Journal Club

Analysis of Alternative Splicing by Affymetrix Exon Arrays

Petr Nazarov

21-11-2008

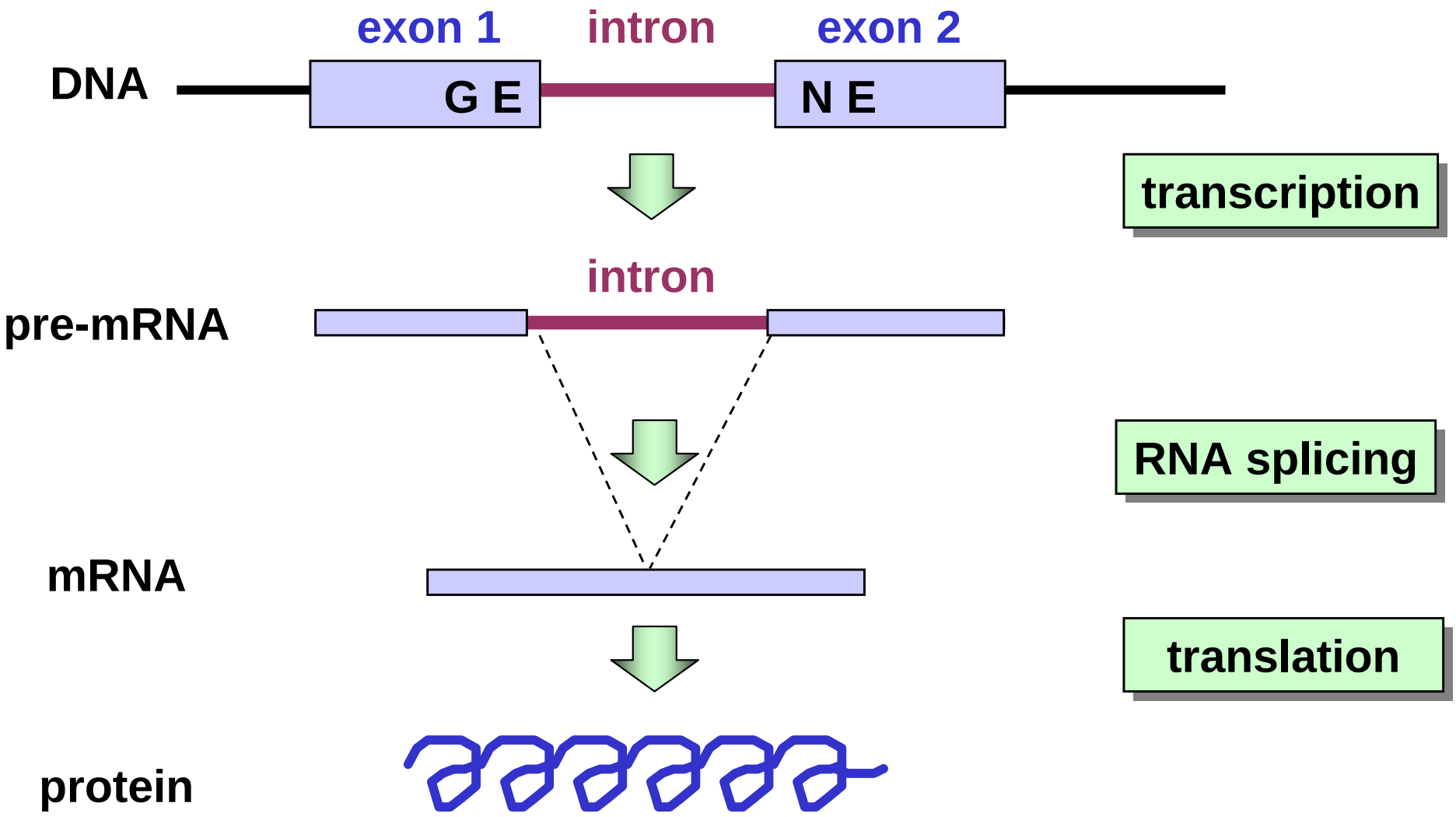
OUTLINE

- **Introduction**
- **Affymetrix Exon Arrays**
- **General scheme for the analysis**
- **Application 1.** Splice variants and novel exons in glial brain tumors
- **Application 2.** Effects of SMN-deficiency on transcription
- **Application 3.** AS in stem cells and neural progenitors
- **Application 4.** AS in lung cancer
- **Future plans**

Importance of alternative splicing:

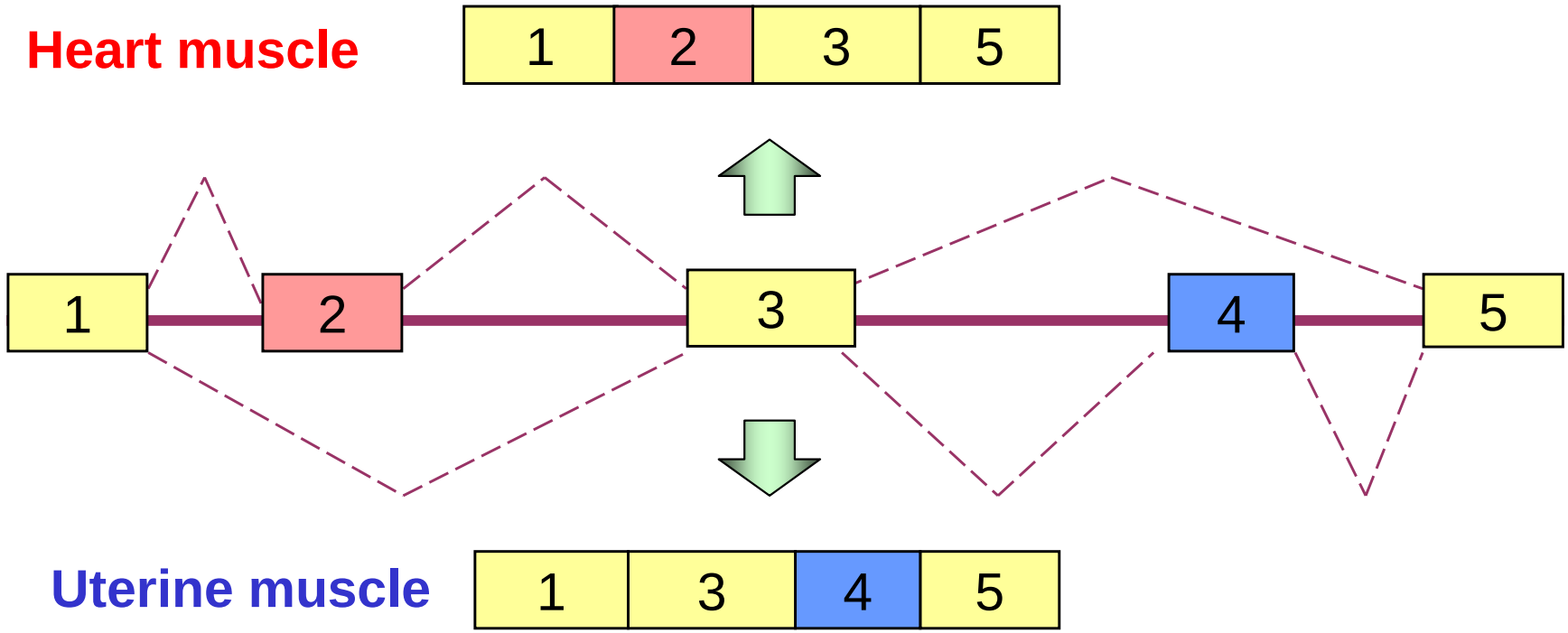
- Approximately 75% of all human multi-exon genes are alternatively spliced
- Defects of the machinery of alternative splicing have been implicated in many diseases, including:
  - neuropathological conditions such as Alzheimer disease
  - cystic fibrosis, those involving growth and developmental defects
  - many human cancers
- Therefore, the detailed understanding of the cell and gene-related diseases must include knowledge of the roles played by alternative splicing and its products

Basic Expression Scheme



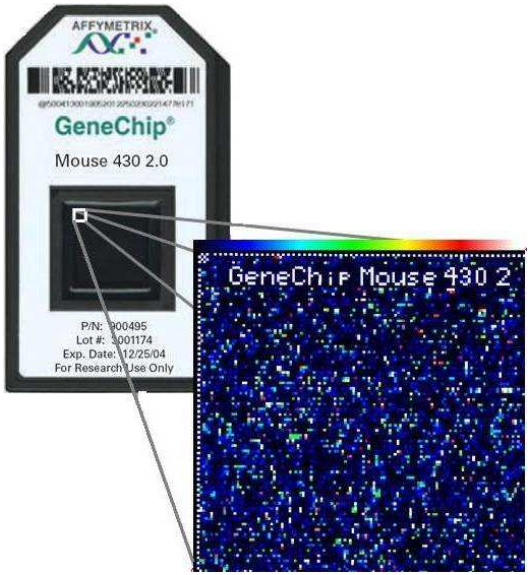
Alternative Splicing

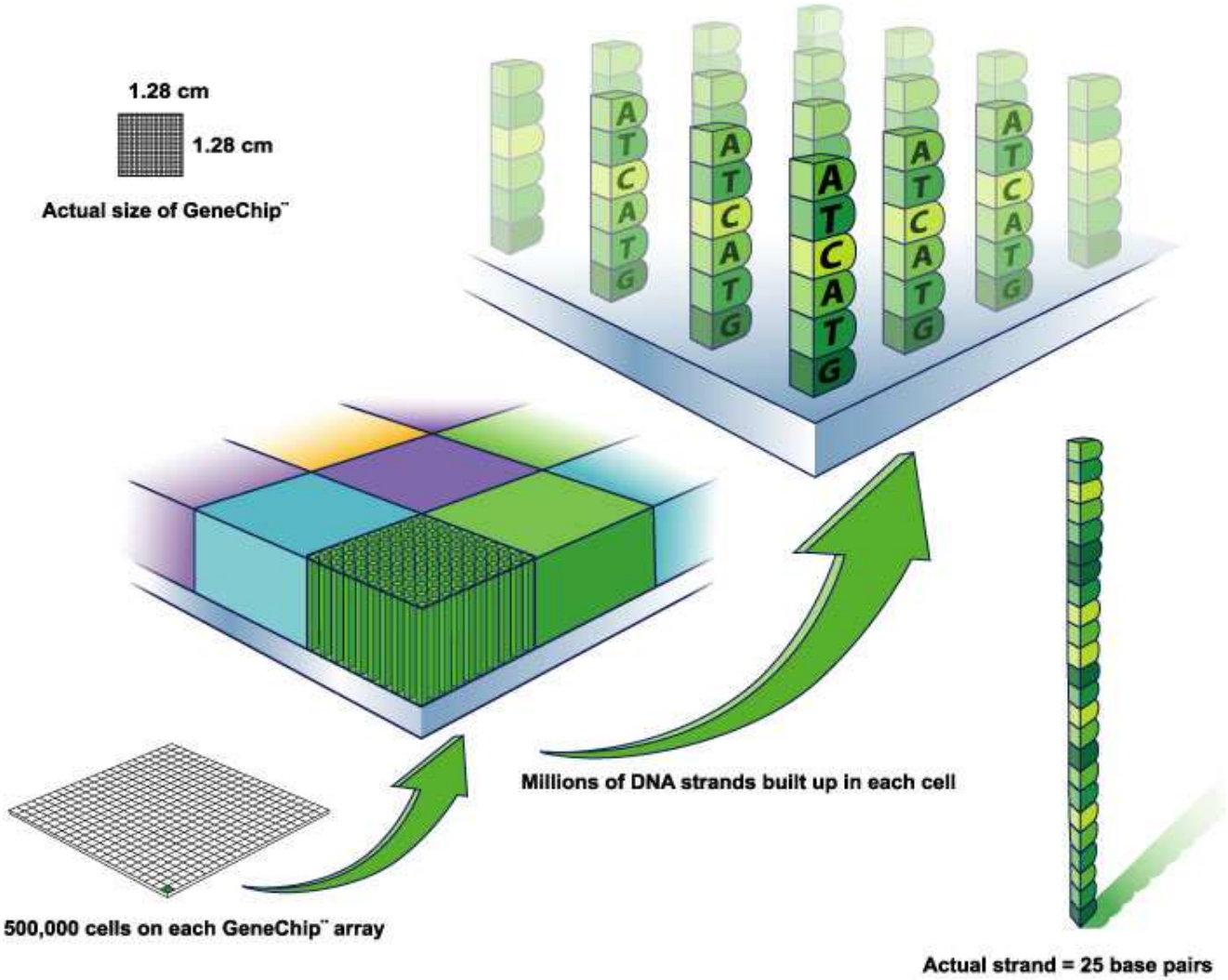
Multiple introns may be spliced differently in different circumstances, for example in different tissues.



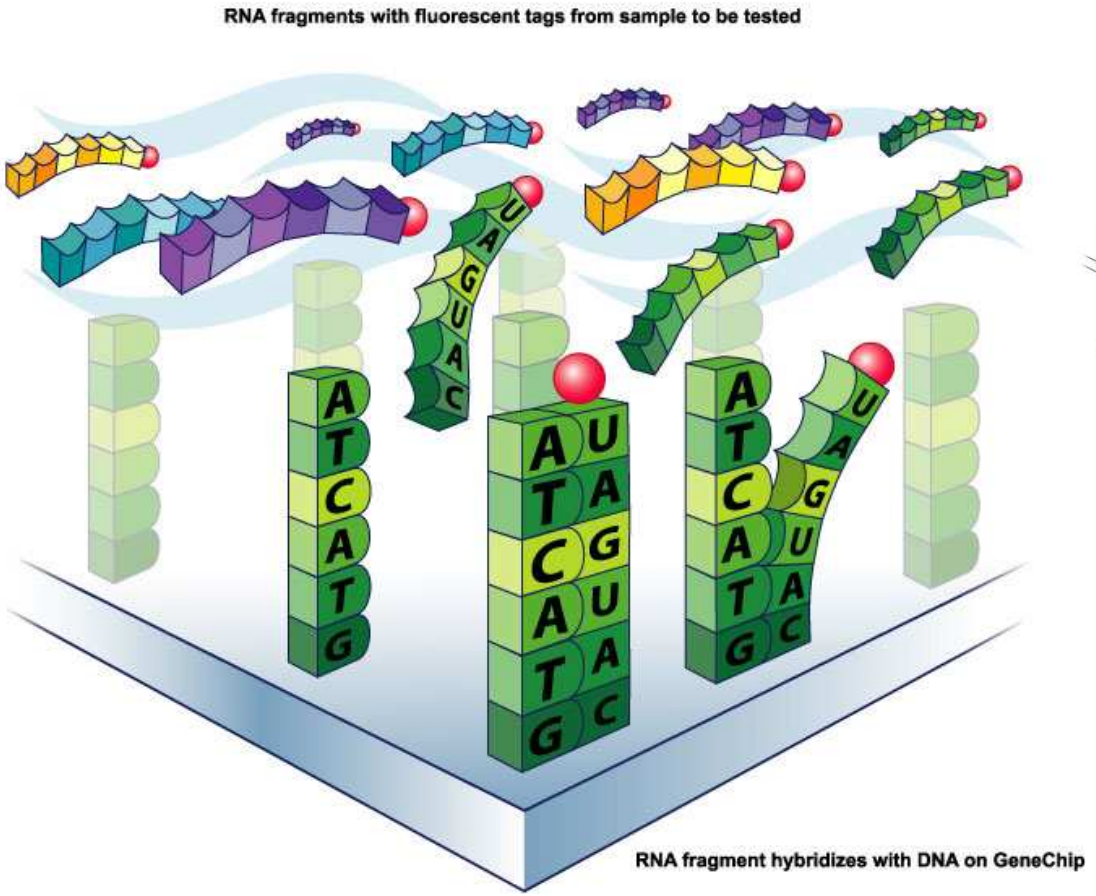
Thus one gene can encode more than one protein. The proteins are similar but not identical and may have distinct properties – an important feature for complex organisms

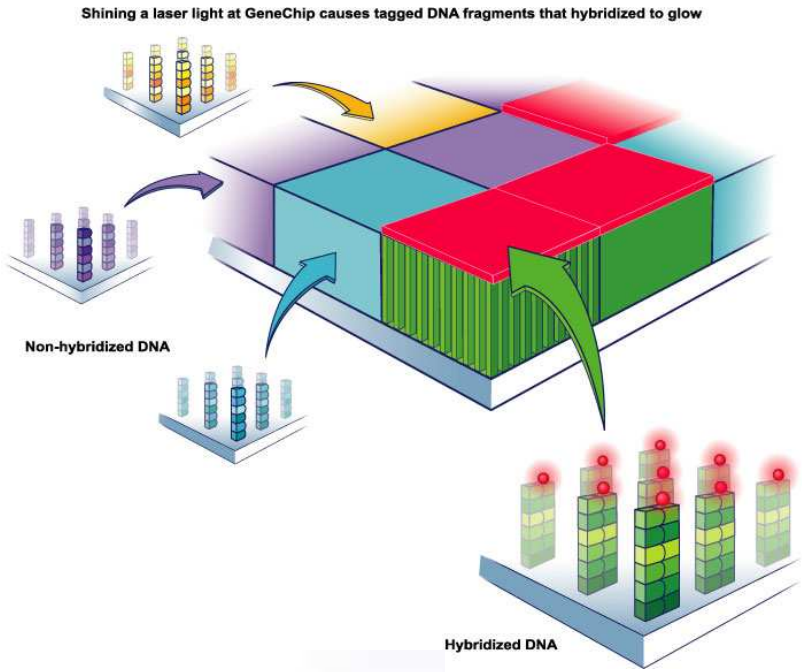
Affymetrix Microarray Design

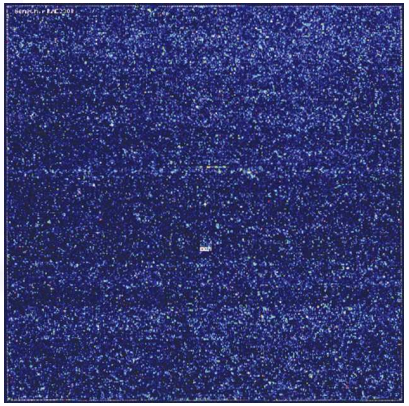




Affymetrix Microarray Design







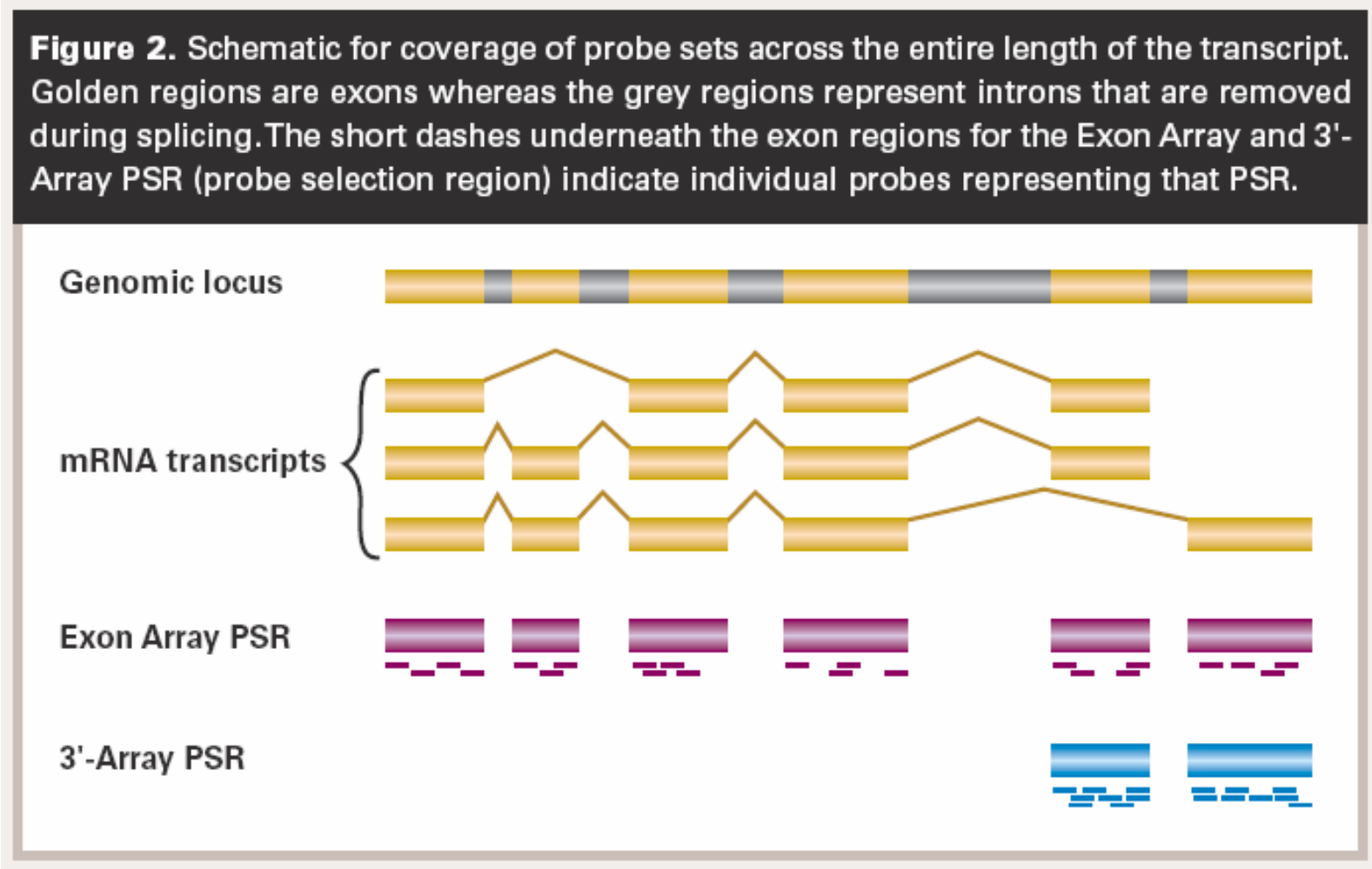
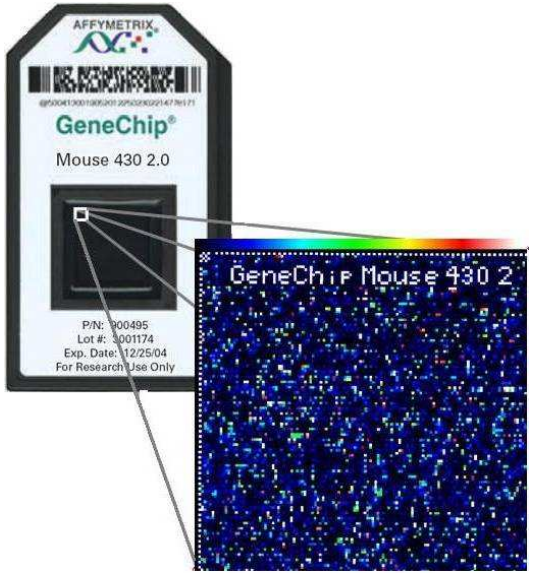


Figure 1. Exon 1.0 ST Arrays probe design and comparison with 3'-Expression Array. Affymetrix, Exon Array Design Datasheet, http://www.affymetrix.com, 2005

Affymetrix Exon Array Design and Main Features

- *Probes*: 25-mer oligonucleotide
- Number of probes: 5.6 M united into 1.4 M *probe-sets*
- Several probe-sets may target the same exon
- Three level of annotation and confidence
  - core (284 000) – highly annotated and validated
  - extended (523 000) – low level of annotation
  - full (580 000) – predicted
- Types of Affymetrix Exon Arrays microarrays
  - Human Exon 1.0 ST Array
  - Mouse Exon 1.0 ST Array
  - Rat Exon 1.0 ST Array



Generalized Data Analysis and AS Detection Flowchart

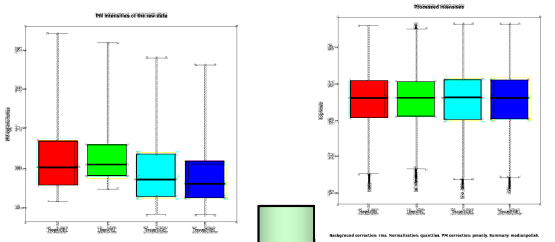
Raw probes intensities (CEL files)

Summarization of probes into probe-sets

probe A1
probe A2
probe A3
probe A4
→ probe-set A

Normalization

- Quantile normalization
- Spatial normalization
- Linear normalization



Calculation of exon expression

Calculation of gene expression

Gene filtering

AS Detection

- Splicing index (SI)
- Pattern-based correlation (PAC)
- Analysis of splice variation (ANOSVA)
- Regression-based exon array protocol (REAP)
- Finding isoforms using robust multichip analysis (FIRMA)

List of AS exons

.....
.....
.....
.....
.....

- Affymetrix Inc, 2005;
- Cline et al., 2005, Bioinformatics;
- French et al., 2007, Cancer research;
- Irizarry et al., 2003, Biostatistics;
- Purdom et al., 2008, Bioinformatics
- Srinivasan et al., 2005, Methods;
- Yeo et al., 2007, PLoS Comput Biology;
- Zhang et al., 2008, Cell

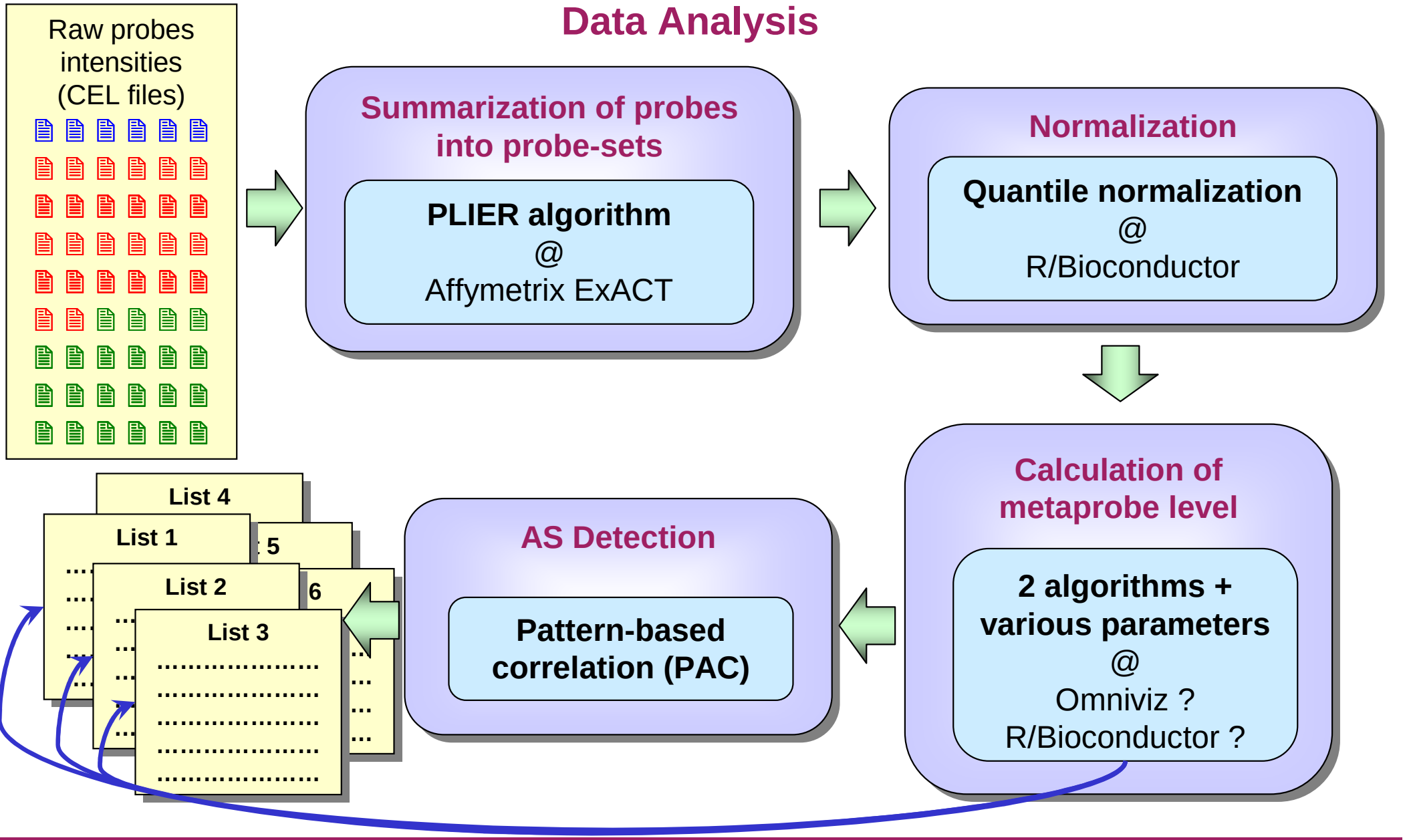
Identification of Differentially Regulated Splice Variants and Novel Exons in Glial Brain Tumors Using Exon Expression Arrays

French, P.J, et al. (2007) Cancer Res.

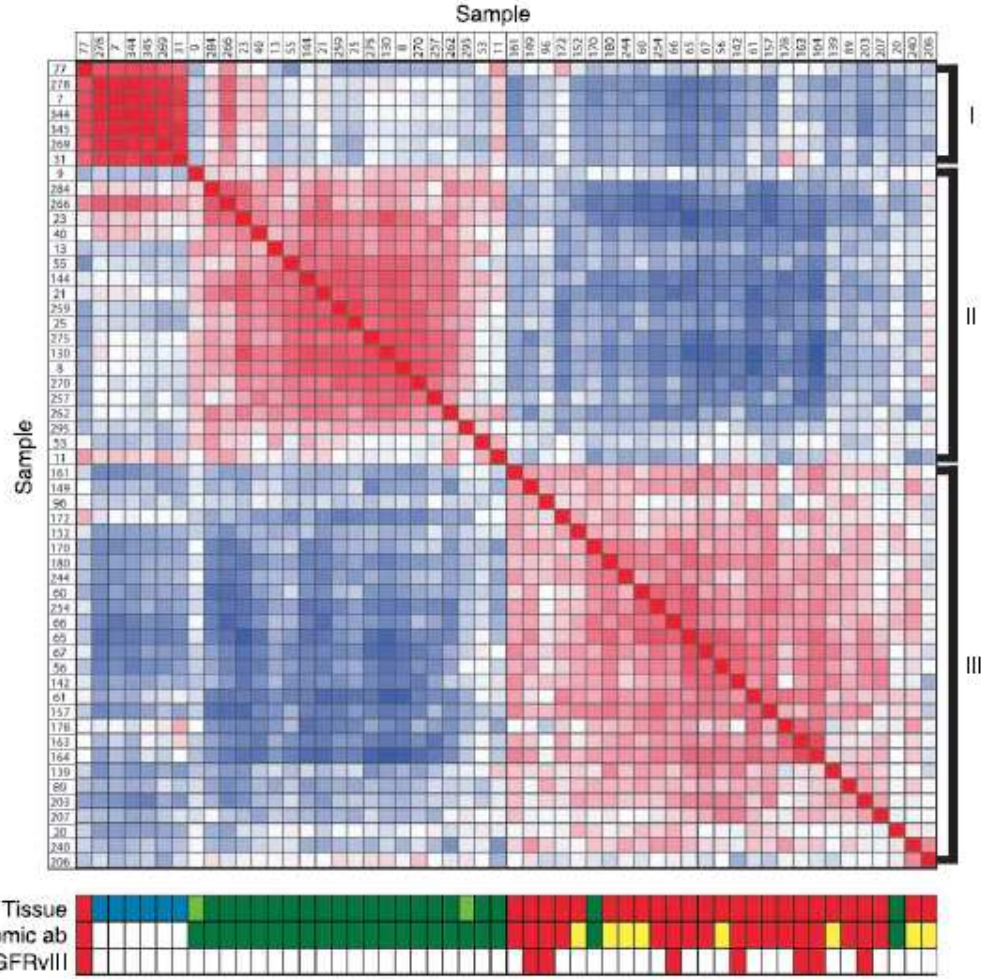
- One of the first study using Affymetrix Human Exon Arrays (sponsored by Affy)
- **Samples:** 26 glioblastomas, 22 oligodendrogliomas, and 6 control brain
- **Microarrays:** Affymetrix GeneChip Human Exon 1.0 ST
- **Goals**: (1) test new microarrays, (2) detect known and new splice variants

Aberrant splice variants are involved in the initiation and/or progression of glial brain tumors. We therefore set out to identify splice variants that are differentially expressed between histologic subgroups of gliomas. Splice variants were identified using a novel platform that profiles the expression of virtually all known and predicted exons present in the human genome. Exon-level expression profiling was done on 26 glioblastomas, 22 oligodendrogliomas, and 6 control brain samples. Our results show that Human Exon arrays can identify subgroups of gliomas based on their histologic appearance and genetic aberrations. We next used our expression data to identify differentially expressed splice variants. In two independent approaches, we identified 49 and up to 459 exons that are differentially spliced between glioblastomas and oligodendrogliomas, a subset of which (47% and 33%) were confirmed by reverse transcription-PCR (RT-PCR). In addition, exon level expression profiling also identified >700 novel exons. Expression of approximately 67% of these candidate novel exons was confirmed by RT-PCR. Our results indicate that exon level expression profiling can be used to molecularly classify brain tumor subgroups, can identify differentially regulated splice variants, and can identify novel exons. The splice variants identified by exon level expression profiling may help to detect the genetic changes that cause or maintain gliomas and may serve as novel treatment targets.

Data Analysis



Data Clustering



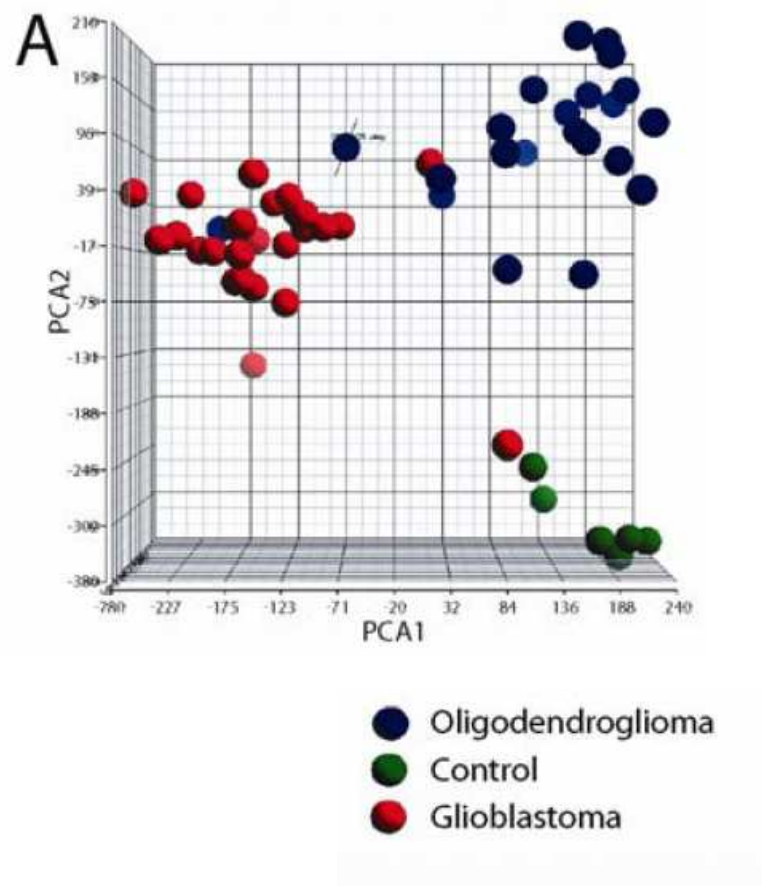


Figure 1. Correlation plot of all samples. Samples are plotted against each other as Pearson's correlation to determine the degree of similarity based on expressed exons. All exons with 4-fold expression difference from the geometric mean are included in the clustering. *Red,* high correlation; *blue,* low correlation. Below the correlation plot is a graphic representation of histologic and patient data. *Tissue*. Origin of sample: ■ control cortex; ■ anaplastic oligodendroglioma (WHO grade III); ■ oligodendroglioma (WHO grade II); and ■ glioblastoma. *Genomic aberrations*. Genomic aberrations of the sample: □ control sample; ■ LOH on 1p and 19q, no amplification of EGFR; ■ no LOH on 1p and 19q but amplification of EGFR; ■ no LOH on 1p and 19q, no amplification of EGFR. *EGFRvIII*: expression of EGFRvIII as determined by RT-PCR: □ no expression; ■ expression. Subgroups identified by Pearsons's correlation plot (*right*; I–III).

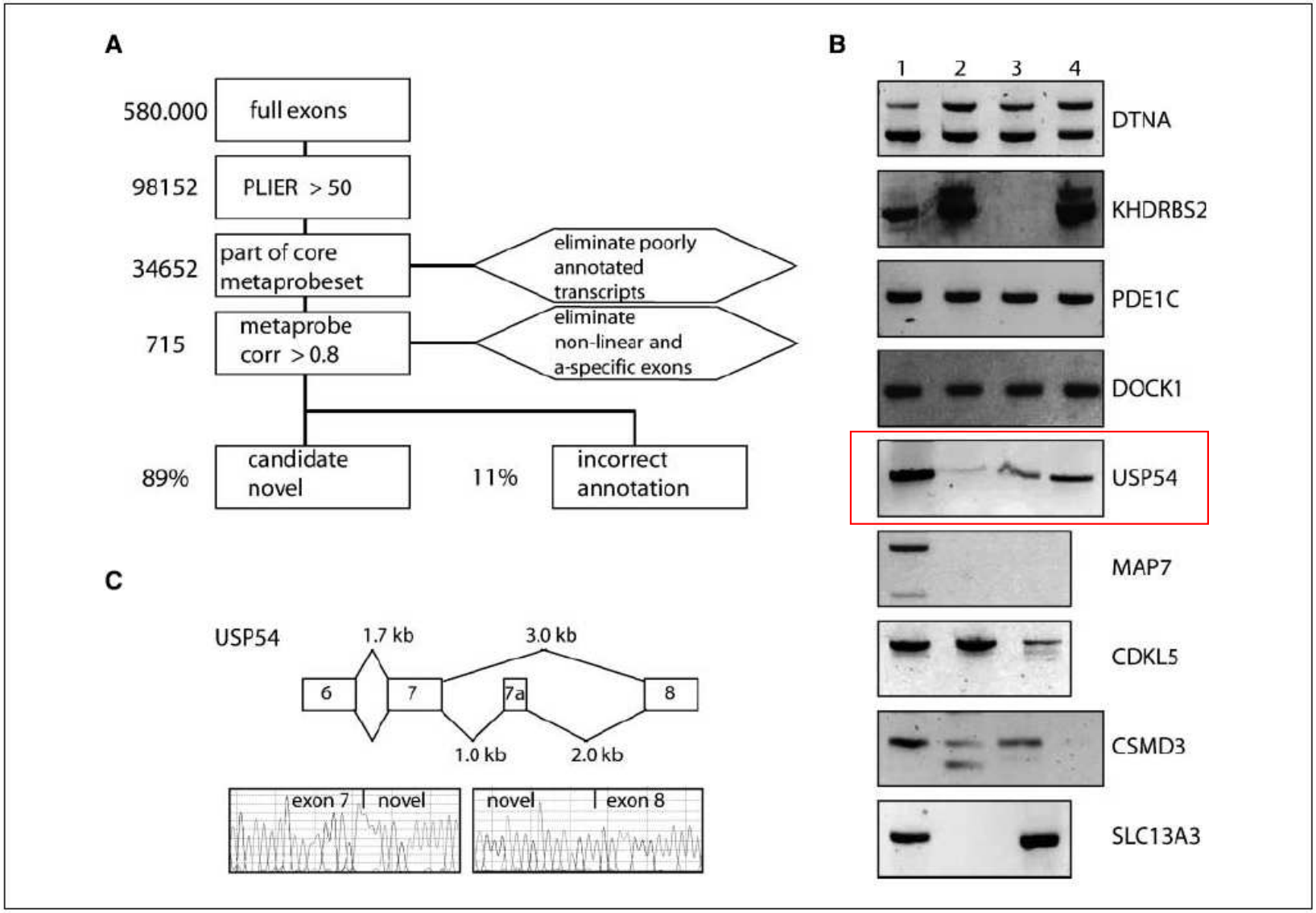


Figure 3. Identification of novel exons by exon level expression profiling. *A*, filtering steps used to identify 715 candidate novel exons. Candidate novel exons are expressed (PLIER) >50 as part of a well-characterized transcript and have a correlation coefficient of >0.8 with its transcript. *B*, RT-PCR of a subset of identified candidates on independent samples (lanes 1–4). *DTNA*, *KHDRBS2*, and *PDE1C* were identified as part of a rare splice variant in public domain databases. Expression of *DTNA* and *KHDRBS2 full* exons was confirmed using exon spanning primers, other *full* exons were confirmed using one primer within the candidate novel exon. Products were sequence verified to exclude a-specific amplifications. *C*, model of splicing of the novel identified exon in *USP54*. Direct sequencing confirmed the presence of the novel exon expressed as part of *USP54*.

Summary

- Identification of differentially expressed splice variants requires rigorous filtering steps to exclude nonlinear and a-specific probe sets. **The proper analysis pipeline is crucial for obtaining robust and reproducible results.**

- The study has identified 715 full exons that are expressed as part of a well-annotated transcript. The confirmation level by RT-PCR ~ 67%.

- The splice variants identified by exon level expression profiling may lead to the identification of causative genetic changes in glial brain tumors. Furthermore, glioma-subgroup specific splice variants may serve as novel treatment targets..

SMN Deficiency Causes Tissue-Specific Perturbations in the Repertoire of snRNAs and Widespread Defects in Splicing

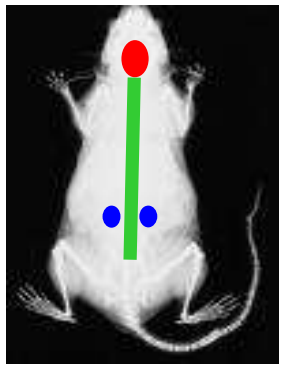
Zhang, Z., et al. (2008) Cell

- Recent high-impact publication. Usage of Partek
- **Samples:** 18 samples, including:
  - Normal tissues: spinal cord (3x), brain (3x), kidney (3x)
  - SMA tissues: spinal cord (3x), brain (3x), kidney (3x)
- **Microarrays:** Affymetrix GeneChip Mouse Exon 1.0 ST
- **Goal**: find the effects of SMN-deficiency on transcription

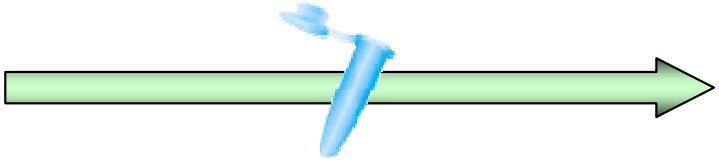
The **survival of motor neurons (SMN)** protein is essential for the biogenesis of small nuclear RNA (snRNA)-ribonucleoproteins (snRNPs), the major components of the pre-mRNA splicing machinery. Though it is ubiquitously expressed, SMN deficiency causes the motor neuron degenerative disease **spinal muscular atrophy (SMA)**. We show here that SMN deficiency, similar to that which occurs in severe SMA, has unexpected cell type-specific effects on the repertoire of snRNAs and mRNAs. It alters the stoichiometry of snRNAs and causes widespread pre-mRNA splicing defects in numerous transcripts of diverse genes, preferentially those containing a large number of introns, in SMN-deficient mouse tissues. These findings reveal a key role for the SMN complex in RNA metabolism and in splicing regulation and indicate that SMA is a general splicing disease that is not restricted to motor neurons.

Experiments

Normal mice



- Spinal cord (3 samples)
- Brain (3 samples)
- Kidney (3 samples)

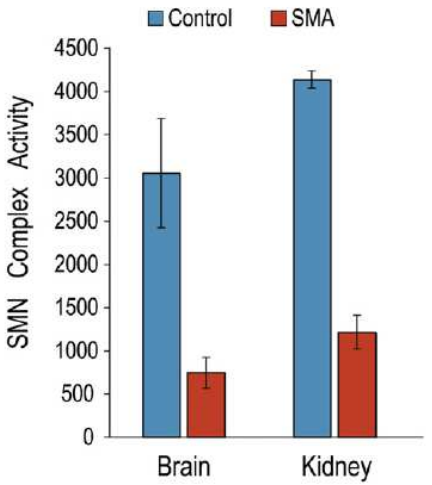


Regulated SMN inhibitor constructed in a lentiviral vector

SMA mice

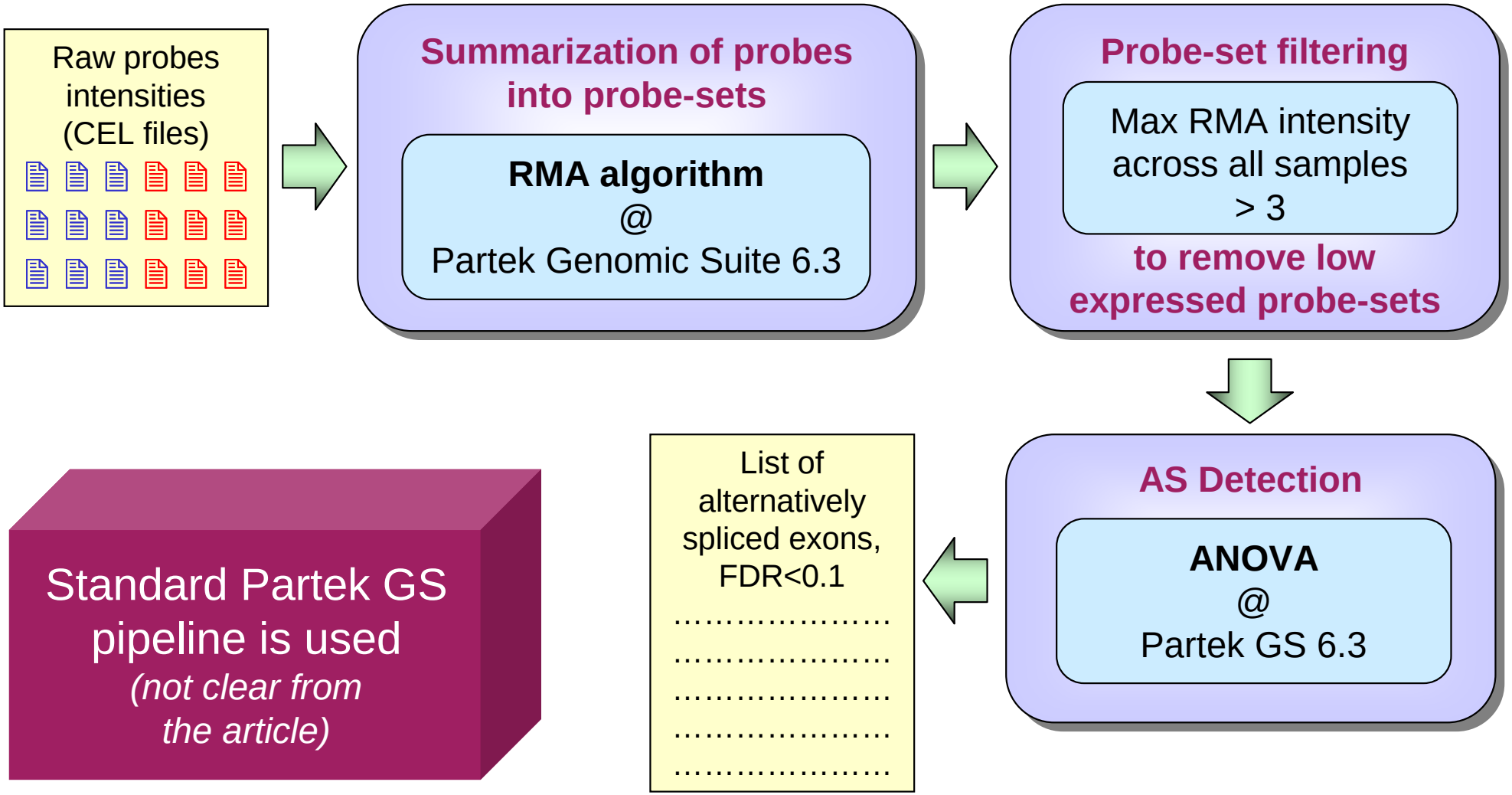
- Spinal cord (3 samples)
- Brain (3 samples)
- Kidney (3 samples)

18 Microarrays



+ Additional check is performed by RT-PCR .

Data Analysis



Results

Figure 5. Summary of Exon Array Data and Validation for Representative Genes

Fold-changes of probe sets detecting exon levels (SMA versus control) are shown on top of the corresponding exon of each gene structure. Note some exons are measured by multiple probesets, while some do not have any targeting probe sets and are not shown in the gene structure. RT-PCR (R), real-time RT-PCR (Q), and sequencing (S) validation reactions are indicated below the gene structure.

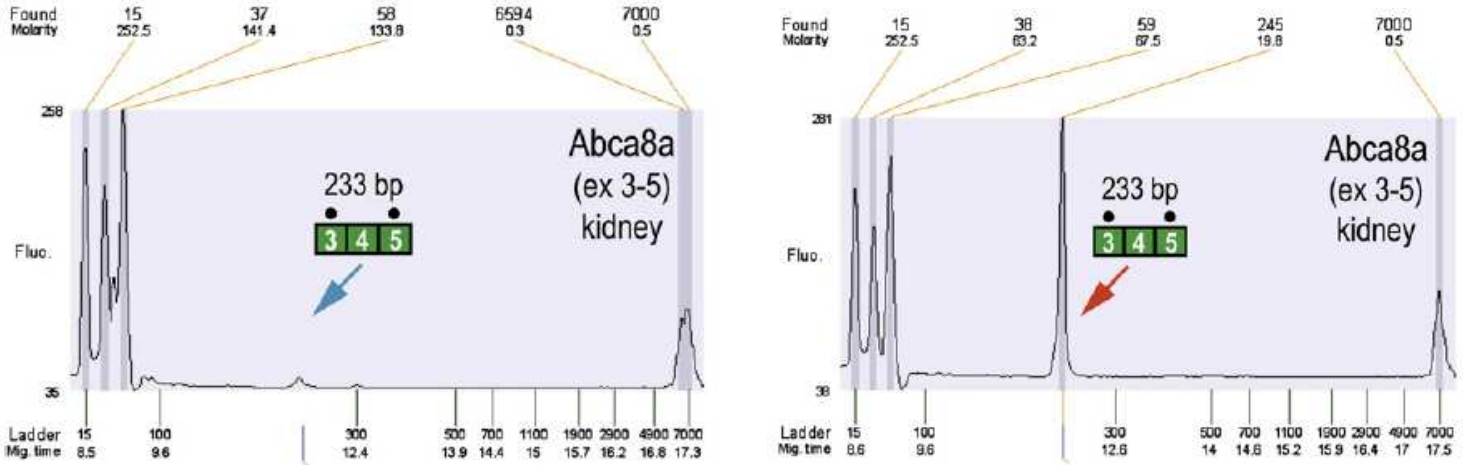


Figure 6. Confirmation of Expression and Splicing Pattern Changes by HT RT-PCR Reactions

Summary 2

- SMN deficiency causes profound changes in cellular RNA metabolism. It alters the repertoire of snRNAs and perturbs pre-mRNA splicing, leading to numerous splicing defects. The defects are widespread and cell type specific, affecting mRNAs of functionally diverse genes.

- A large degree of SMN decrease (>80%) is required to cause a significant change in the levels of snRNAs or cause cell death in cultured cells, suggesting that cells normally contain a large excess capacity of SMN complex to maintain their normal inventory of snRNAs.

- Using the stringent criteria applied to selection of the affected transcripts, only a fraction of genes for which robust signals were obtained showed splicing perturbations (<2% of genes). However the number of affected transcripts is likely to be much higher because the signals for any given exon represent the average for all transcripts in the tissue.

=> more sensitive methods of AS detection should be developed

Critics

(1) It seems that no normalization is performed on the data

(2) Only one type of analysis is performed (standard Partek pipeline), without consideration the effectiveness

Alternative Splicing Events Identified in Human Embryonic Stem Cells and Neural Progenitors

Gene W. Yeo, et al. (2007) PLoS Comput. Biology

- New method for the analysis of AS
- **Samples:** 5 samples, including:
  - Human embryonic stem cells (hESC) of 2 cell lines: Cyt-ES and HUES6-ES
  - Human central nervous system stem cells grown as neurospheres: hCNS-SCns
  - Neural progenitor (NPs) of 2 cell lines: Cyt-NP, HUES6-NP
- **Microarrays:** Affymetrix GeneChip Human Exon 1.0 ST
- **Goal**: understand post-transcriptional changes in NPs in comparison with hESCs

Human **embryonic stem cells (hESCs)** and **neural progenitor (NP)** cells are excellent models for recapitulating early neuronal development in vitro, and are key to establishing strategies for the treatment of degenerative disorders. While much effort had been undertaken to analyze transcriptional and epigenetic differences during the transition of hESC to NP, very little work has been performed to understand post-transcriptional changes during neuronal differentiation. Alternative RNA splicing (AS), a major form of post-transcriptional gene regulation, is important in mammalian development and neuronal function. Human ESC, hESC-derived NP, and human central nervous system stem cells were compared using Affymetrix exon arrays. We introduced an outlier detection approach, **REAP (Regression-based Exon Array Protocol),** to identify 1,737 internal exons that are predicted to undergo AS in NP compared to hESC. Experimental validation of REAP-predicted AS events indicated a threshold-dependent sensitivity ranging from 56% to 69%, at a specificity of 77% to 96%. REAP predictions significantly overlapped sets of alternative events identified using expressed sequence tags and evolutionarily conserved AS events. Our results also reveal that focusing on differentially expressed genes between hESC and NP will overlook 14% of potential AS genes. In addition, we found that REAP predictions are enriched in genes encoding serine/threonine kinase and helicase activities. An example is a REAP-predicted alternative exon in the SLK (serine/threonine kinase 2) gene that is differentially included in hESC, but skipped in NP as well as in other differentiated tissues. Lastly, comparative sequence analysis revealed conserved intronic cis-regulatory elements such as the FOX1/2 binding site GCAUG as being proximal to candidate AS exons, suggesting that FOX1/2 may participate in the regulation of AS in NP and hESC. In summary, a new methodology for exon array analysis was introduced, leading to new insights into the complexity of AS in human embryonic stem cells and their transition to neural stem cells.

Summary 3

- REAP, a regression-based method for analyzing exon array data was introduced, and was applied to discover AS events in hESCs, their derived NPs, and in hCNS-SCns.
- Only a minority of AS events was common between various hESC to NP comparisons.
- They estimate that at least 1,336 of 1,737 REAP exons were true AS events that changed during neuronal differentiation of hESC cells, and/or were different between endogeneous NPs and hESC.
- The results also revealed that focusing on differentially expressed genes between hESC and NP will overlook 14% of potential AS genes.

Critics

Only one type of analysis is performed

Lack of solid conclusions

Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer

Xi L., et al. (2008) Nucl. Acid Res.

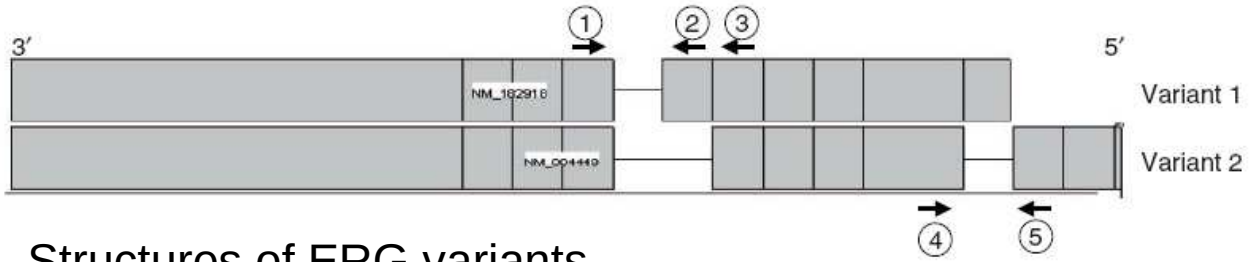
- Modern and solid work. Usage of Partek.
- **Samples:** 115 samples (79 tumour, 36 normal), of which 40 samples used for hybridization:
  - 20 normal lung tissues
  - 20 cancer lung tissues (paired with normal)
- **Microarrays:** Affymetrix GeneChip Human Exon 1.0 ST
- **Goal**: identify cancer-associated alternative splicing events, verify splice variants and to validate differential expression of selected splice variants in independent tissue sets.

Alternative processing of pre-mRNA transcripts is a major source of protein diversity in eukaryotes and has been implicated in several disease processes including cancer. In this study we have performed a genome wide analysis of alternative splicing events in lung adenocarcinoma. We found that 2369 of the 17 800 core Refseq genes appear to have alternative transcripts that are differentially expressed in lung adenocarcinoma versus normal. According to their known functions the largest subset of these genes (30.8%) is believed to be cancer related. Detailed analysis was performed for several genes using PCR, quantitative RT-PCR and DNA sequencing. We found overexpression of ERG variant 2 but not variant 1 in lung tumors and overexpression of CEACAM1 variant 1 but not variant 2 in lung tumors but not in breast or colon tumors. We also identified a novel, overexpressed variant of CDH3 and verified the existence and overexpression of a novel variant of P16 transcribed from the CDKN2A locus. These findings demonstrate how analysis of alternative pre-mRNA processing can shed additional light on differences between tumors and normal tissues as well as between different tumor types. Such studies may lead to the development of additional tools for tumor diagnosis, prognosis and therapy.

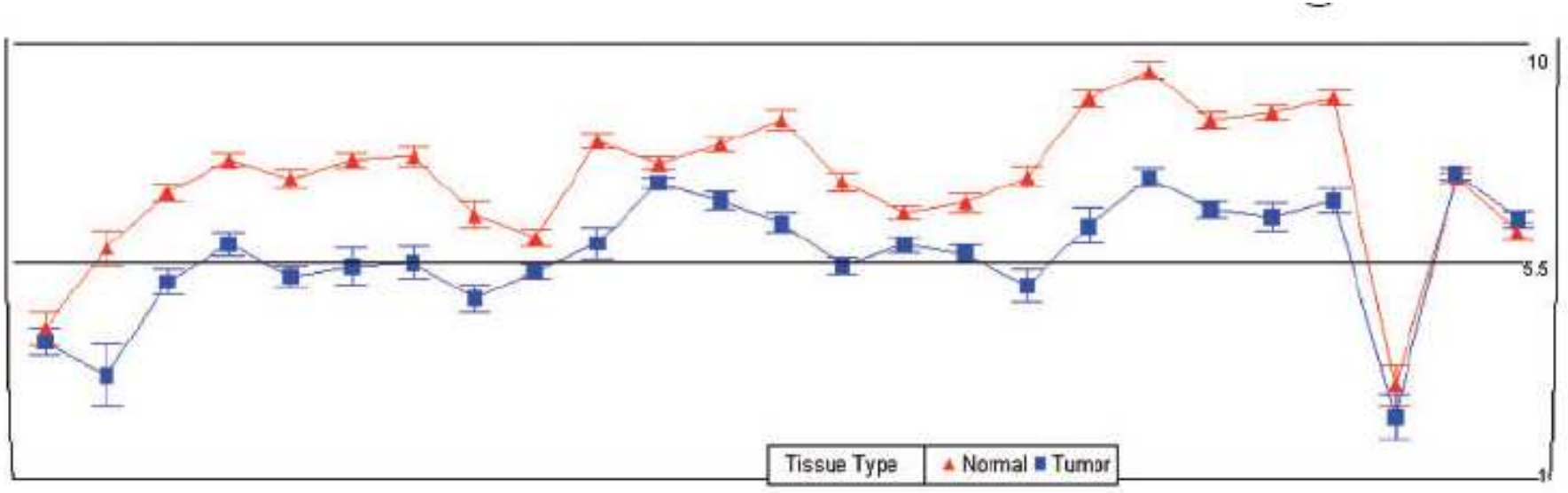
Pipeline

(0) QC with Affymetrix expression console.
(1) Probe-level analysis with GC-RMA in Partek GS.
(2) Alternative splice analysis of exon data with ANOVA in Partek GS.
(3) Function analysis of genes with alternative splicing in Ingenuity Pathway Analysis and generation of a cancer-related gene list.
(4) Manual review of Partek gene view plots to identify alternate splicing forms and determine the frequency of changes observed in the patient set.
(5) Detailed manual analysis focusing on genes with simple forms of alt splicing and high frequency of changes. Reviewed Affymetrix probeset sequences, RefSeq database, Blatted probe sequences in UCSC Genomic Browser.
(6) Identification of genes with alternative splicing at high frequency in this patient set (>50% patients with same change).
(7) Verification and validation.

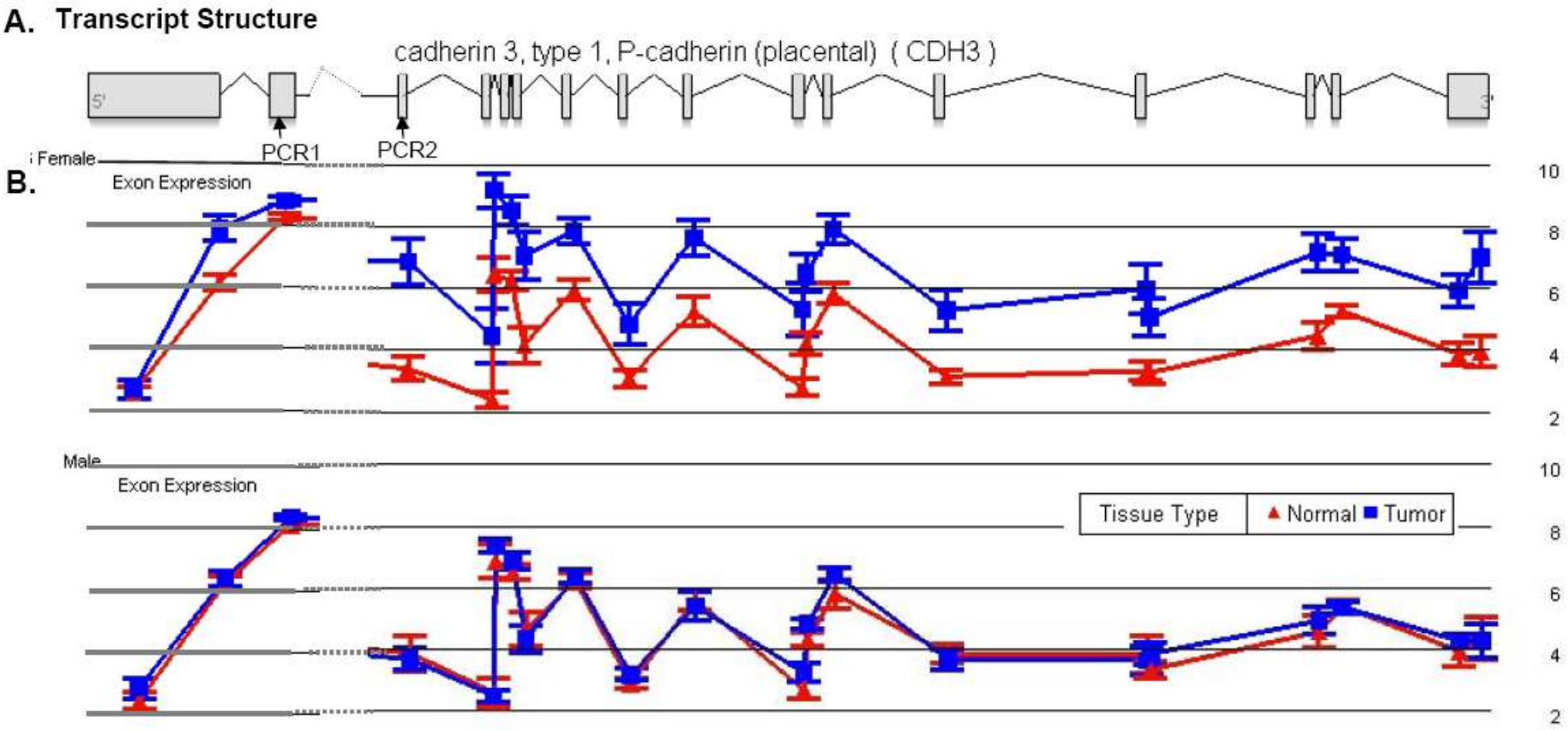
Results: ERG variant



Structures of ERG variants



Exon expression for ERG per patient



Exon expression for CDH3 per exon

Summary 4

- They found that 2369 of the 17 800 core Refseq genes appear to have alternative transcripts that are differentially expressed in lung adenocarcinoma versus normal. According to their known functions the largest subset of these genes (30.8%) is believed to be cancer related.
- Overexpression of ERG variant-2 but not variant-1 in lung tumors and overexpression of CEACAM1 variant-1 but not variant-2 in lung tumors but not in breast or colon.
- A novel novel, overexpressed variant of CDH3 and verified the existence and overexpression of a novel variant of P16 transcribed from the CDKN2A locus.

Critics

(1) It seems that no normalization is performed on the data

(2) Only one type of analysis is performed (standard Partek pipeline), without consideration the effectiveness

Problems and Solutions of AS Detection

- **High variability** in the data significantly **hampers detection of splicing events**. This is crucial for aberrant splicing detection, as it may be a random and hardly reproducible event.

- A number of methods has been proposed recently for splicing detection:
  - Splicing index [Srinivasan, 2005]
  - ANOSVA [Cline, 2005]
  - Clustering by physical adjacency [Fan, 2006]
  - REAP protocol [Yeo, 2007]
  - FIRMA method [Purdom, 2008]

But no benchmarking has been performed yet !

• There is **no software tool, which would satisfy all needs**. Partek is a good candidate, but it is still “wet” and cannot be used for advanced analysis. **The only way is to use the combination** of R/Bioconductor, Partek, etc.

• Often researchers consider existing software as ***panacea***, and forgot to look critically on what they do. They blindly believe in the pipelines and realized algorithms without looking in depth.

Potential Project: Developing and Supporting Statistical Analysis Pipeline for the Study of Gene Alternative Splicing using Affymetrix GeneChip Exon Arrays