GENOMIC AND TRANSCRIPTOMIC STUDIES ON NON-MODEL ORGANISMS

A Dissertation

by

SHUHUA FU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Sing-Hoi Sze
Co-Chair of Committee,	Aaron Tarone
Committee Members,	Margaret Glasner
	Vlad Panin
Head of Department,	Gregory Reinhart

May 2015

Major Subject: Biochemistry

ABSTRACT

As the advance in high-throughput sequencing enables the generation of large volumes of genomic information, it provides researchers the opportunity to study non-model organisms even in the absence of a fully sequenced genome. The hugely advantageous progress calls for powerful sequencing assembly algorithms as these technologies also raise challenging assembly problems: (1) Some RNA products are highly expressed but others may have much lower expression level. (2) Data cannot easily be represented as linear structure, due to post-transcriptional modification like alternative splicing. (3) Conserved sequences in domains in gene families can result in assembly errors, (4) Sequencing errors due to technique limitations. Useful assembly algorithms are required to overcome the difficulties above. In these studies, there is often a need to identify similar transcripts in non-model organisms to transcripts found in related organisms. The traditional approach to address this problem is to perform *de novo* transcriptome assemblies to obtain predicted transcripts for these organisms and then employ similarity comparison algorithms to identify them. I observe it is possible to obtain a more complete set of similar transcripts from transcriptome assembly by making use of evolutionary information. I apply new algorithms to study non-model organisms which play an important role in applied biology.

Moreover, improvement of sequencing technologies and application of current algorithms also help to study interkingdom signals between blow flies and bacteria community. With current computational tools, I annotate genomes of *Proteus mirabilis*

and *Providencia stuartii*, which play an important role in bacteria-insect interaction. The study shows significant features of these strains isolated, which provides useful information to develop and test hypothesis in related interactions in insects and bacteria.

ACKNOWLEDGEMENTS

My committee members Dr. Sing-Hoi Sze, Dr. Aaron Tarone, Dr. Margaret Glasner and Dr. Vlad Panin keep me on the right track and provide valuable suggestions. Single colonies of *Proteus mirabilis* and *Providencia stuartii* were isolated by Dr. Tawni L. Crippen and Dr. Qun Ma. Sequence data of *Melilotus albus* and *Melilotus siculus* were generated by Dr. Peter L. Chang, Dr. Maren L. Friesen and Dr. Natasha L. Teakle. Computations were done on the Whole Systems Genomics Initiative Cluster and the Brazos Cluster at Texas A&M University. This work was supported in part by the National Science Foundation [DBI-0820846, MCB-0951120] and the National Institute of Justice [2012-DN-BX-K024].

NOMENCLATURE

ORF	open reading frame
GO	gene ontology
FDR	false discovery rate

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Central dogma of biology

In the central dogma of biology, there are three general processes which occur in biological systems: DNA replication, transcription and translation. During DNA replication, a complementary strand is synthesized with single-stranded DNA as template and the process is catalyzed by multiple DNA polymerases. DNA segments within gene regions can be transcribed into RNA with catalysis by RNA polymerase. RNA products include non-coding RNA (ncRNA) and coding RNA. Many ncRNAs like ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA (miRNA), small nucleolar RNA (snoRNA) and small nuclear RNA (snRNA) are known to have housekeeping functions (1). For example, miRNAs are single strand ncRNA with length of around 22 nucleotides. They are found to work in concert to inhibit expression of target mRNA. Long non-coding RNAs (lncRNA) are expressed from genomic regions including intergenic regions, the opposite strands of mRNAs and introns of genes. They have been identified to play an important role in epigenetic regulation, transcriptional regulation, post-transcriptional regulation (2). On *Drosophila* male X chromosome, ncRNA *roX* forms complexes with male-specific lethal (MSL) proteins to mediate dosage compensation, a process of transcriptional upregulation on the chromosome X so that males express equal or similar numbers of gene products as females (3). In eukaryotic organisms, the immediate product after transcription is primary transcripts, which are not

functional, therefore posttranscriptional processing is required to produce mRNA (4). Nucleic acid sequence in mRNA is translated into amino acid sequence during protein synthesis. Proteins provide important functions in biological processes, such as catalysis by enzymes, oxygen transport by hemoglobin and oxygen storage by myoglobin (5).

The chapter will review current knowledge about sequencing technologies, mRNA processing in eukaryotes especially alternative splicing and *de novo* assembly strategies for non-model organisms and importance of some non-model organisms.

The expression and splicing of genes can impact many aspects of biology. One important aspect of biology that is critical to understand from numerous applied perspectives is the interaction between bacteria and eukaryotes. For instance, genomes of bacteria isolated from blow flies show genetic differences from clinical strains, which may contribute to physiological distinctions. Blow flies are eukaryotic non-model organisms. There is a study that alternative splicing of PGRP family mediates interaction between flies and bacteria (6). To understand interkingdom signaling between flies and bacteria, studies on genomes from microbial community isolated from flies as well as transcriptomes of flies may give useful hints. The advance of sequencing technologies paves a way for this cross-species study.

Sequencing methods

To get sequence information, people used to reverse transcribe mRNA into cDNA, shear cDNA into small fragments and clone them to get large numbers of random cDNA fragments for Sanger sequencing. The conventional sequencing method, Sanger

sequencing, is named after its inventor, Frederick Sanger. It is also called chain-terminator method, because it utilizes ddNTPs (dideoxynucleotide triphosphates) as sequence terminators. The conventional Sanger sequencing strategies separate DNA samples into four portions for four independent sequence amplification experiments. In each reaction, only one type of the four ddNTPs (ddATP, ddTTP, ddCTP, ddGTP) is added to terminate the polymerization. DdNTPs lack 3-OH compared to dNTPs, and thus cannot form a phosphodiester bond between two adjacent nucleotides. Therefore they terminate DNA strand extension at different positions according to the specific ddNTPs added. The products are DNA fragments with diverse lengths. After electrophoresis, DNA fragments are separated based on their sizes. Each band corresponds to a nucleotide in the overall sequence (7). In this way, sequence data can be retrieved from a gel with the read length of up to approximately 1,000 nucleotides. Sanger sequencing gets information by termination of sequence extension, runs slow, and is relatively expensive compared to next-generation techniques. However, this is also the gold standard for sequence identification, as it has been used successfully for decades and has well characterized error rates.

The next generation methods of sequencing produce many more sequences per dollar; however, they include more errors per sequence read, and their sources of errors are not well understood compared to Sanger sequencing. The main high throughput sequencing methods considered here are 454 pyrosequencing and Illumina/Solexa, SolLiD sequencing and ion torrent technologies.

1) 454 pyrosequencing

The 454 pyrosequencing technique depends on detection of pyrophosphate release. It was invented by 454 Life Science, a biotechnology company in Connecticut (www.454.com). It relies on emulsification PCR (polymerase chain reactions) to guarantee immobilization of DNA fragments during amplification. First, DNA libraries are prepared by shearing long sequences into shorter ones, adding adaptors to both ends and dissociating dsDNA into ssDNA. Then adaptors immobilize DNA fragments on the surface of beads. Each bead carries a unique ssDNA fragment. The fragments will be amplified in a water-and-oil mixture, which is a microreactor. When emulsified PCR finishes, amplified fragments are loaded onto a sequencing instrument. Four types of dNTPs (dATP, dCTP, dGTP, dTTP) are added sequentially to attach to 3' end of the primer if it is complementary to the template, producing pyrophosphate. Pyrophosphate can react with adenosine 5' phosphosulfate to generate ATP, which is then used to convert luciferin to oxyluciferin to emit light. The signals are captured and analyzed to get the sequence. Extra dNTPs are digested by apyrase before next cycle begins. The read length is around 200-400 nts. Per sequence base, 454 pyrosequencing is cheap and fast, compared to Sanger sequencing, but is not sensitive to homopolymers, for which it gives ambiguous base calls because there is no linear relationship between detection signals and the number of identical nucleotides. Also signals of long sequence with identical nucleotides may be above the detection range (8). Such problem can be overcome by Illumina/Solexa.

2) Illumina/Solexa

Illumina/Solexa sequencing is another high throughput sequencing technique. It utilizes four fluorescently-labeled nucleotides to sequence fragments from the surface of a flow well after bridge amplification in parallel. First, long sequences are sheared to get shorter fragments and adapters are added to both ends. The adapters enable fragments to attach to specific positions on the surface of the flow well, where there are two PCR primers attached and one of them has a cleavable site. The fragments hybridize to one primer and serve as template to synthesize complementary strand after polymerase and unlabeled nucleotides are added. After synthesis of new strand, original fragments are denatured and removed. New strands bend to hybridize another PCR primer forming 'bridges'. The primer hybridized extends to form a complementary strand. After cycles of denaturation and extension, ssDNA are still attached to the surface. The strands extended from the primer with the cleavage site are removed. DNA fragments left are loaded on a sequencing device. Four labeled nucleotides are added simultaneously to attach to the template. Each type of nucleotide emits a specific fluorescence if it is attached to the template. A fluorescence signal particular to the addition of each nucleotide is captured to get the sequence data. Illumina/Solexa sequencing costs less per base considering the machinery and chemicals used, and also run at a faster rate compared to traditional method and 454 pyrosequencing because all four types of nucleotides are added simultaneously to synthesize the complementary strand in sequencing; it also overcomes the homopolymer problem by relying on base-by-base sequencing. However the read length is much shorter, around 75-300nts, because it is

harder for longer fragments to “bridge” efficiently on the surface of flow cell, thus the resolution decreases dramatically for longer fragments (9). However, short reads make it more difficult for sequence assembly as they are unable to resolve problems in assembly related to long repeat sequences.

3) SoLiD sequencing

SoLiD is a technique developed by Life Science in 2008. SoLiD depends on 2-base encoding in ligation-based sequencing. After DNA libraries are prepared, fragments are attached to the surface of bead for emulsification PCR like 454 pyrosequencing. When amplification is completed, the adapter attached to the free end attaches to the surface of a flow well via 3' modification to the strand. The surface of the flow well is now coated with beads each attached with a single DNA species. The nucleotide detection is not based on polymerase-driven amplification. Instead, eight-base probes are added. In the probe, named from 3' end, the 1st and 2nd bases are specific which possibly involved in hybridization, the 3th-5th are universal bases which can replace any of the four normal bases (A/T/G/C) without destabilizing duplex interaction, and the last 3 bases are degenerate bases which can replace at least two but not all of the four normal bases. There are 16 possible dual base combinations, so 16 types of probes are added to detect the sequence. Four colors are used to differentiate dinucleotides in 4th and 5th positions of the probe, each corresponding 4 possible dinucleotide combinations. After a universal primer with specific length is attached to the template, if the first dinucleotide in the probe is complement to the template, it ligates to 5' end of the primer to hybridize the

template with ligase, emitting specific fluorescence and the image is captured (Step 1 in Figure 1-1). Unextended strands are protected by phosphatase. Cleavage agent cleaves the 6-8th bases of the probe (Step 2 in Figure 1-1). If dinucleotide from another probe is complement to the template immediately to the 5' end of previous probe, it ligates to 5'end of previous probe (Step 3 in Figure 1-1) and cleavage of 6-8th bases is repeated. The cycle repeats until sequence extension is completed and fluorescence signals are retrieved (Step 4 in Figure 1-1). Then newly synthesized strand is melted and removed, a primer one base shorter than the previous primer is used to repeat the extension to get another set of fluorescence signals. The cycle repeats with primer one base off the previous primer (Step 5-6 in Figure 1-1). After 5 repeats with different-length primers, the sequence information can be retrieved by analyzing the result. It is sensitive, because a single base difference gives rise to 2 separate signal differences, while other technologies only cause one. The read length is ~50-75nts (10). Short reads also make barrier for assembly.

4) Ion torrent semiconductor-based sequencing

Ion torrent sequencing technology was developed by Ion Torrent Systems Inc. The preparation of DNA libraries is similar to 454 pyrosequencing to generate beads attached with short single strands as templates for polymerization. Beads are loaded onto microwells with transistor-based sensors. Four types of unlabeled dNTPs are added separately, if polymerase incorporation occurs, that is, the added deoxynucleotides can be attached to 3'OH end of the growing strand, one of the products hydrogen ions (H^+)

will be released and the pH change will be detected by the sensor as voltage change and translated into readable signals. Unused deoxynucleotides are washed away before the next test to ensure only one type of bases is incorporated in each trial. Considering only one type of deoxynucleotides is added each time, the sequencing time is not short and this technology may be limited by homopolymer detection. When there is a long fragment of identical bases, the signal may be above the detection range. It can produce reads as long as ~400nts (11).

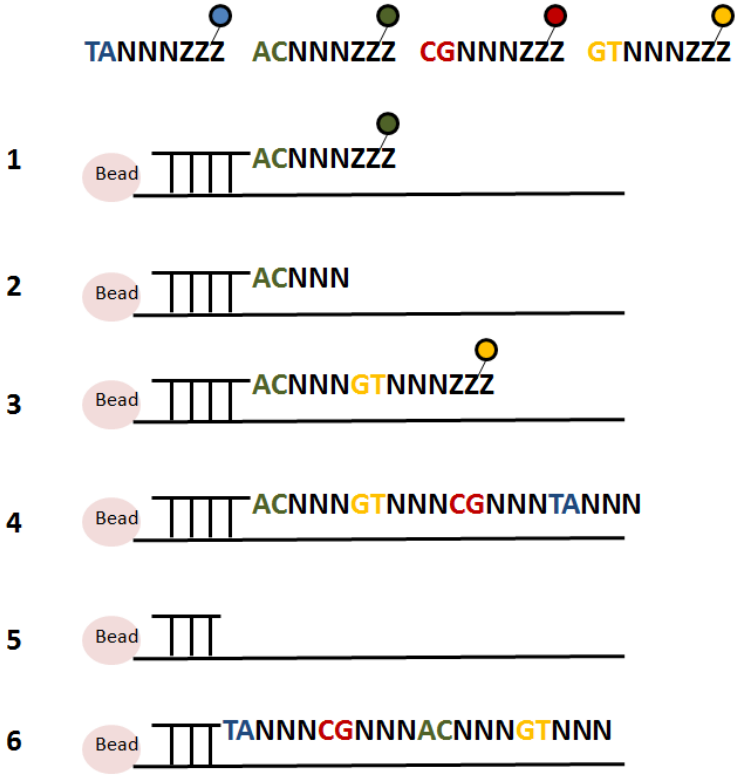


Figure 1- 1 Principle for SoLiD sequencing

Single-molecule sequencing technologies

For the second generation of sequencing, many techniques give short reads after amplification, although technologies like 454 pyrosequencing can detect longer strands, but are not sensitive to homopolymers. Short reads build overwhelming challenges in sequence assembly. The emerging single-molecule DNA sequencing approaches provide new hope in genome research as it provides the potential to produce sequences of very long lengths, though at the moment error rates are very high. They are methods to determine sequences at single base resolution. The new approaches include True Single-Moledule Sequencing (tSMS) developed by Helicos BioSciences, Single Molecule Real Time (SMRT) sequencing by Pacific BioSciences, nanopore sequencing demonstrated by Deamer's group in 1996 and others. The first two approaches have been commercialized.

The tSMS technology is a sequencing-by-synthesis approach. DNA double strands are dissociated, sheared, attached to 3' poly(A) tails, labeled and blocked by terminal transferase. The templates are immobilized with their poly(A) tails covalently bound to poly(T) fragments on a surface. The surface is incubated with solution with one labeled dNTPs. If the dNTPs can be incorporated to 3'end of the growing strand, terminators will be removed and a fluorescence signal will be release. Sequencing information can be deduced from the released signals. Unincorporated nucleotides will be washed away before next test. It does not require PCR amplification, avoiding amplification error and reducing experiment cost. However, the average length of reads is ~30-35 bases, limited

by reversible terminators. It is still challenging to detect strings of consecutive identical base (12).

The SMRT sequencing technology was developed by Pacific Biosciences based on observation of individual fluorophores during DNA synthesis. Different from most sequencing-by-synthesis approaches, fluorescence labels are on the terminal phosphate rather than the bases, with different colored fluorophores on different nucleotides. When a nucleotide is incorporated into the growing strand, fluorescence intensity from the zero-mode waveguides (ZMW) located in SMRT chips is elevated. After the formation of phosphodiester bond, the labeled phosphate group is cleaved from the nucleotide by DNA polymerase. The labeled phosphate group quickly diffuses out of ZMW and ends the fluorescence pulse. It can sequence reads with length of ~3000bp on average with raw error rates of 10-20% limited by photo destruction of DNA polymerase. The accuracy can be improved by repeated sequencing (12).

Nanopore-based sequencing determines base type while ssDNA molecules pass through nanopores with different electrical signals. The sequencing concept was first demonstrated by Deamer et al in 1996. Nanopores are prepared from α-hemo lysine covalently attached with cyclodextrin. Electrical current runs through the pore. After an exonuclease cleaves ssDNA, single bases fall into the nanopores and block the current. Different signals correspond to different nucleotides, thus the signal can be amplified to get the sequence. It can process longer strands even with homopolymers. It does not need DNA labeling, thus it is cheaper compared to fluorescence-label based sequencing. Considering DNA molecules are destroyed as they are read, it is less likely to re-read the

same strand, an improved strategy was proposed to identify individual nucleotides when DNA strands pass intact through nanopores. The possible read length can be ~50,000 bases, without DNA amplification, but the technology is challenged by the cost of the instrument and electrical noise during detection (12).

The progress of sequencing technologies helps to understand post-transcriptional processing such as alternative splicing in eukaryotic organisms. Advanced sequencing technologies provide huge RNA-seq data with splicing information inside at lower cost per base compared to traditional technologies.

Posttranscriptional processing in eukaryotes

Primary transcripts in eukaryotes are still not functional after transcription. They have to undergo removal of non-coding regions by splicing and addition of the 5' cap (m^7G or 7-methyl-guanylate cap) and 3' poly(A) tail, which includes around 100-250 adenosines. The mature mRNA is shown in Figure 1-2. It is composed of 5' cap, 5' untranslated region (UTR), coding sequence, 3' UTR and poly(A) tail from 5' end to 3' end. During processing, nonexpressed intervening sequences called introns are removed, expressed sequences called exons are retained, 5' cap and 3' poly(A) tail are added to 5' end of the first exon and 3' end of the last exon, respectively. Both the cap and poly(A) tail are used to protect mRNA from nucleolytic degradation. The cap consists of a 7-methyguanosine (m^7G) residue which is joined to the nucleotide at 5' end of the transcript. The poly(A) tail, with the length of around 250nt, is appended to 3' end of the

RNA transcript. 5' UTR plays a role in translational control. In 3' UTR, cis-acting elements regulate mRNA stability (13). Both 5' UTR and 3' UTR sequences are not coding sequence and thus not involved in translation. Therefore, both sequences can still be found in transcriptome but not in proteome.

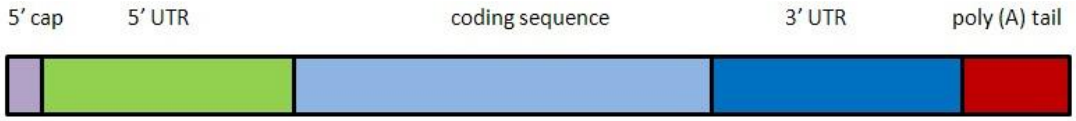


Figure 1- 2 Mature mRNA

Alternative splicing

In eukaryotes, alternative splicing is a common event. It has been found that around 43% genes in fission yeast *Schizosaccharomyces pombe* contain introns (14). In *Drosophila melanogaster*, about 46% of the genes showing different expression patterns during development probably due to alternative usage of promoters or alternative splicing (15). In *Homo sapiens*, about 40-60% of the genes have alternative splicing events (16).

Alternative splicing is not only found frequently in eukaryotic organisms, but also plays a biologically important role. The gene *Dscam* (Down syndrome cell adhesion molecule) in *D. melanogaster* can encode more than 38,000 diverse transcript products due to alternative splicing. Temporal and spatial regulation of alternative splicing of *Dscam* plays an important role in neuronal wiring specificity (17). The expression of

Doublesex (Dsx) transcription factor in *D. melanogaster* is sex-specific, controlled by a cascade of splicing factors which are alternatively spliced themselves (18). Exon skipping of gene BRCA1 caused by nonsense or missense mutations results in breast cancer in humans (19). The expression level of a circadian clock gene called LATE ELONGATED HYPOCOTYL (LHY) in *Arabidopsis thaliana* can be decreased by temperature-associated alternative splicing (20).

This differential use of exons is achieved through a well characterized splicing mechanism. Pre-mRNA is composed of introns and exons. Before splicing, introns have to be recognized via three fundamental signals: 5' donor site, 3' acceptor site and a poly-pyrimidine tract before 3' acceptor site (21). There are multiple types of alternative splicing events: exon skipping, in which an exon can be excluded from a transcript, intron retention, in which an intron can be included after splicing, alternative donor/acceptor, in which donor/acceptor site can be contained in the spliced product. Thus one gene can encode multiple proteins (22).

Splicing events are catalyzed by a splicesome: a complex of U1, U2, U5 and U4/U6 snRNPs (small nuclear ribonucleoproteins), pre-mRNA and various pre-mRNA binding proteins. After assembly of splicesome, splicing occurs in two stages. In the first stage, 2'-OH of an adenosine located close to 3' splice site, nucleophilically attacks the phosphate at the 5' splice site to form 2'5'-phosphodiester bond with 5' end phosphate group, becoming a branch nucleotide. The adenosine is generally located around 20-50 residues upstream of 3' splice site. The second stage includes addition of 3' OH of the previous exon to the 5' end of the next exon forming a phoshodiester bond, as well as

cleavage of RNA at 3' splice site. After exons are joined to each other, the transcript is formed and introns are released in a lariat structure (23).

Alternative splicing mechanism plays an important role in physiological changes to environmental factors, studies on expression of spliced variants under specific conditions and their expressional levels are of great interest to scientists. To understand the splicing information, transcriptomes from eukaryotic organisms need to be sequenced and assembled.

Transcriptome assembly strategies

There are currently two transcriptome assembly strategies. Mapping-first method, such as Cufflinks (24) and Scriptures (25), perform splice-aware alignment of short reads to the reference genome and then assemble transcription products from spliced alignments. It can reconstruct transcripts independent of known splice sites and identify novel mRNA products. But this strategy relies on reference genome and is complicated by sequencing and alignment errors. The alternative strategy, assembly-first approach, also called *de novo* approach, involves software like assemblers Velvet and ABySS, and their post-processing modules Oases and Trans-ABySS respectively. Assemblers assembles RNA reads *de novo*, and then post-processing modules construct predicted transcripts based on assembly data, which can be further aligned to the genome if available by users. It does not require a reference genome, but is less sensitive to construct transcripts which are less abundant and complicated by short reads.

Both strategies above build directed graphs and go through paths in graphs to find diverse transcripts. Graph representation is more suitable than linear structure. First, some positions in pre-mRNA have more than one possibility of splicing, which introduces branches. Second, graph representation contains all possible transcripts and the relation between different transcripts in a concise way. The graph most frequently used is the de Brujin Graph, which is developed by Dr. Pevzner's group. It is different from an overlap graph, in which reads correspond to vertices and edges connecting two vertices correspond overlap (26). In the de Brujin graph, reads from sequencing are decomposed into multiple *k*-mers, with *k* as a parameter called hash length set by users representing the number of nucleotides in a fragment. Each node in the graph represents a *k*-mer, which overlaps with adjacent ones. The overlapping length is *k*-1 nucleotides. Nodes are connected by directed arcs which show the overlap and order (27). Users can also set another parameter called coverage cutoff *c* representing the minimum times that a k-mer appears in the reads. Nodes in linear structure can be merged into single nodes, but there still exist branches where there can be more than one possibility of splicing. Fig 1-3 shows a simple case when *k*=3, reads are rearranged as 3-mers, with 3 nts each. Adjacent 3-mers are overlapped by 2 nts. One 3-mer, GTC, appears 3 times among the reads, thus its coverage is 3. After the construction of a de Brujin graph, linear sequences are merged into one node to get alternative paths. In this case, adjacent nodes, GTC, TCA and CAG are merged together to become a new node GTCAG. In the rearranged graph shown below in Fig 1-3, single nodes are shown as rectangles of A, C and GTCAG.

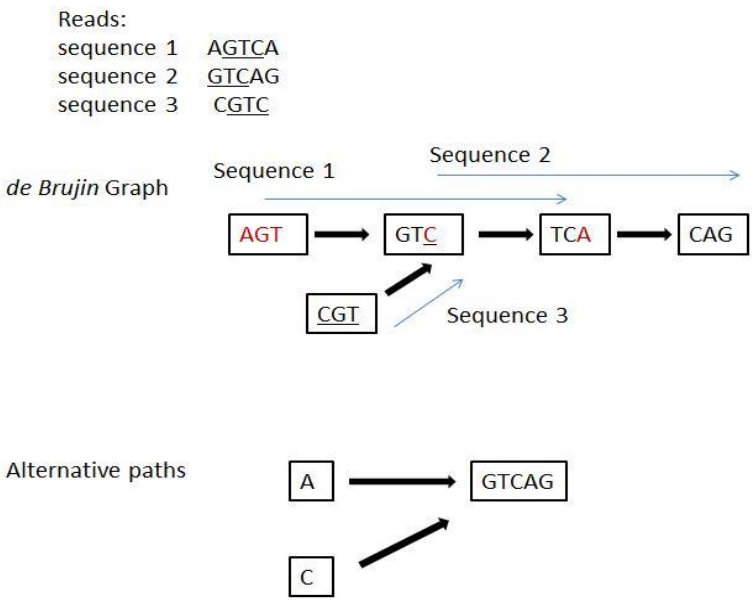


Figure 1- 3 de Bruijn graph construction and alternative paths

After *de novo* assembly by Velvet or ABySS, Oases or Trans-ABySS correspondingly is used to construct predicted transcripts based on coverage from the graph Velvet or ABySS built. However, the prediction process trims out short sequences, which may be part of mRNA products. One way to recover those short sequences is to do construction with BLAST searches. Short sequences may not find a hit in BLAST search, but after connecting to adjacent sequences in the graph, resulting longer sequences may find a better hit in BLAST search than those before connection, and are presumably portion of mRNA products. Thus the more reasonable way is to connect adjacent nodes in the graph based on BLAST result. The process can be conducted by heuristic extension.

Heuristic extension

Performing an exhaustive search to enumerate all possible transcripts from the transcriptome graph is not feasible, but a heuristic algorithm can be used to find a satisfactory solution in reasonable amount of time. One example is shown in Fig 1-4. There are two ways for extension from node 1, extension towards node 2 or node 5. Suppose the path 1→5 after extension towards node 5 shows an improved BLAST score over the path 1→2 after extension towards node 2. To find an optimal path starting from node 1by heuristic extension, extension will not continue in direction towards node 2 and bypass node 3 and node 4. However, exhaustive extension would enumerate all paths starting from node 1, including paths that transverse node 3 or node 4, and it would still identify path 1→5 is the optimal solution. In comparison, heuristic extension can identify path 1→5 as the optimal search solution relatively quickly.

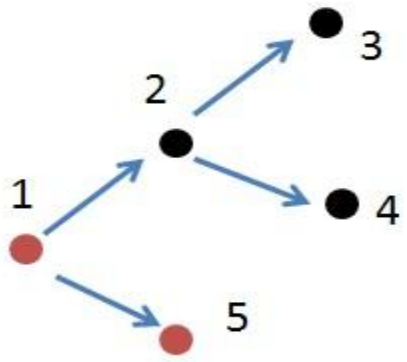


Figure 1- 4 Heuristic extension in the graph

The strategy of *de novo* assembly with heuristic extension can be applied to non-model organism studies. Non-model organisms are organisms which are poorly annotated, because they were not selected for extensive study previously. However, they play an important role in applied biology. Here are the examples of *Melilotus*, *Proteus mirabilis* and *Providencia stuartii* studies that help to answer ecological and evolutionary questions in biology.

Melilotus

Studies on *Melilotus* help to identify transcriptomic features in plants which allow them to tolerate harsh growth conditions. The *Melilotus* genus originating from Eurasia, is a forage legume that fixes nitrogen in root nodules with a symbiotic relationship with rhizobia. It is evolutionarily closely related to *Medicago truncatula*, a legume model organism (28). *Melilotus* can grow in hash environments like high salinity and tolerate waterlogging, while *Medicago* is salt tolerant but susceptible to waterlogging. Studying tolerance of forage legumes plays an important role in risk assessment before recommendation of plants to increase fodder crop production under severe salinity and waterlogging conditions (29).

Among *Melilotus* genus, the species *Melilotus albus* has been identified as productive under saline conditions (28). An earlier study showed that *Melilotus siculus* has high resistance to salinity and waterlogging (30).

Proteus mirabilis and *Providencia stuartii*

Studies on *Proteus mirabilis* and *Providencia stuartii* can advance understanding of bacterial infection in patients and interkingdom signaling between bacteria and flies.

Proteus mirabilis, a Gram-negative rod-shaped pathogen, is a gut commensal bacterium associated with human urinary tract infections (UTI) (31,32). The bacteria produce urease, an enzyme with high molecular weight to catalyze hydroxylation of urea into carbon dioxide and ammonia. Increase of ammonium concentration elevates environmental pH to precipitate normally soluble polyvalent ions like ammonium, phosphate, magnesium and calcium ions, which results in formation of urinary stones (33). Other virulence factors involve flagella which give rise to swarming motility for bacteria to ascend the ureters to the renal tubules, fimbriae which enable bacteria to adhere to kidney epithelium and uroepitheial cells, proteases which avoid host defense as well as hemolysin which causes cytotoxicity(34).

Quorum sensing (QS) is utilized by *P. mirabilis* to sense concentration of secreted small chemical signal molecules (quormons) which reflect cell density and coordinate gene expression (35,36). QS is a process of cell-to-cell communication. Quormons are synthesized within cells and secreted out of the cells. When population density of bacteria community exceeds a specific threshold, there are sufficient quormons to be sensed to initiate concerted actions among bacteria. Different quormons are used by Gram-positive and Gram-negative bacteria to measure density of population (36). A recent study found that some bacteria signaling mechanisms are shared by *Drosophila melanogaster* (37).

Swarming mobility enables *P. mirabilis* to move and spread across surfaces by increasing flagella number and secreting surfactants to reduce surface tension, giving rise to difficulty of isolation from mixed cultures (38,39). When in contact with a solid surface, *P. mirabilis* differentiate into elongated and hyperflagellated swarmer cells from short vegetative swimmer cells (40). Recently, swarming signals associated with *P. mirabilis* have been linked to fly behavior, making the species a model for interkingdom signaling between *P. mirabilis* and *Lucilia sericata* (35).

P. mirabilis has been found in the blow fly *L. sericata* (35,41), a common blow fly used in maggot therapy (42,43), and sterilization of maggot therapy with *P. mirabilis* has been suggested (41,44). Bacteria survive in the guts of flies when added to the flies' diets and may stimulate oviposition for flies by secreting volatile compounds (45). It has been found that maggot secretions contain antimicrobial substances, some of which are metabolic products of *P. mirabilis* (46).

Providencia stuartii has also been found in larvae of blow flies (45) and is phylogenetically closely related to *P. mirabilis* (47), but does not show swarming nature (48). *Providencia* is also distinguished from *Proteus* by producing acid from various sugars and incapability of either hydrolyzing gelatin or producing hydrogen sulfide and lipase (47).

Coinfection of mice with *P. mirabilis* and *P. stuartii* enhances urolithiasis and bacteremia with synergistic induction of urease activity (49). There two species coexist in the catheter biofilm microbial communities (50). Coinfection leads to similar bacterial

load of multispecies infection but urease mutation in *P. mirabilis* results in decreased synergistic induction (49).

Non-model organisms are not well annotated, but studies on non-model organisms show ecological and evolutionary importance and benefit from research on closely related model organisms which are well annotated. Traditional sequencing technologies generate one read per sample, however, next generation sequencing technologies are able to generate millions of reads per sample at lower costs per base. The advance of next generation sequencing technologies enables the genomic and transcriptomic studies on non-model organisms. Advanced bioinformatics algorithms and analysis need to be performed to study the non-model organisms with increased-size read data.

CHAPTER II

IDENTIFYING SIMILAR TRANSCRIPTS IN A RELATED ORGANISM FROM DE BRUIJN GRAPHS OF RNA-SEQ DATA, WITH APPLICATIONS TO THE STUDY OF SALT AND WATERLOGGING TOLERANCE IN *MELILOTUS*

As the advance in high-throughput sequencing enables the generation of large volumes of genomic information, it provides researchers the opportunity to study non-model organisms even in the absence of a fully sequenced genome. These studies often start from sequencing the entire transcriptome, while additional software is applied to process the data. An important mechanism to study is alternative splicing, which is crucial to a variety of biological functions. The goal of these studies is to recover as many isoforms as possible in order to understand the underlying biological processes.

In the presence of a reference database, there are two strategies for analyzing transcriptome data. Mapping-first algorithms perform splice-aware alignment of the reads to the reference genome to reconstruct the transcripts (24,25). While these algorithms can construct transcripts independent of known splice sites and identify novel mRNA products, they only allow very few differences during the alignment. Alternatively, when a reference genome is not available but a reference transcriptome is available, transcript quantification algorithms can be applied to analyze differential expression of genes (51,52).

In the absence of a reference database, an alternative strategy is to employ *de novo* sequence assembly algorithms (27,53-59). A popular strategy of transcriptome assembly

algorithms is to assemble the reads by obtaining a de Bruijn graph that represents the transcriptome (58-61).

Although the de Bruijn graph contains all branching possibilities, an additional step is needed to obtain predicted transcripts from the graph. To obtain information about possible function of these predicted transcripts, a similarity search algorithm such as BLAST (62) is then applied to identify similar transcripts in a related organism. Since the predicted transcripts are constructed based on coverage information, one shortcoming of this approach is that sequences with low coverage are often ignored leading to missed transcripts. The later BLAST step to a related organism then starts from this relatively incomplete set of predicted transcripts.

Instead of performing similarity search from the predicted transcripts, I observe that it is possible to obtain a more complete set of similar transcripts if I start the search from the de Bruijn graph directly (see Figure 2-1). This strategy bypasses the transcript prediction step and makes use of support from evolutionary information. Since the graph retains more information from the transcriptome data, transcripts that have low coverage can still be recovered if they have high similarity with the ones from the related organism. In metagenomics, Wu *et al.* (63) employed a similar idea to extract paths directly from the de Bruijn graph that correspond to homologous genes from closely related species. Recently, Bao *et al.* (64) utilized genomic information from the same organism or a related organism (instead of transcripts from a related organism) to improve *de novo* transcriptome assemblies by first identifying exons from alignments. While the strategy of applying BLAST from each node in a de Bruijn graph to a related

organism can already give a lot of hits, it is possible that some significant hits are missed since the sequence within a node may be too short. There is a need to identify paths in the de Bruijn graph that are similar to transcripts from the related organism. Since the number of possible paths that can be constructed from the de Bruijn graph can be very large, it is not feasible to enumerate all of them.

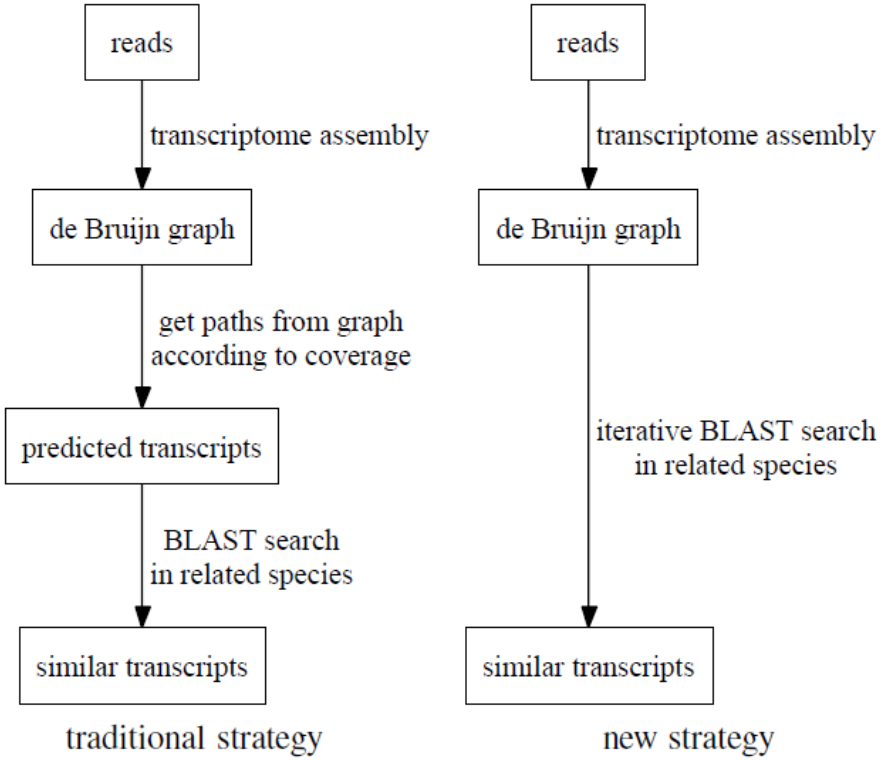


Figure 2- 1 Difference between traditional strategy to obtain similar transcripts and my new strategy that bypasses the transcript prediction step

I develop a heuristic extension algorithm that starts with enumerating short paths in the de Bruijn graph, and iteratively extends these paths in the most promising directions

rather than in all possible directions. This procedure generalizes the BLAST algorithm to allow a non-linear query structure instead of a query sequence. Note that my strategy is different from the one in (63) that uses optimal alignment to extend paths due to the smaller scale of metagenomic data. I compare the performance of my algorithm that starts the search from the de Bruijn graph against existing algorithms that employ the strategy of first obtaining predicted transcripts then applying BLAST to obtain similar transcripts. I validate my algorithm by extracting reads from publicly available RNA-Seq libraries. I construct new RNA-Seq libraries for the non-model organisms *Melilotus albus* and *Melilotus siculus*, and apply my algorithm to study salt and waterlogging tolerance in these two species.

Methods

Given a set of reads and a parameter k, a popular strategy of transcriptome assembly algorithms is to assemble these reads into a de Bruijn graph that represents the transcriptome. By taking each k-mer that appears within the reads as a vertex, and connecting two k-mers by a directed edge if the $(k-1)$-suffix of the first k-mer is the same as the $(k-1)$-prefix of the second k-mer, the de Bruijn graph implicitly assembles the reads by linking together the same k-mer that comes from different reads (65,66). This strategy is very popular among short read assembly algorithms (27,54,55,57,58).

To minimize the effect of sequencing errors, these algorithms remove short tips and further simplify the de Bruijn graph by collapsing similar paths. Each linear path that

contains a sequence of vertices with no branches is collapsed into a single node, and a *k*-mer coverage cutoff *c* is imposed to remove low coverage nodes (27,57,58).

While the resulting de Bruijn graph contains all branching possibilities, it can contain complicated cycles. I cannot consider each connected component as a splicing graph that specifies the alternative splicing paths of a single gene (67). I develop an algorithm to extract paths in the de Bruijn graph that correspond to similar transcripts in a related organism. Each extracted path can be considered as a predicted transcript in the original organism.

Initial choice of contigs to extend

For each transcript in a related organism, my goal is to recover the best path in the de Bruijn graph that corresponds to the transcript. My approach is based on the seed-extension strategy that starts from short paths, and iteratively extends these paths in promising directions. I start the search from nodes in a de Bruijn graph that correspond to contigs from short read assembly algorithms (27,57,58).

Given a de Bruijn graph G=(*V*,*E*), a database of known transcripts in a related organism *T* and an *e*-value cutoff e_f , I first apply BLAST from each node in the de Bruijn graph to the transcript database to obtain all hits with *e*-value below e_i, where $e_i > e_f$ (see Figure 2-2, step 1). The extra *e*-value cutoff e_i is chosen to allow the initial seed nodes to be of lower quality. Some of these nodes can be extended later into longer paths that are of higher quality.

Algorithm extContig $(G=(V,E), T, e_i, e_f, n)$
input: a de Bruijn graph $G=(V,E)$, a transcript database T, initial e-value cutoff e_i, final e-value cutoff e_f, and a parameter n;
output: a set of paths in G that correspond to similar transcripts in T;
1. apply BLAST from all nodes in V to T to find all hits with e-value below e_i;
2. $V' \leftarrow \emptyset$;
3. for each transcript t in T do
4. $V' \leftarrow V' \cup$ (set of top n nodes in V with the best e-value to t);
5. $P \leftarrow V$;
6. for each node u in V' do
7. $p \leftarrow u$; $e \leftarrow$ best e-value of u in step 1;
8. repeat
9. for each outgoing edge $v \rightarrow w$ in E from $p = u \rightarrow \cdots \rightarrow v$ do
10. apply BLAST from path $u \rightarrow \cdots \rightarrow v \rightarrow w$ to T;
11. $P \leftarrow P \cup \{u \rightarrow \cdots \rightarrow v \rightarrow w\}$;
12. if the best e-value in steps 9–10 is less than e then
13. $p \leftarrow$ best path in steps 9–10; $e \leftarrow$ best e-value in steps 9–10;
14. until the e-value of p no longer improves;
15. repeat steps 7-14 starting from the twin node u' of u to get path p';
16. construct path $p'' = y \rightarrow x \rightarrow \cdots \rightarrow u \rightarrow \cdots \rightarrow v \rightarrow w$ from $p = u \rightarrow \cdots \rightarrow v \rightarrow w$ and $p' = u' \rightarrow \cdots \rightarrow x' \rightarrow y'$, where x' (and y') is the twin node of x (and y), and apply BLAST from p'' to T to compute its e-value;
17. $P \leftarrow P \cup \{p''\}$;
18. for each transcript t in T do
19. report the path p in P with the best e-value to t below e_f;

Figure 2- 2 Algorithm extContig that starts the search for similar transcripts from the de Bruijn graph instead of from predicted transcripts. Steps 1–4 choose an initial set of contigs to extend. Steps 5–17 implement the heuristic extension. Steps 18–19 report the results

For each transcript in the database, I extract top n nodes in the de Bruijn graph that give the best BLAST hits to it, where n is a given parameter (see Figure 2-2, steps 2–5). The resulting collection of nodes over all transcripts in the database becomes the set of

all nodes that my heuristic extension algorithm extContig will start from, which are the ones that are most likely to have correspondences with transcripts in the database. Note that more stringent values of *k* and the *k*-mer coverage cutoff *c* can provide longer nodes to start with but can also lead to missed nodes.

Heuristic extension

For each node u in the collection, I extend its sequence by one node along all outgoing edges from u, and apply BLAST from each of these extended sequences to the transcript database. If at least one of these extended sequences gives a better *e*-value, I extract the top extended path that gives the best *e*-value. I repeat the extension procedure starting from this new path until either there are no more outgoing edges to extend from or the *e*-value no longer improves (see Figure 2-2, steps 7–14).

Note that during each extension, only one best direction is chosen. Extending in more than one direction is very time-consuming since the number of possibilities can be exponential even in the absence of cycles. Although it is possible that the real best path may be missed, it is still possible to resolve different isoforms since the heuristic extension procedure starts independently from multiple nodes, some of which may be specific to particular isoforms. The procedure can be applied even in the presence of cycles in the de Bruijn graph since the *e*-value cannot improve indefinitely.

I perform a similar procedure on the node u' that is the twin node of u, which represents the reverse complementary sequence of *k*-mers on the opposite strand, and try to extend it in the opposite direction. In addition to adding these two extended paths

from u and u' to the set of candidate paths, I also merge the twin path that is complementary to the extended path from u' with the extended path from u to obtain a longer path. I add the merged path to the set of candidate paths and identify its best BLAST hit in the transcript database (see Figure 2-2, steps 15–17).

Extraction of similar transcripts

At the end of the procedure, for each transcript in the database, I report the top path that gives the best *e*-value to it among all the candidate paths if such a path exists, where the set of candidate paths includes all paths that BLAST has been applied (see Figure 2-2, steps 18–19). Only the nodes of a path that are in the best BLAST alignment are reported. It is possible that some of these paths may be the same or very similar for different transcripts in the database.

Melilotus RNA-Seq

MRNA was extracted from *Melilotus albus* and *Melilotus siculus* using a Qiagen Oligotex mRNA mini kit. Fragmentation of mRNA was done using an Ambion fragmentation buffer. Construction of the cDNA library was based on the Illumina protocol. First strand cDNA synthesis was done using Random Hexamer Primers (Invitrogen) and second strand synthesized using a DNA Polymerase 1 (Promega). End repair was carried out to create uniform blunt ends (Epicentre End-IT repair kit). Unique 4 bp adaptors (Illumina) were added so that the libraries could be pooled for sequencing. An ‘A’ base was added using a Klenow enzyme (3' to 5' exo minus, NEB) and adaptor

ligation was performed using Epicentre Fast-Link DNA ligation kit. The cDNA template was run on a 2% agarose gel at 120 V for 60 minutes and fragments of approximately 200–500 bp were removed and purified (Zymo gel purification kit). The purified cDNA template was PCR enriched using the Illumina primers and a Phusion polymerase. The library was quantified using an Invitrogen Qubit fluorometer. Libraries were sequenced on an Illumina Genome Analyzer II under normal conditions and conditions associated with salt tolerance or waterlogging tolerance or both as single-end 100 bp reads, which were trimmed to 71 bp.

Results

To assess the performance of my algorithm, I extracted reads from publicly available RNA-Seq libraries (see Table 2-1). I validate my algorithm on model organisms by applying BLAST to a database of annotated transcripts in each model organism itself and in two other related model organisms with varying evolutionary distances, including *Schizosaccharomyces pombe* against another yeast species *Saccharomyces cerevisiae* and another fungus *Neurospora crassa*, *Drosophila melanogaster* against another *Drosophila* species *Drosophila pseudoobscura* and mosquito *Anopheles gambiae*, *Homo sapiens* against squirrel monkey *Saimiri boliviensis* and mouse *Mus musculus*, and *Arabidopsis thaliana* against another *Arabidopsis* species *Arabidopsis lyrata* and rice *Oryza sativa*.

Table 2- 1 Data sets used in the evaluation of my heuristic extension algorithm, with organism indicating the starting organism, related organisms indicating the related model organisms that BLAST is applied to, lib indicating the total number of libraries, size indicating the total number of bases in all the reads after quality trimming, and reference indicating the publication that describes the libraries.

organism	related organism	lib	Size (G, giga base pairs)	reference
Schizosaccharomyces. pombe	*Saccharomyces cerevisiae*	32	17	(59)
	Neurospora crassa		9.6	
Drosophila melanogaster	*Drosophila pseudoobscura*	13	16	(68)
	Anopheles gambiae		16	
Homo sapiens	*Saimiri boliviensis*	4	16	(69)
	Mus musculus			
Arabidopsis thaliana	*Arabidopsis lyrata*	5	16	(70)
	Oryza sativa			
Lucilia. sericata	*D. melanogaster*	9	4.6	(71)
Heterocephalus. glaber	*Homo sapiens*	13	61	(72)
Ctenomys. sociabilis	*Homo sapiens*	10	66	(73)
Cicer arietinum	*Arabidopsis thaliana*	3	8.6	(74)
Melilotus. albus	*Arabidopsis thaliana*	12	5.5	new data
Melilotus. siculus	*Arabidopsis thaliana*	12	5.4	new data

I compare the performance of the algorithms on publicly available RNA-Seq libraries from four non-model organisms. The blow fly *Lucilia sericata* is important in medicine, forensic science and agriculture due to its filth feeding habits, its use in maggot therapy, its colonization of human and animal remains, and its ability to cause myiasis in vertebrates (71). The naked mole rat *Heterocephalus glaber* is important in medicine and in biomedical research due to its resistance to cancer and delayed aging, and its ability to live in adverse conditions (72). The rodent *Ctenomys sociabilis* is

important in the study of social behavior of mammals and the relationship to gene expression (73). The chickpea *Cicer arietinum* is one of the most consumed legume crops that grows in arid areas with low productivity (74). Similarity search is performed from *L. sericata* to the model organism *D. melanogaster*, from *H. glaber* and *C. sociabilis* to the model organism *H. sapiens*, and from *C. arietinum* to the model organism*A. thaliana*. The searches that are applied against the same model organism have varying evolutionary distances. I have constructed new RNA-Seq assemblies for the non-model organisms *Melilotus albus* and *Melilotus siculus*, which are important in the study of salt and waterlogging tolerance of forage plants (28). Genomic information on the species will enable the dissection of coumarin production that can be utilized in pharmaceutical development (75). Similarity search is performed from*M. albus* and*M. siculus* to the model organism *A. thaliana*. I trimmed each read by removing all positions including and to the right of the first position that has a quality score of less than 15. For smaller data sets (including *D. melanogaster*, *L. sericata*, *C. arietinum*, *M. albus* and *M. siculus*), I compare the performance of my heuristic extension algorithm

extVelvet starting from the de Bruijn graph given by Velvet (27) against the performance of Oases (61) that is a postprocessing module of Velvet. Since Oases requires that Velvet is run without coverage cutoff and then applies the coverage cutoff itself, I use the de Bruijn graph within Oases that is modified from Velvet's original de Bruijn graph. For the other larger data sets, I compare the performance of my heuristic extension algorithm extABySS starting from the de Bruijn graph given by ABySS (57) against the performance of Trans-ABySS (60) that is a postprocessing module of ABySS.

I applied each algorithm with *k* as 25 or 31, for smaller data sets *c* as 3, 5 or 10 and for larger data sets *c* as 10, 20 or 50. BLAST is applied to predicted transcripts in Oases and Trans-ABySS, to paths in the de Bruijn graph in extVelvet and extABySS, and to contigs in Velvet/Oases and ABySS. To compare each model organism against itself, nucleotide BLAST search is applied to a database of gene transcripts with initial *e*-value cutoff $e_i=10^{-15}$ and final *e*-value cutoff $e_f=10^{-100}$. For the other cases, translated BLAST search is applied to a database of protein transcripts in a related organism with initial *e*-value cutoff $e_i=10^{-6}$ and final e- value cutoff $e_f=10^{-20}$. For each transcript in the database, top 8 nodes (and their twin nodes) are chosen to form the initial nodes for extension. Additional criteria are imposed to extend past very short nodes.

Transcript recovery

I assess the performance of each algorithm in recovering transcripts by investigating the number of similar transcripts obtained, database coverage, alignment length of shared transcripts, and the number of recovered transcripts that are close to full length. While the absolute performance depends on the amount of RNA-Seq data, the complexity of transcriptomes, the evolutionary distance between organisms and the assembly algorithm that is being used, Figure 2-3 shows that Oases and Trans-ABySS generally recover more similar transcripts than their base algorithms Velvet and ABySS, while extVelvet and extABySS generally recover even more.

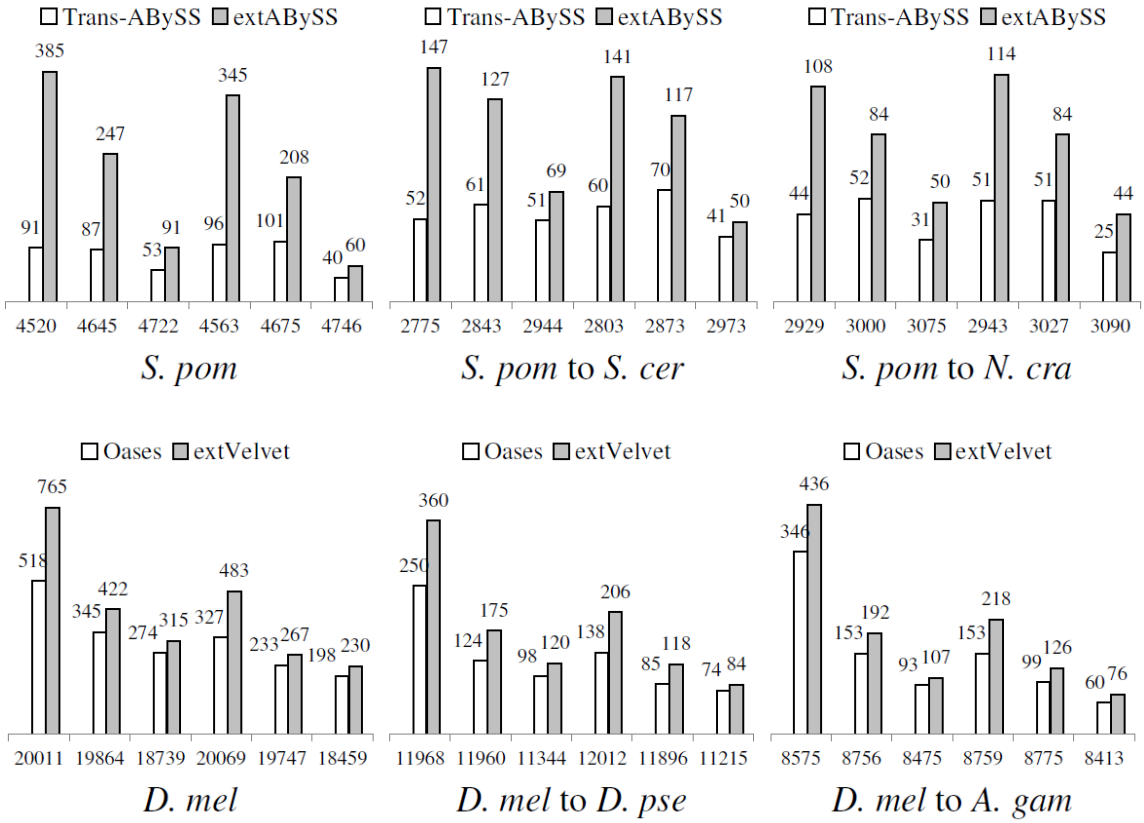


Figure 2- 3 Comparisons of the change in the number of similar transcripts recovered by Oases and Trans-ABySS (shown as white bar) to the change in the number of similar transcripts recovered by extVelvet and extABySS (shown as grey bar) respectively over the numbe r of similar transcripts recovered by Velvet and ABySS (shown under the x-axis) respectively for different values of *k* and *k*-mer coverage cutoffs *c*. Within each graph, the corresponding values of *k*=25/*c*=3, *k*=25/*c*=5, *k*=25/*c*=10, *k*=31/*c*=3, *k*=31/*c*=5, *k*=31/*c*=10 from left to right for smaller data sets, including *D. melanogaster*, *L. sericata*, *C. arietinum*, *M. albus* and *M. siculus*, and *k*=25/*c*=10, *k*=25/*c*=20, *k*=25/*c*=50, *k*=31/*c*=10, *k*=31/*c*=20, *k*=31/*c*=50 from left to right for larger data sets, including *S. pombe*, *H. sapiens*, *A. thaliana*, *H. glaber* and *C. sociabilis*. For comparing each model organism against itself (graphs with a single-species label), nucleotide BLAST search is applied with *e*-value cutoff $e_f = 10^{-100}$. For the other cases, translated BLAST search is applied with *e*-value cutoff $e_f = 10^{-20}$.

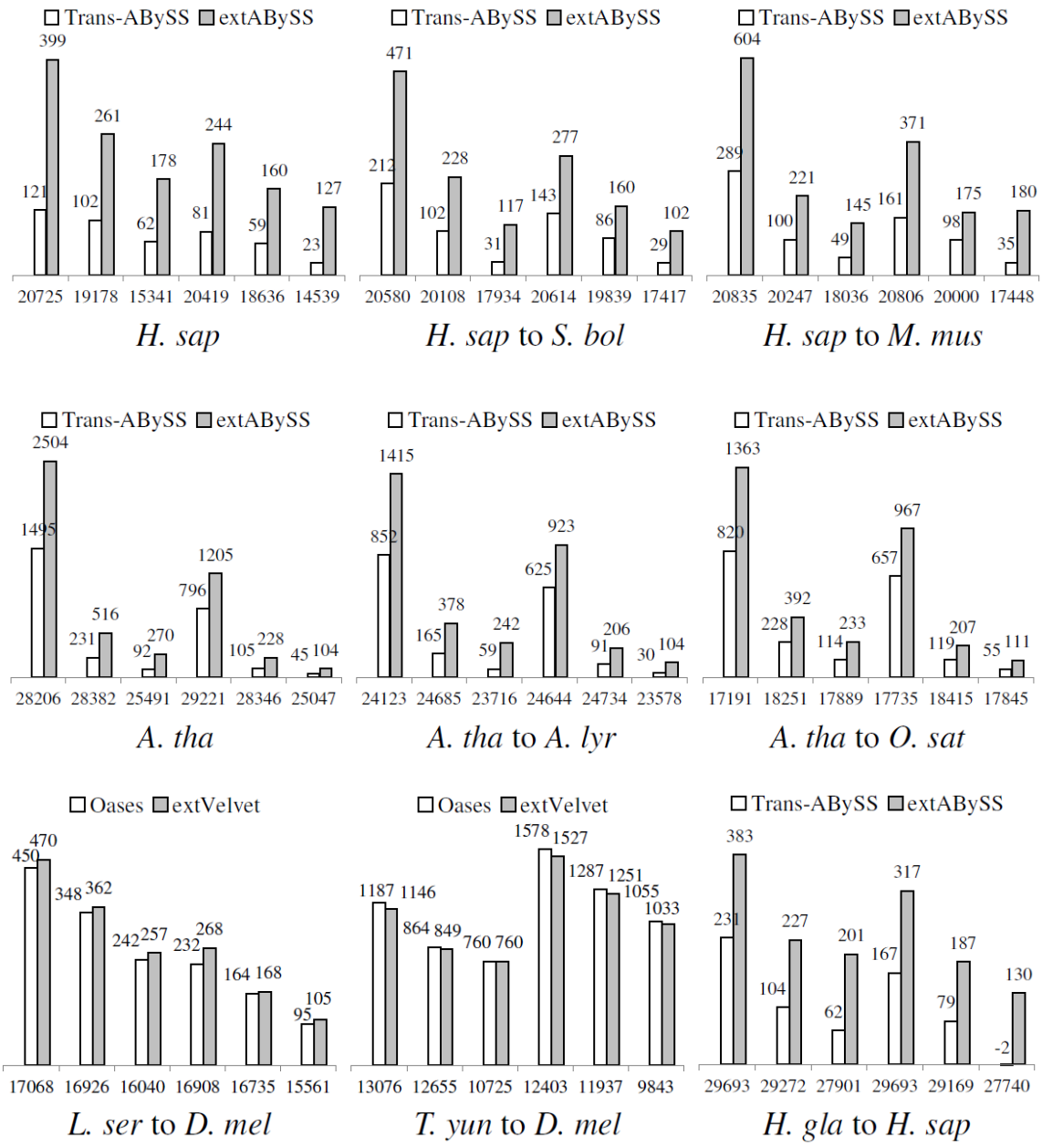


Figure 2- 3 Continued.

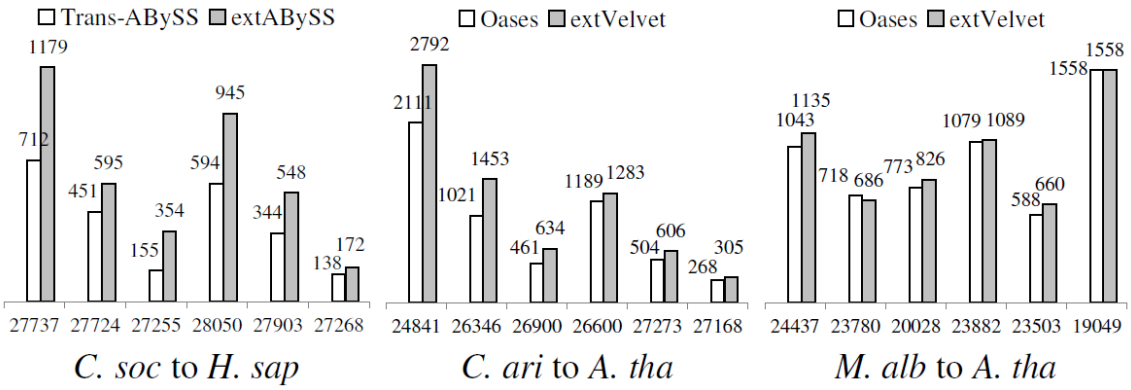


Figure 2- 3 Continued.

Transcript recovery

I assess the performance of each algorithm in recovering transcripts by investigating the number of similar transcripts obtained, database coverage, alignment length of shared transcripts, and the number of recovered transcripts that are close to full length. While the absolute performance depends on the amount of RNA-Seq data, the complexity of transcriptomes, the evolutionary distance between organisms and the assembly algorithm that is being used, Figure 2-3 shows that Oases and Trans-ABySS generally recover more similar transcripts than their base algorithms Velvet and ABySS, while extVelvet and extABySS generally recover even more.

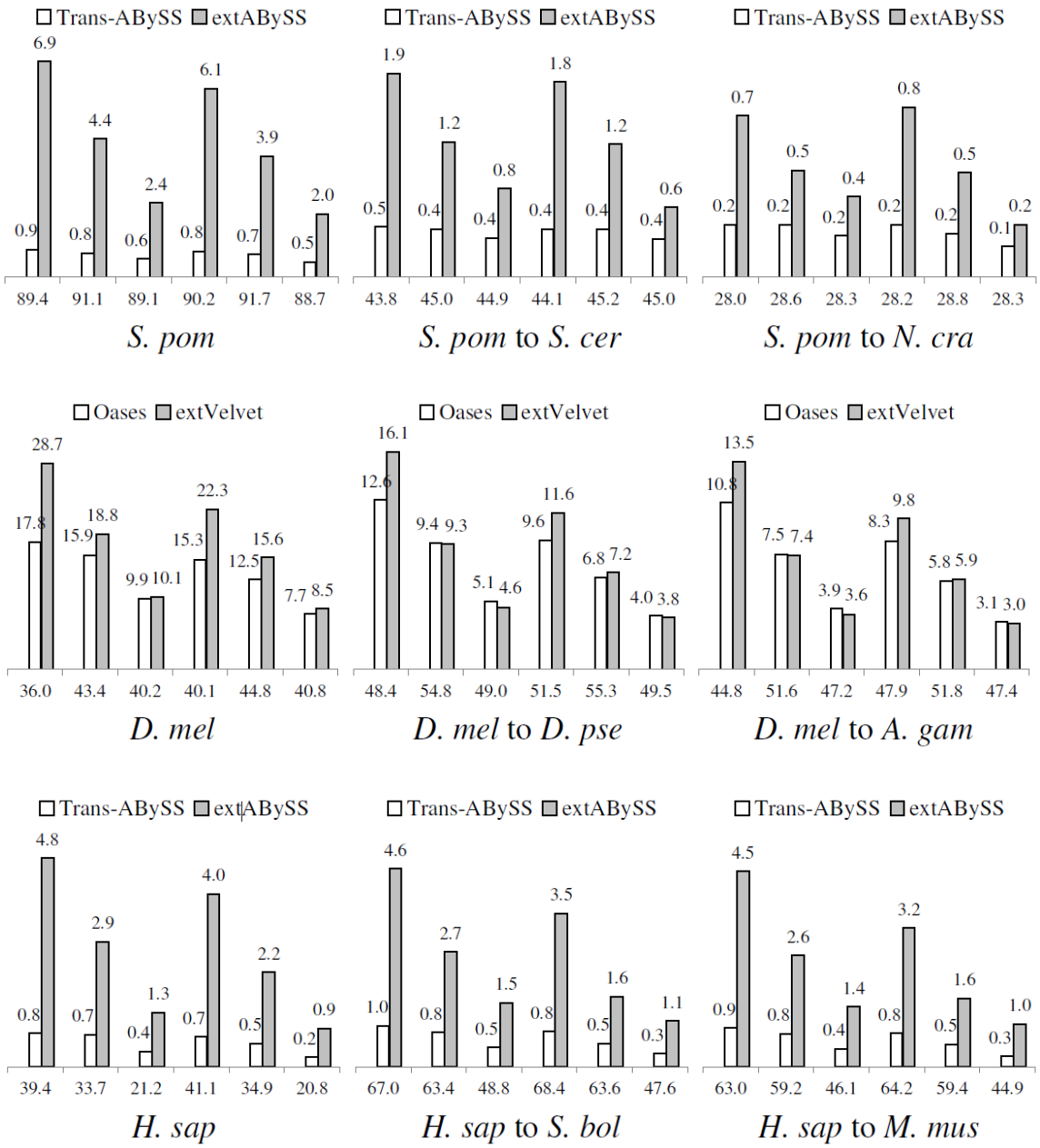


Figure 2- 4 Comparisons of the change in database coverage of Oases and Trans-ABySS to the change in database coverage of extVelvet and extABySS respectively over the database coverage of Velvet and ABySS respectively for different values of *k* and *k*-mer coverage cutoff c. Notations are the same as in Figure 3. Database coverage is defined by the percentage of positions in the transcript database that are included in the best BLAST alignment of each similar transcript.

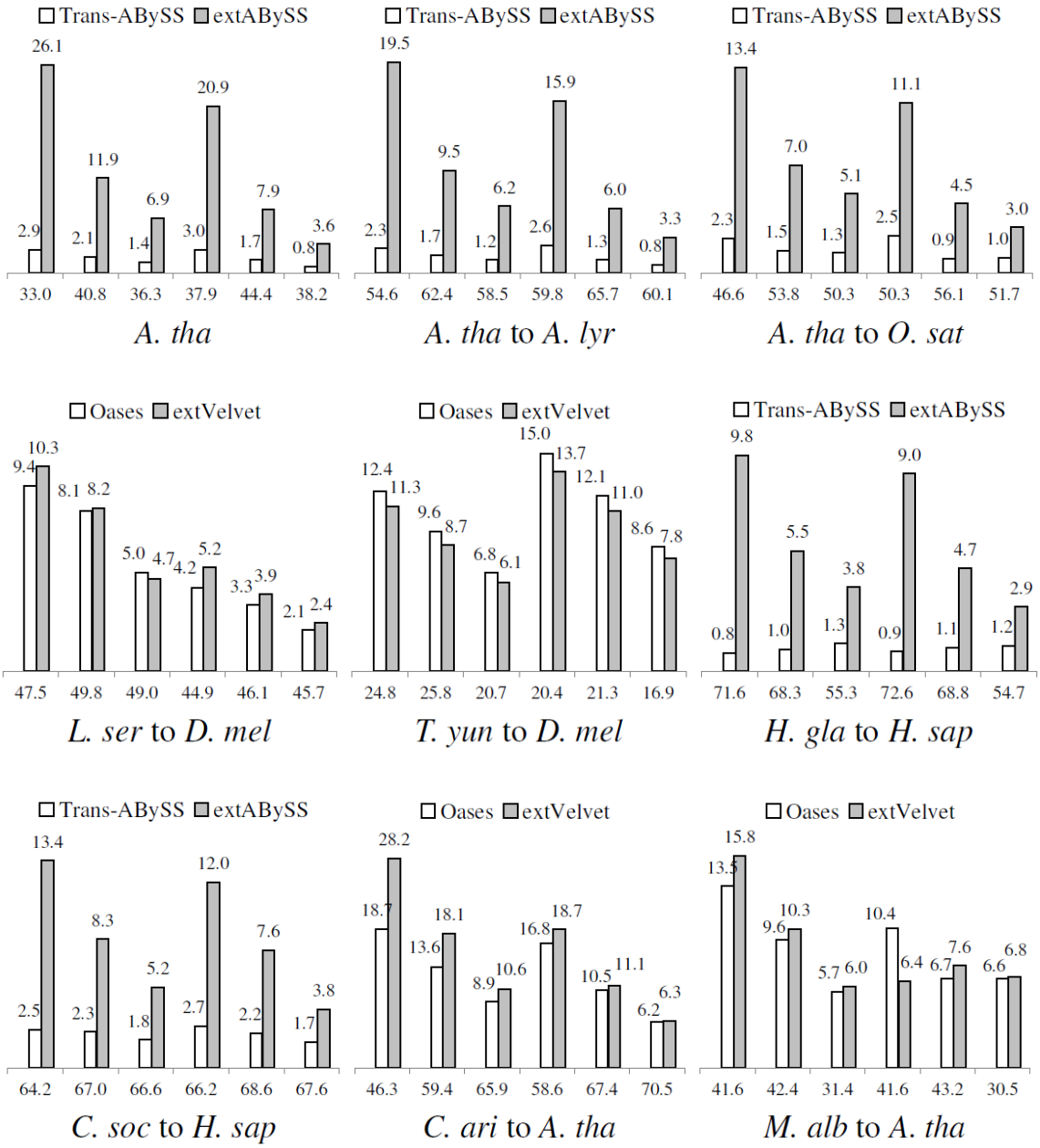


Figure 2- 4 Continued.

While Trans-ABySS recover fewer similar transcripts than ABySS in the case of *H. glaber* to *H. sapiens*, (see Figure 2-3), Figure 2-4 shows that this loss can be offset by

the increase in length of the predicted transcripts over the length of the contigs. While this leads to an overall improvement in database coverage by Oases and Trans-ABySS, extVelvet and extABySS generally improve even more. The improvement in database coverage of Trans-ABySS is small when compared to ABySS, which leads to a much larger improvement of extABySS over Trans-ABySS. These improvements are not absolute since different algorithms can recover different sets of similar transcripts. The base algorithm ABySS already has high performance for *S. pombe* against itself, while the large data set sizes of *H. glaber* and *C. sociabilis* lead to high database coverage for all algorithms (see Table 2-1).

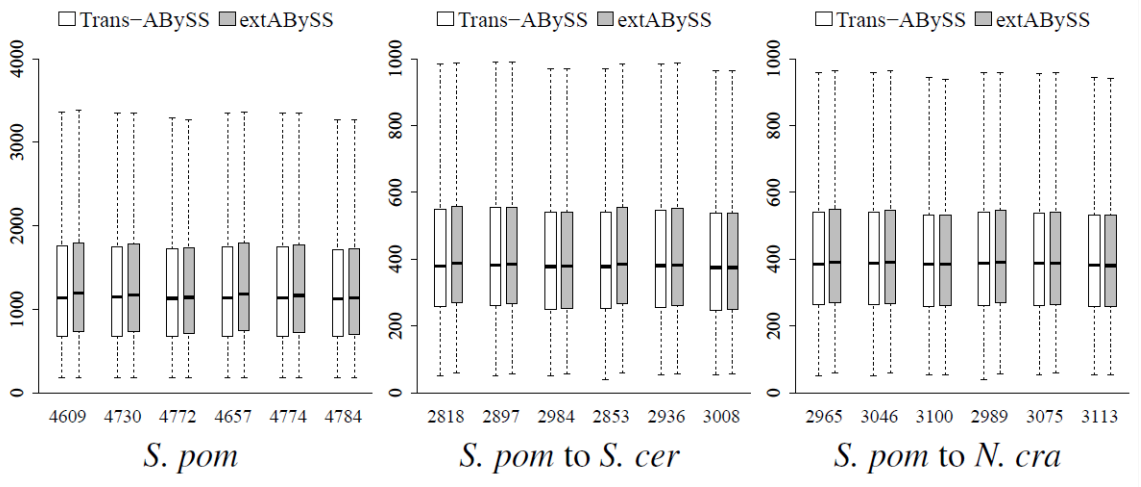


Figure 2- 5 Comparisons of the distributions of the best BLAST alignment length of each similar transcript that is recovered by both Oases and extVelvet (or by both Trans-ABySS and extABySS), with the total number of shared transcripts shown under the x-axis for each value of *k* and *k*-mer coverage cutoff *c*. Y axis shows the distribution of alignment length. Outliers are not shown within each box plot. Other notations are the same as in Figure 3. Alignment length is in nucleotides for comparing each model organism against itself and in amino acids for the other cases.

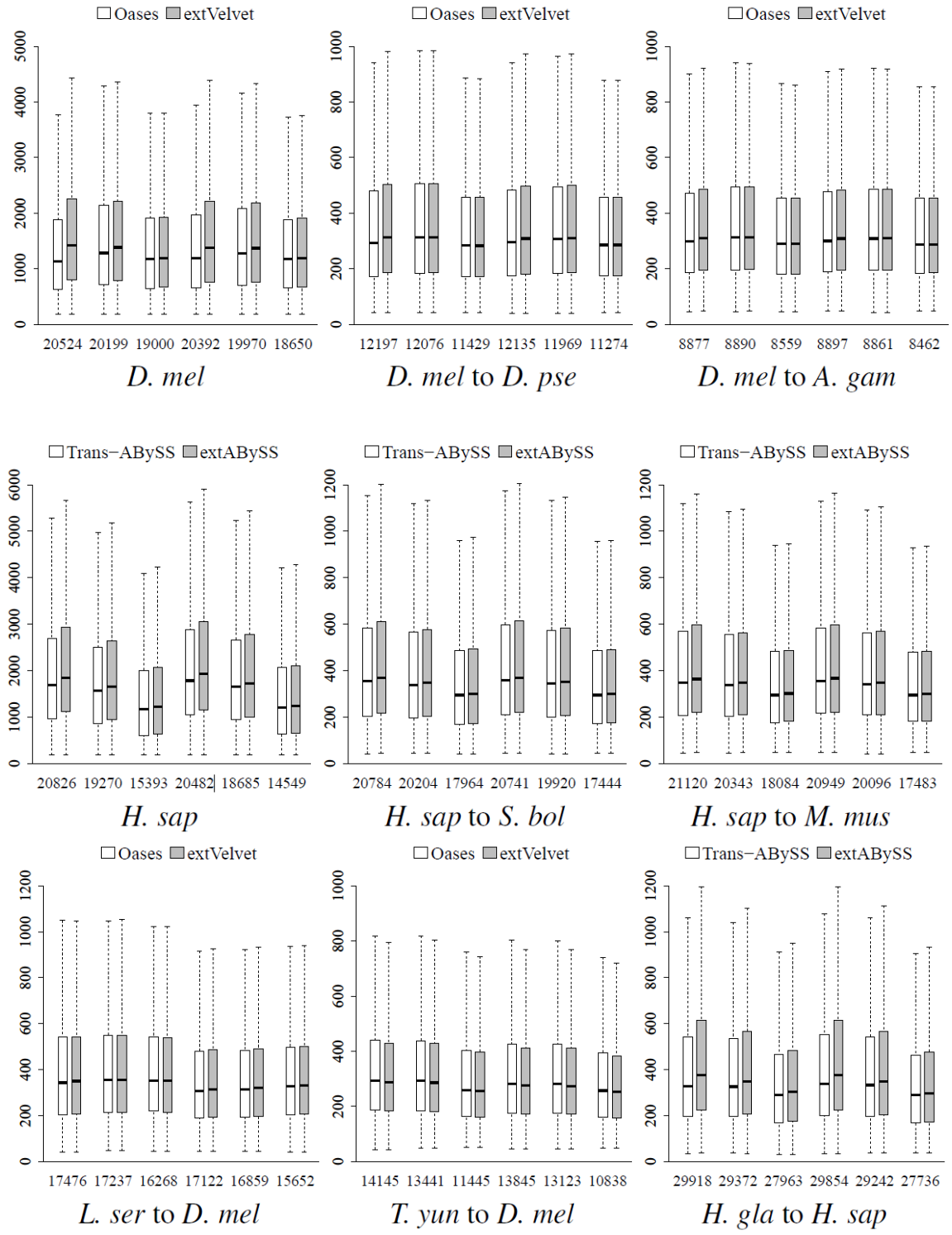


Figure 2- 5 Continued.

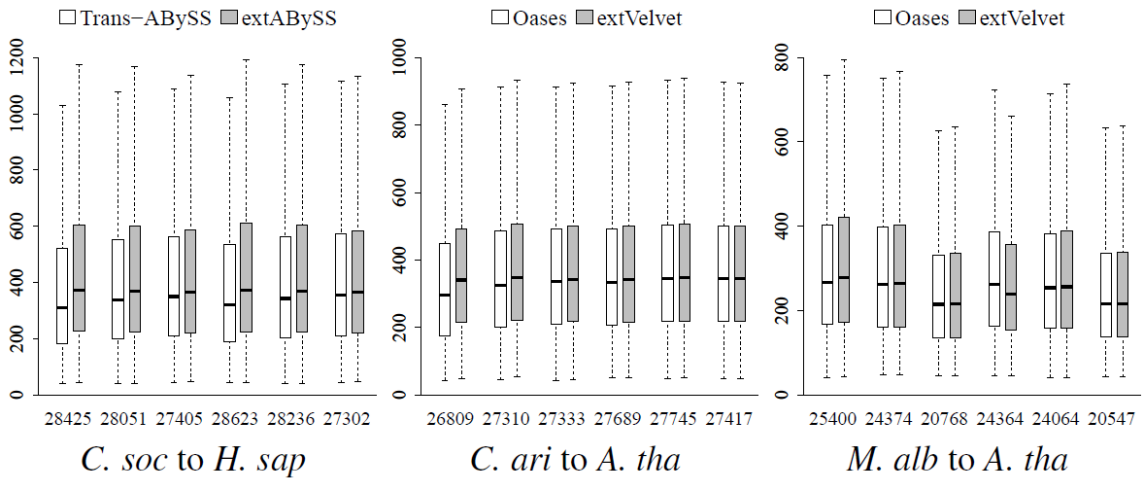


Figure 2- 5 Continued.

Figure 2-5 shows that among the similar transcripts that are recovered by both Oases and extVelvet (or by both Trans- ABySS and extABySS), extVelvet and extABySS can recover longer transcripts in some cases, with large improvements for *A. thaliana*. Most of the recovered transcripts are shared between Oases and extVelvet (or between Trans-ABySS and extABySS) (compare to Figure 2-3).

Figure 2-6 shows that extVelvet and extABySS can recover more similar transcripts that are close to full length than Oases and Trans-ABySS. Both Oases and extVelvet (or Trans-ABySS and extABySS) can recover more full length transcripts than Velvet (or ABySS).

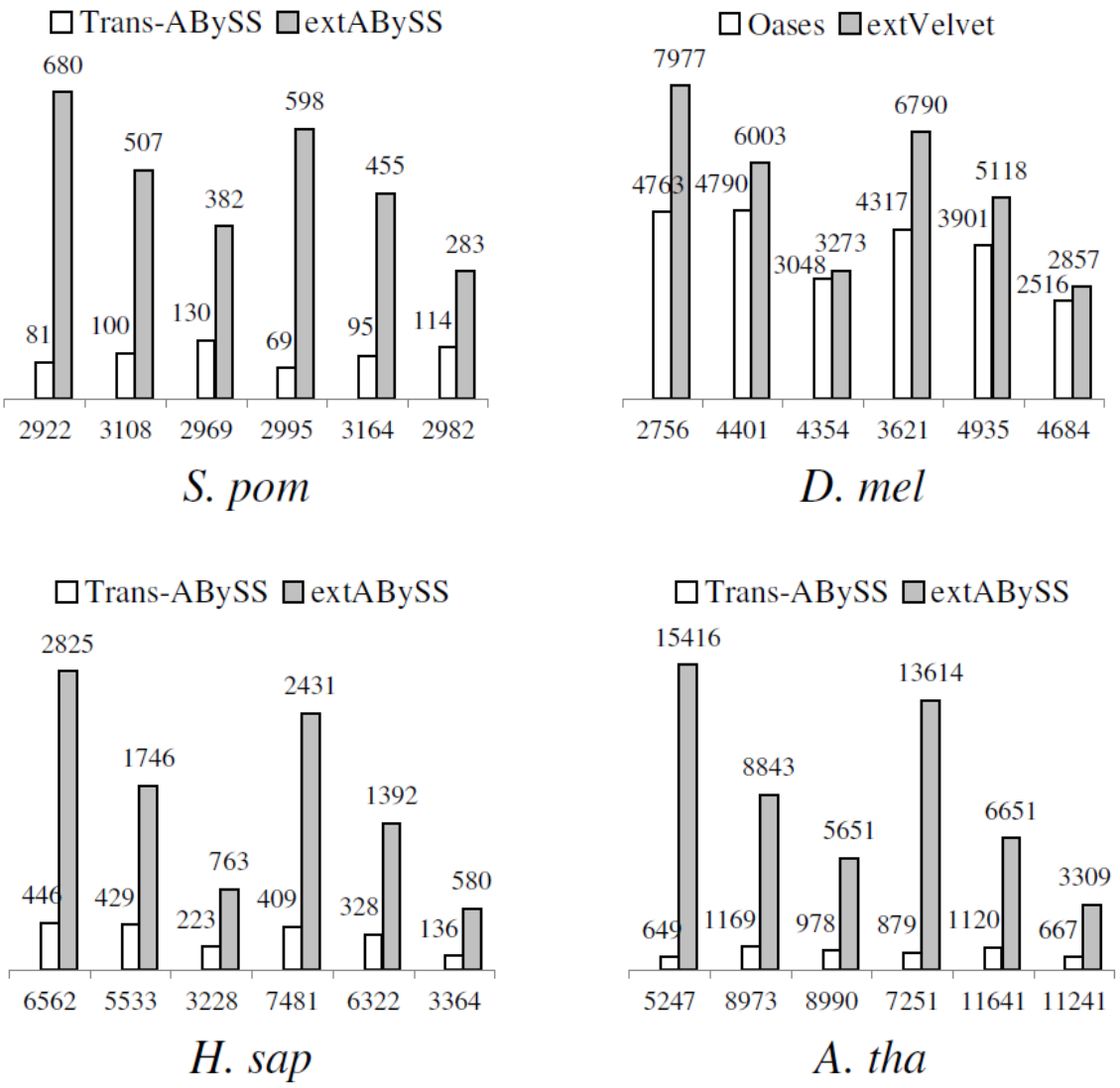


Figure 2- 6 Comparisons of the change in the number of similar transcripts that are 80% full length transcripts (100% full length transcripts for *S. pombe*) and recovered by Oases and Trans-ABySS to the change in the ones recovered by extVelvet and extABySS respectively over the ones recovered by Velvet and ABySS respectively on model organisms for different values of *k* and *k*-mer coverage cutoff c. Notations are the same as in Figure 2-3. These transcripts are the ones in which 80% (100% for *S. pombe*) of the coding region is included in the best BLAST alignment.

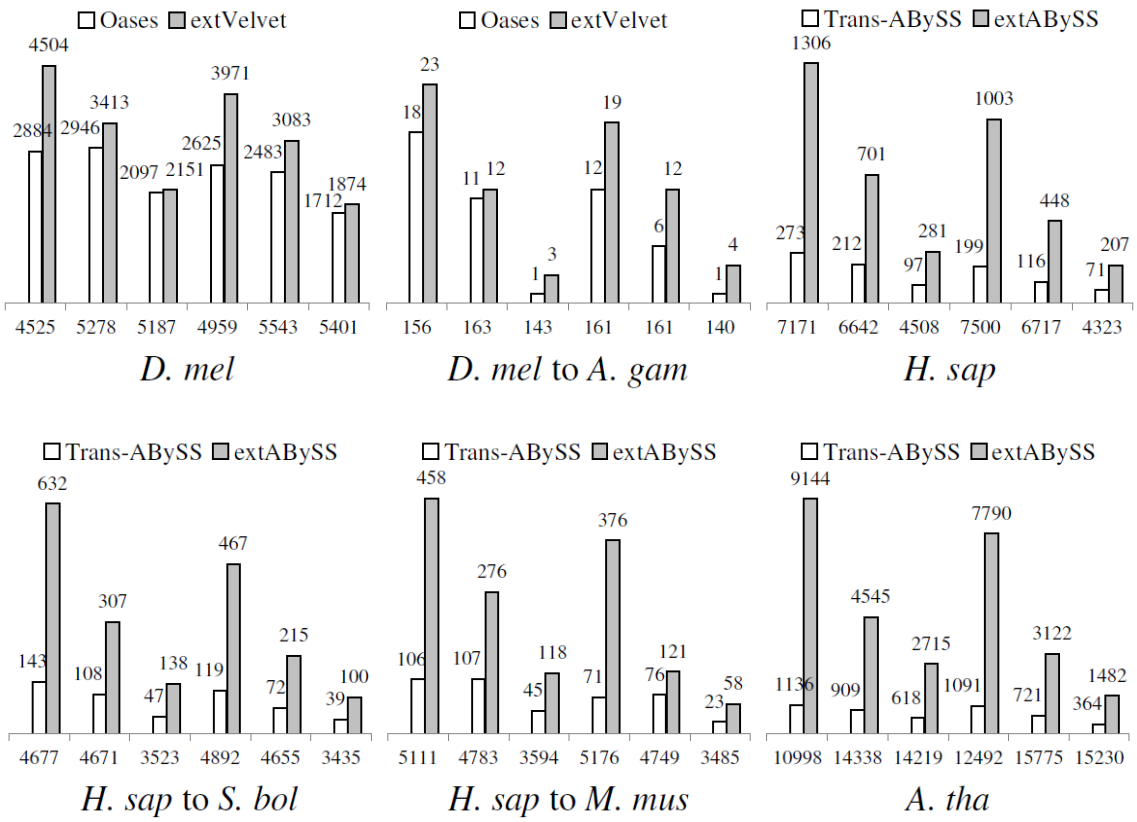


Figure 2- 7 Comparisons of the change in the number of exons that are found in only one annotated transcript of the same gene with multiple isoforms and recovered by Oases and Trans-ABySS to the change in the ones recovered by extVelvet and extABySS respectively ov er the ones recovered by Velvet and ABySS respectively for different values of *k* and *k*-mer coverage cutoff *c*. Notations are the same as in Figure 2-3. Exons within isoforms that do not have the same starting position or the same ending position are considered to be distinct. An exon is recovered if it has some overlap with the best BLAST alignment. Exons within mRNAs are considered for comparing each model organism against itself, while exons within coding regions of the related model organism are considered for the other cases. Results for *S. pombe* are not included since there is little alternative splicing, while a few other results are not included due to poor annotations of alternative splicing in the related model organisms.

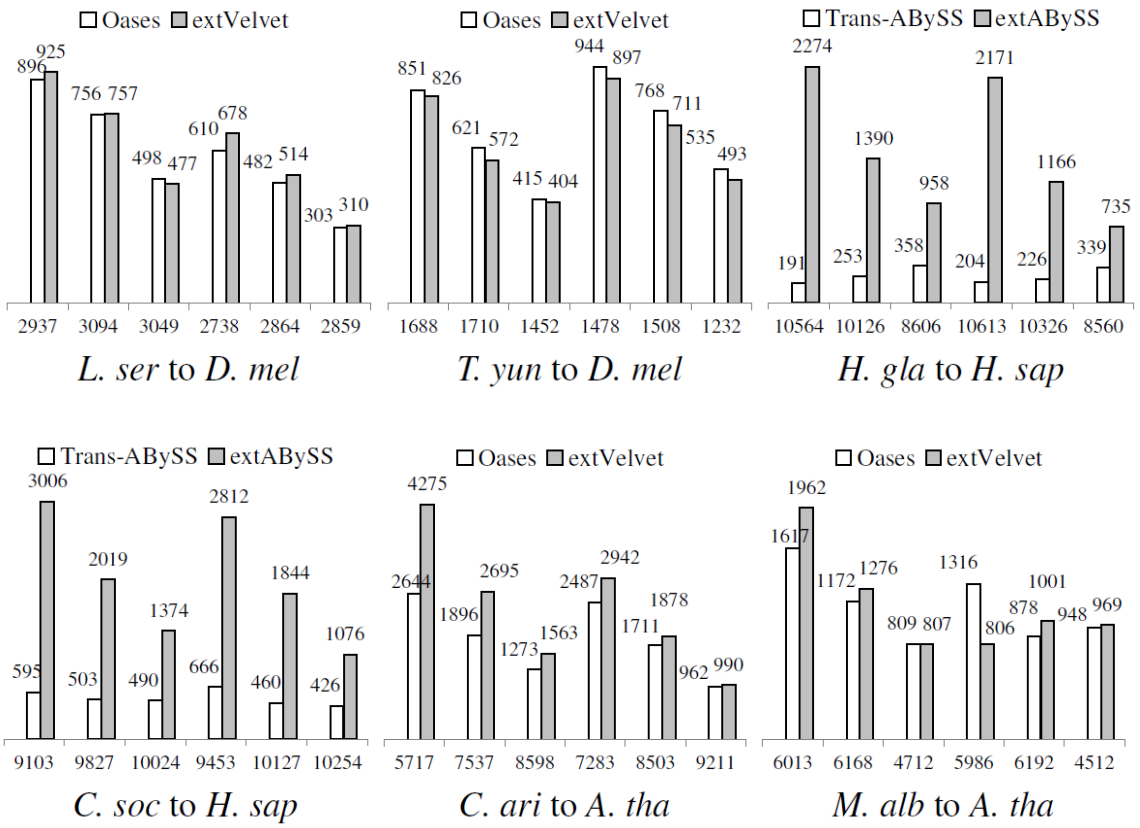


Figure 2- 7 Continued.

Alternative splicing

I assess the ability of each algorithm in distinguishing between isoforms by considering exons that are found in only one annotated transcript of the same gene with multiple isoforms, which are the ones that can resolve them. Figure 2-7 shows that extVelvet and extABySS are able to recover a larger number of such exons in most cases, with large improvements of extABySS over Trans-ABySS. Figure 2-8 shows examples in which extVelvet and extABySS can better resolve isoforms with respect to a related organism, including the *ZDHHC16* gene, which is a zinc finger protein that may be

involved in apoptosis regulation (76); the *dSarm* gene, in which the loss of its function in *D. melanogaster* protects against injury-induced axon death (77); the *STAT3* gene, which is an acute-phase response factor in *H. sapiens* in which the isoforms have unique functions (78); and the *AT4G34660* gene, which is a SH3 domain-containing protein in *A. thaliana* that is involved in clathrin-mediated vesicle trafficking (79).

False positive estimates

I assess the reliability of each algorithm by identifying similar transcripts that are recovered by each algorithm, but are not recovered by a simple protein BLAST search from each model organism to another related model organism with the same *e*-value cutoff. The number of such transcripts serves as the upper limit on the number of false positives (some of these correspondences may actually be correct but not annotated). Figure 2-9 shows that the number of false positives is very small for all algorithms, with extVelvet (or extABySS) having slightly higher values than Velvet and Oases (or ABySS and Trans-ABySS).

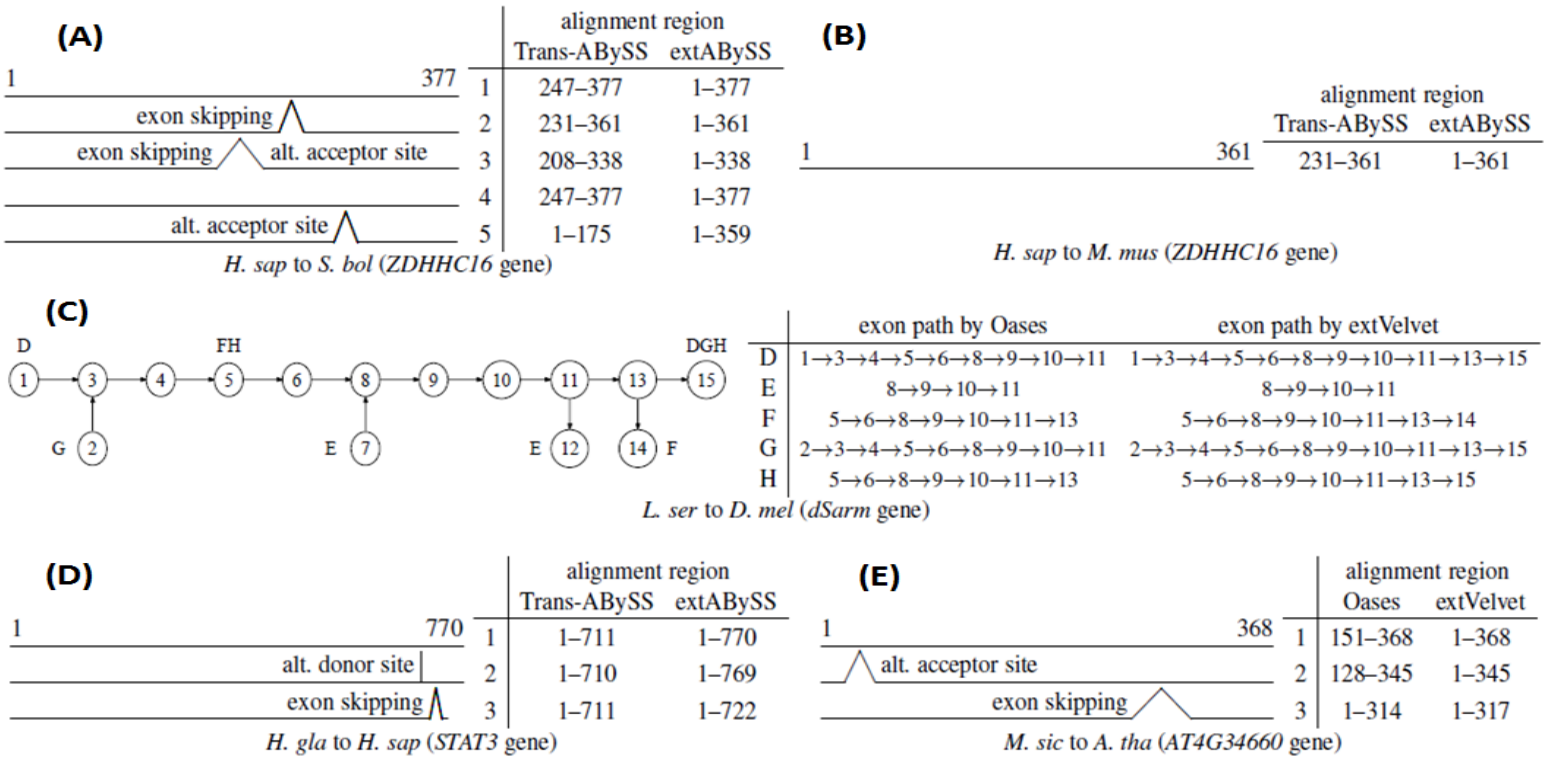


Figure 2- 8 Examples of resolution of alternative splicing with respect to a related organism. The splicing structures are on exons in the coding region of the related organism. For the dSarm gene, uppercase letters indicate isoforms and their start/end exons, with Oases resolving fewer isoforms than extVelvet. In the other lower splicing structures, the isoforms are drawn to scale and the starting and ending amino acid positions of isoform 1 are shown. For the *ZDHHC16* gene, Trans-ABySS cannot resolve between its different isoforms on *S. boliviensis*, and recovers a much shorter segment of it on *M. musculus* with no known alternative splicing. Trans-ABySS cannot resolve isoforms 1 and 3 of the *STAT3* gene, while Oases cannot resolve isoforms 1 and 2 of the *AT4G34660* gene.

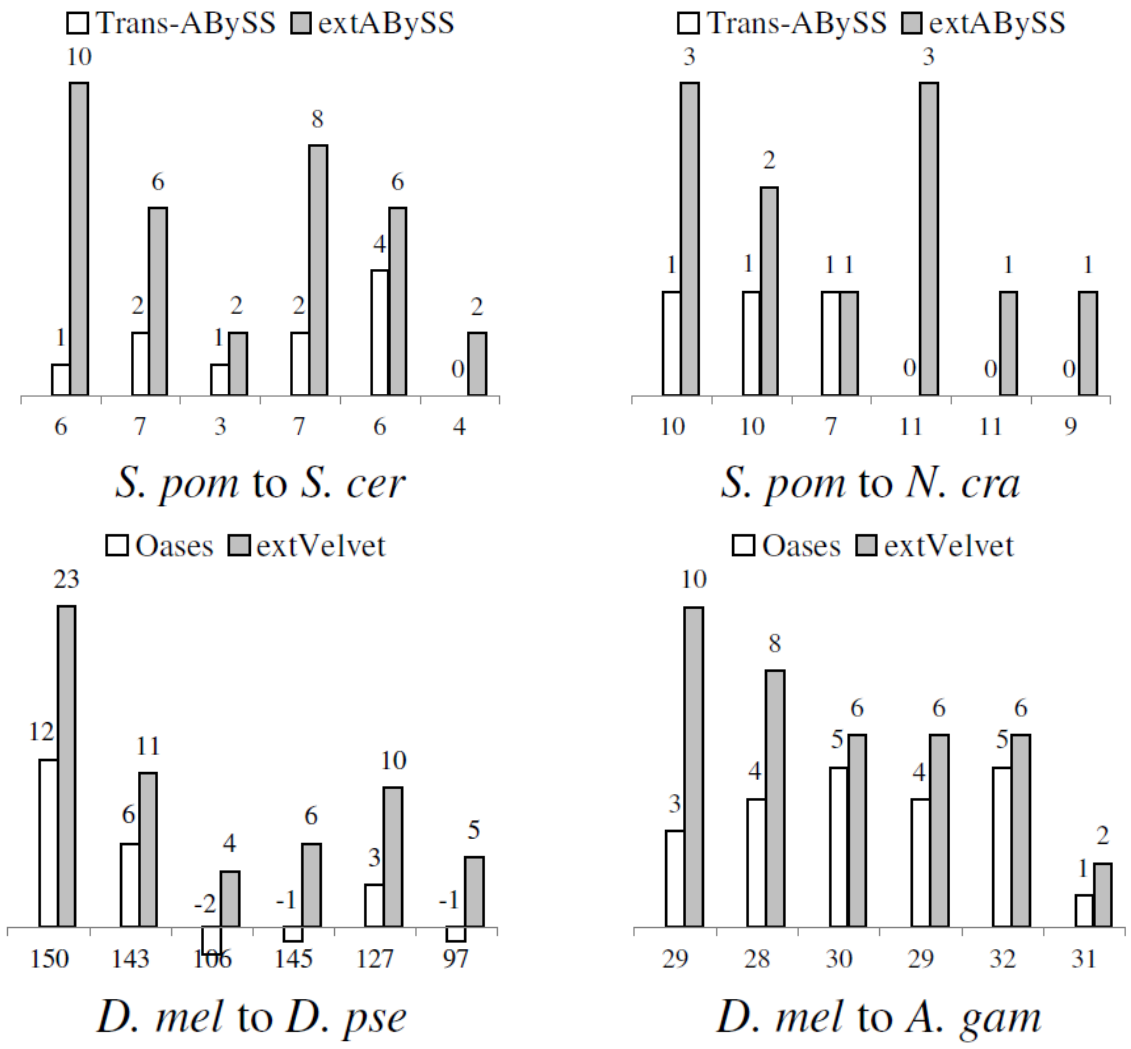
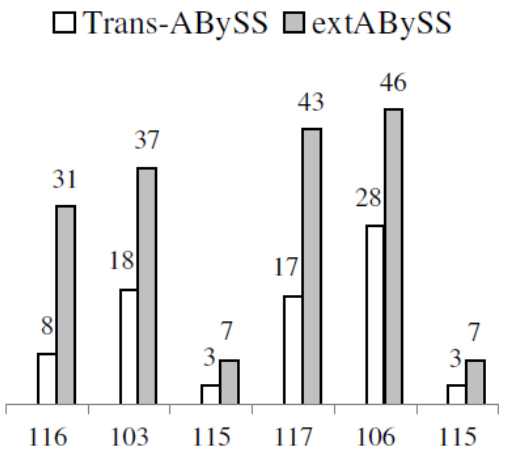
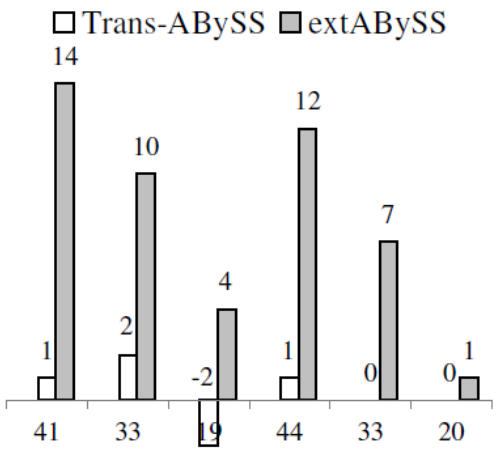


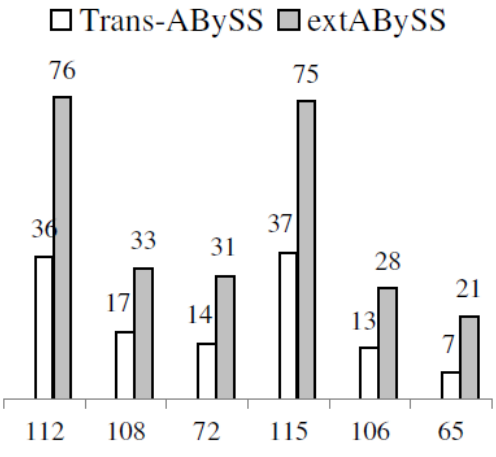
Figure 2- 9 Comparisons of the change in the number of false positive similar transcripts recovered by Oases and Trans-ABySS to the change in the ones recovered by extVelvet and extABySS respectively over the ones recovered by Velvet and ABySS respectively for different values of *k* and *k*-mer coverage cutoff *c*. Notations are the same as in Figure 2-3. A false positive similar transcript is recovered by each algorithm, but is not recovered by a simple protein BLAST search from each model organism to another related model organism with *e*-value cutoff 10^{-20}.



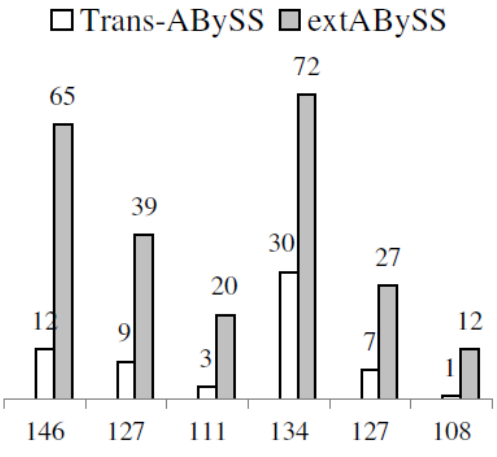
H. sap to *S. bol*



H. sap to *M. mus*



A. tha to *A. lyr*



A. tha to *O. sat*

Figure 2- 9 Continued.

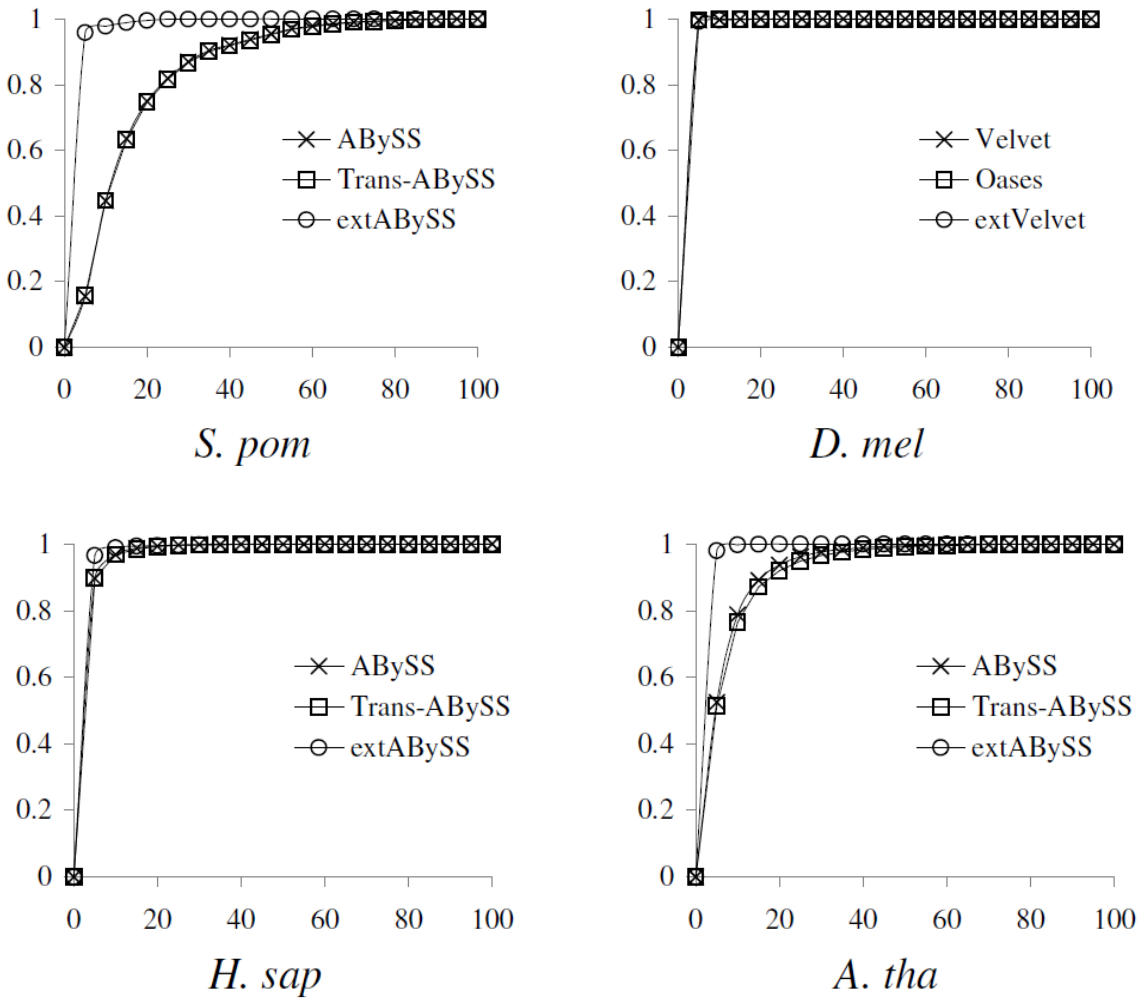


Figure 2- 10 Comparisons of the cumulative distribution of the expression estimates of similar transcripts that are 80% full length transcripts (100% full length transcripts for S. pombe) and recovered by Velvet, Oases and extVelvet (or by ABySS, Trans-ABySS and extABySS) divided into 20 quantiles in model organisms. Y-axis shows fraction of transcripts in different quantiles (5% increment) and x-axis shows expression quantiles. The least stringent values of *k* and *c* are used in each case, which is *k*=25/*c*=3 for *D. melanogaster* and *k*=25/*c*=10 for the other organisms.

Gene expression

I assess the ability of each algorithm to recover transcripts at different expression levels. For each model organism, I apply eXpress (52) to the reads in each data set with respect to the reference transcript database, and obtain expression estimates of similar transcripts that are close to full length and recovered by each algorithm. Figure 2-10 shows that extABySS is able to recover a higher proportion of full length transcripts with low coverage than ABySS and Trans-ABySS.

Melilotus albus and *Melilotus siculus*

In order to study salt and waterlogging tolerance of the two *Melilotus* species, I apply my algorithm starting from each species to the model organism *A. thaliana* and the non-model organism *Medicago truncatula*, which is not as well annotated but closer in evolutionary distance. I assess the differences between the two species by applying GO Term Finder (80) to the two sets of gene names in recovered similar transcripts from *M. albus* and *M. siculus* to *A. thaliana* and *M. truncatula* to identify significantly overrepresented GO terms with Bonferroni corrected p-value below 0.01 within the biological process ontology.

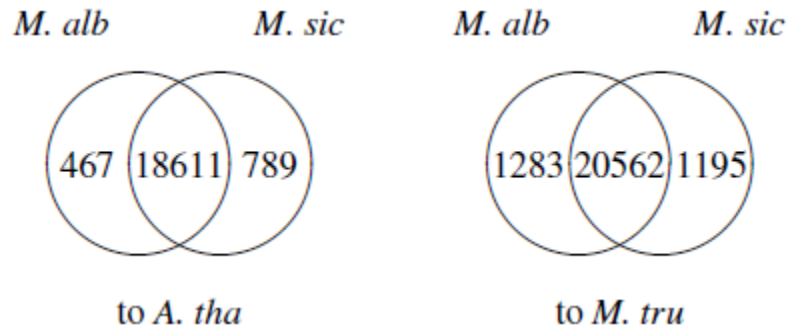


Figure 2- 11 Venn diagrams of the number of genes in recovered similar transcripts from *M. albus* and *M. siculus* to *A. thaliana* and *M. truncatula* in the *k*=25/*c*=3 assembly.

Figures 2-11 and 2-12 show that while a large number of genes in recovered similar transcripts and significantly overrepresented GO terms are shared by the two species, a small number of results that are unique to each species can be found.

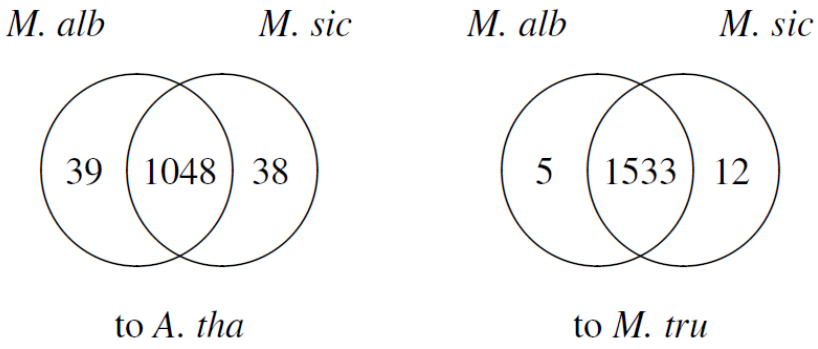


Figure 2- 12 Venn diagrams of the number of significantly overrepresented GO terms from *M. albus* and *M. siculus* to *A. thaliana* and *M. truncatula* in the *k*=25/*c*=3 assembly.

To assess differential gene expression under different conditions, I apply edgeR (81) in the Bioconductor package (82) on the expression estimates given by eXpress (52) to obtain a set of differentially expressed genes under one condition against another condition with q-value below 0.01, and apply GO Term Finder (80) to identify significantly overrepresented GO terms. Tables 2-2 and 2-3 show that differentially expressed genes can be identified under all conditions, with some of them associated with significantly overrepresented GO terms ($q<0.01$). These results provide further basis to study the genes that are responsible for differences in salt and waterlogging tolerance of the two species.

Table 2- 2 Differentially expressed genes recovered from *M. albus* and *M. siculus* to *A. thaliana* and *M. truncatula* from libraries associated with one condition versus another condition in the *k*=25/*c*=3 assembly, with organism indicating the starting organism and its related organism, SvsC indicating salt tolerance versus control, WvsC indicating waterlogging tolerance versus control, SWvsC indicating salt and waterlogging tolerance versus control, SWvsS indicating salt and waterlogging tolerance versus salt tolerance, and SWvsW indicating salt and waterlogging tolerance versus waterlogging tolerance.

Organism	SvsC	WvsC	SWvsC	SWvsS	SWvsW
M. alb to *A.tha*	8	141	81	47	12
M. sic to *A.tha*	39	7	10	45	8
M. alb to *M.tru*	11	220	114	86	17
M. sic to *M.tru*	74	24	31	84	12

Table 2- 3 Significantly overrepresented GO terms recovered from *M. albus* and *M. siculus* to *A. thaliana* and *M. truncatula* from libraries associated with one condition versus another condition in the *k*=25/*c*=3 assembly. Notations are the same as in Table 2-2.

Organism	SvsC	WvsC	SWvsC	SWvsS	SWvsW
M. alb to *A.tha*	0	23	42	7	0
M. sic to *A.tha*	9	0	0	2	0
M. alb to *M.tru*	2	0	1	0	0
M. sic to *M.tru*	0	0	0	0	0

Conclusions

Since the main memory requirement of my algorithm is for storing the de Bruijn graph and performing BLAST searches, my heuristic extension algorithms extVelvet and extABySS are much less memory intensive and more easily parallelizable than the base algorithms Velvet and ABySS (37). Iterative BLAST searches can be performed independently in parallel by assigning disjoint subsets of nodes to different processors for extension.

The running time of my algorithm has large dependence on the number of nodes that are chosen for extension (see Table 2-4). This in turn depends on the size of RNA-Seq data and the complexity of transcriptomes, which are reflected by the number of nodes in the de Bruijn graph and the number of transcripts in the database, and it also depends on the evolutionary distance between the starting organism and the related model organism. When applying to a different related organism, my running time in terms of processor-hours is at most a few to 10 times more than the base algorithm in almost all cases, and it can be much less in some cases.

Table 2- 4 Running time in processor-hours, with the values to the left and to the right of "+" indicating the running time of Velvet and Oases respectively (or ABySS and Trans-ABySS respectively), organism indicating the related model organism, time indicating the running time of extVelvet (or extABySS), chosen indicating the number of nodes that are chosen for extension, de Bruijn indicating the number of nodes in the de Bruijn graph, and database indicating the number of transcripts in the database.

least stringent k_c	organism	time	chosen	de Bruijn	database
S. pom (84+0.2)	*S. pom*	45	40786	536894	5011
	S.cer	12	15252	536894	5907
	N.cra	12	16366	536894	10082
D. mel (6.7+4.4)	*D. mel*	238	139248	466572	22102
	D. pse	67	63982	466572	16071
	A.gam	32	41578	466572	12659
H. sap (45+0.2)	*H.sap*	595	221942	1133348	32787
	S.bol	490	88342	1133348	25621
	M.mus	167	89070	1133348	29617
A. tha (112+0.2)	*A.tha*	2495	397638	3111862	41671
	A.lyr	944	218760	3111862	32549
	O.sat	616	143778	3111862	26777
L. ser (1.2+0.2)	*D. mel*	67	46760	392630	22102
T. yun (2.0+0.4)	*D. mel*	28	47638	330514	22102
H. gla (368+0.2)	*H.sap*	1920	203466	5457924	32799
C. soc (440+0.2)	*H.sap*	1344	175692	5030586	32799
C. ari (4.2+46)	*A.tha*	200	103652	1209068	41671
M. alb (5.8+2.9)	*A.tha*	79	82596	536210	41671

The situation is different in model organisms when similarity searches are performed to the organism itself. Since the BLAST hits are of much higher quality, path extensions can be very time consuming. In such cases, mapping-first algorithms such as Cufflinks (2) or Scripture (1) could be used instead, which often have better performance

since my need to impose a *k*-mer coverage cutoff to simplify the de Bruijn graph for heuristic extension often leads to missed transcripts.

My heuristic extension strategy cannot be applied to all transcriptome assembly algorithms. On algorithms such as Trinity (12) that first clusters the data and constructs a de Bruijn graph individually for each cluster, each of these individual graphs has simple structures. Performing heuristic extension on top of these graphs will not lead to significant improvements.

While my strategy cannot replace transcript predictions in *de novo* assemblies when the goal is to identify novel transcripts that have no similarity to other organisms, I have shown that my strategy can recover more and longer transcripts and can better resolve isoforms when similar transcripts are available from a related organism. By making use of evolutionary information, the sequence similarity support from the BLAST alignments ensures that the correspondences between the similar transcripts in the original organism and in the related organism are real.

CHAPTER III

HEURISTIC PAIRWISE ALIGNMENT OF DE BRUIJN GRAPHS TO FACILITATE SIMULTANEOUS TRANSCRIPT DISCOVERY IN RELATED ORGANISMS FROM RNA-SEQ DATA

There is often a need to investigate the transcriptomes of two related organisms at the same time in order to study their similarities and differences. In these cases, RNA-Seq libraries are obtained from both organisms under different experimental conditions and the goal is to compare their transcriptome assemblies. The traditional approach to address this problem is to perform transcriptome assemblies to obtain predicted transcripts for the two organisms separately (see Figure 3-1). Similarity comparison algorithms such as BLAST (83) are then employed to extract corresponding transcripts that are shared in the two organisms. Since predicted transcripts are constructed independently for each organism based on coverage information only, this strategy is often unreliable. To address this problem, I develop an algorithm to allow direct comparisons between paths in the two intermediate de Bruijn graph structures by an iterative extension strategy (see Figure 3-1). Since sequence similarity information is often more reliable, this strategy allows the direct extraction of shared transcripts based on evolutionary support.

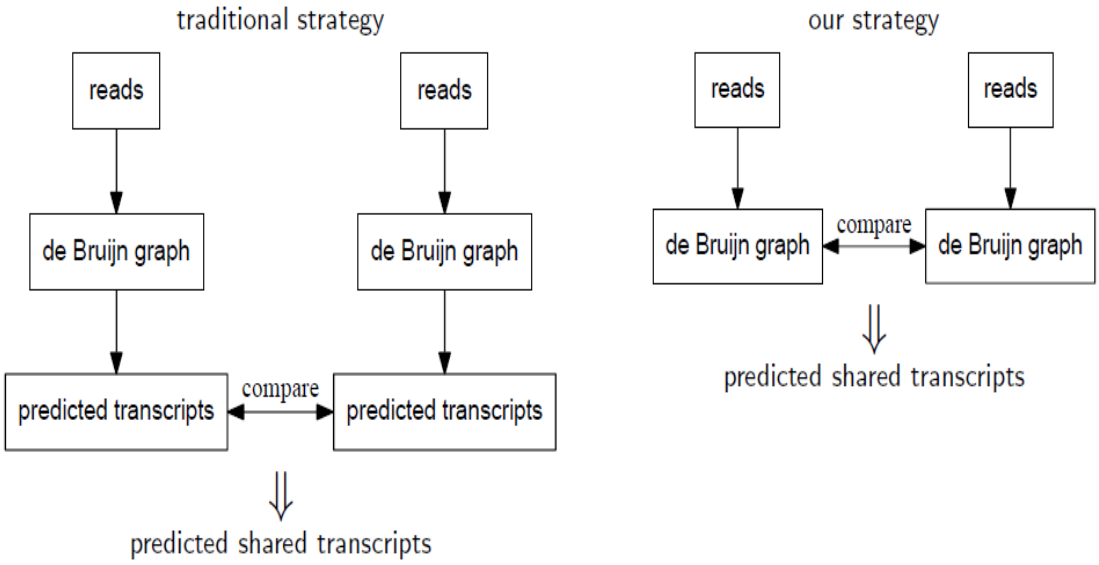


Figure 3- 1 Difference between traditional strategy and my strategy.

Methods

De Bruijn graph

Given a set of reads and a parameter *k*, a de Bruijn graph is constructed by taking each *k*-mer that appears within the reads as a vertex. Two *k*-mers are connected by a directed edge if the (*k*−1)-suffix of the first *k*-mer is the same as the (*k*−1)-prefix of the second *k*-mer (66,84). The de Bruijn graph implicitly assembles the reads by linking together the overlapping parts, and it is employed as the main intermediate structure by most short read assembly algorithms (27,54,55,57,58). To obtain a more compact structure, each linear sequence of vertices that have no branches is collapsed into a single node that corresponds to contigs.

Iterative extension

Given de Bruijn graphs G_1 and G_2 that correspond to transcriptome assemblies of two related organisms, I first apply BLAST to obtain similarity scores between each pair of nodes u from G_1 and v from G_2. I then start the iterative extension process as follows. For each node u from G_1, I extract its most similar node v from G_2 with e-value below a cutoff. If such a node v exists, I retain u as a single-node path. I extend u by one node along all its outgoing edges into multiple paths, and apply BLAST from each of these extended paths from u against v. If at least one of these extended paths gives a better e-value against v, I retain all the paths that have better e-values and continue to extend the top path that gives the best e-value. I repeat the procedure starting from this new path until the e-value no longer improves. Note that only one best direction is chosen since extending in more than one direction is very time-consuming. By starting from each node u in G_1 independently, the probability of missing the real best path is reduced a lot. After the above procedure, I have retained u and all the extended paths from u that have improved e-values, with the top path that gives the best e-value being fully extended. I then retain v as a single-node path and perform a similar extension process starting from v by extending it by one node along all its outgoing edges into multiple paths. I apply BLAST from each of these extended paths from v against all the retained paths from u. If at least one of these extended paths gives a better e-value, I retain all the paths that have better e-values and continue to extend the top path that gives the best e-value. Similar to above, I repeat the procedure starting from this new path until the e-value no longer

improves to obtain a fully extended path and a set of retained paths from *v* that have improved *e*-values (see Figure 3-2).

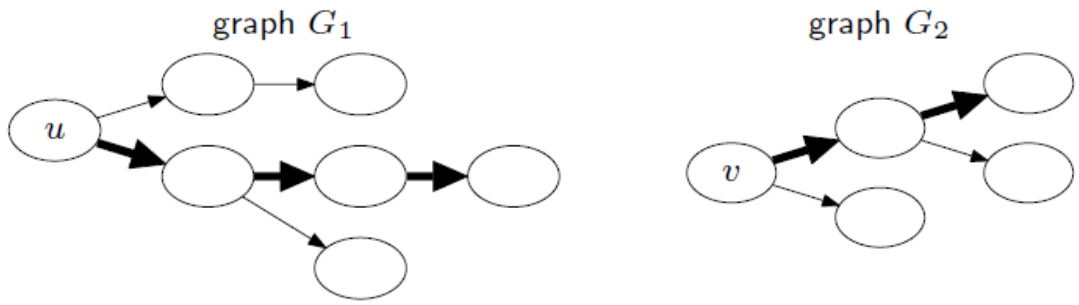


Figure 3- 2 Illustration of the iterative extension procedure. The paths that are fully extended from u in G1 and from v in G2 are marked in bold, while the other retained paths with improved *e*-value are not marked.

I then repeat the entire extension procedure in turn in G_1 and G_2 by replacing *u* by the fully extended path from *u* and comparing against all the retained paths from *v*, and replacing *v* by the fully extended path from *v* and comparing against all the retained paths from *u*. The entire process is repeated until no more improvements can be made, and the algorithm is applied again by switching the role of G_1 and G_2 and repeating all the steps. To obtain longer paths, I consider the retained paths from each node *u* and the retained paths from its twin node *u′*, in which *u′* represents the reverse complementary sequence of *u* on the opposite strand. I merge the twin paths that are complementary to the retained paths from *u′* with the retained paths from *u*, and keep those paths with improved *e*-values.

Extraction of predicted transcripts

I consider all the retained paths in G_1 as predicted transcripts in the first organism and all the retained paths in G_2 as predicted transcripts in the second organism. Since the collection of all these retained paths can be very big, I only keep a path if it contains a node in the de Bruijn graph that is not covered by another path with a better *e*-value according to the top BLAST alignment. In this condition, a node is covered by a path if it contains the node itself or its twin node. To avoid a large number of incorrectly predicted isoforms, I remove paths with worse *e*-values so that each node in the de Bruijn graph along with its twin node appears at most 10 times within the final set of paths.

Extraction of predicted shared transcripts

To obtain predicted shared transcripts that have correspondences between the two organisms, I apply BLAST from each predicted transcript in one organism against the set of all predicted transcripts in the other organism as database. I retain a predicted transcript as a predicted shared transcript if it appears both as a query with BLAST hits from one direction and as a subject BLAST hit in the other direction.

Results and discussion

Validation

I implement my algorithm Mutual as a postprocessing module of Velvet (27), which is a popular sequence assembly algorithm that returns a set of contigs along with the de

Bruijn graph. I compare my performance to Oases (61), which uses output from Velvet to construct predicted transcripts. I validate my algorithm by applying it to simultaneously recover transcripts in mouse against rat and in mouse against human from publicly available RNA-Seq libraries at the sequence read archive (85), including two libraries from mouse in (24) (SRX017794), one library from rat in (86) (SRX076903), and four libraries from human in (87) (SRX011545). I perform quality trimming by removing all positions including and to the right of the first position that has a quality score of less than 15, resulting in a size of 1.3 G for the mouse libraries, 2.5 G for the rat libraries and 1.1 G for the human libraries. I apply each algorithm over k=25 and k=31, and over k-mer coverage cutoff c=3, 5 and 10. In my algorithm Mutual, iterative extension is applied twice with an e-value cutoff of 0.1 using the bl2seq (BLAST 2 Sequences) variant of BLAST, once with translated BLAST and once with nucleotide BLAST. Velvet and Oases are applied independently in each organism. Since Oases applies the coverage cutoff itself to obtain a de Bruijn graph by modifying Velvet's original de Bruijn graph without coverage cutoff, Mutual is applied on the two de Bruijn graphs given by Oases to obtain predicted transcripts. To obtain predicted shared transcripts for both Oases and Mutual, I apply both translated BLAST and nucleotide BLAST with an e-value cutoff of 10^{-7} or 10^{-20} from each predicted transcript in one organism with the set of all predicted transcripts in the other organism as database. The predicted transcripts that appear both as a query with BLAST hits from one direction and as a subject BLAST hit in the other direction are retained as predicted shared transcripts. To evaluate the accuracy of the predicted shared transcripts, I apply

nucleotide BLAST to compare them against known mouse, rat or human transcriptome databases using the same *e*-value cutoff as the one used to obtain the transcripts, which is 10^{-7} or 10^{-20}. To assess the extent of translocated transcripts, I apply GMAP (88) to map the predicted shared transcripts to known mouse, rat or human genomes.

Table 3- 1 Comparisons of the number of predicted transcripts in the test on mouse against rat from Oases and from Mutual over different values of *k* and *k*-mer coverage cutoff *c*. Note that these numbers are not directly comparable between Oases and Mutual since the predicted transcripts from Mutual are obtained by extending similar paths that appear in the two organisms with an *e*-value cutoff of 0.1 from bl2seq, while the predicted transcripts from Oases are obtained independently in each organism without such constraints

		mouse		rat	
k	*c*	Oases	Mutual	Oases	Mutual
25	3	51218	40657	100317	56409
25	5	27873	18511	33396	22538
25	10	10557	6104	7669	5639
31	3	48841	29778	82090	38141
31	5	25947	14073	28047	15981
31	10	8224	3954	5145	3485

Predicted transcripts

Tables 3-1 and 3-2 show that Mutual constructed fewer predicted transcripts than Oases. Note that the predicted transcripts from Mutual are obtained by extending similar paths that appear in the two organisms through iterative BLAST, while the predicted transcripts from Oases are obtained independently in each organism. The similarity constraints in Mutual ensure that a predicted transcript in one organism has a similar

counterpart in the other organism, albeit with a loose *e*-value cutoff. The later reciprocal BLAST is needed to enforce more stringent *e*-value cutoffs. On the other hand, the predicted transcripts from Oases have no such constraints, and reciprocal BLAST is used to obtain shared transcripts.

Predicted shared transcripts

When compared to Tables 3-1 and 3-2, Tables 3-3 and 3-4 show that only a small percentage of predicted transcripts were shared in the two organisms, with a smaller decrease by Mutual than by Oases. The decrease by Mutual is due to more stringent *e*-value cutoffs, while the decrease by Oases is due to imposing similarity constraints between the two organisms. While the actual amount of predicted shared transcripts that can be recovered depends on the size of libraries, the evolutionary distance between the two organisms and the experimental conditions, Tables 3-3 and 3-4 show that Mutual recovered more predicted shared transcripts than Oases. Almost all these predicted shared transcripts are found in the corresponding known transcriptome database, with comparable percentages between Mutual and Oases. The percentages are lower for rat, probably due to the fact that the rat genome is less well annotated. The number of predicted shared transcripts decreases as the assembly parameters become more stringent, but these transcripts are of higher quality.

Table 3- 2 Comparisons of the number of predicted transcripts in the test on mouse against human. Notations are the same as in Table 3-1.

		mouse		human	
k	*c*	Oases	Mutual	Oases	Mutual
25	3	51218	34514	100317	36268
25	5	27873	18561	33396	17519
25	10	10557	7020	7669	5405
31	3	48841	23510	82090	23263
31	5	25947	13433	28047	12867
31	10	8224	4358	5145	3182

Table 3- 3 Comparisons of the number of predicted shared transcripts (shared) and the number of predicted shared transcripts that have BLAST hits from each organism to its known transcriptome database (found) in the test on mouse against rat from Oases and from Mutual over different values of *k* and *k*-mer coverage cutoff *c* and over different *e*-value cutoffs 10^{-7} and 10^{-20}. The number in parentheses is the percentage of predicted shared transcripts that have BLAST hits from each organism to its known transcriptome database

	mouse (10^{-7})				rat (10^{-7})			
k_c	Oases		Mutual		Oases		Mutual	
	shared	found	shared	found	shared	found	shared	found
25_3	27671	26756 (97%)	35230	34011 (97%)	24489	21844 (89%)	39287	34298 (87%)
25_5	12729	12366 (97%)	14924	14520 (97%)	10092	9245 (92%)	15287	13639 (89%)
25_10	3955	3823 (97%)	4589	4465 (97%)	2994	2835 (95%)	3955	3705 (94%)
31_3	22635	22046 (97%)	25035	24396 (97%)	20917	19008 (91%)	27484	24744 (90%)
31_5	10229	10028 (98%)	11039	10825 (97%)	8398	7815 (93%)	11225	10332 (92%)
31_10	2597	2545 (98%)	2871	2815 (98%)	2013	1939 (96%)	2489	2382 (96%)

Table 3- 3 Continued

	mouse (10^{-20})				rat (10^{-20})			
k_c	Oases		Mutual		Oases		Mutual	
	shared	found	shared	found	shared	found	shared	found
25_3	22936	22290 (97%)	28705	27881 (97%)	19282	17719 (92%)	29923	26898 (90%)
25_5	10904	10608 (97%)	12648	12336 (98%)	8242	7669 (93%)	12087	10999 (91%)
25_10	3077	3253 (96%)	3901	3790 (97%)	2510	2388 (95%)	3254	3070 (94%)
31_3	18052	17627 (98%)	20026	19567 (98%)	15835	14699 (93%)	20943	19264 (92%)
31_5	8429	8261 (98%)	9157	8964 (98%)	6623	6623 (94%)	8886	8251 (93%)
31_10	2196	2150 (98%)	2438	2386 (98%)	1681	1681 (97%)	2041	1959 (96%)

Table 3- 4 Comparisons of the number of predicted shared transcripts and the number of predicted shared transcripts that have BLAST hits from each organism to its known transcriptome database in the test on mouse against human. Notations are the same as in Table 3-3

	mouse (10^{-7})				human (10^{-7})			
k_c	Oases		Mutual		Oases		Mutual	
	shared	found	shared	found	shared	found	shared	found
25_3	20763	20406 (98%)	25630	25189 (98%)	22499	22084 (98%)	28364	27911 (98%)
25_5	11914	11685 (98%)	12956	12784 (99%)	12037	11786 (98%)	12806	12643 (99%)
25_10	4644	4520 (97%)	5226	5114 (98%)	3844	3762 (98%)	4121	4047 (98%)
31_3	14631	14440 (99%)	16226	16041 (99%)	16498	16348 (99%)	18482	18482 (99%)
31_5	8351	8241 (99%)	8920	8825 (99%)	9250	9171 (99%)	9841	9753 (99%)
31_10	2727	2686 (98%)	2924	2887 (99%)	2326	2308 (99%)	2438	2420 (99%)

Table 3- 4 Continued.

	mouse (10^{-20})				human (10^{-20})			
k_c	Oases		Mutual		Oases		Mutual	
	shared	found	shared	found	shared	found	shared	found
25_3	15532	15335 (99%)	18418	18165 (99%)	17104	16799 (99%)	19840	19558 (99%)
25_5	9534	9356 (98%)	10249	10137 (99%)	9718	9541 (98%)	10120	10000 (99%)
25_10	3965	3854 (97%)	4452	4358 (98%)	3344	3278 (98%)	3593	3529 (98%)
31_3	10165	10045 (99%)	11250	11127 (99%)	12052	11960 (99%)	13138	13043 (99%)
31_5	6262	6183 (99%)	6728	6654 (99%)	7267	7216 (99%)	7615	7557 (99%)
31_10	2245	2209 (98%)	2419	2385 (99%)	2003	1989 (99%)	2083	2069 (99%)

Top BLAST hits to databases

By applying BLAST from each set of predicted shared transcripts in each organism to its known transcriptome database, Tables 3-5 and 3-6 show that Mutual recovered more shared transcripts than Oases, with many more shared transcripts recovered when the assembly parameters are less stringent.

Length distribution of transcripts

Figures 3-3 and 3-4 show that the lengths of predicted shared transcripts recovered by Mutual were comparable to the ones recovered by Oases, which are slightly shorter for mouse but have slightly higher medians for rat. These transcripts are generally longer when the *k*-mer coverage cutoff *c* increases.

Table 3- 5 Comparisons of the number of top unique BLAST hits to different transcripts from each set of predicted shared transcripts in each organism to its known transcriptome database in the test on mouse against rat from Oases and from Mutual over different values of *k* and *k*-mer coverage cutoff *c* and over different *e*-value cutoffs 10^{-7} and 10^{-20}. Only the top hit with *e*-value below the cutoff is considered. The number in parentheses is the change by Mutual over Oases.

10^{-7}	mouse		rat		10^{-20}	mouse		rat	
k_c	Oases	Mutual	Oases	Mutual	*k_c*	Oases	Mutual	Oases	Mutual
25_3	7780	8349 (+569)	7382	8061 (+679)	25_3	7035	7547 (+512)	6608	7148 (+540)
25_5	5310	5563 (+253)	4863	5158 (+295)	25_5	4715	4929 (+214)	4319	4538 (+219)
25_10	2361	243 (+102)	2011	2094 (+83)	25_10	2008	2094 (+86)	1769	1833 (+64)
31_3	6645	6854 (+209)	6392	6660 (+268)	31_3	5780	5997 (+217)	5527	5802 (+275)
31_5	4286	4368 (+82)	3933	4103 (+170)	31_5	3713	3804 (+91)	3454	3557 (+103)
31_10	1705	1740(+35)	1462	1517 (+55)	31_10	1443	1484 (+41)	1287	1320 (+33)

Table 3- 6 Comparisons of the number of top unique BLAST hits to different transcripts from each set of predicted shared transcripts in each organism to its known transcriptome database in the test on mouse against human. Notations are the same as in Table 3-5

10^{-7}	mouse		human		10^{-20}	mouse		human	
k_c	Oases	Mutual	Oases	Mutual	*k_c*	Oases	Mutual	Oases	Mutual
25_3	7090	7474 (+384)	7123	7548 (+425)	25_3	6169	6402 (+233)	6317	6539 (+222)
25_5	5308	5392 (+84)	5244	5318 (+74)	25_5	4666	4700 (+34)	4679	4696 (+17)
25_10	2781	2818 (+37)	2591	2612 (+21)	25_10	2452	2476 (+24)	2376	2385 (+9)
31_3	5490	5647 (+157)	5198	5387 (+189)	31_3	4421	4557 (+136)	4416	4547 (+131)
31_5	3918	3971 (+53)	3662	3732 (+70)	31_5	3221	3275 (+54)	3180	3222 (+42)
31_10	1796	1805 (+9)	1573	1594 (+21)	31_10	1531	1540 (+9)	1403	1410 (+7)

Table 3- 7 Comparisons of the number of predicted shared transcripts that are 80% full length transcripts in the test on mouse against rat from Oases and from Mutual over different values of *k* and *k*-mer coverage cutoff *c* and over different *e*-value cutoffs 10^{-7} and 10^{-20}. These transcripts are the ones in which 80% of the coding region is included in the best BLAST alignment from each organism to its known transcriptome database. The number in parentheses is the change by Mutual over Oases.

10^{-7}	mouse		rat		10^{-20}	mouse		rat	
k_c	Oases	Mutual	Oases	Mutual	*k_c*	Oases	Mutual	Oases	Mutual
25_3	1900	1840 (-60)	2066	1777 (-289)	25_3	1802	1743 (-59)	1870	1611 (-259)
25_5	1705	1677 (-28)	1739	1581 (-158)	25_5	1595	1561 (-34)	1577	1429 (-148)
25_10	1119	1097 (-22)	862	848 (-14)	25_10	984	975 (-9)	798	788 (-10)
31_3	1144	1158 (+14)	1407	1179 (-228)	31_3	1061	1077 (+16)	1226	1042 (-184)
31_5	1054	1062 (+8)	1240	1095 (-145)	31_5	966	990 (+24)	1092	978 (-114)
31_10	719	724 (+5)	662	662 (0)	31_10	638	646 (+8)	607	602 (-5)

Table 3- 8 Comparisons of the number of predicted shared transcripts that are 80% full length transcripts in the test on mouse against human. Notations are the same as in Table 3-7.

10^{-7}	mouse		rat		10^{-20}	mouse		rat	
k_c	Oases	Mutual	Oases	Mutual	*k_c*	Oases	Mutual	Oases	Mutual
25_3	1851	1808 (-43)	1529	1553 (+24)	25_3	1733	1686 (-47)	1450	1477 (+27)
25_5	1716	1666 (-50)	1534	1536 (+2)	25_5	1605	1552 (-53)	1454	1459 (+5)
25_10	1250	1241 (-9)	1178	1183 (+5)	25_10	1124	1112 (-12)	1114	1126 (+12)
31_3	1085	1099 (+14)	739	746 (+7)	31_3	995	1008 (+13)	686	700 (+14)
31_5	1009	1018 (+9)	734	736 (+2)	31_5	923	932 (+9)	678	683 (+5)
31_10	720	723 (+3)	627	628 (+1)	31_10	654	656 (+2)	579	585 (+6)

Recovery of full length transcripts

The situation is different when considering predicted shared transcripts that are close to full length. Tables 3-7 and 3-8 show that Mutual recovered more or a

comparable number of 80% full length transcripts as Oases when the assembly parameters are more stringent, and less 80% full length transcripts than Oases when the assembly parameters are less stringent. Although Mutual performs worse for rat that recovers less 80% full length transcripts than Oases, its predicted shared transcripts have slightly higher median lengths when considering all the transcripts together (see Figure 3-3), instead of just the ones that are 80% full length transcripts.

Presence of translocated transcripts

Reconstructed transcripts covering fragments from different chromosomes or different loci far away on the same chromosomes may be considered translocated transcripts and identified as assembly errors because they are rare. As reported by GMAP, Tables 3-9 and 3-10 show that Mutual recovered a much larger number of predicted shared transcripts that are uniquely mapped than Oases, while at the same time returning more translocated transcripts that can be considered to be errors due to their rare occurrences (89). The ratio of the number of translocated transcripts to the number of uniquely mapped transcripts is at most about twice as much for Mutual when compared to Oases. This ratio increases when *k* decreases or when the *k*-mer coverage cutoff *c* increases.

Table 3- 9 Comparisons of the number of predicted shared transcripts that are uniquely mapped (unique) or translocated (transloc) as reported by GMAP in the test on mouse against rat from Oases and from Mutual over different values of *k* and *k*-mer coverage cutoff *c* and over different *e*-value cutoffs 10^{-7} and 10^{-20}. The number in parentheses is the ratio of the number of translocated transcripts to the number of uniquely mapped transcripts.

	mouse(10^{-7})				rat(10^{-7})			
	Oases		Mutual		Oases		Mutual	
k_c	unique	transloc	unique	transloc	unique	transloc	unique	transloc
25_3	24635	599 (0.024)	30713	1475 (0.048)	21335	986 (0.046)	33566	2337 (0.067)
25_5	10718	436 (0.041)	12071	1011 (0.084)	8509	438 (0.051)	12676	971 (0.077)
25_10	2913	218 (0.075)	3197	409 (0.128)	2353	122 (0.052)	3042	257 (0.084)
31_3	20360	242 (0.012)	22229	483 (0.022)	18236	497 (0.027)	23818	795 (0.033)
31_5	8778	189 (0.022)	9263	388 (0.042)	7132	251 (0.035)	9453	388 (0.041)
31_10	1914	99 (0.052)	2026	176 (0.087)	1553	65 (0.042)	1888	113 (0.060)

	mouse(10^{-20})				rat(10^{-20})			
	Oases		Mutual		Oases		Mutual	
k_c	unique	transloc	unique	Transloc	unique	transloc	unique	transloc
25_3	20209	544 (0.027)	24662	1368 (0.055)	16880	746 (0.044)	25851	1536 (0.059)
25_5	9070	396 (0.044)	10067	931 (0.092)	7021	332 (0.047)	10097	718 (0.071)
25_10	2431	188 (0.077)	2631	372 (0.141)	1977	98 (0.050)	2499	214 (0.086)
31_3	16077	213 (0.013)	17610	415 (0.024)	13866	376 (0.027)	18290	516 (0.028)
31_5	7136	156 (0.022)	7538	347 (0.046)	5656	177 (0.031)	7572	243 (0.032)
31_10	1590	85 (0.053)	1701	146 (0.086)	1299	51 (0.039)	1559	83 (0.053)

Table 3- 10 Comparisons of the number of predicted shared transcripts that are uniquely mapped or translocated as reported by GMAP in the test on mouse against human. Notations are the same as in Table 3- 9.

	mouse(10^{-7})				human(10^{-7})			
	Oases		Mutual		Oases		Mutual	
k_c	unique	transloc	unique	transloc	unique	transloc	unique	transloc
25_3	18157	531 (0.029)	21931	1209 (0.055)	19912	224 (0.011)	25142	592 (0.024)
25_5	10036	393 (0.039)	10760	763 (0.071)	10353	150 (0.014)	11088	334 (0.030)
25_10	3582	203 (0.057)	3838	420 (0.109)	3114	78 (0.025)	3281	221 (0.067)
31_3	12899	196 (0.015)	14105	370 (0.026)	14748	65 (0.004)	16499	126 (0.008)
31_5	7084	147 (0.021)	7392	302 (0.041)	8101	43 (0.005)	8536	94 (0.011)
31_10	2029	93 (0.046)	2095	167 (0.080)	1858	30 (0.016)	1919	58 (0.030)

	mouse(10^{-20})				human(10^{-20})			
	Oases		Mutual		Oases		Mutual	
k_c	unique	transloc	unique	transloc	unique	transloc	unique	transloc
25_3	13313	499 (0.037)	15285	1073 (0.070)	14928	195 (0.013)	17259	518 (0.030)
25_5	7877	373 (0.047)	8286	713 (0.086)	8301	130 (0.016)	8638	315 (0.036)
25_10	2980	188 (0.063)	3152	400 (0.127)	2699	73 (0.027)	2822	211 (0.075)
31_3	8736	181 (0.021)	9504	330 (0.035)	10690	57 (0.005)	11621	106 (0.009)
31_5	5183	137 (0.026)	5408	281 (0.052)	6325	35 (0.006)	6580	84 (0.013)
31_10	1618	91 (0.056)	1671	161 (0.096)	1591	19 (0.012)	1623	55 (0.034)

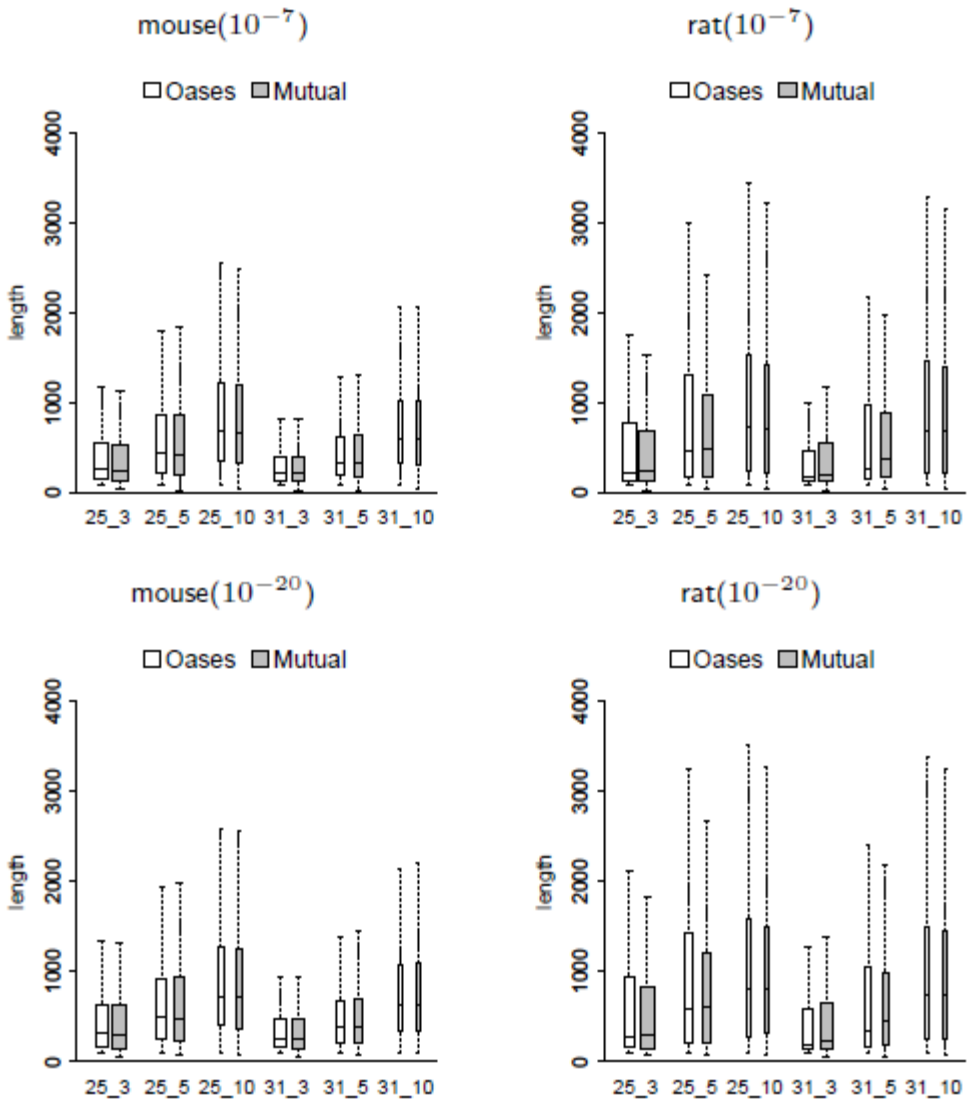


Figure 3- 3 Length distribution of predicted shared transcripts in the test on mouse against rat from Oases and from Mutual over different values of *k* and *k*-mer coverage cutoff *c* (represented by *k_c*) and over different *e*-value cutoffs 10^{-7} and 10^{-20}. The width of each box is proportional to the square root of the size of each group, while outliers are ignored

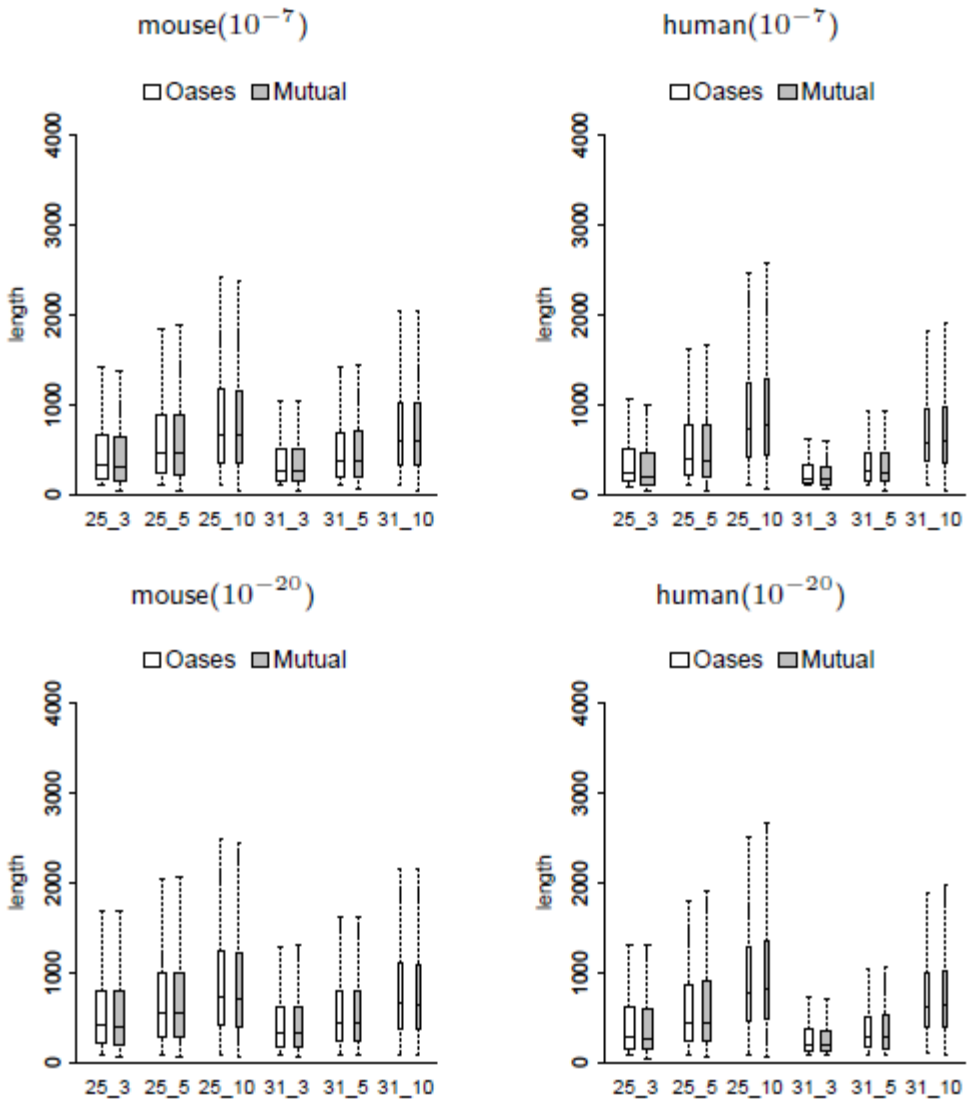


Figure 3- 4 Length distribution of predicted shared transcripts in the test on mouse against human. Notations are the same as in Figure 3-3.

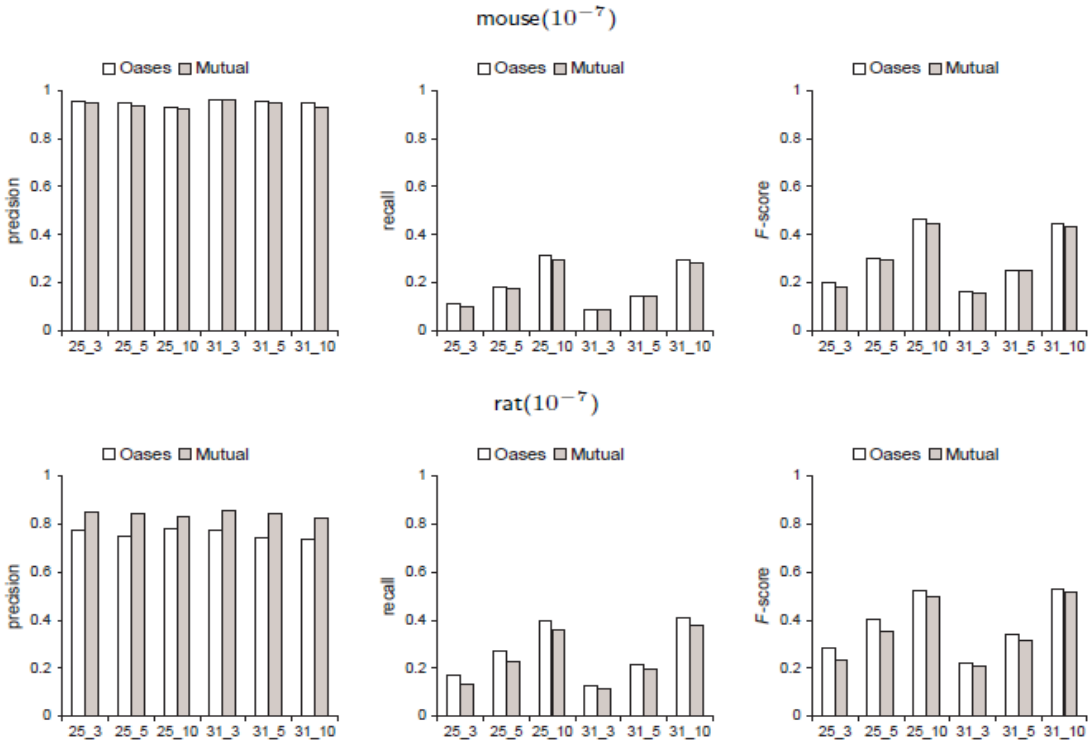


Figure 3- 5 Precision, recall and F-score with respect to the accuracy of shared transcript reconstruction in the test on mouse against rat from Oases and from Mutual over different values of *k* and *k*-mer coverage cutoff *c* (represented by k_c) and over different *e*-value cutoffs 10^{-7} and 10^{-20}. Precision is defined to be the fraction of query positions from predicted shared transcripts that are included in BLAST alignments from each organism to its known transcriptome database. Recall is defined to be the fraction of subject positions from database sequences that are included in BLAST alignments from each organism to its known transcriptome database. *F*-score is the harmonic mean of precision and recall.

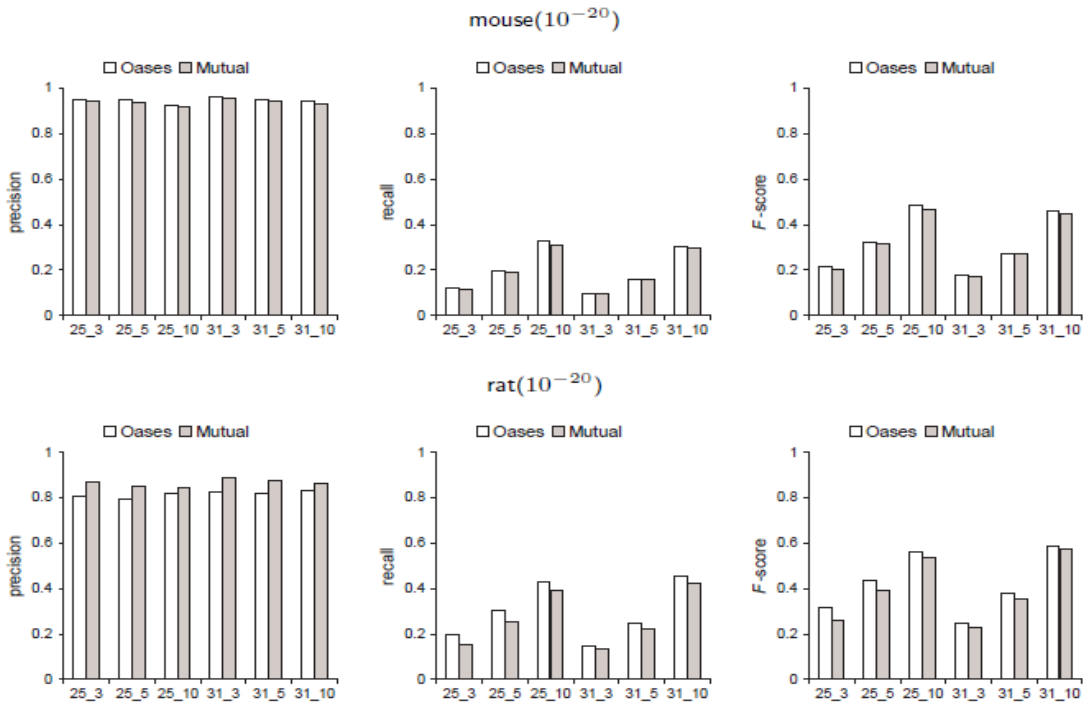


Figure 3- 5 Continued.

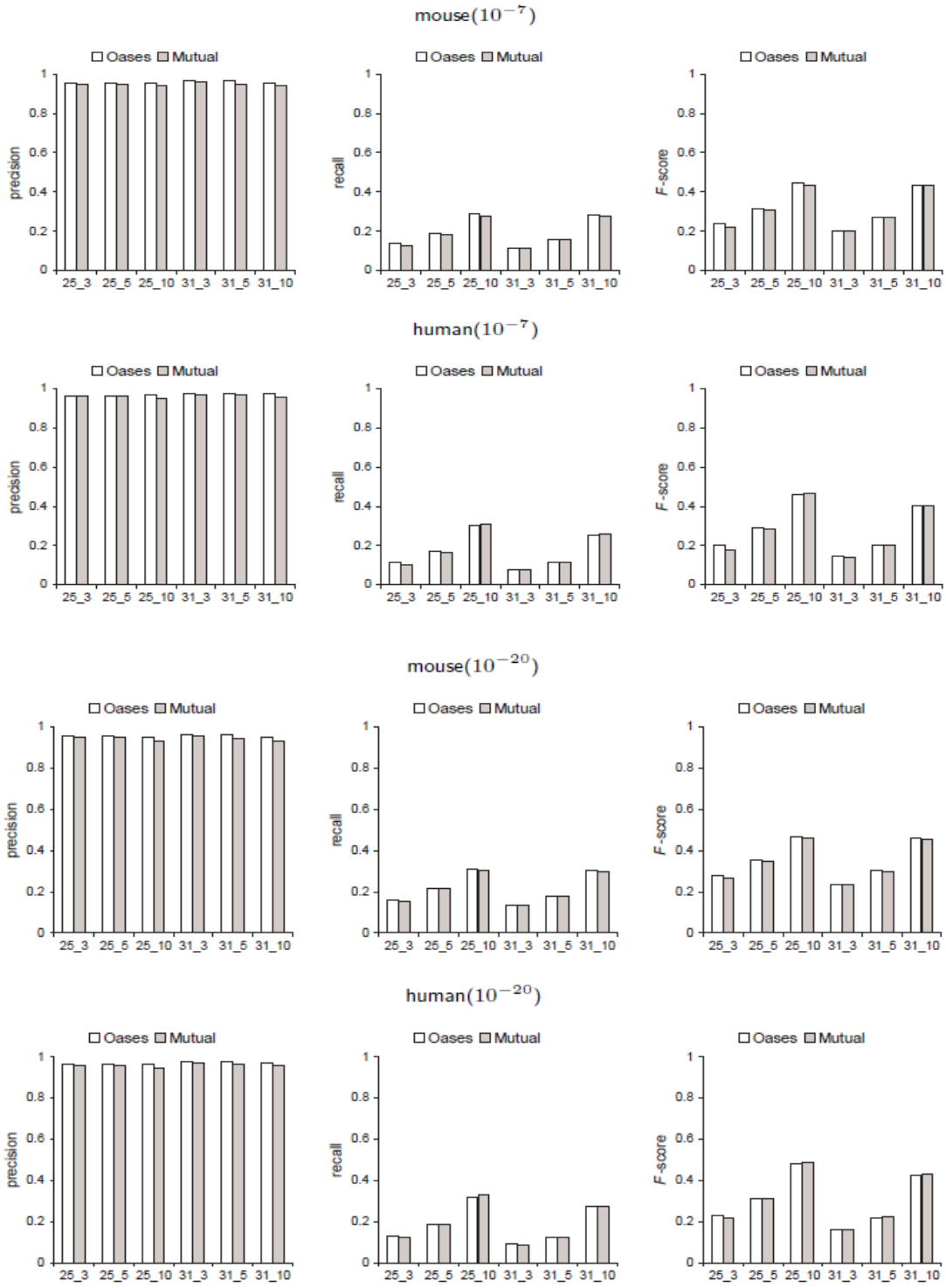


Figure 3- 6 Precision, recall and F-score with respect to the accuracy of shared transcript reconstruction in the test on mouse against human. Notations are the same as in Figure 3- 5.

Accuracy of transcript reconstruction

By investigating the fitness of the alignment between the predicted shared transcripts and the known transcriptome database sequences, Figures 3-5 and 3-6 show that with respect to the accuracy of shared transcript reconstruction, there are tradeoffs between precision and recall by Mutual when compared to Oases. Mutual has slightly lower *F*-scores than Oases in most cases.

Conclusions

I have developed an algorithm that makes use of evolutionary information to simultaneously recover significantly more shared transcripts from RNA-Seq data in two related organisms that may be missed by traditional *de novo* approaches. While more shared transcripts are recovered due to the smaller evolutionary distance between mouse and rat, my algorithm can be applied to related organisms that are evolutionarily farther away, such as between mouse and human. While known transcriptomes are used as databases during validation, one important characteristic of my algorithm is that no reference transcriptomes or a closely related model organism is needed. My algorithm can be used to recover shared transcripts that are specific to two closely related non-model organisms, which may not be present in a related model organism that is evolutionarily farther away. Depending on the size of the de Bruijn graphs, my algorithm can take many processor-hours to run. It takes more than 600 processor-hours to obtain all the predicted transcripts in mouse against rat or in mouse against human for the least stringent values of *k* and the *k*-mer coverage cutoff *c*. Although my algorithm can take

much more computational time than the *de novo* algorithms Velvet or Oases, the iterative BLAST searches can be run independently in parallel on a computing cluster. While an additional 60 processor-hours are needed to obtain predicted shared transcripts from the predicted transcripts, a similar procedure is also needed for Oases. No special memory requirement is needed after the de Bruijn graphs are obtained.

One drawback of my algorithm is that only a weak recovery of non-coding regions of mRNA is possible since these regions may not be conserved. Due to the use of similarity information between two related organisms to extend transcripts, my algorithm cannot identify extended transcripts that are not shared between the two organisms.

CHAPTER IV

GENOME ANNOTATIONS OF *PROTEUS MIRABILIS* AND *PROVIDENCIA STUARTII*

Many bacteria are widespread and exhibit various interactions with insects. For example, insects can serve as important vectors. Some of these bacteria exhibit wide host ranges, while some have preferred host targets. For example, *Bacillus sphaericus* is selectively pathogenic to mosquitoes while *Bacillus papillae* only infects Scarabaeid beetles (90). Understanding how such bacteria interact with insects and other eukaryotes is currently an area of intense interest in biology.

The interaction between insects and bacteria can be considered symbiotic. A symbiotic relationship is an interaction between different species. It can be divided into parasitism, commensalism and mutualism. Parasitism is a relationship when one species benefits from association while its partner gets harmed. Commensalism occurs when one species get benefit from interactions with no significant effect on its partner. Mutualism is the association where species living together mutually benefit each other (91). For example, pea aphid *Acyrthosiphon pisum* is able to diet on plant phloem sap with low content of essential amino acids, in presence of their symbiotic bacteria, *Buchnera aphidicola* by providing these nutrients (92). A virulent pathogen *Photorhabdus* produce an antibiotic (E)-1,3-dihydroxy-2-(isopropyl)-5- (2-phenylethenyl)benzene to inhibit phenoloxidase in the insect *Manduca sexta* to suppress host defenses (93).

There are also studies on symbiotic relationships between phages and their bacterial hosts. Phages rely on their bacterial hosts to complete their life cycle. Lytic phages introduce their genetic material into host cells and utilize host cell machinery for replication of viral genetic material and production of viral proteins. Then viral proteins self-assemble to package viral genetic material into capsids. When a sufficient number of virions are produced, host cells are lysed with lytic bacteriophage enzymes and release progeny viruses. The new cycle starts when released phages contact with new host cells (94). For example, bacteriophage BRK20 is a lytic phage of *Brevibacterium flavum*, an industrial producer of lysine (95). Unlike lytic phages, lysogenic phages can integrate their genetic material into host genome which can be transmitted to host progenies without killing host cells. Lysogenic phages can enter lytic cycle under stressful conditions (94). Lamboid phage Gifsy-1 is an example of a lysogenic phage that integrates its genetic material into its host *Salmonella enterica* serovar Typhimurium (96). Phages not only make use of bacterial hosts for their life cycle, but also have impacts on their abundance, competitive ability, changes in physiology, as well as gene transfer (94,97).

Proteus mirabilis, a Gram-negative bacterium, is an endosymbiont of blow flies. It produces antibacterial agents including phenylacetaldehyde (PAL) and phenylacetic acid (PAA) (98), which are supposed to benefit insects by controlling external and internal microbe community and repressing growth of bacteria which compete with the larvae (99). On the other hand, maggot excretions/secretions of insects inhibit biofilm formation of would pathogens such as *Staphylococcus aureus* and *Enterobacter cloacae*,

but protect biofilms of *P. mirabilis* (100). Biofilms help bacteria resist antimicrobial agents (101).

P. mirabilis has been isolated from salivary glands of the *Lucilia sericata*, a blow fly used in maggot therapy due to antibacterial, antibiofilm, and wound debridement properties. Some of these antibacterial agents are synthesized by *P. mirabilis* (41). An earlier study showed that swarming *P. mirabilis* produces small molecules to attract *L. sericata* to lay eggs and twenty genes associated with the swarming phenotype have been identified in *P. mirabilis*. Several swarming mutations can be complemented with fly attractants like ammonia and putrescine and one mutant has been shown to differentially impact fly oviposition and attraction (35).

Providencia stuartii has also been found to colonize maggots of *L. sericata* along with *P. mirabilis* (35). *P. stuartii* is a Gram-negative pathogen giving rise to human infections like meningitis (102) and causing blockages of urinary catheters (48). It does not swarm like *P. mirabilis* (48). A recent study found that *P. stuartii* shares a common cell-to-cell communication system with *D. melanogaster*. An inner membrane protein AarA in *P. stuartii* is required for exporting extracellular signals. AarA is homologous to rhomboid protein (RHO) in *D. melanogaster*, a serine protease required for stimulation of epidermal growth factor receptor ligands (103) and plays an important role in many developmental processes such as organization of the fly eye and proper wing vein development (37). *D. melanogaster rho* mutant can be complemented with expression of AarA from *P. stuartii* to exhibit normal wing vein development and *P. stuartii aarA*

mutant can be rescued with expression of RHO from *D. melanogaster* to overcome cell-to-cell communication defects (37).

Interkingdom communication between bacteria and their hosts involves hormones and hormone-like compounds, which may help bacteria to recognize the immune system on their hosts and activate their virulence genes (104). For example, binding of an outermembrane protein OprF in *Pseudomonas aeruginosa* to their host human interferon-γ leading to expression of both type I *P. aeruginosa* lectin (PA-I) and pyocyanin, which enables disruption of epithelial cell function (105).

In the present work, I have assembled, annotated and compared the draft genomes of *P. mirabilis* and *P. stuartii* isolated from larvae of *L. sericata* to reference genomes of clinical strains to identify unique genes which are absent in the reference genomes. I annotated gene content which probably contributes to physiological differences between *P. mirabilis* and *P. stuartii* isolated from flies and genes with evidence of recombination or positive selection among *Proteus* or *Providencia* tested. I also identified insertion sequences from other strains into these draft genomes to hypothesize the novel phenotypes studied strains may show.

Methods

Genome assembly

Sequencing was performed using Ion Torrent after preparation with NEBNext® Fast DNA Fragmentation & Library Prep Set to produce approximately $1.38x10^6$ reads of an average length of 211bp with 23x coverage of the genome for *Providencia stuartii* and

$2.77x10^6$ reads of an average length of 218 bp with 41x coverage for *P. mirabilis*. A total of ninety-seven and seventy-five contigs were assembled by CLC *de novo* assembly workbench for *P. stuartii* and *P. mirabilis* respectively. Scaffolds were assembled with CONTIGuator (106) from contigs with well annotated reference genomes *P. stuartii* MRSN 2154 for *P. stuartii* and *P. mirabilis* HI4320 and BB2000 for *P. mirabilis*.

Gene annotation method

Coding sequences (CDS) were predicted from scaffolds and unassembled contigs with PRODIGAL (107). Predicted gene sequences were aligned to genes from related reference genomes including *Providencia stuartii* MRSN 2154 (accession no NC_017731), *Proteus mirabilis* HI4320 (NC_010554 and NC_010555) and *Proteus mirabilis* BB2000 (accession no NC_022000) at *e*-value cutoff of 1e-20. For those unaligned genes, the NCBI non-redundant (nr) database was used to infer their potential functions at an *e*-value cutoff of 1e-5.

Identification of orthologs

Orthologous genes were identified as shared among *Proteus* and *Providencia* of my bacteria strains and strains from NCBI including my *P. mirabilis* draft genomes, *P. stuartii* draft genome, *P. stuartii* MRSN 2154, *P. mirabilis* HI4320, *P. mirabilis* BB2000, *Proteus hauseri* ZMd44, *P. mirabilis* ATCC 29906, *P. mirabilis* WGLW4, *P. mirabilis* WGLW6, *Proteus penneri* ATCC 35198, *Providencia alcalifaciens* 205/92, *P. alcalifaciens* DSM 30120, *P. alcalifaciens* Dmel2, *P. alcalifaciens* F90-2004, *P.*

alcalifaciens PAL-1, *P. alcalifaciens* PAL-2, *P. alcalifaciens* PAL-3, *P. alcalifaciens* R90-1475, *P. alcalifaciens* RIMD 1656011, *Providencia burhodogranariea* DSM 19968, *Providencia rettgeri* DSM 1131, *P. rettgeri* Dmel1, *Providencia rustigianii* DSM 4541, *Providencia sneebia* DSM 19967, *P. stuartii* ATCC 25827 plus the outgroup *Escherichia coli* str. K-12 substr. MG1655. Ortholog identification was performed with PanOCT (108).

Alignment of orthologs

Protein alignment of ortholog clusters identified above was produced with ClustalO (109) for phylogentic analysis, recombination analysis and positive selection analysis. Codon alignment from protein alignment in positive selection analysis was done with PAL2NAL (110).

Phylogenetic analysis

Phylogenetic analysis helps to understand evolutionary relationships among different species. The alignment of all 1322 orthologous clusters that cross all species tested were concatenated and used as input in RAxML (111) with 100 boostraps and with *Escherichia coli* str. K-12 substr. MG1655 set as an outgroup.

SNP and indel identification

Single nucleotide polymorphism (SNP) is variation of a single base in DNA (112). Indel is insertion and deletion of DNA sequences (113). Small genetic variations in

DNA sequences help to reveal evolutionary adaption and develop personal medicines (114). SNP and indels were identified between *P. stuartii* draft scaffold to *P. stuartii* MRSN 2154, *P. mirabilis* draft scaffold to *P. mirabilis* HI4320 or *P. mirabilis* BB2000 with MUMMer (115). Densities of SNP and indel across assembled scaffolds were shown with Circos with window size of 5000nt.

COG annotation

COG annotation is a genome-wide tool to predict protein functions and evolution (116). Predicted CDSs were aligned to NCBI database of Clusters of Orthologous Groups (COG) of proteins(117) collection (118) with BLAST at *e*-value cutoff of 1e-5 to detect enrichment in COG families.

Insertion sequence (IS) analysis

Insertion sequences are mobile genetic elements introduced to host genomes, which may involve gene exchange and reassortment (119). The genome sequence including scaffold and unaligned contigs were aligned to the IS finder database of a collection of bacterial insertion sequences (119) with BLASTx at *e*-value cutoff of 1e-5 to identity potential insertion sequence. Each insertion region was assigned to the top hit of alignment. Sequences in IS database with significant alignment scores were annotated by alignment to microbial protein database from NCBI.

GO term assignment and pathway analysis

Gene ontology terms (GO terms) assignment is a tool to identify functions of gene clusters to reflect important biological aspects (120). GO terms were assigned to predicted genes with Blast2GO (121). Gene function distribution was studied with GO classification of predicted genes with WEGO (122). Predicted genes were mapped to KEGG database from Blast2GO to predict pathway activities of my strains.

Identification of tRNA, rRNA and phage genes

tRNA, as the link between mRNA and proteins, delivers amino acids to the ribosome for peptide synthesis directly by triplet nucleotides (123). rRNA is an important component of ribosome, a complex that catalyzes protein synthesis (124). Bacteria and phages interact with each other biologically (94). Identification of tRNA and rRNA was performed with tRNA-scan (125) and RNAmmer (126) respectively. Phage genes were identified and classified with PHAST (127).

Synteny comparisons

Syntentic blocks are segments of sequences that exhibit conservation across species or within a chromosome. Collinearity reflects conserved orientation and conserved adjacency of genes (128). Synteny comparsions were performed with MUMMer (115) and CONTIGuator between *P. stuatii* draft genome and *P. stuartii* MRSN 2154, *P. mirabilis* draft genome and *P. mirabilis* HI4320 or *P. mirabilis* BB2000.

Recombination analysis

Recombination is a process of genetic material exchange between DNA strands to rearrange genes or parts of genes (129). Only clusters of orthologs with gene orthology across all *Proteus* or all *Providencia* strains used were retained for recombination analysis. Four statistical analyses were performed: GENECONV (inner fragments) (130) , Pairwise homoplasy Index (PHI), Maximum Chi square and Neighbor Similar Score. The latter three were performed with PhiPack (131) with 1000 permutations. Window size was set as 50 in PHI test. The p-values are corrected with program Q-value (132), and significant clusters were reported with FDR of 10%.

Positive selection analysis

Positive selection is selection of advantageous alleles to increase fitness (133). Positive selection analysis was performed with PAML (134) on 2213 and 1965 ortholog clusters across all tested *Proteus* and *Providencia* species respectively. Site-model studies were implemented with codeml to compare model M1a (nearly neutral) to M2a (positive selection). The likelihood ratio test statistics was compared with the chi square distribution with two degrees of freedom. Computed p-values were corrected with program Q-value (132) with FDR of 20%, cases with $p_2=0$ or $w_2=1$ are ignored. Amino acid sites were predicted with Bayes Empirical Bayes inference (BEB inference) (135).

Results

Basic genomic information

Table 4- 1 Basic genomic information

species	sequenced size (Mb)	number of contige	N50 contig	assembled genome	number of CDS	Average GC %
Providencia stuartii	1.38	97	139,103	draft	4522	40.35
Proteus mirabilis	2.77	75	174,883	draft 1	3725	38.48
				draft 2	3719	38.15

The basic genomic information is shown in Table 4-1. *P. stuartii* draft is assembled with reference of *P. stuartii* MRSN 2154, the only *P. stuartii* strain available with complete genome sequence. There are 4522 CDS predicted among draft genome, which are more than the reference genome with 4125 CDS. In total 38 contigs are unassembled in the scaffold assembly with 152 CDS among them. There are 14 rRNA and 74 tRNA regions identified in *P. stuartii* draft genome, with their positions shown on the draft genome in Figure 4-1.

There are two *P. mirabilis* draft genomes, because there are two *P. mirabilis* strains available with complete genome sequences, HI4320 and BB2000. The draft 1 is assembled with reference of *P. mirabilis* HI4320 and contains 3725 CDS, comparable to the corresponding reference genome with 3747 CDS. The draft 2 genome assembled with reference of BB2000 contains six less CDS than draft 1, the CDS number is more than the CDS of BB2000 with 3465 CDS. In total 19 contigs are unassembled in the

scaffold assembly with 109 CDS among them in either way of assembly. There are 15 rRNA and 78 tRNA regions identified in draft 1 genome and 17 rRNA and 78 tRNA regions in draft 2 genome, with their positions shown on draft genomes in Figure 4-2 and Figure 4-3 respectively.

All these three draft genomes show clear GC skew as Figure 4-4-4-6 show, which is an indication of DNA replication origin and terminus (136). Unique genes in draft genomes which are not found in their corresponding clinical reference genomes are also shown in these figures (4th ring), possibly associated with symbiosis with *L. sericata* and the microbial community.

Identification of phage genes

As shown in Figure 4-1, there are eight phage regions identified in *P. stuartii* draft genome. Among these phage regions, five are intact. They are from *Escherichia* phage HK75 (accession no. NC_016160), *Salmonella* phage vB_SosS_Oslo (accession no. NC_018279), *Shigella* phage Sf6 (accession no. NC_005344), *Pseudomonas* phage B3 (accession no. NC_006548) and *Pectobacterium* phage ZF40 (accession no. NC_019522), ordered by their positions on the scaffold beginning from the first nucleotide. Bacteriophage sf6 has been identified to express gene *oac* to change antigenic properties of O-antigen polysaccharide on surface of its host *Shigella flexneri* (137). Pili on *Pseudomonas* surface is required for page B3 adsorption (138).

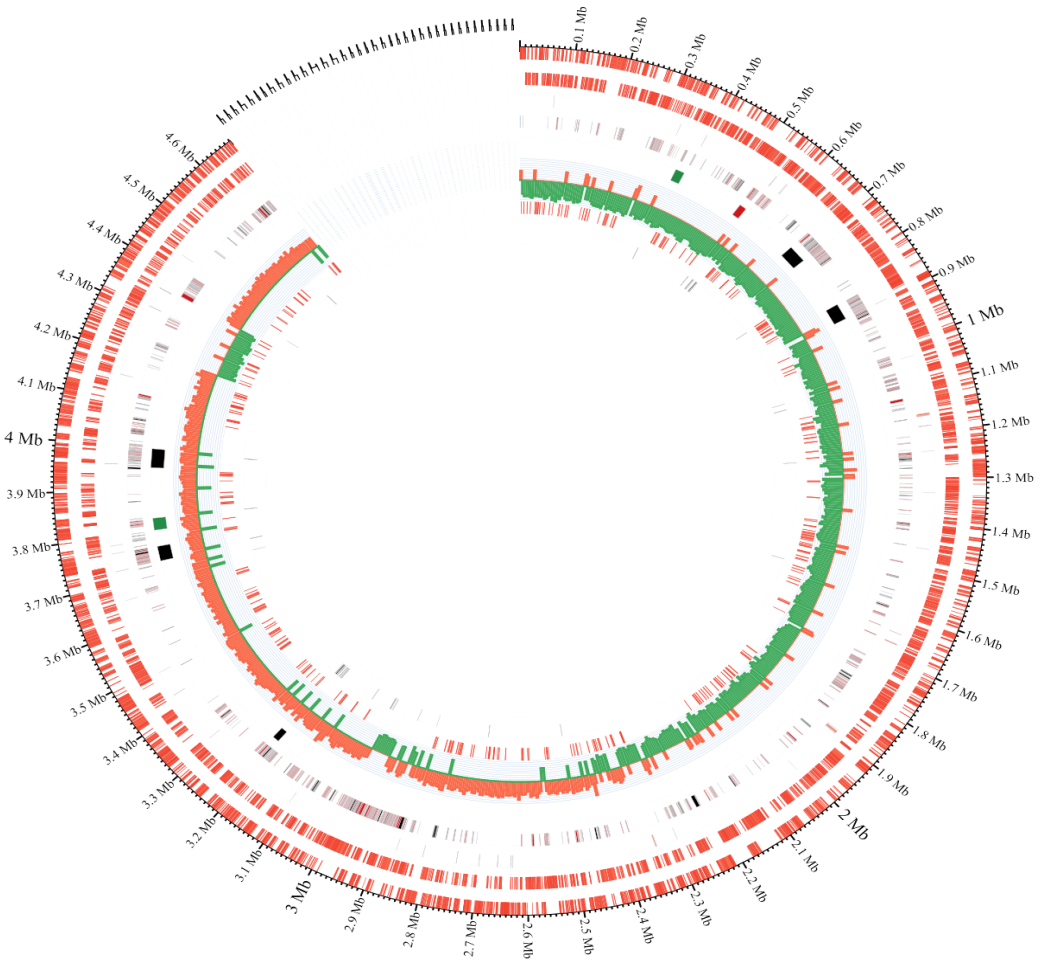


Figure 4- 1 Map of the *Providencia stuartii* draft genome. Unassembled contigs are shown as gaps with unknown positions. Rings from outermost to the center: (1) genes on the forward strand. (2) Genes on the reverse strand. (3) tRNA (black) and rRNA (red) genes. (4) Unique genes when compared to the corresponding reference genomes (P. mirabilis HI4320 and B2000) (5) intact (black), incomplete (red) and questionable (green) phage genes. (6) GC skew with window size of 2000nt with above average region in red and below average region in green. (7) Distribution of orthologous genes with evidence of recombination (8) insertion sequence regions.

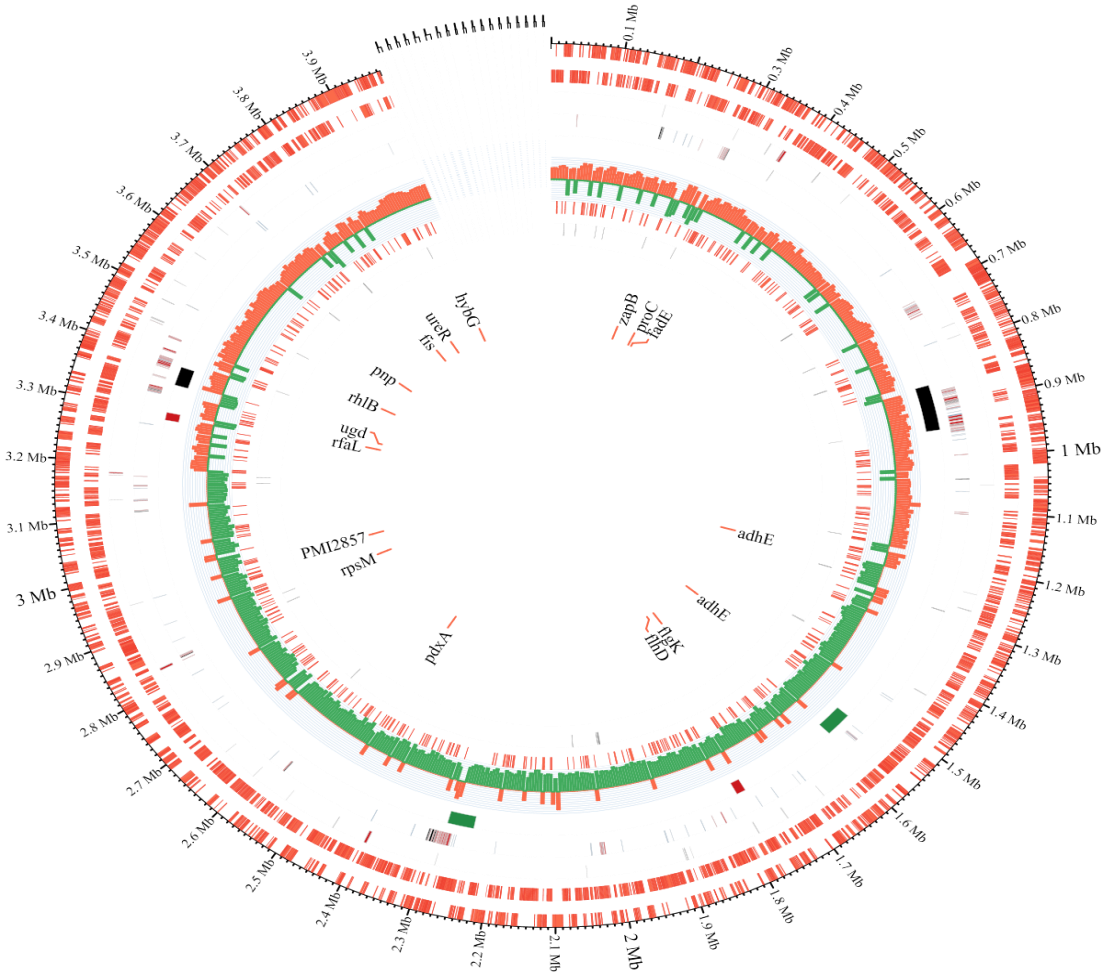


Figure 4- 2 Map of the *Proteus mirabilis* draft 1 genome. Unassembled contigs are shown as gaps with unknown positions. Annotation of rings from outermost to the center are the same as annotation of 1st-8th rings in Figure 4-1. (9) Genes related to swarming.

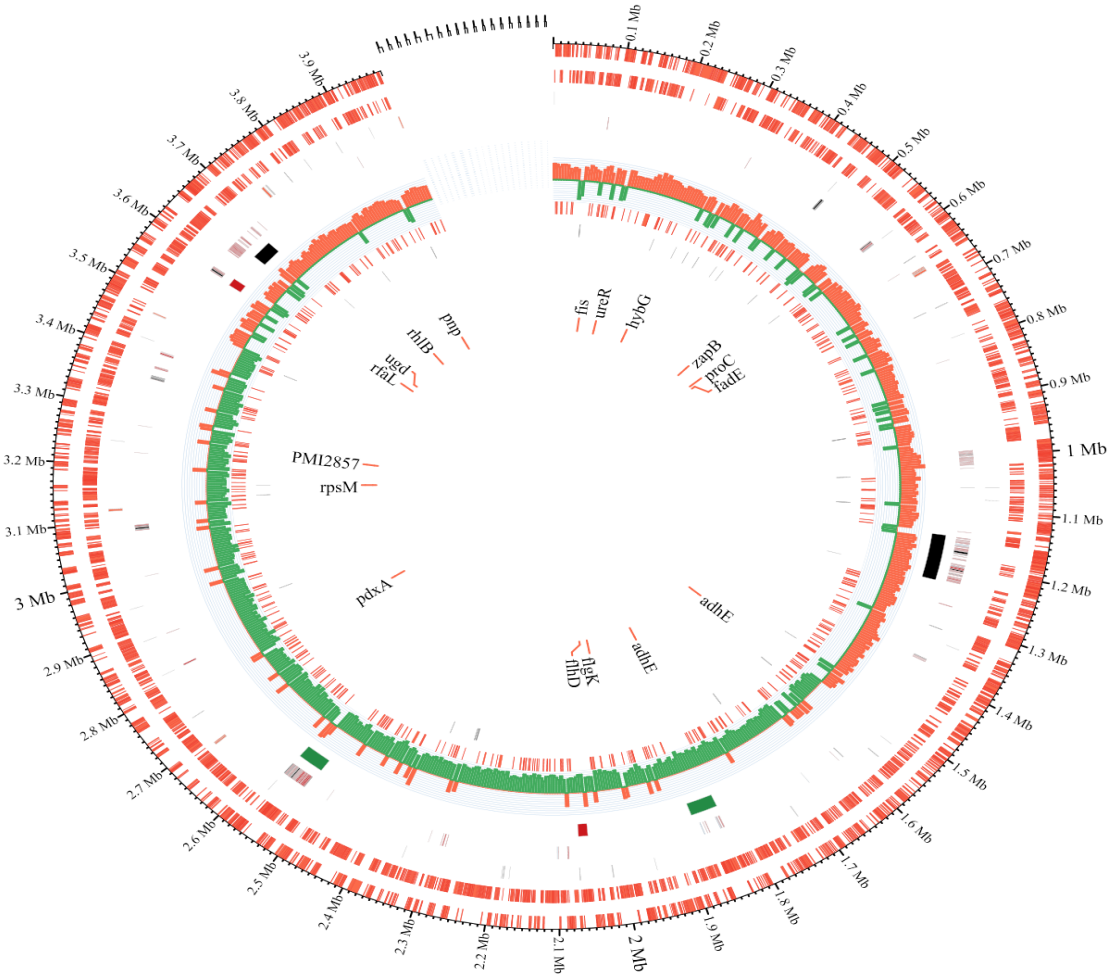


Figure 4- 3 Map of the *Proteus mirabilis* draft 2 genome. Annotation is the same as Figure 4-2.

Both *P. mirabilis* draft genome contain six phage regions, among which two are intact shown in black in the 5th ring of Figure 4-2 -4-3. They are from *Enterobacteria* phage mEp460 (accession no. NC_019716) and *Salmonella* phage Fels-2 (accession no NC_010463), ordered by their positions on the scaffold beginning from the fist nucleotide. Interestingly, within either *P. mirabilis* draft genome, a phage region from *Salmonella* phage Fels-2 is located closely to loci of two genes related to swarming,

UDP-glucose 6-dehydrogenase (ugd) and O-antigen ligase (rfaL) (35). Fels-2 prophage has been identified to be responsible for lethality phenotype of *lexA* null mutants of *Salmonella* (139). I hypnotize that the phage may impact *P. mirabilis* swarming and biochemical tests need to be performed to verify it.

Phylogenetic analysis

I determined the phylogenetic relationship from a concatenated alignment of 1322 orthologs shared by *Proteus* and *Providencia* strains with information available for analysis. *Escherichia coli* str. K-12 substr. MG1655 was set as an outgroup. As Figure 4-4 shows, the studied *P. stuartii* strain is closely related to *P. stuartii* ATCC 25827 and MRSN 2154, with only the latter strain having complete genome. The studied *P. mirabilis* strain is closely related to *P. mirabilis* BB2000 with either way of scaffold assembly. *P. mirabilis* BB2000 is a spontaneous rifampin-resistant mutant of PRM1 (140) with no plasmid (141). Compared to *P. mirabilis* HI4320, *P. mirabilis* BB2000 contains unique CDS related to toxin elements, self recognition and phages but no iron acquisition proteins or transfer (*tra*) genes (141).

Synteny comparison

Synteny relationship was examined between *P. stuartii* draft genome and *P. stuartii* MRSN 2154 reference genome, *P. mirabilis* draft 1 genome and *P. mirabilis* HI4320 reference genome, as well as P. mirabilis draft 2 genome and *P. mirabilis* BB2000 reference genome, as Figure 4-5 shows. There is some conserved synteny between *P.*

stuartii MRSN2154 and the draft genome. For *Proteus mirabilis*, draft 2 genome shows higher synteny to BB2000 than draft 1 to HI4320. The result also validates that the studied *P. mirabilis* draft genome is more evolutionarily closely related to BB2000 as shown in Figure 4-4.

MUMmer-based comparative genomic alignment analysis is shown in Figure 4-6. *P. stuartii* draft genome is somewhat collinear to MRSN2154. For *P. mirabilis* comparison, draft 2 genome shows more colinearity to BB2000 than draft 1 to HI4320.

SNP and indel analysis

SNPs and indels can serve as genetic markers to characterize variants and identify potential evolutionary mutations (142). As Figure 4-7 shows, there are a large number of SNPs and indels between *P. stuartii* draft and MRSN2154. There are relatively fewer SNPs and indels between *P. mirabilis* draft genomes and corresponding reference genomes. It seems the studied draft genomes are more closely related to *P. mirabilis* reference genomes than *P. stuartii* draft genome to its relevant reference genome, which is also reflected in synteny comparison in Figure 4-5 and 4-6. The indels between draft genomes and corresponding reference genomes suggests genome reorganization which may be associated with insect-bacteria association.

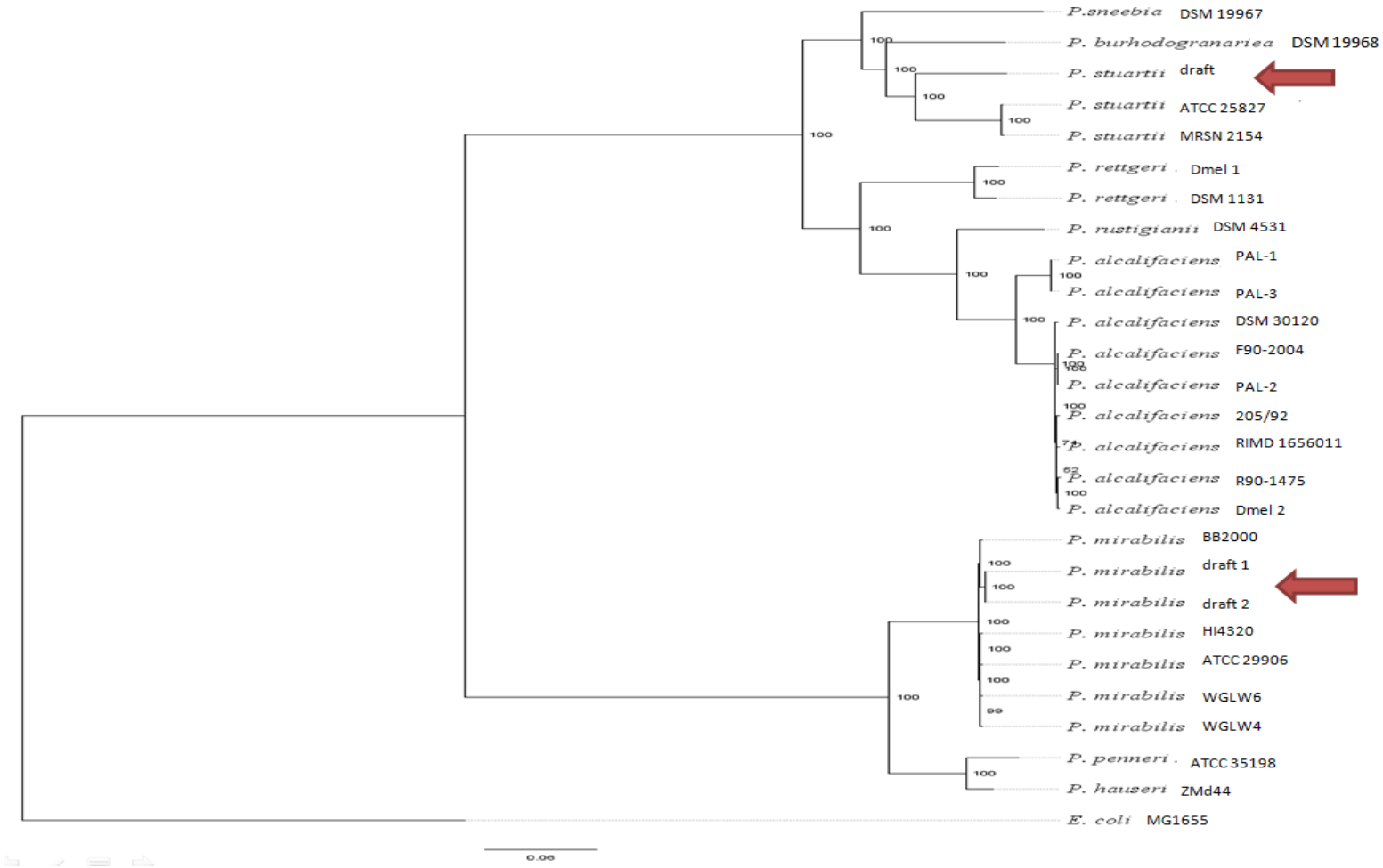


Figure 4- 4 Phylogenetic tree of *Providencia* and *Proteus* strains and the outgroup *E coli* with bootstrap as 100. Draft genomes are labeled with arrows.

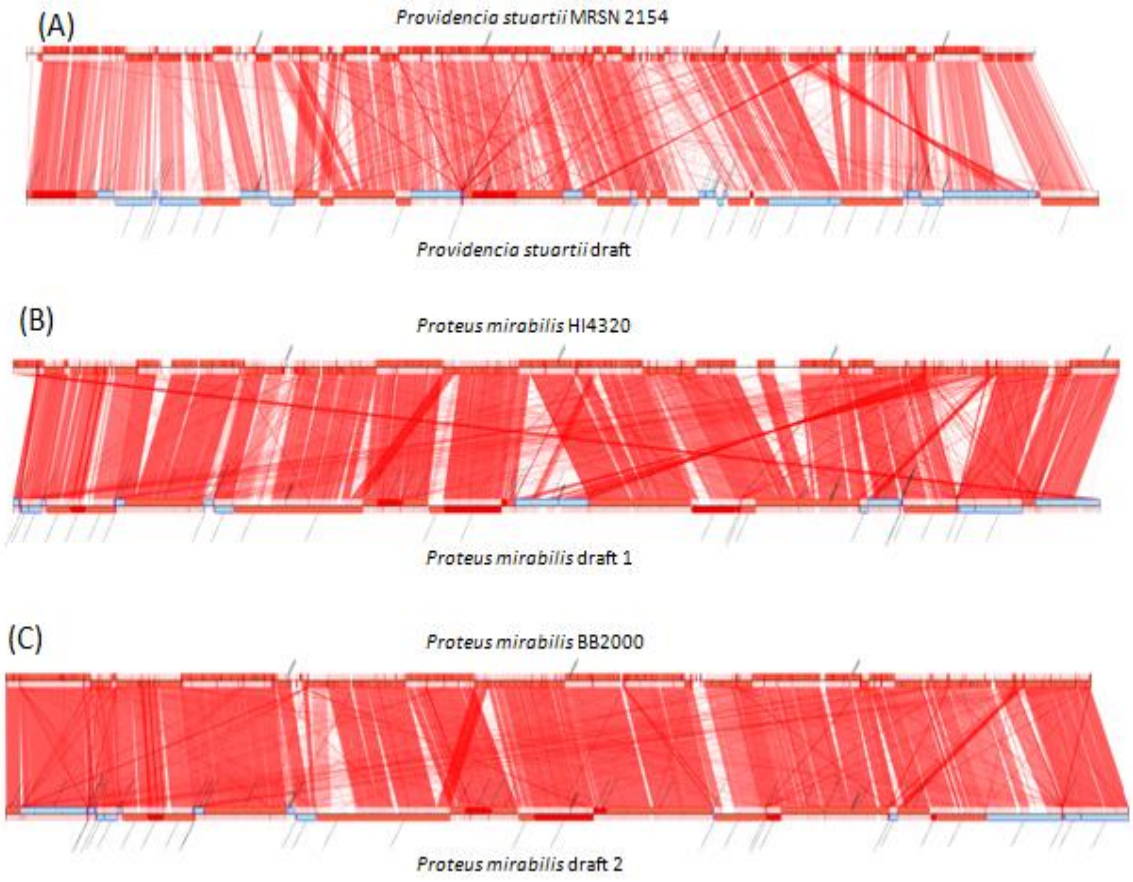


Figure 4- 5 Synteny comparison between *P. stuartii* draft to *P. stuartii* MRSN2154, as well as *P. mirabilis* draft to *P. mirabilis* HI4320 or *P. mirabilis* B2000. Contigs of draft genomes shown with solid red are overlapped with other contigs with two ends, those in light red are overlapped with one end, while those in blue do not show overlap with other contigs.

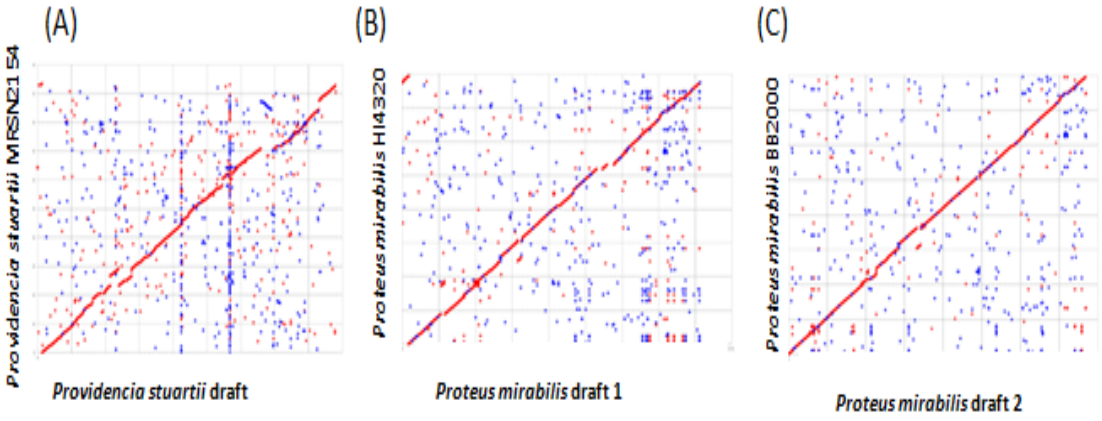


Figure 4- 6 Alignment between (A) *P. stuartii* MSRN2154 and *P. stuartii* draft (B) P. mirabilis HI4320 and P. mirabilis draft 1 (C) *P. mirabilis* B2000 and *P. mirabilis* draft 2. Red dots show alignment in the same orientation in a genomic pair while blue dots show alignment with opposite orientation.

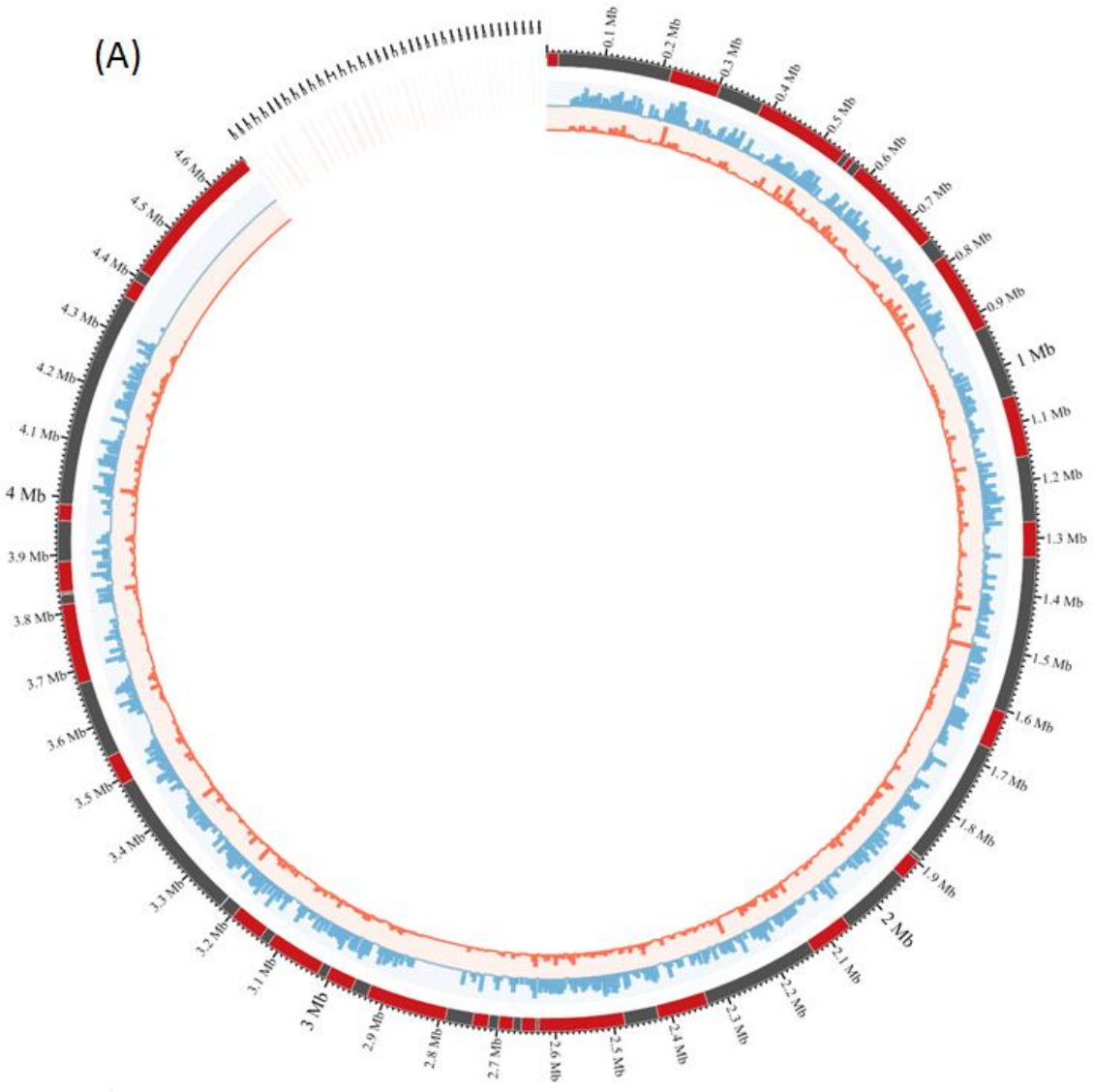


Figure 4- 7 Map of (A) *P. stuartii* draft (B) *P. mirabilis* draft 1 (C) *P. mirabilis* draft 2 genomes. Rings from outermost to the center: (1) contigs in scaffold assembly. (2) SNP. (3) Indel.

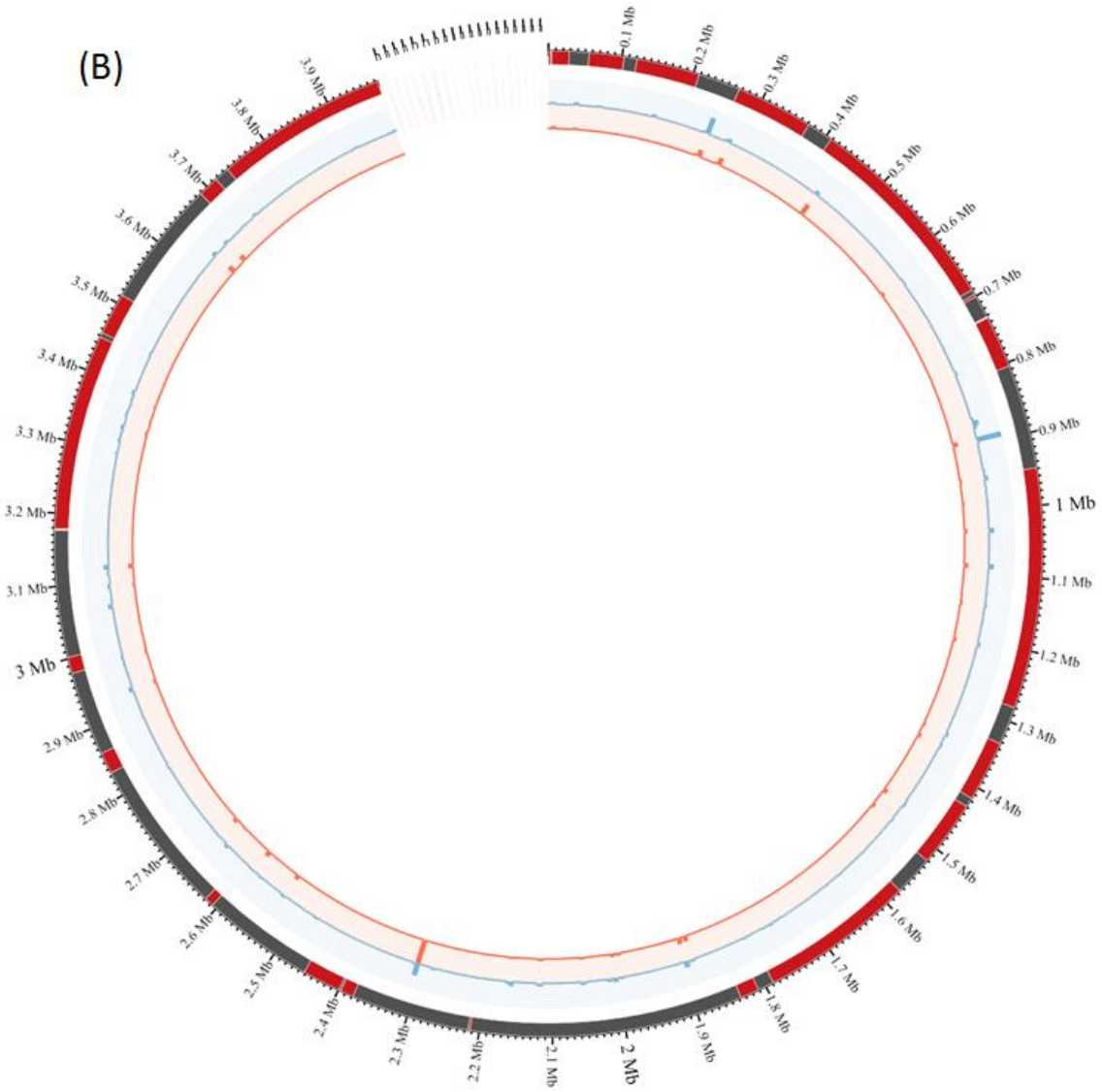


Figure 4- 7 Continued.

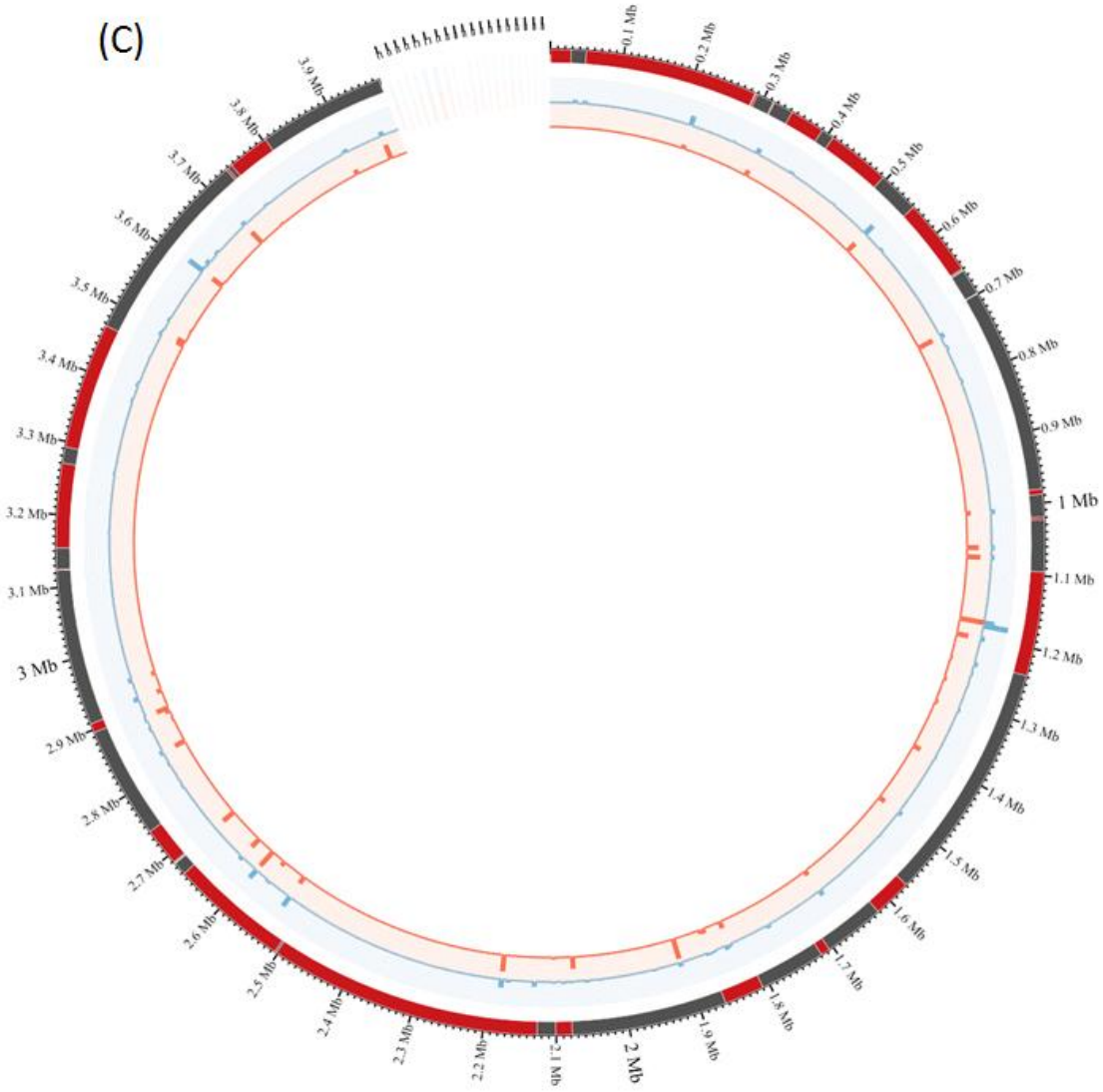
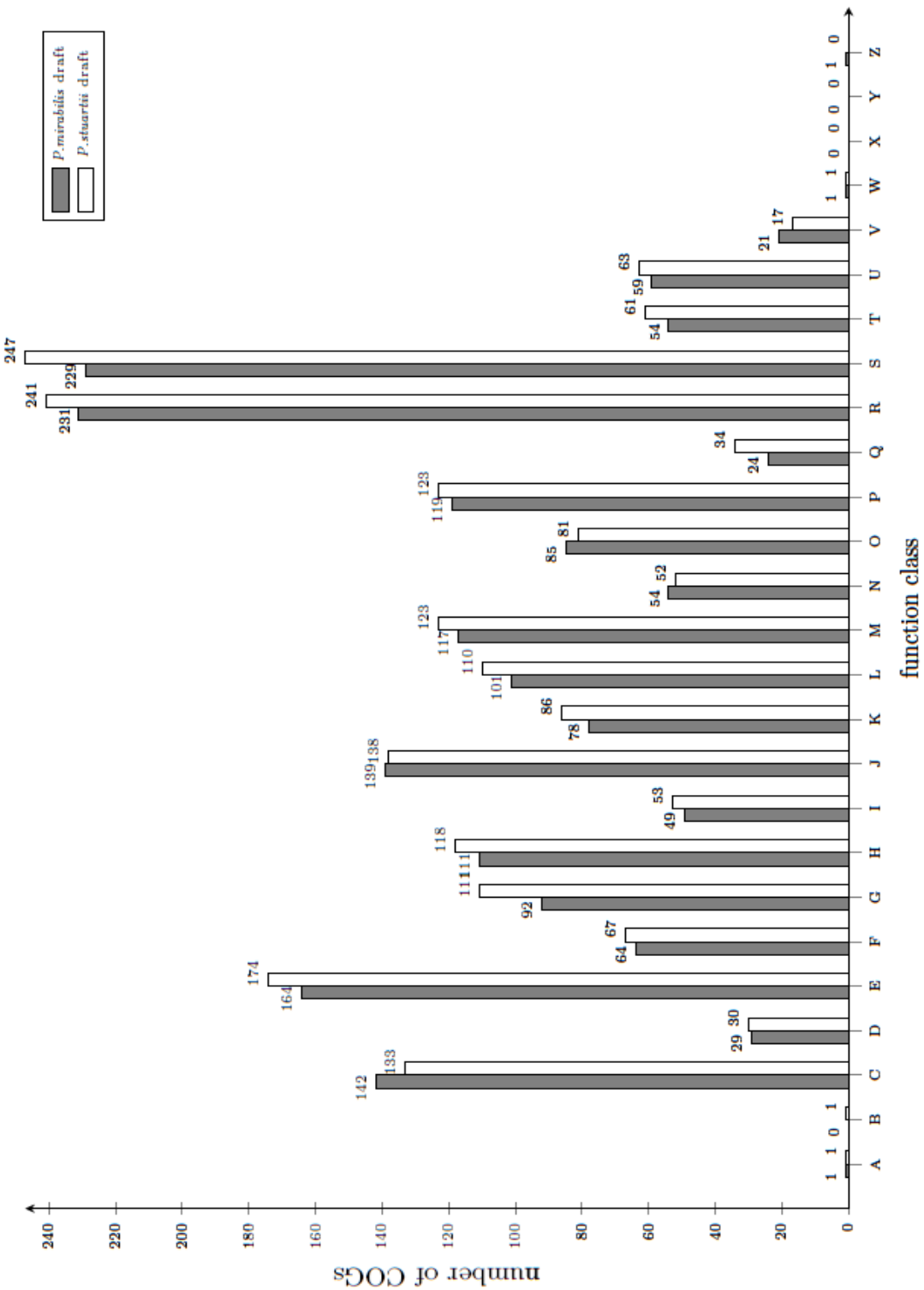


Figure 4- 7 Continued.

COG analysis

Figure 4-8 shows COG distribution of studied *P. mirabilis* and *P. stuartii*. The top five functional classes with most COGs are: [S] Function unknown, [R] General function prediction only, [E] Amino acid transport and metabolism, [C] Energy production and conversion, [J] Translation, ribosomal structure and biogenesis. Genes in the latter three function classes play an essential role in basic cellular functions. Interestingly, in the class [B] Chromatin structure and dynamics, *P. stuartii* draft genome has a COG, deacetylases including yeast histone deacetylase and acetoin utilization protein, which is not found in either *P. mirabilis* draft genome. Histone deacetylase is identified to reconstitute positive charge of lysine by catalyzing removal of the acetyl group from its side chain to stabilize interaction between histone and DNA (143). In class [Z] Cytoskeleton, *P. mirabilis* draft genome has a COG, myosin heavy chain, which is absent in *P. stuartii* draft genome. A protein with high molecular-weight in *E. coli* has been identified to share structural homology to a yeast heavy-chain myosin and supposed to play a role in movement of cell division and nucleoid segregation (144). In class [N] Cell motility, *P. mirabilis* draft genome has COG Tfp pilus assembly protein PilF, which is not found in *P. stuartii* draft genome. *Pseudomonas aeruginosa* mutant with knock-out homolog PilF does not exhibit swarming mobility (145). *P. mirabilis* exhibits swarming mobility but not for *P. stuartii (48)*, this COG absence in *P. stuartii* draft may shed light to that phenotypic difference.

Figure 4- 8 COG analysis of the draft genomes of *P. mirabilis* and *P. stuartii*. The draft 1 genome and draft 2 genome of *P. mirabilis* has the same distribution of COG, so only one is shown in the figure. [A] RNA processing and modification. [B] Chromatin structure and dynamics. [C] Energy production and conversion. [D] Cell cycle control, cell division, chromosome partitioning. [E] Amino acid transport and metabolism. [F] Nucleotide transport and metabolism. [G] Carbohydrate transport and metabolism. [H] Coenzyme transport and metabolism. [I] Lipid transport and metabolism. [J] Translation, ribosomal structure and biogenesis. [K] Transcription. [L] Replication, recombination and repair. [M] Cell wall/membrane/envelope biogenesis. [N] Cell motility. [O] Post-translational modification, protein turnover, and chaperones. [P] Inorganic ion transport and metabolism. [Q] Secondary metabolites biosynthesis, transport, and catabolism. [R] General function prediction only. [S] Function unknown. [T] Signal transduction mechanisms. [U] Intracellular trafficking, secretion, and vesicular transport. [V] Defense mechanisms. [W] Extracellular structures. [Y] Nuclear structure. [Z] Cytoskeleton.


P.mirabilis draft
P.stuartii draft
number of COGs
function class
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
0 20 40 60 80 100 120 140 160 180 200 220 240
1 1 0 1
142 133
29 30
164 174
64 67
92
111 111 118
49 53
139 138
78 86
101 110
117 123
54 52
85 81
119 123
24 34
231 241 229 247
54 61
59 63
21 17
1 1 0 0 0 0 1 0

GO term assignment

GO term assignment result is as Figure 4-9 shows. There are some GO terms unique to *P. mirabilis* draft 2 genome or both *P. mirabilis* draft genomes, including virion, antioxidant, electron carrier, transcription regulator and reproductive process. Considering some of these GO terms play an essential role in organisms and the relatively farther evolutionary distance between *P. stuartii* draft and the corresponding reference genome MRSN2154 used in scaffold assembly compared to *P. mirabilis* assembly, absence of these GO terms in *P. stuartii* draft may be caused by misassembly with reference genome. In GO term virion, there is a gene chitin-binding protein in *P. mirabilis* draft 2 genome. Chitin-binding protein is required for virus infection to host (146).

Recombination and positive selection

I examined 2213 *Proteus* orthologs and 1965 *Providencia* orthologs. They are used to identify recombination and positive selection among all *Proteus* and *Providencia* species tested respectively. Among those orthologs, there are 411 *Proteus* orthologs and 373 *Providencia* orthologs that show evidence of recombination with FDR <20%. The genes from my strains in those ortholog clusters are distributed evenly in the physical genomes, as Figure 4-1-Figure 4-3 show (the 7th ring).

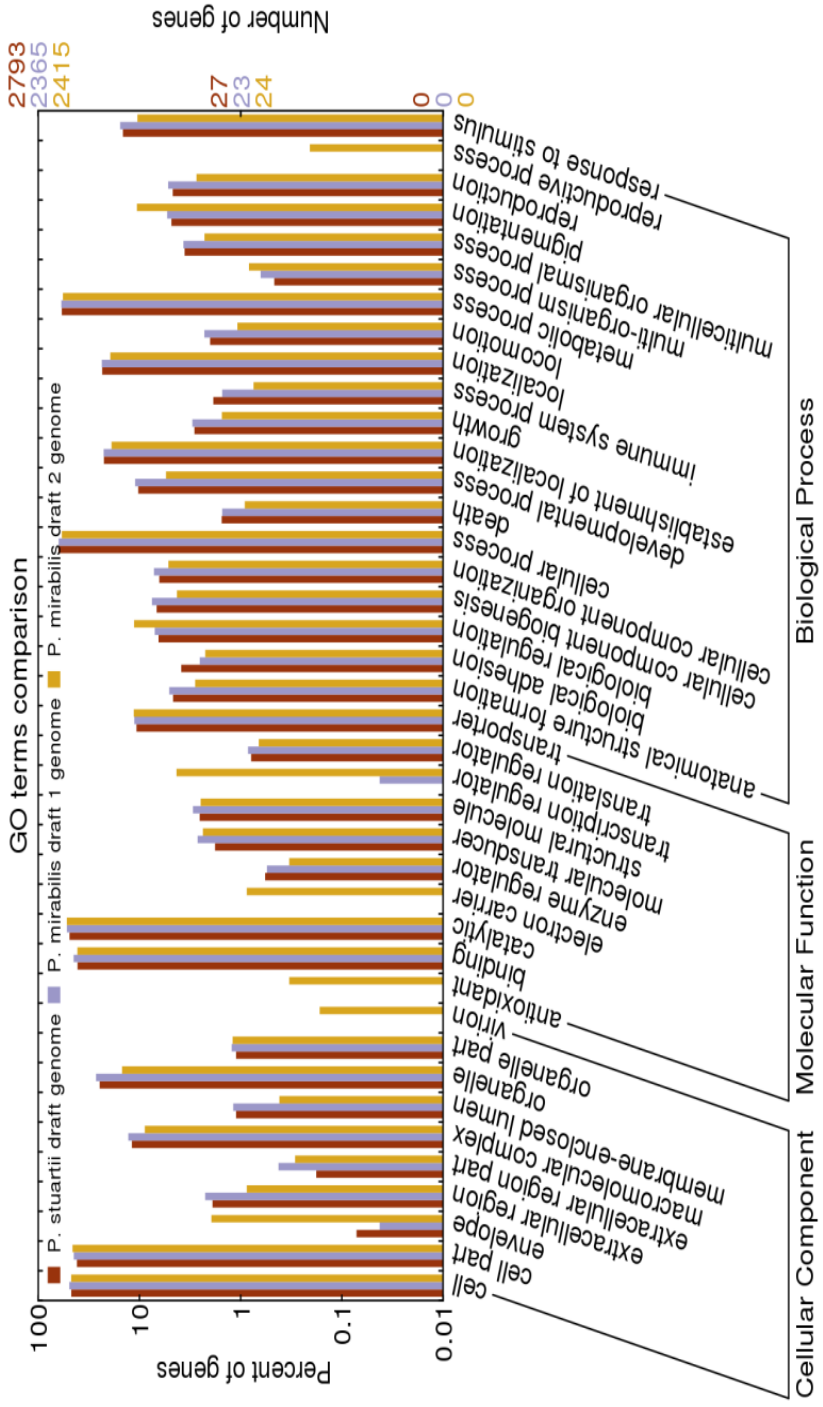


Figure 4- 9 GO term comparison at level 2 between *P. saurartii* draft genome, *P. mirabilis* draft 1 genome and *P. mirabilis* draft 2 genome.

To study positive selection among orthologs, I test *Proteus* and *Providencia* orthologs which do not show significant evidence of recombination, because recombination may violate the assumption of these models (147). It left 1802 *Proteus* orthologs and 1384 *Providencia* orthologs for positive selection test. I used Codeml in PAML to compare likehood of a neutral model M1a to a positive selection model M2a. I found 6 *Proteus* genes but no *Providencia* genes with significant evidence of positive selection with FDR<20%, as Table 4-2 shows.

Among *Proteus* genes with significant evidence of positive selection, 50S ribosomal protein L9 exhibits most significant p-value and q-value in the test. Ribosomal proteins are highly conserved, however, there are examples of ribosomal protein with evidence of positive selection. For example, LSU ribosomal protein L9p exhibits evidence for positive selection in site model tests of four *Providencia* species isolated from *Drosophila melanogaster* (148) .

The other gene exhibiting evidence for positive selection is a virulence factor intimin/invasin. The gene intimin encodes an outer membrane protein as adhesion for bacteria attachment to host cells and its homolog invasin, plays a role in mediating invasion. There is evidence for positive selection in intimin domains in *E. coli* (149), suggesting amino acid substitution to generate novel protein variants to prevent recognition by the host immune system.

The protein called Z-ring-associated protein shows significant evidence of positive selection. Z-ring-associated protein ZapA in *Bacillus subtilis* stimulates the assembly

and stabilization of Z-ring during cell division (150), but it is not essential for division. Positive selection on Z-ring-associated protein may help to modulate growth rate of cells.

The gene hofC encoding assembly protein in type IV pilin biogenesis (151) exhibits evidence of positive selection (p=3.408e-4, q=0.103). It is an outer membrane protein and also found to be positively selected in a Gram-negative bacterium, *Helicobacter pylori* (152).

Glutaredoxin-like protein also exhibits weak evidence for positive selection. The glutaredoxin-like protein NrdH in *E. coli* behaves like thioredoxin as hydrogen donor with the capability of reducing insulin disulfides (153). Positive selection on glutaredoxin-like proteins in *Proteus* may play a role in modulation of redox rate.

IS analysis

As Figure 4-10 shows, IS analysis result reflects among the annotated sources of these insertion sequences in IS database, six fragments show high similarity to sequences from *Escherichia coli* in *P. stuartii* draft genome and four fragments to sequences from *Pseudomonas putida* in either *P. mirabilis* draft genome. *E. coli* and *P. putida* are part of diet for flies (154,155) and coexist in the microbe community with *P. mirabilis* and *P. staurtii*. It is not surprising that insertion sequences from *E. coli* and *P. putida* are found in the draft genomes.

Table 4- 2 Details of genes with positive selection evidence.

annotation	p-value	q-value	$2\triangle\ell$[a]	parameter estimates[b]	Positive selective sites[c]
50S ribosomal protein L9	1.303e-08	2.348e-05	36.312	P_0=0.820, p_1=0, (p_2=0.180) ω_0=0, (ω_1=1), ω_2=999.000	44I, 45E, 48E, 49A, 50R, 51R, 52A
Intimin/invasion	1.870e=04	0.101	17.169	P_0=0.970, p_1=0.021, (p_2=0.009) ω_0=0.013, (ω_1=1), ω_2=326.885	45A
Z-ring-associated protein	2.251e-04	0.101	16.798	P_0=0.848, p_1=0, (p_2=0.152) ω_0=0.013, (ω_1=1), ω_2=326.885	73R, 74D, 75Y, 77Y, 78N, 79M, 80E, 81E, 82K
protein transport protein	3.408e-04	0.103	15.968	P_0=0.791, p_1=0.203, (p_2=0.006) ω_0=0.042, (ω_1=1), ω_2=44.614	22M
glutaredoxin-related protein	6.514e-04	0.168	14.673	P_0=0.965, p_1=0.005, (p_2=0.030) ω_0=0.072, (ω_1=1), ω_2=999.000	69A

a. Twice of the difference of likelihood values between neutral and positive selection models
b. Parameter estimates are from M2a model. Three sites class including under purifying selection, neutral and under positive selection are indicated with their proportions p_0, p_1 and p_2, and the nonsynonymous-synonymous substitution rate ration ω_0, ω_1 and ω_2, respectively. Parameters in paranthese are not free.
c. Positions of positive selection sites in alignment used in Codeml are identified using BEB inference with posterior probabilities >95%. Amino acids are referred to the first reference sequence in alignment.

Details of insertion sequences are mentioned in Table 4-3-Table 4-5. Most of these mobile elements are transposase, integrase and resolvase, which help insertion of foreign sequences into bacteria genomes. There are some inserted sequences with interest and labeled with gray in the tables. They are genes coding chloramphenicol exporter from *Corynebacterium striatum* and genes coding MerR family transcriptional regulator as well lipoprotein signal peptidase from *Pseudomonas putida.* For example, it seems these three assembled draft genomes contain genes coding chloramphenicol exporter which has been identified in *Corynebacterium striatum* and genes coding MerR family

transcriptional regulator as well lipoprotein signal peptidase identified in *Pseudomonas putida.*

Conclusions

I assembled and annotated draft genomes of *P. mirabilis* and *P. stuartii* of fly-isolated strains with reference of corresponding clinical strains. It shows studied *P. mirabilis* is more closely related to *P. mirabilis* BB2000 than *P. mirabilis* HI4320. I identified a phage *Salmonella* phage Fels-2 in either *P. mirabilis* draft genome which is located close to loci of two genes related to swarming, ugd and rfaL, and hypothesize the presence of this phage may play a role in bacteria swarming to attract flies. I found a COG Tfp pilus assembly protein PilF present in *P. mirabilis* draft but not in *P. stuartii* draft, which may partially account for swarming phenotype of *P. mirabilis*.

The present work utilized current bioinformatics analysis approaches to identify genomic features of *P. stuartii* and *P. mirabilis* strains isolated from larva of *L. sericata* and provided some hypotheses on functional differences between these strains and reference strains. Biochemical tests are needed to validate these hypotheses in the future.

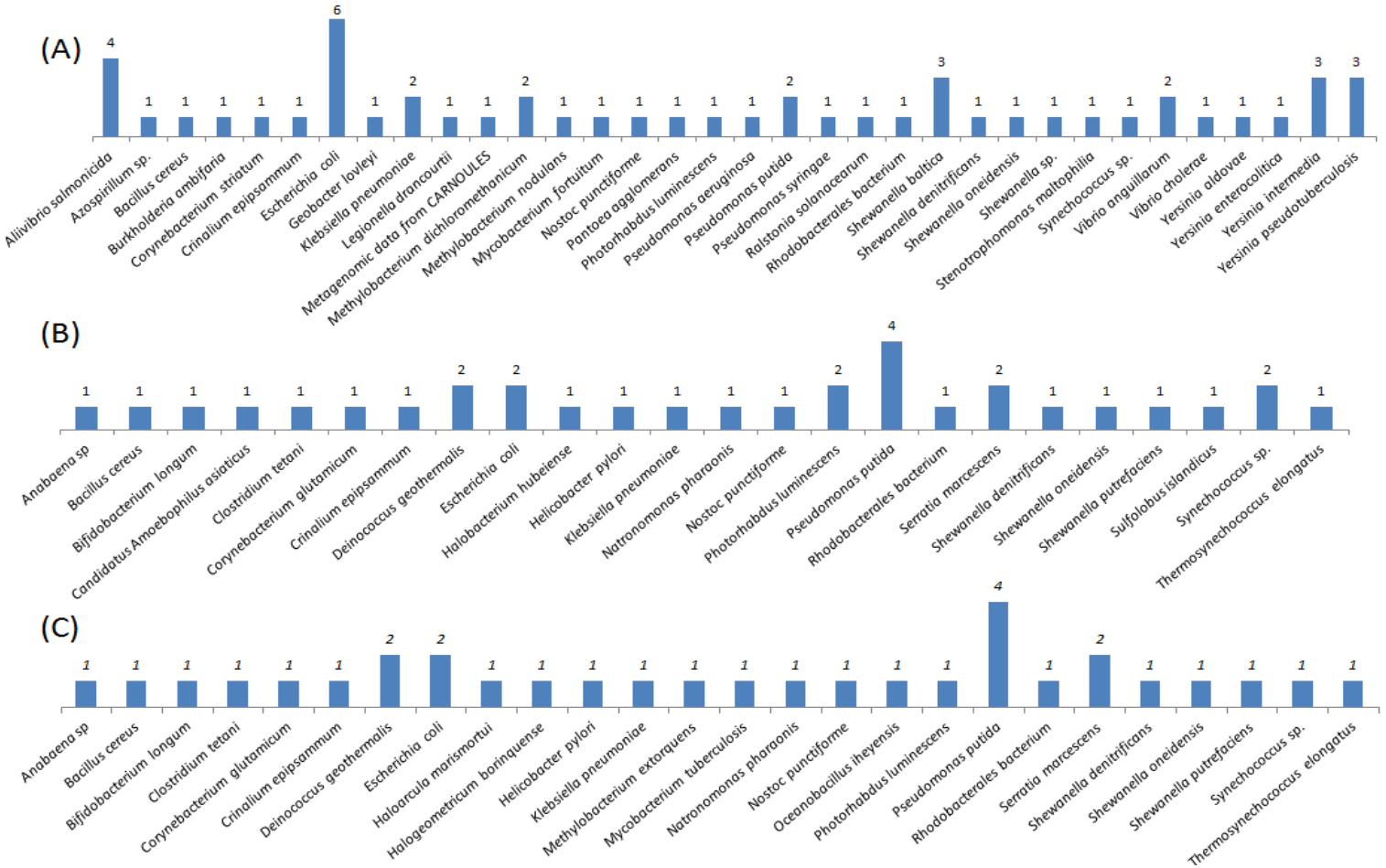


Figure 4- 10 Distribution of insertion sequence source annotated in database for (A) *P. stuartii* draft (B) *P. mirabilis* draft 1 and (C) *P. mirabilis* draft 2 genome.

Table 4- 3 Details of insertion sequence for *P. sauratii* draft genome.

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISEc12	IS21		*Escherichia coli*	0	transposase	scaffold	3021803	3020562
ISPsy30	Tn3		*Pseudomonas syringae*	0	transposase Tn3 family protein	scaffold	3220752	3217753
ISYal1	IS3	IS407	*Yersinia aldovae*	1.00E-178	transposase	scaffold	3831430	3830576
IS3000	Tn3	-	*Escherichia coli*	1.00E-174	transposase Tn3 family protein	scaffold	561159	559582
ISVch4	IS3	IS3	*Vibrio cholerae*	1.00E-174	transposase OrfAB subunit B	scaffold	3212622	3213539
IS1635	IS6	-	*Yersinia intermedia*	1.00E-117	putative transposase	c72	1569	835
ISEc12	IS21		*Escherichia coli*	1.00E-114	transposase	scaffold	3020434	3019688
IS1635	IS6	-	*Yersinia intermedia*	1.00E-109	putative transposase	scaffold	3199946	3199263
ISEc32	IS110		*Escherichia coli*	1.00E-100	IS110 family transposase	c61	104	1057
ISRhba4	IS481		*Rhodobacterales bacterium*	1.00E-97	integrase	scaffold	3207015	3206194
ISEc32	IS110		*Escherichia coli*	3.00E-97	IS110 family transposase	scaffold	542753	543706
ISBam1	IS3	IS150	*Burkholderia ambifaria*	8.00E-87	integrase catalytic subunit	scaffold	366278	367168
ISYps7	IS1		*Yersinia pseudotuberculosis*	2.00E-80	IS1 family transposase orfB	scaffold	3197694	3197254
ISVsa17	ISNCY	ISPlu 15	*Aliivibrio salmonicida*	2.00E-79	transposase	scaffold	577896	576973

Table 4-3 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISSba15	IS3	IS3	*Shewanella baltica*	4.00E-66	integrase catalytic subunit	scaffold	3221815	3222396
ISSde6	IS3	IS3	*Shewanella denitrificans*	2.00E-64	integrase catalytic subunit	scaffold	3214567	3215049
ISVsa19	Tn3		*Aliivibrio salmonicida*	6.00E-54	transposase	scaffold	575032	572699
IS6100	IS6	-	*Mycobacterium fortuitum*	1.00E-50	IS6100 transposase	scaffold	3011637	3011278
ISVa3	IS91		*Vibrio anguillarum*	3.00E-46	transposase	scaffold	3202150	3201668
ISSysp7	ISKra4	ISAzba1	*Synechococcus sp.*	2.00E-45	resolvase	scaffold	563481	564032
ISVsa9	IS91		*Aliivibrio salmonicida*	1.00E-44	transposase	scaffold	3205592	3206083
ISVsa17	ISNCY	ISPlu15	*Aliivibrio salmonicida*	2.00E-43	transposase	c80	2	646
ISPlu15	ISNCY	ISPlu15	*Photorhabdus luminescens*	2.00E-42	ISNCY family transposase	c97	375	1
IS3000	Tn3	-	*Escherichia coli*	7.00E-40	transposase, TnpA family	c94	340	11
ISEhe5	IS1	-	*Pantoea agglomerans*	5.00E-34	insertion element protein	scaffold	3013498	3013767
ISMdi3	IS3	IS407	*Methylobacterium dichloromethanicum*	5.00E-33	integrase catalytic subunit	scaffold	3993426	3994172
ISYps3	Tn3		*Yersinia pseudotuberculosis*	2.00E-29	hypothetical protein plu3296	c58	247	801

Table 4-3 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISCep1	ISKra4	ISAzba1	*Crinalium epipsammum*	6.00E-29	integrase family protein	scaffold	3082868	3083416
IS5564	IS481		*Corynebacterium striatum*	9.00E-29	chloramphenicol exporter	scaffold	4589174	4590094
ISPa38	Tn3		*Pseudomonas aeruginosa*	1.00E-26	resolvase	scaffold	1247898	1248422
ISPpu12	ISL3		*Pseudomonas putida*	1.00E-26	lipoprotein signal peptidase	scaffold	1563130	1562696
ISYen2A	IS21		*Yersinia enterocolitica*	6.00E-25	ISPsy4, transposition helper protein	scaffold	2869723	2869962
ISGlo3	IS481		*Geobacter lovleyi*	3.00E-24	Integrase catalytic region	scaffold	1074088	1074936
IS1635	IS6	-	*Yersinia intermedia*	4.00E-24	putative transposase	c80	1619	1828
ISShes11	Tn3		*Shewanella sp.*	2.00E-22	Transposon Tn21 resolvase	scaffold	3220927	3221484
ISMno23	IS91		*Methylobacterium nodulans*	5.00E-19	integrase family protein	scaffold	797390	796830
ISRso14	IS3	IS407	*Ralstonia solanacearum*	3.00E-17	transposase ISRSO14	scaffold	3831672	3831436
ISSod6	IS5	IS427	*Shewanella oneidensis*	4.00E-15	ISSod6, transposase	scaffold	4399671	4399480
ISVa3	IS91		*Vibrio anguillarum*	3.00E-14	transposase	c48	247	492
ISKpn21	ISNCY	IS1202	*Klebsiella pneumoniae*	8.00E-13	putative transposase	scaffold	2487730	2487434
ISSba14	Tn3		*Shewanella baltica*	1.00E-12	resolvase domain-containing protein	c59	4235	4495

Table 4-3 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISPpu12	ISL3		*Pseudomonas putida*	2.00E-12	MerR family transcriptional regulator	scaffold	1055390	1054989
ISLdr1	ISKra4	ISKra 4	*Legionella drancourtii*	4.00E-11	reverse transcriptase (RNA-dependent DNA polymerase)	c47	1397	1816
ISAzs36	IS481		*Azospirillum sp.*	2.00E-10	transposase	scaffold	3735114	3734158
ISCARN53	ISNCY	IS1202	Metagenomic data from CARNOULES	3.00E-10	Sea24	scaffold	2488005	2487838
ISYps3	Tn3		*Yersinia pseudotuberculosis*	3.00E-10	hypothetical protein plu3296	c80	1357	758
ISSba3	IS3	IS3	*Shewanella baltica*	5.00E-10	transposase IS3/IS911 family protein	scaffold	3018941	3019216
IS231K	IS4	IS231	*Bacillus cereus*	2.00E-08	ribosomal-protein-alanineacetyltransferase	scaffold	3696845	3696573
ISStma11	ISL3		*Stenotrophomonas maltophilia*	2.00E-08	transposase	scaffold	2930434	2930069
ISKpn28	IS481		*Klebsiella pneumoniae*	1.00E-07	hypothetical protein KPN_pKPN3p05990	scaffold	3223595	3223281
ISNpu13	Tn3		*Nostoc punctiforme*	2.00E-06	site-specific recombinase XerD-like protein	scaffold	2709485	2710078

Table 4-3 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISMdi3	IS3	IS407	*Methylobacterium dichloromethanicum*	1.00E-05	transposase of ISMdi3, IS3family (ORF 1)	scaffold	3993097	3993363

Table 4- 4 Details of insertion sequence for *P. mirabilis* draft genome 1

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
IS609	IS200/IS605		*Escherichia coli*	0	putative transposase	scaffold	639358	640551
ISSde6	IS3	IS3	*Shewanella denitrificans*	7.00E-67	integrase catalytic subunit	scaffold	291770	292222
IS609	IS200/IS605		*Escherichia coli*	4.00E-64	putative transposase	scaffold	3747161	3746736
ISDge10	IS200/IS605		*Deinococcus geothermalis*	5.00E-59	transposase, IS891/IS1136/IS1341	scaffold	3747181	3748326
ISPlu15	ISNCY	ISPlu15	*Photorhabdus luminescens*	3.00E-57	ISNCY family transposase	scaffold	2850047	2849298
IS606	IS200/IS605		*Helicobacter pylori*	1.00E-44	IS200 insertion sequence fromSARA17	scaffold	3512519	3512929
ISSm4	ISL3		*Serratia marcescens*	1.00E-43	hypothetical protein CAP2UW1_4293	scaffold	741125	740310
ISPpu12	ISL3		*Pseudomonas putida*	4.00E-28	lipoprotein signal peptidase	scaffold	3918609	3919046
ISCep1	ISKra4	ISAzba1	*Crinalium epipsammum*	4.00E-28	integrase family protein	scaffold	219697	219137

Table 4-4 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISSs1	IS200/IS605	IS1341	*Synechococcus sp.*	5.00E-27	transposase	scaffold	3639013	3639966
ISSoc3	IS200/IS605		*Synechococcus sp.*	7.00E-27	ISSoc3, orfA transposase	scaffold	102862	102479
ISKpn25	ISL3		*Klebsiella pneumoniae*	6.00E-26	putative type Irestriction-modification system DNA methylase	scaffold	1999932	1998736
ISTel2	IS200/IS605		*Thermosynechococcus elongatus*	2.00E-25	transposase	scaffold	1417342	1417743
ISSm4	ISL3		*Serratia marcescens*	9.00E-24	hypothetical protein KPN_pKPN4 p07084	scaffold	1996374	1994497
ISBlo15	IS200/IS605	IS1341	*Bifidobacterium longum*	1.00E-22	transposase, IS605 OrfB family	scaffold	805399	805989
ISRhba4	IS481		*Rhodobacterales bacterium*	1.00E-21	integrase	scaffold	938810	939022
IS891	IS200/IS605	IS1341	*Anabaena sp*	2.00E-21	transposase	scaffold	2606823	2606098
ISSod25	IS91		*Shewanella oneidensis*	4.00E-21	ISSod25 integrase Int_ISSod25	scaffold	2057434	2056598
ISCgl1	IS481		*Corynebacterium glutamicum*	2.00E-19	chloramphenicol exporter	scaffold	1266629	1265523
ISClte2	IS200/IS605		*Clostridium tetani*	2.00E-17	transposase-related protein	scaffold	120767	120369
ISDge19	IS200/IS605	IS1341	*Deinococcus geothermalis*	2.00E-15	transposase, IS605 OrfB	scaffold	805371	804511

Table 4-4 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISPpu12	ISL3		*Pseudomonas putida*	3.00E-14	MerR family transcriptional regulator	scaffold	2874511	2874107
ISPlu2	IS200/IS605	IS200	*Photorhabdus luminescens*	9.00E-13	IS200 family transposase	scaffold	3178639	3179052
ISNph21	IS200/IS605	IS1341	*Natronomonas pharaonis*	6.00E-12	IS1341-type transposase	scaffold	28181	28849
ISNpu13	Tn3		*Nostoc punctiforme*	4.00E-10	site-specific recombinase XerD-like protein	scaffold	1177440	1176892
ISPpu12	ISL3		*Pseudomonas putida*	6.00E-10	MerR family transcriptional regulator	c45	1812	1426
ISSpu9	IS4	IS50	*Shewanella putrefaciens*	7.00E-10	transposase Tn5 dimerisation subunit	scaffold	3405095	3405355
ISSis2	IS200/IS605		*Sulfolobus islandicus*	4.00E-08	transposase	scaffold	55464	56123
ISCaa10	IS200/IS605	IS200	*Candidatus Amoebophilus asiaticus*	7.00E-08	hypothetical protein Aasi_1732	scaffold	639302	638934
ISHhu2	IS200/IS605		*Halobacterium hubeiense*	2.00E-07	transposase, IS605 OrfB family	scaffold	1416427	1415792
ISPpu12	ISL3		*Pseudomonas putida*	3.00E-07	MerR family transcriptional regulator	c51	283	95
IS231K	IS4	IS231	*Bacillus cereus*	4.00E-07	HAD superfamily hydrolase	scaffold	3163243	3163959

Table 4- 5 Details of insertion sequence for *P. mirabilis* draft genome 2.

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
IS609	IS200/IS605		*Escherichia coli*	0	putative transposase	scaffold	930701	931894
ISSde6	IS3	IS3	*Shewanella denitrificans*	7.00E-67	integrase catalytic subunit	scaffold	579118	579570
IS609	IS200/IS605		*Escherichia coli*	4.00E-64	putative transposase	scaffold	59776	59351
ISDge10	IS200/IS605		*Deinococcus geothermalis*	5.00E-59	transposase, IS891/IS1136/IS1341	scaffold	59796	60941
ISPlu15	ISNCY	ISPlu15	*Photorhabdus luminescens*	3.00E-57	ISNCY family transposase	scaffold	3140468	3139719
IS606	IS200/IS605		*Helicobacter pylori*	1.00E-44	IS200 insertion sequence fromSARA17	scaffold	3803561	3803971
ISSm4	ISL3		*Serratia marcescens*	1.00E-43	hypothetical protein CAP2UW1_4293	scaffold	1030472	1029657
ISMex20	IS200/IS605	IS200	*Methylobacterium extorquens*	1.00E-29	transposase	scaffold	390128	389772
ISPpu12	ISL3		*Pseudomonas putida*	4.00E-28	lipoprotein signal peptidase	scaffold	231224	231661
ISCep1	ISKra4	ISAzba1	*Crinalium epipsammum*	4.00E-28	integrase family protein	scaffold	507044	506484
ISSs1	IS200/IS605	IS1341	*Synechococcus sp.*	5.00E-27	transposase	scaffold	3930055	3931008

Table 4-5 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISKpn25	ISL3		*Klebsiella pneumoniae*	6.00E-26	putative type I restriction-modification system DNA methylase	scaffold	2289290	2288094
ISTel2	IS200/IS605		*Thermosynechococcus elongatus*	2.00E-25	transposase	scaffold	1706694	1707095
ISSm4	ISL3		*Serratia marcescens*	9.00E-24	hypothetical protein KPN_pKPN4p07084	scaffold	2285732	2283855
ISBlo15	IS200/IS605	IS1341	*Bifidobacterium longum*	1.00E-22	transposase, IS605 OrfB family	scaffold	1094747	1095337
ISRhba4	IS481		*Rhodobacterales bacterium*	1.00E-21	integrase	scaffold	1228158	1228370
IS891	IS200/IS605	IS1341	*Anabaena sp*	2.00E-21	transposase	scaffold	2895996	2895271
ISSod25	IS91		*Shewanella oneidensis*	4.00E-21	ISSod25 integrase Int_ISSod25	scaffold	2346792	2345956
ISCgl1	IS481		*Corynebacterium glutamicum*	2.00E-19	chloramphenicol exporter	scaffold	1555978	1554872
ISClte2	IS200/IS605		*Clostridium tetani*	2.00E-17	transposase-related protein	scaffold	408112	407714
IS1535	IS607		*Mycobacterium tuberculosis*	2.00E-15	putative TRANSPOSASE	scaffold	1706665	1705748

Table 4-5 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISDge19	IS200/IS605	IS1341	*Deinococcus geothermalis*	2.00E-15	transposase, IS605 OrfB	scaffold	1094718	1093858
ISPpu12	ISL3		*Pseudomonas putida*	3.00E-14	MerR family transcriptional regulator	scaffold	3164933	3164529
ISHbo5	IS200/IS605	IS605	*Halogeometricum borinquense*	2.00E-12	transposase	scaffold	930660	930274
ISNph21	IS200/IS605	IS1341	*Natronomonas pharaonis*	6.00E-12	IS1341-type transposase	scaffold	315524	316192
ISNpu13	Tn3		*Nostoc punctiforme*	4.00E-10	site-specific recombinase XerD-like protein	scaffold	1466789	1466241
ISOih2	IS200/IS605	IS200	*Oceanobacillus iheyensis*	6.00E-10	transposase for IS657	scaffold	286600	286232
ISPpu12	ISL3		*Pseudomonas putida*	6.00E-10	MerR family transcriptional regulator	c45	1812	1426
ISSpu9	IS4	IS50	*Shewanella putrefaciens*	7.00E-10	transposase Tn5 dimerisation subunit	scaffold	3693058	3693318
ISPpu12	ISL3		*Pseudomonas putida*	3.00E-07	MerR family transcriptional regulator	c51	283	95
IS231K	IS4	IS231	*Bacillus cereus*	4.00E-07	HAD superfamily hydrolase	scaffold	3453667	3454383

Table 4-5 Continued

significant alignment	IS family	group	origin	*e*-value	annotation	query	begin	end
ISHma7	IS200/IS605		*Haloarcula marismortui*	8.00E-07	transposase	scaffold	342808	343476

CHAPTER V

CONCLUSION

As the advance in high-throughput sequencing enables the generation of large volumes of genomic and transcriptomic information, it provides researchers the opportunity to study non-model organisms even in the absence of a fully sequenced genome. This progress calls for powerful sequencing assembly algorithms because there are some challenging assembly problems related to the production of genomic and transcriptomic data from organisms whose genomes are not already known: (1) Some RNA products are highly expressed but others may have much lower expression level. (2) Data cannot easily be represented as a linear structure, due to post-transcription modification like alternative splicing. (3) Conserved sequences in domains in gene families make it difficult to understand whether a *de novo* sequence can be attributed to a single gene or several genes in a family, (4) sequencing errors due to technique limitations can interfere with the ability to develop effective *de novo* assemblies.

These assembly problems can be partially overcome by powerful assembly algorithms for non-model organisms. For those transcripts which are lowly expressed and may be ignored by traditional post-processing algorithms, they can be recovered by my algorithms extcontig and mutual. For transcripts generated from the same genes with alternative splicing events, branched structures are required to show relationship between these splicing products. Conserved sequences in domains in gene families may cause cyclic structures which brings trouble in post-processing and may be ignored in

transcript prediction. Partial sequencing errors can be corrected by increasing coverage cutoff of k-mers in de Bruijn graphs.

Advanced development in transcriptome sequencing calls for powerful sequencing assembly algorithms. I have developed algorithms that make use of evolutionary information to recover more similar transcripts from RNA-seq data especially those with low expression levels compared to traditional algorithms like Oases and Trans-ABySS. When my algorithms are applied to model organisms, it may be more time-consuming compared to current mapping-first algorithms which are based on fast alignment. The performance of my algorithms may not outcompete to the mapping-first algorithms because they make use of reference information, which is not required in my algorithms. The performance of my algorithms is affected by evolutionary distance between related organisms and complexity of their transcriptomes. For example, for the RNA-seq dataset of mouse transcriptome, my algorithms identify more and longer similar transcripts with transcript information of rat than that of human. Its performance is better on dataset of plants than yeast which shows less splicing events. My algorithms can be implemented in parallel by assigning disjoint subsets of nodes to different processors for extension, requiring less memory compared to current algorithms. This capability is advantageous for implementation by smaller research groups that lack access to higher-level computing systems. With their application to non-model organisms, computing systems with small memory are sufficient to identify similar transcripts, which may be longer and with higher resolution compared to current memory-intensive algorithms.

Although my algorithms requires less memory to identify similar transcripts from de Bruijn graphs compared to current post-processing algorithms, construction of de Bruijn graphs with large-scale sequence dataset still takes large memory. Future research can be performed on improvement of memory requirement of de Bruijn graph generation.

Current algorithms also help to annotate non-model organisms. For example, Prokaryotic dynamic programming gene-finding algorithm (Prodigal) can be applied to predict gene regions accurately. It shows high-quality gene prediction with low false positive rates. Therefore I apply different existing algorithms to annotate genomes of *P. mirabilis* and *P. stuartii*. The genome annotation work helps to understand interkingdom signaling between bacteria community and insects. My study on those genomes shows the differences between my strains isolated from larvae of blow flies and reference strains isolated from patients before, which may give hints to research of fly influence on bacteria community.

Non-model organisms are not well studied but the advance of high throughput sequencing technologies enables the genomic and transcriptomic studies by providing large volumes of sequence data. Application of existing and new algorithms paves the way to identifying genotypes that correspond to phenotypic features which play an important role in applied biology and broad view of scientists.

REFERENCES

1. Galiveti, C.R., Rozhdestvensky, T.S., Brosius, J., Lehrach, H. and Konthur, Z. (2010) Application of housekeeping npcRNAs for quantitative expression analysis of human transcriptome by real-time PCR. *RNA*, 16, 450-461.

2. Greco, S., Gorospe, M. and Martelli, F. (2015) Noncoding RNA in age-related cardiovascular diseases. *Journal of molecular and cellular cardiology*.

3. Ilik, I.A., Quinn, J.J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R.C., Luscombe, N., Backofen, R. *et al.* (2013) Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila. *Molecular cell*, 51, 156-173.

4. Wu, S.H. (2014) Gene expression regulation in photomorphogenesis from the perspective of the central dogma. *Annual review of plant biology*, 65, 311-333.

5. Wang, Y.Q., Zhang, H.M. and Cao, J. (2014) Binding of hydroxylated single-walled carbon nanotubes to two hemoproteins, hemoglobin and myoglobin. *Journal of photochemistry and photobiology. B, Biology*, 141, 26-35.

6. Werner, T., Liu, G., Kang, D., Ekengren, S., Steiner, H. and Hultmark, D. (2000) A family of peptidoglycan recognition proteins in the fruit fly Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 13772-13777.

7. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 5463-5467.

8. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376-380.

9. Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, 39, e90.

10. Lin, B., Wang, J. and Cheng, Y. (2008) Recent Patents and Advances in the Next-Generation Sequencing Technologies. *Recent patents on biomedical engineering*, 2008, 60-67.

11. Merriman, B. and Rothberg, J.M. (2012) Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, 33, 3397-3417.

12. Xu, M., Fujita, D. and Hanagata, N. (2009) Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small*, 5, 2638-2649.

13. Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002) Untranslated regions of mRNAs. *Genome biology*, 3, REVIEWS0004.

14. Ast, G. (2004) How did alternative splicing evolve? *Nature reviews. Genetics*, 5, 773-782.

15. Barberan-Soler, S. and Zahler, A.M. (2008) Alternative splicing regulation during C. elegans development: splicing factors as regulated targets. *PLoS genetics*, 4, e1000001.

16. Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999) Frequent alternative splicing of human genes. *Genome research*, 9, 1288-1293.

17. Celotto, A.M. and Graveley, B.R. (2001) Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. *Genetics*, 159, 599-608.

18. Hodgkin, J. (1989) Drosophila sex determination: a cascade of regulated splicing. *Cell*, 56, 905-906.

19. Liu, H.X., Cartegni, L., Zhang, M.Q. and Krainer, A.R. (2001) A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nature genetics*, 27, 55-58.

20. James, A.B., Syed, N.H., Bordage, S., Marshall, J., Nimmo, G.A., Jenkins, G.I., Herzyk, P., Brown, J.W. and Nimmo, H.G. (2012) Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. *The Plant cell*, 24, 961-981.

21. Iwata, H. and Gotoh, O. (2011) Comparative analysis of information contents relevant to recognition of introns in many species. *BMC genomics*, 12, 45.

22. Rajan, P., Elliott, D.J., Robson, C.N. and Leung, H.Y. (2009) Alternative splicing and biological heterogeneity in prostate cancer. *Nature reviews. Urology*, 6, 454-460.

23. Green, M.R. (1986) Pre-mRNA splicing. *Annual review of genetics*, 20, 671-708.

24. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28, 511-515.

25. Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28, 503-510.

26. Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 9748-9753.

27. Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18, 821-829.

28. Rogers, M.E., Colmer, T.D., Frost, K., Henry, D., Cornwall, D., Hulm, E., Deretic, J., Hughes, S.R. and Craig, A.D. (2008) Diversity in the genus Melilotus for tolerance to salinity and waterlogging. *Plant Soil*, 304, 89-101.

29. Al Sherif, E.A. (2009) Melilotus indicus (L.) All., a salt-tolerant wild leguminous herb with high potential for use as a forage crop in salt-affected soils. *Flora*, 204, 737-746.

30. Rogers, M.E., Colmer, T.D., Nichols, P.G.H., Hughes, S.J., Frost, K., Cornwall, D., Chandra, S., Miller, S.M. and Craig, A.D. (2011) Salinity and waterlogging tolerance amongst accessions of messina (Melilotus siculus). *Crop Pasture Sci*, 62, 225-235.

31. Mann, P.G. (1972) Proteus urinary infections in childhood. *Journal of clinical pathology*, 25, 551.

32. Mobley, H.L. and Warren, J.W. (1987) Urease-positive bacteriuria and obstruction of long-term urinary catheters. *Journal of clinical microbiology*, 25, 2216-2217.

33. Ahlinder, J., Ohrman, C., Svensson, K., Lindgren, P., Johansson, A., Forsman, M., Larsson, P. and Sjodin, A. (2012) Increased knowledge of Francisella genus diversity highlights the benefits of optimised DNA-based assays. *BMC microbiology*, 12, 220.

34. Bahrani, F.K., Johnson, D.E., Robbins, D. and Mobley, H.L. (1991) Proteus mirabilis flagella and MR/P fimbriae: isolation, purification, N-terminal analysis, and serum antibody response following experimental urinary tract infection. *Infection and immunity*, 59, 3574-3580.

35. Ma, Q., Fonseca, A., Liu, W., Fields, A.T., Pimsler, M.L., Spindola, A.F., Tarone, A.M., Crippen, T.L., Tomberlin, J.K. and Wood, T.K. (2012) Proteus mirabilis interkingdom swarming signals attract blow flies. *The ISME journal*, 6, 1356-1366.

36. Daniels, R., Vanderleyden, J. and Michiels, J. (2004) Quorum sensing and swarming migration in bacteria. *FEMS microbiology reviews*, 28, 261-289.

37. Waters, C.M. and Bassler, B.L. (2005) Quorum sensing: cell-to-cell communication in bacteria. *Annual review of cell and developmental biology*, 21, 319-346.

38. Falkinham, J.O., 3rd and Hoffman, P.S. (1984) Unique developmental characteristics of the swarm and short cells of Proteus vulgaris and Proteus mirabilis. *Journal of bacteriology*, 158, 1037-1040.

39. Kearns, D.B. (2010) A field guide to bacterial swarming motility. *Nature reviews. Microbiology*, 8, 634-644.

40. Darnton, N.C., Turner, L., Rojevsky, S. and Berg, H.C. (2010) Dynamics of bacterial swarming. *Biophysical journal*, 98, 2082-2090.

41. Jaklic, D., Lapanje, A., Zupancic, K., Smrke, D. and Gunde-Cimerman, N. (2008) Selective antimicrobial activity of maggots against pathogenic bacteria. *Journal of medical microbiology*, 57, 617-625.

42. Cickova, H., Cambal, M., Kozanek, M. and Takac, P. (2013) Growth and Survival of Bagged Lucilia sericata Maggots in Wounds of Patients Undergoing Maggot Debridement Therapy. *Evidence-based complementary and alternative medicine : eCAM*, 2013, 192149.

43. Horobin, A.J., Shakesheff, K.M., Woodrow, S., Robinson, C. and Pritchard, D.I. (2003) Maggots and wound healing: an investigation of the effects of secretions from Lucilia sericata larvae upon interactions between human dermal fibroblasts and extracellular matrix components. *The British journal of dermatology*, 148, 923-933.

44. Mohd Masri, S., Nazni, W.A., Lee, H.L., TA, T.R. and Subramaniam, S. (2005) Sterilisation of Lucilia cuprina Wiedemann maggots used in therapy of intractable wounds. *Tropical biomedicine*, 22, 185-189.

45. Ahmad, A., Broce, A. and Zurek, L. (2006) Evaluation of significance of bacteria in larval development of Cochliomyia macellaria (Diptera: Calliphoridae). *Journal of medical entomology*, 43, 1129-1133.

46. Sherman, R.A., Hall, M.J.R. and Thomas, S. (2000) Medicinal maggots: An ancient remedy for some contemporary afflictions. *Annu Rev Entomol*, 45, 55-81.

47. Manos, J. and Belas, R. (2006) The Genera Proteus, Providencia, and Morganella. *Prokaryotes: A Handbook on the Biology of Bacteria, Vol 6, Third Edition*, 245-269.

48. O'Hara, C.M., Brenner, F.W. and Miller, J.M. (2000) Classification, identification, and clinical significance of Proteus, Providencia, and Morganella. *Clinical microbiology reviews*, 13, 534-546.

49. Armbruster, C.E., Smith, S.N., Yep, A. and Mobley, H.L. (2014) Increased incidence of urolithiasis and bacteremia during Proteus mirabilis and Providencia stuartii coinfection due to synergistic induction of urease activity. *The Journal of infectious diseases*, 209, 1524-1532.

50. Macleod, S.M. and Stickler, D.J. (2007) Species interactions in mixed-community crystalline biofilms on urinary catheters. *Journal of medical microbiology*, 56, 1549-1557.

51. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12, 323.

52. Roberts, A. and Pachter, L. (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*, 10, 71-U99.

53. Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome research*, 17, 1697-1706.

54. Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. and Jaffe, D.B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18, 810-820.

55. Chaisson, M.J. and Pevzner, P.A. (2008) Short read fragment assembly of bacterial genomes. *Genome research*, 18, 324-330.

56. Hernandez, D., Francois, P., Farinelli, L., Osteras, M. and Schrenzel, J. (2008) De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome research*, 18, 802-809.

57. Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G., Morin, R.D., Zhao, Y.J., Hirst, M., Schein, J.E. *et al.* (2009) De novo transcriptome assembly with ABySS. *Bioinformatics*, 25, 2872-2877.

58. Li, R.Q., Zhu, H.M., Ruan, J., Qian, W.B., Fang, X.D., Shi, Z.B., Li, Y.R., Li, S.T., Shan, G., Kristiansen, K. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20, 265-272.

59. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q.D. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29, 644-U130.

60. 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.* (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods*, 7, 909-U962.

61. Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28, 1086-1092.

62. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of molecular biology*, 215, 403-410.

63. Wu, Y.W., Rho, M., Doak, T.G. and Ye, Y.Z. (2012) Stitching gene fragments with a network matching algorithm improves gene assembly for metagenomics. *Bioinformatics*, 28, I363-I369.

64. Bao, E., Jiang, T. and Girke, T. (2013) BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences. *Bioinformatics*, 29, 1250-1259.

65. Pevzner, P.A. (1989) 1-Tuple DNA Sequencing - Computer-Analysis. *Journal of biomolecular structure & dynamics*, 7, 63-73.

66. Idury, R.M. and Waterman, M.S. (1995) A new algorithm for DNA sequence assembly. *Journal of computational biology : a journal of computational molecular cell biology*, 2, 291-306.

67. Heber, S., Alekseyev, M., Sze, S.H., Tang, H. and Pevzner, P.A. (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, 18 Suppl 1, S181-188.

68. Daines, B., Wang, H., Wang, L.G., Li, Y.M., Han, Y., Emmert, D., Gelbart, W., Wang, X., Li, W., Gibbs, R. *et al.* (2011) The Drosophila melanogaster transcriptome by paired-end RNA sequencing. *Genome research*, 21, 315-324.

69. Bahn, J.H., Lee, J.H., Li, G., Greer, C., Peng, G.D. and Xiao, X.S. (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome research*, 22, 142-150.

70. Marquez, Y., Brown, J.W.S., Simpson, C., Barta, A. and Kalyna, M. (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome research*, 22, 1184-1195.

71. Sze, S.H., Dunham, J.P., Carey, B., Chang, P.L., Li, F., Edman, R.M., Fjeldsted, C., Scott, M.J., Nuzhdin, S.V. and Tarone, A.M. (2012) A de novo transcriptome assembly of Lucilia sericata (Diptera: Calliphoridae) with predicted alternative splices, single nucleotide polymorphisms and transcript expression estimates. *Insect Mol Biol*, 21, 205-221.

72. Kim, E.B., Fang, X.D., Fushan, A.A., Huang, Z.Y., Lobanov, A.V., Han, L.J., Marino, S.M., Sun, X.Q., Turanov, A.A., Yang, P.C. *et al.* (2011) Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, 479, 223-227.

73. MacManes, M.D. and Lacey, E.A. (2012) The Social Brain: Transcriptome Assembly and Characterization of the Hippocampus from a Social Subterranean Rodent, the Colonial Tuco-Tuco (Ctenomys sociabilis). *PloS one*, 7.

74. Garg, R., Patel, R.K., Tyagi, A.K. and Jain, M. (2011) De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. *DNA Res*, 18, 53-63.

75. Brown, S.A., Towers, G.H.N. and Wright, D. (1960) Biosynthesis of the Coumarins - Tracer Studies on Coumarin Formation in Hierochloe-Odorata and Melilotus-Officinalis. *Can J Biochem Phys*, 38, 143-156.

76. Li, B.J., Cong, F., Tan, C.P., Wang, S.X. and Goff, S.P. (2002) Aph2, a protein with a zf-DHHC motif, interacts with c-Abl and has pro-apoptotic activity. *Journal of Biological Chemistry*, 277, 28870-28876.

77. Osterloh, J.M., Yang, J., Rooney, T.M., Fox, A.N., Adalbert, R., Powell, E.H., Sheehan, A.E., Avery, M.A., Hackett, R., Logan, M.A. *et al.* (2012)

dSarm/Sarm1 Is Required for Activation of an Injury-Induced Axon Death Pathway. *Science*, 337, 481-484.

78. Maritano, D., Sugrue, M.L., Tininini, S., Dewilde, S., Strobl, B., Fu, X.P., Murray-Tait, V., Chiarle, R. and Poli, V. (2004) The STAT3 isoforms and have unique and specific functions. *Nat Immunol*, 5, 401-409.

79. Lam, B.C.H., Sage, T.L., Bianchi, F. and Blumwald, E. (2001) Role of SH3 domain-containing proteins in clathrin-mediated vesicle trafficking in Arabidopsis. *The Plant cell*, 13, 2499-2512.

80. Boyle, E.I., Weng, S.A., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. and Sherlock, G. (2004) GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20, 3710-3715.

81. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139-140.

82. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5, R80.

83. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of molecular biology*, 215, 403-410.

84. Pevzner, P.A. (1989) 1-Tuple DNA sequencing: computer analysis. *Journal of biomolecular structure & dynamics*, 7, 63-73.

85. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S. *et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 38, D5-16.

86. Coon, S.L., Munson, P.J., Cherukuri, P.F., Sugden, D., Rath, M.F., Moller, M., Clokie, S.J.H., Fu, C., Olanich, M.E., Rangel, Z. *et al.* (2012) Circadian changes in long noncoding RNAs in the pineal gland. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 13319-13324.

87. Heap, G.A., Yang, J.H., Downes, K., Healy, B.C., Hunt, K.A., Bockett, N., Franke, L., Dubois, P.C., Mein, C.A., Dobson, R.J. *et al.* (2010) Genome-wide

analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human molecular genetics*, 19, 122-134.

88. Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859-1875.

89. Chu, H.T., Hsiao, W.W.L., Chen, J.C., Yeh, T.J., Tsai, M.H., Lin, H., Liu, Y.W., Lee, S.A., Chen, C.C., Tsao, T.T.H. *et al.* (2013) EBARDenovo: highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. *Bioinformatics*, 29, 1004-1010.

90. Aronson, A.I., Beckman, W. and Dunn, P. (1986) Bacillus thuringiensis and related insect pathogens. *Microbiological reviews*, 50, 1-24.

91. Sanchez-Contreras, M. and Vlisidou, I. (2008) The Diversity of Insect-bacteria Interactions and its Applications for Disease Control. *Biotechnol Genet Eng*, 25, 203-243.

92. Gunduz, E.A. and Douglas, A.E. (2009) Symbiotic bacteria enable insect to use a nutritionally inadequate diet. *P R Soc B*, 276, 987-991.

93. Eleftherianos, I., Boundy, S., Joyce, S.A., Aslam, S., Marshall, J.W., Cox, R.J., Simpson, T.J., Clarke, D.J., ffrench-Constant, R.H. and Reynolds, S.E. (2007) An antibiotic produced by an insect-pathogenic bacterium suppresses host defenses through phenoloxidase inhibition. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 2419-2424.

94. Diaz-Munoz, S.L. and Koskella, B. (2014) Bacteria-phage interactions in natural environments. *Advances in applied microbiology*, 89, 135-183.

95. Bukovska, G., Klucar, L., Vlcek, C., Adamovic, J., Turna, J. and Timko, J. (2006) Complete nucleotide sequence and genome analysis of bacteriophage BFK20--a lytic phage of the industrial producer Brevibacterium flavum. *Virology*, 348, 57-71.

96. Stanley, T.L., Ellermeier, C.D. and Slauch, J.M. (2000) Tissue-specific gene expression identifies a gene in the lysogenic phage Gifsy-1 that affects Salmonella enterica serovar typhimurium survival in Peyer's patches. *Journal of bacteriology*, 182, 4406-4413.

97. Rohwer, F. and Thurber, R.V. (2009) Viruses manipulate the marine environment. *Nature*, 459, 207-212.

98. Nigam, Y., Bexfield, A., Thomas, S. and Ratcliffe, N.A. (2006) Maggot therapy: the science and implication for CAM part II-maggots combat infection. *Evidence-based complementary and alternative medicine : eCAM*, 3, 303-308.

99. Erdmann, G.R. (1987) Antibacterial Action of Myiasis-Causing Flies. *Parasitol Today*, 3, 214-216.

100. Bohova, J., Majtan, J., Majtan, V. and Takac, P. (2014) Selective Antibiofilm Effects of Lucilia sericata Larvae Secretions/Excretions against Wound Pathogens. *Evid-Based Compl Alt*.

101. Kwiecinska-Pirog, J., Bogiel, T. and Gospodarek, E. (2013) Effects of ceftazidime and ciprofloxacin on biofilm formation in Proteus mirabilis rods. *J Antibiot*, 66, 593-597.

102. Sipahi, O.R., Bardak-Ozcem, S., Ozgiray, E., Aydemir, S., Yurtseven, T., Yamazhan, T., Tasbakan, M. and Ulusoy, S. (2010) Meningitis Due to Providencia stuartii. *Journal of clinical microbiology*, 48, 4667-4668.

103. Klambt, C. (2002) EGF receptor signalling: roles of star and rhomboid revealed. *Current biology : CB*, 12, R21-23.

104. Hughes, D.T. and Sperandio, V. (2008) Inter-kingdom signalling: communication between bacteria and their hosts. *Nature Reviews Microbiology*, 6, 111-120.

105. Wu, L., Estrada, O., Zaborina, O., Bains, M., Shen, L., Kohler, J.E., Patel, N., Musch, M.W., Chang, E.B., Fu, Y.X. *et al.* (2005) Recognition of host immune activation by Pseudomonas aeruginosa. *Science*, 309, 774-777.

106. Galardini, M., Biondi, E.G., Bazzicalupo, M. and Mengoni, A. (2011) CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine*, 6, 11.

107. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, 119.

108. Fouts, D.E., Brinkac, L., Beck, E., Inman, J. and Sutton, G. (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research*, 40, e172.

109. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7, 539.

110. Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, 34, W609-612.

111. Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312-1313.

112. Seal, A., Gupta, A., Mahalaxmi, M., Aykkal, R., Singh, T.R. and Arunachalam, V. (2014) Tools, resources and databases for SNPs and indels in sequences: a review. *International journal of bioinformatics research and applications*, 10, 264-296.

113. Kondrashov, A.S. and Rogozin, I.B. (2004) Context of deletions and insertions in human coding sequences. *Human mutation*, 23, 177-185.

114. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409, 928-933.

115. Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome biology*, 5, R12.

116. Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28, 33-36.

117. Coghlan, P. (2005) The prodigal and his brother: impartiality and the equal consideration of interests. *Theoretical medicine and bioethics*, 26, 195-206.

118. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, 278, 631-637.

119. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34, D32-36.

120. Hill, D.P., Smith, B., McAndrews-Hill, M.S. and Blake, J.A. (2008) Gene Ontology annotations: what they mean and where they come from. *BMC bioinformatics*, 9.

121. Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21, 3674-3676.

122. Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R. and Bolund, L. (2006) WEGO: a web tool for plotting GO annotations. *Nucleic acids research*, 34, W293-297.

123. Kirchner, S. and Ignatova, Z. (2015) Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nature Reviews Genetics*, 16, 98-112.

124. Smit, S., Widmann, J. and Knight, R. (2007) Evolutionary rates vary among rRNA structural elements. *Nucleic acids research*, 35, 3339-3354.

125. Schattner, P., Brooks, A.N. and Lowe, T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic acids research*, 33, W686-689.

126. Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T. and Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*, 35, 3100-3108.

127. Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J. and Wishart, D.S. (2011) PHAST: a fast phage search tool. *Nucleic acids research*, 39, W347-352.

128. Ghiurcuta, C.G. and Moret, B.M. (2014) Evaluating synteny for improved comparative studies. *Bioinformatics*, 30, i9-18.

129. Tonjum, T., Havarstein, L.S., Koomey, M. and Seeberg, E. (2004) Transformation and DNA repair: linkage by DNA recombination. *Trends in microbiology*, 12, 1-4.

130. Sawyer, S. (1989) Statistical tests for detecting gene conversion. *Molecular biology and evolution*, 6, 526-538.

131. Bruen, T.C., Philippe, H. and Bryant, D. (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172, 2665-2681.

132. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100, 9440-9445.

133. Biswas, S. and Akey, J.M. (2006) Genomic insights into positive selection. *Trends in Genetics*, 22, 437-446.

134. Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24, 1586-1591.

135. Yang, Z., Wong, W.S. and Nielsen, R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology and evolution*, 22, 1107-1118.

136. Sahyoun, A.H., Bernt, M., Stadler, P.F. and Tout, K. (2014) GC skew and mitochondrial origins of replication. *Mitochondrion*, 17, 56-66.

137. Casjens, S., Winn-Stapley, D.A., Gilcrease, E.B., Morona, R., Kuhlewein, C., Chua, J.E., Manning, P.A., Inwood, W. and Clark, A.J. (2004) The chromosome of Shigella flexneri bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *Journal of molecular biology*, 339, 379-394.

138. Roncero, C., Darzins, A. and Casadaban, M.J. (1990) Pseudomonas aeruginosa transposable bacteriophages D3112 and B3 require pili and surface growth for adsorption. *Journal of bacteriology*, 172, 1899-1904.

139. Bunny, K., Liu, J. and Roth, J. (2002) Phenotypes of lexA mutations in Salmonella enterica: Evidence for a lethal lexA null phenotype due to the Fels-2 prophage. *Journal of bacteriology*, 184, 6235-6249.

140. Belas, R., Erskine, D. and Flaherty, D. (1991) Transposon mutagenesis in Proteus mirabilis. *Journal of bacteriology*, 173, 6289-6293.

141. Sullivan, N.L., Septer, A.N., Fields, A.T., Wenren, L.M. and Gibbs, K.A. (2013) The Complete Genome Sequence of Proteus mirabilis Strain BB2000 Reveals Differences from the P. mirabilis Reference Strain. *Genome announcements*, 1.

142. Godoy, T.F., Moreira, G.C., Boschiero, C., Gheyas, A.A., Gasparin, G., Paduan, M., Andrade, S.C., Montenegro, H., Burt, D.W., Ledur, M.C. *et al.* (2015) SNP and INDEL detection in a QTL region on chicken chromosome 2 associated with muscle deposition. *Animal genetics*.

143. Leipe, D.D. and Landsman, D. (1997) Histone deacetylases, acetoin utilization proteins and acetylpolyamine amidohydrolases are members of an ancient protein superfamily. *Nucleic acids research*, 25, 3693-3697.

144. Casaregola, S., Norris, V., Goldberg, M. and Holland, I.B. (1990) Identification of a 180kd Protein in Escherichia-Coli Related to a Yeast Heavy-Chain Myosin. *Molecular microbiology*, 4, 505-511.

145. Yeung, A.T.Y., Torfs, E.C.W., Jamshidi, F., Bains, M., Wiegand, I., Hancock, R.E.W. and Overhage, J. (2009) Swarming of Pseudomonas aeruginosa Is Controlled by a Broad Spectrum of Transcriptional Regulators, Including MetR. *Journal of bacteriology*, 191, 5592-5602.

146. Huang, P.Y., Leu, J.H. and Chen, L.L. (2014) A newly identified protein complex that mediates white spot syndrome virus infection via chitin-binding protein. *J Gen Virol*, 95, 1799-1808.

147. Anisimova, M., Nielsen, R. and Yang, Z. (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, 164, 1229-1236.

148. Galac, M.R. and Lazzaro, B.P. (2012) Comparative genomics of bacteria in the genus Providencia isolated from wild Drosophila melanogaster. *BMC genomics*, 13, 612.

149. Tarr, C.L. and Whittam, T.S. (2002) Molecular evolution of the intimin gene in O111 clones of pathogenic Escherichia coli. *Journal of bacteriology*, 184, 479-487.

150. Gueiros-Filho, F.J. and Losick, R. (2002) A widely conserved bacterial cell division protein that promotes assembly of the tubulin-like protein FtsZ. *Genes & development*, 16, 2544-2556.

151. Sauvonnet, N., Gounon, P. and Pugsley, A.P. (2000) PpdD type IV pilin of Escherichia coli K-12 can Be assembled into pili in Pseudomonas aeruginosa. *Journal of bacteriology*, 182, 848-854.

152. Giannakis, M., Chen, S.L., Karam, S.M., Engstrand, L. and Gordon, J.I. (2008) Helicobacter pylori evolution during progression from chronic atrophic gastritis to gastric cancer and its impact on gastric stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 4358-4363.

153. Jordan, A., Aslund, F., Pontis, E., Reichard, P. and Holmgren, A. (1997) Characterization of Escherichia coli NrdH. A glutaredoxin-like protein with a

thioredoxin-like activity profile. *The Journal of biological chemistry*, 272, 18044-18050.

154. Alam, M.J. and Zurek, L. (2004) Association of Escherichia coli O157 : H7 with houseflies on a cattle farm. *Applied and environmental microbiology*, 70, 7578-7580.

155. Sacchetti, P., Ghiardi, B., Granchietti, A., Stefanini, F.M. and Belcari, A. (2014) Development of probiotic diets for the olive fly: evaluation of their effects on fly longevity and fecundity. *Ann Appl Biol*, 164, 138-150.

156. Charles, I.G., Harford, S., Brookfield, J.F. and Shaw, W.V. (1985) Resistance to chloramphenicol in Proteus mirabilis by expression of a chromosomal gene for chloramphenicol acetyltransferase. *Journal of bacteriology*, 164, 114-122.

157. Brown, N.L., Stoyanov, J.V., Kidd, S.P. and Hobman, J.L. (2003) The MerR family of transcriptional regulators. *FEMS microbiology reviews*, 27, 145-163.