Bioinformatics and Computational Biology

Mohammed El-Kebir

Computer Science • University of Illinois at Urbana-Champaign

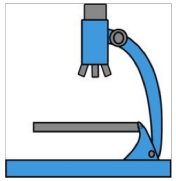
What is Computational Biology/Bioinformatics?

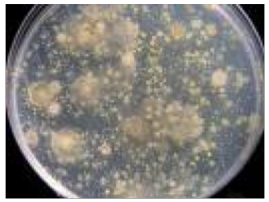
Computational biology and **bioinformatics** is an interdisciplinary field that develops and applies **computational methods** to analyze large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or **discover new biology**.

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

Technology and Bioinformatics are Transforming Biology

Until late 20th Century





Hypothesis Generation and Validation

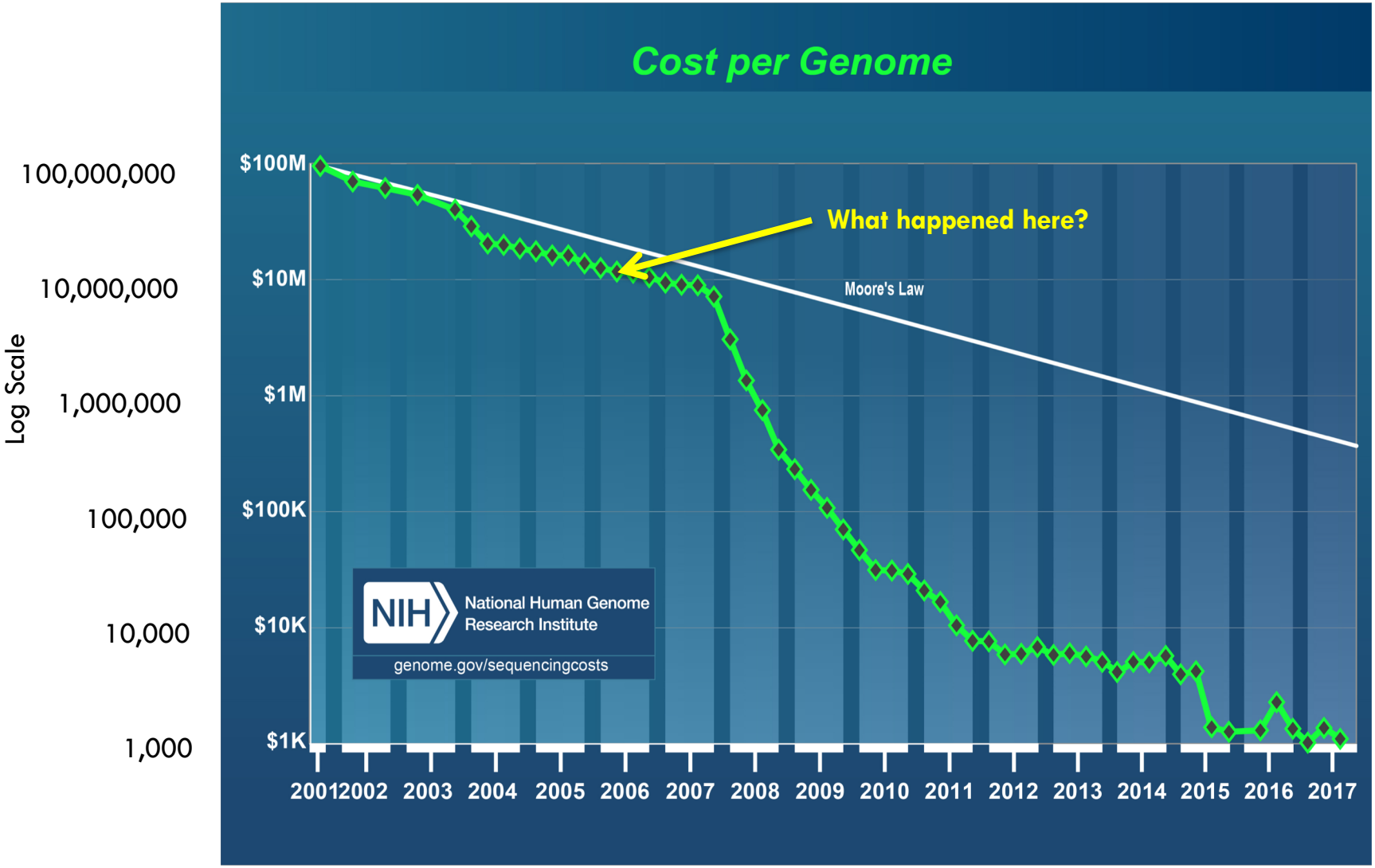
21th Century and Beyond

Algorithms

Hypothesis Generation and Validation

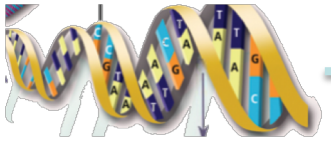
High throughput technologies

A Deluge of Data



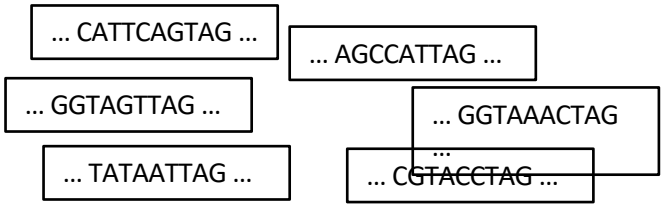
Question: What does it mean that we can sequence a genome?

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



Genome
Millions -billions nucleotides

Next-generation DNA sequencing



10-100's million noisy ***reads***
R*eads:* 30-1000 nucleotides

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	Worst Jobs
1. Actuary	200. Newspaper reporter
2. Audiologist	199. Lumberjack
3. Mathematician	198. Enlisted Military Personnel
4. Statistician	197. Cook
5. Biomedical Engineer	196. Broadcaster
6. Data Scientist	195. Photojournalist
7. Dental Hygienist	194. Corrections Officer
8. Software Engineer	193. Taxi Driver
9. Occupational Therapist	192. Firefighter
10. Computer Systems Analyst	191. Mail Carrier

http://www.careercast.com/jobs-rated/jobs-rated-report-2015-ranking-top-200-jobs

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."

Background for Bioinformatics Research

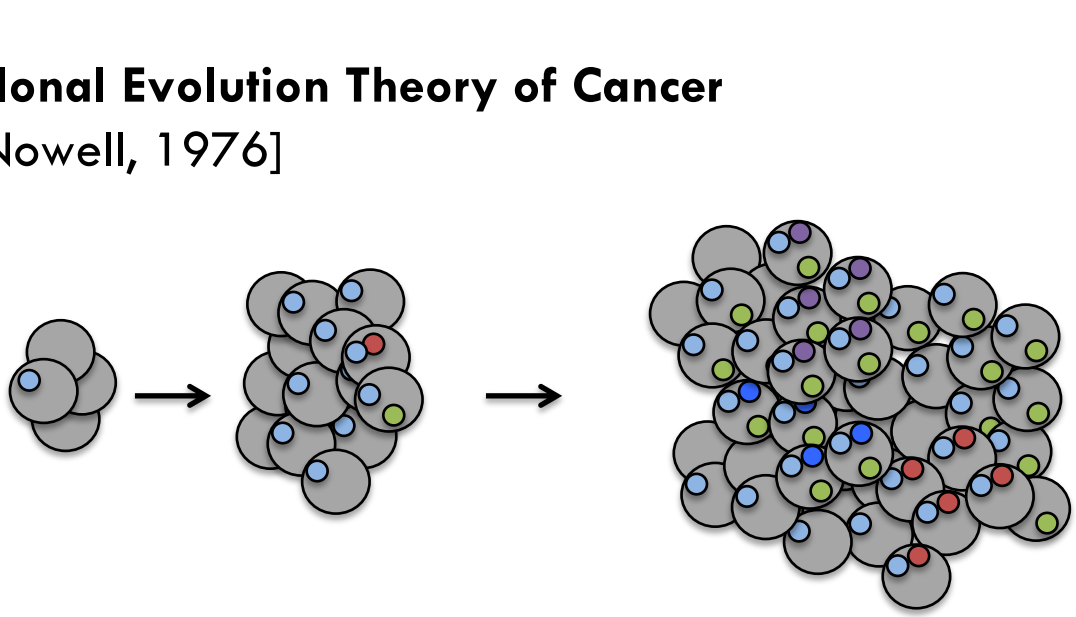
The usual computer science stuff, but especially

- Programming
- Statistics
- Algorithms and theory

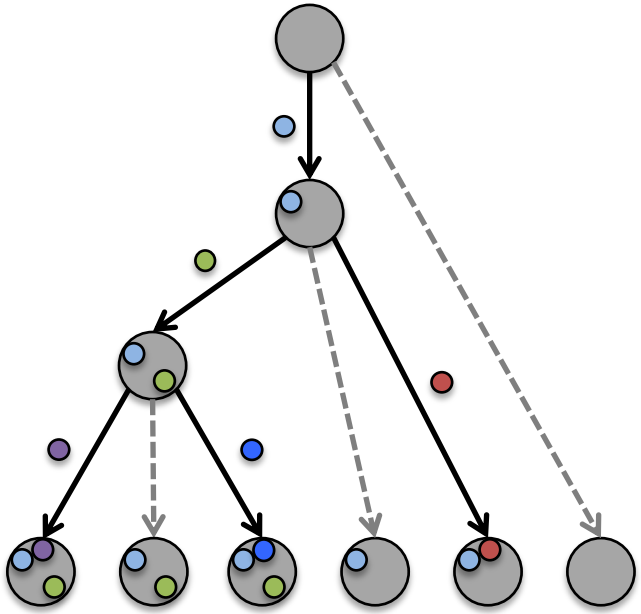
CS 466: Introduction to Bioinformatics! Good if you know some biology, but you can take CS 466, and learn it there!

Tumorigenesis: Cell Mutation & Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Intra-Tumor Heterogeneity

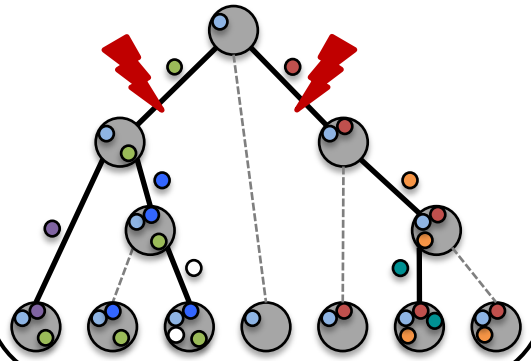


Phylogenetic Tree ***T***

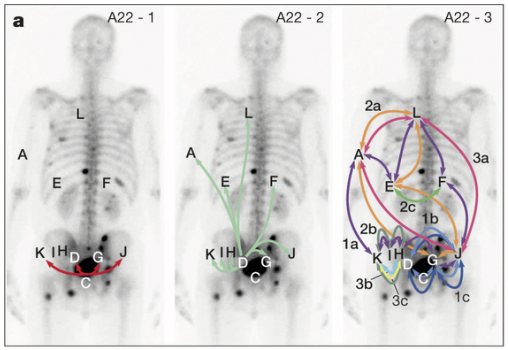
Question: Why are tumor phylogenies important?

Phylogenies are Key to Understanding Cancer

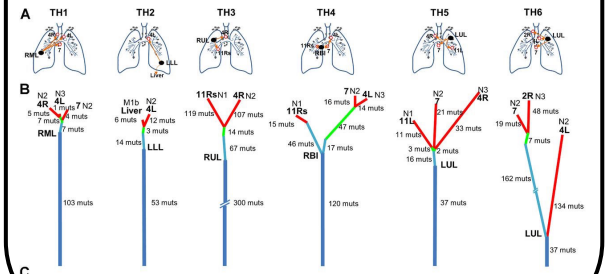
Identify targets for treatment



Understand metastatic development

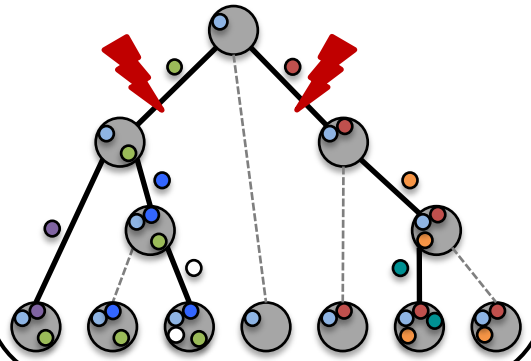


Recognize common patterns of tumor evolution across patients

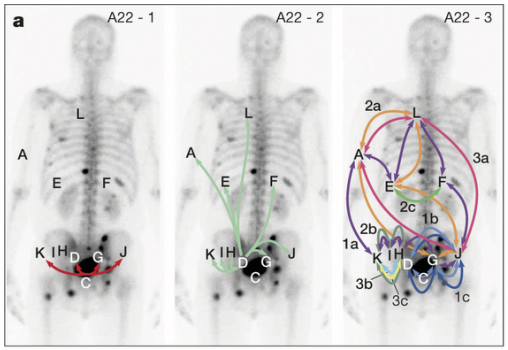


Phylogenies are Key to Understanding Cancer

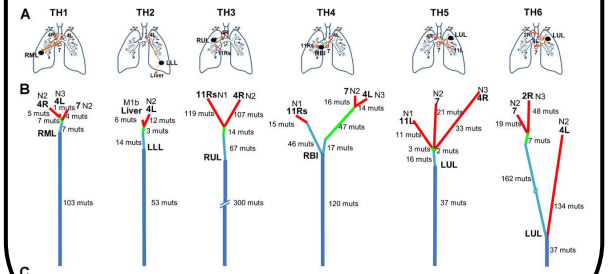
Identify targets for treatment



Understand metastatic development



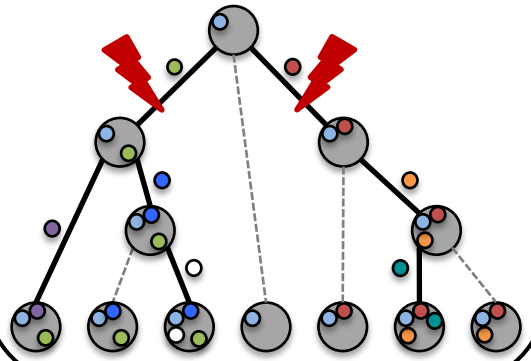
Recognize common patterns of tumor evolution across patients



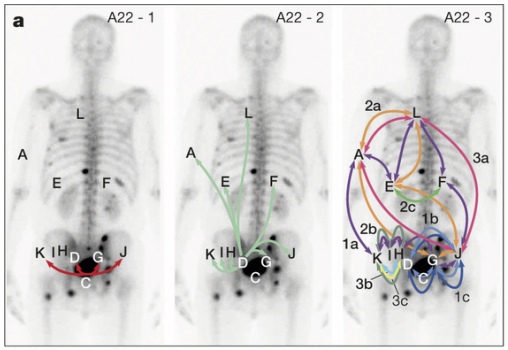
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Phylogenies are Key to Understanding Cancer

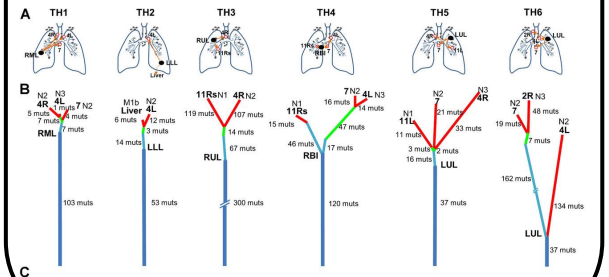
Identify targets for treatment



Understand metastatic development



Recognize common patterns of tumor evolution across patients

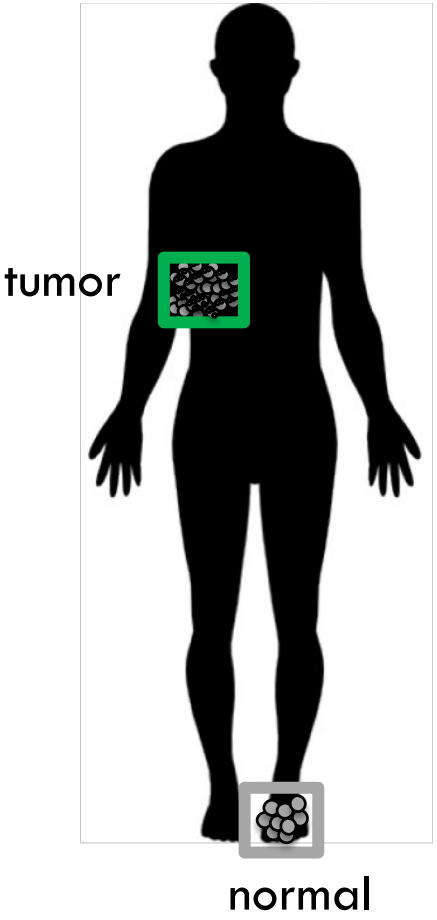


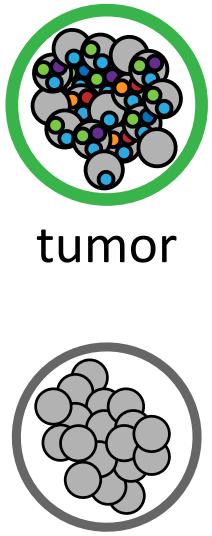
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Key challenge in phylogenetics:

Accurate phylogeny inference from data at present time

Additional Challenge in Cancer Phylogenetics



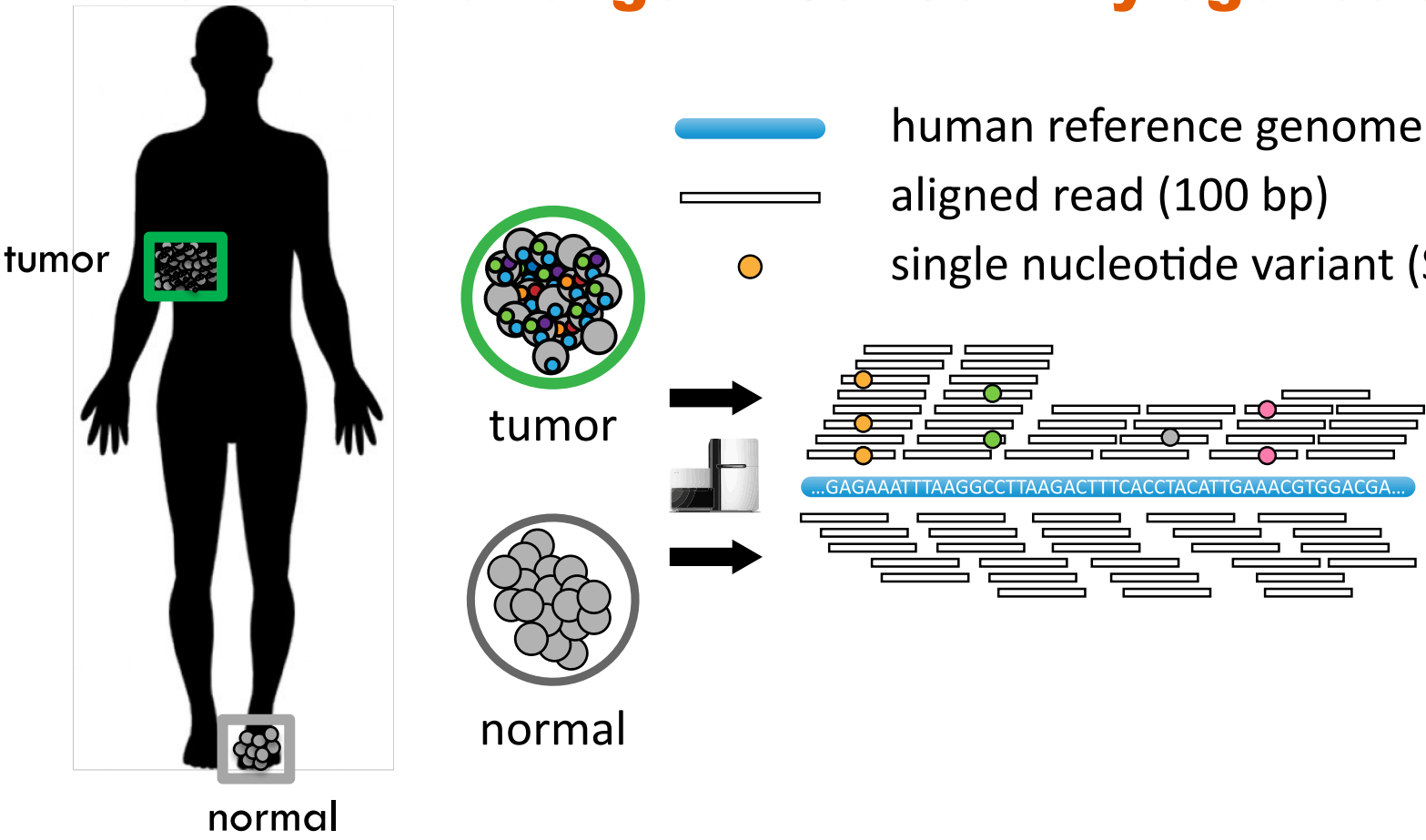


human reference genome (3*10^9 bp)

aligned read (100 bp)

single nucleotide variant (SNV)

Additional Challenge in Cancer Phylogenetics

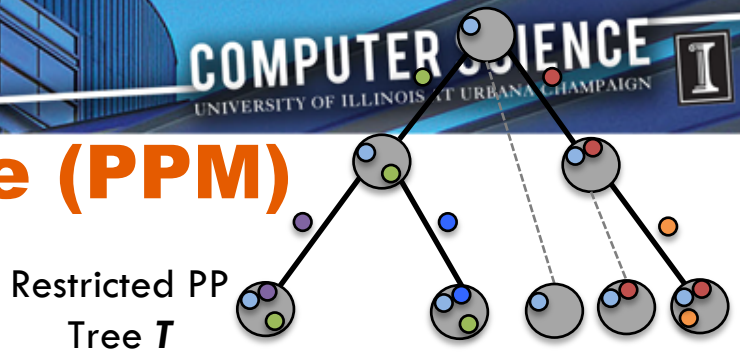


Additional challenge in cancer phylogenetics:
Phylogeny inference from **mixed bulk samples** at present time

Perfect Phylogeny Mixture (PPM)

Assumptions:

- Infinite sites assumption:
  a character changes state once
- Error-free data



1-1 Equivalent

n mutations

clones

m samples

S_1 S_2 S_3

$$\begin{pmatrix} 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Frequency Matrix ***F*** | Mixture Matrix ***U*** | Restricted PP Matrix ***B***

Rows of ***U*** are proportions:

$$u_{pj} \geq 0 \text{ and } \sum_j u_{pj} \leq 1$$

Perfect Phylogeny Theorem
[Estabrook, 1971]
[Gusfield, 1991]

Perfect Phylogeny Mixture: [El-Kebir*, Oesper* et al., 2015]
Given ***F***, find ***U*** and ***B*** such that $F = UB$

Complexity of #PPM

Question 1: Can we determine the number of solutions?

Question 2: Can sample solutions uniformly at random?

#PPM: Given F, count the number of pairs (U, B) composed of mixture matrix U and perfect phylogeny matrix B such that $F = UB$

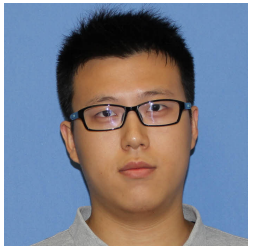
#P is the complexity class of counting problems whose decision problems are in NP

Every problem in #P can be reduced in polynomial time to any problem in #P-complete, preserving cardinalities

Theorem: #PPM is #P-complete

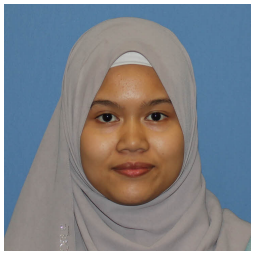
Theorem: There is no FPRAS for #PPM

Theorem: There is no FPAUS for PPM

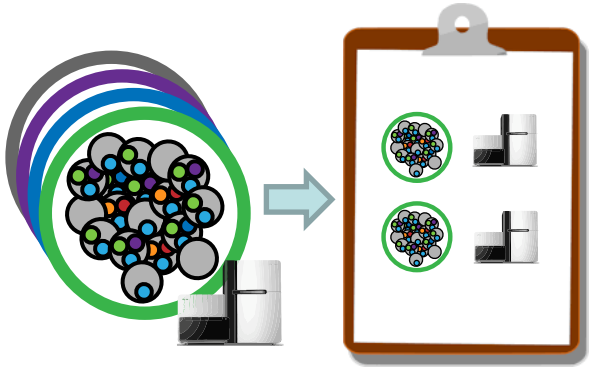


Yuanyuan Qi

Experimental Sequencing Study Design



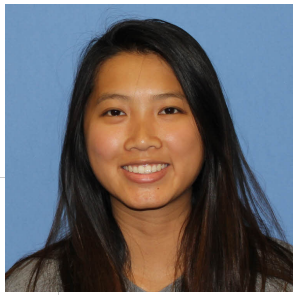
Nuraini Aguse



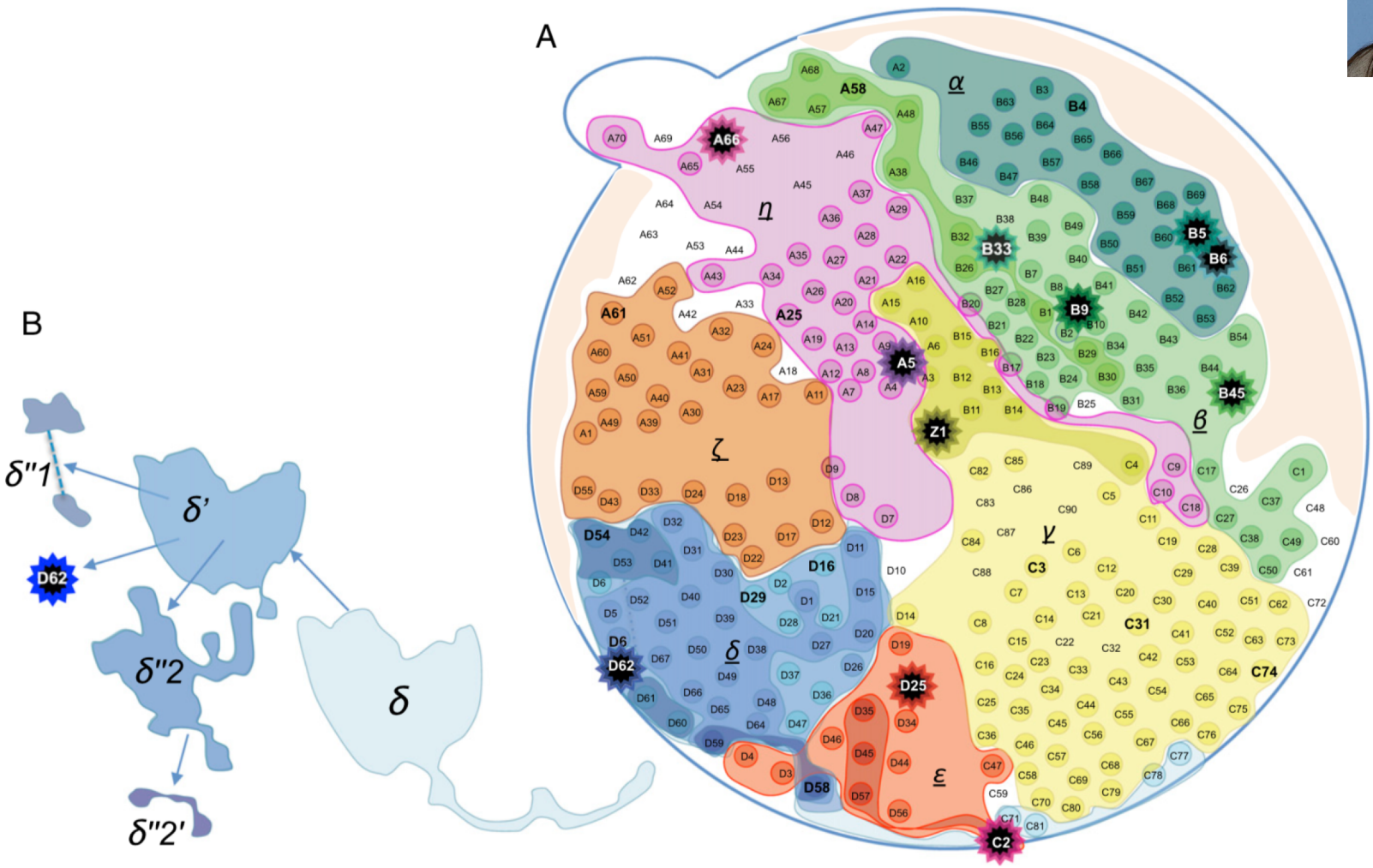
Problem Statement:

Develop a computational method to suggest follow-up sequencing experiments given preliminary sequencing data with the aim of reducing ambiguity.

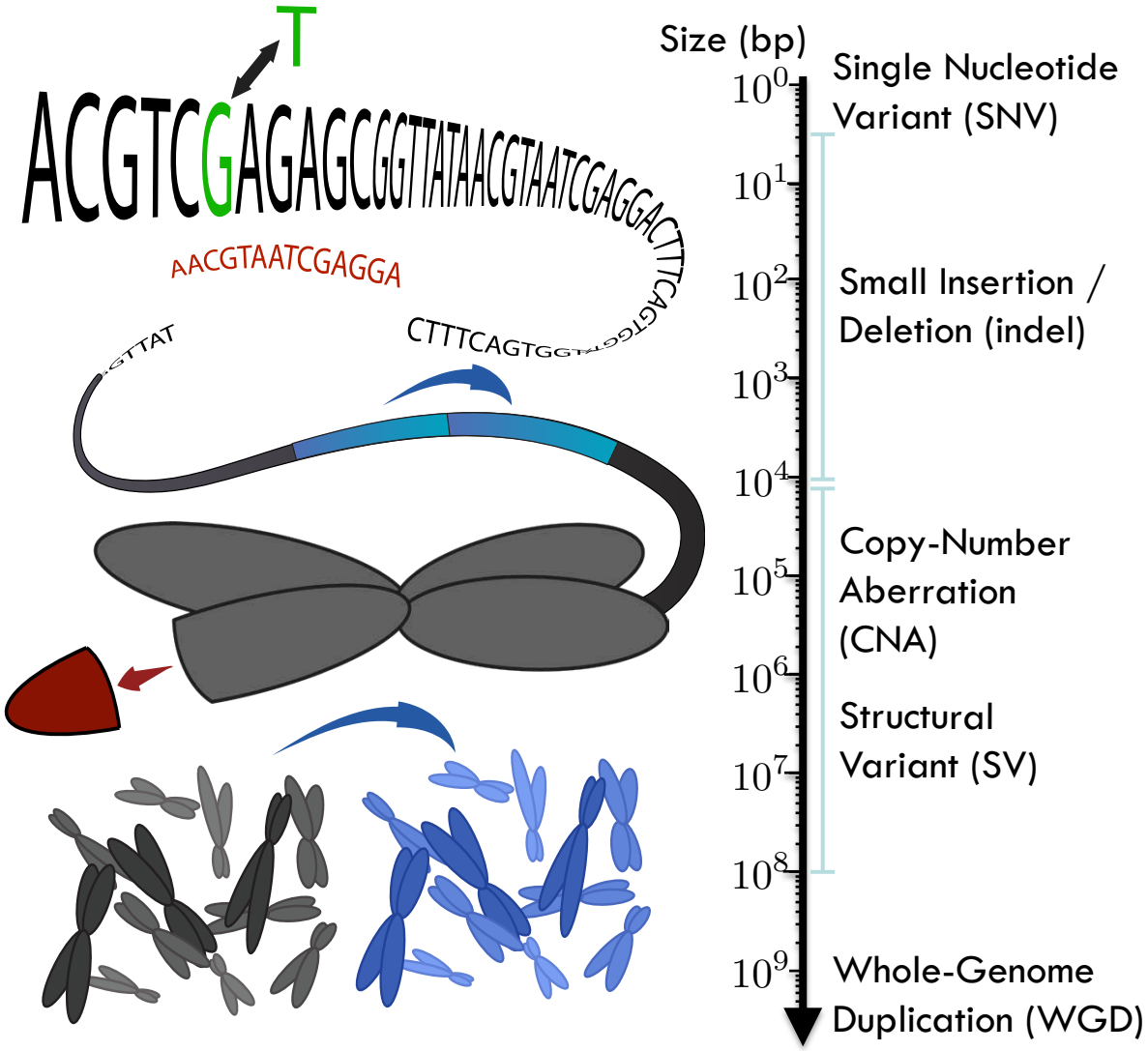
Visualizing Tumor Structure



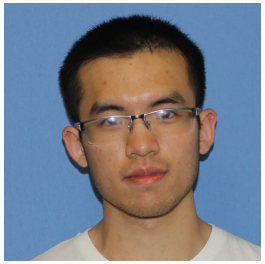
Jiaqi Wu



Somatic Mutations in Cancer



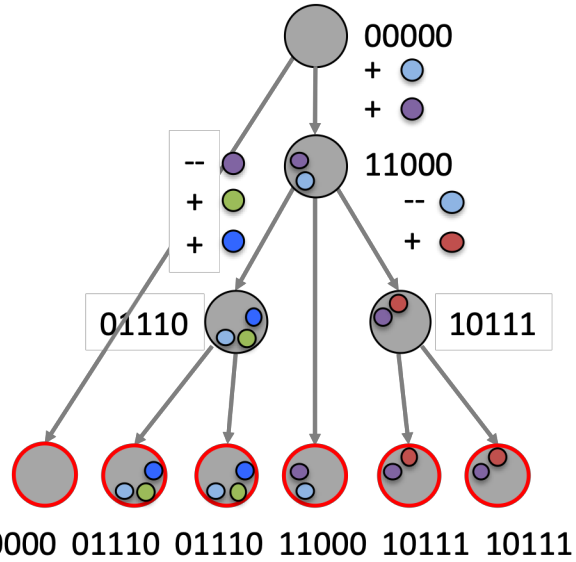
Combinatorial Characterization



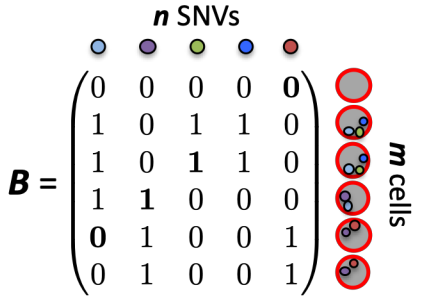
Shunping Xie

Definition 1. A *k-Dollo phylogeny* T is a rooted, node-labeled tree subject to the following conditions.

1. Each node v of T is labeled by a vector $\mathbf{b}_v \in \{0,1\}^n$.
2. The root r of T is labeled by vector $\mathbf{b}_r = [0,\ldots,0]^T$.
3. For each character $c \in [n]$, there is exactly one *gain edge* (v,w) in T such that $b_{v,c} = 0$ and $b_{w,c} = 1$.
4. For each character $c \in [n]$, there are at most k *loss edges* (v,w) in T such that $b_{v,c} = 1$ and $b_{w,c} = 0$.



k-Dollo Phylogeny problem (k-DP). Given a binary matrix $B \in \{0,1\}^{m\times n}$ and parameter $k \in \mathbb{N}$, determine whether there exists a k-Dollo phylogeny for B, and if so construct one.



Advising style

- Try to encourage project ownership
- Very hands-on when close to deadline
  - Happy to code/write together

http://www.el-kebir.net