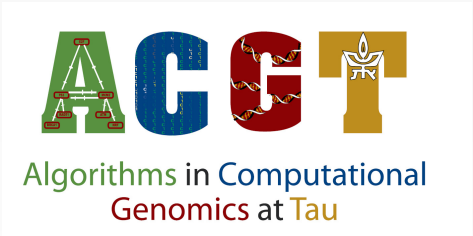
On the relation between gene organization, functional gene groups and cancer

Annelyse Thevenin, Liat Ein-Dor (IBM),
Michal Ozery-Flato (IBM) and Ron Shamir



Background

Until early 2000s

- Gene order in eukarayotes was believed to be random
- Relatively little was known about gene functionality
- Limited data about chromosomal location of genes

What changed?

- Availability of data
- Gene expression
- Whole-genome sequences
- Gene functionality

Previous Studies

Main findings for non-random gene order

- Genes from the same metabolic **pathway** tend to cluster (e.g. J.M. Lee, E.L. Sonnhammer 2003)
- Genes with similar and/or coordinated **expression** tend to cluster (e.g. Cho et al. 2005)
- Adjacent genes are often co-regulated by the same **transcription factor** (Hershberg et al. 2005)

Methods

Pathways - Proximity of genes within each pathway.

Expression - Correlation between genes within a sliding window.

Transcription networks - Motifs in the integrated network of transcriptional relations and chromosomal adjacency.

What is still missing?

- **Whole** genome and whole network study.
- **Uniform methodology**, applicabale to all functional relations.
- Analysis of **additional functional relations** (e.g. PPIs, Complexes).
- Integrating the results into a **unified model** explaining gene organization.

I - Data

II - Concentration of co-functioning genes

III - Intra-chromosomal distances

IV - Spatial measures

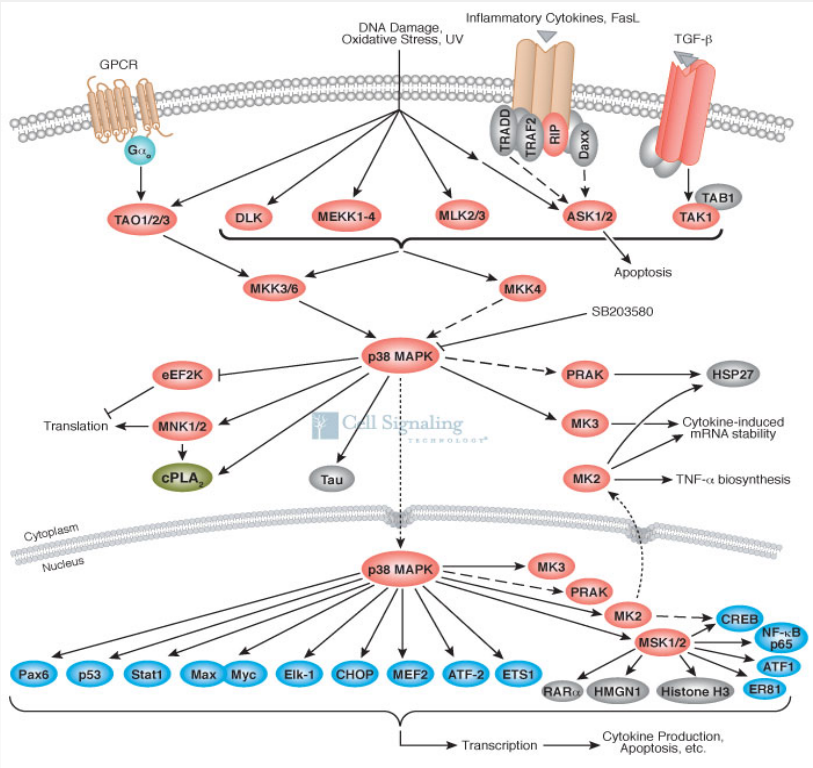
V - Chromosome pairs and cancer

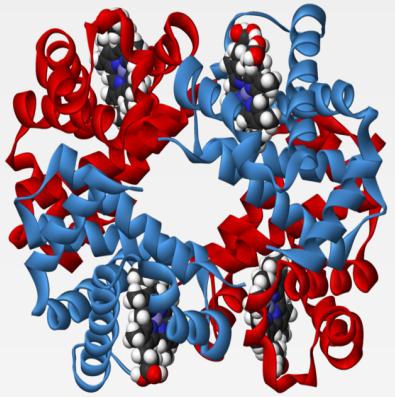
I - Data

Functional relations between genes: Expression, transcriptional relations, transcription factor and target gene, pathways, complexes, PPIs, miRNA networks...

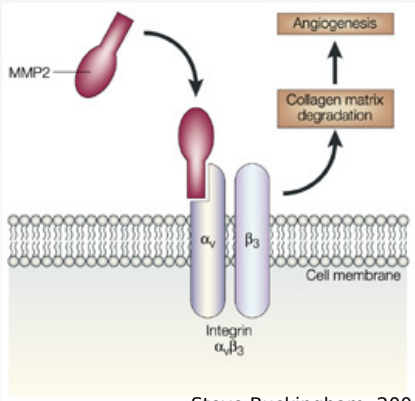
Data: PPI, protein complexes and pathways

Signaling Pathways Activating p38 MAPK





Hemoglobin



Steve Buckingham, 2004

$\sim$ 20,000 protein-coding genes from NCBI.

Data sets

We study 3 datasets of **groups**, a group is either

- Protein-Protein Interaction (**PPI**) from `IntAct`
  - 31,375 sets of Human genes whose products interact.
  - **Size**: 2 genes.
- **Protein complexes** from `CORUM`
  - 1,520 sets of genes whose products form a Human protein complex;
  - **Size**: between 2 and 142 genes (average=5.1 genes).
- **Pathways** from `Reactome` and `KEGG`
  - 650 and 206 sets of genes whose products are present in the same Human pathway.
  - **Size**: between 2 and 345/1122 genes (average=24/69.1 genes).

Data sets

We study 3 datasets of **groups**, a group is either

- Protein-Protein Interaction (**PPI**) from `IntAct`
  - 31,375 sets of Human genes whose products interact.
  - **Size**: 2 genes.
- **Protein complexes** from `CORUM`
  - 1,520 sets of genes whose products form a Human protein complex;
  - **Size**: between 2 and 142 genes (average=5.1 genes).
- **Pathways** from `Reactome` and `KEGG`
  - 650 and 206 sets of genes whose products are present in the same Human pathway.
  - **Size**: between 2 and 345/1122 genes (average=24/69.1 genes).

Data sets

We study 3 datasets of **groups**, a group is either

- Protein-Protein Interaction (**PPI**) from `IntAct`
  - 31,375 sets of Human genes whose products interact.
  - **Size**: 2 genes.
- **Protein complexes** from `CORUM`
  - 1,520 sets of genes whose products form a Human protein complex;
  - **Size**: between 2 and 142 genes (average=5.1 genes).
- **Pathways** from `Reactome` and `KEGG`
  - 650 and 206 sets of genes whose products are present in the same Human pathway.
  - **Size**: between 2 and 345/1122 genes (average=24/69.1 genes).

II - Concentration of co-functioning genes

Test

Relative position of co-functioning genes whole-genome whole-network analysis.

- **Define test function**: Average of number of chromosomes involved in each group.
- Calculate the function value in the **real genome**.
- Estimate the probability to observe a value or higher (or smaller) for **random gene order**.

1 mega-simulations

A permutation of the whole genome.

We change the order of gene in the whole genome.

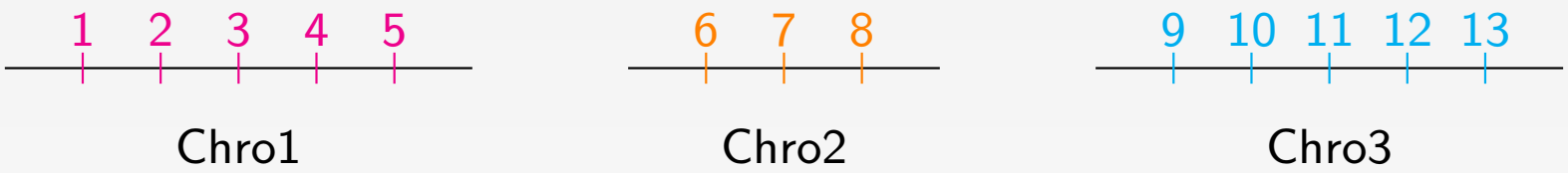


Figure: One genome with 3 chromosomes and 13 genes.

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

1 mega-simulations

A permutation of the whole genome.

We change the order of gene in the whole genome.

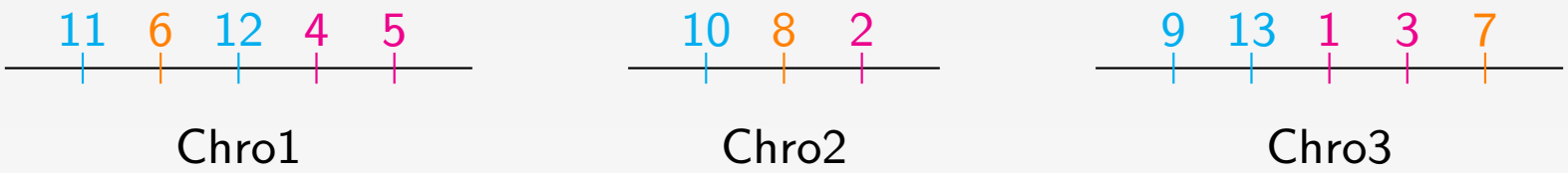


Figure: The genome after a genome permutation

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

Significant results: Number of chromosomes involved

The groups are significanlty involved in few chromosomes for `IntAct`, `KEGG` and `CORUM`. (P-values: 0.0107 for `IntAct`, 0.0788 for `CORUM`, < 0.00001 for `KEGG` and 0.0072 for `Reactome`).

Conclusion

On average, pathways and PPIs tend to involve significantly smaller number of chromosomes than expected at random.

Can we find a more informative test function?

Significant results: Cumulative distribution

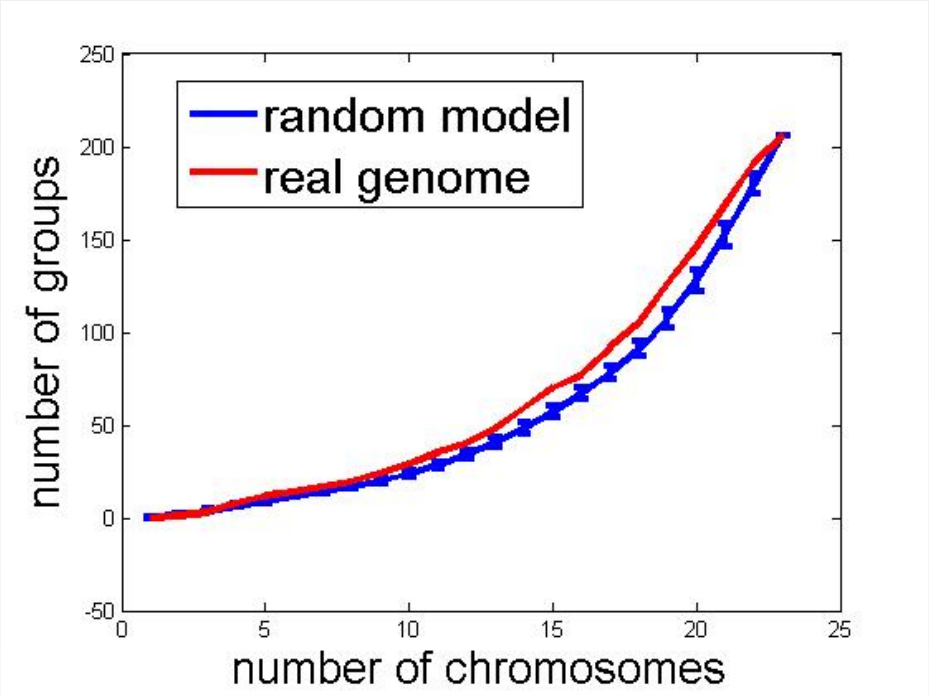


Figure: Cumulative distribution of the number of chromosomes involved in a pathway for the real genome (red curve) and for the average over 10^6 random genome (blue curve). The error-bars denote the standard deviation.

Significant results: Cumulative distribution

- Complexes - Tend to involve a single chromosome (Pvalue 0.0025).
- PPIs - Pairs tend to lie on the same chromosome (Pvalue 0.0083).
- Pathways - Tend to involve fewer chromosomes than expected at random. The effect becomes statistically significant from 5 chromosomes for Kegg and from 2 chromosomes for Reactome (Pvalue 0.013, 0.04 after correction for multiple test).

III - Intra-chromosomal distances

Intra-chromosomal distances - Definition

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

Intra-chromosomal distances

Pairs of genes along the same chromosome.

- **For one pair of genes** (i, j) **along one chromosome**: $d_{i,j}$ such that i preceded j, is equal to the number of *bases* between the end of i and the beginning of j.

Intra-chromosomal distances - Definition

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

Intra-chromosomal distances

Pairs of genes along the same chromosome.

- **For one group and one chromosome**: The distance is $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{i,j}/(n(n-1)/2)$.

Intra-chromosomal distances - Definition

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

Intra-chromosomal distances

Pairs of genes along the same chromosome.

- **For one group in the genome**: The group distance is the average of chromosomal distances for all chromosomes containing at least two genes of this group.

Intra-chromosomal distances - 1 mega-simulations

One permutation by chromosome.

We change the order of genes inside each chromosome.

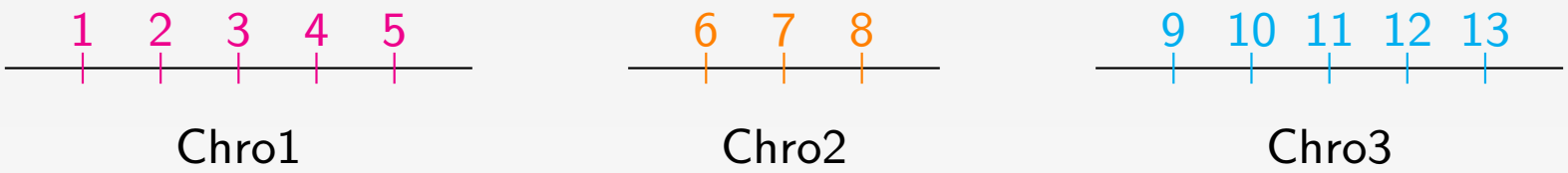


Figure: One genome with 3 chromosomes and 13 genes.

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

Intra-chromosomal distances - 1 mega-simulations

One permutation by chromosome.

We change the order of genes inside each chromosome.

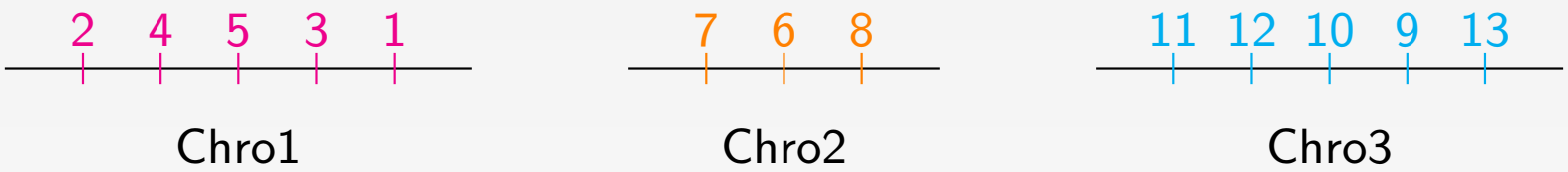


Figure: The genome after chromosome permutations

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

Intra-chromosomal distances - Results

Data set	p-value
Reactome	$< 10E^{-05}$
KEGG	$< 10E^{-05}$
CORUM	$7.2E^{-04}$
IntAct	0.0986

Table: $P(d_i' \leq d)$ for intra-chromosomal distance of 4 data sets.

Conclusion

The co-functioning genes from pathways or protein complexes are **close** along chromosomes.

Significant results: Histogramme

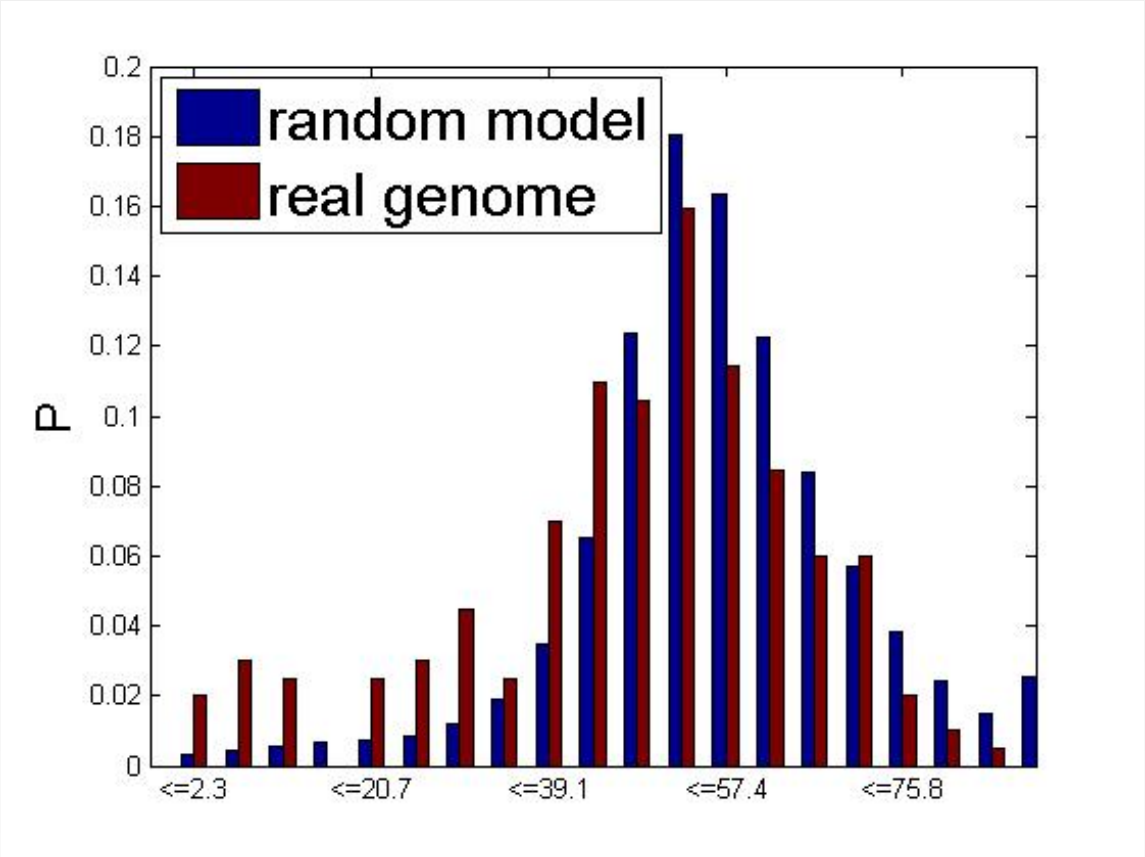


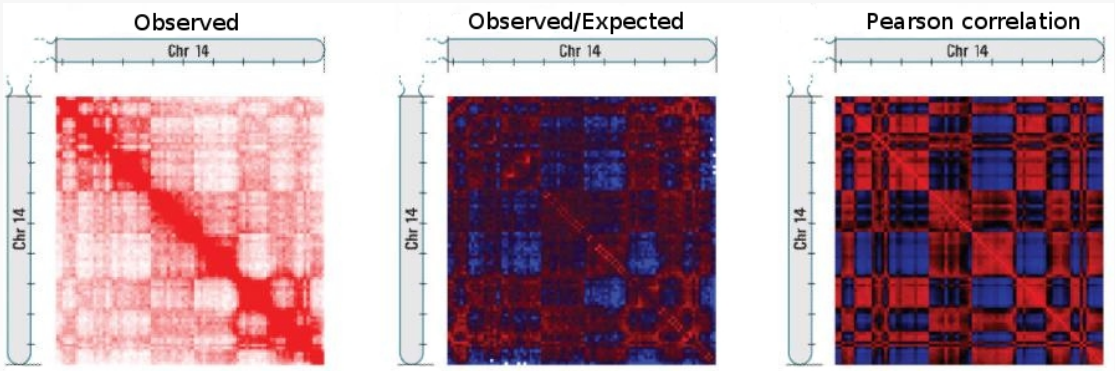
Figure: Normalized histograms of the average intra-chromosomal distance in Mega BP between genes from the same group compared to average over random genomes for KEGG.

IV - Spatial measures

Spatial measures - 3D Human genome

3D genome [Lieberman-Aiden *et al.*, 2010, Eitan and Amos, 2011]

- The **Hi-C** method probes the 3-dimensional architecture of whole genome.
- They obtain intra- and interchromosomal **contact probability**.
- Every chromosome is decomposed into **regions of 1 mega-bases**.



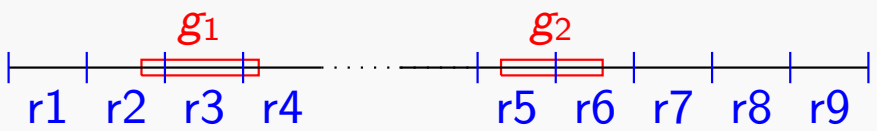
Spatial measures - 3 measures

We compute 3 spatial measures.
Let $m_{i,j}$ is one of these measures between the genes i and j.

3 spatial measures

The genes of one pair can be in different chromosomes.

- **For one pair of genes**: Average of Pearson correlation
- **For one group**: Average of $m_{i,j}$



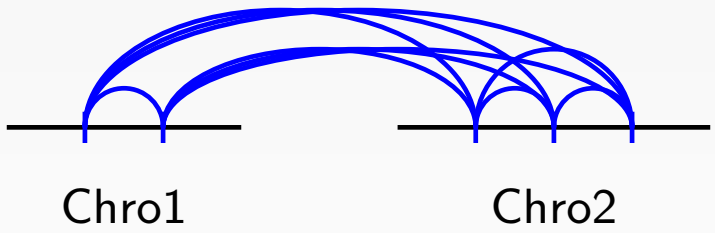
Spatial measures - 3 measures

We compute 3 spatial measures.
Let $m_{i,j}$ is one of these measures between the genes i and j.

3 spatial measures

The genes of one pair can be in different chromosomes.

- **For one pair of genes**: Average of Pearson correlation
- **For one group**: Average of $m_{i,j}$
  measureAll: for all i and j.



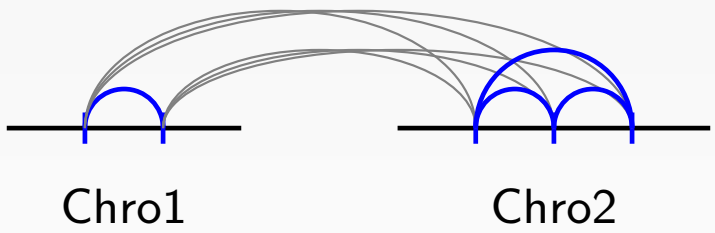
Spatial measures - 3 measures

We compute 3 spatial measures.
Let $m_{i,j}$ is one of these measures between the genes i and j.

3 spatial measures

The genes of one pair can be in different chromosomes.

- **For one pair of genes**: Average of Pearson correlation
- **For one group**: Average of $m_{i,j}$
  measureIntra: for all i and j in the same chromosome.



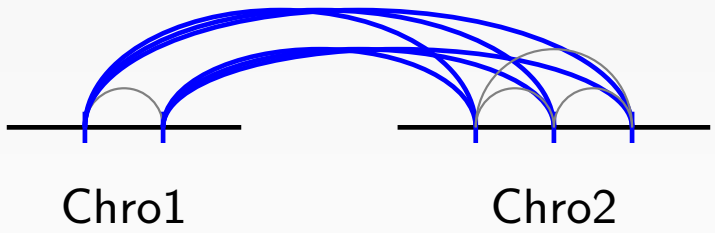
Spatial measures - 3 measures

We compute 3 spatial measures.
Let $m_{i,j}$ is one of these measures between the genes i and j.

3 spatial measures

The genes of one pair can be in different chromosomes.

- **For one pair of genes**: Average of Pearson correlation
- **For one group**: Average of $m_{i,j}$
  measureInter: for all i and j in different chromosomes.



Spatial measures - 1 mega-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 genes inside each chromosome.
↪ For *measureIntra* and *measureInter*.

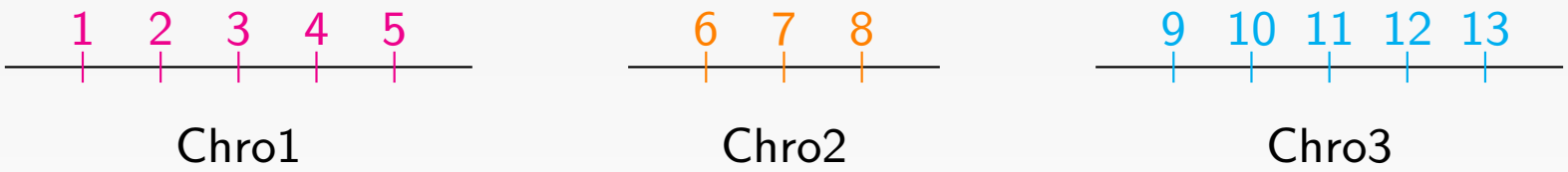


Figure: One genome with 3 chromosomes and 13 genes.

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

Spatial measures - 1 mega-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 genes inside each chromosome.
↪ For *measureIntra* and *measureInter*.

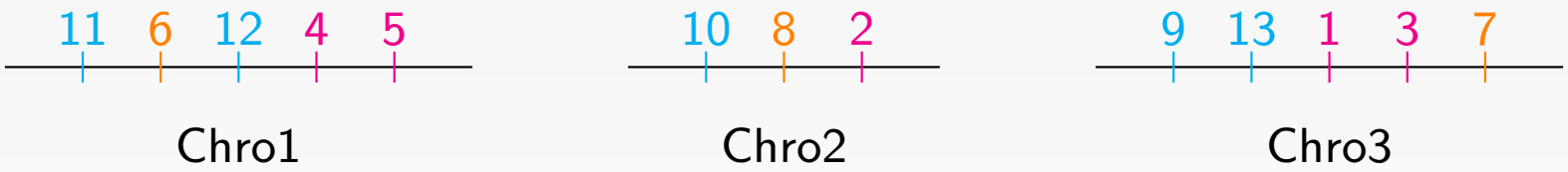


Figure: The genome after a genome permutation

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

Spatial measures - 1 mega-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 genes inside each chromosome.
↪ For *measureIntra* and *measureInter*.

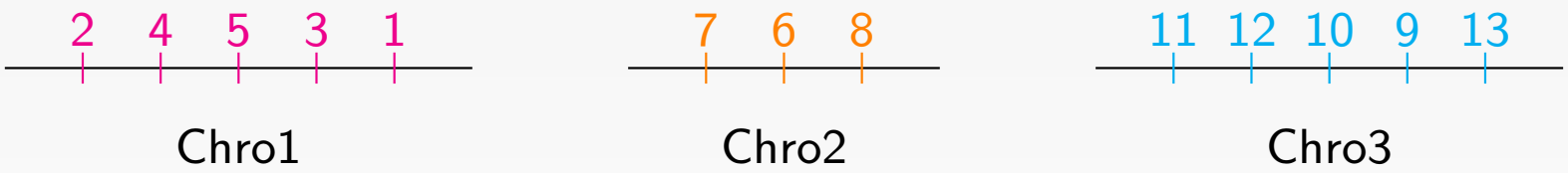


Figure: The genome after chromosome permutations

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

Spatial measures - Results (1 Mega simulations)

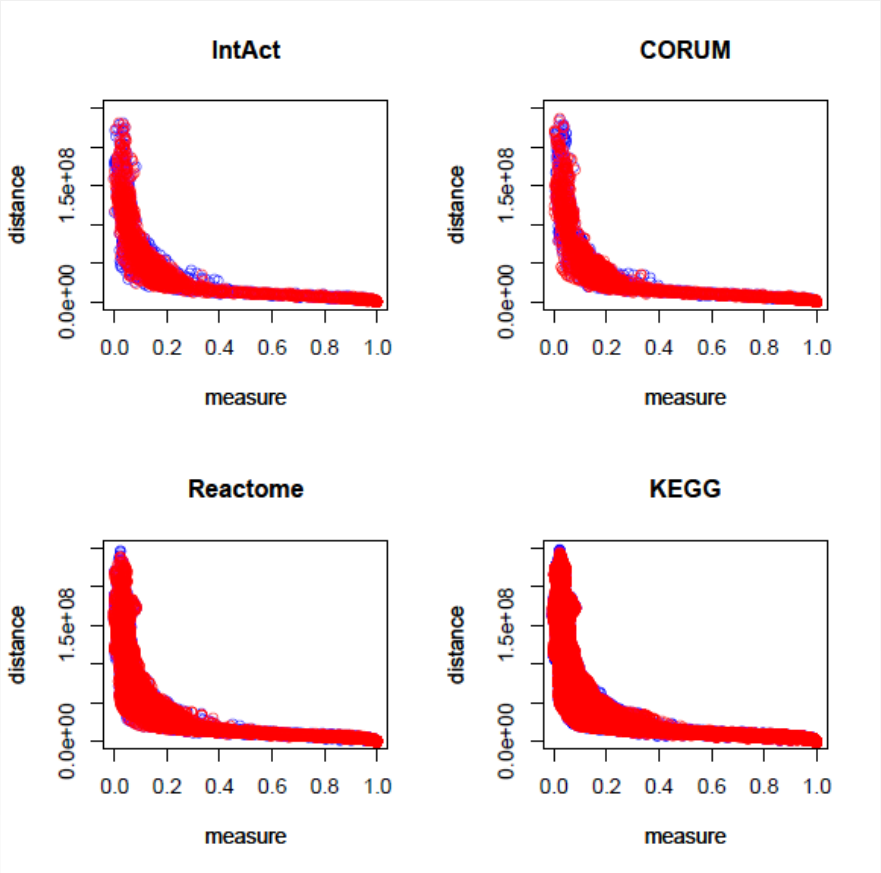
	measureAll	measureIntra	measureInter
Reactome	$< 10E^{-06}$	$< 10E{-}06$	2.81E-01
KEGG	$< 10E{-}06$	$< 10E{-}06$	6.18E-01
CORUM	1.00E-05	1.00E-06	1.05E-01
IntAct	1.50E-05	8.36E-04	1.63E-02

Table: $P(m'_i \leq m)$ for 3 spatial measures of 4 data sets.

Conclusion

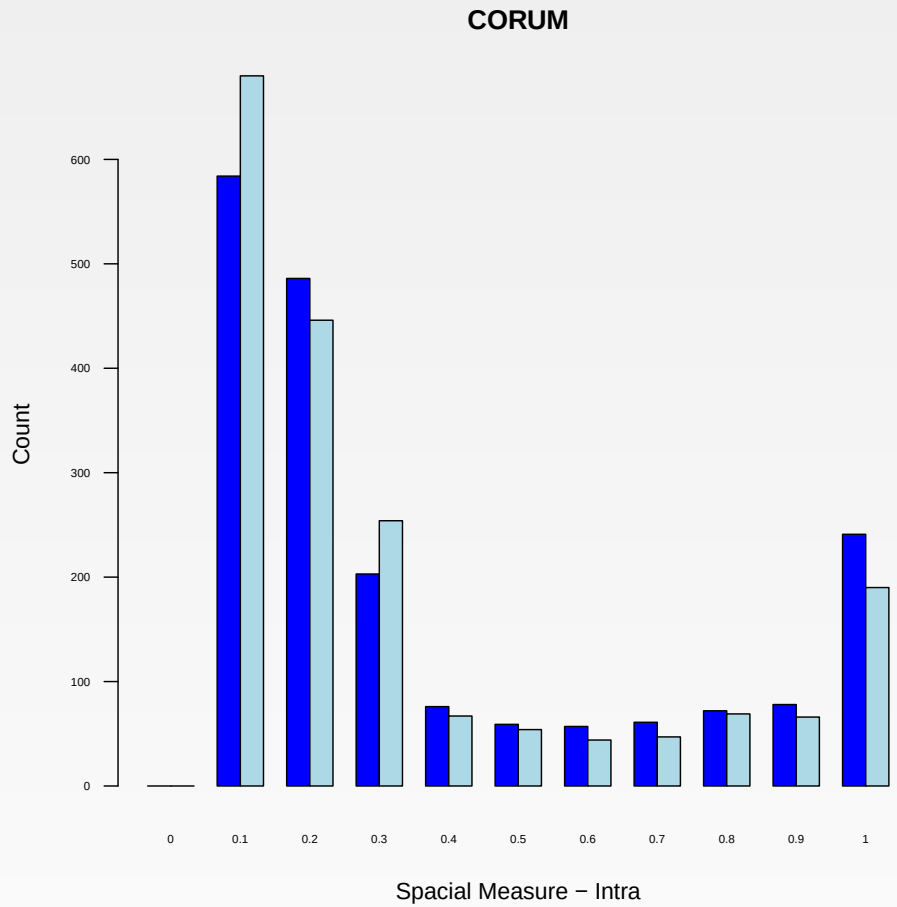
Co-functioning genes are significantly **close** in the space (in particular for genes along same chromosome).

Comparison between intra-chromosomal distance and intra-chromosomal spacial measure



red: Co-functioning pairs of genes, blue: random pairs of genes.

Distribution of spacial measure between pairs of genes



blue: True genome, light blue: random genome.

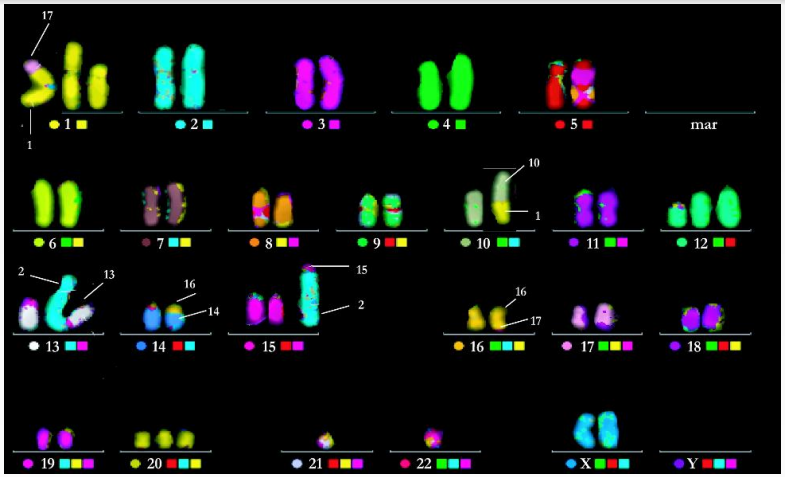
V - Chromosome pairs and cancer

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.

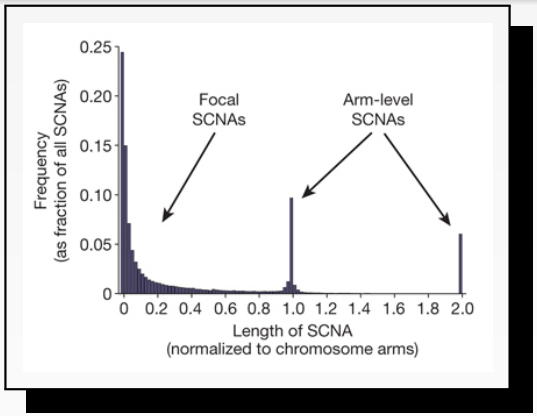


Cancers

Definitions

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

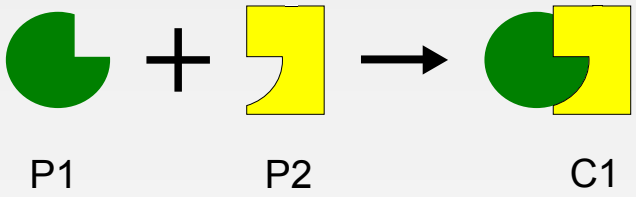
Cancer The cells are distancing themselves from the lineage of origin - malignant tumor.

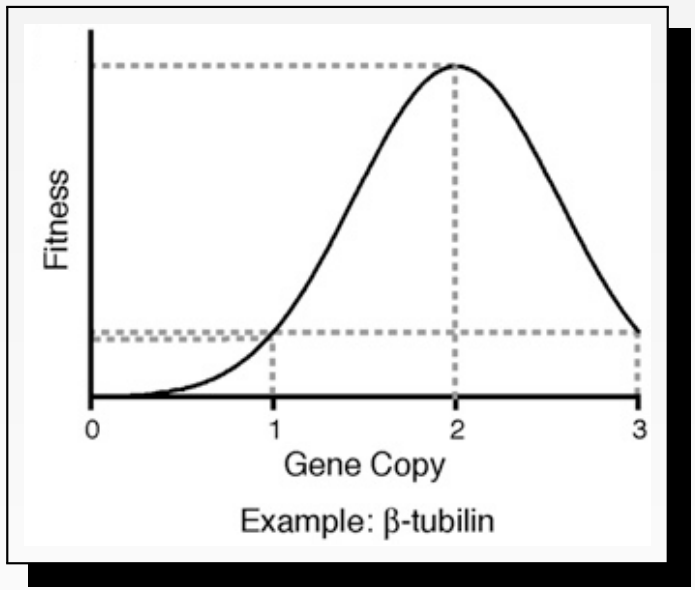


Nature 2009, Beroukhim *et al.*

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]

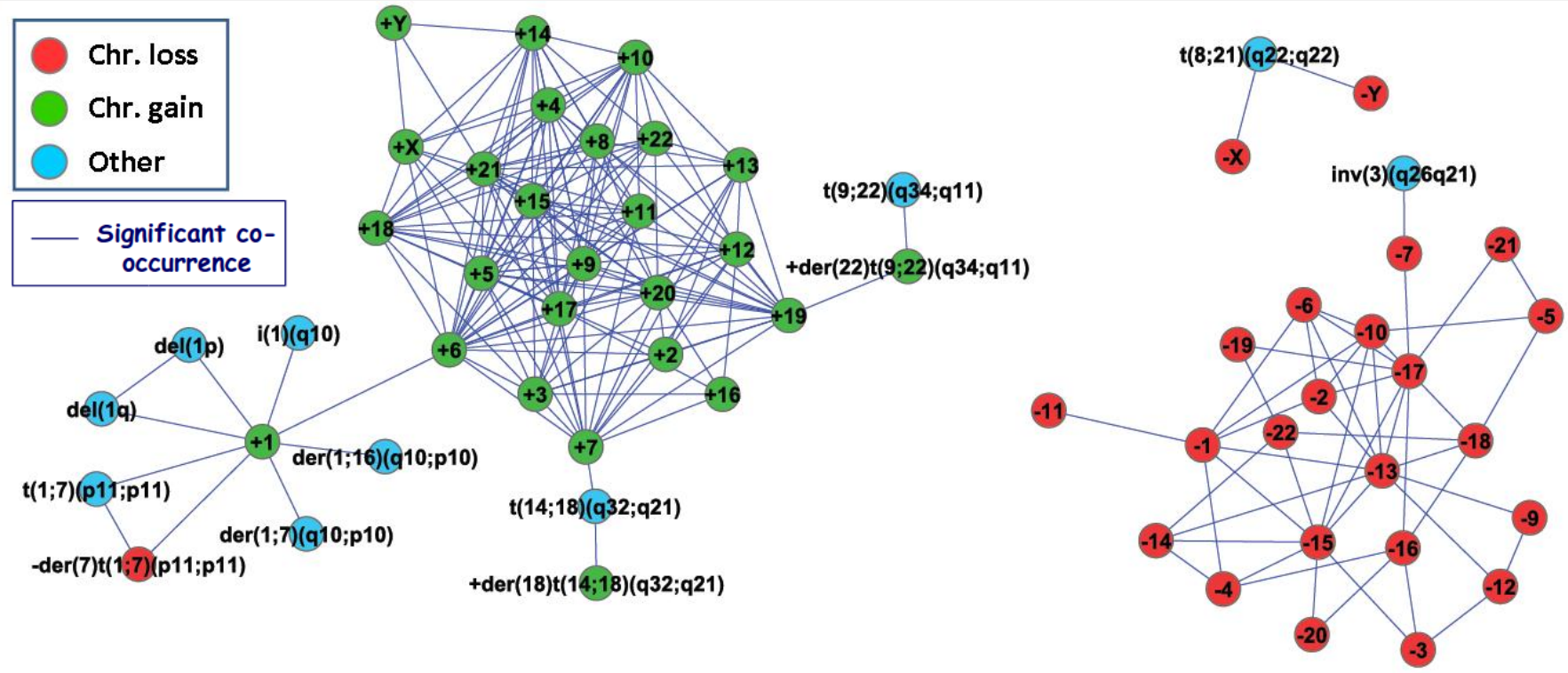
STACK - STatistical Associations in Cancer Karyotypes

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. 2011.

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.

STACK - STatistical Associations in Cancer Karyotypes



Data - STACK

- For each pair of chromosomes, a **P-value** expresses its tendency to be:
  - co-gained or lost (Both),
  - co-gained (Gain) or
  - co-lost (Loss).
- **Cancer type specificity**:
  - For all tissues (All) or for one type of cancer:
  - Hematology (Hema),
  - Lymphoma (Lymph),
  - Solid (Solid).

Data - Number of crossing interactions

For each pair of chromosomes, we compute the number of PPI (from `IntAct`) with one gene in each chromosome, i.e. the **number of crossing interactions**.

We compute this number for the Human genome and for 10^6 randomly permuted genomes to obtain a **P-value** which expresses statistical significance of the number of crossing interactions between the two chromosomes.

We want to evaluate the **correlation** between vectors of P-values from STACK and this new vector of P-values.

Result: Correlations

- **positive correlation** for chromosome gains in solid tumor (Spearmans correlation = 0.23, p-value = 0.0002),
- **negative correlation** for chromosome gains in lymphoid disorders (Spearmans correlation= -0.22, p-value = 0.0004).

Second test: GSEA method.
Enrichment of the set of most-significantly co-gained (resp. co-lost) chromosome pairs among the most functionally related chromosomes.

Results: Co-gained chromosome pair in solid tumors

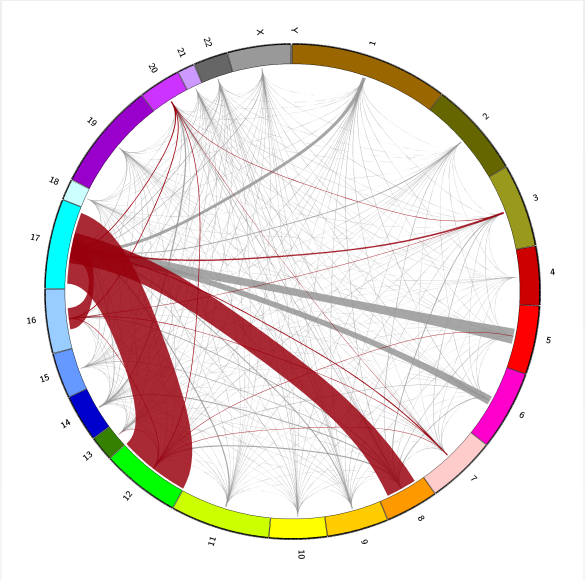


Figure: Illustration of the functional linkage between chromosomes, with a red highlight of the 16 most-significantly co-gained chromosome pairs in solid tumors. **Links width** corresponds to the strength of their functional linkage.

For solid tumors, the set of most-significantly co-gained chromosomes is **remarkably enriched** among the highly functionally-related chromosomes (p-value $< 1E^{-08}$).

Conclusions

Conclusions

We study the organization of co-functioning genes in the Human genome.

Conclusion

- Co-functioning genes tend to involve a **small** number of chromosomes.
- Co-functioning genes tend to be **close** to each other along chromosome and in the space.
- Correlation between **co-gained in solid tumors** and number of **crossing interactions**; depending on cancer type.

Prospects

- Take into account **tandem duplicated genes**.
- Study of **specific** set of co-functioning genes.
- Analysis of **additional functional relations** (e.g. miRNA, expression, transcriptional network).
- Similar analysis in different **species**.
- Integrating the results into a **unified model** explaining gene organization.