A Variational Model for Joint Segmentation of Copy Number Data

Sandro Morganella, Michele Ceccarelli
University of Sannio
Biogem, Bioinformatis Lab

CNA: copy number alterations

- **Copy Number** (CN): The number of times a segment of DNA is repeated throughout a genome

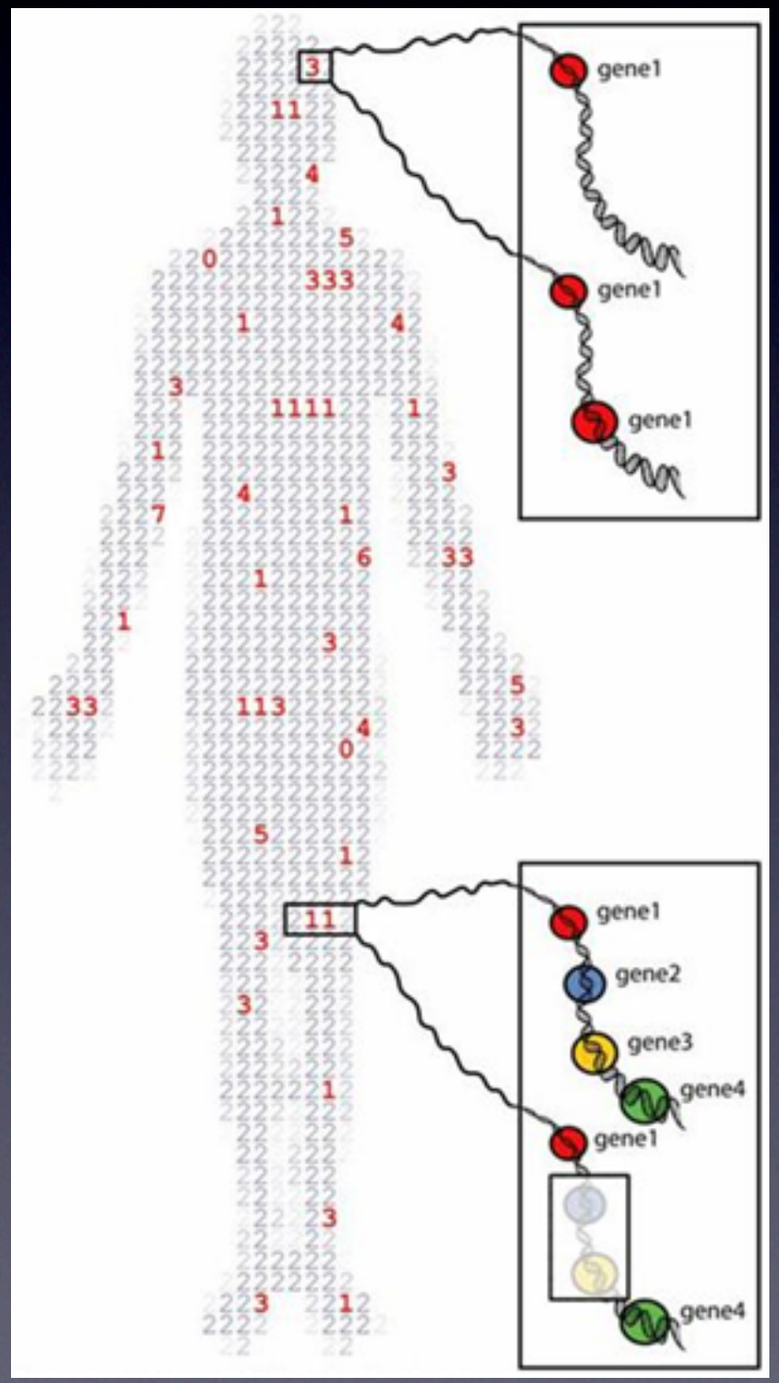
- **Humans are diploid**: cells have two homologous copies of each chromosome (and consequently of each gene)

- CNAs are defined as genomic regions larger than 1 kb in which copy number differences are observed between two or more genomes:

  - **Deletion (loss):** chromosomal region with a CN less then 2

  - **Amplification (gain):** chromosomal region with a CN greater then 2

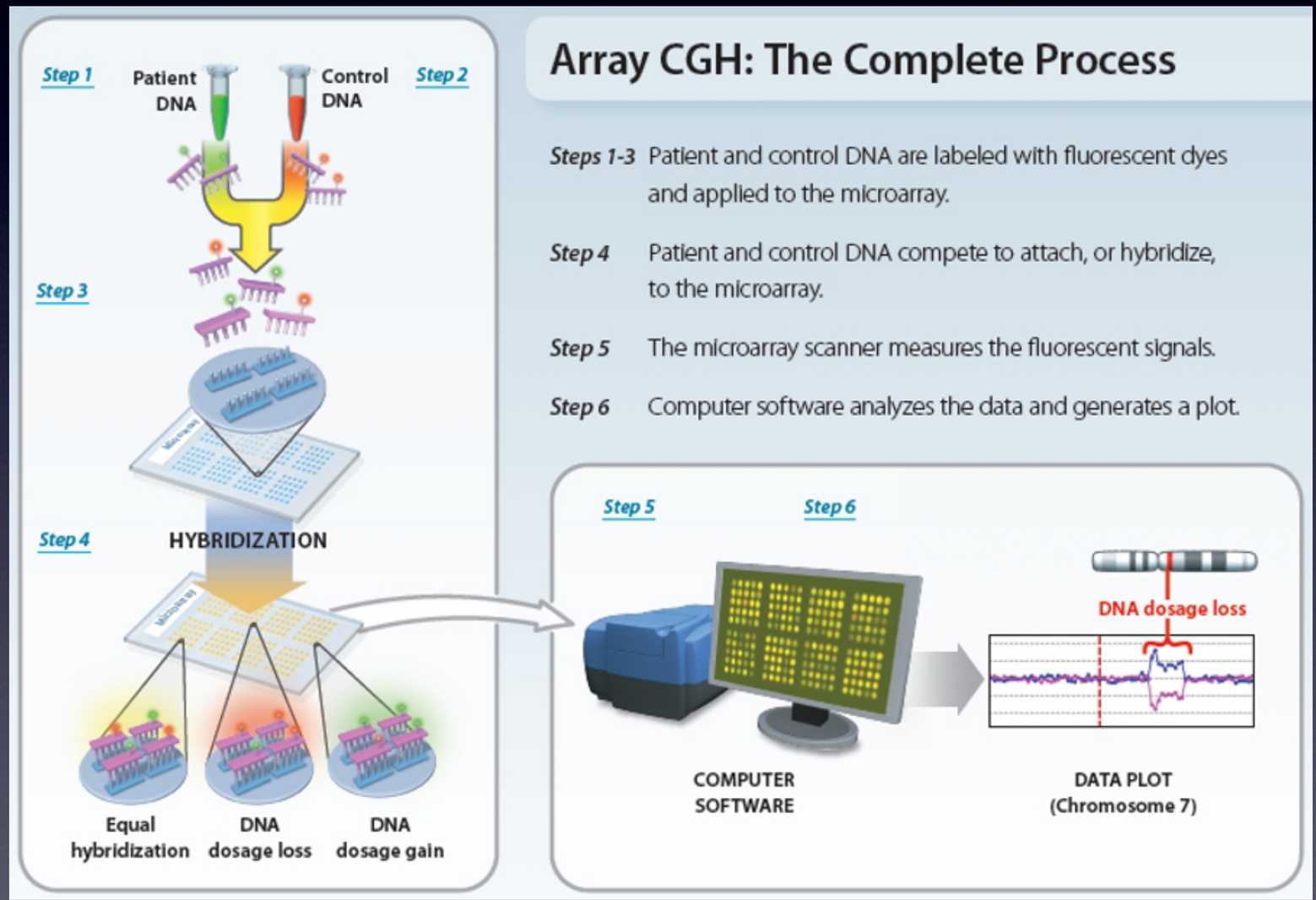
- It was observed that **oncogenes** are often located in regions that show a gain in their copy number, in contrast, **oncosuppressor genes** are found in lost chromosomal regions



Array Comparative Genomic Hybridization

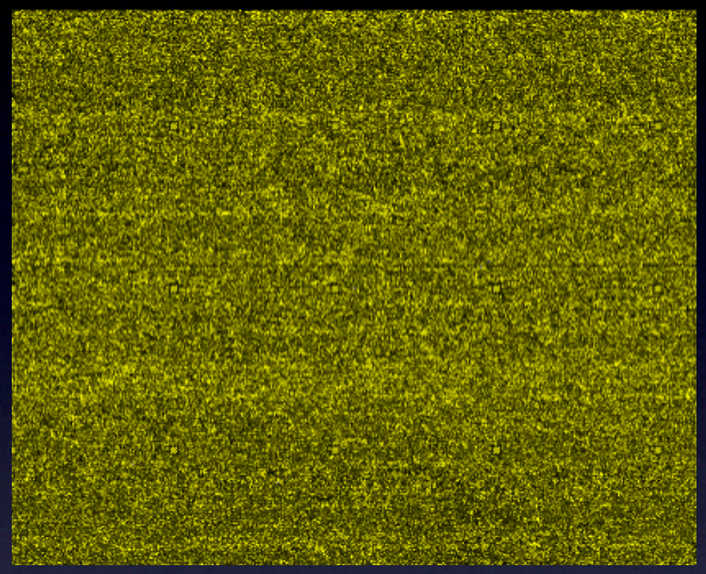
- aCGH technology enables the monitoring of changes at DNA level for more than one million of chromosomal loci (**probes**) of a genome

- In particular, aCGH provides an indirect measure of copy number for each probe, this measure is known as Log R Ratio (**LRR**) and it is computed by the ratio of observed to expected hybridization intensities



SNP Array

Tumor



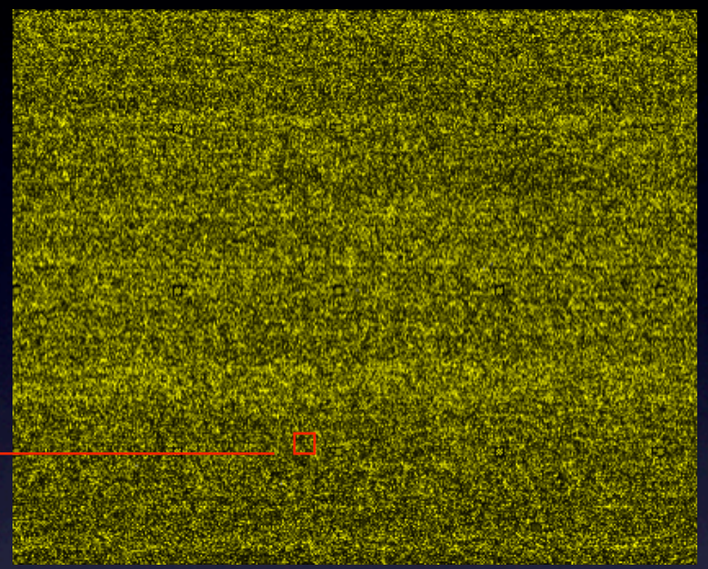
Affymetrix Mapping 250K
Sty-I chip

~250K probe sets

~250K SNPs

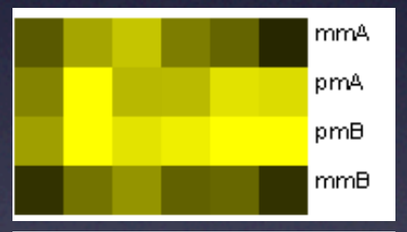
probe set (24 probes)

Normal



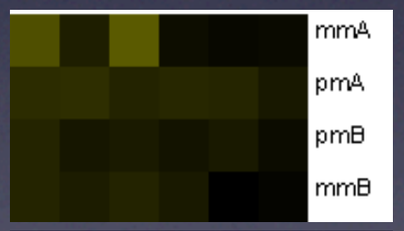
CN=1

Deletion



CN=2

CN=0



Deletion

CN=2

CN>2

Amplification

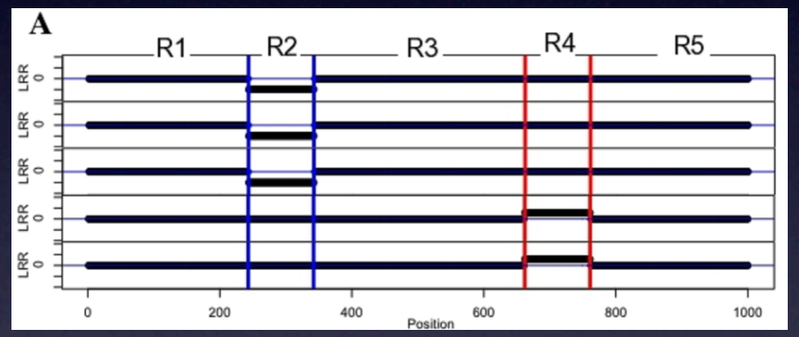
CN=2

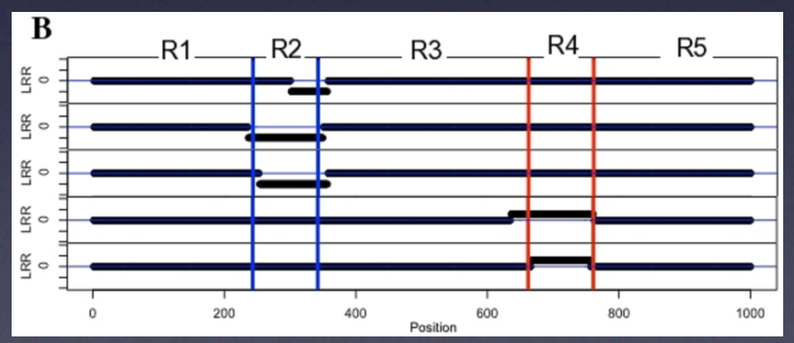
more DNA copy number ⟺ more DNA hybridization ⟺ higher intensity

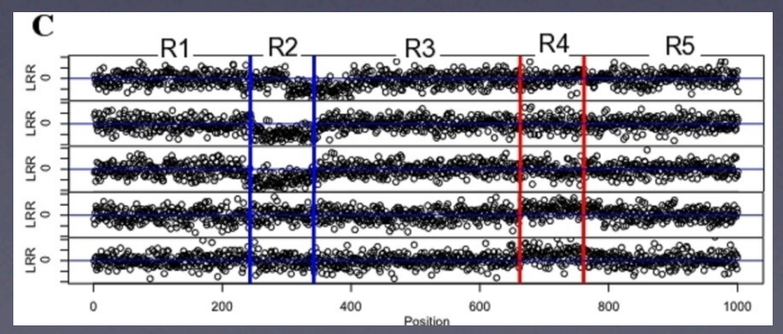
The Problem:

Identification of CNAs shared among a cohort of subjects

- Assumption: Many samples of the dataset reflect the copy number structure of a given disease. Therefore, by a joint analysis of these samples we can pursue the aim of the identification of the recurrent CNA signature of this disease.







Suppose that we have the dataset depicted in Figure A. This dataset is composed of five samples. The first three samples show a loss around the position 300, whereas, the last two samples have a gain around the position 700. So, in this dataset we can distinguish five regions.

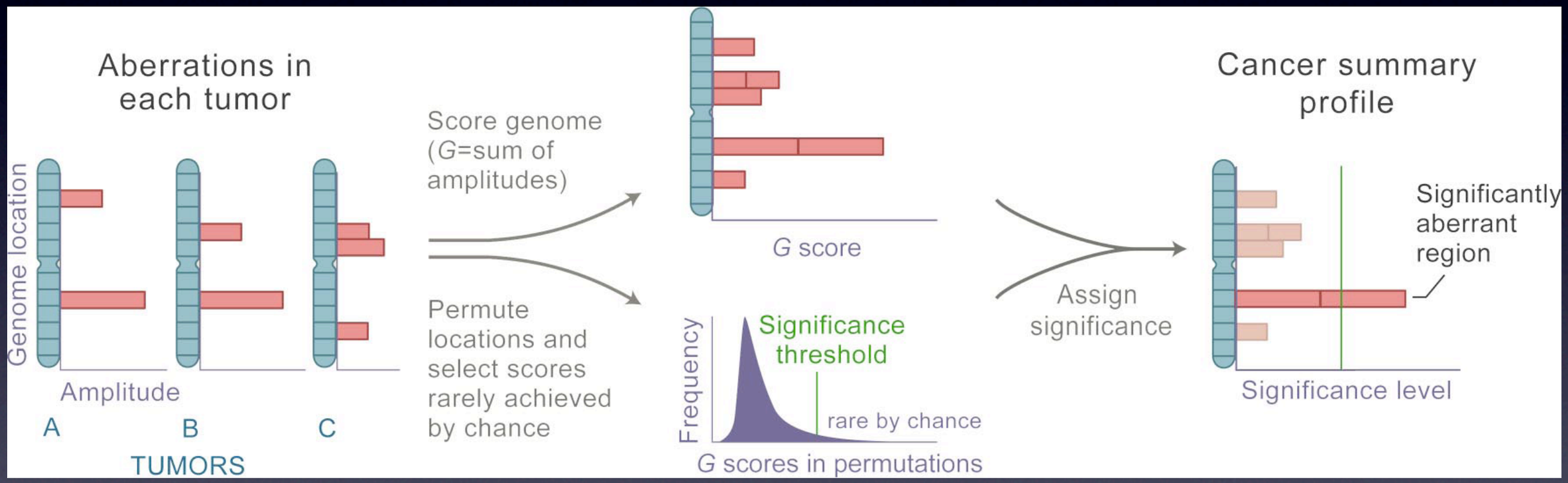
<u>Obstacles for accurate detection of CNAs</u>

- **Biological Noise:** This kind of perturbation is frequent in real data, and it can be due by the mix of tumor and normal tissue specific of each sample
  - CNAs have different position in each sample (Fig B)

- **Experimental Noise:** observed LRR is the ratio of two fluorescence intensities and the measurement process is highly affected by noise
  - Fluctuation of the LRR values (Fig C)

- **Number of probes that have to be analyzed**
  - for example Affyemtrix Genome-Wide Human SNP Array 6.0 produces ~ 1.8 million of probes

Basic (Gistic) idea



Analysis Framework

raw data

e.g. Affymetrix GenomeWide Human SNP arrays

normalizarion and calculation of LLRs

e.g. Affymetrix GenomeWide Human SNP arrays

Wang *et al.:* PennCNV: an integrated hidden Markov model designed for high resolution copy number variation detection in whole-genome SNP genotyping data, *Genome Research* 2007

segmentation of samples

e.g. VEGA algorithm

Morganella S. Ceccarelli M., **VEGA: variational segmentation for copy number detection**, *Bioinformatics* 2010

identification of recurrent CNA

e.g. GAIA algorithm

Morganella S. Ceccarelli M., **Finding recurrent copy number alterations preserving within-sample homogeneity**, *Bioinformatics* 2011

Analysis Framework

raw data

e.g. Affymetrix GenomeWide Human SNP arrays

↓

normalizarion and calculation of LLRs

e.g. Affymetrix GenomeWide Human SNP arrays

Wang *et al.:* PennCNV: an integrated hidden Markov model designed for high resolution copy number variation detection in whole-genome SNP genotyping data, *Genome Research* 2007

↓

segmentation of samples

e.g. VEGA algorithm

Morganella S. Ceccarelli M., **VEGA: variational segmentation for copy number detection**, *Bioinformatics* 2010

↓

identification of recurrent CNA

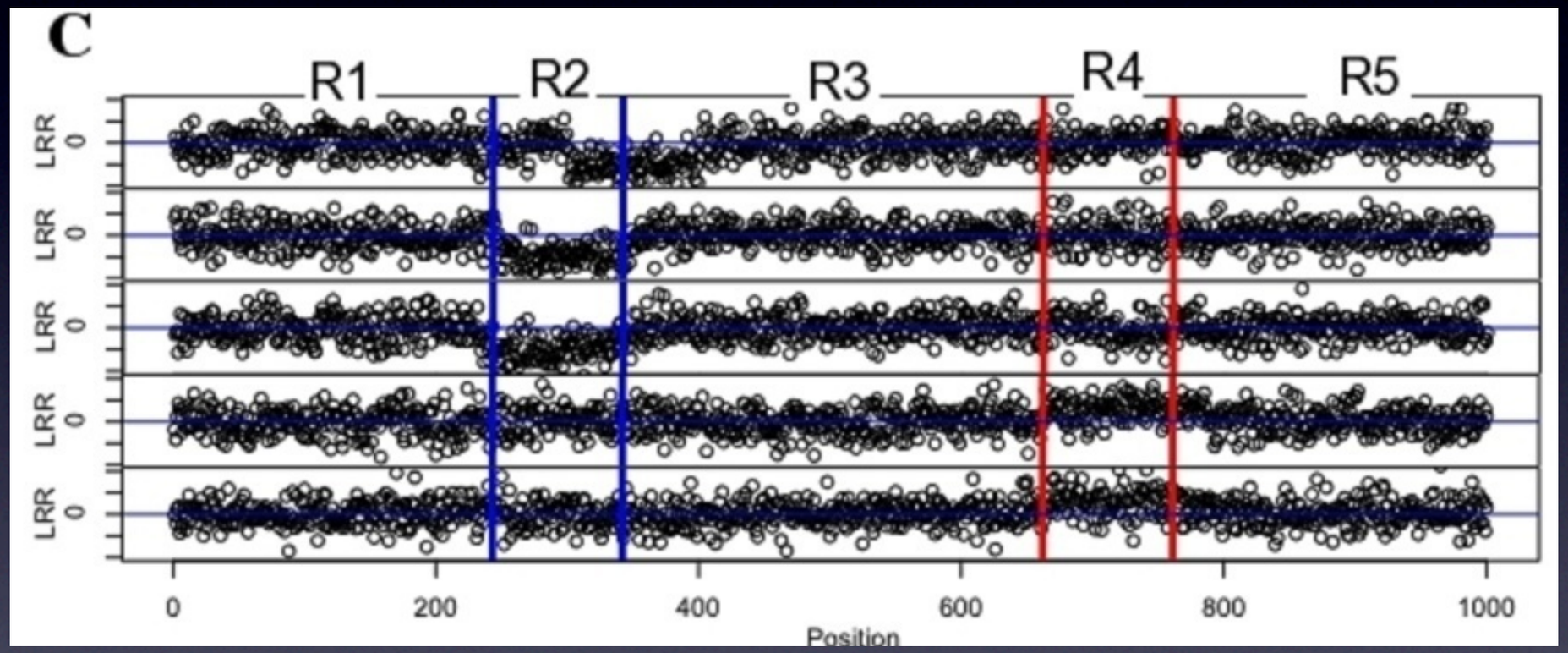
e.g. GAIA algorithm

Morganella S. Ceccarelli M., **Finding recurrent copy number alterations preserving within-sample homogeneity**, *Bioinformatics* 2011

vegaMC:

Joint segmentation of all samples

Detection of recurrent CNA: a segmentation problem



$$LRR = \log \frac{I_{sample}}{I_{control}}$$

The Mumford-Shah Model

- given a multivalued function g defined over a domain Ω, find an approximation u of g over a partition:

$$\Omega = \Omega_1 \cup \Omega_2 \cup \ldots \Omega_n$$

- in order to minimize

$$U(u, \Gamma) = \alpha \int_{\Omega} (u - g)^2 dx\ dy + \int_{\Omega \setminus \Gamma} |\bigtriangledown u|^2 dx\ dy + \lambda |\Gamma|$$

Piecewise Constant Mumford-Shah Model

u_i is a vector of piecewise constant functions

$$U(u,\Gamma) = \sum_i \int_{\Omega_i} (g - u_i)^2 dx\ dy + \lambda|\Gamma| \qquad u_i = \frac{1}{|\Omega_i|}\int_{\Omega_i} g$$

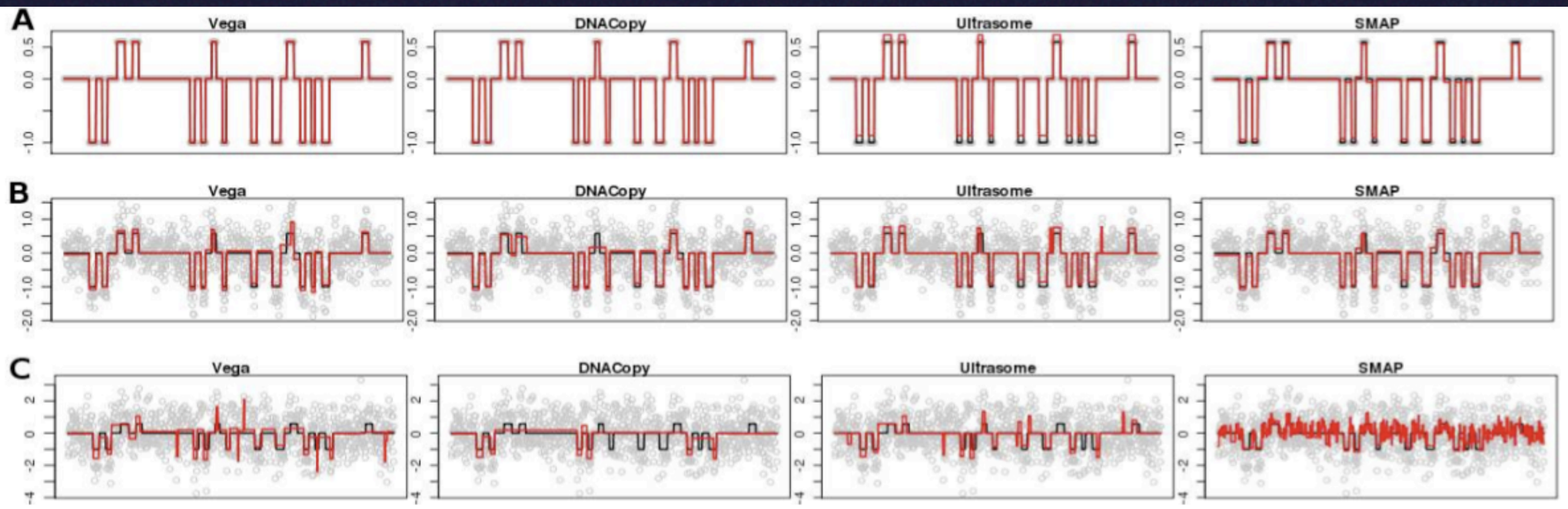


Fig. 2. Segmentation results on a simulated chromosome profile of 1 000 probes. Red line reports the segmentation computed by the algorithm, black line is the true ground. In **A** the results on the simulated data without the noise component. In **B** the results on the simulated data with Gaussian noise, zero mean and variance equal to 0.5. In **C** the results on the simulated data with Gaussian noise, zero mean and variance equal to 1.

Variational Segmentation algorithm

- Greedy region growing
  - small regions are progressively merged to create larger ones
  - Energy differential after merging

$$U(u, \Gamma \setminus \Omega_i \cup \Omega_j) - U(u, \Gamma)) = \frac{|\Omega_i| \ |\Omega_j|}{|\Omega_i| + |\Omega_j|} ||u_i - u_j||^2 - \lambda$$

  - Steepest descent region growing

Steepest Descent minimization

1. Start with the finest segmentation and set $\lambda=0$
2. Choose the next pair or regions to be merged producing the maximum decrease of the energy function

$$E(u, \Gamma \setminus R_i \cup R_j) - E(u, \Gamma)) = \frac{|R_i|\ |R_j|}{|R_i| + |R_j|} ||\mathbf{u_i} - \mathbf{u_j}||^2 - \lambda$$

3. If no pair of region exists producing a decrease of energy, then increase λ
4. Go to 2 until convergence

λ-Schedule

- λ-schedule is the sequence of λ's used in the minimization process
- The cost required for merging of two adjacent regions R_i and R_{i+1} is

$$\hat{\lambda}_i = \frac{|R_{i-1}|\,|R_i|}{|R_{i-1}| + |R_i|} ||\mathbf{u_{i-1}} - \mathbf{u_i}||^2$$

  - λ-update: the smallest available among all pairs
- **Stopping Criterion:**

$$\frac{\Delta\,\lambda}{m} \leq \beta\ ||\nu||$$

$$\Delta\,\lambda = \overline{\lambda}_{l+1} - \overline{\lambda}_l$$

Standard deviation

Identification of aberrant regions

- After segementation, we need to classify each region as normal or aberrant (with its subclasses)

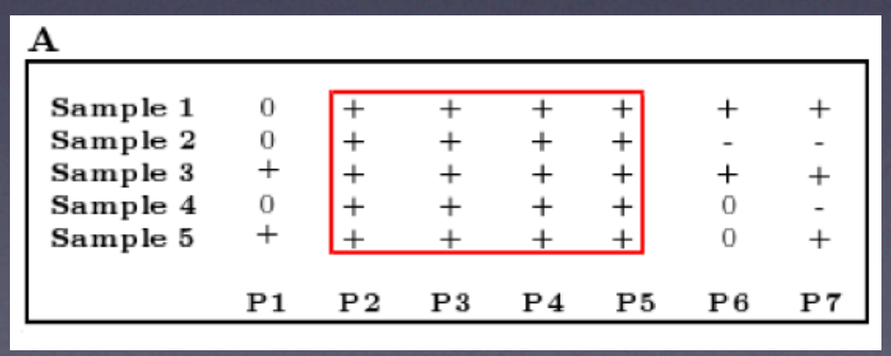
$$Label(R_i) = \begin{cases} \text{loss} & \text{if } ||\mu_i|| < \tau_{loss} \\ \text{gain} & \text{if } ||\mu_i|| > \tau_{gain} \\ \text{normal} & \text{otherwise} \end{cases}$$

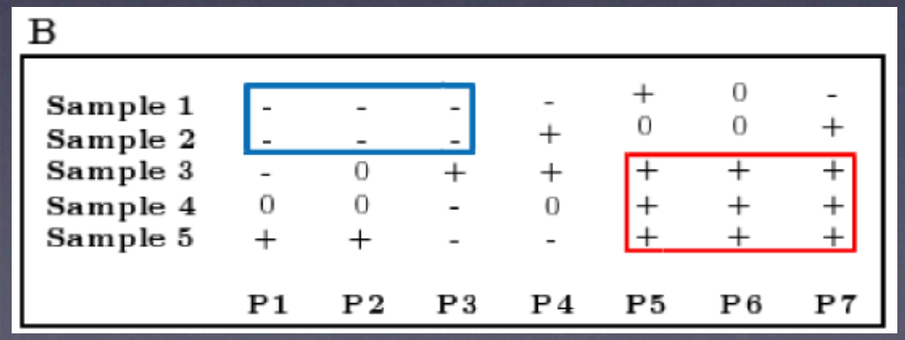
- where $||\mu_i||$ is the L_2 norm of the PWC approximating function in the i-th region

τ_{loss} = -0.2, τ_{gain} = +0.2

Simulated Data

- Strategy for Generation of Synthetic Data
  - Simulation of two fundamental CNA patterns (Figure A and B)
  - Chromosome size of 1000 probes
  - Simulation of different resolution scenarios by increasing CNA widths
- Models for Data Perturbation
  - Dataset 1: **Intensity Noise,** perturbs the data as a white Gaussian process ~ $\mathcal{N}(0, \sigma)$
  - Dataset 2: **Intensity + Spatial Noise**, in addition randomically resizes and move the boundaries of CNAs
- Data available in GAIA home page: http://bioinformatics.biogem.it/download/gaia





Considered approaches for comparison

- GAIA: Morganella and Ceccarelli: Finding recurrent copy number alterations preserving within-sample homogeneity, *Bioinformatics* 2011
  - Uses as input a discrete representation of the observed LRRs
  - statistical framework based on a conservative permutation test
  - CNAs having a high evidence to be sites of CNAs are extracted by an iterative procedure known as peel-off where both statistical significance and within-sample homogeneity are considered
- GADA: Pique-Regi et al.: Joint estimation of copy number variation and reference intensities on multiple DNA array using GADA, *Bioinformatics* 2009
  - decomposition of the observed LRR in three components
  - Based on the PWC assumption uses an expectation maximization framework to jointly estimate all three components
- JISTIC: Sanchez-Garcia et al. : JISTIC: Identification of Significant Targets in Cancer, *BMC Bioinformatics* 2010
  - Uses as input a smoothed representation of the observed LRRs
  - statistical framework based on a conservative permutation test
  - CNAs having a high evidence to be sites of CNAs are extracted by peel-off
- cghMCR: Aguirre et al.: High-resolution characterization of the pancreatic adenocarcinoma genome, *PNAS* 2004
  - Uses as input a smoothed LRRs
  - Smoothed data are used to distinguish between normal and altered probes by a percentile-based approach

Results on Simulated Data

F-measure: Harmonic mean of *Precision* and *Recall* which capture information on the completeness and of exactness of the results

Results on Gastrointestinal Stromal Tumor (GIST)

Cytoband	Expected Alteration	$\|\|\mu\|\|$	Computed Alteration	Target Genes	Description
1p36.23	loss	-0.01	normal	*ENO1*	Enolase 1, (alpha)
5q35.3	gain	0.25	gain	*GNB2L1*	Guanine nucleotide-binding protein-β
7p11.2	gain	0.56	gain	-	-
9p21.3	loss	-0.43	loss	*CDKN2A*, *CDKN2B*	Cyclin-dependent kinase inhibitors
12q15	gain	0.01	normal	-	-
13q14.3	loss	-0.35	loss	*INTS6*	Integrator complex subunit 6
14q23.1	loss	-0.35	loss	*PPM1A*	Protein phosphatase 1Aα isoform
14q32.33	loss	-0.38	loss	*PACS2*	Phosphofurin acidic cluster sorting protein
22q12.2	loss	-0.03	normal	NF2	Neurofibromin 2 (merlin)
22q13.31	loss	-0.02	normal	-	-

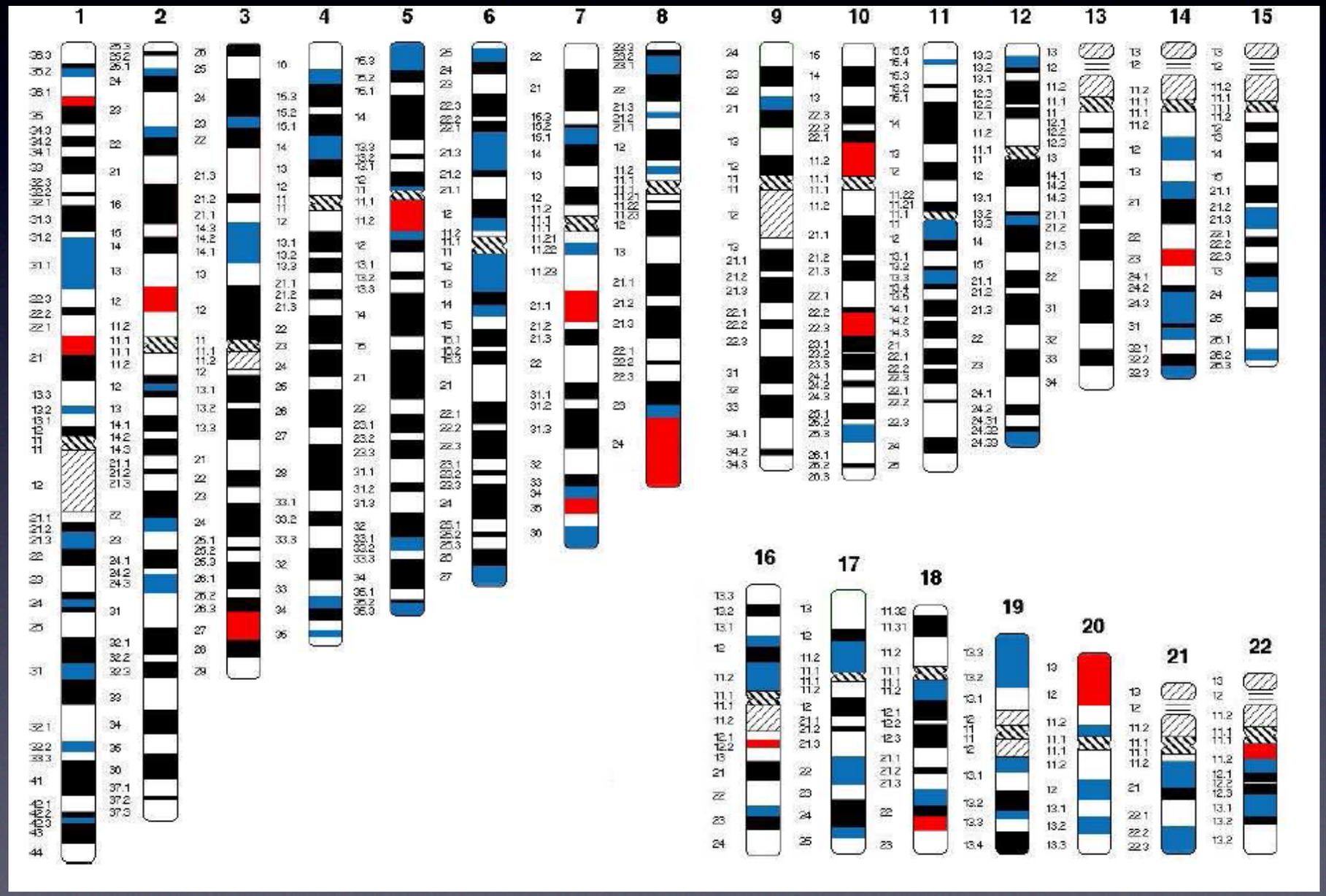
- GISTs are the most common mesenchymal tumors of the gastrointestinal tract
- **25** fresh tissue specimens of GISTs were collected and hybridized by Affymetrix Genome Wide SNP 6.0 (GEO identifier GSE20710)
- Raw data were preprocessed by PennCNV tool obtaining the LRRs for about **1.6 million of probes**
- VegaMC found high aplification of **7p11.2** and low intensities for several target genes: ***CDKN2A, CDKN2B, INTS6, PPM1A*** and ***NF2***.
- The execution time required by VegaMC on this dataset is **1'23"**
  - GAIA: 61'20" - GADA : 38'38" - JISTIC 14'21" - cghMCR 0'35"

Results on Lung Cancer Dataset

Cytoband	Expected Alteration	$\|\|\mu\|\|$	Computed Alteration	Target Genes	Description
3p14	loss	-0.33	loss	*FHIT*	Fragile histidine triad gene
3p21	loss	-0.11	normal	*RASSF1*	Ras association (RalGDS/AF-6) domain family member 1
3q26.33	gain	0.12	normal	*SOX2*	SRY (sex determining region Y)-box 2
6p25.2	loss	-1.14	loss	-	-
8p11.23	gain	0.13	normal	*FGFR1*	Fibroblast growth factor receptor 1
8q24	gain	0.8	gain	*MYC*	v-myc myelocytomatosis viral oncogene homolog (avian)
9p21.3	loss	-0.56	loss	*CDKN2A, CDKN2B*	Cyclin-dependent kinase inhibitors
13q14.2	loss	-0.46	loss	*RB1*	Retinoblastoma 1
20p13	gain	0.31	gain	-	-
22q11.21	gain	0.63	gain	*MAPK1*	Mitogen-activated protein kinase 1

- Lung Cancer is a leading cause of cancer death in industrialized countries
- 155 primary squamous cell lung cancer hybridized by Affymetrix 6.0 SNP arrays (GEO identifier GSE25016)
- Raw data were preprocessed by PennCNV tool obtaining the LRRs for about 1.7 million of probes
- VegaMC found high amplifications of MAPK1 and MYC oncogenes and low intensities of RB1, CDKN2A, CDKN2B and 6p25.2 are important evidences of a well-performed analysis
- Execution time required by VegaMC is 4’35”

Overview of Identified CNAs in Lung Cancer



Availability

- VegaMC as well as Vega and GAIA have been implemented as R/Bioconductor packages
- They can be downloaded from the Bioconductor page:
  http://www.bioconductor.org/
- ... or from the web page of our lab:
  http://bioinformatics.biogem.it/download

Conclusions

- Quantitative and qualitative results show an important aspect of VegaMC: the computed optimal segmentation seems to reflect the "ground truth" of the analyzed data
- This suggests that the copy number structure is well approximated by a PWC function
- Assessment in different noise conditions show that approaches working directly on the original LRRs (as VegaMC and GADA) seem to be more robust with respect to intensity noise than approaches that need of a preprocessing step (as GAIA, JISITC and cghMCR)
- The main advantages of VegaMC are:
  - accuracy increases as the chip resolution increases
  - execution time required to perform the analysis. From our experience we can state that the proposed approach can be considered as one of the fastest algorithm aimed at the identification of recurrent CNAs
- These advantages make VegaMC a valid alternative to other algorithms for high-resolution data analysis

Available Approaches

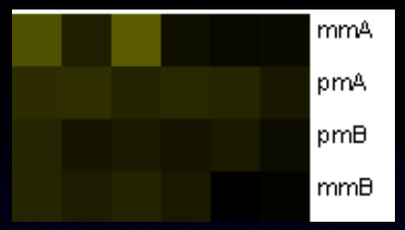
Algorithm	Preprocessing	Input	Output	Availability
VegaMC	No	Original LRR	Weighted mean	R Package
BSA	No	Original LRR	Statistical evidence (q-value)	R scripts
chgMCR	Yes	Smoothed LRR	None	R/Bioconductor Package
CNAnova	No	Original LRR	Statistical evidence (p-value)	R scripts
DiNAMIC	Yes	Smoothed or Discretized Data	Statistical evidence (q-value)	R scripts
GADA	No	Original LRR	None	R Package
GAIA	Yes	Discretized Data	Statistical evidence (q-value)	R/Bioconductor Package
GISTIC	Yes	Smoothed LRR	Statistical evidence (q-value)	Standalone Application based on Matlab
GISTIC 2.0	Yes	Smoothed LRR	Statistical evidence (q-value)	Standalone Application based on Matlab
JISTIC	Yes	Smoothed LRR	Statistical evidence (q-value)	Java Application
KCsmart	No	Original LRR	Statistical evidence (q-value)	R/Bioconductor Package
pREC-A	Yes	Probability of alteration	Probability of alteration	R Package
pREC-S	Yes	Probability of alteration	Probability of alteration	R Package

- Aim of preprocessing step is the identification of breakpoints in the genome that identify regions in which probes have a similar copy number profile (Smoothed LRR) or share the same state (Discretized Data)

- Preprocessing is often performed by a segmentation algorithm

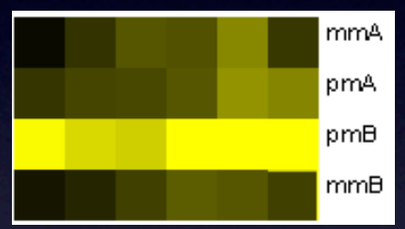
Genotyping and Copy Number Calling

CN=0



2 copy deletion, genotype (_//_)

CN=1



1 copy deletion, genotype (_//B)

CN=2

Normal , genotype (A//B)

CN=3

1 copy amplification, genotype (AA//B)

CN=4

2 copy amplification, genotype (AA//BB)