IntraTumor Heterogeneity and Cancer Evolution

S. Cenk Sahinalp

Bogaz’da Yaz Okulu 2018

Lab for Bioinformatics and Computational Genomics

1. ***Algorithmic infrastructure for genomics***
   - ***mapping, indexing, compression of big omic data to accommodate 250M human genomes to be sequenced by 2030***
2. ***Interpretation of genomic sequence data to resolve the sequence composition of repetitive genomic loci (e.g. immunoglobulin heavy locus, pharmacogenes)***
3. ***Large scale (expressed) genomic alteration detection in heterogeneous tumor samples and tumor evolution modeling***
4. ***Cancer network discovery and (rare) cancer driver prioritization***
5. ***The role of lncRNA based regulation in tumor emergence or progression***

INDIANA UNIVERSITY

VANCOUVER PROSTATE CENTRE
A UBC & VGH Centre of Excellence

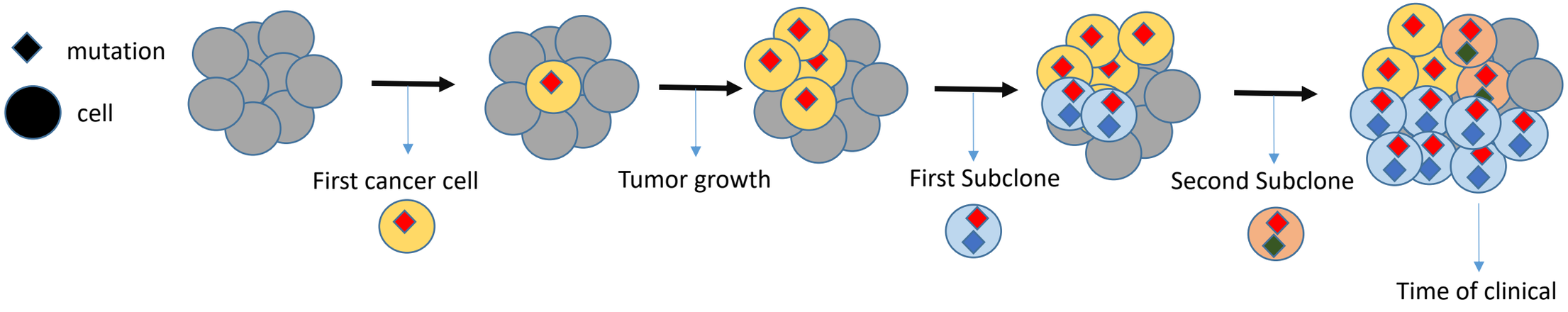
SIMON FRASER UNIVERSITY

Integrative inference of (sub)clonal tumor evolution from bulk and single-cell sequencing data



Simon Fraser University
Vancouver Prostate Centre
A UBC & VGH Centre of Excellence
ETH zürich
Indiana University

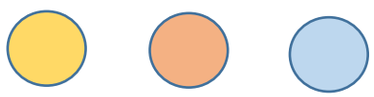
Clonal theory of cancer evolution



Computational problems in intro-tumor heterogeneity

1. Number of distinct cancer cell populations

3



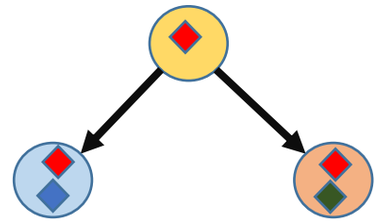
2. For each population set of mutations it harbors

3. Tumor purity and cancer cell fraction of each population

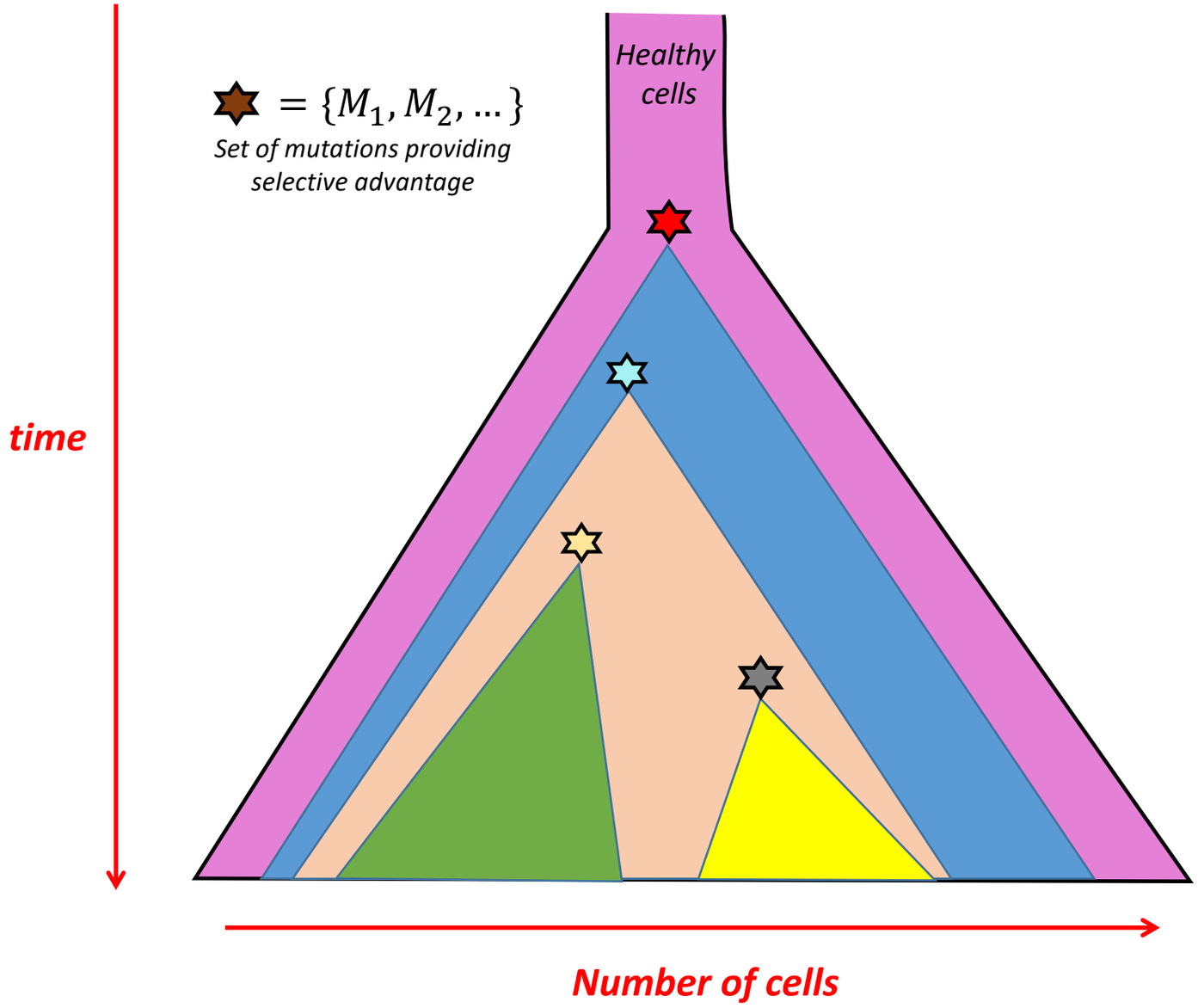
Tumor purity: 0.86

0.25 0.17 0.58

4. Tumor evolutionary tree

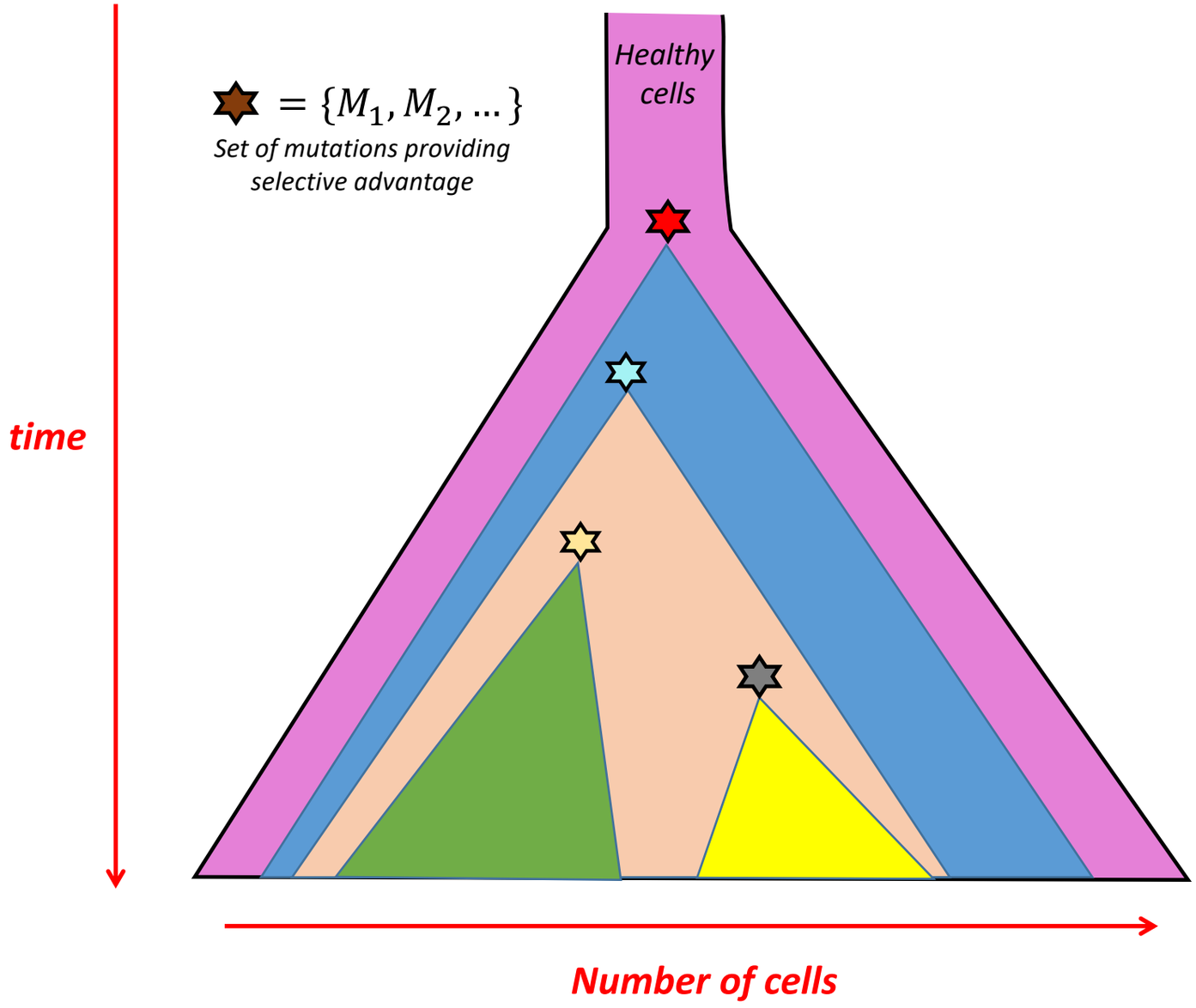


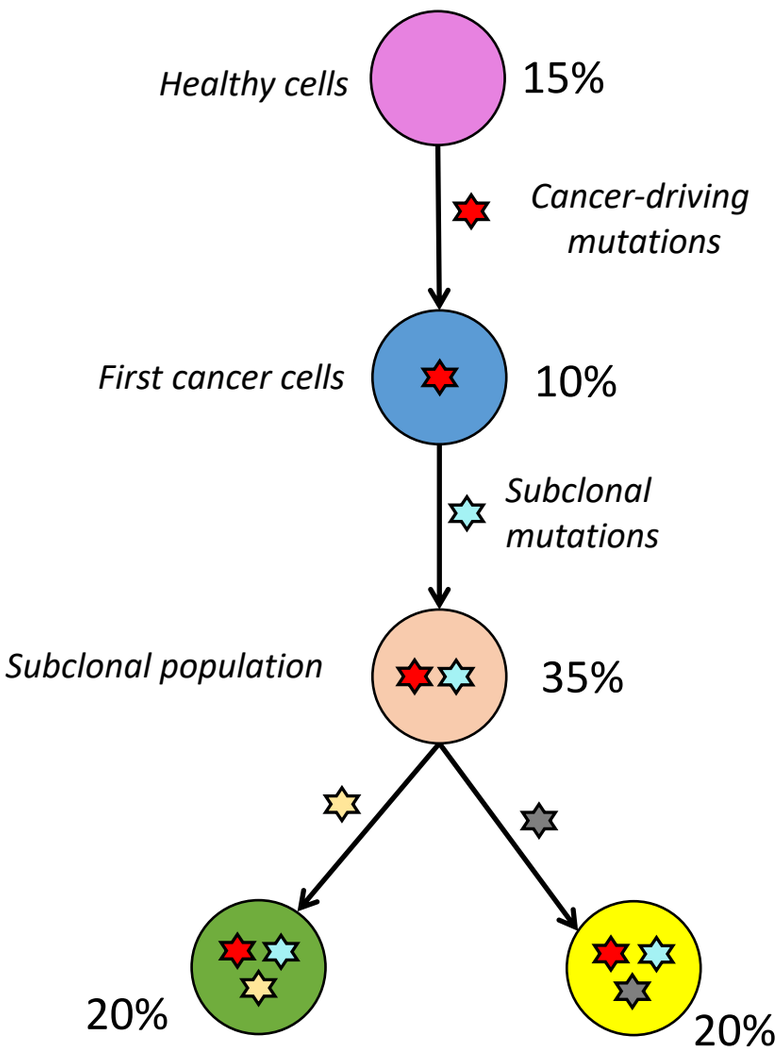
Clonal theory of cancer evolution



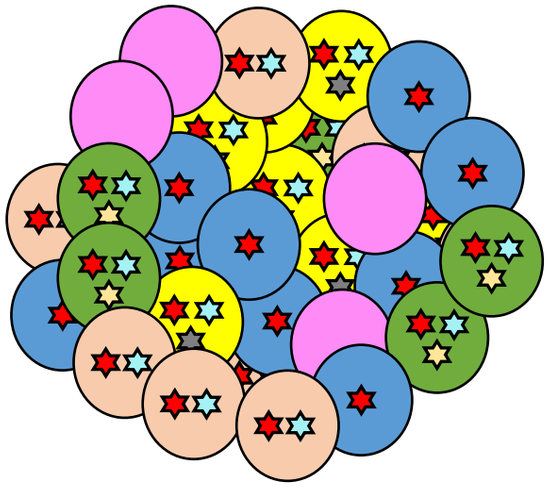
Nowell PC, *Science* 1976

Tree of tumor evolution

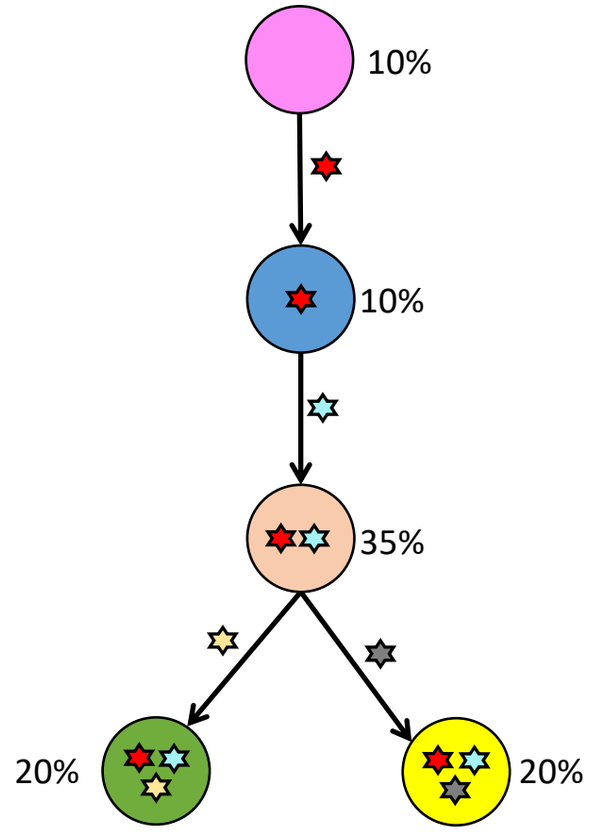




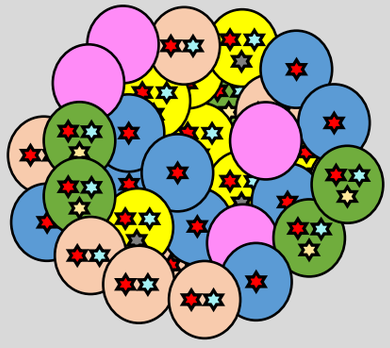
Deciphering tumor evolution and subclonal composition



1. ***Sequencing***
2. ***Mutation calling***
3. ***Tree inference***



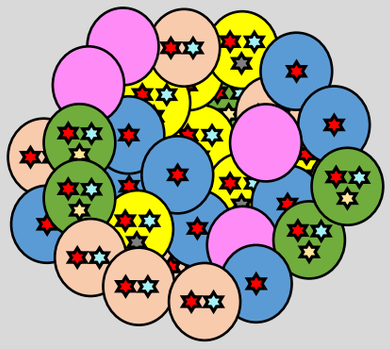
Studying tumor evolution by the use of bulk sequencing data



sequencing and mutation calling

Mutation	Var. reads	Ref. reads	(Avg) CN
M1	820	1100	2
M2	324	3700	4
M3	1215	2800	3
⋮	⋮	⋮	⋮

Studying tumor evolution by the use of bulk sequencing data



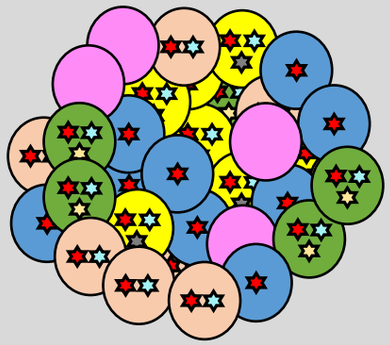
sequencing and mutation calling

Mutation	Var. reads	Ref. reads	(Avg) CN
M1	820	1100	2
M2	324	3700	4
M3	1215	2800	3
⋮	⋮	⋮	⋮

Software	Year	Reference	Phylogeny	Multiple samples	Inference
TrAp	2013	[37]	Y	N	Exhaustive search
Clomial	2014	[31]	N	Y	Binomial / EM
PhyloSub	2014	[32]	Y	Y	Tree-structured stick-breaking / MCMC
PyClone	2014	[38]	N	Y	Dirichlet process, beta-binomial / MCMC
RecBTP	2014	[39]	Y	N	Approximation algorithm
SciClone	2014	[40]	N	N	Beta mixture model
AncesTree	2015	[41]	Y	Y	Optimisation / MILP
CITUP	2015	[42]	Y	Y	Optimisation / QIP
LICHeE	2015	[43]	Y	Y	Heuristic
BayClone	2015	[44]	N	Y	Gibbs sampling / Metropolis-Hastings
CTPsingle	2016	[45]	Y	N	Dirichlet process, beta-binomial / MCMC
Cloe	2016	[46]	Y	Y	Metropolis-coupled MCMC
CHAT	2014	[54]	N	N	Dirichlet process Gaussian mixture model / MCMC
CloneHD	2014	[55]	N	Y	HMM / local optimisation
SubcloneSeeker	2014	[56]	Y	Y	Exhaustive enumeration
PhyloWGS	2015	[58]	Y	Y	Tree-structured stick-breaking / MCMC
SCHISM	2015	[57]	Y	Y	Likelihood ratio tests / genetic algorithm
SPRUCE	2016	[59]	Y	Y	Exhaustive enumeration
CANOPY	2016	[60]	Y	Y	MCMC

Kuipers et al., *BBA-Reviews on Cancer*,2017

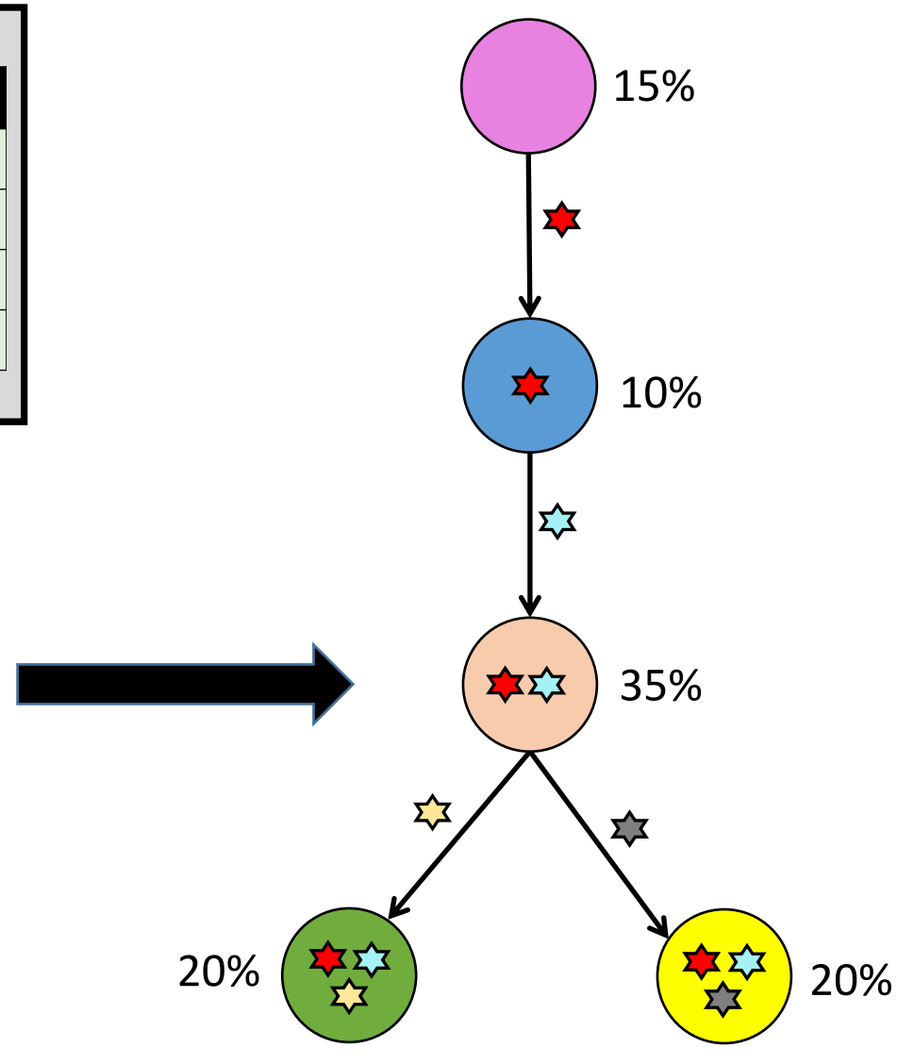
Studying tumor evolution by the use of bulk sequencing data



sequencing and mutation calling

Mutation	Var. reads	Ref. reads	(Avg) CN
M1	820	1100	2
M2	324	3700	4
M3	1215	2800	3
⋮	⋮	⋮	⋮

Software	Year	Reference	Phylogeny	Multiple samples	Inference
TrAp	2013	[37]	Y	N	Exhaustive search
Clomial	2014	[31]	N	Y	Binomial / EM
PhyloSub	2014	[32]	Y	Y	Tree-structured stick-breaking / MCMC
PyClone	2014	[38]	N	Y	Dirichlet process, beta-binomial / MCMC
RecBTP	2014	[39]	Y	N	Approximation algorithm
SciClone	2014	[40]	N	N	Beta mixture model
AncesTree	2015	[41]	Y	Y	Optimisation / MILP
CITUP	2015	[42]	Y	Y	Optimisation / QIP
LICHeE	2015	[43]	Y	Y	Heuristic
BayClone	2015	[44]	N	Y	Gibbs sampling / Metropolis-Hastings
CTPsingle	2016	[45]	Y	N	Dirichlet process, beta-binomial / MCMC
Cloe	2016	[46]	Y	Y	Metropolis-coupled MCMC
CHAT	2014	[54]	N	N	Dirichlet process Gaussian mixture model / MCMC
CloneHD	2014	[55]	N	Y	HMM / local optimisation
SubcloneSeeker	2014	[56]	Y	Y	Exhaustive enumeration
PhyloWGS	2015	[58]	Y	Y	Tree-structured stick-breaking / MCMC
SCHISM	2015	[57]	Y	Y	Likelihood ratio tests / genetic algorithm
SPRUCE	2016	[59]	Y	Y	Exhaustive enumeration
CANOPY	2016	[60]	Y	Y	MCMC



Kuipers et al., *BBA-Reviews on Cancer*,2017

CTPsingle: Clustering of mutations based on read counts

M_i = heterozygous SNV from diploid region

t_i = total number of reads covering genomic position of M_i

$\Rightarrow$ total number of cells in the sample $\sim \frac{t_i}{2}$

v_i = total number of reads supporting M_i

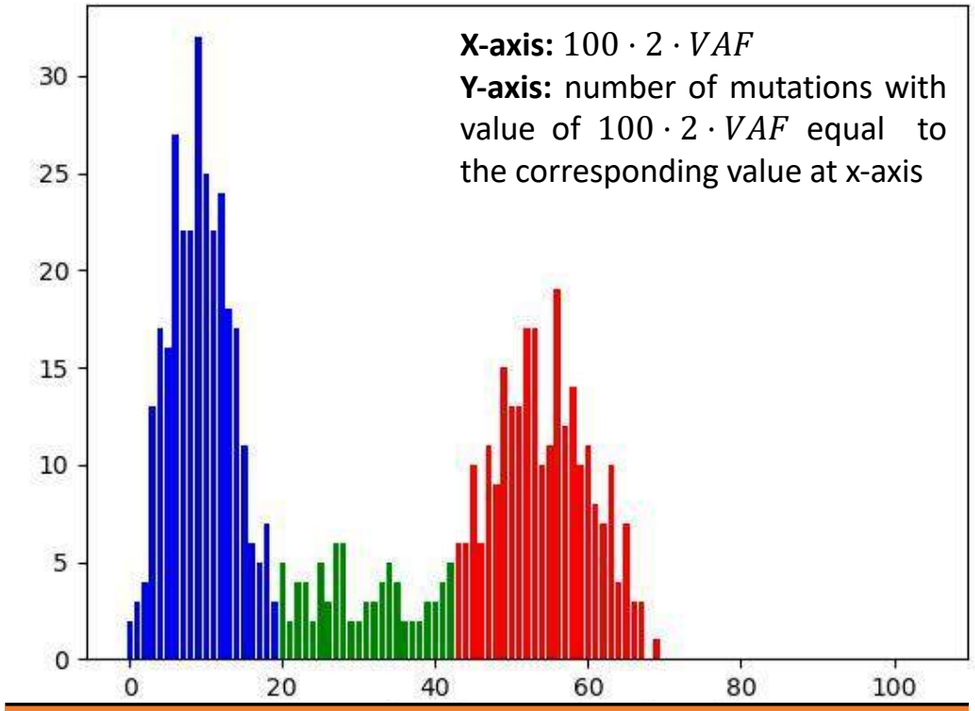
$\Rightarrow$ total number of cells harboring M_i is $\sim v_i$

Expected fraction of cells harboring M_i

$$\frac{\frac{v_i}{t_i}}{2} = \frac{2v_i}{t_i} = 2 \cdot VAF(M_i)$$

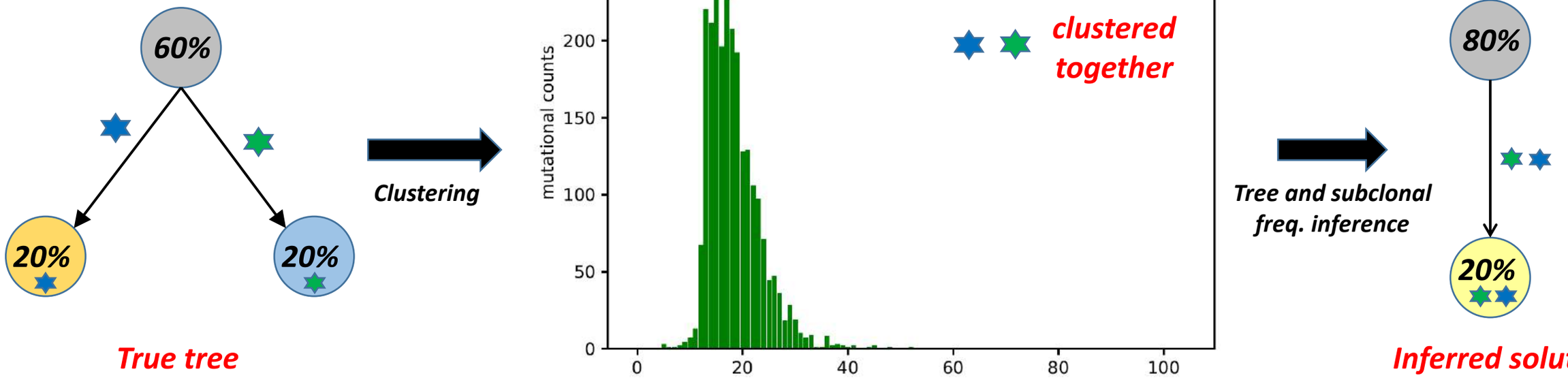
THE MAIN ASSUMPTIONS:

1. Mutations having similar $2 \cdot VAF(M_i)$ occur for the first time at the same cellular population.
2. The existence of clusters of mutations with similar $2 \cdot VAFs$.



ClusterID	Frequency	Mutations
Cluster1	10%	
Cluster 2	30%	
Cluster 3	55%	

Clustering ambiguity: subclones with similar cellular prevalence

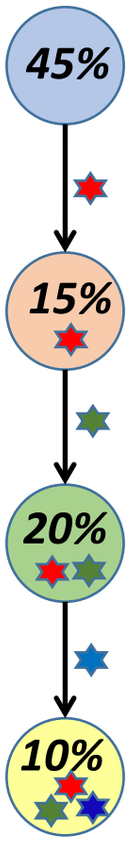


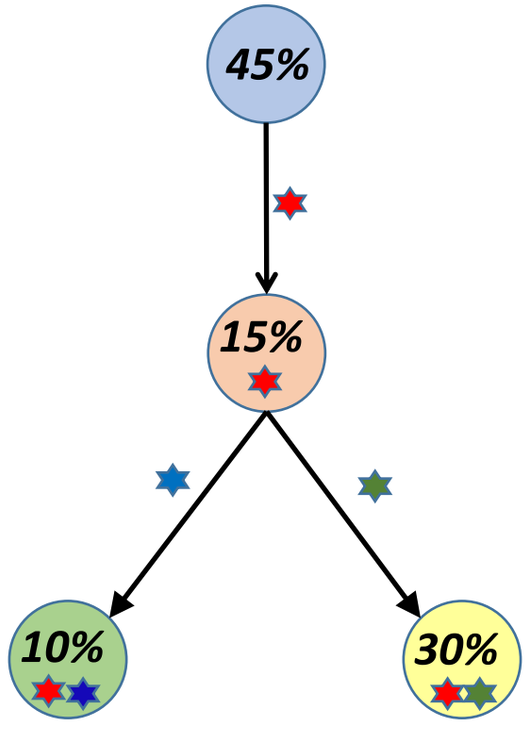
1. *During the clustering step mutations emerging at subclonal populations with similar cellular prevalence are clustered together*
2. *Inaccurate clustering influences the inference of subclonal prevalences and tree of tumor evolution*

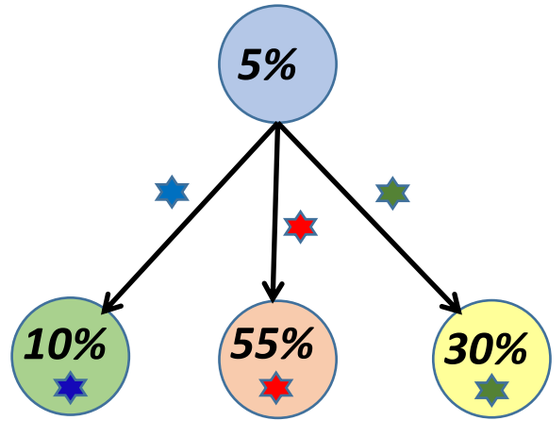
Phylogeny inference ambiguity: multiple equally likely trees

Cluster1	10%	
Cluster 2	30%	
Cluster 3	55%	

1. **Linear (chain) topology is usually among solutions**
2. **In many cases, in addition to linear topology, we also have other solutions with score equal to 0.**







Time Series Liquid Biopsy Data Can Help

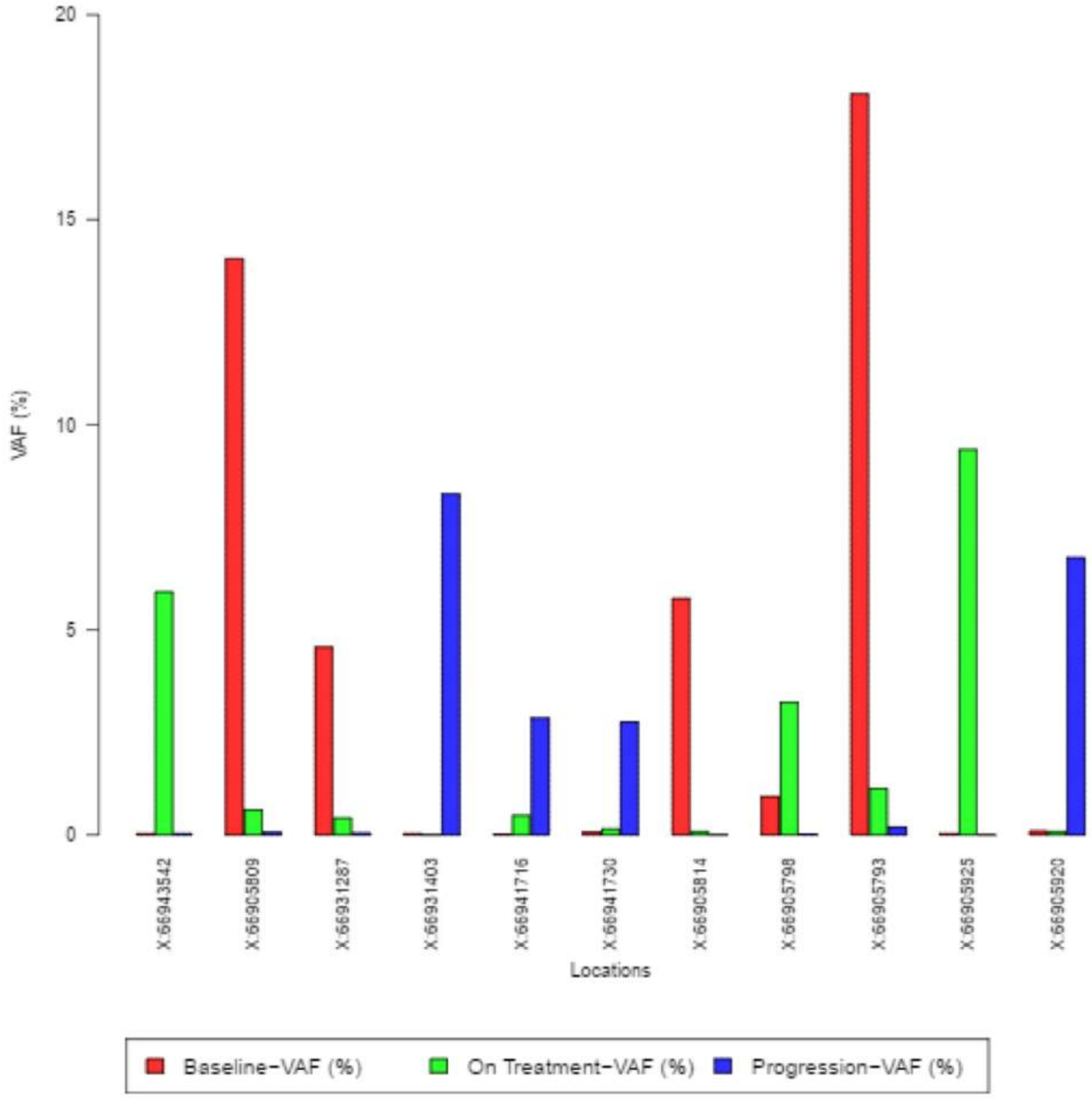
- 12 patients sequenced at three time points of interest

- Baseline, On-Treatment (12-weeks), and Progression

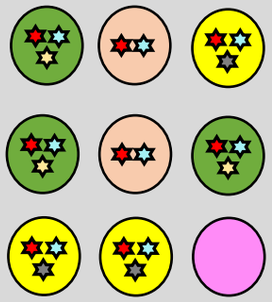
- Sensitively obtain mutation calls (through SiNVICT)

- For each mutation, check:
  - whether treatment has eliminated subclones,
  - whether new and more aggressive subclones emerged

Time Series Liquid Biopsy Data Can Help



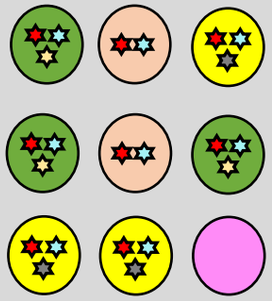
Studying tumor evolution by the use of single-cell sequencing (SCS) data



sequencing and mutation calling

mut\cell	cell1	cell2	…
M1	1	0	---
M2	0	1	---
M3	1	NA	---
⋮	⋮	⋮	⋮

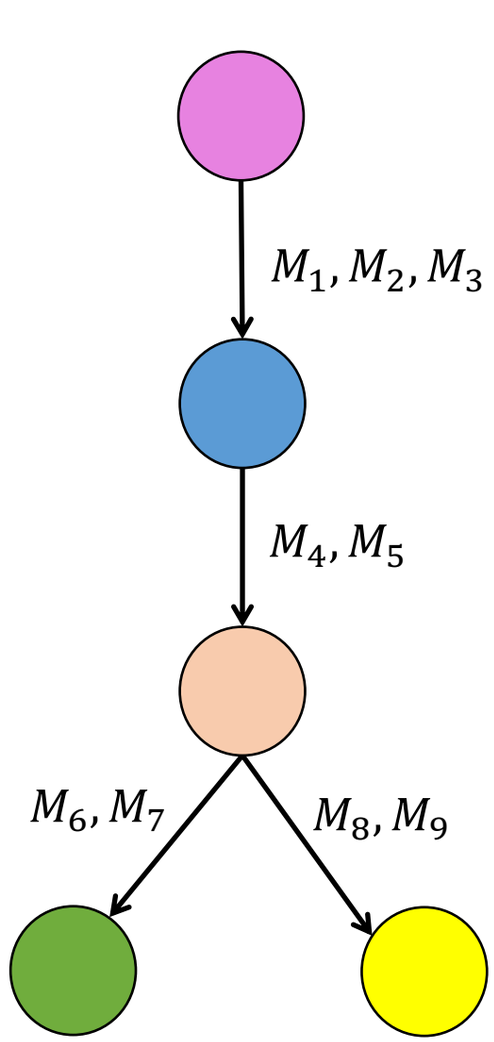
Studying tumor evolution by the use of single-cell sequencing (SCS) data



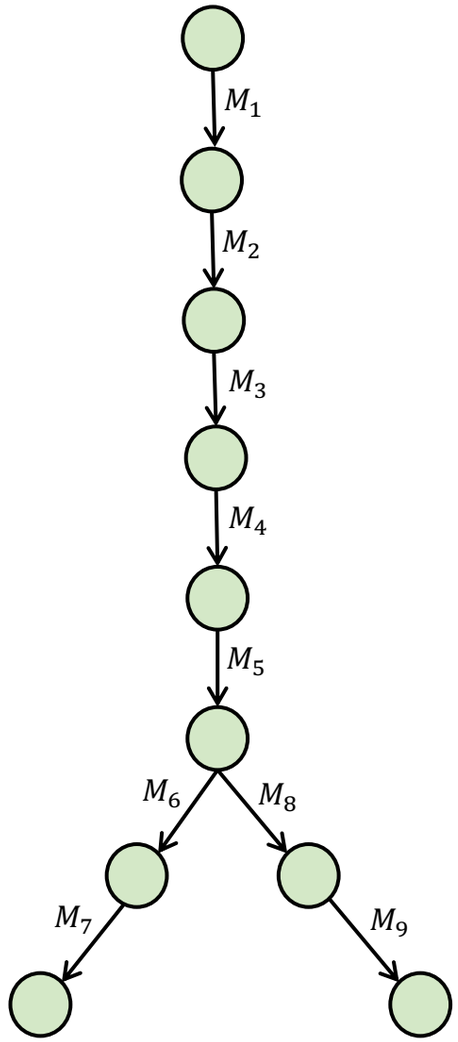
sequencing and mutation calling

mut\cell	cell1	cell2	…
M1	1	0	---
M2	0	1	---
M3	1	NA	---
⋮	⋮	⋮	⋮

Name of method	Authors	Journal	Year
-	Kim and Simon	BMC Bioinformatics	2013
BitPhylogeny	Yuan et al.	Genome Biology	2015
OncoNEM	Ross & Markowetz	Genome Biology	2016
SCITE	Jahn et al.	Genome Biology	2016
SiFit	Zafar et al.	Genome Biology	2017

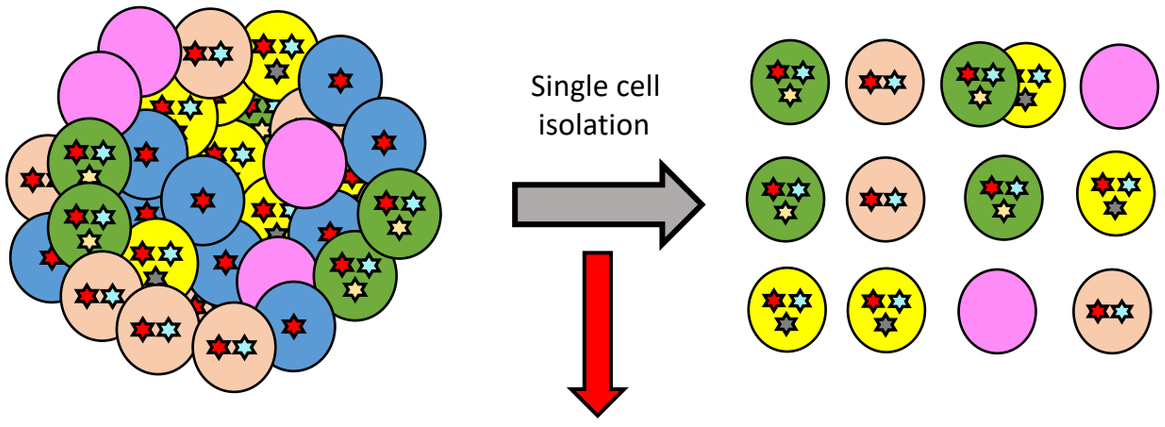


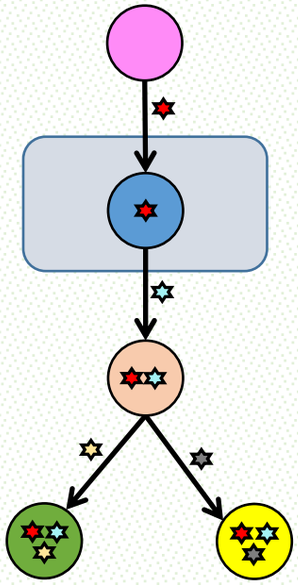
Clonal tree



Mutation tree

Single-cell sequencing data

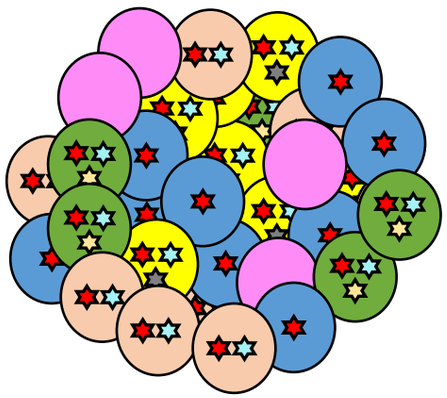




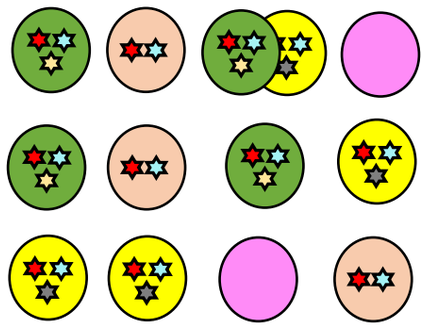
SINGLE-CELL ISOLATION

- Doublets
- Subclones with zero sampled cells
- Non-uniform sampling

Single-cell sequencing data



Single cell isolation

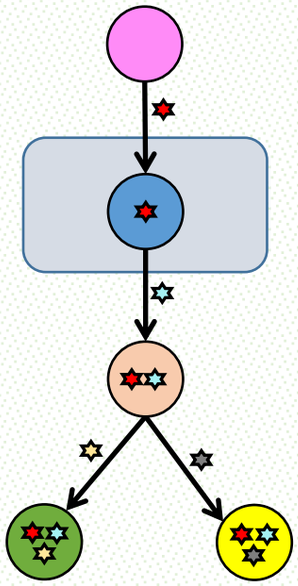


1. DNA extraction
2. DNA amplification
3. DNA sequencing
4. mutation calling

m cells / n mut.				...
M_1	1	**NA**	1	...
M_2	**0**	0	1	...
M_3	0	**1**	1	...
M_4	0	0	1	...
⋮	⋮	⋮	⋮	⋱

$= [M_1, M_2]$

$= [M_3, M_4]$



SINGLE-CELL ISOLATION

- Doublets
- Subclones with zero sampled cells
- Non-uniform sampling

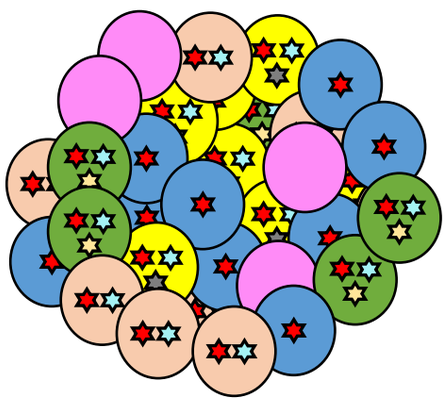
DNA AMPLIFICATION

- Amplification errors (false positives)
- Unequal amplification (false negatives, NAs)

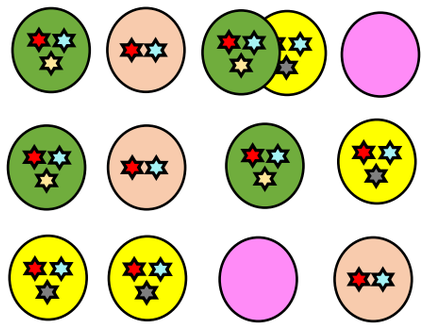
DNA SEQUENCING AND MUTATION CALLING

- False positives

Single-cell sequencing data



Single cell isolation

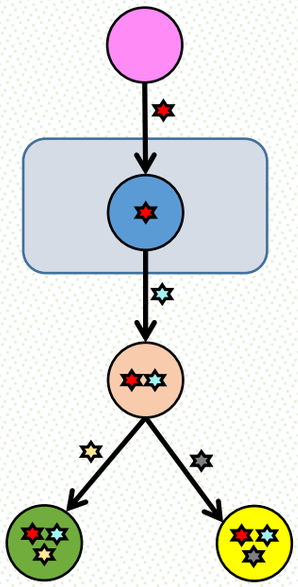


1. DNA extraction
2. DNA amplification
3. DNA sequencing
4. mutation calling

m cells / n mut.				...
M_1	1	**NA**	1	...
M_2	**0**	0	1	...
M_3	0	**1**	1	...
M_4	0	0	1	...
⋮	⋮	⋮	⋮	⋱

$= [M_1, M_2]$

$= [M_3, M_4]$



SINGLE-CELL ISOLATION

- Doublets
- Subclones with zero sampled cells
- Non-uniform sampling

DNA AMPLIFICATION

- Amplification errors (false positives)
- Unequal amplification (false negatives, NAs)

DNA SEQUENCING AND MUTATION CALLING

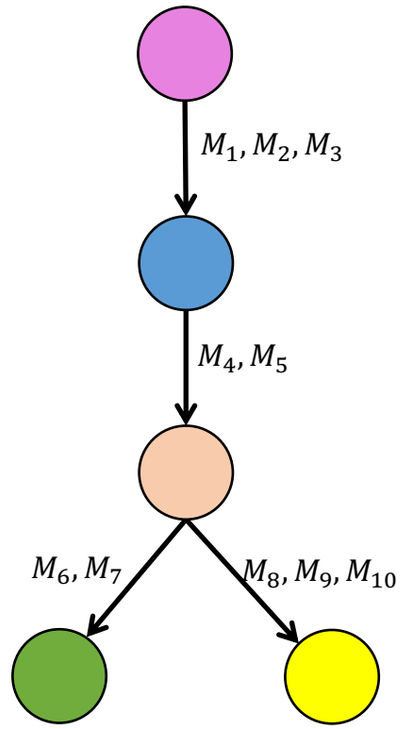
- False positives

Main types of noise in SCS data

1. **False positive (FP)**
   usually $\leq 10^{-5}$
2. **False negative (FN)**
   in the range $0.1 - 0.3$
3. **Missing entries (NA)**
   varies between 0.05-0.50
4. **Doublets**
   varies between 0 and 0.30

Tree inference by the use of SCS data - overview

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
M_1	0	**0**	1	1	**1**	1	1	**0**	1
M_2	**NA**	1	1	1	1	**0**	1	1	**0**
M_3	0	**0**	1	1	1	1	1	**0**	1
M_4	0	1	1	1	1	1	**0**	1	1
M_5	0	1	**0**	1	1	1	1	1	1
M_6	0	0	0	**NA**	1	1	0	0	0
M_7	0	0	0	1	1	**0**	0	0	0
M_8	0	0	**NA**	0	0	0	1	1	**0**
M_9	0	0	0	0	0	0	1	1	**0**
M_{10}	**1**	0	0	0	0	0	**NA**	1	1



$D_{n\times m}$ – single-cell data mutation matrix

n – number of mutations

m – number of cells

$(T, \boldsymbol{\theta})$

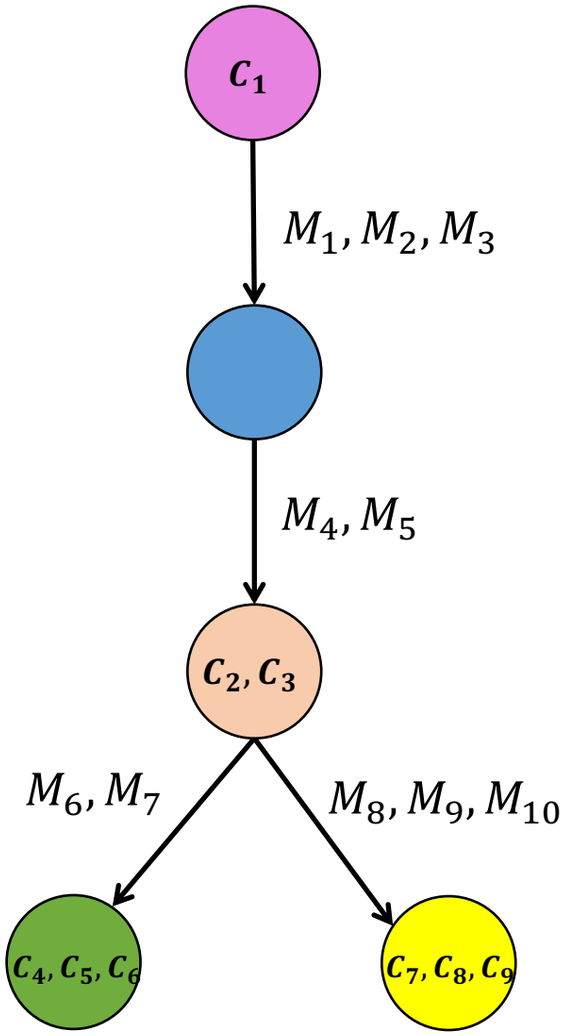
T – tree topology

$\boldsymbol{\theta} = (\alpha, \beta)$

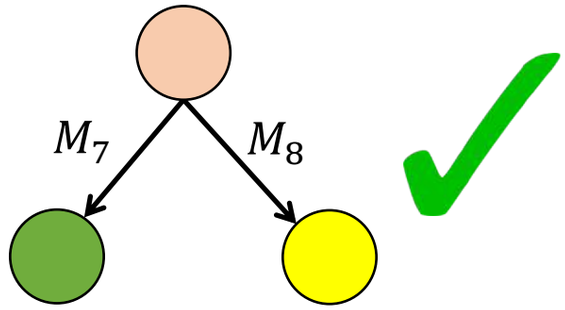
α – false positive rate

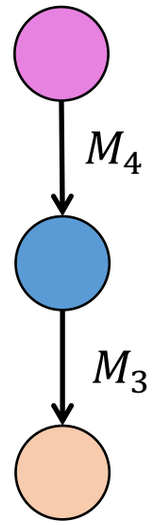
β – false negative rate

Strengths and weaknesses of SCS data



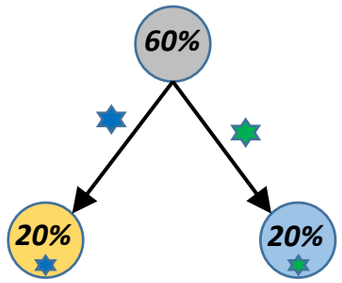
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
M_1	0	**0**	1	1	**1**	1	1	**0**	1
M_2	**NA**	1	1	1	1	**0**	1	1	**0**
M_3	0	**0**	1	1	1	1	1	**0**	1
M_4	0	1	1	1	1	1	**0**	1	1
M_5	0	1	**0**	1	1	1	1	1	1
M_6	0	0	0	**NA**	1	1	0	0	0
M_7	0	0	0	1	1	**0**	0	0	0
M_8	0	0	**NA**	0	0	0	1	1	**0**
M_9	0	0	0	0	0	0	1	1	**0**
M_{10}	**1**	0	0	0	0	0	**NA**	1	1





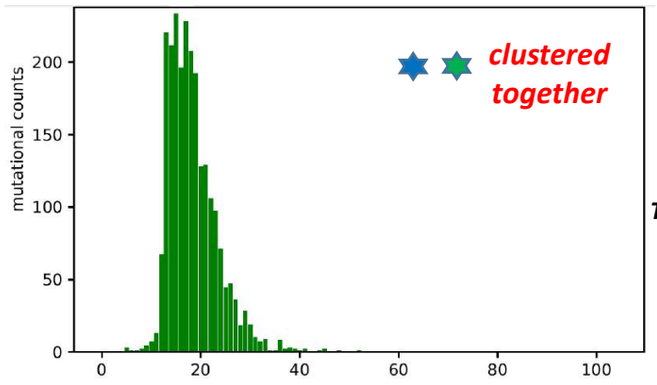
Advantages of combining bulk and SCS data

Bulk data



True tree

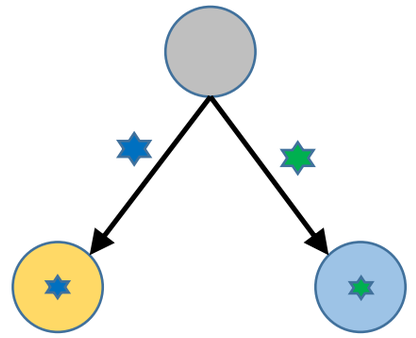
Clustering



Tree and subclonal freq. inference

Inferred solution

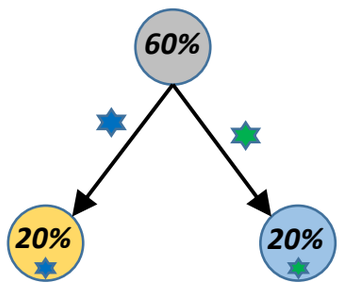
Single-cell data



Identifying branching helps in resolving clustering ambiguity

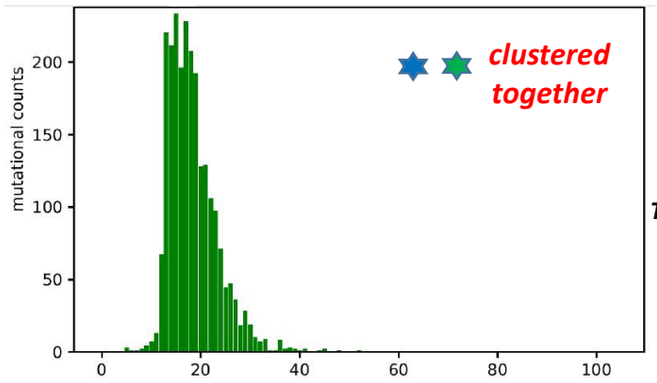
Advantages of combining bulk and SCS data

Bulk data



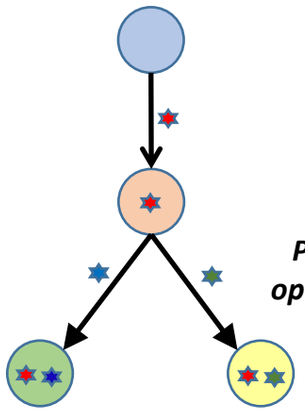
True tree

Clustering



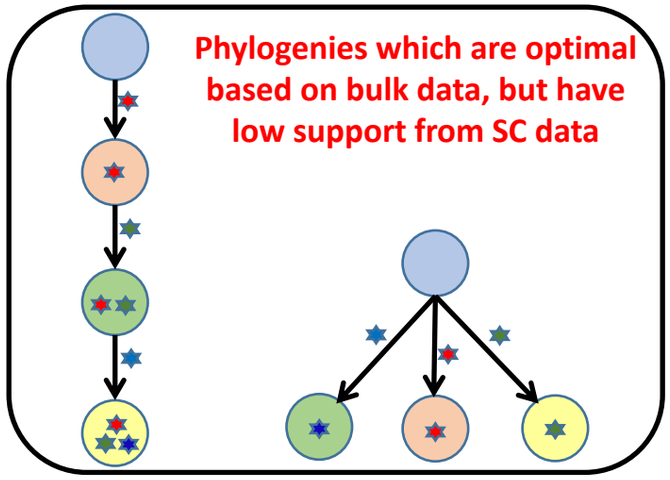
Tree and subclonal freq. inference

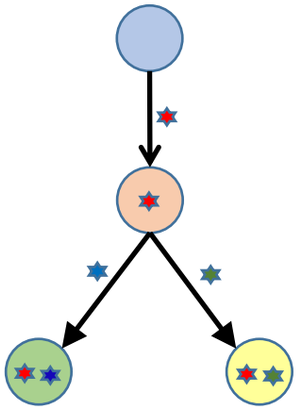
Inferred solution



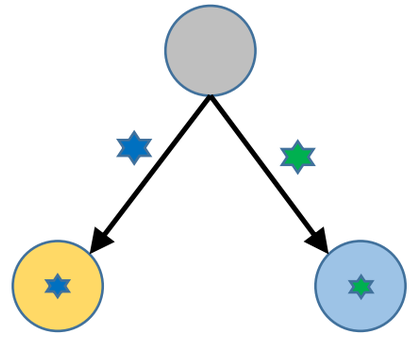
True tree

Possible set of optimal solutions





Single-cell data

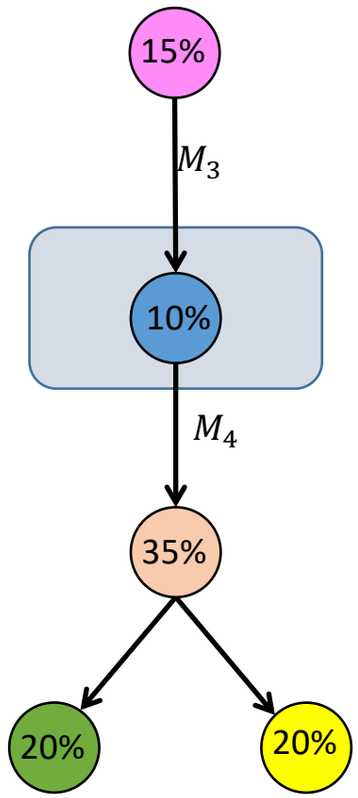


Identifying branching helps in resolving clustering ambiguity

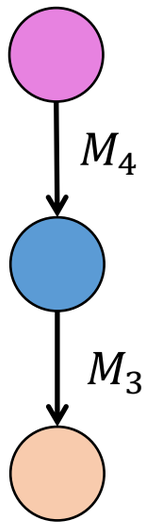
Many of the multiple topologies that are equally likely based on the bulk data have very low support from SC data

Advantages of combining bulk and SCS data

Single-cell data



True tree



Based on SC data *M_4 is more likely to occur before M_3*

Bulk data

$$2 \cdot VAF(M_3) = 0.85$$
$$2 \cdot VAF(M_4) = 0.75$$

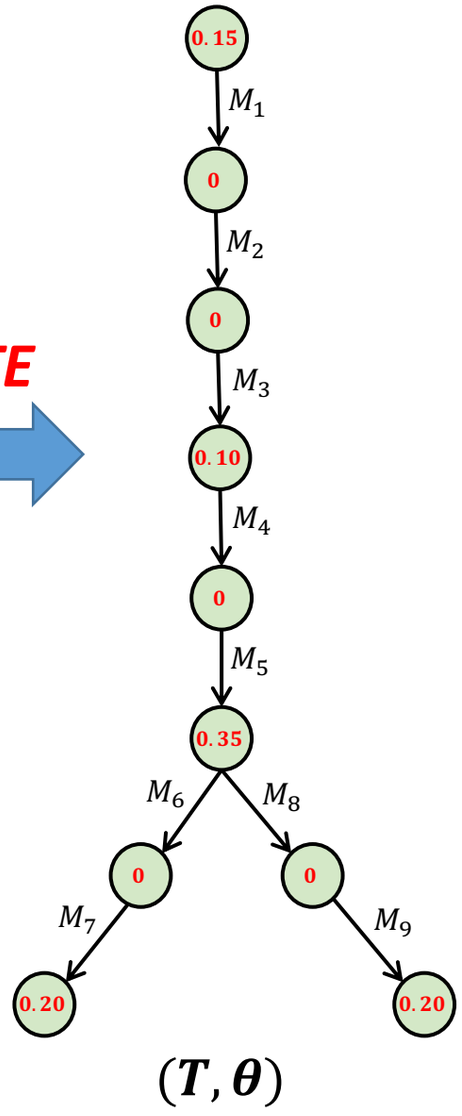
From bulk data we have that it is very likely that M_4 does not occur before M_3

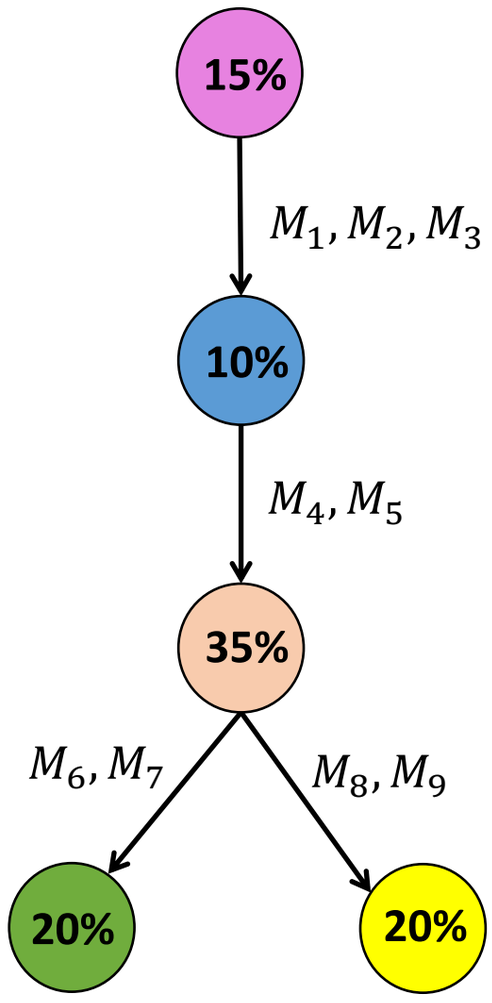
B-SCITE – input and output

Mutation	Variant reads	Reference reads
M_1	1100	2587
M_2	804	2710
M_3	537	3211
⋮	⋮	⋮

	C_1	C_2	C_3	C_4	C_5	C_8
M_1	0	0	1	1	1	…
M_2	NA	1	1	1	1	…
M_3	0	0	1	1	1	…
⋮	⋮	⋮	⋮	⋮	⋮	⋱

B-SCITE





INPUT DATA REQUIREMENTS:

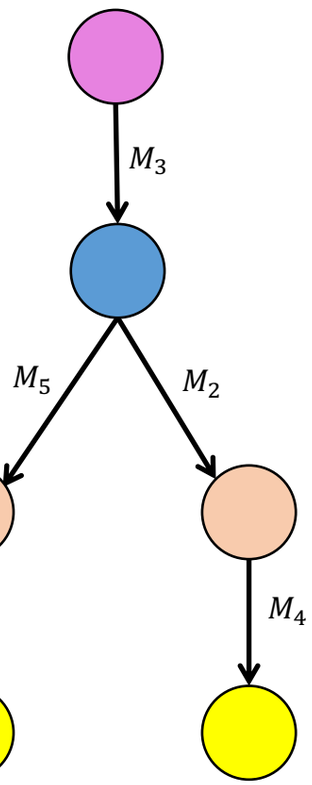
- SNVs from regions not affected by CNAs
- Consider only mutations present in at least one single cell
- Targeted deep sequencing (≥ 1000x) of bulk sample

Integrated SC and Bulk data score

$$\boldsymbol{S_{joint}}(\boldsymbol{T}^*, \boldsymbol{\theta}^*) = \underset{(T,\theta)}{\text{argmax}} \; [S_{SC}(T, \theta) + S_{bulk}(T)]$$

MCMC step

Current state $(\boldsymbol{T}, \boldsymbol{\theta})$

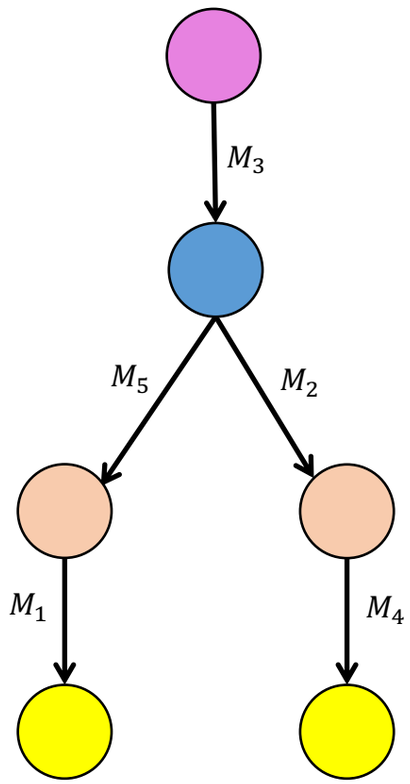


1. Propose (T', θ') state

First decide whether new T or new θ is proposed

MCMC step

Current state $(\boldsymbol{T}, \boldsymbol{\theta})$



1. Propose (T', θ') state

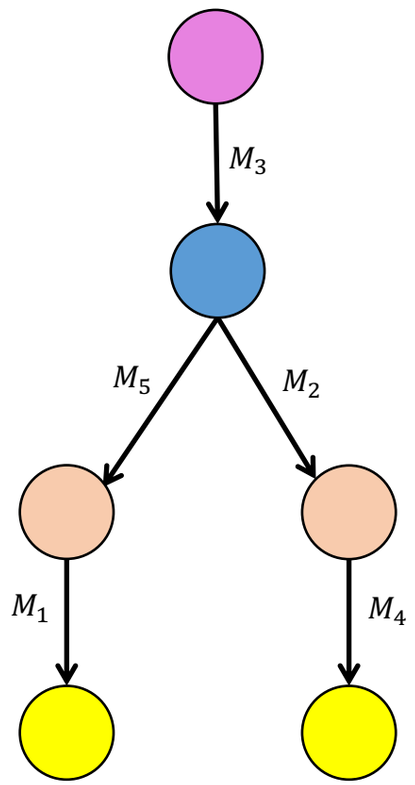
First decide whether new T or new θ is proposed

CASE 1: New θ is proposed

- $T' = T$
- $\theta' = (\alpha', \beta')$ is proposed via simple Gaussian walk
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$
  note that computation of bulk score is not required

MCMC step

Current state $(\boldsymbol{T}, \boldsymbol{\theta})$



1. Propose (T', θ') state

First decide whether new T or new θ is proposed

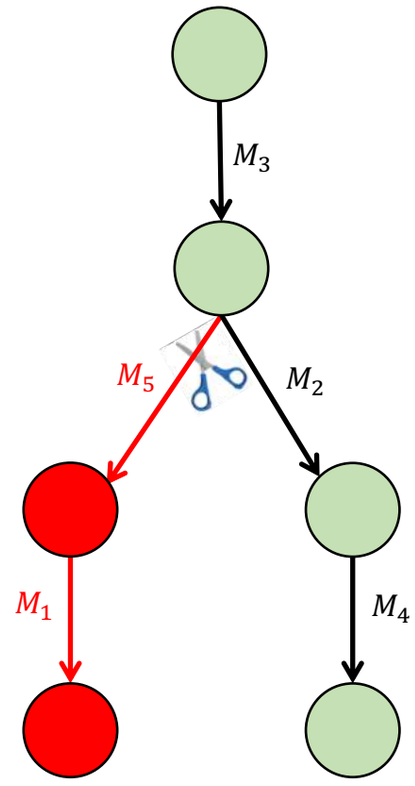
CASE 1: New θ is proposed

- $T' = T$
- $\theta' = (\alpha', \beta')$ is proposed via simple Gaussian walk
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$
  note that computation of bulk score is not required

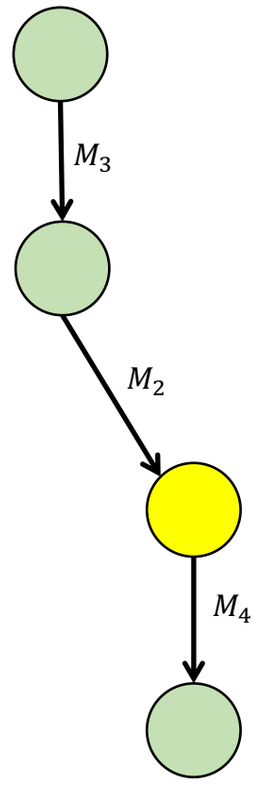
CASE 2: New T is proposed

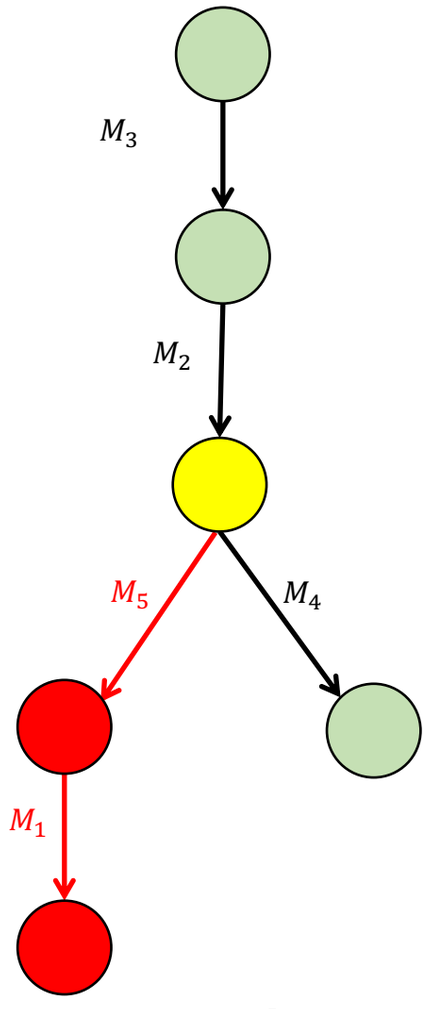
- T'= propose new mutation tree

An example of proposing new mutation tree



T

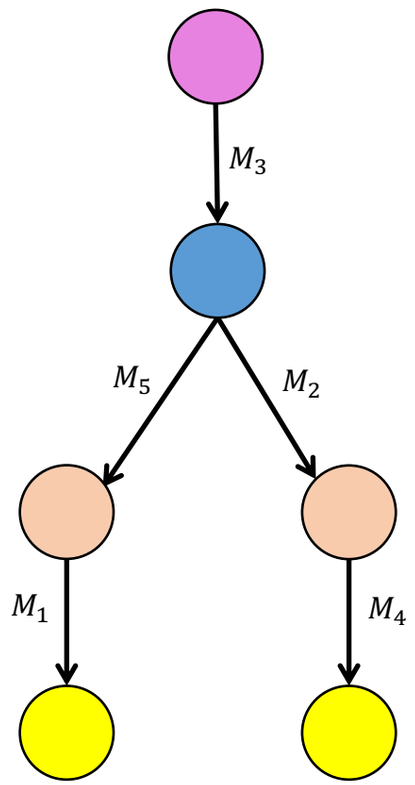




T'

MCMC step

Current state $(\boldsymbol{T}, \boldsymbol{\theta})$



1. Propose (T', θ') state

First decide whether new T or new θ is proposed

CASE 1: New θ is proposed

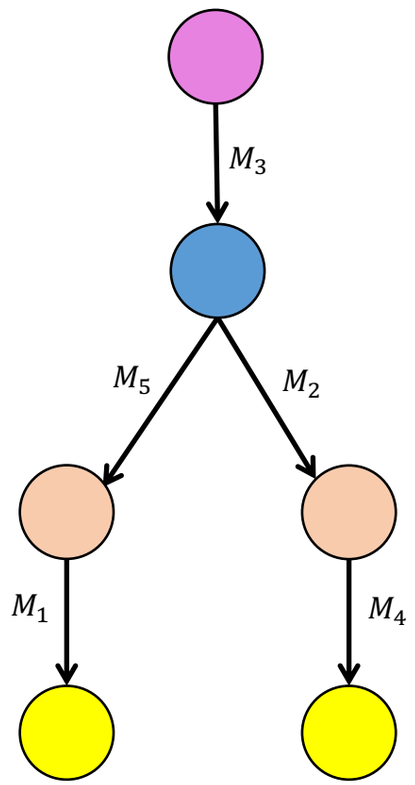
- $T' = T$
- $\theta' = (\alpha', \beta')$ is proposed via simple Gaussian walk
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$
  note that computation of bulk score is not required

CASE 2: New T is proposed

- T'= propose new mutation tree
- $\theta' = \theta$
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$

MCMC step

Current state $(\boldsymbol{T}, \boldsymbol{\theta})$



1. Propose (T', θ') state

First decide whether new T or new θ is proposed

CASE 1: New θ is proposed

- $T' = T$
- $\theta' = (\alpha', \beta')$ is proposed via simple Gaussian walk
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$
  note that computation of bulk score is not required

CASE 2: New T is proposed

- T'= propose new mutation tree (steps described later)
- $\theta' = \theta$
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$

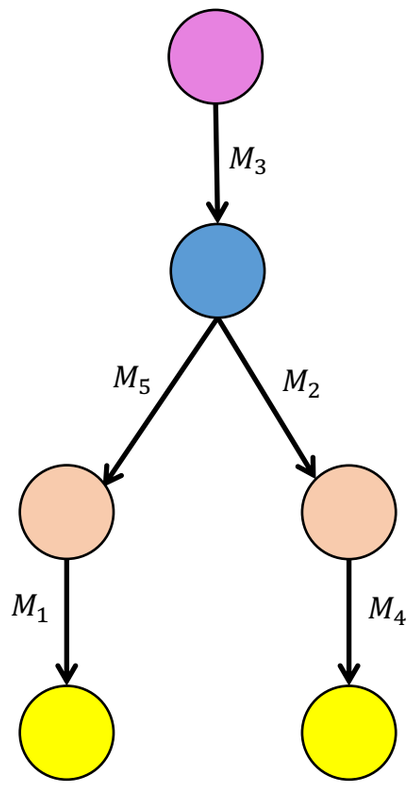
2. Accept or decline proposed (T', θ')

Accept the proposed (T', θ') with the probability

$$\min\left\{1, \frac{q(T, \theta \mid T', \theta')P(T', \theta' \mid D)}{q(T', \theta' \mid T, \theta)P(T, \theta \mid D)}\right\}$$

MCMC step

Current state $(\boldsymbol{T}, \boldsymbol{\theta})$



If move accepted
$(T, \theta) \rightarrow (T', \theta')$

1. Propose (T', θ') state

First decide whether new T or new θ is proposed

CASE 1: New θ is proposed

- $T' = T$
- $\theta' = (\alpha', \beta')$ is proposed via simple Gaussian walk
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$
  note that computation of bulk score is not required

CASE 2: New T is proposed

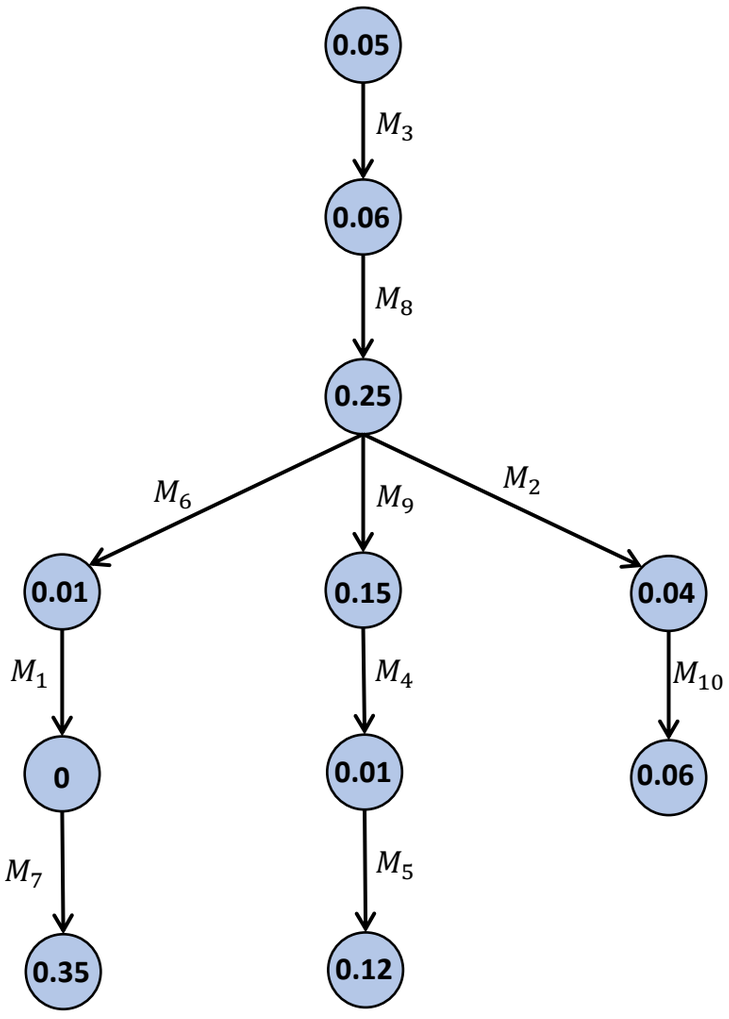
- T'= propose new mutation tree (steps described later)
- $\theta' = \theta$
- Compute $S_{joint}(T', \theta') = S_{SC}(T', \theta') + S_{bulk}(T')$

2. Accept or decline proposed (T', θ')

Accept the proposed (T', θ') with the probability

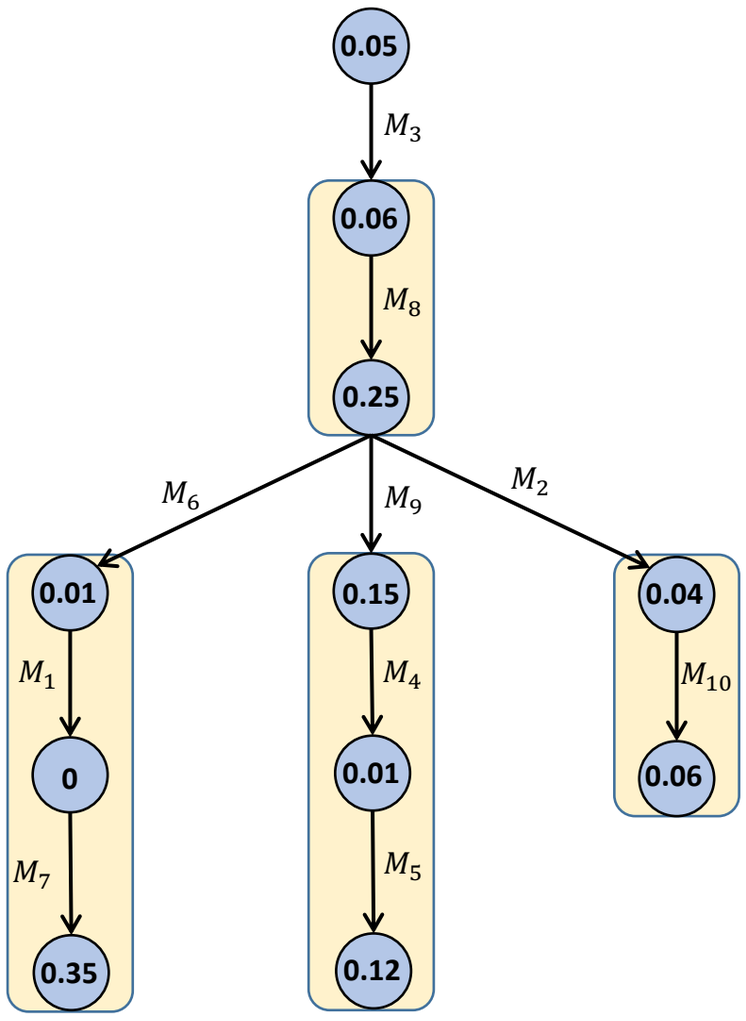
$$\min\left\{1, \frac{q(T, \theta \mid T', \theta')P(T', \theta' \mid D)}{q(T', \theta' \mid T, \theta)P(T, \theta \mid D)}\right\}$$

Mutation trees → clonal trees



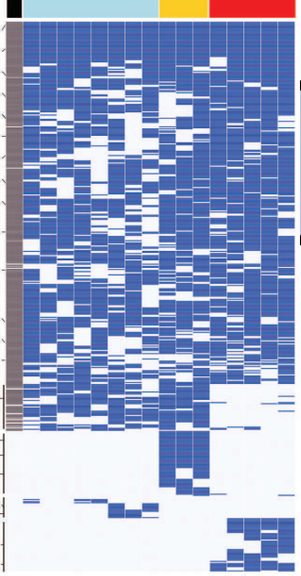
Inferred mutation tree

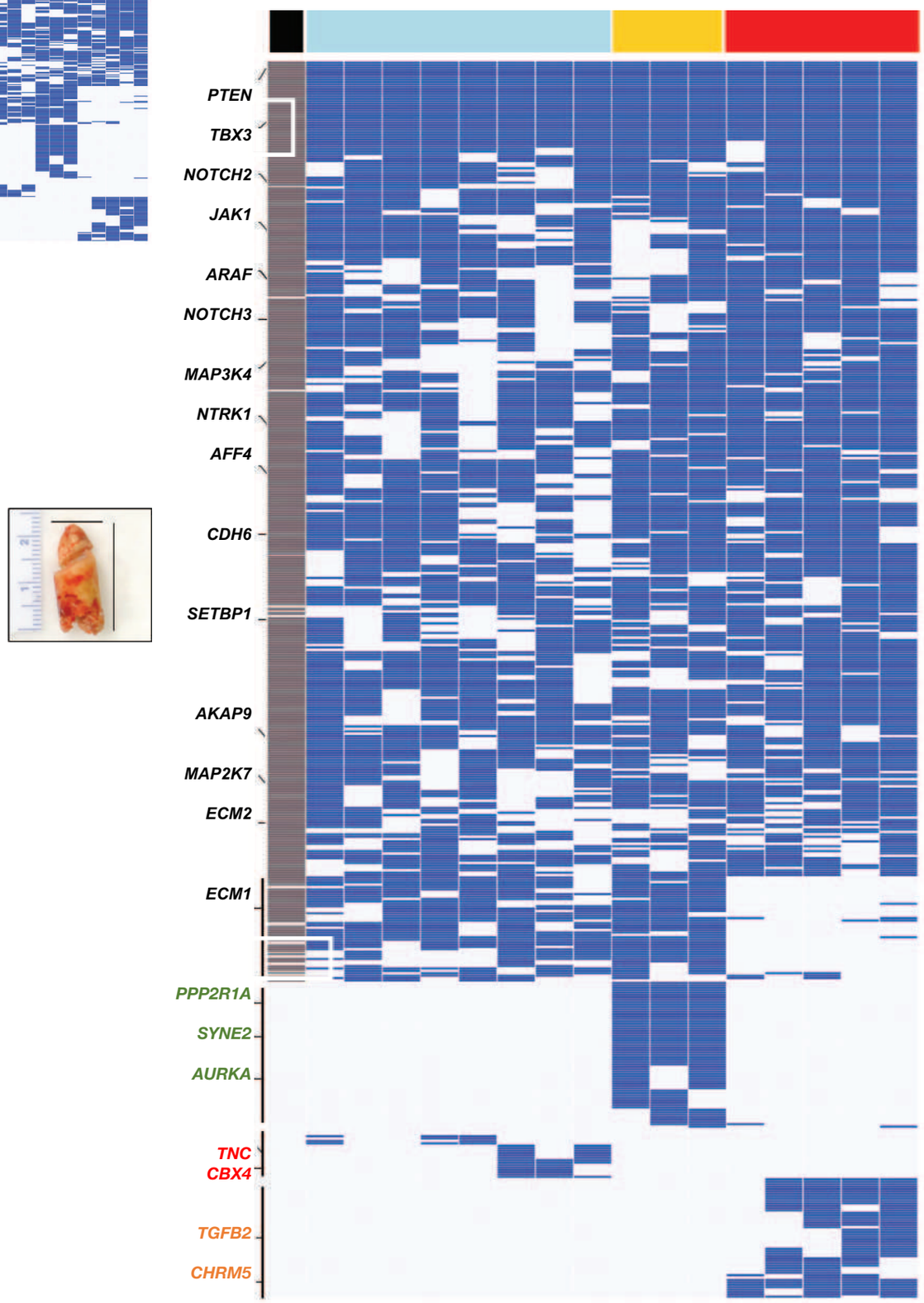
Mutation trees → clonal trees

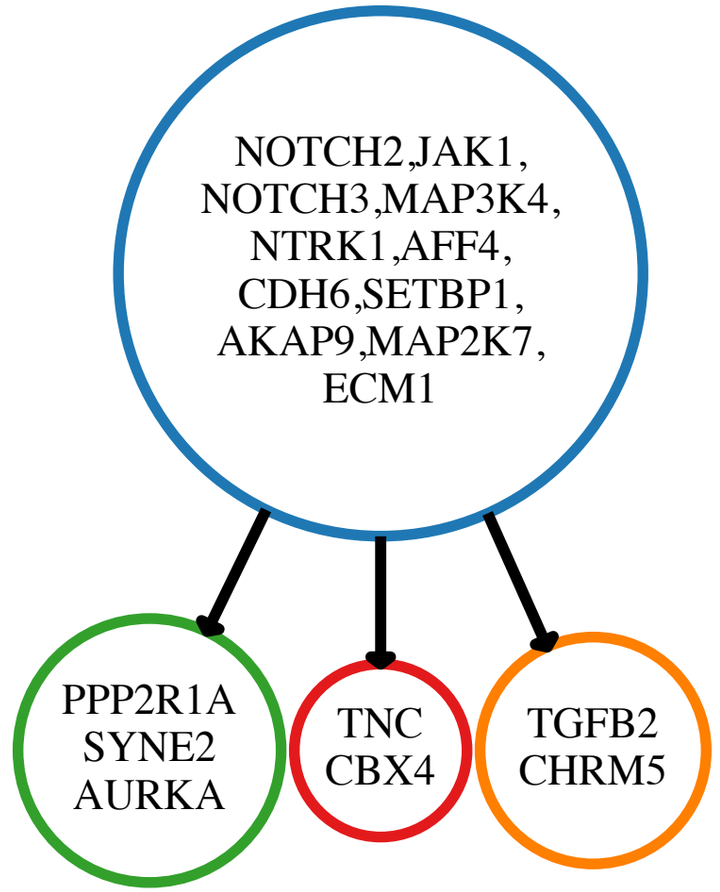


Inferred mutation tree

B-SCITE applied to Triple Negative BC

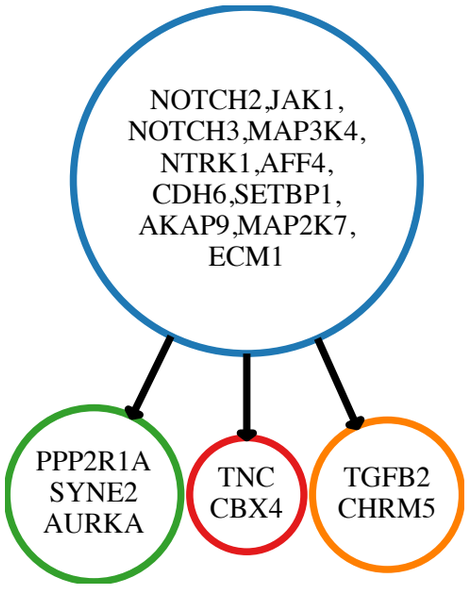




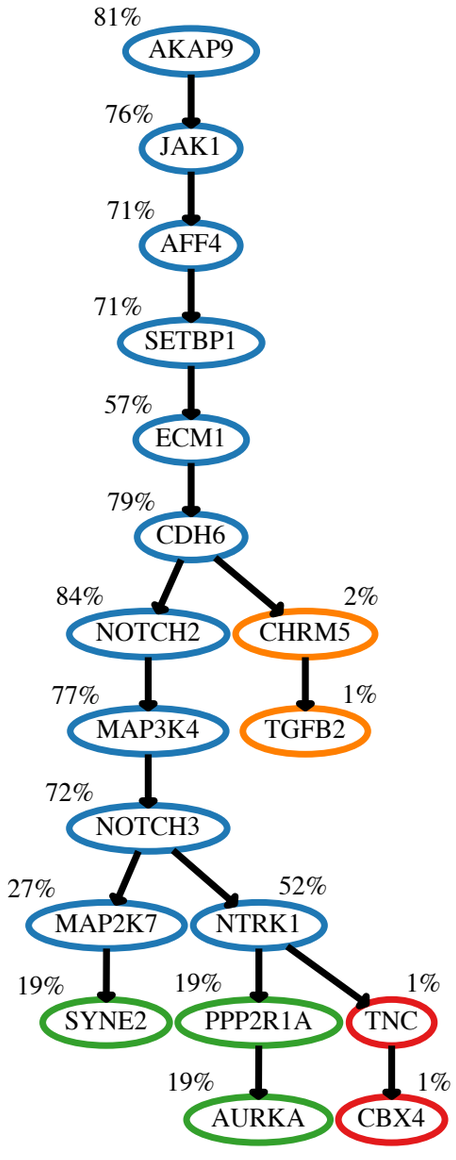


Ground truth inferred by Wang et al. 2014: on a TNBC specimen: three subclones inferred through hierarchical clustering of 374 mutations extracted from 144 single cell whole exome sequence data

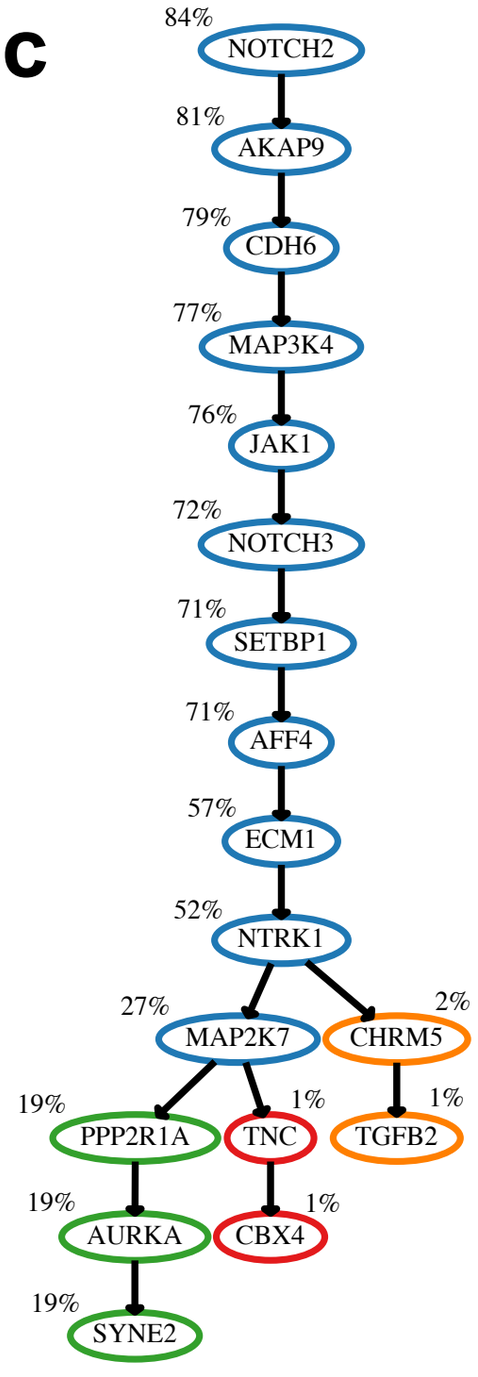
a



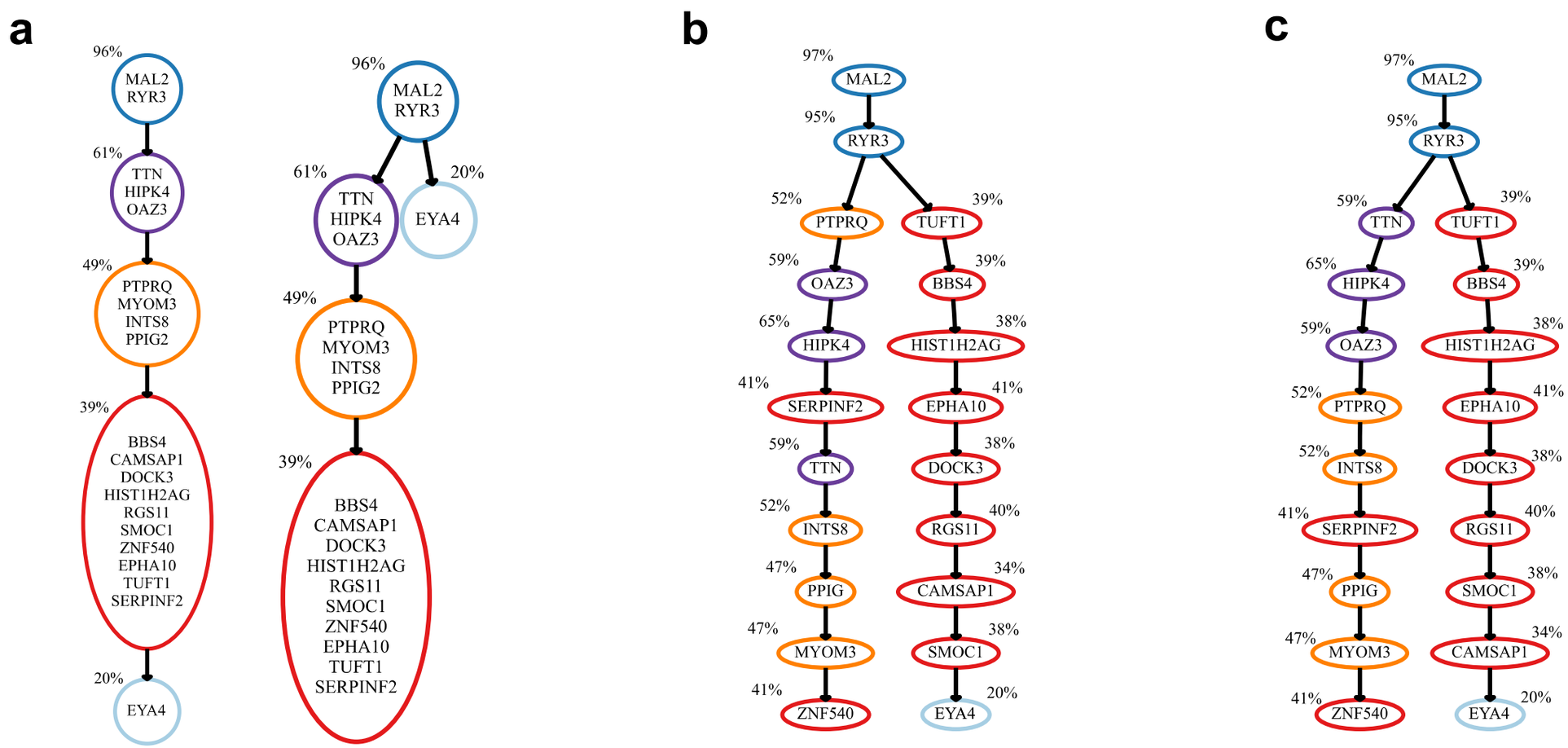
b



c



Results for TNBC Patient 1 from Wang et al. 2014: Input data consists of 72 single-cells and 18 mutations.
(a) Ground truth tree from Wang et al. 2014. **(b)** Tree obtained by SCS data only. **(c)** Tree reported by B-SCITE.



Results for ALL Patient 1 from Gawad et al. 2014: (a) trees obtained by clustering bulk-data read counts (coverage ~ 2000).
(b) Tree obtained by SCS data only **(c)** Tree reported by B-SCITE. Input data consists of 111 single-cells and 20 mutations.

Conclusions

- Methods to infer clonal trees of evolution by the use of single (CTP- Single), multi-site (CITUP) bulk and integrated single-cell (B-SCITE) sequencing data
- Robust to the presence of various types of noise in both types of input data
- Achieve high accuracy, including on tumors consisting of tens of subclones
- Extend to the cases with multiple bulk-sequencing data
- Outperform existing methods on all measures of accuracy

More details: https://www.biorxiv.org/content/early/2017/12/15/234914 (Malikic et al. RECOMB 2018)

CTP-Single & CITUP available at https://github.com/nlgndnmz/CTPsingle

B-SCITE available at https://github.com/smalikic/B-SCITE

ReMixT:
Reconstructing Clone Specific Genomic Structure in Heterogeneous Tumor Samples via Bulk WGS

SIMON FRASER UNIVERSITY

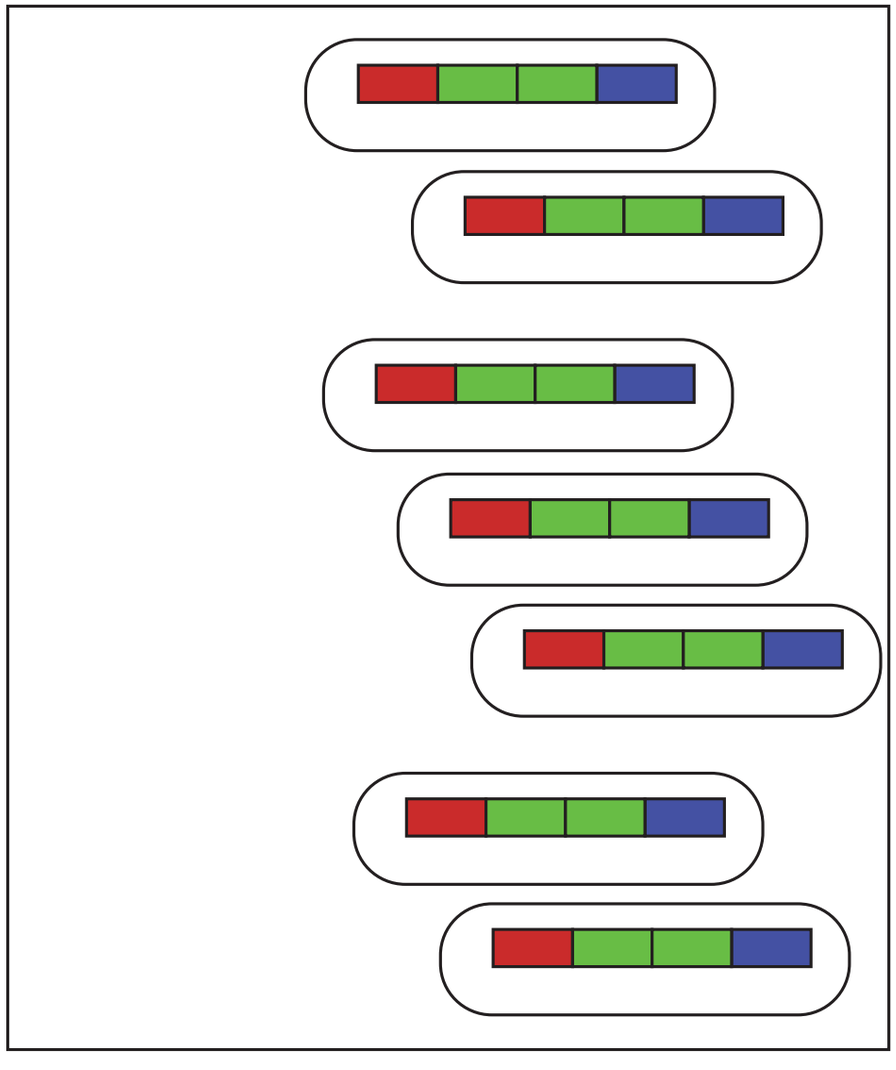
VANCOUVER PROSTATE CENTRE
A UBC & VGH Centre of Excellence

BC Cancer Agency
CARE + RESEARCH

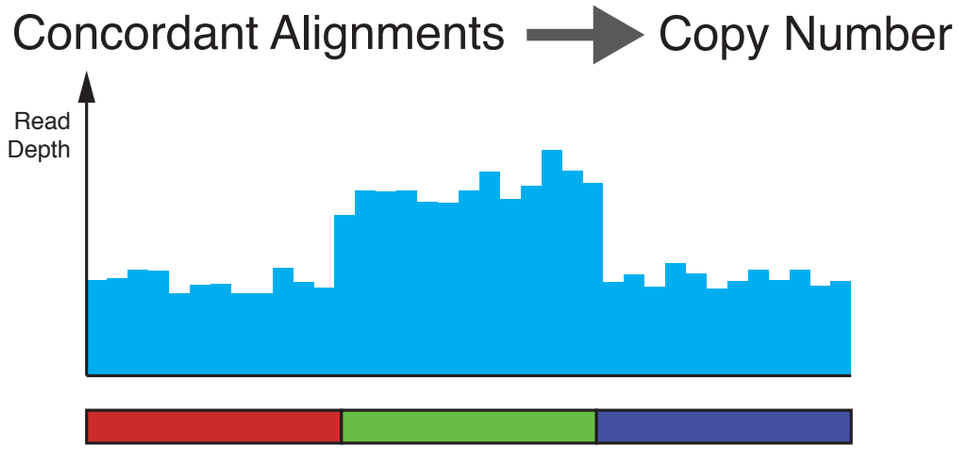
INDIANA UNIVERSITY

Segment Copy Number Change Evident in Whole Genome Sequencing Read Depths

Tumour Sample

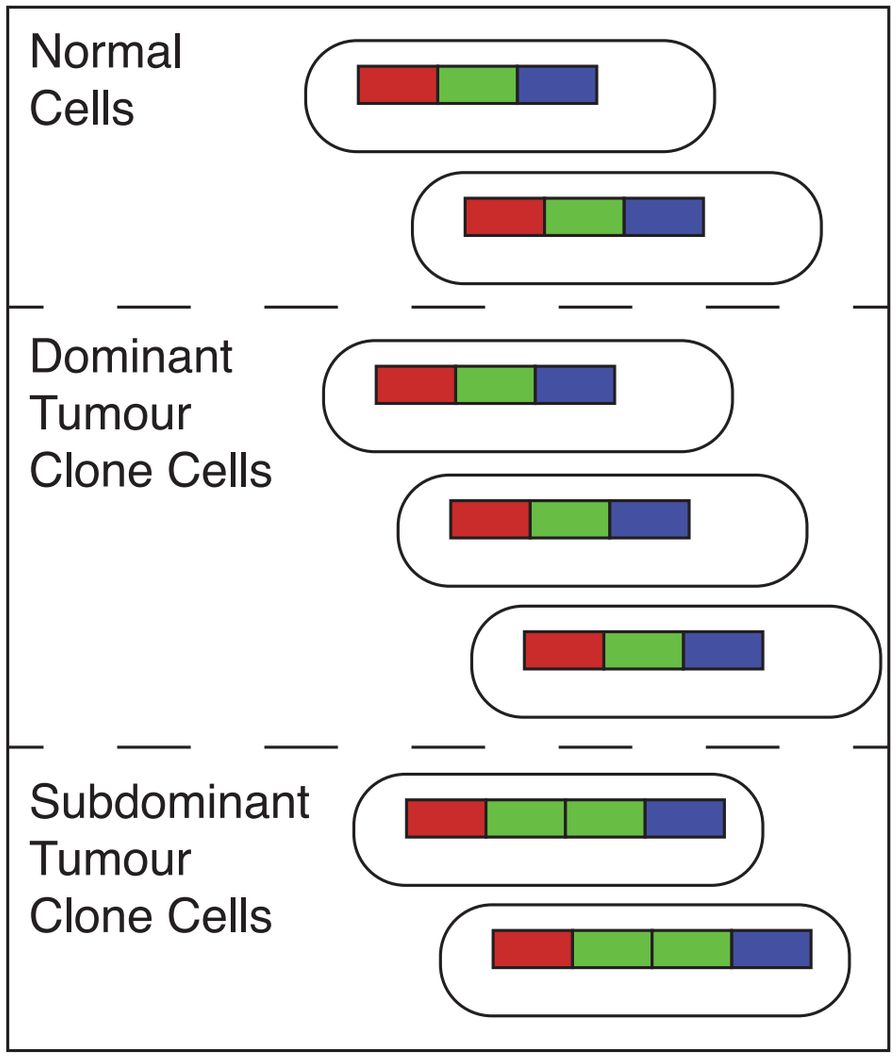


Whole Genome Sequence Data

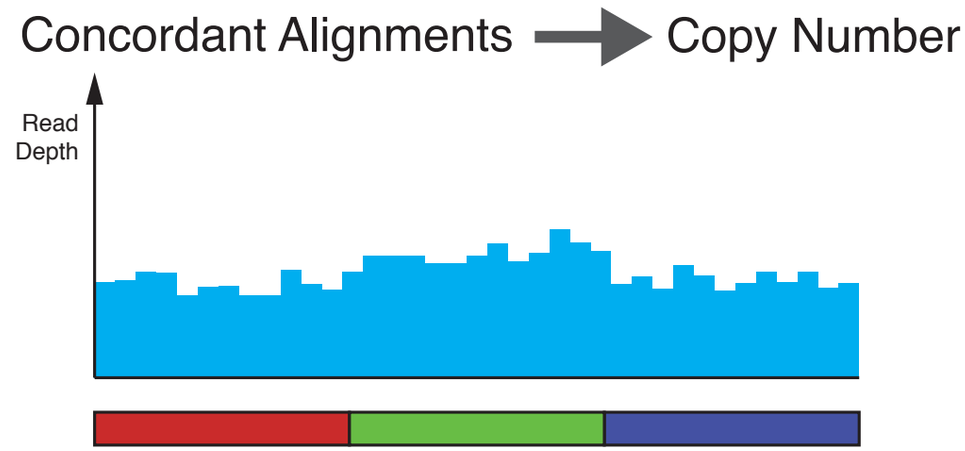


Normal Contamination and Clonal Diversity Dilute the Signal of Copy Number Changes

Tumour Sample

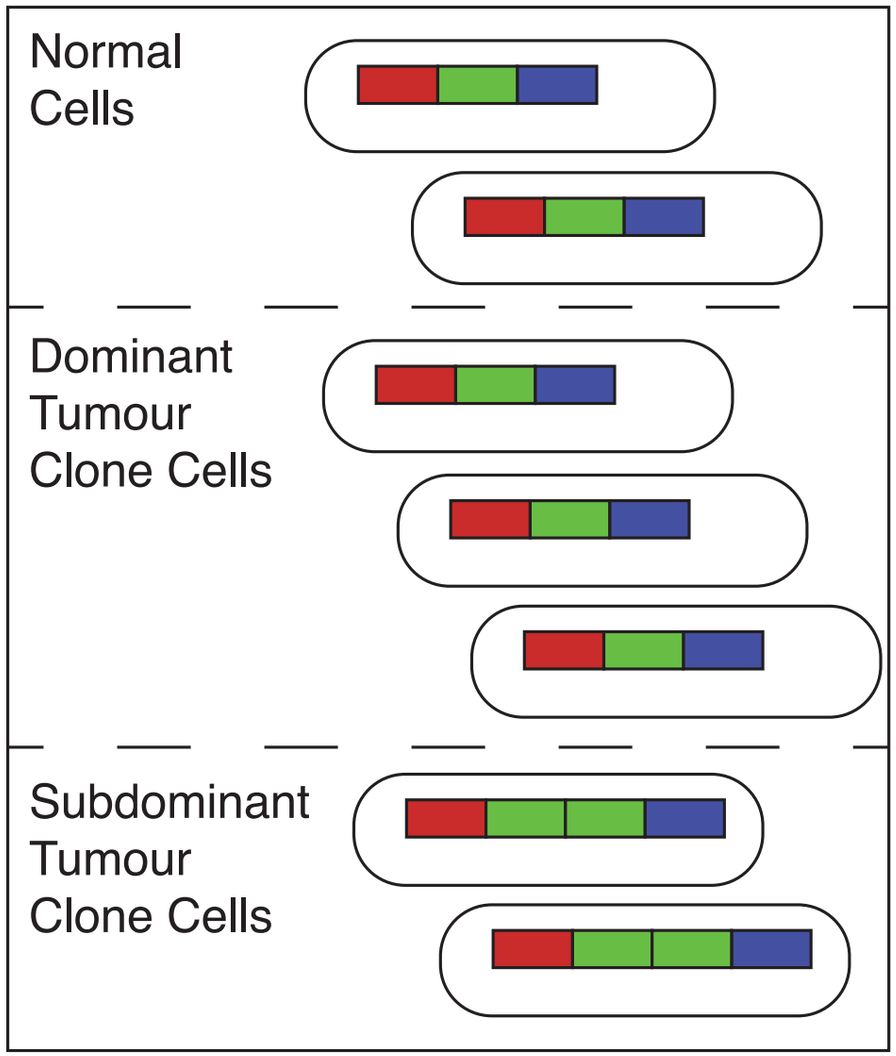


Whole Genome Sequence Data

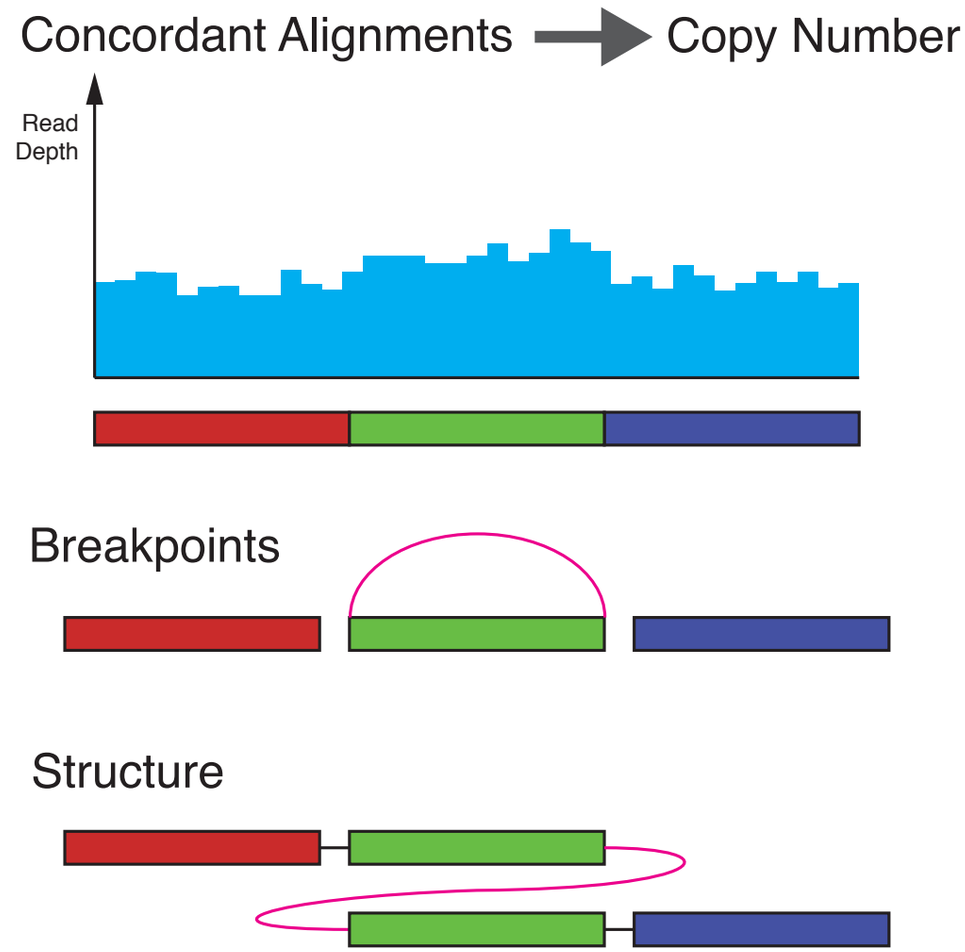


Joint Analysis to Increase Statistical Strength for Identifying Subclonal Copy Number Changes

Tumour Sample

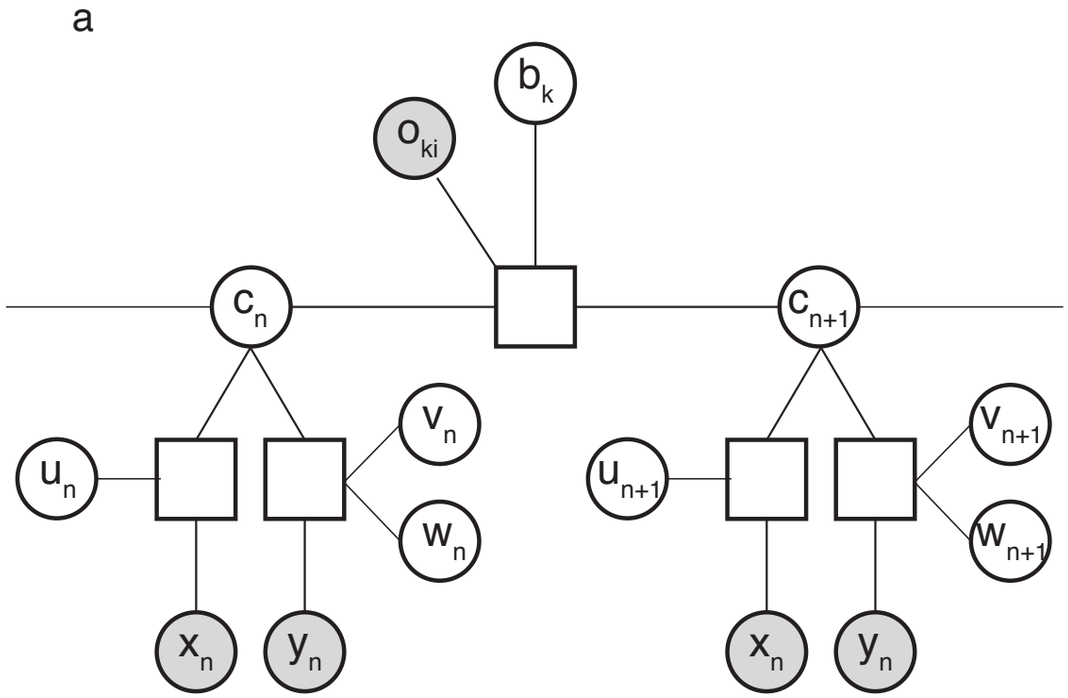


Whole Genome Sequence Data



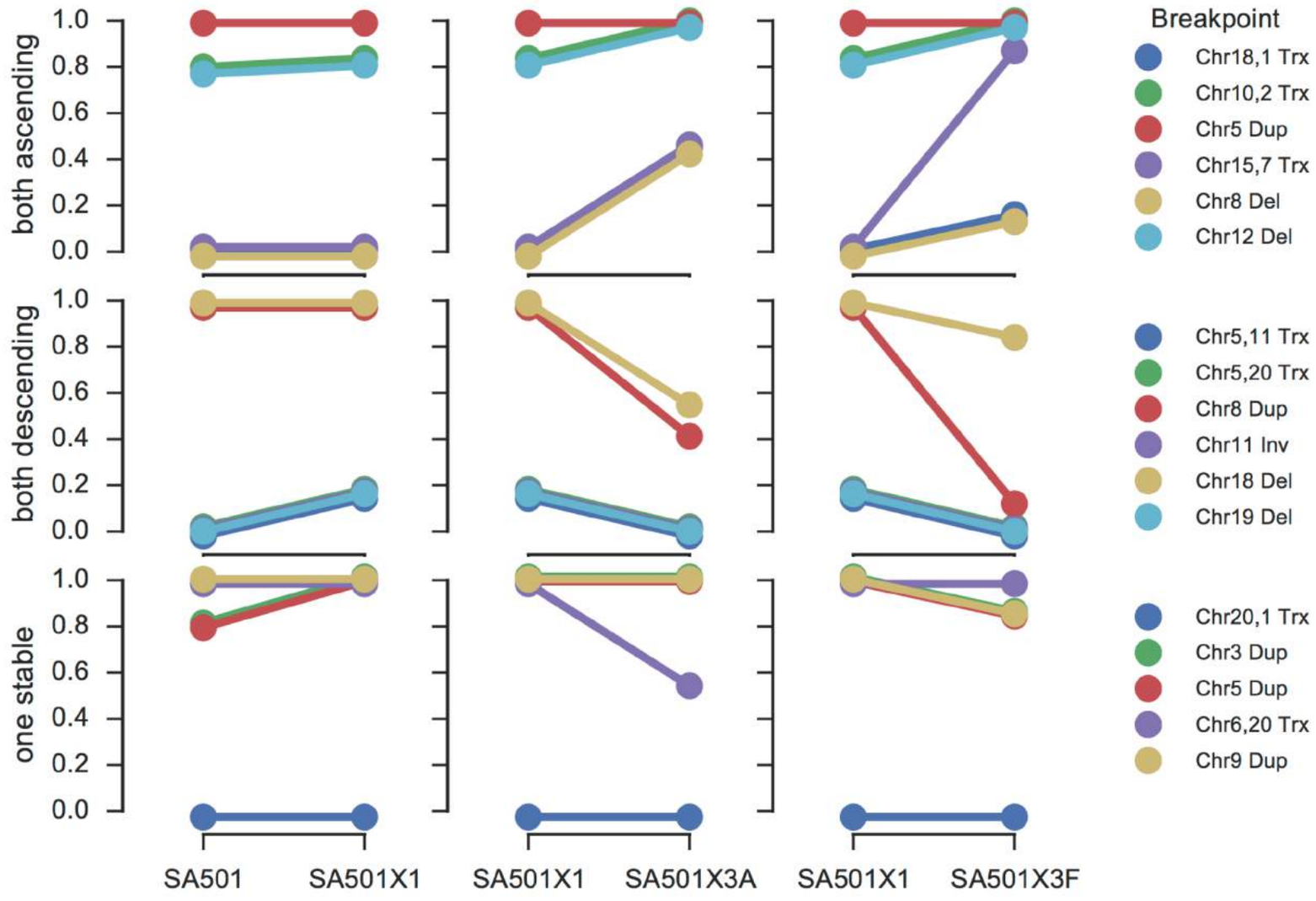
ReMixT: Probabilistic Genome Graph Model

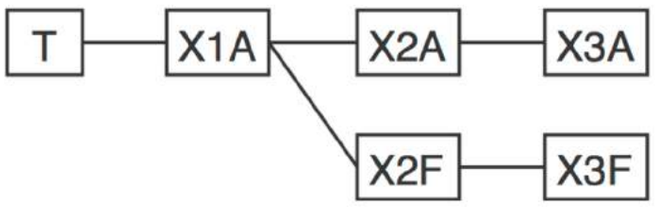
- Allele and clone specific copy number model
  - HMM augmented with breakpoint dependencies
  - Unified state space
  - O(K^2) transition calculations by exploiting symmetry
  - Outlier modeling
  - Allele uncertainty modelling



ReMixT: Reproducible Clonal Dynamics in Replicate Xenografts

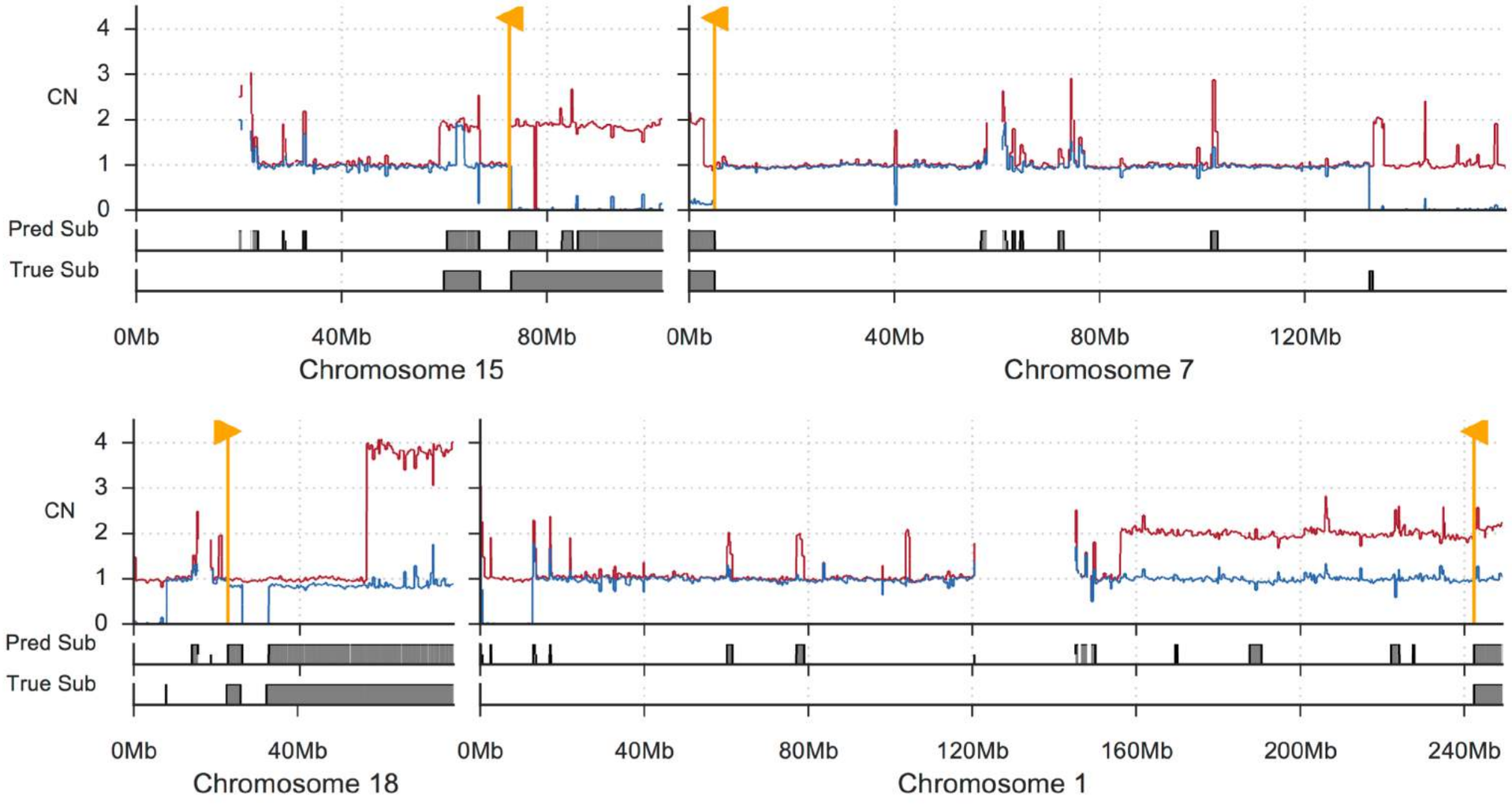
- Cellular Frequency of Breakpoints recapitulates SNV clonal dynamics
- All breakpoints either ascend or descend between through successive passaging





ReMixT: Validation with SA501X3F Whole Genome Single Cell

- Single cell data validates subclonal segments associated with clone specific breakpoints



Conclusion

- ReMix-T simultaneously infers clone specific breakpoints and associated copy number alterations in a heterogeneous tumor sample from bulk sequencing data.
- Can predict breakpoint and associated subclonal frequencies

More details at: McPherson A. et al. Genome Biology 2017

Remix-T available at http://bitbucket.org/dranew/remixt.

Current/Future Directions

- Exact solutions for SCS+Bulk HTS based on ILP and CSP instead of MCMC
- Perfect phylogeny with infinite sites model to be replaced with Dollo parsimony
- Clone specific SNV + SV + CNV composition of heterogeneous tumor samples from integrated bulk and single cell sequencing data
- The use of long read/single molecule sequencing technologies to associate two or more breakpoints for better structural inference
- Algorithms that can scale up to accommodate thousands of single cell WGS data
- Integration with liquid biopsy sequencing