Dissecting an alternative splicing analysis workflow for GeneChip® Exon 1.0 ST Affymetrix arrays

Ljubljana 17 April 2009

raffaele.calogero@unito.it

Bioinformatics & Genomics unit

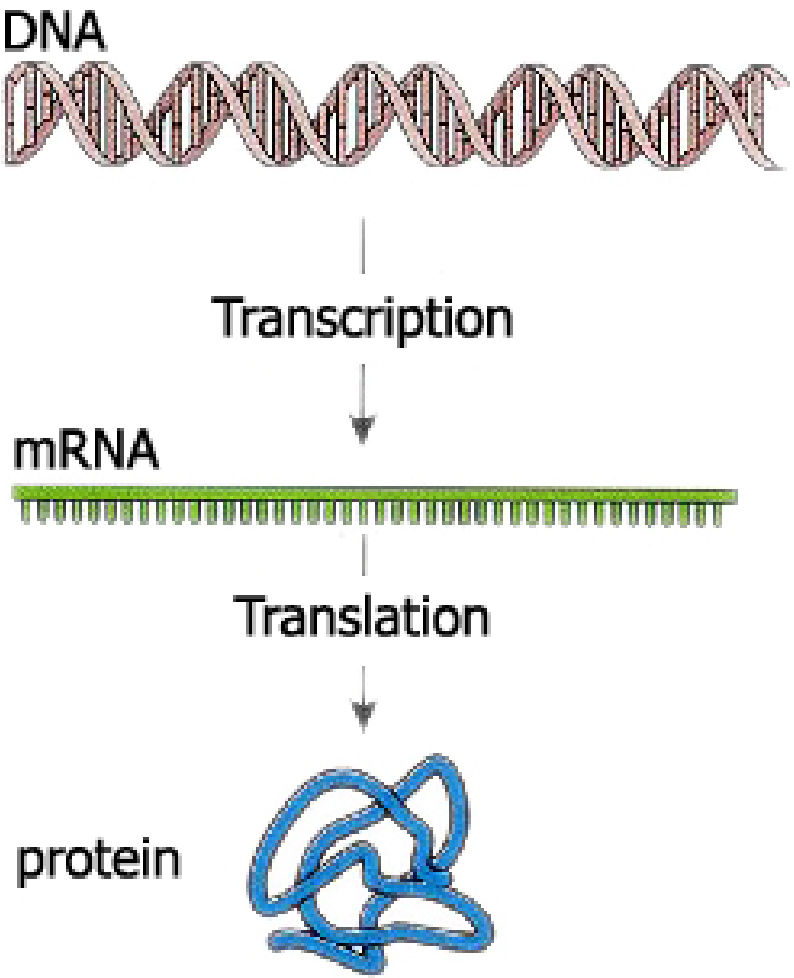
Agenda

- First part
  - Dissecting alternative splicing workflow
- Second part
  - Softwares for exon-array analysis

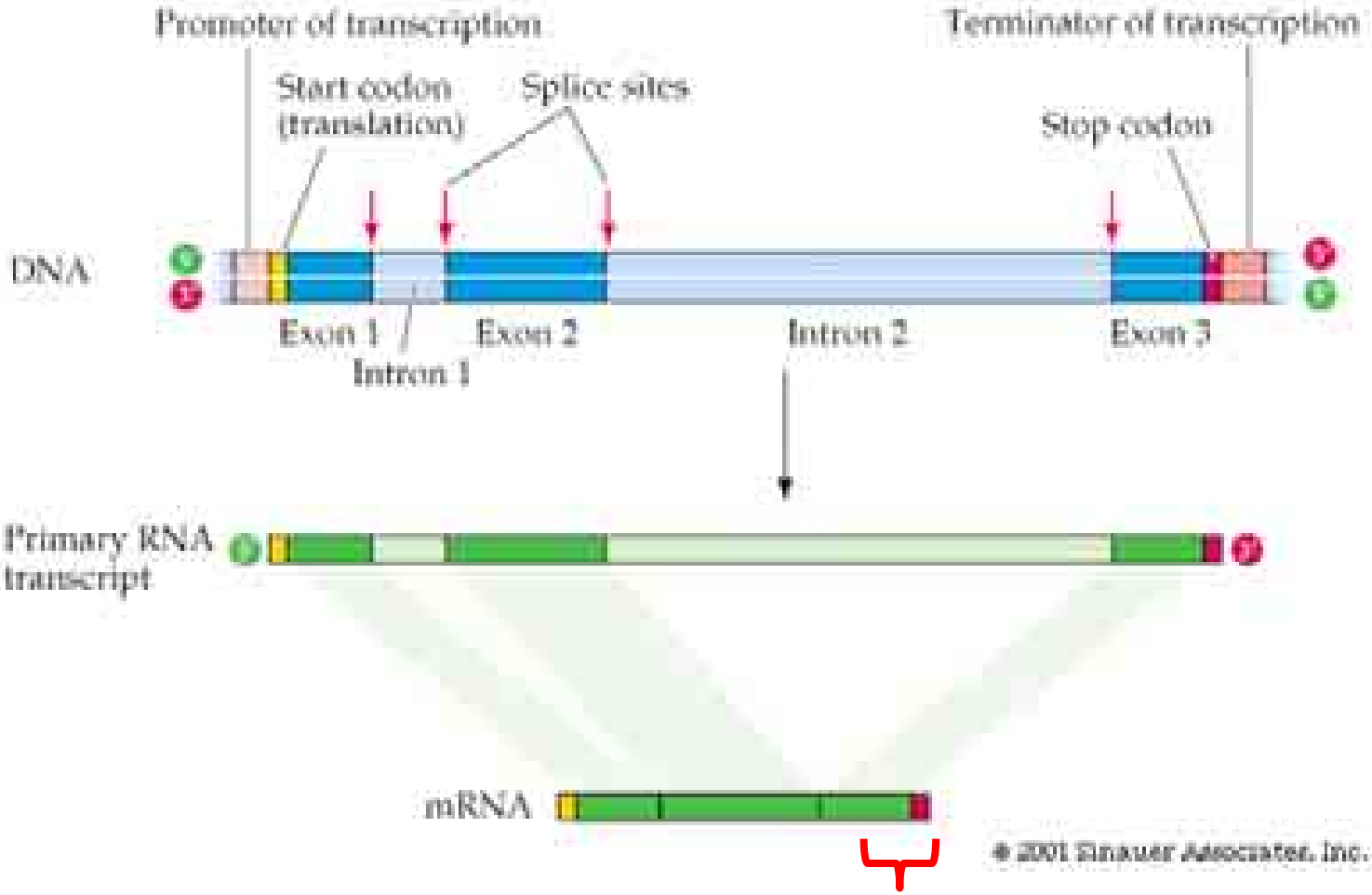
Agenda

- First part
  - Dissecting alternative splicing workflow
- Second part
  - Softwares for exon-array analysis

Central dogma of Biology

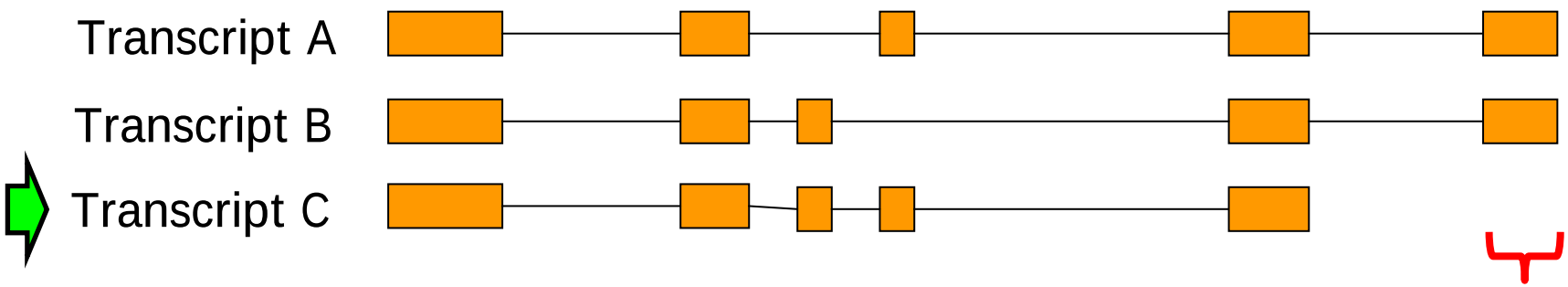


Structure and transcription of a Eukaryotic gene



Measuring transcript level expression

Alternative splicing

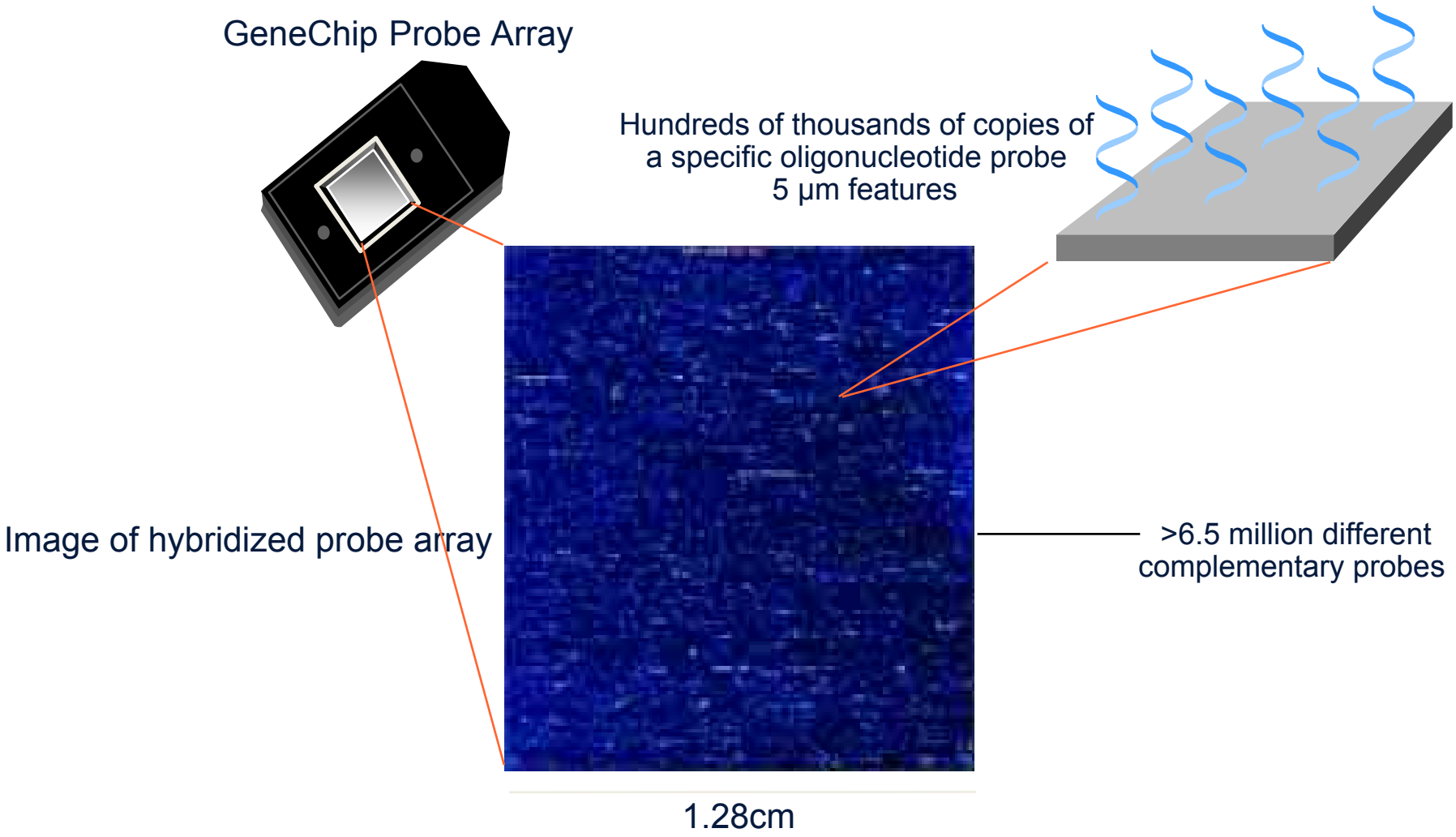


- Objective:
  - Detect the concentration of each transcript isoform.
- Available instrument:
  - Microarray detecting exons concentration
- General issues:
  - Isoforms changes in relative concentrations.
  - Yes/No events are very rare.
  - Gene-level effect should be removed in calculating concentration variation at exon-level
  - Microarray measurements are quite noisy
  - Multiple testing issues

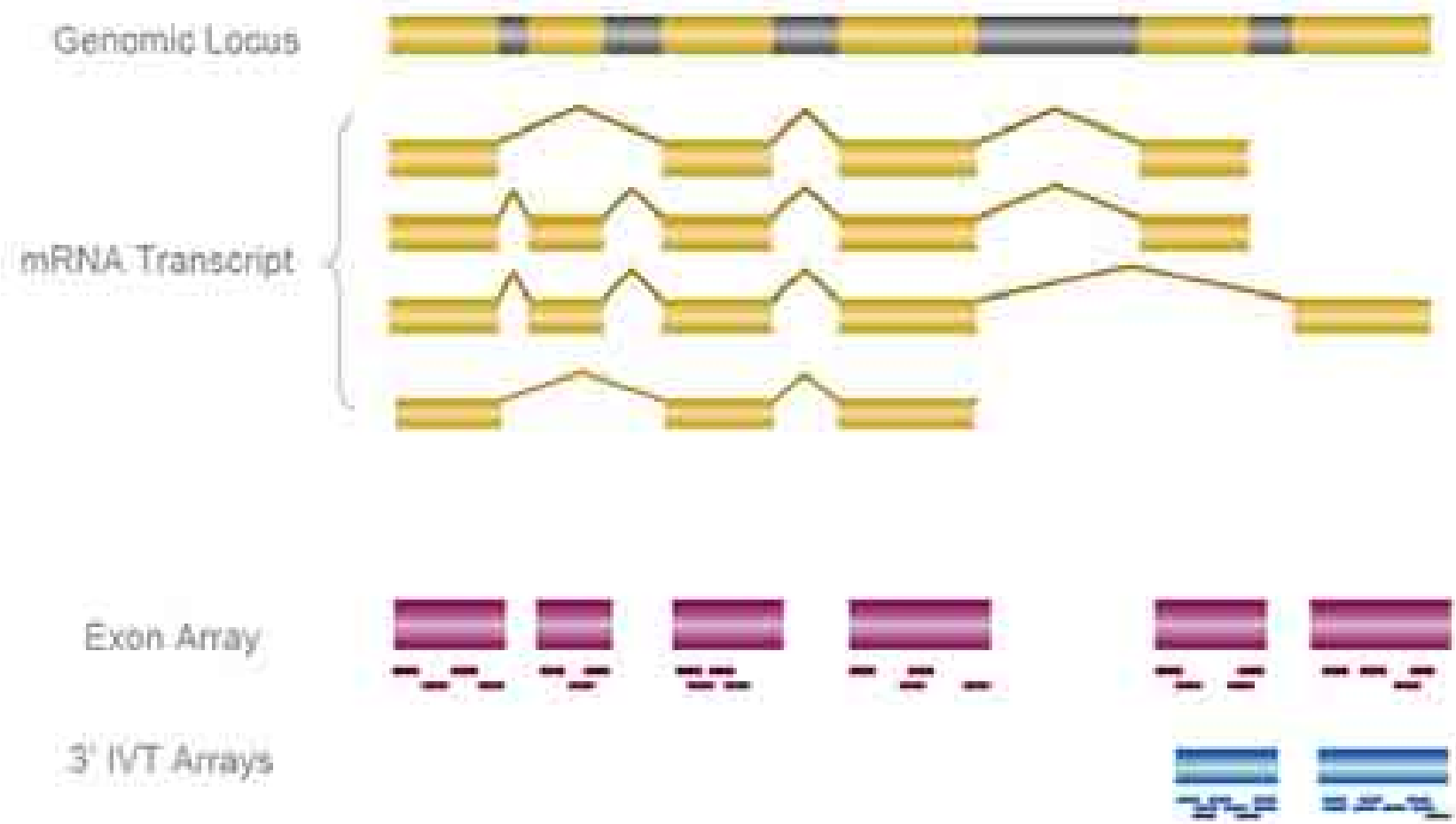
What is Microarray

- A powerful technology for biological exploration which enables to simultaneously measure the level of activity of thousands transcripts.
- The amount of mRNA for each gene in a given sample (or a pair of samples) is measured.
- Microarrays are:
  - Parallel
  - High-throughput
  - Large-scale
  - Genomic scale

GeneChip® Probe Arrays



Exon 1.0 ST genechips

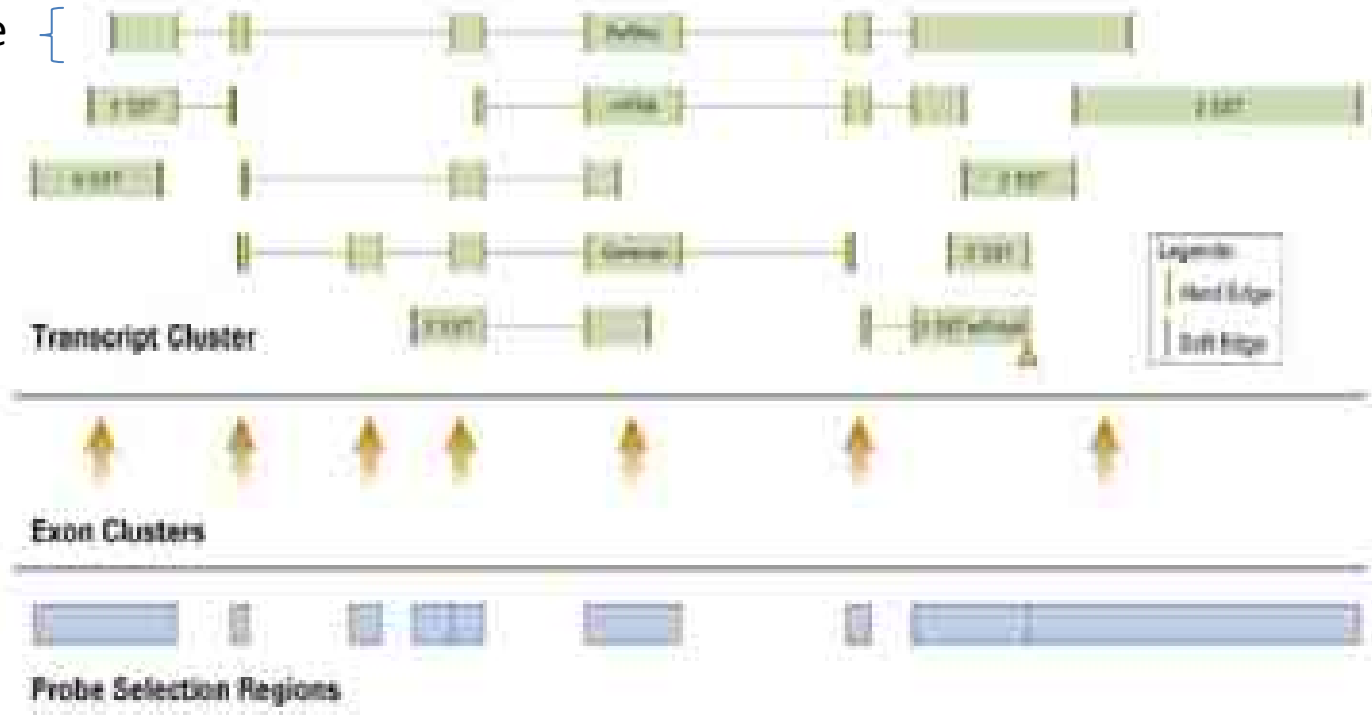


Exon 1.0 ST genechips

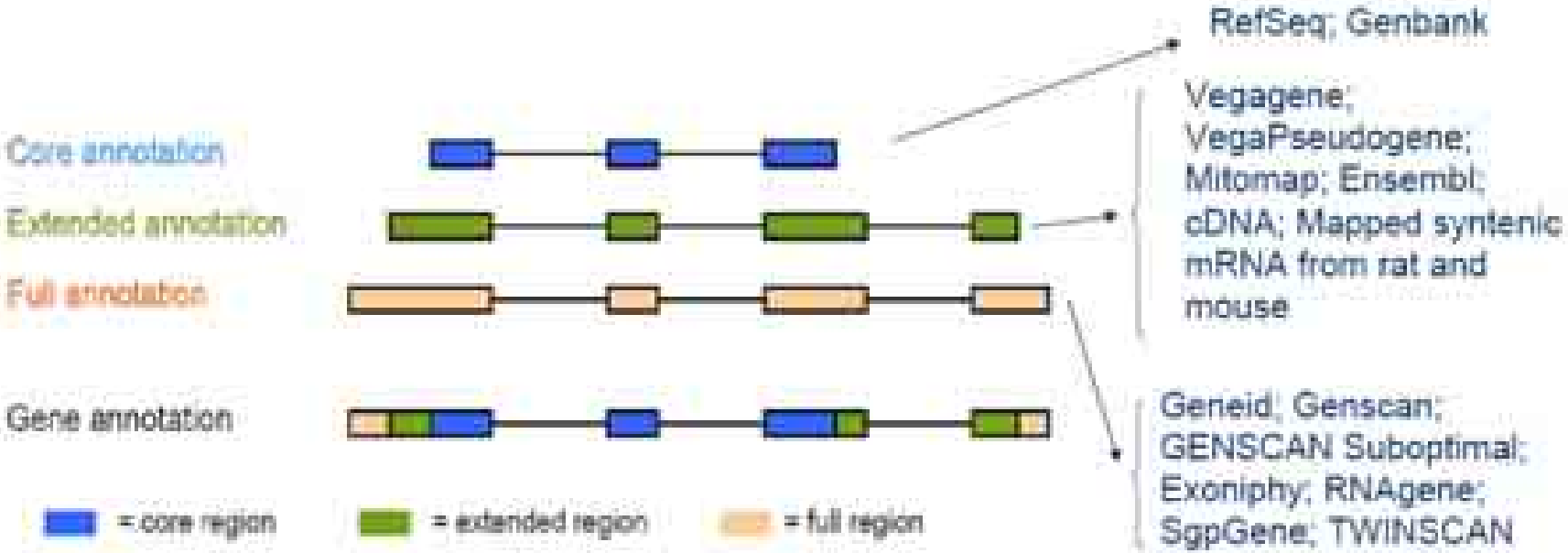
Full

Extended

Core



Exon Annotation Levels



- Core annotation used to study changes of known isoforms
- Extended/full annotations might be used for new isoforms discovery

Affymetrix exon arrays

- Probably still most dense and complex general purpose expression microarray since its introduction to the market

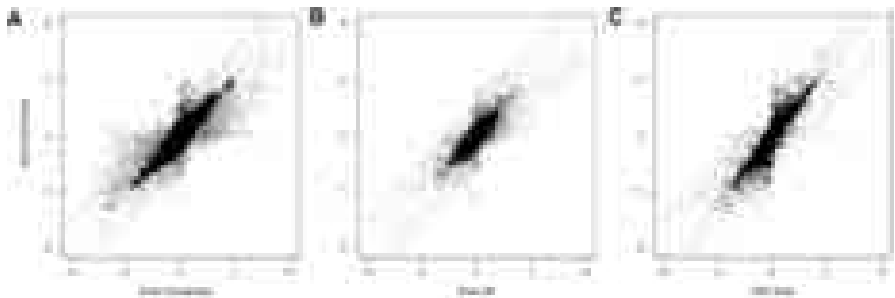
- Unique combination of genomic evidence to design the chip
  - over 1 M known and predicted exons
  - more than 12 databases of genomic evidence combined
  - 70-80% of genes are alternatively spliced, so there is no “gene level expression”

- Gives the chance to get new types of biological questions answered

- Advanced transcriptomics and advanced computing required

Exon arrays - advantages

- Coverage
  - > 6M probes
  - > 1M exons
- Granularity
  - On average >25x more probesets per gene
- Quality



- The chance to answer new biological questions
  - Splicing (Wow, a new exon in my favourite gene!)
  - Isoforms (Is the long isoform more related to my type of cancer than the short?)
- One can still do standard (gene-oriented) expression studies

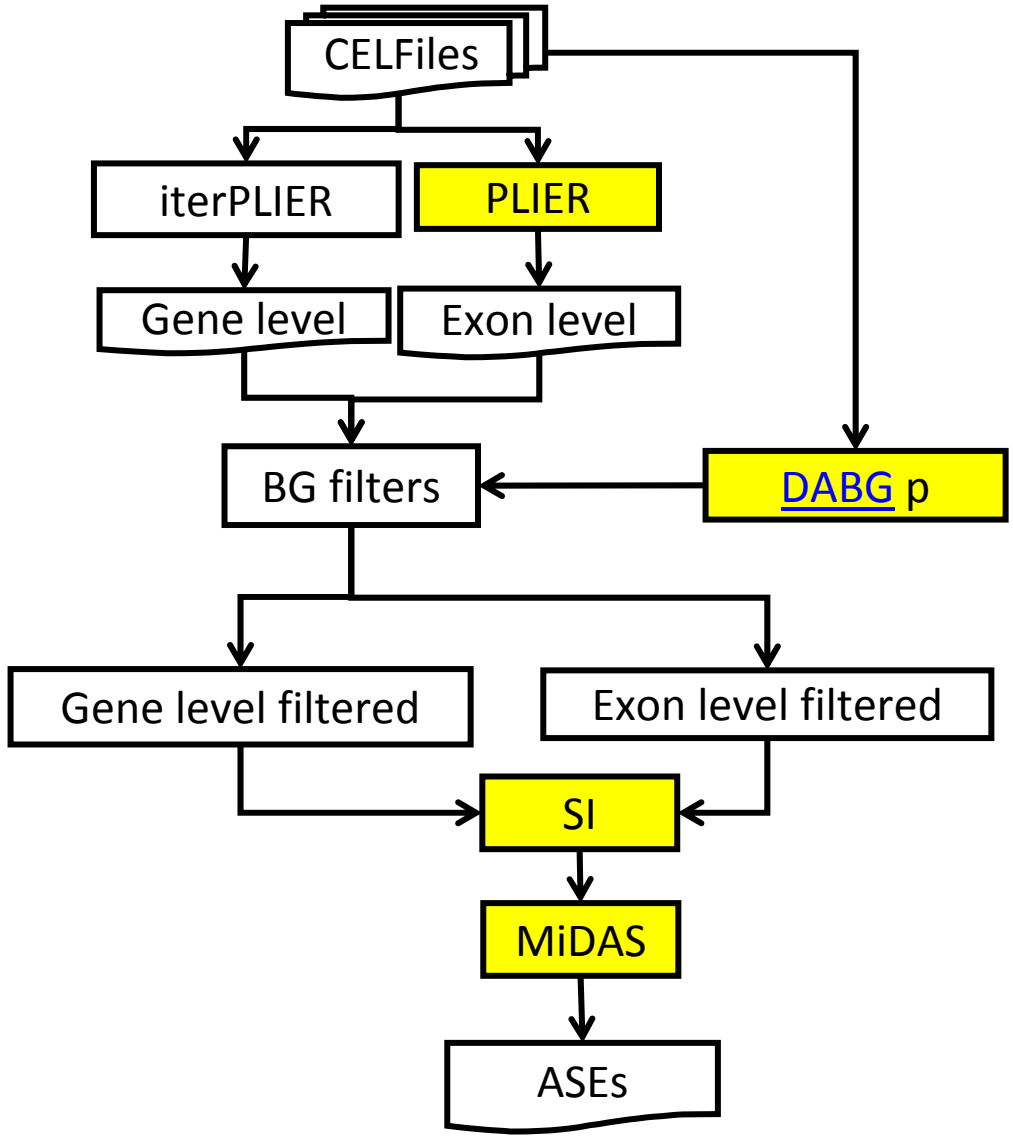
Okoniewski et al,Biotechniques 2007

Exon arrays -challenges

- The need for sophisticated bioinformatics
  - Controlling gene-transcript-exon-probeset-probe structures
  - Cross-hybridization and multiple targeting (probes that hit in many places of the genome)
  - Evolving annotations
- Huge amount of data
  - A database of annotations needed
  - A database to store results...?
- Great variety of splicing events to analyse
  - Exon skipping
  - Mixtures of isoforms
  - Functional domains....

Alternative splicing events (ASEs) detection

- Affymetrix proposed the following exon-level data analysis workflow:
  - Gene/exon level signal calculation.
  - Removal of non-informative signals.
  - Calculation of Splice Index.
  - Detection of alternative splicing events via ANOVA method.
- **No information on the efficacy of this workflow is given by Affymetrix.**



Alternative splicing analysis open issues:

- Intensity signal calculation affects alternative splicing analysis?
- Which is the efficacy of statistics used in alternative splicing analysis?
- Is it possible to moderate multiple testing errors?

Alternative splicing analysis

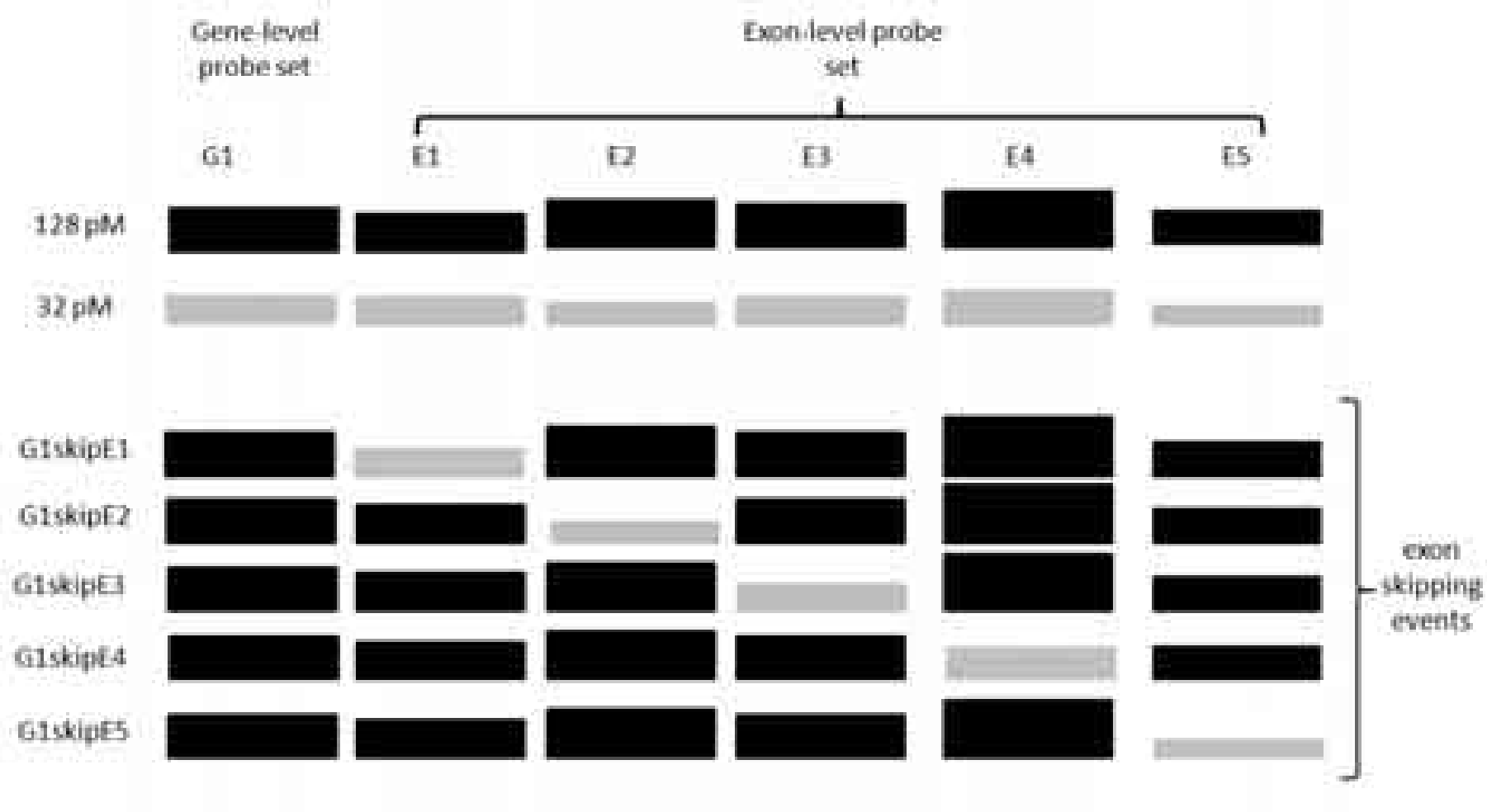
- To compare statistical methods for alternative splicing as well as the effects of various filtering we have developed a semi-synthetic exon skipped data set.
- To create this data set we started from the latin-square experiment of Abdueva (2007):
  - 25 genes were selected as ideal spike-in genes due to their expression absence in HeLa cells.
  - The spike-in concentration were 0, 2, 32, 128 and 512 pM
  - The 25 genes were grouped in 5 subset.
  - Each experimental point was technically replicated three times for a total of 15 arrays.

Alternative splicing analysis

- For the construction of the exon-skipping benchmark experiment we used 4 out of the 5 groups of spike-in genes.
- We focus on those because they were all part of the Exon 1.0 ST core annotation subset.
- For each of the PSR (exonic Probe Selection Region) of the 20 genes we produced three sets of synthetic exon skipping events exchanging:
  - the intensities associated to the 128 pM spike-in with those of the 32 pM (128-32),
  - the intensities associated to the 32 pM spike-in with the 2 pM (32-2),
  - the intensities associated to the 2 pM spike-in with 0 pM (2-0).

Benchmark dataset

- Semi-synthetic exon-skipping benchmark experiment embeds a total of 268 exon skipping events.

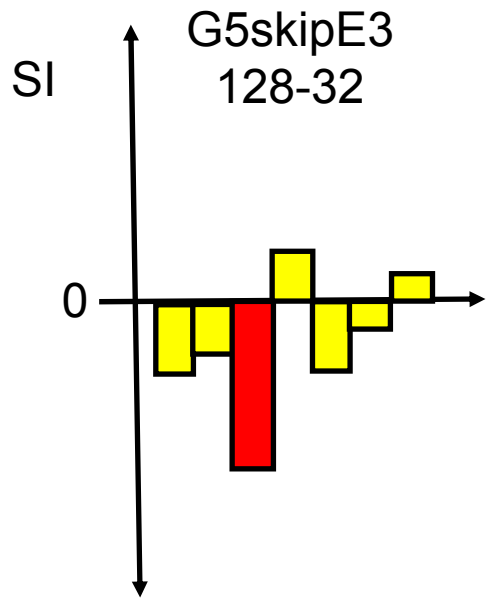


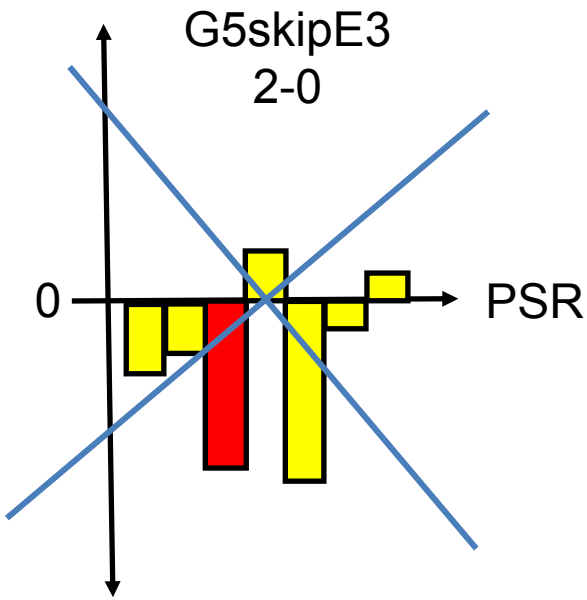
Alternative splicing analysis

- The theoretical differential expression of the spliced exons, expressed as delta splice index (ΔSI), is respectively:
  - 2 (128-32),
  - 4 (32-2),
  - >4 (2-0) $\log_2$(folds)
- associated at the presence of a gene-level differential expression of 2 (512 versus 128-32 and 128 versus 32-2) or > 2 (32 versus 2-0) $\log_2$(folds).

Alternative splicing analysis

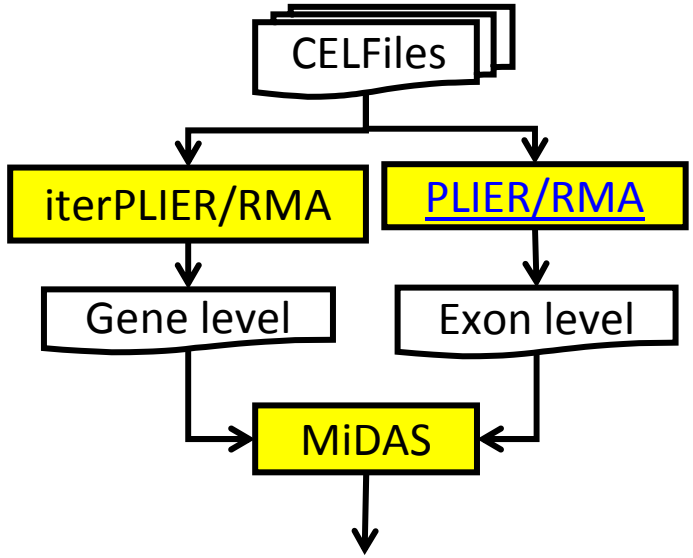
- Furthermore, the skipping events were manually inspected to check that the skipping event represents the most changing event within the synthetic gene.



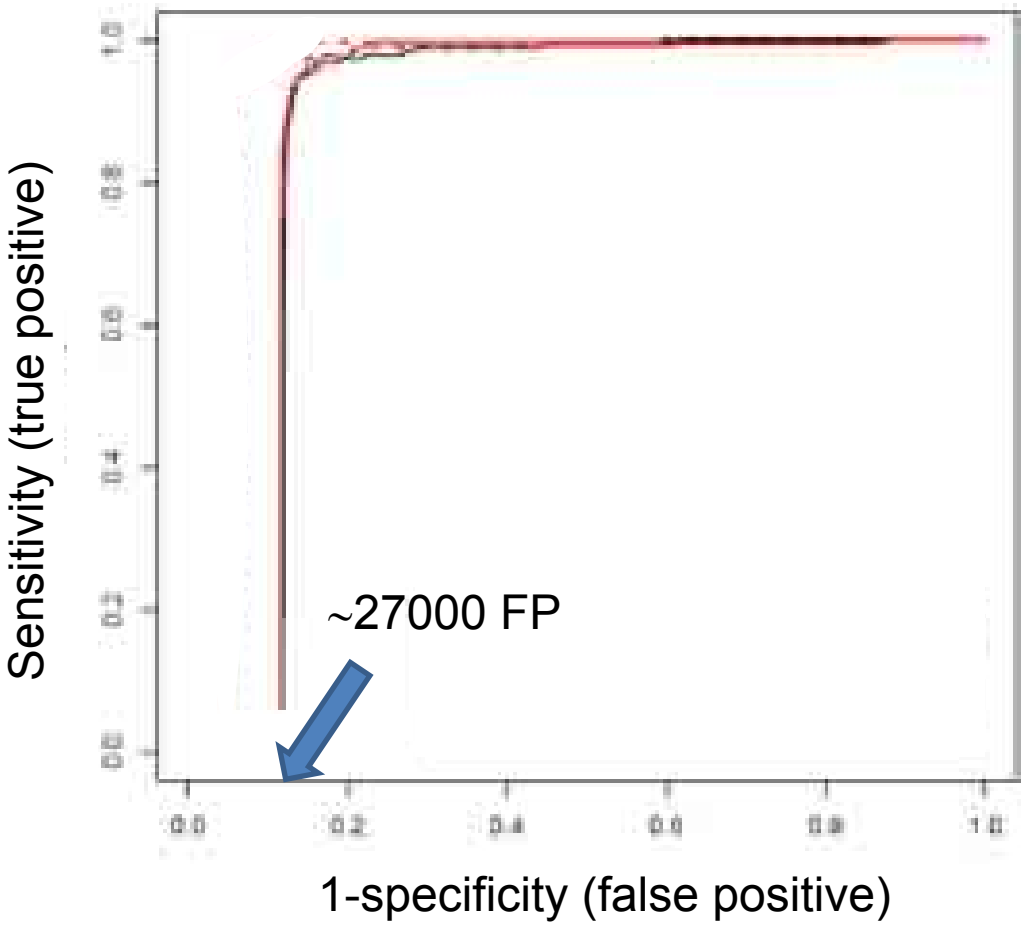


RMA or PLIER?

- To compare the effect of RMA and PLIER on the detection of alternative splicing events we check the ability to detect splicing events on our exon skipping data set by MiDAS.
- Receiver Operating Characteristic (ROC) curve was used to evaluate the effect of intensity summary on alternative splicing detection.
- Inspected Exon-probesets: 228264.



RMA or PLIER?



$$Sensitivity = \frac{TP_{MiDAS}}{TP}$$

$$1 - specificity = \frac{FP_{MiDAS}}{TN}$$

At exon-level, RMA and PLIER produce similar results on an analysis performed on our semi-synthetic data set.

BMC Bioinformatics

Research article

Filtering for increased power for microarray data analysis

Amber J Hackstadt* and Ann M Hess

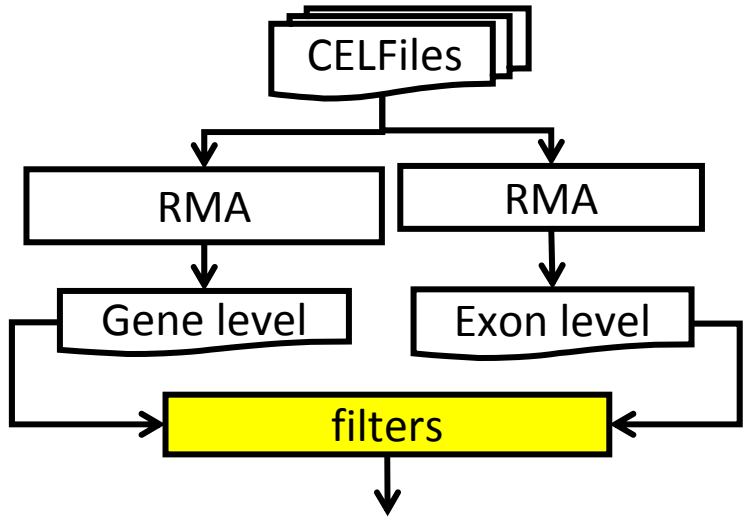
BMC Bioinformatics 2009, 10:11

- The case studies show that both detection call and variance filtering are viable methods of filtering which can increase the number of differentially expressed genes identified.
- The simulation study demonstrates that when paired with a false discovery rate method, filtering by variance can increase power while still controlling the false discovery rate.

Data filtering

- A critical issue is the important number of multiple testing errors that are accumulated if the full set of Exon 1.0 core data is used for the detection of ASEs.

- To moderate this issue, we decided to reduce the complexity of the data set, testing the efficacy of filtering non-informative data at annotation or intensity level:
  - cross hybridization filter
  - DABG filter ($p \leq 0.05$)
  - ENSEMBL filter



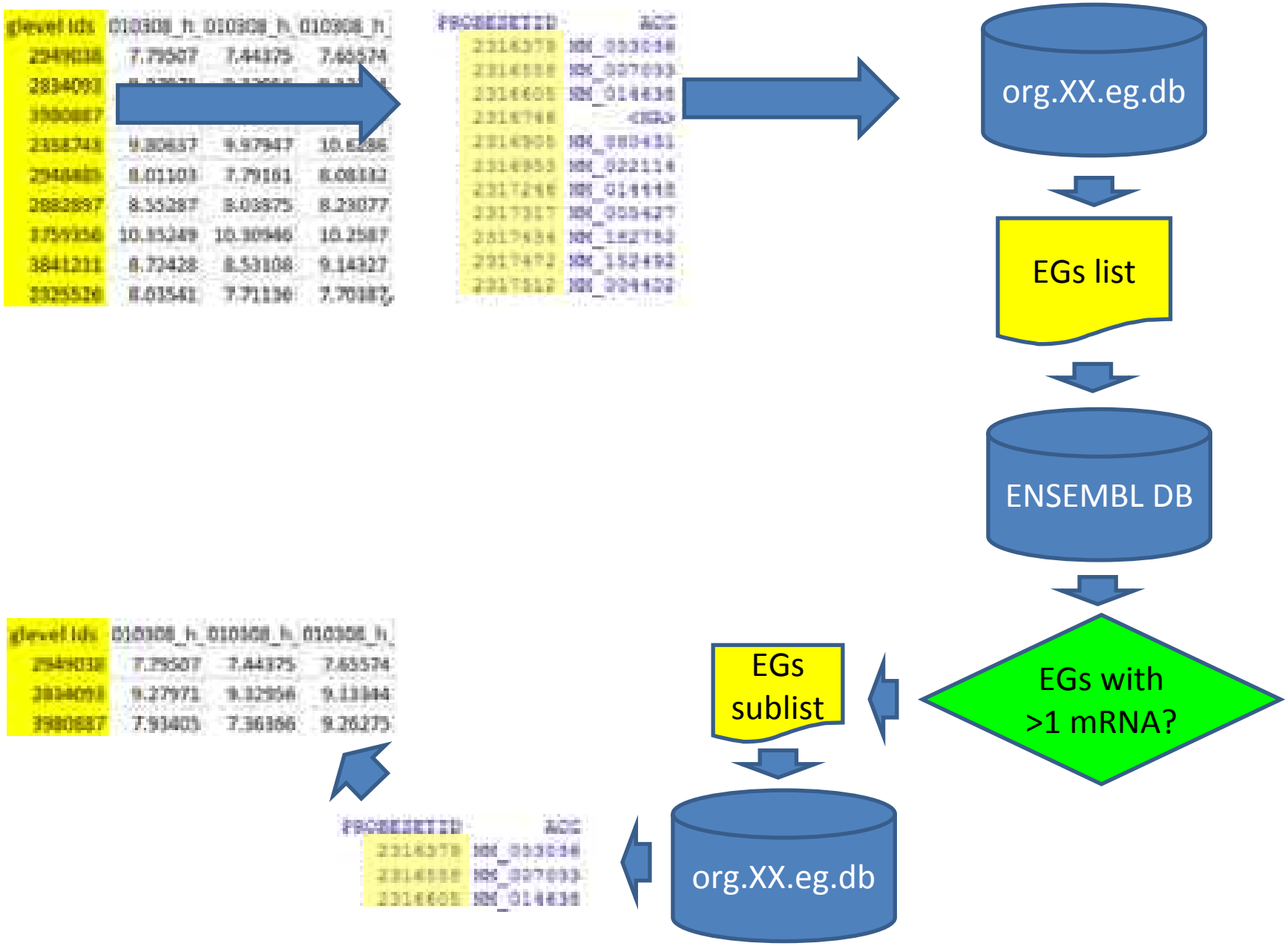
Data filtering

- cross hybridization filter:
  - using the exon-level probe set annotation information provided by Affymetrix, we removed all probe sets where all the probes in the probe set perfectly match more than one sequence in the putatively transcribed array design content as well as those where the probes either perfectly match or partially match more than one sequence in the putatively transcribed array design content.
- DABG filter:
  - DABG pvalue filter, used in this work, is designed to retain only probe sets characterized by a DABG p-value ≤ 0.05 in 90% the arrays.
- ENSEMBL filter:
  - retains only exons of genes which are linked to multiple transcripts in the ENSEMBL database

Data filtering

Table 1

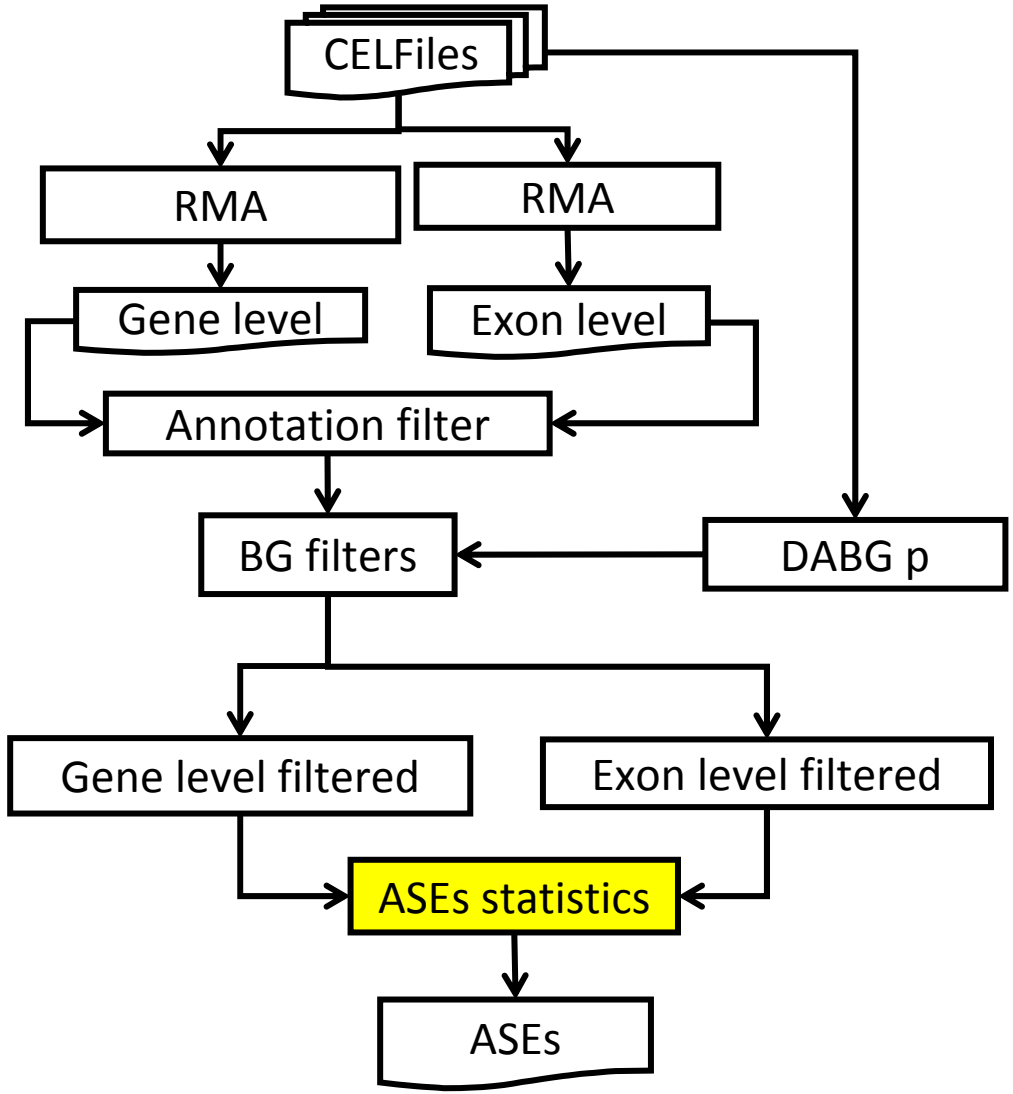
	128.32 vs 512		32.2 vs 128		2.0 vs 32	
	TP (Sensitivity)	TN (1-Specificity)	TP (Sensitivity)	TN (1-Specificity)	TP (Sensitivity)	TN (1-Specificity)
Cross Hybridization filter	172 (1.00)	228264 (1.00)	195 (1.00)	228264 (1.00)	179 (1.00)	228264 (1.00)
Multiple mRNAs filter ENSEMBL	172 (1.00)	71037 (0.31)	195 (1.00)	71037 (0.31)	179 (1.00)	71037 (0.31)
DABG filter (DABG p-value ≤ 0.05 in 90% arrays)	172 (1.00)	197951 (0.86)	185 (0.95)	197951 (0.86)	170 (0.95)	197951 (0.86)

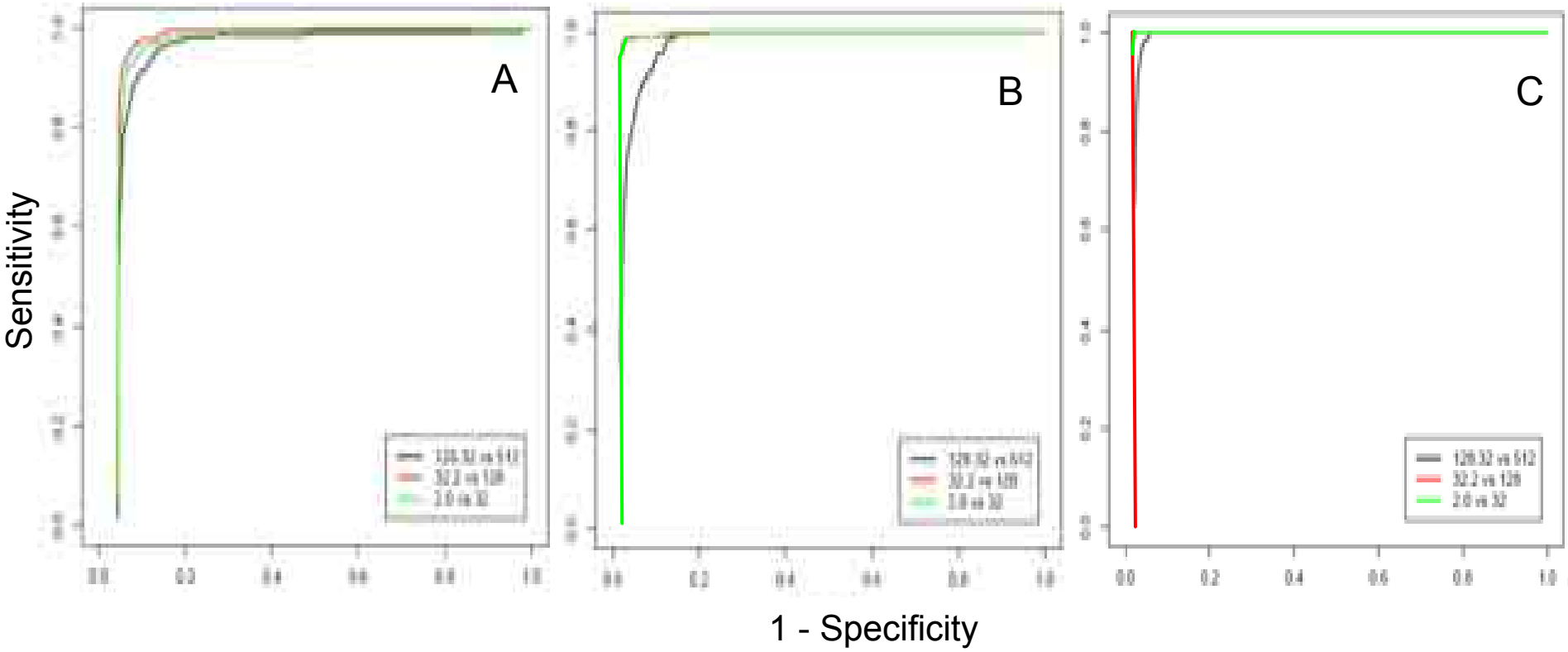


This filter take advantage of the biomaRt library which allows the interrogation of ENSEMBL DB.

ASEs statistical detection

- Detection of Alternative Splicing Events (ASEs) was done using:
  - an ANOVA based algorithm (MiDAS) applied on SI transformed data
  - a permutation based algorithm (RP) applied on SI, RP_{SI}, or directly to exon intensity signals, RP_{I}.





ROC curves were used to detect the efficacy of MiDAS and RP in the detection of ASEs.
A) ROC curves for ASEs detection using MiDAS.
B) ROC curves for ASEs detection using RP_{SI}. RP was calculated using exon signal normalized with respect to gene signal, i.e. SI.
C) ROC curves for ASEs detection RP_I. RP_I was calculated using exon intensity signal without any further normalization.

Table 2

	128.32 vs 512		32.2 vs 128		2.0 vs 32	
	TP (Sensitivity)	FP (1-Specificity)	TP (Sensitivity)	FP (1-Specificity)	TP (Sensitivity)	FP (1-Specificity)
MiDAS ($p \leq 0.05$)	119 (0.68)	2416 (0.03)	176 (0.91)	2319 (0.03)	138 (0.84)	2338 (0.03)
RP$_i$ ($p \leq 0.05$)	172 (1.00)	12941 (0.18)	195 (1.00)	11883 (0.17)	179 (1.00)	9989 (0.14)

Since the two methods are based on completely different assumptions, it is feasible that random events (FPs) contaminating the TPs will not be the same.

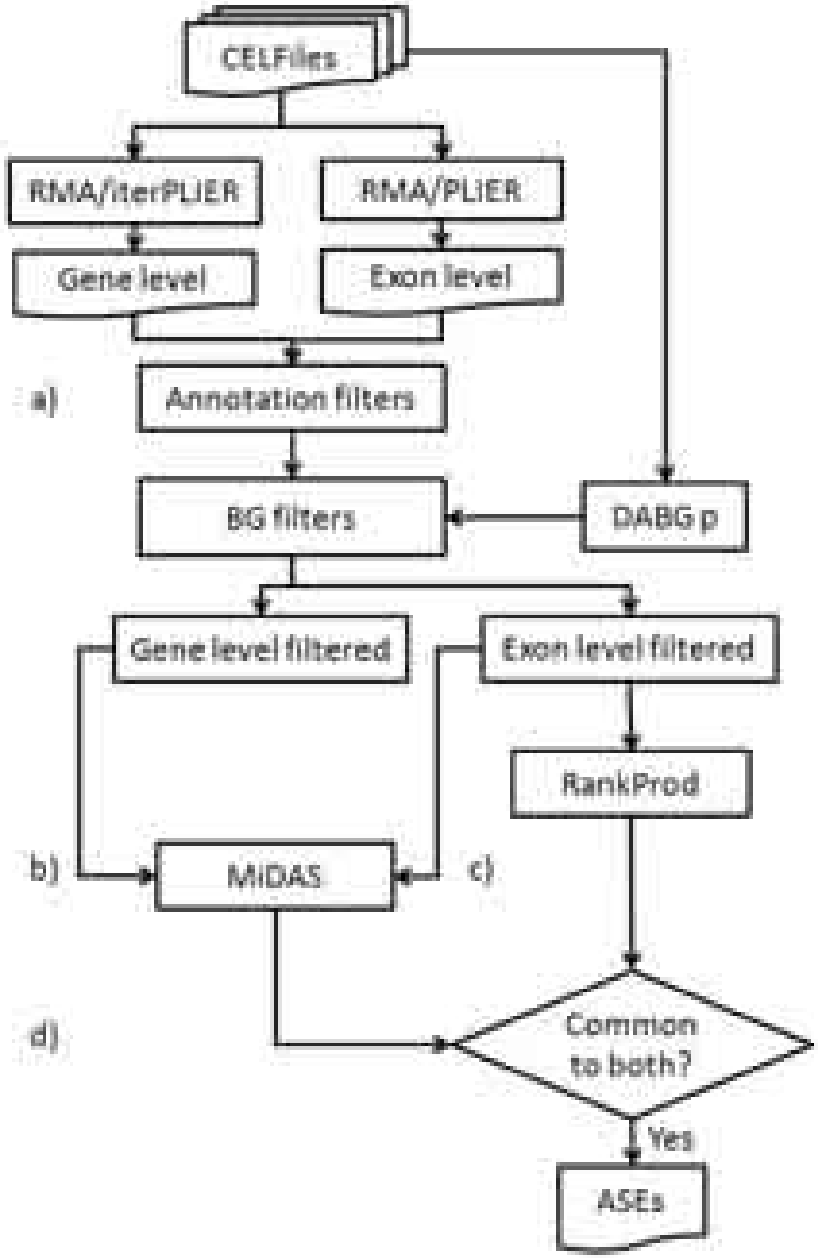
Therefore, the intersection of the results obtained by both statistics, given an arbitrary p-value threshold, might effectively reduces FPs.

Table 2

	128.32 vs 512		32.2 vs 128		2.0 vs 32	
	TP (Sensitivity)	FP (1-Specificity)	TP (Sensitivity)	FP (1-Specificity)	TP (Sensitivity)	FP (1-Specificity)
MiDAS	119 (0.68)	2416 (0.03)	176 (0.91)	2319 (0.03)	138 (0.84)	2338 (0.03)
RP_i	172 (1.00)	12941 (0.18)	195 (1.00)	11883 (0.17)	179 (1.00)	9989 (0.14)
MiDAS & RP intersection	119 (0.68)	436 (0.006)	176 (0.91)	424 (0.006)	138 (0.84)	375 (0.005)

Since at the present time statistics specifically devoted to the detection of ASEs which also address the multiple tests problem are not available, our approach represents an efficient temporary solution for moderating FP.

Further reduction of FP can be realized selecting only those ASEs with a certain level of average signal variation between the two experimental conditions under analysis.



Della Beffa et al. BMC Genomics 2008, 9:571

Algorithms for ASEs detection

BIOINFORMATICS ORIGINAL PAPER

Gene expression

FIRMA: a method for detection of alternative splicing from exon array data

E. Purdom[1,*], K. M. Simpson[2], M. D. Robinson[2,3], J. G. Conboy[4], A. V. Lapuk[4] and T. P. Speed[1,2]

MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays

YI XING,[1,2] PETER STOILOV,[1,3] KAREN KAPUR,[1] ARDUM HAN,[4] HUI JIANG,[5] SHIHAO SHEN,[6]

Research

Open Access

REMAS: a new regression model to identify alternative splicing events from exon array data

Hao Zheng[†1], Xingyi Hang[†2], Ji Zhu[3], Minping Qian[1], Wubin Qu[2], Chenggang Zhang[*2] and Minghua Deng[*1]

Algorithms for ASEs detection

BIOINFORMATICS ORIGINAL PAPER

Gene expression

FIRMA: a method for detection of alternative splicing from exon array data

E. Purdom[1,*], K. M. Simpson[2], M. D. Robinson[2,3], J. G. Conboy[4], A. V. Lapuk[4] and T. P. Speed[1,2]

Implementation in aroma.affymetrix package.

FIRMA uses an additive model which includes the possibility of alternative splicing or different levels of expression per exon

$$\log_2\left(PM_{ijk(j)}\right) = c_i + e_j + d_{ij} + p_{k(j)} + \varepsilon_{ijk(j)}$$

c_i is the chip effect (expression level) for chip i,
e_j is the relative change in exon expression for exon j,
d_{ij} is the interaction between chip and exon giving the relative change for sample i in exon j,
$p_{k(j)}$ is the nested relative probe effect for the k-th probe in exon j.

d_{ij} indicates the discrepancy of a given sample in exon j from the expected expression for that exon.
Large values of d_{ij} indicates differential alternative splicing.

Algorithms for ASEs detection

MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays

Implementation in R available

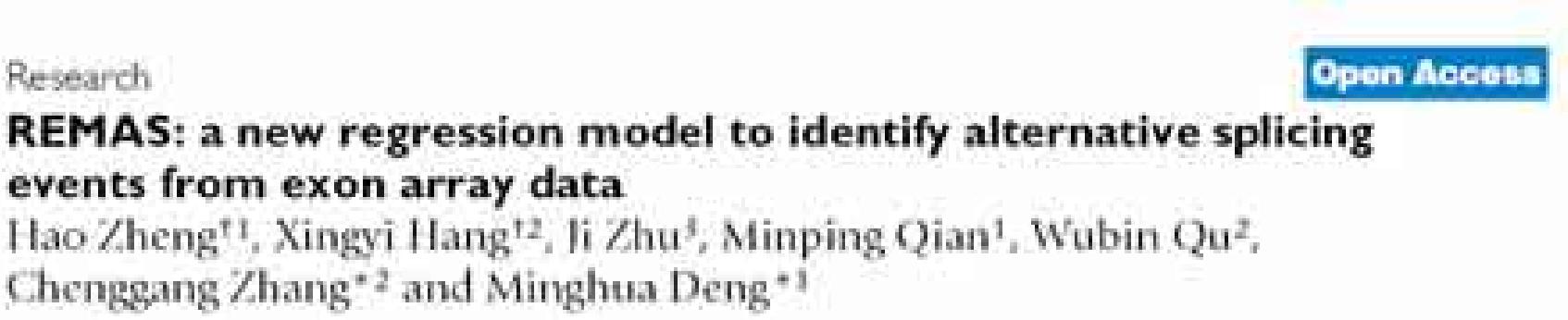
YI XING,[1,2] PETER STOILOV,[1,4] KAREN KAPUR,[1] ARJUN HAN,[4] HUI JIANG,[2] SHIHAO SHEN,[4] DOUGLAS L. BLACK,[1,3] and WING HUNG WONG[2]

Table 3

	MADs golden standard	
	% TP	% FP
MiDAS	21	0
RP,	21	0
MADs t-test	37.5	0

- MADS uses a series of low-level analysis algorithms to construct an efficient statistic for differential splicing:
  - Background correction
  - Iterative probe selection for expression index calculation.
  - Detection/removal of sequence-specific cross-hybridization to off-target transcripts.
- The correction of the major source of noise allows a more efficient detection of differential splicing.
- Splicing events are detected at probe level by t-test.

Algorithms for ASEs detection


Research

Open Access

REMAS: a new regression model to identify alternative splicing events from exon array data

Hao Zheng[†1], Xingyi Hang[†2], Ji Zhu[3], Minping Qian[1], Wubin Qu[2], Chenggang Zhang*[2] and Minghua Deng*[1]

BMC Bioinformatics. 2009 Jan 30;10 Suppl 1:S18.

- REMAS is a regression method for AS detection:
  - Features of alternatively spliced exons are scaled by reasonably defined variables.
  - A hierarchical model, which can represent gene structure and transcriptional influence to exons, and the lasso type penalties is introduced in calculation because of huge variable size.
  - An iterative two-step algorithm was developed to select alternatively spliced genes and exons.

No implementation is available.

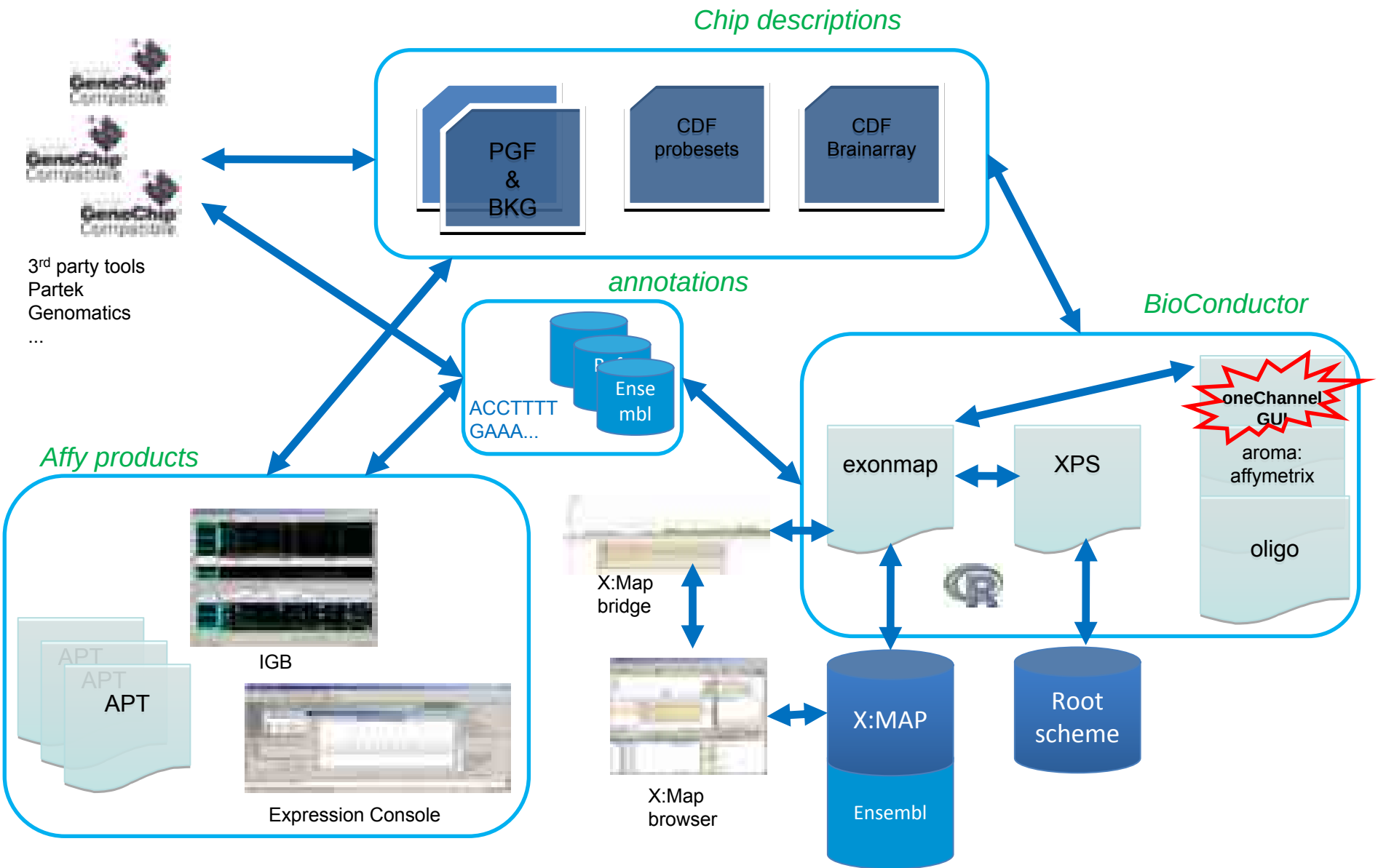
Conclusions

- Our analysis pipe-line represents a temporary solution for ASE detection for final users.
- Available ASEs detection tools mainly focus on optimization of signal data.
- Very little is available as optimized statistics for ASEs detection.
- Experimental reference benchmark is needed for efficient methods comparison.

Agenda

- First part
  - Dissecting alternative splicing workflow
- Second part
  - Softwares for exon-array analysis

Software infrastructure overview

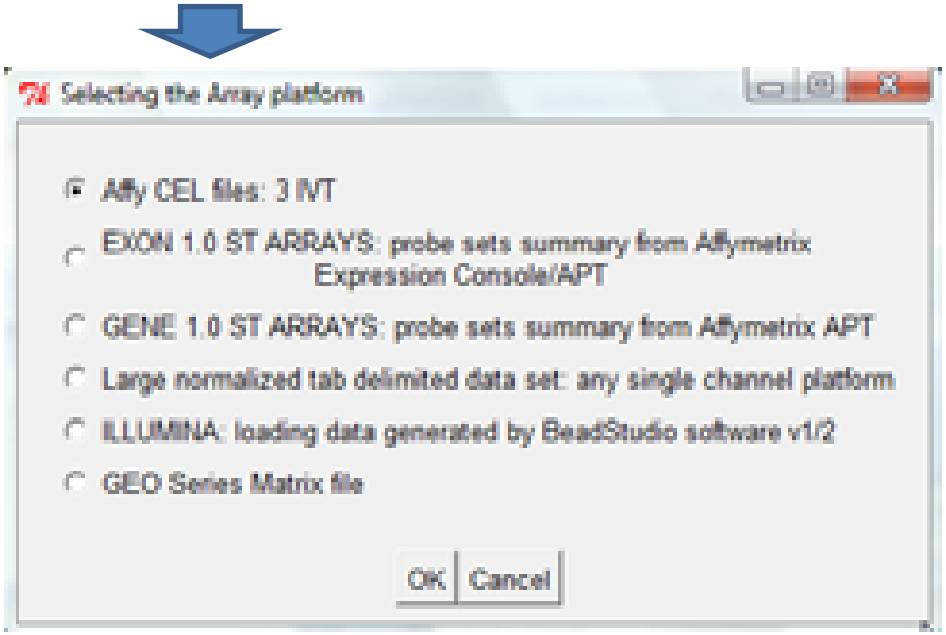


BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3406–3408
doi:10.1093/bioinformatics/btm469

Gene expression

oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language

Remo Sanges[1,†], Francesca Cordero[2,†] and Raffaele A. Calogero[3,*]



BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3406–3408
doi:10.1093/bioinformatics/btm469

Gene expression

oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language

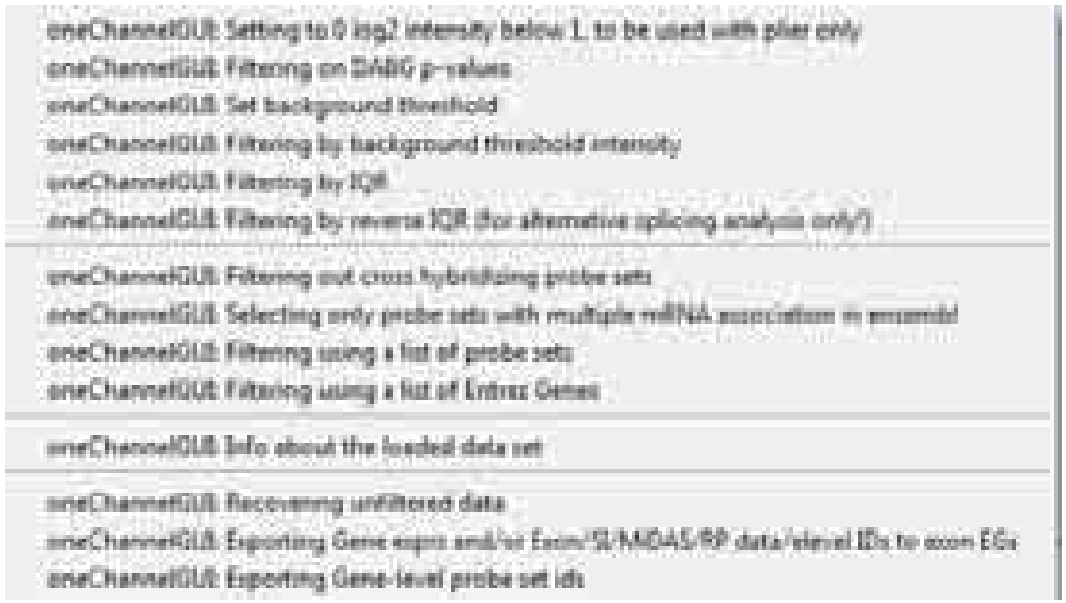
Remo Sanges[1,†], Francesca Cordero[2,†] and Raffaele A. Calogero[3,*]

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3406–3408
doi:10.1093/bioinformatics/btm469

Gene expression

oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language

Remo Sanges[1,†], Francesca Cordero[2,†] and Raffaele A. Calogero[3,*]



BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3406–3408
doi:10.1093/bioinformatics/btm469

Gene expression

oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language

Remo Sanges[1,†], Francesca Cordero[2,†] and Raffaele A. Calogero[3,*]

Gene-level modelling statistics

affylmGUI: Compute Linear Model Fit
affylmGUI: Compute Contrasts
oneChannelGUI: Raw p-value distribution plot
affylmGUI: View Existing Contrasts Parameterization
affylmGUI: Delete Contrasts Parameterization
oneChannelGUI: Table of Genes Ranked in order of Differential Expression

affylmGUI: Quantile-Quantile / Statistic Plot (for one contrast)
oneChannelGUI: Venn Diagram between probe set lists

oneChannelGUI: Create an edesign for maSigPro
oneChannelGUI: Execute maSigPro
oneChannelGUI: View maSigPro results

Gene-level permutation statistics

oneChannelGUI: SAM analysis
oneChannelGUI: Rank product analysis

Exon-level statistics

oneChannelGUI: Calculating MDAS p-value (APT)
oneChannelGUI: Calculating splice index
oneChannelGUI: Rank Product alternative splicing detection

oneChannelGUI: Selecting alternative splicing events by MDAS p-values
oneChannelGUI: Selecting alternative splicing events by RankProd p-values
oneChannelGUI: Filtering gene/exon data by absolute SI mean or min difference
oneChannelGUI: Selecting alternative splicing by RP/MDAS p-values and mean/min delta SI

oneChannelGUI: Inspecting splice indexes
oneChannelGUI: Inspecting splice indexes of one gievel probe set
oneChannelGUI: Mapping exon level probe sets to Reference Sequences
oneChannelGUI: Mapping exon level probe sets to the corresponding exon

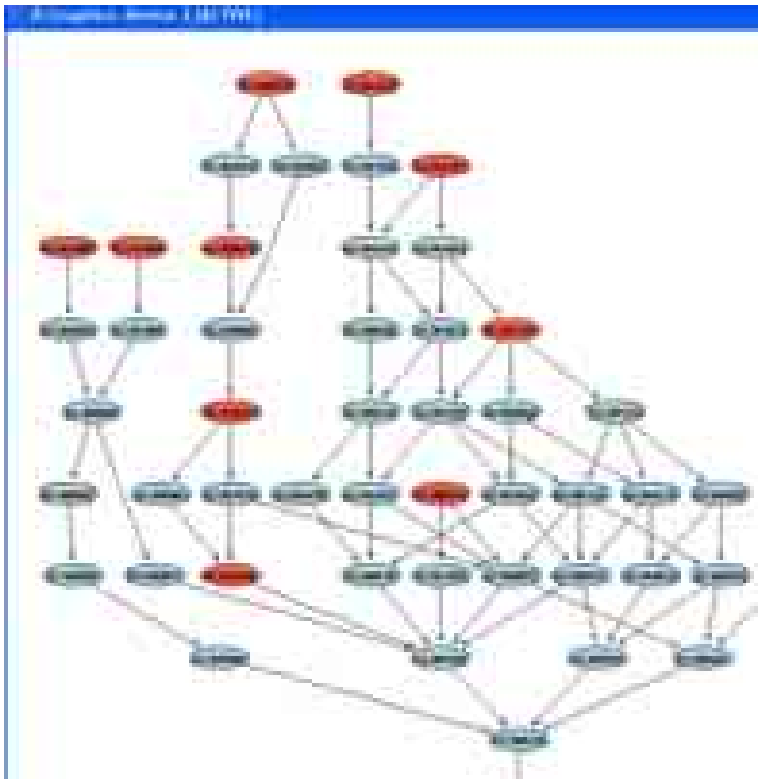
oneChannelGUI: Exporting Gene expre and/or Exon/SI/MDAS/RP data/ alevel IDs to exon FGs
oneChannelGUI: Recovering unfiltered data

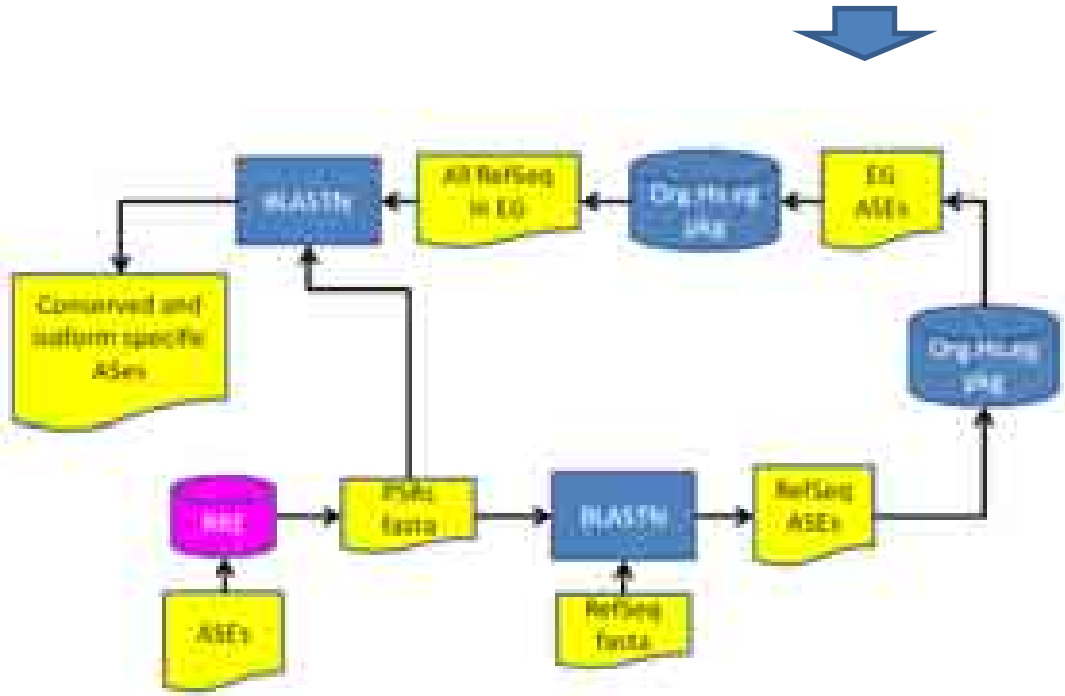
BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 24 2007, pages 3406–3408
doi:10.1093/bioinformatics/btm469

Gene expression

oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language

Remo Sanges[1,†], Francesca Cordero[2,†] and Raffaele A. Calogero[3,*]






University of Torino
Raffaele A. Calogero
Francesca Cordero
Cristina Della Beffa