# A suite of R packages for the analysis of DNA copy number microarray experiments
# Application in cancerology

Philippe Hupé[1,2]

[1]UMR144 Institut Curie, CNRS

[2]U900 Institut Curie, INSERM, Mines Paris Tech

The R User Conference 2009
Rennes




# Outline


1. Biological / clinical context
2. R packages description
3. End-user interfaces / automatic workflow

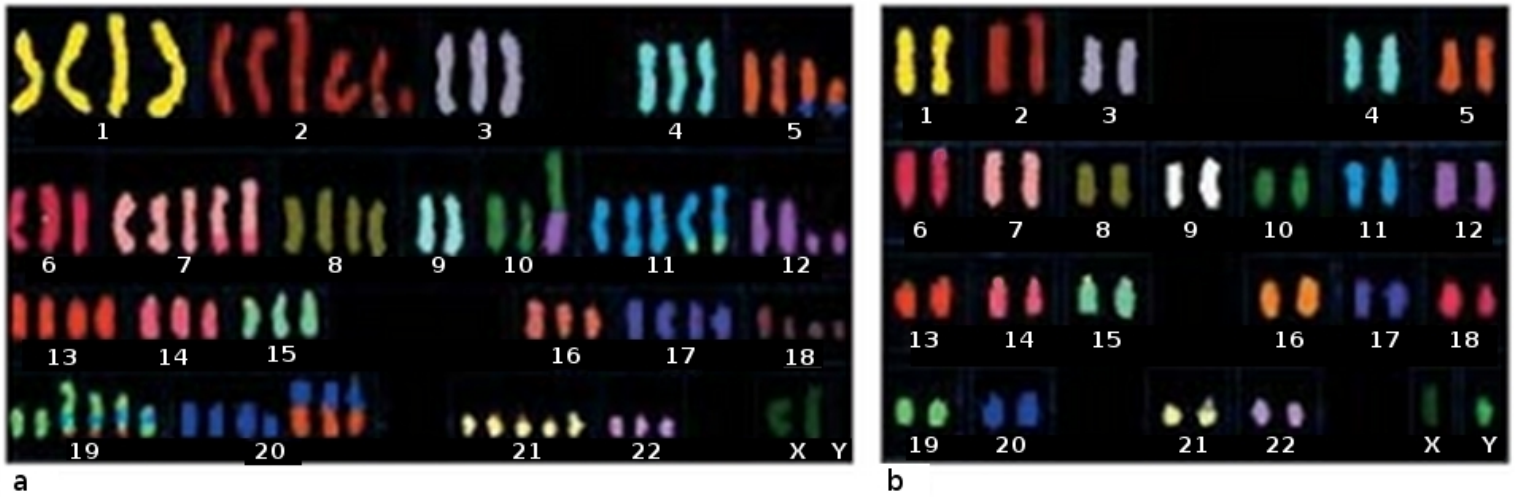
# Outline

1 Biological / clinical context

2 R packages description

3 End-user interfaces / automatic workflow

# DNA copy number alteration in tumour




tumoral cell

normal cell

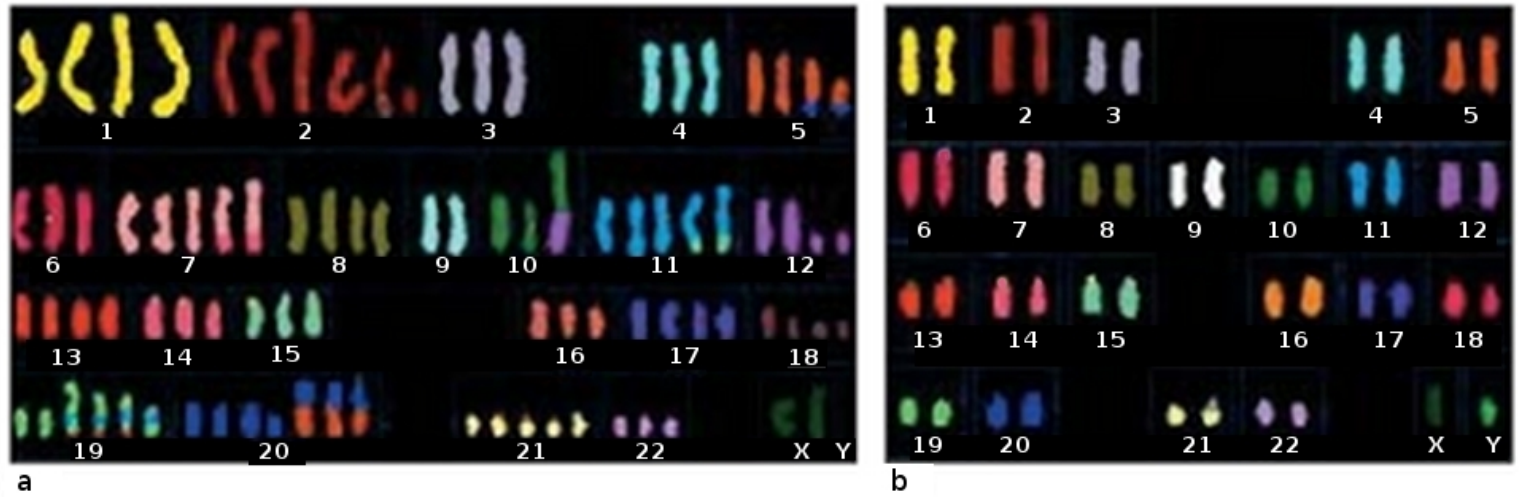
## Chaos in cancer cells

gain, loss or amplification of chromosomes or pieces of chromosomes.

## Molecular profiling of tumours

- Identification of DNA copy number alterations in each patient
- Is the pattern of alterations is related to patient outcome (e.g. relapse, metastasis)?

# DNA copy number alteration in tumour




tumoral cell

normal cell

## Chaos in cancer cells

gain, loss or amplification of chromosomes or pieces of chromosomes.

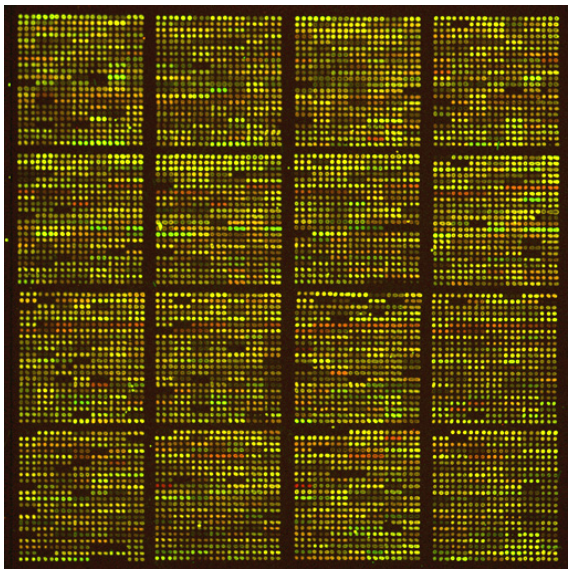
## Molecular profiling of tumours

- Identification of DNA copy number alterations in each patient
- Is the pattern of alterations is related to patient outcome (e.g. relapse, metastasis)?

# High-throughput quantification of DNA copy number

## Microarray technology

- DNA copy number for $5 \times 10^3$ up to $2 \times 10^6$ genomic loci
- Probes spotted on a glass array (i.e. the microarray)



microarray

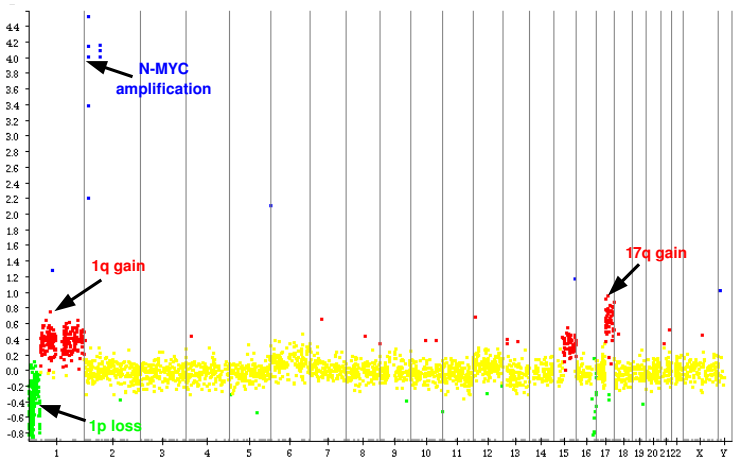
Colour study: squares with concentric circles

*Wassily Kandinsky*, 1913

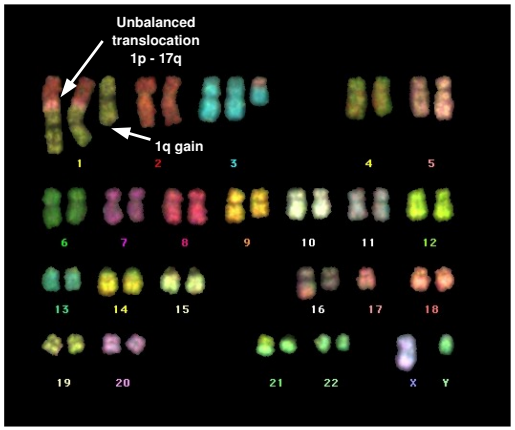
# High-throughput quantification of DNA copy number

## Microarray technology

- DNA copy number for $5 \times 10^3$ up to $2 \times 10^6$ genomic loci
- Probes spotted on a glass array (i.e. the microarray)




DNA copy number profile of the tumour

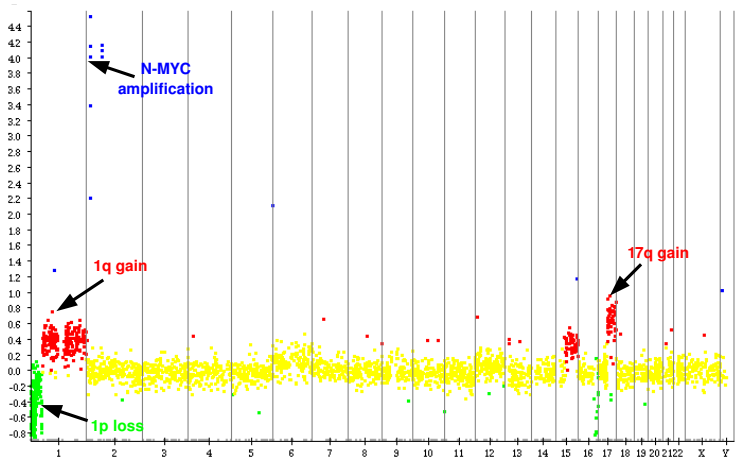



Karyotype of the tumour

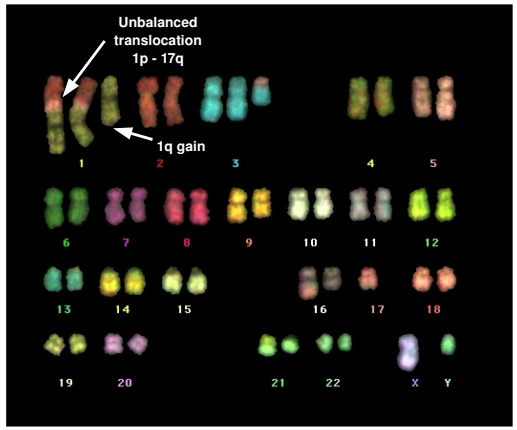
# High-throughput quantification of DNA copy number

## Microarray technology

- DNA copy number for $5 \times 10^3$ up to $2 \times 10^6$ genomic loci
- Probes spotted on a glass array (i.e. the microarray)




DNA copy number profile of the tumour

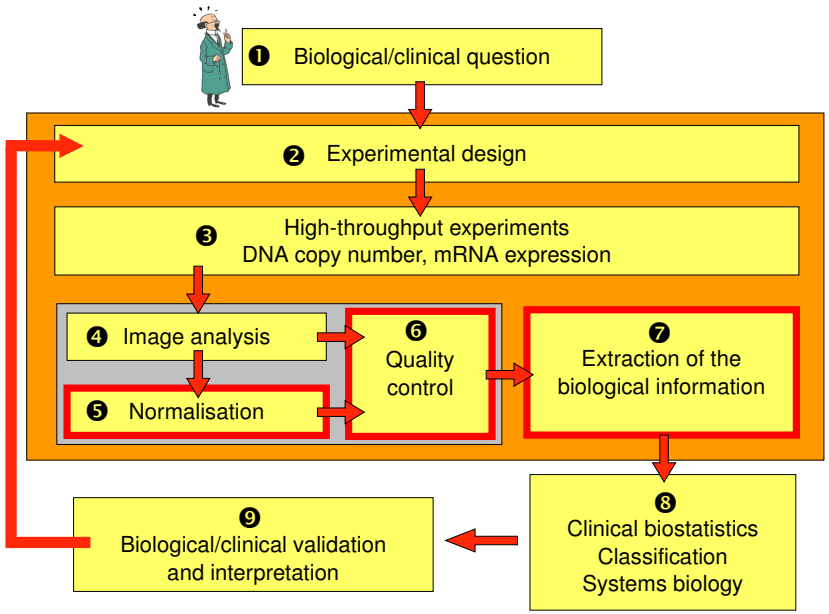



Karyotype of the tumour

## Huge amount of data ($\sim 2 \times 10^6$ variables for each patient)

Need for biostatistical algorithms and automatic bioinformatic pipelines

# Biostatistical workflow




## R packages available from www.bioconductor.org

- MANOR: spatial normalisation
- GLAD: extraction of the biological information
- ITALICS: normalisation + extraction of the biological information

# Outline

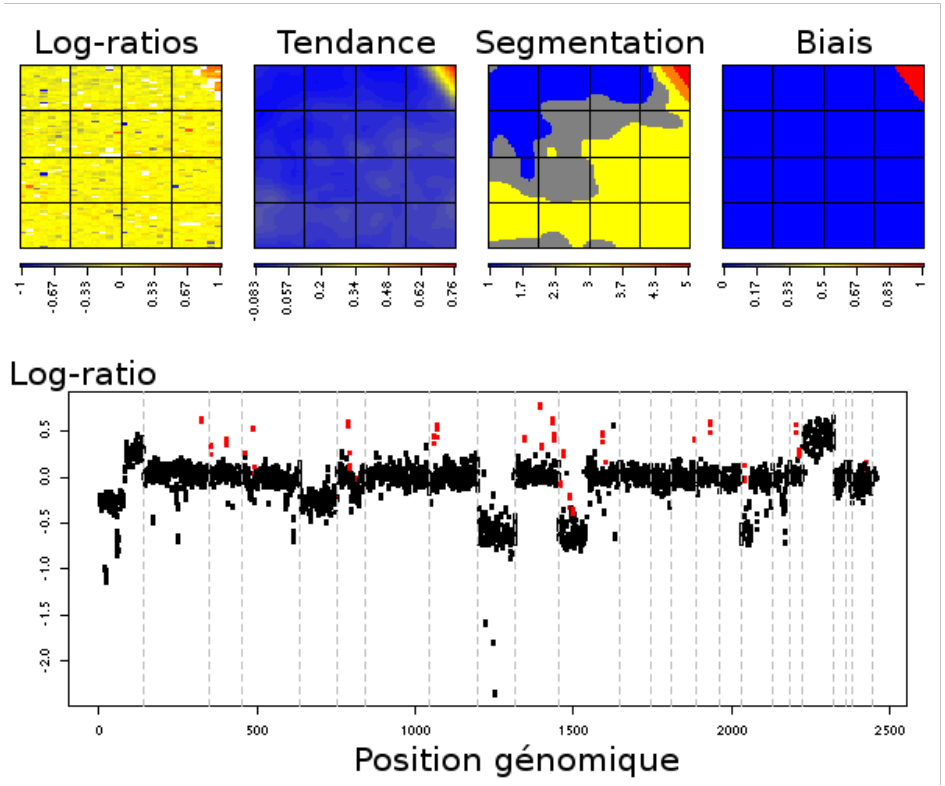
1 Biological / clinical context

2 R packages description

3 End-user interfaces / automatic workflow

# MANOR: an algorithm to detect spatial bias

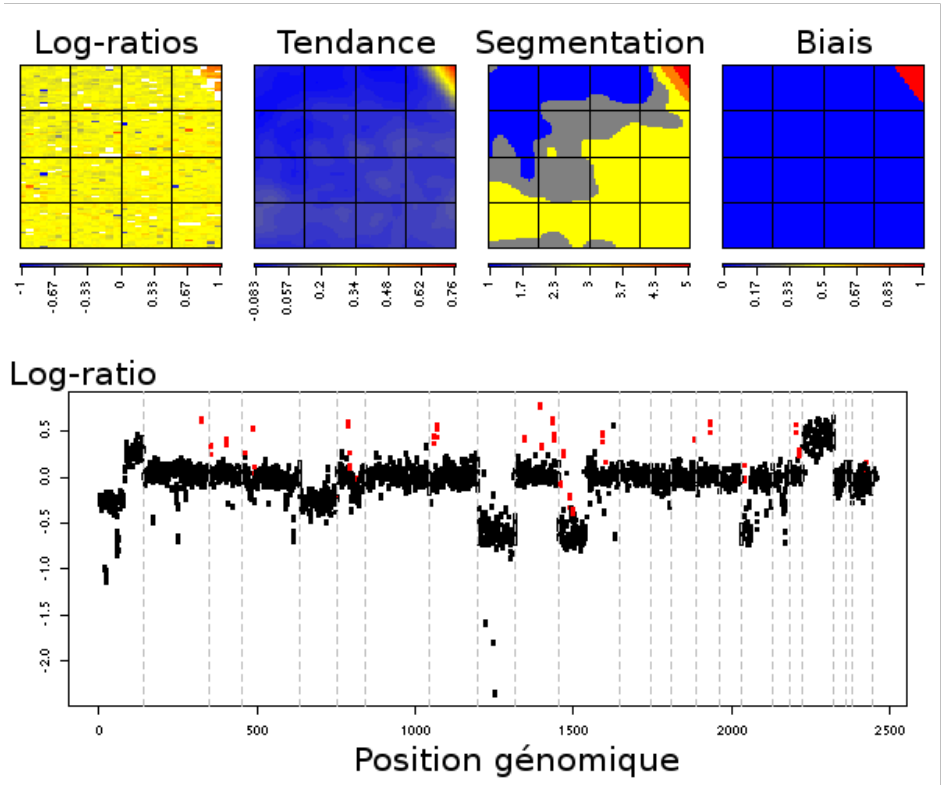
## Neuvial et al., BMC Bioinformatics, 2006




1. Abnormal Log-Ratio in the corner
2. Spatial trend estimation by 2D-LOESS
3. Spatial segmentation
4. Bias area are removed
5. Spots are **outliers** in the genomic profile

# MANOR: an algorithm to detect spatial bias

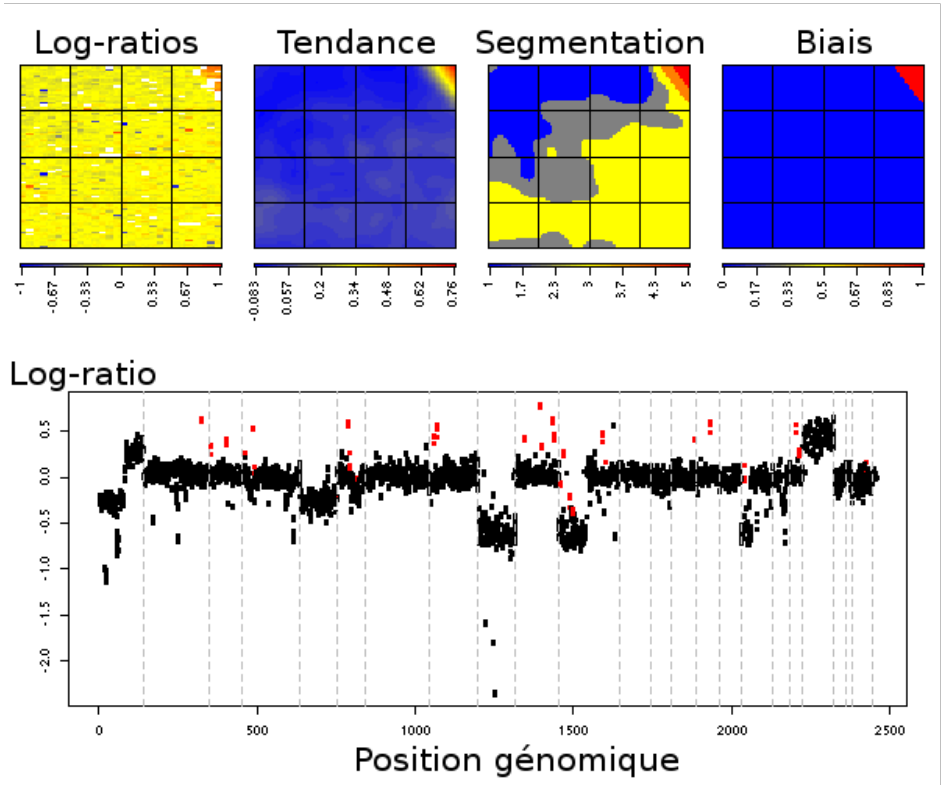
Neuvial et al., BMC Bioinformatics, 2006




1. Abnormal Log-Ratio in the corner
2. Spatial trend estimation by 2D-LOESS
3. Spatial segmentation
4. Bias area are removed
5. Spots are **outliers** in the genomic profile

# MANOR: an algorithm to detect spatial bias

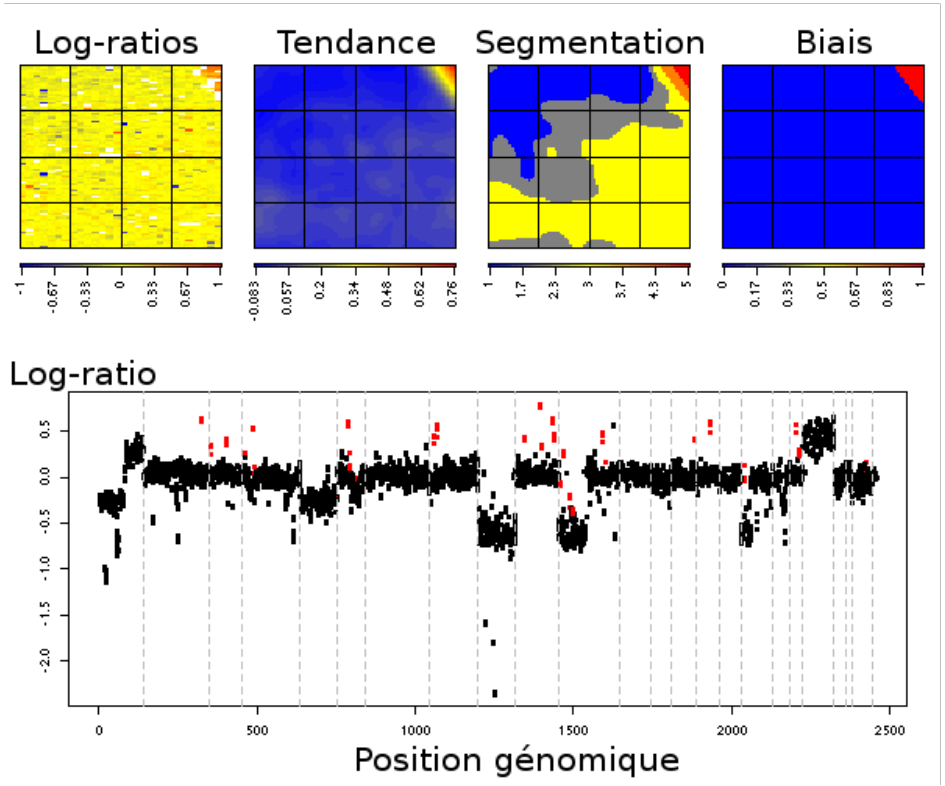
Neuvial et al., BMC Bioinformatics, 2006




1. Abnormal Log-Ratio in the corner
2. Spatial trend estimation by 2D-LOESS
3. Spatial segmentation
4. Bias area are removed
5. Spots are **outliers** in the genomic profile

# MANOR: an algorithm to detect spatial bias

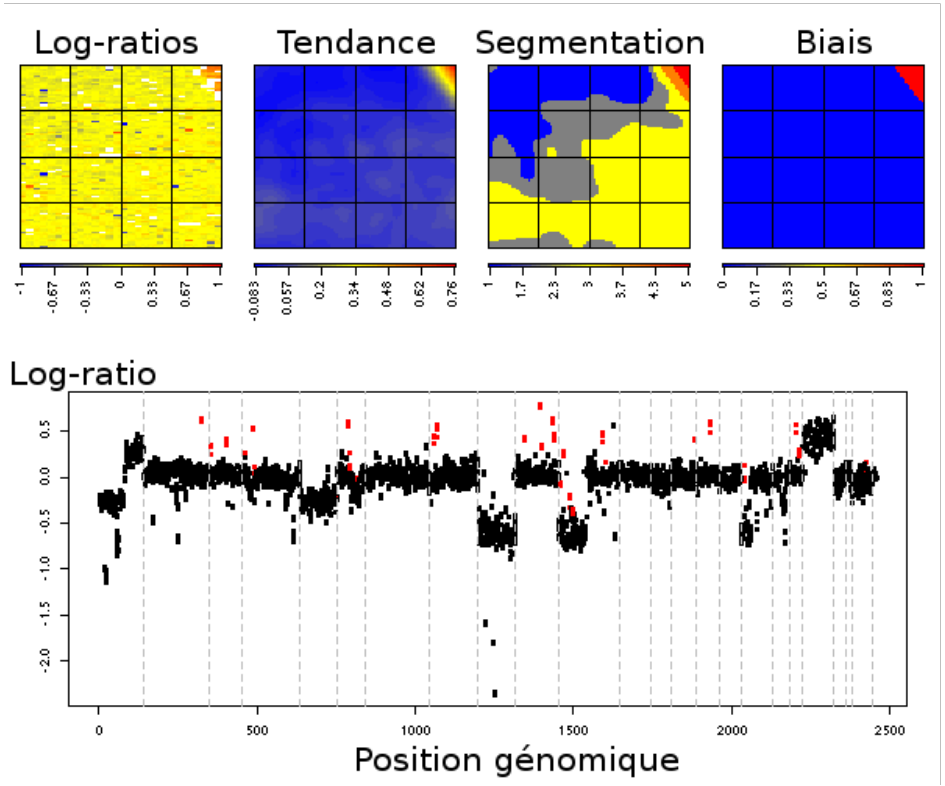
## Neuvial et al., BMC Bioinformatics, 2006




1. Abnormal Log-Ratio in the corner
2. Spatial trend estimation by 2D-LOESS
3. Spatial segmentation
4. Bias area are removed
5. Spots are **outliers** in the genomic profile

# MANOR: an algorithm to detect spatial bias

Neuvial et al., BMC Bioinformatics, 2006



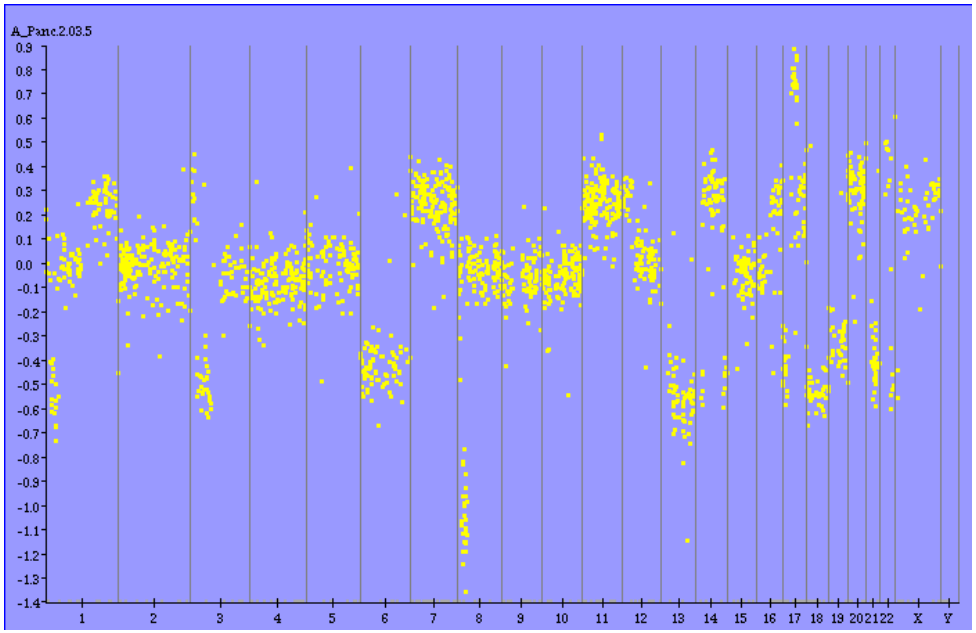

1. Abnormal Log-Ratio in the corner
2. Spatial trend estimation by 2D-LOESS
3. Spatial segmentation
4. Bias area are removed
5. Spots are **outliers** in the genomic profile

# GLAD: Gain and Loss Analysis of DNA

Hupé et al., Bioinformatics, 2004

## Profile segmentation

- The GLAD algorithm aims at identifying chromosomal regions with identical DNA copy number.




1. Log-Ratio profile
2. Smoothing line estimation
3. Breakpoint detection
4. Status assignment
5. Outliers detection

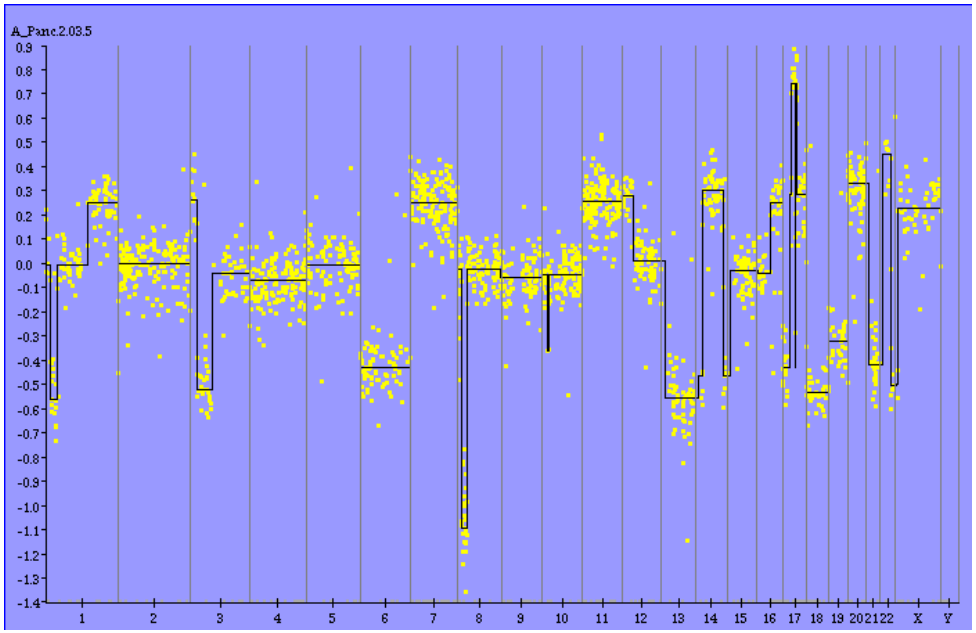
It works with BAC array, cDNA array, oligonucleotide array (Affymetrix, Agilent, Nimblegen, Illumina)

# GLAD: Gain and Loss Analysis of DNA

Hupé et al., Bioinformatics, 2004

## Profile segmentation

- The GLAD algorithm aims at identifying chromosomal regions with identical DNA copy number.




1. Log-Ratio profile
2. Smoothing line estimation
3. Breakpoint detection
4. Status assignment
5. Outliers detection

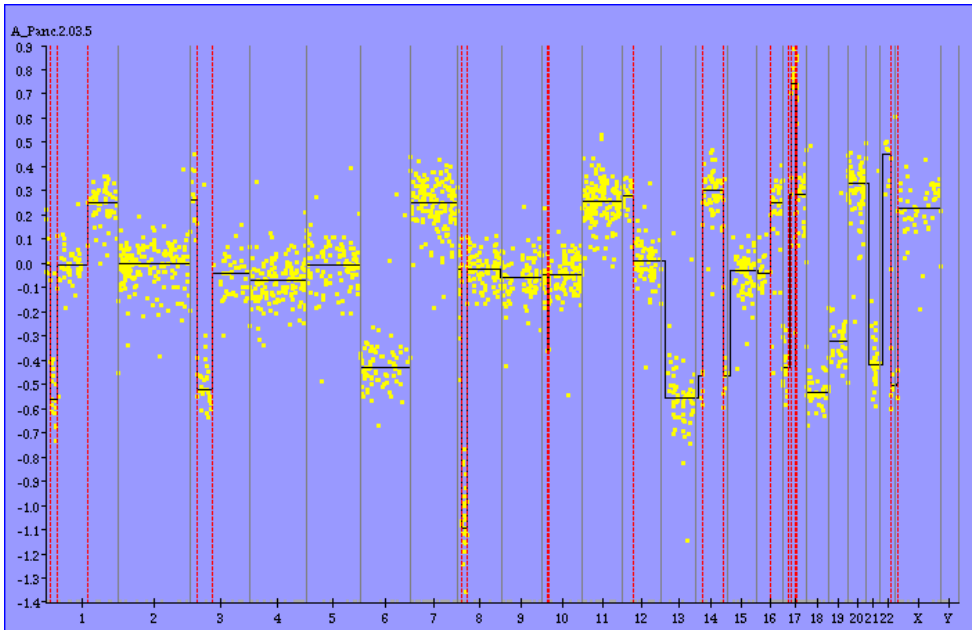
It works with BAC array, cDNA array, oligonucleotide array (Affymetrix, Agilent, Nimblegen, Illumina)

# GLAD: Gain and Loss Analysis of DNA

Hupé et al., Bioinformatics, 2004

## Profile segmentation

- The GLAD algorithm aims at identifying chromosomal regions with identical DNA copy number.




1. Log-Ratio profile
2. Smoothing line estimation
3. Breakpoint detection
4. Status assignment
5. Outliers detection

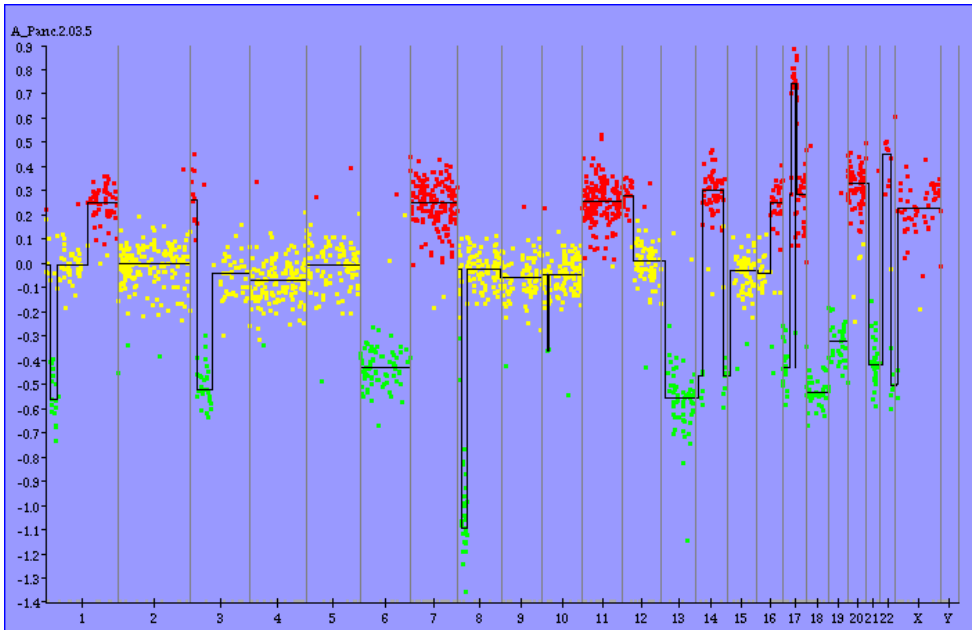
It works with BAC array, cDNA array, oligonucleotide array (Affymetrix, Agilent, Nimblegen, Illumina)

# GLAD: Gain and Loss Analysis of DNA

Hupé et al., Bioinformatics, 2004

## Profile segmentation

- The GLAD algorithm aims at identifying chromosomal regions with identical DNA copy number.




1. Log-Ratio profile
2. Smoothing line estimation
3. Breakpoint detection
4. Status assignment
5. Outliers detection

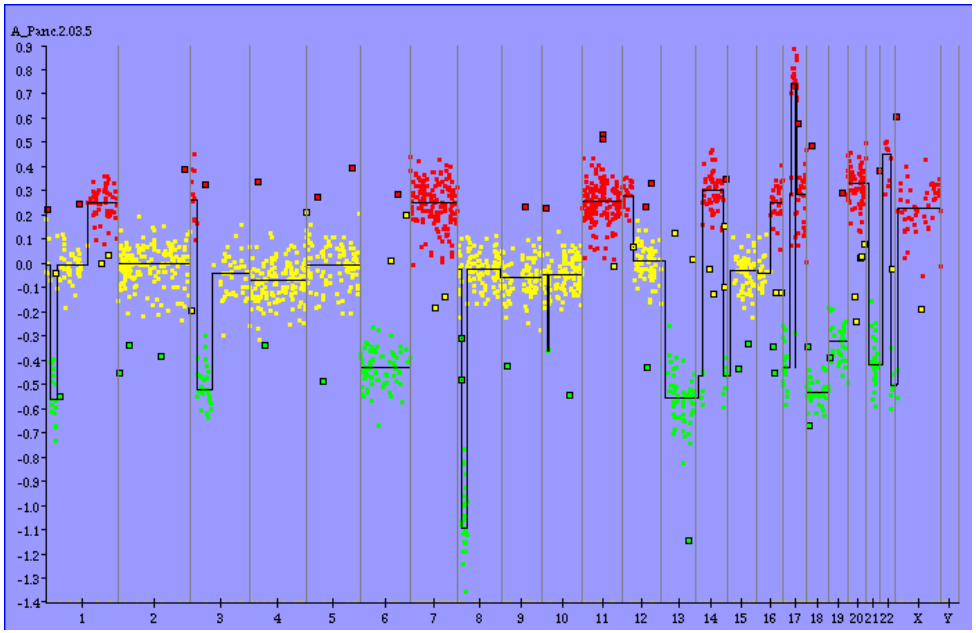
It works with BAC array, cDNA array, oligonucleotide array (Affymetrix, Agilent, Nimblegen, Illumina)

# GLAD: Gain and Loss Analysis of DNA

Hupé et al., Bioinformatics, 2004

## Profile segmentation

- The GLAD algorithm aims at identifying chromosomal regions with identical DNA copy number.




1. Log-Ratio profile
2. Smoothing line estimation
3. Breakpoint detection
4. Status assignment
5. Outliers detection

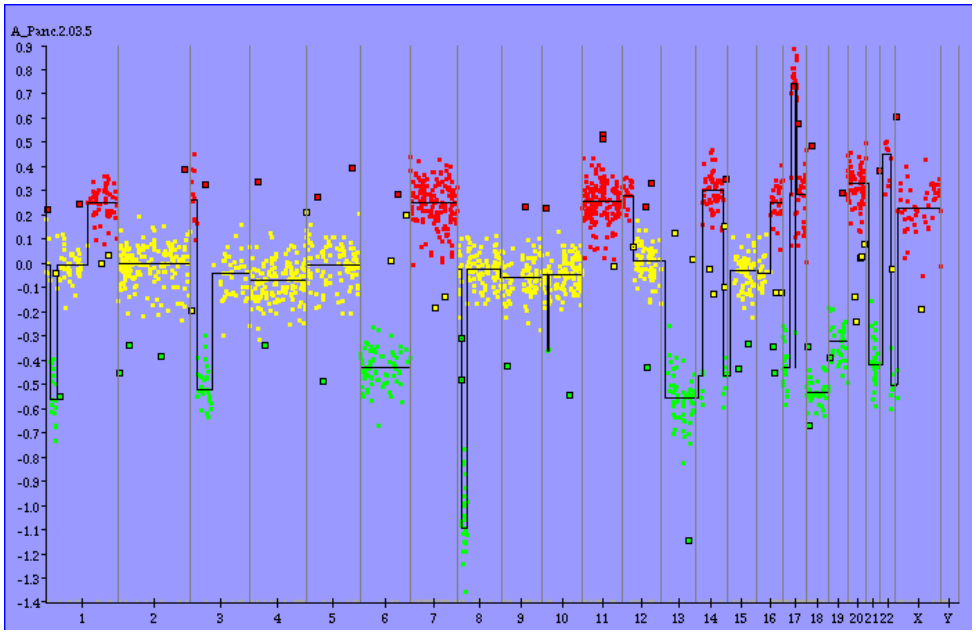
It works with BAC array, cDNA array, oligonucleotide array (Affymetrix, Agilent, Nimblegen, Illumina)

# GLAD: Gain and Loss Analysis of DNA

Hupé et al., Bioinformatics, 2004

## Profile segmentation

- The GLAD algorithm aims at identifying chromosomal regions with identical DNA copy number.




1. Log-Ratio profile
2. Smoothing line estimation
3. Breakpoint detection
4. Status assignment
5. Outliers detection

It works with BAC array, cDNA array, oligonucleotide array (Affymetrix, Agilent, Nimblegen, Illumina)

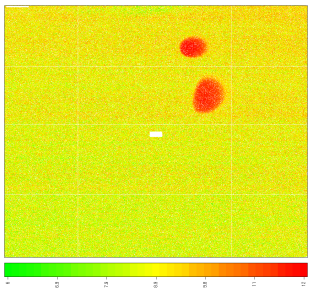
## ITALICS: ITerative and Alternative normaLIzation of Copy number Snp array

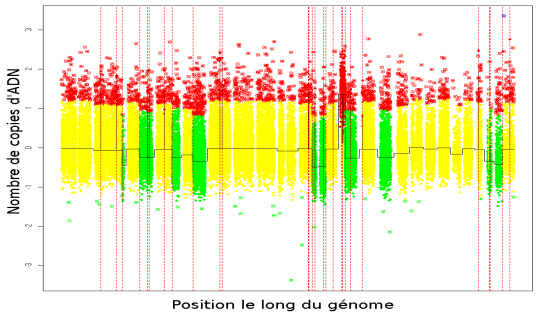
Rigaill et al., Bioinformatics, 2008

## Devoted to the analysis of Affymetrix Genome-Wide SNP chip

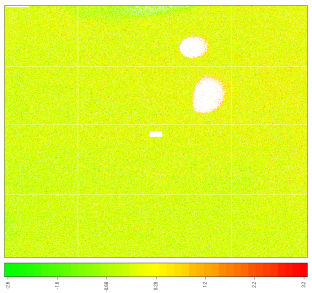
- the specificities of the affymetrix technology are taken into account in the algorithm
- the signal to noise ratio is better
- the breakpoint location is more accurate

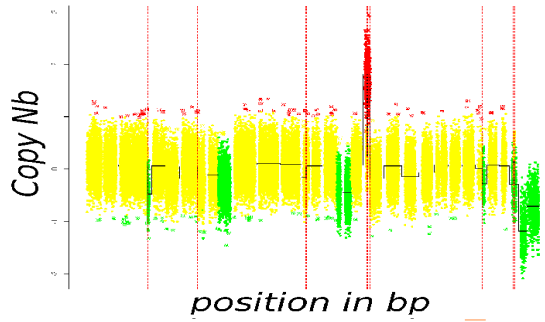
Spatial artifact $\Rightarrow$ 1600 aberrant values






ITALICS $\Rightarrow$ 90 aberrant values, 13 SNP discarded





# Outline

1 Biological / clinical context

2 R packages description

3 End-user interfaces / automatic workflow

# Biologist / Clinician end-users

- need to visualise their data
- biological interpretation of their data
- not necessarly familiar with R programing language
- no biostatistician/bioinformatician in their lab
- need easy-to-use interfaces

**Diffusion of statistical methods within the scientific community**

If we want our statistical methods to be used, we need to package them properly.

# Biologist / Clinician end-users

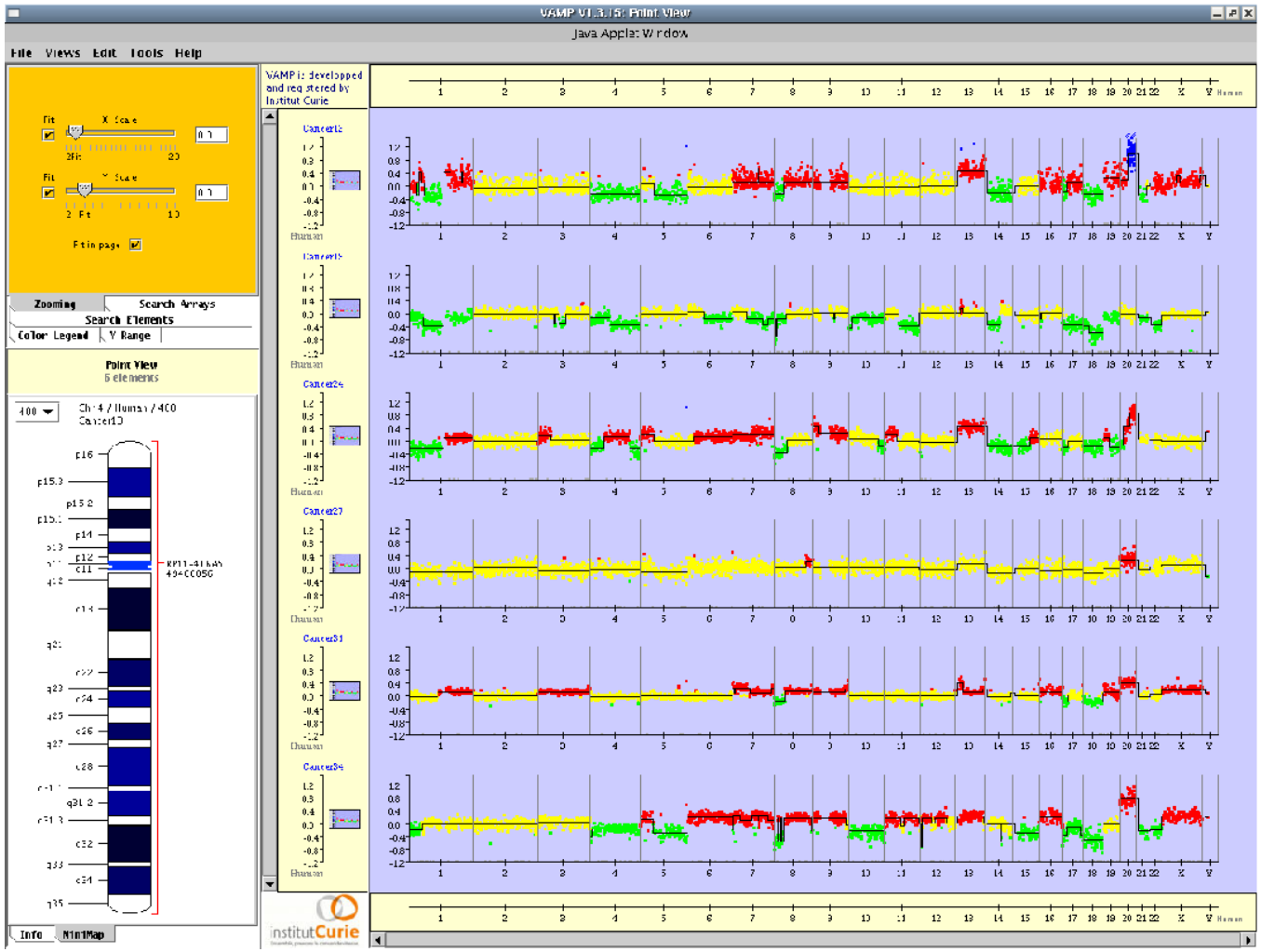
- need to visualise their data
- biological interpretation of their data
- not necessarly familiar with R programing language
- no biostatistician/bioinformatician in their lab
- need easy-to-use interfaces

**Diffusion of statistical methods within the scientific community**

If we want our statistical methods to be used, we need to package them properly.

# VAMP: a software to visualise and analyse data

## La Rosa et al., Bioinformatics, 2006




http://bioinfo.curie.fr/vamp

# Our tools fo DNA copy number experiments

- R packages (MANOR, GLAD, ITALICS) for biostatistical analysis
- VAMP java software for visualisation (and analysis)

**Need for an integrated environment**

CAPweb is a web interface which allows the use of all the previous tools.

# Our tools fo DNA copy number experiments

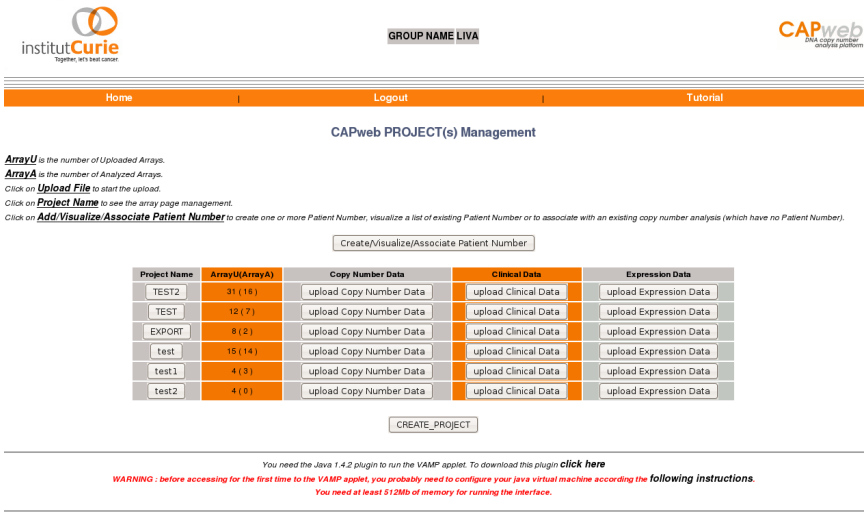
- R packages (MANOR, GLAD, ITALICS) for biostatistical analysis
- VAMP java software for visualisation (and analysis)

**Need for an integrated environment**

CAPweb is a web interface which allows the use of all the previous tools.

# CAPweb: an end-user web platform

## Liva et al., Nucleic Acids Research, 2006

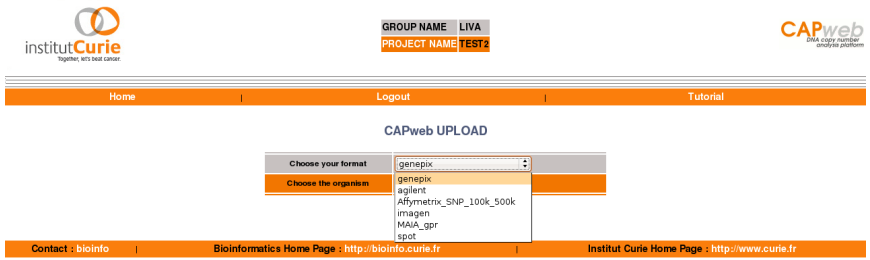



http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# CAPweb: an end-user web platform

Liva et al., Nucleic Acids Research, 2006

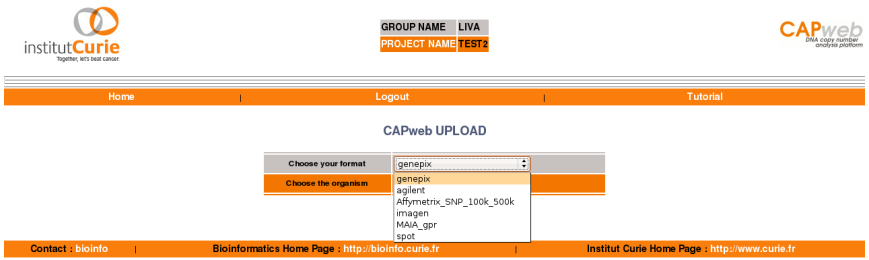



http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# CAPweb: an end-user web platform

Liva et al., Nucleic Acids Research, 2006

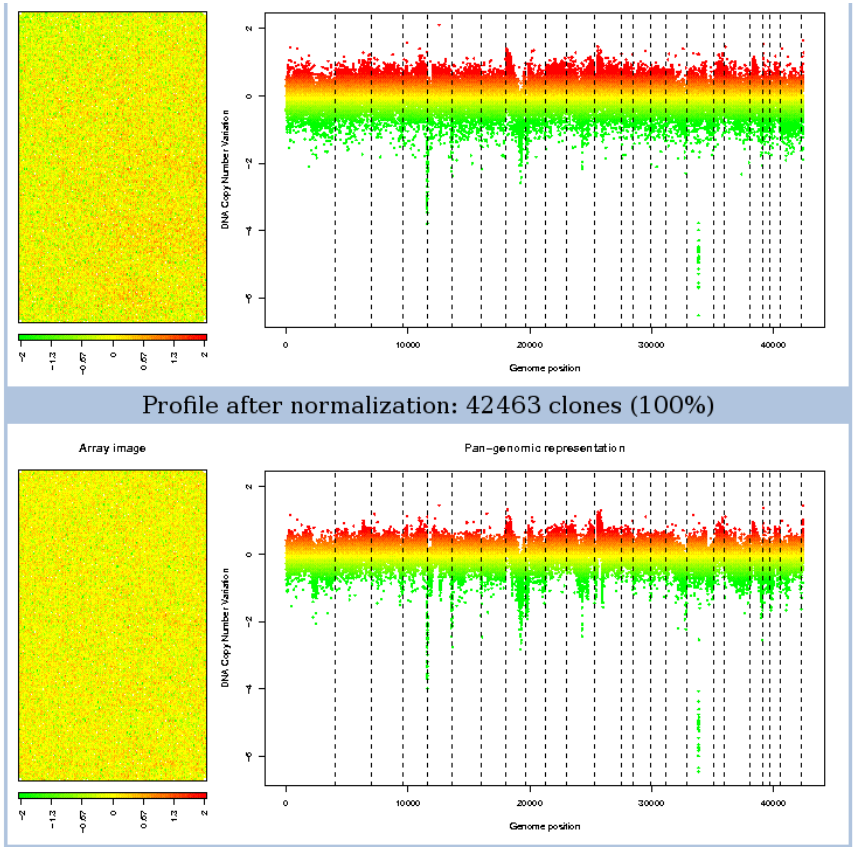



http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# CAPweb: an end-user web platform

## Liva et al., Nucleic Acids Research, 2006

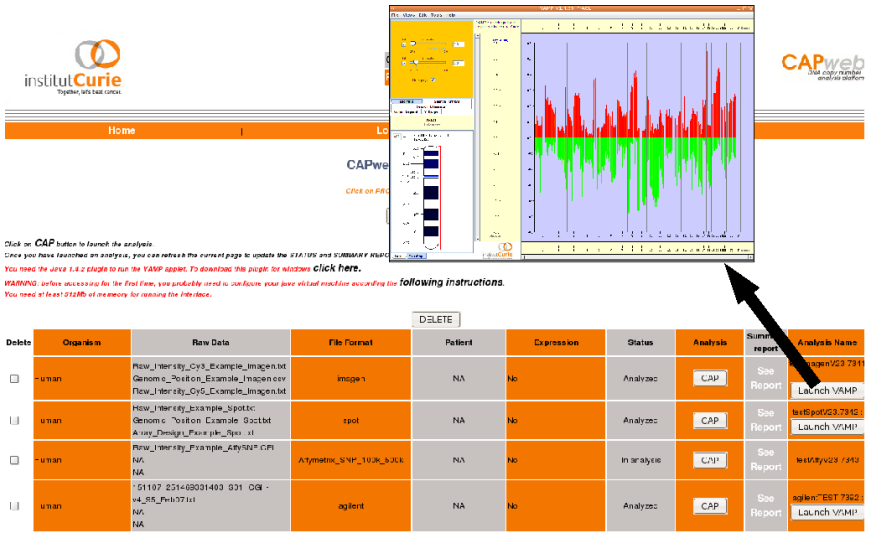



http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# CAPweb: an end-user web platform

## Liva et al., Nucleic Acids Research, 2006

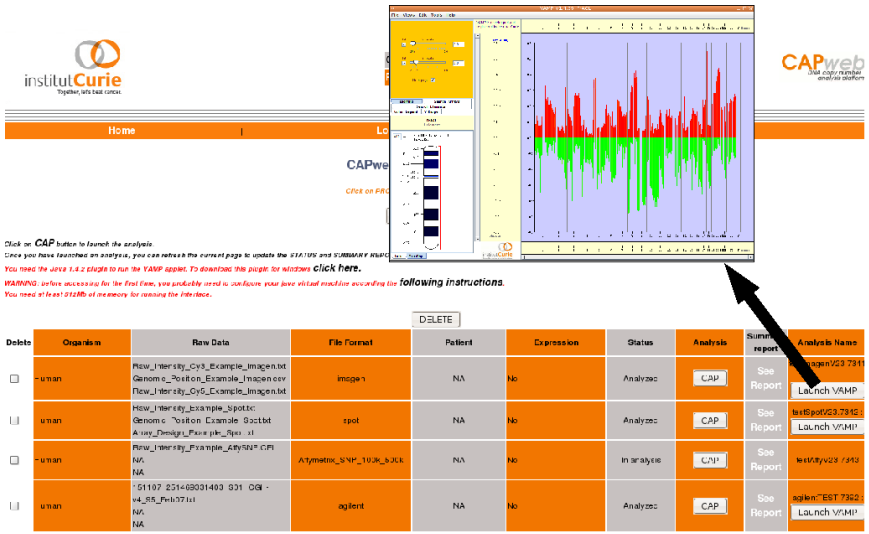



http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# CAPweb: an end-user web platform

## Liva et al., Nucleic Acids Research, 2006

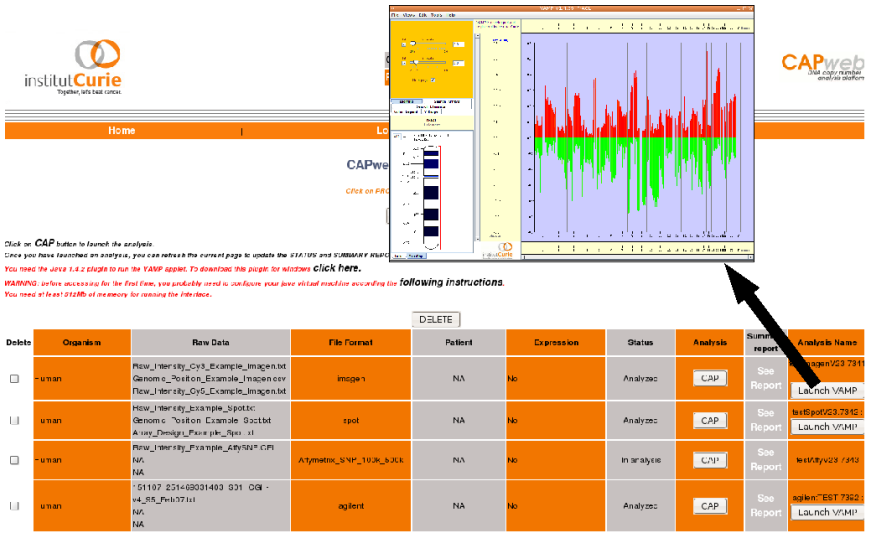



http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# CAPweb: an end-user web platform

## Liva et al., Nucleic Acids Research, 2006

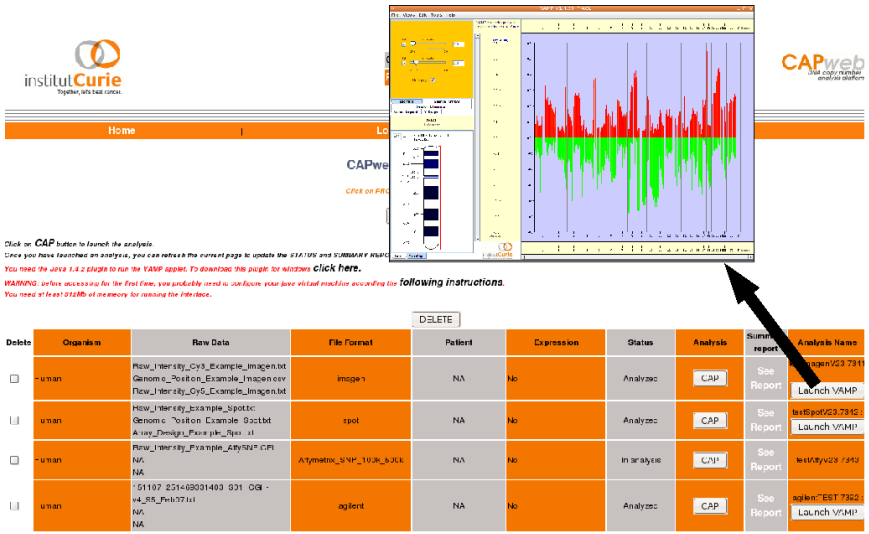



http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# CAPweb: an end-user web platform

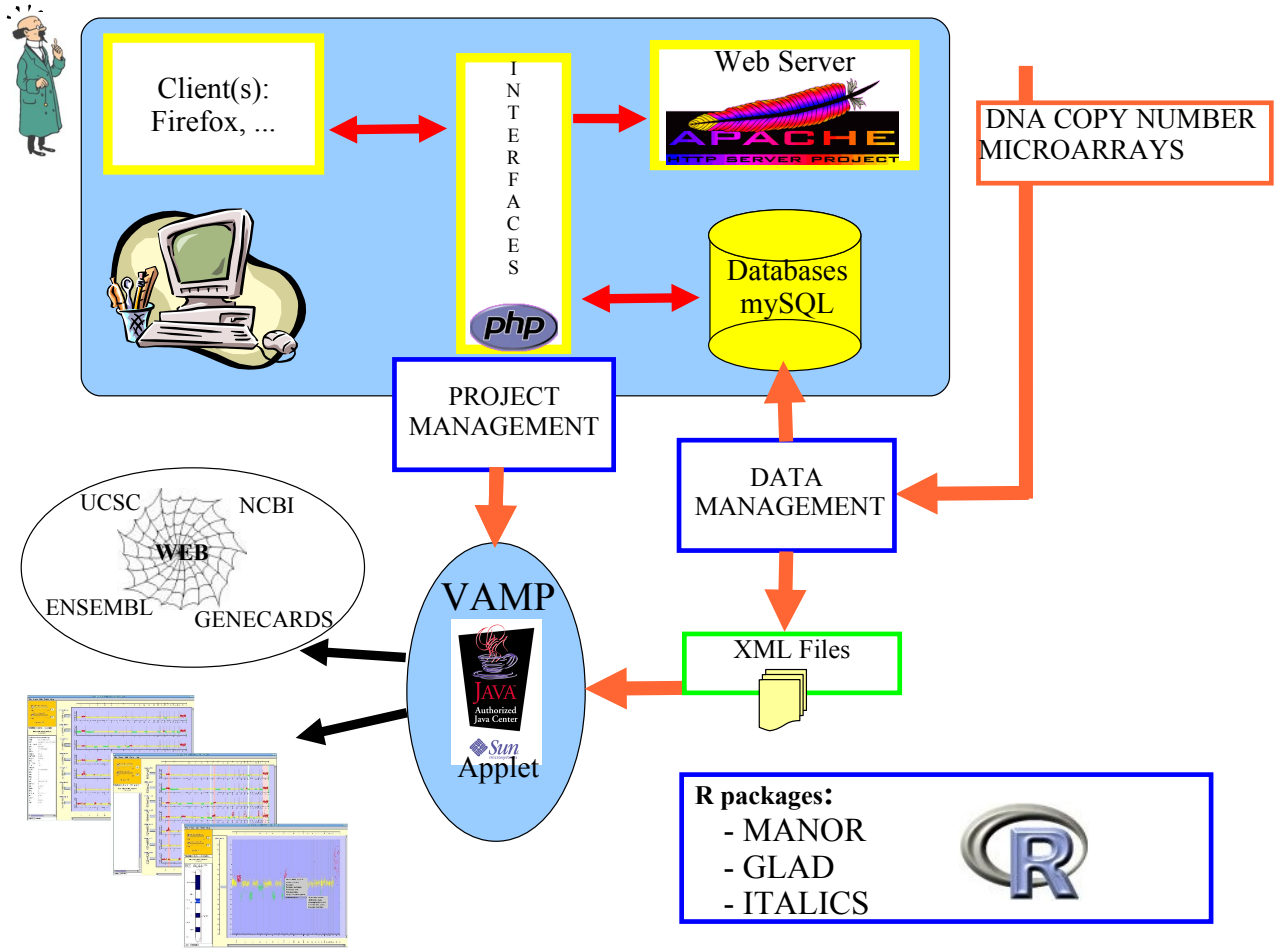
## Liva et al., Nucleic Acids Research, 2006




http://bioinfo.curie.fr/capweb

- user registration
- project management
- input file: Genepix, spot, Imagen, Feature Extraction, CEL
- normalisation: MANOR, ITALICS
- breakpoint detection: GLAD
- summary report
- visualise the data with VAMP
- array technology: BAC, cDNA, Agilent, Affymetrix (100K, 500K) (Illumina, Nimblegen soon)
- integration of clinical data
- integration of mRNA data

# Client / Server Architecture




Our R packages are used calling CGI from any web browser

# Recent evolutions and perspectives

## Recent changes

- Possibility to use HaarSeg algorithm (Ben-Yaacov and Eldar, Bioinformatics, 2008) in GLAD $\rightarrow$ 2 millions genomic profiles can be analysed within 1 minute
- Use C/C++ in order to reduce computing time

## On-going work

- Improvement of ITALICS in order to analyse Affymetrix Genome Wide SNP 6.0
- Extension to Next-Generation Sequencing technologies (Terabytes of data!!)

aroma.affymetrix (Bengtsson et al.) R package offers many functionalities for affymetrix data analysis

# Recent evolutions and perspectives

## Recent changes

- Possibility to use HaarSeg algorithm (Ben-Yaacov and Eldar, Bioinformatics, 2008) in GLAD $\rightarrow$ 2 millions genomic profiles can be analysed within 1 minute
- Use C/C++ in order to reduce computing time

## On-going work

- Improvement of ITALICS in order to analyse Affymetrix Genome Wide SNP 6.0
- Extension to Next-Generation Sequencing technologies (Terabytes of data!!)

aroma.affymetrix (Bengtsson et al.) R package offers many functionalities for affymetrix data analysis

# Recent evolutions and perspectives

## Recent changes

- Possibility to use HaarSeg algorithm (Ben-Yaacov and Eldar, Bioinformatics, 2008) in GLAD $\rightarrow$ 2 millions genomic profiles can be analysed within 1 minute
- Use C/C++ in order to reduce computing time

## On-going work

- Improvement of ITALICS in order to analyse Affymetrix Genome Wide SNP 6.0
- Extension to Next-Generation Sequencing technologies (Terabytes of data!!)

aroma.affymetrix (Bengtsson et al.) R package offers many functionalities for affymetrix data analysis

# Acknowledgements


- Institut Curie / Bioinformatics U900

# THANKS