COMPUTATIONAL APPROACHES FOR COMPARATIVE GENOMICS AND TRANSCRIPTOMICS USING 454 SEQUENCING TECHNOLOGY

By

VANDHANA KRISHNAN

A thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science in Computer Science

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

AUGUST 2009

To the Faculty of Washington State University:

The members of the Committee appointed to examine the thesis of VANDHANA KRISHNAN find it satisfactory and recommend that it be accepted.

_____

Ananth Kalyanaraman, Ph.D., Chair

_____

Amit Dhingra, Ph.D.

_____

Min Sik Kim, Ph.D.

_____

Cynthia Haseltine, Ph.D.

ACKNOWLEDGMENT

First and foremost I am deeply grateful to my advisors Dr. Ananth Kalyanaraman and Dr. Amit Dhingra without whom this thesis would not have been possible. I thank them for their continued support and patience throughout my research in the interdisciplinary field of computational biology. I would like to thank Dr Cynthia Haseltine for her support, providing me an opportunity to work with her research and serving as a member of the committee. I would like to thank Dr. Min Sik Kim for his constant encouragement and support right from the day I joined WSU and for serving as a committee member.

I am very thankful to my fellow research members particularly Scott Schaeffer, Derick Jiwan, Tyson Koepke and Christopher Hendrickson for their immense support and patience at various stages of the project. Lastly, I would like to offer my heartiest thanks to my parents, extended family members and friends for their everlasting support that gave me the strength that carried me through difficult times.

COMPUTATIONAL APPROACHES FOR COMPARATIVE GENOMICS AND TRANSCRIPTOMICS USING 454 SEQUENCING TECHNOLOGY

Abstract

by Vandhana Krishnan, M.S.
Washington State University
August 2009

Chair : Ananth Kalyanaraman

The development and application of computational tools has consistently played an integral role in the advancement of genomics research, leading to the sequencing, assembly and annotation of hundreds of genomes over the last two decades. However, over the last two years, the research community has embraced an array of new high-throughput, cost-effective technologies, referred to as the "next-generation sequencing technologies", which are poised to alter the landscape of genomics research in the path to accelerated biological discovery. This imminent scientific revolution can, however, happen only if a) new computational tools are developed to cater to the type of data generated by these new technologies; and b) efficient computational frameworks are implemented to seamlessly integrate both new and old tools and enable biologists to answer domain-specific questions. The goal of this thesis is to address the latter need in the context of two projects – one toward the sequencing the genome of a new extremophile microbe and another toward enabling transcriptomics in tree fruit species for

which the underlying genomics is not yet fully understood. Both projects use the 454 sequencing technology, which is one of the next-generation sequencing technologies. More specifically, the contributions of the thesis are in the development and application of computational frameworks that have led to the generation of the first known 454-sequencing based draft assembly of the *Sulfolobus solfataricus* strain 98/2 genome along with its comparative genomics, and to the enabling of 454-sequencing based transcriptomics characterization of apple, pear and cherry. It is expected that the methods and computational frameworks implemented as part of this thesis will have a broader applicability in the near future for projects in comparative genomics and transcriptomics that use next-generation sequencing technologies. The thesis is the result of the collaboration among the WSU Department of Horticulture, WSU School of Molecular Biosciences and the WSU School of EECS.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 3

.

CHAPTER 4

LIST OF FIGURES

APPENDIX

Dedication

This thesis is dedicated to my beloved parents and grandparents for without their strong support I would not have reached this far.

CHAPTER ONE

INTRODUCTION

Genomics is at the forefront of modern day biological research. Hundreds of new genomes are being sequenced from all three domains of life (archaea, bacteria and eukaryotes) with the hope and promise that they will collectively lead to a better understanding of the cellular and molecular mechanisms that govern life forms. Computational tools are at the heart of the genome discovery process and without them none of the genome projects that have been completed over the last decade would have been conceivable. And yet, it is the putting together of a coherent *computational pipeline* and making it work so as to enable scientific discovery in a genome sequencing project that serves as one of the primary impediments in modern day genomics projects. This is due to factors that are both experimental and computational in nature.

Genome sequencing and assembly: Firstly, the experimental technology that is applied for sequencing DNA and generating data for genome assembly is evolving at an exponential scale. Until a couple of years ago, Sanger sequencing (Sanger et al. 1977) was the only experimental platform of choice for DNA sequencing (Shendure et al. 2004, Swerdlow et al. 1990) and with the costs of sequencing substantially high, genome projects had remained the prerogative of a selected few genome sequencing centers. But now, with the advent of the several "next generation sequencing technologies" such as 454 pyrosequencing (Margulies et al. 2005), Illumina/Solexa (Fedurco et al. 2006, Turcatti et al. 2008) and SoLiD (Shendure et al. 2005), sequencing (and hence data generation) has become so cheap and cost-effective that even labs

with modest academic resources are beginning to take part in genomics research. Consequently the software needs to analyze these *new types* of sequenced data have also drastically changed to the extent that biologists who are often untrained computationally are solely dependent on trained computer scientists to fill in for their informatics needs.

Downstream analysis: Secondly, several downstream analysis functionalities post-sequencing and assembly are required to be implemented for deriving species specific (organismal) as well as across-species (comparative) information. These tasks include (but not limited to) finding genes within the sequenced genomes, annotating them both structurally and functionally, characterizing the gene expression and functional space ("transcriptomics"), identifying species specific mutations (single nucleotide polymorphisms), and understanding evolutionary mechanisms at gene and species levels. The number of software tools and options that support these core functions is overwhelming and substantial work is needed to select and integrate these different software options.

Data intensive computing: One of the primary differences between genomics now and ten years back is in the sheer scale of data required to be analyzed. Due to the advent of high throughput next generation sequences, biological sequence databases have grown exponentially every 18 months (Benson et al. 1998 and 2008), and consequently, every new genome project is now faced with the challenge of integrating immense volumes of previously acquired information present in these databases into their pipeline. Developing such an analytical capability requires deployment of high performance computing solutions and hardware platforms in genomics research.

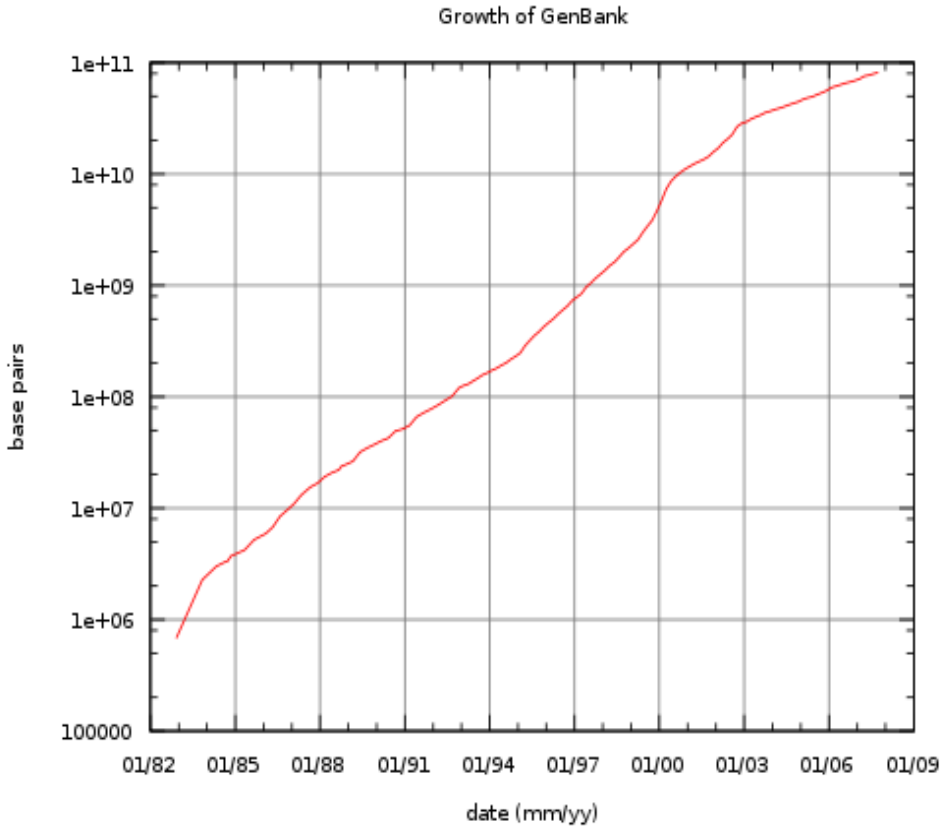


Figure 1.3 Growth of data in GenBank on a semi-log scale indicative of exponential growth (doubling roughly every 18 months). (Source : Veterinary Dictionary: Saunders Comprehensive Veterinary Dictionary 3rd Edition. Copyright © 2007 by D.C. Blood, V.P. Studdert and C.C. Gay, Elsevier.)

The purpose of this thesis research is to address all the emerging concerns outlined above in the context of structural and functional genomics – one toward the sequencing of a new microbial genome (*Sulfolobus solfataricus,* strain 98/2) and another toward enabling comparative transcriptomics in any organism. More specifically, the **contributions of this thesis** are as follows:

i) Development and application of a computational framework that has led to the

generation of the first known 454-sequencing based draft assembly of *Sulfolobus solfataricus* (strain 98/2), structural and functional characterization of the newly assembled genome, and comparative genomics of the 98/2 strain against three other closely related strains.

ii) Design and development of a computational framework for characterizing the transcriptomics of a set of tree fruits within the *Rosaceae* family (apple, pear, cherry) whose genomes are not yet sequenced; such species are called "orphan species".

The work conducted as part of this thesis has culminated in the preparation of two journal articles (pending submission):

Paper #1:

i) **Title** : "454 Sequencing of *Sulfolobus solfataricus* strain 98/2 and comparative analysis with *Sulfolobales*".

Targeted journal : BMC Genomics.

Expected date of submission: July 31, 2009.

Attribution : Vandhana Krishnan, Derick Jiwan and Tyson Koepke performed comparative analysis and genome assembly and mapping. Vandhana Krishnan implemented the computational pipeline with necessary programs and scripts. Derick Jiwan, Tyson Koepke, Scott Schaeffer and Amit Dhingra developed the libraries and performed 454 sequencing. Amit Dhingra and Ananth Kalyanaraman guided data analysis. Vandhana Krishnan, Derick Jiwan, Amit Dhingra, Ananth Kalyanaraman, Cynthia Haseltine and Michael Rolfsmeier prepared the first draft. Andrew Galbraith isolated the DNA. Amit Dhingra, Ananth Kalyanaraman, Michael Rolfsmeier and Cynthia Haseltine supervised the project. All authors read and approved the final manuscript.

Journal format:

The format is as specified below

(http://www.biomedcentral.com/bmcgenomics/ifora/#defaulttype).

The sections to be included if it is a research article are Title page, Abstract, Background, Results, Discussion, Conclusions, Methods (can also be placed after Background), List of abbreviations used (if any), Authors' contributions , Authors' information (if any), Acknowledgements, References, Figure legends (if any), Tables and captions (if any), Description of additional data files (if any).The accession numbers of any kind of sequences is cited using square brackets with the database name specified.

Paper #2:

ii) **Title** : "Enabling comparative and quantitative transcriptome profiling in orphan species."

Targeted journal: Nucleic Acids Research .

Expected date of submission: July 31, 2009.

Attribution :

Vandhana Krishnan developed the computational framework with the necessary programs and scripts. Tyson Koepke, Ananth Kalyanaraman and Vandhana Krishnan developed the algorithm and program to identify an experimentally feasible combination of restriction enzymes for given set of sequences. Scott Schaeffer, Tyson Koepke and Amit Dhingra developed the libraries and performed 454 sequencing. Scott Schaeffer, Christopher Hendrickson, Tyson Koepke and Vandhana Krishnan conducted statistical analyses. Scott Schaeffer performed the q-PCR validation. Amit Dhingra and Ananth Kalyanaraman supervised aspects of the project. Vandhana Krishnan, Amit Dhingra and Ananth Kalyanaraman prepared the first draft.

Journal format:

The format is as specified below (http://www.oxfordjournals.org/our_journals/nar/for_authors/msprep_submission.html). The following sections are to be included in the manuscript : Title page, Abstract, Introduction, Materials and Methods, Results, Discussion, Funding, Acknowledgements, References, Figure Legends. An additional supplementary data section can be made.

Figures 1.1 and 1.2 are an illustrative summary of the different modules developed and applied for achieving the objectives for these two projects. The thesis is the result of collaboration among the WSU Department of Horticulture, WSU School of Molecular Biosciences and the WSU School of EECS. Given the interdisciplinary nature of the thesis, a separate section titled "Biological background" has been added in the appendix to provide the basic molecular biological background that will be required to understand the contents of this thesis.

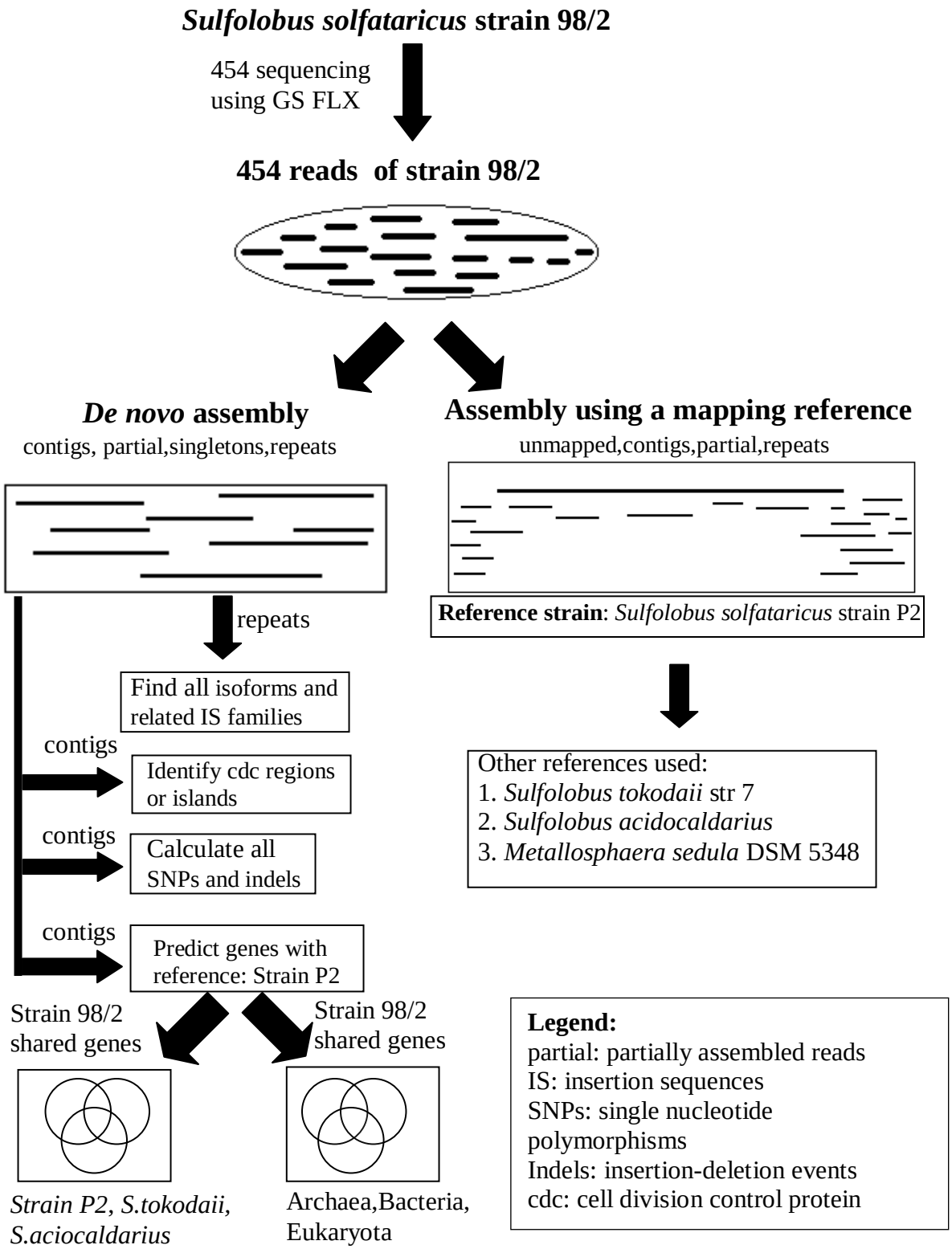


Figure 1.1: Schematic illustration of the overall research workflow for the genome sequencing and comparative analysis of strain 98/2.

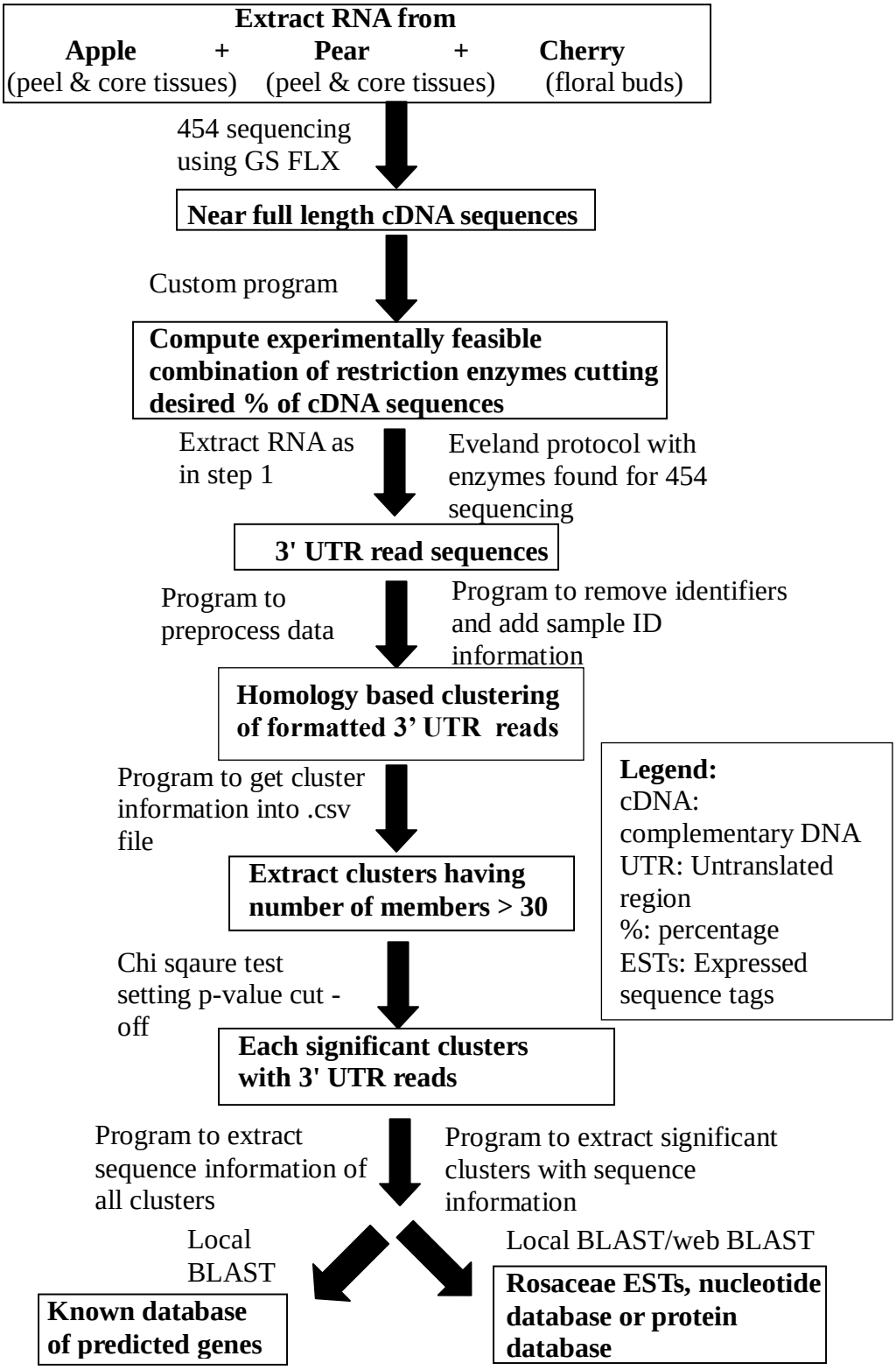


Figure 1.2 Schematic illustration of the overall workflow of trancriptome profiling implemented for tree fruit orphan species.

1.1 Biological Motivation

Comparative genomics of *Sulfolobales*: The three domains of life are Archaea, Bacteria and Eukarya (Woese et. al 1990). A number of archaeal organisms are thermophiles and extremophiles. Extremophiles are organisms that withstand extreme chemical or physical environmental conditions not suited to majority of the living organisms. A thermophile is an extremophile able to withstand high temperature variations not favorable to most living organisms. *Sulfolobus solfataricus* strain 98/2 is an extremophile classified in the domain: Archaea, phylum: Crenarchaeota and order : *Sulfolobales*. *S.solfataricus* strain 98/2 was isolated from the hot springs of the Yellowstone National Park, USA (Brock et. al 1972). The ability of extremophiles to thrive in such harsh environments interests microbiologists and evolutionary biologists to study these organisms more closely. This thesis focuses on comparative genomics of the following members in *Sulfolobales* : *Sulfolobus solfataricus* strain P2, *Sulfolobus tokodaii* strain 7, *Sulfolobus acidocaldarius* and *Metalosphaera sedula.* All of their genome sizes fall within the range of 2 – 4Mbp (Zibat et. al 2005, She et. al 2001, Kawarabayasi et. al 2001, Auernik et. al 2008). Though these organisms may have common characteristics, each of them have their own set of distinct genes encoding different functions, metabolic pathways, and DNA repair mechanisms. Having access to the genome sequence of strain 98/2 can enable comparative genomics studies and highlight the biological processes that make it unique within *Sulfolobales.*

Comparative transcriptomics in orphan organisms: In plant sciences, *Arabidopsis thaliana* has served as a model system for most biological studies so far. However, *Arabidopsis* is an annual plant with a much smaller and simpler genome than many other plant genomes including rice and maize.. Since gene expression varies significantly between different plant species, a model

plant such as *A.thaliana* is not sufficient for unraveling all physiological processes specific to other crops. Apple, pear and cherry, all of which belong to the *Rosaceae* family, constitute a class of economically important tree fruit crops. Apple and pear belong to the same subfamily, *Maloideae,* but cherry is from another subfamily, *Prunoideae*. To date no genomes have been fully sequenced in the *Rosaceae* family. Their genomes are complex and size estimates range from 750 Mbp in apple, 500 Mbp in pear to 250 Mbp in cherry. Due to the absence of a sequenced model organism apple, pear and cherry are termed "orphan species". Apples, pears and cherries are some of the most widely produced and consumed fruits in the world markets. Washington State ranks number one with respect to the production of fruits belonging to the *Rosaceae* family. In particular, apple production in Washington state is about 42%, pear and cherry represent 58% of the total US production respectively (NAAS). Washington state alone has farmgate value of $8 billion (NAAS) for the production of these three fruits. In addition, these trees represent unique biology. The trees exhibit different ploidy levels indicating complex genomic organization, perennial habit, juvenility, various fruit types like pome (apple and pear), stone fruit (cherry), different growth habits to name a few. It is therefore essential to study *Rosaceae* genes and the resulting economically and nutritional traits generated by their products. An in-depth understanding of important physiological traits in these orphan species will prove not only beneficial to growers and breeders, but also to consumers at the other end who look for traits like taste, shelf life, etc.

The central dogma of molecular biology (Crick et. al 1958, 1970) is that DNA is transcribed into RNA which in turn is translated into a protein. Every cell has messenger RNA (mRNA) that represents of the functional unit of a gene. "Transcriptome" is the collective term used to refer to the set of all mRNA transcripts that are expressed in a particular organism. The

mRNA of eukaryotic organisms has a cap region, 5' UTR (untranslated region), start codon, coding sequence, stop codon, 3'UTR followed by a polyA tail.

The genomic content of each cell is identical in the number of genes present in all cells and tissues. However, different cell types show different gene expression patterns. Studying the transcriptomics, therefore, would provide a deeper insight to the differentially expressed genes and corresponding traits exhibited in a particular species. As opposed to genome sequencing, transcriptome sequencing focuses on the different genes expressed in an organism. The study conducted in this thesis focuses on the 3'UTR region of mRNAs, as this region has been experimentally proved to consist of elements responsible for regulating translation efficiency and mRNA stability. In our method, we will generate transcriptome profiling using the 454 sequencing technology. Through this method, we will address the following questions related to peel and core tissues of apple and pear, and developing floral buds in cherry:

1) To understand the differentially expressed genes in the peel and core tissues of Golden delicious and Honey crisp apple varieties.
2) To understand the ripening mechanism using the peel and core tissues of D' Anjou and Bartlett pear varieties.
3) To understand the effect of rootstocks on the floral bud numbers of scions, such as Rainier and Bing, at different time points.

1.2 Computational Motivation

Comparative Genomics of *Sulfolobales*: The common method of sequencing bacterial genomes involve host organisms or clones. With the advent of high-throughput technologies such as the *de*

novo 454 pyrosequencing, cloning has become no longer necessary. Molecular biology techniques such as northern blots , differential display, micro-arrays, serial analysis of gene expression(SAGE), Q-PCR are not suitable to identify multiple genes and their expression simultaneously. The drawbacks of other sequencing methods for gene expression studies have been addressed in detail in Chapter 4.

One of the primary challenges in embracing these next generation high-throughput sequencing technologies for genomics and transcriptomics studies stems from the computational side. Given that these technologies have been adopted in the market only over the last couple of years, computational tools and protocols for performing the data analysis on the data generated from these machines hardly exist. Available software options only cater to old technologies such as Sanger sequencing and implement functionalities for downstream analysis in a genome project (e.g., gene finding, repeat identification, etc.). For this project, however, we are dealing with a) a new genome to be sequenced directly using the 454 sequencing technology, and b) analysis of transcriptome data from orphan species (i.e., those without already sequenced model organisms) – thereby necessitating the development and implementation of new, effective computational pipelines that would allow us to answer the scientific questions posed in this thesis.

The goal of this thesis is to implement coherent computational pipelines that serve the analytical needs for these two projects. This is achieved by incorporating an array of already developed software tools and combined with custom written programs and scripts that ultimately implement an automated pipeline. The tools and pipelines implemented as part of this thesis have a wider applicability to a range of projects in comparative genomics and transcriptomics that use the 454 sequencing technology.

1.3 Organization of thesis

Chapter 2 discusses the related work. Chapter 3 is the manuscript on the sequencing and comparative analysis of the bacterial genome of *Sulfolobus solfataricus* strain 98/2.

Chapter 4 is the manuscript of transcriptome profiling in orphan species. Finally, Chapter 5 presents the conclusions and future work. An appendix has been added at the end that provides the required biological background for understanding the thesis.

CHAPTER TWO

RELATED WORK AND RELEVANT LITERATURE

This chapter explains how data from 454 sequencing technology was used for identifying the structure of the *Sulfolobus solfataricus* strain 98/2 genome and also for the transcriptome profiling of apple, pear and cherry fruits of the *Rosaceae* family. The underlying software modules used and the need for them is explained. The drawbacks of related tools that were not used at various stages of the project are also listed in this chapter. Also, the necessary scripts written for parsing data output by various software modules are given in detail. The actual parameters used while running the necessary tools have been discussed in Chapter 3 and Chapter 4 respectively.

2.1 The 454 Sequencing Technology

The common method of sequencing bacterial genomes or a transcriptome involves cloning and propagation of genomic pieces or cDNAs in a heterogeneous host. Similarly, transcriptome profiling has been done using northern blots, differential display, microarrays, serial analysis of gene expression(SAGE),etc. The various experimental complications and time constraints in all the above prevalent methods have been circumvented with the advent of technologies (Shendure and Ji, 2008) such as the *de novo* 454 high-throughput sequencing provided by 454 Life Sciences Corporation.

The Genome Sequencer FLX system (GS FLX) provided by Roche Inc. is based on sequencing by synthesis that employs pyrophosphate as one of the main ingredients and is thus termed as pyrosequencing. The photons emitted from the release of pyrophosphate bonds

releases energy in the form of light which is captured by a high resolution camera. Utilizing these images, the software does the base calling to produce read sequences of the DNA fragments in each well of a PicoTiterPlate (Genome Sequencer Data Analysis Software Manual, 2008). The length of the individual reads is approximately 100bp -200bp or 200bp – 300bp (used in sequencing *Sulfolobus solfataricus* strain 98/2) depending on the sequencing platform used. The average length of the strain 98/2 read sequences output by the sequencer was 228.78bp and the longest length of a read sequence was 358bp. Similarily, the average length of the apple transcriptome raw read sequences was 83.69bp, pear trancriptome raw read sequences was 82.54bp and that of cherry trancriptome was 84.94bp. The longest length of the trancriptome raw read sequences for apple, pear and cherry were 188bp, 214bp and 507bp respectively. Since, the GS FLX system had assemblers customized for accepting its own data we chose the same. Further reasoning on choosing the GS *De novo* Assembler application (gsAssembler) and GS Reference Mapper application (gsMapper) are explained in the following section.

2.2 Genome Assembly and Reference map applications

Genome assembly is the problem of assembling the nucleotide sequence of a genome from a set of reads sequenced from it using one or a combination of sequencing technologies. The computational problem is analogous to the problem of assembling a jigsaw puzzle from its pieces except that here the final picture of the puzzle (genome) is not known until construction. There are numerous genome assembly software programs (list not exhaustive): CAP3 (Huang and Madan 1999), Arachne (Batzoglu et. al 2002), Phrap (Green et. al 1996), Celera assembler (Myers et. al 2000) and EULER (Chaisson et. al 2004), to name a few. However, all these programs were built for reads from Sanger sequencing (i.e., ~1Kbp) and the genome coverage while initial sequencing is less than 10X. If the next-generation sequencing technologies such as

454, Solexa or SOLiD are used, then the read lengths are much shorter (anywhere from 50 to 400bp) and the genome coverage during sequencing is typically much larger than 10X because the sequencing is both cheap and high throughput. These two factors complicate the algorithms required thereby necessitating new type of software tools that cater more specifically to these "short reads" generated from these high throughput facilities.

There are assembly software such as gsAssembler and gsMapper Genome Sequencer Data Analysis Software Manual, 2008) which are specifically built to analyze data from the 454 sequencer. Moreover, the parametric study on the gsAssembler (in the conclusion section of the thesis) did not give a better set of parameters in case of this project. The resulting contig files from both the gsAssembler and gsMapper are in FASTA format and enables compatibility with other third party tools or softwares for anlaysis. The 454 sequencer generates a set of standard flowgram files (.sff) files that contain quality scores and base calls for all high quality reads in a run. Here, the length of a homopolymer in a sequence is determined. Then, using the lengths of subsequent homopolymer stretches a based called sequence is derived (Genome Sequencer Data Analysis Software Manual, 2008). Thus, this process differentiates the individual base call measurements in other sequencing technologies like Sanger technology. These .sff files are used as input for the gsAssembler and gsMapper to form contigs of better quality. The *de novo* assembler creates assemblies from the GS FLX generated reads using selected .sff files.

Paired end runs and reads in FASTA format from Sanger sequencing can also be incorporated into assemblies for forming better contigs and scaffolds. The Reference Application is used to align reads from one or several sequencing runs and using a reference sequence to form consensus sequence or contigs. The read information is obtained from the corresponding .sff files of interest. Sanger reads can also be incorporated during analysis.

2.3 Comparative analyses of the archaeal genome

This section deals with all the tools used in the analysis of *Sulfolobus solfataricus* strain 98/2 genome and the reason for choosing them. The related programs/scripts are also described in this section.

2.3.1 Finding the GC content

An important aspect of all genomes is the GC (guanine – cytosine) content as it gives an idea about the organisms' evolution. The GC content is ratio of the total number of guanine and cytosine bases to the total number of adenine, guanine, cytosine and thymine bases. Several tools were tested to find GC content of the contigs formed by the gsAssembler for strain 98/2 genome.

"CpG Ratio and GC Content Plotter" (http://mwsross.bms.ed.ac.uk/public/cgi-bin/cpg.pl) is a tool that did not accept the entire set of 509 contigs generated by gsAssembler as input. Since the tool's sequence limit could not accommodate our contig size, the overview of the GC plot of the genome could not be obtained. "GC Content/GC Skew Diagrams" tool (http://nbc11.biologie.uni-kl.de/framed/left/menu/auto/right/GC/) gives a list of positions and GC content values based on the window step specified. These need to be imported into an Microsoft Excel application where the graph can be plotted and using Excel functions, the average GC content can be computed. The tool was not useful as it did not directly provide a profile of the GC content across the genome. Using BioPerl modules such as 'Bio::Graphics::Glyph::generic' for creating a glyph involves installation of related libraries and specifying the start-end positions of the DNA sequences in the script, etc. Thus, GC profile tool (Gao et. al 2006) was chosen to overcome the above limitations. It gives the cumulative GC profile with segmentation points

and also a GC plot of the entire set of assembler's contigs. The GC plot was not used as the graph obtained was not helpful due to the numerous gaps in the draft genome whose locations are unknown. The segmentation points indicate the possible islands or origins of replication present in a bacterial genome. The direct number for the GC content in strain 98/2 was given by the GeneMark tool (Lukashin and Borodovsky, 1998 , Besemer and Borodovsky, 2005) and was used later to predict genes/ORFs.

2.3.2 Identifying Single Nucleotide Polymorphisms and indels in strain 98/2

Detection of Single Nulcleotide Polymorphisms (SNPs) is important for the understanding of genome level differences among evolutionarily related species because of high sequence identity among them. So is the case with insertion-deletion (indels) events occurring in different members of the same species.

SNPs are the single nucleotide differences seen during comparison of two DNA sequences and occur in 12 forms such as AT, TA, GC, CG, AG, GA, GT, TG, TC, CT, AC and CA. and. This means in a specific fragment of DNA sequence identical between two genomes, one of them has a base pair A for which corresponding position in the second genome is a base pair G. Indels are like SNPs except that for a missing base pair in one genome there is corresponding base pair (A or G or T or C) in the second genome. In other words a deletion event in one genome is an insertion event in the other.

SNPs Finder (Song et al. 2005) is a tool used to detect SNPs and indels in genomes of microbes. It lists all the types of SNPs with respect to the reference genome and a link for each SNP found showing the aligned sequences with the site of a SNP/indel marked. A disadvantage

of this output format is that there are no direct numbers given for the kind of SNPs or indels identified.

MUMmer (Delcher et al. 1999, 2002 and Kurtz et. al 2004) was very quick in generating the output in the form of percentages of each type of SNP and indels detected. Hence, MUMmer was chosen to predict the different kind of SNPs and indels in strain 98/2 using the contigs generated by gsAssembler. MUMmer did not report 'GT' SNPs so a custom script was written to process the ouput file generated by MUMmer to obtain the number of 'GT' SNPs.

2.3.3 Sequence similarity searches

At many steps in the analyses, BLAST (Basic Local Alignment Search Tool) has been used for sequence similarity searches. BLAST (Altschul et. al 1990) is one of the well established and extensively used tool in sequencing. The input sequence or "query" is searched for matches against a nucleotide database or a protein database. The query sequence can be either a DNA or a protein sequence. If a DNA sequence is queried against a protein database then the query sequence is translated into all its 6 reading frames before comparison. Both web BLAST (NCBI site: http://blast.ncbi.nlm.nih.gov/Blast.cgi) and local BLAST were used. Local BLAST is useful especially when custom database specific to a project is to be used. It is also useful when the input sequences are very large (sequence limit on the web server), to obtain results faster. For instance, the assembler's contigs of strain 98/2 were aligned to against strain P2 genome and other *Sulfolobales*.

Also, in the case of finding genes in apple, pear and cherry common to other identified plant genes local BLAST searches were done. To find the predicted ORFs of strain 98/2 that match the genes in archaea, bacteria and eukaryota NCBI web BLAST was used.

2.3.4 Identifying insertion sequence elements

Insertion sequences (IS) are genetic mobile elements that change positions within the genome. They are found to occur in most bacterial genomes (Chandler and Mahillon 2002). IS elements are embedded within transposons (found in eukaryotes like maize). In bacteria, IS elements contain only one gene encoding the transposase enzyme (Genes VIII, Lewin) responsible for mobility of the sequence. Generally the length of IS elements varies between 700bp – 2500bp and approximately 21 families have been classified (http://www-is.biotoul.fr/is.html). Many IS elements are known to activate neighboring genes in terms of their expression (Galas and Chandler 1989).

In order to identify the insertion sequences in strain 98/2 genome we need to find those repeat elements in the genome. The procedure is described in Chapter 3 methods section. To find the type of IS elements, we need to have a database of already known elements for which IS finder tool (http://www-is.biotoul.fr/is.html) is the only one available. The BLAST option against the IS database present in IS finder tool was used to identify ISs in repeat sequences of strain 98/2 as described in Chapter 3.

2.3.5 Predicting genes in strain 98/2

Several tools like EasyGene (Larsen and Krogh et. al 2003), Glimmer (Delcher et. al 1999, 2000), GeneMark.hmm (Lukashin and Borodovsky, 1998 , Besemer and Borodovsky, 2005), Genscan (Burge et. al Burge and Karlin 1997,1998; Burge et. al 1998) etc, that predict genes from DNA sequences exist. GenScan (http://genes.mit.edu/GENSCAN.html) can predict the exon- intron stretches in a genome based on 3 model organisms listed in the tool, namely

vertebrate, *Arabidopsis* and maize. These criteria disable it from being used for bacterial gene prediction. WebGene (Milanesi and Rogozin 1998; Milanesi et. al 1999) incorporates two tools CpG and GeneBuilder used in gene prediction and CpG islands identification respectively. However, both these tools cater to complex mammalian genomes proving to be of no use in this project. GeneAlign (Hsieh et. al 2005) is a tool used for prediction and alignment of coding exons but is customized to human and mouse genomes making it unsuitable for our purpose. EasyGene 1.2b on the other hand is specifically used for prokaryotes. EasyGene has 138 model organisms to choose from, *Sulfolobus solfataricus* was one making it desirable. Each contig number (as in assembler's contigs of strain 98/2) has predicted coding sequence (CDS) or suboptimal coding sequence (CDSsub) along with start-end positions in the sequence, orientation of the strand(forward or reverse), predicted start codon and a score given as output. Glimmer (version 3.02) is tool designed to locate genes in bacterial, archaeal and viral genomes. Also, the topology of the genome – linear or circular can be selected as a parameter. Glimmer lists all the input contig headers and if present the predicted ORFs in each of them, start and end of the frame within that contigs, frame number with orientation in that contigs and score. A common disadvantage in both Glimmer and EasyGene is individual tracing of the ORF sequences required to provide more information to biologists.

GeneMark.hmm (version 2.4) handles prokaryotic sequences to predict genes and also has *Sulfolobus solfataricus* as an option for model organism. It gives GC content of the input set of sequences and number of genes predicted directly. For each predicted gene , the strand it is located on, corresponding start and end positions, length of the gene and gene class are specified. The gene class indicates which Markov model was used in prediction. In addition to all these, it indicates:

1. possible ORFs forming CDSs with orientation, frame number and start-end postions.
2. possible frameshifts with starting base position.
3. protein translation of the genes predicted.
4. nucleotide sequences of the genes predicted.

The above mentioned features make GeneMark.hmm ideal to predict the genes present in *Sulfolobus solfataricus* strain 98/2.

2.4 Analyses of transcriptomes

454 sequencing has been extensively used for transcriptome work (Morozova and Marra et. al 2008). This section describes the tools used for analyses of the transcript datasets belonging to apple, pear and cherry. Also, new biological methods adopted and related computational framework developed for extracting useful information from transcriptomes of these orphan species has been discussed.

2.4.1 Restriction enzyme analysis

As mentioned earlier, the established protocol (Eveland et.al 2007) for transcriptome profiling using 454 reads forms the basis for our project. In the existing protocol, the enzyme *MspI* was found to digest at a site approximately 100 bp from the 3' end of UTR (Eveland et.al 2007). The *Msp*I enzyme's cut site was absent in most of the sequences generated in the project. It is not feasible to experimentally test several restriction enzymes against hundreds of thousands of sequences. Moreover, no computational methods exist for calculating the best enzyme or set of enzymes that cleaves a given set of sequences.

NEB cutter (Vincze et. al 2003) and Watcut (http://watcut.uwaterloo.ca/watcut

/watcut/template.php?act=restriction_new) are two tools used for restriction analysis of given DNA sequences but they accept sizes of upto 1Mb and 50Kb respectively. Hence, they cannot be used for huge set of DNA sequences such as the 454 reads we have. 'Restriction enzyme digest of DNA' (http://insilico.ehu.es/restriction/main/)was one tool found but it catered only to completely sequenced bacterial genomes.

Restriction enzyme analysis was enabled through a custom script that finds the restriction enzymes that cut a desired percentage of given input sequences. An enzyme restriction table file contains information of whether an enzyme cuts a given sequence (value= 1) or not (value =0) in a binary format. The table has information for all input raw read sequences and chosen set of enzymes. This binary table is modified using Microsoft Excel application to find the required number of enzymes whose cut percentage is chosen by the experimenter. Further details on this script are included in Chapter 4 methods section.

2.4.2 Parsing the 3' UTR 454 reads

Using the combination of enzymes, a 454 sequence run is performed to obtain the 3' UTR sequences for each apple, pear and cherry using the existing transcriptome protocol.
A preprocessing script is run on the data to format the header information and bring the sequences to a single line below its header. This makes it easy for further analysis of the sequences. A set of custom scripts have been written to process the 454 reads obtained, the details of which are included in Chapter 4. When samples are sequenced in a single run in the sequencer, multiplexing identifiers (MIDs) are used to differentiate among them. For instance golden delicious apple peel, golden delicious apple core, honey crisp apple peel and honey crisp apple core samples can be sequenced in a single run using MIDs.

These MID sequences are provided by Roche Inc. and during assemblies using gsAssembler are removed by the application. Since, we do not require assemblies at this stage; the need to filter these MIDs using a custom script is the only solution which otherwise would have been filtered by gsAssembler. Also, if different samples or different time points of the same sample are being observed, the corresponding information is also added to the headers of each sequence set in order to identify them later with the use of custom programs.

In certain cases, instead of MIDs a large number of primers are used that are designed to specifically bind to certain sequences. A similar custom program was written to filter the primers and separate sequences belonging to each primer. More details on the above programs can be found in Chapter 4.

2.4.3 Clustering based on unique transcripts using PaCE

Our main aim is to identify unique transcripts and this in turn can give information on different alleles present in a species. Different forms of a single gene are termed alleles of that gene. For example, a single gene can be responsible for the height of a certain plant species but few plants are tall and some short. This is because tall and short are two alleles of the gene responsible for height and are expressed differently within the members of the same species.

In order to identify such unique information we discard the process of assembly and use the tool PaCE (Kalyanaraman et. al 2003) instead. PaCE stands for Parallel Clustering of ESTs (Expressed Sequence Tags) and is used on large sets .ESTs are portions of cDNAs used to identify transcripts thereby discovering genes and their corresponding sequences. The main aim of the algorithm used in PaCE is to reduce run time and memory overheads encountered while using assemblers or EST clustering tools such as Phrap (Green et. al 1996) , TIGR assembler (Sutton

et. al 1995) and CAP3 (Huang et. al 1999).

Now, we have the sequences belonging to each PaCE cluster and these are separated out into a .csv file containing information of how many members are present and the frequency of each member.

2.4.4 Statistical Analyses using Chi-square tests

The comma separated file obtained by parsing the information from PaCE output file is used further for some statistical analysis. There are numerous tests to determine the 'goodness of fit' for a set of different values for some given variables. Of these, the Chi square test is the most broadly accepted. This test is used when one has two nominal variables (can be used for greater than nominal values but tends to be complex and is out of scope with respect to this project).

In apple and pear there will be two independent chi-square tests in each resulting in a total of four chi-square tests given as follows:

1. Comparison of two nominal variables that are transcript frequencies of peels in Golden Delicious and Honey Crisp varieties for each cluster.
2. Comparison of two nominal variables that are transcript frequencies of cores in Golden Delicious and Honey Crisp varieties for each cluster.
3. Comparison of two nominal variables that are transcript frequencies of peels in Bartlett and D'Anjou varieties for each cluster.
4. Comparison of two nominal variables that are transcript frequencies of cores in Bartlett and D'Anjou varieties for each cluster.

In cherry there will be a total of four independent chi-square tests given by:

1. Comparison of two nominal variables that are transcript frequencies of Gisela rootstock

in Rainier and Bing varieties for each cluster at time point 1.

2. Comparison of two nominal variables that are transcript frequencies of Mazzard rootstock in Rainier and Bing varieties for each cluster at time point 1.
3. Comparison of two nominal variables that are transcript frequencies of Gisela rootstock in Rainier and Bing varieties for each cluster at time point 2.
4. Comparison of two nominal variables that are transcript frequencies of Mazzard rootstock in Rainier and Bing varieties for each cluster at time point 2.

We can thus determine the deviation of the observed frequencies from the expected frequencies. Once, we obtain the chi- square values for the above datasets, we need to calculate the p-value corresponding to each chi- square value. The p-value gives the probability of the proportion of transcript type considered occurring by chance and not affected by other factors. The p-value for a chi-square number can be obtained using the universal chart of chi-square numbers and p-values or online statistical calculators (Soper et.al 2009). The p-value can be used to determine the "significant" clusters in the experiment which should be the focus for further analysis through gene expression levels.. A cut off value for extracting "significant" clusters was determined as follows: First, the density of the chi square value distribution was plotted for each individual test. When this plot was converted to a normal curve, the resulting plot showed that the experimental values were not close to normal values. The experimental values were then transformed using Box-Cox test and resulting graph again re-plotted as for a normal distribution. As a preliminary experiment, we computed a graph for one of the sample types in cherry. The graph shows the property of Kurtosis (see Figure 2.1).

Kurtosis is the property by which the normal distribution curve shows a high peak and smaller

tails, as most values are cluttered around the mean. This could result in hardly any clusters to be deemed “significant”, even for a pvalue cut off of 0.05.

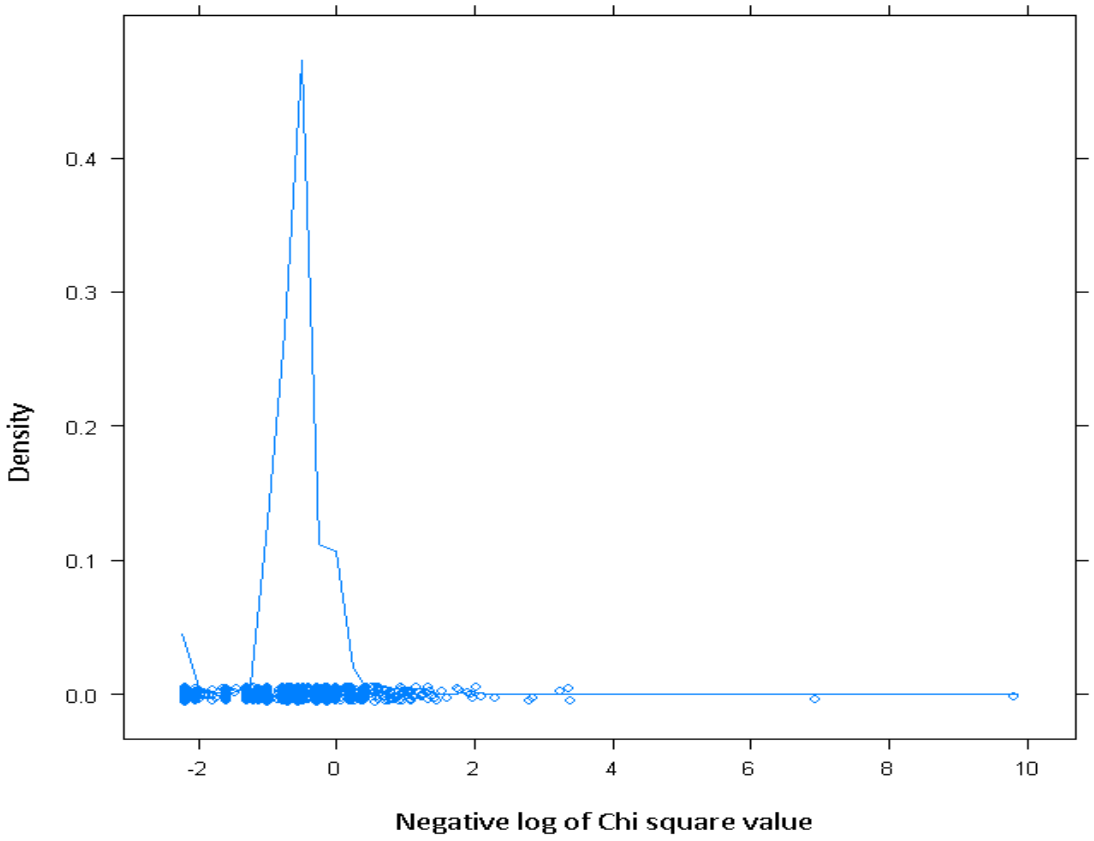


Figure 2.1 Preliminary results of the graph obtained for a sample type in cherry showing Kurtosis.

It is known that the lower the p-value the more significant the cluster is. Moreover, chi-square test is more accurate for larger sample sizes. Using the above mentioned facts, the clusters were sorted according to their sizes (total number of members in a cluster). Then, we filtered the small clusters having sizes less than 30. This reduction of clusters based on their size is essential to obtain a correct result when using chi-square test. The remaining set of

clusters was sorted based on their p-values in an ascending order. Further, we picked a p-value threshold observing the trend in the data set under consideration. The actual numbers of the significant clusters in each of the fruit samples can be obtained from Chapter 4.

CHAPTER THREE

454 SEQUENCING OF *Sulfolobus solfataricus* STRAIN 98/2 AND COMPARATIVE ANALYSIS WITH *SULFOLOBALES*

Vandhana Krishnan[1§], Derick Jiwan[2§], Tyson Koepke[2], Scott Schaeffer[2], Ananth Kalyanaraman[1], Michael Rolfsmeier[3], Andrew Galbraith[3], Amit Dhingra[2*] and Cynthia Haseltine[3*]

[1]School of Electrical Engineering and Computer Science, Washington State University, Pullman, USA

[2]Horticulture and Landscape Architecture, Washington State University, Pullman, USA

[3]School of Molecular Biosciences, Washington State University, Pullman, USA

§ These authors contributed equally to the work

* Corresponding authors

Email addresses:

VK: vandhana_k@wsu.edu

DJ: djiwan@wsu.edu

TK: tkoepke@wsu.edu

SS: smschaeffer@wsu.edu

AK: ananth@eecs.wsu.edu

MR: mrolfsmeier@wsu.edu

AG: agalbraith@wsu.edu

AD*: adhingra@wsu.edu

CH*: chaseltine@wsu.edu

Abstract

Background: *Sulfolobales* is an order within the archaea, a branch of life that is distinct from eukaryotes and bacteria. *Sulfolobus solfataricus* strain 98/2 is an important organism within this order, since it serves as the primary genetic background for a growing number of molecular investigations into the physiology and lifecycle of hyperthermophilic crenarchaeota. However, lack of a sequenced genome for strain 98/2 is an obstacle to further research.

Results: Using 454 GS FLX technology, we generated approximately 65.4 Mbp of sequence information that were assembled into 509 contigs representing 2.68 Mbp of genomic sequence from the *S. solfataricus* strain 98/2. The data are hereby released [NCBI Project ID: 33857] together with comparative genomic analysis with other members of *Sulfolobales*. *S. solfataricus* strain 98/2 has a GC content of 35.43%, with 3055 predicted open reading frames (ORFs). Based on results obtained from gsMapper, the strain 98/2 genome has 84.97% sequence identity with the previously sequenced *S. solfataricus* strain P2. Compared to strain P2, strain 98/2 contains several unique IS elements and 31 unique ORFs.

Conclusions: The addition of the *S. solfataricus* strain 98/2 genome to the set of sequenced *Sulfolobales* adds significant phylogenetic and genomic resources to the crenarchaea, and is expected to support further investigations into several cellular mechanisms including transcription, gene regulation, central metabolism, and DNA damage repair.

Background

Sulfolobus solfataricus strain 98/2 was isolated from the hot springs of Yellowstone National Park, USA [1]. This hyperthermophilic acidophilic aerobe has an optimal growth

temperature of 80˚C and grows in conditions ranging from pH 2 – 4. Like *Sulfolobus solfataricus* strain P2, it can grow both chemolithoautotrophically using sulfur as an energy source and heterotrophically with defined carbon and energy sources [1-4]. *S. solfataricus* species are unique in containing both eubacterial and eukaryotic gene homologues. Thus they have become popular archaeal models for studying numerous cellular processes, including transcription, gene regulation, central metabolism, and DNA damage repair. Strain 98/2 is particularly important since it is the only *Sulfolobus* strain that has been used for both targeted gene disruption and allele replacement through homologous recombination [5, 6], establishing its value as a model organism for archaeal hyperthermophiles. Comparative analyses of the strain 98/2 genome with those of other members of the order *Sulfolobales* contributes insight into evolution of these microbes and helps to identify variations between individual crenarchaeal branch members.

All currently available *Sulfolobales* genome sequences (*S. solfataricus* strain P2, *Sulfolobus acidocaldarius*, *Sulfolobus tokodaii*, and *Metallosphaera sedula*) were determined using traditional Sanger sequencing techniques requiring the generation of clones and propagation of DNA fragments in a heterologous host [7-11]. All but one of these organisms has a relatively low GC content, ranging from approximately 33 – 37% [7-9]. With the exception of *S. acidocaldarius,* these archaea contain numerous repeat sequences that appear to be insertion sequence elements (IS) [8]. Here we report the genome sequence of *S. solfataricus* strain 98/2 generated using 454 GS FLX sequencing platform. This approach used a single preparation of genomic DNA and did not require the gene cloning that is required for Sanger sequencing and we found that most regions of the genome were adequately represented without any cloning bias.

Results and Discussion

454 sequencing and genomic features of strain 98/2. Purified total genomic *S. solfataricus* strain 98/2 DNA was sequenced using 454 GS FLX platform. A total of 285,931 sequence reads, with an average read length of 228.78 bp and a total length of 65,415,544 bases were obtained. The sequences were assembled with gsAssembler, using default assembly parameters (Roche Inc.). A minimum overlap length of 40 bp and a minimum overlap identity of 90% were used for assembly and the contigs obtained were used for all subsequent analyses. The draft assembly of the genome has an approximate size of 2.68 Mb as computed using gsAssembler (default parameters) and with strain P2 as a reference; overall draft sequence coverage was calculated to be approximately 21X. The assembled size excludes repeat regions represented mainly by IS elements. *De novo* assembly of the reads without a reference genome resulted in 509 contigs **(Table 1)**, with an average contig length of 5.2 Kbp. When the *S. solfataricus* strain P2's genome was used as a reference, the gsMapper generated 497 contigs and identified a considerable number of repeats, partial sequences, and unmapped regions **(Table 1)**. This difference provides evidence that the two *S. solfataricus* strains are not identical. The N50 value for both assembly approaches has been listed (**Table 1**).

GeneMark [12, 13] predicted the GC content to be 35.43% for the 509 contigs generated using gsAssembler. This percentage is considerably lower than the 46% GC content reported for *M. sedula* [10], but within the range of the other sequenced *Sulfolobales*. Segmentation points of differential GC content were seen at three positions **(Figure 1)**, at 518,818 bp, 523,711 bp, and 2,410,038 bp, and may represent genomic islands [14]. The entire set of contigs from gsAssembler was aligned to cdc6 domains 1, 2 and 3 using BLAST [15]. Based upon the observed sequence identity, three cdc6 regions in 98/2 strain corresponding to the segmentation

points were retrieved. Contig number 00119, 00208, and 00291 represented complete cdc6 domains 1, 2 and 3 respectively.

Comparative genomic analyses with other members of *Sulfolobales.* Previously published genome sequences from four other *Sulfolobales* were compared to *S. solfataricus* strain 98/2 sequence using gsMapper (Roche Inc.) **(Table 2)**. 84.97% of sequences from *S. solfataricus* strain 98/2 mapped onto the *S. solfataricus* strain P2 genome. Only 2.15%, 0.76%, and 0.41% of *S. solfataricus* strain 98/2 sequences mapped to *S. tokodaii*, *S. acidocaldarius*, and *M. sedula* genomes, respectively.

Since *S. solfataricus* strain 98/2 appears to be closely related to *S. solfataricus* strain P2, we performed global sequence alignment of two strains using MUMmer **(Table 3)** [16-18]. The breakpoints were found to be different between the two strains with 372 defined translocations specific to strain P2. Breakpoints are defined as large regions of sequence bounded by gaps on either side and are generally similar between the two strains. The number of breakpoints may be somewhat different if repeat elements are integrated in their biological position in the genome. Insertions, representing sequences present in one genome but absent in the other genome, were observed to be different between the two strains. When compared with *S. solfataricus* strain P2, a total of 412 indels were identified in *S. solfataricus* strain 98/2 **(Table 4)** and number of single nucleotide polymorphisms (SNPs) was also calculated. Transition mutations (A←→G and T←→C) are far more commonplace than transversion mutations, indicating close evolutionary relation between the two strains. The calculated R value (number of transitions/transversions) is 2.9729 **(Figure 2)**. The genome of *S. solfataricus* strain 98/2, like the closely related strain P2, harbors multiple insertion sequence (IS) elements. We found 159 distinct IS element sequences that together comprise approximately 5% of the draft genome, which is nearly half the

percentage of IS element genome occupation (approximately 11%) observed in *S. solfataricus* strain P2 [8]. While there are fewer IS elements, a large number of these have not been previously reported in either *S. solfataricus* strain P2 or *S. tokodaii* **(Table 5)**.

GeneMark [12, 13] predicted a total of 3,055 ORFs in *S. solfataricus* strain 98/2 using strain 98/2 contigs generated by gsAssembler, 234 fewer ORFs than reported for *S. solfataricus* strain P2 by GeneMark. The number of predicted ORFs in *S. solfataricus* strain 98/2 is very similar to that reported for *S. tokodaii* (3,015), but more than that of *S. acidocaldarius* a (2,344) **(Figure 3A)**. The software predicted protein translated regions for 3053 ORFs in strain 98/2, 3287 ORFs in strain P2, 2344 ORFs in *S. acidocaldarius*, 3015 ORFs in *S. tokodaii* and these were used for all future comparative studies. Further examination of these predicted ORFs indicates that there are a number of unique sequences in *S. solfataricus* strain 98/2 (**Table 6**). The predicted ORFs in the *S. solfataricus* strain 98/2 genome common to ORFs of other sequenced members of the order *Sulfolobales* are presented in **Figure 3B**. A total of 3,020 ORFs were shared between the two *S. solfataricus* species, confirming their close relationship. In all, 2929 ORFs were shared amongst all of the *Sulfolobales*. These probably include both metabolic and physiological pathways important for ecological niche adaptation and survival that would be required for all these hyperthermophiles.

Additionally, we determined the number of predicted ORFs in the *S. solfataricus* strain 98/2 genome that could be categorized as common in each of the three branches of life, namely archaea, eukaryota and bacteria **(Figure 4)**. We found a total of 2,153 ORFs that are common to the three branches, but relatively few predicted ORFs that could be assigned to either the bacterial or eukaryotic branch exclusively. Instead, most of the predicted ORFs are shared across two domains, with the largest number of ORFs shared between the archaeal and bacterial

branches of life. Some ORFs of strain 98/2 shared with archaea encoded some metabolic enzymes like glutamine amidotransferase, keto-acid reductoisomerase, s-adenosylmethionine decarboxylase, carbamoyl phosphate synthase responsible for glutamine metabolism, valine biosynthesis, spermidine biosynthesis and arginine biosynthesis to name a few. The unique ORFs in strain 98/2 with respect to the three members in *Sufolobales* are 16 and the unique ORFs in strain 98/2 with respect to the three domains of life are 97 as observed in Figure 3 and Figure 4 respectively. The discrepancy in the numbers could be partially explained by the fact that the ORF set from *Sulfolobales* members were also derived from running GeneMark software and hence could contain predictions absent in the NCBI nr database. These two sets of unique ORFs in strain 98/2 overlap in just 7 ORFs as shown in **Figure 5**, implying a total of 106 unique ORFs in strain 98/2 that have hits neither with the domains nor with any other *Sulfolobales* member could be unknown/hypothetical proteins predicted by GeneMark, subject to further investigation and validation.

Some members of *Sulfolobales* have been shown to harbor plasmids and viruses. However, these extrachromosomal elements have not been observed experimentally in strain 98/2 (M. Rolfsmeier and C. Haseltine, unpublished). A BLAST search was performed between the known plasmid and virus sequences (obtained from http://www.sulfolobus.org/cbin/mutagen.pl?page=sequence) and the entire set of assembled contigs from strain 98/2. Using parameters of 95% identity and sequences greater than 500 bp, no significant sequences representing *Sulfolobales*-specific plasmids or viruses were retrieved, thereby supporting the experimental observations.

Conclusions

The sequence of the *S. solfataricus* strain 98/2 genome was determined using high-throughput 454 sequence technology. 65.4 Mbp of sequence reads were obtained, resulting in an estimated 21X sequence coverage with strain P2 as a reference. The sequence coverage represented here does not include repeat elements. Unlike other previously reported *Sulfolobales* genomes, strain 98/2's genome was sequenced directly from total DNA without cloning or propagation of genomic DNA in a heterologous host. We expect that circumvention of cloning resulted in an unbiased representation of strain 98/2 genome during sequencing. Overall genome size, obtained by assembling the sequences without the repeat elements, was estimated to be approximately 2.68 Mb, with a GC content of 35.43%. The genome contains 3,055 predicted open reading frames. Comparative genomic analysis indicated that *S. solfataricus* strain 98/2 is closely related to *S. solfataricus* strain P2, with 84.97% sequence *identity*. The frequency of IS element sequences was about half of that observed in *S. solfataricus* strain P2, and many of the IS elements identified had not been reported previously in the *Sulfolobales*. This difference may be a reflection of geographic isolation of the strains, where evolutionary acquisition and retention of IS elements has been distinct. Further comparative genomic studies and additional genetic analyses, including complete genome annotation, will give a deeper insight into how these archaeal organisms adapt to thrive in harsh environments.

Methods

Strain cultivation and DNA preparation. *S. solfataricus* strain 98/2 [4] was the kind gift of Paul Blum (University of Nebraska – Lincoln) *S. solfataricus* strain 98/2 has been deposited in the American Type Culture Collection and assigned accession number ATCC BAA-1849 and was grown as described previously [4] at 80˚C in screw-cap flasks and aerated by vigorous shaking.

The medium used contained 10 mM ammonium sulfate, 4 mM dibasic potassium phosphate, 1.2 mM magnesium sulfate, 1 mM calcium chloride, 1 μM iron chloride, 5 μM manganese chloride, 8 μM zinc chloride, 4 μM cobalt chloride. Sucrose was added at a final concentration of 0.2% (wt/vol) and tryptone was added at a final concentration of 0.2% (wt/vol). The growth medium was adjusted with sufficient sulfuric acid to yield a pH of 3.0. Growth was monitored spectrophotometrically at a wavelength of 540 nm. Cells were harvested at mid-log phase and total cellular DNA was isolated as described [11]. The resulting DNA was then subjected to two additional phenol chloroform extraction steps before being processed for 454 sequencing.

454 sequencing, data assembly and mapping. A shotgun DNA library was prepared from 5 μg of genomic DNA using the 454 Library Preparation Kit (Roche Inc.). Pyrosequencing was performed on a GS FLX instrument according to manufacturer's protocol (Roche Inc.). Sequence reads obtained from one run of the 454 GS FLX platform were assembled into contigs using gsAssembler software (Roche Inc.). The sequence reads were aligned to *S. solfataricus* strain P2 using gsMapper (Roche Inc.). Similar reference mapping was performed with other *Sulfolobales*. The results were tabulated and are summarized in Table 2. All software used in this study were implemented using default parameters. GC content was calculated using GeneMark software. GC Profile software [14] was used to identify segmentation points. The following parameters were used: halting = 50.00, filtered gap size = 26,813 bp, minimum length of segmentation = 1,000 bp and no gaps in the genome were filtered. Contigs from gsAssembler served as the input for the GC plot. Using a custom script, the partial headers of predicted repeats were extracted from the gsAssembler file. A second custom script identified repeat sequence information which was analyzed using IS Finder software http://www-is.biotoul.fr/is.html. A custom script to extract the top isoform hit, which represents the most significant hit, for each

repeat sequence was used. The output was used to calculate the frequency of each isoform.

Comparative genomic analysis. The genome sequence of *S. solfataricus* strain P2 was downloaded from NCBI (NC 002754). *S. tokodaii*, *S. acidocaldarius*, and *M. sedula* genome sequences were obtained from www.sulfolobus.org. Whole genome alignment and summary of location and characteristics of differences between the *S. solfataricus* strain 98/2 and *S. solfataricus* strain P2 contig sets were determined using the dnadiff program of the MUMmer software [16-18]. gsMapper (Roche Inc.) was used to map the genome reads and arrange contigs using the *S. solfataricus* strain P2 genome as a reference. Similarly, sequence identity comparisons were performed using gsMapper for the other genome sequences. gsMapper output was parsed using a custom script into text files representing partial, repeat, and unmapped reads. Each text file was mapped against the original reads input file to retrieve the entire header information along with sequence. A second custom script was used to obtain the final FASTA output and identify unmapped sequences. Repeat sequences were identified using gsAssembler and extracted using a custom script, yielding 14,871 sequences. Isoforms matching each repeat sequence were identified using IS Finder software (http://www-is.biotoul.fr/is.html). The frequency of each isoform was calculated and the isoforms were categorized into their respective families.

Predicted ORFs in the *S. solfataricus* strain 98/2 genome were identified using GeneMark [12, 13], which is a program suited for prokaryote gene finding and outputs information about protein translations, transcripts and possible frameshifts for the predicted ORFs. The program was run with the following options: Print GeneMark 2.4 predictions; GeneMark.hmm predictions; Translate predicted genes into proteins; and Sequences of predicted genes. Similar runs were performed on the other members of *Sulfolobales* namely *S. solfataricus* P2, *S.*

acidocaldarius and *S. tokodaii* respectively. The protein translated regions for the predicted ORFs given by GeneMark was used for finding the shared ORFs between strain 98/2 and each other member of *Sulfolobales*. Protein BLAST [15] was run in each case and a script extracted the top local BLAST hits. The resulting comma separated file generated contains the translation coordinates of the hit between organisms under comparison and their identity value for that hit. Another script was written to compute the number of ORFs shared between these organisms using the filtered BLAST hits in their comma separated files. Similarly, derived protein sequences of strain 98/2 were subjected to protein BLAST using NCBI web blast with the "Organism" options: Archaea, Eukaryota, and Bacteria, restricted to the top 10 hits. A list of gene superfamilies representing three branches of life was derived from BLAST searches **(Table 7)**. A custom script was used to filter the topmost hit and generate a comma separated file containing the protein translation coordinates, name of the protein, score, and e-value. Another custom script calculated the genes/proteins shared and unique among the Archaeal, Eukaryotic, and Bacterial hits.

Nucleotide sequence accession number

The genome sequence has been deposited at NCBI under Project ID: 33857.

Author contributions

VK, DJ and TK performed comparative analysis and genome assembly and mapping. DJ and VK submitted all the data to NCBI short read archive and GenBank. VK and AK developed custom scripts. DJ, TK, SS and AD developed the libraries and performed 454 sequencing. AD and AK guided data analysis. VK, DJ, AD, AK, CH and MR prepared the first draft. AG isolated the DNA. AD, AK, MR, and CH supervised aspects of the project. All authors read and approved

the final manuscript.

Acknowledgements

Washington State University's School of Molecular Biosciences startup funds to CH and, Agriculture Research Center and Department of Horticulture start up funds to AD supported this work. VK was supported in part by USDA-NRI grant # 2008-2268 to AD and AK. TK and SMS are grateful for the NIH Protein Biotechnology training grant and ARCS fellowship. We thank Kevin Kipp for assistance in data analysis, Christopher Hendrickson and Marian Laughery for helpful comments on the manuscript. Dr. Michael Kahn is thanked for useful discussions and critical comments on the manuscript.

References

1. TD Brock, KM Brock, RT Belly, RL Weiss: **Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature**. *Arch Mikrobiol* 1972, **84**:54-68.
2. M de Rosa, A Gambacorta, JD Bu'lock: **Extremely thermophilic acidophilic bacteria convergent with Sulfolobus acidocaldarius**. *J Gen Microbiol* 1975, **86**:156-64.
3. C Haseltine, M Rolfsmeier, P Blum: **The glucose effect and regulation of alpha-amylase synthesis in the hyperthermophilic archaeon Sulfolobus solfataricus**. *J Bacteriol* 1996, **178**:945-50.
4. M Rolfsmeier, P Blum: **Purification and characterization of a maltase from the extremely thermophilic crenarchaeote Sulfolobus solfataricus**. *J Bacteriol* 1995, **177**:482-5.
5. P Worthington, V Hoang, F Perez-Pomares, P Blum: **Targeted disruption of the alpha-amylase gene in the hyperthermophilic archaeon Sulfolobus solfataricus**. *J Bacteriol* 2003, **185**:482-8.
6. J Schelert, M Drozda, V Dixit, A Dillman, P Blum: **Regulation of mercury resistance in the crenarchaeote Sulfolobus solfataricus**. *J Bacteriol* 2006, **188**:7141-50.

7. L Chen, K Brugger, M Skovgaard, P Redder, Q She, E Torarinsson, B Greve, M Awayez, A Zibat, HP Klenk, et al: **The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota**. *J Bacteriol* 2005, **187**:4992-9.
8. Q She, RK Singh, F Confalonieri, Y Zivanovic, G Allard, MJ Awayez, CC Chan-Weiher, IG Clausen, BA Curtis, A De Moors, et al: **The complete genome of the crenarchaeon Sulfolobus solfataricus P2**. *Proc Natl Acad Sci U S A* 2001, **98**:7835-40.
9. Y Kawarabayasi, Y Hino, H Horikawa, K Jin-no, M Takahashi, M Sekine, S Baba, A Ankai, H Kosugi, A Hosoyama, et al: **Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, Sulfolobus tokodaii strain7**. *DNA Res* 2001, **8**:123-40.
10. KS Auernik, Y Maezato, PH Blum, RM Kelly: **The genome sequence of the metal-mobilizing, extremely thermoacidophilic archaeon Metallosphaera sedula provides insights into bioleaching-associated metabolism**. *Appl Environ Microbiol* 2008, **74**:682-92.
11. M Rolfsmeier, C Haseltine, E Bini, A Clark, P Blum: **Molecular characterization of the alpha-glucosidase gene (malA) from the hyperthermophilic archaeon Sulfolobus solfataricus**. *J Bacteriol* 1998, **180**:1287-95.
12. AV Lukashin, M Borodovsky: **GeneMark.hmm: new solutions for gene finding**. *Nucleic Acids Res* 1998, **26**:1107-15.
13. J Besemer, M Borodovsky: **GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses**. *Nucleic Acids Res* 2005, **33**:W451-4.
14. F Gao, CT Zhang: **GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences**. *Nucleic Acids Res* 2006, **34**:W686-91.
15. SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403-10.
16. AL Delcher, A Phillippy, J Carlton, SL Salzberg: **Fast algorithms for large-scale genome alignment and comparison**. *Nucleic Acids Res* 2002, **30**:2478-83.
17. AL Delcher, S Kasif, RD Fleischmann, J Peterson, O White, SL Salzberg: **Alignment of whole genomes**. *Nucleic Acids Res* 1999, **27**:2369-76.
18. S Kurtz, A Phillippy, AL Delcher, M Smoot, M Shumway, C Antonescu, SL Salzberg: **Versatile and open software for comparing large genomes**. *Genome Biol* 2004, **5**:R12.

Figure legends

Figure 1: Plot of Cumulative GC profile across 98/2 genome. The red line indicates the GC Profile. Segmentation points are indicated by green boxes and highlighted by arrows.

Figure 2: A graph representing the number and type of transitions and transversions that occur in strain 98/2 using 98/2 contigs from gsAssembler and reference strain P2 as input for MUMmer

Figure 3: (A) Histogram indicating the number of ORFs predicted in strain 98/2 and other members in *Sulfolobales* using GeneMark. (B) Venn diagram (not to scale) showing how 3053 (out of 3055) predicted genes of strain 98/2 that corresponded to protein translated regions are distributed among other members of *Sulfolobales*. The numbers in parentheses below each strain are the total number of genes of strain 98/2 shared with that member of *Sulfolobales*.

Figure 4: Venn diagram (not to scale) showing the distribution of 3053 (out of 3055) predicted genes with protein translations in strain 98/2 among archaea, bacteria and eukaryotes. The numbers in parentheses below each domain are the total number of genes of strain 98/2 shared with that domain.

Figure 5: Diagram showing the distribution of 3053 (out of 3055) predicted genes with protein translations in strain 98/2 among the 3 domains of life indicated by 'Domain' and *Sulfolobus* family (*S. tokodaii, S. acidocaldarius* and *S. solfataricus* P2) indicated by 'Family'.

Tables

Table 1: Comparison of *de novo* assembly of *S. solfataricus* strain 98/2 genome using gsAssembler and guided assembly with *S. solfataricus* strain P2 as a reference using gsMapper.

Type of read	Number of sequences with no reference (gsAssembler)	Number of sequences with *S. solfataricus* strain P2 as a reference(gsMapper)
Partially mapped/assembled	4,248	14,316
Repeats	14,871 (3,557,066 bp)	16,107 (3,541,989 bp)
Singletons/Unmapped	663	17,758
Contigs	509 (2,681,380 bp)	497 (2,499,824 bp)
N50 value	24,184 bp	26,517 bp

Table 2: Identity of *S. solfataricus* strain 98/2 to sequenced *Sulfolobales* using gsMapper.

Reference Strain	Percentage reference coverage	Genome Size (Mbp)
Sulfolobus solfataricus strain P2	84.97%	2.99
Sulfolobus tokodaii	2.15%	2.69
Sulfolobus acidocaldarius	0.76%	2.23
Metallosphaera sedula	0.41%	2.19

Table 3: Genomic feature detail differences between *S. solfataricus* strain P2 and *S. solfataricus strain* 98/2. * indicates that the translocation events recorded for strain P2 were not recorded in strain 98/2. That does not indicate that strain 98/2 does not have any translocations.

Genome Elements	*S. solfataricus* strain P2	*S. solfataricus* strain 98/2
Breakpoints	1,488	1,432
Translocations	372	0*
Insertions	920	151

Table 4: Different types of indels detected in strain 98/2 in comparison with strain P2. Deletion of base A in strain 98/2 with respect to strain P2 is indicated by A→ **.** and an insertion of base A in strain 98/2 with respect to strain P2 is indicated by **.** →A.

Deletions	Insertions
A→**.** 94	**.** →A 45
C→**.** 50	**.** →C 33
G→**.** 52	**.** →G 14
T→**.** 72	**.** →T 52

Table 5: Types of insertion elements prevalent in *Sulfolobales*. IS Family indicates the broad family of insertion elements. The repeat elements identified in de novo assembly were processed through IS Finder to identify the family members. The last column represents the number of similar IS elements present in strain 98/2 common to *S. solfataricus* and *S. tokadaii*.

IS Family	Number of occurrences of isoforms in family for strain 98/2	Unique isoforms in family for strain 98/2	Unique Isoforms in other members of *Sulfolobales* present in strain 98/2
IS1	279	6	*Sulfolobus solfataricus* (2) *Sulfolobus tokodaii* (1)
IS110	2261	12	*Sulfolobus solfataricus* (3) *Sulfolobus tokodaii* (3)
IS1182	132	9	0
IS1380	2	2	0
IS1595	33	7	0
IS1634	2	2	0
IS200/ IS605	1217	10	*Sulfolobus solfataricus* (1)
IS21	24	10	0
IS256	31	5	*Sulfolobus solfataricus* (2)
IS3	30	13	0
IS30	21	3	0
IS4	70	11	0
IS481	5	1	0
IS5	3169	13	*Sulfolobus solfataricus* (4)
IS6	71	3	*Sulfolobus tokodaii* (1)

IS607	1004	5	*Sulfolobus solfataricus* (2) *Sulfolobus tokodaii* (1)
IS630	310	6	*Sulfolobus solfataricus* (3)
IS66	11	6	0
IS701	2	2	0
IS91	6	1	0
IS982	37	5	0
ISAs1	18	5	0
ISH3	4330	8	*Sulfolobus solfataricus* (3) *Sulfolobus tokodaii* (2)
ISL3	10	4	0
ISNCY	1770	8	*Sulfolobus solfataricus* (2) *Sulfolobus tokodaii* (1)
UNKNOWN	1	1	0
Tn3	23	2	0

Table 6: Number of genes unique to strain 98/2 relative to each other member of *Sulfolobales*.

Name of strain	Unique genes in strain 98/2
Sulfolobus solfataricus P2	33
Sulfolobus tokodaii	76
Sulfolobus acidocaldarius	79
Metallosphaera sedula	220

Table 7: List of protein superfamilies found in predicted genes of strain 98/2 during Eukaryota, Archaea and Bacterial BLAST results.

Contig ID of strain 98/2 containing predicted gene region	Number of Amino acids in the predicted gene	Superfamily	Multi-Domain	Site(s)
Superfamilies of Eukaryota in 98/2 ORFs				
00148 00149 00150	164	1.Ser_Recombinase 2.SSF Superfamily	- -	Catalytic residues -
00147	692	PHA00735	-	-
00208	852	-	COG1361	-
00209	935	1.nt_trans 2. nt_trans 3.Anticodon_1	leuS	1. LeuRS core with active site and HIGH motif 2. active site, nuclotide binding site and KMSKS motif
00225	865	1. P-loop NTPase 2. P-loop NTPase	SbcC	1. ABC_Rad50 hit with ATP binding site, Walker A/P- loop, Q-

				loop/lid 2. ATP binding site,ABC Transporter signature motif Walker B, D- loop, H-loop/switch region
00227	895	Polysacc_deac_1	COG1449	-
00227	418	PotE	-	-
00227	308	Na_Ca_ex	ECM 27	-
00227	136	Restriction_Endonuclease_like	-	Archeal HJR with active site, dimer interface and DNA binding cleft
00227	357	AAT_I	Aminotran_1_2	AAT-like hit with Pyridoxal 5' - phosphate binding,Catalytic residue, homer dimer interface
00225	340	NurA	-	-
00227	356	PBPb	-	-
00291	1001	1. FAD_binding_4 2. HCP_like	GlcD, GlpC, PRK11230, COG1139,MurB,FrDB,RnfC, glpC	Non specific hits: 1. FAD_binding_4 2. ACS_1
Superfamilies of Archaea in 98/2 ORFs				
00147	692	PHA00735	- -	-
00148 00149 00150	164	1.Ser_Recombinase 2 .SSF	- -	- -
00208	1052	1.CPSase_L_chain 2.Dala_Dala_lig_C 3.CPSase_L_D3	CarB	-
Superfamilies of Bacteria in 98/2 ORFs				
00025	914	DEXDc	Lhr	ATP binding site, putative Mg++ binding site

FIGURES

FIGURE 1

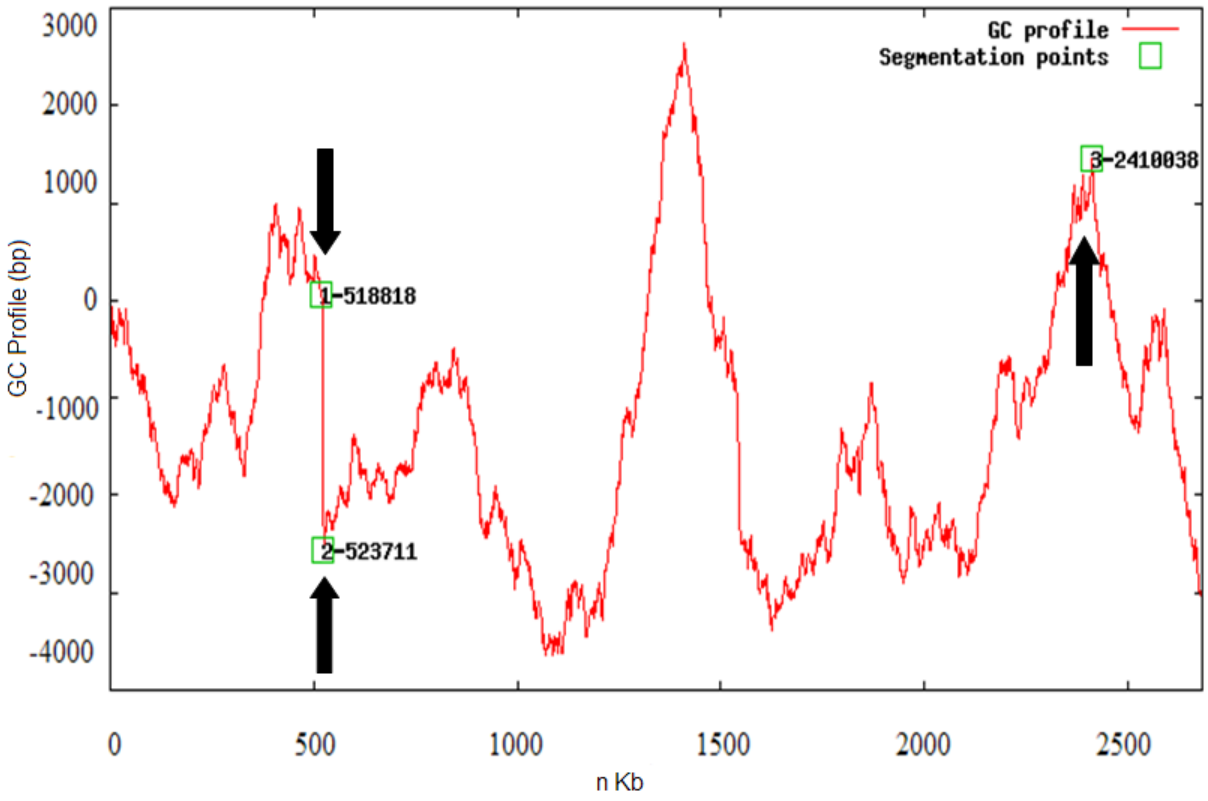


FIGURE 2

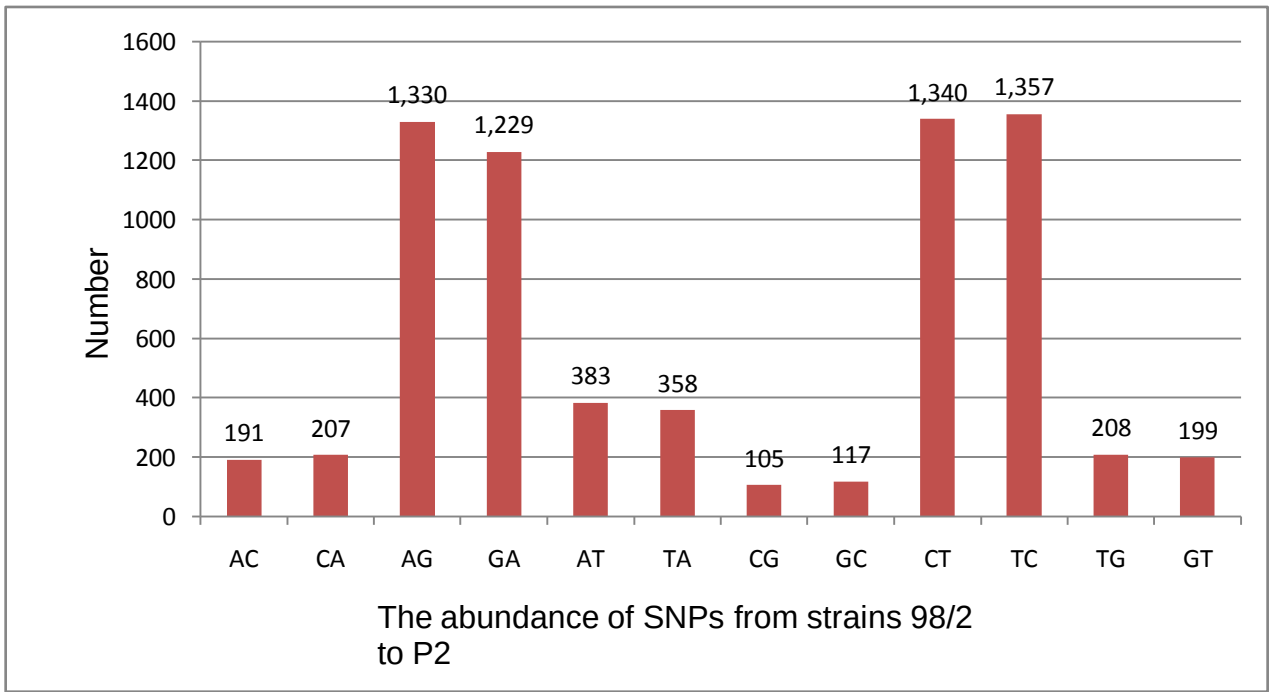

1600
1400
1200
1000
800
600
400
200
0
Number
191
207
1,330
1,229
383
358
105
117
1,340
1,357
208
199
AC
CA
AG
GA
AT
TA
CG
GC
CT
TC
TG
GT
The abundance of SNPs from strains 98/2 to P2

FIGURE 3

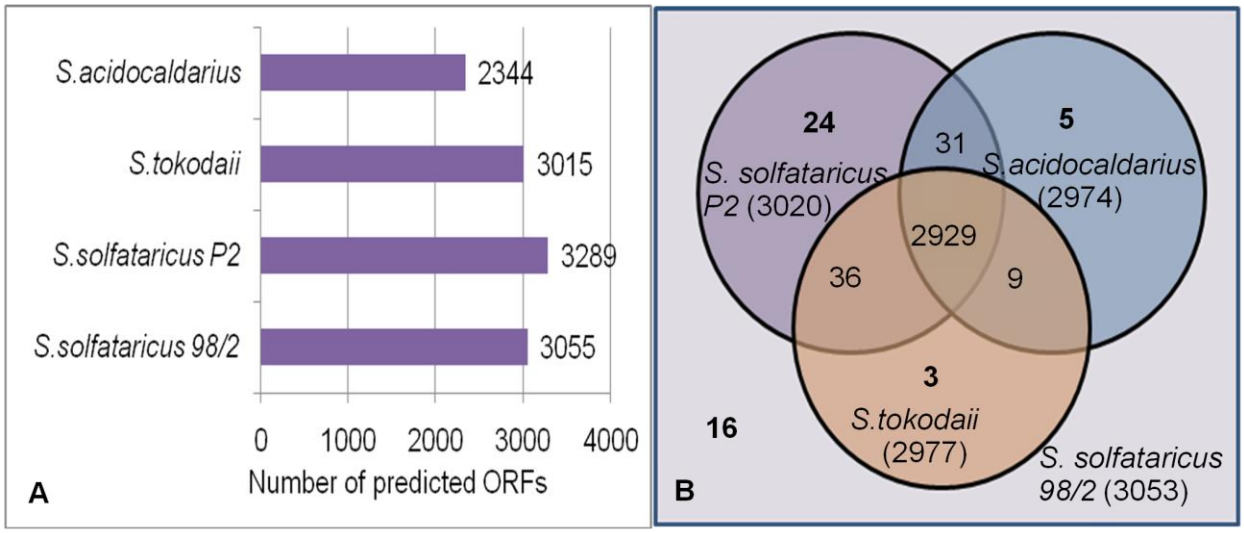

S.acidocaldarius
2344
S.tokodaii
3015
S.solfataricus P2
3289
S.solfataricus 98/2
3055
0
1000
2000
3000
4000
Number of predicted ORFs
A
24
S. solfataricus P2 (3020)
31
5
S.acidocaldarius (2974)
2929
36
9
3
S.tokodaii (2977)
16
S. solfataricus 98/2 (3053)
B

FIGURE 4

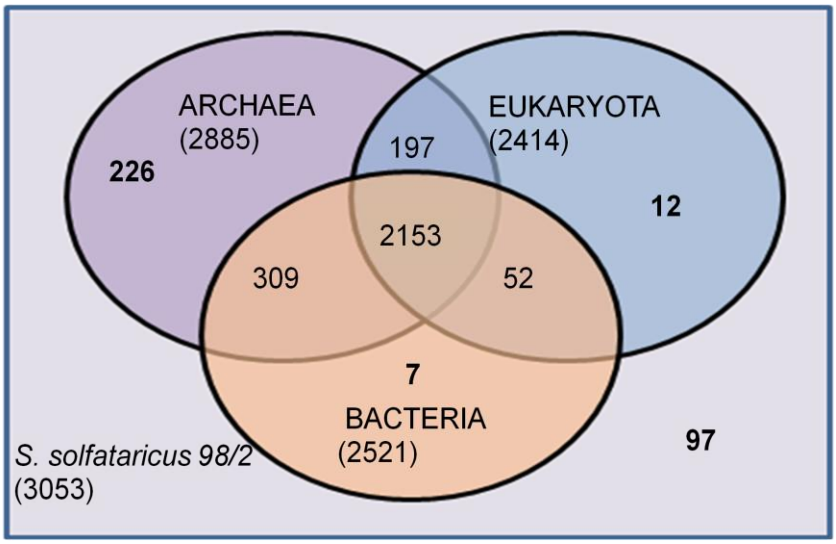
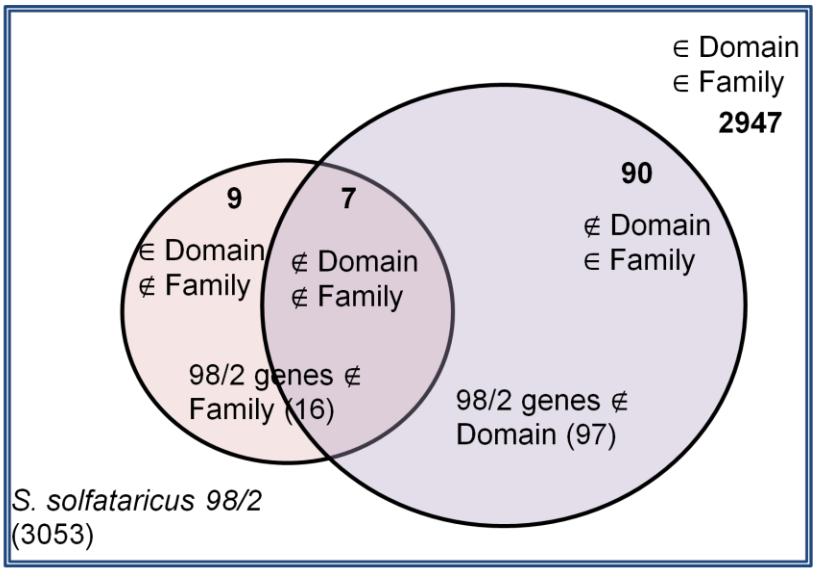


FIGURE 5



CHAPTER FOUR

ENABLING COMPARATIVE AND QUANTITATIVE TRANSCRIPTOME PROFILING IN ORPHAN PLANT SPECIES

Vandhana Krishnan[1], Tyson Koepke[2], Scott Schaeffer[2], Christopher Hendrickson[2], Ananth Kalyanaraman[1], Amit Dhingra[2]

[1]School of Electrical Engineering and Computer Science, Washington State University, Pullman, USA

[2]Horticulture and Landscape Architecture, Washington State University, Pullman, USA

VK: vandhana_k@wsu.edu

TK: tkoepke@wsu.edu

SS: smschaeffer@wsu.edu

CH: chendrikson@gmail.com

AK*: ananth@eecs.wsu.edu

AD*: adhingra@wsu.edu

Abstract:

A biological and computational framework has been established for comparative and functional genomics in orphan species without using any prior genomic resources. The underlying biological approach consists of generating qualitative and quantitative transcriptome profiles from biological samples representing important physiological conditions or a carefully selected trait. The framework includes a program used to obtain an appropriate set of enzymes to maximally digest sequences of a sample to achieve longer transcript reads.

Next-generation *de novo* sequencing of the entire transcriptome provides a reference

dataset for the quantitative 3'-UTR sequencing established in the laboratory. 3'UTR sequencing plays an important role in identifying expression levels of each transcript based on read frequency. Through the method developed in this paper, we can obtain and compare transcript information within species having no reference sequence information. The quality of statistical results obtained has been validated further using prevalent molecular biology techniques.

Introduction:

Gene expression studies have been extensively undertaken to study the different traits seen among organisms of the same species. Several genomes have been sequenced so far and most of the information on the recently sequenced organisms are based on standard models such as *Arabidopsis thaliana* in plant species.

However, genes can be very divergent among different species and sequencing their genomes is not sufficient to understand the numerous traits displayed in such organisms. Also, the sequence information from closely related species is not adequate to identify the unique characteristics of the organism under study. This is the reason why *de novo* sequencing has evolved to play a vital role in the field of genomics.

We know that for a particular species , their genes can express themselves differently and hence the traits vary from one another. This can be due to external factors influencing the expression of the genes and the organism adapting itself to the environment in which it lives. The first form of gene expression studies involved use of northern blots [1] in which one we needs to know the sequence of the gene to design probes. Moreover, northern blots are time consuming for analyzing several genes as the probes need to be designed for each one of them. Also, it involves radioactivity not desirable for long term purposes.

Differential display is another method used to compare the genes that are differentially expressed [2]. An advantage of this method is that it can be used to discover unknown genes. The random primers designed could leave out a few cDNA sequences and we may miss some gene information or have results containing redundant transcripts. The process of identifying the differences in the bands (due to varying gene expression) , extracting the cDNA from the gel to clone and sequence the genes can be a tedious process. The method is laborious and may take many years to establish good results.

Microarrays on the other hand enables study of a large amount of genes simultaneously by using DNA oligonucleotides of known sequences attached to fluorescent probes in micro-spots of a chip [3, 4]. It is useful in identifying SNPs and gene expression levels of different organisms. Again, the limitation is that study of orphan organisms is not possible especially when EST information is scarce. The orphan organism's *ESTs* would otherwise be used to the design the chip. The length of the oligonucleotide designed varies from 20 -50bp. Depending on the intensities of the color in each spot, the expression levels are quantified. The protocols, fabrication of the chip and analysis methods are not standardized making it difficult to use.

Serial Analysis for Gene Expression (SAGE) [5,6] is a method in which fragments of different mRNA molecules (tags) extracted from a desired sample are linked together. These tags are cloned using vectors and later can be sequenced using high throughput sequencers. Further, using statistical analyses the number of times these tags appear is calculated to find the expression levels. This is more efficient than microarrays as it does not involve quantitative analysis based on spot intensities which are prone to miscalculations. The main drawback of SAGE is that it is not applicable for large scale gene expression studies due to the costs involved.

Semi - quantitative PCR [7] can also be used to study the expression of genes. It involves

using cDNA pools of desired samples and running a regular PCR. Quantifiable amounts of the amplified sequences are extracted at defined incremental cycles until the end of the PCR reaction. These samples collected at every incremental cycle are run on a gel and depending on the band intensities, the transcript that is most expressed between samples is estimated. The main disadvantage here is that the initial level of transcripts in each sample is unknown. The experiment is based on the fact that a particular transcript if present in high numbers initially, after every cycle of PCR amplification the same transcript outnumbers the others. To overcome this limitation a Q- PCR can be performed.

Q- PCR involves use of a control for all the given samples based on which one can estimate the initial sample concentration. All the samples should be of equal concentration showing a thick single control band to avoid false positives of transcript expression levels. It inherently plots a graph of the level of transcripts in each sample at every cycle. Unfortunately, setting up samples of equal concentration, designing gene specific primers and appropriate reaction times for the experiment is a tedious task. In this method too, identifying unique genes in orphan species is not possible.

Current breakthrough sequencing technologies such as Solexa (Illumina Inc.) and 454 sequencing (Roche Inc.) provide a better coverage of the sequences than Sanger sequencing [8,9]. The length of the sequences obtained is higher compared to all previous methods for instance, Solexa geneartes 75 bp reads for analysis. Solexa performs *de novo* sequencing and sequencing based on references but the length of the reads generated is smaller compared to 454 sequencing.

Using the 454 sequencing, read lengths of 100bp - 300bp can be obtained depending on the sequencing platform is used. Also, with the recent upgrade to titanium series reads of 400bp

length can be obtained using 454 sequencing. Moreover, there is the method [10] to sequence 3' UTR sequences efficiently using 454 technology. Hence, we propose a modified technique to perform transciptome profiling in orphan species coupling the advantages of the above mentioned facts.

Materials and Methods:

The reads from apple, pear, cherry near full-length cDNA sequencing runs are assembled using the gsAssembler software (Roche Inc.). This builds a reference library for the genes we are trying to identify in orphan species.

To prepare cDNA libraries for 3' UTR sequencing, we need to find the restriction enzyme sites. First, those sequences which have a 3' poly A tail or a 5' poly T header are extracted and oriented in their respective 3'-5' directions. This is achieved as follows: the sequences with a poly A tail are trimmed of their polyAs and reversed; and the sequences with a poly T header are trimmed of their polyTs and complemented.

Identification of Restriction Enzyme Cut Sites

In the next step, we identify restriction enzyme cut sites for each sequence present in the polyA trimmed fasta file, within a specified distance from their 3' ends towards their 5' ends. The range is a user specifiable parameter to the program. In our experiments, we used a range of 100bp-200bp from the 3' end. The purpose of specifying this range is to provide the flexibility to accommodate the use of different sequencing technologies that support different read lengths (e.g., 454, Solexa/Illumina, Sanger).

Let $R=\{r_1,r_2,\ldots r_m\}$ denote the input set of restriction enzymes with known cut sites, and $S=\{s_1,s_2,\ldots s_n\}$ denote the set of all input sequences.

1) Using the restriction enzymes in *R*, we first identified the coordinates of their cut sites inside individual reads in *S* such that the cut site is present within the specified range. This step was implemented using simple regular expression-based pattern matching in Perl.
2) Next, we build a binary matrix *X* of size *m x n*, where X[i,j] is set to 1 if enzyme r_i cuts read s_j; and 0 otherwise (**Table 1**).

The Restriction Enzyme Selection ("RES") Problem

Problem statement:

Select a smallest subset of restriction enzymes from R that collectively cut at least a user-specified percentage (ζ%) number of sequences in S.

Lemma: The RES problem is NP-Hard.

Proof:

The sequences cut by each restriction enzyme represent a subset of S. This implies that the *Set Covering problem* [11] which is a well-known NP-Hard problem in computational theory, can be reduced to the RES problem by simply setting ζ=100% . Therefore, the RES problem is also NP-Hard.

An Efficient Algorithm

Our algorithm for restriction enzyme greedy selection can be illustrated as shown in Figure 1. The main idea of the approach is to maintain a running list for selected restriction enzymes

("selected list") and another running list for sequences yet to be covered ("pending list"). At every iteration of the algorithm, a restriction enzyme that covers most of the sequences in the pending list is selected. As a result of the selection, all the sequences that are covered in the pending list are removed, and the selected restriction enzyme itself becomes unavailable for future selection. The process is repeated until either ζ% of the sequences in *S* are covered or no such cover exists. The algorithm runs in $O(mn^2)$ time and $O(mn)$ space.

Processing the 3' UTR reads

Utilizing this suite of enzymes, the procedure to sequence 3'UTR reads follows the established protocol [10]. The output file obtained from the 454 sequencing contained the 3'UTR reads. These reads were formatted using the preprocessing script as mentioned previously. Another customized script removes the 'Multiplexing Identifiers' (MIDs) from all the reads and attaches appropriate header information to the existing header for each read in a particular file. (MIDs are unique set of nucleotide sequences that act as a bar code identifier for the sequencing machine.) This removal of MIDs is not required if one desires assembly of the reads using gsAssembler.

Each file represents a separate dataset indicating a) a different sample collection date b) different sample varieties or c) different samples run simultaneously. In order to separate data from within samples , unique header information is attached to the individual reads from each file. Then, individual files belonging to a single sample are put into a single file. Subsequently, PaCE [12] is run on each sample separately. PaCE generates clusters with a distinct gene represented by each cluster. Thus, we avoid losing transcript information by not performing assembly at this stage.

Next, a customized script is run on each sample's PaCE output file to give a comma separated

file (.csv) with the following details:

ClusterNumber, Number of Members, Number of Sample1type_1Members, Number of Sample1type_2Members,…… Number of Sample1type_nMembers.

This .csv file is imported to an Excel sheet and a Chi- square analysis is performed on them grouping the sample types we are looking for gene comparison. Next the p-value for the corresponding chi-squares is calculated. Further, sorting clusters in their ascending order based on their p-values gives the clusters for a particular sample in their order of significance. The user can specify a cut-off to limit the number of significant clusters to be studied for that sample.

The desired set of significant cluster numbers as identified by PaCE [12] is exported to a text file. To obtain the sequence information of all the transcripts within each cluster in a fasta form, another module built for PaCE output is used. Further, using the text file of the desired significant cluster numbers along with the directory containing the fasta information of all clusters, a script retrieves the required cluster files. The resulting 3'UTR sequences in a cluster can now be mapped against the original near full length c-DNA sequences using local BLAST [13]. A custom script extracts only the c-DNA sequences that have a hit with the 3'UTR contigs generating a fasta file. The script performs two functions: a) attaches the contig name of the UTR that has a hit with the corresponding c-DNA sequence header b) extracts the c-DNA sequence information in FASTA format from the original c-DNA library file. The above step gives the cDNA sequence to which a 3' UTR read maps and this cDNA sequence larger than the 3' UTR read sequence can improve the local BLAST hit results on other databases such as nr/nt or a database of interest for eg; flowering genes, predicted genes from same family as the sample, etc. These BLAST results can be used for further annotation studies. Q-PCR or RT-PCR is run to confirm the expression of some genes.

Thus, Q-PCR performed validates the data obtained computationally and statistically as described above.

Results:

The computational framework successfully helped identify some differentially expressed genes in Honey crisp and Golden delicious apples, genes responsible for ripening in Bartlett and D'Anjou pear varieties and the effect of flowering genes in Rainier and Bing cherry varieties for different rootstocks.

The restriction cut program calculated the minimum number of enzymes required to cut the desired set of sequences for a given cut percentage as listed in **Table 2-5**. These tables list the cut percentage of the enzymes in a cumulative order, the last entry indicating the total cut percentage obtained for the combination of the enzymes listed.

The number of reads used as input to PaCE for the different fruit samples and the resulting clusters obtained in listed in **Table 6**.

The details on significant clusters and p-value thresholds chosen for apple fruit tissue type within the varieties under comparison is listed in **Table 7**. The trend of the negative log of p-values for the significant clusters in apple fruit can be seen in the Figure 2 and Figure 3.

Discussion:

Combining high throughput sequencing technologies and computational methods yield good results in transcriptome profiling. The existing transcriptome profiling using 3' UTR reads [10] gave good results in case of maize but the same technique could not be applied to orphan organisms. This is because the enzyme *Msp*I was not effective in the digestion of all the cDNAs from interested fruit tissues to the same extent as in maize. In order to overcome this limitation,

we developed a program that can give the user a feasible set of restriction enzymes that maximally cuts a give set of input sequences. Thus, the generic program can be used for any transcriptome studies in the future as it is not dependent on the organism being studied or the type of restriction enzyme. Also, the program greatly reduces the amount of costs incurred in finding the appropriate enzyme(s) for preparing quality full length cDNA library using molecular biology techniques. An added advantage of using the program is that the range for the cut site in the input 3'UTR can be specified and we can obtain a longer transcript length if desired.

The homology based clustering using parallelization software - PaCE is a direct approach that has linear space complexity and high scalability as opposed to using serial software – BLAST that performs sequence similarity searches and has exponential complexity in terms of word size.

The strength of our approach in contrast to the existing protocol [10] is that we can cater to orphan organisms using RES algorithm to obtain transcripts of any size using any sequencing technology.

Acknowledgements:

VK is thankful to Derick Jiwan for help in understanding the biological aspects of the project.

Author Contributions:

TK, AK and VK developed the algorithm and program to identify an experimentally feasible combination of restriction enzymes for given set of sequences. SS, TK and AD developed the libraries and performed 454 sequencing. SS, CH, TK and VK conducted statistical analyses.

SS performed the q-PCR validation. AD and AK supervised aspects of the project.

VK, AD and AK prepared the first draft.

References

1. Alwine, J.C., D.J. Kemp, G.R. Stark. (1977). **Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes**. *Proc Natl Acad Sci USA*. Vol. 74, No.12, pp. 5350-5354.
2. Liang, P., L. Averboukh, and A.B. Pardee. (1993). **Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization**. *Nucleic Acids Res.* Vol. 21, No.14, pp. 3269–3275.
3. Schena, M., D. Shalon, R.W. Davis, P.O. Brown. (1995). **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science*. Vol. 270, pp. 467–470.
4. Lashkari, D.A., J.L. DeRisi, J.H. McCusker, A.F. Namath, C. Gentile, S.Y. Hwang, P.O. Brown, R.W. Davis. (1997). **Yeast microarrays for genome wide parallel genetic and gene expression analysis**. *Proc Natl Acad Sci USA*. Vol. 94, pp. 13057–13062.
5. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu, VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Hieter P, Vogelstein B, and Kinzler KW. (1997). **Characterization of the yeast transcriptome**. *Cell*. Vol. 88, pp. 243-251.
6. Velculescu, V.E., L. Zhang , B. Vogelstein , and K.W. Kinzler. (1995). **Serial Analysis Of Gene Expression**. *Science* Vol. 270, pp. 484-487.
7. Marone, M., S. Mozzetti, D. De Ritis, L. Pierelli and G. Scambia. (2001). **Semiquantitative RT-PCR analysis to assess the expression levels of multiple transcripts from the same sample**. *Biol Proced Online*. Vol. 3, pp. 19-25.
8. F. Sanger, S. Nicklen and A.R. Coulson. (1977). **DNA sequencing with chain-terminating inhibitors**, *Proc. Natl. Acad. Sci. U.S.A.* Vol.74, pp. 5463–5467.
9. A.M. Maxam and W. Gilbert. (1977). **A new method for sequencing DNA**. *Proc. Natl. Acad. Sci. U.S.A.* Vol. 74 , pp. 560–564.
10. Andrea L. Eveland,Donald R. McCarty,Karen E. Koch. (2008). **Transcript Profiling by 3'- Untranslated Region Sequencing resolves Expression of gene Families**. *Plant Physiology* Vol. 146, pp. 32-44.
11. Cormen, T., C. Lieserson, R. Rivest, C. Stein. **Introduction to Algorithms, McGraw-Hill Science/EngineeringMath**; 2nd edition, 2003.
12. Anantharaman Kalyanaraman,Srinivas Aluru,Suresh Kothari and Volker Brendel**.** (2003) **Efficient Clustering of large EST data sets on parallel computers.**

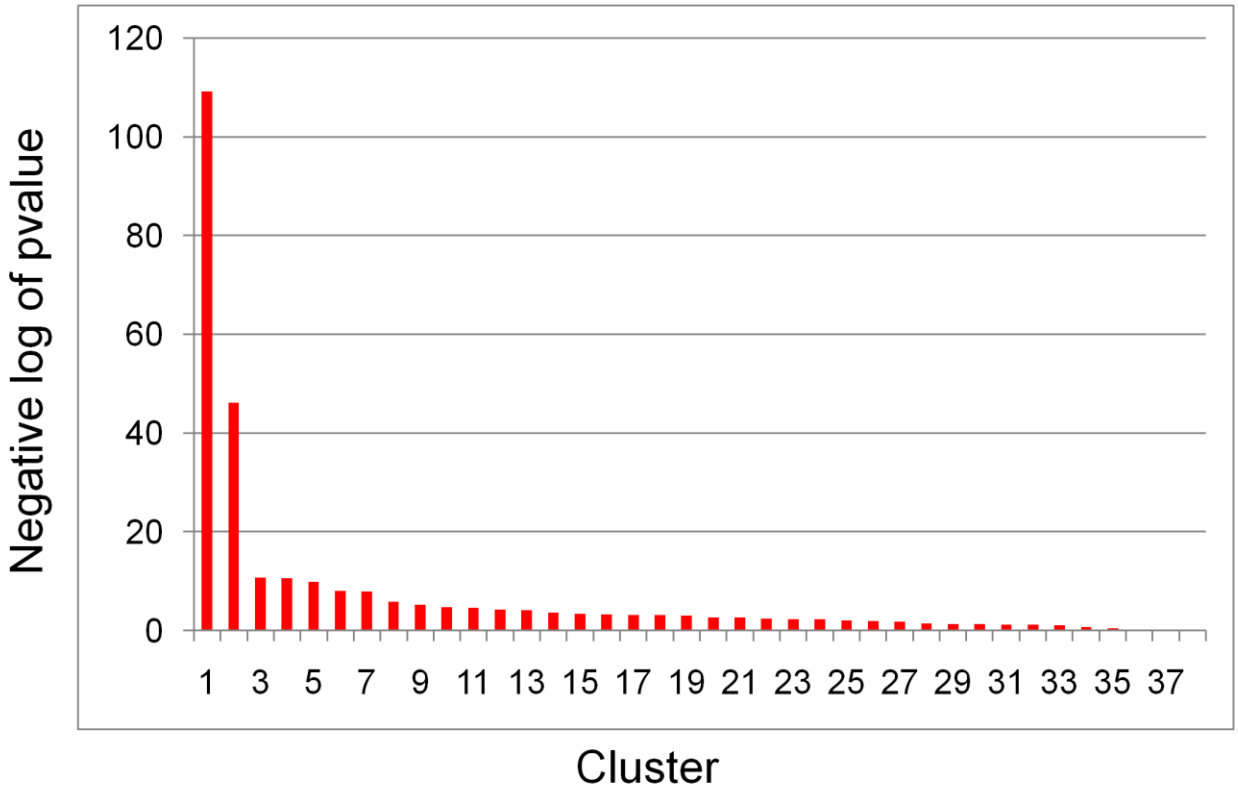
Nucleic Acids Research, Vol. 31, No. 11 2963-2974.
13. SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman.(1990) **Basic local alignment search tool**. *Mol Biol*, **215**, 403-10.

Figures

Figure 1 Pseudocode for the restriction enzyme selection program.

```
RestrictionEnzymeSelection (Table X, cutoff ζ )
{
    Let Enzyme_Selection_List ← Φ;
    Let  Pending_Sequence_List  ← {s1,s2,……. sn };

    Let X' ← X;

    LOOP:

        Let ri be the row with maximum number of 1s in X';
        Enzyme_Selection_List ← Enzyme_Selection_List ∪ {ri };
        ∀ sequences sj in S cut by ri
        {
                    Remove j from Pending_Sequence_List;
                    Remove column j from Table X' ;
                    Remove row i from Table X' ;
        }

    UNTIL (Table X' is empty) or ( |Pending_Sequence_List| < ( 1-ζ ) ) or
    (Enzyme_Selection_List is Φ )

    If (|Pending_Sequence_List| >= 1- ζ ) or (Enzyme_Selection_List is Φ ) Then
            Output "A valid enzyme selection does not exist";
    Else
            Output the Enzyme_Selection_List;
}
```

Figure 2: A graph of the negative log of pvalues against the clusters in apple tissues.



List of Tables

Table 1: Template of result file of the program calculating the cut sites of the input sequences for a given set of enzymes.

	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence n
Enzyme 1	0	1	1	0	1
Enzyme 2	0	1	0	1	1
Enzyme 3	0	0	1	0	1
Enzyme 4	1	1	0	0	0
Enzyme 5	0	0	1	1	1
Enzyme 6	1	0	0	0	0
Enzyme n	0	1	0	1	1

Table 2: Results of the program that finds the optimal combination of enzymes for a set of Apple predicted genes 0bp to 400bp from the 3'UTR end. The enzyme used is listed in the first column, the enzymes' cut site is specified in the second column and the percentage of cDNA sequences it cuts is the third column.

Enzyme	Cut site	Total cut percentage
*Taqα*I	T/CGA	68.10%
*Msp*I	C/CGG	82.70%
*Hpy*CH4IV	A/CGT	89.00%

Table 3: Results of the program that finds the optimal combination of enzymes for a set of Apple ESTs 0bp to 400bp from the 3'UTR end. The enzyme used is listed in the first column, the enzymes' cut site is specified in the second column and the percentage of cDNA sequences it cuts is the third column.

Enzyme	Cut site	Total cut percentage
*Taqα*I	T/CGA	60.66%
*Msp*I	C/CGG	76.97%
*Hpy*CH4IV	A/CGT	86.09%
*Hin*P1I	GR/CGYC	89.86%

Table 4: Results of the program that finds the optimal combination of enzymes for a set of Pear predicted genes 0bp to 400bp from the 3'UTR end. The enzyme used is listed in the first column, the enzymes' cut site is specified in the second column and the percentage of cDNA sequences it cuts is the third column.

Enzyme	Cut site	Total cut percentage
*Taqα*I	T/CGA	38.62%
*Msp*I	C/CGG	60.14%
*Hpy*CH4IV	A/CGT	52.21%
*Hin*P1I	GR/CGYC	64.93%

Table 5: Results of the program that finds the optimal combination of enzymes for a set of Peach ESTs in the range 0bp to 400bp from the 3'UTR end. The enzyme used is listed in the first column, the enzymes' cut site is specified in the second column and the percentage of cDNA sequences it cuts is the third column.

Enzyme	Cut site	Total cut percentage
*Taqα*I	T/CGA	44.98%
*Msp*I	C/CGG	66.90%
*Hpy*CH4IV	A/CGT	77.43%
*Hin*P1I	GR/CGYC	82.72%

Table 6: Results of running PaCE on number of 3'UTR reads output from the GS GLX. The last column is the number of clusters formed by PaCE for these reads.

Sample	Total number of 454 reads(3'UTR)	Total number of clusters
Apple	49689	28298
Pear	36239	22163
Cherry	580450	114145

Table 7: Chi-square analysis on apple peel and core tissues resulted in the following number of significant clusters for each sample with p-value cut off for the significant clusters listed in the second column..(GDP –HCP : Golden delicious peel to Honey crisp peel comparison, GDC-HCC : Golden delicious core to Honey crisp core comparison).

Sample	p-value cut off	No of significant clusters (after p-value cut off)
GDP-HCP	0.005	17
GDC-HCC	0.005	48

CHAPTER FIVE

CONCLUSIONS

Genomics and transcriptomics are rapidly advancing with an array of high throughput technologies available these days. The sequence data generated per run by these advanced systems is enormous and at the same time the quality of the reads generated is high in terms of read length. An added advantage of these technologies is their ability to sequence several samples in parallel during a single run. The various goals for a desired project can be accomplished through the establishment of a computational framework in terms of handling the data to obtain useful biological information. This framework is not a straightforward implementation by aggregating several computational tools available today as seen in the two projects dealt with in this thesis. The data from the sequencer has to be parsed in many instances and then fed into the appropriate tools for analysis. Further, custom programs have been developed to analyze and obtain required results from the output generated by the tools at every stage. For instance, the gsAssembler identifies certain reads as repeats by providing a portion of the read header. We need to process that information and obtain the sequence information of those repeats. Then, as per the organism under study one can choose an appropriate tool to analyze these repeat structures. Every organism has its own set of unique traits and one cannot completely base their studies on information from closely related sequenced organisms or a standard model organism. The purpose of *de novo* assemblers is to avoid this above bias. One such assembler - the gsAssembler specifically designed to analyze 454 sequencing data has been used in our studies. The data from this assembler was used for identifying most of the structural and functional characteristics of the *Sulfolobus solfataricus* strain 98/2 archaeal genome. In

case of transcriptome profiling work, we used PaCE as a pre-assembler to identify potential transcripts of a gene(s). Thus, we were able to obtain useful information for orphan species (apple, pear and cherry) which have no prior genomic information available. Moreover, handling huge data sets for eukaryotic genomes such as these have been addressed through parallel computation implemented in PaCE. Certain stages of research can be impeded by the absence of tools to perform a certain task. It is here that generic programs such as the restriction enzyme finder developed for initial preparation of cDNA libraries required for transcriptome profiling play a vital role. This program has a wide scope of usage as an experimentally feasible set of restriction enzymes can be found for any set of input sequences for different base – pair ranges or various sequencing technologies. Thus, such programs reduce the time and costs involved in obtaining the results through molecular biology techniques and the need for them is constantly rising.

We have overcome a number of challenges by employing certain techniques some of which are stated above but others such as automation of the computational pipeline continue to exist. The transcriptome profiling work helped identify and predict the functions of genes responsible for certain interested traits depending on the fruit tissue type.

The next section deals with future prospects of the general challenges faced during comparative analysis.

5.1 Future work

At every step of genome or transcriptome analysis, a conventional pipeline cannot be set up by simply providing the output of one stage to a tool in the next stage. It involves pre-processing or post-processing of sequence data that transcends different interfaces or environments. An

example for the above scenario can be seen in various stages of the transcriptome profiling work as in Figure 1.2. The computational method of finding the feasible set of enzymes leads to the preparation of cDNA libraries with those enzymes using biological techniques making these two consecutive stages impossible to automate. Even in the case of processing the 3' UTR reads beforehand and consequently feeding the data for clustering using PaCE, it is not easy to automate as it involves movement of data from a single processor to a cluster of parallel processors. Thus, we see the problems posed by the amalgamation of computational and biological techniques. The recent advancements such as cloud computing (Vaquero et. al 2009, Buyya et. al 2008) can be used to overcome the barriers of crossing different computational environments to process and analyze data. Although, cloud computing can be considered as a good alternative in the future, data security and storage issues can arise (Vouk et. al 2008).

The next two sections describe the future work with respect to the archaeal genome and transcriptome profile works.

5.1.1 Future work for Sulfolobus solfataricus strain 98/2

One cannot filter unique repeat sequences as the *S.solfataricus* strain 98/2 genome contains duplication of IS elements as in strain P2 leading to complex assembly. Thus, one of the following methods can be used to orient the repeat elements within the rest of the genome contigs:

1. A hybrid approach implementing Solexa sequencing to fill the gaps in strain 98/2 sequences obtained from 454 technology can increase the quality of the draft genome sequence (Aury et. al 2008).
2. Alternatively a pair – end BAC (Bacterial Artificial Chromosome) sequencing can be

used to obtain a finished genome sequence.

Preliminary studies were conducted, adjusting the overlap and percentage identity parameters in gsAssembler. The results were compared for characteristics like average length of contig, number of contigs, singletons, repeats and partially assembled reads generated in each case as seen in the following Figures 5.1- 5.6 .

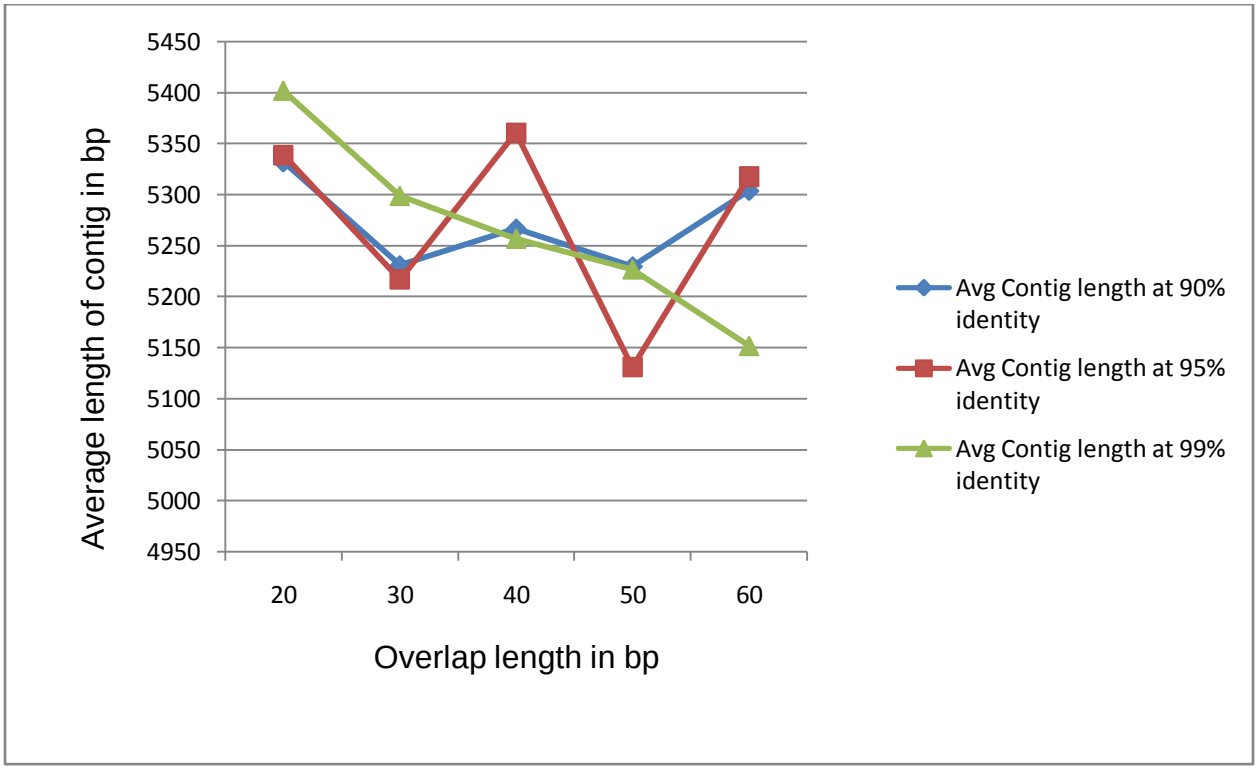


Figure 5.1 Graph of average length of contigs in base pairs for varying overlap lengths and identities.

Average length of contigs at 90% identity (used in our study) is very close to that at 99% identity for 40 bp overlap.

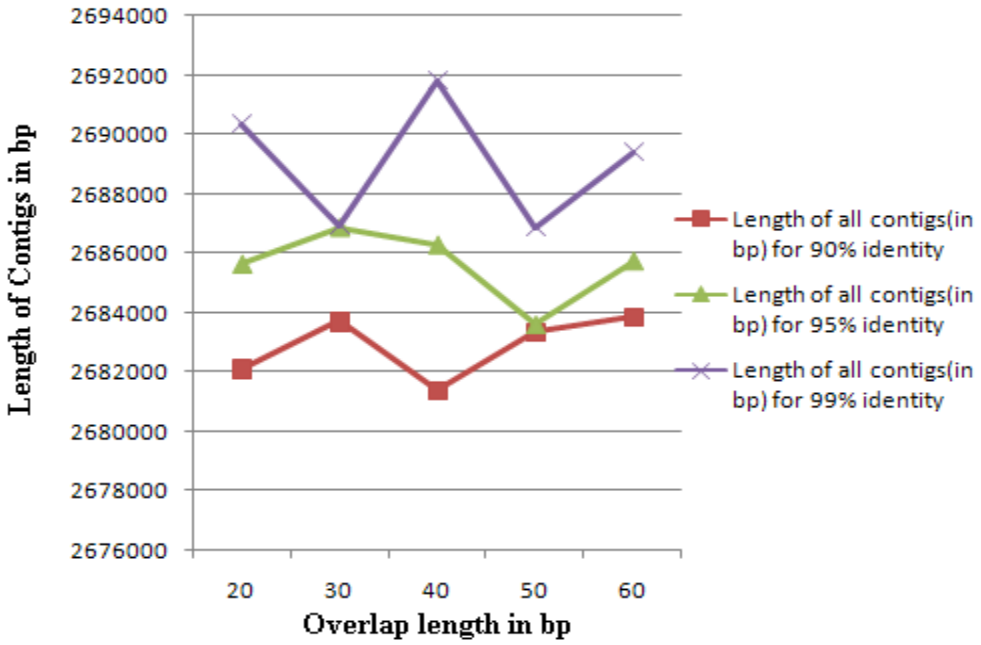


Figure 5.2 Graph of total length of all contigs in base pairs for varying overlap lengths and identities.

The length of all contigs at 90% identity (used in our study) is the lowest at 40 bp overlap.

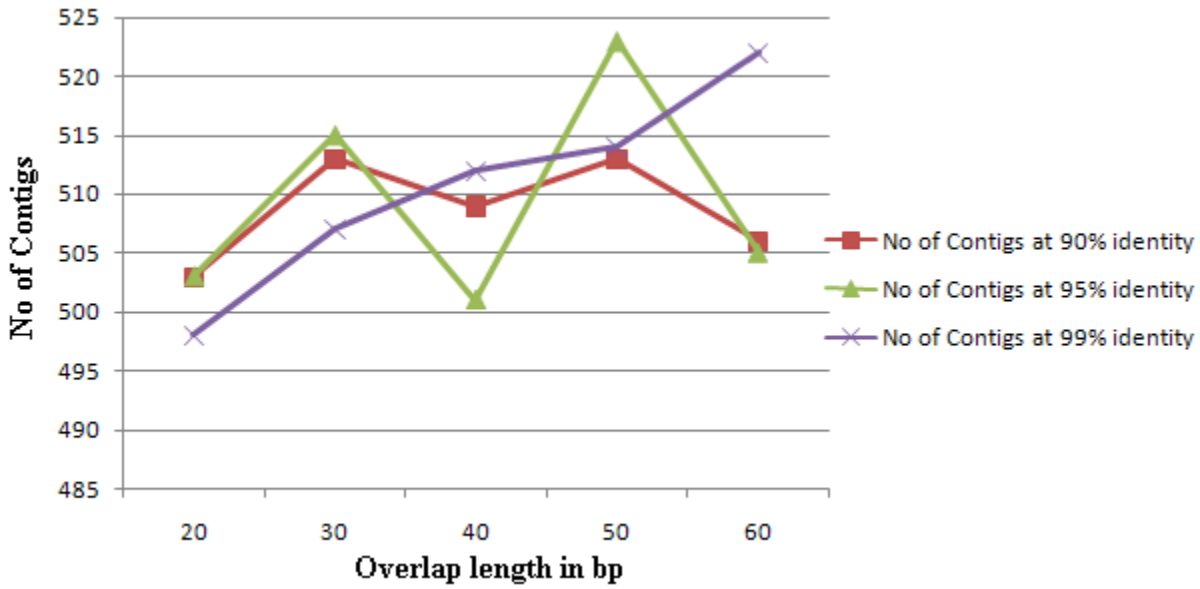


Figure 5.3 Graph of total number of contigs for varying overlap lengths and identities.

We observe that the number of contigs at 90% identity (used in our study) is very close to that at 99% identity at 40 bp and 50 bp overlaps.

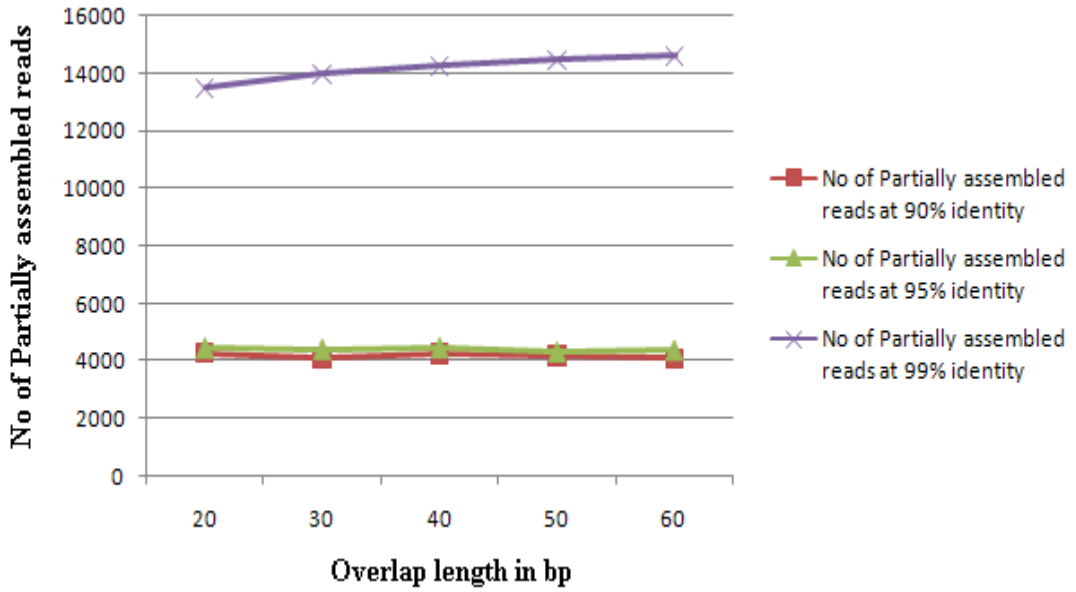


Figure 5.4 Graph of number of partially assembled reads for varying overlap lengths and identities.

The number of partially assembled reads at 90% identity (used in our study) is identical to that at 95% identity for all overlaps.

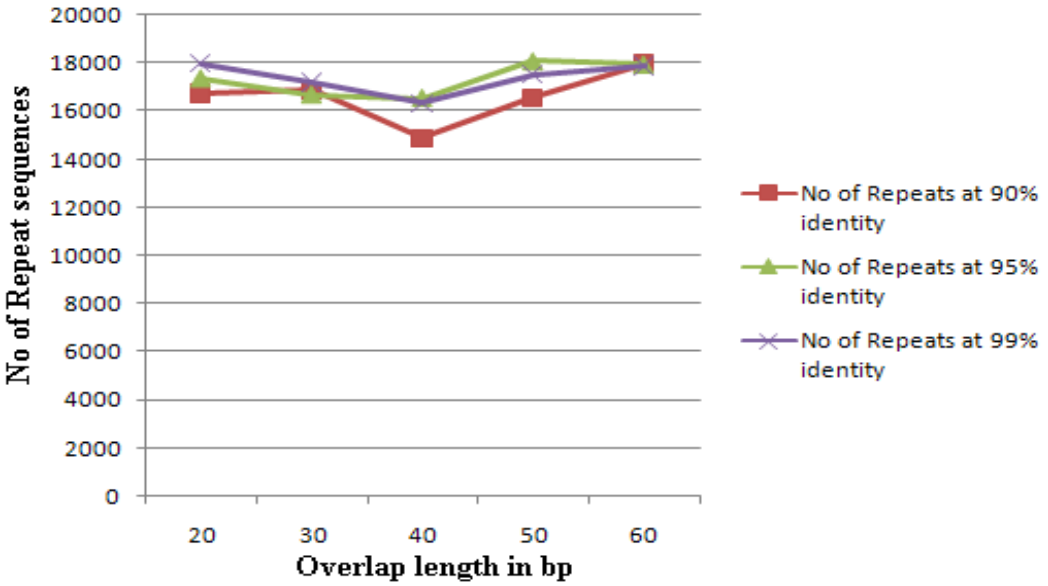


Figure 5.5 Graph of number of repeat elements for varying overlap lengths and identities.

It is observed that the number of repeat reads at 90% identity (used in our study) is identical to that at 95% identity and 99% for at 60bp overlap but lowest at 40 bp overlap (used in our study).

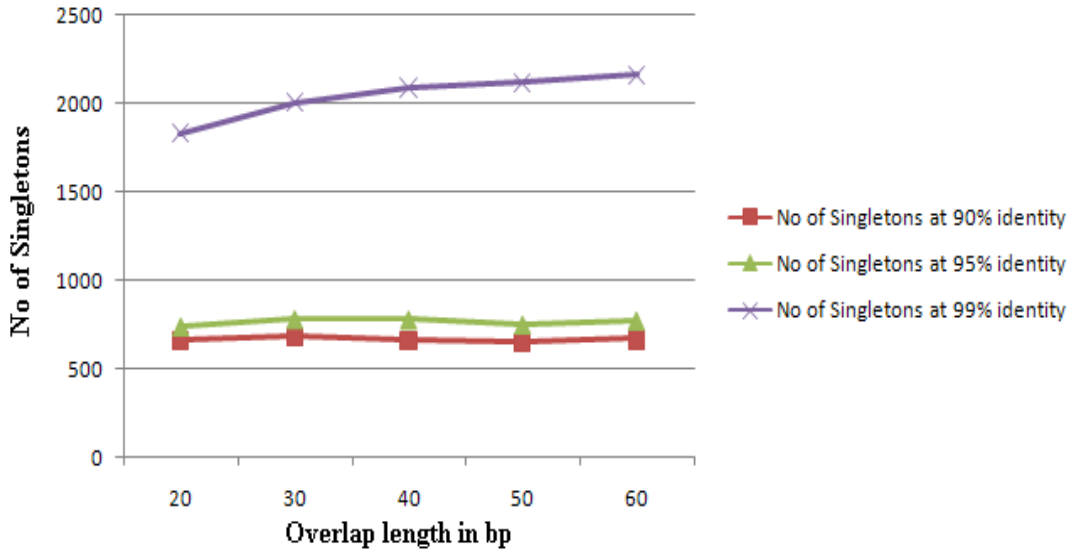


Figure 5.6 Graph of number of singletons for varying overlap lengths and identities

We observe that the number of singletons at 90% identity (used in our study) is very close to that at 95% identity throughout but lowest among all overlap parameters.

Hence, depending on the user two or more feasible characteristics for a project can be chosen and depending on their trends for varying identities and overlaps, the appropriate parameters can be set for running the gsAssembler. Similar studies can be conducted on the parameters for the gsMapper if so desired. The only constraint in these studies is that we cannot choose several characteristics to be compared simultaneously as one genomic characteristic is dependent on the numbers of the other. In our study if average length of contigs and number of contigs are considered, both 90% and 99% identities at 40bp overlap work well. Observing the number of singletons, we can eliminate the 99% identity parameter as it causes a very high number of singletons to be formed. Parametric studies if conducted prior to assembly of genomes can possibly yield better results.

Visualization of the genome especially when few annotated features are found after the

gaps are filled can prove to be very useful for genomicists. This can be enabled through tools like Artemis (Rutherford et.al 2000) and GBrowse (Stein et.al 2002).

5.1.2 Future work for transcriptome profiling in orphan species

We have successfully demonstrated a new way to study the differentially expressed gene(s) for apple, pear and cherry all of which are classified as orphan species.

In the trancriptome profiling work in this thesis only the comparison between various apple and pear sample varieties were taken into consideration but the research can be extended to compare the expression of their genes collected at various time points or growth stages. The statistical analysis extended to the information is all clusters is too complex and is currently out of scope for this thesis. Since no methods exist to evaluate the nature of such data, we need to devise a complex statistical approach to solve our problem of missing good gene information and discarding false positives.

Further exploration of other molecular biology techniques for the purpose of validation of statistical results is required.

BIBLIOGRAPHY

*Altschul, S.F., W. Gish, W. Miller, E.W. Myers, D.J. Lipman.(1990). Basic local alignment search tool. J Mol Biol . 215: 403-10.

**Alwine, J.C., D.J. Kemp, G.R. Stark. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc Natl Acad Sci USA. 74(12): 5350-5354.

*Auernik, K.S., Y. Maezato, P.H. Blum, R.M. Kelly. (2008). The genome sequence of the metal-mobilizing, extremely thermoacidophilic archaeon Metallosphaera sedula provides insights into bioleaching-associated metabolism. Appl Environ Microbiol. 74: 682-92.

Aury, J.M., C. Cruaud, V. Barbe, O. Rogier, S. Mangenot, G. Samson, J. Poulain, V. Anthouard, C. Scarpelli, F. Artiguenave, P. Wincker. (2008). High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics. 9: 603.

Batzoglou, S., D.B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J.P. Mesirov, and E.S. Lander. (2002). ARACHNE: A Whole-Genome Shotgun Assembler. Genome Res. 12(1): 177-189.

Benson, D.A. , M.S. Boguski, D. J. Lipman, J. Ostell, B.F.F. Ouellette, B.A. Rapp and D.L. Wheeler. GenBank. (1999). Nucleic Acids Research. 27(1): 12-17.

Benson, D.A. , I.K. Mizrachi, D. J. Lipman, J. Ostell and E.W. Sayers.GenBank. (2008). Nucleic Acids Research. 37: D26-D31.

*Besemer, J., M. Borodovsky. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 33: W451-4.

Bikandi, J., R. San Millán, A. Rementeria, and J. Garaizar. (2004). *In silico* analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction. Bioinformatics**.** 20(5): 798-799.

*Brock, T.D., K.M. Brock, R.T. Belly, R.L. Weiss. (1972). Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. Arch Mikrobiol. 84: 54-68.

Burge, C.B., and S. Karlin (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268: 78-94.

Burge, C. B. (1998). Modeling dependencies in pre-mRNA splicing signals. In Salzberg, S., D. Searls, and S. Kasif, eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam. 127-163.

Burge, C. B., and Karlin, S. (1998) Finding the genes in genomic DNA. Curr Opin Struct Biol. 8: 346-354.

Buyya, R., C.S. Yeo, S.Venugopal. (2008). Market-oriented cloud computing : Vision, hype and reality for delivering IT services as computing utilities. Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications.

Chaisson, M., P. Pevzner and H. Tang. (2004). Fragment assembly with short reads. Bioinformatics. 20(13): 2067-2074.

Chandler, M., and Mahillon, J.(2002) Insertion Sequences Revisited. In: Craig N.L., R.Craigie, M. Gellert, A.M. Lambowitz, editors. Mobile DNA II. Washington D.C.: American Society for Microbiology. 305–366.

*Chen L., K. Brugger, M. Skovgaard, P. Redder, Q. She, E. Torarinsson, B. Greve, M. Awayez, A. Zibat, H.P. Klenk. (2005). The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota. J Bacteriol. 187: 4992-9.

**Cormen, T., C. Lieserson, R. Rivest, C. Stein. Introduction to Algorithms, McGraw-Hill Science/EngineeringMath; 2nd edition, 2003.

Crick , F.H.C. and J.D. Watson. (1954). Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. 223(1152): 80-96.

Crick, F. (1958). Ideas on protein synthesis. Symp. Soc. Exp. Biol. XII, 139-163.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature. 227: 561-563.

Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. (1999). Improved microbial gene identification with GLIMMER . Nucleic Acids Research. 27(23): 4636-4641.

Delcher, A.L., K.A. Bratke, E.C. Powers, and S.L. Salzberg. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 23(6): 673-679.

*Delcher, A.L., A. Phillippy, J. Carlton, S.L. Salzberg. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30: 2478-83.

*Delcher, A.L., S. Kasif, R.D. Fleischmann, J. Peterson, O. White, S.L. Salzberg. (1999). Alignment of whole genomes. Nucleic Acids Res. 27(23): 69-76.

**Eveland, A.L., D.R. McCarty, K.E. Koch. (2008). Transcript Profiling by 3'-Untranslated Region Sequencing resolves Expression of gene Families. Plant Physiology. 146: 32-44.

Fedurco, M., A.Romieu, S. Williams, I. Lawrence and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. (2006) Nucleic Acids Res. 34(3): e22.

Florea, L., C. Riemer, S. Schwartz, Z. Zhang, N. Stojanovic, W. Miller and M. McClelland. (2000). Web-based visualization tools for bacterial genome alignments. Nucleic Acids Res. 28(18): 3486-3496.

Galas,D.J., and Chandler,M. (1989) Mobile DNA. American Society for Microbiology, Washington, DC. 109-162.

*Gao, F., C.T. Zhang. (2006). GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. Nucleic Acids Res. 34: W686-91.

Genome Sequencer Data Analysis Software Manual. (Software version 2.0.00, 2008). Available at http://sequence.otago.ac.nz/download/GS_FLX_Software_Manual.pdf .

Green, P . (1996). Phrap documentation. Available at http://www.phrap.org/phredphrap/phrap.html .

Grogan, D.W. (1989). Phenotypic Characterization of the Archaebacterial Genus Sulfolobus: Comparison of Five Wild-Type Strains. J Bacteriol. 171(21): 6710-6719.

*Haseltine, C., M. Rolfsmeier, P. Blum.(1996). The glucose effect and regulation of alpha-amylase synthesis in the hyperthermophilic archaeon Sulfolobus solfataricus. J Bacteriol. 178: 945-50.

Hsieh, S.J., C.Y. Lin, Y.S. Chung, and C.Y. Tang. (2005). Comparative Exon Prediction based on Heuristic Coding Region Alignment. Proceedings of the International Symposium on Parallel Architectures, Algorithms, and Networks. 14-19.

**Huang, X., and A. Madan. (1999). CAP3: A DNA sequence assembly program .Genome Res. 9: 868-877.

Jacquemet, A., J. Barbeau, L. Lemiègre, T. Benvegnu. (2009). Archaeal tetraether bipolar lipids: Structures, functions and applications. Biochimie. 91 : 711-717.

Kalyanaraman A., S. Aluru, S. Kothari and V. Brendel. (2003). Efficient clustering of large EST data sets on parallel computers. Nucleic Acids Res. 31(11): 2963 -2974.

*Kawarabayasi, Y., Y Hino, H Horikawa, K Jin-no, M Takahashi, M Sekine, S Baba, A Ankai, H Kosugi, A Hosoyama.(2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, Sulfolobus tokodaii strain7. DNA Res . 8: 123-40.

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

Larsen, T.S., and A. Krogh. (2003). EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics. 4: 21.

Lashkari, D.A., J.L. DeRisi, J.H. McCusker, A.F. Namath, C. Gentile, S.Y. Hwang, P.O. Brown, R.W. Davis. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA. 94: 13057–13062.

Lewin, B. (2004). Genes VIII. Pearson- Prentice Hall, NJ.

Liang, P., L. Averboukh, and A.B. Pardee. (1993). Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization. Nucleic Acids Res. 21(14): 3269–3275.

*Lukashin, A.V., M Borodovsky. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26: 1107-1115.

Mahillon, J., and M. Chandler. (1998). Insertion Sequences. Microbiology and Molecular Biology Reviews. 62(3): 725-774.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Goodwin, W. He, Scott Helgesen, C.H. Ho, G.P. Irzyk , S.C. Jando, M.L.I. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, BP. Puc, M T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M. P. Weiner, P. Yu, R.F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. Nature. (2005). 437: 376–380.

Marone, M., S. Mozzetti, D. De Ritis, L. Pierelli and G. Scambia. (2001). Semiquantitative RT-PCR analysis to assess the expression levels of multiple transcripts from the same sample. Biol Proced Online. 3: 19-25.

Milanesi, L., and Rogozin I.B. (1998). Prediction of human gene structure. In: Guide to Human Genome Computing (2nd ed.) (Ed. M.J.Bishop) Academic Press, Cambridge. 215-259.

Milanesi, L., D. D'Angelo, I.B. Rogozin. (1999). GeneBuilder: interactive in silico prediction of genes structure. Bioinformatics. 15 (7): 612-621.

Morozova, O., and M.A. Marra. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics. 92(5): 255-264.

Myers, E.W., G.G. Sutton , A.L. Delcher, I.M. Dew , D.P. Fasulo, M.J. Flanigan,S.A.Kravitz, C.M. Mobarry, K.H. Reinert , K.A. Remington, E.L. Anson, R.A. Bolanos, H.H. Chou, C.M. Jordan, A.L. Halpern, S. Lonardi, E.M. Beasley, R.C. Brandon, L. Chen, P.J. Dunn, Z. Lai, Y. Liang, D.R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G.M.Rubin, M.D. Adams, J.C. Venter. (2000). A whole-genome assembly of Drosophila. 287(5461) : 2196-204.

Murray, P.R., K.S. Rosennthal, M.A. Pfaller. (2005). Medical Microbiology. 5th Edition, Elsevier.

Nielsen, P., and A. Krogh. (2005). Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics. 21: 4322-4329.

Palmer, M. An on-line tool for restriction analysis, silent mutation scanning, SNP-RFLP analysis. Available at http://watcut.uwaterloo.ca/watcut/watcut/template.php.

Redder, P., R.A. Garrett. (2006). Mutations and rearrangements in the genome of Sulfolobus solfataricus P2. J Bacteriol. 188: 4198-206.

*Rolfsmeier, M., P. Blum. (1995). Purification and characterization of a maltase from the extremely thermophilic crenarchaeote Sulfolobus solfataricus. J Bacteriol. 177: 482-5.

*Rolfsmeier, M., C. Haseltine, E. Bini, A. Clark, P. Blum. (1998). Molecular characterization of the alpha-glucosidase gene (malA) from the hyperthermophilic archaeon Sulfolobus solfataricus. J Bacteriol. 180: 1287-95.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen & P. Nyren. (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242: 84–89

*Rosa, M.D., A. Gambacorta, J.D. Bu'lock. (1975). Extremely thermophilic acidophilic bacteria convergent with Sulfolobus acidocaldarius. J Gen Microbiol. 86: 156-64.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.A. Rajandream, B. Barrell. (2000). Artemis: sequence visualization and annotation. Bioinformatics. 16: 944-5.

** Saha, S., A.B. Sparks, C. Rago , V. Akmaev , C.J. Wang, B. Vogelstein, K.W. Kinzler, V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, P. Hieter, B. Vogelstein, and Kinzler KW. (1997). Characterization of the yeast transcriptome. Cell. 88: 243-251. Salzberg, S., A. Delcher, S. Kasif and O. White. (1998). Microbial gene identification using interpolated Markov models. Nucleic Acids Research. 26(2): 544-548.

Sanger, F., G.M. Air, B.G. Barrell, N.L. Brown, A.R. Coulson, J.C. Fiddes , C.A. Hutchison, P.M. Slocombe & M. Smith. (1977). Nucleotide sequence of bacteriophage ΦX174 DNA. Nature. 265: 687 – 695.

**Sanger, F., S. Nicklen and A.R. Coulson. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74: 5463–5467.

*Schelert, J., M. Drozda, V. Dixit, A. Dillman, P. Blum. (2006). Regulation of mercury resistance in the crenarchaeote Sulfolobus solfataricus. J Bacteriol. 188: 7141-50.

**Schena, M., D. Shalon, R.W. Davis, P.O. Brown. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 270: 467–470.

*She, Q., R.K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M.J. Awayez, C.C. Chan-Weiher, I.G. Clausen, B.A. Curtis, A.D. Moors. (2001). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci USA . 98: 7835-40.

Shendure, J., R.D. Mitra, C. Varma & G.M. Church. (2004). Advanced sequencing technologies: methods and goals. Nature Reviews Genetics. 5: 335–344.

Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, G.M. Church.(2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 309:1728–1732.

Shendure, J., and H. Ji. (2008). Next –generation DNA sequencing. Nature Biotechnology. 26: 1135-1145.

Song, J., Y. Xu, S. White, K.W.P. Miller and M. Wolinsky.(2005). SNPsFinder – A Web-Based Application for Genome-Wide Discovery of Single Nucleotide Polymorphisms in Microbial Genomes. Bioinformatics. 21(9): 2083-2084.

Soper, D.S., (2009). The Free Statistics Calculators Website.Online Software. Available at http://www.danielsoper.com/statcalc/.

Stein, L.D. (2002) The generic genome browser: a building block for a model organism system database. Genome Res. **12**: 1599-610.

Swerdlow, H., S. Wu, H. Harke & N.J. Dovichi. Capillary gel electrophoresis for DNA sequencing: Laser-induced fluorescence detection with the sheath flow cuvette. (1990). J. Chromatogr. 516: 61–67.

Sutton G. G., O. White, M.D. Adams, A.R. Kerlavage. (1995). TIGR Assembler: A new tool for assembling large shotgun sequencing projects. Genome Science and Technology. 1(1): 9-19.

Turcatti ,G. , A. Romieu, M. Fedurco & A.P Tairi. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res. 36, e25.

Vaquero, L.M., L. Rodero-Merino, J. Caceres, M. Lindner. (2009). A break in the clouds: Towards a cloud definition. ACM SIGCOMM Computer Communication Review. 39(1): 50-55.

**Velculescu, V.E., L. Zhang , B. Vogelstein, and K.W. Kinzler. (1995). Serial Analysis Of Gene Expression. Science. (270): 484-487.

Veterinary Dictionary. Blood D.C, V.P. Studdert and C.C. Gay, Elsevier. (2007).

Vincze, T., Posfai, J. and Roberts, R.J.(2003). NEBcutter: a program to cleave DNA with restriction enzymes.Nucleic Acids Res. 31: 3688-3691. Available at http://tools.neb.com/NEBcutter2/index.php.

Vouk, M.A. (2008). Cloud Computing- Issues, Research and Implementations. Proceedings of ITI 30th Int. Conf. on Information Technology Interfaces.Cavat,Croatia.

Woese, C.R., O. Kandler, M.L. Wheelis (1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. 87: 457- 4579.

*Worthington, P., V. Hoang, F. Perez-Pomares, P. Blum.(2003). Targeted disruption of the alpha-amylase gene in the hyperthermophilic archaeon Sulfolobus solfataricus. J Bacteriol. 185: 482-8.

*** References from manuscript in Chapter 3**

**** References from manuscript in Chapter 4**

APPENDIX

BIOLOGICAL BACKGROUND

This section gives a brief background about the fundamentals of molecular biology useful to understand certain concepts presented in this thesis.

Classification of all living organisms

All living organisms can be classified into one of the three domains namely ‘Bacteria’ , ‘Archaea’ and ‘Eukaryota’.

In general, organisms that do not have a nucleus in their cells are called *prokaryotes*. Bacteria are prokaryotic micro-organisms that are single celled and do not have a nucleus. All the genetic material they possess are in the cytoplasm of their cells. Typically bacteria contain a single chromosome that is circular in shape. There are both useful and harmful classes of bacteria. A few examples include *Lactobacillus* (used in yoghurt and cheese making), *salmonella* (causes food poisoning) and *Escherichia coli* (some strains cause food poisoning) (Medical Microbiology , 5th edition, Murray, Rosenthal, Pealler).

Archaea are also unicellular and lack nucleus and other cell organelles in their cell. The individuals belonging to this domain are termed 'archeon' and they are similar to bacteria but have some genes and enzymes present in eukaryotes. Archeons have a single circular chromosome too. Archaea, bacteria and eukaryotes can have extra-chromosomal elements called plasmids. In bacteria plasmids can provide antibiotic resistance. Plasmids may also help in fixation of compounds or elements such as nitrogen and sulfur. In genetic labs, plasmids are very useful as they are used to make several copies of a gene or express a specific gene. Two phyla namely ‘Crenarchaeota’ and the

‘Euryarchaeota’ under the domain archaea are the most studied classes of organisms since members of these branches are the only archaeons that have been successfully cultivated. Archeons have a wide range of habitats and also thrive in extreme environmental conditions. They have numerous uses in the field of biotechnology (Jacquemet et. al 2009).

Several thermophiles (resistance to high temperature ranges) and hyperthermophiles in Crenarchaeota have been studied especially the *Sulfolobus* species (Grogan et. al 1989). This thesis discusses the sequencing and comparative genetic studies of *Sulfolobus solfataricus* strain 98/2, an extremophile belonging to the Phylum: Crenarchaeota.

Eukaryotes are single celled or multicellular organisms that have a well defined nucleus and membrane bound organelles like mitocondria, Golgi apparatus , endoplasmic recticulum and chloroplasts. The process of cell division in eukaryotes is different than bacteria and archaea. Plants , animals, fungi and protists (algae, amoebae) are classified as eukaryotes. In the Plant Kingdom , flowering plants termed 'Angiosperms' exist. Of these we are interested in a division of land plants called *Magnoliophyta*. In this division, the family *Rosaceae* is where apple, pear and cherry fruits belong. In this thesis, we perform the trancriptome analysis of the above mentioned fruits, the genomes of which have not yet been sequenced.

DNA

DNA stands for deoxyribonucleic acid present in every living cell. It has the hereditary or genetic information required for construction and working of living organisms. DNA contains genes and other elements that regulate the use of the genetic material present. DNA is a long chain of nucleotides supported by a framework of sugar and phosphate bonds. It consists of two strands running in opposite directions forming a double helix structure as proposed by Watson and Crick using data collected by Rosalind Franklin and Maurice Wilkins (Crick and Watson 1954). There are four types of *nucleotides*

namely adenine, guanine, cytosine and thymine (A, G, T and C). Each nucleotide or *base* is attached to the sugar. A gene has a coding sequence within it that codes for an RNA ("ribonucleic acid") or protein product through a process known as *transcription*. The process follows the *genetic code* which is a set of nucleotide triplets in the coding sequence of a gene called "codon". There are 64 possible codons that map to 20 amino acid residues. A protein is a sequence of amino acid residues. Plants store most of their DNA in the nucleus of their cells and some in mitochondria and chloroplasts. Bacteria and archaea store DNA in the cytoplasm portion of their cells.

The structure of DNA is as shown below in figure 1 with 2 strands , one running in the 5'-3' direction and the other in the 3'-5' direction.

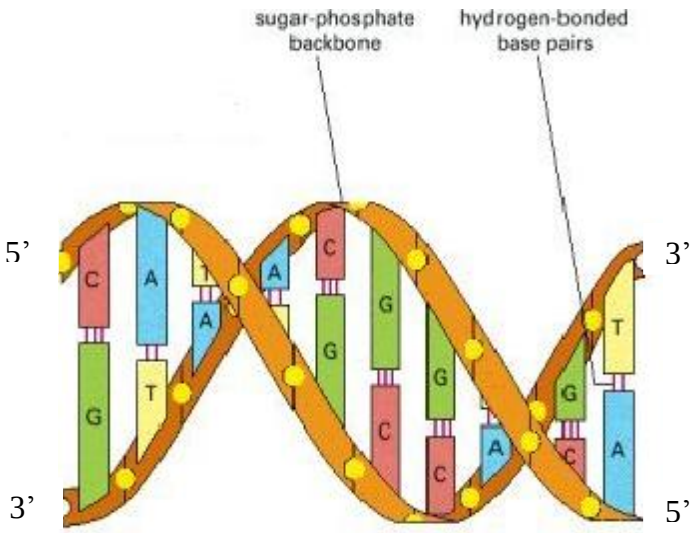


Figure A.1 Structure of DNA with top strand in 5'- 3' direction and bottom strand in 3'- 5' direction. (Source : Google Images)

The base G binds with base C in triple hydrogen bonds whereas base A binds with base T in double hydrogen bonds. This base-pairing rule is referred to as *base complementarity.* The GC content of organisms is the percentage of Gs and Cs with respect to all nucleotides.

RNA

RNA is Ribonucleicacid, single stranded and is the bridge between DNA and protein formation. The nucleotides present in RNA are the same as in DNA expect that the base Thymine (T) is replaced by Uracil (U).

RNA polymerase enzymes help in transcribing DNA to RNA. There are many kinds of RNA like mRNA (messenger RNA) , rRNA (ribosomal RNA) , tRNA (transfer RNA). The mRNA is carries genetic information to the ribosomes which are composed of rRNAs and proteins. These ribosomes in turn translate the information in mRNA to proteins. The structure of DNA and RNA is illustrated in Figure 2.

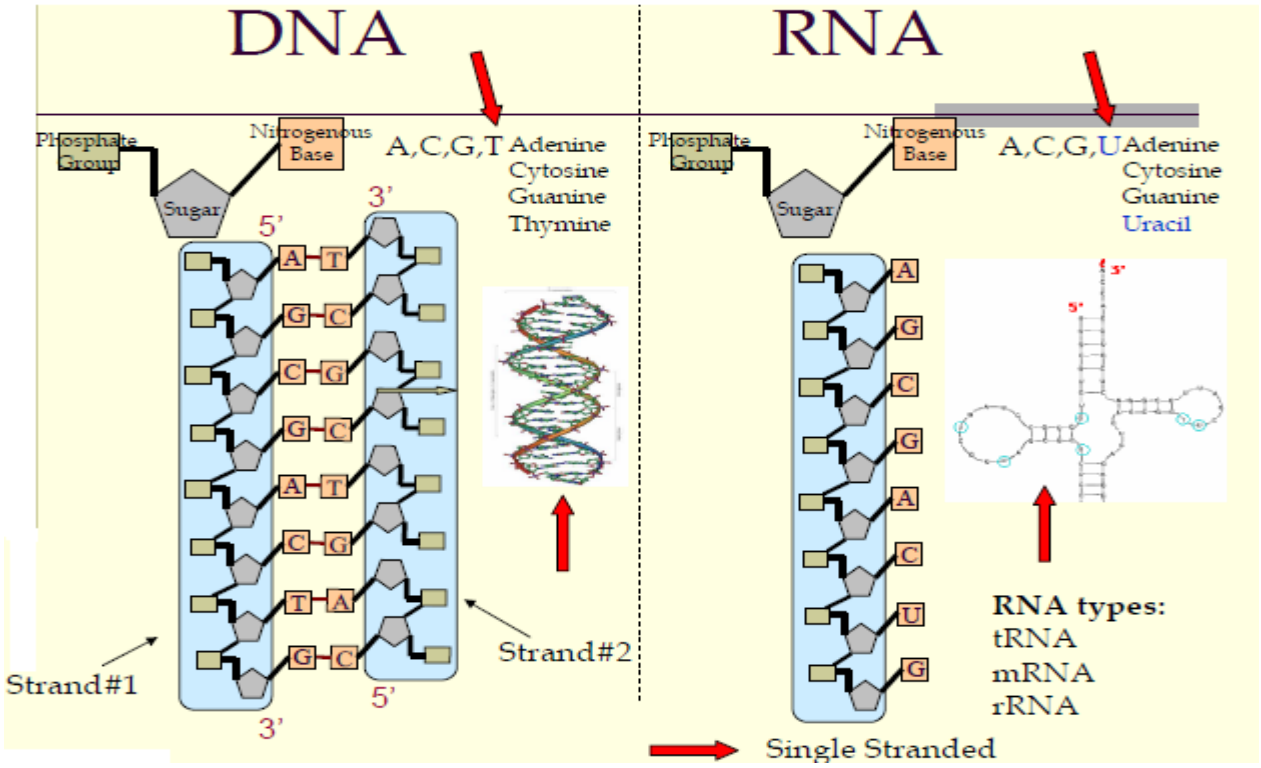


Figure A.2 Comparison the structures of DNA and RNA.

(Source : http://www.eecs.wsu.edu/~ananth/CptS580/Lectures/IntroToCompBio.pdf)

Genes

Genes is the basic unit of heredity in all living organisms and is responsible for a given trait. A gene is a contiguous stretch within a DNA molecule which codes (or "transcribes") for either an RNA product or eventually a protein product. The coding sequences are termed *exons* and the non coding sequences are *introns*.

Alleles

Alternate forms of a gene are termed as *alleles*. For instance, if there is a gene that controls the height of a plant some of the member may be short while others are tall. Thus, the gene trait seen in each plant depends on which allele is present in them.

Genotype

The entire genetic or hereditary information of a living organism regardless of what traits are expressed is termed *genotype* of that organism.

Phenotype

What can be observed in an organism in terms of physical form or structure, growth and behavior is termed its *phenotype*.

Transcription

Transcription is the process by which a fully processed or mature mRNA is obtained from a gene using RNA polymerase.

Translation

In prokaryotes, the mRNA does not require any transportation as all genetic material is present in the cytoplasm. Hence, the ribosome can begin translation of the mRNA code to proteins as soon as transcription is done and termed co-transcription. However, in Eukaryotic cells mRNA moves out of the nucleus into the cytoplasm and the translation happens either in the free ribosomes of the cytoplasm or the ribosomes within the endoplasmic recticulum (organelle).

The central dogma of molecular biology as proposed by Francis Crick in 1958 is that sequence information in biological systems flows in only one direction that is from DNA to RNA and RNA to proteins (Crick et. al 1958, 1970). This theory is simplified for the purpose of this thesis and has been diagrammatically represented in Figure 3.

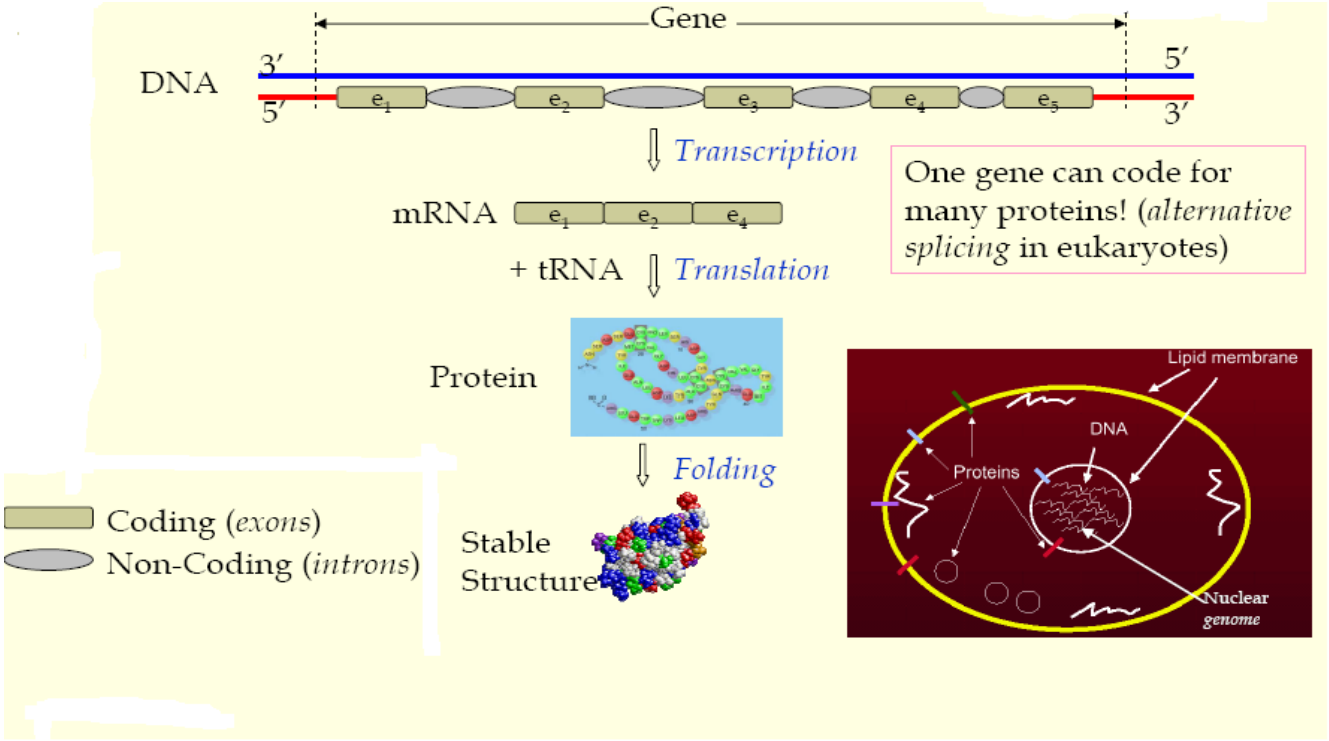


Figure A. 3 Schematic representation of Central Dogma of Molecular Biology (DNA → RNA → Proteins)

(Source : http://www.eecs.wsu.edu/~ananth/CptS580/Lectures/IntroToCompBio.pdf)

The typical structure of mRNA is as seen in Figure 4, but the mRNA especially in eukaryotic cells could be in the pre-processed state or degraded form. The 5' and 3' ends have *untranslated regions* (*UTR*) as seen. The coding sequence begins with a specific start codon (AUG) and ends with a specific stop codon (UGA,UAA,UAG). The start and stop codons help the DNA polymerase to recognize the start and end points of the genetic code during translation. The 5' cap region consists of a modified version of guanine nucleotide used for the binding of mRNA to ribosome. A poly A tail is a stretch of adenine bases used for protecting mRNA from degradation. An mRNA molecule is referred as transcript.

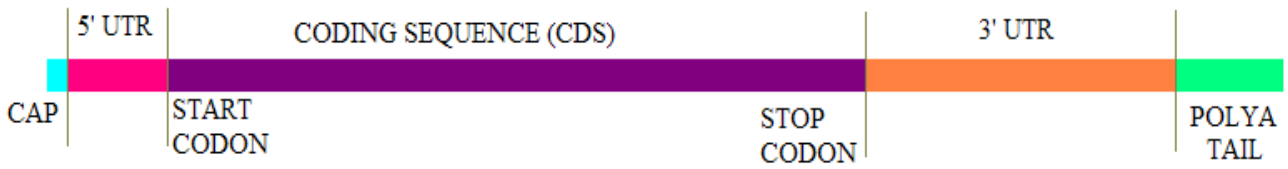


Figure A.4 General structure of mRNA in Eukaryotes.

The untranslated regions are believed to contribute to factors like stability and localization of mRNA. Also, the efficiency or at times inhibition of the translation process is controlled by untranslated regions. We are interested in the 3' UTR portion of the mRNA for transcriptome studies as it is considered to affect the expression of genes.

cDNA

cDNA stands for complementary DNA and is synthesized from mRNA by using the enzyme reverse transcriptase. cDNA libraries can be prepared from specific tissues of an organism and contains only

the expressed genes.

Expressed sequence tags (*ESTs*) are portions derived from cDNA sequences. They are used in gene discovery, and gene expression studies.

Genome

The collection of all genetic information present in the cell of an organism including nucleus and other organelles is collectively termed as the organism's *genome*. It may include non-chromosomal elements such as plasmids, viruses and mobile genetic elements such as transposons.

Transcritptome

A collection of mRNA molecules in a particular organism is termed *trancriptome*. Equivalently, it collectively refers to the collection of all transcripts in a particular tissue or cell type.

FASTA format

The sequence information can be stored in a FASTA file format where the sequence has a ">" symbol followed header information in one line and the next line as the base pair information. These files have extensions as .fa, .fna, .fasta and .fas .An example for this format is as seen below:

```
>Contig00001 length=234 numreads=250
AGTCGATCGTAGCTAGCTAGTCGATCTATCGTATCGTATCGTCTATCTATTCGCGCCGCGGC
CTGCTAGCTGATCGTAGCTAGCT
```

Paired end reads

The two ends of a single DNA molecule can be sequenced with appropriate adaptors for these ends. These DNA sequences are termed as paired end reads.