Introduction

The ocean contains a striking amount of diversity, particularly at the microbial level, but much of this diversity has yet to be characterized. The majority of these organisms remain difficult to culture and study using classic microbial ecology techniques. Advancements in sequencing over the past decade have made it feasible instead to investigate this vast diversity via bulk shotgun sequencing of seawater samples. These techniques have already led to fundamental recharacterizations of microbial diversity and a reevaluation of total diversity on Earth, including recent publication of a bacterial-dominated tree of life[1]. The ability to analyze the wealth of information present in marine sequencing data has the potential to revolutionize our understanding of the diversity and function of ocean communities.

The emergence of sequencing as an affordable, standard approach in studying both ecological diversity and community function in the ocean has created a secondary issue: analysis remains complex and computationally intensive, and the enormity of data being produced threatens to overwhelm current strategies for analysis, sharing, and management of data. As part of my dissertation work, I built *de novo* transcriptome resources from an ecologically and economically important squid species, *Doryteuthis opalescens*. This project highlighted major issues facing researchers working in genomics of non-model species: the lack of a standardized, open access, biologist-friendly pipeline has resulted in a patchwork of published transcriptomes varying in accuracy and completeness, often assembled using methods that are not fully documented. In practice, even well-annotated published sequencing data is of limited accessibility and utility; Researchers must be aware of the data generated and must have sufficient computational skills and resources to run comparative analysis with the many potentially relevant datasets. These issues fall into two main categories: suboptimal techniques for analysis of new data, and organizational and analytical impediments to comparative analyses using reference data.

Analyses of new sequence data nearly always involve functional annotation of genes and transcripts via comparative mapping to identify similarities with existing data. Today's gold standard mapping methods (i.e. BLAST, HMMER) are designed to accurately recover sequence similarity and homology, favoring accuracy over speed[2,3]. These tools work best when comparing sequences to appropriate references (i.e. a squid *de novo* transcriptome to the published *Octopus* genome[4]), but function poorly when trying to map to very large databases, i.e. non-redundant NCBI database, NR[5]. With the increased utilization and output of sequencing technologies, there is an urgent need for tools that can achieve similar accuracy while processing large amounts of data quickly and without excessive computational or memory footprints. Two computational techniques, *de bruijn* graphs and indexing via sequence bloom trees or minimum hashing, may allow us to apply the algorithms underlying current best practices tools into a framework that enables fast searching of large databases.

The study proposed here aims to apply *de bruijn* graph mapping methods and indexing techniques to transcriptomic data in order to enable database-wide comparative analyses. This is a novel application of these methods that has the capacity to vastly improve assembly, annotation, expression, and sharing of sequencing data. By testing these methods on the wide array of sequencing data found in the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP), we will assess accuracy and scalability of these methods, and their potential to help us characterize the vast diversity of the ocean[6].

Objectives and Specific Aims

The overarching goal of this project is to improve RNA sequencing analysis by both improving tools for rapid and accurate analysis of newly sequenced data, and by developing methods for comparative analysis with existing reference data. In the project proposed here, I will apply and extend existing computational methods to RNA-Seq data, testing success through computational experiments using RNA-Seq data from the Marine Microbial Eukaryotic Transcriptome Sequencing Project. To accomplish this objective, I consider three specific aims:

Aim 1 *(new data)*: Develop open-source pipelines that utilize novel streaming graph-based approaches to improve assembly, annotation, and expression analysis of RNA-Seq.

In Aim 1, I will modify existing graph-based expression and annotation algorithms to function well for RNA-Seq analysis. I hypothesize that graph-based annotation and expression analysis of RNA-Seq data will improve sensitivity and accuracy while reducing analysis time.

Aim 2 *(comparative sequence analysis)*: Leverage computational indexing methods to allow rapid comparative analysis of new data with large existing reference data.

In Aim 2, I will focus on applying computational indexing methods (i.e. Sequence Bloom Tree, MinHash) to index reference data (i.e. MMETSP, SRA). I hypothesize that the ability to rapidly compare new data with reference data will enable accurate and comprehensive annotation of new sequencing data, thereby standardizing and improving transcriptome analysis.

Aim 3 *(comparative expression analysis)*: Extend indexing methods to include expression data and enable comparative expression analyses with reference data.

In Aim 3, I will extend the indexing framework from Aim 2 to include relative expression information. I hypothesize that comparative analysis of RNA-Seq expression patterns will facilitate identification and functional characterization of gene co-expression suites.

Study System

The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) focused on building pure transcriptomes of cultured marine microbial eukaryotes, but is remarkably phylogenetically diverse[6]. This database forms a rich reference that can be queried in order to functionally characterize genes from environmental samples containing species that cannot be cultured. The project's focus on transcriptome data recognizes that eukaryotic genomes tend to be larger and more complex than prokaryotic genomes, and that expressed gene sequencing adds the ability to discern functional differences between communities in addition to sequence diversity differences.

The MMETSP database provides a unique opportunity to assess RNA-Seq analysis methods using a wide variety of organisms with relatively well-studied transcriptomes. The first aim will utilize new tools to improve assembly and analysis of each individual transcriptome. The second and third aims will index all data in the MMETSP using tools that facilitate rapid yet accurate comparative analysis. While MMETSP provides an ideal study system for this project, successful completion of each aim will increase the quality, accessibility and utility of the transcriptome data within MMETSP, thus ultimately furthering the initial project goals.

Approach and General Methodology

Aim 1 *(new data)*: Develop open-source pipelines utilizing streaming graph-based approaches to improve assembly, annotation, and expression analysis of RNA-Seq.

De novo transcriptome generation via RNA Sequencing (RNA-Seq) has become a standard approach in genomic studies of non-model organisms, particularly in population genomics, adaptation and evolution. However, the absence of easily accessible end-to-end workflows for analysis and consensus transcriptome quality assessment tools has resulted in a patchwork of transcriptome analyses that vary in quality and completeness[7,8]. As part of my PhD, I attempted to solve this problem by combining best practices open-source tools into a user-friendly pipeline that conducted transcriptome assembly, quality assessment, annotation, and expression analysis with a single command from the user. My own pipeline (MakeMyTranscriptome), and a number of other recently-released pipelines (OmicsPipe, Pipeliner, Dammit and Trinity/Trinotate), that also do all or part of transcriptome analysis, all attempt to solve accuracy and repeatability issues in RNA-Seq analysis, yet none have emerged as a gold standard, and none have yet solved the additional issue of how to cope with extremely large data[9-13]. Labs working on a single (or a few) particular species are now generating new RNA-Seq datasets with each additional study, but may not have the time or computational resources to reassemble the transcriptome using all available data. As it is impossible to assess expression of genes or transcripts unless they are first assembled properly in the transcriptome, these hurdles will greatly limit future analyses.

Aim 1 is designed to tackle some of these major issues by developing the *de bruijn* graph, rather than the consensus transcriptome, as the reference for annotation and expression. *De bruijn* graphs (DBG) are already essential in de novo RNA-Seq analysis: nearly all *de novo* transcriptome assemblers now rely on *de bruijn* graphs to link short reads together into larger contiguous sequences ('contigs') based on sequence similarity. Transcriptome "assembly" then involves collapsing these information-rich graphs into the consensus contigs that then form the basis for all downstream analyses (i.e. transcript quantification and annotation). However, recent work has begun to take advantage of the additional information present in the *de bruijn* graph to improve assembly, most notably via read error correction, which has been shown to improve transcriptome quality[14,15], and digital normalization, which can drastically reduce required computational resources for assembly[16]. In this aim, I will build from elements of the khmer software, developed in the Data Intensive Biology Lab to implement the graph annotation and quantification strategies[17].

Graph-based annotation: Typical annotation relies on sequence similarity search via BLAST, or HMM model searches, and two functioning prototypes have applied these mapping methods at the DBG level. BlastGraph conducts BLAST searches over the DBG, and Xander uses HMM models of reference genes to extend the DBG in a reference-guided fashion[18,19]. I will extend the algorithms described in these programs using the advancements recently made in tools such as DIAMOND to reduce search time and optimize graph-based expression[20].

Graph-based expression: While programs developed in the lab of sponsoring scientist CT Brown use kmer coverage quantification for both digital normalization and read error correction, these methods are not optimal for conducting expression analyses[17]. To accurately quantify RNA-Seq data, I will modify Salmon, a fast, accurate, reference-based expression quantification program, to function at the DBG level[21]. Recent publication of two graph mapping programs demonstrates feasibility of this goal[22,23]. Using the MMETSP data, I will compare DBG mapping with results from traditional mappers (BWA, Bowtie2) and quantification programs (Kallisto, Salmon) to ensure accuracy[21,24-26].

Testing: As all MMETSP datasets have already been analyzed twice via traditional assembly methods – first prior to initial publication in 2014, and second in the DIB lab in 2016, I will be able to compare the results of this new methodology against current gold-standard transcriptome analysis[27].

Expected Outputs: This aim will produce an open-source user-friendly pipeline to conduct graph-based assembly, annotation, and expression of transcriptomes with a single input step. This pipeline will then be used to reanalyze all datasets in the MMETSP database, after which all of these transcriptomes will be made publically available, along with metrics describing improvement over previous methods.

Aim 2 ***(comparative sequence analysis)*****: Leverage computational indexing methods to allow for rapid comparative analysis of new data with all existing reference data.**

All sequence data is inherently related via the same phylogenetic relationships that link all forms of life, but very few repositories incorporate this information alongside sequencing data deposition, which means this information is only usable in its most basic form – i.e. via BLAST-like searches for similarity. This sort of data storage relies on individual researchers to query the vast (and growing) data repositories to identify relevant datasets for interesting and appropriate comparative analysis. However, with appropriate indexing and search methods, we can transfer the burden of this task from the researcher to the computer. **The goal of this aim is to use computational indexing techniques to create a reduced representation of existing reference data to facilitate rapid computational querying with new data.**

There are a number of ways to create reduced representations of large datasets, which can vary both in required computational resources and in the level of detail the representations retain. Minimum Hashing (MinHash) and Sequence Bloom Trees (SBT) are two methods that have promising applications for comparative sequence analysis. **MinHash** techniques utilize the kmer content of sequencing data to create small unique signatures for each dataset that correlate with average nucleotide identity[28]. While these signatures cannot be re-resolved into the original kmers used to create a dataset, signatures are very small and easily comparable across datasets. The utility of MinHash lies in quickly identifying relevant and interesting datasets for more detailed comparisons. **Sequence Bloom Trees (SBT)** have currently been applied to RNA-Seq data by storing kmer content information in a binary tree containing a hierarchy of compressed bloom filters[29]. Importantly, these trees allows for sequence-based searching of short read reference data, and thus can allow resolution of the specific search sequence. SBT and MinHash indexing methods differ in their level of specificity and compression and therefore differ in their end utility, particularly in the applications just described. However, the SBT structure is highly flexible, and can alternatively be extended beyond binary tree format and used to store MinHash signatures, rather than kmer information[30]. In this application, MinHash signatures can be efficiently searched via a SBT structure.

Testing with MMETSP: I will apply MinHash and SBT indexing to the MMETSP transcriptomes, with the goal of querying with respect to sequencing similarity and functional similarity. **Sequence Similarity:** As the kmer content of a transcriptome mirrors sequence similarity, the MinHash signature based upon that kmer content is more similar in closely related species as compared with distant species. This property enables use of these indexed data representations to identify datasets built from closely related species, including building a rough phylogenetic relationship of a novel transcriptome with related species in the indexed database. A functional prototype of this algorithm shows proof-of-concept for this application[31]. **Functional Similarity:** These indexing techniques need not be exclusively applied to sequence similarity. If instead, we can build MinHash signatures based on all genes in a functional pathway (i.e. nitrogen fixation) we may instead be able to use this type of indexing to identify these pathways in any transcriptome (or metatranscriptome). This would necessarily require abstraction to protein level similarity (using BLOSUM amino acid substitution matrices) to compare functional similarity without comparing synonymous substitutions at the nucleotide level. I will test this application using well-annotated pathways from the MMETSP transcriptomes, before attempting to use it to identify functional pathways in metatranscriptome datasets. **Resource optimization:** Finally, for each application, I will test whether subsets of the RNA-Seq data (i.e. 1, 5, 10 million, etc) are sufficiently representative of the entire dataset to create a reduced representation. Preliminary analysis suggests this technique may be able to maintain sequence similarity relationships while significantly reducing the computational resources required for indexing[32].

Expected Outputs: This aim will produce an open-source tool to index new sequencing data using the MinHash and Sequence Bloom Tree methods, and a comparison tool that will enable rapid comparison between indexed novel data and indexed reference databases. These tools will be used to index all MMETSP data, and the indices will be made available for future comparisons. Upon successful completion of these outputs, the tool will then be applied to all SRA data to expand the reference database to improve comparative analyses.

Aim 3 *(comparative expression analysis)*: Extend indexing methods to include expression data to enable comparative expression analyses with reference data.

The majority of existing comparative analysis focuses on sequence similarity, with many programs and algorithms designed to use reference data to help characterize the structure and function of genes found within new RNA sequencing. However, while sequence can be used to predict function, it cannot be used to infer true functionality in the organism. Only by pairing sequence data with expression data can we gain an understanding of the function of genomes within the organisms or environment. Several gene expression reference databases already exist, i.e. BodyMap[33], though these databases tend to be limited by organism. In Aim 3, I will test two different approaches for incorporating expression information into the data structures utilized in Aim 2, MinHash Graph Annotation and weighted MinHash. **Graph Annotation** would focus on annotating the graph structure containing the MinHash information with abundance data. In this method, MinHash signatures would still be related primarily by sequence content, with the added benefit of retaining information on the abundance or expression of sequences within the original dataset. **Weighted MinHash,** on the other hand, follows a similar kmer indexing approach to regular MinHash, with the exception that sequences are assigned additional hashes according to the abundance of that sequence within the database[34,35]. These hashes are then aggregated into a similar fashion, with the caveat that closely-related hashes will now be related to abundance/expression similarity rather than sequence similarity.

Testing with MMETSP: I will test the utility of these extended indexing methods in enabling query search for species similarity, functional similarity, and expression similarity within MMETSP.
Expected Outputs: This aim will extend the tool produced in Aim 2 to facilitate similarity search for expression similarity in addition to sequence similarity (i.e. identify both species and functional groups). The tool will then be applied to MMETSP and subsequently to SRA data.

Significance

Sequencing analysis is currently limited by lack of appropriate tools, insufficient training, and poor data accessibility. The Sequence Read Archive (SRA) now contains over ~3 petabases of data, rendering it extremely difficult for researchers to identify, download, access, and use relevant and appropriate reference data for comparative analysis[29]. **It is becoming clear that we must rely on computational indexing and approximations if we are to be able to properly utilize the growing repository of information.** The project proposed here aims to address this issue, first by improving analysis methods for each individual transcriptome (aim 1), and then by increasing the accessibility and utility of reference data (aims 2, 3). Accomplishing these aims will contribute to the development of fast, standardized approaches for RNA-Seq analysis and indexed searching of reference databases, essential for the advancement of the field. If sequencing continues at the current rate, reference databases will soon contain sequence data from most common inhabitants in accessible environments, linked with functional information. Access to these comprehensive indexed reference databases will allow biologists to use sequencing data to tackle larger biological problems, such as microbial community functional response to altered abiotic ocean conditions brought about by global change.

Training Objectives, Justification of Sponsoring Scientist, and Career Goals

Training Objectives and Career Goals: The proposed project is a logical expansion of my PhD work and presents an excellent opportunity to learn new methods and techniques in data intensive biology. By immersing myself in a bioinformatics lab, I will be able to continue my training as a computational biologist and hone my skills in writing reusable and extensible scripts according to best practices for software development. My current expertise within bioinformatics lies in eukaryotic transcriptome analysis, but I aim to develop this skill set beyond single-organism transcriptome research to integrative transcriptomics to investigate functional processes in microbial and eukaryotic communities. As part of my dissertation research, I wrote scripts to link existing open source tools together into a pipeline for *de novo* RNA-Seq analysis. My research objective in my postdoc work is to gain experience in developing, modifying, and applying analysis and indexing algorithms to apply to large biological data. The project aims and the experience I gain through code review and interactions with computer scientists working on biological projects will allow me to apply my expanded toolset to understanding responses of marine organisms in a changing ocean. Throughout this project and beyond, I will continue my commitment to helping inspire, guide, and train young scientists as they pursue their own scientific goals (also see broadening participation section). My ultimate goal is to gain the necessary training and experience to apply for a tenure-track faculty position in Marine Bioinformatics by the end of my time as a Fellow.
Justification of Sponsoring Scientist and Host Institution: UC Davis, and in particular the Genome Center, has become a world leader in genomics research and training. Sponsoring scientist C. Titus Brown has expertise both in marine developmental biology and in computation, and is considered one of the leading researchers in non-model genomics and transcriptomics analysis. My proposed research is highly synergistic with Dr. Brown's interests and current research directions, and will specifically leverage the extensive codebase and computational resources available in the Data-Intensive Biology (DIB) lab. Dr. Brown and his lab have been actively exploring the applications of the MinHash signatures and Sequence Bloom Trees to sequence data, and thus are ideally suited to provide support, guidance, and collaboration as I apply these tools to RNA-Seq data. In addition, Dr. Brown has expressed a strong commitment to facilitating my independent research and making my professional training and growth a priority. Dr. Brown's lab philosophy explicitly supports my goals in enhancing diversity and inclusion in STEM fields. His open science policy is both exciting and inspiring – the open sharing of code and data

both enhances reproducibility and furthers the goals of equity, diversity, and inclusion by removing cost barriers that block access to high quality research papers.

Broader Impacts and Broadening Participation of Underrepresented Groups in Biology

Throughout my educational career, I have developed a strong commitment to teaching, training, and fostering enthusiasm for scientific research. As a Hispanic woman and a first generation American, I am acutely aware of the obstacles that could have prevented me from reaching graduate school, and continue to affect other students like me. I am committed to working to improve diversity, equity, and inclusion in science in meaningful and lasting ways. Through 2014-2016, I have worked with others to conceptualize and write position specifications for a Diversity Initiatives Coordinator at Scripps Institution of Oceanography. I feel privileged to have had the opportunity to work with an administration that has made a strong commitment to the idea that hiring students, faculty, and staff from a diversity of backgrounds, cultures, experiences and identities can enhance our ability to tackle large-scale social and scientific problems. I am proud to have contributed to the vision and work that led to the creation of this position and the hiring of a highly qualified candidate to work towards increasing diversity and retention of diverse students, faculty, and staff by fostering an inclusive and equitable culture at SIO.

In order to build a more inclusive and diverse academe, we must also work to expose a wider array of young students to scientific research. To that end, my 2016 outreach efforts included helping with an SIO-led project with San Diego's Ocean Discovery Institute to expose underserved young students to rigorous research in ocean acidification and deoxygenation, and co-leading one of 19 select researcher-designed workshops at Expanding Your Horizons, an annual conference designed to expose young girls to hands-on science and engineering. At UC Davis, I see great opportunities to continue my commitment by participating in and expanding the strong data analysis training programs developed by Dr. Brown and the UC Davis Genome Center. In 2015, I trained to become a certified instructor for Software Carpentry, a nonprofit organization that, along with related organization Data Carpentry, focuses on running accessible, inclusive data workshops to teach software and data analysis to promote efficient, sharable and reproducible data practices. I will continue my commitment to data training with these workshops but I also hope to meld my commitment to data analysis training with marine science outreach. As exposure to coding can reduce the barrier to entry, teaching basic coding and discussing its use and importance in marine biology will also be important in training and recruiting the next generation of marine biologists, for whom bioinformatics will be an essential skill. To address this need, I hope to collaborate with the postdoc-led ISOpod (Inquiry-based Science Outreach Pod) team at UC Davis and researchers at Bodega Marine Lab, as well as educators at the Ocean Discovery Institute and EYH and to develop a hands-on learning module focusing on analyzing temperature variation and its effects on intertidal organisms using simple Thermochron ibutton loggers, python coding, and traditional physiological metrics[36].

Timeline

If supported for this postdoctoral work, I hope to receive a 3-year (36 month) fellowship to allow for completion of research aims while fulfilling my commitment to training, mentoring, and leading outreach efforts to introduce a diverse array of students to the field of marine bioinformatics. In Summer 2017, I will begin by being an instructor and participant (for different areas of expertise) at a new UC Davis summer bioinformatics institute organized by Dr. Brown. Year 1 will focus on extending *de bruijn* graph methods for aim 1, with full reanalysis of the MMETSP transcriptomes and working open-source software made publically available by end of Year 1. In Summer 2018, I will develop prototype ISOPod educational lessons, and test them over the following year with Ocean Discovery Institute and at the EYH Conference. Year 2 will focus on modifying existing algorithms (and in one case, a working prototype) of MinHash and SBT indexing for sequence (Aim 2) and expression (Aim 3) data. As the structure and optimization of these tasks will be similar, work on the computational structure underlying each Aim will be conducted simultaneously. Year 3 will be devoted to optimization and testing of the new indexing software using the MMETSP and SRA data. I will present results at scientific meetings annually and produce at least one peer-reviewed publication for each Aim by the end of my tenure.