An Introduction to Biology with Computers

Brittany N. Lasseigne, PhD
HudsonAlpha Intstitute for Biotechnology

4 June 2018

@bnlasse   blasseigne@hudsonalpha.org

- My background

- ‘Genomical’ Data: the Necessity of Biology with Computers

- Introduction to Bioinformatics and Computational Biology

- Applications of Computational Biology in Genomics

- **My background**

My Education

The Mississippi School for Mathematics and Science

An Opportunity for Excellence

My Education

The Mississippi School
for Mathematics and Science
An Opportunity for Excellence

STATE
MISSISSIPPI STATE UNIVERSITY™
JAMES WORTH
BAGLEY
COLLEGE OF ENGINEERING

BS: Biological Engineering

My Education

The Mississippi School
for Mathematics and Science
An Opportunity for Excellence

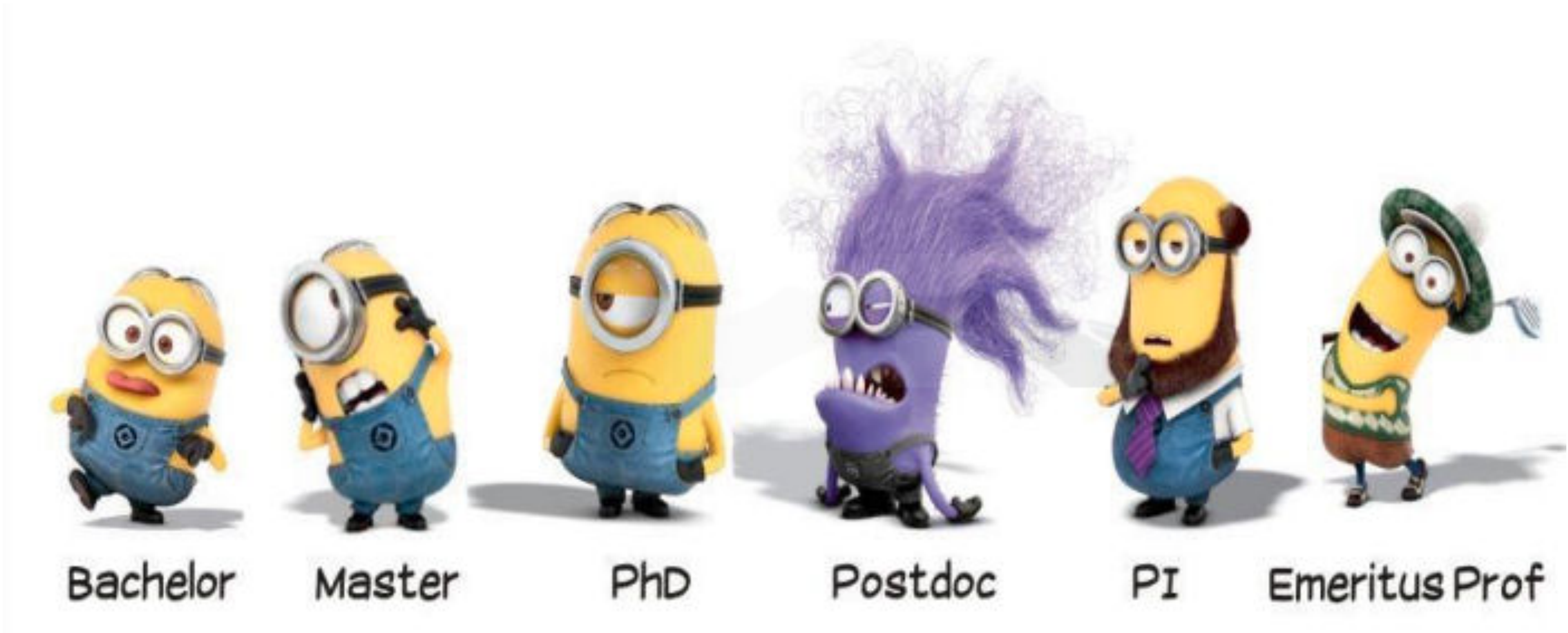
STATE
MISSISSIPPI STATE UNIVERSITY™
JAMES WORTH
BAGLEY
COLLEGE OF ENGINEERING

BS: Biological Engineering

UAH
The University of Alabama in Huntsville

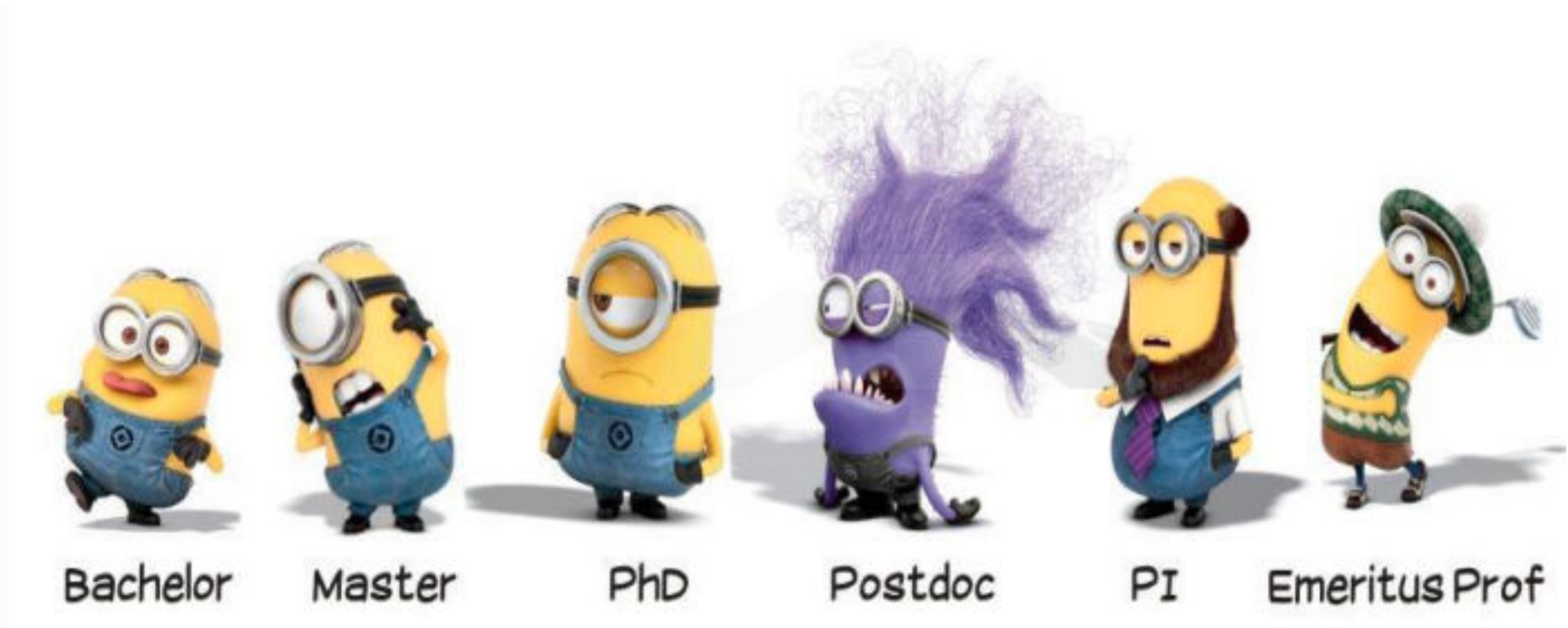
PhD: Biotechnology Science and Engineering

Postdoctoral Fellow & Senior Scientist

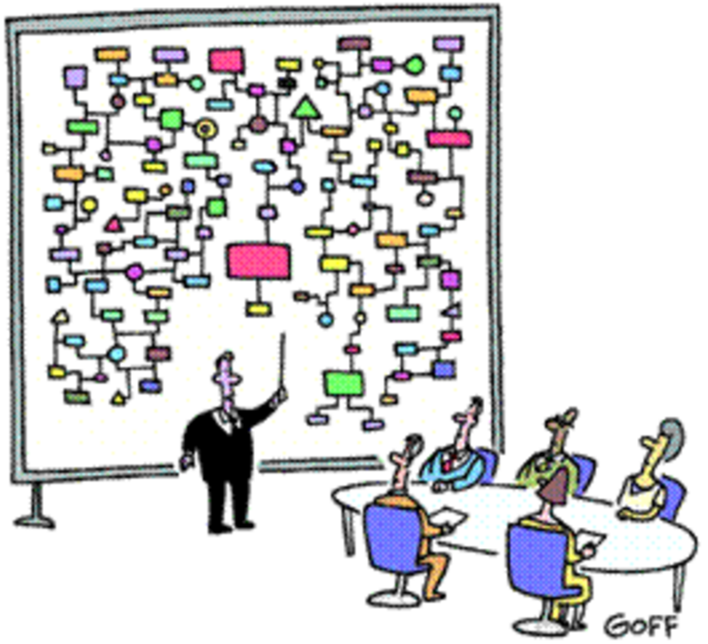


Postdoctoral Fellow & Senior Scientist

- **HudsonAlpha Institute for Biotechnology, 2014-present**
  - Applying machine learning, big data integraiton and genomics to complex human disease to improve disease prevention, detection, treatment, and monitoring



- My background
- **‘Genomical’ Data: the Necessity of Biology with Computers**
- Introduction to Bioinformatics and Computational Biology
- Applications of Computational Biology in Genomics



"And that's why we need a computer."

Complex Human Diseases:

combination of genetic, environmental and lifetyle factors
(most of which have not yet been identified)

American Cancer Society, 2015 & Harvard NeuroDiscovery Center, 2017.

Complex Human Diseases:

combination of genetic, environmental and lifetyle factors
(most of which have not yet been identified)

Cancer:

- Men have a 1 in 2 lifetime risk of developing cancer
- Women have a 1 in 3 lifetime risk of developing cancer

American Cancer Society, 2015 & Harvard NeuroDiscovery Center, 2017.

Complex Human Diseases:

combination of genetic, environmental and lifetyle factors
(most of which have not yet been identified)

Cancer:

- Men have a 1 in 2 lifetime risk of developing cancer
- Women have a 1 in 3 lifetime risk of developing cancer

Psychiatric Illness:

- 1 in 4 adults suffer from a diagnosable mental disorder each year
- ~6% suffer serious disabilities as a result

American Cancer Society, 2015 & Harvard NeuroDiscovery Center, 2017.

Complex Human Diseases:

combination of genetic, environmental and lifetyle factors
(most of which have not yet been identified)

Cancer:

- Men have a 1 in 2 lifetime risk of developing cancer
- Women have a 1 in 3 lifetime risk of developing cancer

Psychiatric Illness:

- 1 in 4 adults suffer from a diagnosable mental disorder each year
- ~6% suffer serious disabilities as a result

Neurodegenerative Disease:

- ~6.5M Americans suffer from a neurodegenerative disease; expected to rise to 12M by 2030

American Cancer Society, 2015 & Harvard NeuroDiscovery Center, 2017.

Complex Human Diseases:

combination of genetic, environmental and lifetyle factors
(most of which have not yet been identified)

Cancer:

- Men have a 1 in 2 lifetime risk of developing cancer
- Women have a 1 in 3 lifetime risk of developing cancer

Psychiatric Illness:

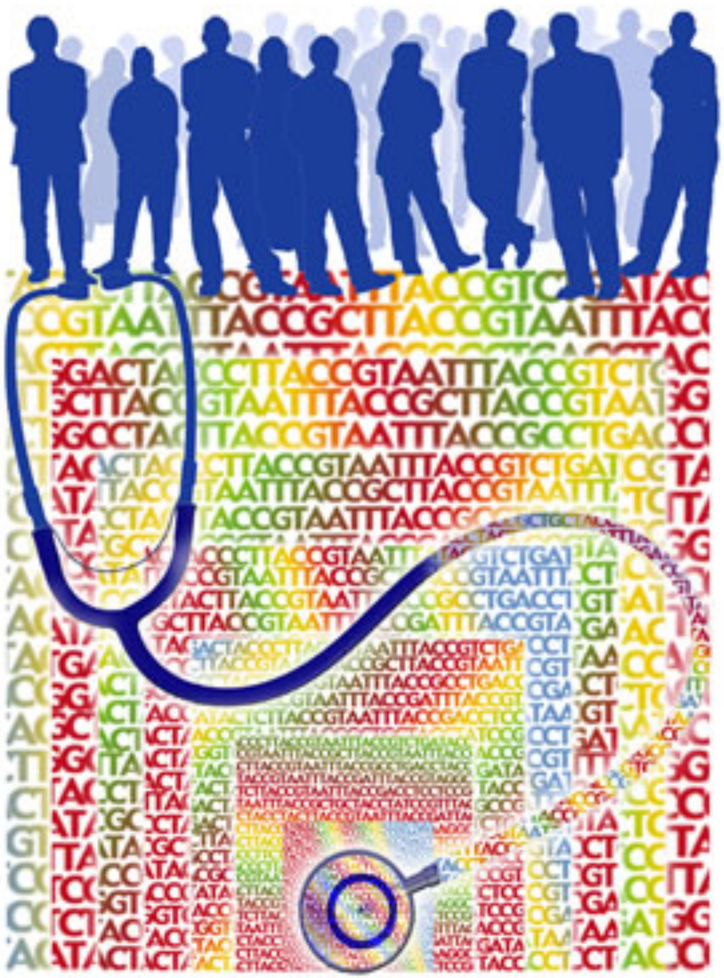
- 1 in 4 adults suffer from a diagnosable mental disorder each year
- ~6% suffer serious disabilities as a result

Neurodegenerative Disease:

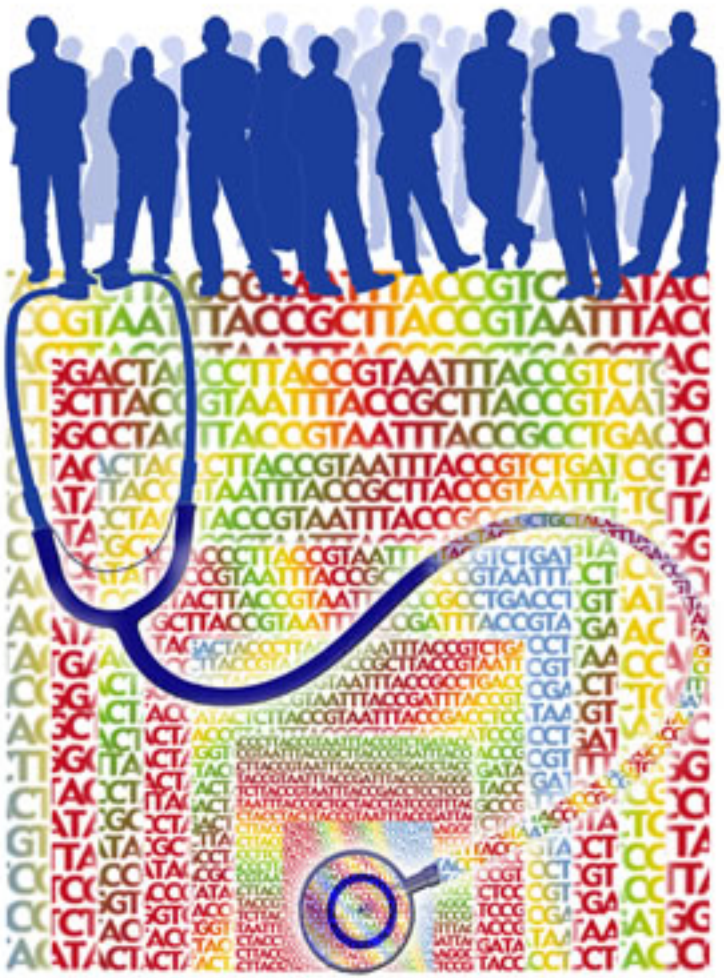
- ~6.5M Americans suffer from a neurodegenerative disease; expected to rise to 12M by 2030

American Cancer Society, 2015 & Harvard NeuroDiscovery Center, 2017.

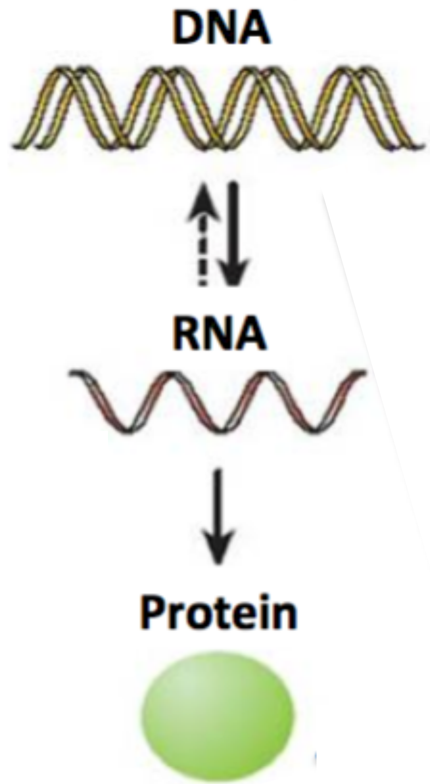
Identify genetic/genomic variation associated with disease to improve patient care

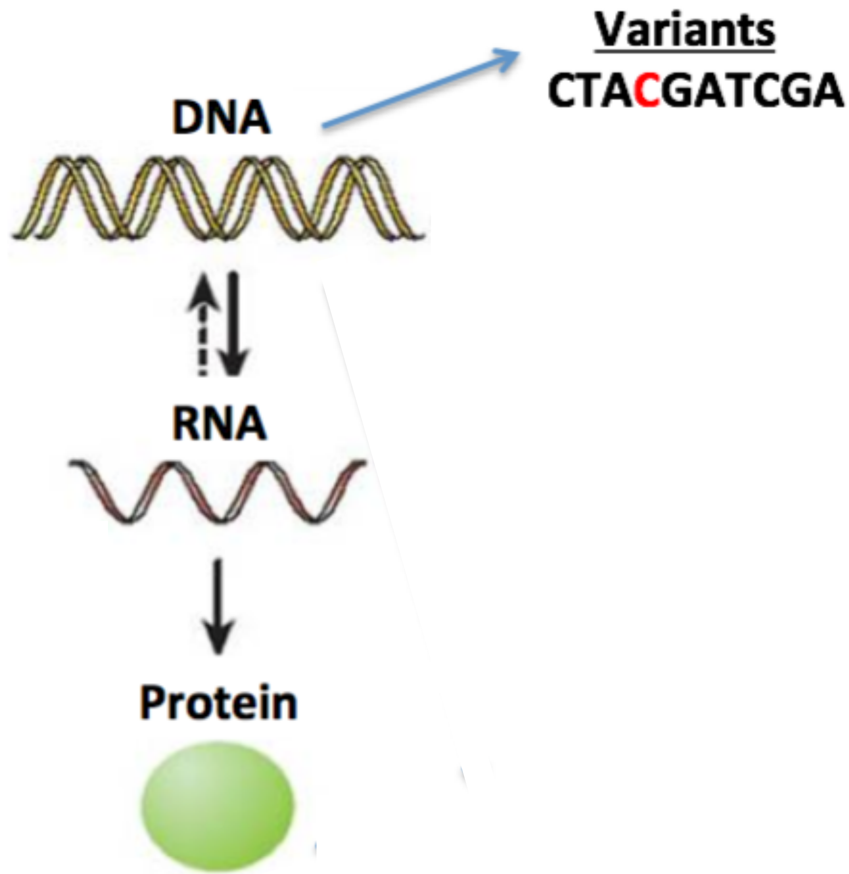


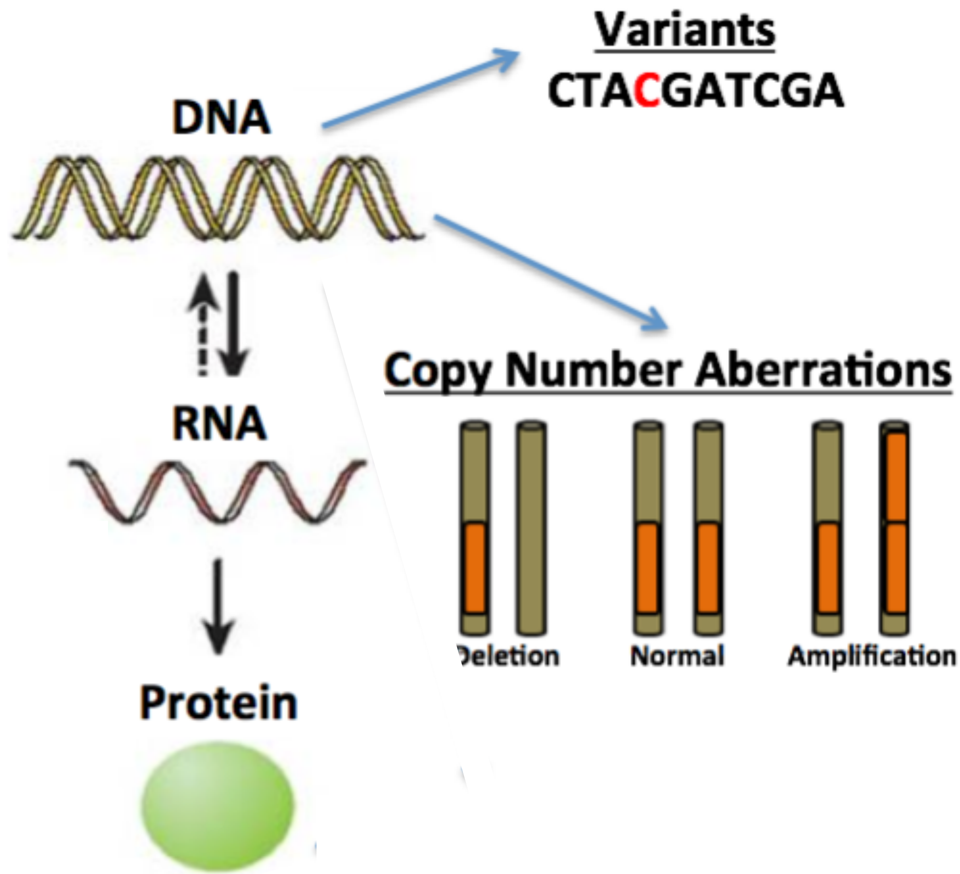
Identify genetic/genomic variation associated with disease to improve patient care

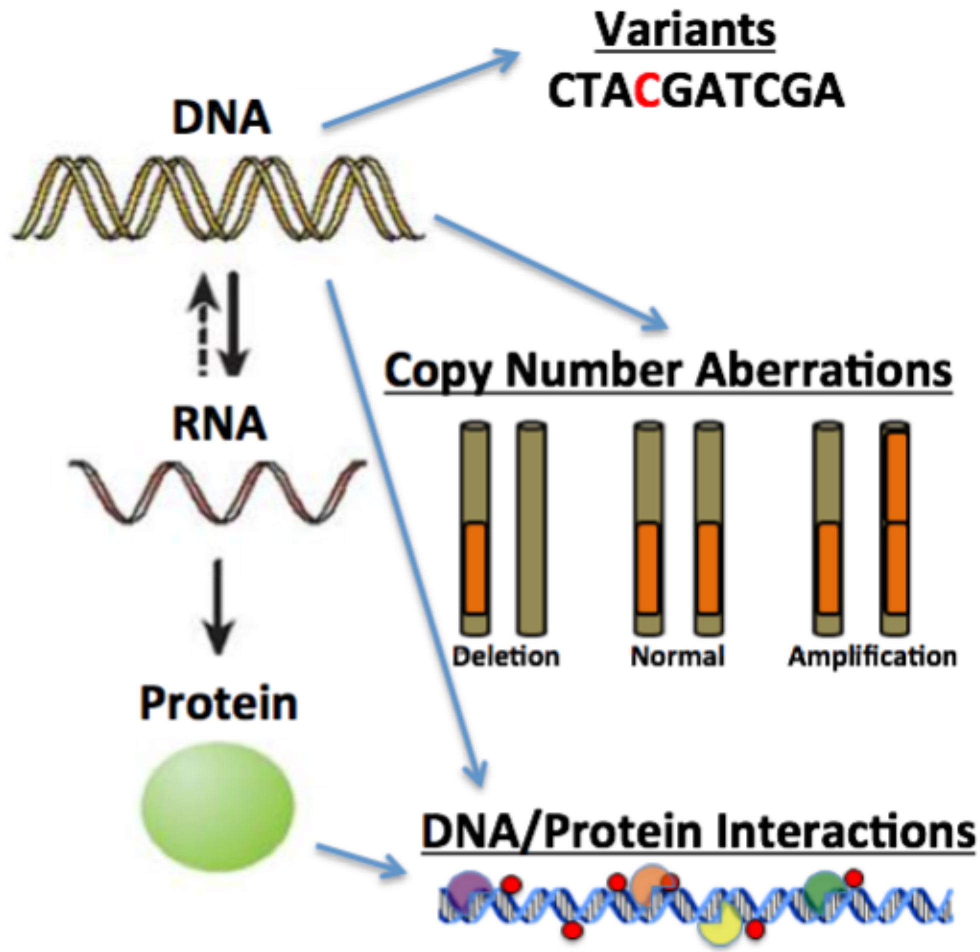


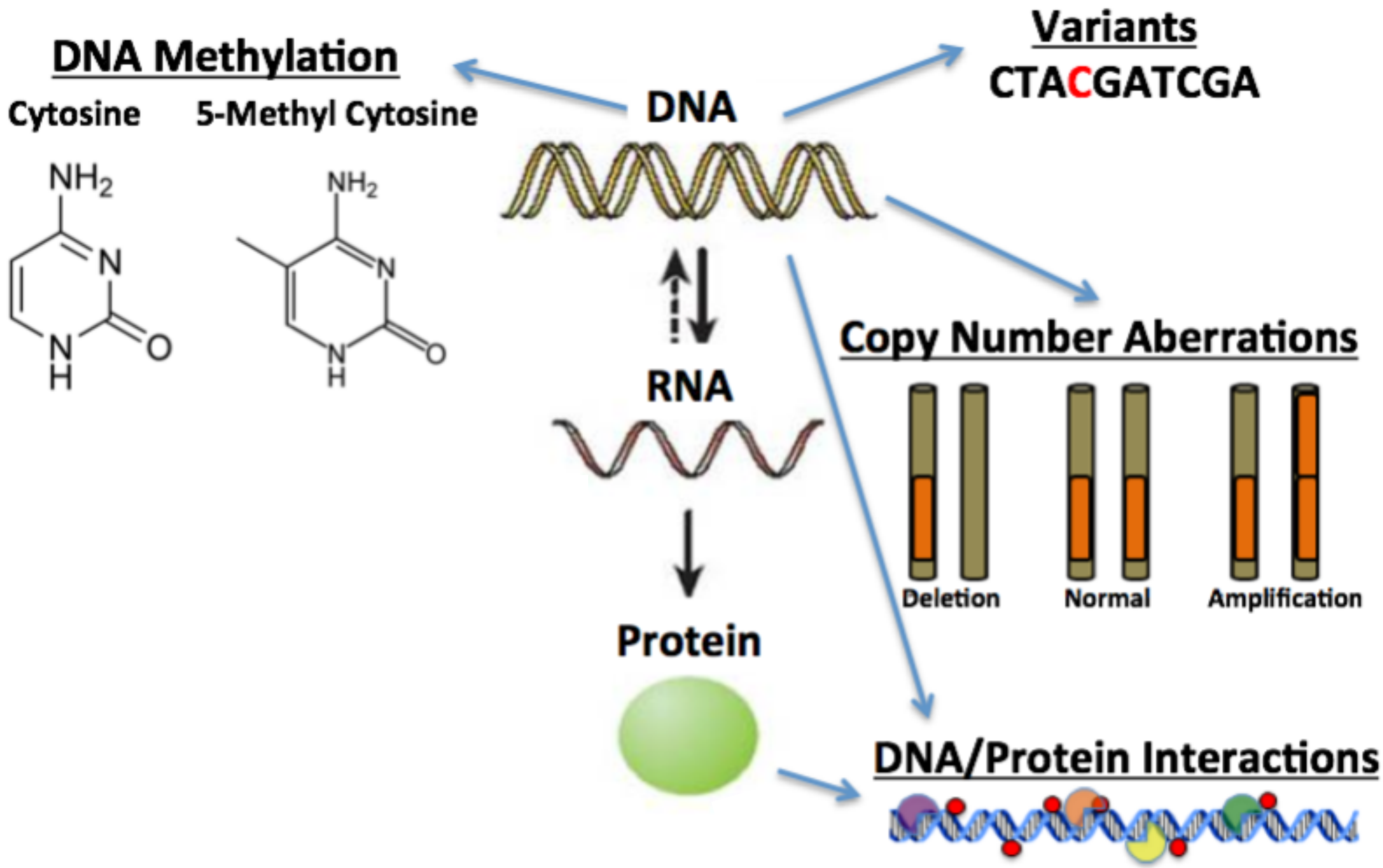
- Which patients are high risk for developing cancer?
- What are early biomarkers of cancer?
- Which patients are likely to be short/long term cancer survivers?
- What chemotherapeutic might a cancer patient benefit from?

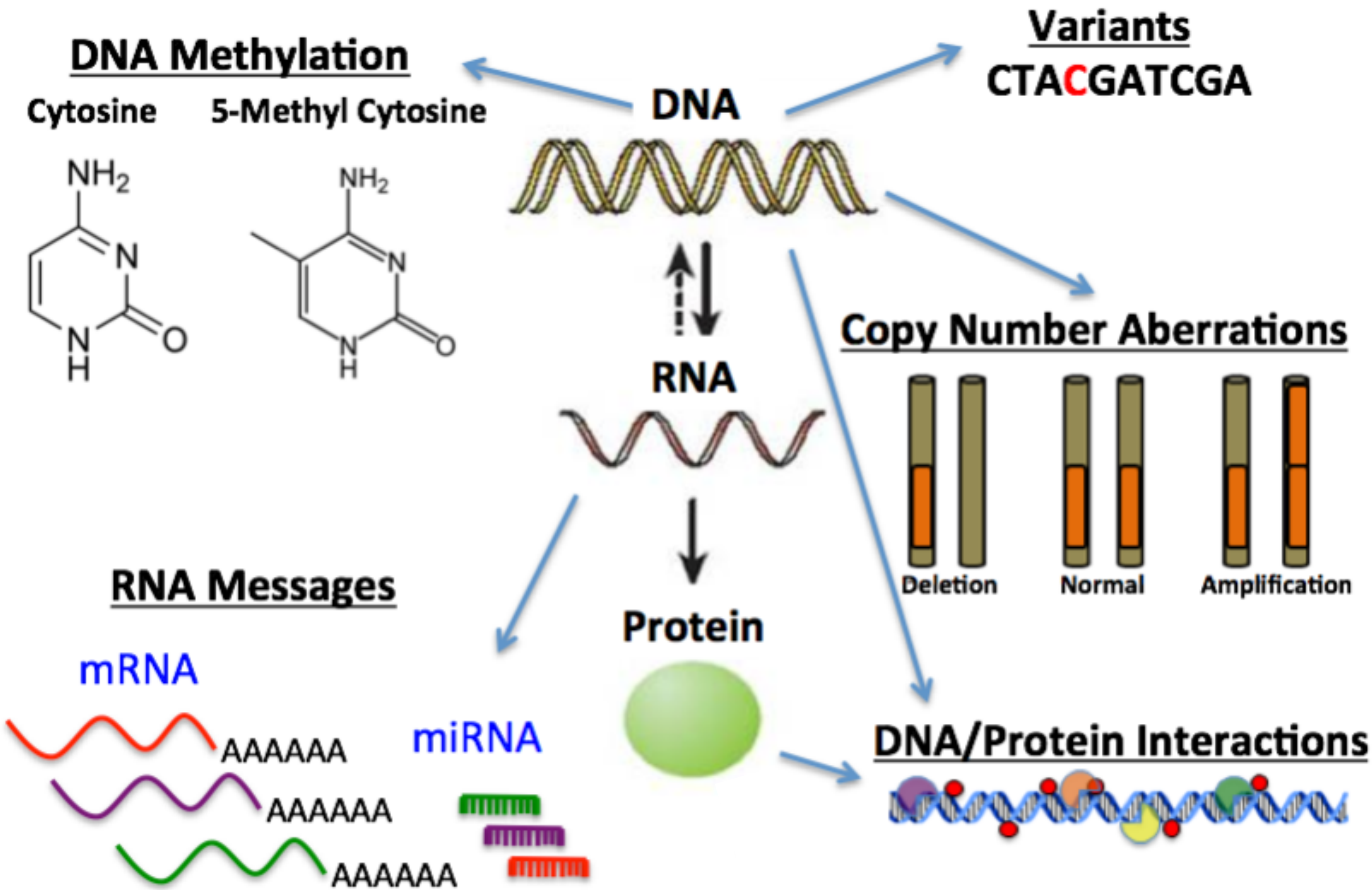

DNA
RNA
Protein

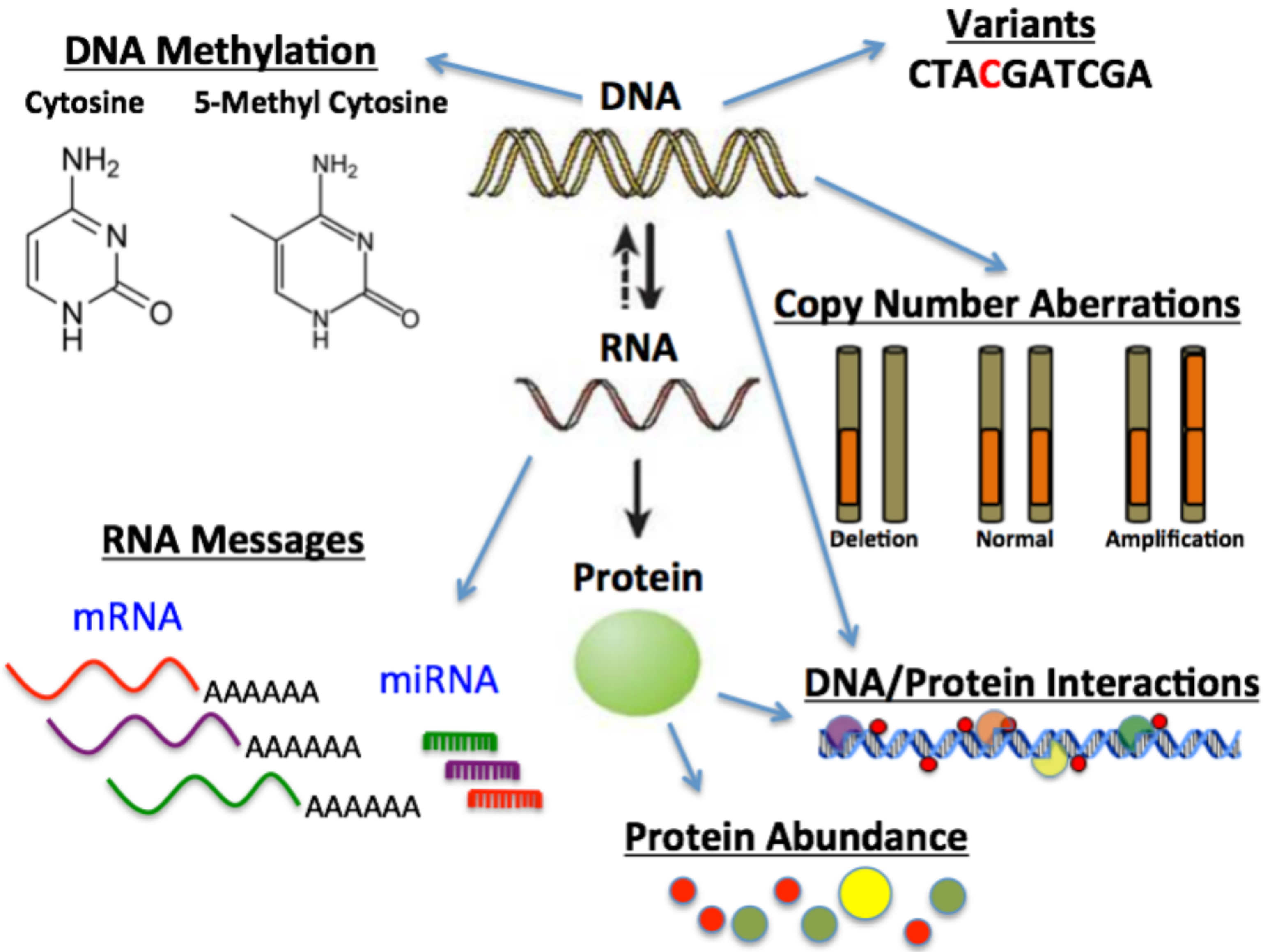

Variants
CTACGATCGA
DNA
RNA
Protein

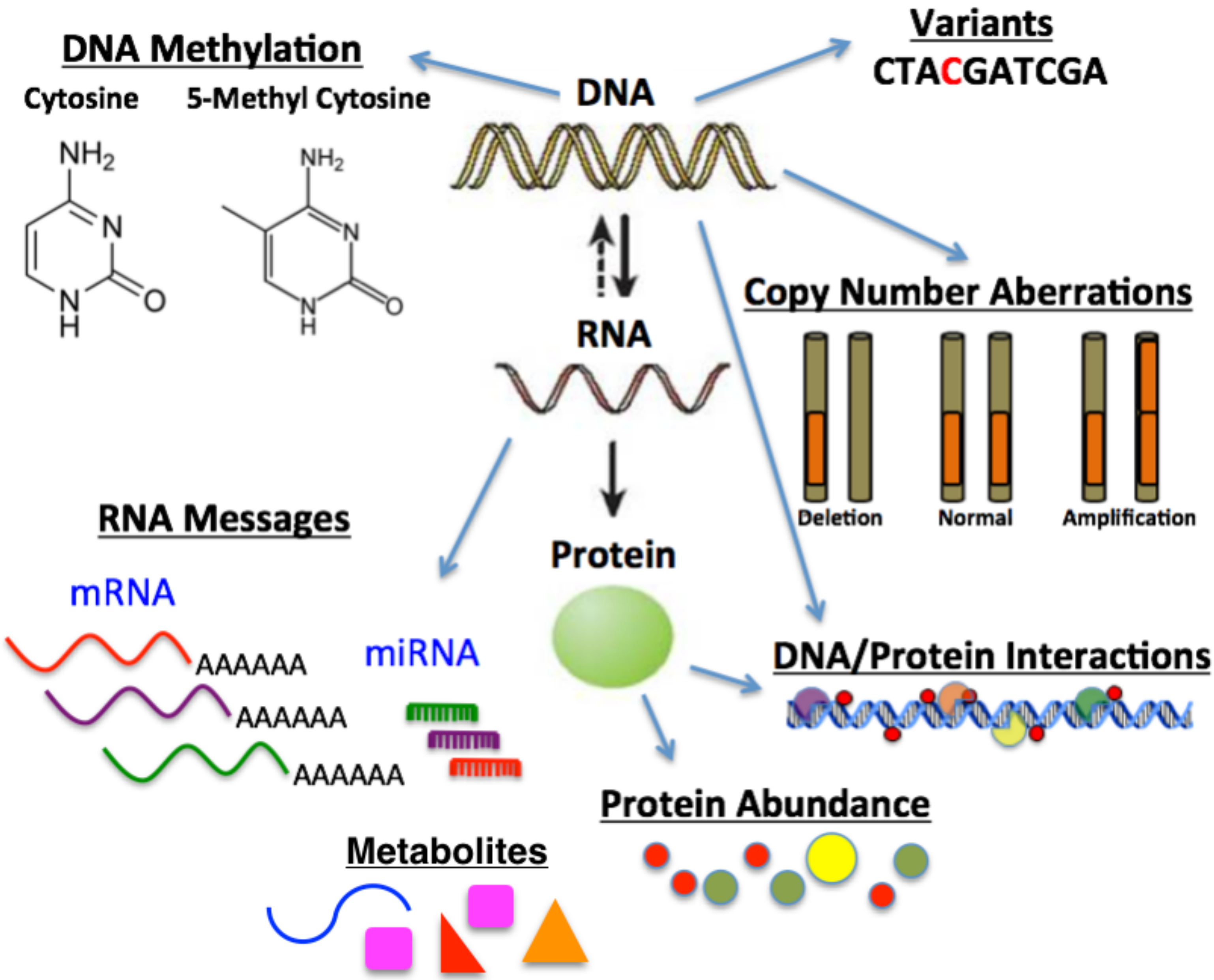

Variants
CTACGATCGA
DNA
RNA
Protein
Copy Number Aberrations
Deletion
Normal
Amplification


DNA
RNA
Protein
Variants
CTACGATCGA
Copy Number Aberrations
Deletion
Normal
Amplification
DNA/Protein Interactions


DNA Methylation
Cytosine
5-Methyl Cytosine
NH_2
N
N
H
O
DNA
Variants
CTACGATCGA
RNA
Copy Number Aberrations
Deletion
Normal
Amplification
Protein
DNA/Protein Interactions


DNA Methylation
Cytosine
5-Methyl Cytosine
NH2
N
N
H
O
DNA
Variants
CTACGATCGA
RNA
Copy Number Aberrations
Deletion
Normal
Amplification
RNA Messages
mRNA
AAAAAA
AAAAAA
AAAAAA
miRNA
Protein
DNA/Protein Interactions


DNA Methylation
Cytosine
5-Methyl Cytosine
NH_2
N
N
H
O
NH_2
N
N
H
O
DNA
Variants
CTACGATCGA
RNA
Copy Number Aberrations
Deletion
Normal
Amplification
RNA Messages
mRNA
AAAAAA
AAAAAA
AAAAAA
miRNA
Protein
DNA/Protein Interactions
Protein Abundance



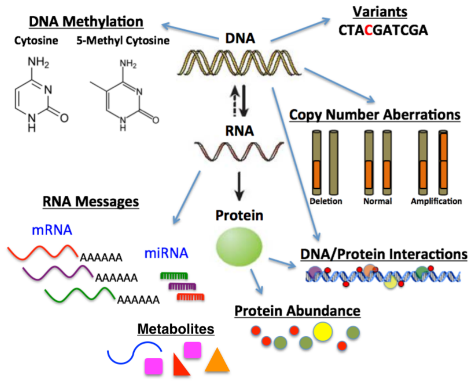
DNA Methylation
Cytosine
5-Methyl Cytosine
NH_2
N
O
N
H
NH_2
N
O
N
H
DNA
Variants
CTACGATCGA
RNA
Copy Number Aberrations
Deletion
Normal
Amplification
RNA Messages
mRNA
AAAAAA
AAAAAA
AAAAAA
miRNA
Protein
DNA/Protein Interactions
Protein Abundance
Metabolites

Improve disease prevention, diagnosis, prognosis, and treatment efficacy

Stephens, et al. *PLOS Biology*, 2015. Images from encodeproject.org and nanalyze.com.

Improve disease prevention, diagnosis, prognosis, and treatment efficacy

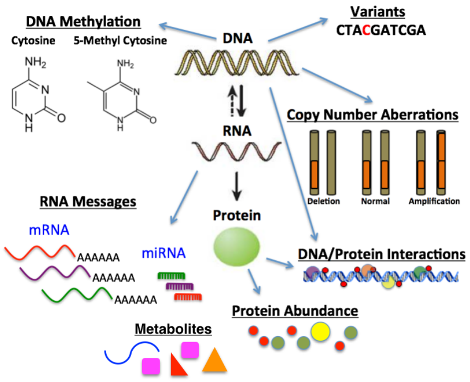
Multidimensional Data Sets



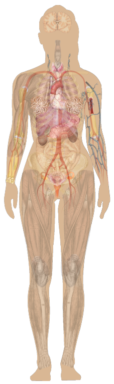
Stephens, et al. *PLOS Biology*, 2015. Images from encodeproject.org and nanalyze.com.

Improve disease prevention, diagnosis, prognosis, and treatment efficacy

Multidimensional Data Sets



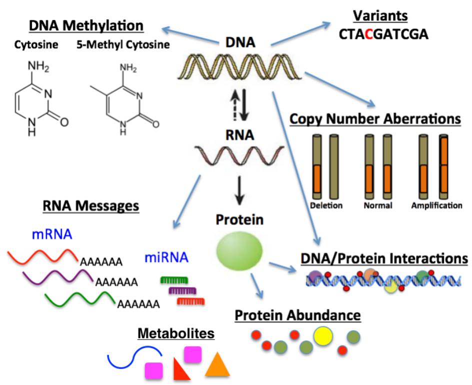
Cells, Tissues, & Diseases



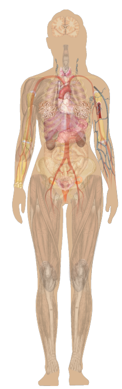
Stephens, et al. *PLOS Biology*, 2015. Images from encodeproject.org and nanalyze.com.

Improve disease prevention, diagnosis, prognosis, and treatment efficacy

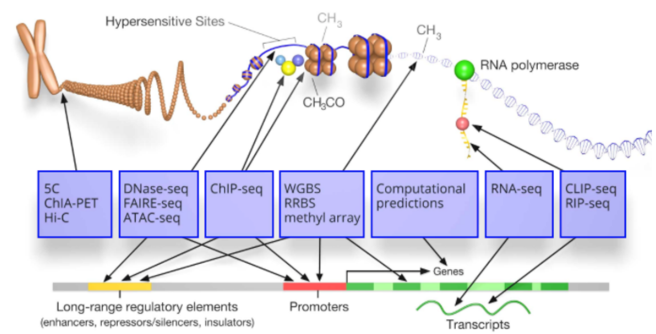
Multidimensional Data Sets



Cells, Tissues, & Diseases



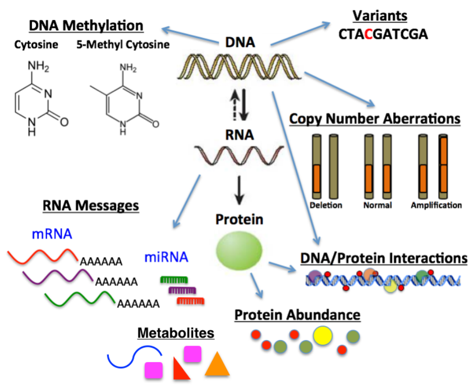
Functional Annotations



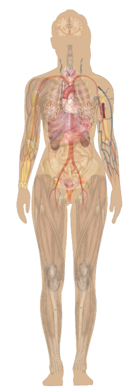
Stephens, et al. *PLOS Biology*, 2015. Images from encodeproject.org and nanalyze.com.

Improve disease prevention, diagnosis, prognosis, and treatment efficacy

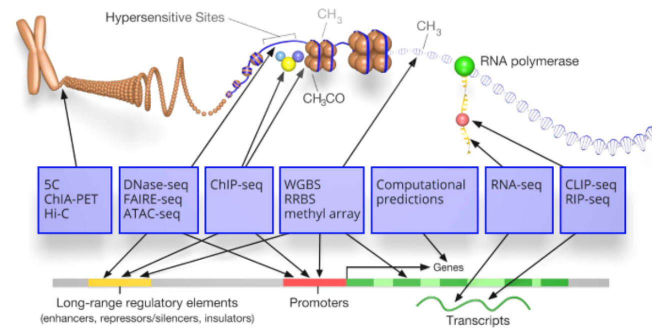
Multidimensional Data Sets



Cells, Tissues, & Diseases



Functional Annotations

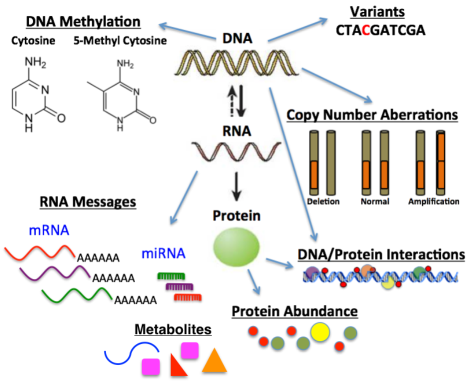


Big Data

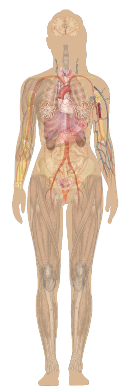
Stephens, et al. *PLOS Biology*, 2015. Images from encodeproject.org and nanalyze.com.

Improve disease prevention, diagnosis, prognosis, and treatment efficacy

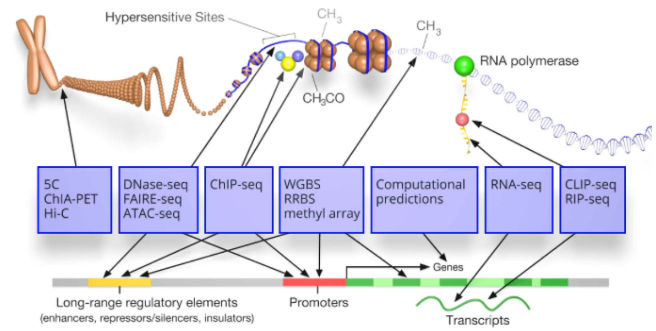
Multidimensional Data Sets



Cells, Tissues, & Diseases



Functional Annotations



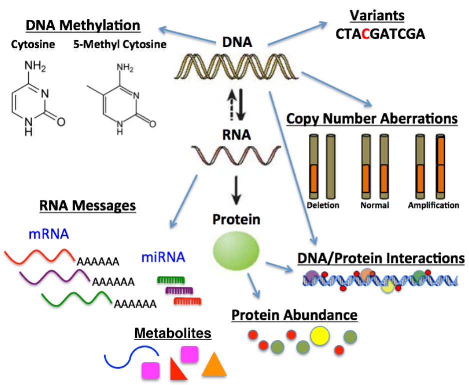
Big Data

Case Study: The Cancer Genome Atlas

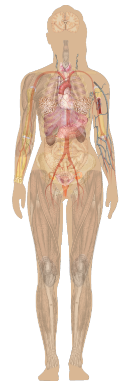
- Mulitiple data types for 11,000+ patients
- 549,625 files with 2000+ metadata attributes
- **>2.5 Petabytes of data**

Improve disease prevention, diagnosis, prognosis, and treatment efficacy

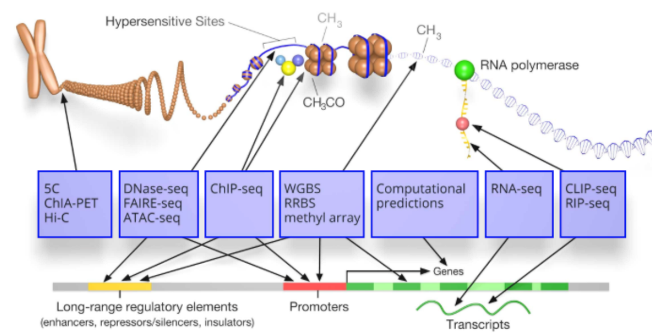
Multidimensional Data Sets



Cells, Tissues, & Diseases

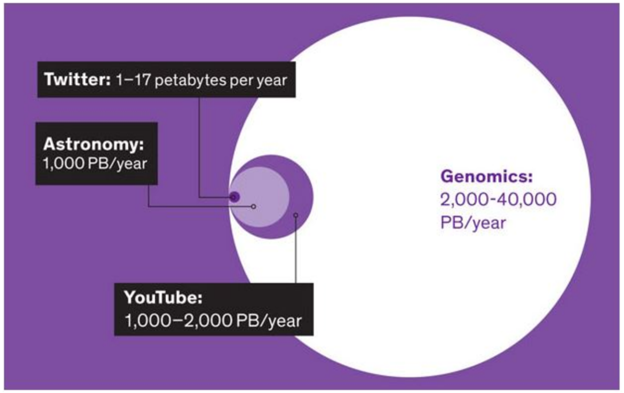


Functional Annotations



Big Data

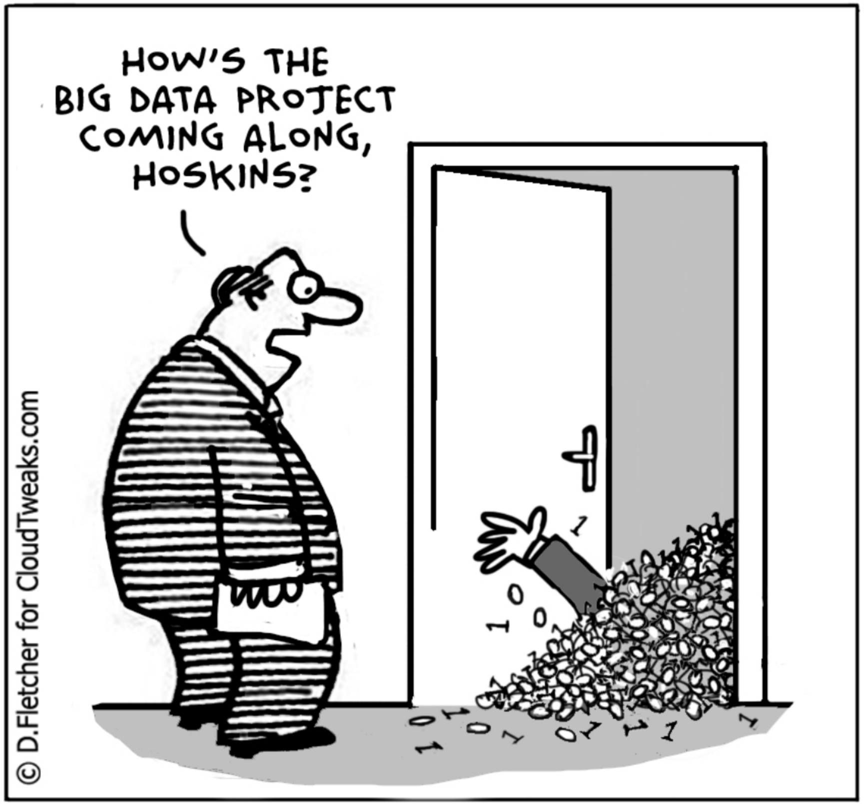
2025 Projection

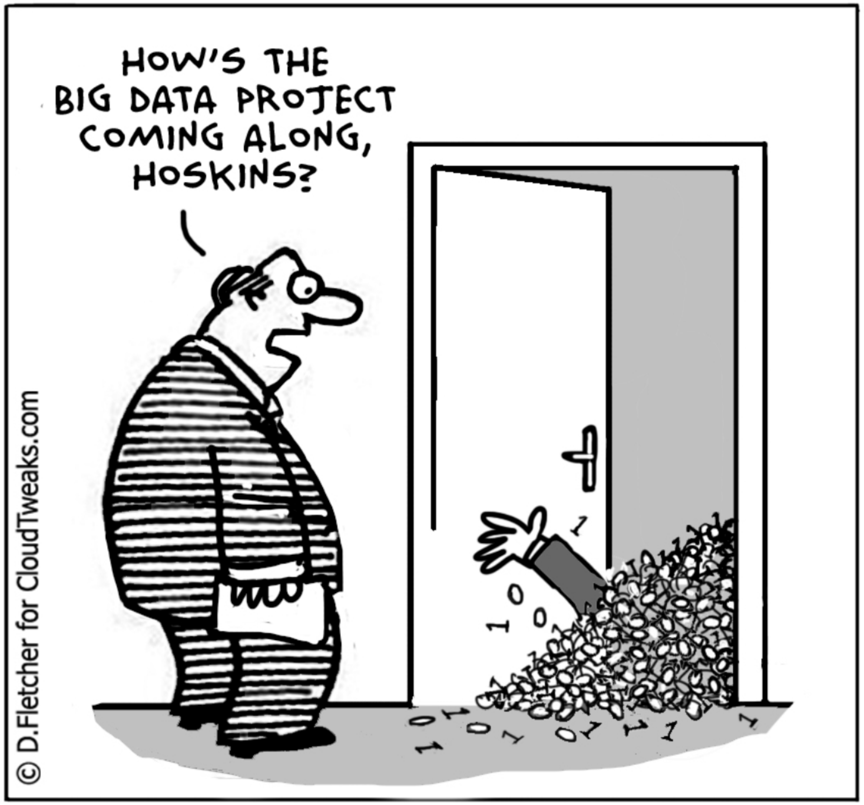


Case Study: The Cancer Genome Atlas

- Mulitiple data types for 11,000+ patients
- 549,625 files with 2000+ metadata attributes
- **>2.5 Petabytes of data**

Stephens, et al. *PLOS Biology*, 2015. Images from encodeproject.org and nanalyze.com.


HOW'S THE BIG DATA PROJECT COMING ALONG, HOSKINS?
© D.Fletcher for CloudTweaks.com



1 Petabyte of Data =

20M four-drawer filing cabinets filled with text

or

13.3 years of HD-TV video

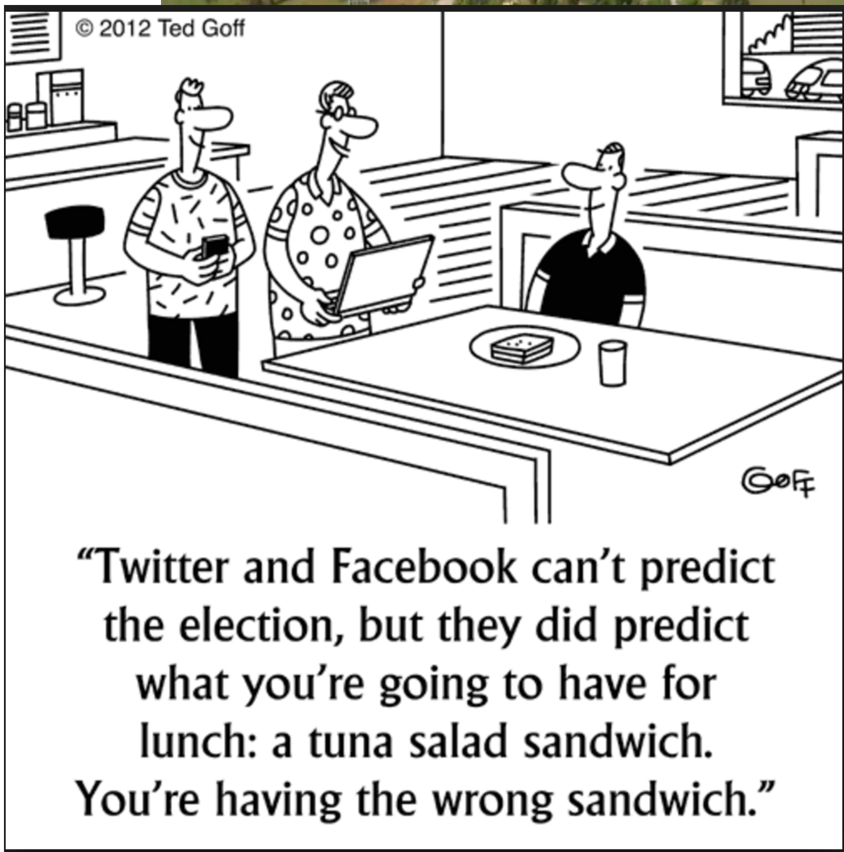
or

~7 billion Facebook photos

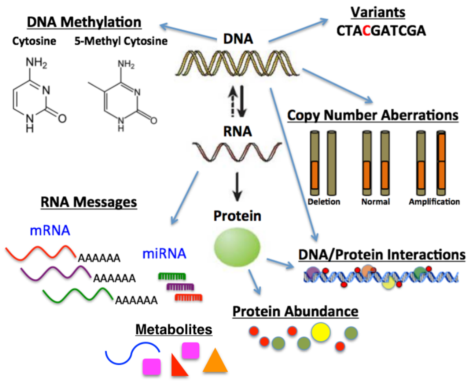
or

1 PB of MP3 songs requires ~2,000 years to play

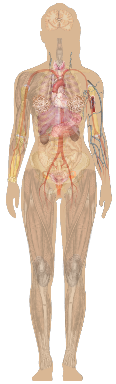
- My background
- ‘Genomical’ Data: the Necessity of Biology with Computers
- **Introduction to Bioinformatics and Computational Biology**
- Applications of Computational Biology in Genomics



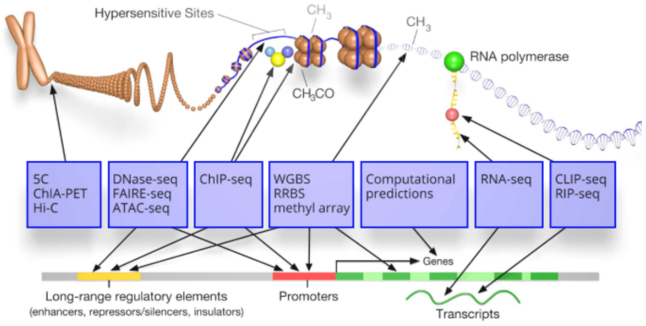
Multidimensional Data Sets



Cells, Tissues, & Diseases



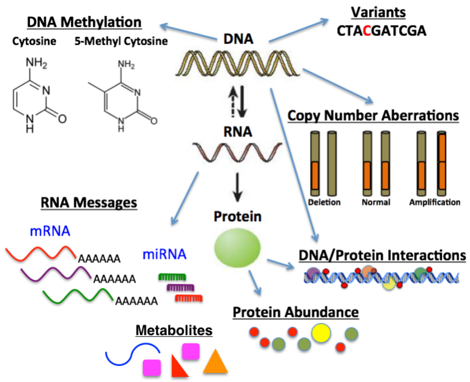
Functional Annotations

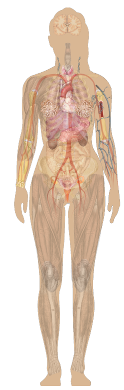


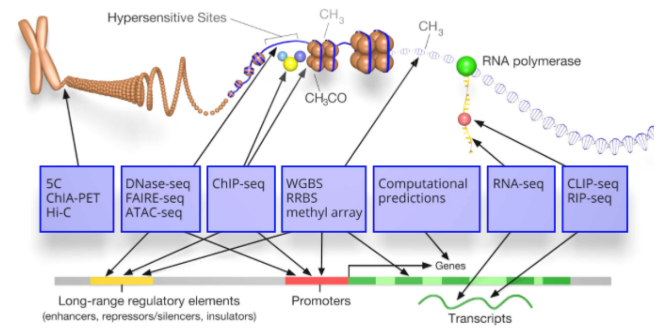
Improve disease prevention, diagnosis, prognosis, and treatment efficacy

mage from encodeproject.org and xorlogics.com.

Multidimensional Data Sets | **Cells, Tissues, & Diseases** | **Functional Annotations**





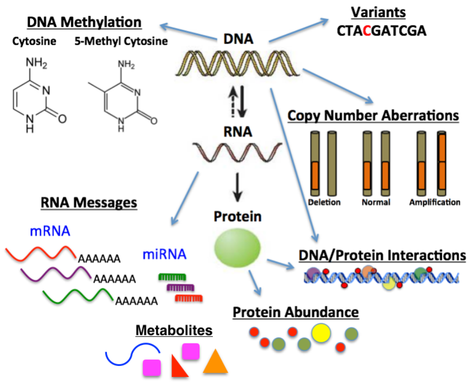


Improve disease prevention, diagnosis, prognosis, and treatment efficacy

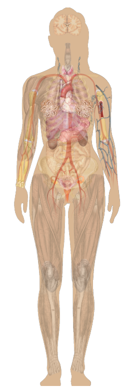
- We have lots of data and complex problems
- We want to manage lots of data and make data-driven predictions

mage from [encodeproject.org]() and [xorlogics.com]().

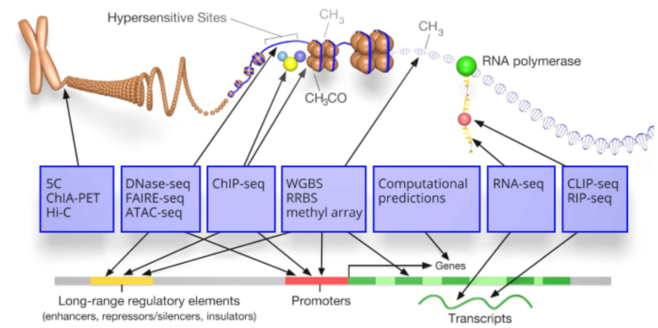
Multidimensional Data Sets



Cells, Tissues, & Diseases



Functional Annotations



Complex problems + Big Data —>

Computer Science + Mathematics

mage from encodeproject.org and xorlogics.com.

Computational Biology and Bioinformatics

*Disclaimer: My Opinion

Computational Biology and Bioinformatics

*Disclaimer: My Opinion

- **Computational biology is the application of computer science and mathematics to problems in biology**

Computational Biology and Bioinformatics

*Disclaimer: My Opinion

- **Computational biology is the application of computer science and mathematics to problems in biology**
  - Not just genomics! e.g., biophysics, biochemistry, etc.

Computational Biology and Bioinformatics

*Disclaimer: My Opinion

- **Computational biology is the application of computer science and mathematics to problems in biology**
  - Not just genomics! e.g., biophysics, biochemistry, etc.

- **The terms 'bioinformatics' and 'computational biology' are often used interchangeably:**

Computational Biology and Bioinformatics

*Disclaimer: My Opinion

- **Computational biology is the application of computer science and mathematics to problems in biology**
  - Not just genomics! e.g., biophysics, biochemistry, etc.

- **The terms 'bioinformatics' and 'computational biology' are often used interchangeably:**
  - **Bioinformatics** is often associated with the development of software tools, databases, and visualization methods

Computational Biology and Bioinformatics

*Disclaimer: My Opinion

- **Computational biology is the application of computer science and mathematics to problems in biology**
  - Not just genomics! e.g., biophysics, biochemistry, etc.

- **The terms 'bioinformatics' and 'computational biology' are often used interchangeably:**
  - **Bioinformatics** is often associated with the development of software tools, databases, and visualization methods
  - **Computational biology** is often used to describe data analysis, algorithm development, and mathematical modeling

Computational Biology and Bioinformatics

*Disclaimer: My Opinion

- **Computational biology is the application of computer science and mathematics to problems in biology**
  - Not just genomics! e.g., biophysics, biochemistry, etc.

- **The terms 'bioinformatics' and 'computational biology' are often used interchangeably:**
  - **Bioinformatics** is often associated with the development of software tools, databases, and visualization methods
  - **Computational biology** is often used to describe data analysis, algorithm development, and mathematical modeling

Other terms you might hear to describe the interdiscipinary field of biology/math/computer science:

Computational Biology and Bioinformatics

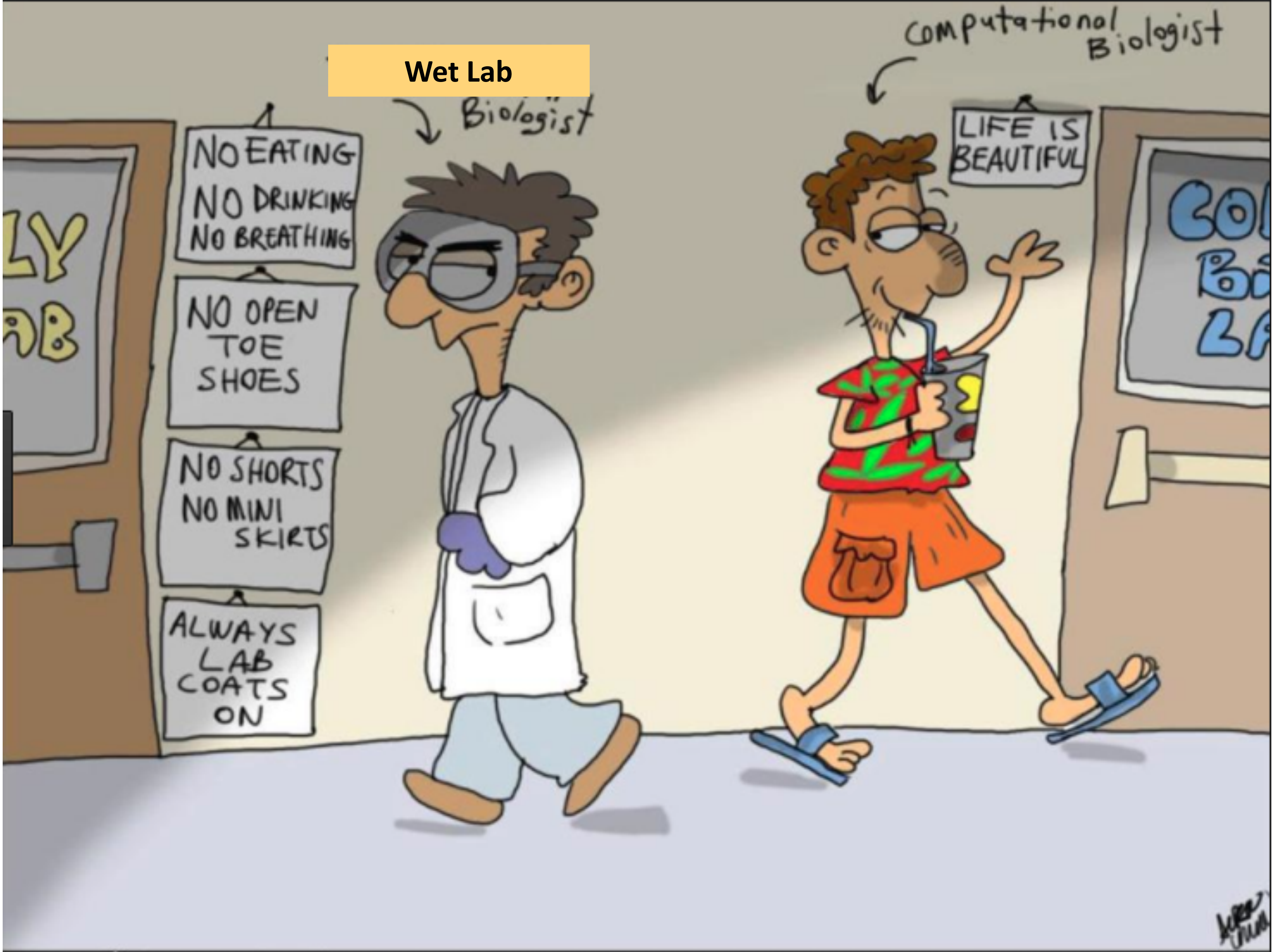
*Disclaimer: My Opinion

- **Computational biology is the application of computer science and mathematics to problems in biology**
  - Not just genomics! e.g., biophysics, biochemistry, etc.

- **The terms 'bioinformatics' and 'computational biology' are often used interchangeably:**
  - **Bioinformatics** is often associated with the development of software tools, databases, and visualization methods
  - **Computational biology** is often used to describe data analysis, algorithm development, and mathematical modeling

Other terms you might hear to describe the interdisciplinary field of biology/math/computer science:

Data Science, Systems Biology, Statistical Biology, Biostatistics, and Genomics (implicit)


Wet Lab
Biologist
Computational Biologist
NO EATING
NO DRINKING
NO BREATHING
NO OPEN TOE SHOES
NO SHORTS
NO MINI SKIRTS
ALWAYS LAB COATS ON
LIFE IS BEAUTIFUL

Computational people can work from anywhere…
but that also means they can work from anywhere

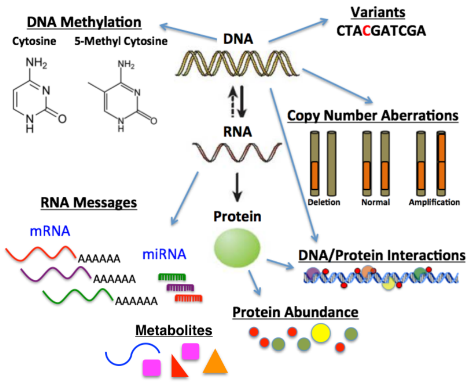
Computational people can work from anywhere…
but that also means they can work from anywhere

Generally computational skills are:

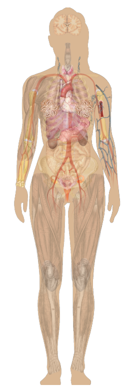
- In demand
- Flexible
- Highly transferable

Computational Biology <u>IS</u> Biology!

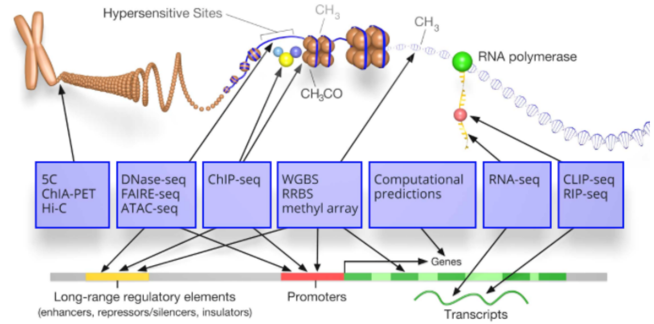
Multidimensional Data Sets



Cells, Tissues, & Diseases



Functional Annotations



Complex problems + Big Data —>

Machine Learning!

mage from encodeproject.org and xorlogics.com.

Machine Learning

- data analysis method that automates analytical model building

Machine Learning

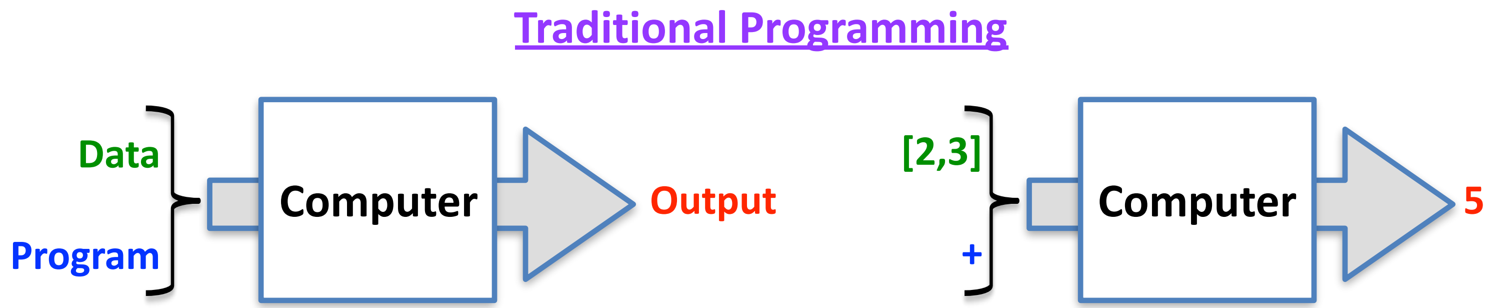
- data analysis method that automates analytical model building
- make data driven **predictions** or discover **patterns** without explicit human intervention

Machine Learning

- data analysis method that automates analytical model building
- make data driven **predictions** or discover **patterns** without explicit human intervention
- Useful when have complex problems and lots of data (‘big data’)

Machine Learning

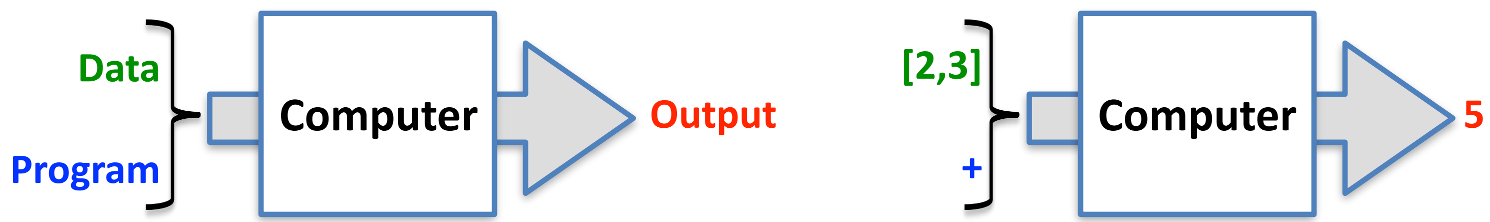
- data analysis method that automates analytical model building
- make data driven **predictions** or discover **patterns** without explicit human intervention
- Useful when have complex problems and lots of data ('big data')



Machine Learning

- data analysis method that automates analytical model building
- make data driven **predictions** or discover **patterns** without explicit human intervention
- Useful when have complex problems and lots of data ('big data')

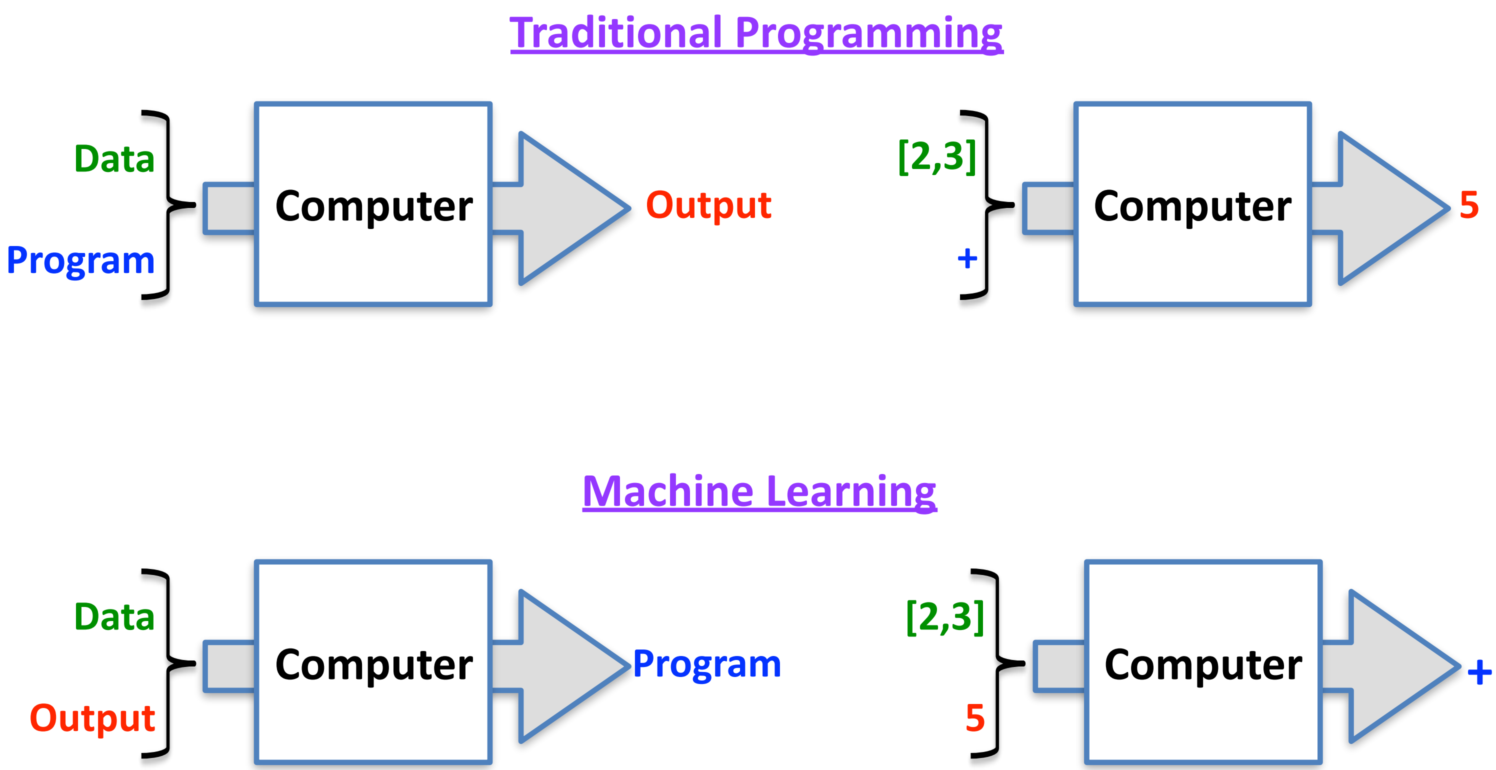
Traditional Programming



Machine Learning

Machine Learning

- data analysis method that automates analytical model building
- make data driven **predictions** or discover **patterns** without explicit human intervention
- Useful when have complex problems and lots of data ('big data')



- Our goal isn't to make perfect guesses, but to make useful guesses—we want to build a model that is useful for the future

Supervised Learning:

-Prediction

Ex. linear & logistic regression

Unsupervised Learning:

-Find patterns

Ex. Clustering, Principle Component Analysis

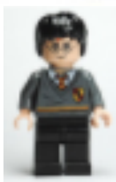
Supervised Learning:

-Prediction

Ex. linear & logistic regression

Known Data + Known Response

YES



NO

Unsupervised Learning:

-Find patterns

Ex. Clustering, Principle Component Analysis

Supervised Learning:

-Prediction

Ex. linear & logistic regression

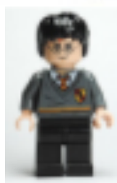
Unsupervised Learning:

-Find patterns

Ex. Clustering, Principle Component Analysis

Known Data + Known Response

YES



NO

MODEL

Supervised Learning:

-Prediction

Ex. linear & logistic regression

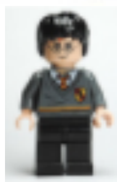
Unsupervised Learning:

-Find patterns

Ex. Clustering, Principle Component Analysis

Known Data + Known Response

YES



NO

MODEL

NEW DATA

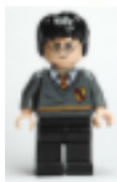
Supervised Learning:

-Prediction

Ex. linear & logistic regression

Known Data + Known Response

YES



NO

MODEL

NEW DATA

Predict Response

Unsupervised Learning:

-Find patterns

Ex. Clustering, Principle Component Analysis

Supervised Learning:

-Prediction

Ex. linear & logistic regression

Known Data + Known Response

YES

NO

MODEL

NEW DATA

Predict Response

Unsupervised Learning:

-Find patterns

Ex. Clustering, Principle Component Analysis

Uncategorized Data

Supervised Learning:

-Prediction

Ex. linear & logistic regression

Unsupervised Learning:

-Find patterns

Ex. Clustering, Principle Component Analysis

Known Data + Known Response

YES

NO

MODEL

NEW DATA

Predict Response

Uncategorized Data

Clusters of Categorized Data

Real-World Machine Learning Applications

Real-World Machine Learning Applications

Self-Driving Car

Real-World Machine Learning Applications

Self-Driving Car

Real-World Machine Learning Applications

Self-Driving Car

Mail Sorting

Real-World Machine Learning Applications

Self-Driving Car

Mail Sorting

NETFLIX

Recommendation Engine

- **Applications of Computational Biology in Genomics**

Example Computational Biology Experiments and Tasks:

Example Computational Biology Experiments and Tasks:

- **Example 1: Identify Variants Associated with a Predisposition to ALS**

Amyotrophic Lateral Sclerosis (ALS)

- Also known as Lou Gehrig's disease
- Progressive neurodegenerative disease causing muscle weakness and atrophy due to degeneration of motor neurons
- ~5,600 new cases in the US annually
- Median survival time from onset to death is 39 months

Photo: phillysportshistory.com

89% of sporadic ALS cases are not explained by known genetic alterations

Heterogeneous symptoms, progression, and genetic mutations

↓

20+ Distinct ALS Subtypes

Genetic subtype	Chromosomal locus	Gene	Protein	Onset	Inheritance	Clinical feature	Other diseases caused by the gene
ALS1	21q22.1	SOD1	Cu/Zn SOD-1	Adult	AD/AR	Typical ALS	NA
ALS2	2q33-2q35	Alsin	Alsin	Juv	AR	Slowly progressive, predominantly UMN signs like limb, & facial spasticity	PLS IAHSP
ALS3	18q21	Unknown	Unknown	Adu	AD	Typical ALS with limb onset especially lower limb	NA
ALS4	9q34	SETX	Senataxin	Juv	AD	Slowly progressive, distal hereditary motor neuropathy with pyramidal signs	SCAR 1 and AOA2
ALS5	15q15-21	SPG 11	Spatacsin	Juv	AR	Slowly progressive	HSP
ALS6	16p11.2	FUS	Fused in Sarcoma	Juv/Adu	AD/AR	Typical ALS	NA
ALS8	20q13.3	VAPB	VAPB	Adu	AD	Typical and atypical ALS	SMA
ALS9	14q11.2	ANG	Angiogenin	Adu	AD	Typical ALS, FTD and Parkinsonism	NA
ALS10	1p36.2	TARDBP	DNA-binding protein	Adu	AD	Typical ALS	NA
ALS11	6q21	FIG 4	Phosphoinositide-5phosphatease	Adu	AD	Rapid progressive with prominent corticospinal tract signs	CMT 4 J
ALS12	10p13	OPTN	Optineurin	Adu	AD/AR	Slowly progressive with limb onset and predominant UMN signs	Primary Open Angle Glaucoma
ALS14	9p13.3	VCP	VCP	Adu	AD	Adult onset, with or without FTD	IBMPFD
ALS15/ ALSX	Xp11	UBQLN2	Ubiquilin 2	Adu/Juv	XD	UMN signs proceeding LMN signs	NA
ALS16	9p13.2-21.3	SIGMAR1	SIGMAR1	Juv	AR	Juvenile onset typical ALS	FTD
ALS-FTD1	9q21-22	unknown	unknown	Adu	AD	ALS with FTD	FTD
ALS-FTD2	9p21	C9ORF72	C9ORF72	Adu	AD	ALS with FTD	FTD
NA	2p13	DCTN1	Dynactin	Adu	AD	Distal hereditary motor neuropathy with vocal paresis	NA
Other rare-occurring ALS genes							
ALS3	18q21	Unknown	Unknown	Adu	AD	Typical ALS with limb onset especially lower limb	NA
ALS7	20ptel-p13	Unknown	Unknown	Adu	AD/AR	Typical ALS	NA
NA	12q22-23	DAO	DAO	Adu	AD	Typical ALS	NA

Figure: Chen, et al. 2013.

Neurotoxic Protein Aggregates in >95% of ALS Patients

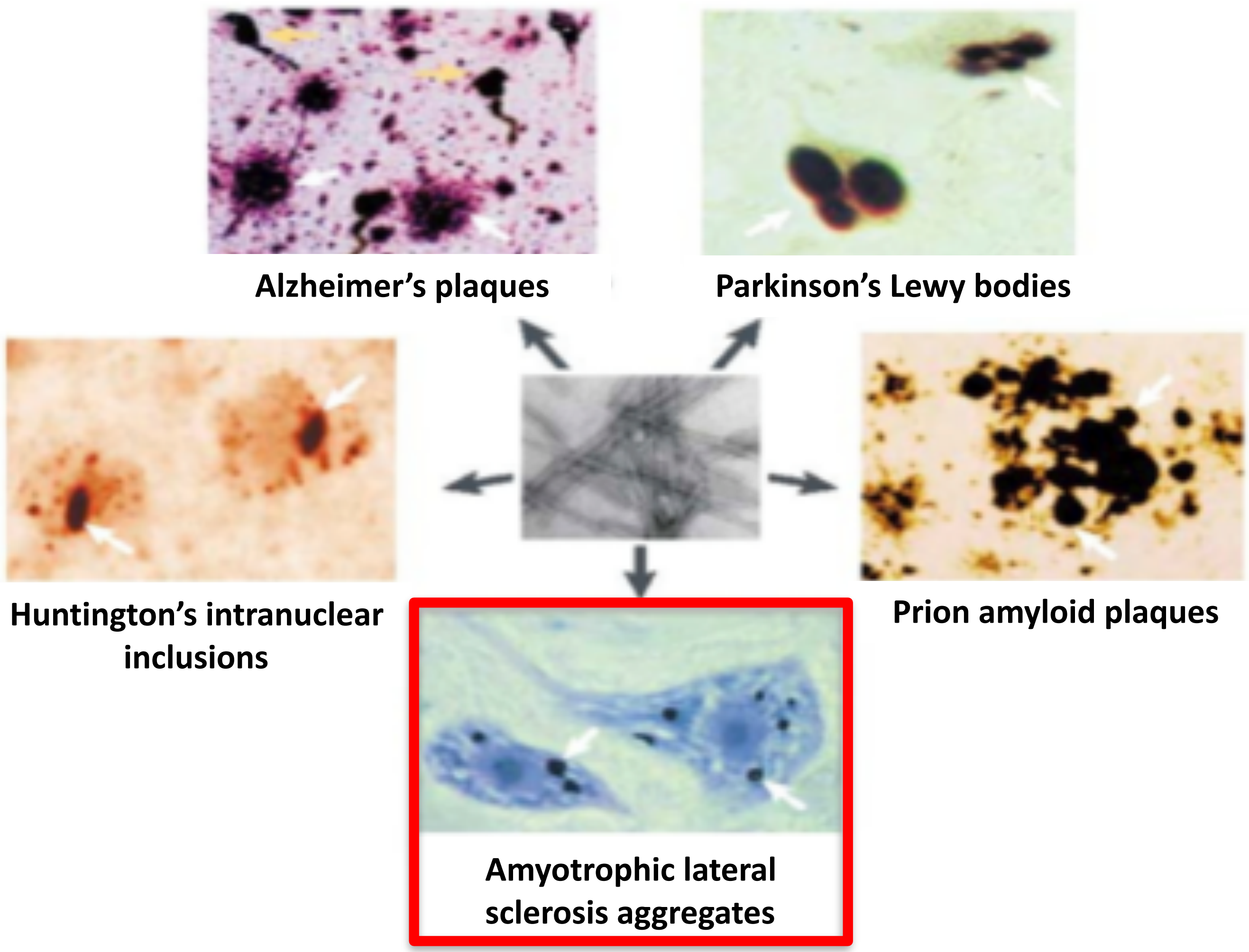


Image: QR Pharma.

ALS Genome Sequencing Consortium

biogen idec

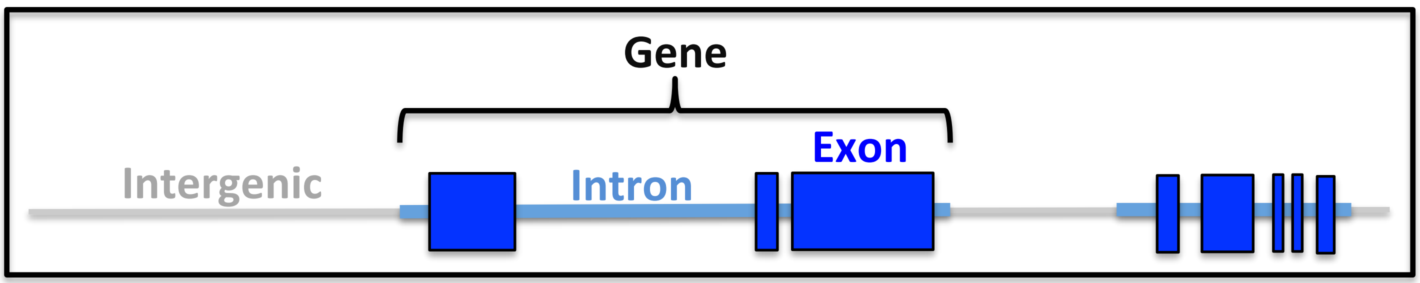
ALS Genome Sequencing Consortium

Project Goals

Identify rare coding variants and new genes/pathways associated with sporadic ALS

Identifying Variants with Exome Sequencing

- Exome Sequencing: Identify variation in coding regions (genes)
- Advantage: Interpretability and lower cost compared to whole genome sequencing



Compare Variants

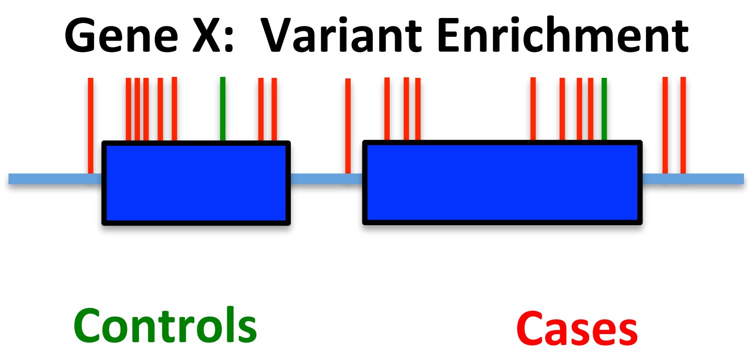
CTACGATCGA Control Group (n=~6500)

CTAGGATCGA Affected Patient Group (n=~3000)

Gene Burden Testing of Rare Variants

Count Qualifying Variants:

- Count qualifying variants in a gene-based collapsing analysis including exons meeting coverage benchmarks
  Example: Loss of Function (splice, nonsense, or frameshift)



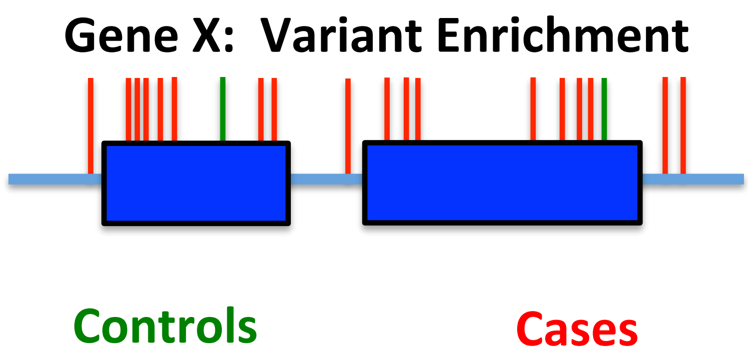
Compare Frequency Distributions

- Significant enrichment of qualifying variants between groups

Gene Burden Testing of Rare Variants

Count Qualifying Variants:

- Count qualifying variants in a gene-based collapsing analysis including exons meeting coverage benchmarks
  Example: Loss of Function (splice, nonsense, or frameshift)



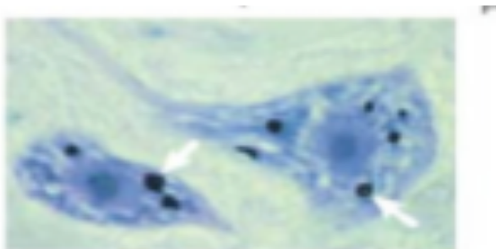
✔ *SOD1*: First gene associated with familial ALS (enzyme that destroys free superoxide radicals)

Compare Frequency Distributions

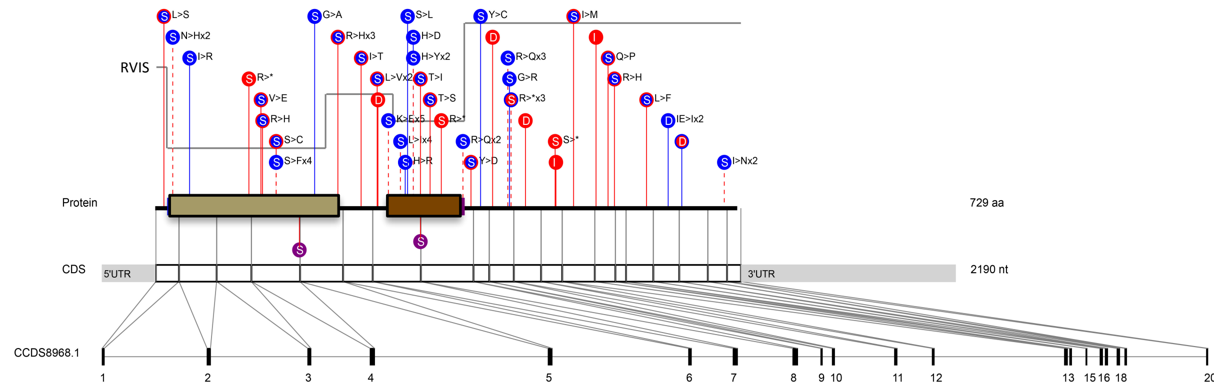
- Significant enrichment of qualifying variants between groups

Identifying Novel ALS Genes: *TBK1*

- TBK1 interacts with other ALS-associated genes that play important roles in autophagy and inflammation



Amyotrophic lateral sclerosis aggregates

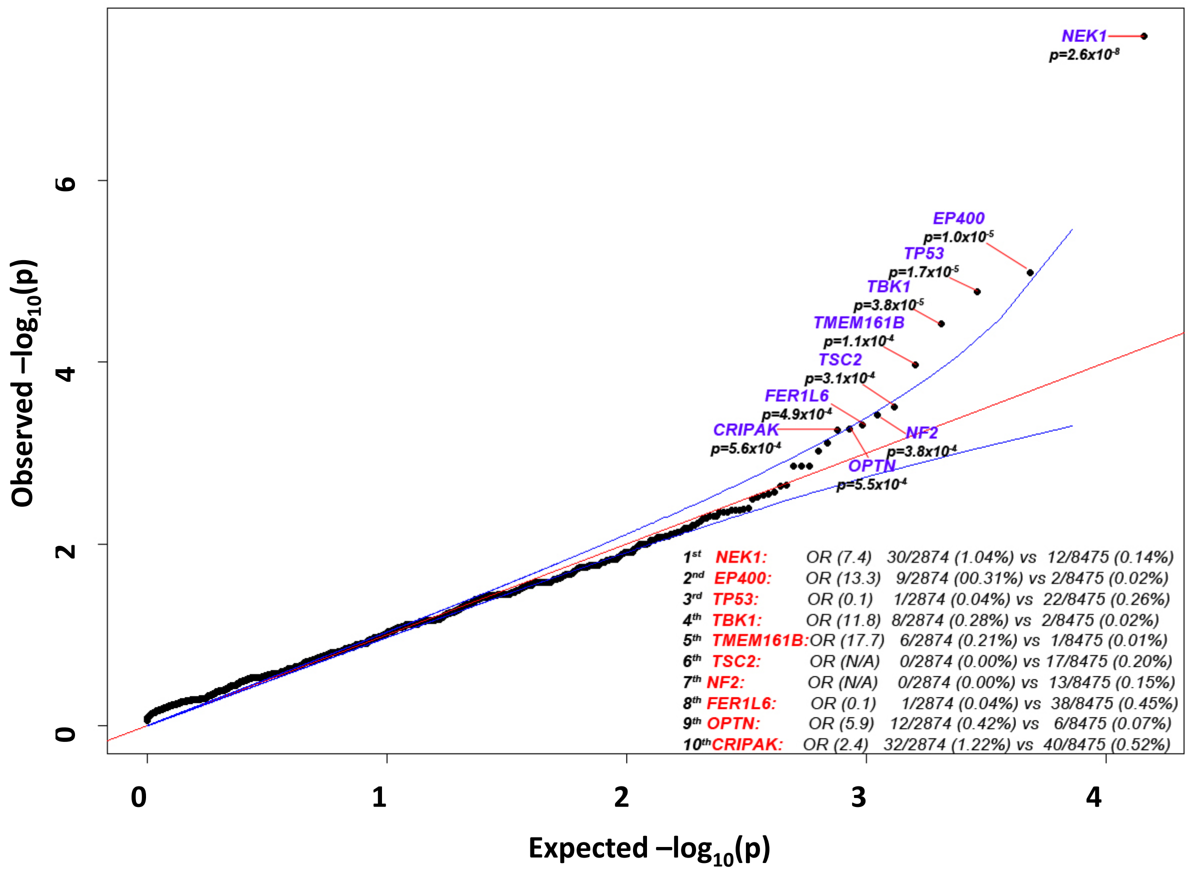


LOF variant
Missense variant
Splice variant
Case variant
Control variant
Case/control variant

- Non-benign variants: 1.097% of cases
- LoF variants: 0.382% of cases

Cirulli & Lasseigne, et al. 2016. Wild, et al. 2011. Gleason, et al. 2011. Pilli, et al. 2012. Kachaner, et al. 2012. Komatsu, et al. 2012.

Identifying Novel ALS Genes: *NEK1*

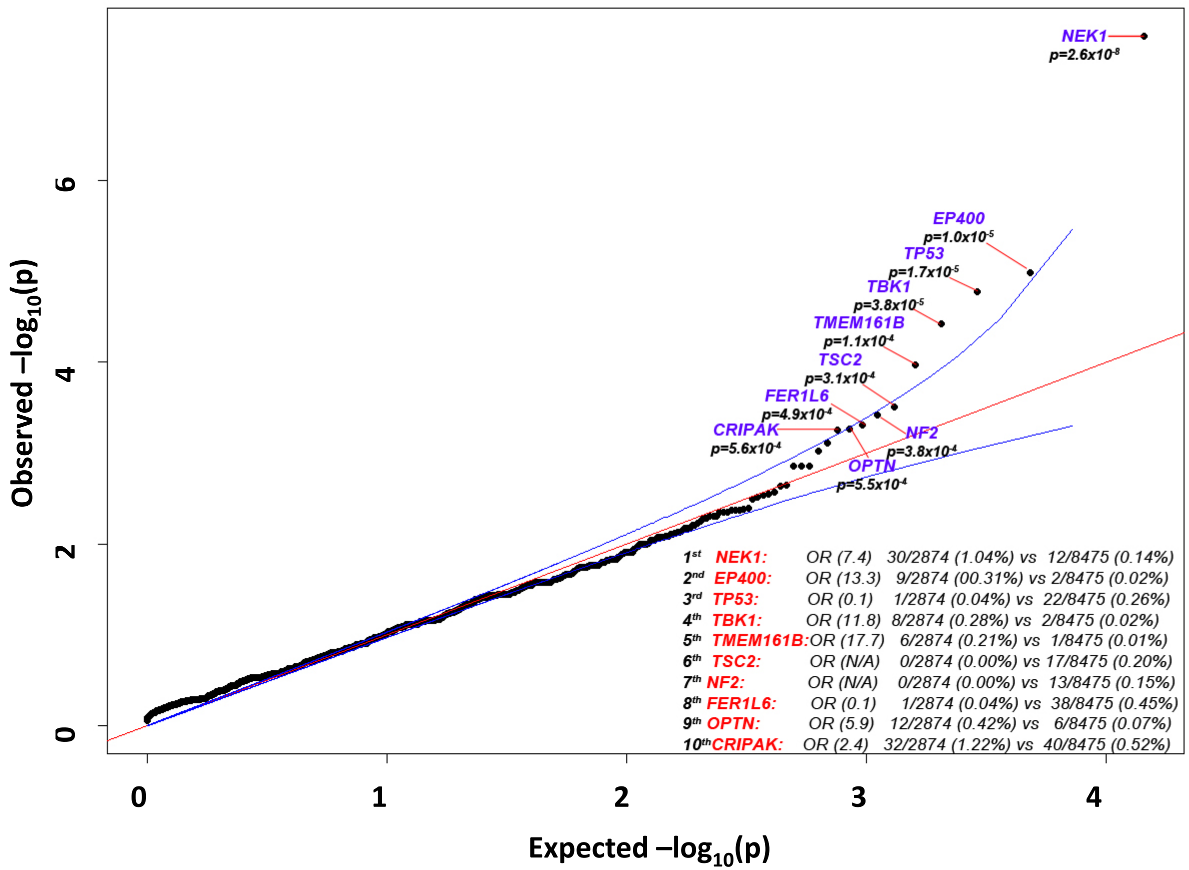


QQ plot: Dominant LoF model

Cirulli & Lasseigne, et al. 2015.

Identifying Novel ALS Genes: *NEK1*

- *NEK1*: multi-functional kinase, role in cilia formation and centrosome function, never previously linked to ALS
- Follow-up cohort (1,318 additional cases and 2,371 additional controls) further supports *NEK1*'s role in ALS predisposition



QQ plot: Dominant LoF model

Cirulli & Lasseigne, et al. 2015.

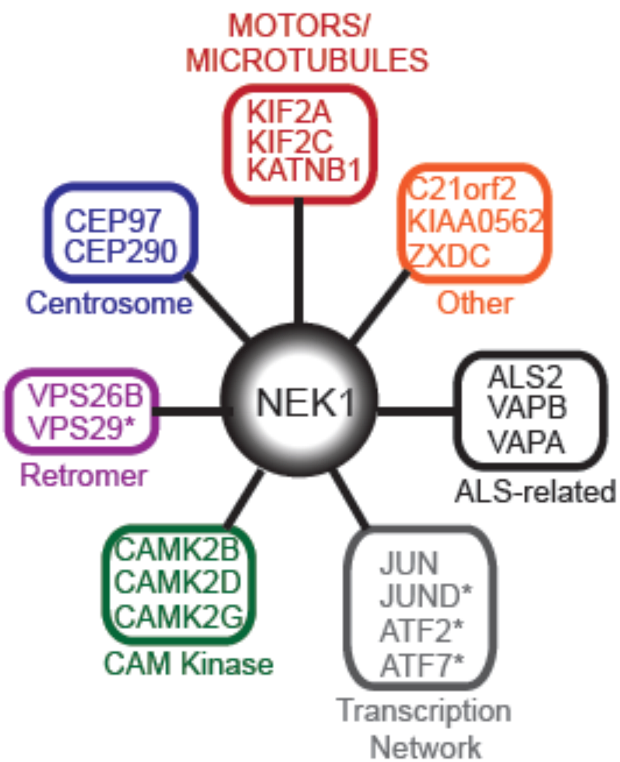
NEK1 associates with ALS2 and VAPB

- To investigate binding partners, we performed an unbiased screen of NEK1-interacting proteins in human kidney epithelial cells via AP-MS

Cirulli & Lasseigne, et al. 2015.

NEK1 associates with ALS2 and VAPB

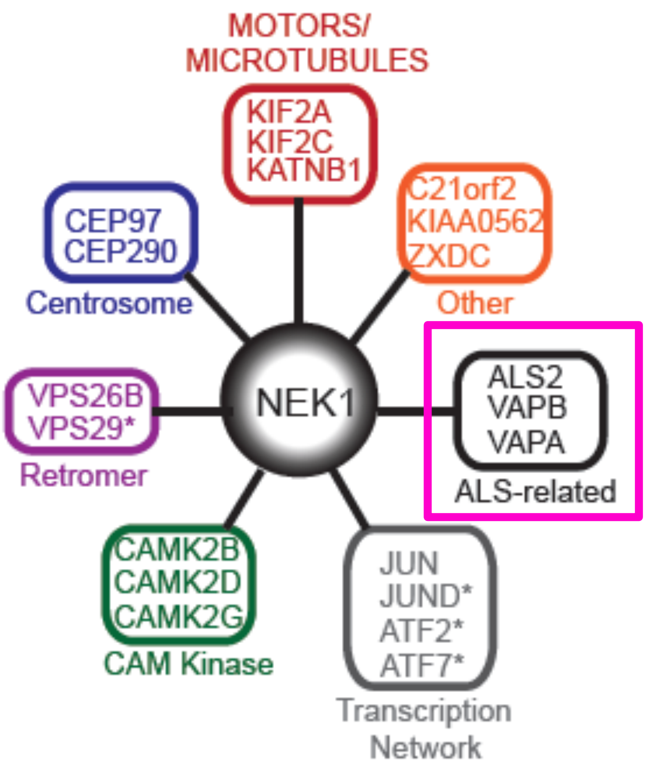
- To investigate binding partners, we performed an unbiased screen of NEK1-interacting proteins in human kidney epithelial cells via AP-MS



Cirulli & Lasseigne, et al. 2015.

NEK1 associates with ALS2 and VAPB

- To investigate binding partners, we performed an unbiased screen of NEK1-interacting proteins in human kidney epithelial cells via AP-MS
- Interactions validated by immunoprecipitation followed by western blotting of co-expressed proteins in neuronal NSC-34 cells

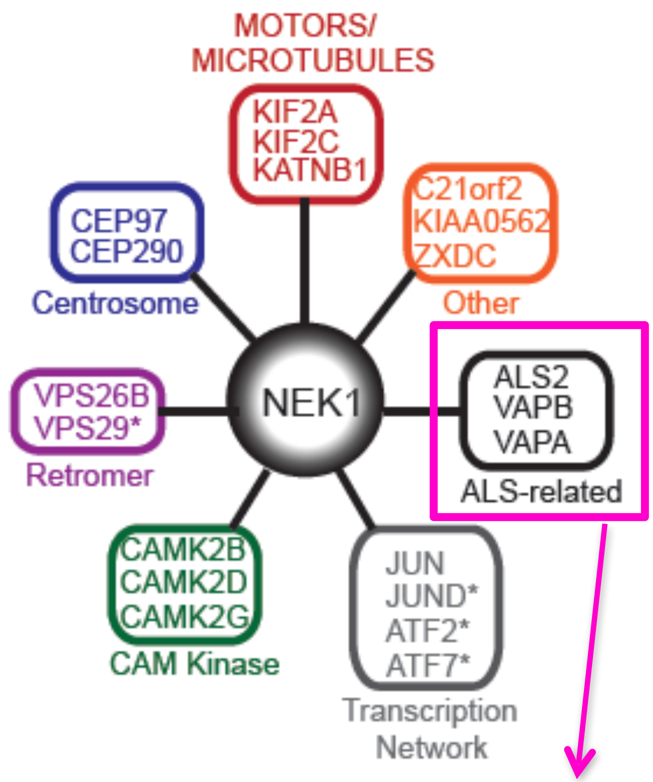


Recessive causes of ALS when mutated:
ALS2: RAB guanine nucleotide exchange factor
VAPB/VAPA: transmembrane proteins that transfer lipids from the ER to the plasma membrane

Cirulli & Lasseigne, et al. 2015.

NEK1 associates with ALS2 and VAPB

- To investigate binding partners, we performed an unbiased screen of NEK1-interacting proteins in human kidney epithelial cells via AP-MS
- Interactions validated by immunoprecipitation followed by western blotting of co-expressed proteins in neuronal NSC-34 cells
- Suggests *NEK1* may contribute to ALS through multiple mechanisms:
  - ALS2 and VAPB control cytoplasmic trafficking of endosomes and lipids in diverse cell lineages, respectively, both biological functions that are now appreciated as important in other neurodegenerative diseases



Recessive causes of ALS when mutated:
ALS2: RAB guanine nucleotide exchange factor
VAPB/VAPA: transmembrane proteins that transfer lipids from the ER to the plasma membrane

Cirulli & Lasseigne, et al. 2015.

Example Computational Biology Experiments and Tasks:

- **Example 1: Identify Variants Associated with a Predisposition to ALS**
  - **Annotation**
  - **Databasing**
  - **Statistical Programming (analysis + visualization)**
  - **Hypothesis-Generating Research**

Example Computational Biology Experiments and Tasks:

- **Example 1: Identify Variants Associated with a Predisposition to ALS**
  - Annotation
  - Databasing
  - Statistical Programming (analysis + visualization)
  - Hypothesis-Generating Research
- **Example 2: Develop Biomarkers for Kidney Cancer Diagnosis**

Kidney Cancer Diagnosis and Treatment

- ~65,000 new cases in the United States each year (10^{th} most common cancer)

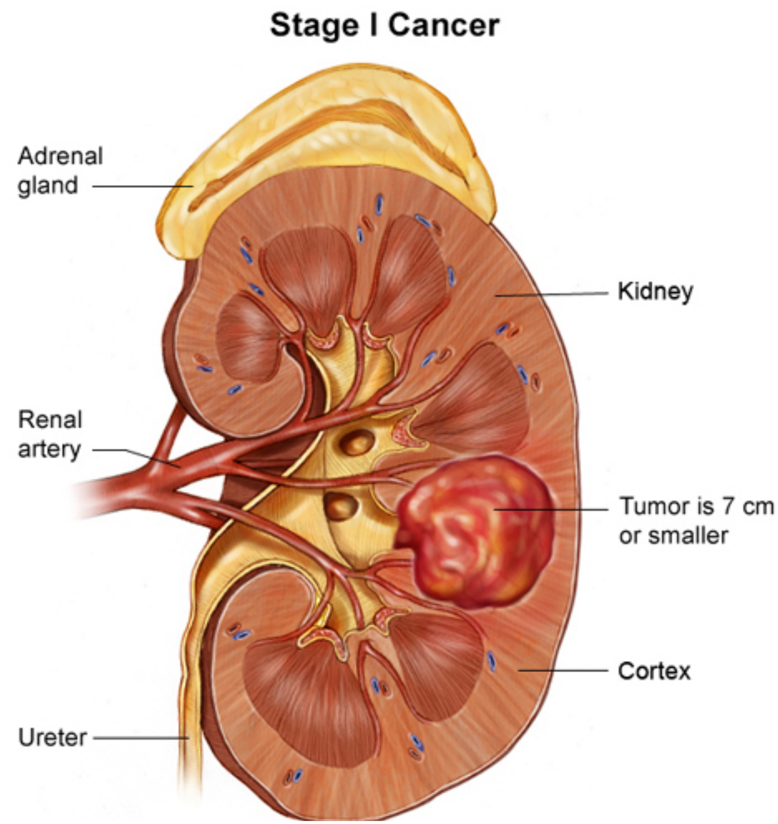
- If caught early, patients typically do well

- Treatment for advanced cases has improved in recent years, but the best drugs only increase disease free progression after resection by months and have harsh side effects

- Considered non-responsive to traditional radiation and chemotherapies

Kidney Cancer Diagnosis and Treatment

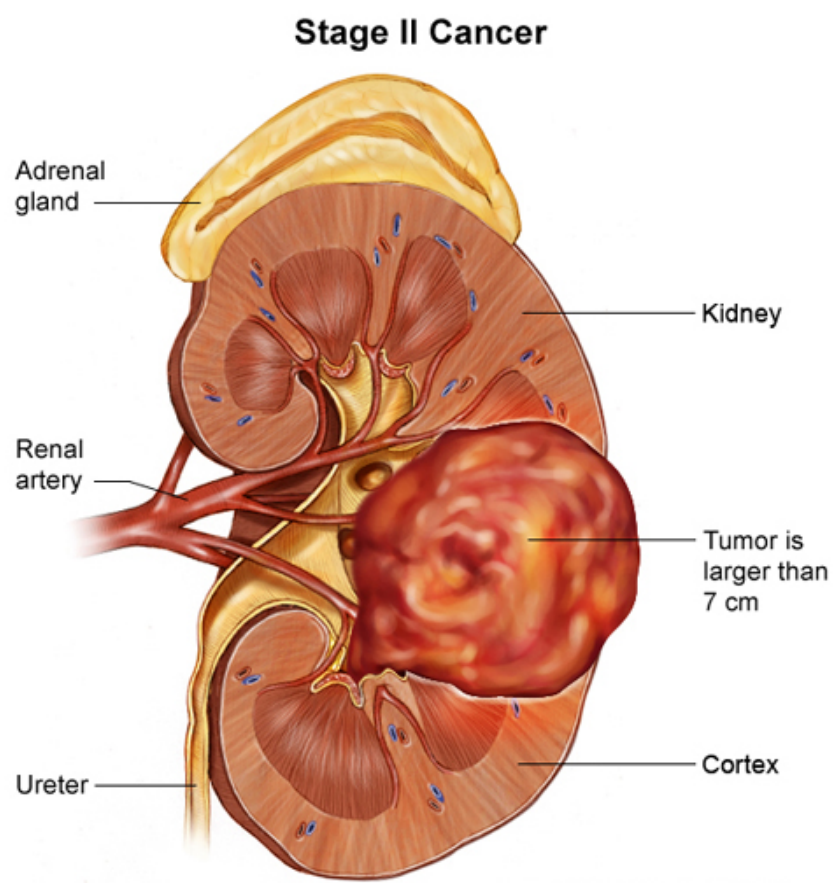
- ~65,000 new cases in the United States each year (10th most common cancer)
- If caught early, patients typically do well
- Treatment for advanced cases has improved in recent years, but the best drugs only increase disease free progression after resection by months and have harsh side effects
- Considered non-responsive to traditional radiation and chemotherapies



81% Survival at 5 years

Kidney Cancer Diagnosis and Treatment

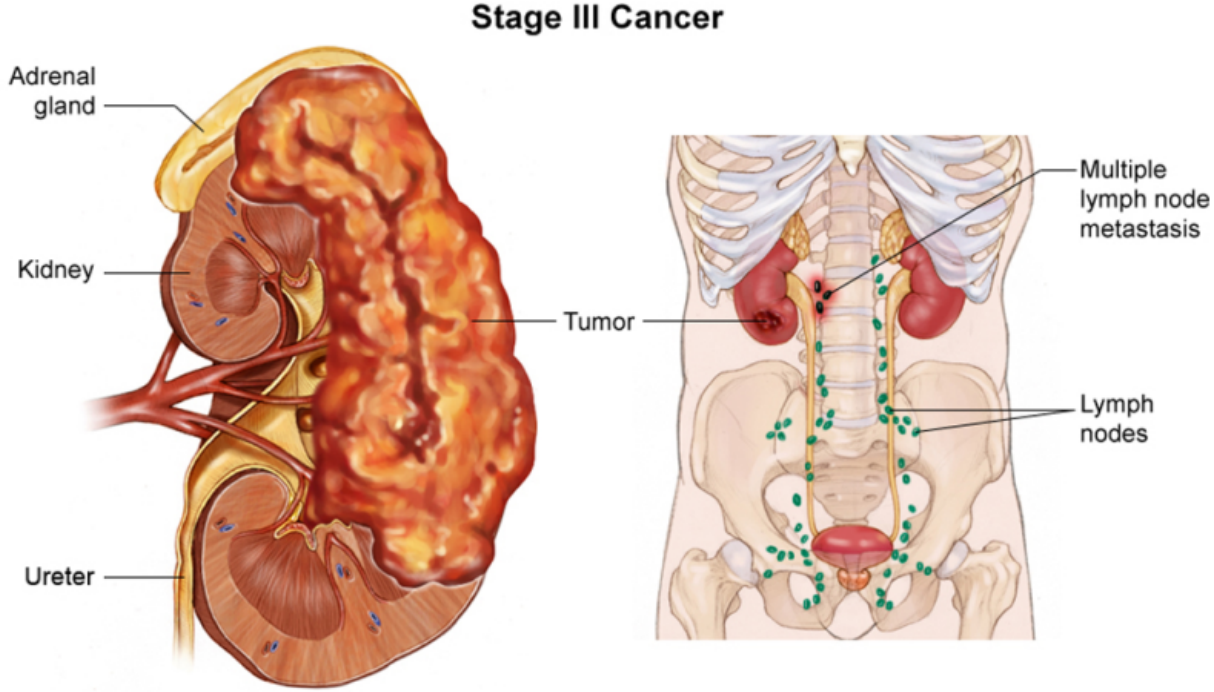
- ~65,000 new cases in the United States each year (10th most common cancer)
- If caught early, patients typically do well
- Treatment for advanced cases has improved in recent years, but the best drugs only increase disease free progression after resection by months and have harsh side effects
- Considered non-responsive to traditional radiation and chemotherapies



74% Survival at 5 years

Kidney Cancer Diagnosis and Treatment

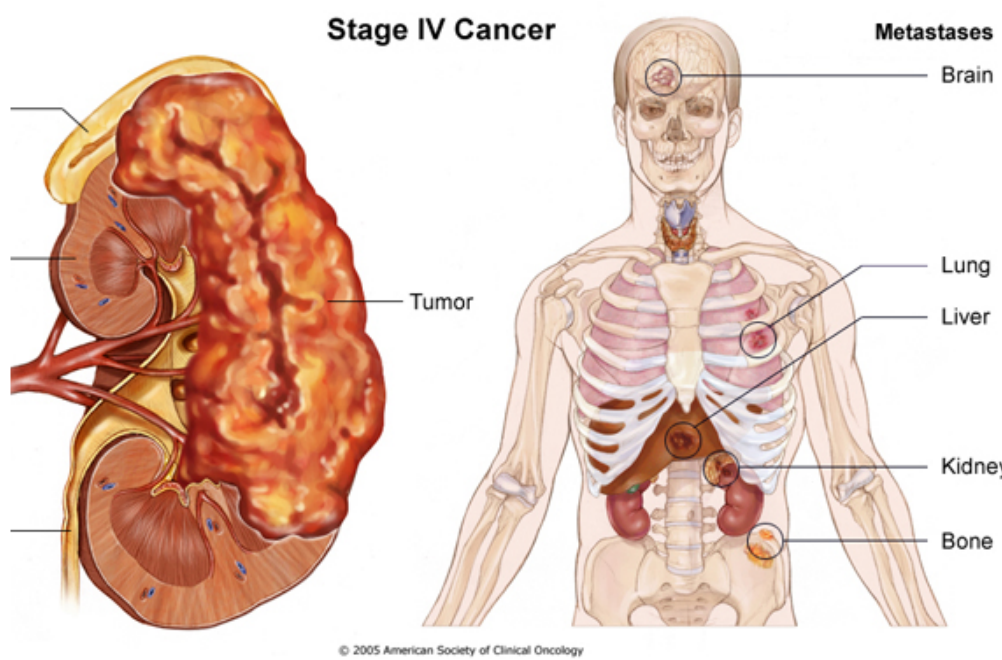
- ~65,000 new cases in the United States each year (10th most common cancer)
- If caught early, patients typically do well
- Treatment for advanced cases has improved in recent years, but the best drugs only increase disease free progression after resection by months and have harsh side effects
- Considered non-responsive to traditional radiation and chemotherapies



53% Survival at 5 years

Kidney Cancer Diagnosis and Treatment

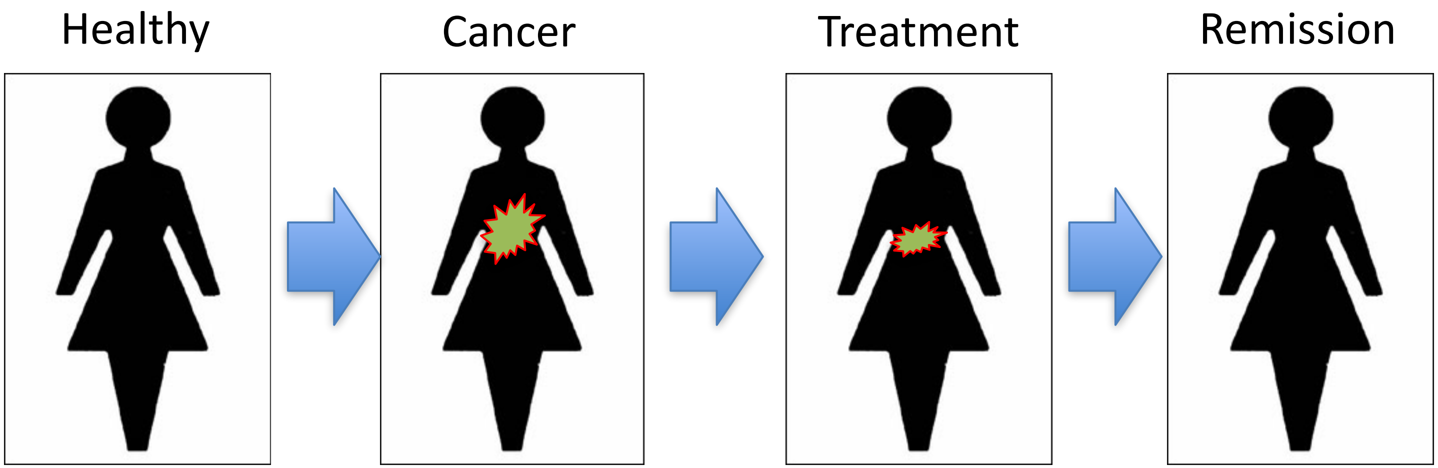
- ~65,000 new cases in the United States each year (10th most common cancer)
- If caught early, patients typically do well
- Treatment for advanced cases has improved in recent years, but the best drugs only increase disease free progression after resection by months and have harsh side effects
- Considered non-responsive to traditional radiation and chemotherapies



8% Survival at 5 years

Cancer Genomics Research: Identifying Genomic Changes Relevant to Patient Care

101 Tumor and Normal Kidney Samples



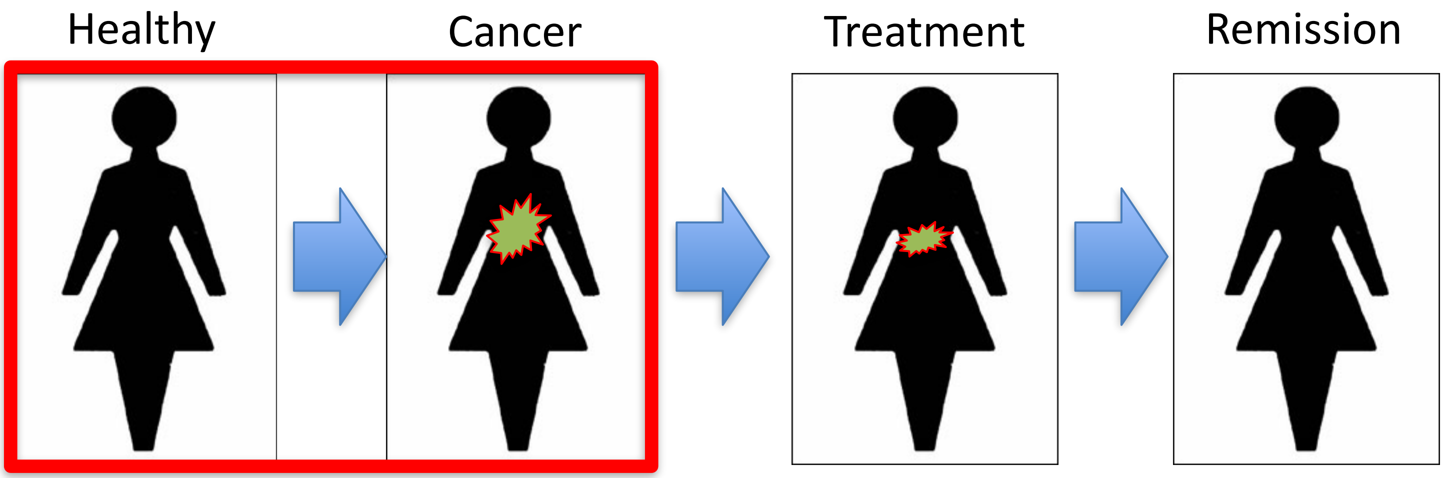
<u>Early Diagnosis</u>
cancer-specific molecular defects

<u>Prognosis & Treatment</u>
molecular defects predicting survival or personalized treatment

<u>Treatment Efficacy</u>
monitor molecular signatures of response or resistance to treatment

Cancer Genomics Research:
Identifying Genomic Changes Relevant to Patient Care

101 Tumor and Normal Kidney Samples



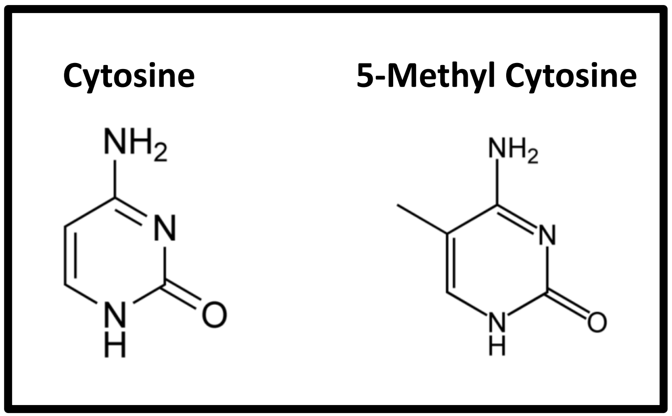
<u>Early Diagnosis</u>
cancer-specific
molecular defects

<u>Prognosis &</u>
<u>Treatment</u>
molecular
defects
predicting survival
or personalized
treatment

<u>Treatment Efficacy</u>
monitor molecular
signatures of response or
resistance to treatment

DNA Methylation at CpGs: The "Fifth" Base

Regulates biological processes without altering genetic blueprint (DNA sequence)

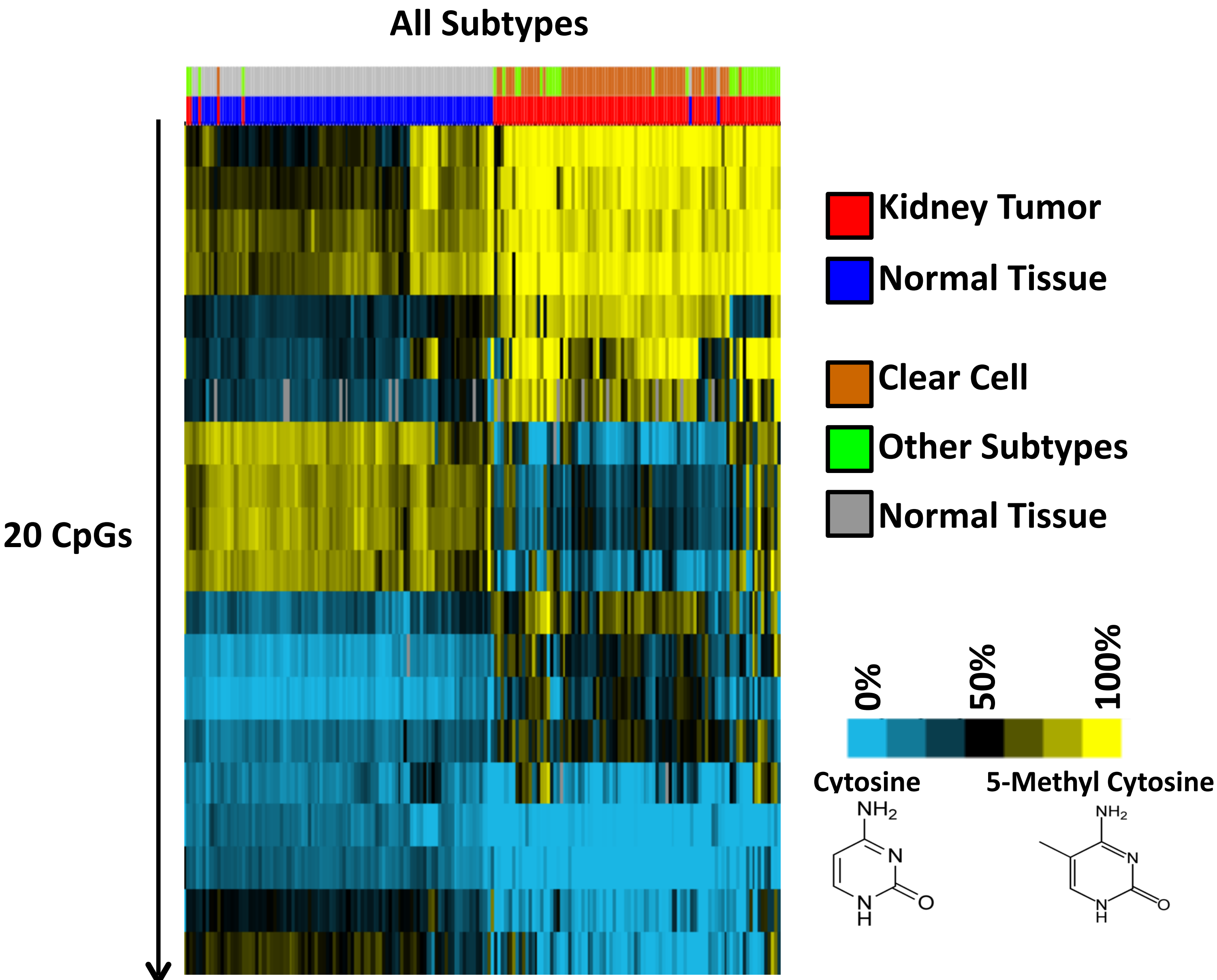


DNA Methylation Functions:

- DNA-protein interactions
- Cellular differentiation
- Transposable element suppression
- X-inactivation
- Genomic imprinting
- Gene regulation

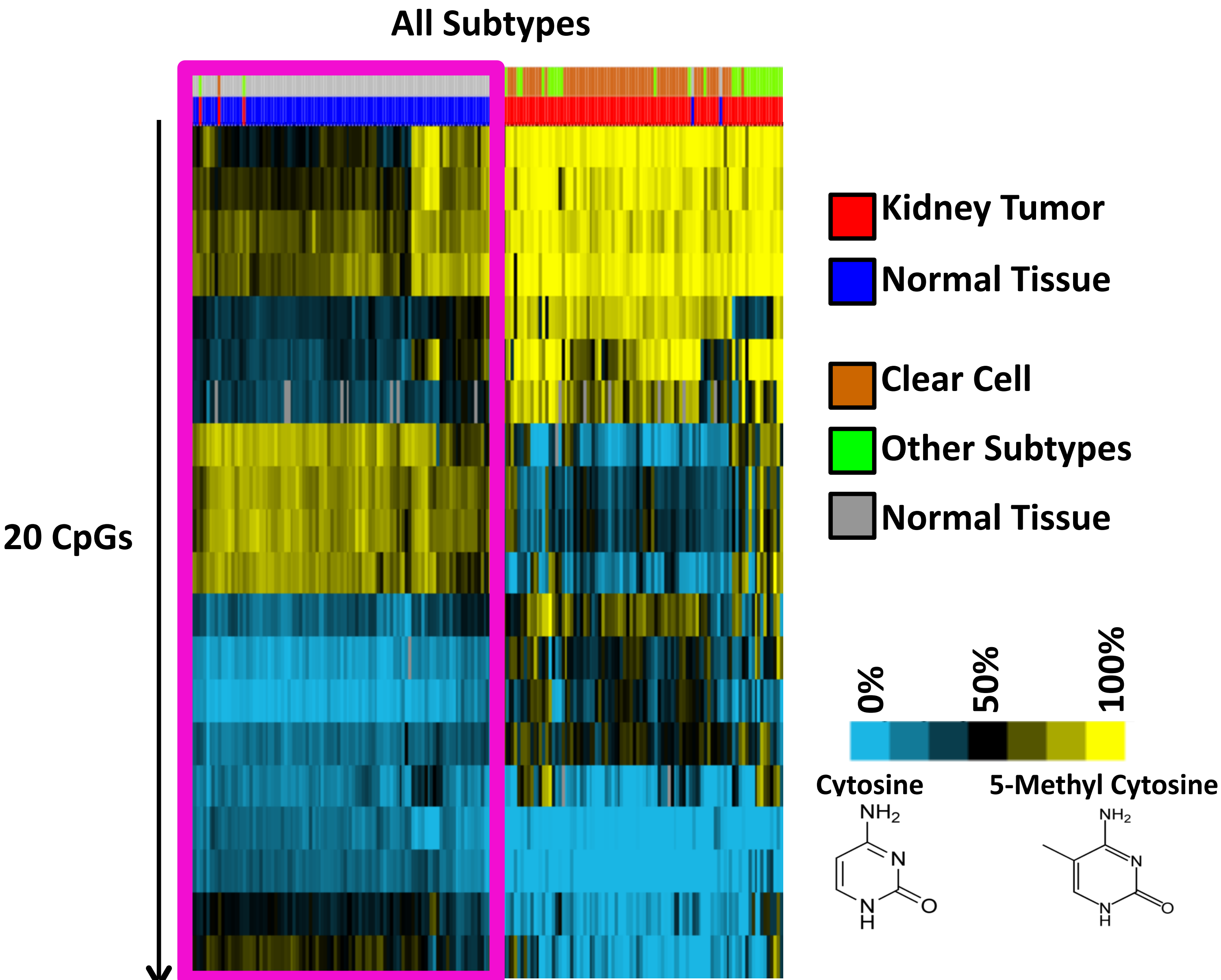
- DNA methylation as early diagnostic biomarkers:
  - Early events in carcinogenesis
  - Stable DNA mark and can be quantitatively measured

Diagnostic DNA Methylation Biomarkers: Kidney Cancer



Lasseigne, et al. BMC Cancer, 2014.

Diagnostic DNA Methylation Biomarkers: Kidney Cancer



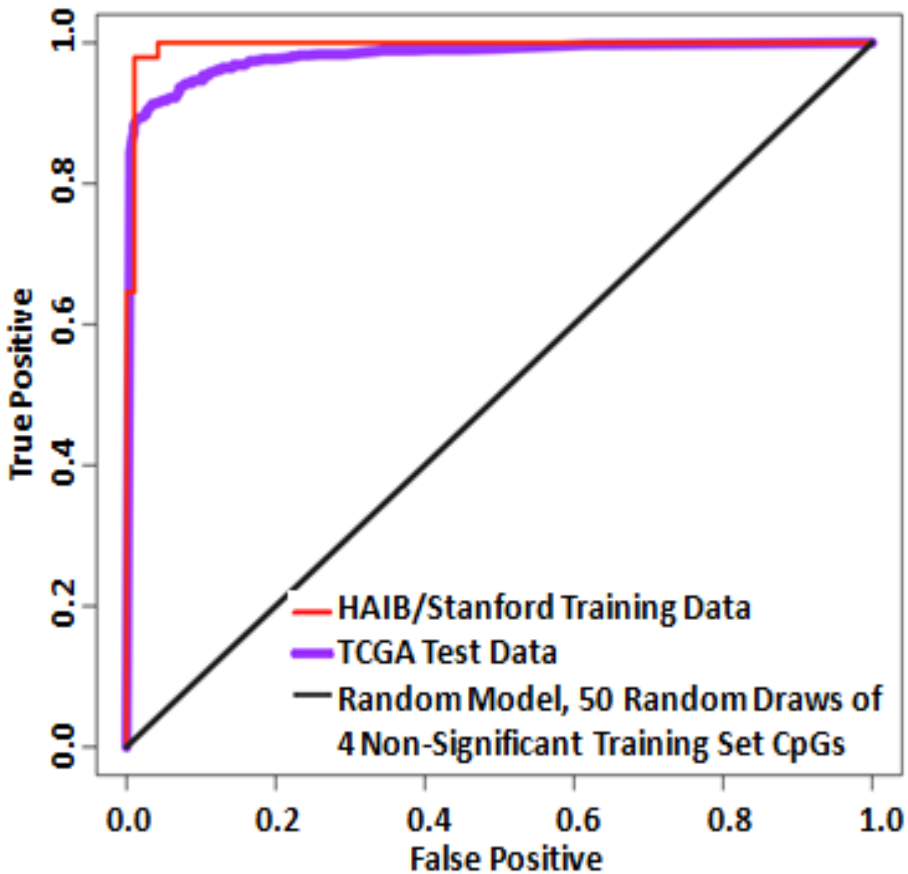
Lasseigne, et al. BMC Cancer, 2014.

Kidney Cancer Diagnostic Model

TCGA data as a validation test set:
-732 kidney cancer tissues
(3 subtypes!)
-410 normal kidney tissues

Lasseigne, et al. BMC Cancer, 2014.

Kidney Cancer Diagnostic Model



TCGA data as a validation test set:

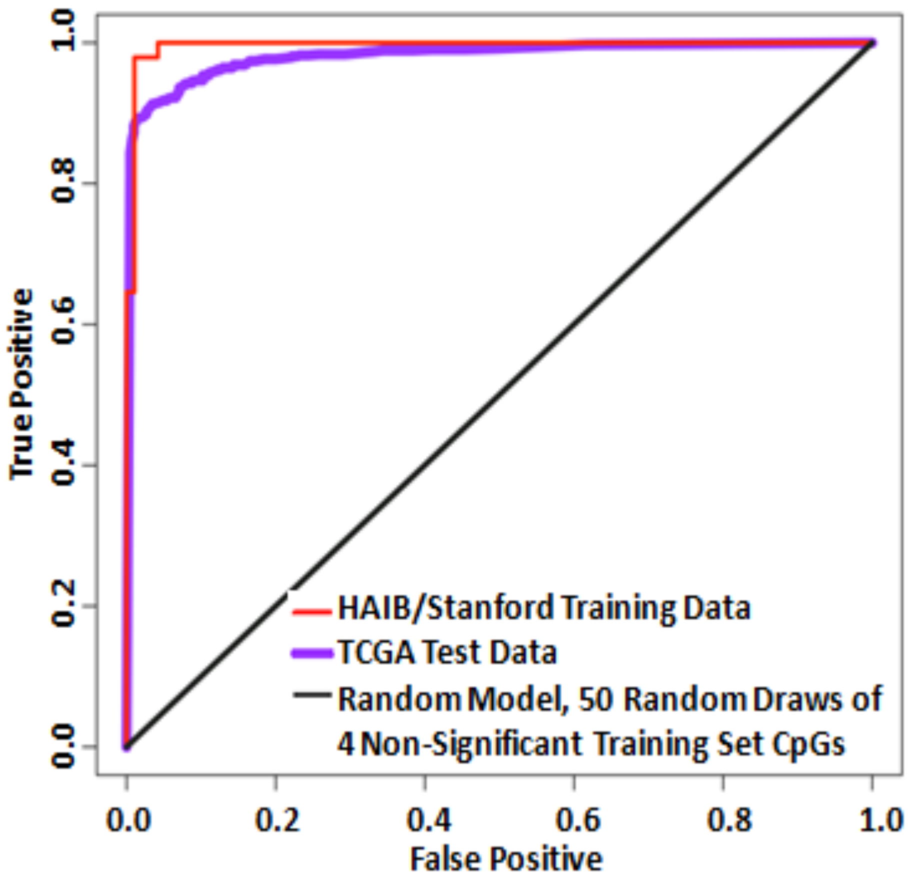
-732 kidney cancer tissues

(3 subtypes!)

-410 normal kidney tissues

Lasseigne, et al. BMC Cancer, 2014.

Kidney Cancer Diagnostic Model



TCGA data as a validation test set:

-732 kidney cancer tissues

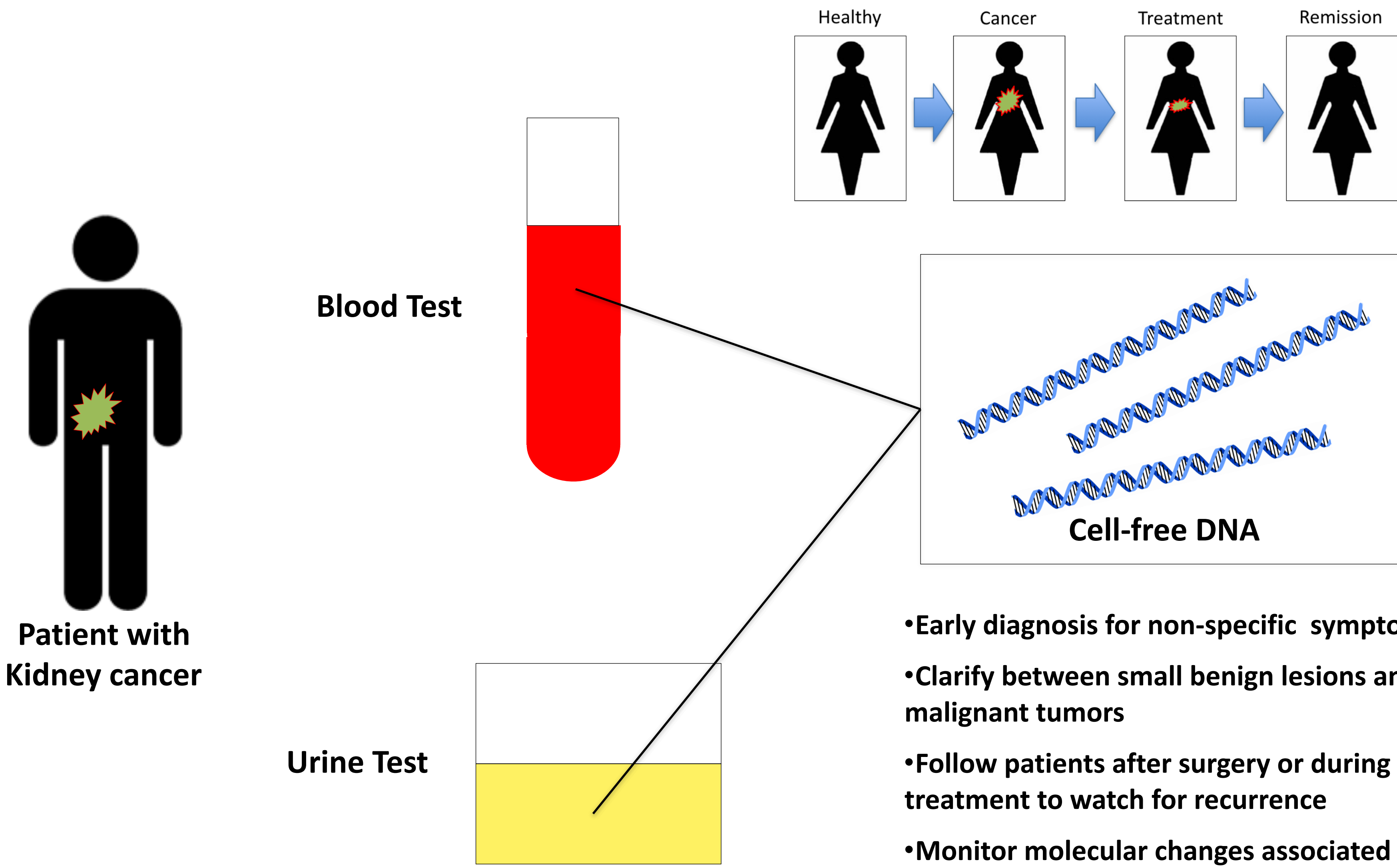
(3 subtypes!)

-410 normal kidney tissues

Correctly predict 87.8% of the normal tissues and 96.2% of the tumor tissues in the TCGA data

Lasseigne, et al. BMC Cancer, 2014.

From Bench To Bedside:
’liquid biopsies’ from peripheral fluids



- Early diagnosis for non-specific symptoms
- Clarify between small benign lesions and malignant tumors
- Follow patients after surgery or during treatment to watch for recurrence
- Monitor molecular changes associated with patient outcome

Example Computational Biology Experiments and Tasks:

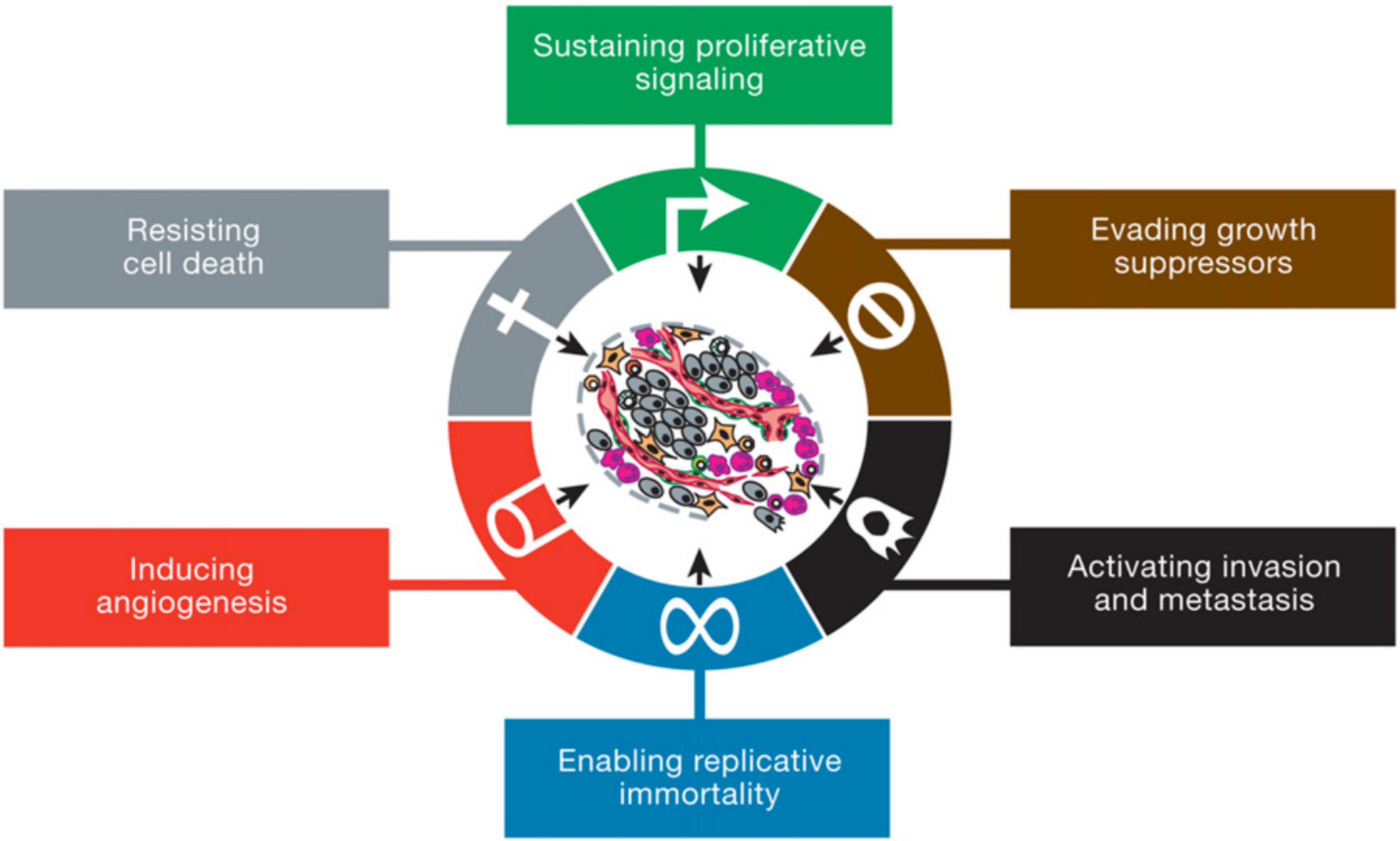
- Example 1: Identify Variants Associated with a Predisposition to ALS
  - Annotation
  - Databasing
  - Statistical Programming (analysis + visualization)
  - Hypothesis-Generating Research

- Example 2: Develop Biomarkers for Kidney Cancer Diagnosis
  - **Statistical Programming (analysis + visualization)**
  - **Machine Learning**
  - **Direct Clinical Application**
  - **Interdependent and Complementary ‘Wet’/‘Dry’ Biology Research**

Example Computational Biology Experiments and Tasks:

- Example 1: Identify Variants Associated with a Predisposition to ALS
  - Annotation
  - Databasing
  - Statistical Programming (analysis + visualization)
  - Hypothesis-Generating Research
- Example 2: Develop Biomarkers for Kidney Cancer Diagnosis
  - Statistical Programming (analysis + visualization)
  - Machine Learning
  - Direct Clinical Application
  - Interdependent and Complementary ‘Wet’/‘Dry’ Biology Research
- Example 3: Generate Pan-Cancer Models of Patient Prognosis

Cell proliferation is fundamental to cancer



Hanahan & Weinberg, Cell, 2000.

Measuring cell proliferation from RNA-seq data

- Venet, et al. cell proliferation 'metagene':

  -Median of top 1% of genes associated with PCNA expression (essential for replication)

<u>'Proliferative Index' (PI)</u>:

relative expression of proliferation-associated genes

PI/metaPCNA: Ge, et al, Genomics 2005 and Venet, et al, PLOS Computational Biology, 2011

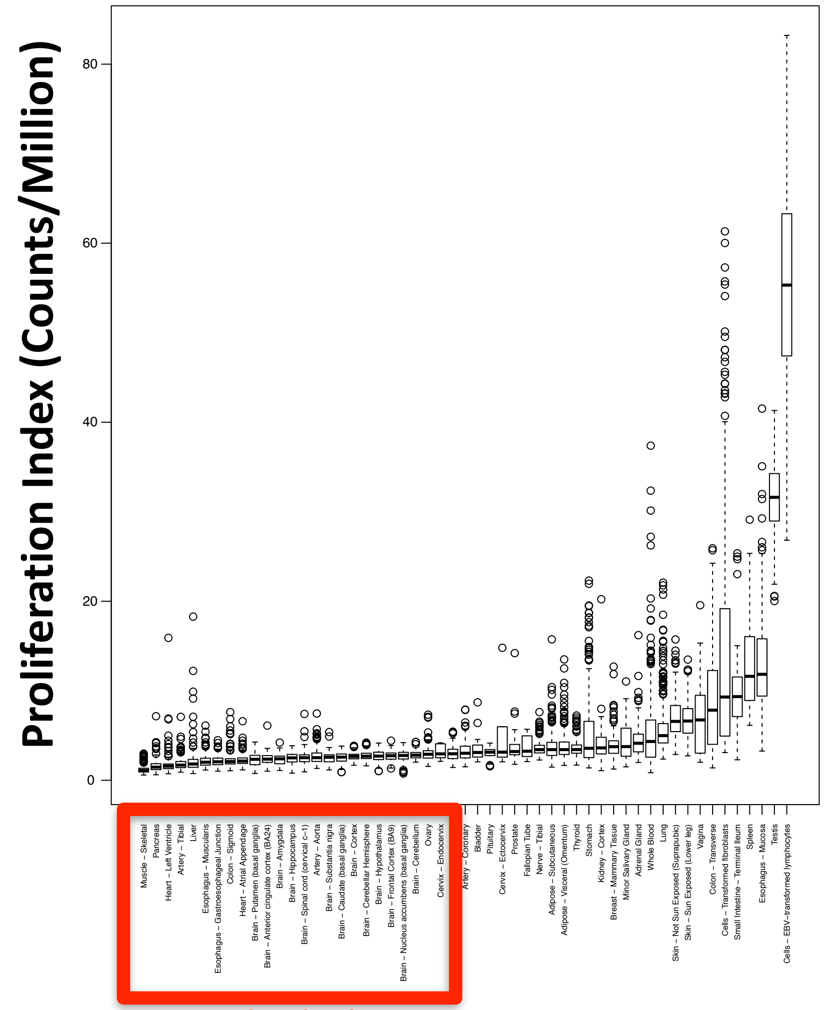
Measuring cell proliferation from RNA-seq data

- Venet, et al. cell proliferation ‘metagene’:
  - -Median of top 1% of genes associated with PCNA expression (essential for replication)

<u>‘Proliferative Index’ (PI)</u>:

relative expression of proliferation-associated genes

‘Healthy’ GTEx Tissues



<u>post-mitotic tissues</u>

ex. skeletal muscle

PI/metaPCNA: Ge, et al, Genomics 2005 and Venet, et al, PLOS Computational Biology, 2011

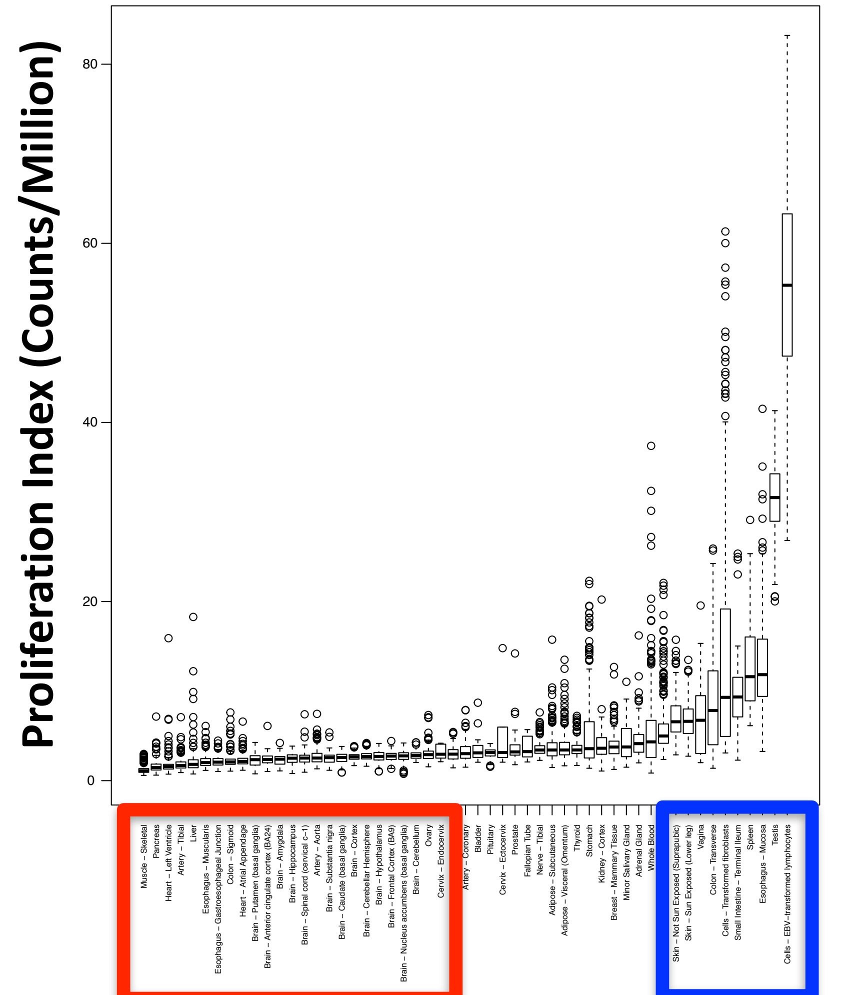
Measuring cell proliferation from RNA-seq data

- Venet, et al. cell proliferation ‘metagene’:
  - -Median of top 1% of genes associated with PCNA expression (essential for replication)

<u>‘Proliferative Index’ (PI)</u>:

relative expression of proliferation-associated genes

‘Healthy’ GTEx Tissues



<u>post-mitotic tissues</u>
ex. skeletal muscle

<u>high cell turnover</u>
ex. skin

PI/metaPCNA: Ge, et al, Genomics 2005 and Venet, et al, PLOS Computational Biology, 2011

Examine the role of cell proliferation in patient outcomes across cancers catalogued by The Cancer Genome Atlas

The TCGA Dataset

Abbreviation	Cancer	n
ACC	Adrenocortical Carcinoma	79
BLCA	Bladder Urothelial Carcinoma	385
BRCA	Breast Invasive Carcinoma	1038
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	393
ESCA	Esophageal Carcinoma	163
GBM	Glioblastoma Multiforme	144
HNSC	Head and Neck Squamous Cell Carcinoma	508
KIRC	Kidney Renal Clear Cell Carcinoma	525
KIRP	Kidney Renal Papillary Cell Carcinoma	266
LAML	Acute Myeloid Leukemia	148
LGG	Brain Lower Grade Glioma	463
LIHC	Liver Hepatocellular Carcinoma	355
LUAD	Lung Adenocarcinoma	493
LUSC	Lung Squamous Cell Carcinoma	479
MESO	Mesothelioma	72
OV	Ovarian Serous Cystadenocarcinoma	252
PAAD	Pancreatic Adenocarcinoma	167
SARC	Sarcoma	248
STAD	Stomach Adenocarcinoma	403

Total: 19 Cancers, 6581 Patients

‘Common Survival Genes’ across 19 cancers

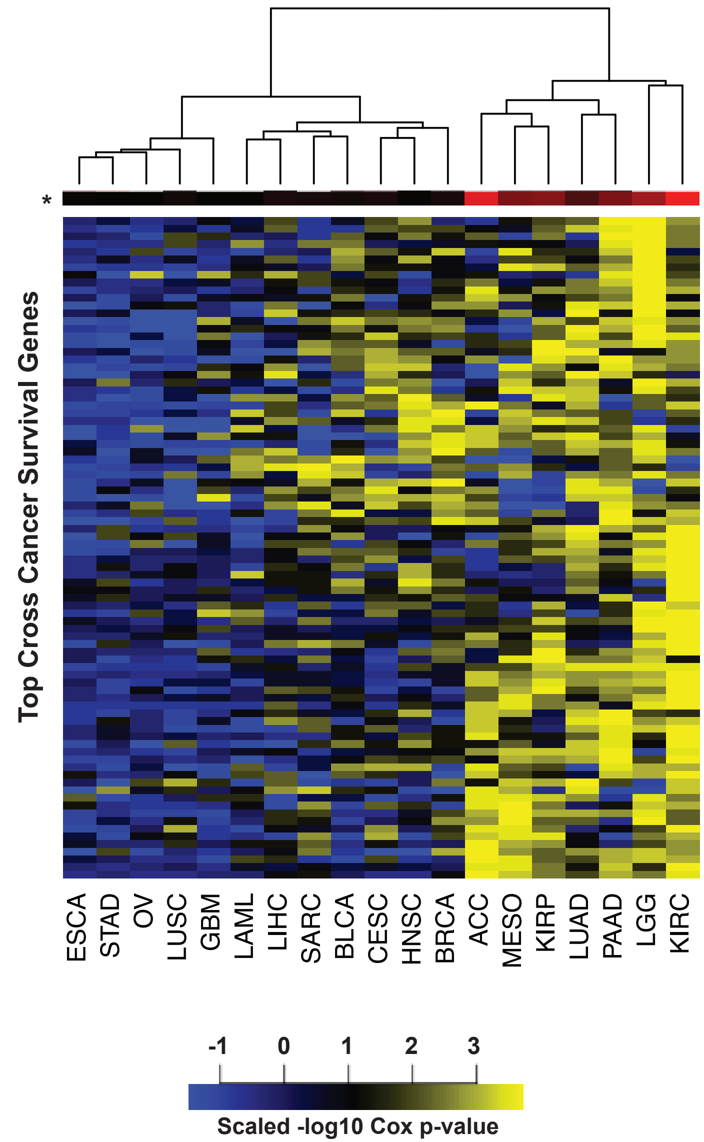
- **‘Common Survival Genes’**
  Cox regression uncorrected p-value <0.05 for a gene in at least 9/19 cancers:
  - **84 genes, enriched for proliferation-related processes including mitosis, cell and nuclear division, and spindle formation**

Ramaker & Lasseigne, et al. 2017.

‘Common Survival Genes’ across 19 cancers

- **‘Common Survival Genes’
Cox regression uncorrected p-value <0.05 for a gene in at least 9/19 cancers:**
    - **84 genes, enriched for proliferation-related processes including mitosis, cell and nuclear division, and spindle formation**

- **Clustering by Cox regression p-values:**

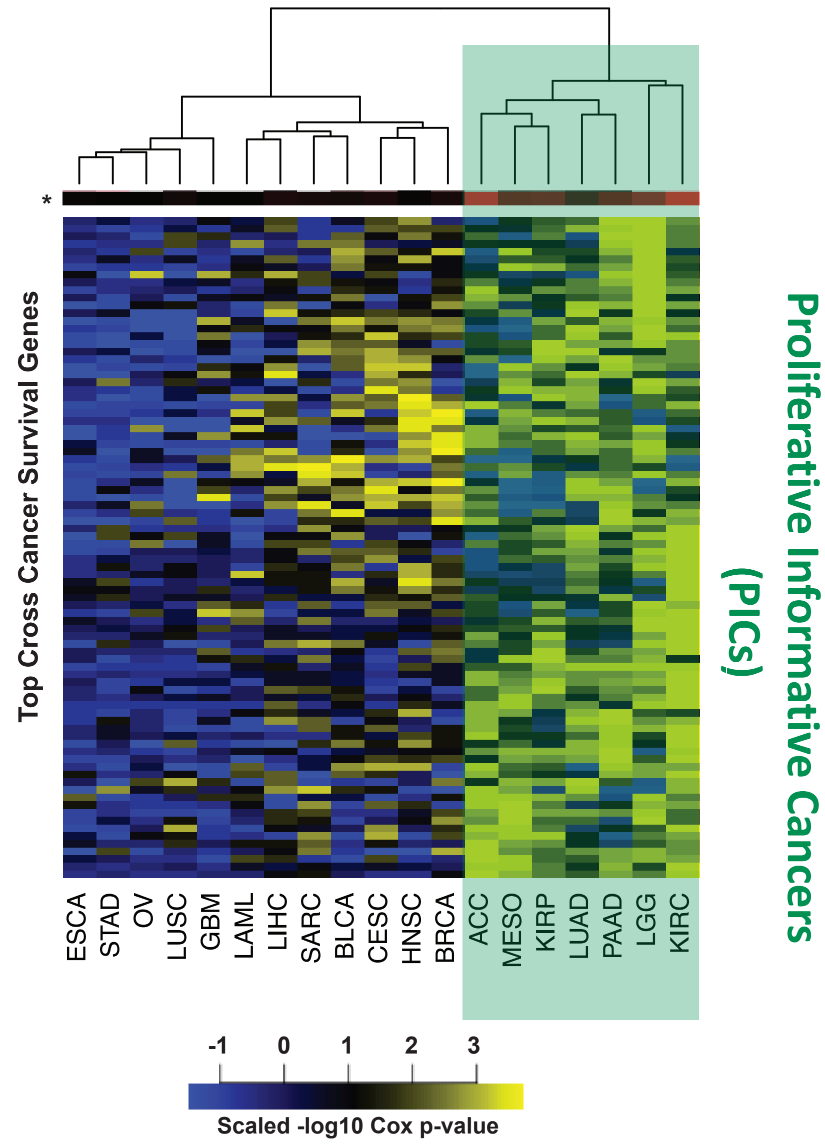


Ramaker & Lasseigne, et al. 2017.

‘Common Survival Genes’ across 19 cancers

- **‘Common Survival Genes’**
  Cox regression uncorrected p-value <0.05 for a gene in at least 9/19 cancers:
    - **84 genes, enriched for proliferation-related processes including mitosis, cell and nuclear division, and spindle formation**

- **Clustering by Cox regression p-values:**
  7 ‘Proliferative Informative Cancers’ and 12 ‘Non-Proliferative Informative Cancers’

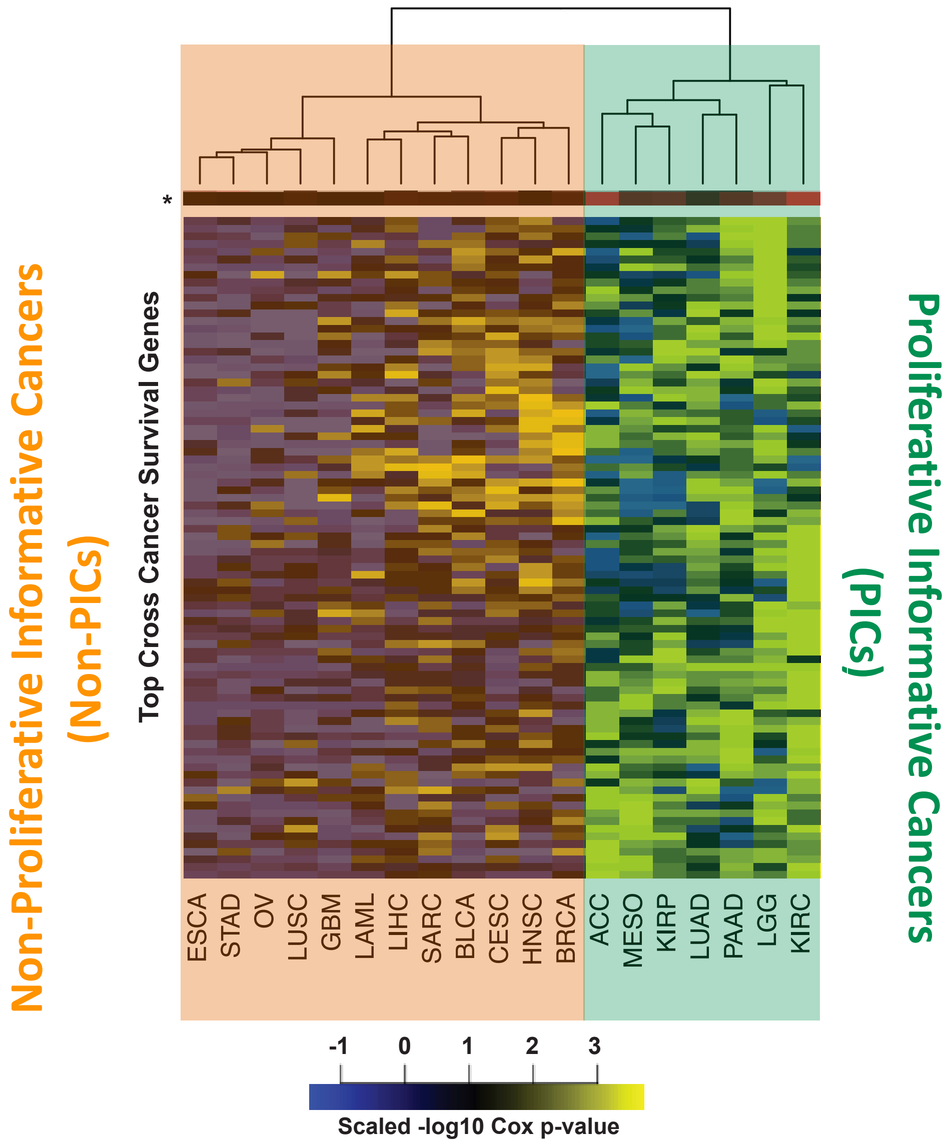


Ramaker & Lasseigne, et al. 2017.

‘Common Survival Genes’ across 19 cancers

- **‘Common Survival Genes’**
  Cox regression uncorrected p-value <0.05 for a gene in at least 9/19 cancers:
  - **84 genes, enriched for proliferation-related processes including mitosis, cell and nuclear division, and spindle formation**

- **Clustering by Cox regression p-values:**
  7 ‘Proliferative Informative Cancers’ and 12 ‘Non-Proliferative Informative Cancers’



Ramaker & Lasseigne, et al. 2017.

Cross-Cancer Patient Outcome Model

~20,000 gene expression values
Cancer Patient Survival

Cox regression with LASSO feature selection

Survival~ -0.104 + 0.086*ADAM12 + 0.037*CKS1 - 0.088*CRYL1 + 0.056*DNA2 + 0.013*DONSON + 0.098*HJURP - 0.022*NDRG2 + 0.031*RAD54B + 0.040*SHOX2 - 0.155*SUOX

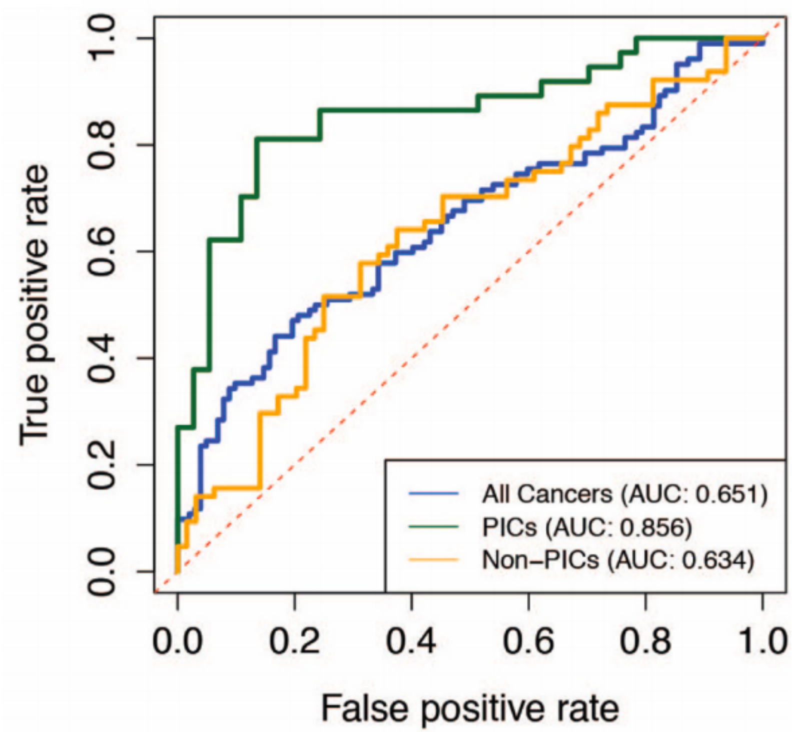
Ramaker & Lasseigne, et al. 2017.

Cross-Cancer Patient Outcome Model

~20,000 gene expression values
Cancer Patient Survival

Cox regression with LASSO feature selection

Survival~ -0.104 + 0.086*ADAM12 + 0.037*CKS1 - 0.088*CRYL1 + 0.056*DNA2 + 0.013*DONSON + 0.098*HJURP - 0.022*NDRG2 + 0.031*RAD54B + 0.040*SHOX2 - 0.155*SUOX



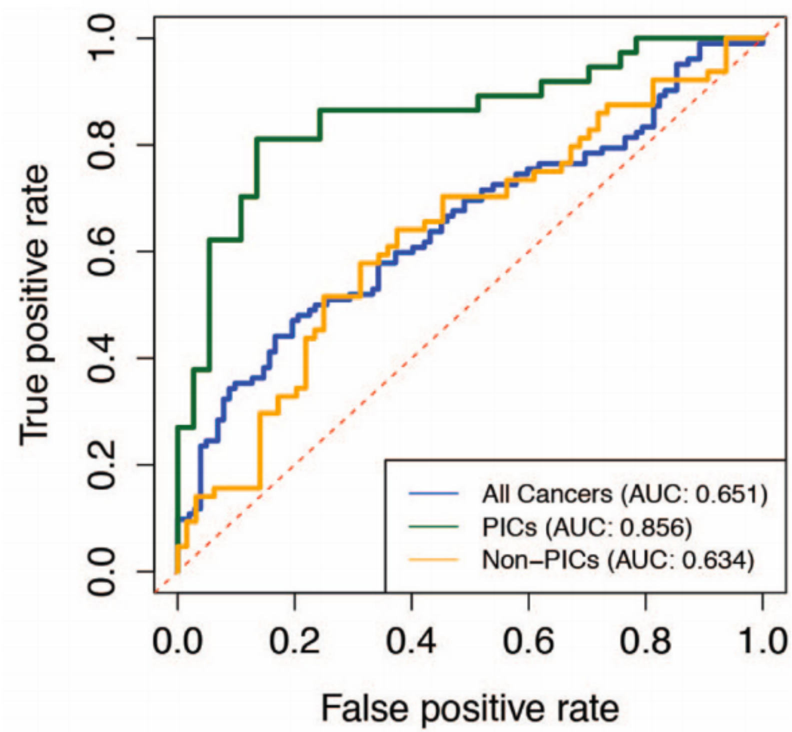
Ramaker & Lasseigne, et al. 2017.

Cross-Cancer Patient Outcome Model

~20,000 gene expression values
Cancer Patient Survival

Cox regression with LASSO feature selection

Survival~ -0.104 + 0.086*ADAM12 + 0.037*CKS1 - 0.088*CRYL1 + 0.056*DNA2 + 0.013*DONSON + 0.098*HJURP - 0.022*NDRG2 + 0.031*RAD54B + 0.040*SHOX2 - 0.155*SUOX



Utility:

- Predict patient prognosis
- Potentially inform on treatments
- Use of metagenes to infer molecular profiles from gene expression data

Ramaker & Lasseigne, et al. 2017.

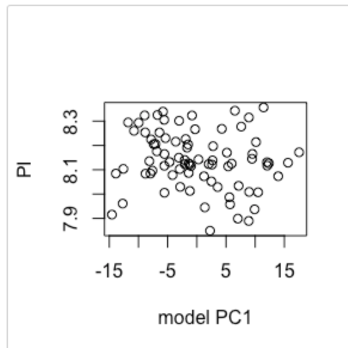
Analysis Packages: e.g. ‘ProliferativeIndex’

- Analytical R package available on CRAN and GitHub (continuous integration with Travis CI)
- Documented functions and a vignette with examples
- Provides users with R functions for calculating and analyzing the proliferative index (PI) from an RNA-seq dataset

compareModeltoPI function

The function `compareModeltoPI` will take, as input, the user's data and model identifiers and compare to PI:

```
modelComparison<-compareModeltoPI(exampleTCGAData, proliferativeIndices)
```



	SpearmanRho	SpearmanPvalue	PCAPropOfVariance
PC1	-0.1706670	0.1324595	0.44799
PC2	0.1009250	0.3753928	0.08169
PC3	0.0541626	0.6347829	0.04912
PC4	-0.2893379	0.0099231	0.04025
PC5	-0.1059396	0.3520354	0.03288
PC6	-0.1822055	0.1079531	0.02686
PC7	-0.4116115	0.0001866	0.02272
PC8	0.1556962	0.1703124	0.02070
PC9	-0.2600779	0.0208781	0.01918
PC10	-0.0916504	0.4210060	0.01803

Ramaker & Lasseigne, et al. *Oncotarget*, 2017.

Example Computational Biology Experiments and Tasks:

- **Example 1: Identify Variants Associated with a Predisposition to ALS**
  - Annotation
  - Databasing
  - Statistical Programming (analysis + visualization)
  - Hypothesis-Generating Research

- **Example 2: Develop Biomarkers for Kidney Cancer Diagnosis**
  - Statistical Programming (analysis + visualization)
  - Machine Learning
  - Clinical Application
  - Interdependent and Complementary 'Wet'/'Dry' Biology Research

- **Example 3: Generate Pan-Cancer Models of Patient Prognosis**
  - **Statistical Programming (analysis + visualization)**
  - **Machine Learning**
  - **Software Development**
  - **Computational Research**

Take-Home Message

Take-Home Message

- Genomics generates big data to address complex biological problems, e.g., improving human disease prevention, diagnosis, prognosis, and treatment efficacy

Take-Home Message

- Genomics generates big data to address complex biological problems, e.g., improving human disease prevention, diagnosis, prognosis, and treatment efficacy
- Computers and math are necessary to advance biological research (may be referred to as computational biology, bioinformatics, systems biology, data science, statistical biology, biostatistics, etc.)

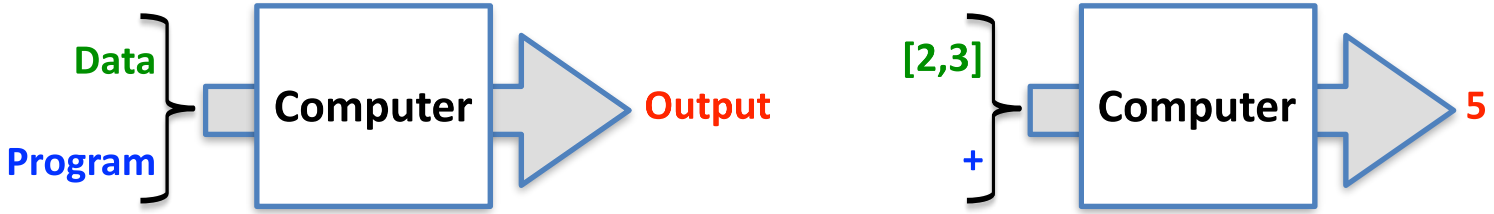
Take-Home Message

- Genomics generates big data to address complex biological problems, e.g., improving human disease prevention, diagnosis, prognosis, and treatment efficacy
- Computers and math are necessary to advance biological research (may be referred to as computational biology, bioinformatics, systems biology, data science, statistical biology, biostatistics, etc.)
- Machine learning is a data analysis method (and subfield of computer science) that automate analytical model building to make data driven **predictions** or discover **patterns** without explicit human intervention (algorithms are implemented in code)

Take-Home Message

- Genomics generates big data to address complex biological problems, e.g., improving human disease prevention, diagnosis, prognosis, and treatment efficacy
- Computers and math are necessary to advance biological research (may be referred to as computational biology, bioinformatics, systems biology, data science, statistical biology, biostatistics, etc.)
- Machine learning is a data analysis method (and subfield of computer science) that automate analytical model building to make data driven **predictions** or discover **patterns** without explicit human intervention (algorithms are implemented in code)

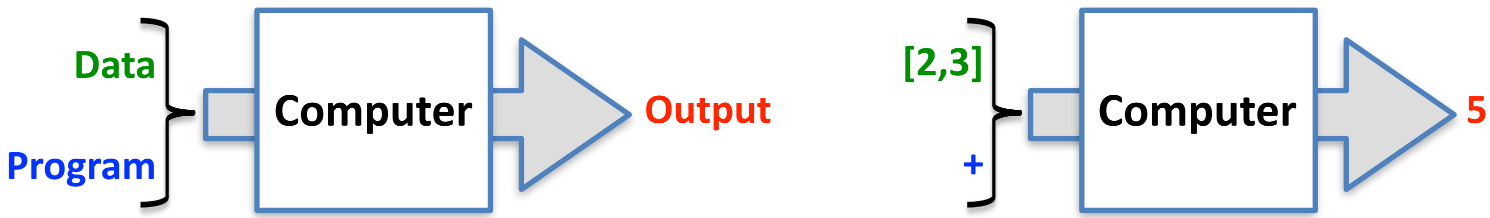
Traditional Programming



Take-Home Message

- Genomics generates big data to address complex biological problems, e.g., improving human disease prevention, diagnosis, prognosis, and treatment efficacy
- Computers and math are necessary to advance biological research (may be referred to as computational biology, bioinformatics, systems biology, data science, statistical biology, biostatistics, etc.)
- Machine learning is a data analysis method (and subfield of computer science) that automate analytical model building to make data driven **predictions** or discover **patterns** without explicit human intervention (algorithms are implemented in code)

Traditional Programming

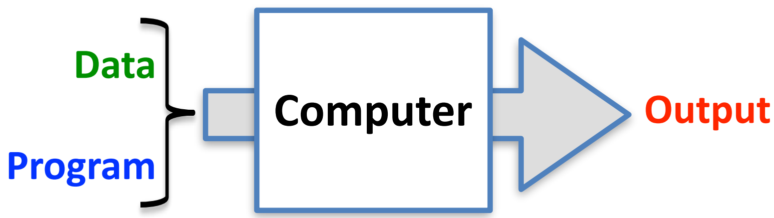


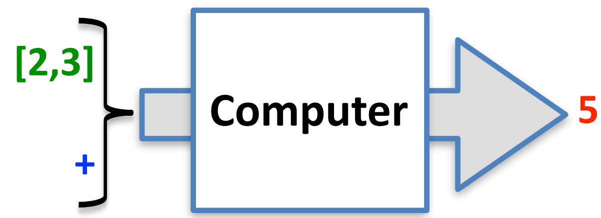
Machine Learning

Take-Home Message

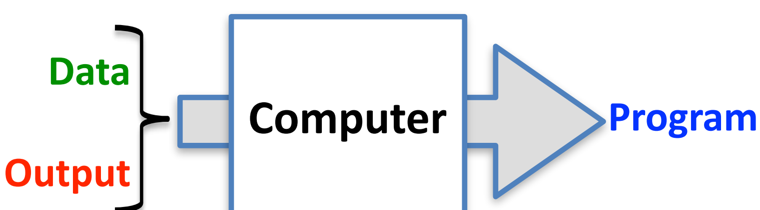
- Genomics generates big data to address complex biological problems, e.g., improving human disease prevention, diagnosis, prognosis, and treatment efficacy
- Computers and math are necessary to advance biological research (may be referred to as computational biology, bioinformatics, systems biology, data science, statistical biology, biostatistics, etc.)
- Machine learning is a data analysis method (and subfield of computer science) that automate analytical model building to make data driven **predictions** or discover **patterns** without explicit human intervention (algorithms are implemented in code)
- Machine learning is useful when we have complex problems with lots of 'big' data

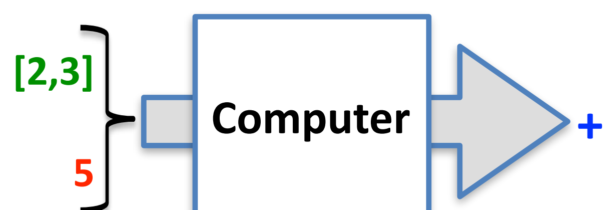
Traditional Programming





Machine Learning

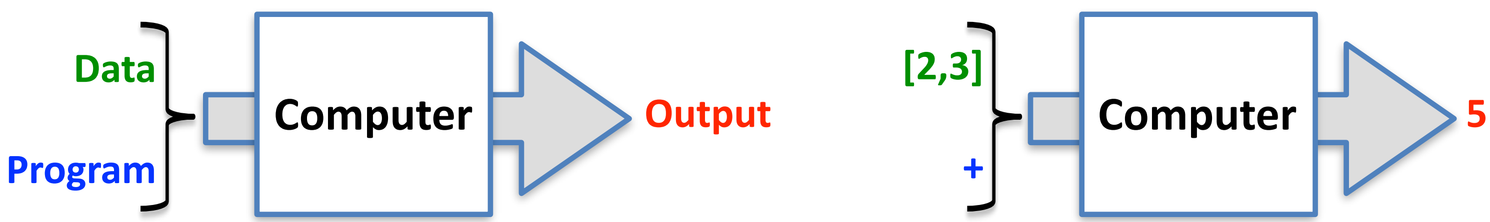




Take-Home Message

- Genomics generates big data to address complex biological problems, e.g., improving human disease prevention, diagnosis, prognosis, and treatment efficacy
- Computers and math are necessary to advance biological research (may be referred to as computational biology, bioinformatics, systems biology, data science, statistical biology, biostatistics, etc.)
- Machine learning is a data analysis method (and subfield of computer science) that automate analytical model building to make data driven **predictions** or discover **patterns** without explicit human intervention (algorithms are implemented in code)
- Machine learning is useful when we have complex problems with lots of 'big' data
- 'Wet lab' and 'dry lab' biology inform one another—>both are biology!

Traditional Programming



Machine Learning

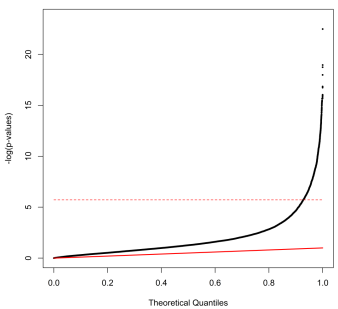
Genomics Requires Team and Individual Expertise in Many Disciplines Because We Are Addressing Complicated Questions

```
MethylationAnalysis = function(Data,Vars,perms=3000,seed=1
# Source the libraries
library(samr)
library(MASS)
BonLine=-log(0.05/nrow(Data$Data),10)
# initilize the variables
pvals=NULL # for samr
x=NULL #our retrun variable
#  Cut down the info file to only the data of interest
Data$info=Data$info[,match(Vars,colnames(Data$info))]
```

Programming

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

Mathematics



Computational Biology

GENOMICS

Engineering

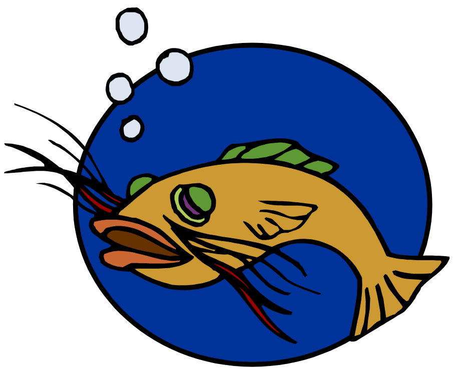
etc....

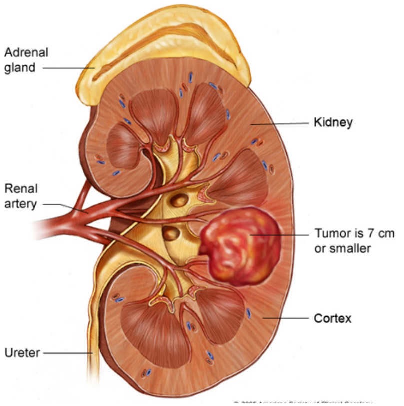
Computational Infrastructure
(IT)

Molecular Biology/
Genetics

Communication

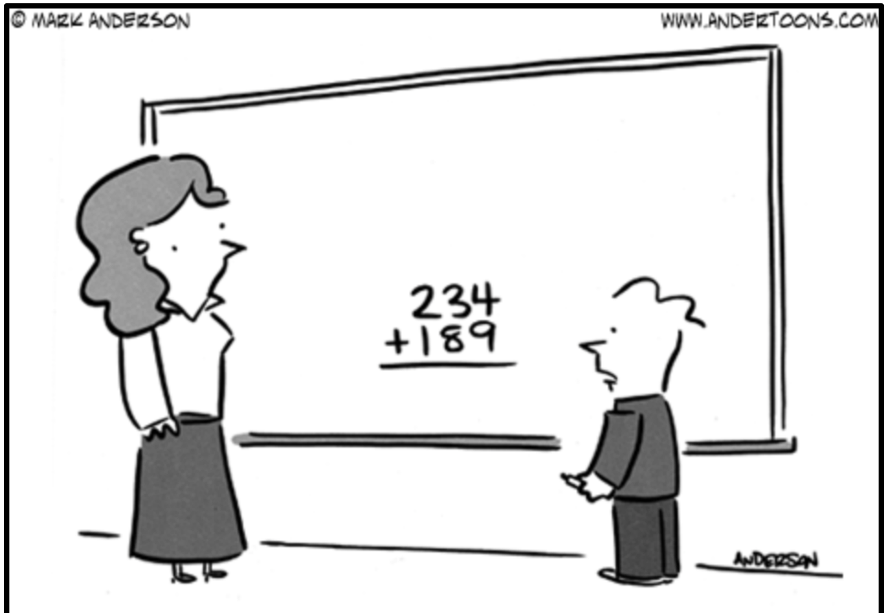
PSA: Any Research Experience is Useful When You're Starting Out





- Catfish virus genes increasing disease susceptibility
- Using bacteria to clean hydrocarbons from ship bilge water
- Hormone effect on kidney mitochondria and obesity
- Reverse engineering electromagnetic flow probes
- Using bacteria to produce ethanol
- Mechanisms of oxidative stress in the brain

Thanks! Slides available at https://www.lasseigne.org/post/2018-06-04-biotraincompbioworkshop2018/



"Does this count as big data?"

Brittany N. Lasseigne, PhD

@bnlasse **blasseigne@hudsonalpha.org**