Novel bioinformatics methods for the identification of coexpressed, differentially expressed, and differentially coexpressed genes with application to cancer

Obi L. Griffith
Supervisor: Dr. Steven Jones
Thursday Trainee Seminar
June 14, 2007

How can we use gene expression data to investigate cancer?

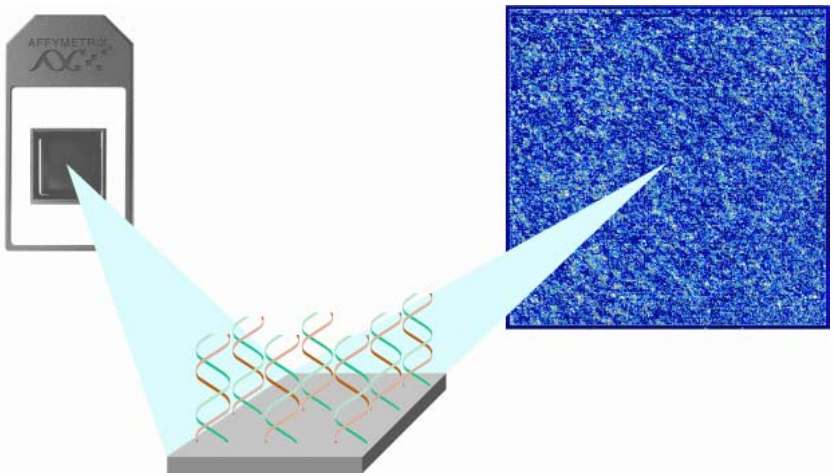
I. Multi-platform Coexpression

II. Multi-platform differential expression – Thyroid cancer

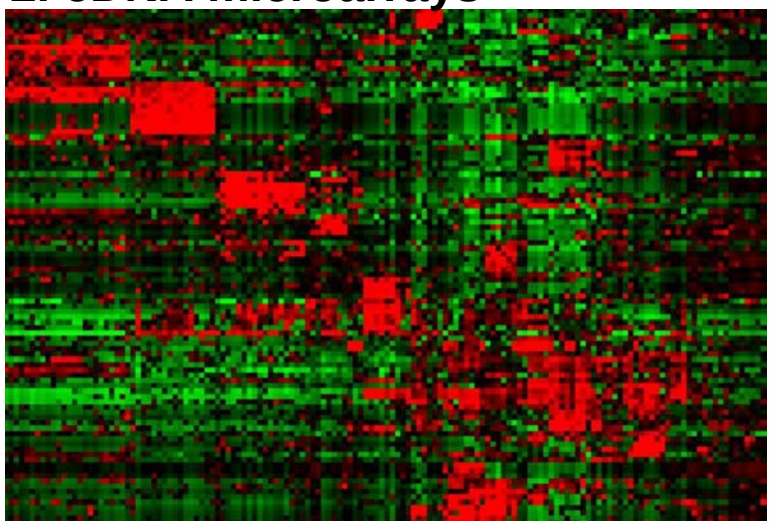
III. Differential Coexpression – Prostate cancer

Three major expression platforms

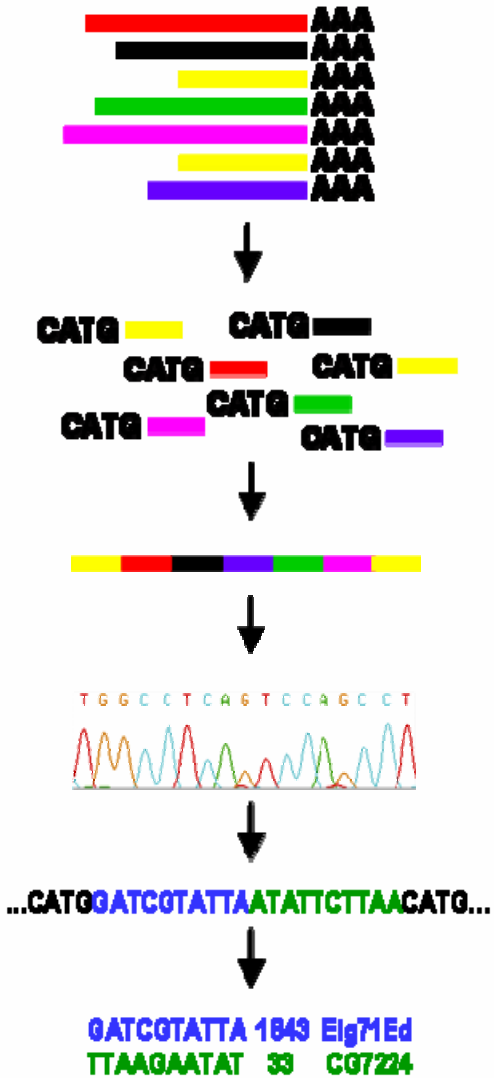
1. Oligonucleotide arrays



2. cDNA microarrays



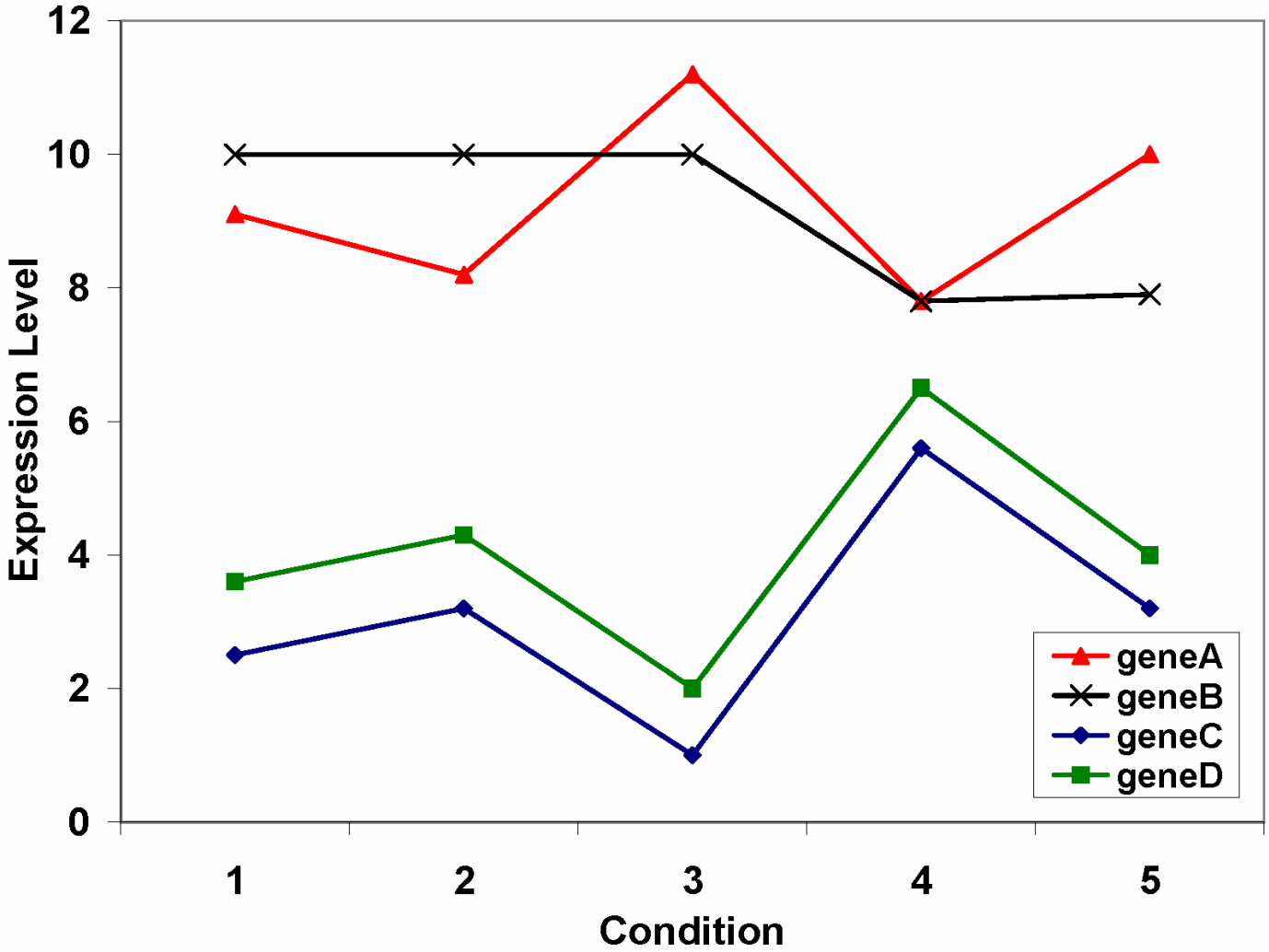
3. SAGE



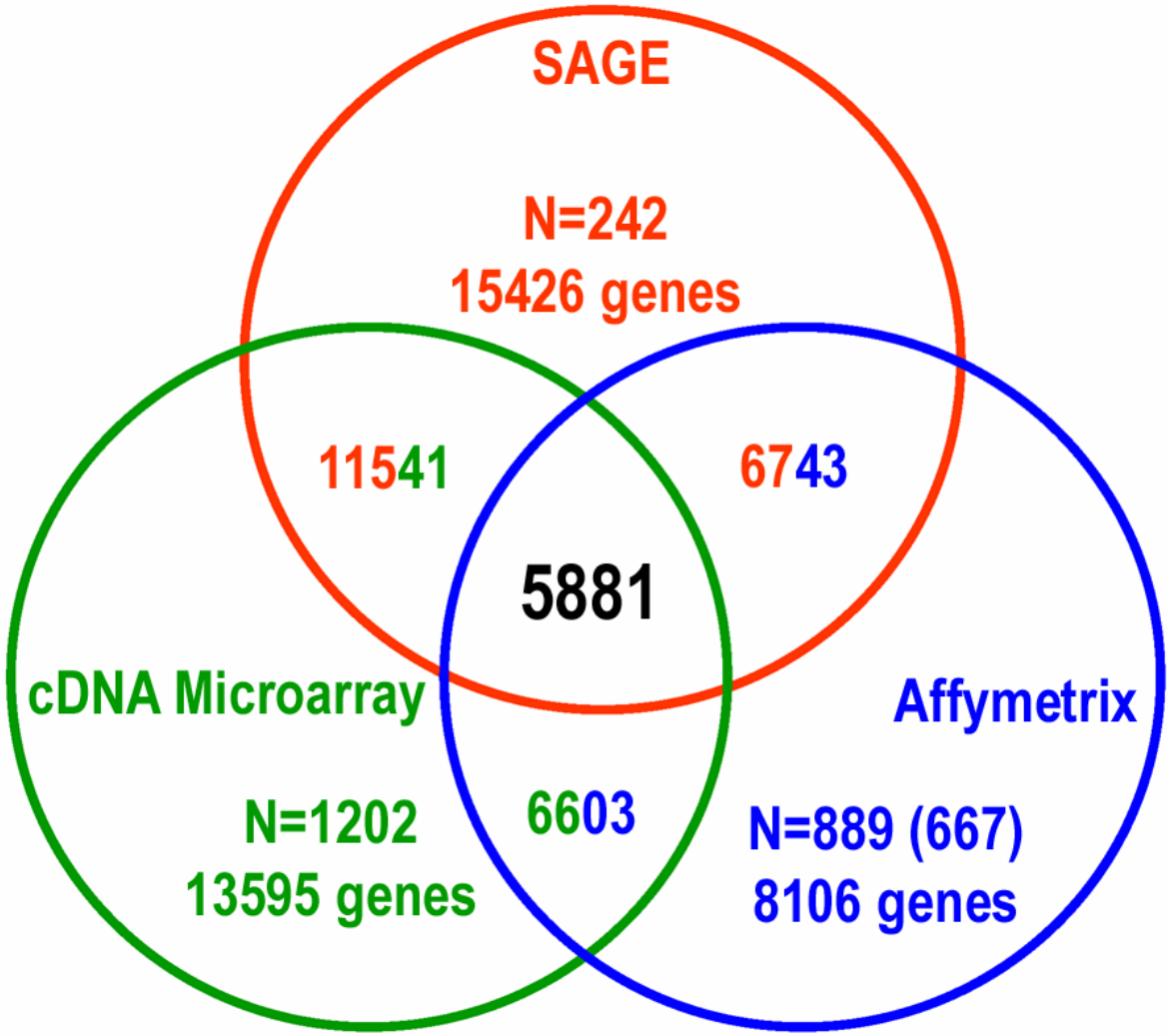
I) Multi-platform coexpression

- Coexpression can be used to
  - define clusters of genes with common biological processes
  - infer functional associations between genes
  - for integration with other large-scale datasets
  - for the generation of high-quality biological interaction networks
  - to identify co-regulation
  - identify groups of related genes that are important in specific cancers or represent common tumour progression mechanisms

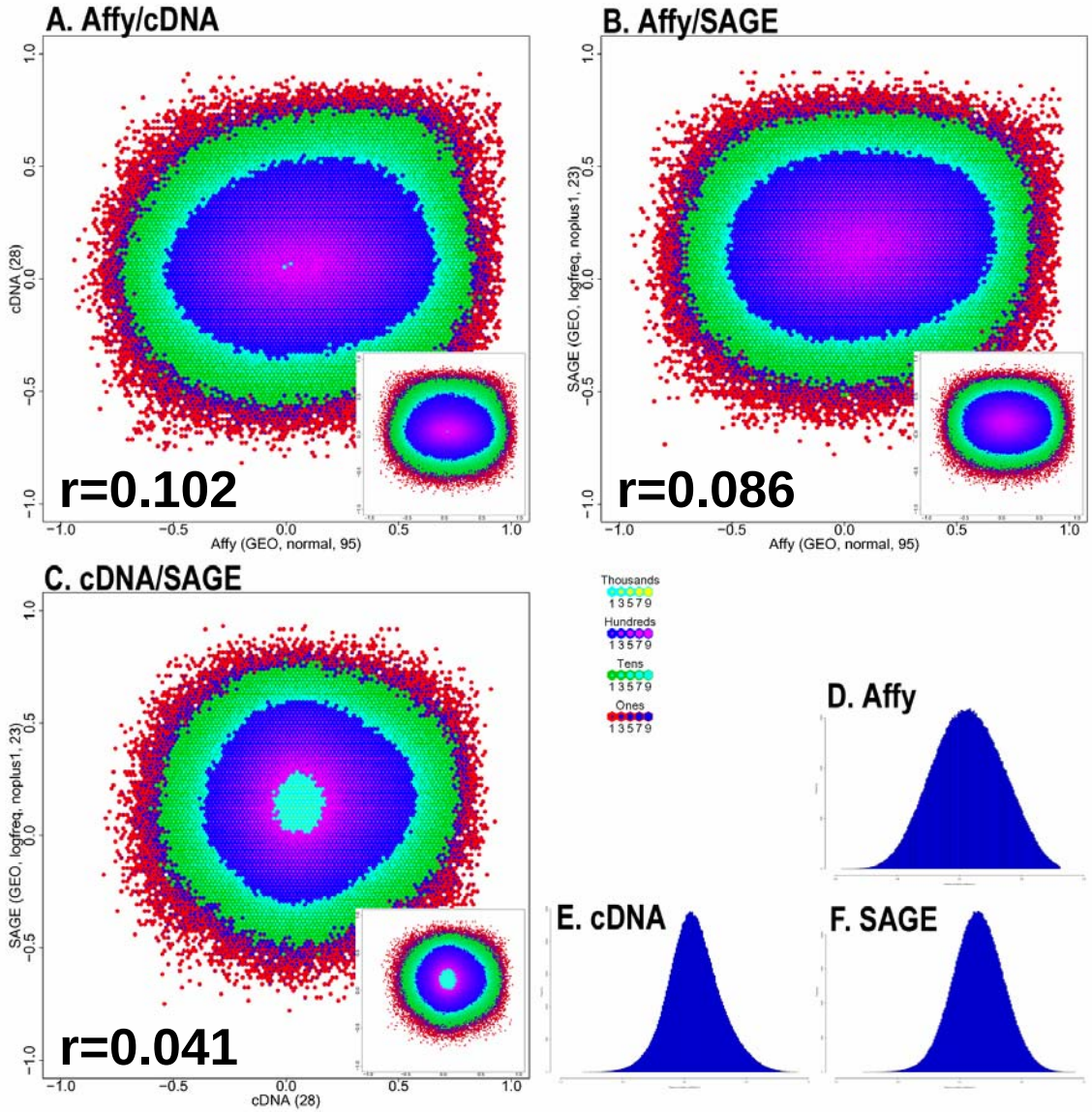
What is Coexpression?



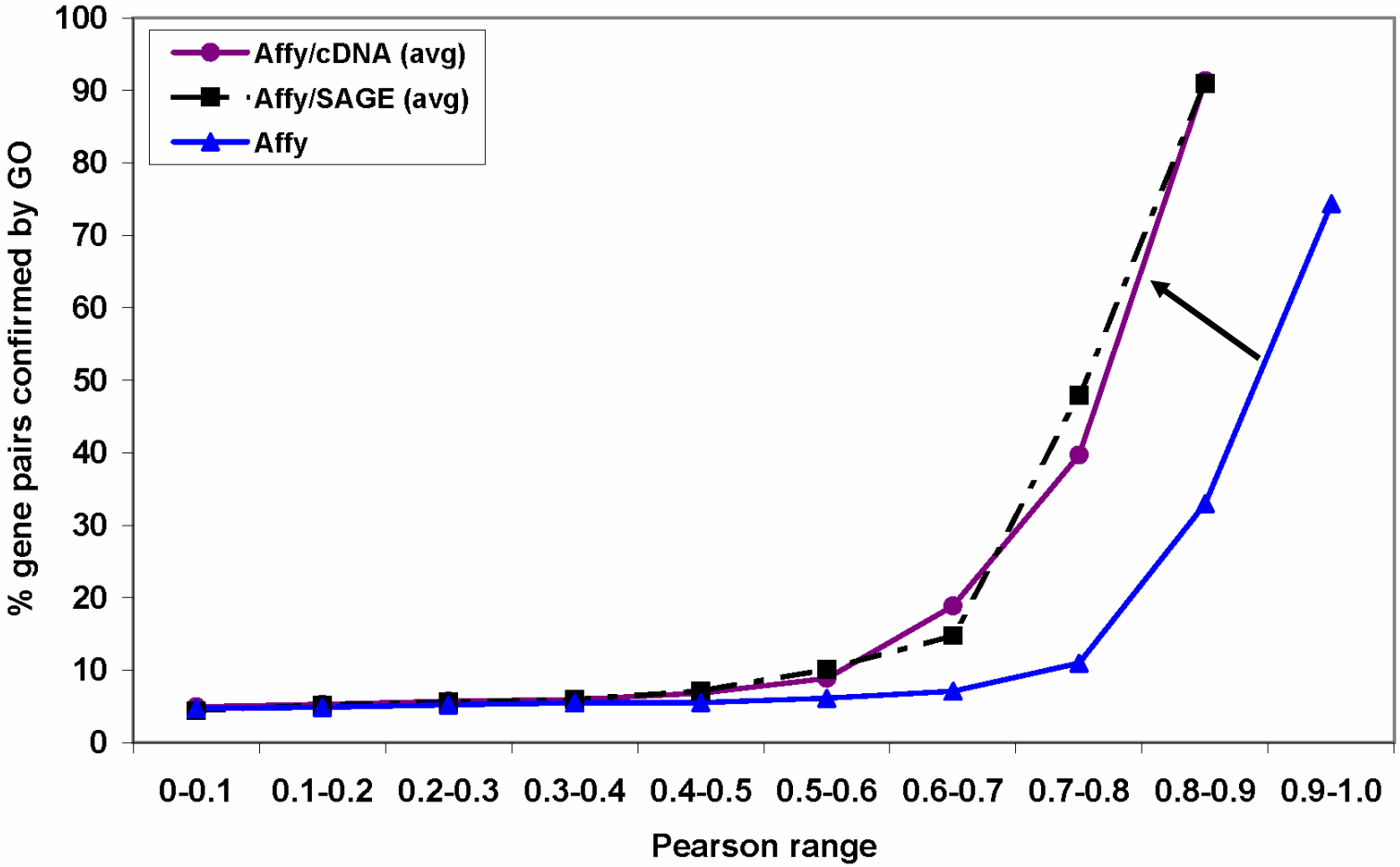
Available data



Platform Comparisons



Coexpression methods that combine different platforms or datasets improve quality of predictions (according to GO)



OL Griffith, ED Pleasance, DL Fulton, M Oveisi, M Ester, AS Siddiqui, SJM Jones. 2005. Assessment and Integration of Publicly Available SAGE, cDNA Microarray, and Oligonucleotide Microarray Expression Data for Global Coexpression Analyses. Genomics. 86:476-488

Conclusions

- Platforms compare significantly better than random but in general correlations are poor
- GO analysis indicates that all 3 platforms identify some biologically relevant gene pairs
- Higher Pearson indicates increased biological relevance
- Combining different platforms improves quality of predictions

II) Multi-platform differential expression in thyroid cancer

- Thyroid nodules are extremely common
  - 4-7% of North American adult population
- Fine needle aspiration biopsy (FNAB) is most important initial test
  - **10-20% indeterminate or suspicious → Surgery**
- After thyroid surgery as little as 20% are confirmed as malignant

Rationale

- Improved diagnostic markers are needed
- Gene expression profiling attempts to identify such markers
- A large number of thyroid cancer expression profiling studies exist
- Hundreds/thousands of potential markers (genes) have been identified
- What markers are most consistently reported?

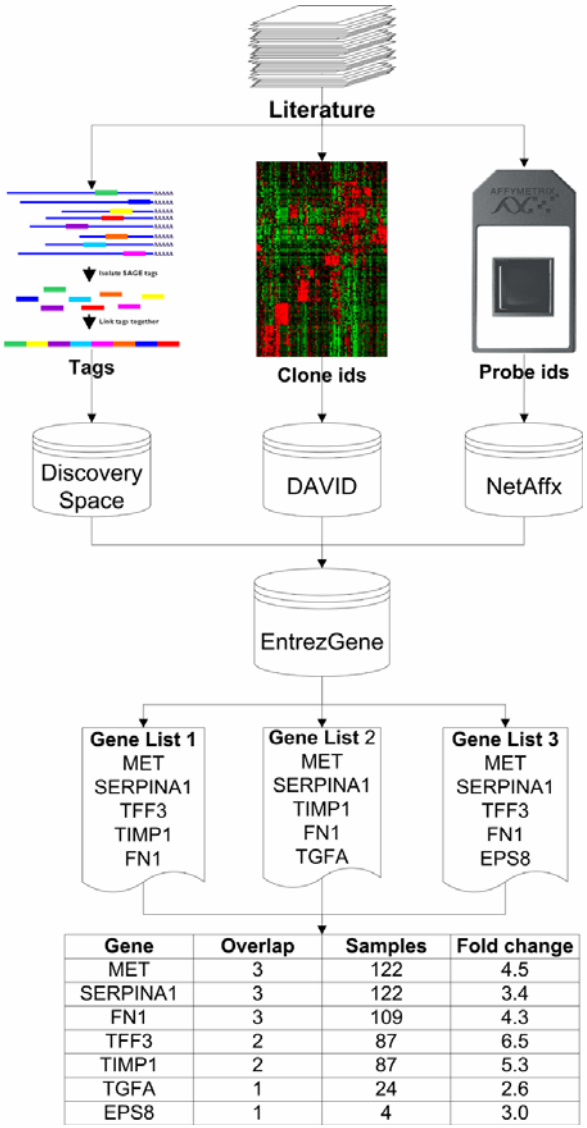
Literature review reveals 21 studies

Study	Platform	Genes/ features	Comparison		Up-/down
			Condition 1 (No. samples)	Condition 2 (No. samples)	
Chen *et al.* 2001	Atlas cDNA (Clontech)	588	M (1)	FTC (1)	18/40
Arnaldi *et al.* 2005	Custom cDNA	1807	FCL(1)	Norm (1)	9/20
			PCL(1)	Norm (1)	1/8
			UCL(1)	Norm (1)	1/7
			FCL(1), PCL(1), UCL(1)	Norm (1)	3/6
Huang *et al.* 2001	Affymetrix HG-U95A	12558	PTC (8)	Norm (8)	24/27
Aldred *et al.* 2004	Affymetrix HG-U95A	12558	FTC (9)	PTC(6), Norm(13)	142/0
			PTC (6)	FTC(9), Norm(13)	0/68
Cerutti *et al.* 2004	SAGE	N/A	FA(1)	FTC(1), Norm(1)	5/0
			FTC(1)	FA(1), Norm(1)	12/0
Eszlinger *et al.* 2001	Atlas cDNA (Clontech)	588	AFTN(3), CTN(3)	Norm(6)	0/16
Finley *et al.* 2004	Affymetrix HG-U95A	12558	PTC(7), FVPTC(7)	FA(14), HN(7)	48/85
Zou *et al.* 2004	Atlas cancer array	1176	MACL(1)	ACL(1)	43/21
Weber *et al.* 2005	Affymetrix HG-U133A	22283	FA(12)	FTC(12)	12/84
Hawthorne *et al.* 2004	Affymetrix HG-U95A	12558	GT(6)	Norm(6)	1/7
			PTC(8)	GT(6)	10/28
			PTC(8)	Norm(8)	4/4
Onda *et al.* 2004	Amersham custom cDNA	27648	ACL(11), ATC(10)	Norm(10)	31/56
Wasenius *et al.* 2003	Atlas cancer cDNA	1176	PTC(18)	Norm(3)	12/9
Barden *et al.* 2003	Affymetrix HG-U95A	12558	FTC(9)	FA(10)	59/45
Yano *et al.* 2004	Amersham custom cDNA	3968	PTC(7)	Norm(7)	54/0
Chevillard *et al.* 2004	custom cDNA	5760	FTC(3)	FA(4)	12/31
			FVPTC(3)	PTC(2)	123/16
Mazzanti *et al.* 2004	Hs-UniGem2 cDNA	10000	PTC(17), FVPTC(15)	FA(16), HN(15)	5/41
Takano *et al.* 2000	SAGE	N/A	FTC(1)	ATC(1)	3/10
			FTC(1)	FA(1)	4/1
			Norm(1)	FA(1)	6/0
			PTC(1)	ATC(1)	2/11
			PTC(1)	FA(1)	7/0
			PTC(1)	FTC(1)	2/1
Finley *et al.* 2004	Affymetrix HG-U95A	12558	FTC(9), PTC(11), FVPTC(13)	FA(16), HN(10)	50/55
Pauws *et al.* 2004	SAGE	N/A	FVPTC(1)	Norm(1)	33/9
Jarzab *et al.* 2005	Affymetrix HG-U133A	22283	PTC(16)	Norm(16)	75/27
Giordano *et al.* 2005	Affymetrix HG-U133A	22283	PTC(51)	Norm(4)	90/151
21 studies	**10 platforms**		**34 comparisons (473 samples)**		**1785**

21 cancer vs. non-cancer comparisons

Study	Platform	Genes/ features	Comparison		Up-/down
			Condition 1 (No. samples)	Condition 2 (No. samples)	
Chen *et al.* 2001	Atlas cDNA (Clontech)	588	M (1)	FTC (1)	18/40
Arnaldi *et al.* 2005	Custom cDNA	1807	FCL(1)	Norm (1)	9/20
			PCL(1)	Norm (1)	1/8
			UCL(1)	Norm (1)	1/7
			FCL(1), PCL(1), UCL(1)	Norm (1)	3/6
Huang *et al.* 2001	Affymetrix HG-U95A	12558	PTC (8)	Norm (8)	24/27
Aldred *et al.* 2004	Affymetrix HG-U95A	12558	FTC (9)	PTC(6), Norm(13)	142/0
			PTC (6)	FTC(9), Norm(13)	0/68
Cerutti *et al.* 2004	SAGE	N/A	FA(1)	FTC(1), Norm(1)	5/0
			FTC(1)	FA(1), Norm(1)	12/0
Eszlinger *et al.* 2001	Atlas cDNA (Clontech)	588	AFTN(3), CTN(3)	Norm(6)	0/16
Finley *et al.* 2004*	Affymetrix HG-U95A	12558	PTC(7), FVPTC(7)	FA(14), HN(7)	48/85
Zou *et al.* 2004	Atlas cancer array	1176	MACL(1)	ACL(1)	43/21
Weber *et al.* 2005	Affymetrix HG-U133A	22283	FA(12)	FTC(12)	12/84
Hawthorne *et al.* 2004	Affymetrix HG-U95A	12558	GT(6)	Norm(6)	1/7
			PTC(8)	GT(6)	10/28
			PTC(8)	Norm(8)	4/4
Onda *et al.* 2004	Amersham custom cDNA	27648	ACL(11), ATC(10)	Norm(10)	31/56
Wasenius *et al.* 2003	Atlas cancer cDNA	1176	PTC(18)	Norm(3)	12/9
Barden *et al.* 2003	Affymetrix HG-U95A	12558	FTC(9)	FA(10)	59/45
Yano *et al.* 2004	Amersham custom cDNA	3968	PTC(7)	Norm(7)	54/0
Chevillard *et al.* 2004	custom cDNA	5760	FTC(3)	FA(4)	12/31
			FVPTC(3)	PTC(2)	123/16
Mazzanti *et al.* 2004	Hs-UniGem2 cDNA	10000	PTC(17), FVPTC(15)	FA(16), HN(15)	5/41
Takano *et al.* 2000	SAGE	N/A	FTC(1)	ATC(1)	3/10
			FTC(1)	FA(1)	4/1
			Norm(1)	FA(1)	6/0
			PTC(1)	ATC(1)	2/11
			PTC(1)	FA(1)	7/0
			PTC(1)	FTC(1)	2/1
Finley *et al.* 2004*	Affymetrix HG-U95A	12558	FTC(9), PTC(11), FVPTC(13)	FA(16), HN(10)	50/55
Pauws *et al.* 2004	SAGE	N/A	FVPTC(1)	Norm(1)	33/9
Jarzab *et al.* 2005	Affymetrix HG-U133A	22283	PTC(16)	Norm(16)	75/27
Giordano *et al.* 2005	Affymetrix HG-U133A	22283	PTC(51)	Norm(4)	90/151
21 studies	**10 platforms**		**34 comparisons (473 samples)**		**1785**

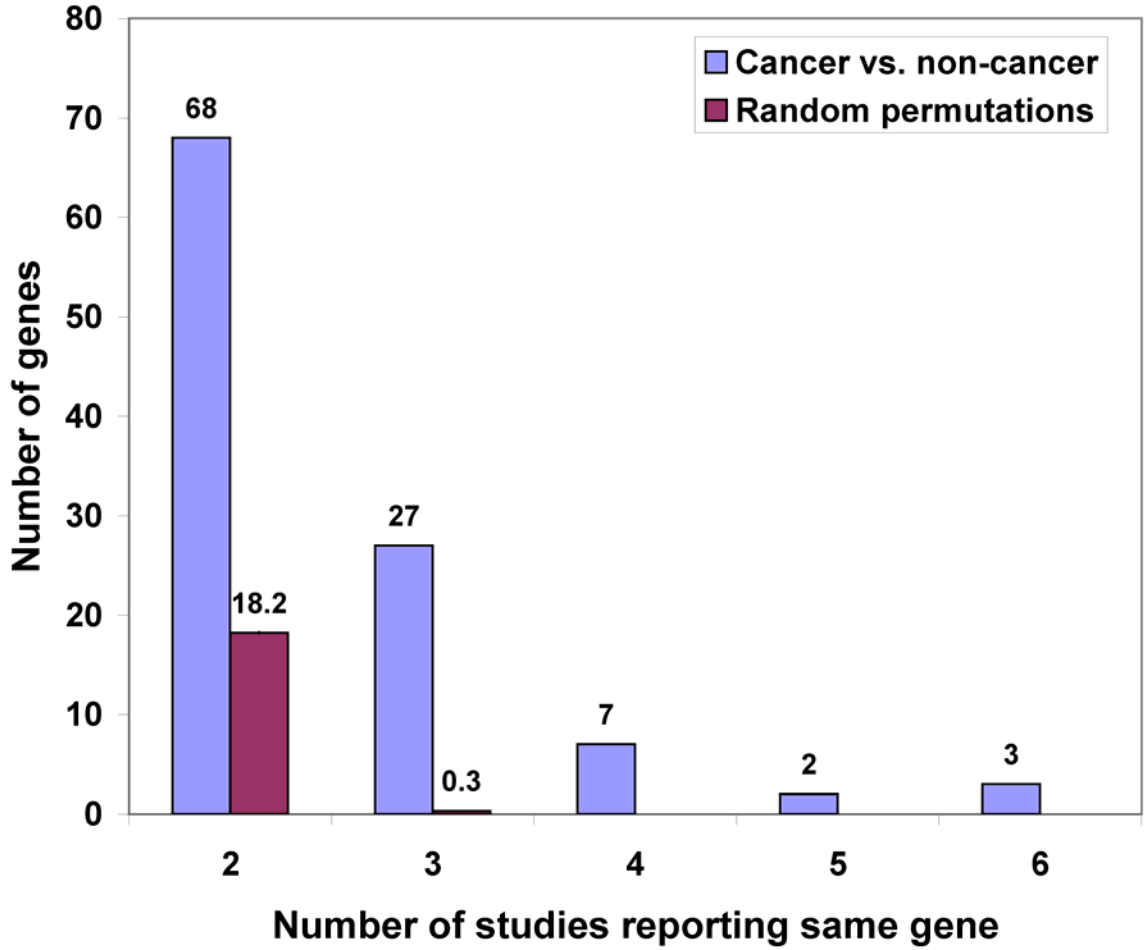
Multi-platform approach



Gene	Overlap	Samples	Fold change
MET	3	122	4.5
SERPINA1	3	122	3.4
FN1	3	109	4.3
TFF3	2	87	6.5
TIMP1	2	87	5.3
TGFA	1	24	2.6
EPS8	1	4	3.0

- Collect and curate data from over 20 studies
- Map various IDs to Entrez Gene ID
- Analyze datasets for overlap
- Rank genes according to:
  - amount of overlap
  - size of studies
  - fold change
- Assess significance of result

A significant number of genes are consistently reported as differentially expressed from multiple independent studies



Griffith OL, Melck A, Jones SJM, Wiseman SM. 2006. JCO. 24(31):5043-5051.

Top 12 most consistently differentially expressed genes (cancer vs. non-cancer)

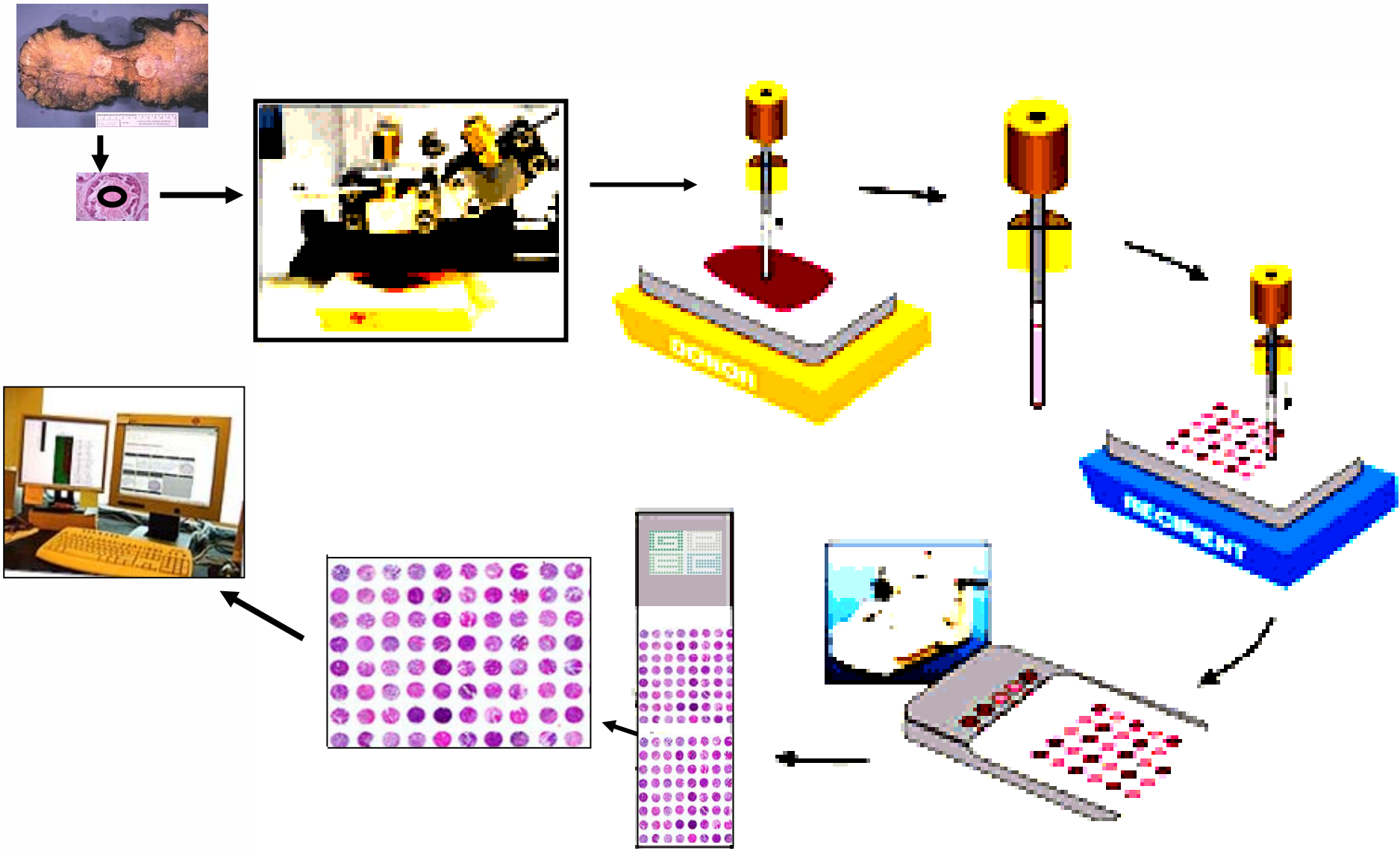
Gene	Description	Comps Up/Down	N	Mean FC (Range)
MET	met proto-oncogene (hepatocyte growth factor receptor)	6/0	202	4.54 (2.60 to 6.60)
TFF3	trefoil factor 3 (intestinal)	0/6	196	-22.04 (-63.55 to -3.80)
SERPINA1	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	6/0	192	15.84 (5.00 to 27.64)
EPS8	epidermal growth factor receptor pathway substrate 8	5/0	186	3.14 (2.10 to 3.80)
TIMP1	tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)	5/0	142	5.37 (3.20 to 10.31)
TGFA	transforming growth factor, alpha	4/0	165	6.18 (3.20 to 7.91)
QPCT	glutaminyl-peptide cyclotransferase (glutaminyl cyclase)	4/0	153	7.31 (3.40 to 11.67)
PROS1	protein S (alpha)	4/0	149	5.76 (3.40 to 7.39)
CRABP1	cellular retinoic acid binding protein 1	0/4	146	-11.54 (-24.45 to -2.20)
FN1	fibronectin 1	4/0	128	7.67 (5.20 to 10.30)
FCGBP	Fc fragment of IgG binding protein	0/4	108	-3.20 (-3.30 to -3.10)
TPO	thyroid peroxidase	0/4	91	-6.25 (-8.60 to -2.70)

Griffith OL, Melck A, Jones SJM, Wiseman SM. 2006. JCO. 24(31):5043-5051.

What's next? Tissue microarrays

- Two arrays (Dr. Sam Wiseman):
  - 100 Benign versus 105 Cancer patient samples
    - 57 markers stained
  - 12 differentiated vs. 12 undifferentiated samples
    - Matched samples from patients with extremely rare and aggressive Anaplastic cancer
    - A model for cancer progression
    - 62 markers stained

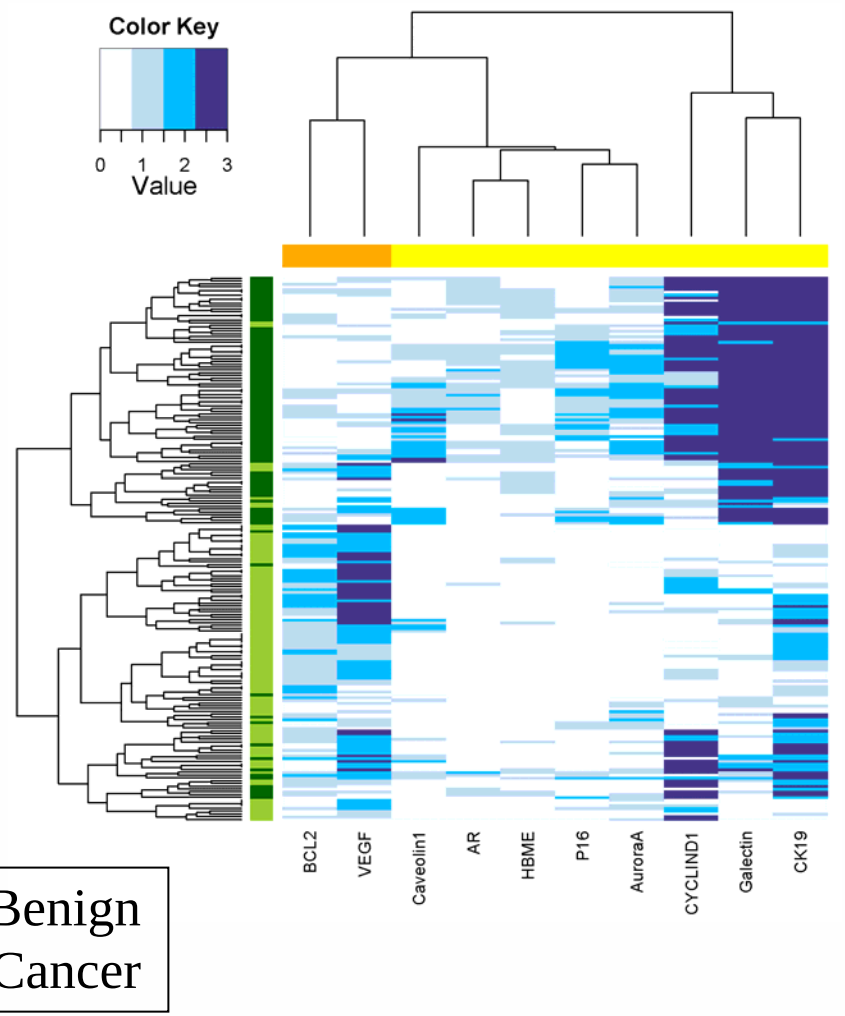
Methods: Tissue Array Construction



Benign versus cancer array: Results for top 25 markers

Marker	Benign Mean Rank	Malignant Mean Rank	Change	P-value	Corr. P-value	Variable Imp.
VEGF	130.3	65.3	Down	0.0000	0.0000	6.909
Galectin	59.1	139.6	Up	0.0000	0.0000	15.895
CK19	60.1	138.5	Up	0.0000	0.0000	13.942
AR	74.7	123.3	Up	0.0000	0.0000	5.048
AuroraA	68.1	123.2	Up	0.0000	0.0000	4.437
HBME	74.4	123.6	Up	0.0000	0.0000	5.309
P16	73.8	123.5	Up	0.0000	0.0000	4.174
BCL2	121.1	71.1	Down	0.0000	0.0000	2.383
CYCLIND1	67.0	115.5	Up	0.0000	0.0000	2.852
Caveolin1	77.5	119.1	Up	0.0000	0.0000	2.308
ECAD	120.2	75.9	Down	0.0000	0.0000	3.186
CYCLINE	77.1	118.0	Up	0.0000	0.0000	1.633
CR3	77.5	113.9	Up	0.0000	0.0000	1.045
Clusterin	79.6	117.0	Up	0.0000	0.0000	2.478
IGFBP5	79.0	112.2	Up	0.0000	0.0000	1.144
P21	81.0	113.4	Up	0.0000	0.0000	0.549
BetaCatenin	89.5	107.9	Up	0.0000	0.0000	0.295
IGFBP2	82.1	109.7	Up	0.0000	0.0001	1.051
Caveolin	78.8	109.0	Up	0.0001	0.0002	2.359
HER4	82.7	112.6	Up	0.0001	0.0003	1.273
TG	104.0	87.7	Down	0.0001	0.0003	1.268
CKIT	104.8	88.6	Down	0.0002	0.0004	0.810
S100	89.0	101.6	Up	0.0002	0.0004	0.230
KI67	86.9	101.6	Up	0.0003	0.0007	0.793
AuroraC	79.7	104.7	Up	0.0007	0.0015	1.059

TMA marker data can be used to attempt to classify benign vs. cancer patient samples



Random Forests classifier performance:

- overall accuracy=91.3%
- sensitivity=88.5%
- specificity=94.0%
- Misclassification:
  - 6 benign; 11 cancer

Thyroid cancer: Conclusions and future work

Conclusions:

- A significant number of genes are consistently identified by multiple expression profiling studies
- Both known and novel markers
- Preliminary IHC analysis on TMAs show promising results

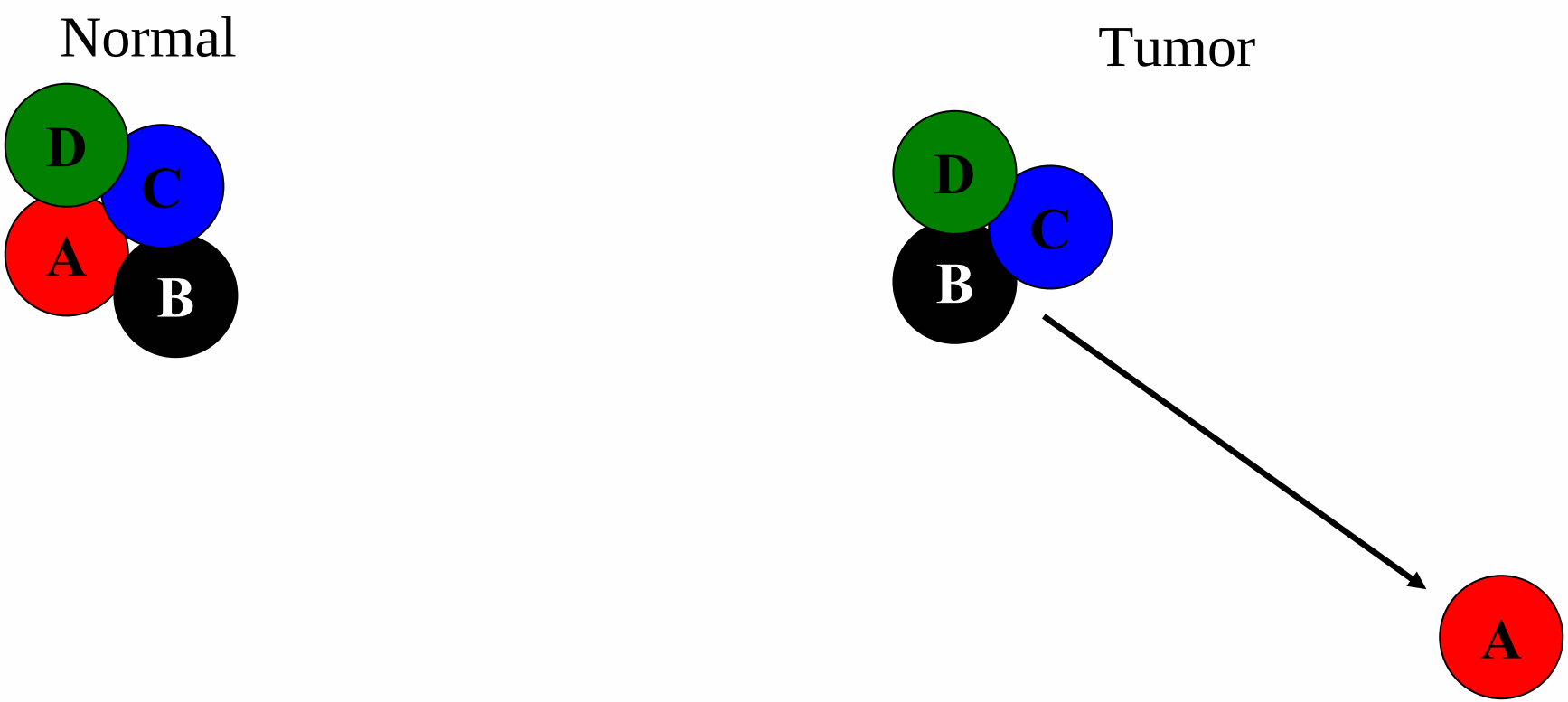
Future work:

- Addition of candidate genes from the meta-analysis to TMA analysis
- Development of a clinically useful classifier for thyroid tissue based on results of TMA

III) Differential coexpression in cancer

- Hypothesis: In some cases progression of cancer is mediated through changes in genetic regulatory regions that can be detected through gene expression studies and bioinformatics analyses.
- Specific hypothesis: Genes with significant changes in coexpression patterns will represent good candidates for regulatory changes
- Objective: Develop methods to assess differential coexpression.

Genes in coexpression space – differential coexpression



Difference in Mean correlation

Norm	Exp1	Exp2	Exp3	Exp4	Exp5	…
geneA	1.2	1.3	-1.4	0.1	2.2	…
geneB	1.3	1.3	-0.9	0.1	2.3	…
geneC	-1.2	1.0	0.1	0.5	1.4	…
…	…	…	…	…	…	…

Tumor	Exp1	Exp2	Exp3	Exp4	Exp5	…
geneA	11	35	2	4	50	…
geneB	12	35	0	3	47	…
geneC	0	10	4	15	20	…
…	…	…	…	…	…	…

Calculate all PCCs for each gene

Norm	geneA	geneB	geneC	geneD	…
geneA	NA	0.91	0.01	0.99	…
geneB	0.91	NA	-0.03	0.87	…

Tumor	geneA	geneB	geneC	geneD	…
geneA	NA	0.31	0.01	0.23	…
geneB	0.31	NA	-0.03	0.90	…

Find n nearest genes in normal and compare to tumor

Norm	geneD	geneB	geneX	geneY	…
geneA	0.99	0.91	0.90	0.89	…

Tumor	geneD	geneB	geneX	geneY	…
geneA	0.23	0.31	0.18	0.01	…

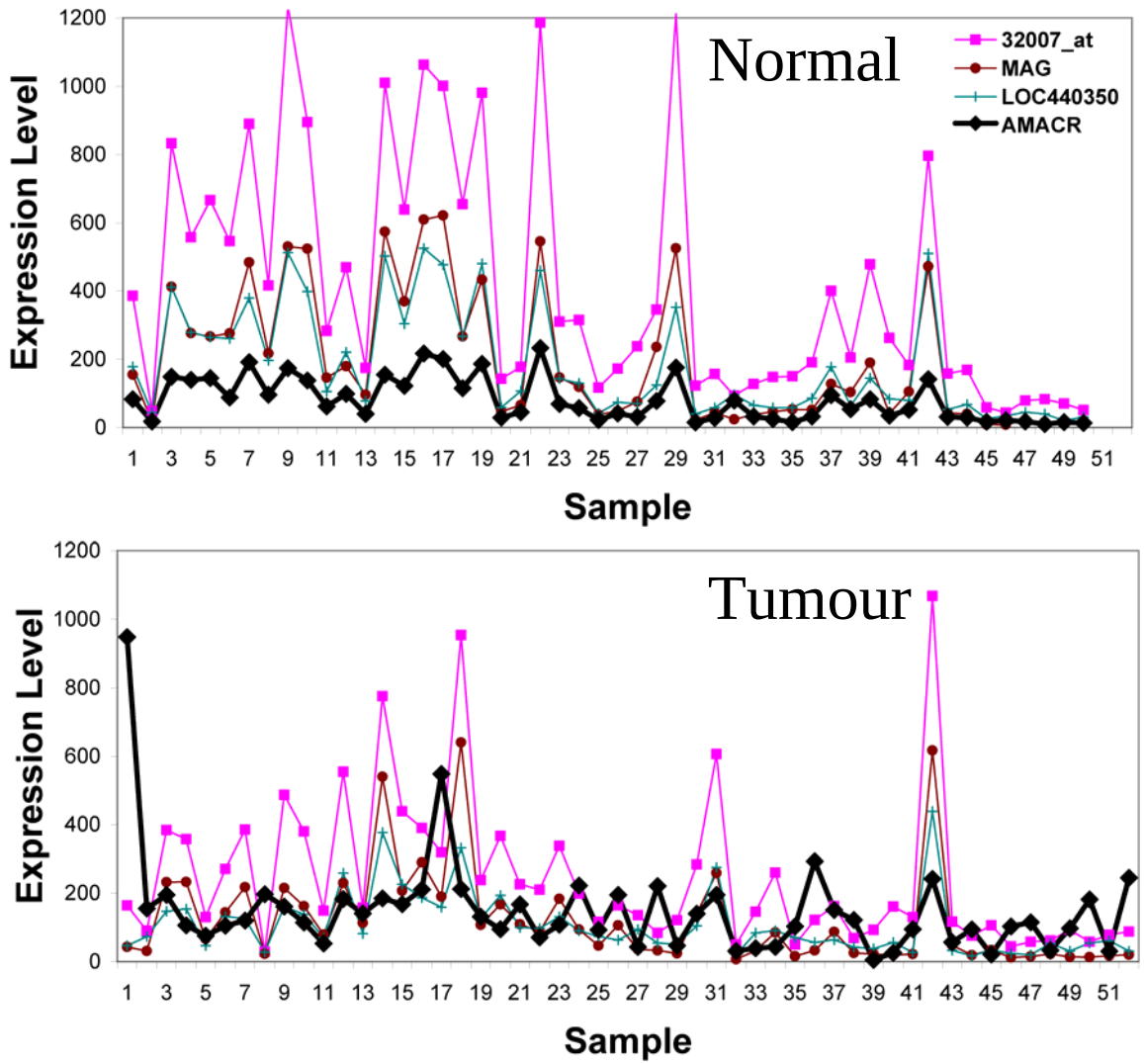
Calculate difference in mean PCC

Differential coexpression analysis

Expression Data

- Singh et al (2002)
- 52 prostate tumor
- 50 normal prostate
- Affymetrix U95Av2
- ~12,500 genes

An example of differential coexpression in prostate cancer (AMACR)



Candidate prostate cancer genes

Symbol	Comments
SLC26A3	Protein downregulated in colon adenoma and adenocarcinoma
CELSR1	Developmentally regulated, neural-specific gene which plays an unspecified role in early embryogenesis
AMACR	Proven biomarker that can help distinguish cancer from benign cells, with high sensitivity and specificity for prostate carcinoma.
PEX5	peroxisomal biogenesis factor
G1P2	Induced by camptothecin and Retinoic Acid in human tumor cells
SOX9	Overexpression results in suppression of growth and tumorigenicity in the prostate tumor cell line M12
ATP6V1E1	ATPase
LOC153561	function unknown
SEMG1	Interacts with PSA
MGC5576	function unknown
RAD51C	mRNA and protein levels elevated ~2- to 5-fold in malignant prostate cell lines
RAD23B	Highly expressed in the human testis and in ejaculated spermatozoa. Down-regulated during early apoptosis in human hepatoma cells exposed to Paeoniae Radix extract in vitro
SEMG2	Interacts with PSA
SNX4	not well characterized (only 4 pubmed)
DLGAP2	putative tumor suppressor gene. Chromosomal region (8p23.2) frequently deleted in prostate cancer.
TFDP2	Differential expression shown in some cancer cell lines
HGF	c-Met/HGF receptor have roles in prostate neoplasm progression
PEX10	peroxisomal biogenesis factor
ABL1	Important in leukemia - Bcr-Abl translocation
GSPT1	Overexpressed in gastric cancer
DNAJA2	function unknown
C7orf24	function unknown
GRM5	glutamate receptor, metabotropic

Cancer
Prostate Cancer

Summary

- Differential coexpression analysis represents a useful and complementary method to traditional differential expression methods for identifying potentially relevant cancer genes.
- Such genes may represent novel prostate cancer genes and would make good candidates for regulatory mutation analysis.

Acknowledgements

Supervisor
Dr. Steven Jones

Coexpression analysis
Yuliya Prychyna
Maggie Zhang
Yan Jia Pan
Erin Pleasance
Debra Fulton

Differential coexpression
Erin Pleasance
Malachi Griffith

Thyroid meta-analysis
Sam Wiseman
Adrienne Melck

CIHR IRSC

Michael Smith Foundation for Health Research

GENOME Sciences Centre

BC Cancer Agency
CARE & RESEARCH