From the Institute of Environmental Medicine,
Division of Molecular Toxicology
Karolinska Institutet, Stockholm, Sweden

CANCER BIOMARKER DISCOVERY BY IN VITRO SYSTEMS BIOLOGY

Rebecca Ceder

Stockholm 2011

Till Mamma och Pappa

ABSTRACT

This Thesis was made with the intention to mechanistically assess and further develop a multi-stage cell line-based (*in vitro*) model for oral cancer development. Efforts of establishing additional tumor cell lines for expanding the model were coupled with the application of systems biology technologies for characterization of the three entities of the start-up model, including: 1) normal, 2) immortal and non-tumorigenic, versus 3) immortal and tumorigenic stages. Omics data integration from assessment of cell lines as unique entities, and model-driven *in vitro* manipulations formed the basis for construction of two bioinformatics-based pipelines for this task. Altered phenotypic and genotypic characteristics and the event of non-functional cell differentiation (a hallmark of cancer development) was analyzed broadly among the transformed stages of the model relative the normal counterpart, testing the overarching hypothesis that thorough analysis of cell line data might contribute clinically useful tumor biomarkers potentially hidden in existing genome-wide assessments of clinical tissue samples.

The separate papers forming the Thesis, in order, generated: 1) a review of existing data from the start-up model under a selected standardized serum-free condition, 2) an omics-integrative tumor biomarker discovery pipeline based on the start-up model, 3) a model-driven tumor biomarker discovery pipeline based on assessment of influences of confluency (high cell density and cell-to-cell contact) in the seemingly most differentiation-deficient cell line in the start-up model, 4) a novel tumor cell line applicable to expand the number of serum-free entities of the model, 5) an expanded model-driven tumor biomarker discovery pipeline based on assessment of serum-induced influences of the extended model (now with four entities), and finally, 6) an analysis of the novel cell line under a further expanded omics-integrative tumor biomarker discovery pipeline. The overall results included broad description of the multiple alterations at gene, pathway and ontology levels that coupled with the transformed phenotypes and non-functional cell differentiation in the cell line models. The bioinformatics-driven assessment using overall six different processing tools of differential expression of 44 proteins and thousands of transcripts from these analyses suggested multiple potential biomarker signatures in head and neck squamous cell carcinoma. Overall, five *in vitro*-based signatures could be validated for clinical significance in independent data from tumor tissue analysis, including multiple oral and non-oral patient data sets as well as body-wide transcriptomics and proteomics expression databases. The taken approaches elucidated basic mechanisms of cell transformation while simultaneously generating paradigms/protocols generally applicable to cancer biomarker discovery. Proving the hypothesis under testing, the results show that the *in vitro*-derived biomarkers are complementary, often with superior accuracy, to those generated from direct assessment of cancer tissue specimens. Overall, the application of technologies and methods as described possibly generated a first description of an "*in vitro* systems biology model of oral cancer development" with potential for wide further application in experimental and translational research.

LIST OF PUBLICATIONS

This thesis is based on the following publications, referred to in the text by their Roman numerals.

I. **Ceder, R.,** Merne, M., Staab, C.A., Nilsson, J.A., Höög, J.O., Dressler, D., Engelhart, K., and Grafström, R.C. 2007. The application of normal, SV40 T-antigen-immortalised and tumor-derived oral keratinocytes, under serum-free conditions, to the study of the probability of cancer progression as a result of environmental exposure to chemicals. *Altern Lab Anim* 35:621-39

II. Staab, C.A., **Ceder, R.,** Jägerbrink, T., Nilsson, J.A., Roberg, K., Jörnvall, H., Höög, J.O., and Grafström, R.C. 2007. Bioinformatics processing of protein and transcript profiles of normal and transformed cell lines indicates functional impairment of transcriptional regulators in buccal carcinoma. *J Proteome Res* 6:3705-17

III. **Ceder, R.,** Haig, Y., Merne, M., Hansson, A., Zheng, X., Roberg, K., Nees, M., Iljin, K., Bloor, B.K., Morgan, P.R., Fadeel, B., and Grafström, R.C. 2011. Differentiation-promoting culture of competent and non-competent keratinocytes identifies biomarkers for head and neck cancer. *Am J Pathol*, in press

IV. Roberg, K., **Ceder, R.,** Farnebo, L., Norberg-Spaak, L., and Grafström, R.C. 2008. Multiple genotypic aberrances associate to terminal differentiation-deficiency of an oral squamous cell carcinoma in serum-free culture. *Differentiation*. 76:868-80

V. **Ceder, R.,** Merne, M., Jerhammar, F., Nees, M., Roberg, K., and Grafström, R.C. Serum-exposure of normal and transformed keratinocytes associates novel genes to head and neck cancer, manuscript

VI. **Ceder, R.,** Staab, C.A., Roberg, K., Nees, M., Jörnvall, H., Fadeel, B., Höög, J-O., and Grafström, R.C. Gene signatures generated from a unique serum-free medium-derived tumor line accurately classifies normal and tongue squamous cell carcinoma tissue, manuscript

PUBLICATIONS NOT INCLUDED IN THE THESIS

1. Staab, C.A., **Ceder, R.,** Roberg, K., Grafström, R.C., and Höög, J.O. 2008. Serum-responsive expression of carbonyl-metabolizing enzymes in normal and transformed human buccal keratinocytes. *Cell Mol Life Sci* 65:3653-63

2. Ansell, A., Jerhammar, F., **Ceder, R.,** Grafström, R.C., Grénman, R., and Roberg, K. 2009. Matrix metalloproteinase-7 and -13 expression associate to cisplatin resistance in head and neck cancer cell lines. *Oral Oncol* 45:866-71

3. Thompson, C.M., **Ceder, R.,** and Grafström, R.C. 2010. Formaldehyde dehydrogenase: beyond phase I metabolism. *Toxicol Lett* 193:1-3

4. Jerhammar, F*., **Ceder, R*.,** Garvin, S., Grénman, R., Grafström, R.C., and Roberg, K. 2011. Fibronectin 1 is a potential biomarker for radioresistance in head and neck squamous cell carcinoma. *Cancer Biol Ther* 10:1244-51
   *These authors contributed equally to this work.

5. Kwiecinska, A., Ottoson-Wadlund, A., **Ceder, R.,** Grafström, R.C., Björck, E., Nordenskjöld, M., Porwit, A., and Fadeel, B. 2011. HAX-1 expression in human B lymphoma. *Leukemia* 25:868-72

6. Farnebo, L., Jerhammar, F., **Ceder, R.,** Grafström, R.C., Vainikka, L., Thunell, L., Grénman, R., Johansson, A.C., and Roberg, K. 2011. Combining factors on protein and gene level to predict radioresponse in head and neck cancer cell lines. *J Oral Pathol Med* 40:739-46

CONTENTS

LIST OF ABBREVIATIONS

CASP	Caspase
CFE	Colony forming efficiency
CK	Cytokeratin
2D-PAGE	Two-dimensional polyacrylamide gel electrophoresis
EMHA	Epithelial medium with high levels of amino acids
FBS	Fetal bovine serum
GO	Gene Ontology
GOTM	Gene Ontology tree machine
GSATK	Gene set analysis tool kit
HGEM	Human gene expression map
HNSCC	Head and neck squamous cell carcinoma
HPA	Human protein atlas
HPV	Human papillomavirus
IPA	Ingenuity pathway analysis
IST	In silico transcriptomics
IVDGS	*In vitro*-derived differentiation-related gene set
MALDI-TOF	Matrix-assisted laser desorption-ionizing time-of-flight
MAS 5.0	Microarray suite 5.0
NOK	Normal oral keratinocytes
OMICS	The science of large scale profiling to retrieve holistic information from one or several biological levels in the cell e.g., genome, transcriptome, proteome, metabolome and interactome, omics was not capitalized in the text
OSCC	Oral squamous cell carcinoma
3R	Principle for replacement, refinement and reduction of animal experiments
RMA	Robust multi-array average
SET	Signature evaluation tool
SIGS	Serum-induced gene set
SV40	Simian virus 40
TGF-β	Transforming growth factor-β
TSCC	Tongue squamous cell carcinoma

1 INTRODUCTION

1.1 THE BIOLOGICAL FATE OF KERATINOCYTES

1.1.1 The oral mucosa – structure and function

The oral mucosa serves as a barrier to maintain tissue homeostasis and protect the underlying tissue from environmental factors such as bacterial, chemical and physical agents (1). The oral epithelium, encompassing several layers of keratinocytes, creates the surface of the oral mucosa, and it is supported by adjacent connective tissue, the lamina propria. Based on morphology and differentiation patterns, the oral mucosa can be divided into three major parts (2, 3). First, the major surface area of the oral mucosa encompasses the lining mucosa, a non-keratinized, stratified squamous epithelium that includes the buccal mucosa and floor of the mouth. Second, the masticatory mucosa, a keratinized, stratified squamous epithelium, includes the dorsum of the tongue, hard palate and gingival mucosa. Third, the specialized mucosa encompasses the taste buds on the dorsal surface of the tongue. The current thesis has primarily focused on the lining mucosa for the *in vitro* studies, encompassing around 60% of the total surface area of the oral lining, and therefore serves as an important target for carcinogen exposure and subsequent cancer development in the head and neck region (4-6). The lining mucosa, can be further divided into three layers; [1] the inner basal layer that attach to the basement membrane and contain stem cells and undifferentiated, mitotic keratinocytes [2] the intermediate layer that contains maturing and migrating keratinocytes that eventually end up in [3] the superficial layer that contains flattened, terminally differentiated cells (1, 7). A majority of the oral epithelium encompasses keratinocytes, but other cell types are also present such as Langerhans cells, Merkel cells and inflammatory cells (3). The migration and maturation from the basal layer to the upper layers take 14-20 days for the lining mucosa, and the keratinocytes eventually lose their proliferative potential as they move up in the epithelial layers and become terminally differentiated (3). The function and balance between these layers is disrupted during in cancer development.

1.1.2 Terminal differentiation

A balance between cell growth and cell death is essential for proper development of the tissue (3, 8). Terminal differentiation (TD) represents the primary biological fate for cell death in the oral epithelium (1, 7, 8). Migration of the basal keratinocytes to the upper layers to form the cross-linked protein envelope, which serves as a protective layer, associates with a multifaceted differentiation program. This transition is

accompanied by morphological changes, cytoskeletal rearrangements and influences on other biological fates such as proliferation and apoptosis (1, 3, 8). Expression of cytokeratins (CK), structural proteins that form intermediate filaments, is determined by the differentiation stage (1, 7) Keratinocytes in the basal layer primarily express CK5, CK14 and CK19, while the upper layers express CK4 and CK13. Constituents of the cross-linked envelope, e.g., involucrin, small-proline rich protein 1 and 3, and transglutaminase 1 increase in expression as the cells start to differentiate (1, 3). Involucrin in particular serves as a marker for commitment to terminal differentiation (9). Initiation in differentiation circuits lead to decreased proliferation in order to maintain the tissue homeostasis (8). This balance is perturbed in immortalized cells due to terminal differentiation resistance (10). There are numerous components that regulate normal tissue homeostasis including growth factors, adhesion molecules and toxicants (3). Means of inducing TD *in vitro* include, serum factors (e.g., Ca^{2+}, retinoic acid, vitamin D_3, and transforming growth factor-β), confluency (high cell density and cell-to-cell contact) tumor promoting factors (e.g., 12-O-tetradecanoylphorbol-13-acetate) and toxicants (e.g., aldehydes) (11-16). To mirror the complete differentiation program that emerges *in vivo*, different means of inducing TD *in vitro* are likely needed. The importance of perturbations in differentiation circuits for head and neck squamous cell carcinoma (HNSCC) pathogenesis was recently emphasized by whole-exome sequencing of 74 tumor specimens where 30% of the mutational landscape involved genes regulating differentiation (17). These findings indicate that similar to previous knowledge on p53 inactivation, which is consistent among this heterogeneous malignancy, aberrances in the differentiation program seem equally comprehensive for HNSCC carcinogenesis (17-19).

1.1.3 Apoptosis

Despite previous perplexity regarding the distinction between apoptosis and TD, it is now established that these processes represent separate cellular fates, even if some molecular features are similar (8). Apoptosis is characterized by cytoplasmic condensation and plasma membrane blebbing generating apoptotic bodies that can be recognized and engulfed by macrophages (20). The apoptotic process can be initiated by the extrinsic or intrinsic pathway. The extrinsic pathway is activated by binding of a ligand to a death receptor (e.g., Fas) that triggers a caspase (CASP) cascade (CASP8 and CASP3) resulting in cell degradation. The intrinsic pathway is activated upon cellular stress (e.g., DNA damage) by the release of pro-apoptotic factors such as cytochrome c from the mitochondria. Binding of cytochrome c to the apoptotic

protease-activating factor-1 activates a caspase cascade (CASP9 and CASP3) resulting in cell degradation (20). There are several regulators of apoptosis acting either in a pro-apoptotic (e.g., BAX, BAK) or anti-apoptotic (e.g., BCL2, BCL2L1, MCL1) manner. Although apoptosis and terminal differentiation involve some similar signal transduction pathways, the activation patterns of the involved transcription factors and proteins differ clearly on the mechanistic level (8). For example, TP53 is imperative for apoptosis in response to cellular stress such as DNA damage, while the family member TP63 is central for terminal differentiation (8, 17). Second, caspases are crucial for execution of both the intrinsic and extrinsic pathway, but appear to have less relevance for differentiation, despite the involvement of CASP14 in formation of the outermost layer of the skin (20, 21). Recent mutational profiling does however display that 8% of the assessed HNSCC samples harbor mutations in CASP8 (17).

1.1.4 Head and neck / oral carcinogenesis

Head and neck cancer is a broad term applied for cancers in the oral cavity, pharynx and larynx and it is the sixth most common cancer worldwide (5, 22, 23). More than 90% of the tumors are of epithelial origin and about 50% establish in the oral mucosa (24). Known risk factors include alcohol consumption, tobacco usage and human papilloma virus (HPV) infections (5, 18, 22). Patients are currently diagnosed by clinical examination, radiology and histopathological staging. Despite recent advances in aggressive multidisciplinary treatment approaches, including surgery, chemotherapy and radiotherapy, the mortality rate is still around 50% after 5 years and patients develop locoregional recurrences, distant metastasis and second primary tumors (5). The extensive molecular heterogeneity of HNSCC provides an additional level of biological complexity for prognosis and treatment, including utilization of targeted therapies, emphasizing the importance of developing molecular markers for individualized treatment (25-27). To date, there are no molecular tests that provide higher diagnostic accuracy compared to the conventional pathological scoring methods (28).

Carcinogenesis in the head and neck region is a multi-step process by which genes and pathways acquire genetic and epigenetic changes that enable a transition from a normal epithelium to an invasive carcinoma via dysplasia and carcinoma in situ (5, 10, 18, 22). Recently reviewed by Hanahan and Weinberg, cancer development is generally typified

by a set of biologically distinguishable hallmarks (29). These hallmarks can be translated to HNSCC development and include:

- **Evasion of growth suppression,** by mutations and altered expression of the cell cycle regulators TP53, CCND1 and CDKN2A in primarily HPV-negative HNSCC, while HPV positive cancers display changes in TP53 and the Rb family genes (18).

- **Self-sufficiency in growth signals,** by increased expression of the epidermal growth factor receptor, also serving as a marker of early carcinogenesis and predictor for poor outcome (30-32).

- **Insensitivity to anti-growth signals**, by down-regulation of the growth inhibitory transforming growth factor β (TGF- β), for example via down-regulation of disabled homolog 2 permitting TGF-β mediated cell motility and tumor growth (33).

- **Evasion of cell death,** by mutations in terminal differentiation-associated genes (e.g., IRF6, TP63, NOTCH1, RIPK4) and apoptotic genes (e.g., TP53, PTEN, CASP8, and PIK3CA) (17, 19).

- **Invasion and metastasis,** via alterations in genes involved in epithelial-mesenchymal transition, such as the hub-gene E-cadherin, to acquire plasticity and facilitate migration (34).

- **Induction of angiogenesis,** by increasing the vascular endothelial growth factor receptor expression, which associates with poor outcome (35)

- **Reprogramming of metabolism,** by expression of mutant mitochondrial DNA (NADH dehydrogenase subunit 2) which induces reactive oxygen species production, aerobic glycolysis and tumor growth (36).

- **Immune escape,** by increased expression of STAT3, resulting in suppressed expression and excretion of proinflammatory cytokines and chemokines (37).

Genetic instability in the cancer cells is also an important feature for driving the multi-step progression from normal to tumorigenic (5, 29).

1.1.5 *In vitro* models for head and neck / oral carcinogenesis

Cell lines constitute a renewable source of well-defined starting material for exploring carcinogenesis and biomarkers. The epithelia constitute the major target for tumor development (5). Multiple cell line models exist for head and neck carcinogenesis studies, where the absolute majority applies serum to the culture conditions (38, 39). Differently, most recent methods for culture of cell lines from normal (non-transformed) tissue employ serum-free conditions (11, 40). Thus, the comparison of normal and transformed states is commonly hampered by cultivation of the respective

4

cell phenotypes in different conditions. Interestingly, normal, proliferative and basal-like cells in culture undergo terminal differentiation when exposed to serum, typically fetal bovine serum (FBS) (11, 15, 16). Normal cells in culture are typified by a finite lifespan, and thus laboratory-based efforts for immortalization of cells, or generating lines from tumor tissue, have been beneficial to acquire renewable, broadly applicable sources of cell lines (11, 41, 42). Of particular importance to this thesis project, the transformed systems proliferate reproducibly on the basis that they are unable to undergo TD (10). Considering a central role of perturbed differentiation in HNSCC carcinogenesis, and the inability of transformed cells to undergo their normal and most common cellular fate, argues against the application of transformed model systems only (10, 17). Cancer development, even at its early stages, may also involve altered regulation of other cellular fates such as apoptosis (8, 20). The definition of whether, and to what extent, cell line models are proficient of relevant forms of cell death therefore needs to be addressed at genomic and functional level.

Limited cell lines are available for modeling the intermediary stages in cancer progression, mainly because cancers that establish *in vitro* associate with poor prognosis (43). Immortalization is a crucial step in the transformation of normal cells. Efforts to laboratory-immortalize normal keratinocytes have primarily been made using viral oncogenes such as Simian virus 40 (SV40) and HPV that perturb the p53 and Rb pathways (42, 44). Viral oncogene immortalization generates predominantly non-tumorigenic cell lines (38). Alterations of other cancer-associated genes for immortalization are successfully demonstrated by cyclin D1 overexpression and p53 inactivation, however failing to produce growth in soft agar and tumors in nude mice (45, 46). Additional overexpression of epithelial growth factor receptor and c-myc as well as telomerase reactivation is needed for complete transformation generating colonies in soft agar and tumors in immunodeficient nude mice (46).

To avoid the complication that arises from differing culture conditions, a triad of model systems, i.e., normal oral keratinocytes (NOK), the SV40 T-antigen immortalized line SVpgC2a and the carcinoma line SqCC/Y1 have been utilized for *in vitro* modeling of oral carcinogenesis under a standardized serum-free condition (Figure 1, and reviewed in PAPER I). NOK cultures from the mucosal lining are typically derived from biopsies of patients undergoing maxillofacial surgery (47). *In vitro* cultures of NOK aim at deriving highly proliferative and basal-like cells that include a low number of oral keratinocyte stem cells, among a pre-dominant population of transiently-amplifying

5

progenitor cell that eventually commit to terminal differentiation, in particular from stimuli such as confluency culture (high cell density and cell-to-cell contact) and from exposure to FBS at levels of 2 to 10% supplementation (11, 15, 16).

The SVpgC2a cell line was derived from transfection of SV40 T-antigen to NOK from a female donor, resulting in a number of different cell lines with extended longevities (42). One of these cell lines survived two crises that resulted in the establishment of the SVpgC2a cell line. The SVpgC2a cell line is immortal (cultured for >150 passages), non-tumorigenic when injected into nude mice and highly terminal-differentiation deficient (42). The SqCC/Y1 cell line was originally derived from a verruous squamous cell carcinoma of a female donor using a standard serum-containing medium. Adaptation of the culture to serum-free media required two months of culture effort, including signs of crises after three weeks, for generation of the serum-free strain of SqCC/Y1 (41). The SqCC/Y1 cell possesses an immortal, terminal differentiation-deficient and tumorigenic phenotype.

Figure 1. Morphologies and key features of normal and transformed oral keratinocytes in the start-up model grown in monolayer under standardized serum-free conditions
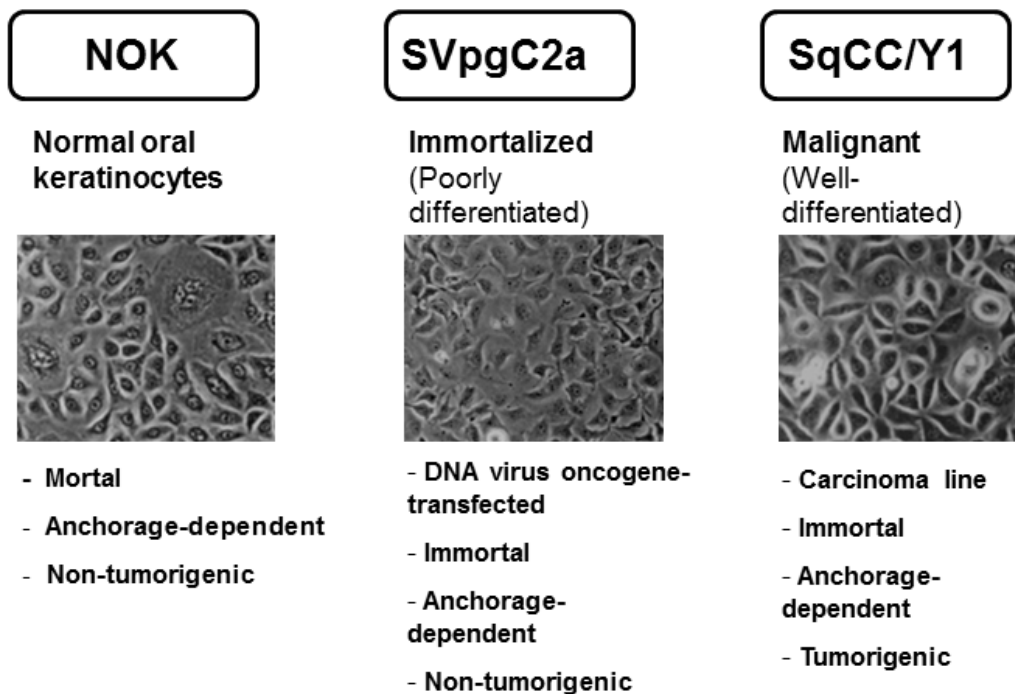
Major characteristics of the cell line model were summarized in PAPER I

1.2 "OMICS" AND BIOINFORMATICS ANALYSES

1.2.1 Cancer systems biology

The definition and application of the term "systems biology" covers technological, infrastructural as well as philosophical aspects (48-50). Systems biology aims at providing a holistic view of biological systems by multi-disciplinary integration of data from different technological platforms and experimental sources. The human genome project with subsequent development of high-throughput technologies for genomics, transcriptomics, proteomics and metabolomics enable generation of global data sets from cells and tissues. Novel integrative tools are needed to interpret and decipher the true biology from noise in these huge data sets, requiring multi-disciplinary collaboration and appropriate infrastructure (48, 51). Systems biology argues for a philosophical shift in the way we perceive biological systems, moving from a reductionist approach by assessment of individual components in the cells, to a systems-oriented approach, where components and networks are put together (49).

"Cancer systems biology" then adds further complexity (50). Cancer is a heterogeneous disease with a strong genetic component (29). It is well recognized that different mutations in various molecular pathways take place in a multi-step progression towards cancer development, including for HNSCC (5, 18, 22, 29). To elucidate the pathways that make a cancer cell functionally aberrant from normal cells and the interrelationship of the perturbed pathways, integration of different cellular level data is crucial (52). Cancer genome sequencing effort have shown that the number of mutations vary considerably between different tumor types, including inter tumor heterogeneity (53, 54). The mutational spectrum in HNSCC displays on average 130 mutations, and 25 % of these mutations seem to appear consistently (17). Despite a general low overlap, mutations are believed to alter expression of genes within similar functional classes (gene ontologies) or pathways leading to the hallmarks associated with cancer (29, 54, 55). There are several systems approaches using cancer omics data for delineating perturbed networks in tumors, as well as for prediction of poor outcome (56-58). Limited systems biology-based approaches for elucidation of mechanisms underlying transformation and diagnostics are available for head and neck cancer (59, 60). The ultimate goal of systems biology is taking the concepts to the clinic for application of systems medicine to eventually enable a more personalized cancer diagnosis and therapy (61-63). To achieve this goal, there is a need to elucidate if deregulated

molecular networks can associate genes and signatures with higher precision for potentially forecasting the multiple sides of prognosis and treatment.

1.2.2 Transcriptomics

1.2.2.1 A technological overview

The transcriptome is defined as the relative expression levels of the complete set of transcripts in a cell or a tissue. The microarray technology allows for analysis of thousands of genes simultaneously and provides a biological snap-shot of the transcript levels in the cell at a particular time or condition (64). Microarray analysis has broad range of application areas and has served as a powerful tool for e.g., hypothesis generation, sample classification, and clinical assessments. The basic principle underlying all microarray analyses is the presence of a variable number of nucleotides (the probes) corresponding to the respective genes on an array surface and subsequent addition of "sample" messenger RNA, which has been labeled with some fluorescent dye (64, 65). The amount of sample hybridizing to each probe is proportional to the abundance of the transcript, which is detected by the intensity of the dye, corresponding to the signal. Two different types of arrays are available for assessment of transcript levels. Two-channel arrays are characterized by hybridization of two samples labeled with different fluorescent dyes to the same array, and differential gene expression is calculated from the relative intensity of each fluorophore bound to the respective probes (65). In contrast, one-channel, oligonucleotide arrays only involve hybridization of one sample per array, and differential gene expression is obtained from probe intensity comparisons relative another array with the same probe setup, hybridized with a different sample of interest (66). The latter type was applied for the work in this thesis, in particular the Affymetrix technology (Santa Clara, CA). Each probe on the Affymetrix chips contains 25 nucleotides targeting the 3'-end of genes, which are spotted on a silica substrate using a photolithographic technique that enables production of high-density arrays (66-68). Eleven pairs of oligonucleotide probes are utilized to measure the abundance of each transcript to correct for unspecific binding. In addition, a mismatch probe is available for every perfect match probe, which is represented by a nucleotide at the 13th position. This is later taken into account for certain pre-processing methods (69).

1.2.2.2 Experimental design

The experimental design is fundamental for successful microarray studies since the amount of data generated is enormous. The first important issue to consider is the level

of replication, usually reflected by the biological question and funding. There are two major types of replication i.e., biological and technical (70). Biological replicates refer to independent samples from the same cell line or tissue. In general, analyses of normal cell and tissue samples encompass a larger interindividual variation compared to cells in culture, hence require more biological replicates (71). Technical replicates refer to repeated analyses of the same biological sample, however with today's technological advances this is of less importance (72). Following RNA isolation, labeling, hybridization and scanning of the arrays, the statistical processing and data mining commences To this end, replicates are also important for obtaining proper power in the downstream analyses.

1.2.2.3 *Pre-processing – the path towards finding the actual biological effects*

The first step towards revealing true biological differences between selected comparisons, while reducing technical differences, is the pre-processing that encompass background correction, normalization and summarization of the probe sets (73). Background correction primarily serves to adjust for background noise and un-specific binding, allowing the expression values to cover a suitable range. The normalization serves to reduce the non-biological variation within or between arrays, while the summarization generates a final probe set value corresponding to a single gene on the array. There are various methods to implement these steps, but the studies in this thesis primarily applied the MicroArray Suite (MAS 5.0) algorithm and the Robust Multi-array Average (RMA) (69, 74). The MAS 5.0 applies a background correction that divides the array into 16 regions; the mean intensity of the lowest 2% of the signal spot for each region is calculated, to provide a weighted sum of the zone background, which is further subtracted from each spot. The Tukey biweight robust median then is utilized for deriving a signal value for mismatch probes and perfect match probes, which are then subtracted from each other to obtain gene-specific probes on the array (69). This is followed by a global scaling for the individual arrays to a set value. The RMA algorithm applies a global background correction, quantile normalization and a robust linear model (median polish) for summarization. This method only takes perfect match probes into account (74). To date, there is no golden standard regarding which method to apply, and evaluation studies provide different results (75-77). Image analysis is also an important step in the pre-processing, commonly performed using software provided by the array company. Further quality

control measures are available on the Affymetrix array to assess amplification, labeling and hybridization (73, 78).

1.2.2.4 Statistical tests to enable identification of differently expressed genes

Similar to the pre-processing step, there are various methods to find significantly differentially expressed genes between the analyzed samples. Microarray studies typically include small sample sizes from which enormous amounts of data points are generated, and this poses a challenge for statistical testing (79, 80). Parametric tests such as t-tests are commonly applied, but require assumptions regarding the distribution of the data. Since microarray data may be noisy, nonparametric test may therefore be more applicable e.g., the Wilcoxon signed-rank test, but can also be less sensitive compared to the parametric tests. Other tests include Significance Analysis of Microarray that applies a similar method to t-tests, although making fewer assumptions regarding the data distribution. It performs permutations of repeated measurements to estimate the false discovery rate (81). This correction for multiple testing is useful for decreasing the identification of false-positive genes. The limma package in R/Bioconductor is also applicable and performs a linear model fit on the data, calculates a moderated t-statistics and to give a stable estimate for small sample sizes, an empirical Bayes step is used (82). Different statistical tests may provide different answers, and the overlap between the methods is sometimes limiting (83). Subsequent data mining is therefore important since the differential expression of groups of genes likely reflect the biology more accurately compared to single genes.

1.2.3 Proteomics

The proteome is defined as the complete set of proteins expressed in a cell or tissue. Proteomics analyses enable assessment of biological programs that are executed, while transcriptomics analyses provide an overview of all ongoing processes. While the human genome codes for approximately 20 000 to 25 000 genes, the proteome is much larger due to post-translational modifications such as phosphorylation, acetylation and cleavage likely generating millions of different proteins (84, 85). Assessment of global protein profiles is still a technical challenge compared to transcriptomics, even under circumstances of rapid technological methods development (86). Proteomics methodologies can be divided into two groups, the first assumes no prior knowledge on the identity of the protein and the second assumes prior knowledge on the protein identity.

1.2.3.1 Methods for finding proteins with unknown identity

Two dimensional gel electrophoresis and mass spectrometry are commonly used for global profiling of unknown proteins. The two dimensional electrophoresis separate the sample in two dimensions, the first according to charge by isoelectric focusing and the second according to molecular weight by sodium dodecyl sulfate polyacrylamide gel electrophoresis (87). Subsequent separation, proteins are detected by reversible post-electrophoresis methods such as coomassie blue or silver staining, resulting in spots that correspond to the respective proteins. The spot sizes are quantified digitally for detection of differently expressed protein using specific image software such as PDQuest (Bio-Rad, Hercules, CA) (88). A mass spectrometer, including an ionizing source, an analyzer and a detector, is subsequently applied to identify the proteins by converting neutral molecules to ions and separating them by their mass to charge ratio in an electromagnetic field (89). For example, the matrix-assisted laser desorption-ionizing time-of-flight (MALDI-TOF) mass spectrometry analysis generates peptide fragments with various molar masses that are compared to theoretical peptide fragments in databases for identification of the right protein (90). Gel-based methods are difficult from a technical perspective and have a fairly low dynamic range, affecting the sensitivity (87). Several non-gel based techniques employing high-performance liquid chromatography are currently being developed including improved quantification using isotope labeling for enhanced reproducibility and dynamic range (86). To this end, global proteomics methods provide means to identify high abundance proteins, yet low abundance transcripts are still difficult to detect (91, 92). Moreover, the correlation between protein and transcript abundance is low, while higher correlations are identified for functional categories (93-95). Novel bioinformatics processing approaches are needed to bridge the gap between protein and transcript data for potential alleviation of limitations with both methodologies.

1.2.3.2 Methods for finding proteins with known identity

Under circumstances when the protein is known, several antibody based techniques are available e.g., immunohistochemistry, Western blot and tissue microarray. Tissue microarrays enable simultaneous assessment of up to 1000 specimens, initially collected from individual paraffin-embedded tissue samples for assembly onto a common-paraffin block (96). The latter can be cut into hundreds of sections for individual assessment of various proteins on the complete tissue set. Tissue microarray coupled with immunohistochemistry is a powerful high throughput technology

enabling generation of for example a proteomics *in silico* tool (96-98). Specific antibodies are a limiting factor for this method.

1.2.4 Bioinformatics: protocols, tools and databases

Bioinformatics can be defined as the application of computational and informatics technologies to biology and medicine (99). Bioinformatics analyses commonly take two paths either "supervised" implying application of prior knowledge to drive the analyses or "unsupervised" referring to analyses without prior knowledge (100). Analyses of omics data tend to start with "unsupervised" approaches to obtain an overview of the complete set of generated data. Hierarchical cluster analyses can be applied to sort samples and genes, allowing visualization of patterns and trends in two dimensions (101) . Different distance metrics and linkage rules are utilized to organize the data and depending on selected method, diverse cluster tress may appear from the same data set (102). Commonly applied distance or correlation metrics include the for example the Euclidian distance or the Pearson correlation. To compare the nodes, different linkage rules are applied e.g., single linkage or average linkage. Refined clustering approaches are also available where the number of clusters is pre-defined, e.g., self-organizing maps and k-means clustering. Principal components analysis is an alternative applicable method for multidimensional data reduction and visualization that does not rely on pairwise comparisons (103). "Supervised" approaches comprise support vector machine (primarily for sample classification) as well as assessment of genes with known biological function and action (100).

1.2.4.1 Gene Ontologies and networks

The Gene Ontology (GO) nomenclature is powerful for providing functional biological analyses to data sets. The GO provides description of gene products under three main branches, i.e., biological process, molecular function and cellular component, and a gene may be linked to one or several categories under the respective branch (104). The ontology is displayed as directed acyclic graphs to visualize the relationship among categories (105, 106). There are several tools (Table 1) allowing functional profiling by making use of the GO-nomenclature to quantitatively calculate enrichment (over-representation) of categories among a selected set of differently expressed genes using statistical methods such as Fisher's exact test or Hypergeometric distribution (105-107). For these analyses, a proper background or reference is crucial, however for certain high throughput experiments it may be difficult to select the most appropriate

background (107). To this end, larger backgrounds provide more significant p-values. Several tools include tests to correct for multiple testing, though these may be quite conservative and impair the sensitivity (106, 107). GO-categories are also applicable for visualization of transcripts on the microarray with similar functions, displayed by the AffyAnnotator tool (Table 1). The GO-nomenclature is constantly developed and particularly valuable for identifying biological patterns of already known genes, since the results are dependent on annotated genes. To find novel interesting genes and gene sets, molecular network analysis is a promising approach. There are several commercially available network software that generate networks from a defined set of omics data relative to a knowledge database. The Ingenutiy Pathway Analysis (IPA) software has primarily been used in these studies and is therefore further elaborated on (Table 1). The selected "omics data set" is overlaid onto a global molecular network of information in the Ingenuity Pathway Knowledge Base. Network are further algorithmically derived based on the connectivity of the input data. Fischer's exact test is applied to determine the significance of associated biological function. Network analyses facilitate identification of centrally located and highly connected genes compared to ontology-based analyses. Another available molecular network and pathway database is the Kyoto encyclopedia of genes and genomes, also applicable for enrichment analyses relative to omics data sets (106, 108).

Table 1. Bioinformatics processing tools applied for the omics analyses in this Thesis

Function	Tool	URL
Gene Ontology enrichment analysis	Gene Ontology Tree Machine/ Gene Set Analysis Tool Kit	http://bioinfo.vanderbilt.edu/webgestalt
	Database for Annotation, Visualization and Integrated Discovery	http://david.abcc.ncifcrf.gov
Gene Ontology-based visualization of transcripts on the microarray	AffyAnnotator	http://www.bea.ki.se/jnlp
Network analysis and key regulator identification	Ingenuity Pathway Analysis	http://www.ingenuity.com
Body-wide gene expression maps	In Silico Transcriptomics	http://www.genesapiens.org
	Human Gene Expression Map	http://www.ebi.ac.uk/gxa/array/U133A
Body-wide proteomics resource	Human Protein Atlas	http://www.proteinatlas.org
Signature evaluation and classification	Signature Evaluation Tool	http://microarray.ym.edu.tw/tools/set

1.2.4.2 Standards and public omics resources

Data sharing and standards is an evolving and crucial concept for omics studies to enable reuse and integration of available data. The microarray analysis standards initiative commenced by the Microarray Gene Expression Data Society, nowadays called Functional Genomics Data Society, suggested a minimum of information about a microarray experiment, e.g., definition of platform, description of samples, and characterization of hybridization and labeling protocol (109). Similar standards are available for proteomics data, minimum information about a proteomics experiment (110). The application of omics technologies in clinical settings has initiated the MicroArray Quality Control consortium involving organizations such as the US Food and Drug Administration to establish quality control and standards for data comparability across platforms (72). The collection of transcriptomics data into public repositories such as the Gene Expression Omnibus and ArrayExpress has enabled the development of several novel data mining tools and serve as an excellent basis for validation of particularly *in vitro* data, as well as biological signature reduction (111, 112). The In Silico Transcriptomics (IST) data base and the Human Gene Expression Map (HGEM) are examples of two databases that have assembled expression profiles from thousands of microarray experiments (Table 1) (113, 114). The IST contains information from 10 000 samples with an average of 11 500 genes per sample including 43 healthy tissues, 68 malignant tissues and 64 non-tumor diseases (113). In total, data was retrieved from 5 different generation of Affymetrix platforms and comparisons were enabled by a novel normalization method (115). Body-wide expression profiles can be derived for single genes and co-expression analyses between selected genes. The HGEM includes 5372 human samples representing 369 different cell and tissue types, disease states and cell lines from one Affymetrix platform (114). The data is divided into different biological groups e.g., 96 biological groups that include at least ten biological replicates and empower statistical analyses among groups for differential expression of single genes. The lack of direct correlation between transcript and protein abundance often necessitates assessment of protein expression levels for identified genes (94, 95). The Human Protein Atlas (HPA), a proteomics database covering more than 50% of all human protein encoding genes contain high-resolution immunohistochemical images of the protein expression in 46 different normal human tissues, 20 different cancer types and 47 different human cell lines (Table 1) (97, 116). The simultaneous antibody generation and quality control of this proteomics resource is a key feature since the proteomics field has struggled more with

unspecific probes compared to the transcriptomics field. To this end, the expanding domain of public omics data and novel high throughput resources allow for integration of tools and development of approaches to ultimately increase the clinical translation of findings in early phases.

1.3 BIOMARKER DISCOVERY IS A RAPIDLY EVOLVING RESEARCH FIELD

Realizing the huge diversity among cancer types and impact of inter-individual differences, the wish for effective management of cancer patients becomes reformed from a "one size fits all" approach to personalized medicine (28, 117). The latter strategy requires development of biomarkers, an indicator of a biological state, to allow the physician to construct the optimal treatment plan. Extensive cancer biomarker discovery efforts have so far yielded minor support to the clinics need for point-of-care devices for making the correct cancer diagnosis and person-specific prognosis (118-120). Limited numbers of biomarker kits are available for clinical testing, and even fewer are commercially available (121-123). Apparently lacking for HNSCC, one approved gene signature for breast cancer is the 70-gene signature MammaPrint® (Agendia, Amsterdam, The Netherlands) that predicts metastasis-free survival (120, 121). The MammaPrint® signature was recently re-evaluated in independent population data sets, and in contrast to initial validation with limited data sets, the performance of the signature is now considered overestimated (119). The sensitivity of the signature remains at 90%, but the specificity (serving to predict which patients are at lower risk of metastasis) is only around 50%. One way to address this problematic issue in the biomarker discovery process is the consideration of early referencing of results to publically deposited transcriptomics and proteomics data, partly in databases, from normal and tumors tissues (98, 113, 114).

1.3.1 Cells lines and tissues in biomarker discovery

Well known to the cancer community, tissue analyses for cancer biomarker discovery are coupled with hurdles (117, 124). The patient can generally only be probed for specimen samples once or a limited number of times, and moreover many patients receive neo-adjuvant therapy to shrink the tumor before surgical procedures. Although partially circumvented by micro-dissection, tumor samples are generally mixed with some normal tissue cells and variably with other cell types than of tumor origin (124-

126). Many tissues like the oral cavity and the head and neck region display large structural heterogeneity although considered as one clinical entity (18, 25). Especially relevant to HNSCC development, the process of field cancerization, involving precancerous genetic alterations without notable histological changes in normal appearing tissues, serves to question the appropriateness of applying adjacent and even distant normal-appearing tissue as control reference material in the tumor analyses (127). Cell lines of tumor and normal tissue origin are promising complementary sources to biomarker discovery, providing opportunity for repetitive or even unlimited sampling, and abundant material of often relatively larger homogeneity (than tissue) (117). Moreover, cell cultures can be manipulated in multiple ways before analysis. However, cell lines are naturally poised with the problematic of correctly reflecting *in vivo* functions under selected culture protocols, or might not even establish in culture from certain tumors or tumor types (43, 128).

1.3.2 Gene expression profiling of HNSCC for biomarker discovery

Gene expression profiling of normal and HNSCC specimen and cell lines has generated biomarker signatures proposed for tumor classification (129-137). For example, large transcriptomics studies of primary versus normal HNSCC tissues have generated accurate signatures, including quite large number of genes, however with limited independent validation (130, 131, 134, 136). The insufficiency of HNSCC studies with outcome data is a limiting factor for the translation of biomarker signatures. Three major studies from HNSCC tumors with outcome data are available in the public domain, with variable assessment relative to differentiation (129, 133, 137). An effort of comprehensive molecular classification of HNSCC identified four sub-types of tumors that associated with poor outcome, including with differentiation-related changes (129). Another gene expression study in a large HNSCC primary tumor data set confirmed inconsistency between the differentiation state and outcome, including assessments relative to metastasis (133). The most recent study including normal oral epithelium and primary HNSCC associated primarily immune response defects and adhesion to survival (137). These studies provide valuable data sets and biomarker suggestions for independent validation of *in vitro* identified genes as potentially clinically relevant.

1.3.3 Strategies for biomarker discovery

Data-driven (unsupervised) and knowledge-based (supervised) approaches have commonly been applied to anchor gene profiles/signatures with biological responses (117). Limited attention have been given to model-driven *in vitro* strategies, which include induction of a biological process by a selected agent or precise culture protocol, and utilization of the transcriptional response of the cells to predict tumor properties (138). Integrative *in vitro* strategies of RNA and protein data are likely to increase the reliability also within biomarker discovery (139). A need of analyses of cellular differentiation states in the context of biomarker discovery is apparent from recent genomic analyses of HNSCC (17). Cultures of NOK can be directed to mimic physiologically driven processes such as differentiation, e.g., by confluency (contact)-mediated inhibition of growth or serum-exposure of cultures otherwise routinely grown without serum (11, 12, 14-16). Serum-exposure also adds the dimension of assessing differentiation from seeing "cancers as wounds that do not heal" (140, 141). These protocols allow for testing if impaired contact inhibition and absence of differentiated features reflects *in vivo* generation of tumor mass by hyper-proliferation, adjacent tissue invasion and metastasis (29). Central to the hypothesis testing of this thesis work, cell lines are viewed as complementary to tissue analysis, addressing the possibility that thorough analysis of even a limited number of cell models might reveal data hidden in the multitude of current relatively unsuccessful tissue analyses made so far in the field of cancer biomarker discovery. The translation from *in vitro* to *in vivo* can potentially be addressed by standardized serum-free culture protocols, relevant human cell types (epithelial cells are the source of the most common human tumor types), control cultures from multiple donors in each analysis, unique, well-characterized cell lines, and finally, extensive omics profiling and bioinformatics assessment and validation in public data sets and databases.

2 AIMS OF THIS THESIS

The overall aim of the current thesis project was to elucidate if genomic analyses of an *in vitro* model of normal and transformed human epithelium can elucidate basic mechanisms of cancer development and simultaneously generate paradigms/protocols generally applicable to cancer biomarker discovery.

An additional important dimension of this Thesis work has been to assess the general applicability of the applied *in vitro* and *in silico* technologies to alternative methods development in environmental medicine. This separate aim considers especially the societal need of defining novel cost-effective strategies for analyzing the safety of the multiple existing, or to be synthesized, untested chemicals and nanomaterials.

Two major approaches were taken to address oral carcinogenesis, and from a bioinformatics perspective, broadly the many sites that encompass head and neck squamous cell carcinoma. First, transcriptomics and proteomics data from the cell lines were integrated and bioinformatically processed to the Gene Ontology category and molecular network level into different gene signatures. Second, attempts of driving squamous differentiation (the major biological fate of most normal oral epithelia) were made by confluency culture and/or serum exposure of the cell lines within the model system. The applicability of the *in vitro*-generated signatures was validated in multiple oral and non-oral data patient data sets, including body-wide transcriptomics and proteomics expression databases.

The specific aims were:
- To compile existing results generated from a start-up cell line model of cancer development to facilitate anchoring of the novel "omics"-based characterization of the model
- To establish and characterize novel tumor cell lines under serum-free conditions with the intent of potentially expanding the number of entities in the model
- To derive expression profiles for broad overview of the basis for phenotype differences related primarily to studies of functional and non-functional cell differentiation

- To develop "omics-integrative" (combining protein and transcript data) and omics model-driven strategies (changing conditions of culture) for applying bioinformatics processing tools to biomarker discovery, defining altered genes, molecular networks and Gene Ontologies

- To assess broadly that the concepts of tumor biomarker discovery are applicable in concept also to the current needs of environmental medicine

3 EXPERIMENTAL PROCEDURES

This section provides an overview of the materials and methods that have been used in the present thesis. Detailed descriptions can be found in the Materials and Method section of the attached publications and manuscripts.

3.1 CELL CULTURE PROTOCOLS

Normal oral keratinocytes (NOK) were cultured and transferred in the laboratory-fabricated serum-free media (EMHA) or the commercial Keratinocyte-SFM (Gibco), both which contained growth-promoting supplements like epidermal growth factor and pituitary extract, were used interchangeably without detectable differences in growth among cell lines or outcome of experiments. Cells in passage 2 or 3 were used throughout the experiments. The SV40 T antigen-immortalized oral keratinocyte line SVpgC2a and the buccal squamous cell carcinoma line SqCC/Y1, were cultured under identical conditions as the NOK. Passages 60-72 were used for SVpgC2a and passages 125-135 for SqCC/Y1 for all experiments. Characteristics of both cell lines were extensively reviewed in PAPER I.

The tongue squamous cell carcinoma cell line LK0412 was established and cultured under conditions identical to the normal counterpart (PAPER IV). Morphology of all cells was evaluated under phase-contrast microscopy. For the LK0412 cell line, transmission electron microscopy images were also generated.

3.2 ASSESSMENT OF KERATINOCYTE BIOLOGICAL FATES

3.2.1 Terminal differentiation

Two established protocols were applied to induce terminal differentiation (TD). Cells were grown to 100% confluency and kept confluent for up to four day or by cultivation to 100% confluency followed by 5% fetal bovine serum (FBS)-exposure for four days. Commitment to TD was determined by assessment of established markers for TD e.g., involucrin expression, and by microarray analysis (single gene and Gene Ontology level) and immunochemical analysis.

3.2.2 Apoptosis

Scoring of apoptosis was based on morphological hallmarks i.e., condensed chromatin indicative of pyknosis. Cells were formalin-fixed and deposited on coverslips with the fluorescent DNA staining dye DAPI or propidium iodide. A florescent microscope was

used to analyze and score the cells. Apoptosis was also assessed by microarray analysis at the Gene Ontology level.

3.3 ANALYSES OF CELL GROWTH

Proliferative capacity was assessed by manual counting under the microscope or by the colony forming efficiency (CFE) assay. For the CFE assay, cells were seeded at cell type specific densities and incubated until surviving colonies could be detected and scored under phase contrast.

3.4 TRANSFORMATION ASSESSMENTS

Anchorage-independent growth was analyzed by soft-agar colony growth over a wide range of seeding densities and colonies reaching a pre-selected size were counted under phase contrast. NOK did not generate soft agar colonies and was therefore used as a negative control. The tumorigenicity in an immunodeficient host was assessed in BALB/c (nu/nu) mice subcutaneously injected with cells. Tumor tissue was analyzed by routine histopathological protocols. Injection of NOK served as a negative control.

3.5 MUTATION AND "OMICS" ANALYSES

3.5.1 DNA mutation analysis

Single strand polymorphism analysis followed by sequence analysis was applied to identify mutations in cells and tissues.

3.5.2 Transcriptomics

Transcriptomics profiles of the cell lines were generated using the oligonucleotide Human Genome Focus array (Affymetrix). Raw data files were processed and subjected to data mining using the tools listed in Table 1 and in the bioinformatics processing section.

3.5.3 Proteomics

Proteomics profiles were generated using two-dimensional gel electrophoresis (2D-PAGE) followed by in gel digestion and matrix-assisted laser/desorption ionizing-time of flight (MALDI-TOF) mass spectrometry or liquid chromatography-mass spectrometry/mass spectrometry for mass finger printing. Western blot analysis was applied to verify low abundance proteins and selected transcripts identified from the microarray analysis.

3.6 BIOINFORMATICS PROCESSING

3.6.1 Quality control and preprocessing

All CEL files underwent basic quality control using the simpleaffy package in the R-environment from the Bioconductor project (http://www.bioconductor.org). Pre-processing was performed using MAS 5.0 or RMA algorithms.

3.6.2 Assessment of differential gene expression

Various statistical tests were applied to find significantly differently expressed transcripts, e.g., empirical Bayes statistics, Wilcoxon's signed rank test and Significance Analysis of Microarray.

3.6.3 Gene Ontology analyses

Transcript characterization using the Gene Ontology nomenclature under biological process, molecular function and cellular component was applied by the GO-enrichment programs Gene Ontology Tree Machine (GOTM) / Gene Set Analysis Tool Kit (GSATK) or the Database for Annotation, Visualization and Integrated Discovery. Visualization of transcripts on the microarray chip sorted according to the Gene Ontology nomenclature was enabled by the AffyAnnotator program.

3.6.4 Network analyses

The network analysis tool, Ingenuity Pathway Analysis (IPA) was applied to generate molecular networks from selected gene products based on information in a curated data base encompassing millions of publications. To pinpoint centrally located genes, the concept key regulator gene was defined representing a gene with at least three interactions with significantly differently expressed transcripts.

3.6.5 Validation in public repositories and databases

The *in vitro*-derived gene expression profiles were validated using data selected data sets from the public microarray repositories, ArrayExpress and Gene Expression Omnibus. The compiled transcriptomics databases In Silico Transcriptomics (IST) and Human Gene Expression Map (HGEM) were also applied. The proteomics database, the Human Protein Atlas (HPA) was utilized to assess the findings at the protein level. Selected findings were also validated in relation to the healthy plasma proteome and whole saliva from healthy and oral cancer patients.

3.6.6 Gene signature evaluation

The Signature Evaluation Tool (SET) was applied to evaluate and refine the discriminatory power of the *in vitro*-derived signatures using Golub's weighted voting algorithm.

3.6.7 Patient survival analyses

Survival differences among individual genes and gene sets were assessed by Kaplan-Meier analysis and log-rank test. For individual genes, the median gene expression levels were applied to divide samples into two groups. For gene sets, the concept of "survival points" was applied, taking each gene into account. Accordingly, points were provided to samples with gene expression levels that correlated with good outcome. In contrast, gene expression levels that correlated with poor outcome were given no points. The survival points were subsequently summarized for each sample. The samples were then further divided into two groups based on a cut-off level of half of the maximum of assigned points.

3.6.8 The biomarker discovery pipelines

The "**integrative**" pipeline included processing of proteomics and transcriptomics data from normal and transformed cells. Significantly differently expressed proteins and transcripts were integrated by Gene Ontology (GOTM) and network analyses (IPA) via the AffyAnnotator tool. *In vitro* derived profiles were further assessed relative to a normal and tumor tissue training data sets for signature evaluation (SET). The herein refined signatures were further analyzed relative to various independent oral and non-oral transcriptomics data, as well as global transcriptomics (IST, HGEM) and proteomics (HPA) databases, and saliva and plasma datasets.

The "**model-driven**" pipeline included induction of biological processes in the normal and transformed cells by means of confluency and/or serum with subsequent transcriptomics profiling. Significantly differently expressed transcripts were assessed by application of Gene Ontology enrichment (GOTM / GSATK) and network analyses (IPA). The derived signatures were then further assessed relative to selected HNSCC data sets and signatures including survival data, as well as transcriptomics (HGEM) and proteomics (HPA) databases.

4 RESULTS

4.1 PAPER I: ASSESSMENT OF THE START-UP CELL LINE MODEL

The main goal was to evaluate features of the initial cell line model, to mirror the progression of cancer from normal to a fully malignant state, and to broadly identify potential application domains, including future biological and bioinformatics modeling.

Preservation of features typical of normal tissue epithelia was demonstrated for NOK, e.g., finite life-span, positive response to growth factors, ability to undergo growth arrest and TD as a result of stimuli such as TGF-β, Ca^{2+} and FBS. The SVpgC2a and SqCC/Y1 cell lines displayed immortal phenotypes, partly or fully-resistant to differentiation-inducing agents. SVpgC2a showed signs of extensive loss of differentiation, typical of aggressive tumors. Organotypic cultures of SVpgC2a encompassed primarily basaloid non-differentiated cells, while SqCC/Y1 formed relatively more preserved normal-appearing mucosal structures. The influence of toxic agents was assessed in terms of cellular pathology including growth and cloning, as well as mitochondrial and plasma membrane function. Overall, the results displayed differences in relation to cell survival, growth and differentiation in the respective normal, immortalized and malignant states. Notably, SVpgC2a retained normal tissue functions for metabolism of xenobiotics, in contrast to SqCC/Y1, allowing assessment of sensitivity to cell transformation by chemicals at different stages of cancer progression.

Several components showed the advantage of using this triad of cell lines for mimicking oral cancer progression. First, p53 tumor-suppressor protein was functional in NOK, but not in SVpgC2a and SqCC/Y1. Second, compared to NOK, the apoptosis rate was increased in SVpgC2a and decreased in SqCC/Y1, reflecting results obtained from tissue analyses, including a comparison of normal, dysplastic and malignant states. Third, immunochemical profiling of cytokeratins in organotypic cultures indicated that NOK express the same keratins as the tissue of origin, whereas SVpgC2a and SqCC/Y1 showed keratin profiles similar to an oral epithelial dysplasia and well-differentiated squamous cell carcinomas, respectively. Taken together, comprehensive global characterization by omics-technologies would be useful for extended application of this cell line model to precisely define the differences between cultured and non-cultured states.

4.2 PAPER II: ESTABLISHMENT OF THE INTEGRATIVE BIOMARKER PIPELINE

The main goal was to derive a bioinformatics approach for integrating proteomics and transcriptomics data from the start-up cell line model, and identify perturbed low abundance transcriptional regulators. Proteomics analyses of SVpgC2a and SqCC/Y1 identified 19 differently expressed proteins relative to the normal state, in one or both cell lines. Single-gene transcript assessment was feasible for 17 of the 19 proteins by microarray analysis, where 9 of the 17 transcript displayed a significant change relative to the normal state.

Protein and transcript data were linked in a two-step manner. First, the GOTM tool was applied for GO enrichment-analysis; the respective protein profiles enriched eight categories in SVpgC2a and four in SqCC/Y1. The AffyAnnotator tool was next applied to link the categories with the transcripts, by selecting differently expressed transcripts within protein-enriched GO-categories. Overall, the GO-categories captured 582 transcripts in SVpgC2a and 146 transcripts in SqCC/Y1.

The IPA tool was next applied to investigate if low-abundance transcriptional regulators could be identified from the protein-derived transcript and molecular networks. Overall, 10 molecular networks were identified for SVpgC2a and 5 networks for SqCC/Y1. Cancer, cell death and cellular growth and proliferation were among the top three functions for a majority of the networks in both cell lines. A total of 18 genes from the networks fulfilled the criteria for consideration as a key regulator. The analysis identified CDKN2A, MYC, MYCN, SP1 and TP53 as key regulator genes in both cell lines. Additional key regulators included 10 specific key regulators for SVpgC2a, and three for SqCC/Y1. Five of the transcription factors were further assessed at the protein level by Western blot analysis. Relative to the normal state, the expression of Sp1, Sp3 and c-myc was increased in both SVpgC2a and SqCC/Y1. Differently, Hif-1α was selectively increased in SqCC/Y1 and p16 in selectively increased in SVpgC2a.

4.3 PAPER III: MODEL-DRIVEN BIOMARKER PIPELINE: ASSESSMENT OF CONFLUENCY-DRIVEN CHANGES

The main goal was to define transcript profiles of altered responsiveness to the growth-inhibitory and differentiation inducing effects of cell-cell contact in normal and immortalized oral keratinocytes (NOK and SVpgC2a).

Initial analyses of contact inhibition and differentiation-promoting culture of NOK and SVpgC2a displayed altered growth, cloning and saturation density in the immortalized versus normal state, including absence of differentiated morphological features and differential regulation of apoptosis.

Transcript profiles of NOK and SVpgC2a at the sub-confluent and confluent state were generated. The six days confluent state of NOK relative to the sparse state generated 120 differently expressed genes, including categories and networks confirming association to development, differentiation and adhesion. Confluent cultures of SVpgC2a, displayed 12 differently expressed genes, none corresponding to the changes induced by NOK confluency. The SVpgC2a versus NOK transcriptome including 341 genes enriched altogether 52 GO categories, 18 networks and 39 key regulator genes, several of which associated to epithelial-mesenchymal transition.

An "*in vitro*-derived differentiation-related gene set" (IVDGS) encompassing 476 genes was constructed from fusion of the three *in vitro*-derived signatures. Overlap analysis of the IVDGS with the Rickman *in vivo* differentiation and metastasis gene sets displayed limited overlapping genes. Evaluation of the IVDGS relative to survival data in the same data set identified 31 significant genes. Assessment of these genes relative to the HNSCC samples in the HGEM indicated four and 14 genes, that displayed concordant changes in tumors with poor and good outcome respectively. A significant impact on overall survival was further obtained in an independent HNSCC data set containing 71 samples from applying the 4-gene signature (COX7A1, MFAP5, MPDU1 and POLD1).

4.4 PAPER IV: ESTABLISHMENT OF AN OSCC CELL LINE UNDER SERUM-FREE CONDITIONS

The main goal was to establish and characterize an oral squamous cell carcinoma (OSCC) cell line in conditions devised for NOK, to ultimately extend the initial cell line model. The cell line LK0412 was established from a protocol involving four passages and two periods of cultural degradation. Subsequently, cultures retained identical morphology without noticeable signs of degeneration or crisis for over 50 passages, including weekly transfers at split ratios of around 1:3. The epithelial morphology was verified by phase contrast microscopy, as well as transmission electron microscopy at the sub-cellular level displaying cytoplasmic intermediate filaments and desmosomal junctions. Immunohistochemical analyses further verified consistent presence of cytokeratins among the cells. Growth characterization showed that LK0412 exhibited a cloning efficiency of 25-35% and a growth rate of 0.5±0.1 population doublings per day. Assessment of biological fates compared to NOK demonstrated increased indices for proliferation and apoptosis, while the terminal differentiation index was decreased. FBS inhibited growth and increased apoptosis, but failed to induce terminal differentiation, the latter clearly differed from the response in NOK.

A transformed state of LK0412 was indicated from anchorage-independence and growth to high saturation density *in vitro*, including tumor development following inoculation of cultures into an immune-deficient host. Histological examination of the tumors indicated a moderate differentiation, similar to the histology of the original tumor. p53 protein analysis displayed increased levels with a missense mutation detectable in both the cell line and original tumor specimen.

Transcriptomics characterization by microarray analyses identified 225 differently expressed genes. GO-enrichment analysis generated eleven categories, the majority under biological process. Protein-level assessment of selected transcripts previously associated to OSCC development indicated altered expression levels in LK0412 relative to NOK. Moreover, the GO-analysis as well as "top ten deregulated transcripts" in LK0412 relative to NOK, suggested potentially novel OSCC biomarker genes such as BST2 and ISG15.

4.5 PAPER V: MODEL-DRIVEN BIOMARKER PIPELINE: ASSESSMENT OF SERUM-DRIVEN CHANGES

The main goal was to assess the effects of FBS in the extended cell line model, including NOK and the transformed keratinocyte cell lines SVpgC2a, SqCC/Y1 and LK0412. Accordingly, exposure to 5% FBS induced expected signs of squamous differentiation in NOK. The serum-exposure displayed some effects on the transformed cell morphologies however overall lacking signs of morphological differentiation.

Transcriptomics profiles were derived to elucidate changes induced from FBS. NOK displayed 99 differently expressed genes, 13 GO-categories and six molecular networks, including coupling to development, differentiation and injury effects. Overall, the transformed cells displayed around 3-fold lower number of differently expressed transcripts. In SVpgC2a, 31 genes were differentially expressed, contributing to one GO-category i.e., AT-binding. SqCC/Y1 displayed 23 differently expressed genes, and 16 GO-categories, encompassing biological functions such as development, metabolism, and inflammation and wounding. LK0412 exhibited 29 differently expressed genes and 54 GO-categories including, association to cell death, cell migration, inflammatory and immune responses, and wounding. Network analyses for the transformed cell line variably supported the GO-analyses. Lipid metabolism was shared as a top functions for both the normal and transformed states.

A serum-induced gene set (SIGS), encompassing 180 genes was constructed from fusing the *in vitro*-derived gene sets. Assessment of the SIGS to *in vivo* differentiation and metastasis-related gene sets from Rickman displayed limited overlap. Further evaluation relative to survival indicated 17 genes with impact on overall survival. Assessment of the 17 genes relative to HNSCC in the HGEM showed concordant expression for 10 genes relative outcome in the Rickman set. The HPA indicated increased intensities for five genes, and decreased intensity of one relative to the normal state. Higher than median expression of PDGFRL associated with decreased overall survival an independent HNSCC cohort. Significant survival differences were not noted for the other genes. Finally, a supervised review of published literature implicated previous connection to HNSCC for 12 of the 17 genes.

4.6 PAPER VI: EXTENSION AND *IN VIVO* ASSESSMENT OF THE INTEGRATIVE BIOMARKER PIPELINE

To main goal was to extend and apply the integrative biomarker pipeline to clinical data sets. Proteomics analyses identified 27 differently expressed proteins in LK0412 relative to NOK (Signature A). GO-analysis of these proteins generated 10 categories including 119 transcripts (Signature B). Molecular network analyses of these transcripts derived six networks, including 11 key regulator genes (the latter Signature C).

The three *in vitro*-derived Signatures A-C were next assessed relative to a training data set encompassing 38 normal and 57 tumor tongue tissues for determination of the classification power using the signature evaluation tool. The intact versions of Signature A-C, in order, classified the normal and tumor tissues with an overall accuracy of 80, 84 and 86%. Evaluation of all variants below the complete set of genes in each signature generated refined signatures (termed D, E and F respectively) which displayed even higher classification power, i.e., between 83 and 96%. Signature E which generated the highest classification power involved the genes, AURKA, CENPA, CYP27B1, ISG15, KRT15, MMP9, MX1 and TPX2.

The refined signatures (Signature D-F) were next validated in four independent oral/HNSCC data sets. A total of 171 normal and tumor tissue samples were classified with accuracies between 82-96% by Signature E. The predictive power of signature E was further analyzed relative to non-oral normal and tumor tissues from blood, brain, breast, kidney, lung and sarcoma. The non-epithelial tumors were classified correctly to only 20% or below. Differently, breast ductal carcinomas and non-small lung cancers showed higher classification power around 80%.

The 18 genes included in refined signatures D-F were next explored in two transcriptomics databases (IST and HGEM) and the proteomics resource, HPA. Elevated expression at the transcript level was found for eight genes in one or both transcriptomics databases. Considering immunostaining-based intensity and/or quantity, alterations were implied for nine genes in HNSCC relative to the normal state. Further analysis of the 18 genes relative to the saliva proteome from healthy and OSCC patients, as well as plasma proteome from healthy individuals, identified seven of the genes in either saliva, plasma or both.

5 DISCUSSION

Omics technologies have provided extensive information on the expression of suites of genes and proteins that are altered in cell lines and tumors. The genetic portrait in tissues is not completely reflected by cell lines (128). Questions remain how this can be addressed and potentially utilized for reducing the complexity in elucidating mechanisms underlying cancer development and biomarker discovery. The studies in this thesis aimed at generating broadly applicable integrative and model-induced biomarker pipelines based on application and extension of a multi-stage model of normal and transformed epithelium.

Evaluation and development of the cell line model provides a basis for biomarker discovery

Well-characterized cell line models are imperative for anchoring gene expression changes derived by high-throughput technologies. In PAPER I, the application and characteristics of the start-up *in vitro* model containing NOK, SVpgC2a and SqCC/Y1 was evaluated. Overall, extensive biochemical and morphological analyses supported that the cell lines mirrored aberrances in proliferation, terminal differentiation and apoptosis, in a manner similar to cancer progression *in vivo* (8, 20, 29). The inclusion of normal cells especially emphasized the applicability of the model system for assessment of terminal differentiation for the purpose of determining retained and aberrant normal tissue responses. Given the importance of terminal differentiation as the primary biological fate for keratinocytes, the definition of such response is also important as a reference for toxicity testing (8, 10). Especially since the majority of existing *in vitro* toxicity tests are utilizing cell lines that cannot turnover in a normal manner, and likely reflect early or late stages of cancer development (10). Elaborated characterization of this model system by broad "omics" analyses would emphasize the understanding of to what extent these human cell lines may reflect a specific cell type, tissue or even "strain of human representation" including complex biological endpoints.

The results in PAPER I pointed towards a general lack of serum-free OSCC cell lines and the unexploited approach of establishing tumor tissue-derived oral cell lines under serum-free conditions. To this end, the start-up model was extended by the first reported establishment of an OSCC line (LK0412) under conditions developed for the normal counterpart in PAPER IV. The carcinoma origin of LK0412 was verified from expression of multiple epithelial characteristics on the genomic and phenotypic level.

Histopathological assessments of the resultant tumors in nude mice agreed with that of the original tumor, as did the DNA sequence analysis of p53 of the cell line and tumor respectively. These findings support that that the resultant cell lined emanated from the tumor. Characterization of the LK0412 cell line indicated numerous features associated with malignancy, some traits reflecting tumor states which seldom appear in culture. For example, the LK0412 cell line emanated from a tumor exhibiting a moderately differentiated phenotype, but poorly differentiated states constitute the majority of current tumor cell lines (43). The transcriptomics profile of LK0412 was matched to previously proposed markers for oral cancer development from two major studies, however only seven out of 70 displayed similar deregulated expression, including some also at the protein level (22, 142). The LK0412-enriched GO-categories displayed a higher degree of overlap to previous OSCC studies e.g., for the categories cell adhesion, cell division, cell proliferation, immune response and extracellular components (143, 144). This further supports that genes do not act as single units, but rather as components and that the perturbation of a category or network may occur from varying mechanisms (52). In addition, the unique genotype of LK0412, as well as the presence and influences of adjacent normal tissue in tumor analyses may yield expression differences (125, 126). Unfortunately, there was not enough material of the original tumor to generate a transcriptomics profile to compare the resultant cell line with the original tumor specimen. The phenotypic changes in LK0412 also corroborated with the transcriptomics-associated GO-categories e.g., deregulated "response to stimulus" associated to alterations in genes involved in differentiation. In addition, the top differently expressed genes entailed a number of genes that may serve as potential novel biomarkers for tongue/oral cancer e.g., BST2 and ISG15. Additional serum-free cell lines would be useful to confirm findings implicated by LK0412. Furthermore, it would be of particular interest to assess if the expression profiles of serum-free generated cell lines potentially display higher similarity to the corresponding tissue of origin relative to serum-grown cell lines (128).

The integrative "omics" approach identifies low abundance transcriptional regulators and gene signatures for tumor classification

The initial evaluation of the start-up model in PAPER I inspired omics profiling efforts, and in PAPER II the integrative biomarker pipeline was established by combining proteomics and transcriptomics data from the transformed cell lines (SVpgC2a and SqCC/Y1) relative to NOK. Proteomics profiling identified 19 differently expressed proteins in one or both of the transformed cell lines. At the time, eight of the proteins

were seemingly novel to oral cancer, following publication of this study, ACAT1, PRDX3 and TMP3 have also been linked to oral cancer directly or indirectly (145-147). The corresponding transcript levels from the microarray analysis were incoherent for the majority of the proteins, in line with previous studies displaying a lack of overlap among protein and transcript abundance on the single gene level (94, 95, 148). The latter finding supported the taken bioinformatics approach since gene expression is considered to better mirror protein abundance in terms of functional categories (93). In total, 18 key regulator genes were identified from the networks of which five were novel to oral cancer at the time, but associated to other malignancies e.g., leukemia, sarcoma, breast and renal cancer (149-153). More recently, CEBPA has proven association to HNSCC (154). Western blot analysis of selected low-abundance key regulator genes confirmed deregulated expression in one or both transformed cell lines relative to the normal state for the majority of the assessed transcripts. The transcript level changes, either qualitatively or quantitatively, did not agree completely with noted changes in protein levels. To this end, the integrative omics approach enabled identification of deregulated genes for which probes were not present on the array, e.g., SP1 and SP3. The major result derived from the pipeline concept forwarded herein indicates that expression changes in only a few abundant proteins can identify key changes in cellular pathways.

In PAPER VI application of the integrative biomarker pipeline (PAPER II) was extended from analysis of the novel LK0412 cell line (PAPER IV), thus *in vitro* findings consisting of transcript and protein data were validated in independent oral and non-oral microarray data sets, as well as public transcriptomics and proteomics databases. Proteomic analyses identified 27 abundant proteins (Signature A), of which 24 displayed increased expression in LK0412 relative to the normal state, which is beneficial in terms of being biomarkers. Direct comparisons of the proteins to the transcripts showed low coherence similar to findings in Paper II. The GO-enrichment analysis of the protein profile provided limited overlap with the transcript analysis, however this was not unexpected due to the higher number of assessed transcripts compared to proteins. The initial classification of the normal and TSCC samples provided accuracies between 80-86%, where Signature C displayed the highest accuracy, however following signature reduction, the refined Signature B i.e., Signature E showed higher accuracy rates. The classification power was retained for Signature E following assessment in four independent oral/HNSCC data sets. Overall, this supports the initial idea with the integrative biomarker approach that transcripts within perturbed

categories reflect the biology with higher accuracy compared to highly abundant proteins. To this end, the proteins are still crucial for identifying important categories. Assessment in non-oral cancer data sets overall provided lower classification power in non-epithelial cancers, suggesting that the signature has an "epithelial specific" component. Moreover, the majority of the eight genes was previously coupled to TSCC or could be verified in the transcriptomics, proteomics and/or saliva and plasma protein databases as well as in published literature (134, 155-157). Notably, ISG15, an interferon-regulated gene (158), was already proposed as a biomarker in PAPER IV and was now also included in the most predictive signature. To this end, the integrative biomarker approach enabled classification of TSCC specimen from applying "omic" profiles from only one cell line to levels comparable or even better than what was previously reported for larger gene sets and other studies.

Model-driven approaches for differentiation induction in operative and non-operative cells identify genes with impact on overall survival in HNSCC

In PAPER III, a model-driven approach was taken to define gene signatures that covered functional and non-functional differentiation by confluency culture of NOK and SVpgC2a. Initial, morphological and biochemical assessment of biological fates in confluent NOKs indicated that the proliferation potential was decreased, with prominent commitment to terminal differentiation, and with only minor involvement of apoptosis. Differently, compared to NOK, SVpgC2a consistently showed a lack of epithelial morphology, elevated cloning ability, sustained growth potential, absence of commitment to terminal differentiation and higher rates of apoptosis. Taking into account that minor changes in the balance of biological fates may have significant impact on the tissue homeostasis, the concurrent reduction of apoptosis agrees with the concept that contact-inhibited proliferation may result in fewer mitotic errors and apoptotic abortion of aberrant cells (20, 159). The increased apoptotic frequency in SVpgC2a may thus be due to an increased proliferation rate and a perturbed differentiation capacity.

Differentiation-induction in confluent NOKs was confirmed by downstream Gene Ontology and network analyses of the differently expressed genes. Accordingly, enrichment for "tissue development" and "differentiation"-related categories and functions were indicated. The phenotypic comparison of NOK and SVpgC2a provided the major contribution to the IVDGS and additionally pointed to a negative regulation of differentiation and development-related categories. The latter finding was supported by network-based top functions and key regulator analysis that implicated epithelial-

mesenchymal transition from e.g., decreased expression of E-cadherin (160). The IVDGS was next assessed relative to four extensive microarray data sets from HNSCC patients (129, 132, 133, 137). Totally, 31 survival-associated genes were identified from the Rickman data set, of which 18 genes corroborated with the directional changes (14 good outcome and 4 poor outcome) for HNSCC in the HGEM (132, 133). To this end, the 4-gene signature (COX7A1, MFAP5, MPDU1 and POLD1) displayed significant impact in one of the two independent HNSCC data sets (129, 137). Interestingly, the gene MFAP5, encoding a cell motility-stimulating matrix glycoprotein was associated to HNSCC for the first time (161). Overall, although all signatures did not prove to have predictive power, the results agree with the frequent finding that global transcriptomics studies commonly generate highly different biomarker signatures (60, 119). Availability of larger data sets for initial training of the *in vitro* signatures, including complementation of additional sets to the body-wide expression database, would likely provide higher precision in future analyses. Nevertheless, by applying an *in vitro*-derived signature as starting point, this study was able to reveal prognostic genes that were concealed in the initial *in vivo* analyses.

In PAPER V, a further dimension was added to the model-driven strategy. Differentiation-induction was studied again, this time in the extended *in vitro* model and from exposing the cells to FBS. The microscopic and bioinformatics analyses by single-genes and GO-categories confirmed TD-induction from this selected serum-exposure protocol, and the absence of this response in all of the transformed states. Interestingly, 35 genes overlapped among the serum and confluency-induced responses in NOK, and GO-categorization of these genes enriched only differentiation and development-related categories. A subject for future studies, this signature may be applicable for assessing the role of differentiation-induction capacity as a toxicity outcome. In contrast, serum-induced GOs found in the transformed models pointed to other fates such as epithelia-mesenchymal transition in SVpgC2a i.e., the category "AT-binding" included altered expression of the architectural transcription factors HMGA1 and HMGA2 (162, 163). No overlapping genes were found between the serum and confluency response in SVpgC2a. The malignant cell lines SqCC/Y1 and LK0412 enriched multiple GOs related to stress and wounding, underscoring that malignant cell lines constantly cultured in serum may reflect and enhance their constitutive aggressive nature. Lipid metabolism was implied for all models, and similar enrichment was also noted from overlay and GO-category analyses of the SIGS with a published serum-dependent fibroblast signature that predicted poor outcome in

breast cancer (138, 164). This result indicates preserved response among different cell types and a role of altered lipid metabolism in malignant transformation (165). Overall, taking into account the morphological and gene expression data, serum seems to act primarily to drive differentiation in normal oral keratinocytes. In contrast, the loss of differentiation capacity in the transformed states couples to changes in the biological fate induction typically associated with wound healing. Evaluation of the SIGS to the Rickman data set identified 17 genes with impact on survival. The transcriptomics database HGEM supported the involvement of 10 of these 17 genes in HNSCC, applying a requirement for the same directional change in gene expression as in the Rickman data set (114, 133). The slightly higher degree of precision, compared to PAPER III may stem from application of the extended *in vitro* model, reflecting more entities and stages in cancer progression. Additional validation of the 17 genes in the HPA supported involvement of six of the genes to the protein level. Interestingly, PDGFRL displayed significant impact on poor survival in an independent data set containing 71 HNSCC, differently to the complete 17-gene based signature set (137). The overall results agreed with a role for wound healing-associated effects in cancer development, and provided further evidence that model-driven approaches with cell cultures can reveal novel biomarkers in clinical samples.

An alternative strategy to biomarker discovery using cell cultures and computational biology provide means to replace animal experiments

Especially addressed in PAPER I, this project have had a significant component of addressing the general applicability of cell lines models and the applied technologies also to the need of developing alternative methods under the 3R principle (Replacement, Refinement, Reduction). This principle includes the consideration of replacing traditional animal experiments and animal-based toxicity testing with *in vitro* and *in silico* methods (166). Taking this approach, the Thesis work can be viewed as "an experimental animal-free biomarker discovery pipeline" useful to the consideration of replacing painful and often long-term experiments that are common in cancer and toxicity studies with animals. Recent studies have shown that short-term transcriptional profiling data serve accurately to predict long-term cancer-related safety of a multitude of environmental and industrial chemicals (167). "Standardized serum-free cell culture conditions", "studying both normal and transformed human epithelial cell lines side-by-side", "gene expression profiling" and "bioinformatics processing" are lead terms of this work that makes the pipeline concepts generally applicable to accurate definition of molecular signatures also for chemicals and nanomaterials toxicity, being central areas

within the environmental medicine field. Clearly, bioinformatics processing of all expressed transcripts in the cell serves as a way to describe and overview cellular actions relative to accumulated knowledge in databases (e.g., collected from around three millions of publications in the IPA database). Expression of results in terms of Gene Ontology constitutes a relatively unexplored way to describe toxicity-induced influences on cells. Application of this standardized nomenclature in general toxicology research, hazard characterization and eventually risk assessment is bound to facilitate the application of computational transcript analysis into safety assessments. Additionally, serum-effects were analyzed in models normally cultured without serum (PAPER V). The aspiration for increased reproducibility and reduced experimental variability that comes with the absence of serum is in line with the societal wish for complete removal of animal products from toxicity testing protocols and biomedical research generally. Studies with epithelial cells notably correspond to a clinically relevant cell type for studying toxicity effects of environmental agents since epithelia often serve as a first line of defense. Overall, the research program underlying this Thesis was carried out with the consistent aim of assessing the models from this 3R perspective.

6 CONCLUSIONS

A unique serum-free three-stage cell line model of cancer development was initially shown to exhibit multiple of the phenotypic changes typically noted for cancer development in humans. The model was thereafter considered as a start-up model for further analyses under systems biology concepts, involving detailed gene expression profiling analysis with emphasis on tumor biomarker discovery relative to functional and non-functional cell differentiation. Efforts to extend the model with additional entities were successful from incorporation of a novel cell line, making the overall model progress from a normal-appearing state via an immortalized pre-malignant state further into malignant states, including both a well-differentiated and a moderately differentiated phenotype. Variable assessments, including visual microscopic, transcript and protein expression, as well as functionality-based assays demonstrated impaired cell differentiation in the transformed states. This result demonstrated that the model could be applied to address tumor biomarker discovery as planned. Cell genotype comparisons and comparisons of gene expression changes from influences of different culture protocols under two bioinformatics-based pipeline strategies implicated then multiple transcripts and proteins as potential tumor biomarkers. Implying true validity of the model and for the hypotheses being tested, multiple, but not all, of the differentially expressed single genes, pathways, molecular networks and gene ontologies had previous association to oral cancer, broadly considering also the complete head and neck region. Integration of transcriptomics and proteomics data served not surprisingly to enhance the precision of the biomarker discovery effort, including naturally to decrease the uncertainty of findings from applying one type of omics analysis only. Highly significant to biomarker discovery, the identification of only a handful of abundant, easily to detect proteins enabled elucidation of the majority of pathways altered by cell transformation. This result included the identification of key regulator genes, most of which seem to be transcription factors. Processing of the *in vitro*-derived gene expression profiles via a number of bioinformatics tools, including relative publically available cancer patient data sets and body-wide gene expression databases, generated the necessary training and validation sets that allowed for accurate predictions related to cancer prognosis. Overall, five novel biomarker genes or gene signatures (encompassing between 1 to 9 genes) were generated that variably proved to accurately distinguish normal tissue from tumor tissue, or alternatively to accurately predict outcome in patient-based survival plots. The results therefore proved the overarching hypothesis under testing, namely that systems biology studies under the

concept of *in vitro* modeling of cancer development can serve to identify tumor biomarker genes hidden in complex analysis of cancer patient material. Overall, an extended cell culture model for cancer development was derived, with potential to serve in multiple manners to future wishes of addressing both normal and aberrant cell functions. Especially, the model proved to elucidate further the complex mechanisms of biological fate induction in normal and cancer tissues, as well as to generally demonstrate the usefulness of attempting translational application of *in vitro* results all the way to the human situation. Viewing and analyzing thoroughly human cell lines as unique entities seems to have great potential for contributing valuable information to the many issues and complexities that underlie the future need of personalized medicine in the treatment of human cancer.

7 ACKNOWLEDGEMENTS

bringing a happy smile to the analyses; Lovisa Farnebo for never-ending enthusiasm; Lena Norberg-Spaak for nice collaboration.

Colleagues at the Department of Medical Biochemistry and Biophysics, KI: Professor Hans Jörnvall for providing an excellent proteomics facility; Professor Jan-Olov Höög and Therese Jägerbrink for fruitful collaboration and discussions.

Colleagues at Department of Oral Pathology, King's College London Dental Institute, Professor Peter R. Morgan and Balvinder Bloor for fruitful collaboration.

The Bioinformatic and Expression Analysis Core Facility, KI; David Brodin, for incredible kindness and excellent introduction to the world of bioinformatics; Marika Rönnholm for wonderful technical assistance.

Berit Råsmark, my biology teacher at Göteborgs Högre Samskola, for making biological sciences more interesting than golf.

Märtha and Simo Merne, for providing me with an apartment to rent in Turku, and for making me feel so welcome to Finland.

I would also like to thank my dearest friends for support, love and happiness. The Kai-Larsen family: Ylva for all the fun and crazy adventures during the years and for always caring so much about me – I can't thank you enough! Karin, for encouragement and a crazy sense of humor that light up the day. Love and Hanna for being the cutest kids! A special thanks to Love for occasionally keeping my mind focused on Lego building instead. My running partners for life: Daniel Johansson for your constant positive attitude and ability to make everything seem uncomplicated, and Adrian Solitander for finally moving to Stockholm and bringing unlimited laughter to the running tracks around Djurgården; Elin Raask, my deepest appreciation for your tremendous encouragement and devotion to arrange everything from sports events to amazing dinners, and making me believe that we can succeed in all our pursuits. Karin Aspenberg, for motivation to write this thesis; Malin Eriksson, for being a true source of inspiration and making me see the really important things; Christofer Boman, without you I would not be where I am today; Caroline Liden, for helping me put things in perspective; Olof Carlsson, for endless support and always making me laugh;

Roza Ghotbi, for always being there when needed; Sara Jonsson, for constant encouragement and brilliance.

Mina älskade syskon; Miriam för din ovärderliga omtanke för mig var än i världen du befinner dig. David, för alla fantastiska middagar som du lagar och din inspirerande tro på att ingenting är omöjligt.

Min älskade Mamma och Pappa, ett innerligt tack från hela mitt hjärta, det finns inga ord som kan beskriva hur tacksam jag är för all er kärlek, omtanke och stöd!

8 REFERENCES

1.   Presland, R.B., and Dale, B.A. 2000. Epithelial structural proteins of the skin and oral cavity: function in health and disease. *Crit Rev Oral Biol Med* 11:383-408.
2.   Gartner, L.P. 1994. Oral anatomy and tissue types. *Semin Dermatol* 13:68-73.
3.   Squier, C.A., and Kremer, M.J. 2001. Biology of oral mucosa and esophagus. *J Natl Cancer Inst Monogr*:7-15.
4.   Collins, L.M., and Dawes, C. 1987. The surface area of the adult human mouth and thickness of the salivary film covering the teeth and oral mucosa. *J Dent Res* 66:1300-1302.
5.   Argiris, A., Karamouzis, M.V., Raben, D., and Ferris, R.L. 2008. Head and neck cancer. *Lancet* 371:1695-1709.
6.   Baez, A. 2008. Genetic and environmental factors in head and neck cancer genesis. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 26:174-200.
7.   Presland, R.B., and Jurevic, R.J. 2002. Making sense of the epithelial barrier: what molecular biology and genetics tell us about the functions of oral mucosal and epidermal tissues. *J Dent Educ* 66:564-574.
8.   Lippens, S., Denecker, G., Ovaere, P., Vandenabeele, P., and Declercq, W. 2005. Death penalty for keratinocytes: apoptosis versus cornification. *Cell Death Differ* 12 Suppl 2:1497-1508.
9.   Eckert, R.L., Yaffe, M.B., Crish, J.F., Murthy, S., Rorke, E.A., and Welter, J.F. 1993. Involucrin--structure and role in envelope assembly. *J Invest Dermatol* 100:613-617.
10.   Grafström, R.C. 1990. Carcinogenesis studies in human epithelial tissues and cells in vitro: emphasis on serum-free culture conditions and transformation studies. *Acta Physiol Scand Suppl* 592:93-133.
11.   Sundqvist, K., Liu, Y., Arvidson, K., Ormstad, K., Nilsson, L., Toftgård, R., and Grafström, R.C. 1991. Growth regulation of serum-free cultures of epithelial cells from normal human buccal mucosa. *In Vitro Cell Dev Biol* 27A:562-568.
12.   Poumay, Y., and Pittelkow, M.R. 1995. Cell density and culture factors regulate keratinocyte commitment to differentiation and expression of suprabasal K1/K10 keratins. *J Invest Dermatol* 104:271-276.
13.   Eckert, R.L., and Welter, J.F. 1996. Epidermal keratinocytes - genes and their regulation. *Cell Death Differ* 3:373-383.
14.   Poumay, Y., Jolivet, G., Pittelkow, M.R., Herphelin, F., De Potter, I.Y., Mitev, V., and Houdebine, L.M. 1999. Human epidermal keratinocytes upregulate expression of the prolactin receptor after the onset of terminal differentiation, but do not respond to prolactin. *Arch Biochem Biophys* 364:247-253.
15.   Nilsson, J.A., Hedberg, J.J., Vondracek, M., Staab, C.A., Hansson, A., Höög, J.O., and Grafström, R.C. 2004. Alcohol dehydrogenase 3 transcription associates with proliferation of human oral keratinocytes. *Cell Mol Life Sci* 61:610-617.
16.   Staab, C.A., Ceder, R., Roberg, K., Grafström, R.C., and Höög, J.O. 2008. Serum-responsive expression of carbonyl-metabolizing enzymes in normal and transformed human buccal keratinocytes. *Cell Mol Life Sci* 65:3653-3663.
17.   Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A., et

al. 2011. The mutational landscape of head and neck squamous cell carcinoma. *Science* 333:1157-1160.

18. Leemans, C.R., Braakhuis, B.J., and Brakenhoff, R.H. 2011. The molecular biology of head and neck cancer. *Nat Rev Cancer* 11:9-22.

19. Agrawal, N., Frederick, M.J., Pickering, C.R., Bettegowda, C., Chang, K., Li, R.J., Fakhry, C., Xie, T.X., Zhang, J., Wang, J., et al. 2011. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* 333:1154-1157.

20. Fadeel, B., and Orrenius, S. 2005. Apoptosis: a basic biological phenomenon with wide-ranging implications in human disease. *J Intern Med* 258:479-517.

21. Eckhart, L., Declercq, W., Ban, J., Rendl, M., Lengauer, B., Mayer, C., Lippens, S., Vandenabeele, P., and Tschachler, E. 2000. Terminal differentiation of human keratinocytes and stratum corneum formation is associated with caspase-14 activation. *J Invest Dermatol* 115:1148-1151.

22. Hunter, K.D., Parkinson, E.K., and Harrison, P.R. 2005. Profiling early head and neck cancer. *Nat Rev Cancer* 5:127-135.

23. Parkin, D.M., Bray, F., Ferlay, J., and Pisani, P. 2005. Global cancer statistics, 2002. *CA Cancer J Clin* 55:74-108.

24. Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., and Thun, M.J. 2009. Cancer statistics, 2009. *CA Cancer J Clin* 59:225-249.

25. Severino, P., Alvares, A.M., Michaluart, P., Jr., Okamoto, O.K., Nunes, F.D., Moreira-Filho, C.A., and Tajara, E.H. 2008. Global gene expression profiling of oral cavity cancers suggests molecular heterogeneity within anatomic subsites. *BMC Res Notes* 1:113.

26. Matta, A., and Ralhan, R. 2009. Overview of current and future biologically based targeted therapies in head and neck squamous cell carcinoma. *Head Neck Oncol* 1:6.

27. Schmitz, S., and Machiels, J.P. 2010. Molecular biology of squamous cell carcinoma of the head and neck: relevance and therapeutic implications. *Expert Rev Anticancer Ther* 10:1471-1484.

28. Harris, T.J., and McCormick, F. 2010. The molecular pathology of cancer. *Nat Rev Clin Oncol* 7:251-265.

29. Hanahan, D., and Weinberg, R.A. 2011. Hallmarks of cancer: the next generation. *Cell* 144:646-674.

30. Grandis, J.R., and Tweardy, D.J. 1993. Elevated levels of transforming growth factor alpha and epidermal growth factor receptor messenger RNA are early markers of carcinogenesis in head and neck cancer. *Cancer Res* 53:3579-3584.

31. Ang, K.K., Berkey, B.A., Tu, X., Zhang, H.Z., Katz, R., Hammond, E.H., Fu, K.K., and Milas, L. 2002. Impact of epidermal growth factor receptor expression on survival and pattern of relapse in patients with advanced head and neck carcinoma. *Cancer Res* 62:7350-7356.

32. Pectasides, E., Rampias, T., Kountourakis, P., Sasaki, C., Kowalski, D., Fountzilas, G., Zaramboukas, T., Rimm, D., Burtness, B., and Psyrri, A. 2011. Comparative prognostic value of epidermal growth factor quantitative protein expression compared with FISH for head and neck squamous cell carcinoma. *Clin Cancer Res* 17:2947-2954.

33. Hannigan, A., Smith, P., Kalna, G., Lo Nigro, C., Orange, C., O'Brien, D.I., Shah, R., Syed, N., Spender, L.C., Herrera, B., et al. 2010. Epigenetic downregulation of human disabled homolog 2 switches TGF-beta from a tumor suppressor to a tumor promoter. *J Clin Invest* 120:2842-2857.

34. Chang, J.Y., Wright, J.M., and Svoboda, K.K. 2007. Signal transduction pathways involved in epithelial-mesenchymal transition in oral cancer compared with other cancers. *Cells Tissues Organs* 185:40-47.

35. Kyzas, P.A., Stefanou, D., Batistatou, A., and Agnantis, N.J. 2005. Prognostic significance of VEGF immunohistochemical expression and tumor angiogenesis in head and neck squamous cell carcinoma. *J Cancer Res Clin Oncol* 131:624-630.

36. Zhou, S., Kachhap, S., Sun, W., Wu, G., Chuang, A., Poeta, L., Grumbine, L., Mithani, S.K., Chatterjee, A., Koch, W., et al. 2007. Frequency and phenotypic implications of mitochondrial DNA mutations in human squamous cell cancers of the head and neck. *Proc Natl Acad Sci U S A* 104:7540-7545.

37. Albesiano, E., Davis, M., See, A.P., Han, J.E., Lim, M., Pardoll, D.M., and Kim, Y. 2010. Immunologic consequences of signal transducers and activators of transcription 3 activation in human squamous cell carcinoma. *Cancer Res* 70:6467-6476.

38. Sacks, P.G. 1996. Cell, tissue and organ culture as in vitro models to study the biology of squamous cell carcinomas of the head and neck. *Cancer Metastasis Rev* 15:27-51.

39. Lin, C.J., Grandis, J.R., Carey, T.E., Gollin, S.M., Whiteside, T.L., Koch, W.M., Ferris, R.L., and Lai, S.Y. 2007. Head and neck squamous cell carcinoma cell lines: established models and rationale for selection. *Head Neck* 29:163-188.

40. Izumi, K., Tobita, T., and Feinberg, S.E. 2007. Isolation of human oral keratinocyte progenitor/stem cells. *J Dent Res* 86:341-346.

41. Sundqvist, K., Kulkarni, P., Hybbinette, S.S., Bertolero, F., Liu, Y., and Grafström, R.C. 1991. Serum-free growth and karyotype analyses of cultured normal and tumorous (SqCC/Y1) human buccal epithelial cells. *Cancer Commun* 3:331-340.

42. Kulkarni, P.S., Sundqvist, K., Betsholtz, C., Höglund, P., Wiman, K.G., Zhivotovsky, B., Bertolero, F., Liu, Y., and Grafström, R.C. 1995. Characterization of human buccal epithelial cells transfected with the simian virus 40 T-antigen gene. *Carcinogenesis* 16:2515-2521.

43. Masters, J.R. 2000. Human cancer cell lines: fact and fantasy. *Nat Rev Mol Cell Biol* 1:233-236.

44. Smeets, S.J., van der Plas, M., Schaaij-Visser, T.B., van Veen, E.A., van Meerloo, J., Braakhuis, B.J., Steenbergen, R.D., and Brakenhoff, R.H. 2011. Immortalization of oral keratinocytes by functional inactivation of the p53 and pRb pathways. *Int J Cancer* 128:1596-1605.

45. Opitz, O.G., Suliman, Y., Hahn, W.C., Harada, H., Blum, H.E., and Rustgi, A.K. 2001. Cyclin D1 overexpression and p53 inactivation immortalize primary oral keratinocytes by a telomerase-independent mechanism. *J Clin Invest* 108:725-732.

46. Goessel, G., Quante, M., Hahn, W.C., Harada, H., Heeg, S., Suliman, Y., Doebele, M., von Werder, A., Fulda, C., Nakagawa, H., et al. 2005. Creating oral squamous cancer cells: a cellular model of oral-esophageal carcinogenesis. *Proc Natl Acad Sci U S A* 102:15599-15604.

47. Grafström, R.C. 2002. Human Oral Epithelium. In *Culture of Epithelial Cells*. Edited by Freshney, I.R., and Freshney, M.G. New York, Wiley-Liss, 2002, p195-255

48. Weston, A.D., and Hood, L. 2004. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 3:179-196.

49. Noble, D. 2008. *The music of life - biology beyond genes*: Oxford University Press.

50. Faratian, D., Bown, J.L., Smith, V.A., Langdon, S.P., and Harrison, D.J. 2010. Cancer systems biology. *Methods Mol Biol* 662:245-263.

51. Ideker, T., Dutkowski, J., and Hood, L. 2011. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* 144:860-863.

52. del Sol, A., Balling, R., Hood, L., and Galas, D. 2010. Diseases as network perturbations. *Curr Opin Biotechnol* 21:566-571.

53. Stratton, M.R. 2011. Exploring the genomes of cancer cells: progress and promise. *Science* 331:1553-1558.

54. Bell, D.W. 2010. Our changing view of the genomic landscape of cancer. *J Pathol* 220:231-243.

55. Ali, M.A., and Sjoblom, T. 2009. Molecular pathways in tumor progression: from discovery to functional understanding. *Mol Biosyst* 5:902-908.

56. Lin, B., White, J.T., Lu, W., Xie, T., Utleg, A.G., Yan, X., Yi, E.C., Shannon, P., Khrebtukova, I., Lange, P.H., et al. 2005. Evidence for the presence of disease-perturbed networks in prostate cancer cells by genomic and proteomic analyses: a systems approach to disease. *Cancer Res* 65:3081-3091.

57. Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., Basso, K., Dalla-Favera, R., and Califano, A. 2008. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol* 4:169.

58. Carro, M.S., Lim, W.K., Alvarez, M.J., Bollo, R.J., Zhao, X., Snyder, E.Y., Sulman, E.P., Anne, S.L., Doetsch, F., Colman, H., et al. 2010. The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463:318-325.

59. Järvinen, A.K., Autio, R., Kilpinen, S., Saarela, M., Leivo, I., Grenman, R., Makitie, A.A., and Monni, O. 2008. High-resolution copy number and gene expression microarray analyses of head and neck squamous cell carcinoma cell lines of tongue and larynx. *Genes Chromosomes Cancer* 47:500-509.

60. Viet, C.T., and Schmidt, B.L. 2010. Understanding oral cancer in the genome era. *Head Neck* 32:1246-1268.

61. Gonzalez-Angulo, A.M., Hennessy, B.T., and Mills, G.B. 2010. Future of personalized medicine in oncology: a systems biology approach. *J Clin Oncol* 28:2777-2783.

62. Wang, K., Lee, I., Carlson, G., Hood, L., and Galas, D. 2010. Systems biology and the discovery of diagnostic biomarkers. *Dis Markers* 28:199-207.

63. Hood, L., and Friend, S.H. 2011. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* 8:184-187.

64. Quackenbush, J. 2006. Microarray analysis and tumor classification. *N Engl J Med* 354:2463-2472.

65. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J.M. 1999. Expression profiling using cDNA microarrays. *Nat Genet* 21:10-14.

66. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14:1675-1680.

67. McGall, G., Labadie, J., Brock, P., Wallraff, G., Nguyen, T., and Hinsberg, W. 1996. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci U S A* 93:13555-13560.

68. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. 1999. High density synthetic oligonucleotide arrays. *Nat Genet* 21:20-24.

69. Hubbell, E., Liu, W.M., and Mei, R. 2002. Robust estimators for expression analysis. *Bioinformatics* 18:1585-1592.

70. Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 7:55-65.

71. King, H.C., and Sinha, A.A. 2001. Gene expression profile analysis by DNA microarrays: promise and pitfalls. *JAMA* 286:2280-2288.

72. Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y., et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 24:1151-1161.

73. Mohapatra, S.K., and Krishnan, A. 2011. Microarray data analysis. *Methods Mol Biol* 678:27-43.

74. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249-264.

75. Millenaar, F.F., Okyere, J., May, S.T., van Zanten, M., Voesenek, L.A., and Peeters, A.J. 2006. How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics* 7:137.

76. Seo, J., and Hoffman, E.P. 2006. Probe set algorithms: is there a rational best bet? *BMC Bioinformatics* 7:395.

77. Gyorffy, B., Molnar, B., Lage, H., Szallasi, Z., and Eklund, A.C. 2009. Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLoS One* 4:e5645.

78. Wilson, C.L., and Miller, C.J. 2005. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* 21:3683-3685.

79. Somorjai, R.L., Dolenko, B., and Baumgartner, R. 2003. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* 19:1484-1491.

80. Saeys, Y., Inza, I., and Larranaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23:2507-2517.

81. Tusher, V.G., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98:5116-5121.

82. Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3.

83. Jeffery, I.B., Higgins, D.G., and Culhane, A.C. 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* 7:359.

84. International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945.

85. Jensen, O.N. 2004. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol* 8:33-41.

86. Lottspeich, F. 2009. Introduction to proteomics. *Methods Mol Biol* 564:3-10.

87. Monteoliva, L., and Albar, J.P. 2004. Differential proteomics: an overview of gel and non-gel based approaches. *Brief Funct Genomic Proteomic* 3:220-239.

88. Berth, M., Moser, F.M., Kolbe, M., and Bernhardt, J. 2007. The state of the art in the analysis of two-dimensional gel electrophoresis images. *Appl Microbiol Biotechnol* 76:1223-1243.

89. Domon, B., and Aebersold, R. 2006. Mass spectrometry and protein analysis. *Science* 312:212-217.

90. Lubec, G., Afjehi-Sadat, L., Yang, J.W., and John, J.P. 2005. Searching for hypothetical proteins: theory and practice based upon original data and literature. *Prog Neurobiol* 77:90-127.

91. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., and Aebersold, R. 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* 97:9390-9395.

92. Kuruma, H., Egawa, S., Oh-Ishi, M., Kodera, Y., and Maeda, T. 2005. Proteome analysis of prostate cancer. *Prostate Cancer Prostatic Dis* 8:14-21.

93. Greenbaum, D., Jansen, R., and Gerstein, M. 2002. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* 18:585-596.

94. Tian, Q., Stepaniants, S.B., Mao, M., Weng, L., Feetham, M.C., Doyle, M.J., Yi, E.C., Dai, H., Thorsson, V., Eng, J., et al. 2004. Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol Cell Proteomics* 3:960-969.

95. Maier, T., Guell, M., and Serrano, L. 2009. Correlation of mRNA and protein in complex biological samples. *FEBS Lett* 583:3966-3973.

96. Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M.J., Sauter, G., and Kallioniemi, O.P. 1998. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 4:844-847.

97. Uhlen, M., Björling, E., Agaton, C., Szigyarto, C.A., Amini, B., Andersen, E., Andersson, A.C., Angelidou, P., Asplund, A., Asplund, C., et al. 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* 4:1920-1932.

98. Pontén, F., Schwenk, J.M., Asplund, A., and Edqvist, P.H. 2011. The Human Protein Atlas as a proteomic resource for biomarker discovery. *J Intern Med* 270:428-446.

99. Luscombe, N.M., Greenbaum, D., and Gerstein, M. 2001. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* 40:346-358.

100. Quackenbush, J. 2001. Computational analysis of microarray data. *Nat Rev Genet* 2:418-427.

101. Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-14868.

102. Gollub, J., and Sherlock, G. 2006. Clustering microarray data. *Methods Enzymol* 411:194-213.

103. Li, L. 2010. Dimension reduction for high-dimensional data. *Methods Mol Biol* 620:417-434.

104. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25-29.

105. Zhang, B., Schmoyer, D., Kirov, S., and Snoddy, J. 2004. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* 5:16.

106. Zhang, B., Kirov, S., and Snoddy, J. 2005. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* 33:W741-748.

107. Huang da, W., Sherman, B.T., and Lempicki, R.A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44-57.

108. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 38:D355-360.

109. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29:365-371.

110. Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P.A., Julian, R.K., Jr., Jones, A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., et al. 2007. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* 25:887-893.

111. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., et al. 2011. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res* 39:D1005-1010.

112. Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., et al. 2011. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res* 39:D1002-1004.

113. Kilpinen, S., Autio, R., Ojala, K., Iljin, K., Bucher, E., Sara, H., Pisto, T., Saarela, M., Skotheim, R.I., Bjorkman, M., et al. 2008. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol* 9:R139.

114. Lukk, M., Kapushesky, M., Nikkila, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., and Brazma, A. 2010. A global map of human gene expression. *Nat Biotechnol* 28:322-324.

115. Autio, R., Kilpinen, S., Saarela, M., Kallioniemi, O., Hautaniemi, S., and Astola, J. 2009. Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. *BMC Bioinformatics* 10 Suppl 1:S24.

116. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. 2010. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol* 28:1248-1250.

117. van't Veer, L.J., and Bernards, R. 2008. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452:564-570.

118. Sotiriou, C., and Piccart, M.J. 2007. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer* 7:545-553.

119. Koscielny, S. 2010. Why most gene expression signatures of tumors have not been useful in the clinic. *Sci Transl Med* 2:14ps12.

120. Ziober, A.F., D'Alessandro, L., and Ziober, B.L. 2010. Is gene expression profiling of head and neck cancers ready for the clinic? *Biomark Med* 4:571-580.

121. van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530-536.

122. Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., et al. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 351:2817-2826.

123. Ferte, C., Andre, F., and Soria, J.C. 2010. Molecular circuits of solid tumors: prognostic and predictive tools for bedside use. *Nat Rev Clin Oncol* 7:367-380.

124. Wistuba, II, Gelovani, J.G., Jacoby, J.J., Davis, S.E., and Herbst, R.S. 2011. Methodological and practical challenges for personalized cancer therapies. *Nat Rev Clin Oncol* 8:135-141.

125. Knezevic, V., Leethanakul, C., Bichsel, V.E., Worth, J.M., Prabhu, V.V., Gutkind, J.S., Liotta, L.A., Munson, P.J., Petricoin, E.F., 3rd, and Krizman, D.B. 2001. Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics* 1:1271-1278.

126. Nassiri, M., and Nadji, M. 2006. Tissue detection of biomolecular predictors in breast cancer. *Expert Rev Anticancer Ther* 6:1225-1232.

127. Roesch-Ely, M., Nees, M., Karsai, S., Ruess, A., Bogumil, R., Warnken, U., Schnolzer, M., Dietz, A., Plinkert, P.K., Hofele, C., et al. 2007. Proteomic analysis reveals successive aberrations in protein expression from healthy mucosa to invasive head and neck cancer. *Oncogene* 26:54-64.

128. Sandberg, R., and Ernberg, I. 2005. The molecular portrait of in vitro growth by meta-analysis of gene-expression profiles. *Genome Biol* 6:R65.

129. Chung, C.H., Parker, J.S., Karaca, G., Wu, J., Funkhouser, W.K., Moore, D., Butterfoss, D., Xiang, D., Zanation, A., Yin, X., et al. 2004. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell* 5:489-500.

130. Kuriakose, M.A., Chen, W.T., He, Z.M., Sikora, A.G., Zhang, P., Zhang, Z.Y., Qiu, W.L., Hsu, D.F., McMunn-Coffran, C., Brown, S.M., et al. 2004. Selection and validation of differentially expressed genes in head and neck cancer. *Cell Mol Life Sci* 61:1372-1383.

131. Toruner, G.A., Ulger, C., Alkan, M., Galante, A.T., Rinaggio, J., Wilk, R., Tian, B., Soteropoulos, P., Hameed, M.R., Schwalb, M.N., et al. 2004. Association between gene expression profile and tumor invasion in oral squamous cell carcinoma. *Cancer Genet Cytogenet* 154:27-35.

132. O'Donnell, R.K., Kupferman, M., Wei, S.J., Singhal, S., Weber, R., O'Malley, B., Cheng, Y., Putt, M., Feldman, M., Ziober, B., et al. 2005. Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity. *Oncogene* 24:1244-1251.

133. Rickman, D.S., Millon, R., De Reynies, A., Thomas, E., Wasylyk, C., Muller, D., Abecassis, J., and Wasylyk, B. 2008. Prediction of future metastasis and molecular characterization of head and neck squamous-cell carcinoma based on transcriptome and genome analysis by microarrays. *Oncogene* 27:6607-6622.

134. Ye, H., Yu, T., Temam, S., Ziober, B.L., Wang, J., Schwartz, J.L., Mao, L., Wong, D.T., and Zhou, X. 2008. Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics* 9:69.

135. Yu, Y.H., Kuo, H.K., and Chang, K.W. 2008. The evolving transcriptome of head and neck squamous cell carcinoma: a systematic review. *PLoS One* 3:e3215.

136. Estilo, C.L., P, O.c., Talbot, S., Socci, N.D., Carlson, D.L., Ghossein, R., Williams, T., Yonekawa, Y., Ramanathan, Y., Boyle, J.O., et al. 2009. Oral tongue cancer gene expression profiling: Identification of novel potential prognosticators by oligonucleotide microarray analysis. *BMC Cancer* 9:11.

137. Thurlow, J.K., Pena Murillo, C.L., Hunter, K.D., Buffa, F.M., Patiar, S., Betts, G., West, C.M., Harris, A.L., Parkinson, E.K., Harrison, P.R., et al. 2010. Spectral clustering of microarray data elucidates the roles of microenvironment remodeling and immune responses in survival of head and neck squamous cell carcinoma. *J Clin Oncol* 28:2881-2888.

138. Chang, H.Y., Sneddon, J.B., Alizadeh, A.A., Sood, R., West, R.B., Montgomery, K., Chi, J.T., van de Rijn, M., Botstein, D., and Brown, P.O. 2004. Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* 2:E7.

139. Fagan, A., Culhane, A.C., and Higgins, D.G. 2007. A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics* 7:2162-2171.

140. Dvorak, H.F. 1986. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. *N Engl J Med* 315:1650-1659.

141. Pedersen, T.X., Leethanakul, C., Patel, V., Mitola, D., Lund, L.R., Dano, K., Johnsen, M., Gutkind, J.S., and Bugge, T.H. 2003. Laser capture microdissection-based in vivo genomic profiling of wound keratinocytes identifies similarities and differences to squamous cell carcinoma. *Oncogene* 22:3964-3976.

142. Schliephake, H. 2003. Prognostic relevance of molecular markers of oral cancer--a review. *Int J Oral Maxillofac Surg* 32:233-245.

143. Ginos, M.A., Page, G.P., Michalowicz, B.S., Patel, K.J., Volker, S.E., Pambuccian, S.E., Ondrey, F.G., Adams, G.L., and Gaffney, P.M. 2004. Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res* 64:55-63.

144. Yan, B., Yang, X., Lee, T.L., Friedman, J., Tang, J., Van Waes, C., and Chen, Z. 2007. Genome-wide identification of novel expression signatures reveal distinct patterns and prevalence of binding motifs for p53, nuclear factor-kappaB and other signal transcription factors in head and neck squamous cell carcinoma. *Genome Biol* 8:R78.

145. Vachani, A., Nebozhyn, M., Singhal, S., Alila, L., Wakeam, E., Muschel, R., Powell, C.A., Gaffney, P., Singh, B., Brose, M.S., et al. 2007. A 10-gene classifier for distinguishing head and neck squamous cell carcinoma and lung squamous cell carcinoma. *Clin Cancer Res* 13:2905-2915.

146. Xie, H., Onsongo, G., Popko, J., de Jong, E.P., Cao, J., Carlis, J.V., Griffin, R.J., Rhodus, N.L., and Griffin, T.J. 2008. Proteomics analysis of cells in whole saliva from oral cancer patients via value-added three-dimensional peptide fractionation and tandem mass spectrometry. *Mol Cell Proteomics* 7:486-498.

147. Cervigne, N.K., Reis, P.P., Machado, J., Sadikovic, B., Bradley, G., Galloni, N.N., Pintilie, M., Jurisica, I., Perez-Ordonez, B., Gilbert, R., et al. 2009. Identification of a microRNA signature associated with progression of leukoplakia to oral carcinoma. *Hum Mol Genet* 18:4818-4829.

148. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. 2011. Global quantification of mammalian gene expression control. *Nature* 473:337-342.

149. Folpe, A.L. 2002. MyoD1 and myogenin expression in human neoplasia: a review and update. *Adv Anat Pathol* 9:198-203.

150. Lucas, B., Grigo, K., Erdmann, S., Lausen, J., Klein-Hitpass, L., and Ryffel, G.U. 2005. HNF4alpha reduces proliferation of kidney cells and affects genes deregulated in renal cell carcinoma. *Oncogene* 24:6418-6431.

151. Miyamoto, K., Fukutomi, T., Akashi-Tanaka, S., Hasegawa, T., Asahara, T., Sugimura, T., and Ushijima, T. 2005. Identification of 20 genes aberrantly methylated in human breast cancers. *Int J Cancer* 116:407-414.

152. Borghini, S., Vargiolu, M., Di Duca, M., Ravazzolo, R., and Ceccherini, I. 2006. Nuclear factor Y drives basal transcription of the human TLX3, a gene overexpressed in T-cell acute lymphocytic leukemia. *Mol Cancer Res* 4:635-643.

153. Schuster, M.B., and Porse, B.T. 2006. C/EBPalpha: a tumour suppressor in multiple tissues? *Biochim Biophys Acta* 1766:88-103.

154. Bennett, K.L., Hackanson, B., Smith, L.T., Morrison, C.D., Lang, J.C., Schuller, D.E., Weber, F., Eng, C., and Plass, C. 2007. Tumor suppressor activity of CCAAT/enhancer binding protein alpha is epigenetically down-regulated in head and neck squamous cell carcinoma. *Cancer Res* 67:4657-4664.

155. Chi, L.M., Lee, C.W., Chang, K.P., Hao, S.P., Lee, H.M., Liang, Y., Hsueh, C., Yu, C.J., Lee, I.N., Chang, Y.J., et al. 2009. Enhanced interferon signaling pathway in oral cancer revealed by quantitative proteome analysis of microdissected specimens using 16O/18O labeling and integrated two-dimensional LC-ESI-MALDI tandem MS. *Mol Cell Proteomics* 8:1453-1474.

156. Chen, Y.J., Lin, S.C., Kao, T., Chang, C.S., Hong, P.S., Shieh, T.M., and Chang, K.W. 2004. Genome-wide profiling of oral squamous cell carcinoma. *J Pathol* 204:326-332.

157. Shigeishi, H., Fujimoto, S., Hiraoka, M., Ono, S., Taki, M., Ohta, K., Higashikawa, K., and Kamata, N. 2009. Overexpression of the receptor for hyaluronan-mediated motility, correlates with expression of microtubule-associated protein in human oral squamous cell carcinomas. *Int J Oncol* 34:1565-1571.

158. Jeon, Y.J., Yoo, H.M., and Chung, C.H. 2010. ISG15 and immune diseases. *Biochim Biophys Acta* 1802:485-496.

159. Foo, R.S., Mani, K., and Kitsis, R.N. 2005. Death begets failure in the heart. *J Clin Invest* 115:565-571.

160. Kalluri, R., and Weinberg, R.A. 2009. The basics of epithelial-mesenchymal transition. *J Clin Invest* 119:1420-1428.

161. Mok, S.C., Bonome, T., Vathipadiekal, V., Bell, A., Johnson, M.E., Wong, K.K., Park, D.C., Hao, K., Yip, D.K., Donninger, H., et al. 2009. A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. *Cancer Cell* 16:521-532.

162. Reeves, R., Edberg, D.D., and Li, Y. 2001. Architectural transcription factor HMGI(Y) promotes tumor progression and mesenchymal transition of human epithelial cells. *Mol Cell Biol* 21:575-594.

163. Watanabe, S., Ueda, Y., Akaboshi, S., Hino, Y., Sekita, Y., and Nakao, M. 2009. HMGA2 maintains oncogenic RAS-induced epithelial-mesenchymal transition in human pancreatic cancer cells. *Am J Pathol* 174:854-868.

164. Chang, H.Y., Nuyten, D.S., Sneddon, J.B., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H., He, Y.D., van't Veer, L.J., Bartelink, H., et al. 2005. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A* 102:3738-3743.

165. Menendez, J.A., and Lupu, R. 2007. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat Rev Cancer* 7:763-777.

166. Collins, F.S., Gray, G.M., and Bucher, J.R. 2008. Toxicology. Transforming environmental health protection. *Science* 319:906-907.

167. Thomas, R.S., Bao, W., Chu, T.M., Bessarabova, M., Nikolskaya, T., Nikolsky, Y., Andersen, M.E., and Wolfinger, R.D. 2009. Use of short-term transcriptional profiles to assess the long-term cancer-related safety of environmental and industrial chemicals. *Toxicol Sci* 112:311-321.