Chapter 8


**Discussion**

8.1 Serial Analysis of Gene Expression

The major advantage of SAGE over other gene expression profiling techniques is its comprehensiveness. Unlike other techniques SAGE does not only study a set of genes based on prior availability of transcript information. Because of this feature SAGE makes it possible to study the expression levels of all genes expressed, including genes that are not yet characterized. The essential step towards a biologically relevant SAGE expression profile is TAG-to-GENE identification. Typically, this is performed by comparing a database of SAGE generated tags to a database of tags generated from a collection of transcripts like the UniGene database [www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene]. UniGene is a system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene. UniGene tags are generated automatically by screening all sequences present in a UniGene cluster for a 3’signature. Commonly, a polyadenylation signal combined with a polyA-tail is used for such a signature, as in NCBI’s SAGEmap program [1]. Sequences with a valid 3’signature are screened for NlaIII-restriction sites (CATG) and the flanking sequence 10bp downstream of the most 3’CATG-site is considered to be the SAGE tag corresponding to this particular transcript.

In theory, transcripts from a single gene are represented by one and the same SAGE tag. In practice, UniGene clusters can have up to several hundred tags assigned to them. This can be caused by several reasons. Sequence errors in EST sequences, wrongly assigned sequences to a cluster, internal polyA sequences, alternative splicing or polymorphisms. As is shown in Chapter 4 of this thesis, the thyroglobulin gene can generate 3 different tags resulting from alternative polyadenylation cleavage sites. Since this particular type of alternative splicing takes place in the 3’region of the mRNA where a SAGE tag is typically located, this can lead to TAG-to-GENE identification problems. When looking at the most recent GENE-to-TAG identification for the thyroglobulin UniGene cluster (accession number Hs.305916) 1317 sequences generate 247 different tags. Although the thyroglobulin gene spans 48 exons and alternative splicing is abundant [2, 3], to the best of our knowledge only 3 tags represent all possible thyroglobulin transcripts. The large number of possible tags generated for this UniGene cluster makes it impossible to perform a reliable TAG-to-GENE identification only to be solved ultimately by hand curation. This problem does not only occur in this particularly example, it can be observed in almost all UniGene clusters.

A recent program called SAGE Genie tries to solve some of these problems in its GENE-to-TAG identification [cgap.nci.nih.gov/SAGE][4]. This program however only gives the most reliable tag for a given cluster. Although this might seem very handy, alternative transcripts with a different tag are difficult to analyse. The thyroglobulin cluster in this case only identifies 1 out of the 3 possible SAGE tags.

After all of these problems are solved, every gene will be represented by different tags and vice versa because it is known that single genes can be represented by several transcripts generating different SAGE tags. Which transcripts are expressed depends on the type cell or tissue and can be different in identical cells under different (experimental) conditions. In the human SAGEmap database 152 SAGE libraries from a range of normal tissues, tumours and cell lines generated a total of 2,518,369 tags. From this collection 475,694 unique tags could be distinguished corresponding to 83,789 UniGene clusters. Even when incorporating a 10% error rate because of sequencing errors, on average 5 tags per gene are expected.
to be found. When considering that the latest estimates on the total number of human genes is much lower (30-35,000) more than 10 tags per gene can be expected [5, 6]. So, either the number of human genes is underestimated by at least a factor 2, or alternative splicing is even more frequent than expected. To achieve a reliable TAG-to-GENE identification every possible tag needs to be assigned to a cluster of transcripts corresponding to a single gene. In this way, SAGE generated tags can be assigned with high certainty to a single gene. Further improvements that lead to a definitive list of TAG-to-GENE identifications can be:
- A strictly non-redundant transcript database
- Removal of sequence errors from the transcript database
- Incorporation of SNPs in tag-containing 3' sequences
- Hand curation of TAG-to-GENE identifications

The last point seems to be the most promising since it can attain a final decision on every TAG-to-GENE identification and avoids the use of intricate logarithms which will never be able to assign tags with a 100% accuracy, simply because it's design is based on similarity of characteristics and transcript characteristics are by definition different per gene. The feasibility of hand-curation is real. Experience in NCBI's RefSeq project has shown that hand-curation is in the end the only way to annotate genes. The labour intensity is high but will finally end in a comprehensively annotated list of SAGE tags. Since many researchers have been doing hand-curation for their subsets of interesting tags already, pooling these efforts globally will leave only a small number of tags to be assigned to human genes still uncharacterised at this time. When a definitive list is formed in this way a final parameter should be incorporated; tags that cannot be assigned to a gene transcript with a high level of certainty should be annotated as NoMatch tag, since an erroneous TAG-to-GENE identification can greatly complicate the translation of SAGE data to biologically relevant data. As long as the human transcriptome is incomplete, NoMatch tags are an important tool to characterize novel genes [Chapter 6][7]. To relate both Match and NoMatch tags to specific tissue expression or co-expression with specific pathways, the Tissue Preferential Expression (TPE) analysis can be used [8]. Using the TPE algorithm one can assign a value to a given tag giving information about the specificity of it for a certain tissue or experimental condition. Chapters 5 & 7 show examples of this analysis.

8.2 SMAP31

SMAP31 is expressed in placenta, brain and lung as well as in thyroid and heart. As is discussed in Chapter 6, the expression in normal thyroid is low and an overexpression of SMAP31 can be observed in thyroid carcinoma. Furthermore, normal tumour-surrounding thyroid tissue shows a slight overexpression of SMAP31. In spite of the overexpression in thyroid carcinoma, normal thyroid cells do express SMAP31, making this gene difficult to use as a diagnostic or prognostic marker. However, the fact that SMAP31 encodes a homeobox transcription factor can give valuable information about changes in thyroid-specific gene expression in normal thyroid and thyroid carcinoma.

Recent studies in heart development have investigated the protein HOP (homeobox only protein) [9, 10]. HOP is identical to SMAP31. HOP is located in the cell nucleus where it acts as a co-repressor. In spite of its conserved homeobox domain, HOP does not bind to DNA, instead, HOP modulates transcriptional activation of gene promoters. In heart, HOP interacts with SRF (serum response factor) to inhibit SRF-dependent cardiac gene expression (Figure 1).
Expression of HOP is induced by another transcription factor Nkx2.5, which is an inducer of cardiac-related gene expression and is reported to play an essential role in the early embryonal development of heart, pharynx and thyroid [11, 12, 13]. Furthermore, Nkx2.5/Nkx2.6 double mutants have been reported to show absence of thyroid development in mice [14]. In mice development, HOP is expressed in embryonal heart and in placental trophoblasts. HOP-/⁻ mice show cardiac malformations with a partially penetrant lethal phenotype in the foetal period. At E11.5 10% of mice foetuses show arrested development with a hypocellular and thin myocardium. Foetuses that survive beyond E11.5 have an increased cardiac size as a result of cardiomyocyte hyperplasia. Both defects seem to be a result of a disturbed equilibrium between proliferation and differentiation of cardiomyocytes. HOP is a modulator of cardiac growth and development. The role of SMAP31/HOP in thyroid embryogenesis and foetal development is unknown but the embryonic development of the heart and the thyroid show similarities in the expression of early transcription factors. Nkx2.5 is expressed in embryonal heart as well as in embryonal thyroid in the mouse from E7.5 in gestation. Essential thyroid-specific transcription factors Nkx2.1, FKHL15 and Pax8 regulate thyroid development (migration and proliferation) from E9 but also induce thyroid-specific gene expression (e.g. thyroglobulin, TPO) for differentiation at E11.5 (Figure 2). From studies into the molecular background of thyroid agenesis in man and mouse it is evident that there must be more, currently unidentified, factors involved in development of the thyroid gland [15].
The equilibrium between proliferation and differentiation in thyroid development could be regulated by SMAP31. It is tempting to speculate that SMAP31 can modulate the thyroid-specific expression via the thyroid-specific transcription factors Nkx2.1, FKHL15 and Pax8. This mechanism can work, like in the heart, via competitive binding of SMAP31 to these transcription factors directly or via an SRF-like thyroid-specific co-factor regulating these transcription factors. Binding assays to study the ability of SMAP31 to bind to thyroid specific transcription factors will elucidate this. Figure 3, on page 128, summarizes the postulated hypothesis indicating a central role for SMAP31 in the modulation of thyroid-specific gene expression in embryonal development. The observation that SMAP31 is overexpressed in thyroid carcinoma could also imply a role for this co-repressor in dedifferentiation processes in the progression of thyroid tumours.

In light of the pivotal role that HOP plays in the development of the heart it has been suggested that mutations in the HOP gene could account for congenital cardiac malformations. At the same time several reports have linked congenital cardiac malformations with thyroid dysgenesis [16, 17, 18, 19]. In patients with congenital hypothyroidism (CH) as a result of thyroid dysgenesis, cardiac malformations are more prevalent than in the normal population. The similarities of heart and thyroid embryonal and foetal development, especially the role of SMAP31 in these processes makes the SMAP31 gene an interesting candidate gene to screen for mutations in patient with a combined heart and thyroid phenotype.
Chapter 8

8.3 ECM1

The association of ECM1 with hyaluronan-related molecules like hyaluronidase and RHAMM, as described in Chapter 7, suggests a role for ECM1 in remodelling of the extracellular matrix surrounding epithelial thyroid tumour cells. All cells in solid tissue are surrounded by an extracellular matrix, which is composed of proteins and polysaccharides. Different tissues have different compositions of their extracellular matrix.

In general, extracellular matrix has two main functions:
- The stable positioning of cells in tissues through cell matrix adhesion
- Extracellular communication that is assisted by cell adhesion molecules

The two main groups of molecules that make up the basic extracellular matrix are complex chains of carbohydrates and polysaccharides joined to protein (glycoproteins). Embedded in this structure can be various types and amounts of structural and insoluble collagen fibres. The extracellular matrix has long been thought of as an inert entity providing structural support for tissues. Recently several reports have identified other functions of the extracellular matrix related to signal transduction, cell growth and embryogenesis but also to pathological processes such as transformations and metastasis [20].

The glycosaminoglycan molecule hyaluronan is a major compound of the epithelial extracellular matrix. Hyaluronidases catalyse its degradation. The glycoprotein aggrecan is linked to hyaluronan, forming a matrix like structure. Aggrecan has a double-loop structure for hyaluronan binding. The structure of the epithelial extracellular matrix is depicted in Figure 4.
Invasion of cancer cells can be viewed as cellular motility coupled to regulated adhesion and detachment from the extracellular matrix by proteolysis of extracellular matrix molecules. Tumour cells can subsequently reattach to an inappropriate extracellular matrix in, for example, lymph nodes.

All human epithelial tumours are surrounded by hyaluronan-enriched extracellular matrix. The proteinase hyaluronidase is elevated in many tumours. The elevated levels of hyaluronidase in urine are reliable markers for the presence and grade of bladder carcinoma [21]. Cell surface receptor RHAMM associates with the cytoskeleton via hyaluronan and is involved in the ras signal transduction pathway. RHAMM receptors connect to the cytoskeleton and co-regulate cell motility and migration, possibly in concert with integrin receptors [22]. CD44 is another hyaluronan binding transmembrane molecule responsible for cell-matrix adhesion. Like RHAMM it is also able to transduce growth factor signalling. Reduced levels of CD44 are a prognostic factor for metastasis and unfavourable outcome of thyroid cancer [23]. As is indicated in figure 4, the downregulation of hyaluronan-binding molecules like CD44 and RHAMM together with the elevated levels of hyaluronidase can induce extracellular matrix breakdown and promote tumour invasion. The observed downregulation of RHAMM combined with the upregulation of hyaluronidase by ECM1 in Chapter 7 points to a role for ECM1 in the breakdown of the epithelial extracellular matrix. The exact role of ECM1 and hyaluronan in the remodelling of the extracellular matrix needs to be studied to be able to predict whether the overexpression of ECM1 in thyroid and others tumours can be associated with extracellular matrix related tumour progression.
8.4 Applications of SAGE in thyroid cancer

In thyroid cancer, the lack of powerful diagnostic markers that distinguish between benign and malignant thyroid disease cause many patients to undergo unnecessary surgery. Although the general prognosis of thyroid carcinoma is favourable, subgroups are at risk for recurrent disease, metastasis or death. Because the identification of these groups is difficult, almost all patient will undergo intensive initial therapy, which may not be necessary in all patients when suitable risk factors are available.

Available clinical and histopathological diagnostic and prognostic tools try to divide patients into different groups according to clinical symptoms, tumour sub-type, prognosis and therapy. Several staging systems can be used for this purpose. This way, standardization of patient care can be achieved. In all human cancers cases occur which are difficult to assign to pre-specified subgroups. Diagnosis in this case is difficult and prognostic information can only be given with great care.

Advances in molecular biology have improved the knowledge on molecular mechanisms behind tumorigenesis. The activation of small GTPases like ras and gsp is described as the underlying mechanism for many tumours from different tissues, including thyroid. However, not all tumours can be explained by mutations in ras and gsp as is described in Chapter 1. The available information on the genetic changes associated with development and progression of thyroid follicular neoplasia has no practical value for clinical diagnosis and does not provide any guidance for prognosis or therapy.

Recent technological breakthroughs making expression profiling of whole genomes possible make it feasible to link molecular knowledge to clinical applications. Microarray profiling in breast cancer has resulted in the identification of gene clusters able to predict clinical behaviour [24]. SAGE profiling in lung cancer has identified a novel gene PGP9.5 which not only correlates with lung carcinoma, but has prognostic value in showing increasing expression levels as the disease progresses to advanced stages [25]. Another SAGE study has revealed that the PRL-3 gene is important for colorectal cancer metastasis and provides a new therapeutic target [26]. In general, the increasing databases containing expression profiling data from disease studies performed all over the world will eventually give us great insight in the behaviour of human cancer [27]. SAGE expression profiles of thyroid carcinoma presented in this thesis have identified several genes, both already characterised genes and novel genes that augment the diagnostic and prognostic scope for thyroid cancer.

Techniques like microarray and SAGE can in the future be used for the expression profiling of individual tumours. Combined with classic diagnostic tools this can lead to a set of highly informative clinical parameters for every type of cancer, which can diagnose with high accuracy, predict disease outcome and select therapy tailored to individual patients.
References


