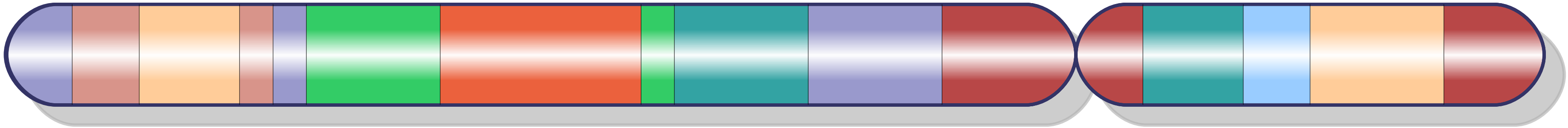
PhD defense

Chromosomal rearrangements in mammalian genomes : characterising the breakpoints



Claire Lemaitre

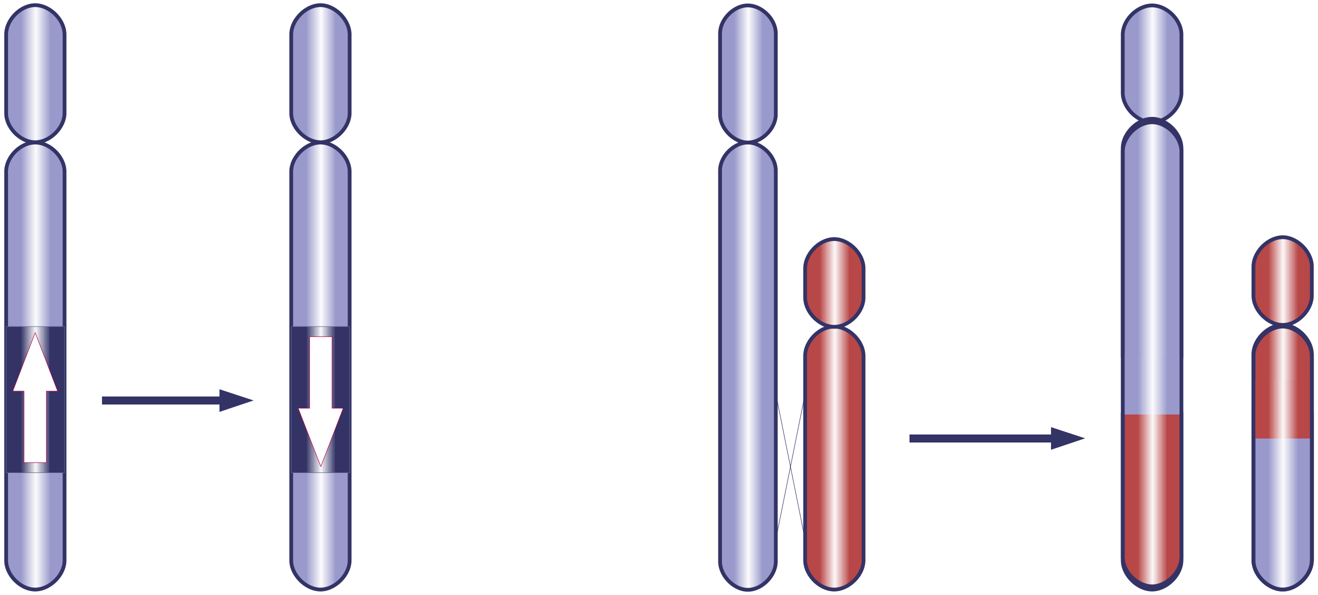
Laboratoire de Biométrie et Biologie Évolutive
Université Claude Bernard Lyon 1

6 novembre 2008

Genome dynamics

- Point mutations: insertion, deletion, substitution
- Large-scale modifications: chromosomal rearrangements

inversions, translocations, transpositions, fusions, fissions, duplications, deletions



Genome dynamics

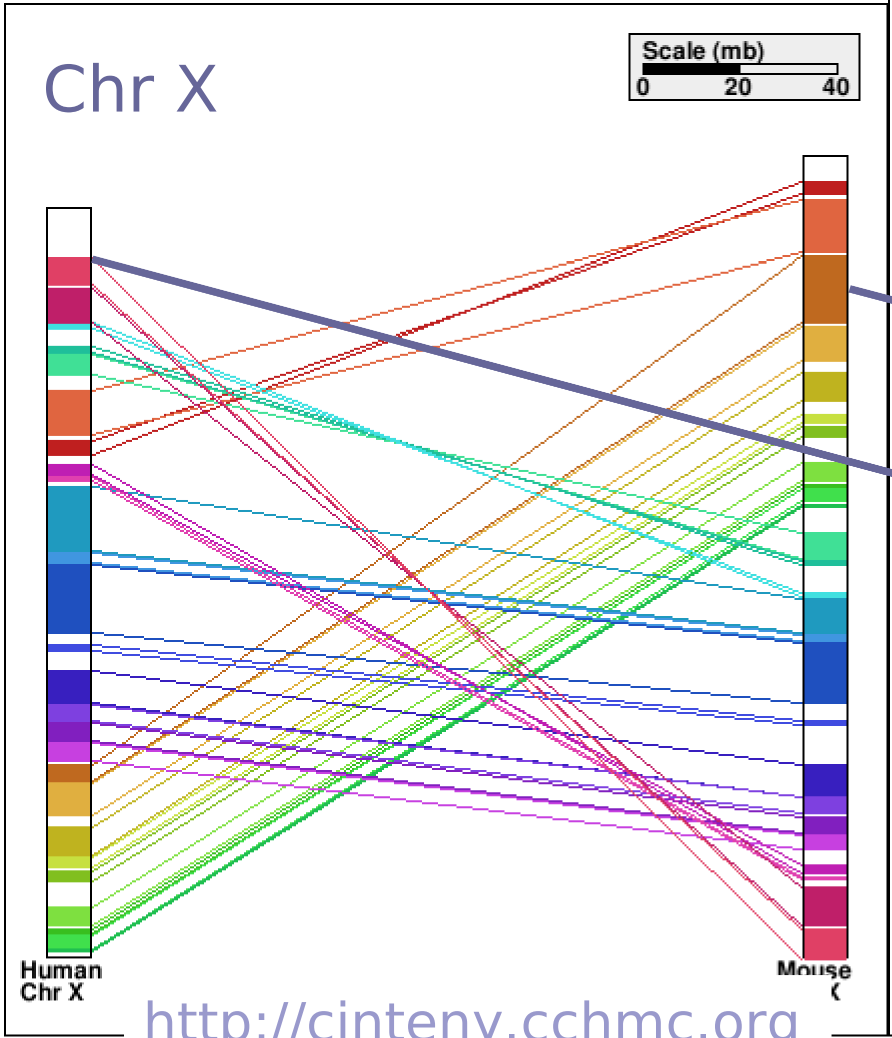
- Functional impacts of rearrangements
  - duplication / deletion
  - breakage in functional sequences
  - modification of the genome organisation
- Rearrangements are found in:
  - inherited diseases
  - polymorphism
  - evolution
- Also : cancer, speciation

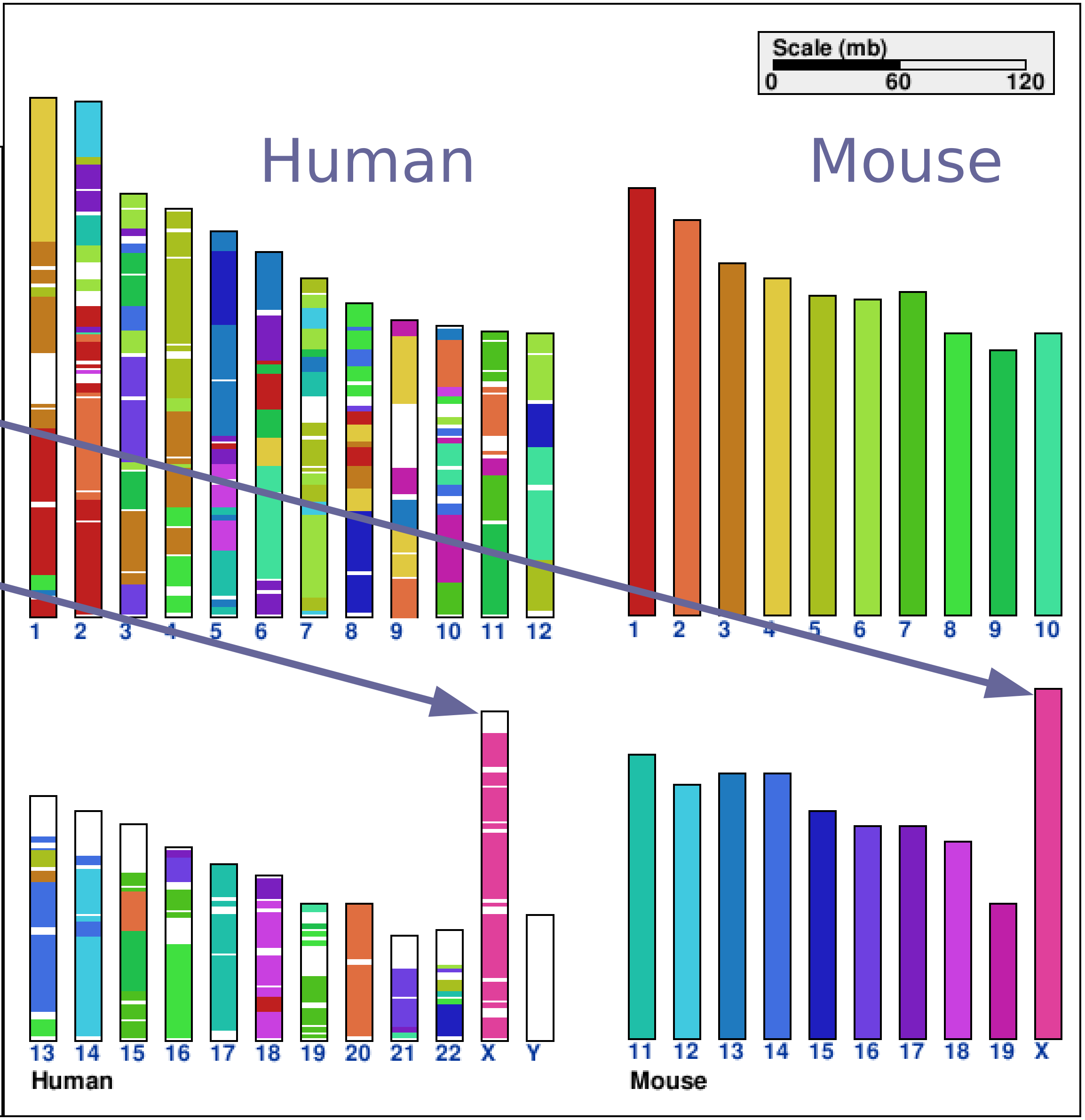
Genome evolution

- Structural differences between species:
  - in germ-line cell
  - inheritance
  - fixation in the population
- Examples in mammals:
  - number of chromosomes: $2n=6 \rightarrow 2n=102$
  - Human-chimpanzee:
    - 1.2 % sequence divergence
    - 9 large inversions and 1 fusion and many smaller rearrangements

Chromosomal evolution : example

- Human vs Mouse





http://cinteny.cchmc.org

Open questions in chromosomal evolution

- Diversity in rates and types of rearrangements in different lineages
- Localisation of rearrangements along the genomes:
  - Random Breakage Model

    size of conserved segments (Nadeau & Taylor, 1984)
  - Fragile Breakage Model

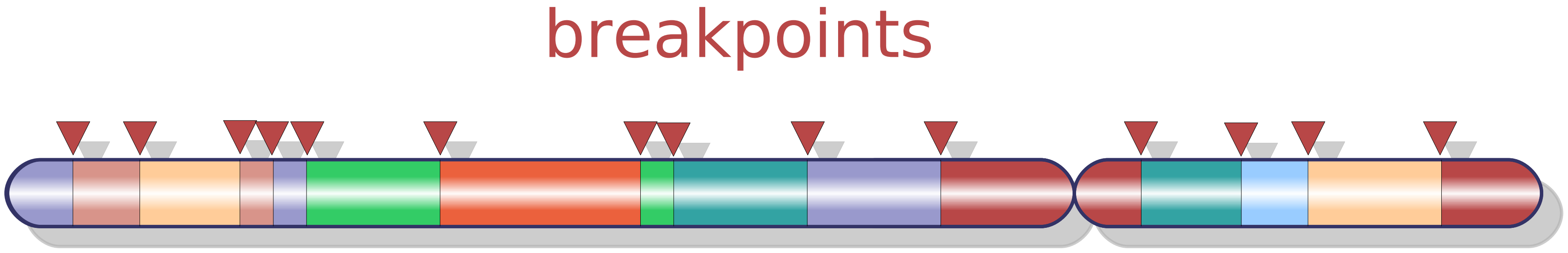
    more data thanks to whole genome sequencing

    many small segments

    re-use of breakpoints in different lineages

Pevzner, 2003
Kent, 2003
Murphy, 2005
...

Motivations



- Do the breakpoint sequences show some characteristics? Is it possible to characterise the breakpoint sequences?

  base composition, repeated elements, motifs...

- Is the breakpoint distribution along the genome linked to some genome organisation?

  isochores, gene distribution, recombination, replication, chromatin structure...

Strategy

1. Localising very precisely the breakpoints along one genome
2. Analysing:
   - breakpoint sequences
   - breakpoint distribution

Strategy

1. Localising very precisely the breakpoints along one genome
   - review on the computational methods to detect rearrangement breakpoints
   - whole genome sequences: precision expected
   - ... but disappointing

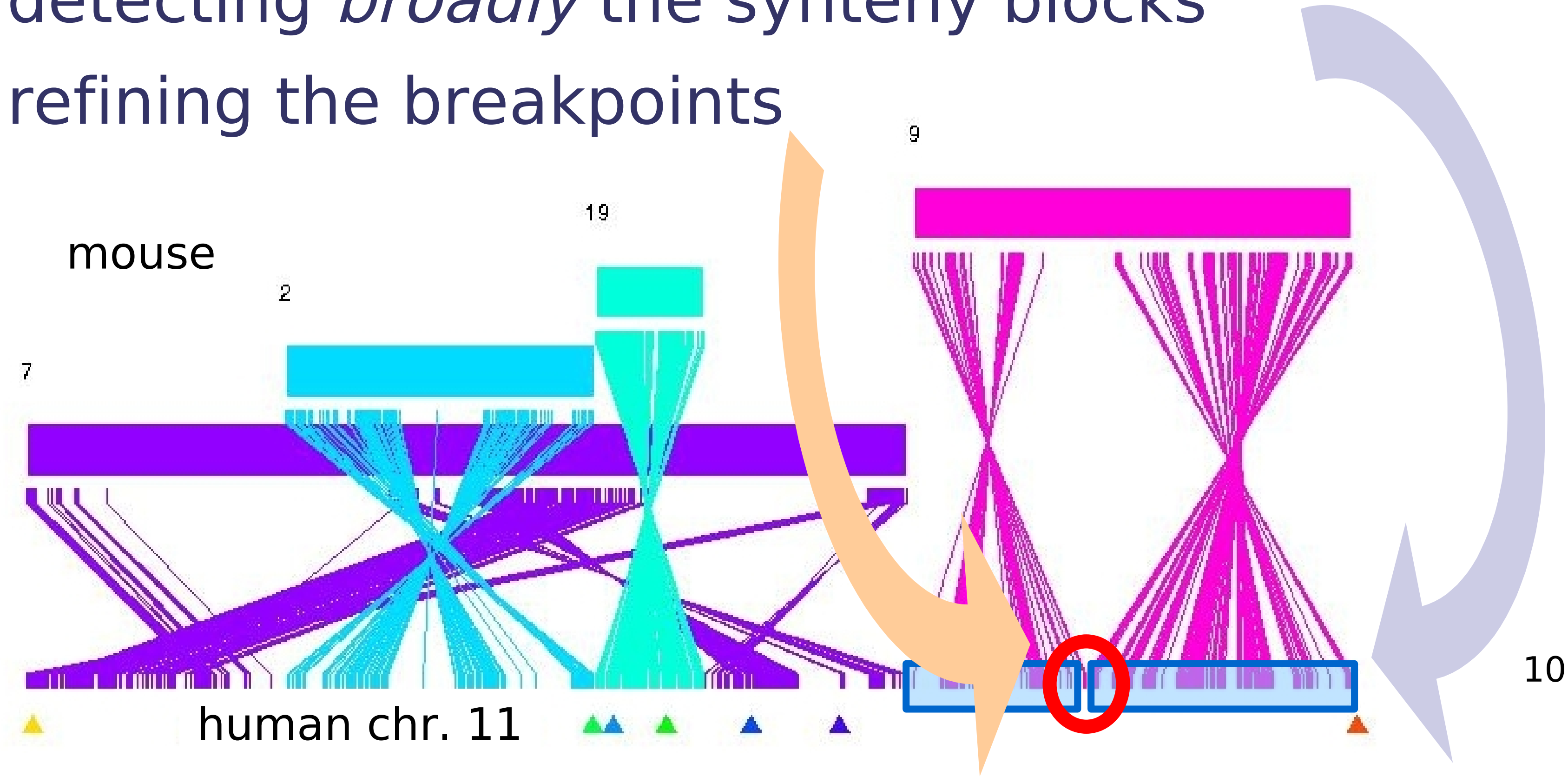
Lemaitre & Sagot. A small trip in the untranquil world of genomes. *TCS* 2008

Strategy

1. Localising very precisely the breakpoints on a genome:

development of a method in 2 steps

1. detecting *broadly* the synteny blocks
2. refining the breakpoints



Synteny blocks detection

- Def: orthologous regions between 2 genomes which have not been rearranged

  => conserved order and orientation of orthologous markers

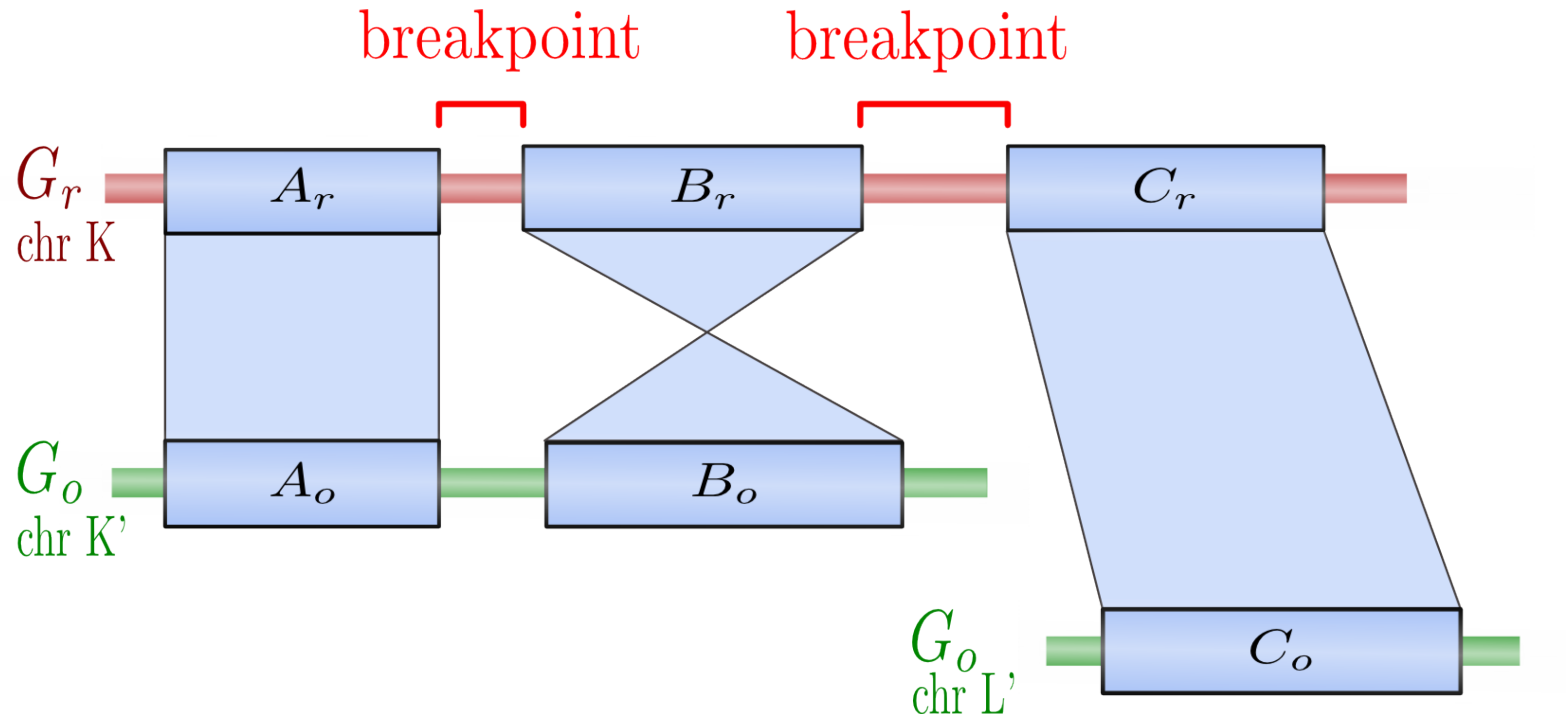
- Our contribution:
  - formal definition of synteny blocks
  - flexibility
  - blocks without conflicts (no overlap)
- markers = genes

Method to refine breakpoints

- INPUT:
  - the synteny blocks between 2 genomes G_r and G_o
  - the sequences of genomes G_r and G_o
- OUTPUT:
  - the breakpoints regions on G_r

Breakpoint refinement

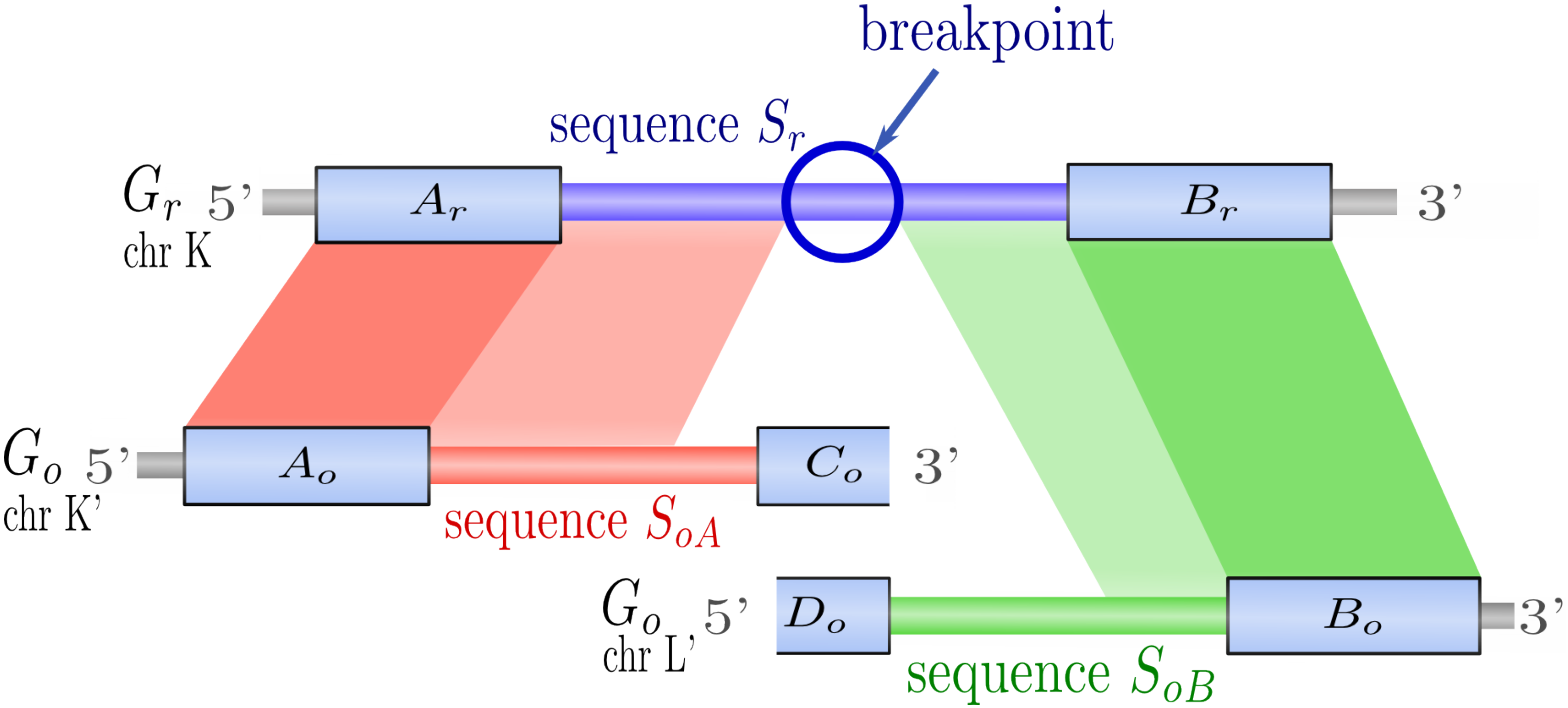
- The breakpoint = between 2 consecutive synteny blocks on G_r, rearranged on G_o



- Asymmetry + origin of the breakage event

Alignments

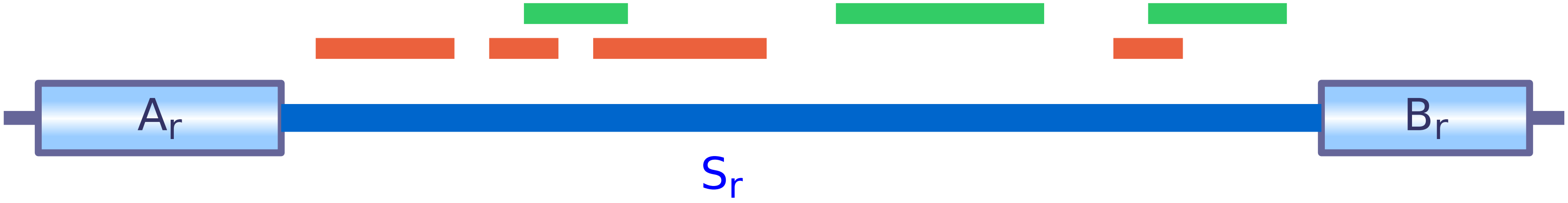
- Alignment of the *inter-block* sequences



alignment: local, sensitive and fast -> BlastZ

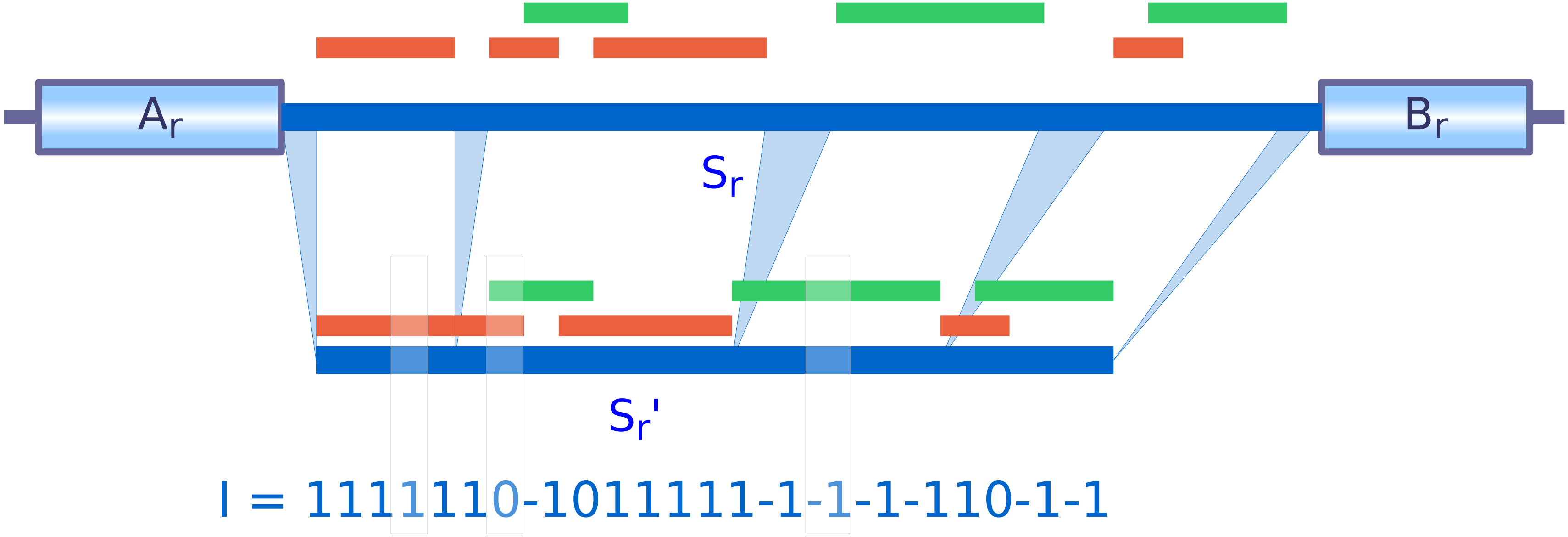
Alignments (2)

- 2 lists of hits:
  - 🟥 hits between S_r and S_{oA}
  - 🟩 hits between S_r and S_{oB}
- The hits are mapped on sequence S_r



Segmentation

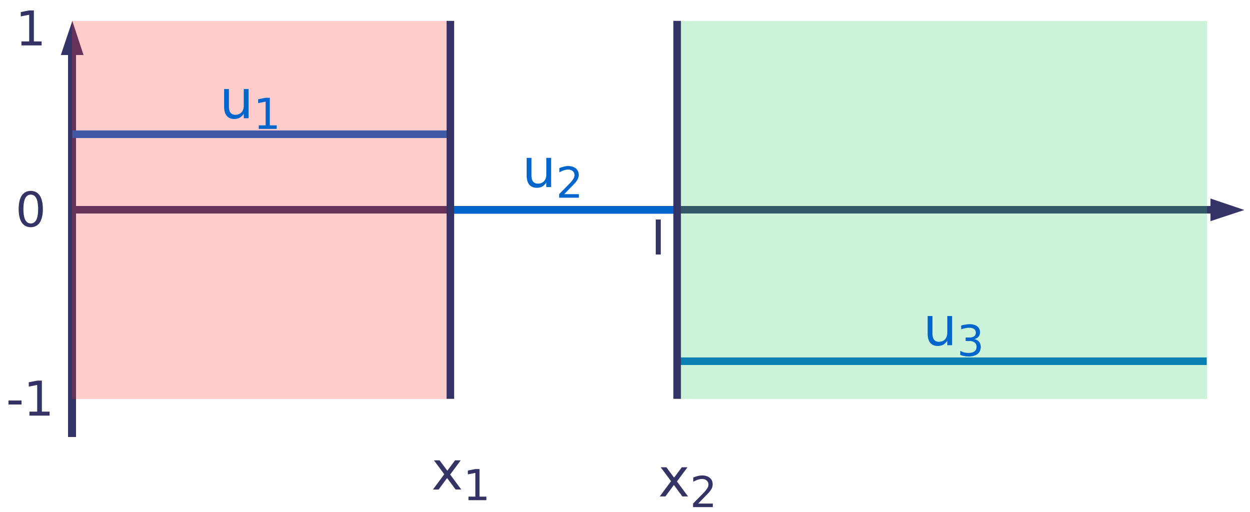
- Coding the hits information:
  - only the positions covered by hits
  - numerical sequence I



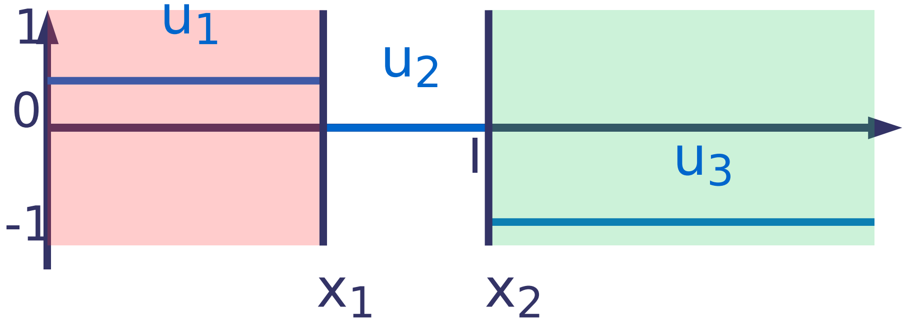
I = 1111110-1011111-1-1-1-110-1-1

Segmentation (2)

- looking for 3 segments:
  - segment 1: homology with S_{oA}
  - segment 2: breakpoint
  - segment 3: homology with S_{oB}



Segmentation (3)



- 3 segments:
  - segment 1 : $u_1 = \begin{cases} \text{mean } (I[1..x_1]) \text{ if } > 0 \\ + \infty \text{ otherwise} \end{cases}$
  - segment 2 : $u_2 = 0$
  - segment 3 : $u_3 = \begin{cases} \text{mean } (I[x_2+1..n]) \text{ if } < 0 \\ + \infty \text{ otherwise} \end{cases}$
- Find x_1 et x_2 such that:

$$\min \quad f(x_1, x_2) = \sum_{j=1}^{3} \sum_{k=x_{j-1}+1}^{x_j} (I[k] - u_j)^2$$

(with $x_0=0$ and $x_3=n$)

Segmentation - algorithm

- Classical algorithm:

  dynamic programming => $O(n^2)$

- Speed-up:

  two independent minimisations => O(n)

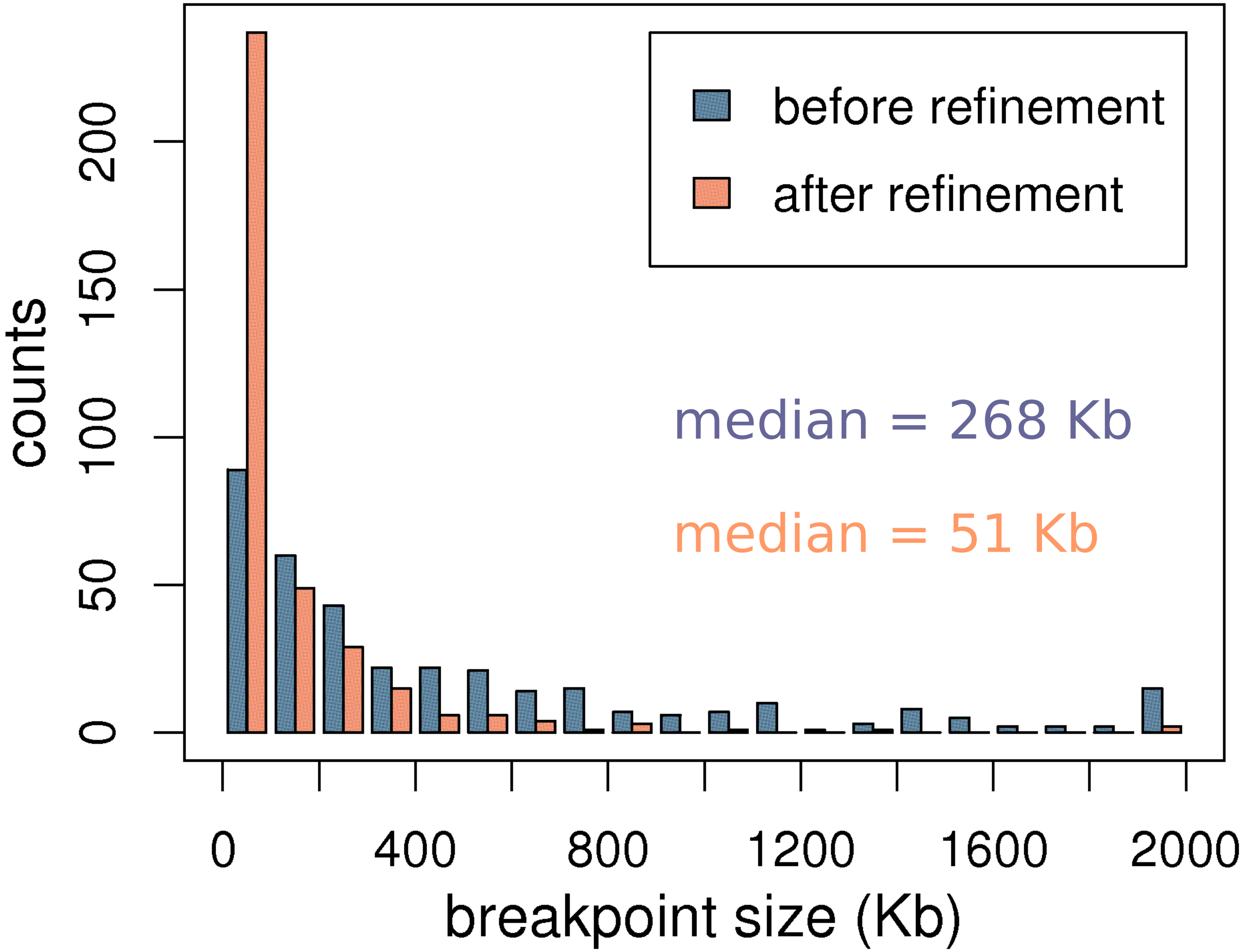
- Evaluation:

  estimation of a p-value

  random sequences (I) by shuffling the hits

Results

- Comparison human-mouse
- 354 breakpoints
- breakpoints reduced on average by 536 Kb
- 171 breakpoints < 50 Kb



Comparisons with other methods

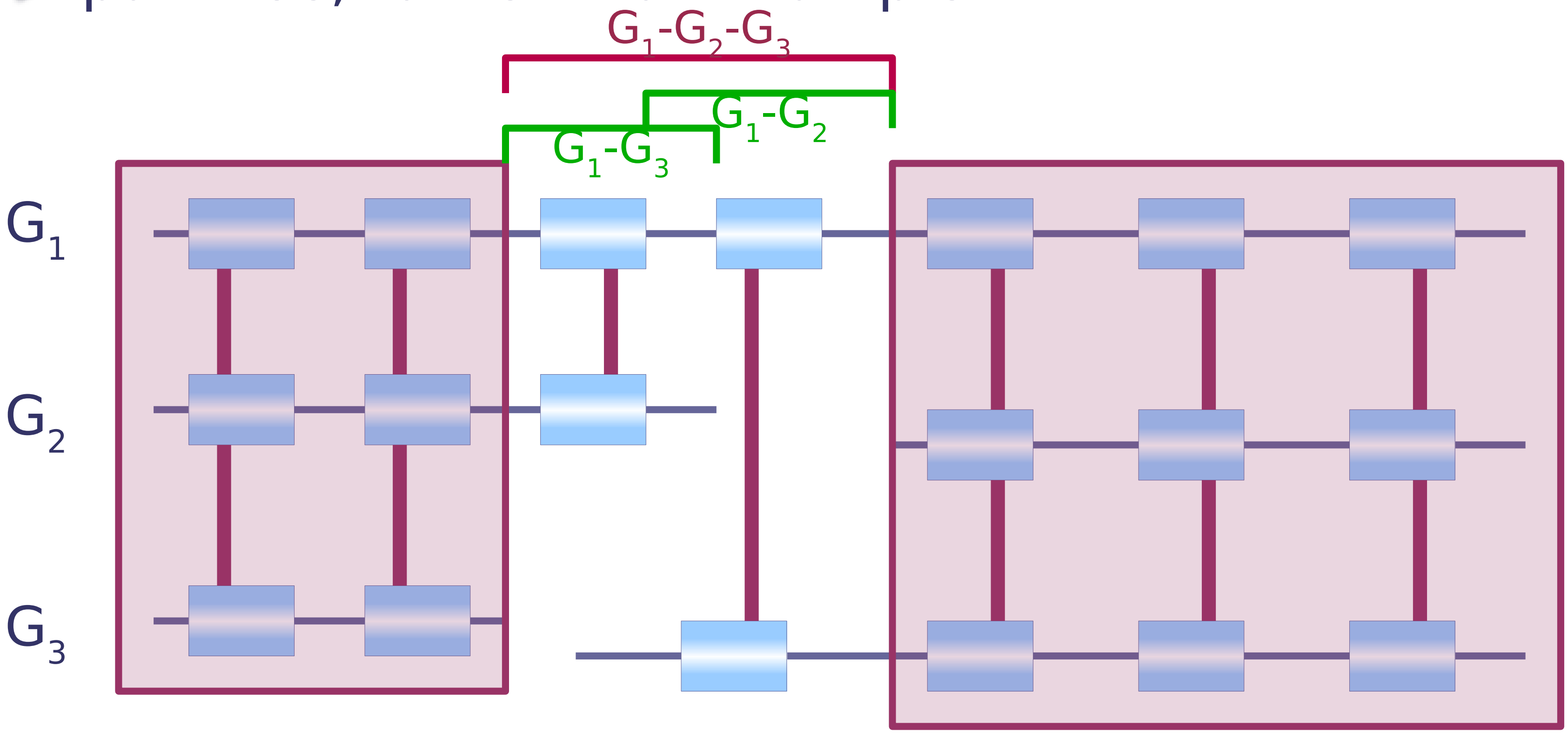
- Whole genome alignments:
  - pairwise and multiple
  - human-mouse

	Breakpoint size (Kb)	
	median	mean
Refined	51	129
Grimm2 (Pevzner & Tesler, 2003)	156	364
Grimm3 (Bourque, 2004)	268	454
Ensembl (Hubbard, 2007)	95	223

Lemaitre *et al.* 2008. *BMC Bioinformatics*

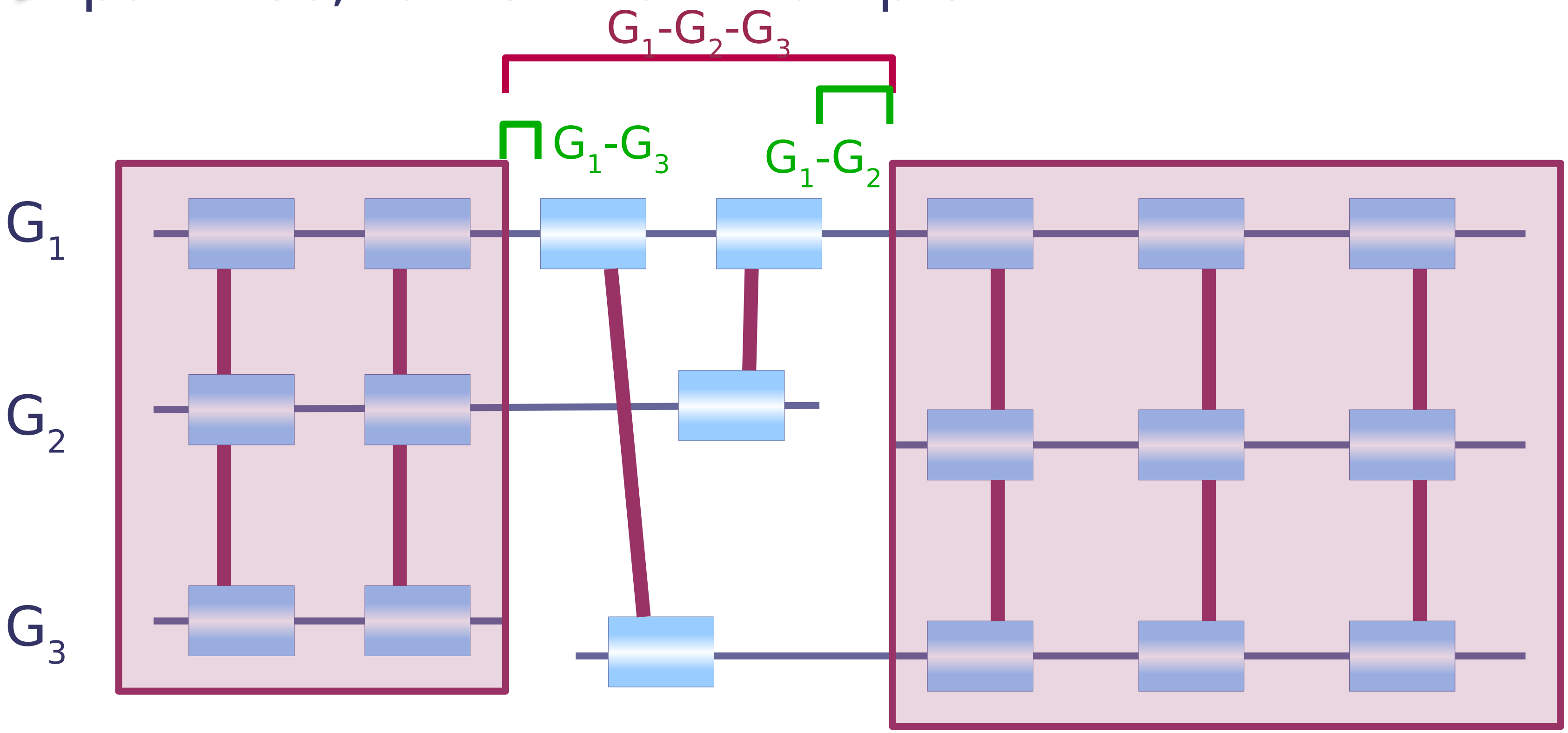
Discussion

- Better precision because:
  - limitation of the search space (2 steps), rather than whole genome alignments
  - pairwise, rather than multiple



Discussion

- Better precision because:
  - limitation of the search space (2 steps), rather than whole genome alignments
  - pairwise, rather than multiple



Characterising the breakpoints

- In a systematic way:
  - mammalian breakpoints:

    human vs mouse, rat, dog, macaque, chimpanzee
  - compared to:
    - artificial breakpoints
    - flanking sequences
    - randomised points

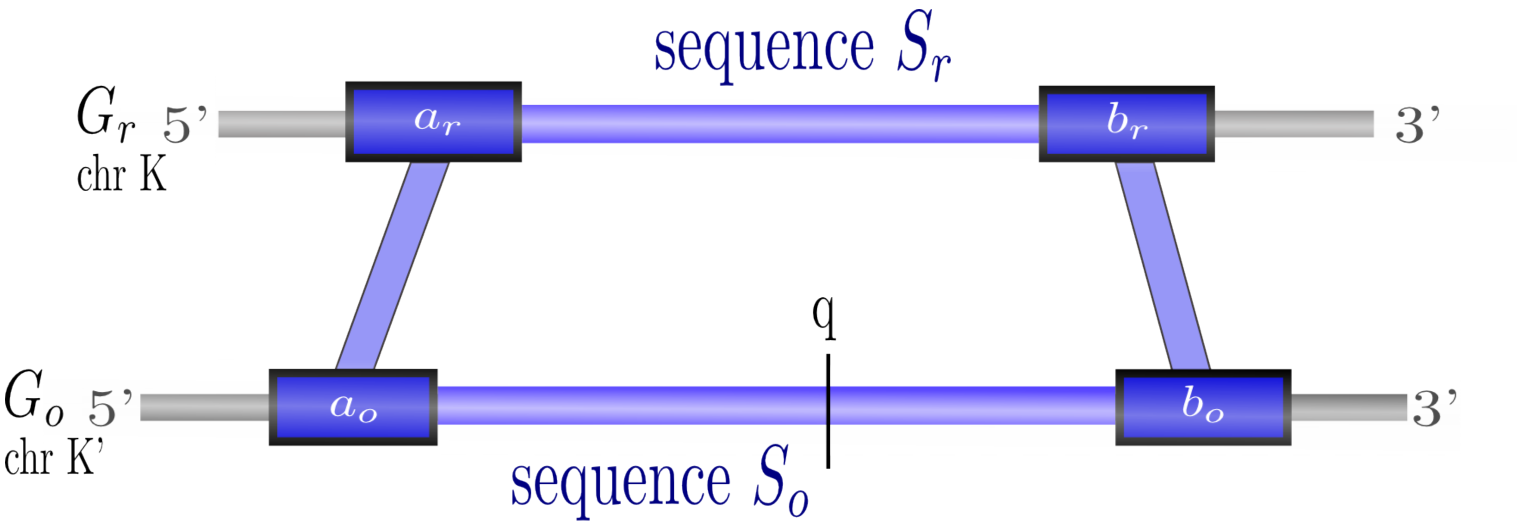
segmentation properties

sequence characteristics

distribution

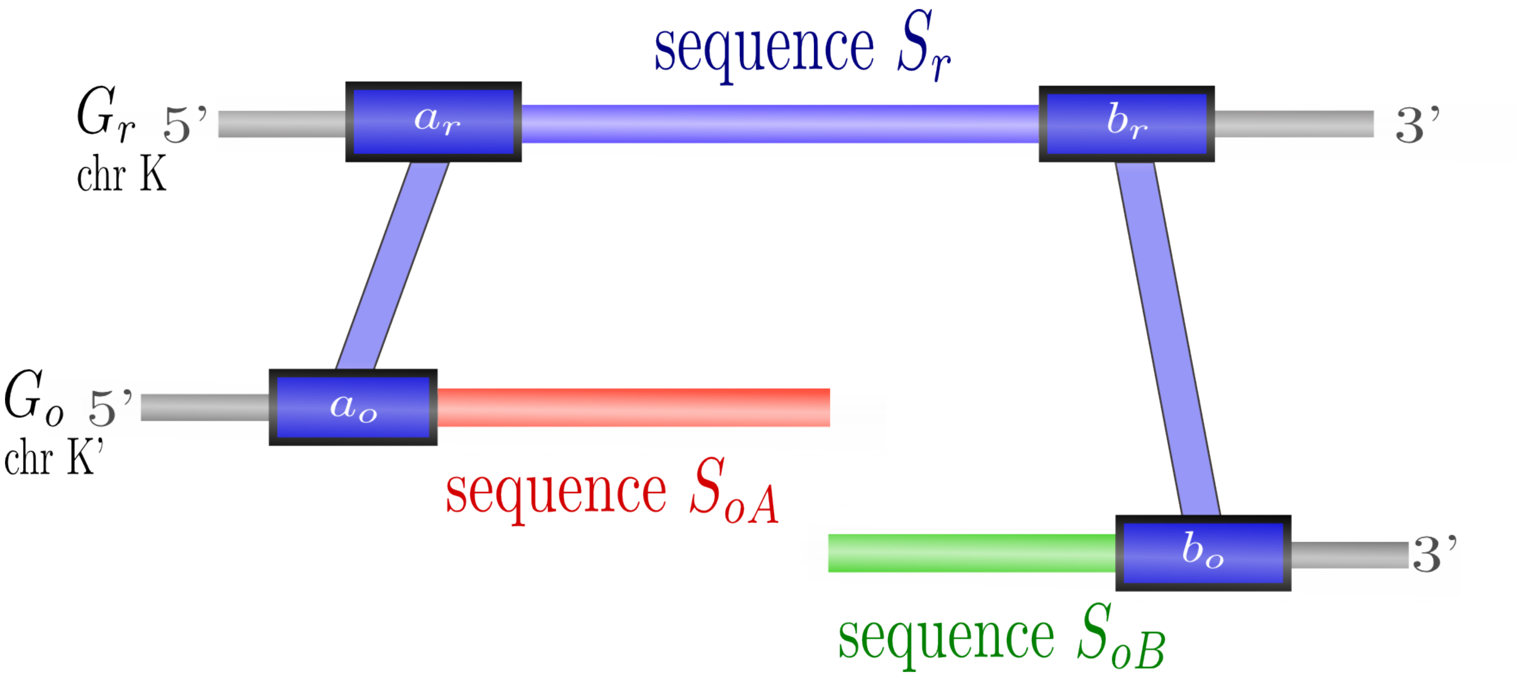
Artificial breakpoints

- Artificial breakpoints



Artificial breakpoints or a null model of breakage

- Artificial breakpoints



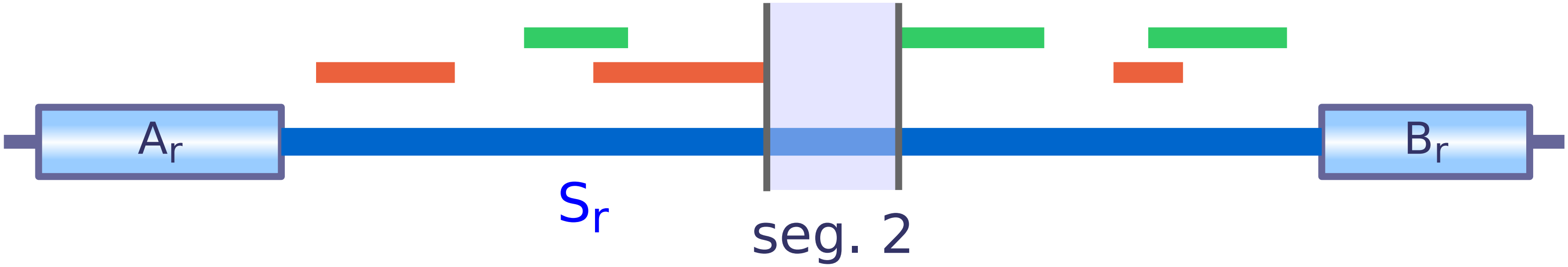
- model = breakage + sequence evolution as if no rearrangement

Segmentation differences

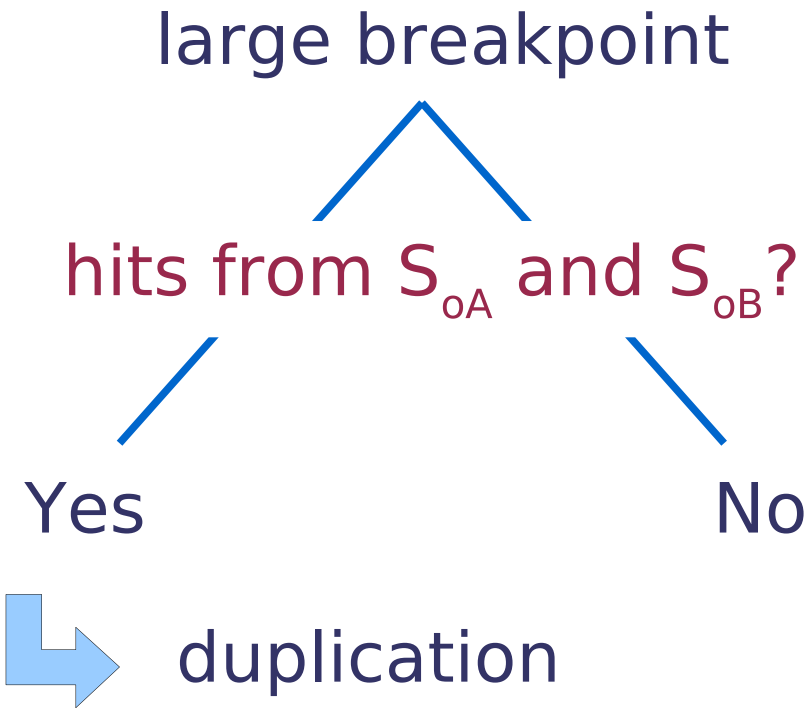
- Size of the « breakpoint » : point or region ?
- Similarity of the adjacent sequences (seg. 1 & 3)

	Size of segment 2	Hits coverage over segments 1 and 3
Breakpoints	128 Kb	50 %
Artificial points	7 Kb	59 %

human-mouse



Investigating large breakpoints



large breakpoint

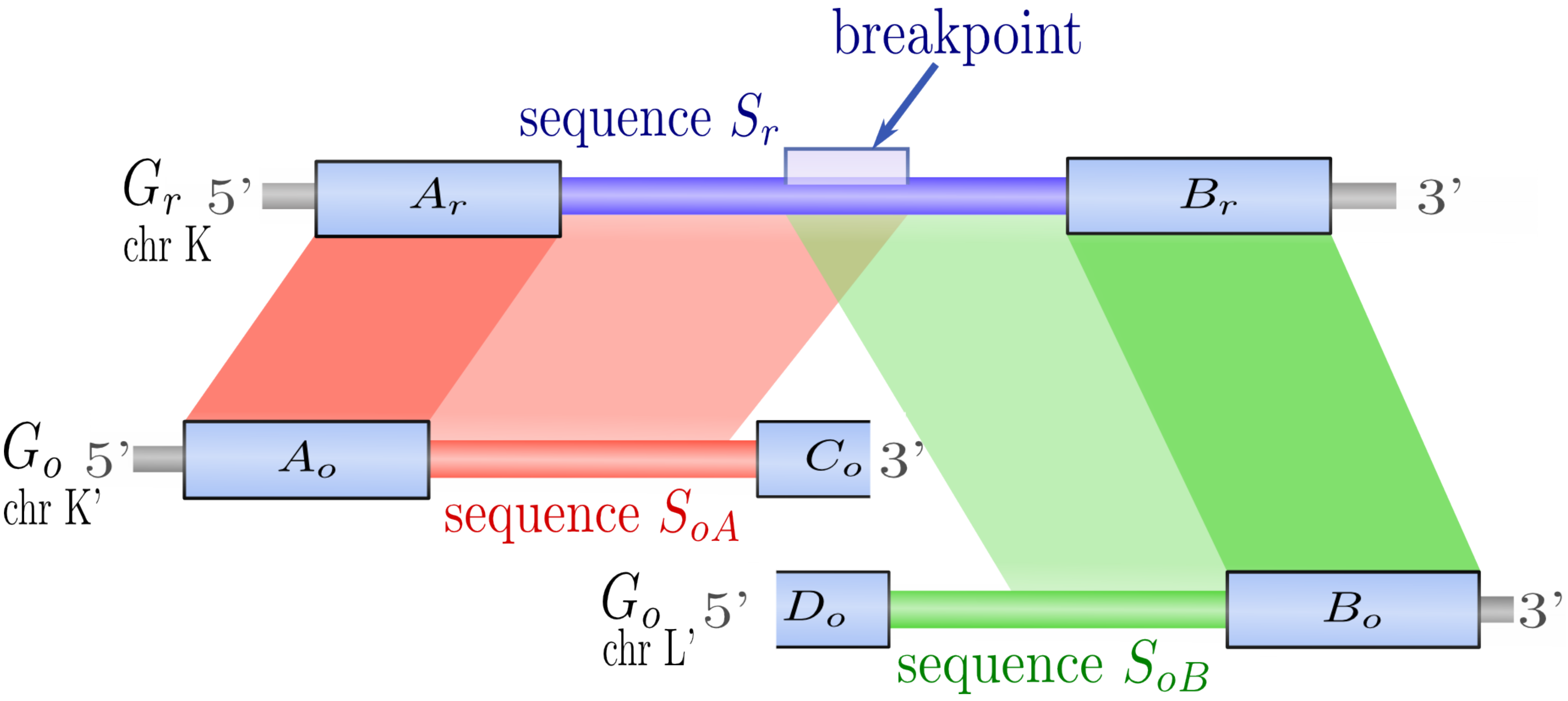
hits from S_{oA} and S_{oB}?

Yes

No

duplication

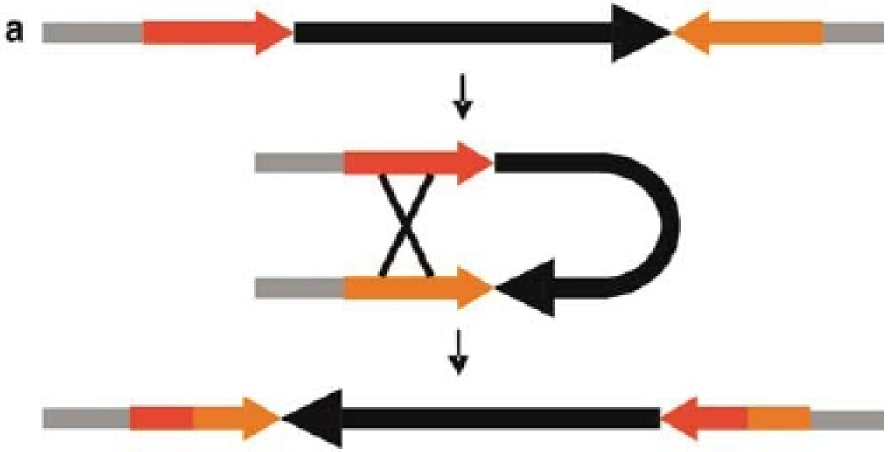
Duplication at the breakpoint



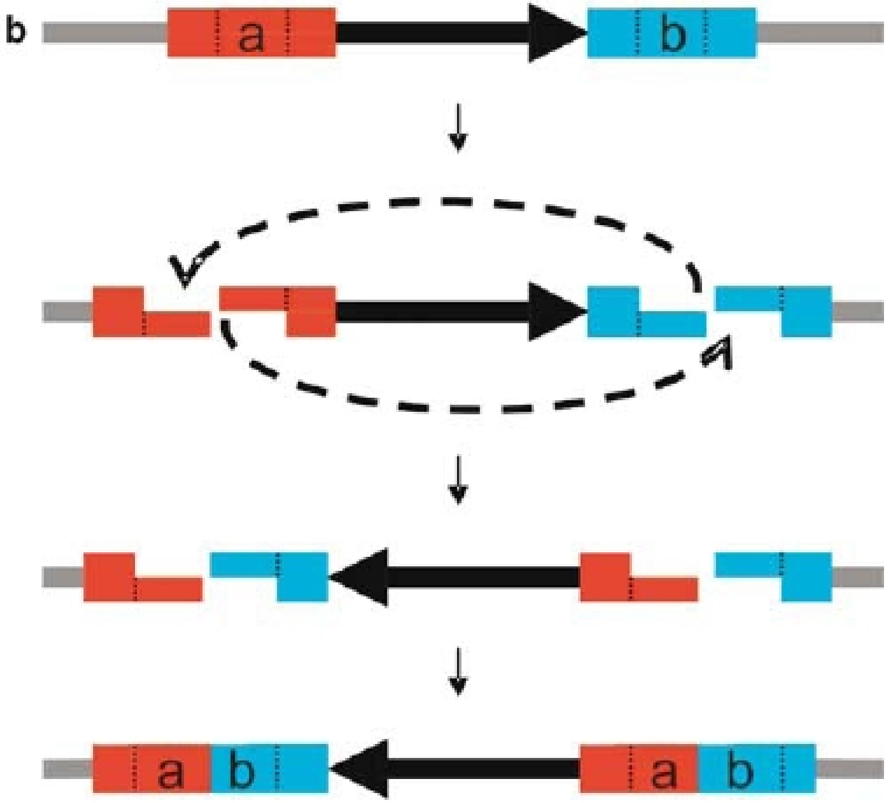
- explaining the size of some breakpoints
- related to the molecular mechanisms of rearrangement

Duplication at the breakpoint (2)

2 mechanisms of rearrangements involving duplications



ectopic recombination



staggered breaks

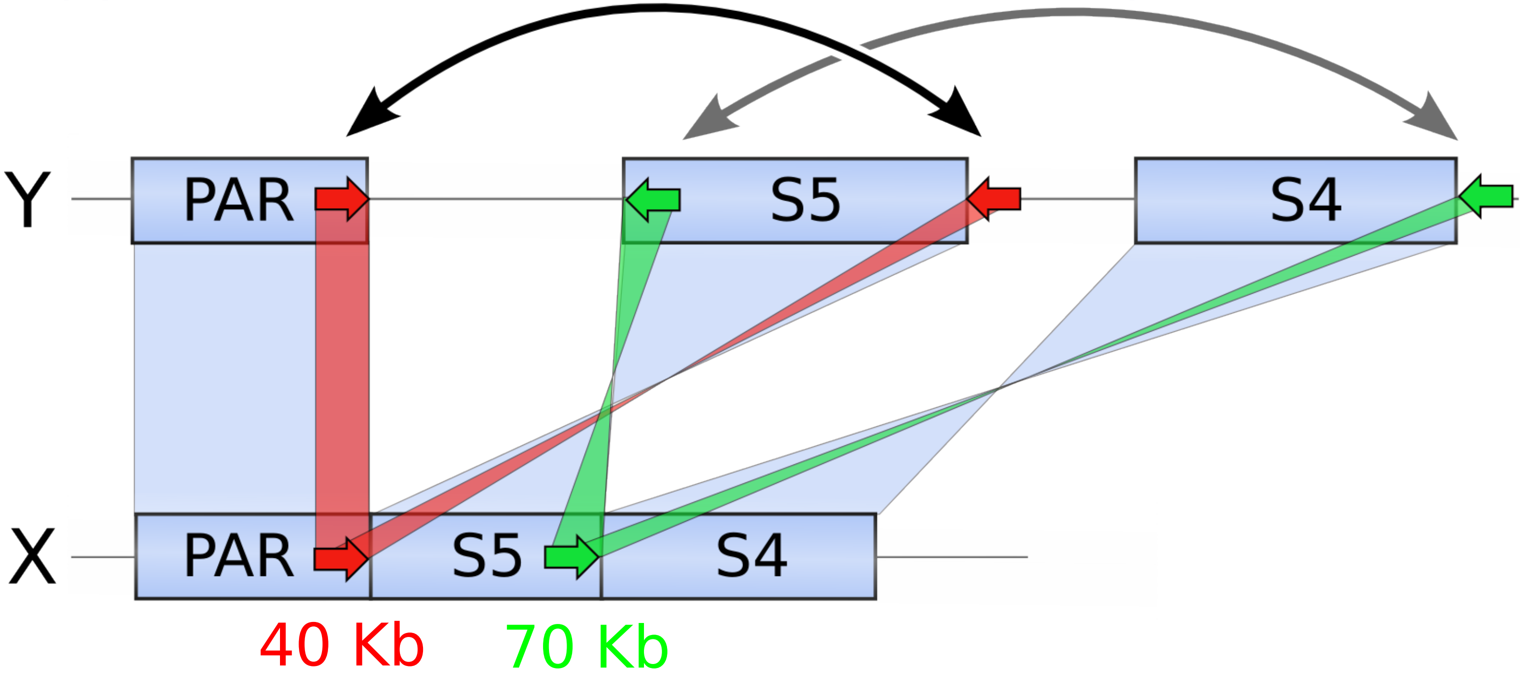
Casals, 2007

Duplication at the breakpoint (3)

- Application to the human X-Y comparison:

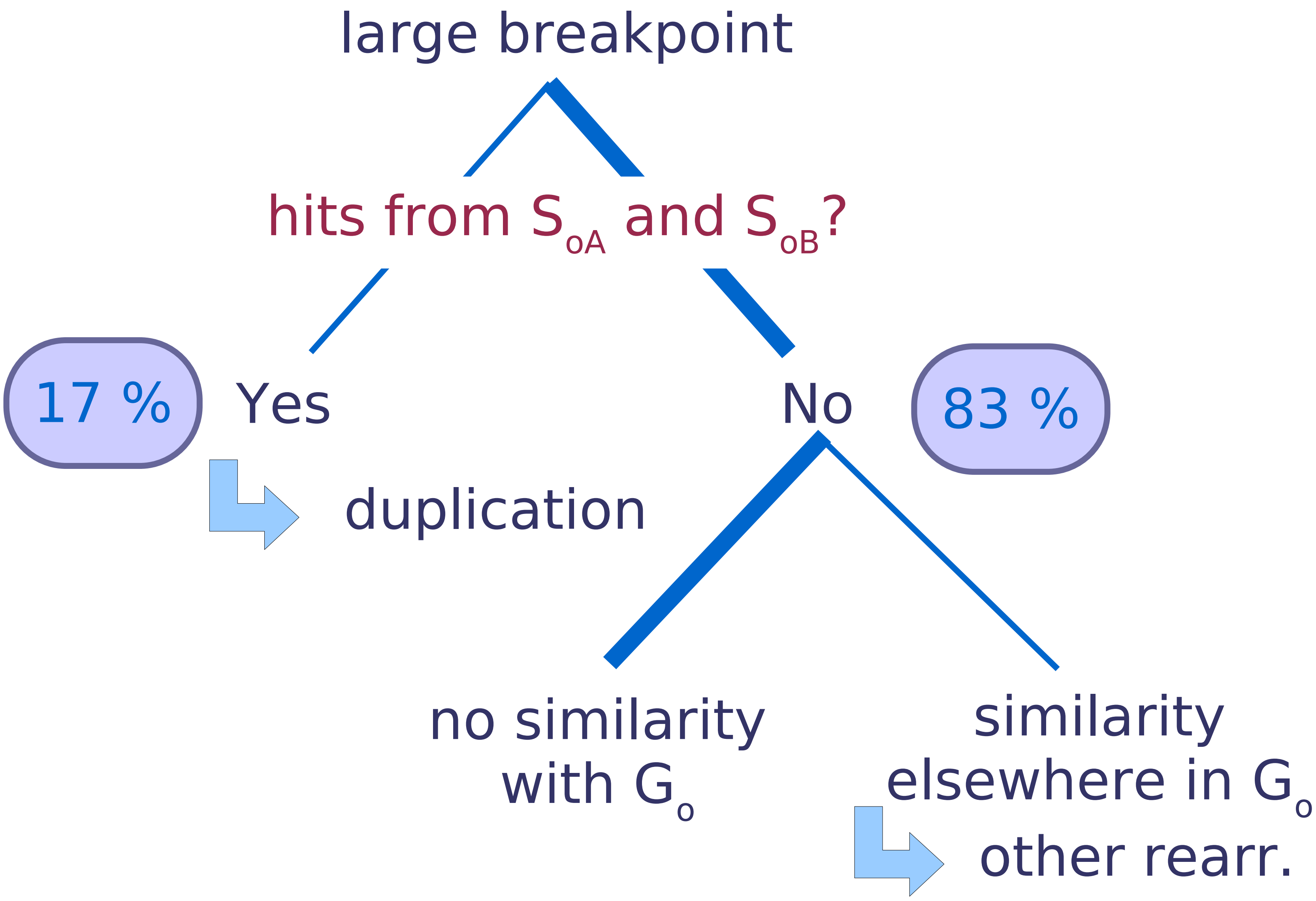
  identification of 2 duplications = footprints of 2 inversions + temporal ordering

  hypothesis of sex differenciation by inversions



Lemaitre *et al.* (M. Braga, G. Marais). 2008. sub. to *Mol Biol Evol.*

Investigating large breakpoints



large breakpoint

hits from S_{oA} and S_{oB}?

17 % Yes

No 83 %

duplication

no similarity with G_o

similarity elsewhere in G_o

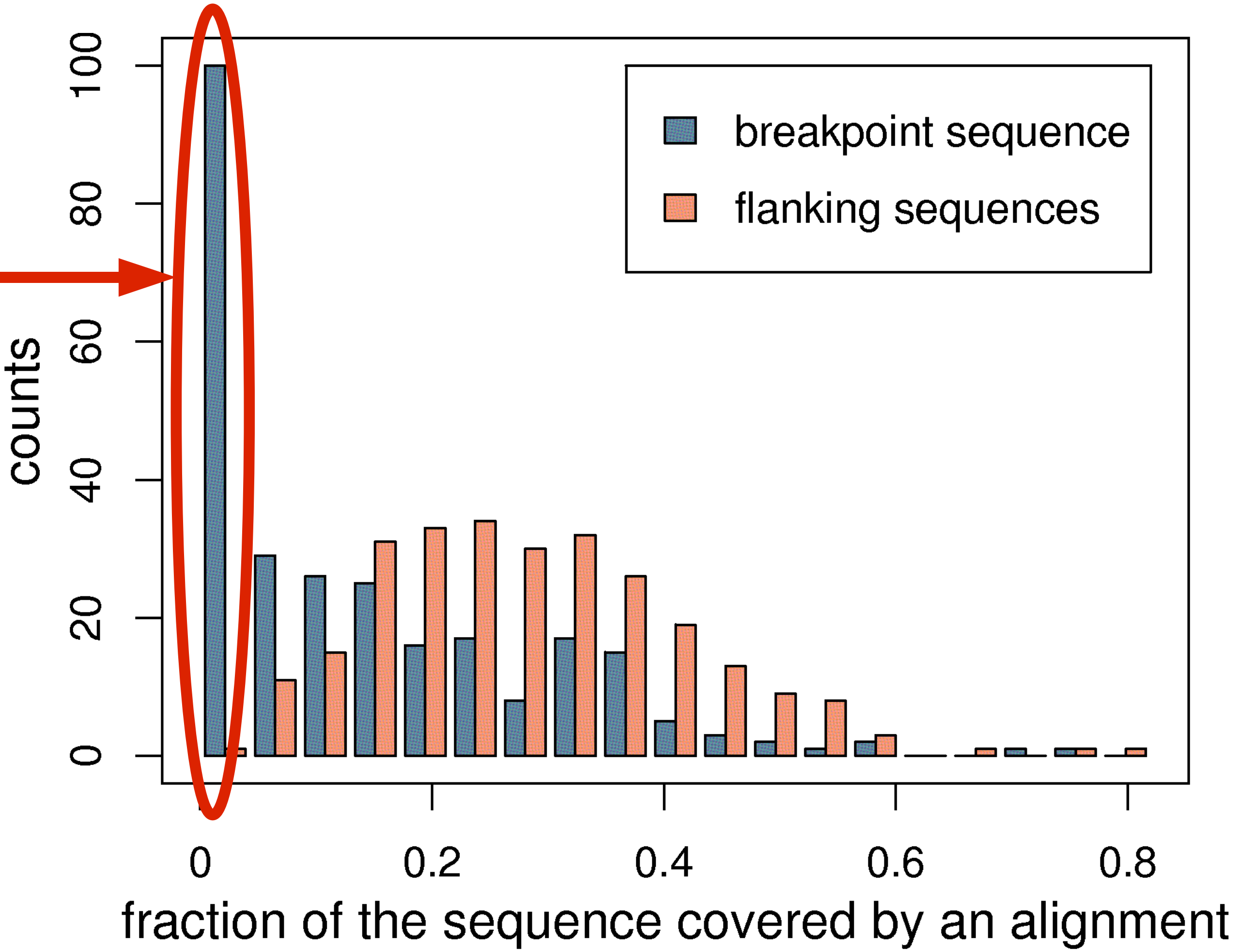
other rearr.

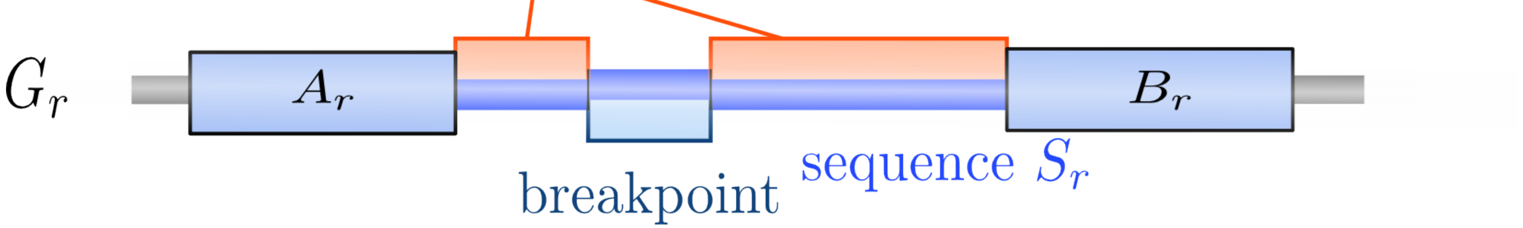
Similarity elsewhere

UCSC whole genome alignments

no similarity:

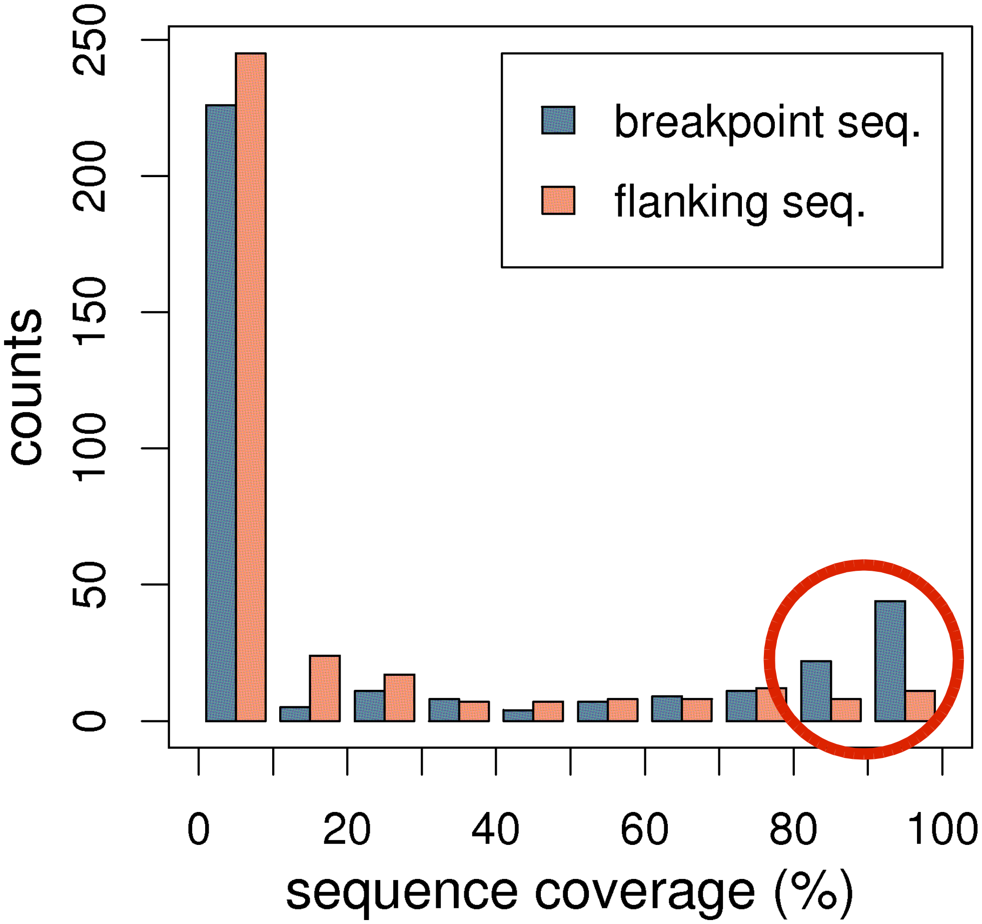
- rapid divergence
- insertion of new sequence in human
- deletion in mouse



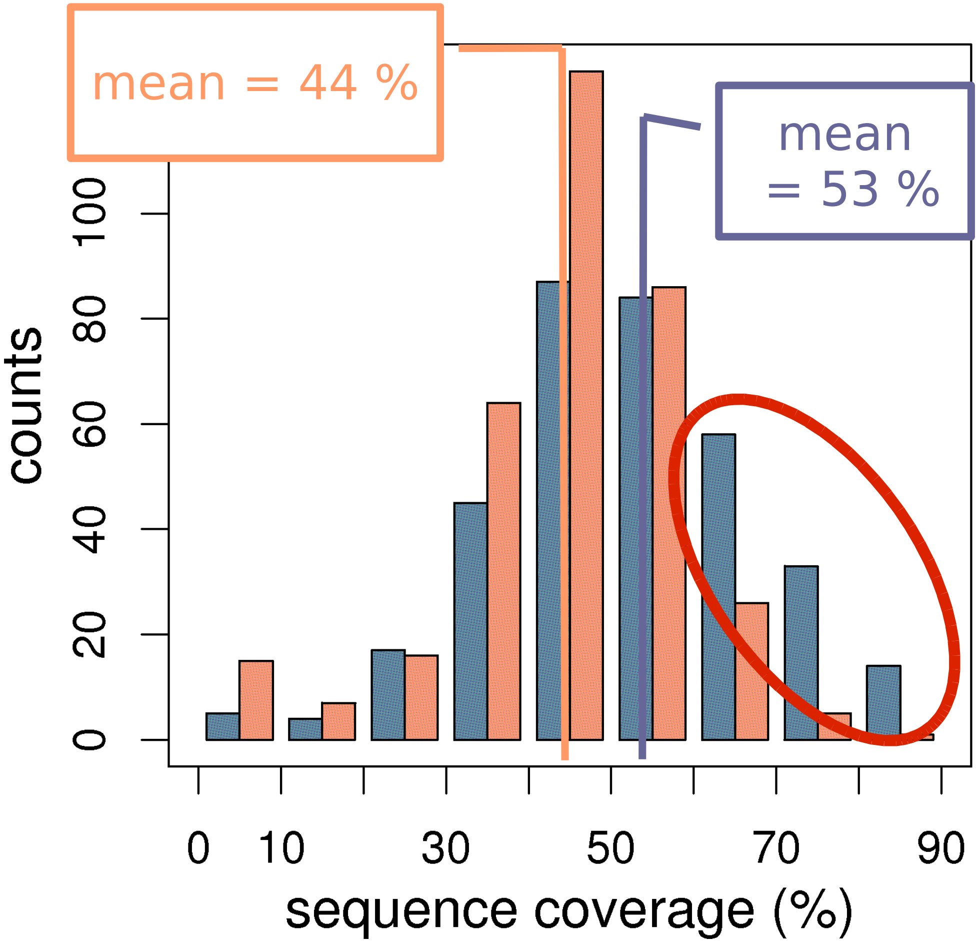


Other characteristics

Human segmental duplications



Transposable elements



mean = 44 %

mean = 53 %

Breakpoints features

- Results :
  - Loss of similarity inside and outside breakpoints
  - Duplications and repeated elements

- Complexity of breakpoints :
  - not only punctual breakage
    - sequence evolution more complex « after » the rearrangement
    - or: sequence properties « before » the rearrangement

Distribution of breakpoints along the human genome

- Are the breakpoints distributed uniformly and independently along the genome?
  - Random Breakage Model
- Are there some « forbidden » regions?
  - negative selection preventing breakage inside genes and functional regions
  - Intergenic Breakage Model
- Are there some « fragile » regions?
  - neutral model, regions more prone to breakage
  - Hotspots or Fragile Breakage Model

Breakpoint data

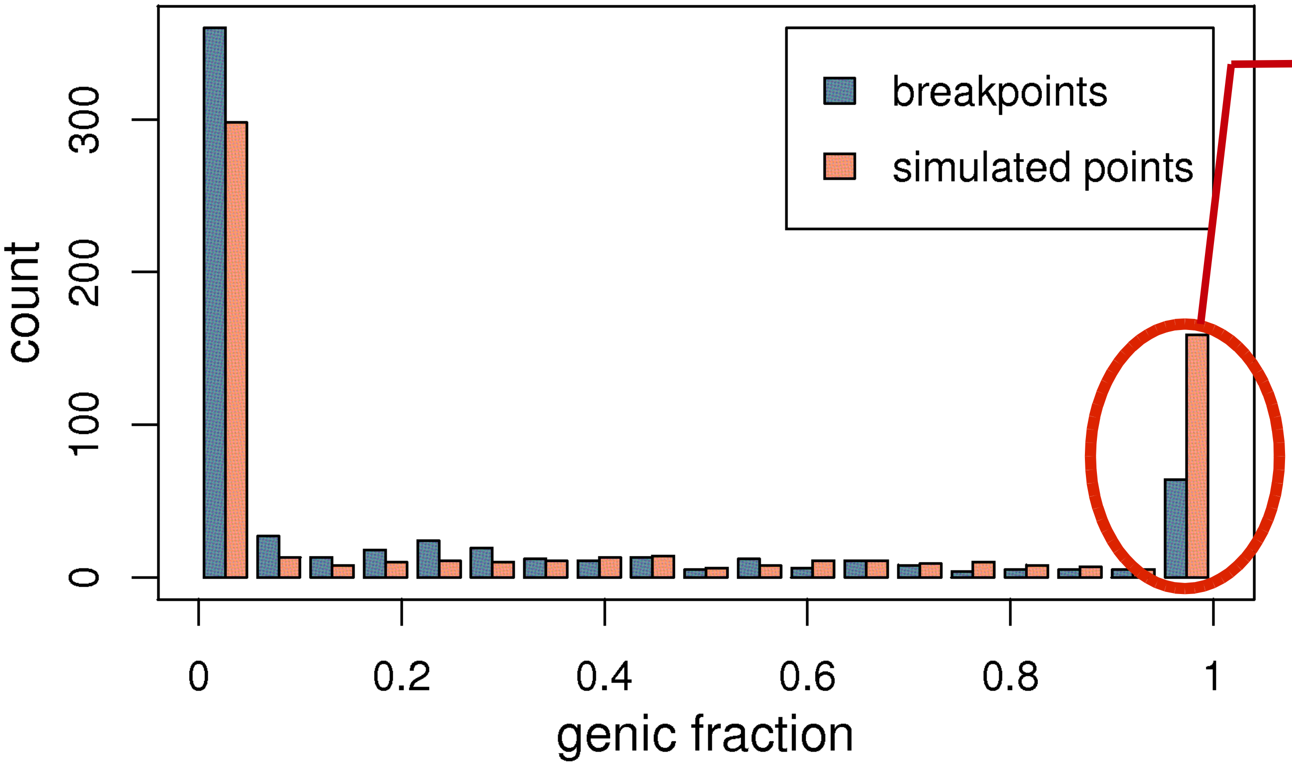
- Mammalian breakpoints:
  - 5 pairwise comparisons Human – X

    X= mouse, rat, dog, macaque, chimpanzee
  - 622 breakpoints mapped on the human genome
  - median size of 26.6 Kb

- Simulations: random breakage model:
  - 1000 data sets: 622 breakpoints uniformly redistributed on the human genome (same size, without overlap)

Breakpoints and genes

- Comparison of the genic fraction of breakpoints



57 breakpoints fully genic

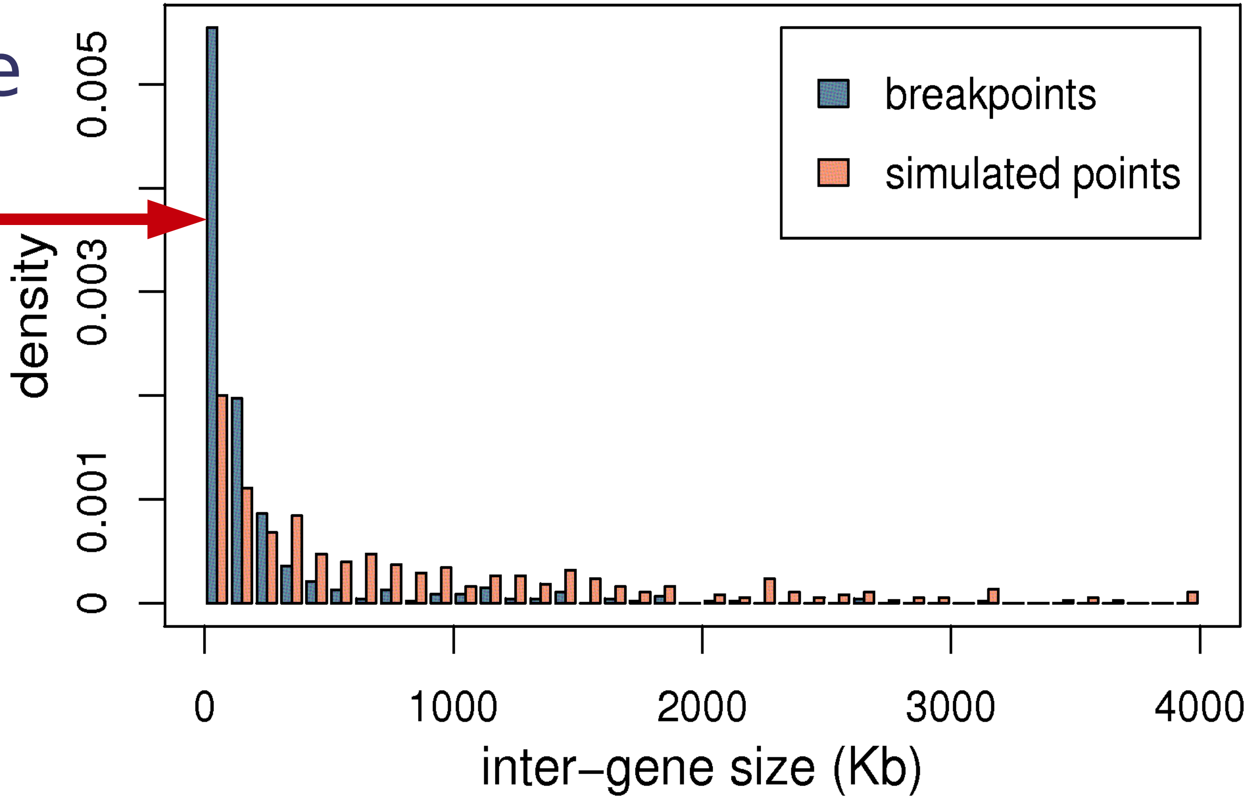
no simulation with ≤ 57
p-value < 0.001

- under-representation of breakpoints inside genes ⟹ RBM + negative selection

Breakpoints and genes (2)

- Intergenic breakpoints
- Inter-gene size

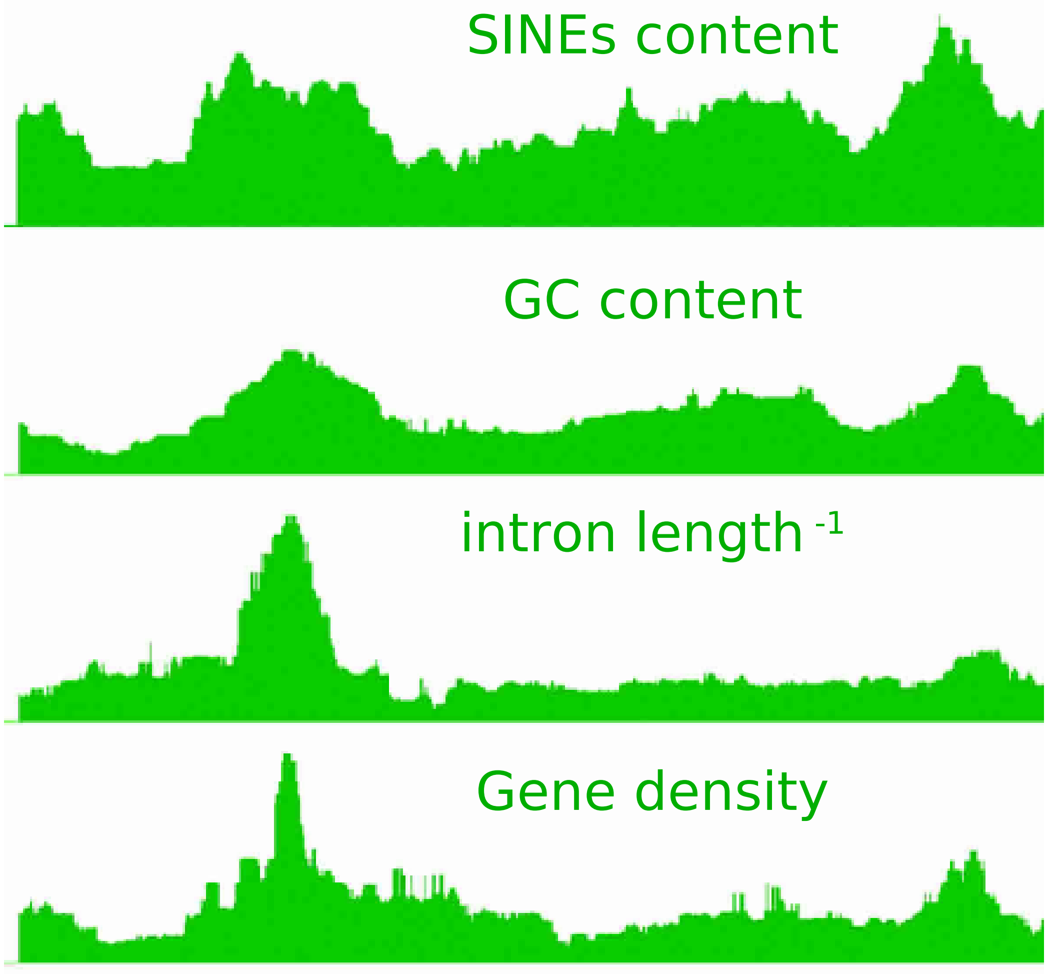
breakpoints are located in smaller inter-genes



- Can not be explained by random breakage + natural selection

The isochore organisation

- Genomic landscape
  - isochores : homogeneous GC content
  - correlations with other genomic features



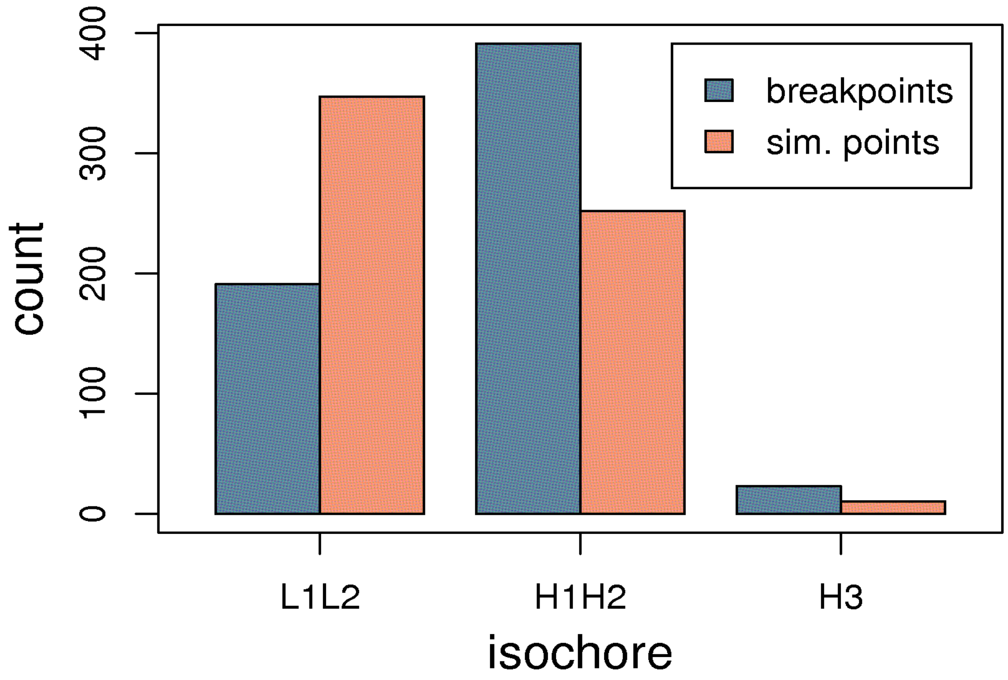
Part of human chromosome 9

Versteeg *et al.* 2003

Other correlations

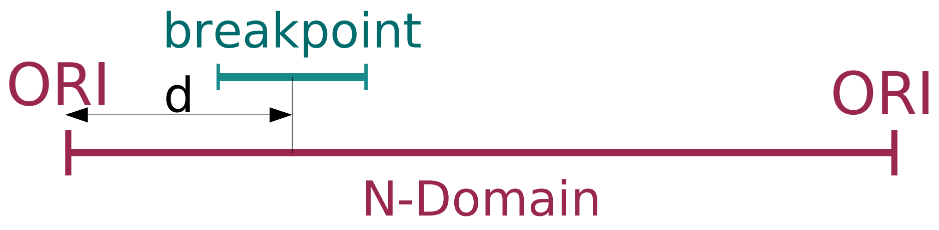
	breakpoints	sim. points	test (p-value)
GC content	44 %	41 %	e-12
Gene density (#/Mb)	14.9	8.3	<2e-16
SINEs	19.2 %	12.6 %	e-13

⇨ Isochores

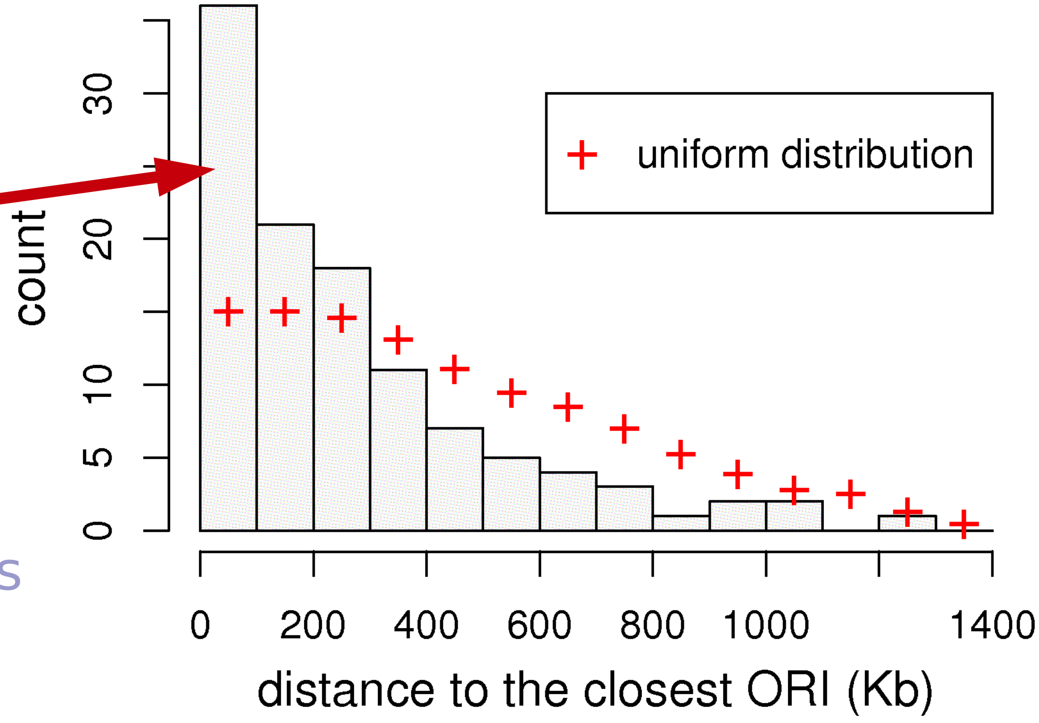


Replication origins

- Detection *in silico,* based on GC + AT skew profiles
  - 578 N-domains
  - 1060 putative origins
  - ~20% of the genome



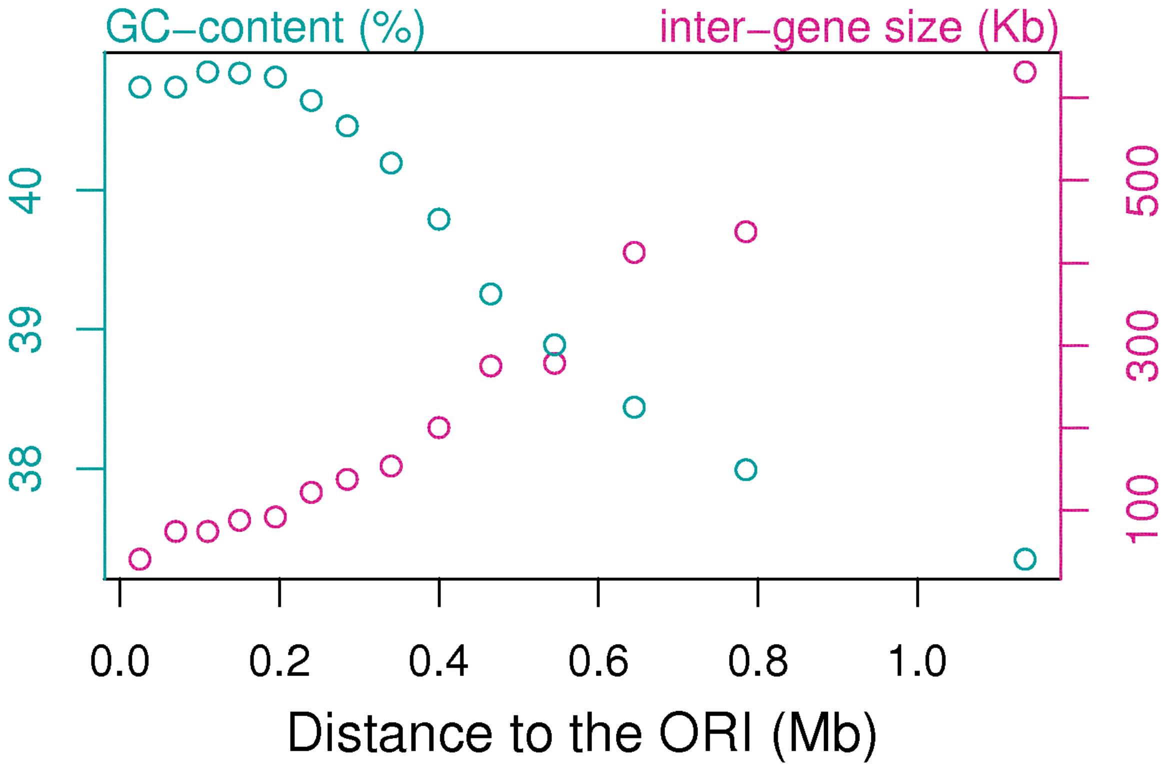
breakpoints are over-represented close to the ORIs



Collaboration with A. Arneodo's team at ENS of Lyon
Huvet *et al.*, 2007.

Replication origins (2)

- ORIs contain small intergenes and are richer in G+C



A new model

- Breakpoints are over-represented in regions with :
  - high transcriptional activity,
  - replication initiation,
- Open chromatin hypothesis:

these regions are « open » and thus more susceptible to breakage

- Model:

neutral mutational bias + natural selection in genes

Lemaitre *et al.* (Arneodo's team) 2008. sub. to *Genome Res.*

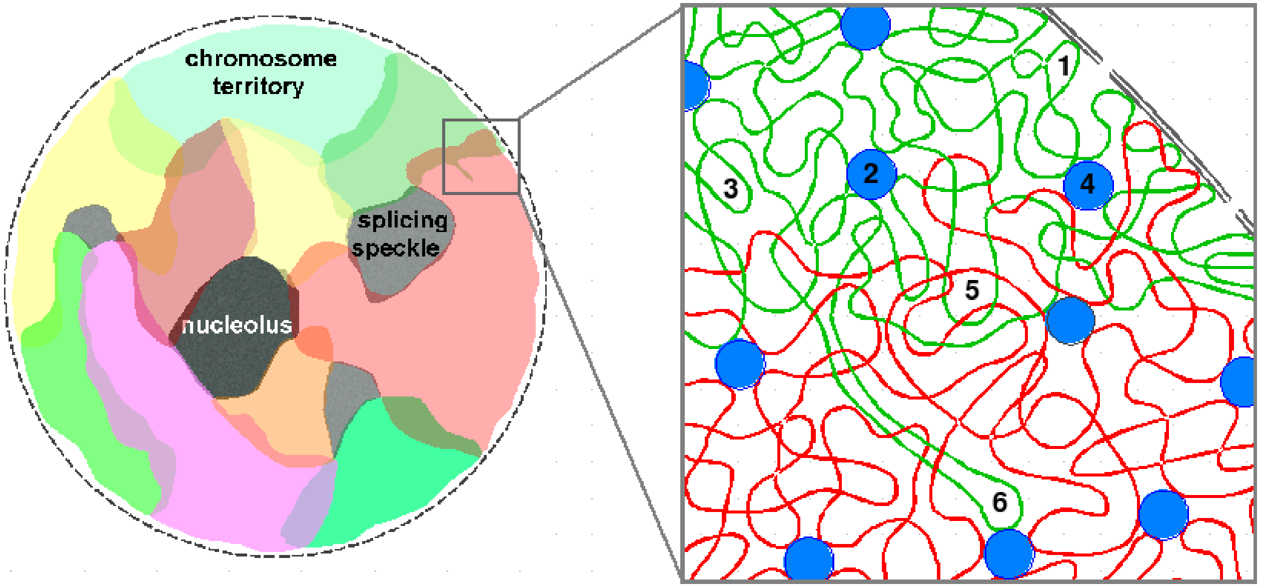
Conclusion and future work

- A method allowing to analyse precisely breakpoint structure and distribution
  - automatically detection of duplications
  - analysing the similarity decrease around the breakpoint
- Characterisation of breakpoints: duplications, loss of similarity, motifs…
  - comparing different types of rearrangements
  - take into account the evolutionary origin

To continue...

- A new model of breakpoint localisation along the human genome
  - expression and chromatin data
  - investigating cases of breakage inside genes
- 1D → 3D

spatial genome
organisation
inside the nucleus

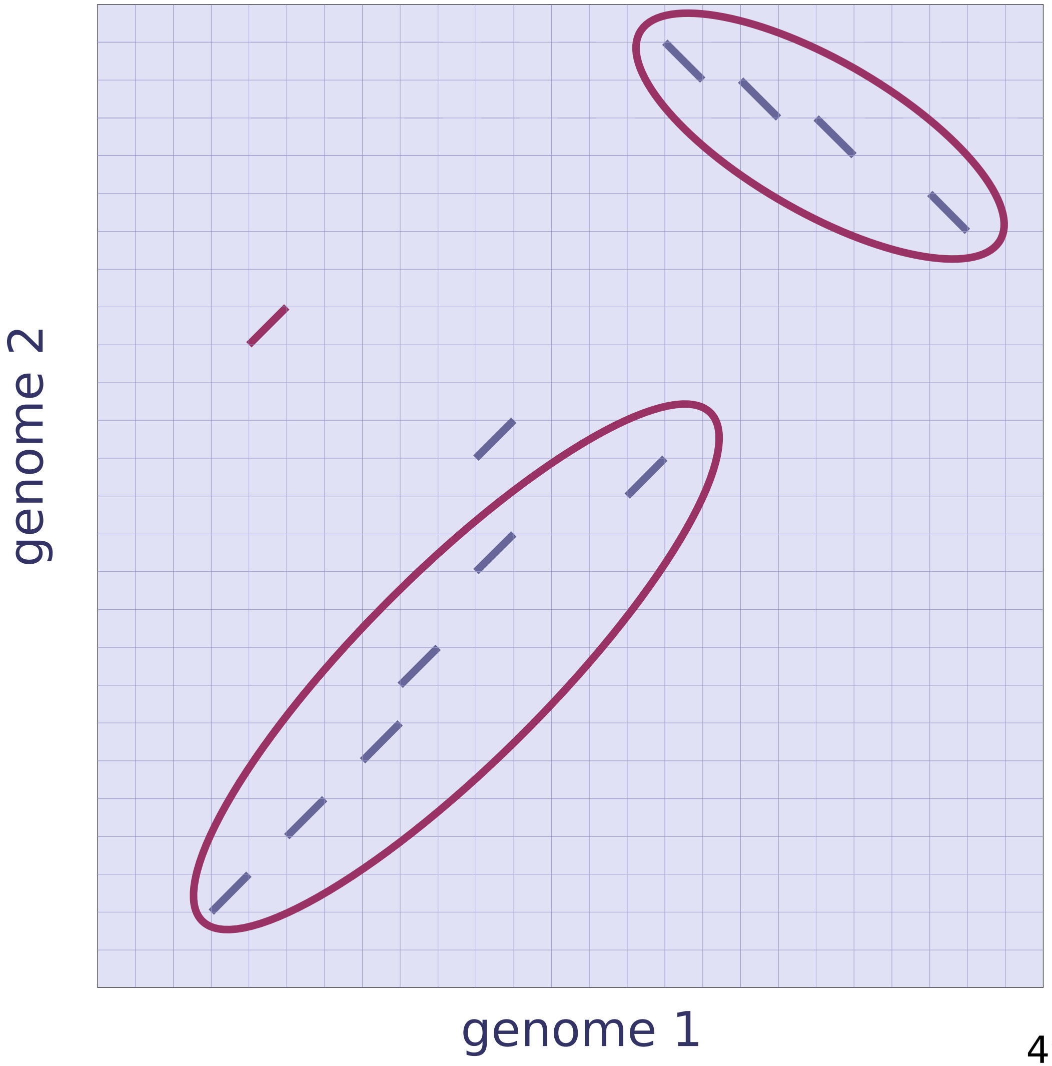


Branco and Pombo, 2006

merci !!!

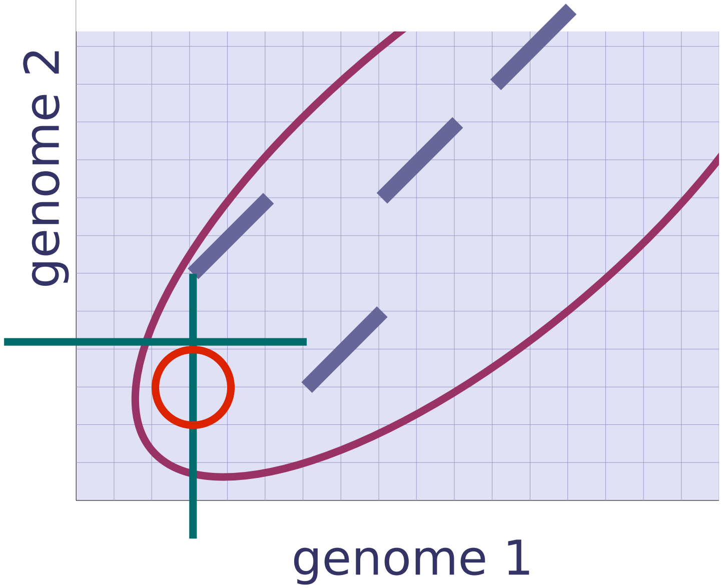
Synteny blocks

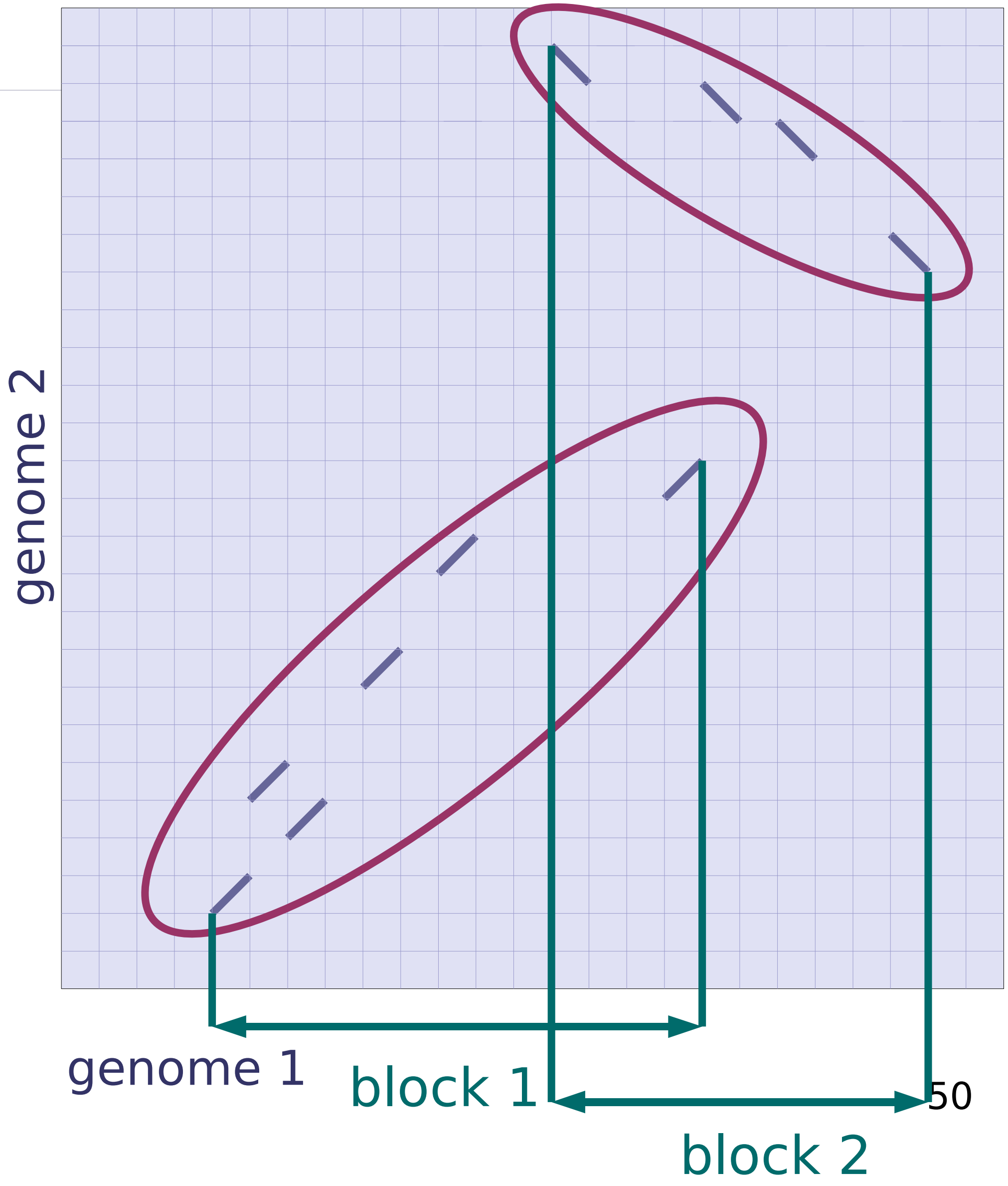
- Flexibility
- Chaining principle:
    - colinearity
    - distance criteria
    - size criteria



Conflicts

- overlaps between blocks
- orthology at the extremities

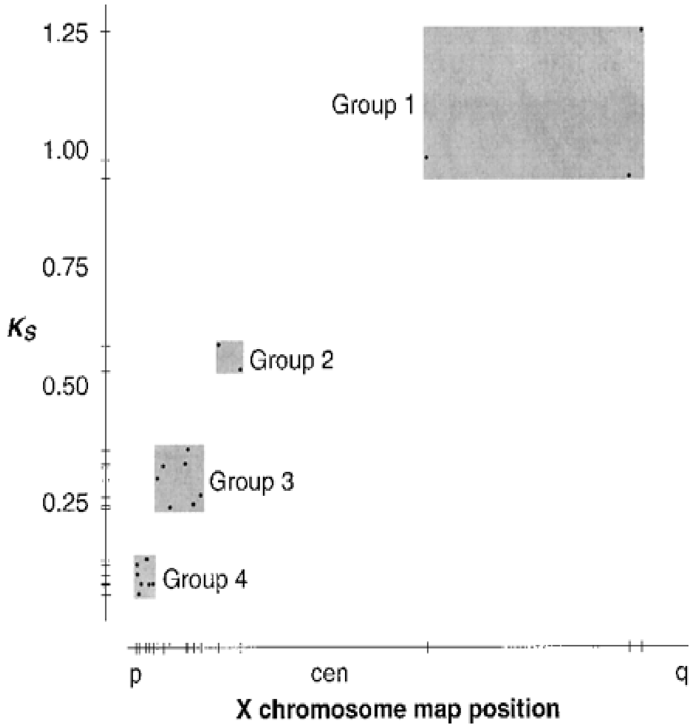




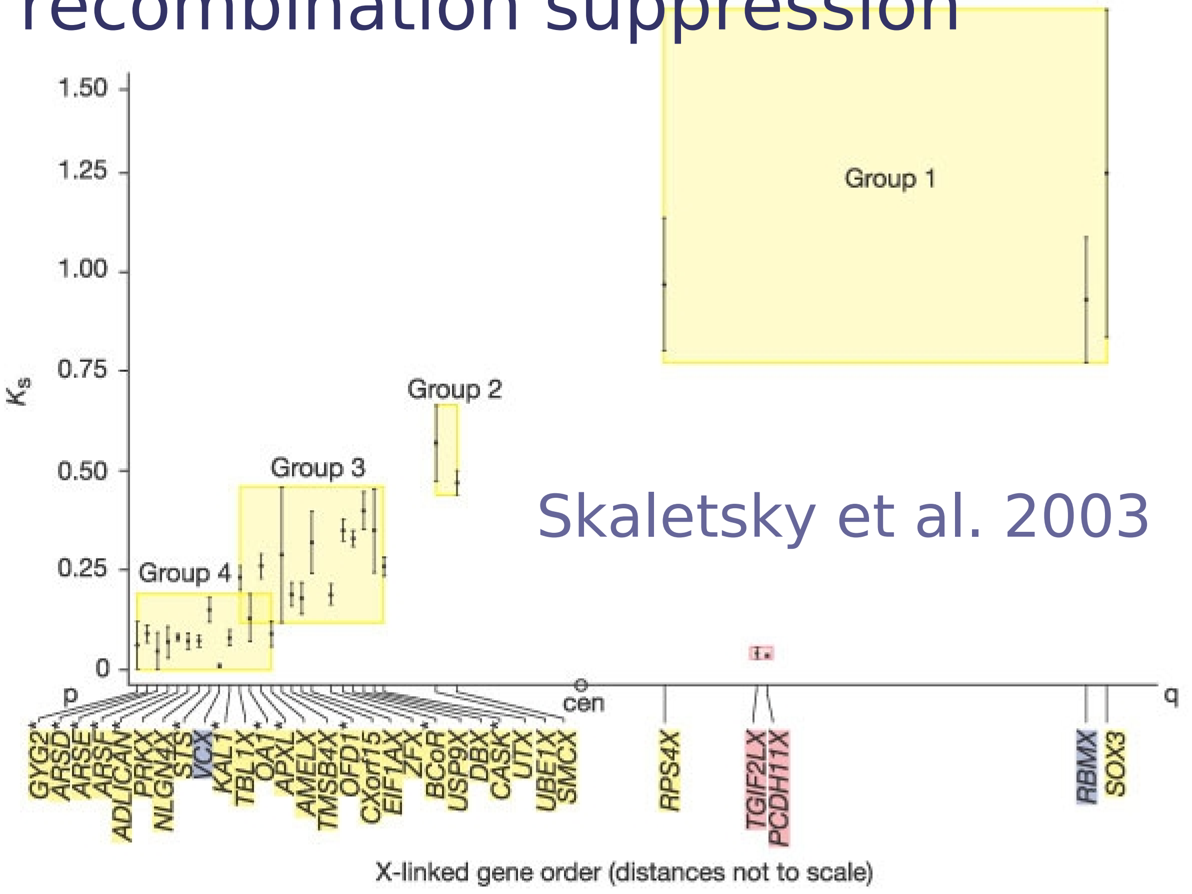
Human X-Y divergence

- stair-shape of the divergence between X-Y genes along X

  => several steps of recombination suppression



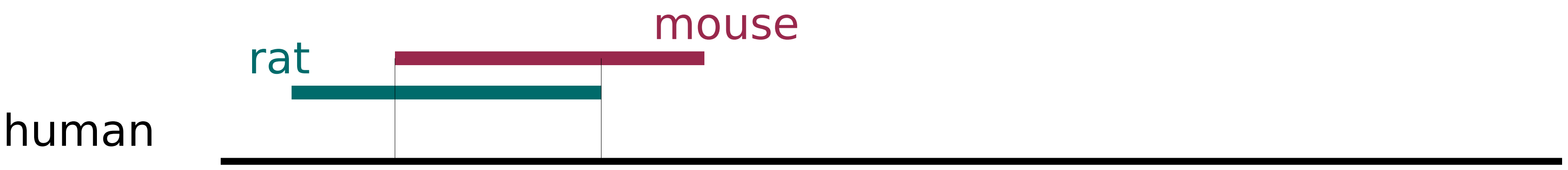
Lahn and Page, 1999

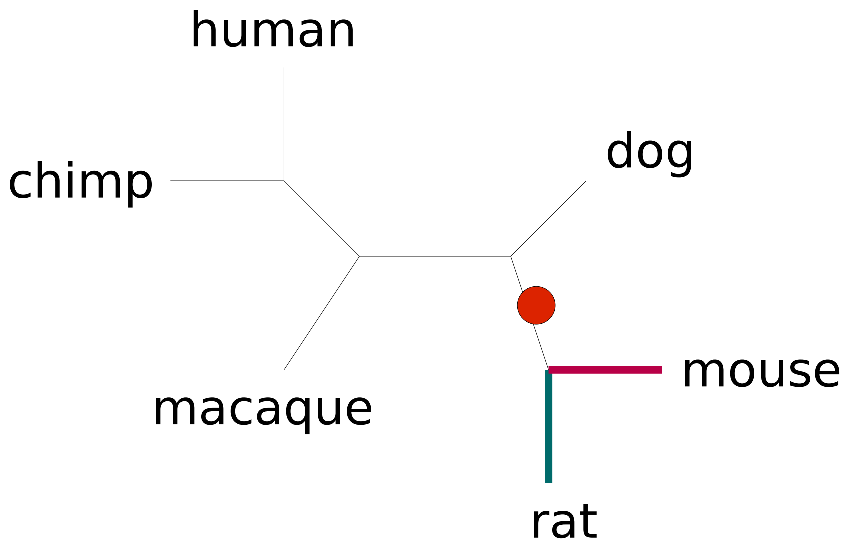


Skaletsky et al. 2003

Combining pairwise datasets

- Example :


mouse
rat
human


human
chimp
dog
mouse
macaque
rat