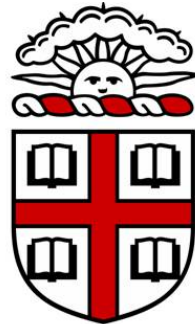


# Inferring Intra-Tumor Heterogeneity from Whole-Genome/Exome Sequencing Data

Layla Oesper  
Gryte Satas and Ben Raphael

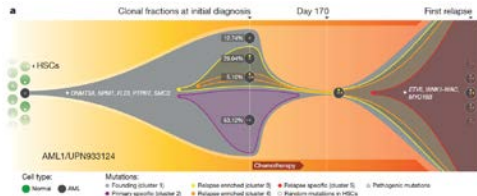
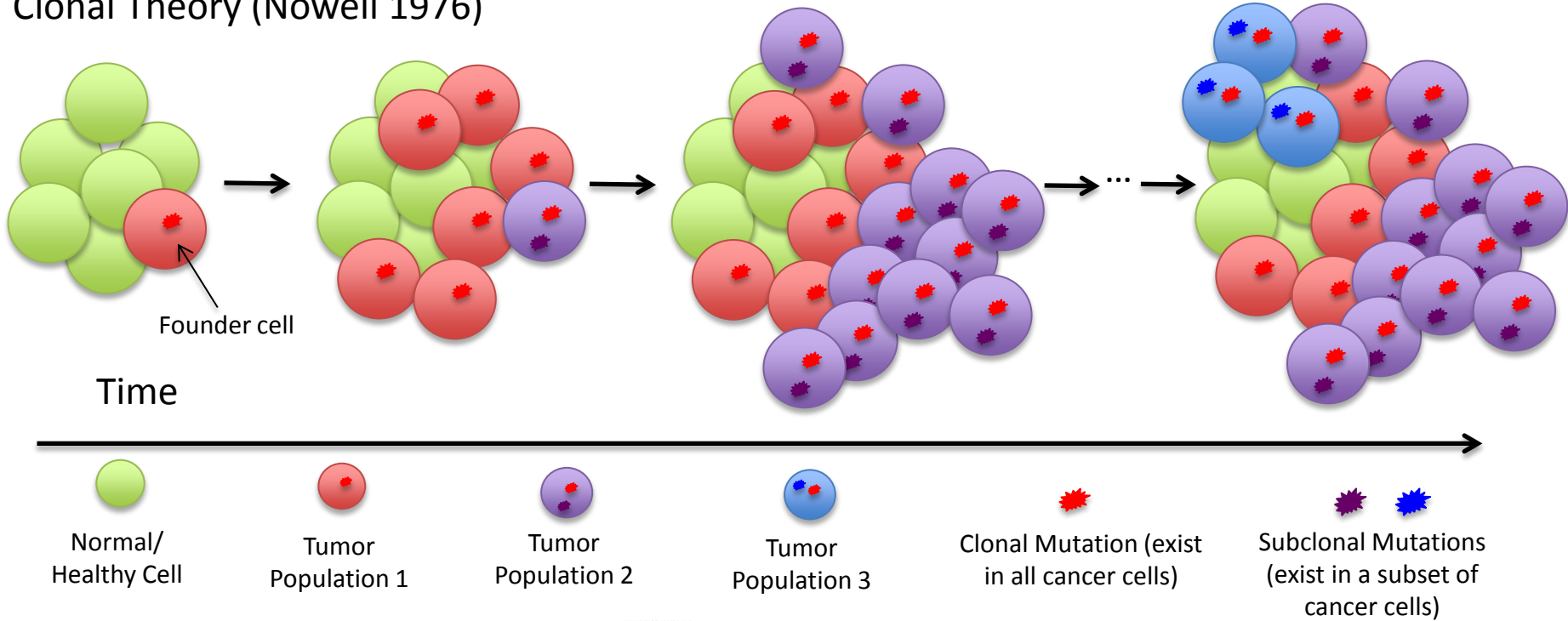


BROWN

Department of Computer Science  
Center for Computational Molecular Biology

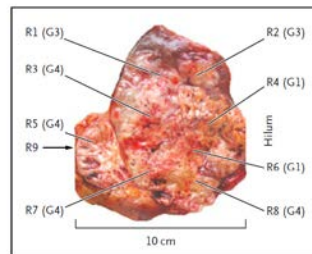
# Intra-Tumor Heterogeneity

Clonal Theory (Nowell 1976)

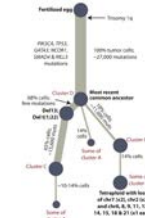


[Ding et al., Nature, 2012]

A Biopsy Sites



[Gerlinger et al., NEJM, 2012]



etc ...

[Nik-Zainal et al., Cell, 2012]

# Intra-Tumor Heterogeneity



Infer tumor composition from *single*,  
*mixed* tumor sample.

# Intra-Tumor Heterogeneity



## SNV Based Methods:

PyClone – Roth *et al.*, *Nature Methods* (2014)

SciClone – Miller *et al.* (In Press)

Nik-Zainal *et al.*, *Cell* (2012)

...

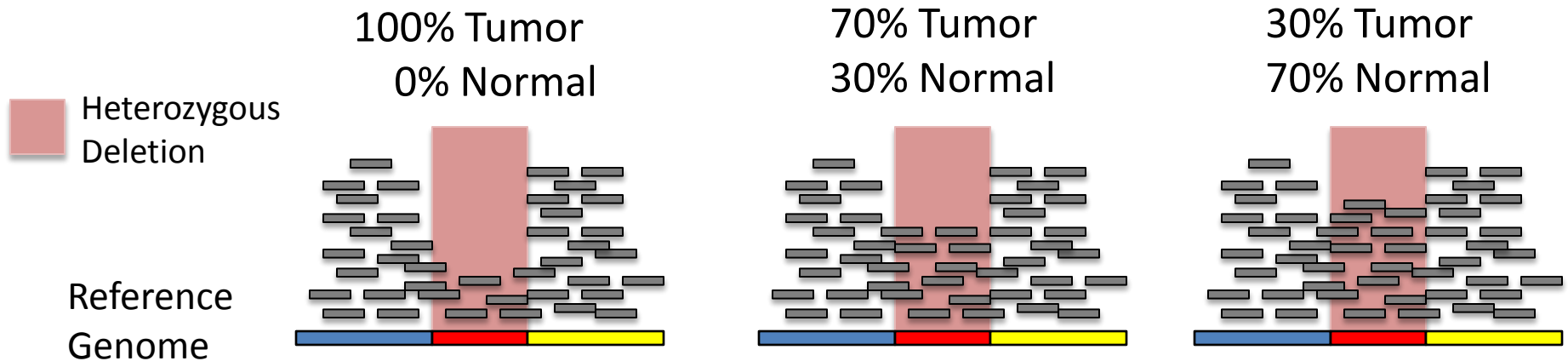
## CNA Based Methods:

Originally designed for SNP Array

ABSOLUTE – Carter *et al.*, *Nat. Biotechnol.* (2012)  
 ASCAT – Van Loo *et al.*, *PNAS* (2010)

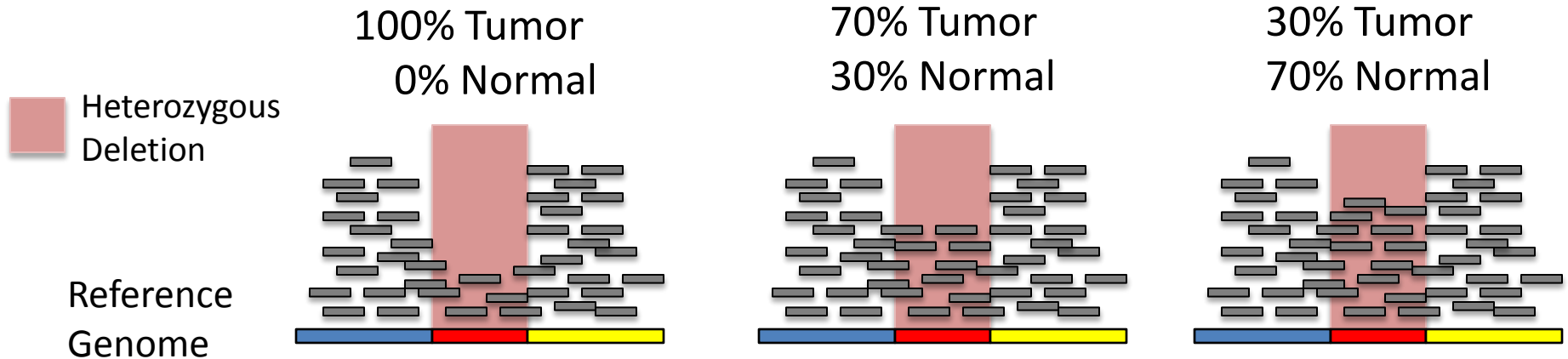
...

# Copy Number Aberrations in Tumors

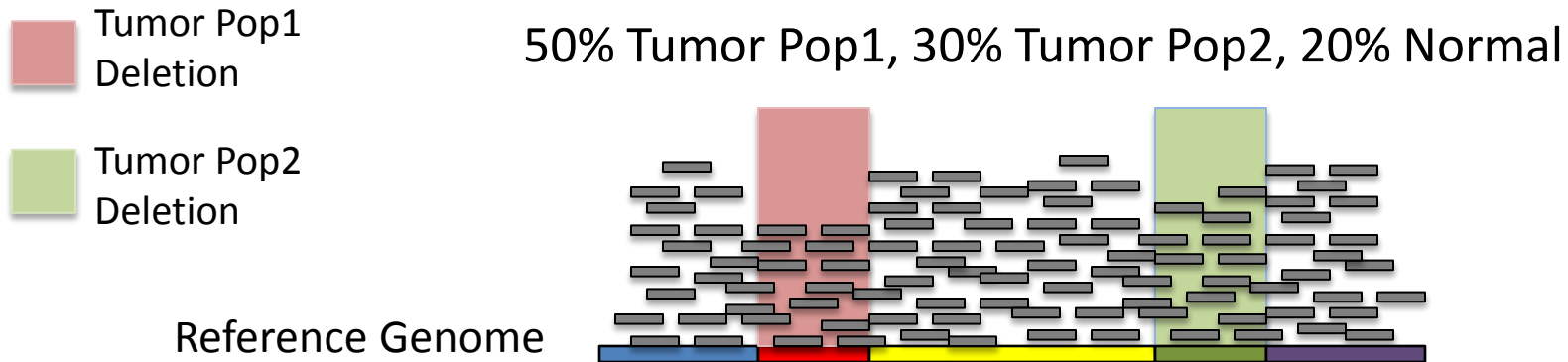


*Decrease in read-depth in deleted region  $\propto$  fraction of tumor cells*

# Copy Number Aberrations in Tumors

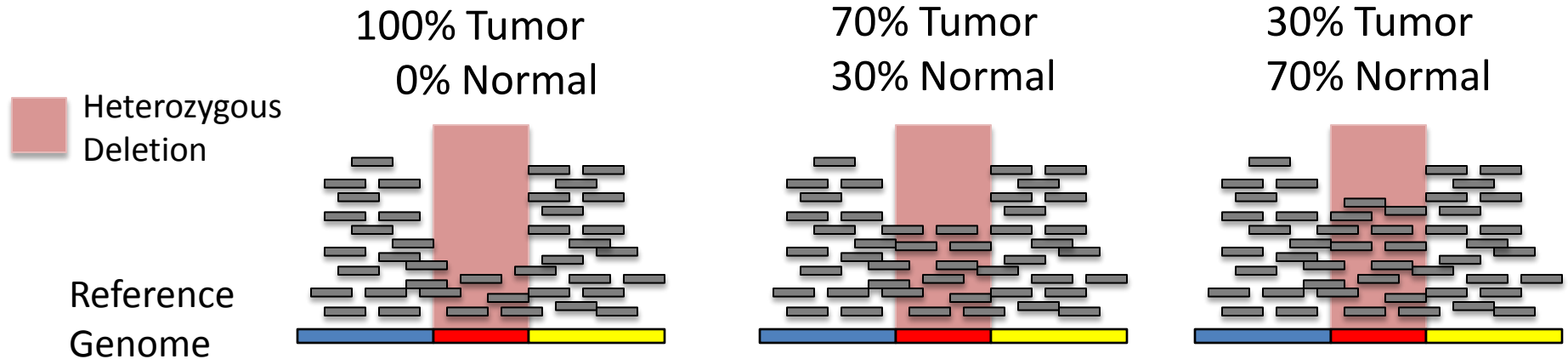


*Decrease in read-depth in deleted region*  $\propto$  fraction of tumor cells



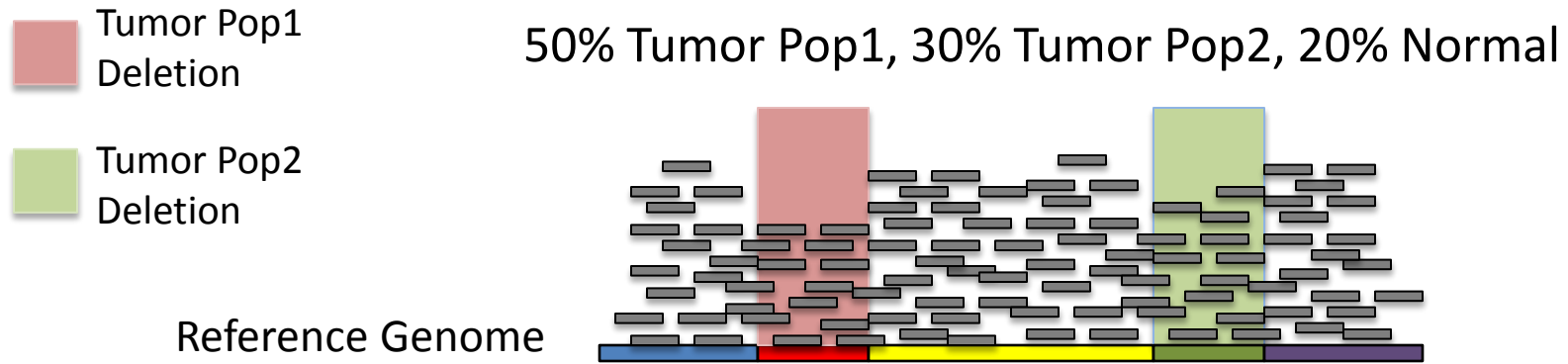
May be more than one tumor subpopulation.

# Copy Number Aberrations in Tumors

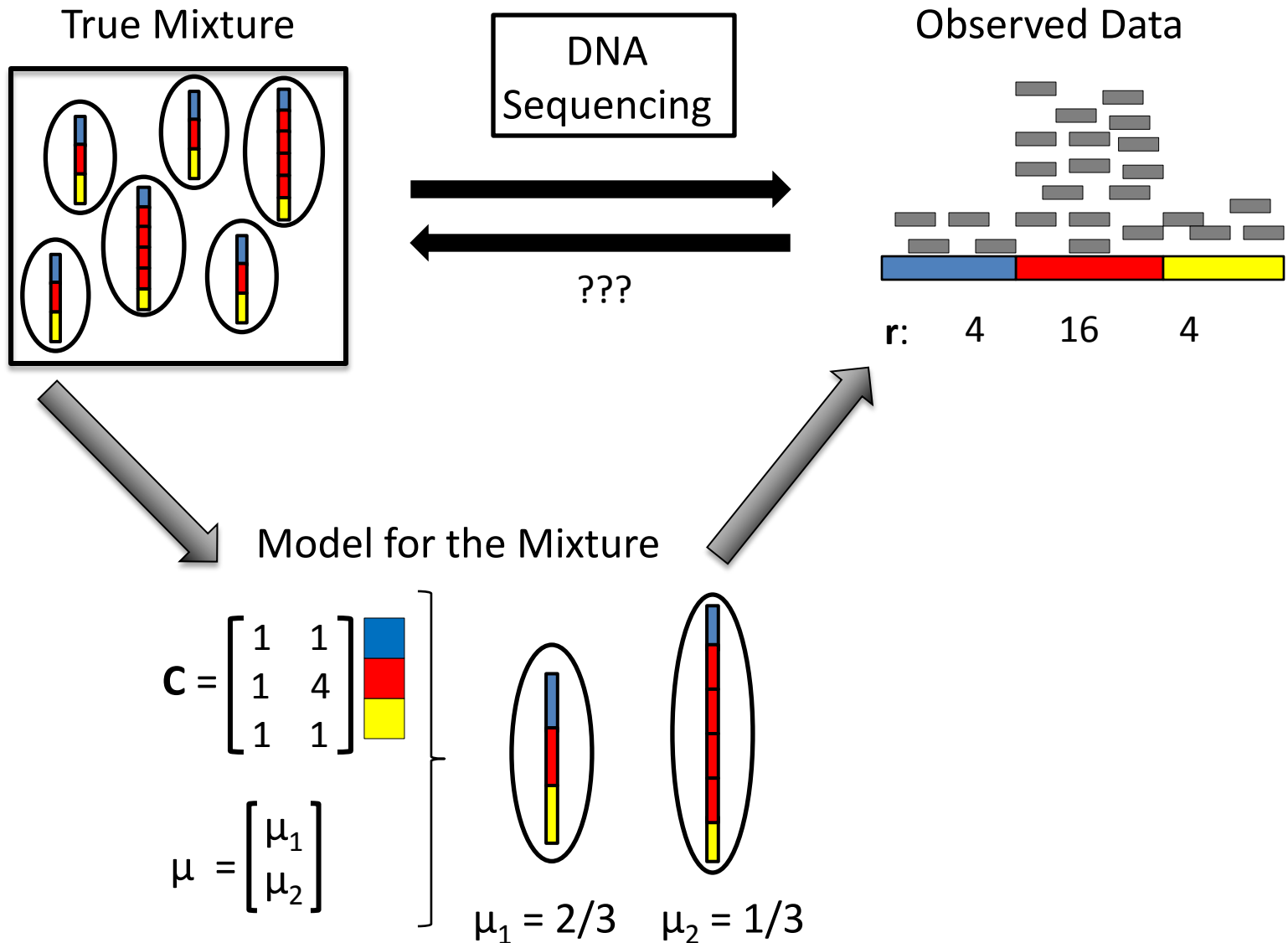


Copy number aberrations give **strong** signal in sequencing data

Signal can be combined across aberrations to infer tumor composition.



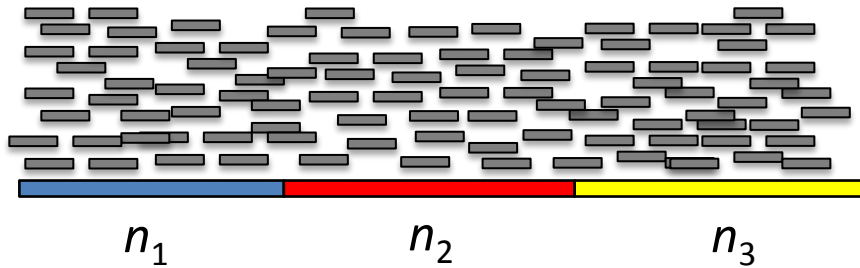
# Probabilistic Model



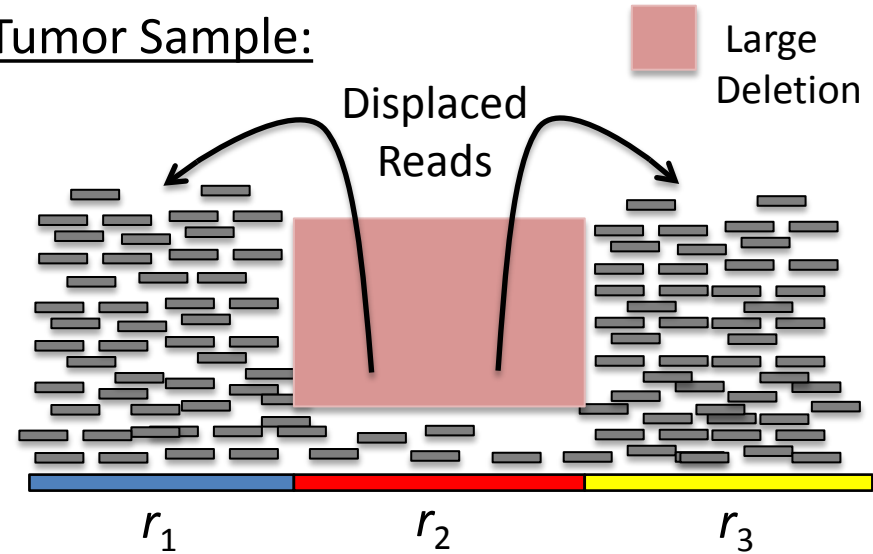


# Modeling Read Depth

Matched Normal Sample:



Tumor Sample:



Not independent!

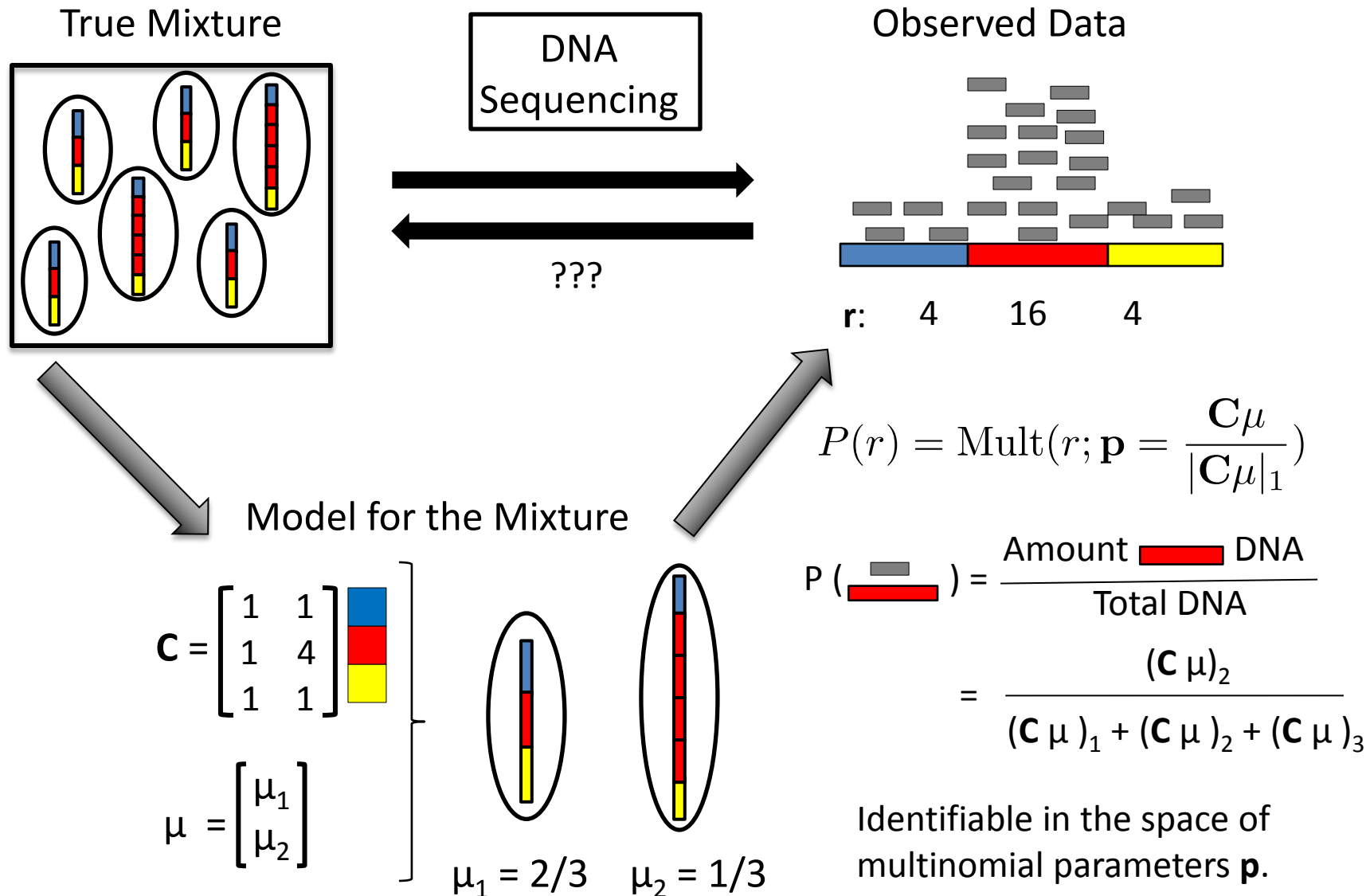
$$P(\mathbf{r}) = \text{Multinomial}(\mathbf{r}, \mathbf{p})$$

$$p_1 = P \left( \begin{array}{c} \text{grey bar} \\ \text{blue bar} \end{array} \right) = \frac{\text{Amount } \text{blue bar} \text{ DNA}}{\text{Total DNA}}$$

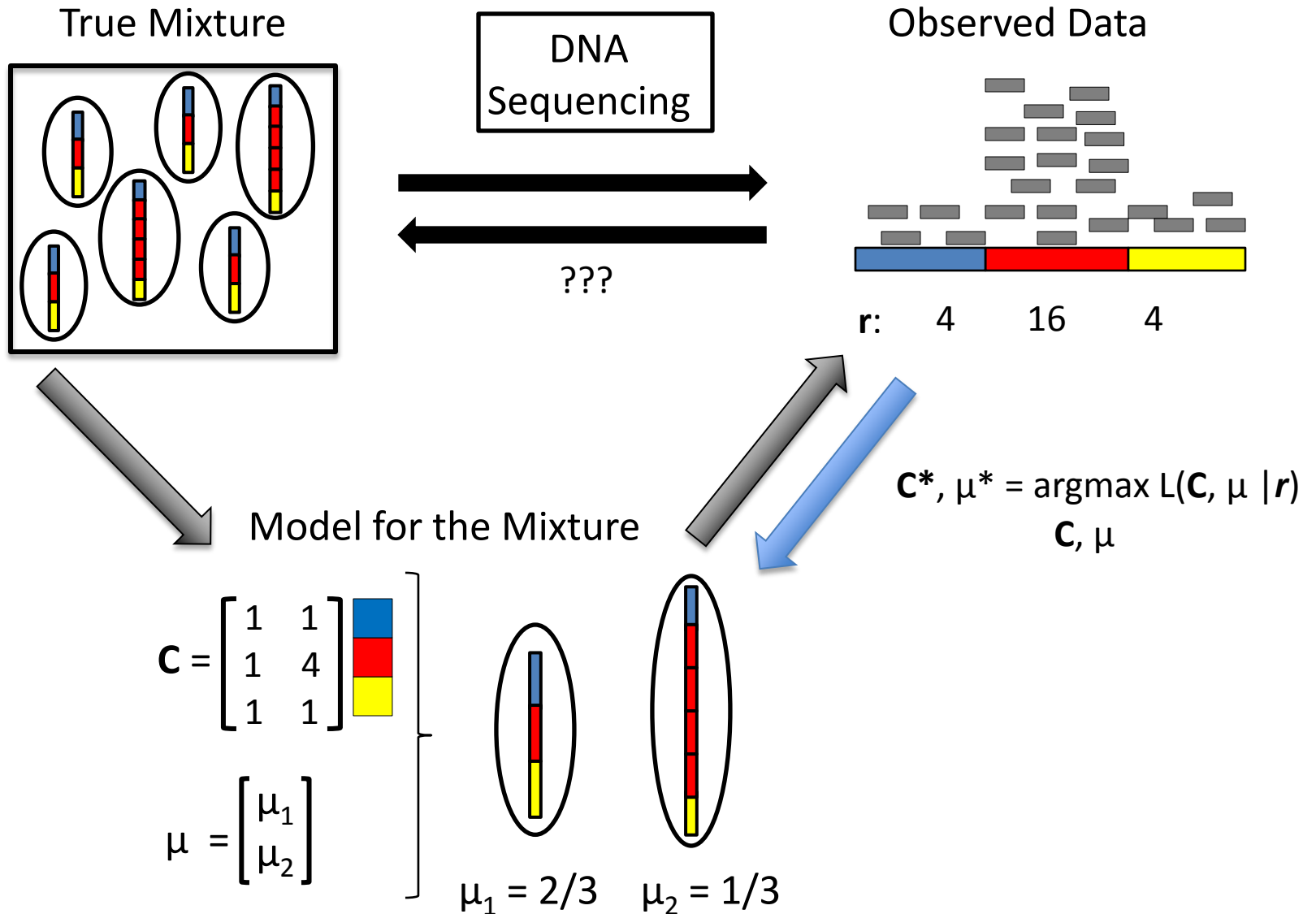
$$p_2 = P \left( \begin{array}{c} \text{grey bar} \\ \text{red bar} \end{array} \right) = \dots$$

$$p_3 = P \left( \begin{array}{c} \text{grey bar} \\ \text{yellow bar} \end{array} \right) = \dots$$

# Probabilistic Model



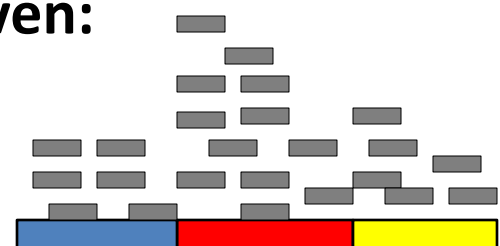
# Probabilistic Model



# Tumor Heterogeneity Analysis (THetA)


Finds the *most likely* tumor composition ( $\mathbf{C}$ ,  $\mu$ ) from measured read depth  $r$ .

**Given:**



$r$ : 6 12 6

**Find:**

$$\mathbf{C} = \begin{bmatrix} 1 & c_{12} \\ 1 & c_{22} \\ 1 & c_{32} \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

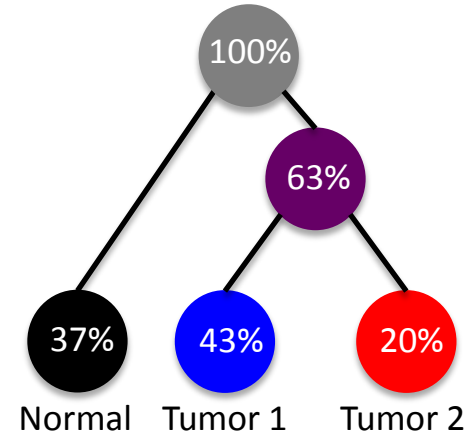
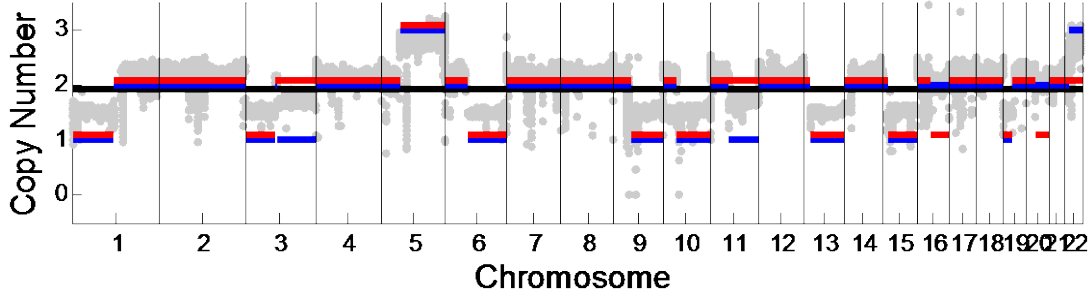
Such that  $L(\mathbf{C}, \mu \mid r)$  is maximized.

- (1) THetA is **efficient** (polynomial-time) for mixtures containing normal cells and *single* tumor subpopulation.
- (2) THetA can infer the composition of a mixture containing normal cells and **any number** of tumor subpopulations.

# Subclonal Simulation

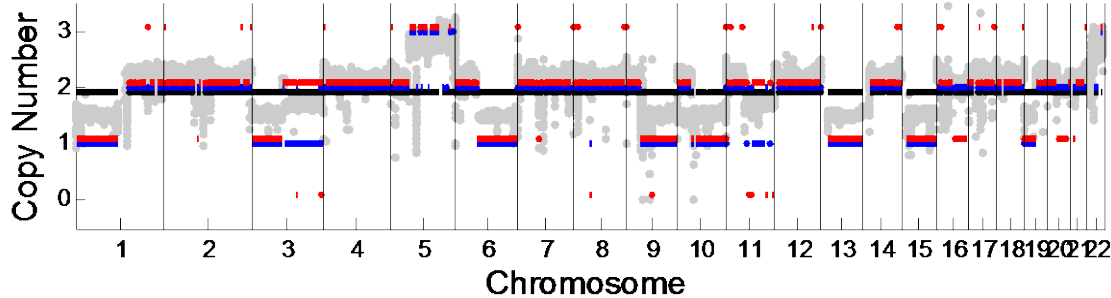
**Simulated Mixture of 3 subpopulations:**

Simulation – Normal:37%, Tumor1:43%, Tumor2:20%



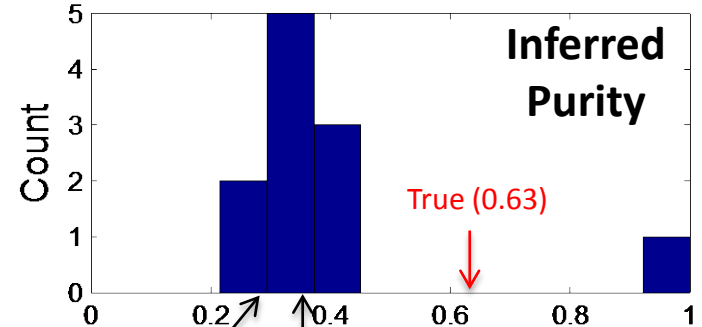
**THetA:**

THetA – Normal:40.1%, Tumor1:43.6%, Tumor2:16.3%



**ABSOLUTE:**

All 12 ABSOLUTE Solutions



BIC-seq [Xi *et al.*, PNAS (2011)] for segmentation

ABS Most Likely  
Karyotype (0.28)

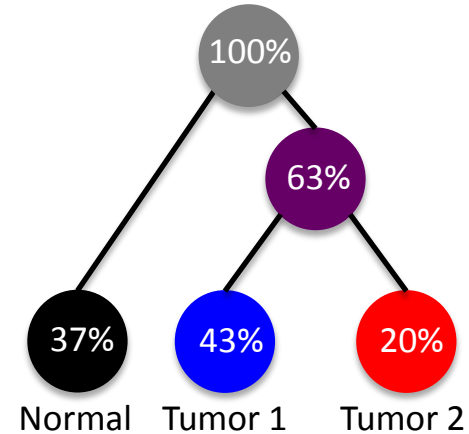
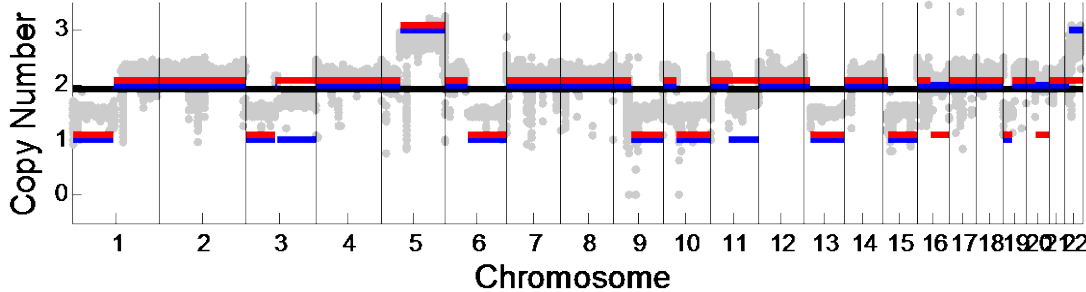
ABS Most  
Likely (0.35)

THetA (0.6)

# Subclonal Simulation

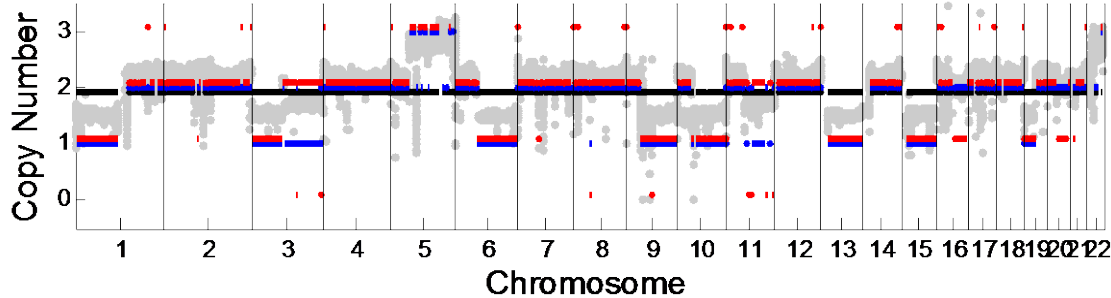
**Simulated Mixture of 3 subpopulations:**

Simulation – Normal:37%, Tumor1:43%, Tumor2:20%



**THetA:**

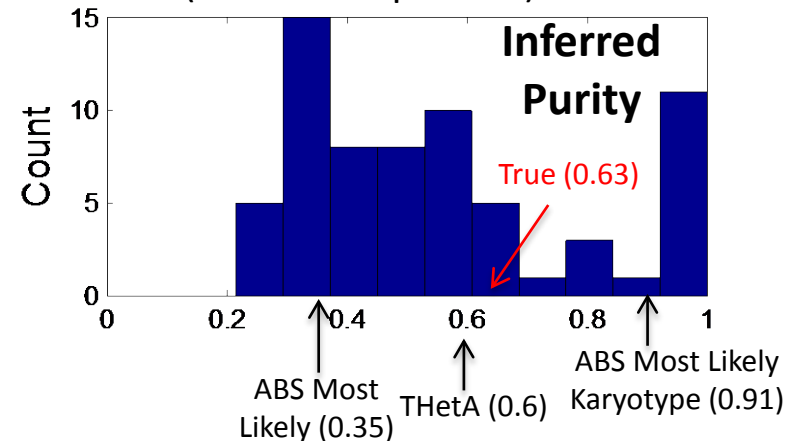
THetA – Normal:40.1%, Tumor1:43.6%, Tumor2:16.3%



BIC-seq [Xi *et al.*, PNAS (2011)] for segmentation

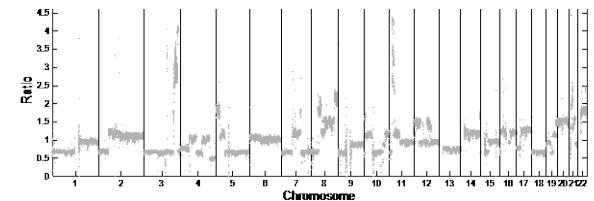
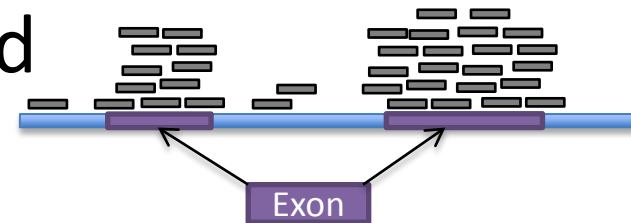
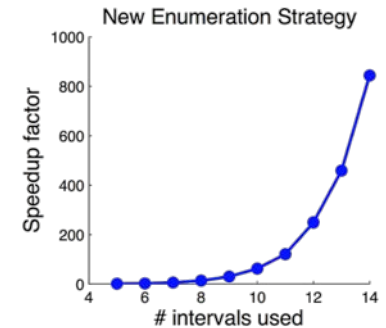
**ABSOLUTE:**

All 68 ABSOLUTE Solutions  
(subclonal up to 0.5)

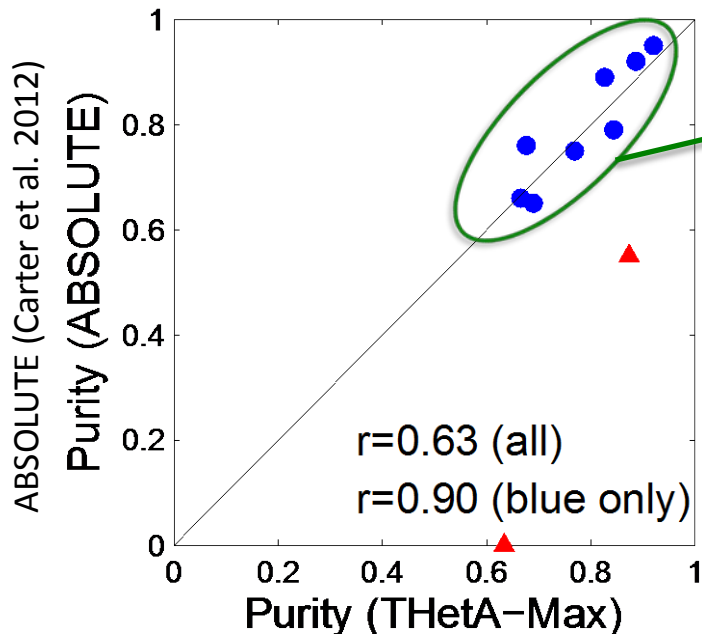


# THetA: Next-Generation

1. Improved optimization for multiple tumor subpopulations (> 1000X faster)
2. Extension to whole-exome and low pass (~7X) WGS data
3. Analysis of highly-rearranged genomes



# THetA for whole-exome data



Similar results across methods

## Segmentation:

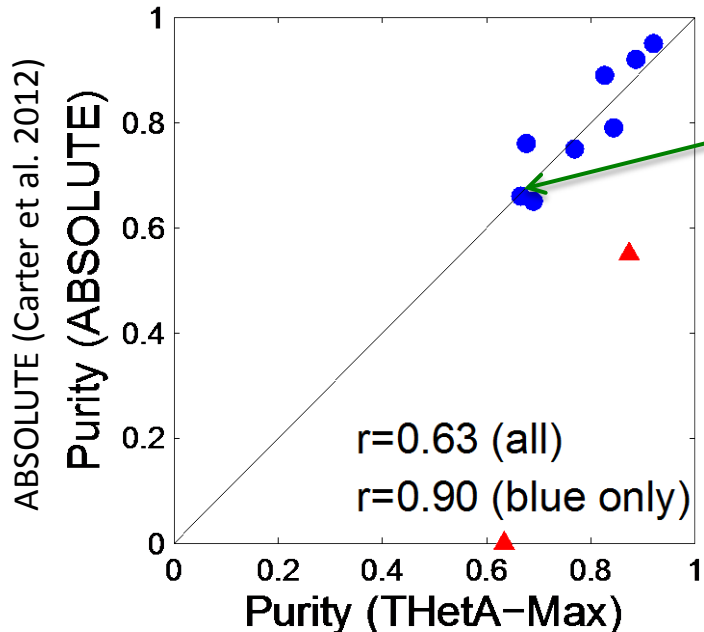
ExomeCNV [Sathirapongsasuti *et al.*, *Bioinformatics* (2011)]

EXCAVATOR [Magi *et al.*, *Genome Biology* (2013)]

TCGA



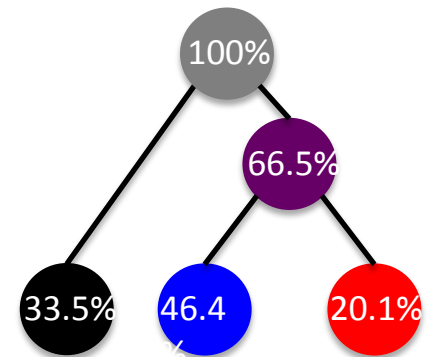
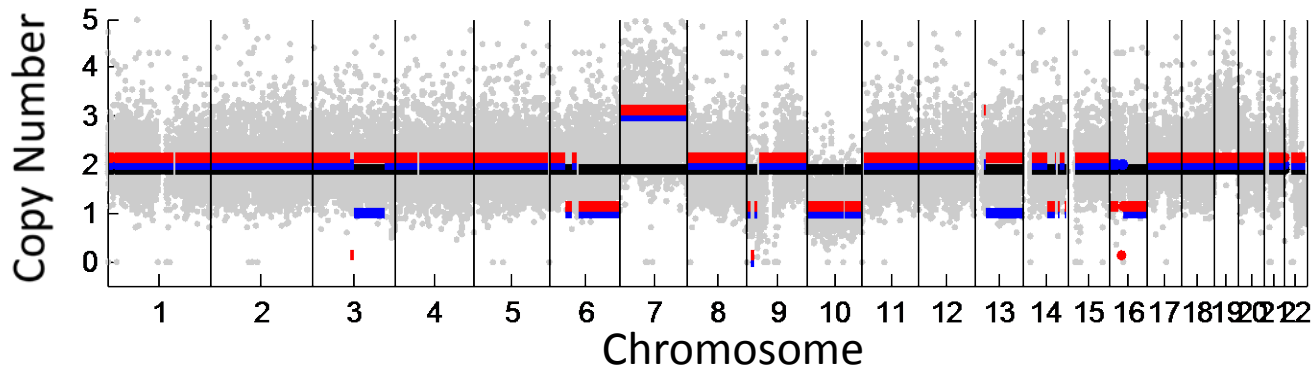
# THetA for whole-exome data



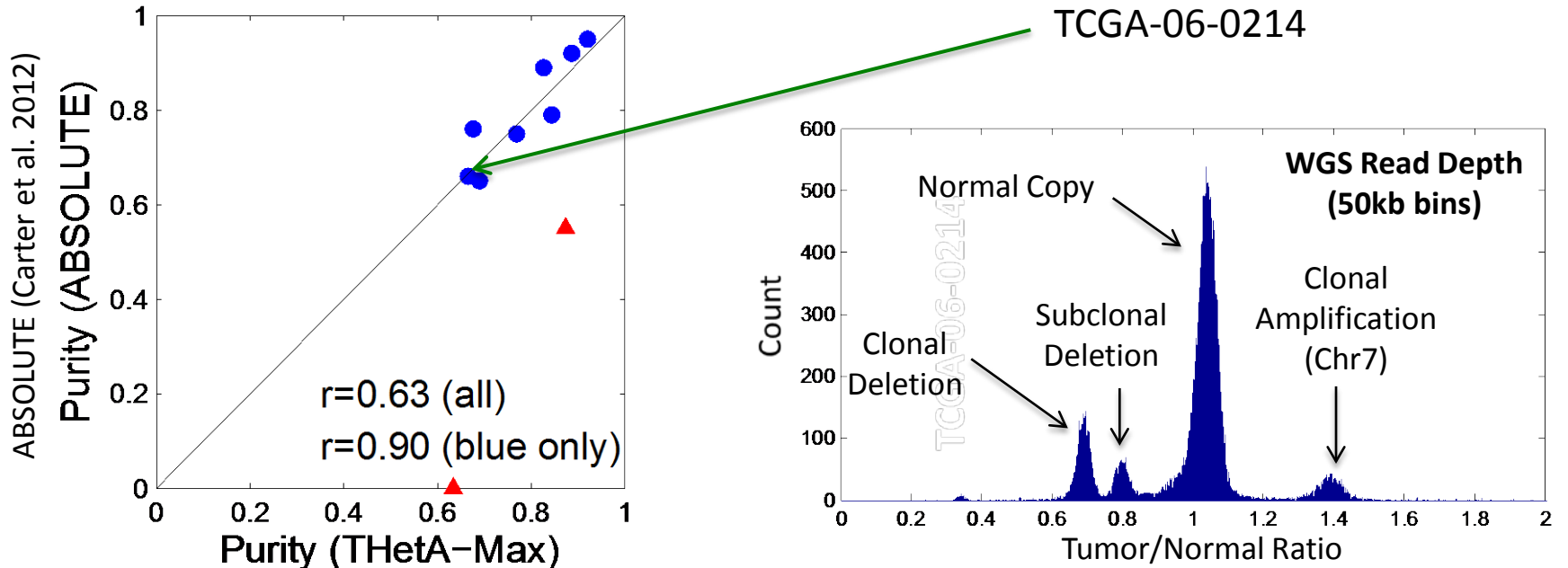
TCGA-06-0214

TCGA-06-0214	Method	Purity
	THetA-wxs	0.67
	THetA-wgs	0.67
	ABSOLUTE	0.66
TCGA Histopathology		0.25 – 0.8

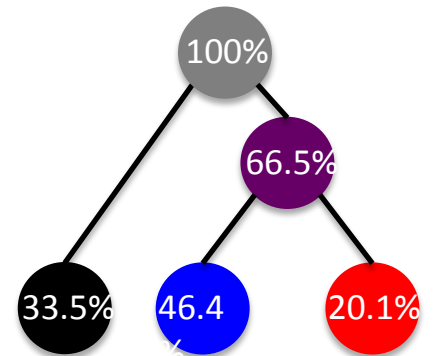
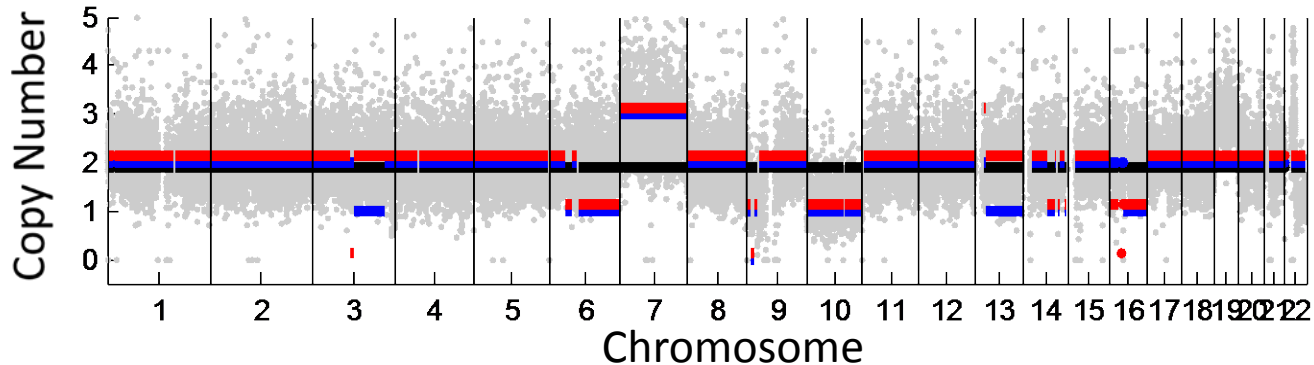
TCGA-06-0214-exome – Normal: 33.5%, Tumor1: 46.4%, Tumor2:20.1% (Glioblastoma)



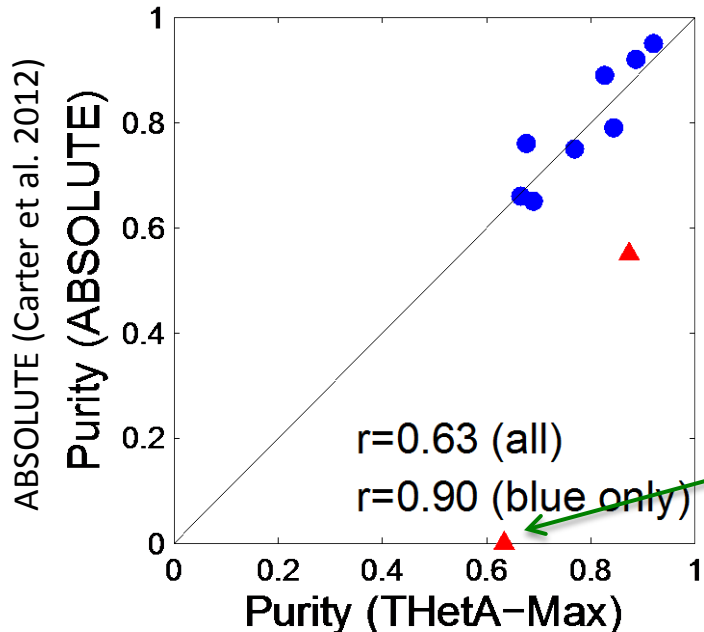
# THetA for whole-exome data



TCGA-06-0214-exome – Normal: 33.5%, Tumor1: 46.4%, Tumor2:20.1% (Glioblastoma)

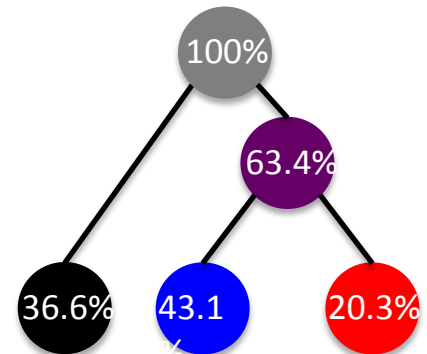
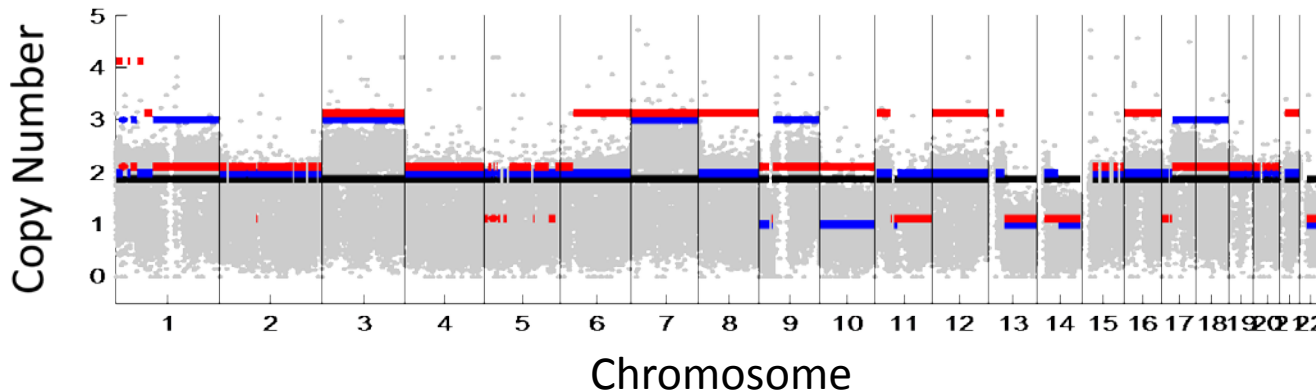


# THetA for whole-exome data

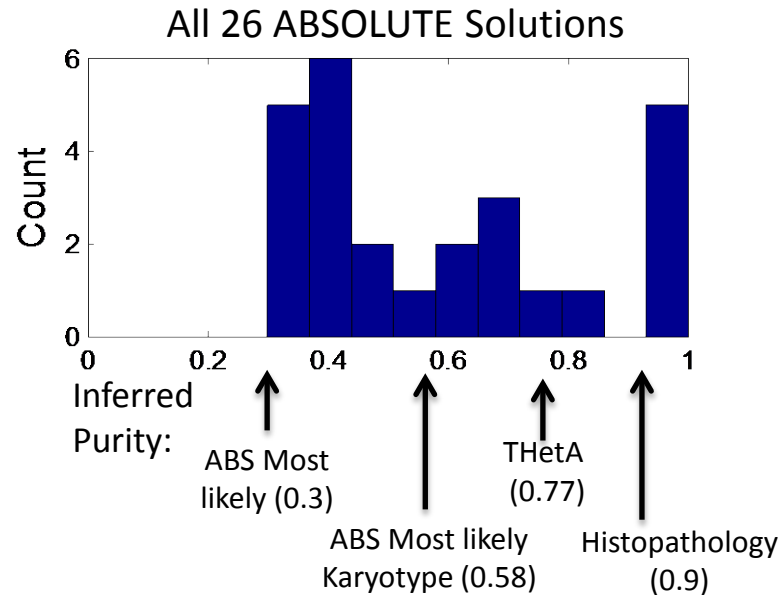
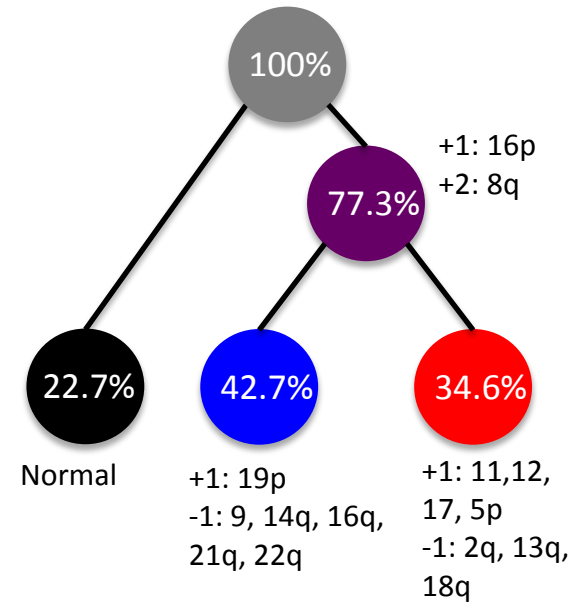
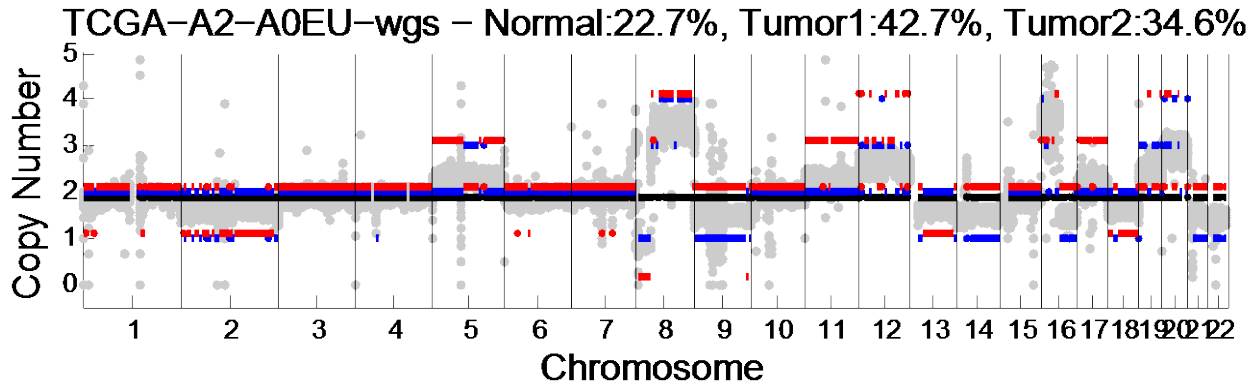


TCGA-06-0188:  
ABSOLUTE says  
highly subclonal

TCGA-06-0188-exome – Normal: 36.6%, Tumor1: 43.1%, Tumor2: 20.3% (Glioblastoma)

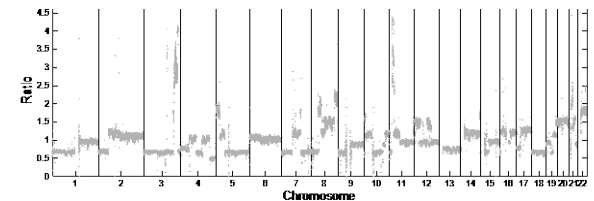
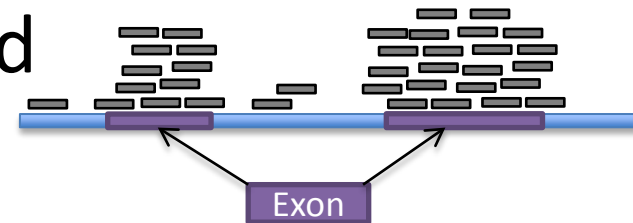
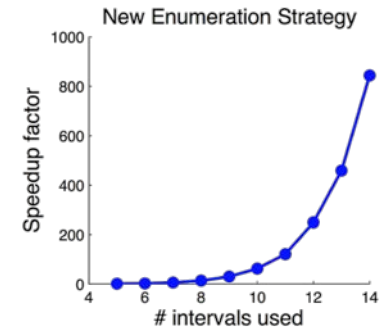


# Low-Pass Breast Cancer Genome

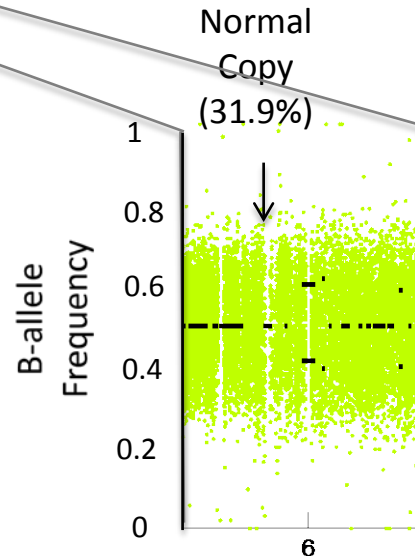
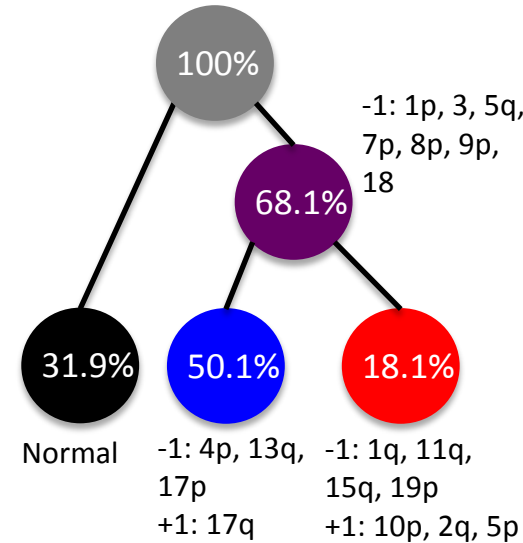
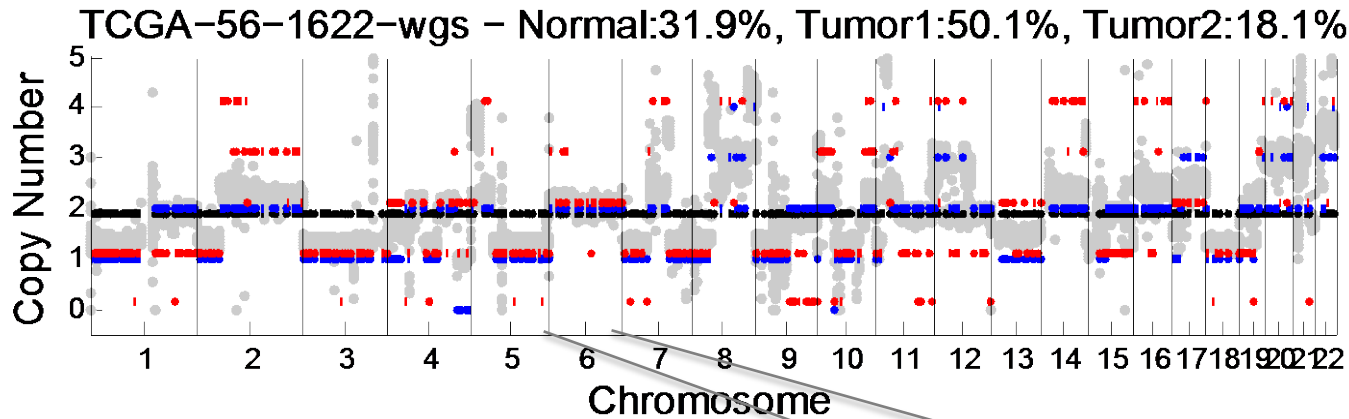


# THetA: Next-Generation

1. Improved optimization for multiple tumor subpopulations (> 1000X faster)
2. Extension to whole-exome and low pass (~7X) WGS data
3. Analysis of highly-rearranged genomes

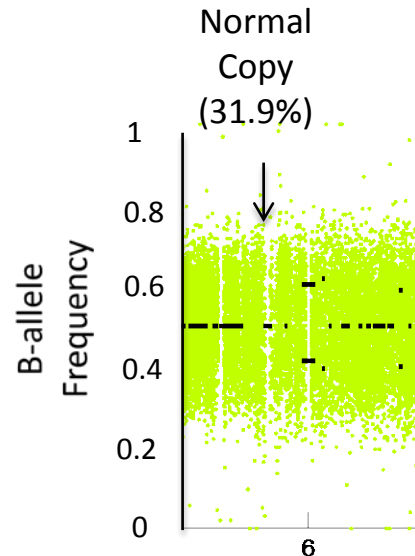
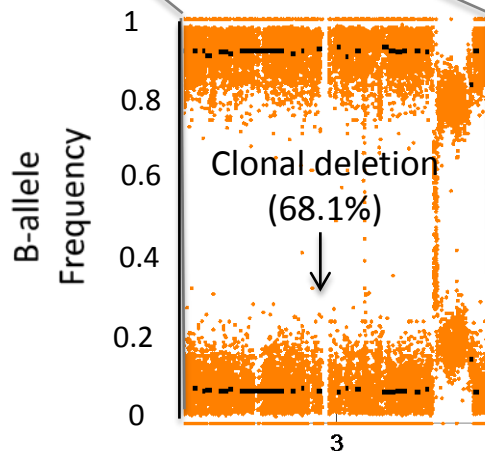
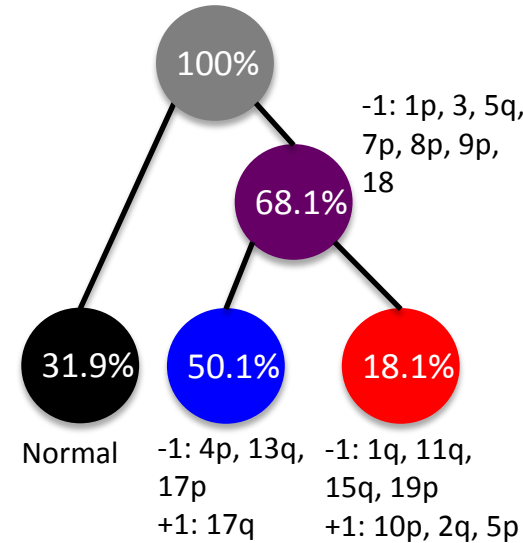
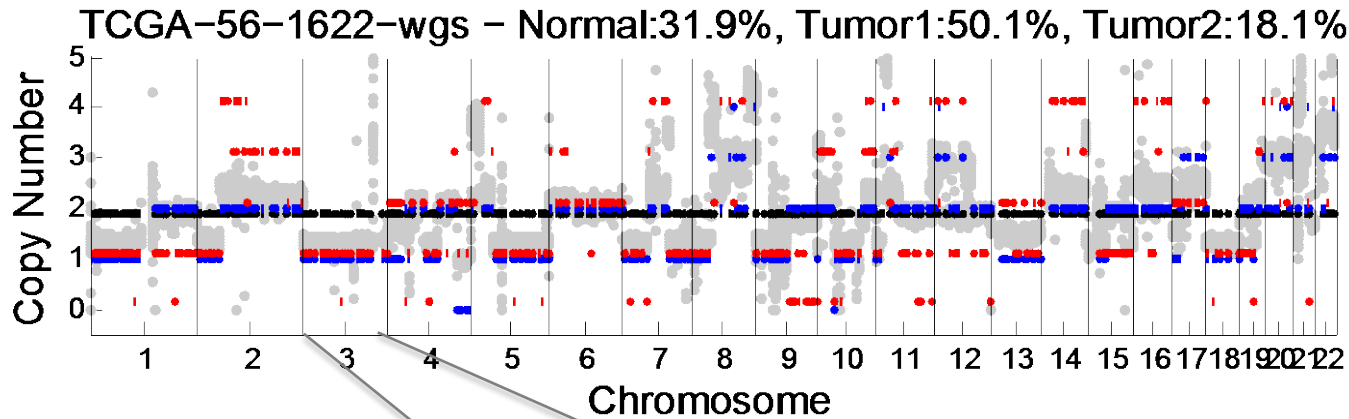


# Highly rearranged LUSC tumor



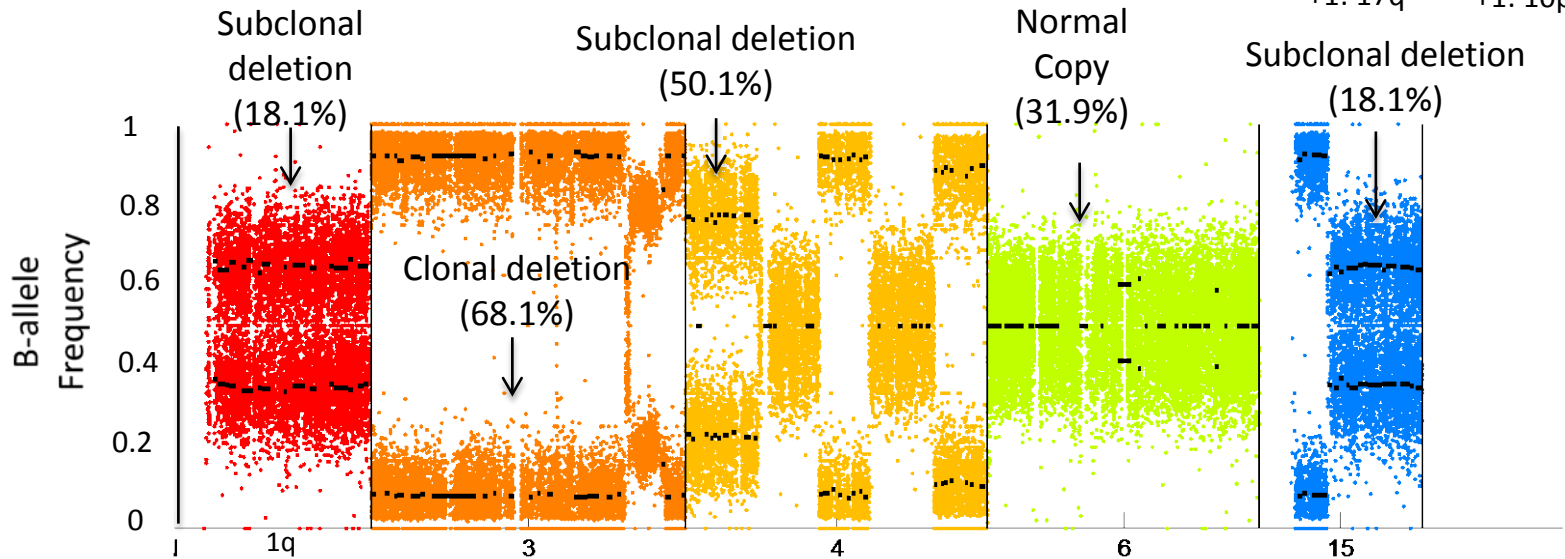
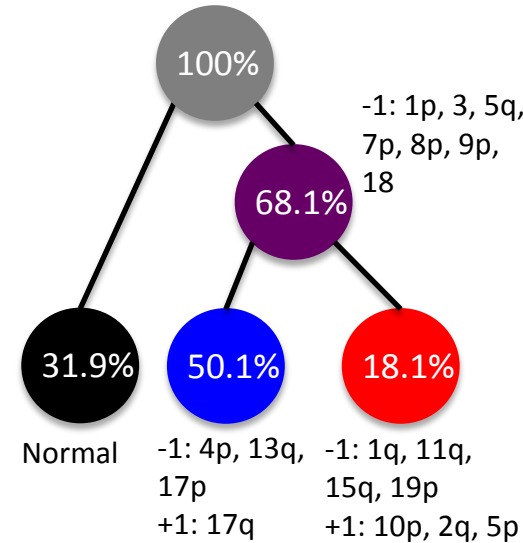
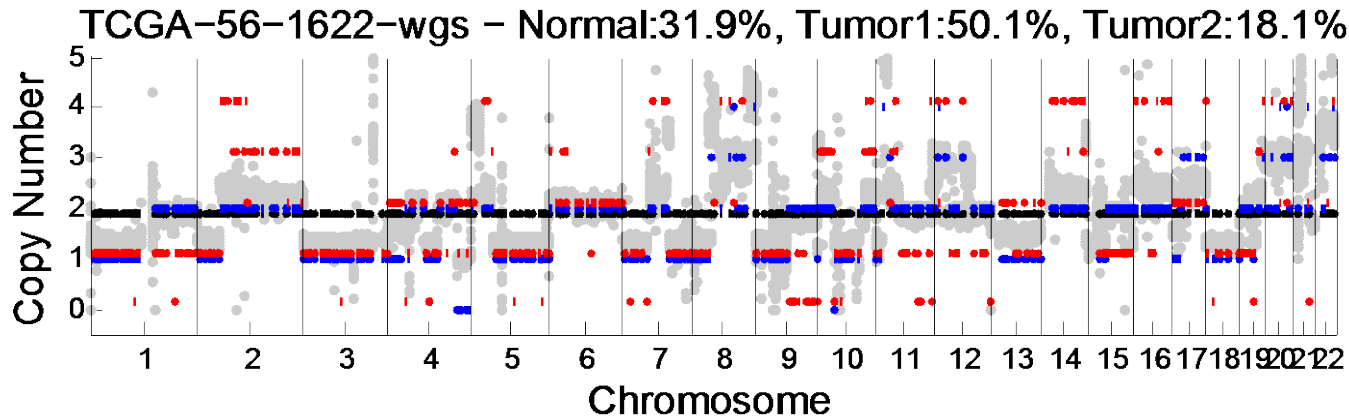
— Mean B-allele frequency suggested by data

# Highly rearranged LUSC tumor



— Mean B-allele frequency suggested by data

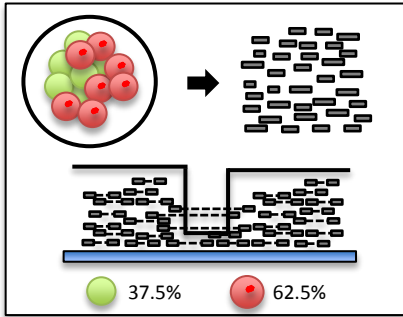
# Highly rearranged LUSC tumor



— Mean B-allele frequency suggested by data

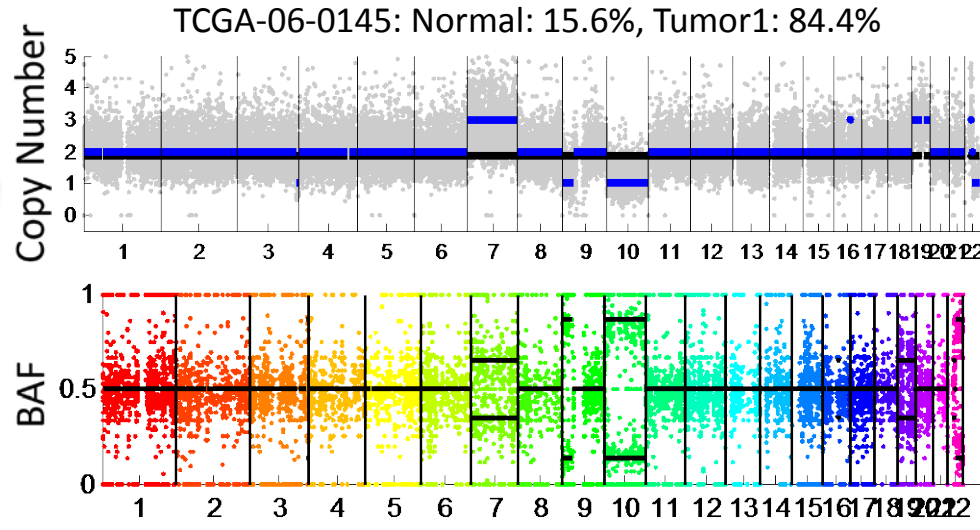


# THetA: Using B-Allele Frequencies

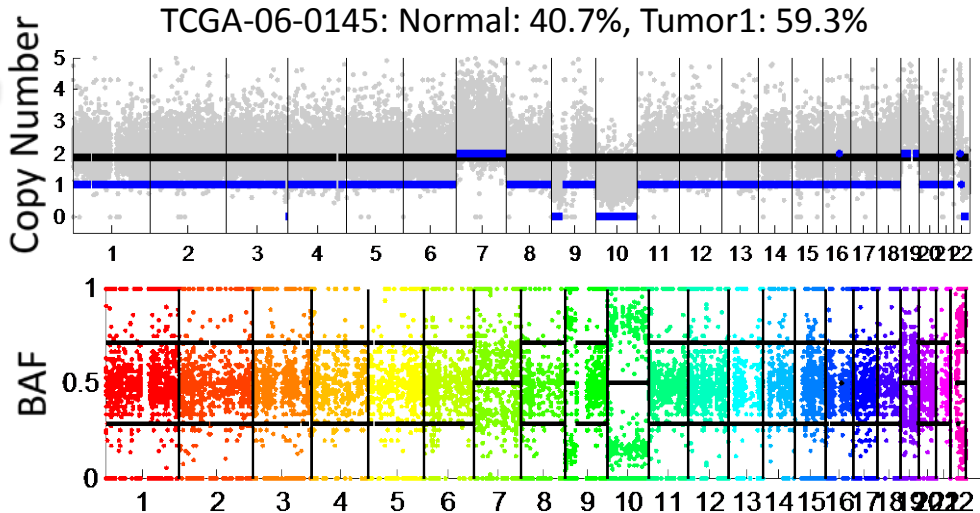


THetA returns two equally likely reconstructions for GBM TCGA-06-0145.

**Future work:** Incorporate BAFs and SNVs directly into THetA's model.



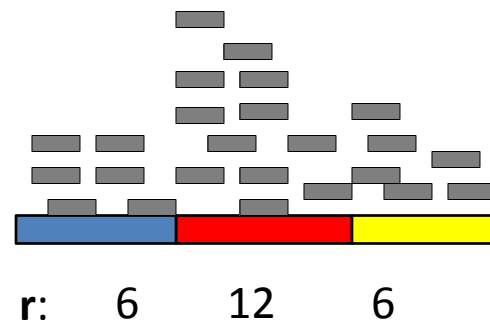
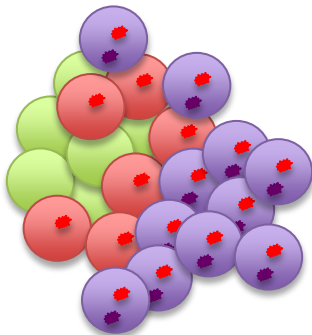
More likely solution using a probabilistic model of BAFs.



Less likely solution using a probabilistic model of BAFs.

# Summary

- Describe THetA – infers tumor sample purity and cancer subpopulations.
- Introduce improvements allowing THetA to be applied to a range of datatypes: WGS (including low pass), and WXS.



# Acknowledgments

Advisor:

Ben Raphael

Collaborators:

**Gryte Satas**

Ahmad Mahmoody

Others:

Max Leiserson

Hsin-ta Wu

Iman Hajirisouliha

Jason Dobson

Fabio Vandin



The Raphael Research Group



THetA is available at  
<http://compbio.cs.brown.edu>

NSF Graduate Research  
Fellowship DGE0228243