Using the genome, transcriptome and proteome to better understand species: from *Homo sapiens* to *Homarus americanus*

A thesis submitted by

Prarthana Khanna

in partial fulfillment of the requirements for the degree of

PhD

in

Genetics

Tufts University

Sackler School of Graduate Biomedical Sciences

February 2018

Advisors: David R. Walt, PhD, Jill L. Maron, MD, MPH

Abstract

The field of genetics encompasses genomics, transcriptomics and proteomics and entails understanding the role of organisms' genes. These three omes along with numerous genetics based tools including next generation sequencing platforms and ultrasensitive protein assays can be utilized to inform and expand our understanding of different species and solve a multitude of problems. Using two different model organisms (*Homo sapiens* and *Homarus americanus*), we utilized genomic, transcriptomic and proteomic data to address the objectives of our projects. For both projects, we used the transcriptome in different ways to address our goals. For the first project, we used the neonatal transcriptome to independently inform us about the proteome. For the second project, we used the American lobster transcriptome as a reference for assembling the *de novo* genome.

The aim of the first project was to determine if we could detect oral feeding readiness from neonatal saliva. A two-tiered approach was taken to address this goal. First, research was conducted to translate a previously identified transcriptomic panel (informative of oral feeding readiness) into a rapid proteomic platform to provide objective, near real-time assessment of oral feeding skills, to better inform care and improve neonatal outcomes. Assays for proteins involved in sensory integration, hunger signaling and facial development were developed. This study provided the foundation for the development of an informative rapid proteomic platform to assess neonatal oral feeding maturation. Simultaneously, an RNA sequencing platform was used to advance the development of feeding assessment tools by reviewing the entire human

transcriptome using a high-throughput and quantitative approach. This work was undertaken to better understand the disrupted developmental patterns in the premature newborn. Moreover, we conducted a supplementary study to identify reliable reference genes to use for normalization of transcriptomic data in neonatal salivary diagnostics. These projects were all conducted in neonatal saliva and helped elucidate the complexities involved in neonatal oral feeding maturation with the help of transcriptomic and proteomic data.

The goal of the second project was to hone my bioinformatics skills and provide a complete American lobster genome and in turn, a foundation for crustacean genetics studies. Among crustaceans, the American lobster is an iconic species that is integral to many marine ecosystems and is an important commercial fishery in the Northwest Atlantic. By using the sequenced transcriptome as a reference for assembling the *de novo* genome we were able to provide the first complete American lobster genome. We annotated approximately 37,000 genes from the genome and 24,000 from the transcriptome. A 70% to 80% concordance between the genome and transcriptome was observed signifying complete sequencing of the genome.

Utilizing the genomic, transcriptomic and proteomic information collectively, we were able to study the smallest of humans to the largest of crustaceans. Not only did we use this approach to address challenges in the clinical setting but also in the commercial fishing industry. Hence, it is important to consider the genome, transcriptome and proteome simultaneously and how they can help better inform each other to address numerous genetic enigmas.

Acknowledgements

This thesis would not have been possible without the support and guidance of many people. I would like to start by thanking my mentors, Drs. David Walt and Jill Maron. I would like to thank Dr. Walt for always having my back, giving me the freedom to work on what made me happy and being available to mentor me even when he was halfway across the world. Not only has he has taught me to think outside the box but also made me believe I’m invincible if I try hard enough. I would like to thank Dr. Maron for being a wonderful role-model and always being my knight in shining armor. Not only has she supported my professional development but also my personal development (and taught me how to look good while balancing both aspects of my life). Both my advisors have taught me to be ambitious and reach for the stars and I will be eternally grateful for having such brilliant mentors during my PhD experience.

I would next like to thank my committee members, Drs. Diana Bianchi, Jonathan Garlick and Mitch McVey, for being very generous with their time and expertise, and for helping me stay on track and providing me with much needed guidance and advice. I would like to add a special thank you to Dr. B. for providing me with the inspiration to join Tufts, introducing me to Dr. Maron, and helping me give birth to my PhD project!

Additionally, I would also like to thank Dr. Albert Tai for spending hours teaching me all I know about bioinformatics and always being patient with me even when I repeatedly asked for his help with new projects. And I need to thank Dr. Kirby Johnson for his ethical, statistical and research advice.

I would also like to thank the Maron and Walt lab members for helping me through this entire process. Drs. Austin Rosner, Ruby Bartolome, Faycal Guedj, Tove Lekva, Milena Milutinovic, Stephanie Walter, Shonda Gaylord, Trinh Dinh, Limor Cohen, Lauren Gibson and Shazia Baig for their endless scientific help (experimental and written), snacks, lunches, dinners, drinks, shoulders for crying on, tight hugs, celebrations and most importantly, their friendship. All of you made my PhD experience filled with happiness and love and I will treasure these memories forever. Dr. Austin Rosner gets a special shout-out for spending hours repeatedly looking over my gibberish and helping me edit it.

I also need to thank the Genetics program students for being very supportive and helpful, especially at the beginning of my PhD when I needed it the most. And most importantly, a huge thank-you to Di Pierce for always looking out for me and helping me solve all my grad school problems.

I would not be here today without the love, support and guidance of my family and friends. I need to thank Raine Vollor for being my friend and helping me survive these past five years by being there for me. I am also grateful to my sister-in-law, Sanya Khanna, and my baby brother, Vidhur Sethi, for providing me with your endless love and support, I am very lucky to have you both in my life. And the most important thank you is for the best Mum in the world, Jyoti Sethi. I got so lucky to have not only a perfect mother but a best friend in you too. Words are not enough to thank you for all that you have done for me. Thank you for being such a strong woman and offering me with the greatest role-model to look up to, any goodness I have in me is because of you.

To my Ada and Ronak, thank you for keeping me sane and efficient and making me realize that a bad science day is not the end of the world with your beautiful smiles and cuddles.

And finally, the most vital person in my life, my best friend without whom I would have been unable to function a single day in this PhD, Tarun Khanna. Thank you for picking up my slack as a parent, a partner and a human being when I needed you to. Thank you for the endless cups of coffee and tea, the home-cooked (and restaurant) meals and the alcoholic drinks. Thank you for pushing me when I was being lazy, hugging me when I wanted to quit, listening to me vent about yet another failed experiment and giving me space when I needed it. I could write another thesis with all the things I need to thank you for but needless to say, I am very lucky to have you as the love of my life.

Table of Contents

List of Tables

List of Figures

List of Copyrighted Materials Produced by Author

1. Salivary diagnostics using a portable point-of-service platform: a review
   P Khanna, DR Walt
   Clinical therapeutics 37 (3), 498-504
2. Optimal reference genes for RT-qPCR normalization in the newborn
   P Khanna, KL Johnson, JL Maron
   Biotechnic & Histochemistry, 1-8
3. Development of a rapid salivary proteomic platform for oral feeding readiness in the preterm newborn
   P Khanna, JL Maron, D Walt
   Frontiers in Pediatrics 5, 268

List of Abbreviations

AMPK	adenosine-monophosphate-activated protein kinase
ELISA	Enzyme Linked Immunosorbent Assay
NICU	Neonatal Intensive Care Unit
NPHP4	nephronophthisis 4
NPY2R	Neuropeptide Y2 receptor
PCA	Post Conceptional Age
PLXNA1	Plexin A1
POC	Point of Care
SiMoA	Single Molecule Array
WNT3	wingless-type MMTV integration site family, member 3

Chapter 1: Introduction

1.1 The evolution of this thesis

This introduction is to set the context of this thesis and specific introductions are later provided to better understand the individual aspects of the various projects. My primary PhD project was focused on better understanding the developmental processes involved during oral feeding maturation and developing a rapid proteomic assay to detect this maturation in the neonatal population. Simultaneously, I started working on providing the first complete sequenced, annotated, and characterized genome for *Homarus americanus* to hone my bioinformatics skills as a genetics PhD student. Towards the completion of my PhD, when I was trying to link my projects together I realized that in this thesis, I have used the transcriptome in multiple ways to address my projects' goals. I utilized previously identified transcriptomic information to develop a proteomic platform to determine oral feeding readiness. At the same time, I used the assembled transcriptome of the American lobster as a reference to assemble and annotate the *de novo* genome. I had the opportunity to study the mechanics of the upstream (genome) and downstream (proteome) effects of the transcriptome and was able to study these in two different model organisms, *Homo sapiens* and *Homo americanus*. In turn, I was able to utilize genomic, transcriptomic and proteomic information to tackle both these projects about different species while addressing incredibly different challenges.

1.2 The relationship between the genome, transcriptome and proteome

The genome, transcriptome and proteome all convey a different story about an individual [1]. The genome contains all the coding (1-2%) and non-coding (98-99%) DNA of an organism and comprises all of the information needed to build and maintain the structure and function of that organism [2]. For *Homo sapiens*, the genome is made up of approximately three billion DNA base pairs encoded within the 23 chromosome pairs in cell nuclei. Humans contain approximately 20,000 protein-coding genes [3]. DNA contains the entire collection of genes that remain somewhat constant throughout the lifetime of an organism barring epigenetic effects [4]. Conversely, the transcriptome and proteome provide a window into an organism's ongoing functionality. The transcriptome consists of RNA transcripts of all coding and non-coding regions, however it can also include multiple RNA transcripts for the same gene due to alternative splicing and gene editing [5]. Hence, this adds a layer of dynamic complexity to the transcriptome compared to the genome and enables the transcriptome to inform us about which genes are involved in different pathways [6]. The proteome is even more complex with a multitude of post-translational modifications and molecular interactions, making it a challenge to get an accurate understanding of proteomic expression. In addition, there are many technical difficulties in handling proteins, including the inability to amplify proteins [7]. Hence, the transcriptome provides a bridge between the genome and proteome, rendering a connection between the genetic code and functional proteins [8]. Rather than considering these three -omes independently, it is important to think of them as a whole as they work together and can help inform each other's expression patterns.

1.3 Using the transcriptome to inform us about the proteome

The transcriptome is translated to the proteome, which comprises of all the proteins present in a biological sample at a particular point of time. Unlike DNA or RNA, proteins have a vast array of chemical modifications that directly impact function making proteins enormously more complex than DNA and RNA [9]. Multiple studies have conducted parallel expression level comparisons between the transcriptome and proteome [8, 10, 11]. While some studies have identified a direct correlation between RNA transcripts and generated protein expression levels, others have shown low correlation between their expression patterns [11-14]. We chose to take a novel approach by using the transcriptomic expression levels to independently inform us about the proteome in the newborn population. Previous transcriptomic studies identified five key regulatory genes responsible for oral feeding maturity that were differentially expressed between successful and unsuccessful oral feeders, including *Plexin A1 (PLXNA1), Neuropeptide Y2 receptor (NPY2R), adenosine-monophosphate-activated protein kinase (AMPK), wingless-type MMTV integration site family, member 3 (WNT3),* and *nephronophthisis 4* (*NPHP4*) [15]. Using this transcriptomic information, we hypothesized that the protein expression for these targets would parallel their gene expression. The goal of this study was to develop a rapid proteomic bedside assay platform to assess oral feeding readiness.

Furthermore, we aimed to utilize a next generation sequencing platform, RNA sequencing, to increase the previously discovered transcriptomic biomarker panel's accuracy. This exploratory research targeted not only discovering additional biomarkers that may improve the overall diagnostic accuracy of the existing platform, but also identifying alternative splicing of gene targets to better understand transcriptional regulatory process of the previously identified targets.

Moreover, we endeavored to identify reference gene targets that could serve as valid markers for appropriate normalization of transcriptomic data in the newborn. The Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines dictates the use of multiple reference genes for relative gene expression quantification [16, 17]. To date, no set of ideal reference genes has been established for use in the neonatal population, who present unique challenges due to their rapid and ongoing development (2). Without due diligence to identify stable, constituently expressed reference genes for use in this population, interpretation of gene expression data will remain limited. To our knowledge, this large-scale clinical study was the first to examine gene expression variability of reference gene candidates in neonates across the preterm and term neonatal age spectrum, providing valuable insight in the developing newborn population.

In addition, these numerous projects were all conducted in one particular bio-fluid, neonatal saliva. A comprehensive review of salivary diagnostics was initially performed and published prior to beginning these multiple studies utilizing transcriptomic and proteomic platforms with neonatal salivary samples.

The clinical correlate: the premature newborn

The majority of infants born preterm (< 37 weeks' gestational age) lack the developmental maturity to successfully orally feed [18]. It is estimated that each year in the US, 630,000 newborns are affected by oral feeding complications [19]. Inappropriate feeding attempts can lead to acute and long-term morbidities, as well as prolonged hospitalizations with associated healthcare costs [20, 21]. Determining feeding maturity is a clinical challenge because available feeding assessment tools are subjective, qualitative and not validated [22]. Two Cochrane Reviews confirmed the futility of available feeding assessment tools in 2012 and 2016, concluding there was “no evidence to inform clinical practice” [19, 23]. To address this knowledge gap, we utilized a relatively understudied human bio-fluid (saliva) in a population that is not well understood (neonates) to study the salivary transcriptome in an innovative and clinically relevant manner.

1.4 Using the transcriptome to inform us about the genome

Nucleated cells in an organism contain the same genome but not every gene is transcriptionally active in every cell. This discrepancy allows for different gene expression patterns, resulting in different cell types and functions [24]. While the *Homo sapiens* genome is relatively well characterized, there are many organisms that have still not been fully sequenced [4, 25]. *De novo* sequencing is a way to sequence novel genomes when there is no reference sequence available [26]. We used a novel method for *de novo* sequencing the *Homarus americanus* (American lobster) genome by using its sequenced and assembled transcriptome as a reference for alignment [27].

The de novo species correlate: Homarus americanus

Crustaceans are a heterogeneous set of organisms and include the *Homarus americanus* or American lobster [28]. There is no complete sequenced crustacean genome currently available, making a fully sequenced, annotated and well-defined American lobster genome very attractive to the field of crustacean genomics [29]. Due to the lack of a well-defined crustacean genome, we used the completely sequenced transcriptome of the American lobster as a reference to inform the process of *de novo* sequencing of the *Homarus americanus* genome [30].

Chapter 2: Salivary diagnostics using a portable point-of-service platform: A Review

Salivary diagnostics using a portable point-of-service platform: a review, P Khanna, DR Walt, Clinical therapeutics 37 (3), 498-504

Reprinted here with permission of publisher.

2.1 Introduction

Human saliva is comprised of 99.5% water containing electrolytes, proteins, nucleic acids, peptides, polynucleotides, hormones, enzymes, cytokines, antibodies, and other components [31, 32]. Saliva is primarily secreted from the parotid, the sub-mandibular and the sublingual salivary glands [33]. Saliva also contains serumnal components that are transported from blood capillaries into saliva by diffusion, active transport and/or ultra-filtration via gingival crevices [34]; hence, saliva can beconsidered to be a partial filtrate of blood and can provide a window into the health status of an individual. Saliva is an attractive diagnostic biofluid because its collection is relatively non-invasive, stress-free, inexpensive, and requires minimally trained personnel [35]. Diagnostic methods that use biomarkers from biofluids such as blood and saliva are essential for clinical analyses. Protein and nucleic acid biomarkers can be used to detect medical conditions rapidly, ideally even before the disease presents symptoms in the patient [36]. Many researchers have reported using saliva as a diagnostic fluid. For example, mRNA detection using saliva samples for oral cancer diagnosis was reported by Wong and coworkers [37]. It has also been established in multiple studies that various proteins in saliva correlate with the pathophysiological state of certain medical conditions [38-41].

Laboratory testing of clinical samples can be time-intensive and expensive [42]. Diagnostic testing is moving towards point-of-service (POS) devices due to the rapid results possible from POS devices where early detection is paramount [43]. Multiple POS devices have been developed over the past decade for various diagnostic applications [44-49]. Our laboratory has developed a portable POS device that is capable of automated, multiplexed and sensitive detection of biomarkers present in saliva [49] (Fig. 2.1). This review will describe the benefits and disadvantages of using saliva over blood in a clinical environment, and how a POS device can be used effectively in certain settings. We will briefly describe saliva sample preparation including saliva collection and extraction—a critical aspect of the overall diagnostic process. Next, we discuss the use of protein and nucleic acid biomarkers. We will then describe the principles of sandwich immunoassays, enzyme linked immunosorbent assays (ELISAs), digital ELISAs and microsensor arrays. Finally, we will discuss the multiplexing capabilities of these assays and their applications to salivary diagnostics.

2.2 Sample Preparation

Two essential components of the overall salivary diagnostic process are the saliva collection and processing steps. Saliva collection can be done by three established ways include suctioning, ejecting whole saliva and swabbing [50]. Saliva can be collected under resting or 'unstimulated' conditions or can be 'stimulated' by varied methods, with gustatory and masticatory stimuli being the most common [51]. The method utilizing ejection of whole saliva is recommended for collection of both unstimulated and stimulated whole saliva due to its reproducibility and reliability [51].

Once saliva is collected, it can be analyzed either as whole saliva or separated by centrifugation into the supernatant and pellet and analyzed independently. The supernatant contains dissolved proteins, nucleic acids, organic metabolites, and ions while the pellet contains bacteria and viruses, human cells, debris (such as food particles) and other insoluble components. The composition of whole saliva is complex and, due to its highly proteolytic nature, presents challenges to sample preservation, particularly for disease biomarkers [52]. Unless whole saliva is utilized immediately, centrifugation is encouraged to separate the cells from the protein-containing supernatant to delay protein degradation [53]. Oftentimes, protease and nuclease inhibitors are added to prevent degradation of proteins and nucleic acids, respectively. In addition, cooling the sample on ice reduces the rate of proteolysis and other degradative processes.

Transcriptional and post-transcriptional mechanisms of gene regulation can cause protein and RNA expression levels to be dissimilar [54]. Due to this uncorrelated expression, using proteins as biomarkers for disease detection can sometimes present an incomplete picture. Detecting nucleic acids is not only a potential alternative for clinical biomarker detection in saliva but also a way of providing a more comprehensive picture of the biomarker expression activity by studying the full transcriptome. Numerous studies have successfully detected human mRNA transcripts in cell-free saliva by various methods including oligonucleotide microarray profiling and reverse transcriptase-quantitative polymerase chain reaction (RT-qPCR) [55-59]. Nucleic acid detection can be used to complement protein detection methods and provide additional information about the biomarkers and their post-transcriptional activity.

Proteins are commonly used as clinical biomarkers. The salivary proteome has been characterized and was recently reviewed by Oppenheim and coworkers [52]. Analyzing proteins in saliva involves standardizing saliva collection, sample preparation and protein extraction. Extracting and stabilizing proteins for detection typically involves separating saliva supernatant by centrifugation to prevent protein degradation and analyzing the sample immediately or storing it at 0 to 4°C for short-term use or -80°C if the sample can't be analyzed within a few hours. There are many aspects about saliva collection, storage, and composition that cannot be covered here. The remainder of this review will focus on protein biomarker detection.

2.3 Sandwich immunoassays

Sandwich immunoassays are frequently used to detect clinical biomarkers. These assays detect the presence and/or the concentration of proteins of interest using an antibody as a recognition element. Sandwich immunoassays depend on an antibody's ability to recognize and bind to a specific site, called an epitope present on the antigen of interest. A capture antibody is first attached to a solid surface and the sample containing the protein of interest is then incubated with the immobilized antibody. The bound protein is then detected by using a second antibody, called the detection antibody, which recognizes and binds to a different epitope on the protein. This second antibody carries a label that produces a signal after binding to the protein. Before conducting sandwich immunoassays, the saliva sample is typically diluted with a blocking buffer to prevent nonspecific binding, thereby minimizing background signal [60]. One way of implementing the sandwich immunoassay is to immobilize the capture antibodies onto microspheres. Our laboratory has successfully developed numerous microsphere-based immunoassays for various protein targets [60]. Using microsphere-based immunoassays has been shown to improve assay sensitivity, reproducibility and analysis times [47, 61-65]. In a microsphere-based immunoassay, the target-specific capture antibodies are chemically coupled to microspheres. These antibody-labeled microspheres are then sequentially incubated with a saliva sample to capture the target protein molecules, biotinylated detection antibodies, and a streptavidin-conjugated fluorescent probe (such as phycoerythrin) to label the sandwich complex (Fig. 2.2). The fluorescence intensities are then quantified to detect concentrations of the protein biomarkers [49, 64].

The microsphere-based protein sandwich immunoassay described above can be carried out on optical fiber arrays [66]. This array platform uses an optical fiber bundle containing approximately 50,000 individual 3.1 μm fibers. The distal ends of these optical imaging fibers are chemically etched with an acid solution to create microwells for depositing the capture microspheres [67, 68]. Microspheres are randomly deposited in the etched wells before sample exposure. The assembled arrays are then exposed to the sample to capture the target molecules and the remaining steps in the assay are performed as described above [67]. Fluorescent signals emitted from the individual microspheres in the array are detected using a conventional fluorescence microscope [68]. This array technology, which is highly sensitive and capable of multiplexed detection, has been successfully used to detect protein biomarkers for different medical conditions such as asthma and other respiratory and inflammatory diseases [66, 69, 70].

2.4 Multiplexing Capabilities

Oftentimes, a single biomarker is not sufficient for diagnosis. In these situations, detection of multiple biomarkers is required to ensure that a definitive diagnosis can be made. By encoding the microspheres to differentiate them from one another, multiplexed immunoassays can be developed where several biomarkers are detected simultaneously. There are a variety of methods for microsphere encoding including physical encoding, electronic encoding, graphical encoding, and spectrometric encoding [71]. Physical encoding differentiates between different microspheres by their physical properties such as shape, composition and density [72]. Electronic encoding uses radiofrequency memory tags for encoding the microspheres [73]. Graphical encoding is where two dimensional patterns, such as barcodes, are encoded in the microspheres [74]. Finally, spectral encoding methods encode microspheres with chemical tags that can be decoded by spectrometry [75]. Chemical tags can include different luminescent materials such as lanthanides, quantum dots (QDs), fluorescent dyes, and combinations of these materials that emit light at different wavelengths to generate uniquely identifiable signatures [76]. Fluorescent dyes are most commonly used for microsphere-based protein microarrays [77]. The fluorescent dye encoding mechanism involves staining the microspheres with various concentrations of multiple dyes and decoding the microspheres according to their distinct spectral signatures (Fig. 2.3). The staining can be done either by labeling the surface of the microsphere with the dye or by entrapping the dye in the microsphere via solvent swelling [60].

2.5 Integration into a POS Device

POS devices ideally use small sample volumes, detect multiple biomarkers, require minimal training and perform the analysis in an automated fashion. These devices are designed for portability and rapid diagnosis. Portable POS devices have many advantages over larger laboratory devices, including providing clinical care in remote locations [78] and convenient in-home testing such as the self-monitoring of blood glucose [79]. Using saliva as a biofluid in conjunction with a POS device ensures better patient compliance because of the simple collection method.

Our laboratory has developed multiple POS diagnostic studies using human saliva. In 2008, we used colorimetric test strips to monitor the effect of hemodialysis on salivary nitrite and uric acid in renal disease patients [80]. Using the test strips, reductions in salivary nitrite and uric acid were successfully observed in renal disease patients during dialysis. This assay was useful for evaluating dialysis progress and monitoring renal disease progress noninvasively.

Our lab has also been developing multiplexed assays for detecting inflammatory protein biomarkers in human saliva that have led to the development of a portable POS device. In 2009, our laboratory studied inflammatory cytokines in human saliva relevant to pulmonary inflammatory diseases [70]. A multiplexed cytokine array was developed to examine endogenous mediator patterns in saliva supernatants from patients with pulmonary inflammatory diseases such as chronic obstructive pulmonary disease and asthma. This assay was useful in inflammatory disease research and diagnostics and led to a study in which a multiplexed assay was developed to measure a panel of six salivary biomarkers for diagnosing asthma and cystic fibrosis exacerbation [69]. This successful assay demonstrates the potential to use the multiplexed protein array for respiratory disease diagnosis, which may be applicable to other protein biomarkers and diseases. Last year, our lab migrated these assays onto an automated, integrated POS platform [49]. Six inflammatory protein biomarkers were used for the multiplexed assay. Ten µl of a saliva sample is loaded onto a microfluidic chip (containing all the necessary reagents required for the immunoassay) and the chip is inserted into the automated POS device. The results are read optically and are available within 70 minutes. This POS device was used in two hospitals on over 250 human saliva samples collected from different individuals to successfully diagnose various respiratory conditions. This study demonstrated the POS device's potential to assist with the diagnosis of respiratory diseases and potentially provide a platform to diagnose additional medical conditions.

2.6 Future Directions

Our laboratory has demonstrated a POS device capable of automated, multiplexed and sensitive biomarker detection. To enhance our POS device’s sensitivity, we can improve the detection limit by implementing digital ELISAs using a single molecule array (SiMoA) technology. The SiMoA platform, originating from our lab, has previously been used to capture and analyze individual target molecules in the microwell optical fiber arrays [68]. Briefly, SiMoA involves adding capture antibody microspheres to a sample containing the target protein molecules but where there are more microspheres than molecules. After sandwich formation with an enzyme-labeled detection antibody, the microspheres are loaded into femtoliter reaction wells and sealed with a fluorogenic substrate. Determining the number of wells that display increased fluorescence after incubation with the fluorogenic substrate quantifies the protein concentration from the original sample [81]. This SiMoA technology is capable of detecting biomarkers at the attomolar (aM) to femtomolar (fM) range [82]. These detection limits are hundreds to thousands of times more sensitive than conventional ELISAs, enabling the detection of target proteins at concentrations that have not been measured before. SiMoA technology was developed in our laboratory and recently extended to nucleic acids, thereby providing a sensitive detection alternative to existing technologies that utilize amplification, such as the polymerase chain reaction (PCR) [83]. To summarize, SiMoA technology is being used to detect virus particles, proteins, and nucleic acid biomarkers [82, 84] and has the potential to improve the sensitivity of our POS device for future applications.

Currently, our laboratory is implementing the POS device to assess oral feeding maturity in the newborn. Premature births affect an estimated 11.5% of all pregnancies in the United States resulting in medical costs exceeding $26 billion annually [85]. Most of these infants do not have the developmental maturity to successfully feed by mouth and must confront the challenges of learning to feed orally before they can be discharged from the hospital. Currently, there are no strategies available to objectively assess oral feeding maturity in newborns [19]. This lack of an objective assessment tool has resulted in missed opportunities to feed infants with mature oral feeding skills, while placing immature oral feeders at risk for significant feeding-associated morbidities [86]. Each failed approach results in a prolonged length of stay with an aggregated loss of millions of dollars over the neonatal population in associated health care costs [19]. We are using the POS platform to provide a sensitive and multiplexed readout of the multiple biomarkers responsible for oral feeding maturity [87]. This project represents an important opportunity to integrate advances in salivary molecular diagnostics into the neonatal population to enhance clinical decisions and improve patient outcomes.

The POS device developed in our laboratory using saliva has the potential to be beneficial outside the hospital environment in venues such as remote resource limited locations, areas where blood collection is not possible because of cultural issues, locations where biohazard containment is not possible and venues for individual monitoring. Future improvements to this POS device will include battery power as an option and manual methods that avoid using a pipet to make it more accessible to other venues outside the hospital.

2.7 Conclusion

A completely automated and portable POS device using noninvasively collected samples is an ideal diagnostic platform. The ability to noninvasively assess patients rapidly at the point of care is a major goal of modern medicine. Saliva is an attractive bio-fluid to assess health, disease and development [88]. Furthermore, saliva is preferred over other types of biological samples due to its convenience and ease of collection. It may be collected repeatedly, even in the most vulnerable patients, without risk of harm [89].

Technological advances are now permitting the high-throughput analysis of saliva for thousands of genes, proteins and metabolites from a single sample source [32, 66]. Recently, our laboratory developed an integrated automated diagnostic platform based on an antibody-based multiplexed protein microarray [49, 60]. This platform fulfills the requirements of a POS device, including integration, automation, multiplexed detection ability, small sample size, fast analysis, and minimal training. The next step for this technology is translating this device into clinical care for relevant correlates requiring real-time noninvasive assessment such as for oral feeding maturity in the neonatal population.

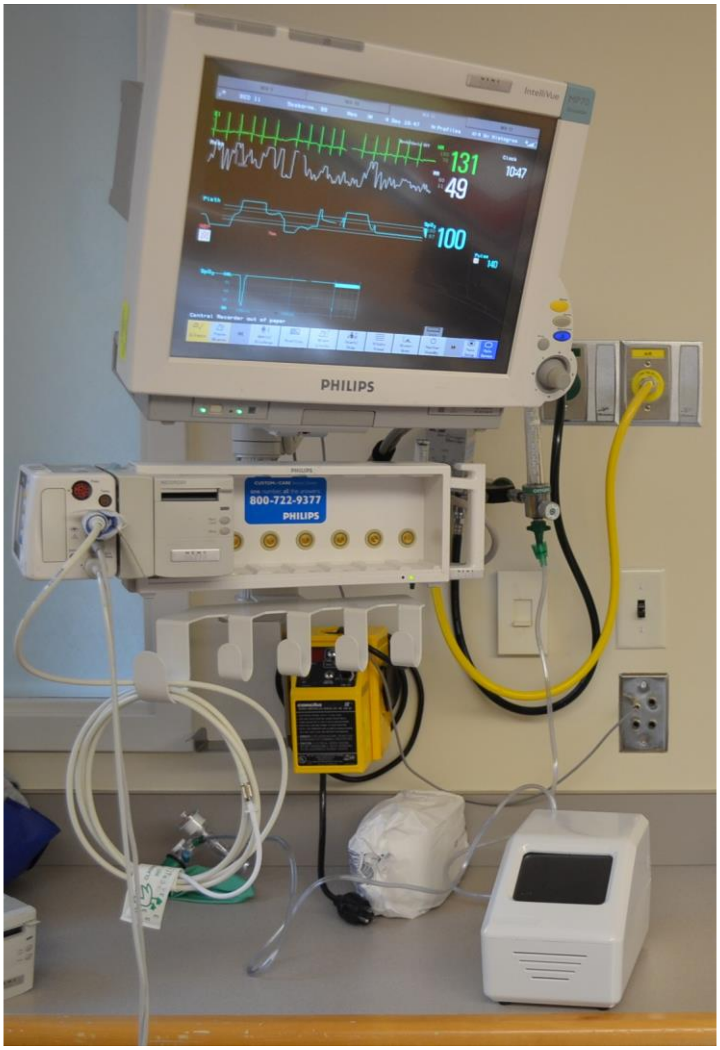


Figure 2.1: Point of service (POS) device, developed in our laboratory for salivary diagnostics, present at a patient’s hospital bedside. The device is powered by a conventional AC adapter providing 12 VDC at 8.5 A (not shown), and the case is 15 cm (wide) × 15 cm (tall) × 25 cm (deep). The weight of the device is 2.7 kg. The concentrations of multiple protein biomarkers in the saliva sample are quantified via sandwich immunoassays conducted on a microfluidic chip present in the device.

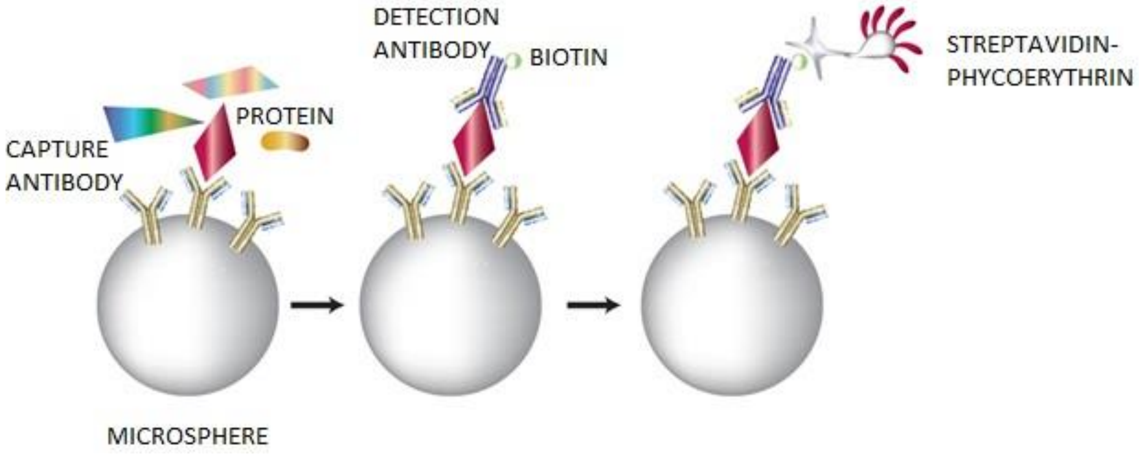
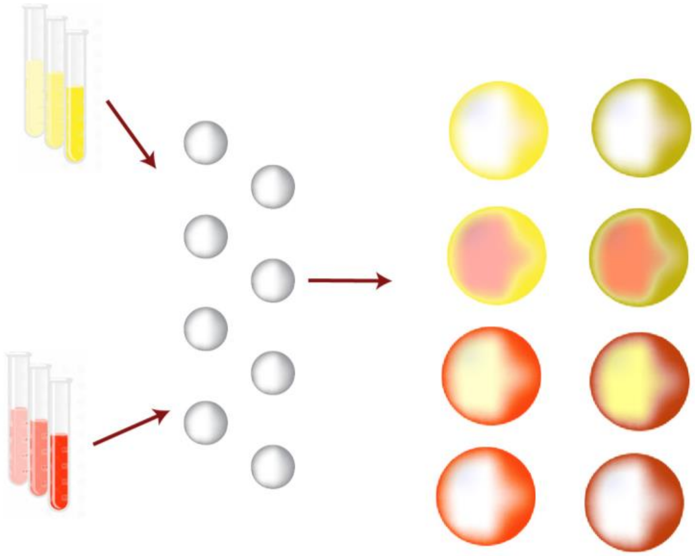


Figure 2.2: Workflow for the protein sandwich immunoassay. Microspheres are coupled with capture antibodies, incubated with a saliva sample and then biotinylated detection antibodies, and finally a streptavidin-conjugated fluorescent probe is added to the complex for a quantified readout.



BEAD TYPE		Eu- TTA (mM)		
		0	10	100
C30 (mM)	100	🟨	🟨	🟨
	10	🟧	🟧	🟫
	0		🟫	🟥

Figure 2.3: Depiction of encoding beads with multiple fluorescent dyes to produce numerous bead types for multiplexed arrays. Microspheres are encoded with different concentrations of two fluorescent dyes, Coumarin (C30) and Europium (Eu-TTA) to produce distinct bead types to be used for detection of multiple biomarkers simultaneously.

Chapter 3: Development of a rapid salivary proteomic platform for oral feeding readiness in the preterm newborn

Development of a rapid salivary proteomic platform for oral feeding readiness in the preterm newborn, P Khanna, JL Maron, D Walt, Frontiers in Pediatrics 5, 268

Reprinted here with permission of publisher.

3.1 Introduction

Although oral feeding competency is a discharge requirement from the neonatal intensive care unit (NICU), there is currently a paucity of objective assessment tools to determine oral feeding maturity in this population [19, 85]. Rather, standard of care is largely limited to subjective assessment of an infant's feeding cues (i.e. ability to suck on a pacifier) once an infant corrects to > 32 weeks' post-conceptional age (PCA) and has a stable respiratory status [22, 90-92]. The absence of an objective assessment tool to determine oral feeding readiness has not only placed immature oral feeders at risk for significant feeding associated morbidities, including choking, poor growth, impaired short- and long-term neurodevelopmental outcomes and feeding aversions, but it has also resulted in prolonged hospitalization, increased healthcare costs, and parental anxiety [20, 21, 86, 93, 94]..

Previously, salivary gene expression analyses on hundreds of premature infants' at both pre- and post-oral feeding success, identified five genes involved in oral feeding maturity that were differentially expressed between successful and unsuccessful oral feeders [15]. These biomarkers included *NPY2R* (hunger signaling), *AMPK* (energy homeostasis), *PLXNA1* (olfactory neurogenesis), *NPHP4* (visual behavior) and *WNT3* (facial development). While this prior study demonstrated the feasibility of utilizing saliva as a noninvasive biofluid to detect transcriptomic biomarkers associated with neonatal developmental milestones, the ability to monitor these markers in a timely manner to inform care remains a challenge. Proteins have numerous benefits over mRNA transcripts including their relative abundance and stability [95]. Combined with their relative ease for detection and quantification compared to genes, proteins are more appealing for the development of a rapid proteomic platform. This study aimed to translate the previously-described gene expression panel to a salivary proteomic platform in order to objectively assess oral feeding readiness in a more rapid format to limit neonatal morbidities and improve outcomes.

3.2 Materials and Methods

3.2.1 Infant Recruitment and Saliva Collection

This study was approved by the Tufts Medical Center Institutional Review Board. Parents of premature infants ranging from 33 to 39 weeks' PCA were consented for enrollment. The Tufts Medical Center (TMC) NICU utilizes the cue-based feeding assessment protocol of Ludwig and Waitzman [96]. In accordance to this protocol, infants with a stable respiratory status and PCA of >32 weeks were assessed for feeding capability as part of their routine vital signs by the nursing staff. This cue-based feeding assessment protocol scores infants from 1-5 (1 signifies oral feeding ready and 5 signifies no oral feeding cues present). Infants who demonstrate a score ranging from 1 to 2 consecutively for three assessment points are allowed to attempt oral feeds. No infant less than 32 weeks' PCA was offered oral feeding in the NICU, thus, saliva was not collected until the infant was mature enough to attempt oral feeds (e.g. >32 weeks' PCA). Saliva samples were collected equally from successful (n=10) and unsuccessful (n=10) oral feeders at a single time-point. Unsuccessful oral feeders, termed 'non-feeders', took < 50% of feeds by mouth; successful oral feeders, termed 'feeders', took full (100%) oral feeds. The < 50% of feeds by mouth cutoff was utilized to ensure that extraneous factors (e.g. nursing staff ratios or acute medical emergencies that may have prohibited an oral feeding session for the infant) did not contribute to an infant's designated feeding status. Only those infants who consistently demonstrated an oral intake of <50% of full enteral nutrition were deemed unsuccessful oral feeders. The infants from both cohorts were matched for sex, gestational age, PCA at time of sample collection and ethnicity to limit the potential confounding effects of these variables on protein expression (Table 3.1).

Two saliva samples were collected for protein analysis from a single time point in all subjects. Salivary protein was collected, stabilized and extracted from each sample using established protocols [60]. Final elution volume following extraction was 250 µl, making it necessary to collect multiple samples from each infant to have sufficient volumes for downstream experiments (required volume: 120µl per biomarker). Total protein extracted from the salivary samples was stored at -80°C pending analysis.

3.2.2 Development of rapid proteomic platform

Immunoassays were used to measure target protein molecules in a sample. These assays were based on the specific recognition of target molecules by both capture and detection antibodies. For detection of proteins, the sandwich format was used due to its high specificity and ability to analyze bio-fluids in complex matrices [69]. For protein immunoassays, antibody pairs and recombinant protein standards were utilized. In brief, capture antibodies were immobilized on microspheres that could be suspended in solution. The microspheres were incubated for 40 minutes with the sample to allow for protein-specific antibody binding. Subsequently, detection antibody was added to the solution that then bound to another epitope on the target protein. The microspheres were washed and then incubated for 20 minutes with 2.5 μg/mL of a streptavidin−phycoerythrin conjugate (Columbia Biosciences) to generate the fluorescent complex. After a final wash and resuspension in 75 μL phosphate buffer saline tween-20 (PBST), the assay results were evaluated on the Tecan Infinite M200 Plate Reader platform. Serial dilutions of recombinant protein standards for all biomarkers were run to generate calibration curves and to determine assay sensitivity, antibody performance and protein detection range. Additionally, salivary sample dilution series and spike-in and recovery experiments were run for all developed assays to exclude non-specific binding and to determine the effect of the saliva matrix on the proteins. Manufacturer's details and catalog numbers for all antibodies and recombinant protein standards used are provided in Table 3.4.

3.2.3 Testing of rapid proteomic platform

Sandwich assays were performed for detection of protein biomarkers in neonatal saliva as described above. Alongside the recombinant protein standards assays, the microspheres were also incubated with neonatal saliva samples for feeders and non-feeders. The fluorescence intensity for all samples was measured on the Tecan Infinite M200 Plate Reader and their protein concentrations were determined based upon the calibration curves on the same plates.

3.2.4 Data Analysis

GraphPad Prism was used to generate calibration curves for all biomarkers. These curves were fit using a 4PL fit with 1/y2 weighting factor. The calibration curves were used to determine concentrations of all protein biomarkers from the neonatal saliva samples. GAPDH and YWHAZ were used as reference proteins to normalize varying neonatal saliva volume input and serve as quality control indices. Detection of both reference proteins was required for a sample to be considered in our analysis [97-99]. Samples were normalized for comparative analysis with the use of the geometric mean (GM) of the two reference protein concentrations, using the following formula (example shown below is for AMPK):

$$\Delta \text{AMPK (n)} = \text{AMPK (n)} / \text{GM [GAPDH (n)} + \text{YWHAZ (n)]}$$

3.3 Results

Immunoassays were successfully developed for five of the seven biomarkers: AMPK, NPY2R, WNT3, GAPDH and YWHAZ. No compatible antibody pairs were found for NPHP4 and PLXNA1. Assay development for all biomarkers is summarized in Table 3.2. Results for salivary sample dilution series and spike-in and recovery experiments are summarized in Table. 3.5. Calibration curves for all assays are shown in Figure 3.1. Raw and normalized data are depicted in Figures 3.2 and 3.3, respectively.

All neonatal saliva samples met quality control as defined by detection of both reference proteins in a sample. Subject demographics are summarized in Table 3.1. Protein levels of both AMPK and NPYR mirrored the gene expression profiles reported previously. Namely, AMPK was either undetectable (n=6) or decreased (n=4) in unsuccessful oral feeders while expression levels of NYP2R were increased in unsuccessful oral feeders (n=10). Median concentrations for AMPK and NPY2R in 20 neonatal saliva samples split by feeders and non-feeders are summarized in Table 3.3. WNT3 was undetectable in the neonatal saliva samples analyzed.

3.4 Discussion

The transcriptome and proteome, unlike the genome, provide insight into biological function and phenotype. Transcription and translation are complex mechanisms consisting of stochastic expression levels of RNA and protein over time. Numerous studies have shown that there is not a direct correlation between the levels of mRNA and protein [100-103]. However, to date, no study has used transcriptomic information as a guide for determining whether or not proteins are expressed. Using information from the neonatal salivary transcriptome, we hypothesized not only that the same proteins would be present and correlate with mRNA expression levels, but also that this information would advance our understanding of the dynamic relationship between the transcriptome and the proteome in the developing premature newborn.

A rapid proteomic bedside platform for assessing oral feeding readiness has the potential to limit hospital length of stay and dramatically reduce health care costs. Two specific groups of neonates, in particular, may largely benefit from such a diagnostic assay. First, utilizing this assay to identify neonates with mature oral feeding skills, in a timely fashion, who could begin oral feeding trials without the fear of deleterious side effects, would likely reduce hospital length of stay. Second, this assay could be utilized to understand the developmental pathways limiting oral feeding success in infants who struggle to orally feed despite an advancing PCA. This approach would allow caregivers to personalize care plans based specifically on an infant's salivary profile. Here too, there is an important opportunity to expedite oral feeding maturation and reduce time spent in the NICU. With average NICU costs at $3500 per day in the US, successful development of this rapid $5 assay to assess oral feeding maturation may result in healthcare cost savings of billions of dollars per year.

Previously, we identified five key regulatory genes responsible for oral feeding maturity that were differentially expressed between successful and unsuccessful oral feeders, including *NPY2R*, *AMPK*, *PLXNA1*, *NPHP4* and *WNT3* [15]. In this prior work, each biomarker was expressed in a binary fashion (i.e. it was either present or absent as ascertained by amplification). With mRNA expression, an infant demonstrated a mature oral feeding pattern when *AMPK, PLXNA1,* and *NPHP4* were present, and when *NPY2R* and *WNT3* were absent in neonatal saliva. Using these data, we hypothesized that the protein expression for these targets would parallel their gene expression and allow for translation to a more rapid proteomic bedside assay platform to assess oral feeding readiness.

In this pilot study, assays were successfully developed for five of the seven biomarkers including AMPK, NPY2R, WNT3, GAPDH and YWHAZ. We were unable to develop protein based assays for all the biomarkers previously identified because of lack of suitable binding reagents for all the proteins. When these five immunoassays were carried out on newborn saliva, only the two reference biomarkers (GAPDH, YWHAZ) and the two proteins involved in hunger signaling (NPY2R, AMPK) were detectable. WNT3, associated with facial development, was undetectable. Each of the two detectable proteins paralleled the gene expression results previously described. In contrast to the gene expression levels, where the biomarkers were informative in a binary fashion, proteins could be measured at levels such that we could quantify them. This discrepancy may be due to each respective assay's detection limit or because of differences between mRNA expression levels and protein abundance in neonatal saliva. The biological significance of this finding is unknown.

Of the three proteomic biomarkers assessed in this study, only those involved in hunger signaling were readily detectable in neonatal saliva. Their presence suggests not only that they may play an important role in regulating feeding behavior in the newborn, but also implies that protein levels in saliva may be required for a maturing gut-brain axis necessary for successful oral feeding. NPY2R was first described in 1996; it is an appetite hormone and candidate gene for obesity development and control of food intake [87, 104, 105]. Although one feeder's NPY2R expression (F6) remained an outlier after normalization, this result could be attributed to the subject's earlier gestational age (GA), 31 5/7, compared to the other subjects (GA: 32 1/7 to 38 2/7).

Similar to NPY2R, AMPK expression may also play a key regulatory role in feeding maturation. AMPK detects and maintains metabolic energy balance by promoting ATP production and facilitating the pathways involved in circadian rhythms of metabolism and feeding behavior [106, 107]. Expression of these two protein biomarkers corresponded not only to their gene expression profiles previously reported, but were predictive of the feeding status of the newborn. Thus, they have the potential to serve as independent and reliable biomarkers of neonatal oral feeding success.

Limitations of this study include the small sample size and the inability to detect all the biomarkers previously shown to be indicative of oral feeding readiness in the newborn. The assay's ability to detect only biomarkers involved in hunger signaling hinders its current applicability at the bedside. Biomarker proteins corresponding to other key developmental milestones required for oral feeding success including neurodevelopment, gastrointestinal maturation and sensory integration will need to be identified before the assay will reach its full diagnostic potential. Nevertheless, this study is a promising first step towards the development of a NICU bedside device to assess oral feeding maturation to improve care and outcomes in this population. Finally, it is important to note that the assay requires only a small amount of neonatal saliva, which is easy to obtain and avoids blood collection—a cause of morbidity in neonates.

In conclusion, this pilot study is not only clinically relevant because we show concordance between protein and gene expression, but we also demonstrate that protein expression can be informative of oral feeding status. Thus, for these particular biomarkers, a rapid proteomic assay may be utilized to assess real-time hypothalamic feeding development using neonatal saliva.

Table 3.1: Patient Demographics

FEEDER	SEX	PCA	GA	NON-FEEDER	SEX	PCA	GA
1	F	39 2/7	37 0/7	1	F	38 5/7	38 2/7
2	F	33 4/7	32 1/7	2	F	32 6/7	32 2/7
3	F	37 2/7	34 0/7	3	F	37 0/7	36 2/7
4	F	34 4/7	33 4/7	4	F	34 2/7	33 1/7
5	F	35 1/7	32 5/7	5	F	35 2/7	33 5/7
6	M	35 2/7	31 5/7	6	M	34 6/7	34 1/7
7	M	37 2/7	36 6/7	7	M	34 1/7	33 0/7
8	M	35 0/7	33 1/7	8	M	34 3/7	32 6/7
9	M	39 2/7	37 4/7	9	M	39 1/7	37 6/7
10	M	35 3/7	34 5/7	10	M	34 6/7	33 0/7

Table 3.2: Assay Overviews

Protein	Assay Developed	Protein Detected
NPY2R	Y	Y
AMPK	Y	Y
PLXNA1	N	N
NPHP4	N	N
WNT3	Y	N
GAPDH	Y	Y
YWHAZ	Y	Y

Table 3.3: Median concentrations (pM) of the two measured biomarkers in 20 neonatal saliva samples split by feeders and non-feeders

Biomarkers	AMPK (Feeders)	AMPK (Non-Feeders)	NPY2R (Feeders)	NPY2R (Non-Feeders)
Median (pM)	510	271	31.95	415
Detectable Samples	10/10	4/10	10/10	10/10

Table 3.4: Manufacturer's details and catalog numbers for all antibodies and recombinant protein standards

Target Protein	Capture Antibody	Detection Antibody	Recombinant Protein
AMPK	Fisher/R&D Systems DYC3197	Fisher/R&D Systems DYC3197	Fisher/R&D Systems DYC3197
GAPDH	Fisher/R&D Systems DYC5718	Fisher/R&D Systems DYC5718	Fisher/R&D Systems DYC5718
YWHAZ	Fisher/R&D Systems DY2669	Fisher/R&D Systems DY2669	Fisher/R&D Systems DY2669
NPY2R	Sigma-Aldrich SAB2500707	LSBio (Direct) LS-C264678	ABCAM AB152580
WNT3	ABCAM AB52568	Novus Biologicals, LLC H00007473-D01PB	ABCAM AB132336

Table 3.5: Assay validations

Protein	0pM + 250ul saliva	0pM + 125ul saliva	10pM	100pM	1000pM	1000pM (+ 1hr)
	Protein Detected		% Recovery			
AMPK (in neonatal saliva)	Y	N	108.20%	87.70%	105.50%	100.70%
GAPDH (in neonatal saliva)	Y	N	98.40%	111.30%	82.30%	89.30%
NPY2R (in neonatal saliva)	Y	N	117.50%	109.40%	109.90%	100.10%
YWHAZ (in neonatal saliva)	Y	N	102.10%	99.30%	97.00%	91.90%

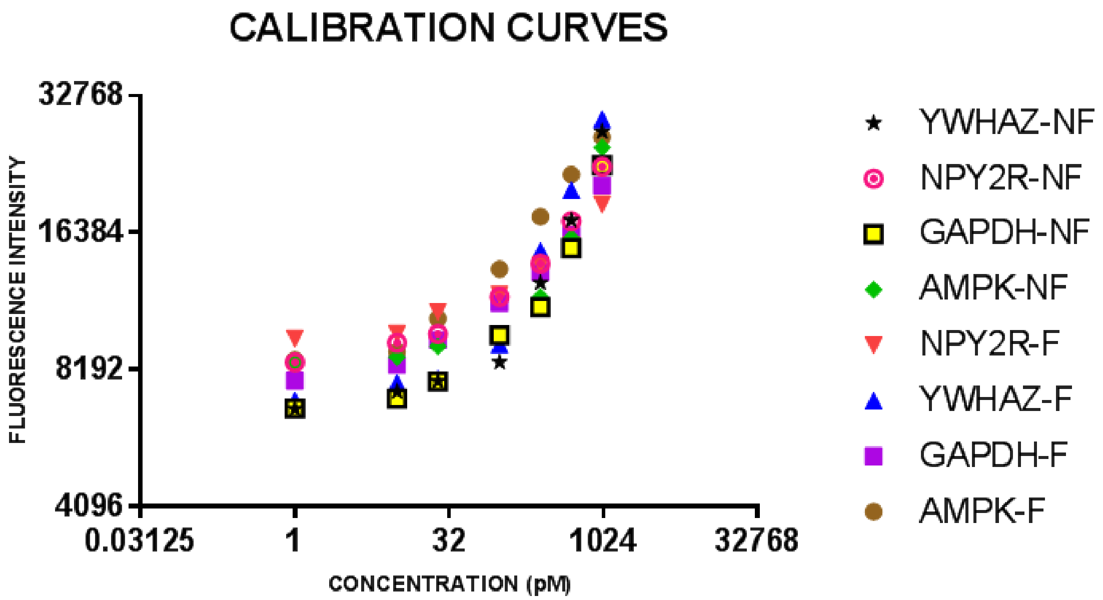


Figure 3.1: Calibration curves for eight bulk assays for salivary neonatal biomarkers. Serial dilutions of recombinant protein standards for AMPK and NPY2R were run on all plates for feeder and non-feeder infant samples. Error bars depict the standard deviations for the values as they were all run in triplicate.

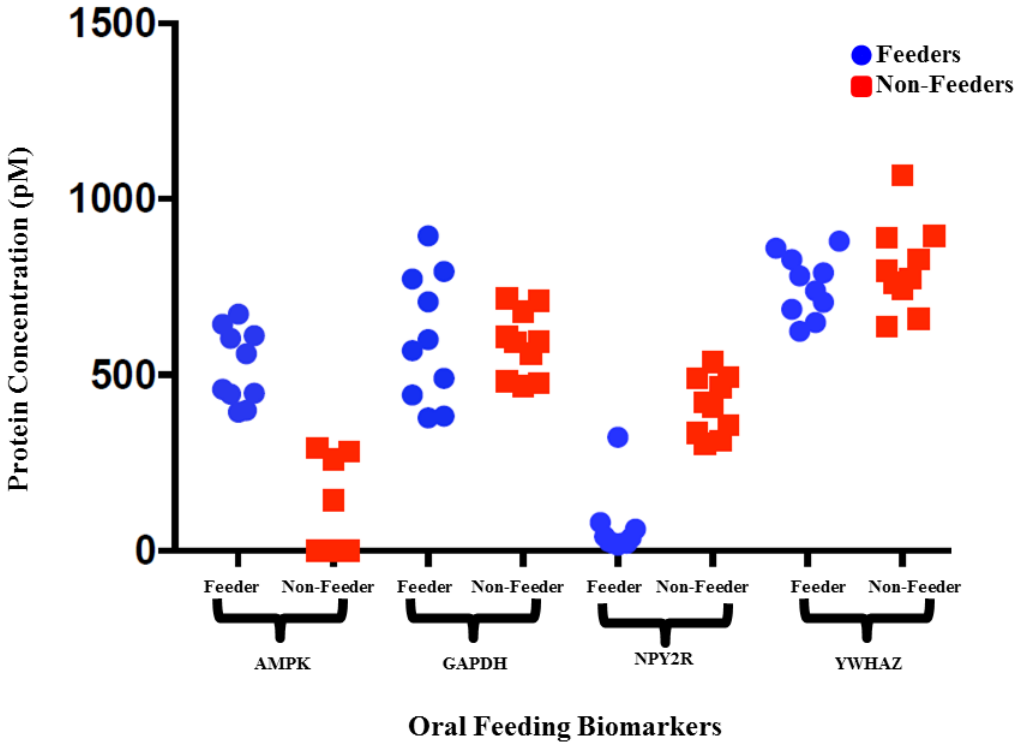


Figure 3.2: Raw protein expression levels for salivary neonatal biomarkers. Saliva samples from ten feeders and ten non-feeders were run using bulk immunoassays for AMPK, GAPDH, NPY2R and YWHAZ. Salivary protein concentrations (pM) were calculated using the calibration curves and plotted. The wide variation observed in protein expression of reference biomarkers GAPDH and YWHAZ signify varying protein input between samples.

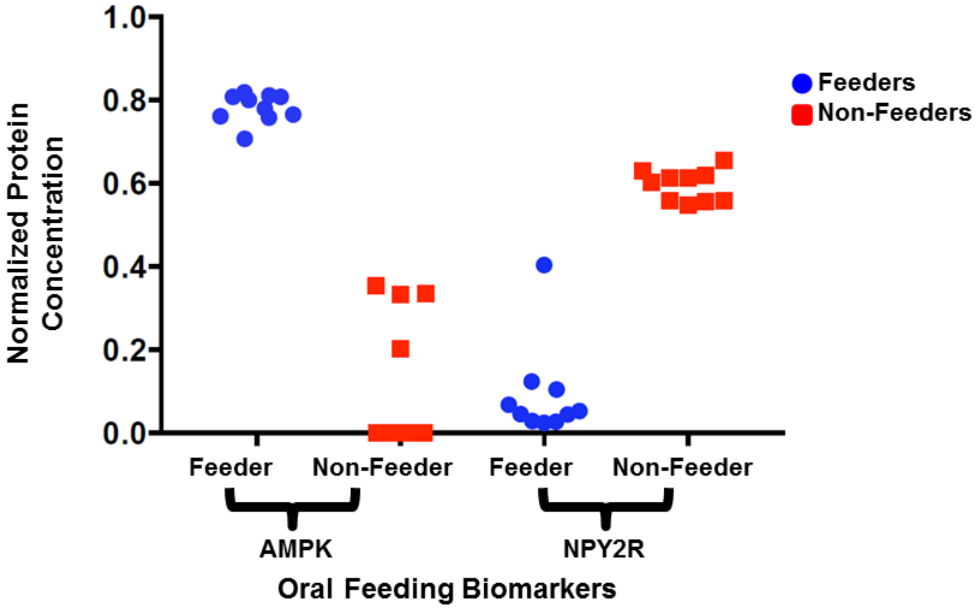


Figure 3.3: Normalized protein expression levels for salivary neonatal biomarkers. The protein concentrations present in the clinical samples for AMPK and NPY2R were normalized against GAPDH and YWHAZ and plotted for all ten feeders and all ten non-feeders.

Chapter 4: Salivary RNA Sequencing Analysis Highlights Sex Specific Developmental Pathways Involved in the Maturation of Oral Feeding in the Preterm Infant

Salivary RNA Sequencing Analysis Highlights Sex Specific Developmental Pathways Involved in the Maturation of Oral Feeding in the Preterm Infant. Prarthana Khanna, PhD, Kaley Jenney, Albert Tai, PhD, David R. Walt, PhD, Jill L. Maron, MD, MPH. To be submitted to JAMA PEDS.

4.1 Introduction

To date, next-generation sequencing has largely been used in the neonate for either whole genome or whole exome sequencing in critically ill newborns or those with suspected monogenetic disorders. However, the vast majority of infants born prematurely are affected neither by genetic mutations nor syndromes. Rather, neonatal morbidities are largely a result of disrupted developmental pathways that are a direct consequence of preterm birth. Thus, utilizing high-throughput sequencing technology to explore gene expression in the newborn may provide a near real-time window into ongoing development and may help to identify unique developmental pathways associated with neonatal pathology and developmental impairments. These targets may elucidate opportunities to personalize care and treatment strategies based upon an infant's gene expression profile.

Attainment of oral feeding competency is a major determinant of length of stay in the Neonatal Intensive Care Unit (NICU) and represents a developmental challenge for the majority of the 15 million infants born prematurely (<37 weeks gestational age) worldwide each year. Inappropriate feeding attempts can lead to acute and long-term morbidities, as well as prolonged hospitalizations with associated healthcare costs [18, 20, 90, 94, 108, 109]. Further, infants who fail to successfully orally feed by corrected term gestational age (GA) are at increased risk for developmental delays throughout infancy, childhood and beyond. Despite the prevalence of oral feeding morbidities and their long-term health consequences, our ability to assess oral feeding maturity, and more importantly determine the biological mechanisms limiting oral feeding success, remains a clinical challenge.

Successful oral feeding is dependent upon the simultaneous maturation and integration of the gut-brain axis, as well as sensorimotor, neurodevelopmental and gastrointestinal systems. Developmental maturation of these systems varies among infants and is believed to be affected, in part, by sex and GA. Males typically learn to orally feed at older post-conceptional ages (PCAs) compared to females. Moreover, infants born at earlier GAs learn to orally feed at older PCAs compared to infants born later in gestation. In order to provide relevant information to caregivers to personalize treatment strategies and improve oral feeding outcomes, assessment tools must be able to evaluate multiple developmental systems. Cochrane Reviews conducted in both 2012 and 2016 confirmed the futility of available feeding assessment tools for use in the newborn, concluding both times that there is currently "no evidence to inform clinical practice" [19, 23].

This study utilized RNA sequencing in order to advance our understanding of disrupted developmental pathways and sex-specific differences involved in oral feeding maturation in the premature newborn. Sequencing was performed on whole saliva, a rich source of systemic gene expression, previously shown to be an informative bio-fluid capable of discerning between successful and unsuccessful neonatal oral feeders. We hypothesized that this novel approach to noninvasive assessment of the developing newborn could significantly impact our clinical approach to oral feeding difficulties and ultimately improve short and long-term neonatal outcomes for millions of premature neonates born each year.

4.2 Materials & Methods

4.2.1 Subject Selection and Recruitment

This prospective, observational, single-center study was approved by the Tufts Medical Center Institutional Review Board. Informed consent was obtained from parents of premature infants ranging from 32 to 36 weeks' PCA. To limit the potential confounding effects of ethnicity on the sequencing data, only Caucasian infants were asked to participate in this pilot study. There was equal representation of male and female infants. The Tufts Medical Center NICU utilizes the cue-based feeding assessment protocol of Ludwig and Waitzman[96]. In accordance with this protocol, no infant less than 32 weeks' PCA is offered oral feeding in the NICU. Infants were considered unsuccessful oral feeders if they took < 50% of feeds by mouth (non-feeder); successful oral feeders took 100% of enteral nutrition by mouth and did not have a nasogastric tube in place (feeder).

4.2.2 Saliva Collection & Quality Control

Saliva samples were obtained using previously described techniques [110]. Briefly, saliva was collected with a 1 mL syringe attached to low wall suction. Saliva was placed immediately in RNAProtect Saliva (QIAGEN) at the bedside, vortexed and put on ice. Samples were stored for a minimum of 48 hours but no longer than 28 days at 4°C prior to RNA extraction with the RNAProtect Saliva Mini Kit (QIAGEN) per manufacturer's instructions. On column DNase treatment was performed for each sample to eliminate

DNA contamination. Extracted total RNA was stored at -80°C pending RNA sequencing analysis.

Prior to RNA sequencing analysis, the quality and quantity of extracted total RNA was assessed on the Agilent Bioanalyzer. Only those samples that met pre-established criteria designed specifically to target cell-free RNA in saliva were subsequently sequenced. Criteria included: (1) a minimum of 500 ng of total RNA per sample; (2) a RNA Integrity Number (RIN) between 5 and 8, suggestive of cell-free rather than cellular RNA; and (3) 28S peak > 18S ribosomal RNA peaks on the qualitative analysis.

4.2.3 RNA Sequencing

Samples that met criteria underwent next generation sequencing at the Tufts University Genomic Core facility. We used Illumina sequencing in this study. RNA sequencing with the Illumina platform involves library preparation with the Illumina TruSeq Stranded Total RNA with Ribo-Zero Globin kit. After library preparation is completed the libraries are denatured, introduced into the flow cell, and subjected to “bridge amplification” in order to create clonal clusters of single stranded cDNA molecules. Next, the DNA is sequenced using Reversible Terminator chemistry. Samples are run in HiSeq flow cells and sequenced via paired end 150 bps in rapid-mode in one lane with approximately 200 million reads per sample. RNA sequencing is significantly more advanced than microarrays for several reasons including unbiased detection of novel transcripts, broader dynamic range, increased specificity and sensitivity and easier detection of rare and low-abundance transcripts. Unlike microarrays, RNA sequencing does not utilize transcript-specific probes, allowing a much broader dynamic range to discover new genes.

4.2.4 Data Analysis

Raw data were obtained in the form of FASTQ files. Bioinformatics analyses were performed on the Tufts University Linux Research Cluster (Boston, MA). Quality control checks were run on the data using the FASTQC assessment. Reads were truncated to remove base positions that scored lesser than a low median score. A median quality score of < 20 was deemed unusable. Processed reads were mapped to the UCSC human genome 19 (hg19). The NGS mapper, Tophat, was used to map RNA sequencing reads. Reads were assembled into complete transcripts that were analyzed for differential expression. Differential expression analyses on samples occurred with Qlucore for each dataset. Genes differentially expressed between feeding time points were identified and further explored with the use of Ingenuity® Pathway Analysis (IPA) software. Alternative splicing of gene targets and transcriptional regulatory elements were also examined. Qlucore was used to visualize the data and generate 3D Principal Component Analysis (PCAs) and heat maps [111].

4.3 Results

Subject demographics are summarized in Table 4.1. Thirty-two infants were recruited for this study; 24 subjects met RNA sequencing quality criteria and ultimately were sequenced. Mean RNA sequencing read alignment rates averaged 40% (Table 4.3). Demographics for the 24 patients are summarized in Table 4.1. The initial analysis by

Qlucore included the entire cohort of successful feeders vs. unsuccessful feeders with the p value set at 0.05. This comprehensive analysis identified 63 genes that were differentially expressed between the successful (N=12) and unsuccessful oral feeders' (N=12). Subsequently, a different series of analyses were conducted by removing the variable of sex. Instead of comparing all 24 infants, these samples were split by sex and only male feeders were compared with male non-feeders and only female feeders were compared with female non-feeders. Both of these analyses were also conducted using Qlucore and the p value was set consistently at 0.05 through all three analyses. The drawback to only examining the data as a whole and not split by sex would have involved masking any effects of sex on the gene expression levels. When separated by sex, 88 differentially expressed genes were identified among the female cohorts (N=12), 15 of which overlapped with the original 63 genes (comparing all feeders vs. non-feeders). Comparatively, 78 differentially expressed genes were identified among the male cohorts (N=12), six of which overlapped with the original 63 genes. No overlap of genes was observed in the 88 and 78 differentially expressed genes between the females and males, respectively. Differentially expressed genes identified via all three evaluations are presented in Table 4.2.

Principal component analysis (PCA) displayed distinct clustering of successful feeders vs. unsuccessful feeders (Fig. 4.1). Across all three dataset comparisons (all, females, and males), the male subgroup showed the most distinctive clustering based on feeding status. The heat maps displayed distinctive signatures for the two different feeding cohorts for all three dataset comparisons (Fig. 4.2).

For the 63 genes that were differentially expressed between feeders and non-feeders, IPA identified the following most statistically significant developmental pathways, independent of sex: nervous system development and function (p values: < 0.02 to < 0.0001, n = 7 genes), tissue morphology (p values: < 0.03 to < 0.002, n = 10 genes), embryonic development (p values: < 0.03 to < 0.002, n = 12 genes), hematologic development and function (p values: < 0.03 to < 0.002, n = 9 genes) and hematopoiesis (p values: < 0.03 to < 0.002, n = 6 genes) [93, 112-115].

For the female only cohort, IPA identified hematologic development and function (p values: < 0.01 to < 0.0001, n = 12 genes), immune cell trafficking (p values: < 0.01 to < 0.0001, n = 8 genes), lymphoid tissue structure and development (p values: < 0.01 to < 0.0001, n = 7 genes), digestive system development and function (p values: < 0.01 to < 0.0002, n = 7 genes) and humoral immune response (p values: < 0.01 to < 0.0002, n = 4 genes) [114-118]. Among these pathways, digestive system development and function was deemed most relevant to feeding. Differentially-expressed biomarkers associated with gastrointestinal and digestive development included *COMMD3-BM11, BCL2, FST*, *GSKIP, NECTIN3, PTK6,* and *PDCD1* [117, 119, 120]. IPA analysis characterized these biomarkers as being associated with atypical neurogenesis of the intestine, tooth development (i.e. root development, incisor development), development of the secondary and hard palate and an increase of the intestinal villus [116, 120, 121].

For the male cohort, IPA identified nervous system development and function (p values: < 0.04 to < 0.0008, n = 6 genes), cardiovascular system development and function (p values: < 0.05 to < 0.003, n = 7 genes), embryonic development (p values: < 0.05 to <

0.003, n = 6 genes), connective tissue development and function (p values: < 0.05 to < 0.003, n = 8 genes) and hair and skin development (p values: < 0.05 to < 0.003, n = 4 genes) [78, 118, 120, 122, 123]. Additionally, IPA also highlighted genes related to memory and learning, disruption in palatal shelf formation, maturation of circadian rhythms, abnormal morphology of hindgut and mesenchyme and development of the abdomen [107, 124, 125]. Differentially-expressed biomarkers associated with neurodevelopment included *G6M6B*, *TNFRSF21, XPC, FOXO3, GPM6B, SLC39A13*, and *BRINP1* [122, 123]. IPA analysis characterized these biomarkers as being associated with abnormal myelin sheath development, a decreased size of the olfactory bulb (associated with reduced sense of smell), decreased size of dentate gyrus, and decreased size of the anterior commissure (which contains decussating fibers from the olfactory tract) [126-129]. In addition, IPA analysis also characterized these biomarkers as being associated with the thinning of the cornea and the abnormal morphology of CA1 pyramidal neurons in the hippocampus (a key component in memory) [119, 123].

4.4 Discussion

Premature birth results in an acute disruption of normal development of the newborn and may cause significant morbidities with lifelong impairments [15]. Challenges associated with the achievement of oral feeding competency affect the majority of infants born prematurely, and may lead to morbidities in the neonatal period including choking, desaturations and parental anxiety, or more significant complications including aspiration, feeding aversion and failure to thrive [108]. Importantly, an infant’s ability to orally feed has been directly associated with later developmental milestones, including speech emergence and language impairments [130]. Historically, our ability to assess disruption of the developmental pathways responsible for these complications has been limited [90]. However, technological advances now permit rapid, high-throughput analysis of entire genomes, exomes and transcriptomes, with ever decreasing costs. To our knowledge this is the first study to perform high-throughput RNA sequencing on the premature newborn to better understand the molecular mechanisms and sex differences affecting oral feeding maturation. This research identified novel potential pathways involved in oral feeding, as well as important areas of differential gene expression based upon sex.

Clinically, it is well established that female infants will achieve oral feeding competency prior to males of the same gestational and post-conceptional ages [20]. Sex specific maturation of oral motor function and development has been seen as early as 15 weeks gestation [131]. By using ultrasound assessment of oral-upper airway regions in 85 healthy infants in utero, Miller *et al.*, demonstrated significant differences in the development of lingual and pharyngeal structures, as well as pharyngeal motor activity and oral-lingual movements between male and female fetuses, concluding that complex oral-motor and upper airway skills emerged earlier in females [131]. Differential maturation of other systems based upon sex has also been described [112, 126]. Female infants have been shown to reach mature lecithin/sphingomyelin (L/S) ratios, an indicator of lung maturation and appropriate surfactant production, 1.4 weeks earlier than male counterparts [126]. And, despite advances in perinatal care, males continue to suffer significantly more respiratory morbidity and mortality compared to female infants [132].

These prior studies suggest that development of males and females follow different time course trajectories and that applicable future therapies must be sex specific in order to truly improve care. However, in order to develop targeted, sex-specific therapies, we must be able to clearly demonstrate that near real-time monitoring of development can be performed in this vulnerable population. By applying RNA sequencing to noninvasively obtained saliva samples, we are able to move beyond merely reporting epidemiological associations, and delve much deeper into the biological mechanisms that are responsible for them.

When performing a combined comparative analysis of all successful and unsuccessful oral feeders, developmental systems involving the nervous, tissue and embryonic systems are all involved in oral feeding maturation. Specifically, pathways involved in cranial nerve (CN) development (CN I, III and IV), sensory integration, and facial development were all identified as being differentially expressed between successful and unsuccessful oral feeders [132]. These pathways are not only biologically relevant, but have been shown previously by our group to be essential for oral feeding. In 2015, we demonstrated that expression profiles of genes involved in olfactory (*PLXNA1)* and vision (*NPHP4*), as well as facial development (*WNT3*), could be used to predict oral feeding maturation in the newborn [15]. The prospective validation of the importance of these developmental pathways in oral feeding further substantiates their potential for personalized treatment strategies. For example, in order to expedite oral feeding and/or treat deficiencies, increased kangaroo care with parents could be utilized in infants lagging in sensory integration, while the use of the FDA-approved NTrainer System®, designed to improve nutritive sucking, could be used in infants with delayed facial development [130]. However, when separated by sex, males and females revealed distinct salivary profiles. In fact, there was no overlap in differentially expressed genes between successful and unsuccessful oral feeders when split by sex. Male infants who could not successfully feed appeared to be more affected by nervous system aberrations, particularly as it related to memory and learning, while unsuccessful female infants were more affected by structural impairments involving intestinal, tooth and palate development.

Previously, we have shown the important role of the gut-brain axis on oral feeding [87]. Maturation of the hypothalamus, and specifically genes involved in hunger signaling (e.g. *NPY2R* and *AMPK*), are required in order for an infant to demonstrate a mature oral feeding profile. Those prior studies utilized both gene expression microarrays and reverse-transcription quantitative polymerase chain reaction (RT-qPCR) to associate gene expression in neonatal saliva to feeding outcomes. However, through RNA sequencing analysis we have identified, for the first time, genes and pathways associated with the learning process of oral feeding itself. Through IPA analysis, unsuccessful male feeders demonstrated associations with aberrant expression patterns involved in the neurogenesis of the hippocampus, as well as abnormal migration and morphology of the Cajal-Reelin neurons [122, 133]. Reelin secreting neurons are located in the marginal zone of the neocortex and hippocampus and have been an area of investigation in the setting of memory disorders, including Alzheimer’s disease [134]. Furthermore, unsuccessful male feeders exhibited associations with abnormal morphology of hippocampal CA1 regions that are required for contextual memory retrieval and detailed episodic memories, as well as abnormal myelination [128].

Conversely, IPA analysis showed associations for unsuccessful female oral feeders' gene expression patterns with abnormal hard and secondary palate development, as well as disrupted morphology and neurogenesis of the intestine [94]. None of these infants displayed palate malformations. Rather, these data suggest a delay of infant palate maturation, an essential component for proper oral feeding. Of note, female infants were also shown to have dysregulation of pathways involved in hematopoiesis and immune response [115]. While it is possible that genes within these pathways play a role in oral feeding or gut development, it should be noted that the most common cell types in human saliva are epithelial cells, leukocytes and erythrocytes [35]. While future research will need to be conducted to determine what roles, if any, these pathways may play in oral feeding maturity, it is possible that differential cell counts within the whole saliva samples collected may have resulted in these findings. Nevertheless, these data suggest that premature newborns may have sex-specific, time sensitive and distinct maturation of developmental pathways that lead to oral feeding success. Understanding these differences is essential to developing and implementing targeted and personalized therapies to improve feeding and associated long-term outcomes.

One of the most significant aspects of this research is that the data were derived from noninvasively obtained saliva samples. No infant in this study was subjected to invasive phlebotomy, experimentation or testing. Rather, comprehensive gene expression profiles were analyzed based upon an average of 10 μL to 20 μL of noninvasively collected whole saliva. Previously, we have demonstrated that saliva is one of the ideal bio-fluids to explore near real-time development in the premature newborn, particularly on microarray and RT-qPCR platforms [135]. Despite the promise of salivary diagnostics, there are unique considerations when utilizing whole saliva on the RNA sequencing platform. These considerations include the impact of cellular material on sequences reads as well as the impact of microbial debris. Despite attempts in this study to perform a targeted sequencing analyses of cell free RNA, we saw a wide range of sample alignment to hg19, with a mean rate of 44%. In comparison, alignment rates for human datasets typically exceed 90%. It is probable that the varying rates of alignment seen in this study may be attributed to the degraded quality of RNA found in saliva, varying cellular contents, small sample volumes, and bacterial sequences present in the saliva as part of the oral microbiome. Investigators must be cautious of these variables when performing and interpreting RNA sequencing data derived from saliva samples.

Other limitations to this study include the small sample size and lack of ethnic diversity in subjects. As such, it is unlikely that these data can be universally applied to all newborns across PCAs. Nevertheless, our findings continue to contribute to a body of literature demonstrating the developmental complexity and sex-specific pathways involved in oral feeding maturation in the neonate.

In conclusion, RNA sequencing of neonatal saliva to assess oral feeding competency is feasible, informative and provides near real-time information regarding ongoing development, as well as aberrant gene regulation, in the preterm infant. While it has been long known clinically that males and females have a distinct timeline for the development of oral feeding maturation [130], integrating gene expression sequencing platforms into neonatal care will allow us to better understand these differences at an organ system level. Data generated may be used to develop targeted and personalized treatment strategies for the millions of infants affected by oral feeding difficulties born each year. Importantly, this proof of principle study lays the foundation for noninvasive assessment of a multitude of other morbidities affecting the preterm newborn and holds significant promise for improving care and outcomes for this vulnerable population.

Table 4.1. Patient Demographics

Subgroups	Numbers
Sex	
Females	16
Males	12
Feeding Status	
Feeders	14
Non-Feeders	14
Age (weeks)	
PCA	34 1/7 – 39 1/7
GA	29 0/7 – 35 4/7
Ethnicity	
Caucasian	25
Hispanic	2
Unknown	1

Table 4.2. List of differentially expressed genes (DEG) identified between all feeders vs. non-feeders, only the female cohort and only the male cohort

COHORT	DIFFERENTIALLY EXPRESSED GENES
ALL INFANTS	ACP5, ARSD, BCYRN1, BLM, BRI3BP, C16orf93, C6orf226, C9orf93, CCR4, CENPL, CPA4, DBP, EFNB1, FAM83D, FOXO3, GIPR, GJA9, GPR22, HIST1H3H, IMPG1, JAKMIP1, KANK3, KDR, KRI1, LAMC1, LOC100130954, LOC100506321, LOC100506688, LOC100652999, LOC283404, LOC550112, LOC646278, MLXIPL, MMP17, MPI, MUC20, NAGPA, NR6A1, NUP35, OR8U1, PAQR4, PAQR6, PARP3, PLEKHA1, PPIL6, PSORS1C2, PVRL3, RASD1, RRP7B, SH3BP5L, SIRT2, SLC4A4, SMOX, SNORA6, SYNPO2, TCTN2, TIGIT, TMPRSS11BNL, VSIG4, ZNF324B, ZNF382, ZNF699, ZNF714
FEMALES	ABHD12, ABTB2, ACD, ACP5, ALPPL2, ATG9B, BCL2, BIRC3, C14orf129, C19orf54, CD28, CENPL, CEP70, CLEC18B, CMKLR1, COMMD3-BMI1, CPA4, CTTNBP2NL, DNAH8, DNAJB7, EEPD1, EML3, EN1, FAM22G, FAM3D, FBXW4P1, FST, GJB4, HAP1, HIST1H2BF, HIST1H4J, HSD11B1L, HSF4, IFT140, KCNC2, KIAA1239, KLF8, LOC100287015, LOC100506136, LOC100652999, LOC401093, LOC441454, LOC550112, LOC645513, LOC728377, MAPK10, MKRN3, NAGPA, NCS1, NIPAL1, NUP35, OGDH, OLFML2B, OR8U1, PAQR6, PDCD1, PHF13, PHLPP2, PI4K2A, PLTP, PPIL6, PRKCH, PRSS8, PTK6, PVRL3, RAD52, RAD54L, RASD1, RBM43, SAPCD2, SLC35A2, SLCO2B1, STIL, SULT1A2, THOC3, TMEM102, TNFSF4, TP53INP2, VWA1, ZNF235, ZNF280C, ZNF382, ZNF414, ZNF594, ZNF714, ZNF827, ZNF83, ZNRF1
MALES	AACS, ARHGEF12, BLM, BOLA3, C6orf132, CCBP2, CCDC137, CCDC14, CDH13, CHML, COQ4, CROCCP2, DBC1, DCLRE1B, EMG1, EML5, FAM83D, FAM84B, FOXO3, GLRX2, GPM6B, GPR125, GPR22, GSG1, HEATR3, HSH2D, KLF3, LOC100128590, LOC100130451, LOC100507299, LOC152217, LOC202181, LOC642236, LOC84989, MAP1S, METTL15, MOCS3, MRPL41, MUM1, MYL9, NACA, NDUFA12, NDUFA7, NDUFAB1, NGEF, NPIPL3, OPLAH, PAK1IP1, PLEKHA5, PNPLA7, POLR3E, PRICKLE3, RBM41, RGAG4, RNF26, RPP38, S100A13, SEPX1, SHQ1, SLC35E3, SLC39A13, SMOX, SNORD21, SORBS3, TGIF1, THYN1, TIGD7, TM6SF1, TMEM160, TMPRSS11BNL, TNFRSF21, TRMT2A, UBA7, XPC, ZFPL1, ZNF32, ZNF600, ZNF613

Table 4.3. Read alignment/mapping rates for all infants

Female Feeders	Mapping Rate	Female Non-feeders	Mapping Rate	Male Feeders	Mapping Rate	Male Non-feeders	Mapping Rate
F1	55.40%	**NF1**	21.90%	**F9**	42.90%	**NF9**	19.60%
F2	20.20%	**NF2**	43.40%	**F10**	42.60%	**NF10**	24.80%
F3	5.90%	**NF3**	48.90%	**F11**	41.40%	**NF11**	51.10%
F4	16.10%	**NF4**	40.60%	**F12**	16.30%	**NF12**	7.30%
F5	28.00%	**NF5**	32.80%	**F13**	34.90%	**NF13**	37.50%
F6	93.60%	**NF6**	17.40%	**F14**	4.50%	**NF14**	22.40%
F7	43.70%	**NF7**	94.30%				
F8	50.10%	**NF8**	39.60%				

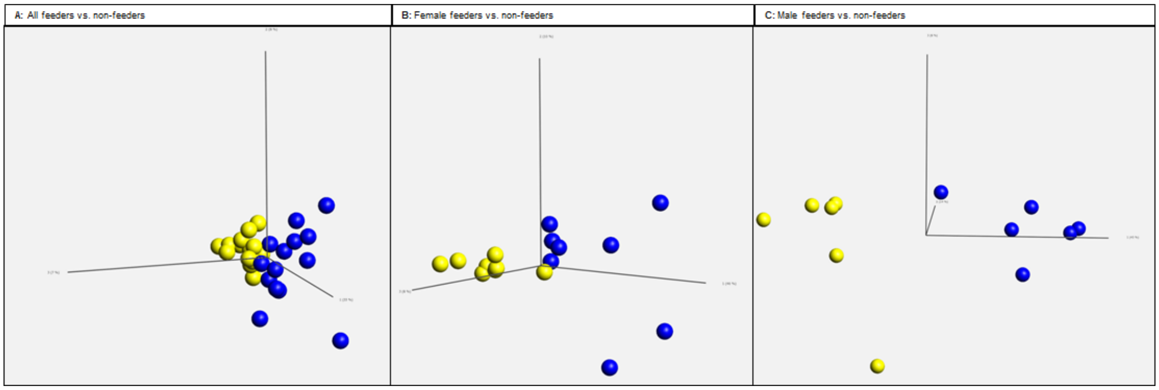


Figure 4.2: Principal component analysis (PCA) of successful feeders vs. unsuccessful feeders. (A) All infants recruited in the study (B) Females only and (C) Males only.

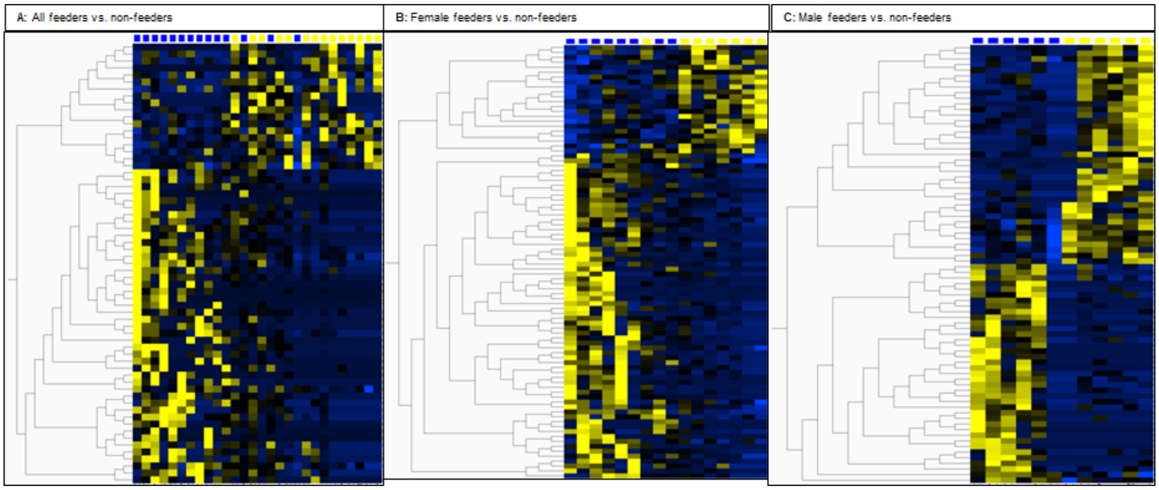


Figure 4.2: Heat maps of successful feeders vs. unsuccessful feeders. (A) All infants recruited in the study (B) Females only and (C) Males only.

Chapter 5: Optimal reference genes for RT-qPCR normalization in the newborn

Optimal reference genes for RT-qPCR normalization in the newborn, P Khanna, KL Johnson, JL Maron, Biotechnic & Histochemistry, 1-8

Reprinted here with permission of publisher.

5.1 Introduction

The Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines were established in 2009 to produce more consistent results and increased experimental transparency [16]. MIQE dictates the use of multiple reference genes for relative gene expression quantification to ensure data integrity and to allow for reliable comparative analyses between publications [16, 17]. To date, no set of ideal reference genes has been established for use in the neonatal population [136].

Newborns present unique challenges to investigators. Their vulnerable physiological state, combined with rapid and ongoing development, is distinct in the human lifespan [137]. For nearly a decade, our laboratory has utilized gene expression platforms to explore physiological and aberrant development in the neonate [15, 87, 110, 135]. Specifically, we have evaluated gene expression changes in neonatal saliva. This research is dependent upon normalization to control for variability in initial salivary volume collected as well as total RNA concentration in each sample. Proper normalization of a reverse transcription - quantitative polymerase chain reaction (RT-qPCR) assay involves reporting the ratio of the mRNA concentration of the gene of interest to that of the reference genes [138-140]. Normalizing RT-qPCR data to genes with inherent variability generates background noise and affects the accuracy of quantified mRNA expression levels [141].

RT-qPCR studies commonly use the reference genes beta-actin (*ACTB*), glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) and tyrosine 3-monooxygenase/tryptophan 5-

monooxygenase activation protein zeta (*YWHAZ*) [142-145]. We studied the gene expression levels of these three reference biomarkers in neonatal saliva to determine their relative variability and applicability in this patient population. Saliva was collected across the preterm and term neonatal age spectrum for both sexes to specifically explore their expression levels during rapid development and to identify potential sex-specific gene expression differences.

5.2 Materials and Methods

5.2.1 Saliva Collection

This study was approved by the Tufts Medical Center Institutional Review Board. Informed parental consent was acquired for all enrolled infants. Saliva samples were obtained using techniques described previously [110]. Saliva was immediately stabilized in 500 μL of RNAProtect Saliva (Qiagen), vortexed and stored at 4°C pending total RNA extraction with the use of RNeasy Mini Kit (Qiagen) per manufacturer's instructions. All samples underwent DNAse treatment to eliminate potential DNA contamination. Extracted total RNA samples (14 μL) were stored at -80°C prior to gene expression analyses.

5.2.2 Characteristics of neonate study population

There were 298 infants who provided a total of 400 saliva samples for this study. A subset of infants provided more than one saliva sample for analysis, collected at varying post-conceptional ages (PCA). Thus, each sample represented a specific moment in development and was considered individually. There were 217 male and 144 female samples with a PCA range of 32 to 48 weeks.

5.2.3 RT-qPCR Platform Experiments

The RT-qPCR assay was designed to adhere to the established MIQE guidelines [16]. All primer probe sets were inventoried at Life Technologies (Carlsbad, CA USA). Each spanned exons; only *GAPDH* had a known SNP contained within the amplicon. Following total RNA extraction, samples were converted to cDNA using the SuperScript VILO cDNA Synthesis kit (Life Technologies) per manufacturer's protocols. Subsequently, samples underwent a targeted pre-amplification using a pooled custom assay mix prepared by Life Technologies based upon their inventoried amplicons. Following pre-amplification, samples were run on the Applied Biosystems ViAA 7 RT-qPCR platform with the following thermal cycle profile: uracil-N-glycosylate was incubated at 50°C for 2 minutes followed by 40 cycles of denaturation at 95°C for 1 second and annealing and extension at 60°C for 20 seconds. Negative and positive controls were run on each plate. All samples were run in duplicate.

Normalized Ct values were calculated for all three genes *ACTB (*G1*)*, *GAPDH (*G2*)* and *YWHAZ (*G3) and for all samples analyzed in the study (n) using the raw Ct values and their geometric mean (GM) with the following formula (example shown below is for *ACTB*):

$$\Delta \mathrm{Ct}\ (\mathrm{G1}\ n) = \mathrm{Ct}\ (\mathrm{G1}\ n) - \mathrm{GM}\ [\mathrm{Ct}\ (\mathrm{G2}\ n) + \mathrm{Ct}\ (\mathrm{G3}\ n)]$$

5.2.4 Statistical Analyses

Statistical analyses were performed using SPSS with significance set at p < 0.05. Descriptive statistics were calculated for each gene (mean ± standard deviation). Scatter plots of the normalized Ct values for each gene were generated. To determine variability in expression across the age spectrum, the homogeneity of variance was calculated using the Levene's test, embedded as a One-Way ANOVA. Data were further analyzed based upon sex.

5.3 Results

5.3.1 Variability in reference genes

ACTB showed greatest variation when the normalized Ct values were plotted against the PCAs of the infants, while both *GAPDH* and *YWHAZ* displayed a more compact expression pattern (Fig. 5.1). The same trend was observed when the normalized Ct values were plotted as a boxplot (Fig. 5.2) where *ACTB* had a wider range and larger standard deviation bars compared to *GAPDH* and *YWHAZ.* Comparing the three reference genes in the total population generated a statistically significant variance (p < 0.0001), consistent with the trend observed in the plots. Pairwise comparisons between the targets showed statistically significant variation (p < 0.0001) for both *GAPDH* and *YWHAZ* when paired with *ACTB.* The Levene's test comparing expression values for *GAPDH* versus *YWHAZ* was not statistically significant (p = 0.068) (Table 5.1).

5.3.2 Sex differences in reference gene performance

Splitting the data by sex and plotting the normalized Ct values against the PCAs of the infants showed a wider spread and greater variation in males compared to females (Fig. 5.3). Homogeneity of variance was also performed on the data based on sex (Table 5.1). In females, significant variance ($p < 0.040$) was only observed when comparing *ACTB* versus *GAPDH*. For males, there was an overall significant difference ($p < 0.0001$) in variance across all targets, but also when comparing *ACTB* vs *GAPDH* and *ACTB* vs *YWHAZ*. The only pairwise comparison for males that was not statistically significant was *GAPDH* vs *YWHAZ* ($p = 0.560$).

Descriptive statistics in SPSS revealed that the standard deviation for *ACTB* was the greatest at 4.37, which was confirmed as the high variance trend in the gene expression plots. In comparison, *GAPDH* standard deviation was 2.31 and *YWHAZ* was 1.43, (Table 5.2).

5.3.3 RT-qPCR Performance

There was sample loss of approximately 10% (N=39), defined as the failure of an individual sample to amplify all three reference genes [135]. Ultimately, 361 samples were used for the analyses from the total cohort of 400. No amplification occurred in the negative control wells. The positive control amplified successfully on each plate.

5.4 Discussion

Our study aimed to identify reference gene targets that could serve as valid markers for appropriate normalization of RT-qPCR data in the newborn. Three of the most commonly used reference genes cited in the literature for gene expression normalization are *ACTB*, *GAPDH*, and *YWHAZ* [142, 146-148]. *ACTB* is a highly conserved cytoskeleton protein involved in cell motility, structure and integrity [149]. *GAPDH* is essential for glucose metabolism, transcription activation, initiation of apoptosis, endoplasmic reticulum to golgi vesicle shuttling, and fast axonal or axoplasmic transport [150, 151]. *YWHAZ* is central to cell survival and plays a key role in a number of cancers and neurodegenerative diseases [106, 115, 152, 153]. Presumably, these genes would maintain their expression throughout development given their essential roles in cell survival. However, our data suggest that each gene shows variation across the neonatal age spectrum and may also be differentially expressed based upon sex. *ACTB* expression is highly variable in neonatal saliva and demonstrates sex-specific expression profiles, making it an inappropriate reference gene for relative normalization and quantification. Alternatively, *GAPDH* and *YWHAZ* serve as more stable reference gene targets across the premature and neonatal age spectrum, and *YWHAZ* specifically is more reliable and preferential for use in female cohorts.

Our data validate other investigators who have clearly shown the relative instability of *ACTB* in various populations and cell types. Despite being widely used as a reference gene, there are numerous sources citing that *ACTB* expression can change during growth and differentiation [154-157]. Studies in other organisms have also shown variation in

ACTB levels between male and female cohorts [158-162]. Similarly, our own data reveal that *ACTB* appears to be more constituently expressed in females compared to males. The overall variance in *ACTB* expression appears to be driven by the male subset in the data (Fig. 5.3). This finding suggests that *ACTB* may play a specific role in the development of a male newborn. Recently, Panahi *et al*., reported that *ACTB* was an unsuitable reference gene in saliva of male children and that the combination of *GAPDH* and *YWHAZ* was far more stable [148]. While this study examined a much smaller, single sex population (males with autism; n=9), their results confirm that *ACTB* is not a suitable reference gene, especially in developing children. Further investigation will help understand the functionality of *ACTB* in the neonatal population leading to its variable and sex-specific expression profile.

Technological advancements now permit high throughput screening of thousands of genes in a single, small volume source with extraordinary assay sensitivity. Utilizing this technology to develop a robust salivary gene expression assay could provide valuable insight into the developing newborn [15]. Neonates pose unique challenges to the exploration and interpretation of differential gene expression. Not only are most biological specimens, like blood, difficult to obtain in this vulnerable population, but newborns also undergo rapid development and growth after birth, greatly impacting gene expression changes. Neonatal salivary diagnostics has shown to yield enormous amounts of near real-time developmental information [15, 87]. Indeed, salivary diagnostics is a rapidly emerging field throughout medicine [95]. Due to its safe biohazard profile and ease of attainment, especially for longitudinal studies, salivary diagnostic platforms have been described for a wide range of medical conditions, including breast, pancreatic and oral cancers, as well as immune-mediated diseases such as Sjögren's Disease [40, 55, 95, 110, 163]. Integrating similar salivary diagnostic platforms into clinical practice to explore development and disease into the newborn population holds great promise for improving care. However, without due diligence to identify stable, constituently expressed reference genes, interpretation of gene expression data will remain limited in this population.

Strengths of this study include our large sample size that was representative of a wide range of PCAs. Furthermore, this study provides insight into those infants born prematurely, when ongoing 'fetal' organ and tissue differentiation is being directly impacted by their new aberrant environment, including noxious stimuli such as light, needle sticks and touch that directly impact gene expression [164]. Never before have investigators had access to such a comprehensive database of expression data to inform prospective trials and study design, and to examine gene expression variability of reference gene candidates in neonates across the preterm and term neonatal age spectrum.

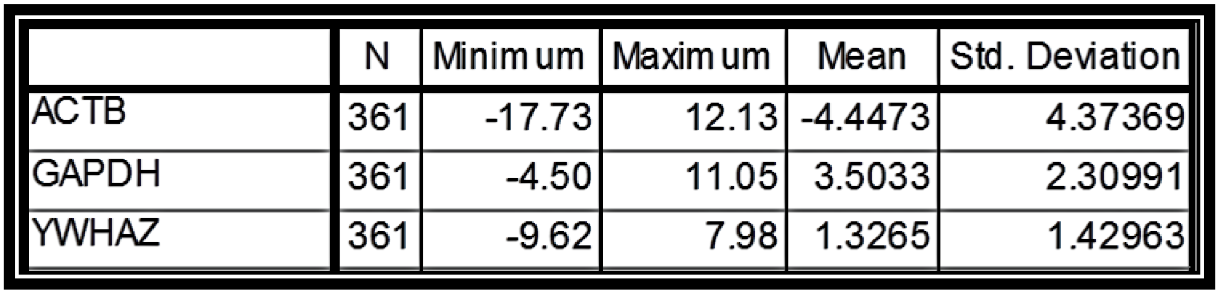
In summary, we have explored the expression profiles of three reference genes in the newborn and found two out of three reference genes to be suitable candidates for RT-qPCR normalization in this population. This study emphasizes the need to reevaluate the reference genes traditionally being used in RT-qPCR and to recognize and understand the importance of sex differences in the newborn. Failure to appreciate these unique characteristics may result in misinterpretation of data that could deleteriously impact clinical care.

Table 5.1. Test of homogeneity of variances.

Test of Homogeneity of Variances (Levene’s Statistic)			
Comparison Groups	Sig. (Total)	Sig. (Females)	Sig. (Males)
ACTB vs GAPDH vs YWHAZ	0.0001*	0.139	0.0001
ACTB vs GAPDH	0.0001*	0.040*	0.0001
ACTB vs YWHAZ	0.0001	0.315	0.0001
GAPDH vs YWHAZ	0.068	0.342	0.560

(* indicates significance level of *p* < 0.05).

Table 5.2. Descriptive statistics for the reference genes with the number of samples (N) and the minimum, maximum, mean and standard deviation values for their dCTs.



	N	Minim um	Maxim um	Mean	Std. Deviation
ACTB	361	-17.73	12.13	-4.4473	4.37369
GAPDH	361	-4.50	11.05	3.5033	2.30991
YWHAZ	361	-9.62	7.98	1.3265	1.42963

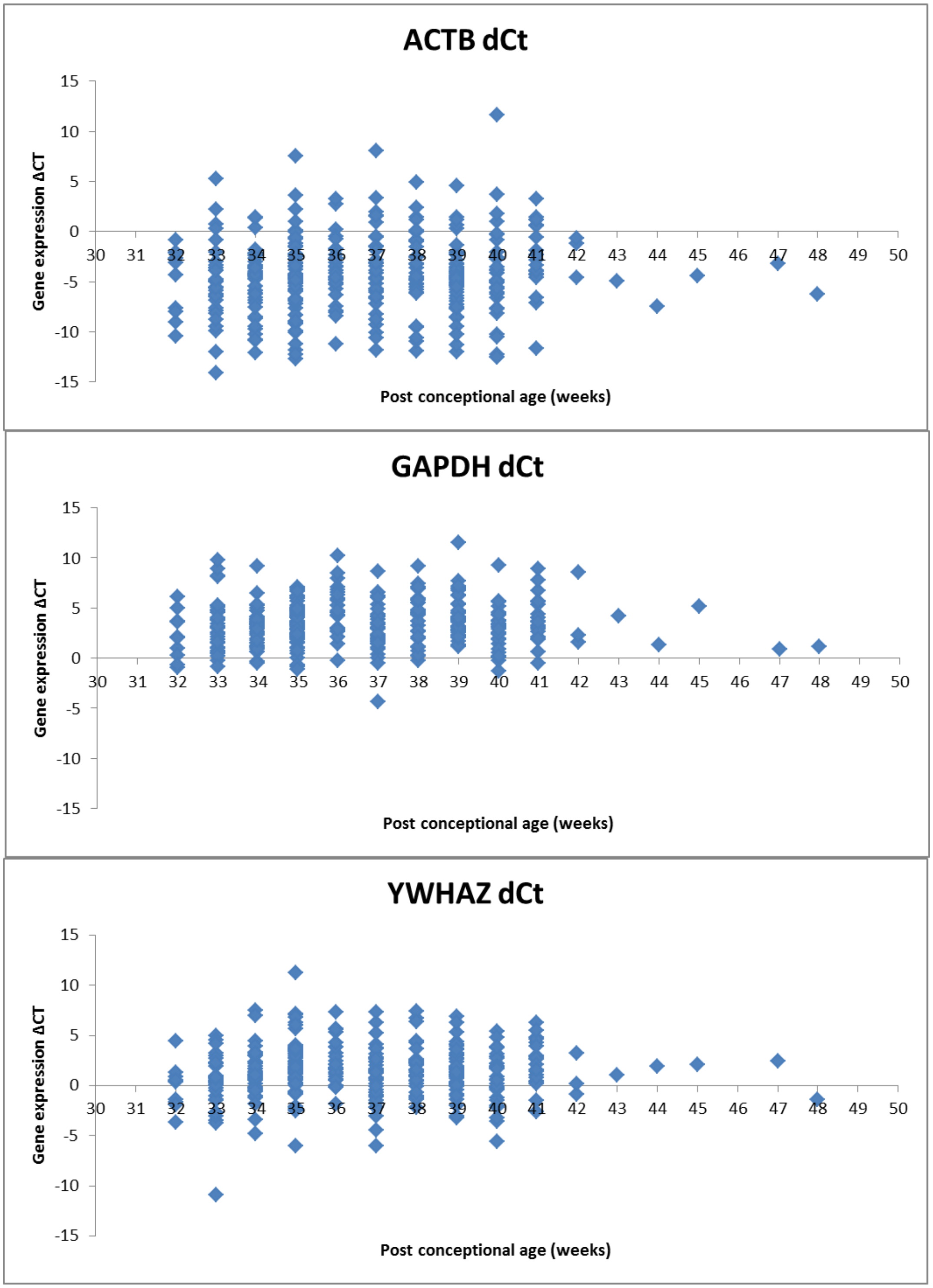


Figure 5.1: Comparison of each candidate reference gene expression level (ΔCT) with postconceptional age. For the x axis, the ages of the infants were rounded to the closest whole number (scatter plots with the raw ages are included in the supplemental data).

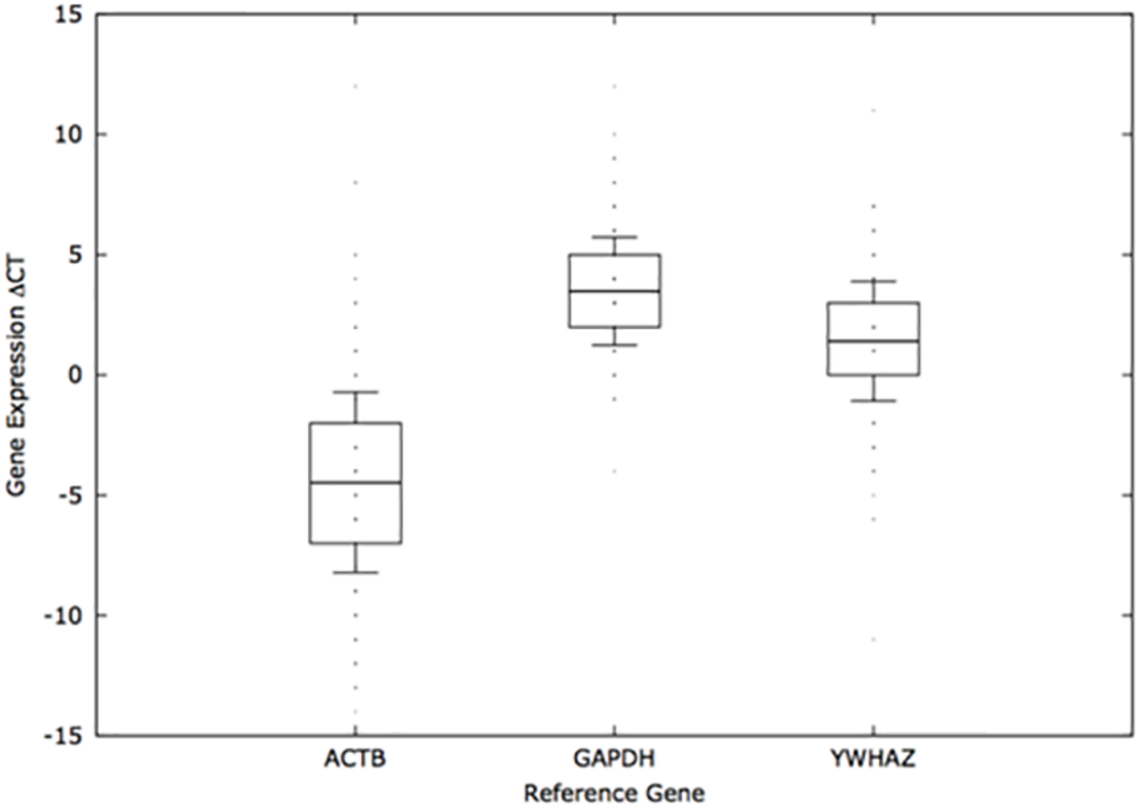


Figure 5.2: Boxplot comparing each candidate reference gene expression level (ΔCT).

The plots represent the spread of the ΔCt values by RT-qPCR with the median ΔCt values of *ACTB*, *GAPDH* and *YWHAZ* being -4.46, 3.48 and 1.41, respectively. The error bars represent the standard deviation for *ACTB*, *GAPDH* and *YWHAZ* of 4.37, 2.31 and 1.43, respectively.

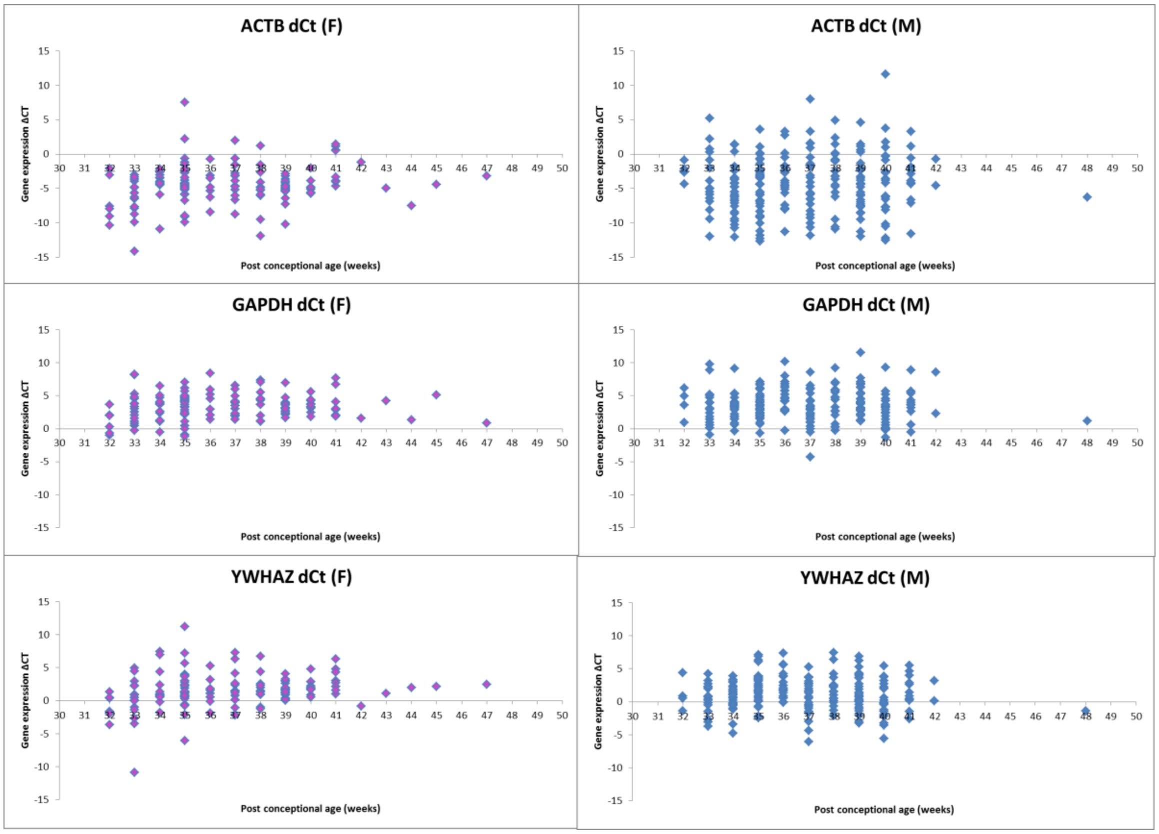


Figure 5.3: Comparison of each candidate reference gene expression level (ΔCT) with postconceptional age separated by sex.

Panel (a) shows the gene expression patterns for the females (F) and panel (b) shows the gene expression patterns for the males (M).

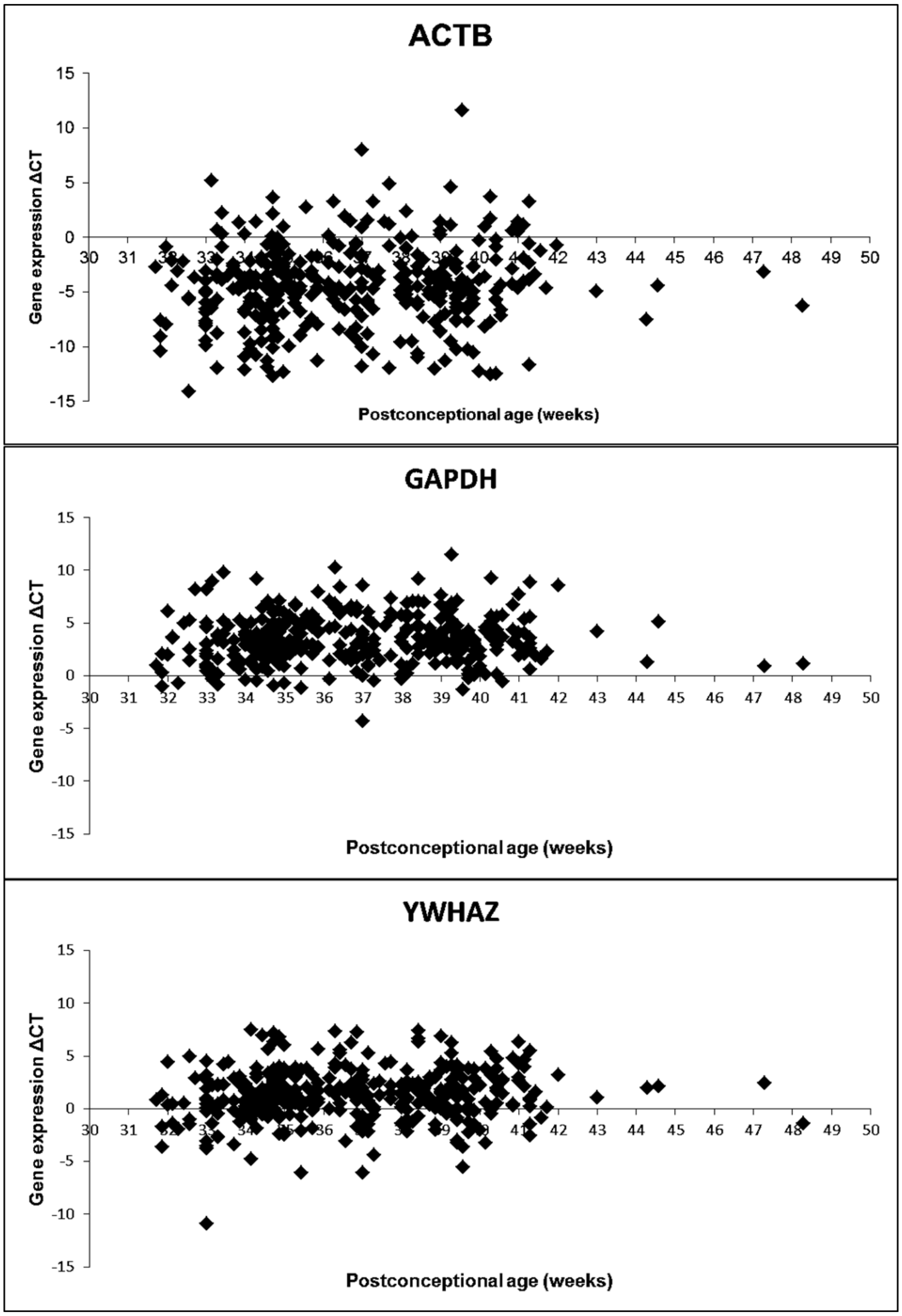


Figure 5.4: Comparison of each candidate reference gene expression level (ΔCt) with raw postconceptional age.

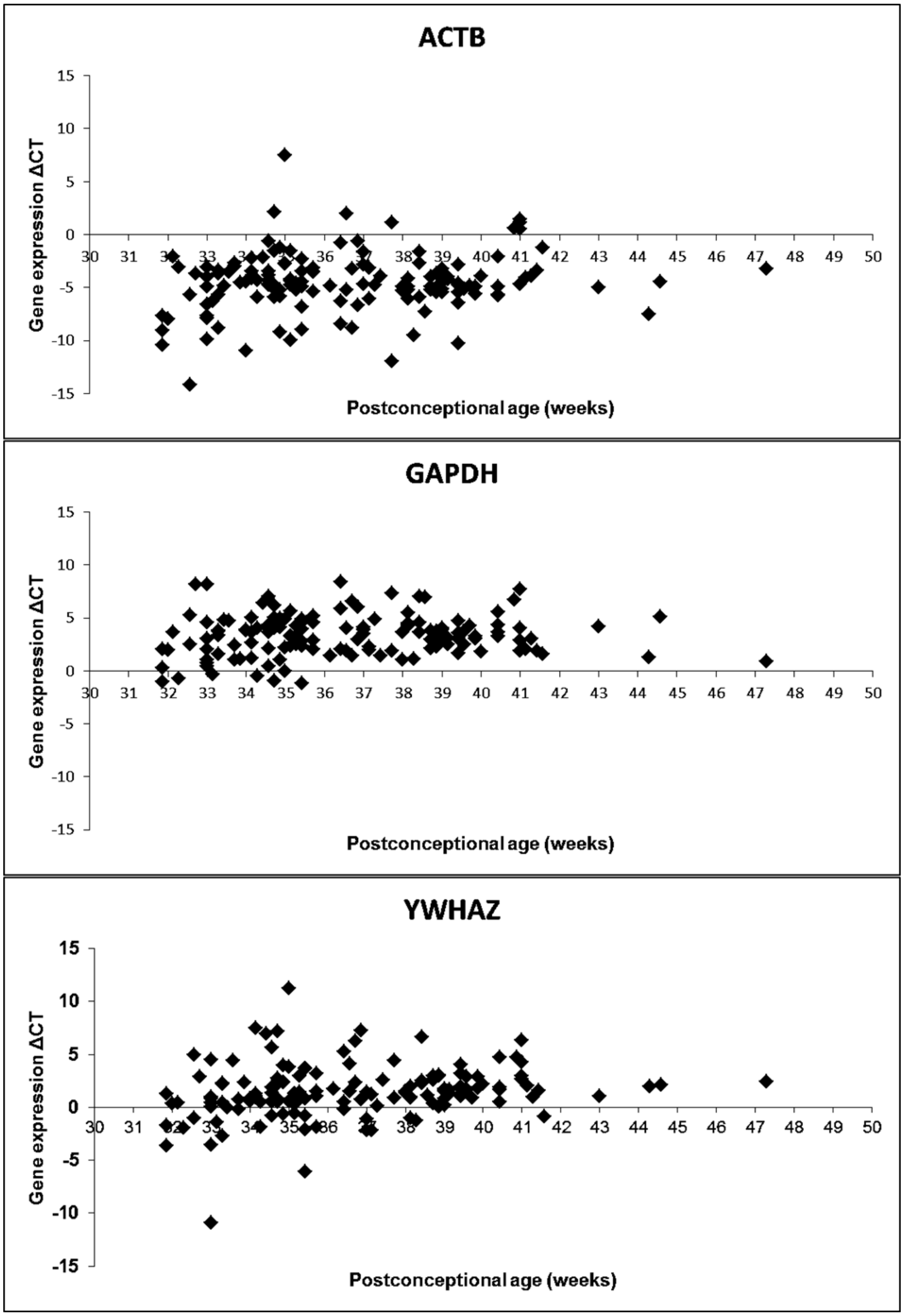


Figure 5.5: Comparison of each candidate reference gene expression level (ΔCt) with raw postconceptional age (Females).

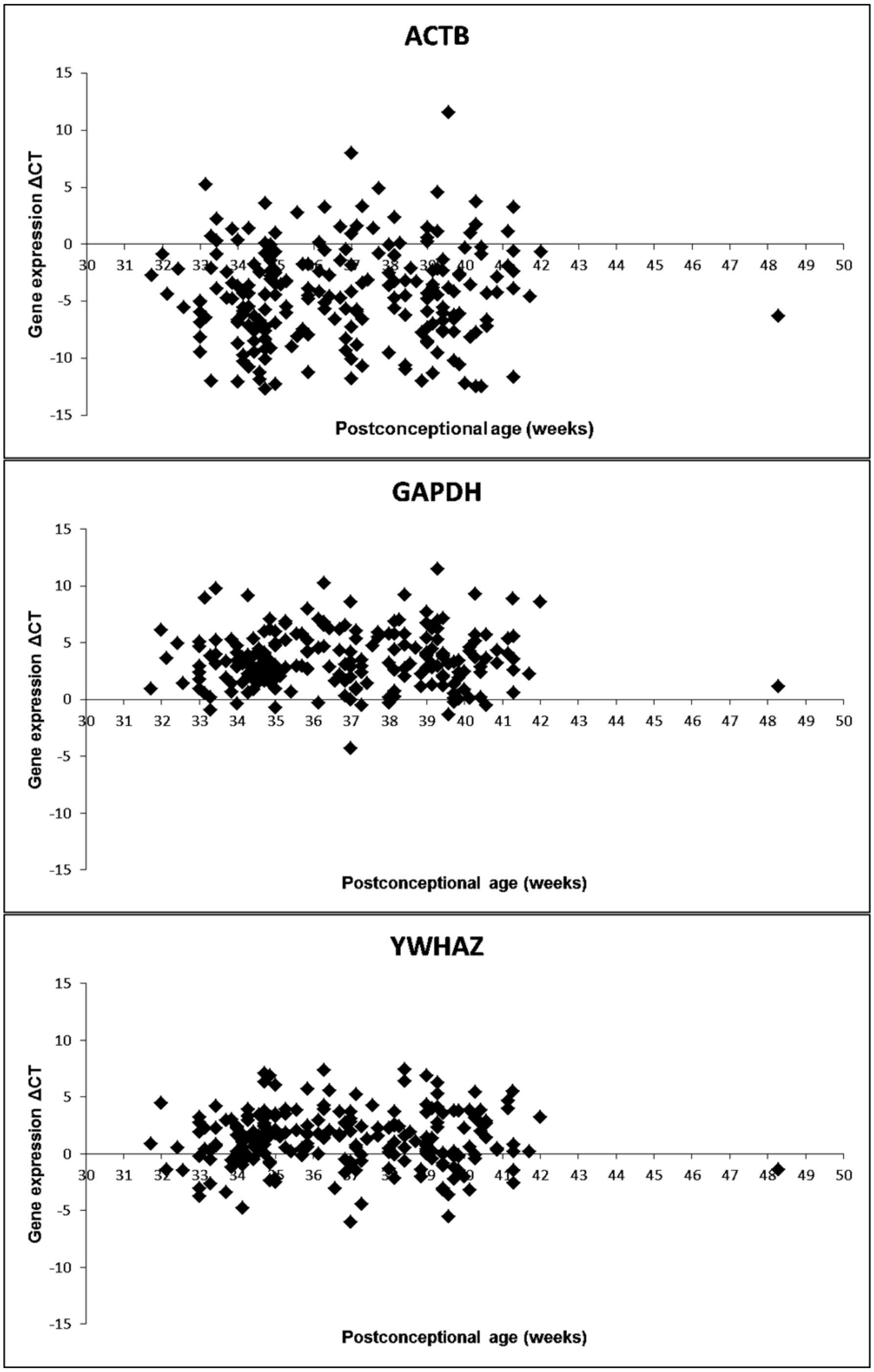


Figure 5.6: Comparison of each candidate reference gene expression level (ΔCt) with raw postconceptional age (Males).

Chapter 6: Cracking the lobster genome: *De novo* sequencing, assembly, annotation and characterization of the *Homarus americanus* genome and transcriptome

Cracking the lobster genome: *De novo* sequencing, assembly, annotation and characterization of the *Homarus americanus* genome and transcriptome. P. Khanna, A. Tai1, C. Munkholm, M. Hartley, T. Dickinson, B. Rice, M. Vierra, F. Clark, S. Greenwood, D. Walt. To be submitted to SCIENCE.

6.1 Introduction

Crustaceans are a heterogeneous group of organisms comprised of approximately 50,000 different species including both terrestrial and aquatic animals [165-167]. The American lobster, *Homarus americanus*, is not only the heaviest known crustacean and arthropod species, but it is a commercially and environmentally important inhabitant of the Atlantic coast of North America [28, 168-170]. The American lobster commercial fishery is vital in both the US and Canada with many coastal communities relying on the economic benefits of this fishery [169]. Diseases in marine organisms are on the rise, and recent studies have highlighted the American lobster as a model system for the study of behavior, diseases, ecology and fisheries [171, 172]. Additionally, shell diseases highlight the complexity of the dynamic internal and external environments of lobsters [173].

The American lobster is a polyploid organism with an estimated genome size of 4.75 Gb [174]. To date, no complete genome sequence is available for the American lobster due to its large genome. Additionally, the American lobster's genome contains repetitive DNA sequences presenting further technical challenges for sequence alignment and assembly. For many crustacean species, some portions of DNA and RNA sequence data exist, such as expressed sequence tag (EST) libraries and mitochondrial genomes, but there is no complete and well-defined crustacean genome currently available [29]. Crustacean genetics are relatively undeveloped compared to the genetics of most other organisms. Thus, there is a distinct need for a well sequenced and annotated genome to provide a foundational base to move this field forward.

This study provides the first complete sequenced, annotated, and characterized genome for *Homarus americanus* and delivers a reference genome for understanding crustacean biology. Unraveling the genome can facilitate our understanding of the lobster's growth and development, and provide a foundation for lobster health and sustainability issues. It may also provide insight into the relationship between environmental factors and epistatic interactions among genes that lead to environmental challenges and give rise to complex traits such as disease susceptibility [175].

6.2 Materials and Methods

6.2.1 .Chicago library preparation and sequencing for genome

Four Chicago libraries were prepared as previously described [176]. Briefly, for each library, 500ng of HMW gDNA (mean fragment length = ~50kbp for libraries 1-3 and ~150kbp for library 4) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with MboI, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq 2500 (rapid run mode). The number and length of read pairs produced for each library was: 177 million, 2x100bp for library 1; 179 million, 2x150 bp for library 2; 84 million, 2x 150bp for library 3; and 126 million, 2x 100bp for library 4. Together, these Chicago library reads provided 30x physical coverage of the genome (1-50kb pairs).

6.2.2 De Novo Assembly of the Homarus americanus genome:

A *de novo* assembly was constructed using a combination of paired end (mean insert size ~500bp) and mate pair (insert sizes of 3.4 -8.7 kb) libraries. *De novo* assembly was performed using Meraculous 2 (2.0.2) [177] with a kmer size of 61. The input data consisted 950 million read pairs sequenced from paired end libraries (totaling 341 Gbp) and 59 million read pairs sequenced from mate pair libraries (spanning 322 Gbp). Reads were trimmed for quality, sequencing adapters, and mate pair adapters using Trimmomatic [178].

6.2.3 Scaffolding the de novo assembly with with HiRise

The input *de novo* assembly, shotgun reads, and Chicago library reads were used as input data for HiRise, a software pipeline designed specifically for using Chicago data to scaffold genome assemblies [176]. Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins and to score prospective joins. After scaffolding, shotgun sequences were used to close gaps between contigs.

6.2.4 Genome annotation

The MAKER v2.31.6 pipeline was used for annotating the assembled genome [27]. Initially, genes were predicted using ab initio to predict genes. Next, genes were predicted using both SNAP and Augustus v2.5.5, both trained with a hidden-Markov model developed from the predictions of the initial maker step. This second step also included the EST-based evidence as described in the first step. The *de novo* assembled transcriptome from Trinity was also used as the EST-based evidence for the gene prediction with MAKER2. All runs of MAKER2 included the masking of repetitive regions using repeatmasker v4.0.3 against the repbase v19.07 library. For each gene prediction, evidence was selected with an annotation edit distance (AED) < 0.75.

Using the longest isoform for each protein sequence, each gene was functionally annotated using a combination of blastp v2.2.30 (http://ncbi.nlm.nih.gov/blast) and interproscan 5.7.48. Blast searches were performed against protein sequences from the Swiss-Prot database with an e-value cut-off of 10−3, and only the top 20 hits were retained. Annotations were stored as Gene Ontology (GO) terms for each sequence.

6.2.5 Transcriptome library preparation, sequencing and assembly

Stage IV lobster larvae, previously preserved in RNAlater®, were thawed and residual RNAlater® was removed. Larval lobsters were homogenized in 1 mL of RNA preservation reagent (1.4 M guanidine isothiocyanate, 38% phenol (pH 4.5), 5% glycerol and 0.1 M sodium acetate) using an Omni Homogenizer (Omni International, Kennesaw, GA, USA) followed by the addition of 200 µL of chloroform, vigorous inversion and a 3 min incubation at room temperature. The homogenate was centrifuged at 15,000g for 15 min at 4°C. The resulting aqueous phase was removed and mixed, via inversion, with an equal volume of 100% ethanol. RNA was isolated from the resulting solution using RNeasy spin columns (Qiagen, Toronto, ON, Canada) using the manufacturer's instructions and the optional DNaseI on-column digestion. RNA was quantified using spectrophotometry (NanoDrop ND-1000, Thermo-Fisher, Ottawa, ON, Canada) and its quality was assessed using an Agilent Bioanalyzer 2100 and the RNA 6000 nano cartridge (Agilent Technologies, Mississauga, ON, Canada). RNA samples were stored at -80°C until ready for use.

High quality RNA from 40 stage IV larval lobsters was sent to Genome Quebec (Montreal, PQ, Canada) for TruSeq RNA library preparation (Illumina, Victoria, BC, Canada) according to the manufacturer's instructions. Paired-end 100 sequencing was performed on a HiSeq2000 instrument (Illumina, San Diego, CA, USA). Sequencing reads were filtered by removal of all reads with overall quality scores of < 36.

Raw sequence reads were generated using the Illumina CASAVA pipeline with base quality encodes in phred 33. Reads were trimmed using Trimmomatic software [178] to remove Illumina sequence adapters and bases with phred scores of 30 or less. Minimal read length for sequences was set at 32 nt. Trimmed reads were normalized using Trinity normalization utility [179] prior to the *de novo* generation of the *H. americanus* transcriptome using the Trinity assembler [26, 179].

6.2.6 Transcriptome annotation

Annotation was performed using Blast2go v3.2 and the TRINOTATE pipeline (https://trinotate.github.io/). All assembled genes were searched against several databases (the NCBI (non-redundant) protein database (Nr) (ftp://ftp.ncbi.nih.gov/blast/db/ 29-02-2015), Swissprot-Uniprot database, Kyoto Encyclopedia of Genes and Genomes (KEGG), GO (Gene Ontology), EggNog and InterproScan) using BlastX with an E-value cut-off set to 10−5. Gene open reading frames (ORFs) were predicted using Transdecoder v2.0.1 (http://transdecoder.sourceforge.net/). Only predicted ORFs that were at least 100 amino acids long were retained, whether partial or complete. Obtained ORFs were blasted using BlastP against the NCBI Uniref90 database with an E-value cut-off of 10−6. The remaining functional annotation was achieved using Blast2GO and TRINOTATE. The TRINOTATE pipeline uses several software: Hmmer v3.1b1, a protein domain identification (PFAM) software, Tmhmm v2.0c prediction of transmembrane helices in proteins, Rnammer v1.2 to predict ribosomal RNA, SignalP v4.1 predict signal peptide cleavage sites, prediction of gene ontology GOseq, Eggnog v3.0 search for orthologous group. The gene completeness of the assembled transcriptome was assessed using the BUSCO (Benchmarking Universal Single-Copy Orthologs) library (http://busco.ezlab.org/). Blast2GO uses the KEGG database and InterProScan software.

6.3 Results and Discussion

6.3.1 Homarus americanus genome and transcriptome statistics

For the first time, we report the complete, assembled, annotated and characterized genome sequence and transcriptome for *Homarus americanus*—the North American lobster. Brain tissue from the North American lobster was used to sequence the genomic DNA and generate a completed HiRise assembly. Conventional de novo assembly (shotgun data) was combined with Dovetail data to be scaffolded (genomic sequence consisting of contigs that have been ordered and oriented relative to each other) by Dovetail's "HiRise" software pipeline. While, conventional assembly processes typically result in assemblies with N50's (scaffold length such that the sum of the lengths of all scaffolds of this size or less is equal to 50% of the total assembly length) much less than a megabase, after Dovetail scaffolding the resulting assembly N50's were much larger than a megabase (Table 6.1). Approximately 30X physical coverage of the Chicago libraries was achieved with a scaffold N50 of 908kb (a 38-fold increase over the 24kb starting assembly). The lobster genome will provide a scientific foundation for all crustacean species. The MAKER2 and TRINOTATE pipelines were used to annotate the genome and transcriptome, respectively. Approximately 37,000 genes were identified from the genome and 24,000 from the transcriptome with a 70% to 80% concordance between the genome and transcriptome (Table 6.1).

The *H. americanus* is known to be a is a polyploid organism with an estimated genome size of 4.75 Gb. Polyploidy is vital to eukaryotic genomes by providing opportunities for functional divergence between duplicated genes [180]. The sequencing data genome revealed its size to be approximately 1.5 Gb. The 70%-80% concordance of the genome with the transcriptome confirms complete sequencing of the genome hinting towards a triploidy genome. While, there is no certainty even with the supplementing long read sequencing data, further investigation is warranted. Additionally, triploidy genomes have been linked with having an advantage in surviving subarctic temperatures, especially in crustaceans [30, 181].

6.3.2 BLAST analysis of top-hit species

All annotated genes were blasted to identify the top hits for species with the highest homology and the top 20 species were plotted (Fig. 6.1). *Scylla olivacea* was discovered as the most homologous to the *Homarus americanus* genome, while *Zootermopsis nevadensis* (Dampwood termite) was second. Interestingly, when the annotated transcriptome genes were blasted, the Dampwood termite emerged as the most homologous to the lobster transcriptome (Fig. 6.5). This result is in accordance with the long-established fact that crustaceans are the closest relatives to insects, with both groups evolving from a shared pancrustacean ancestor [29]. Unfortunately, these data are biased towards species with already sequenced genomes and transcriptomes; therefore, species whose genomes have not been sequenced may exhibit a higher degree of homology. While many of the homologous organism top-hits were expected, as they were aquatic species, some other unexpected organisms of interest along with the Dampwood termite included *Pediculus humanus subspecies Corporis* (Human lice), *Stegodyphus mimosarum* (African social velvet spider) and *Tribolium castaneum* (Red flour beetle).

6.3.3 Functional gene annotation via gene ontology

Studying the gene ontology, we classified the functions into three categories: Biological processes, Cellular components, and Molecular functions (Fig. 6.2). Overall, the terms "regulation of transcription", "metabolic process", "integral components of the membrane", "ATP binding" and "nucleic acid binding" were the most abundantly represented. To further evaluate the reliability of the genome annotation process, the same functionality analysis was conducted for the transcriptome also revealing "ATP binding" as the top-hit for the Molecular function category. "Translation" emerged as the top-hit for Biological processes and "cytoplasm" for Cellular components (Fig. 6.6). These annotations provide a valuable resource for investigating specific developmental pathways during crustacean research.

6.3.4 Targeted functional analyses

Following a comprehensive analysis of the gene ontology functional classifications, we delved further into more specific functionality analyses based on publications about crustacean species. The immune system is of particular interest for understanding crustacean physiology due to the unique environmental and pathogenic pressures crustaceans face [182, 183]. For lobster fisheries management and distribution in the food chain, the ability to identify the crustacean genes responsible for disease-resistance and susceptibility requires an in-depth knowledge of crustacean innate immune systems with assigned functional annotations. We were able to identify 16 different biological processes associated with the immune system (Fig. 6.3) and construct a list of 23 genes with assigned functional annotations (Table 6.2).

Subsequently, we focused on chitin, a component of the cell walls of crustaceans known to play an important role in shell disease [184]. Over the past decade, the prevalence of shell disease has increased drastically and a complete functional annotation of the different processes of chitin would provide scientists with a foundational basis to better understand the progression and development of this disease [185, 186]. We identified five different biological processes associated with chitin (Fig. 6.4) and 108 unique genes involved in these five processes (Table 6.3). Identification of these functionally annotated genes from the *Homarus Americanus* genome provides the opportunity to address gaps in our understanding of crustacean biology.

Table 6.1.

Summary of the assembly and annotation statistics.

Genome Size	36,929 genes	
Transcriptome Size	24,334 genes	
Genome-Transcriptome Concordance	70-80%	
	Input Assembly	**Dovetail HiRise Assembly**
Longest Scaffold	616,323	11,769,346
Number of scaffolds	205,311	39,404
Number of scaffolds > 1kb	205,065	39,322
Percentage of genome in gaps	20.72%	21.63%

Table 6.2.

GO identification of biological processes associated with the immune system and a list of associated genes.

Genes	Function
E2RMC6	activation of innate immune response; cellular response to exogenous dsRNA; cyclic nucleotide biosynthetic process; positive regulation of defense response to virus by host
A0A131Z9D6	activation of innate immune response; positive regulation of type I interferon production
H0W0D1	adaptive immune response; defense response to bacterium; Fc-gamma receptor signaling pathway; innate immune response; T cell differentiation involved in immune response
M4AFX7	antigen processing and presentation; immune response
C3Z0F6	cell surface receptor signaling pathway; immune response
U3KP35	defense response to bacterium; innate immune response
M3YLV7	defense response to fungus; defense response to Gram-negative bacterium; dendritic cell migration; helper T cell enhancement of adaptive immune response; positive regulation of interleukin-6 secretion; positive regulation of interleukin-8 secretion; positive regulation of T-helper 1 type immune response; positive regulation of T-helper 17 type immune response
F6W4T0	defense response to virus; innate immune response; negative regulation of NF-kappaB transcription factor activity; negative regulation of type I interferon-mediated signaling pathway; positive regulation of interferon-gamma-mediated signaling pathway; positive regulation of MHC class I biosynthetic process; positive regulation of transcription from RNA polymerase II promoter; positive regulation of type I interferon-mediated signaling pathway; regulation of kinase activity

H9CIE5	evasion or tolerance by virus of host immune response; pathogenesis
V5HHL2	humoral immune response; positive regulation of transcription, DNA-templated; regulation of osteoclast differentiation
A0A0P4W4Y6, A0A0P4VXW9, A0A0P4W1C0, H9GLN1	immune response
W0USX2	immune response; inflammatory response
I3J7X5	inflammatory response; innate immune response; signal transduction
Q32S45	innate immune response; peptidoglycan catabolic process
A0A0L8GRS1	
A0A0A7M0R5	innate immune response; signal transduction
C3YPC3	
G5CJV7	
B5BRC1	
S9WX48	regulation of immune system process; transport

Table 6.3.

GO identification of biological processes associated with chitin and a list of associated genes.

Function	Genes
carbohydrate metabolic process; chitin catabolic process	A0A0H4LSY8, A0A059VQH9, A0A024BU17, B0LY40, A0A034V5L4, A0A067RWA3, C4P6W4, H8YI21, H8YI18, A0A0B4WEJ4, A0A0P4XMK2, A0A0P5B874, A8JD68, D6WZW3, U4UG49, T1JB05, A0A0P4ZGI1, V9IIB8, A0A0C5Q4X3, A0A060KSS6, A0A0P5K310, A0A0P5AYS9, A0A0P4WCZ4, A0A0H4LRH2, Q2TTG1, L0AUF3, A5YVK0, A0A0P4VXH9, E2ACU7, A0A0P5IZF6, H9NID9, B5B0D2
carbohydrate metabolic process; chitin catabolic process; inflammatory response; negative regulation of cytokine production involved in inflammatory response	F1NER5
carbohydrate metabolic process; chitin metabolic process	A0A0P5AXF0, A5YVK1, A0A084VNM0, D7NXS1, A0A0M5JDJ0, A0A0G2RLP0, B2YHF7
cell adhesion; chitin metabolic process	R9R483, Q75WG0
chitin metabolic process	A0A0P4WTM1, A0A0P4W2G1, A0A0B4NFP6, N6UD83, E9FTU3, D6X360, A0A0P5W1Z2, A0A0P4WLF2, U5XK53, A0A0F6PMG3, A0A067QWD6, A0A0P4WC75, A0A0P5B7G1, A0A0P5WNK4, A0A0N0BBW1, T1J4M2, B0WKF5, A0A0P5WWB8, A0A0P5JAL7, A0A0K2SXV6, A0A0K2U3K3, A0A0N0BH61, A0A0N8CH73, B1P1W0, A0A0K2U5A1, A0A0P4WJU3, D1MAI2, A0A131XLL7, A0A0P4WRW9, A0A0P4WZ80, A0A0P4WJV2, A0A087UP73, A0A0P4WKJ1, E0VPA9, A0A0K2V0R5, A0A0K2U295, T1JAE3, Q16I33, A0A0P4VZ52, Q4SHN1, A0A0A1WJ14, A0A0K8RHU4, V5GK57, E2ACM5, B3M4M6, E9G7M1, A0A0P4WFL4, E9HF51, A0A0P4WHV8, N6T9B0, A0A0Q9X5D9, A0A0L7QKL0, A0A067QQV9, A0A139WF85, T1J0Q9, A0A087TED0, A0A131Z4B5, A0A084WM71, E9GC79, A0A0P6APJ6, A0A0P5SP99, A0A0T6AZZ5, E2A6H3, A0A0J7LA17, T1J0R8, T1IK24

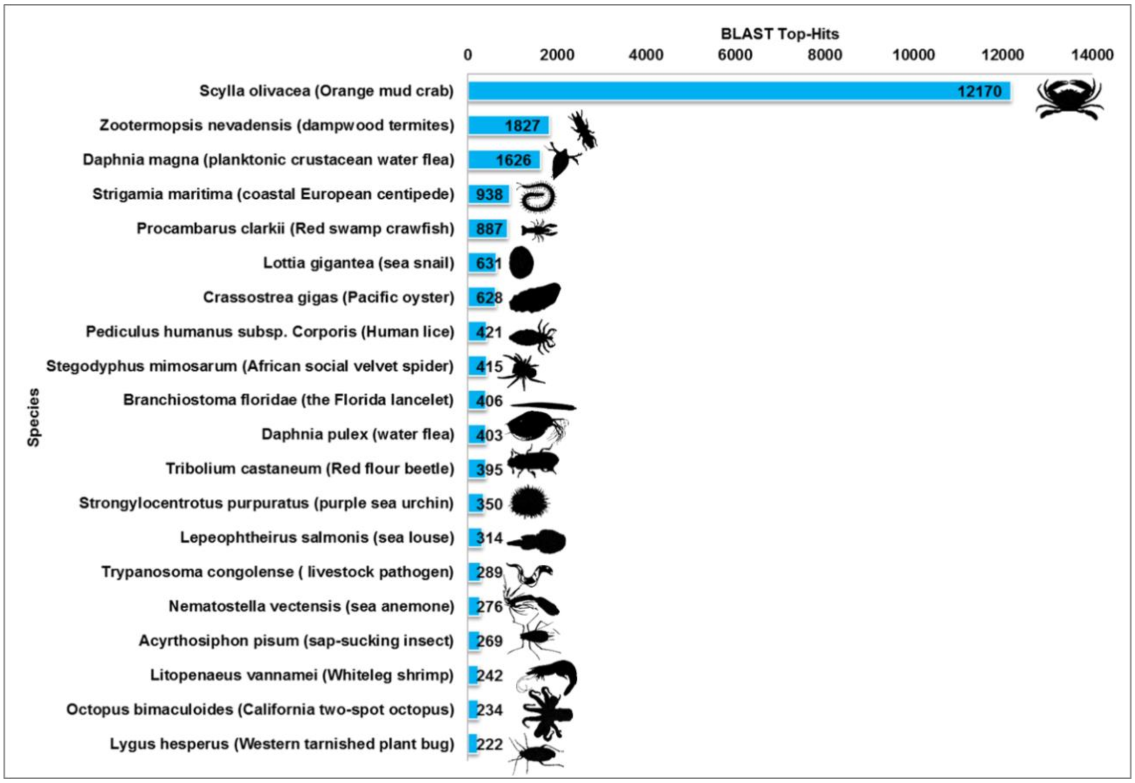


Figure 6.1.

Distribution graph of top-hit BLAST match species for the *Homarus americanus* genome.

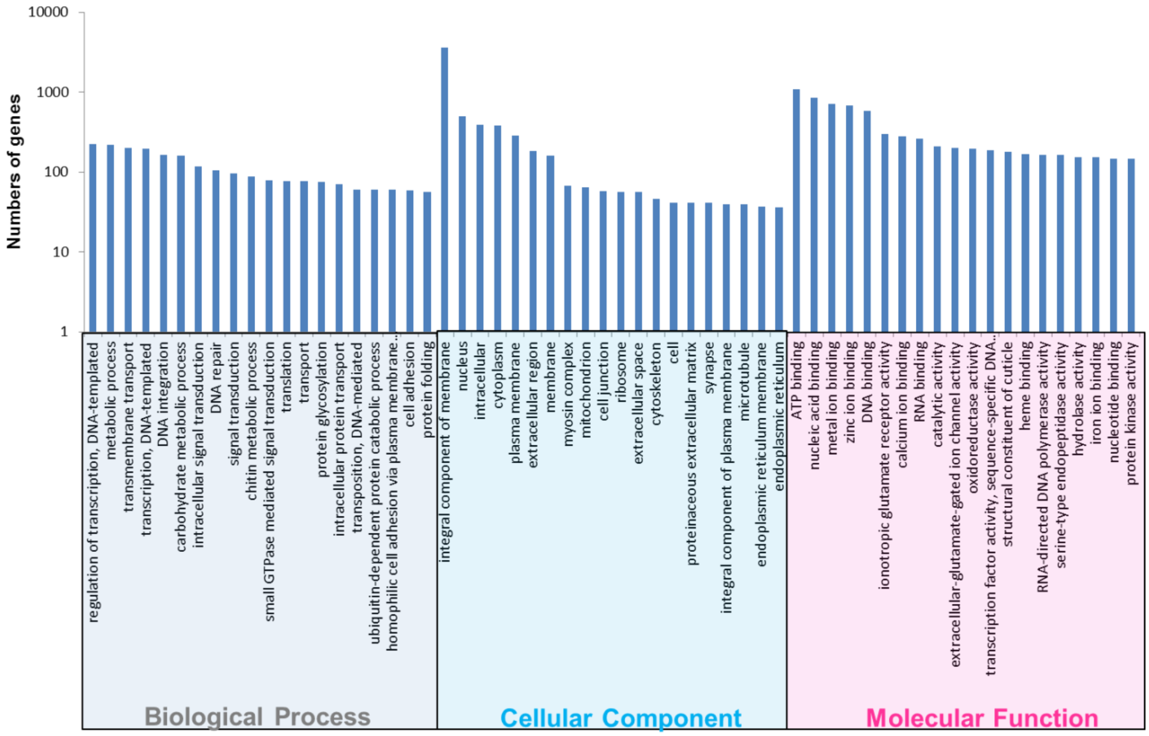


Figure 6.2.

Gene Ontology (GO) distribution of the genome of *Homarus americanus* in three GO categories: biological process, cellular component, and molecular function.

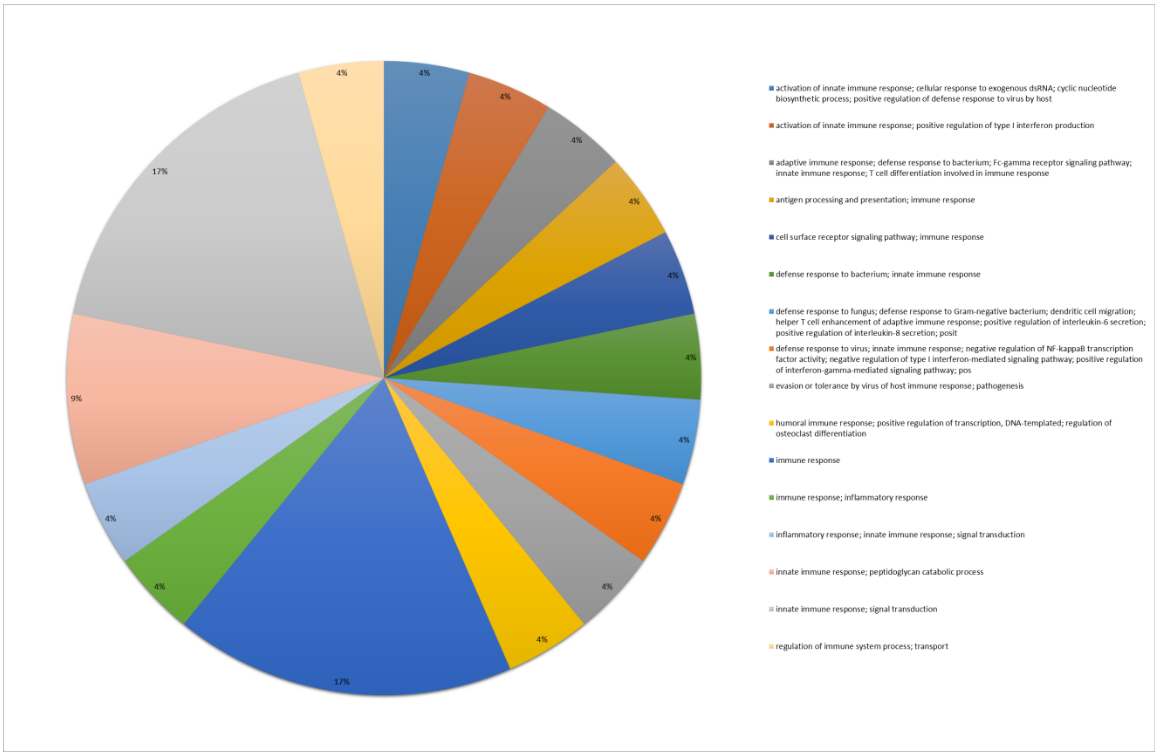


Figure 6.3.

Targeted GO identification of the biological processes and the percentages of annotated genes associated with the immune system.

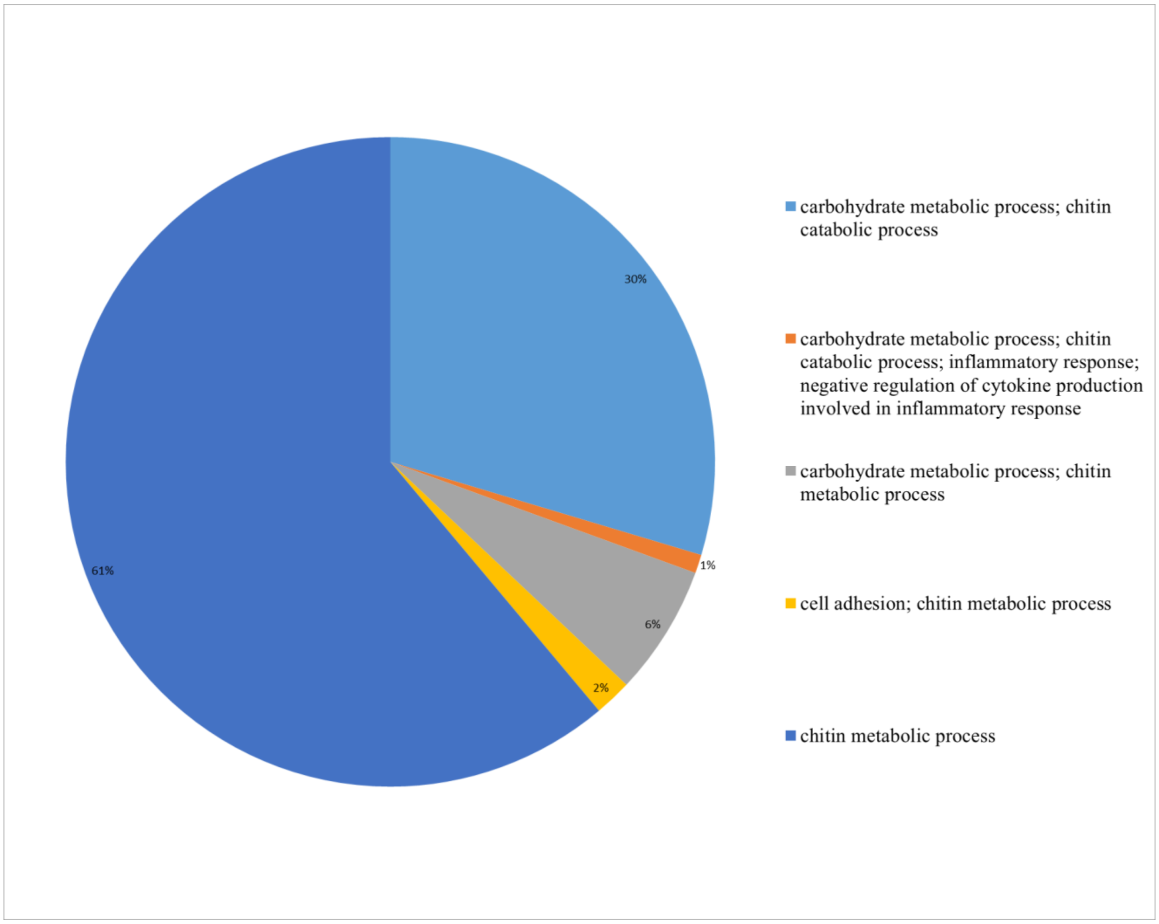


Figure 6.4.

Targeted GO identification of the biological processes and the percentages of annotated genes associated with chitin.

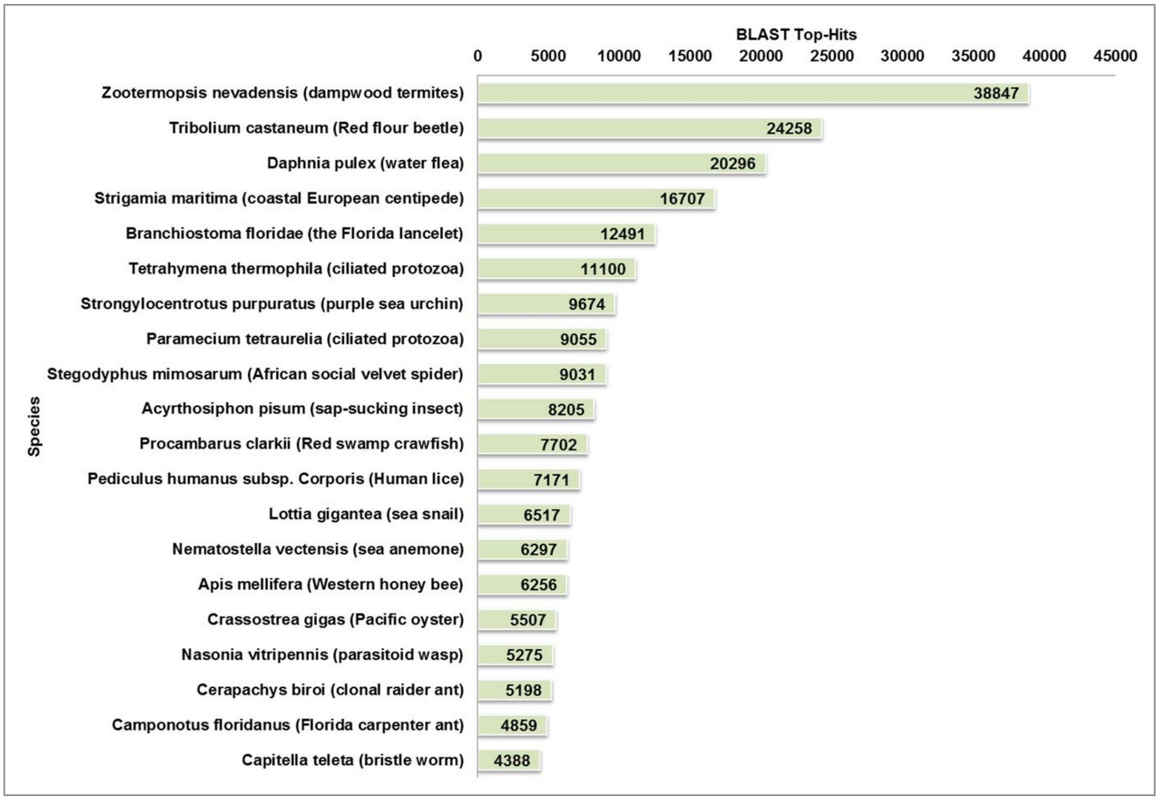


Figure 6.5

Distribution graph of top-hit BLAST match species for the *Homarus americanus* transcriptome.

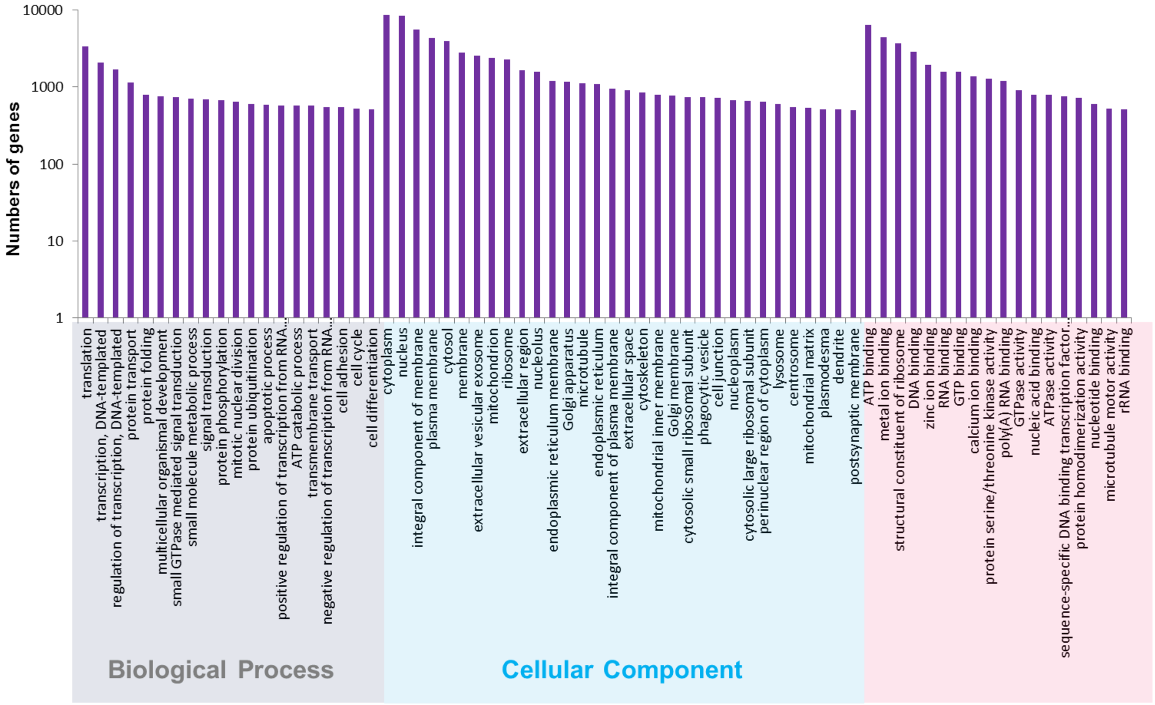


Figure 6.6

Gene Ontology (GO) distribution of the transcriptome of *Homarus americanus* in three GO categories: biological process, cellular component, and molecular function.

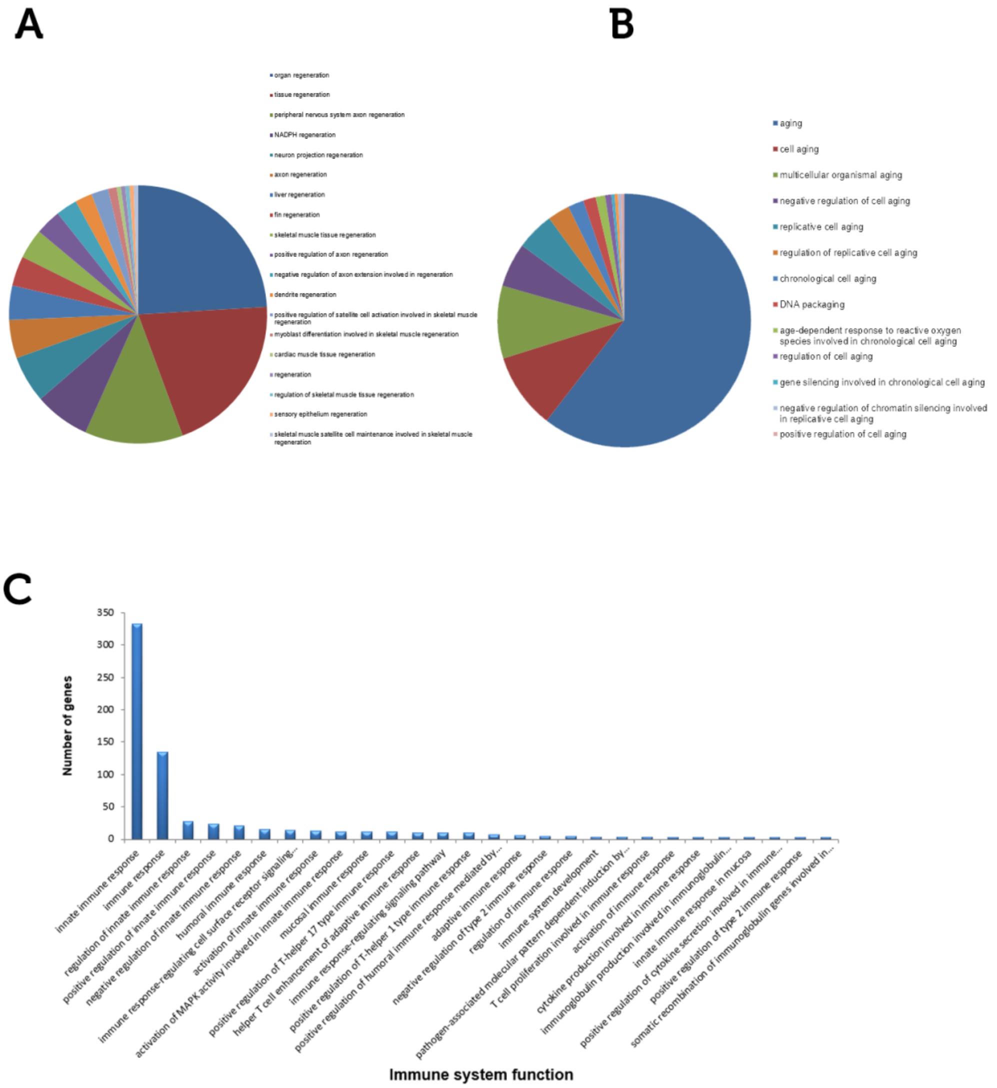


Figure 6.7

Transcriptome targeted GO identification of the biological processes and the number of annotated genes associated with A. regeneration, B. aging and C. immune system function.

Chapter 7: Discussion

During the past five years of this PhD I have worked on multiple studies that have resulted in five manuscripts. While these studies led to individual publications, the projects were all interlinked by the power of genetics. These studies involved utilizing the genome, transcriptome and proteome, along with numerous genetics based tools to inform and address our questions in two different model organisms (*Homo sapiens* and *Homarus americanus*). My primary research was focused on oral feeding readiness and was all conducted in neonatal saliva.

The initial publication was a comprehensive review of salivary diagnostics, a bio-fluid I utilized for most of my research during the past five years. This review reported how clinical diagnostics can be improved by faster and more accessible disease detection. Our laboratory has developed a point-of-service (POS) device capable of rapid, sensitive, automated, and multiplexed biomarker detection that uses human saliva instead of other bio-fluids. Here, we reviewed the technology that led to the development of this POS device. This POS technology can advance clinical diagnostics by saving time because of faster diagnosis, saving money because of a shorter hospital stay, and ultimately improving clinical care.

The next study was conducted to translate a previously identified transcriptomic panel (informative of oral feeding readiness) into a rapid proteomic platform to provide objective, near real-time assessment of oral feeding skills, to better inform care and improve neonatal outcomes. Assays for proteins involved in sensory integration, hunger signaling and facial development were developed. This study provided the foundation for the development of an informative rapid proteomic platform to assess neonatal oral feeding maturation. Oral feeding competency is a major determinant of length of stay in the neonatal intensive care unit (NICU). An infant must be able to consistently demonstrate the ability to take all required enteral nutrition by mouth prior to discharge home. Most infants born prematurely (<37 weeks) will require days, if not weeks, to master this oral feeding competency skill. Inappropriately timed feeding attempts can lead to acute and long-term morbidities, prolonged hospitalizations and increased healthcare costs. Previously, a panel of five genes involved in essential developmental pathways including sensory integration (*NPHP4*, *PLXNA1*), hunger signaling (*NPY2R*, *AMPK*) and facial development (*WNT3*) required for oral feeding success were identified in neonatal saliva. This current study aimed to translate these five transcriptomic biomarkers into a rapid proteomic platform to provide objective, real-time assessment of oral feeding skills, to better inform care and to improve neonatal outcomes. Total protein was extracted from saliva of ten feeding-successful and ten feeding-unsuccessful infants matched for age, sex and post-conceptional age (PCA). Development of immunoassays was attempted for five oral feeding biomarkers and two reference biomarkers (GAPDH, YWHAZ) to normalize for starting protein concentrations. Normalized protein concentrations were correlated to both feeding status at time of sample collection and previously described gene expression profiles. Only the reference proteins and those involved in hunger signaling were detected in neonatal saliva at measurable levels. Expression patterns for NPY2R and AMPK correlated with the gene expression patterns previously seen between successful and unsuccessful feeders and predicted feeding outcome. Salivary proteins associated with hunger signaling are readily quantifiable in neonatal saliva and may be utilized to assess oral feeding readiness in the newborn. This study lays the foundation for the development of an informative, rapid, proteomic platform to assess neonatal oral feeding maturation.

Simultaneously, an RNA sequencing platform was used to advance the development of feeding assessment tools by reviewing the entire human transcriptome using a high-throughput and quantitative approach. This work was undertaken to better understand the disrupted developmental patterns in the premature newborn. This study used high-throughput sequencing to understand a long-standing problem of oral-feeding maturation in neonatal care. While pathways discovered in our initial findings were reiterated in this study when comparing the feeders and non-feeders, splitting them by sex revealed that males and females had different feeding trajectories with certain developmental pathways evolving at a different pace. Overall, molecular pathways most related to feeding that were identified in the analyses were neurodevelopment in male infants and digestive and gastrointestinal in the female cohort. These data suggest that different developmental pathways did not reach maturation simultaneously in both sexes. Altered maturation has previously been seen between the sexes in neonatal pulmonary development. These data reiterate our findings of males and females not being exactly aligned in developmental pathway progression. Additionally, for the first time, hippocampal development (associated with memory and learning) was observed in feeding maturation for males. While researchers have identified adult male and female brains as having distinctly different compositions and pathways, our study has observed these sex differences in newborns. Furthermore, this study proves that neonatal saliva sequencing is a feasible platform for scientific research and clinical care, as results show biomarker discovery, sex differences and developmental pathway recognition. This project highlights important sex differences in oral feeding development and how they need to be incorporated in neonatal care moving forward. While developing treatments for the neonatal population, males and females cannot be treated as a whole population, but rather need to be considered based on their sex and possibly other confounding factors including age and ethnicity. In conclusion, this study has improved our understanding of oral feeding maturation in the neonatal population. Using RNA sequencing, we detected the presence of differentially expressed mRNA gene transcripts related to different developmental pathways, as well as those found in our previous studies. Furthermore, grouping the data by sex highlighted significant developmental differences, suggesting a clinically interesting finding. Additionally, expanding upon the current dataset for a better understanding of neonatal oral feeding maturation should validate these findings and further explore the different developmental pathways.

Moreover, we conducted a supplementary study to identify reliable reference genes to use for normalization of transcriptomic data in neonatal salivary diagnostics. Normalization of RT-qPCR expression data depends upon stable and consistently expressed genes. There is a paucity of data identifying reliable reference genes in the newborn population, where rapid and ongoing development directly impacts dynamic expression changes. Total RNA was extracted from 400 neonatal saliva samples (postconceptional ages: 32 5/7 to 48 3/7 weeks), converted to cDNA, pre-amplified and analyzed by qPCR for three commonly used reference genes, *ACTB*, *GAPDH* and *YWHAZ.* Relative quantification was determined using the Δ Ct method. All statistical analyses were performed using

SPSS. Data were analyzed as a whole and also stratified by age and sex. Descriptive statistics and homogeneity of variance (Levene's test) were performed to identify optimal reference genes. Data analyzed across all ages and sexes showed significant expression variation for *ACTB* (p < 0.0001), while, *GAPDH* and *YWHAZ* showed greater stability (p = 0.068). Divided by sex, male infants showed increased expression variation (p < 0.0001) compared to females (p = 0.139) for *ACTB*, but neither *GAPDH* nor *YWHAZ* showed significant variance (females: p = 0.342; males: p = 0.560). Our data suggests that *ACTB* is an unreliable reference gene for use in the newborn population. Males showed significantly more variation in *ACTB* expression compared to females, suggesting a sex-specific developmental role for this biomarker. *GAPDH* and *YWHAZ* were less variable and therefore are preferred reference genes for use in the neonate. This knowledge has the potential to improve the use of reference genes for the RT-qPCR platform in the newborn.

And finally, we utilized the transcriptome in an upstream direction to inform the genome and worked on presenting the first completely sequenced, annotated, and characterized genome for the *Homarus americanus* and provide a foundation for crustacean genetics. Crustaceans are a diverse taxon with both terrestrial and marine members and are closely related to insects. Among crustaceans the American lobster is an iconic species that is integral to many marine ecosystems and is an important commercial fishing industry in the Northwest Atlantic. In this paper, we provide the first complete American lobster genome and transcriptome sequences. We annotated approximately 37,000 genes from the genome and 24,000 from the transcriptome. A 70% to 80% concordance between the genome and transcriptome was observed. *Scylla olivacea* and *Zootermopsis nevadensis* were revealed as the most homologous species. Gene ontology revealed interesting functionally annotated genes with 16 different biological processes associated with the immune system and 23 genes with assigned functional annotations. The *Homarus americanus* genome provides a foundation for performing additional crustacean genetics studies.

In conclusion, the genome, transcriptome and proteome together had the potential to address genetics based questions in the smallest of humans to the largest of crustaceans. We used the transcriptome to inform downstream to the proteome and upstream to the genome. Utilizing transcriptomic and proteomic information simultaneously along with high-throughput transcriptomic technology and ultra-sensitive proteomic assays we were able to better understand oral feeding maturation in the human neonatal population while using salivary diagnostics and addressing challenges in the clinical setting. Moreover, we were able to use the transcriptome as a reference to perform *de novo* sequencing of the American lobster genome along with transcriptomic and proteomic bioinformatics pipelines to provide a solid foundation for crustacean genetics and support the commercial fishing industry. Hence, it is important to consider species on all three levels of the genome, transcriptome and proteome simultaneously and utilize the power of genetics for impact across a multitude of disciplines and fields.

Chapter 8: Appendix

8.1 Additional experiments for development of a rapid salivary proteomic platform for oral feeding readiness in the preterm newborn

Over the past decade, using microarray analysis and high-throughput reverse transcription-quantitative polymerase chain reaction (RT-qPCR), the Maron laboratory has performed salivary gene expression analyses on hundreds of infants at both pre- and post-oral feeding success to identify biomarkers that inform caregivers of neonatal feeding maturity [15]. The first phase of this research was exploratory and involved whole-transcriptome microarray analysis, comparing 12 neonates that were unable to orally feed with 12 infants demonstrating oral feeding readiness. Neonatal saliva was used to extract total RNA, which was then converted to cDNA and run on the Affymetrix microarray format. Affymetrix microarrays use base pairing complementarity, or hybridization, to identify the mRNA expression profile of genes present in the sample. Computational modeling and systems biology analysis were used to identify 421 genes associated with oral feeding success. Of these genes, 21 were selected that were identified within the Bayesian Network and/or were associated with oral feeding success. For the second phase, the RT-qPCR platform was utilized to run 400 neonatal saliva samples from infants divided into two cohorts: successful and unsuccessful oral feeders. After controlling for age and sex, the extensive research identified five key regulatory genes responsible for oral feeding maturity that were differentially expressed in successful and unsuccessful oral feeders. These genes included *Plexin A1 (PLXNA1), Neuropeptide Y2 receptor (NPY2R), adenosine-monophosphate-activated protein kinase (AMPK), wingless-type MMTV integration site family, member 3 (WNT3)*, and *nephronophthisis 4* (*NPHP4*). These five genes gave an area under the receiver operator characteristic curve (AUROC) score of 0.78 when combined into a model to predict successful oral feeders. Our study used this previously identified transcriptomic biomarker panel that is informative of neonatal oral feeding readiness to translate to a proteomic assay. While proteins are more complex compared to DNA and RNA, their relative stability makes them easier to work with and gives a more informed picture of gene expression. In addition, protein assays are quicker and easier to run than nucleic acid assays—an important feature for some clinical tests.

To better study the proteome in a clinical setting, the Walt laboratory has developed multiple sensitive protein detection platforms. These include a point-of-service (POS) device capable of automated, multiplexed, and sensitive biomarker detection and a digital ELISA platform called single molecule array (SiMoA) technology. Another platform to study the proteome in the clinical setting developed by the Walt laboratory is the SiMoA platform, which, was previously used to capture and analyze individual target molecules in the microwell optical fiber arrays [68, 82]. Briefly, SiMoA involves adding capture antibody microspheres to a sample that contains the target protein molecules, but with more microspheres than molecules. After sandwich formation with an enzyme-labeled detection antibody, the microspheres are loaded into femtoliter reaction wells and sealed with a fluorogenic substrate. Determining the number of wells that display increased fluorescence after incubation with the fluorogenic substrate quantifies the protein concentration from the original sample [81]. This SiMoA technology is capable of detecting biomarkers at the attomolar to femtomolar range [82]. These detection limits are hundreds to thousands of times more sensitive than conventional ELISAs, enabling the detection of target proteins at concentrations that were not measured before. This successful assay indicates the potential to use the multiplexed protein array for respiratory disease diagnosis, which may be applicable to other protein biomarkers and diseases.

Our study bridged the gap between neonatology and chemistry with a genetics based link. We used the sensitive protein detection tools developed in the Walt Laboratory to address the neonatal oral feeding complications identified by the Maron Laboratory. In this study, we used the previously identified transcriptomic neonatal oral feeding biomarker panel and attempted to translate it to a rapid and more stable proteomic assay. We used an advanced, state-of-the art, high-tech approach in order to tackle a long-standing problem in neonatal care. Our multi-disciplinary team afforded an important opportunity to create an impact in clinical care by translating bench side technological advances to the neonatal bedside to objectively assess feeding maturity in the premature newborn.

8.1.1 Oral Feeding Maturation in Neonates

Premature birth affects an estimated 11.5% of all pregnancies in the United States resulting in medical costs exceeding $26 billion annually [85]. Of these infants, most do not suffer from a genetic disorder or syndrome. Rather, their overall development has been disrupted due to their prematurity. This disruption begins with aberrant gene expression that leads to aberrant protein translation and ultimately results in phenotypes and disease that are unique to this population. Understanding this disrupted development at a transcriptomic and proteomic level may allow for targeted therapies and personalized care plans based upon an infant’s gene and/or protein expression.

One well known area of disrupted development that affects the premature infant and presents a significant burden to families and the healthcare system is the maturation of oral feeding. The majority of premature infants do not have the developmental maturity to successfully feed by mouth and must overcome the challenge of learning to orally feed before they can be discharged from the hospital. Failure of a premature infant to properly transition from nasogastric to oral feeds may result in choking, feeding aversion, prolonged hospitalization, parental anxiety, poor growth, and impaired short- and long-term neurodevelopmental outcomes [20, 21, 93, 187]. Research has shown that up to 40% of children in feeding disorder programs are former preterm infants, strongly linking long-term feeding difficulties to failed oral feeding trials in the neonatal intensive care unit (NICU) [94]. In addition, infants either born at term gestation or who correct to term post-conceptional age (PCA) who cannot successfully orally feed have been shown to be at increased risk for developmental disabilities and may require surgical insertion of a gastrostomy tube to provide adequate enteral nutrition [118, 188].

Currently, there are no strategies available to objectively assess oral feeding maturity in newborns [19]. Standard of care is limited to subjective assessment of an infant’s feeding cues (i.e. ability to suck on a pacifier) once an infant corrects to > 32 weeks’ PCA and has a stable respiratory status [22, 90-92]. This lack of an objective assessment tool has resulted in missed opportunities to feed infants with mature oral feeding skills, while placing immature oral feeders at risk for significant feeding associated morbidities [86].

Each failed approach results in prolonged length of stay with millions of dollars in associated health care costs [19].

Previous studies, using microarray analysis and high-throughput reverse transcription-quantitative polymerase chain reaction (RT-qPCR), have performed salivary gene expression analyses on hundreds of infants at both pre- and post-oral feeding success to identify biomarkers that inform caregivers of neonatal feeding maturity [87]. The result of this considerable study has been the identification of five key regulatory genes responsible, in part, for oral feeding maturity that were differentially expressed in successful and unsuccessful oral feeders. These biomarkers are involved in different developmental pathways and include *NPY2R* (involved in hunger signaling), *AMPK* (involved in energy homeostasis), *PLXNA1* (involved in olfactory neurogenesis), *NPHP4* (involved in visual behavior) and *WNT3* (involved in facial development) [15]. Each biomarker was informative in a binary fashion (+/- gene expression). An infant demonstrated a mature oral feeding pattern when *AMPK*, *PLXNA1*, and *NPHP4* were positively expressed and *NPY2R* and *WNT3* were negatively expressed. This combination of gene expression for these five biomarkers provided an accuracy of 78% at discerning oral feeding status (area under the receiver operating curve [AUROC] = 0.78). Importantly, this newly developed biomarker panel may have the capability to not only predict feeding maturity, but also to elucidate disrupted developmental patterns limiting feeding success. For example, while some unsuccessful oral feeders may have immature hunger signaling pathways, others may have disruption in facial developmental pathways. Discerning between these infants, in near real-time, will allow caregivers to develop personalized care plans to expedite feeding success. For example, an infant could be introduced to sensory therapies like the NTrainer System® (a biomedical device programmed to synthesize pneumatic pulse trains through a Soothie® silicone pacifier (Soothie® New Brighton, MN, USA) and has been correlated to improved sucking and feeding skills in pre-term infants) if we could identify that his/her somatosensory pathway was not fully developed [130]. This individualized care based on an infant's gene expression pattern would be a significant advancement over current subjective feeding assessment tools. Translating such a platform from a more burdensome transcriptome to a proteome assay would expedite the development of a point-of-service device to aid caregivers in their decision making regarding feeding practice and personalized interventions. However to date, no research has examined protein expression in neonatal saliva to determine oral feeding maturity. We have applied a novel approach in the field of human genetics by using RNA expression levels detected in neonatal saliva to guide us in the selection of potential protein biomarkers.

8.1.2 Development of additional biomarker assays

Protein expression for two additional biomarkers was considered in this study. These markers included FOXP2, indicative of neonatal oral feeding status, and ACTB, a reference biomarker [15]. ACTB was later excluded from this reference panel based on its poor performance in regards to its variable expression as a reference biomarker [189]. Immunoassays were successfully developed for both biomarkers. FOXP2, a gene involved in the development of speech and language, was not part of the initial biomarker panel but was later included due to its association with oral feeding readiness [15, 190].

FOXP2 was not detected in the neonatal saliva samples. The assay was subsequently transferred to a more sensitive platform developed in the Walt laboratory called Simoa™ (Single Molecule Array), in an attempt to improve the assay sensitivity, which would aid in protein detection [82]. Different concentrations of recombinant protein standards and protease inhibitor were added to the saliva collection process to rule-out protein degradation during the collection process. None of these steps facilitated protein detection for this target.

8.1.3 Translating the protein assays indicative for oral feeding readiness to a point-of-service (POS) platform

The objective of this project was to adapt the POS device from the Walt laboratory to detect biomarkers indicative of neonatal oral feeding status at the multiplex protein level. For this purpose, we aimed to collaborate with researchers from different fields, including microfluidic device development, automated device assembly, image processing, and data analysis. Finally, we intended to prospectively validate the newly optimized and enhanced POS assay in a blinded fashion to determine the assay's overall accuracy. It was essential to demonstrate not only the accuracy of the platform, but also its universal applicability. For this validation, parents of infants with a diverse range of gestational ages (≥ 24 to 42 weeks) and clinical sequelae were going to be asked to participate in this phase of the study (n=100). Saliva samples would have been collected from these neonates once they reached a PCA ≥ 32 up to 44 weeks. Samples would have been collected equally from successful (n=50) and unsuccessful (n=50) oral feeders and coded upon collection to blind the investigator to feeding status. Protein expression profiles would have been retrospectively correlated to feeding status after all samples had been run on the platform to determine its accuracy in predicting feeding maturity. Due to the inability to detect only two of the five biomarkers indicative of oral feeding readiness in the newborn, the assay's translation to the POS device seemed premature. A bedside assay capable of detecting only two biomarkers, with both involved in hunger signaling hindered its translation to a POS device. Before translating to a POS device, additional assays should be developed that include biomarkers involved in other significant developmental milestones required for oral feeding success, including neurodevelopment, gastrointestinal maturation and sensory integration.

8.1.4 Biological & Technical Challenges

The goal of this project was to develop protein assays for all eight biomarkers indicative of oral feeding and multiplex them for rapid detection. Consequently, this multiplexed assay was to be optimized and transferred to the POS platform. Months were spent trying to find matched antibody pairs for the development of assays for the eight protein targets. Antibodies from numerous vendors were tested including Abcam, LS Bio, Novus USA, Novus Bio, Aviva systems, Somalogic and R&D Systems. We tested these antibodies on two different sandwich assay platforms: conventional ELISA and SiMoA. We tried modifying experimental parameters to improve performance, including numerous different concentrations and incubation times. Ultimately, only targets that had matched antibody pair duo-sets available from R&D Systems were successful in developing protein assays. Once the assays were developed, we encountered issues where we were unable to detect the target protein in the neonatal saliva samples. Troubleshooting steps included testing saliva from infants of varying ages (from very pre-term to full-term) and adults to ensure that we did not miss the window of protein expression. Saliva samples were also spiked with protease inhibitors at different concentrations and time-points to confirm that proteases were not destroying the proteins of interest before detection. For some protein biomarkers (NPHP4 and PLXNA1) we were unable to find good matched pair antibodies, but other proteins (WNT3 and FOXP2) were not detectable in the neonatal saliva samples. These findings could hint at the underlying biology as both salivary biomarkers (AMPK and NPY2R) detected were involved in hunger signaling, potentially suggesting some form of interaction for these two proteins in the hunger signaling pathway, and that they may play an important role in regulating feeding in the newborn. It is also possible that these salivary proteins are playing some role in the maintenance and regulation of this pathway within the oral cavity and other tissues as observed in other proteins [95, 191]. Further research on the neonatal salivary proteome is warranted to better understand and confirm these findings.

8.2 Additional experiments for salivary RNA sequencing analysis of oral feeding maturation in the preterm infant

8.2.1 RNA sequencing of neonatal saliva using Oxford Nanopore's MinION

Oxford Nanopore produces a nanopore technology-based sequencer and our lab gained access to it under a pre-commercial and broad early access program. We attempted to use their MinION (a miniaturized sequencing device) to conduct RNA sequencing on neonatal saliva. To date, there is no published data using this device to sequence RNA content of saliva. This project aimed to be the first to successfully conduct RNA sequencing on neonatal saliva and potentially bring the portable miniaturized sequencer to the infant's bedside in the NICU. After receiving the configuration pack (that included the MinION device, USB cable and a Configuration Test Cell) and successfully installing and checking the MinION and software, sequencing kits, a wash kit to enable multi-sample usage of MinION™ Flow Cells, return kits and two experimental flow cells were received. Next, Burn-in experiments (generic control experiments that Oxford Nanopore runs internally as a matter of routine) were successfully run. The purpose of these 'Burn-in' runs was to enable participants to master the technology and fully understand it from end-to-end without the complication of novel biology. Subsequently, we converted total RNA from neonatal saliva to cDNA and ran it on the MinION platform linked to a laptop.

Once the run was complete (run-time of approximately 48 hours), data were extracted using a Python script after selecting the Windows-based option. All base-called fast5 files generated by MinKNOW were then converted into a single fastq file. All base-called files were sent to the "fail" folder instead of the "pass" folder by the Metrichor agent software. This happened because the reads did not pass the quality filter threshold specified by MinKNOW, as they were too short. Neonatal saliva samples are limited in volume. Additionally, besides the quantity, we are also limited in terms of quality as human bio-fluids contain degraded RNA compared to human tissue materials. Recent publications indicate that MinIONs are now capable of RNA counting in real time due to upgraded software and hardware [192, 193]. While this development is a promising next step for this project, these projects require RNA extracted from tissue, which still leaves the problem of the quality of degraded RNA.

8.2.2 Biological & Technical Challenges

This research was dependent upon the recruitment of human subjects, which presented a unique set of challenges. Months were spent waiting for babies to be born that met our respective research study criteria. Infants that displayed neonatal abstinence syndrome (NAS) because they had been exposed to addictive illegal or prescription drugs while in utero were excluded from the study. Additionally, no subject who was intubated or receiving continuous positive airway pressure (CPAP) was included in the study. Overall, we only used healthy infants without significant morbidities, which significantly extended the study timeline, as premature infants with no major health conditions are relatively uncommon. Additionally, we restricted this study to Caucasian infants to reduce genetic variability introduced by different ethnicities, which also extended the study timeline. Finally, the infants were all matched for sex and age, which also contributed to lengthening the recruitment time period.

8.2.3 Saliva QC

Another aspect of this project that presented a challenge and lengthened the study timeframe was using neonatal saliva, which has its own technical limitations. Two saliva samples were collected at each time point and stored in separate freezers to ensure the safety of samples in the event of power loss. Numerous selection criteria were established because sample loss is inherent in a bio-fluid like saliva that already has degraded RNA. The selection criteria were evaluated based on the output from the Agilent Bioanalyzer on which the saliva samples were analyzed. These criteria included having at least 500ng of total RNA in the saliva sample, the RNA Integrity Number (RIN) of the sample be between 5 and 8 (anything less than 5 indicated completely degraded RNA and anything over 8 likely indicated mostly epithelial cells). And finally, the 28s peak be greater than the 18s peak in the electropherogram output. With such strict parameters, many samples in the bio-bank had to be discarded adding to the recruitment time period.

8.3 Biological & technical challenges for identifying optimal reference genes for RT-qPCR normalization in the newborn

The most challenging aspect of this study was finding the most useful statistical test to analyze and comprehend the data. We had gene expression values for three reference biomarkers from 400 neonatal salivary samples but were uncertain of which statistical test would be the most accurate one for comparing their performance as a reference biomarker for normalization. Months were spent utilizing all available statistical resources to identify which test to use. One-way ANOVA, linear regression, Kruskal-Wallis, Mann-Whitney and multivariate analyses were all considered before choosing the Levene's test (subset of ANOVA). The Levene's test assessed the homogeneity of variance across each gene, which helped us accomplish the study aims.

8.4 Biological & technical challenges for *de novo* sequencing, assembly, annotation and characterization of the *Homarus americanus* genome and transcriptome

Sequencing, assembling and annotating *de novo* genomes and transcriptomes is an active field with multiple pipelines being developed in parallel. Every pipeline has its own set of advantages and disadvantages, which is why we tried multiple pipelines simultaneously. Every step in these pipelines takes days to weeks before we find out if that particular step worked and in the interest of saving time we ran them simultaneously. Additionally, because the lobster datasets were considerably larger than other species' genomes (one of the major reasons no one has attempted to sequence them yet) they took longer than the average time estimated for these pipelines. Furthermore, we also encountered unexpected computer crashes partly due to the large size of the datasets. For assembling the genome, we ultimately used the MAKER2 pipeline but had been running PASA and Exonerate simultaneously. Fortuitously, there is a well-established pipeline in the literature for transcriptomic annotation called Trinotate, which we were able to use successfully in a time-efficient manner.

8.4.1 American Lobster Genome Ploidy Estimation

We suspected that the *Homarus americanus* genome was triploid due to an estimated physical genome size of 4.75 Gb [174], and the sequenced genome was about 1.5 Gb indicating a genome with many repeated regions. The high concordance (70-80%) of the genome and transcriptome suggests the genome is almost completely sequenced and a genome ploidy estimation was required to confirm our hypothesis. We ran the ploidy estimation pipeline, ConPADE, developed by Gabriel R. A. Margarido and David Heckerman [194].

Initially, we ran ConPADE allowing up to a ploidy of six and the majority of the contigs were reported as six (which indicated that six as the upper limit might not be enough). Hence, we gradually increased the maximum ploidy estimation to try and reach a point where all the contigs had not reached their limit. We maxed out at 16, indicating problems with the data including spurious alignments and possible sequence collapses. After reaching out to the creators of the software we reached the conclusion that ConPADE might not be the best ploidy estimation pipeline to use with this dataset and other pipelines like ploidyNGS will be undertaken [195].

Chapter 9: Bibliography

1. Manzoni, C., et al., *Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences.* Brief Bioinform, 2016.

2. Paigen, K., *One hundred years of mouse genetics: an intellectual history. II. The molecular revolution (1981-2002).* Genetics, 2003. 163(4): p. 1227-35.

3. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. 431(7011): p. 931-45.

4. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. 22(9): p. 1760-74.

5. Claverie, J.M., *Fewer genes, more noncoding RNA.* Science, 2005. 309(5740): p. 1529-30.

6. Frith, M.C., M. Pheasant, and J.S. Mattick, *The amazing complexity of the human transcriptome.* Eur J Hum Genet, 2005. 13(8): p. 894-7.

7. Gimelbrant, A., et al., *Widespread monoallelic expression on human autosomes.* Science, 2007. 318(5853): p. 1136-40.

8. Hack, C.J., *Integrated transcriptome and proteome data: the challenges ahead.* Brief Funct Genomic Proteomic, 2004. 3(3): p. 212-9.

9. Kalia, A. and R.P. Gupta, *Proteomics: a paradigm shift.* Crit Rev Biotechnol, 2005. 25(4): p. 173-98.

10. Haider, S. and R. Pal, *Integrated Analysis of Transcriptomic and Proteomic Data.* Current Genomics, 2013. 14(2): p. 91-110.

11. Ghazalpour, A., et al., *Comparative analysis of proteome and transcriptome variation in mouse.* PLoS Genet, 2011. 7(6): p. e1001393.

12. Chen, G., et al., *Discordant protein and mRNA expression in lung adenocarcinomas.* Mol Cell Proteomics, 2002. 1(4): p. 304-13.

13. Gygi, S.P., et al., *Correlation between protein and mRNA abundance in yeast.* Mol Cell Biol, 1999. 19(3): p. 1720-30.

14. Yeung, E.S., *Genome-wide correlation between mRNA and protein in a single cell.* Angew Chem Int Ed Engl, 2011. 50(3): p. 583-5.

15. Maron, J.L., et al., *Computational gene expression modeling identifies salivary biomarker analysis that predict oral feeding readiness in the newborn.* J Pediatr, 2015. 166(2): p. 282-8.e5.

16. Bustin, S.A., et al., *The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.* Clin Chem, 2009. 55(4): p. 611-22.

17. Bustin, S.A., *Why the need for qPCR publication guidelines?--The case for MIQE.* Methods, 2010. 50(4): p. 217-26.

18. Mizuno, K. and A. Ueda, *The maturation and coordination of sucking, swallowing, and respiration in preterm infants.* J Pediatr, 2003. 142(1): p. 36-40.

19. Crowe, L., A. Chang, and K. Wallace, *Instruments for assessing readiness to commence suck feeds in preterm infants: effects on time to establish full oral feeding and duration of hospitalisation.* Cochrane Database Syst Rev, 2012. 4: p. CD005586.

20. Mizuno, K. and A. Ueda, *Neonatal feeding performance as a predictor of neurodevelopmental outcome at 18 months.* Dev Med Child Neurol, 2005. 47(5): p. 299-304.

21. Samara, M., et al., *Eating problems at age 6 years in a whole population sample of extremely preterm children.* Dev Med Child Neurol, 2010. 52(2): p. e16-22.

22. Howe, T.-H., et al., *A review of psychometric properties of feeding assessment tools used in neonates.* J Obstet Gynecol Neonatal Nurs, 2008. 37(3): p. 338-49.

23. Crowe, L., A. Chang, and K. Wallace, *Instruments for assessing readiness to commence suck feeds in preterm infants: effects on time to establish full oral feeding and duration of hospitalisation.* Cochrane Database Syst Rev, 2016(8): p. Cd005586.

24. Mata, J., S. Marguerat, and J. Bähler, *Post-transcriptional control of gene expression: a genome-wide perspective.* Trends Biochem Sci, 2005. 30(9): p. 506-14.

25. Venter, J.C., H.O. Smith, and M.D. Adams, *The Sequence of the Human Genome.* Clin Chem, 2015. 61(9): p. 1207-8.

26. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity.* Nature protocols, 2013. 8(8): p. 10.1038/nprot.2013.084.

27. Holt, C. and M. Yandell, *MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.* BMC Bioinformatics, 2011. 12: p. 491.

28. VanHook, A.M. and N.H. Patel, *Crustaceans.* Curr Biol, 2008. 18(13): p. R547-50.

29. Stillman, J.H., et al., *Recent advances in crustacean genomics.* Integrative and Comparative Biology, 2008. 48(6): p. 852-868.

30. Dufresne, F. and N. Jeffery, *A guided tour of large genome size in animals: what we know and where we are heading.* Chromosome Research, 2011. 19(7): p. 925-938.

31. Carpenter, G.H., *The Secretion, Components, and Properties of Saliva.* Annual Review of Food Science and Technology, 2013. 4(1): p. 267-276.

32. Wong, D.T., *Salivary diagnostics powered by nanotechnologies, proteomics and genomics.* J Am Dent Assoc, 2006. 137(3): p. 313-21.

33. Mandel, I., *The role of saliva in maintaining oral homeostasis.* The Journal of the American Dental Association, 1989. 119(2): p. 298-304.

34. Humphrey, S.P. and R.T. Williamson, *A review of saliva: Normal composition, flow, and function.* The Journal of Prosthetic Dentistry, 2001. 85(2): p. 162-169.

35. Mandel, I.D. and S. Wotman, *The salivary secretions in health and disease.* Oral sciences reviews, 1976(8): p. 25-47.

36. Shipp, G., *Ultrasensitive measurement of protein and nucleic Acid biomarkers for earlier disease detection and more effective therapies.* Biotechnol Healthc, 2006. 3(2): p. 35-40.

37. Matse, J.H., et al., *Discovery and Prevalidation of Salivary Extracellular microRNA Biomarkers Panel for the Noninvasive Detection of Benign and Malignant Parotid Gland Tumors.* Clinical Cancer Research, 2013. 19(11): p. 3032-3038.

38. Smith, D.J., et al., *Effect of Age on Immunoglobulin Content and Volume of Human Labial Gland Saliva.* Journal of Dental Research, 1992. 71(12): p. 1891-1894.

39. Huang, C.-M., *Comparative proteomic analysis of human whole saliva.* Archives of Oral Biology, 2004. 49(12): p. 951-962.

40. Nagler, R.M., et al., *Saliva analysis in the clinical setting: revisiting an underused diagnostic tool.* J Investig Med, 2002. 50(3): p. 214-25.

41. Nagler, R.M. and O. Hershkovich, *Relationships between age, drugs, oral sensorial complaints and salivary profile.* Archives of Oral Biology, 2005. 50(1): p. 7-16.

42. Mascini, M. and S. Tombelli, *Biosensors for biomarkers in medical diagnostics.* Biomarkers, 2008. 13(7-8): p. 637-657.

43. Weigl, B., et al., *Towards non- and minimally instrumented, microfluidics-based diagnostic devices.* Lab on a Chip, 2008. 8(12): p. 1999-2014.

44. Hartmann, M., et al., *Expanding Assay Dynamics: A Combined Competitive and Direct Assay System for the Quantification of Proteins in Multiplexed Immunoassays.* Clinical Chemistry, 2008. 54(6): p. 956-963.

45. Hartwell, S. and K. Grudpan, *Flow based immuno/bioassay and trends in micro-immuno/biosensors.* Microchimica Acta, 2010. 169(3-4): p. 201-220.

46. Sia, S.K. and G.M. Whitesides, *Microfluidic devices fabricated in Poly(dimethylsiloxane) for biological studies.* ELECTROPHORESIS, 2003. 24(21): p. 3563-3576.

47. Derveaux, S., et al., *Synergism between particle-based multiplexing and microfluidics technologies may bring diagnostics closer to the patient.* Analytical and Bioanalytical Chemistry, 2008. 391(7): p. 2453-2467.

48. Spisak, S. and A. Guttman, *Biomedical Applications of Protein Microarrays.* Current Medicinal Chemistry, 2009. 16(22): p. 2806-2815.

49. Nie, S., et al., *An automated integrated platform for rapid and sensitive multiplexed protein profiling using human saliva samples.* Lab Chip, 2014. 14(6): p. 1087-98.

50. Michishige, F., et al., *Effect of saliva collection method on the concentration of protein components in saliva.* The Journal of Medical Investigation, 2006. 53(1,2): p. 140-146.

51. Navazesh, M., *Methods for Collecting Saliva.* Annals of the New York Academy of Sciences, 1993. 694(1): p. 72-77.

52. Helmerhorst, E.J. and F.G. Oppenheim, *Saliva: a Dynamic Proteome.* Journal of Dental Research, 2007. 86(8): p. 680-693.

53. Thomadaki, K., et al., *Whole-saliva Proteolysis and Its Impact on Salivary Diagnostics.* Journal of Dental Research, 2011. 90(11): p. 1325-1330.

54. Proudfoot, N.J., A. Furger, and M.J. Dye, *Integrating mRNA processing with transcription.* Cell, 2002. 108(4): p. 501-12.

55. Hong, J., et al., *Nucleic acid from saliva and salivary cells for noninvasive genotyping of Crohn's disease patients.* Genet Test, 2008. 12(4): p. 587-9.

56. Li, Y., et al., *Salivary transcriptome diagnostics for oral cancer detection.* Clin Cancer Res, 2004. 10(24): p. 8442-50.

57. Li, Y., et al., *RNA profiling of cell-free saliva using microarray technology.* J Dent Res, 2004. 83(3): p. 199-203.

58. St John, M.A.R., et al., *Interleukin 6 and interleukin 8 as potential biomarkers for oral cavity and oropharyngeal squamous cell carcinoma.* Arch Otolaryngol Head Neck Surg, 2004. 130(8): p. 929-35.

59. Hu, S., et al., *Salivary proteomic and genomic biomarkers for primary Sjögren's syndrome.* Arthritis Rheum, 2007. 56(11): p. 3588-600.

60. Nie, S., et al., *Multiplexed fluorescent microarray for human salivary protein analysis using polymer microspheres and fiber-optic bundles.* J Vis Exp, 2013(80).

61. Jason, A.T., et al., *Polymeric microbead arrays for microfluidic applications.* Journal of Micromechanics and Microengineering, 2010. 20(11): p. 115017.

62. Barbee, K.D., et al., *Multiplexed protein detection using antibody-conjugated microbead arrays in a microfabricated electrophoretic device.* Lab on a Chip, 2010. 10(22): p. 3084-3093.

63. Derveaux, S., et al., *Layer-by-layer coated digitally encoded microcarriers for quantification of proteins in serum and plasma.* Anal Chem, 2008. 80(1): p. 85-94.

64. Rissin, D.M. and D.R. Walt, *Duplexed sandwich immunoassays on a fiber-optic microarray.* Anal Chim Acta, 2006. 564(1): p. 34-9.

65. Thompson, J.A. and H.H. Bau, *Microfluidic, bead-based assay: Theory and experiments.* J Chromatogr B Analyt Technol Biomed Life Sci, 2010. 878(2): p. 228-36.

66. Walt, D.R., *Protein measurements in microwells.* Lab Chip, 2014. 14(17): p. 3195-200.

67. Szurdoki, F., K.L. Michael, and D.R. Walt, *A duplexed microsphere-based fluorescent immunoassay.* Anal Biochem, 2001. 291(2): p. 219-28.

68. Pantano, P. and D.R. Walt, *Ordered Nanowell Arrays.* Chemistry of Materials, 1996. 8(12): p. 2832-2835.

69. Nie, S., et al., *Multiplexed salivary protein profiling for patients with respiratory diseases using fiber-optic bundles and fluorescent antibody-based microarrays.* Anal Chem, 2013. 85(19): p. 9272-80.

70. Blicharz, T.M., et al., *Fiber-optic microsphere-based antibody array for the analysis of inflammatory cytokines in saliva.* Anal Chem, 2009. 81(6): p. 2106-14.

71. Braeckmans, K., et al., *Encoding microcarriers: present and future technologies.* Nat Rev Drug Discov, 2002. 1(6): p. 447-56.

72. Trau, M. and B.J. Battersby, *Novel Colloidal Materials for High-Throughput Screening Applications in Drug Discovery and Genomics.* Advanced Materials, 2001. 13(12-13): p. 975-979.

73. Service, R.F., *CHEMISTRY - RADIO TAGS SPEED COMPOUND SYNTHESIS.* Science, 1995. 270(5236): p. 577-577.

74. Xiao, X.-y., et al., *Combinatorial Chemistry with Laser Optical Encoding.* Angewandte Chemie International Edition in English, 1997. 36(7): p. 780-782.

75. Czarnik, A.W., *Encoding methods for combinatorial chemistry.* Current Opinion in Chemical Biology, 1997. 1(1): p. 60-66.

76. Gerver, R.E., et al., *Programmable microfluidic synthesis of spectrally encoded microspheres.* Lab Chip, 2012. 12(22): p. 4716-23.

77. Han, M., et al., *Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules.* Nat Biotechnol, 2001. 19(7): p. 631-5.

78. Tideman, P.A., et al., *Impact of a regionalised clinical cardiac support network on mortality among rural patients with myocardial infarction.* Med J Aust, 2014. 200(3): p. 157-60.

79. Olansky, L. and L. Kennedy, *Finger-stick glucose monitoring: issues of accuracy and specificity.* Diabetes Care, 2010. 33(4): p. 948-9.

80. Blicharz, T.M., et al., *Use of colorimetric test strips for monitoring the effect of hemodialysis on salivary nitrite and uric acid in patients with end-stage renal disease: a proof of principle.* Clin Chem, 2008. 54(9): p. 1473-80.

81. Rissin, D.M., et al., *Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations.* Nat Biotechnol, 2010. 28(6): p. 595-9.

82. Rissin, D.M., et al., *Simultaneous detection of single molecules and singulated ensembles of molecules enables immunoassays with broad dynamic range.* Anal Chem, 2011. 83(6): p. 2279-85.

83. Song, L., et al., *Direct detection of bacterial genomic DNA at sub-femtomolar concentrations using single molecule arrays.* Anal Chem, 2013. 85(3): p. 1932-9.

84. Zhang, H., et al., *Oil-sealed femtoliter fiber-optic arrays for single molecule analysis.* Lab Chip, 2012. 12(12): p. 2229-39.

85. Martin, J.A., M.J.K. Osterman, and P.D. Sutton, *Are preterm births on the decline in the United States? Recent data from the National Vital Statistics System.* NCHS Data Brief, 2010(39): p. 1-8.

86. Muraskas, J. and K. Parsi, *The Cost of Saving the Tiniest Lives: NICUs versus Prevention.* Virtual Mentor, 2008. 10(10): p. 655-8.

87. Maron, J.L., et al., *Neuropeptide Y2 receptor (NPY2R) expression in saliva predicts feeding immaturity in the premature neonate.* PLoS One, 2012. 7(5): p. e37870.

88. Granger, D.A., et al., *Integration of salivary biomarkers into developmental and behaviorally-oriented research: problems and solutions for collecting specimens.* Physiol Behav, 2007. 92(4): p. 583-90.

89. Forde, M.D., et al., *Systemic assessments utilizing saliva: part 1 general considerations and current assessments.* Int J Prosthodont, 2006. 19(1): p. 43-52.

90. da Costa, S.P. and C.P. van der Schans, *The reliability of the Neonatal Oral-Motor Assessment Scale.* Acta Paediatr, 2008. 97(1): p. 21-6.

91. Palmer, M.M., K. Crawley, and I.A. Blanco, *Neonatal Oral-Motor Assessment scale: a reliability study.* J Perinatol, 1993. 13(1): p. 28-35.

92. Bingham, P.M., T. Ashikaga, and S. Abbasi, *Relationship of Neonatal Oral Motor Assessment Scale to Feeding Performance of Premature Infants.* J Neonatal Nurs, 2012. 18(1): p. 30-36.

93. Delaney, A.L. and J.C. Arvedson, *Development of swallowing and feeding: prenatal through first year of life.* Dev Disabil Res Rev, 2008. 14(2): p. 105-17.

94. Lau, C., *[Development of oral feeding skills in the preterm infant].* Arch Pediatr, 2007. 14 Suppl 1: p. S35-41.

95. Khanna, P. and D.R. Walt, *Salivary diagnostics using a portable point-of-service platform: a review.* Clin Ther, 2015. 37(3): p. 498-504.

96. Ludwig, S.M. and K.A. Waitzman, *Changing feeding documentation to reflect infant-driven feeding practice.* Newborn and Infant Nursing Reviews, 2007. 7(3): p. 155-160.

97. Collins, M.A., et al., *Total protein is an effective loading control for cerebrospinal fluid western blots.* Journal of Neuroscience Methods, 2015. 251: p. 72-82.

98. Murphy, R.M. and G.D. Lamb, *Important considerations for protein analyses using antibody based techniques: down-sizing Western blotting up-sizes outcomes.* The Journal of Physiology, 2013. 591(Pt 23): p. 5823-5831.

99. Li, X., et al., *Identification and validation of rice reference proteins for western blotting.* J Exp Bot, 2011. 62(14): p. 4763-72.

100. Vogel, C., et al., *Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line.* Mol Syst Biol, 2010. 6: p. 400.

101. Harris, J.J., C.W. Duarte, and M.C. Mossing, *Using protein abundance to indicate underlying mRNA expression levels in E.coli: an SEM modelling approach.* Int J Comput Biol Drug Des, 2011. 4(4): p. 387-95.

102. Taniguchi, Y., et al., *Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells.* Science, 2010. 329(5991): p. 533-8.

103. Wang, H., et al., *Systematic investigation of global coordination among mRNA and protein in cellular society.* BMC Genomics, 2010. 11: p. 364.

104. Ammar, D.A., et al., *Characterization of the human type 2 neuropeptide Y receptor gene (NPY2R) and localization to the chromosome 4q region containing the type 1 neuropeptide Y receptor gene.* Genomics, 1996. 38(3): p. 392-8.

105. Hunt, S.C., et al., *Polymorphisms in the NPY2R gene show significant associations with BMI that are additive to FTO, MC4R, and NPFFR2 gene effects.* Obesity (Silver Spring), 2011. 19(11): p. 2241-7.

106. Weerasekara, V.K., et al., *Metabolic-stress-induced rearrangement of the 14-3-3zeta interactome promotes autophagy via a ULK1- and AMPK-regulated 14-3-3zeta interaction with phosphorylated Atg9.* Mol Cell Biol, 2014. 34(24): p. 4379-88.

107. Hardie, D.G., F.A. Ross, and S.A. Hawley, *AMPK: a nutrient and energy sensor that maintains energy homeostasis.* Nat Rev Mol Cell Biol, 2012. 13(4): p. 251-62.

108. Hawdon, J.M., et al., *Identification of neonates at risk of developing feeding problems in infancy.* Dev Med Child Neurol, 2000. 42(4): p. 235-9.

109. Dole, N., et al., *Maternal stress and preterm birth.* Am J Epidemiol, 2003. 157(1): p. 14-24.

110. Dietz, J.A., et al., *Optimal Techniques for mRNA Extraction from Neonatal Salivary Supernatant.* Neonatology, 2012. 101(1): p. 55-60.

111. Schurch, N.J., et al., *How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?* Rna, 2016. 22(6): p. 839-51.

112. Shors, T.J., C. Chua, and J. Falduto, *Sex Differences and Opposite Effects of Stress on Dendritic Spine Density in the Male Versus Female Hippocampus.* The Journal of Neuroscience, 2001. 21(16): p. 6292-6297.

113. Hu, S., et al., *Systems biology analysis of Sjögren's syndrome and mucosa-associated lymphoid tissue lymphoma in parotid glands.* Arthritis Rheum, 2009. 60(1): p. 81-92.

114. Bhandari, V., et al., *Hematologic profile of sepsis in neonates: neutrophil CD64 as a diagnostic marker.* Pediatrics, 2008. 121(1): p. 129-34.

115. Liang, R., et al., *Increased 14-3-3zeta expression in the multidrug-resistant leukemia cell line HL-60/VCR as compared to the parental line mediates cell growth and apoptosis in part through modification of gene expression.* Acta Haematol, 2014. 132(2): p. 177-86.

116. Aps, J.K., J. Delanghe, and L.C. Martens, *Salivary electrolyte concentrations are associated with cystic fibrosis transmembrane regulator genotypes.* Clin Chem Lab Med, 2002. 40(4): p. 345-50.

117. Blanchet, P., et al., *MYT1L mutations cause intellectual disability and variable obesity by dysregulating gene expression and development of the neuroendocrine hypothalamus.* PLoS Genet, 2017. 13(8): p. e1006957.

118. Wilson, L. and M. Oliva-Hemker, *Percutaneous endoscopic gastrostomy in small medically complex infants.* Endoscopy, 2001. 33(5): p. 433-6.

119. van der Kooij, M.A., et al., *Role for MMP-9 in stress-induced downregulation of nectin-3 in hippocampal CA1 and associated behavioural alterations.* Nat Commun, 2014. 5: p. 4995.

120. Fukuda, T., et al., *Aberrant cochlear hair cell attachments caused by Nectin-3 deficiency result in hair bundle abnormalities.* Development, 2014. 141(2): p. 399-409.

121. Chance, P.F., et al., *Clinical nosologic and genetic aspects of Joubert and related syndromes.* J Child Neurol, 1999. 14(10): p. 660-6; discussion 669-72.

122. Ji, J. and S. Maren, *Differential roles for hippocampal areas CA1 and CA3 in the contextual encoding and retrieval of extinguished fear.* Learn Mem, 2008. 15(4): p. 244-51.

123. Bartsch, T., et al., *CA1 neurons in the human hippocampus are critical for autobiographical memory, mental time travel, and autonoetic consciousness.* Proc Natl Acad Sci U S A, 2011. 108(42): p. 17562-7.

124. Jack, C.R., et al., *Age, sex, and apoe ε4 effects on memory, brain structure, and β-amyloid across the adult life span.* JAMA Neurology, 2015. 72(5): p. 511-519.

125. Andreano, J.M. and L. Cahill, *Sex influences on the neurobiology of learning and memory.* Learn Mem, 2009. 16(4): p. 248-66.

126. Fleisher, B., et al., *Lung profile: sex differences in normal pregnancy.* Obstet Gynecol, 1985. 66(3): p. 327-30.

127. Wu, C.J., et al., *[Long-term effect of oligodendrocyte precursor cell transplantation on a rat model of white matter injury in the preterm infant].* Zhongguo Dang Dai Er Ke Za Zhi, 2017. 19(9): p. 1003-1007.

128. Xue, S., et al., *Loss-of-Function Mutations in LGI4, a Secreted Ligand Involved in Schwann Cell Myelination, Are Responsible for Arthrogryposis Multiplex Congenita.* Am J Hum Genet, 2017. 100(4): p. 659-665.

129. Liang, X.P., et al., *[Peripheral nerve injury in LAMA2-related congenital muscular dystrophy patients].* Zhonghua Er Ke Za Zhi, 2017. 55(2): p. 95-99.

130. Poore, M., et al., *Patterned orocutaneous therapy improves sucking and oral feeding in preterm infants.* Acta Paediatr, 2008. 97(7): p. 920-7.

131. Miller, J.L., C. Macedonia, and B.C. Sonies, *Sex differences in prenatal oral-motor function and development.* Dev Med Child Neurol, 2006. 48(6): p. 465-70.

132. Townsend, L.B. and S.L. Smith, *Genotype- and sex-dependent effects of altered Cntnap2 expression on the function of visual cortical areas.* J Neurodev Disord, 2017. 9: p. 2.

133. Booler, H.S., et al., *Degree of Cajal-Retzius Cell Mislocalization Correlates with the Severity of Structural Brain Defects in Mouse Models of Dystroglycanopathy.* Brain Pathol, 2016. 26(4): p. 465-78.

134. Joo, Y., et al., *Involvement of 14-3-3 in tubulin instability and impaired axon development is mediated by Tau.* Faseb j, 2015. 29(10): p. 4133-44.

135. Maron, J.L., *Exploring the neonatal salivary transcriptome: technical optimization and clinical applications.* Clin Biochem, 2011. 44(7): p. 467-8.

136. Huggett, J., et al., *Real-time RT-PCR normalisation; strategies and considerations.* Genes Immun, 2005. 6(4): p. 279-84.

137. Challis, J., et al., *Fetal sex and preterm birth.* Placenta, 2013. 34(2): p. 95-9.

138. Gachon, C., A. Mingam, and B. Charrier, *Real-time PCR: what relevance to plant studies?* J Exp Bot, 2004. 55(402): p. 1445-54.

139. Pfaffl, M.W., *A new mathematical model for relative quantification in real-time RT-PCR.* Nucleic Acids Res, 2001. 29(9): p. e45.

140. Guenin, S., et al., *Normalization of qRT-PCR data: the necessity of adopting a systematic, experimental conditions-specific, validation of references.* J Exp Bot, 2009. 60(2): p. 487-93.

141. Bustin, S.A., et al., *Quantitative real-time RT-PCR--a perspective.* J Mol Endocrinol, 2005. 34(3): p. 597-601.

142. Su, X., et al., *Optimization of Reference Genes for Normalization of Reverse Transcription Quantitative Real-Time Polymerase Chain Reaction Results in Senescence Study of Mesenchymal Stem Cells.* Stem Cells Dev, 2016. 25(18): p. 1355-65.

143. Du, M., et al., *Selection of reference genes in canine uterine tissues.* Genet Mol Res, 2016. 15(2).

144. Chen, I.H., et al., *Selection of reference genes for RT-qPCR studies in blood of beluga whales (Delphinapterus leucas).* PeerJ, 2016. 4: p. e1810.

145. Solano, M.E., et al., *Identification of suitable reference genes in the mouse placenta.* Placenta, 2016. 39: p. 7-15.

146. Eisenberg, E. and E.Y. Levanon, *Human housekeeping genes, revisited.* Trends in Genetics. 29(10): p. 569-574.

147. Li, X., et al., *Identification of Suitable Reference Genes for Normalization of Real-Time Quantitative Polymerase Chain Reaction in an Intestinal Graft-Versus-Host Disease Mouse Model.* Transplant Proc, 2015. 47(6): p. 2017-25.

148. Panahi, Y., et al., *Selection of Suitable Reference Genes for Analysis of Salivary Transcriptome in Non-Syndromic Autistic Male Children.* Int J Mol Sci, 2016. 17(10).

149. Gunning, P.W., et al., *The evolution of compositionally and functionally distinct actin filaments.* J Cell Sci, 2015. 128(11): p. 2009-19.

150. Tarze, A., et al., *GAPDH, a novel regulator of the pro-apoptotic mitochondrial membrane permeabilization.* Oncogene, 2007. 26(18): p. 2606-20.

151. Zala, D., et al., *Vesicular glycolysis provides on-board energy for fast axonal transport.* Cell, 2013. 152(3): p. 479-91.

152. Nishimura, Y., et al., *Overexpression of YWHAZ relates to tumor cell proliferation and malignant outcome of gastric carcinoma.* Br J Cancer, 2013. 108(6): p. 1324-31.

153. Matta, A., K.W. Siu, and R. Ralhan, *14-3-3 zeta as novel molecular target for cancer therapy.* Expert Opin Ther Targets, 2012. 16(5): p. 515-23.

154. Deindl, E., et al., *Differential expression of GAPDH and beta3-actin in growing collateral arteries.* Mol Cell Biochem, 2002. 236(1-2): p. 139-46.

155. Moshier, J.A., T. Cornell, and A.P. Majumdar, *Expression of protease genes in the gastric mucosa during aging.* Exp Gerontol, 1993. 28(3): p. 249-58.

156. Serazin-Leroy, V., et al., *Semi-quantitative RT-PCR for comparison of mRNAs in cells with different amounts of housekeeping gene transcripts.* Mol Cell Probes, 1998. 12(5): p. 283-91.

157. Sellars, M.J., et al., *Real-time RT-PCR quantification of Kuruma shrimp transcripts: a comparison of relative and absolute quantification procedures.* J Biotechnol, 2007. 129(3): p. 391-9.

158. Verma, A.S. and B.H. Shapiro, *Sex-dependent expression of seven housekeeping genes in rat liver.* J Gastroenterol Hepatol, 2006. 21(6): p. 1004-8.

159. Das, R.K., S. Banerjee, and B.H. Shapiro, *Extensive sex- and/or hormone-dependent expression of rat housekeeping genes.* Endocr Res, 2013. 38(2): p. 105-11.

160. Chen, G., et al., *beta-Actin protein expression differs in the submandibular glands of male and female mice.* Cell Biol Int, 2016. 40(7): p. 779-86.

161. McCurley, A.T. and G.V. Callard, *Characterization of housekeeping genes in zebrafish: male-female differences and effects of tissue type, developmental stage and chemical treatment.* BMC Mol Biol, 2008. 9: p. 102.

162. Derks, N.M., et al., *Housekeeping genes revisited: different expressions depending on gender, brain area and stressor.* Neuroscience, 2008. 156(2): p. 305-9.

163. Zhang, L., et al., *Discovery and preclinical validation of salivary transcriptomic and proteomic biomarkers for the non-invasive detection of breast cancer.* PLoS One, 2010. 5(12): p. e15573.

164. Victoria, N.C. and A.Z. Murphy, *The long-term impact of early life pain on adult responses to anxiety and stress: Historical perspectives and empirical evidence.* Exp Neurol, 2016. 275 Pt 2: p. 261-73.

165. Porter, M.L., M. Perez-Losada, and K.A. Crandall, *Model-based multi-locus estimation of decapod phylogeny and divergence times.* Mol Phylogenet Evol, 2005. 37(2): p. 355-69.

166. Rota-Stabelli, O., et al., *Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda.* Genome Biol Evol, 2010. 2: p. 425-40.

167. Mykles, D.L. and J.H. Hui, *Neocaridina denticulata: A Decapod Crustacean Model for Functional Genomics.* Integr Comp Biol, 2015. 55(5): p. 891-7.

168. Maynard, J., et al., *Improving marine disease surveillance through sea temperature monitoring, outlooks and projections.* Philos Trans R Soc Lond B Biol Sci, 2016. 371(1689).

169. Steneck, R.S., et al., *Creation of a gilded trap by the high economic value of the Maine lobster fishery.* Conserv Biol, 2011. 25(5): p. 904-12.

170. Benestan, L., et al., *Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (Homarus americanus).* Mol Ecol, 2016. 25(20): p. 5073-5092.

171. Harvell, D., et al., *The rising tide of ocean diseases: unsolved problems and research priorities.* Frontiers in Ecology and the Environment, 2004. 2(7): p. 375-382.

172. Factor, J.R., K.M. Castro, and B.A. Somers, *Sea Grant 3rd Annual Science Symposium Lobsters as Model Organisms for Interfacing Behavior, Ecology, and Fisheries: Discussion Session Summary on Disease and Population Level Impacts.* Journal of Crustacean Biology, 2006. 26(4): p. 661-662.

173. Cawthorn, R.J., *Diseases of American lobsters (Homarus americanus): A review.* Journal of Invertebrate Pathology, 2011. 106(1): p. 71-78.

174. Jeffery, N.W. and T.R. Gregory, *Genome size estimates for crustaceans using Feulgen image analysis densitometry of ethanol-preserved tissues.* Cytometry Part A, 2014. 85(10): p. 862-868.

175. Carlborg, O. and C.S. Haley, *Epistasis: too often neglected in complex trait studies?* Nat Rev Genet, 2004. 5(8): p. 618-625.

176. Putnam, N.H., et al., *Chromosome-scale shotgun assembly using an in vitro method for long-range linkage.* Genome Res, 2016. 26(3): p. 342-50.

177. Chapman, J.A., et al., *Meraculous: de novo genome assembly with short paired-end reads.* PLoS One, 2011. 6(8): p. e23501.

178. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. 30(15): p. 2114-20.

179. Grabherr, M.G., et al., *Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data.* Nature biotechnology, 2011. 29(7): p. 644-652.

180. Jiang, Q., et al., *Inheritance and Variation of Genomic DNA Methylation in Diploid and Triploid Pacific Oyster (Crassostrea gigas).* Mar Biotechnol (NY), 2016. 18(1): p. 124-32.

181. Dufresne, F. and P.D.N. Hebert, *Temperature-related differences in life-history characteristics between diploid and polyploid clones of the Daphnia pulex complex.* &#xc9;coscience, 1998. 5(4): p. 433-437.

182. Lazzaro, B.P. and T.J. Little, *Immunity in a variable world.* Philos Trans R Soc Lond B Biol Sci, 2009. 364(1513): p. 15-26.

183. Robinson, G.E., et al., *Creating a buzz about insect genomes.* Science, 2011. 331(6023): p. 1386.

184. Hsu, A.C. and R.M. Smolowitz, *Scanning electron microscopy investigation of epizootic lobster shell disease in Homarus americanus.* Biol Bull, 2003. 205(2): p. 228-30.

185. Raabe, D., et al., *Preferred crystallographic texture of alpha-chitin as a microscopic and macroscopic design principle of the exoskeleton of the lobster Homarus americanus.* Acta Biomater, 2007. 3(6): p. 882-95.

186. Romano, P., H. Fabritius, and D. Raabe, *The exoskeleton of the lobster Homarus americanus as an example of a smart anisotropic biological material.* Acta Biomater, 2007. 3(3): p. 301-9.

187. Lau, C., et al., *Characterization of the developmental stages of sucking in preterm infants during bottle feeding.* Acta Paediatr, 2000. 89(7): p. 846-52.

188. Tsai, S.-W., C.-H. Chen, and M.-C. Lin, *Prediction for developmental delay on Neonatal Oral Motor Assessment Scale in preterm infants without brain lesion.* Pediatr Int, 2010. 52(1): p. 65-8.

189. Khanna, P., K.L. Johnson, and J.L. Maron, *Optimal reference genes for RT-qPCR normalization in the newborn.* Biotechnic & Histochemistry, 2017: p. 1-8.

190. Zimmerman, E., M. Maki, and J. Maron, *Salivary FOXP2 expression and oral feeding success in premature infants.* Cold Spring Harb Mol Case Stud, 2016. 2(1): p. a000554.

191. Mandel, A.L., H. Ozdener, and V. Utermohlen, *Brain-derived Neurotrophic Factor in Human Saliva: ELISA Optimization and Biological Correlates.* Journal of immunoassay & immunochemistry, 2011. 32(1): p. 18-30.

192. Garalde, D.R., et al., *Highly parallel direct RNA sequencing on an array of nanopores.* bioRxiv, 2016.

193. Batovska, J., et al., *Metagenomic arbovirus detection using MinION nanopore sequencing.* J Virol Methods, 2017. 249: p. 79-84.

194. Margarido, G.R. and D. Heckerman, *ConPADE: genome assembly ploidy estimation from next-generation sequencing data.* PLoS Comput Biol, 2015. 11(4): p. e1004229.

195. Augusto Correa Dos Santos, R., G.H. Goldman, and D.M. Riano-Pachon, *ploidyNGS: visually exploring ploidy with Next Generation Sequencing data.* Bioinformatics, 2017. 33(16): p. 2575-2576.