Genome Matrices and The Median Problem

Joao Meidanis[1]

University of Campinas, Brazil

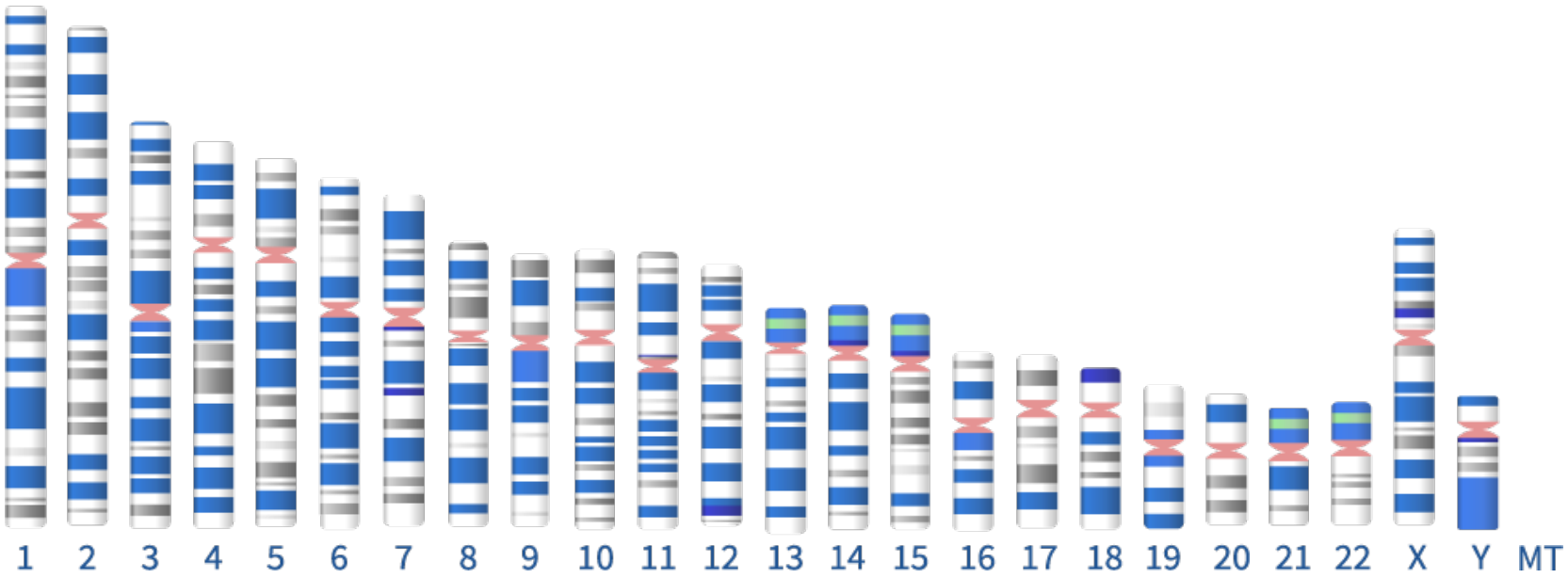
October 2018

[1] Joint work with P.Biller, J.Zanetti, L.Chindelevitch

Summary

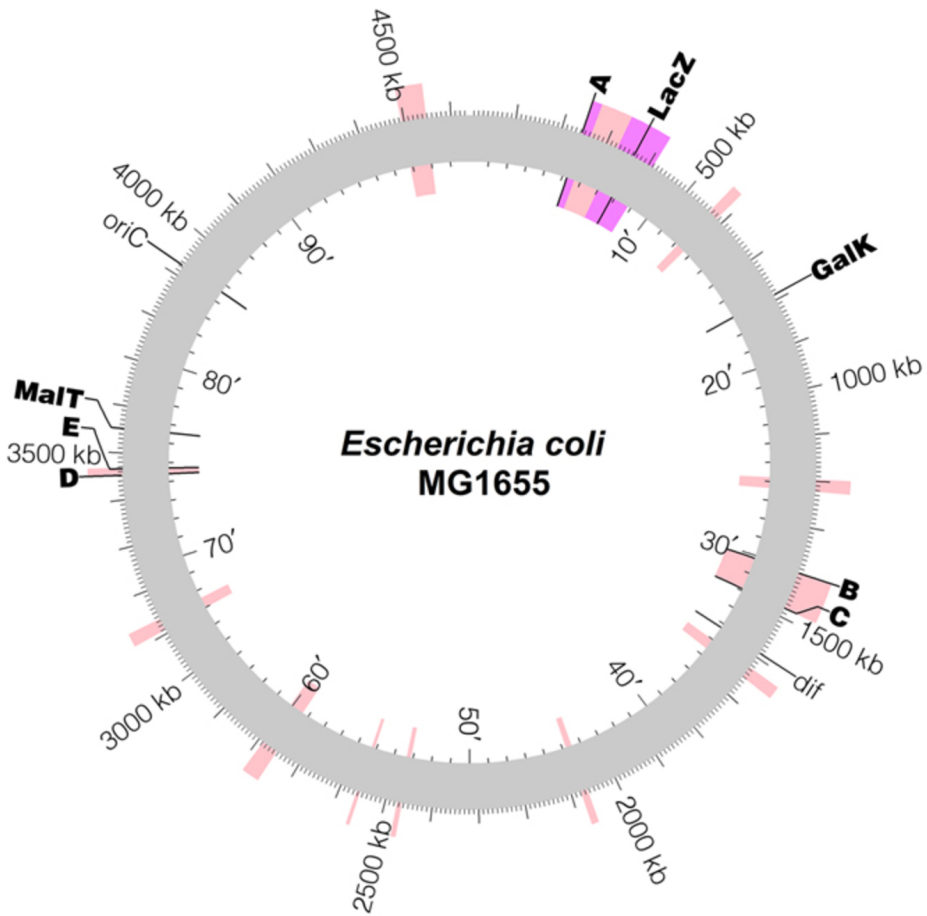
Genome Rearrangements

The Human Genome



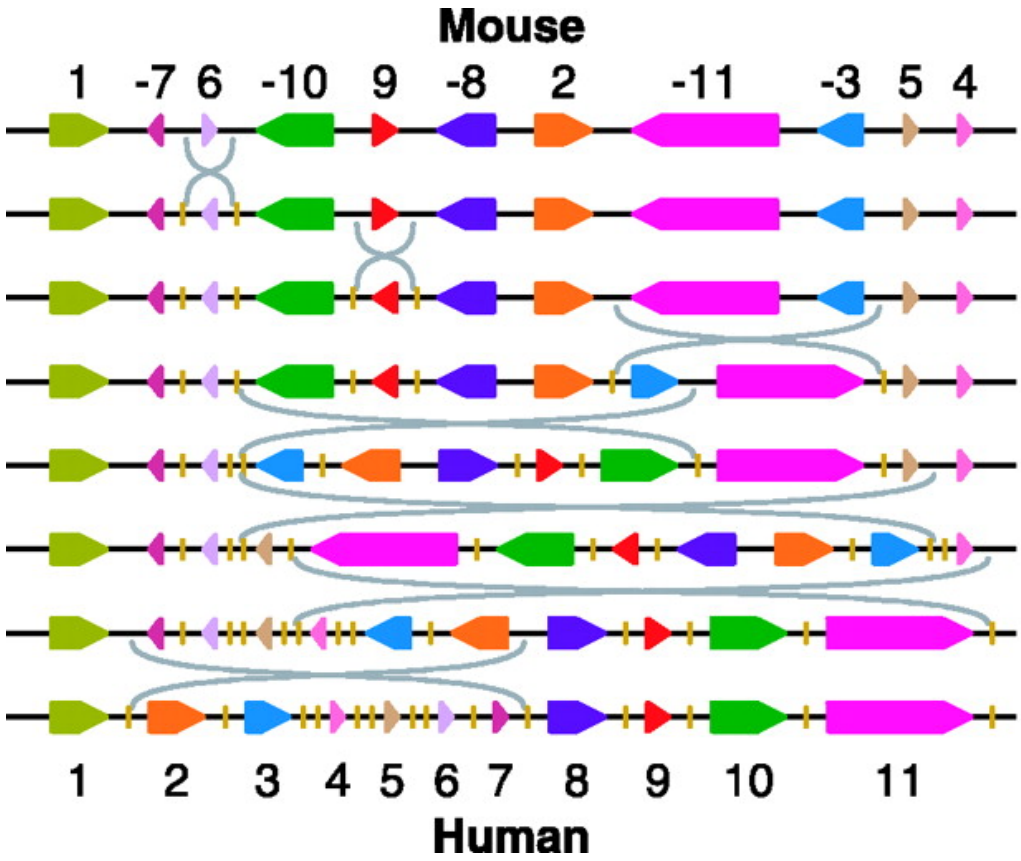
Source: National Center for Biotechnology Information (NCBI), USA

A Bacterial Genome: *E. coli*



Source: P J Enyeart *et al.*,Molecular Systems Biology (2013) 9, 685

X chromosome: Mouse vs. Human



Source: P Pevzner, G Tesler; PNAS June 24, 2003 100 (13) 7672–7677

Genome Evolution

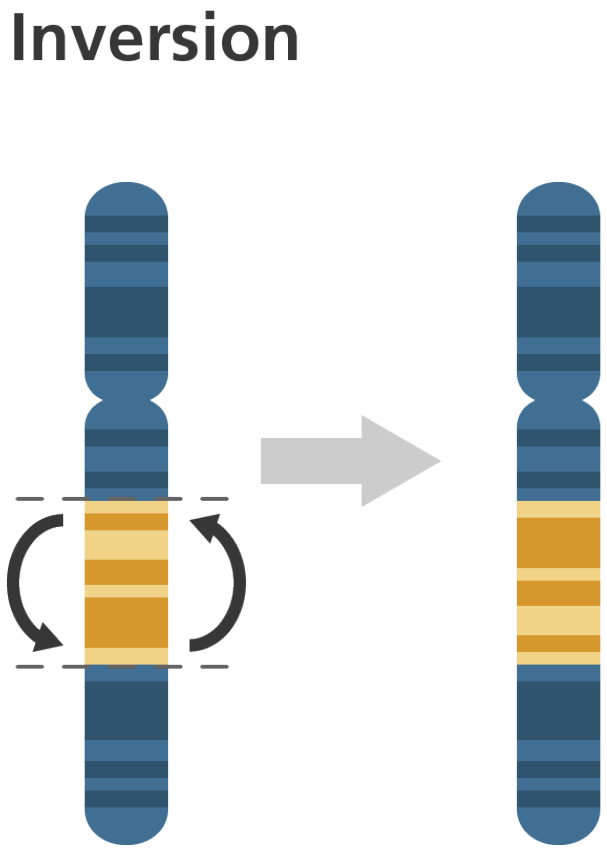
Events

- Point mutations
- Inversions
- Translocations
- Transpositions
- Duplications
- Gain/loss
- Horizontal transfer
- Many others

Our focus

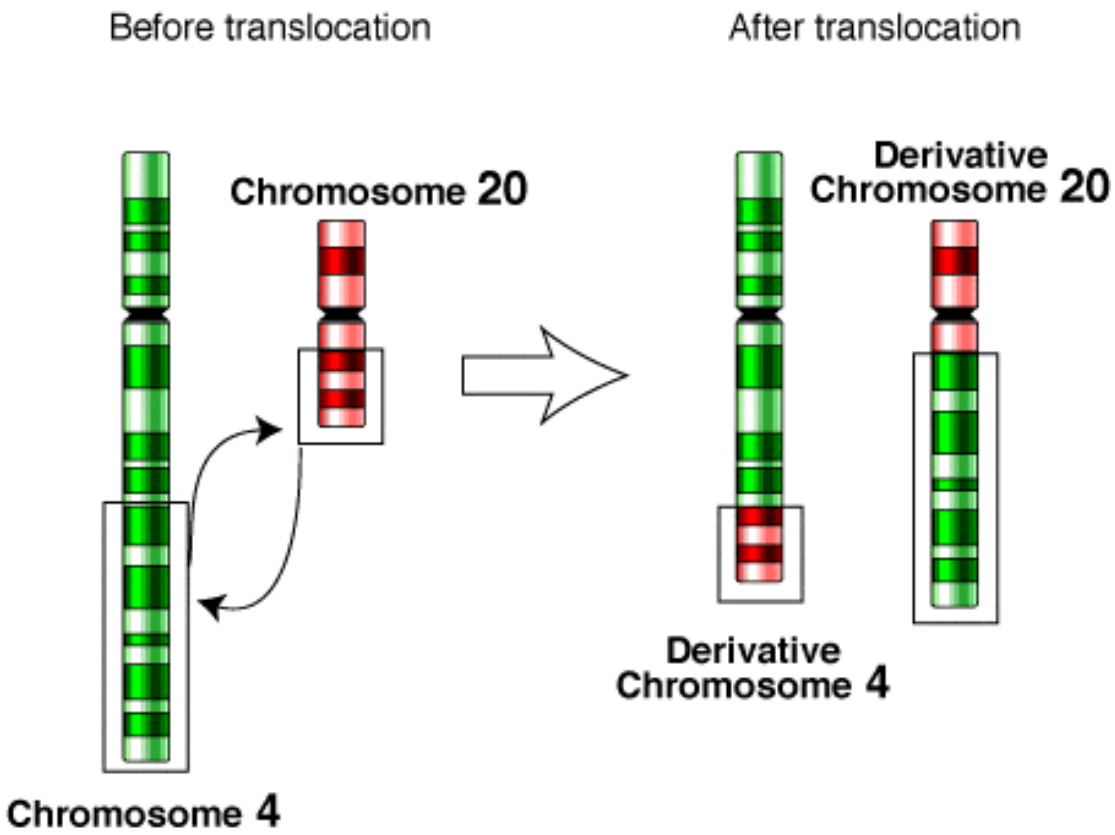
- Genome rearrangements that don't change gene content

Inversion



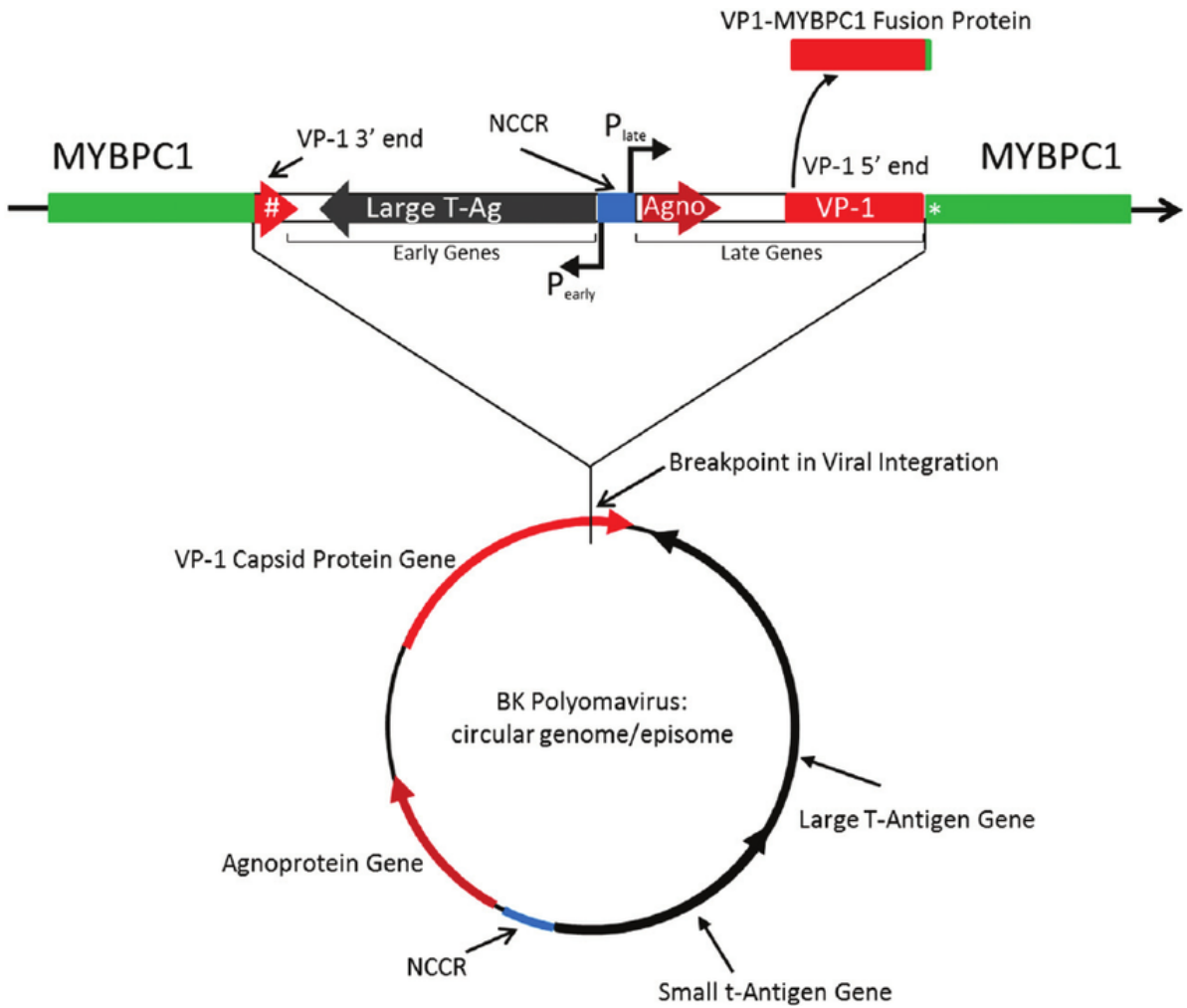
Source: yourgenome, Public Engagement Team, Wellcome Genome Campus, accessed 2017-11-08

Translocation



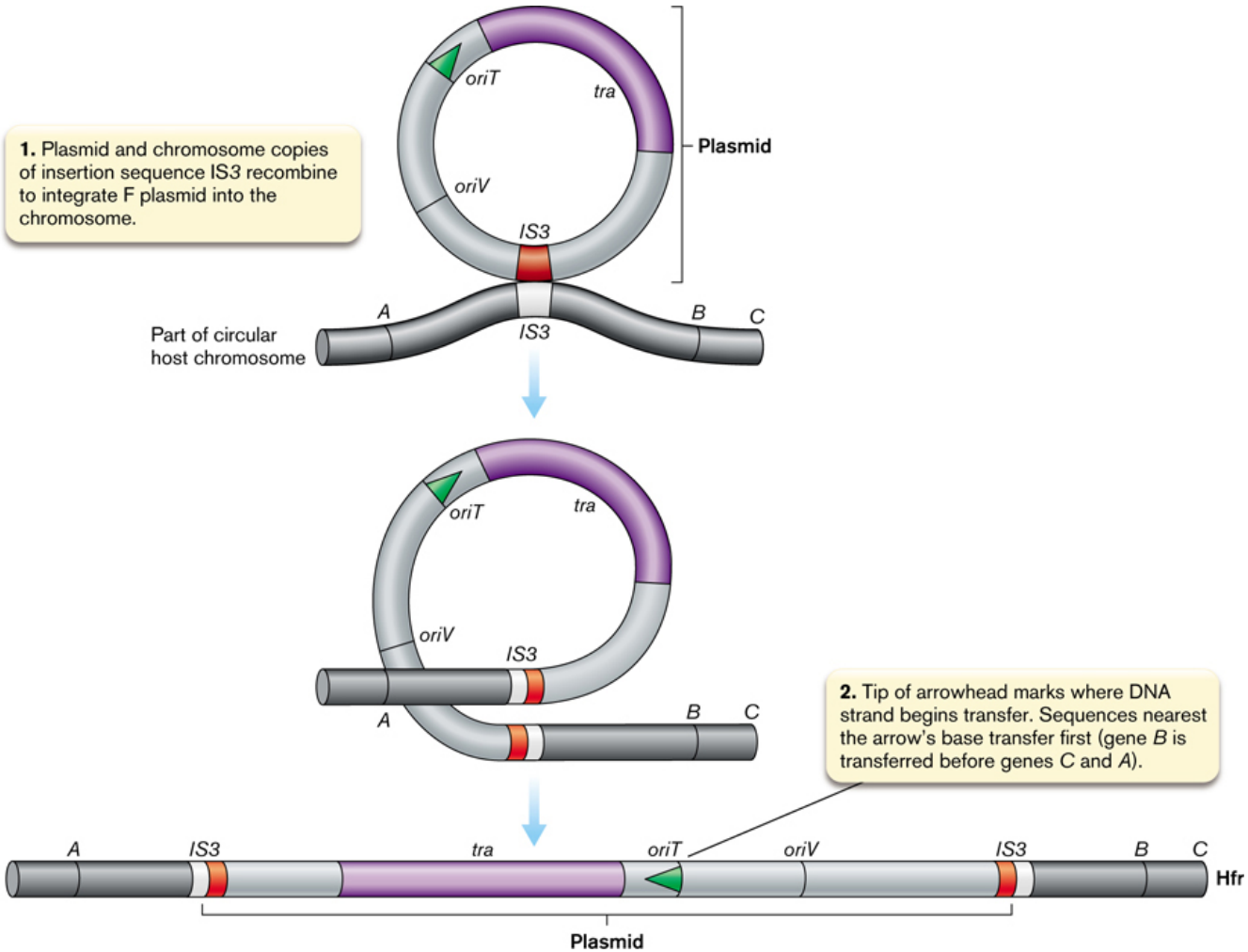
Source: Wikipedia, Chromosomal translocation, accessed 2017-11-08

Integration of circular virus into human genome



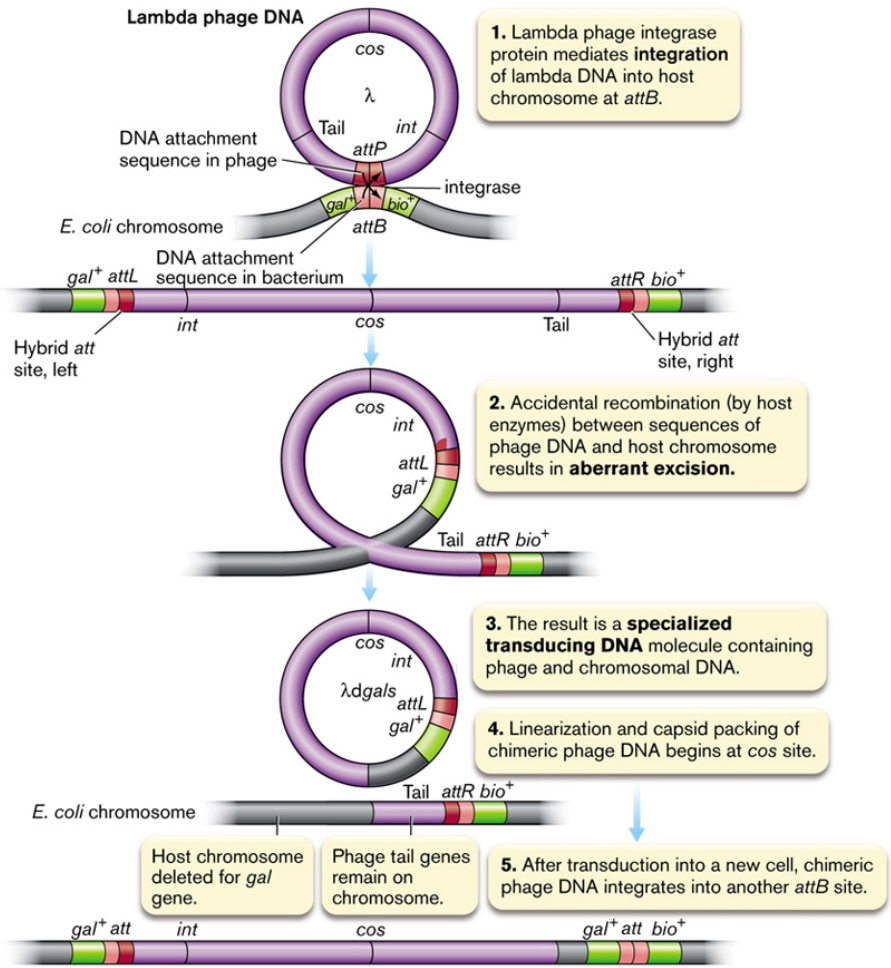
Source: Kenan DJ, Mieczkowski PA, Burger-Calderon R, Singh HK, Nickeleit V., J Pathol. 2015 Nov 237(3):379–389

Integration of plasmid into bacterial genome



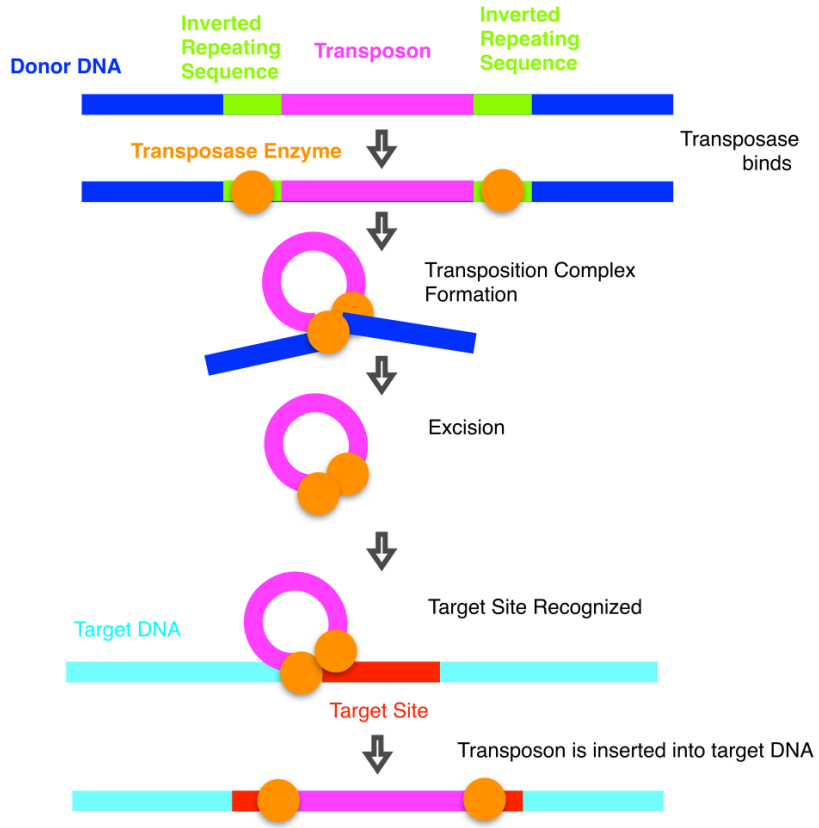
Foster J, Aliabadi Z, Slonczewski J., Microbiology: The Human Experience, W. W. Norton & Company, Inc., Indep. Publ., 2017

Integration/excision of phage lambda



Foster J, Aliabadi Z, Slonczewski J., Microbiology: The Human Experience, W. W. Norton & Company, Inc., Indep. Publ., 2017

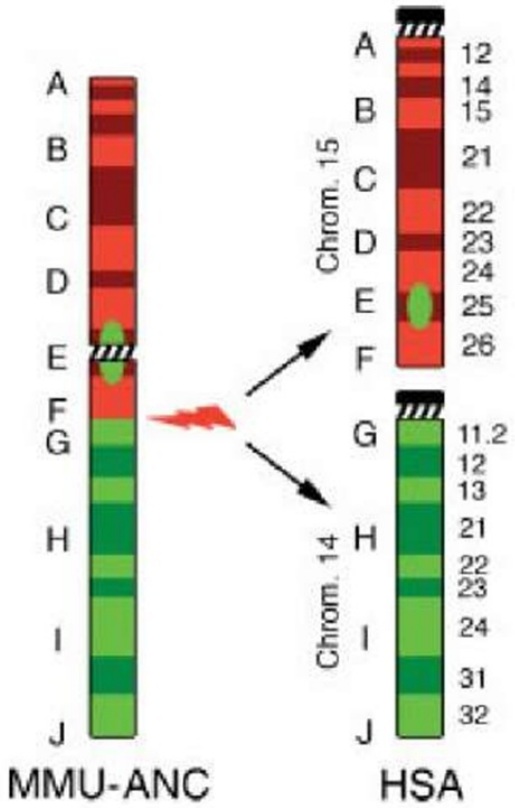
Transposition



Source: Created by Alana Gyemi; accessed in Wikipedia, Chromosomal translocation, 2017-11-12

Chromosome Fission

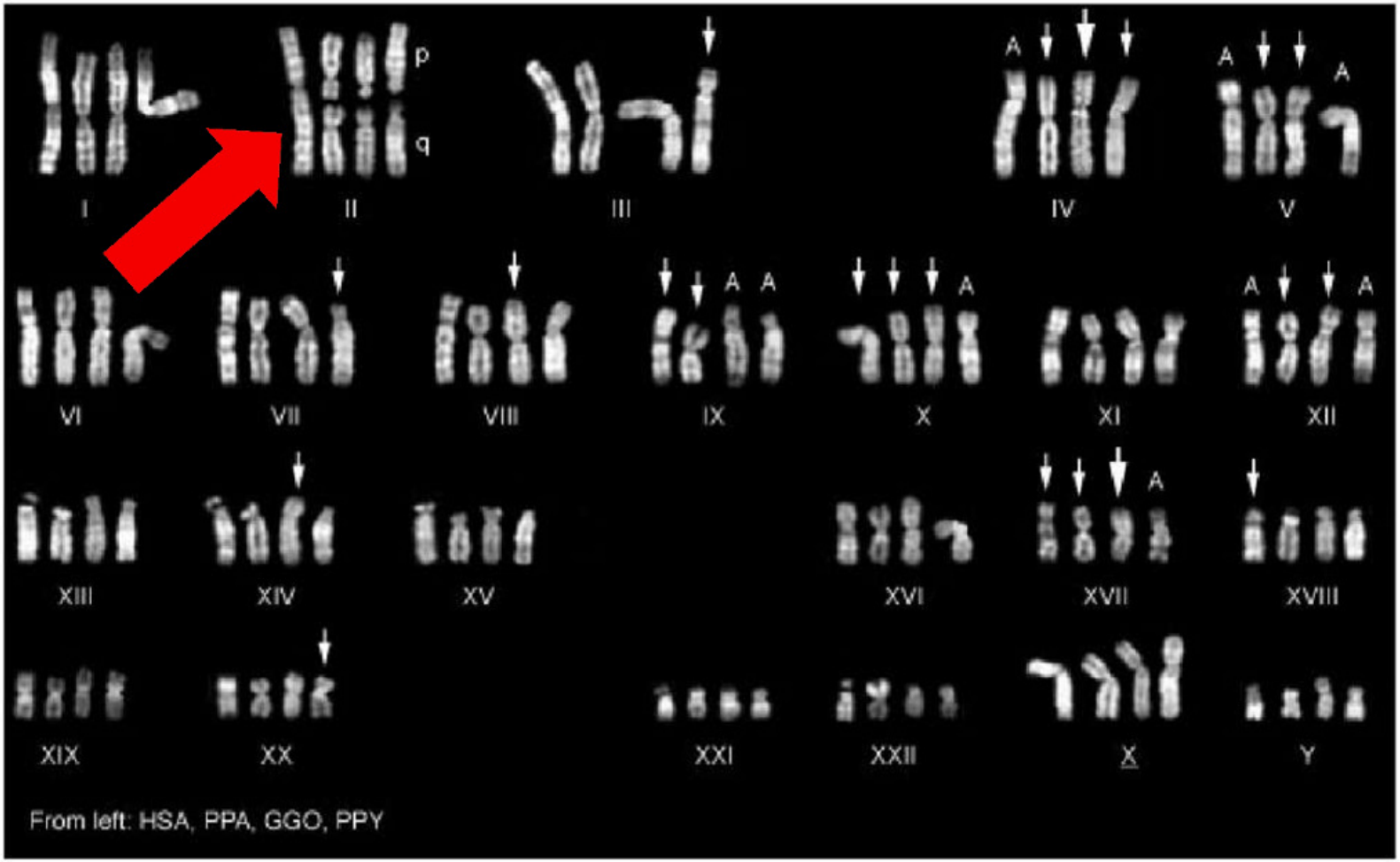
Macaca mulatta
(Rhesus macaque)



Homo sapiens

Source: what-when-how, Genomics, Comparisons with primate genomes; accessed on 2017-11-14

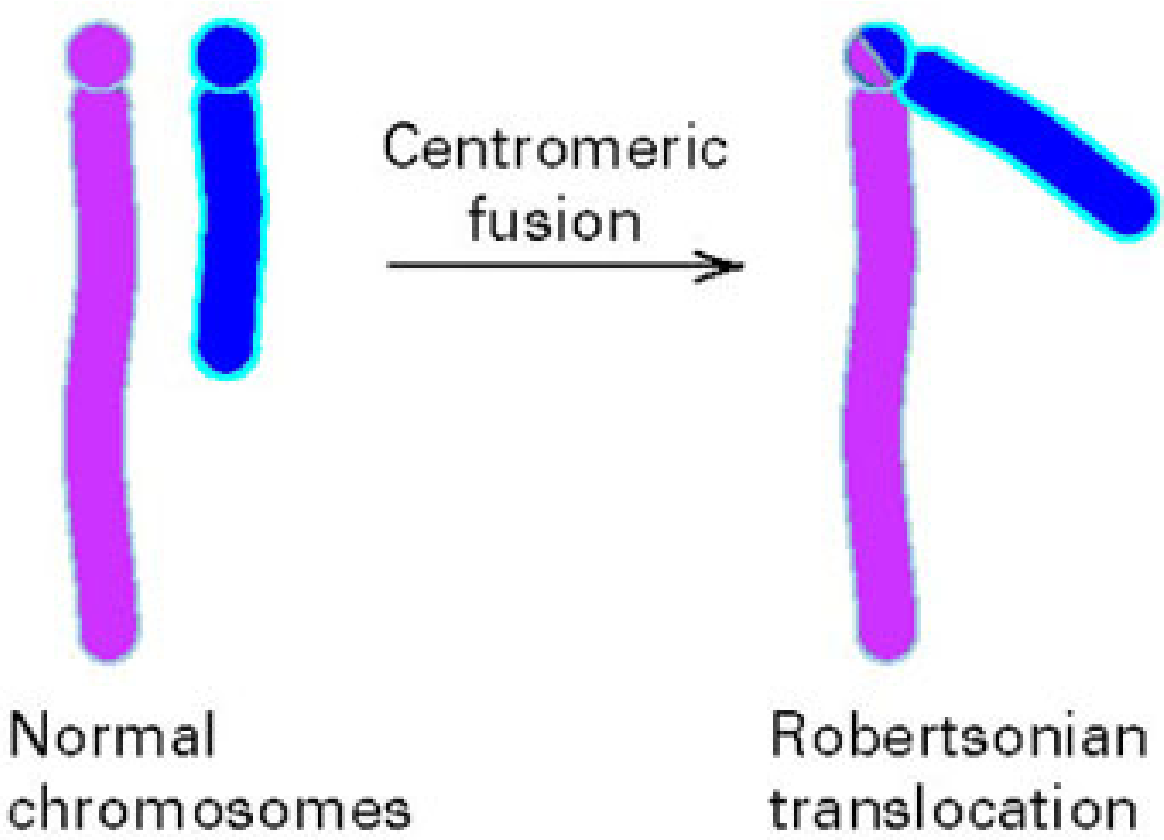
Chromosome Fusion



Homo sapiens, Pan paniscus, Gorilla gorilla, Pongo pygmaeus pygmaeus

Source: what-when-how, Genomics, Comparisons with primate genomes; accessed on 2017-11-14

Chromosome Fusion



Source: Dr. Dana M. Krempels, University of Miami, Course: Genetics (BIL250), Fall 2017 Lecture Notes, Lecture 8: Mutations at the Chromosome Level; accessed on 2017-11-14

Linearization

Journal List › EMBO Rep › v.8(2); 2007 Feb › PMC1796773

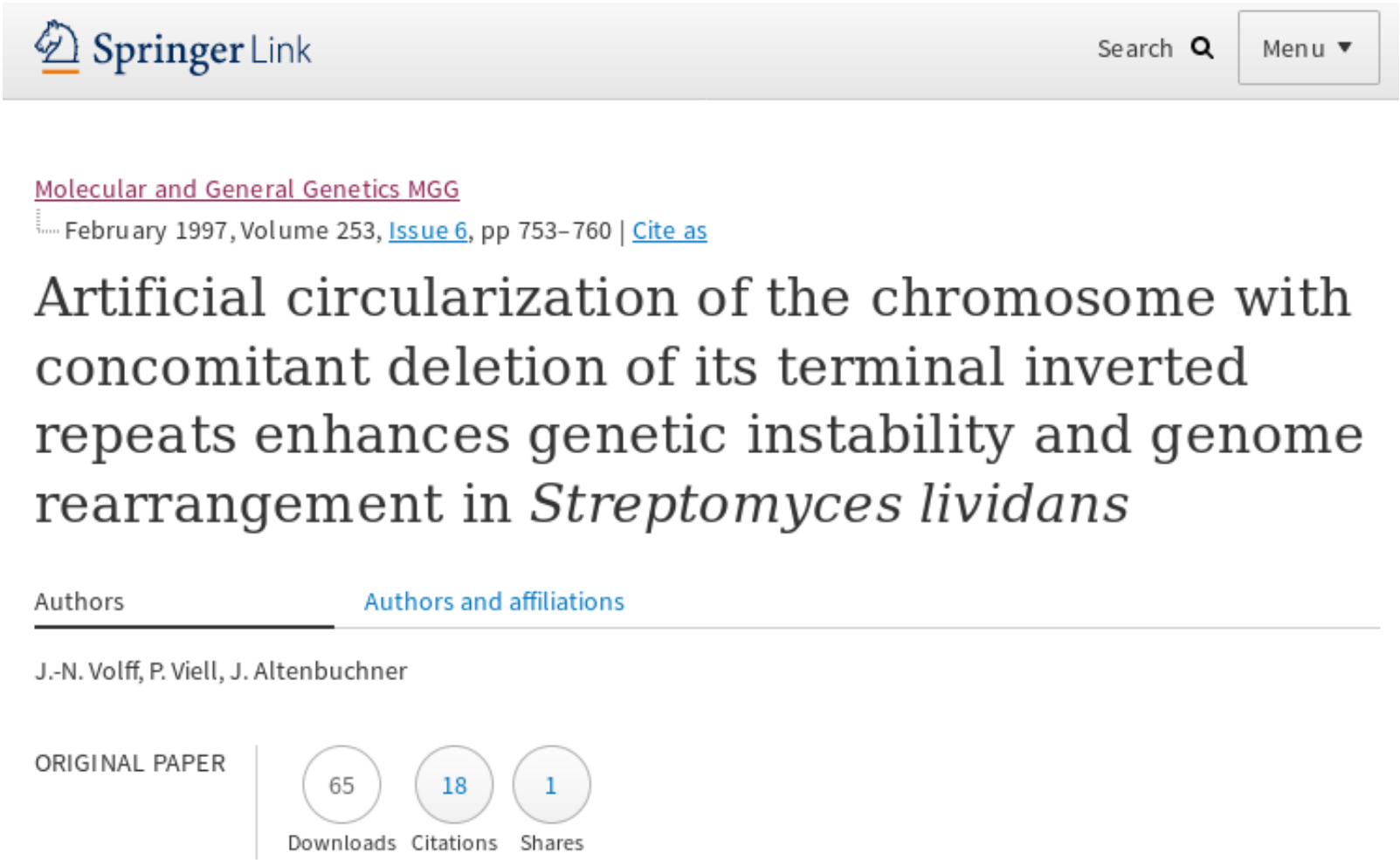
EMBO Rep. 2007 Feb; 8(2): 181–187. PMCID: PMC1796773
Published online 2007 Jan 12. doi: 10.1038/sj.embor.7400880
Scientific Report

Escherichia coli with a linear genome

Tailin Cui,[1] Naoki Moro-oka,[1] Katsufumi Ohsumi,[1] Kenichi Kodama,[1] Taku Ohshima,[2] Naotake Ogasawara,[2] Hirotada Mori,[2] Barry Wanner,[3] Hironori Niki,[4] and Takashi Horiuchi[1,a]

Author information ► Article notes ► Copyright and License information ►

Circularization



Genome Rearrangement Problems

- **Distance:** Minimum # of rearrangements from A to B ?
- **Scenario:** Which rearrangements ?
- **Phylogeny:** How did the genomes evolve ?
- **Reconstruction:** What did the ancestors look like ?

Some Milestones

Year	Milestone
1995	polynomial-time algorithm for inversion distance (first polynomial-time algorithm)
2000	algebraic distance: several events (circular genomes only)
2005	double-cut-and-join distance (contemplates linear chromosomes)
2012	algebraic distance extended to linear chromosomes (another extension to linear chromosomes)
2016	rank distance (twice the algebraic distance; uses matrices)

Genomes in Rearrangement Studies

Positional View

- $P : [1..n] \mapsto$ genes (unsigned)
- $P : [1..n] \mapsto$ gene ends (signed)

Algebraic View

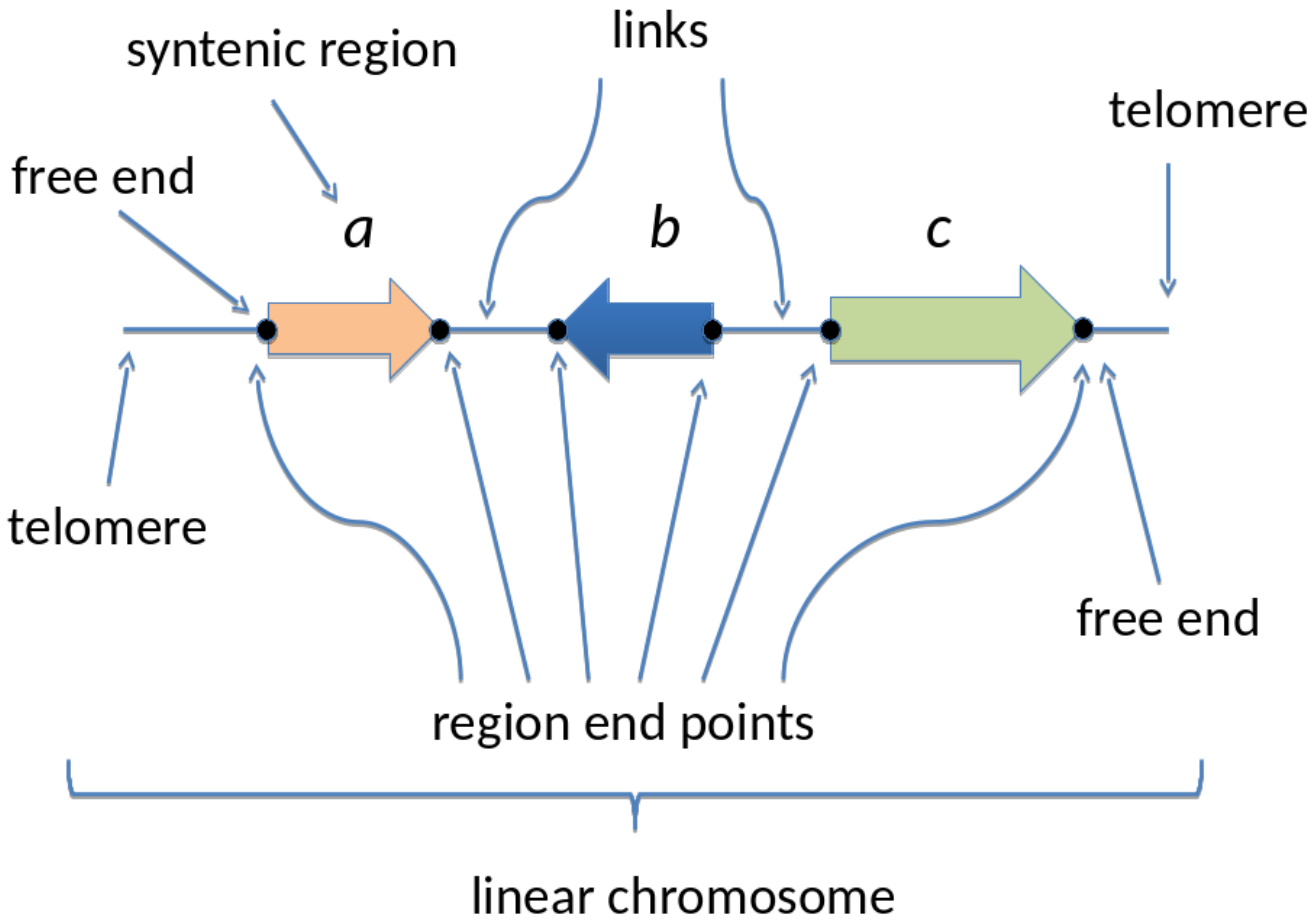
- $P :$ genes $\mapsto$ genes (unsigned)
- $P :$ gene ends $\mapsto$ gene ends (signed)

Comparing Two (or more) Genomes

- Breakpoint Graph
- Adjacency Graph

Genomes Matrices and Rank Distance

Genome elements



- Links: $\{a_h, b_h\}, \{b_t, c_t\}$; free ends: a_t, c_h

Representing genomes as matrices

- Links: $\{a_h, b_h\}, \{b_t, c_t\}$; free ends: a_t, c_h

$$
\begin{array}{c}
 \\
a_t \\
a_h \\
b_t \\
b_h \\
c_t \\
c_h
\end{array}
\begin{array}{c}
\begin{array}{cccccc} a_t & a_h & b_t & b_h & c_t & c_h \end{array} \\
\left[\begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}\right]
\end{array}
$$

Properties

- symmetric matrix ($A = A^\top$)
- orthogonal matrix ($A^\top = A^{-1}$)
- involution ($A^2 = I$)

Permutation matrices

- binary (just 0's and 1's)
- orthogonal ($A^\top = A^{-1}$)
- not necessarily symmetric

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 & 5 & 1 & 4 & 6 \end{bmatrix}$$

Rank Distance

- Distance between two genome matrices is the rank of their difference

$$d(A, B) = r(A - B)$$

Properties

- Rank: maximum number of linearly independent rows
- $d(A, B) = 0$ if and only if $A = B$
- $d(A, B) = d(B, A)$
- $d(A, C) \leq d(A, B) + d(B, C)$

Biological Significance of Rank Distance

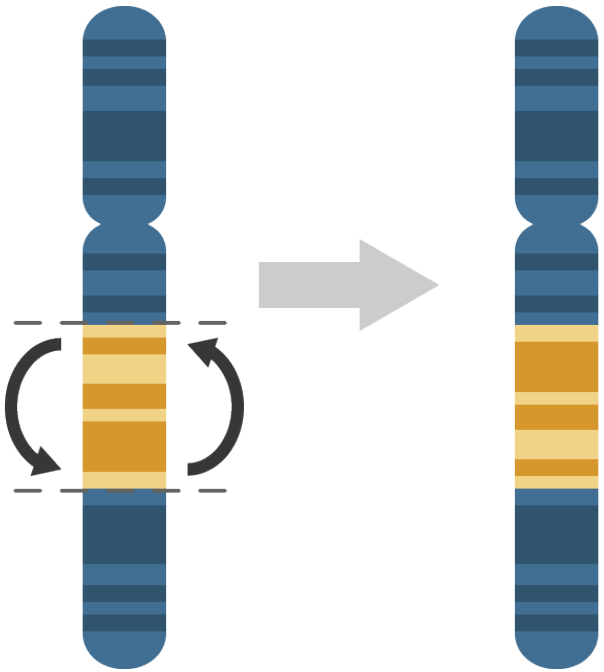
frequent rearrangements $\approx$ small rank

rank distance $\approx$ composition of small rank operations

rank distance $\approx$ amount of evolution

Inversion RANK = 2

Inversion



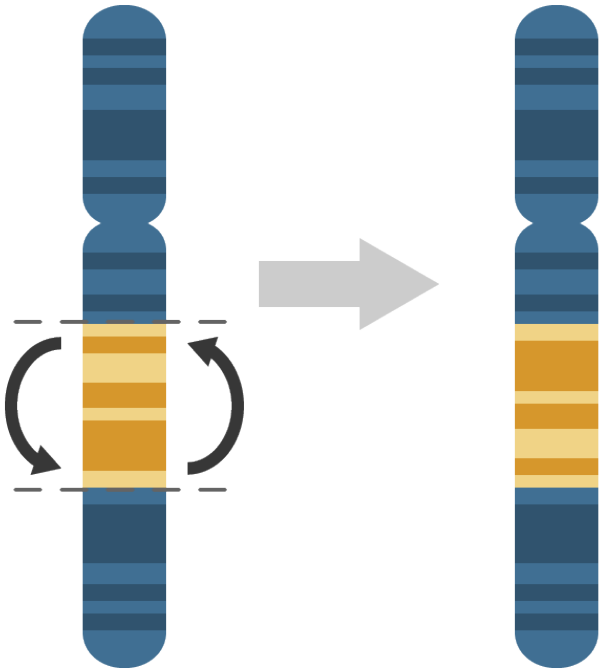
$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

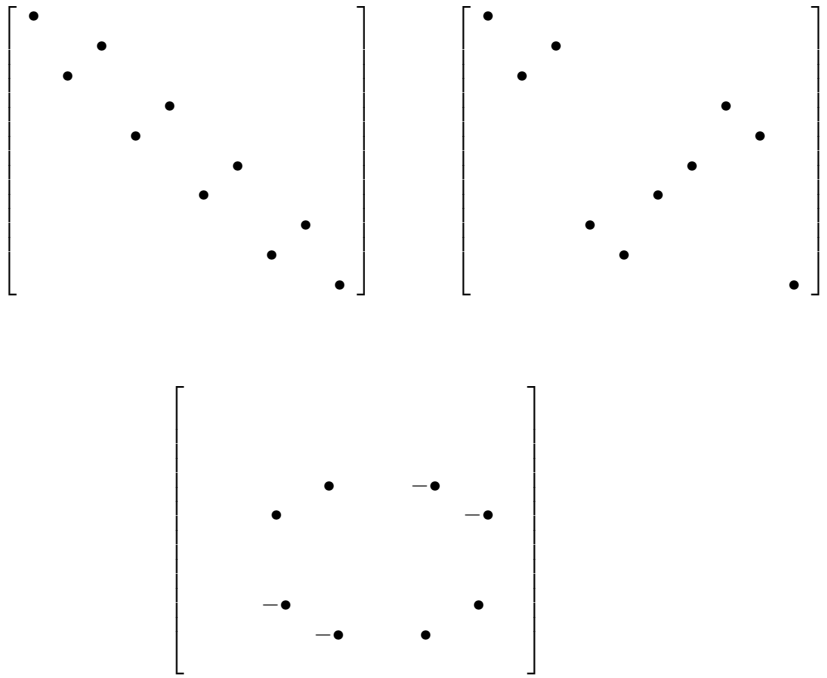
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Source: yourgenome, Public Engagement Team, Wellcome Genome Campus, accessed 2017-11-08

Inversion RANK = 2

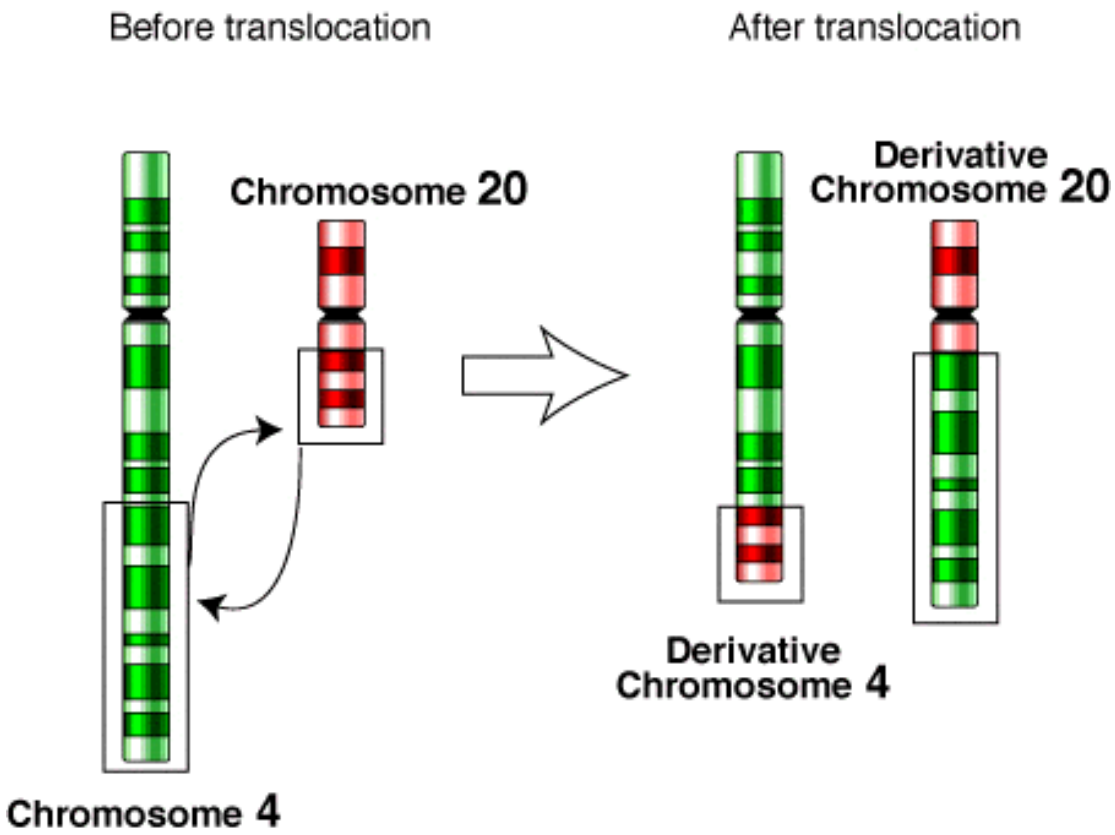
Inversion





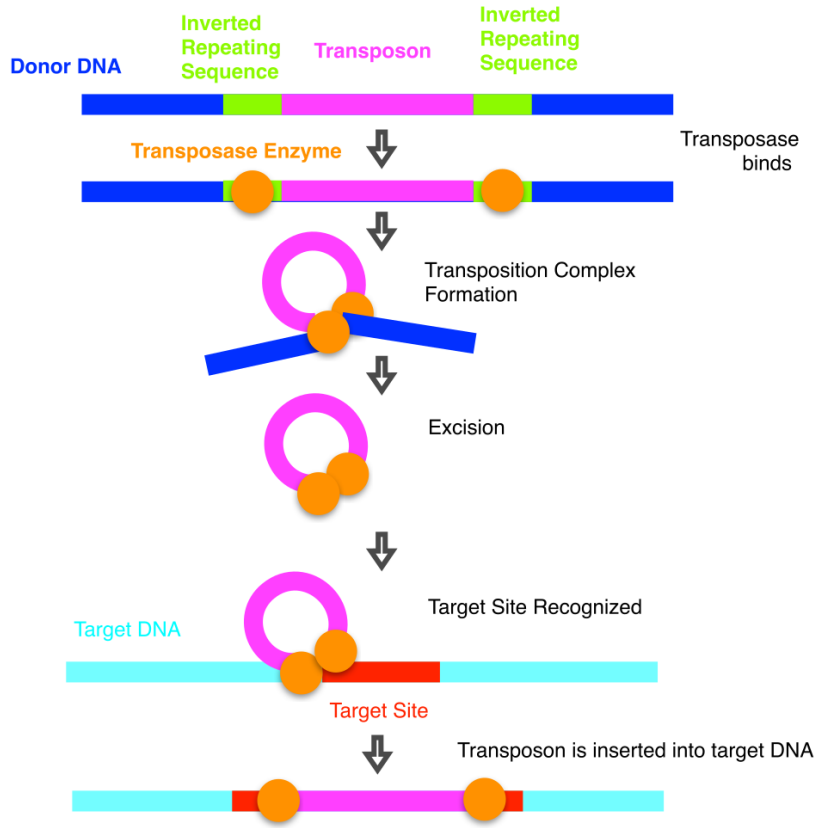
Source: yourgenome, Public Engagement Team, Wellcome Genome Campus, accessed 2017-11-08

Translocation RANK = 2



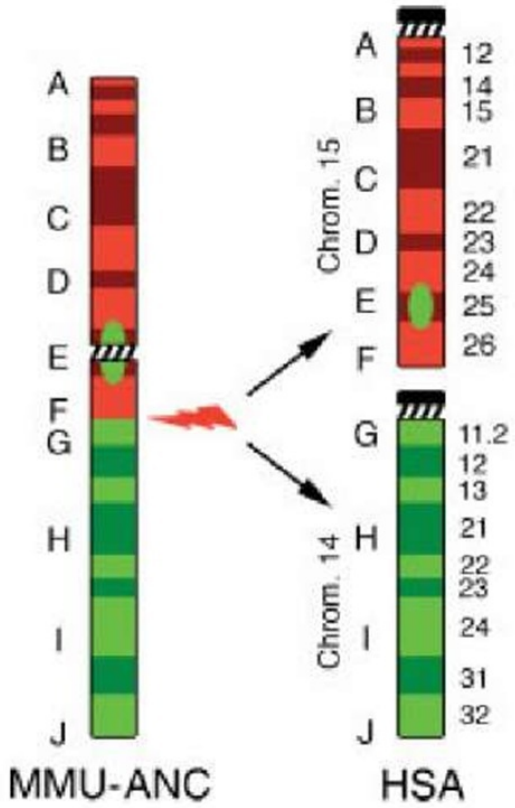
Source: Wikipedia, Chromosomal translocation, accessed 2017-11-08

Transposition RANK = 4



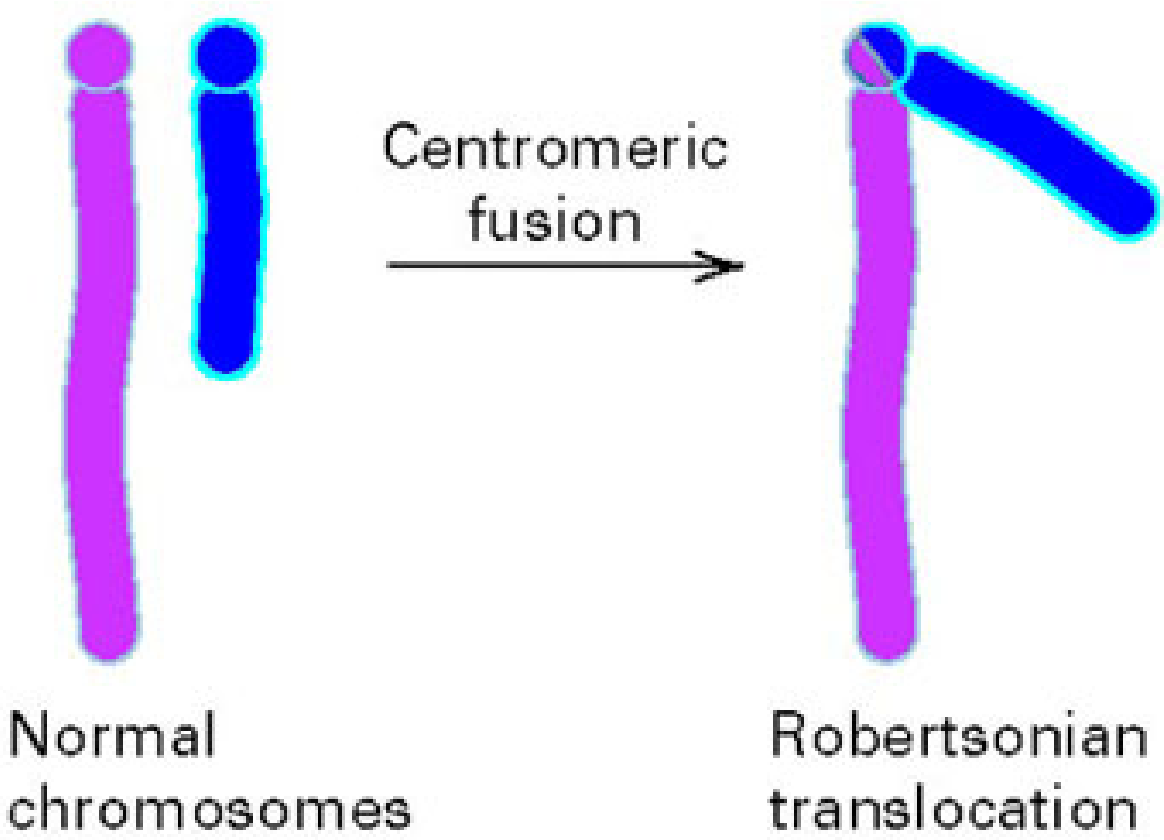
Source: Created by Alana Gyemi; accessed in Wikipedia, Chromosomal translocation, 2017-11-12

Chromosome Fission RANK = 1



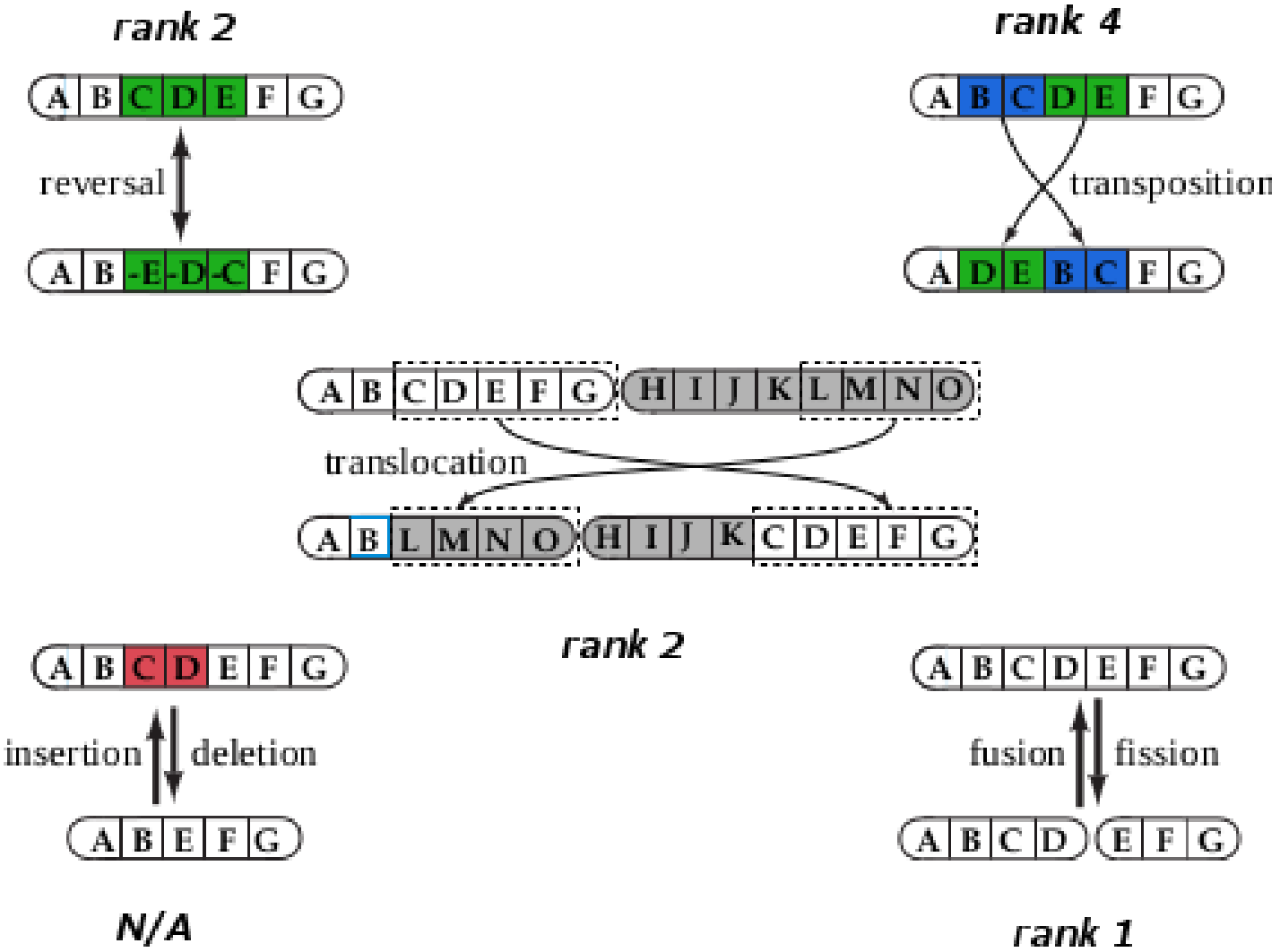
Sorce: what-when-how, Genomics, Comparisons with primate genomes; accessed on 2017-11-14

Chromosome Fusion RANK = 1



Source: Dr. Dana M. Krempels, University of Miami, Course: Genetics (BIL250), Fall 2017 Lecture Notes, Lecture 8: Mutations at the Chromosome Level; accessed on 2017-11-14

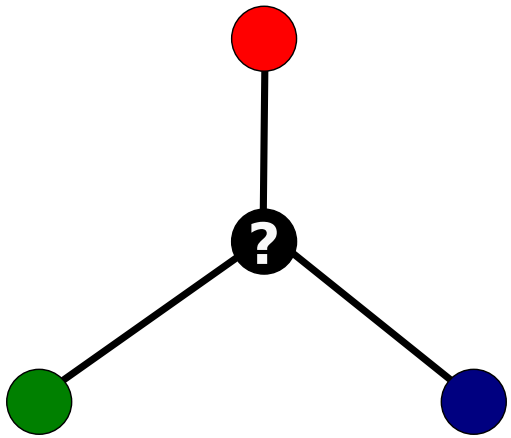
Genome Rearrangements



Ancestral Reconstruction: The Median Problem

Genome Median Problem

Given three input genomic matrices A, B, and C:

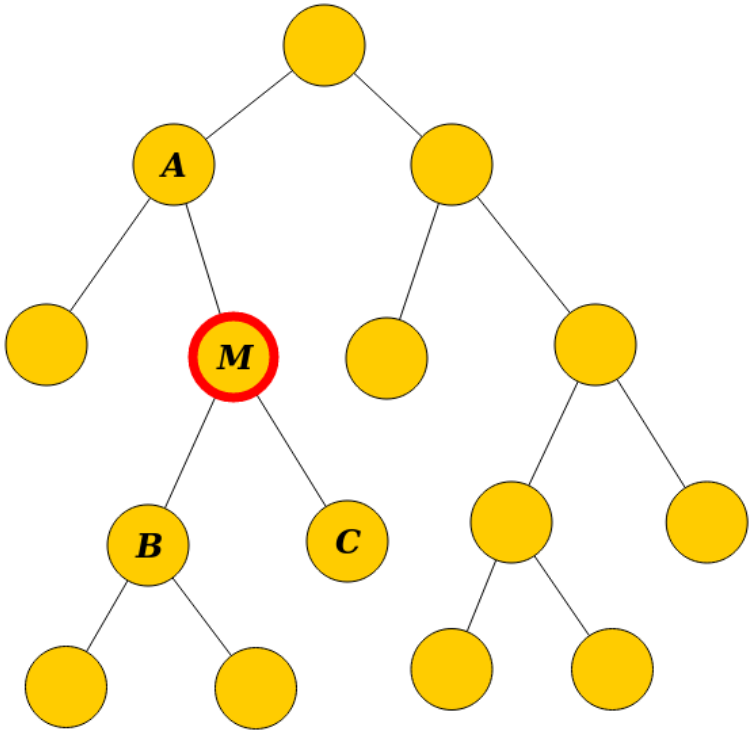


Find matrix M minimizing:

$$\text{score}(M; A, B, C) = d(M, A) + d(M, B) + d(M, C).$$

Application

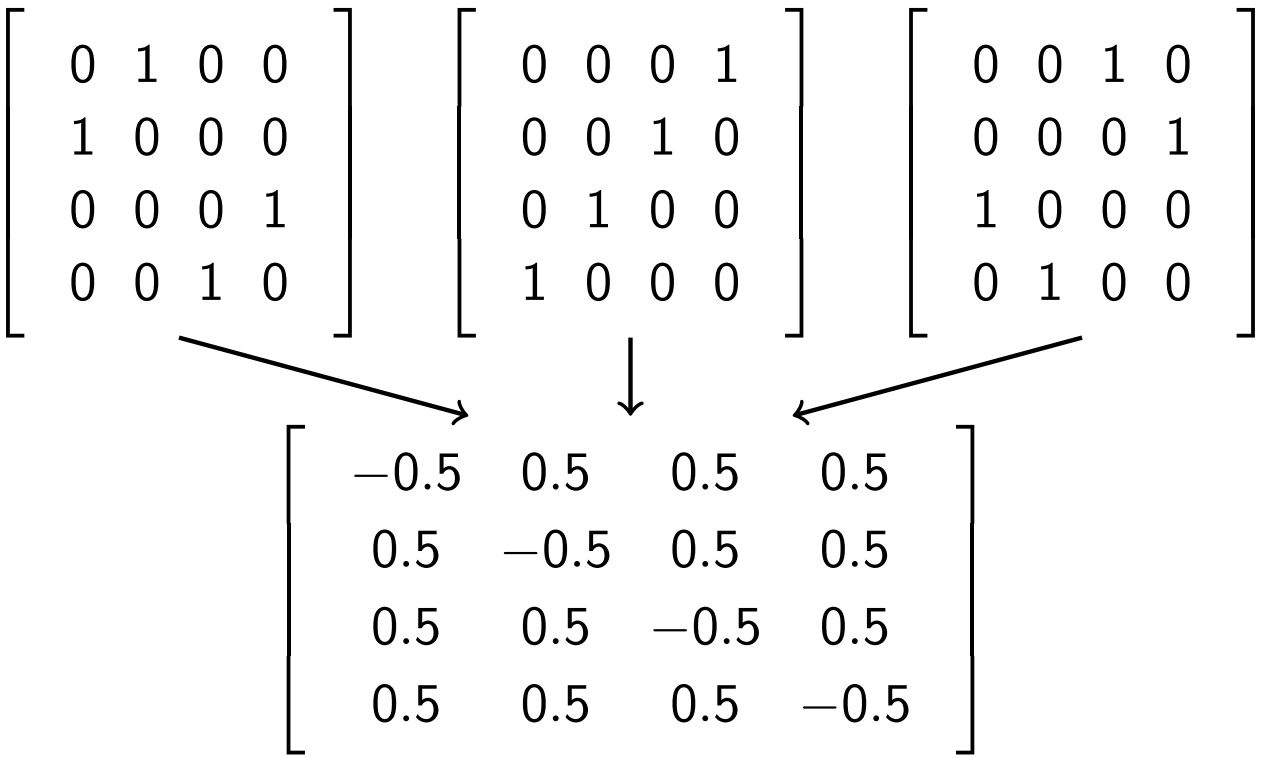
Refine ancestral genomes



Repeat for all internal nodes until convergence

Matrix median may not be genomic

- Example of input matrices and their unique median



Once you go to matrices ...

What kinds of medians?

Median Problems

Input	Output **Genomic**	**Permutation**	**General**
Genomic	NP-hard		$O(n^{\omega})$
Permutation	NP-hard	NP-hard	$O(n^{\omega+1})$
General			

$O(n^{\omega})$: matrix multiplication/inverse

- Need a way to go back from general matrices to genomes

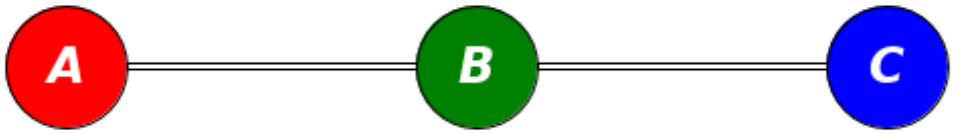
Algorithms for the Median Problem

Preliminary Observations:

- **Lower bound**

$$\text{score}(M; A, B, C) \geq \frac{d(A, B) + d(B, C) + d(C, A)}{2}$$

- Genome B **intermediate** between A and C:



$$d(A, C) = d(A, B) + d(B, C).$$

- Median M has score equal to the lower bound $\Longrightarrow$ M is intermediate between any two of A, B, C.

Algorithms for the Median Problem

- Interesting Property

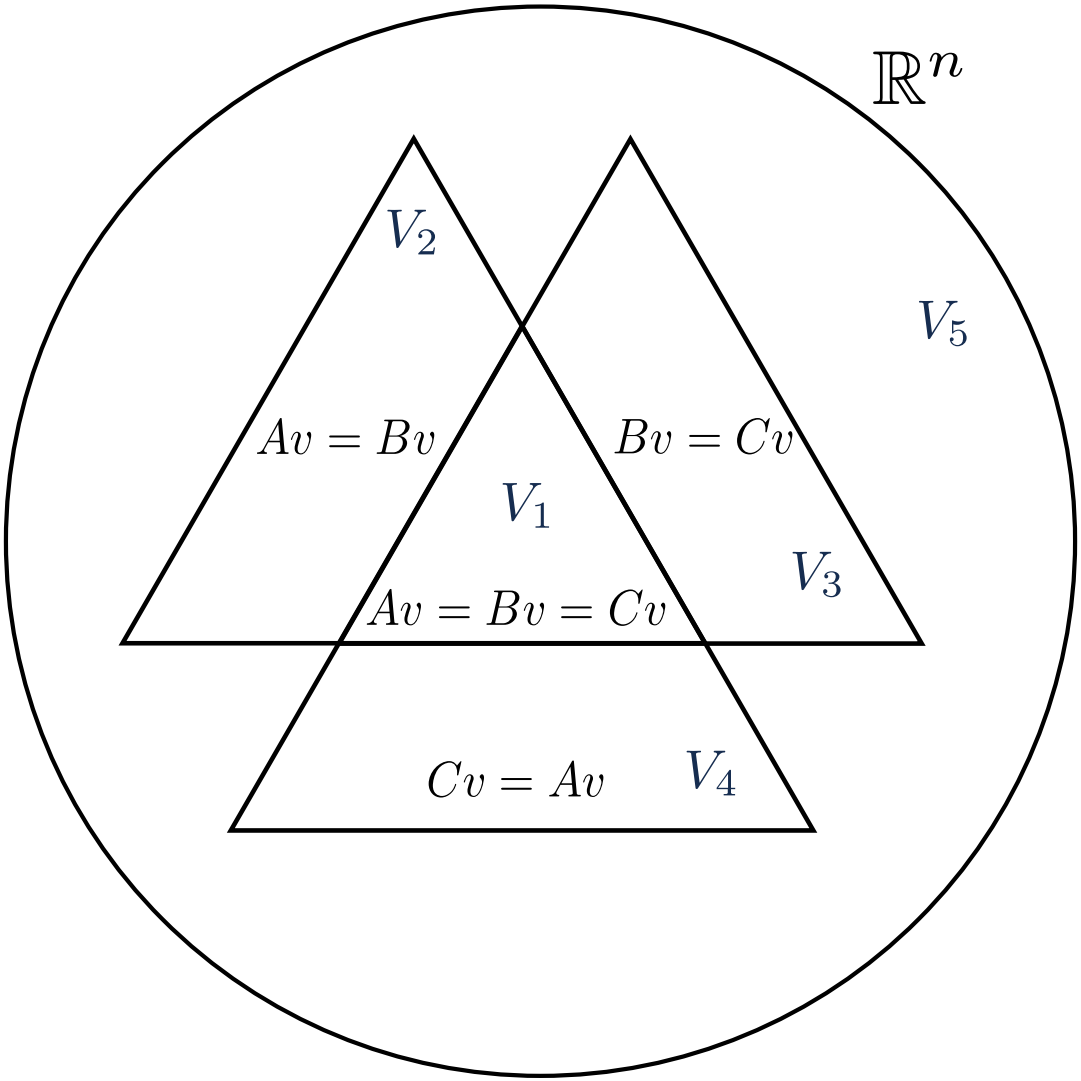
Theorem

For any three $n \times n$ genome matrices A, B, and C there is a median M satisfying: for all vectors $v \in \mathbb{R}^n$ such that $Av = Bv = Cv$, we have $Mv = Av$.

- Can we say the same if just $Av = Bv$? Open (with partial results).
- However, we can act on this idea.

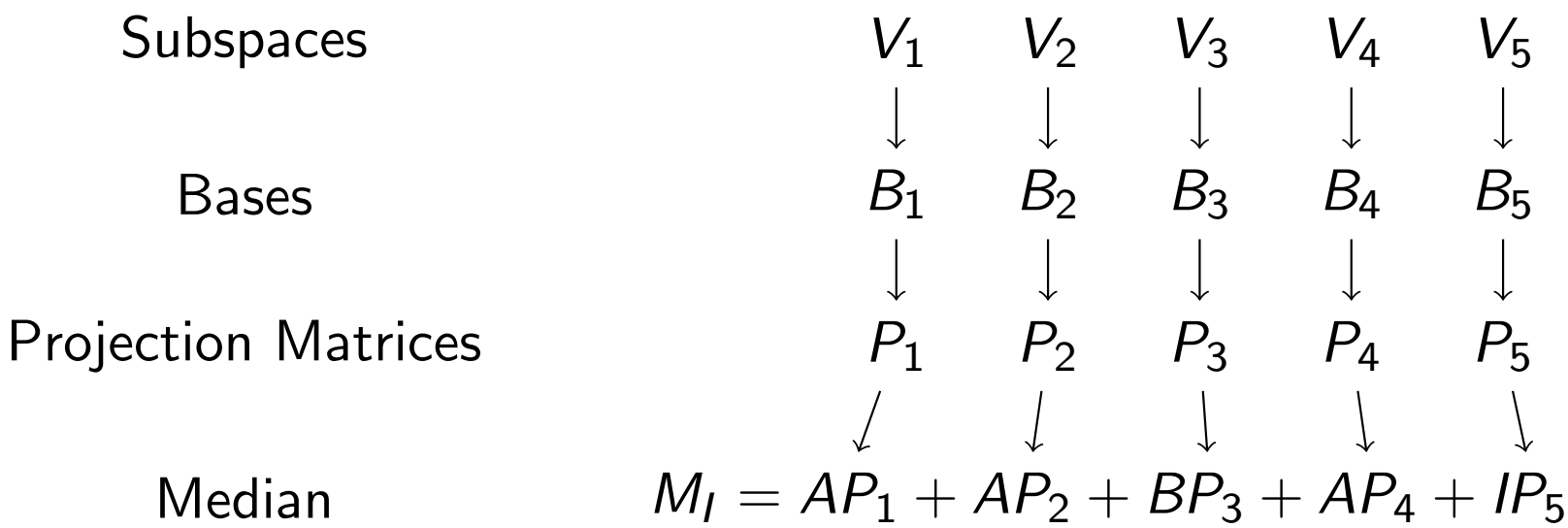
Algorithms for the Median Problem

Division into subspaces



M_I Median — $O(n^\omega)$

- Specific for genome matrices
- M_I follows majority in V_1 through V_4
- M_I follows I in V_5



Efficient Computation

Technical Improvements

- B_i don't need to be orthonormal, $i = 1..4$
- B_i's computed from permutation **vectors** and DFS
- B_i's all binary
- Improved formula

$$M_I = I + ([AB_1\ AB_2\ BB_3\ AB_4] - B_{14})(B_{14}^T B_{14})^{-1} B_{14}^T$$

  where $B_{14} = [B_1\ B_2\ B_3\ B_4]$
- B_5 not needed

Orthogonal medians

Theorem

For any three orthogonal matrices A, B, and C, all general medians satisfy the lower bound.

- Algorithm

function MEDIAN(A, B, C)
  while $d(A, B) + d(B, C) > d(A, C)$ **do**
    Find non-zero $u \in \text{im}(A - B) \cap \text{im}(C - B)$
    $B \leftarrow B - 2uu^T B / u^T u$
  return B

- Nondeterministic
- "Walks" from B towards median

Implementation

Hardware

- Laptop
- 8 GB memory
- 4 AMD A8-7410 cores

Software

- Windows 10 + WSL
- GNU Octave 3.8.1 (Matlab)
- Also code in R, Python

Data Sets

Simulation

- Start with random genome
- Apply random rearrangement operations
- Repeat to get A, B, C

Parameters

- sizes: 12, 16, 20, 30, 50, 100, 200, 300, 500 gene ends
- type of operation: Add/remove adjacencies (near) or DCJ (far)
- number of operations: 5% to 30%, in 5% increments
- 10 $\times$ each
- 1,080 instances: 540 near, 540 far

Results — $M_{/}$ algorithm

Near

- For all 540 cases, the algorithm finds a median
- Median is genomic in 535 cases

Far

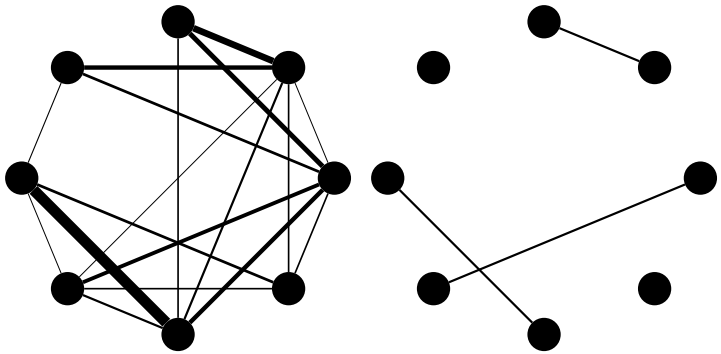
- For all 540 cases, the algorihtm finds a median
- Median is genomic in 254 cases

Times (in minutes) to run all instances of a given size

size	mi-Near	mi-Far
500	9:52	8:24
300	3:26	2:34
200	1:40	1:08
100	0:30	0:24

From general matrices to genomes

$$\begin{bmatrix} 0.2 & 0.8 & 0.5 & 0 & 0 & 0.4 & 0 & 0.1 \\ 0.4 & 0 & 0 & 0 & 0 & -0.3 & 0 & 0.6 \\ 0.3 & 0 & -0.5 & 0.2 & 0 & 0 & 0 & -0.3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -0.1 & 0 & 0 & 0.1 & 0.1 & 0.4 & 0.2 & 0.7 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -0.3 & 0 & 0 & 0.5 & -0.1 & 0 & 0.4 & 0.1 \\ 0 & 0.8 & 0.2 & 0 & 0 & 0.8 & 0.2 & 0.3 \end{bmatrix}$$



$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

- Assign weight $|a_{ij}| + |a_{ji}|$ to edge ij
- Take a maximum weight matching as your solution
- A genome is a matching of gene ends

Extensions and Other Properties

How to deal with indels

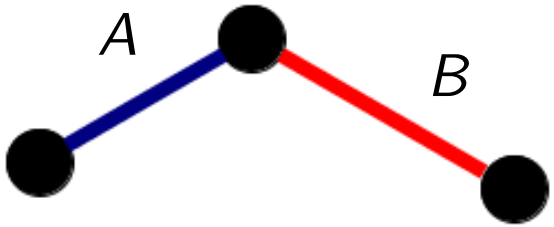
- Matrices with empty rows/columns

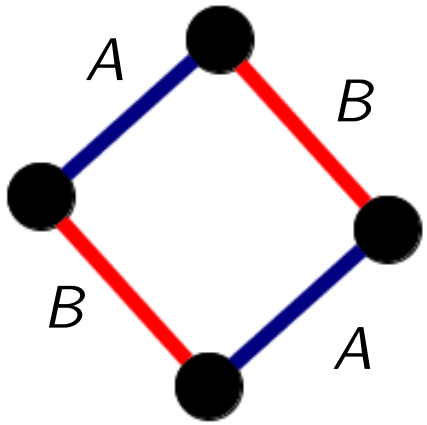
$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- New operations: inserion/deletion of genes or gene ends
- Biological significance with mathematical soundness

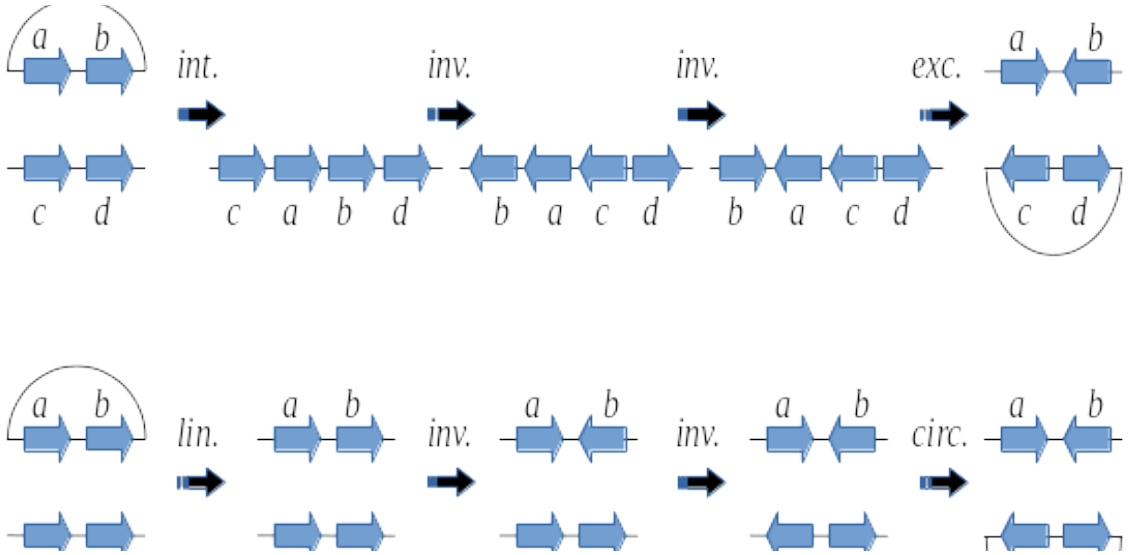
Graph components

- breakpoint graph: no caps
- gene ends: vectors
- components: minimal invariant subspaces: $AV = V$, $BV = V$
- linear components: 1 AB-orbit
- circular components: 2 AB-orbits





No recombination



- Both scenarios are DCJ-optimal
- Only the second scenario is rank-optimal
- Rank-optimal scenarios don't allow chromosome mixing

Counting intermediate genomes

- For a linear graph component with k gene ends:

$$I_p(k) = \binom{k+1}{\lfloor (k+1)/2 \rfloor}$$

- For a circular graph component with $2k$ gene ends:

$$I_c(2k) = \frac{1}{k+1}\binom{2k}{k}$$

Challenges

Future Work

Application to real ancestral reconstruction problems (fungi)

Ongoing study of insertions, deletions, and duplications

More ways of transforming a general matrix into a close, genomic matrix

Do all medians of A, B, C form a manifold ?

If yes, can we "walk" on this manifold to a binary matrix ?

References

- Zanetti, J.P.P., Biller, P., Meidanis, J.
  Median approximations for genomes modeled as matrices.
  Bull Math Biol (2016) 78: 786.

- Chindelevitch, L., Zanetti, J.P.P., Meidanis, J.
  On the Rank-Distance Median of 3 Permutations.
  BMC Bioinformatics (2018) 19 (Suppl 6): 142.

- Chindelevitch, L., Meidanis, J.
  A cubic algorithm for the generalized rank median of three genomes.
  RECOMB CG (2018).

- Meidanis, J., Biller, P., Zanetti, J.P.P.
  A Matrix-Based Theory for Genome Rearrangements.
  U. of Campinas Tech Report IC-18-10 (2018).

Thanks!!

P Biller

JPP Zanetti

L Chindelevitch

Get this presentation:

`http://www.ic.unicamp.br/~meidanis/research/rear/`