Proteogenomics of the Spotted Hyena (*Crocuta crocuta*)

By

Aidan Swartz

Thesis presented in fulfilment of the requirements for the degree of Molecular Biology in the Faculty of Medicine and Health at Stellenbosch University

Supervisor: Professor David L. Tabb

Co-Supervisor: Professor Michele Miller

March 2021

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2021

Abstract

The spotted hyena (*Crocuta crocuta*) is an important yet understudied organism that could provide insights into the fields of disease resistance, pathogen movement and disease evolution. They exist in matrilineally controlled, transient, clan-like groups that feed on a variety of organic matter and, subsequently, control the spread of pathogenic infections within an environment. Due to this, they appear to possess a high degree of resistance to pathogens. In this project, RNA-Seq data were utilized to assemble a transcriptome for the spotted hyena and tissue samples were further used to acquire protein data via MS/MS analysis. The aim of this study was to produce an accurate assembly via the transcriptomic data and subsequently further validate this assembly through the use of proteomics to better prove the quality therein. The assembly was produced using the Trinity *de novo* assembly software tool and assessed via the BUSCO and TransRate analysis tools. Orthology detection was carried out using ProteinOrtho, using closely related species (tiger, house cat, leopard, cheetah). Finally, LC-MS/MS data (consisting of tissue samples from peripheral, abdominal, head and thoracic lymph, as well as lung and liver tissue), and fractionated data from the sample containing the most diverse spectra, were searched against both the assembly itself and the translated genome data from the NCBI. These data served as the means by which the proteomic data were assessed and to determine whether the fractionation was successful, based on the comparative quantity of spectra between initial and fractionated analyses, in diversifying the sample. Further, these data were utilized to determine whether the translated transcriptome assembly could be successfully aligned against the proteomic data. The analysis of the quality control results found that the assembly was of appropriate quality when compared to the standards found within NCBI and within those described by the quality analysis tools. This coupled with the analysis of the proteomic data suggest that the assembly is useable, though requires further refinement. Based on the above, the inclusion of more data for assembly, is required for it to be a completely viable and ideal model assembly, however, current results are promising.

Opsomming

Alhoewel die huidige tydlyn dit verhoed het, sou daar data oor hiëna-reekse voor hierdie projek beskikbaar wees, die analise sal verder uitgebrei word. Die eerste stap sou 'n meer uitgebreide snywerk en daaropvolgende kwaliteitsbeoordelingsstap gewees het, wat sou bepaal of die snystap suksesvol is om die kwaliteit van die samestelling van die begin af te verbeter. Die voordeel van die beskikbaarheid van 'n genoom sou die gebruik van 'n ander samesteller noodsaak, moontlik deur die verwysingsgebaseerde samestellingsinstrument te gebruik, wat die genoom sou benut om 'n beter samestelling te bewerkstellig. 'n Verdere assessering, met behulp van 'n versameling monteerinstrumente, kan voordelig wees, aangesien een instrument waarskynlik onvoldoende is om al die data vas te lê. Die toets van 'n toepaslike instrument vir versoeningsversameling volg die vorige stap, wat die navorser in staat stel om te ondersoek of elkeen van die gemeentes saam beter resultate lewer as wanneer dit afsonderlik gebruik word.

Toetsing van kwaliteit behou die gebruik van BUSCO en TransRate, maar kon nie so maklik vir verwysingsgebaseerde analise gebruik word nie. In hierdie geval is dit die beste om 'n vergelykende stap met die NCBI-samestelling uit te voer of instrumente te ondersoek wat meer geskik is vir hierdie tipe analise, hoewel TransRate steeds gebruik kan word, aangesien dit die samestelling op die oorspronklike fastq-lêers karteer. Daar is verskeie ander instrumente vir genoomassessering, soos GAGE, maar dit is onseker of dit korrek van toepassing kan wees op 'n RNA-Seq-vergadering of 'n versoenende vergadering met behulp van RNA-Seq-data.

Na versoening en kwaliteitsbeoordeling is verdere ontleding nodig met behulp van die proteïendata. Hierdie stap sal die NCBI-proteïendata insluit vanaf die begin van die analise. Dit kan eenvoudiger wees, aangesien proteogenomiese navorsing met RNA, DNA en proteïene uitgevoer is, in plaas daarvan om slegs met RNA-Seq-data of genomiese data te begin. Een metode behels die bepaling van die vlak van oorvleueling tussen die twee proteïenstelle, sowel as tussen die proteïenstelle en die verskillende samestellings, as 'n vorm van vergelykende analise. Die bestryding kan in hierdie geval 'n meer verwante organisme wees, 'n lid van die Felidae-familie, of 'n selfs verder verwante spesie, soos 'n mens, wat 'n uitgebreide vergadering beskikbaar het.

Table of Contents

Acknowledgements

I would like to thank the Centre for Proteomic and Genomic Research (CPGR) for their excellent work in providing me with my LC-MS/MS data, which was utilized in this project.

I would like to extend this appreciation to Professor Paul van Helden, for providing the funding for this analysis and allowing this project to be carried out.

I would, finally, like to thank the South African Tuberculosis Bioinformatics Initiative (SATBBI) for providing the bursary for my studies and allowed me to continue with my thesis.

Opinions expressed, and conclusions arrived at, are my own and do not necessarily represent the views of those funding or supporting this project.

List of Figures

List of Tables

List of Abbreviations

ABySS	Assembly By Short Sequences
BLAST	Basic Local Alignment Search Tool
bp	Base Pairs
BUSCO	Benchmarking Universal Single-Copy Orthologs
CD-HIT	Cluster Database at High Identity with Tolerance
cDNA	Complementary Deoxyribonucleic Acid
CID	Collision Induced Dissociation
CPGR	Centre for Proteomics and Genomic Research
CPU	Central Processing Unit
DBG	De Bruijn Graph
ddNTP	Dideoxynucleotide Triphosphate
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleoside Triphosphate
EDTA	Ethylenediaminetetraacetic Acid
FDR	False Discovery Rate
FTICR	Fourier Transform Ion-Cyclotron
GC-MS/MS	Gas Chromatography Tandem Mass Spectrometry
HCD	Higher energy Collisional Dissociation
LC-MS/MS	Liquid Chromatography Tandem Mass Spectrometry
m/z	Mass-to-Charge
MIRA	Mimicking Intelligent Read Assembly
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
NCBI	National Center for Biotechnology Information
NGS	Nex Generation Sequencing

NIL	Refers to Negative Control Tubes with No Additives
OLC	Overlap Layout Consensus
PCR	Polymerase Chain Reaction
PSM	Peptide-Spectrum Match
QC	Quality Control
RAM	Random Access Memory
RNA	Ribonucleic Acid
RNA-Seq	RNA sequencing
RSA	Republic of South Africa
SCX	Strong Cation Exchange
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SOAPdenovo	Short Oligonucleotide Analysis Package
SPAdes	St. Petersburg Genome Assembler
ToF	Time of Flight
WGS	Whole Genome Sequencing

Chapter 1

Summary

The spotted hyena (*C. crocuta*) is a highly disease resistant species of scavenger organism. While the organism itself does not suffer, in most cases, from severe symptoms of infection, they can act as disease carriers. Alongside this, their role as scavengers serves to reduce the risk of disease infection through carcasses within an environment. In spite of this, the information regarding their genomic, transcriptomic or proteomic structure is sparse and not readily available. The main purpose of this research project was to assemble and validate a transcriptomic assembly for the spotted hyena.

The first component to this research project involved the assessment of quality for the original reads, which were then used to assemble the transcriptome. These data were acquired from a previous opportunistic extraction of hyena RNA, which was utilized for RNA-Seq analysis. The quality assessment analysis involved the use of FastQC, through which it was decided that further steps, such as trimming, were unnecessary. The reads were thus assembled using Trinity *de novo* assembly tool, which is a well know software tool used to assemble transcriptomic reads when there is a lack of a reference genome available.

The second component was the assessment of this assembly based on metric analysis using reference free assessment tools. In this case BUSCO and TransRate were used to best identify whether the assembly had been produced successfully and accurately. TransRate assesses assembly quality based on how well it aligns to the original reads from which it was produced, while BUSCO examines the completeness of the assembly based on how many single copy universal orthologs are present within the assembly, from the lineage database provided.

The third component specifically dealt with proteomics, in that proteomic data was acquired from lymph, liver and lung tissue, using LC-MS/MS with a subsequent fractionation step prior to further analysis. These data were analyzed using MS-GF+ which was provided a database of proteins drawn both from NCBI and the main assembly, which was converted to protein coding regions using TransDecoder. Selections of peptides identified from both were then compared via ProteinOrtho to determine overlapping regions with the original protein data collected from the NCBI.

The final component of this thesis is the discussion and conclusion. Here it could be ascertained that the assembly could be considered viable for further use and was accurately assembled. The fractionation process successfully introduced further diversity in terms of peptides. Furthermore, the use of proteomic data did prove that it was possible to align transcriptome-derived protein sequences to tandem mass spectra. Ultimately, further analyses would benefit the assembly, such using other tools, specifically those built to be used for proteogenomic research, such Galaxy or Spritz. That said, the assembly produced by this project is viable to support further hyena research.

Chapter 2

Introduction

Taxonomy of the Hyena

The family *Hyaenidae* consists of four species falling under the suborder of *Feliformia*. The family include four species: the brown hyena (*Parahyaena brunnea*), the striped hyena (*Hyaena hyaena*) and the spotted hyena (*Crocuta crocuta*), and the aardwolf (*Proteles cristata*). This study focuses on the spotted hyena (*Crocuta crocuta*), which is a social, scavenger species that lives in matrilineal based groups[1]. The aardwolf belongs to a separate subfamily of the *Hyaenidae* family, and it subsists on termites, whereas the other species within the *Hyaenidae* family use hunting and scavenging for surivival[2]. *C. crocuta* are known to be highly resilient to disease[3]. Despite their exposure to pathogens as a scavenger species, they rarely become infected[4], and they are often found to be asymptomatic in the cases when they are infected[5].

Social Structure and Hierarchy

The social structure is matriarchal with members formed into clans[6], which consists of transient males and philopatric females[1]. Similarly to baboons and primate species, the *C. crocuta* engage in highly social dynamics and competitive behaviour both with their fellow clan-mates and members that are not related to their clan[7]. Use of resources within a clan (such as food, space, mating partners, etc.) is determined by the rank of the individual within the clan[7]. Clans are sometimes found living near human settlements and scavenge on human waste products[8], which results in pathogen exposure and other human influences. Due to the regular exposure to pathogens through scavenging and predation of infected prey, the unusual occurrence of disease suggests that they may have evolved a robust immune system.

Exposure to Pathogens through Scavenging

Hyenas belong to the category of predatory animals, which appear to be highly resistant to diseases, displaying the capability to hunt and eat a large variety of animals, garbage, and dung. The hyena is known as an opportunistic scavenger, however, the majority of their calories are still collected via hunting[8]. Furthermore, spotted hyenas are highly efficient, in terms of their consumption, with certain studies determining that they eat the majority of the animal, leaving very little remaining organic matter on the carcass, with only the hair products being excluded[8]. In fact, their role as a scavenger species serves mainly to remove infectious material from an ecosystem, thus preventing the exposure of less resistant species.

Acquisition of Spotted Hyena RNA-Seq Data

There has been relatively little published regarding hyena sequence information. Fortunately, we were able to obtain transcriptomic data for the spotted hyena, as part of a previous project. This presented an opportunity to study this organism and further develop the knowledge base around this species of hyena, which could be a benefit for future research, such as investigation of disease resistance, or pathogen movement within an ecosystem, given the high disease resistance found within the spotted hyena, and from their regular exposure to pathogens. Furthermore, we attempted to improve upon these data with the inclusion of proteomic data.

Problem Statement and Research Questions

Rationale

The hyena is an apparently highly disease-resistant scavenger/predator, acting to remove carcasses from the environment. Information regarding the immune system of this species may improve our understanding of mammalian disease resistance.

Problem Statement

The hyena is an organism that, at the time of writing, has only recently had sequence data published to public databases, as part of a whole genome study by Yang et al[9], which greatly expanded the representation of this species in sequence databases and which can facilitate future research in the spotted hyena. Understanding sequence and function would advance knowledge of the biology of the species. The study described in this thesis is different in that it attempts to combine proteomics and transcriptomics to validate genome annotations and establish gene expression data for the spotted hyena.

Research Questions

Can a well-annotated transcriptomic assembly be produced, *de novo*, for the spotted hyena?

Can proteomic data be extracted from spotted hyena tissue and analysed by liquid chromatography tandem mass spectrometry (LC-MS/MS)?

Can transcriptomic data be used in conjunction with proteomic data to improve the quality of annotation for this species?

Hypothesis

Extracted and sequenced spotted hyena (*C. crocuta*) data can provide the foundation for an accurate transcriptome assembly through *de novo* assembly and be validated via proteomic data derived from the same species.

Aim

The aim of this study was to produce a well-annotated transcriptome assembly of the *C. crocuta* which has been validated via proteomic data.

Objective

1. To produce a well-annotated assembly from short sequencing reads of cDNA using a *de novo* assembler, and to evaluate assembly quality.

2. To describe the extraction of peptides and subsequently derived protein information from spotted hyena tissue in order to acquire useable proteomic data for proteogenomic assessment of gene expression.
3. To perform proteogenomic validation of the RNA assembly using the protein data in order to ensure a high-quality assembly.

Chapter 3

Literature Review

The Sequencing Boom and the Challenges that Followed

DNA Sequencing is a well-known, well documented set of processes which can be used to determine genomic sequences from DNA extracted from tissue samples. This process permits the direct examination of gene sequences, or the genome through assistance from an assembly tool. These data allow the researcher to gain a better understanding into the function, placement, and expression of genes.

Sequencing technology is generally divided into two categories, first generation, more commonly known as Sanger Sequencing, and next generation sequencing (NGS), which can also be described, more accurately, as massively parallel sequencing. Next generation sequencing technology is a broad field, which cannot be classified with a singular category. For the sake of brevity, in this review, all technologies post Sanger will be classified as NGS.

Rapid advancements in sequencing[10] have provided new avenues for studying biological processes. For example, there has been an increase in whole genome sequencing, which was originally a long and expensive endeavour[11]. An example of where this is used is in diagnosis, which has allowed for personalized medicine and can even possibly allow for the determination of patient predisposition to certain diseases[12]. Furthermore, WGS allows for a high level of discrimination, such as differentiating between antibiotic resistance in some bacterial organisms[13]. Another example is the use of NGS techniques in personalized medicine, which can improve disease diagnosis through discovery of gene alterations[14]. The uses of NGS also include understanding viral processes[15] or sequencing non-model organisms, which is used in situations where there is very little genomic data available for the organism[16]. NGS processes can also be applied to determine the transcriptome of an organism[16]. The transcriptome consists of the coding and non-coding RNA sequences which are transcribed from an organism's DNA.

The development of sequencing brought significant changes in research approaches when it was first introduced. The Human Genome Project[17] is a well-known example that began in the 1980's[17] and it can be argued that this was the push needed for advancement in sequencing technology[18]. Following the completion of the Human Genome project, the advances made were used to improve the existing sequencing technology, which resulted in rapid progress in the field[19].

The initial processes involved whole genome shotgun sequencing, which was performed using Sanger Sequencers[20]. The process behind this is shown in Figure 3.1. Sanger utilized a process known as chain termination[21]. This process involves the use of labelled dideoxynucleotide triphosphates (ddNTP), which are used in DNA polymerase reactions alongside normal deoxynucleoside triphosphates (dNTP), as well as a primer for a select DNA segment. During the extension process, when the reaction randomly incorporates a ddNTP instead of a dNTP, it results in a termination of the reaction for that specific strand, while continuing the reaction along other strands until they incorporate ddNTP. In this way sequence variants are produced, each with a different length[21]. Four parallel reactions may be carried out, each using a single type of ddNTP. The sequence can be inferred by running each reaction on a separate well in a polyacrylamide gel and determining the position of the labelled ddNTPs[21]. The detection step depends on the label used, such as autoradiography, in the case of radiolabelling[21]. In the case of fluorescently labelled ddNTPs, however, the process may be carried out in the same well or container. However, this was an older method used by Sanger, with the current process being known as capillary electrophoresis. The basic process behind this involves using a series of ddNTPs that are inserted into wells that are connected to a negatively charged cathode. Ultra-thin capillary tubes are then used to connect the cathode component to a positively charged anode. When a high voltage is applied, it leads to the migration of ddNTPs through the capillaries. As above, the smaller fragments move more quickly than the larger fragments. Any dyes on the nucleotides are detected by a laser and a photometer is used to read the fluorescence.

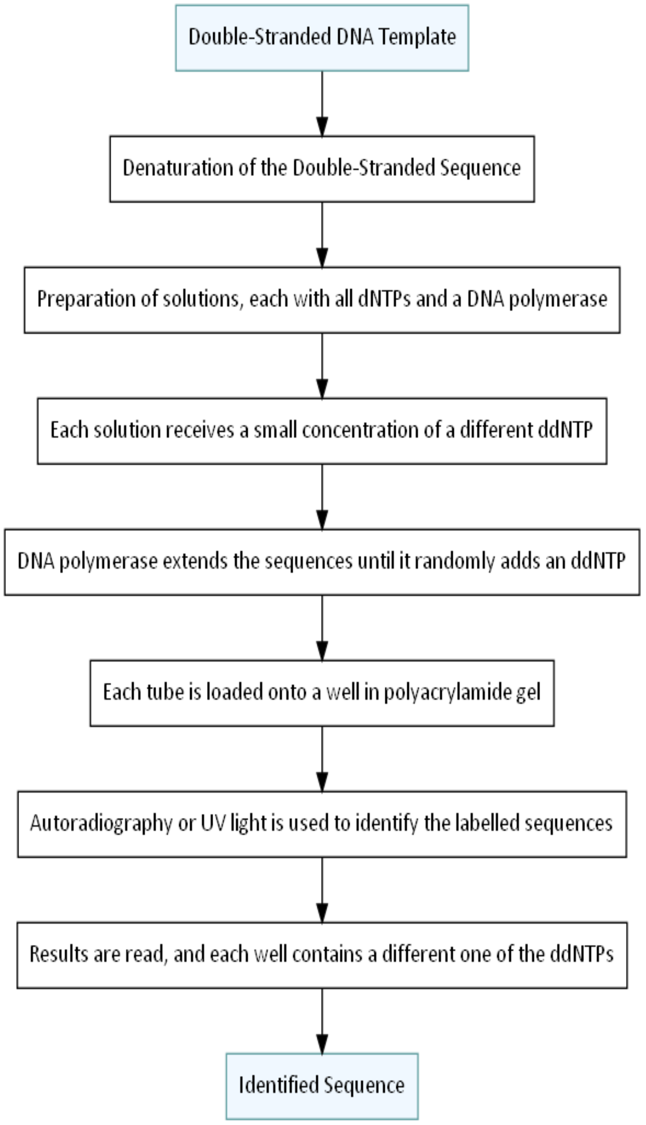
Double-Stranded DNA Template
Denaturation of the Double-Stranded Sequence
Preparation of solutions, each with all dNTPs and a DNA polymerase
Each solution receives a small concentration of a different ddNTP
DNA polymerase extends the sequences until it randomly adds an ddNTP
Each tube is loaded onto a well in polyacrylamide gel
Autoradiography or UV light is used to identify the labelled sequences
Results are read, and each well contains a different one of the ddNTPs
Identified Sequence

Figure 3.1. Basic Sanger Sequencing Procedure. The technique utilizes the physical properties of dideoxy nucleotide triphosphates (ddNTPs), which leads to termination DNA polymerisation. Each of the ddNTPS used are labelled in some fashion, such as with fluorescence, or with radiolabelling.

Sanger sequencing was eventually succeeded, though not replaced, by NGS (Next Generation Sequencing)[20]. One of the differences between Sanger and the more recent NGS technologies is that NGS carries out many millions of sequencing procedures in parallel[22]. This method usually produces many more shorter reads than Sanger sequencing. The end result is considered more accurate and possesses a higher coverage of the

DNA strand, due to the larger quantity of reads. In addition, this process occurs at a faster rate and can be performed for both forward and reverse sequences simultaneously[23].

Both Sanger and NGS processes have their own advantages and limitations. NGS systems, for example, are more rapid than the Sanger systems. However, these advantages come with the limitation of very short reads (in the range of 25 – 300 bp[24], in contrast to Sanger with 600 – 900bp[25]), when using NGS, which makes assembly of sequences more challenging[22]. The reason for this is that the means by which the technology works is by producing millions of shorter reads which can be a challenge to map to a genome due to repetitive sequences, high levels of guanine and cytosine, or a high degree of structural variation[26]. Alternatively, Sanger sequencing was once regarded as the ideal means to detect novel mutations within a sequence, however these advantages came with limited sensitivity, as well as an inability to carry out multiple parallel sequencing procedures[27]. One of the challenges that arose from the new sequencers was the ability to generate data at a greater speed than the ability to analyse it, and thus a large quantity of available data which required further study[19].

More recently, there has been a shift to a new variant in next generation sequencing that returns to the use of longer reads and a relatively high number of reads (approximately several hundred thousands, when compared to the millions produced by a technology such as Illumina[28], and the tens of thousands when compared to Sanger[29]), but is believed to overcome many of the difficulties that were originally associated with the original long read sequencers[26]. It is suggested that the longer read length may overcome the challenges with repetitive regions associated with short read sequencers, and because of the read length, can improve mapping certainty and ability to detect structural variants [30]. The reason behind this is due to the length of a standard nucleic acid sequence, which when reduced and re-assembled from short read sequences might lead to a loss of quality in the form of mapping certainty and isoform identification[30].

Another sequencing approach is RNA-Seq technology, which forms part of transcriptomics. Transcriptomics is the process used to determine and catalogue all RNA from a particular source in order to gain an understanding of the coding regions, the dynamic quantities of particular transcripts, and post-transcriptional modifications within the selected RNA[31]. RNA-Seq is a process whereby a RNA library sample is converted to a cDNA library, adapters are ligated to either end of each fragment within the library, and this is subsequently analyzed via NGS technology. RNA-Seq is utilized as a means to determine novel transcripts[32], while identifying the splicing patterns present within the sequence[33]. DNA sequences tend to be challenging to assemble *de novo*, due to the length of the sequences used, which can be too short to be able to bridge repetitive regions, thus exposing the assembly to possible errors in the form of fragmentation[34].

Although sequencing generation technology has allowed improved genomic analysis, this requires the ability to assemble and interpret these sequences. Therefore, it is essential to understand the aspects of data analysis that follow the sequencing process.

Assemblers

Assembly is the process of inferring a long sequence (contig) from short reads by recognizing the overlap between pairs of reads. Numerous types of assemblers exist, such as SPAdes[35], MIRA[36], Trinity[37] and SOAPdenovo[38] to name a few, each with their own specific bias and strengths. These assemblers often fall into one of two basic types, either reference-based assembly, or *de novo* assembly, with variations on the read length and/or single or paired end reads. The ultimate type of assembler selected for any project is highly dependent on the format of the sequence data, whether the reads are based on DNA or RNA, which is sequenced via cDNA, as well as the organism from which they originated. Assembling sequence data requires knowledge of the quality of the sequence data, as well as whether the source of the data was RNA or DNA.

Genomic assemblers are used with DNA. These assemblers have to account for challenges such as the spacing and length of repetitive regions, as well as the short reads used in the majority of sequencing in the present day[39]. Liao *et al*, in a recent study which attempted to solve some of the difficulties within *de novo* assembly, described the phenomenon of sequencing bias. Sequencing bias was described as the likelihood of a sequencer to introduce errors, such as substitutions, insertions and deletions, within a particular region of a sequence depending on GC or AT content[39]. In terms of genome assemblers, a few examples include ABySS[40], SPAdes[35] and DISCOVAR[41].

Transcriptomic assembly, using RNA data, is regarded as the more challenging of the assembly processes, in some cases, due to differences between contigs, which affects overall read coverage[42]. Available resources, such as computational power and server access, create a further limitation on transcriptomic assemblies[43]. These resources are especially necessary in the case of *de novo* assembly[43], which often requires a substantially larger amount of server space and random access memory (RAM), and which can process for longer periods of time[39].

Reports in the literature have attempted to compare the different assembly technologies, using different parameters and limitations[44]. Any comparison between assemblers, based purely on quality determined by external tools, would not necessarily provide an accurate definition of the properties of the various assembly tools. Various limitations must be considered, such as the sequencing tool used, and the organism examined. MIRA[36] is one such example of this. The MIRA tool, which is an acronym for Mimicking Intelligent Read Assembly, was specifically built to be used for smaller assemblies. The authors state, within the MIRA website, that for an assembly that is larger than approximately 100 megabases or more than 20 million reads, users should consider using a different assembler. These are limiting factors to any comparative study, and therefore, these studies can be highly specific to a given set of parameters[42]. When comparing the quality of an assembly, there is an argument that current methods for comparison are insufficient, with a large amount of emphasis placed on the contig length[45]. However, the full quality of the assembly must also take into account the sequencer used to produce the data. An assembler designed specifically for Sanger sequenced reads will not necessarily perform well against an assembler designed for the shorter read lengths produced by NGS technology.

Variation Between Assembly Strategies

The selection of the assembler depends on the format of the data examined as well as the type of assembly performed. This includes reference-guided assembly, *de novo* assembly, or a hybridization of both methods. Whether there are data available, along with the data type, can change the quality of the resulting assembly.

Reference-guided assembly, or reference mapping, is a process which relies upon pre-existing data for a particular organism[46], which must be a reference dataset derived from a closely related organism[47]. Some of the major advantages associated with this strategy are a lower error rate, which has been described as being below 1/10000, and the ability to use this process to detect divergence between individuals of the reference species[48]. There are two main strategies for this type of assembly. The initial approach makes use solely of the reference genome and the reference-guided assembler, where the reads are mapped against the reference genome, which provides the basis for the final consensus of the assembly[49]. The alternative strategy makes use of a *de novo* assembler in order to first produce the contigs before this is mapped against the reference, which serves to identify errors and correct them in line with the reference[49]. Regardless of the strategy utilized, however, the reference genome will induce a level of bias into the resulting assembly[50], although it has been suggested that a means to reduce this bias is to introduce various references, which are representative of different individuals or strains of the organism[48].

Within *de novo* (reference-free approaches) assembly, there are four algorithms which are most commonly used and a final variation which serves as a hybrid of the others[23]. These strategies are De Bruijn Graphs (DBG), Overlap Layout Consensus (OLC), string graphs, and greedy algorithm[23]. The hybrid variation is not a separate algorithm but a mix of the others which is meant to account for the limitations within the individual algorithms[23].

De Bruijn Graph strategies are specifically suited to the newer line of technologies, such as NGS. It makes use of shorter reads by first shortening them into fragments of a similar length and arranging them around a node[51]. Despite the association of shorter reads and NGS, it has been suggested that it could also be reliably applied with Sanger sequencers[52]. Each node within this graph contains the overlapping fragments of a length *k*, which are arranged so that the next fragment overlaps at a rate of *k-1*[53], known as a de Bruijn graph. This requires that these *k* length fragments, which constitute what is known as kmers, must be capable of aligning to the previous node and overlapping with it for the majority of nucleotide bases. The underlying sequences of these fragments, or kmers, is the eventual end product resulting from the complete overlap between all the sequences[53].

Overlap Layout Consensus is similar to DBG in that it also seeks consensus among the reads, however, it does not shorten them or break them into lengths of *k*[51]. OLC considers each of the reads as a separate node on the graph, while the nodes are connected based on consensus[53]. Unfortunately, this process is not suited to shorter reads produced from NGS[53]. Another aspect is that it tends to be a slower process compared to DBG[23]. The reason behind this may be attributed to large amounts of data introduced by the NGS strategies, which subsequently increases the amount of data which must be computed[54]. However, this strategy is more successful with Sanger sequencing technology[23].

String Graphs share similarities with both OLC and DBG. Their functions are to create stepwise overlapping sequences from the reads[55]. Similar to the DBG, it uses the overlapping sections and converts them into singular sequences based on the overlaps[56]. Where it differs, however, is that it does not first convert the reads into the smaller kmers, and also removes transitive regions on the sequence in the process[55].

Greedy strategies work by first utilising the shortest reads available. When it detects overlap between these sequences, it begins to form contigs[23]. Each sequence adds the next best aligned sequences and so forth, until no more sequences remain to be joined in consensus[57]. This approach may be described as similar to OLC, however, there is simultaneous alignment and consensus during the graph construction[58]. Assemblers that utilise this algorithm often appear to have a high degree of memory consumption due to how they store their reads[59].

Despite the rise of data assembly technology, it has been suggested that many final assemblies (in the case of genomes, at least) are often disorganised and incomplete[60]. Alhakami *et al* describe a complete assembly as a single sequence per chromosome[60]. They define this as being due to the difficulties associated with the actual assembly process, and the algorithms used[60]. In figure 3.2., the basic characteristics of the two main assembly strategies are depicted. This consists of overlap layout consensus and De Bruijn graph methods. The majority of assembly tools use a variation of one of the two algorithms.

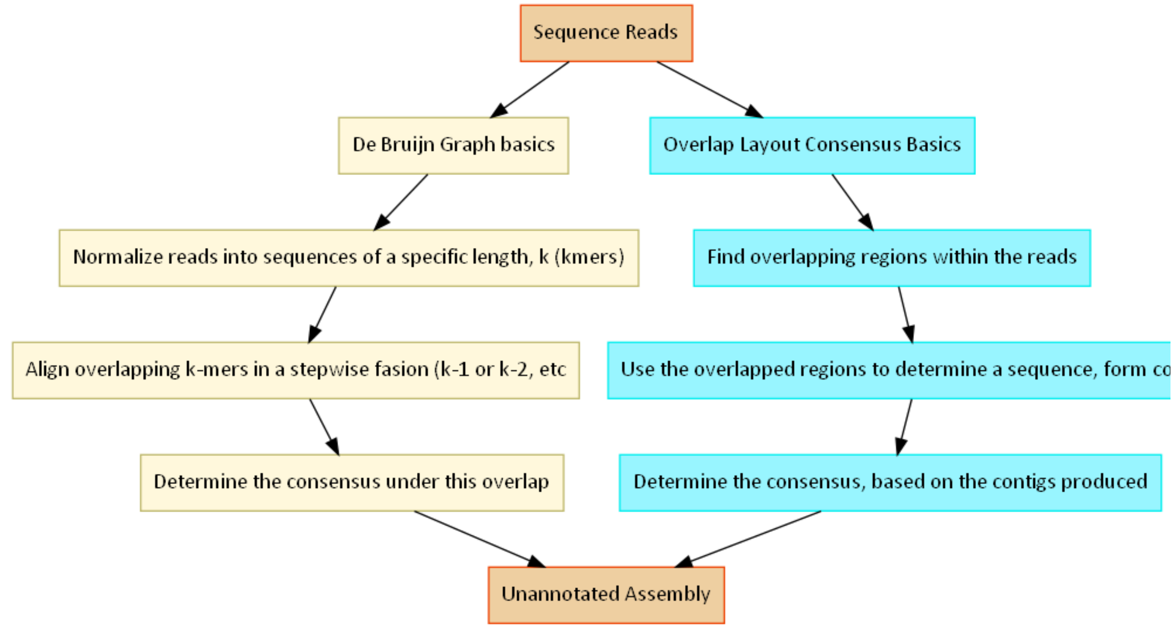
Sequence Reads
De Bruijn Graph basics
Overlap Layout Consensus Basics
Normalize reads into sequences of a specific length, k (kmers)
Find overlapping regions within the reads
Align overlapping k-mers in a stepwise fasion (k-1 or k-2, etc
Use the overlapped regions to determine a sequence, form co
Determine the consensus under this overlap
Determine the consensus, based on the contigs produced
Unannotated Assembly

Figure 3.2. Basic Steps to Assembly via the Major Algorithms. The diagram above depicts the basic process behind the two major algorithms used for the assemblers, De Bruijn Graph and Overlap Layout Consensus. The major difference between the two is depicted between how it interacts with the reads initially, with De Bruijn Graphs first reducing read size to kmers prior to alignment.

De Novo and Reference-Guided Assembly - Background

A decrease in the cost of sequencing and the development of *de novo* assembly have made the process more viable over the last few decades. This has made assembly the first step of many analyses, as the short-read fragments must first be connected into readable data prior to further analysis[61].

Reference-guided assemblies are based on similarity between the target and reference genomes[49]. In the case that little information is available, the reference sequences can include closely related organisms[62] which may be utilized to assist in some areas of the assembly. The more closely related the species, the more likely that target sequences will map to a larger component of the reference organism[49].

Current assembly technology has been developed around the new data formats, making use of the shorter reads encountered with NGS[63]. Despite this, the *de novo* and reference guided assembly technologies are not equal, with the current *de novo* technology experiencing a loss of quality when compared against reference-based tools[64]. This may be due to some of the challenges experienced by *de novo* assemblers, such as how they handle shorter read lengths, and shorter insert lengths, which are known issues, specifically with highly repetitive sequences[47]. Furthermore, any contamination can lead to an error prone assembly.[47]

Assembly Process

Assembling a collection of reads has several fundamental steps that are essential to the process. Data acquisition precedes all bioinformatic analysis, which means that the DNA or RNA may need to first be extracted and subsequently processed via the sequencing tool of choice[65], or alternately, the data could be publicly available for use[44]. In the case of newly acquired data and public data, the raw data may have primers, which are additional bases included on a sequence, generally for ligation purposes, known as adapters. These are removed through a process known as trimming, as they can introduce errors in read alignment and base calling[66]. The assembler must be selected based on the desired criteria, whether the assembly is single or paired-end[67], and whether the assembly has any prior data[49] or is a non-model organism, thus requiring a *de novo* assembly method[67]. Following assembly, it is necessary to assess the assembly produced, which can be carried out via several statistical approaches and tools to assess the completeness, as well as the base assembly statistics such as the contig count or the N50 score[68]. Annotation is the final step of an assembly, to ensure that the genes and coding regions have been identified.

Assembler Comparison

In this section, comparisons of a small selection of available assemblers are described, both for transcriptomic and genomic assemblers. The functions of assemblers, and how they relate to other assemblers, are discussed. More specifically, the challenges of selecting the correct assembly tool will be highlighted and why the tool in question should be selected with care.

Read Length as a Quality Determinant

With the use of NGS software, read length became shorter and the quantity of data became larger. However, this is not ideal for assembly because, as mentioned above, this can lead to the introduction of errors. The assemblers that utilize DBG algorithms are most commonly used for *de novo* assembly strategies. These are assemblers such as Trinity[37], Trans-ABySS[69] and Velvet[53]. Although OLC was commonly used in early sequencing tools, it is not considered suitable for *de novo* assembly of the shorter reads produced by NGS tools[53]. These include tools such as Celera[70], ARACHNE[71] and Atlas[72]. This is reported in a study by Li *et al*[51], where the authors compared the major assembly algorithms of OLC and DBG, and concluded that tools using OLC were better suited to the assembly of longer reads with lower coverage, while the DBG tools worked best with shorter reads, but with a higher coverage. This was not the final determining factor, as the authors also stated that the sequencing tools played a major role in determining the overall quality of the final assembly[51]. Another study by Zhang *et al*[73] reinforced the point regarding algorithm bias in assembler selection. The authors suggested that the OLC could be used for shorter read lengths but did suggest that this be used solely for smaller genomes, since OLC benefits from longer reads[73]. In conclusion, the read length is a necessary factor to consider when selecting an assembler, but context must also be given to the assembler itself, and whether it may be biased towards a particular type of data.

Sequencers as a Quality Determinant

The choice of sequencer was shown to impact quality during a comparison of assemblers used to process DNA reads collected from *E.coli* and sequenced using a Nanopore system, which produces read lengths of between 500 and 50000bp[74]. In this study, the outputs of three assemblers were compared using metrics such as N50 score, number of contigs, total length of contigs, and contig mean length. The researchers used the OLC algorithm, via Celera[70], and consistently achieved better results when compared to the other assemblers[74]. However, it must be pointed out that this study specifically focused on nanopore data, which is known to perform single molecule sequencing[75], and produce longer read lengths. This could create a bias towards an assembler such as Celera, [70] which is an OLC based assembler, and one which specifically focuses on the assembly of haploid sequences. This tool functions by focusing on variant detection and does not produce a singular consensus strand, but rather a set of consensus sequences. This characteristic may explain why it performed well in the comparative analysis, since this assembler was specifically designed for these types of sequences.

In another study by Yahave *et al*[34], a comparative analysis of two assemblers, SOAPdenovo[38] and SPAdes[35], was performed on assembled genomic data from species of the order *Hymenoptera*. The researchers tested these tools using both diploid and haploid data. These data were selected to determine the benefits and compare the quality of the results. They described their rationale for using multiple assemblers to accommodate any errors that might arise with any single assembler, such as missing sequences or bias. The authors concluded that the use of a haploid dataset improved the overall contiguity of the assembly produced[34]. However, in an article by Kleiman *et al*, the authors suggested that the diploid nature of organisms ensures redundancy of sequences[76], however, biases within the diploid assembly are also compounded, which would explain the higher degree of error in the diploid compared to the haploid assembly.

Quality of Alternative Algorithms

While the De Bruijn graph does see extensive usage in short read assembly, this does not, however, mean that the De Bruijn graph is the best method of assembly. There is an argument against De Bruijn graph algorithm based assemblers in that they lead to a significant increase in computational expenditure as the amount of data increases, and subsequently introduce errors into the process[58]. Another issue with De Bruijn graphs arises from how they handle repetitive sequences, especially those that are longer than any of the kmers they produce[77]. Due to these challenges, researchers have considered alternative algorithms, which improve upon this design. The greedy algorithm was introduced to overcome these challenges. BASE[77] is a tool which was developed based on this premise, and can assemble larger genomes, with longer NGS reads. However, it has been suggested that BASE does not perform as well as other assemblers when it does not have paired end information[77]. Furthermore, BASE was found to have a lower coverage in terms of contigs.

SSAKE[78] is another such assembler, which in contrast to BASE, uses shorter reads, specifically for assembling viral target sequences or short sequences. The assembler was capable of providing a suitable sequence from multitudinous smaller reads, which could represent the non-repetitive component of the genome[78]. However, there has not been a comparative analysis of how well this assembler performs when compared against other assemblers. This tool was developed during a period of time when *de novo* assembly was still a relatively recent process, so tools for comparative purposes may have been sparse[59].

Transcriptomics Assemblers

Transcriptomic data are often used for *de novo* assembly. This is at least partly due to the reduction in cost for RNA-Seq analysis and because RNA can provide a direct correlation with protein expression levels[79]. Examples of transcriptomic assemblers are Trinity[37], SOAPDenovo-Trans[80], and Trans-ABySS[69].

An Ideal *De Novo* Transcriptome Assembler

The ideal assembler should be capable of accounting for factors such as variable expression levels as well as isoformic variants and can efficiently derive a pattern from millions of diverse sequences. The quality metrics would depend on whether a reference genome is available or not. Yang *et al*[81] described two possible methods for assessing *de novo* assemblies. They mention that a purely metric based method can only describe the data, which would then require a degree of inference. However, without a reference, it cannot provide a concrete determination of whether the assembly fulfils all the necessary concepts, such as accuracy of the assembly, or the number of mis-assemblies within the sequence[81]. The second method, the use of reference models, is dependent on the quality of the reference used.

Other researchers have assessed the quality of different methods to assemble a non-model organism[82]. This evaluation was based on metrics, rather than model organisms, given that non-model organisms would not have a reference genome. These authors based a large part of their assessment on the basic metrics for the assembly, such as contig length and contig count, but also made use of BLASTX to determine how well represented the dataset was against a closely related species[82]. The metrics have been reported for the quality of *de novo* assembly techniques when approached with 454 data, which has a longer read length compared to Illumina systems[83]. Although, the examination was focused on the assemblers, rather than the final data, they extended the metrics analyzed to the N50 scores, as well as the time the assembly required to output a finished product.

Some tools provide metrics but can also employ alternative means of assessing *de novo* sequence data. These tools include BUSCO[84], Transrate[85] and Detonate[86]. An article by Hölzer *et al*[44] describes an extensive evaluation of various transcriptomic assemblers, where the selected assemblers were evaluated against several datasets, and analyzed using a variety of analysis tools, including BUSCO[84], TransRate[85] and DETONATE[86]. In this study[44], Trinity was statistically measured to be second best, below Trans-ABySS. Additionally, the top scoring tools, which produced the most consistent results in terms of assembly quality, included Trinity[37], Trans-ABySS[69] and SPAdes[35]. Another study[87], which specifically assessed how the kmers contributed to the assembly results, did not provide a conclusive score, but instead identified areas where each of the chosen assemblers contributed better results than the others. It was concluded that Bridger[88] was the preferred tool for their assembly, based on its consistently high score results, although this study was specifically focused on the effect of kmers in regards to sequence quality[87].

As can be observed in these studies, assemblers vary in quality and how close they are to the ideal criteria. Bridger[88] was designed as a combination of the methods used by a reference based assembler, Cufflinks and Trinity. This tool had improved transcript counts and reduced errors when compared against other assemblers[88]. The researchers reported improved metrics compared to all the tested assemblers, however, the analysis process seemed to only look at sensitivity.

Assembly Reconciliation

As mentioned previously, the majority of genomic assemblies are incomplete[89]. In this situation, assembly reconciliation may be used. This is the process of examining multiple assemblies, and converging the most representative components into a singular assembly[89]. When an attempt is made to reconcile various assemblies, the aim is to diversify the assembly through the use of different assembly tools, and merge or unify the assemblies in order to reduce the bias or weaknesses associated with the different tools[90]. Unfortunately, different assemblers will each introduce their own bias into the assembly[90]. Furthermore, the reduction of the assembly and removal of repetitive regions and, subsequently, unique sections within these repetitive regions, is an issue with using an unfinished or “draft” genome[90]. This section will provide a brief quality comparison of current assembly reconciliation methods.

Comparison of Assembly Reconciliation Output

The final outcome of assembly reconciliation is to produce an assembly which has better quality than the individual assemblies used[91]. This is accomplished through detection of errors and the use of various other assemblies to then repair these errors[91].

In 2008[91], a study was published which detailed a method for assembly reconciliation for genomic data. The focus was on merging draft genomes in order to improve the overall quality of the genome assembly[91]. This study was performed on *Drosophila* datasets, which were all tested before and after reconciliation, using N50 score and ‘CE statistic’, which was used to detect mis-assemblies[91]. The results showed that each of the post-reconciliation assemblies had a large increase in their metrics, and each of the reconciled assemblies had a large reduction in errors, as depicted by the ‘CE statistic’.

An additional comparison analysis was published for several of the currently used reconciliation tools[60]. In this study, although the benefits of the different reconciliation tools were described[60], the results could not provide a conclusive overall best quality reconciler tool. The authors tested assembly reconciliation using different assembly algorithms to determine whether high quality inputs were affected[60]. They found that each of the reconciled assemblies produced consistently better results than the original input assemblies[60]. Unfortunately, there are very few comparative analyses for assembly reconciliation software. This reflects the frequency with which this method is utilized, although it is a fairly recent development.

Proteomic Analysis

Proteomics consist of both the scientific approach as well as the technologies used for identifying and characterizing the complex mixtures of proteins for a given organism, environment, or tissue. This is a highly useful source of information regarding organisms and their biological functions[92]. The interest in proteomics has spurred the same sort of leaps in technology associated with genomic and transcriptomic work[93]. The technology behind proteomic acquisition has progressed over time, starting with the process of 2-dimensional gel electrophoresis. Eventually, the invention of tandem mass spectrometry (MS/MS) lead to further ability to identify proteins[94], especially when the two methods were coupled together[95]. Limitations within this method, however, included such as difficulties with the dynamic range of detection and increasing the number of runs required per analysis[95]. Thus, this saw the development of later, methods such as Two-dimensional liquid chromatography, when coupled with tandem mass spectrometry. This Two-dimensional liquid chromatography has subsequently lead to the decline in the older gel electrophoresis based methods.

Mass spectrometry is the means by which proteins are identified within a proteome[96]. It accomplishes this through detection of the mass-to-charge ratio of peptide ions and the fragments produced from peptide ions. Although a single mass analyser may be useful (as in GC-MS), multiple mass analysers may be connected in tandem to carry out tandem mass spectrometry (MS/MS). Each subsequent tandem mass analyser used in sequence increases the specificity of the analysis[96]. Tandem mass spectrometry events are separated in several ways, one of which is the kinetic energy of the ions, and the other is by analysis and excitation techniques[97]. The kinetic energy of an ion is an important factor, since the collision of an ion and gas molecule can produce different results, depending on whether the reaction is high or low energy[97]. Kinetic energy from the colliding particles is used to induce various reactions, including collision induced dissociation (CID), charge permutation and collision cooling[97].

Analysis and excitation events can be further separated based on the location they are performed, or the time they occur[97]. As an example, in a “beam” mass analyser (where ions are separated based on a spatial feature within the mass analyser[97]), the technique induces molecular collision to introduce fragmentation, and the analysis of these fragments may each be carried out in a separate mass analyser. In contrast, time based separation takes place more often within ion trap mass analysers, where each step may be carried out in a single mass analyser, but occur at different times in the analysis[97]. Aside from CID mentioned above, another method known as higher energy collision dissociation (HCD) is also used. HCD is found, for example, within LTQ Orbitrap, where fragmentation of the ions take place within a collision cell[98]. As per the name HCD uses higher energy than CID for dissociation, but through this induces a more diverse fragmentation[98].

The Various Mass Spectrometers

Quadrupole mass analysers, which have four symmetrical, parallel rods, have been used frequently in Liquid Chromatography Mass Spectrometry (LC-MS)[99]. This MS system works by filtering ions to retain only those within a narrow range of mass-to-charge (m/z) values[99]. In quadrupole MS systems, the ions continuously move through the system in a beam after forming in the ion source, detecting the ions that make it through the mass filter by their impacts on an electron multiplier detector, with the intensity of signal attributable to a particular m/z value[100].

Ion traps are a version of mass analysers which utilize magnetic fields or a low RF voltage through a ring electrode, which traps ions[101]. Ion traps have a high level of sensitivity and can be used to perform repetitive analysis on a sequence of ions[101].

Time-of-Flight (ToF) mass analysers determine the amount of time required for ions to traverse a flight tube of fixed length[99], from which the mass-to-charge ratio can be inferred[102]. ToF MS systems tend to generate a large quantity of spectra[102].

Fourier transform ion-cyclotron mass analysers (FTICRs) determine the mass-to-charge ratio by trapping ions to orbit within an ion trap; the frequencies in their induced current can be used to infer their m/z values very accurately[99]. These mass spectrometers are known to have a high level of accuracy and resolution, specifically due to the frequency parameters they measure[103].

The Orbitrap is a mass analyser variant of Fourier Transforms mass analyser[104] that makes use of electrostatic fields, which work to contain ions for analysis. The movement of ions up and down the spindle electrode occurs at a particular frequency for a given mass-to-charge ratio[104].

All mass analysers have their own biases and strengths associated with their usage. A mixture of different types of mass analysers can achieve differing results compared to using only a single type. Hybrid mass spectrometry functions using a combination of two or more mass analysers[105] connected in sequence. In a study by Michalski et al[106], they evaluated the performance of a hybrid mass spectrometer composed of a quadrupole and an Orbitrap system, known as a Q Exactive. The authors described that the system was only able to utilize Higher energy Collisional Dissociation (HCD), but found that, despite this, the system was not particularly limited and actually improved efficiency and speed of the system. Furthermore, they found that the technology could multiplex both MS and MS/MS mass ranges and did not display any notable limitations in this regard[106].

Fractionation Techniques

Fractionation chromatography separates peptides used by the mass spectrometers, which serves to increase the specificity of the peptides selected for analysis. Liquid chromatography serves several purposes in the MS/MS procedure and it produces an increased level of specificity and higher throughput[107]. The purpose of chromatography is to allow the desired peptides to be isolated and better identified and quantified[108].

Peptide fractionation is a procedure which is utilized to improve the quantification and identification of peptides within a sample. This is done via a reduction in the complexity of the components, which enables easier identification and differentiation[109]. It is used to separate peptides into smaller molecules via phase transition (basic change between physical states of matter) which reduces molecular complexity for analysis[110]. In this fashion a sample is reduced to several fractions, each with a different selection of peptides that could potentially increase peptide diversity, based on the sample used. This method is popular in mass spectrometry as it is highly reproducible, while introducing further selection parameters into the process[109].

Like mass spectrometry, it can be challenging to compare the quality of results from various techniques. This is especially true since fractionation can have different purposes depending on the sample. In this review, the focus will be on peptide fractionation as it is used in mass spectrometry for proteomic analysis. A comparative analysis of SDS-PAGE, strong cation exchange, and Off-Gel^{TM} isoelectric focusing was performed using *E. coli* and human datasets in order to diversify the testing samples[111]. These techniques are the commonly seen methods reported in the literature. Isoelectric focusing is a method that makes use of electrophoresis to separate amphoteric compounds[112]. Strong cation exchange is a form of ion exchange chromatography, which separates peptides based on the charge within their molecules[112]. Finally, SDS-PAGE is a separation technique based on the molecular weight of the samples. In this specific study[111], the authors acknowledged that due to the nature of the procedures, it could be challenging to perform a fair analysis between the samples. To compensate, they maintained the parameters for each of the techniques at optimum conditions, but kept other mitigating conditions, such as mass spectrometry time, constant. Their results showed that the strong cation exchange method seemed to have the best sequence coverage as well as highest amount of peptide identifications, although they largely attributed the identifications to a larger amount of protein sample used compared to the other methods[111]. Another study by Chiu *et al*[113], evaluated the efficacy of various fractionation strategies. They found the strong cation exchange process identified a large number of peptides, however, they compare the efficacy of fractionation techniques when salt was present in the samples.

Another separation technique, also known as reversed phase liquid chromatography, is one of the more common methods of separation used in fractionation[112] with the process being found to possess a higher resolving power and higher peak capacities than other methods[114]. According to a study by Tanveer *et al*[115] this strategy was found to be comparatively better than the more common strong-cation exchange (SCX) strategies.

A Combination of Proteomics and Transcriptomics

Proteogenomics is the application of genomic, or transcriptomic, and proteomic data combined using a bioinformatics approach. This may be used to provide improved annotation for an assembly[116]. According to Castellana *et al*[117] one of the major goals of any genome based project is to produce a collection of protein annotations for the sequence data. A large component of this work has been accomplished via the use of gene prediction, sequences of cDNA and comparative genomics[117]. Baerenfaller *et al*[118] found that using a proteome alongside genome based methods was beneficial for genome annotation and gene prediction. The insights into the amino acid sequences provided by proteomics can serve several functions, such as improving the gene models created by the genomic sequences[119], or leading to the discovery of novel coding domains[120]. One of the important functions attributed to proteogenomic analysis comes from its use in error correction, where the sequence is annotated again, using new proteomic data[120]. Inclusion of transcriptomics into this approach can also be used to validate the genomic data[116], while providing a means to determine the coverage of RNA against the DNA reference[121].

Proteogenomics have been used extensively in the field of comparative studies, specifically for variant detection, as seen in a study by Mertins *et al*[122], where they examined the variation between unmatched RNA and DNA data. Through this, they could detect various mutations which had not been seen when using just the RNA-Seq data. The researchers also only saw a low number of their genomic and transcriptomic variants confirmed via the MS/MS data. However, based on previous work, this low number of confirmations was not unusual[122]. Furthermore, this low level of variant confirmation seems to be a relatively common issue, as is noted by Lazar *et al*[123]. This is one of the known difficulties in proteomics experiments which involve sequence data. The reason for this could be multi-fold, as Lazar *et al*[123] suggests, including low sequence coverage to a lack of correlating data between the proteins and the comparative samples (RNA or DNA).

Despite challenges associated with variant detection via proteogenomics, it is still a valid tool for use. In a review by Nesvizhskii *et al* [119], a benefit of variant detection includes decreasing the investigative and resource requirements of researchers. Information is gained regarding variant peptide sequences, thus eliminating the need for extensive searching and preparatory lab work. Sequence correction and verification can play another role through proteogenomics. According to Castellana *et al*[117], they describe the means by which they determined the error rate of *Arabidopsis thaliana* gene models.

From the above it can be inferred that proteogenomics is a process whereby the chemical evidence is generated using a mass spectrometer, which may then be applied to either transcriptomic or genomic data available, and the predictions within those transcriptomic or genomic datasets. Furthermore, proteogenomics may also be considered all the tools which are used to detect and identify variations within amino acid sequences, through the use of nucleotide sequences[124].Fortunately, to aid in this, several tools exist which can facilitate this part of the research. Tools such as Spritz[124] and Galaxy[125] were built to function with proteogenomics studies. Galaxy, which is a platform for integrative analysis workflows, and possesses a large number of tools for procedures such as peptide spectrum matching and post processing[126]. One such tool associated with the galaxy platform is the web-based tool, Galaxy Integrated Omics (GIO), which allows for an easier means of working with proteomics studies wherein transcripts are used[126]. Another such tool is Spritz. Spritz is used to produce proteogenomic databases used to identify peptide variations[124].

Chapter 4

Materials and Methods

Animal Samples

Between 2011 and 2017, tissue samples were collected opportunistically during post-mortem examinations from two female and three male spotted hyenas (*Crocuta crocuta)*, which had been euthanized by veterinarians in the Kruger National Park, South Africa, which is endemic for *M. bovis.* Animals were immobilized via a tranquilizer gun using a plastic dart, which contained 5 mg kg^{-1} tiletamine-zolazepam (Zoletil; Virbac RSA (Pty) Ltd, Centurion, South Africa), or 0.5-1 mg kg^{-1} tiletamine-zolazepam along with 0.03–0.05 mg kg^{-1} medetomidine (Kyron Laboratories (Pty) Ltd, Benrose, South Africa). This component of the project was carried out as part of a separate project performed by Higgitt *et al*[127]. Permission and ethics approval for these projects were obtained from the Stellenbosch University and the South African National Parks service. The project number for this was ACU-2019-10347.

Title	Individual	Tissue Sample
Hyena 1	17/571	Brain
		Liver
Hyena 2	17/572	Testes
Hyena 3	17/575	Brain
Hyena 4	14/418	Unstimulated Whole Blood
Hyena 5	15/261	Unstimulated Whole Blood

Table 4.1. Samples Used for RNA-Seq Analysis. This table describes the individual hyenas from which samples were collected for RNA-Seq analysis and transcriptome assembly.

RNA sequence analysis

The process of sequencing the extracted mRNA was performed by the Centre for Proteomics and Genomic Research (CPGR) in Cape Town, South Africa. RNA was extracted prior to analysis as part of a separate project[127]. The samples used consisted of whole blood, testes, brain and liver samples, which had previously been stored frozen. The extraction procedure had taken place prior to delivery, however there is no indication as to how this procedure took place or how the tissue samples were processed, except that all samples were initially frozen and stored in RNAlater prior to processing.

In brief extracted RNA was prepared for sequencing using a TruSeq stranded mRNA kit (Illumina, Inc, San Diego, California, United States), as well as Ribo-Zero (Illumina, Inc, San Diego, California, United States). The Illumina nextseq 500 (Illumina, Inc, San Diego, California, United States) was used to analyse and produce the fastq files that were used for the assembly.

As per the CPGR analytical report, in preparation for the sequencing step, the libraries were normalized to 4 nM with 10 mM Tris-HCl (pH 8.5) and combined at equal volumes to obtain an equimolar library pool. Quality control for cluster generation, sequencing and alignment, was performed using the Illumina® PhiX library. For the sequencing process, a 4 nM concentration was prepared from a 10 nM stock solution, through the use of 10 mM Tris-HCl (pH 8.5) as the diluent.

The diluted equimolar library pool (4 nM) and the diluted PhiX control (4 nM) were initially denatured through the use of 0.2 N NaOH, prior to being neutralized (200 mM Tris-HCl, pH 7.0) and finally diluted to a concentration of 1.8 pM via a hybridization buffer (HT1). The PhiX control (1.8 pM) was inserted into the library pool (1.8 pM) at 1%, and placed onto a NextSeq 500/550 Mid Output Kit v2 (150 Cycle).

The Illumina® NextSeq 500 system was programmed so that it would perform a paired-end, dual-indexed 2x 76 cycle sequencing procedure. Configuration for the run was set to integrate with BaseSpace Sequencing Hub to ensure de-multiplexing and conversion into the FASTQ file formats.

The raw RNA-Seq FASTQ files have been submitted to the NCBI database and are publicly available at the following accession numbers: SAMN15877773, SAMN15877774, SAMN15877775, SAMN15877776, SAMN15877777, SAMN15877778.

Assembly and Evaluation of the Hyena Transcriptome

Individual Sample Assessment and Assembly

Once the sequences from the biological samples, in the form of fastq files, were received from the CPGR, they were examined using FastQC[128] (Babraham Bioinformatics) to determine if they were suitable for assembly. Adapter content[129] was examined and noted as being below 0.1% and was thus regarded as negligible by FastQC.

Each sequence of the individual samples was assembled using Trinity (Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA), version 2.8.4, using the parameters as listed below. Quality was determined using BUSCO[84] (Benchmarking Universal Single-Copy Orthologs)(Department of Genetic Medicine and Development, Swiss Institute of Bioinformatics, University of Geneva Medical School, Geneva, Switzerland), version 3, and TransRate[85] (Department of Plant Sciences, University of Cambridge, United Kingdom; Department of Computer Science, Stony Brook University, USA; Department of Plant Sciences, University of Oxford, United Kingdom), version 1.01, using the default criteria as detailed on the respective websites and manuals, using eukaryota_odb9 super kingdom dataset.

BUSCO[84] is a tool that searches for genes that it assumes will exist within the assembly. These genes are single copy orthologues, which were found to exist within the majority of the species selected to form part of the library[84]. The BUSCO[84] process requires a lineage to be selected as part of the analysis. BUSCO[84] was run in every instance using mainly the default features for a transcriptomic dataset, using *eukaryota_odb9* super kingdom orthologue dataset, as listed below, drawn from the BUSCO website.

TransRate[85] assesses the quality of the assembly based on the alignment of the original sequenced reads from which it was originally assembled. This tool measures the accuracy and completeness for both the overall assembly, as well as the individual contigs within the assembly[85].

Before each assembly, the fastq files for the R1 and R2 category were collected and concatenated into individual files, representative of the R1 and R2 groups, using the Linux *cat* command, in bash scripting. Each sample was run using a limited amount of memory in order to allow for multiple samples to be run at the same time.

The quality assessment performed via BUSCO[84] (and TransRate[85]) was carried out using the default parameters and was run using a mammalian lineage, besides the eukaryote lineage mentioned above, as provided by the BUSCO website. TransRate[85] was similarly run using default settings, using the concatenated fastq files for the final assembly and 4 cores with no specified RAM limit

Final Assembly

Trinity[37] was run on all the samples, using a similar script as that mentioned above. The script above represents the use of Trinity to assembly the entire collection of reads from each of the individual tissue samples. The "--*left*" and "--*right*" functions represent reverse and forward fastq files, while the values used for the two commands are representative of the path to the concatenated fastq file for each of the reverse and forward file types. It was run using more RAM, and more cores as more time was available for assembling the single assembly, as opposed to the individual assemblies. This provided more resources for the main assembly when compared to the individual assemblies. The assembly products were analyzed with the BUSCO [84] and Transrate[85] tools, using the same configuration as the individual samples assembled above.

Alignments

ProteinOrtho[130] (http://www.bioinf.uni-leipzig.de/Software/proteinortho) was utilized to determine the overlapping sequences between the main assembly and orthologous sequences. This was performed against the closely related organisms (house cat, leopard, cheetah, tiger), as well as the Trinity assembly, which had been converted into peptide sequences via TransDecoder[131].

This indicates that the number of cores provided were 1 and the alignment tool used was BLASTP+ (National Center for Biotechnology Information, https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download)[132]. The *samples.fasta* represents the fasta files containing the sequences from the organisms or datasets included within the analysis. Each of the datasets used were first clustered using CD-HIT[133] to reduce the level of redundancy within the closely related species dataset and the assembly protein translation. This was performed using the default CD-HIT[133] (http://weizhongli-lab.org/cd-hit) script at a sequence identity threshold of 0.95. The results from this analysis were visualized using the UpSetR[134] package, on Rstudio.

Tandem Mass Spectrometry and Fractionation

For protein acquisition and analysis, 15 tissue samples were sent to the CPGR, after which proteins were extracted from the tissues and subsequently analyzed via LC-MS/MS (Liquid Chromatography Tandem Mass Spectrometry).

For the protein extraction step, tissue pieces were removed using a scalpel. The removed sample was weighed, ensuring a range between 80 – 150 mg. Samples were washed in 1 ml of Phosphate Buffered Saline (PBS, Sigma P4417) and transferred to tubes containing Zirconium beads (Benchmark Scientific D1032-15) where 400 µl of extraction buffer (50mM Tris-HCl (Merck 1.08382.0500), 150mM NaCl (Sigma 13565) pH 7.5, protease inhibitor (Thermo Fisher 1861278)) was added. Samples were homogenised at 4 m/s for 45 seconds using a BeadRuptor (Omni International, USA). Homogenisation was repeated once more. Following homogenisation, 100 µl of 10% Sodium dodecyl sulphate (SDS, Sigma 71736) was added and samples were stored at 95°C for 10 minutes. Once cooled, 500U of Benzonase (Sigma E8263) was included in each sample and incubated for 20 minutes at room temperature to remove nucleic acids. Following this, samples were centrifuged at 10 000 x g for 10 minutes. The supernatant was transferred to a new tube and quantified using the QuantiPro BCA assay kit (Sigma QBCA) according to the manufacturer’s instructions.

LC-MS/MS was followed by high pH reverse phase fractionation. Only a single sample was used, which displayed the largest quantity of distinct peptides. Fractionation was followed by another 12 LC-MS/MS analyses on this fractionated sample. The tissue samples sent for the first step of LC-MS/MS consisted of abdominal, head, peripheral and thoracic lymph nodes, and liver from three different hyenas. As per the CPGR analytical report, the analysis was performed on a Thermo Q-Exactive hybrid quadrupole-orbitrap mass analyser (Thermofisher, Waltham, Massachusetts, United States), which was coupled with a Dionex Ultimate 3000 nano-UPLC system. Acquisition of data was accomplished using an Xcalibur v4.1.31.9, Chromeleon v6.8 (SR13), Orbitrap MS v2.9 (build 2926) and Thermo Foundations 3.1 (SP4). Peptides were first dissolved in 0.1% Formic Acid (FA, Sigma56302), 2% Acetonitrile (ACN, Burdick & Jackson BJLC015CS) and subsequently loaded on a C18 trap column (PepMap100, 9027905000, 300 µm × 5 mm × 5 µm). The peptide injection amounted to approximately 400 ng. Samples were first trapped on the column before being washed for 3 minutes. After this, the valve was switched, and peptides eluted onto the analytical column as described below.

For chromatographic separation, a Waters nanoEase (Zenfit) M/Z Peptide CSH C18 column (186008810, 75 µm × 25 cm × 1.7 µm) was used. The solvent system employed was solvent A: LC water (Burdick and Jackson BJLC365), 0.1% FA and solvent B: ACN, 0.1% FA. The multi-step gradient used for peptide separation was generated at 300 nL/min as follows: time change 5 min, gradient change: 2 – 5% Solvent B, time change 40 min, gradient change 5 – 18% Solvent B, time change 10 min, gradient change 18 – 30% Solvent B, time change 2 min, gradient change 30 – 80% Solvent B. The gradient was then held at 80% Solvent B for 10 minutes before returning it to 2% Solvent B and equilibrating the column for 15 minutes. All data acquisitions were obtained using Proxeon stainless steel emitters (Thermo Fisher TFES523).

The raw mzml files have not yet been submitted and are still in storage at the time of writing this thesis.

The table below provides a summary of the samples sent to the CPGR for analysis.

As per the table, the samples were selected based on the premise of the project, which was to assemble an accurate and well-annotated transcriptome for the hyena. Disease resistance plays a part in this project, as it relates to the natural resistance of the spotted hyena, and thus lymph node proteins were felt to be important in gaining an accurate representation of the hyena.

Organism No.	Sample	Sex	Age
17/571	Abdominal Lymph	Male	Young Adult
17/572	Abdominal Lymph	Male	Young Adult
17/575	Abdominal Lymph	Female	Young Adult
17/571	Peripheral Lymph	Male	Young Adult
17/572	Peripheral Lymph	Male	Young Adult
17/575	Peripheral Lymph	Female	Young Adult
17/571	Head Lymph	Male	Young Adult
17/572	Head Lymph	Male	Young Adult
17/575	Head Lymph	Female	Young Adult
17/571	Thoracic Lymph	Male	Young Adult
17/572	Thoracic Lymph	Male	Young Adult
17/571	Liver	Male	Young Adult
17/572	Liver	Male	Young Adult
17/575	Liver	Female	Young Adult
17/572	Lung	Male	Young Adult

Table 4.2 Summary of the samples submitted to the CPGR for proteomic analysis. Lymph node tissue was submitted based on availability.

High pH reverse phase fractionation was used prior to the second set of LC-MS/MS analyses, using a pooled head lymph node sample that had maximized the number of distinct peptides in unfractionated experiments. This accomplished by the CPGR through the use of a Dionex Ultimate 3000 micro-HPLC system for High pH reverse phase fractionation. Solvent A: Millipore water, 20mM Ammonium Hydroxide (Sigma 338818) and Solvent B: Acetonitrile, 20mM Ammonium Hydroxide was utilized for the solvent system. Finally, 120 µg of peptide was injected onto a Phenomenex Gemini C-18 column (00F-4435-B0, 5 µm x 150mm x 2mm) was utilized for fractionation. UV detection was measured at 214nm, while fractions were collected in intervals of 60 seconds for the entire run time of the LC run. Following fractionation, the fractions at different gradients were combined and the combined fractions were dried before being resuspended in FA(8 µl of 2% ACN, 0.1%). These fractions were run on the Q-Exactive mass spectrometer as described prior. This produced 12 fractions.

Once the Raw files were received from CPGR, the analysis began by converting them into mzML files using MSConvert[135] (http://proteowizard.sourceforge.net/tools.shtml), and the peak-picking parameter from the Proteowizard[136] package. This was followed by analysis of the converted files by MS-GF+[137] (Department of Computer Science and Engineering, University of California San Diego, La Jolla, USA, https://github.com/MS-GF+/msgfplus/releases), using a tryptic and semi-tryptic search respectively, to ascertain which protein samples might be appropriate for further analysis, based on the distinct peptides and peptide matches present within the sample. This analysis also determined whether quality was consistent across both searches.

Protein sequences generated by Yang *et al*[9] were used as the database against which the spectra results were compared. Decoys were introduced to the databases by reversing the sequences drawn from the database in the MSGF+ tool. The original and reversed sequences are then concatenating into a single FASTA for search[137]. Reversed sequences, denoted by a prefix string on their accessions, make it possible to estimate the fraction of erroneous PSMs in a collection of identifications, typically output as an aggregate FDR. The FDR is calculated from the total number of decoys that were found divided by the total number of decoys that were smaller than a particular E-value threshold.

MS-GF+ was operated on a computer with 96Gb of Ram and 8 threads. The protein database FASTA used was the NCBI proteome produced by Yang et al[9], however, the assembly was translated via TransDecoder and used as the search database for the fractionated data, alongside the NCBI data. In each situation the MS-GF+[137] tool was run on the Deep Thought server, housed at Stellenbosch University, Tygerberg Campus.

The search analysis was likewise performed using the translated final assembly from RNA-Seq as a protein database file, providing only 3500M of RAM and using the decoy prefix of “rev_.” All other parameters remained the same, as per above.

The fractionated data was analyzed in a tryptic and semi-tryptic search against the proteins generated by Yang *et al*[9]. The script used for the fractionated files followed that used for the individual files, again using the decoy prefix "rev_." This was performed for the final assembly produced from RNA-Seq. A semi-tryptic search was also performed on the fractionated data, using the translated Trinity assembly, as above, using the decoy prefix "Cntm_".

Following acquisition of the mzid files from these analyses, they were examined using IDPicker[135], which provided the distinct proteins present within each sample. The mzid files were further examined using the MSnID (version 1.22.0, r version 4.0.2) R package, which was used to determine the evalue per peptide.

The amino acid sequences were drawn from the IDPicker[135] results for fractionated data, as well as the semi-tryptic and tryptic results of the initial search results. These were analyzed via ProteinOrtho[130], and subsequently interpreted via the R package, ggpubr (version 0.4.0), on Rstudio using r version 4.0.2, to determine the overlap between the different searches, and gain an idea of whether the fractionated samples improved sensitivity.

Finally, results from the semi-tryptic search of the fractionated and the individual sample data were compared using ProteinOrtho[130], using a subset fasta file collected from IDPicker[135] search tool. This was initially only used for determining the overlaps between samples searched using the same database. The analysis was extended to include the complete collection of semi-tryptic search results to determine the overlapping values between the searched data from the NCBI database as well as the *de novo* assembly. The assembly had been converted in protein sequences via the TransDecoder tool. In both cases using ProteinOrtho above, the -singles command, but otherwise using default commands. TransDecoder[131] is a tool that function by searching for several criteria. It searches for a minimum length open reading frame (ORF) within a transcript sequence. It tries to determine if the log-likelihood score is similar to that which is computed by GeneID software. The software determines if the score is highest when the ORF is scored in the first reading frame as compared to the two other forward reading frames. It then determines in the selected ORF is surrounded by another candidate ORF, and if this is the case, the larger ORF is selected[131]. By default the TransDecoder tool attempts to identify an ORF consisting of at least 100 amino acids in length[131].

Chapter 5

Results

Assembly Quality Control

Following the acquisition of the data from the CPGR as fastq files, the overall quality of the files was determined using fastqc analysis tool. Results are shown in Figure 5.1.

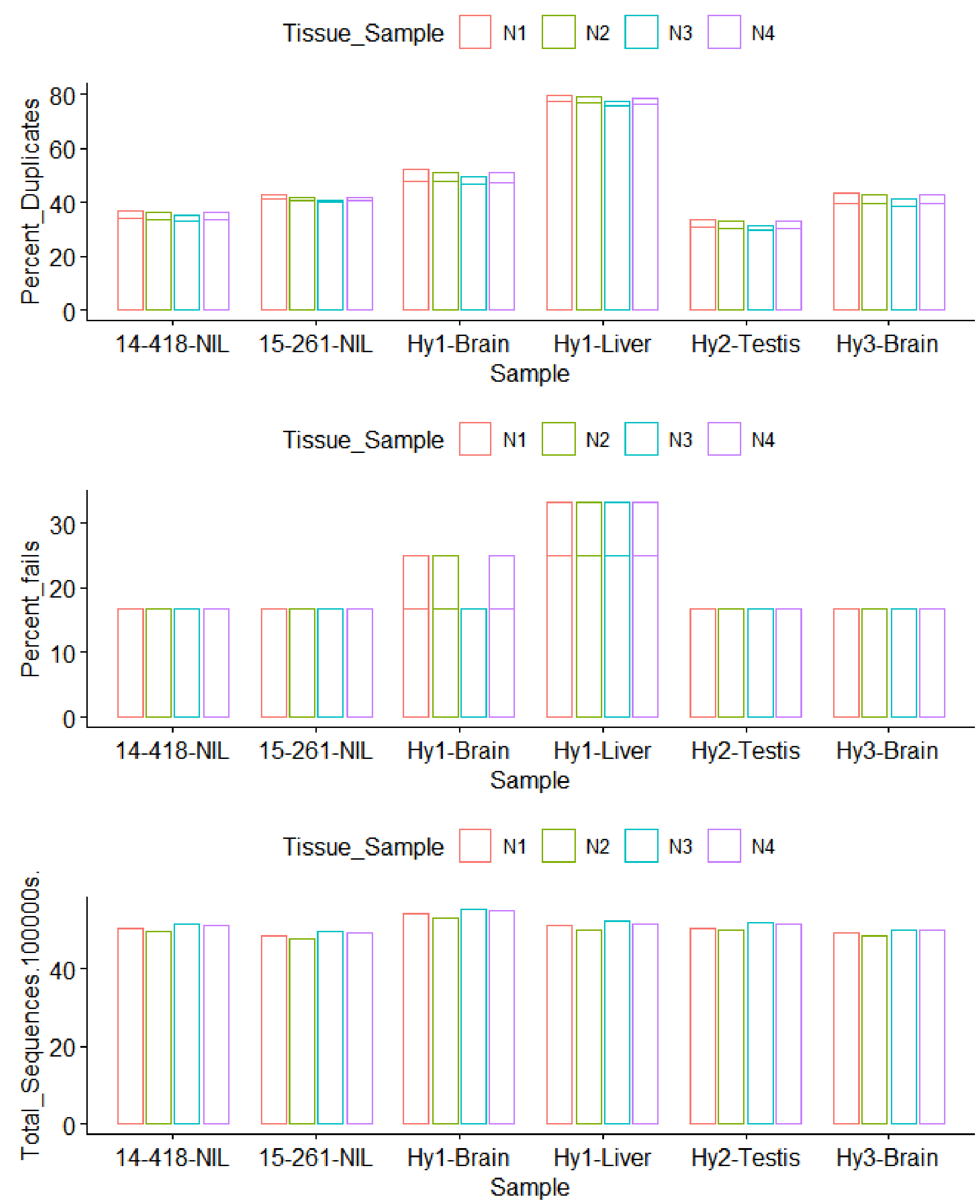

Figure 5.1. Comparative Results from the Fastq Analysis. The graph above represents the collected fastqc results of the 48 fastq files that were returned following sequencing. Results are grouped according the sample that they are associated with and all results for a particular statistic are represented within its respective bar. Each bar is representative of eight separate files which are grouped together, with reverse and forward values overlapping within the graphs. Percent Duplicates is a percentage value. Total sequences are set such that each single sequence is representative of 100000 sequences.

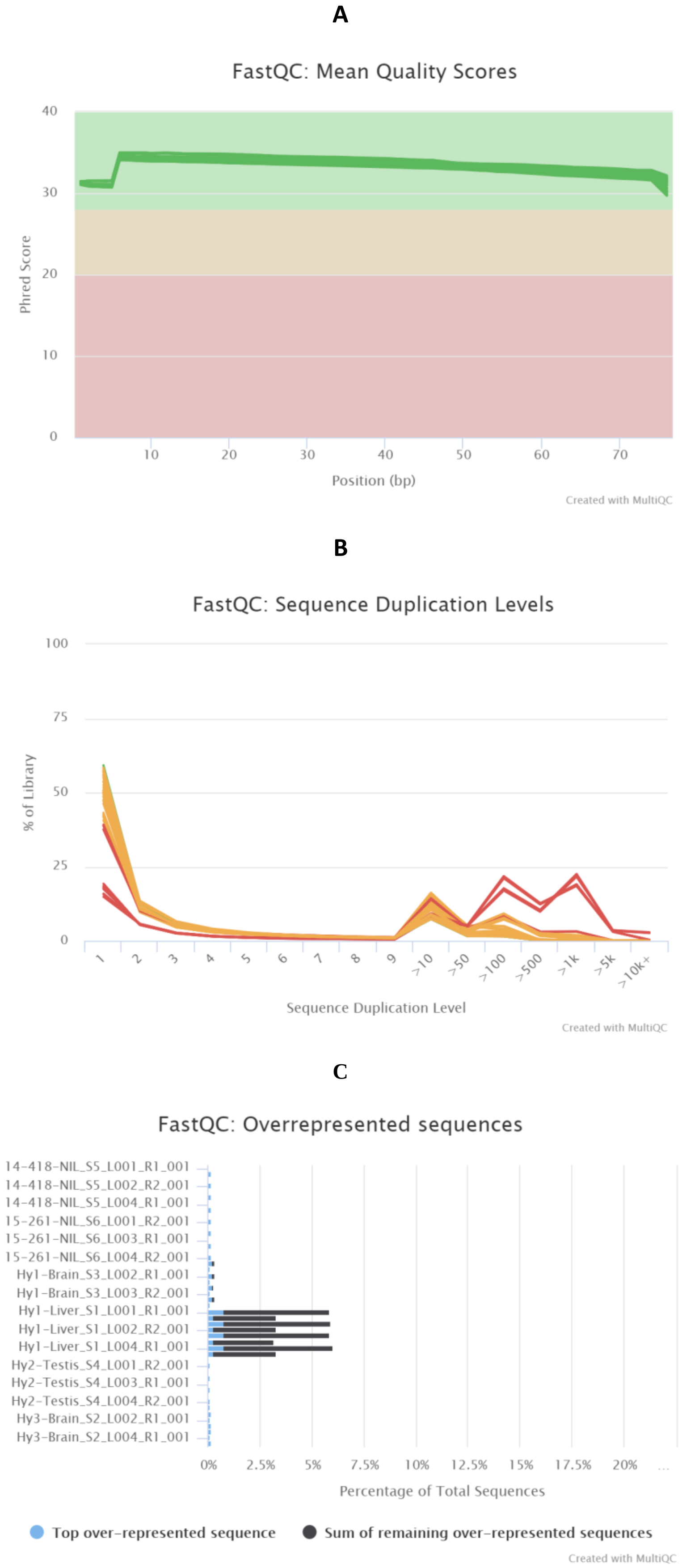

Figure 5.2. MultiQC Graph Metrics. The graphs above are representative of the sequence duplication levels, the overrepresented sequences, and the mean quality per sequence for the original fastq files.

Results in Figure 5.2 show sequence duplicates since this statistic can often be used to represent sequence coverage, low duplication percentage, or the introduction of bias during enrichment, which can be determined from a high level of duplication. The liver sample total sequences were included to determine if any one of the samples produced a disproportionately large number of sequences compared to the others. Adaptor content was not included in the above because the FastQC[128] analysis could not detect any adaptor content that the tool considered in great enough quantity to describe. FastQC[128] described the adapter content as less than 0.1% of the total sequences.

In Figure 5.2, the FastQC[128] analysis showed that each of the individual files were within parameters to be considered appropriate quality, as determined by the FastQC[128] manual. Liver tissue, however, displayed a severe drop in quality, compared to the other tissue samples analyzed, with a higher degree of sequence duplicates. Further examination of the FastQC files indicated that the liver samples had a large quantity of overrepresented sequences, which could indicate either highly biologically significant results, contamination, or a lack of diversity.

The BUSCO[84] and TransRate[85] analyses for the individual assemblies returned results comparable to those found within their respective literature. Based on this, they may be deemed successful in the case of each assembly. The exception to this the liver which returned a decreased value comparative to the remaining sequences.

Organism	Sample	Transcripts	N50
Hyena 1 – Young Adult Female	Brain	62884	1323
	Liver	32798	557
Hyena 2 – Young Adult Male	Testes	123062	1196
Hyena 3 – Young Adult Male	Brain	85792	1582
Hyena 4 – Young Adult Female	Unstimulated Blood	115332	1948
Hyena 5 – Adult Female	Unstimulated Blood	93374	1637

Table 5.1. Basic Metric Breakdown for the Initial Assemblies Produced from Individual Tissues. The above table is representative of the scores of the individual Trinity assemblies produced, in order to test the various separate tissue samples, and assess points of weakness in the assembly going forward.
Five organism were utilized in total for the RNA sequencing procedure; however six samples were extracted, with liver and one of the brain samples being taken from the same animal. All other samples were from separate organisms. N50 scores are the shortest contig that may be used to produce a sum of lengths equal to 50% of the total sequence. Longer scores are generally preferred.

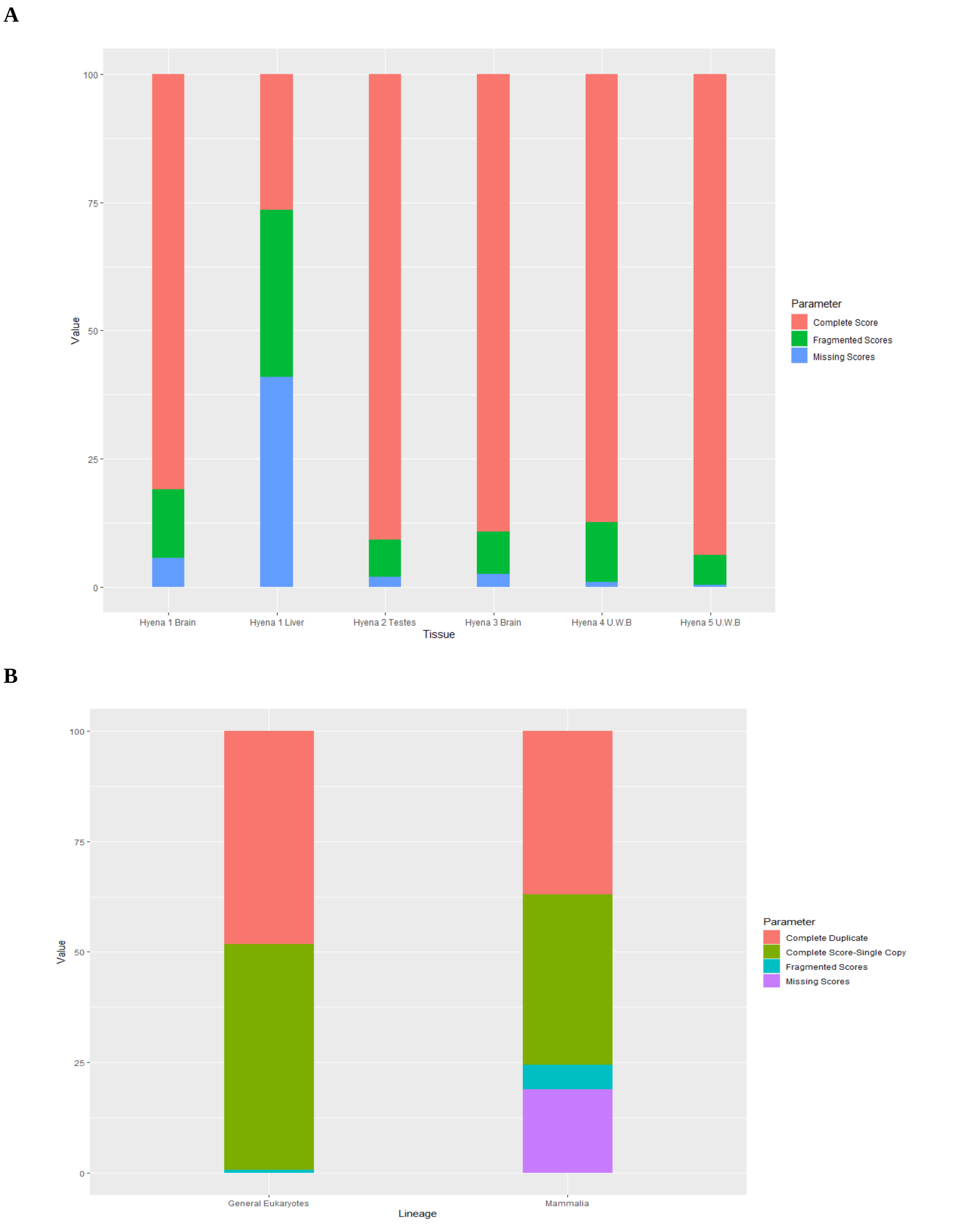

Figure 5.3. BUSCO Results for Individual and the Main Assembly. In the figures above, the graph in A.) represents a stacked barplot of BUSCO results for each of the individual assemblies. The values are in percentages and amount to total of 100. In graph B.) the BUSCO results for the main in-house assembly is represented, and separated via the lineage used to analyze the assembly. In A.) U.W.B. represents unstimulatred whole blood.

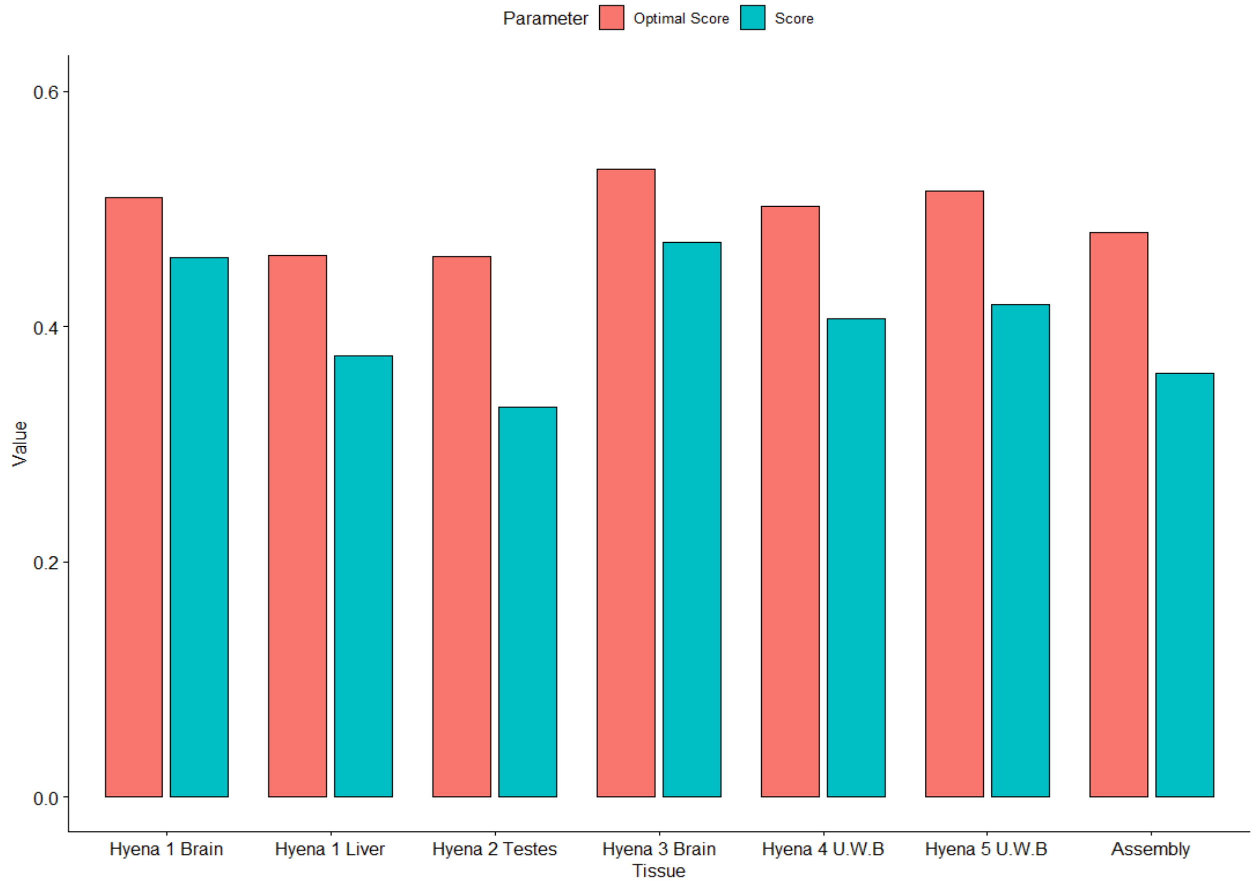

Figure 5.4. TransRate Results for the Individual and Main Assembly. In the graph above, TranRate scores are represented for the optimal and normal score. These values are a proportion between 0 and 1. Both graphs are grouped by the tissue which made up their assembly. The label "Assembly" is representative of the main assembly, while the individual assemblies are represented by the tissue they were produced from. U.W.B is representative of unstimulated whole blood.

As shown in Table 5.1, the majority of N50 values were above 1000, with the highest value being the unstimulated blood of Hyena 4, which had a value of 1948. The lowest value was seen in the liver tissue, which had a value of 557, which is less than half the second lowest value of the testes. The N50 score may be considered a measure of the assembly contiguity, or rather the ability of ability of the contigs to connect or overlap other contigs. The value of the liver sample suggests that it was not an ideal assembly, especially given the large difference between that value and the other closest value. The TransRate and BUSCO results are represented in Figure 5.3 and 5.4.

The BUSCO results, depicted in graph A, figure 5.3, are derived from the concept that certain orthologues should remain conserved among species, which determines the transcriptome completeness. Generally, the scores represented are above 80% complete orthologues, with the minimum percentage being the missing sequences. The highest of these values was for the testes sample, while the lowest was seen with the liver sample, where the greatest percentage was seen in the missing sequences, and the complete sequences being the lowest value. In graph B, figure 5.3, the

BUSCO results for the main assembly are represented using both a Mammalia lineage as well as the more general eukaryote lineage. From figure 5.3 it can be seen that the eukaryotic lineage displayed a majority of sequences as being complete, and only a small quantity were found to be fragmented, with none missing. However, a large proportion of the total consisted of duplicated sequences, but the majority, at 51.16 percent were found to be single copy. The Mammalia results presented a more diverse range, with a larger quantity of missing sequences to the eukaryotic lineage. However, the majority of the sequences were still found to be complete, with 38.6 percent being single copy and 37 percent being duplicate sequences, comparatively. As with the eukaryotic lineage, few of the sequences were found to be fragmented.

TransRate analysis provided proportionate scores of the assembly. TransRate uses its scores as a measure of the accuracy and correctness of the assembly, based on the reads used. The values, depicted in Figure 5.4, fall within a similar or higher value to that displayed in the article by TransRate[85], with the highest values seen within the Hyena 3 brain samples, and the lowest values displayed by hyena two in the testes.

After determining that each of the individual assemblies were of suitable quality, the main assembly was assembled by a similar method and examined via the same tools. The results for the analysis of the Trinity assembly returned a transcript count of 298954 and N50 score of 1745. The BUSCO and TransRate results are depicted in Figure 5.2. The BUSCO scores, depicted in Figure 5.3, B, depict that, within the eukaryote lineage, 99.34% of the sequences are present, while no missing sequences were found, and the remainder were fragmented. The results of the eukaryote lineage differs from the mammalian lineage BUSCO results, which depicts that 18.9% of the universal orthologues within this lineage are missing, however the majority of orthologues (75.6%) are present, with the remainder being fragmented sequences. Both lineages are not, however, comparable due to differing level of orthologues within both lineages. The TransRate analysis returned a proportion between 1 and 0 for both score results. Based on the results derived from the TransRate literature and other sources that used TransRate, the results were within expected values.

ProteinOrtho Alignment Analysis

In order to further validate the results from the BLAST run, the Trinity assembly was analyzed against four closely related species: cheetah (*Acinonyx jubatus*), domestic cat (*Felis catus*), leopard (*Panthera pardus*), and tiger (*Panthera tigris altaica*). This was accomplished via ProteinOrtho which determined orthologous sequences between the datasets provided by different species. ProteinOrtho[130] was utilised in an attempt to provide a direct comparison between the closely related species against the assembly.

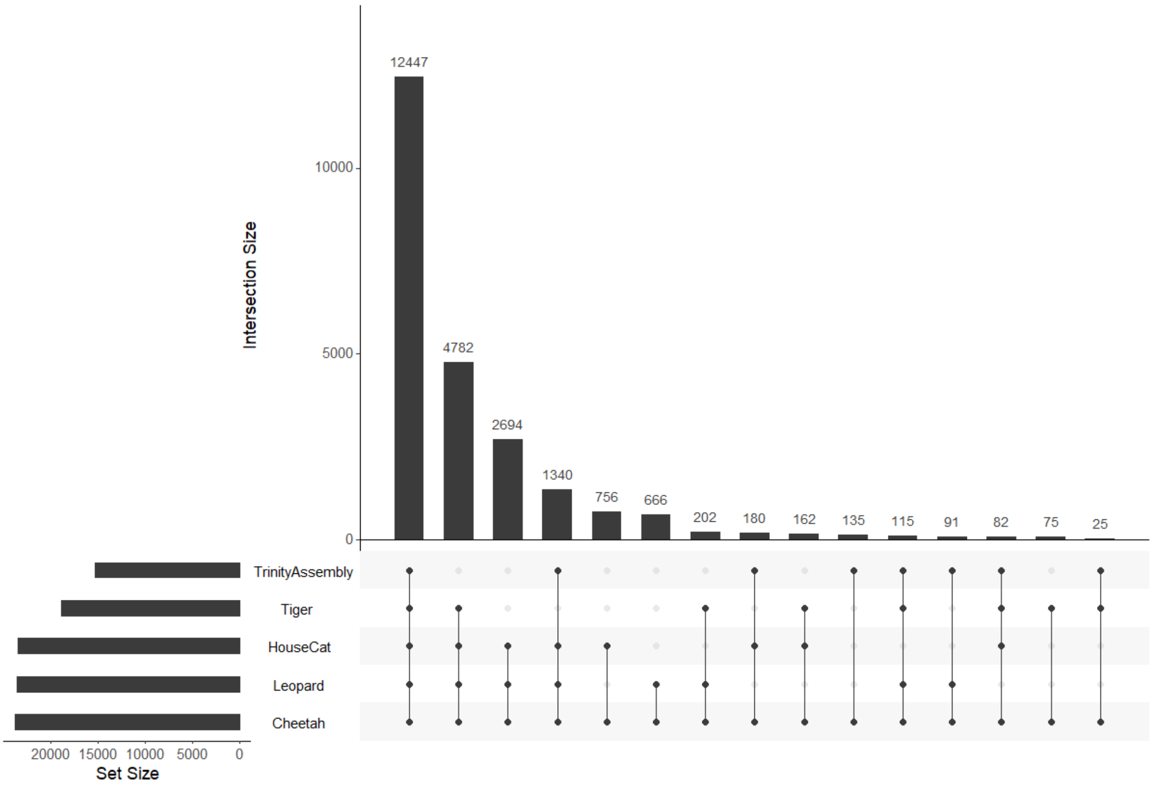

Figure 5.5. ProteinOrtho Results between the Main Assembly and Closely Related Species. The above graphs are representative of the ProteinOrtho analysis Trinity assembly. The graph is representative of the translated main assembly aligned against proteins from each of the most closely related species available, such as House Cat, Tiger, Leopard and Cheetah. The samples are further represented as overlapping values between different individual tissues and the main assembly. Total values of each column are represented above each of the bars. The results are set to only display the hits that overlapped with Trinity and are ordered in a descending order within that set.

In Figure 5.5, the total number of overlapping hits present from all samples was equal to 12447, while the total hits excluding tiger equated to 1505. The Trinity Assembly further had 293 orthologues detected within tiger sample alone, as opposed to the 133 non-shared hits within cat. The assembly itself produced a total of 127125 sequences, after being converted to protein sequences. This proved a total of 32599 hits in total across all the orthologous organisms. Approximately one quarter of the sequences matched with an orthologue sequence.

Overall, the highest quantity of hits can be attributed to the cheetah derived proteins, with 14372. However, the tiger hits were the lowest value, at 13131. This does not match the BLASTP results depicted in Figure 5.4, which displayed the house cat as having the most hits against the assembly, while the cheetah was found to have the lowest quantity at a low average quality per hit.

Mass Spectrometry Analysis

The MS-GF+ results for the semi-tryptic searches for the individual samples are shown in Table 5.2.

Semi-Tryptic Search of the Individual Samples Used for LC-MS/MS, Prior to Fractionation

Sample	Distinct Peptides	Filtered Spectra	Distinct Matches	Protein Groups
Abdominal Lymph1	4019	5239	4955	1192
Abdominal Lymph2	3937	4869	4674	1183
Abdominal Lymph3	5651	7003	6677	1467
Total Abdominal	7196	17111	9162	1609
Head Lymph 1	3311	4018	3848	1048
Head Lymph2	5026	6175	5914	1380
Total Head	5806	10193	6946	1455
Liver Sample 1	3536	4529	4282	934
Liver Sample 2	3277	4162	3966	886
Total Liver	4234	8691	5247	1025
Lung	3358	4061	3869	1078
Peripheral Lymph1	5452	6733	6488	1431
Peripheral Lymph2	4127	5040	4830	1259
Peripheral Lymph3	4863	5952	5723	1323
Total Peripheral	7414	17725	9043	1626
Pooled Head Tissue*	6658	8005	7724	1608
Pooled Liver Tissue*	5926	7609	7218	1329
Thoracic Lymph1	4611	5691	5436	1309
Thoracic Lymph2	2811	3360	3233	952
Total Thoracic	5103	9051	6122	1363
Total	15270	82446	19817	2067

Table 5.2 Peptide Search Results Metrics for the Individual Tissues Prior to Fractionation. This depicts the tryptic and semi-tryptic search results of the raw files against the NCBI *C. crocuta* draft genome, produced by Yang *et al*. The statistics represented are the counts of distinct peptides, spectra, distinct protein groups and distinct matches within each category of sample returned to us from the CPGR. The totals beneath them are representative of the total counts from those categories. Each of the samples were collected from one of three individuals of the *C. crocuta* species, with the sample number representing the individual the sample was collected from. Samples are grouped according to similarity/ whether similar samples were used in the analysis. The totals are representative of these similar groups, but were not, themselves used in the analysis. The groups with the asterisk next to them are representative of the pooled samples in this table.

The results in Table 5.2 indicate the distinct matches found within the analyzed tissue samples. From the above analysis, the category for the pooled head lymph node sample had the highest category for distinct matches and distinct peptides, if looking at a single sample, in both the tryptic and semi-tryptic search categories. Large changes were observed between tissue samples, which were all derived from separate individuals. The protein FDR of the semi-tryptic search results was 3.35%. The search found 1835 protein clusters, 2067 protein groups, and 2151 proteins.

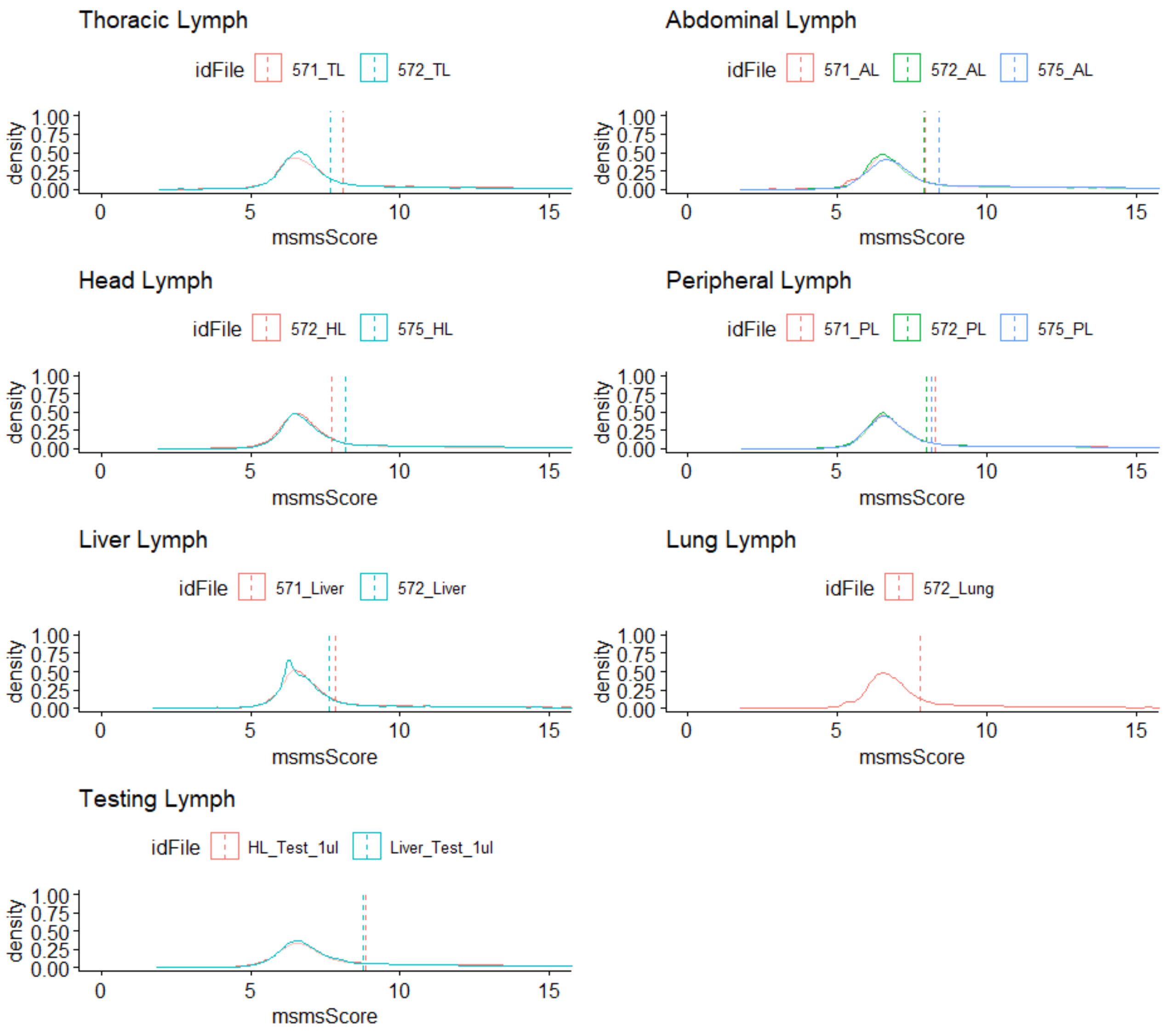

Figure 5.6. Analysis of the Original Lymph LC-MS/MS Results Prior to Fractionation. The graphs above depict the number of hits within particular "msmsScore" groups. "msmsScore" refers to the evalue converted to more readable values via a formula of *-log10*. Each of the graphs have had their count values normalized to 20000 counts to better compare them against each other. Each of the graphs have been groups according to the type of tissue the peptides were extracted from and were differentiated based on the different individuals the tissue was originally collected from. The mean for each sample is represented by the dotted line which bisects the graph.

From the results in Figure 5.6, the evalue per spectra collected from the initial LC-MS/MS analysis could be determined. The scores reflect the evalue per PSM multiplied by -log10. From the scores, the tissue from the pooled head lymph nodes seemed to have an overall lower quantity of spectra in terms of the quantity close to a value greater than or close to the mean value, although it still seemed to have a larger quantity when compared to the liver sample. The majority of the individual samples showed only small variation in quantities between them. This quantitative change could be due to the pooled samples being used for testing purposes to initially prepare for the LC-MS/MS analysis. However, the pooled samples also had a higher mean value than other samples, though this might be due to a lower quantity of PSMs which could affect the spread of values and influence the mean value.

A

Source	Distinct Peptides	Filtered Spectra	Distinct Matches	Protein Groups
Fraction1	2450	3004	2837	1289
Fraction2	3242	4098	3865	1565
Fraction3	4681	6332	5842	1879
Fraction4	2732	3498	3268	1395
Fraction5	2286	2766	2598	1229
Fraction6*	4360	5916	5503	1877
Fraction7	809	1762	922	546
Fraction8	640	1385	699	424
Fraction9	3245	4002	3777	1571
Fraction10*	158	183	173	135
Fraction11	17	24	22	23
Fraction12	25	28	28	30
Total	19880	32998	24369	2797

B

Source	Distinct Peptides	Filtered Spectra	Distinct Matches	Protein Groups
Fraction01	2465	3046	2823	1399
Fraction02	3304	4167	3928	1740
Fraction03	4814	6516	6027	2088
Fraction04	2716	3502	3265	1500
Fraction05	2371	2899	2710	1344
Fraction6*	4337	6034	5547	2059
Fraction07	908	1952	1041	643
Fraction08	394	737	438	299
Fraction09	3327	4126	3872	1717
Fraction10*	149	173	164	132
Fraction11	17	23	21	24
Fraction12	15	18	18	17
Total	20104	33175	24783	3132

Table 5.3. Results Metrics of the Peptide Search for Fractionated Tissues Used in LC-MS/MS. These data were searched by both the NCBI data produced Yang *et al* (Graph A) and the Trinity Assembly, produced as part of this project, following translation into peptide sequences (Graph B). The above table represents the semi-tryptic search results of the fractionated protein data. The data represented is the quantity of distinct peptides, as well filtered spectra, distinct matches and protein groups. This provides an initial indicator of the quality of each fraction. Fraction10 and Fraction6 produced initially unsatisfactory results during quality control, by the CPGR, and were re-analyzed using a higher concentration. The groups with an asterisk next to them are representative of the fractions that were run at a higher concentration.

The data represented in Table 5.3 provide an initial indication of how well each of the fractionated raw files performed during the analysis. There is a high degree of consistency between the results found in graph A and B. In both cases the highest quantities of distinct peptides were seen within fraction 3, while fraction 11 was the lowest value. In graph A, the search identified 2856 proteins, 2797 protein groups and 2499 cluster, with a protein FDR of 0.91%, and 19880 distinct peptides. While graph B identified 6249 proteins, 3146 protein groups, 2788 cluster and had an FDR of 0.22%, as well as 20104 distinct peptides. From the above data it can be seen that the Trinity assembly generated a lower FDR value and a larger number of proteins, when compared to the NCBI database. It can also be determined that, based on the table above, the Transcript based search results found more overall distinct peptides and identified more proteins than the results from the NCBI data. This would mean that, of the 127125 proteins predicted by TransDecoder, four percent of them were identifiable as proteins, when the assembly was used as the database.

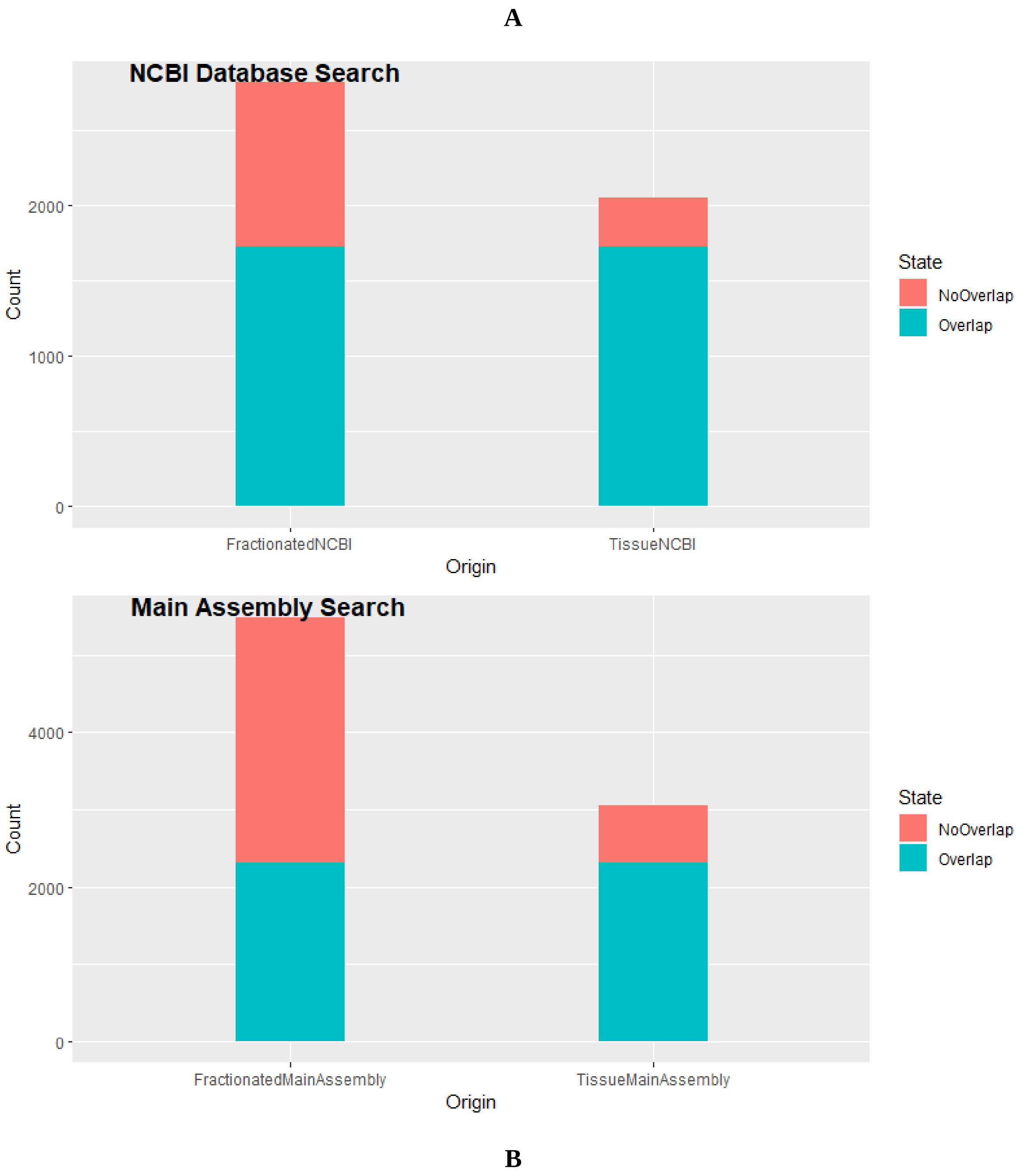

Figure 5.7 ProteinOrtho Results Between the Fractionated and Individual Tissue LC-MS/MS Results. The diagram above depicts overlapping proteins shared between the semi-tryptic "Fractionated" search results and the semi-tryptic results of the individual tissues, referred to as "Tissue", as determined by ProteinOrtho6. The stacked bar charts are separated to two groups, based on the database used to search them. The bar graph at the top, referred to as A, is representative of the results searched using the NCBI *C. crocuta* protein data, while the bottom bar graph, referred to as B, represents the results searched using the translated assembly produced as part of this project. The terms Overlap vs NoOverlap are representative of the number of peptides within the sample that found an overlapping result within the search database, while the NoOverlap represents peptides which could not find a match within the database. Overlapping regions are split between the tissue and fractionated categories, they are a shared quantity and remain constant within both graphs.

The results shown in Figure 5.7 represent the overlapping peptides shared between each of the fasta files collected from IDPicker for each of the separate searches run on the mzid files produced both pre and prior to the fractionation step of the peptides produced from the LC-MS/MS analysis. This step was expected to produce more diverse results post fractionation. The graph A, in figure 5.7, depicts the differences that could be seen between the fractionated sample and semi-tryptic search in a database searched using the NCBI *C. crocuta* protein data. Although both search results depict results that did not overlap, there was a large discrepancy between the non-overlapping results of the tissue category, when compared to the fractionated search results. However, the majority of hits did still fall within the overlap category. Furthermore, it appeared that the fractionated samples had a greater quantity of non-overlapping peptides compared to the tissue results.

The results in graph B, in Figure 5.7, represent the overlapping peptides between semi-tryptic search results of the fractionated and initial sample results. These were aligned with the *de novo* assembly and arranged as per the methods mentioned at the bottom of the proteomics component of the methods section. From the results, it appeared to reflect similar results to those seen in graph A in Figure 5.7 which suggest that the fractionation of the initial peptides was successful in increasing the peptide and spectra diversity The majority of the hits appeared to be non-overlapping peptides that fall within the fractionated section, although, there was still a large proportion of overlap between the two. This disparity of overlapping hits display a contrast between the fractionated and non-fractionated results, in the main assembly. These results indicate that the assembly can detect proteins which were identified using the *C. crocuta* draft genome.

Chapter 6

Discussion

In this project, we attempted to assemble the transcriptome of the hyena via the *de novo* Trinity assembly tool. The quality of these data was subsequently assessed via BUSCO[84], TransRate[85] and BLAST[132] analysis. Annotation was carried out using a BLAST[132] database consisting of closely related species, as well as a draft genome for *C. crocuta* from NCBI. Orthological detection was carried out using ProteinOrtho[130], using data for the closely related species, as well as the final Trinity[37] assembly. This was followed by LC-MS/MS of lymph node samples which provided further information and validity to the transcriptome assembly. The assembly produced was found to have a higher quality than the standard for metrics for the average transcriptome assembly, as determined by the TransRate score[85] and represented a high score for a general BUSCO eukaryote lineage. The inclusion of proteomics further validated the assembly, based against the NCBI assembly results. However, a lack of analysis in other areas, such as intensive mapping comparison using both the in-house Trinity RNA-Seq assembly and the NCBI genome assembly, weakens the final conclusion that the assembly is the best possible quality.

Hyena fastq quality assessment, Trimmomatic, and individual assembly

The analysis of each of the individual assemblies and the original fastq files was necessary to determine the quality of the assembly going forward, as well as whether pre-assembly adjustments would be necessary, such as trimming, or removing a sample from the main assembly based on its quality. This was accomplished via FastQC[128], as well as assembly and analysis of the individual tissue samples.

The analysis of the FastQC[128] results, from the fastq files, suggested that the quality of the raw files followed a standard, expected distribution, as per the FastQC[128] tutorial results for the per base quality. Per sequence quality scores also followed this trend. This was not, however, true for the sequence duplication levels, in which 11 of the fastq files were regarded as failures by the FastQC[128] tool. According to the FastQC[128] tool, any sequence which makes up more than one percent of the total content can be considered an error, or failed result. This implies the sequence is overrepresented. In this case, however, that did not necessarily represent a complete failure of the file, although it is common for highly abundant sequences to generate a failed result for this test. What was interesting, was that all the failed results were derived from samples collected from the same individual, including brain and liver samples. Overrepresented sequences are similar to the above in that they can exist within RNA data and still not constitute a failure in the sequence, as determined from the FastQC[128] tutorial instructions. Highly abundant sequences are more common in RNA than DNA, and it is not the complete failure it would have been had it been present within a DNA sequence, rather than an RNA sequence. The expression of highly repetitive elements can be tissue specific[138] and this may explain the observed results. The RNA-Seq data were based on cDNA generated from an RNA strand,

and the RNA-Seq process itself produces vast quantities of duplicate, but variable length sequences. Based on this information, it is possible that the expression pattern within a particular tissue emphasizes a specific product or pattern over most others which would lead to what could be considered overrepresented sequences. A further analysis of the individual assemblies produced from each of the samples, however, suggests that a problem may have been present. Although, whatever the cause for this lower quality or concern it might be better to exclude this failed sample from any further analyses in future, especially if they might lead to a drop in quality.

The brain and liver of individual 1 (17/571) both seemed to possess a smaller quantity of transcripts, compared to the other samples. This, coupled with the results from the analysis in figure 5.1, implied that both of the samples might have had a lower diversity of transcripts, compared to the other samples, although from the literature, this seems to be a greater concern when including proteomic analysis[139]. The diversity of transcripts pertains to and affects spectral searching. However, these results are also likely to be affected by how the samples were extracted and other factors such as storage or extraction difficulties, but it is challenging to determine without additional analysis and more samples.

Further analysis of the results via TransRate[85], which assesses quality of the assembly based on the initial reads, implied that although the sequences might have lacked diversity, the sequences themselves were still well assembled, based by the assembly tool. All of the individual assemblies seemed to have assembled well, as determined from results depicted in the TransRate study[85]. However, the testes sample from individual 2 (17/572) displayed the lowest results for the analysis. TransRate[85] fails a particular mapping when it cannot meet all of the four metrics: both the pairs (forwards and reverse) must align to each other, the aligned pair must be in the correct orientation, the aligned pair must be within the same contig, and the aligned pair must not overlap the ends of the contig. The low score of the testes sample suggests that a large proportion of the reads did not meet the above requirements. Further investigation must be carried out to determine which of the above four parameters have been breached in the greatest quantity. Furthermore, a more intensive comparative analysis, using the NCBI DNA assembly and the Trinity RNA-Seq assembly, could yield data which might be better used to infer the quality of the assembly.

It could be argued that the quantity of highly duplicated sequences seen within the individual samples of the brain and liver of individual 1 might not affect the results as severely as it might using a different analysis tool which tests diversity of sequences. A BUSCO[84] analysis of the individual assemblies provided more illumination as to the overall quality, as it tested the assembly completeness via the percentage of universal single copy orthologues that appeared within the individual assemblies. In this case it reflected the concerns regarding the liver and brain samples' diversity, as both had a lower number of complete sequences and a higher number of missing or fragmented sequences when compared to each of the other individual sample assemblies produced. The liver proved the most concerning in this regard given that a full 40% of the expected orthologues were missing from the assembly, with only 26% of them complete. In contrast, the brain sample assembly still possessed over 80% of the expected orthologues. This analysis may be explained by the results from the FastQC[128] analysis as liver had both a highly overrepresented collection of sequences as well as a high level of duplication. Despite the results of the individual assembly, the various original files were concatenated, and the main assembly was produced and assessed in a similar fashion.

Hyena assembly metrics and assembly process

The results from the initial analysis might have adversely affected the quality of the results based on the metrics associated with individual 1 (17/571). Thus, it might have been prudent to exclude that individual from the analysis. However, given that the TransRate[85] scores and TransRate[85] optimal score were 0.36 and 0.48, respectively, this suggests that it was not appropriate to exclude since the results fell below the average established by the individual assemblies. What this implies is that while resultant assembly might have been correct based on the original fastq files, some method in the assembly process or parameters might have been inefficient, possibly due to how the original files were selected, or at least the number of files that were excluded.

Alternatively, another reason for this relatively low quality might simply be that there was a high quantity of low scoring contigs, as can be seen from the 0.12 increase between the normal score and the optimal. The BUSCO analysis instead presented a high score of complete single copy orthologues with 99.4% of the 304 universal copy orthologue sequences complete within the assembly. There were no missing sequences found within the broader Eukaryotic lineages used. However, when the same analysis was performed using the specific Mammalian lineage, the results were more specific and provided a less general overview of the assembly completeness by using sequences which can only be found within Mammalian species. These data provided a larger quantity of both missing sequences and fragmented sequences and a lower number of complete sequences. What this implies is that the assembly was generally complete when compared against the orthologues found within eukaryotes in general, however in the complete mammalian comparison, this was not the case. There was a far larger quantity of missing as well as fragmented orthologues in this selection. The number of complete sequences were still far greater than the missing or fragmented sequences. The difference in the sequence quantities might be due to, both the more specific collection of sequences and the larger quantity collection of sequences found within the Mammalian lineage.

The assembly may be improved upon and either trimmed or added to with further data. It is likely that the limited samples have influenced this result, and further samples could alleviate this problem and improve upon this result. Despite these metrics, it is challenging to state whether the assembly was produced successfully, as there were few comparative points to other successful assemblies. There are more analyses that can be performed using the assembly. When compared to another assembly, such as one produced for the spiny mouse by Mamrot *et al*[140], the results were within expected quantities. This goes for both the BUSCO[84] analysis as well as the TransRate[85]. Furthermore, the authors of this study mentioned that the average for TransRate scores on the NCBI tended towards 0.2, which was in line with our results. Therefore, in this case, it could be argued that it was successfully and accurately assembled. However, it should still be aligned against different species, either to determine overlap, or reflect the quality of the hits which align to this assembly.

Hyena proteinOrtho alignment

The results of the ProteinOrtho[130] analysis found that the cheetah had the highest quantity of hits, although the largest quantity consisted of single, unmatched reads that did not align against any of the sequences. This may be due to a less studied proteome vs genome for each of the selected species, in which case a more diverse selection of proteins might prove more effective at removing the unaligned peptides. Alternatively, it might be due to the method by which the protein sequences for the assembly were acquired, or a further step could be taken to reduce the assembly transcripts based on a reduction of lower quality reads, or further clustering, bearing in mind that the assembly was already clustered before the alignment. A nucleotide alignment might also be more efficient at elucidating the unknown sequences since the nucleotide databases currently are often more extensive than available protein databases.

What was interesting about this analysis was that it revealed a selection of proteins which overlapped only between the *de* novo Trinity assembly and single tissue samples. However, these were often in such a small quantity that they might be negligible. These results also reflect well on the data since they produced high scoring hits against the aligned species, although the majority of hits still expressed fairly low -log10 (evalue) results. A nucleotide alignment, as stated above, might reflect differing results and could include greater variance between nucleotide sequences, as protein and amino acid sequences tend to retain their identity between different species and sequences compared to DNA and especially RNA. Further analysis can be determined via analysis of actual protein results.

Hyena LC-MS/MS (tissue and fractionated)

The initial examination of the MSGF+[137] results examined the difference between tryptic and semi—tryptic results from the search data. The analysis found consistent results between both the fractionated and initial individual tissue sample searches. Furthermore, based on the results, the pooled head lymph node sample was selected for further fractionation, due to the quantity of distinct peptides and spectra within the sample. During both of the above analyses it is important to supply as complete a protein sequence database as possible, as this directly impacts the quality of the database search results[94]. The better the quality and greater number of proteins within the database the more accurate the proteins within this database the lower the chance of a false positive discovery within the database search.

The examination of the LC-MS/MS results suggested that the fractionation was successful in increasing the detection of diverse peptides within the tissue, based on the comparison between the semi-tryptic results of fractionated and individual tissue search results. The fractionated proteins database search results, that was based on the Trinity assembly, were found to have a comparatively lower FDR and a higher quantity of proteins when compared to the results of the fractionated proteins searched against the NCBI data. This does imply that the assembly is accurate, but these results may also be connected to the fact that the same specimens that were used for the proteomic analysis as the RNA-seq data assembly. It does however, display that the assembly can be used to inform on the proteomic data. For the ProteinOrtho[130] comparison results the majority of the results overlapped between the two search results, a large portion of peptides did not overlap with the main assembly search, in figure 5.7, and the searched protein samples. The difference in results between the two searches was three times higher for the fractionated sample. This was partially reflected in the results from the ProteinOrtho[130] analysis of the semi-tryptic results for the proteins searched using the translated main assembly. This analysis was in contrast to the previous alignment in which the majority of search results were non-overlapping, despite both individual tissue and fractionated samples being determined using the in-house assembly.

What these results suggest is that the fractionation of the pooled head lymph node samples was successful in increasing peptide diversity. Finally, all the semi-tryptic results from both search databases were overlapped using ProteinOrtho[130]. This analysis showed a majority of hits overlapped between all four analyses, and only a small quantity of non-overlapping hits between each of the four search results. The exception was the results from the fractionated samples which were searched against the in-house assembly. This produced a large quantity of non-overlapping hits compared to the other analyses, and had a large proportion of overlapping hits with the results of the fractionated sample that were searched against the NCBI *C. crocuta* database.

The results of the proteome analysis suggested that the assembly was well assembled, or that it was similar to results in the NCBI database. This was due to the number of hits found within the search, which amounted to a greater amount compared to that of the NCBI data which was translated genomic data, although there is some concern as to the level of redundant hits within the assembly, and whether split reads were an issue with only partial alignment. Overall, it could be argued that the assembly was successfully validated by the proteome based on these data, though it is necessary to perform further tests to ensure that this assessment is accurate.

Limitations to the study identified areas where the study may be improved. The quantity of hyena samples used, and the limited tissue types restricted what could be examined via RNA-Seq. Additional samples could improve the output of the RNA-Seq analysis and subsequently improve the quality of the assembly. The assembly tool was also a limitation in the study, as it required testing against other assembly tools, and it was unknown what biases may have been introduced.

In a study by Ma et al[139], they report the use of Trinity as an assembly tool, and how the tool was not designed with proteomics in mind. According to this article, the use of Trinity attempts to minimize the presence of false positive hits, by reducing the transcript diversity, which affects spectral searching. During the project, a draft genome was released by another research group, which could have altered the initial study design had it been available at the start. Regardless, this information was incorporated retrospectively.

The analysis tools, BUSCO and TransRate, could have been expanded to include further analysis, such as through the Detonate tool. These tools are purely metric based assessments, and the quality may only be compared against currently available values used by other analyses. These metrics are limiting factors in that deviation outside of the norm provided by the tutorial or literature can produce difficult-to-interpret results. It was found that the lack of any reference sequence during a large portion of the study prevented certain assessments, and thus the project was forced to mainly rely on metrics, such as TransRate[85]. Metrics may have led to confusing methods of analysis which could have been affected by the software and available server resources. Time and processing resources limited what analysis could be accomplished as well as how much or to what extent, especially towards the end of the project where most analysis had to be performed without the server.

Ultimately, however, the assembly contributes to the field of transcriptomics. Further analysis is required for validation purposes, but the assembly was well assembled, and displayed good metric scores for each of the analyses. The proteomics results from the study, which, although searched using the assembly, requires further analysis and validation. However, the results from the assembly search aligned with the results from the NCBI database search. Given that these are fractions, a more accurate and clear view of the results could likely be presented via multiple alignments using the same parameters, but multiple fractions.

Although it was not possible with this project, with the current data on hyena sequences available, the resultant analyses could easily be expanded upon in any future projects. There would be similarity in the initial steps to this project, whereby the fastq file quality control would be the main point. An extensive trimming procedure and subsequent quality assessment step would ascertain whether the trimming step is successful in improving quality of the assembly from the start. This would ensure that whatever assembly produced would be produced using as high quality initial reads as can be produced, with as little interference from adapter content. In this step the tools Trimmomatic[66] and FastQC[128] are important in this step.

The benefit of having a genome available would have influenced the second step, necessitated the use of a different assembler, possibly using the reference-based assembly tool, which would leverage the genome to produce a better assembly. A tool such as RNA-eXpress[141] could be used. A further assessment, using a collection of assembly tools could be beneficial, as a single tool is probably insufficient to capture all the data. This would include *de novo* assembly tools, such as Trinity[37] and SOAPdenovo[80]. Testing an appropriate assembly reconciliation tool follows the previous step, which would allow the researcher to examine whether each of the assemblies in tandem provide better results than when used individually. In this case a tool such as TransBorrow[142] might be efficient, even though it is still guided by a genome. In this case, it might allow the reference-guided step to be skipped entirely.

Quality control of the assemblies, prior and post merge may be carried out using BUSCO[84] and TransRate[85], though with the inclusion of a tool such as DETONATE[86] to expand the analysis. In this case it might be best to perform a comparative step with the NCBI assembly or investigate tools more suited to this type of analysis, although TransRate[85] could still be used, as it maps the assembly against the original fastq files.

The final step is the proteogenomics component, which requires we begin with peptide searching, which can use MS-GF+. As before, the database can be the NCBI data. Although, another option could be *de novo* peptide sequencing. This method does not require a database to infer peptides. A tool such as ScanRanker[146] could be used to assess the best quality spectra prior to an analysis by PepNovo[143] which should perform the actual *de novo* peptide sequencing process. This can be compared against the searching using a database. ScanRanker would be used before hand to reduce the run time of the PepNovo tool and ensure only the best quality spectra are used. This could account for sequences that could not be determined using a database search or validate those that were already determined. The next step would be to compare how the quality of the assembly compares to the genome, which is where this project is probably weakest. One method would involve determining the level of overlap between the two protein sets (Assembly produced as part of the project and NCBI assembly), and the two assemblies themselves, when converted to proteins via a tool such as TransDecoder[131]. Following annotation it should also be possible to carry out direct comparison of the annotations, as was seen in an article by Zhu *et al*[144]. An alternative solution is to follow a more set workflow, such as the one mentioned by Sheynkman *et al*[145], where they made use of Galaxy-P[125] to make use of RNA-Seq data alongside MS/MS analysis to improve the ability to discover novel peptide sequences. It would require testing to determine how effective it would be with this dataset. Besides this, it is stated in the article that many of the tools seem to require reference and index data, which is more limited within this dataset.

Conclusion

The project attempted to produce a well annotated transcriptome assembly which could provide a foundation for future research into the biology of the spotted hyena and specifically, disease resistance. To an extent, this was achieved and is likely the most well assembled transcriptome assembly for the spotted hyena currently produced. The metrics suggest a high degree of sequence completeness and a high score when aligned against the original files. What the proteomics data search and alignment (via ProteinOrtho figure 5.5 and 5.7) suggest is that translated transcriptome data are viable for further analysis in proteogenomics analysis, sans an available genome for comparison. Furthermore, the availability of this assembly is beneficial as the spotted hyena is still a relatively understudied organism, even though a draft genome now exists for this species.

This study has provided transcriptome data, and some further protein data which contributes to knowledge of this species. The lymph node tissue were also successfully processed to produce protein data, which could prove beneficial to future research on hyena immune responses. More specifically, these data provide a resource on which others may build a more extensive analysis going forward. However, further refinement of these data is required, especially regarding expression levels and how they relate to other closely related species. It is less certain whether these data were actually benefitted from the use of solely transcriptomic data over solely genomic data, as the source of peptide searching methods. While the fractionated peptide samples did produce more overlapping hits from transcriptomic data, when compared to the genomic data, it should also be stated that the peptides were collected from the same individuals that the original RNA-Seq samples were collected from, which would likely influence the results.

In the future, analysis could be refined by employing multiple assemblers, and perhaps assembly reconciliation to account for the biases of different assembly tools. Furthermore, the use of reference-based assembly from the start would be beneficial, given the access to a draft genome, although the quality of this assembly would then be dependent upon the reference. Both of the above points are salient purely as a single datapoint does not make a dataset. This is especially relevant because, while Trinity has been shown to perform well in the past, this does not mean that another assembler may not perform better or result in hits that did not appear within the Trinity assembly, or perhaps provide a lesser degree of split reads, which were observed within the Trinity assembly. Further, exclusion of low quality individual tissue samples, such as the liver or brain tissue from individual 1 (17/571) might improve the assembly quality or redundancy. Alternatively, clustering the individual assemblies produced, such as via CD-HIT[133], prior to assembly reconciliation could offer further options for producing a more accurate assembly, especially if this is tested using different assembly tools.

Trimming tools are a further path to improving the assembly. While the FastQC[128] analysis suggested that no adaptor content was present, this does not mean it would not be beneficial to test this assumption. Producing and assessing several assemblies which had been trimmed would reflect on the quality of the final assembly produced. Importantly, it would allow an assessment of the accuracy of the FastQC[128] analysis to ensure that the results were accurate. It becomes especially important when assessing how the quality of the quality control metrics affect the end result of a translated or annotated transcriptome. Further analyses via BUSCO using a mammalian lineage would likely provide a different collection of results, compared to the more broad or general lineage we utilized for this analysis.

The current assembly lacks refinement overall. It requires further analysis in the form of Gene Ontology analysis, as well as further comparison against the NCBI assembly. Further analysis using the proteomic data, with a more in-depth examination of the overrepresented sequences, would improve any further examination of the gene sequences which the protein samples were based on, such as lymph node tissue, both within the current hyena data as well as the closely related orthologues.

This type of search may further be improved if access to more closely related species, such as the mongoose and meerkat become available. If this were to include more spotted hyena individuals, it would both increase the complexity and likely the diversity of the transcriptome that could be collected via RNA-Seq analysis. The BUSCO[147] results should also be taken into account, and, while the missing sequences cannot necessarily be completely corrected without further data, the fragmented sequences should be examined and determine whether they already exist within the current assembly but within multiple different sequences that, through error, did not form a complete sequence.

In summary, the current assembly can provide a basis for future research and may facilitate investigation of disease resistance and identifying more difficult to find sequences.

References

(1) Trinkel, M. Prey Selection and Prey Preferences of Spotted Hyenas Crocuta Crocuta in the Etosha National Park, Namibia. *Ecol. Res.* **2010**, *25* (2), 413–417.

(2) Rohland, N.; Pollack, J. L.; Nagel, D.; Beauval, C.; Airvaux, J.; Pääbo, S.; Hofreiter, M. The Population History of Extant and Extinct Hyenas. *Mol. Biol. Evol.* **2005**, *22* (12), 2435–2443.

(3) Flies, A. S.; Maksimoski, M. T.; Mansfield, L. S.; Weldele, M. L.; Holekamp, K. E. Characterization of Toll-like Receptors 1–10 in Spotted Hyenas. *Vet. Res. Commun.* **2014**, *38* (2), 165–170.

(4) Siembieda, J. L.; Kock, R. A.; McCracken, T. A.; Newman, S. H. The Role of Wildlife in Transboundary Animal Diseases. *Anim. Heal. Res. Rev.* **2011**, *12* (1), 95–111.

(5) Califf, K. J.; Ratzloff, E. K.; Wagner, A. P.; Holekamp, K. E.; Williams, B. L. Forces Shaping Major Histocompatibility Complex Evolution in Two Hyena Species. *J. Mammal.* **2013**, *94* (2), 282–294.

(6) Theis, K. R.; Venkataraman, A.; Dycus, J. A.; Koonter, K. D.; Schmitt-Matzen, E. N.; Wagner, A. P.; Holekamp, K. E.; Schmidt, T. M. Symbiotic Bacteria Appear to Mediate Hyena Social Odors. *Proc. Natl. Acad. Sci.* **2013**, *110* (49), 19832–19837.

(7) HOLEKAMP, K. E.; SMITH, J. E.; STRELIOFF, C. C.; VAN HORN, R. C.; WATTS, H. E. Society, Demography and Genetic Structure in the Spotted Hyena. *Mol. Ecol.* **2012**, *21* (3), 613–632.

(8) Yirga, G.; De Iongh, H. H.; Leirs, H.; Gebrihiwot, K.; Deckers, J.; Bauer, H. Adaptability of Large Carnivores to Changing Anthropogenic Food Sources: Diet Change of Spotted Hyena (Crocuta Crocuta) during Christian Fasting Period in Northern Ethiopia. *J. Anim. Ecol.* **2012**, *81* (5), 1052–1055.

(9) Yang, C.; Li, F.; Xiong, Z.; Koepfli, K. P.; Ryder, O.; Perelman, P.; Li, Q.; Zhang, G. A Draft Genome Assembly of Spotted Hyena, Crocuta Crocuta. *Sci. Data* **2020**, *7* (1), 1–10.

(10) Abeles, S. R.; Pride, D. T. Molecular Bases and Role of Viruses in the Human Microbiome. *J. Mol. Biol.* **2014**, *426* (23), 3892–3906.

(11) Shangguan, L.; Han, J.; Kayesh, E.; Sun, X.; Zhang, C.; Pervaiz, T.; Wen, X.; Fang, J. Evaluation of Genome Sequencing Quality in Selected Plant Species Using Expressed Sequence Tags. *PLoS One* **2013**, *8* (7), e69890.

(12) Hou, Y.-C. C.; Yu, H.-C.; Martin, R.; Cirulli, E. T.; Schenker-Ahmed, N. M.; Hicks, M.; Cohen, I. V.; Jönsson, T. J.; Heister, R.; Napier, L.; et al. Precision Medicine Integrating

Whole-Genome Sequencing, Comprehensive Metabolomics, and Advanced Imaging. *Proc. Natl. Acad. Sci.* **2020**, *117* (6), 3053–3062.

(13) Saltykova, A.; Wuyts, V.; Mattheus, W.; Bertrand, S.; Roosens, N. H. C.; Marchal, K.; De Keersmaecker, S. C. J. Comparison of SNP-Based Subtyping Workflows for Bacterial Isolates Using WGS Data, Applied to Salmonella Enterica Serotype Typhimurium and Serotype 1,4,[5],12:I:. *PLoS One* **2018**, *13* (2), e0192504.

(14) Lu, Y.-Q.; Lu, K.-H. Advancements in Next-Generation Sequencing for Diagnosis and Treatment of Non-Small-Cell Lung Cancer. *Chronic Dis. Transl. Med.* **2017**, *3* (1), 1–7.

(15) Sood, A.; Chauhan, R. S. Comparative NGS Transcriptomics Unravels Molecular Components Associated with Mosaic Virus Infection in a Bioenergy Plant Species, Jatropha Curcas L. *BioEnergy Res.* **2017**, *10* (1), 129–145.

(16) Zhang, S.; Sui, Z.; Chang, L.; Kang, K.; Ma, J.; Kong, F.; Zhou, W.; Wang, J.; Guo, L.; Geng, H.; et al. Transcriptome de Novo Assembly Sequencing and Analysis of the Toxic Dinoflagellate Alexandrium Catenella Using the Illumina Platform. *Gene* **2014**, *537* (2), 285–293.

(17) Hood, L.; Rowen, L. The Human Genome Project: Big Science Transforms Biology and Medicine. *Genome Med.* **2013**, *5* (9), 79.

(18) Collins, F.; Galas, D. A New Five-Year Plan for the U.S. Human Genome Project. *Science (80-.).* **1993**, *262* (5130), 43–46.

(19) Levy, S. E.; Myers, R. M. Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* **2016**, *17* (1), 95–115.

(20) Cao, Y.; Fanning, S.; Proos, S.; Jordan, K.; Srikumar, S. A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies. *Front. Microbiol.* **2017**, *8* (1829).

(21) Heather, J. M.; Chain, B. The Sequence of Sequencers: The History of Sequencing DNA. *Genomics* **2016**, *107* (1), 1–8.

(22) van Dijk, E. L.; Auger, H.; Jaszczyszyn, Y.; Thermes, C. Ten Years of Next-Generation Sequencing Technology. *Trends Genet.* **2014**, *30* (9), 418–426.

(23) Khan, A. R.; Pervez, M. T.; Babar, M. E.; Naveed, N.; Shoaib, M. A Comprehensive Study of De Novo Genome Assemblers: Current Challenges and Future Prospective. *Evol. Bioinforma.* **2018**, *14*, 117693431875865.

(24) Chhangawala, S.; Rudy, G.; Mason, C. E.; Rosenfeld, J. A. The Impact of Read Length on

Quantification of Differentially Expressed Genes and Splice Junction Detection. *Genome Biol.* **2015**, *16* (1), 131.

(25) Wommack, K. E.; Bhavsar, J.; Ravel, J. Metagenomics: Read Length Matters. *Appl. Environ. Microbiol.* **2008**, *74* (5), 1453–1463.

(26) Mantere, T.; Kersten, S.; Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. *Front. Genet.* **2019**, *10* (MAY), 1–14.

(27) Arsenic, R.; Treue, D.; Lehmann, A.; Hummel, M.; Dietel, M.; Denkert, C.; Budczies, J. Comparison of Targeted Next-Generation Sequencing and Sanger Sequencing for the Detection of PIK3CA Mutations in Breast Cancer. *BMC Clin. Pathol.* **2015**, *15* (1), 20.

(28) Cui, J.; Shen, N.; Lu, Z.; Xu, G.; Wang, Y.; Jin, B. Analysis and Comprehensive Comparison of PacBio and Nanopore-Based RNA Sequencing of the Arabidopsis Transcriptome. *Plant Methods* **2020**, *16* (1), 85.

(29) Alidjinou, E. K.; Deldalle, J.; Hallaert, C.; Robineau, O.; Ajana, F.; Choisy, P.; Hober, D.; Bocket, L. RNA and DNA Sanger Sequencing versus Next-Generation Sequencing for HIV-1 Drug Resistance Testing in Treatment-Naive Patients. *J. Antimicrob. Chemother.* **2017**, *72* (10), 2823–2830.

(30) Amarasinghe, S. L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M. E.; Gouil, Q. Opportunities and Challenges in Long-Read Sequencing Data Analysis. *Genome Biol.* **2020**, *21* (1), 1–16.

(31) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* **2009**, *10* (1), 57–63.

(32) Patterson, J.; Carpenter, E. J.; Zhu, Z.; An, D.; Liang, X.; Geng, C.; Drmanac, R.; Wong, G. K.-S. Impact of Sequencing Depth and Technology on de Novo RNA-Seq Assembly. *BMC Genomics* **2019**, *20* (1), 604.

(33) Costa, V.; Angelini, C.; De Feis, I.; Ciccodicola, A. Uncovering the Complexity of Transcriptomes with RNA-Seq. *J. Biomed. Biotechnol.* **2010**, *2010*, 1–19.

(34) Yahav, T.; Privman, E. A Comparative Analysis of Methods for de Novo Assembly of Hymenopteran Genomes Using Either Haploid or Diploid Samples. *Sci. Rep.* **2019**, *9* (1), 1–10.

(35) Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A. A.; Dvorkin, M.; Kulikov, A. S.; Lesin, V. M.; Nikolenko, S. I.; Pham, S.; Prjibelski, A. D.; et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **2012**, *19* (5), 455–477.

(36) Chevreux, B.; Wetter, T.; Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma. '99, GCB, Hann. Ger.* **1999**, 45–56.

(37) Grabherr, M. G. .; Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., N.; Friedman, and A. R. Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data. *Nat. Biotechnol.* **2013**, *29* (7), 644–652.

(38) Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler. *Gigascience* **2012**, *1* (1), 18.

(39) Liao, X.; Li, M.; Zou, Y.; Wu, F. X.; Yi-Pan; Wang, J. Current Challenges and Solutions of de Novo Assembly. *Quantitative Biology*. **2019**, *7* (2), 90–109.

(40) Simpson, J. T.; Wong, K.; Jackman, S. D.; Schein, J. E.; Jones, S. J. M.; Birol, I. ABySS: A Parallel Assembler for Short Read Sequence Data. *Genome Res.* **2009**, *19* (6), 1117–1123.

(41) Weisenfeld, N. I.; Yin, S.; Sharpe, T.; Lau, B.; Hegarty, R.; Holmes, L.; Sogoloff, B.; Tabbaa, D.; Williams, L.; Russ, C.; et al. Comprehensive Variation Discovery in Single Human Genomes. *Nat. Genet.* **2014**, *46* (12), 1350–1355.

(42) Kumar, S.; Blaxter, M. L. Comparing de Novo Assemblers for 454 Transcriptome Data. *BMC Genomics* **2010**, *11* (1), 571.

(43) Vijay, N.; Poelstra, J. W.; Künstner, A.; Wolf, J. B. W. Challenges and Strategies in Transcriptome Assembly and Differential Gene Expression Quantification. A Comprehensive in Silico Assessment of RNA-Seq Experiments. *Mol. Ecol.* **2013**, *22* (3), 620–634.

(44) Hölzer, M.; Marz, M. De Novo Transcriptome Assembly: A Comprehensive Cross-Species Comparison of Short-Read RNA-Seq Assemblers. *Gigascience* **2019**, *8* (5).

(45) Narzisi, G.; Mishra, B. Comparing De Novo Genome Assembly: The Long and Short of It. *PLoS One* **2011**, *6* (4).

(46) Marchant, A.; Mougel, F.; Mendonça, V.; Quartier, M.; Jacquin-Joly, E.; da Rosa, J. A.; Petit, E.; Harry, M. Comparing de Novo and Reference-Based Transcriptome Assembly Strategies by Applying Them to the Blood-Sucking Bug Rhodnius Prolixus. *Insect Biochem. Mol. Biol.* **2016**, *69*, 25–33.

(47) Cattonaro, F.; Policriti, A.; Vezzi, F. Enhanced Reference Guided Assembly. In *2010 IEEE*

International Conference on Bioinformatics and Biomedicine (BIBM); IEEE, **2010**; pp 77–80.

(48) Schneeberger, K.; Ossowski, S.; Ott, F.; Klein, J. D.; Wang, X.; Lanz, C.; Smith, L. M.; Cao, J.; Fitz, J.; Warthmann, N.; et al. Reference-Guided Assembly of Four Diverse Arabidopsis Thaliana Genomes. *Proc. Natl. Acad. Sci.* **2011**, *108* (25), 10249–10254.

(49) Lischer, H. E. L.; Shimizu, K. K. Reference-Guided de Novo Assembly Approach Improves Genome Reconstruction for Related Species. *BMC Bioinformatics* **2017**, *18* (1), 474.

(50) Alonge, M.; Soyk, S.; Ramakrishnan, S.; Wang, X.; Goodwin, S.; Sedlazeck, F. J.; Lippman, Z. B.; Schatz, M. C. RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes. *Genome Biol.* **2019**, *20* (1), 224.

(51) Li, Z.; Chen, Y.; Mu, D.; Yuan, J.; Shi, Y.; Zhang, H.; Gan, J.; Li, N.; Hu, X.; Liu, B.; et al. Comparison of the Two Major Classes of Assembly Algorithms: Overlap-Layout-Consensus and de-Bruijn-Graph. *Brief. Funct. Genomics* **2012**, *11* (1), 25–37.

(52) Boisvert, S.; Laviolette, F.; Corbeil, J. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. *J. Comput. Biol.* **2010**, *17* (11), 1519–1533.

(53) Zerbino, D. R.; Birney, E. Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs. *Genome Res.* **2008**, *18* (5), 821–829.

(54) Yang, X.; Charlebois, P.; Gnerre, S.; Coole, M. G.; Lennon, N. J.; Levin, J. Z.; Qu, J.; Ryan, E. M.; Zody, M. C.; Henn, M. R. De Novo Assembly of Highly Diverse Viral Populations. *BMC Genomics* **2012**, *13* (1), 475.

(55) Simpson, J. T.; Durbin, R. Efficient Construction of an Assembly String Graph Using the FM-Index. *Bioinformatics* **2010**, *26* (12), i367–i373.

(56) Gonnella, G.; Kurtz, S. Readjoiner: A Fast and Memory Efficient String Graph-Based Sequence Assembler. *BMC Bioinformatics* **2012**, *13* (1), 82.

(57) Miller, J. R.; Koren, S.; Sutton, G. Assembly Algorithms for Next-Generation Sequencing Data. *Genomics* **2010**, *95* (6), 315–327.

(58) Yoon, S.; Kim, D.; Kang, K.; Park, W. J. TraRECo: A Greedy Approach Based de Novo Transcriptome Assembler with Read Error Correction Using Consensus Matrix. *BMC Genomics* **2018**, *19* (1), 653.

(59) Zhu, X.; Leung, H. C. M.; Chin, F. Y. L.; Yiu, S. M.; Quan, G.; Liu, B.; Wang, Y. PERGA: A Paired-End Read Guided de Novo Assembler for Extending Contigs Using SVM and Look Ahead Approach. *PLoS One* **2014**, *9* (12), 1–27.

(60) Alhakami, H.; Mirebrahim, H.; Lonardi, S. A Comparative Evaluation of Genome Assembly Reconciliation Tools. *Genome Biol.* **2017**, *18* (1), 93.

(61) Nagarajan, N.; Pop, M. Sequence Assembly Demystified. *Nature Reviews Genetics*. **2013**, *14* (3), 157–167.

(62) Silva, G. G. Z.; Dutilh, B. E.; Matthews, T.; Elkins, K.; Schmieder, R.; Dinsdale, E. A.; Edwards, R. A. Combining de Novo and Reference-Guided Assembly with Scaffold_builder. *Source Code Biol. Med.* **2013**, *8* (1), 23.

(63) Ruby, J. G.; Bellare, P.; DeRisi, J. L. PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3 Genes, Genomes, Genet.* **2013**, *3* (5), 865–880.

(64) Marchant, A.; Mougel, F.; Mendonça, V.; Quartier, M.; Jacquin-Joly, E.; da Rosa, J. A.; Petit, E.; Harry, M. Comparing de Novo and Reference-Based Transcriptome Assembly Strategies by Applying Them to the Blood-Sucking Bug Rhodnius Prolixus. *Insect Biochem. Mol. Biol.* **2016**, *69*, 25–33.

(65) Moreno-Santillán, D. D.; Machain-Williams, C.; Hernández-Montes, G.; Ortega, J. De Novo Transcriptome Assembly and Functional Annotation in Five Species of Bats. *Sci. Rep.* **2019**, *9* (1), 6222.

(66) Williams, C. R.; Baccarella, A.; Parrish, J. Z.; Kim, C. C. Trimming of Sequence Reads Alters RNA-Seq Gene Expression Estimates. *BMC Bioinformatics* **2016**, *17* (1), 103.

(67) Ungaro, A.; Pech, N.; Martin, J.-F.; McCairns, R. J. S.; Mévy, J.-P.; Chappaz, R.; Gilles, A. Challenges and Advances for Transcriptome Assembly in Non-Model Species. *PLoS One* **2017**, *12* (9), e0185020.

(68) Haak, M.; Vinke, S.; Keller, W.; Droste, J.; Rückert, C.; Kalinowski, J.; Pucker, B. High Quality de Novo Transcriptome Assembly of Croton Tiglium. *Front. Mol. Biosci.* **2018**, *5* (62).

(69) Robertson, G.; Schein, J.; Chiu, R.; Corbett, R.; Field, M.; Jackman, S. D.; Mungall, K.; Lee, S.; Okada, H. M.; Qian, J. Q.; et al. De Novo Assembly and Analysis of RNA-Seq Data. *Nat. Methods* **2010**, *7* (11), 909–912.

(70) Denisov, G.; Walenz, B.; Halpern, A. L.; Miller, J.; Axelrod, N.; Levy, S.; Sutton, G. Consensus Generation and Variant Detection by Celera Assembler. *Bioinformatics* **2008**, *24* (8), 1035–1040.

(71) Batzoglou, S. ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Res.* **2002**, *12* (1),

177–189.

(72) Kieser, S.; Brown, J.; Zdobnov, E. M.; Trajkovski, M.; McCue, L. A. ATLAS: A Snakemake Workflow for Assembly, Annotation, and Genomic Binning of Metagenome Sequence Data. *BMC Bioinformatics* **2019**, *21* (257).

(73) Zhang, W.; Chen, J.; Yang, Y.; Tang, Y.; Shang, J.; Shen, B. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLoS One* **2011**, *6* (3), e17915.

(74) Cherukuri, Y.; Janga, S. C. Benchmarking of de Novo Assembly Algorithms for Nanopore Data Reveals Optimal Performance of OLC Approaches. *BMC Genomics* **2016**, *17* (S7), 507.

(75) Branton, D.; Deamer, D. W.; Marziali, A.; Bayley, H.; Benner, S. A.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X.; et al. The Potential and Challenges of Nanopore Sequencing. *Nat. Biotechnol.* **2008**, *26* (10), 1146–1153.

(76) Kleiman, M.; Tannenbaum, E. Diploidy and the Selective Advantage for Sexual Reproduction in Unicellular Organisms. *Theory Biosci.* **2009**, *128* (4), 249–285.

(77) Liu, B.; Liu, C.-M.; Li, D.; Li, Y.; Ting, H.-F.; Yiu, S.-M.; Luo, R.; Lam, T.-W. BASE: A Practical de Novo Assembler for Large Genomes Using Long NGS Reads. *BMC Genomics* **2016**, *17* (S5), 499.

(78) Warren, R. L.; Sutton, G. G.; Jones, S. J. M.; Holt, R. A. Assembling Millions of Short DNA Sequences Using SSAKE. *Bioinformatics* **2007**, *23* (4), 500–501.

(79) Carruthers, M.; Yurchenko, A. A.; Augley, J. J.; Adams, C. E.; Herzyk, P.; Elmer, K. R. De Novo Transcriptome Assembly, Annotation and Comparison of Four Ecological and Evolutionary Model Salmonid Fish Species. *BMC Genomics* **2018**, *19* (1), 32.

(80) Xie, Y.; Wu, G.; Tang, J.; Luo, R.; Patterson, J.; Liu, S.; Huang, W.; He, G.; Gu, S.; Li, S.; et al. SOAPdenovo-Trans: De Novo Transcriptome Assembly with Short RNA-Seq Reads. *Bioinformatics* **2014**, *30* (12), 1660–1666.

(81) Yang, Y.; Gribskov, M. The Evaluation of RNA-Seq de Novo Assembly by PacBio Long Read Sequencing. *bioRxiv* **2019**.

(82) Huang, X.; Chen, X.-G.; Armbruster, P. A. Comparative Performance of Transcriptome Assembly Methods for Non-Model Organisms. *BMC Genomics* **2016**, *17* (1), 523.

(83) Mundry, M.; Bornberg-Bauer, E.; Sammeth, M.; Feulner, P. G. D. Evaluating Characteristics of De Novo Assembly Software on 454 Transcriptome Data: A Simulation Approach. *PLoS One* **2012**, *7* (2), e31410.

(84) Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* **2015**, *31* (19), 3210–3212.

(85) Smith-Unna, R.; Boursnell, C.; Patro, R.; Hibberd, J. M.; Kelly, S. TransRate: Reference-Free Quality Assessment of de Novo Transcriptome Assemblies. *Genome Res.* **2016**, *26* (8), 1134–1144.

(86) Li, B.; Fillmore, N.; Bai, Y.; Collins, M.; Thomson, J. A.; Stewart, R.; Dewey, C. N. Evaluation of de Novo Transcriptome Assemblies from RNA-Seq Data. *Genome Biol.* **2014**, *15* (12), 553.

(87) Rana, S. B.; Zadlock, F. J.; Zhang, Z.; Murphy, W. R.; Bentivegna, C. S. Comparison of De Novo Transcriptome Assemblers and K-Mer Strategies Using the Killifish, Fundulus Heteroclitus. *PLoS One* **2016**, *11* (4), e0153104.

(88) Chang, Z.; Li, G.; Liu, J.; Zhang, Y.; Ashby, C.; Liu, D.; Cramer, C. L.; Huang, X. Bridger: A New Framework for de Novo Transcriptome Assembly Using RNA-Seq Data. *Genome Biol.* **2015**, *16* (1), 1–10.

(89) Aganezov, S. S.; Alekseyev, M. A. CAMSA: A Tool for Comparative Analysis and Merging of Scaffold Assemblies. *BMC Bioinformatics* **2017**, *18* (S15), 496.

(90) Ruggles, K. V; Tang, Z.; Wang, X.; Grover, H.; Askenazi, M.; Teubl, J.; Cao, S.; McLellan, M. D.; Clauser, K. R.; Tabb, D. L.; et al. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol. Cell. Proteomics* **2016**, *15* (3), 1060–1071.

(91) Zimin, A. V.; Smith, D. R.; Sutton, G.; Yorke, J. A. Assembly Reconciliation. *Bioinformatics* **2008**, *24* (1), 42–45.

(92) Patterson, S. D.; Aebersold, R. H. Proteomics: The First Decade and Beyond. *Nat. Genet.* **2003**, *33* (S3), 311–323.

(93) Wang, M.; Wang, J.; Carver, J.; Pullman, B. S.; Cha, S. W.; Bandeira, N. Assembling the Community-Scale Discoverable Human Proteome. *Cell Syst.* **2018**, *7* (4), 412-421.e5.

(94) Graves, P. R.; Haystead, T. A. J. Molecular Biologist's Guide to Proteomics. *Microbiol. Mol. Biol. Rev.* **2002**, *66* (1), 39–63.

(95) Graumann, J.; Dunipace, L. A.; Seol, J. H.; McDonald, W. H.; Yates, J. R.; Wold, B. J.; Deshaies, R. J. Applicability of Tandem Affinity Purification MudPIT to Pathway Proteomics in Yeast. *Mol. Cell. Proteomics* **2004**, *3* (3), 226–237.

(96) Mittal, R. D. Tandem Mass Spectroscopy in Diagnosis and Clinical Research. *Indian J. Clin. Biochem.* **2015**, *30* (2), 121–123.

(97) Glish, G. L.; Burinsky, D. J. Hybrid Mass Spectrometers for Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (2), 161–172.

(98) Jedrychowski, M. P.; Huttlin, E. L.; Haas, W.; Sowa, M. E.; Rad, R.; Gygi, S. P. Evaluation of HCD- and CID-Type Fragmentation Within Their Respective Detection Platforms For Murine Phosphoproteomics. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.009910.

(99) Choi, B. K.; Hercules, D. M.; Zhang, T.; Gusev, A. I. Comparison of Quadrupole, Time-of-Flight, and Fourier Transform Mass Analyzers for LC-MS Applications. *LCGC North Am.* **2001**, *19* (5), 514–524.

(100) Fitzgerald, R. L.; O'Neal, C. L.; Hart, B. J.; Poklis, A.; Herold, D. A. Comparison of an Ion-Trap and a Quadrupole Mass Spectrometer Using Diazepam as a Model Compound. *J. Anal. Toxicol.* **1997**, *21* (6), 445–450.

(101) Soler, C.; Mañes, J.; Picó, Y. Comparison of Liquid Chromatography Using Triple Quadrupole and Quadrupole Ion Trap Mass Analyzers to Determine Pesticide Residues in Oranges. *J. Chromatogr. A* **2005**, *1067* (1–2), 115–125.

(102) Vázquez Peláez, M.; Costa-Fernández, J. M.; Sanz-Medel, A. Critical Comparison between Quadrupole and Time-of-Flight Inductively Coupled Plasma Mass Spectrometers for Isotope Ratio Measurements in Elemental Speciation. *J. Anal. At. Spectrom.* **2002**, *17* (8), 950–957.

(103) Shi, S. D. H.; Drader, J. J.; Freitas, M. A.; Hendrickson, C. L.; Marshall, A. G. Comparison and Interconversion of the Two Most Common Frequency-to-Mass Calibration Functions for Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Int. J. Mass Spectrom.* **2000**, *195–196*, 591–598.

(104) Makarov, A.; Denisov, E. Dynamics of Ions of Intact Proteins in the Orbitrap Mass Analyzer. *J. Am. Soc. Mass Spectrom.* **2009**, *20* (8), 1486–1495.

(105) Makarov, A.; Denisov, E.; Lange, O.; Horning, S. Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer. *J. Am. Soc. Mass Spectrom.* **2006**, *17* (7), 977–982.

(106) Michalski, A.; Damoc, E.; Hauschild, J.-P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass Spectrometry-Based Proteomics Using Q Exactive, a High-Performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* **2011**, *10* (9), M111.011015.

(107) Zhang, Y. V.; Wei, B.; Zhu, Y.; Zhang, Y.; Bluth, M. H. Liquid Chromatography–Tandem

Mass Spectrometry: An Emerging Technology in the Toxicology Laboratory. *Clin. Lab. Med.* **2016**, *36* (4), 635–661.

(108) Chelius, D.; Bondarenko, P. V. Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *J. Proteome Res.* **2002**, *1* (4), 317–323.

(109) Millioni, R.; Franchin, C.; Pivato, M.; Tessari, P.; Arrigoni, G. Sample Loading Influences Studies Comparing Isoelectric Focusing vs. Strong Cation Exchange Peptide Fractionation. *Journal of Chromatography A*. **2013**, *1307*, 207–208.

(110) Solovyeva, E. M.; Lobas, A. A.; Kopylov, A. T.; Ilina, I. Y.; Levitsky, L. I.; Moshkovskii, S. A.; Gorshkov, M. V. FractionOptimizer: A Method for Optimal Peptide Fractionation in Bottom-up Proteomics. *Anal. Bioanal. Chem.* **2018**, *410* (16), 3827–3833.

(111) Mostovenko, E.; Hassan, C.; Rattke, J.; Deelder, A. M.; van Veelen, P. A.; Palmblad, M. Comparison of Peptide and Protein Fractionation Methods in Proteomics. *EuPA Open Proteomics* **2013**, *1*, 30–37.

(112) Manadas, B.; Mendes, V. M.; English, J.; Dunn, M. J. Peptide Fractionation in Proteomics Approaches. *Expert Rev. Proteomics* **2010**, *7* (5), 655–663.

(113) Chiu, C. W.; Chang, C. L.; Chen, S. F. Evaluation of Peptide Fractionation Strategies Used in Proteome Analysis. *J. Sep. Sci.* **2012**, *35* (23), 3293–3301.

(114) Yang, F.; Shen, Y.; Camp, D. G.; Smith, R. D. High-PH Reversed-Phase Chromatography with Fraction Concatenation for 2D Proteomic Analysis. *Expert Rev. Proteomics* **2012**, *9* (2), 129–134.

(115) Batth, T. S.; Francavilla, C.; Olsen, J. V. Off-Line High-PH Reversed-Phase Fractionation for In-Depth Phosphoproteomics. *J. Proteome Res.* **2014**, *13* (12), 6176–6186.

(116) Guillot, L.; Delage, L.; Viari, A.; Vandenbrouck, Y.; Com, E.; Ritter, A.; Lavigne, R.; Marie, D.; Peterlongo, P.; Potin, P.; et al. Peptimapper: Proteogenomics Workflow for the Expert Annotation of Eukaryotic Genomes. *BMC Genomics* **2019**, *20* (1), 56.

(117) Castellana, N. E.; Payne, S. H.; Shen, Z.; Stanke, M.; Bafna, V.; Briggs, S. P. Discovery and Revision of Arabidopsis Genes by Proteogenomics. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (52), 21034–21038.

(118) Baerenfaller, K.; Grossmann, J.; Grobei, M. A.; Hull, R.; Hirsch-Hoffmann, M.; Yalovsky, S.; Zimmermann, P.; Grossniklaus, U.; Gruissem, W.; Baginsky, S. Genome-Scale Proteomics Reveals Arabidopsis Thaliana Gene Models and Proteome Dynamics. *Science (80-.).* **2008**, *320* (5878), 938–941.

(119) Nesvizhskii, A. I. Proteogenomics: Concepts, Applications and Computational Strategies. *Nat. Methods* **2014**, *11* (11), 1114–1125.

(120) Ang, M. Y.; Low, T. Y.; Lee, P. Y.; Wan Mohamad Nazarie, W. F.; Guryev, V.; Jamal, R. Proteogenomics: From next-Generation Sequencing (NGS) and Mass Spectrometry-Based Proteomics to Precision Medicine. *Clin. Chim. Acta* **2019**, *498*, 38–46.

(121) Ruggles, K. V.; Tang, Z.; Wang, X.; Grover, H.; Askenazi, M.; Teubl, J.; Cao, S.; McLellan, M. D.; Clauser, K. R.; Tabb, D. L.; et al. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol. Cell. Proteomics* **2016**, *15* (3), 1060–1071.

(122) Mertins, P.; Mani, D. R.; Ruggles, K. V.; Gillette, M. A.; Clauser, K. R.; Wang, P.; Wang, X.; Qiao, J. W.; Cao, S.; Petralia, F.; et al. Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer. *Nature* **2016**, *534* (7605), 55–62.

(123) Lazar, I. M.; Karcini, A.; Ahuja, S.; Estrada-Palma, C. Proteogenomic Analysis of Protein Sequence Alterations in Breast Cancer Cells. *Sci. Rep.* **2019**, *9* (1), 1–13.

(124) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. *J. Proteome Res.* **2021**, *20* (4), 1826–1834.

(125) Afgan, E.; Baker, D.; Batut, B.; van den Beek, M.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B. A.; et al. The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Res.* **2018**, *46* (W1), W537–W544.

(126) Fan, J.; Saha, S.; Barker, G.; Heesom, K. J.; Ghali, F.; Jones, A. R.; Matthews, D. A.; Bessant, C. Galaxy Integrated Omics: Web-Based Standards-Compliant Workflows for Proteomics Informed by Transcriptomics*. *Mol. Cell. Proteomics* **2015**, *14* (11), 3087–3093.

(127) Higgitt, R. L.; Buss, P. E.; van Helden, P. D.; Miller, M. A.; Parsons, S. D. Development of Gene Expression Assays Measuring Immune Responses in the Spotted Hyena (Crocuta Crocuta). *African Zool.* **2017**, *52* (2), 99–104.

(128) Andrews, S. FASTQC A Quality Control Tool for High Throughput Sequence Data [Online]. *Babraham Inst.* Available at: Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ **2015**, https://qubeshub.org/resources/fastqc.

(129) Turner, F. S. Assessment of Insert Sizes and Adapter Content in Fastq Data from NexteraXT

Libraries. *Front. Genet.* **2014**, *5* (5).

(130) Lechner, M.; Findeiß, S.; Steiner, L.; Marz, M.; Stadler, P. F.; Prohaska, S. J. Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis. *BMC Bioinformatics* **2011**, *12* (1), 124.

(131) Haas, B.; Papanicolaou, A. TransDecoder. Available at: https://github.com/TransDecoder/TransDecoder/wiki.

(132) McGinnis, S.; Madden, T. L. BLAST: At the Core of a Powerful and Diverse Set of Sequence Analysis Tools. *Nucleic Acids Res.* **2004**, *32* (Web Server), W20–W25.

(133) Li, W.; Godzik, A. Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* **2006**, *22* (13), 1658–1659.

(134) Conway, J. R.; Lex, A.; Gehlenborg, N. UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties. *Bioinformatics* **2017**, *33* (18), 2938–2940.

(135) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920.

(136) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, *24* (21), 2534–2536.

(137) Kim, S.; Pevzner, P. A. MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. *Nat. Commun.* **2014**, *5* (1), 5277.

(138) Billingsley, K. J.; Lättekivi, F.; Planken, A.; Reimann, E.; Kurvits, L.; Kadastik-Eerme, L.; Kasterpalu, K. M.; Bubb, V. J.; Quinn, J. P.; Kõks, S.; et al. Analysis of Repetitive Element Expression in the Blood and Skin of Patients with Parkinson's Disease Identifies Differential Expression of Satellite Elements. *Sci. Rep.* **2019**, *9* (1), 4369.

(139) Ma, J.; Saghatelian, A.; Shokhirev, M. N. The Influence of Transcript Assembly on the Proteogenomics Discovery of Microproteins. *PLoS One* **2018**, *13* (3), e0194518.

(140) Mamrot, J.; Legaie, R.; Ellery, S. J.; Wilson, T.; Seemann, T.; Powell, D. R.; Gardner, D. K.; Walker, D. W.; Temple-Smith, P.; Papenfuss, A. T.; et al. De Novo Transcriptome Assembly for the Spiny Mouse (Acomys Cahirinus). *Sci. Rep.* **2017**, *7* (1), 8996.

(141) Forster, S. C.; Finkel, A. M.; Gould, J. A.; Hertzog, P. J. RNA-EXpress Annotates Novel Transcript Features in RNA-Seq Data. *Bioinformatics* **2013**, *29* (6), 810–812.

(142) Yu, T.; Mu, Z.; Fang, Z.; Liu, X.; Gao, X.; Liu, J. TransBorrow: Genome-Guided Transcriptome Assembly by Borrowing Assemblies from Different Assemblers. *Genome Res.*

2020, *30* (8), 1181–1190.

(143) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.* **2005**, *77* (4), 964–973.

(144) Zhu, Y.; Engström, P. G.; Tellgren-Roth, C.; Baudo, C. D.; Kennell, J. C.; Sun, S.; Billmyre, R. B.; Schröder, M. S.; Andersson, A.; Holm, T.; et al. Proteogenomics Produces Comprehensive and Highly Accurate Protein-Coding Gene Annotation in a Complete Genome Assembly of Malassezia Sympodialis. *Nucleic Acids Res.* **2017**, *45* (5), 2629–2643.

(145) Sheynkman, G. M.; Johnson, J. E.; Jagtap, P. D.; Shortreed, M. R.; Onsongo, G.; Frey, B. L.; Griffin, T. J.; Smith, L. M. Using Galaxy-P to Leverage RNA-Seq for the Discovery of Novel Protein Variations. *BMC Genomics* **2014**, *15* (1), 703.

(146) Ma, Z. Q.; Chambers, M. C.; Ham, A. J. L.; Cheek, K. L.; Whitwell, C. W.; Aerni, H. R.; Schilling, B.; Miller, A. W.; Caprioli, R. M.; Tabb, D. L. ScanRanker: Quality Assessment of Tandem Mass Spectra via Sequence Tagging. *J. Proteome Res.* **2011**, *10* (7), 2896–2904.

(147) Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* **2015**, *31* (19), 3210–3212.