gpps: an ILP-based approach for inferring cancer progression with mutation losses from single cell data

Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha and Paola Bonizzoni

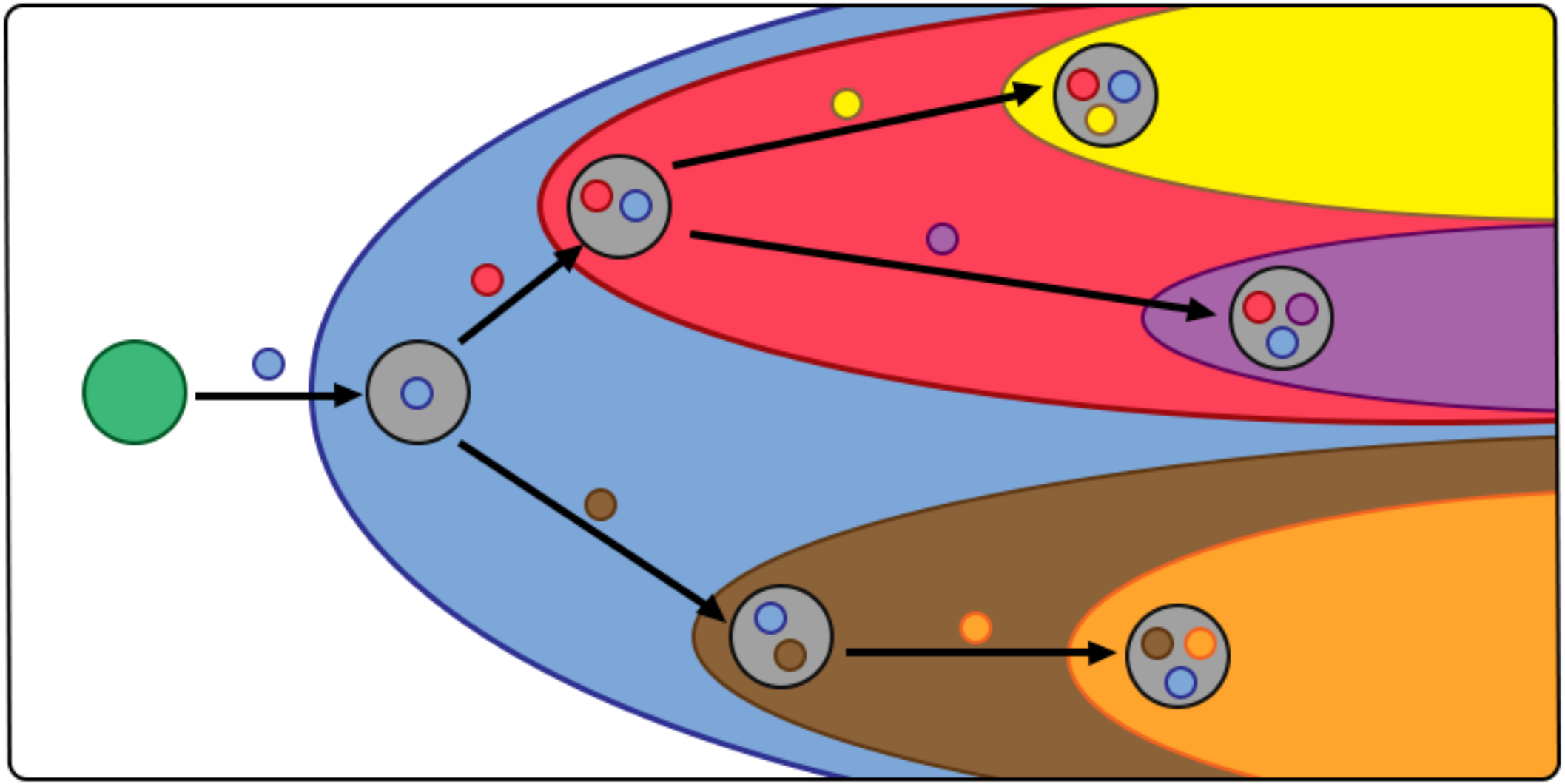
ICCABS, Las Vegas 2018

Weill Cornell Medicine

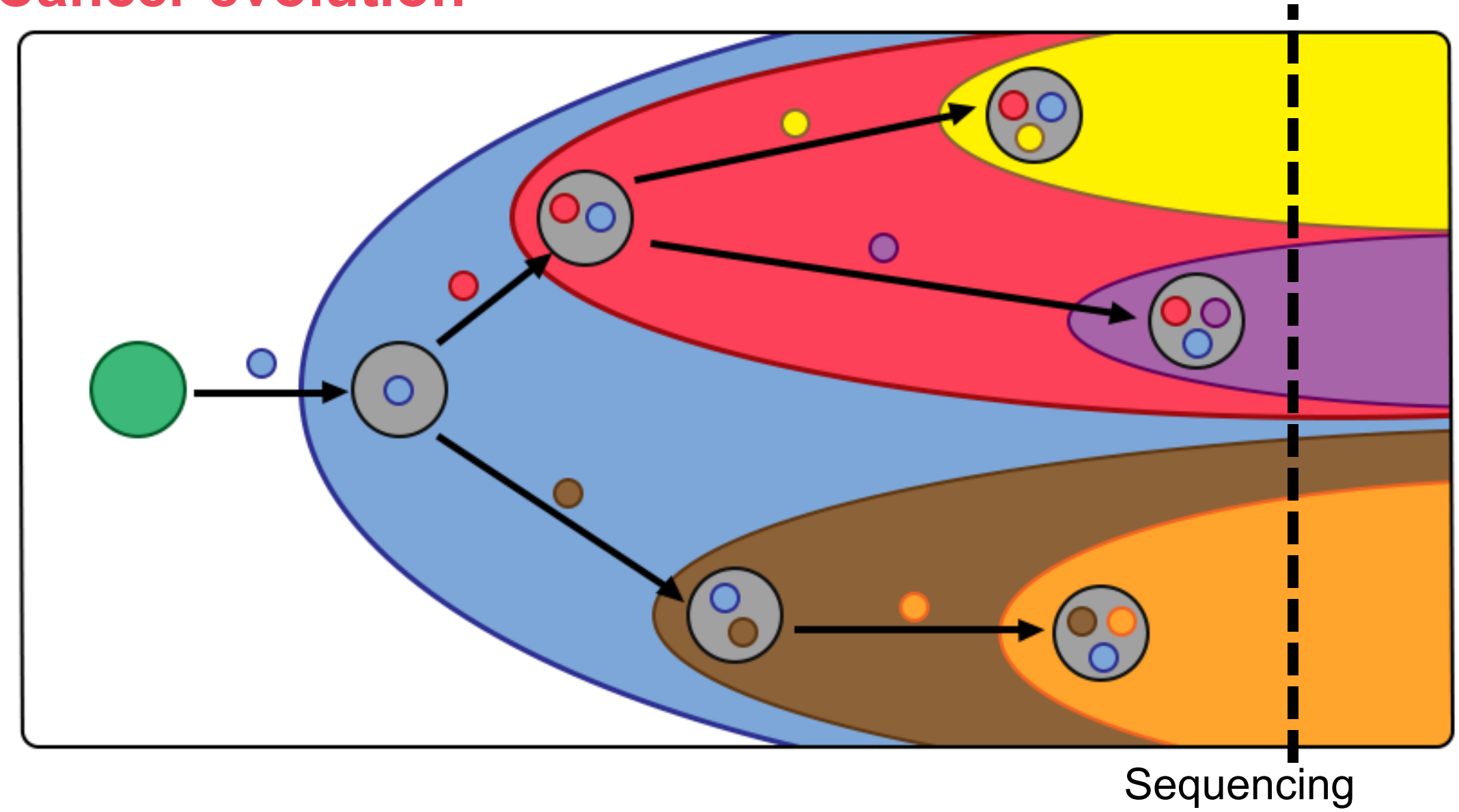
Introduction

- Cancer phylogeny
- Mutation losses
- gpps
- Experimental results

Cancer evolution



Cancer evolution



Infinite Sites Assumption

No two mutations can occur at the same locus (site). *Kimura*, Genetics, 1969.

- The most used assumption for the inference of cancer evolutions
- Permits the use of the simplest phylogeny model

Infinite Sites Assumption

No two mutations can occur at the same locus (site). *Kimura*, Genetics, 1969.

- The most used assumption for the inference of cancer evolutions
- Permits the use of the simplest phylogeny model

- "Our results **refute the general validity of the infinite sites assumption** and indicate that more complex models are needed to adequately quantify intra-tumor heterogeneity for more effective cancer treatment."

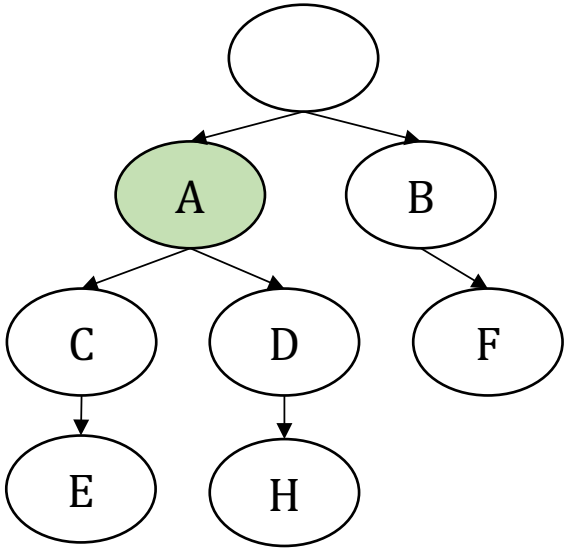
Kuipers et al., Genome Research, 2017.

- "In genomically unstable cancers, **deletion of large chromosomal segments is common**."

Brown et al., Nature 8, 2017.

Phylogenies: Perfect vs Dollo

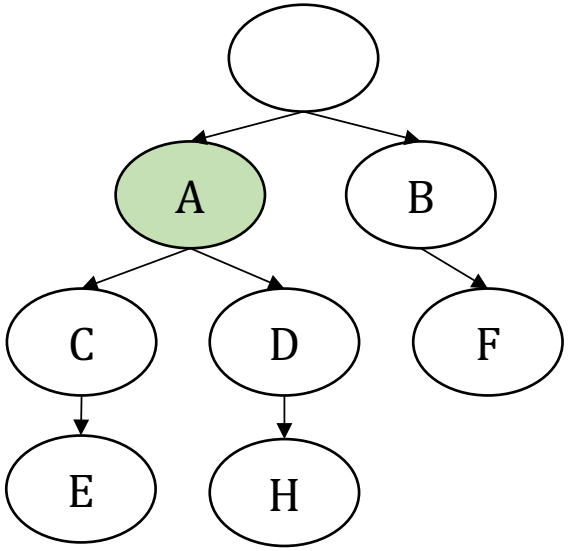
Perfect Phylogeny



Each mutation is **acquired** *once* in the evolutionary history

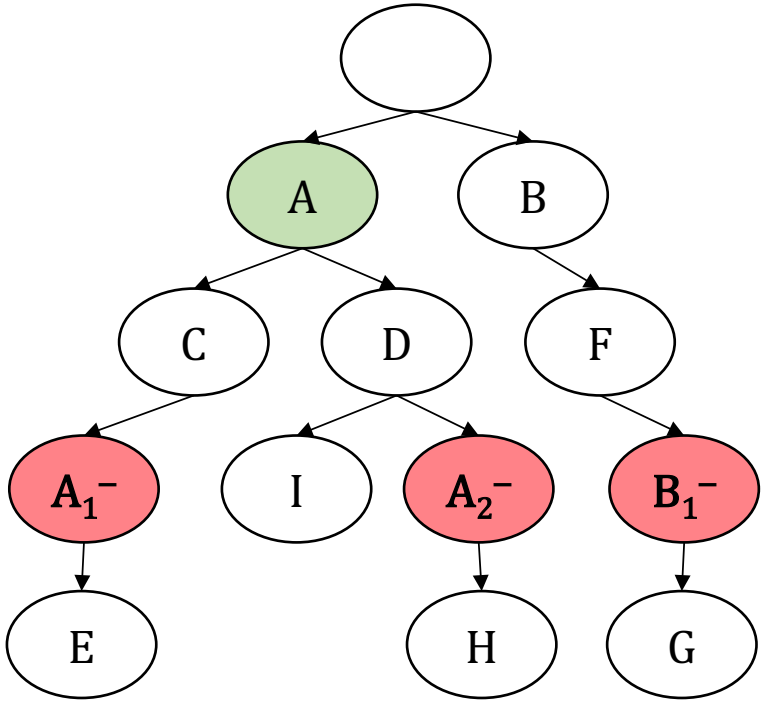
Phylogenies: Perfect vs Dollo

Perfect Phylogeny



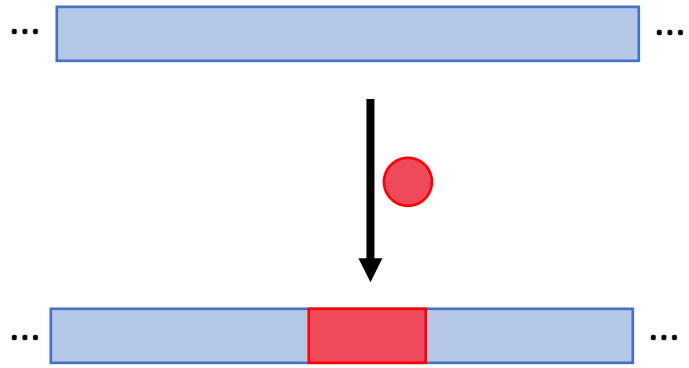
Each mutation is **acquired** *once* in the evolutionary history

Dollo-k Phylogeny

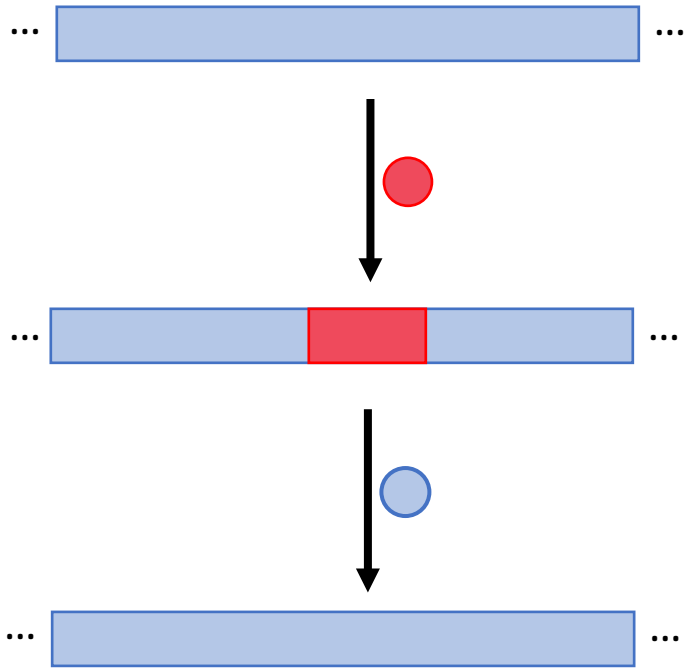


Each mutation is **acquired** *once*, but it can be **lost** *at most k times* in the evolutionary history

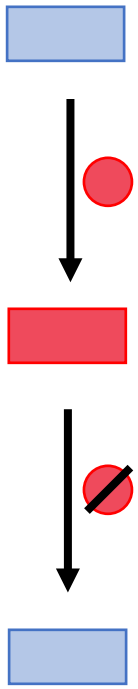
Loss of a mutation



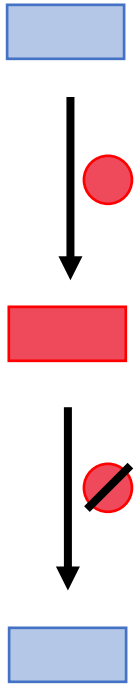
Loss of a mutation

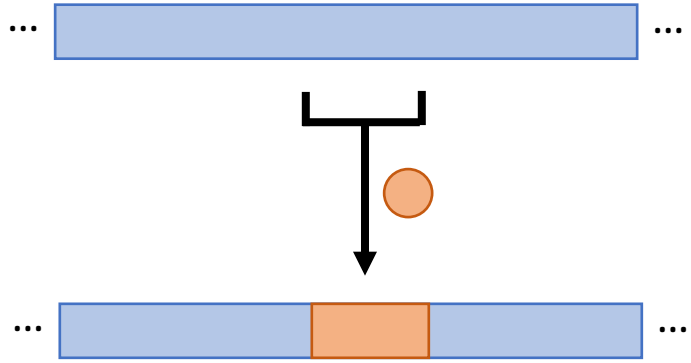


Loss of a mutation

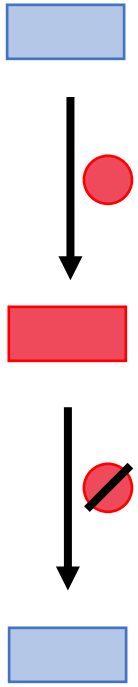


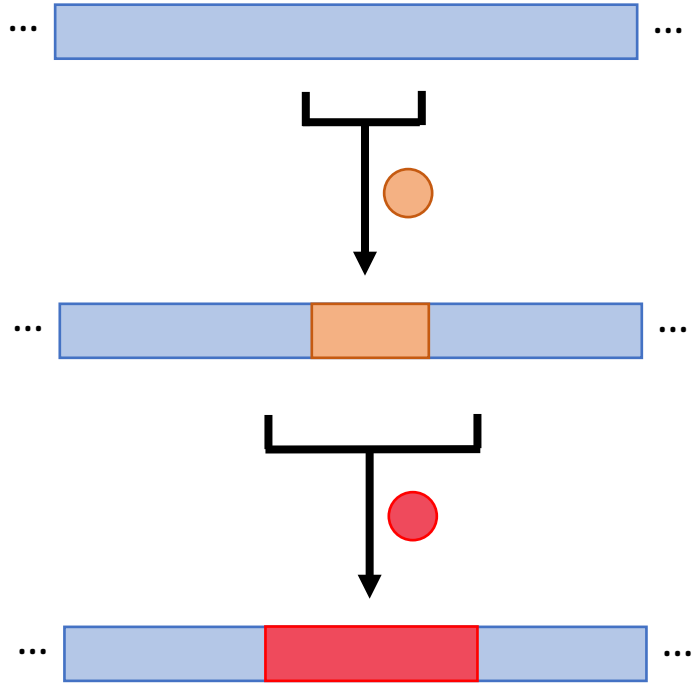
Loss of a mutation



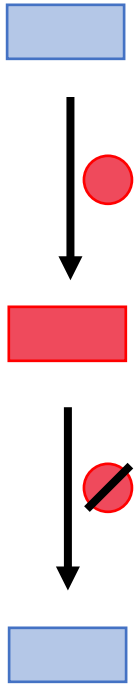


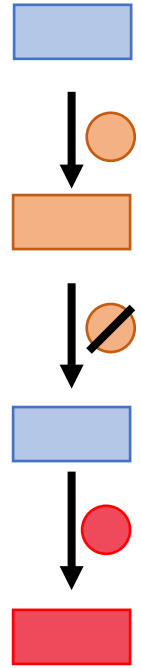
Loss of a mutation





Loss of a mutation





Single Cell Sequencing

Available methods for SCS

SCITE:

- Markov Chain Monte Carlo (MCMC) maximum likelihood tree search
- Relies on the Perfect Phylogeny model
- Produces solutions with respect to the Infinite Site Assumption

Tree inference for single-cell data.
Jahn K., Kuipers J. and Beerenwinkel N.,
Genome Biology, 2016.

Available methods for SCS

SCITE:

- Markov Chain Monte Carlo (MCMC) maximum likelihood tree search
- Relies on the Perfect Phylogeny model
- Produces solutions with respect to the Infinite Site Assumption

SiFit:

- Hidden Markov Model (HMM) maximum likelihood tree search
- Does not impose any specific phylogeny model
- Can produce solutions that violate the Infinite Site Assumption

Tree inference for single-cell data.
Jahn K., Kuipers J. and Beerenwinkel N.,
Genome Biology, 2016.

SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models.
Zafar H., Tzen A., Navin N., Chen K. and Nakhleh L.,
Genome Biology, 2017.

Infinite Sites Assumption

- Once a mutation arises it is not possible to lose it

For each pair of columns (mutations) there are no three rows with all the three configurations *(0,1), (1,0), (1,1)*

	A	B	C	D	E
s_1	0	1	1	1	1
s_2	1	0	1	0	1
s_3	0	0	1	0	0
s_4	0	0	1	1	0
s_5	0	1	1	1	0

Infinite Sites Assumption

- Once a mutation arises it is not possible to lose it

For each pair of columns (mutations) there are no three rows with all the three configurations *(0,1), (1,0), (1,1)*

	A	B	C	D	E
s_1	0	1	1	1	1
s_2	1	0	1	0	1
s_3	0	0	1	0	0
s_4	0	0	1	1	0
s_5	0	1	1	1	0

Infinite Sites Assumption

- Once a mutation arises it is not possible to lose it

For each pair of columns (mutations) there are no three rows with all the three configurations *(0,1), (1,0), (1,1)*

	A	B	C	D	E
s_1	0	1	1	1	1
s_2	1	0	1	0	1
s_3	0	0	1	0	0
s_4	0	0	1	1	0
s_5	0	1	1	1	0

Infinite Sites Assumption

- Once a mutation arises it is not possible to lose it

For each pair of columns (mutations) there are no three rows with all the three configurations *(0,1), (1,0), (1,1)*

	A	B	C	D	E
s_1	0	1	1	1	1
s_2	1	0	1	0	1
s_3	0	0	1	0	0
s_4	0	0	1	1	0
s_5	0	1	1	1	0

A binary matrix M is a Dollo Phylogeny if the extended matrix M_e is a Perfect Phylogeny

Bonizzoni et al., TCBB 2018

Dollo extended matrix

	A	B	C	D	E
s_1	0	1	1	1	1
s_2	1	0	1	0	1
s_3	0	0	1	0	0
s_4	0	0	1	1	0
s_5	0	1	1	1	0

Dollo extended matrix

	A	**A-**	B	**B-**	C	**C-**	D	**D-**	E	**E-**
s_1	0		1		1		1		1	
s_2	1		0		1		0		1	
s_3	0		0		1		0		0	
s_4	0		0		1		1		0	
s_5	0		1		1		1		0	

Dollo extended matrix

	A	A-	B	B-	C	C-	D	D-	E	E-
s_1	0	0	1	1	1	0	1	0	1	1
s_2	1	0	0	0	1	0	0	0	1	0
s_3	0	0	0	0	1	1	0	0	0	0
s_4	0	0	0	0	1	0	1	0	0	0
s_5	0	0	1	0	1	0	1	1	0	0

Dollo extended matrix

	A	A-	B	B-	C	C-	D	D-	E	E-
s_1	0	0	1	1	1	0	1	0	1	1
s_2	1	0	0	0	1	0	0	0	1	0
s_3	0	0	0	0	1	1	0	0	0	0
s_4	0	0	0	0	1	0	1	0	0	0
s_5	0	0	1	0	1	0	1	1	0	0

Mutation **A** in s_1 is never acquired and never lost

Dollo extended matrix

	A	A-	B	B-	C	C-	D	D-	E	E-
s_1	0	0	1	1	1	0	1	0	1	1
s_2	1	0	0	0	1	0	0	0	1	0
s_3	0	0	0	0	1	1	0	0	0	0
s_4	0	0	0	0	1	0	1	0	0	0
s_5	0	0	1	0	1	0	1	1	0	0

Mutation **C** in s_5 is acquired but never lost

Dollo extended matrix

	A	A-	B	B-	C	C-	D	D-	E	E-
s_1	0	0	1	1	1	0	1	0	1	1
s_2	1	0	0	0	1	0	0	0	1	0
s_3	0	0	0	0	1	1	0	0	0	0
s_4	0	0	0	0	1	0	1	0	0	0
s_5	0	0	1	0	1	0	1	1	0	0

Mutation **E** in s_1 is acquired and then lost

gpps: ILP

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$ if $I(c,m) = 1$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$ if $I(c,m) = 0$

gpps: ILP

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1 - \alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$ if $I(c,m) = 1$

2) $w(c,m) = \alpha F(c,m) + (1 - \beta)\big(1 - F(c,m)\big)$ if $I(c,m) = 0$

α = false negative rate

gpps: ILP

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$ if $I(c,m) = 1$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$ if $I(c,m) = 0$

α = false negative rate

β = false positive rate

gpps: ILP

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$ if $I(c,m) = 1$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$ if $I(c,m) = 0$

α = false negative rate

β = false positive rate

I = Input matrix

gpps: ILP

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$ if $I(c,m) = 1$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$ if $I(c,m) = 0$

α = false negative rate

β = false positive rate

I = Input matrix

F = Inferred matrix

gpps: ILP

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$ if $I(c,m) = 1$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$ if $I(c,m) = 0$

α = false negative rate

β = false positive rate

I = Input matrix

F = Inferred matrix

$$= \max \sum_{c} \sum_{m} \log(P(I(c,m)|F(c,m))$$

gpps: ILP

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c, m)$$

- Additional set of constraints used to ensure a Dollo model on the tree

- *Gusfield, RECOMB 2002*
- *Gusfield et al., ReCombinatorics 2014*
- *Bonizzoni et al., TCBB 2018*

gpps: Hill Climbing

Continues the search process after the ILP is stopped

ILP Solver

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$

...

gpps: Hill Climbing

Continues the search process after the ILP is stopped

ILP Solver

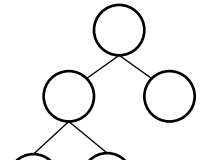
$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$

...

Timeout



Sub-optimal solution

gpps: Hill Climbing

Continues the search process after the ILP is stopped

ILP Solver

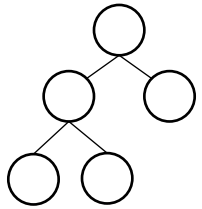
$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m)$$

1) $w(c,m) = (1-\alpha)F(c,m) + \beta\big(1 - F(c,m)\big)$

2) $w(c,m) = \alpha F(c,m) + (1-\beta)\big(1 - F(c,m)\big)$

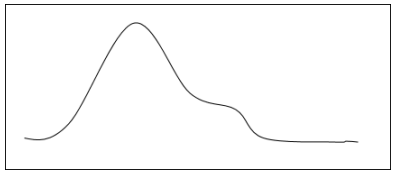
...

Timeout



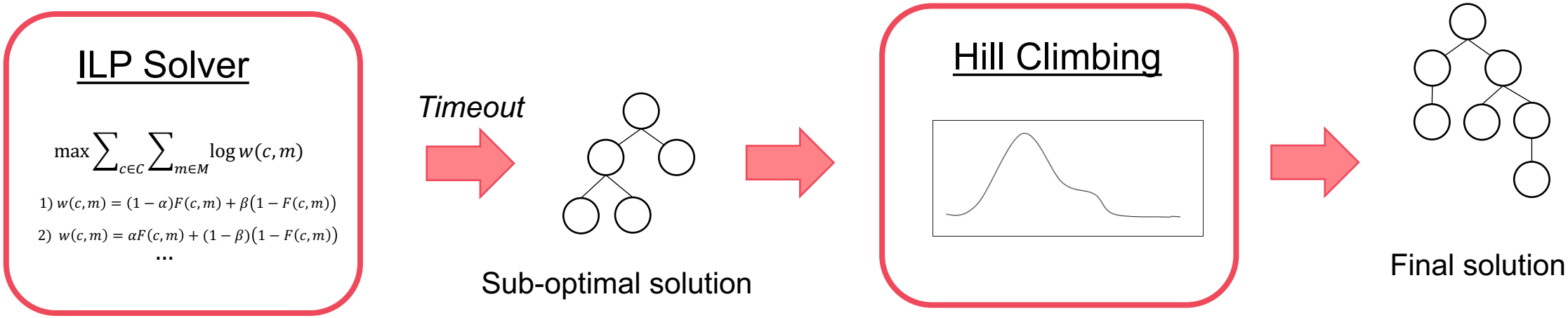
Sub-optimal solution

Hill Climbing



gpps: Hill Climbing

Continues the search process after the ILP is stopped

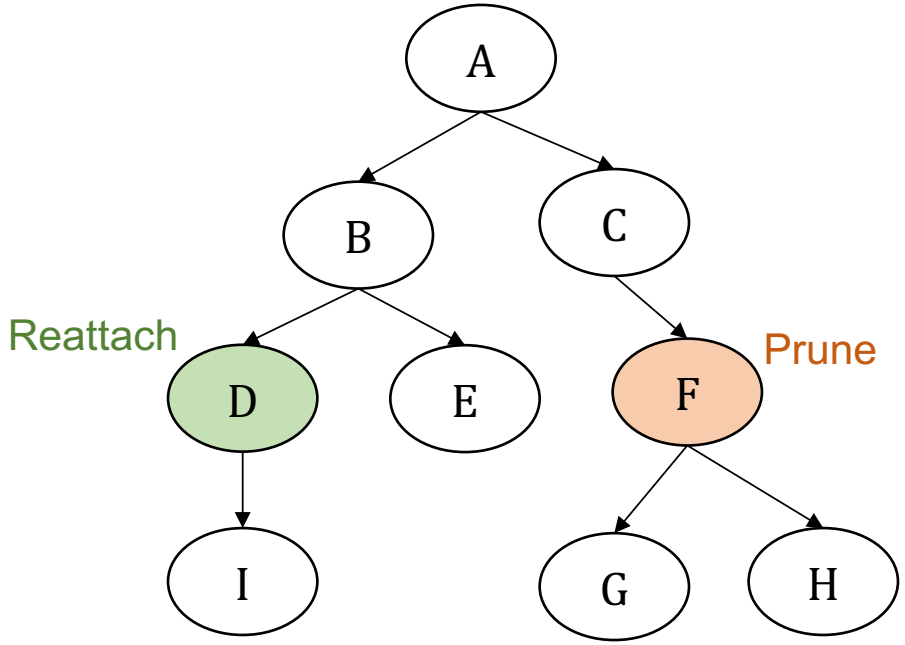


gpps: Hill Climbing

Continues the search process using the Subtree Prune and Reattach operation

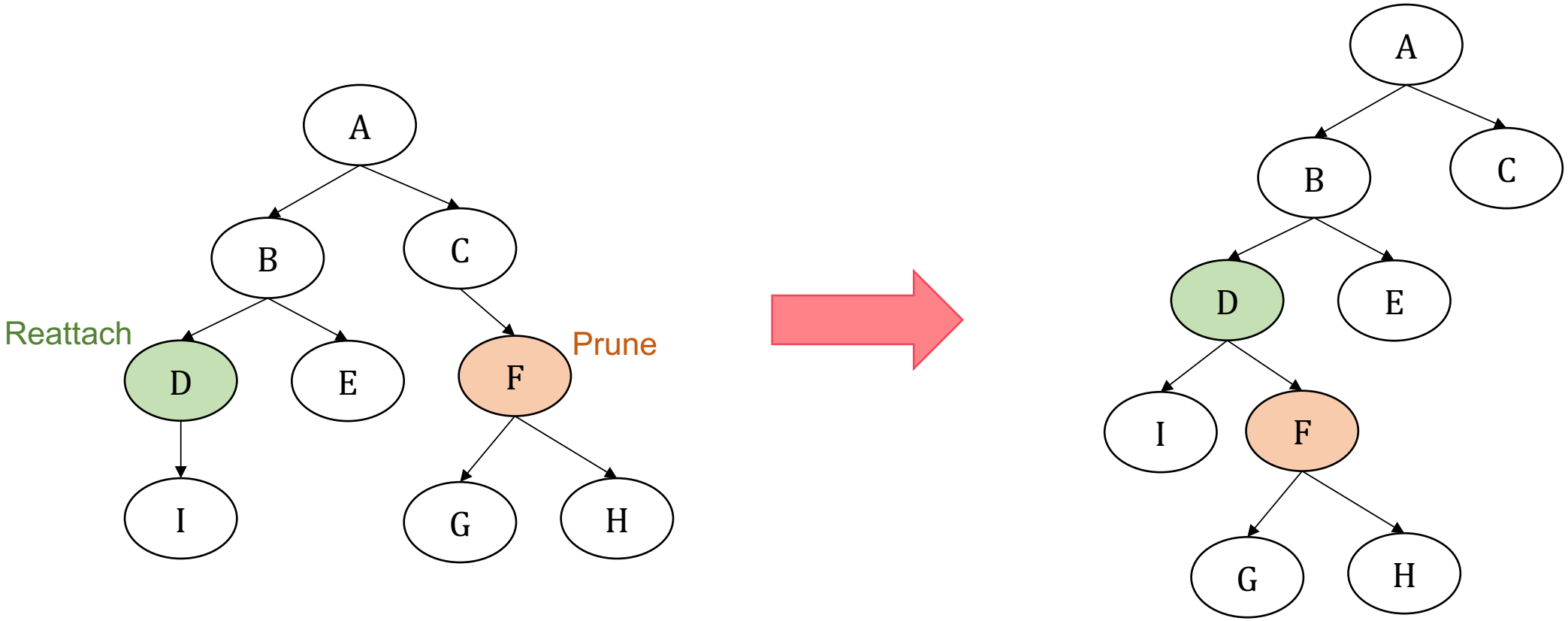
gpps: Hill Climbing

Continues the search process using the Subtree Prune and Reattach operation



gpps: Hill Climbing

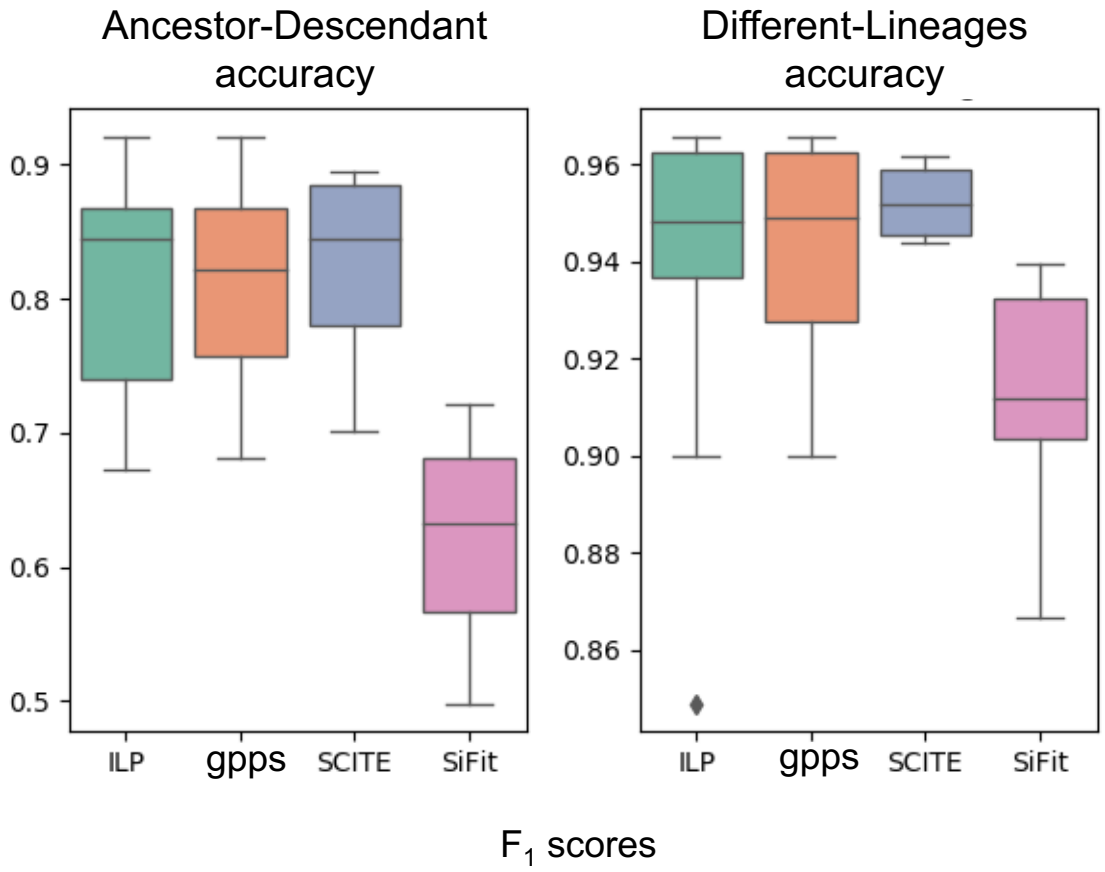
Continues the search process using the Subtree Prune and Reattach operation



Results: Simulated data

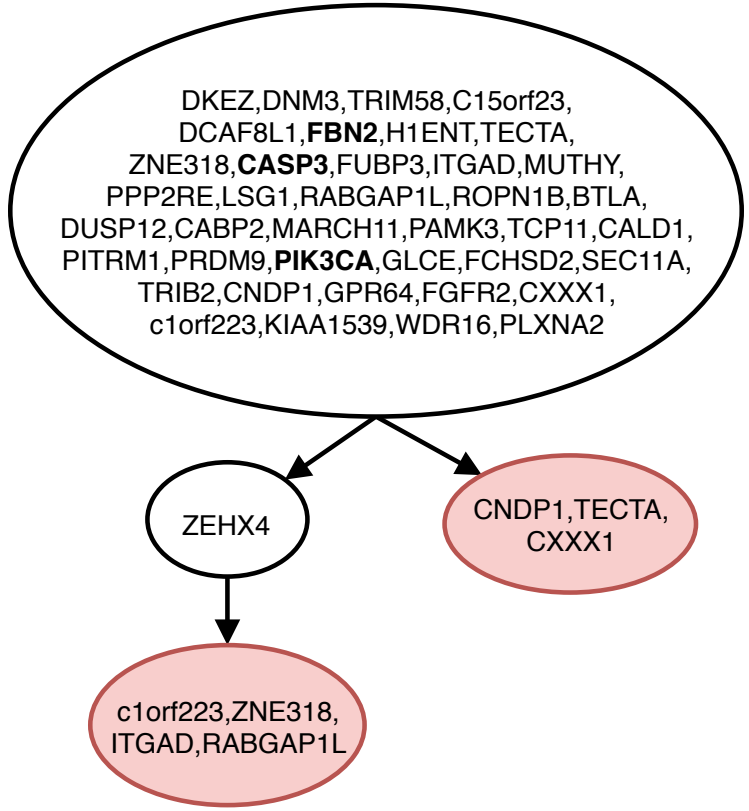
- **Ancestor-Descendant** accuracy:
  Pairs of mutations in Ancestor-Descendant relationship correctly inferred

- **Different lineages** accuracy:
  Pairs of mutations in different branches correctly inferred

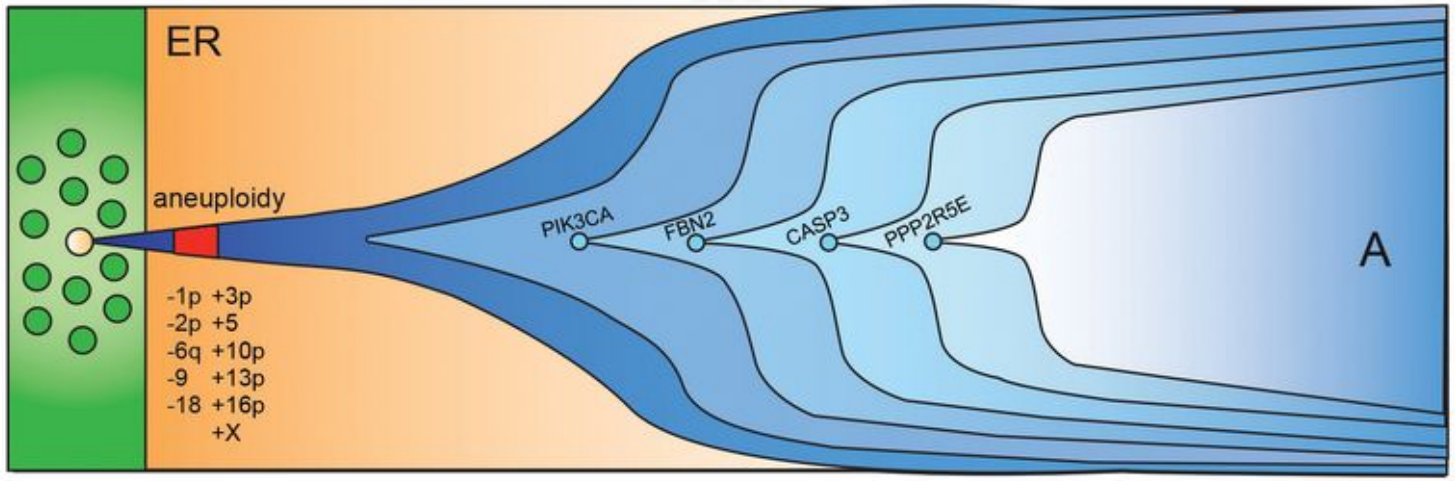


Real data: ER Breast Cancer

Tree inferred by **gpps**



Manually curated clonal evolution



Bold-faced *mutations are drivers and red nodes are losses*

Data from: Wang et al. *Nature, 2014*

Conclusions

- gpps is an accurate tool for inferring intra-tumor progression and subclonal composition from SCS data

- gpps infers mutation losses employing a Dollo model

- gpps provides a new progression model on Single Cell data

- **Future directions**:
  1. Explore different heuristics (Genetic Programming)
  2. Include the possible presence of doublets to the model

Thank you!