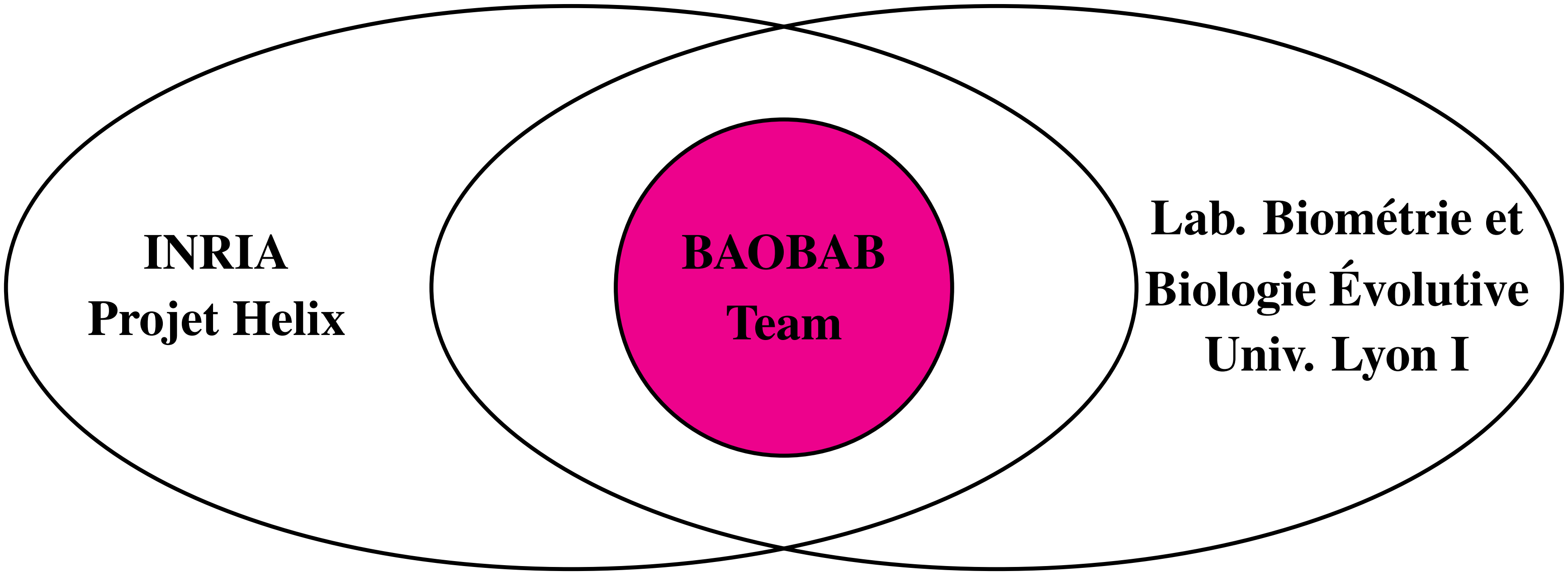
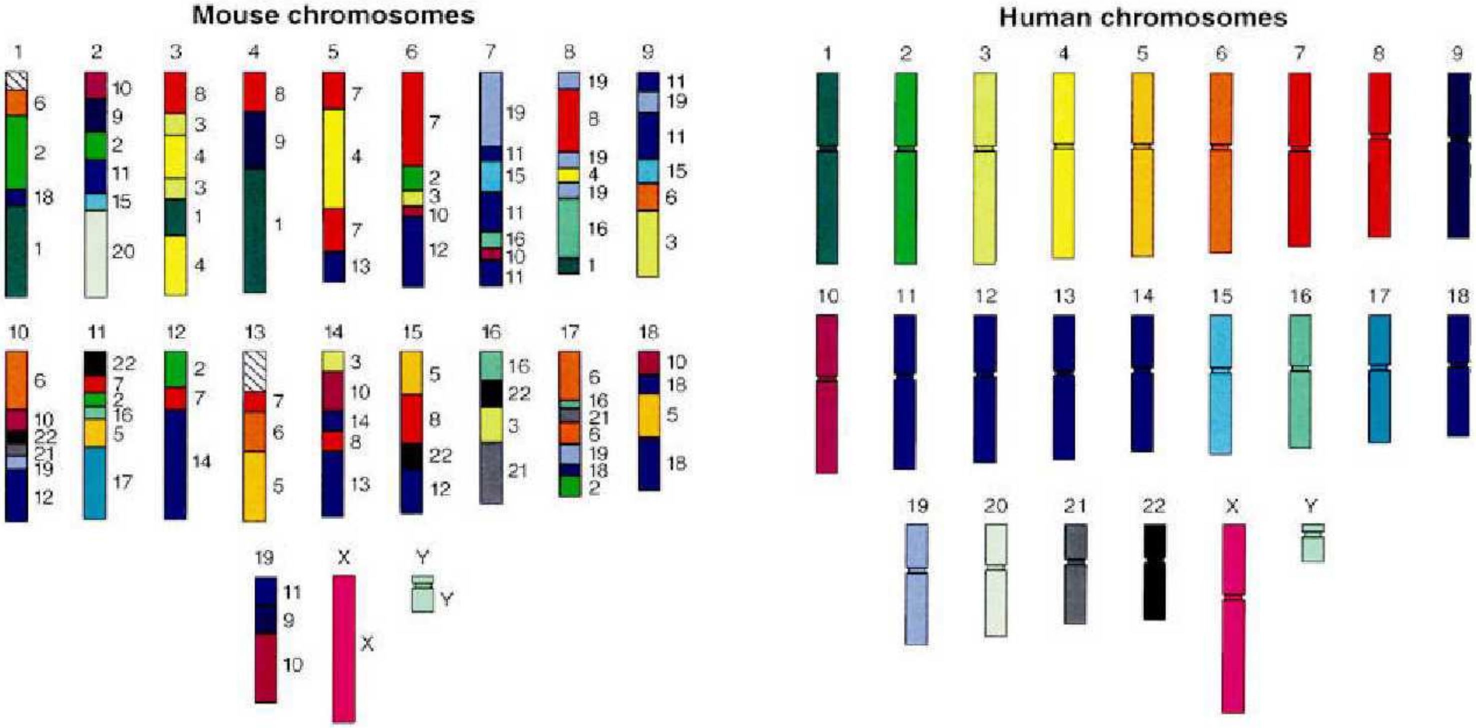
Some questions around genome rearrangements

Marie-France Sagot



Why genome rearrangements?

Mouse and Human Genetic Similarities



My personal motivation: biology

Combinatorial evolution

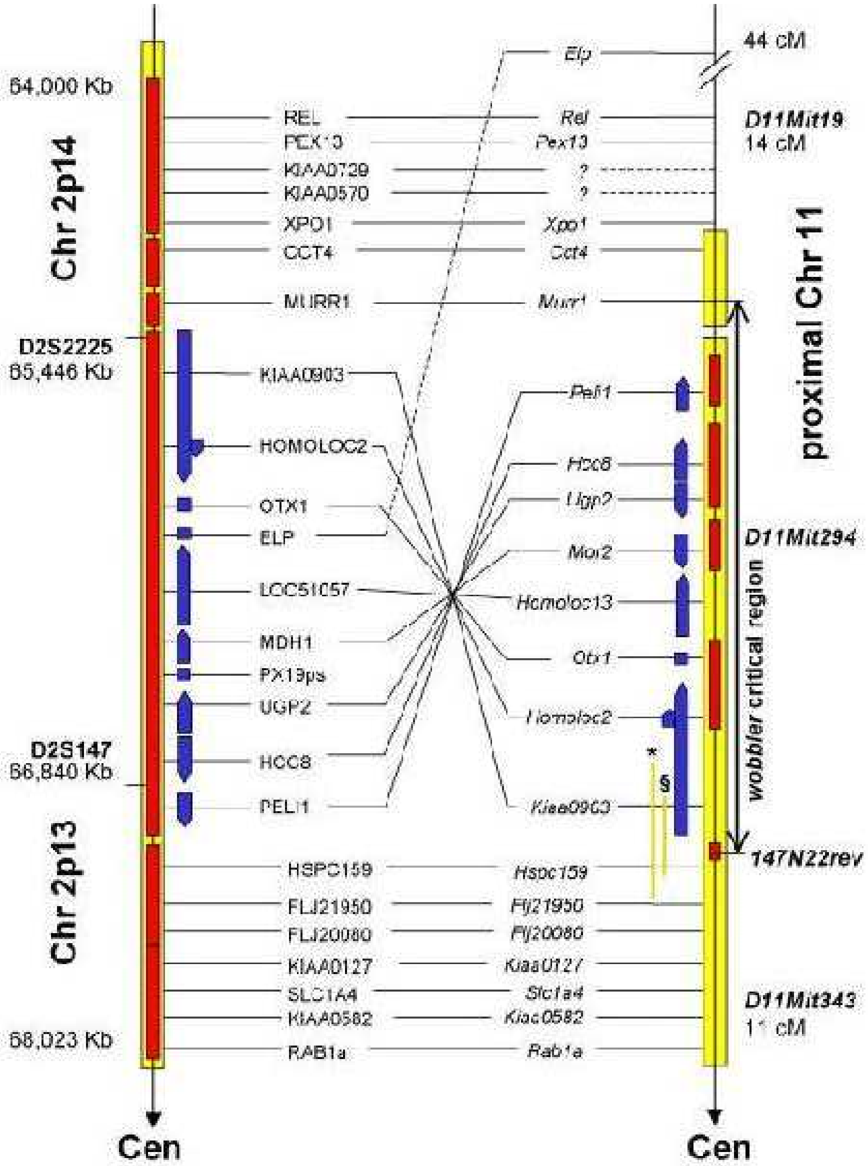
"By analogy with human language, biological evolution proceeds not only by making many words (molecules) from a single set of letters (genetic code). By taking a limited vocabulary of words, it can rearrange those into an even more enormous number of distinctive sentences, and then can continually and simultaneously reorganize paragraphs and pages and chapters and books (the phenotype at different levels of organization) without increasing the volume of the basic vocabulary at all."

(Mary-Jane West-Eberhard in "Developmental Plasticity and Evolution", Oxford University Press, 2003)

Three main questions around genomic rearrangements

- **Calculating rearrangement distance (between genomes and between trees representing evolution of genomes/sequences where rearrangements occurred)**
- **Identifying genomic segments "conserved" by rearrangements**
- **Understanding rearrangements: which ones took really place, why, what other factors may be at play etc.**

Inversions



64,000 Kb
Chr 2p14
D2S2225
65,446 Kb
D2S147
66,840 Kb
Chr 2p13
68,023 Kb
Cen
REL
PEX13
KIAA0729
KIAA0570
XPO1
CCT4
MURR1
KIAA0903
HOMOLOC2
OTX1
ELP
LOC51057
MDH1
PX19ps
UGP2
HCC8
PELI1
HSPC159
FLJ21950
FLJ20080
KIAA0127
SLC1A4
KIAA0582
RAB1a
Elp
Rel
Pex13
?
?
Xpo1
Cct4
Murr1
Peli1
Hcc8
Ugp2
Mor2
Homoloc13
Otx1
Homoloc2
Kiaa0903
Hspc159
Flj21950
Flj20080
Kiaa0127
Slc1a4
Kiaa0582
Rab1a
44 cM
D11Mit19
14 cM
proximal Chr 11
D11Mit294
wobbler critical region
147N22rev
D11Mit343
Cen

Sorting by inversions / reversals

0 7 5 3 1 6 2 4 8

minimum

number of reversals

0 1 2 3 4 5 6 7 8

Sorting by signed reversals

$$
\begin{array}{ccccccccc}
0 & 7 & 5 & 3 & -1 & -6 & -2 & 4 & 8
\end{array}
$$

minimum number | of signed reversals

$$
\downarrow
$$

$$
\begin{array}{ccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8
\end{array}
$$

Sorting by signed reversals

0	7	5	3	-1	-6	-2	4	8
0	1	-3	-5	-7	-6	-2	4	8

Sorting by signed reversals

0	7	5	3	-1	-6	-2	4	8
0	1	-3	-5	-7	-6	-2	4	8
0	1	-3	-5	-4	2	6	7	8

Sorting by signed reversals

0	7	5	3	-1	-6	-2	4	8
0	1	-3	-5	-7	-6	-2	4	8
0	1	-3	-5	-4	2	6	7	8
0	1	-3	-2	4	5	6	7	8

Sorting by signed reversals

0	7	5	3	-1	-6	-2	4	8
0	1	-3	-5	-7	-6	-2	4	8
0	1	-3	-5	-4	2	6	7	8
0	1	-3	-2	4	5	6	7	8
0	1	2	3	4	5	6	7	8

State of the art

1995 **Hannenhali and Pevzner – first polynomial algorithm,** $O(n^4)$

1996 **Berman and Hannenhali –** $O(n^2\alpha(n))$**,** α**: inverse Ackerman**

1997 **Kaplan, Shamir, Tarjan –** $O(n^2)$

1997 **Caprara – proved non signed version is NP**

2001 **Bergeron – "simpler" formulation but** $O(n^3)$

2001 **Bader, Moret and Yan –** $O(n)$ **but for calculating distance only**

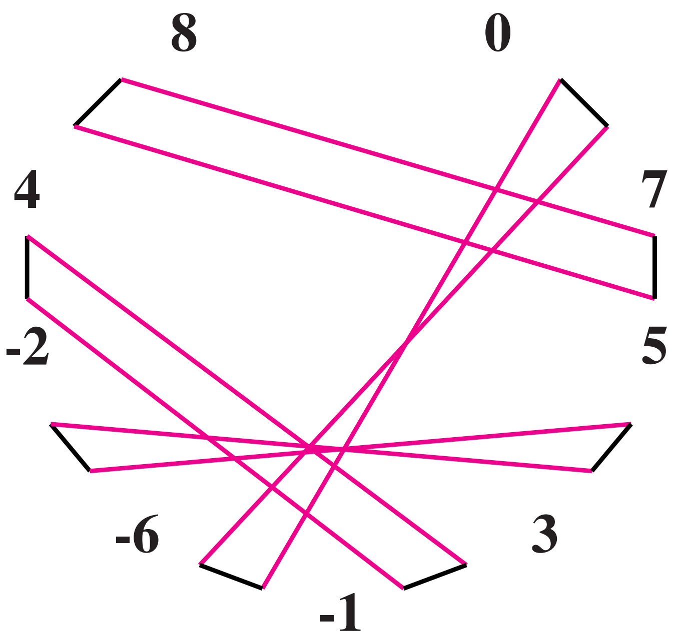
2003 **Kaplan and Verbin –** $O(n\sqrt{n \log n})$ **but heuristic**

2003 **Ozery-Flato and Shamir**
"A central question in the study of genome rearrangements is whether one can obtain a sub-quadratic algorithm for SBR"

Hannenhali's and Pevzner's theory

Dream and reality

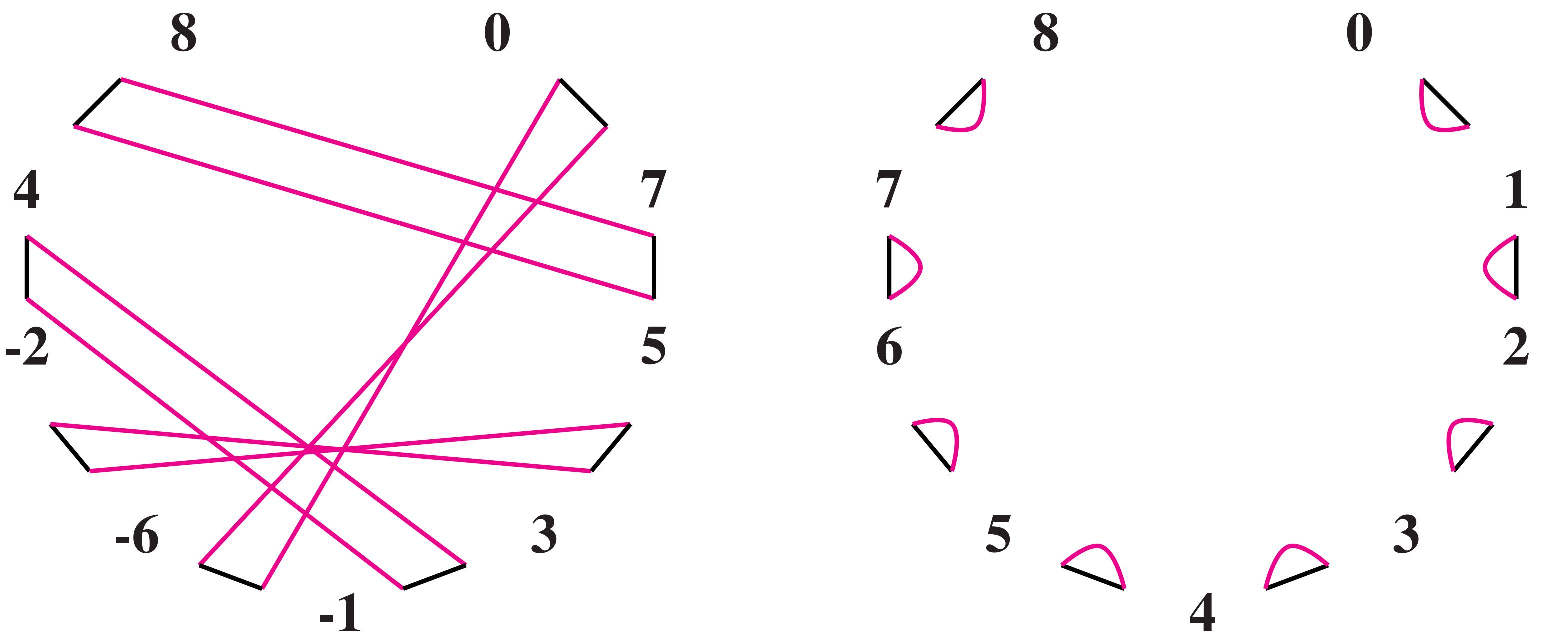
0 7 5 3 -1 -6 -2 4 8



Hannenhali's and Pevzner's theory

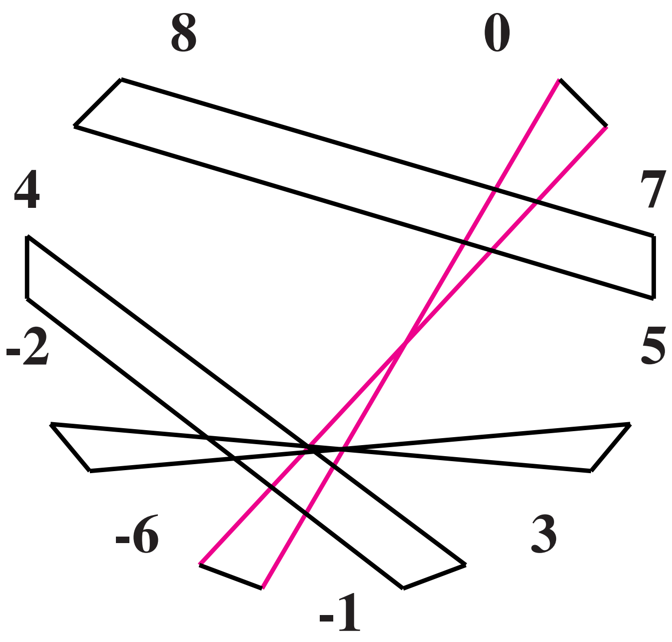
Dream and reality

0	7	5	3	-1	-6	-2	4	8
0	1	2	3	4	5	6	7	8



Oriented pairs

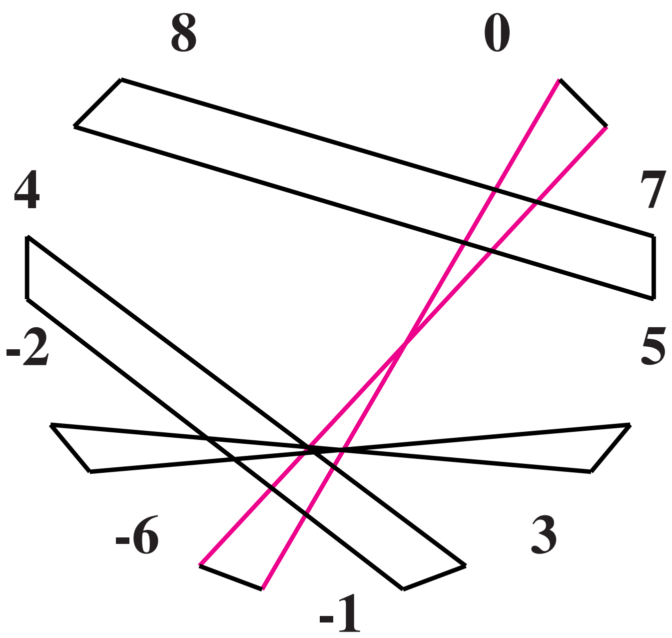
0 7 5 3 -1 -6 -2 4 8



Oriented pairs

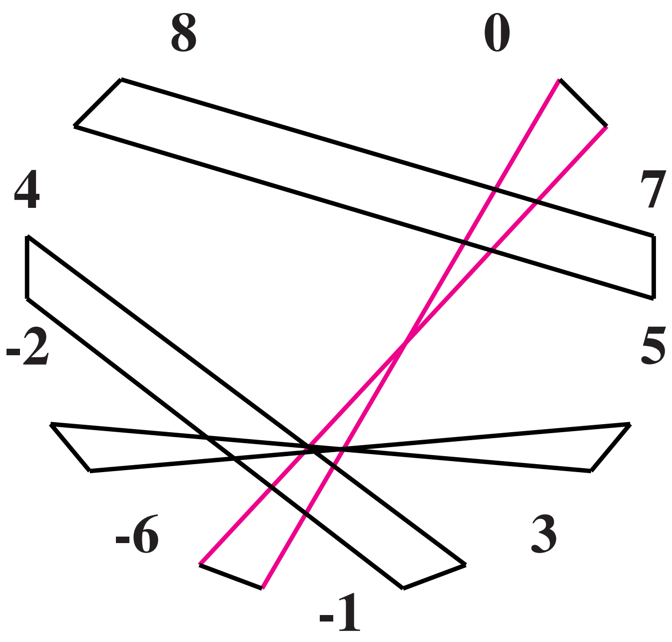
0 7 5 3 -1 -6 -2 4 8

0 1 -3 -5 -7 -6 -2 4 8



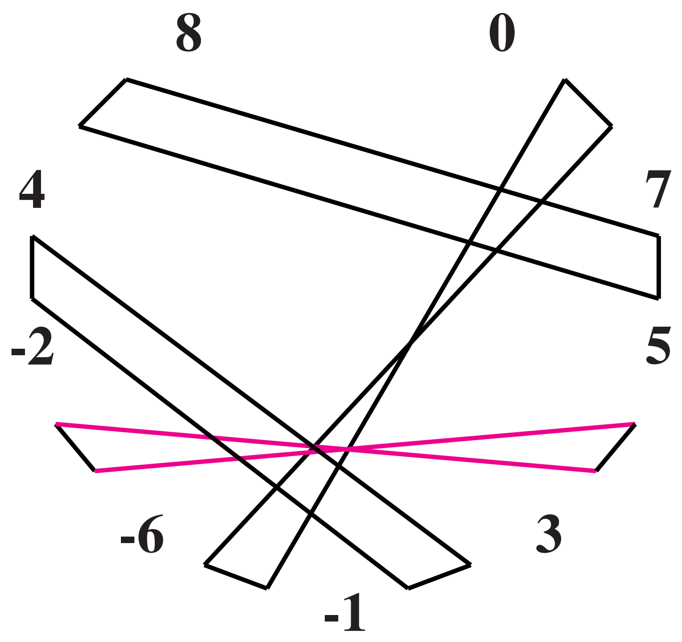
Oriented pairs

0 7 5 3 -1 -6 -2 4 8



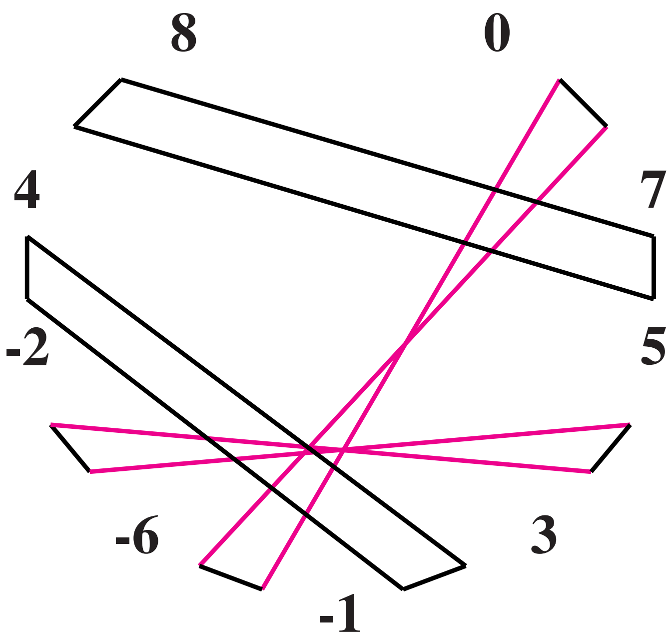
Another oriented pair

0 7 5 3 -1 -6 -2 4 8



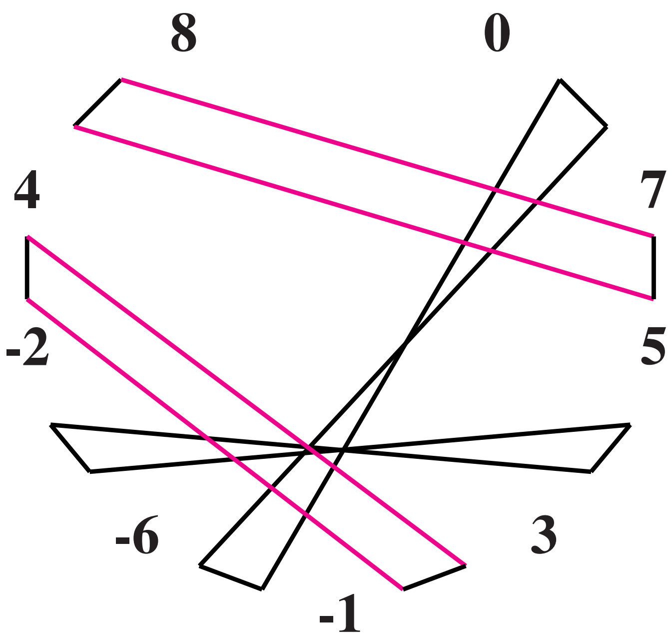
“Good” cycles

0 7 5 3 -1 -6 -2 4 8



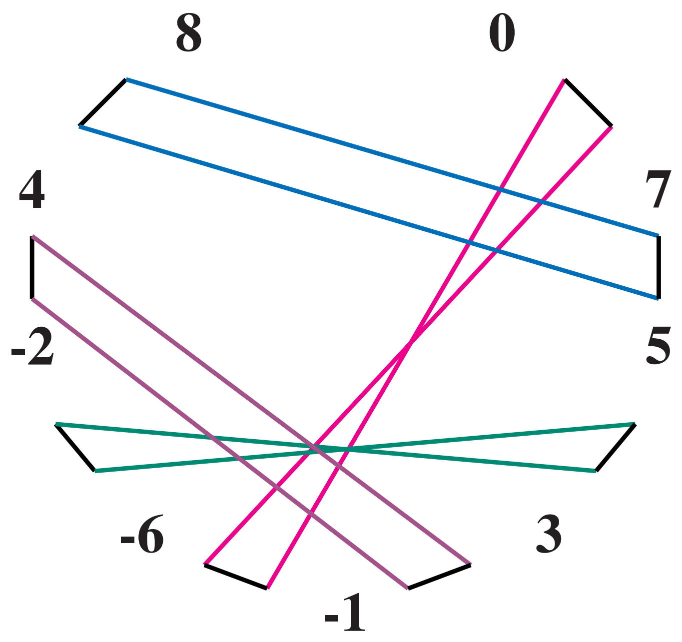
"Bad" cycles

0 7 5 3 -1 -6 -2 4 8



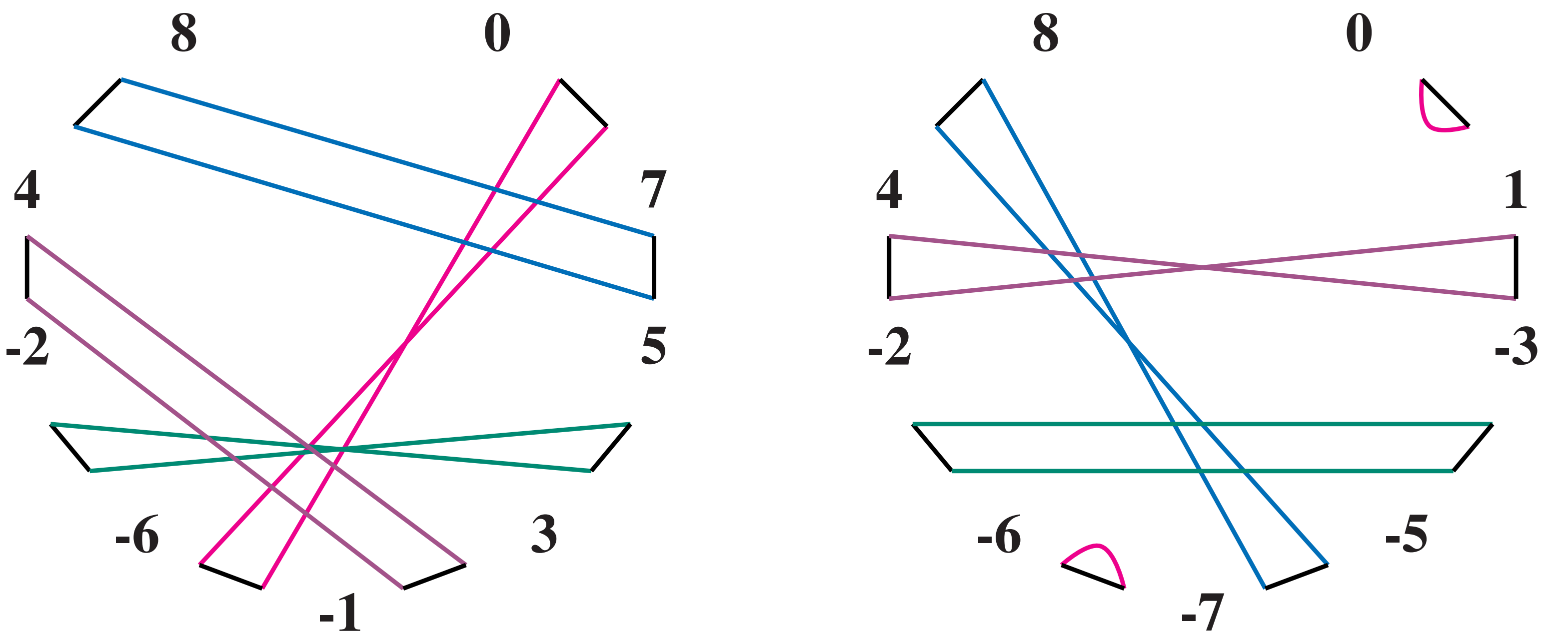
Breakpoint graph

0 7 5 3 -1 -6 -2 4 8



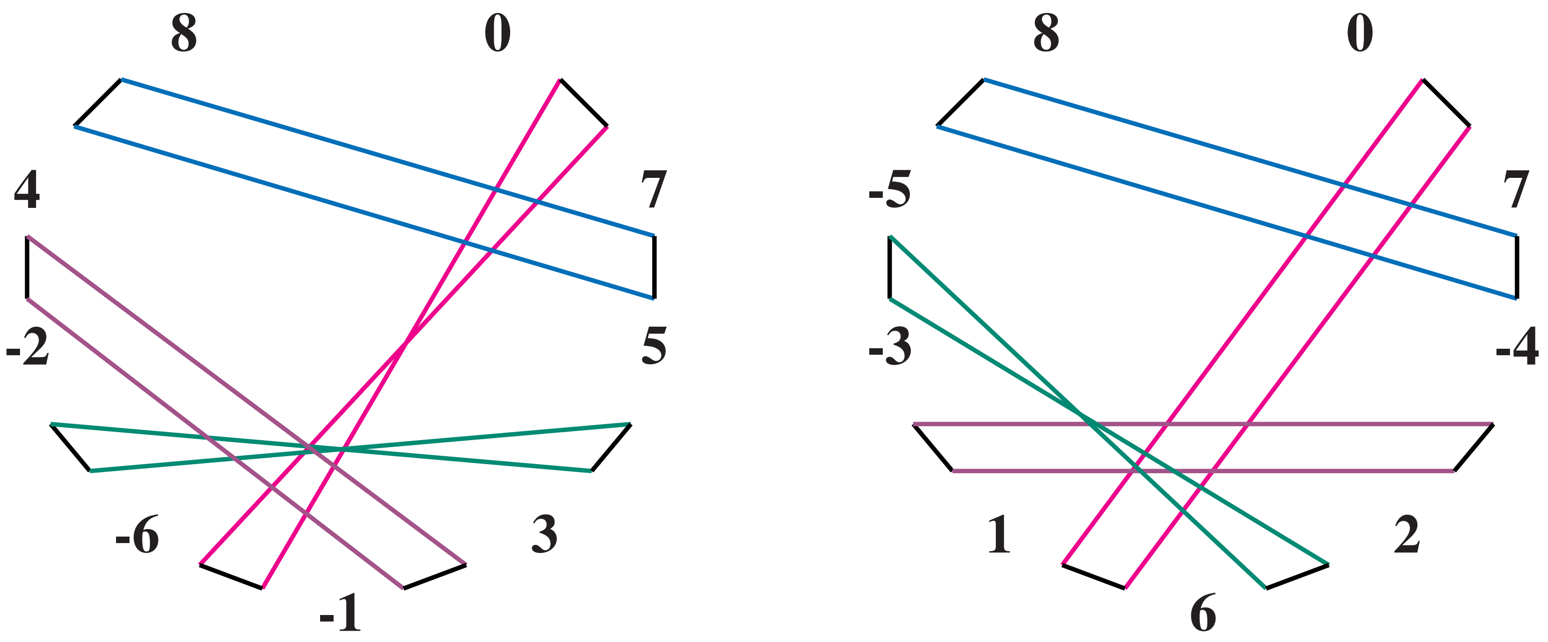
Consequences of an inversion: good cycle

0	7	5	3	-1	-6	-2	4	8
0	1	-3	-5	-7	-6	-2	4	8



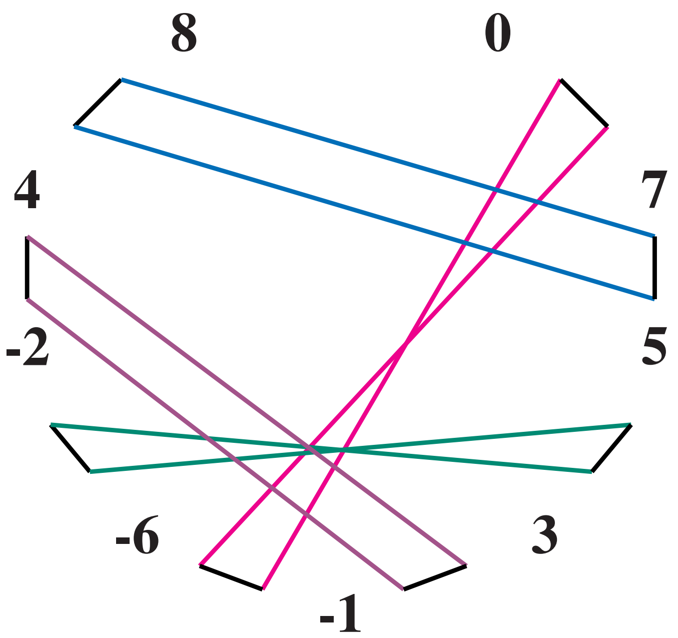
Consequences of an inversion: bad cycle

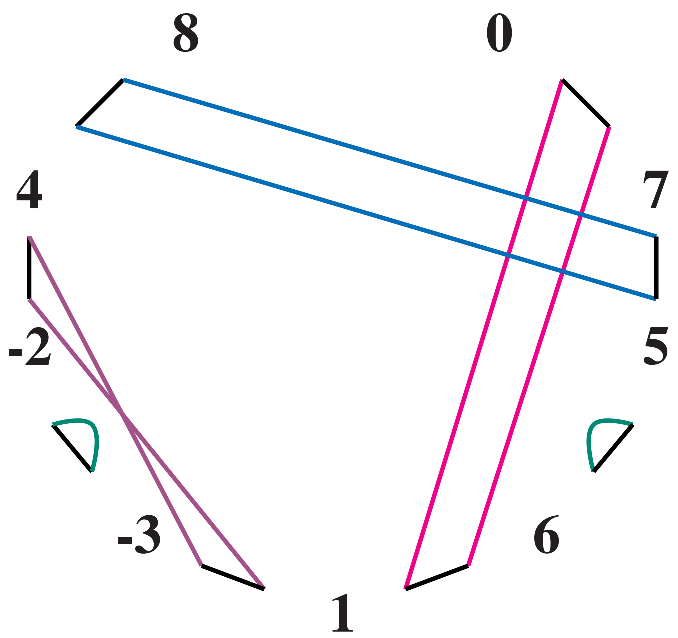
0	7	5	3	-1	-6	-2	4	8
0	7	-4	2	6	1	-3	-5	8



Bad choice of a good cycle

0 7 5 $\underline{3 \quad -1 \quad -6}$ -2 4 8

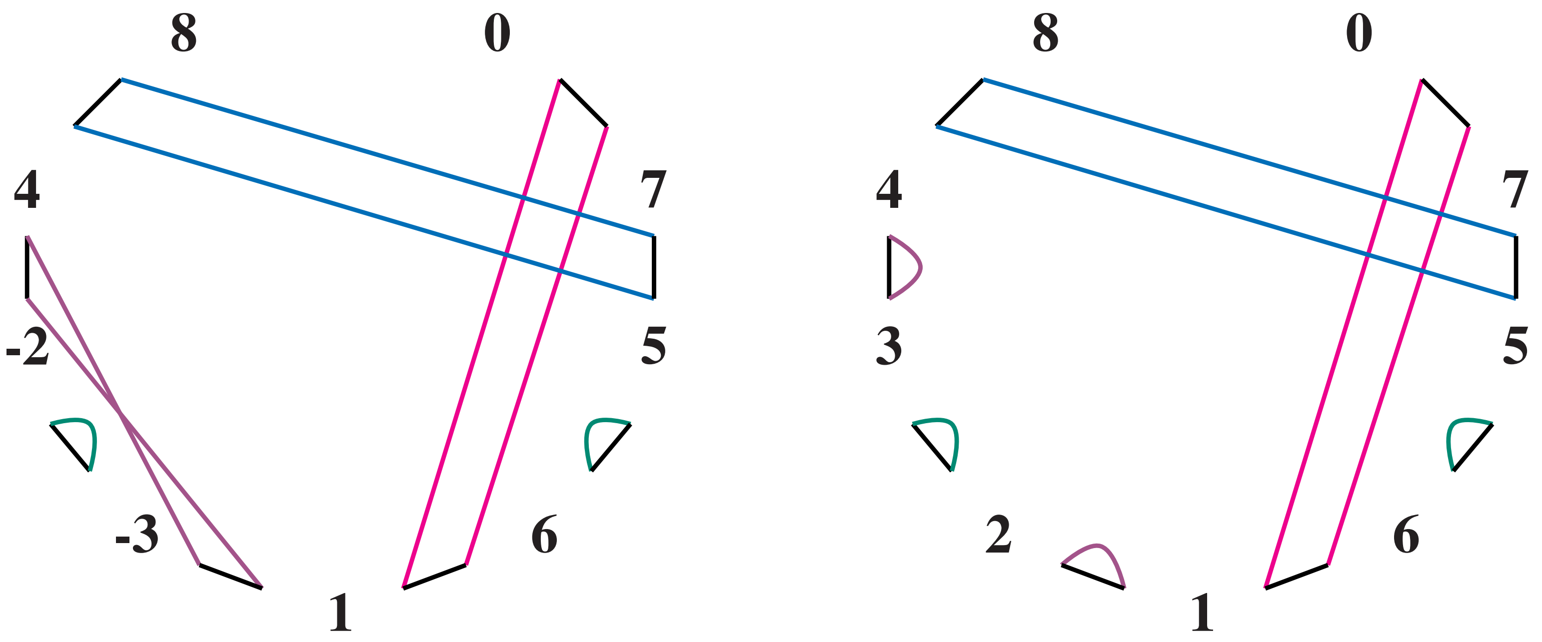




Looks good at first... but leads to a bad component

$$0 \quad 7 \quad 5 \quad 6 \quad 1 \quad \underbrace{-3 \quad -2} \quad 4 \quad 8$$

$$0 \quad 7 \quad 5 \quad 6 \quad 1 \quad 2 \quad 3 \quad 4 \quad 8$$



Components in the breakpoint graph

Component

set of cycles which do not cross the other cycles in the graph

Bad component

component which contains only bad cycles

Good component

component which contains at least one good cycle

Hannenhali and Pevzner's theorem

size of permutation π

number inversions needed to treat bad components

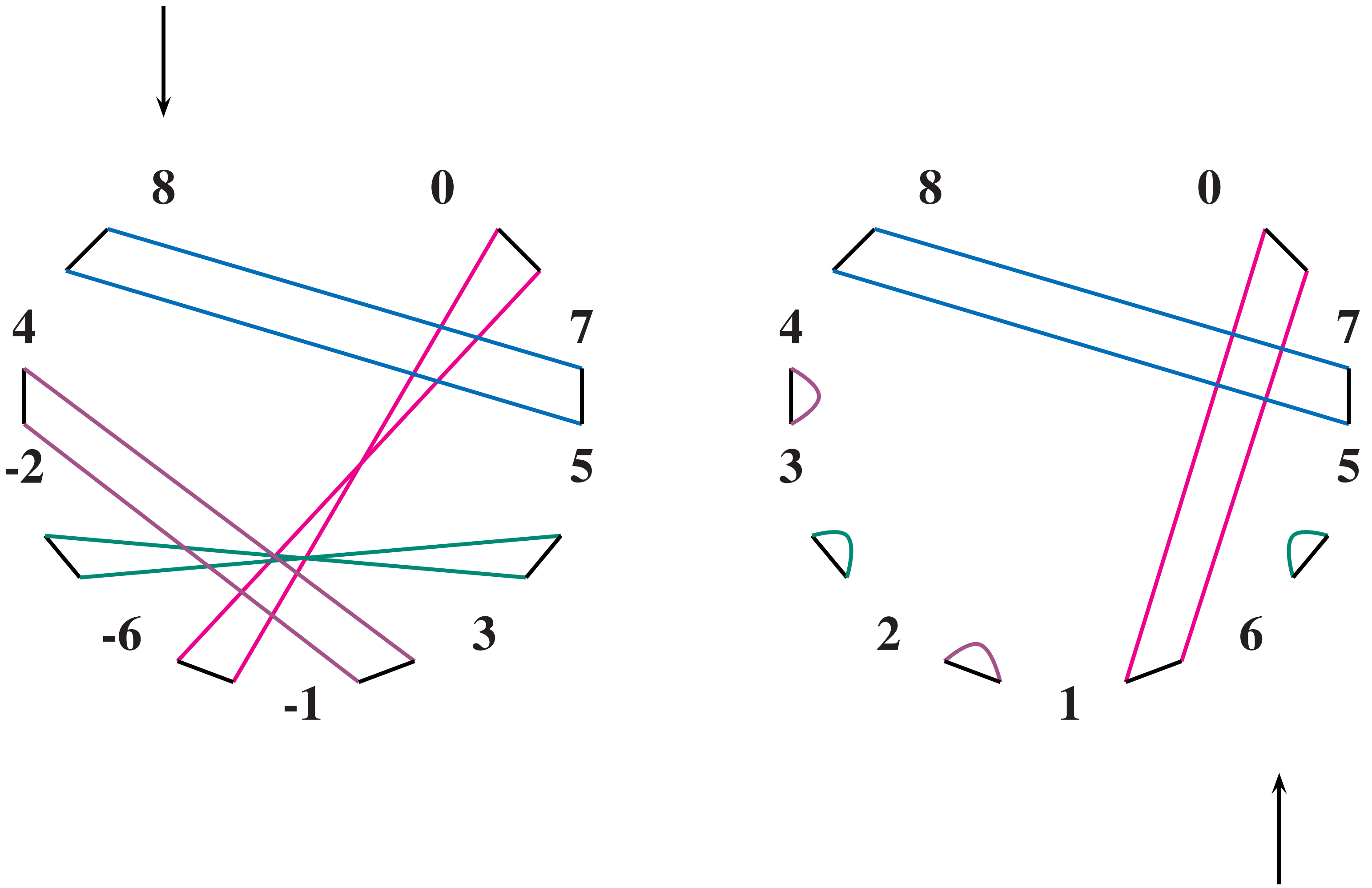
$$d(\pi) \quad = \quad n-1 \quad - \quad c(\pi) \quad + \quad t(\pi)$$

min. number inversions to sort π

number cycles in breakpoint graph

The good, the bad and the ugly

0 7 5 3 -1 -6 -2 4 8



0 7 5 6 1 2 3 4 8

Observation

If the initial permutation has more than one component, each can be treated separately

Treat bad components

size of permutation π

$$\underbrace{d(\pi)}_{\text{min. number inversions to sort } \pi} = \overbrace{n-1}^{\text{size of permutation } \pi} + \underbrace{c(\pi)}_{\text{number cycles in breakpoint graph}}$$

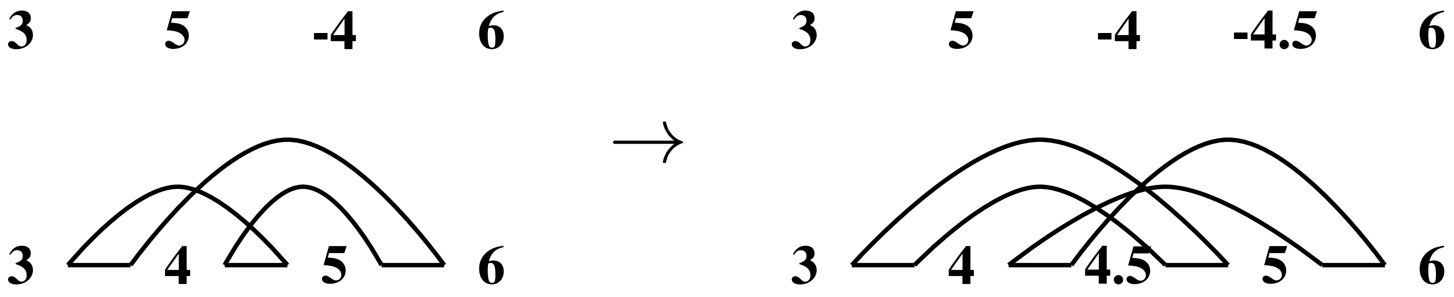
min. number inversions to sort π

number cycles in breakpoint graph

Can be done in linear time (Hannenhali and Pevzner)

Transform permutation into “simple permutation”

The breakpoint graph of simple permutations contains only 1-cycles (adjacencies) and 2-cycles



Can be done in linear time

(Tannier and S. adapting Badet, Moret and Yan’s idea)

Hannenhali and Pevzner's theorem for simple permutations

size of permutation π

$$d(\pi) \quad = \quad n-1 \quad + \quad c(\pi) \quad = \quad c(\pi)$$

min. number inversions to sort π

number cycles in breakpoint graph

What remains to be done?

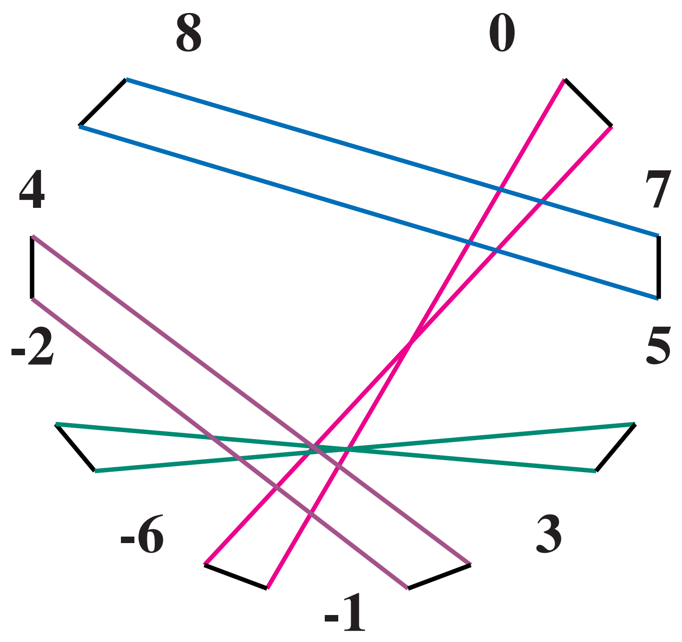
Choose well the good cycles

How to choose cycles: different approaches

Hannenhali: test properties of cycles to choose good one – $O(n^2)$

Bergeron: test all cycles – $O(n^3)$

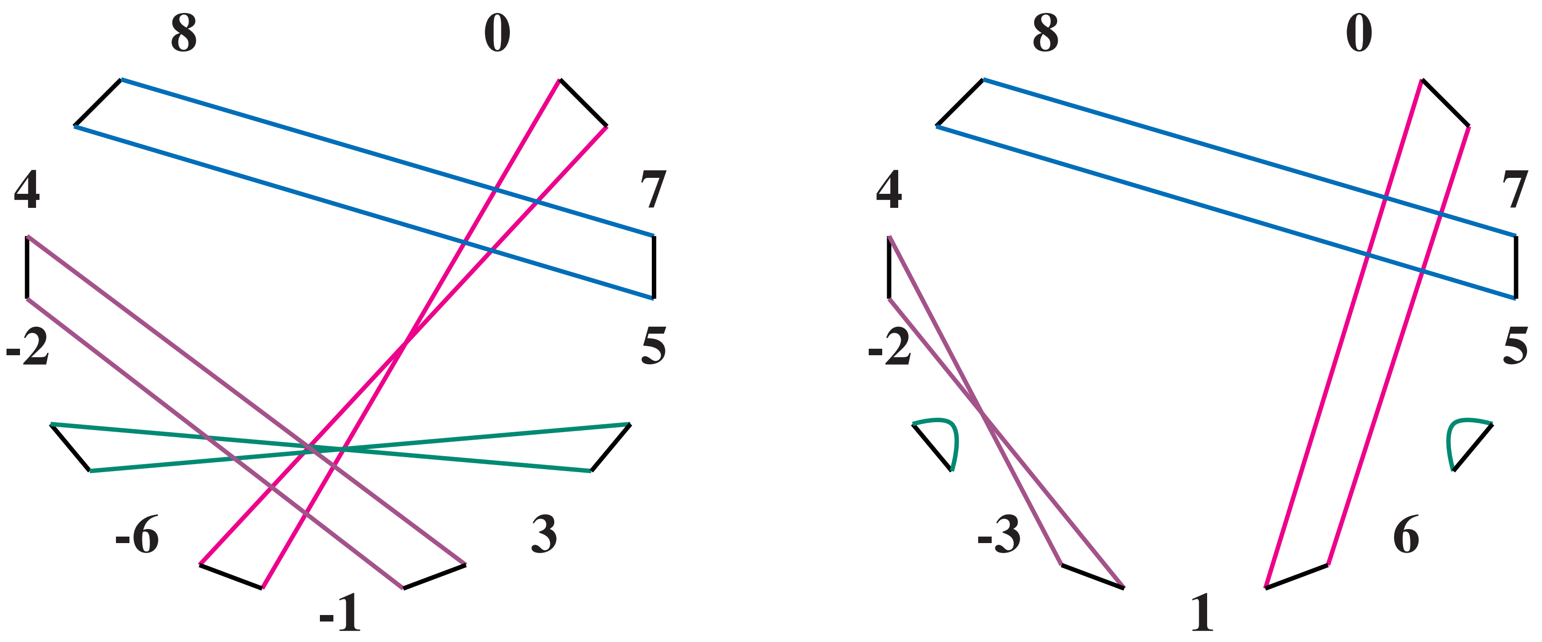
Kaplan: pick any cycle as bad choices are rare – heuristic



Tannier & S.: bad choices will become good later

Tannier and S.'s Algorithm

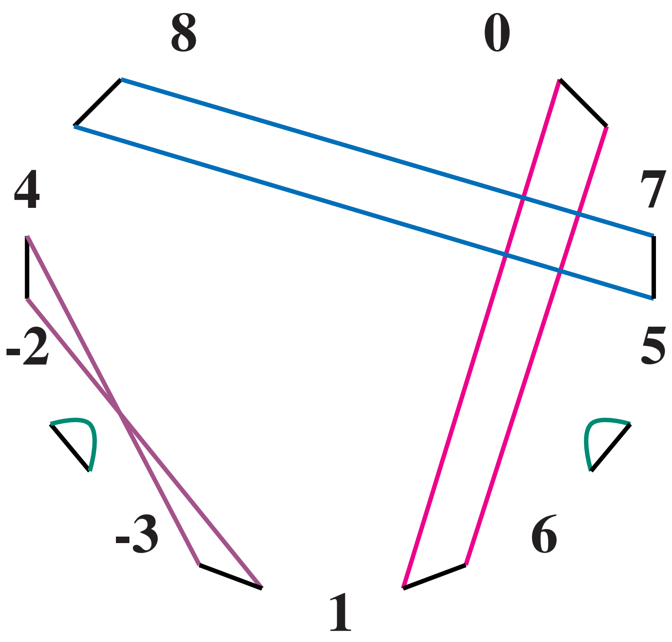
pick green cycle

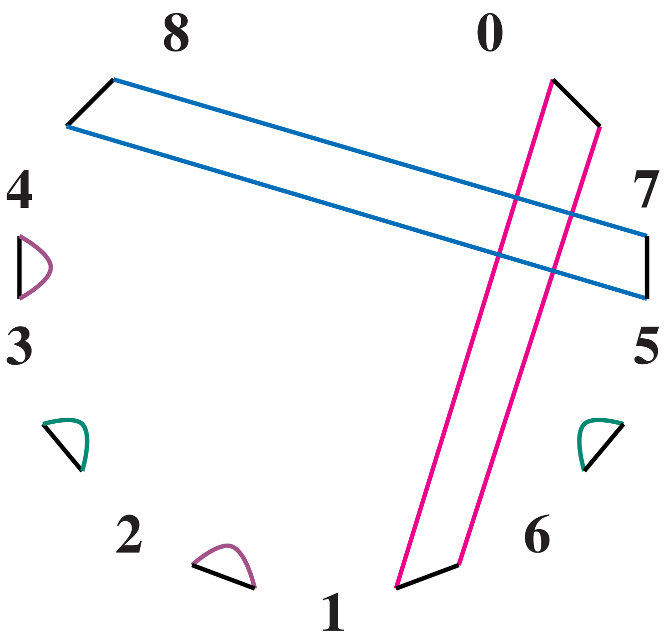


Algorithm (ctd.)

pick green cycle

then pick violet cycle

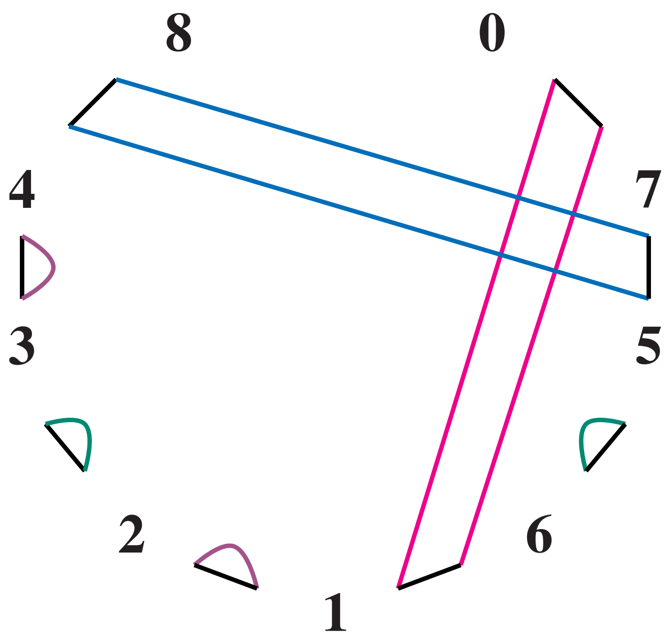




Algorithm (ctd.)

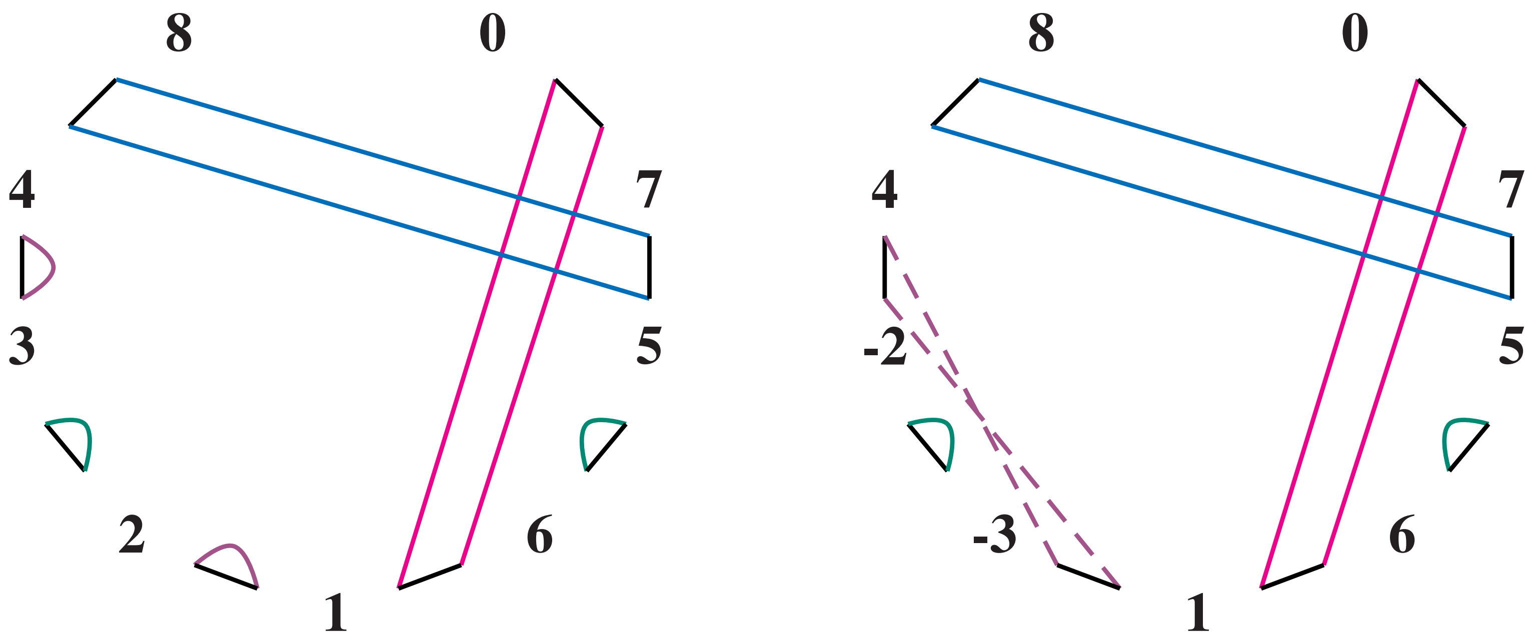
partial solution green violet

but then we are stuck with a bad component



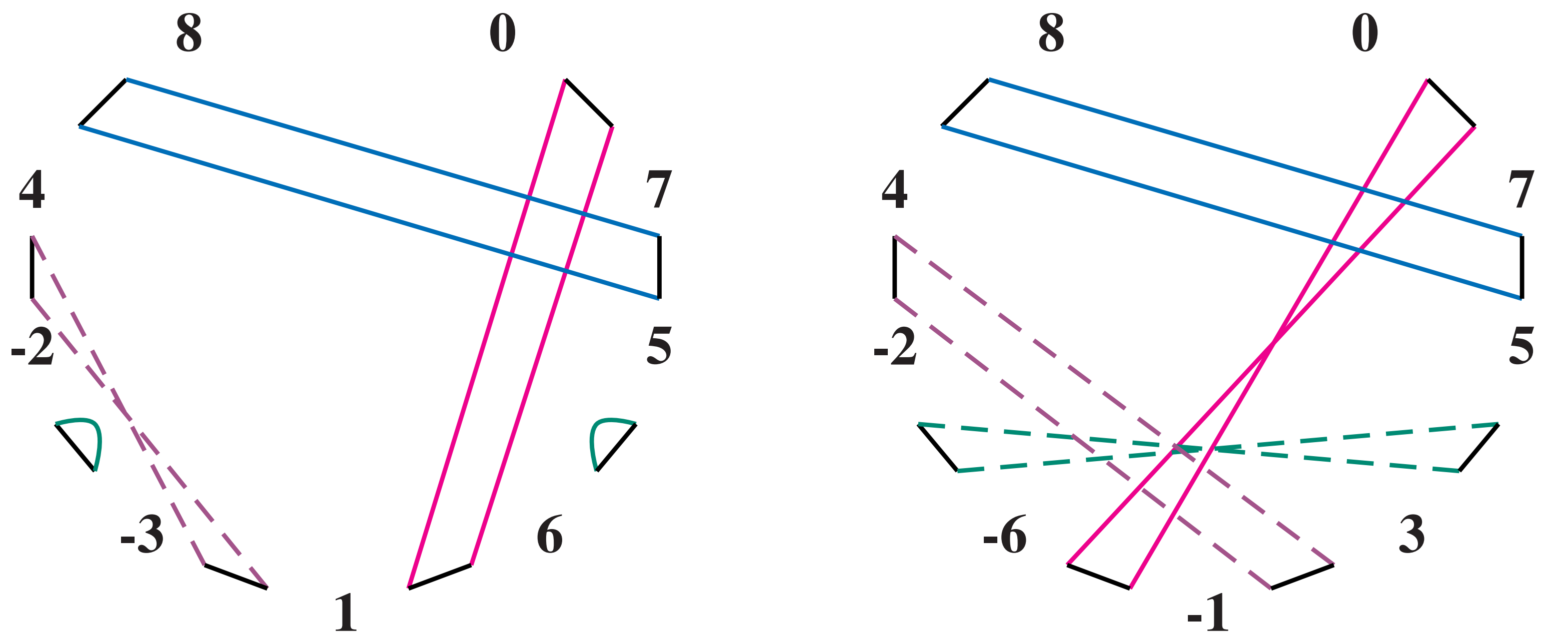
Algorithm (ctd.)

return “putting back” cycles



Algorithm (ctd.)

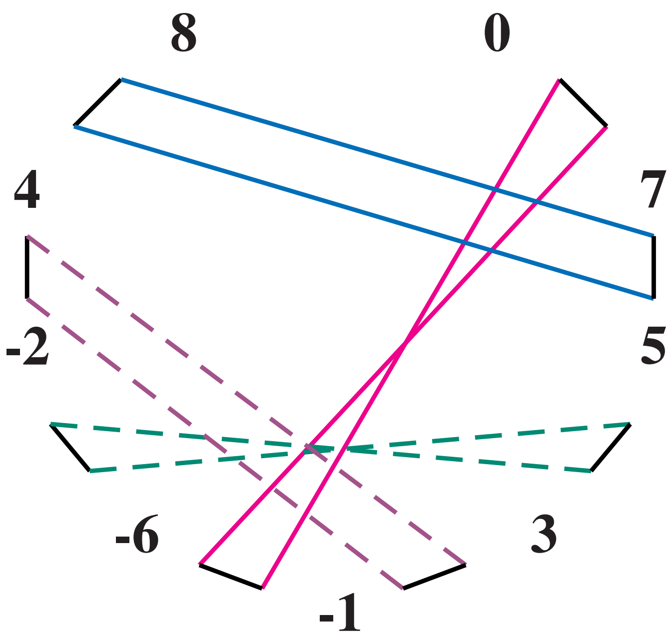
return "putting back" cycles
until the first that intersects
the component(s) left

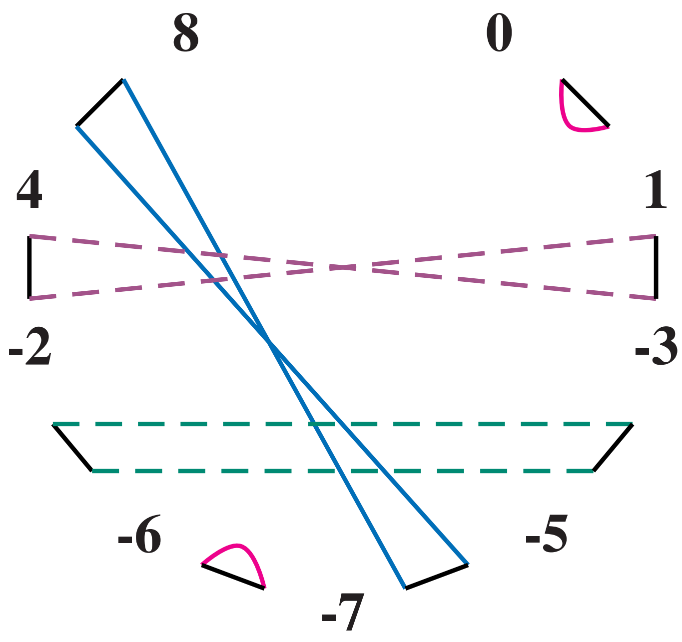


at least one cycle
in component is good

Algorithm

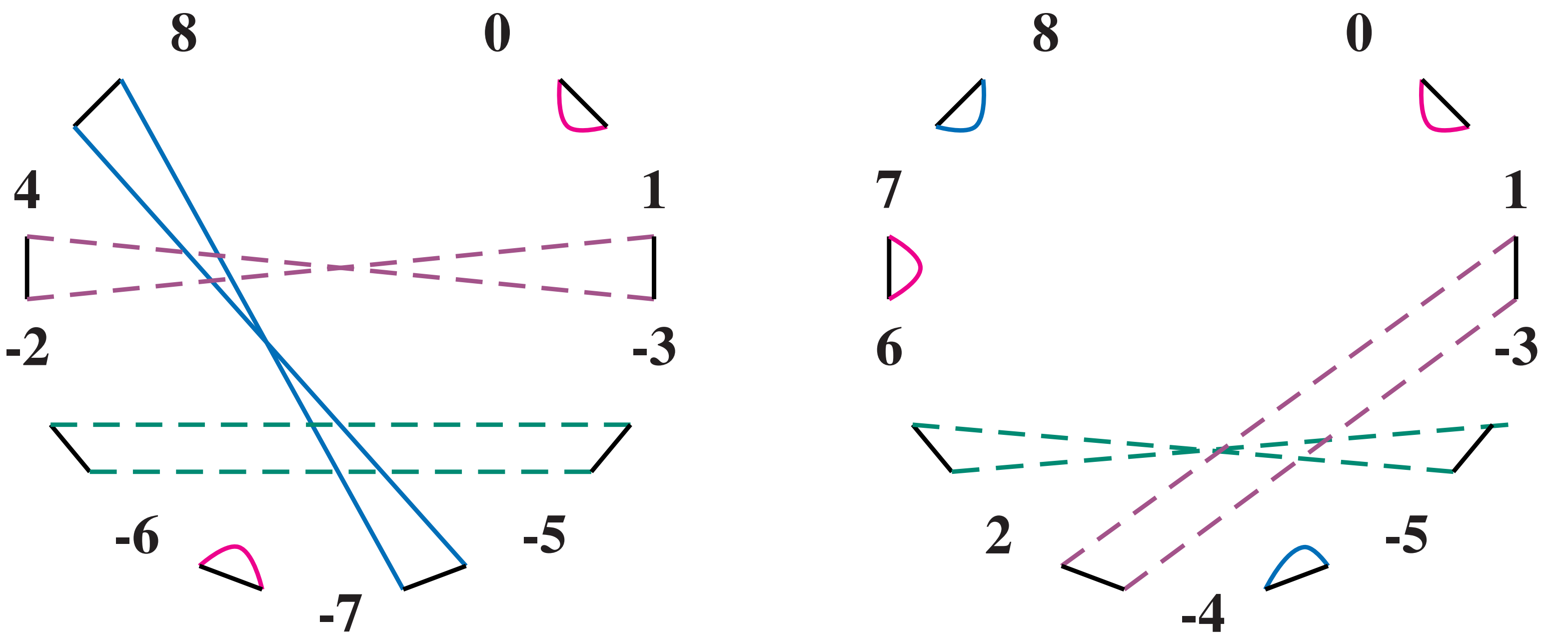
pick a good cycle
not among those already chosen





Algorithm (ctd.)

continue picking good cycles
not among those already chosen
until possibly stuck again or no untouched cycle remains



put back the part of the solution
already identified

Solution: pink blue green violet

Tannier and S.'s time complexity

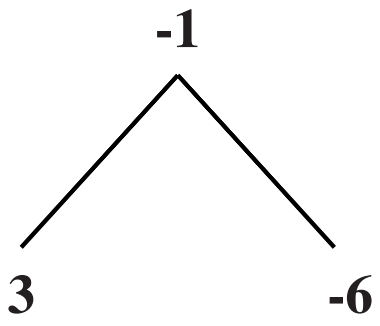
- **with a classical data structure (array):**
  $O(n)$ **per inversion,** $O(n^2)$ **for the whole sorting**

- **with an adaptation of Kaplan and Verbin's structure:**

$$0 \quad 7 \quad 5 \quad \underbrace{\Big|\ 3 \quad -1 \quad -6\ \Big|}_{} \quad -2 \quad 4 \quad 8$$

$$\tfrac{1}{2}\sqrt{n\log n} \leq l \leq \tfrac{3}{2}\sqrt{n\log n}$$

splay trees
or any balanced trees



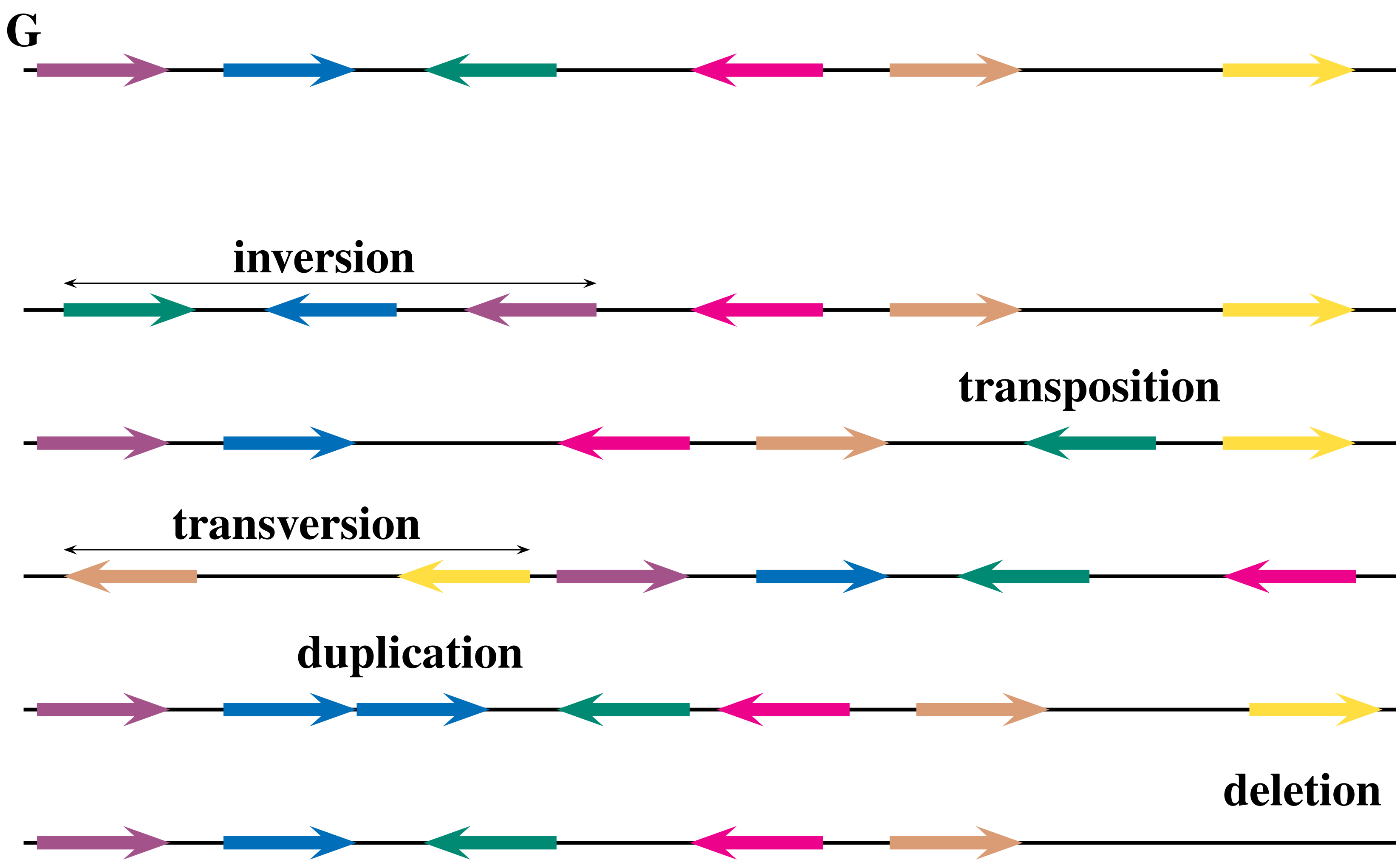
$O(n\sqrt{n\log n})$

Open problem

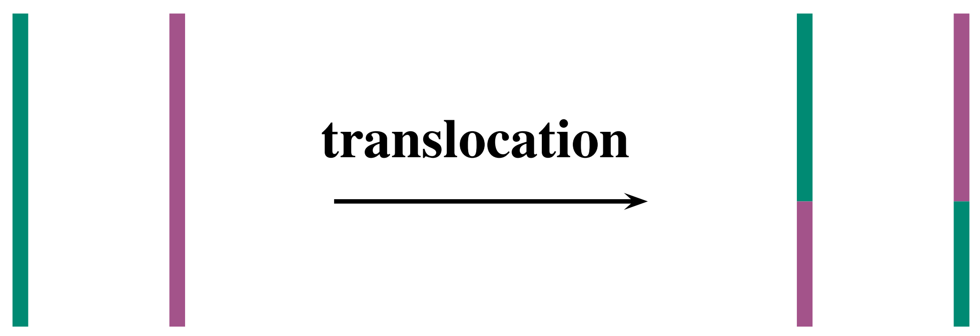
Can a better (linear) complexity be obtained?

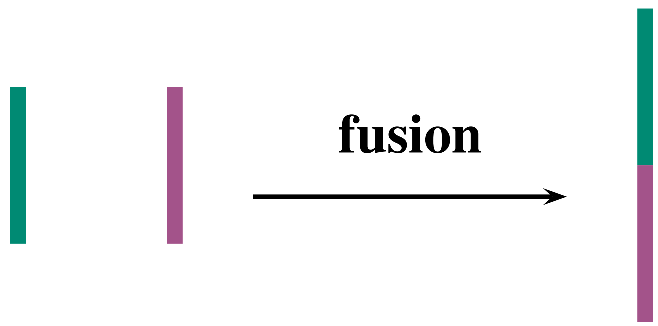
Almost there with $O(n\sqrt{n})$

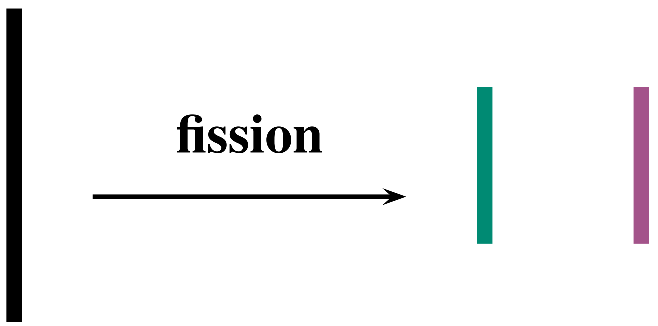
Other open problems: many rearrangements possible



And, in the case of multichromosomal genomes







Three main questions around genomic rearrangements

- **Calculating rearrangement distance (between genomes and between trees representing evolution of genomes/sequences where rearrangements occurred)**
- **Identifying genomic segments "conserved" by rearrangements**
- **Understanding rearrangements: which ones took really place, why, what other factors may be at play etc.**

The BAOBAB Team

At Lyon

Said S. Adi	Fred Boyer	Christian Gautier
Laurent Guéguen	Vincent Lacroix	Claire Lemaitre
Giulio Marcon	Jeane C.B. de Melo	Christelle M. de Lima
Vincent Navratil	Leonor Palmeira	Eric Tannier

Students outside Lyon

Julien Allali	Alexandra Carvalho	Paulo G.S. Fonseca
Benoît Olivieri	Pierre Peterlongo	Susana Vinga

Thank you for your attention