Insulin





A Smart Health System



Module 1: image-based food recognition though mobile apps

Module 4: active learning based on user social network

Dietary intervention

food type and nutrient

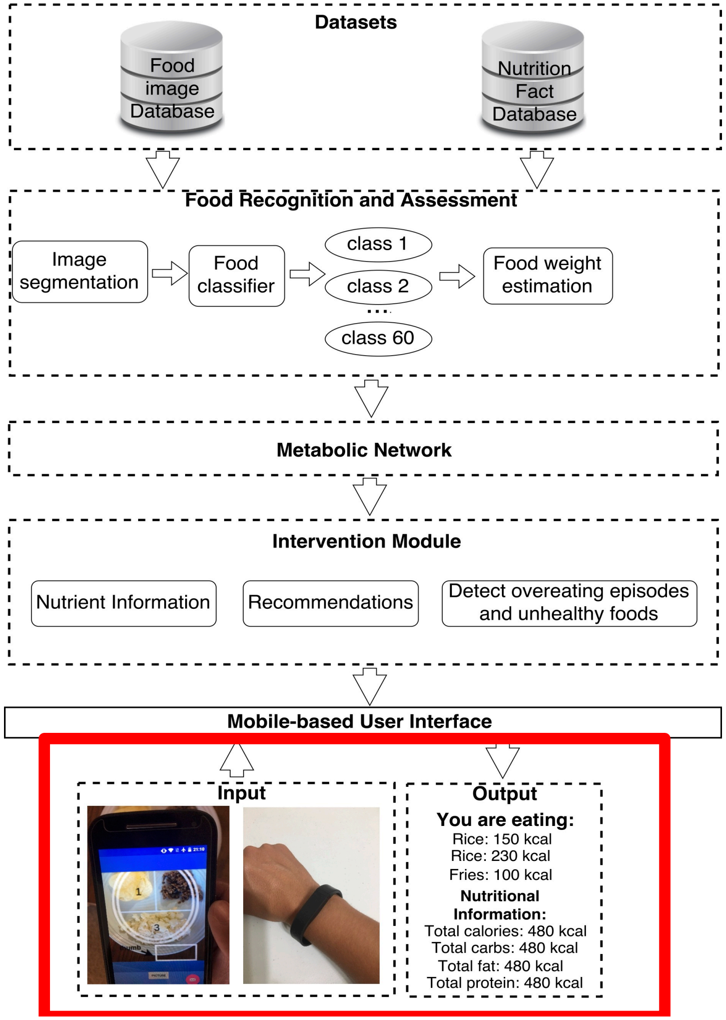
Energy production

Energy expenditur e

Module 2: energy production by individual's metabolic system

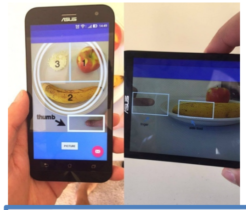
Module 3: physical activity data collected through wearable devices

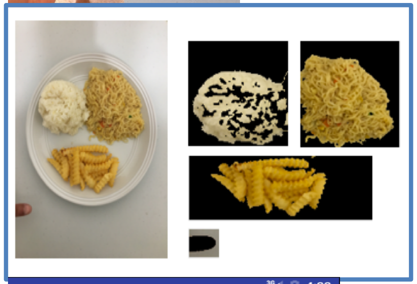
Automated Nutrient Intake Logging Based on Food Images

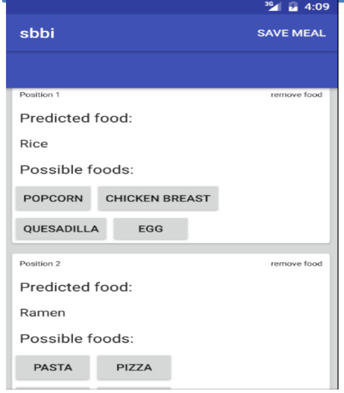


You are eating:
Rice: 150 kcal
Rice: 230 kcal
Fries: 100 kcal

Nutritional Information:
Total calories: 480 kcal
Total carbs: 480 kcal
Total fat: 480 kcal
Total protein: 480 kcal







sbbi SAVE MEAL

Position 1 remove food

Predicted food:

Rice

Possible foods:

POPCORN CHICKEN BREAST

QUESADILLA EGG

Position 2 remove food

Predicted food:

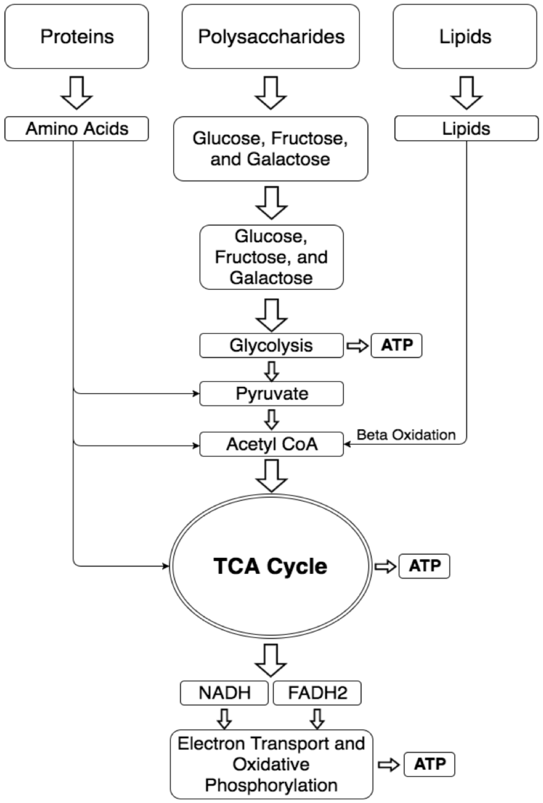
Ramen

Possible foods:

PASTA PIZZA

Silva, et al. , J. of Health and Medical Informatics, 2018

Simulation of Energy Production



Proteins

Polysaccharides

Lipids

Amino Acids

Glucose, Fructose, and Galactose

Lipids

Glucose, Fructose, and Galactose

Glycolysis

ATP

Pyruvate

Acetyl CoA

Beta Oxidation

TCA Cycle

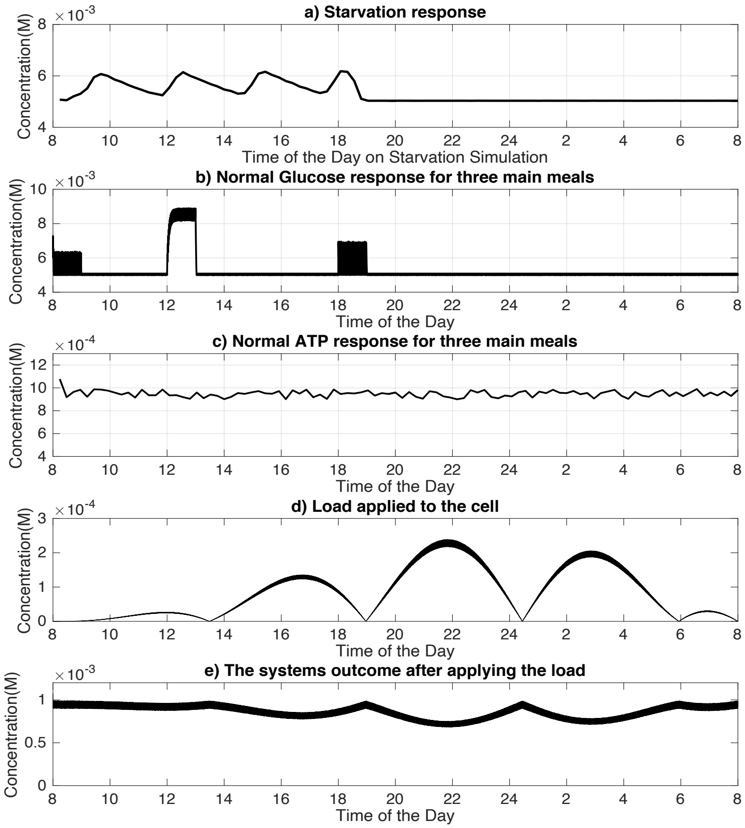
ATP

NADH

FADH2

Electron Transport and Oxidative Phosphorylation

ATP



Summary

- Integration of multi-omics data from various sources is key to understanding important but complex processes biology.

- Various learning models facilitate the mechanistic discoveries in complex human diseases, especially when complete kinetics models are not available.

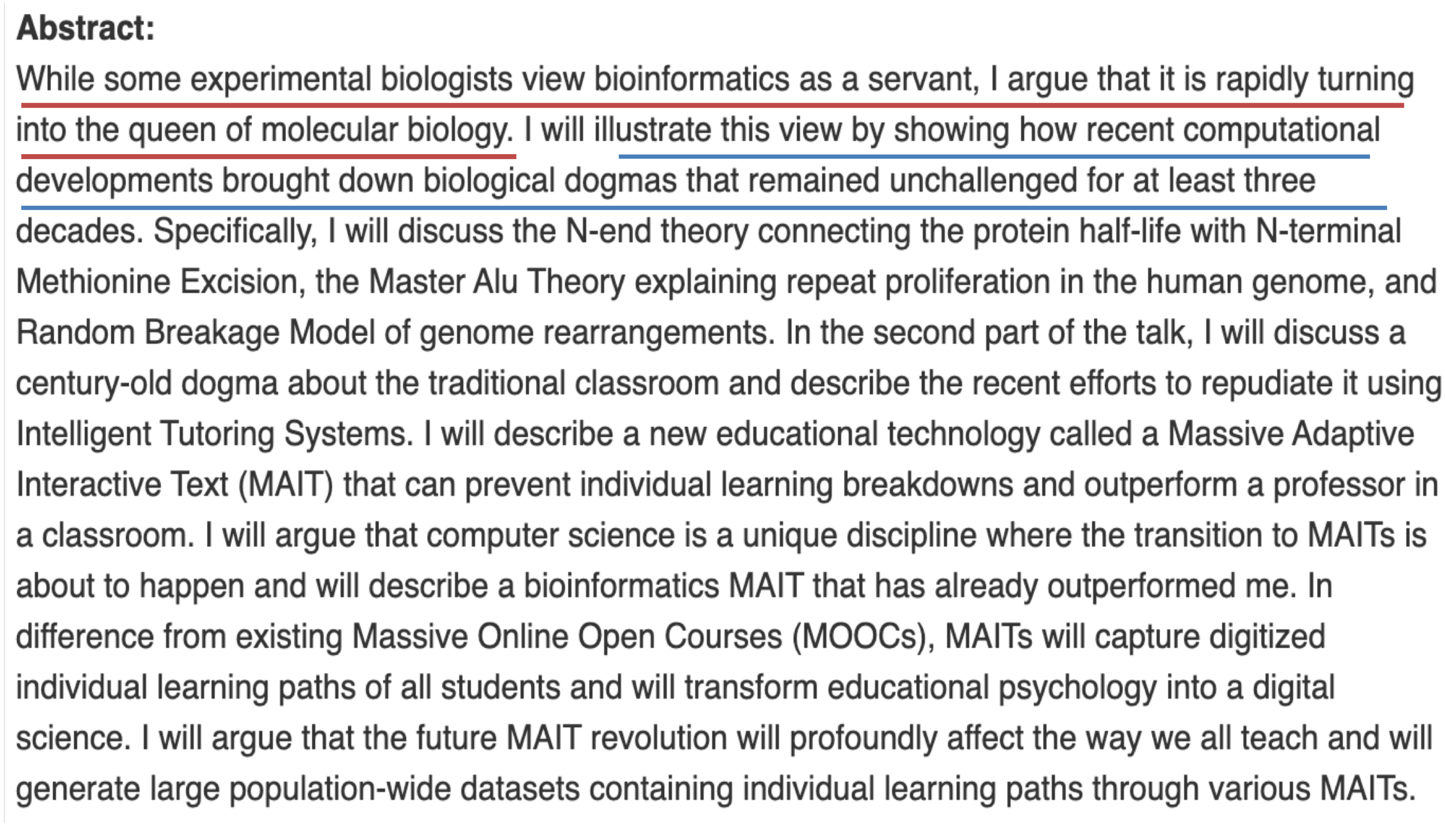
- More interpretable deep learning frameworks by coupling the structure of the neural network with the internal workings of cell are desired.

A Widely Accepted Saying

- *What computational science is to molecular biology is like what mathematics has been to physics* ……

Bioinformatics: a Servant or the Queen of Molecular Biology?

- Pavel Pevzner, BIBM 2020



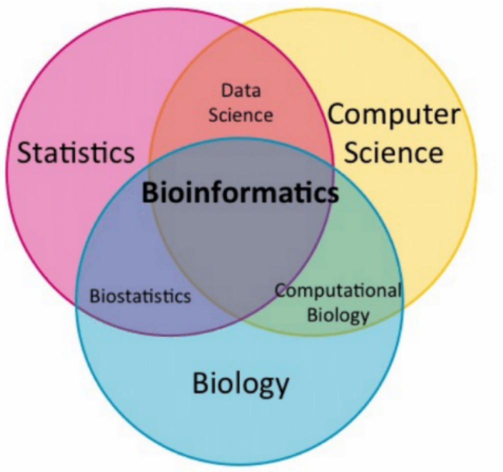
Abstract:

While some experimental biologists view bioinformatics as a servant, I argue that it is rapidly turning into the queen of molecular biology. I will illustrate this view by showing how recent computational developments brought down biological dogmas that remained unchallenged for at least three decades. Specifically, I will discuss the N-end theory connecting the protein half-life with N-terminal Methionine Excision, the Master Alu Theory explaining repeat proliferation in the human genome, and Random Breakage Model of genome rearrangements. In the second part of the talk, I will discuss a century-old dogma about the traditional classroom and describe the recent efforts to repudiate it using Intelligent Tutoring Systems. I will describe a new educational technology called a Massive Adaptive Interactive Text (MAIT) that can prevent individual learning breakdowns and outperform a professor in a classroom. I will argue that computer science is a unique discipline where the transition to MAITs is about to happen and will describe a bioinformatics MAIT that has already outperformed me. In difference from existing Massive Online Open Courses (MOOCs), MAITs will capture digitized individual learning paths of all students and will transform educational psychology into a digital science. I will argue that the future MAIT revolution will profoundly affect the way we all teach and will generate large population-wide datasets containing individual learning paths through various MAITs.

Bioinformatics Programs and Courses at CSE

- **Computational Biology and Bioinformatics (CBB) minor**
- **PhD/MS in CS with Bioinformatics specialization**

- CSCE496/896 Computational Methods in Bioinformatics (Renamed to CSCE 471/871 Introduction to Bioinformatics)
  - A general introduction to the field of bioinformatics
  - A way of thinking -- tackling “biological problem” computationally
  - Some exposure to computational biology and bioinformatics research, covering multiple aspects of computational genomics, proteomics and systems biology
- CSCE971 Advanced Bioinformatics
  - Fundamental machine learning and state-of-the-art deep learning
  - Probabilistic modeling
- CSCE155T Programming in Python
- CSCE311 Data Structures and Algorithms for Informatics


Data
Science
Statistics
Computer
Science
Bioinformatics
Biostatistics
Computational
Biology
Biology