Tumor methylation markers and clinical outcome of primary oral squamous cell carcinomas
Clausen, Martijn

DOI:
10.33612/diss.113437849

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 7

General Discussion
In recent years great progress has been made in improving detection of LN metastasis in OSCC using clinical and histological parameters in the primary tumor. However, the detection of the presence of LN using these parameters, cannot be applied on a specific subset of OSCC that appear to behave differently than the majority of the OSCC. OSCC of this particular subset initially appear as low risk and as clinically negative for nodal spread, but eventually develop LN metastases. Since these OSCC cannot be distinguished from the other cN0 OSCC using the current clinical and histological parameters, other biomarkers might prove to be helpful in identifying this subset of tumors that potentially develop subclinical metastases in the neck ensuring the proper elective treatment for these LN metastases.

In this thesis the identification and application of markers regulated by DNA methylation for use as predictor for biological behavior of Oral Squamous Cell Carcinomas is reported.

In chapter 2, we first report on the evolution of DNA methylation markers from the literature associated with LN metastasis in our cohort of OSCC [278]. From a panel of 28 genes only the methylation status of DAPK and MGMT were predictive for pN-status of the neck in OSCC. Although the negative predictive value of the methylation status of DAPK1 and MGMT combined was 76%, it did not outperform current clinical nodal staging techniques such as sentinel lymph node biopsy with a reported negative predictive value 88 to 95% [25], [26], [193], [219]. DAPK1 and MGMT are two of the most widely studied methylated genes and have been reported the most for clinical application. However, these two genes aren’t sufficient to solve the clinical negative neck dilemma [209], [395], [396]. Thus, to improve upon the predictive value using a DNA methylation gene panel predictive for N-status in OSCC, new DNA methylation markers are necessary and genome-wide discovery approaches are needed to identify these novel DNA methylation markers.

Using genome-wide discovery MethylCap-Seq analysis followed by statistical analysis as well as in silico validation, we identified and characterized three different methylation markers predictive for the presence of lymph node metastasis in OSCC. WISP1, RAB25 and S100A9. WISP1 was found to be hypomethylated and overexpressed in pN+ OSCC (Chapter 3), while RAB25 (Chapter 4) and S100A9 (Chapter 5) are hypermethylated and down regulated in pN+ OSCC. WISP1 and S100A9 aberrant protein levels and DNA methylation levels were also found to be related to patients’ survival. Our analyses revealed that all three DNA methylation biomarkers can be used to improve current diagnostic modalities and potentially provide additional treatment options. Finally, WISP1 is involved in the Wnt-pathway, RAB25 in the Raf/MEK/ERK pathways and S100A9 is a well-known calcium-binding protein, and calcium levels are both involved in the Wnt-pathway and the Raf/MEK/ERK pathways.

Unfortunately, the studies reported in this thesis revealed that the predictive value for the individual patient of these newly discovered biomarkers is not outperforming current diagnostic modalities [77], [397], [398]. Especially, when comparing the negative predictive values of these DNA methylation biomarkers to recent advancements as the Sentinel Node Biopsies [25], [26]. Summarizing, the methylation status of WISP1, RAB25 and S100A9 were found to be not clinically applicable biomarkers as compared to for example MGMT in glioblastomas [114]. There are several potential reasons for why the impact of methylation of these three genes is not as predictive for the phenotype of OSCC as that of MGMT in glioblastomas.
First of all, the OSCC methylome is more variable between OSCC cases compared to other cancers. This broader variation in methylomes could be due to OSCC life-style related risk factors which drive carcinogenesis also happen to influence genome-wide methylation levels. The main tumorigenic factors in OSCC are cigarette smoking and alcohol consumption [2]. The combination of both drinking and smoking is especially hazardous and synergistic [2]. Interestingly, both alcohol consumption and cigarette smoking have been shown to be related to global DNA hypomethylation in head and neck cancer which is one of the new generation of hallmark of cancer as defined by Hanahan and Weinberg in 2011 [97], [375], [399]. More specifically, both alcohol and cigarette smoke are known to cause inhibition of the DNMT proteins responsible for DNA methylation maintenance, causing passive global demethylation (reviewed by [400]. For example, alcohol consumption is known to down regulate the mRNA levels of DNMT3a and DNMT3b [401]. In addition, DNMT1 activity can be down-regulated as a result of interaction with ethanol resulting passive genome-wide demethylation (reviewed by [401]). Specifically, alcohol consumption was found to be related to hypermethylation of several genes involved in HNSCC such as CDKN2A and E-cadherin [402]. Besides alcohol, DNMT1 activity is also downregulated by tobacco smoke contributing to the same passive genome-wide hypomethylation as influenced by ethanol (reviewed by [403],[404]).

These aberrations of a normal methylation machinery is exemplary of a carcinogenic methylome [403]. Global hypomethylation might be further stimulated by cigarette smoking through DNMT3b downregulation [405]. Finally, several studies report cigarette smoking-induced hypermethylation of several specific genes in different types of cancers [406], [407]. The direct effects of the etiologic effects of OSCC, and especially the effect on the DNMT family, could cause a wider variability of methylation changes in OSCC that drive tumor growth compared to other cancers in tissue less frequently exposed to cigarette smoke and alcohol.

The diversity of carcinogenic aberrations in oral squamous cell carcinoma is further supported by the presence of field cancerization in the oral cavity [3]. Field cancerization is the process where several (pre-)neoplastic lesions develop simultaneously at multiple locations, under the assumption that these premalignant sites develop independently [3]. Field cancerization explains the occurrence of second primary tumors in this area. This field-effect illustrates the diversity and scale of the collection of genetic aberrations in the oral cavity in relation to the etiology of the variability in OSCC methylomes. The high impact on DNA methylation of alcohol consumption and cigarette smoking also explains the absence of a universal marker like MGMT in glioblastoma because risk factors for glioblastomas are radiation exposure and certain genetic syndromes which could cause more homogenous drivers of carcinogenesis [408].

A second reason for not finding a universal methylation marker associated with pN-status in OSCC is the focus on more conventional genomic loci as well as large changes in epigenetic changes in our experimental design. Traditionally, methylation studies focused on changes in methylation levels in CpG rich regions, specifically CpG islands and especially those CpG islands associated with Transcription Start Sites. To prevent being limited to CpG islands we chose MethylCcap-Seq analysis, a genome-wide platform as a discovery tool for DNA methylation biomarkers [136]. However, the protein that was used for methylated DNA fragment enrichment (MBD2) only enriches DNA fragments that have a certain minimum amount of methylated CpG within a certain minimum distance of each other [136]. This means that certain
DNA fragments with either a single or just a few methylated CpG or with too much space between these methylated CpGs, could be missed by this MBD2 enrichment. The potential impact of this MBD2 bias for CpG rich DNA fragments has become more apparent over the last few years due to the discovery of the biological relevance of DNA methylation of CpG-poor regions. Originally, DNA methylation research was mainly directed at CpG Islands. However, meta-analysis of Illumina 450k data revealed high transcriptional regulation by DNA methylation in CpG island neighboring regions. Specifically, DNA methylation of regions up to 2000 bp adjacent on either sides of CpG islands, referred to as CpG Island Shores, were identified as having a strong regulatory function on gene expression [193]. Additionally, the next 2000 bp further upstream and downstream of these CpG Island Shores were also found to be involved in epigenetic regulation of gene transcription and were dubbed CpG shelves (Figure 7.1) [246].

Besides regions of differential methylation, some gene regulatory regions were identified that across different normal tissues have very stable methylation levels. These regions have been called CpG Canyons and CpG Ravines [409], [410]. The consistency of these methylation patterns across tissue suggests an important biological function of these regions and hypo- or hypermethylation of these regions could be especially pathogenic.

Besides the before mentioned technical bias of MBD2 to more highly methylated fragments, we have introduced biases towards conventional regions of DNA methylation in our algorithms and in silico analyses. First of all, we selected regions located 2000 bp upstream to 500 bp downstream of TSS. Additionally, we focused on differential methylation of Methylation Cores identified in the map of the human Methylome [168], [382]. The recent addition of these Shores, Shelves, Ravines and Canyons to the targets of DNA methylation research expand and complicate the selection of DNA methylation techniques. Moreover, enhancer sequences, which lie distal to TSS but still impact gene expressions can also be epigenetically regulated [411]. Their distance to CpG islands associated with genes complicates the epigenetic regulation of a gene and could be easily overlooked in experiment design and data analysis [412]. While the ENCODE [413], [414] and FANTOM5 [415] projects have made leaps forward in identifying these distal regulatory regions, these enhancers are often not included in methylation studies. For example, the widely used Illumina 450k lacks these probes covering these long-distance enhancers [412]. In our experimental design we focused only on the transcriptional impact of methylation of regions in the proximity of gene transcription start sites. So, most studies might only cover a very limited part of any the epigenetic iceberg.

**Figure 7.1. The relative locations of CpG shelves, CpG shores and CpG islands.** CpG islands are genomic regions that confirm to certain statistics that define these loci as especially CpG site rich. The adjacent 2 kb upstream and downstream of these CpG islands are referred to as CpG Island Shores. The next 2kb further downstream or upstream of CpG islands and CpG island shores are called CpG shelves. Any regions further located from CpG islands are referred to as the Open Sea.
Thirdly, our studies focused on methylated cytosines as a very stable compound. However, the methylation of cytosines is a much more complicated process. The epigenetic machinery, consistent besides DNA methyltransferases also of the TET protein family. The three ten-eleven translocation (TET) proteins, TET1, TET2 and TET3, facilitate the oxidation of methylated cytosines to hydroxymethyl-cytosines, formyl-cytosines and carbo-cytosines (reviewed by [416], [417]), which are intermediates between methylated and unmethylated cytosines (Figure 7.2).

Interestingly, there is much support of a role of TET proteins in carcinogenesis. TET proteins are frequently mutated in tumors and TET proteins has been shown to function as tumor suppressor genes [94], [416], [417], [419]. Regardless of their role in tumorigenesis, TET proteins are often neglected in DNA methylation studies. While occurrences of these TET mediated cytosine modifications is much more rare than DNA methylation, about 0.03 to 0.7% of all cytosines, compared to a 5% of methylated cytosines [420], a major concern is that the widely used bisulfite-based methods of methylation detection such as MSP and bisulfite Sequencing do not differentiate between methylated and the three TET catalyzed intermediates as all different modified cytosines are converted to uracil by bisulfite treatment [421]. Moreover, in general the effects of DNA methylation and the TET DNA methylation intermediates on gene expression are opposite, causing inaccurate methylation detection as well as wrong conclusions of the effects of the detected methylation of biological behavior [417]. In contrast, MethylCap-Seq analysis does distinguish between these different cytosine modifications because the MBD2 enrichment is only specific for methylated cytosines. However, because we performed further gene validation by MSP and since bisulfite Sequencing can’t distinguish between the different cytosine modifications [422], the presence of these non-methylation modified cytosines might have distorted the relation between methylation and protein expression found in our studies resulting in an overestimate of the actually DNA methylation levels and

Figure 7.2. TET protein mediated methylated cytosine (SmC) modifications. The family of TET proteins further modify methylated cytosines (SmC) to hydroxymethyl-cytosines (SmhC), formyl-cytosines (SmF) and carbo-cytosines (SmC). Both SmF and SmC are intermediates for active demethylation of cytosines by base excision repair (BER). Adapted from [418].
thus underestimate of the epigenetic regulation of the studies genes.

In general DNA methylation of gene regulatory sequences results in the down-regulation of the associated genes’ expression. However, the relation between methylation and gene expression is further complicated by the observation that some genes appear to be upregulated by hypermethylation. For example, the hTERT gene, a gene involved in regulating telomerase activity, has been found to be upregulated by an increase in DNA methylation [423], [424]. Furthermore, traditionally, gene expression regulation by CpG site methylation had been thought to be confined to the CpG islands. However, as discussed above studies have shown that methylation of CpG island shores might have an even greater influence on gene expression regulation [193]. The influence of DNA methylation on gene expression has been shown to also be capable of long-range interaction between CpGs and target genes [425], [426]. These findings contest the traditionally model that CpG methylation influences gene expression solely by methylation of CpG islands that overlap with regulatory regions. More importantly, a recent genomic study reported that the influence of long-range DNA methylation on its long-range target genes is far greater than the influence of CpG island methylation [427]. Additionally, the same study that found tumor and normal tissue display far greater differences in methylation levels of the long-range regulatory regions [427]. Therefore, it is not prudent to assume that CpG Islands always influence the gene that is closest. To properly identify epigenetic regulation of genes by DNA methylation, direct correlation of mRNA levels and DNA methylation is insufficient. This might explain why we did not find a significant correlation between RAB25 expression and RAB25 TSS methylation. Incorporation of long-range regulatory regions data should be part of algorithms for the identification and characterization of DNA methylation biomarkers.

Besides huge differences between different primary tumors, within a single tumor also lots of variability are present due to tumor heterogeneity. A single tumor can originate from various sub clones that individually collected different genetic aberrations. Differences in these carcinogenic drives can cause variability in molecular and biological differences within a single tumor [428]. A recent study found that the amount of tumor heterogeneity itself is correlated with the presence of LN metastases [429]. Even though, our studies took tumor heterogeneity into account during biomarkers validation by taking at least three different biopsies from a single tumor sample based on HE-coupes for our tissue microarray studies, it’s not possible to completely compensate for tumor heterogeneity. The true variability within a tumor is unclear and difficult to assess, and probably would require sequencing of many different tumor subpopulations acquired by macro-dissection to identify clonal subpopulations [430]. Possible heterogeneity of the tumor DNA used for our MethylCap-Seq could have obscured consistent methylation data due to the presence of several tumor subclones. The chance of heterogeneity in OSCC is especially high due to field cancerization as discussed earlier [3]. Because of the nature of the etiology and risk-factors of OSCC such as tobacco smoking and alcohol consumption, the whole oral epithelial is simultaneously subjected to the same risk factors causing a lot of different pre-cancerous genetic aberrations can arise within the oral epithelial layers increasing the risk of tumor subclones.

Finally, the microenvironment of a cancer is an important contributor to the tumorigenesis and might provide another variable to connect our OSCC methylome with the metastatic phenotype [431].
influence of macrophages, fibroblasts and endothelial cells are known to partially drive the nodal spread of OSCC, through direct cell-to-cell interactions and paracrine signaling (reviewed by [431]). Additionally, it is possible that these cells were present in the OSCC tissue used for MethylCap-Seq analysis we cannot exclude any influence of these cells in the assessment of our methylome.

While OSCC methylation markers are regularly reported in the literature, a universal OSCC N-status methylation marker has currently not been found yet. Major bottleneck in the process of marker identification is the intra-center validation. Often major variations, improper or incompletely reported in patient population and selection, technique selection and execution, sample size or experimental design prevent successful marker validation. Recent endeavors such as The Cancer Genome Atlas, The International Cancer Genome Consortium, and the Gene Expression Omnibus have provided platforms to make intra-cohort validation easier [409], [432], [433]. Both databases provide huge amounts of publicly available datasets as well as analysis tools for easy validation of markers [434], [435].

A possible biological model for the OSCC metastatic methylome

We report on three different approaches for the identification and selection of DNA methylation markers that are predictive for the presence of nodal metastases in OSCC. All three approaches have led to the identification of a different biomarker: WISP1, RAB25 and S100A9. Initially, these three genes have different functions and are involved in different pathways, showing no direct connection or overlapping pathway. However, all three proteins and their cellular functions can eventually be traced to calcium signaling and calcium levels. For example, WISP1 is part of the Wnt-pathway in which calcium has been shown to act as a second messenger [345], [346]. Additionally, RAB25 is part of the Raf/MEK/ERK pathway which is involved in cellular differentiation and this process is known to be induced by calcium stimulation [347]. And finally, S100A9 is a well-known protein directly binding calcium signaling pathway components [316].

Additionally, we performed a pathway analysis using The Database for Annotation, Visualization and Integrated Discovery (DAVID) [349], [350] on 887 differentially methylated genes identified and selected with MethylCap-Seq. This analysis revealed that the calcium signaling pathway is the only significantly enriched pathway amongst these genes. This confirms the calcium signaling pathway as the main contributor to the metastatic phenotype of pN+ OSCC.

In the introduction we discussed nine major pathways in which the biomarkers predictive for LN metastasis in OSCC can be classified: Cell cycle regulation, proliferation and apoptosis; Cell motility, cell adhesion and microenvironment; Transcription factors, immune system and angiogenesis [75]. Interestingly, all of these nine pathways have been shown to be related to the calcium signaling: cell cycle regulation [351], [352], cell proliferation [353], [354], apoptosis [351], [352], [354], cell motility [355], cell adhesion [352], [356], [357], microenvironment [354], [358] transcription factors [354], [359], the immune system [352], [360], and angiogenesis [352], [361], [362].
Summarizing, the calcium pathway appears to be the interconnecting factor between all three identified DNA methylation markers. Interestingly, the calcium pathway can be linked through several mechanisms to metastasis, specifically to mesenchymal-to-epithelial-transition as well as epithelial-to-mesenchymal-transition [436].

In addition, calcium promotes proteolysis of the extracellular matrix detaching cells from the matrix, enhanced migration, trailing end contraction of cells allowing cell motility, phosphorylation of contractile proteins activating them, enhancing actin assembly promoting a dynamic cytoskeleton and thus motility of cells, as well as enhancing turnover of adhesion molecules [437]. On the other side, calcium is also involved in other process associated with carcinogenesis such as mesenchymal-to-epithelial transition, apoptosis, inducing cell differentiation, inhibiting cell proliferation and inducing autophagy-like cell death[437].

Furthermore, changes in the DNA methylation patterns occur in close association with calcium signaling during carcinogenesis [438]. And finally, all three DNMTs have been found to be Ca2+ ion- and redox state-dependent, providing a direct link between the aberrant methylation in OSCC as well as the Calcium related genes found to be epigenetically deregulated in OSCC [439]. It must be noted though that very high Ca2+ concentrations are required before this effect is seen and therefore the clinical relevance of this finding might prove limited.

We hypothesize, that the changes of DNA methylation during carcinogenesis causes a shift in calcium-induced metastasis by inhibiting tumor suppressor genes through promoter hypermethylation of RAB25 and S100A9 as well as hypomethylation and upregulation of WISP1 (Figure 7.3) based upon the identification and characterization of these three new DNA methylation biomarkers and their role in the calcium pathway analysis as well as the pathway analysis of our shortlists of genes annotated by MethylCap-Seq analysis.
OSCC DNA methylation potential therapeutic targets

The three novel methylation markers might provide viable therapeutic targets for pN+ OSCC patients. Therapeutic upregulation of WISP1 might already be soon clinically possible due to the nature of the epigenetic changes that lead to overexpression of the WISP1 protein. Although no direct medication for WISP1 is available, drugs targeting the Wnt-pathway, of which WISP1 is a part, are being developed such as Compound LGK974 (WNT974) which specifically is being developed for treatment of several cancers including HNSCC [440], [441]. The epigenetic repression of RAB25 and S100A9 might be less easily directly targeted due to the lowered levels of these proteins as a result of hypermethylation. The epigenetic down-regulation of RAB25 and S100A9 could be treated with general demethylating medication. Two inhibitors of DNMT activity have already been approved by the FDA. Both Azacitidine and Decitabine are being used in the clinic to reduce overall DNA methylation in myelodysplastic syndromes [365], [366]. In 2014 a clinical trial (NCT02178072) started where HNSCC patients were treated with Azacitidine [367]. In fact, several preclinical studies have been performed where great promise of Azacitidine treatment of HNSCC was shown through the reversal of chemoresistance and the induction of apoptosis [226]. However, genome-wide methylation might also induce harmful side effects such as the demethylation of epigenetically silenced oncogenes of metastasis promoting gene [174] such as we have shown is the case for WISP1 (chapter 3).

Additional treatment options for epigenetically silenced tumor suppressor genes such as S100A9 and RAB25 might be provided through advancements in the clinical application of CRISPR-Cas9. The Crispr-Cas9 complex is a RNA-guided DNA endonuclease that originated from bacterial immune systems (reviewed by [368]). The guide RNA is 20 nucleotides long and can be replaced with any other 20 nt long sequence to target any desired DNA sequence. This has led to a great interest for the CRISPR-Cas9 complex as a genomic editing tool. A recent study has shown that the CRISPR-Cas9 inactive Cas9 endonuclease domain can be fused with a DNMT3A functional domain to specifically methylate a the guide RNA's target sequence[442]. Additionally, by fusing the RNA guided enzymatic system with the catalytic domain of the demethylation enzyme TET1, the CRISPR-Cas9 system has been used to unmethylated the targeted DNA [369]. These two targeted epigenetic editing tools might provide future therapeutic options to reverse pathogenic methylation levels of the hypomethylated WISP1 as well as the hypermethylated RAB25 and S100A9.

Finally, since all three biomarkers share a relation with calcium induced metastatic potential, the aberrant calcium-dependent pathways in tumor cells could also provide a therapeutic target for treatment of pN+ OSCC. For example, carboxyamido-triazole (CAI), a drug that alters Ca2+ concentration through inhibition of receptor-mediated Ca2+ influx and was shown to affect malignant proliferation and metastasis [362]. The relevance of these new treatment strategies however needs additional clinical validation as well as biological characterization.
CHAPTER 7

Reflection on choosing MethylCap-Seq combined with Methylation Specific PCR

When constructing our experimental designs, we choose to use MethylCap-Seq to establish the OSCC-specific methylome. Our reasoning for choosing MethylCap-Seq was its preference of CG-rich hypermethylated sequences which would focus on promoter regions which would lend itself especially well for identifying epigenetically silenced genes. However, this traditional view on DNA methylation induced gene silencing has been contested after those initial experiments due to the identification of regions that are less CpG-rich than CpG islands but heavily impact gene expression such as CpG island shores and CpG shelves [193], [246]. In hind-sight, an algorithm that would better include these regulatory regions would give a more complete view of the OSCC methylome. Unfortunately, we introduced a bias towards CpG Islands by only selection MC associated with transcription start sites. Additionally, the bias of MBD2, the enrichment protein used in MethylCap-Seq, only binds fragments with a high amount of methylated CpG site. Possibly, the bias eliminated the less CpG-rich CpG shores and CpG shelves. However, these insights were not available at the time of the MethylCap-Seq selection. Moreover, MethylCap-Seq was the best choice at the time that was less constrained to certain regions such as the Illumina 27k platform that was available at the time. Other available options of profiling techniques were also not more suited at the time. While frozen sample DNA used for MethylCap-Seq, the source of the DNA of the same patients used for the validation originated from FFPE. DNA extracted from FFPE is associated with a relatively lower DNA quality, the required high amount and high-quality DNA that is required for e.g. Whole-Genome Shotgun Bisulfite Sequencing, WGSBS was not suited for our validation OSCC cohort.

The other DNA methylation discovery tools available at the time of study design like MeDIP and Reduced representation bisulfite sequencing were comparable to the MethylCap-Seq. The deciding factors for MethylCap-Seq included the bias of MethylCap-Seq for DNA fragments with a certain threshold of number of CpG sites and distance between CpG sites compared to MeDip and RRBS. Additionally, MethylCap-Seq was chosen for the compatibility of this data with the data available in the Map of the Human Methylome. This, in hindsight, was probably one of the essential pros of MethylCap-Seq for the outcome of this thesis. Inclusion of validation of biomarker selection with the data of the Map of the Human Methylome phase proven to be essential of both the LN metastasis study as well as the saliva study as described in Chapter 6.

Future prospective

The extensive use of available public databases of genome-wide OSCC methylation data provided in both the TCGA and GEO databases to validate and further refine our long list of potential DNA methylation biomarkers predictive for the presence of LN metastasis in OSCC, was an essential part to select a restricted number of candidate methylation markers. Validation of most of the very first selected methylation markers from our own MethylCap-Seq data with other methylation detection techniques failed on our larger OSCC validation cohort. The sample size of six pN0 and six pN+ OSCC used for our discovery experiment proved too small for error-free biomarker identification often leading to false positives and false negatives. By using publicly available 450k data of pN0 and pN+ OSCC provided by the TCGA studies as a first validation set, we could save a lot of time and experimental testing by
removing any underperforming methylation markers from our prior selections. Additionally, the TCGA 450k data originates from an array of different centers and patient populations, assisting in decreasing any bias introduced by using a single center patient cohort. Overall, the public access to such vast DNA methylation data greatly assisted us in identifying more promising OSCC N-status methylation markers. Not only did the TCGA portal provide methylation data but also mRNA and mutation data to further elucidate the biology behind metastatic OSCC. We also found that methylation is not the only mechanism underlying the deregulation of the gene of interest. Therefore, endeavors like the TCGA and The International Cancer Genome Consortium (ICGC) are the way forward in biomarker studies. The TCGA and ICGC projects give smaller labs access to greater patient populations and datasets then they could muster by themselves. Additionally, the amount of data and research questions that could be answered with these data is to vast to be done by a single institute. For example, the different labs working on the first TCGA papers concerning OSCC did not investigate a potential pN predictive marker at all while all the data for identifying these pN status predictive marker were present [443].

The GEO datasets provided invaluable updated annotation of the 450k probe locations. Not only do these open platforms provide lots of data for other labs to use, it also supports open and verifiable research by making all data publicly available as well as providing a platform for comparable analyses. However, there is still some progress to be made in the application and accessibility of these platforms. Reanalysis of these data still requires above-novice bioinformatic skills to which most labs do not have access. GEO does a great job in providing some R analysis on their GEO cloud with GEO2R but this only allows analysis of a single GEO dataset and not on multiple similar GEO datasets (https://www.ncbi.nlm.nih.gov/geo/geo2r/). Comparably, TCGA does provide some simple analysis tools on their cBioportal website but for larger more complex datasets such as the 450k data is not possible with the provided tools. Essential steps such as data normalization, especially when dealing with multiple 450k experiments from different centers, are not provided. Also, while a great amount of clinical data is available for the TCGA patients some clinical parameters such as patient survival is incomplete or inconsistently reported between different participating centers and treatment-response data are generally lacking or incomplete. For GEO datasets the extensive individual clinical data is generally lacking where patient group designation is the only parameter provided.

Nevertheless, the possibilities with these open datasets are endless. I might even foresee some labs completely dedicating themselves to answering research questions using only publicly available data. Some major insights have already been achieved with this approach. For example, the before-mentioned CpG Island canyons have been identified by analyzing publicly available 450k datasets. I also believe that extensive mining of these datasets will provide many more insights for which these datasets have not been used yet. The list of all TCGA project publications can be found online (https://tcga-data.nci.nih.gov/docs/publications/) and only a single paper is present for the characterizations of HNSCC but that paper does not compare pNO and pN+ OSCC but compares all HNSCC samples together as on group [443]. Additional examples of reusage of the same TCGA data, in our lab we have performed a pilot study for the detection of local-recurrences by analyzing DNA methylation patterns in saliva of patient undergoing OSCC follow-up (Chapter 6). For the identification of OSCC recurrence markers
we have used the same MethylCap-Seq and 450k but instead of looking for differentially methylated regions between the pN0 and pN+ we have compared all OSCC with normal oral epithelial cells to identify universal hypermethylation markers in OSCC. Using the vast majority of the same datasets as used for the methylation N-status markers, we have solved another research question.

This thesis has created blueprints for DNA methylation biomarker identification experimental designs. We have laid the foundations for other studies to use when trying to identify biologically relevant biomarkers. The steps taken and described in this thesis have set an example for other groups how to incorporate the data available in both the Gene Expression Omnibus and The Cancer Genome Atlas. We have shown how to use these databases to both validate the predictive value of the selected DNA methylation markers and the biological validity of these markers. These steps can now be applied to different cancers and research questions. These steps will both reduce the costs for biomarker discovery studies as well as increase the odds of finding a biomarker with a high predictive value.

In our lab we have applied this same approach to identifying DNA methylation biomarkers to predict radiotherapy response in Oropharyngeal and laryngeal tumors (Clausen et al. in preparation). For this study another MethylCap-Seq was performed on three radioresistant and three radiosensitive HNSCC cell lines to identify DNA methylation markers that predict the tumors sensitivity to radiotherapy. It must be noted that employing TCGA data to this study is more difficult because the radiotherapy data in the TCGA database is less well defined and doesn’t allow for precise determination of the tumors radiotherapy sensitivity.

As we have shown in this thesis, no expected universal predictive marker is available for predicting the nodal status of OSCC. It is therefore expected that a gene panel consisting of various DNA methylation biomarkers as well as other biomarkers, is a more feasible approach for clinical application. The size of this panel is yet unclear. In the literature a well-defined mRNA expression profile is described and this study consists of 696 genes to achieve a negative predictive value of 89% for determining nodal spread. The size of the panel greatly determines the DNA methylation detection technique that can be used for diagnostics. When assuming a gene panel of 228 genes a multiplexing on for example on the Ion Torrent S5 could be employed or upcoming techniques such as digitalMLPA with allows multiplexing of up 1500 regions of interest would be greatly suited for translating a gene panel to the clinic [444]. Using either of these NGS based methods would be much more expensive than for example MSP or Q-MSP, because MSP is most likely not quantitative enough to be applicable to the OSCC methylome that seems highly variable and requires strict cut-offs. This might make such a large desirable gene panel unsuited for large spread clinical application.

By developing a DNA methylation marker panel that for example focusses on improving on the major downsides of SLNB by providing a non-invasive and high positive predictive panel might greatly contribute to a clinical modality consisting of both SLNB and a methylation biomarker. Potentially, in the future the TCGA 450k data could be used to construct a DNA methylation gene signature predictive for LN-metastasis.