NCI Center for Biomedical Informatics and Information Technology
(CBIIT) Speaker Series, October 14, 2015

Northwestern University Feinberg School of Medicine, Chicago, IL, USA

# Cancer genomes are too complex: It is time to move away from simple gene-centric approaches;

# *and adapt to isoform-centric approaches*

Ramana V Davuluri, PhD
Department of Preventive Medicine – Division of Health and Biomedical Informatics
Department of Neurological Surgery
Robert H Lurie Comprehensive Cancer Center

Robert H. Lurie
Comprehensive Cancer Center
of Northwestern University

Northwestern University
Feinberg
School of Medicine

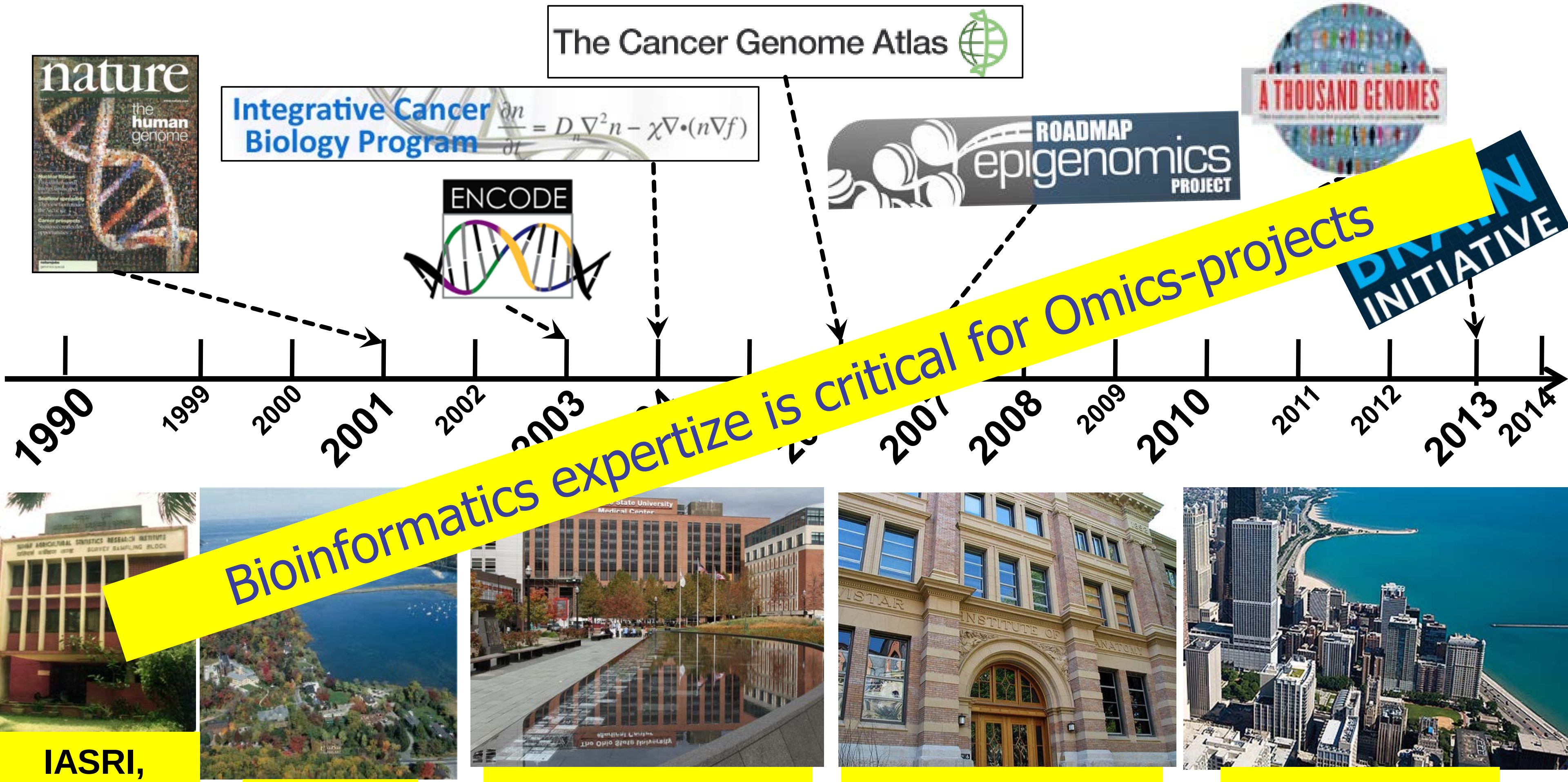
# Topics of Discussion

1. Grows of multi –omics data
2. Why “gene” as a unit of measure is too simplistic?
3. Exon-arrays and RNA-seq methods
4. Gene-level Vs Isoform-level analysis
   A. Cancer Vs Non-cancer cell-line grouping
   B. Isoform-level gene signatures for brain tumor sub-typing
5. Evaluation of isoform-level expression estimation algorithms for RNA-seq and exon-array platforms

# Topics of Discussion

1. Grows of multi –omics data
2. Why study “gene expression” at isoform-level?
3. Exon-arrays and RNA-seq methods
4. Gene-level Vs Isoform-level analysis
   A. Cancer Vs Non-cancer cell-line grouping
   B. Isoform-level gene signatures for brain tumor sub-typing
5. Evaluation of isoform-level expression estimation algorithms for RNA-seq and exon-array platforms

# Growth of multi –omics Data



nature the human genome

Integrative Cancer Biology Program $\frac{\partial n}{\partial t} = D_n\nabla^2 n - \chi\nabla\cdot(n\nabla f)$

ENCODE

The Cancer Genome Atlas

ROADMAP epigenomics PROJECT

A THOUSAND GENOMES

INITIATIVE

1990 1999 2000 2001 2002 2003 2007 2008 2009 2010 2011 2012 2013 2014

Bioinformatics expertize is critical for Omics-projects

**IASRI, New Delhi**

Statistics & Computer Science

**CSHL, NY**

Biology

**OSU, Columbus, OH**

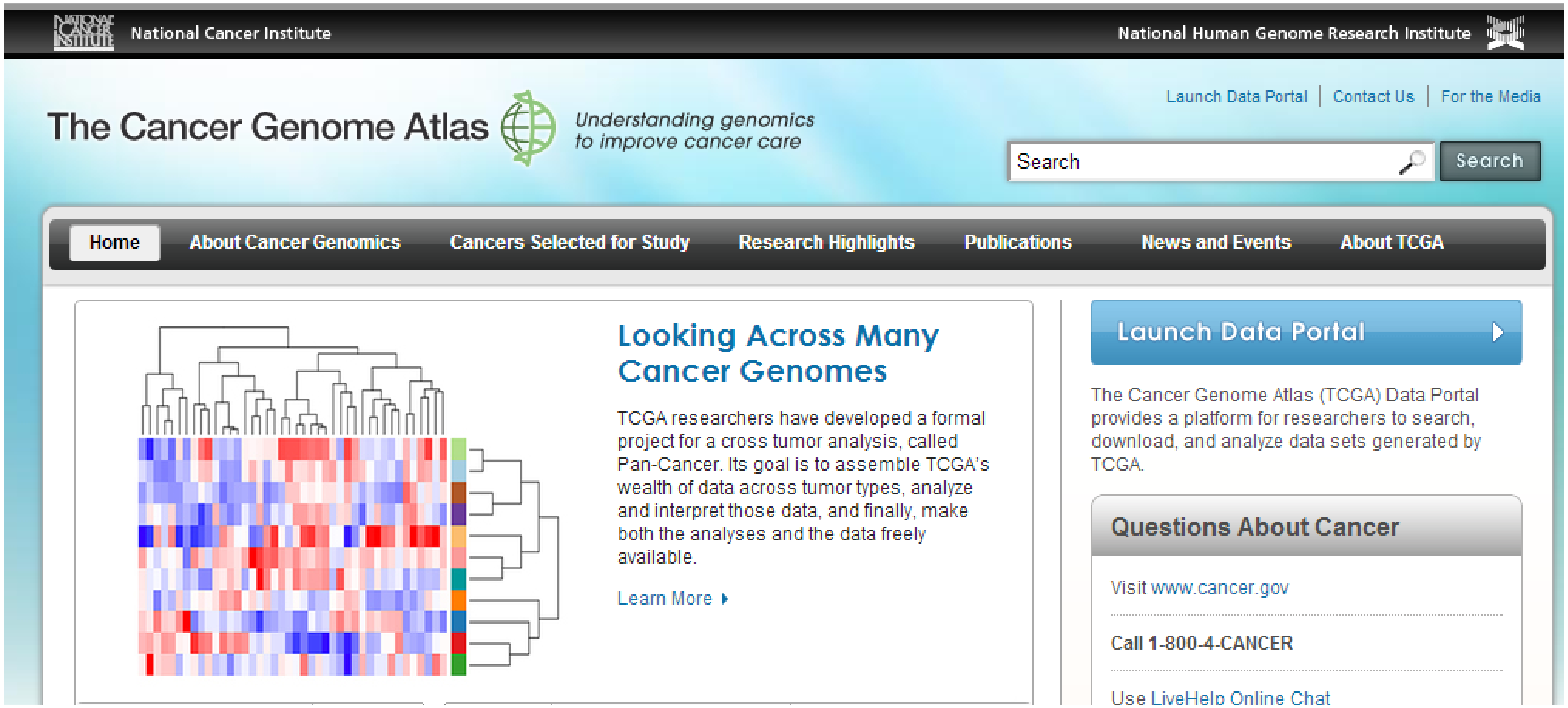
Medicine

**Wistar, Philadelphia**

Biology

**NUFSM, Chicago**

Medicine


National Cancer Institute
National Human Genome Research Institute
Launch Data Portal | Contact Us | For the Media
Search
Search
Home
About Cancer Genomics
Cancers Selected for Study
Research Highlights
Publications
News and Events
About TCGA
Launch Data Portal
Learn More

The Cancer Genome Atlas

*Understanding genomics to improve cancer care*

## Looking Across Many Cancer Genomes

TCGA researchers have developed a formal project for a cross tumor analysis, called Pan-Cancer. Its goal is to assemble TCGA's wealth of data across tumor types, analyze and interpret those data, and finally, make both the analyses and the data freely available.

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA.

### Questions About Cancer

Visit www.cancer.gov

Call 1-800-4-CANCER

Use LiveHelp Online Chat

- TCGA pilot started in 2006
  - NCI & NHGRI (with an investment of $50 million each)
  - Atlas of genomic changes created for specific cancer types
- Expanded to >20 additional tumor types
- New approaches to the detection, diagnosis, treatment, and possibly prevention of the disease

# TCGA datasets currently available

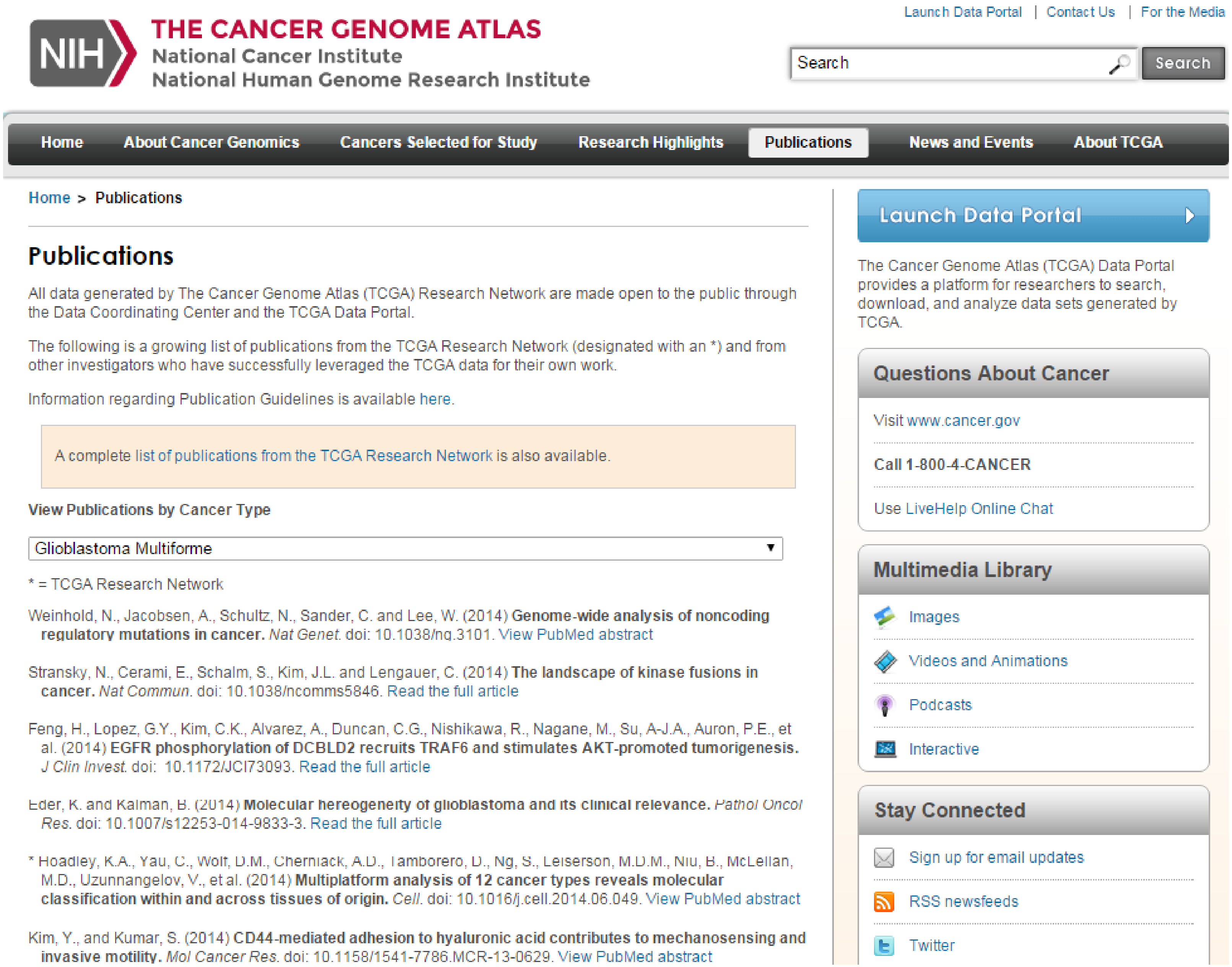
Total Cancers: 42 Total Live File Count: **106527**

Total Size of All Live Files: **2,309,174.2 Gigabytes**

| disease | disease (abbr) | file_counts | size_in_gigabytes |
|---|---|---|---|
| Glioblastoma multiforme | GBM | 3137 | 85856.6 |
| Brain Lower Grade Glioma | LGG | 3817 | 47202.1 |
| Lung adenocarcinoma | LUAD | 4620 | 67961.1 |
| Lung squamous cell carcinoma | LUSC | 4139 | 73908.1 |
| Breast invasive carcinoma | BRCA | 10151 | 142070.7 |
| Ovarian serous cystadenocarcinoma | OV | 6197 | 130444.3 |
| Prostate adenocarcinoma | PRAD | 3336 | 40841.1 |

https://cghub.ucsc.edu/summary_stats.html

# Publications from analyses of TCGA datasets



NIH
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute
Launch Data Portal | Contact Us | For the Media
Search
Search
Home
About Cancer Genomics
Cancers Selected for Study
Research Highlights
Publications
News and Events
About TCGA
Home > Publications

## Publications

All data generated by The Cancer Genome Atlas (TCGA) Research Network are made open to the public through the Data Coordinating Center and the TCGA Data Portal.

The following is a growing list of publications from the TCGA Research Network (designated with an *) and from other investigators who have successfully leveraged the TCGA data for their own work.

Information regarding Publication Guidelines is available here.

A complete list of publications from the TCGA Research Network is also available.

**View Publications by Cancer Type**

Glioblastoma Multiforme ▾

* = TCGA Research Network

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. and Lee, W. (2014) **Genome-wide analysis of noncoding regulatory mutations in cancer.** *Nat Genet.* doi: 10.1038/ng.3101. View PubMed abstract

Stransky, N., Cerami, E., Schalm, S., Kim, J.L. and Lengauer, C. (2014) **The landscape of kinase fusions in cancer.** *Nat Commun.* doi: 10.1038/ncomms5846. Read the full article

Feng, H., Lopez, G.Y., Kim, C.K., Alvarez, A., Duncan, C.G., Nishikawa, R., Nagane, M., Su, A-J.A., Auron, P.E., et al. (2014) **EGFR phosphorylation of DCBLD2 recruits TRAF6 and stimulates AKT-promoted tumorigenesis.** *J Clin Invest.* doi: 10.1172/JCI73093. Read the full article

Eder, K. and Kalman, B. (2014) **Molecular hereogeneity of glioblastoma and its clinical relevance.** *Pathol Oncol Res.* doi: 10.1007/s12253-014-9833-3. Read the full article

* Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunnangelov, V., et al. (2014) **Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin.** *Cell.* doi: 10.1016/j.cell.2014.06.049. View PubMed abstract

Kim, Y., and Kumar, S. (2014) **CD44-mediated adhesion to hyaluronic acid contributes to mechanosensing and invasive motility.** *Mol Cancer Res.* doi: 10.1158/1541-7786.MCR-13-0629. View PubMed abstract

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA.

# Topics of Discussion

1. Grows of multi –omics data
2. Why “gene” as a unit of measure is too simplistic?
3. Exon-arrays and RNA-seq methods
4. Gene-level Vs Isoform-level analysis
   - A. Cancer Vs Non-cancer cell-line grouping
   - B. Isoform-level gene signatures for brain tumor sub-typing
5. Evaluation of isoform-level expression estimation algorithms for RNA-seq and exon-array platforms

# We need to re-think

1. “Gene” as a unit of measure in the human genome

    - Gene Expression
    - Gene Regulation

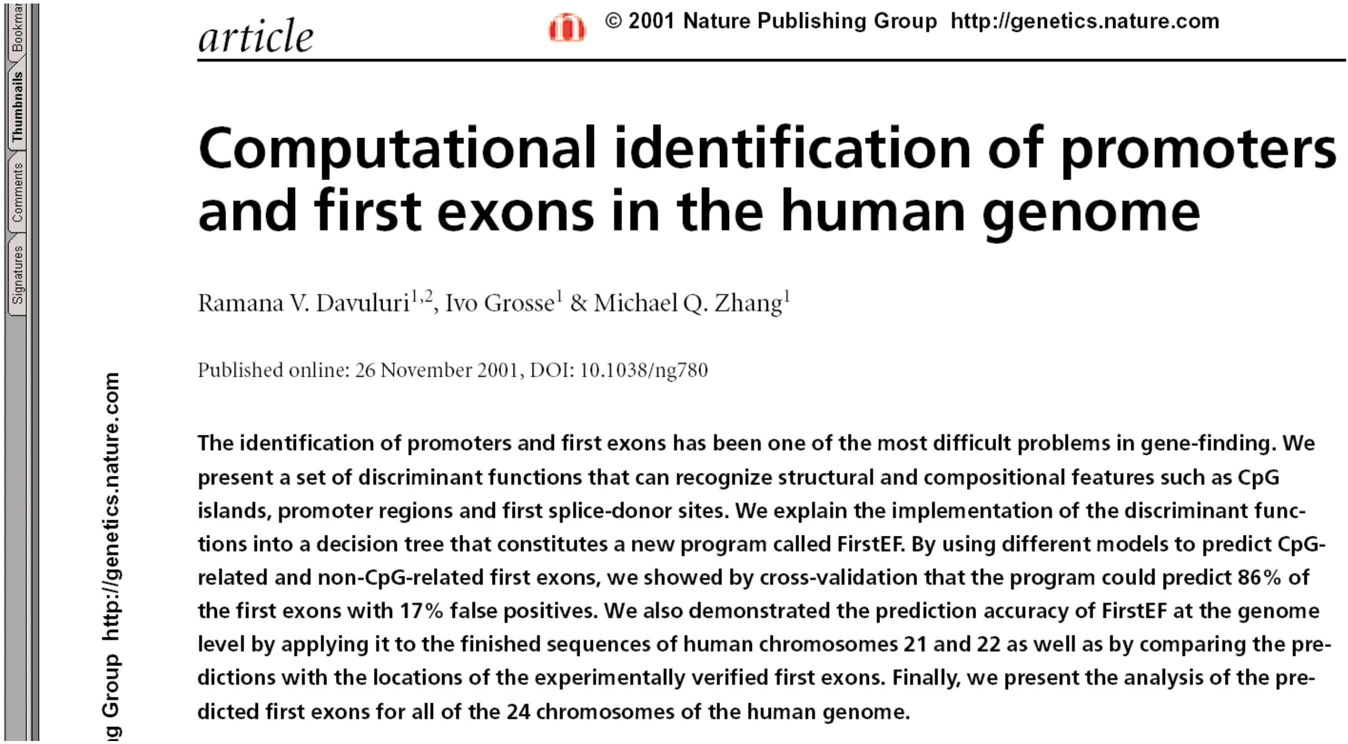
~~“one gene → one mRNA → one functional protein product”~~

“one gene → multiple mRNAs → multiple **protein isoforms** and/or **ncRNAs**”

**Pal, Gupta & Davuluri (2012) Pharmacology & Therapeutics**

# EXAMPLE – 1

## Promoter and First Exon predictions in the human genome



*article*

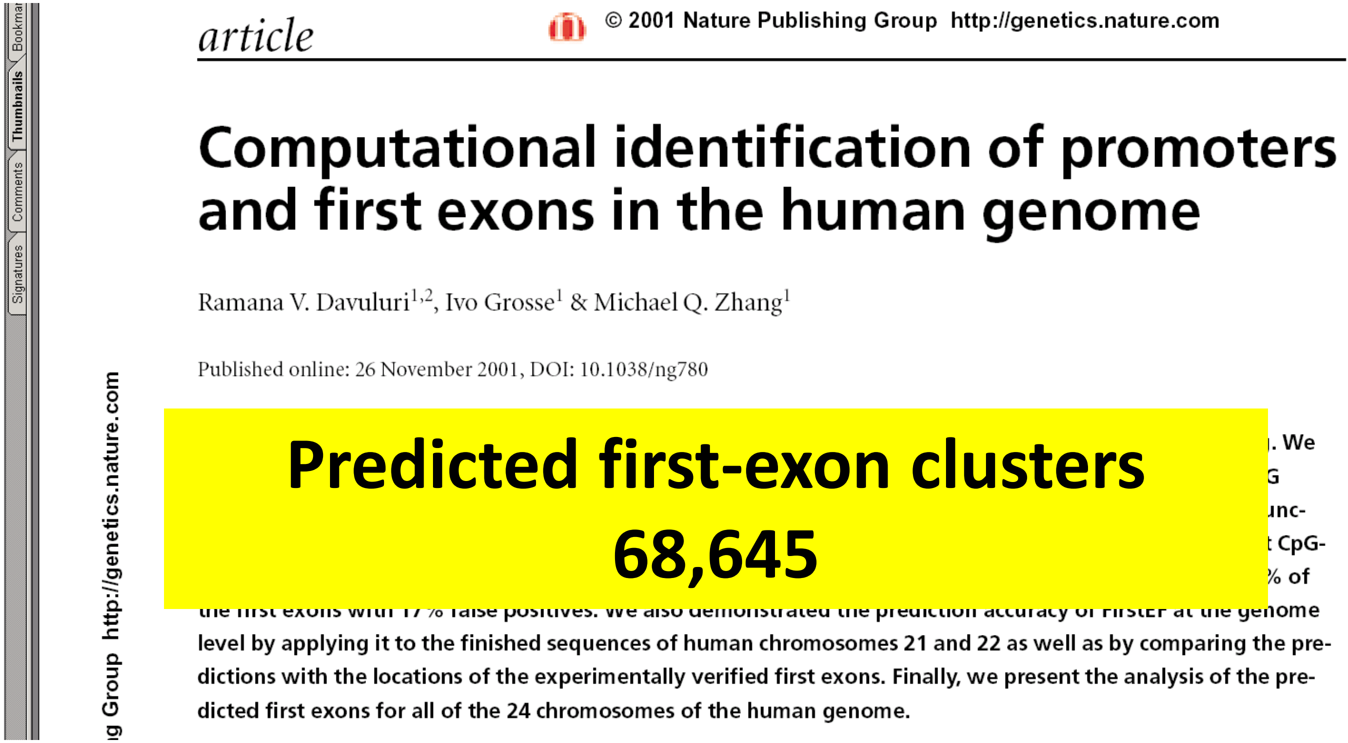
### Computational identification of promoters and first exons in the human genome

Ramana V. Davuluri[1,2], Ivo Grosse[1] & Michael Q. Zhang[1]

Published online: 26 November 2001, DOI: 10.1038/ng780

**The identification of promoters and first exons has been one of the most difficult problems in gene-finding. We present a set of discriminant functions that can recognize structural and compositional features such as CpG islands, promoter regions and first splice-donor sites. We explain the implementation of the discriminant functions into a decision tree that constitutes a new program called FirstEF. By using different models to predict CpG-related and non-CpG-related first exons, we showed by cross-validation that the program could predict 86% of the first exons with 17% false positives. We also demonstrated the prediction accuracy of FirstEF at the genome level by applying it to the finished sequences of human chromosomes 21 and 22 as well as by comparing the predictions with the locations of the experimentally verified first exons. Finally, we present the analysis of the predicted first exons for all of the 24 chromosomes of the human genome.**

# FirstEF (First Exon Finder) Program



g Group http://genetics.nature.com

article © 2001 Nature Publishing Group http://genetics.nature.com

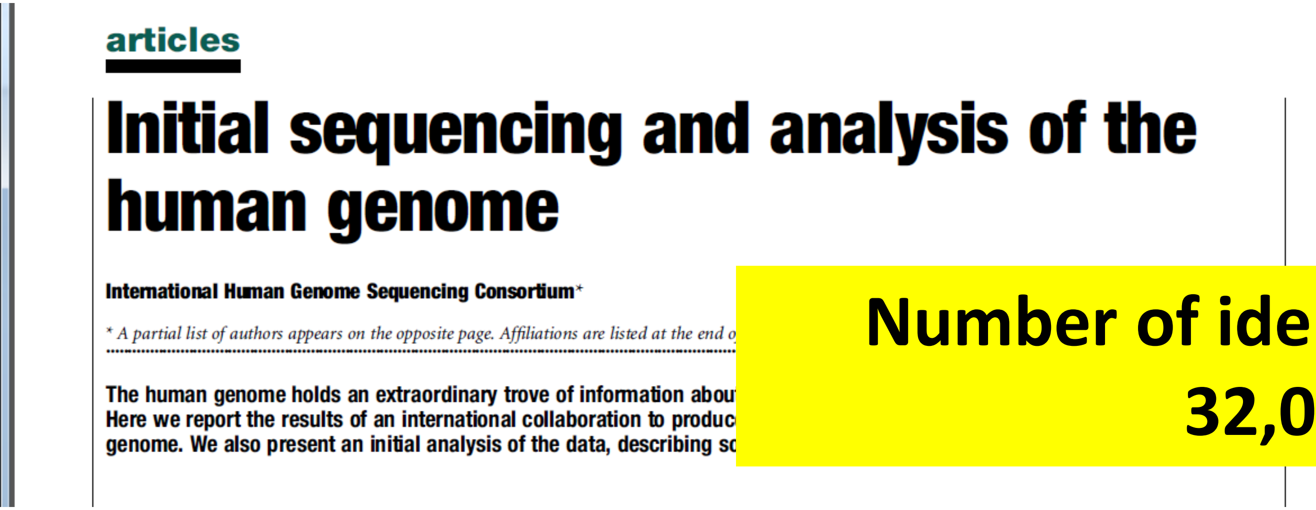
Computational identification of promoters and first exons in the human genome

Ramana V. Davuluri[1,2], Ivo Grosse[1] & Michael Q. Zhang[1]

Published online: 26 November 2001, DOI: 10.1038/ng780

. We
G
unc-
t CpG-
% of
the first exons with 17% false positives. We also demonstrated the prediction accuracy of FirstEF at the genome level by applying it to the finished sequences of human chromosomes 21 and 22 as well as by comparing the predictions with the locations of the experimentally verified first exons. Finally, we present the analysis of the predicted first exons for all of the 24 chromosomes of the human genome.

**Predicted first-exon clusters**
**68,645**



articles

Initial sequencing and analysis of the human genome

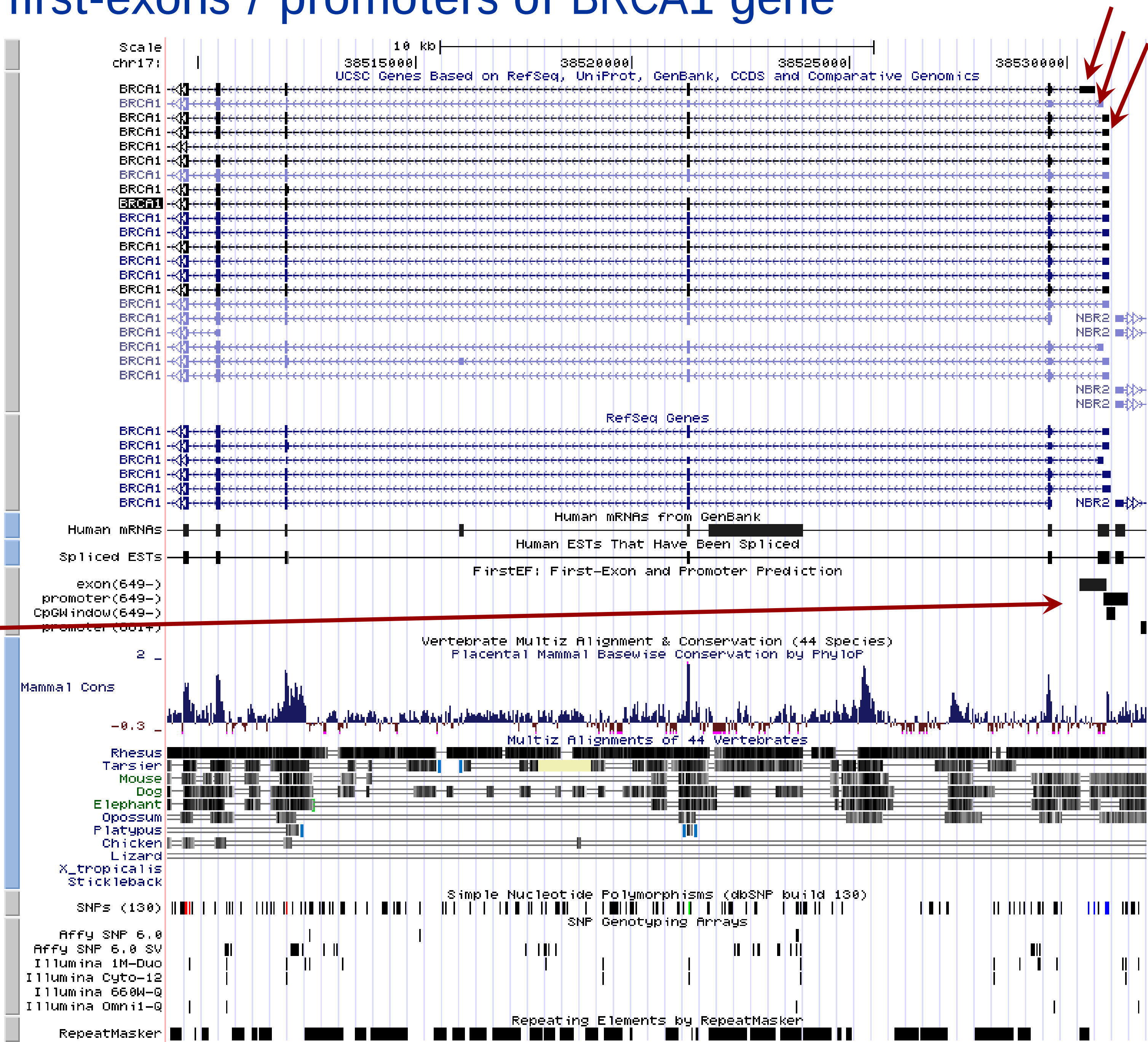
International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end o

The human genome holds an extraordinary trove of information abou
Here we report the results of an international collaboration to produc
genome. We also present an initial analysis of the data, describing sc

**Number of identified genes**
**32,000**

# Alternative first-exons / promoters of BRCA1 gene



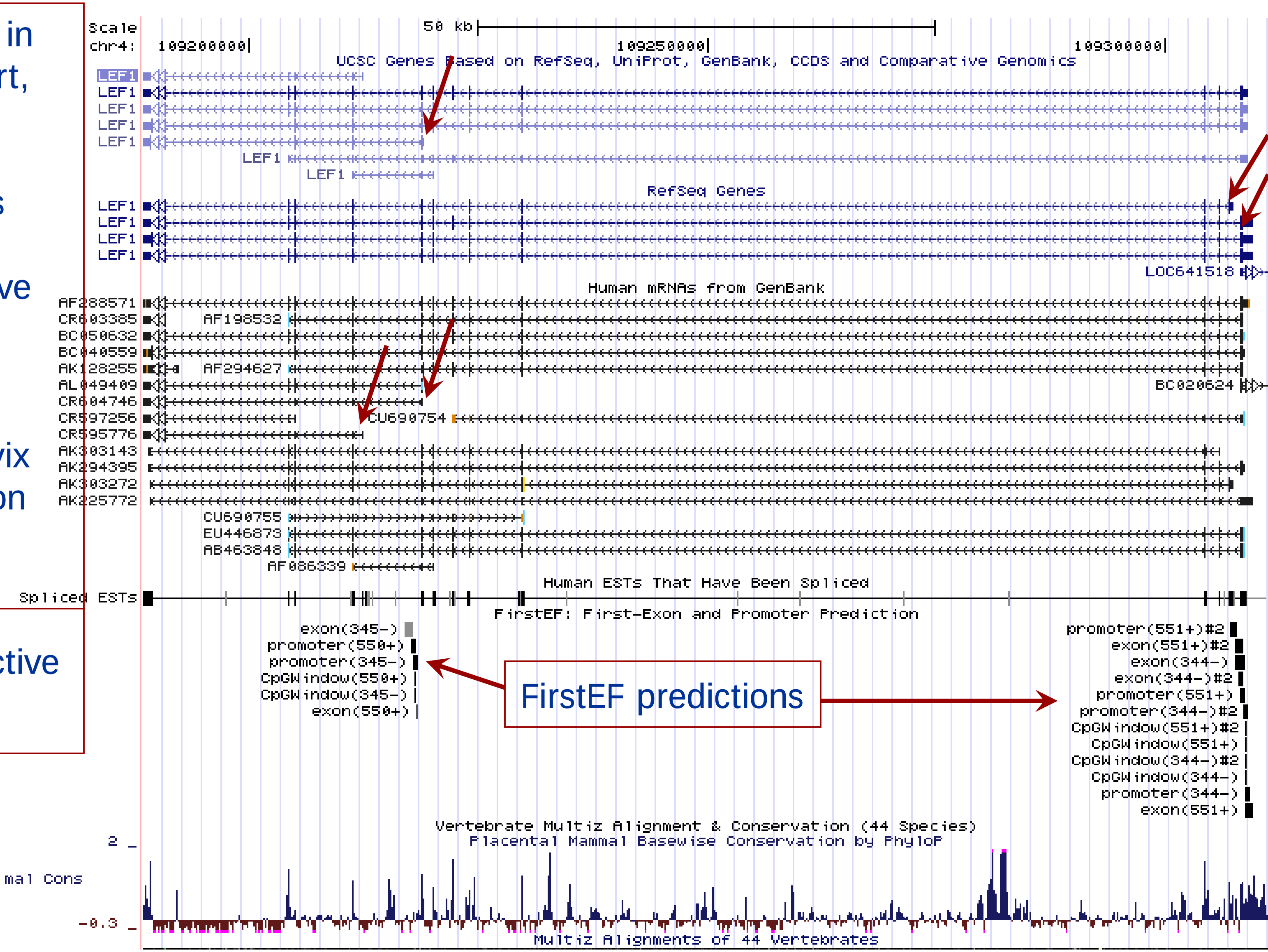
FirstEF predictions

# Production of different protein isoforms with distinct functional activities (e.g., LEF1)

Normally active in testis, fetal heart, nasopharynx, prostate, and pregnant uterus

Abnormally active in embryonal carcinoma, melanotic melanoma, cervix tumor, CLL, colon cancer

preferentially active in cancer cells



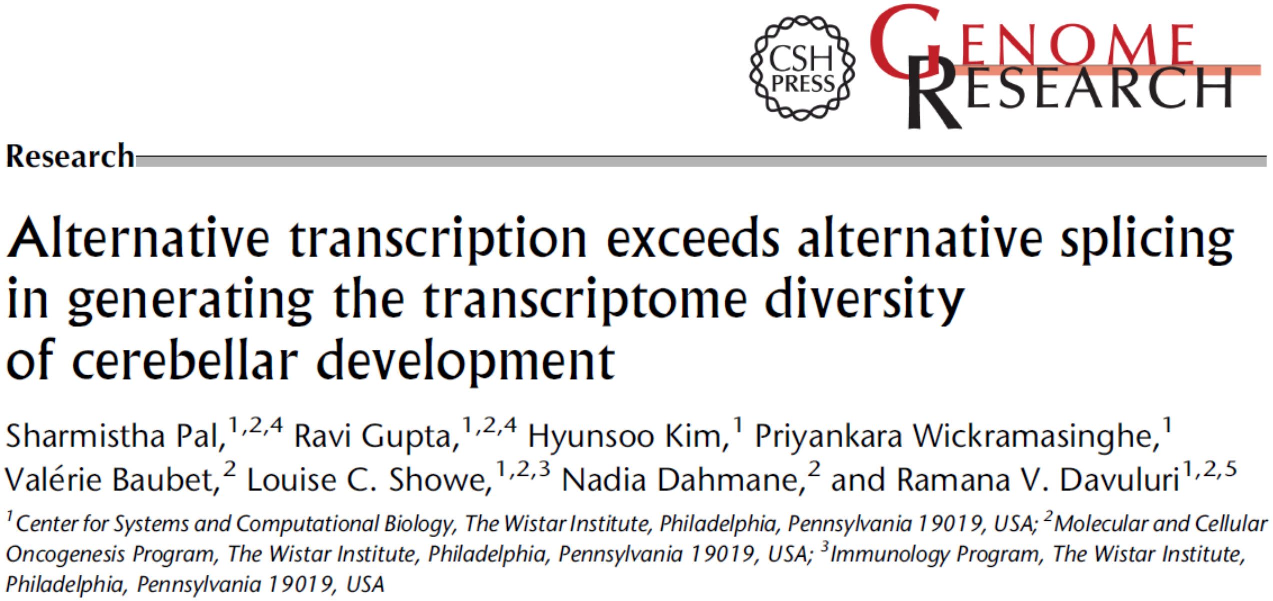
FirstEF predictions

# EXAMPLE – 2

Multiple isoforms are produced and differentially expressed in different developmental stages during brain development.

Ravi Gupta

Sharmista Pal



CSH PRESS GENOME RESEARCH

Research

## Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development

Sharmistha Pal,[1,2,4] Ravi Gupta,[1,2,4] Hyunsoo Kim,[1] Priyankara Wickramasinghe,[1] Valérie Baubet,[2] Louise C. Showe,[1,2,3] Nadia Dahmane,[2] and Ramana V. Davuluri[1,2,5]

[1] Center for Systems and Computational Biology, The Wistar Institute, Philadelphia, Pennsylvania 19019, USA; [2] Molecular and Cellular Oncogenesis Program, The Wistar Institute, Philadelphia, Pennsylvania 19019, USA; [3] Immunology Program, The Wistar Institute, Philadelphia, Pennsylvania 19019, USA

# Next-Gen. DataSets for the Study

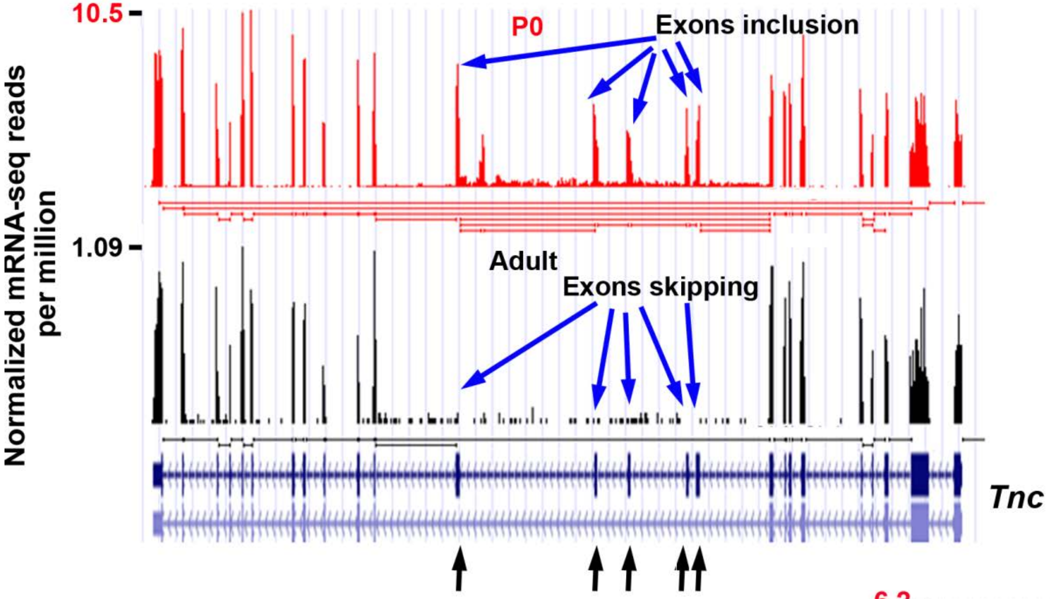
Development

| Mouse Cerebellum | mRNA Expression mRNA-Seq | IgG Control ChIP-Seq | RNA-Pol2 ChIP-Seq | H3K4me3 ChIP-Seq | H3K27me3 ChIP-Seq |
|---|---|---|---|---|---|
| Post-Natal Day0 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Post-Natal Day5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Post-Natal Day15 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Adult (Day 56) | ✓ | ✓ | ✓ | ✓ | ✓ |

# Summary – Cerebellar Transcriptome Study

- A total of 61,525 (12,796 novel) distinct mRNAs transcribed by 29,589 (4,792 novel) promoters corresponding to 15,669 protein-coding and 7,624 non-coding genes were identified.
- Aberrant use of alternative promoters in medulloblastoma.
- Gene isoforms that are specifically active in early development (no expression in adult stags) are over-expressed in cancer.
- Numerous gene isoforms are differentially expressed (but not at gene-level) during normal development and in cancer.

**Pal et al., Genome Research 2011**

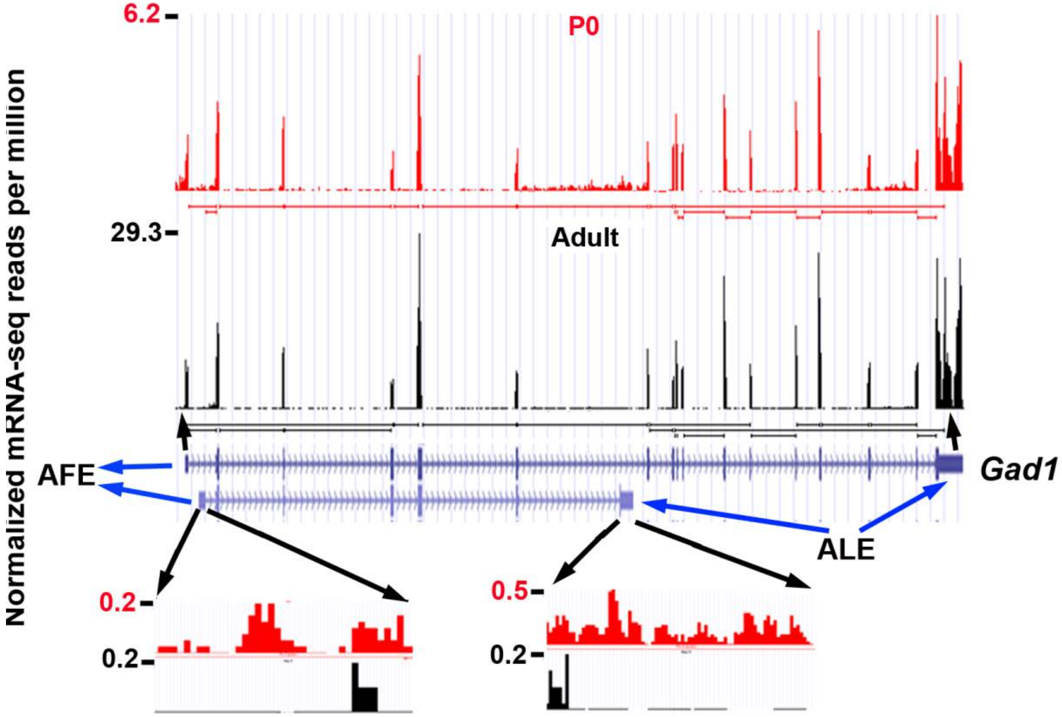


**Exon skipping** is used by tenascin-C to generate alternative mRNAs that are differentially used during early development and adult stages.

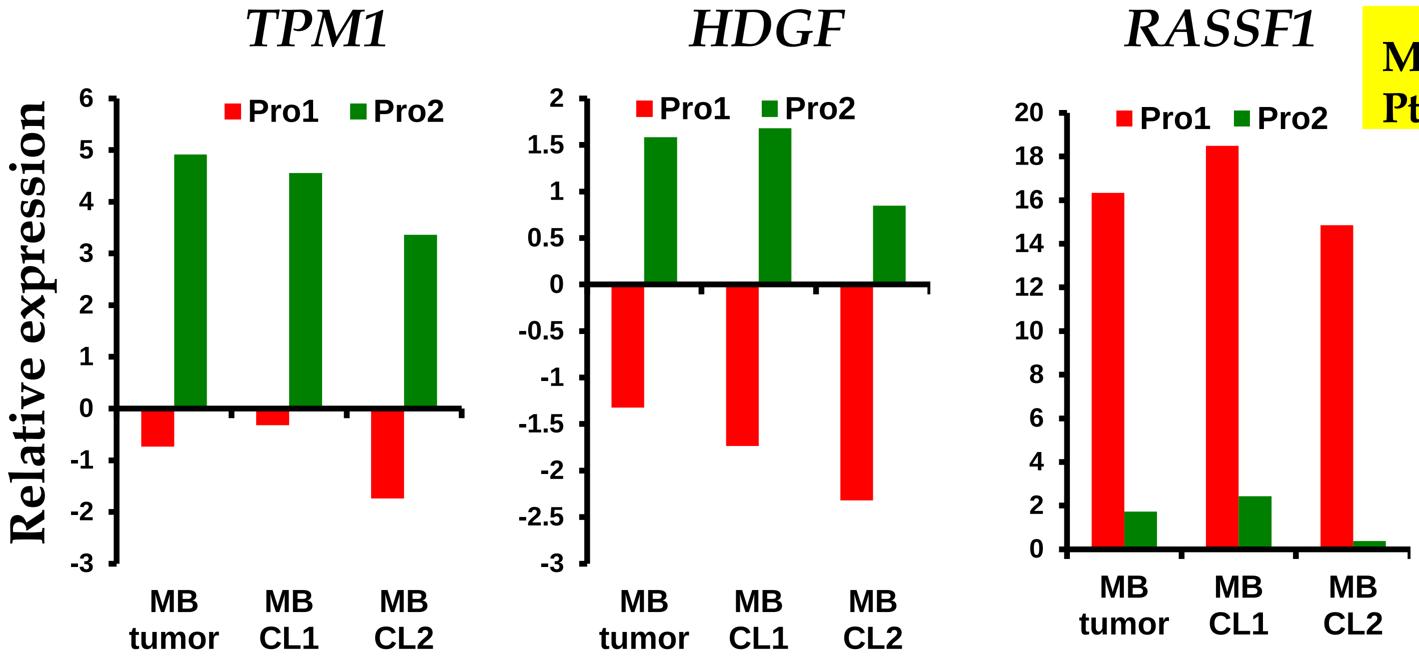
TNC is implicated in guidance of migrating neurons as well as axons during development, synaptic plasticity, and neuronal regeneration.

**Alternative transcription** is used by **Gad-1** (**glutamate decarboxylase 1 (brain, 67kDa)**)

Generate alternative pre-mRNAs that are differentially used during early development and adult stages.



# Opposite behavior of Alternative Promoters/Transcripts in Primary Medulloblastoma Tumor & derived Cell Lines



**Mouse Ptch+/-;p53-/- model**

**Promoters active during early development were turned "ON" in medulloblastoma**

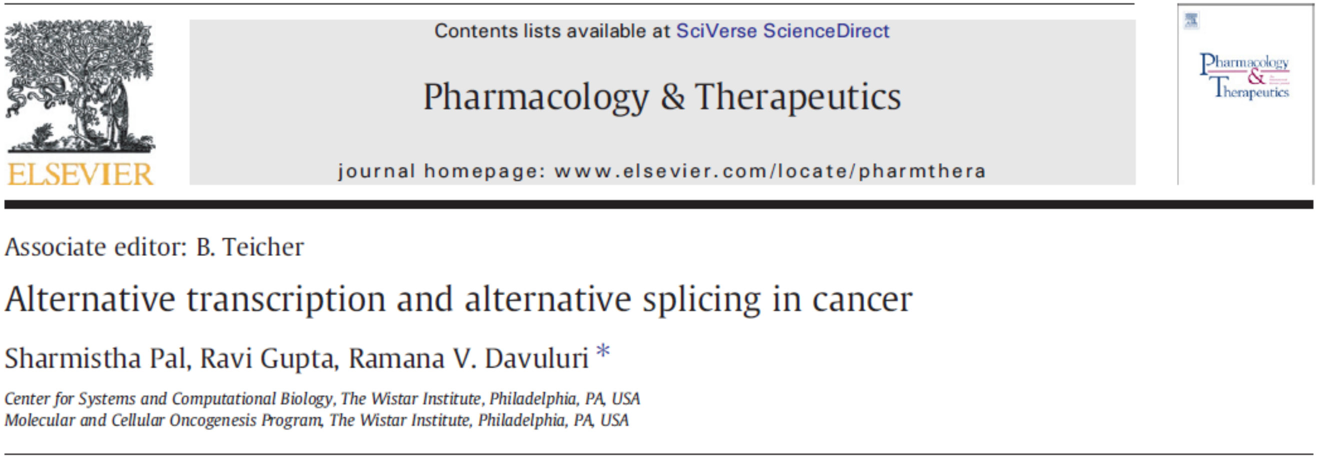
**Menghi et al, 2011, Cancer Res-" Genome-wide analysis of altersnative splicing in medulloblastoma identifies splicing patterns characteristic of normal cerebellar development.“**

**Pal et al., Genome Research 2011**

# EXAMPLE – 3

Protein isoforms are prevalent among commonly targeted genes for anti-cancer therapy.

Sharmista Pal



Contents lists available at SciVerse ScienceDirect

Pharmacology & Therapeutics

journal homepage: www.elsevier.com/locate/pharmthera

ELSEVIER
Pharmacology & Therapeutics

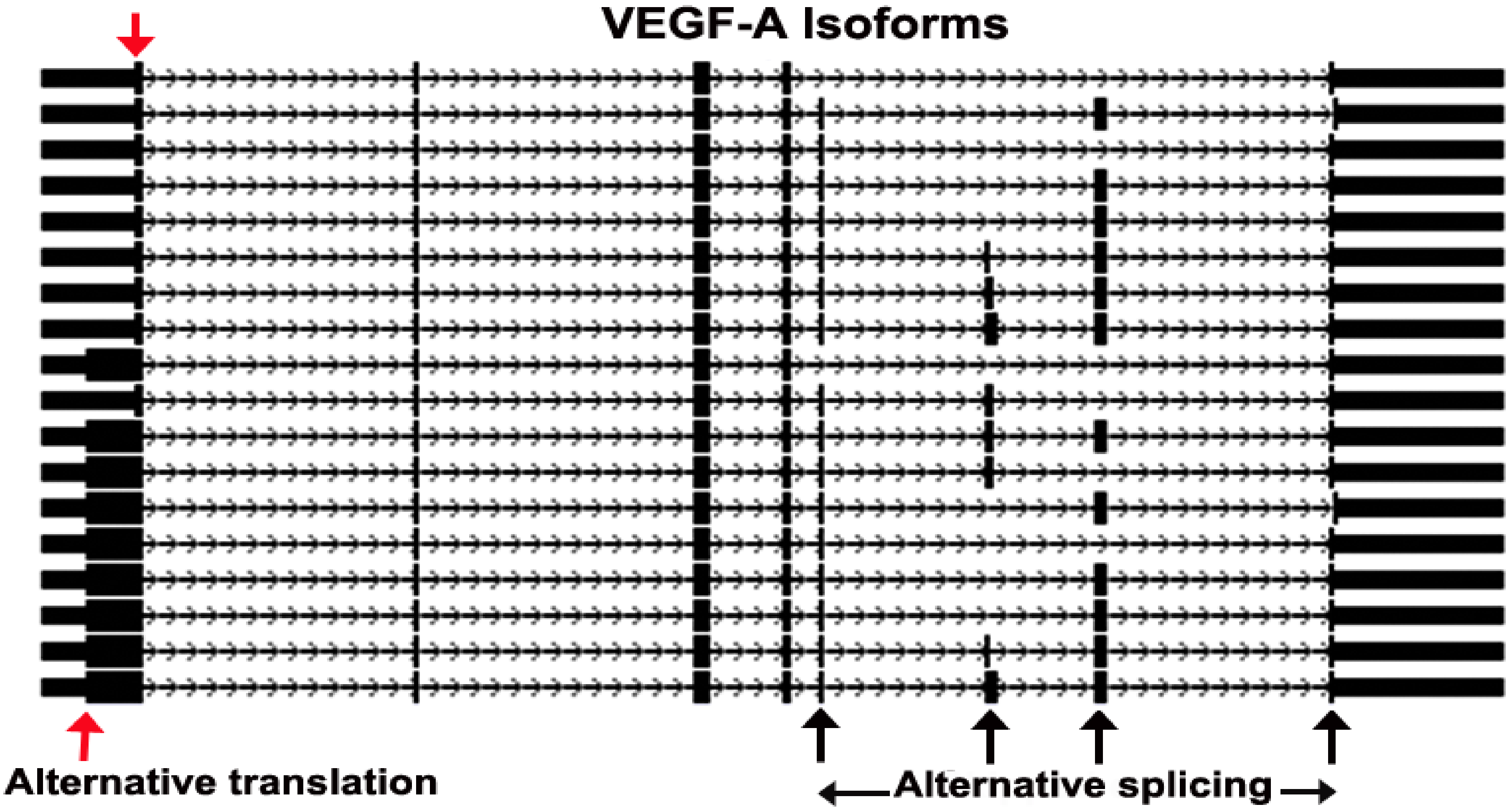
Associate editor: B. Teicher

Alternative transcription and alternative splicing in cancer

Sharmistha Pal, Ravi Gupta, Ramana V. Davuluri *

*Center for Systems and Computational Biology, The Wistar Institute, Philadelphia, PA, USA*
*Molecular and Cellular Oncogenesis Program, The Wistar Institute, Philadelphia, PA, USA*

# Molecularly targeted therapies
# (e.g. Avastin binds to circulating VEGF-A rendering it inactive)



- VEGF gene alternative splicing: pro- and anti-angiogenic isoforms in cancer (Biselli-Chicote PM et al. *J Cancer Res Clin Oncol*. 2011 Nov).

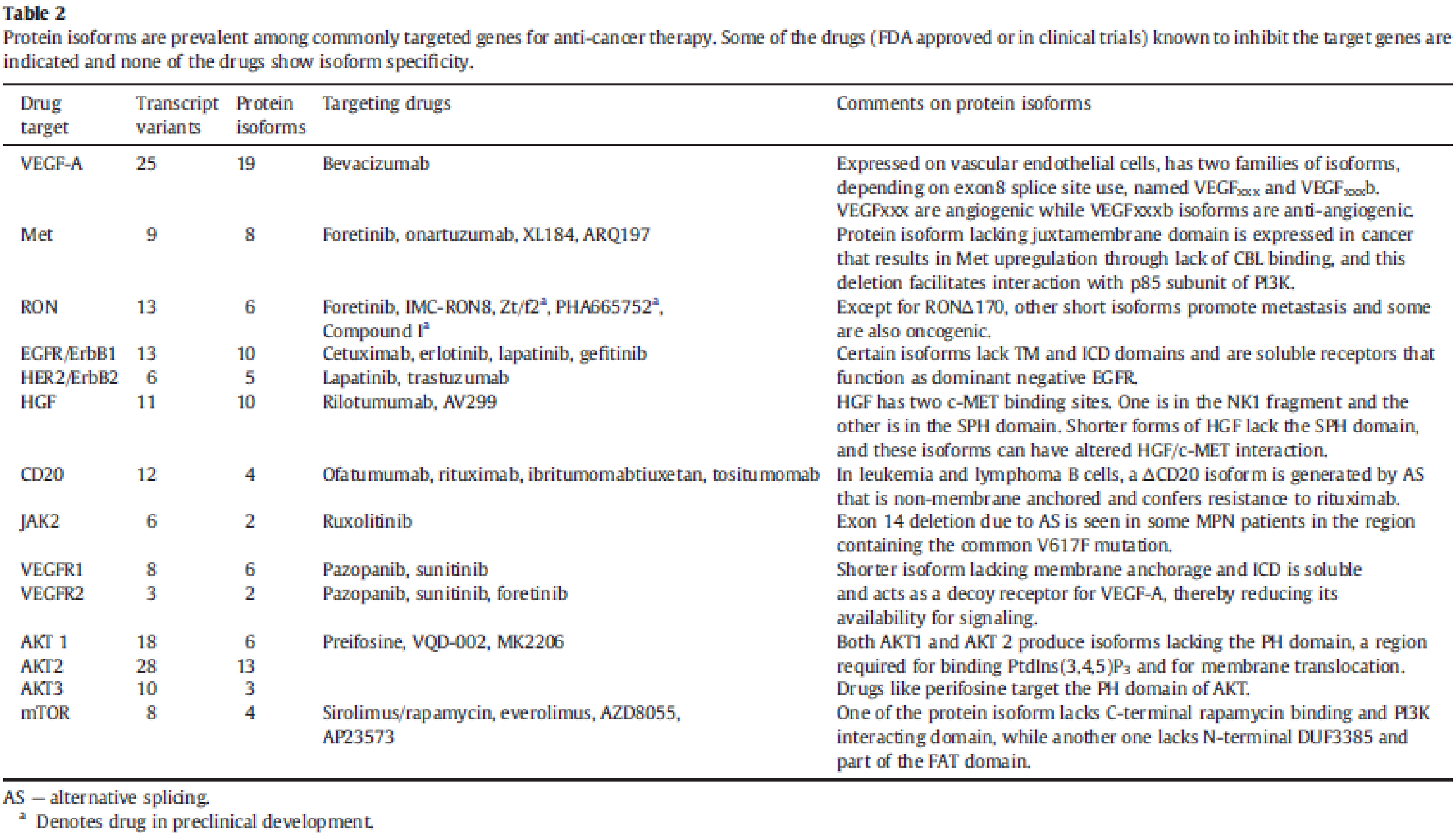


**Table 2**

Protein isoforms are prevalent among commonly targeted genes for anti-cancer therapy. Some of the drugs (FDA approved or in clinical trials) known to inhibit the target genes are indicated and none of the drugs show isoform specificity.

| Drug target | Transcript variants | Protein isoforms | Targeting drugs | Comments on protein isoforms |
|---|---|---|---|---|
| VEGF-A | 25 | 19 | Bevacizumab | Expressed on vascular endothelial cells, has two families of isoforms, depending on exon8 splice site use, named VEGF$_{xxx}$ and VEGF$_{xxx}$b. VEGFxxx are angiogenic while VEGFxxxb isoforms are anti-angiogenic |
| Met | 9 | 8 | Foretinib, onartuzumab, XL184, ARQ197 | Protein isoform lacking juxtamembrane domain is expressed in cancer that results in Met upregulation through lack of CBL binding, and this deletion facilitates interaction with p85 subunit of PI3K. |
| RON | 13 | 6 | Foretinib, IMC-RON8, Zt/f2[a], PHA665752[a], Compound I[a] | Except for RONΔ170, other short isoforms promote metastasis and some are also oncogenic. |
| EGFR/ErbB1 | 13 | 10 | Cetuximab, erlotinib, lapatinib, gefitinib | Certain isoforms lack TM and ICD domains and are soluble receptors that |
| HER2/ErbB2 | 6 | 5 | Lapatinib, trastuzumab | function as dominant negative EGFR. |
| HGF | 11 | 10 | Rilotumumab, AV299 | HGF has two c-MET binding sites. One is in the NK1 fragment and the other is in the SPH domain. Shorter forms of HGF lack the SPH domain, and these isoforms can have altered HGF/c-MET interaction. |
| CD20 | 12 | 4 | Ofatumumab, rituximab, ibritumomabtiuxetan, tositumomab | In leukemia and lymphoma B cells, a ΔCD20 isoform is generated by AS that is non-membrane anchored and confers resistance to rituximab. |
| JAK2 | 6 | 2 | Ruxolitinib | Exon 14 deletion due to AS is seen in some MPN patients in the region containing the common V617F mutation. |
| VEGFR1 | 8 | 6 | Pazopanib, sunitinib | Shorter isoform lacking membrane anchorage and ICD is soluble |
| VEGFR2 | 3 | 2 | Pazopanib, sunitinib, foretinib | and acts as a decoy receptor for VEGF-A, thereby reducing its availability for signaling. |
| AKT 1 | 18 | 6 | Preifosine, VQD-002, MK2206 | Both AKT1 and AKT 2 produce isoforms lacking the PH domain, a region |
| AKT2 | 28 | 13 | | required for binding PtdIns(3,4,5)$P_3$ and for membrane translocation. |
| AKT3 | 10 | 3 | | Drugs like perifosine target the PH domain of AKT. |
| mTOR | 8 | 4 | Sirolimus/rapamycin, everolimus, AZD8055, AP23573 | One of the protein isoform lacks C-terminal rapamycin binding and PI3K interacting domain, while another one lacks N-terminal DUF3385 and part of the FAT domain. |

AS – alternative splicing.

[a] Denotes drug in preclinical development.

Pal, Gupta & Davuluri (2012) Pharmacology & Therapeutics

# “one gene → one mRNA → one protein” model is too simplistic in the human genome

Gene counts

http://useast.ensembl.org

| **Coding genes** | **20,300** |
|---|---:|
| Small NC genes | 7,715 |
| Long NC genes | 14,863 |
| Misc NC genes | 2,307 |
| Pseudogenes | 14,424 |
| **Gene transcripts** | **198,457** |

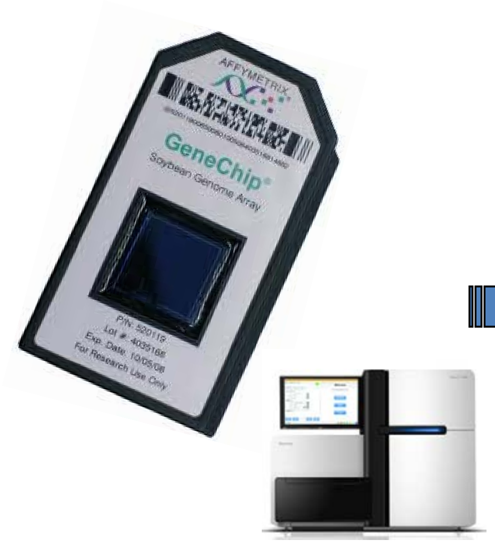
Consensus CDS counts

| **Gene IDs** | **18,826** |
|---|---:|
| CCDS IDs | 31,826 |
| Genes with >1 CCDS ID | 7,058 |

http://www.ncbi.nlm.nih.gov/CCDS/

# Sample X Gene expression data matrix

**Samples** x **Genes/Transcripts Matrix**



$$\left(X_{ij}\right)_{N \times M} =$$

| ID | Sample 1 | Sample 2 |
|---|---|---|
| ENSG00000185518 | 3.23 | 1.68 |
| ENSG00000147676 | 2.68 | 1.34 |
| ENSG00000006116 | 1.95 | 1.95 |
| ENSG00000072657 | 1.21 | 1.85 |
| ENSG00000102468 | 2.39 | 1.85 |
| ENSG00000166111 | 2.53 | 1.28 |
| ENSG00000164588 | 2.30 | 2.66 |
| ENSG00000137766 | 1.77 | 2.57 |
| ENSG00000104888 | 3.96 | 1.81 |

N – Number of genes
M – Number of samples

Gene-level analysis

$$\left(X_{ij}\right)$$

**20,000 x M**

Isoform-level analysis

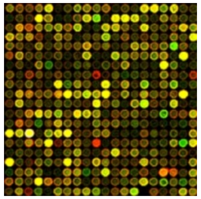
$$\left(X_{ij}\right)$$

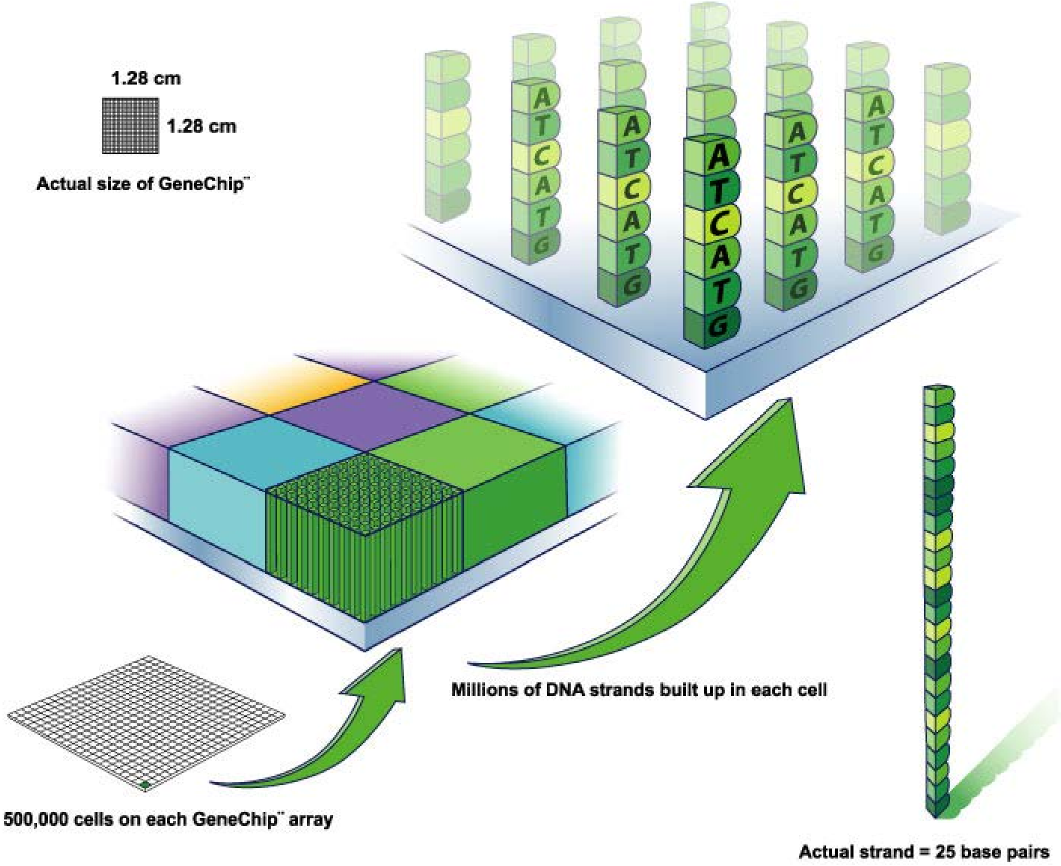
**200,000 x M**

# Topics of Discussion

1. Grows of multi –omics data
2. Why “gene” as a unit of measure is too simplistic?
3. Exon-arrays and RNA-seq methods
4. Gene-level Vs Isoform-level analysis
    A. Cancer Vs Non-cancer cell-line grouping
    B. Isoform-level gene signatures for brain tumor sub-typing
5. Evaluation of isoform-level expression estimation algorithms for RNA-seq and exon-array platforms

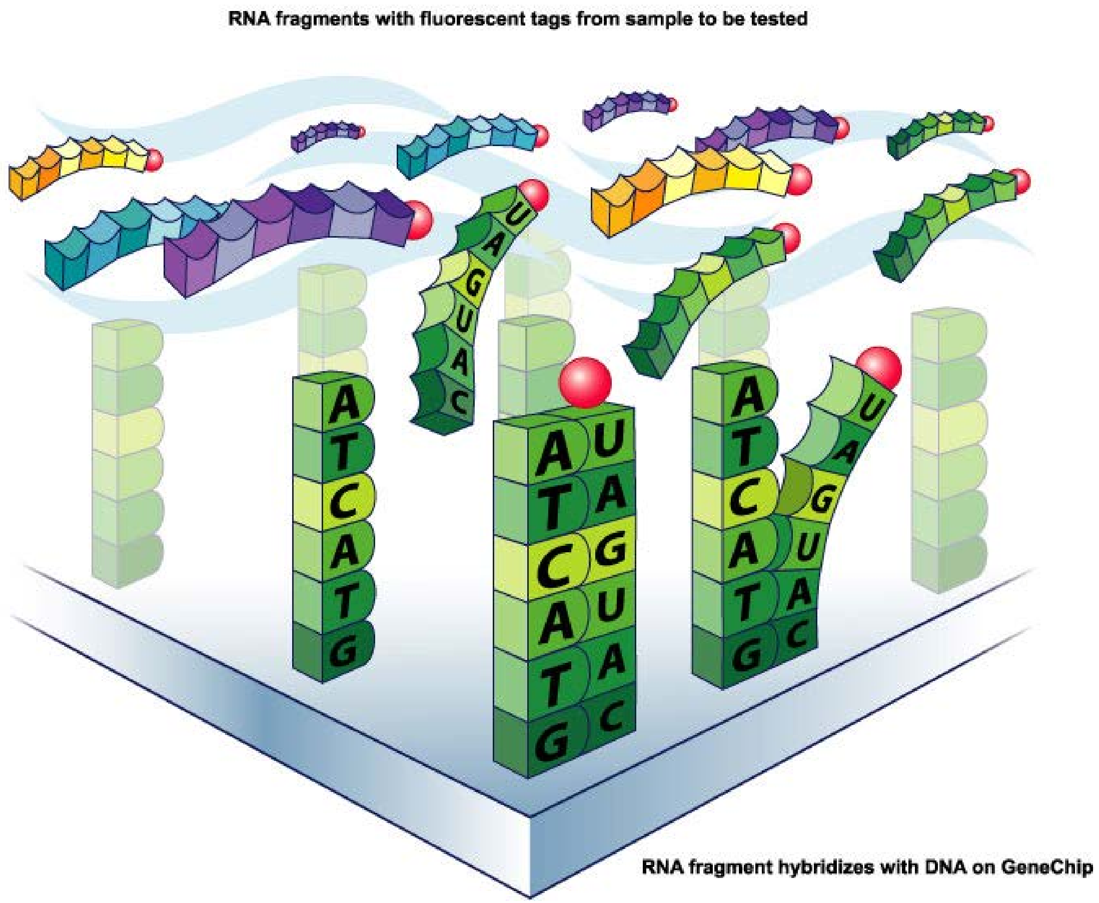
# Early days of molecular profiling – Microarrays




1.28 cm
1.28 cm
Actual size of GeneChip™
A
T
C
A
T
G
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

Cartoon of spotting/growing oligonucleotide probe on a silicon wafer. Courtesy of Affymetrix

# Hybridization to its complementary oligonucleotide probe:


RNA fragments with fluorescent tags from sample to be tested
U
A
G
U
A
C
A
T
C
A
T
G
A
T
C
A
T
G
U
A
G
U
A
C
A
T
C
A
T
G
U
A
G
U
A
C
RNA fragment hybridizes with DNA on GeneChip

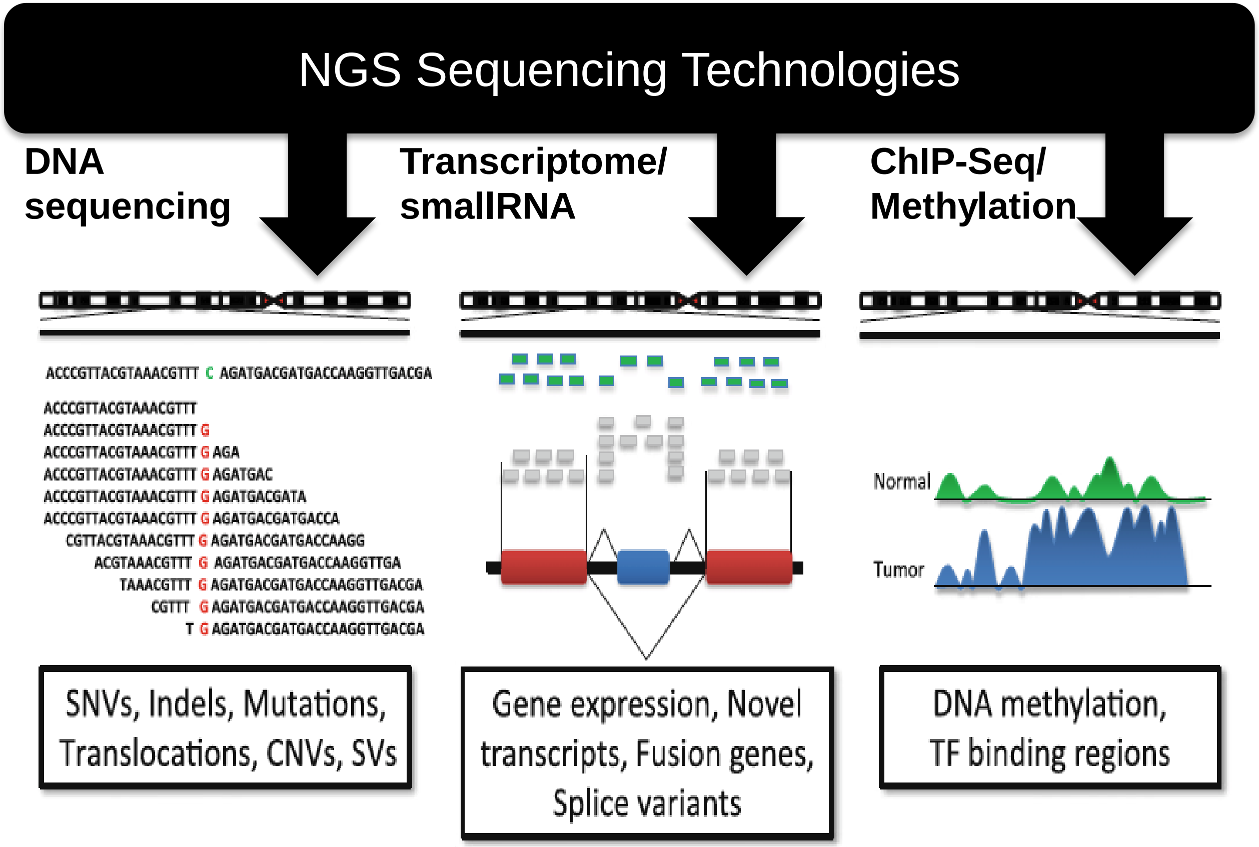
- The experimental sample, which can be either RNA or DNA, is amplified and labelled with a fluorescent tag.
- The tagged sample is then applied to the microarray.
- The tagged sample can then hybridise to its complementary oligonucleotide probe, as each feature contains millions of oligonucleotide probe, the amount of tagged sample that binds within the feature is comparable to the amount contained within the original sample

Cartoon of hybridisation of fluorescently tagged samples. Courtesy of Affymetrix

# Software to analyze gene chip data

- Estimating gene expression indices and finding significantly different genes between conditions
  - BRB-Arraytools (http://linus.nci.nih.gov/BRB-ArrayTools.html)
  - dCHIP (http://www.hsph.harvard.edu/cli/complab/dchip/)
  - SAM (http://www-stat.stanford.edu/~tibs/SAM/)
  - MMBGX (http://www.bgx.org.uk/software/mmbgx.html)
- Clustering (finding groups of samples with similar expression profiles)
  - Cluster analysis can be performed using CLUSTER software and visualize by TREEVIEW software (http://www.eisenlab.org/eisen/)
- Open Source Software for Bioinformatics
  - BioConducter (http://www.bioconductor.org/)

# Next-Generation Sequencing Technologies



NGS Sequencing Technologies

DNA sequencing

Transcriptome/ smallRNA

ChIP-Seq/ Methylation

ACCCGTTACGTAAACGTTT C AGATGACGATGACCAAGGTTGACGA
ACCCGTTACGTAAACGTTT
ACCCGTTACGTAAACGTTT G
ACCCGTTACGTAAACGTTT G AGA
ACCCGTTACGTAAACGTTT G AGATGAC
ACCCGTTACGTAAACGTTT G AGATGACGATA
ACCCGTTACGTAAACGTTT G AGATGACGATGACCA
CGTTACGTAAACGTTT G AGATGACGATGACCAAGG
ACGTAAACGTTT G AGATGACGATGACCAAGGTTGA
TAAACGTTT G AGATGACGATGACCAAGGTTGACGA
CGTTT G AGATGACGATGACCAAGGTTGACGA
T G AGATGACGATGACCAAGGTTGACGA
Normal
Tumor

SNVs, Indels, Mutations, Translocations, CNVs, SVs

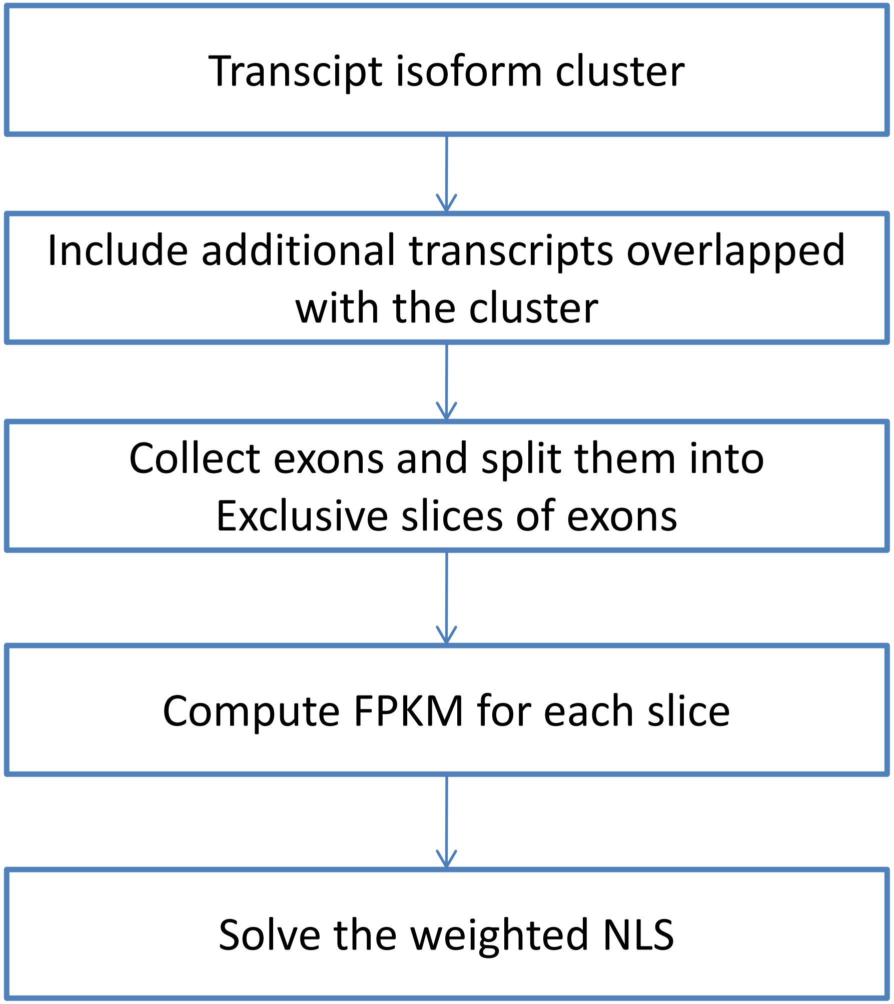
Gene expression, Novel transcripts, Fusion genes, Splice variants

DNA methylation, TF binding regions

# List of transcript abundance estimation algorithms from RNA-seq

| Algorithm | version | Reference | Estimation method | URL |
|---|---|---|---|---|
| **Cufflinks** | v2.0.2 | (Trapnell, et al., 2010), Nature biotechnology | EM | http://cufflinks.cbcb.umd.edu/ |
| **RSEM** | v1.2.3 | (Li, et al., 2010), Bioinformatics | EM | http://deweylab.biostat.wisc.edu/rsem/ |
| **eXpress** | v.1.4.0 | (Roberts and Pachter, 2013), Nature methods | online_EM | http://bio.math.berkeley.edu/eXpress/index.html |
| **IsoformEx** | v1.0.0 | (Kim, et al., 2011), BMC Bioinformatics | Weighted none-negative least squares | http://bioinformatics.wistar.upenn.edu/isoformex |
| **MMBGX** | v0.99.20 | (Turro, et al., 2010), Nucleic acids research | Bayesian | http://www.bgx.org.uk/software/mmbgx.html |

# IsoformEx Algorithm



Transcipt isoform cluster

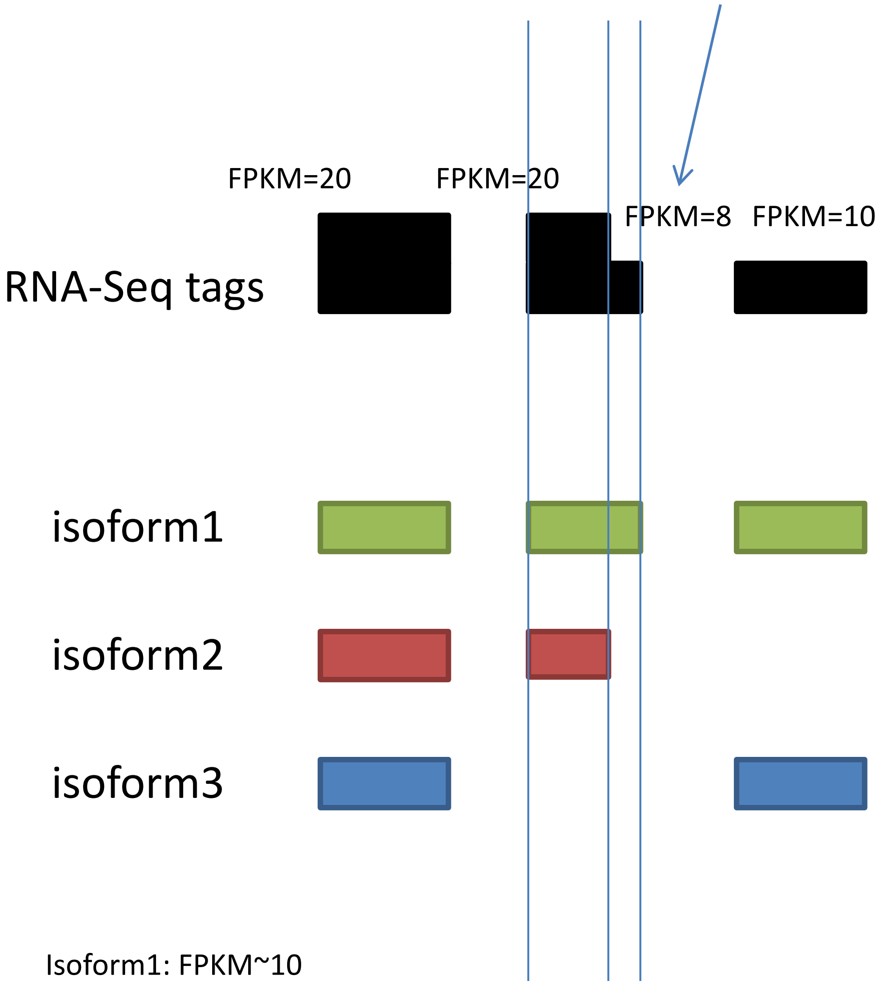
Include additional transcripts overlapped with the cluster

Collect exons and split them into Exclusive slices of exons

Compute FPKM for each slice

Solve the weighted NLS

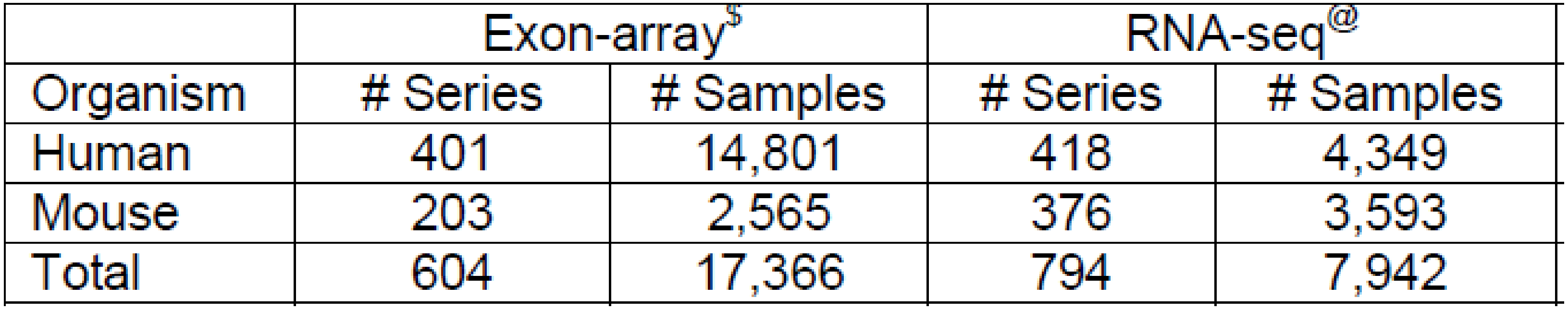
Apply low weight since this slice is too short. (Let's suppose it is too short for this presentation. But, when it is not too short, the weight is high since it is a discriminative slice.)



Isoform1: FPKM~10
Isoform2: FPKM~10
isoform3: FPKM=0

Kim, et al., BMC Bioinformatics 2011, 12:305

# Summary of available datasets (series) and samples for human and mouse in different data sources, including GEO



| | Exon-array[$] | | RNA-seq[@] | |
|---|---|---|---|---|
| Organism | # Series | # Samples | # Series | # Samples |
| Human | 401 | 14,801 | 418 | 4,349 |
| Mouse | 203 | 2,565 | 376 | 3,593 |
| Total | 604 | 17,366 | 794 | 7,942 |

*[$] Exon-array platforms:* *Affymetrix Human Exon 1.0 ST Array and Affymetrix Mouse Gene 1.0 ST Array*

*[@] NGS Platforms:* *Illumina Genome Analyzer, Illumina HiSeq, AB SOLiD and 454 GS FLX*

*Data sources:* *GEO, BROAD, TCGA and ArrayExpress*

# Topics of Discussion

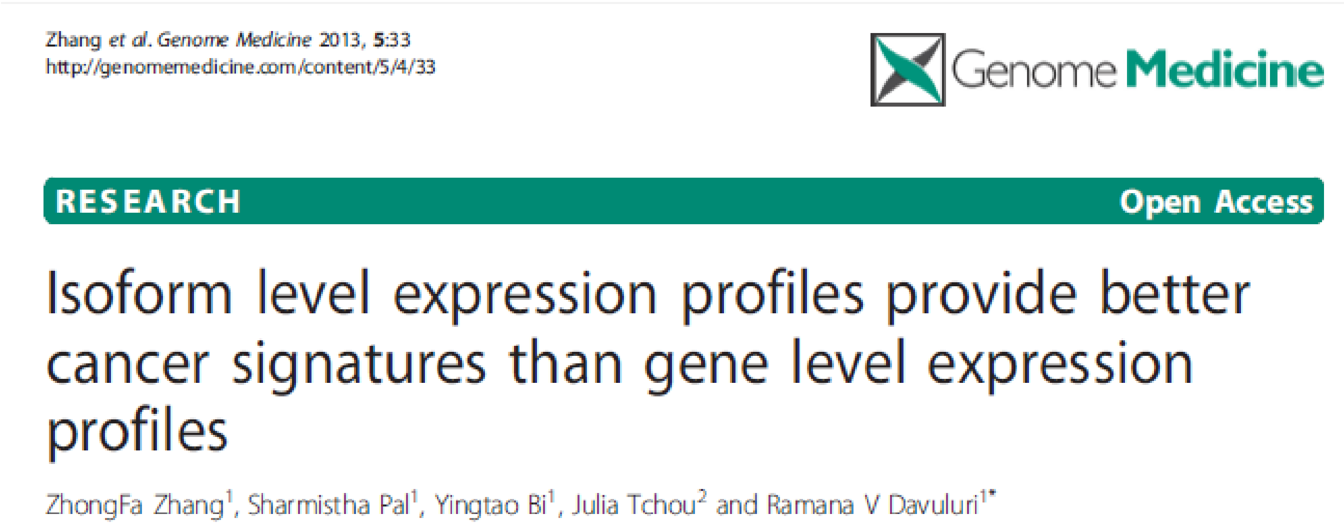
1. Grows of multi –omics data
2. Why "gene" as a unit of measure is too simplistic?
3. Exon-arrays and RNA-seq methods
4. Gene-level Vs Isoform-level analysis
   - A. Cancer Vs Non-cancer cell-line grouping
   - B. Isoform-level gene signatures for brain tumor sub-typing
5. Evaluation of isoform-level expression estimation algorithms for RNA-seq and exon-array platforms

# Cancer Vs Non-cancer cell line grouping

Cancer cell lines, regardless of their tissue of origin, can be effectively discriminated from non-cancer cell lines at isoform level, but not at gene level.

Jacob Zhang

Sharmista Pal



Zhang *et al. Genome Medicine* 2013, **5**:33
http://genomemedicine.com/content/5/4/33

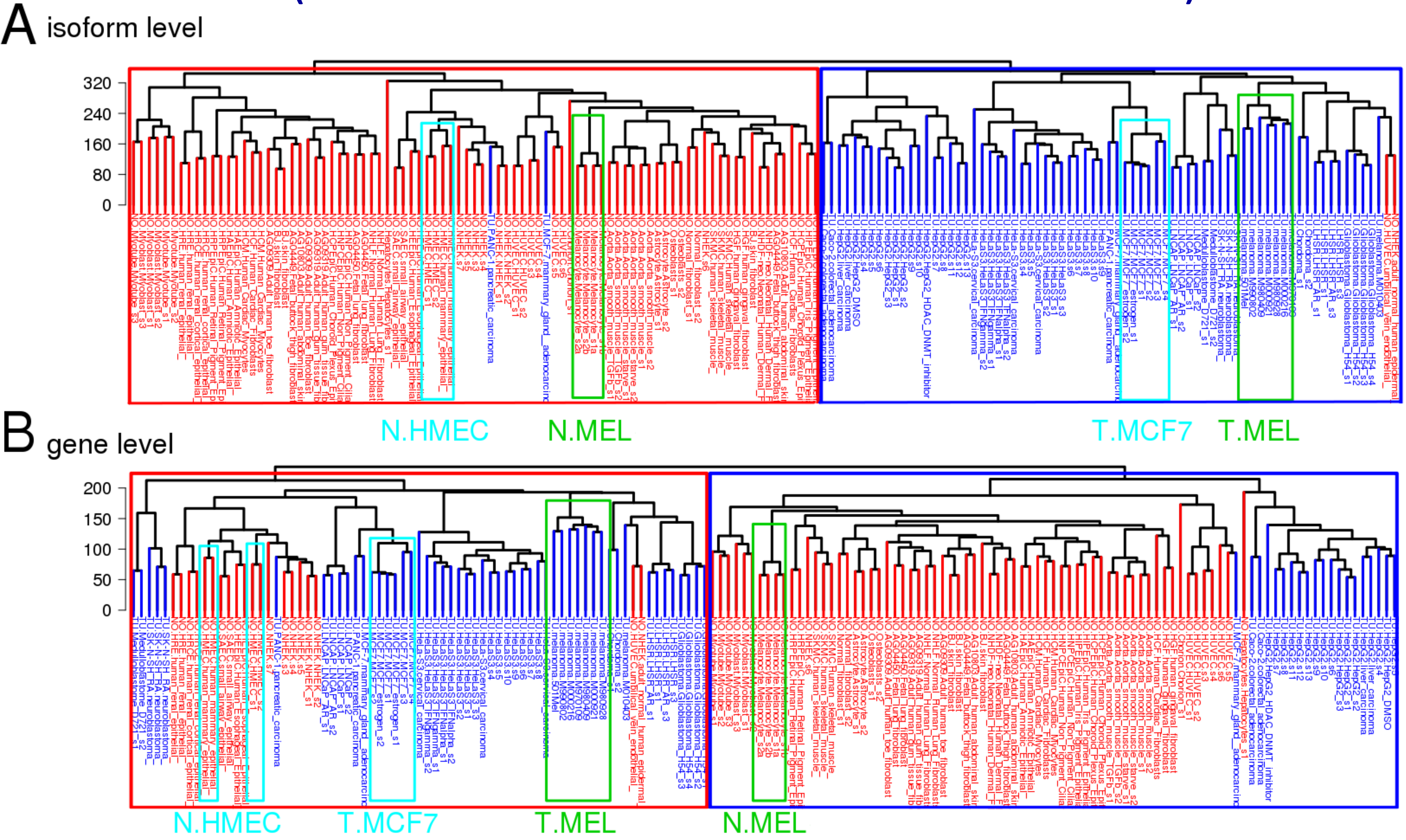
Genome Medicine

RESEARCH Open Access

Isoform level expression profiles provide better cancer signatures than gene level expression profiles

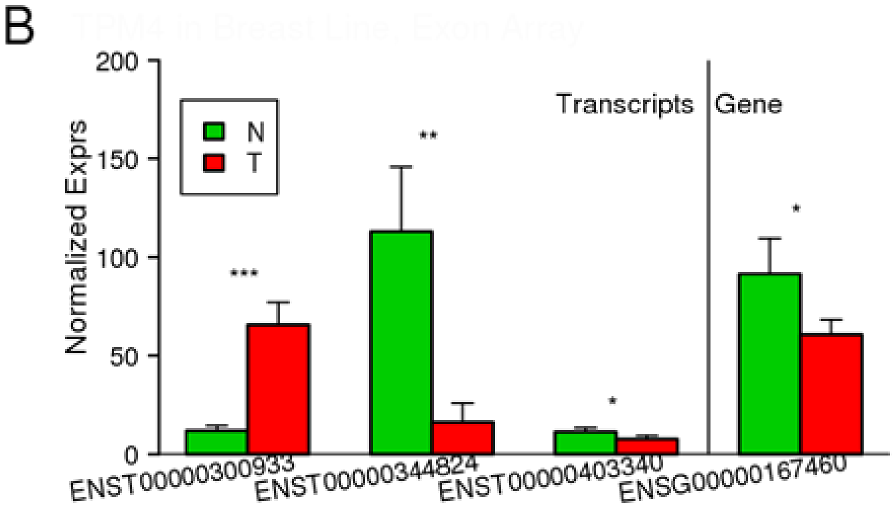
ZhongFa Zhang[1], Sharmistha Pal[1], Yingtao Bi[1], Julia Tchou[2] and Ramana V Davuluri[1*]

# Hierarchical clustering dendrograms of 160 datasets (73 cancer and 87 non-cancer cell-lines)



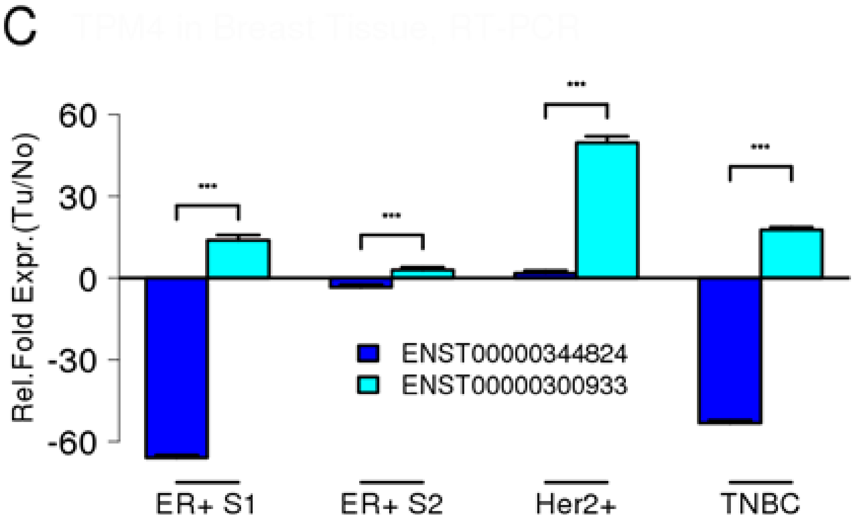
Affymetrix Human Exon 1.0 ST Array (whole-transcript GeneChip) platform, were downloaded from Gene Expression Omnibus (GEO) data depository

# Isoform-level expression profiles provide better cancer signatures than gene-level expression profiles

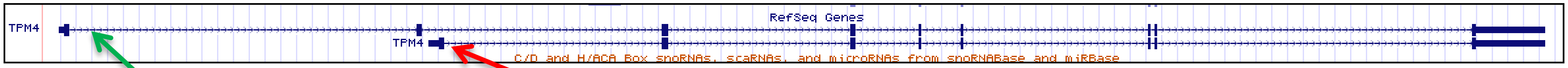


exon-array data

Mean normalized expression estimates of *TPM4* and its transcript variants in HMEC (N) and MCF7 (T) cell-lines



RT-qPCR data in breast cancer tissues


TPM4
RefSeq Genes
TPM4
C/D and H/ACA Box snoRNAs, scaRNAs, and microRNAs from snoRNABase and miRBase

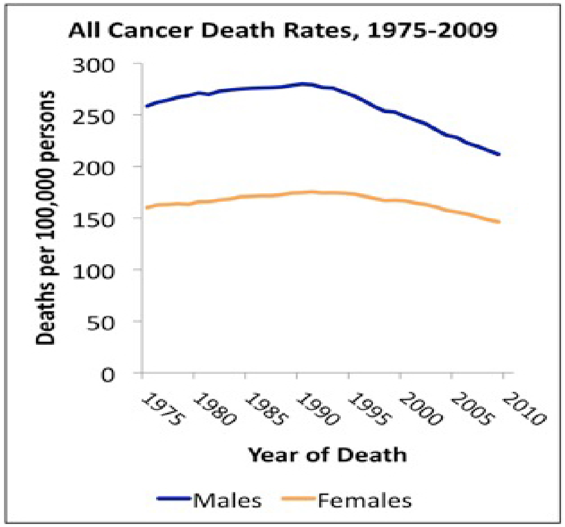
ENST00000344824

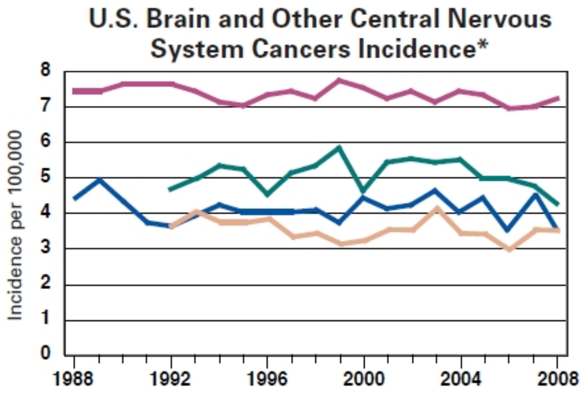
ENST00000300933

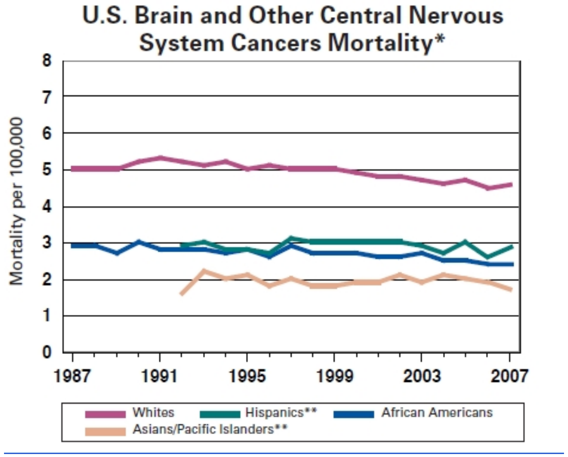
Zhang et al., Genome Medicine 2013

# Glioblastoma Multiforme (GBM) – A Deadly Brain Tumor

- **Statistics**
  - **Estimated new cases (23,130) and death (14,080) from brain and other nervous system cancer for 2013. (http://cancer.gov).**
  - **GBM accounts for 12% to 15% of all intracranial tumors and 50% to 60% of astrocytic tumors (http://www.braintumor.org)**
  - **About 9% of childhood brain tumors are glioblastomas.**
- **Incidence - annually 2 to 3 per 100,000 people (in US or Europe)**
- **Survival info**
  - **The median survival time of GBM patients is 12-14 months (Smith and Jenkins, 2000).**



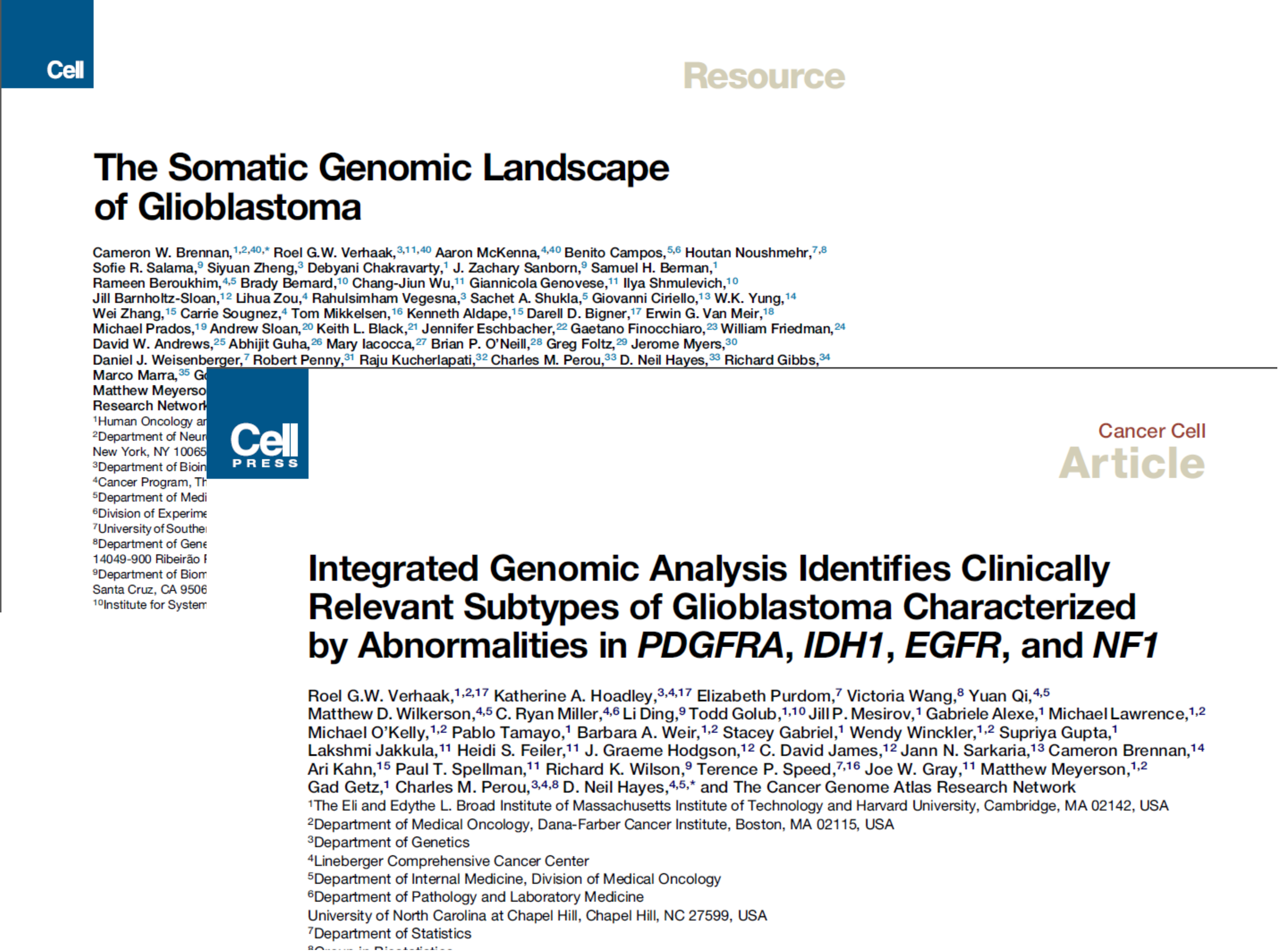




# GBM sub-typing (Gene level vs Isoform-level)

| Molecular sub-type | Number of samples (n) | | |
|---|---|---|---|
| | Core | Other | Total |
| Classical (C) | 37 | - | 173 |
| Mesenchymal (M) | 55 | - | |
| Neural (N) | 27 | - | |
| Proneural (PN) | 54 | - | |
| Other GBM (subtype not known) | | 246 | 246 |
| Total GBM samples | | | 419 |
| Normal brain | | 10 | 10 |

- Verhaak et. al. (Cancer Cell 2010): Classified GBM into 4 groups- Proneural (PN), Neural (N), Mesenchymal (M), And Classical (CL). **Identified a 840 gene based signature, uses 210 genes per class**.



Cell

Resource

# The Somatic Genomic Landscape of Glioblastoma

Cameron W. Brennan,[1,2,40,*] Roel G.W. Verhaak,[3,11,40] Aaron McKenna,[4,40] Benito Campos,[5,6] Houtan Noushmehr,[7,8]
Sofie R. Salama,[9] Siyuan Zheng,[3] Debyani Chakravarty,[1] J. Zachary Sanborn,[9] Samuel H. Berman,[1]
Rameen Beroukhim,[4,5] Brady Bernard,[10] Chang-Jiun Wu,[11] Giannicola Genovese,[11] Ilya Shmulevich,[10]
Jill Barnholtz-Sloan,[12] Lihua Zou,[4] Rahulsimham Vegesna,[3] Sachet A. Shukla,[5] Giovanni Ciriello,[13] W.K. Yung,[14]
Wei Zhang,[15] Carrie Sougnez,[4] Tom Mikkelsen,[16] Kenneth Aldape,[15] Darell D. Bigner,[17] Erwin G. Van Meir,[18]
Michael Prados,[19] Andrew Sloan,[20] Keith L. Black,[21] Jennifer Eschbacher,[22] Gaetano Finocchiaro,[23] William Friedman,[24]
David W. Andrews,[25] Abhijit Guha,[26] Mary Iacocca,[27] Brian P. O'Neill,[28] Greg Foltz,[29] Jerome Myers,[30]
Daniel J. Weisenberger,[7] Robert Penny,[31] Raju Kucherlapati,[32] Charles M. Perou,[33] D. Neil Hayes,[33] Richard Gibbs,[34]
Marco Marra,[35] G
Matthew Meyerso
Research Network
[1]Human Oncology ar
[2]Department of Neur
New York, NY 10065
[3]Department of Bioin
[4]Cancer Program, Th
[5]Department of Medi
[6]Division of Experime
[7]University of Southe
[8]Department of Gene
14049-900 Ribeirão F
[9]Department of Biom
Santa Cruz, CA 9506
[10]Institute for System

Cell PRESS

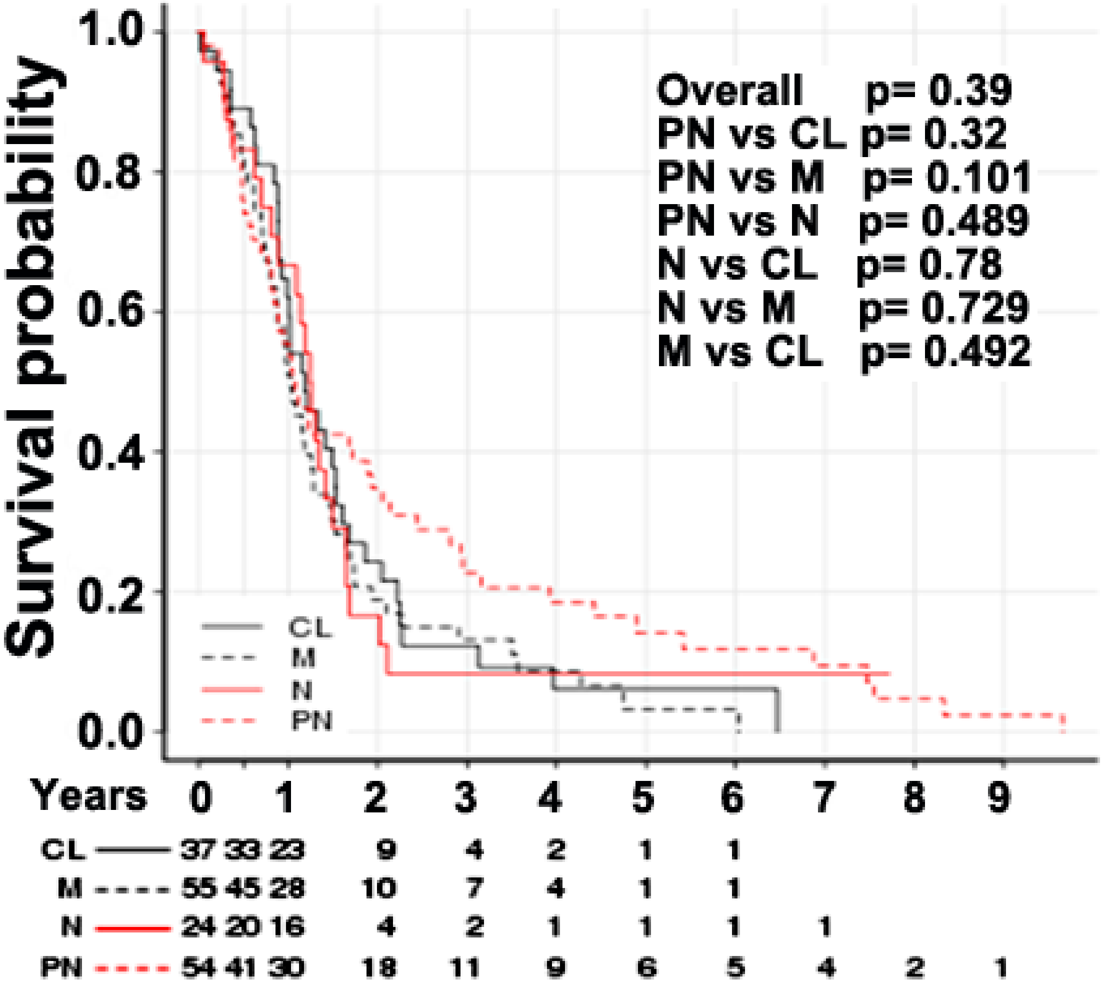
Cancer Cell

Article

# Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*

Roel G.W. Verhaak,[1,2,17] Katherine A. Hoadley,[3,4,17] Elizabeth Purdom,[7] Victoria Wang,[8] Yuan Qi,[4,5]
Matthew D. Wilkerson,[4,5] C. Ryan Miller,[4,6] Li Ding,[9] Todd Golub,[1,10] Jill P. Mesirov,[1] Gabriele Alexe,[1] Michael Lawrence,[1,2]
Michael O'Kelly,[1,2] Pablo Tamayo,[1] Barbara A. Weir,[1,2] Stacey Gabriel,[1] Wendy Winckler,[1,2] Supriya Gupta,[1]
Lakshmi Jakkula,[11] Heidi S. Feiler,[11] J. Graeme Hodgson,[12] C. David James,[12] Jann N. Sarkaria,[13] Cameron Brennan,[14]
Ari Kahn,[15] Paul T. Spellman,[11] Richard K. Wilson,[9] Terence P. Speed,[7,16] Joe W. Gray,[11] Matthew Meyerson,[1,2]
Gad Getz,[1] Charles M. Perou,[3,4,8] D. Neil Hayes,[4,5,*] and The Cancer Genome Atlas Research Network
[1]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA 02142, USA
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[3]Department of Genetics
[4]Lineberger Comprehensive Cancer Center
[5]Department of Internal Medicine, Division of Medical Oncology
[6]Department of Pathology and Laboratory Medicine
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[7]Department of Statistics
[8]Group in Biostatistics

# TCGA classification has no prognostic significance
# GBM patients (173 core group) into 4 groups



Verhaak et. al. (Cancer Cell 2010):

# PIGExClass – Platform-independent Isoform-level Gene-Expression based Classification-system

Pal & Bi *et al*. *Nucleic Acids Res*. 2014

Yingtao Bi, Ph.D.
Staff Scientist
(Statistics, UC Riverside)

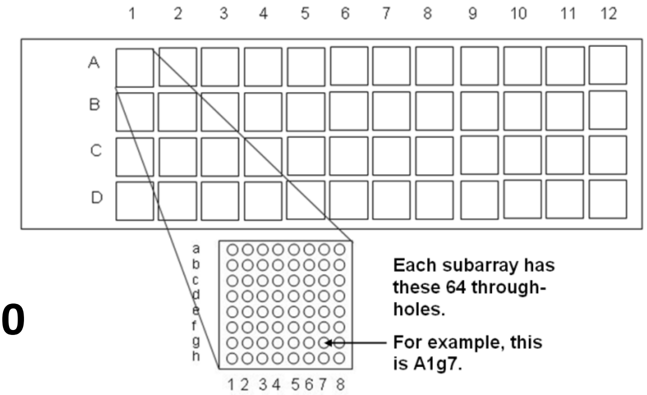
Sharmista Pal, Ph.D.
Staff Scientist
(Mol Bio., OSU, Columbus)

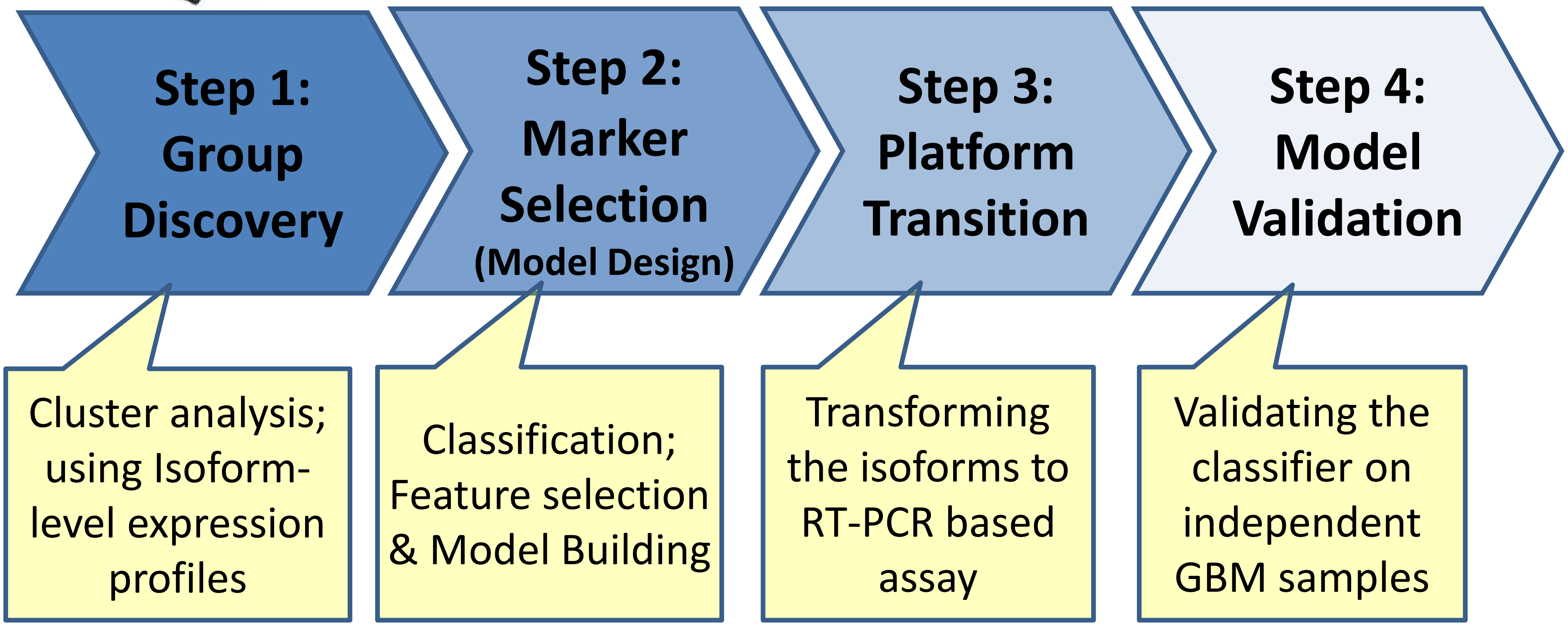
**Samples** x **Genes/Transcripts Matrix**

$$\begin{pmatrix} X_{ij} \end{pmatrix} \rightarrow \begin{pmatrix} Y_{ij} \end{pmatrix}$$

400 x **115,000**   400 x **200**

**OpenArray® RT-qPCR Platform**


1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
a b c d e f g h
1 2 3 4 5 6 7 8
Each subarray has these 64 through-holes.
For example, this is A1g7.



**Step 1: Group Discovery**

Cluster analysis; using Isoform-level expression profiles

**Step 2: Marker Selection (Model Design)**

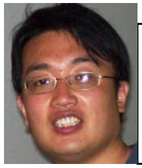
Classification; Feature selection & Model Building

**Step 3: Platform Transition**

Transforming the isoforms to RT-PCR based assay

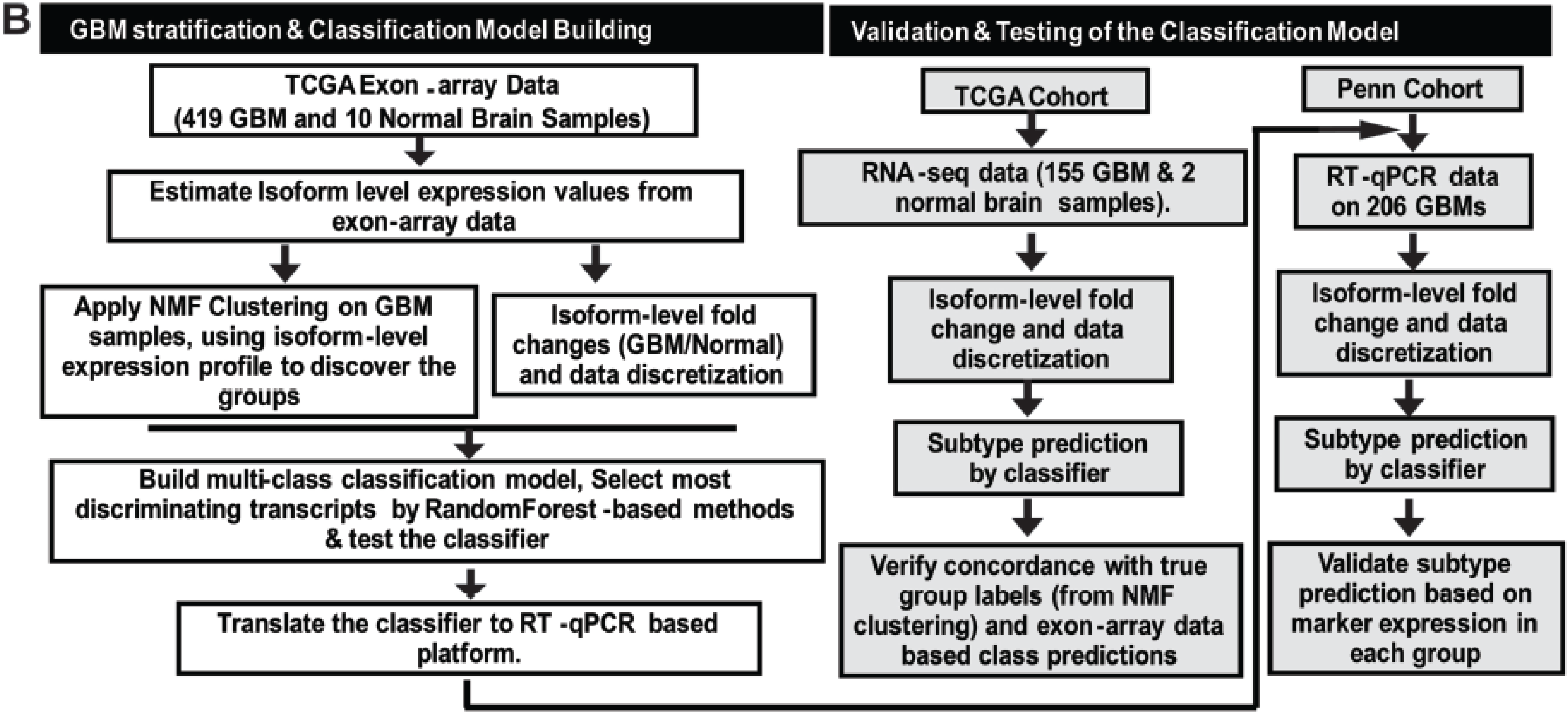
**Step 4: Model Validation**

Validating the classifier on independent GBM samples



Yingtao Bi, Ph.D.
Postdoctoral fellow
(Statistics, UC Riverside)

Sharmista Pal, Ph.D.
Staff Scientist
(Mol Bio., OSU, Columbus)



B

GBM stratification & Classification Model Building

TCGA Exon .array Data (419 GBM and 10 Normal Brain Samples)

Estimate Isoform level expression values from exon-array data

Apply NMF Clustering on GBM samples, using isoform-level expression profile to discover the groups

Isoform-level fold changes (GBM/Normal) and data discretization

Build multi-class classification model, Select most discriminating transcripts by RandomForest -based methods & test the classifier

Translate the classifier to RT -qPCR based platform.

Validation & Testing of the Classification Model

TCGA Cohort

RNA-seq data (155 GBM & 2 normal brain samples).

Isoform-level fold change and data discretization

Subtype prediction by classifier

Verify concordance with true group labels (from NMF clustering) and exon-array data based class predictions

Penn Cohort

RT-qPCR data on 206 GBMs

Isoform-level fold change and data discretization

Subtype prediction by classifier

Validate subtype prediction based on marker expression in each group

Pal & Bi *et al. Nucleic Acids Res*. 2014

# TCGA datasets analyzed by our group

| Sample type | Data-type | Number of samples |
|---|---|---|
| Normal brain (control samples) | Gene expression (exon-array data) | 10 |
| GBM tumor | Gene expression (exon-array data) | 419 |
| GBM tumor | Gene expression (RNA-seq) | 169 |
| GBM tumor | Exome sequencing | 323 |
| GBM matched blood | Exome sequencing | 259 |
| LGG tumor | Gene expression (RNA-seq) | ?? |
| LGG tumor | Exome sequencing | 180 |
| LGG matched blood | Exome sequencing | 160 |

76 common

https://tcga-data.nci.nih.gov/

# Gene-level and Isoform-level analysis of transcriptome changes

| TCGA Exon-array Data Analysis (q≤0.001 and fold-change ≥2.0) | | |
|---|---|---|
| | Gene-level | Isoform (transcript variant)-level |
| Upregulated | 912 | 2085 |
| Downregulated | 1922 | 5228 |

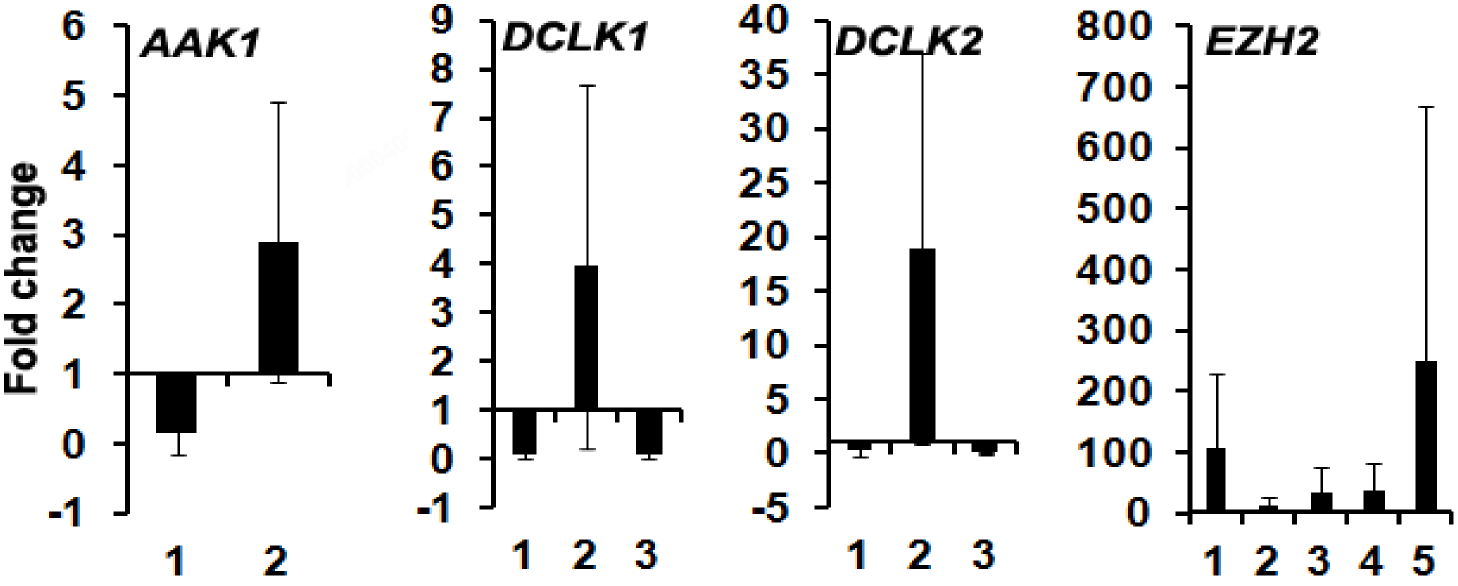
| symbol | FC |
|---|---|
| AAK1 | -2.09 |
| DCLK1 | -2.49 |
| DCLK3 | -2.01 |

Gene-level fold changes

| symbol | FC |
|---|---|
| AAK1-001 | -6.77 |
| AAK1-004 | -2.62 |
| AAK1-011 | 3.52 |
| DCLK1-001 | 3.17 |
| DCLK1-006 | -5.04 |
| DCLK1-013 | -2.47 |
| DCLK1-201 | -5.66 |
| DCLK2-201 | 7.31 |
| DCLK2-202 | -3.52 |
| DCLK3-001 | -2.15 |

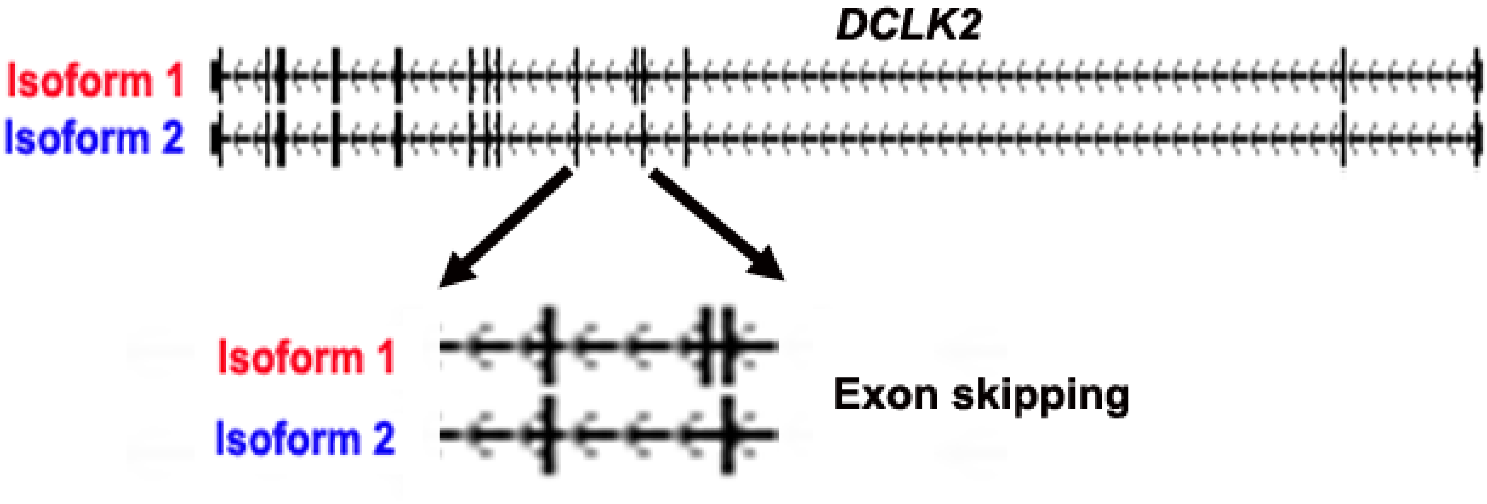
Isoform-level fold changes

# Validation in independent brain tumor cohort (UPenn Neurosurgery Dept)



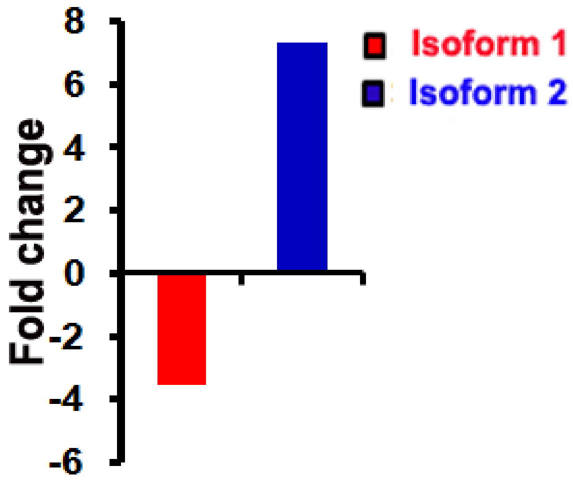
- Validated the isoform-level expression changes by RT-qPCR in primary GBM samples for 15 of 16 isoform transcripts corresponding to 6 genes

# An example showing isoform specific dysregulation

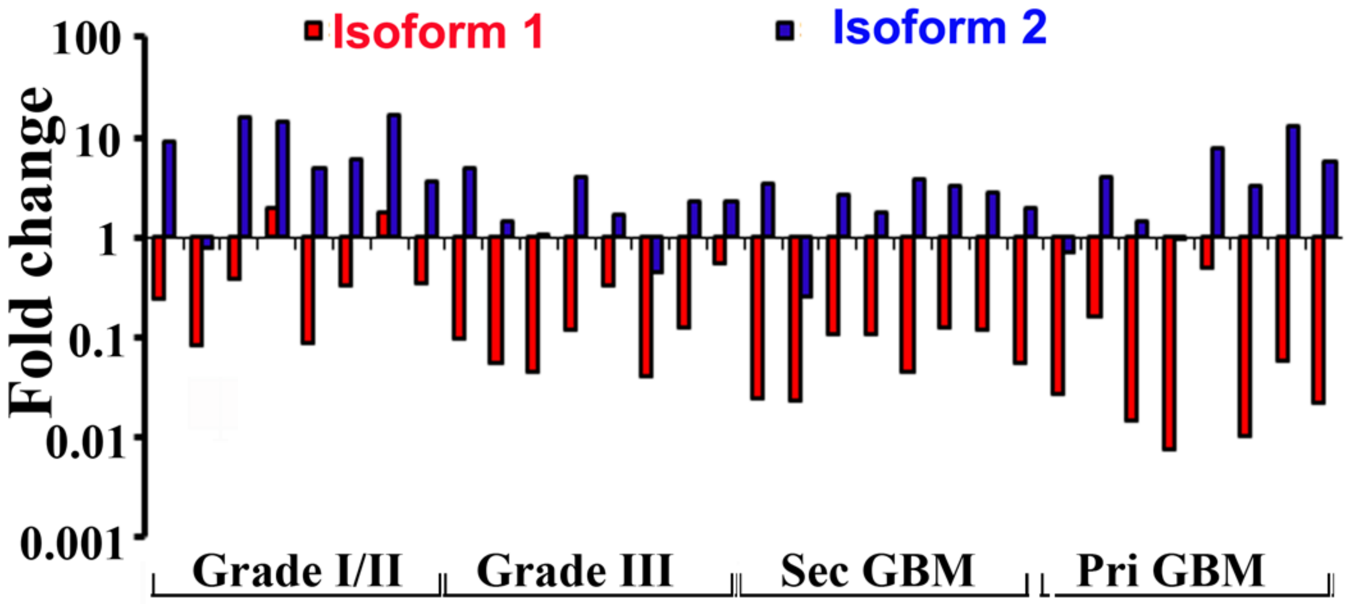


# *DCLK2* isoforms show opposite patterns of expression in gliomas versus normal brain

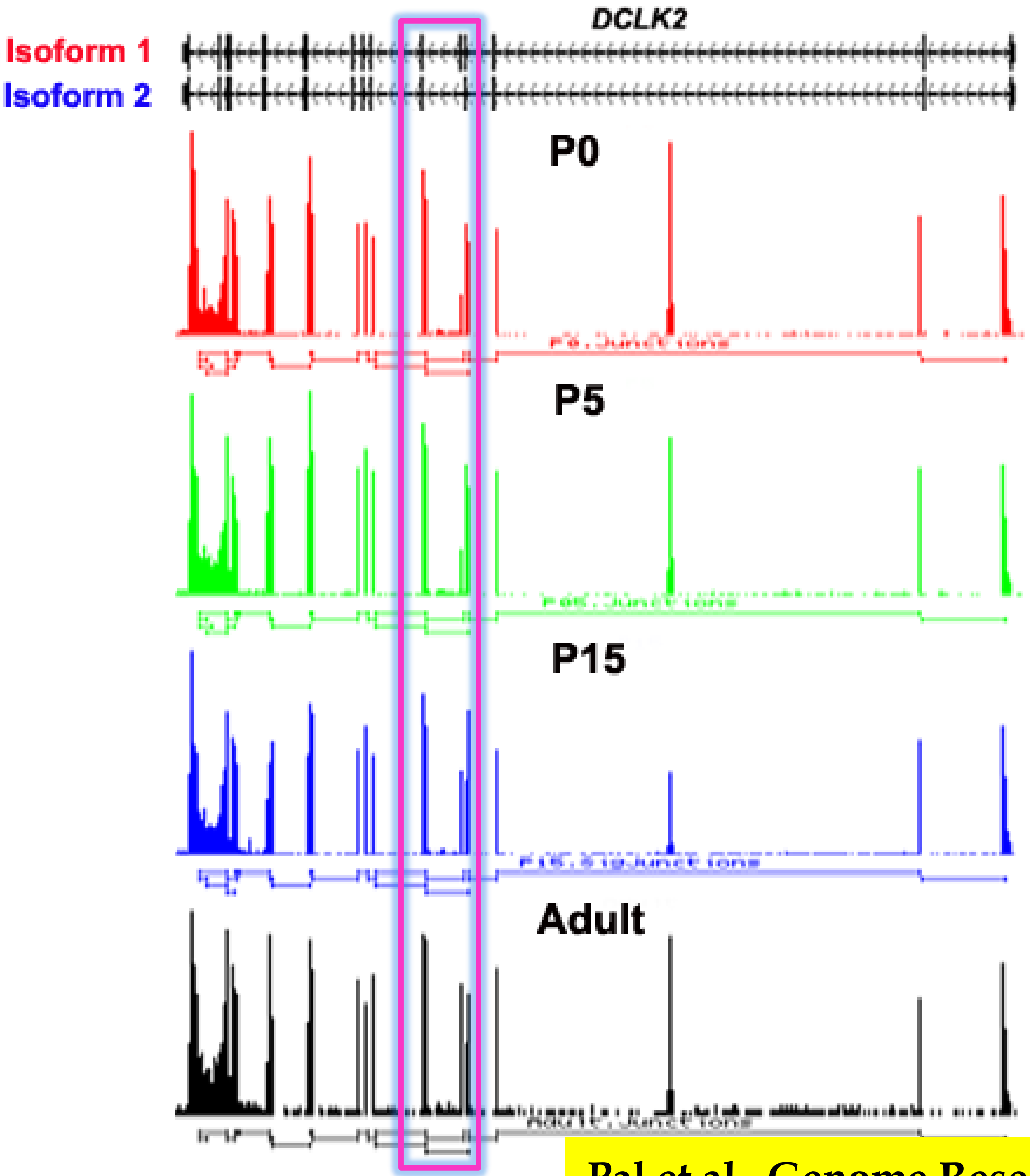
Analysis of TCGA data

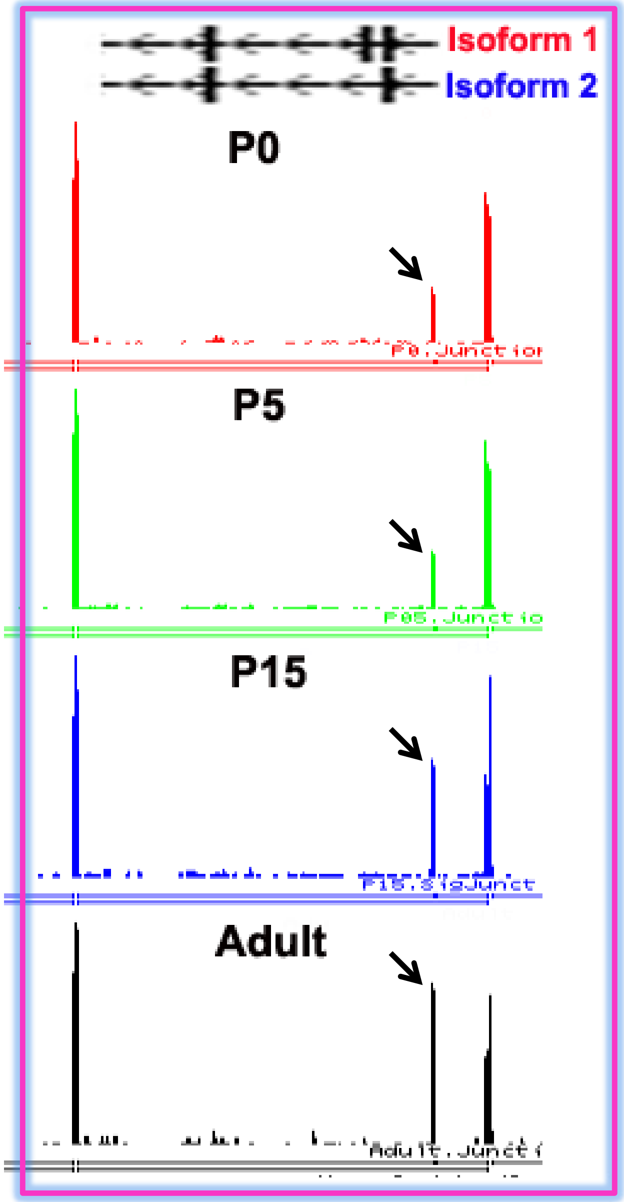


Validation in independent cohort of gliomas



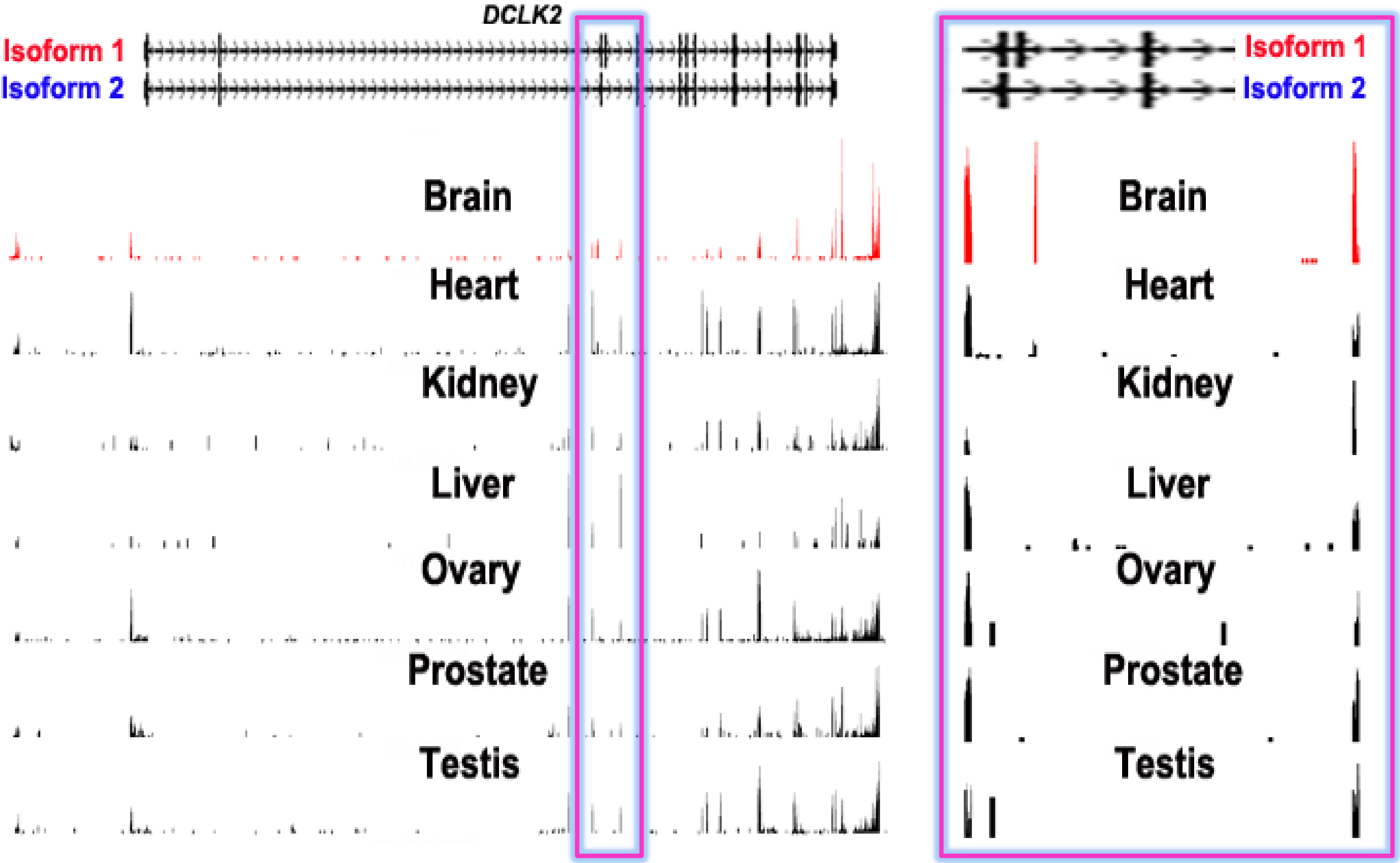
# *Dclk2* isoforms are developmentally regulated





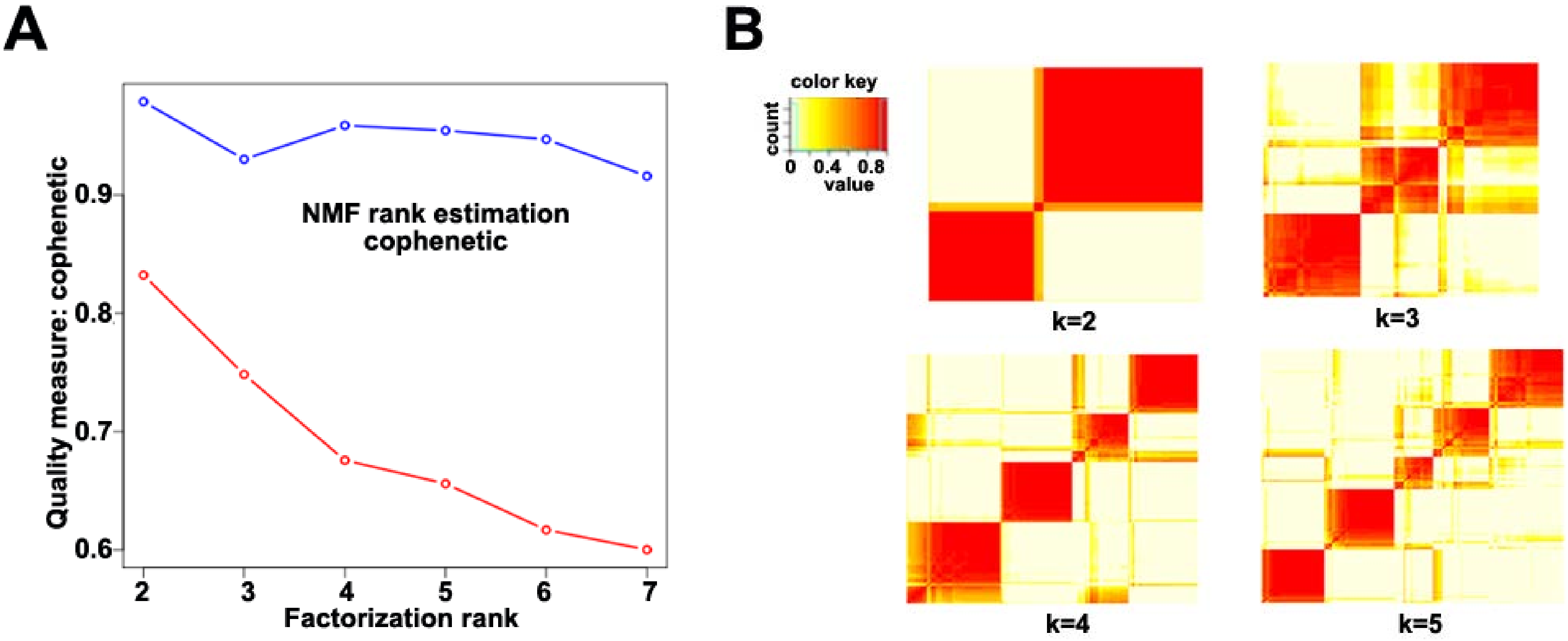
Pal et al., Genome Research 2011

# *DCLK2* isoform 1 is tissue specific in humans



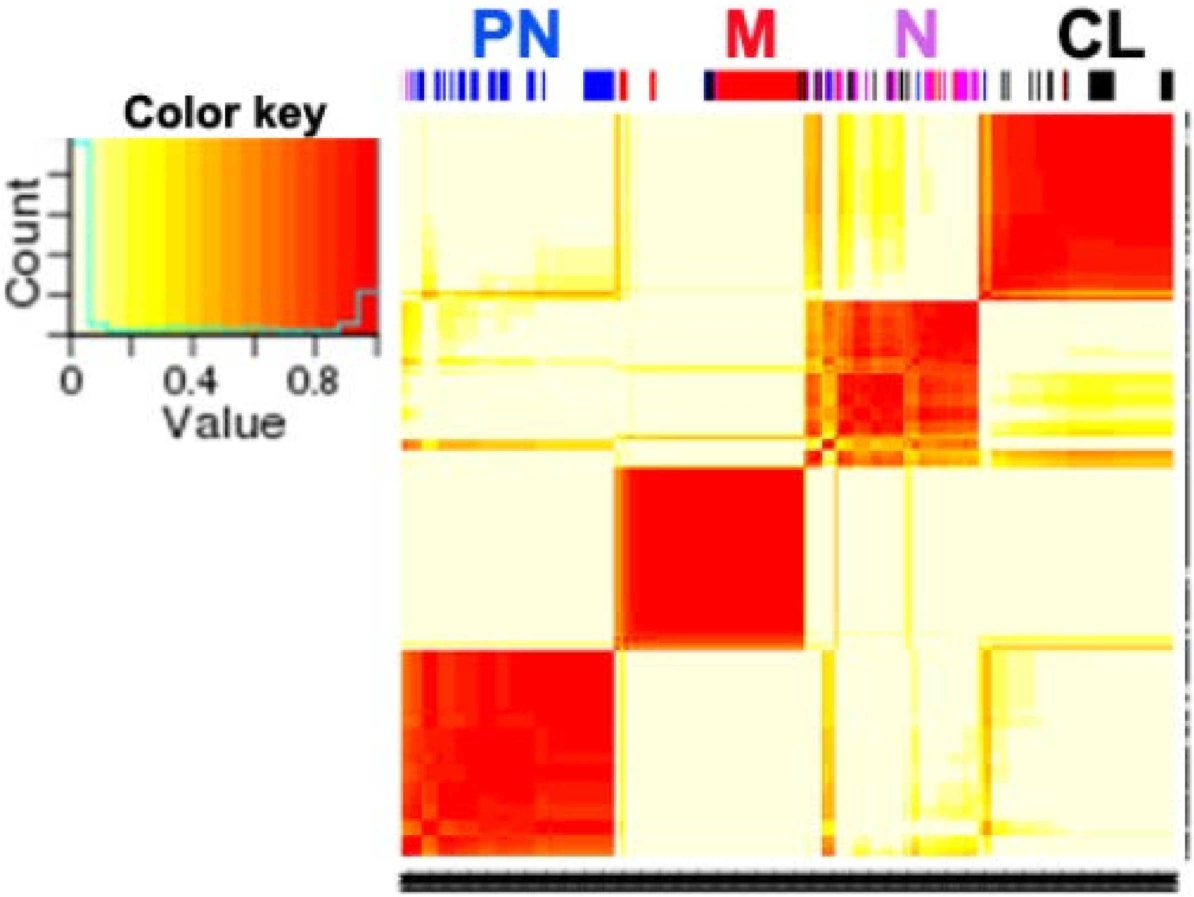
- DCLK2 isoform 1, which is brain specific and expressed higher in adult brain than in early development is down-regulated in cancer (GBM)

# Stable clustering at isoform-level can be achieved in four groups



- Data matrix – isoform expression data of 197 (or 419) samples and 1600 isoforms
- Consensus non-negative matrix factorization (NMF) clustering method
- Silhouette width was computed to filter out samples that were included in a subclass, but that were not a robust representative of the subclass

# NMF clustering of 419 GBM patient samples based on the expression of 1,600 of the most variable isoforms across the patients



| | | |
|---|---|---|
| PN | 54 | 173 |
| M | 55 | |
| N | 27 | |
| CL | 37 | |

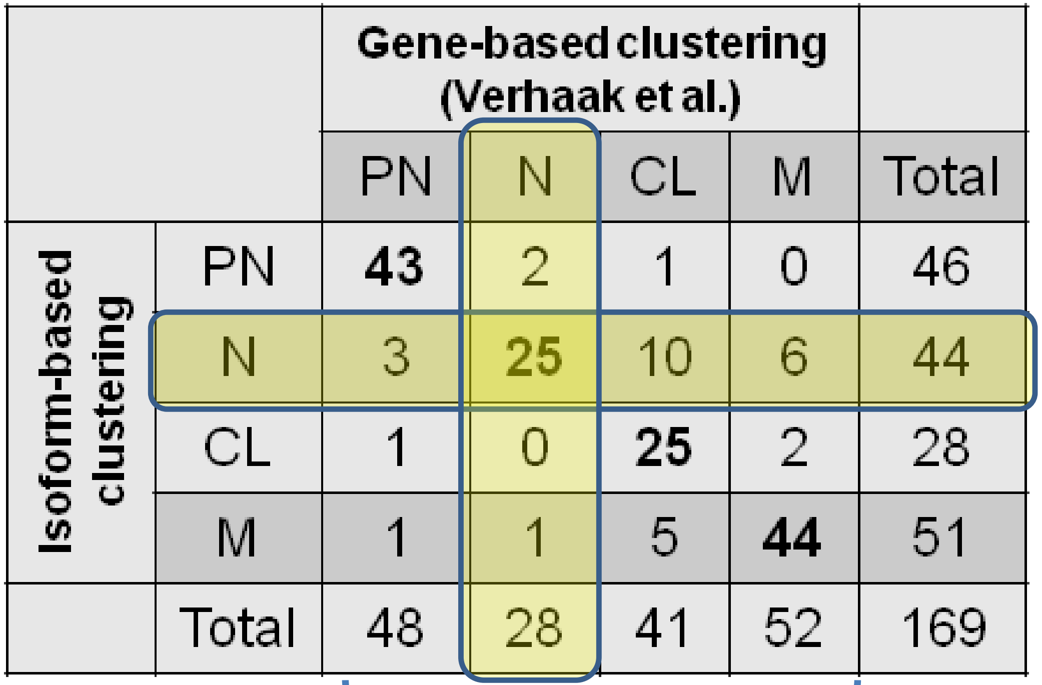
TCGA core samples

95 (PN) 85 (M) 75 (N) 87 (CL)

342

A total of **342** as most representative of the four groups, "**isoform-based core samples**"

# Concordance in cluster membership calls between our isoform-based and gene-based groupings in the TCGA publication

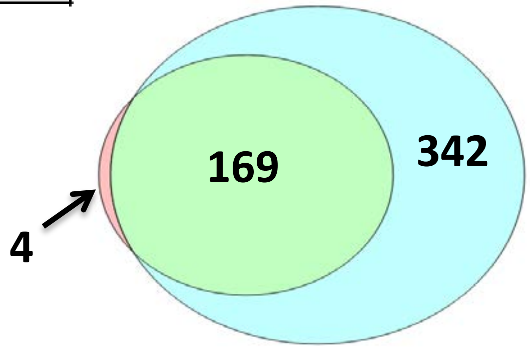


| | | Gene-based clustering (Verhaak et al.) | | | | |
|---|---|---|---|---|---|---|
| | | PN | N | CL | M | Total |
| Isoform-based clustering | PN | **43** | 2 | 1 | 0 | 46 |
| | N | 3 | **25** | 10 | 6 | 44 |
| | CL | 1 | 0 | **25** | 2 | 28 |
| | M | 1 | 1 | 5 | **44** | 51 |
| | Total | 48 | 28 | 41 | 52 | 169 |

**Isoform-based clustering (Our Grouping)**

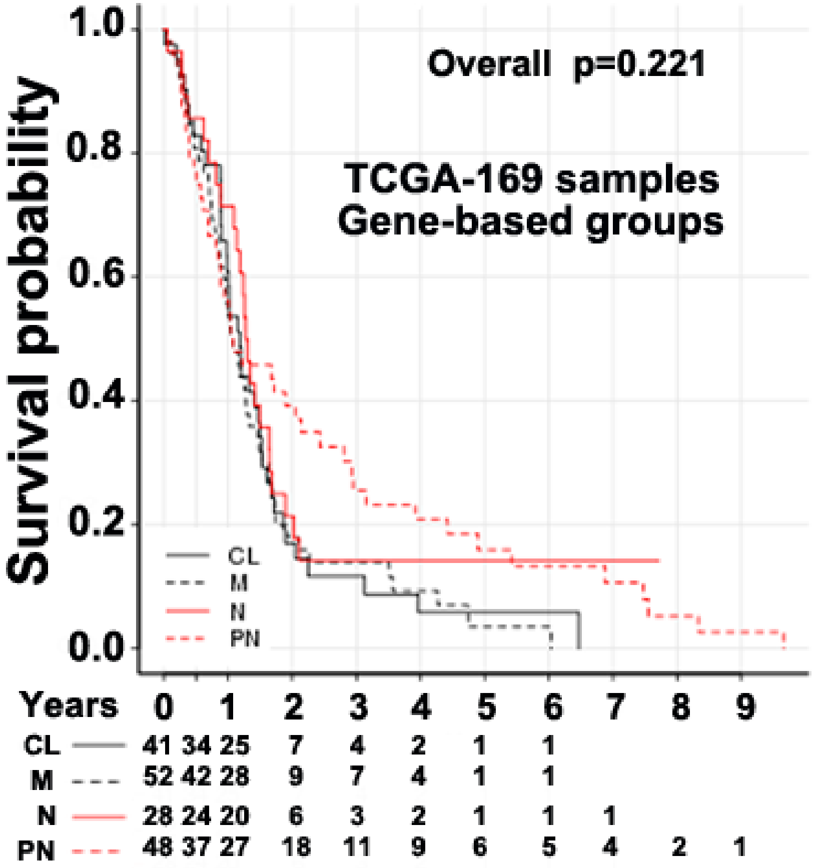
**Gene-based clustering (Verhaak et al Grouping)**

32 (~20%) were reassigned to a different subgroup by our isoform-based signature.



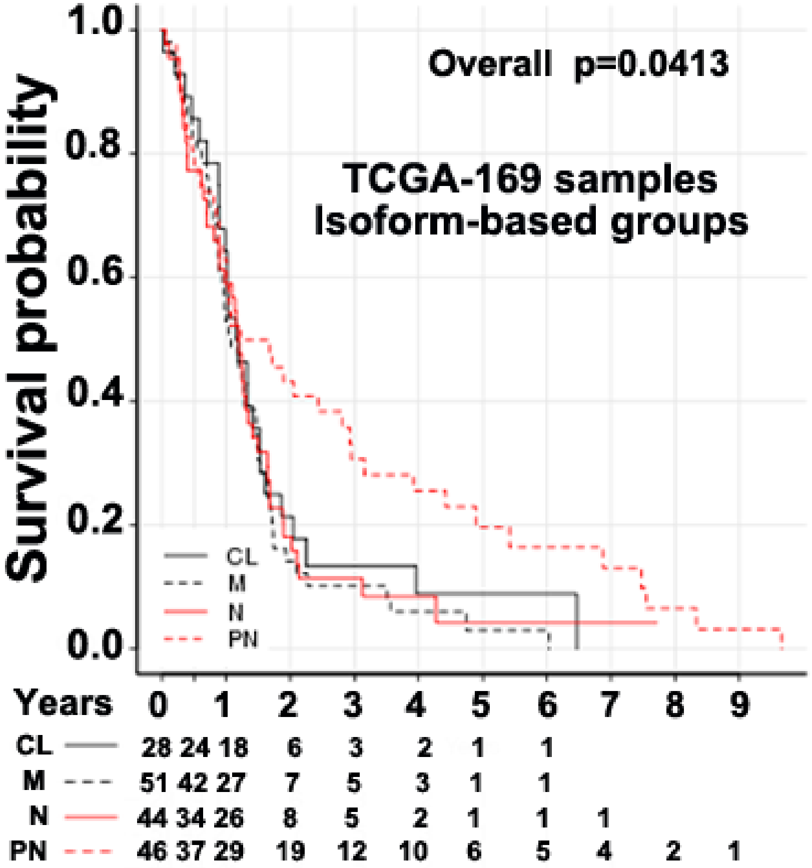
**Overlap of Our & TCGA Core Samples**

# Survival plots of gene vs isoform-level grouping of 169 samples



| Years | 0 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CL | 41 | 34 | 25 | 7 | 4 | 2 | 1 | 1 | | | |
| M | 52 | 42 | 28 | 9 | 7 | 4 | 1 | 1 | | | |
| N | 28 | 24 | 20 | 6 | 3 | 2 | 1 | 1 | 1 | | |
| PN | 48 | 37 | 27 | 18 | 11 | 9 | 6 | 5 | 4 | 2 | 1 |

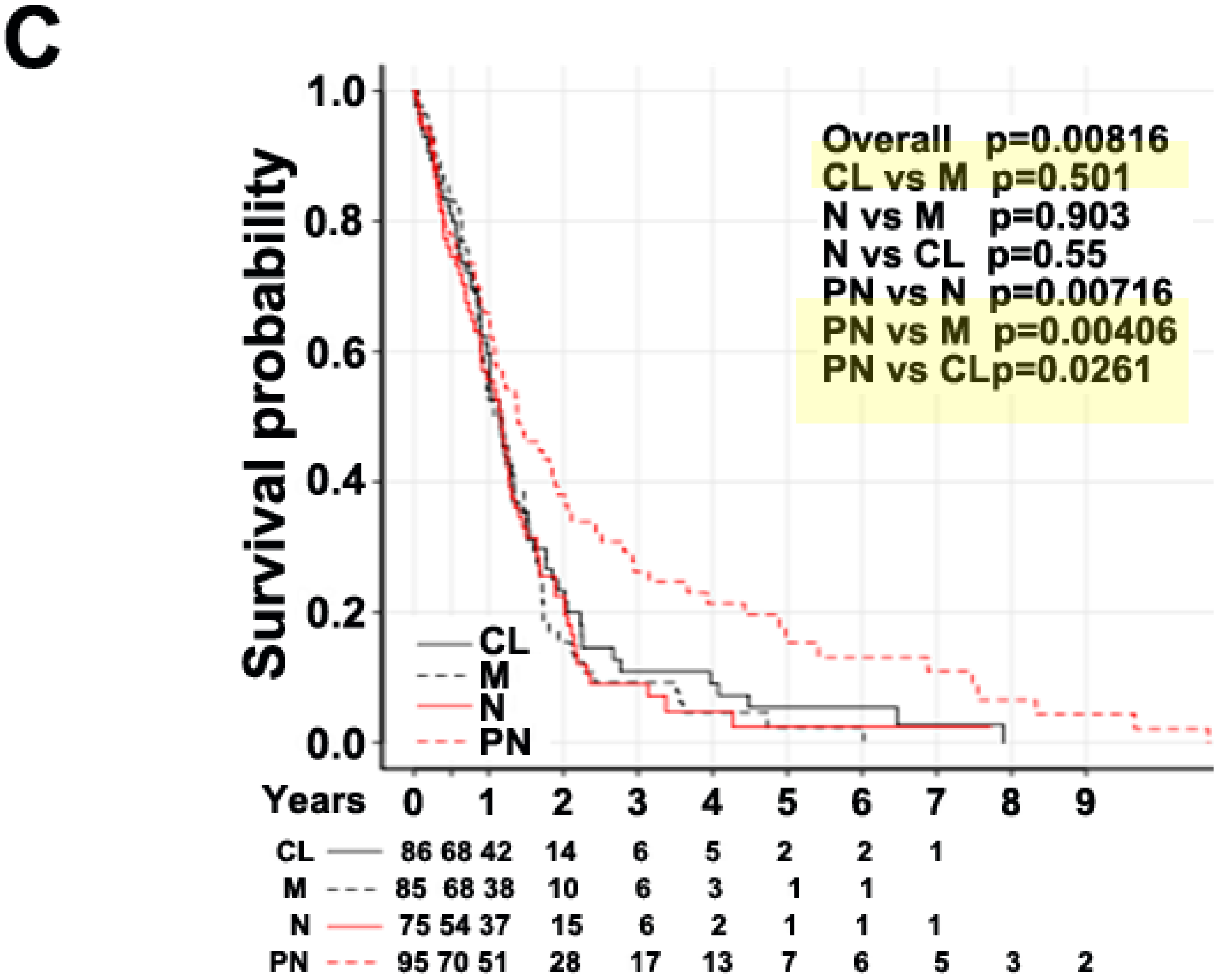
Gene-based clustering of 169 samples
(Verhaak et al Grouping)



| Years | 0 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CL | 28 | 24 | 18 | 6 | 3 | 2 | 1 | 1 | | | |
| M | 51 | 42 | 27 | 7 | 5 | 3 | 1 | 1 | | | |
| N | 44 | 34 | 26 | 8 | 5 | 2 | 1 | 1 | 1 | | |
| PN | 46 | 37 | 29 | 19 | 12 | 10 | 6 | 5 | 4 | 2 | 1 |

Isoform-based clustering of 169 samples
(Our Grouping)

# Survival plot for the four groups based on isoform-level clustering



| Years | 0 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CL —— | 86 | 68 | 42 | 14 | 6 | 5 | 2 | 2 | 1 | | |
| M - - - | 85 | 68 | 38 | 10 | 6 | 3 | 1 | 1 | | | |
| N —— | 75 | 54 | 37 | 15 | 6 | 2 | 1 | 1 | 1 | | |
| PN - - - | 95 | 70 | 51 | 28 | 17 | 13 | 7 | 6 | 5 | 3 | 2 |

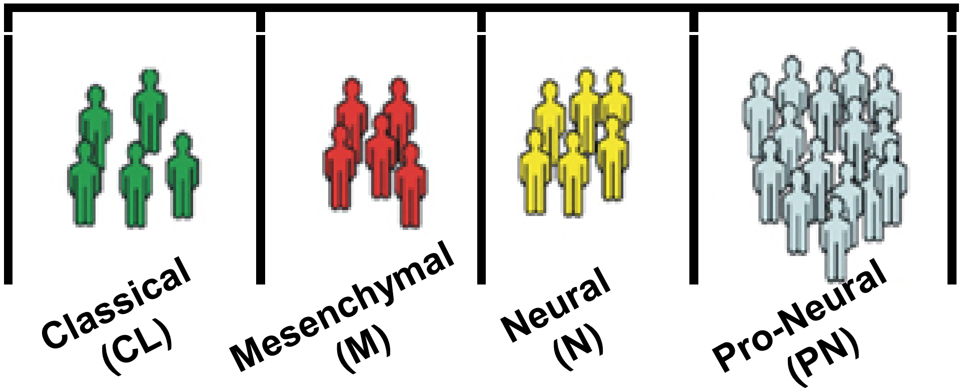
Isoform-based clustering of 341 core-samples
(Our Grouping)

# Brain tumor sub-typing → Precision Medicine

**Predictive classifiers – composite gene signatures as biomarkers**

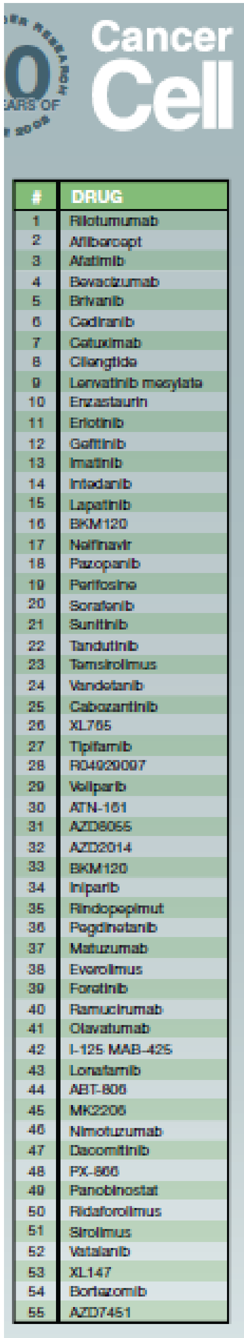
GBM patient group

**Isoform-level classifier** for GBM patient stratification



A diagnostic assay to predict the molecular subtype of a future GBM patient is currently lacking

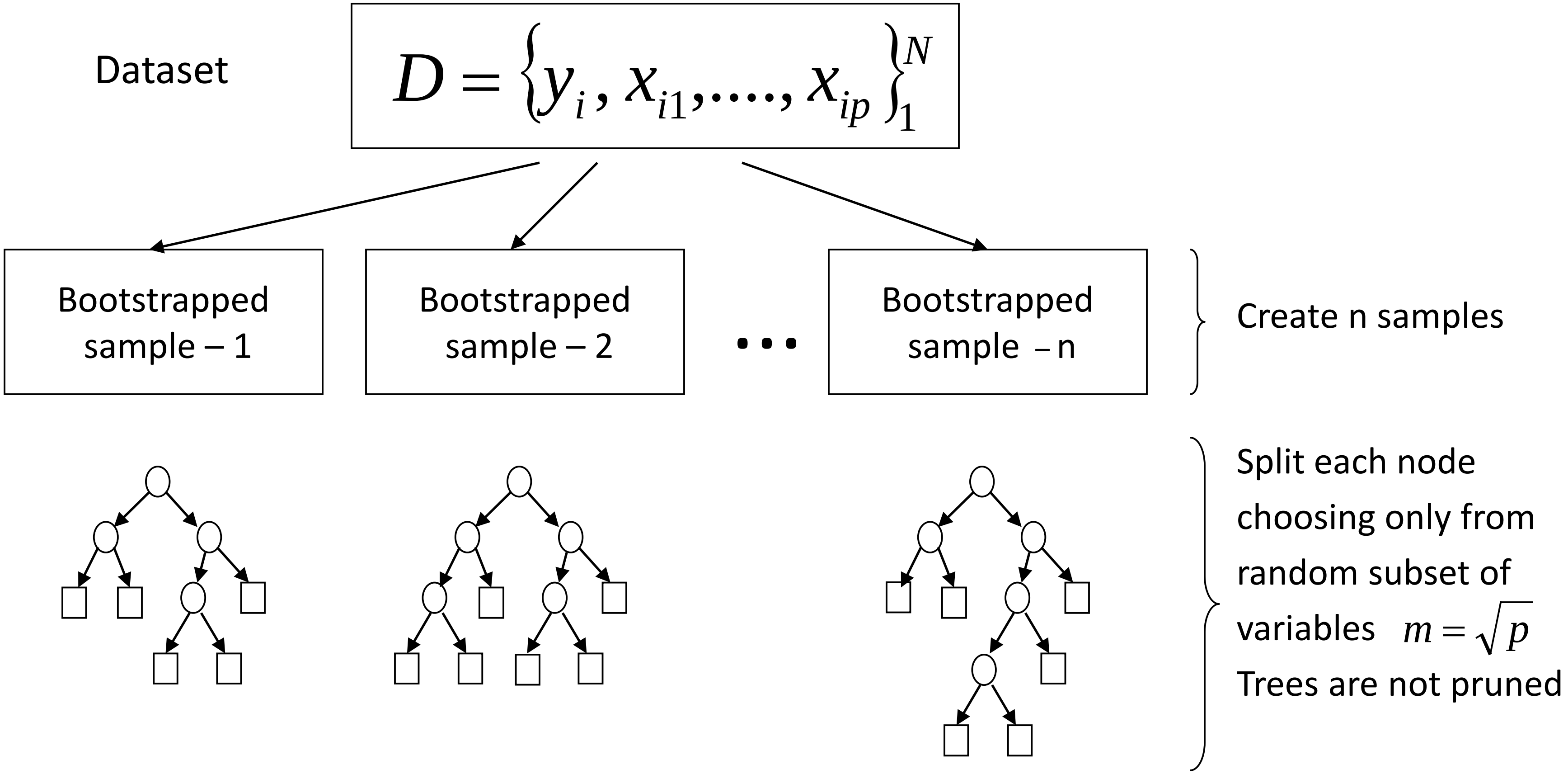
Kotliarova & Fine (2012) SnapShot: glioblastoma multiforme. ***Cancer Cell.***



Cancer Cell

| # | DRUG |
|---|---|
| 1 | Rilotumumab |
| 2 | Aflibercept |
| 3 | Afatinib |
| 4 | Bevacizumab |
| 5 | Brivanib |
| 6 | Cediranib |
| 7 | Cetuximab |
| 8 | Cilengtide |
| 9 | Lenvatinib mesylate |
| 10 | Enzastaurin |
| 11 | Erlotinib |
| 12 | Gefitinib |
| 13 | Imatinib |
| 14 | Intedanib |
| 15 | Lapatinib |
| 16 | BKM120 |
| 17 | Nelfinavir |
| 18 | Pazopanib |
| 19 | Perifosine |
| 20 | Sorafenib |
| 21 | Sunitinib |
| 22 | Tandutinib |
| 23 | Temsirolimus |
| 24 | Vandetanib |
| 25 | Cabozantinib |
| 26 | XL765 |
| 27 | Tipifarnib |
| 28 | R04929097 |
| 29 | Veliparib |
| 30 | ATN-161 |
| 31 | AZD8055 |
| 32 | AZD2014 |
| 33 | BKM120 |
| 34 | Iniparib |
| 35 | Rindopepimut |
| 36 | Pegdinetanib |
| 37 | Matuzumab |
| 38 | Everolimus |
| 39 | Foretinib |
| 40 | Ramucirumab |
| 41 | Olaratumab |
| 42 | I-125 MAB-425 |
| 43 | Lonafarnib |
| 44 | ABT-806 |
| 45 | MK2206 |
| 46 | Nimotuzumab |
| 47 | Dacomitinib |
| 48 | PX-866 |
| 49 | Panobinostat |
| 50 | Ridaforolimus |
| 51 | Sirolimus |
| 52 | Vatalanib |
| 53 | XL147 |
| 54 | Bortezomib |
| 55 | AZD7451 |

# Feature Selection & Classification: RandomForest



Dataset

$$D = \{y_i, x_{i1}, \dots, x_{ip}\}_1^N$$

Bootstrapped sample – 1

Bootstrapped sample – 2

…

Bootstrapped sample – n

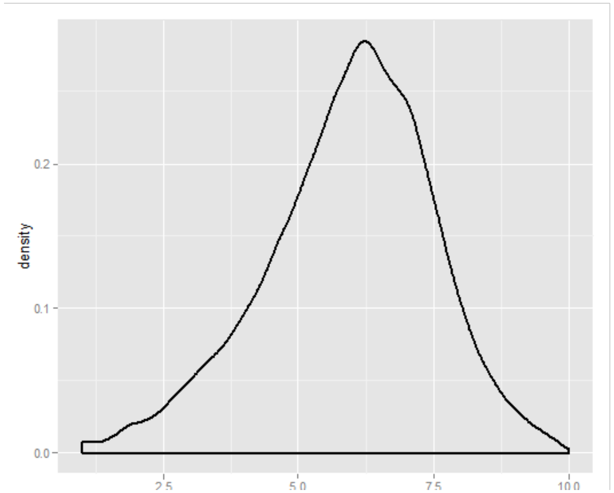
Create n samples

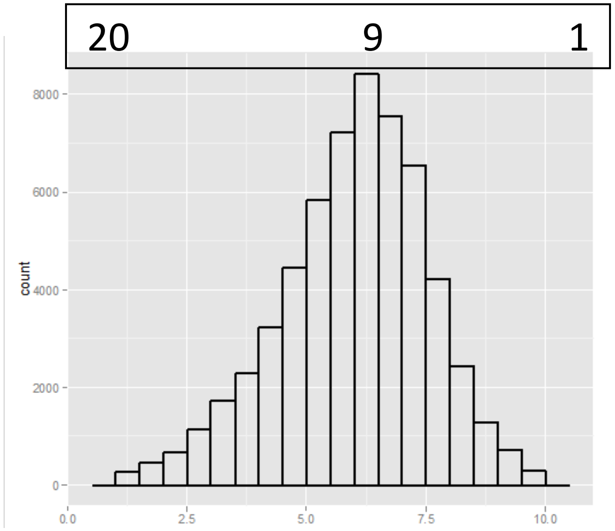
Split each node choosing only from random subset of variables $m = \sqrt{p}$

Trees are not pruned

- The majority vote of the trees determines the classification result of an observation.
- An estimate of the classification error is supplied by the out-of-bag sample

# Platform Transition: Converting FCs to discrete values





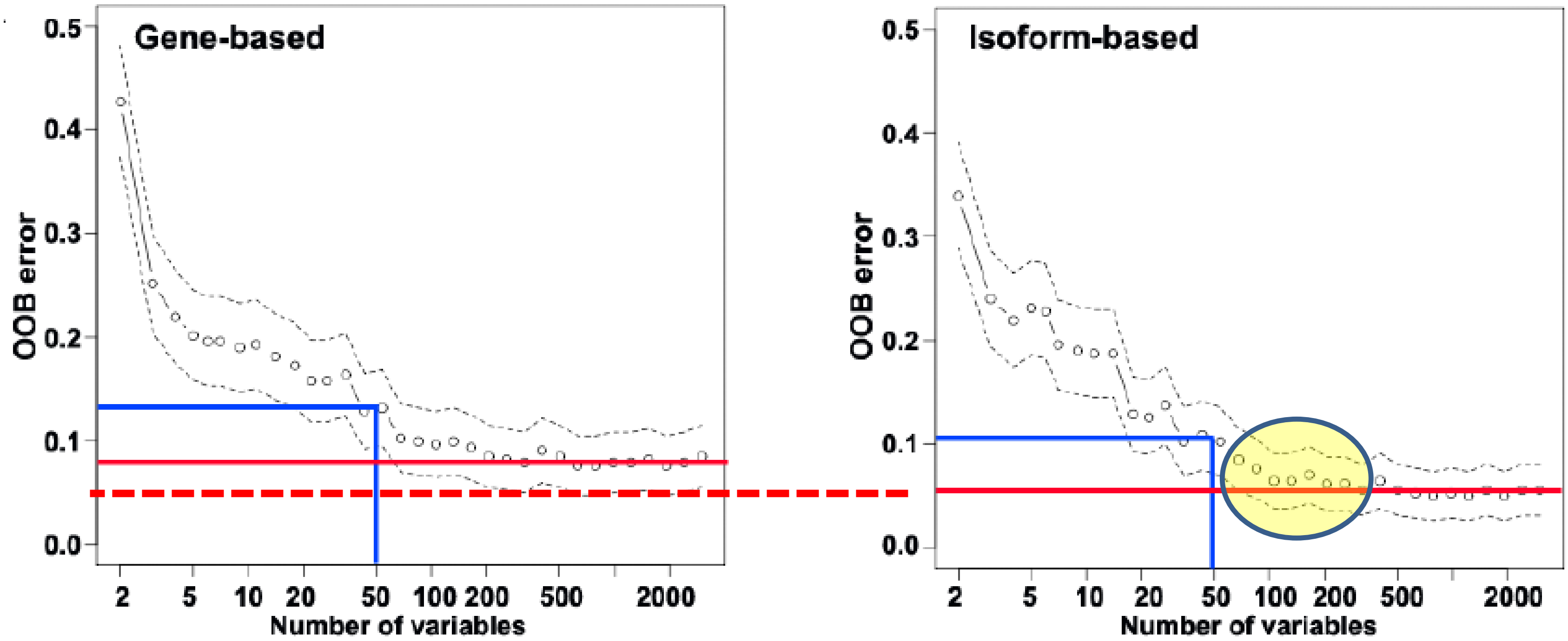
$$\left( Y_{ij} \right) =$$

| ID | Sample 1 | Sample 2 |
|---|---|---|
| ENSG00000185518 | 3.23 | 1.68 |
| ENSG00000147676 | 2.68 | 1.34 |
| ENSG00000006116 | 1.95 | 1.95 |
| ENSG00000072657 | 1.21 | 1.85 |
| ENSG00000102468 | 2.39 | 1.85 |
| ENSG00000166111 | 2.53 | 1.28 |
| ENSG00000164588 | 2.30 | 2.66 |
| ENSG00000137766 | 1.77 | 2.57 |
| ENSG00000104888 | 3.96 | 1.81 |

| ID | Sample 1 | Sample 2 |
|---|---|---|
| ENSG00000185518 | 1 | 2 |
| ENSG00000147676 | 2 | 3 |
| ENSG00000006116 | 4 | 2 |
| ENSG00000072657 | 5 | 2 |
| ENSG00000102468 | 3 | 2 |
| ENSG00000166111 | 2 | 3 |
| ENSG00000164588 | 3 | 1 |
| ENSG00000137766 | 4 | 1 |
| ENSG00000104888 | 1 | 2 |

Data-discretization is an important step in platform transition

# Performance of gene-based vs isoform-based model to discriminate the four molecular subgroups of GBM



While the isoform-based randomForest model achieved 90% accuracy with as few as 50 isoforms as feature variables, the gene-based model required more than 100 genes as feature variables for comparable accuracy to the isoform-based model

# Classification model from RandomForest

| Number of variables/ features selected by RandormForest feature selection | OOB error rate | Error rate based on independent test set |
|---|---|---|
| 213 transcript variants | 0.0661 | 0.07 |

## Assay design- Open array platform

**121 variable transcripts - 18 Non-coding transcripts**

8 transcripts- consistently up

7 transcripts- consistently down

4 house keeping genes- Polr2a, GAPDH, B2M, b-Actin

# Accuracy of 121 transcript-based classifier on exon-array data

Predicted labels

True labels

| | N | PN | M | CL | Class Error |
|---|---|---|---|---|---|
| N (78) | 63 | 5 | 3 | 5 | 0.17 |
| PN (95) | 0 | 92 | 1 | 2 | 0.03 |
| M (85) | 3 | 0 | 82 | 0 | 0.04 |
| CL (86) | 4 | 1 | 1 | 80 | 0.07 |

Confusion matrix based on 121 selected transcripts (Number of bins equal to 15)

OOB estimate of error rate: 7.31%

# Accuracy of 121 transcript-based classifier on RNA-seq data (76 samples)

Predicted labels

True labels

| | N | PN | M | CL | Class Error |
|---|---|---|---|---|---|
| N (22) | 16 | 1 | 1 | 4 | 0.27 |
| PN (18) | 0 | 18 | 0 | 0 | 0.00 |
| M (20) | 0 | 0 | 20 | 0 | 0.00 |
| CL (16) | 0 | 0 | 0 | 16 | 0.00 |

Confusion matrix based on 121 selected transcripts (Number of bins equal to 15)

OOB estimate of error rate: 7.89%

# Sub-typing of 206 GBM patients using RT-qPCR assay (based on 121 assays/transcripts)

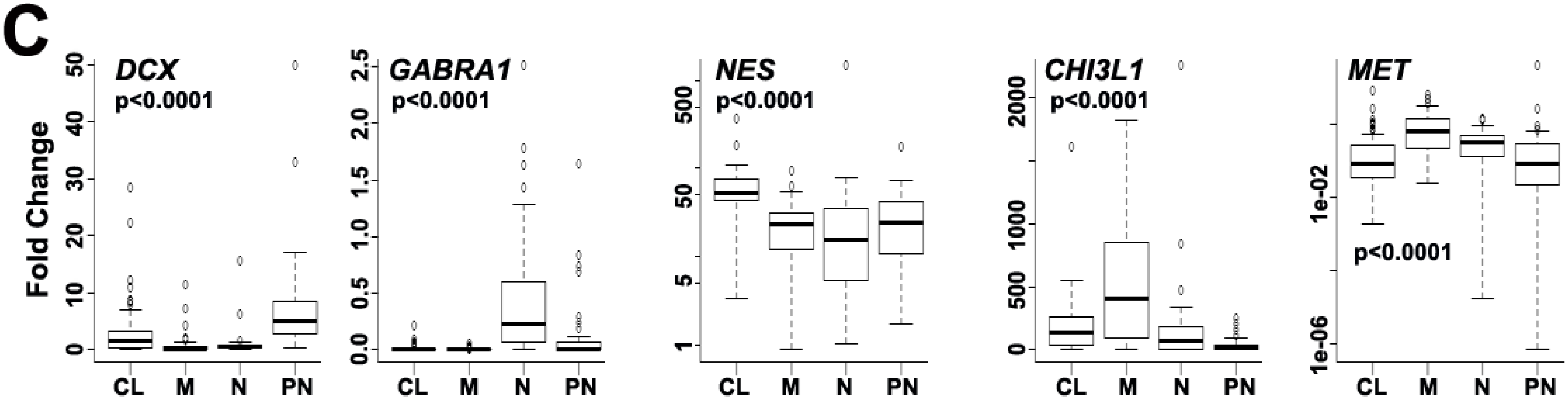
| Sample ID | Probability of sample in sub-type | | | | Predicted Sub-type |
|---|---|---|---|---|---|
| | CL | M | N | PN | |
| 1409 | 0.16 | 0.16 | **0.43** | 0.25 | N |
| 1470 | 0.02 | **0.96** | 0.01 | 0.01 | M |
| 1621 | 0.02 | 0.01 | **0.88** | 0.09 | N |
| 1716 | 0.04 | 0.02 | 0.17 | **0.77** | PN |
| 1770 | 0.08 | 0.01 | 0.36 | **0.55** | PN |
| 1817 | **0.53** | 0.23 | 0.10 | 0.14 | CL |
| 1961 | **0.87** | 0.05 | 0.05 | 0.03 | CL |
| 1659 | 0.03 | 0.02 | 0.49 | 0.46 | N |
| 1730 | 0.09 | 0.11 | 0.39 | 0.41 | PN |

High-confidence predictions
**91%**

Low-confidence predictions
**9%**

| | N | PN | M | CL | Total |
|---|---|---|---|---|---|
| TCGA | 76<br>(22%) | 95<br>(27.8%) | 85<br>(24.9%) | 86<br>(25.2%) | 342 |
| PENN | 41<br>(19.9%) | 52<br>(25.2%) | 50<br>(24.2%) | 63<br>(30.5%) | 206 |

# Validation of our classifier-PENN GBM cohort



Expression of specific markers for each subgroup

| Group | Marker gene |
|---|---|
| PN | *DCX* |
| N | *GABRA1* |
| CL | *NES* |
| M | *CHI3L1 and MET* |

# Summary

- Isoform-level expression clustering identified four GBM subgroups with significant (p=0.0103) survival differences

- A four-class classifier, built with 121 transcript-variants, assigns GBM patients' molecular subtype with 92% accuracy

- The GBM classifier was translated to an RT-qPCR-based assay and validated on an independent cohort of 206 glioblastoma samples, and maintained high-confidence subtype calls for 91% of the patients.

- We found the proneural subtype to have the worst prognosis for patients, except for the younger group (<40 years) who showed significantly better survival (p=0.007), while a better prognosis for the neural subtype was observed (p=0.02) in older patients (≥40 years).

# Clinical Significance of the Assay

- This assay could be used in prospective clinical trials to select specific groups of GBM patients for treatment with drugs targeting subtype-specific pathways

- GBM patients can be stratified into 4 subgroups, so that patients within a group can receive treatments that have been tailored specifically for them

# Topics of Discussion

1. Grows of multi –omics data
2. Why “gene” as a unit of measure is too simplistic?
3. Exon-arrays and RNA-seq methods
4. Gene-level Vs Isoform-level analysis
   A. Cancer Vs Non-cancer cell-line grouping
   B. Isoform-level gene signatures for brain tumor sub-typing
5. Evaluation of isoform-level expression estimation algorithms for RNA-seq and exon-array platforms

# Comparative assessment of isoform-level expression estimation algorithms (for RNA-Seq, exon-array)

Matthew Dapas

1. **TCGA data:**
   - 103 tumor- and 4 normal-tissue glioblastoma multiforme (GBM) samples
   - Samples feature both RNA-seq and exon array data available in TCGA
2. **Exon array analysis:**
   - Estimates obtained using Multi-Mapping Bayesian Gene eXpression (MMBGX)
   - Ensembl 70 (GRCh37.p8) reference annotation
3. **RNA-seq analysis:**
   - Genome alignments were made using Bowtie2, Ensembl 70.
   - Tested the following tools: **TopHat/Cufflinks**, **RSEM**, **eXpress**, and **Sailfish**.
4. **RT-qPCR:**
   - GBM samples obtained from the Human Brain Tumor Tissue bank at the University of Pennsylvania
   - RT-qPCR performed on 159 transcripts previously selected for tumor subtyping
5. **Expression and fold change correlations:**
   - Sample-by-sample correlations between RNA-seq and exon array evaluated using Spearman's correlation.
   - Fold changes calculated using mean values from 4 normal-tissue GBM samples.
   - RNA-seq expression estimates (FPKM) were normalized using upper quartile normalization.
   - For RT-qPCR correlations, estimates were further normalized by POL2A expression.

# Correlations between RNA-Seq expression estimates

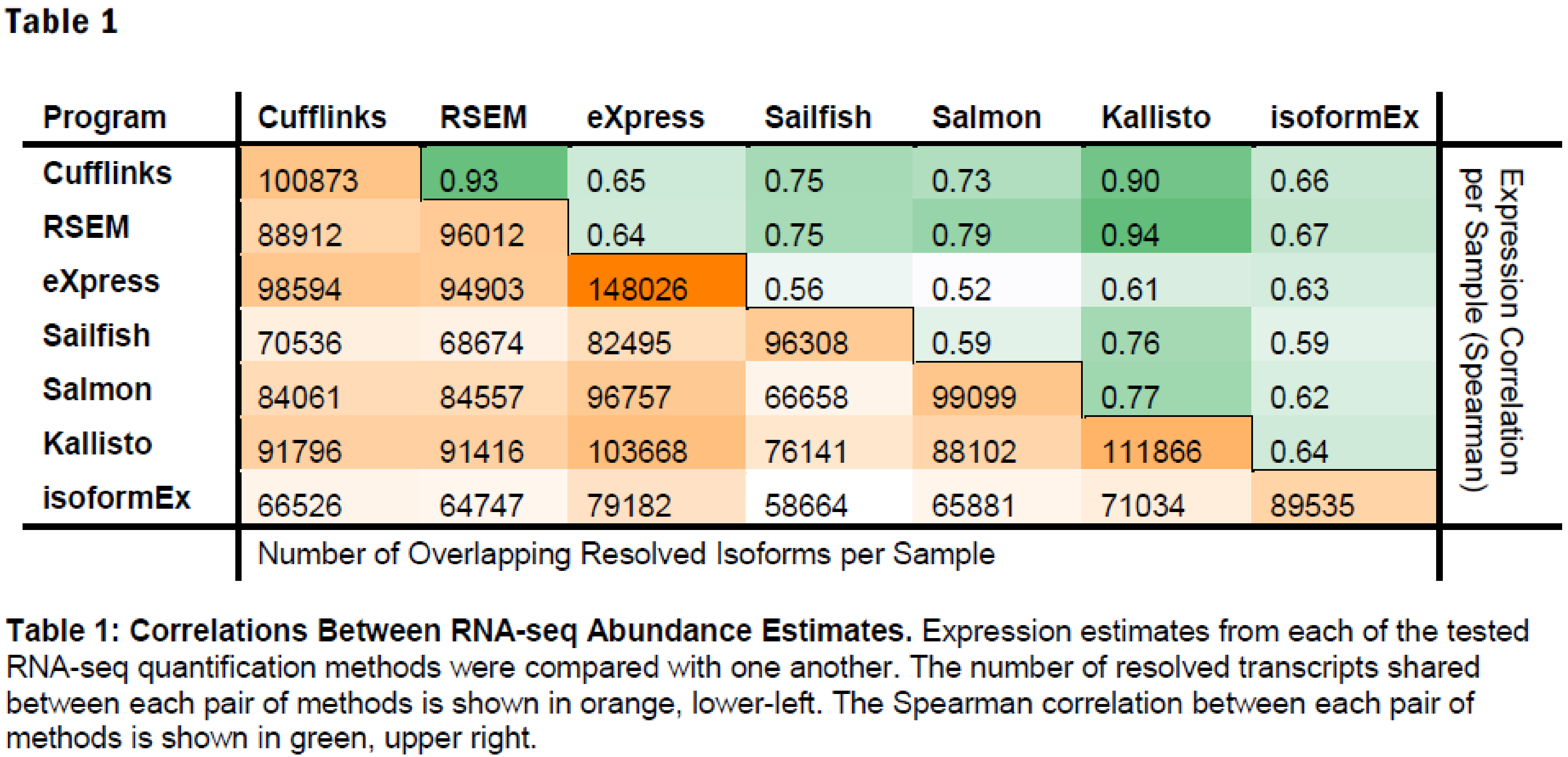


**Table 1**

| Program | Cufflinks | RSEM | eXpress | Sailfish | Salmon | Kallisto | isoformEx |
|---|---|---|---|---|---|---|---|
| Cufflinks | 100873 | 0.93 | 0.65 | 0.75 | 0.73 | 0.90 | 0.66 |
| RSEM | 88912 | 96012 | 0.64 | 0.75 | 0.79 | 0.94 | 0.67 |
| eXpress | 98594 | 94903 | 148026 | 0.56 | 0.52 | 0.61 | 0.63 |
| Sailfish | 70536 | 68674 | 82495 | 96308 | 0.59 | 0.76 | 0.59 |
| Salmon | 84061 | 84557 | 96757 | 66658 | 99099 | 0.77 | 0.62 |
| Kallisto | 91796 | 91416 | 103668 | 76141 | 88102 | 111866 | 0.64 |
| isoformEx | 66526 | 64747 | 79182 | 58664 | 65881 | 71034 | 89535 |

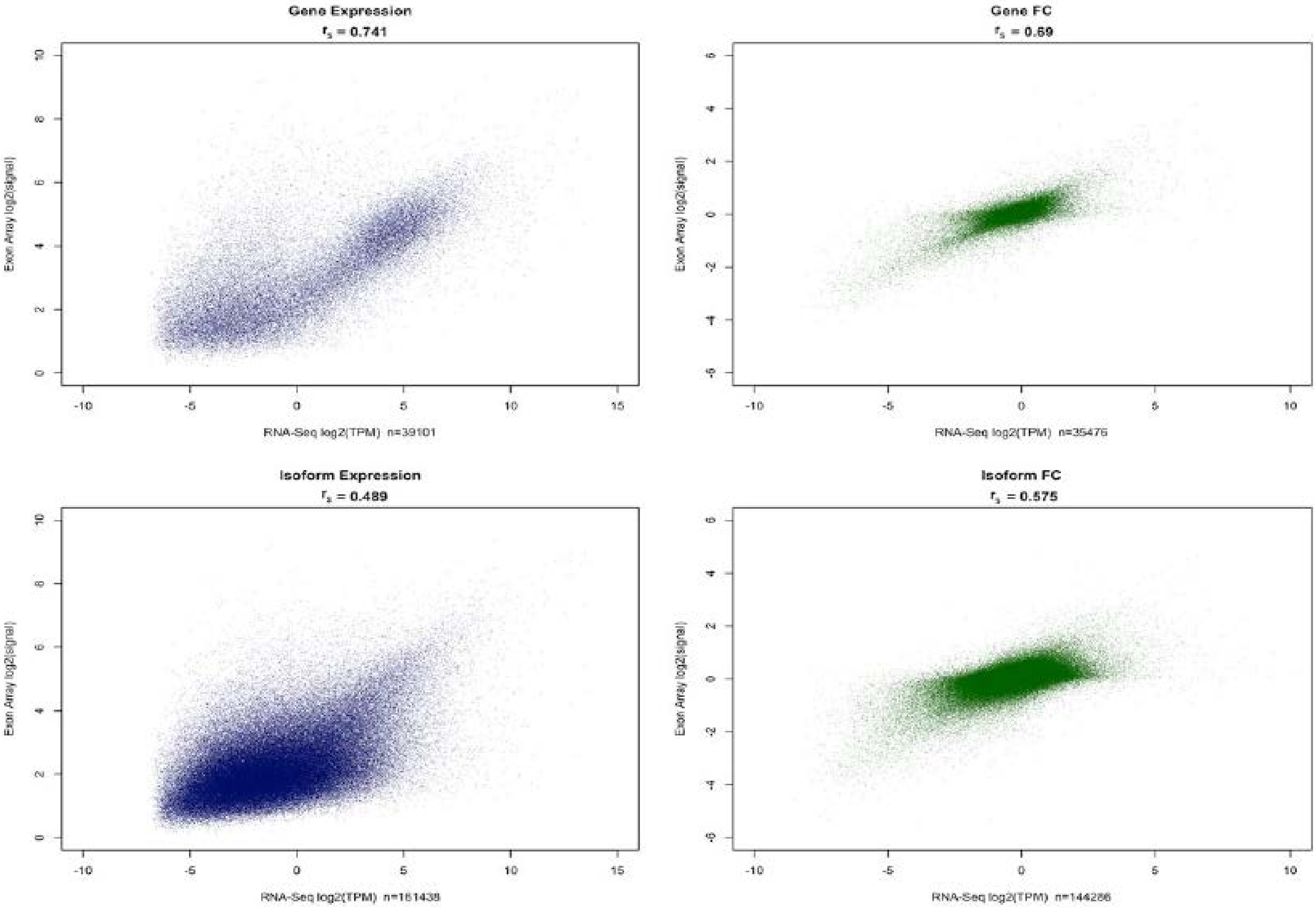
Number of Overlapping Resolved Isoforms per Sample

Expression Correlation per Sample (Spearman)

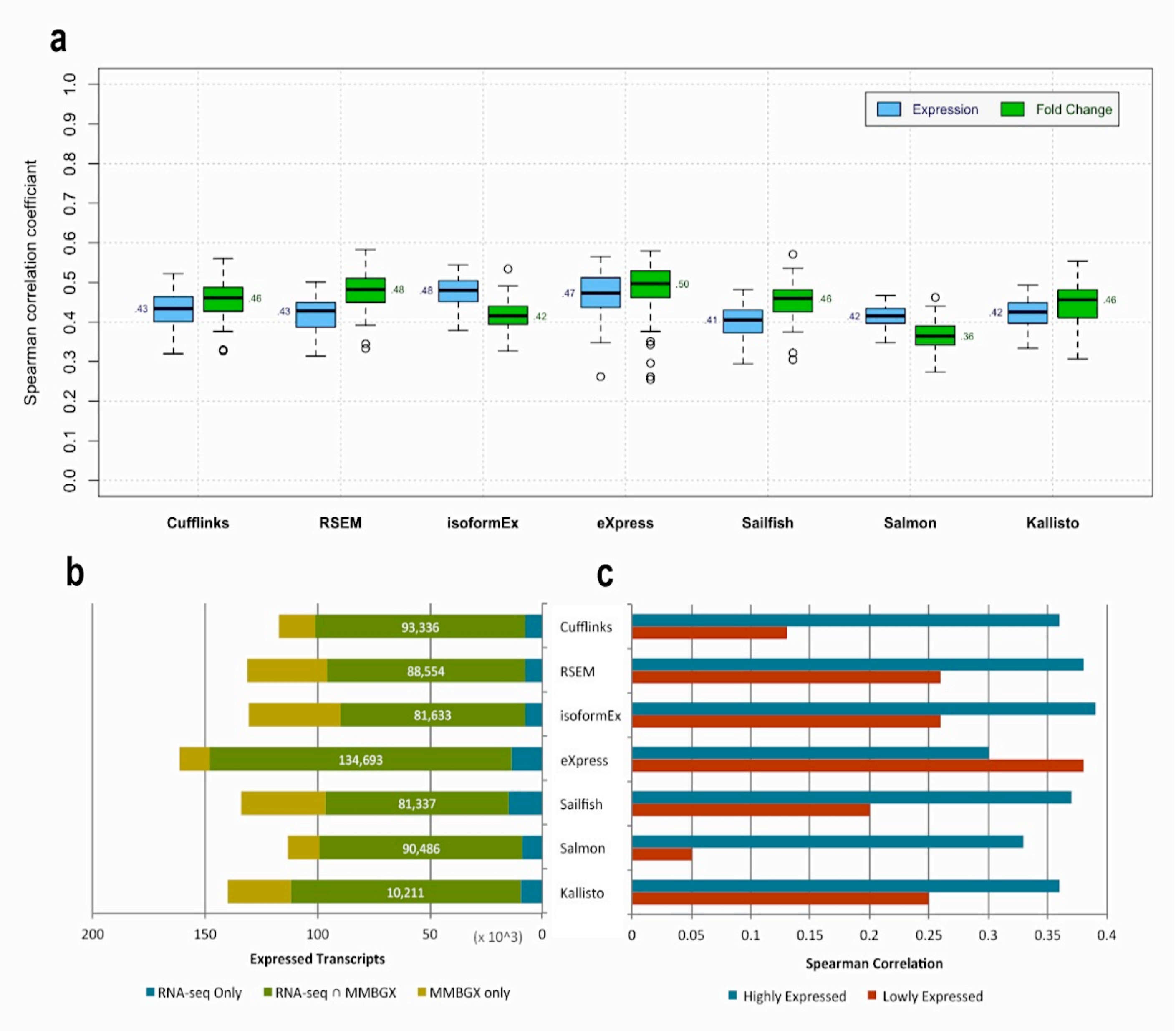
**Table 1: Correlations Between RNA-seq Abundance Estimates.** Expression estimates from each of the tested RNA-seq quantification methods were compared with one another. The number of resolved transcripts shared between each pair of methods is shown in orange, lower-left. The Spearman correlation between each pair of methods is shown in green, upper right.

NORTHWESTERN UNIVERSITY

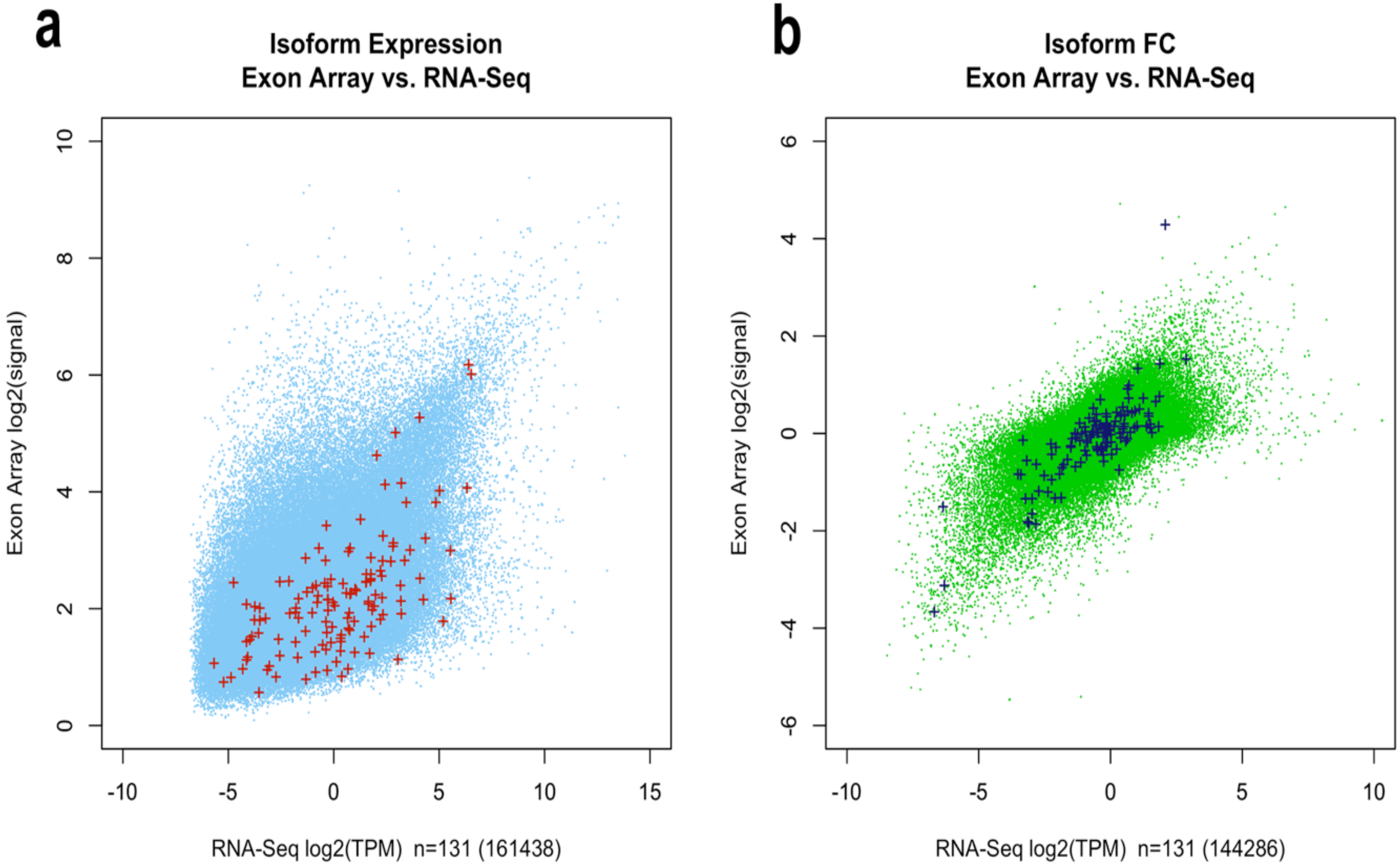
# Scatter plots of average expression and fold change (tumor vs. normal) estimates between exon array and RNA-seq



# Spearman correlation coefficients between MMBGX and different RNA-seq quantification methods



# RT-qPCR Correlations



The transcripts included in RT-qPCR analysis (red), according to their average expression estimates (a) and fold-changes (b) from the RNA-seq and MMBGX exon array tumor results

# RT-qPCR Correlations

| Algorithm | Expression Correlation ($r_S$) | Fold Change Correlation ($r_S$) | # Transcripts |
|---|---|---|---|
| eXpress | **0.470** | **0.900** | **139** |
| isoformEx | 0.292 | 0.873 | 127 |
| Salmon | 0.115 | 0.864 | 131 |
| Kallisto | 0.287 | 0.860 | 132 |
| TopHat/ Cufflinks | 0.223 | 0.849 | 133 |
| Exon Array - MMBGX | **0.424** | **0.836** | **142** |
| RSEM | 0.231 | 0.835 | 132 |
| Sailfish | 0.259 | 0.812 | 126 |

The Spearman correlations and number of shared, resolved transcripts between the various programs tested and the RT-qPCR estimates

# Summary

- Better concordance between RNA-seq/exon-array and RT-qPCR platforms for fold change estimates than for raw abundance estimates, suggesting that fold-change normalization against a control is an important step for integrating expression data across platforms.

- Potentially important isoform-level expression changes can be masked by gene-level estimates

- While eXpress and MMBGX programs achieved the best performance for RNA-seq and exon-array platforms respectively for deriving the isoform-level fold change values, there is an urgent need to improve the methods for abundance estimation.

# Acknowledgement

## Funding :

- NHGRI/NIH (R01)
- Pennsylvania State Dept of Health
- Philadelphia Healthcare Trust Professorship, Wistar Institute, Philadelphia, PA
- Tobin Kestenbaum Family Professorship, Wistar Institute, Philadelphia, PA
- Zell Scholar, Robert H. Lurie Comprehensive Cancer Center, NU-FSM, Chicago, IL.
- NLM/NIH (R01 LM011297)
- NCI SPORE in Prostate Cancer (P50 CA090386)
- Intelligence Advanced Research Projects Activity

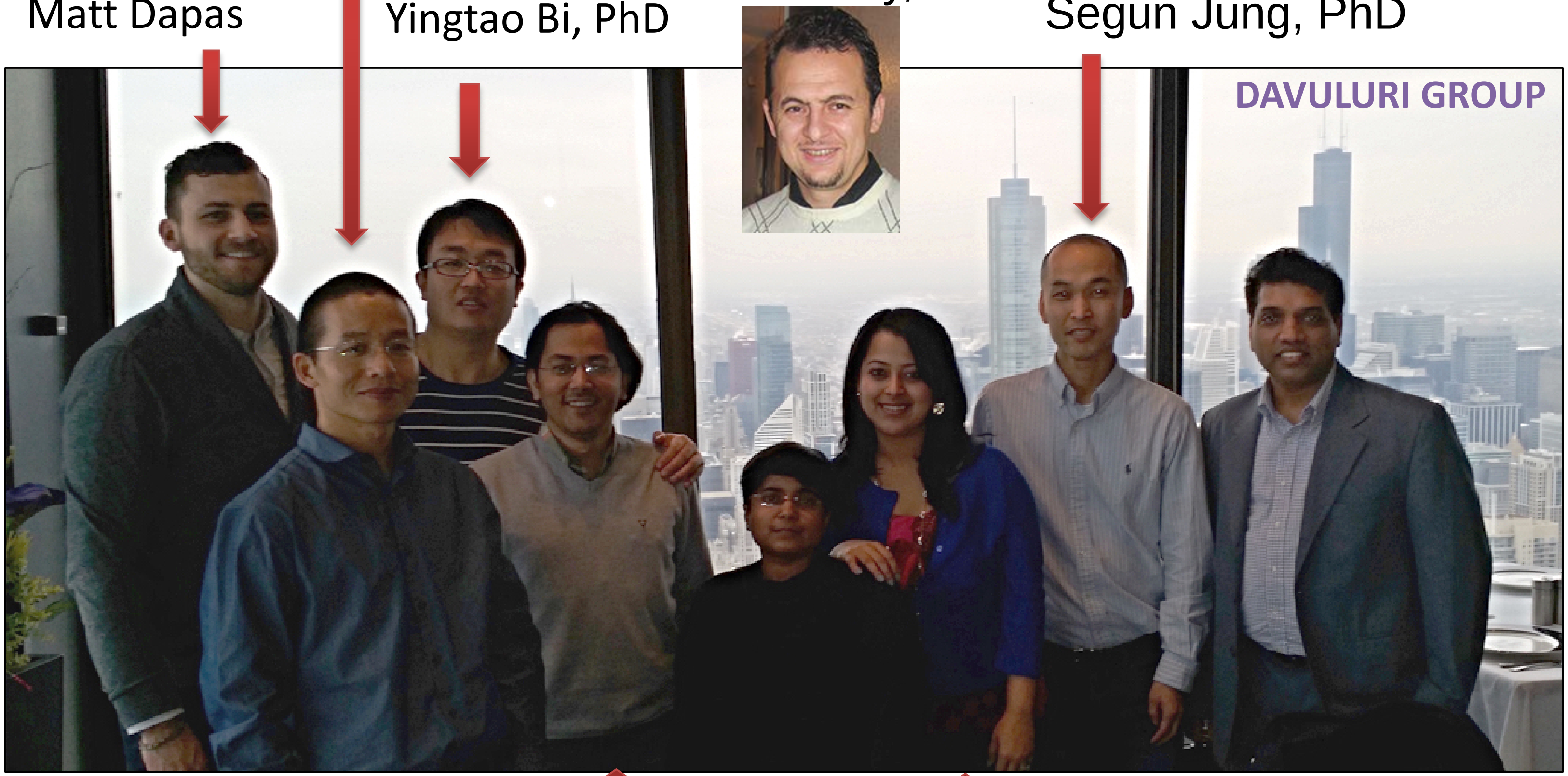
Hongjian Jin, PhD

Matt Dapas

Yingtao Bi, PhD

Ferhat Ay, PhD

Segun Jung, PhD



Manoj Kandpal, PhD

Arunima Shilpi, PhD Student

Auditi Debroy, PhD

NORTHWESTERN UNIVERSITY
FEINBERG
SCHOOL OF MEDICINE

Center for Data Science and Informatics

NORTHWESTERN
UNIVERSITY