

# Topics of today:

Singular DCJ-indel distance and sorting:

1. Review
2. Capped relational graph of canonical genomes
3. Capped relational graph of singular genomes
4. Indel-potential of cycles via transitions

# Components of a relational graph

Cycle with  $k$  extremity-edges:  $k$ -cycle or  $c_k$

Path with  $k$  extremity-edges:  $k$ -path or  $p_k$

if  $k = 0$  the component is a **singleton**

Disjoint sets of components:  $\left\{ \begin{array}{l} \mathcal{C} = \{c_k : k \geq 2\} : \text{set of cycles } (k \text{ is even}) \\ \mathcal{S} = \{c_k : k = 0\} : \text{set of circular singletons} \\ \mathcal{P}_{\mathbb{A}\mathbb{A}} = \{p_k : \text{starts and ends in } \mathbb{A}\} : \\ \quad \text{set of } \mathbb{A}\mathbb{A}\text{-paths } (k \geq 0 \text{ is even}) \\ \mathcal{P}_{\mathbb{B}\mathbb{B}} = \{p_k : \text{starts and ends in } \mathbb{B}\} : \\ \quad \text{set of } \mathbb{B}\mathbb{B}\text{-paths } (k \geq 0 \text{ is even}) \\ \mathcal{P}_{\mathbb{A}\mathbb{B}} = \{p_k : \text{starts in } \mathbb{A} \text{ and ends in } \mathbb{B}\} : \\ \quad \text{set of } \mathbb{A}\mathbb{B}\text{-paths } (k \geq 1 \text{ is odd}) \end{array} \right.$

**DCJ-sorted (or short) components:** 2-cycles and 1-paths (and 0-cycles and 0-paths)

**Long components:**  $k$ -cycles (with  $k \geq 4$ ) and  $k$ -paths (with  $k \geq 2$ )

**DCJ-sorting a long component  $C$ :** transforming  $C$  into a set of DCJ-sorted components with DCJ-operations

# Types of DCJ operation

**With respect to the position of the cuts:**

**Internal:** either a single-cut operation or two cuts applied in the same component

**Recombination:** each cut is applied in a distinct component

**With respect to the effect on the relational graph:**

**Gaining:** creates one cycle or two  $\mathbb{A}\mathbb{B}$ -paths  
 $\Delta_{\text{DCJ}} = 0$

**Neutral:** preserves the number of cycles and of  $\mathbb{A}\mathbb{B}$ -paths  
 $\Delta_{\text{DCJ}} = 1$

**Losing:** destroys one cycle or two  $\mathbb{A}\mathbb{B}$ -paths  
 $\Delta_{\text{DCJ}} = 2$

# Each component can be sorted separately...

...with an internal gaining DCJ at each step:

Cycle: creates a new cycle at each step



AB-path: creates a new cycle at each step



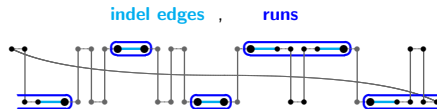
AA-path: creates a new cycle at each step, eventually one step is a single cut (on B) that creates two AB-paths



BB-path: analogous to AA-path

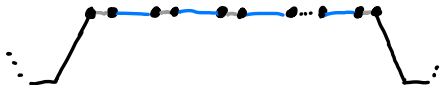
# Accumulating runs

One indel-enclosing cycle:



$$\Lambda = 4$$

Accumulated run:



Each **run** can be **accumulated** with internal gaining DCJ operations and then inserted/deleted at once

⇒ Second upper bound:

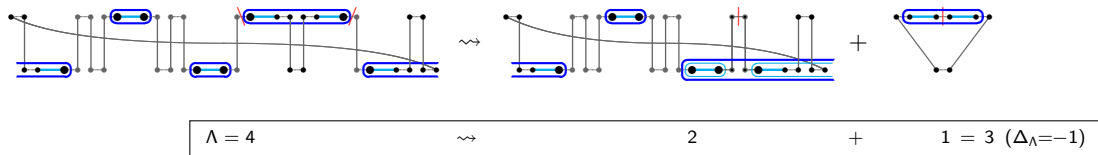
$$d_{\text{DCJ}}^{\text{ID}}(\mathbb{A}, \mathbb{B}) \leq n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{A}\mathbb{B}}|}{2} + \sum_{C \in \text{RG}(\mathbb{A}, \mathbb{B})} \Lambda(C)$$

# Merging runs with internal gaining DCJ operations

DCJ operations can modify the number of runs by at most two:

$$\text{A DCJ operation can have } \begin{cases} \Delta_{\Lambda} = -2 & \text{(merges two pairs of runs)} \\ \Delta_{\Lambda} = -1 & \text{(merges one pair of runs)} \\ \Delta_{\Lambda} = 0 & \text{(preserves the runs)} \\ \Delta_{\Lambda} = 1 & \text{(splits one run)} \\ \Delta_{\Lambda} = 2 & \text{(splits two runs)} \end{cases}$$

A **gaining DCJ operation** applied to two adjacency-edges belonging to the same indel-enclosing component can **decrease** the number of runs:



**Indel-potential  $\lambda(C)$  of a component  $C$ :**

minimum number of runs that we can obtain by DCJ-sorting  $C$  with internal gaining DCJ operations

# Indel-potential of a cycle $C$ - with $\Lambda(C) = 0, 1, 2, 4, 6, 8, \dots$

We will show that  $\lambda(C)$  depends only on the value  $\Lambda(C)$ : denote  $\lambda(C) = \lambda(\Lambda(C))$

$$\Lambda(C) = 0 \Rightarrow \lambda(0) = 0$$

$$\Lambda(C) = 1 \Rightarrow \lambda(1) = 1$$

$$\Lambda(C) = 2 \Rightarrow \lambda(2) = 2$$

$$\Lambda(C) = 4 \Rightarrow \lambda(4) = 3 \text{ (can be verified by listing all cases)}$$

$\Lambda(C) \geq 6$  : extract 3 runs from  $C$  into a new cycle  $\rightarrow$  guarantees that  $\Delta_\Lambda = -2$

two resulting cycles:  $\begin{cases} \text{one with 2 runs} \\ \text{one with } \Lambda(C) - 4 \text{ runs} \end{cases}$

$$\Rightarrow \lambda(6) = \lambda(2) + \lambda(2) = 2 + 2 = 4$$

$$\Rightarrow \lambda(8) = \lambda(2) + \lambda(4) = 2 + 3 = 5$$

$$\Rightarrow \lambda(10) = \lambda(2) + \lambda(6) = 2 + 4 = 6$$

$\Lambda$	$\lambda$
0	0
1	1
2	2
4	3
6	4
8	5
⋮	⋮
⋮	⋮

Induction:  $\begin{cases} \text{hypothesis: } \lambda(\Lambda(C)) = \frac{\Lambda(C)}{2} + 1 \\ \text{base cases: } \lambda(1) = 1, \lambda(2) = 2 \text{ and } \lambda(4) = 3 \end{cases}$

Induction step: in general, for  $\Lambda(C) \geq 6$ , we can state  $\lambda(\Lambda(C)) = \lambda(2) + \lambda(\Lambda(C) - 4)$

$$\begin{aligned} &= 2 + \left( \frac{\Lambda(C) - 4}{2} + 1 \right) \\ &= \frac{\Lambda(C)}{2} + 1 \end{aligned}$$

# Indel-potential $\lambda$ of a path $P$ - with $\Lambda(P) = 0, 1, 2, 3, 4, 5, 6, 7, 8, \dots$

Since  $\lambda(P)$  depends only on the value  $\Lambda(P)$ , we can denote  $\lambda(P) = \lambda(\Lambda(P))$

$$\Lambda(P) = 0 \Rightarrow \lambda(0) = 0$$

$$\Lambda(P) = 1 \Rightarrow \lambda(1) = 1$$

$$\Lambda(P) = 2 \Rightarrow \lambda(2) = 2$$

$$\Lambda(P) \geq 3 : \begin{cases} \text{if } \Lambda(P) \text{ is even, then } \lambda(\Lambda(P)) = \frac{\Lambda(P)}{2} + 1 \\ \text{else } \lambda(\Lambda(P)) = \lambda(\Lambda(P) - 1) \end{cases}$$

$\Lambda$	$\lambda$
0	0
1	1
2	2
3	2
4	3
5	3
6	4
7	4
$\vdots$	$\vdots$
$\vdots$	$\vdots$

In general, for  $\Lambda(P) \geq 2$ , we have

$$\lambda(\Lambda(P)) = \left\lceil \frac{\Lambda(P) + 1}{2} \right\rceil$$



# Indel-potential $\lambda$ of a component $C$

If  $C$  is a singleton:  $\lambda(C) = 1$

If  $C$  is a cycle:

$$\lambda(C) = \begin{cases} 0 & \text{if } \Lambda(C) = 0 \text{ (} C \text{ is indel-free)} \\ 1 & \text{if } \Lambda(C) = 1 \\ \frac{\Lambda(C)}{2} + 1 & \text{if } \Lambda(C) \geq 2 \end{cases}$$

If  $C$  is a path:

$$\lambda(C) = \begin{cases} 0 & \text{if } \Lambda(C) = 0 \text{ (} C \text{ is indel-free)} \\ \lceil \frac{\Lambda(C)+1}{2} \rceil & \text{if } \Lambda(C) \geq 1 \end{cases}$$

In general, for any component  $C$ :

$$\lambda(C) = \begin{cases} 0 & \text{if } \Lambda(C) = 0 \text{ (} C \text{ is indel-free)} \\ \lceil \frac{\Lambda(C)+1}{2} \rceil & \text{if } \Lambda(C) \geq 1 \end{cases}$$

$\Lambda$	$\lambda$
0	0
1	1
2	2
3	2
4	3
5	3
6	4
7	4
$\vdots$	$\vdots$
$\vdots$	$\vdots$

paths and cycles  
 paths, cycles and singletons  
 paths and cycles  
 paths  
 paths and cycles  
 paths  
 paths and cycles  
 paths

**Third upper bound:**

$$d_{DCJ}^{\text{ID}}(\mathbb{A}, \mathbb{B}) \leq n - |C| - \frac{|\mathcal{P}_{AB}|}{2} + \sum_{C \in RG} \lambda(C)$$

(gaining DCJ operations + indels sorting components separately)

## Effect of a DCJ operation on the third upper bound:

$$\text{DCJ-types of DCJ operation} \begin{cases} \Delta_{\text{DCJ}} = 0 \text{ (gaining):} & \text{creates one cycle or two } \mathbb{A}\mathbb{B}\text{-paths} \\ \Delta_{\text{DCJ}} = 1 \text{ (neutral):} & \text{preserves the numbers of cycles and of } \mathbb{A}\mathbb{B}\text{-paths} \\ \Delta_{\text{DCJ}} = 2 \text{ (losing):} & \text{destroys one cycle or two } \mathbb{A}\mathbb{B}\text{-paths} \end{cases}$$

$$\text{Indel-types of DCJ operation} \begin{cases} \Delta_{\lambda} = -2 & : \text{decreases the overall indel-potential by two} \\ \Delta_{\lambda} = -1 & : \text{decreases the overall indel-potential by one} \\ \Delta_{\lambda} = 0 & : \text{does not change the overall indel-potential} \\ \Delta_{\lambda} = 1 & : \text{increases the overall indel-potential by one} \\ \Delta_{\lambda} = 2 & : \text{increases the overall indel-potential by two} \end{cases}$$

Effect of a DCJ operation  $\rho$  on the third upper bound:  $\Delta_{\text{DCJ}}^{\lambda}(\rho) = \Delta_{\text{DCJ}}(\rho) + \Delta_{\lambda}(\rho)$

$$\text{DCJ Operations that can decrease the third upper bound:} \begin{cases} \Delta_{\text{DCJ}} = 0 \text{ (gaining) and } \Delta_{\lambda} = -2 & : \Delta_{\text{DCJ}}^{\lambda} = -2 \\ \Delta_{\text{DCJ}} = 0 \text{ (gaining) and } \Delta_{\lambda} = -1 & : \Delta_{\text{DCJ}}^{\lambda} = -1 \\ \Delta_{\text{DCJ}} = 1 \text{ (neutral) and } \Delta_{\lambda} = -2 & : \Delta_{\text{DCJ}}^{\lambda} = -1 \end{cases}$$

- ▶ By definition: any internal gaining DCJ operation  $\rho$  (applied to a single component) has  $\Delta_{\lambda}(\rho) \geq 0$  and, consequently,  $\Delta_{\text{DCJ}}^{\lambda}(\rho) \geq 0$
- ▶ Any losing DCJ operation  $\rho$  has  $\Delta_{\text{DCJ}}^{\lambda}(\rho) \geq 0$

# DCJ operations involving cycles

$\Lambda$	$\lambda$
0	0
1	1
2	2
4	3
6	4
8	5
⋮	⋮
⋮	⋮

- ▶ Any recombination involving two cycles is losing and has  $\Delta_{\text{DCJ}}^{\lambda} \geq 0$  (cannot decrease the DCJ-indel distance)
  - ▶ An internal DCJ operation  $\rho$  applied to a cycle  $C$  can be:
    - ▶ Gaining, with  $\Delta_{\text{DCJ}}^{\lambda}(\rho) \geq 0$  (cannot decrease the DCJ-indel distance)
    - ▶ Neutral ( $\Delta_{\text{DCJ}}(\rho) = 1$ ):
      - If  $\Lambda(C) \geq 4$ , the DCJ  $\rho$  can merge at most two pairs of runs:  $\Delta_{\Lambda}(\rho) \geq -2$  and  $\Delta_{\lambda}(\rho) \geq -1$
- $\Rightarrow$  Any internal neutral DCJ operation applied to a cycle has  $\Delta_{\text{DCJ}}^{\lambda} \geq 0$  (cannot decrease the DCJ-indel distance)

If singular genomes  $\mathbb{A}$  and  $\mathbb{B}$  are circular, the graph  $RG(\mathbb{A}, \mathbb{B})$  has only cycles (and eventually singletons).

In this case:

$$d_{\text{DCJ}}^{\text{ID}}(\mathbb{A}, \mathbb{B}) = n - |\mathcal{C}| + \sum_{C \in \mathcal{C}} \lambda(C)$$

# DCJ operations involving paths

- ▶ Any recombination involving a path and a cycle is losing and has  $\Delta_{\text{DCJ}}^\lambda \geq 0$  (cannot decrease the DCJ-indel distance)

$\Lambda$	$\lambda$
0	0
1	1
2	2
3	2
4	3
5	3
6	4
7	4
⋮	⋮
⋮	⋮

- ▶ An internal DCJ operation  $\rho$  applied to a path  $P$  can be:
  - ▶ Gaining, with  $\Delta_{\text{DCJ}}^\lambda(\rho) \geq 0$  (cannot decrease the DCJ-indel distance)
  - ▶ Neutral ( $\Delta_{\text{DCJ}}(\rho) = 1$ ):
    - If  $\Lambda(P) \geq 4$ , the DCJ  $\rho$  can merge at most two pairs of runs:  $\Delta_\Lambda(\rho) \geq -2$  and  $\Delta_\lambda(\rho) \geq -1$
- ⇒ Any internal neutral DCJ operation applied to a path has  $\Delta_{\text{DCJ}}^\lambda \geq 0$  (cannot decrease the DCJ-indel distance)

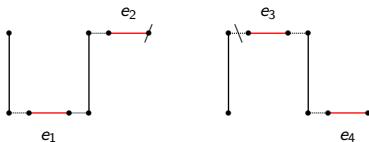
# Path recombinations can have $\Delta_{DCJ}^\lambda \leq -1$

A gaining (**deducting**) path recombination with  $\Delta_{DCJ}^\lambda = -2$ :

Sources

$$(\sum \lambda = 2 + 2 = 4)$$

AA + BB  
2 runs + 2 runs

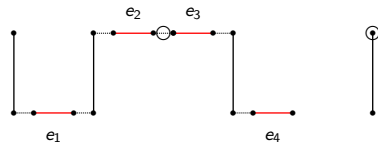


→  
gaining  
DCJ

Resultants

$$(\sum \lambda = 2 + 0 = 2)$$

AB + AB  
3 runs + no run



$$\cancel{AA} + \cancel{BB} = \begin{cases} AB_{\mathcal{B}AB} + AB_{\epsilon} \\ (AB_{\mathcal{A}BA} + AB_{\epsilon}) \\ (AB_{\mathcal{A}} + AB_{\mathcal{B}}) \end{cases} \quad (\text{all variants have } \Delta_{DCJ}^\lambda = -2)$$

**Deducting path recombinations**

have  $\Delta_{DCJ}^\lambda \leq -1$

**General DCJ-indel distance formula:**

$$d_{DCJ}^{ID}(\mathbb{A}, \mathbb{B}) = n - |C| - \frac{|P_{AB}|}{2} + \sum_{C \in RG} \lambda(C) - \delta,$$

where  $\delta$  is the value obtained by optimizing deducting path recombinations

# Optimizing deducing path recombinations (for computing $\delta$ )

$$\text{Run-type of a path} \begin{cases} \varepsilon & \equiv \varepsilon \text{ (empty)} \\ \mathcal{A}\mathcal{B}\mathcal{A}\mathcal{B} \dots \mathcal{A} & \equiv \mathcal{A} \text{ (odd)} \\ \mathcal{B}\mathcal{A}\mathcal{B}\mathcal{A} \dots \mathcal{B} & \equiv \mathcal{B} \text{ (odd)} \\ \mathcal{A}\mathcal{B}\mathcal{A}\mathcal{B} \dots \mathcal{A}\mathcal{B} & \equiv \mathcal{A}\mathcal{B} \text{ (even)} \\ \mathcal{B}\mathcal{A}\mathcal{B}\mathcal{A} \dots \mathcal{B}\mathcal{A} & \equiv \mathcal{B}\mathcal{A} \text{ (even)} \end{cases} \quad \text{Path types} \begin{cases} \mathcal{A}\mathcal{A}_\varepsilon, \mathcal{A}\mathcal{A}_\mathcal{A}, \mathcal{A}\mathcal{A}_\mathcal{B}, \mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} (\equiv \mathcal{A}\mathcal{A}_{\mathcal{B}\mathcal{A}}) \\ \mathcal{B}\mathcal{B}_\varepsilon, \mathcal{B}\mathcal{B}_\mathcal{A}, \mathcal{B}\mathcal{B}_\mathcal{B}, \mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}} (\equiv \mathcal{B}\mathcal{B}_{\mathcal{B}\mathcal{A}}) \\ \mathcal{A}\mathcal{B}_\varepsilon, \mathcal{A}\mathcal{B}_\mathcal{A}, \mathcal{A}\mathcal{B}_\mathcal{B}, \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}, \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}} \\ \Rightarrow \text{an } \mathcal{A}\mathcal{B}\text{-path is always read from } \mathcal{A} \text{ to } \mathcal{B} \end{cases}$$

Deducting path recombinations that allow the best reuse of the resultants:

sources	resultants	$\Delta_\lambda$	$\Delta_{\text{DCJ}}$	$\Delta_{\text{DCJ}}^\lambda$
$\mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} + \mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\bullet + \bullet$	-2	0	-2
$\mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} + \mathcal{B}\mathcal{B}_\mathcal{A}$	$\bullet + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	-1	0	-1
$\mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} + \mathcal{B}\mathcal{B}_\mathcal{B}$	$\bullet + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	-1	0	-1
$\mathcal{A}\mathcal{A}_\mathcal{A} + \mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\bullet + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	-1	0	-1
$\mathcal{A}\mathcal{A}_\mathcal{B} + \mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\bullet + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	-1	0	-1
$\mathcal{A}\mathcal{A}_\mathcal{A} + \mathcal{B}\mathcal{B}_\mathcal{A}$	$\bullet + \bullet$	-1	0	-1
$\mathcal{A}\mathcal{A}_\mathcal{B} + \mathcal{B}\mathcal{B}_\mathcal{B}$	$\bullet + \bullet$	-1	0	-1

sources	resultants	$\Delta_\lambda$	$\Delta_{\text{DCJ}}$	$\Delta_{\text{DCJ}}^\lambda$
$\mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} + \mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}}$	$\mathcal{A}\mathcal{A}_\mathcal{A} + \mathcal{A}\mathcal{A}_\mathcal{B}$	-2	+1	-1
$\mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}} + \mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\mathcal{B}\mathcal{B}_\mathcal{A} + \mathcal{B}\mathcal{B}_\mathcal{B}$	-2	+1	-1
$\mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\bullet + \mathcal{A}\mathcal{A}_\mathcal{A}$	-2	+1	-1
$\mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	$\bullet + \mathcal{A}\mathcal{A}_\mathcal{B}$	-2	+1	-1
$\mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}} + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\bullet + \mathcal{B}\mathcal{B}_\mathcal{B}$	-2	+1	-1
$\mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}} + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	$\bullet + \mathcal{B}\mathcal{B}_\mathcal{A}$	-2	+1	-1
$\mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}} + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	$\bullet + \bullet$	-2	+1	-1

$$\bullet \begin{cases} \mathcal{A}\mathcal{B}_\varepsilon \\ \mathcal{A}\mathcal{B}_\mathcal{A} \\ \mathcal{A}\mathcal{B}_\mathcal{B} \end{cases}$$

Sources:

$\mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}} : \mathcal{W}$   
 $\mathcal{A}\mathcal{A}_\mathcal{A} : \bar{\mathcal{W}}$   
 $\mathcal{A}\mathcal{A}_\mathcal{B} : \underline{\mathcal{W}}$   
 $\mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}} : \mathcal{M}$   
 $\mathcal{B}\mathcal{B}_\mathcal{A} : \bar{\mathcal{M}}$   
 $\mathcal{B}\mathcal{B}_\mathcal{B} : \underline{\mathcal{M}}$   
 $\mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}} : \mathcal{Z}$   
 $\mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}} : \mathcal{N}$

Path recombinations with  $\Delta_{\text{DCJ}}^\lambda = 0$  creating resultants that can be used in deducing recombinations:

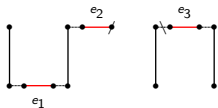
sources	resultants	$\Delta_\lambda$	$\Delta_{\text{DCJ}}$	$\Delta_{\text{DCJ}}^\lambda$
$\mathcal{A}\mathcal{A}_\mathcal{A} + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	$\bullet + \mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}}$	-1	+1	0
$\mathcal{A}\mathcal{A}_\mathcal{B} + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\bullet + \mathcal{A}\mathcal{A}_{\mathcal{A}\mathcal{B}}$	-1	+1	0
$\mathcal{B}\mathcal{B}_\mathcal{A} + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\bullet + \mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	-1	+1	0
$\mathcal{B}\mathcal{B}_\mathcal{B} + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	$\bullet + \mathcal{B}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	-1	+1	0

sources	resultants	$\Delta_\lambda$	$\Delta_{\text{DCJ}}$	$\Delta_{\text{DCJ}}^\lambda$
$\mathcal{A}\mathcal{A}_\mathcal{A} + \mathcal{B}\mathcal{B}_\mathcal{B}$	$\bullet + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	0	0	0
$\mathcal{A}\mathcal{A}_\mathcal{B} + \mathcal{B}\mathcal{B}_\mathcal{A}$	$\bullet + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	0	0	0
$\mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}} + \mathcal{A}\mathcal{B}_{\mathcal{A}\mathcal{B}}$	$\mathcal{A}\mathcal{A}_\mathcal{A} + \mathcal{B}\mathcal{B}_\mathcal{B}$	-2	+2	0
$\mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}} + \mathcal{A}\mathcal{B}_{\mathcal{B}\mathcal{A}}$	$\mathcal{A}\mathcal{A}_\mathcal{B} + \mathcal{B}\mathcal{B}_\mathcal{A}$	-2	+2	0

# Optimizing deducing path recombinations (for computing $\delta$ )

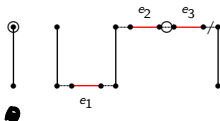
Deducing chain of path recombinations  $\left\{ \begin{array}{l} \text{transforming } 2 \times AA_{AB} + BB_A + BB_B \\ \text{into } 3 \times AB_\varepsilon + AB_B \\ \text{with overall } \Delta_{DCJ}^\lambda = -3 \end{array} \right.$

$AA_{AB} + BB_A$   
2 runs + 1 run  
 $\lambda = 2 + \lambda = 1$



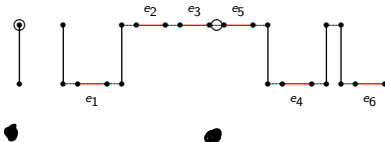
$(\Delta_{DCJ}^\lambda = -1)$   
gaining DCJ

$AB_\varepsilon + AB_{BA}$   
no run + 2 runs  
 $\lambda = 0 + \lambda = 2$

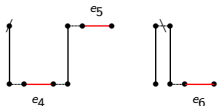


$AB_\varepsilon + AB_B$   
no run + 3 runs  
 $\lambda = 0 + \lambda = 2$

$(\Delta_{DCJ}^\lambda = -1)$   
neutral DCJ

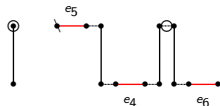


$AA_{AB} + BB_B$   
2 runs + 1 run  
 $\lambda = 2 + \lambda = 1$



$(\Delta_{DCJ}^\lambda = -1)$   
gaining DCJ

$AB_\varepsilon + AB_{AB}$   
no run + 2 runs  
 $\lambda = 0 + \lambda = 2$



id		sources			resultants			$\Delta_{DCJ}^\lambda$	scr	
$\mathcal{P}$	WM	$AA_{AB}$	$BB_{AB}$	—	—	—	$2 \times \bullet$	-2	-1	
$\mathcal{Q}$	WWM	$2 \times AA_{AB}$	$BB_A + BB_B$	—	—	—	$4 \times \bullet$	-3	-3/4	
	MMW	$AA_A + AA_B$	$2 \times BB_{AB}$	—	—	—	$4 \times \bullet$	-3	-3/4	
$\mathcal{T}$	WZM	$AA_{AB}$	$BB_A$	$AB_{AB}$	—	—	$3 \times \bullet$	-2	-2/3	
	WWM	$2 \times AA_{AB}$	$BB_A$	—	$AA_B$	—	$2 \times \bullet$	-2	-2/3	
	WNM	$AA_{AB}$	$BB_B$	$AB_{BA}$	—	—	$3 \times \bullet$	-2	-2/3	
	WWM	$2 \times AA_{AB}$	$BB_B$	—	$AA_A$	—	$2 \times \bullet$	-2	-2/3	
	MNW	$AA_A$	$BB_{AB}$	$AB_{BA}$	—	—	$3 \times \bullet$	-2	-2/3	
	MMW	$AA_A$	$2 \times BB_{AB}$	—	—	$BB_B$	$2 \times \bullet$	-2	-2/3	
	MZW	$AA_B$	$BB_{AB}$	$AB_{AB}$	—	—	$3 \times \bullet$	-2	-2/3	
	MMW	$AA_B$	$2 \times BB_{AB}$	—	—	$BB_A$	$2 \times \bullet$	-2	-2/3	
$\mathcal{S}$	ZN	—	—	$AB_{AB} + AB_{BA}$	—	—	$2 \times \bullet$	-1	-1/2	
	WM	$AA_A$	$BB_A$	—	—	—	$2 \times \bullet$	-1	-1/2	
	WM	$AA_B$	$BB_B$	—	—	—	$2 \times \bullet$	-1	-1/2	
	WM	$AA_{AB}$	$BB_A$	—	—	$AB_{BA}$	$\bullet$	-1	-1/2	
	WM	$AA_{AB}$	$BB_B$	—	—	$AB_{AB}$	$\bullet$	-1	-1/2	
	WZ	$AA_{AB}$	—	$AB_{AB}$	$AA_A$	—	$\bullet$	-1	-1/2	
	WN	$AA_{AB}$	—	$AB_{BA}$	$AA_B$	—	$\bullet$	-1	-1/2	
	WW	$2 \times AA_{AB}$	—	—	$AA_A + AA_B$	—	—	-1	-1/2	
	MW	$AA_A$	$BB_{AB}$	—	—	—	$AB_{AB}$	$\bullet$	-1	-1/2
	MW	$AA_B$	$BB_{AB}$	—	—	—	$AB_{BA}$	$\bullet$	-1	-1/2
	MZ	—	$BB_{AB}$	$AB_{AB}$	—	$BB_B$	—	$\bullet$	-1	-1/2
	MN	—	$BB_{AB}$	$AB_{BA}$	—	$BB_A$	—	$\bullet$	-1	-1/2
	MM	—	$2 \times BB_{AB}$	—	—	$BB_A + BB_B$	—	-1	-1/2	



	id	sources			resultants				$\Delta_{DCJ}^\lambda$	scr
$\mathcal{M}$	$ZZ\bar{W}\bar{M}$	$AA_B$	$BB_A$	$2 \times AB_{AB}$	—	—	—	$4 \times \bullet$	-2	-1/2
	$NN\bar{W}\bar{M}$	$AA_A$	$BB_B$	$2 \times AB_{BA}$	—	—	—	$4 \times \bullet$	-2	-1/2
$\mathcal{N}$	$Z\bar{W}\bar{M}$	$AA_B$	$BB_A$	$AB_{AB}$	—	—	$AB_{BA}$	$2 \times \bullet$	-1	-1/3
	$ZZ\bar{W}$	$AA_B$	—	$2 \times AB_{AB}$	$AA_A$	—	—	$2 \times \bullet$	-1	-1/3
	$ZZ\bar{M}$	—	$BB_A$	$2 \times AB_{AB}$	—	$BB_B$	—	$2 \times \bullet$	-1	-1/3
	$N\bar{W}\bar{M}$	$AA_A$	$BB_B$	$AB_{BA}$	—	—	$AB_{AB}$	$2 \times \bullet$	-1	-1/3
	$NN\bar{W}$	$AA_A$	—	$2 \times AB_{BA}$	$AA_B$	—	—	$2 \times \bullet$	-1	-1/3
	$NN\bar{M}$	—	$BB_B$	$2 \times AB_{BA}$	—	$BB_A$	—	$2 \times \bullet$	-1	-1/3

Sources:

$W$  :  $AA_{AB}$

$\bar{W}$  :  $AA_A$

$\underline{W}$  :  $AA_B$

$M$  :  $BB_{AB}$

$\bar{M}$  :  $BB_A$

$\underline{M}$  :  $BB_B$

$Z$  :  $AB_{AB}$

$N$  :  $AB_{BA}$

DCJ-indel distance formula:

$$d_{DCJ}^{ID}(A, B) = n - |C| - \frac{|P_{AB}|}{2} + \sum_{C \in RG} \lambda(C) - \delta,$$

where  $\delta$  is the value obtained by optimizing deducting path recombinations:

$$\delta = 2P + 3Q + 2T + S + 2M + N$$

the values  $P$ ,  $Q$ ,  $T$ ,  $S$ ,  $M$  and  $N$  refer to the corresponding number of chains of deducting path recombinations of each type and can be obtained by a greedy approach (simple top-down screening of the table)

## Singular DCJ-indel model - summary

**DCJ-indel distance:**  $d_{\text{DCJ}}^{\text{ID}}(\mathbb{A}, \mathbb{B}) = n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{A}\mathbb{B}}|}{2} + \sum_{C \in \mathcal{R}\mathcal{G}} \lambda(C) - \delta$ , where  $\delta$  is the value obtained by optimizing deducting path recombinations

**A and B are circular:**  $d_{\text{DCJ}}^{\text{ID}}(\mathbb{A}, \mathbb{B}) = n - |\mathcal{C}| + \sum_{C \in \mathcal{R}\mathcal{G}} \lambda(C)$

Computing the distance and sorting can be done in **linear time**.

# Quiz 1

1 Which of the following statements is correct?

A Any DCJ operation has  $\Delta_{\text{DCJ}}^\lambda \geq 0$ .

B Any gaining DCJ operation has  $\Delta_{\text{DCJ}}^\lambda \geq 0$ .

C Any internal gaining DCJ operation has  $\Delta_{\text{DCJ}}^\lambda \geq 0$ .

2 Which of the following statements about the DCJ-indel model are true?

A Any DCJ that decreases the number of runs has  $\Delta_\lambda < 0$ .  $\Delta_\lambda < 0$  ✓

B If the input genomes are circular, sorting each component of the relational graph separately is an optimal approach.

C An optimal sequence of DCJ operations and indels sorting one singular genome into another can have gaining, neutral and losing DCJs.

D The triangular inequality holds for the DCJ-indel distance.

E The DCJ-indel distance can be distinct from the restricted DCJ-indel distance.

# Capped relational graph

Capping is a procedure that circularizes all paths of a relational graph by adding **caps (artificial genes)**:

- ▶ if the capping is optimal, the genomic distance is preserved
- ▶ from the capped relational diagram we can derive genomes composed only of circular chromosomes

A capping may require adjacencies between caps:

$\Gamma_{\mathbb{A}}$ : represents an adjacency between caps in genome  $\mathbb{A}$

$\Gamma_{\mathbb{B}}$ : represents an adjacency between caps in genome  $\mathbb{B}$ .

# Capped relational graph of canonical genomes

Optimally linking paths from  $RG(\mathbb{A}, \mathbb{B})$  of canonical genomes  $\mathbb{A}$  and  $\mathbb{B}$  into cycles can be done as follows:

id	paths	linking cycle		$\Delta n$	$\Delta c$	$\Delta(2\mathbb{A}\mathbb{B})$	$\Delta_{\text{DCJ}}$
1	$\mathbb{A}\mathbb{B}$	$(\mathbb{A}\mathbb{B})$		+0.5	+1	-0.5	0
2	$\mathbb{A}\mathbb{A} + \mathbb{B}\mathbb{B}$	$(\mathbb{A}\mathbb{A}, \mathbb{B}\mathbb{B})$		+1	+1	0	0
3	$\mathbb{A}\mathbb{A}$	$(\mathbb{A}\mathbb{A}, \Gamma_{\mathbb{B}})$	$\cup$	+1	+1	0	0
4	$\mathbb{B}\mathbb{B}$	$(\mathbb{B}\mathbb{B}, \Gamma_{\mathbb{A}})$	$\cap$	+1	+1	0	0

- { Closing an  $\mathbb{A}\mathbb{A}$ -path (over-represented in genome  $\mathbb{A}$  and marked with a  $\cup$ ) requires an adjacency  $\Gamma_{\mathbb{B}}$ .
- { Closing a  $\mathbb{B}\mathbb{B}$ -path (over-represented in genome  $\mathbb{B}$  and marked with a  $\cap$ ) requires an adjacency  $\Gamma_{\mathbb{A}}$ .

Any capping producing linking cycles as indicated on the table above is optimal:

- ▶ The value  $\Delta_{\text{DCJ}} = \Delta n - \Delta c - \Delta(2\mathbb{A}\mathbb{B})$  is the DCJ-effect produced by each type of linking cycle.
- ▶ All given linking cycles have  $\Delta_{\text{DCJ}} = 0$ , therefore they preserve the DCJ distance.

Let  $\begin{cases} \kappa_{\mathbb{A}}: \text{number of linear chromosomes in } \mathbb{A} \\ \kappa_{\mathbb{B}}: \text{number of linear chromosomes in } \mathbb{B} \end{cases}$

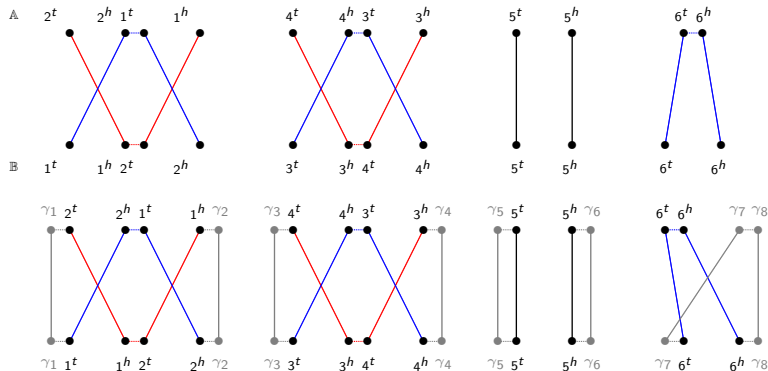
The difference between the number of  $\mathbb{A}\mathbb{A}$ - and of  $\mathbb{B}\mathbb{B}$ -paths is equal to the difference between  $\kappa_{\mathbb{A}}$  and  $\kappa_{\mathbb{B}}$ .

An optimal capping that maximizes the number of linking cycles of type **2** minimizes the number of caps:

- { The number of caps to be added is exactly  $p_* = \max\{\kappa_{\mathbb{A}}, \kappa_{\mathbb{B}}\}$ .
- { The number of adjacencies between caps is exactly  $a_* = |\kappa_{\mathbb{A}} - \kappa_{\mathbb{B}}|$ .

# Capped relational graph of canonical genomes - example

$$\mathbb{A} = [2^t \ 1] \ [4^h \ 3] \ [5] \ (6) \quad \text{and} \quad \mathbb{B} = [1^t \ 2] \ [3^h \ 4] \ [5] \ [6] \quad ; \quad p_* = 4 \quad \text{and} \quad a_* = 1$$



$$\begin{aligned} d_{\text{DCJ}} &= n - |\mathcal{C}| - \frac{|\mathcal{P}_{\mathbb{A}\mathbb{B}}|}{2} \\ &= 6 - 0 - 1 \\ &= 5 \end{aligned}$$

$$\begin{aligned} d_{\text{DCJ}} &= n + p_* - |\mathcal{C}| \\ &= 6 + 4 - 5 \\ &= 5 \end{aligned}$$

Any way of pairing the cap extremities  $\gamma_1, \gamma_2, \dots, \gamma_8$  is valid; possible derived circular genomes are:

$$\begin{aligned} \mathbb{A}_o &= (2^t \ 1^h \ W) \ (4^h \ 3^t \ X) \ (5^t \ Y) \ (6) \ (Z) \quad \text{and} \quad \mathbb{B}_o = (1^t \ 2^h \ W) \ (3^h \ 4^t \ X) \ (5^t \ Y) \ (6) \ (Z) \\ &\quad (W^h = \gamma_1, W^t = \gamma_2, X^h = \gamma_3, X^t = \gamma_4, Y^h = \gamma_5, Y^t = \gamma_6, Z^h = \gamma_7, Z^t = \gamma_8) \\ &\quad \text{or} \end{aligned}$$

$$\begin{aligned} \mathbb{A}_o &= (2^t \ 1^h \ W \ 4^h \ 3^t \ X \ 5^t \ Y \ Z) \ (6) \quad \text{and} \quad \mathbb{B}_o = (1^t \ 2^h \ W \ 3^h \ 4^t \ X \ 5^t \ Y \ 6) \ (Z) \\ &\quad (W^h = \gamma_3, W^t = \gamma_2, X^h = \gamma_5, X^t = \gamma_4, Y^h = \gamma_7, Y^t = \gamma_6, Z^h = \gamma_1, Z^t = \gamma_8) \end{aligned}$$

# Capping the relational graph - singular genomes

The sources of each chain of deducing recombinations must be properly linked together into a single cycle.

{ Unbalanced chains over-represented in genome  $\mathbb{A}$  are marked with a  $\cup$   
 $\mathbb{B}\mathbb{B}_\varepsilon \prec \Gamma_{\mathbb{B}}$ : a path  $\mathbb{B}\mathbb{B}_\varepsilon$  is preferred to close a  $\cup$ -unbalanced chain; if it does not exist, an adjacency  $\Gamma_{\mathbb{B}}$  is used  
 { Unbalanced chains over-represented in genome  $\mathbb{B}$  are marked with a  $\cap$   
 $\mathbb{A}\mathbb{A}_\varepsilon \prec \Gamma_{\mathbb{A}}$ : a path  $\mathbb{A}\mathbb{A}_\varepsilon$  is preferred to close a  $\cap$ -unbalanced chain; if it does not exist, an adjacency  $\Gamma_{\mathbb{A}}$  is used

In order to give the correct order of linking { a path  $\mathbb{A}\mathbb{B}_{\mathbb{A}\mathbb{B}}$  can be represented by  $\mathbb{B}\mathbb{A}_{\mathbb{B}\mathbb{A}}$   
 a path  $\mathbb{A}\mathbb{B}_{\mathbb{B}\mathbb{A}}$  can be represented by  $\mathbb{B}\mathbb{A}_{\mathbb{A}\mathbb{B}}$

	id	sources	linking cycle		$\Delta n$	$\Delta c$	$\Delta(2\mathbb{A}\mathbb{B})$	$\Delta\lambda$	$\Delta\lambda_{\text{DCJ}}^\lambda$
$\mathcal{P}$	$\overline{\text{WM}}$	$\mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}} + \mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}}$	$(\mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{B}\mathbb{A}})$		+1	+1	0	-2	-2
$\mathcal{Q}$	$\overline{\text{W}\overline{\text{M}}\overline{\text{M}}}$	$2 \times \mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}} + \mathbb{B}\mathbb{B}_{\mathbb{A}}$	$(\mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{B}}, \mathbb{A}\mathbb{A}_{\mathbb{B}\mathbb{A}}, \mathbb{B}\mathbb{B}_{\mathbb{A}})$		+2	+1	0	-4	-3
	$\overline{\text{M}\overline{\text{M}}\overline{\text{W}}}$	$2 \times \mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}} + \mathbb{A}\mathbb{A}_{\mathbb{A}} + \mathbb{A}\mathbb{A}_{\mathbb{B}}$	$(\mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}}, \mathbb{A}\mathbb{A}_{\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{B}\mathbb{A}}, \mathbb{A}\mathbb{A}_{\mathbb{A}})$		+2	+1	0	-4	-3
$\mathcal{T}$	$\overline{\text{WZ}\overline{\text{M}}}$	$\mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}} + \mathbb{B}\mathbb{B}_{\mathbb{A}} + \mathbb{A}\mathbb{B}_{\mathbb{A}\mathbb{B}}$	$(\mathbb{A}\mathbb{B}_{\mathbb{A}\mathbb{B}}, \mathbb{A}\mathbb{A}_{\mathbb{B}\mathbb{A}}, \mathbb{B}\mathbb{B}_{\mathbb{A}})$		+1.5	+1	-0.5	-3	-2
	$\overline{\text{W}\overline{\text{W}}\overline{\text{M}}}$	$2 \times \mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}} + \mathbb{B}\mathbb{B}_{\mathbb{A}}$	$(\mathbb{A}\mathbb{A}_{\mathbb{B}\mathbb{A}}, \mathbb{B}\mathbb{B}_{\mathbb{A}}, \mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_\varepsilon \prec \Gamma_{\mathbb{B}})$	$\cup$	+2	+1	0	-3	-2
	$\overline{\text{WN}\overline{\text{M}}}$	$\mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}} + \mathbb{B}\mathbb{B}_{\mathbb{B}} + \mathbb{A}\mathbb{B}_{\mathbb{B}\mathbb{A}}$	$(\mathbb{A}\mathbb{B}_{\mathbb{B}\mathbb{A}}, \mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{B}})$		+1.5	+1	-0.5	-3	-2
	$\overline{\text{W}\overline{\text{W}}\overline{\text{M}}}$	$2 \times \mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}} + \mathbb{B}\mathbb{B}_{\mathbb{B}}$	$(\mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{A}}, \mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_\varepsilon \prec \Gamma_{\mathbb{B}})$	$\cup$	+2	+1	0	-3	-2
	$\overline{\text{MN}\overline{\text{W}}}$	$\mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}} + \mathbb{A}\mathbb{A}_{\mathbb{A}} + \mathbb{A}\mathbb{B}_{\mathbb{B}\mathbb{A}}$	$(\mathbb{A}\mathbb{B}_{\mathbb{B}\mathbb{A}}, \mathbb{A}\mathbb{A}_{\mathbb{A}}, \mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}})$		+1.5	+1	-0.5	-3	-2
	$\overline{\text{M}\overline{\text{M}}\overline{\text{W}}}$	$2 \times \mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}} + \mathbb{A}\mathbb{A}_{\mathbb{A}}$	$(\mathbb{B}\mathbb{B}_{\mathbb{B}\mathbb{A}}, \mathbb{A}\mathbb{A}_{\mathbb{A}}, \mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}}, \mathbb{A}\mathbb{A}_\varepsilon \prec \Gamma_{\mathbb{A}})$	$\cap$	+2	+1	0	-3	-2
	$\overline{\text{MZ}\overline{\text{W}}}$	$\mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}} + \mathbb{A}\mathbb{A}_{\mathbb{B}} + \mathbb{A}\mathbb{B}_{\mathbb{A}\mathbb{B}}$	$(\mathbb{A}\mathbb{B}_{\mathbb{A}\mathbb{B}}, \mathbb{A}\mathbb{A}_{\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{B}\mathbb{A}})$		+1.5	+1	-0.5	-3	-2
	$\overline{\text{M}\overline{\text{M}}\overline{\text{W}}}$	$2 \times \mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}} + \mathbb{A}\mathbb{A}_{\mathbb{B}}$	$(\mathbb{B}\mathbb{B}_{\mathbb{A}\mathbb{B}}, \mathbb{A}\mathbb{A}_{\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{B}\mathbb{A}}, \mathbb{A}\mathbb{A}_\varepsilon \prec \Gamma_{\mathbb{A}})$	$\cap$	+2	+1	0	-3	-2

	id	sources	linking cycle		$\Delta n$	$\Delta c$	$\Delta(2AB)$	$\Delta\lambda$	$\Delta\lambda_{DCJ}^\lambda$
$\mathcal{S}$	ZN	$AB_{AB} + AB_{BA}$	$(AB_{AB}, AB_{BA})$		+1	+1	-1	-2	-1
	$\overline{WM}$	$AA_A + BB_A$	$(AA_A, BB_A)$		+1	+1	0	-1	-1
	$\underline{WM}$	$AA_B + BB_B$	$(AA_B, BB_B)$		+1	+1	0	-1	-1
	$\overline{W\overline{M}}$	$AA_{AB} + BB_A$	$(AA_{BA}, BB_A)$		+1	+1	0	-1	-1
	$\underline{W\overline{M}}$	$AA_{AB} + BB_B$	$(AA_{AB}, BB_B)$		+1	+1	0	-1	-1
	WZ	$AA_{AB} + AB_{AB}$	$(AA_{BA}, BB_\varepsilon \prec \Gamma_B, AB_{AB})$	U	+1.5	+1	-0.5	-2	-1
	WN	$AA_{AB} + AB_{BA}$	$(AA_{AB}, BB_\varepsilon \prec \Gamma_B, AB_{BA})$	U	+1.5	+1	-0.5	-2	-1
	WW	$AA_{AB} + AA_{AB}$	$(AA_{AB}, BB_\varepsilon \prec \Gamma_B, AA_{BA}, BB_\varepsilon \prec \Gamma_B)$	U	+2	+1	0	-2	-1
	$\overline{M\overline{W}}$	$BB_{AB} + AA_A$	$(AA_A, BB_{AB})$		+1	+1	0	-1	-1
	$\underline{M\overline{W}}$	$BB_{AB} + AA_B$	$(AA_B, BB_{BA})$		+1	+1	0	-1	-1
	MZ	$BB_{AB} + AB_{AB}$	$(BB_{BA}, AB_{AB}, AA_\varepsilon \prec \Gamma_A)$	$\cap$	+1.5	+1	-0.5	-2	-1
	MN	$BB_{AB} + AB_{BA}$	$(BB_{AB}, AB_{BA}, AA_\varepsilon \prec \Gamma_A)$	$\cap$	+1.5	+1	-0.5	-2	-1
	MM	$BB_{AB} + BB_{AB}$	$(BB_{AB}, AA_\varepsilon \prec \Gamma_A, BB_{BA}, AA_\varepsilon \prec \Gamma_A)$	$\cap$	+2	+1	0	-2	-1
$\mathcal{M}$	$\underline{ZZ\overline{W\overline{M}}}$	$2 \times AB_{AB} + AA_B + BB_A$	$(AB_{AB}, AA_B, BA_{BA}, BB_A)$		+2	+1	-1	-4	-2
	$\underline{NN\overline{W\overline{M}}}$	$2 \times AB_{BA} + AA_A + BB_B$	$(AB_{BA}, AA_A, BA_{AB}, BB_B)$		+2	+1	-1	-4	-2
$\mathcal{N}$	$\overline{Z\overline{W\overline{M}}}$	$AB_{AB} + AA_B + BB_A$	$(AB_{AB}, AA_B, BB_A)$		+1.5	+1	-0.5	-2	-1
	$\underline{ZZ\overline{W}}$	$2 \times AB_{AB} + AA_B$	$(AB_{AB}, AA_B, BA_{BA}, BB_\varepsilon \prec \Gamma_B)$	U	+2	+1	-1	-3	-1
	$\underline{ZZ\overline{M}}$	$2 \times AB_{AB} + BB_A$	$(BA_{BA}, BB_A, AB_{AB}, AA_\varepsilon \prec \Gamma_A)$	$\cap$	+2	+1	-1	-3	-1
	$\overline{N\overline{W\overline{M}}}$	$AB_{BA} + AA_A + BB_B$	$(AB_{BA}, AA_A, BB_B)$		+1.5	+1	-0.5	-2	-1
	$\underline{NN\overline{W}}$	$2 \times AB_{BA} + AA_A$	$(AB_{BA}, AA_A, BA_{AB}, BB_\varepsilon \prec \Gamma_B)$	U	+2	+1	-1	-3	-1
	$\underline{NN\overline{M}}$	$2 \times AB_{BA} + BB_B$	$(BA_{AB}, BB_B, AB_{BA}, AA_\varepsilon \prec \Gamma_A)$	$\cap$	+2	+1	-1	-3	-1



	remaining paths	linking cycle		$\Delta n$	$\Delta c$	$\Delta(2\mathbb{A}\mathbb{B})$	$\Delta\lambda$	$\Delta_{DCJ}^\lambda$
1	$\mathbb{A}\mathbb{B}_*$	$(\mathbb{A}\mathbb{B}_*)$		+0.5	+1	-0.5	0	0
2	$\mathbb{A}\mathbb{A}_* + \mathbb{B}\mathbb{B}_*$	$(\mathbb{A}\mathbb{A}_*, \mathbb{B}\mathbb{B}_*)$		+1	+1	0	0	0
3	$\mathbb{A}\mathbb{A}_*$	$(\mathbb{A}\mathbb{A}_*, \Gamma_{\mathbb{B}})$	$\cup$	+1	+1	0	0	0
4	$\mathbb{B}\mathbb{B}_*$	$(\mathbb{B}\mathbb{B}_*, \Gamma_{\mathbb{A}})$	$\cap$	+1	+1	0	0	0

Any capping producing linking cycles following a top-down screening of the table above is optimal:

- ▶  $\Delta_{DCJ}^\lambda = \Delta n - \Delta c - \Delta(2\mathbb{A}\mathbb{B}) + \Delta\lambda$  gives the DCJ-indel-effect produced by each type of linking cycle.
- ▶ All given linking cycles have  $\Delta_{DCJ}^\lambda$  equivalent to the respective chain of deducting recombinations, therefore they achieve the optimal DCJ-indel distance.

**P1:** After identifying chains of recombinations  $\left\{ \begin{array}{l} \text{either there are no unbalanced chains} \\ \text{or there are only } \cup\text{-unbalanced chains (over-repr. in } \mathbb{A}) \\ \text{or there are only } \cap\text{-unbalanced chains (over-repr. in } \mathbb{B}) \end{array} \right.$

**P2:** When an unbalanced chain is being linked

$\left\{ \begin{array}{l} \text{if there is a remaining indel-free } \mathbb{A}\mathbb{A}_\varepsilon/\mathbb{B}\mathbb{B}_\varepsilon \text{ (of the under-repr. genome), it is used to link the chain} \\ \text{otherwise there is no remaining } \mathbb{A}\mathbb{A}_*/\mathbb{B}\mathbb{B}_* \text{ (of the under-repr. genome) and an adjacency } \Gamma_{\mathbb{A}/\mathbb{B}} \text{ links the chain} \end{array} \right.$

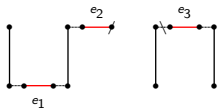
Any optimal capping that links all possible chains of deducting recombinations as described above and, for the remaining paths, maximizes the number of linking cycles of type 2 minimizes the number of caps:

$\left\{ \begin{array}{l} \text{The number of caps to be added is exactly } p_* = \max\{\kappa_{\mathbb{A}}, \kappa_{\mathbb{B}}\}. \\ \text{The number of adjacencies between caps is exactly } a_* = |\kappa_{\mathbb{A}} - \kappa_{\mathbb{B}}|. \end{array} \right.$

# Capped relational graph of singular genomes - example

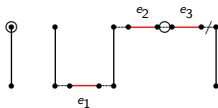
Deducting chain of path recombinations  $\left\{ \begin{array}{l} \text{transforming} \quad 2 \times AA_{AB} + BB_A + BB_B \\ \text{into} \quad 3 \times AB_\varepsilon + AB_B \\ \text{with} \quad \text{overall } \Delta_{DCJ}^\lambda = -3 \end{array} \right.$

$AA_{AB} + BB_A$   
2 runs + 1 run  
 $\lambda = 2 + \lambda = 1$



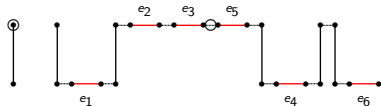
$(\Delta_{DCJ}^\lambda = -1)$   
gaining DCJ

$AB_\varepsilon + AB_{BA}$   
no run + 2 runs  
 $\lambda = 0 + \lambda = 2$

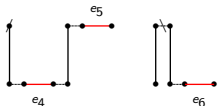


$\searrow$

$AB_\varepsilon + AB_B$   
no run + 3 runs  
 $\lambda = 0 + \lambda = 2$

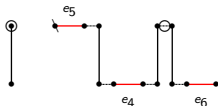


$AA_{AB} + BB_B$   
2 runs + 1 run  
 $\lambda = 2 + \lambda = 1$



$(\Delta_{DCJ}^\lambda = -1)$   
gaining DCJ

$AB_\varepsilon + AB_{AB}$   
no run + 2 runs  
 $\lambda = 0 + \lambda = 2$

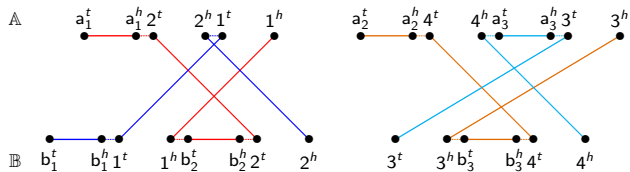


$(\Delta_{DCJ}^\lambda = -1)$   
neutral DCJ

$\nearrow$

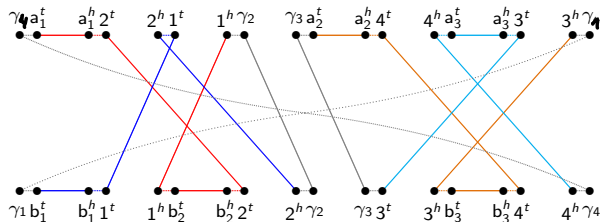
# Capped relational graph of singular genomes - example

$$\mathbb{A} = [a_1 \ 2 \ 1] \ [a_2 \ 4 \ a_3 \ 3] \quad \text{and} \quad \mathbb{B} = [b_1 \ 1 \ b_2 \ 2] \ [3 \ b_3 \ 4] \quad ; \quad p_* = 2 \quad \text{and} \quad a_* = 0$$



Components:  $2 \times \mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{A}}, \mathbb{B}\mathbb{B}_{\mathbb{B}}$

$$\begin{aligned} d_{\text{DCJ}}^{\text{ID}} &= n - |C| - \frac{|\mathcal{P}_{\mathbb{A}\mathbb{B}}|}{2} + \sum \lambda(C) - \delta \\ &= 4 - 0 - 0 + 6 - 3 \\ &= 7 \end{aligned}$$



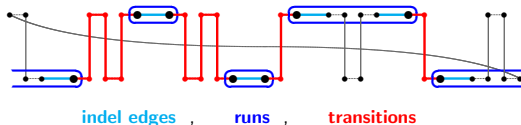
Linking cycle:  $(\mathbb{A}\mathbb{A}_{\mathbb{A}\mathbb{B}}, \mathbb{B}\mathbb{B}_{\mathbb{B}}, \mathbb{A}\mathbb{A}_{\mathbb{B}\mathbb{A}}, \mathbb{B}\mathbb{B}_{\mathbb{A}})$

$$\begin{aligned} d_{\text{DCJ}}^{\text{ID}} &= n + p_* - |C| + \sum \lambda(C) \\ &= 4 + 2 - 1 + 2 \\ &= 7 \end{aligned}$$

The four sources of a chain of deducing recombinations are optimally linked into a single cycle.

# Indel-potential via transitions

One indel-enclosing cycle:



$$\aleph = \Lambda = 4$$

$\Lambda(C)$  is the number of **runs** in cycle  $C$

$\aleph(C)$  is the number of **transitions** in cycle  $C$

**Indel-potential** of a component  $C$ :

$$\lambda(C) = \begin{cases} 0 & \text{if } \Lambda(C) = 0 \text{ (} C \text{ is indel-free)} \\ 1 & \text{if } \Lambda(C) = 1 \\ \frac{\Lambda(C)}{2} + 1 & \text{if } \Lambda(C) \geq 2 \end{cases}$$

$\Lambda$	$\aleph$	$r$	$\lambda$
0	0	0	0
1	0	1	1
2	2	1	2
4	4	1	3
6	6	1	4
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

cycles  
cycles and singletons  
cycles  
cycles  
cycles  
⋮  
⋮

$$\lambda(C) = \frac{\aleph(C)}{2} + r(C)$$

$$r(C) = \begin{cases} 1, & \text{component } C \text{ is indel-enclosing} \\ 0, & \text{component } C \text{ is indel-free} \end{cases}$$

# Quiz 2

1 Which of the following statements about the capped relational graph are true?

- A In an optimal capping, the distance computed based on the capped relational diagram must be equivalent to the distance computed based on the original relational diagram.
- B Let  $RG(\mathbb{A}, \mathbb{B})$  be a relational graph of **canonical** genomes.  
An optimal capping of  $RG(\mathbb{A}, \mathbb{B})$  that maximizes the number of cycles linking a pair  $\mathbb{A}\mathbb{A} + \mathbb{B}\mathbb{B}$  has a minimum number of caps ( $= \max\{\kappa_{\mathbb{A}}, \kappa_{\mathbb{B}}\}$ ).
- C Let  $\max\{\kappa_{\mathbb{A}_s}, \kappa_{\mathbb{B}_s}\} = \max\{\kappa_{\mathbb{A}_c}, \kappa_{\mathbb{B}_c}\}$ .  
An optimal capping of the relational graph of **singular** genomes  $\mathbb{A}_s$  and  $\mathbb{B}_s$  requires more caps than an optimal capping of the relational graph of **canonical** genomes  $\mathbb{A}_c$  and  $\mathbb{B}_c$ .
- D The indel-potential can be equivalently computed based on the number of runs or based on the number of transitions.

# References

Double Cut and Join with Insertions and Deletions

(Marília D.V. Braga, Eyla Willing and Jens Stoye)

JCB, Vol. 18, No. 9 (2011)

Computing the Rearrangement Distance of Natural Genomes

(Leonard Bohnenkämper, Marília D. V. Braga, Daniel Doerr and Jens Stoye)

LNCS, 12074, pp 3-18 (2020)