

CS 598MEB

Computational Cancer Genomics

Lecture 1

Mohammed El-Kebir

January 26, 2021



Course Staff

Instructor:

- Mohammed El-Kebir (melkebir)
- Office hours: Tuesdays, 3:15-4:15pm



Developing combinatorial algorithms for studying all stages of cancer progression.

Course Information

Course website:

- www.el-kebir.net/teaching/cs598MEB

Piazza: (please sign up)

- <https://piazza.com/illinois/spring2020/cs598meb>

Description:

- This course focuses on **recent algorithmic methods in cancer genomics**, including somatic variant calling, phylogeny inference and identification of driver mutations. Students will study the underlying principles of these methods and the application of these methods to cancer genomics data.

Course Objectives

Learn:

- Learn underlying ideas of common algorithms in cancer genomics.
- Learn to translate a biological problem into a computational problem.
- Learn to read and critique scientific papers.
- Learn to propose and conduct independent research.
- Learn to present key ideas of a paper to other people.
- Learn to ask critical questions.

Not learn:

- Will not learn to run popular cancer genomics packages.
- Will not learn how to program.

Grading

- Class participation (20%)
 - Peer reviews
 - Asking questions
- Paper presentation (30%)
- Course project (50%)
 - Proposal
 - Report/paper
 - Presentation

Tentative Course Schedule

Introductory lectures (Jan 26 to Mar 11)

- Molecular biology and cancer biology
- Fundamental algorithms in computational biology
- Algorithms in computational genomics

Paper presentations (Mar 16 to Apr 13)

- Student presentation of research/survey paper

Course projects (Apr 15 to May 6)

- Proposal presentation
- Final presentation + report

Paper Presentation

- Each student will present a paper picked by the student. The goal of the presentation is to facilitate a discussion, focusing on:
 - Presenting the biological problem and corresponding computational problem
 - How did the authors solve the problem?
 - Did they manage to answer the original biological question?
 - How can we improve the results? What are future directions?
- The remaining students are required to write a short peer review
 - Summary
 - Major and minor comments
 - Outlook/future directions

Course Project

- 1-2 students per project
- First write a proposal, which will receive feedback from instructor and fellow students
- Then, conduct research and write a paper
- Pick venue (conference/journal) and use LaTeX style for your paper
- Students will anonymously peer review submitted papers using EasyChair (if time permits)

Lecture Outline

- Primer on Molecular Biology
- Primer on Computational Biology
- Primer on Cancer Biology
- Tumor Phylogeny Inference

Reading

- “Biology for Computer Scientists” by Lawrence Hunter
(http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)

Primer on Molecular Biology

Molecular Biology is the field of **biology** that studies the composition, structure and interactions of cellular **molecules** – such as nucleic acids and proteins – that carry out the **biological** processes essential for the cell's functions and maintenance.

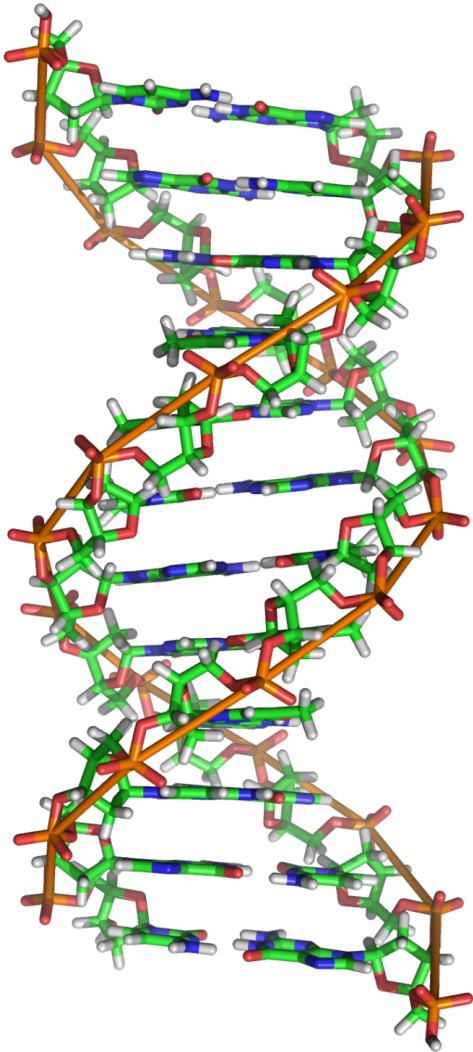
<https://www.nature.com/subjects/molecular-biology>

Cellular molecules:

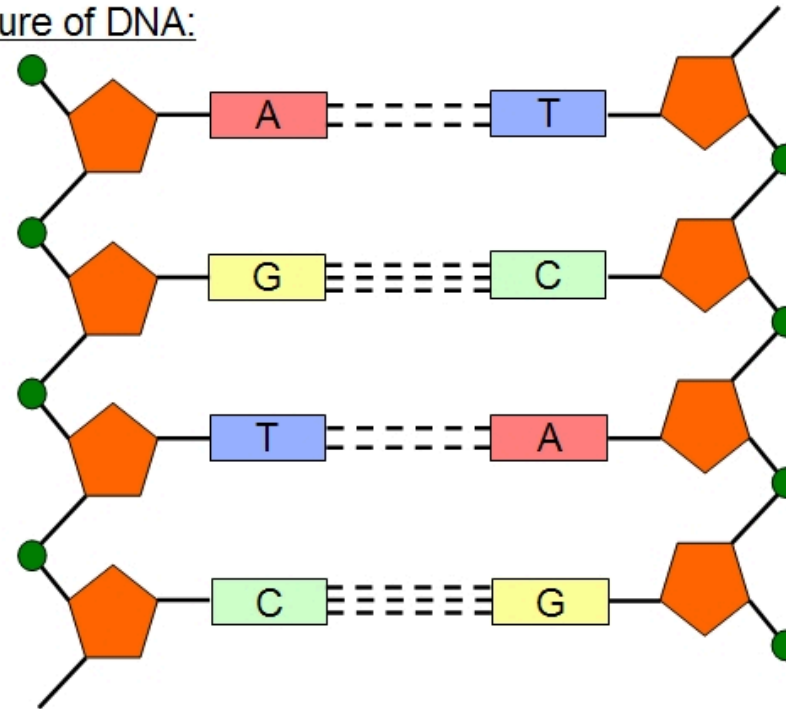
1. DNA
2. RNA
3. Protein

DNA

Each strand composed of sequence of covalently bonded **nucleotides (bases)**.



Structure of DNA:



$A \leftrightarrow T$, $C \leftrightarrow G$ Watson-Crick base-pairing

Four nucleotides:

A (adenine)

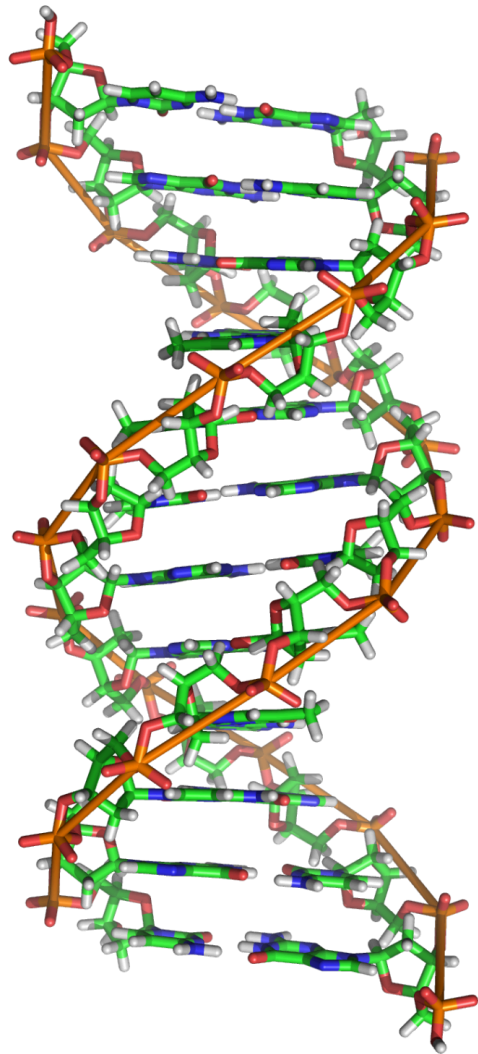
C (cytosine)

T (thymine)

G (guanine)

DNA

Each strand composed of sequence of covalently bonded **nucleotides (bases)**.



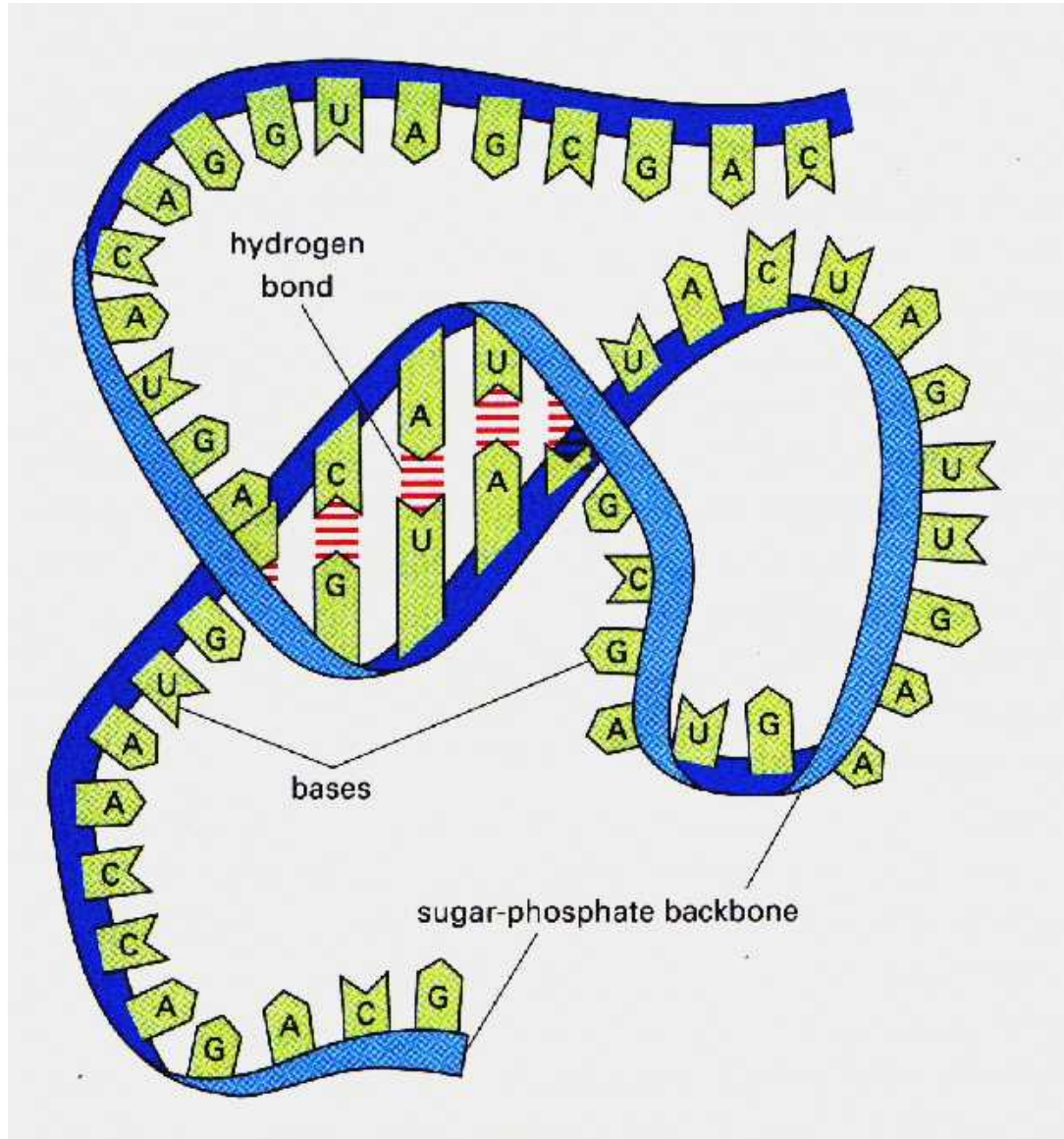
5' ...ACGTGACTGAGGACCGTG... 3'
... ||||| ||||| ||||| ||||| ...
3' ...TGCCTGACTCCTGGCAC... 5'

Pair of strings
from 4 character
alphabet

5' ...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
TGATTTTAAAAAAATATT... 3'

Single string
from 4 character
alphabet

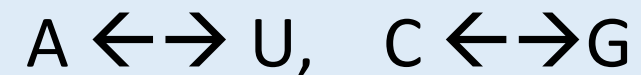
RNA



- **Single-stranded**

- A (adenine)
- C (cytosine)
- U (uracil)
- G (guanine)

- Can fold into **structures** due to base complementarity.



- Comes in many flavors:

mRNA, rRNA, tRNA, tmRNA, snRNA,
snoRNA, scaRNA, aRNA, asRNA, piwiRNA, etc.

Protein

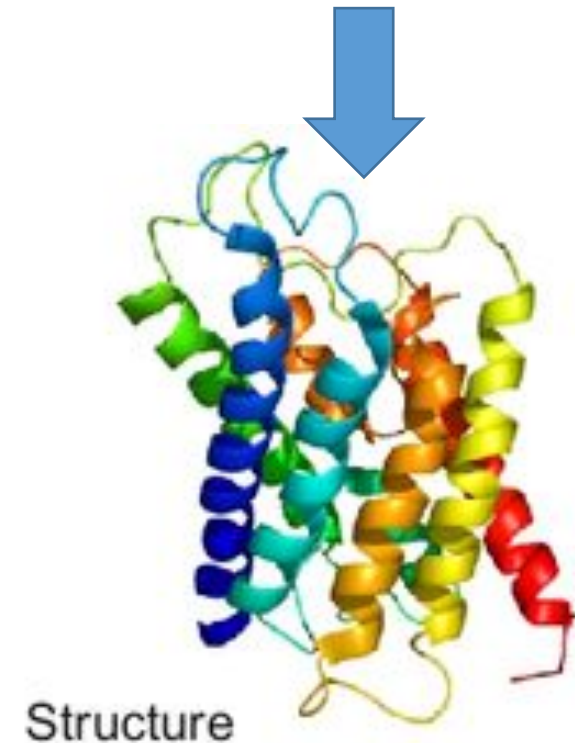
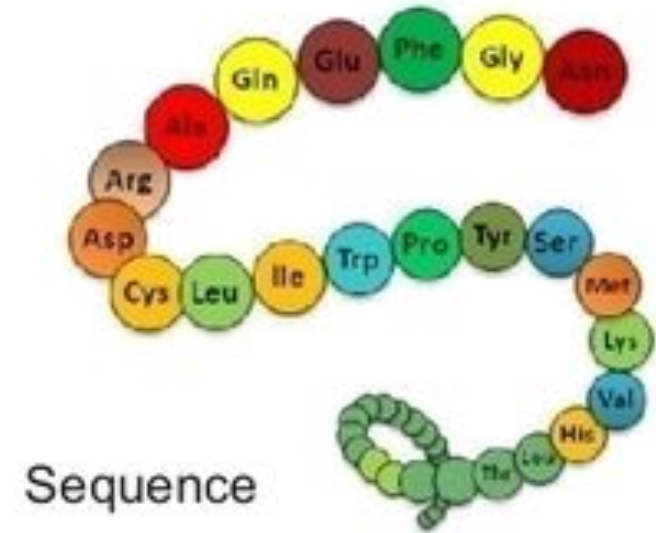
- String of amino acids: 20 letter alphabet

...DTIGDWNSPSFFGIQLVSSVHT
TLWYRENAFPVLGGFSWLSWFNW
HNMGYYPVYHIGYPMIRCGTHL
VPMQFAFQSIARSFALVHWNAPM
VLKINPHERQDPVFWPCLYYSVD
IRSMHIGYPMIRCYQA...

Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Protein

- String of amino acids: 20 letter alphabet
- Folds into 3D structures to perform various functions in cells



Primer on Molecular Biology

Three fundamental molecules:

1. DNA

Information storage.

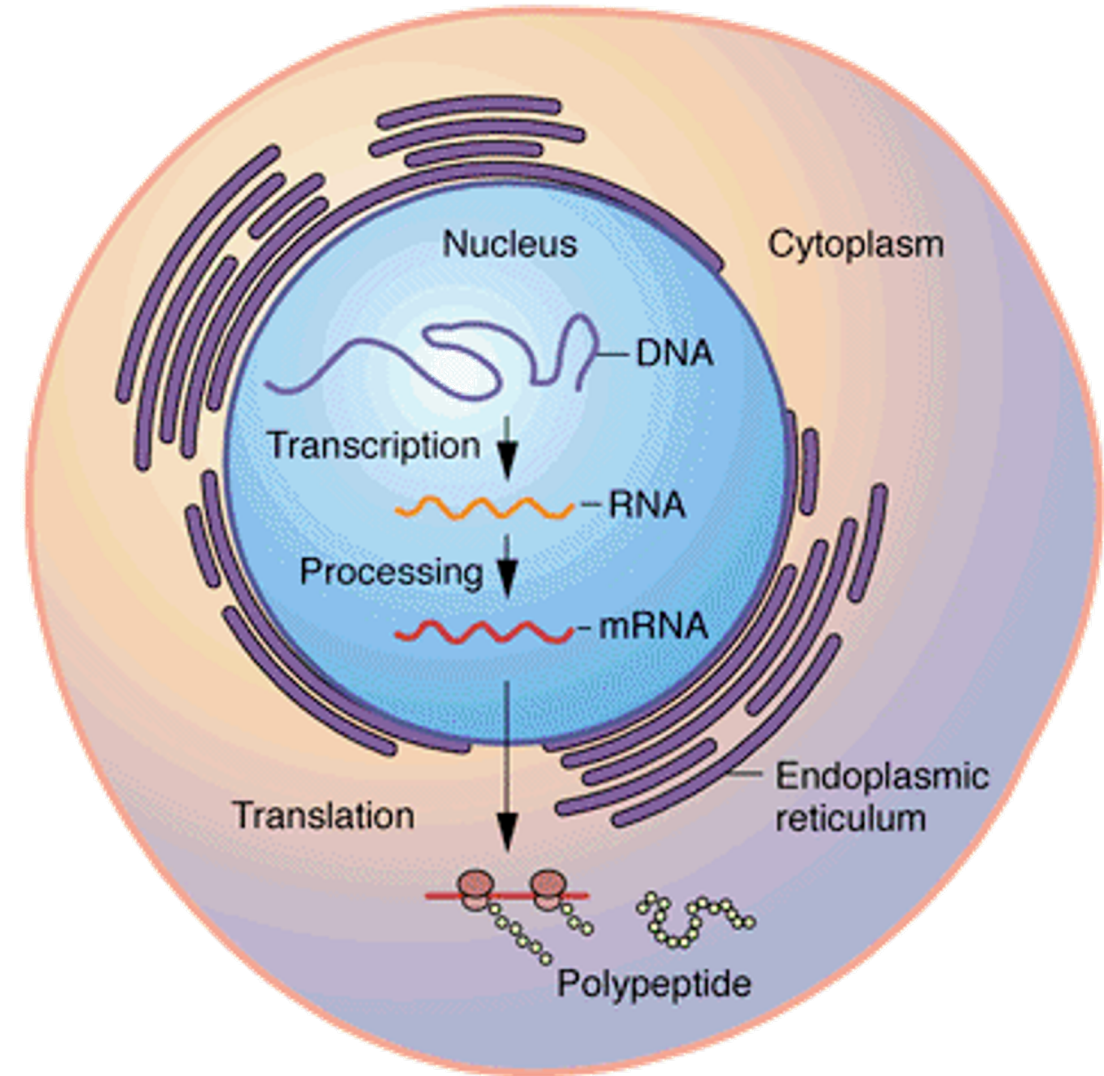
2. RNA

Old view: Mostly a “messenger”.

New view: Performs many important functions.

3. Protein

Perform most cellular functions
(biochemistry, signaling, control, etc.)

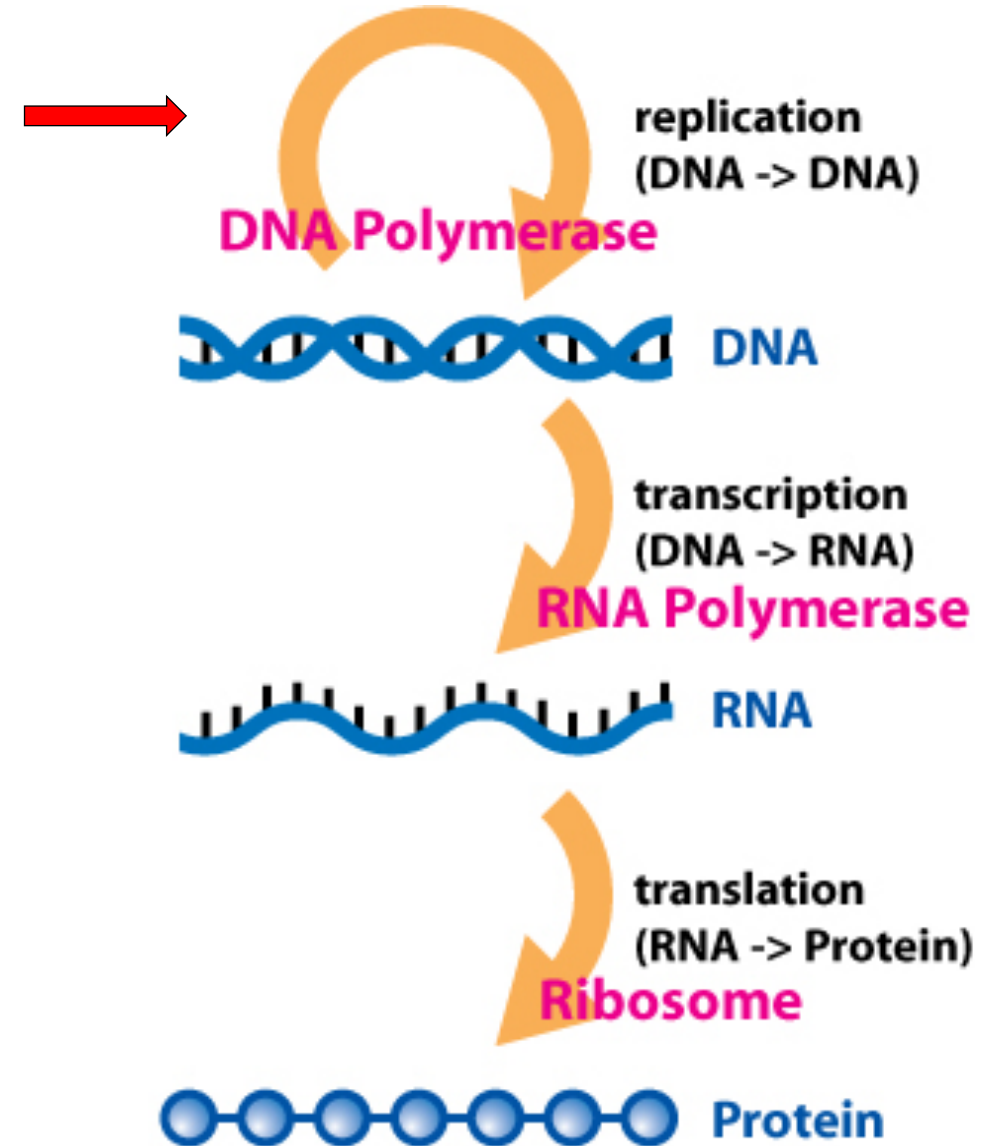


Central Dogma of Molecular Biology

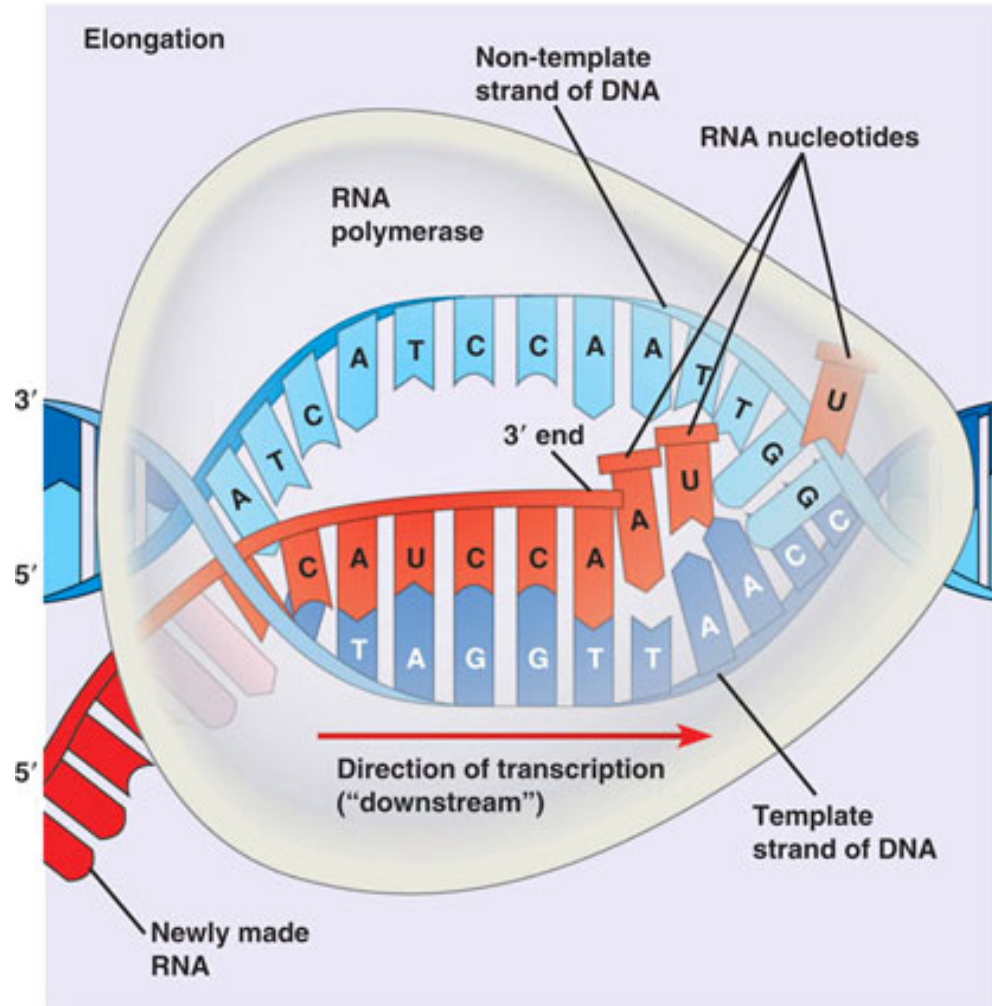
DNA → RNA → Protein:
The process by which cells
“read” the genome

First proposed by Francis Crick in 1956.

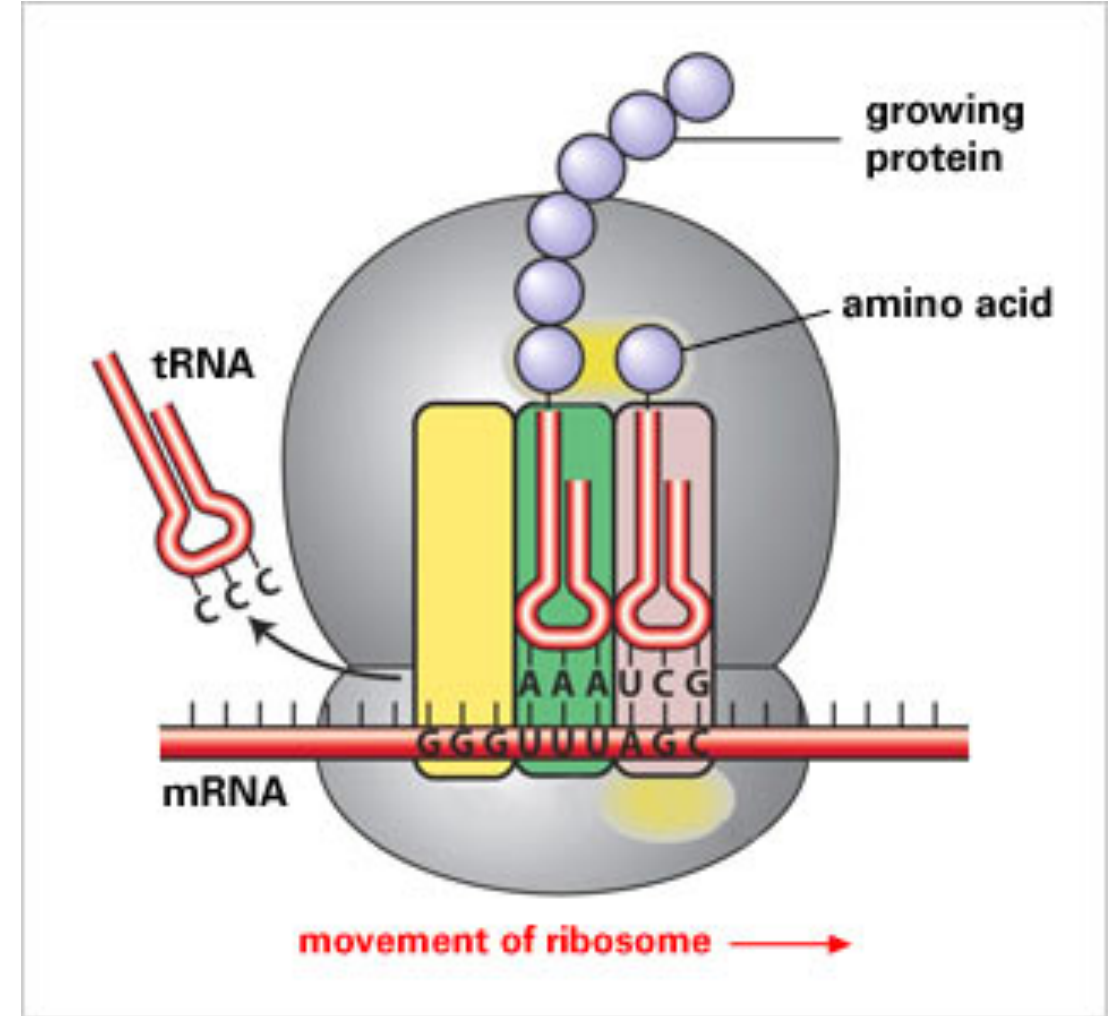
Start here



Transcription and Translation

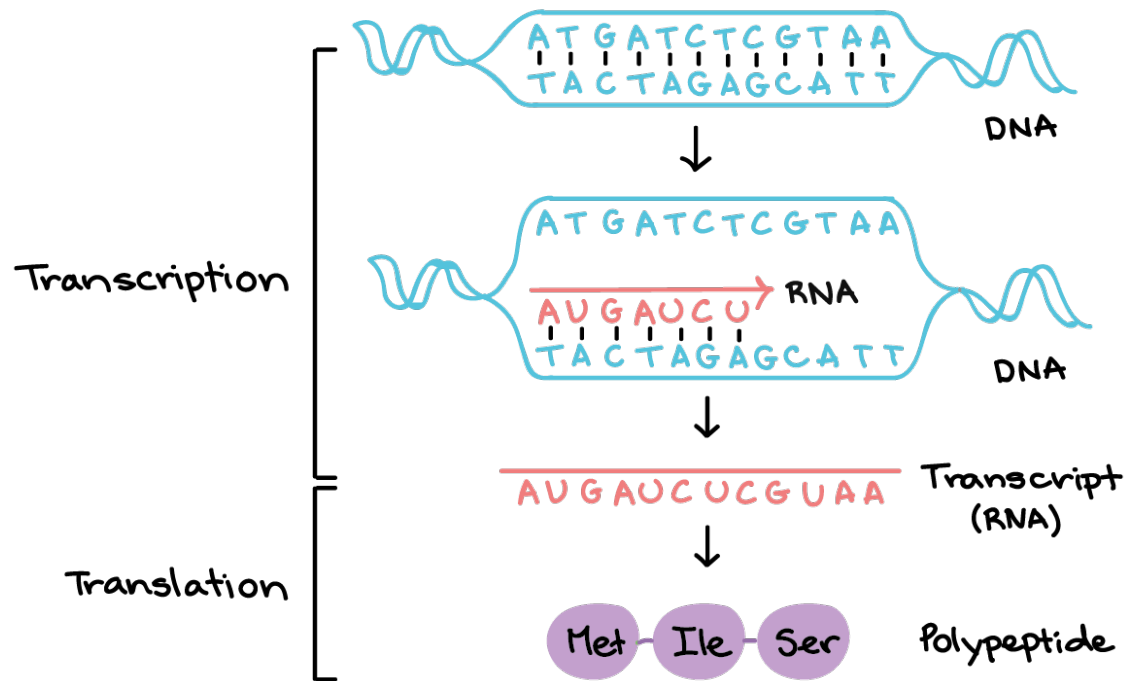


<http://dna-rna.net/wp-content/uploads/2011/08/rna-transcription2.jpg>



http://www.frontiers-in-genetics.org/en/pictures/translation_1.jpg

Transcription and Translation



		Second base				
		U	C	A	G	
First base	U	UUU } Phenyl-alanine F UUC } UUA } Leucine L UUG }	UCU } Serine S UCC } UCA } UCG }	UAU } Tyrosine Y UAC } UAA } Stop codon UAG } Stop codon	UGU } Cysteine C UGC } UGA } Stop codon UGG } Tryptophan W	U C A G
	C	CUU } Leucine L CUC } CUA } CUG }	CCU } Proline P CCC } CCA } CCG }	CAU } Histidine H CAC } CAA } Glutamine Q CAG }	CGU } Arginine R CGC } CGA } CGG }	U C A G
	A	AUU } Isoleucine I AUC } AUA } AUG } Methionine start codon M	ACU } Threonine T ACC } ACA } ACG }	AAU } Asparagine N AAC } AAA } Lysine K AAG }	AGU } Serine S AGC } AGA } Arginine R AGG }	U C A G
	G	GUU } Valine V GUC } GUA } GUG }	GCU } Alanine A GCC } GCA } GCG }	GAU } Aspartic acid D GAC } GAA } Glutamic acid E GAG }	GGU } Glycine G GGC } GGA } GGG }	U C A G

<https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/overview-of-transcription>

<http://bioinfo.bisr.res.in/project/crat/pictures/codon.jpg>

Lecture Outline

- Primer on Molecular Biology
- Primer on Computational Biology
- Primer on Cancer Biology
- Tumor Phylogeny Inference

Reading

- “Biology for Computer Scientists” by Lawrence Hunter
(http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)

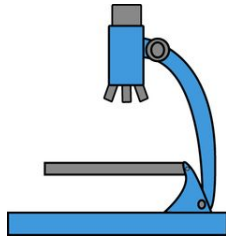
What is Computational Biology/Bioinformatics?

Computational biology and **bioinformatics** is an interdisciplinary field that develops and applies **computational methods** to analyze large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or **discover new biology**.

<https://www.nature.com/subjects/computational-biology-and-bioinformatics>

Technology and Bioinformatics are Transforming Biology

Until late 20th Century



Hypothesis Generation
and Validation

21th Century and Beyond



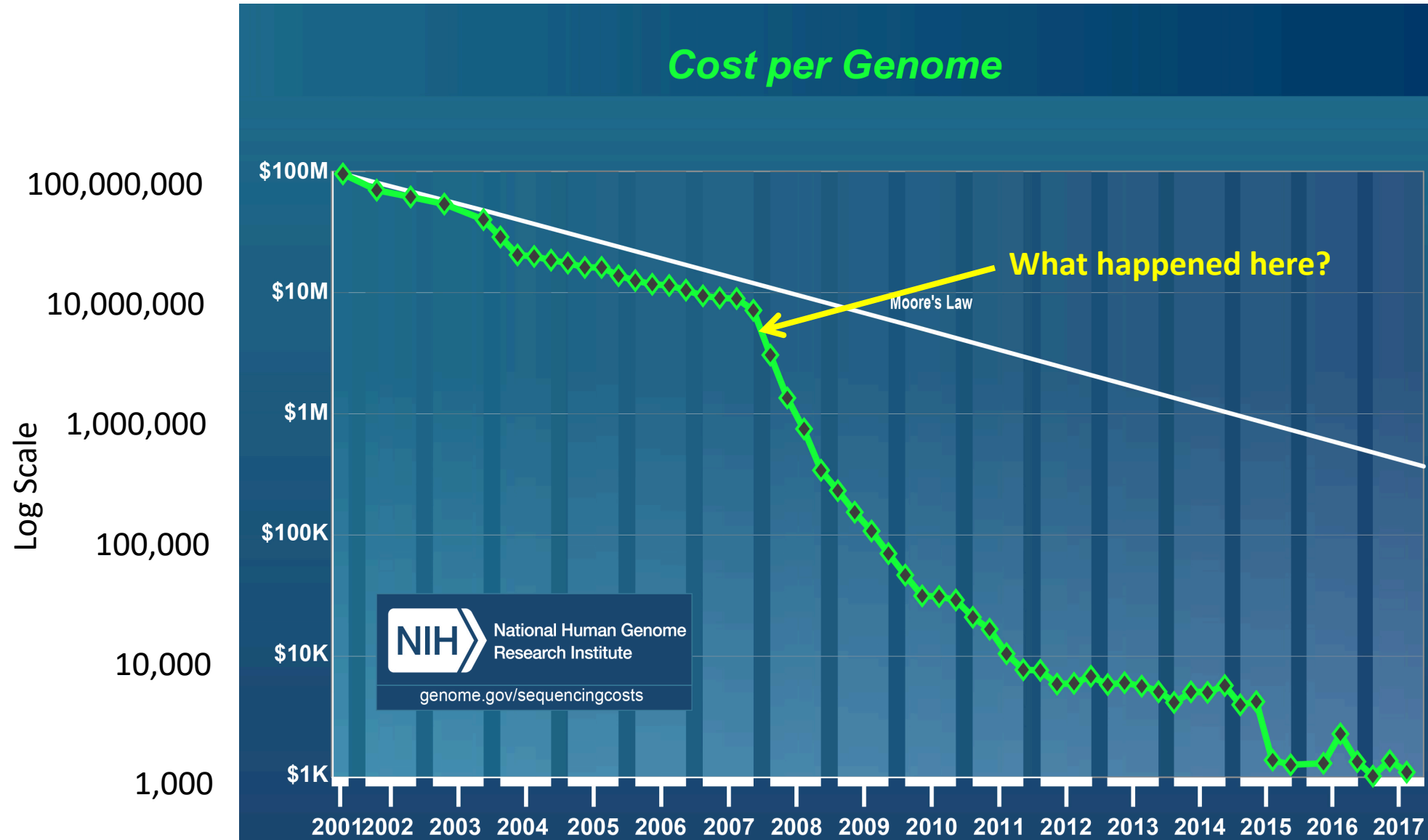
Algorithms



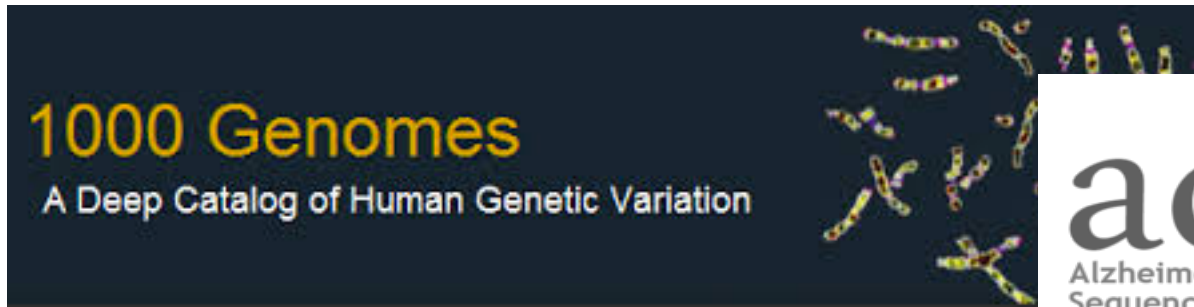
Hypothesis Generation
and Validation

High throughput technologies

A Deluge of Data



A Deluge of Data



1000 Plant Genomes



International Cancer Genome Consortium

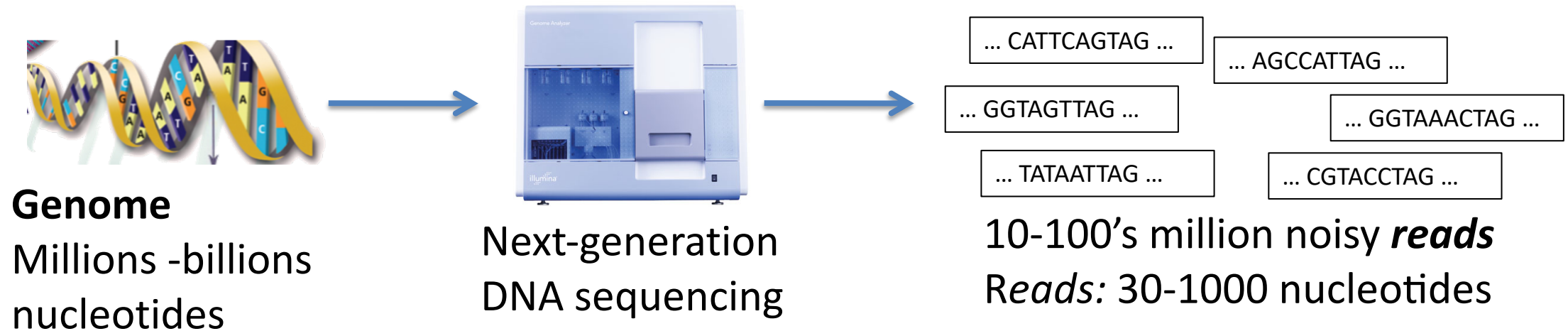


NIH HUMAN MICROBIOME PROJECT



Question: What does it mean that we can sequence a genome?

No technology exists that can sequence a complete (human) genome from end to end!



Making sense of this data absolutely requires the use and development of **algorithms!**

Why Study Computational Biology?

Interdisciplinary

Biology

Computer Science

Mathematics

Statistics

= FUN!



Why choose just 1?

Best Jobs

1. Actuary
2. Audiologist
3. Mathematician
4. Statistician
5. Biomedical Engineer
6. Data Scientist
7. Dental Hygienist
8. Software Engineer
9. Occupational Therapist
10. Computer Systems Analyst

Worst Jobs

200. Newspaper reporter
199. Lumberjack
198. Enlisted Military Personnel
197. Cook
196. Broadcaster
195. Photojournalist
194. Corrections Officer
193. Taxi Driver
192. Firefighter
191. Mail Carrier



Donald Knuth

Professor emeritus of Computer Science at Stanford University

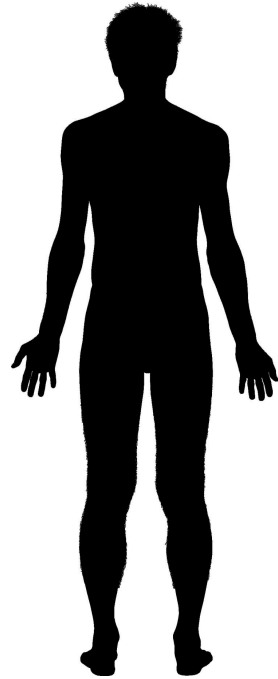
Turing Award winner

“father of the analysis of algorithms.”

*“I can’t be as confident about computer science as I can about biology. **Biology easily has 500 years of exciting problems to work on.** It’s at that level.”*

Computational Biology: Sequence Alignment

Question: How do we compare two genes/genomes?

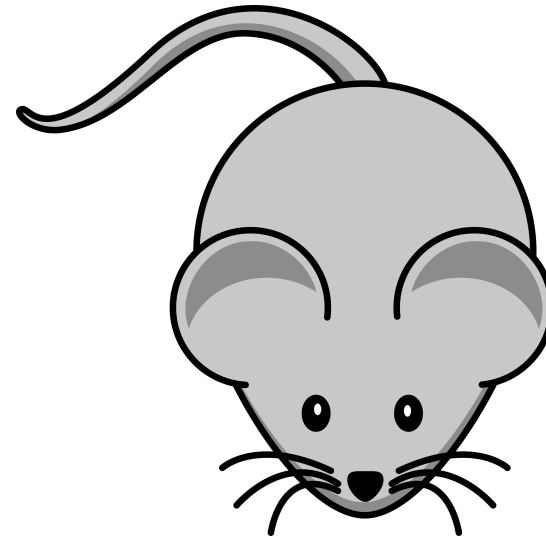


Human Genome:

...ACTCGACTGAGAGGATTTTCGAGCATGA...

$\approx 3.2 \times 10^9$ bp

vs.

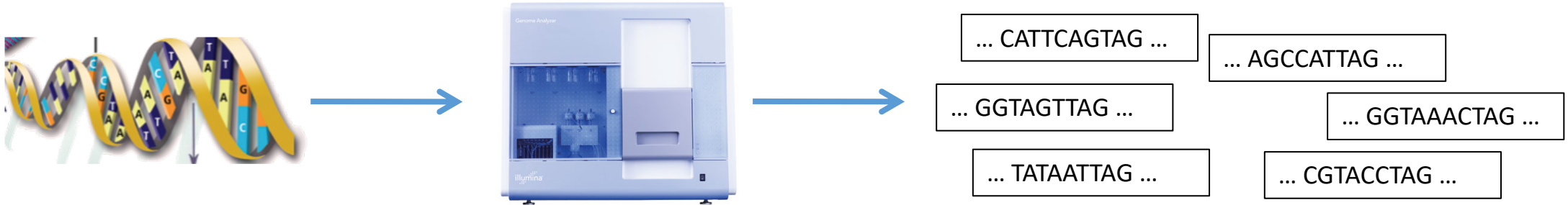


Mouse Genome:

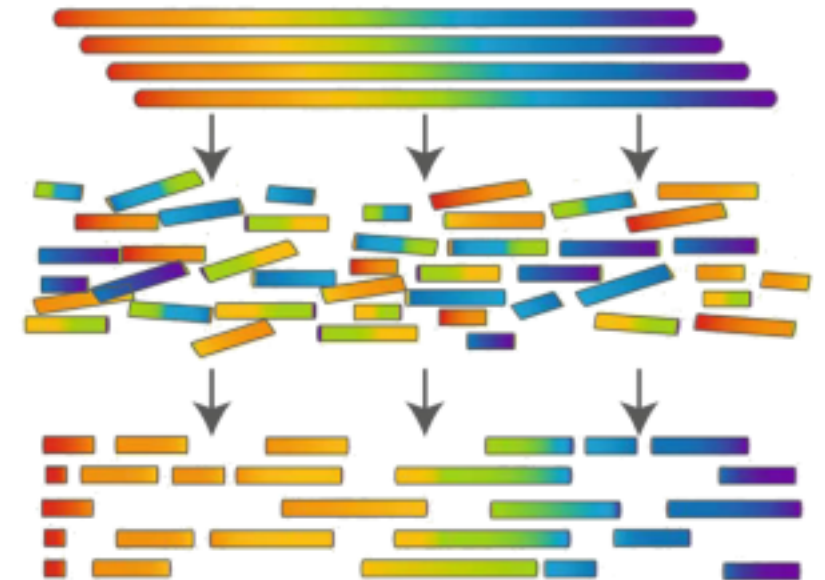
...ACTCAACTGAGATTCGAGCTTCAATGA...

$\approx 2.8 \times 10^9$ bp

Computational Biology: Genome Assembly

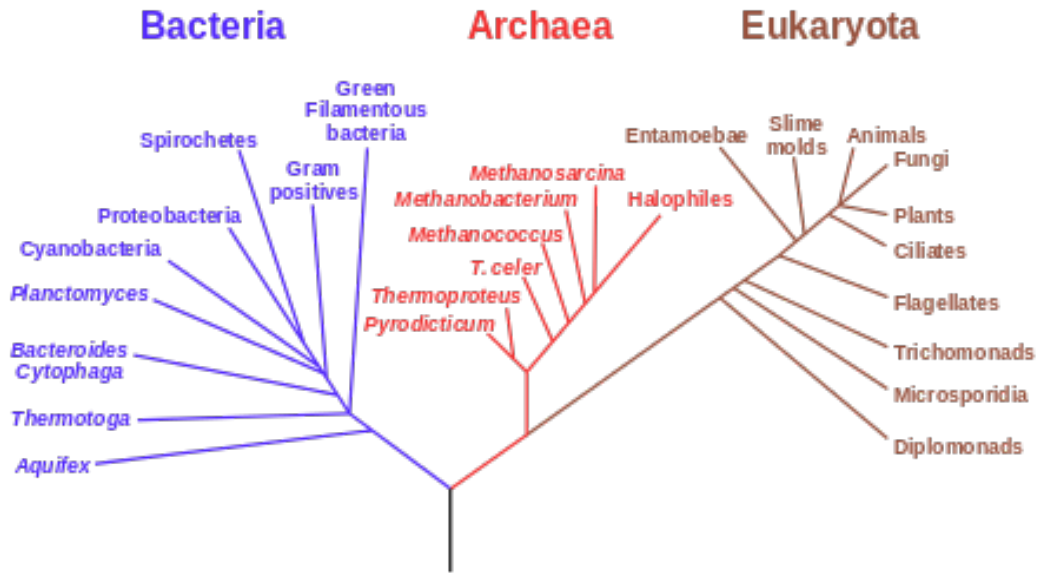


Question: How do we put all the pieces back together?



Computational Biology: Phylogenetics

Phylogenetic Tree of Life

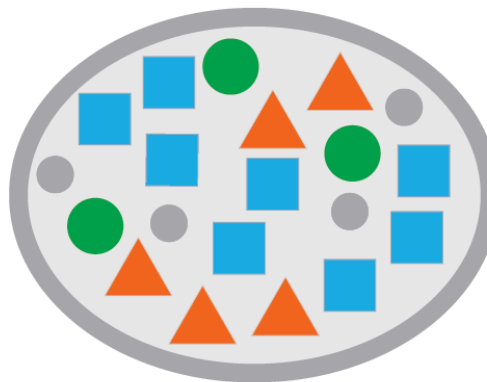


https://en.wikipedia.org/wiki/Phylogenetic_tree

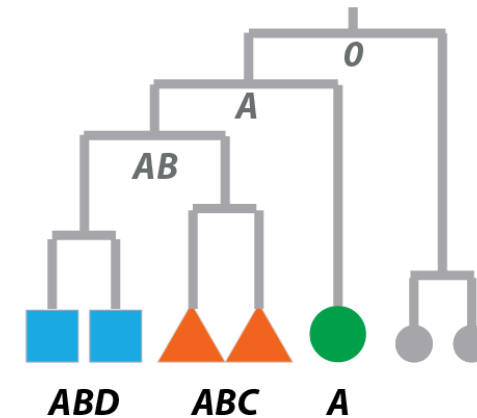
Question: Can we reconstruct the evolutionary history of different species?

Question: Can we recover how a tumor has evolved overtime?

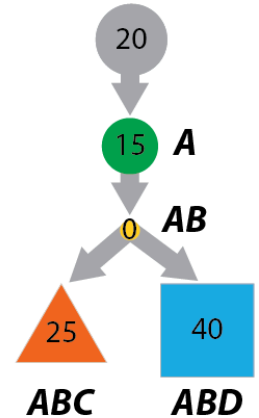
Poly-clonal tumor at sampling



Classical phylogenetic tree

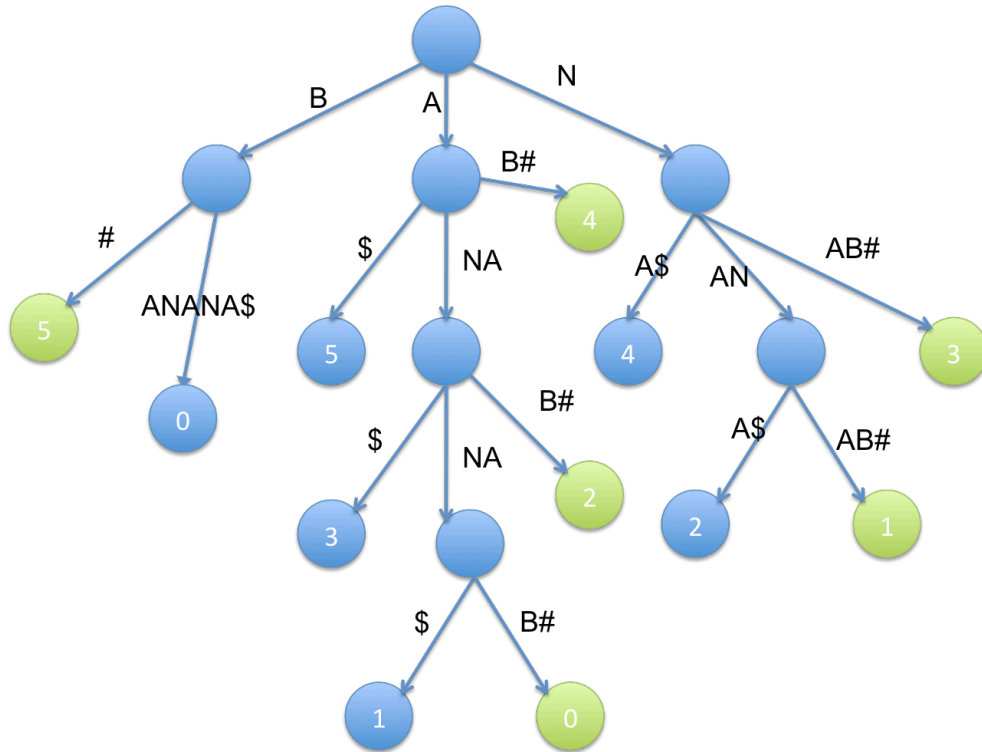


Clonal evolution tree

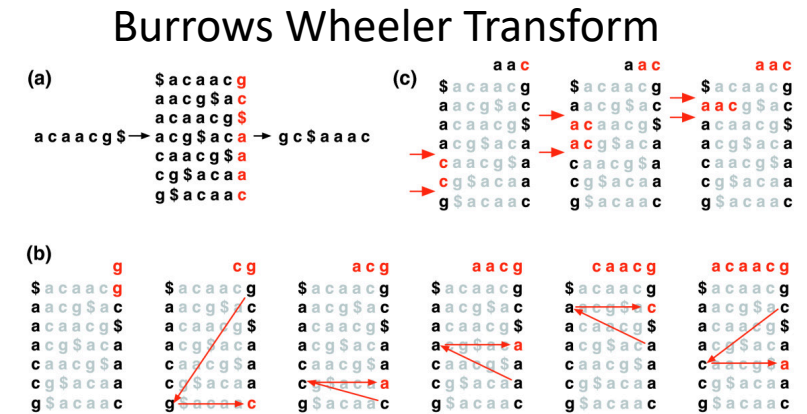


Computational Biology: Pattern Matching

Question: How do we start to make sense of all these sequences?

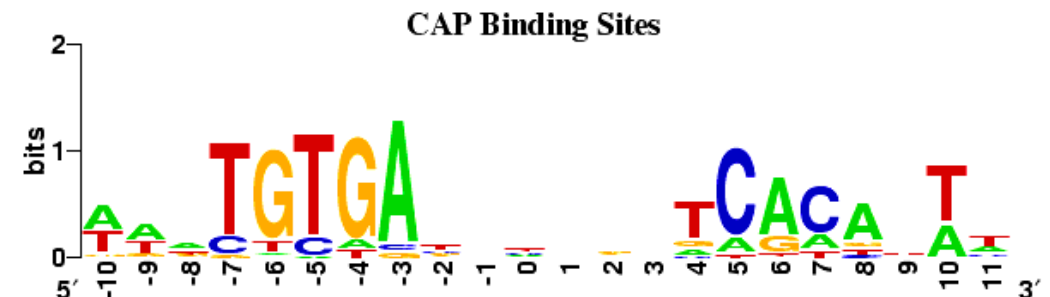


Suffix Trees



<http://www.genomebiology.com/2009/10/3/R25/figure/F1?highres=y>

Motif Finding



Computational Biology is Computer Science

1. Sequence alignment
'How do we compare two genes/genomes?'
Dynamic programming: edit distance
2. Genome assembly
'How do we put all the pieces back together?'
Graphs: de Bruijn graph, Eulerian and Hamiltonian paths
3. Phylogenetics
'What is the evolutionary history of different sequences?'
Trees and distances: distance matrices, neighbor joining, hierarchical clustering, Sankoff/Fitch algorithms, perfect phylogeny and compatibility
4. Pattern matching
'How do we start to make sense out of all these sequences?'
Suffix trees/arrays. Burrows-Wheeler transform, Hidden Markov Models (HMMs)

Pet Peeve: Problem \neq Algorithm

Problem Π with instance X and solution set $\Pi(X)$:

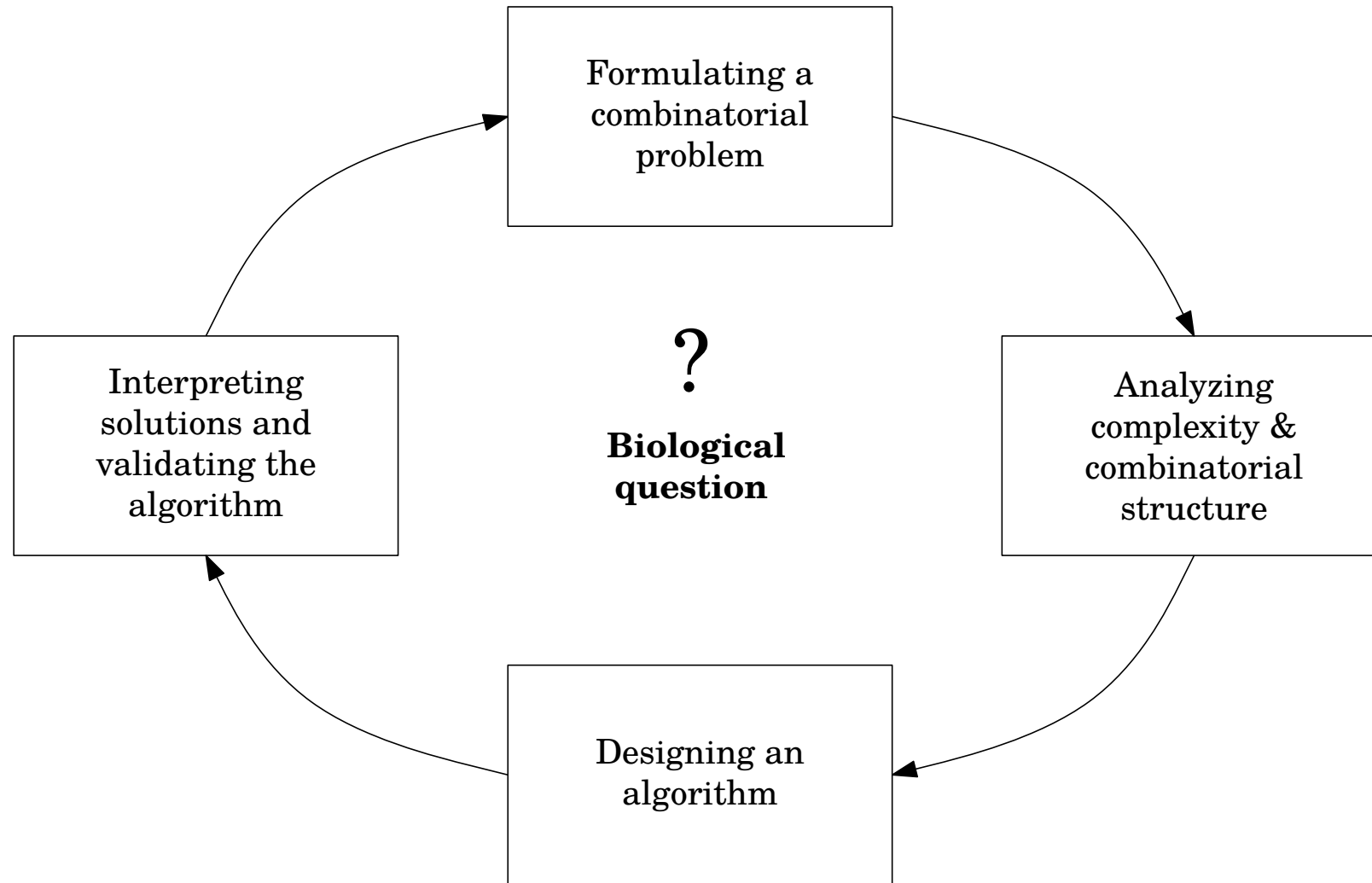
- Decision problem:
 - Is $\Pi(X) = \emptyset$?
- Optimization problem:
 - Find $y^* \in \Pi(X)$ s.t. $f(y^*)$ is optimum.
- Counting problem:
 - Compute $|\Pi(X)|$.
- Sampling problem:
 - Sample uniformly from $\Pi(X)$.
- Enumeration problem:
 - Enumerate all solutions in $\Pi(X)$

Algorithms:

Set of instructions for solving problem.

- Exact
- Heuristic

Key Challenge in Computational Biology



Translating a biological problem into a computational biology

Lecture Outline

- Primer on Molecular Biology
- Primer on Computational Biology
- Primer on Cancer Biology
- Tumor Phylogeny Inference

Reading

- “Biology for Computer Scientists” by Lawrence Hunter
(http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)

Cancer Statistics: Incidence and Mortality

The Burden of Cancer in the United States

- In 2018, an estimated 1,735,350 new cases of cancer will be diagnosed in the United States and 609,640 people will die from the disease.
- The number of new cases of cancer (cancer [incidence](#)) is 439.2 per 100,000 men and women per year (based on 2011–2015 cases).
- The number of cancer deaths (cancer [mortality](#)) is 163.5 per 100,000 men and women per year (based on 2011–2015 deaths).
- Approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes (based on 2013–2015 data).

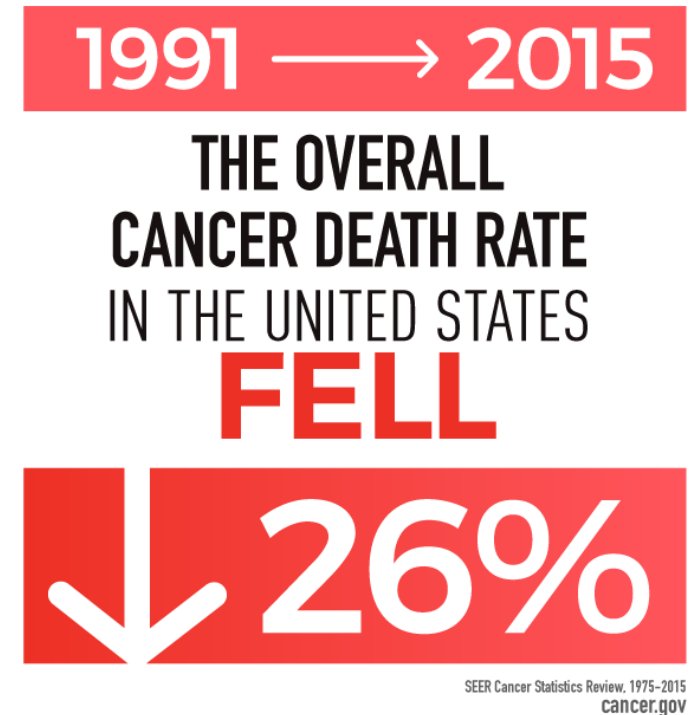
Source: Surveillance, Epidemiology, and End Results (SEER) Program

Cancer Statistics: Incidence and Mortality

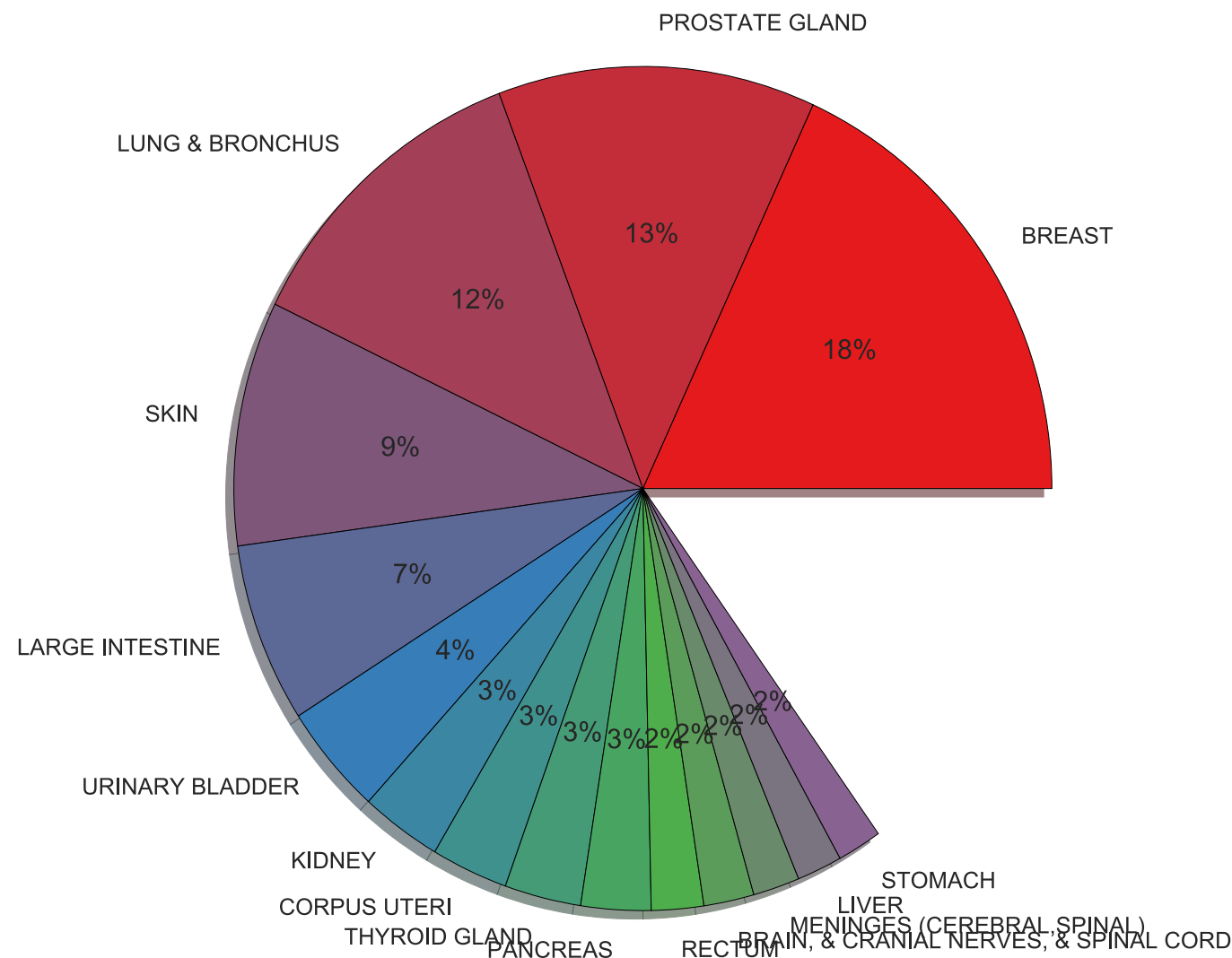
The Burden of Cancer in the United States

- In 2018, an estimated 1,735,350 new cases of cancer will be diagnosed in the United States and 609,640 people will die from the disease.
- The number of new cases of cancer (cancer [incidence](#)) is 439.2 per 100,000 men and women per year (based on 2011–2015 cases).
- The number of cancer deaths (cancer [mortality](#)) is 163.5 per 100,000 men and women per year (based on 2011–2015 deaths).
- Approximately 38.4% of men and women will be diagnosed with cancer at some point during their lifetimes (based on 2013–2015 data).

Source: Surveillance, Epidemiology, and End Results (SEER) Program



Cancer Statistics: Primary Tumor Location



90% of cancer patients die of **metastasis** [Gupta, G. P. & Massagué, Cell, 2006]

Hallmarks of Cancer

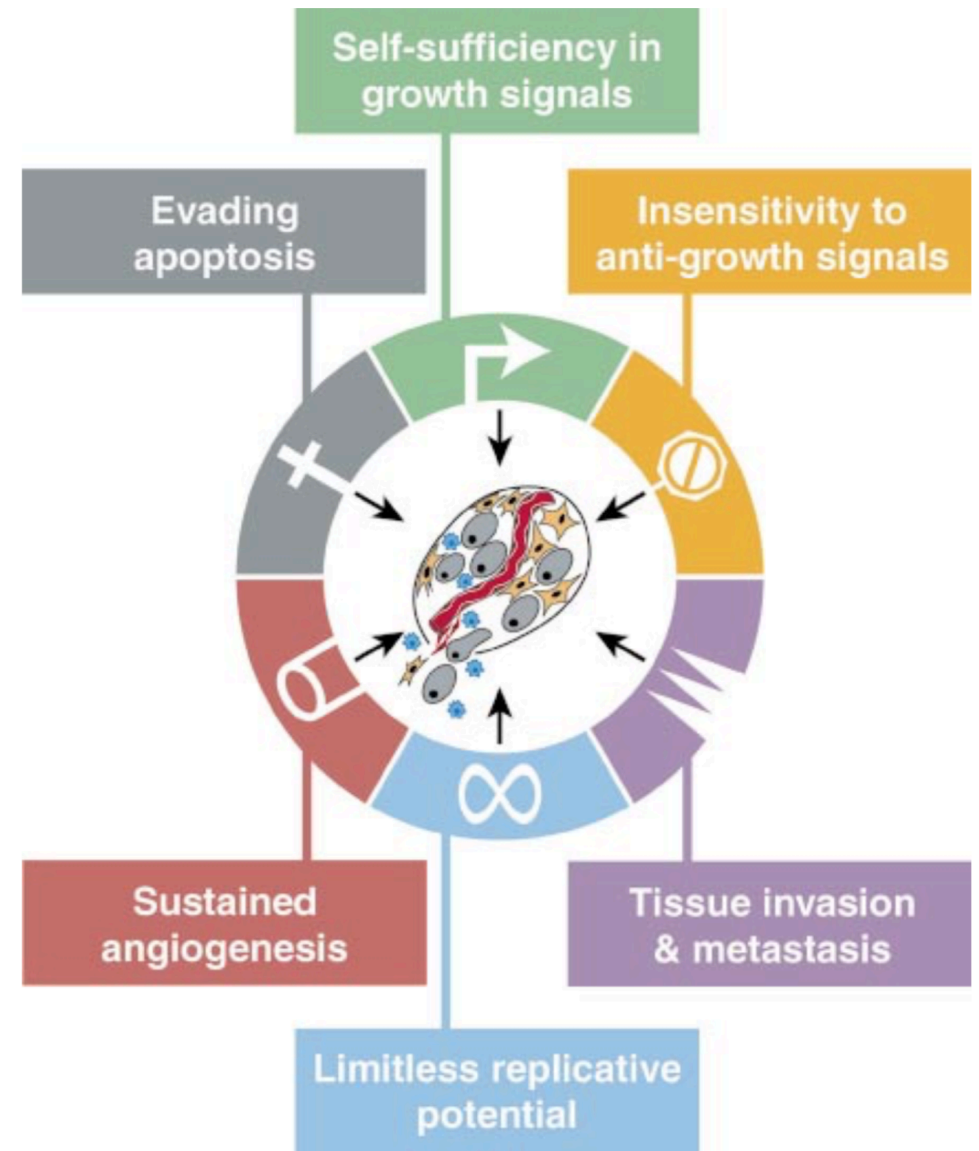


Figure 1. Acquired Capabilities of Cancer

We suggest that most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies.

Hallmarks of Cancer



Inter-tumor heterogeneity:
Every tumor is different!

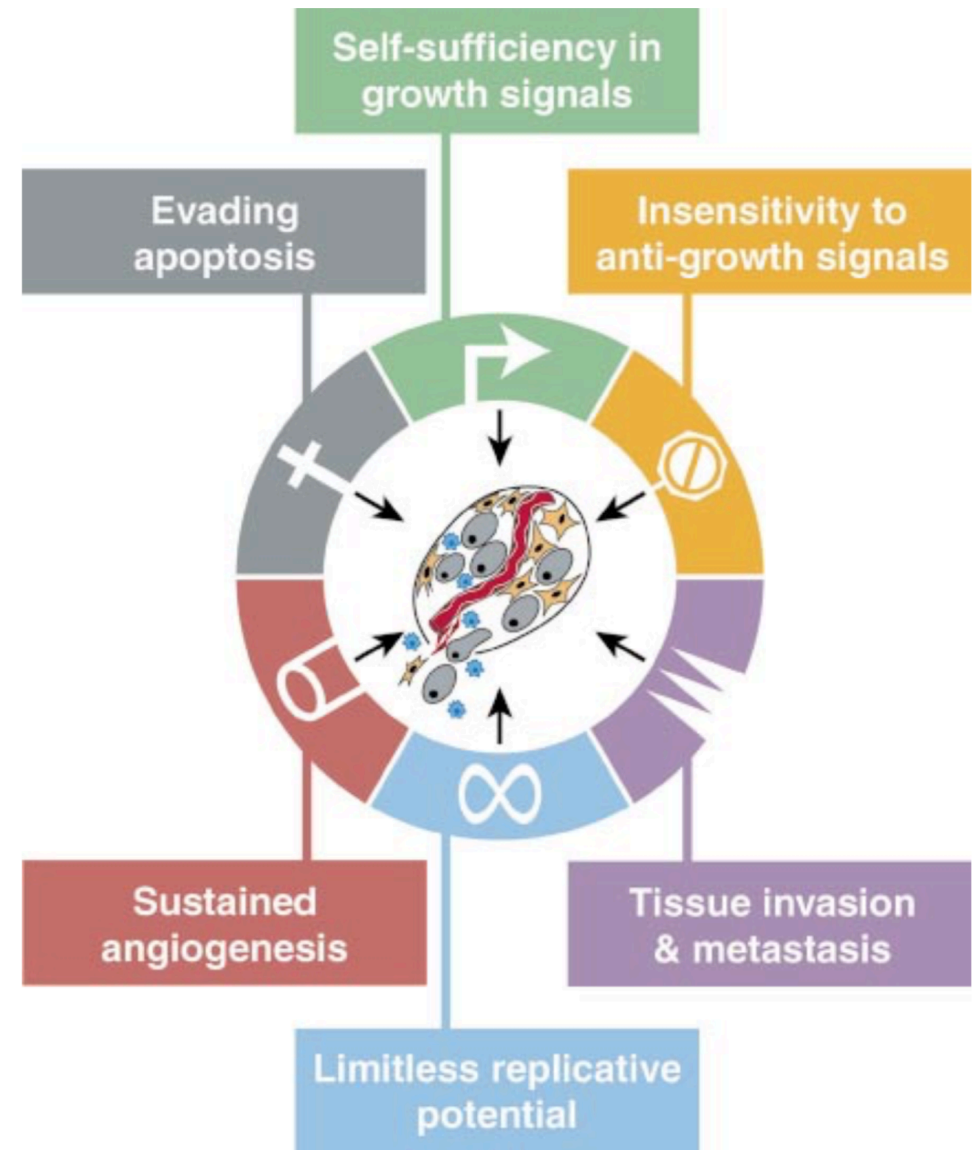


Figure 1. Acquired Capabilities of Cancer

We suggest that most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies.

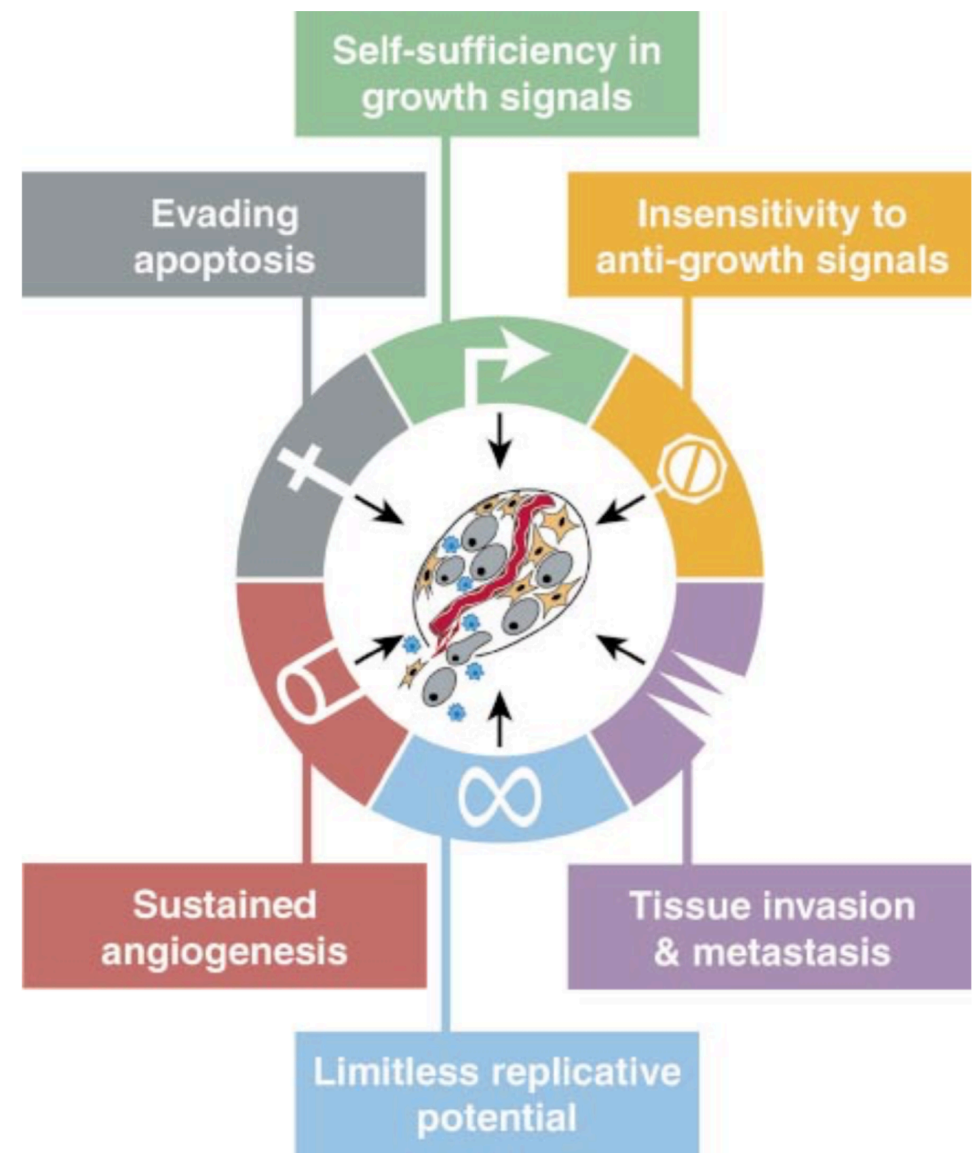
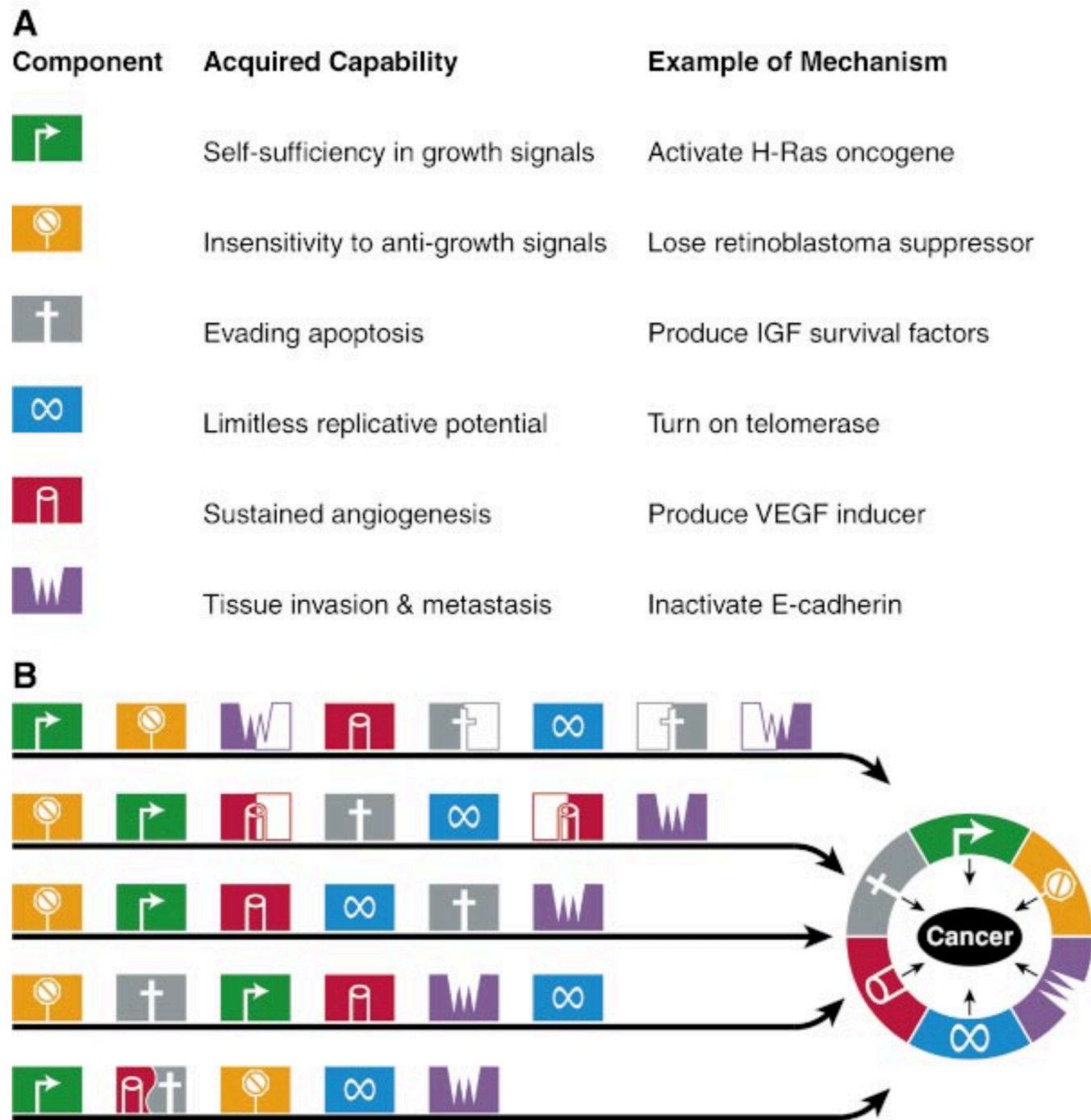
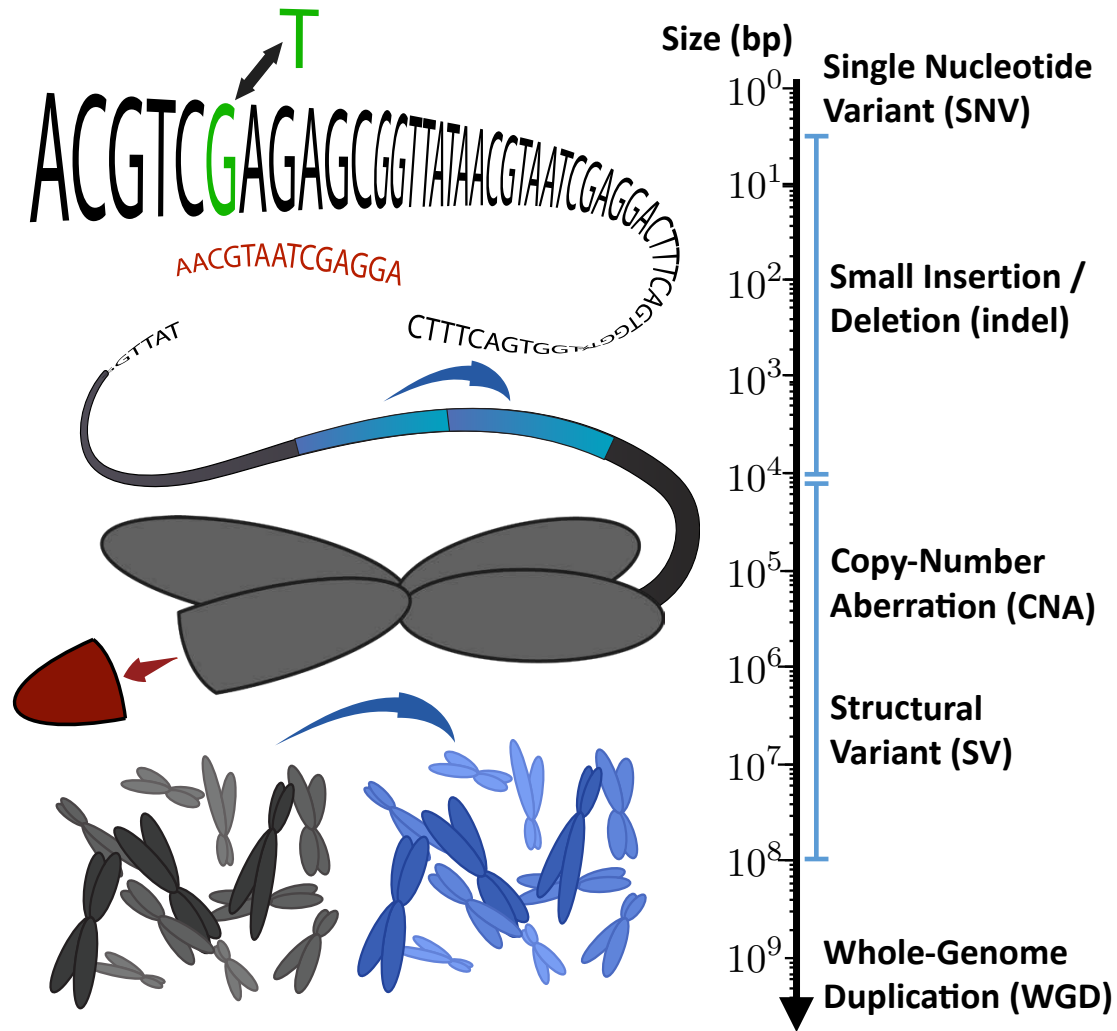


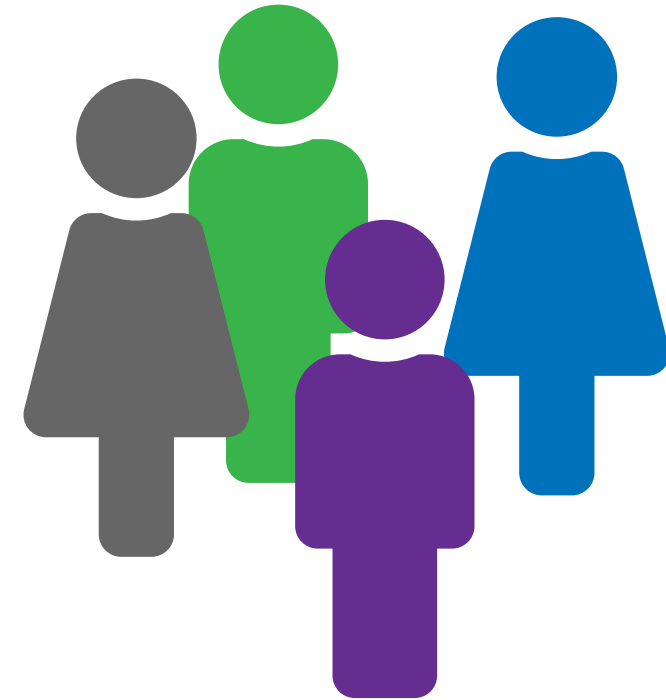
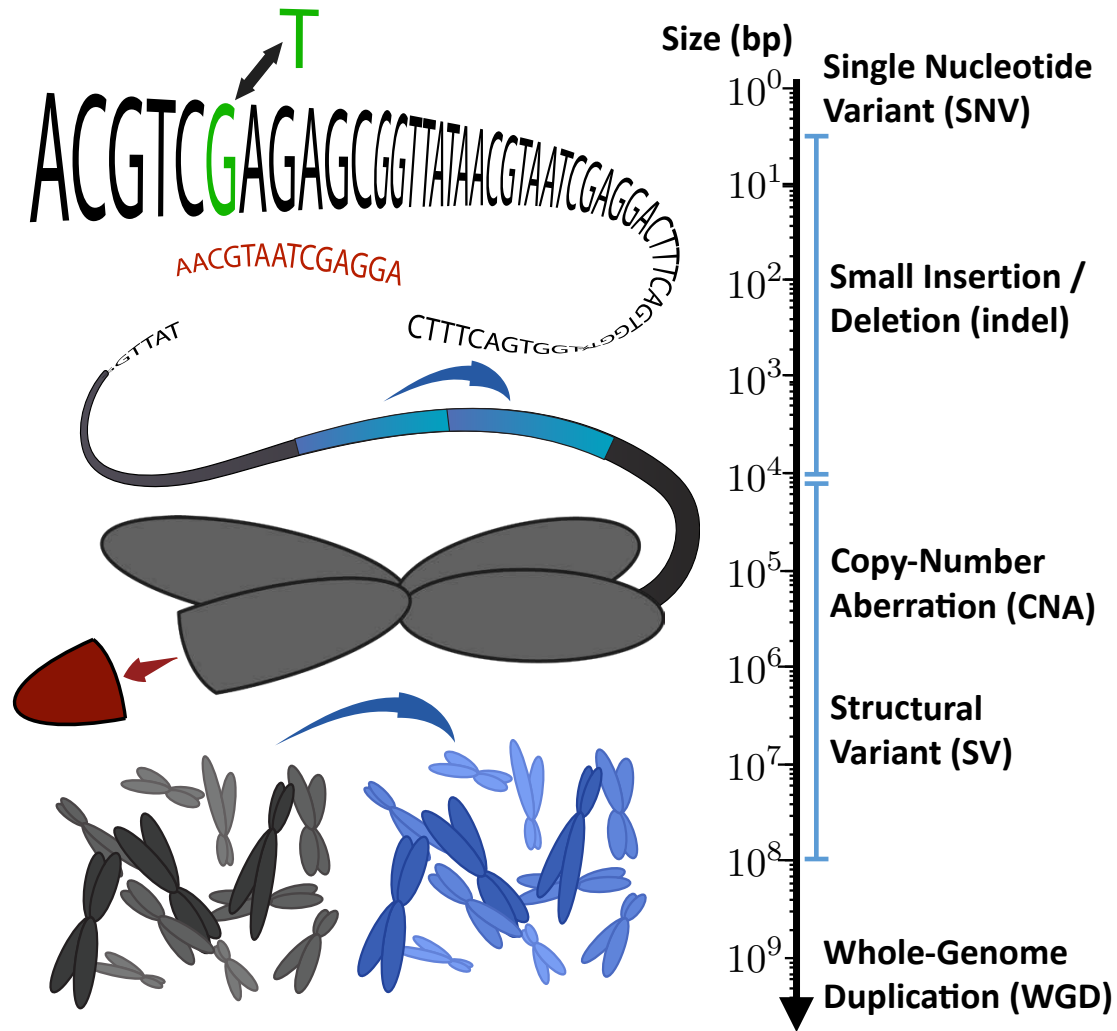
Figure 1. Acquired Capabilities of Cancer

We suggest that most if not all cancers have acquired the same set of functional capabilities during their development, albeit through various mechanistic strategies.

Cancer is Caused by Somatic Mutations

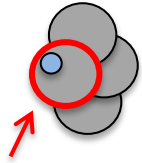


Cancer is Caused by Somatic Mutations



Question: Why is there inter-tumor heterogeneity?

Tumorigenesis: Cell Mutation



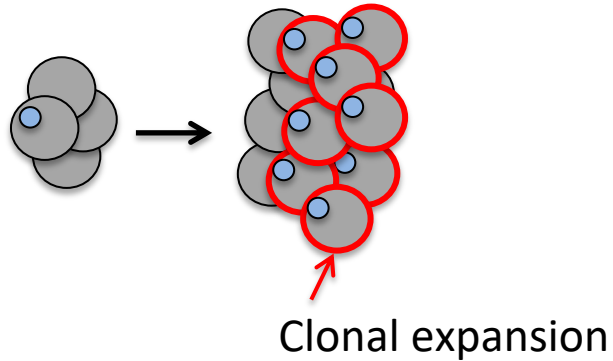
Founder
tumor cell


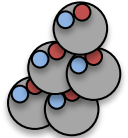


with somatic mutation: ●
(e.g. BRAF V600E)

Tumorigenesis: Cell Mutation, Division

Clonal Evolution Theory of Cancer

[Nowell, 1976]

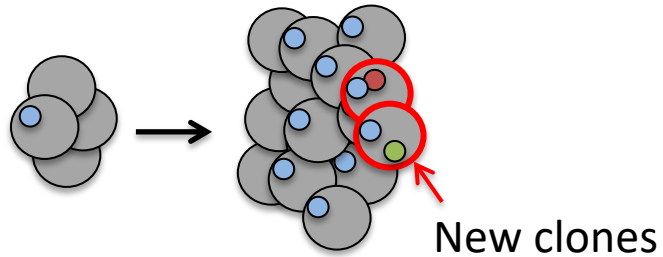



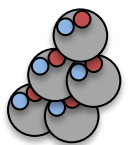


Clone  is a group  of cells with the same mutations { ,  }

Tumorigenesis: Cell Mutation, Division

Clonal Evolution Theory of Cancer

[Nowell, 1976]

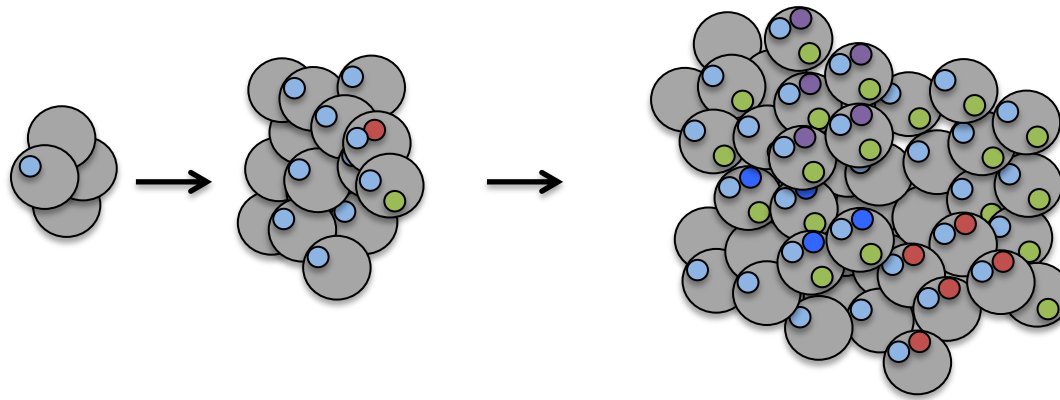


Clone  is a group  of cells with the same mutations { ,  }

Tumorigenesis: Cell Mutation, Division

Clonal Evolution Theory of Cancer

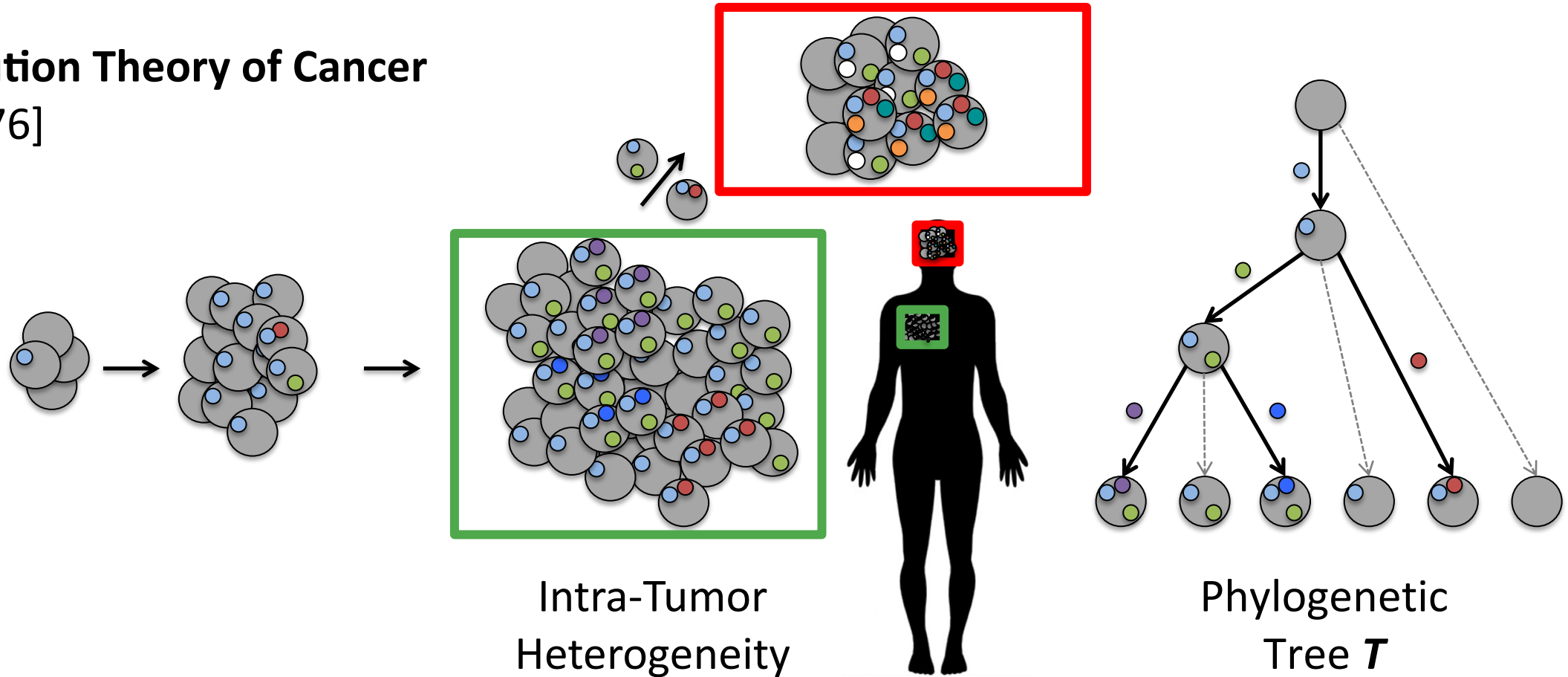
[Nowell, 1976]



Intra-Tumor
Heterogeneity

Tumorigenesis: Cell Mutation, Division & Migration

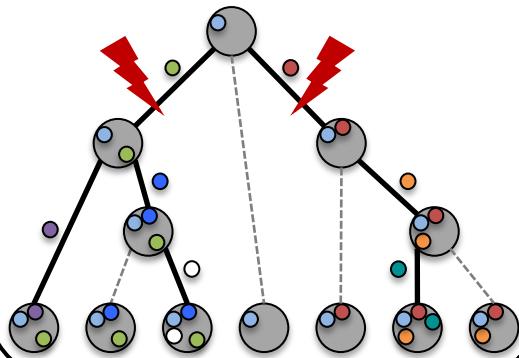
Clonal Evolution Theory of Cancer [Nowell, 1976]



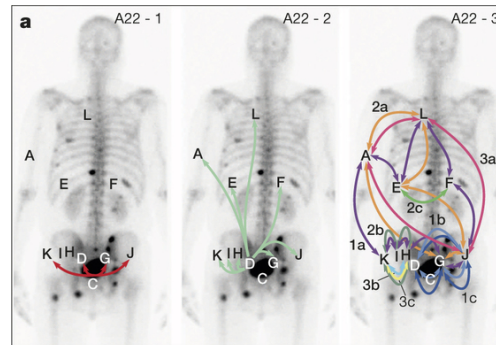
Question: Why are tumor phylogenies important?

Phylogenies are Key to Understanding Cancer

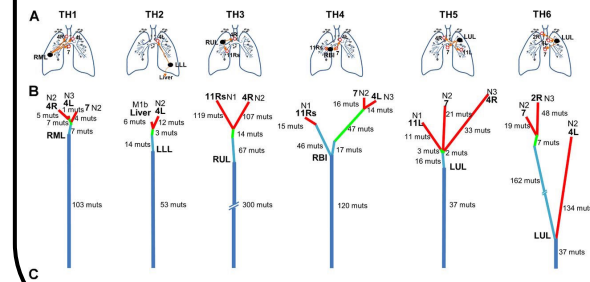
Identify targets for treatment



Understand metastatic development

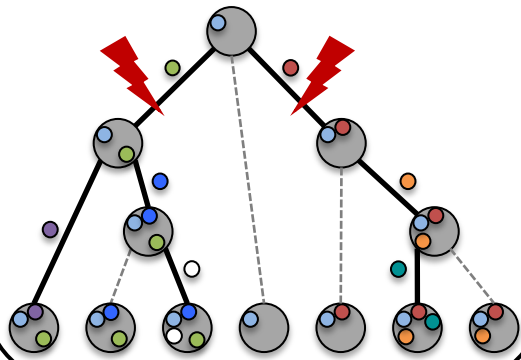


Recognize common patterns of tumor evolution across patients

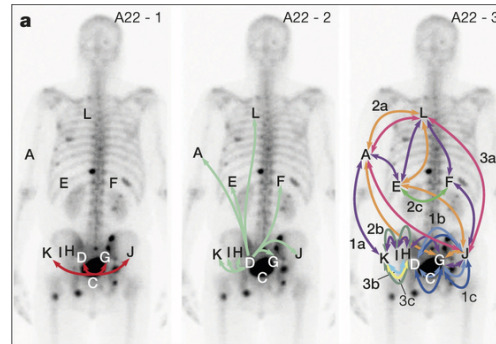


Phylogenies are Key to Understanding Cancer

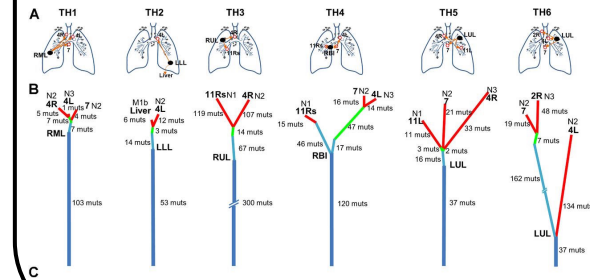
Identify targets for treatment



Understand metastatic development



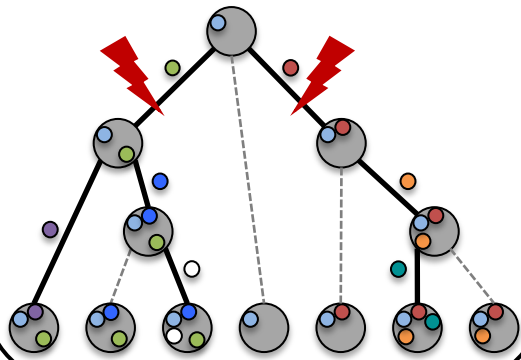
Recognize common patterns of tumor evolution across patients



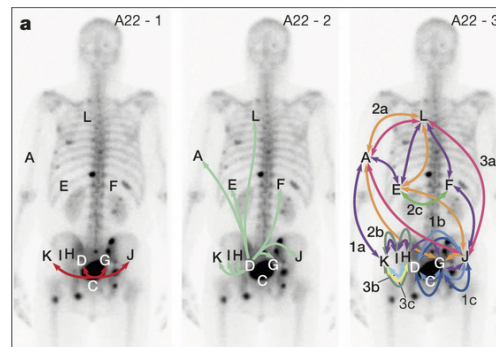
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Phylogenies are Key to Understanding Cancer

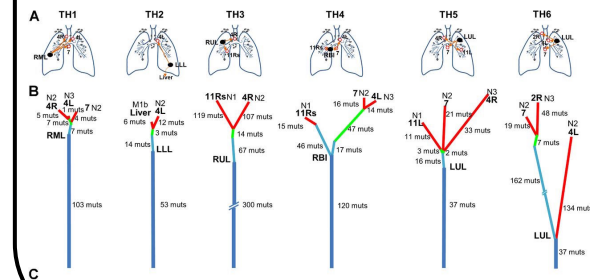
Identify targets for treatment



Understand metastatic development



Recognize common patterns of tumor evolution across patients



These downstream analyses **critically rely** on accurate tumor phylogeny inference

Key challenge in phylogenetics:
Accurate phylogeny inference from data at present time

Lecture Outline

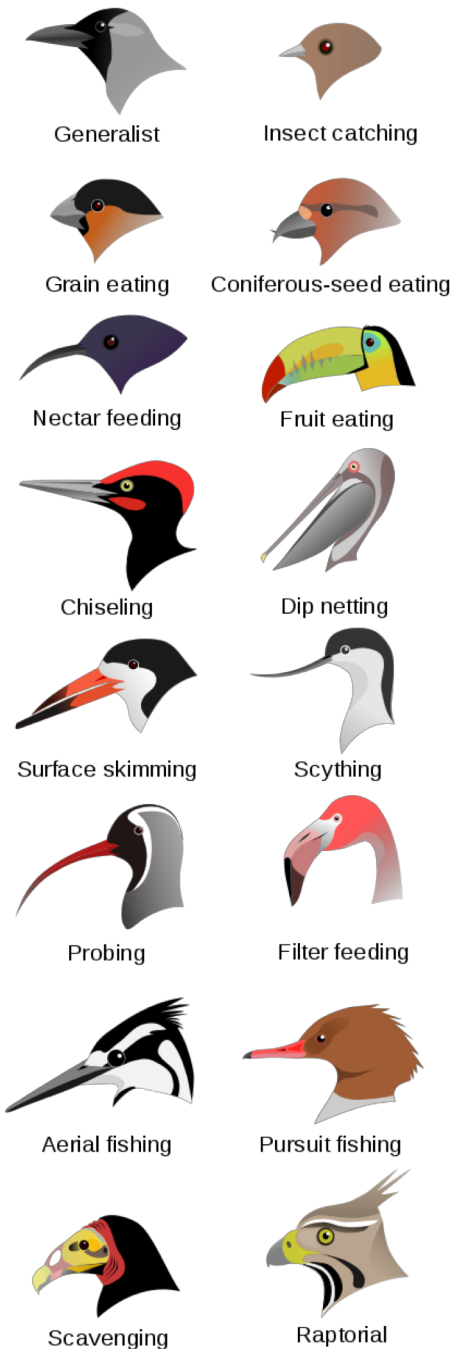
- Primer on Molecular Biology
- Primer on Computational Biology
- Primer on Cancer Biology
- Tumor Phylogeny Inference

Reading

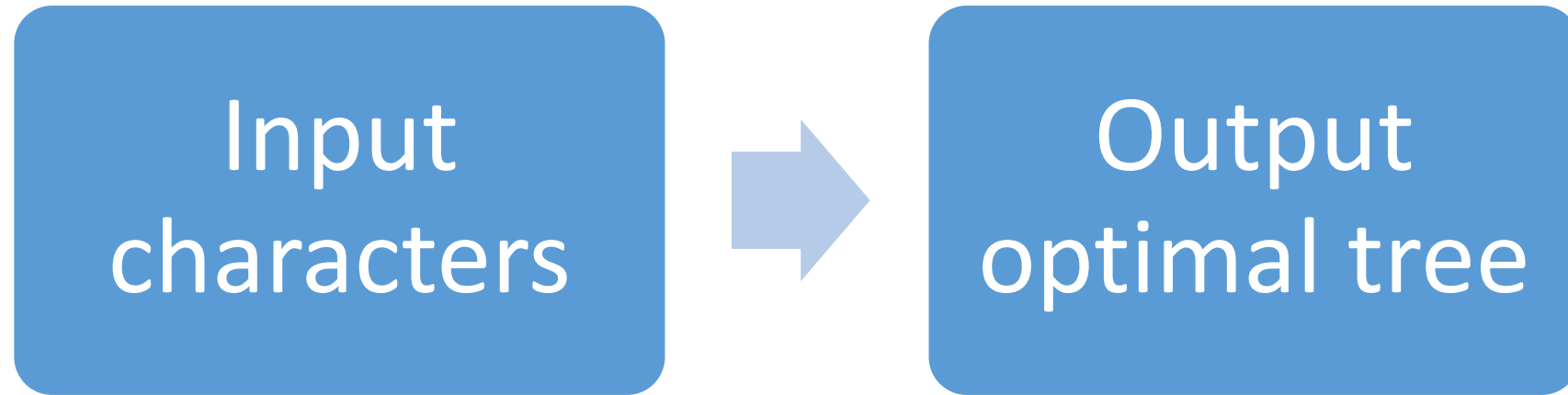
- “Biology for Computer Scientists” by Lawrence Hunter
(http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)

Character-Based Tree Reconstruction

- Characters may be morphological features
 - Shape of beak {generalist, insect catching, ...}
 - Number of legs {2,3,4, ..}
 - Hibernation {yes, no}
- Character may be nucleotides/amino acids
 - {A, T, C, G}
 - 20 amino acids
- Values of a character are called states
 - We assume discrete states

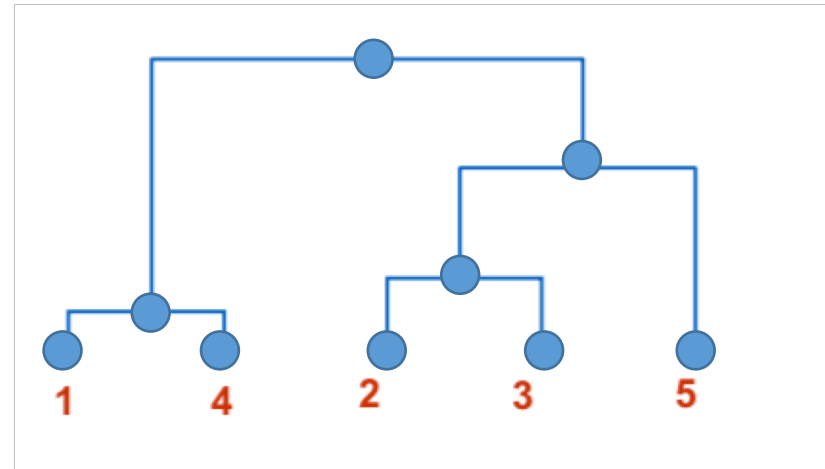


Character-Based Phylogeny Reconstruction

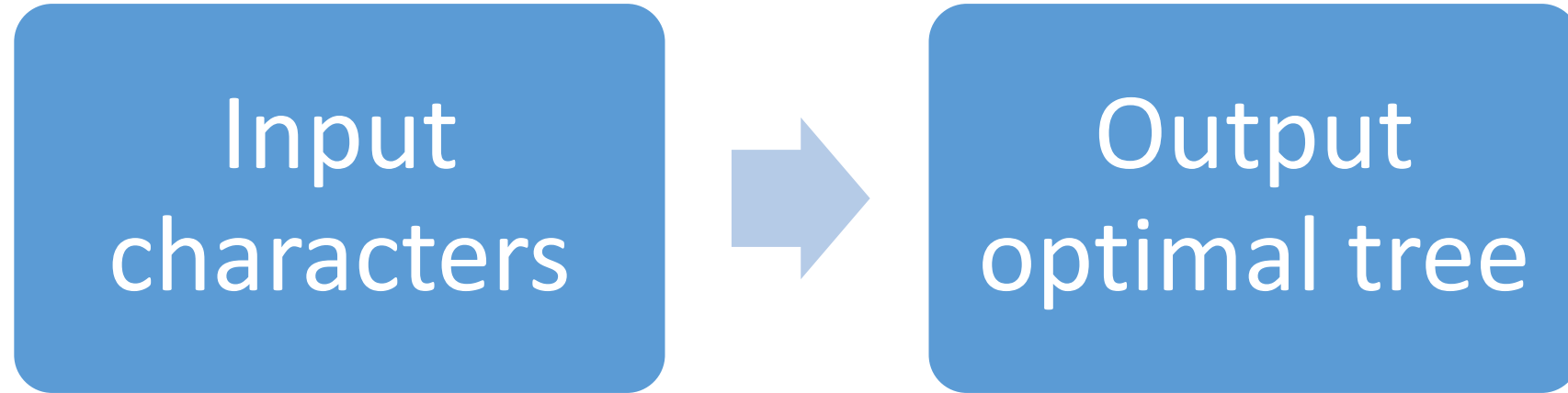


Question: What is optimal?

Want: Optimization criterion



Character-Based Phylogeny Reconstruction

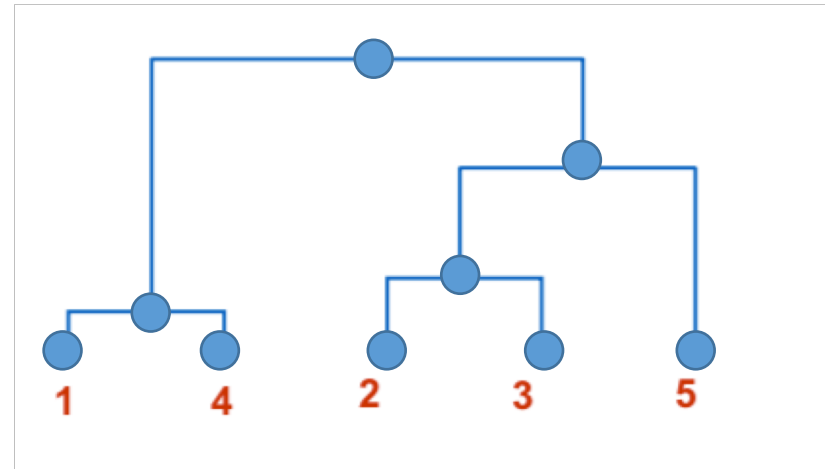


Question: What is optimal?

Want: Optimization criterion

Question: How to optimize this criterion?

Want: Algorithm

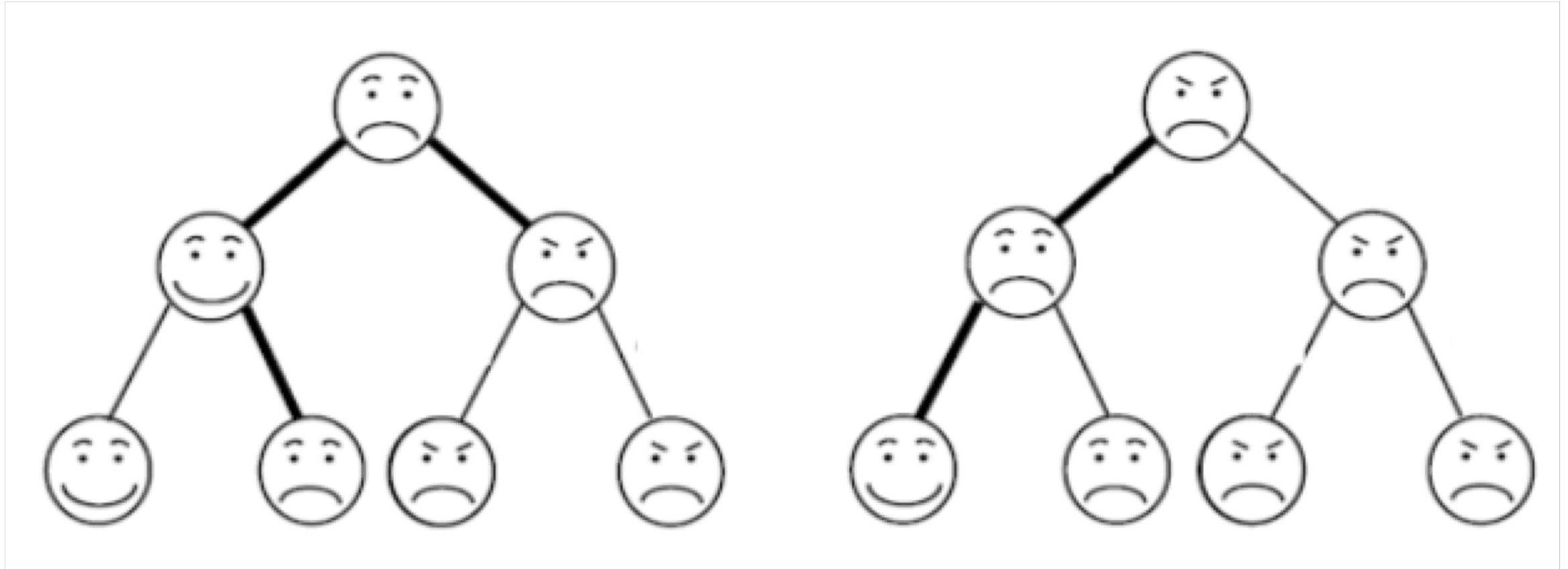


Character-Based Phylogeny Reconstruction: Input

Characters / states	State 1	State 2
Mouth	Smile	Frown
Eyebrows	Normal	Pointed

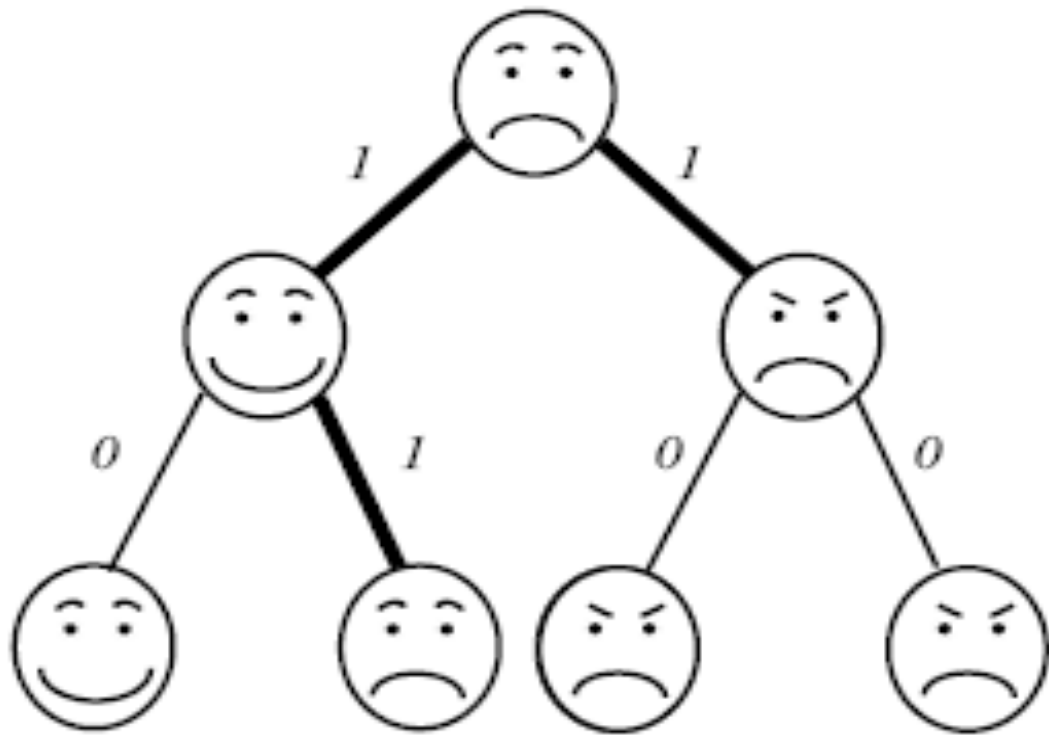


Character-Based Phylogeny Reconstruction: Criterion

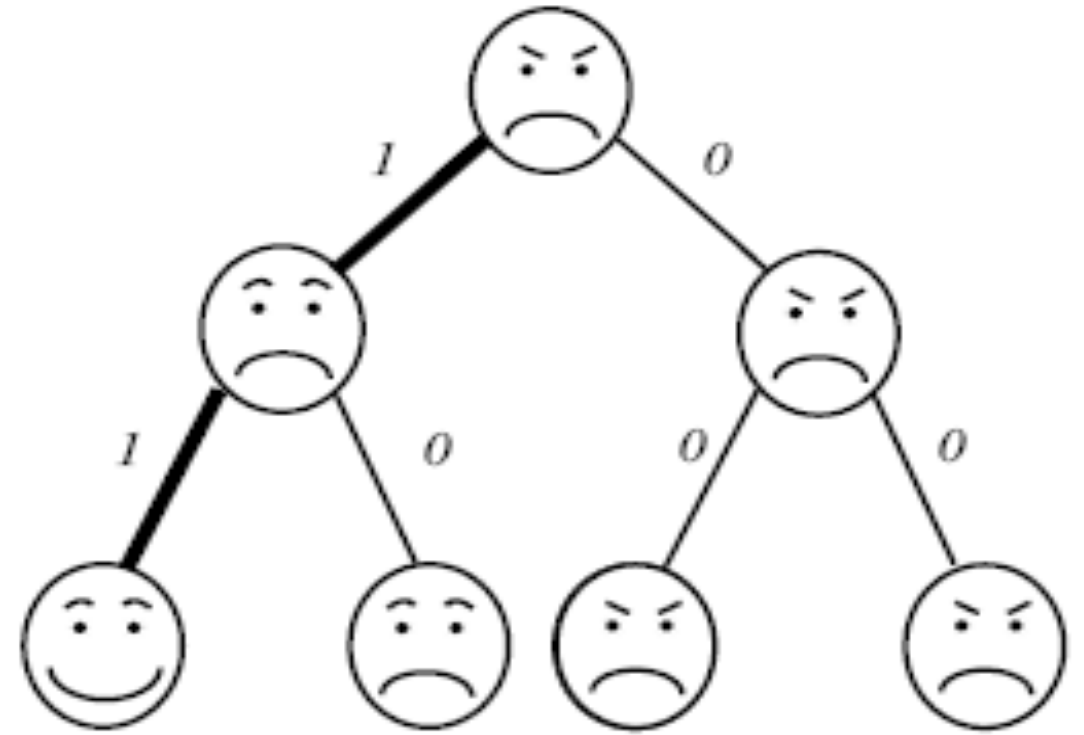


Question: Which tree is better?

Character-Based Phylogeny Reconstruction: Criterion



(a) *Parsimony Score=3*

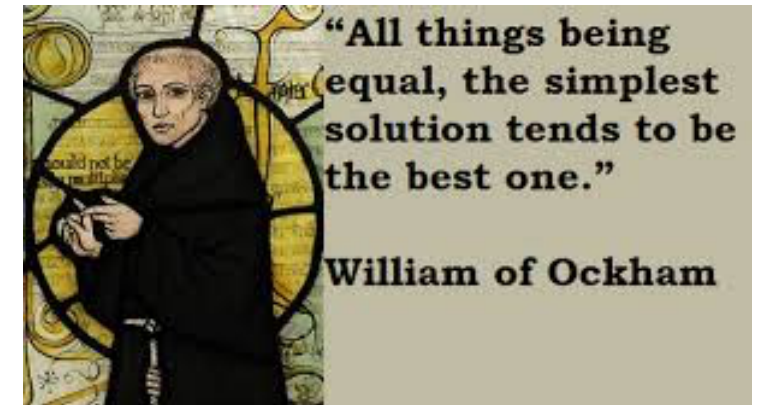


(b) *Parsimony Score=2*

Parsimony: minimize number of changes on edges of tree

Why Parsimony?

- Ockham's razor: “simplest” explanation for data
- Assumes that observed character differences resulted from the fewest possible mutations
- Seeks tree with the lowest **parsimony score**, i.e. the sum of all (costs of) mutations in the tree.



A Small and a Large Problem

Small Maximum Parsimony Phylogeny Problem:

Given $m \times n$ matrix $A = [a_{i,j}]$ and tree T with m leaves, find assignment of character states to each internal vertex of T with minimum parsimony score.

Large Maximum Parsimony Phylogeny Problem:

Given $m \times n$ matrix $A = [a_{i,j}]$, find a tree T with m leaves labeled according to A and an assignment of character states to each internal vertex of T with minimum parsimony score.

A Small and a Large Problem

Small Maximum Parsimony Phylogeny Problem:

Given $m \times n$ matrix $A = [a_{i,j}]$ and tree T with m leaves, find assignment of character states to each internal vertex of T with minimum parsimony score.

Large Maximum Parsimony Phylogeny Problem:

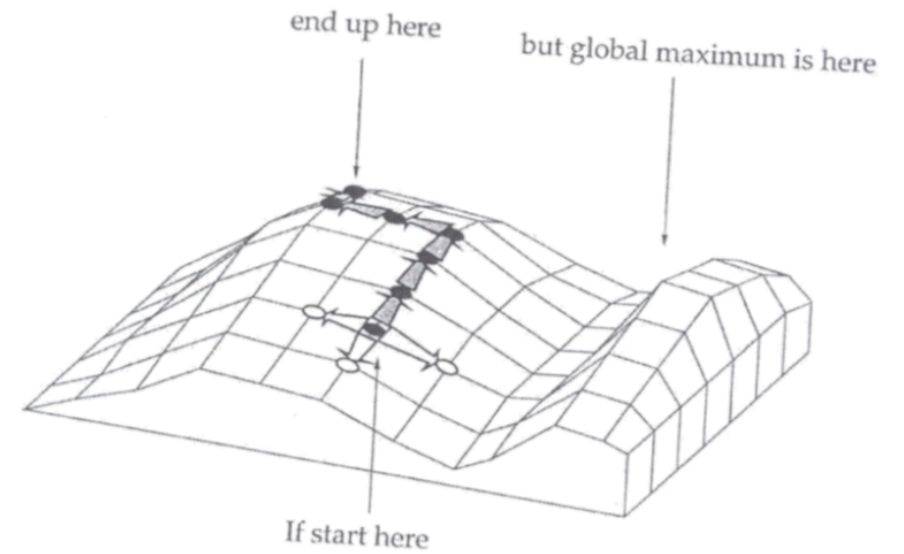
Given $m \times n$ matrix $A = [a_{i,j}]$, find a tree T with m leaves labeled according to A and an assignment of character states to each internal vertex of T with minimum parsimony score.

Question: Are both problems easy (i.e. in P)?

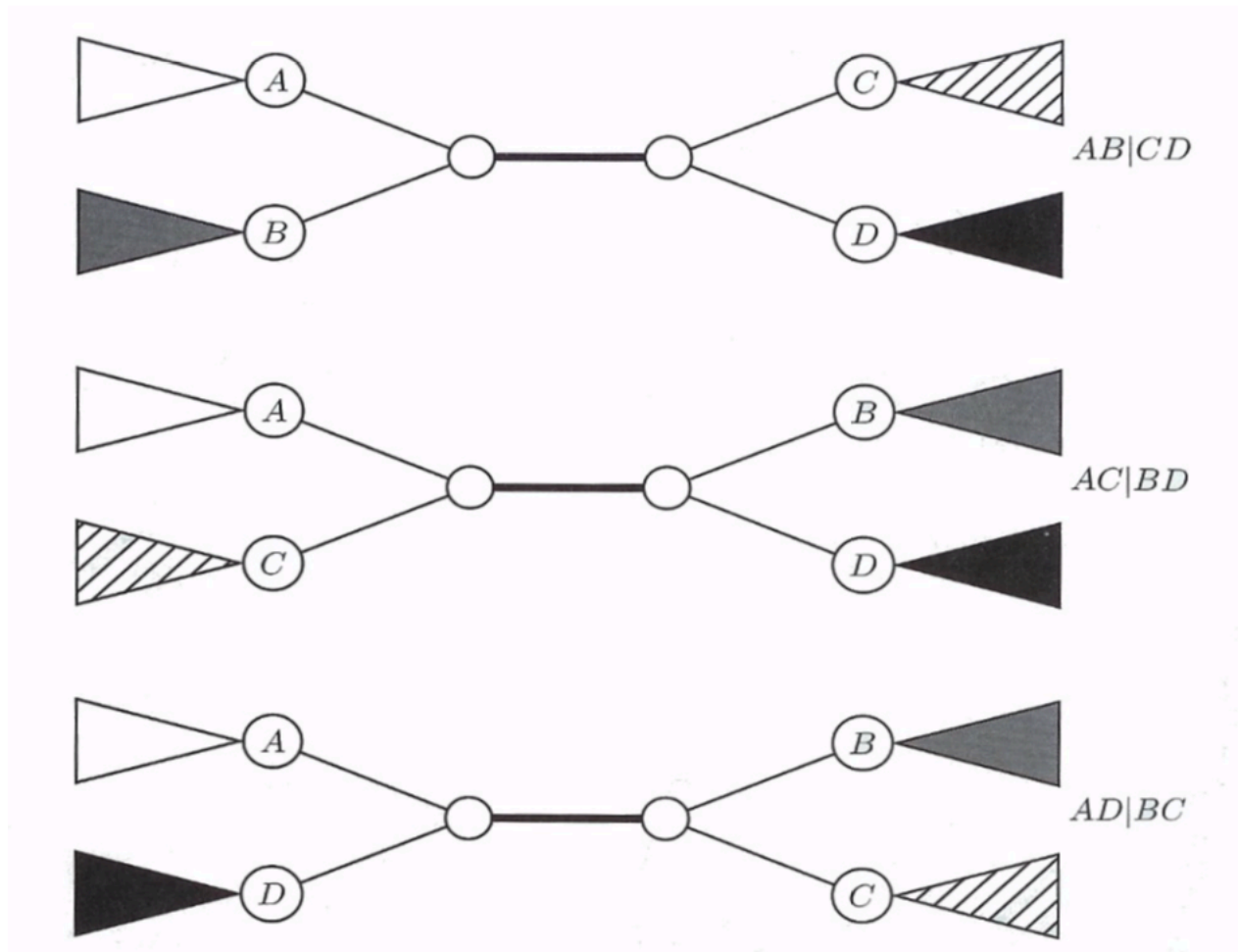
Large Maximum Parsimony Phylogeny

- This problem is NP-hard
- Heuristics using local search (tree moves)
 1. Start with an arbitrary tree T .
 2. Check “neighbors” of T .
 3. Move to a neighbor if it provides the best improvement in parsimony/likelihood score.

Caveats:
Could be stuck in **local** optimum, and not achieve global optimum



Example: Nearest-Neighbor Interchange (NNI)

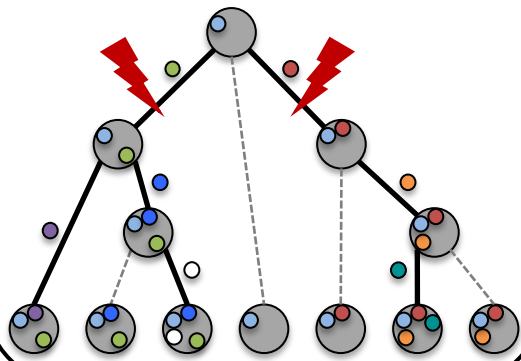


Rearrange four subtrees
defined by one
internal edge

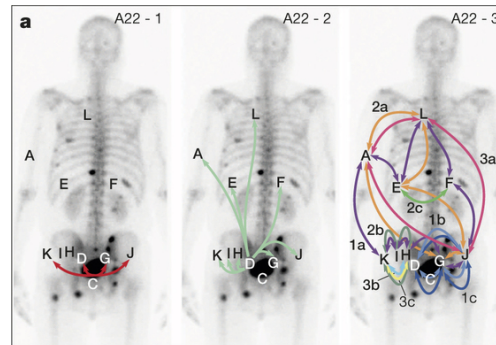
Figure: Jones and Pevzner

Phylogenies are Key to Understanding Cancer

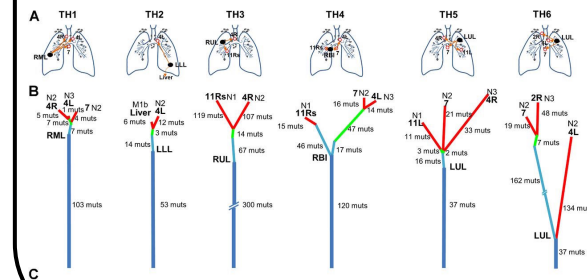
Identify targets for treatment



Understand metastatic development



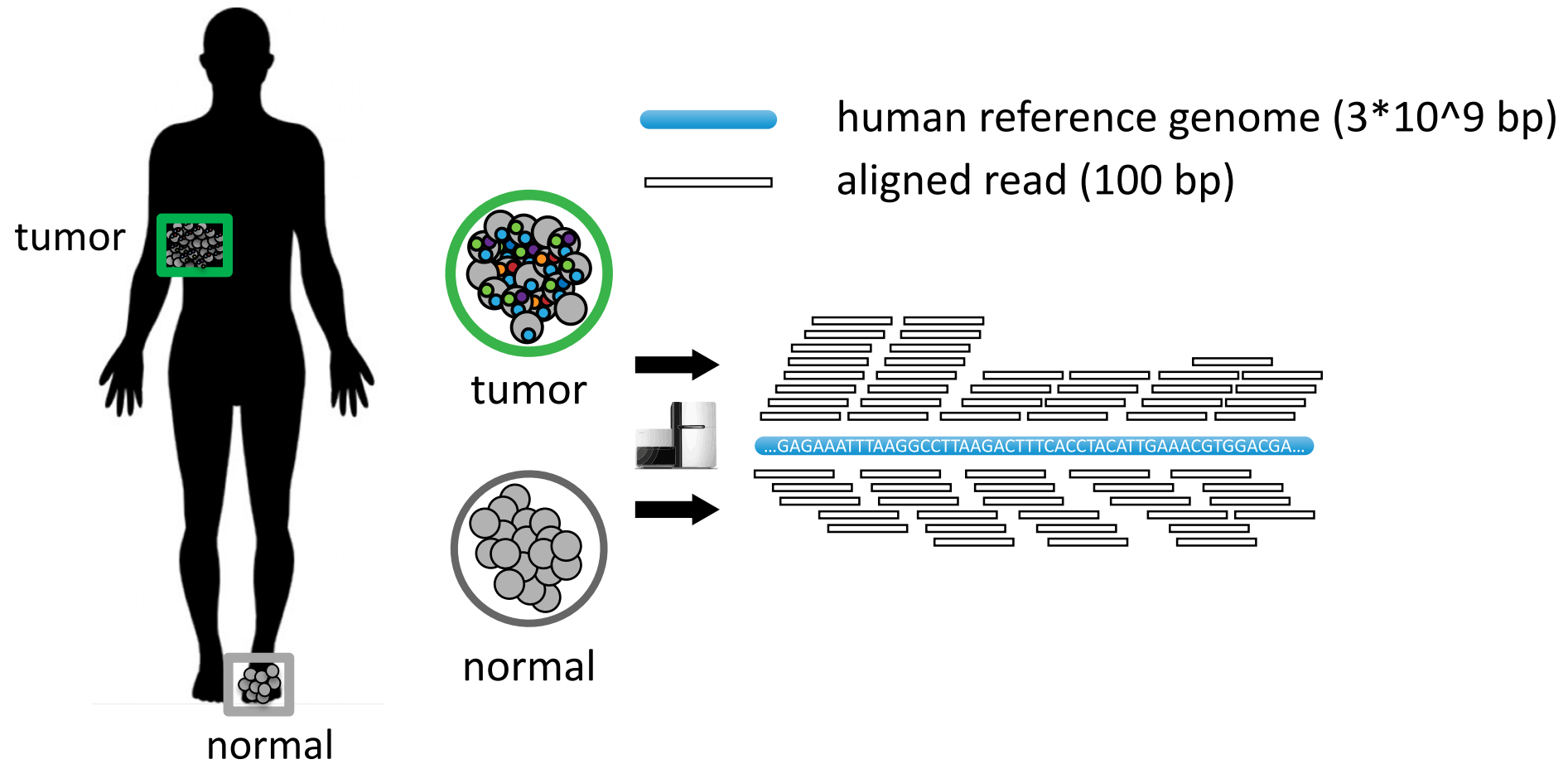
Recognize common patterns of tumor evolution across patients



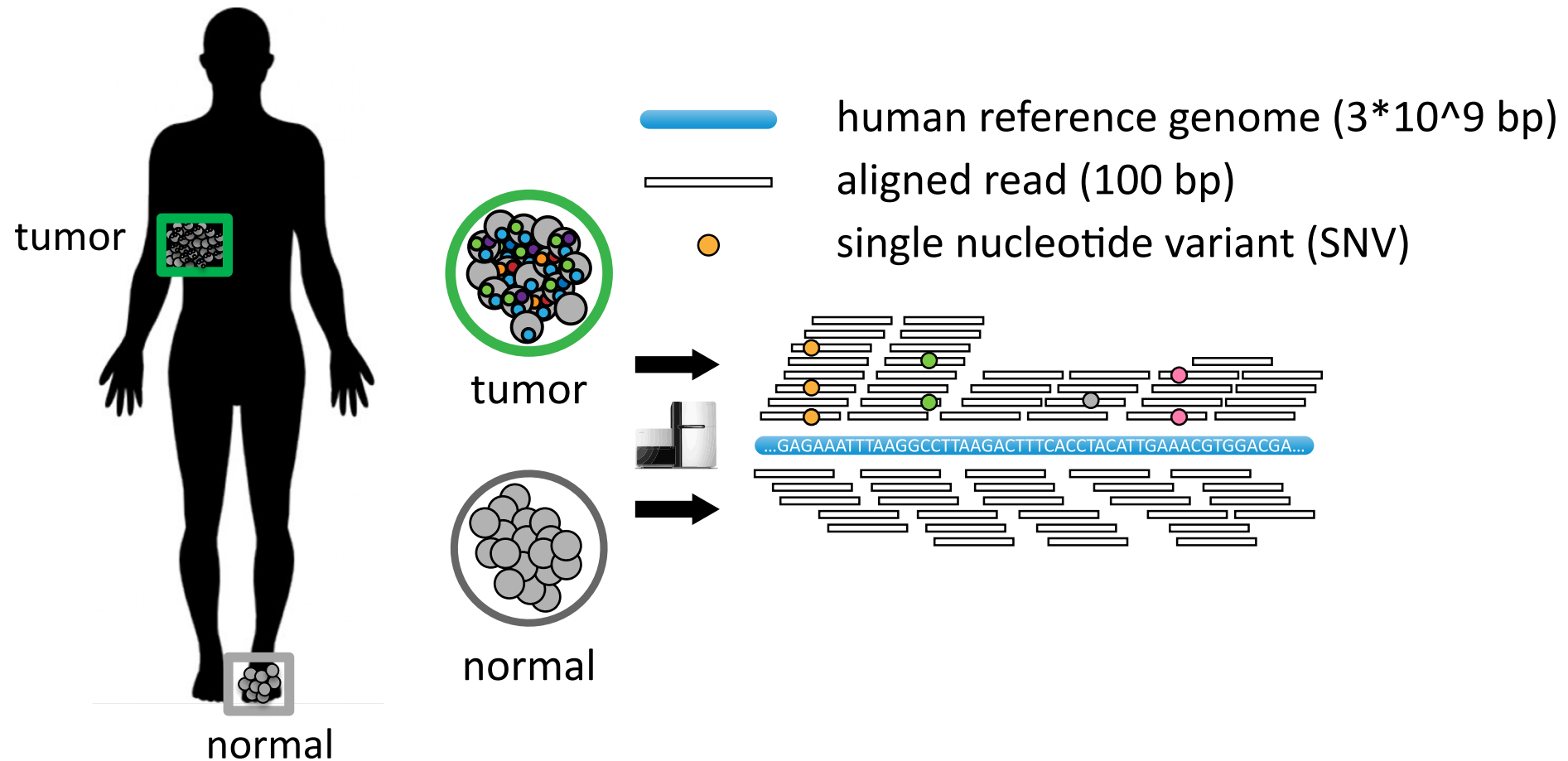
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Key challenge in phylogenetics:
Accurate phylogeny inference from data at present time

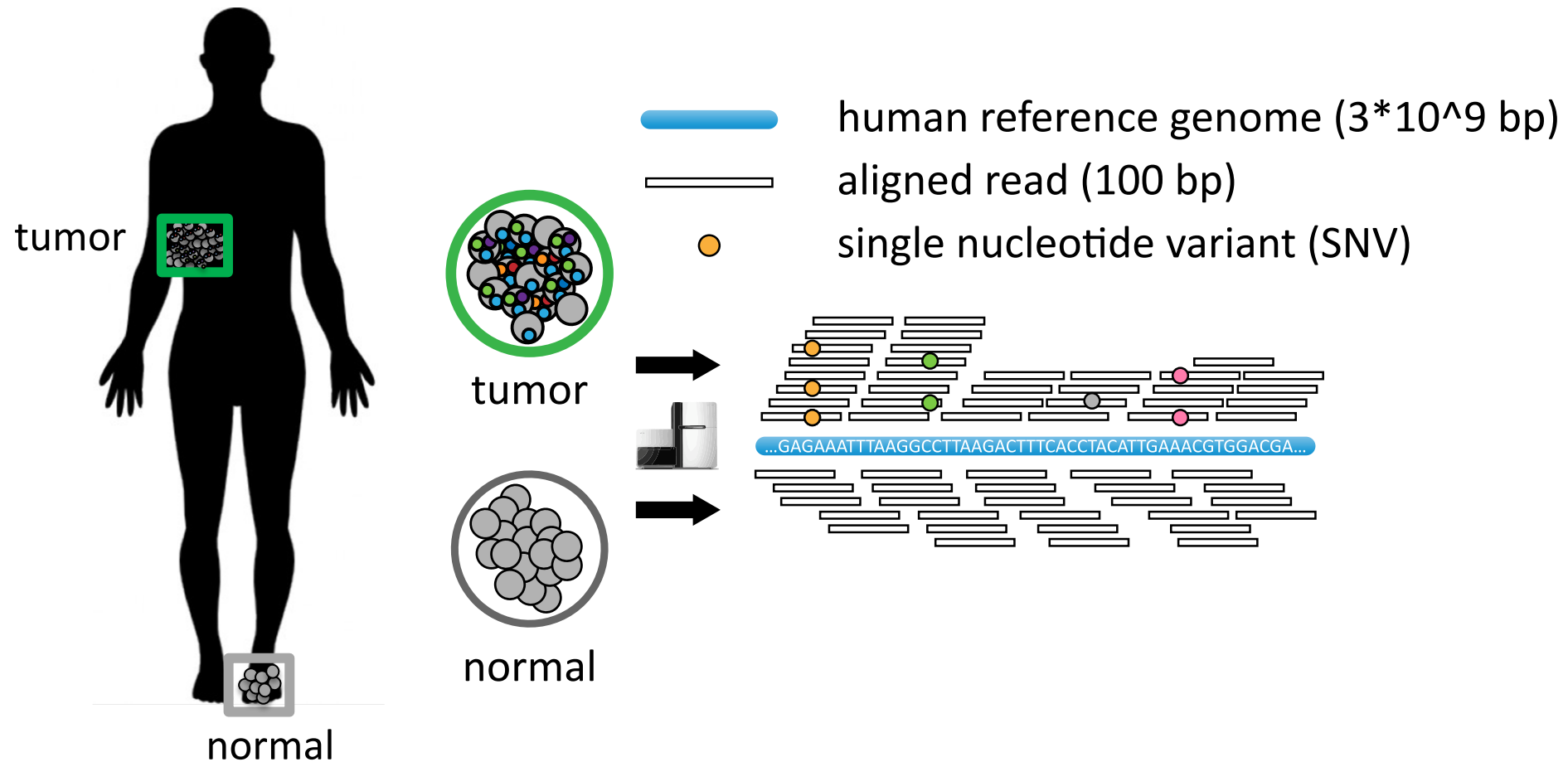
Additional Challenge in Cancer Phylogenetics



Additional Challenge in Cancer Phylogenetics



Additional Challenge in Cancer Phylogenetics



Additional challenge in cancer phylogenetics:
Phylogeny inference from **mixed bulk samples** at present time

Tumor Phylogeny Inference







Metastatic Colorectal Cancer (Patient CRC2)

[Kim et al., *Clin Cancer Res* 21(19), 2015]:

- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)

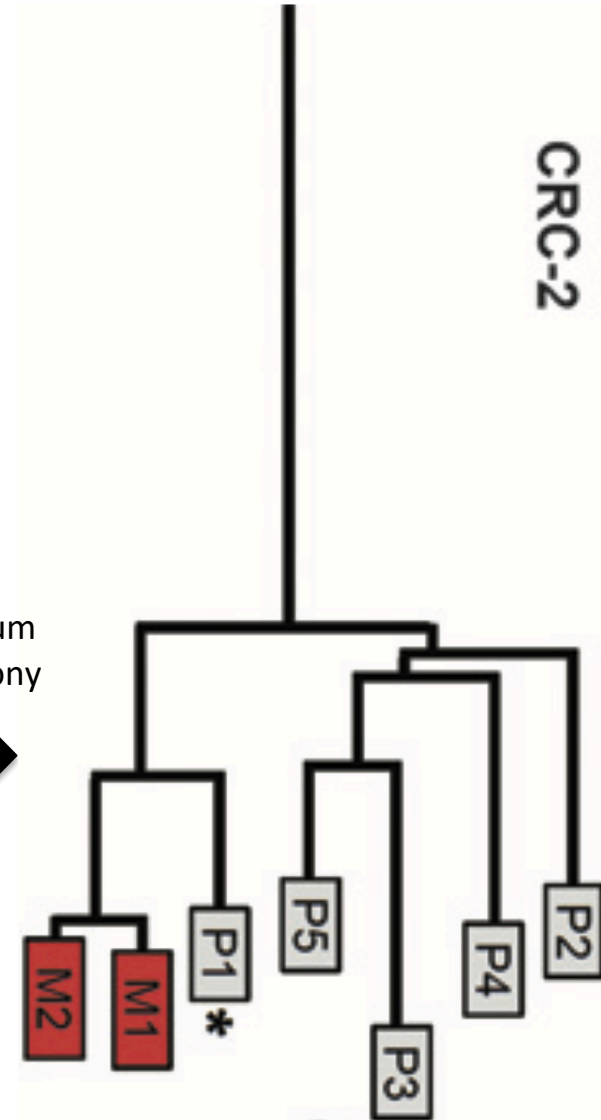
m samples

n mutations

						
P1	1	1	1	0	0	0
P2	1	1	0	1	0	0
P3	1	0	1	0	1	1
P4	0	1	1	0	0	0
P5	0	1	0	1	0	1
M1	1	1	0	0	1	0
M2	0	1	1	1	1	1

Binary Matrix B

Maximum
Parsimony









Tumor Phylogeny Inference

Metastatic Colorectal Cancer (Patient CRC2)

[Kim et al., *Clin Cancer Res* 21(19), 2015]:

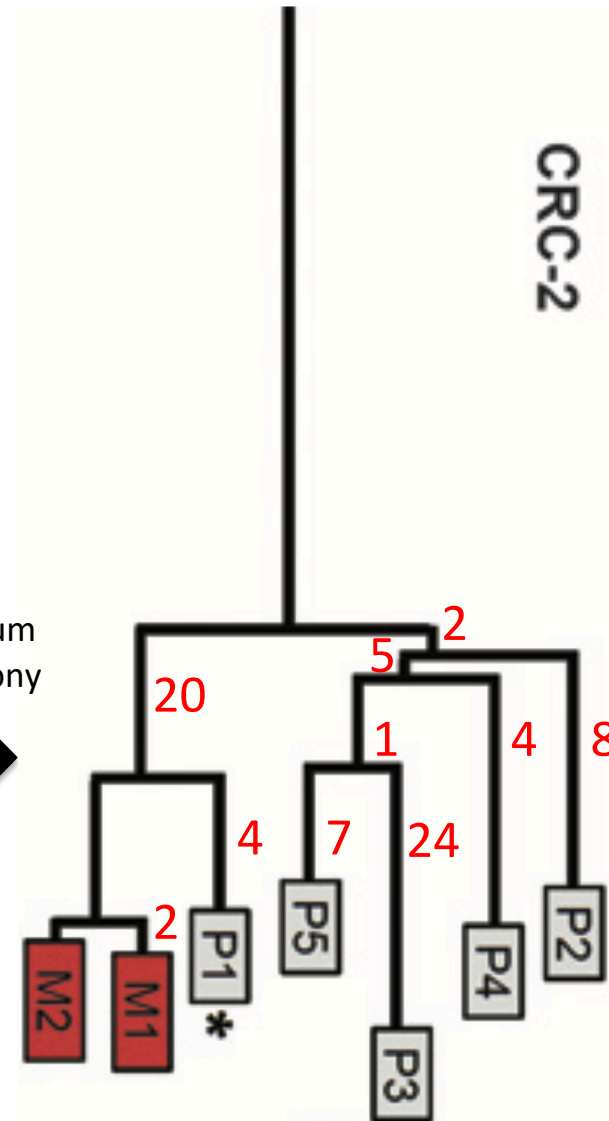
- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)
- 41 mutate more than once (**homoplasy**)

m samples

		n mutations					
							
P1	1	1	1	1	0	0	0
P2	1	1	0	1	0	0	0
P3	1	0	1	0	1	1	1
P4	0	1	1	0	0	0	0
P5	0	1	0	1	0	1	1
M1	1	1	0	0	1	0	0
M2	0	1	1	1	1	1	1

Binary Matrix B

Maximum
Parsimony

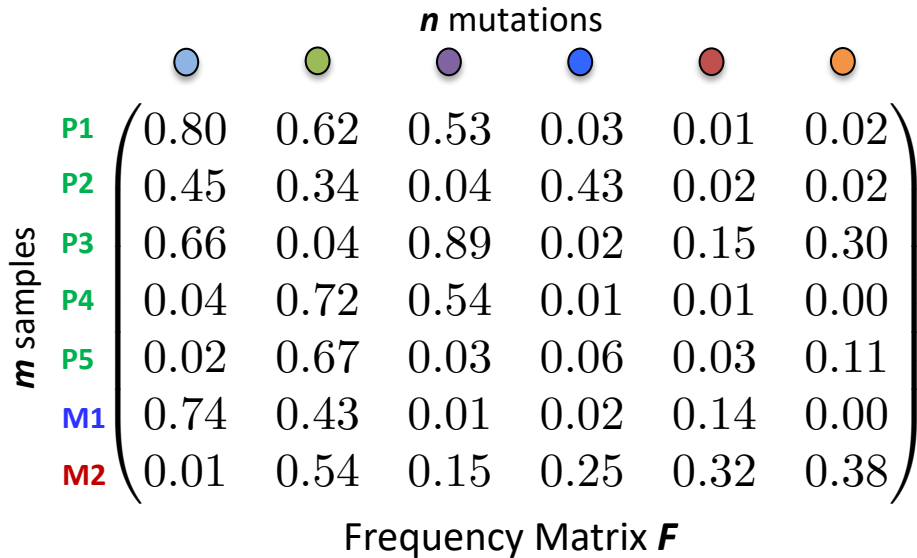


Heuristic for Tumor Phylogeny Inference

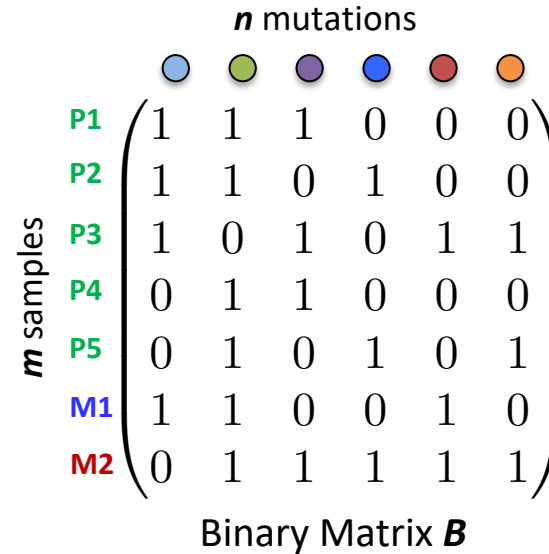
Metastatic Colorectal Cancer (Patient CRC2)

[Kim et al., *Clin Cancer Res* 21(19), 2015]:

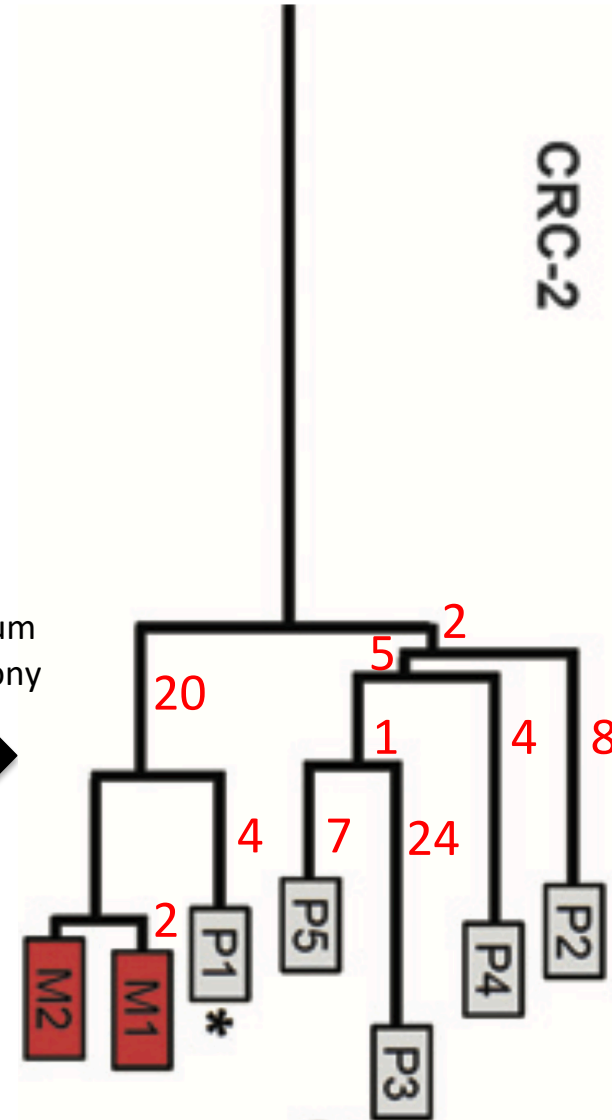
- 5 primary samples (P1-P5)
- 2 metastases (M1-M2)
- 412 single-nucleotide variants (SNVs)
- 41 mutate more than once (**homoplasy**)



Discretize



Maximum Parsimony



Resulting **sample tree** is **not** representative of the division/mutation history or the migration history

Summary

- DNA, RNA and proteins are sequences
 - Central dogma of molecular biology: DNA -> RNA -> protein
- Problem != algorithm
- Key challenge in computational biology is translating a biological problem into a computational problem
- Cancer is a genetic disease caused by somatic mutations
- Inter-tumor heterogeneity and intra-tumor heterogeneity:
 - *Not only is every tumor different, but so is every tumor cell...*
- Reading:
 - “Biology for Computer Scientists” by Lawrence Hunter
(http://www.el-kebir.net/teaching/CS466/Hunter_BIO_CS.pdf)