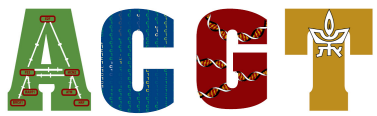
Organization of genes from complexes and pathways in cancer research

Liat Ein-Dor[1] Michal Ozery-Flato[1] Ron Shamir[2]
Annelyse Thévenin[2]

[1]Machine Learning and Data Mining group, IBM Haifa Research Lab, Israel

[2]Algorithms in Computational Genomics in TAU, Tel Aviv University, Israel



Algorithms in Computational Genomics at Tau

Cancers

Definitions

Tumor An abnormal proliferation of tissues. Cell transformation results including loss of cell cycle control, insensitivity to apoptosis, abnormalities in DNA repair.

Cancer The cells are distancing themselves from the lineage of origin - malignant tumor.

STACK - STatistical Associations in Cancer Karyotypes

Previous work: *A systematic assessment of associations among chromosomal aberrations in cancer karyotypes*, Michal Ozery-Flato, Chaim Linhart, Luba Trakhtenbrot, Shai Izraeli and Ron Shamir.

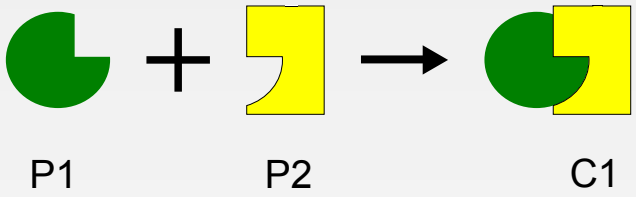
Methods: Observation of aberrations in 15 000 karyotypes.

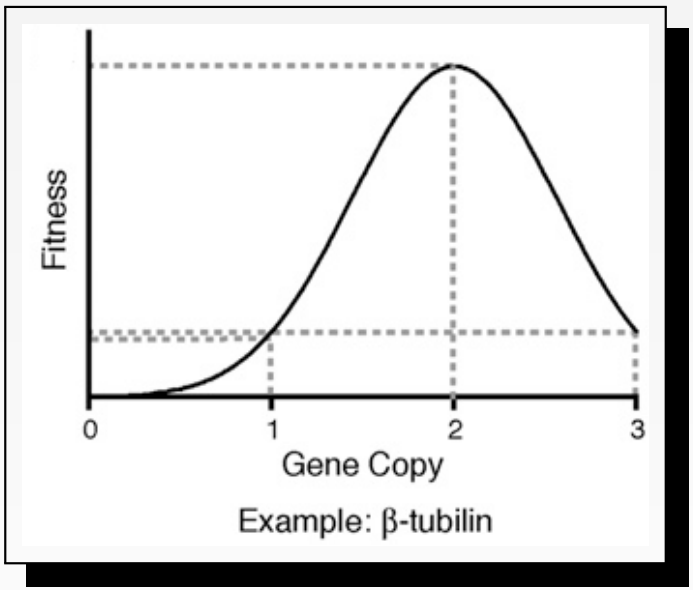
Results: There exist pairs of aberrations significantly correlated. The majority of them are gain/gain or loss/loss.

http://acgt.cs.tau.ac.il/stack/

Protein stoichiometry

Two genes g_1 and g_2 with 2 copies of each in the genome, and their products, the proteins P_1 and P_2.





[Katz *et al.* 1990]

Aims

2 aims

- Is there a link between two chromosomes which could explain the gains and lost correlated observed in the STACK study?
- Are the position of genes from the same complex/pathway significantly close?

I - Data

Data: complexes and pathways

We study 3 datasets of **groups**, a group is either a protein complex or a pathway.

Protein complexes: CORUM

- Set of genes whose products form a protein complex;
- **Number**: 1587 human complexes;
- **Size**: between 2 and 143 genes (average=5.05 genes).

Pathways: KEGG and REACTOME

- Set of genes whose products are present in the same pathway.
- **Number**: 224 and 68 human pathways;
- **Size**: between 2 and 1126/872 genes (average=71.44/122.87 genes).

II - Correlation inside two chromosomes

Definitions

- *Number of groups* involved in one(two) specific chromosome(s).
- *Number of chromosomes* involved in a specific group.
- Measure of *entropy* 2^s for groups with the same number of genes such that s is equal to $-\sum_{i=1}^{n_c} p_i \log(p_i)$, n_c is the number of chromosomes and p_i is the number of genes of this group involved in the chromosome i divided by the number of genes of this group.
- A *fullGroup* is a group whose genes are in the same chromosome.

	Numbers of genes					
	Chro1	Chro2	Chro3	Chro4	Total	$\mathbf{2^s}$
Group 1	1	1	1	19	22	1.46
Group 2	1	1	10	10	22	1.99

Simulation

A permutation of the whole genome.

We change the order of gene in the whole genome.

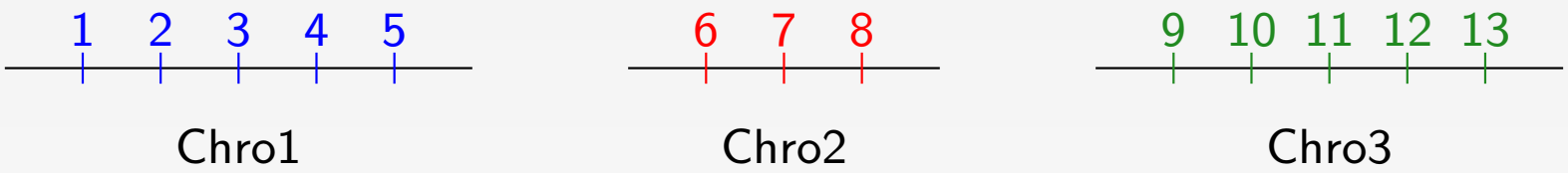


Figure: One genome with 3 chromosomes.

One group: (3, 4, 8, 13).

Simulation

A permutation of the whole genome.

We change the order of gene in the whole genome.

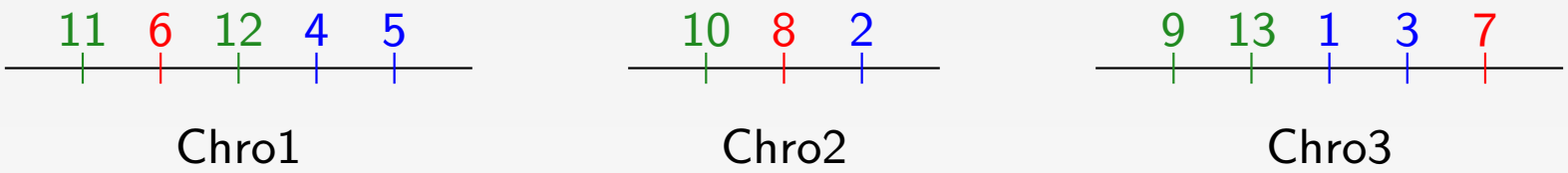


Figure: The genome after a genome permutation

One group: (3, 4, 8, 13).

Protein complexes - Significant results

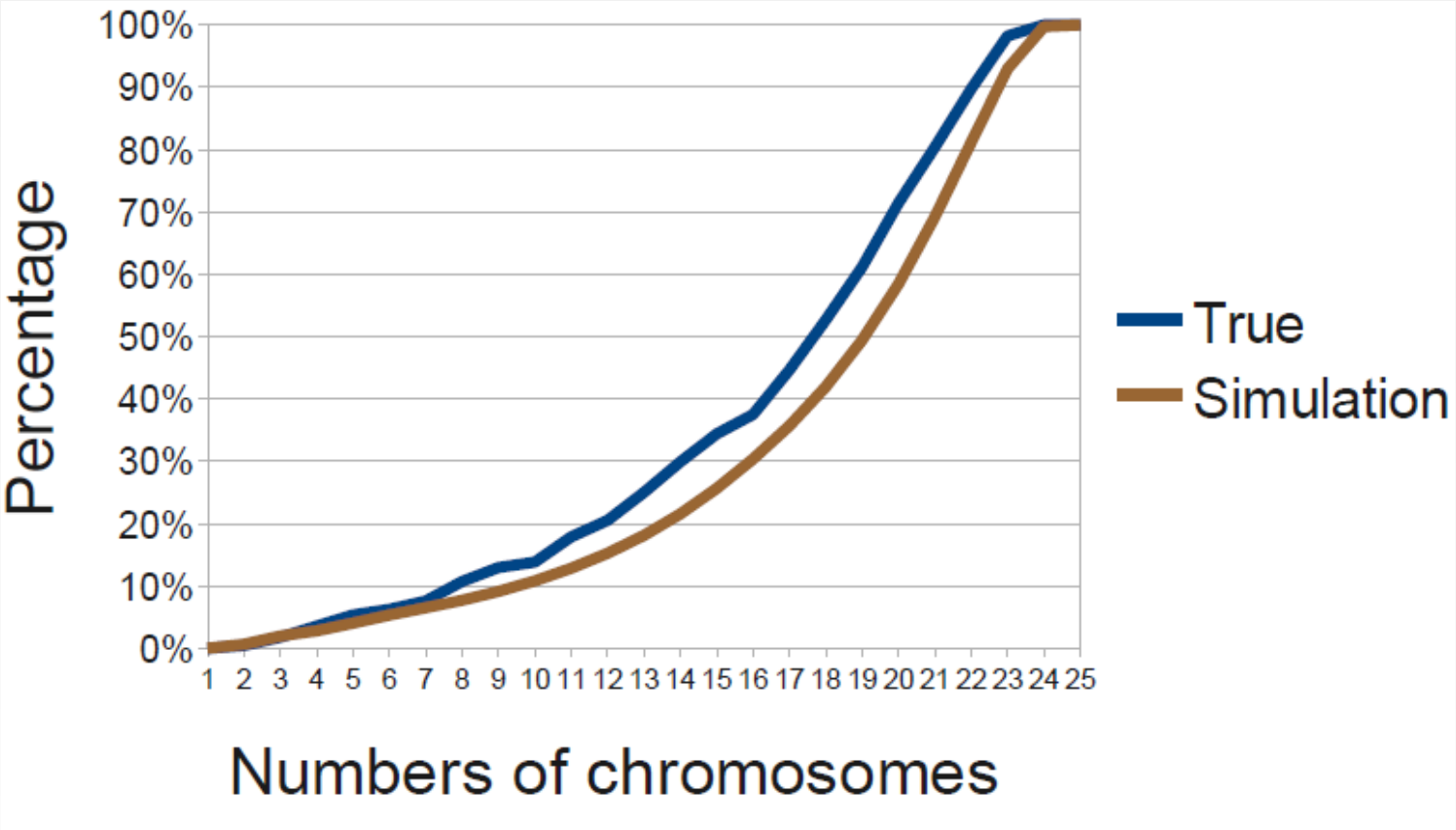
- The 432 complexes with only 2 genes are involved in significantly few chromosomes (True data=1.91 chromosomes, Simulation=1.96, p-value=0.002).
- Measure of entropy is significantly small for the complexes with 2 genes (True data=1.91 chromosomes, Simulation=1.93, p-value=0.002).
- Number of fullComplexes significantly large (True data=41 complexes, Simulation=24.85, p-value=0.001).

Pathways - Significant results

- There are 9 chromosomes with significantly few pathways: 8, 11, 13, 18, 19, 20, 21 and X.
- 73 (32.5%) pathways are significantly involved in few number of chromosomes. 11 (4.9%) are significantly involved in lot of chromosomes.
- 66 (29.4%) pathways have a low entropy (which 43 among the 73 previous pathways).
- 41 set of pathways with the same number of genes (41 among 113 - 36.2%, = 97 pathways) have a low entropy.
- There is a significantly low entropy for pathways with at more x genes ($x \in [7; 1126]$).

Pathways - Results

There are significantly lot of pathways involved in at more x chromosomes ($x \in [4; 24]$)



Protein complexes and STACK - Definitions

Let i and j two chromosomes.

$c_{i,j}$ Number of complexes involved only in i and j, in average $c_{i,j} = 1.59$.

$\mathcal{C} = \{c_{i,j}\}$.

$\mathcal{C}^* = \{c_{i,j} | \text{ significantly large}\}$, $|\mathcal{C}^*| = 22$.

$R^c_{i,j}$ Set of ranks for all (i,j) in $\mathcal{C}$ in function of $c_{i,j}$.

$s_{i,j}$ P-value from STACK.

$\mathcal{S} = \{s_{i,j}\}$.

R^s Set of ranks for all (i,j) in $\mathcal{S}$ in function of $s_{i,j}$.

Protein complexes and STACK - Definitions

- Average $s_{i,j}$ for the 22 chromosome pairs in $\mathcal{C}^*$: 0.014 (1,000,000 random runs $\rightarrow$ 0.018, p-value=0.526).
- Average rank from R^s for the chromosome pairs in $\mathcal{C}$ (84.5) *vs* the average rank for the 22 chromosome pairs in $\mathcal{C}^*$ (81.63) (after 1,000,000 random runs $\rightarrow$ 84.51, p-value=0.61).
- Pearson correlation between $\mathcal{C}$ and $\mathcal{S}$ (-0.054) *vs* between $\mathcal{C}^*$ and $\mathcal{S}$ (-0.00076).
- Similar analysis for $\mathcal{S}^*$ (the 22 pairs of chromosomes with the largest value $s_{i,j}$).

$\hookrightarrow$ No significant result.

Gene clusters

A **gene cluster** is a set of two or more genes that serve to encode for the same or similar products. They are generally close on the genome.

HLA - Human Leukocyte Antigen

All genes of this cluster are in the chromosome 6.

- CORUM: 56 complexes with exactly 1 HLA gene.
- KEGG: 16 pathways with at least one HLA gene (15 in average)
- REACTOME: 1 pathway with 18 HLA genes.

Gene clusters

A **gene cluster** is a set of two or more genes that serve to encode for the same or similar products. They are generally close on the genome.

Olfactory receptor

More than half in the chromosomes 1 and 11.

- CORUM: 274 complexes with in average 1.28 olfactory receptors.
- KEGG: 1 pathway with 388 olfactory receptors.
- REACTOME: 1 pathway with 339 olfactory receptors.

III - Distances and measures

1) Genomic distances

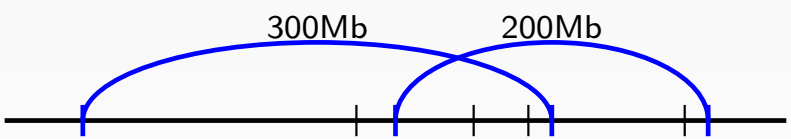
Genomic distances - Definition

Let $d_{i,j}$ a distance between the genes i and j.

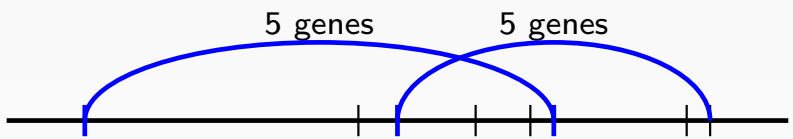
4 genomics distances

All pairs of genes are in the same chromosome.

- **For one pair of genes**: $d_{i,j}$ is equal to the number of *bases* or the number of *genes*.



DistanceBase



DistanceGene

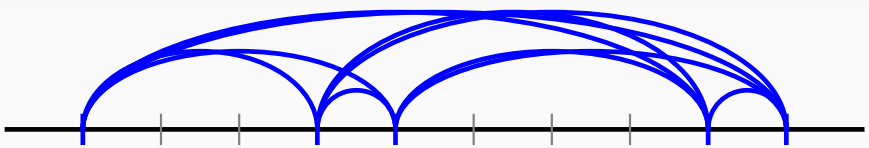
Genomic distances - Definition

Let $d_{i,j}$ a distance between the genes i and j.

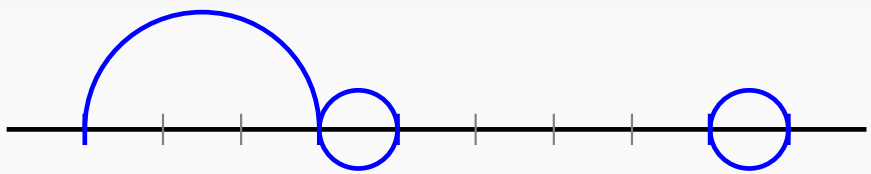
4 genomics distances

All pairs of genes are in the same chromosome.

- **For one group** (complex or pathway): The distance is either $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{i,j}$ (*distanceAll*), or $\sum_{i=0}^{n} d_{i,k_i}$ with k_i the closer gene of i (*distanceClose*).



DistanceAll



DistanceClose

Distances - Simulations

One permutation by chromosome.

We change the order of gene inside each chromosome.

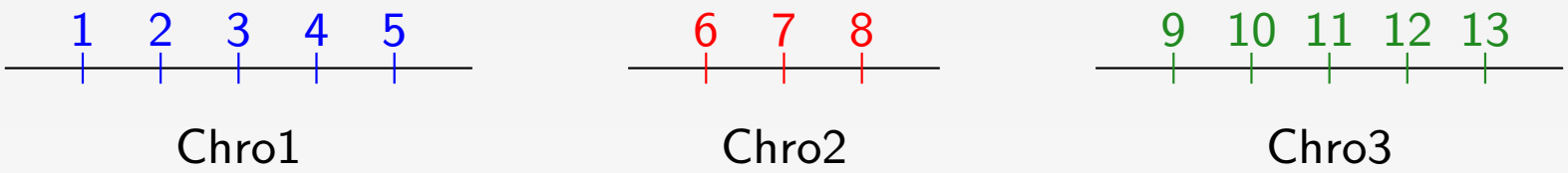


Figure: One genome with 3 chromosomes.

One group: (3, 4, 8, 13).

Distances - Simulations

One permutation by chromosome.

We change the order of gene inside each chromosome.

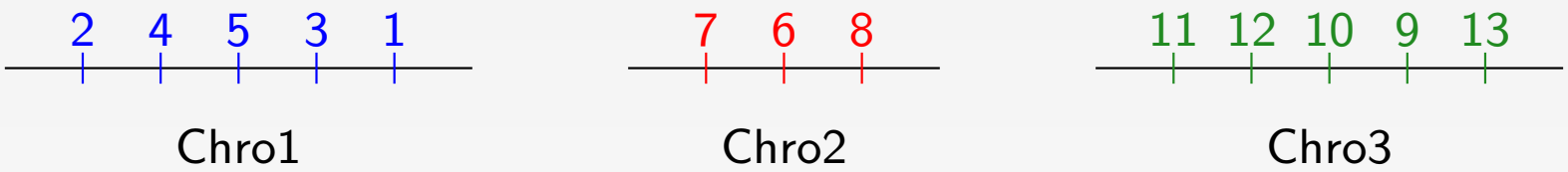


Figure: The genome after chromosome permutations

One group: (3, 4, 8, 13).

Genomic distances - Results

Complexes: We do not have significant result.

Pathways:

	All		Close	
	Base	Gene	Base	Gene
KEGG	0	0	0	0.36
REACTOME	0.34	0	0.96	0.16

Figure: The distance is significantly small if the p-value is ≤ 0.05.

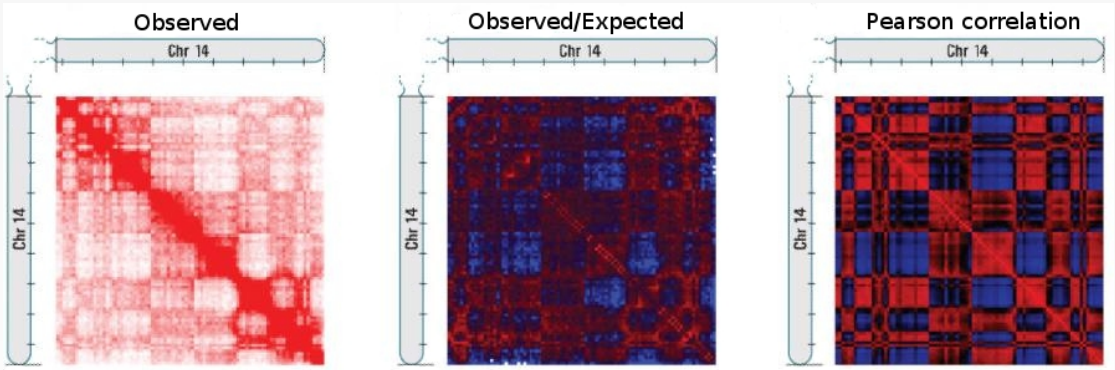
III - Distances and measures

2) Spatial measures

Data: genome fold

3D genomes [Lieberman-Aiden *et al.*, 2010]

- The **Hi-C** method probes the 3-dimensional architecture of whole genome.
- They obtain intra- and interchromosomal **contact probability**.
- Every chromosomes are decomposed in **regions of 1,000,000 bases**.



Spatial measures - Definition

We compute 4 spatial measures.

Let $m_{i,j}$ is one of these measures between the genes i and j.

4 spatial measures

The genes of one pair could be in different chromosomes.

- **For one pair of genes**: Average of pearson correlation
- **For one group**: $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_{i,j}$

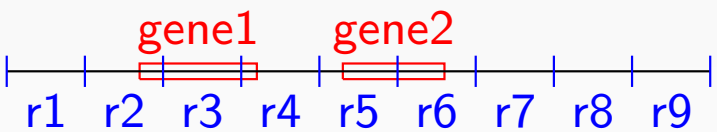


Figure: Two genes and ten regions.

Spatial measures - Definition

We compute 4 spatial measures.

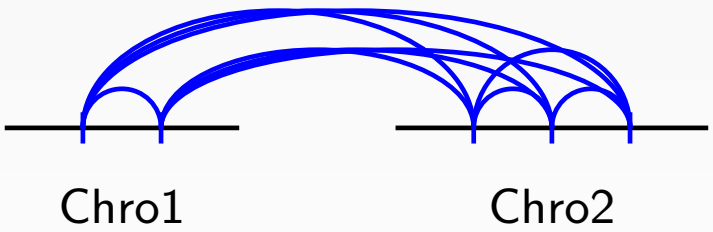
Let $m_{i,j}$ is one of these measures between the genes i and j.

4 spatial measures

The genes of one pair could be in different chromosomes.

- **For one pair of genes**: Average of pearson correlation
- **For one group**: $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_{i,j}$

  measureAll: for all i and j.



Spatial measures - Definition

We compute 4 spatial measures.

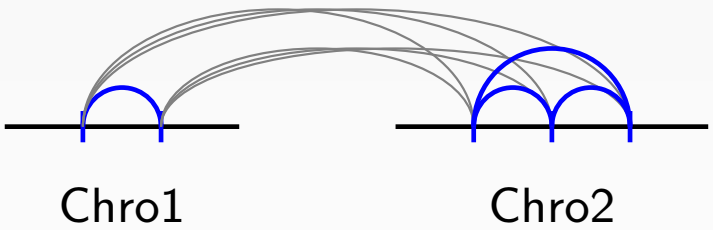
Let $m_{i,j}$ is one of these measures between the genes i and j.

4 spatial measures

The genes of one pair could be in different chromosomes.

- **For one pair of genes**: Average of pearson correlation
- **For one group**: $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_{i,j}$

  measureIntra: for all i and j in the same chromosome.



Spatial measures - Definition

We compute 4 spatial measures.

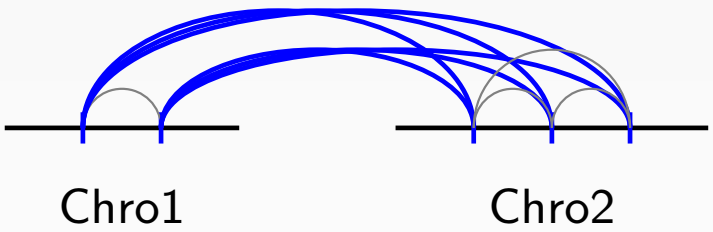
Let $m_{i,j}$ is one of these measures between the genes i and j.

4 spatial measures

The genes of one pair could be in different chromosomes.

- **For one pair of genes**: Average of pearson correlation
- **For one group**: $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_{i,j}$

  measureInter: for all i and j in different chromosomes.



Spatial measures - Definition

We compute 4 spatial measures.

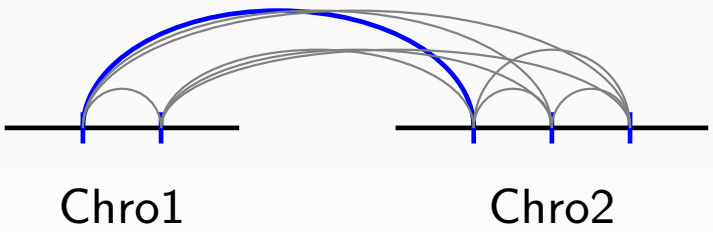
Let $m_{i,j}$ is one of these measures between the genes i and j.

4 spatial measures

The genes of one pair could be in different chromosomes.

- **For one pair of genes**: Average of pearson correlation
- **For one group**: $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} m_{i,j}$

  measureInterMax: for the pair i and j in different chromosomes with the highest correlation.



Spatial measures - Simulations

Two types of simulation: chromosome permutations and genome permutation.

We change the order of gene in the whole genome.
↪ For *measureAll*.

We change the order of gene inside each chromosome.
↪ For *measureIntra*, *measureInter* and *measureInterMax*.

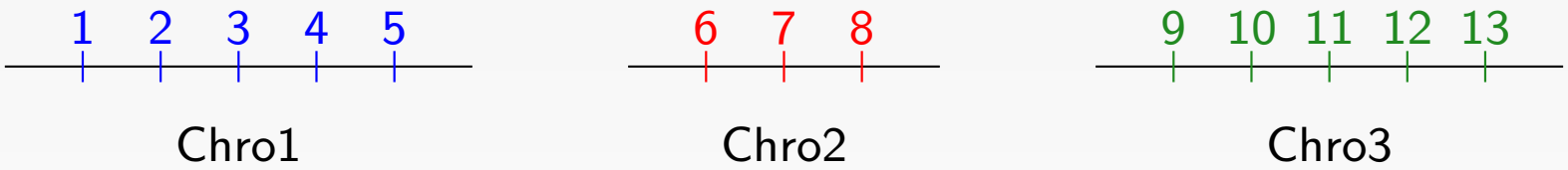


Figure: One genome with 3 chromosomes.

One group: (3, 4, 8, 13).

Spatial measures - Simulations

Two types of simulation: chromosome permutations and genome permutation.

We change the order of gene in the whole genome.
↪ For *measureAll*.

We change the order of gene inside each chromosome.
↪ For *measureIntra*, *measureInter* and *measureInterMax*.

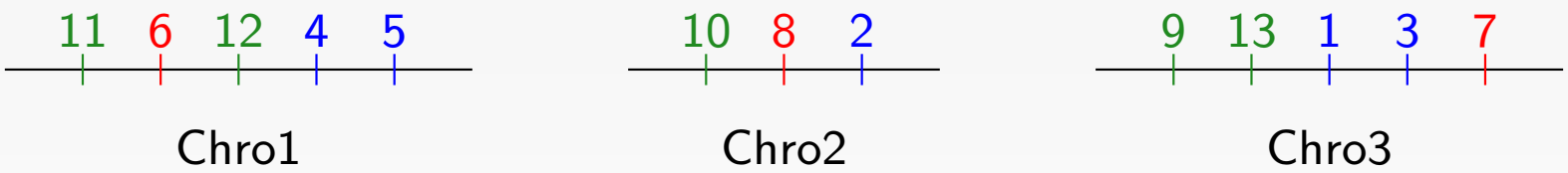


Figure: The genome after a genome permutation

One group: (3, 4, 8, 13).

Spatial measures - Simulations

Two types of simulation: chromosome permutations and genome permutation.

We change the order of gene in the whole genome.
↪ For *measureAll*.

We change the order of gene inside each chromosome.
↪ For *measureIntra*, *measureInter* and *measureInterMax*.

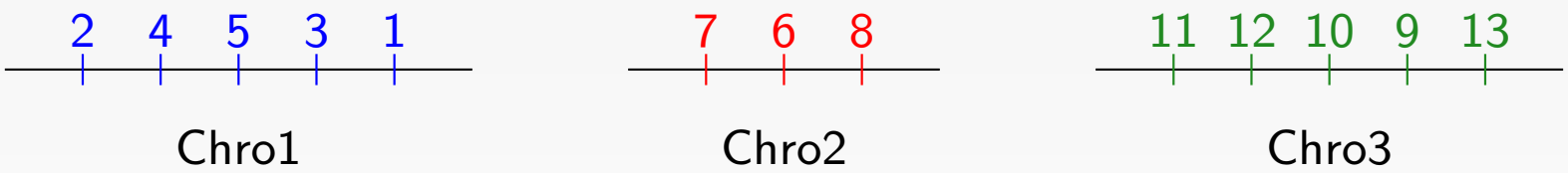


Figure: The genome after chromosome permutations

One group: (3, 4, 8, 13).

Spatial measures - Results

For KEGG:

	True	Simulations	p-value (S≤T)	p-value (T≤S)
MeasureAll	0.041	0.027	1.000	0.000
MeasureIntra	0.365	0.263	1.000	0.000
MeasureInter	0.013	0.014	0.101	0.898
MeasureInterMax	0.582	0.612	0.035	0.965

Table: Spatial measures and p-values for 100,000 simulations.

Conclusions

Conclusions - Correlation inside two chromosomes

Study of organization of human genes whose the products are in protein complexes or in pathways.

- **Not enough groups** involved only in a specific pair of chromosomes.
- Groups involved in 1 or 2 chromosomes are **significantly concentrated**.
- The significant results could be explained by **gene clusters**.

Conclusions - Distances and measures

Protein complexes - CORUM

No significant result (genomic distance) or no result (spatial measures).

Pathways - KEGG and REACTOME

- Genes, in the **same** chromosome, from the same pathway are **close** (genomic distance and spatial measures).
- Genes, in **different** chromosomes, from the same pathway are **far**.

Prospects

- Operate with **gene clusters**.
- Study **individually** each group.
- Spatial measure: similar study with **coordinates**.
- **Tissue** specific.
- Other **species**.