Bioinformatics and Computational Biology

Professor Mohammed El-Kebir

Computer Science • University of Illinois at Urbana-Champaign

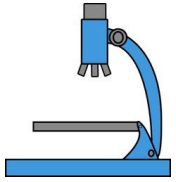
What is Computational Biology/Bioinformatics?

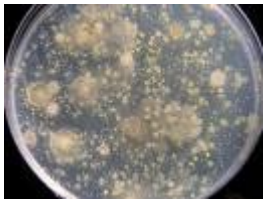
Computational biology and **bioinformatics** is an interdisciplinary field that develops and applies **computational methods** to analyze large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or **discover new biology**.

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

Technology and Bioinformatics are Transforming Biology

Until late 20th Century





Hypothesis Generation and Validation

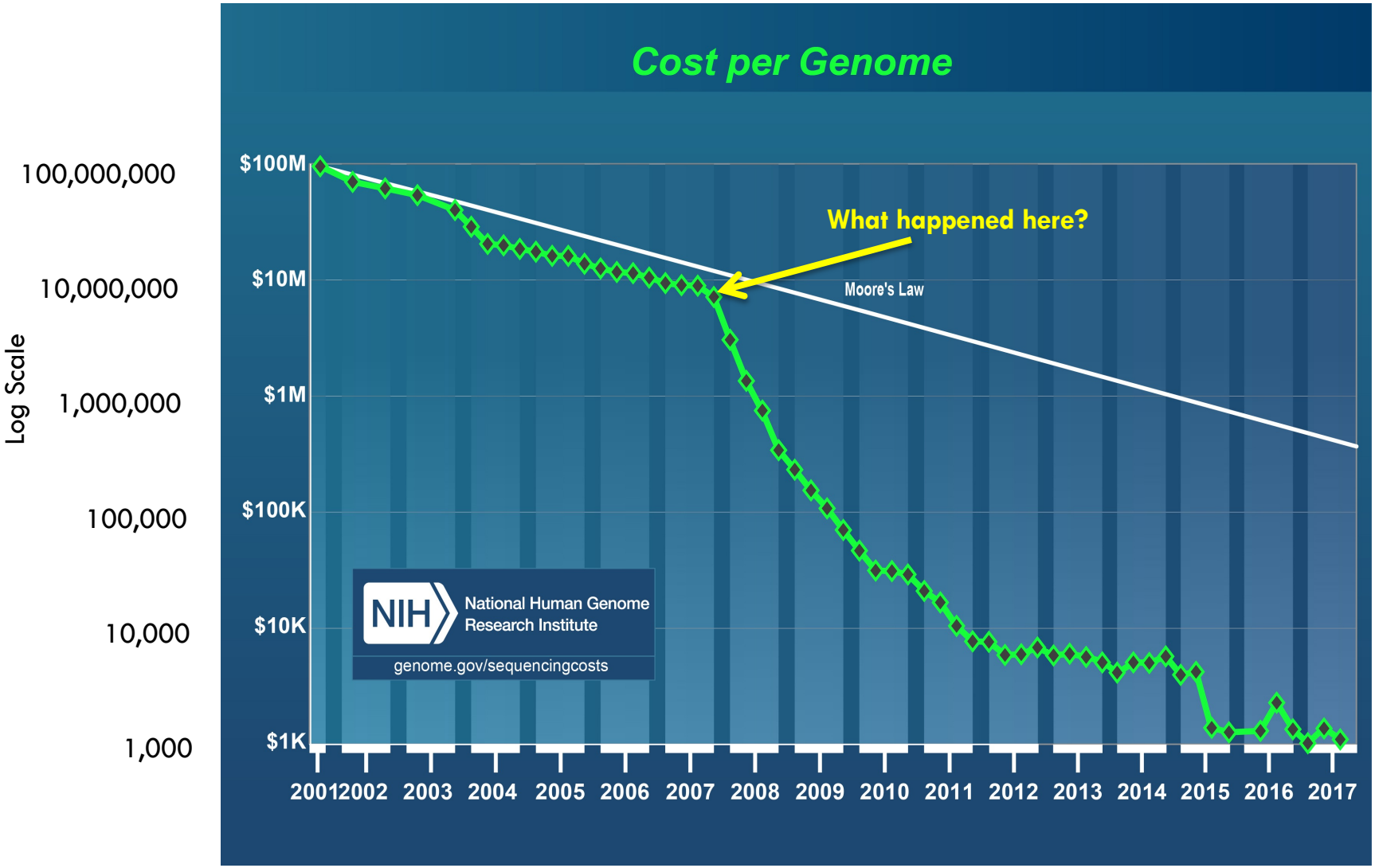
21th Century and Beyond

Algorithms

Hypothesis Generation and Validation

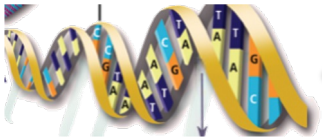
High throughput technologies

A Deluge of Data



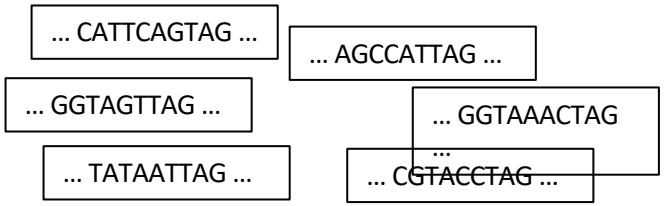
Question: What does it mean that we can sequence a genome?

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."

Coursework for Bioinformatics Research

The usual computer science stuff, but especially

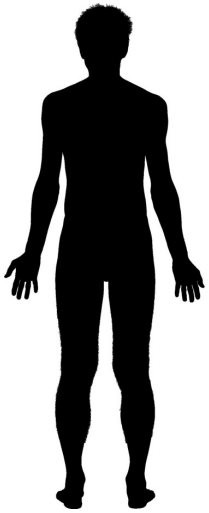
- CS 125 (programming)
- CS 173 (abstract thinking)
- CS 225 (data structures)
- CS 374 (algorithms and models for computation)

A bit of statistics is helpful (e.g., CS 361)

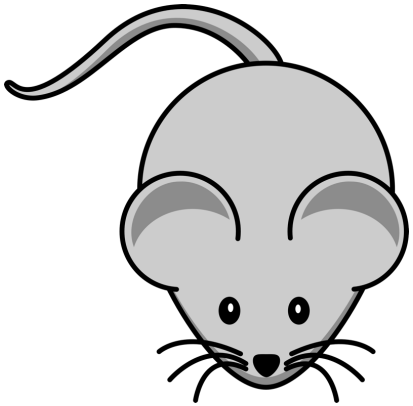
CS 466: Introduction to Bioinformatics! Good if you know some biology, but you can take CS 466, and learn it there!

Course Topic #1: Sequence Alignment

Question: How do we compare two genes/genomes?



vs.



Human Genome:

...ACTCGACTGAGAGGATTTCGAGCATGA...

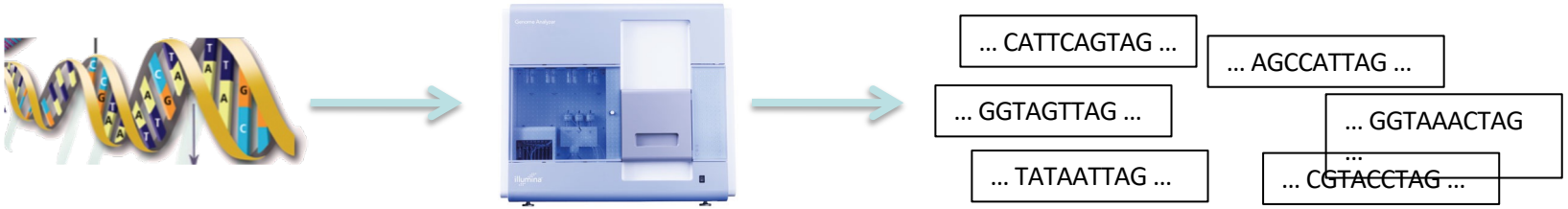
≈3.2 x 10^9 bp

Mouse Genome:

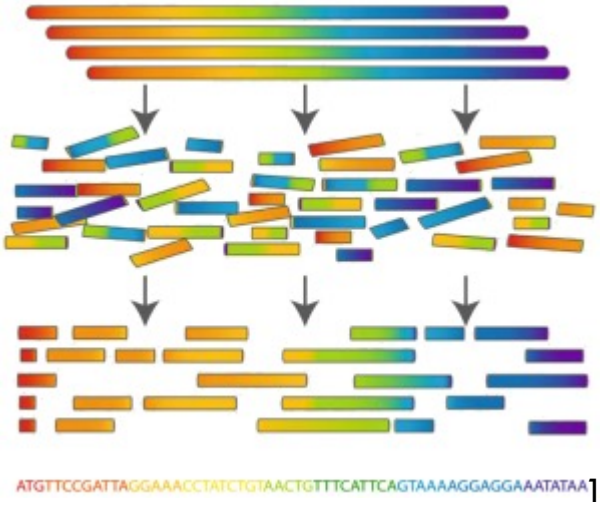
...ACTCAACTGAGATTCGAGCTTCAATGA...

≈2.8 x 10^9 bp

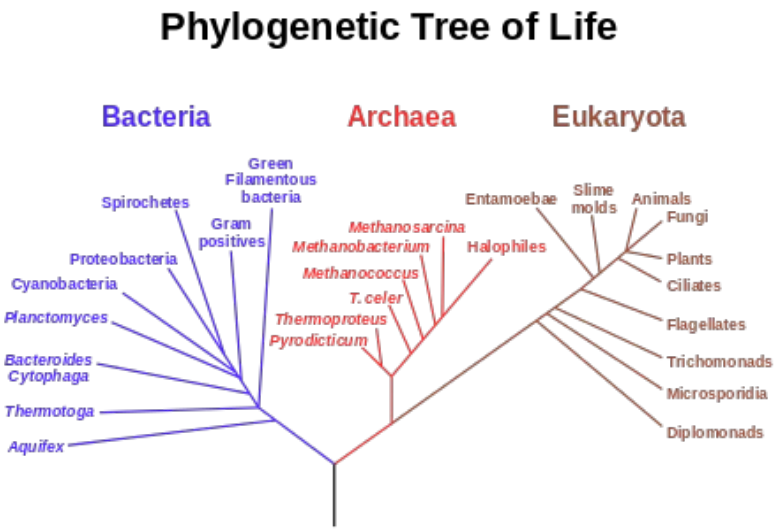
Course Topic #2: Genome Assembly



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



Course Topic #3: Phylogenetics

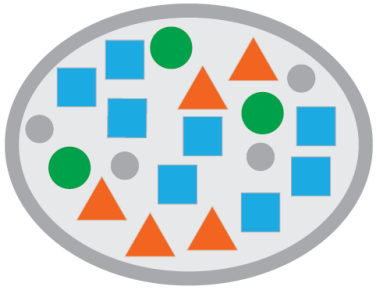


https://en.wikipedia.org/wiki/Phylogenetic_tree

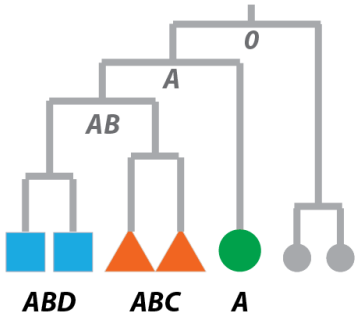
Question: Can we reconstruct the evolutionary history of different species?

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

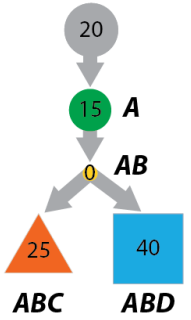
Poly-clonal tumor at sampling



Classical phylogenetic tree

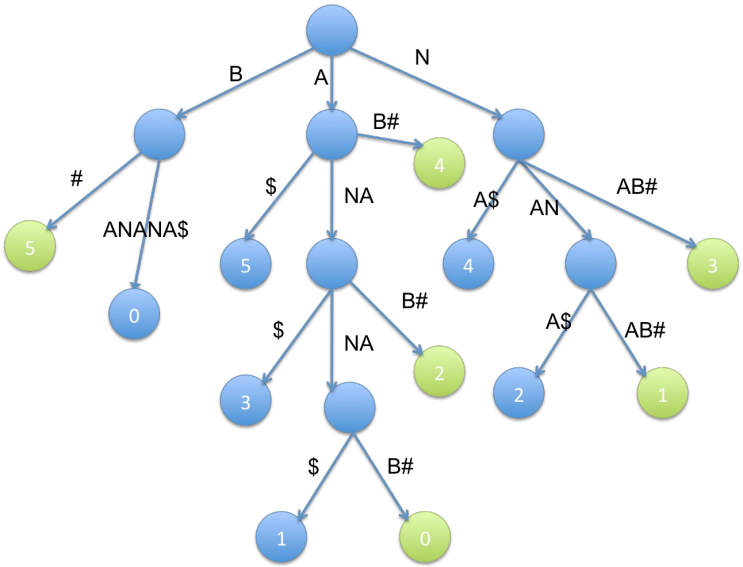


Clonal evolution tree



https://scientificbsides.wordpress.com/2014/06/09/inferring-tumour-evolution-2-comparison-to-classical-phylogenetics/

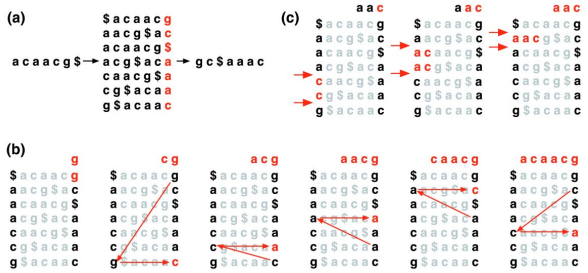
Course Topic #4: Pattern Matching



Suffix Trees

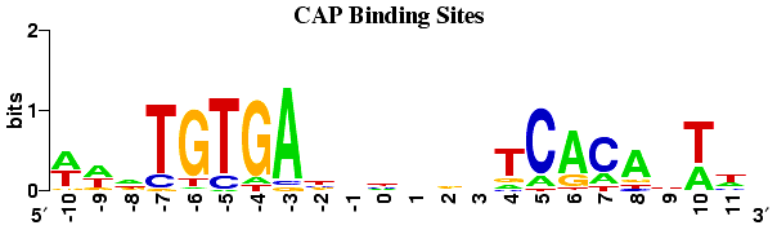
Question: How do we start to make sense of all these sequences?

Burrows Wheeler Transform



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

Motif Finding



Course Topics

1. Sequence alignment
   'How do we compare two genes/genomes?'

2. Genome assembly
   'How do we put all the pieces back together?'

3. Phylogenetics
   'What is the evolutionary history of different sequences?'

4. Pattern matching
   'How do we start to make sense out of all these sequences?'

Course Topics

1. Sequence alignment
   Dynamic programming: edit distance

2. Genome assembly
   Graphs: de Bruijn graph, Eulerian and Hamiltonian paths

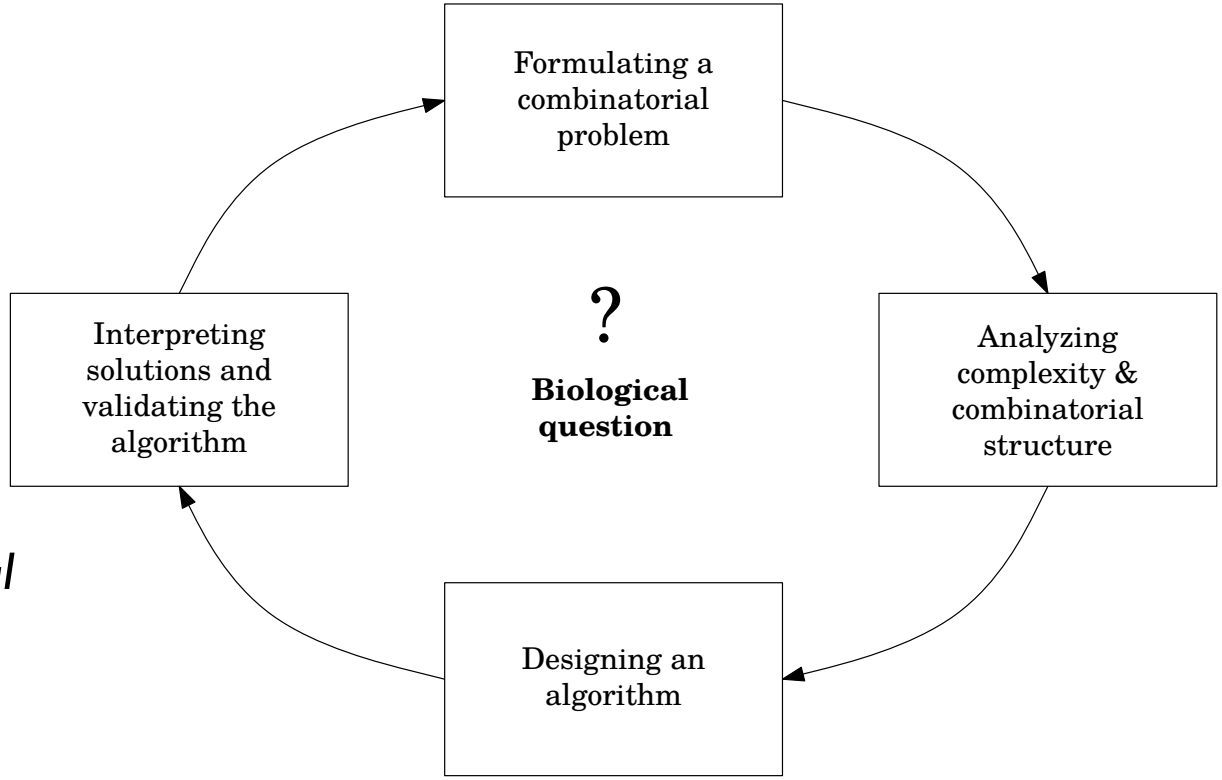
3. Phylogenetics
   Trees and distances: distance matrices, neighbor joining, hierarchical clustering.
   Phylogenies: Sankoff/Fitch algorithms, perfect phylogeny and compatibility

4. Pattern matching
   Suffix trees/arrays. Burrows-Wheeler transform, Hidden Markov Models (HMMs)

Research Statement & Approach

Lab focus:

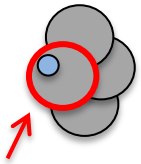
Application of combinatorial optimization techniques to answers questions and solve problems in biology.



Problem != Algorithm

Tumorigenesis: Cell Mutation

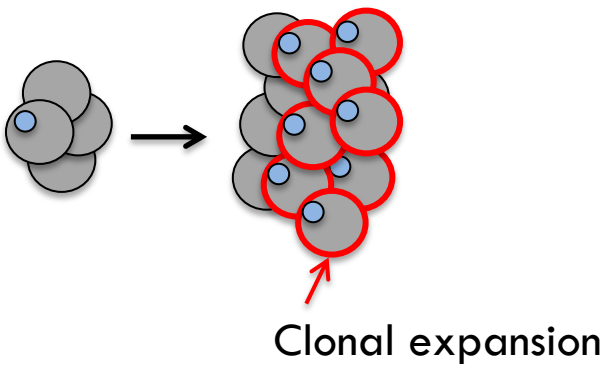
Clonal Evolution Theory of Cancer
[Nowell, 1976]



Founder
tumor cell
with somatic mutation:
(e.g. BRAF V600E)

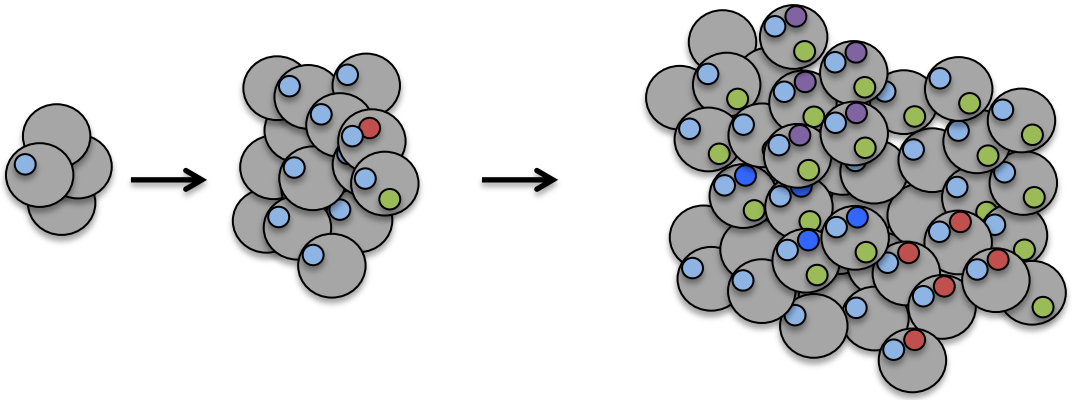
Tumorigenesis: Cell Mutation

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Tumorigenesis: Cell Mutation & Division

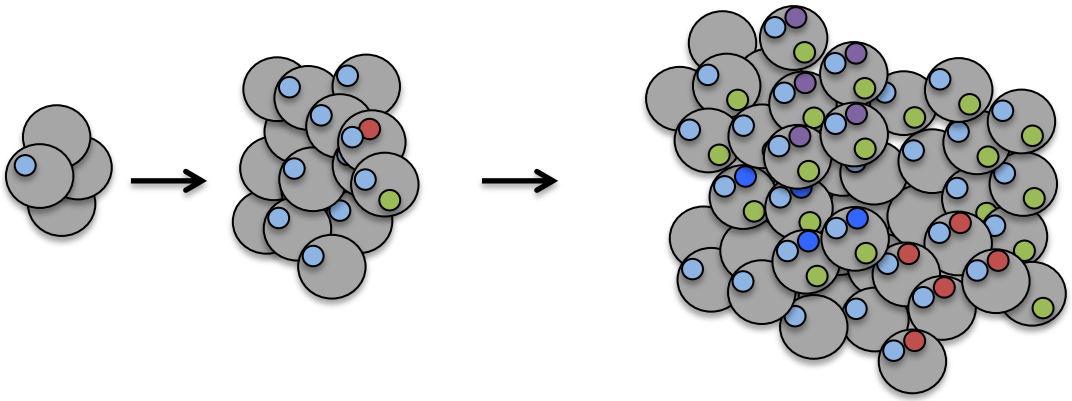
Clonal Evolution Theory of Cancer
[Nowell, 1976]



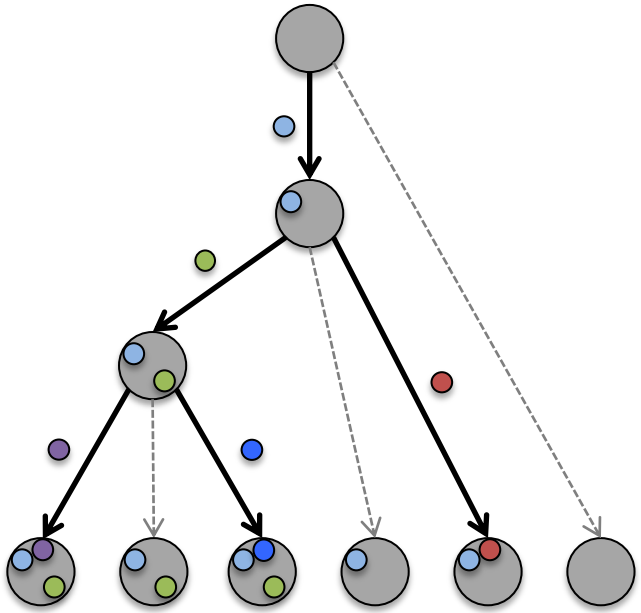
Intra-Tumor Heterogeneity

Tumorigenesis: Cell Mutation & Division

Clonal Evolution Theory of Cancer
[Nowell, 1976]



Intra-Tumor Heterogeneity

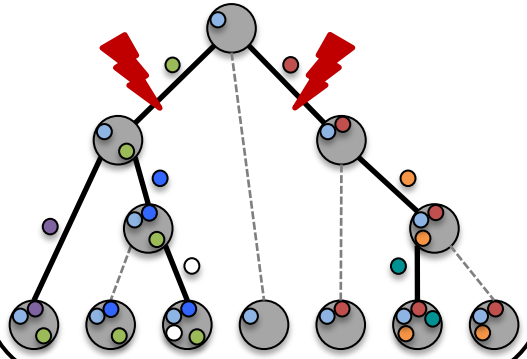


Phylogenetic Tree ***T***

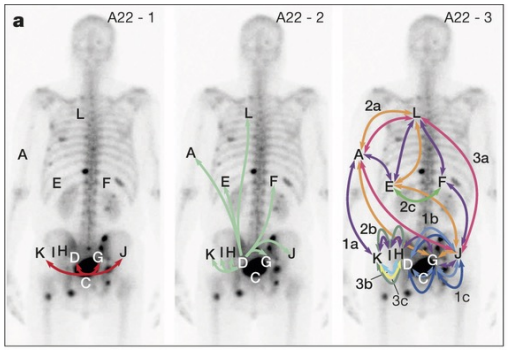
Question: Why are tumor phylogenies important?

Phylogenies are Key to Understanding Cancer

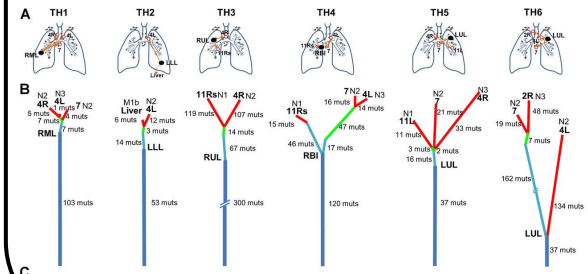
Identify targets for treatment



Understand metastatic development

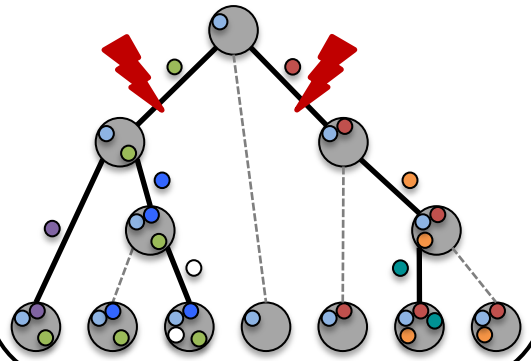


Recognize common patterns of tumor evolution across patients

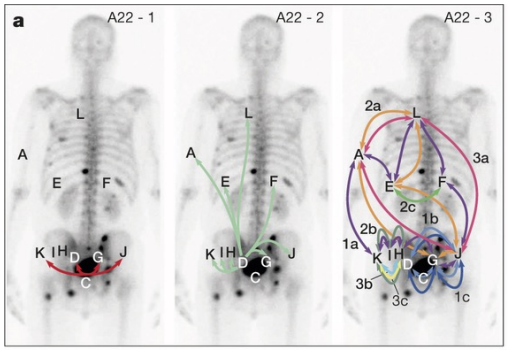


Phylogenies are Key to Understanding Cancer

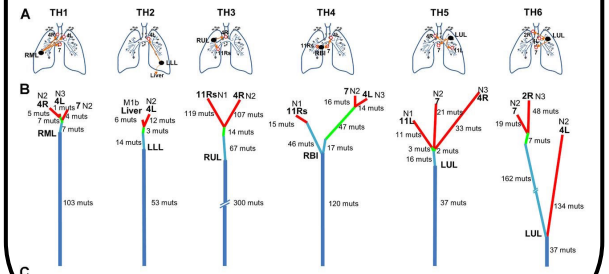
Identify targets for treatment



Understand metastatic development



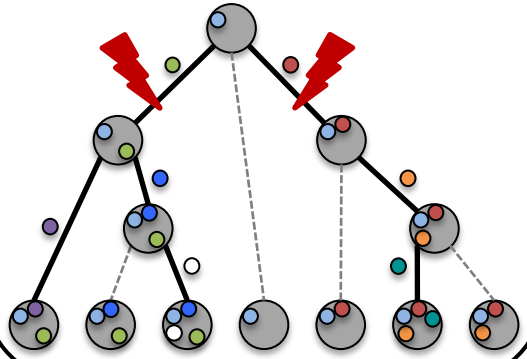
Recognize common patterns of tumor evolution across patients



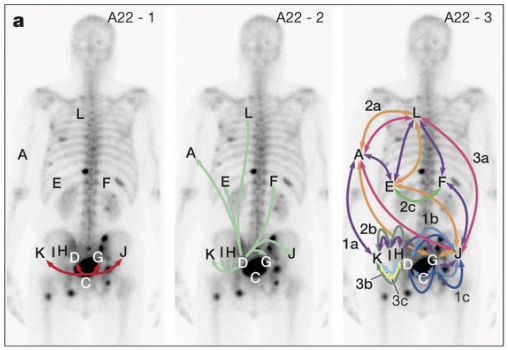
These downstream analyses **critically rely** on accurate tumor phylogeny inference

Phylogenies are Key to Understanding Cancer

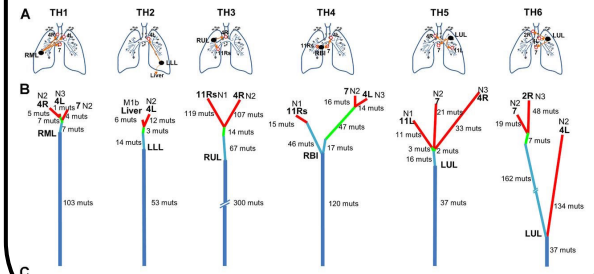
Identify targets for treatment



Understand metastatic development



Recognize common patterns of tumor evolution across patients

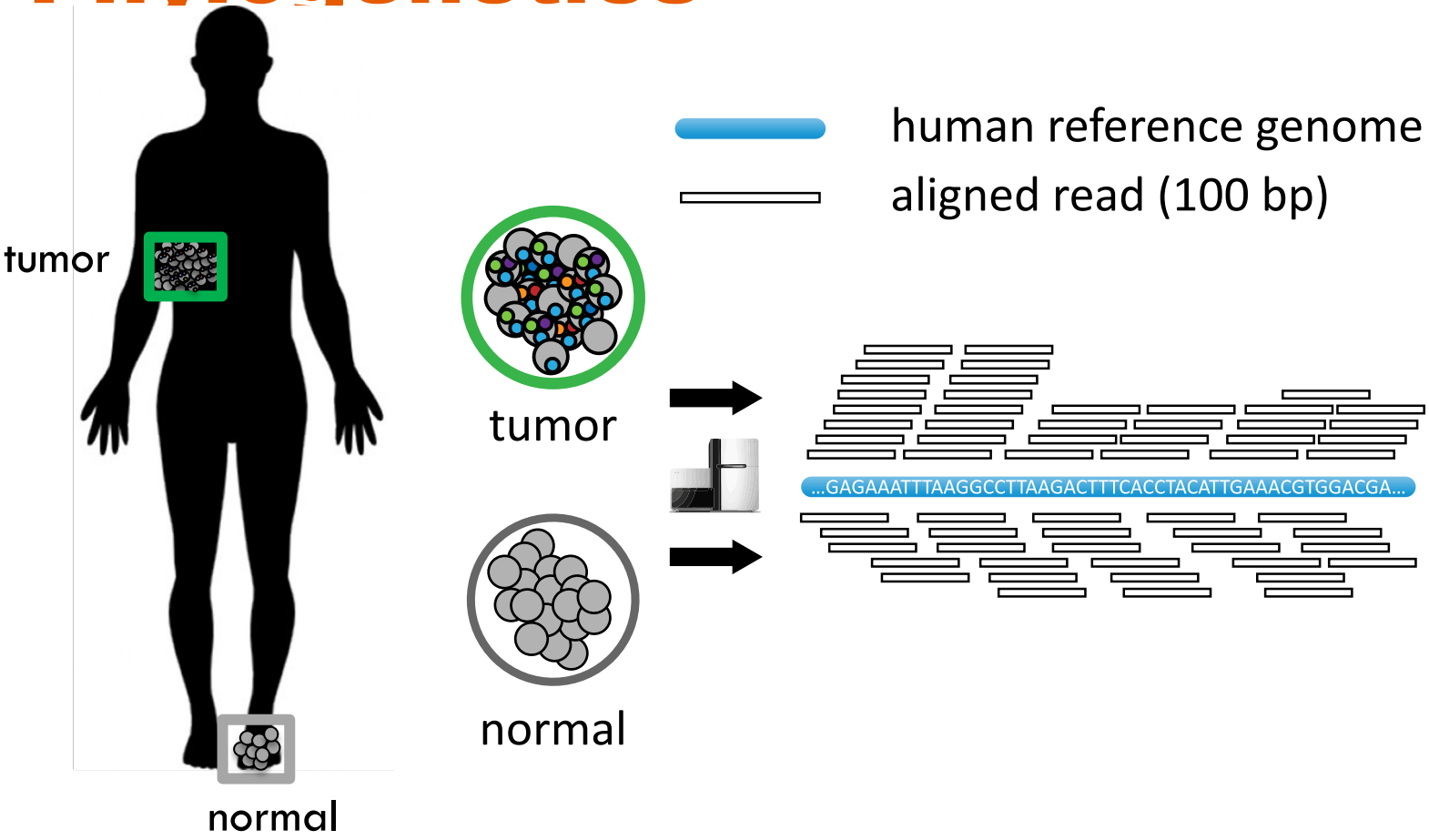


These downstream analyses **critically rely** on accurate tumor phylogeny inference

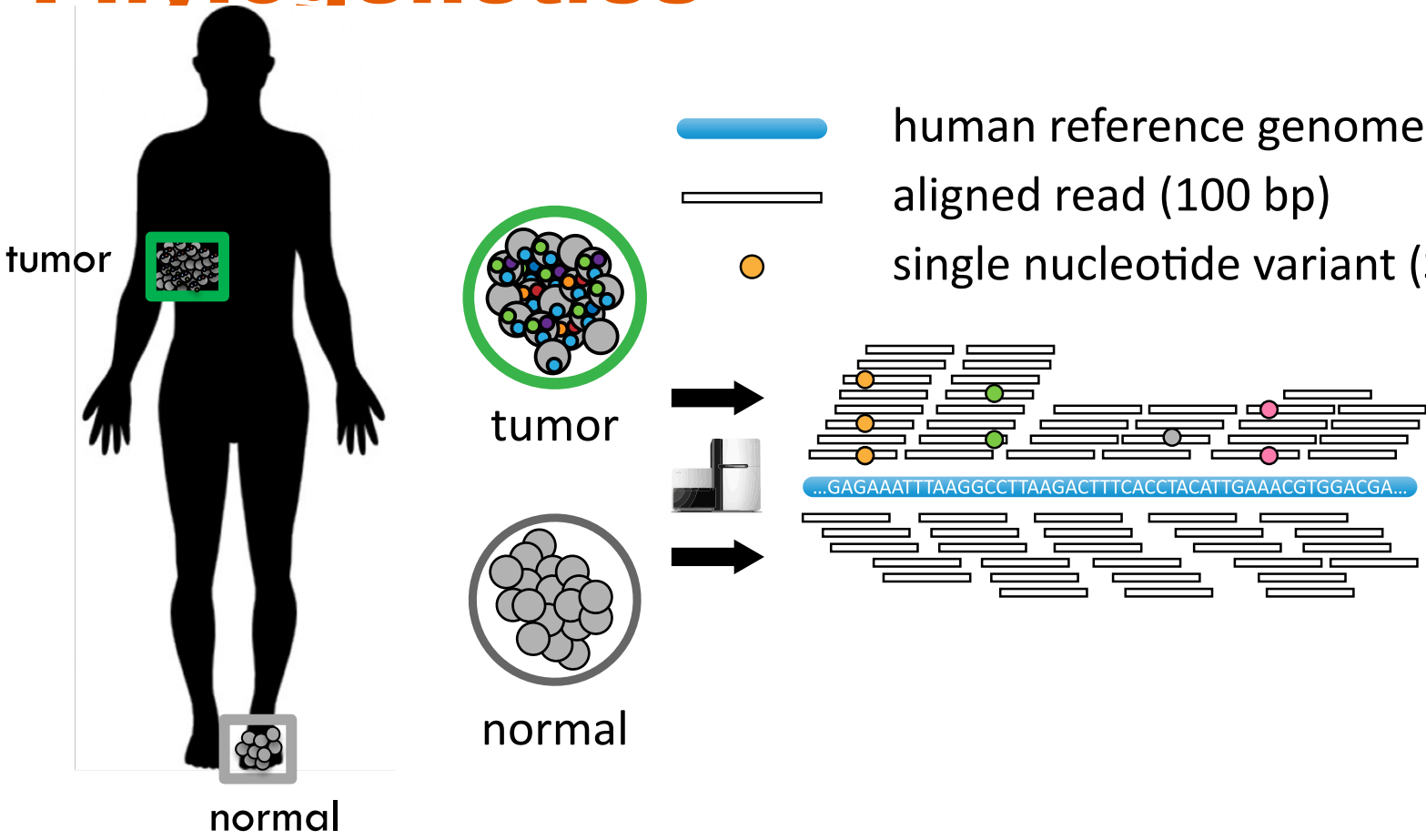
Key challenge in phylogenetics:

Accurate phylogeny inference from data at present time

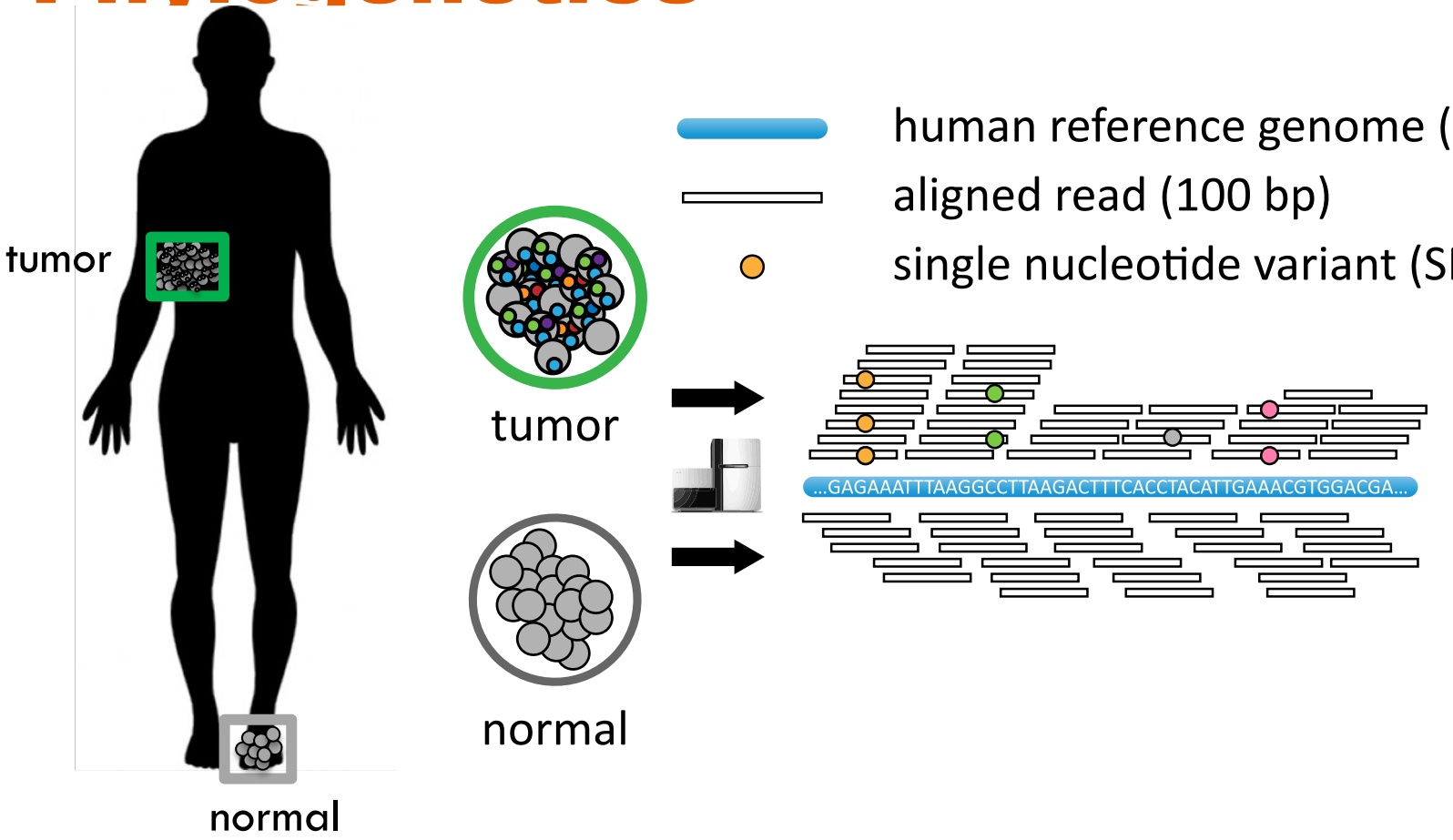
Additional Challenge in Cancer Phylogenetics



Additional Challenge in Cancer Phylogenetics



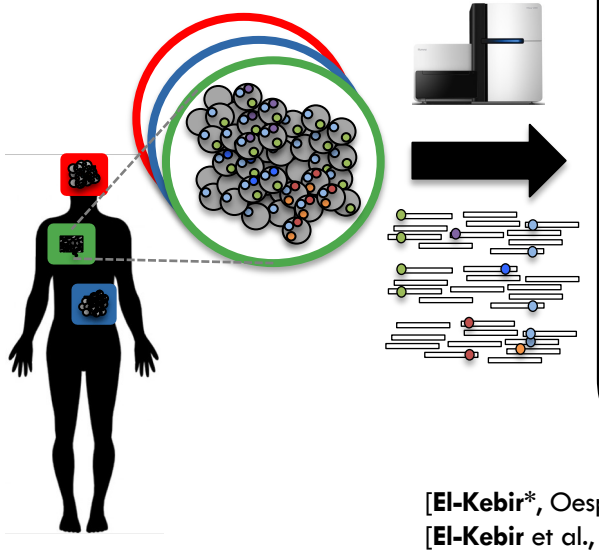
Additional Challenge in Cancer Phylogenetics



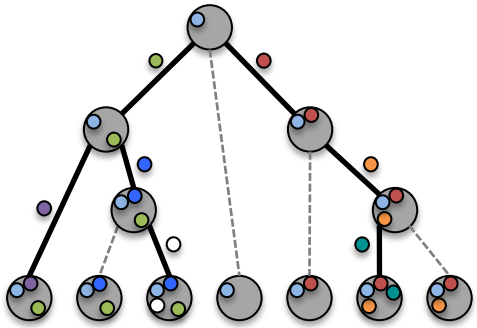
Additional challenge in cancer phylogenetics:

Phylogeny inference from **mixed bulk samples** at present time

Overview of Current Research



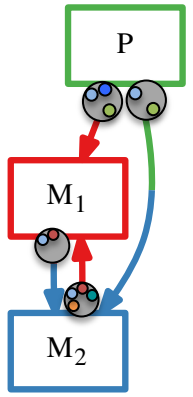
Cell division and mutation



Tumor Phylogeny Inference

[**El-Kebir***, Oesper* et al., *ISMB 2015/Bioinformatics*]
[**El-Kebir** et al., *RECOMB 2016*]; [**El-Kebir***, Satas* et al., *Cell Systems 2016*]
[**El-Kebir** et al., *WABI 2016*]; [Zaccaria*, **El-Kebir*** et al., *RECOMB 2017*]
[**El-Kebir**, *Bioinformatics/ECCB 2018*], [Pradhan and **El-Kebir**, *RECOMB-CG 2018*]

Cell migration

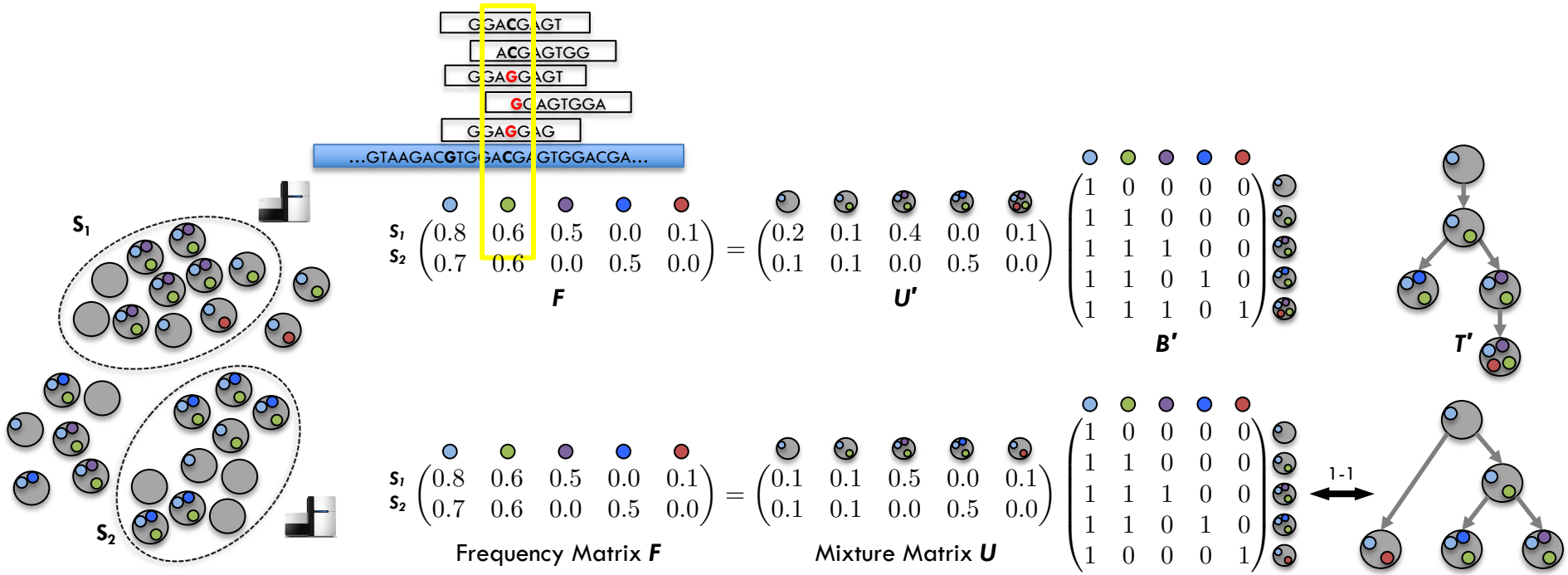


Tumor Migration Analysis

[**El-Kebir** et al., *Nature Genetics* 2018]
[**El-Kebir**, WABI 2018]

Thesis: Precise mathematical models are needed to describe the evolutionary process in cancer

Non-uniqueness of Solutions in Bulk DNA



Question 1: Can we determine the number of solutions?

Question 2: Can we sample solutions uniformly at random?

Question 3: Can we design follow-up experiments to reduce ambiguity?

Question 4: Inclusion of prior knowledge/additional data?

Quantifying Extent of Non-uniqueness

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 $F = UB$

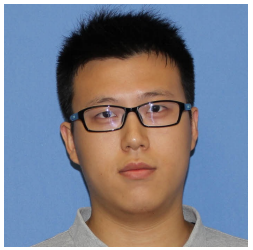
#P is the complexity class of counting problems whose decision problems are in NP

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

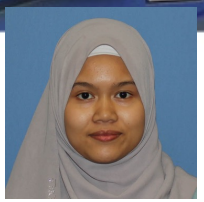


Yuanyuan Qi

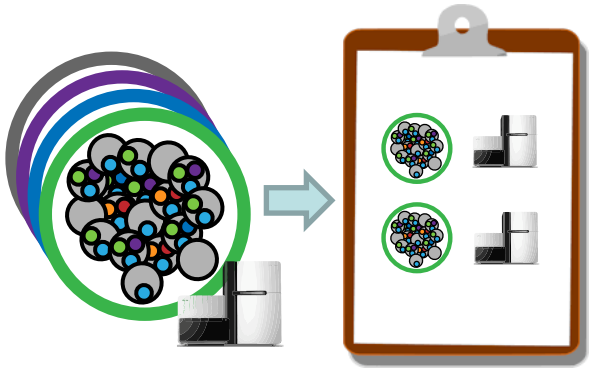
Sequencing Study Design

Question 3: Can we design follow-up experiments to reduce ambiguity? [Funded by CCBGM]

Leah Weber

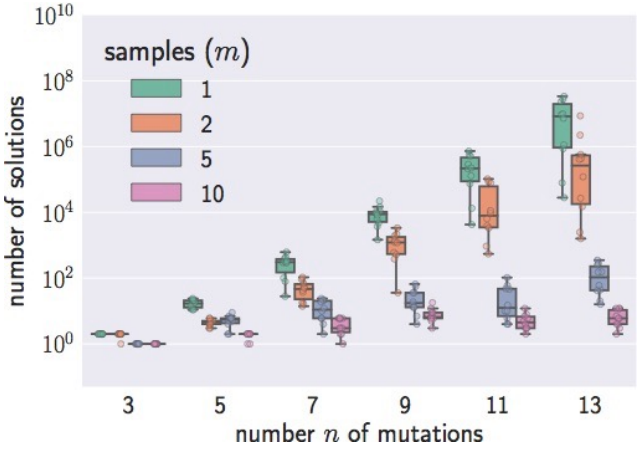


Nuraini Aguse

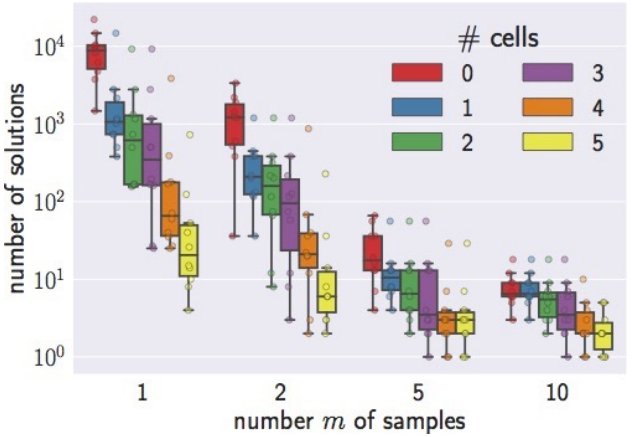


Problem Statement:

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

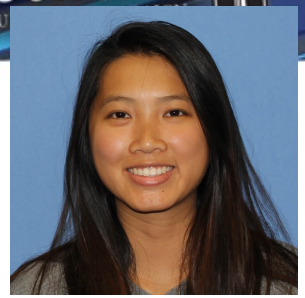


Effect of ***n*** and ***m***

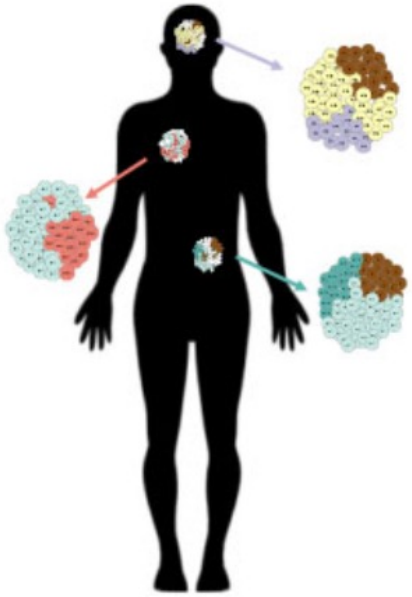


Effect of single-cell sequencing

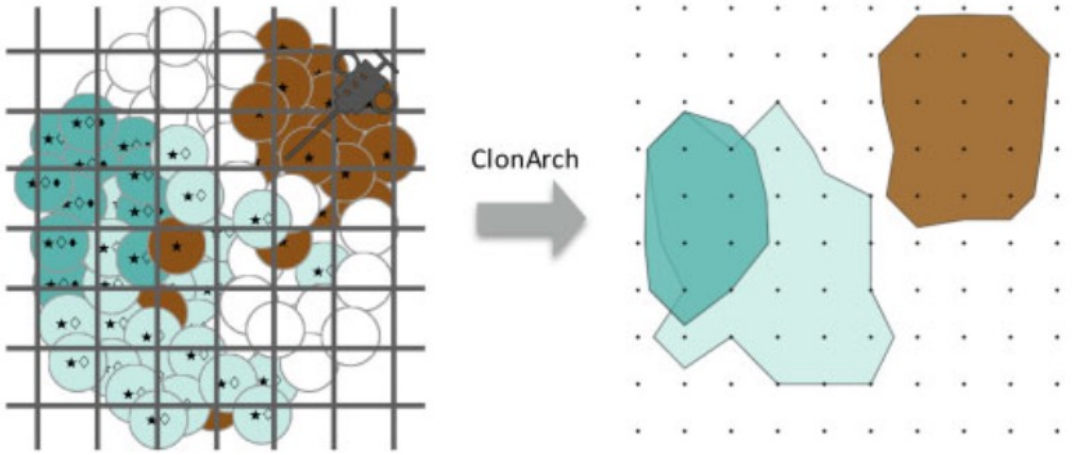
Visualizing Tumor Structure



Jiaqi Wu

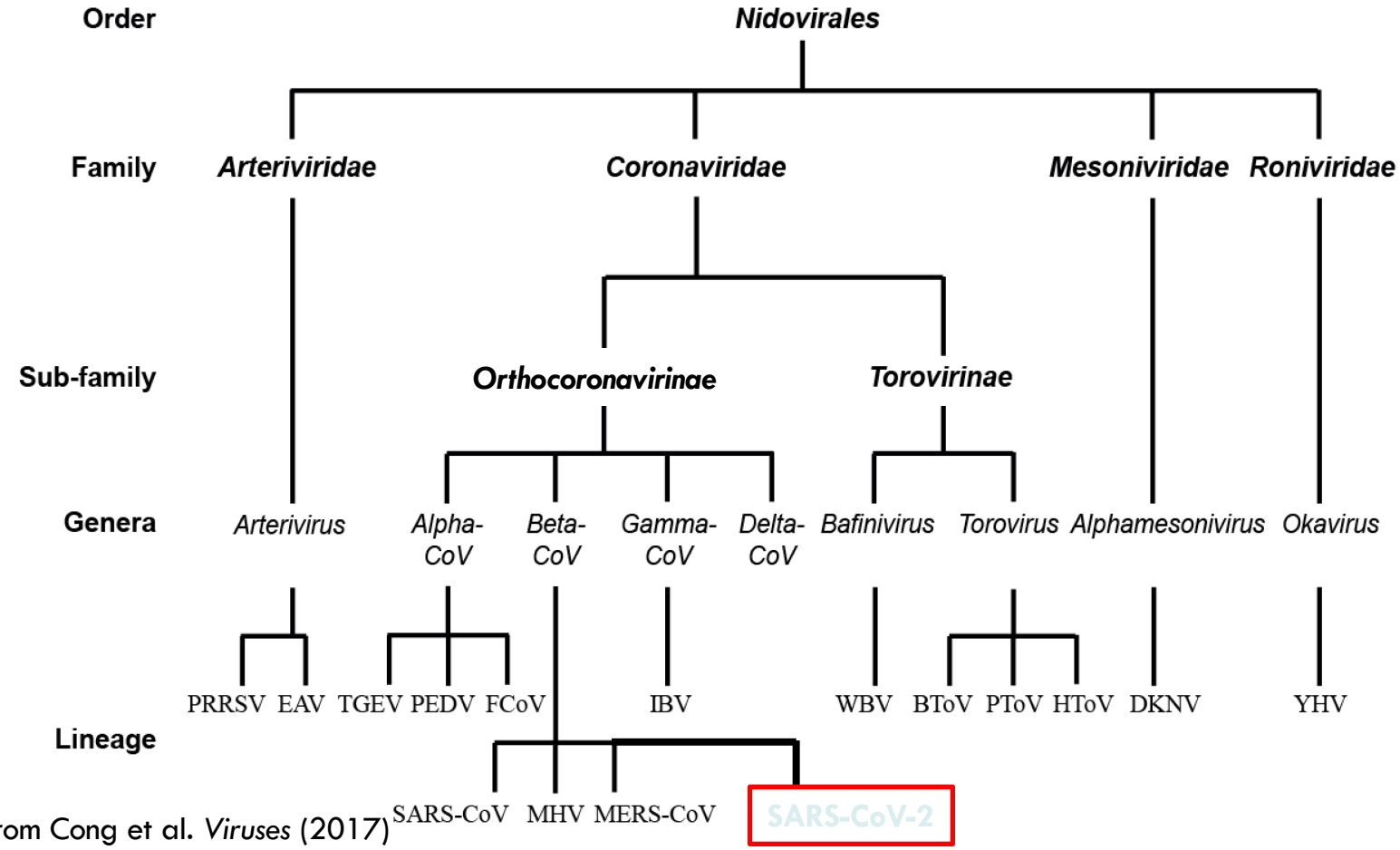


(d)



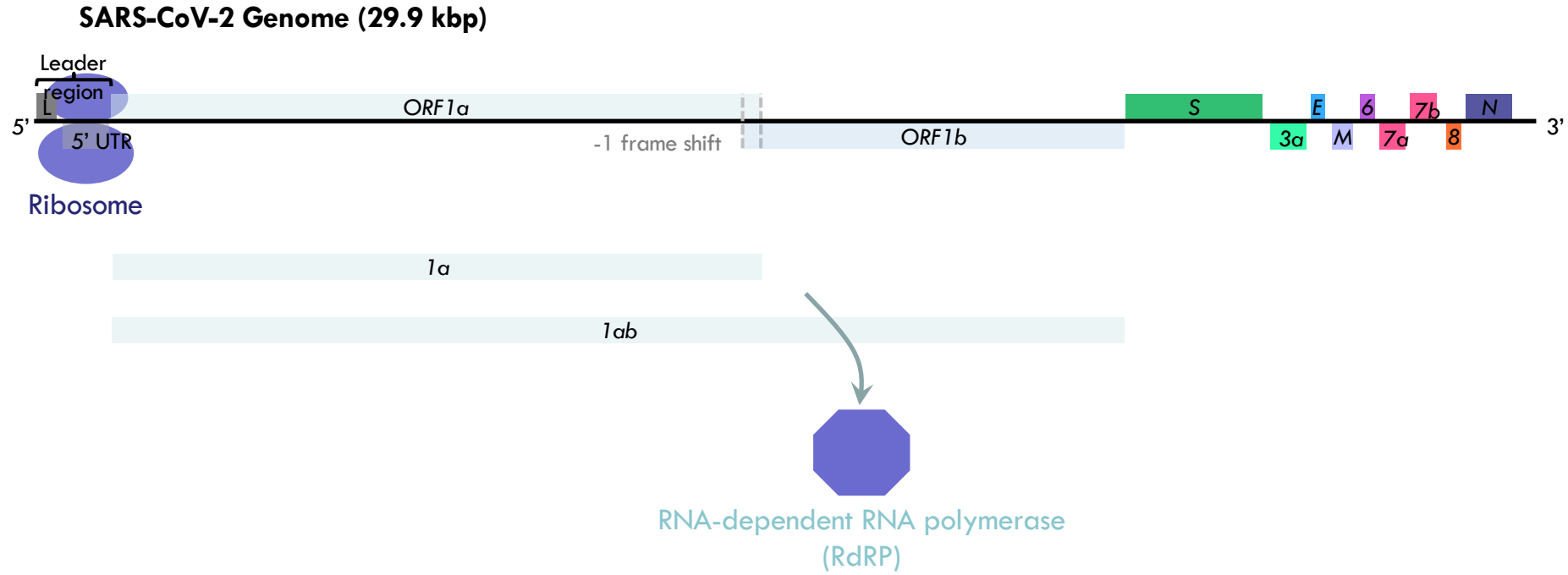
(e)

Coronaviruses – Background

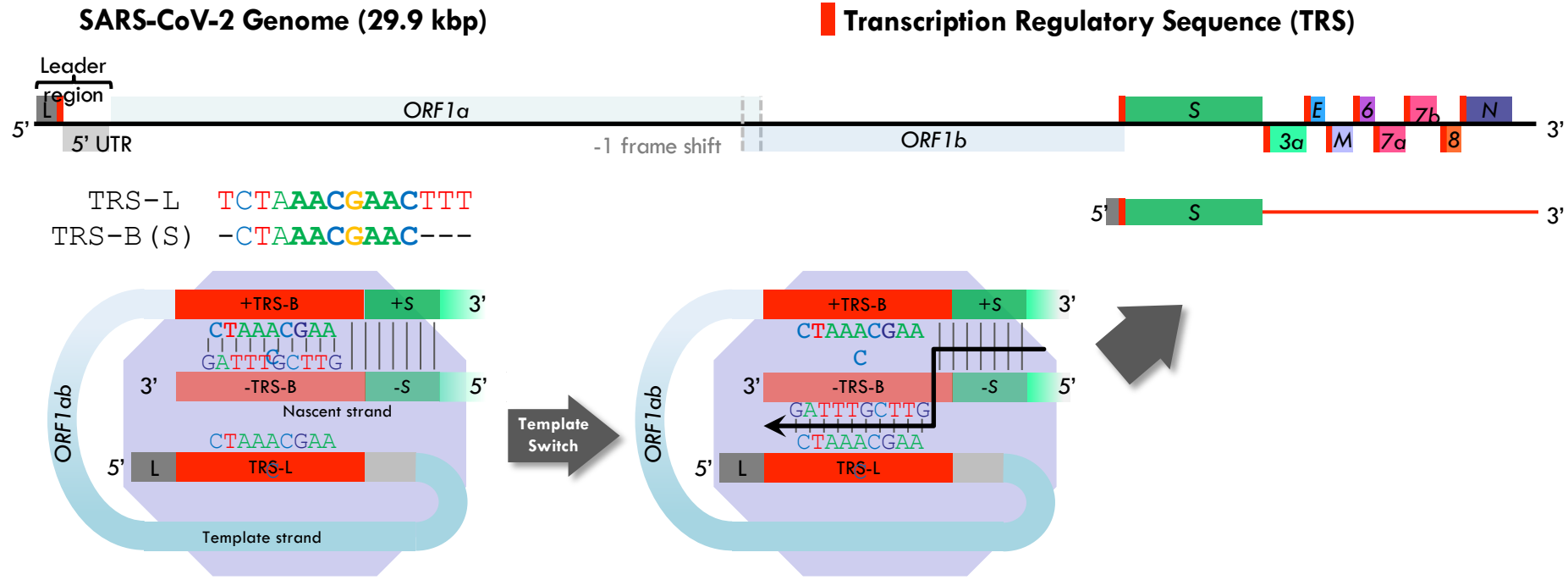


Adapted from Cong et al. *Viruses* (2017)

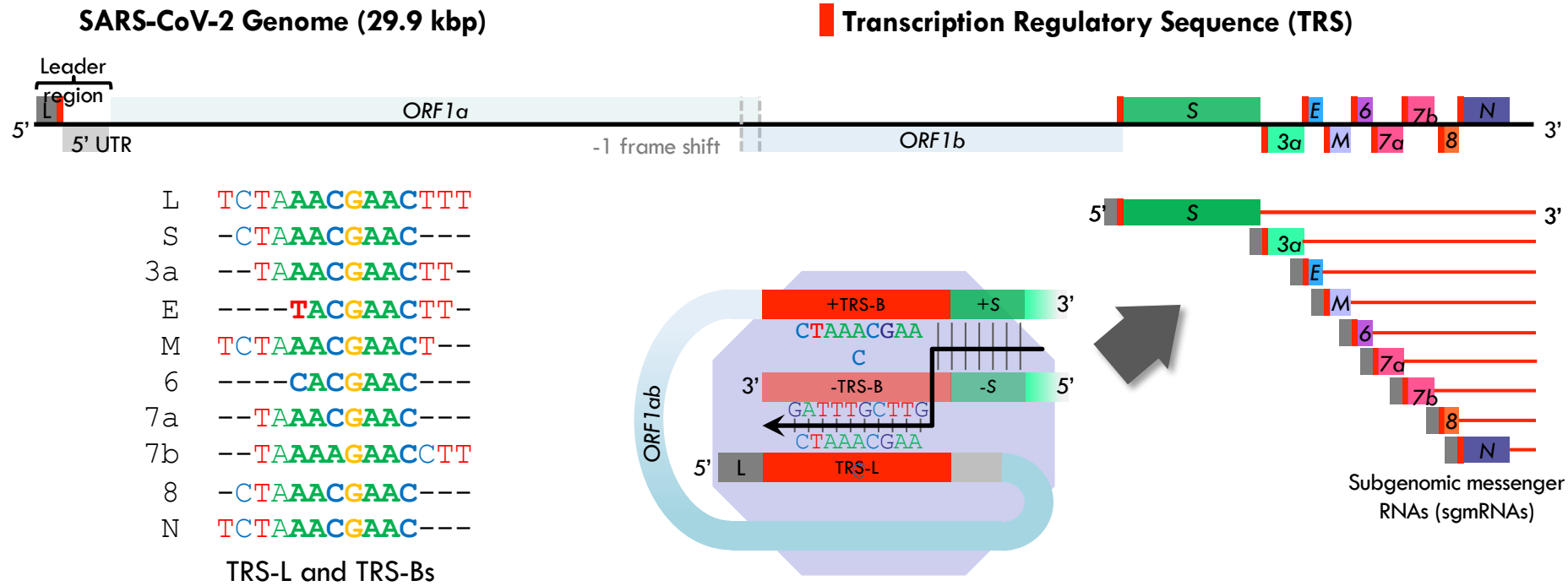
Coronaviruses – Background



Coronaviruses – Background



Coronaviruses – Background



Discontinuous transcription due to template switching of RdRP at transcription regulatory sequences

Sola et al., *Annual Review of Virology,* (2015)

Coronaviruses – Two Questions

Question 1. Can we identify TRS-L and TRS-Bs in **annotated** genomes?

Annotated genome $\mathbf{v}$ with genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$

TRS-L $\mathbf{a}_0$ and TRS-Bs $\mathbf{a}_1, \ldots, \mathbf{a}_n$

Question 2. Can we identify TRS-L, TRS-Bs and their **corresponding genes** in **unannotated** genomes?

Unannotated genome $\mathbf{v}$

TRS-L $\mathbf{a}_0$, TRS-Bs $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and associated genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$

Bioinformatics & Computational Biology Group



Top: Mohammed El-Kebir, Jian Peng, Tandy Warnow

Bottom: ChangXiang Zhai, Jiawei Han, and Olgica Milenkovic

And others!



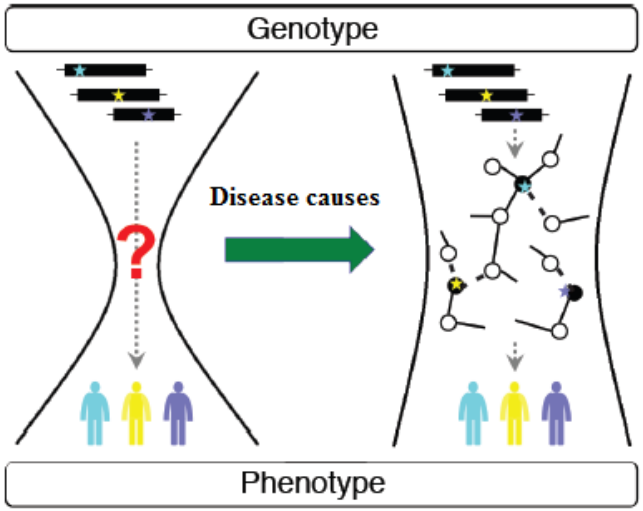
https://cs.illinois.edu/research/bioinformatics-and-computational-biology

Algorithmic Network Medicine

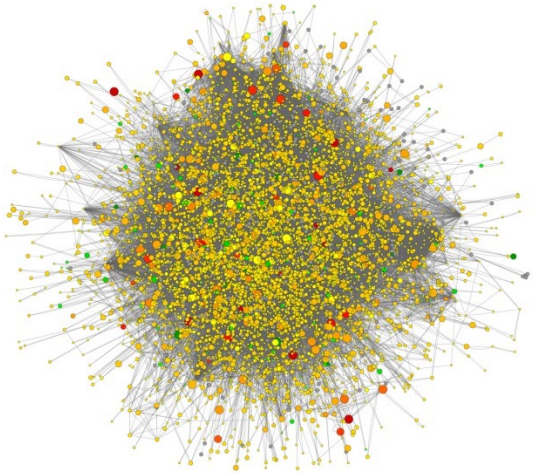
- Understanding human diseases from gene network and DNA
- Patient stratification for personalized medicine
- Acceleration of drug design

DNA data

Predictive Modeling



Gene Network

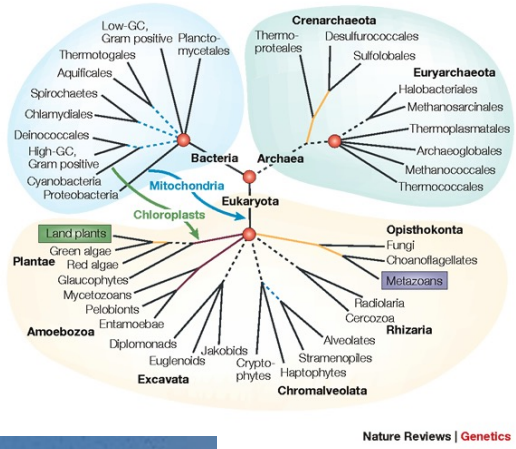


Precision Treatment

http://jianpeng.web.engr.illinois.edu

CS @ ILLINOIS

Tandy Warnow
The Tree of Life: *Multiple Challenges*

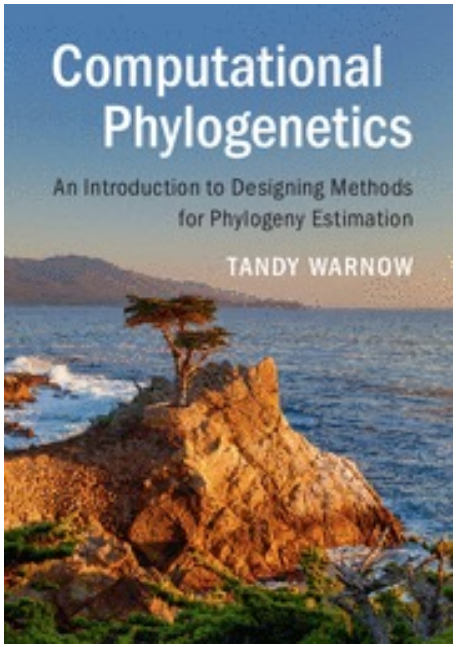


Nature Reviews | Genetics

Large datasets:

- 100,000+ sequences
- 10,000+ genes

"BigData" complexity



Large-scale statistical phylogeny estimation

- Ultra-large multiple-sequence alignment
- Estimating species trees from incongruent gene trees
- Supertree estimation
- Genome rearrangement phylogeny
- Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

http://tandy.cs.illinois.edu