Meta-Analysis of Thyroid Cancer Expression Profiling Studies
Identification of Most Promising Biomarkers For Tissue Microarray Analysis

Obi L Griffith[1], Sam M Wiseman[2], and Steven JM Jones[1]

1. Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency
2. Genetic Pathology Evaluation Center, Prostate Research Center of Vancouver General Hospital & British Columbia Cancer Agency

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

1. Abstract

Objective and Design
It is estimated that 5-10% of the population will develop a clinically significant thyroid nodule during their lifetime. In one third or more of these patients, pre-operative diagnoses by needle biopsy are inconclusive. In many cases, a patient will undergo unnecessary surgery for what ultimately proves to be a benign lesion. Thus, there is a clear need for improved diagnostic tests to distinguish malignant from benign samples. The recent development of tissue microarray techniques should allow the rapid evaluation of potential new markers. However, researchers are faced with an overwhelming number of potential markers from numerous thyroid cancer profiling and classification studies. We present a systematic and comprehensive selection of potential thyroid cancer biomarkers from published studies by meta-analysis for use in tissue microarray analysis (TMA).

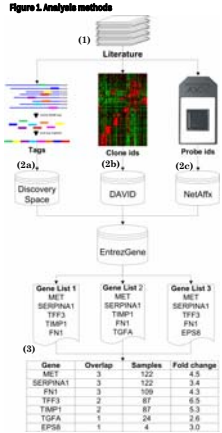
Materials & Methods
A total of 21 published studies were identified from the literature. Each study reported differentially expressed genes for at least one comparison type (eg. Normal versus PTC). The following information was recorded wherever possible: unique identifier (probe/tag/accession), gene name, gene description, gene symbol, tissue types compared, fold change (magnitude and/or direction), p-value, validation (RT-PCR, IHC, Western), and pubmed ID. Whenever possible, the mapping of clone accession, probe id or SAGE tag was updated using NCBI mapping files, Affymetrix annotation files, and the DiscoverySpace SAGE tag mapping tool respectively. A heuristic system was devised to identify the most promising markers, taking into consideration the number of studies reporting the potential marker, sample sizes and fold-change.

Results
In total, 1,785 potential biomarkers were identified (not considering overlap) from 21 gene expression studies considering 34 different tumour or tissue type comparisons. This resource allows the identification of markers that consistently differentiate one tumour/tissue type (e.g. papillary thyroid cancer) from others (e.g. normal, benign, other thyroid cancer subtypes). From this analysis, an informed selection of markers will be made for tissue microarray analysis, optimizing the chance of finding clinically relevant markers.

Conclusion
Bioinformatics meta-analysis and tissue microarray analysis represents a powerful approach to identifying new thyroid cancer biomarkers. Such markers could prove invaluable for the diagnosis and prognosis of tumours in the clinical setting. The meta-analysis of published thyroid studies should prove a useful resource for many thyroid cancer researchers.

2. Methods

Figure 1. Analysis methods



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

Fig 1: **(1)** Lists of differentially expressed genes were collected and curated from published studies. Each study consists of one or more comparisons between pairs of conditions (e.g. PTC vs. norm). The following information was recorded wherever possible: Unique identifier (probe, tag, accession); gene description; gene symbol; comparison conditions; sample numbers for each condition; fold change; direction of change. **(2)** SAGE tags, cDNA clone ids and Affymetrix probe ids were mapped to Entrez Gene using: (a) the DiscoverySpace software package[1]; (b) the DAVID Resource[2]; (c) the Affymetrix annotation files[3]. **(3)** Genes are ranked according to several criteria in the following order of importance: (a) number of comparisons in agreement (ie. listing the same gene as differentially expressed and with a consistent direction of change); (b) total number of samples for comparisons in agreement; and (c) average fold change reported for comparisons in agreement.

Table 1: Lists all abbreviations used to describe the samples and conditions compared in the various studies.

Table 2: A total of 34 comparisons were available from 21 studies, utilizing at least 10 different expression platforms. Platforms can be generally grouped into cDNA arrays (blue), oligonucleotide arrays (purple) and SAGE (pink). The numbers of 'up-/down-regulated' genes reported are for condition 1 relative to condition 2 for each comparison as provided. Only genes that could be mapped to a common identifier were used in our subsequent overlap analyses (see Analysis methods).

Table 1. Abbreviations for sample descriptions

ACL	Anaplastic thyroid cancer cell line
APTN	Autonomously functioning thyroid nodules
ATC	Anaplastic thyroid cancer
CTN	Cold thyroid nodule
FA	Follicular adenoma
FCL	Follicular carcinoma cell line
FTC	Follicular thyroid carcinoma
FVPTC	Follicular variant papillary carcinoma
GT	Goiter
HCC	Hurthle cell carcinoma
HN	Hyperplastic nodule
M	Metastatic
MACL	Anaplastic thyroid cancer cell line with metastatic capacity
Norm	Normal
PCL	Papillary carcinoma cell line
PTC	Papillary thyroid carcinoma
TCVPTC	Tall-cell variant PTC
UCL	Undifferentiated carcinoma cell line

3. Thyroid cancer expression data

Table 2. Thyroid cancer profiling studies included in analysis

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)	PTC(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	PTC(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
			PTC(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	PTC(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**

4. Overlap analysis results

Table 3. Comparison groups analyzed for overlap

Overlap analysis group	condition set 1	Condition set 2	comps	# genes (multi-study)
All	Any	Any	34	
Cancer vs. non-cancer	ACL, ATC, FCL, PTC, FVPTC, HCC, M, MACL, PCL, PTC, TCVPTC, UCL	AFTN, CTN, FA, GT, HN, Norm	21	755 (107)
Cancer vs. normal	ACL, ATC, FCL, PTC, FVPTC, HCC, M, MACL, PCL, PTC, TCVPTC, UCL	Norm	12	478 (53)
Cancer vs. benign	ACL, ATC, FCL, PTC, FVPTC, HCC, M, MACL, PCL, PTC, TCVPTC, UCL	AFTN, CTN, FA, GT, HN	8	332 (38)
Normal vs. benign	Norm	AFTN, CTN, FA, GT, HN	3	19 (1)
Papillary cancer vs. non-cancer	FVPTC, PCL, PTC, TCVPTC	AFTN, CTN, FA, GT, HN, Norm	12	503 (82)
Papillary cancer vs. normal	FVPTC, PCL, PTC, TCVPTC	Norm	8	369 (49)
Papillary cancer vs. benign	FVPTC, PCL, PTC, TCVPTC	AFTN, CTN, FA, GT, HN	4	183 (13)
Papillary cancer vs. other	FVPTC, PCL, PTC, TCVPTC	Any other	15	528 (107)
FVPTC vs. other	FVPTC	Any other	2	157 (0)
PTC vs. FA	PTC	FA	6	222 (3)
Follicular cancer vs. other	FTC, FCL	Any other	10	403 (15)
Aggressive cancer vs. other	ACL, ATC, M, MACL	Any other	4	145 (4)
Anaplastic cancer vs. other	ACL, ATC, MACL	Any other	3	91 (6)

Figure 2. Gene overlap for cancer vs. non-cancer analysis

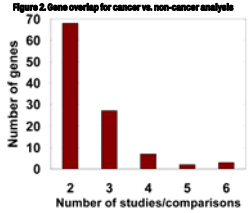


Table 3: Each overlap analysis group defines an artificial group of comparisons for which gene overlap was analyzed. In all groups considered except for one, we identified one or more genes that were reported in two or more studies. For example, the "cancer vs. non-cancer" group (highlighted) includes all comparisons between what we would consider 'cancer' (as in condition set 1) and 'non-cancer' (as in condition set 2). In this case, 21 comparisons met the criteria and produced a list of 755 potential cancer markers, 107 of which were identified in multiple studies. These 'multi-study cancer versus non-cancer markers' are summarized further in figures 2-3 and tables 4-5.

Fig. 2: A breakdown of the 107 genes found in multiple studies for the cancer versus non-cancer analysis. Some genes were observed in as many as six studies.

4. Overlap analysis results (cont'd)

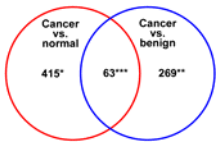


Fig 3. Of the 478 genes in the cancer/normal comparison and 332 genes of the cancer/benign group, a total of 63 genes were found in both.

Table 4: shows a partial list (genes identified in 4 or more comparisons) from the cancer vs. non-cancer analysis. A complete table for this group and all others are available as supplementary data (www.bcgsc.ca/bioinfo/gethytroid/).

Table 5: Of the 107 genes with multi-study confirmation from the cancer versus non-cancer overlap analysis group, 102 were present in the Gene Ontology set of 15240 human genes. From this list, a total of 12 GO terms were found to be statistically over-represented: 3 biological process (B); 3 cellular component (C); and 6 molecular function (F). The 'obs/total' column shows the number of genes from the list found associated with each GO term over the total number of genes annotated to that term in GO. The p-value was calculated using the BINGO[4] plugin for Cytoscape[5]. P-values are for a hypergeometric test, corrected with a Benjamini & Hochberg False Discovery Rate (FDR) correction, and a cut off of 0.05 applied to the result.

Table 4. Cancer versus non-cancer genes identified in 4 or more independent studies

Gene	Description	Comp's (Up/Down)	N	Fold Change
MET	met proto-oncogene (hepatocyte growth factor receptor)	6/0	202	3.03
TFF3	trefoil factor 3 (intestinal)	0/6	196	-14.70
SERPINA1	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	6/0	192	15.84
EPS8	epidermal growth factor receptor pathway substrate 8	5/0	186	3.15
TIMP1	tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor)	5/0	142	5.38
TGFA	transforming growth factor, alpha	4/0	165	4.64
QPCT	glutaminyl-peptide cyclotransferase (glutaminyl cyclase)	4/0	153	7.31
PROS1	protein S (alpha)	4/0	149	4.32
CRABP1	cellular retinoic acid binding protein 1	0/4	146	-11.55
FN1	fibronectin 1	4/0	128	7.68
FCGBP	Fc fragment of IgG binding protein	0/4	108	-2.41
TPO	thyroid peroxidase	0/4	91	-4.69

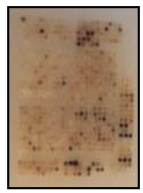
Table 5. GO analysis of multi-study genes from the cancer versus non-cancer overlap analysis group

GO term	Ontology	obs/total	p-value	Genes in test set
extracellular region	C	23/1093	4.09E-04	CYR61, CHI3L1, TIMP1, TNFRSF11B, ADM, GPC3, LOX, PLAU, TFF3, LGALS3, TGFA, CCL21, SPINT1, SERPINA1, RNASE1, PROS1, TNC, DPT, MATN2, IGFBP3, COL9A3, BMP1, FN1
cadmium ion binding	F	3/4	4.09E-04	MT1F, MT1E, MT1A
thyroid hormone generation	P	3/5	6.78E-04	DIO2, DIO1, TPO
thyroid hormone metabolism	P	3/6	1.01E-03	DIO2, DIO1, TPO
selenium binding	F	3/7	1.41E-03	DIO2, SELENBP1, DIO1
hormone metabolism	P	5/48	1.97E-03	HSD17B6, DIO2, DIO1, TPO, ADM
MAP kinase phosphatase activity	F	3/10	3.40E-03	DUSP6, DUSP4, DUSP1
thyroxine 5'-deiodinase activity	F	2/3	1.16E-02	DIO2, DIO1
copper ion binding	F	4/41	1.25E-02	MT1F, MT1E, MT1A, LOX
extracellular matrix (sensu Metazoa)	C	9/342	3.45E-02	CHI3L1, TIMP1, TNC, DPT, MATN2, COL9A3, GPC3, LOX, FN1
extracellular matrix	C	9/347	3.49E-02	CHI3L1, TIMP1, TNC, DPT, MATN2, COL9A3, GPC3, LOX, FN1
retinoic acid receptor activity	F	2/6	3.82E-02	RXRG, RARA

5. Conclusions and Future work

Conclusions:
> A significant number of genes are consistently identified in the literature as differentially expressed between different thyroid tissue and tumour subtypes
> These consistent genes represent a useful starting point for a large-scale tissue microarray analysis to identify useful prognostic and diagnostic markers in a clinical setting

Future work:
> Probability simulation to assess significance
> Meta-analysis starting from raw expression data
> Selection of final candidates for tissue microarray (TMA) analysis
> Development of classifier for thyroid tissue based on results of TMA



6. Acknowledgments

funding | Natural Sciences and Engineering Council of Canada (OG); Michael Smith Foundation for Health Research (OG, SW, and SJ); Canadian Institutes of Health Research (OG); BC Cancer Foundation

references | 1. Varhol et al, unpublished, http://www.bcgsc.ca/discoveryspace/; 2. Dennis et al. 2003, http://david.abcc.ncifcrf.gov/; 3. Affymetrix, http://www.affymetrix.com/support/index.affx; 4. Maere et al. 2005, http://www.psb.ugent.be/cbd/papers/BiNGO/; 5. Shannon et al. 2003, http://www.cytoscape.org/

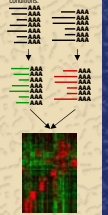
SAGE

Serial analysis of gene expression (SAGE) is a method of large-scale gene expression analysis that involves sequencing small segments of expressed transcripts ("SAGE tags") in such a way that the number of times a SAGE tag sequence is observed is directly proportional to the abundance of the transcript from which it is derived.

A description of the protocol and other references can be found at www.sagenet.org.

cDNA Microarrays

cDNA Microarrays simultaneously measure expression of large numbers of genes based on hybridization to cDNAs attached to a solid surface. Measures of expression are relative between two conditions.



For more information, see www.microarrays.org.

Affy Oligo Arrays

Affymetrix oligonucleotide arrays make use of tens of thousands of carefully designed oligos to measure the expression level of thousands of genes at once. A single labeled sample is hybridized at a time and an intensity value reported. Values are the based on numerous different probes for each gene or transcript to control for non-specific binding and chip inconsistencies.

For more information, see www.affymetrix.com.