Informing the Oral Squamous Cell Carcinoma Biomarker Search by Exudate Proteomics

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Joel Allan Kooren

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Timothy Griffin

April 2013

Acknowledgements

Pratik Jagtap played an integral role in aiding with mass spectrometry data analysis
Susan Van Riper created software (RIPPER) that was instrumental in determining protein candidates from Orbitrap data to the greatest extent possible

Jill Goslinga aided in finding the relevant protein level insights from peptide level analysis

Nelson Rhodus and Patricia Fernandez provided me with access to clinical samples for my research

Chuanning Tang provided useful data through her proteomic analysis of brush biopsy samples

Ebbing de Jong, Matthew D Stone, Sricharan Bandhakavi, and Mikel Roe all provided valuable advice and sometimes collaboration as fellow laboratory members during my thesis work

Also, thanks to the Center for Mass Spectrometry and Proteomics at the University of Minnesota for invaluable work on and providing access to the Orbitrap XL and Velos, as well as sample fractionation.

Thesis Abstract

Oral cancer is the sixth most common cancer worldwide ahead of Hodgkin's lymphoma, leukemia, brain, stomach, or ovarian cancers, with about 41,000 Americans being diagnosed annually. More than 90% of oral cancers are oral Squamous cell carcinomas (OSCC). While the overall 5-year survival rate is about 60%, the survival rate when diagnosed early is higher than 80%. Currently the standard for diagnosis of OSCC is early visual detection of a suspicious oral lesion followed by scalpel biopsy with histology. However, the invasiveness, expense, and required expertise involved prevents consistent application on at risk individuals. Chapter 1 discusses the methods that are being investigated for sampling and discovering biomarkers of OSCC that address some of these limitations. Protein biomarkers contained in samples collected non-invasively and directly from at-risk oral premalignant lesions (OPML) would address current needs in a uniquely targeted fashion.

Chapter 2 of this thesis describes work evaluating the potential of a novel method using commercial PerioPaper absorbent strips for the collection of oral lesion exudate fluid coupled with mass spectrometry based proteomics for OSCC biomarker discovery. This research focuses on demonstrating the feasibility of using oral lesion exudates in proteomic research, exploring the proteome of exudate samples, discriminating between exudates collected from clinically different sources, with supplemental table 1 showing which proteins distinguish healthy and OPML sources. Furthermore, to ensure that the best possible marker candidates are selected given clinical sample availability, multiple methods were explored enable and improve quantitative proteomic analysis of exudates in chapter 3 (Identified proteins in supplemental files 2 and 3). Our label-free quantitative proteomics strategy analyzed paired control and OPML exudates (figure 8), identifying differentially abundant proteins between sample types. Next, we selected several [exudate] differentially abundant proteins for testing in while saliva, comparing their relative abundance levels in healthy, OPML and oral Squamous cell carcinoma (OSCC) subjects. Two proteins, CK10 and A1AT, showed differences in saliva. Our results provide a demonstration of the value of tissue exudate analysis for guiding

salivary biomarker discovery in oral cancer, as well as providing promising biomarker candidates for future evaluation.

Dedication

My father, mother, and sisters

Your love and support has been central to my life. This dissertation represents years of work, thought and occasional struggle. Your patience and understanding have made this possible.

My friend, confidant, honorary family member, and early career scientific mentor Adi

Thank you for being the brother I never had, and for introducing me to the wonder of examining life at its most basic level.

My thesis advisor Timothy

These advancements required the effort and skill of many people other than myself. Most centrally you have been a guide, an example, and a critic to keep my research on the right path.

To all my mentors and advisors over the years

Your guidance, encouragement, and advice have inspired and pushed me to new discoveries.

Table of Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

oral squamous cell carcinoma (OSCC)

oral pre-malignant lesion (OPML)

mass spectrometry (MS)

human papillomavirus (HPV)

micro RNA (miRNA)

nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB)

matrix-assisted laser desorption ionization (MALDI)

tandem mass spectrometry (MS/MS)

liquid chromatography (LC)

nano-scale reversed-phase liquid chromatography (nanoLC)

mass to charge ratio (m/z)

peptide sequence match (PSM)

intensity-based label-free quantification (ILFQ)

area under the curve (AUC)

surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF)

strong cation exchange (SCX)

high-performance liquid chromatography (HPLC)

sodium dodecyl sulfate (SDS)

SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

parts per million (ppm) [used in terms of mass accuracy of mass spectrometers]

dithiothreitol (DTT)

stable isotope labeling of amino acids in cell culture (SILAC) [quantitative MS method]

false discovery rate (FDR)

electrospray ionization-ion trap type mass spectrometer (ESI-TRAP)

human oral microbial database (HOMD)

proximity based intensity normalization (PIN)

horseradish peroxidase (HRP)

Total Ion Current (TIC)

alpha-1-antitrypsin (A1AT)

selected reaction monitoring (SRM) [a targeted MS method]

Chapter 1: Introduction to Thesis

This Chapter is adapted from published work (Kooren, 2011, Emerging Non-Invasively Collected Genomic and Proteomic Biomarkers for the Early Diagnosis of Oral Squamous Cell Ccarcinoma (OSCC)

Adapted from: Nova Science Publishers inc., Squamous Cell Carcinoma, Chapter 10 Emerging Non-Invasively Collected Genomic and Proteomic Biomarkers for the Early Diagnosis of Oral Squamous Cell Ccarcinoma (OSCC), pp 197-210, 2011 Joel A. Kooren, Nelson L. Rhodus, Timothy J. Griffin. With the permission of Nova Science Publishers inc.
ISBN: 978-1-61209-929-3

As mentioned earlier, oral cancer and specifically delayed diagnosis of OSCC is a major public health concern that is currently dealt with primarily by biopsy based histology. However, there are multiple limitations associated with biopsies: being invasive clinicians are hesitant to perform them, and patients may not agree to them due to the pain and discomfort of the procedure; the subsequent histology requires expert analysis and is therefore expensive; and issues such as under-sampling add uncertainty to diagnosis. An ideal alternative to scalpel biopsy would be non-invasively collected samples containing biomarkers which can distinguish between oral pre-malignant lesions (OPMLs) and OSCC, and potentially predict the transition from pre-malignancy to malignancy. Methods for sampling and discovering biomarkers of OSCC in a non-invasive fashion have been emerging, including those focusing on whole saliva and cells and other specimens collected directly from oral lesions. These samples are ideally suited for system-wide analysis using genomic and proteomic technologies for biomarker discovery. In this introductory chapter, the current state of the clinical diagnosis of oral cancer is described, with an emphasis on emerging genomic and proteomic strategies seeking to identify non-invasively collected biomarkers that could improve the early diagnosis of OPML transition to OSCC.

Etiology, Progression and Diagnosis of Oral Cancer:

Etiology and Progression of OSCC

The genesis of OSCC involves genetic, epigenetic and metabolic changes usually due to insult and injury brought on by carcinogens such as tobacco use (Sudbo, 2004; Lippman et al., 2005), and/or viral infections such as HPV (Marur et al., 2010). Generally, the causative insult or injury generates an oral pre-malignant lesion (OPML) which can be a small point occurrence or a field covering a large area in the oral cavity. For leukoplakia, the primary type of OPML, the risk of malignant transformation is between 5% and 17% (Silverman et al., 1984; Silverman, 2001; Rhodus, 2005) with some subsequent studies reporting more variation in these estimated risks, and a higher risk with other types of less common OPML such a erythroplakia (Greenspan & Jordan, 2004). Since most OSCC is preceded by OPML, (see figure 1), a primary need in the clinic is to reliably diagnose a transition to malignancy in its earliest stage (Rhodus, 2005; ACS, 2013).

Current Gold Standard Method for Oral Cancer Detection: Incisional Biopsy

Unfortunately, even a conventional oral examination by a highly trained dental practitioner cannot determine if an OPML has transitioned to cancer, or is likely to do so. The current gold standard for OSCC diagnosis is invasive incisional biopsy followed by histopathology (Lingen et al., 2008). Despite being the standard, this method for OSCC diagnosis has a number of limitations. A main limitation to the scalpel biopsy with histopathology is lack of frequent testing of suspicious lesions. In many cases diagnosis of OSCC is delayed, despite patients visiting clinicians, because the clinicians did not biopsy an abnormality they previously identified as OPML (Lumerman et al., 1995; Axell et al., 1996). In fact, a retrospective study found that of those dentists who had diagnosed a patient with OPML, only 14.1% followed the lesion up with a second biopsy within a 3

year period (Rhodus, 2005). Patients are often resistant to invasive procedures such as a scalpel biopsy, particularly for multiple subsequent biopsies which are often necessary for monitoring the progression of OPML. The fact that histological examination is necessary increases expense and effort required to perform a proper scalpel biopsy analysis of a suspicious lesion (OPML or OSCC) further decreasing the frequency that biopsies are performed. Even when a biopsy is performed, lesions which spread over a large area are often require multiple site biopsies which may be taken in a specific location(s) which is not truly representative of the status of the lesion (Rhodus, 2005). The very fact that oral cancer currently only has a 50% 5-year survival rate, when it could be as high as 90% if detected in the earliest stage, is a testimony to the inadequacy of the current standard of practice (Rhodus, 2009).

Rather than histopathology, numerous studies have been undertaken on tissue collected via incisional biopsy seeking to identify molecular biomarkers that could be useful for diagnosis of oral cancer. Some work has focused on detecting chromosomal abnormalities (Reshmi & Gollin, 2005). Beyond diagnosis, many studies focusing on chromosomal abnormalities have investigated their utility in cancer prognosis, and have explored sample collection strategies that are at least minimally invasive (Sato et al., 2010). Additionally, there have been numerous biomarker studies in biopsied tissue using genomic and proteomic approaches (Viet & Schmidt, 2010). Although many of these studies on biopsied tissue have provided new insights into OSCC-associated biomolecules, as clinical tests for oral cancer they all suffer from their dependence on the invasive incisional biopsy, with its inherent limitations as described above.

Figure 1

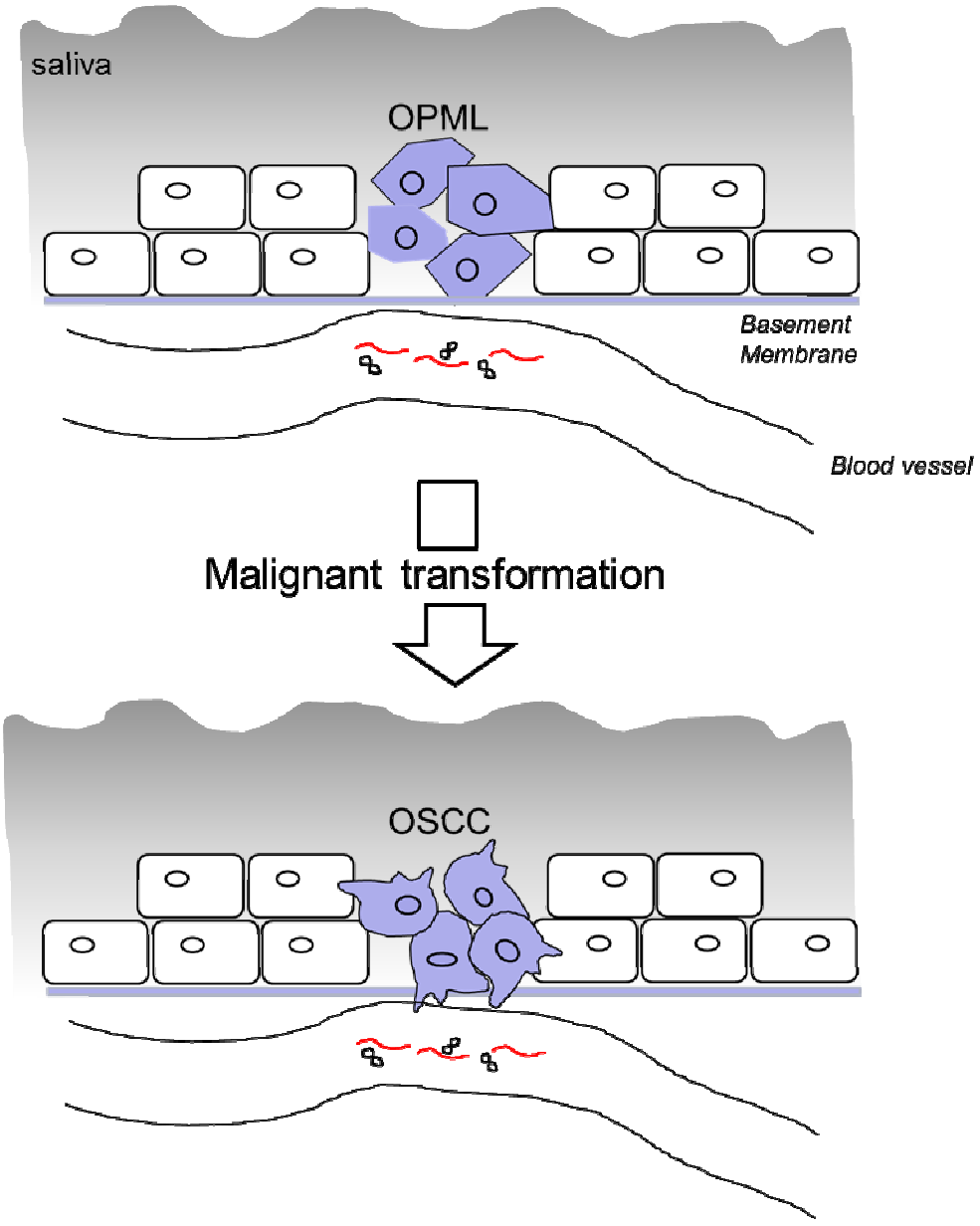

saliva
OPML
Basement
Membrane
Blood vessel
Malignant transformation
OSCC

Figure 1 legend: Oral Pre-Malignant Lesion to Oral Squamous Cell Carcinoma clinical transition

Healthy oral mucosa at some point may become an oral pre-malignant lesion (OPML), due to insult or injury (e.g. tobacco use). Though the OPML is in an inflamed, pre-malignant state, the abnormal OPML cells do not compromise the basement membrane. The resulting OPML may at some point transform into oral Squamous cell carcinoma (OSCC) by threatening the integrity of the basement membrane.

Current State of Non-Invasive Testing of OSCC: Screening Versus Case-Finding

The limitations of invasive incisional biopsy have motivated the development of non-invasive methods for diagnosing oral cancer. At the outset, it is worth noting that in the area of non-invasive testing methods for oral cancer, there are two distinct types of testing, each with different goals. One type seeks to detect the presence of a suspicious lesion within the oral cavity; the other type seeks to diagnose an already detected oral lesion as pre-malignant or malignant. A review (Lingen et al., 2008) has defined these two types of methods as either screening (detection of an oral lesion) or case-finding (diagnosis of an oral lesion). Both of these methods seek to maximize the specificity (the ability to distinguish individuals with a condition from those without) and sensitivity (the ability to reliably detect a condition in subjects who truly have it) to distinguish between healthy individuals (no lesion) and those with OPML or OSCC. The focus of this overview will be mainly on non-invasive tests for the case-finding of oral cancer which we will refer to as diagnosis, with only a brief description below of the current state of non-invasive screening tests.

A number of methods exist for non-invasive screening for suspicious oral lesions. These tests seek to detect suspicious oral lesions within a population of asymptomatic individuals. These screening methods have been well-described in a review by (Lingen et al., 2008), summarized briefly here. The most commonly used test is the conventional oral examination, in which a clinician searches the oral cavity for visible irregularities in the oral epithelium. However, there are limitations to this approach. It cannot determine whether lesions are benign or on a path towards potential malignancy, and it might simply miss small lesions in an early stage of development. Therefore other methods have been developed to improve upon the conventional oral examination. Toluidine Blue (also known as tolonium chloride) is a dye used to stain abnormal tissues, sometimes including OSCC, which would not otherwise be diagnosed by upon visual inspection. Though sensitivity is high (~100%) for OSCC, specificity is low (~60% varying between studies). Another way of enhancing a visual inspection is to use an Aceto-white method. Aceto-white methods of screening use a 1% acetic acid solution rinse followed by a

visual inspection of the oral cavity under blue and white light. Two commercially available types of aceto-white screening methods are ViziLite Plus, and MicroLux DX. Like Toluidine blue, aceto-white methods are sometimes capable of detecting OSCC that might not be readily visible (high sensitivity ~100%), but also makes many innocuous lesions visible and thereby decreases the specificity of the test. Tissue fluorescence is used in a variety of formats for screening, and case-finding in both diagnostic and prognostic fashions. A major commercially available screening tool using fluorescence-based detection is the VELscope. Studies have reported high sensitivity and specificity, though detection of lesions that would not be apparent with a conventional oral examination is quite rare. VELscope does have utility in detecting dysplastic tissue (OPML) that may extend beyond the location of a visible lesion (Lingen et al., 2008).

Emerging Genomic and Proteomic Biomarkers in Non-Invasively Collected Samples for Early Diagnosis and Monitoring

Nucleic acid based Studies in Whole Saliva

The availability of non-invasive, inexpensive, and accurate tests would foster more frequent testing and improve the early diagnosis of oral cancer. The most prominently used non-invasively collected sample in OSCC biomarker research is whole saliva. This is because saliva is highly available and its composition is influenced by many diseases, drug treatments, and general conditions relevant to health. These advantages have already been used to develop salivary tests for other applications including HIV detection and drug monitoring . This is largely due to several readily apparent advantages (Rhodus, 2005). In addition to being easy to collect in a non-invasive fashion, saliva has direct contact with oral lesions, and can be collected in moderately large volumes (> 1 milliliter) in an on demand manner. Given these advantages, a number of whole saliva-based tests already exist, prominently for HIV detection and also drug monitoring (Kaufman & Lamster, 2002).

Finding new molecular biomarkers of OSCC in saliva is ideally suited for a "discovery-science" approach (Aebersold et al., 2000) using technologies capable of identifying cancer-dependent molecular changes on a system-wide level. For analyses in whole saliva seeking to identify diagnostic biomarkers of OSCC a number of system-wide analysis approaches have been applied. For this thesis, these are categorized as either genomics approaches, analyzing DNA sequence or RNA expression, or as proteomic approaches, analyzing the protein complement.

In recent years, genomic-based tests for OSCC biomarker discovery in whole saliva have focused on DNA methylation, mRNA expression analysis, and analysis of microRNA (miRNA) targets, which are seeing increased research interest (Viet & Schmidt, 2010). Although most biomarker studies using genomics have focused on invasively collected biopsy samples, there are some examples in non-invasively collected whole saliva. In one recent study, (Viet et al., 2007) demonstrated that DNA methylation associated with OSCC can be detected and analyzed in saliva with results at least comparable to that of tissue. A study by (Carvalho et al., 2008) compared the oral rinses from patients with HNSCC to healthy controls. Using a technique to detect hypermethylation, the authors identified one panel of altered promoters which gave high specificity with low sensitivity (90% and 35%, respectively), and another panel of promoters which achieved high sensitivity but with poor specificity (85% and 30%, respectively). Regrettably, no group of biomarkers could be identified which offered both high sensitivity and specificity. However, if biologically distinct paths to OSCC exist as many expect (Schaaij-Visser et al., 2010), then it may be necessary to use a combination of biomarker tests that achieve high specificity by correctly tracking a route to OSCC, while having individually low sensitivities. Currently, the limited amount of mechanistic insights into OSCC development makes such an approach prohibitively difficult.

At the level of mRNA expression, (Zimmermann et al., 2007) used cDNA arrays to quantify mRNA transcripts in saliva from healthy and OSCC patients. They achieved 91% specificity and 91% sensitivity with their panel of four mRNAs indicative of OSCC. In addition, they also tested RNA stabilization reagents to generate a protocol that should allow salivary sample collection in situations where immediately freezing samples might

not be feasible. More recently, the same group investigated the presence of miRNA in whole saliva, and their potential as diagnostic markers for OSCC (Park et al., 2009). They found over 300 miRNAs in whole saliva, with two of these showing a statistically significant drop in abundance in patients with OSCC compared to healthy controls.

Mass Spectrometry-based Proteomics and Protein-based studies in whole saliva

Despite revealing numerous discoveries of promising OSCC biomarkers in whole saliva, the genomic approaches above do not capture post-transcriptional molecular events that may be associated with OSCC development. These molecular events are best captured by analysis of proteins. As biomarkers, proteins have numerous advantages, given that they are the biological effector molecules that may be directly involved in the mechanisms of OSCC, and they are amenable to antibody-based clinical assay development once validated. Additionally, beyond their use as biomarkers, proteins also provide potential targets for therapies aimed at treating OPML or OSCC.

A number of studies have demonstrated the potential of proteins in whole saliva as biomarkers of OSCC. Examples include the interleukin proteins 1, 6 and 8 whose levels are all affected by NF-κB signaling (Rhodus et al., 2004; Rhodus et al., 2005). As a proof-of-principle for point-of-care-clinical assay development, an optical protein sensor was developed targeting interleukin 8 in whole saliva (Tan et al., 2008). Although encouraging, these candidate-based studies only capture a very small number of proteins that may be useful for oral cancer diagnosis, and may miss those that are of most value. A more system-wide analysis of proteins via a proteomics approach can provide such an analysis.

For proteomic studies seeking to discover new protein biomarkers in whole saliva, mass spectrometry (MS) is the analytical platform of choice. The technologies that have made protein mass spectrometry a valid option start with methods to ionize proteins and/or peptides so that they can be introduced into a mass spectrometer. These approaches, developed in the 1980s are Matrix-Assisted Laser Desorption Ionization (MALDI) (Tanaka et al., 1988) and Electrospray Ionization (ESI) (Fenn et al., 1989).

There are two main approaches to MS-based proteomics, termed “top down” and “bottom up”; in “top down” mass analysis is typically performed directly on intact proteins, whereas in “bottom up” methods proteins are typically digested to peptides, which are analyzed by MS. The top-down approach has many advantages for identifying protein isoforms and post-translational modifications with confidence, as well as estimating or quantifying their stoichiometry of different isoforms, (Stastna & Van Eyk, 2012).

The bottom-up approach, despite some limitations, has proven more amenable to scaling up to approach a system-wide view. This is also referred to as the “shotgun” approach, and is done by piecing together information on peptide sequences to infer the presence of proteins (Chen & Yates, 2007) illustrated in Figure 2. Here, proteins are digested to peptides, followed by preparative and analytical scale separations, tandem mass spectrometry (MS/MS) analysis and sequence database searching. Thus, the bottom-up approach actually integrates several key technologies.

Figure 2

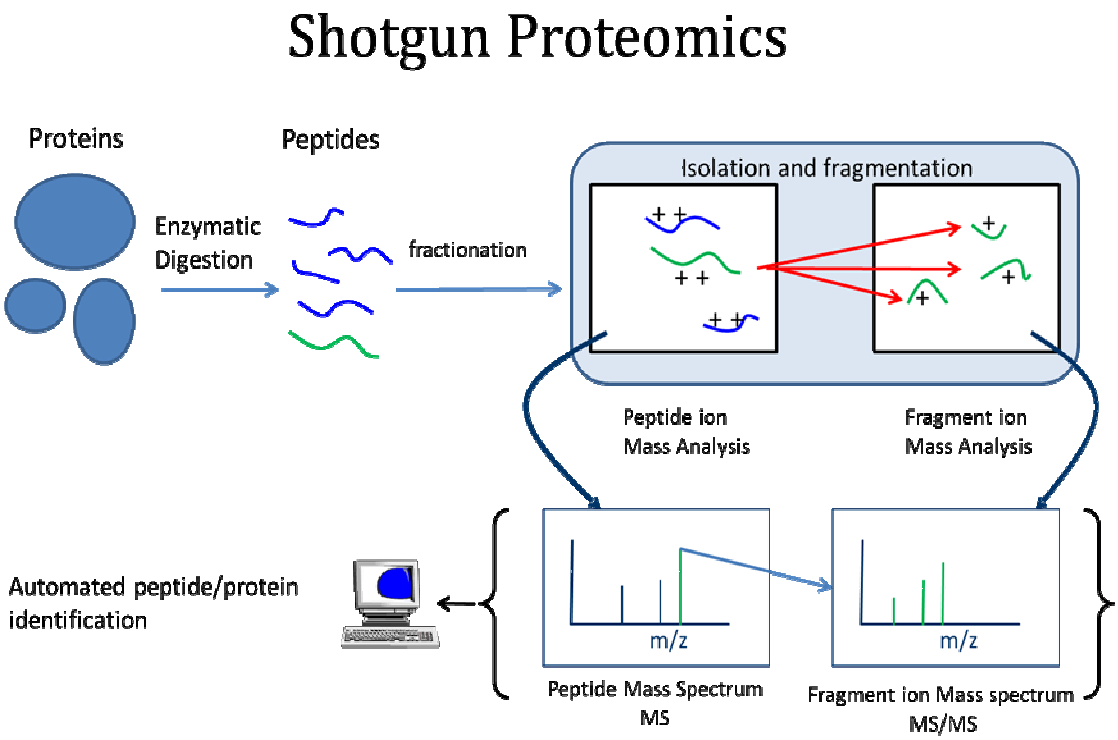

Shotgun Proteomics
Proteins
Peptides
Enzymatic
Digestion
fractionation
Isolation and fragmentation
Peptide ion
Mass Analysis
Fragment ion
Mass Analysis
Automated peptide/protein
identification
m/z
m/z
Peptide Mass Spectrum
MS
Fragment ion Mass spectrum
MS/MS

Figure 2 legend: Shotgun Proteomics

In Shotgun proteomics proteins are digested to peptides prior to analysis. MS analysis is then performed on peptide ions determining their mass to charge (m/z) ratios. Generally fragmentation of the peptide ion, with subsequent MS analysis of the resulting fragments (MS/MS), allows the determination of sequence information. The combination of peptide ion data (primary MS) and fragment ion data (MS/MS) can be used to identify the peptide ion.

One key technology is nano-scale reversed-phase liquid chromatography (nanoLC) with MS detection (Deterding et al., 1991). NanoLC works by eluting a peptide mixture off of a chromatographic column (at very small scale) into a mass spectromer. In this way, a complex peptide mixture can be introduced to MS analysis over time, such that a less complex portion is analyzed at any given moment, reducing signal suppression effects. The miniaturization to a nanoscale results in more efficient electrospray ionization because a small sample can be concentrated (into a sub picoliter volume) and achieve high chromatographic resolution, increasing sensitivity. While coupling nanoLC to electrospray ionization MS/MS does allow a simpler sample to be introduced into a mass spectrometer, the sample complexity and dynamic range may still be a limiting factor for many sample types, including those discussed in subsequent chapters (e.g. human saliva, plasma, tissue samples). In these cases, multidimensional fractionation methods can be applied upstream of MS analysis to further simplify complex mixtures. Classically, pre-LC fractionation often involved 2D gel electrophoresis, followed by in-gel digestion (Hu et al., 2008). However, due to several inherent limitations of 2D gels, gel-free fractionation of peptide mixtures arose using LC-based methods (Link et al., 1999; Washburn et al., 2001; Gygi et al., 2002). These methods have been applied to good effect on clinical samples including saliva (Xie et al., 2005; Guo et al., 2006).

Another key technology for bottom-up proteomics is MS/MS analysis. Here, peptide ions are first recorded at distinct mass-to-charge (m/z) values as they elute from the LC column, and each detected peptide is isolated and fragmented in the instrument, collecting a mass spectrum (also known as MS/MS spectrum) of peptide fragments' m/z values. These fragmentation spectra eventually enable sequence information to be determined for selected peptide ions (Hunt et al., 1986). This is generally achieved by automated matching of peptide sequences to MS/MS spectra, generating what is termed a peptide sequence match (PSM). This is generally achieved using a computer program, such as SEQUEST (Eng et al., 1994) or Mascot (Perkins et al., 1999). These are two of the earliest programs, numerous other choices are now available. These programs search large databases of known peptide sequences, generating their expected MS/MS spectra, and match these to the MS/MS spectra obtained. If the match in the database passes a

threshold for quality, the PSM is then considered for that MS/MS spectrum. Those deemed to have high confidence are then aggregated by the protein sequences from which they are derived, and protein identities in the starting mixture are inferred. Generated PSMs are often subjected to methods such as reversed database searching (Peng et al., 2003), which is used to estimate the false positive rate for the set of acquired PSMs that satisfy the previously set scoring criteria threshold. The reversed database is a set of theoretical peptides generated by reversing the sequences of known peptides in a traditional database. Because the reverse database contains an equal quantity of false theoretical peptides to the potential true matches in the traditional [forward] database, the number of PSMs matched to the reverse database serves as an estimate of how many false matches are likely to exist among forward database PSMs.

Proteins are dynamic, with their abundances regulated by synthesis and degradation. Because of this, the ability to quantify, even in only a relative manner, rather than just catalogue their presence can be essential to finding the proteomic differences corresponding to a condition of interest. Within a mass spectrometer, ions that differ by chemical identity, due to their amino acid sequence in the case of peptides, and other properties such as charge-state vary in their detected signal intensities. Because these differences are difficult or impossible to predict and/or model, signal detected in the mass spectrometer can't be directly converted to an absolute abundance (Gygi et al., 1999). However, signals compared between samples (e.g. different patient's saliva, different treatment conditions in a cell culture, etc) can be quantified in a relative manner.

One popular method for deriving relative quantitative information from MS-based proteomics experiments uses stable isotope labeling. These methods label proteins and/or peptides with stable but less common isotopes (15N, 13C, 18O) to create one or more labeled experimental condition, which can then be compared with another condition wherein proteins and/or peptides contain naturally occurring isotope composition (Becker, 2008; Gevaert et al., 2008). These methods involve either labeling through cell culture, in which some cells are grown in media containing heavy stable isotopes, (Oda et al., 1999; Ong et al., 2002) or chemical labeling of some form (Gygi et al., 1999; Reynolds & Fenselau, 2004; Ross et al., 2004; Dayon et al., 2008). Labeling through cell

culture is effective when applicable, but unsuitable for clinical samples (or other situations where cell culture isn't used), whereas chemical labeling is more versatile but requires more sample handling and exact protein or peptide quantification before labeling. Both approaches have limitations in terms of scaling when comparing many conditions against each other. In the case of cell culture methods, proteins extracted from 'heavy' (isotopically labeled) and 'light' (natural isotope abundance) cell cultures are combined so that total protein quantity is equivalent. After preparation (digestion, and likely fractionation) peptides common to both samples, though differentially isotopically labeled, retain the same chemical properties and behave similarly in fractionation and mass spectrometric detection. Thus the 'heavy' and 'light' forms of any given peptide are detected simultaneously, and relative abundance can be measured from the MS signal displaying 'heavy' and 'light' peptides at differential m/z. In the case of chemical labeling techniques, which can be used on clinical samples from humans, either proteins or peptides are chemically labeled. Again, accurate knowledge of starting quantities of proteins or peptides in separate samples to be compared is essential and potentially a limiting factor in accuracy. Additionally, partial labeling must be minimized, and for labels that allow multiplexed analysis via MS/MS fragmentation (e.g. isobaric peptide labeling (Ross et al., 2004)), co fragmentation of multiple peptides at one time is another documented limitation (Ting et al., 2011).

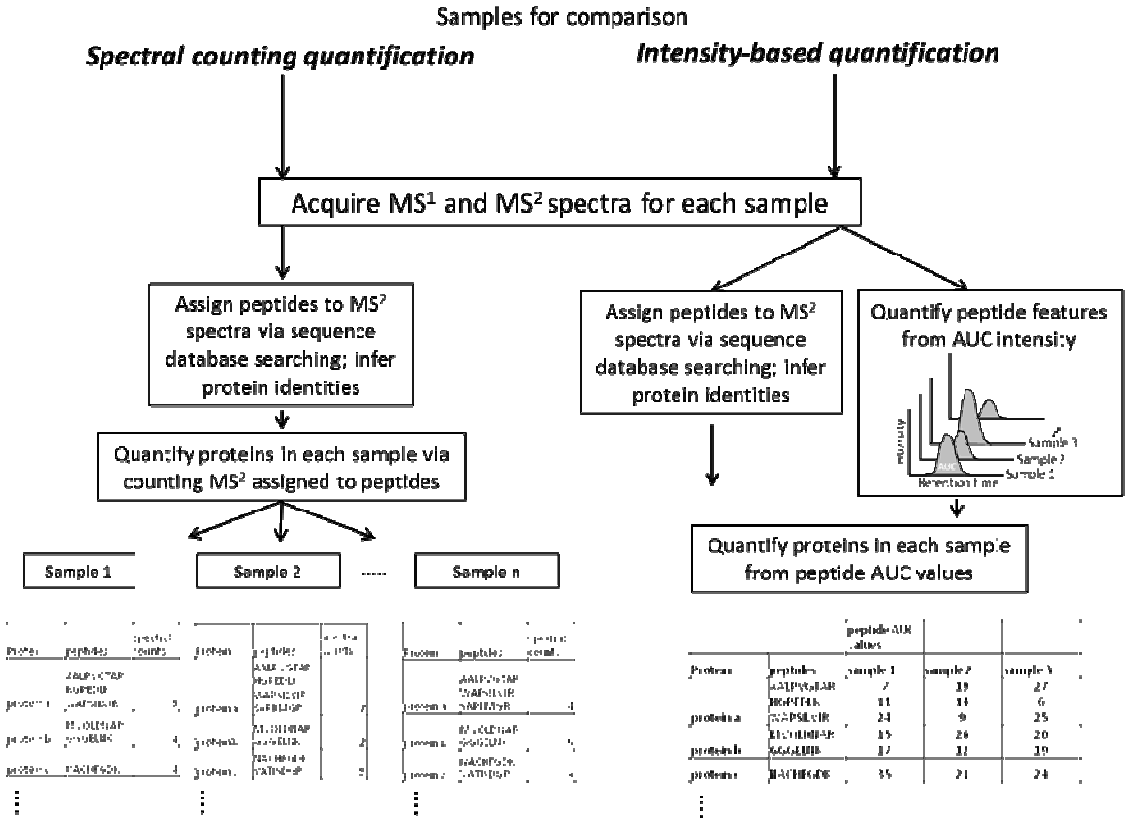
Label-free approaches offer another method for quantitative proteomics, without the requirement of stable isotope labeling. There are two primary methods that comprise label-free approaches: spectral counting and intensity based measurements. Figure 3 demonstrates key differences between the two methods. Spectral counting is directly dependent on the ability to acquire PSMs, using these as a proxy for quantity. The underlying idea of spectral counting is that the number of PSMs derived from a protein is proportional to its abundance in the starting mixture; that is, high abundance proteins produce many PSMs and lower abundance proteins produce fewer. Many simple spectral counting methodologies have often been employed where peptide identifications for each protein were summed. However, many more advanced techniques have been developed

that normalize for a number of factors, such as protein length, that may lead to spectral counting inaccuracies (Neilson et al., 2011).

Intensity based methods, also known as intensity-based label-free quantification (ILFQ), offer an alternative to spectral counting. ILFQ is based on the fact that during a LC-MS/MS analysis, the signal intensity of any peptide, having a distinct m/z, is recorded over the entire chromatographic peak. Using this information it is possible to reconstruct a chromatographic peak for each peptide based on the recorded signal intensities for its m/z value (Figure 3 b). An area under the curve (AUC) can be calculated and then compared to the AUC for the same peptide detected in additional samples run separately, providing information on relative quantity. Normalization to account for introduced errors, including unequal loading amounts between samples being compared, is critical and can be achieved via a variety of methods (Callister et al., 2006). As with spectral counting, peptide ions are matched to sequences via MS/MS and sequence database searching, but because this is separate from the quantification, ILFQ is not dependent on successful PSM. Thus ILFQ offers some benefits to overcoming situations where the number of PSMs does not reflect accurately the relative abundance of a protein. However, PSM information becomes critical if one desires to know the peptide sequence, and protein from which it is derived, for any measured AUC.

Figure 3 (a and b) Spectral counting vs. intensity based label-free quantification methods (ILFQ).

a



b

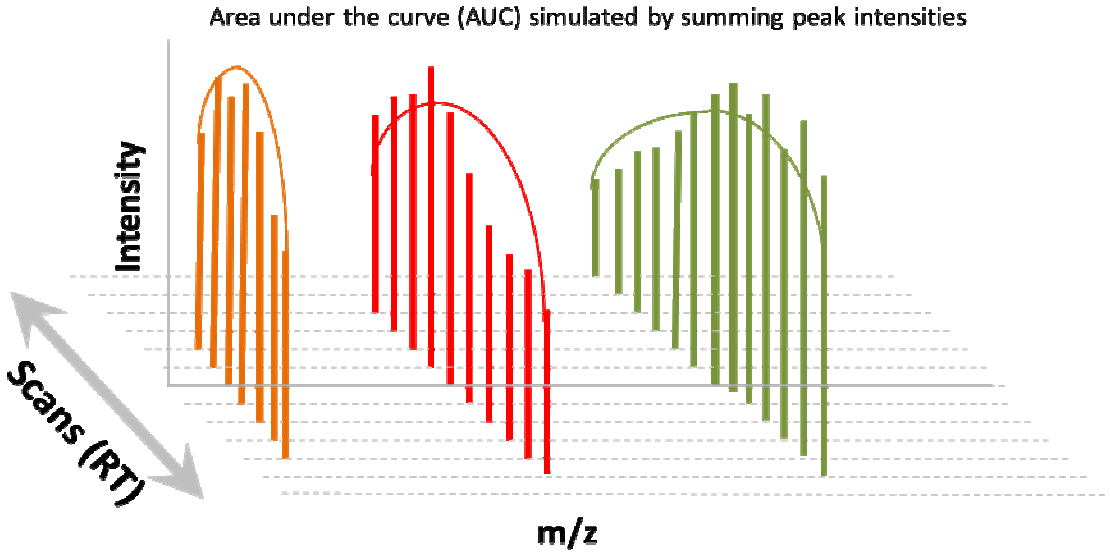


Figure 3 Legend: Spectral counting versus ILFQ

Spectral counting based label-free quantitation techniques rely entirely on peptide identifications for quantitative information. The intensity information in the MS primary scan is not used in spectral counting approaches. For intensity based label-free quantitation (ILFQ) the primary scan intensity information is measured, and the identifications are used to annotate the intensity measurements.

Mass spectrometry based proteomics studies

Several recent studies have utilized advanced MS-based proteomics approaches to discover diagnostic protein biomarkers of OSCC in whole saliva. In one notable study by (Hu et al., 2008), MS-based proteomic analysis was used to profile whole saliva proteins from 16 pooled healthy individuals and 16 pooled OSCC patients. Using a semi-quantitative approach based on the number of MS/MS spectra matched to proteins in the healthy or OSCC pooled samples, 23 proteins were classified as differentially expressed. After rigorous biochemical verification in individual patient samples, five proteins (M2BP, MRP14, CD59, catalase, and profilin) displayed statistically significant differential abundance between healthy and OSCC samples. Next they measured the abundance levels of these proteins in 48 healthy and 48 OSCC whole saliva samples; the combined panel of biomarkers provided a sensitivity of 0.90 and specificity of 0.83.

In another study, (Jou et al., 2010) identified the protein transferrin in whole saliva as a potential biomarker for OSCC diagnosis. Starting with 2D gel electrophoresis they identified differentially abundant proteins following up with MS and eventually ELISA verification of their selected marker, transferrin. They showed that increasing abundance of transferrin was associated with both severe and early stage OSCC. A receiver-operator curve was used to evaluate the sensitivity and specificity of transferrin as a marker. In such an analysis the area under the [receiver-operator] curve reflects the probability that a randomly chosen positive event (OSCC in this case) would have a higher measure (transferrin level) than a randomly chosen negative event (healthy control). The curve can be used to demonstrate what level of sensitivity is achieveable for a given specificity and vice-versa for a transferrin based test to classify individuals as OSCC or healthy. The constructed a receiver-operator curve had an area under the curve of 0.91+ when comparing (similar numbers of) healthy and OSCC samples.

Using the MS method of surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF), which focuses on the detection of relatively small proteins and peptides, (Shintani et al., 2010) analyzed whole saliva from healthy individuals and those with OSCC. Among a number of possibilities, the authors chose to

focus on one detected peptide which they determined to be a truncated form of salivary cystatin SA-I. However, its value as a biomarker is difficult to judge given the lack of reported values of sensitivity, and specificity as well as the lack of a study of Cystatin SA-I levels in a larger number of subjects using non-MS-based (e.g. immunoblotting or ELISA) verification.

One component lacking in all of the whole saliva biomarker studies above is the inclusion of samples collected from individuals with an oral pre-malignant lesion (OPML). Comparison of OPML patients with OSCC patients is highly valuable, as it would provide insights into which biomarkers may distinguish these lesion types, revealing those biomarkers that may have potential for diagnosing a transition to malignancy early. To this end (de Jong et al., 2010) used quantitative MS-based proteomics to compare proteins in whole saliva from OPML and OSCC patients. For the discovery phase of this study, pooled whole saliva from four OPML patients was compared to pooled whole saliva from four OSCC patients. Among almost 1000 identified proteins, about 200 showed differential abundance between the two patient groups. Based on bioinformatic analysis, the cytoskeletal proteins actin and myosin were selected for further verification. Independently, their known roles in cell motility and invasion for epithelial cancers support their use (Hall, 2009). Western blotting against actin and myosin was undertaken in 12 additional OPML and 12 additional OSCC whole saliva samples. For actin, a sensitivity of 100% and specificity of 75% was reported, whereas for myosin the values were 67% and 83% respectively. Further validation was performed on the exfoliated cells in whole saliva given the potential diagnostic utility of these cells (Xie et al., 2008). Within the exfoliated cells, actin and myosin showed similar differences in abundance between OPML and OSCC patients, linking the results observed in soluble whole saliva to concomitant changes within the epithelial cells.

Studies in Other Non-Invasively Collected Sample Types

Although saliva is an optimal medium for biomarkers in terms of accessibility, its complexity makes the initial detection of possible markers difficult. Identifying markers

from more direct sampling, at the point of disease, may provide options for tests that are either saliva based, or use other ideally non-invasively collected sample types. A unique and advantageous feature of OSCC is that the epithelial lesion where cancer formation takes place is usually accessible to the clinician. Unlike other cancer types affecting internal organs and tissues, OSCC thus offers the possibility of directly collecting specimens in the form of cells and fluids from the oral lesion. Such specimens offer a potentially valuable sample type for identification of diagnostic biomarkers of OSCC.

One example is cell samples collected by brush biopsy (Mehrotra et al., 2009), which can be performed in a minimally invasive manner. Although the commercial brush biopsy kit is designed for cytological analysis of collected cells for OSCC diagnosis, these collected cells can be analyzed using genomic or proteomic approaches to discover new molecular biomarkers. One proteomics-based study sought to identify protein biomarkers from cells collected via brush biopsy from OPML and OSCC patients (Driemel et al., 2007). The MS analysis was performed using SELDI instrumentation, analyzing 49 patients with OPML and 49 with OSCC. Three peaks in the SELDI MS spectra were found to discriminate the patient groups with an area under the receiver operator curve of more than 0.9, indicating fairly high sensitivity and specificity. Two of these peaks were identified as S100A8 and S100A9; however the third was not identified. The use of OPML samples in this study is an advantage, providing evidence that S100A8 and S100A9 are potentially biomarkers capable of diagnosing early transition to OSCC. In a similar fashion, others have measured RNA abundance levels in cells collected via brush biopsy, with one study (Toyoshima et al., 2009)identifying Cytokeratin 17 as a potential biomarker of OSCC within this sample type.

Motivations for this thesis project

The motivations of my thesis work were driven by two competing factors: the potential of whole saliva as a transformative clinical specimen for oral cancer diagnostics, and the difficulties of biomarker discovery in the inherently complex whole saliva proteome. To address the challenge of whole saliva proteomics and realize its

promise, my work took advantage of the ability to access the site of oral cancer development, and non-invasively collect tissue exudate, which is composed of fluid and cells from the surface of these tissues. The overall objective was to evaluate the potential for the proteomic analysis of tissue exudates to guide the search for biomarkers in whole saliva. To achieve this objective, we first developed a method to maximize the amenability of PerioPaper exudate samples to MS-based proteomic analysis. With this method in hand, we next characterized the exudate proteome, and compared exudates from healthy individuals to those from individuals with OPML using exudates samples directly from the lesions demonstrating the amenability of exudate samples to quantitative proteomics. This work is described in Chapter 2.

While characterizing the exudate proteome, and comparing healthy and OPML individuals' exudates demonstrated the promise of exudate analysis, the methods used in our first study faced a few inherent limitations. First of all, the fact that exudates were sampled from a small number of individuals (6 individuals, 3 OPML and 3 healthy) presented too much opportunity for individual variation to be misinterpreted as features defining healthy versus OPML states; Second, the spectral counting approach used, while fitting our needs for a label-free method, allowed relative protein quantities to be estimated only roughly. Finally, it did not prove that proteins identified n exudates could be used to guide evaluation in whole saliva. In response, my follow up study sought to answer these concerns by using paired samples from OPML individuals representing diseased and equivalent healthy locations, using an ILFQ method to improve results, and evaluating candidates from the exudate analysis in while saliva samples.

Chapter 4 provides a final discussion on the significance of the findings from my studies. It also reflects on the strengths and weaknesses of the approaches taken within my studies, and thoughts on future studies to extend my findings to date.

Thesis Chapter 2:

This Chapter is adapted from published work, following the applicable open access policy:

Kooren et al.: Evaluating the potential of a novel oral lesion exudate collection method coupled with mass spectrometry-based proteomics for oral cancer biomarker discovery. *Clinical Proteomics* 2011 8:13.

The early diagnosis of Oral Squamous Cell Carcinoma (OSCC) increases the survival rate of oral cancer. For early diagnosis, molecular biomarkers contained in samples collected non-invasively and directly from at-risk oral premalignant lesions (OPMLs) would be ideal. Therefore In this pilot study we evaluated the potential of a novel method using commercial PerioPaper absorbent strips for non-invasive collection of oral lesion exudate material coupled with mass spectrometry-based proteomics for oral cancer biomarker discovery.

Our evaluation focused on three core issues. First, using an "on-strip" processing method, we found that protein can be isolated from exudate samples in amounts compatible with large-scale mass spectrometry-based proteomic analysis. Second, we found that the OPML exudate proteome was distinct from that of whole saliva, while being similar to the OPML epithelial cell proteome, demonstrating the fidelity of our exudate collection method. Third, in a proof-of-principle study, we identified numerous, inflammation-associated proteins showing an expected increase in abundance in OPML exudates compared to healthy oral tissue exudates. These results demonstrate the feasibility of identifying differentially abundant proteins from exudate samples, which is essential for biomarker discovery studies. Collectively, our findings demonstrate that our exudate collection method coupled with mass spectrometry-based proteomics has great

potential for transforming OSCC biomarker discovery and clinical diagnostics assay development.

Chapter 2 Introduction

Oral cancer occurs most commonly (~90%) in the form of OSCC and develops in stages starting with healthy oral epithelium progressing to an OPML and on to OSCC. The survival rate of OSCC is about 60%. However, where malignancy is detected soon after the transition from OPML, treatments are more effective and survival is as high as 80% (Altekruse et al., 2010; ACS, 2013). Despite the clinical need to distinguish between OSCC and OPML, lesion types are not readily classified by simple visible inspection and more invasive tests are used instead. Currently the gold standard for classifying lesions is to use an incisional biopsy coupled with histological analysis (Rhodus, 2005; Lingen et al., 2008). Yet biopsies have numerous limitations: being invasive clinicians are hesitant to perform them, and patients are hesitant to agree to them due to the pain and discomfort of the procedure; the following histology requires expert analysis and is therefore expensive; and issues such as under-sampling can lead to misdiagnosis (Pentenero et al., 2003).

An ideal alternative to scalpel biopsy would be a non-invasively collected sample rich in molecular biomarkers which distinguish OPML and OSCC, and potentially predict the transition from pre-malignancy to malignancy. One such alternative is the use of protein or nucleic acid biomarkers in saliva that are secreted or shed from the oral lesions (Lee et al., 2009; de Jong et al., 2010). However, despite its benefits, whole saliva is not the direct source of potential biomarkers, and the complexity of the fluid (Bandhakavi et al., 2009) makes identification of potential biomarkers challenging. In contrast to whole saliva, some have directly analyzed incisional biopsy tissues (Ralhan et al., 2008b; Ralhan et al., 2008a), but for clinical diagnostics this approach suffers from the same limitations described above for scalpel biopsy.

Given the ongoing need for improved oral cancer detection, we describe here a promising alternative method for the direct and non-invasive sampling of oral lesions,

which can be coupled with mass spectrometry (MS)-based proteomics. Our method uses commercially available PerioPaper Strips, traditionally used for oral fluid sampling relevant to periodontal disease (Johnson et al., 1999; Grant et al., 2010) to directly collect oral lesion exudate. Exudate is defined as the fluid and cellular material present on the surface of inflamed tissue (2007). Our results show that the exudate samples contain ample protein for large-scale proteomics analysis, and that the exudate proteome of OPMLs is distinct from whole saliva, while being highly similar to the proteome of lesion-associated epithelial cells. We also undertook a pilot study comparing healthy tissue and OPML exudates, demonstrating that the method is amenable to quantitative proteomic analysis necessary for biomarker discovery studies. Collectively our results demonstrate the great potential of our exudate collection method for oral cancer biomarker discovery and clinical diagnostics.

Chapter 2 Experimental Methods

Patient information

Exudates and brush biopsies were collected from three patients diagnosed with a dysplastic OPML at the University of Minnesota Dental School. Exudates and brush biopsies from buccal mucosa, and whole saliva were also collected from three healthy volunteers. All samples were collected with written consent using an IRB protocol approved by the University of Minnesota. Three different lesion and three healthy samples were analyzed to provide some statistical significance for measurements of differential protein abundance between tissue types while balancing the time and cost of large-scale proteomic analysis of individual patient samples.

Exudate sample collection and protein processing

To collect the exudate we first used rolled cotton to swab away ambient saliva around tissue to be sampled (e.g. OPML). The rolled cotton was then moved adjacent to

the area to be sampled to block flow of additional saliva onto the tissue. A PerioPaper strip (Oraflow, Smithtown, New York) was left in position for ~30 seconds. Immediately after collection the strip was placed in a microcentrifuge tube, on ice, and then transferred to a -20^{O} freezer within minutes. The PerioPaper strips were subjected to on-strip trypsin digestion in which each PerioPaper strip containing exudate was submerged in 100 µl buffer containing 5 mM DTT, 100 mM Tris pH 8.0 and boiled for 5 min. After cooling to room temperature, 2 µg of sequencing grade trypsin (Promega, Madison, WI) was added and the microcentrifuge tube was placed in 37^{O} C water bath to digest for 12 hours. The protein digest was then purified and concentrated using Waters Sep-Pak 3cc cartridges as described (de Jong et al., 2010) drying the purified peptides by vacuum centrifuge (~2hrs or until dry). The peptides were analyzed either directly by mass spectrometry or subjected to strong cation exchange (SCX) HPLC fractionation as described below.

Brush Biopsy sample collection and protein processing

To collect brush biopsies we first dried the tissue to be sampled as with exudate collection. Next, we collected transepithelial cells from the tissue using an OralCDx brush test kit (OralCDx laboratories, Inc. Suffern, NY) and following manufacturer's suggested procedure. After collection of cells, the brush head was cut off from the handle and submerged in 250µL of 2× SDS cell lysis buffer (4% SDS, 20% glycerol, 10% 2-mercaptoethanol, and 100 mM Tris-HCl pH 6.8) and 1x protease inhibitors (Complete Mini, Roche Applied Science, Indianapolis, IN, USA) in a 2mL microcentrifuge tube. In order to extract proteins from the cell lysate while removing detergents and minimizing other impurities, the proteins were precipitated with acetone added at a 5:1 ratio and left overnight at -20^{O} C. Precipitated protein was centrifuged at 6000 rpm for 10 min at 4^{O} C, then rinsed and re-centrifuged with pure acetone twice. Proteins were redissolved in trypsin digestion buffer, quantified using the BCA protein assay (Pierce, Rockford, IL, USA), and digested with trypsin as described above for the exudate samples.

Collection and processing of control whole saliva samples

To collect the PerioPaper saliva samples we placed the PerioPaper strip at a location in a healthy volunteer's oral cavity where ambient saliva had pooled (the back of the lower lip). The strip was allowed to saturate with saliva (<20 sec) before being removed. Once collected the PerioPaper saliva samples were immediately placed on ice. On-strip digestion, as described above, was implemented within several minutes of sample collection.

SCX HPLC fractionation

Peptide digests from exudates, brush biopsies, and the whole saliva samples were subjected to offline SCX HPLC fractionation essentially as in previous studies (Bandhakavi et al., 2009). A UV chromatogram (215 nm absorbance) was generated for every different sample. For all samples, SCX fractions containing UV signals indicating the presence of peptides were combined into 9 fractions for subsequent analysis by mass spectrometry. Loading amounts from each peptide fraction were normalized between different samples based on UV absorbance units, to ensure loading of relatively equal amounts of peptides across all different samples.

Shotgun proteomics analysis: Tandem mass spectrometric analysis and sequence database searching

Peptide mixtures from all sample types (exudates, brush biopsy and whole saliva) were analyzed using online capillary liquid chromatography coupled with tandem mass spectrometry (MS/MS) using an LTQ-Orbitrap XL mass spectrometer (Thermo Scientific, San Jose, CA). The chromatography conditions and instrumental parameters used have been described(Bandhakavi et al., 2009). The .RAW files generated by the LTQ-Orbitrap XL were converted to .MSM file format peaklists using "Quant" module

from MaxQuant's (v 1.0.13.13, Max-Planck Institute for Mass Spectrometry, Martinsried, Germany(Cox & Mann, 2008; Cox et al., 2009). The .MSM files are peaklists with high precursor mass accuracy and limited product ion 'noise' peaks. MaxQuant achieves high precursor mass accuracy by using information from LC-MS precursor peaks. Top 6 MS/MS peaks per 100 Da are selected to generate peaklists with limited background noise peaks. The .MSM peaklists were searched using Mascot (v2.1, Matrix Sciences, London, United Kingdom) Daemon and with following parameters : Orbitrap/FT as the instrument, no SILAC labeling, Methionine oxidation as the only variable modification, no fixed modifications, Trypsin as the enzyme, two missed cleavages, MS tolerance at 7 ppm and MS/MS tolerance at 0.5 Da and searched against target-decoy version of Human IPI database (v3.52, Nov 2008) plus contaminant proteins (148372 forward plus reversed sequences). Mascot search generates an output in .dat format that contains peptide-spectrum matching information.

Mascot output .dat files were subjected to statistical validation and protein inference using Scaffold Q+ v 3.0 (Proteome Software, Portland, OR). For peptide identification the false discovery rate threshold was maintained at 1 %. For quantification using spectral counts, total spectra identified in a dataset were normalized with spectra identified in dataset that was to be compared. This normalization, which is achieved by using a display option called "Quantitative value" in Scaffold v 3.0, is used to determine relative abundance of proteins within datasets.

Normalized Spectral Counting and statistical analysis of quantitative proteomics data

Relative abundance levels of identified proteins in healthy and OPML exudates were determined via normalized spectral counting (Lundgren et al., 2010), using the quantitative analysis feature in the Scaffold data viewer software (Version 3, Portland, OR). Quantitative values for each protein were compared in the healthy individuals to those in the OPML individuals differences were determined via assigned P-values using the two-tailed student's t-test (type 2). All proteins with a P-value of less than 0.05 from the healthy exudate to OPML comparison are included in Supplemental File/Table 1. When screening for inflammation-associated proteins showing differential abundance (Table 1), a P-value threshold of < 0.1 was used.

Chapter 2 RESULTS

Our objective was to determine whether exudate collection from oral lesions coupled with MS-based shotgun proteomics is a viable option for oral cancer biomarker discovery.

To achieve our objective three fundamental questions needed to be answered: 1) Are non-invasively collected tissue exudates compatible with MS-based shotgun proteomics? 2) What is the composition and extent of contamination by saliva of the exudate proteome? 3) Can the differential abundance of protein within exudates collected from different tissue (e.g. healthy tissue vs. OPML) be measured?

To answer the first question, we initially explored methods for isolating intact proteins from the PerioPaper strip. For these experiments, we used representative exudate samples collected from healthy oral tissue using the PerioPaper strip as described in Experimental Methods and shown in Figure 4. We first attempted to recover intact proteins from the strips using SDS containing buffers. However, the amount of protein recovered from the strips was very small, at or below the limit of detection for protein quantification using the BCA assay or even reliable detection via SDS-PAGE (data not shown).

Given our inability to isolate ample amounts of intact protein, we instead tested an alternate “on-strip” digestion method. Here, we submerged the PerioPaper strip in buffer containing trypsin, and collected liberated peptides after overnight incubation. Initially, we analyzed a 5% aliquot of peptides by MS/MS to evaluate whether adequate amounts of peptides were captured via the on-strip digestion procedure to enable large-scale protein identification. We identified approximately 140 proteins on average from these samples, which is a reasonable number of proteins when analyzing < 1 microgram of peptides directly by LC-MS/MS. Based on these results we estimated that each exudate sample contained tens of micrograms of total peptides.

Figure 4

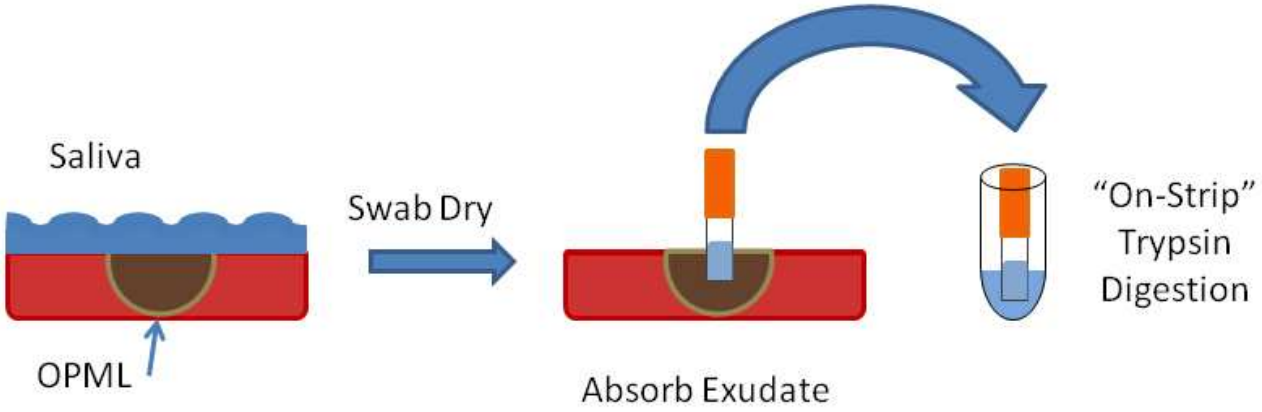

Saliva
OPML
Swab Dry
Absorb Exudate
"On-Strip"
Trypsin
Digestion

Figure 4 Legend: "On-Strip" digestion method

To produce a peptide solution MS-based proteomics analysis, we dried oral epithelium to remove ambient saliva, then placed a PerioPaper strip on the location of interest (Oral Pre-Malignant lesion or normal oral epithelium) and allowed it to absorb exudate. The PerioPaper was next placed in trypsin solution for digestion of proteins to peptides for subsequent processing and mass spectrometry analysis. See Experimental Methods for details.

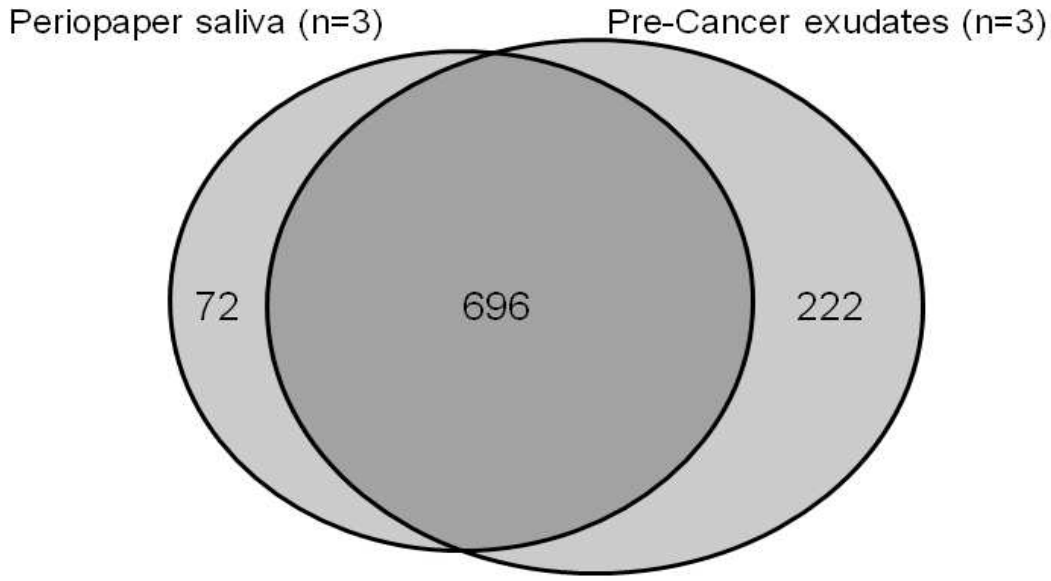
We next employed offline SCX HPLC peptide fractionation (Gygi et al., 2002) in order to increase the number of proteins identified. Fractionation of a representative exudate sample greatly increased our number of protein identifications and sequence coverage for their identifications, producing ~700 identified proteins (< 1% estimated peptide False Discovery Rate).

We then moved on to answering our second question: What is the composition of the OPML proteome and extent of contamination by whole saliva? One initial concern was that the exudate strip would simply absorb saliva, despite our attempts to remove excess ambient salivary fluid from the tissue prior to exudate collection. The high abundant salivary proteins would potentially obscure the identification of lesion-associated proteins. To explore this issue, we compared the protein composition of exudates collected from three different OPML patients, to the composition of whole saliva collected from three different individuals. For this comparison, each whole saliva sample was collected and processed in a similar manner to the exudates, by absorbing saliva onto a PerioPaper strip, followed by on-strip digestion, SCX HPLC fractionation and LC-MS/MS analysis.

We first compared all proteins identified from three saliva samples to all proteins identified from the OPML exudates (Figure 5a). These results showed that the vast majority of proteins from whole saliva were also identified in the exudate samples, indicating that proteins from whole saliva are still prominent within the exudate samples. Next we focused on some of the highest abundance proteins in whole saliva (salivary amylase, lysozyme, proline-rich proteins, and cystatin proteins). We sought to determine whether the relative amounts of these highly abundant proteins were different between whole saliva and exudates. For this investigation, we used normalized spectral counting as a means to assess the relative abundance of selected high abundance saliva proteins within each sample. As shown in Figure 5b, all of the selected salivary proteins were present in significantly higher relative amounts in the whole saliva samples compared to exudate samples.

Figure 5

a


Periopaper saliva (n=3)
Pre-Cancer exudates (n=3)
72
696
222

b

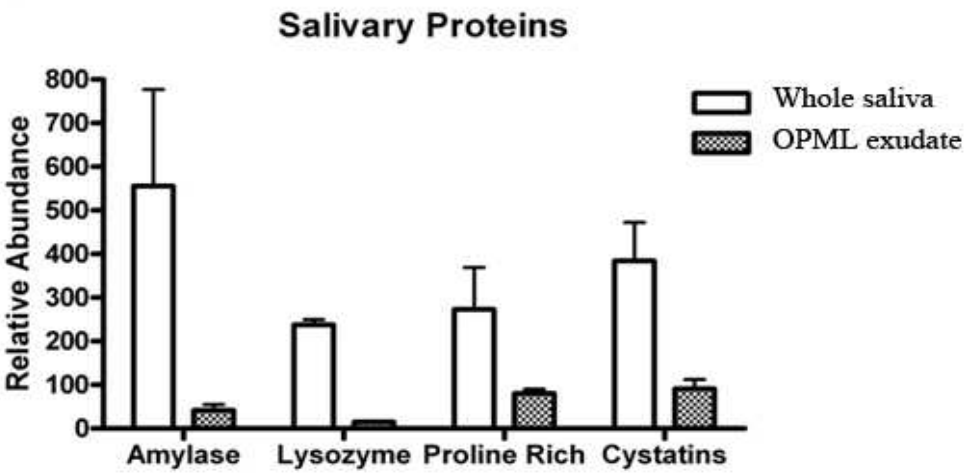

Salivary Proteins
Relative Abundance
800
700
600
500
400
300
200
100
0
Whole saliva
OPML exudate
Amylase
Lysozyme
Proline Rich
Cystatins

Figure 5 Legend: comparison of exudates and saliva

Exudate proteome is distinct from whole saliva: **5a** Venn diagram showing overlap of total protein identifications from PerioPaper collected whole saliva from three individuals compared to OPML exudate proteins from three individuals. **5b** Figure showing the relative proportion of major salivary proteins in OPML exudates compared saliva. See Experimental Methods for dataset generation details.

To further elucidate the composition of the exudate proteome, we compared it to the proteome derived from the epithelial cells collected from the OPMLs. Here, we collected OPML cells via brush biopsy (see Chapter 2 Experimental Methods) from the same three patients that we collected exudates, and analyzed the isolated protein using shotgun proteomics. The proteins identified from the three brush biopsy samples were then compared to the proteins identified from the three exudate samples (Figure 6). The results show that these two sample types are highly similar, with 96% of the proteins found in the exudate samples also present in the brushed cells.

Finally, we sought to answer the question of whether we could measure differential abundance of proteins within exudates collected from different tissue types. Here we decided to compare two distinct tissue types: healthy oral tissue and OPML tissue. Our objective in these experiments was to provide proof-of-principle for conducting quantitative proteomic studies in exudate samples, rather than discover new biomarker candidates. Therefore, we focused on expected protein abundance differences within the samples compared that could serve as a benchmark to determine whether we could reliably measure differential protein abundance in exudate samples. Based on prior studies of OPML and similar inflammatory epithelial lesions (Driemel et al., 2007; He et al., 2008; Ralhan et al., 2009) there are numerous proteins which we would expect to show increased abundance within these inflammatory lesions compared to healthy tissues. We analyzed tissue exudates from three different healthy individuals, and three different individuals with OPML. We focused on the numerous proteins showing increased relative abundance in the OPML samples determined via normalized spectral counting and statistical analysis (see Experimental Methods section). The supplemental File/Table 1 of Kooren et al. (Kooren et al., 2011) shows all proteins determined to show differential abundance between the two groups. Table 1 shows selected proteins with increased relative abundance in the OPML tissue compared to the healthy tissues. As detailed in Table 1, prior studies have established the increased abundance of all of these proteins either in OPMLs or related inflammatory epithelial tissues. Figure 7 graphically shows the measured abundance levels of the proteins in Table 1 as determined via spectral counting.

Figure 6

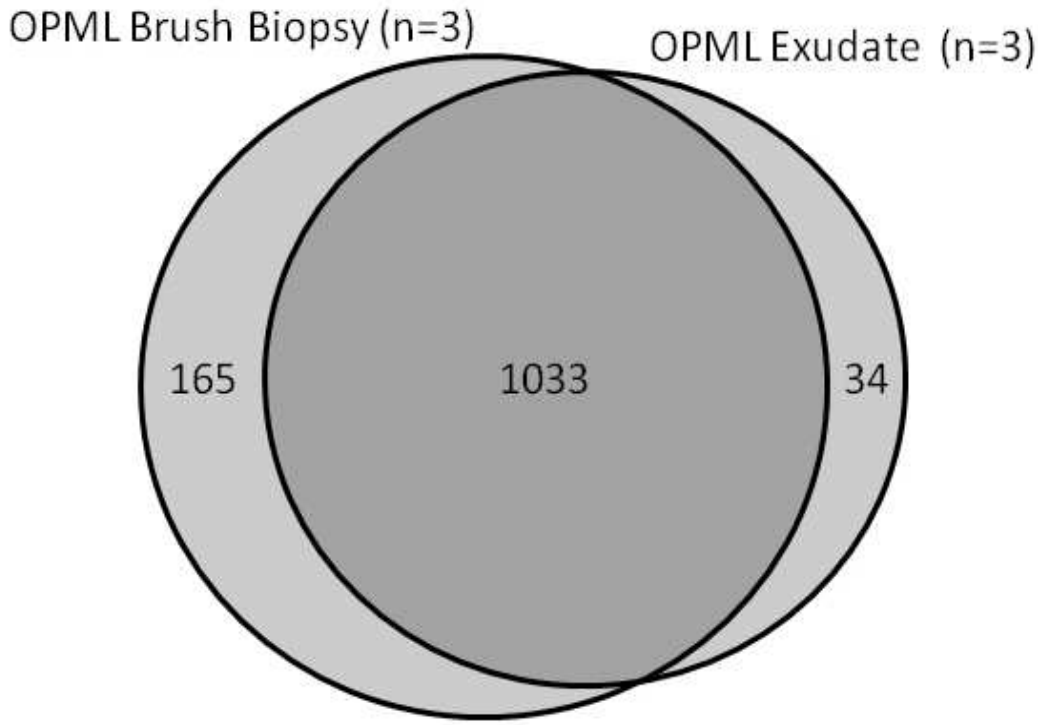


Figure 6 Legend**:** Comparison of exudates and cellular brush biopsies
Identified proteins from brush biopsies vs. exudates from the same subjects (n=3); Exudate proteome is similar to cellular proteome of pre-malignant lesions. See Experimental Methods for dataset generation details.

Table 1. Selected proteins showing increased relative abundance in OPML tissue compared to healthy

Protein	Quantitative ratio[1]	P value[1]	Evidence of association with OPML and/or epithelial inflammation
hnRNPM	8.00	0.091	RNA binding, splicing and inflammation signaling; increased abundance of hnRNPs has been measured in OPML tissue (Ralhan et al., 2009)
IL1F6	22.00	0.016	Cytokine involved in inflammation and immune response; increases in abundance in inflamed epithelial tissues (Blumberg et al., 2007); (Ichii et al., 2010)
LCN2	4.00	0.021	Iron transporter involved in immune response and apoptosis; activated in inflammatory and pre-malignant tissues (Bolignano et al., 2010); (Nielsen et al., 1996)
S100A8	2.02	0.088	Calcium binder and pro-inflammatory factor; increases in abundance in OPML tissue (Driemel et al., 2007)
NQO1	10.00	0.050	Quinone reductase; induced under inflammatory conditions (Rushworth et al., 2008)
XRCC5/6	4.00/8.00	0.016/0.001	Protein complex involved in DNA repair; DNA damage response proteins known to be activated in OPML (He et al., 2008) and other dysplastic epithelial lesions (Raynaud et al., 2008)

[1]Quantitative values and P values determined as described in Chapter 2 Experimental Methods

Table 1 Legend: OPML associated proteins

Selected proteins with increased relative abundance in the OPML tissue compared to the healthy tissues. As detailed, prior studies have established the increased abundance of all of these proteins either in OPMLs or related inflammatory epithelial tissues.

Figure 7

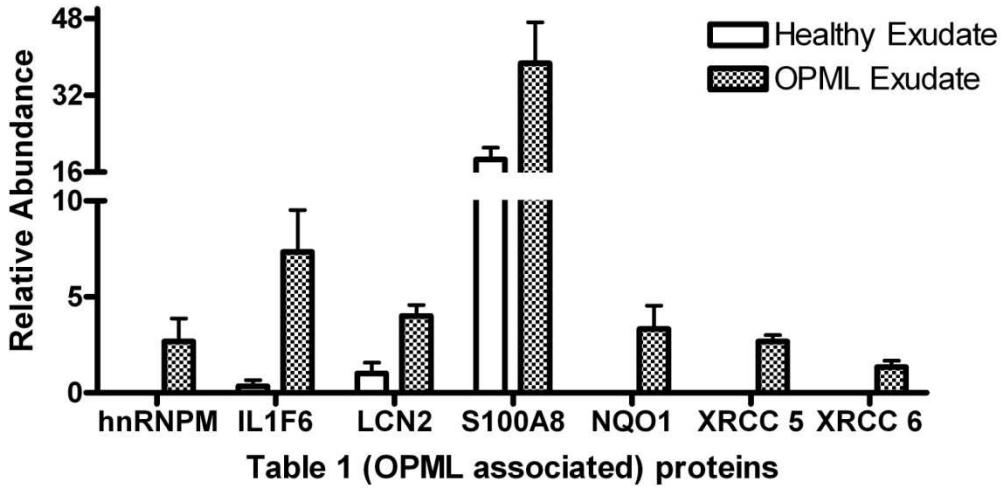

Relative Abundance
48
32
16
10
5
0
Healthy Exudate
OPML Exudate
hnRNPM
IL1F6
LCN2
S100A8
NQO1
XRCC 5
XRCC 6
Table 1 (OPML associated) proteins

Figure 7 legend: Inflammation associated proteins in OPML exudates

Plot of abundance levels of inflammation-associated proteins identified in healthy and OPML tissue exudates. See Table I for more details on each protein.

Chapter 2 Discussion

In this study, the first fundamental question addressed was whether exudates collected via PerioPaper strips can be analyzed using MS-based proteomics. Due to the small quantity of exudate collected and proteomic novelty of the sample, the possibility that it may not contain ample protein amounts to facilitate their analysis using MS needed to be addressed. Although isolating intact proteins from the strips proved difficult, our on-strip digestion method liberated ample amounts of peptides for large-scale shotgun proteomic analysis.

We next investigated the composition of the OPML exudate proteome. One concern was absorbance of ambient saliva, and the high abundance proteins contained therein which may obscure the identification of lower abundance exudate proteins. However, comparison of exudate samples to whole saliva samples revealed that high abundance salivary proteins were greatly decreased in their relative amounts within the exudate samples. Thus, our exudate collection procedure sufficiently removes ambient saliva, which should enable identification of proteins sampled directly from the lesion tissue.

Interestingly, the exudate proteome was highly similar to the proteome of epithelial cells collected directly from the lesions via brush biopsy. This indicated that the PerioPaper strip also absorbs cells from the surface of the lesion, whose protein contents can be detected using our processing method. The presence of cellular proteins is in keeping with the accepted definition of an exudate, which is a mixture of fluid, cells and cellular debris on the surface of inflamed tissue (2007). Thus, exudate collection with PerioPaper strips may offer an alternative to brush biopsy collection, with the advantage that the processing of the samples via on-strip trypsin digestion being simpler than processing intact cells collected via brush biopsy, which includes additional steps of cell lysis and protein precipitation. Additionally, a paper-strip based collection method also has potential use in micro- or nano-scale devices for point-of-care clinical testing (Klasner et al., 2010).

Finally we investigated the feasibility of quantitative proteomic analysis of exudate samples collected from different tissue types. By comparing exudates from healthy oral tissue to OPML tissue we expected to identify inflammation-related proteins at increased abundance in the lesion exudates. Indeed this was the case, demonstrated by the proteins listed in Table 1. Thus we conclude that our method of exudate sampling and MS-based proteomic analysis for profiling is compatible with profiling differential protein abundance, necessary for biomarker discovery studies.

In conclusion, we demonstrate here a promising new method for the non-invasive, direct sampling of oral lesions, compatible with MS-based proteomics. In future studies, we envision application of this method to comparative analysis of OPMLs and malignant oral lesions. This should provide a powerful means to identify protein biomarkers distinguishing these lesion types that may be useful for early detection of malignant transformation. Given our findings, lesion exudates should be amenable to the full suite of proteomic analysis tools, including those aimed at identifying post-translational modifications or sequence variants (Menon et al., 2009) that may serve as powerful biomarkers of oral cancer. Additionally, exudate samples analyzed by MS-based proteomics should be amenable to a metaproteomics approach (Rudney et al., 2010) seeking to identify bacterial or viral components of oral lesions which may play a role in pathogenesis. Proteins identified within the exudate samples may also serve as a guide to identifying lesion-derived proteins shed into the saliva that could be used for oral cancer detection in this easily collected fluid (Hu et al., 2008; de Jong et al., 2010). Finally, the PerioPaper strips could provide the foundation for point-of-care clinical devices for oral cancer diagnostics, given the emergence of such devices designed for paper-based fluid sampling and analysis(Klasner et al., 2010).

Thesis Chapter 3:

Guiding salivary OSCC biomarker discovery via proteomic analysis of lesion exudates

Salivary protein biomarkers offer potentially optimal means to diagnose and monitor oral cancer progression. Unfortunately, the complexity of the salivary proteome challenges the discovery of promising biomarker candidates. Quantitative proteomic analysis of tissue exudates collected directly from oral lesions could address this challenge, by identifying proteins at the site of cancer development that could be tested in whole saliva and thereby guide the discovery of promising biomarkers. The objective of this proof-of-principle study was to determine the feasibility of this approach. Using a comprehensive, label-free quantitative proteomics strategy, we first analyzed paired (from the same individual) control and OPML exudates (figure 8), seeking to identify differentially abundant proteins between these tissue types. After prioritization, we selected several differentially abundant proteins for testing in while saliva, comparing their relative abundance levels in healthy, OPML and oral Squamous cell carcinoma (OSCC) subjects. Two proteins, CK10 and A1AT showed notable differences between saliva samples compared. CK10 showed a small increase from healthy to OPML patients, with a larger decrease in OSCC patients, consistent with its proposed role as a tumor suppressor. A1AT shoed a differential SDS-PAGE migration pattern, especially between OPML and OSCC patients, consistent with published findings suggesting its post-translational modification within tumors. Collectively, our results provide a demonstration of the value of tissue exudate analysis for guiding salivary biomarker discovery in oral cancer, as well as providing promising biomarker candidates for future evaluation.

Chapter 3 Introduction

Saliva is a potentially excellent non-invasive sample for oral cancer diagnostics for several reasons, including: availability, ease of access, contact with the disease location, and the potential for use in point-of-care diagnostic methods or devices (Hofman, 2001; Wong, 2006). Proteins found in whole saliva may make highly promising biomarkers if their abundance levels discriminate between patients with premalignant lesions (OPML) and those with malignant lesions (OSCC). Despite this potential promise, profiling the salivary proteome can be difficult. A main reason for this difficulty is the complexity of whole saliva(Bandhakavi et al., 2009), with most of the endogenous protein content from sources other than OPML or OSCC locations, which suppresses detection of cancer-associated proteins and adds to biological variation.

While saliva maintains advantages as an ideal sample medium for a diagnostic test, it may be helpful to use a more direct sampling method for biomarker discovery, and validate it in saliva. Due to direct access to the disease site, valuable protein content, and ease of sample collection, oral lesion exudates are an ideal aide to salivary biomarker discovery as our proof of principle studies in Chapter 2 demonstrated (Kooren et al., 2011). The work in this chapter extends these initial results, seeking to demonstrate the use of oral exudate sampling coupled with quantitative proteomics analysis to guide biomarker discovery in whole saliva.

As in our initial work with oral exudates (Chapter 2), here we used a label-free approach for quantitative proteomic analysis. The rather simple spectral counting method, used in our initial studies and summarized in the introductory chapter, can indicate the possibility of a relative protein abundance difference. However, there are important sources of uncertainty that limit its usefulness. Use of an exclusion duration in sampling a precursor peptide for fragmentation and analysis by tandem mass spectrometry (MS/MS) prevents abundant peptides from being sampled more than a few times so that less abundant peptides are more readily identified. It is a valuable tool for identifying large numbers of peptides, but limits the usefulness of spectral counting quantification. Also, the uncertainty of whether the obtained MS/MS spectrum of a given

ion will lead to a positive identification when it's sampled adds another level of uncertainty.

Exclusion duration is an experimental parameter used data dependent in liquid chromatography (LC)- MS/MS analysis of complex peptide mixtures, where each peptide ion selected for MS/MS is prevented from being re-selected for a set amount of time, thus enabling more comprehensive selection of peptides, including those of lower abundance (Gatlin et al., 2000). While the greater number of subsequent peptide sequence matches to MS/MS spectra are desired, linearity between ion abundance and the number (or counts) of corresponding MS/MS spectra is lost. Additionally, the uncertainty of whether an MS/MS sampling of a given ion will lead to a confident match to a peptide sequence each time it's sampled, also decouples MS/MS spectral counts from abundance because a peptide may be either identified or not in separate analyses even when signal intensity and abundance are similar. These constraints led us to pursue a more comprehensive quantitative proteomics approach, using not only spectral counting but also intensity-based label-free quantification (ILFQ), a concept introduced in Chapter 1.

The objective of the work presented in this chapter was this: to determine whether quantitative proteomic analysis of exudates collected from different oral tissues could identify differentially abundant proteins, which in turn showed parallel abundance changes in whole saliva. Achieving this objective would provide proof-of-principle demonstration of the value of oral lesion exudates for guiding salivary protein biomarker discovery.

Chapter 3 Methods

Sample collection and preparation

We collected paired samples from 6 volunteers in the Ear, Nose and Throat Clinic at the University of Minnesota in accordance with our Institutional Review Board guidelines (Approved IRB study number 0001M34501). These samples were from laterally symmetric locations in the oral cavity, an OPML and a healthy control sample

were collected from each patient (figure 7). Prior to exudate collection, whole saliva samples were collected from patients with an OPML, as well as healthy control patients without an oral lesion. Control individuals were selected for similar age, sex and smoking status to the OPML and OSCC individuals. Table 2 provides information on patients enrolled in this study. These samples were frozen immediately upon collection and subsequently underwent on-strip digestion (Kooren et al., 2011). The digested exudate samples were then fractionated by offline (strong cation exchange) SCX HPLC peptide fractionation as in previous studies (Bandhakavi et al., 2009).

A UV chromatogram (215 nm absorbance) was generated from eluted strong cation exchange fractions collected from each sample (OPML and paired control). Seven fractions were created from each sample with identical chromatographic retention times (15 min to 78 min retention total 9 min per fraction). The 215 nm absorbance of each fraction was used to normalize the relative quantity of peptide in each fraction. Paired fractions contained the same peptide quantity by 215 nm absorbance, with all fractions having similar total quantities (~ 1µg) loaded onto the mass spectrometer.

On-line HPLC and electrospray ionization MS/MS analysis

The fractionated peptide samples were analyzed by online capillary liquid chromatography coupled with tandem mass spectrometry (MS/MS) using an LTQ-Orbitrap XL mass spectrometer (Thermo Scientific, San Jose, CA). The chromatography conditions were as described (Bandhakavi et al., 2009), except with the reversed phase HPLC gradient was lengthened to 90 min (same gradient over longer time).

Data analysis methods

Acquired data was analyzed via five methods, each being described in a point-by-point fashion below.

Table 2:

Patient	samples	type	MS used	MS sample	sex	age	tobacco
196	196a	OPML saliva	n		F	60	non-smoker
	196f	control exudate	y	P1C			
	196g	lesion exudate	y	P1L			
197	197a	OPML saliva	n		F	49	non-smoker
	197d	control exudate	y	P2C			
	197e	lesion exudate	y	P2L			
198	198a	OPML saliva	n		M	66	non-smoker
	198d	control exudate	y	P3C			
	198e	lesion exudate	y	P3L			
199	199d	control exudate	y	P4C	F	71	ex- smoker
	199e	lesion exudate	y	P4L			
200	200f	control exudate	y	P5C	F	56	non-smoker
	200g	lesion exudate	y	P5L			
201	201a	OPML saliva	n		F	70	non-smoker
	201g	control exudate	y	P6C			
	201f	lesion exudate	y	P6L			
231	231a	OPML saliva	n		M	62	ex-smoker
232	232a	OPML saliva	n		F	88	non-smoker
268	268a	Healthy control saliva	n		M	40	chew user
269	269a	Healthy control saliva	n		F	66	non-smoker
270	270a	Healthy control saliva	n		F	84	non-smoker
271	271a	Healthy control saliva	n		F	65	non-smoker
272	272a	Healthy control saliva	n		M	63	non-smoker
274	274a	Healthy control saliva	n		F	57	ex-smoker
275	275a	Healthy control saliva	n		F	57	non-smoker
245	245a	OSCC saliva	n		M	76	current smoker
248	248a	OSCC saliva	n		M	47	current smoker
256	256a	OSCC saliva	n		M	55	ex-smoker
258	258a	OSCC saliva	n		M	39	ex-smoker
260	260a	OSCC saliva	n		F	71	ex-smoker
261	261a	OSCC saliva	n		M	42	non-smoker
262	262a	OSCC saliva	n		M	66	ex-smoker

Table 2 Legend: Clinical samples used in Chapter 3

Sample information for Chapter 3 Healthy, OPML and OSCC samples; for those used in the MS based portion of the research project the MS sample name is provided. This table includes salivary samples that were used for western blotting testing of candidate markers (salivary samlples not used in MS analysis).

Data analysis method 1: Quant/Mascot/Scaffold 3 data analysis

The .RAW files generated were first converted to .MSM file format peaklists using the "Quant" module from MaxQuant (v1.0.13.13, Max-Planck Institute for Mass Spectrometry, Martinsried, Germany) (Cox & Mann, 2008; Cox et al., 2009). This created .MSM files peaklists that feature higher precursor mass accuracy and reduced product ion 'noise' peaks. To search the .MSM peaklists, we used Mascot Daemon (v2.1, Matrix Sciences, London, United Kingdom), as previously reported (Kooren et al., 2011), except with a few exceptions. These are the use of ESI-TRAP as the instrument parameter, no fixed modifications, the combination of all .MSM files for each sample (OPML and Control separate) for each search, and a concatenated human-Uniprot and the human oral microbial database (HOMD). The HOMD data was included for more comprehensive peptide sequence matching, including possible peptides from microbial sources. The database was generated by using a two-step approach wherein matches derived from a primary search against a large database were used to create a smaller subset database. The second search was performed against a target-decoy version of this subset database merged with a host database, as in (Jagtap et al. 2013). The .dat format files generated by this search contain peptide-spectrum matching information. We next subjected mascot .dat files to statistical validation and protein inference using Scaffold 3 (Proteome software, Portland, OR) with a false discovery rate threshold of 1%, again as previously described in chapter 2 (Kooren et al., 2011). This resulted in 979 protein identifications (supplemental file/table 2). Where the Quant/Mascot/Scaffold 3, or simply Scaffold 3 analysis is mentioned in this paper, SC-based normalization using a Scaffold 3 feature "Quantitative value" was applied to estimate relative abundance of proteins within datasets.

Data analysis method 2: Proximity-based normalization via RIPPER

For the peptide feature based quantification we employed the Proximity based Intensity Normalization (PIN) method developed by Susan Van Riper, initially described

in de Jong et al (de Jong et al., 2011) and implemented within the RIPPER software framework. Publications describing PIN/RIPPER are currently in submission, and a provisional patent (registration number: 69874) has been filed.

In order to adapt the software to a fractionated proteomic dataset, we had to make a number of modifications. One key need was matching up signals from the same peptide ion detected across different patient samples. This involved specifying a retention time (RT) and m/z tolerance for any given peptide ion to define its signal within one LC-MS run, and match it to the LC-MS data from another patient using these same RT and m/z parameters, ultimately desiring to compare signal intensities for the peptide between the different samples. After much examination of possible windows, $\pm$ 0.005 m/z and $\pm$ 2 min were selected as the acceptable RT and m/z windows within and between runs.

RIPPER was then used for analysis of all data, first converting each .RAW LC-MS data file to mzXML using the program msConvert (available from proteowizard.sourceforge.net). This analysis created from each LC-MS run a list of normalized peptide signals with associated intensities, m/z, and retention time characteristics. We then used Genedata Analyst (Genedata Inc) to select those peptide signals that showed statistically significant differential abundance between categories (OPML or control) considering pairwise comparison of the patient samples (paired healthy tissue compared to OPML tissue in each patient). We kept fractionation in mind by comparing equivalent fractions collected for each separate sample (example: Peptide signal A in Patient 1 OPML tissue SCX fraction 2 was compared to Patient 1 paired control tissue SCX fraction 2).

Data analysis method 3: ProteinPilot searches

OPML and healthy control samples were searched against the concatenated human-Uniprot and HOMD database using ProteinPilot v 4.0 software. Briefly, MGF files created from MaxQuant “Quant” module were used for ProteinPilot search (Jagtap et al 2012) with following parameters:: Instrument: LTQ/Orbitrap subppm; Digestion:

Trypsin; ID Focus: Biological Modifications; Search effort: Thorough; Protein identification threshold: 10% Conf. Peptide spectrum matches (PSMs) identified at 1% local FDR were further processed for their modifications and these distinct peptide sequences were represented with their spectral counts and protein information (supplemental file/table 3). The ProteinPilot identifications were aligned with RIPPER normalization and quantification analysis by using the observed m/z values and information regarding the sample, fraction and san number (see below). In particular, scan number was used to identify the correct retention time. We found that these set of values were sufficient and accurate enough to align the RIPPER quantification to Protein Pilot identifications.

Data analysis method 4: Genedata Analyst applied to RIPPER and Protein Pilot results.

The results of RIPPER analysis (peptide ion m/z and RT range values) were imported into Genedata Analyst. These normalized peptide ion feature intensities were sorted by the two groups paired statistical test using the following nomenclature for patient comparisons: Patient n control lesion was annotated as PnC, patient n OPML was annotated as PnL. Thus a paired effect size and p-value were generated for each peptide signal compared between healthy tissue and OPML tissue in each patient (P1-P6). To generate a list of candidate differentially abundant peptides for further examination, all peptide signals with a paired effect > 2.0 and a p-value <0.05 were extracted, and matched to the peptide sequence associated with that m/z and RT from the ProteinPilot analysis. The ion feature m/z had to match within 0.005 Da, with an MS/MS acquired and matched to the peptide sequence within the RT window tracked by RIPPER. Protein identification inference relied on ProteinPilot, and other PSMs derived from the inferred proteins were examined in light of RIPPER results. The presence of other peptides that were deemed differentially abundant in the same comparison using slightly relaxed criteria ($p<0.1$, paired effect>1.7) were noted. Additionally we made sure that there were no conflicting results from any given inferred protein considered for further analysis,

such as having PSMs with abundance changes that showed a reverse abundance difference between the healthy and OPML tissue within the same patient.

Data analysis method 5: Refiner MS and Analyst analysis

In the case of data analysis by Refiner MS and Analyst (Genedata), the .RAW files were uploaded into Refiner MS. Refiner MS was used to produce chromatograms from .RAW files, which then underwent noise subtraction, retention time alignment, detection of sustained peaks over time, and de-isotoping, followed by a Mascot search, all within the Refiner MS framework. The results, in the form of a .GDA file, were imported into analyst and treated similarly to the RIPPER data. The main exception was using protein level rather than peptide level analysis to determine paired effect size and p-value.

Western Blotting methodology

Commercial antibodies used for western blot validation experiments were as follows (product numbers in parentheses): Cytokeratin 10 monoclonal antibody (Abcam Inc. ab9025), and alpha 1 Antitrypsin monoclonal antibody (Abcam Inc. ab9400). Five µg of protein from each individual subject analyzed in validation experiments (7 healthy individuals, 7 OSCC individuals, and 6 OPML individuals) and molecular weight markers was loaded in separate lanes and separated via SDS-PAGE on a Bio-Rad minigel apparatus. The proteins were transferred to a PVDF membrane and the membrane blocked with 1xPBS containing 2.5 mg/ml BSA and 0.25% (v/v) Tween 20, followed by incubation with the primary antibody, diluted in PBS at concentrations suggested by the antibody manufacturer. After washing, the membranes were incubated with secondary antibodies conjugated to horseradish peroxidase (HRP). For detection, membrane-bound proteins recognized by the HRP conjugated secondary antibody were visualized by chemiluminescence using the SuperSignal West Pico ECL Substrate (Thermo Scientific).

Membrane was performed by first treating the PVDF membrane with Coomasie Stain solution (50% methanol, 10% acetic acid, 0.25% Coomasie Brilliant Blue), submerged with 10 RPM rocking for 20 min. Next the membrane is rinsed with deionized water for 2 min, followed by treatment with destain solution (40% methanol, 10% acetic acid), again submerged and rocking at 10 RPM for 30 min, and repeated with fresh destain solution.

Chapter 3 Results

Our objective was to analyze healthy and OPML lesion exudates and use label free quantification to find differentially abundant proteins directly associated with lesion tissue, and determine whether these proteins were also differentially abundant in whole saliva, thereby revealing promising salivary biomarker candidates.

Achieving this objective required completing three steps: 1) Designing and applying label-free relative quantification methods to determine differentially abundant proteins in different tissue types; 2) For selected differentially abundant proteins, testing their relative abundance in whole saliva from healthy volunteers versus OPML patients; 3) Further evaluation of promising candidates from step 2 to determine differential abundance between OPML and OSCC patient's whole saliva.

Step 1: Designing and using label-free relative quantification for determining differentially abundant proteins in paired healthy and OPML tissue exudates

For the first step we chose a label-free approach to quantity our data. Given the nature of the exudate sample there are many reasons for this approach. Exudate samples were collected with PerioPaper strips, which absorb the exudate fluid and associated proteome. The best protein identification results from exudate samples involve on-strip digestion, which digests proteins to liberate a peptide sample from the PerioPaper strip, as in chapter 2, (Kooren et al., 2011). Because a peptide rather than protein solution is formed, more accurate protein quantification methods cannot be used. Chemical labeling

(e.g. iTRAQ) requires accurate sample quantification before labeling; both the small quantities and tryptic peptide nature of exudate samples limits the effectiveness and compromises our ability to normalize labeling. We instead chose label-free approaches, wherein we could use data normalization methods to account for small differences in loading amounts to provide quantitative information. We also used the UV-profile from the SCX fractionation step to estimate peptide amounts in each fraction, and equalizes loading for LC-MS/MS analysis between paired tissue samples.

In our previous studies, we compared exudates from healthy and OPML individuals in order to test the value of exudate samples (Kooren et al., 2011). However, the fact that healthy and OPML exudate samples were from different individuals (e.g. not paired tissue samples from the same individual) introduced complexity and differences that were not related to clinical status. Therefore in this study we analyzed exudate samples taken from paired OPML diseased and healthy oral locations (using lateral symmetry, see figure 8) so that observed differences were more likely due to disease state.

We analyzed six paired samples from OPML patients (see materials and methods). The spectral counting first used the available tools within Scaffold 3^{TM} (data analysis method 1) to develop a list of candidates that stood out as being differentially abundant in either the lesion tissue or paired healthy tissue exudates from each patient. Initial analysis with Scaffold 3^{TM} used p-values generated for normalized spectral counts, and normalized Total Ion Current (TIC) of spectral counts. We also examined fold changes observed within individual sample pairs, focusing on those proteins that showed significant fold change as defined by Scaffold 3 statistics in three or more patients with no contradicting results. While some initial candidates were generated, the three quantification methods had very few overlaps (Table 3). One protein that stood out in the pairwise analysis was Alpha-1-Antitrypsin (A1AT).

Because the different SC-based methods in Scaffold had little overlap, we also wanted to use ILFQ in order to compare relative abundance profiles. In order to do so we applied the PIN method, implemented in the software RIPPER (data analysis method 2) originally developed for ILFQ in LC-MS peptidomics experiments (de Jong et al., 2011)

and modified it for use with a fractionated proteomic dataset. Using RIPPER we created first a candidate list of peptide signals (MS1 M/Z signals with retention time duration and intensity) and subsequently matched these to high confidence PSMs resulting from their selection for MS/MS, thereby generating a peptide candidate list. To generate a more comprehensive listing of PSMs we used Protein Pilot to identify spectra in addition to the previously mentioned Mascot/Scaffold3 workflow (data analysis method 3).

With PSMs in hand for many of the peptide signals, we prioritized the proteins to select those of highest interest for possible evaluation in saliva (data analysis method 4). We examined the protein sources of the peptide signals that indicated differential abundance between the OPML-location and matched healthy location samples. We next narrowed this protein candidate list by including only those with two separate peptides tracked as peptide signals, in which at least one signal indicated a significant abundance difference, and no other peptide signals from the protein contradicted the differential abundance (e.g. all up in OPML, or all down in OPML). In order to properly interpret the RIPPER output we used Genedata Analyst™ to organize results and evaluate pair-wise comparisons (matched healthy versus OPML abundance) across all patients. This generated a condensed list of candidate proteins deemed significant by RIPPER analysis (Table 4).

Figure 8

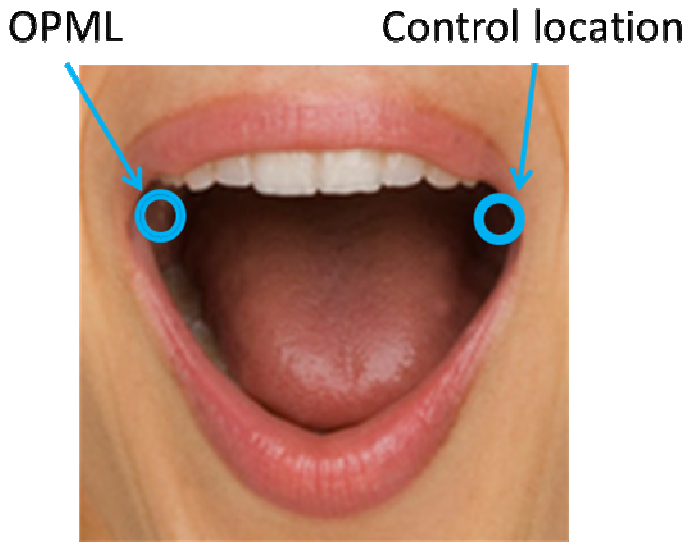


Figure 8 legend: Paired Samples

Diagram depicting paired OPML and control location site selection. For OPML patients, the paired healthy exudate control was taken from healthy tissue at a location laterally symmetric to the OPML location.

Table 3:

Uniprot protein ID	normalized spectral count	Total TIC	significant fold change in 3 or more pairs, without contradiction	OPML or Control associated	grouping ambiguity
UBB_HUMAN	yes	no	no	OPML	no
AACT_HUMAN	yes	yes	no	OPML	no
IF4H_HUMAN	yes	no	no	Control	no
RL18A_HUMAN	yes	yes	no	Control	no
ARC1A_HUMAN	yes	no	no	Control	no
AK1BA_HUMAN	yes	no	yes	Control	no
RABP2_HUMAN	yes	no	no	OPML	no
PTMA_HUMAN	yes	no	no	OPML	no
GELS_HUMAN	yes	no	no	Control	yes
IDHC_HUMAN	yes	no	no	OPML	yes
LGUL_HUMAN	yes	yes	no	Control	no
MYH14_HUMAN	no	yes	no	OPML	yes
KRT84_HUMAN	no	yes	no	OPML	yes
VIME_HUMAN	no	yes	no	Control	yes
CLIC1_HUMAN	no	yes	yes	OPML	no
PDIA6_HUMAN	no	yes	no	OPML	no
PDIA4_HUMAN	no	yes	no	OPML	no
LMO7_HUMAN	no	yes	no	OPML	no
ATP5J_HUMAN	no	yes	no	OPML	no
TCPA_HUMAN	no	yes	no	OPML	no
RL18A_HUMAN	no	yes	no	Control	no
CPNS1_HUMAN	no	yes	no	OPML	yes
NIBL1_HUMAN	no	yes	no	OPML	no
RL23A_HUMAN	no	yes	no	OPML	no
B2MG_HUMAN	no	yes	no	OPML	no
K22O_HUMAN	no	yes	no	Control	yes
A1AT_HUMAN	no	no	yes	OPML	no
KV402_HUMAN	no	no	yes	OPML	no
FAM25_HUMAN	no	no	yes	Control	no
GDIB_HUMAN	no	no	yes	Control	yes
RTN4_HUMAN	no	no	yes	Control	no

Table 3 Legend: Max Quant/Mascot/Scaffold approach candidates

Protein candidates from analysis using the Quant module (Max Quant) with Mascot and Scaffold 3; The spectral counting, Total Ion Current of spectral counts, and pairwise comparison of spectral counts Scaffold 3 (including their normalization tools) did not display significant overlap.

Column headings:

Accession # protein accession number

Normalized spectral count: was $P<0.05$ when comparing measurements of spectral counts in Scaffold 3^{TM} with normalization?

Total TIC: was $P<0.05$ when comparing measured Total Ion Current (TIC) of spectral counts in Scaffold 3^{TM}

Significant change: When individual .sf3 files were created for each pair of OPML and healthy location samples, was there a significant fold change observed in the same direction for 3+ out of the six samples pairs, and with no pairs showing the opposite trend at all (whether significant fold change or not)?

OPML or Control associated: Was the candidate more abundant (associated with) the OPML or Control portions of sample pairs?

Grouping Ambiguity: Were the peptides used to identify the protein subject to grouping ambiguity?

We also conducted a separate analysis of pair-wise protein abundance differences using the ILFQ functionality build into the Genedata workflow (data analysis method 5), coupling the Refiner MS and Analyst modules. The output from this analysis supported the use of one previously noted potential candidate Alpha 1 Antitrypsin (Table 5) but was otherwise not central to our generation of candidates. We instead focused on the RIPPER-derived results, based on our confidence in its superior normalization routing for better accuracy (Van Riper, unpublished results), and the spectral counting results from Scaffold as a complementary method to RIPPER's intensity-based measurements.

Comparison of the most promising candidates by these methods indicated a trend wherein promising candidates generated by RIPPER, or even those detected by the software at all, were mostly higher abundance proteins with multiple peptide identifications. Curiously, the spectral counting quantification of these same candidates often did not display any significant abundance differences between the tissue types (Figure 9a).

Step 2: Testing relative abundance of selected protein candidates in whole saliva from healthy volunteers versus OPML patients

For the second step, we turned to testing for corresponding differential abundance in whole saliva by western blotting of a select few candidates generated from the quantitative proteomics data (Table 6). Candidates for testing were chosen both by the quality of the MS data (e.g. statistical significance of abundance difference, confidence in PSMs based on manual confirmation), and also based on factors such as amenability to western blotting (antibody availability, molecular weight) and potential linkage to cancer mechanisms.

For this testing, we focused on the soluble supernatant of whole saliva, as this sample is easily collected via low-speed centrifugation, without the need for further processing. Initially, we established whether or not proteins could be detected reliably by western blotting in salivary supernatant, either in sample collected from healthy controls or patients with an OPML. Based on these initial experiments, some proteins were

excluded from further testing, based on either complete lack of detection or only sporadic detection in only a few samples, which would preclude any assessment of differential abundance trends between patients. In the case of AHNAK, exclusion was based on high molecular weight, which both lessens the confidence that the few RIPPER tracked peptides were representative, and is a possible confounding factor in western blotting. Cytokeratin 4 seemed to only be present in the cellular portion of saliva, which is itself more variable and only intermittently a part of saliva samples. Anti-chymotrypsin was only sporadically detectable in OPML, and not at all in controls.

For those that could be detected reliably, we next examined whether or not these showed a trend towards differential abundance corresponding to those that we measured in the exudates. One protein, CK10 demonstrated a clear trend of increasing abundance between healthy and OPML patients (Figure 10a), similar to our exudate measurements. Another candidate A1AT, although showing a less obvious differential abundance trend, showed a possible shift in migration pattern in the gel, indicating potential post-translational modification (Figure 11 a and b). Therefore we decided to proceed with further evaluation of this candidate in OPML and OSCC samples.

Step 3: Evaluating relative abundance of promising candidates in OPML and OSCC patient's whole saliva.

For the final step, we sought to determine whether the two most promising proteins from our initial validations, CK10 and A1AT, displayed differential abundance in saliva from OPML and OSCC patients. In the case of CK10, after an increase between healthy and OPML, it shows a trend towards decreased abundance in OSCC, being barely detectable in a number of the OSCC samples (Figure 9b). Meanwhile A1AT shows an interesting differential migration pattern between OPML and OSCC, suggesting possible differential post-translation modification states between the patient groups.

Table 4:

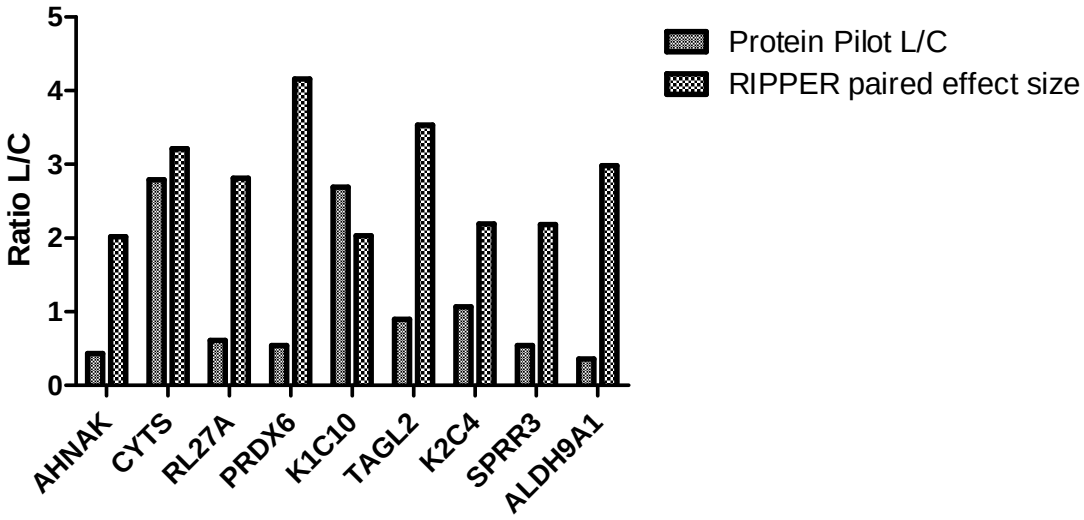
Uniprot Entry name	Protein name	# of peptides with paired effect >1.7 P value < 0.1	Top peptide (paired effect >2, P value < 0.05)
AHNK_HUMAN	Neuroblast differentiation-associated protein AHNAK	2	FSMPGFK Oxidation(M)@3
K1C14_HUMAN	Cytokeratin 14	1	NHEEEMNALR Deamidated(N)@1; Oxidation(M)@6
K1C9_HUMAN	Keratin, type I cytoskeletal 9	1	IKFEMEQNLR Oxidation(M)@5, missed K-F@2
CYTS_HUMAN	Cystatin-S	3	RPLQVLR
B8ZZW7_HUMAN	Thymosin alpha-1	1	AAEDDEDDDVDTK K, missed K-K@13
E9PLX7_HUMAN	60S ribosomal protein L27a	1	TGAAPIIDVVR
F5GZ12_HUMAN	Small proline-rich protein 3	2	VPEQGYTKVPVPGY TK, missed K-V@8
PRPC_HUMAN	Salivary acidic proline-rich phosphoprotein 1/2	3	GRPQGPPQQGGHQQ GPPPPPPGKPQ, cleaved Q-G@C-term
FIBB_HUMAN	Fibrinogen beta chain	1	TPCTVSCNIPVVSGK Oxidation(C)@3; Dehydrated(T)@4
K1C10_HUMAN	Keratin, type I cytoskeletal 10	2	SEITELR
TAGL2_HUMAN	Transgelin-2	1	TLMNLGGLAVAR Oxidation(M)@3
K2C4_HUMAN	Keratin, type II cytoskeletal 4	9	DVDAAYLNKVELEA K Deamidated(N)@8, missed K-V@9
PRDX6_HUMAN	Peroxiredoxin-6	1	VVFVFGPDK
ALBU_HUMAN	Serum albumin	1	SLHTLFGDK
B9EKV4_HUMAN	Aldehyde dehydrogenase 9 family, member A1	2	VEPADASGTEK

Table 4 Legend: RIPPER ILFQ candidates

Protein candidates by Proximity based Intensity Normalization (PIN)/RIPPER analysis. These are proteins that had at least two peptides that were tracked by RIPPER that showed significant differences between healthy and OPML samples. Identification of peptides and proteins relied on Protein Pilot™ and Genedata Analyst™ was used to find pairwise statistical significance.

Figure 9

a


Protein Pilot L/C
RIPPER paired effect size
Ratio L/C
0
1
2
3
4
5
AHNAK
CYTS
RL27A
PRDX6
K1C10
TAGL2
K2C4
SPRR3
ALDH9A1

b

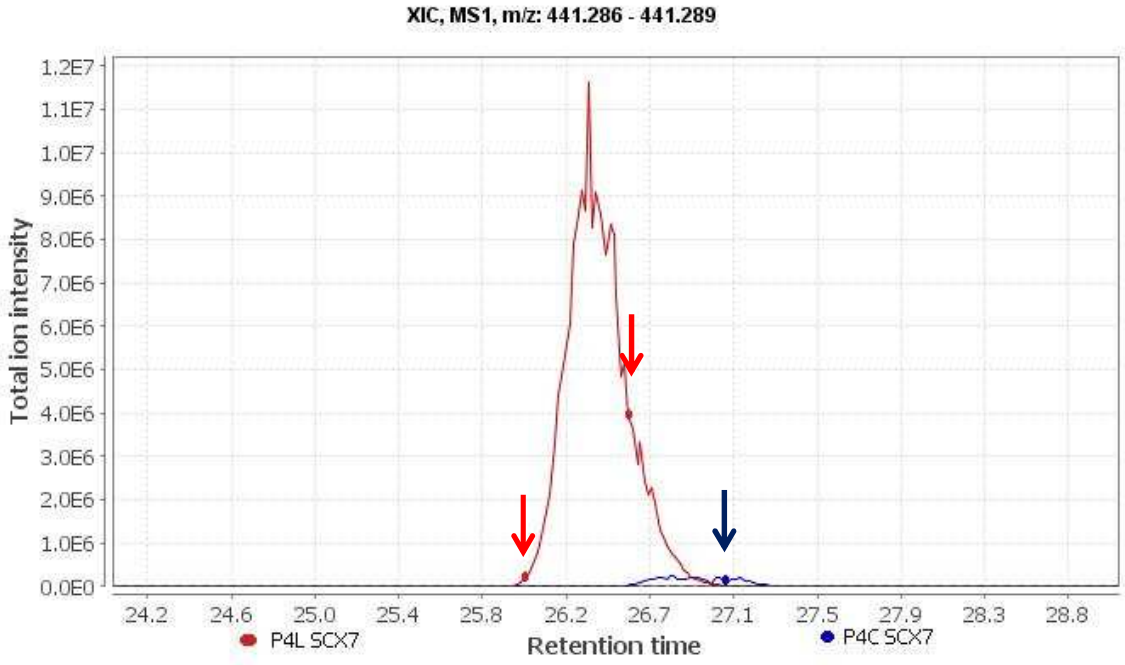

XIC, MS1, m/z: 441.286 - 441.289
Total ion intensity
1.2E7
1.1E7
1.0E7
9.0E6
8.0E6
7.0E6
6.0E6
5.0E6
4.0E6
3.0E6
2.0E6
1.0E6
0.0E0
24.2
24.6
25.0
25.4
25.8
26.2
26.7
27.1
27.5
27.9
28.3
28.8
Retention time
P4L SCX7
P4C SCX7

Figure 9 legend: ILFQ and spectral counting divergence

9a Spectral counting results of RIPPER candidates; Candidates from RIPPER analysis tend to those with multiple identifications, but without clear abundance differences in spectral counts.

9b: Divergence of spectral counts and MS intensity possible with a spectral counting analysis. Extracted Ion Chromatograms for peptide RPLQVLR from protein Cystatin-S are compared. In the run from Patient 4 lesion sample, SCX fraction 7: there are two identifications, and for Patient 4 control sample SCX 7 there is one Identification. However the ratio of total intensity (area under the curve) >>2.

Information about the peptides tracked in Figure 9a is shown in Table 5

Table 5

Protein	peptide	m/z	Analyst paired effect p-value
AHNAK	FSMPGFK Oxidation(M)@3	415.1996	0.03572
CYTS	RPLQVLR	441.2874	0.0195
RL27A	TGAAPIIDVVR	556.3316	0.04803
PRDX6	VVFVFGPDK	504.2837	0.02403
K1C10	SEITELR	424.228	0.00706
TAGL2	TLMNLGGLAVAR Oxidation(M)@3	616.3401	0.02438
K2C4	DVDAAYLNKVELEAK Deamidated(N)@8, missed K-V@9	839.9333	0.02487
SPRR3	VPEQGYTKVPVPGYTK, missed K-V@8	588.3182	0.03077
ALDH9A1	VEPADASGTEK	552.2644	0.01956

Table 5 Legend**:** Spectral counting results of RIPPER candidates

RIPPER tracked peptides with m/z values and paired affect p-values for each of the protein comparisons in Figure 9a.

Table 6

Protein ID	Peak Count	Peptide	Mass	Score	Chromatogram Name	Scan ID
		VTIASLPR				
				21.47		3979
LRMP_HUMAN	3	VDLGALLR	855.517792	32.27	P2L3 P3L3	3790
						2502
						4140
		FLENEDR		39.27		3616
				29.83		4127
		DTEEEDFH-		34.51		4203
		VDQVTTVK		100.81		4204
				51.94		4216
		FLEDVKK		56.62		2537
				70.05		3035
		FLENEDRR		31.67		2336
			921.419189	42.47	P3L3 P3L4	1684
		TDTSHHDQ-	1890.848343	30.15	P2L4 P5C4	5019
		DHPTFNK	877.490906	23.68 31.5	P6C4 P5L4	5412
			1077.520294	45.86	P5L5 P6L5	2805
		LQHLENEL-	1778.760864	31.55	P3L6 P6L6	3149
A1AT_HUMAN	19	THDIITK	1802.952637	27.31 22.4	P5L6 P5C6	2412
				57.75		3909
				44.99		3885
COX5A_HUMAN	2	LNDFASTVR	1021.519257	52.94 48.58	PIL3 P2C3 P2L3 P6L3	3899 4319
SULF2_HUMAN	2	KWPEMK	833.410538	25.18	PIL6	3856

Table 6 Legend: Genedata Refiner MS/Analyst method output

Resulting significant proteins by Genedata Refiner MS and Analyst processing of the raw data. This data justified continuing to consider Alpha-1-antitrypsin as a candidate despite not being flagged by RIPPER (also was a candidate by Scaffold analysis).

Table 7:

Protein	Protein Pilot-RIPPER-Analyst	MaxQuant-Mascot-Scaffold3	Refiner MS-Analyst	Protein Pilot spectral counts	Quantitative measurement details (lesion/Control)
Cytokeratin 4	Yes	no	no	no	2.2 increased abundance, P = 0.02 (RIPPER)
Cytokeratin 10	Yes	no	no	no	2.0 increased abundance, p = 0.007 (RIPPER)
AHNAK	Yes	no	no	no	2.0 increased abundance, p = 0.04 (RIPPER)
Cystatin-S	Yes	No	no	no	3.3 increase abundance, p = 0.019 (RIPPER)
Salivary acidic proline-rich phospho-protein	Yes	No	no	no	3.0 increased abundance, p = 0.007 (RIPPER)
Aldehyde dehydrogenase 9-A1	Yes	No	no	no	3.0 increased abundance, p = 0.02 (RIPPER)
Alpha 1-antichymotrypsin	No	yes	No	Yes	both SC and TIC from Scaffold significant (19+ lesion IDs from 5/6 samples, 0 control IDs)
Alpha-1-antitrypsin	No	yes	Yes	Yes	Significantly Lesion associated in all spectral counting/TIC approaches, plus top candidate of RefinerMS/Analyst approach

Table 7 Legend: Initial western blotting candidates

Western blotting candidates; these were selected from the RIPPER candidates deemed most amenable to western blotting in saliva, as well as Alpha-1-Antitrypsin, and Alpha-1-Antichymotrypsin which were considered the top usable candidates from other analyses (Both Scaffold and RefinerMS/Analyst). The use of spectral counting as a relative protein quantification technique is abbreviated here as SC due to space concerns

Chapter 3 Discussion

The overall goal of this study was to provide a proof-of-principle demonstration of an approach addressing the challenges of protein biomarker discovery in whole saliva, the optimal fluid for developing non-invasive, cheap and simple clinical tests. The challenges include proteome complexity, biological variation, and the possibility of pursuing misleading candidate biomarkers not associated directly with the site of cancer development.

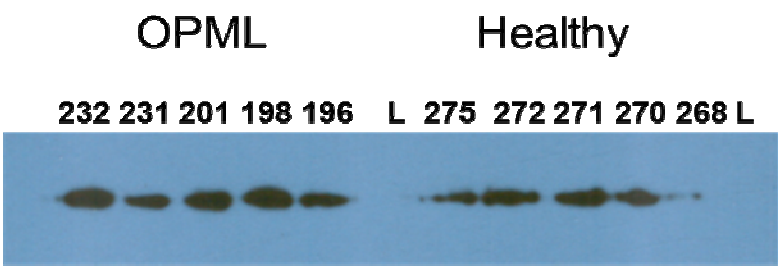
To address this challenge, we first analyzed exudates collected non-invasively from oral lesions. The use of exudates provided an initial quantitative proteomic analysis of samples collected directly from the site of disease. In addition to decreasing the likelihood of candidate biomarkers being proteins that did not originate at the disease site, the exudate also reduced the sample complexity by avoiding endogenous salivary background proteins, which suppress detection of lower abundance components (Bandhakavi et al., 2009). To address the challenge of biological variation between samples that may be unrelated to the disease state, we used paired samples from OPML individuals that were collected from both healthy and diseased tissue. Additionally, exudate samples possess other potential advantages to guide salivary biomarker discovery. The exudate allows an analysis of the population of proteins most likely to be both produced by the oral lesion (or tumor) and shed into saliva, as we are collecting proteins associated with the surface of the lesion in contact with the saliva.

Our comprehensive label-free strategy, employing both spectral counting and ILFQ methods, for analyzing the exudate proteomes led to several interesting observations. Specifically, we noticed that each method performed best on different portions of the MS data. It is interesting to note that while spectral counting seemed to find candidate biomarkers from which there was little MS signal in one sample type and more in the other (low to moderate abundance) the ILFQ analysis told us when abundant peptides produced significantly more signal in MS sample analysis runs from one type of

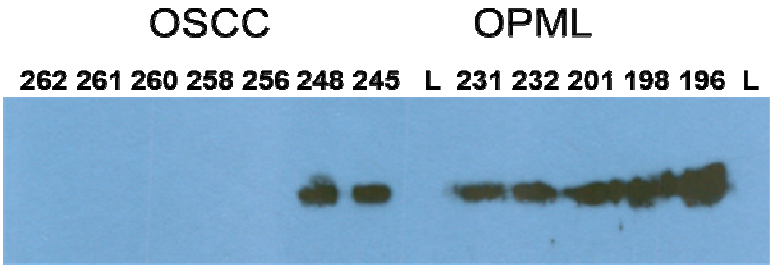
sample than another. This occurred in many places where the number of identifications was very similar between samples being compared. A likely reason for this observation is the dynamic exclusion parameter used in LC-MS/MS based experiments. Here, despite having differential abundance between samples, peptides derived from protein with high absolute abundance may have similarly wide LC peak widths, at a similar or identical multiple of the dynamic exclusion time parameters. These are selected numerous times for MS/MS analysis in both samples, causing a saturation effect when using spectral counting quantification. This effectively decouples the spectral counting measurements from relative abundance (Figure 8a and b). ILFQ via RIPPER however, can effectively quantify these high abundance proteins, as it measures the normalized mass spectral signal intensity which is reflective of relative abundance differences, independent of any MS/MS peptide matching information.

Figure 10: a b c and d

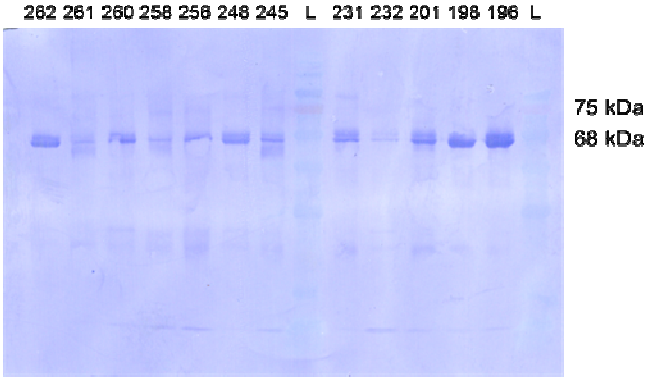
a



b



c



d

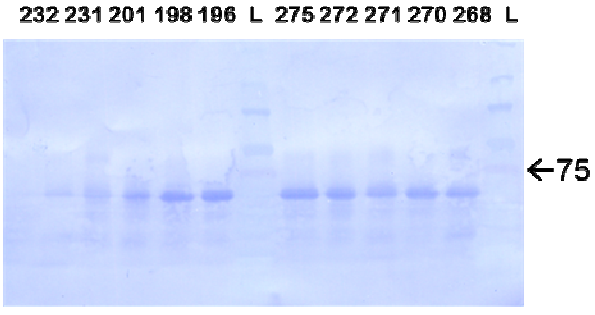
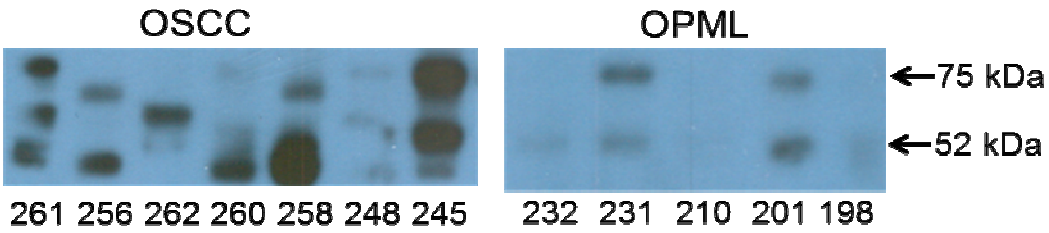


Figure 10 legend: Cytokeratin 10 western blots

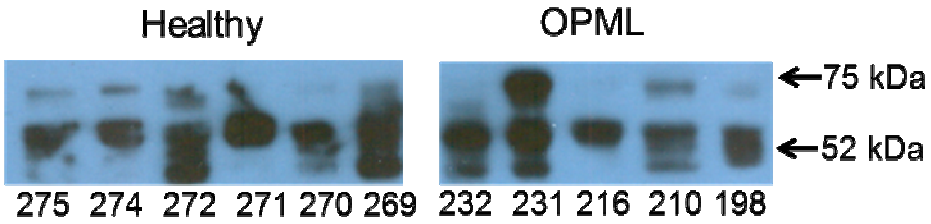
Cytokeratin 10 Western blots: **a** comparison of healthy and OPML results. **b** OSCC and OPML western blotting results. This suggests that Cytokeratin 10 is elevated in OPML, compared to both healthy and OSCC levels. **C** and **d** show the total protein stain of the membranes

Figure 11 a and b:

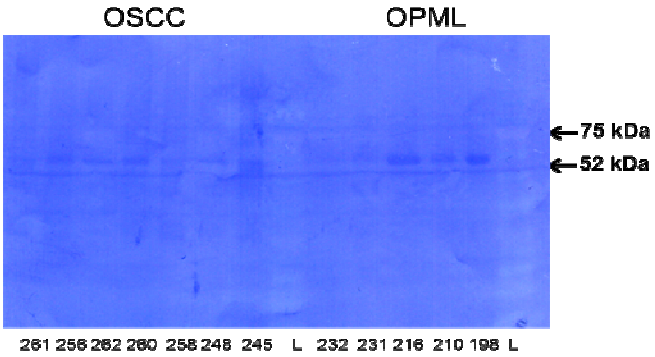
a



b



c



d

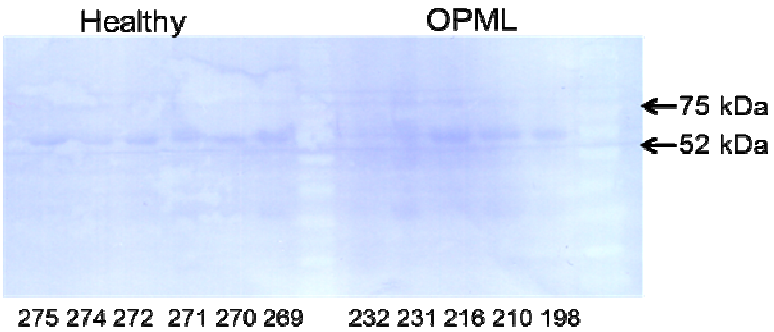


Figure 11 Legend: Western blotting of A1AT

A1AT western blotting Healthy vs. OPML and OPML vs. OSCC. In figure 2A the OSCC to OPML comparison shows considerable variability, but without the intensity of any bands either at 52 kDa, or other mobility shifted bands displaying the intensity that characterizes many OSCC samples. The OSCC/OPML comparison and the Healthy/OPML comparison were from the same blots respectively. Both healthy and OPML western blotting results showed similar abundance levels, with some mobility shifts, but less of either abundance or tendency to display alternate mobilities as the OSCC samples.

Our observations also showed that spectral counting had the most value for those proteins of lower absolute abundance, in scenarios where the protein amount in one sample was low enough such that MS/MS were not matched reliably, while in the other sample being compared there was ample protein amount to lead to numerous confident MS/MS identifications. In some of these cases, ILFQ could not detect differences, because it failed to track the relevant peptide ions at all. Thus our conclusions are that spectral counting and ILFQ can be thought of as complementary approaches, each covering a mostly distinct dynamic range of the proteomes being compared.

When it came to transitioning from quantitative proteomics analysis of exudates to testing for abundance differences of candidates in whole saliva, a number of points were considered. First of all, candidate biomarkers were examined with considerations about previous detection in saliva, including prior studies in our lab (Xie et al., 2008; Bandhakavi et al., 2009; de Jong et al., 2010). Prior literature on possible associations with epithelial cancer development was also considered. Furthermore, in the case of some protein candidates, physiochemical characteristics such as molecular weight was counted against their inclusion (e.g. AHNAK), as extremely large proteins may lead to artifacts in spectral counting quantification, and difficulties in western blotting such proteins. Finally the commercial availability of antibodies was also considered. While we considered RIPPER to be our primary method of selecting candidates due to the low overlap of candidates from the other methods, the proteins of Alpha-1-antitrypsin and alpha-1antichymotrypsin were selected because of their presence in multiple non-RIPPER approach candidate lists (Table 6).

A number of the selected candidates from western blotting failed to produce quantitative results in whole saliva consistent with the exudate samples. These are not necessarily unsurprising results, for a number of reasons. One obvious reason is that proteins in the exudates are not shed into the saliva in enough quantity for their reliable detection. For those proteins detectable in saliva but failing to show any differential abundance, it may be that there are other sources of these proteins in the saliva which obscure any differences contributed by the oral lesion.

Our results did show two proteins displaying differences in whole saliva revealed to us via the exudate analysis. One of these was CK10, which showed a trend of increased abundance in saliva from healthy controls to OPML patients, followed by a dramatic decrease in abundance between OPML and OSCC patients. CK10 is a member of the keratin protein family, involved in the cytoskeletal architecture of oral epithelial cells. Interestingly, numerous studies have shown that CK10 expression is lost in epithelial tumors, suggesting a role in inhibiting cell proliferation and suppressing tumor formation (Santos et al., 1997; Santos et al., 2002; Santos et al., 2005). Our observation of decreased abundance of CK10 in the saliva of OSCC patients is consistent with the past findings.

The second protein of note was A1AT. The detection of bands at both higher and lower molecular weights in the OSCC compared to OPML suggests differential post-translation modification of this protein, possibly via the addition of covalent modifications and/or proteolytic cleavage. Although further experimentation is needed to conclusively understand the cause of this differential migration pattern, our results are consistent with some past studies. A1AT is an acute phase protein, expressed in inflammatory tissue which inhibits serine protease activity. Differential modification via glycoslyation in tumor cells has been suggested (Semaan et al., 2012). Additionally, cleavage of A1AT at the c-terminus has been demonstrated(Zhou et al., 2010).

In conclusion we have demonstrated a proof-of-principle of using quantitative proteomic analysis of oral lesion exudates to guide the discovery of potential salivary biomarkers of oral cancer. In future studies, this method could be applied more directly to the search for salivary biomarkers that can distinguish OPML from OSCC patients by analyzing directly exudates from OSCC patients via quantitative proteomics. Also, more directed analyses of post-translational modifications, could inform both biomarker discovery and the basic biology of the OPML to OSCC transition. Longitudinal proteomic studies of exudates from individuals before and after transition from OPML to OSCC could be invaluable, but unfortunately the availability of these clinical samples is understandably very limited. Metaproteomic and proteogenomic approaches already applied to these sample types (Jagtap et al., 2013), could be expanded and improved to

investigate microbial contributions or the possibly valuable presence of alternative splicing products. Finally, an analysis that featured a larger number of patient samples would be highly valuable to further confirm our interesting findings for CK10 and A1AT, and provide statistical rigor in evaluating these as potential biomarkers with clinical utility.

Chapter 4: Conclusions and Future Directions

Summary and impact of results

Given the low 5 year survival rate of individuals with OSCC, driven by the major discrepancy between early and late diagnosis in terms of survival, and a lack of frequent follow-up monitoring of OPML due to the expense and invasiveness of current diagnostic methods, a simple, non-invasive diagnostic test holds great potential to improve outcomes. In my thesis project I have addressed several of the challenges inherent in the development of such a test by profiling protein changes that define the disease state, and testing the relevance of selected proteins as surrogate salivary biomarkers.

In order to address the need for potential biomarkers of OSCC that are easily and non-invasively available our interdisciplinary research team has identified tissue exudates as an easily accessible source of proteins with biomarker potential. In my work, I have developed sample preparation techniques for quantitative MS-based proteomic analysis of exudates collected in the clinic via absorbent PerioPaper strips, characterized the nature of the oral exudate proteome, and examined the differences between healthy and OPML exudates. Within this work I have evaluated and applied several label-free quantification methods for proteomics data. In order to follow up these proteomics studies, we have endeavored to test the applicability of OPML-associated proteins as surrogate biomarkers distinguishing the OPML to OSCC transition in whole saliva.

In chapter 2 we addressed whether PerioPaper collected exudate samples can be analyzed using MS-based proteomics to generate large-scale protein identification data and to compare different tissue types. The on-strip tryptic digestion method yielded peptide amounts necessary for reliable MS analysis as compared to methods attempting to extract intact proteins prior to proteolysis. In addition, the basic nature of the exudate proteome was characterized. We found that the exudate fluid was distinct from whole saliva, and shared many similarities with brush biopsy collected cellular proteomes. Of particular interest was the greatly decreased relative abundance of major salivary proteins (e.g. amylase) which dominate the saliva proteome and make detection of lower-

abundance proteins difficult for salivary proteomics. Without these proteins present in exudate samples, a wide variety of other potentially interesting targets can be identified more easily. Furthermore, the analysis of proteins in exudates collected directly from the oral lesion immediately increases their value as potential biomarkers of oral cancer development. Finally, we demonstrated the ability to quantify differences in exudate proteins between different tissue types (healthy vs. OPML) using spectral counting label-free quantification, critical if this approach is going to be useful for biomarker discovery.

While providing a valuable demonstration of proof-of-principle, the results in Chapter 2 had a number of shortcomings. One was the use of non-paired tissue samples from healthy individuals and different individuals with OPML, which introduces potential for biological variation between subjects to confound clinical differences between tissue types. Also, the spectral counting based label-free approach was a potential source of uncertainty in our quantitative results, as this method has limitations in accuracy of its measurements. Finally, our results in Chapter 2 did not evaluate whether lesion associated proteins could be detected in whole saliva, and used as possible biomarkers in this preferred clinical sample.

In Chapter 3 we addressed the shortcomings of Chapter 2 with improved clinical sampling (collecting paired healthy and lesion tissue exudates from the same subject for comparison), and considerable work on improvements to our label-free quantification method, employing ILFQ methods in addition to spectral counting methods. We also evaluated the relative abundance of selected proteins in whole saliva from healthy, OPML and OSCC subjects. Notably, we were able to confirm corresponding relative abundance changes in the Cytokeratin 10 protein, and observed an intriguing differential signature of the alpha-1-antitrypsin protein. To our knowledge, this is the first demonstration of using tissue-derived proteomic data to guide discovery of candidate biomarkers of oral cancer in whole saliva. Thus, this work has demonstrated a new strategy in the search for non-invasive, salivary biomarkers of oral cancer. In addition to these exciting results, the data derived from these studies was used by our collaborators for a study developing computational methods for metaproteomic and proteogenomic analyses seeking to identify proteins of microbial origin and of novel sequence,

respectively in clinically-relevant datasets. These results were published recently (Jagtap et al., 2013).

Reflections and future directions

Upon completion of my studies, some reflection is warranted on challenges encountered in this work and possible solutions viewed in retrospect, as well as possible future directions. A clear limitation to our research is the number of human subjects samples available for our studies. In the case of any marker that might be applied in a clinical setting, testing its connection to disease state ultimately involves large numbers of human subjects (hundreds to thousands). While a study of the scale needed for such a quick jump to clinical application might not be feasible for a single researcher (or a small lab), a starting study with dozens of well-characterized research subjects would provide more statistically reliable results, as opposed to our results on only a handful of subjects. Thus the results we present can only be considered preliminary at this point. Unfortunately, infrastructure for collecting large numbers of human samples, especially for OPML subjects, using the non-invasive sampling methods described in this thesis work is lacking. If a larger number of samples could be collected, an OSCC exudate proteomic analysis would have been helpful to complement our results from OPML exudates and should be a consideration for future research.

In terms of the instrumentation and computational approaches used, we made considerable improvements over initial studies (Chapter 2), but a retrospective view reveals a number of possibilities for change. First of all, the use of offline fractionation added some challenge to our ILFQ approach. Difficulties were encountered because individual mass spectrometer runs were performed on SCX fractions that were by their nature chemically distinct, from subject to subject, due to slight differences in retention time from sample to sample, differences in total peptide loading amount etc. Because ILFQ is based on having LC-MS runs conducted on samples that are relatively homogenous, complexities introduced by off-line fractionation prior to LC-MS is not optimal for this quantification approach. This was managed by comparing similar

fractions to each other (e.g. SCX fraction 2 from sample A to SCX fraction 2 from sample B). Though care was taken through the use of UV absorbance levels of each SCX fraction to ensure equal loading amounts for LC-MS analysis, ensuring comparable composition between LS-MS runs from different samples was difficult. A possible solution would be to use a single LC-MS run for each digested clinical sample, without the use of offline fractionation. While fractionation tends to increase total sensitivity, the use of longer gradient times, and possibly a more sensitive mass spectrometer (such as emerging instruments like the Q-Exactive from Thermo Fisher) may maintain comprehensive proteome coverage while aiding and simplifying the applicability of label-free quantification methods to the dataset.

Another limitation is the preliminary nature of the blotting results. This is tied to the sample number issue, in addition to time constraints, the labor-intensiveness of immunoblotting and the available quantity of saliva in each sample. Samples collected from a larger number of subjects could help solidify or disprove the status of CK10 as a putative biomarker. Also, the differential migration pattern of A1AT results suggests the possibility of post-translational modifications, proteolytic cleavage, and/or just antibody cross-reactivity or other issues. Blotting after treatment with deglycoslyation could help elucidate if shifts are due to glycoslyation. Testing a number of monoclonal antibodies against different portions of the protein might aid in determining if truncated forms are present, and/or if the antibody tested was problematic. Although possible, given the nature of Western blotting as a protein detection assay, these experiments would have taken substantial amount of time. As an alternative, targeted MS-based approaches could complement or replace blotting in validation of specific markers. Using Selected Reaction Monitoring (SRM) could provide relative quantification while also elucidating which protein forms are present, with higher throughput and increased accuracy in quantitative measurements.

Another challenge developing proteins discovered in exudates into biomarkers in whole saliva is due to the fact that many of these exist primarily in the cellular portion of saliva rather than the soluble supernatant. Since exudate samples include a cellular portion, this became an issue when examining exudate-derived putative biomarkers in

saliva. While it's possible to solubilize and study the cellular portion of saliva, the variable quantity of cells available from individual saliva samples makes these targets harder to compare in an objective fashion. Using gently collected cells, rather than saliva (for example from a brush biopsy) might be a better approach that is able to consistently study some of the putative exudate markers. Alternatively, continuing with exudate samples may be useful, although low amounts of protein obtained from these can become an issue with non MS-based approaches.

While protein or peptide abundance levels can be a source of much biological information, the contribution of variable protein modifications, and variable forms (e.g. truncations, splice isoforms etc.) could be more valuable as biomarkers of oral cancer, given their potential direct role in cancer progression mechanisms. While modifications on peptides can and were detected by our approach via database searching, and the presence or absence of portions of a protein sequence can be ascertained, more targeted approaches are required to obtain this information more comprehensively. Targeted bottom up MS approaches, or targeted top-down (intact protein) MS could provide this information. Intact protein MS can produce insight into the forms of proteins present in a sample and provide an estimate of their stoichiometry, in a way that bottom-up proteomics cannot. Targeted bottom up approaches could be as simple as directing the mass spectrometer to search for all likely peptides from a protein and detecting possible sequence variants. Alternatively, specific post-translational modifications could be targeted by selective enrichment during sample preparation, although such an approach could compromise the ability to estimate the proportion modified. These methods also rely on relatively large amounts of starting material, which may not be feasible in samples collected in the clinic.

A systems biology-based approach, coupling with other 'omics' approaches, such as an mRNA transciptomics study, might help reinforce some biomarkers, or shed light on cases where transcript and protein abundance diverge, suggesting post-transcriptional regulation. Such information could both improve our knowledge of markers, or highlight what types of cellular signaling are behind proteomic/transcriptomic changes in

metastasis or pre-cancerous inflammation. The connection with mechanism could help in the selection of the most valuable candidate biomarkers.

My studies have contributed towards an ultimate goal: the translation of non-invasively collected biomarkers into the clinic, where they can improve oral cancer diagnosis and management. The next step towards this goal would be transition from preliminary findings presented here, to well-validated biomarkers. Validation will take a concerted effort of recruiting the large number of subjects necessary, and putting in place the assays to rigorously evaluate statistical power of biomarker candidates. For those markers passing this evaluation, non-invasively collected biomarkers, such as those described in this thesis (e.g. salivary proteins), would be well-positioned for development of clinical point-of-care devices. Such devices would finally enable cheaper and more frequent diagnostic testing for oral cancer, helping to decrease the suffering and death from disease.

Bibliography:

(2007). Dorand's Medical Dictionary. Saunders.
ACS (2013). Cancer Facts & Figures 2013. Atlanta: American Cancer Society.
Aebersold, R., Hood, L. E., & Watts, J. D. (2000). Equipping scientists for the new biology. *Nature Biotechnology*, 18(4), 359-359.
Altekruse, S., Kosary, C., Krapcho, M., Neyman, N., Aminou, R., Waldron, W., Ruhl, J., Howlader, N., Tatalovich, Z., Cho, H., Mariotto, A., Eisner, M., Lewis, D., Cronin, K., Chen, H., Feuer, E., Stinchcomb, D., & BK, E. (2010). SEER Cancer Statistics Review, 1975-2007. Bethesda, MD.
Axell, T., Pindborg, J. J., Smith, C. J., & vanderWaal, I. (1996). Oral white lesions with special reference to precancerous and tobacco related lesions: Conclusions of an international symposium held in Uppsala, Sweden, May 18-21 1994. *Journal of Oral Pathology & Medicine*, 25(2), 49-54.
Bandhakavi, S., Stone, M. D., Onsongo, G., Van Riper, S. K., & Griffin, T. J. (2009). A Dynamic Range Compression and Three-Dimensional Peptide Fractionation Analysis Platform Expands Proteome Coverage and the Diagnostic Potential of Whole Saliva. *Journal of Proteome Research*, 8(12), 5590-5600.
Becker, G. W. (2008). Stable isotopic labeling of proteins for quantitative proteomic applications. *Briefings in functional genomics & proteomics*, 7(5), 371-82.
Blumberg, H., Dinh, H., Trueblood, E. S., Pretorius, J., Kugler, D., Weng, N., Kanaly, S. T., Towne, J. E., Willis, C. R., Kuechle, M. K., Sims, J. E., & Peschon, J. J. (2007). Opposing activities of two novel members of the IL-1 ligand family regulate skin inflammation. *Journal of Experimental Medicine*, 204(11), 2603-2614.
Bolignano, D., Donato, V., Lacquaniti, A., Fazio, M. R., Bono, C., Coppolino, G., & Buemi, M. (2010). Neutrophil gelatinase-associated lipocalin (NGAL) in human neoplasias: A new protein enters the scene. *Cancer Letters*, 288(1), 10-16.
Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W. J., Webb-Robertson, B. J. M., Smith, R. D., & Lipton, M. S. (2006). Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of Proteome Research*, 5(2), 277-286.
Carvalho, A. L., Jeronimo, C., Kim, M. M., Henrique, R., Zhang, Z., Hoque, M. O., Chang, S., Brait, M., Nayak, C. S., Jiang, W. W., Claybourne, Q., Tokumaru, Y., Lee, J., Goldenberg, D., Garrett-Mayer, E., Goodman, S., Moon, C. S., Koch, W., Westra, W. H., Sidransky, D., & Califano, J. A. (2008). Evaluation of promoter hypermethylation detection in body fluids as a Screening/Diagnosis tool for head and neck squamous cell carcinoma. *Clinical Cancer Research*, 14(1), 97-107.
Chen, E. I., & Yates, J. R. (2007). Cancer proteomics by quantitative shotgun proteomics. *Molecular Oncology*, 1(2), 144-159.
Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), 1367-1372.
Cox, J., Matic, I., Hilger, M., Nagaraj, N., Selbach, M., Olsen, J. V., & Mann, M. (2009). A practical guide to the MaxQuant computational

platform for SILAC-based quantitative proteomics. *Nature Protocols*, 4(5), 698-705.
Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F., Burkhard, P. R., & Sanchez, J.-C. (2008). Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Analytical Chemistry*, 80(8), 2921-2931.
de Jong, E. P., van Riper, S. K., Koopmeiners, J. S., Carlis, J. V., & Griffin, T. J. (2011). Sample collection and handling considerations for peptidomic studies in whole saliva; implications for biomarker discovery. *Clinica Chimica Acta*, 412(23-24), 2284-2288.
de Jong, E. P., Xie, H. W., Onsongo, G., Stone, M. D., Chen, X. B., Kooren, J. A., Refsland, E. W., Griffin, R. J., Ondrey, F. G., Wu, B. L., Le, C. T., Rhodus, N. L., Carlis, J. V., & Griffin, T. J. (2010). Quantitative Proteomics Reveals Myosin and Actin as Promising Saliva Biomarkers for Distinguishing Pre-Malignant and Malignant Oral Lesions. *Plos One*, 5(6).
Deterding, L. J., Moseley, M. A., Tomer, K. B., & Jorgensen, J. W. (1991). NANOSCALE SEPARATIONS COMBINED WITH TANDEM MASS-SPECTROMETRY. *Journal of Chromatography*, 554(1-2), 73-82.
Driemel, O., Murzik, U., Escher, N., Melle, C., Bleul, A., Dahse, R., Reichert, T. E., Ernst, G., & von Eggeling, F. (2007). Protein profiling of oral brush biopsies: S100A8 and S100A9 can differentiate between normal, premalignant, and tumor cells. *Proteomics Clinical Applications*, 1(5), 486-493.
Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). AN APPROACH TO CORRELATE TANDEM MASS-SPECTRAL DATA OF PEPTIDES WITH AMINO-ACID-SEQUENCES IN A PROTEIN DATABASE. *Journal of the American Society for Mass Spectrometry*, 5(11), 976-989.
Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). ELECTROSPRAY IONIZATION FOR MASS-SPECTROMETRY OF LARGE BIOMOLECULES. *Science*, 246(4926), 64-71.
Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C., & Yates, J. R. (2000). Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Analytical Chemistry*, 72(4), 757-763.
Gevaert, K., Impens, F., Ghesquiere, B., Van Damme, P., Lambrechts, A., & Vandekerckhove, J. (2008). Stable isotopic labeling in proteomics. *Proteomics*, 8(23-24), 4873-4885.
Grant, M. M., Brock, G. R., Matthews, J. B., & Chapple, I. L. C. (2010). Crevicular fluid glutathione levels in periodontitis and the effect of non-surgical therapy. *Journal of Clinical Periodontology*, 37(1), 17-23.
Greenspan, D., & Jordan, R. C. K. (2004). The white lesion that kills - aneuploid dysplastic oral leukoplakia. *New England Journal of Medicine*, 350(14), 1382-1384.
Guo, T., Rudnick, P. A., Wang, W. J., Lee, C. S., Devoe, D. L., & Balgley, B. M. (2006). Characterization of the human salivary proteome by capillary isoelectric focusing/nanoreversed-phase liquid chromatography coupled with ESI-tandem MS. *Journal of Proteome Research*, 5(6), 1469-1478.
Gygi, S. P., Rist, B., Griffin, T. J., Eng, J., & Aebersold, R. (2002). Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags. *Journal of Proteome Research*, 1(1), 47-54.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17(10), 994-999.
He, Y., Chen, Q. M., & Li, B. Q. (2008). ATM in oral carcinogenesis: association with clinicopathological features. *Journal of Cancer Research and Clinical Oncology*, 134(9), 1013-1020.
Hofman, L. F. (2001). Human saliva as a diagnostic specimen. *Journal of Nutrition*, 131(5), 1621S-1625S.
Hu, S., Arellano, M., Boontheung, P., Wang, J. H., Zhou, H., Jiang, J., Elashoff, D., Wei, R., Loo, J. A., & Wong, D. T. (2008). Salivary Proteomics for Oral Cancer Biomarker Discovery. *Clinical Cancer Research*, 14(19), 6246-6252.
Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., & Hauer, C. R. (1986). PROTEIN SEQUENCING BY TANDEM MASS-SPECTROMETRY. *Proceedings of the National Academy of Sciences of the United States of America*, 83(17), 6233-6237.
Ichii, O., Otsuka, S., Sasaki, N., Yabuki, A., Ohta, H., Takiguchi, M., Hashimoto, Y., Endoh, D., & Kon, Y. (2010). Local overexpression of interleukin-1 family, member 6 relates to the development of tubulointerstitial lesions. *Laboratory Investigation*, 90(3), 459-475.
Jagtap, P., Goslinga, J., Kooren, J., McGowan, T., Wrobleski, M., Seymour, S., & Griffin, T. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies.
Johnson, R. B., Streckfus, C. F., Dai, X., & Tucci, M. A. (1999). Protein recovery from several paper types used to collect gingival crevicular fluid. *Journal of Periodontal Research*, 34(6), 283-289.
Jou, Y. J., Lin, C. D., Lai, C. H., Chen, C. H., Kao, J. Y., Chen, S. Y., Tsai, M. H., Huang, S. H., & Lin, C. W. (2010). Proteomic identification of salivary transferrin as a biomarker for early detection of oral cancer. *Analytica Chimica Acta*, 681(1-2), 41-48.
Kaufman, E., & Lamster, I. B. (2002). The diagnostic applications of saliva - A review. *Critical Reviews in Oral Biology & Medicine*, 13(2), 197-212.
Klasner, S. A., Price, A. K., Hoeman, K. W., Wilson, R. S., Bell, K. J., & Culbertson, C. T. (2010). Paper-based microfluidic devices for analysis of clinically relevant analytes present in urine and saliva. *Analytical and Bioanalytical Chemistry*, 397(5), 1821-1829.
Kooren, J. A., Rhodus, N. L., Tang, C., Jagtap, P. D., Horrigan, B. J., & Griffin, T. J. (2011). Evaluating the potential of a novel oral lesion exudate collection method coupled with mass spectrometry-based proteomics for oral cancer biomarker discovery. *Clinical proteomics*, 8, 13-13.
Lee, J. M., Garon, E., & Wong, D. T. (2009). Salivary diagnostics. *Orthodontics & Craniofacial Research*, 12(3), 206-211.
Lingen, M. W., Kalmar, J. R., Karrison, T., & Speight, P. M. (2008). Critical evaluation of diagnostic aids for the detection of oral cancer. *Oral Oncology*, 44(1), 10-22.
Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., & Yates, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, 17(7), 676-682.

Lippman, S. M., Sudbo, J., & Hong, W. K. (2005). Oral cancer prevention and the evolution of molecular-targeted drug development. *Journal of Clinical Oncology*, 23(2), 346-356.
Lumerman, H., Freedman, P., & Kerpel, S. (1995). ORAL EPITHELIAL DYSPLASIA AND THE DEVELOPMENT OF INVASIVE SQUAMOUS-CELL CARCINOMA. *Oral Surgery Oral Medicine Oral Pathology Oral Radiology and Endodontics*, 79(3), 321-329.
Lundgren, D. H., Hwang, S. I., Wu, L. F., & Han, D. K. (2010). Role of spectral counting in quantitative proteomics. *Expert Review of Proteomics*, 7(1), 39-53.
Marur, S., D'Souza, G., Westra, W. H., & Forastiere, A. A. (2010). HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncology*, 11(8), 781-789.
Mehrotra, R., Hullmann, M., Smeets, R., Reichert, T. E., & Driemel, O. (2009). Oral cytology revisited. *Journal of Oral Pathology & Medicine*, 38(2), 161-166.
Menon, R., Zhang, Q., Zhang, Y., Fermin, D., Bardeesy, N., DePinho, R. A., Lu, C., Hanash, S. M., Omenn, G. S., & States, D. J. (2009). Identification of Novel Alternative Splice Isoforms of Circulating Proteins in a Mouse Model of Human Pancreatic Cancer. *Cancer Research*, 69(1), 300-309.
Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., Mariani, M., Assadourian, G., Lee, A., van Sluyter, S. C., & Haynes, P. A. (2011). Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics*, 11(4), 535-553.
Nielsen, B. S., Borregaard, N., Bundgaard, J. R., Timshel, S., Sehested, M., & Kjeldsen, L. (1996). Induction of NGAL synthesis in epithelial cells of human colorectal neoplasia and inflammatory bowel diseases. *Gut*, 38(3), 414-420.
Oda, Y., Huang, K., Cross, F. R., Cowburn, D., & Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12), 6591-6596.
Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1(5), 376-386.
Park, N. J., Zhou, H., Elashoff, D., Henson, B. S., Kastratovic, D. A., Abemayor, E., & Wong, D. T. (2009). Salivary microRNA: Discovery, Characterization, and Clinical Utility for Oral Cancer Detection. *Clinical Cancer Research*, 15(17), 5473-5477.
Peng, J. M., Elias, J. E., Thoreen, C. C., Licklider, L. J., & Gygi, S. P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *Journal of Proteome Research*, 2(1), 43-50.
Pentenero, M., Carrozzo, M., Pagano, M., Galliano, D., Broccoletti, R., Scully, C., & Gandolfo, S. (2003). Oral mucosal dysplastic lesions and early squamous cell carcinomas: underdiagnosis from incisional biopsy. *Oral Diseases*, 9(2), 68-72.
Perkins, D. N., Pappin, D. J. C., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18), 3551-3567.

Ralhan, R., DeSouza, L. V., Matta, A., Tripathi, S. C., Ghanny, S., Gupta, S. D., Bahadur, S., & Siu, K. W. M. (2008a). Discovery and verification of head-and-neck pre-cancer and cancer biomarkers by differential protein expression analysis using iTRAQ-labeling and multidimensional liquid chromatography and tandem mass spectrometry. *Cancer Biomarkers*, 4(3), 158-158.
Ralhan, R., Desouza, L. V., Matta, A., Tripathi, S. C., Ghanny, S., Gupta, S. D., Bahadur, S., & Siu, K. W. M. (2008b). Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. *Molecular & Cellular Proteomics*, 7(6), 1162-1173.
Ralhan, R., DeSouza, L. V., Matta, A., Tripathi, S. C., Ghanny, S., DattaGupta, S., Thakar, A., Chauhan, S. S., & Siu, K. W. M. (2009). iTRAQ-Multidimensional Liquid Chromatography and Tandem Mass Spectrometry-Based Identification of Potential Biomarkers of Oral Epithelial Dysplasia and Novel Networks between Inflammation and Premalignancy. *Journal of Proteome Research*, 8(1), 300-309.
Raynaud, C. M., Jang, S. J., Nuciforo, P., Lantuejoul, S., Brambilla, E., Mounier, N., Olaussen, K. A., Andre, F., Morat, L., Sabatier, L., & Soria, J. C. (2008). Telomere shortening is correlated with the DNA damage response and telomeric protein down-regulation in colorectal preneoplastic lesions. *Annals of Oncology*, 19(11), 1875-1881.
Reshmi, S. C., & Gollin, S. M. (2005). Chromosomal instability in oral cancer cells. *Journal of Dental Research*, 84(2), 107-117.
Reynolds, K. J., & Fenselau, C. (2004). Quantitative protein analysis using proteolytic 18O water labeling. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, Chapter 23, Unit 23.4-Unit 23.4.
Rhodus, N. L. (2005). Oral cancer: leukoplakia and squamous cell carcinoma. *Dent Clin North Am*, 49(1), 143-65, ix.
Rhodus, N. L. (2009). Oral cancer and precancer: improving outcomes. *Compend Contin Educ Dent*, 30(8), 486-8, 490-4, 496-8 passim; quiz 504, 520.
Rhodus, N. L., Ho, V., Miller, C. S., Myers, S., & Ondrey, F. (2004). NF-kappa B dependent cytokine levels in saliva of patients with oral preneoplastic lesions and oral squamous cell carcinoma. *Cancer Detection and Prevention*, 29(1), 42-45.
Rhodus, N. L., Cheng, B., Myers, S., Miller, L., Ho, V., & Ondrey, F. (2005). The feasibility of monitoring NK-kappa B associated cytokines: TNF=alpha, IL-alpha, IL-6, and IL-8 in whole saliva for the malignant transformation of oral lichen planus. *Molecular Carcinogenesis*, 44(2), 77-82.
Ross, P. L., Huang, Y. L. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., & Pappin, D. J. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12), 1154-1169.
Rudney, J. D., Xie, H., Rhodus, N. L., Ondrey, F. G., & Griffin, T. J. (2010). A metaproteomic analysis of the human salivary microbiota by three-dimensional peptide fractionation and tandem mass spectrometry. *Molecular Oral Microbiology*, 25(1), 38-49.

Rushworth, S. A., MacEwan, D. J., & O'Connell, M. A. (2008). Lipopolysaccharide-Induced Expression of NAD(P)H:Quinone Oxidoreductase 1 and Heme Oxygenase-1 Protects against Excessive Inflammatory Responses in Human Monocytes. *Journal of Immunology*, 181(10), 6730-6737.
Santos, M., Ballestin, C., GarciaMartin, R., & Jorcano, J. L. (1997). Delays in malignant tumor development in transgenic mice by forced epidermal keratin 10 expression in mouse skin carcinomas. *Molecular Carcinogenesis*, 20(1), 3-9.
Santos, M., Paramio, J. M., Bravo, A., Ramirez, A., & Jorcano, J. L. (2002). The expression of keratin K10 in the basal layer of the epidermis inhibits cell proliferation and prevents skin tumorigenesis. *Journal of Biological Chemistry*, 277(21), 19122-19130.
Santos, M., Rio, P., Ruiz, S., Martinez-Palacio, J., Segrelles, C., Lara, A. F., Segovia, J. C., & Paramio, J. A. (2005). Altered T cell differentiation and notch signaling induced by the ectopic expression of keratin K10 in the epithelial cells of the thymus. *Journal of Cellular Biochemistry*, 95(3), 543-558.
Sato, H., Uzawa, N., Takahashi, K. I., Myo, K., Ohyama, Y., & Amagasa, T. (2010). Prognostic utility of chromosomal instability detected by fluorescence in situ hybridization in fine-needle aspirates from oral squamous cell carcinomas. *Bmc Cancer*, 10.
Schaaij-Visser, T. B. M., Brakenhoff, R. H., Leemans, C. R., Heck, A. J. R., & Slijper, M. (2010). Protein biomarker discovery for head and neck cancer. *Journal of Proteomics*, 73(10), 1790-1803.
Semaan, S. M., Wang, X., Marshall, A. G., & Sang, Q.-X. A. (2012). Identification of Potential Glycoprotein Biomarkers in Estrogen Receptor Positive (ER+) and Negative (ER-) Human Breast Cancer Tissues by LC-LTQ/FT-ICR Mass Spectrometry. *Journal of Cancer*, 3, 269-84.
Shintani, S., Hamakawa, H., Ueyama, Y., Hatori, M., & Toyoshima, T. (2010). Identification of a truncated cystatin SA-I as a saliva biomarker for oral squamous cell carcinoma using the SELDI ProteinChip platform. *International Journal of Oral and Maxillofacial Surgery*, 39(1), 68-74.
Silverman, S. (2001). Demographics and occurrence of oral and pharyngeal cancers - The outcomes, the trends, the challenge. *Journal of the American Dental Association*, 132, 7S-11S.
Silverman, S., Gorsky, M., & Lozada, F. (1984). ORAL LEUKOPLAKIA AND MALIGNANT TRANSFORMATION - A FOLLOW-UP-STUDY OF 257 PATIENTS. *Cancer*, 53(3), 563-568.
Stastna, M., & Van Eyk, J. E. (2012). Analysis of protein isoforms: Can we do it better? *Proteomics*, 12(19-20), 2937-2948.
Sudbo, J. (2004). Novel management of oral cancer: a paradigm of predictive oncology. *Clin Med Res*, 2(4), 233-42.
Tan, W., Sabet, L., Li, Y., Yu, T., Klokkevold, P. R., Wong, D. T., & Ho, C. M. (2008). Optical protein sensor for detecting cancer markers in saliva. *Biosensors & Bioelectronics*, 24(2), 266-271.
Ting, L., Rad, R., Gygi, S. P., & Haas, W. (2011). MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods*, 8(11), 937-940.
Toyoshima, T., Koch, F., Kaemmerer, P., Vairaktaris, E., Al-Nawas, B., & Wagner, W. (2009). Expression of cytokeratin 17 mRNA in oral squamous cell carcinoma cells obtained by brush biopsy: preliminary results. *Journal of Oral Pathology & Medicine*, 38(6), 530-534.

Viet, C. T., & Schmidt, B. L. (2010). UNDERSTANDING ORAL CANCER IN THE GENOME ERA. *Head and Neck-Journal for the Sciences and Specialties of the Head and Neck*, 32(9), 1246-1268.
Viet, C. T., Jordan, R. C. K., & Schmidt, B. L. (2007). DNA promoter hypermethylation in saliva for the early diagnosis of oral cancer. *J Calif Dent Assoc*, 35(12), 844-9.
Washburn, M. P., Wolters, D., & Yates, J. R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19(3), 242-247.
Wong, D. T. (2006). Towards a simple, saliva-based test for the detection of oral cancer. *Expert Review of Molecular Diagnostics*, 6(3), 267-272.
Xie, H. W., Rhodus, N. L., Griffin, R. J., Carlis, J. V., & Griffin, T. J. (2005). A catalogue of human saliva proteins identified by free flow electrophoresis-based peptide separation and tandem mass spectrometry. *Molecular & Cellular Proteomics*, 4(11), 1826-1830.
Xie, H. W., Onsongo, G., Popko, J., de Jong, E. P., Cao, J., Carlis, J. V., Griffin, R. J., Rhodus, N. L., & 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. *Molecular & Cellular Proteomics*, 7(3), 486-498.
Zhou, J., Trock, B., Tsangaris, T. N., Friedman, N. B., Shapiro, D., Brotzman, M., Chan-Li, Y., Chan, D. W., & Li, J. (2010). A unique proteolytic fragment of alpha1-antitrypsin is elevated in ductal fluid of breast cancer patient. *Breast Cancer Research and Treatment*, 123(1), 73-86.
Zimmermann, B. G., Park, N. J., & Wong, D. T. (2007). Genomic targets in saliva. *Oral-Based Diagnostics*, 1098, 184-191.