RECONCILIATION WITH SEGMENTAL DUPLICATION, TRANSFER, LOSS AND GAIN

Yoann Anselmetti[1] *Mattéo Delabre*[2] Nadia El-Mabrouk[2]

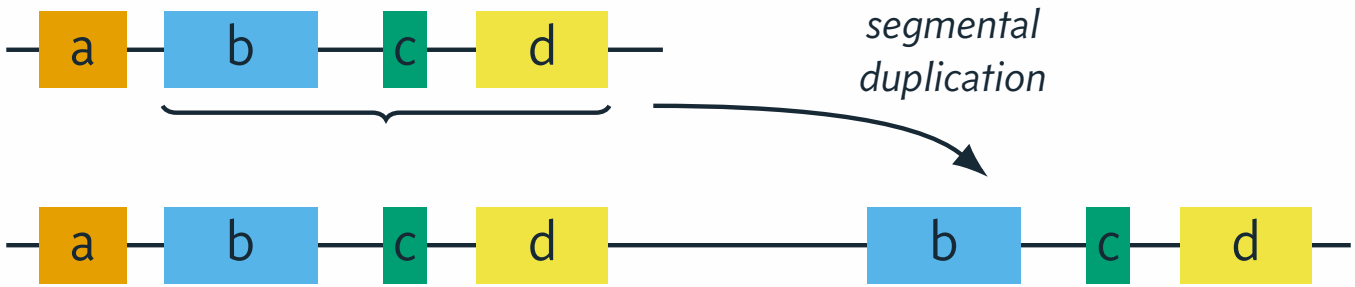
[1]Université de Sherbrooke
[2]Université de Montréal

RECOMB-CG — 19th Annual Satellite Conference
of RECOMB on Comparative Genomics

La Jolla, USA
May 21, 2022

HISTORY OF SYNTENIES

- **Syntenies:** Groups of genes that can be **affected simultaneously** by segmental events *(e.g., because of their location in the genome)*



- **Problem:** Given a family of *homologous* syntenies, **infer a plausible scenario** for their history, made up of segmental events

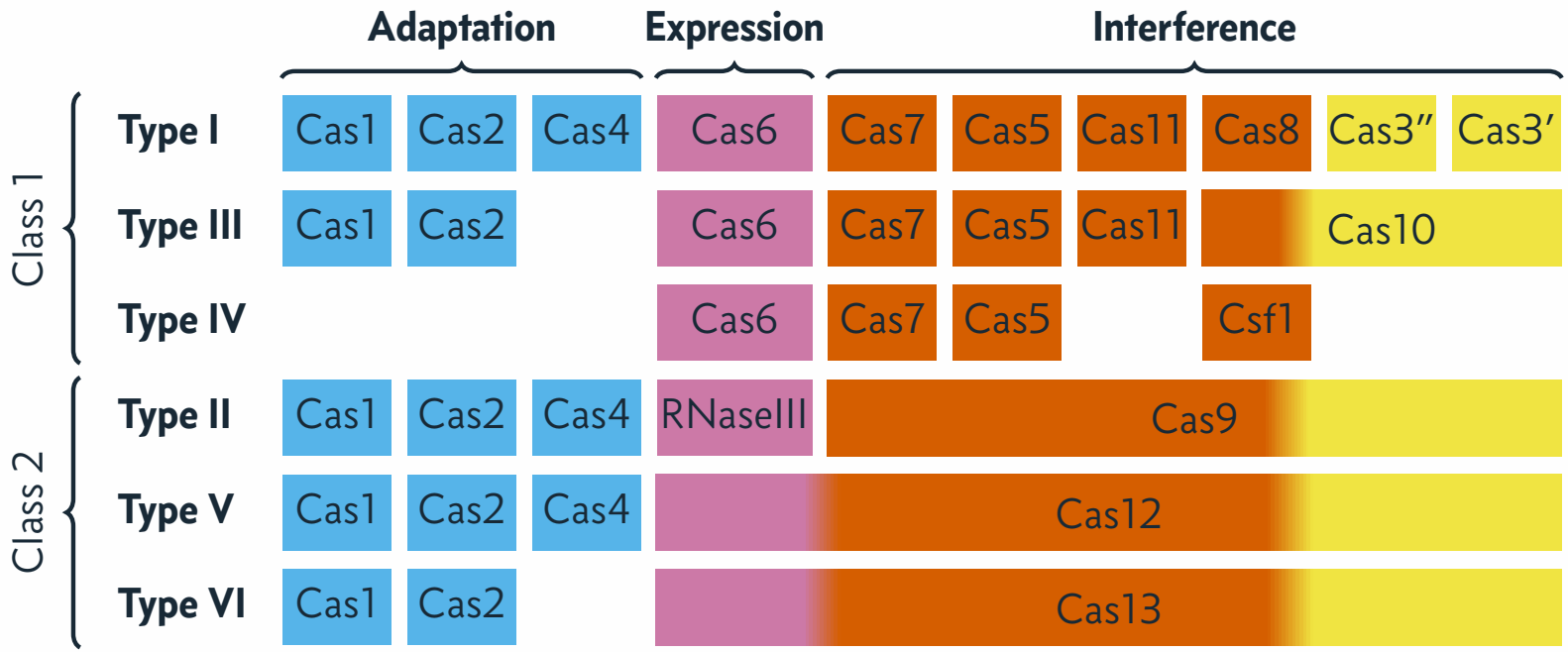
CRISPR-CAS SYSTEMS

- **Adaptive immunity** mechanisms of bacteria and archaea

- **Foreign sequences** are stored as spacers in the **CRISPR array**
- **Cas genes** perform adaptation, expression, and interference

CAS SYNTENIES CLASSIFICATION

- Cas genes form a **diverse family** of syntenies

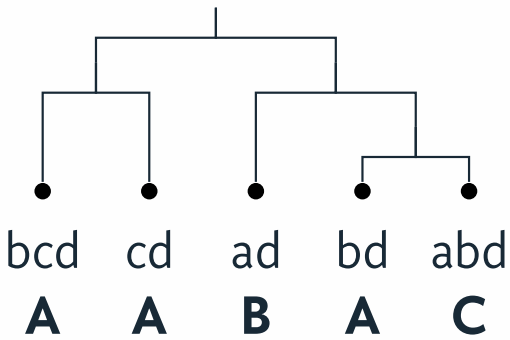


- Can we **infer a scenario** for the history of **Cas syntenies?**

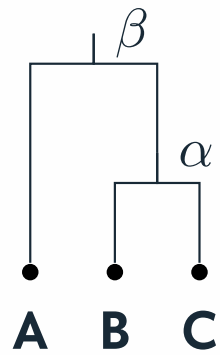
Makarova et al. “Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants” (Feb. 2020).

USING SUPER-RECONCILIATION METHODS

- **Extend** usual reconciliation techniques for **synteny** history inference
- A **synteny tree** replaces the gene tree in the problem input



Synteny tree



Species tree

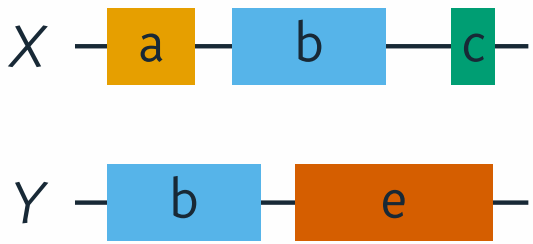
- The synteny tree can be obtained as a **supertree** of individual gene trees, hence **"super-reconciliation"**

OUTLINE

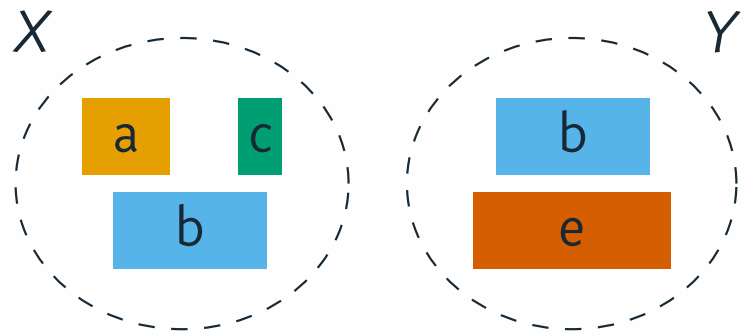
1. **Modelling the Evolution of Syntenies**
2. Super-Reconciliation Approach
3. Preliminary Results on Cas Syntenies
4. Conclusion

MODELLING SYNTENIES

1 As **strings**



2 As **sets**



- Super-reconciliation is **less computationally expensive** on **sets**
- **Strings** require either **"order-consistent"** syntenies or modelling **rearrangement** events

---

Delabre, El-Mabrouk, Huber, Lafond, Moulton, Noutahi, and Castellanos. "Evolution through segmental duplications and losses: a Super-Reconciliation approach" (May 2020).

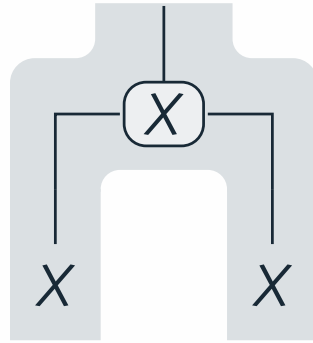
MODELLING SEGMENTAL EVENTS

- BINARY EVENTS
  Create new synteny copies

- UNARY EVENTS
  Change existing syntenies

MODELLING SEGMENTAL EVENTS

- BINARY EVENTS
  Create new synteny copies
  - **Speciation**

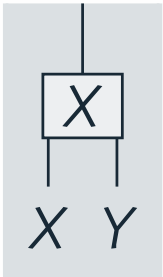


- UNARY EVENTS
  Change existing syntenies

MODELLING SEGMENTAL EVENTS

- BINARY EVENTS
  Create new synteny copies
  - Speciation
  - **Duplication**

- UNARY EVENTS
  Change existing syntenies

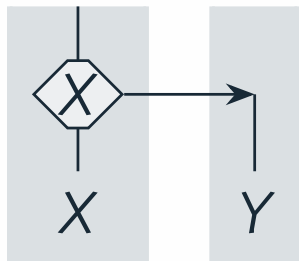


$Y \subseteq X$

MODELLING SEGMENTAL EVENTS

- BINARY EVENTS
  Create new synteny copies
  - Speciation
  - Duplication
  - **Transfer**

- UNARY EVENTS
  Change existing syntenies

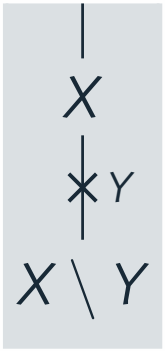


$Y \subseteq X$

MODELLING SEGMENTAL EVENTS

- **BINARY EVENTS**
  Create new synteny copies
  - Speciation
  - Duplication
  - Transfer

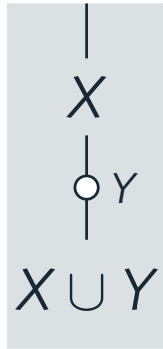
- **UNARY EVENTS**
  Change existing syntenies
  - **Loss**



MODELLING SEGMENTAL EVENTS

- **BINARY EVENTS**
  Create new synteny copies
  - Speciation
  - Duplication
  - Transfer

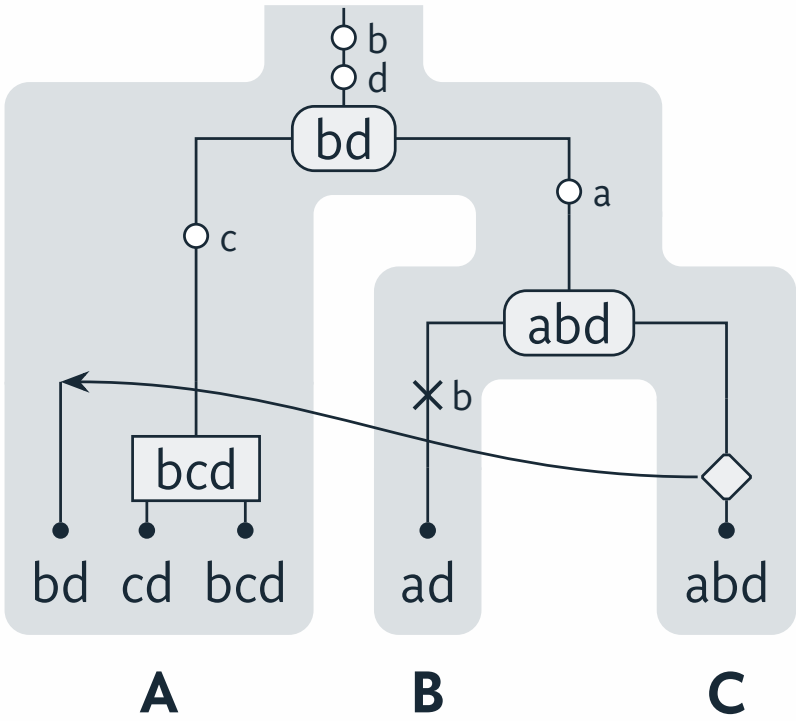
- **UNARY EVENTS**
  Change existing syntenies
  - Loss
  - **Gain**



- Restrictions
  - **Gains** are **non-segmental** ($|Y| = 1$)
  - Each family is gained exactly once (**no convergent evolution**)

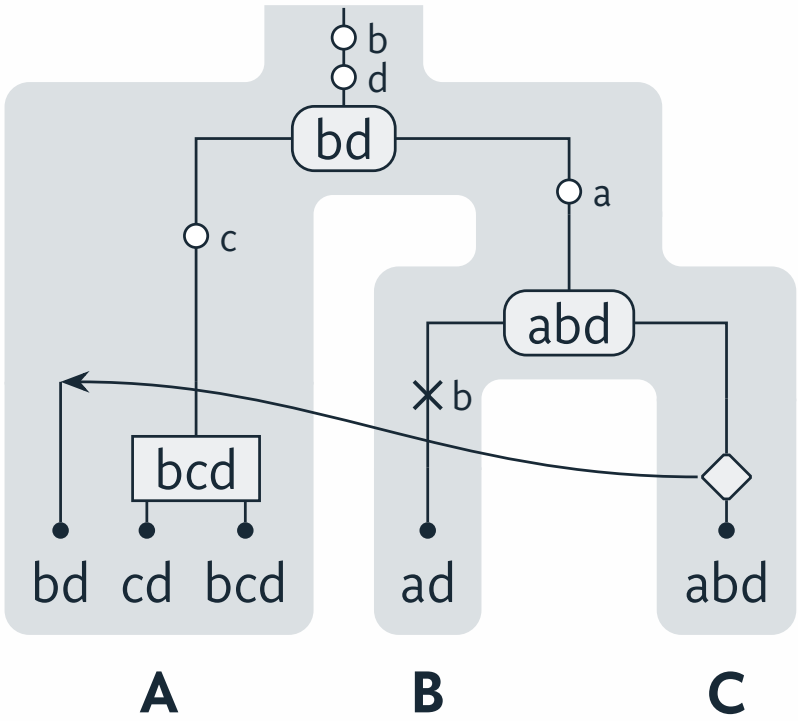
EVOLUTIONARY HISTORY

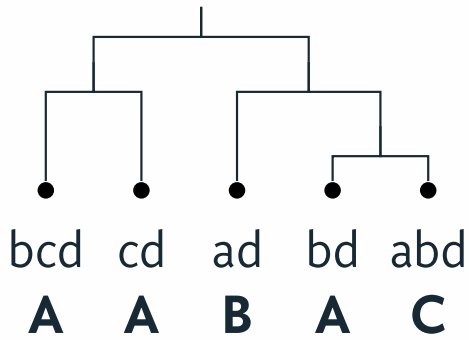
- **Branching sequence** of segmental events



EVOLUTIONARY HISTORY

- **Branching sequence** of segmental events
- **Corresponds to** a synteny tree

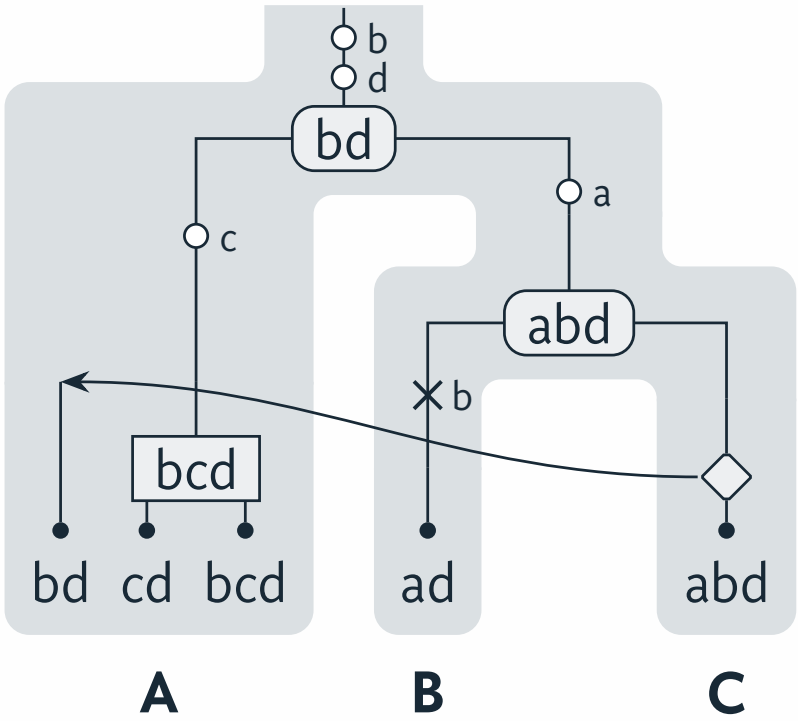


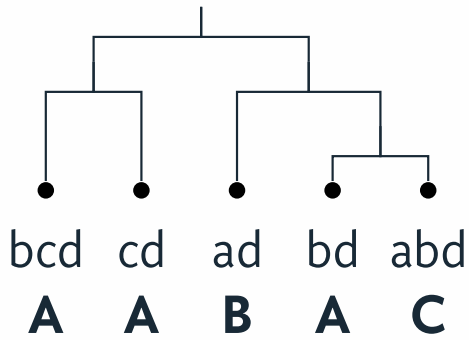


Synteny tree

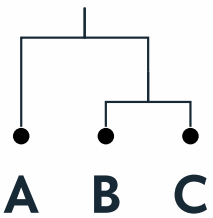
EVOLUTIONARY HISTORY

- **Branching sequence** of segmental events
- **Corresponds to** a synteny tree
- **Fits in** a species tree





Synteny tree



Species tree

PARSIMONY PROBLEM

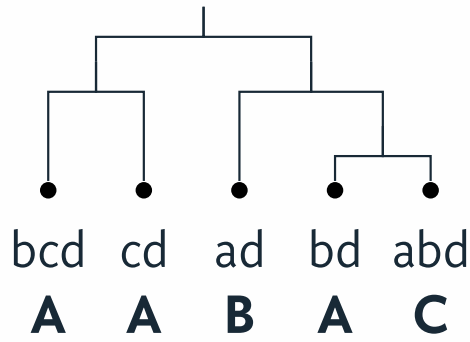
- Find most **plausible histories** for given synteny and species trees
- Each **event** is given a **cost** value (quantifying its "rareness")
- The cost of a history is the **sum of costs** for its events

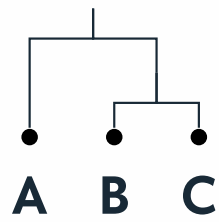
Minimum-Cost History Problem

Input: Synteny tree T, species tree S
Output: Minimum-cost history corresponding to T and fitting in S

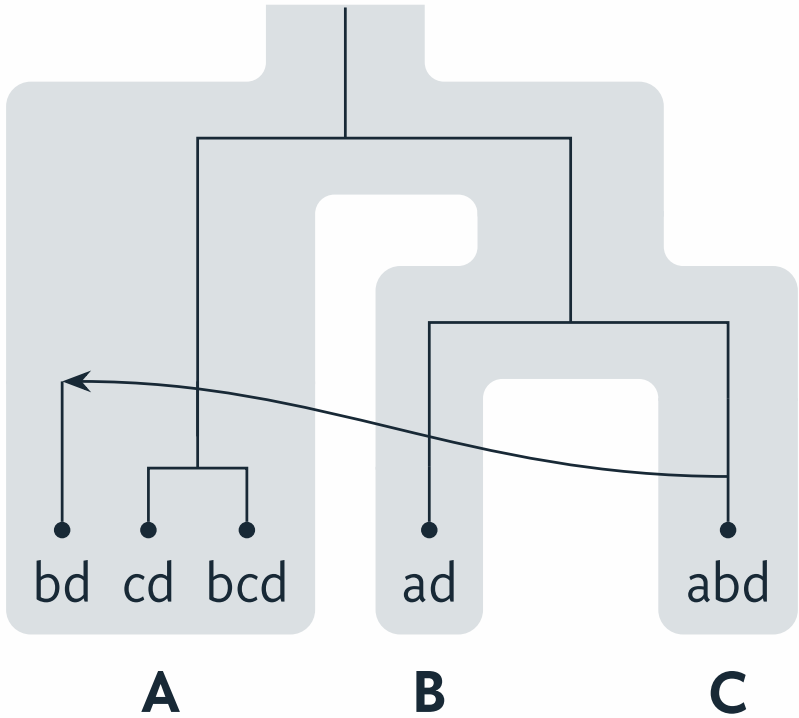
OUTLINE

PARTS OF A HISTORY



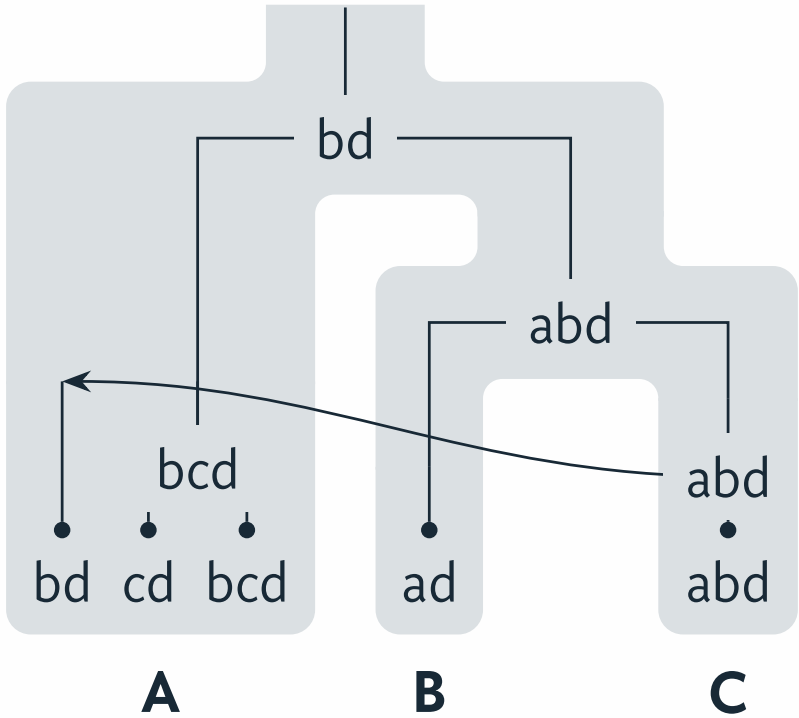


1 Synteny to species tree mapping



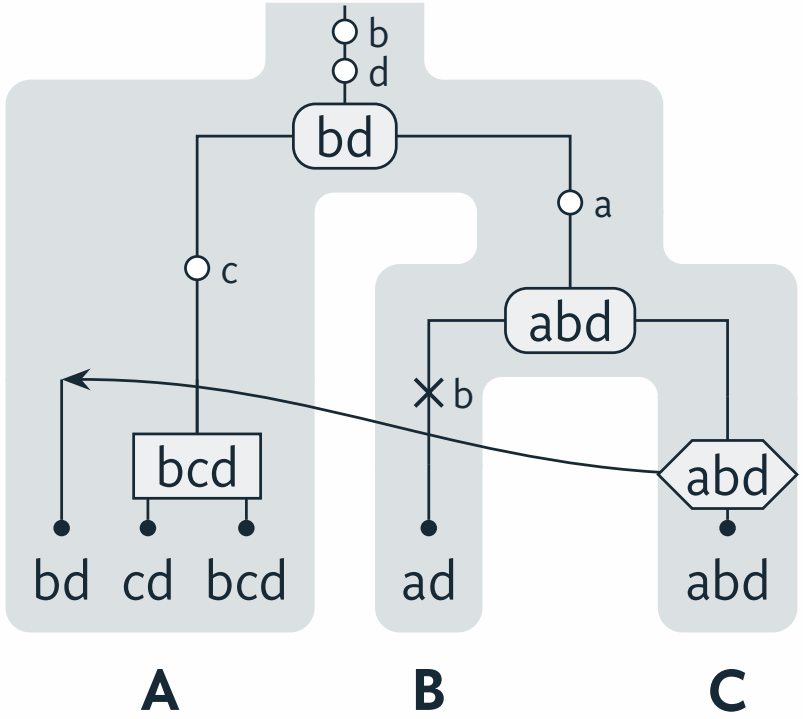
PARTS OF A HISTORY

1. Synteny to species tree mapping
2. **Ancestral syntenies assignment**

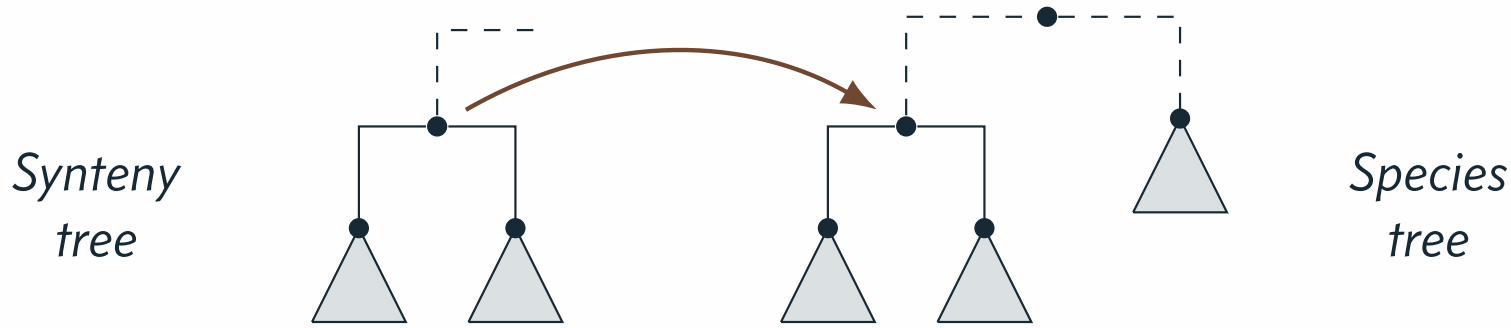


PARTS OF A HISTORY

1. Synteny to species tree mapping
2. Ancestral syntenies assignment
3. **Events (come for free!)**



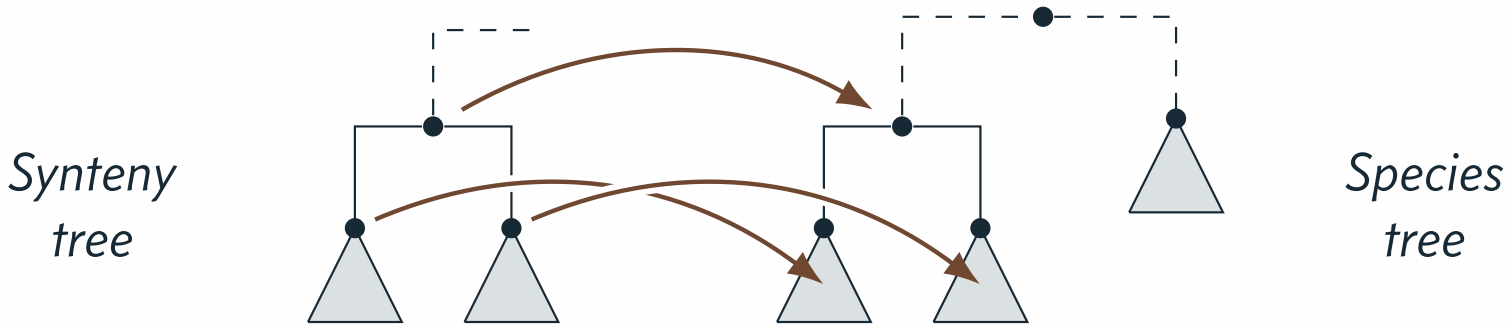
MAPPING SYNTENY TREE TO SPECIES TREE



Tofigh, Hallett, and Lagergren. "Simultaneous identification of duplications and lateral gene transfers" (Mar. 2011).

Bansal, Alm, and Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss" (June 2012).

MAPPING SYNTENY TREE TO SPECIES TREE

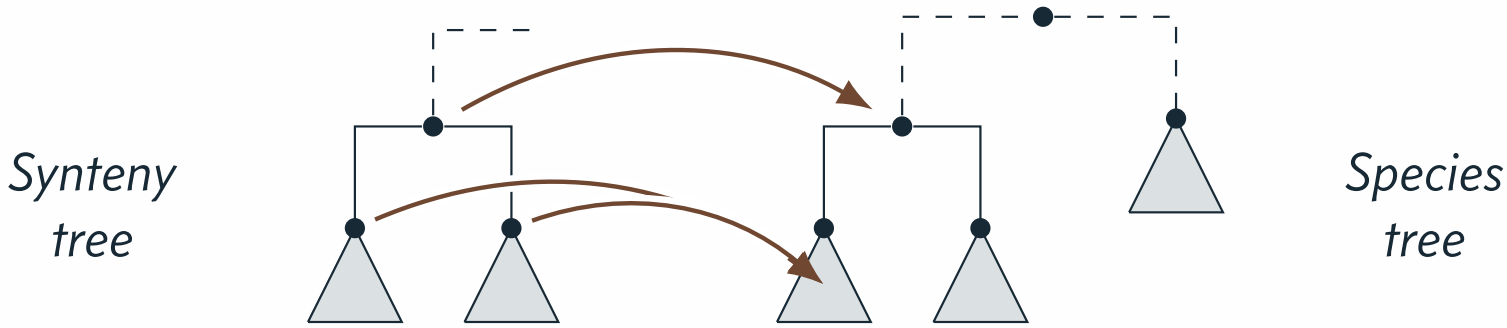


1 **Different descendent subtrees** Speciation

Tofigh, Hallett, and Lagergren. "Simultaneous identification of duplications and lateral gene transfers" (Mar. 2011).

Bansal, Alm, and Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss" (June 2012).

MAPPING SYNTENY TREE TO SPECIES TREE



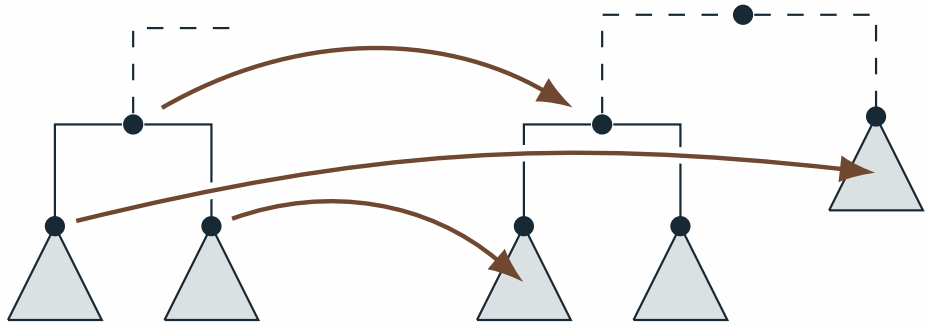
1	Different descendent subtrees	Speciation
2	**Same descendent subtree**	Duplication

Tofigh, Hallett, and Lagergren. "Simultaneous identification of duplications and lateral gene transfers" (Mar. 2011).

Bansal, Alm, and Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss" (June 2012).

MAPPING SYNTENY TREE TO SPECIES TREE

Synteny tree



Species tree

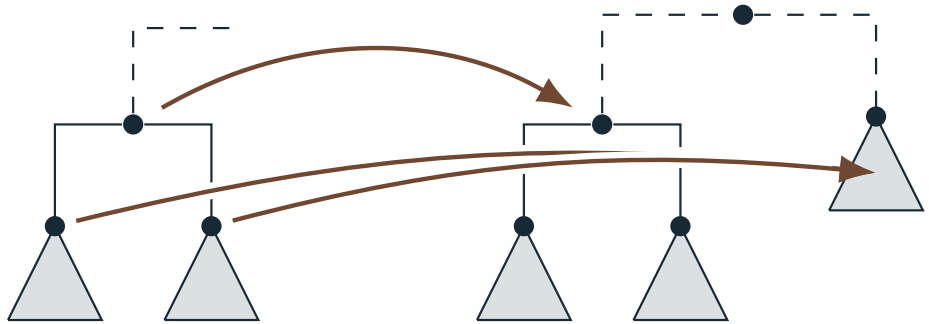
1	Different descendent subtrees	Speciation
2	Same descendent subtree	Duplication
3	**One in separate subtree**	Transfer

Tofigh, Hallett, and Lagergren. "Simultaneous identification of duplications and lateral gene transfers" (Mar. 2011).

Bansal, Alm, and Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss" (June 2012).

MAPPING SYNTENY TREE TO SPECIES TREE

Synteny tree



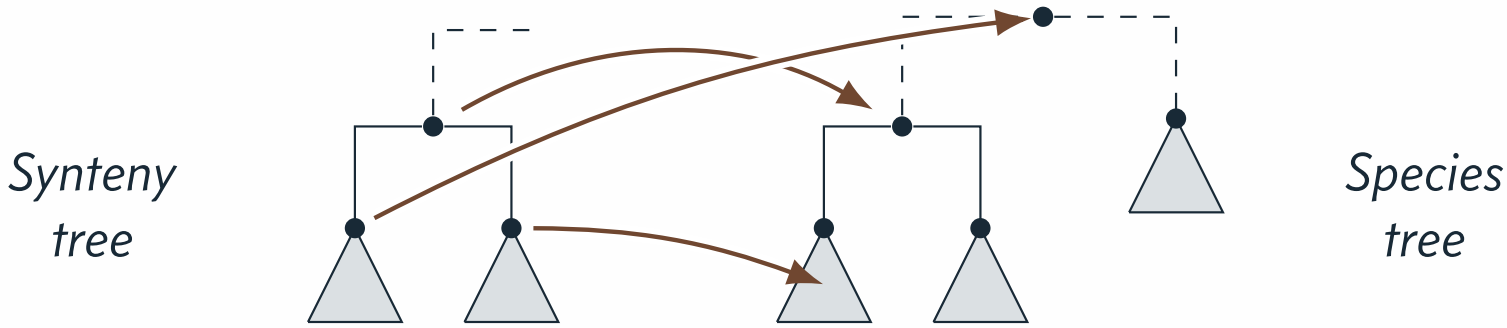
Species tree

1	Different descendent subtrees	Speciation
2	Same descendent subtree	Duplication
3	One in separate subtree	Transfer
4	**Both in separate subtrees**	*(Excluded)*

Tofigh, Hallett, and Lagergren. "Simultaneous identification of duplications and lateral gene transfers" (Mar. 2011).

Bansal, Alm, and Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss" (June 2012).

MAPPING SYNTENY TREE TO SPECIES TREE



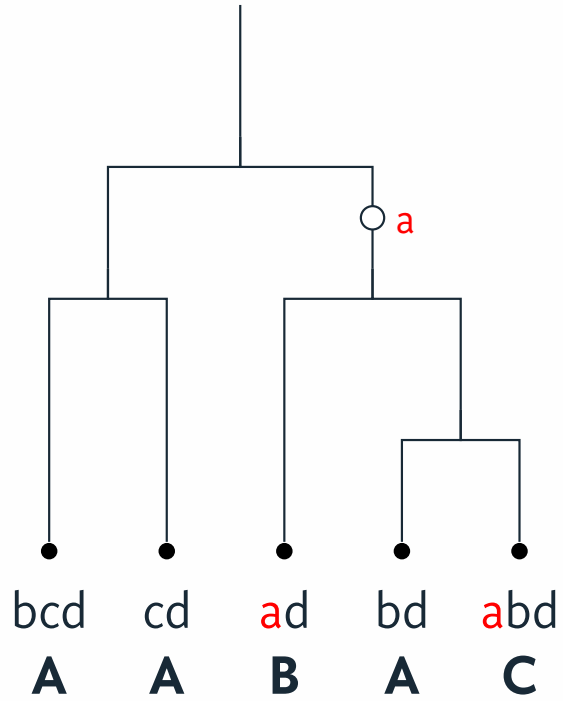
1	Different descendent subtrees	Speciation
2	Same descendent subtree	Duplication
3	One in separate subtree	Transfer
4	Both in separate subtrees	*(Excluded)*
5	**One in ancestor**	*(Inconsistent)*

Tofigh, Hallett, and Lagergren. "Simultaneous identification of duplications and lateral gene transfers" (Mar. 2011).

Bansal, Alm, and Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss" (June 2012).

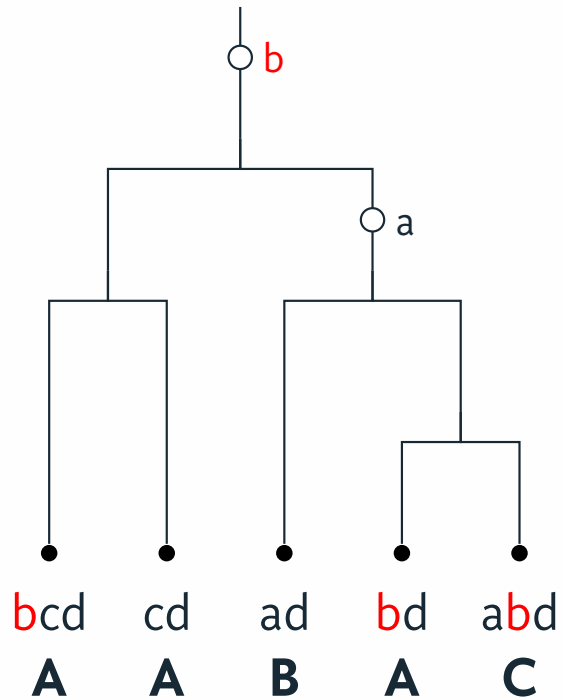
INFERRING GAIN EVENTS

- No family can be gained below the **LCA of the leaves it appears in**
- Minimum-cost history can be obtained by **placing gains exactly at the LCA**



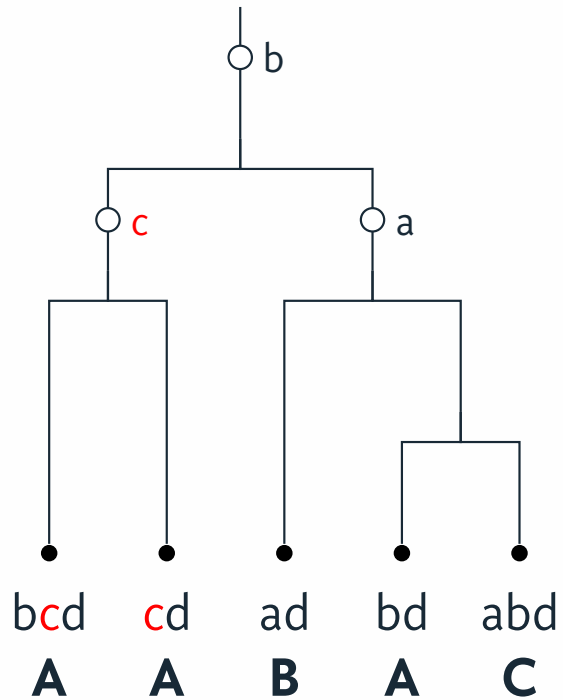
INFERRING GAIN EVENTS

- No family can be gained below the **LCA of the leaves it appears in**
- Minimum-cost history can be obtained by **placing gains exactly at the LCA**



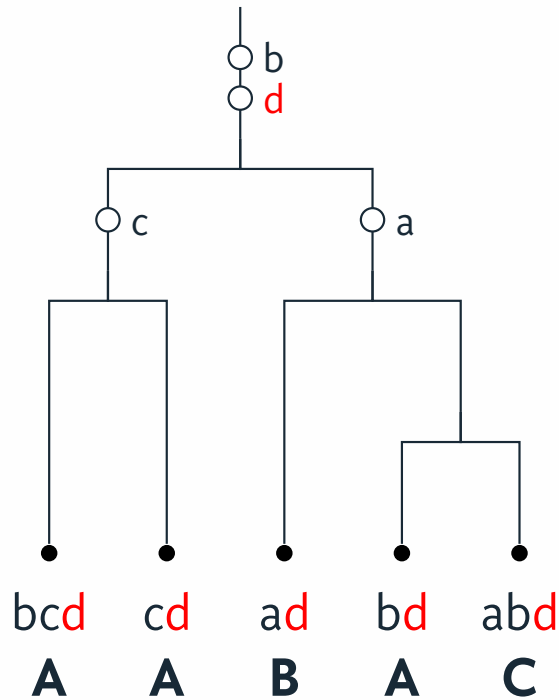
INFERRING GAIN EVENTS

- No family can be gained below the **LCA of the leaves it appears in**
- Minimum-cost history can be obtained by **placing gains exactly at the LCA**



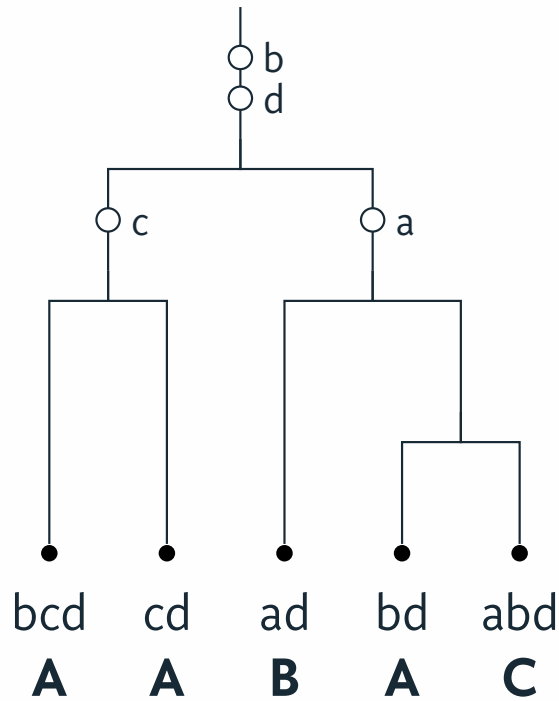
INFERRING GAIN EVENTS

- No family can be gained below the **LCA of the leaves it appears in**
- Minimum-cost history can be obtained by **placing gains exactly at the LCA**



INFERRING ANCESTRAL SYNTENIES

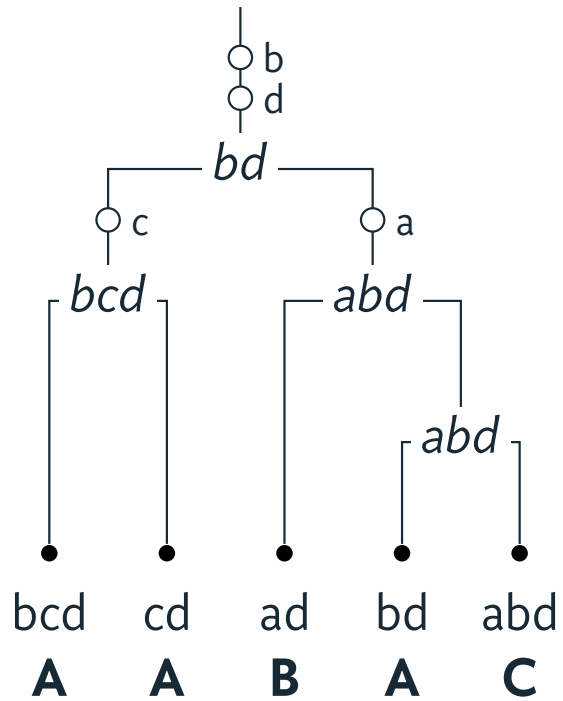
- Choose a **subset of families** for each node so that **losses are minimized**
  - *Exponential choices?*



Delabre, El-Mabrouk, Huber, Lafond, Moulton, Noutahi, and Castellanos. "Evolution through segmental duplications and losses: a Super-Reconciliation approach" (May 2020).

INFERRING ANCESTRAL SYNTENIES

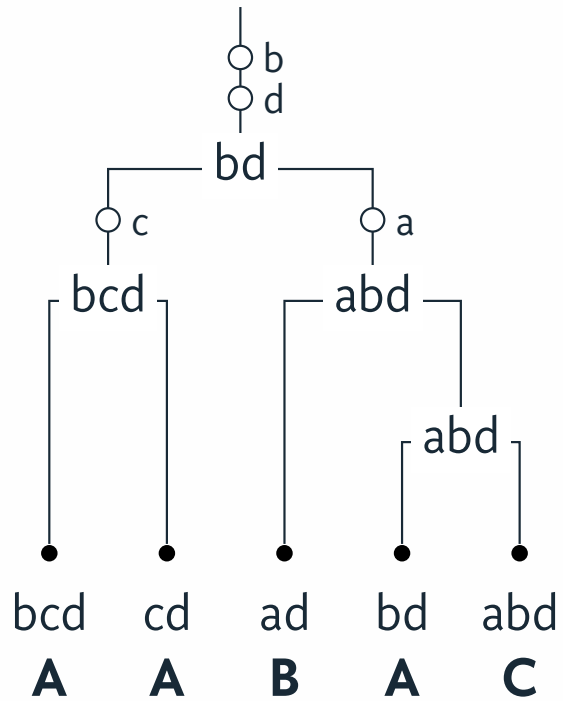
- Choose a **subset of families** for each node so that **losses are minimized**
  - *Exponential choices?*
- Each node must contain a **minimum set** of families, may **contain more**
  - "Extra content"



Delabre, El-Mabrouk, Huber, Lafond, Moulton, Noutahi, and Castellanos. "Evolution through segmental duplications and losses: a Super-Reconciliation approach" (May 2020).

INFERRING ANCESTRAL SYNTENIES

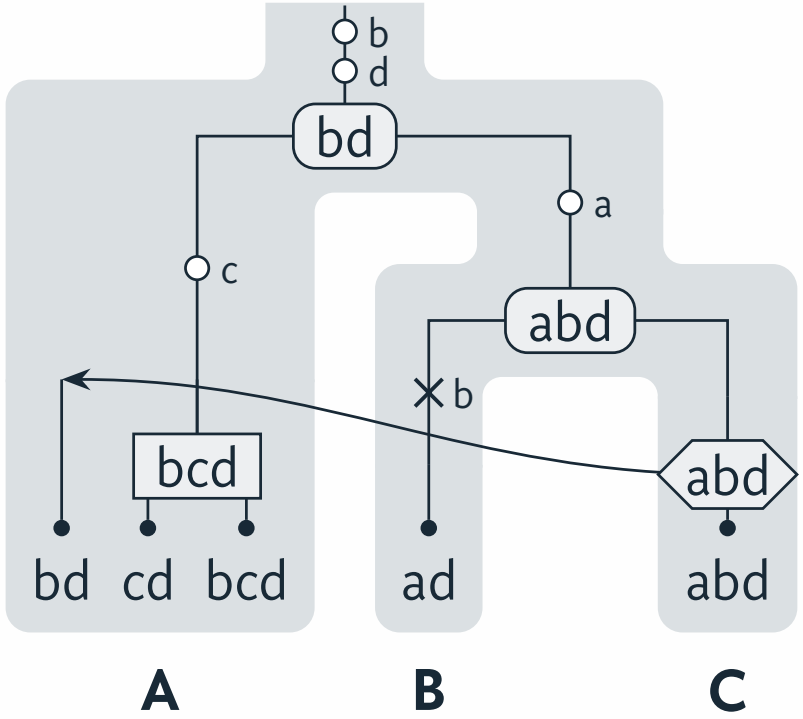
- Choose a **subset of families** for each node so that **losses are minimized**
  - *Exponential choices?*
- Each node must contain a **minimum set** of families, may **contain more**
  - "Extra content"
- Any extra content leads to the same cost, **no matter the actual content**



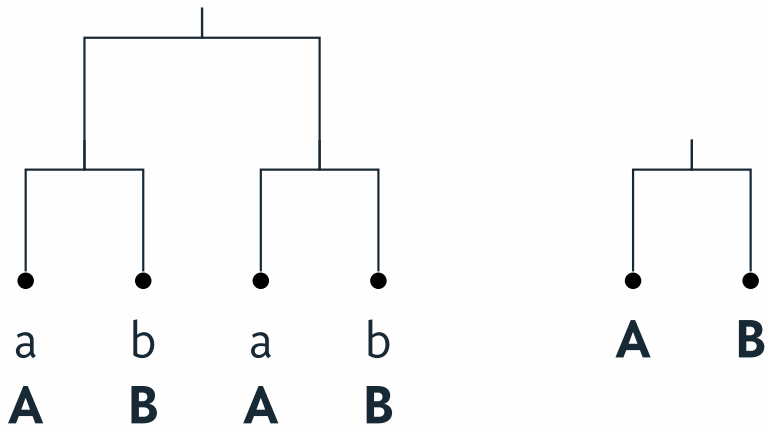
Delabre, El-Mabrouk, Huber, Lafond, Moulton, Noutahi, and Castellanos. "Evolution through segmental duplications and losses: a Super-Reconciliation approach" (May 2020).

PARTS OF A HISTORY

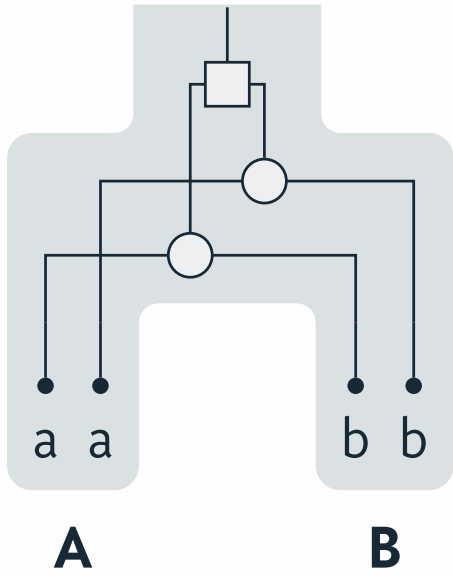
1. Synteny to species tree mapping
2. Ancestral syntenies assignment



NEED FOR JOINT INFERENCE

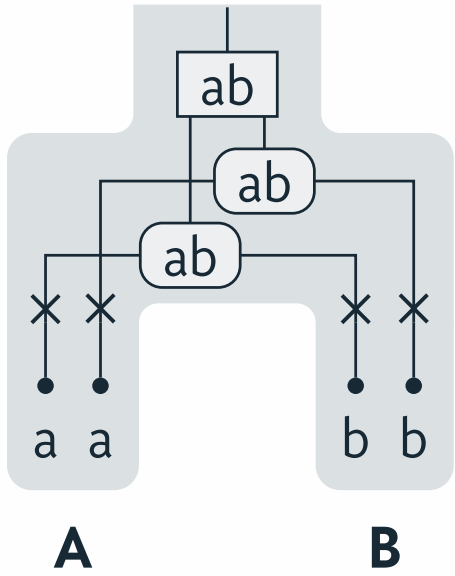


NEED FOR JOINT INFERENCE



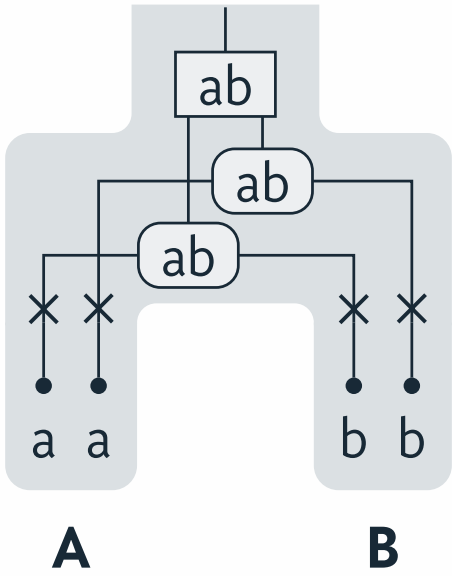
Cost: 1

NEED FOR JOINT INFERENCE

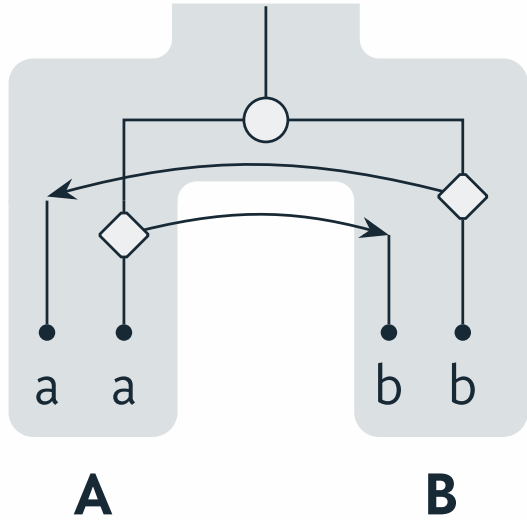


Cost: 1 + 4

NEED FOR JOINT INFERENCE

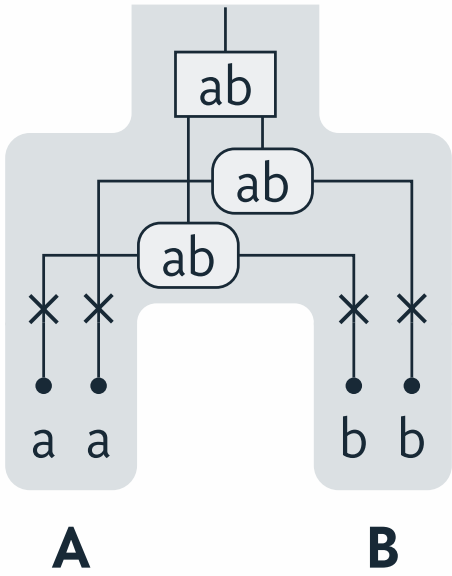


Cost: 1 + 4

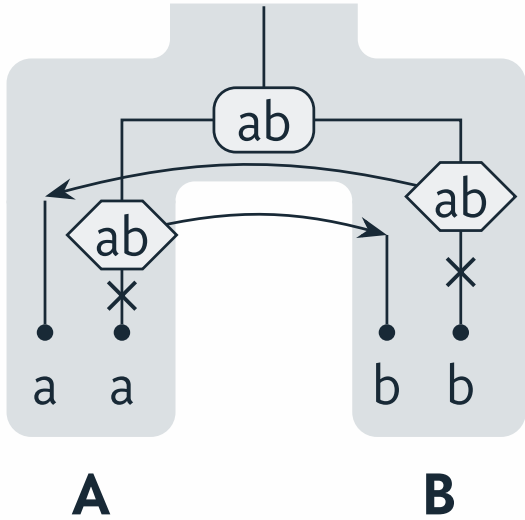


Cost: 2

NEED FOR JOINT INFERENCE



Cost: 1 + 4



Cost: 2 + 2

SUPERDTL ALGORITHM SKETCH

- **For each pair** of nodes (v, σ) in the **synteny** and **species** trees, find minimum cost of **sub-history** below v such that v **is mapped to** σ
  1. Consider **all pairs of species** to which v's children can be mapped
  2. Assign to v either the **minimum synteny** or **any extra content**

- **Overall** minimum cost obtained by looking at values for the **root**
- **Solutions** obtained by backtracing the cost computation

SUPERDTL ALGORITHM SKETCH

- **For each pair** of nodes (v, σ) in the **synteny** and **species** trees, find minimum cost of **sub-history** below v such that v **is mapped to** σ
  1. Consider **all pairs of species** to which v's children can be mapped
  2. Assign to v either the **minimum synteny** or **any extra content**

- **Overall** minimum cost obtained by looking at values for the **root**
- **Solutions** obtained by backtracing the cost computation

- **Complexity:** $\mathcal{O}(NM \times M^2) = \mathcal{O}(NM^3)$ time, $\mathcal{O}(NM)$ space
  N: number of syntenies, M: number of species

OUTLINE

DATASET

- Set of **15 bacterial Cas syntenies** from [Makarova *et al.*, 2020][1]
  - Types I, III and IV (class 1)
  - Gene repetitions discarded
- **Synteny tree** as constructed in the article
  - One synteny per species, one species per subtype
  - **Includes multifurcations**
- **Species tree** based on the topology in [Coleman *et al.*, 2021][2]

[1] Makarova et al. "Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants" (Feb. 2020).

[2] Coleman et al. "A rooted phylogeny resolves early bacterial evolution" (May 2021).

SETTINGS AND PRELIMINARY RESULTS

- Enumerate **all resolutions** of the **synteny tree**, keep best solutions
- **Evaluated event costs**
  - Transfer: 4
  - Duplication: $\{1, 1.5, 2, 2.5, 3\}$
  - Loss: 1

[1] Koonin and Makarova. "Evolutionary plasticity and functional versatility of CRISPR systems" (Jan. 2022).

SETTINGS AND PRELIMINARY RESULTS

- Enumerate **all resolutions** of the **synteny tree**, keep best solutions
- **Evaluated event costs**
  - Transfer: 4
  - Duplication: $\{1, 1.5, 2, 2.5, 3\}$
  - Loss: 1

- **32 optimal solutions**
- All **agree** on the multifurcations **resolution**
- Differences between optimal solutions are not significant
- Inferred scenario **mostly in line** with [Koonin and Makarova, 2022][1]
- See full inferred scenario on poster #10

[1] Koonin and Makarova. "Evolutionary plasticity and functional versatility of CRISPR systems" (Jan. 2022).

OUTLINE

KEY POINTS

- Reconciliation framework is **extended** to include **segmental** (non-ordered) duplication, transfer, loss and gain events
- **Most parsimonious** history under that model **found in** $\mathcal{O}(NM^3)$
- History inference for small **Cas synteny** dataset gives results mostly **in line with current evolutionary hypotheses**

FUTURE WORK

- Unsampled lineages[1]
  - Transfers may only happen between **co-existing species**
  - Some species may be **unsampled** yet act as intermediate hosts
- Dataset extension
  - Include archaeal Cas syntenies
  - Extract data from CRISPRCasDb
- Complexity improvements[2]

---

[1] Weiner and Bansal. "Improved duplication-transfer-loss reconciliation with extinct and unsampled lineages" (Aug. 2021).

[2] Bansal, Alm, and Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss" (June 2012).

Code available at: github.com/UdeM-LBIT/superrec2

YOANN ANSELMETTI
yoann.anselmetti@usherbrooke.ca

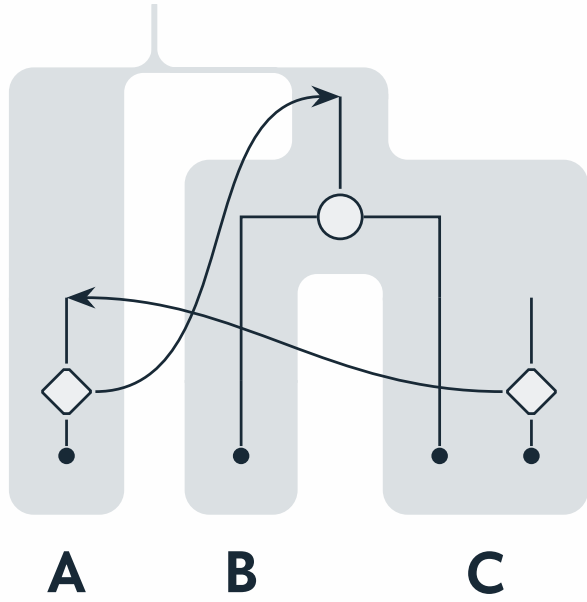
MATTÉO DELABRE
matteo.delabre@umontreal.ca

NADIA EL-MABROUK
mabrouk@iro.umontreal.ca

BACKUP SLIDES

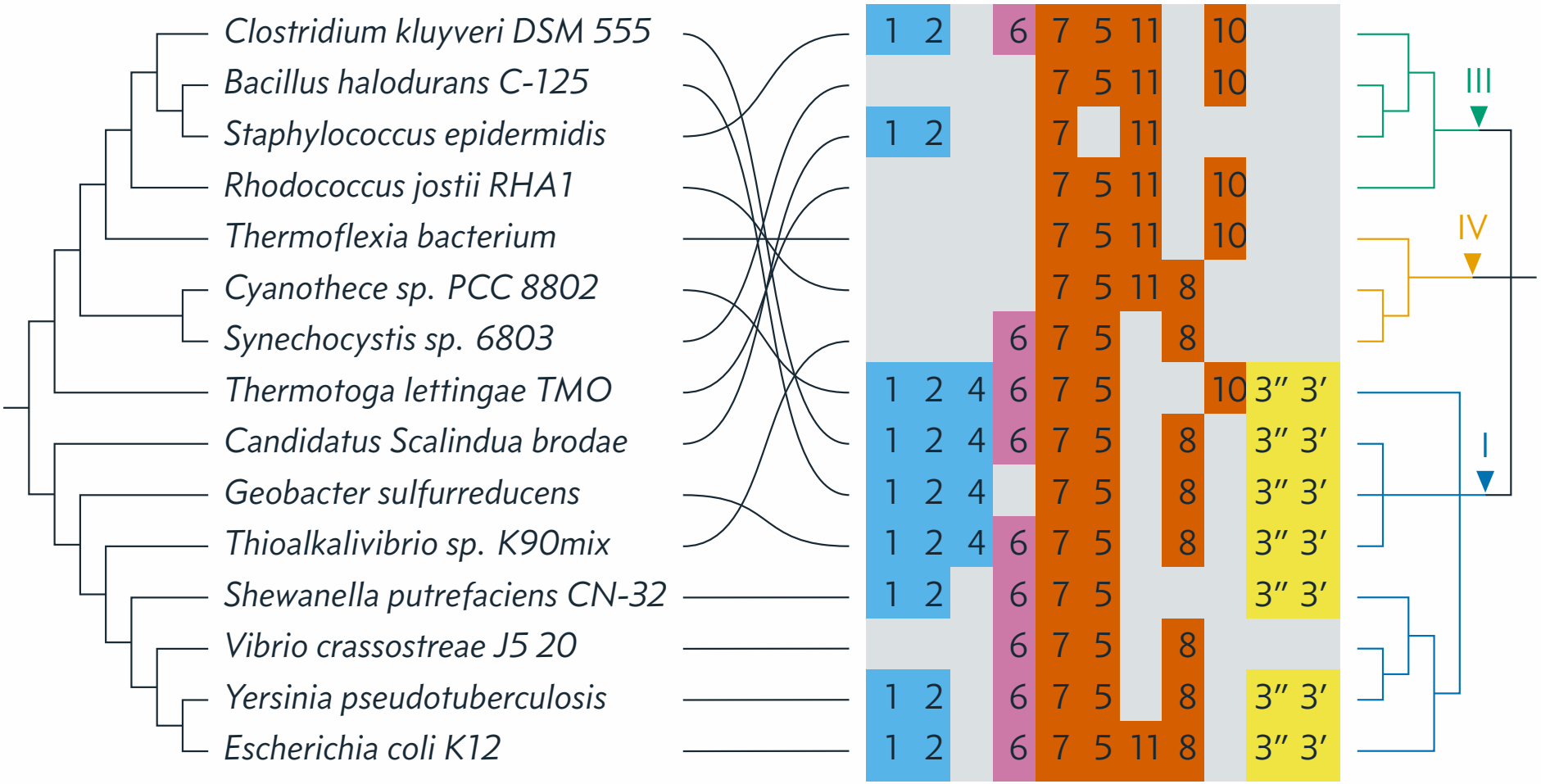
TIME-CONSISTENCY

- Model admits **time-inconsistent** histories
- Forbidding them makes the problem **intractable**



Tofigh, Hallett, and Lagergren. "Simultaneous identification of duplications and lateral gene transfers" (Mar. 2011).

DATASET DETAILS



RESULTS DETAILS

Resolutions of the synteny tree: 27

Solutions per cost value:

$c_T \backslash c_D$	**1**	**1.5**	**2**	**2.5**	**3**
1	1376	1376	1376	1376	1376
2	132	132	132	132	132
3	112	60	60	60	60
4	*32*	*32*	*32*	*32*	*48*
5	288	288	320	32	32
6	320	288	288	288	288