DEA/DES in Bioinformatics 2003-2004

Thesis

Breast Cancer Diagnosis Using Microarray

Training Master:
M. Christos Sotiriou

Training Supervisor:
M. Gianluca Bontempi

Benjamin Haibe-Kains

Table of Contents

- Introduction
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- Methods and Results
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- Discussion
  - Future Works

Table of Contents

- **Introduction**
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- Methods and Results
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- Discussion
  - Future Works

Introduction
Breast Cancer Diagnosis

- Several histological criteria characterize breast tumor
  - Invasive/non-invasive tumor
  - Number of involved lymph nodes
  - Size
  - Tumor grade
  - Estrogen receptor status
  - Oncogene over-expression
  - Margins of resection

Introduction
Breast Cancer Diagnosis (2)

- Appearance of distant metastases in the first 5 years of follow-up
  - Binary classification (relapse/non-relapse)
- Goals
  - Reduce significantly the patients who receive unnecessary treatments
    - Adverse side effects
    - Treatment costs
  - Isolate involved genes

Introduction
Breast Cancer Diagnosis (3)

- Histological criteria fail to classify the tumors
- Development of new predictors based on **gene expression profile**

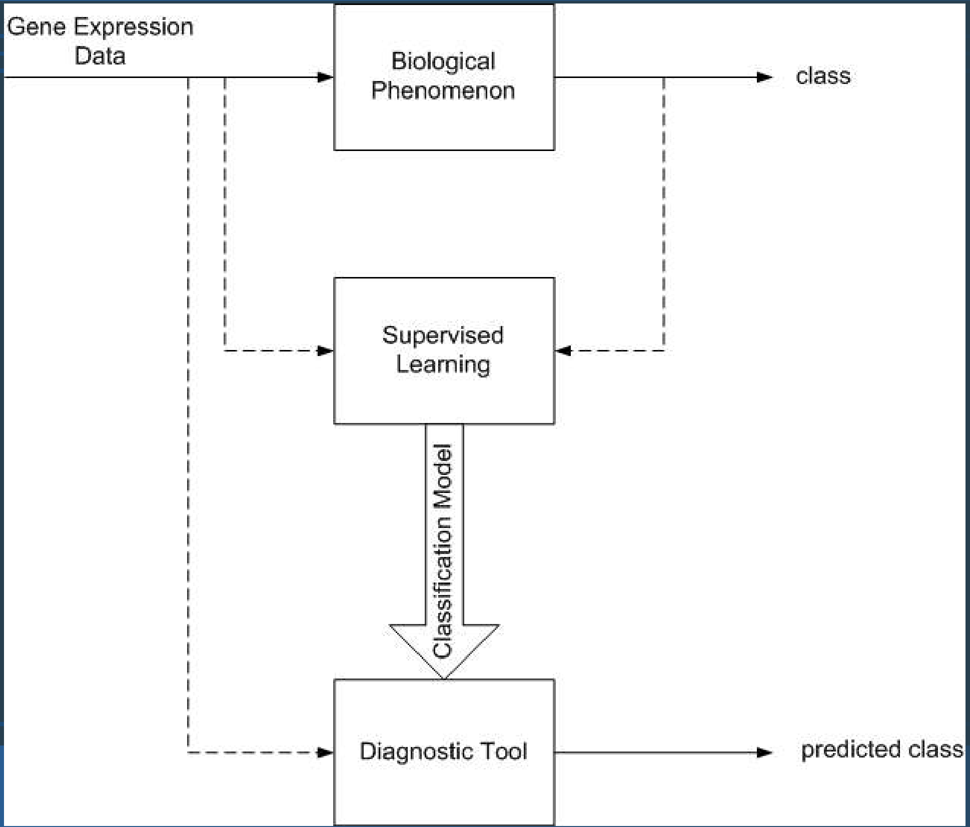


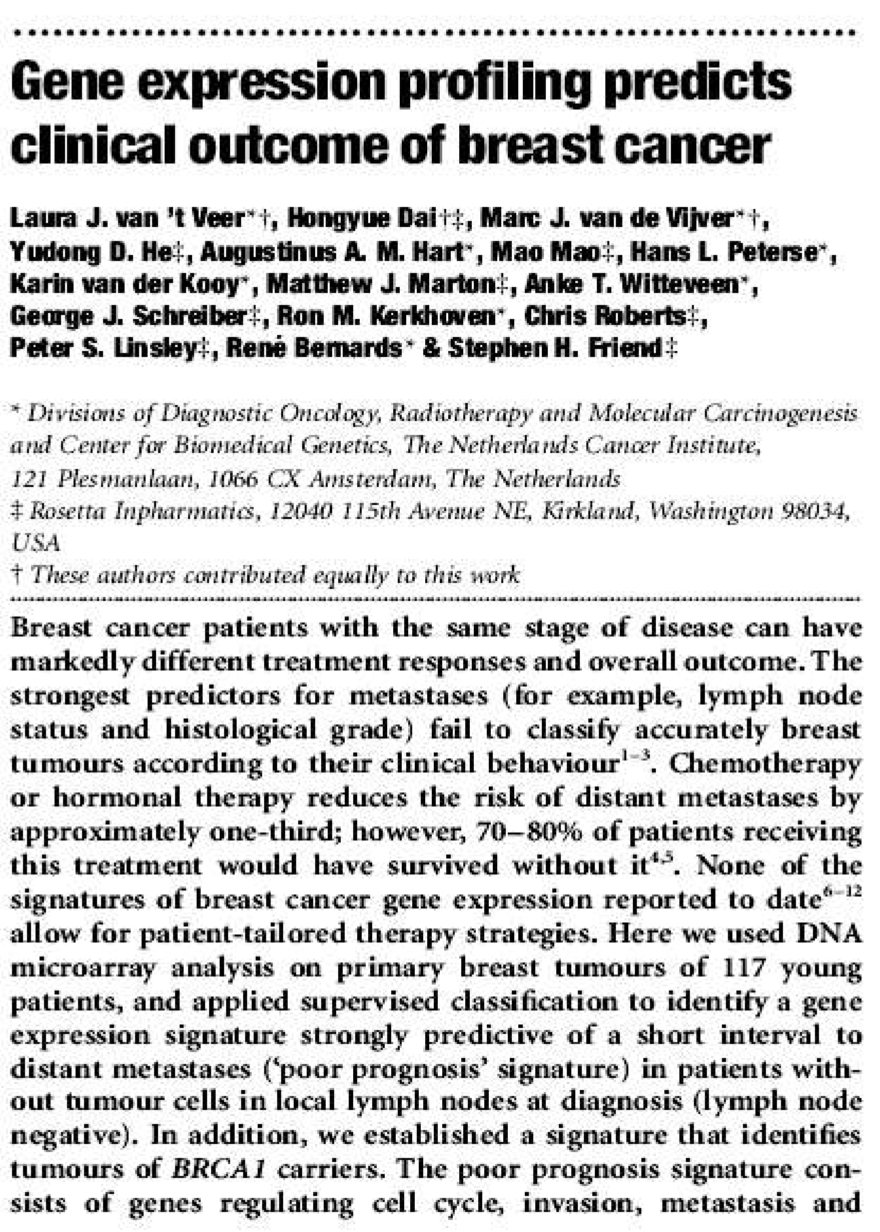
Table of Contents

- **Introduction**
  - **TransBIG Project**
- Materials
  - Populations
  - Microarray Platform
- Methods and Results
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- Discussion
  - Future Works

Introduction
TransBIG Project

- TransBIG project
  - Validation of van't Veer signature
    - Agilent microarray technology
    - 70 maker genes (van't Veer *et al*. 2002)
  - **Development of a new signature**
    - Affymetrix microarray technology
    - Supervised by Christos Sotiriou at the IJB (Microarray Unity)
    - Collaboration with the SIB

Introduction
TransBIG Project (2)



Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

** Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands*

‡ Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA

† These authors contributed equally to this work

Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour[1–3]. Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70–80% of patients receiving this treatment would have survived without it[4,5]. None of the signatures of breast cancer gene expression reported to date[6–12] allow for patient-tailored therapy strategies. Here we used DNA microarray analysis on primary breast tumours of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature) in patients without tumour cells in local lymph nodes at diagnosis (lymph node negative). In addition, we established a signature that identifies tumours of *BRCA1* carriers. The poor prognosis signature consists of genes regulating cell cycle, invasion, metastasis and

70 genes of van't Veer Signature

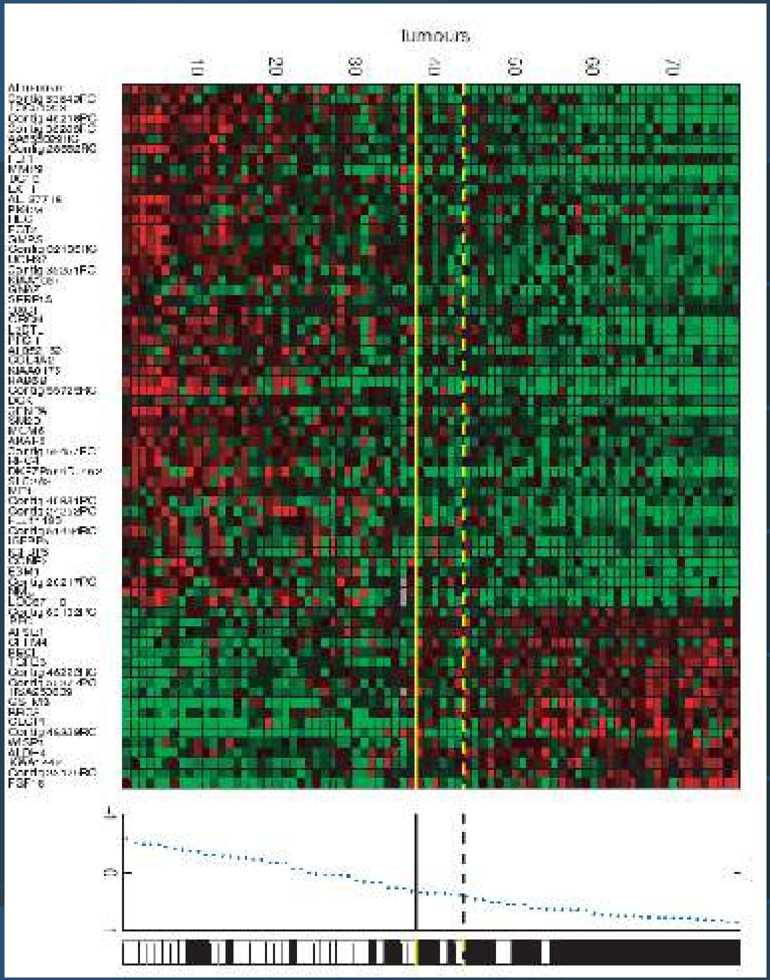


Table of Contents

- Introduction
  - TransBIG Project
- **Materials**
  - **Populations**
  - Microarray Platform
- Methods and Results
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- Discussion
  - Future Works

Materials
Populations

- John Radcliffe Hospital (JRH, Oxford)
  - 77 samples hybridized at IJB
- Gustave Roussy Hospital (IGR, Paris)
  - 65 samples hybridized at IJB
- Karolinska Institute and Hospital (Karolinska, Stockholm)
  - 19 samples hybridized at IJB
  - 68 samples hybridized at Karolinska

Materials
Populations (2)

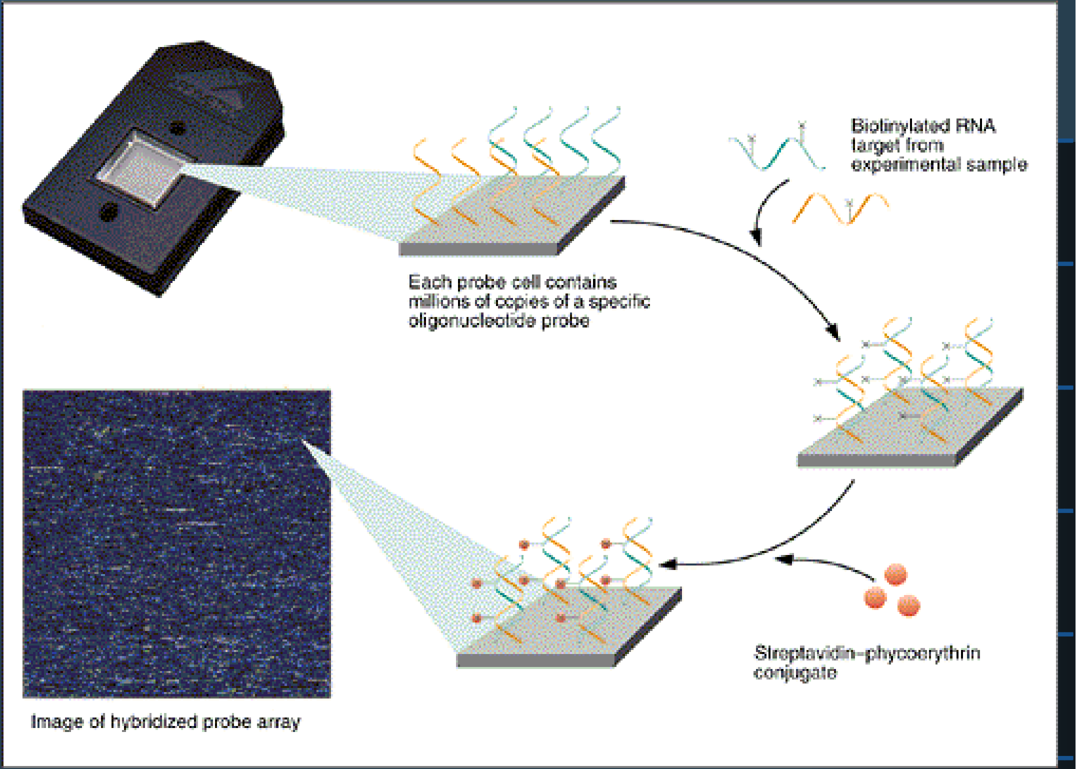
- Highly **unbalanced** class distribution
  - ¼ of relapses (class 1)
  - ¾ of non-relapses (class 0)

Table of Contents

- Introduction
  - TransBIG Project
- **Materials**
  - Populations
  - **Microarray Platform**
- Methods and Results
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- Discussion
  - Future Works

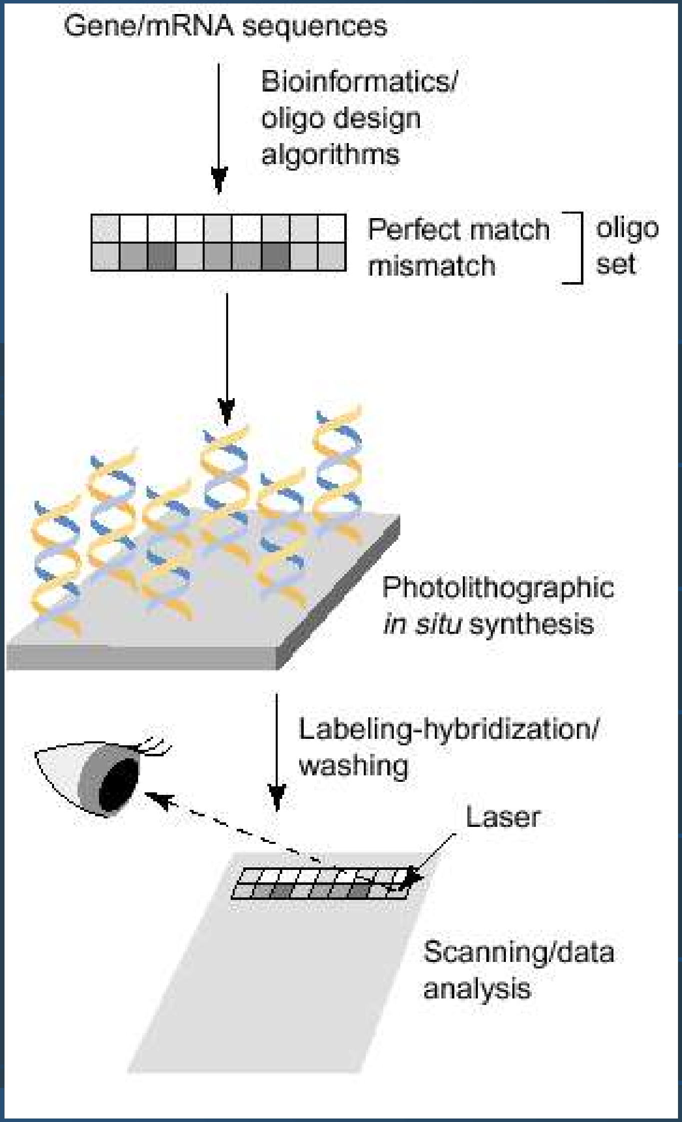
Materials
Microarray Platform

- cRNA microarrays is a recent technique used to determine **genomewide gene expression levels**
- Measurement of the quantity of cRNA, prepared from mRNA, hybridized on the chip


Biotinylated RNA target from experimental sample
Each probe cell contains millions of copies of a specific oligonucleotide probe
Streptavidin-phycoerythrin conjugate
Image of hybridized probe array

Materials
Microarray Platform (2)

- **Affymetrix**: short oligonucleotide technology
- Chip *hgu133a* (22283 probe sets)
- Chip *hgu133b* (22645 probe sets)
- *CEL* files



Gene/mRNA sequences
Bioinformatics/ oligo design algorithms
Perfect match
mismatch
oligo set
Photolithographic *in situ* synthesis
Labeling-hybridization/ washing
Laser
Scanning/data analysis

Table of Contents

- Introduction
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- **Methods and Results**
  - **Development Tools**
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- Discussion
  - Future Works

Methods and Results
Development Tools

- R and Bioconductor
  - Manifold and reliability
  - Completeness
  - Open-source
- Application server installation to carry out large bioinformatics analyzes

Table of Contents

- Introduction
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- **Methods and Results**
  - Development Tools
  - **Quality Assessment**
  - Supervised Classification
  - Gene Ontology
- Discussion
  - Future Works

Methods and Results
Quality Assessment

- Important step in the analysis design
  - During hybridization: tests carried out in laboratory (e.g. tissue purity)
  - After hybridization: quality controls based on Affymetrix *CEL* files
    - Probe array image
    - Average background
    - Spike controls and RNA degradation
    - Detection calls
    - Scaling factor
    - Box plots for PM intensities

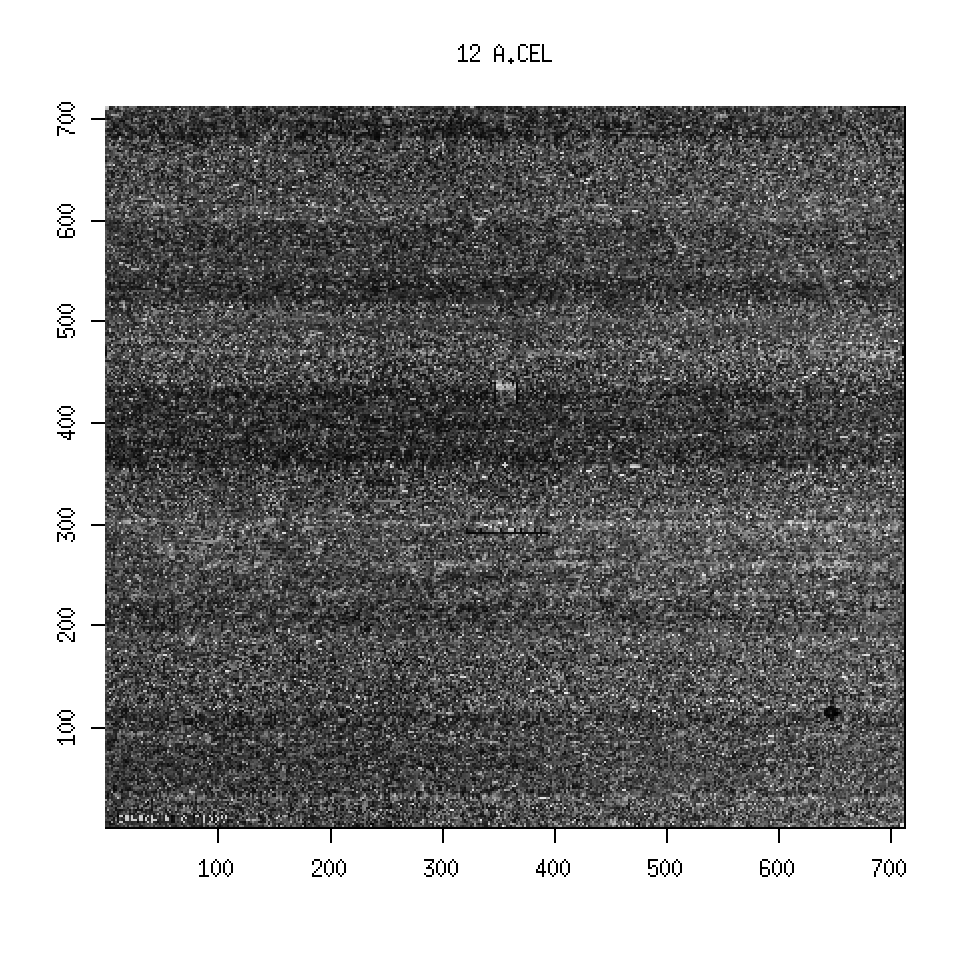
Methods and Results
Quality Assessment (2)

- No standard for quality control
- Affymetrix and Bioconductor guidelines

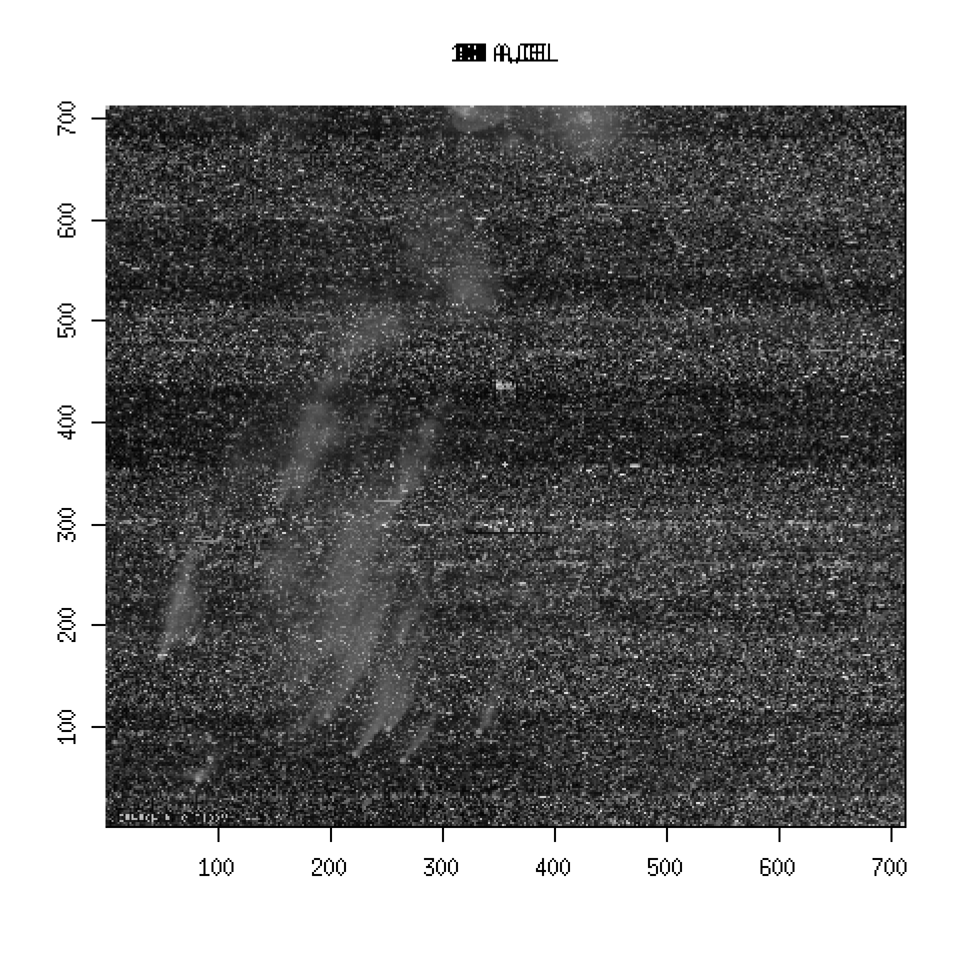
- Probe array image
  - Gray scale images of the chips
  - Gray intensities computed from *CEL* file intensities
  - Visual inspection to detect artifacts

Methods and Results
Quality Assessment (3)

- Good chip



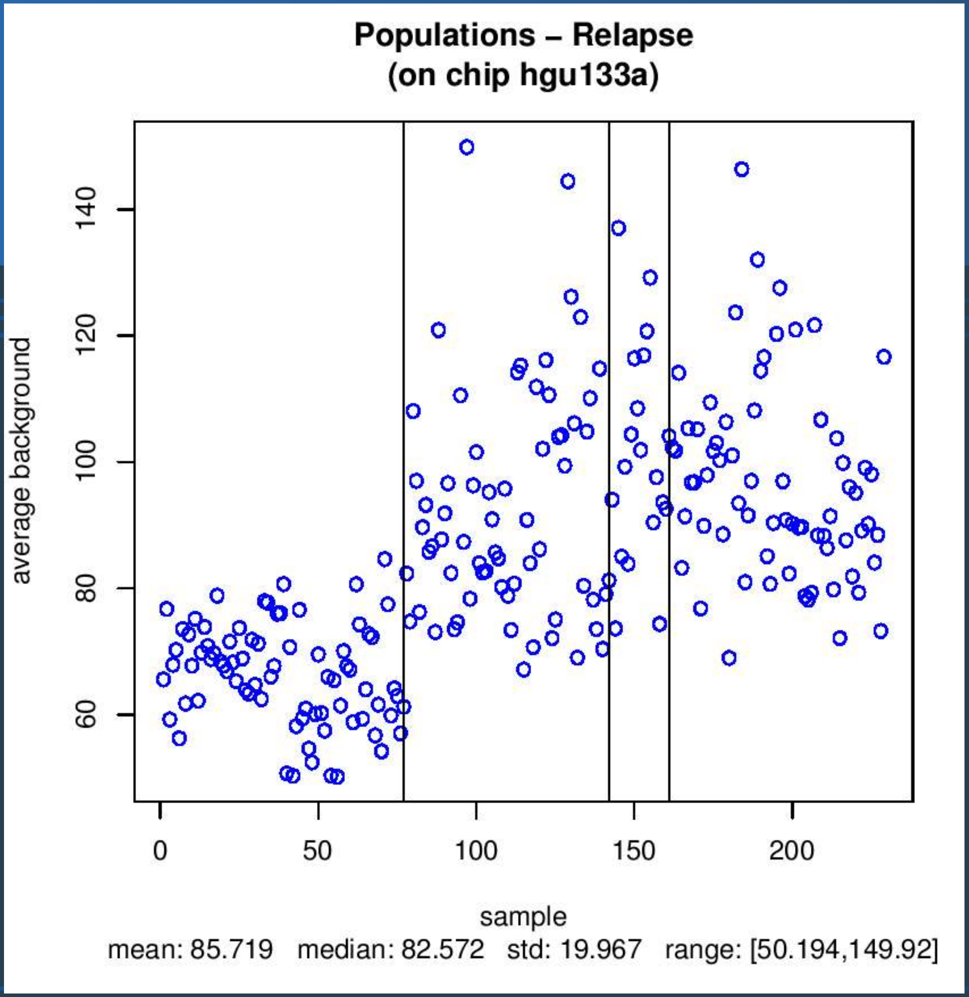
- Bad chip

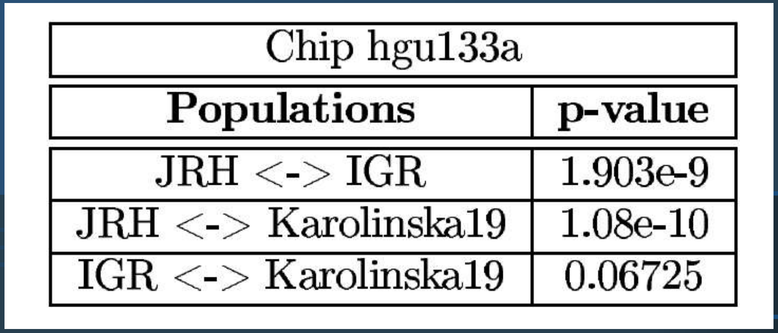


Methods and Results
Quality Assessment (4)

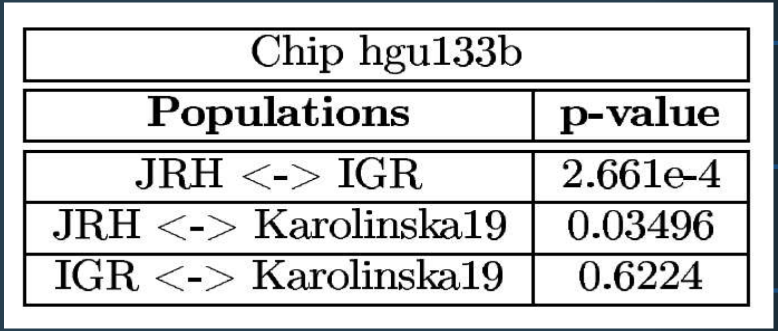
- Average background
  - Assessment of the background intensities in the chip
  - Computed by MAS 5.0 algorithm
  - Affymetrix guidline: values should be similar and < 100
  - Permutation tests to assess difference between populations

Methods and Results
Quality Assessment (5)





Chip hgu133a	
Populations	**p-value**
JRH <-> IGR	1.903e-9
JRH <-> Karolinska19	1.08e-10
IGR <-> Karolinska19	0.06725



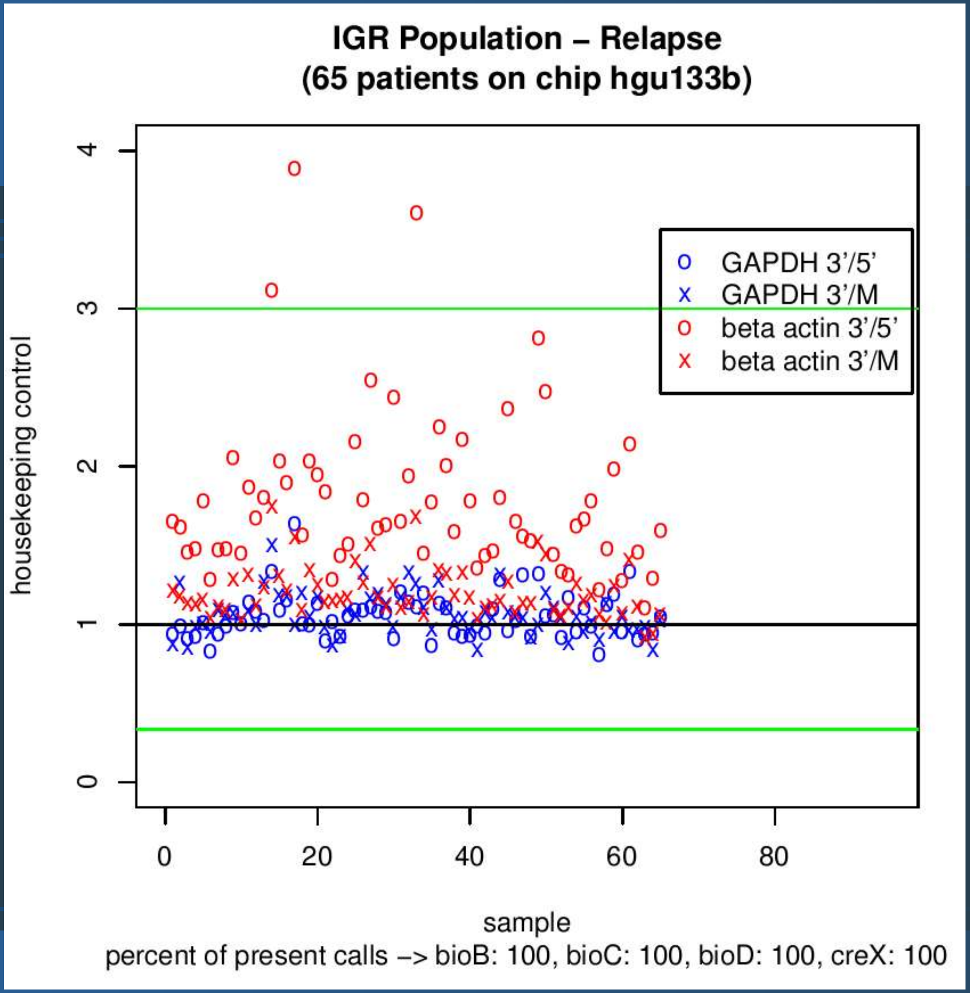
Chip hgu133b	
Populations	**p-value**
JRH <-> IGR	2.661e-4
JRH <-> Karolinska19	0.03496
IGR <-> Karolinska19	0.6224

Methods and Results
Quality Assessment (6)

- RNA degradation
  - Typically starts from the 5' end to the 3' end of the molecule (control with GAPDH and beta actin genes)
  - Affymetrix guideline: ratio 3'/5' < 3
  - RNA quality assessment
- Spike controls
  - Probes spiked during the sample preparation process (BioB, BioC, BioD, CreX should be detected as present)
  - Hybridization efficiency assessment

Methods and Results
Quality Assessment (7)

- Good quality for all the populations



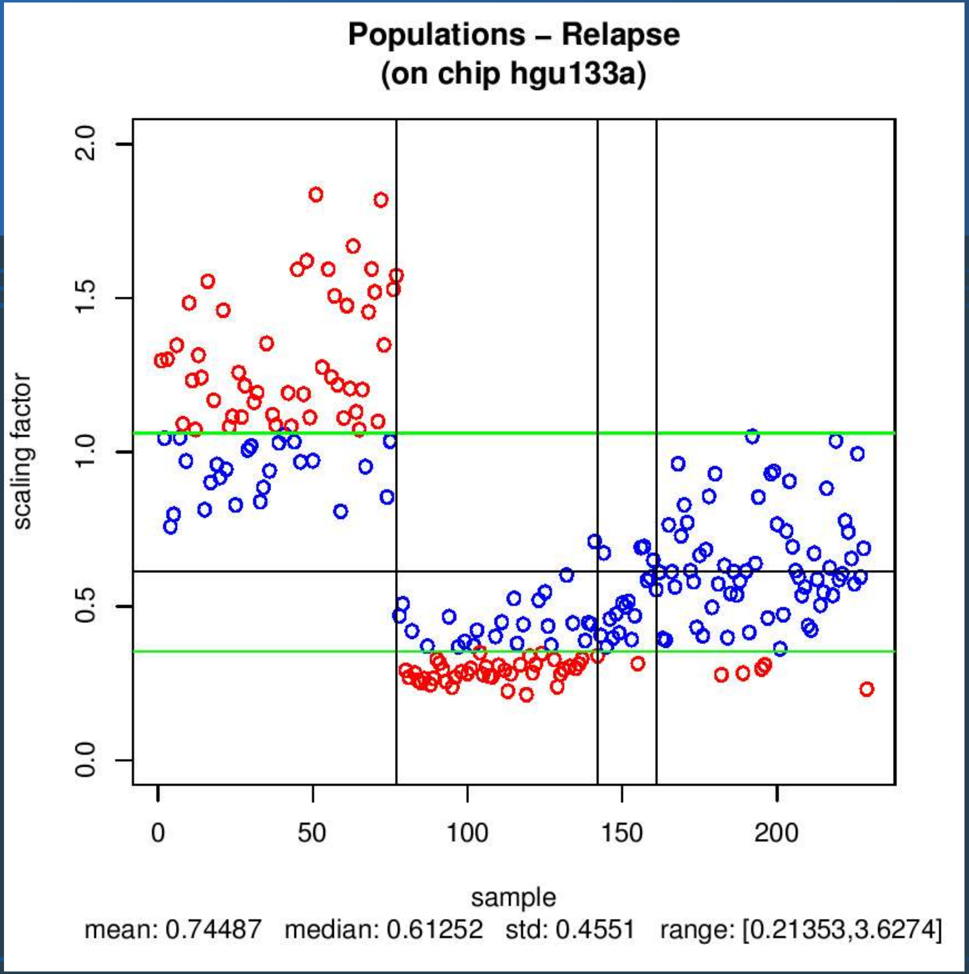
Methods and Results
Quality Assessment (8)

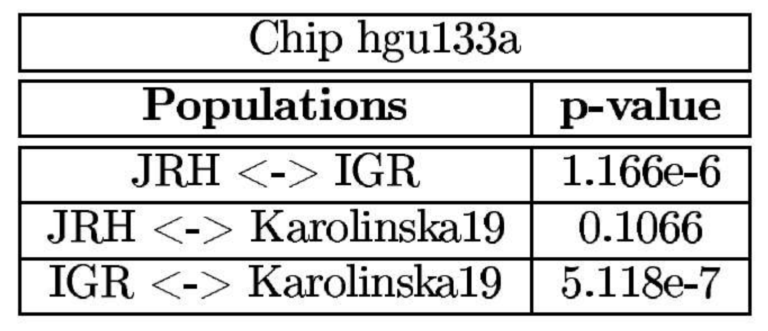
- Detection calls
  - Use of the intensities of the PM and MM probes to test statistically the *presence* or the *absence* of a specific gene
  - Computed by MAS 5.0 algorithm
  - Affymetrix guideline: extremely low percentage of *present* calls may indicate poor quality
  - Good quality for all the populations

Methods and Results
Quality Assessment (9)

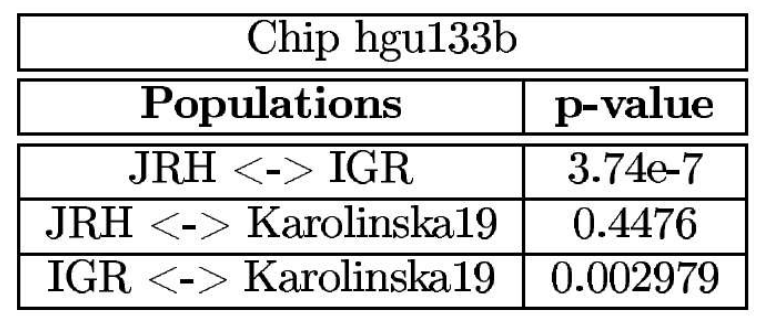
- Scaling factor
  - Assessment of the difference in mean intensity between chips
  - Computed by MAS 5.0 algorithm
  - Affymetrix guideline: recommended value of maximum three-fold scaling factor
  - Permutation tests to assess difference between populations

Methods and Results
Quality Assessment (10)





Chip hgu133a	
Populations	**p-value**
JRH <-> IGR	1.166e-6
JRH <-> Karolinska19	0.1066
IGR <-> Karolinska19	5.118e-7

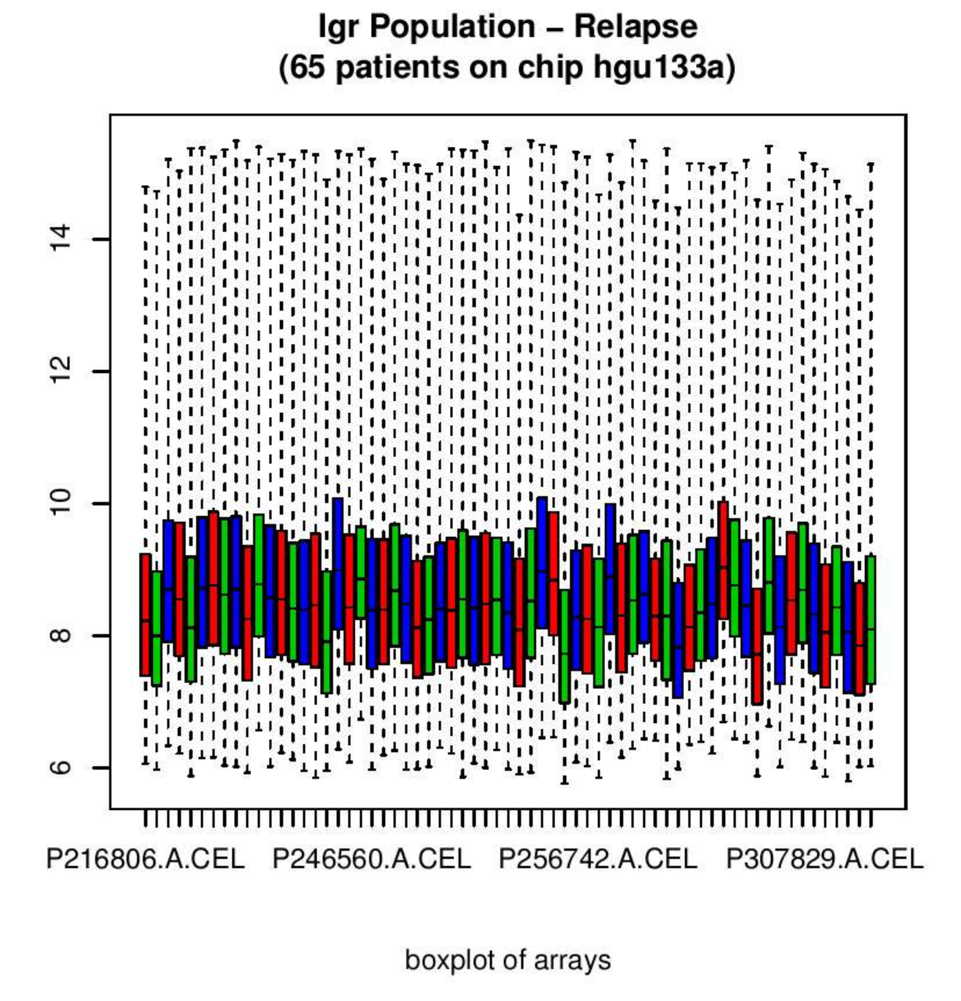


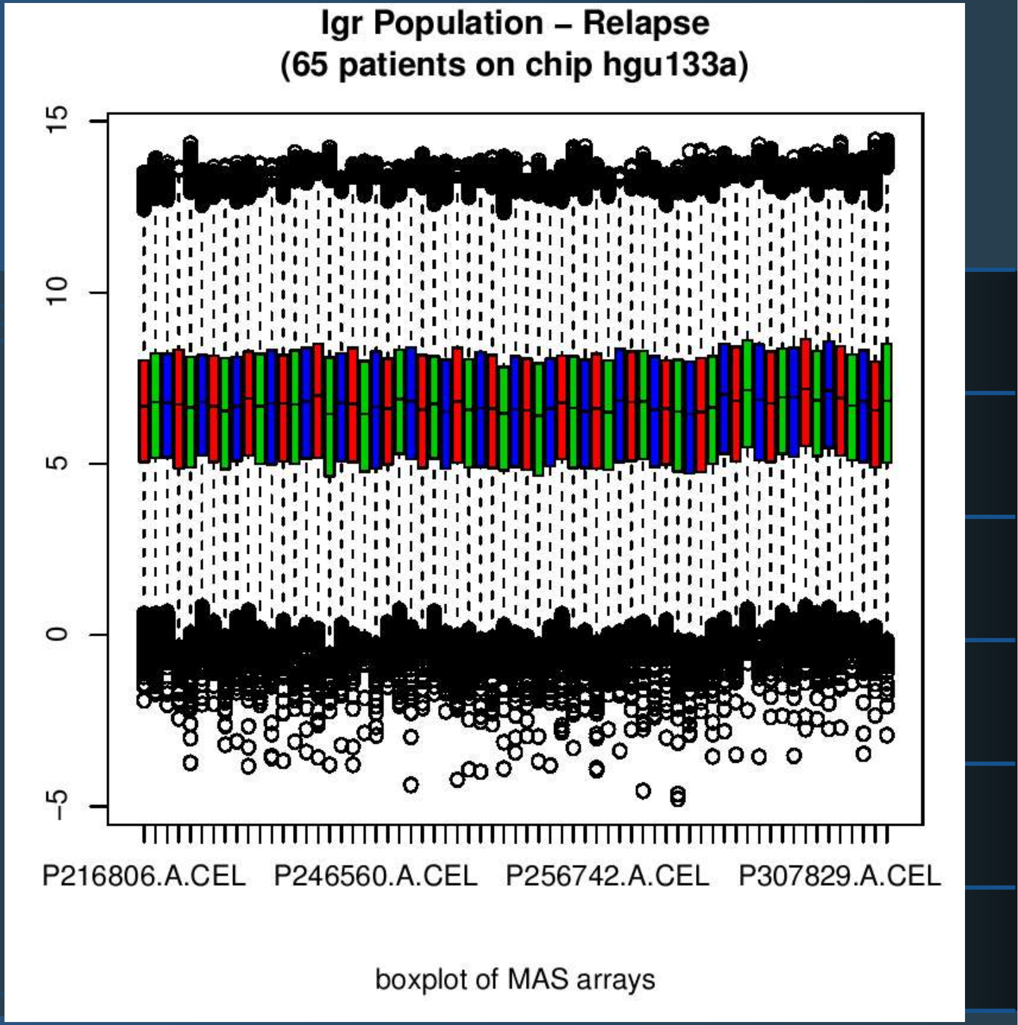
Chip hgu133b	
Populations	**p-value**
JRH <-> IGR	3.74e-7
JRH <-> Karolinska19	0.4476
IGR <-> Karolinska19	0.002979

Methods and Results
Quality Assessment (11)

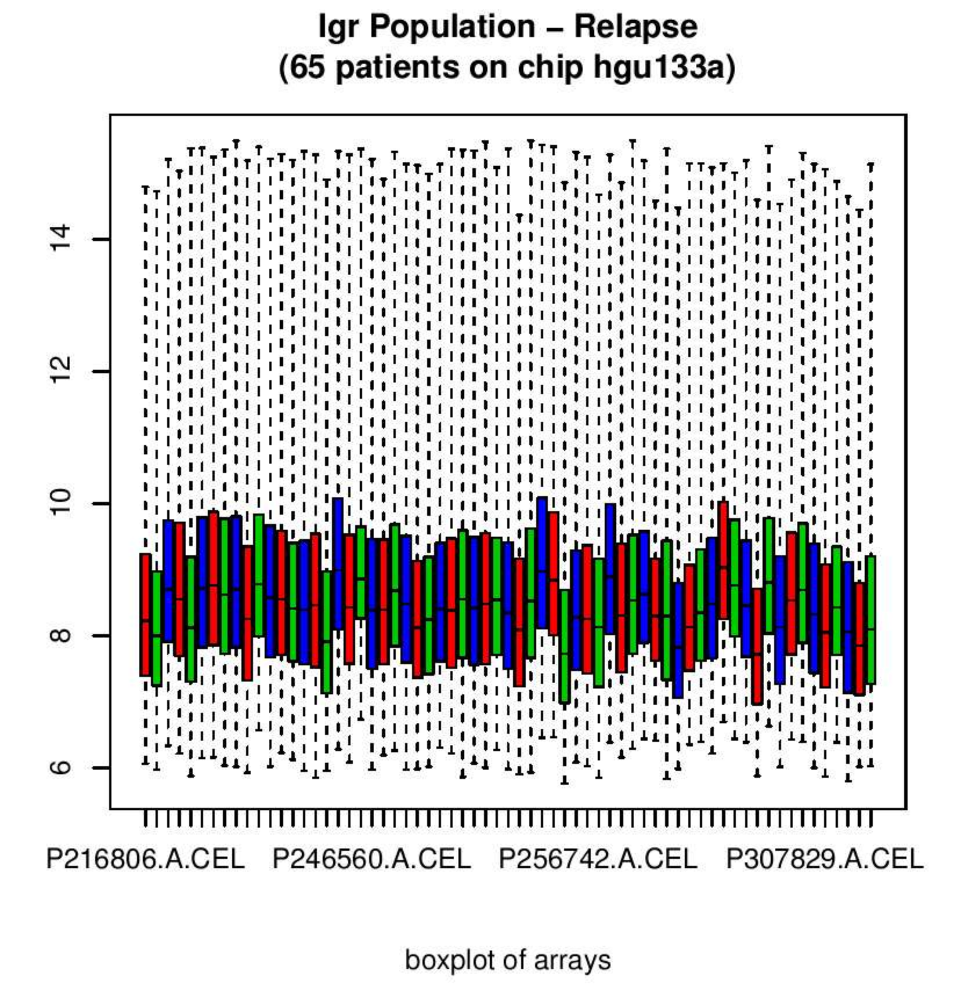
- Box plots for PM intensities
  - Useful to detect outlier and to assess the quality of the normalization
  - Computation of the median and the interquartile range of PM intensities for each chip

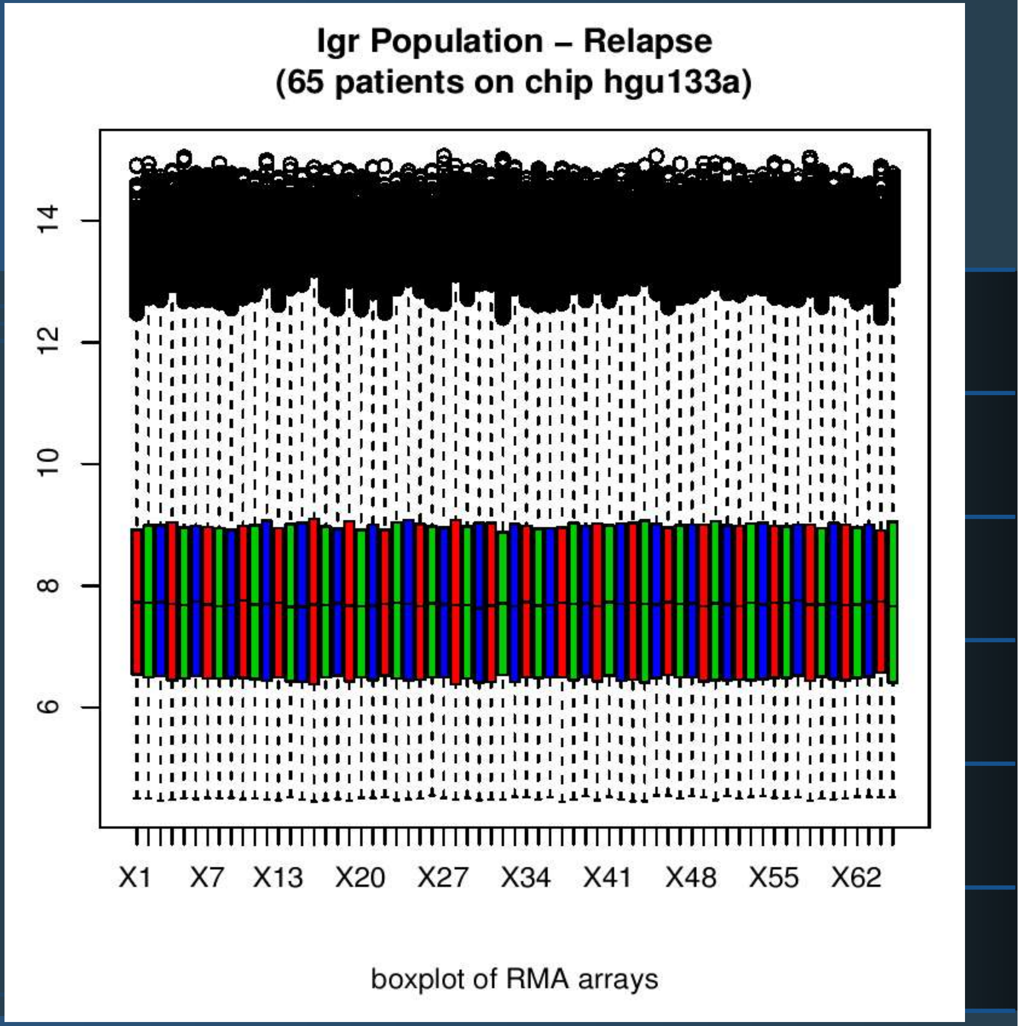
Methods and Results
Quality Assessment (12)





Methods and Results
Quality Assessment (13)





Methods and Results
Quality Assessment (14)

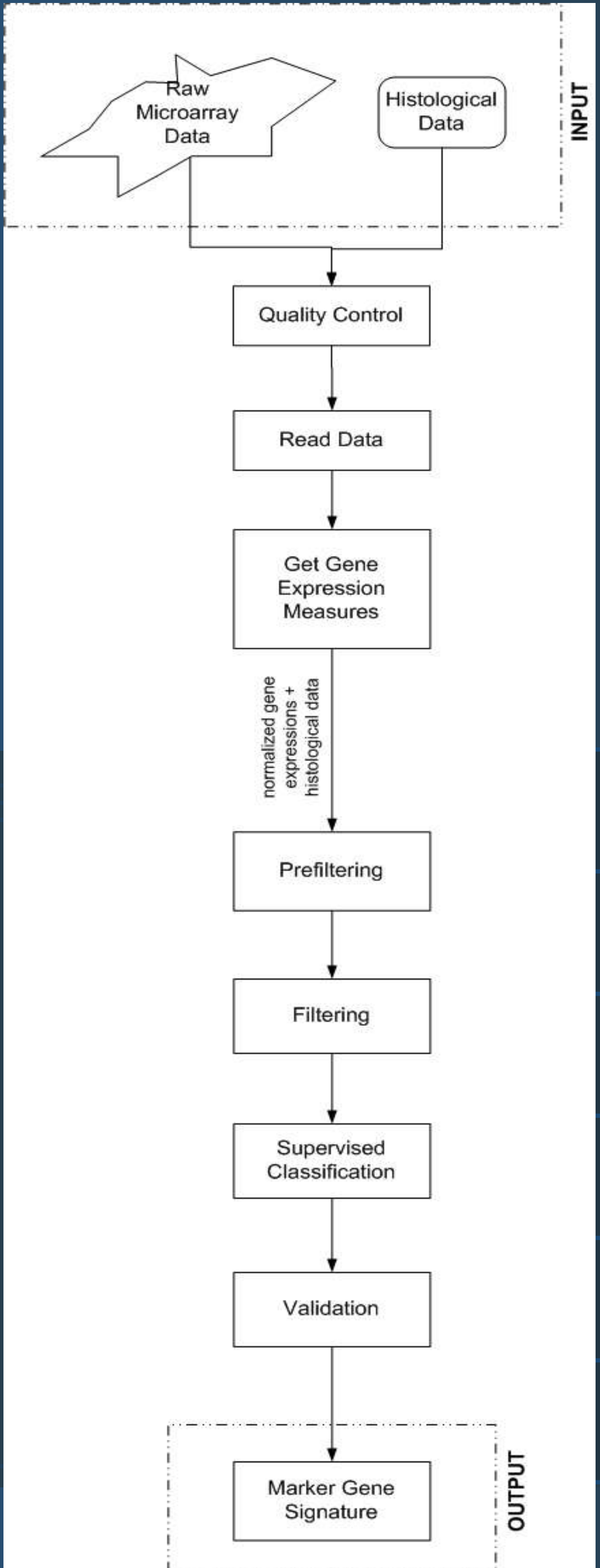
- Preliminary conclusion
  - Statistically significant difference between populations
  - Populations are not necessary comparable
  - Population preprocessing before analysis (not yet investigated)

Table of Contents

- Introduction
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- **Methods and Results**
  - Development Tools
  - Quality Assessment
  - **Supervised Classification**
  - Gene Ontology
- Discussion
  - Future Works

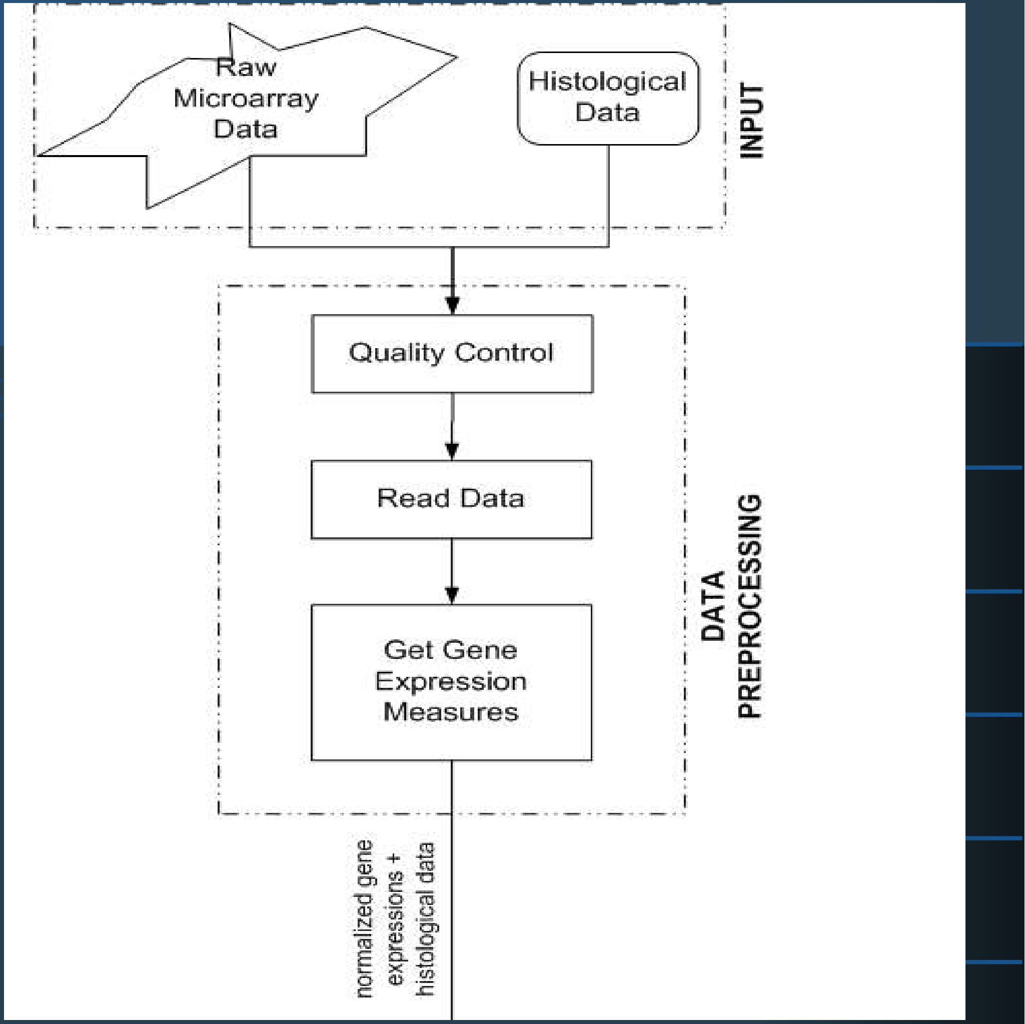
Methods and Results
Supervised Classification

- "Traditional" design of supervised classification in microarray data analysis



Methods and Results
Supervised Classification (2)

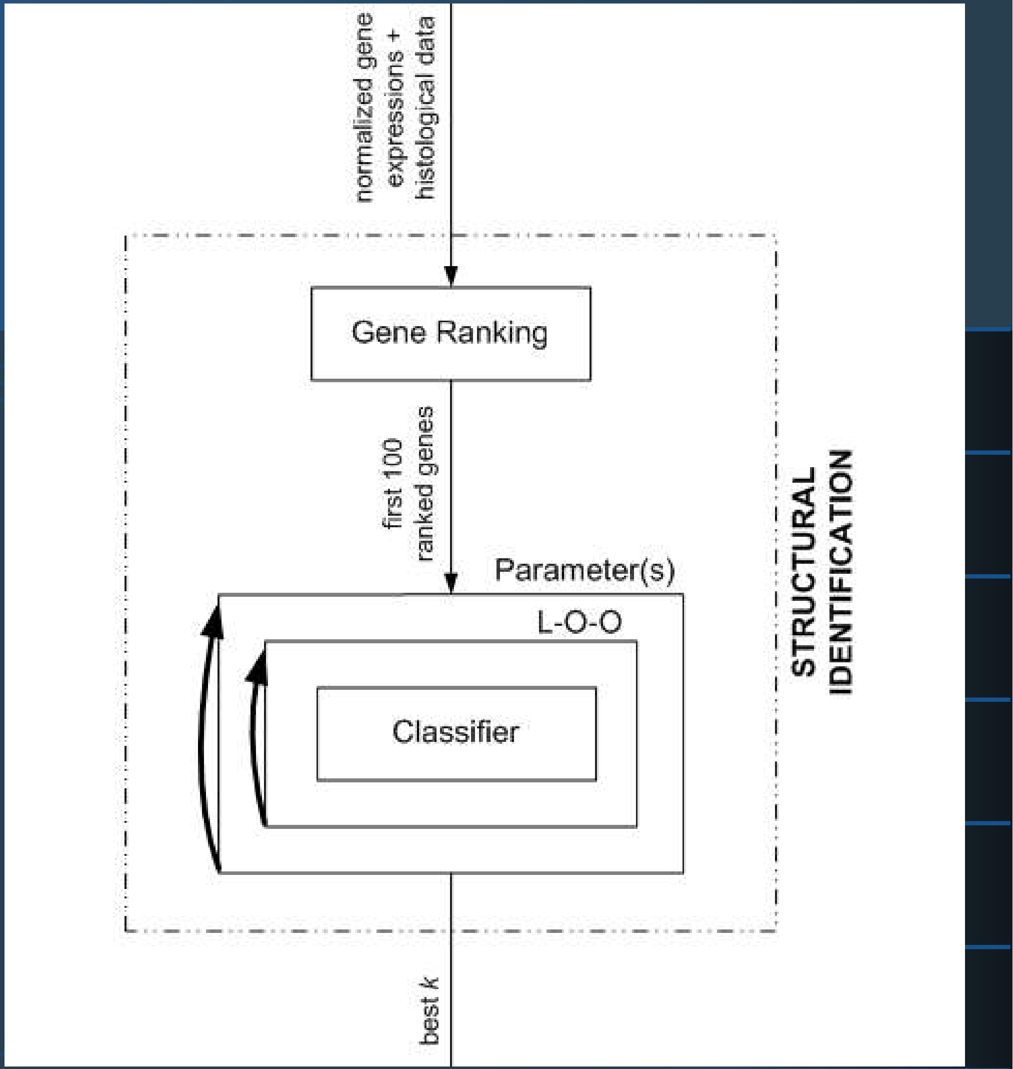
- Preprocessing Affymetrix data
  - Normalized gene expressions
  - Histological data



NB: only 99 patients have been considered in the classification procedure (52 from JRH and 47 from IGR)

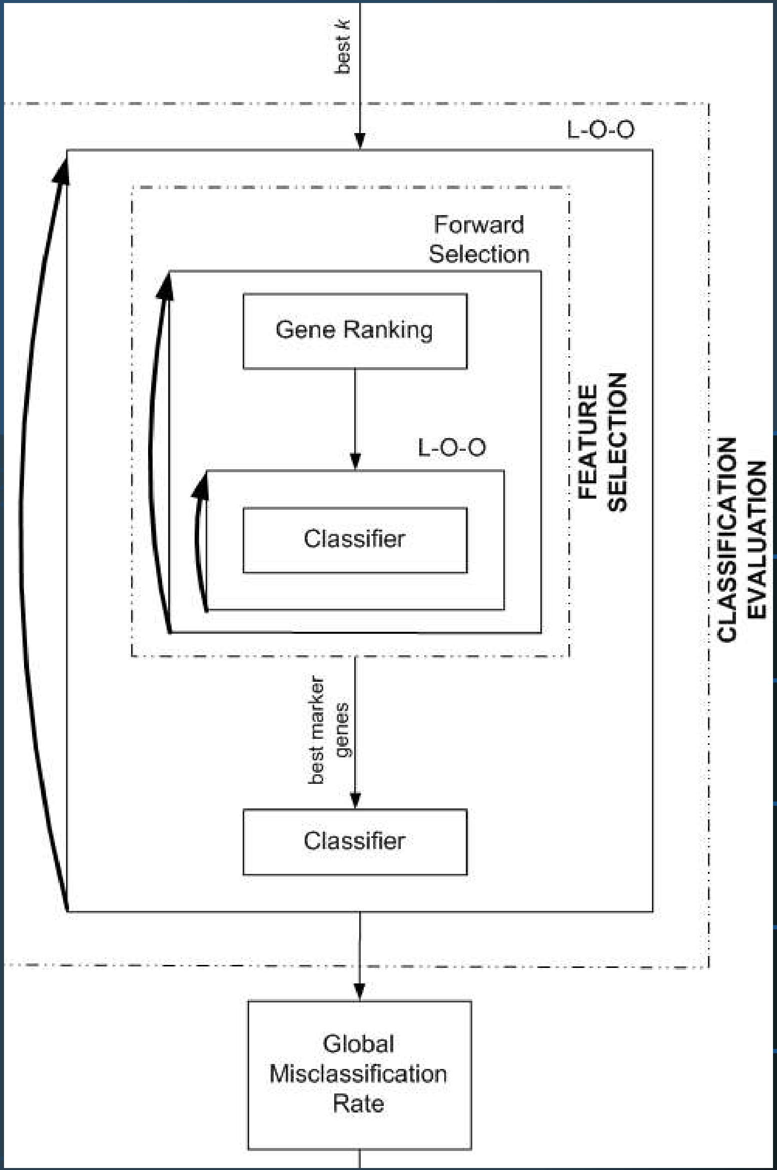
Methods and Results
Supervised Classification (3)

- Structural identification
  - **Gene ranking** by *Pearson* correlation coefficient
  - First 100 ranked genes (arbitrary criteria)
  - Classifier (KNN)
    - Parameter *k*


normalized gene expressions + histological data
Gene Ranking
first 100 ranked genes
Parameter(s)
L-O-O
Classifier
STRUCTURAL IDENTIFICATION
best k

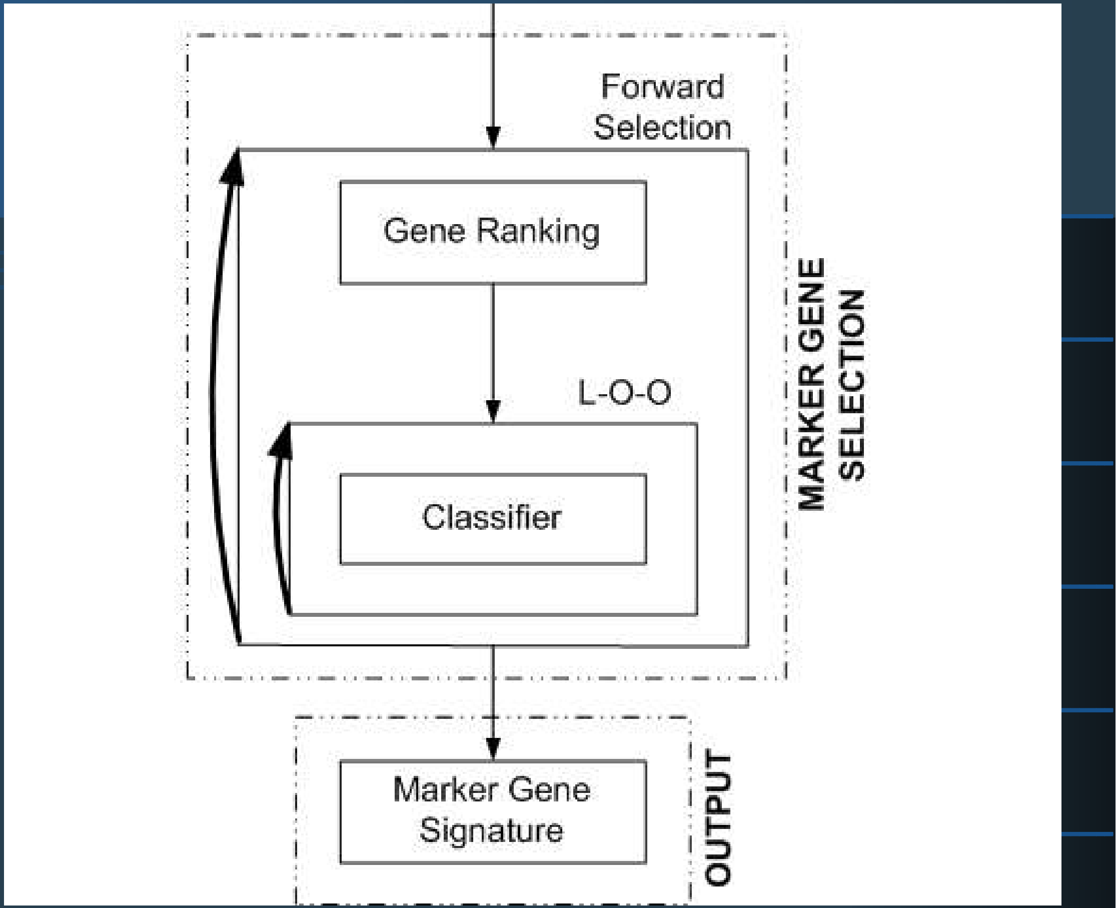
Methods and Results
Supervised Classification (4)

- Classification evaluation
  - **Feature selection** by variable ordering
  - At each L-O-O, a best set of marker genes is selected



Methods and Results
Supervised Classification (5)

- After classification procedure evaluation
- Marker gene selection with all the patients (using the same procedure)
- Assumption: the signature quality increases with the number of patients



Methods and Results
Supervised Classification (6)

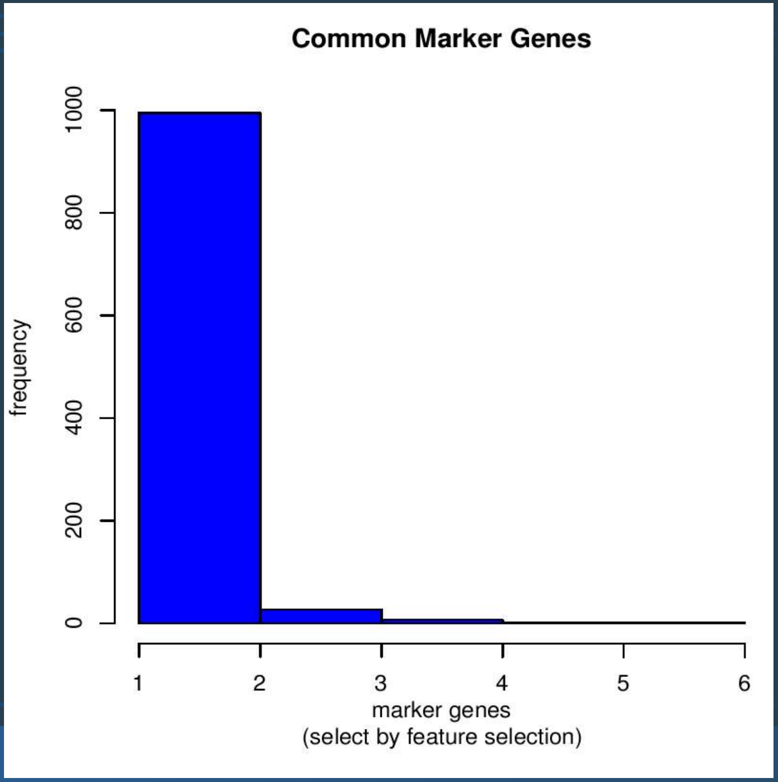
- Misclassification type

	Reality	
Prediction	**relapse (+)**	**non-relapse (-)**
relapse (+)	TP	FP
non-relapse (-)	**FN**	TN

- Class weights (classifier)
  - clw_0 = 1 for non-relapse class
  - **clw_1 = 10 for relapse class**
- Quality estimator (feature selection)
  - $q = clw_0 * \mathrm{FP} + clw_1 * \mathrm{FN}$

Methods and Results
Supervised Classification (7)

- Robustness of marker genes selected by the feature selections: frequency of appearance of each marker gene



Methods and Results
Supervised Classification (8)

- Signature is very dependent to the training set
- Expected result because of the **very small size of signatures** (relative to the number of genes)
  - 10 (mean) for the KNN
- Indication of poor biological information

Methods and Results
Supervised Classification (9)

- Global misclassification rate (KNN)
  - **FN: 21/24**
  - **FP: 4/75**

- Marker gene signature: 2 genes
  - 224529_s_at (C6ORF69)
  - 223176_at (NT5C1A)

Methods and Results
Supervised Classification (10)

- Preliminary conclusion
  - Avoid overfitting as much as possible according to computer resources
  - Tune the classifier to avoid a high false negative rate
  - Poor performance:
    - Arbitrary number of marker genes in the structural identification
    - KNN is sensible to unbalanced data set
    - High variance of the procedure (multiple L-O-O and feature selection)

Table of Contents

- Introduction
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- **Methods and Results**
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - **Gene Ontology**
- Discussion
  - Future Works

Methods and Results
Gene Ontology

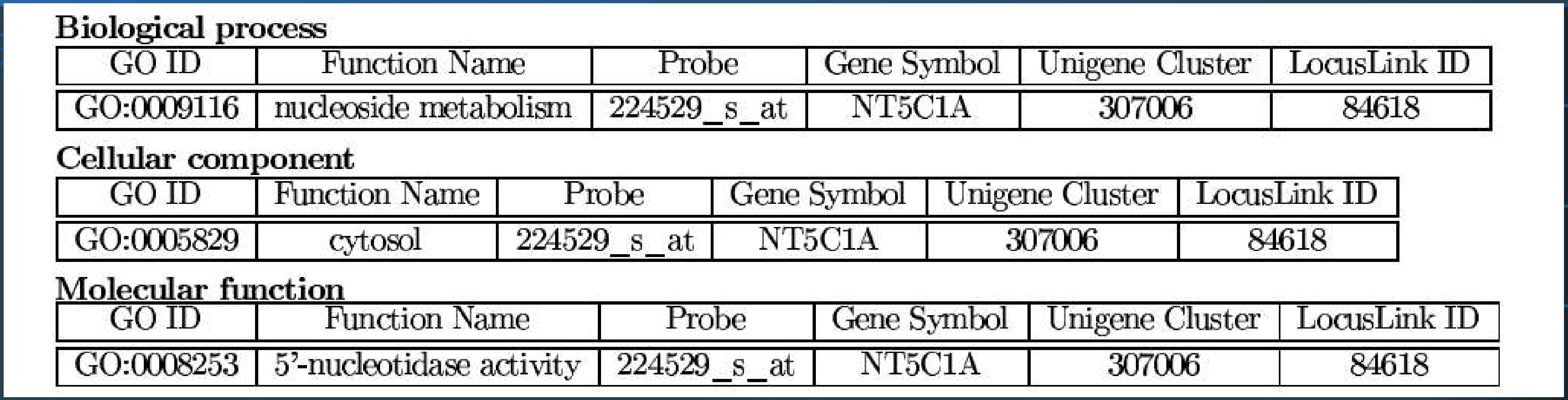
- GO consortium is setting a *dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing*
- Automatic annotation of marker genes in terms of
  - Molecular function
  - Biological process
  - Cellular component

Methods and Results
Gene Ontology (2)

- Onto-Express (Ostermeier *et al*. 2003)
- Statistical framework to assess the significance of gene clusters in each GO functional category
  - Take into account the tested genes (here the whole genome)
  - Take into account the set of marker genes
- Valuable if the number of marker genes in the signature is large (tens or hundreds)

Methods and Results
Gene Ontology (3)

- Not the case here: 2 marker genes
- Only one gene exists in the GO (**224529_s_at**)



Biological process

GO ID	Function Name	Probe	Gene Symbol	Unigene Cluster	LocusLink ID
GO:0009116	nucleoside metabolism	224529_s_at	NT5C1A	307006	84618

Cellular component

GO ID	Function Name	Probe	Gene Symbol	Unigene Cluster	LocusLink ID
GO:0005829	cytosol	224529_s_at	NT5C1A	307006	84618

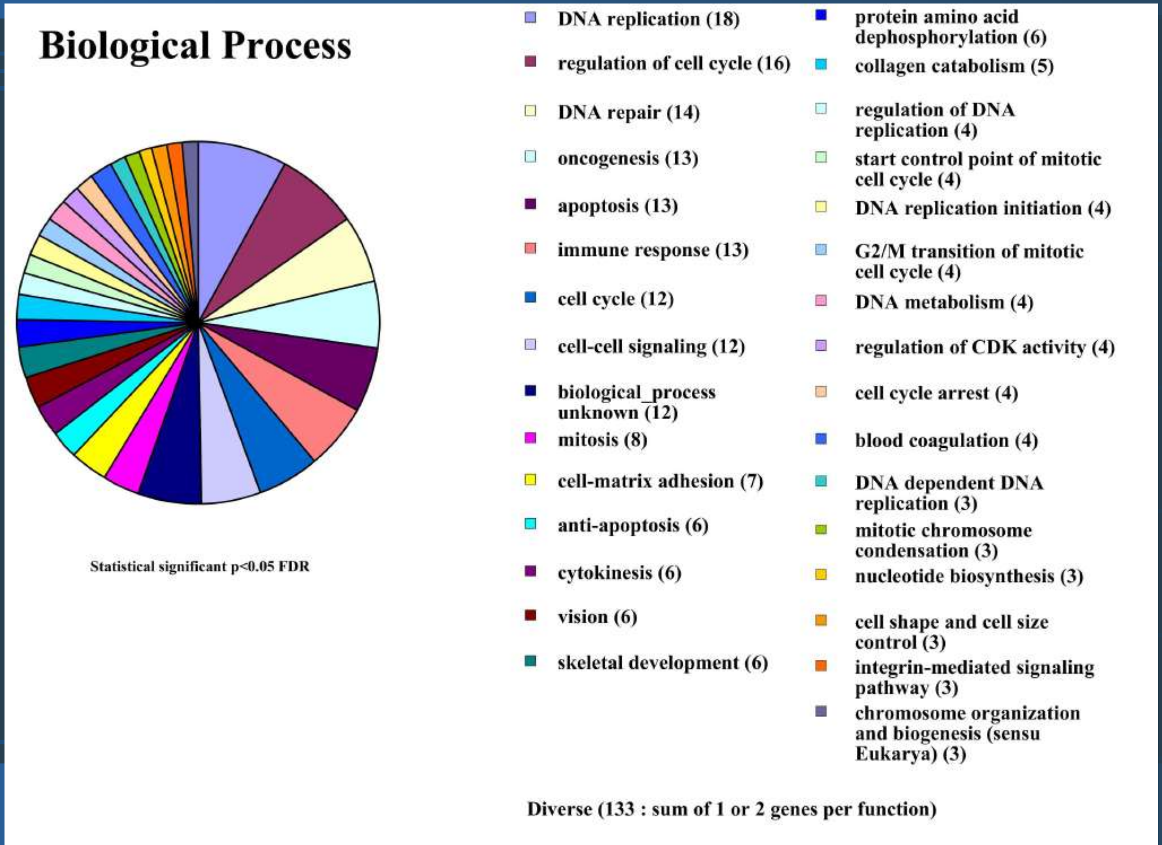
Molecular function

GO ID	Function Name	Probe	Gene Symbol	Unigene Cluster	LocusLink ID
GO:0008253	5'-nucleotidase activity	224529_s_at	NT5C1A	307006	84618

- The two genes are not known in breast cancer literature

Methods and Results
Gene Ontology (4)

- Collaboration: *Gene Regulation by Phorbol 12-myristate 13-acetate (PMA) in two Highly Different Breast Cancer Cell Lines*. Lacroix M, Haibe-Kains B, Laes JF, Hennuy B, Lallemand F, Gonze I, Cardoso F, Piccart M, Leclerq G, and Sotiriou C (in press, Oncology Report)



Statistical significant p<0.05 FDR

Diverse (133 : sum of 1 or 2 genes per function)

Table of Contents

- Introduction
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- Methods and Results
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- **Discussion**
  - Future Works

Discussion

- Microarrays have already provided valuable information about breast cancer
- Promising results in breast cancer diagnosis
- Issues need to be addressed before clinical use
  - Quality standards
  - Multi-populations, multi-platforms and multi-laboratories validation
  - Validation of marker gene expression by an alternative RNA quantitative method (e.g. RT-PCR)

Table of Contents

- Introduction
  - TransBIG Project
- Materials
  - Populations
  - Microarray Platform
- Methods and Results
  - Development Tools
  - Quality Assessment
  - Supervised Classification
  - Gene Ontology
- **Discussion**
  - **Future Works**

Discussion
Future Works

- Step by step complexity of analysis design
- Statistical framework for quality assessment
- Parallelism
- Preprocessing data
- Criterion for misclassification rate
- Marker gene stability
- Feature selection
- Independent validation set
- Signature validation and refinement

Applications to Genomic and Proteomic Data

Thanks for your attention

Benjamin Haibe-Kains

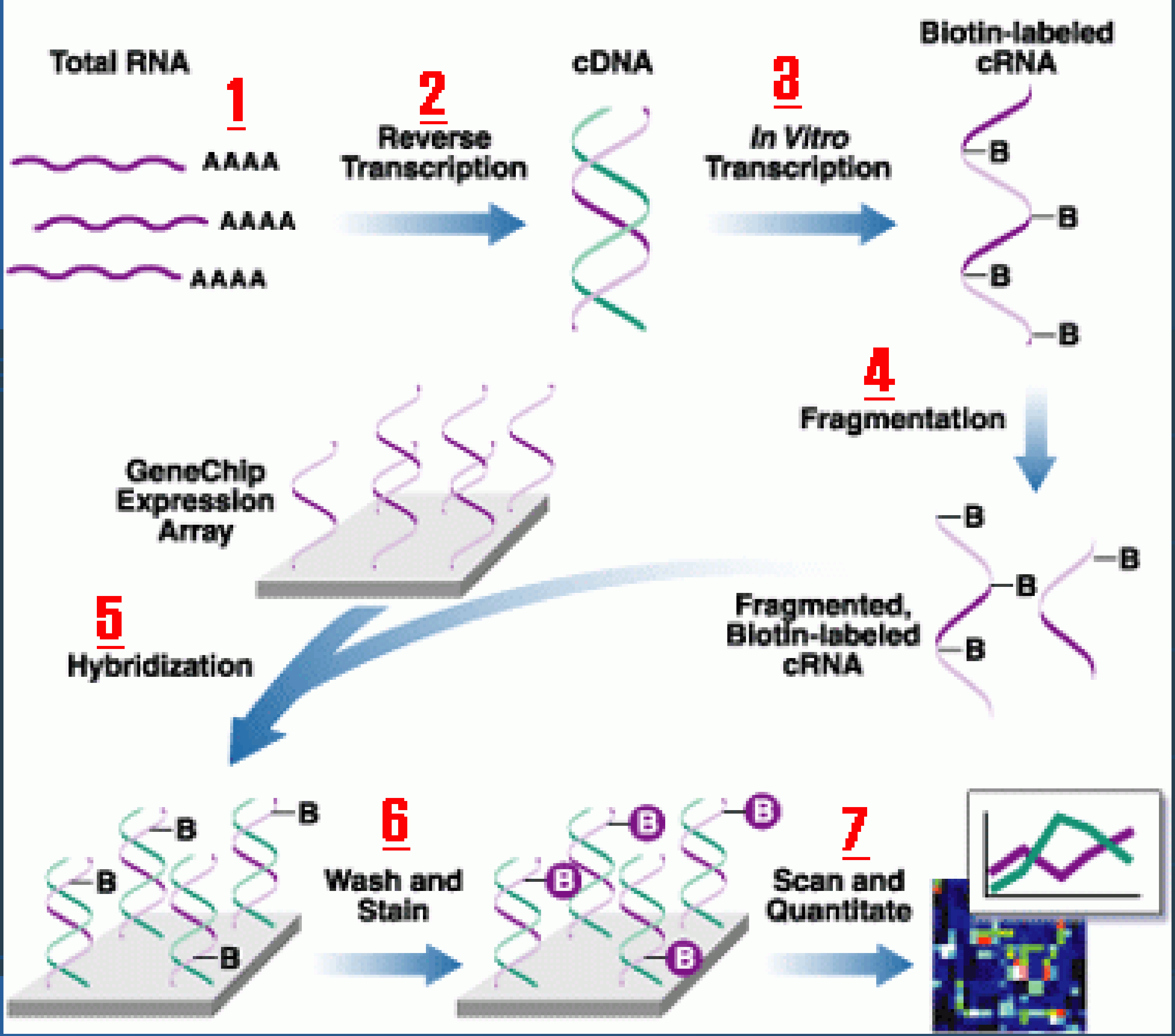
Thesis

Appendix

Materials
Populations

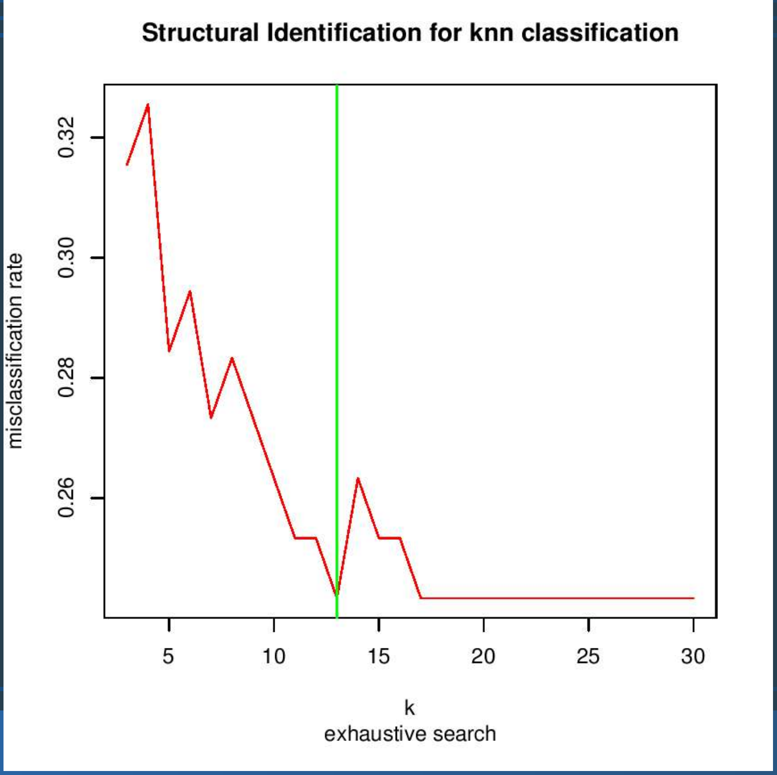
- Lymph node negative
- Not treated by adjuvant treatment

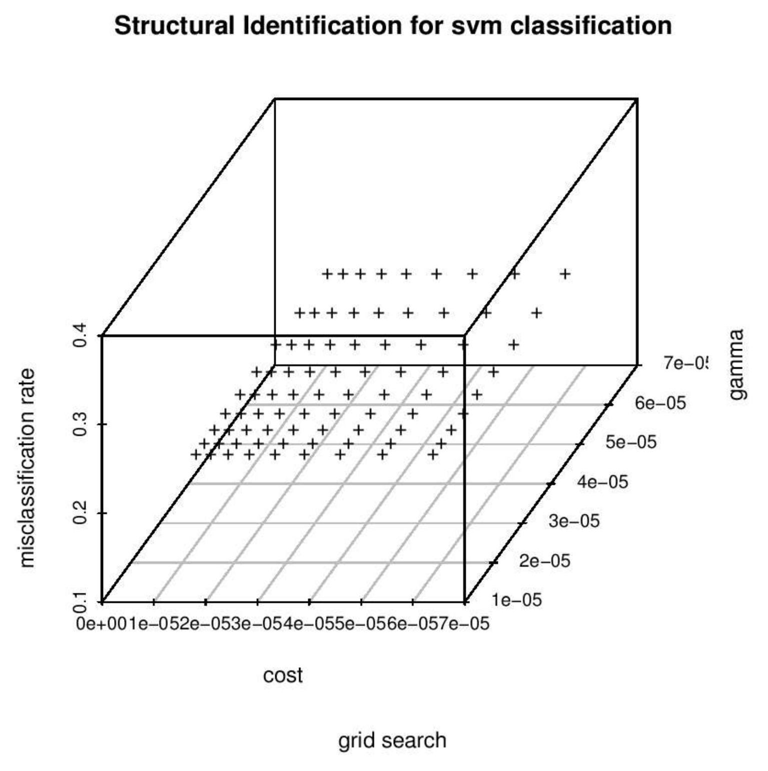
Materials
Microarray Platform



Methods and Results
Structural Identification

- First 100 ranked genes with all patients
  - KNN
  - SVM





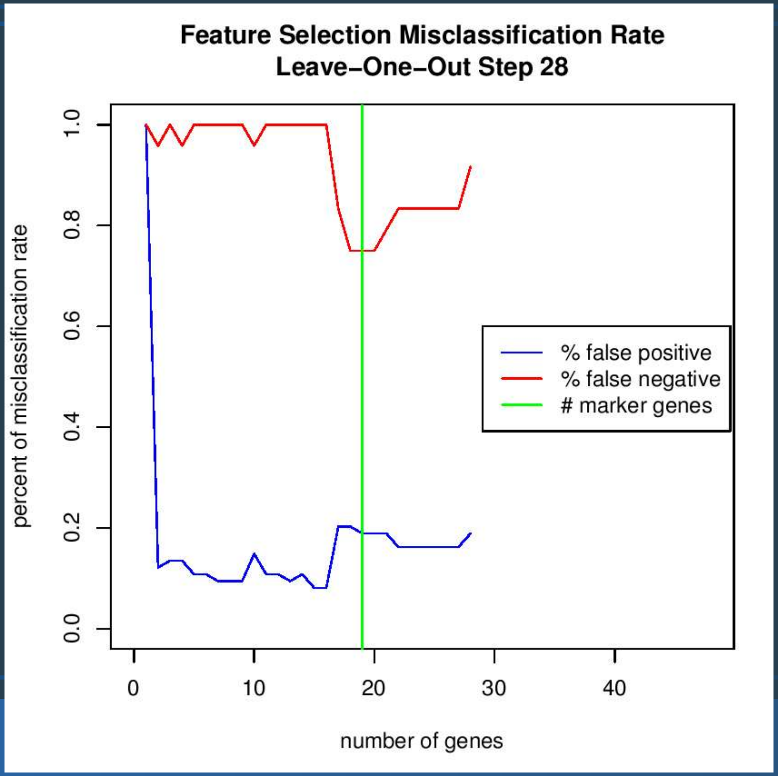
Methods and Results
Structural Identification

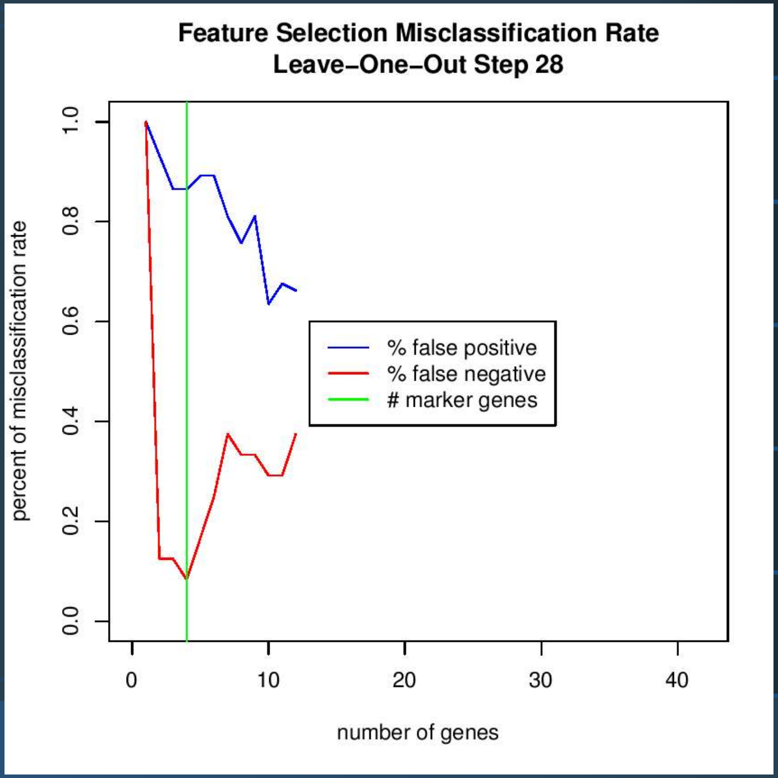
- Use of *tune.foo* R function
  - Low execution time (relative to the complexity)
  - Only global misclassification rate
  - No class weights
  - Leave-one-out cross-validation
  - Approximatively 25% of misclassification

→ **No indication about FN and FP**

Methods and Results
Feature Selection

- Misclassification rate: **opposite trend** between KNN and SVM classifiers
  - KNN
  - SVM





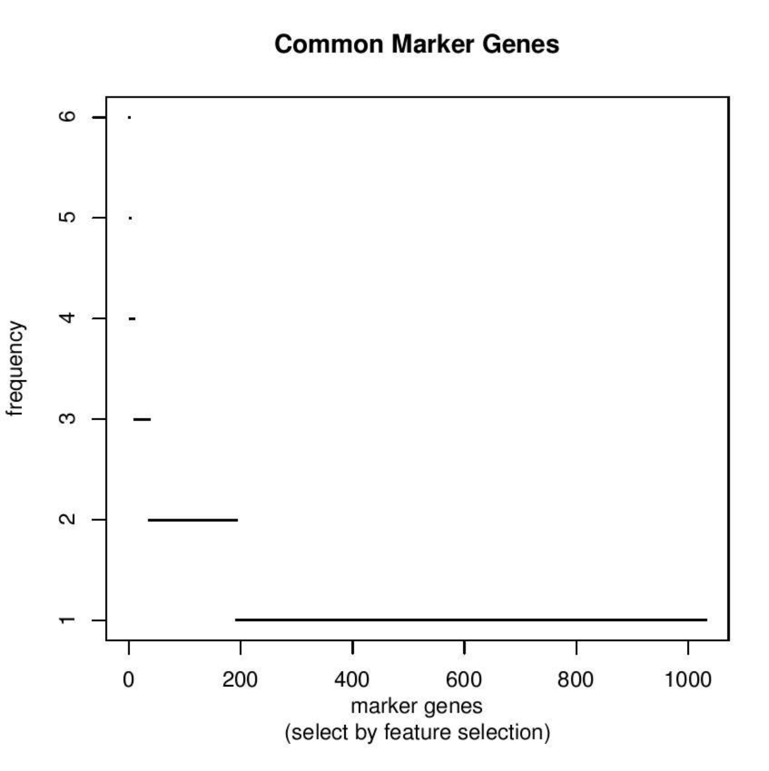
Methods and Results
Feature Selection (2)

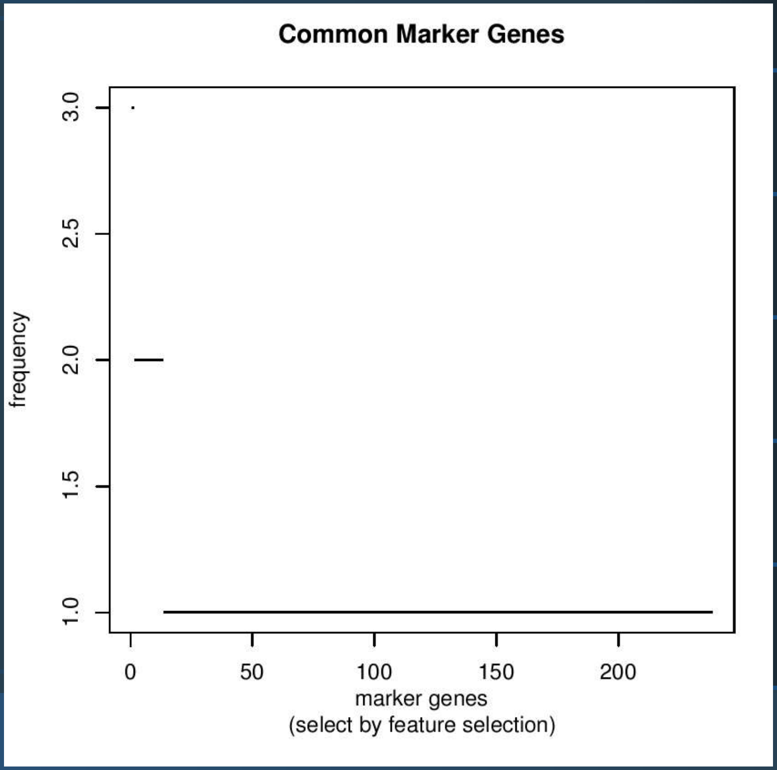
- Due to
  - No class weight for the KNN
  - KNN is more sensible to unbalanced data set

- Robustness of marker genes selected by the feature selections: frequency of appearance of each marker gene

Methods and Results
Feature Selection (3)

- Common marker genes during global leave-one-out
  - KNN
  - SVM



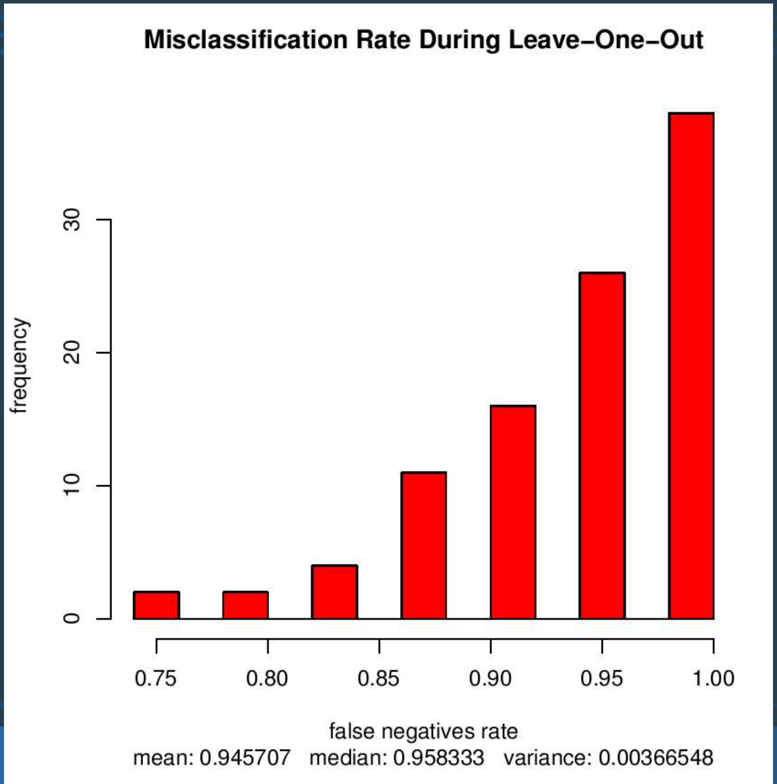


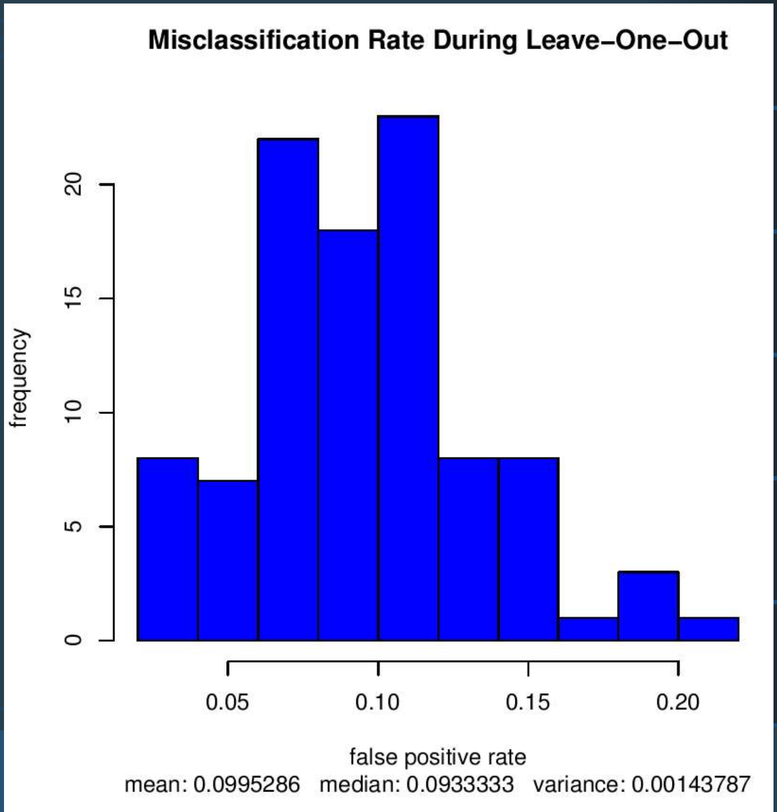
Methods and Results
Feature Selection (4)

- Similar observations for the KNN and the SVM classifiers
  - Signature is very dependent to the training set
  - Expected result because of the **very small size of signatures**
    - 10 (mean) for the KNN
    - 2 (mean) in the SVM
  - Indication of poor biological information

Methods and Results
Misclassification Rate

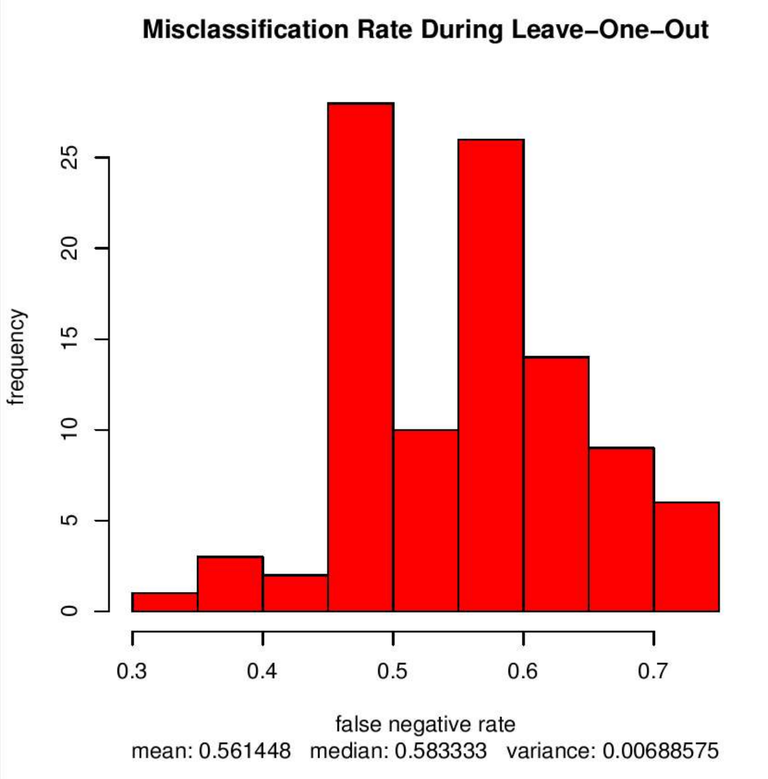
- KNN: misclassification during feature selections (global → **21/24** and **4/75**)
  - False negatives
  - False positives

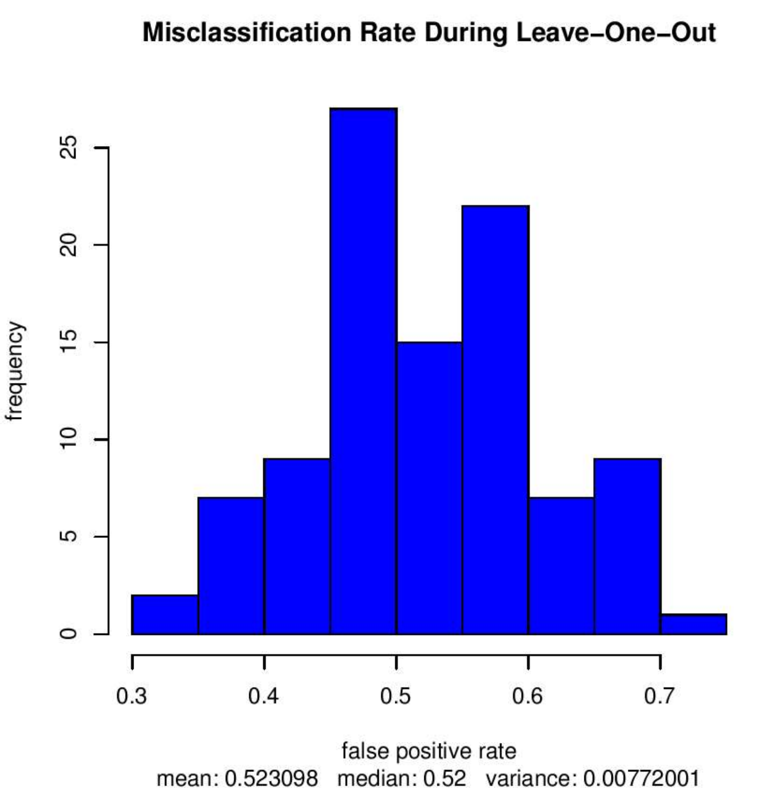




Methods and Results
Misclassification Rate (2)

- SVM: misclassification during feature selections (global → **2/24** and **65/75**)
  - False negatives
  - False positives





Methods and Results
Gene Ontology

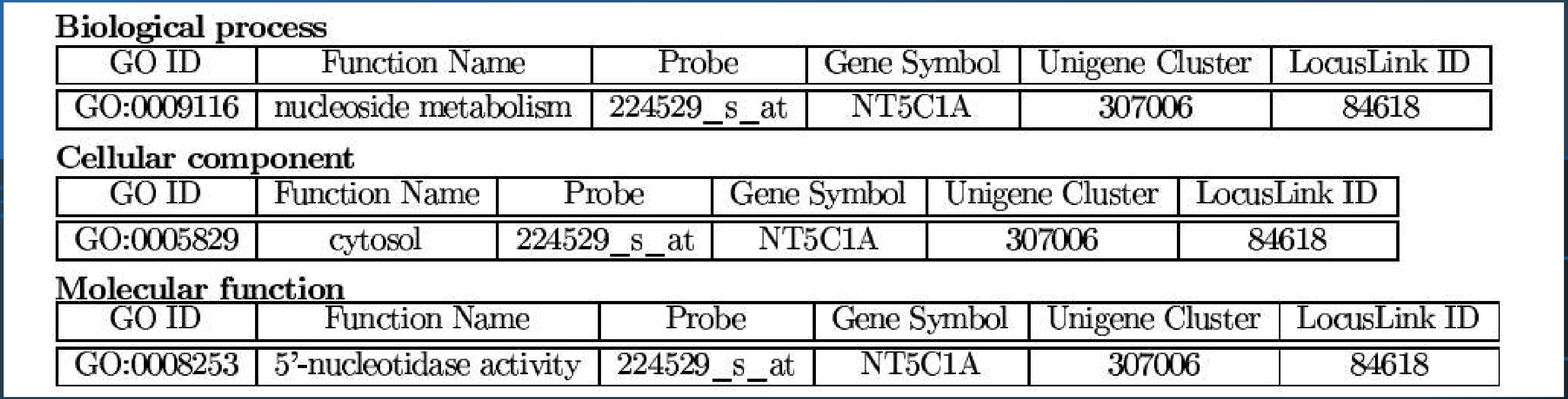
- Probe set id: **223176_at**
- Accession number: BC003697
- Gene name: chromosome 6 open reading frame 69
- Symbol: C6ORF69
- Unigene: Hs.188757

Methods and Results
Gene Ontology

- Probe set id: **224529_s_at**
- Accession number: AY028778
- Gene name: 5'-nucleotidase, cytosolic IA
- Symbol: NT5C1A
- Unigene: Hs.307006

Methods and Results
Gene Ontology (3)

- Only one gene exists in GO (**224529_s_at**)



Biological process

GO ID	Function Name	Probe	Gene Symbol	Unigene Cluster	LocusLink ID
GO:0009116	nucleoside metabolism	224529_s_at	NT5C1A	307006	84618

Cellular component

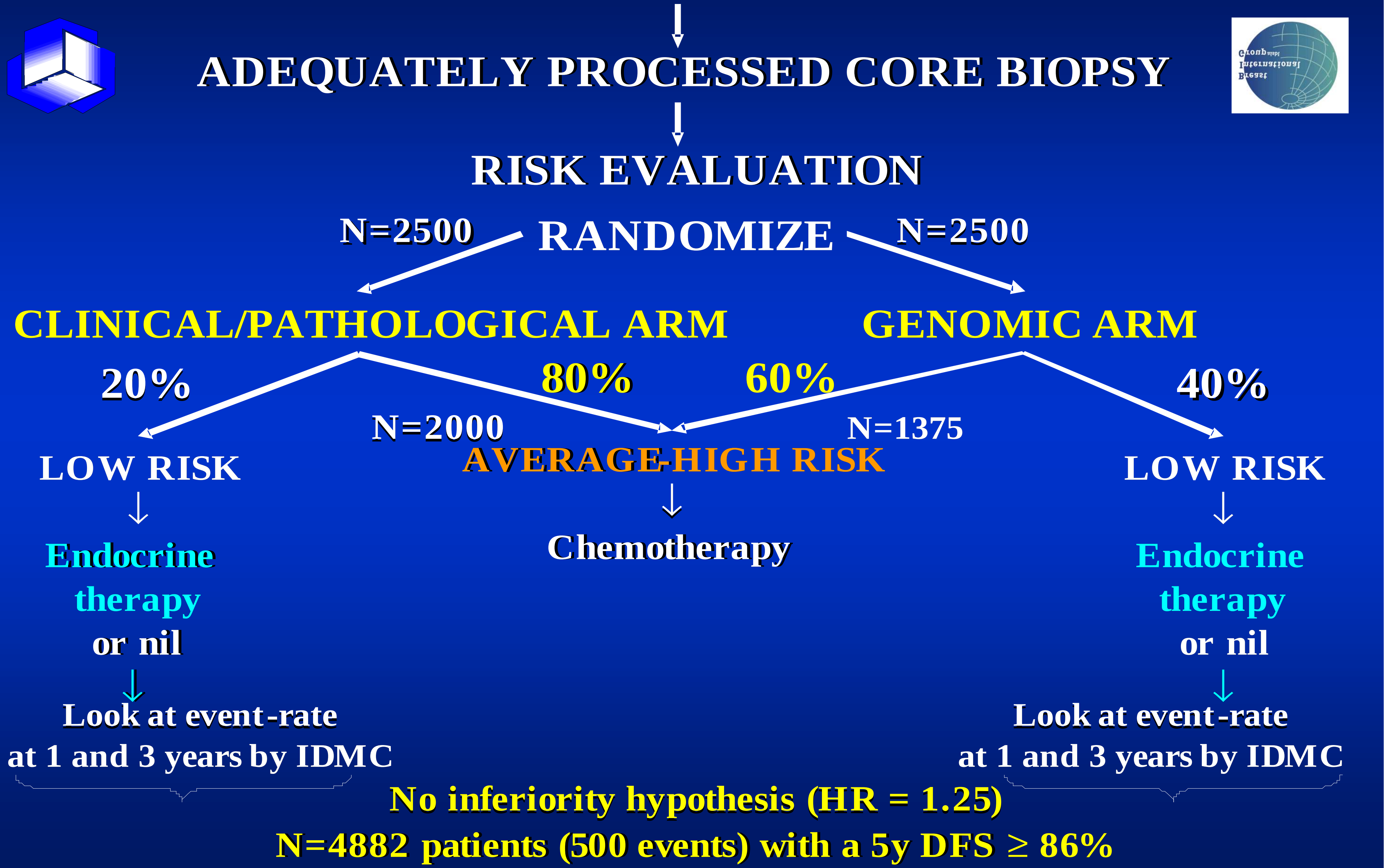
GO ID	Function Name	Probe	Gene Symbol	Unigene Cluster	LocusLink ID
GO:0005829	cytosol	224529_s_at	NT5C1A	307006	84618

Molecular function

GO ID	Function Name	Probe	Gene Symbol	Unigene Cluster	LocusLink ID
GO:0008253	5'-nucleotidase activity	224529_s_at	NT5C1A	307006	84618

- Nucleoside: combination of a base and a sugar without phosphate
- Nucleotide: nucleoside with 1, 2, or 3 phosphate groups
- Nucleotidase: enzyme hydrolizing nucleosides to nucleotides; the proportioning of the serum 5'-nucléotidase is used in digestive pathology

EORTC-BIG NODE NEGATIVE BREAST CANCER TRIAL



ADEQUATELY PROCESSED CORE BIOPSY

RISK EVALUATION

N=2500 RANDOMIZE N=2500

CLINICAL/PATHOLOGICAL ARM

GENOMIC ARM

20%

80%

60%

40%

N=2000

N=1375

LOW RISK

↓

Endocrine therapy or nil

↓

Look at event-rate at 1 and 3 years by IDMC

AVERAGE-HIGH RISK

↓

Chemotherapy

LOW RISK

↓

Endocrine therapy or nil

↓

Look at event-rate at 1 and 3 years by IDMC

No inferiority hypothesis (HR = 1.25)

N=4882 patients (500 events) with a 5y DFS ≥ 86%