Elucidating tumor evolutionary patterns using high-depth molecular data

Caitlin Guccione
Dr. Hossien Khiabanian

How do we treat cancer?

What is cancer?

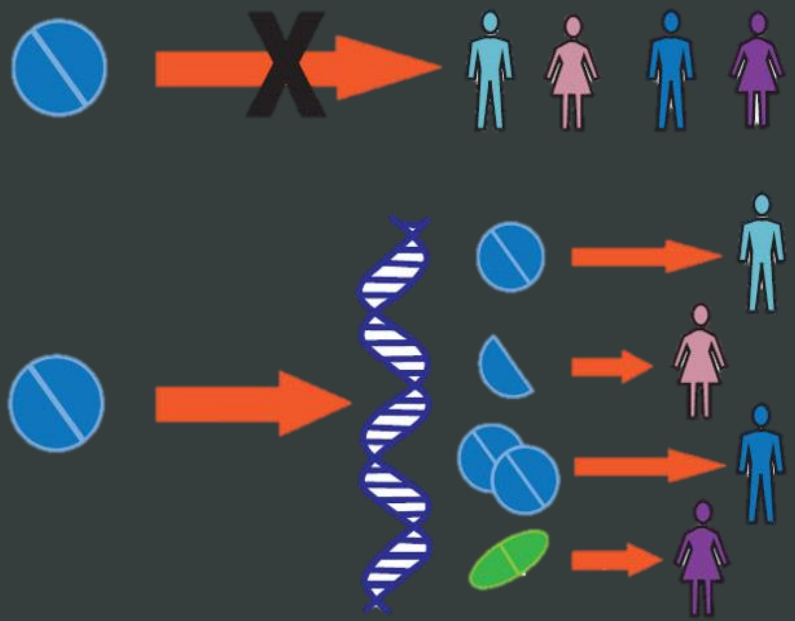
- ❖ A genetic disease caused by mutations in the DNA of particular cells

How is it currently being treated?

- ❖ Chemotherapy: Generally targets cancers cells while also killing normal cells in the process
- ❖ Targeted Therapy: Goes after a specific mutation that leads to cancer

How do we advance treatment?

- ❖ Sequencing cancer tumors to find particular mutations
- ❖ Finding medication to target more mutations
- ❖ Understand the evolution and growth



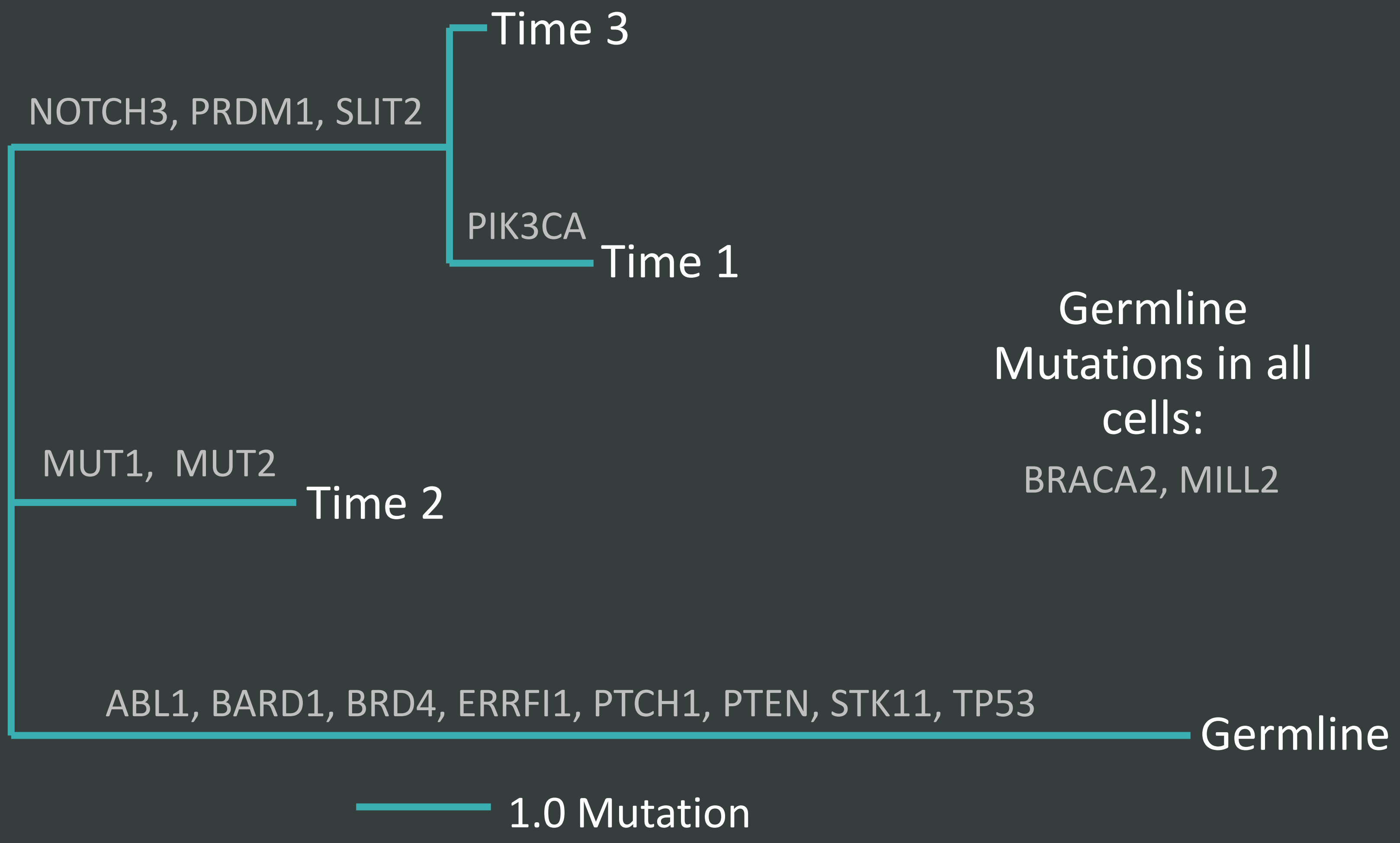
medicine.iupui.edu/IIPM

Basic Tree using Hamming Distance

Gene	Germline	Time 1	Time 2	Time 3
ABL1	0	1	1	1
BARD1	0	1	1	1
BRCA2	1	1	1	1
BRD4	0	1	1	1
ERRFI1	0	1	1	1
MLL2	1	1	1	1
Mut1	0	0	1	0
Mut2	0	0	1	0
NOTCH3	0	1	0	1
PIK3CA	0	1	0	0
PRDM1	0	1	0	1
PTCH1	0	1	1	1
PTEN	0	1	1	1
SLIT2	0	1	0	1
STK11	0	1	1	1
TP53	0	1	1	1

0 - Mutation is not present

1 - Mutation is present



Germline Mutations in all cells:

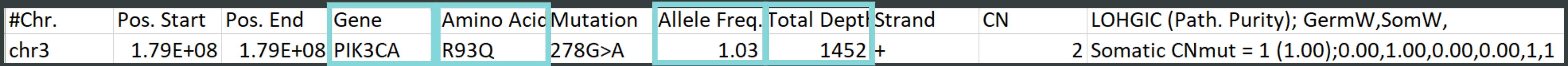
BRACA2, MILL2

What is sequencing data?

What's inside?

- ❖ Basic patient information ex. Gender, Age ect.
- ❖ Estimated purity of sample
    - ❖ Percentage of cancerous cells in sample
- ❖ ID number of sample

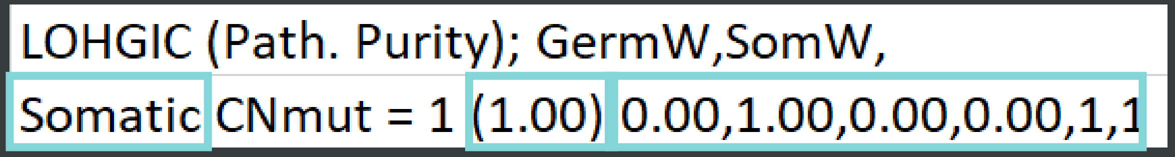
Mutations



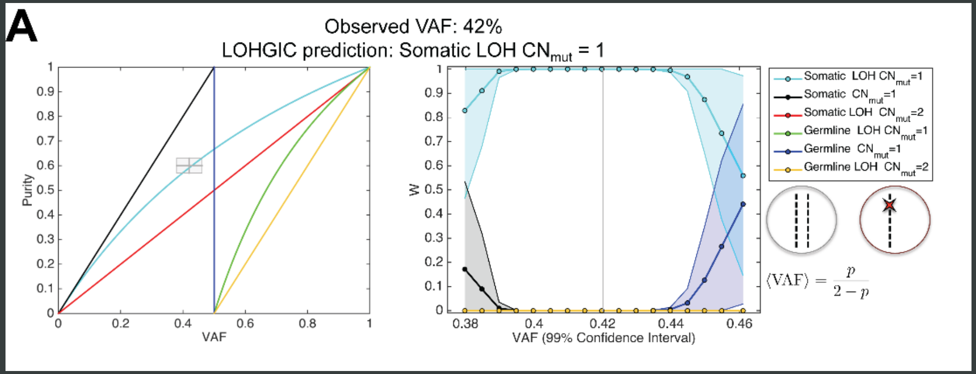
#Chr.	Pos. Start	Pos. End	Gene	Amino Acic	Mutation	Allele Freq.	Total Deptl	Strand	CN	LOHGIC (Path. Purity); GermW,SomW,
chr3	1.79E+08	1.79E+08	PIK3CA	R93Q	278G>A	1.03	1452	+	2	Somatic CNmut = 1 (1.00);0.00,1.00,0.00,0.00,1,1

- ❖ Gene and Amino Acid: location of the mutation
- ❖ Allele Frequency: total percent of alleles with mutation
- ❖ Depth: helps calculate the error on the allele frequency

LOHGIC's* Output

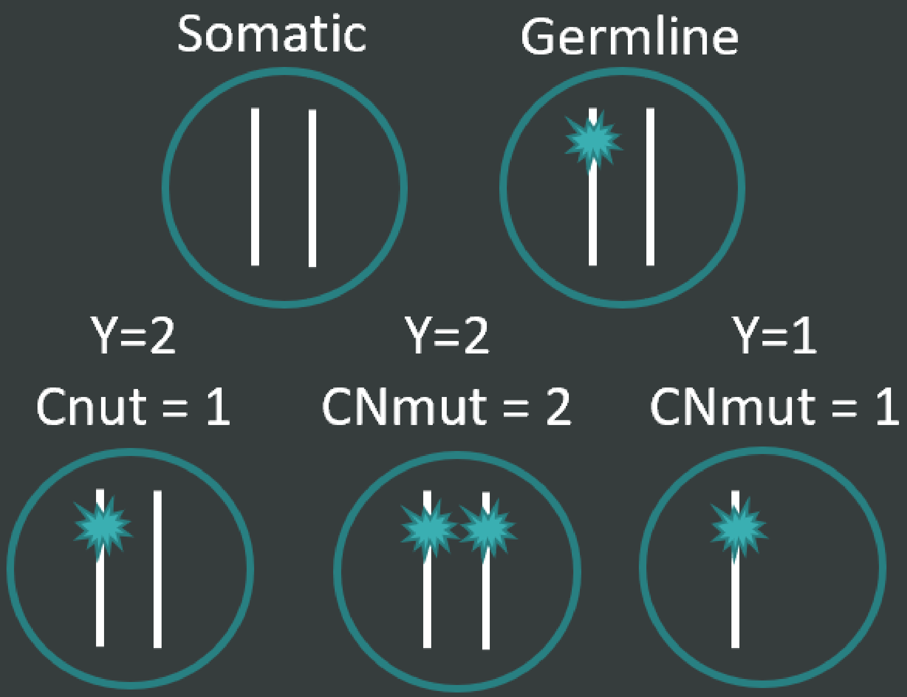


LOHGIC (Path. Purity); GermW,SomW,
Somatic CNmut = 1 (1.00) 0.00,1.00,0.00,0.00,1,1



Somatic vs. Germline

- Germline: Found in every cell, even non-cancerous
  - Allele frequency ≈50 or ≈100
  - Are passed onto children
  - Require more aggressive treatment
- Somatic: Found in some (or all) cancer cells
  - Allele frequency varies



Somatic

Germline

Y=2
Cnut = 1

Y=2
CNmut = 2

Y=1
CNmut = 1

Error

- Probability that the following model is correct

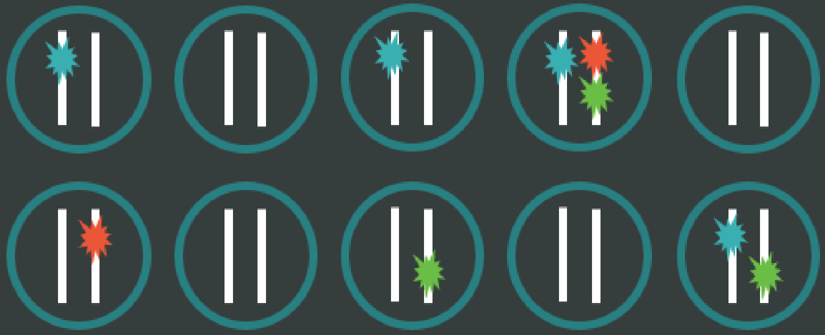
Models

- Y : Total number of copies of a gene per cell with particular mutation
- Copy Number of Mutations (CNmut): Total number of mutated alleles per cell with particular mutation

*Written by previous DIMACS student

Interpreting sequencing data

Find how often ✷ mutations occurred



Allele Frequency $\frac{\text{alleles with n mutation}}{\text{all alleles}} = 4/20 = 20\%$

Cancer Cell Frequency $\frac{\text{cells with n mutation}}{\text{cancerous cells}} = 4/6 = 67\%$

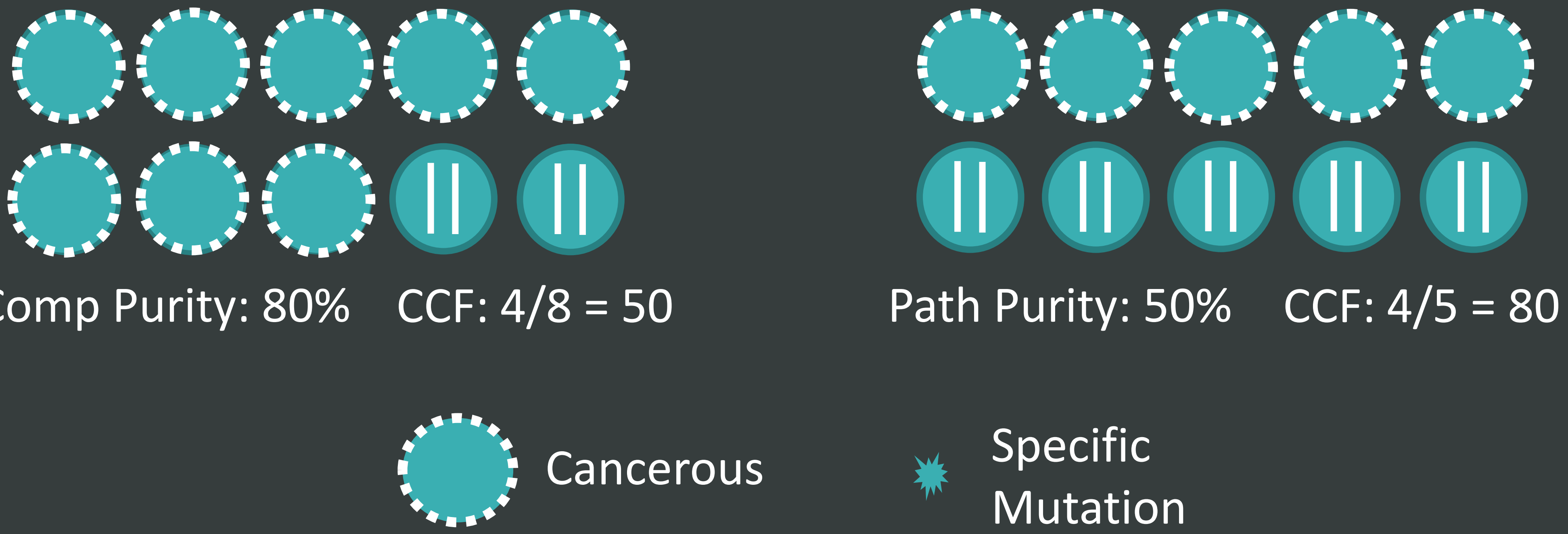
- Allele Frequency changes drastically based on sample purity, CCF is a more stable measurement
- Added error bars to CCF based on depth and purity irregularity in data

Why find CCF?

- Focus on finding drugs for most common and toxic mutations
- Understand how mutations grow to be one step ahead

Errors within Purity

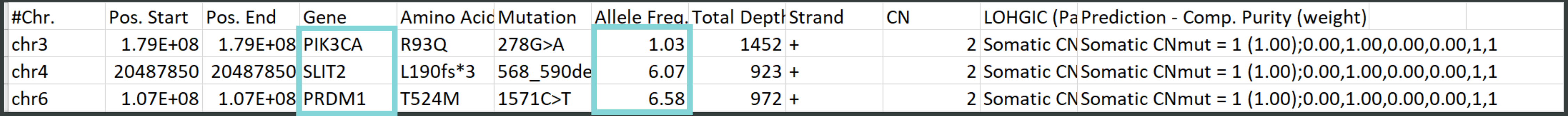
- Often we are given multiple purities that are drastically different
- Purity is found by staining cells and then manually counting to find cancerous cells
- **Error in purity → Error in CCF**


Comp Purity: 80%   CCF: 4/8 = 50

Path Purity: 50%   CCF: 4/5 = 80

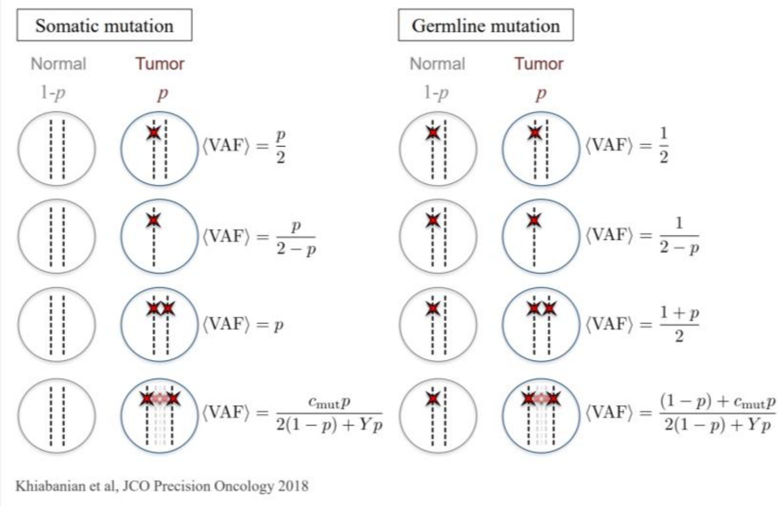
Cleaning Up Purity

For each mutation in the sample



#Chr.	Pos. Start	Pos. End	Gene	Amino Acid	Mutation	Allele Freq.	Total Deptl	Strand	CN	LOHGIC (Pa	Prediction - Comp. Purity (weight)
chr3	1.79E+08	1.79E+08	PIK3CA	R93Q	278G>A	1.03	1452	+	2	Somatic CN	Somatic CNmut = 1 (1.00);0.00,1.00,0.00,0.00,1,1
chr4	20487850	20487850	SLIT2	L190fs*3	568_590de	6.07	923	+	2	Somatic CN	Somatic CNmut = 1 (1.00);0.00,1.00,0.00,0.00,1,1
chr6	1.07E+08	1.07E+08	PRDM1	T524M	1571C>T	6.58	972	+	2	Somatic CN	Somatic CNmut = 1 (1.00);0.00,1.00,0.00,0.00,1,1

1. Calculate the purity, *p* for each model using the following equations

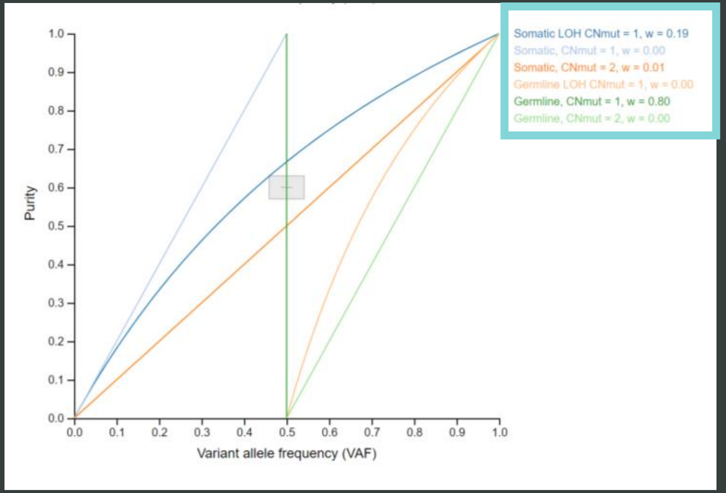

Somatic mutation
Normal
Tumor
1-p
p
$\langle VAF \rangle = \frac{p}{2}$
$\langle VAF \rangle = \frac{p}{2-p}$
$\langle VAF \rangle = p$
$\langle VAF \rangle = \frac{c_{mut}p}{2(1-p)+Yp}$
Germline mutation
Normal
Tumor
1-p
p
$\langle VAF \rangle = \frac{1}{2}$
$\langle VAF \rangle = \frac{1}{2-p}$
$\langle VAF \rangle = \frac{1+p}{2}$
$\langle VAF \rangle = \frac{(1-p)+c_{mut}p}{2(1-p)+Yp}$

Khiabanian et al, JCO Precision Oncology 2018

2. Using the *p* from the left and the given *VAF* calculate the *CCF*'s for each model

$$CCF = \frac{\textit{cells with n mutation}}{\textit{cancerous cells}}$$

3. Run the mutation through LOGIC to get the weights, *W* or probability for each model



4.

$$\sum W_{ij}(CCF_{ij}-1)^2$$

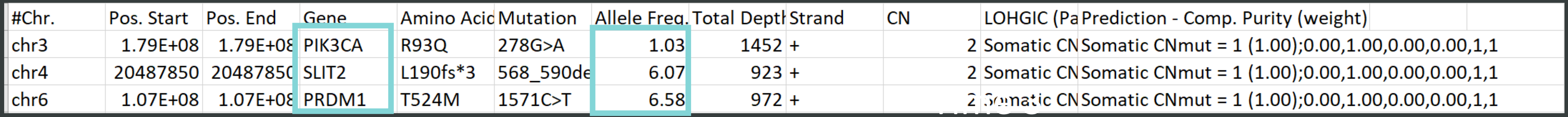
i = 3 mutations

j = 8 possible models

5. Developed a program that produced example data with a hidden purity to test the algorithm above

Nei's Genetic Distance

Calculate purity *p* for entire sample



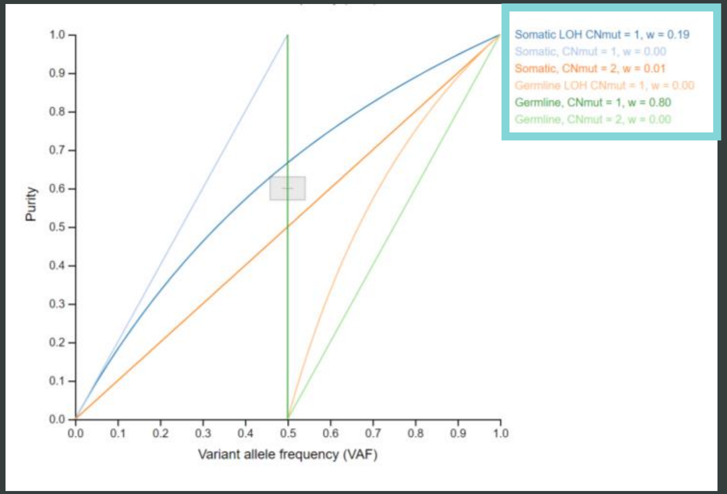
#Chr.	Pos. Start	Pos. End	Gene	Amino Acic	Mutation	Allele Freq.	Total Deptl	Strand	CN	LOHGIC (Pa	Prediction - Comp. Purity (weight)
chr3	1.79E+08	1.79E+08	PIK3CA	R93Q	278G>A	1.03	1452	+	2	Somatic CN	Somatic CNmut = 1 (1.00);0.00,1.00,0.00,0.00,1,1
chr4	20487850	2048785C	SLIT2	L190fs*3	568_590de	6.07	923	+	2	Somatic CN	Somatic CNmut = 1 (1.00);0.00,1.00,0.00,0.00,1,1
chr6	1.07E+08	1.07E+08	PRDM1	T524M	1571C>T	6.58	972	+	2	Somatic CN	Somatic CNmut = 1 (1.00);0.00,1.00,0.00,0.00,1,1

For each mutation in the sample:

1. Using given *p* and the given *VAF* calculate the *CCF*'s for <u>each</u> model

$$\boxed{CCF} = \frac{\textit{cells with n mutation}}{\textit{cancerous cells}}$$

2. Run the mutation through LOGIC to get the weights, $\boxed{W}$ or probability for each model



3.

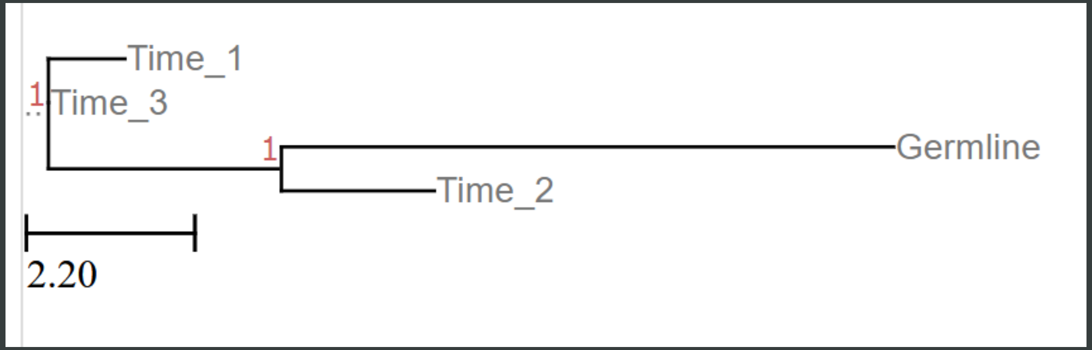
$$\textit{Nei's} = W * \frac{\sum_i Pi * Qi + (1 - Pi) * (1 - Qi)}{\sqrt{\sum_i P_i^2 + (1 - P_i)^2} * \sqrt{\sum_i P_i^2 + (1 - P_i)^2}}$$

Improved Tree Using Hamming Distance

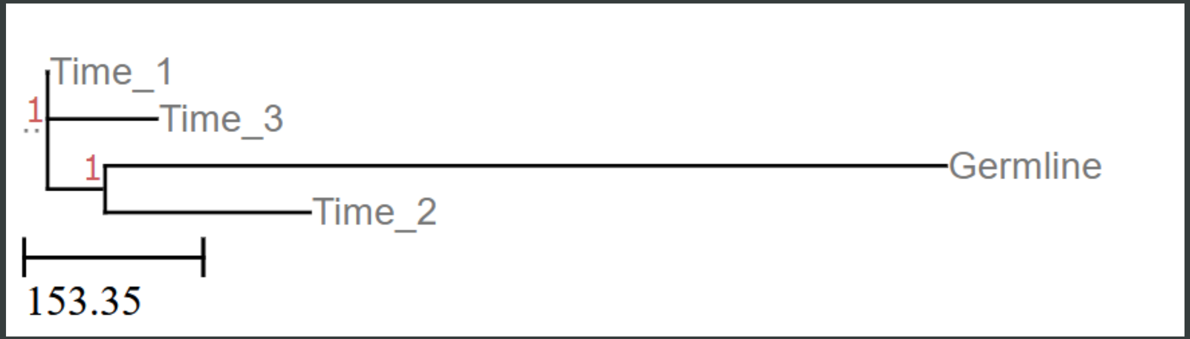
Differences

- Time 1 vs. Time 3
- Scale
- Germline length

Hamming



Nei



End Goal

- Incorporate error bars into trees since input data still has error
- Create trees for a large amount of patient data and track mutations
- Find patterns within patient trees to understand evolution of cells

[2]

Works Cited

DIMACS
Center for Discrete Mathematics & Theoretical Computer Science
Founded as a National Science Foundation Science and Technology Center

[1] Khiabanian et al, JCO Precision Oncology 2018

[2] ETE 3: Reconstruction, analysis and visualization of phylogenomic data.
Jaime Huerta-Cepas, Francois Serra and Peer Bork.
Mol Biol Evol 2016; doi: 10.1093/molbev/msw046

[3] DOI: 10.1200/PO.17.00148 JCO Precision Oncology - published online January 19, 2018

[4] http://ipl.physics.harvard.edu/wp-uploads/2013/03/PS3_Error_Propagation_sp13.pdf

This work was supported by National Science Foundation grant CCF-1559855:

REU Site: DIMACS

RUTGERS
Cancer Institute of New Jersey
at University Hospital
RUTGERS HEALTH

NSF

UNIVERSITY OF RHODE ISLAND
HOPE
1892