Dissertation title: Perils of Genome Assembly: Data types and sequencing platform defines optimal genome assembly in prokaryotes and eukaryotes

Thesis Submitted to AcSIR for the Award of the Degree of

DOCTOR OF PHILOSOPHY

In Biological Science

By

Mathu Malar C

Registration Number: **10BB15J17002**

Under the guidance of

Dr. Sucheta Tripathy

Structural Biology & Bioinformatics Division,

CSIR - Indian Institute of Chemical Biology,

Kolkata

Dedicated to Parents,

Grandparents and PI

ACKNOWLEDGEMENT

Undertaking this PhD is a part of dreams in my life and it would not have been possible to complete the PhD without the visible and invisible support and motivation from many not only human but also other than humans.

First of all, I am extremely grateful to my guide, Dr. Sucheta Tripathy for her continuous support, supervision, guidance and motivation from my first day at her lab. I feel myself luckiest PhD student for getting her guidance and blessings. She is not only a PhD guide for me, she influenced my life a lot and her countless teachings will guide me in my whole life at each and every situation. I learned from her, how to become a good human being, how to become a good leader. I am very grateful to her for providing me full freedom to work; this nature makes her unique among all.

I am very thankful to my collaborator Takao Kasuga (I should say my next mentor or PI), Jennifer Yuzon from UC-Davis, California greatly for providing me the data and guidance and teaching the oomycetes biology. In this collaboration I have earned a very good friend such as Jennifer Yuzon and Kyle Fletcher to discuss about the problems of PacBio and it helped me to solve my research problems.

I would like to acknowledge the UC-DAVIS Genome sequencing facility for providing me the data to analyse.

I would like to specially mention my collaborator Ramesh Vettukuri from Swedish Agricultural University. I am thankful to him for supporting me in all my bad times. His continuous phone calls and emails, watsapp messages that kept me motivated.

I am Thankful to Prof Brett Tyler for teaching me the effector prediction and learnt how to handle data and present data in excel. I thank him for mentoring me in genome project of oomycetes. He helped me to improve the quality of paper and work.

I would like to thank our collaborator Steve Whisson from James Hutton Institute, Dundee for working and improving my manuscript

I would like to thanks Dr Nahid Ali and Dr Suman Khowala my internal Collaborators from IICB, for trusting me and provided me the data. My Hands on with Dr Suman Khowala data was the first time I was introduced to transcriptomics.

I am Thankful to Soumya Di for working with me on Transcriptomics projects. I am also thankful to Roma Di for continuously working with me in *Leishmania* genome projects and constantly supporting me from Australia. I am happy to work with these awesome and cool people.

I am very much inspired and motivated by Dad, Grandpa, Granny and Dr. Abdul Kalam.

I feel myself blessed for getting the chance of staying on the land of sweets, where many great leaders such as Vivekananda, Rabindra Nath Tagore and Subhash Chandra Bose were born.

I would like to gratefully acknowledge my all teachers who taught me everything since my childhood, I learned from them writing, speaking, and even learning also.

Completion of PhD is not an individual's task, many people contributed directly or indirectly in this thesis, I am thankful to all of them.

I am very much thankful to our farmers who produced sufficient amount of food that is the first need to anyone for doing research.

I am very thankful to security staff of IICB who took care of any danger for securing us.

Living in Kolkata for a Non Bengali is a very big challenge but I greatly appreciate my friends, roommates and Ada for teaching me sufficient Bengali that helped me to survive in Kolkata. I appreciate the canteen workers for taking care of my food.

I would like to thank my roommates Moumita and Titli for a delicious Bengali food and fish recepies.

I would like to thank my Gym Instructors and other people who helped me in the journey of fitness.

I would like to thank Pavan my first friend from IICB for helping me to choose right lab for my PhD journey.

I would like to thank SK good friend of me although I know him for a shorter time, but you were with me in all my bad times for motivation.

I would like to thank Murugan from IICB, for approaching me for Badminton tournament and I would like to thank my badminton partners Diya and Nilanjana for baring me. I thank Nilanjana for taking care of me when I was alone. I am thankful to you for our good times in midnight parties at HRC and Roxy and Nepal trip.

I would like to thank Neha (Badi Neha) for being my teacher. She taught me data analysis of NGS. You were my great teacher. I learnt how to work and maintain lab notebooks.

I enjoyed the company of Diya. One of my best friends who helped me in all my bad times and motivated me. I enjoyed her gifts such as various shades of lipstick and dress. I don't forget the Friday nights, marathons, hiking, morning running, breakfast times in south Indian restaurant. She helped me in correcting my thesis, paper and abstract. More than a friend she acted as my sister.

I would like to thank Lubna for helping me in my journey of PhD. She helped me in finding places to stay in Kolkata, provided me the model of thesis to compare, served me delicious food, and made me to visit her hometown.

I would like to thank my landlord Dey's family for renting the flat and treating me with delicious food.

I would like to thank Deeksha, Mayuri (future IAS), Arpita for being my besties and like our lunch times, shopping, movie times we spent in IICB.

I am thankful to Arijit (dirty boy) for sharing the sitting place with me for 5 years. I don't forget the fun, fights (Tom and Jerry), parties and I am thankful to you for making me worse coffee at your room.

I am thankful to my 4 years old friend Ada, my little sunshine whenever I am sad or depressed she made me happy by her smile and hugs. I enjoyed your company from a very first day.

I am thankful to Abhishek (cleanliest boy) for being my stress buster, whenever I am bored, I fight with you and we make fun of each other. More than my lab mate you took a responsibility of brother.

I am thankful to Samrat (Leydhcor and Dirty boy) and Pijush (Dirty boy and gabar) for supporting me in the journey of PhD.

I am thankful to Subhadeep (cleanliest boy) in our lab, who had thought me how to work smart.

I enjoyed the company of other lab mates Aditya, Sashi and I am thankful to them for making my PhD journey memorable.

I am thankful to former lab mates Rajeev Da, Gyan Prakash Mishra (Motu), Neha (choti), Sunanda, Vineeta, Madhavi, Sushma, Arup Ghosh, Ajay, Anindya, and Vikas for the wonderful memories.

I thank my Niper students Anusha, Monika, Madhu Shankar for being a good student.

I would like to thank my Dad, Granny, brother (Hen) for giving my freedom to do PhD. I also would like to thank all my cousins Arun, Preethi, Moj, Indu, Laddu, Janani, Jimmy, prashant, Ram, Ragav, Viji, Anjana, Anjali, Arjun, Dee, Aksha, and Ashok. I would like thank my mama, athai, pinni and chinna for supporting me tirelessly.

I am grateful to the Council of Scientific and Industrial Research (CSIR, Govt. of India) for providing me with financial assistance for continuing the research work. I also express my gratitude to Prof. Siddhartha Roy, Ex-Director, CSIR - IICB, and Prof. Samit Chattopadhyay, Director, CSIR - IICB for letting me, be a part of the wonderful research fraternity at Indian Institute of Chemical Biology (IICB). I am also grateful to Academy of Scientific and Innovative Research (AcSIR) for giving me permission for PhD registration.

I would like to thank my funding agency DST- Inspire for providing me fellowship.

I would like to thank my twitter friends, ISCB friends for having a good suggestions, discussion and parties.

I am also thankful to AcSIR DAC members Dr. Sharmila chattopadhyay, Dr Jayati Sengupta, Dr Chitra Dutta (former member), Dr Siddhartha Roy for their valuable suggestions for my PhD.

Most important my all friends they always stood with me regardless of I was wrong or right. I do not have the words to acknowledge my friends.

I would like to acknowledge all my servers corona, amrit, ada, apala, ajeya, cmacs, Aditya for not ditching me and saved my data carefully. Without this server my PhD Journey would have been very difficult.

Date: 19-06-2018

Mathu Malar C,

Structural Biology and Bioinformatics Division,

CSIR- Indian Institute of Chemical Biology,

4, Raja S C Mullick Road,

Kolkata – 700032, India.

List of abbreviations

PK– PolyKetide

NRPS – Non Ribosomal Peptides

NGS – Next Generation Sequencing

SMRT – Single Molecule Real Time Sequencing

SNP – Single Nucleotide Polymorphism

TR – Transposons

HR – Highly Repetitive

PE- Paired End

MP – Mate Pair

OLC – Overlap Layout Consensus

DBG – De Bruijn Graph

CNV – Copy Number Variants

SOD – Sudden Oak Death

LD – Linkage Disequilibrium

P-ctg – Primary Contig

A-ctg – Alternate Contig

PHI – Pathogen Host Interaction

CAZy- Carbohydrate Active Enzymes

HMM – Hidden Markov Model

COG – Clusters of Orthologous Groups

BWA – Burrows-Wheeler Algorithm

CRN – Crinkler

BLAST – Basic Local Alignment Search Tool

CDS – Coding Sequences

dN – Non Synonymous Substitution

DNA – Deoxyribo Nucleic Acid

PGAP – Prokaryotic Genome Analysis Tool

dS – Synonymous Substitution

tRNA – Transfer Ribo Nucleic Acid

PCR – Polymerase Chain Reaction

mRNA – Messenger Ribo Nucleic Acid

dNTP – Deoxyribo Nucleotide Tri Phosphate

WGS - Whole Genome Sequencing

NCBI – National Center for Biotechnology Information

Contents

List of tables

List of figures

सी.एस.आई.आर - भारतीय रासायनिक जीवविज्ञान संस्थान
CSIR - INDIAN INSTITUTE OF CHEMICAL BIOLOGY
(Govt. of India)
4, राजा एस. सी. मल्लिक रोड, यादवपुर, कोलकाता - 700 032
4, Raja S. C. Mullick Road, Jadavpur, Kolkata-700 032

Dr. Sucheta Tripathy, Ph.D.,
Ramalingaswamy Fellow (2011-2016)
Principal Scientist
Structutal Biology and Bioinfarmatics Division

CERTIFICATE

This is to certify that the work incorporated in this Ph.D. thesis entitled

"**Perils of Genome Assembly: Data types and sequencing platform defines optimal genome assembly in prokaryotes and eukaryotes**" submitted by **Ms. Mathu Malar C** to Academy of Scientific and Innovative Research (AcSIR) in fulfilment of the requirements for the award of **the Degree of Doctor of Philosophy (Ph.D.),** embodies original research work under my supervision. I further certify that this work has not been submitted to any other University or Institution in part or full for the award of any degree or diploma. Research material obtained from other sources has been duly acknowledged in the thesis. Any text, illustration, table etc., used in the thesis from other sources, have been duly cited and acknowledged.

18/6/2018

Mathu Malar C
(Student)

18/6/2018

Dr. Sucheta Tripathy
(Supervisor)

दुरभाष / Phone : 91-33-2499-5894 (Off.), फैक्स / Fax : +91-33-2473-5197
E-Mail : tsucheta@gmail.com ● Website : http://www.iicb.res.in ● https://twitter.com/tsucheta

Preface

Genome sequencing plays an important role in understanding genetics of microbes and developing personalised medicines etc. Rapid advancements in sequencing technology and advancement of bioinformatics tools made understanding genomes more feasible. Genome sequencing is more efficient and affordable now than it was a decade ago. The whole genome shotgun sequencing started with Sanger sequencing (1975). Overall genome sequencing can be classified into three different classes e.g.; 1st generation, 2nd generation and 3rd generation sequencing.

First generation sequencing technology was based on Sanger sequencing technology and Maxam Gilbert's chemical cleavage method. In Sanger sequencing method, dideoxynucleotides were labelled with flurochromes followed by electrophoresis. Picking up colonies and fluorescence detection were then automated using the robotic technologies. This sequencing technology produced reads that were on an average about one kilobase in length. While Maxam Gilbert method involves denaturing of DNA into single stranded chains and each fragment is labelled on 5' end. Here DNA is cleaved at specific points, using different combinations of chemicals to obtain more fragments. Then the reactions are loaded into polyacrylamide gel for differentiating the fragment sizes. The fragments are then visualized by radioactive tag. The bases are called by the banding pattern.

Second generation sequencing technology involves generation of millions of short reads in parallel. Speed of sequencing is very fast when compared with the first generation sequencing methods. Cost of sequencing has dropped significantly using this method as compared to first generation sequencing technologies. Here the read lengths are relatively shorter than the first generation sequencing methods e.g.; from 100 base pairs to 250 base pairs, where as the depth of sequencing is very high

Third generation sequencing technology brought about major revolution in sequencing technology. Here the whole DNA is used for sequencing. Using this method reads of length between 2000 to 50000 bases can be generated seamlessly. Long read sequencing technology therefore provides opportunity for chromosome level resolution at a very affordable price.

Although majority of sequencing is done via short read sequencing method (second generation sequencing technology), this is still marred with base-calling errors (Paszkiewicz

& Studholme, 2010). Due to the short read length, unique overlaps between pairs of reads are much less and the repetitive regions in the genome are sometimes longer than the reads themselves. This causes the real difficulty in maintaining the contiguity, since such regions tend to collapse using many genome assembly methods.

Based on the sequencing technologies, new genome assemblers are developed almost at a very rapid pace. Previously, using illumina short reads draft genome assemblies were generated. Now we have a choice to combine long and short reads to generate a better assembly. An emerging trend is to combine both short and long reads to improve the assembly contiguity reaching up to chromosome scale (Phillippy, 2017).

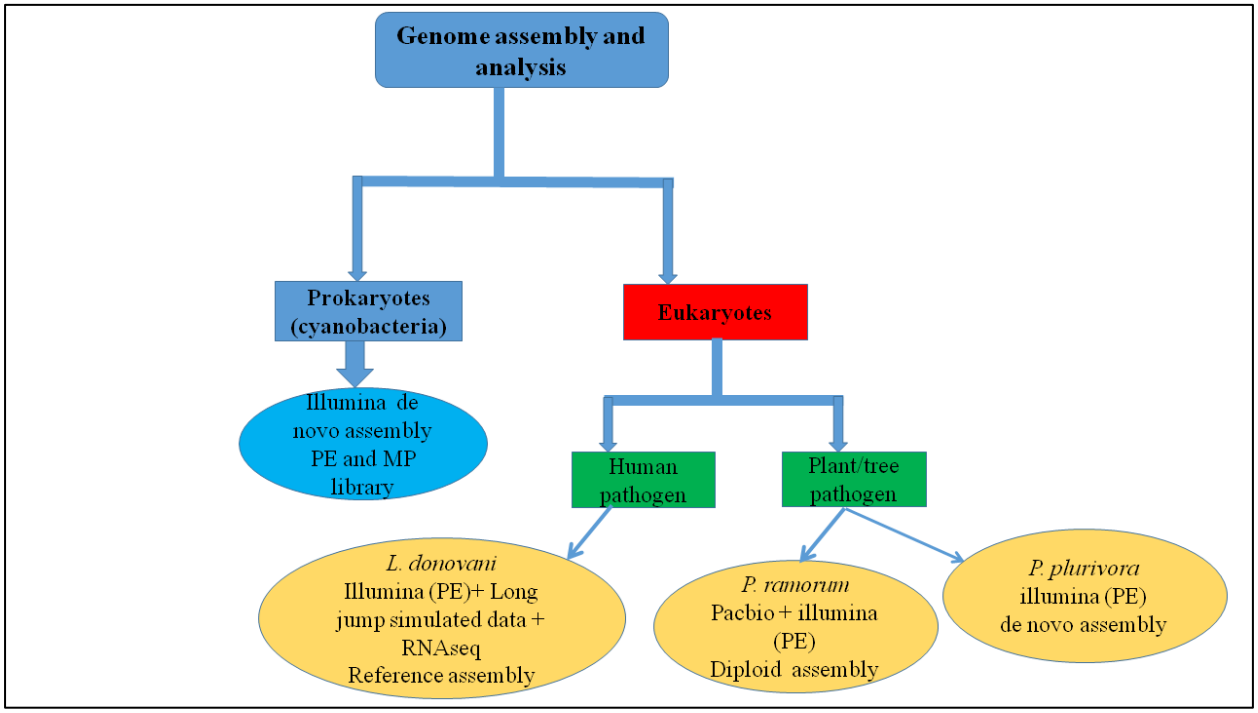
Lately, the ultimate goal of genome assembly is a gapless, haplotype – resolved assembly. Combining illumina and PacBio data with the help of new phasing methods can produce haplotype blocks seamlessly (Edge et al., 2017). However, there are still several hurdles in assembling a genome and it is a critical problem. Read length, library making protocols, read base overlaps between paired end samples, existence of long jump libraries are big factors in determining the assembly protocol. Several assembly methods exist which demands a huge computational space and may not be accessible to smaller labs. At the same time, there are assemblers that are processor driven and take very long linear time to complete a job. It is therefore very essential for a lab to choose the assembly protocol most suitable for their sequencing data type and in most efficient way. This problem is addressed in the present thesis and generated sequences from different library protocols, different organisms, variable GC contents, and read lengths were used to produce an optimal assembly. Here we have handled simple paired end and mate pair libraries of prokaryotes; paired and mate pair data for eukaryotic pathogens; long and short reads from different sequencing technologies in producing haplotype fused mosaic and haplotype phased diploid heterozygous assemblies (Figure 1).

Scope of the study

Genome assembly and genome analysis are great challenges especially dealing with complex genomes of higher organisms and contaminated sequences. We have used different genome assembly methods for optimizing genome assemblies of various photosynthetic prokaryotes, eukaryotic plant and human pathogens.

Genome analysis workflow used in the thesis is presented below.

Figure 1: Genome assembly and genome analysis workflow adopted in this work



We have implemented and tested most of the contemporary assemblers such as MIRA, ALLPATHS, Abyss, CAP3, Ray, Velvet, SOAPdenovo2, Edena, and A5 for assembling the genomes under study.

For Cyanobacterial samples, Illumina mate-pair, paired-end data was used. We first used the above mentioned methods and later refined our methods by using bioinformatically simulated sheared libraries for optimizing assembly output. The two Cyanobacteria species we sequenced were, *Tolypothrix bouteillei*, a fresh water organism and *Lyngbya confervoides*, a marine water Cyanobacteria using illumina sequencing platform. We used many combinations of assemblers to benchmark genome assembly. ALLPATHS assembler produced optimal final assembly which combined a second step assembly including a bioinformatically generated long jump library from the first round assembly.

In our eukaryotic genome assembly, human and plant pathogens were assembled. For plant pathogens such as *Phytophthora ramorum*, there were two isolates sequenced e.g.; Pr102, a pathogen for Oak plants and ND886, a pathogen of ornamental Camellia plants. Both the isolates had Illumina short reads and Pacbio long reads from two different chemistries (P5-C3 for Pr012 and P6-C4 for ND886).In case of P5-C3 chemistry, (used for *P. ramorum* Pr102),

posed serious challenges in assembly due to read errors. Pacbio reads are more erroneous with indel read errors; it is challenging task to identify the polymorphisms by eliminating sequencing artifacts. . For *P. ramorum* ND886, we had Pacbio P6-C4 chemistry data, which is an updated chemistry over P5-C3 and illumina reads. A diploid aware assembler, FALCON was exclusively used to produce the draft contigs. Genotypes were generated after illumina reads were mapped to the primary contigs. Using the Illumina mapped Genotypes and Pacbio raw reads mapped data to the primary contigs, haplotype phased assemblies were generated. Short reads played an important role in correcting indels from the assembled genomes that included Pacbio long reads.

Another tree pathogen *P. plurivora* was sequenced using illumina sequencing platform and genome was assembled using spades assembler to produce draft contigs. After assembling the genome, ploidy level and predicted genes were studied and compared with other related *Phytophthora species.*

In case of human pathogen *Leishmania donovani,* a reference based assembly was performed to get a full chromosome level assembly. Reference genome from NCBI was used for this purpose to accomplish the goal. For this work, we developed an assembler with minimal dependencies on third party software that includes Perl, bash, mummer and BEDtools. Our assembler utilises the reference genome and aligns with the draft contigs, followed by filtering on the basis of length of the aligned regions to remove false positives. Once the contigs are ordered, uncovered regions are padded with gaps and marked as Ns. This assembler works very well when chromosome level assembly is available as a reference.

Aims and objectives

1. To understand the roles of various assembly programs for generating optimal assembly for the chosen haploid Cyanobacterial species.

2. Genome assembly and analysis of complex plant pathogens.

i) To understand the nuances of hybrid assembly involving short and long reads from two different chemistries involving tree pathogens such as *Phytophthora ramorum* isolates for generating gapless diploid heterozygous assembly. Using this information to optimize genome assembly of another tree pathogen *Phytophthora plurivora*.

ii) Optimization of effector prediction methods from the final assemblies and understand the two speed genome concept from the hybrid assemblies.

3. Genome assembly and analysis of human pathogen.

i) To generate chromosome level assembly for human pathogen *Leishmania donovani* using reference based assembly with in-house assembler.

ii) Comparison of the late and early passages of the assembled genomes of *L. donovani* in detecting the fine chromosomal changes that determines pathogenicity.

Abstract

Genome sequencing not only facilitates understanding the biological activities of an organism but also provides a very useful resource for several downstream experiments. Sequencing technology has progressed enormously with the advent of third generation technologies where very long reads are produced compared to second generation sequencing. Long reads are used in larger eukaryotic genomes to close gaps, study complex regions containing repeats, and to generate more complete assemblies. The goal of this doctoral work was to study different genome assembly methods in depth and apply them in generating high quality genome assemblies for prokaryotes and complex eukaryotes. We investigated various strategies for genome assembly for reads generated from various platforms with different library generating strategies.

Chapter 1 contains the general introduction of genome sequencing, genome assembly, assembly algorithms, error correction methods and tools available for correcting long read sequencing data.

Chapter 2 contains optimized *de novo* assembly strategy of cyanobacterial genomes from short reads generated by Illumina sequencing platform. Assembling of cyanobacteria genome using different assemblers, revealed Allpaths-LG worked well on paired-end and mate-pair data. However, Allpaths-LG only worked well with tailor made libraries where both paired-end and mate-pair data is available with a given insert size. For other data types, this assembler did not produce optimal results. Genome mining on these organisms identified several polyketide and NRPs clusters in the assembled genome. When other methods such as Velvet, Mira, Ray, SOAPdenovo2, CAP3, Edena, A5 and Abyss were used, the assembly was largely fragmented and hence are not recommended. Another conclusion drawn from this work is, when assembly is not good enough with the first round of assembly, a second round can be done where the first assembly can be sheared bioinformatically to generate long jump libraries of longer insert size e.g.; 6Kb to 20Kb. In our study, the long jump libraries have improved the assembly statistics substantially for haploid prokaryotic organisms.

In chapter 3 we narrate the assembly methods optimized for oomycetes plant pathogens belonging to Stramenopiles group. In this study, we have included two plant pathogens, *Phytophthora ramorum* and *P. plurivora*. *Phytophthora ramorum* affects the Oak trees (Pr102 isolate) and ornamental flowering plant, *Camellia* (ND886 isolate). We have

generated hybrid genome assembly for Pr102 isolate and diploid genome assembly for the ND886 isolate for the first time. *P. ramorum* was difficult to assemble, since genome is highly heterozygous which also happens to be trisomic in nature. The assembled genome was used for downstream data analysis such as predicting effector polymorphism and curation of unique classes of repeat elements etc. On the other hand, *Phytophthora plurivora* was isolated from European beech (*Fagus sylvatica*) in Malmö, Sweden and only Illumina reads were generated for this species. Several closely related genomes were compared to understand the selection pressure on the RXLR effectors.

Overall, assembly was a challenge for eukaryotic plant pathogens because these species are diploid, heterozygous and have complex repeat rich regions making a bi-partite genomic architecture. Another major issue is the indel errors generated due to earlier sequencing chemistry which led to truncated gene models and pseudogenes, changing the total number of predicted genes. A polishing step using Pilon was introduced here, but it is also necessary to take care that the polishing is not undoing the polymorphisms captured by deep sequencing. As an outcome, we successfully developed pipeline for error correction of P5-C3 Pacbio sequencing chemistry and our optimal draft genome assembly was complete without gaps for Pr102 isolate. For ND886 isolate, haplotype phased diploid genome assembly was generated. This facilitated identification of more tandem repeat elements, paralogs of RXLRs which were not identified in other oomycetes earlier. Genome architecture of these organisms identified "two-speed genome architecture". Genome analysis on *P. plurivora* identified to be a polyploid genome and predicted RXLR genes were detected with ancient duplication events in comparison with other closest *Phytopthora* genomes.

Chapter 4 contains the exclusive study of the human pathogen *Leishmania donovani* isolated from India. The genomes of early passage (immediately taken from the host and grown in the lab; virulent in nature) and late passages (where the culture has been growing in the lab *in vitro* conditions till 25^{th} passages; lost virulence) were sequenced using illumina platform. We developed an in-house reference based assembler, which was used to further improve the assembly up-to chromosome level. Assembled early and late passage genomes revealed there were subtle changes in the genomes during the passages. We detected Copy number variation (CNV) events in both the passages. These CNV events play an important role in maintaining virulence in this organism and helps adapting them to a susceptible environment.

Chapter 5 contains conclusion and future scope of the dissertation

Chapter1: Introduction

1.1. Genome sequencing

The main aim of genome sequencing is to decode the complete DNA of the organism. The sequenced nucleic acids in the polynucleotide chains contain information of heredity and the biochemical properties of terrestrial life (Heather & Chain, 2016). The first sequencing technology was developed in 1977 by Sanger et al. (Sanger et al., 1977) who was awarded Nobel Prize in chemistry in 1980 along with Walter Gilbert (Maxam & Gilbert, 1977) for developing novel nucleotide sequencing methods. Their discovery paved path to study the genetic code of living beings and brought the sequencing technology to the doorsteps of the researcher's. Sequencing has revolutionized science in a greater way now having a newer field called as personalized medicine. Sanger sequencing is one of the widely used techniques (Pareek et al., 2011). The first Human genome project was carried out using Sanger sequencing technology and took almost 15 years to complete (Guzvic, 2013).

1.1.1. First Generation sequencing technology

Initial genome sequencing involved organisms having RNA genomes such as genomes of single-stranded RNA bacteriophages (Heather & Chain, 2016). In 1965, Robert Holley and colleagues were able to produce the first whole nucleic acid sequence of alanine tRNA from *Saccharomyces cerevisiae* (Holley et al., 1965). In the meantime Fred Sanger and colleagues developed a technique based on the detection of radio-labelled partial-digestion fragments after two-dimensional fractionation (Sanger et al., 1965). It is a 2-D fractionation method that Walter Fiers' laboratory produced for the first time to complete protein-coding gene sequences in 1972 of coat protein of bacteriophage MS2 (Fiers et al., 1972). In the mid-1970's Alan Coulson and Sanger's plus minus system (in 1975) and Allan Maxam and Walter Gilbert's chemical cleavage technique was introduced (Sanger & Coulson, 1975). The plus minus sequencing technique involved adding polymerases to synthesize the template using a primer with the radio-labelled nucleotide molecules. By using this technology first bacteriophage of ϕX174 was sequenced.

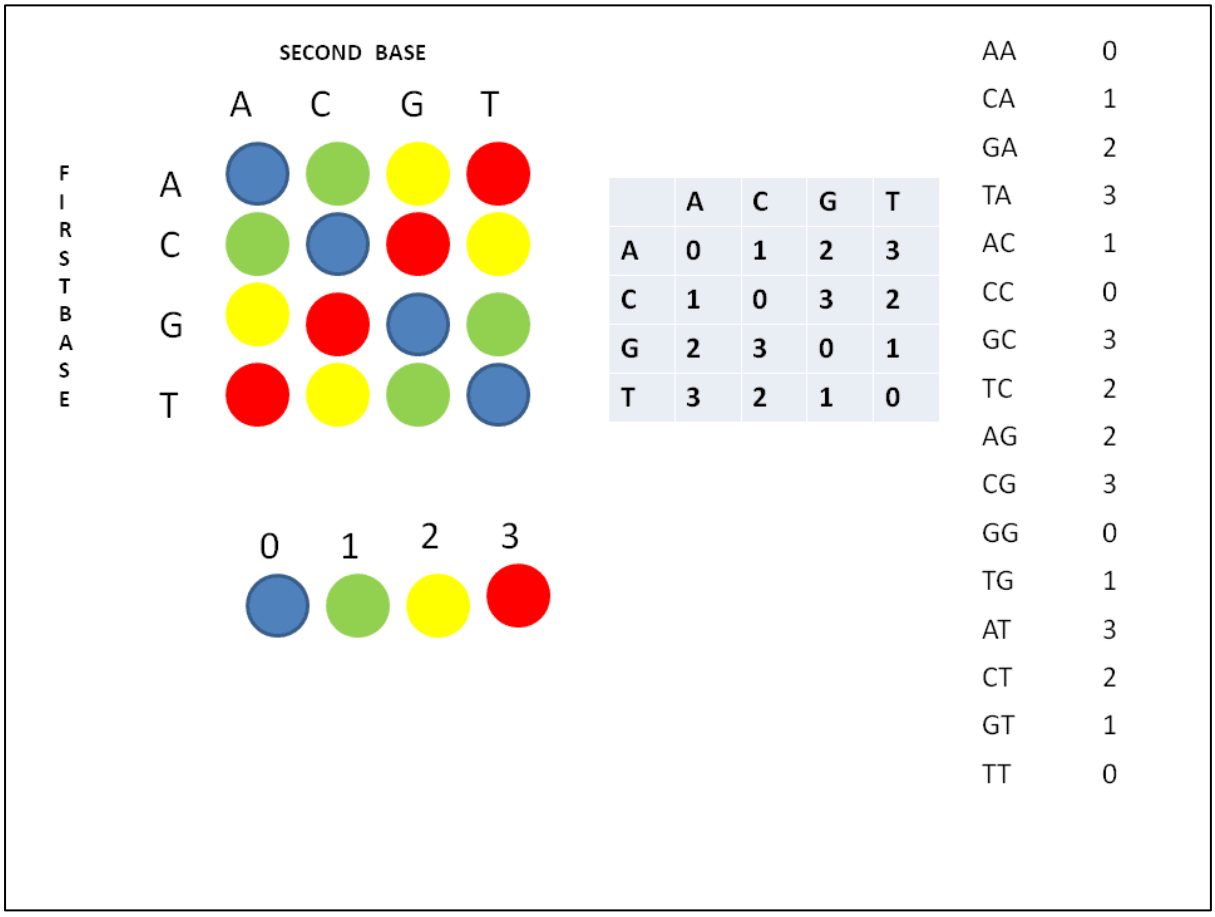
On the other hand, Maxam Gilbert method used to cleave the DNA chain at specific bases that were radio labelled. With this, began the birth of 'First generation sequencing technology'.

More improvements were done on First generation sequencing technology by replacing the phospho - or tritrium-radiolabelling with the flurometric detection and allowing the sequencing reaction in one large vessel instead of four. Then these developments in sequencing led to the development of automated DNA sequencing technologies.

1.1.2. 2nd Generation sequencing technology

This is another remarkable improvement over the first generation sequencing technology. This method doesn't include any radio- or fluorescently-labelled dNTPs or oligonucleotides for visualising under a gel. Here the sequencing is done by using a recently discovered luminescent method for measuring pyrophosphate synthesis (Heather & Chain, 2016). Later improvements were made to this method by attaching DNA to paramagnetic beads and reducing the washing time. There was a major pitfall to this method when the same number nucleotide is incorporated multiple times. The intensity of the fluorescence can't be correctly ascertained to exact number of nucleotides, and this error is known as homopolymer error. This technique is called Pyrosequencing and it was licensed by 454 Life technologies. It became the first successful second generation sequencing technology. For the parallelization process the DNA molecules are first attached to the beads via adapter sequences. The dNTPs were washed over and the release of pyrophosphate molecule was measured using charge. This method produced 400-500 bp long DNA sequences. The first high throughput machine was available in the market for the sequencing which was superior than 454 and available to the users as 454 GS FLX. This method offered more number of wells, more volume of data with better quality. Then the number of sequencing methods increased with solexa, illumina that largely depends on PCR based approach. For improving the sequencing accuracy, paired-end sequencing was developed. SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing method on the other hand was based on ligation of dinucleotides and was introduced by the Applied Biosystems. There were 16 such dinucleotides and four colors were assigned to each dinucleotide beginning with a unique nucleotide (Figure 1.1). Due to its 2 base encoding system, it ensures greater accuracy e.g.; 99.94%. This method used a DNA ligase developed by George Church's group (Shendure et al., 2005). While the SOLiD platform was able to produce high quality reads, data handling and assembly was complicated (Buermans & Den Dunnen, 2014).

Figure 1.1: Colorspace method of DNA sequencing (Source Genome-Array BlogSpot)

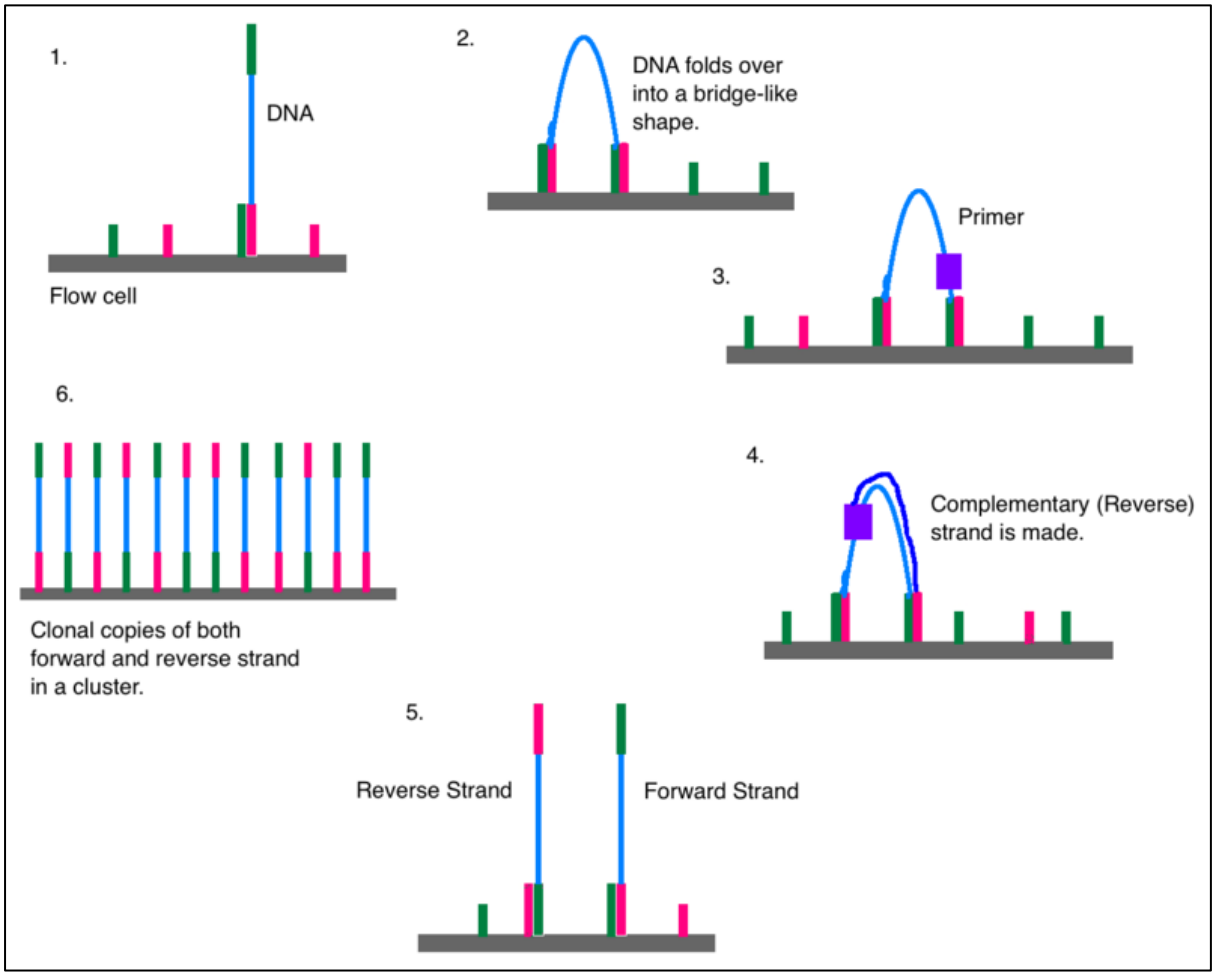


	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

The next improvement in the second generation sequencing technology was developed by Ion torrent called as "Post light sequencing". It does not use any fluorescence or luminance to detect changes in the sequencing bases. It is based on the pH changes occurring (H + ions) during polymerisation and the microprocessor chip is used to detect the signals.

The most widely used is Illumina sequencing or short read sequencing technology. In this technology DNA is sheared into fragments and adapters are ligated to the DNA fragments. Primers are attached to the complementary strands. When sequencing each cluster of DNA molecule emit the signal that is detected. Unlabelled nucleotide base and DNA polymerase are added to lengthen and join the strands of DNA attached to the flowcell. Bridges are created between the double stranded DNA and primers on the flowcell. The double stranded DNA is denatured into fragments of single stranded DNA using heat. Primers and fluorescent labelled terminators are attached to the nucleotide bases. Primers attached to the DNA are sequenced. DNA polymerase then binds to the primer and adds the first fluorescent labelled terminator to the new DNA strand. Laser light are passed through the flowcell to activate fluorescent label on the nucleotide base. The fluorescence is detected by camera and recorded on a computer and each bases with a different color. The process of illumina sequencing is depicted in Figure 1.2.

Figure 1.2: Illumina sequencing process (source: Google)

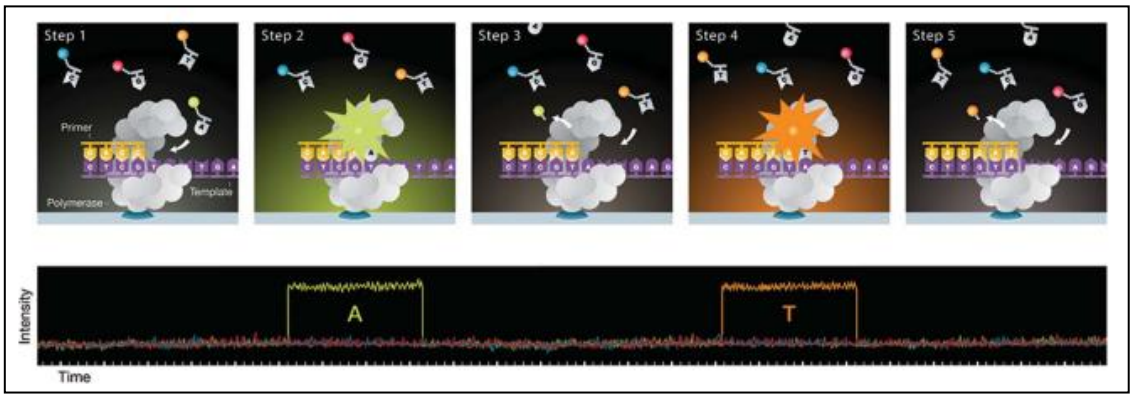


1.1.3. 3rd Generation sequencing technology

Single Molecule or real time sequencing was first developed in the lab of Stephen Quake (Braslavsky et al., 2003) and it was commercialized by Helicos Biosciences, without the PCR amplification process.. The most used third generation sequencing refers to Pacbio and Nanopore sequencing technology. During Single Molecule, Real-Time (SMRT) sequencing, the nucleotide bases are labelled into different colours. Each nucleotide contains the base specific fluorescent label with the phosphate group, which is released when incorporating the polymerase. This can be detected by real time imaging during strand synthesis (Bleidorn, 2016). This whole process of sequencing takes place in the aluminium walls called zero mode wave guide. The single DNA polymerase molecule attaches with these aluminium walls, and the molecules can be monitored and sequenced. The average read length produced ranged from 10 kbp to 54 kbp. Using this technique approximately 2 to 4 nucleotides are sequenced per second. However, the estimated error rate with this sequencing technology is up to 20% (Hackl et al., 2014). The SMRT sequencing assemblers, diploid assemblers, SV calling methods are under active development, but it has not been optimized as yet. The sequencing

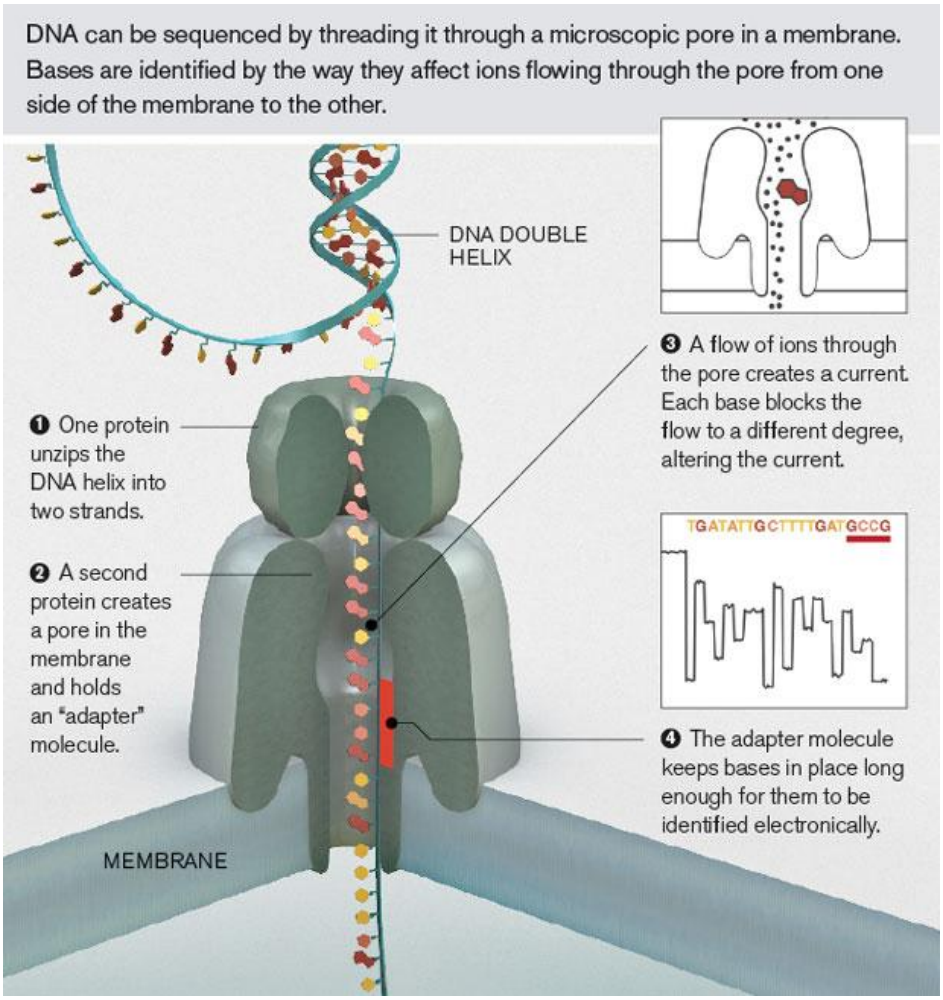
technology also is under active development for minimizing errors. SMRT sequencing process is showed in Figure 1.3.

Figure.1.3: Single molecule sequencing technology (Pacbio sequencing) adopted from *Bleidorn, 2016 et al.*Each of the nucleotides are represented with four different colors. Labelled nucleotide is associated with the template DNA in the polymerase active site, fluroscence intensity activity can be recorded. The emission of light will be detected by the ZMW sensor. Nucleotide incorporated DNA signal is recorded



With the MinION, Oxford Nanopore has developed a new technology to sequence nucleotides using Nanopores. The sequencing devices has the size of a small USB like device and connected with a desktop computer or a laptop. This sequencing is dependent on the biological pores made up of proteins. When these bilayered pores are kept in the salt solution, the electrodes act as anionic gradient. The negatively charged DNA molecule can be passed through the Nanopore. The MinION sequencing contains 512 channels containing the nanopores and detects approximately 10 bp per second (Bleidorn, 2016). The average read length ranges from 1 kb to 5 kb. Here the sequencing can be done in two ways e.g; a single strand sequencing or a double strand sequencing. The Sequencer machine is very small in size and occupies very little working space. Sequencing can be done anywhere and only for basecalling internet is required. Offline basecalling methods are now newly introduced to help the base calling when internet is not there. MinION sequencing process is represented in Figure 1.4. In this method, the signals for every possible 5- mers are recorded. . The raw data thus generated are base callled using the MinKnow cloud based basecallers or Poretools which are recently developed. The basecallers, the assemblers, structural variants (SV) callers are still in developmental stage.

Figure 1.4 : Representation of MinION sequencing



1.2.Error correction tools and algorithms available for long reads

Third generation sequencing technology provides opportunity for obtaining longer reads up to 50,000 bp with approximate error rate of 15% (Salmela et al., 2016). Although long read are successful and useful in producing better genome assembly, but the actual challenge occurs when reducing sequence indel error in the assembly. Several methods are available to correct the errors in the genome assembly. The error correcting methods are based on two main categories, they are self correction and short reads (illumina) based error correction.

Self correction method involves correction by aligning them against each other. PBCR (Berlin et al., 2015), LoRDEC (Salmela & Rivals, 2014) and LoRMA (Salmela et al., 2016) are the available tools for self correction.

Hybrid error correction requires high quality short reads. Illumina sequences are necessary for correcting errors in long reads. It is essential to sequence short reads for the genomes. Proovred

(Hackl et al., 2014), Jabba (Miclotte et al., 2015) along with PBCR and LoRDEC can be used for hybrid correction.

1.3. Scope of Genome sequencing in microbes

The next generation sequencing technology is largely used for understanding bacterial pathogenesis and pathogenic islands among the pathogens. Whole genome sequences are also important for understanding the host pathogen interaction. Genome sequencing of *S. pyogenes* has identified several virulence genes such as C5a peptidase, streptolysin O, and streptolysin S (Nakagawa et al., 2003).

Another important application of genome sequencing is to develop novel antibiotics. When *S. pneumoniae* genome was sequenced for the first time, it provided a great opportunity for discovering novel antibiotic targets (Mitchell, 2006).

As an important finding of genome sequencing include *Pseudomonas chlororaphis* PCL1606 soil born bacteria was identified with antifungal Bio control agent with Phytopathogenic fungi (Calderón et al., 2015).

1.4. Pre-processing of Genomics data

Sequence reads are first quality checked using tools such as FASTQC (Andrews, 2010), MultiQC (Ewels et al., 2016) etc. Poor quality sequences are with less standard quality scores (Usually less than 30 phred quality scores) are removed. Subsequently, sequences are checked for presence of adapters and if present, they are subsequently trimmed using software programs such as Trimmomatic (Bolger et al., 2014). The cleaned reads thus generated are used for genome assembly and analysis.

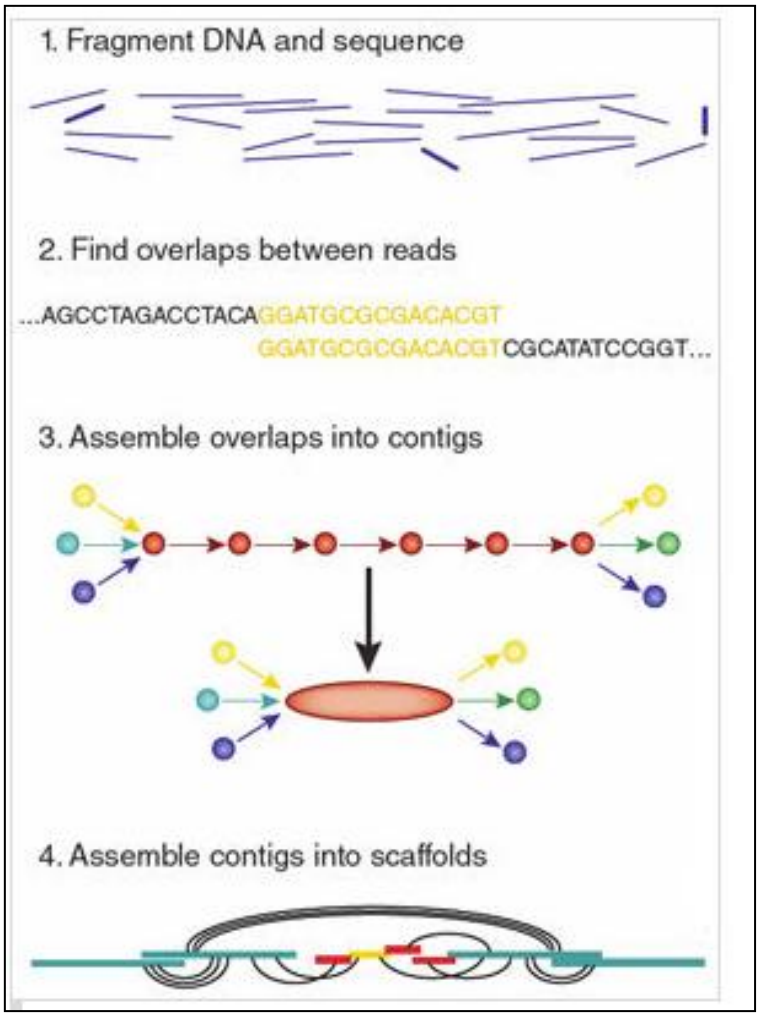
1.5.Genome assembly

Genome assembly is a challenging problem that requires time, computational resources and expertise in basic UNIX or other programming languages (Ekblom & Wolf, 2014). Choice of genome assembler is often dependent on the sequencing platform. The time for completion of genome assembly depends on the size, complexity of the genome and the processors being used. Single or the paired end reads can be used for assembly. Short reads and long reads can be merged producing hybrid assemblies. Assemblers in many cases fail to handle repetitive sequences, polymorphisms, missing data resulting in misassemblies.

Genome assembly is a process where sequenced reads are stitched together to make contiguous fragments. Before assembly, it is essential to remove poor quality sequences, duplicated reads, adapter sequences, vector sequences and sequences representing contamination. There are many bioinformatics tools that can be used for this purpose such as Trimmomatic (Bolger et al., 2014). Cleaned data is then ready to be assembled. Genome assembly is mostly a 2-stage process comprising of i) generation of contigs from reads and ii) joining of contigs into larger fragments called scaffolds often by the incorporation of Ns representing gaps (Hunt et al., 2014). Recent advances in the genome sequencing aids to generate chromosome level assembled genomes using the combination of

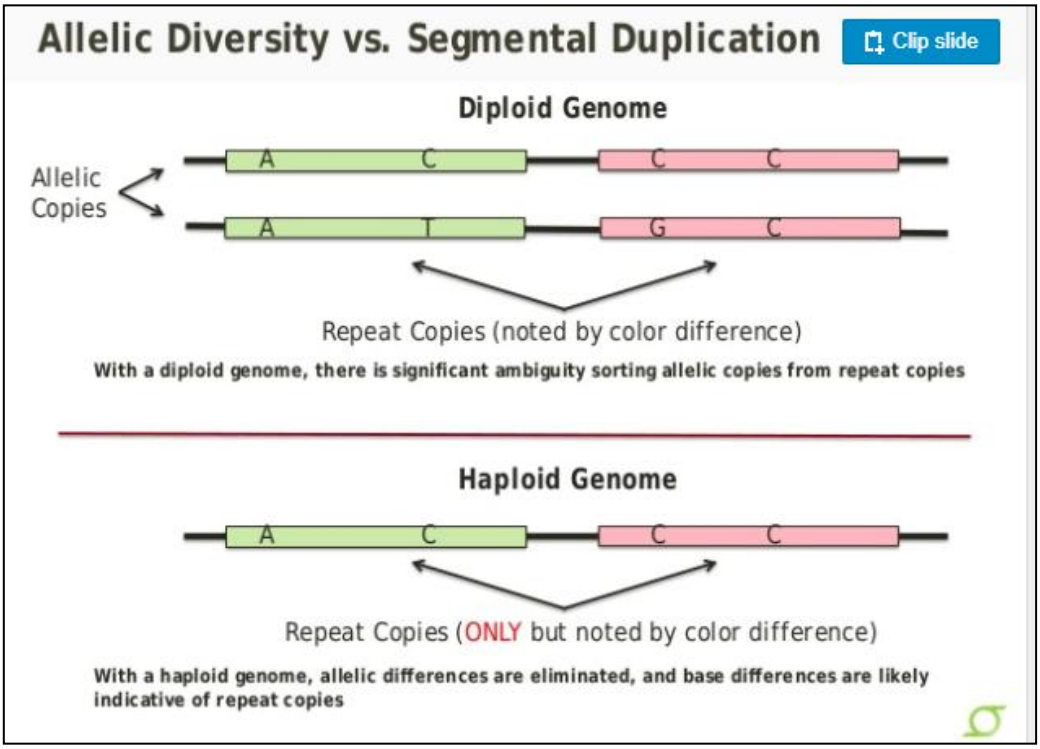
approaches such as BAC sequencing or Linked Read sequencing. Outline of genome assembly process is shown in Figure 1.5.

Figure.1.5: Process of genome assembly from *Baker et al 2012*



Software engineers or computational biologists who develop assemblers believe that for each sequencing technology there should be specific assembly tools. The error rates and the reliability of sequencing motivate developers for creating different types of genome assemblers using various algorithms. Genome assembly can be performed in two ways; a) reference-guided assembly which is done with the help of high quality reference genomes, and b) *de-novo* assembly which is done when no reference genome is available. Reference-guided assembly provides better assembly statistics. Comparison of haploid and diploid genome assembly process is shown in Figure 1.6.

Figure.1.6: The haploid and diploid genome assembly process from *ASHG GRC workshop 2015*



1.5.1. Genome Assembly Quality Assessment

The quality of genome assembly is assessed by certain metrics such as contig N50, number and size of gaps, length of the contigs and size of the genome assembly. There are various tools such as Quast (Gurevich et al., 2013) that helps to assess the assembly or compare with the closest reference genomes. Other methods include checking gene space that includes the number of core genes present in assembled genome using CEGMA (Parra et al., 2008) or BUSCO (Klioutchnikov et al., 2017). Sometimes misassemblies due to highly repetitive sequences are difficult to correct in the genome assembly. Initially most assemblers produced mosaic haploid assemblies, but presently with the help of long read sequencing technology diploid genome assembly is also possible where allele information is included.

1.6. An overview of algorithms and tools for genome assembly

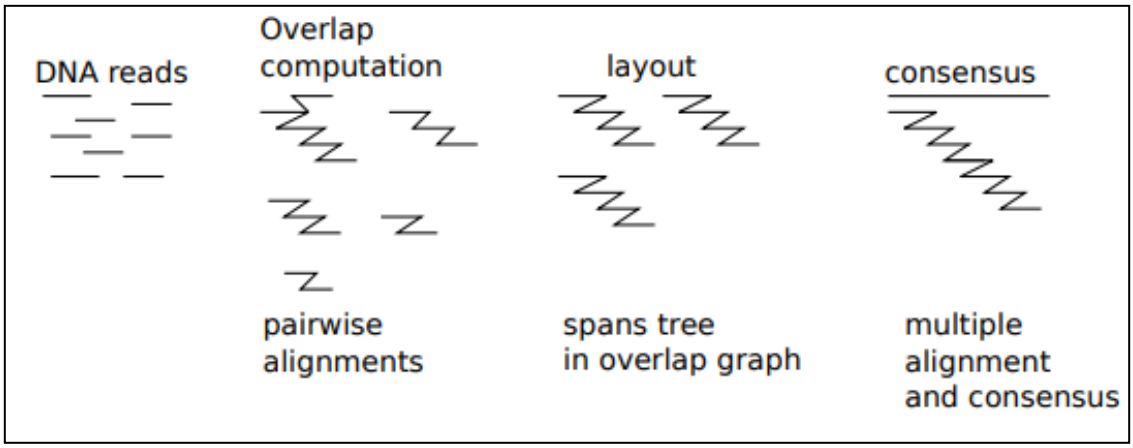
The genome assembly algorithms are mainly organised into 3 categories. **a.** Overlap layout consensus (OLC) which is based on the overlap graphs **b.** de Bruijn Graph (DBG) method based on the K-mer graph **c.** Greedy overlap method which uses DBG or OLC(Miller et al., 2010).

1.6.1. Overlap Layout Consensus algorithm

The Overlap Layout Consensus (OLC) graph is widely used in computer science. It contains a set of nodes and edges called as vertices and arcs. If the edges are transversed in one direction, it is called a directed graph. The graphs are conceptualised using balls and arrows connecting them. Each directed

edge represents the connection from one source node to sink node. Simple path contains only distinct node. This simple node does not intersect or does not connect with any other paths. The nodes and edges maybe assigned to the variety of attributes and semantics. The overlap graph represents the sequences and its corresponding overlaps. The graphs are computationally expensive since it performs pairwise alignments and each of the alignment is stored in the memory. The overlaps are represented as reads and edges. The graph might contain the different attributes to represent the forward and reverse reads. The reverse compliments are represented by the mirror images. Arachne, Celera, PHRAP, CAP, TIGR are the few assemblers based on the OLC graph algorithm (Figure 17).

Figure 1.7: Representation of OLC algorithm from *Google*



1.6.2. De Bruijn Graph algorithm

The de Bruijn graph was originally developed for representing the strings from a finite alphabet. The nodes represent all possible fixed-length strings. An Eulerian path in DBG (k-1) corresponds to a Hamiltonian path in DBG (k) (Figure 1.8 and Figure 1.9).

Figure1.8: Representation of graph containing nodes and edges. [Image source *Google]*

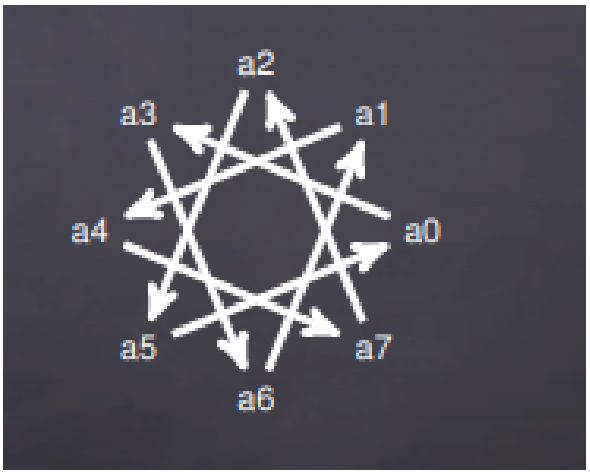
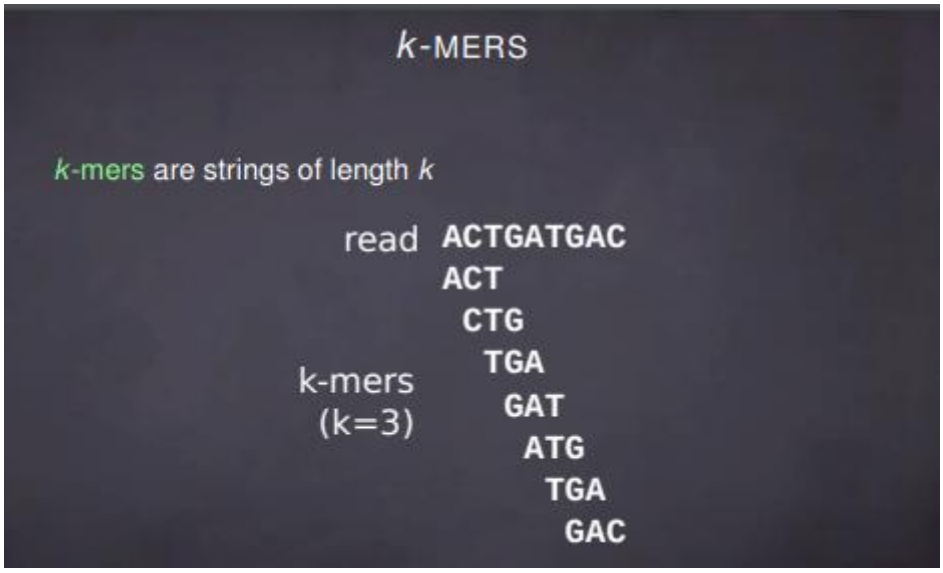


Figure 1.9: Representation of K-mer search process used in genome assembly process



1.6.3. Greedy algorithm

The greedy algorithm always considers higher scores for every overlap. It reduces the false positives by incorporating false positive overlaps into contigs. However, this process is more time consuming and is computationally intensive. Most of the assemblers use a graph based method with greedy algorithm. . For a given set of contigs or reads greedy approach searches iteratively until no new overlapping pairs are found. For each overlaps it uses the highest overlapping score to join the reads. The scores note the number of matches in the overlaps. The greedy algorithm helps to extend the smaller contigs into extended larger contigs. SSAKE is the first designed assembler on the basis of greedy algorithm.

1.6.4. Diploid genome assembly

The genome assemblers designed for haploid assembly generally produce mosaic assemblies contained many mis-assembled regions. Heterozygous genomes are assembled mostly as fragmented genomes. Large scaffolding and chromosome sequencing based methods such as BAC sequencing, linked read sequencing are quite expensive. With the advanced SMRT sequencing technology, the diploid aware genome assemblers such as FALCON (Chin et al., 2016) were developed to phase haplotypes and retain the information of both parents. This assembler initially performs the error correction method by using the inbuilt tool daligner which align and overlap with all raw reads FALCON-SENSE helps to preserve the heterozygous sites. After error correction step, FALCON identifies the overlap between all pairs of preassembled reads and read overlaps are constructed by the Myer's algorithm (Myers et al., 2014).The heterozygosity information is stored as bubbles and chains.

1.7. Analysis of Genomics data

Genome data comes in forms of raw reads in large numbers. This data first need to be re-constructed to form long contiguous fragments that can be further used for predicting genes, finding SNPs, finding structural variations in the genomes, genome co-linearity and so on. This process is called as genome assembly and is a very complicated step. Both prokaryotic as well as eukaryotic organisms have several complex regions in the genomes including repeats, transposons, palindromes, recombination's and translocations just to name a few. Most of the assemblers developed to this date have several issues handling these regions, thereby making assembly fragmented or chimeric. Homozygous sequences are better handled where there are no locus specific variations but heterozygous regions are dealt with great difficulty. The assemblers often can't differentiate a SNP with a heterozygotic sequence. In absence of a proper assembly, all the proposed functional studies can't be done.

My primary work in this thesis involves handling genomes with various degree of complexities, using various assembly techniques individually or in combination with others to optimize the assembly process.

References

Andrews S, 2010. FastQC: a quality control tool for high throughput sequence data.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM, 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology* **33**, 623.

Bleidorn C, 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and biodiversity* **14**, 1-8.

Bolger AM, Lohse M, Usadel B, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20.

Braslavsky I, Hebert B, Kartalov E, Quake SR, 2003. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences* **100**, 3960-4.

Buermans H, Den Dunnen J, 2014. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1842**, 1932-41.

Calderón CE, Ramos C, De Vicente A, Cazorla FM, 2015. Comparative genomic analysis of Pseudomonas chlororaphis PCL1606 reveals new insight into antifungal compounds involved in biocontrol. *Molecular Plant-Microbe Interactions* **28**, 249-60.

Chin C-S, Peluso P, Sedlazeck FJ, *et al.*, 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods* **13**, 1050.

Edge P, Bafna V, Bansal V, 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**, 801-12.

Ekblom R, Wolf JB, 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications* **7**, 1026-42.

Ewels P, Magnusson M, Lundin S, Käller M, 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-8.

Fiers W, Haegeman G, Iserentant D, Min Jou W, 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*.

Gurevich A, Saveliev V, Vyahhi N, Tesler G, 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-5.

Guzvic M, 2013. The History of DNA Sequencing. *Journal of Medical Biochemistry* **32**, 301-12.

Hackl T, Hedrich R, Schultz J, Förster F, 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-11.

Heather JM, Chain B, 2016. The sequence of sequencers: the history of sequencing DNA. *Genomics* **107**, 1-8.

Holley RW, Apgar J, Everett GA, *et al.*, 1965. Structure of a ribonucleic acid. *science*, 1462-5.

Hunt M, Newbold C, Berriman M, Otto TD, 2014. A comprehensive evaluation of assembly scaffolding tools. *Genome biology* **15**, R42.

Klioutchnikov G, Kriventseva EV, Zdobnov EM, 2017. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol*.

Maxam AM, Gilbert W, 1977. A new method for sequencing DNA. *Proceedings of the national academy of sciences* **74**, 560-4.

Miclotte G, Heydari M, Demeester P, Audenaert P, Fostier J. Jabba: Hybrid error correction for long sequencing reads using maximal exact matches. *Proceedings of the International Workshop on Algorithms in Bioinformatics, 2015*: Springer, 175-88.

Miller JR, Koren S, Sutton G, 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-27.

Mitchell TJ, 2006. Streptococcus pneumoniae: infection, inflammation and disease. In. *Hot Topics in Infection and Immunity in Children III.* Springer, 111-24.

Myers G, Brown D, Morgenstern B, 2014. Algorithms in bioinformatics.

Nakagawa I, Kurokawa K, Yamashita A, *et al.*, 2003. Genome sequence of an M3 strain of Streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res* **13**, 1042-55.

Pareek CS, Smoczynski R, Tretyn A, 2011. Sequencing technologies and genome sequencing. *Journal of applied genetics* **52**, 413-35.

Parra G, Bradnam K, Ning Z, Keane T, Korf I, 2008. Assessing the gene space in draft genomes. *Nucleic Acids Research* **37**, 289-97.

Paszkiewicz K, Studholme DJ, 2010. De novo assembly of short sequence reads. *Briefings in bioinformatics* **11**, 457-72.

Phillippy AM, 2017. New advances in sequence assembly. In.: Cold Spring Harbor Lab.

Salmela L, Rivals E, 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**, 3506-14.

Salmela L, Walve R, Rivals E, Ukkonen E, 2016. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* **33**, 799-806.

Sanger F, Brownlee G, Barrell B, 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of molecular biology* **13**, 373IN1-98IN4.

Sanger F, Coulson AR, 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**, 441IN19447-446IN20448.

Sanger F, Nicklen S, Coulson AR, 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences* **74**, 5463-7.

Shendure J, Porreca GJ, Reppas NB, *et al.*, 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *science* **309**, 1728-32.

Chapter 2: Genome Assembly of photosynthetic prokaryotes

2.1. Introduction

Cyanobacteria are one of the diverse and widely distributed bacteria phyla. These tiny organisms existed on earth atleast 2.7 billion years ago (Knoll, 2008). They are considered to be the progenitor of the chloroplast. Cyanobacteria play an important role in fixing the atmospheric nitrogen (Zehr et al., 2008). It also produces various secondary metabolic compounds (Welker & Von Döhren, 2006) which can be further exploited in the biotechnology industry and for other commercial purposes. Genome sequencing opened a new chapter in the cyanobacterial research (Mulkidjanian et al., 2006). Other than its role in biogeochemical cycle, cyanobacteria produce secondary metabolites, called as the natural products (Leao et al., 2017). The secondary metabolites from this organism can be commercially exploited further for treating cancer, inflammation, infections and many other diseases (Newman & Cragg, 2016). Earlier reports on filamentous marine cyanobacteria *Moorea* genus was reported to produce 40% of Natural products (Kleigrewe et al., 2016). The genomic analysis on this cyanobacteria revealed that this organism produce unique and novel secondary metabolites (Moss et al., 2016).

Till 2014, 72 draft genomes are registered in the NCBI database.. Cyanobacterial organisms are known to absorb light through phycobilisomes; it uses membrane bound light harvesting complexes. Earlier studies (La Roche et al., 1996) on cyanobacteria identified increase in transmembrane protein such as Pcb and IsiA which are important for light harvesting. Another study on cyanobacterial whole genome comparisons (Raymond et al., 2002) aimed to identify genes that play an essential role in phototrophy and to understand the advent and development of photosynthesis. Horizontal gene transfer studies on cyanobacteria found that interphylum metabolic gene transfers are more frequent in comparison to informational gene transfers (Zhaxybayeva et al., 2006). Genome mining on *Planktothrix serta* PCC 8927 identified acquisition of new gene cluster called hassalidin with genomic rearrangements. These hassalidins are shown to have antifungal activity (Pancrace et al., 2017).

For our analysis we performed the genome sequencing of beneficial cyanobacteria. We did illumina sequencing of marine cyanobacteria *Lyngbya confervoides* and fresh water cyanobacteria from the stone monuments, *Tolypothrix bouteillei* .Both the organisms are filamentous, heterocystous and were not studied before. *De novo* sequencing and assembly was done for these two species for the first time in our lab. The resulting assembly posed a great challenge since there were no closest reference available and there were no genome reports to compare. Also, there is a common belief that these Cyanobacteria may be harbouring other bacteria inside their cells making the genome reads difficult to assemble. We optimized assembly and charted out a way to remove unlikely sequences from the mixed dataset and performed downstream data analysis for elucidating the genomes.

2.1. Materials and Methods

2.1.1. Genome sequencing

Two libraries, one paired-end (300-bp insert size) and one mate-pair library (3-kb insert size), were constructed for these two organisms. Whole-genome sequencing was carried out using the Illumina HiSeq platform.

For *T. bouteillei,* it was sequenced at 90.0× coverage generating approximately 7.4 million reads with 151 bases read length. A mate-pair library was constructed with a 3 kb insert size and sequenced at 35× coverage generating 6 million reads and an average read length was 101 bp.

For marine water cyanobacteria *L. confervoides,* sequences were generated at 98× coverage (12.3 million reads). For the mate-pair library the insert size was approximately 3 kb and the coverage was 40× (5.6 million reads); with average read length 101 bp.

2.1.2. Sequence filtration

The raw reads of paired end and mate-pair libraries were quality checked, the poor quality contaminated sequences and adapter containing sequences were trimmed and cleaned. Thereafter sequences were used for genome assembly process for both the organisms. Adaptors were removed using tagdust (Lassmann et al., 2009) followed by removing of PCR duplicates and correcting reads using SGA assembler (Simpson & Durbin, 2012).

2.1.3. Genome assembly of *T. bouteillei*

First we tried optimizing genome assembly with various assemblers and protocols. Most of the assemblers failed to produce optimal genome assembly with good assembly metrics. Details of assemblers used for optimizing the genome assembly are given below.

2.2.Different assemblers and algorithms used for this study

2.2.1. Mira

MIRA (Chevreux et al., 1999) is a multi-pass DNA sequence data assembler which assembles reads into contigs. MIRA *de novo* genome assembler was developed for second generation sequencing data. Especially genomes with higher repeat elements can be handled by MIRA assembler. The disadvantage with MIRA is it requires huge memory to assemble a genome. It uses high confidence regions of several aligned read pairs to build a contig.

MIRA genome assembler works with following steps

1. Data pre-processing
2. Read scanning
3. Systematic match inspection
4. Building contigs
5. Pathfinder and contig interaction
6. Consensus approval methods
7. Read extension
8. Contig linking and editing

2.2.2. CAP3 assembler

CAP3 assembler (Huang & Madan, 1999) works on three phases.

In the first phase, 5′ and 3′ poor quality regions of each read are identified and removed. Overlaps between reads are computed false overlaps are identified and removed.

In the second phase, reads are joined to form contigs in decreasing order of overlap scores. Then, forward–reverse constraints are used to make corrections to contigs.

In the third phase, a multiple sequence alignment of reads is constructed and a consensus sequence along with a quality value for each base is computed for each contig. Base quality values are used in computation of overlaps and construction of multiple sequence alignments.

2.2.3. Ray assembler

Ray genome assembler is a plug-in based distributed and parallel compute engine that uses message passing interface for passing the messages. It works on the principle of coverage distribution for K-mers in the de-Bruijn graph and is used for inferring the average coverage depth for unique genomic regions.

2.2.4. A5 assembler

The A5 pipeline is automated pipeline that performs all steps to generate bacterial genome assemblies from illumina reads. This pipeline works on five steps.

Read cleaning:	Sequence adapters and low quality reads are removed or trimmed using Trimmomatic (Lohse et al., 2012).
Contig assembly:	Paired reads are used for contig assembly by using IDBA-UD algorithm (Peng et al., 2012).
Crude scaffolding:	This process is done using generated insert libraries for producing initial crude scaffolds.
Misassembly correction:	Misassembly correction is done on the basis of read pair overlap. The reads which are not overlapping with expected distance of overlap, the regions are corrected.
Final scaffolding:	Final procedure in which stringent parameters are used for bridging the contigs together to get final scaffolds.

2.2.5. Edena assembler

Edena assembler is based on the overlap layout framework. As a first step, the redundant short reads are removed from the sequenced short reads. Then the overlap graph is constructed for the redundant removed short reads. The graph is cleaned for removing any spurious edges by resolving bubbles. In the final step, the contigs with minimum size in the graphs are provided as assembled contigs.

2.2.6. Abyss assembler

Assembly by short read sequences (Abyss) works on the basis of k-mer. The k-mer values are generated for the sequenced reads. K-mer data is processed to remove any read errors and initial contigs are built. As a second step, matepair sequence reads are provided to extend the contigs by resolving ambiguities in contig overlap.

2.2.7. Velvet assembler

Velvet genome assembler is based on de Bruijn graphs. A de Bruijn graph with compact representation of k-mer values are provided for construction of contigs. By using this approach contigs are generated for the short reads.

2.2.8. SOAPdenovo2 assembler

SOAPdenovo2 is based on *de Bruijn* graph construction. While constructing the graphs it requires huge amount of memory. The reads are cut into k-mer values and those values are used for combining into large unique linear K-mer values and used for assembling the reads into contigs.

2.2.9. ALLPATHS assembler

ALLPATHS assembler was initially designed for assembling short reads. Recently, this assembler accepts long reads to assemble. ALLPATHS assembler mainly works on the following two concepts **a**. It finds all possible paths across a read pair i.e., all sequences from one read to other reads which are covered by other reads**. b**. Localizes all read pairs to isolate and assemble smaller regions of the genome and assemble those region independently.

By using all the described assemblers we tried optimizing genome assembly of *Tolypothrix bouteillei*.

Genome Annotation

Genes from optimized final assembly was predicted using NCBI PGAAP pipeline for cyanobacterial genomes. Gene space was done using the blast based searches with cyanobacterial COGs.

2.3.Results and discussion

2.3.1. Characteristics of *T. bouteillei*

Tolypothrix bouteillei belongs to the order Nostocales. They form blue or green patches when grown *in vitro*. Under microscope, it appears acylindrical and filamentous in shape. These organisms have pseudo branching. It contains thick cell wall gelatinous like structure called heterocysts and this helps in nitrogen fixation. This organism grown on aquatic plants and on rock bottoms of flowing waters. It survives in harsh environmental conditions. Microscopic visualization of this organism in provided in Figure 2.1.

Figure 2.1: Microscopic visualization of *T. bouteillei*, fresh water cyanobacteria

2.3.2. QC results for *T. bouteillei*

The paired end and mate pair reads were quality checked for removing the poor quality sequences. After filtration, a total of 7,542,626 paired end reads passed the quality check. While filtering the mate pair reads total 5,596,346 reads passed the quality filtration. Filtered reads were used for assembly process.

2.3.3. Allpaths produced the best assembly among others tested for *T. bouteillei*

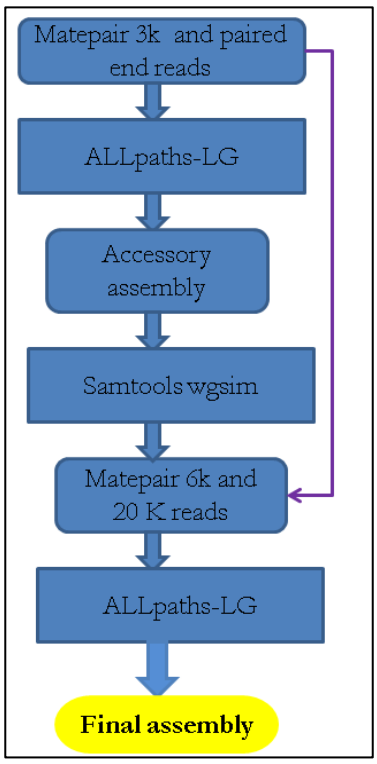
Various combinations of assembly methods were used in all possible combinations to benchmark the genome assembly. Comparison of genome assembly statistics are represented in Table 2.1.

Table 2.1: Genome assembly statistics of various assemblers used for *Tolypothrix bouteillei*

Sequencing Data Type	**Assembler**	**No of contigs**	**N50 (bp)**	**Largest contig (bp)**	**Smallest contig (bp)**	**Total length**	**Number of core genes out of total 384**
PE + MP	Mira	48518	1468	42333	51	45607177	228
“	Mira + cap3	40102	2245	70267	2245	40041359	228
“	Mira + SSPACE	2608	20195	395876	2001	21003902	228
“	Ray + SSPACE	1583	92796	511549	2000	16031125	134
“	A5	3730	23523	240416	2000	33283155	178
“	Edena	726	22345	91061	2000	10943288	151
“	Abyss	4494	3831	32963	2000	16934734	161
“	velvet	3241	5318	64819	2000	15481986	156
“	Soapdenovo2	864	22333	93080	2000	10928734	148
PE+ MP + simulated library	**ALLPATHS**	**70**	**1987087**	**2493347**	**2000**	**11572263**	**378**

Initially cleaned reads were assembled using A5 pipeline (Blin et al., 2013). The cleaned reads of *T. bouteillei* were assembled using all other methods failed to produce the final assembly. Contigs were initially bioinformatically sheared to produce long jump libraries of 6k and 20k libraries using wgsim (Li, 2011b). The sheared libraries and paired, matepair libraries were used to assemble using Allpaths-LG-49856. Flow chart representing genome assembly process is explained in Figure 2.2. Our final assembly resulted in 70 final contigs with total genome length of 11 Mb and containing 378 cyanobacterial core genes (out of total 384 core genes) we considered this assembly as a complete draft assembly.

Figure 2.2: Genome assembly pipeline optimized for *T. bouteillei*



From the above study, most of the assemblers produced fragmented assembly with large number of contigs with poor N50 values. N50 values of genome assembly cannot be compromised when assessing the genome assembly. As additional criteria to assess assembly quality, we checked the number of cyanobacterial core genes present in assembled genomes. We found that ALLPATHS generated assembly had the best statistics of higher N50, minimal number of contigs and high number of core genes in genome assembly. Our comparison revealed that this was more complete genome assembly in comparison to all other assemblies.

Genome assembly optimization is done using many different criteria. Memory consumption, CPU usage and processing time also considered for benchmarking. Consumption of memory was higher in case of MIRA assembler. MIRA not only consumed higher memory but also produced fragmented assembly. The processing time was also very lengthy reaching almost to a whole week for producing a genome assembly size of 12 Mb.

Velvet assembler also had similar issues when handling the genomes. In case of Ray assembler more reads were discarded and fragmented assembly was observed. Though assembler consumes less memory with less processing speed. Many core genes were missing in the assembled genomes.

By comparing all these criteria we have benchmarked ALLPATHS assembler for assembling cyanobacterial genomes.

After benchmarking the assembly of *T. bouteillei* we used the same approach to assemble another cyanobacterial genome *L. confervoides*.

With complete genome assemblies, genome mining was performed as a downstream process.

2.3.4. Genome mining of *T. bouteillei*

The final assembly of assembled genome was 11.5 Mb in length with 70 scaffolds and N50 value of 1,987,087 base pairs. The microscopy image of *T. bouteillei* is represented in Figure 2.2. Largest scaffold was 2,493,347 and smallest scaffold with the length of 2000 bp. Calculated GC content of the genome was 42%.

PGAP annotation (Tatusova et al., 2016) on genome predicted 7,777 protein coding genes, 35 CRISPR, 6 rRNA genes, 92 tRNA, 1,275 pseudogenes and 1 ncRNA. Here also we found more pseudogenes suggesting this genome might evolve faster.

Core ortholog analysis using OrthoMCL (Li et al., 2003) was performed on assembled genome. There are 384 core ortholog genes of cyanobacteria and we found that 378 was present in the assembly. This suggests that assembled genome is more complete.

Our blast searches of predicted genes with other known genomes found that 50% of the genes have no similarity with already annotated genomes. We found a dozen of alcohol dehydrogenase genes present in the genome. Other important genes such as VioC monooxygenase and CheY which codes for antibiotics and quorum sensing genes were found Table 2.2.

Table 2.2: Major genes identified in Genomes

proteins annotated	Copy number in *Tolypothrix bouteillei*	Copy number in *Lyngbya confervoides*
Polyketide	18	2
MFS transporter	51	95
ABC transporter	269	275
Tet R	27	13
Luciferase	2	NA
Cyanophycin	4	2
alcohol dehydrogenase	9	5
Cupin	19	12
chemotaxis protein	42	7
vioC monooxygenase	2	NA
Vioj	NA	2
spermidine synthase	4	3

2.3.5 Commercial potential of *T. bouteillei*

Tolypothrix bouteillei genome was not previously sequenced and thus our genome provides better insights. Studies on other Tolypothrix genomes (Hughes et al 2017) identified that these organism

produce a compound called Tolyporins. The active anticancer agent was identified from 62 tetrapyrrole macro cycle compound called tolyporphin A.

2.4.Genome analysis of *Lyngbya confervoides*

2.4.1. Characteristics of *L. confervoides*

L.confervoides is marine cyanobacteria grows in highly saline condition belongs to order Oscillatoriales. It forms long unbranching filaments inside a rigid mucilaginous sheath. These sheath or mats which intermix with other phytoplankton and grow into patchy thalluses.

2.4.2. QC results of *L. confervoides*

Our quality filtration on paired end reads resulted in total 12,343,980 reads passing the quality parameters and used for assembly process. In case of mate pair data, a total of 4,833,460 reads passed the quality check and used for assembly and downstream analysis.

2.4.3. Genome assembly of *L. confervoides*

Using our optimal assembler (Allpaths) we were able to assemble the genome. Genome assembly on filtered reads resulted in 298 scaffolds, with total assembly length of 8.7 Mb and N50 value of 5,207,129. The largest scaffold was 5,207,129 bp and the smallest scaffold was 3,620 bp in length. The GC content was about 55%.

Assembled length of a genome was comparable with other closest genome such as *Leptolyngbya* from NCBI assembly size is also around 5 to 9 Mb.

2.4.4. Genome mining of *L. confervoides*

Assembled 8.7 Mb genome was used for downstream analysis. PGAP annotation on assembled genome identified total 6093 protein coding genes, 1096 pseudogenes, 4 CRISPR, 70 tRNA and 2 ncRNA. Large numbers of pseudogenes were identified in the genome, stating that maybe this genome is evolving for adapting to various environmental conditions.

Pathway analysis on assembled genome identified a nonribosomal peptide synthetic pathway leading to the production of viomycin (Vioj), which has a property of antituberculosis activity Table 2.1. *Lyngbya* lives in marine environmental conditions. Salt tolerant genes such as spermidine synthase spermidine transporters are found in the genome, this gene might play an important role in growing organism in the salt conditions. Our comparison on this genes identified that it shares the closest similarity with *Leptolyngbya* sp. PCC 6406 of 83% identity.

2.4.5. Economic potential of *L. confervoides*

Earlier reports on *Lyngbya* species confirmed the presence of secondary metabolite biosynthesis, transport, and catabolism (Jones A C et al 2011). There are two major gene clusters which are characterized to be natural products curacin A and barbamide.

2.4.6. Present improvements in cyanobacterial genomics

Previously genomes of *Tolypothrix* and *Lyngbya* were not sequenced. Only 69 complete chromosome assemblies were available in public repositories. Our genomic studies identified various antibacterial and secondary metabolic compounds.

Present genomic reports from NCBI confirmed that total 120 (approx) complete chromosome level assemblies are found in public repositories. A drastic improvement in genome sequencing facilitates to sequence more cyanobacterial genomes. Genome sequencing helps in identifying major gene clusters, biosynthetic gene clusters and pathways.

2.5.Summary

In summary, this chapter throws light on importance of genome sequencing and genome assembly process for beneficial cyanobacteria whose complete genome sequences were not available. Here for assembling the genomes various combinations of genome assemblers were used to optimize the genome assembly. From our benchmarking process ALLPATHS assembler produced better draft assemblies of cyanobacteria species. ALLPATHS work very well for tailor made libraries and essentially need a paired end and a long mate pair library for scaffolding. An absence of a mate pair library other methods can be tried. For the given short reads the ALLPATHS assembler worked well for marine cyanobacteria as well with a higher N50 value. For better scaffolding, a different approach can be adopted that involves creating longer jump read libraries bioinformatically using wgsim. This process when followed the scaffolding involved longer assembly and N50 was also significantly higher. The downstream data analysis is a function of the qualitative genome assembly. This will ascertain better understanding of the biology of the organisms.

Publications resulted from this work

1. Chandrababunaidu MM, Singh D, Sen D, et al. Draft Genome Sequence of *Tolypothrix boutellei* Strain VB521301. Genome Announcements. 2015;3(1):e00001-15. doi:10.1128/genomeA.00001-15.
2. Chandrababunaidu MM, Sen D, Tripathy S. Draft Genome Sequence of Filamentous Marine Cyanobacterium *Lyngbya confervoides* Strain BDU141951. Genome Announcements. 2015;3(2):e00066-15. doi:10.1128/genomeA.00066-15.

References

Blin K, Medema MH, Kazempour D, *et al.*, 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research* **41**, W204-W12.

Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. *Proceedings of the German conference on bioinformatics*, *1999*: Hanover, Germany, 45-56.

Huang X, Madan A, 1999. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-77.

Kleigrewe K, Gerwick L, Sherman DH, Gerwick WH, 2016. Unique marine derived cyanobacterial biosynthetic genes for chemical diversity. *Natural product reports* **33**, 348-64.

Knoll AH, 2008. Cyanobacteria and earth history. *The Cyanobacteria: Molecular Biology, Genomics, and Evolution* **484**.

La Roche J, Van Der Staay G, Partensky F, *et al.*, 1996. Independent evolution of the prochlorophyte and green plant chlorophyll a/b light-harvesting proteins. *Proceedings of the National Academy of Sciences* **93**, 15244-8.

Lassmann T, Hayashizaki Y, Daub CO, 2009. TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* **25**, 2839-40.

Leao T, Castelão G, Korobeynikov A, *et al.*, 2017. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea. *Proceedings of the National Academy of Sciences* **114**, 3198-203.

Li H, 2011. wgsim-Read simulator for next generation sequencing. *Github Repository*.

Li L, Stoeckert CJ, Roos DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-89.

Hughes, Rebecca-Ayme, et al. "Genome sequence and composition of a tolyporphin-producing cyanobacterium—microbial community." Applied and environmental microbiology (2017): AEM-01068.

Jones, Adam C., et al. "Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium Lyngbya majuscula." Proceedings of the National Academy of Sciences 108.21 (2011): 8815-8820.

Lohse M, Bolger AM, Nagel A, *et al.*, 2012. R obi NA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic acids research* **40**, W622-W7.

Moss NA, Bertin MJ, Kleigrewe K, Leão TF, Gerwick L, Gerwick WH, 2016. Integrating mass spectrometry and genomics for cyanobacterial metabolite discovery. *Journal of industrial microbiology & biotechnology* **43**, 313-24.

Mulkidjanian AY, Koonin EV, Makarova KS, *et al.*, 2006. The cyanobacterial genome core and the origin of photosynthesis. *Proceedings of the National Academy of Sciences* **103**, 13126-31.

Newman DJ, Cragg GM, 2016. Natural products as sources of new drugs from 1981 to 2014. *Journal of natural products* **79**, 629-61.

Pancrace C, Jokela J, Sassoon N, *et al.*, 2017. Rearranged biosynthetic gene cluster and synthesis of hassallidin E in Planktothrix serta PCC 8927. *ACS chemical biology* **12**, 1796-804.

Peng Y, Leung HC, Yiu S-M, Chin FY, 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-8.

Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE, 2002. Whole-genome analysis of photosynthetic prokaryotes. *Science* **298**, 1616-20.

Simpson JT, Durbin R, 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**, 549-56.

Tatusova T, Dicuccio M, Badretdin A, *et al.*, 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic acids research* **44**, 6614-24.

Welker M, Von Döhren H, 2006. Cyanobacterial peptides—nature's own combinatorial biosynthesis. *FEMS microbiology reviews* **30**, 530-63.

Zehr JP, Bench SR, Carter BJ, *et al.*, 2008. Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *science* **322**, 1110-2.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT, 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* **16**, 1099-108.

Chapter 3: Genome assembly of Eukaryotic plant pathogen

a. **Hybrid genome assembly of *Phythophthora ramorum* Pr102 using Pacbio P5-c3 sequencing.**
b. **Haplotype phased diploid genome assembly of *Phytophthora ramorum* ND886 using Pacbio P6-C4 sequencing.**
c. **Genome assembly of *Phytophthora plurivora* using short read sequencing.**

A. Hybrid genome assembly of *Phythophthora ramorum* Pr102 using Pacbio P5-c3 sequencing.

3.1.1. Introduction

Phytophthora's are largest threat to the environment. Many genomic studies on Phytopthora's help us to understand their interaction with their host. The plant pathogen *P. ramorum* Pr102 was sequenced in 2006 using Sanger sequencing method with 65 Mb genome size. The great majority of studies on *Phytophthora* species have focussed on those that infect important annual crop plants, for example *P. infestans* that infects potato and tomato, and *P. sojae* that infects soybean(Haas et al., 2009). Oomycetes not only affects plants, it also affects the animals and human. In the last 20 years, seven different Phytophthora species (*P. cambivora*, *P. hibernalis*, *P. citrophthora, P. capsici, P. cactorum, P. drechsleri* and *P. infestans*) have been identified in affecting crops such as citrus, pepper, strawberry, melon, and potato respectively (Biçici & Çinar, 1990). Two of them, *P. citrophthora* and *P. capsici*, which attack citrus and pepper, are very destructive and have the greatest economic importance. *P. citrophthora* has caused approximately 15% fruit losses every year and 8–30% infection on a susceptible lemon cultivar in Cukurova region. *P. capsici* is very dangerous for pepper-spice and pepper-paste production because it causes up to 100% drying and killing of pepper plants under conditions of poor drainage and incorrect irrigation practices. *P. cambiuora*, which causes the ink disease of chestnut, has spread from the Black Sea coast to the Mediterranean; about 20 000 chestnut trees have been killed by this organism from 1952 to 1970.

The PacBio sequencing has become quite affordable, which can be used to close the gaps in a genome. However, the error rate of the reads are high, but the GC bias and the PCR bias was found to very less in the data. The genomic data can be largely used to understand the polymorphisms, evolutions and gene regulations of organism. Certain limitations however exist in correcting the errors in the reads and generating proper genome assembly. We have implemented the methods from already existing error correction tools in different phase that has provided the maximum number of corrected reads. This has improved the number of corrected reads from 33% to 47%. Generating the optimal genome assembly was a great challenge especially for heterozygous organism like oomycetes. Since the repeat content of the genome is high (more than 30%), it greatly interferes with the assembly process. When we sequenced this genome, the sequencing chemistry for longer reads were just introduced. We sequenced the genome with the available chemistry Pacbio P5-C3. There were

very limited numbers of tools to assemble and correct errors in those reads as well. For optimal genome assembly, five different approaches were used to benchmark assembly.

For *P. ramorum* however, the incompleteness of the Pr102 genome assembly posed several problems for comparing genomes of additional isolates for finer scale epidemiology and genetic studies. First, the 2006 assembly contained 2,576 scaffolds encompassing 12 Mb of gaps, compared to the ten to twelve chromosomes of *P. ramorum* (n = 10-12; D. Beattie and C.M. Brasier, personal communication). As a result of the gaps, the higher level architecture of the pathogen genome was obscured and potentially, many effectors and repetitive elements were lost, or their relative locations were unclear.

The 2006 genome assembly of Pr102 contained fragmented scaffolds as well. This fact impedes analyses involving allelic variation and linkages between polymorphisms. Therefore, to provide a foundation for more detailed studies of pathogen evolution, we used Pacific Biosciences long read sequencing technology to produce a more complete genome assembly with fewer gaps.

. Improvements in long read (multi-kilobase) sequencing technologies have enabled rapid and cost-effective assembly of large, complex genomes (Gnerre et al. 2011; Lam et al. 2011). Long read sequencing technologies include PCR-free methods to reduce PCR or GC bias (Rhoads & Au, 2015) and to conserve the fidelity of alleles when assembling the genome. Single molecule real-time (SMRT) sequencing technology (English et al., 2012) produces multi-kilobase reads that enable production of genome assemblies with longer contigs, that can effectively fill gaps, producing a less fragmented assembly (Vij et al., 2016).

3.1.2. Materials and Methods

3.1.2.1. Datasets used for benchmarking the genome assembly

PacBio long reads of *P. ramorum Pr102* were generated from the P5-C3 chemistry with the average read length of 4499 bp. The illumina short reads of transcript and illumina paired end libraries were generated for this study, to assist genome assembly process to reduce the error rate of P5-C3 chemistry. The intermediate assemblies were generated using ALLPATHS (Butler et al., 2008) assembler using illumina libraries. Short reads were used for correcting sequence based errors from the earlier genome assembly of *P. ramorum* Pr102 V1 (Tyler et al., 2006). Assembled Sanger unitigs from Tyler et al 2006 were higher quality contigs; and hence were appropriate for generating long jump matepair libraries bioinformatically. The detailed report on reads generated with sequencing coverage is shown in Table 3.1.

Table 3.1: Details of the number of reads generated for *P. ramorum* Pr102 isolate

Platform	No of raw reads	Read coverage	Isolate
Pacbio	435399	25X	*P. ramorum* Pr102
Illumina library	20942377	10X	*P. ramorum* Pr102
RNAseq Illumina	31781430	15X	*P. ramorum* Pr102
Simulated 10k mate-pair reads 6k mate-pair reads	10468 118360	6X	*P. ramorum* Pr102 V1 unitigs
Simulated 20k mate-pair reads	56758	5X	*P. ramorum* Pr102 V1 unitigs

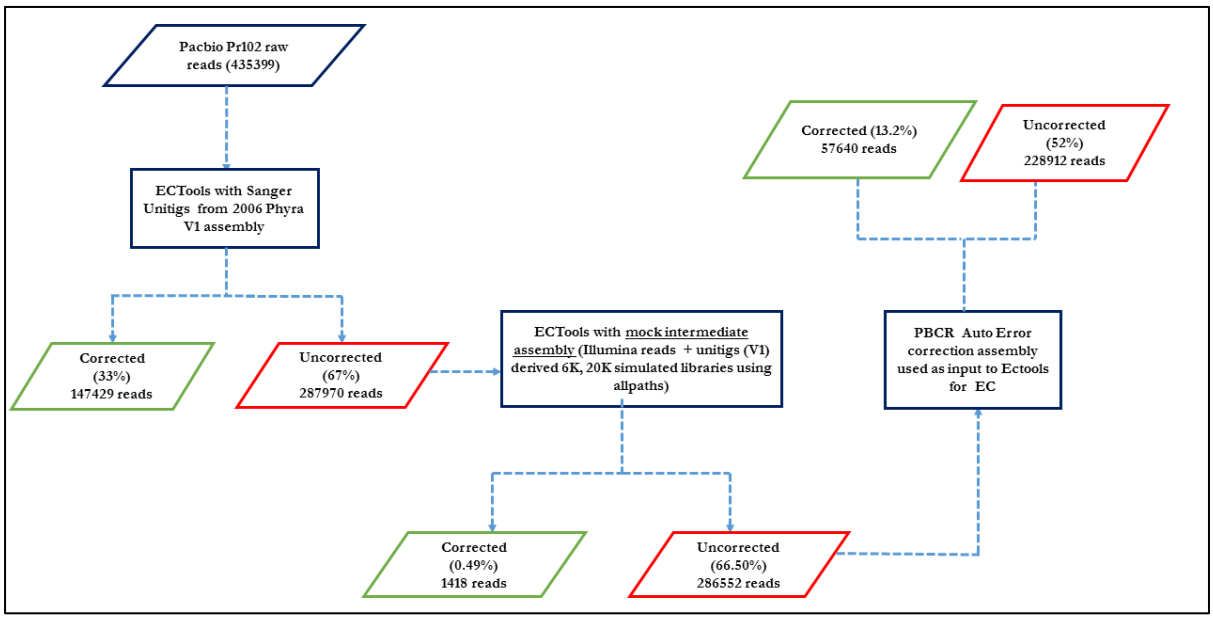
3.1.2.2. Improved 3-way error correction method

Phase 1: PacBio sequenced reads with sequencing base errors were initially corrected using ECTools [https://github.com/jgurtowski/ectools]. Sanger unitigs played an important role in correcting erroneous PacBio reads. However, using this method only 33% of reads were corrected.

Phase 2: Our second round of correction included utilizing illumina reads and Sanger contigs to produce the accessory assembly using ALLPATHS. Assembled contigs were simulated for 6k, 10k and 20k insert matepair library using wgsim (Li, 2011b). By this process we were able to improve the PacBio error corrected reads by 0.49%. Figure 3.1 depicts the workflow of error correction method used for this genome.

Phase 3: Final third round of correction was done using self error correction approach by using PBCR (Koren et al., 2012). By this method, the remaining uncorrected long reads were used to correct the errors. Here we were able to correct additional 13.2 % of reads, making the error correction to 47%.

Figure 3.1: Improved 3-way error correction method for *P. ramorum* Pr102 isolate



3.1.3. Genome assembly optimization

For a gapless hybrid genome assembly, varieties of genome assemblers with different approaches were used to assemble this genome. We have also used Illumina short reads to produce a better quality assembly. Total 47% of error corrected reads were used to assemble the genome. We tried five different combinations of assembly approaches. Each assembly protocols were represented as **V2, V3, V4, V5**, and **V6** in this study.

The transcript assembly was performed on RNA-Seq reads to incorporate in the hybrid assembly process. Our assessment of assembly involved checking the core genes, Avh effector genes and validation using Quast (Gurevich et al., 2013).

3.1.3.1. List of assemblers used in this study

3.1.3.2. Celera assembler

This assembler is based on Overlap layout consensus approach. Reads were trimmed to obtain 98% accuracy of the read quality values. Celera assembly includes on the following steps:

a. Screener, b. Overlapper, c. Unitigger, d. Scaffolder, e. Consensus

This assembler works by screening the poor quality and contaminant sequences and removes them as a first step. This is followed by overlapping the reads and assembling them. In the next step, it generates long contiguous sequences called as contigs. It also needs unitigs generated with high confidence from earlier assembly. By using these unitigs, it tries to scaffold the unitigs to contigs and generates scaffolds. The consensus step fixes the constructed errors of a mosaic haploid representation of a diploid locus.

3.1.3.3. Minimus

This assembler also follows the principle of Overlap Layout Consensus paradigm (Peltola et al., 1984). It works by a combination of three processes. A) By hash-overlap: It increases the speed of and reduces the memory usage. B) Tigger: It identifies the sequence clusters from reads and uniquely assembles the region of the genome. c) Consensus-layout performs the consensus layout and generates multiple sequence alignment of reads.

3.1.3.4. SSPACE

This program is used for scaffolding the contigs. From our workflow, contigs from various programs such as minimus (Sommer et al., 2007) and CAP3 (Huang & Madan, 1999) were used in scaffolding process. This assembler takes preassembled contigs along with illumina libraries to produce the final scaffolds. Bowtie aligner plays an important role in aligning the contigs to reads and produce final scaffolds.

3.1.3.5. SSPACE-long read

Due to improvement in sequencing technology, SSPACE-long read (Boetzer & Pirovano, 2014) assembler was introduced. Disadvantage of SSPACE (Boetzer et al., 2010) is that it can only use illumina reads for scaffolding process. The advantage of using SSPACE-long read assembler is that it can utilise long reads, illumina reads and preassembled contigs.

3.1.3.6. Dedupe

Dedupe is also a part of Celera (Koren et al., 2012) assembler used to remove the duplicated sequences from already assembled genome.

3.1.3.7. Redundans

This assembler is specially designed for handling heterozygous genome, which mainly works on the principle of Reduction, Scaffolding and Gap filling (Pryszcz & Gabaldón, 2016). Due to heterozygosity repeat collapse, fragmented assemblies and misassemblies are the main problems when assembling complex genomes. To solve this problem during assemblies instead of choosing two haploid contigs from heterozygous region, this assembler removes one haploid contigs and retains the other to minimise the errors in assembly.

3.1.3.8. CANU assembler

The successor of Celera assembler which uses the better alignment algorithm of MinHash and a sparse assembly graph construction that avoids collapsing diverged repeats and haplotypes.

3.1.3.9. CAP3

It is a contig extension program; it extends the pre-assembled large number of contigs into larger contigs.

3.1.3.10. LRNA scaffolder

This assembler is an alternate scaffolding approach to scaffold the genome assemblies using transcript sequences (Xue et al., 2013). This assembler uses long transcriptome reads to order, orient and combine genomic fragments into larger sequences.

3.1.4. Assessment of genome assemblies

For checking the quality of genome assemblies CEGMA(Parra et al., 2007) was used to assess the presence of core genes in all versions of assemblies. Quality of genome assemblies were assessed using QUAST (Gurevich et al., 2013).

3.1.5. Downstream Genome analysis

3.1.5.1. Gene Prediction

The Gene prediction was done using Augustus (Stanke & Waack, 2003) and Scipio (Keller et al., 2008). RNA-Seq reads of *P. ramorum* Pr102 was used as hints to predict the genes from the gapless optimal assembly of Pr102 V6.

3.1.5.2. Transposon Mobile elements prediction

The mobile elements from final genome assembly were predicted using the program TransposonPsi (Haas, 2007).

3.1.5.3. CAZymes search

Other than effector proteins, carbohydrate active enzyme also plays an important role in the pathogenicity. To identify the genomic regions, encoding putative carbohydrate-active enzymes (CAZymes) were annotated online using the dbCAN database.

3.1.5.4. Effector prediction

Initially orf's were predicted using EMBOSS, followed by HMM searches of RXLR, WY and L motifs. The sequences which qualified the HMM were again searched for the signal peptides using Signalp3.0 (Bendtsen et al., 2004). Proteins associated with transmembrane were predicted using TMHMM v2.0 (Krogh et al., 2001) and targeted to any of the organelle using TargetP (Emanuelsson et al., 2007) were identified and excluded from our analysis. Effectors which were associated with glycolysilated associated sites were further identified using PredGPI (Pierleoni et al., 2008) and filtered. This was followed by MEME (Bailey et al., 2009) motif analysis for identifying motifs.

3.1.6. Results and Discussion

3.1.6.1. Enhanced Error correction of PacBio reads

Single-molecule sequencing generates several kilobases of long reads, has a great potential to improve the Genome assemblies, but the error rates for the long reads are high (Au et al., 2012). For *P. ramorum Pr102* strain, we have generated 4, 35,399 reads with the read coverage of 30 X in PacBio RS II instrument with the P5-C3 chemistry with the Average read length of 4.5 Kb. PacBio reads are prone to error and need to be error-corrected before assembly.

Phase-1: The Sanger unitigs from the *P. ramorum* Pr102 was used with the PacBio reads generated by the ECTools, the corrected reads were much less, only 147,429 of (37%) of reads were corrected using this approach.

Phase-2: This step includes correcting errors from the uncorrected reads of 67% (287970 reads). The illumina reads and V1 unitigs were used for generating the intermediate assembly using ALLPATHS assembler. The assembly thus obtained was used by the ECTools to correct the read indel errors. By this process we could only corrected as an additional of 0.49% (1418) reads.

Phase-3: This involves self error correction that used the uncorrected reads of 66.50% (286552 reads) from the steps as above. PBCR was used in this method. It corrected an additional 13.2% (57640 reads) reads making the total read correction to 47% from 33%.

From this process, it is inferred that genome assembly quality is heavily dependent on the sequencing chemistry as well. There was no optimal tool to handle the data from P5-C3 chemistry.

By using various combinations of tools and error correction approaches helped to correct errors from longer reads. This approach can be useful for other researcher as well to correct errors from P5-C3 chemistry.

3.1.7. Pr102 V6 is a better genome assembly than all other versions

3.1.7.1. Pr102 V2 assembly

In this method first round of error corrected reads 147429 (33%) were assembled using Celera assembler that produced 6730 contigs (96 Mb).The largest contig was 781846 bp and the smallest contig was around 2734 bp. Then the contigs were merged using minimus, with length cut off 15,000 bp. With this, we obtained around 1545 contigs with and 76 Mb size. For further extension of contigs, SSPACE assembler was used along with the illumina libraries. This step produced an assembly of 1407 scaffolds with 76 Mb size. For further improving the assembly, SSPACE-long read assembler was used. The final V2 assembly had 1114 scaffolds, with largest scaffold of 880281 bp, smallest scaffold of 15009 bp and the total length of 78285078 bp with calculated N50 value of 130116, and total gaps in the assembly was found to be 450621 bases.

3.1.7.2. Pr102 V3 assembly

The 3-way error corrected long reads (47%) were assembled using Celera assembler. Then the assembled contigs were provided as input to the Dedupe program that was used to remove the duplicate contigs if present in the assembly. Then the duplicate removed contigs were passed onto minimus program to produce the contigs. Further SSPACE was used along with the illumina reads to extend the contigs. SSPACE-Long read assembler was used for scaffolding with long reads. Finally, Redundans assembler was used which is designed for handling highly heterozygous genomes, produced less fragmented assembly. The assembly resulted in 2325 scaffolds, with largest scaffold of 781884 bp, and the smallest of 2508 bp and total length of 78.4 Mb. This assembly had gaps of 923395 bp and N50 size was 65030. When comparing the V3 assembly with V2 assembly, we found

no improvement in the assembly statistics. We designed the next assembly pipeline to further improve the assembly statistics.

3.1.7.3.Pr102 V4 assembly

Since the V3 assembly was fragmented with gaps a developed long read assembler Canu (Koren et al., 2017). This assembler produced 59 Mb of contigs. The contigs were further used by SSPACE assembler along with the illumina reads to produce contig-extended assembly. For scaffolding the contigs, with long reads we used SSPACE-Longread assembler. This step resulted in 920 scaffolds. The largest scaffold of 655506 bp, with the smallest scaffold of 3055 bp, with gaps of 431583 bp and the calculated N50 value of 116386. The assembled genome size was smaller than the expected genome size and more number of genes were absent in the assembly indicating an incomplete assembly.

3.1.7.4. Pr102 V5 assembly

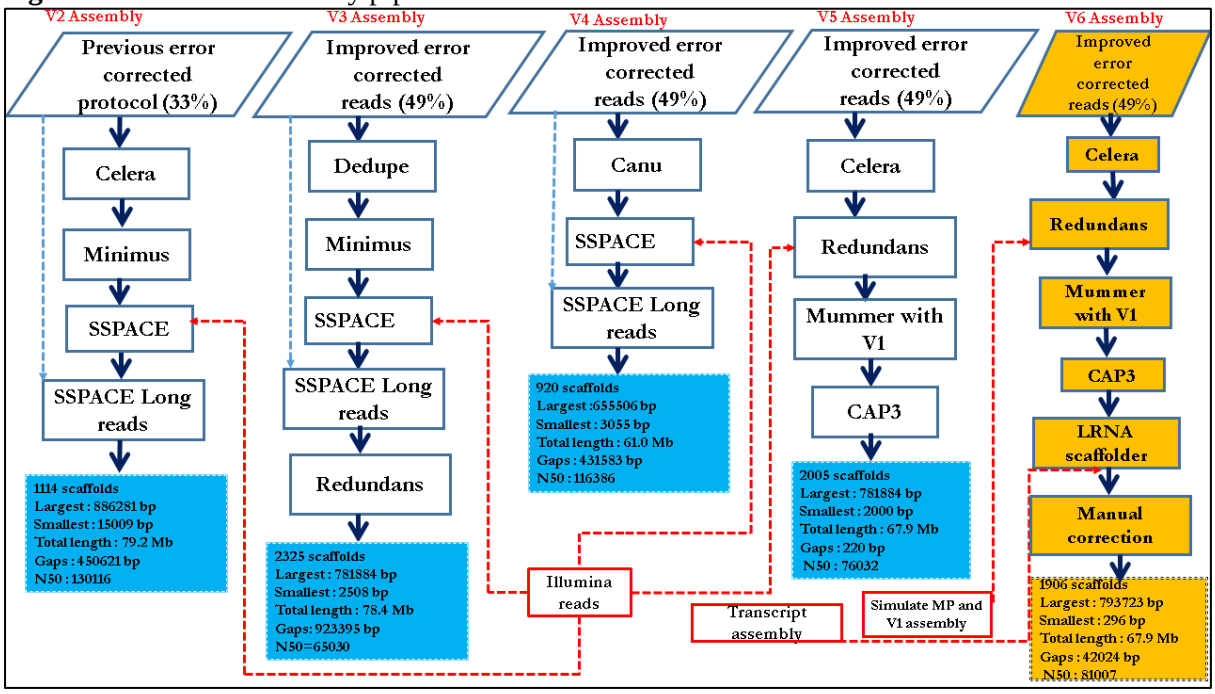
Here we modified the protocols used in V3 assembly to further improve assembly statistics. The 3-way error corrected PacBio reads (47%) were used for assembly with Celera assembler. The contigs were processed with Redundans with illumina reads and simulated matepair libraries using wgsim (6k and 20k from intermediate Allpaths assembly) The resulting assembly was aligned with the *P. ramorum* V1 assembly using Nucmer (Kurtz et al., 2004b). The Nucmer aligned region was checked for unmapped contigs. The unmapped contigs were further assembled using CAP3 resulting in V5 Pr102 assembly, which resulted in 2005 scaffolds, with the largest of 81884 bp, and smallest scaffold of 2000 bp, to the total length of 67.9 Mb with the gap of 220 bases and N50 value of 76032. This assembly also missed several core conserved and effector genes, and hence there was a need for further refinement of the assembly.

3.1.7.5. Pr102 V6 (final assembly)

The error corrected reads (47%) were assembled using Celera. The contigs obtained from this step were again assembled using Redundans assembler along the simulated matepair libraries (As described in Pr102 V5 method section) and the illumina libraries. The resulting assembly was aligned using Nucmer with the *P. ramorum* V1 assembly to identify the unmapped contigs. The unmapped contigs were assembled again using CAP3. Transcript assembly of *P. ramorum* Pr102 from RNA-Seq reads was used with the LRNA assembler to improve the inclusion of genic region in the assembly. The regions containing effectors were seen to have some read errors and were manually corrected. Final Optimal assembly resulted in 1907 scaffolds, with largest scaffold of 793723 bp, smallest scaffold of 296 bp, and gaps of 42024 bases and the N50 value of 81007. Here the effector coding genes were used as a hint to correct some of the scaffolds which was misassembled. After examining the different versions of assemblies, the better draft assembly of V6 was finalized for further

downstream data analysis. The final assembly size was approximately 68 Mb. All versions of assembly protocols are represented in Figure 3.2.

Figure 3.2: Genome assembly pipelines used for all versions of *P. ramorum* Pr102 isolate



From above comparisons, we found that combinations of tools can be used to optimize genome assembly. Higher percentage error corrected reads was used in Celera assembler to produce final assembly. RNAseq reads were also used to optimize genome assembly. RNAseq helped to improve the gene models.

3.1.8. Genome assembly assessments

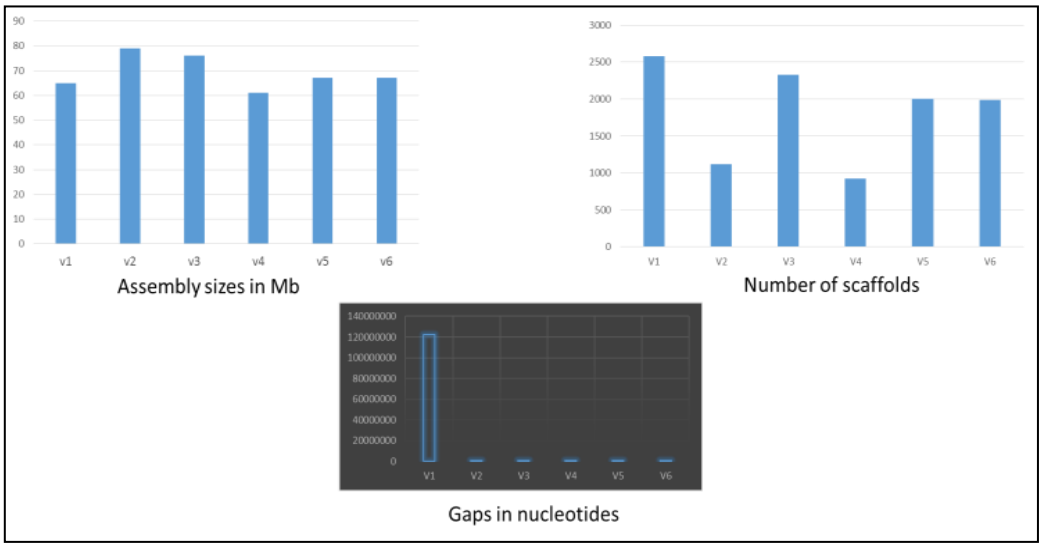
Gene space assessments were performed using Eukaryotic COGs. Our gene space assessments were done for all versions of genome assemblies for checking the completeness of genome including the V1 version of assembly. Comparison shows V5 and V6 assembly having more complete sets of core orthologous genes than other version of assemblies. We had chosen V6 as a final assembly since some effector coding genes were found to be missing in earlier assembly versions. The results of gene space assessments are represented in Table 3.2.

Table 3.2: Number of core genes present in all versions of *P. ramorum* Pr102 assemblies

Assembly version of Pr102	**No of core proteins (248 completely highly conserved CEG)**	**% of completeness**	**Out of 458 core genes present in genome**
V1	236	95.16	412
V2	237	95.56	412
V3	236	95.16	413
V4	237	95.56	416
V5	238	95.97	414
V6(optimal)	**238**	**95.97**	**414**

Other comparisons including gaps, assembly length with all assemblies are plotted and provided in Figure 3. 3.

Figure 3.3: Comparisons of all versions of assemblies with length, gaps and number of scaffolds in *P. ramorum* Pr102



3.1.8.1. Quast assessments

Genome assemblies were quality assessed for contig length, number of gaps, N50 values. Comparison results on different assembly versions shows that V6 assembly contains less number of gaps and fragmented contigs were also reduced in V6 version assembly. The assessments are shown in Figure 3.4.

Figure 3.4: Genome assembly quality validation using Quast for all 6 versions of assemblies of *P. ramorum* Pr102

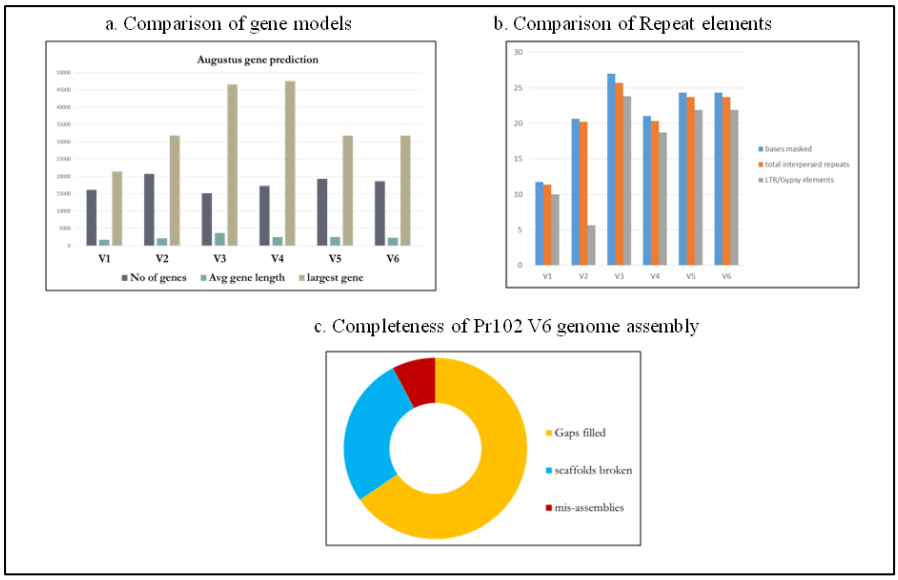
Worst Median Best ☑ Show heatmap

Statistics without reference	Phyra	pr102_V2	pr102_V3	pr102_V4	pr102_V5	2017_V6_Pr102_assembly
# contigs	2576	1114	2325	920	2005	1906
# contigs (>= 0 bp)	2576	1114	2325	920	2005	1907
# contigs (>= 1000 bp)	2576	1114	2325	920	2005	1901
# contigs (>= 5000 bp)	1279	1114	2112	912	1692	1663
# contigs (>= 10000 bp)	573	1114	1811	871	1632	1605
# contigs (>= 25000 bp)	261	702	753	572	598	581
# contigs (>= 50000 bp)	170	430	324	343	308	301
Largest contig	1 244 001	880 281	781 884	655 506	781 884	793 723
Total length	66 652 401	78 299 969	78 463 906	61 035 742	67 998 772	67 904 312
Total length (>= 0 bp)	66 652 401	78 299 969	78 463 906	61 035 742	67 998 772	67 904 608
Total length (>= 1000 bp)	66 652 401	78 299 969	78 463 906	61 035 742	67 998 772	67 900 068
Total length (>= 5000 bp)	62 068 209	78 299 969	77 633 232	61 005 525	67 135 885	67 241 069
Total length (>= 10000 bp)	56 381 157	78 299 969	75 397 344	60 678 098	66 711 976	66 826 183
Total length (>= 25000 bp)	51 982 817	70 584 869	57 986 385	55 490 061	49 900 282	50 213 015
Total length (>= 50000 bp)	48 409 377	60 751 405	43 458 489	47 210 926	40 156 632	40 840 920
N50	308 042	130 116	65 030	116 386	76 032	81 007
N75	40 567	57 197	23 910	55 918	23 644	24 105
L50	63	160	250	131	210	195
L75	206	393	789	316	644	611
GC (%)	53.86	54.32	52.4	54.09	53.98	53.99
Mismatches						
# N's	12 227 865	450 640	923 395	431 614	220	42 024
# N's per 100 kbp	18 346	575.53	1176.84	707.15	0.32	61.89

3.1.8.2. Comparison of genome assemblies

All versions of genome assemblies were compared for assessing of quality and genome completeness using various methods. Comparisons indicate that overall V6 genome assembly is optimal in capturing repeats and complete sets of gene models. The comparison results are represented in Figure 3.5.

Figure 3.5: Comparison of repeat contents and predicted gene statistics among different assembly versions of *P. ramorum* Pr102



3.1.8.3. Comparisons of gene models:

Gene models of Pr102 (V1 to V6) were compared with each other, while comparing we found that gene model from V6 was better. All other assemblies had fragmented genes, certain important housekeeping genes were missing from the assembly. We decided to choose V6 as final assembly containing maximum complete genes.

3.1.8.4. Comparison of Repeat elements:

Repeat elements from the assemblies were compared with each other. While comparing it was found V6 was having consistently more number of repeat elements. Long read sequencing captured the higher number of repeat elements while comparing older assembly of *P. ramorum*, suggesting that gap closed regions might had complex repeats which was not captured in sanger reads in 2006.

3.1.8.5. Completeness of genome assembly

Genome assembly completeness was assessed by comparing V6 assembly to older V1 assembly. In our comparison it was found that more gaps from Sanger V1 assembly was filled and misassemblies were corrected. Fragmented broken scaffolds were reduced in V6 version of the genome.

3.1.9. Downstream analysis results

We predicted 18,705 genes in V6 assembly; largest gene was of 31,832 bp. The results were compared with V1 assembly. The earlier version of genome had 16,134 genes with average length of 1673 bp. The extra predicted genes present in this version were actually present in the gap regions of the earlier published *P. ramorum* genome (Tyler et al., 2006).

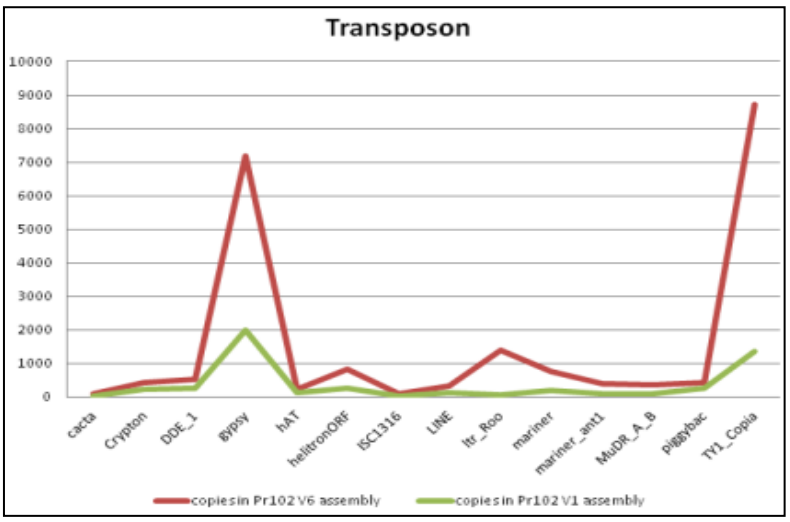
CAZy analysis in the protein coding genes of V6 assembly identified 100 unique categories of CAZymes. Most abundant CAZy family was found to be α-1, 3-glucosyltransferase (GT1), Polysaccharide Lyase family (PL3), Glycoside Hydrolase Family (GH5), Cinnamoyl esterase Family (CE1 and CE10).

Repeat analysis also identified more number of repeat elements (V1: 24% versus V6:40%). The number of transposon elements was also found to be in large numbers compared to V1 assembly version.

Our effector prediction analysis also identified more number of RXLRs (V1: 370 versus V6:375).

To conclude, the latest assembly version generated after inclusion of long reads and several combination assemblers could capture repeat rich regions and transposable elements that were missing earlier. This may be due to the fact the first generation sequencing methods are incapable of capturing repeats and the assemblers were also inefficient in handling large repetitive elements.

Figure 3.6: Comparison of transposon elements in V1 and V6 assemblies of *P. ramorum* Pr102



B. Haplotype phased diploid genome assembly of *Phytophthora ramorum* ND886 using Pacbio P6-C4 sequencing

3.2.1. Introduction

Phytophthora ramorum is a devastating pathogen that infects a wide range of plant species and is invasive to Western Europe and North America. The NA1 lineage of *P. ramorum* ND886 is the causal affects Camellia ornamental plants (Gruenwald et al., 2008). Despite its primary mode of asexual reproduction and the lack of sexual recombination (Ivors et al., 2004), the pathogen shows diverse colony morphology and aggressiveness.

The 2006 assembly contained 12 Mb of gaps. Therefore, the higher level architecture of the pathogen genome was obscured and potentially, many effectors and repetitive elements were lost, or their relative locations were unclear. Moreover the assembly of Pr102 V1 was a consensus assembly comprising of a mosaic of haplotypes based on the most frequent sequence reads at each polymorphic site. This fact impedes analyses involving allelic variation and linkages between polymorphisms.

Therefore, a need for more detailed studies of pathogen evolution motivated us to obtain a more complete diploid genome assembly with fewer gaps. With the help of long read sequencing it is quite possible to attain haplotype block information from the chromosome. It is necessary to have a gold standard control reference genome to study the population genetics with other isolates.
We decided to sequence *P. ramorum* ND886 using PacBio P6-C4 chemistry and illumina sequencing. We sequenced *P. ramorum* ND886 which has one of the host as Camellia flowering plant, a normal healthy euploid.

Haplotype phasing is the process by which alleles are sorted by chromosome. For analysing genetic variation in diploid genomes, it is essential to use phased assemblies. Several promising genome-wide tools, genotyping tools and bioinformatics workflows have been developed to improve the determination of haplotype phase (Browning & Browning, 2011). Phasing provides valuable resources for studying allelic variation, Linkage Disequilibrium (LD), and larger scale genetic variation.

3.2.2. Materials and Methods

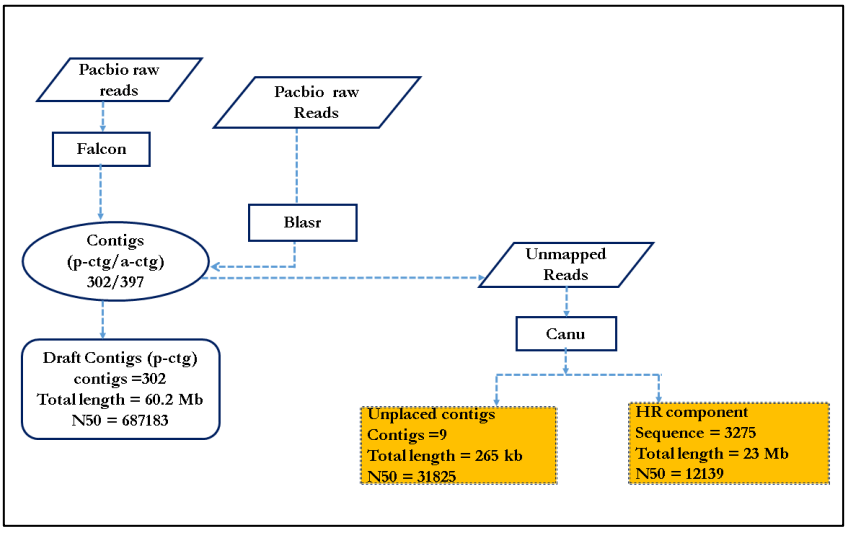
3.2.2.1. Genome sequencing

P. ramorum ND886 isolated from Camellia was sequenced using PacBio P6-C4 chemistry with average read length of 10.5 kb. Long reads of 70X coverage with total 402170 reads were generated for this study. Also we included Illumina reads from 3 different libraries with the coverage of 55X, 63X and 86X respectively.

3.2.2.2. Genome assembly of *P. ramorum* ND886

The ND886 isolate from camellia flowering plant was assembled using FALCON (Chin et al., 2016b) assembler, which is based on the overlap consensus layout algorithm. FALCON produced primary contigs of length 60 MB. Next we mapped the PacBio reads to the primary contigs produced by FALCON using BLASR (Chaisson & Tesler, 2012). A total 234 Mb of PacBio reads were unmapped. The unmapped reads were assembled using Canu (Koren et al., 2017) and were named as unplaced contigs since it is unknown part of the genome that the reads belong to. The unassembled part of the genome was called the 'highly repetitive (HR)' region. The detail of the genome assembly is represented in Figure 3.7.

Figure 3.7: Genome assembly pipeline of *P. ramorum* ND886

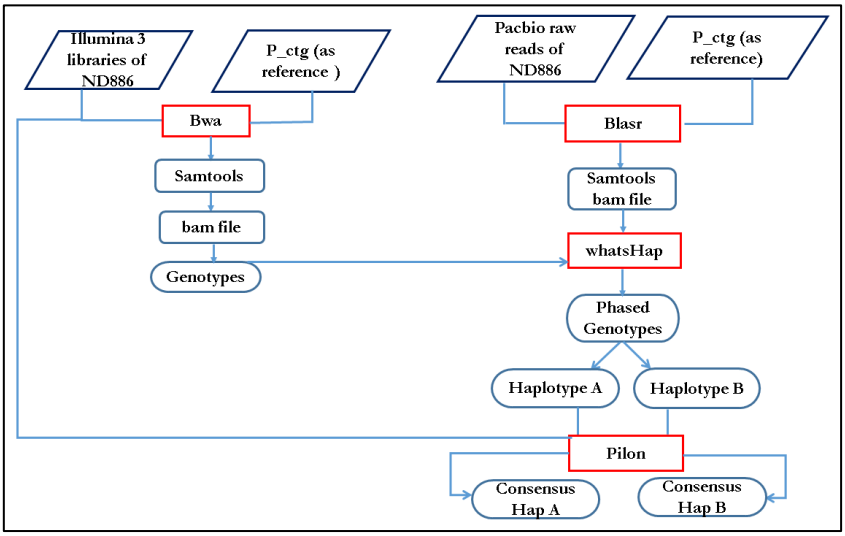


3.2.2.3. Haplotype Phasing in *P. ramorum* ND886 isolate

We phased the genome of ND886 to generate a diploid assembly. We started by mapping reads from three Illumina libraries with average read length of (100 bp, 100 bp, 150bp) to the reference genome assembly produced by FALCON (p-ctg) using BWA-mem (Li, 2013). Genotypes were generated using Samtools (Li et al., 2009) and Bcftools (Li, 2011a). Genotype filtration was done using SnpEff (Cingolani et al., 2012) followed by comparison of all heterozygous loci in all Illumina libraries. A subset of heterozygous sites included loci that had two alternative alleles representing the genotype

(e.g. 1/2).These sites were sequencing errors in the Illumina libraries because the PacBio reference allele was not represented. The PacBio raw reads were mapped with the reference assembly using BLASR (Chaisson & Tesler, 2012). The PacBio read alignment and illumina genotypes were used for phasing using WhatsHap (Garg et al., 2016). The haplotigs were arbitrarily categorized as haplotype **A** and haplotype **B** within each block. The consensus haplotypes from illumina were generated to reduce indel errors using Pilon (Walker et al., 2014). The detailed protocol describing haplotype phasing is presented in Figure 3.8

Figure 3.8: Haplotype phasing of *P. ramorum* ND88 isolate



3.2.3. Downstream analysis

3.2.3.1. Gene prediction

P. ramorum ND886 gene prediction was done using AUGUSTUS (Stanke & Waack, 2003) with gene models from *P. ramorum* V1 (Tyler *et al* 2006) as initial training set. Later the predicted gene models were retrained for obtaining the final gene model. Twice training of genes was done to get optimized accurate gene models.

3.2.3.2. Genome Annotation

Gene space was assessed with BUSCO 2.0 (Simão et al., 2015b) using the stramenopile COG categories to assess the genome assembly of *P. ramorum* ND886 genome. Functional annotation was done with interproscan (Jones et al., 2014) and BLASTP searches were carried out against the nr database. CAZyme proteins that are involved in carbohydrate metabolism were identified by searching the CAZY database at dbCAN (Yin et al., 2012). Further to find CAZYs that are associated with virulence or pathogenicity, the identified CAZY proteins were searched against the pathogen host interaction (PHI) database (Winnenburg et al., 2006).

Repeat Masker 4.0.6 (Smit et al., 2016) was used to identify repetitive elements. The Repeat library of oomycetes was used to identify the repetitive elements from the genomes.

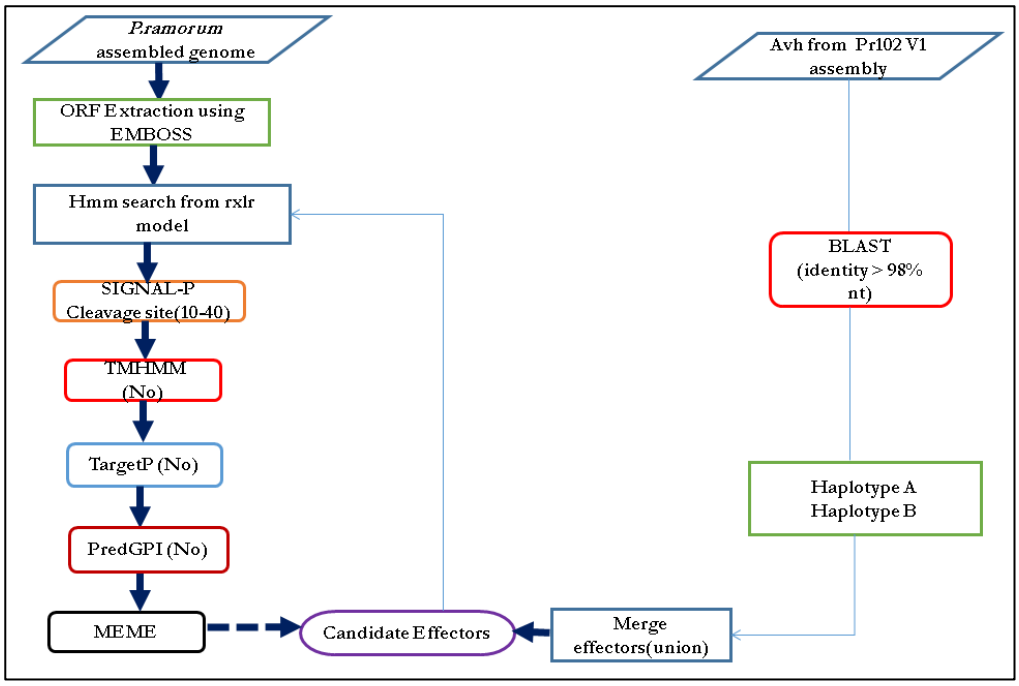
3.2.3.3. Effector Prediction

The haplotype phased consensus assembly was used for effector prediction and downstream analysis of *P. ramorum* ND886 isolate. For ND886 isolates RXLR effector prediction were done using two different approaches,

i) The Galaxy pipeline (Giardine et al., 2005) by using the models of (Win et al., 2007) and (Whisson et al., 2007)

ii) Our in-house pipeline method consists of initial orf extraction using EMBOSS, followed by HMM searches of RXLR, WY and L motifs. The sequences which had the motif by HMM search were again searched for the signal peptides using Signalp3.0 (Bendtsen et al., 2004). Proteins associated with transmembrane regions were predicted using TMHMM v2.0 (Krogh et al., 2001). TargetP (Emanuelsson et al., 2007) prediction which identifies proteins targeted to different organelles was also conducted. These predicted proteins were excluded from our analysis. Avh effectors which were associated with glycolysilated sites were further identified using PredGPI (Pierleoni et al., 2008). This was followed by MEME (Bailey et al., 2009) motif searches for identifying the motifs. First round predicted effectors were used to build new hmm model for searching in emboss orf. This process was continued until no new RXLRs were predicted from EMBOSS orf's. The detailed prediction of RXLR is represented in Figure 3.9.

Figure 3.9: RXLR prediction pipeline from in-house method used for *P. ramorum* ND886 consensus haplotypes



3.2.3.4. Crinkler prediction

Crinklers from the haplotype phased genome assembly of *P. ramorum* ND886 were predicted using HMM profiles which were built from published CRN effectors, as described by (Yin et al., 2015) .

3.2.3.5. Genome architecture studies on *Phytopthora* genomes

The genome architecture was assessed by calculating the flanking intergenic regions (FIRs) of every gene in the genome. To determine the architecture of the genome, the distances of each gene (including effector genes) to its closest gene neighbours in the 5' and 3' directions were calculated from the gff co-ordinate file using R and Perl scripts. Two-dimensional data binning was then performed (Saunders et al., 2014).

3.2.4. Results and Discussion

3.2.4.1. Phased Diploid genome assembly of *P. ramorum* ND886 using FALCON

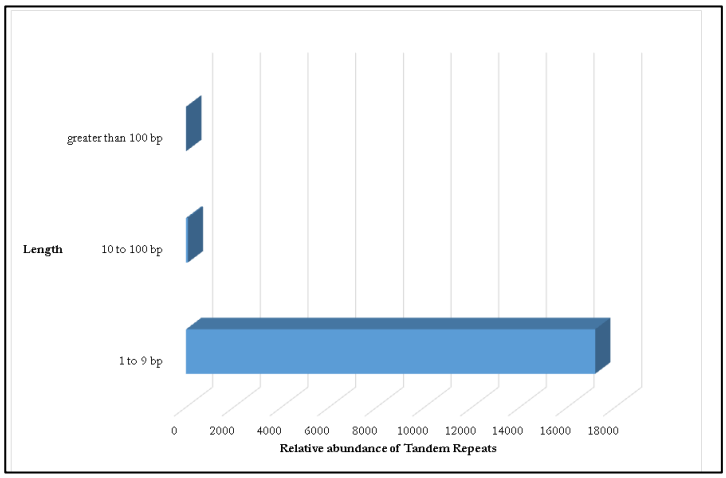
FALCON (Chin et al., 2016b) produces two types of contigs, namely primary contigs and associated contigs (which mainly contain regions with structural variants allelic to the primary contigs). Primary contigs were used for downstream analysis. Using FALCON, 302 primary contigs of length 60,221,882 bp with contig N50 of 687,183 bp were obtained. There were 397 associated contigs (most likely allelic to the primary contigs) comprising 18,850,872 bp with an N50 of 55,130 bp. Assembly statistics are provided in Table 3.3.

Table 3.3: Genome assembly statistics of *P. ramorum* ND886

Assembly Name	**Total contigs**	**Assembly size (in bp)**	**N50**	**Total Gaps**	**No of Busco COG's**	**No of genes**
Pr102 V1 (BM Tyler et al 2006)	2,576	66,600,000	Scaffold N50: 308,042 bp Contig N50: 47,500 bp	12.2 Mb	230	16134
ND886 (primary contigs) + unplaced contigs	**302 + 9**	**60,221,882 + 265,202**	**687,183, 31,753**	**NA**	**155+ NA**	12,337 + 42
ND886 (associated contigs)	397	18,850,872	55,130	NA	66	NA
ND886 (consensus Haplotype A)	302	60,284,349	688,117	NA	231	14,470
ND886 (consensus Haplotype B)	302	60,285,334	687,920	NA	231	14,998
ND886 (Highly Repetitive component)	3,275	23,217,622 (. Most reads are singletons)	12,139	NA	NA	NA

PacBio raw reads were mapped back to the primary contigs and associated contigs. There were a total of 13,300 unmapped reads which were further assembled using Canu (Koren et al., 2017).These unmapped reads produced 9 contigs encompassing about 265 kb (smallest 13,373 bp, largest 76,986 bp and N50 value of 31,753 bp), which we designated as unplaced contigs. These sequences were not incorporated into the initial assembly probably because of the presence of complex repetitive elements. Canu also produced an unassembled low coverage sequences which was named as highly repetitive (HR) component that were included in the analysis for checking for tandem repeats. Tandem repeat units of size >=10 with repeat numbers larger than 100 were also present in this assembly (Figure 3.10).

Figure 3.10: Tandem repeat elements in the HR region of the *P. ramorum* ND886



Assembly of heterozygous genomes has always been a challenge, especially when they are repeat-rich. In this study, PacBio genomic long reads were used to assemble the genome of *P. ramorum* isolate ND886 to produce a relatively better draft contig assembly of 60.5 Mb, compared to the previously published Sanger assembly of the genome of isolate Pr102. The earlier assembly (Pr102 V1) was produced using Sanger reads and the sequence content of the assembly was 54.4 Mb excluding 12.2 Mb of gaps within the 66.6 Mb assembly (Tyler et al., 2006). The ND886 primary contig assembly had more repeat content, compared to the Pr102 V1 assembly; the repeat percentages were 48% and 29% respectively.

Phytophthora species have genomes with highly repetitive regions and long stretches of repeats rendering them very difficult to assemble (Tyler et al., 2006, Haas et al., 2009).

Large numbers of tandem repeat elements were found in the HR component of the ND886 genome. Some of the abundant tandem repeats elements are represented in Table 3.4. We checked for these tandem repeats in the reference genome of *P. sojae* we couldn't find any and Pr102 V1 contained very few copy of these elements. This suggests that these tandem repeat elements are either unique to P.

ramorum or they are missing from the P. sojae assembly because they fell into unsequenceable or unassembled regions.

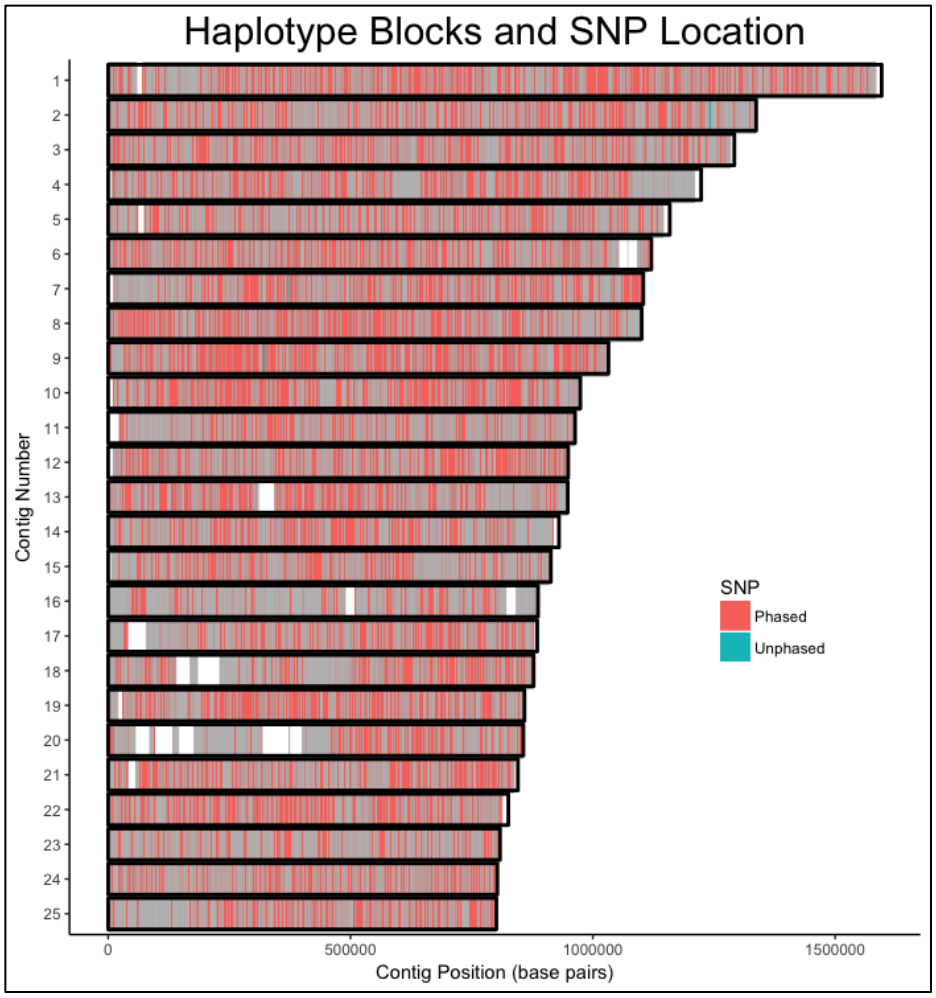
Table 3.4: Abundant Tandem elements in the HR region of *P. ramorum* ND886

Predominant Repeat sequence in HR component	**Unit length bp**	**Number in HR region**	**Comparison with Phyra V1**
GCGACCGTTG	10	948	No match
TCGCCAACGG	10	806	8
CGCCAACCCAT	11	228	No match
AACGGTCGCC	10	195	8
GTTGGCGATGG	11	163	No match
GGAGACCACGACGGGAGAC	19	113	No match
CTTGTTGATTAA	12	72	No match

3.2.4.2. Genome phasing using whatsHap

The reconstruction of haplotypes can capture genetic variants that can be missed in unphased genomes (e.g. indels, Structural Variants, and gene conversion). The assembled primary contigs of ND886 were used for read-based phasing with WhatsHap using vcf files containing ND886 Illumina data as well as bam files from the PacBio alignment (Garg et al., 2016), to produce a haplotype-phased diploid assembly . The phasing resulted in two haplotypes (**A** and **B)** for heterozygous regions in the genome. The haplotype phasing of ND886 identified 223,294 heterozygous variants. Of these, 222,892 variants were phased into 345 blocks, encompassing 54 Mb of the genome assembly. The largest 25 scaffolds, showing phased and unphased SNP's, are plotted in Figure 3.11.

Figure 3.11: Plot representing phased haplotypes from largest 25 contigs of *P. ramorum* ND886



The average number of variants per block was 646 and the median block length was 44,513 bp. Phasing the haplotypes of *P. ramorum* provides a powerful resource for future population genomics studies on large scale polymorphisms, understanding points of mitotic recombination, and to better understand the relationship between DNA sequence and RNA expression. Factors such as the SNPs preserve during asexual lineages, mutation events, and recombination rates can be obtained. The creation of allelic diversity in an asexually reproducing population such as *P. ramorum* can result in phenotypic diversification and the increase of evolutionary potential.

Haplotype phasing of assemblies can be challenging because it relies on marker density, read length, and read accuracy. In our haplotype-phased assembly, relationships between haplotypes in different blocks remain unknown as complete haplotypes could not be inferred. Though our marker density averaged 0.62 SNPs/1000bp, and read length was 10.57 kb on average. In addition, PacBio reads have a higher error rate, which makes it difficult to distinguish between a heterozygous region and error when coverage is low. The reads were still not long enough to span all regions with low marker density. For the first 25 largest contigs, SNP density is plotted in Figure 3.12 and Figure 3.13.

Figure 3.12: Coverage plots for largest 25 contigs of *P. ramorum* ND886

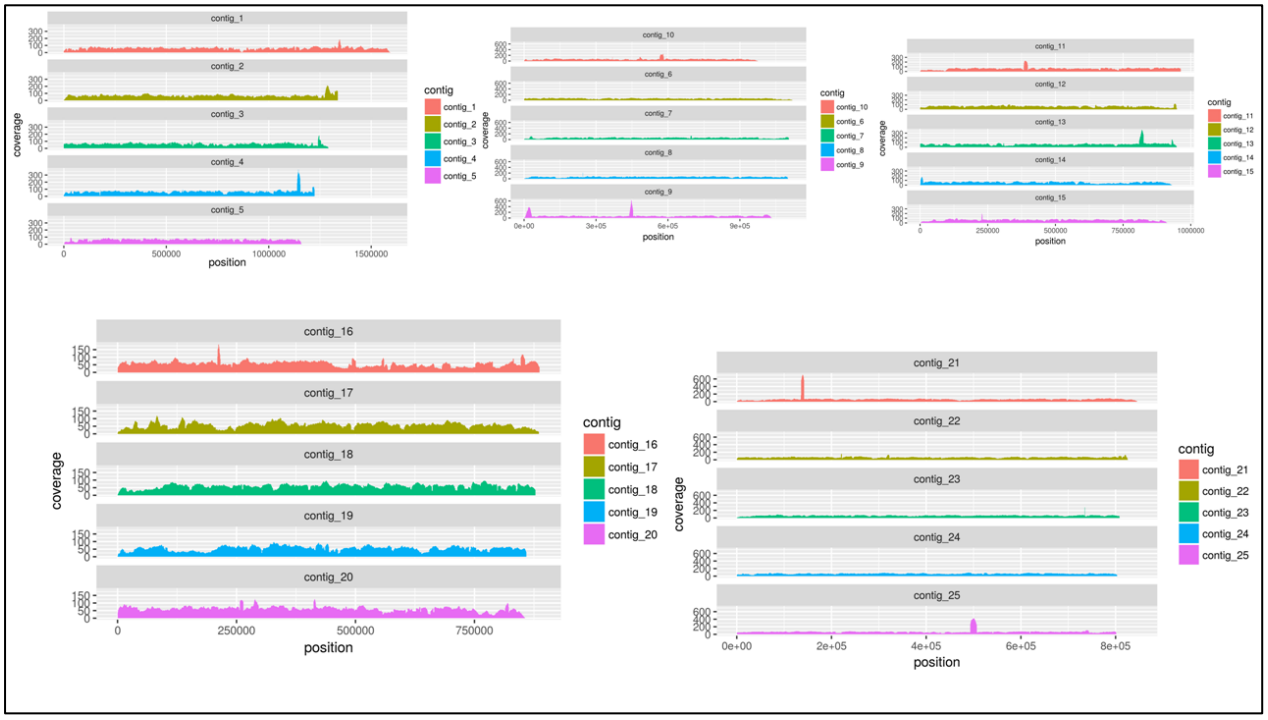
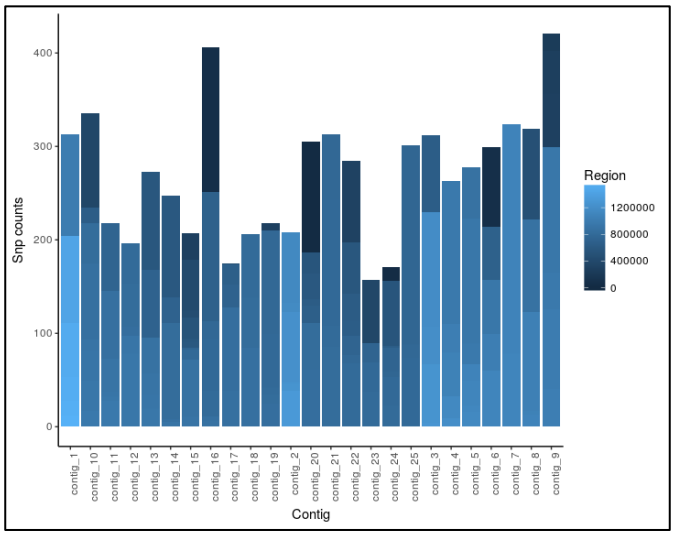


Figure 3.13: SNP density for largest 25 contigs of *P. ramorum* ND886 isolate

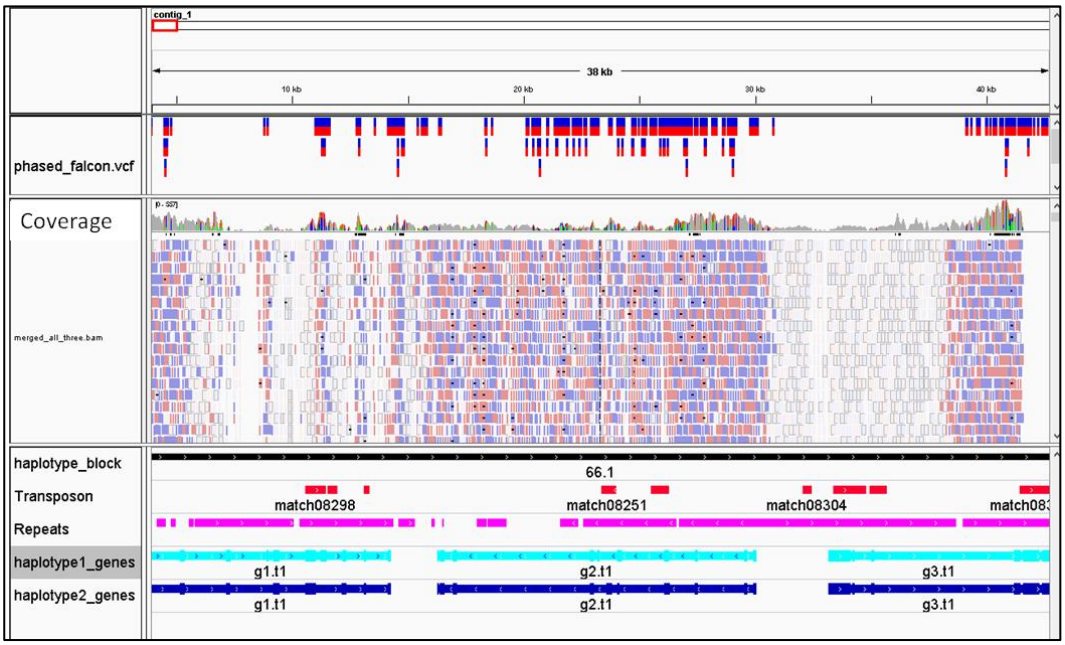


The 345 haplotype blocks overlapped with 11,227 of the protein coding genes, which is about 91% of all the protein coding genes. This includes 328 RXLR and 19 CRN effector genes.

We found 2781 unique Pfam domains in the haplotype blocks. We checked for the remaining 9% of protein coding genes outside the haplotype blocks; those genes were enriched in transposon-associated annotations such as Reverse transcriptase, GAG, Integrase core domain, DUF 4219, and GAG-pre-integrase domain. The annotations also include Zinc Knuckle, ABC transporter, Ankyrin

repeats and Sugar efflux transporter for intercellular exchange. From our analysis, contig_1 (1,595,961 bp) contained the longest haplotype block of size of 1,513,201bp, and contained 7,265 phased variants spanning more than 420 protein coding genes (Figure 3.14 **).**

Figure 3.14: Contig_1 representing the largest haplotype phased block of *P. ramorum* ND886 isolate



3.2.4.3. Genome analysis of *P. ramorum* ND886

The phased haplotypes were polished using the ND886 Illumina reads to reduce the indel errors from the phased assemblies. The consensus haplotypes were used for downstream analysis.

Gene prediction was performed on the genome assembly (primary contigs) and haplotype-phased consensus assembly (A and B) using Scipio (Keller et al., 2008) and Augustus. A total of 12,337 genes were predicted in the primary contigs, and 42 in the 265 kb of unplaced contigs. 14,470 and 14,998 genes were predicted in consensus Haplotypes A and B.

By using oomycete repeat sequence libraries from Repeat Masker 4.0.6 (Smit et al., 2016) we identified repetitive elements comprising 48.11% of both the haplotypes.

The gene space coverage was evaluated using BUSCO V2 (Simão et al., 2015). The stramenopile core consisting of 234 COG's were used for the assessment. We found 230 non-redundant COG's in the *P. ramorum* Sanger assembly. From the polished haplotype assembly, 200 and 201 COG's were identified in the two haplotypes. We observed multiple COGS' were missing due to PacBio indel errors that caused incomplete gene models, which affected the first step in the BUSCO analysis. In order to get around this problem we took the missing COGs and used tblastn to align them back to the assembled haplotypes. This resulted in identification of a total of 231 BUSCOs with only 3 BUSCO genes missing from both haplotypes.

Proteins containing carbohydrate-binding modules have been reported as one family of virulence factors in plant pathogenic oomycetes (Brouwer et al., 2014). Carbohydrate-active enzymes

(CAZymes) encoded in the *P. ramorum* ND886 and *P. sojae* genomes were predicted from the dbCAN database (Yin et al., 2012). 107 CAZy families containing 554 protein sequences were predicted in *P. ramorum* ND886 haplotype A and 571 protein sequences with 107 CAZy families were predicted in consensus haplotype B. The most predominant families were found to be PL3 (Polysaccharide Lyase Family 3), glycosyl transferases GT71 (Glycosyl transferase) and glycosyl hydrolases (GH17). CAZyme proteins predicted to be potentially involved in pathogen-host interactions were 542 and 557 from ND886 phased consensus haplotypes A and B, respectively. Table 3.5 contains the summary of CAZy involved in pathogenicity.

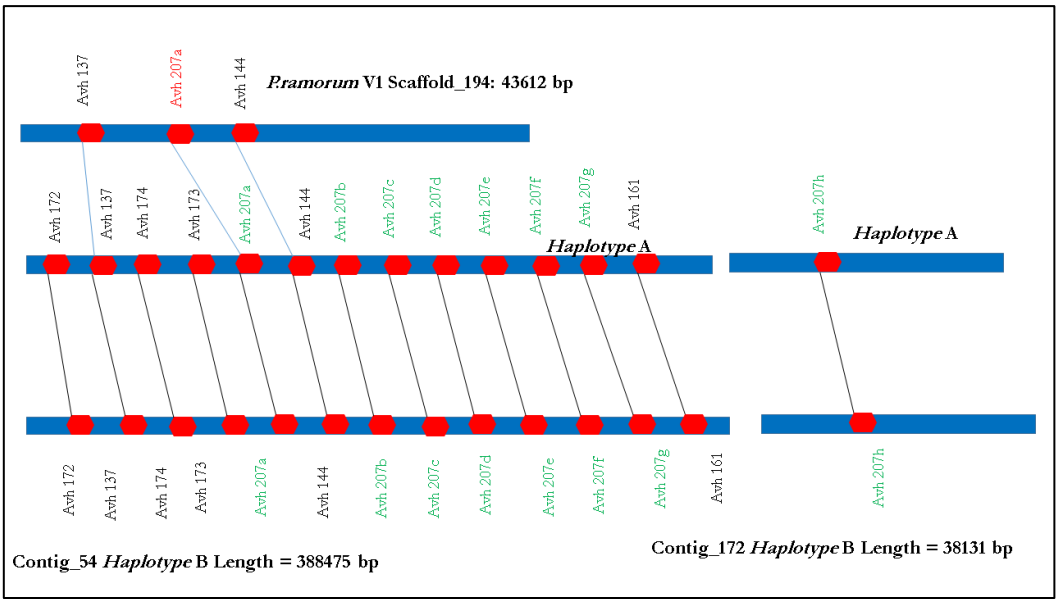
Table 3.5: CAZy associated with the virulence in *P. ramorum* ND886 phased haplotypes

CAZy Family associated with PHI-database	**Copy Number in *P. ramorum* ND886 haplotypes**
glycoside hydrolases	46
carbohydrate esterases	12
carbohydrate-binding modules	11
glycosyl transferases	30
auxiliary activities	8
polysaccharide lyases	4

3.2.4.4. Avh new effector identification in ND886 assembled genome

In Pr102 V1 assembly 370 RXLR effectors were predicted. We predicted a total of 393 and 394 RXLR effectors from the phased haplotypes A and B using our approach. Our HMM search on the emboss ORF's identified 286 WY motifs out of that 176 had signal peptides. There were 224 effectors which were 99 to 100% identical to Pr102 V1 effectors. Additionally our prediction found 14 newly identified paralogs in ND886 corresponding to 8 effectors in Pr102 V1 assembly. There also were 131 RXLRs with indels relative to Pr102 effectors from both the consensus haplotypes. Probably this was due to the sequencing error. We found 24 new effectors in both the consensus haplotypes A and B. One of the interesting effector from our study was Avh207 that had 7 paralogs that were newly identified in the ND886 haplotype phased consensus assembly from the contigs of 54 and 172 (Figure 3.15) represents the effectors from the phased consensus haplotypes. Large number of effectors were predicted in contig 31 (23) and contig 50 (24).

Figure 3.15: Avh207 paralogs from the contigs of 54 and 172 from consensus phased haplotype assembly of *P. ramorum* ND886 and *P. ramorum* Pr102 V1



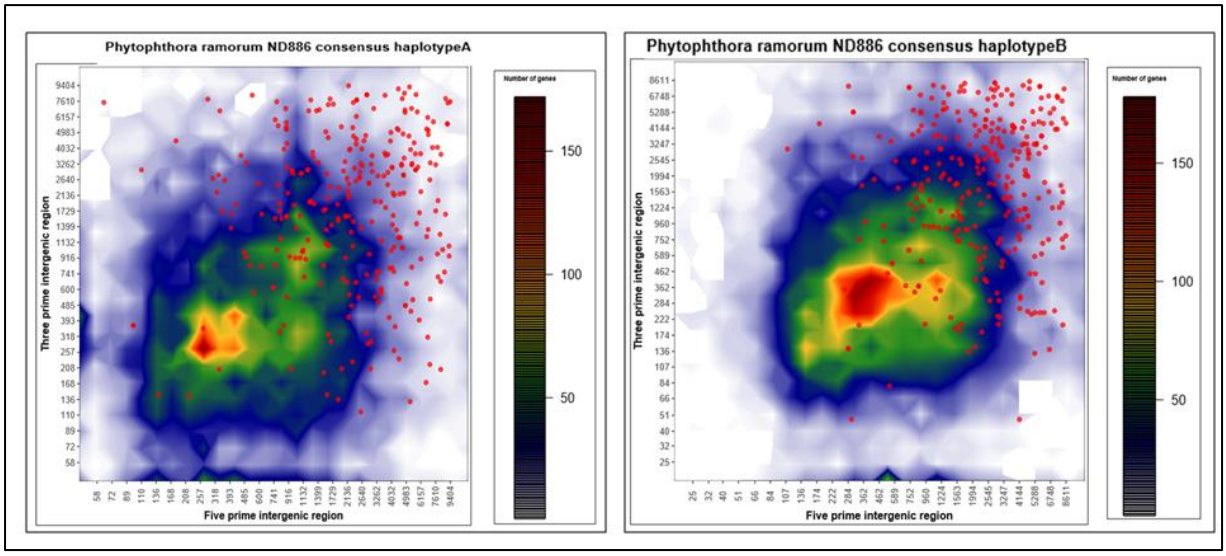
These organisms secrete another group of effectors belonging to the Crinkler (CRN). The CRN effectors contains conserved N-terminal LXLFLAK motif, DWL domain, and diverse C-terminal effector domains (Haas et al., 2009, Schornack et al., 2010, Stam et al., 2013, Yin et al., 2017a). The previous genome *P. ramorum* V1 had about 19 *Crinkler* genes (Haas et al., 2009a). We identified 25 CRN effector genes from the consensus haplotypes A and B by (Yin et al., 2017) and (Jiang & Tyler, 2012), from this total 15 were polymorphic between haplotypes. Approximately 50% of the CRN effectors predicted by our method did not have signal peptides that could be recognized by SignalP 3.0 (Käll et al., 2004).

3.2.4.5. Genome Architecture Analysis

For the phased consensus haplotypes we tried to study the genome architecture. All the *Phytopthora's* and plant pathogens are known to have the bi-partite genome architecture such that the effector genes are present in highly evolving, repeat rich, gene sparse region whereas core genes are found in the slow evolving, gene dense and repeat poor region. . In contrast, those that can be identified by host immune receptors can trigger the plant's defense response. This co-evolutionary struggle constantly selects for changes in the effector genes. Over time, the genes become un-recognizable through standard gene prediction methods. Therefore, in order to reduce false negatives, refined bioinformatics prediction methods are required to predict the RXLR effectors. For studying the architecture, two-dimensional based intergenic distance based binning of 5', 3' intergenic distances of effectors; core genes and all genes were calculated using (Dong et al., 2015). The mean 5' and 3' intergenic distances for all genes were calculated to be 1151 bp and 990 bp respectively in haplotype A. However, for RXLR and CRN effectors, the mean 5' and 3' intergenic distances were 4,638 bp and

4,753 bp for Haplotype A. In Haplotype B, the mean 5' and 3' distances for all genes were 1,176 and 943 bp while the mean 5' and 3' intergenic distances for effectors were 4,263 bp and 4,002 bp respectively. Our intergenic based analysis clearly showed the two-speed genome architecture on phased consensus haplotypes. Figure 3.16 clearly represents the two-speed genome architecture of *P. ramorum* consensus haplotypes.

Figure 3.16: Two speed genome bi-partite architecture of haplotypes of *P. ramorum* ND886



C. Draft genome sequence for the tree pathogen *Phytophthora plurivora*

3.3.1. Introduction

Phytophthora species are known to affect variety of plants and animal species. *Phytophthora* species with a narrow host rage, still it is potential to affect variety of plant species (Erwin & Ribeiro, 1996). In this study we have sequenced soil born root rot pathogen *P. plurivora* which affects larges variety of woody plants such as *Quercus sp*., *Acer sp*., *Alnus sp*., *Vaccinium sp*., *Rhododendron sp*., and *Fagus sp.* Common symptoms of *P. plurivora* disease include collar rots; bark cankers, extensive damage to fine roots, and crown dieback on young and mature trees (Jung et al 2000; Orlikowski et al 2011). Stem inoculations to test the relative susceptibility of conifers and broadleaved tree species common in Sweden demonstrated *P. plurivora* to be highly aggressive on pedunculate oak (*Quercus robur* L.), European beech and black cottonwood (*Populus trichocarpa*), highlighting the overall risk of different *Phytophthora* species to forest trees (Cleary et al 2017). The most recent phylogeny of the genus *Phytophthora* places *P. plurivora* in Clade 2, subclade 2c, together with related pathogens *Phytophthora acerina*, *Phytophthora pachyleura*, *Phytophthora capensis*, *Phytophthora pini*, *Phytophthora multivora* and *Phytophthora citricola* (Yang et al 2017). *P. plurivora* is proposed to be most likely native to Europe based on haplotype and microsatellite data analysis (Schoebel et al 2014). It is now distributed worldwide, aided by dissemination of diseased plant material through the plant nursery trade (Schoebel et al. 2014). In southern Sweden, *P. plurivora*, along with *P. cactorum*, has been recognized as an increasing threat to cultivated plantation forests (Cleary et al. 2017). To understand broad host range pathogen *P. plurivora* this affects larger variety of woody plants in Europe. We sequenced the P. plurivora genome with illumina sequencing which is native to Europe isolated from Sweden. We assembled the genome using SPADES assembler. From assembled genome we identified effector genes (RXLR and CRN) which were responsible for causing infection when infecting host. We found that effectors are largely present in the gene-sparse and repeat-rich highly evolving region.

3.3.2. Materials and Methods

3.3.2.1. Genome sequencing

P. plurivora AV1007 was isolated from *Fagus sylvatica* trees from Malmö, Sweden. Illumina Hiseq paired-end sequencing was used to sequence the genome. Thirty seven million reads from each library with average read length of 200 bp were generated.

3.3.2.2. Genome assembly and assesments of *P. plurivora*

A De-bruijn graph based assembly was used for handling sequenced illumina reads. Spades assembler (Bankevich et al., 2012) was used to assemble the genome. The assembled genome was assesed for the contiguity of contigs with all other genomes which are sequenced from the same clade with the similar genome size using QUAST(Gurevich et al., 2013).

For checking the completeness of genome assembly, core genes were assesed using BUSCO 2.0(Simão et al., 2015) using stramenopile COGs. The BUSCO comparisons was also done with the all other complete genomes of *P. ramorum*, *P. infestans*.

3.3.3. Downstream genome analysis

Gene prediction was done using AUGUSTUS (Stanke & Waack, 2003) as initial training material of *P. capsici* gene as model. First round predicted sequences were retrained again to get the accurate gene models.

Protein sequences were annotated for Pfam domains using interproscan (Jones et al., 2014) searches. Repetative sequences from the genome assembly was annotated using RepeatMasker (Smit et al., 2016) from the consensus file generated from the Repeat Modeler program.

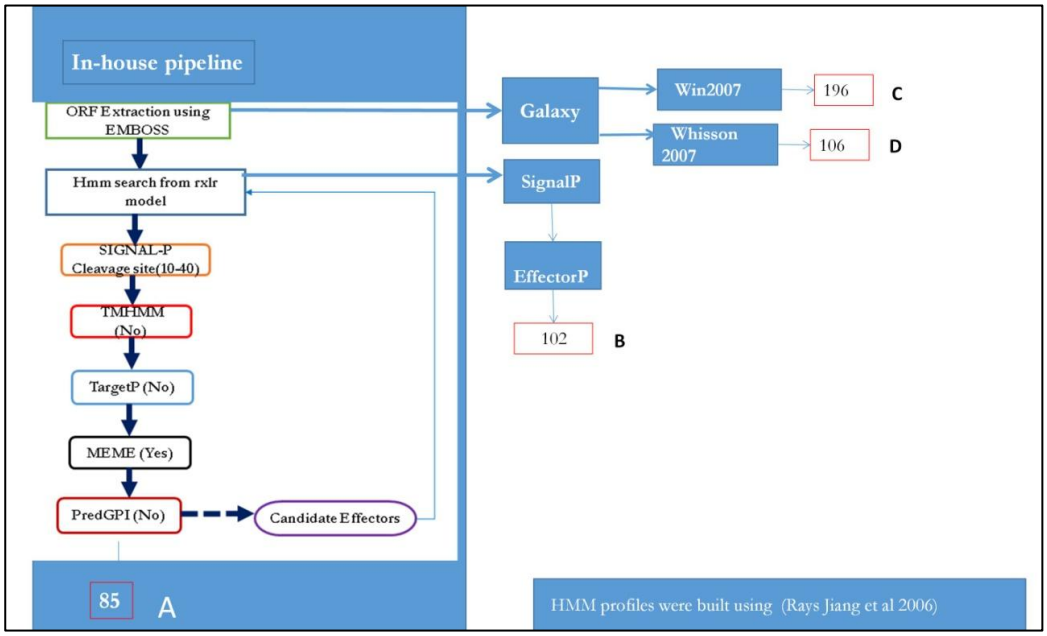
dbCAN (Yin et al., 2012) was used to annotate the carbohydrate active enzyme against the predicted protein sequences. The Pathogen Host interaction (PHI) database (Winnenburg et al., 2006) was used to identify the virulence associated proteins from the annotated proteins.

The proteins associated with secondary metabolites were annotated using ANTISMASH 3.0 webserver (Weber et al., 2015). To estimate the ploidy level in the genome ploidyNGS (Augusto Corrêa dos Santos et al., 2017) was used in the illumina genome sequences.

3.3.3.1. RXLR prediction

The effector prediction was carried out using Four different pipelines. The first pipeline was an in-house based pipeline discussed in above Figure 3.3, second prediction was using EffectorP 1.0 (Sperschneider et al., 2016), third prediction method was based on Galaxy based pipeline (Cock et al., 2013) using Whisson et al 2007 and Win et al 2007 based method. By comparing all other methods, our stringent in-house method was used as a final pipeline for predicting RXLRs from genomes of *P. plurivora*, *P. multivora* isolates, *P. capsici*. The prediction pipeline used for identifying RXLRs are depicted in Figure 3.17 .

Figure 3.17: RXLR prediction pipeline used for identifying candidate effectors from *P. plurivora* genome



3.3.3.2. dN/dS calculation

For our comparative study we clustered the RXLR genes of *P. plurivora*, *P. multivora* isolates, *P. capsici*, *P. cinnamomi* and *P. ramorum* genomes. Orthomcl (Li et al., 2003) was used to cluster the RXLR genes. The obtained ortholog groups with >= 3 members were used for our calculations. CODEML from the PAML (Yang, 2007) was used to calculate the dN/dS ratios from the clusters.

3.3.3.3. Genome architecture and synteny analysis

The 5' and 3' intergenic distance for effectors, genes, BUSCOs were calculated using the method described in Saunders et al 2014. The distances were two dimensionally binned and plotted using various R packages. NUCMER program from MUMMER 3.0 (Kurtz et al., 2004) was used with the cutoff value of maxgap 50 and breaklen 400 was used for comparing the genomes of *P. multivora* and *P. capsici.*

3.3.4. Results and Discussion

3.3.4.1. Genome assembly and assesments of *P. plurivora*

Illumina sequenced *P. plurivora* was assembled using SPADES 3.5.0 9 9(Nurk et al., 2013) to obtain the assembled genome of 41 Mb. There were 1,919 contigs with 1,898 scaffolds with mean coverage of 220X and N50 value of 48,620 bp was the final draft assembly.only 9% of raw reads were not mapping with the assembled genome of *P. plurivora*. Contigs below 2 kb were removed from the assembly, and mitochondrial genome sequences were not screened out; 9 % of sequence reads were unassembled and were discarded.

The assembled genome was quality assesed using QUAST(Gurevich et al., 2013). The assembled genome was compared with other closer genomes of *P. multivora* (two isolates), *P. kernoviae*, and *P. agathidicida* with similar genome sizes shown in Figure 3.18.

Figure 3.18: Genome assembly assesments of *P. plurivora* and other close relatives *Phytopthora's*

Worst Median Best ☑ Show heatmap

Statistics without reference	mul_3378	mul_3448	Phyag	Phyke	plurivora
# contigs	1984	2022	2947	1760	1890
# contigs (>= 0 bp)	2844	2840	3754	7254	1898
# contigs (>= 1000 bp)	2838	2830	3744	2487	1898
# contigs (>= 5000 bp)	1264	1258	1929	1122	1292
# contigs (>= 10000 bp)	838	857	1205	774	921
# contigs (>= 25000 bp)	455	454	376	424	489
# contigs (>= 50000 bp)	220	214	61	189	230
Largest contig	267 541	376 293	131 575	280 632	294 496
Total length	38 825 098	39 146 202	36 186 986	36 145 732	40 433 131
Total length (>= 0 bp)	40 059 192	40 326 536	37 337 699	38 770 482	40 448 196
Total length (>= 1000 bp)	40 054 549	40 320 039	37 331 420	37 183 117	40 448 196
Total length (>= 5000 bp)	36 558 043	36 747 593	32 881 587	34 128 016	38 530 886
Total length (>= 10000 bp)	33 499 376	33 906 063	27 700 357	31 623 084	35 872 391
Total length (>= 25000 bp)	27 300 881	27 475 743	14 509 263	25 929 914	28 999 117
Total length (>= 50000 bp)	18 809 814	18 818 374	4 005 065	17 426 655	19 638 251
N50	48 917	47 383	20 236	48 103	48 620
N75	20 083	19 971	10 593	21 315	21 603
L50	233	230	535	203	242
L75	536	538	1151	475	547
GC (%)	51.95	51.95	52.59	50.36	51.74
Mismatches					
# N's	675	774	925	301	6744
# N's per 100 kbp	1.74	1.98	2.56	0.83	16.68

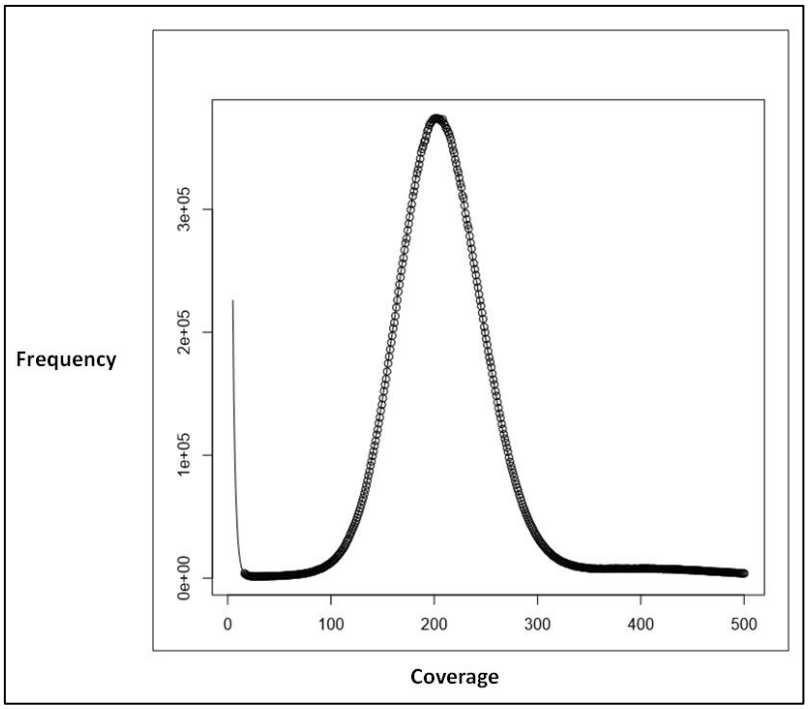
*mul_3378: *P. mutivora* 3378; mul_3348: *P. multivora* 3348; Phyag: *P. agathicida*; Phyke: *P. Kernoviae* , plurivora: *P. plurivora*

In comparisons of all other genome assemblies our assembled genome was having better assembly with the longest 1890 contigs and more contiguous assembly as well. Assessment of assembly quality revealed: N50 = 48620 bp; N75 = 21603 bp; L50 = 242; L75 = 547; longest contig = 294496 bp; number of contigs >25 kb = 489; number of contigs >10 kb = 921).

To assure the estimated genome size, illumina sequences were checked for the K-mer frequency using jellyfish(Liu et al., 2013). The genome size was estimated to be 45 Mb. Maybe the repetative content in the genome were missed in the assembled genome of *P. plurivora*. This also suggests that our genome assembly for P. plurivora spans over 90% of the genome size estimate, and that repetitive sequences are not as prevalent as in some other Phytophthora genomes, especially *P. infestans* (Tyler

et al. 2006; Haas et al. 2009).The k-mer plot for the estimated genome size is represented in Figure 3.19.

Figure 3.19: K-mer frequency plot for the genome size estimation of *P. Plurivora*



The core gene assesment was done using stramenophile BUSCO datasets for the genomes of *P. plurivora* , *P. capsici*, *P. multivora* isolates. while comparing we found 16 missing BUSCOs in *P. capsici*, 10 and 12 BUSCOs were missing in the two isolates of *P. multivora* NZFS 3378 and NZFS 3448, respectively, four missing BUSCOs in *P. ramorum*, and nine missing BUSCOs in *P. infestans.* The BUSCO analysis suggests that our genome assembly is highly representative of the gene space in *P. plurivora*, and compares favourably to other Phytophthora genome assemblies.

3.3.4.2. Genome Analysis and annotation

From the predicted protein sequences about 2.6% proteome had matches with the host pathogen interaction database. CAZY analysis on *P. plurivora* identified glycoside hydrolases (332), glycosyltransferases (271), carbohydrate binding modules (304), polysaccharide lyases (49), and carbohydrate esterases (43). The number of polysaccharide lyases predicted in *P. plurivora* is similar to that found in other Phytophthora species, but our primary analysis here shows elevated numbers of other CAZy proteins in *P. plurivora*, compared to other Phytophthora species (Ospina-Giraldo et al. 2010). Homologs were found in the PHI database for 2.6% of the total predicted *P. plurivora* proteome (308/11749), predominantly with *Phytophthora sojae* and *Fusarium graminearum*.

Phytophthora genomes often contain high levels of repetitive DNA sequences, such as *P. infestans* for which the genome contains over 75% repetitive sequences (Haas et al. 2009; Tyler et al. 2006). Approximately 15% of the *P. plurivora* genome is comprised of repetitive sequences, far less than

many other Phytophthora genomes sequenced to date. The predominant repeat type is interspersed repeats, accounting for 50% of the total repeats. The detailed report on the repeat composition in the genome is available in Table 3.6.

Table 3.6: Classification of Repetative elements in the assembled genome of *P. plurivora*

Repeats	**Number**	**Length occupied**	**Percentage of sequence**
SINEs:	52	6478	0.02%
ALUs	0	0	0.00%
MIRS	0	0	0.00%
LINEs:	181	87398	0.22%
LINE1	114	43446	0.11%
LINE2	0	0	0.00%
L3/CR1	25	20942	0.05%
LTR elements:	1543	945917	2.34%
ERVL	0	0	0.00%
ERVL-MaLRs	0	0	0.00%
ERVL-class I	0	0	0.00%
ERVL-class II	0	0	0.00%
DNA elements:	2430	1115747	2.76%
hAT-charlie	0	0	0.00%
TcMar-Tigger	3	887	0.00%
Unclassified	1427	826032	2.04%
Total interspersed repeats:		2981572	7.37%
Satellites:	0	0	0.00%
Simple repeats	4134	186013	0.46%
Low complexity:	500	26340	0.07%

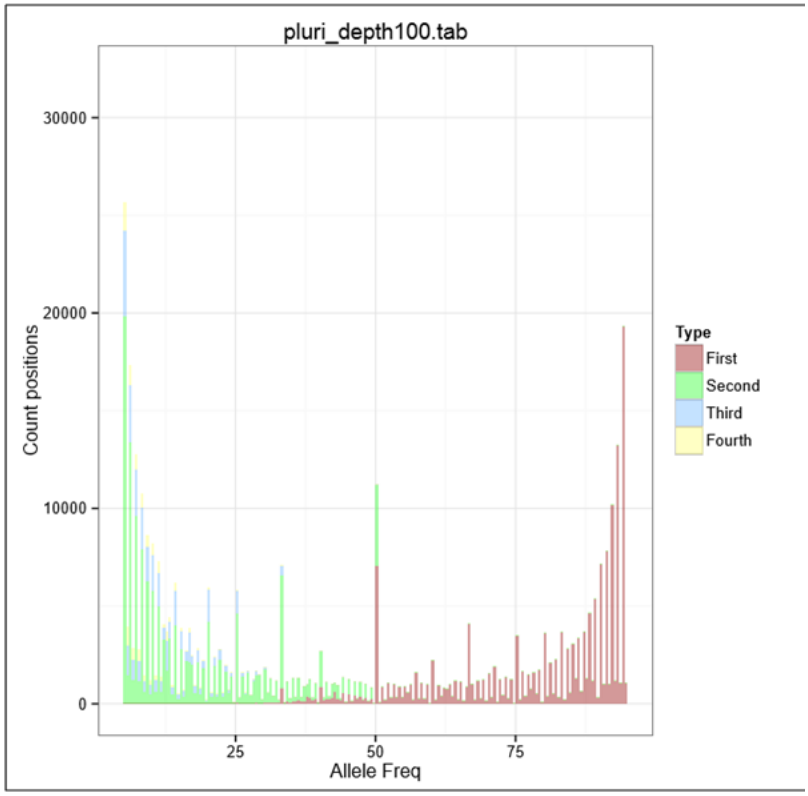
Our analysis on secondary metabolite compounds identified 56 genes encoding the non-ribosomal peptide synthase, and enzymes involved in ectoine and terpene biosynthesis. The details of prediction results are presented in Table 3.7. Compared to fungi, species of Phytophthora are not known to produce many secondary metabolites. The best characterized secondary metabolites from Phytophthora are the mating hormones α1 and α2, which are diterpene molecules (Tomura et al. 2017). It has also been shown that Phytophthora species produce a signal molecule derived from 4,5-dihydroxy-2,3-pentanedione which has quorum-sensing activity in bacteria (Kong et al. 2010). It is possible that the secondary metabolism predictions from the *P. plurivora* genome may be involved in the synthesis of these bioactive secondary metabolites.

Table 3.7: Identified secondary metabolite gene clusters from the *P. plurivora* genome

Cluster No	**Type**	**location**
1	other	Scaffold_109: 37497-72587
2	Nrps	Scaffold_243:1-27126
3	Terpene	Scaffold_271:11958-33151
4	Other	Scaffold_491:1-24842
5	other	Scaffold_556:1-21340
6	Ectoine	Scaffold_576:2257-13544

Our ploidy analysis based on Kolmogorov-Smirnov distance using ploidyNGS identified *P. plurivora* as homothallic (self-fertile, inbreeding) and tetraploid. Figure 3.20 represents the ploidy distribution in assembled *P. plurivora* genome. *P. plurivora* is a homothallic (self-fertile, inbreeding) species, signifying that if its survival in the environment is via sexually derived oospores, then heterozygosity levels will be reduced with each generation, as has been observed in *P. plurivora* strains sampled from different countries (Schoebel et al. 2014).

Figure 3.20: Representation of allele frequency based on the Kolmogorov-Smirnov distance on *P. plurivora* genome



3.3.4.3. Effector prediction

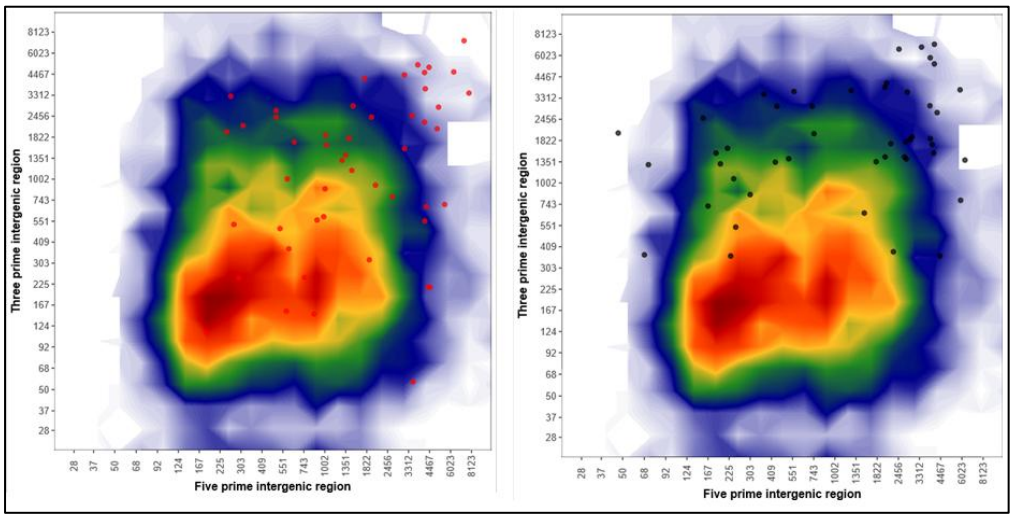
Our in-house prediction pipeline, predicted 84 candidate RXLRs, out of which 80 proteins contain signal peptide, RXLR and EER motifs. These high confidence set of predicted RXLRs were used for further analysis. Crinklers (CRN) were predicted using the method which was described in (Yin et al., 2017) using CRN sequences from *P. infestans* as training material (Haas et al., 2009). We predicted 139 CRN proteins, of which 60 had signal peptides. These effectors are modular proteins that contain an N-terminal signal peptide, a conserved RXLR peptide motif typically within the next 40 amino acids, and often an EER motif near the RXLR. The functional effector peptide region is located between the RXLR-EER and the C-terminus. The majority of effectors in this class that have

been functionally characterized contain all three of these features (Anderson et al. 2015; Whisson et al. 2016).

3.3.4.4. Genome architecture studies

The genome organization was studied for *P. plurivora* genome based on the intergenic distances between the genes. The genome architecture was found to be "two-speed " genome architecture. Figure 3.21 represents the genome organization.

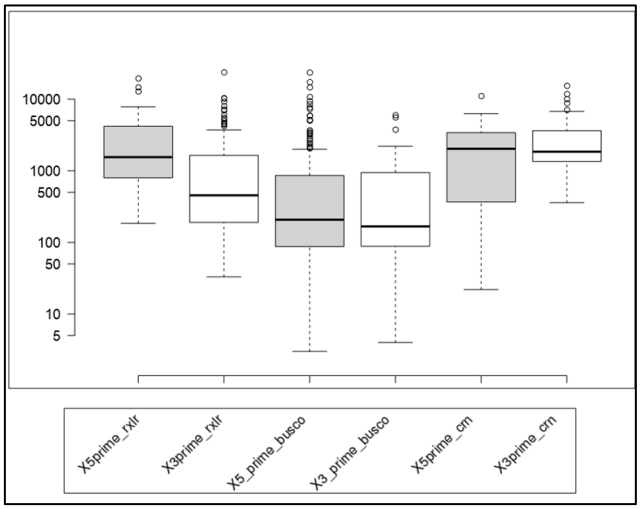
Figure 3.21: *P. plurivora* representing the two speed genome architecture in the genome. Left panel showing predicted RXLRs are found in the gene sparse fast evolving region, the red circles represents RXLRs. Right panel shows predicted CRNs are also found in the fastly evolving region of genome. The black circles represents the predicted CRNs from the genome.



The mean intergenic distance between all *P. plurivora* genes at the 5' was 1107 bp and the 3' mean intergenic distance for the entire predicted gene set was 848 bp. The 5' intergenic distance was 3117 bp and 3' distance was 2128 bp for RXLR coding genes. For CRNS 5' intergenic distance was 2018 bp and 3' intergenic distance was 3087 bp. The median intergenic distance between all predicted genes was significantly less than that for the effector genes classes analyzed. The median 5' intergenic distance between all predicted genes was 582 bp, whereas it was 1920 bp and 1518 bp for the RXLR and CRN effector genes, respectively. Similarly, the median 3' intergenic distance between all the genes was 316 bp, whereas it was 1414 bp and 1925 bp for RXLR and CRN effectors, respectively. Comparison of RXLR class effectors between Phytophthora species from different clades has typically revealed that these effectors have diverse sequences, with many having no homologs in other species (Quinn et al. 2013; McGowan and Fitzpatrick 2017), and thus are evolving rapidly.

The t-test on flanking intergenic distances of effectors and core genes shows that intergenic distances are significantly different from the core and effector genes. The box plot showing significant differences in the intergenic distance are plotted in Figure 3.22.

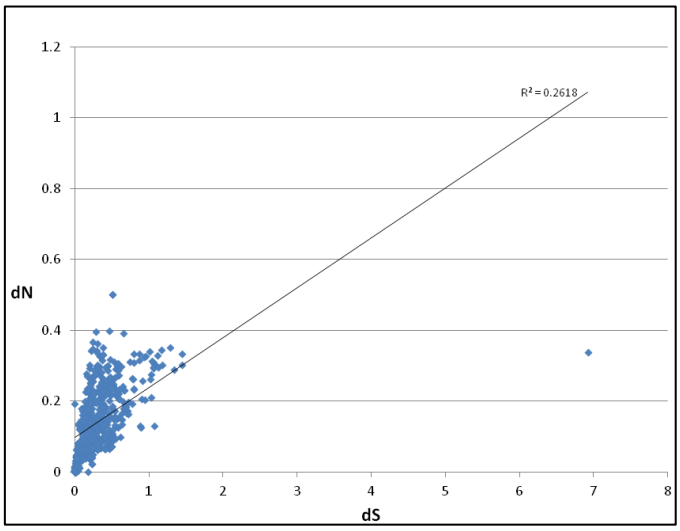
Figure 3.22: *P. plurivora* showing differences in the intergenic distance of effectors (RXLR and CRN) from core BUSCO gene sets



3.3.4.5. dN/dS analysis of RXLRs reveals gene sets undergoing positive selection

The six genomes of broad host range pathogens were chosen for studying the selection pressure. *P. plurivora* RXLRs, along predicted RXLRs from two isolates of *P. multivora*, *P. capsici*, *P. cinnamomi* and *P. ramorum* were chosen for the analysis. Orthomcl clustred 105 gene clusters suitable for the analysis. The higher dN/dS ratios ranged from 1.0 to 3.6 and included five effectors from *P. plurivora*, but only PlRXLR53 exhibited a markedly elevated dN/dS ratio of 1.9 that suggested positive selection. Figure 3.23 shows the corrleation value of dN and dS values from the prediced RXLRs. Taken together, these results suggest that *P. plurivora* possesses RXLR effectors that are under diverse evolutionary pressures, with one subset showing evidence of purifying selection, and a further subset that have evolved rapidly and are specific only to *P. plurivora*. Sequencing of additional Clade 2 species, and species that are more closely related to *P. plurivora*, may provide more resolution in clarifying the mode of selection acting on this class of effectors.

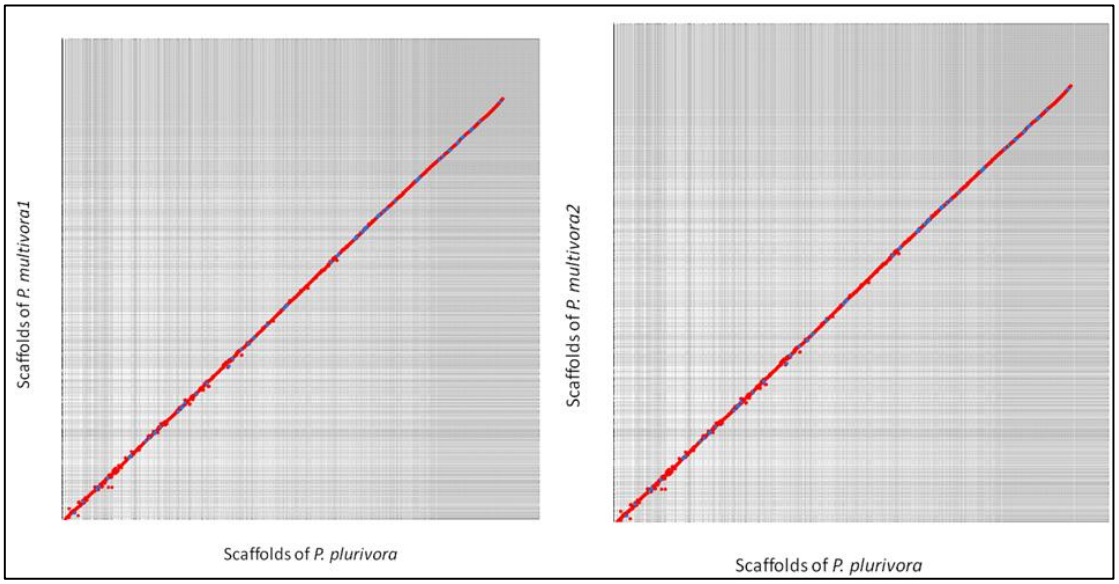
Figure 3.23: Regression plot of dN and dS values for *P. plurivora* RXLR genes



3.3.4.6. Synteny analysis of *P. plurivora* and its closest genome of *P. multivora* isolates

P. plurivora genome was compared with isolates of *P. multivora* NZFS3378 and NZFS3348 from the same phylogeny clade 2 of Pythopthora. More similarity was observed between the genomes. There were 2158 and 1472 *P. plurivora* genes without orthologs in *P. multivora* isolates. Genome synteny was studied between *P. plurivora* and *P. capsici*, and both isolates of *P. multivora*. The largest scaffold (scaffold_1 length of 294496 bp) with 95 protein coding genes was almost fully syntenic with both *P. multivora* isolates, except for three missing genes: g65, g72 and g95. These three genes are highly conserved in the two *P. multivora* isolates and in other *Phytophthora sp* (Figure 3.24).

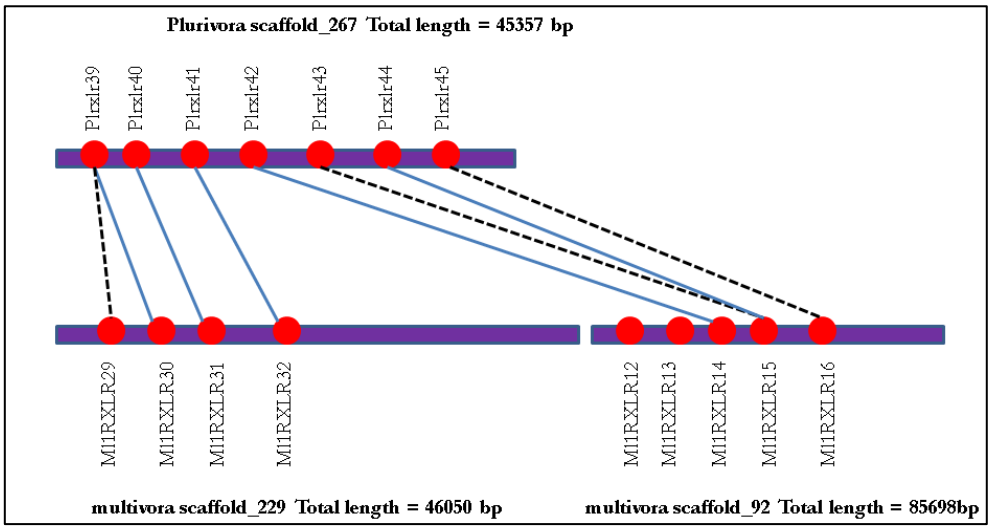
Figure 3.24: Synteny among the isolates of P. *Multivora* and *P. plurivora* exhibiting significant similarity



We also observed there were a large cluster of *P. plurivora* exhibiting colinearity. Seven RXLRs from *P. plurivora* Scaffold_267 was identical with the scaffolds (LGSM010000246.1 and LGSM01000099.1) and (LGSL01000255.1 and LGSL01000075.1) of both the isolates of *P. multivora*. The gene duplications was observed in the assembled genome of *P. plurivora*. Two RXLRs of P. *Multivora* isolate MlRXLR15 and MlRXLR16 were observed with ancient duplication (Figure 3.25).

MlRXLR29 and MlRXLR30 are derived from duplication of PlRXLR39 from *P. plurivora.*

Figure 3.25: Syntenic organization of RXLRs from *P. plurivora* and *P. multivora* genome a closest relative



3.3.5. Summary

To summarise, this chapter is divided into 3a, 3b and 3c. The 3a explains the importance of optimization of long read genome assembly of *P. ramorum* Pr102 using PacBio P5-C3 older chemistry. As well as the handling of errors from reads using illumina sequencing and reducing indel error rate is described in detail. In comparison with V1 assembly the gaps in the genome are closed.

Chapter 3b explains the importance of haplotye phased diloid genome assembly of *P. ramorum* ND886 with advanced P6-C4 chemistry. Importance of genetic haplotype phasing and challenges of pacbio read errors are described. *P. ramorum* ND886 genome contains more tandem repeats and repeative elements were identified in the genome.

Chapter 3c explains the tree pathogen genome which was sequenced for the first time using illumina sequencing. We compared the RXLR effectors from different clade members. We found this pathogen is a tetraploid .

List of Publications :

1. Draft genome sequence for the tree pathogen *Phytophthora plurivora.*
   Ramesh R. Vetukuri [1] *, Sucheta Tripathy[2,6*†,] Mathu Malar C*[† 2,6] , Arijit Panda[2,6] , Sandeep K. Kushwaha[3,4], Aakash Chawade [3] , Erik Andreasson[1] , Laura J. Grenville-Briggs [1 3] , Stephen C. Whisson[5] (Minor Revision in Genome Biology and evolution, *joint first author) .

2. Characterization of phenotypic variation and genome aberrations observed among Phytophthora ramorum isolates from diverse hosts.
   Marianne Elliott[1] , Jennifer Yuzon[2] , Mathu Malar C[3] , Sucheta Tripathy[3] , Mai Bui[4] , Gary A. Chastagner[1] , Katie Coats[1] , David M. Rizzo[2] , Matteo Garbelotto[5] and Takao Kasuga[4*] [Published in BMC genomics, 2018]

3. Haplotype-phased genome assembly of virulent *Phythophthora ramorum* isolate ND886 facilitated by long-read sequencing suggests effector polymorphism.
   Mathu Malar C [1,5#], Jennifer Yuzon [2,3#], Subhadeep Das [1,5] ,Abhishek Das [1,5], Samrat Ghosh [1,5], Arijit Panda [1,5], Brett M. Tyler*[4], Takao Kasuga[2,3*], Sucheta Tripathy[1,5*] [Ready to be submitted in BMC Biology,# joint first author]

4. Genome sequencing and reanalysis of Sudden Oak Death pathogen *Phytophthora ramorum* Pr102 using P5-C3 chemistry PacBio sequencing reads.
   Mathu Malar C [1,5#], Jennifer Yuzon [2,3#], Takao Kasuga[2,3*], Sucheta Tripathy[1,5*][For submission in Scientific reports, #joint first author]

References

Anderson RG, Deb D, Fedkenheuer K, McDowell JM 2015. Recent progress in RXLR effector research. Molecular Plant-Microbe Interactions 28: 1063-1072. doi: 10.1094/MPMI-01-15-0022-CR

Au KF, Underwood JG, Lee L, Wong WH, 2012. Improving PacBio long read accuracy by short read alignment. *PLoS One* **7**, e46679.

Augusto Corrêa Dos Santos R, Goldman GH, Riaño-Pachón DM, 2017. ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics* **33**, 2575-6.

Bailey TL, Boden M, Buske FA, *et al.*, 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202-W8.

Bankevich A, Nurk S, Antipov D, *et al.*, 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* **19**, 455-77.

Bendtsen JD, Nielsen H, Von Heijne G, Brunak S, 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology* **340**, 783-95.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W, 2010. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9.

Boetzer M, Pirovano W, 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics* **15**, 211.

Brouwer H, Coutinho PM, Henrissat B, De Vries RP, 2014. Carbohydrate-related enzymes of important Phytophthora plant pathogens. *Fungal Genetics and Biology* **72**, 192-200.

Browning SR, Browning BL, 2011. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**, 703.

Butler J, Maccallum I, Kleber M, *et al.*, 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* **18**, 810-20.

Chaisson MJ, Tesler G, 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics* **13**, 238.

Chin CS, Peluso P, Sedlazeck FJ, *et al.*, 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050-4.

Cingolani P, Platts A, Wang LL, *et al.*, 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92.

Cock PJ, Grüning BA, Paszkiewicz K, Pritchard L, 2013. Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. *PeerJ* **1**, e167.

Dong S, Raffaele S, Kamoun S, 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Current opinion in genetics & development* **35**, 57-65.

Emanuelsson O, Brunak S, Von Heijne G, Nielsen H, 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols* **2**, 953-71.

Erwin DC, Ribeiro OK, 1996. *Phytophthora diseases worldwide*. American Phytopathological Society (APS Press).

Garg S, Martin M, Marschall T, 2016. Read-based phasing of related individuals. *Bioinformatics* **32**, i234-i42.

Giardine B, Riemer C, Hardison RC, *et al.*, 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-5.

Gruenwald NJ, Goss EM, Press CM, 2008. Phytophthora ramorum: a pathogen with a remarkably wide host range causing sudden oak death on oaks and ramorum blight on woody ornamentals. *Molecular Plant Pathology* **9**, 729-40.

Gurevich A, Saveliev V, Vyahhi N, Tesler G, 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-5.

Haas B, 2007. TransposonPSI: an application of PSI-blast to mine (Retro-) transposon ORF homologies. In.

Haas BJ, Kamoun S, Zody MC, *et al.*, 2009. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature* **461**, 393-8.

Huang X, Madan A, 1999. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-77.

Ivors KL, Hayden KJ, Bonants PJM, Rizzo DM, Garbelotto M, 2004. AFLP and phylogenetic analyses of North American and European populations of *Phytophthora ramorum*. *Mycological Research* **108**, 378-92.

Jiang RH, Tyler BM, 2012. Mechanisms and evolution of virulence in oomycetes. *Annual review of phytopathology* **50**, 295-318.

Jones P, Binns D, Chang H-Y, *et al.*, 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-40.

Jung T, et al. 2017. Six new Phytophthora species from ITS Clade 7a including two sexually functional heterothallic hybrid species detected in natural ecosystems in Taiwan. Persoonia 38: 100-135. doi: 10.3767/003158517X693615

Käll L, Krogh A, Sonnhammer EL, 2004. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology* **338**, 1027-36.

Keller O, Odronitz F, Stanke M, Kollmar M, Waack S, 2008. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC bioinformatics* **9**, 278.

Koren S, Schatz MC, Walenz BP, *et al.*, 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* **30**, 693-700.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM, 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-36.

Kong P, Lee BWK, Zhou ZS, Hong C 2010. Zoosporic plant pathogens produce bacterial autoinducer-2 that affects Vibrio harveyi quorum sensing. FEMS Microbiology Letters 303: 55-60. doi: 10.1111/j.1574-6968.2009.01861.x

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL, 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567-80.

Kurtz S, Phillippy A, Delcher AL, *et al.*, 2004. Versatile and open software for comparing large genomes. *Genome biology* **5**, R12.

Li H, 2011a. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-93.

Li H, 2011b. wgsim-Read simulator for next generation sequencing. *Github Repository*.

Li H, 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li H, Handsaker B, Wysoker A, *et al.*, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-9.

Li L, Stoeckert CJ, Roos DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-89.

Liu B, Shi Y, Yuan J, *et al.*, 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint arXiv:1308.2012*.

Nurk S, Bankevich A, Antipov D, *et al.*, 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of computational biology* **20**, 714-37.

McGowan J, Fitzpatrick DA 2017. Genomic, network, and phylogenetic analysis of the oomycete effector arsenal. mSphere 2. doi: 10.1128/mSphere.00408-17

Orlikowski LB, et al. 2011. Phytophthora root and collar rot of mature Fraxinus excelsior in forest stands in Poland and Denmark. Forest Pathology 41: 510-519. doi: 10.1111/j.1439-0329.2011.00714.x

Ospina-Giraldo MD, Griffith JG, Laird EW, Mingora C 2010. The CAZyome of Phytophthora spp.: A comprehensive analysis of the gene complement coding for carbohydrate-active enzymes in species of the genus Phytophthora. BMC Genomics 11: 525. doi: 10.1186/1471-2164-11-525

Parra G, Bradnam K, Korf I, 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-7.

Peltola H, Söderlund H, Ukkonen E, 1984. SEQAID: A DNA sequence assembling program based on a mathematical model.

Pierleoni A, Martelli PL, Casadio R, 2008. PredGPI: a GPI-anchor predictor. *BMC bioinformatics* **9**, 392.

Pryszcz LP, Gabaldón T, 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic acids research* **44**, e113-e.

Quinn L, et al. 2013. Genome-wide sequencing of Phytophthora lateralis reveals genetic variation among isolates from Lawson cypress (Chamaecyparis lawsoniana) in Northern Ireland. FEMS Microbiology Letters 344: 179-185. doi: 10.1111/1574-6968.12179

Saunders DG, Win J, Kamoun S, Raffaele S, 2014. Two-dimensional data binning for the analysis of genome architecture in filamentous plant pathogens and other eukaryotes. *Plant-pathogen interactions: methods and protocols*, 29-51.

Schoebel CN, Stewart J, Gruenwald NJ, Rigling D, Prospero S 2014. Population history and pathways of spread of the plant pathogen Phytophthora plurivora. PLOS ONE 9: e85368. doi: 10.1371/journal.pone.0085368

Schornack S, Van Damme M, Bozkurt TO, *et al.*, 2010. Ancient class of translocated oomycete effectors targets the host nucleus. *Proc Natl Acad Sci U S A* **107**, 17421-6.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM, 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-2.

Smit A, Hubley R, Green P, 2016. RepeatMasker Open-4.0. 2015. *Google Scholar*.

Sommer DD, Delcher AL, Salzberg SL, Pop M, 2007. Minimus: a fast, lightweight genome assembler. *BMC bioinformatics* **8**, 64.

Sperschneider J, Gardiner DM, Dodds PN, *et al.*, 2016. EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist* **210**, 743-61.

Stam R, Jupe J, Howden AJ, *et al.*, 2013. Identification and characterisation CRN effectors in Phytophthora capsici shows modularity and functional diversity. *PLoS One* **8**, e59517.

Stanke M, Waack S, 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215-ii25.

Tyler BM, Tripathy S, Zhang X, *et al.*, 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261-6.

Tomura T, Molli SD, Murata R, Ojika M 2017. Universality of the Phytophthora mating hormones and diversity of their production profile. Scientific Reports 7: 5007. doi: 10.1038/s41598-017-05380-3.

Walker BJ, Abeel T, Shea T, *et al.*, 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963.

Weber T, Blin K, Duddela S, *et al.*, 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research* **43**, W237-W43.

Whisson SC, Boevink PC, Moleleki L, *et al.*, 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* **450**, 115.

Win J, Morgan W, Bos J, *et al.*, 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *The Plant Cell* **19**, 2349-69.

Winnenburg R, Baldwin TK, Urban M, Rawlings C, Köhler J, Hammond-Kosack KE, 2006. PHI-base: a new database for pathogen host interactions. *Nucleic acids research* **34**, D459-D64.
Xue W, Li J-T, Zhu Y-P, *et al.*, 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC genomics* **14**, 604.
Yang Z, 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-91.
Yin, L. et al. Characterization of the secretome of Plasmopara viticola by de novo transcriptome analysis. Physiol. Mol. Plant Pathol. 91, 1–10 (2015)
Yin L, An Y, Qu J, *et al.*, 2017. Genome sequence of Plasmopara viticola and insight into the pathogenic mechanism. *Scientific Reports* **7**.
Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y, 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research* **40**, W445-W51.

Chapter 4 – Genome Assembly of Eukaryotic Human Pathogen

4.1. Introduction

Leishmania donovani is an intracellular obligate parasite of mammalian macrophages and causes visceral leishmaniasis or *kala-azar*, which is a fatal disease if not treated on time. The disease severity varies from host, region and between different strains of the same parasite. Extensive research by different groups exploiting genomic, proteomic, metabolomic, immunologic, and animal models have pointed towards tripartite determinants mediating the development of this disease *viz.* the vector, host and pathogen. Although vector and host characteristics are important for symptomatic disease, the fate of the disease is mainly determined by the characteristics of the parasite (McCall et al., 2013). Promastigotes are the culturable form of *Leishmania*. Stationary phase culture of *Leishmania* is expected to contain a large number of metacyclic parasites and have been routinely used for experimental infections. Metacyclic promastigotes are injected into the host during blood meal and are the infective stage of the parasite. Studies at genomic level mostly identified parasite evolution in the Indian subcontinent under drug pressure or disease phenotype (Imamura et al., 2016b)¯(Zhang et al., 2014b). When the organism is cultured in vitro, after several passages, it tends to lose virulence. Such cultures are called as late passage. The early passages are the culture that are still virulent and are isolated from infected hamsters. The genomes in the public repositories are invariably from the late passage. We sequenced the genomes of early and late passages to identify the genomic changes that may be causing loss of virulence.

Abbreviations: HTI4- Early passage genome, HTI5 – Late passage genome, LdBPK282A1- reference *L.donovani* LdBPK282A1 from NCBI

4.2. Materials and Methods

4.2.1. Genome sequencing

High quality genomic DNA was extracted from stationary phase promastigotes of the early (HTI4) and late (HTI5) passage of *L. donovani* AG83. Two separate libraries of *L. donovani* AG83 were prepared one each for early and late passages. *De novo* paired end sequencing was done on Illumina HiSeq 2500 with the read length of 125 base pairs with an insert size of 250 bp. A total of 66 and 44 million raw reads were generated from early and late passages, respectively.

4.2.2. Quality check

Quality control analysis of raw data was done using FastQC (Bioinformatics, 2011). The reads were preprocessed using Trimmomatic (Bolger et al., 2014) and the poor quality and adapter sequences, contaminated sequences were removed using the blast and blat based searches using the reference (LdBPK82A1; Bioproject: PRJNA171503) from Genbank. The high quality filtered reads were used for downstream data analyses.

4.2.3. Genome assembly

The early and late passage genomes were assembled using Allpaths-LG assembler (Butler et al., 2008). The 6k, 20k insert matepair libraries were sheared from the reference genome LdBPK82A1 using Wgsim (Li et al., 2008). The generated 6k, 20k and cleaned reads were used to avail the draft scaffolds. We designed an in-house tool (STLab assembler) for our study to resolve the scaffolds to chromosome on the basis of reference genome.

4.2.4. STLab assembler

The STLab assembler is based entirely on the availability of reference genomes. This tool is with very less dependency, which requires only BEDtools, awk, mummer and Perl. The assembled draft genomes with the scaffolds are aligned using mummer to generate the co-ordinate file containing the scaffolds mapping with the reference chromosome. Length filtering cutoff value of 1500 bp was used to eliminate the poorly aligned regions. By this, we were able to stitch the scaffolds into chromosome. For poorly aligned regions or region with less nucleotide, Ns are assigned to designate gaps. The program used to assemble our genome is available in github https://github.com/madhubioinfo/STLab-assembler . By using this early and late passage genomes were assembled to complete chromosome level.

4.2.5. Gene prediction and Genome annotation

BUSCO (Simão et al., 2015a) was used to assess the genome completeness of early and late passages using the eukaryotic core ortholog COG's.

Protein coding genes were predicted from assembled genomes using Augustus (Stanke et al., 2008) and Scipio with LdBPK282A1 coding sequences as initial training sequences.

Predicted genes were annotated using BLAST (Altschul et al., 2009) as well as InterProScan (Quevillon et al., 2005) searches for Pfam domain identification. Whole genome protein sequences were submitted to the gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis. GO enrichment was analyzed in the categories of molecular function, cellular component and biological process. Metabolic pathways analysis was done using the KASS server (Moriya et al., 2007).

Additionally, we used companion server (Steinbiss et al., 2016) for predicting orthologous genes from early and late passage genomes using *Leishmania major* as the reference. Many pseudogenes were identified in the genomes that were not predicted by Augustus. Genomes of two passages were compared using Mummer (Kurtz et al., 2004a).

4.2.6. SNP detection

Illumina reads from the early and late passage genomes were aligned with the reference genome LdBPK282A1 using bwa (Li & Durbin, 2009). The duplicate reads were marked using PICARD tools and the variants are called using Genome Analysis Toolkit (GATK) program (Van der Auwera et al., 2013). The variant filtration was done using the cutoff value of QD < 2.0 || MQ < 40 || FS > 60.0 || ReadPosRankSum < -8.0.

4.2.7. Chromosome copy number variation

The copy number variant analysis was done using CNVnator (Abyzov et al., 2011). First the filtered reads were mapped to NCBI reference genome using bwa. Read depth coverage was estimated for each of the chromosome using the coverage values from bam files using SAMtools (Li et al., 2008). The window size of 100 bp was used to generate histogram bin values for the read depth values. Filtration cut-off e- values of 1e-5 calculated from the t-test values were used to filter the read depth values.

4.3. Results and discussion

4.3.1. Genome Assembly and annotation

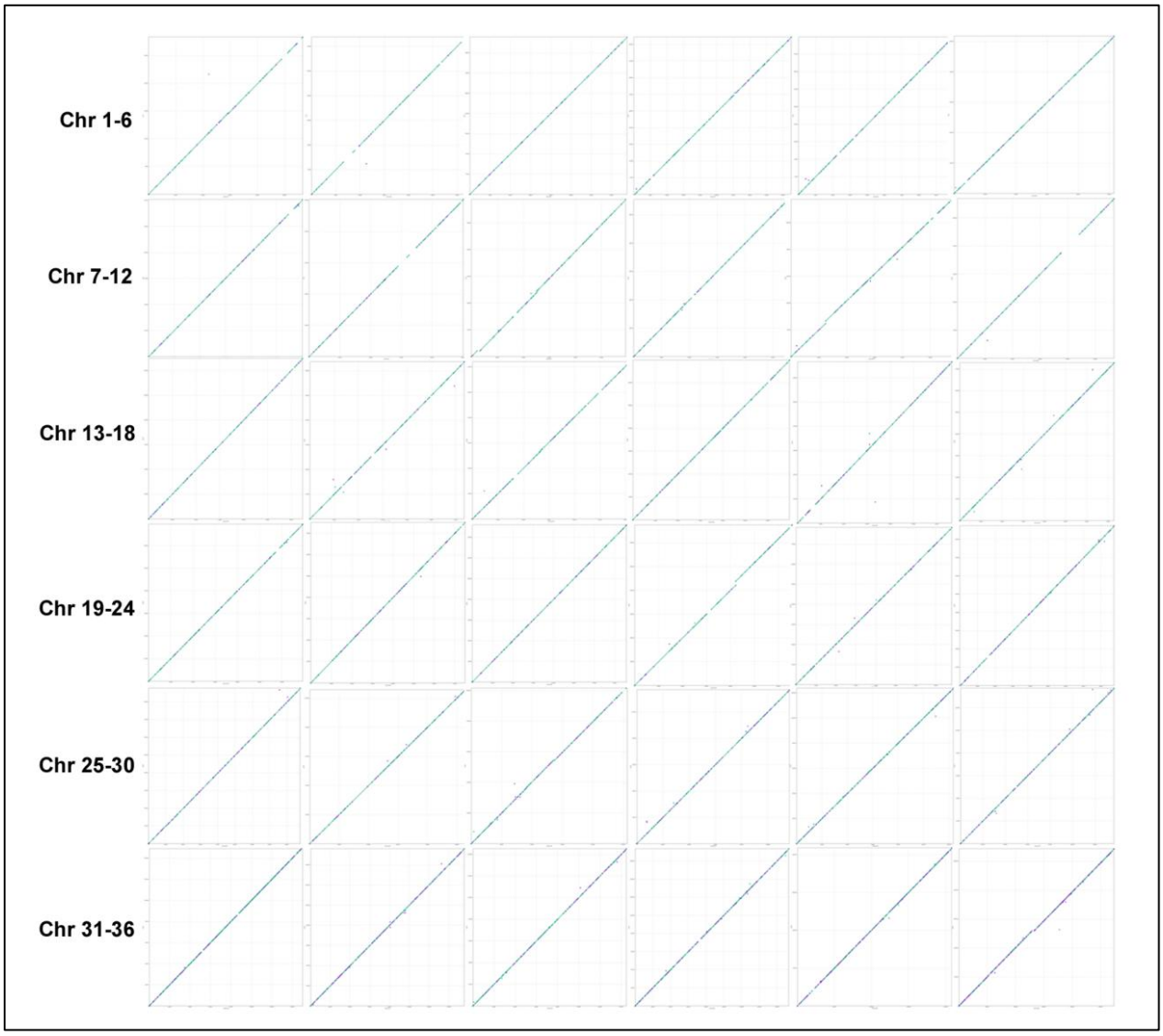
The genomes of early and late passages were assembled into complete 36 chromosomes with a genome size of 32.2 MB and 32.1 MB respectively (Table 4.1).

Table 4.1: Genome assembly statistics of early and late passage genomes of *Leishmania donovani*

Assembly name	**Number of contigs**	**No of chromosomes**	**Total assembly size (in bp)**	**Largest chromosome (in bp)**	**Smallest chromosome (in bp)**	**Gaps (in bp)**	**N50 and GC%**
HTI4 (early passage of *L. donovani*)	2382	36	32196393	2743999	284264	3663498	105808 1, 59%
HTI5 (Late passage of *L . donovani*)	2445	36	32148377	2714535	283355	3653324	105804 3, 58%
LdBPK282A1 (GCA_00022713 5.2)	2152	36	32444968	2713248	283432	1192833	102408 5, 59%

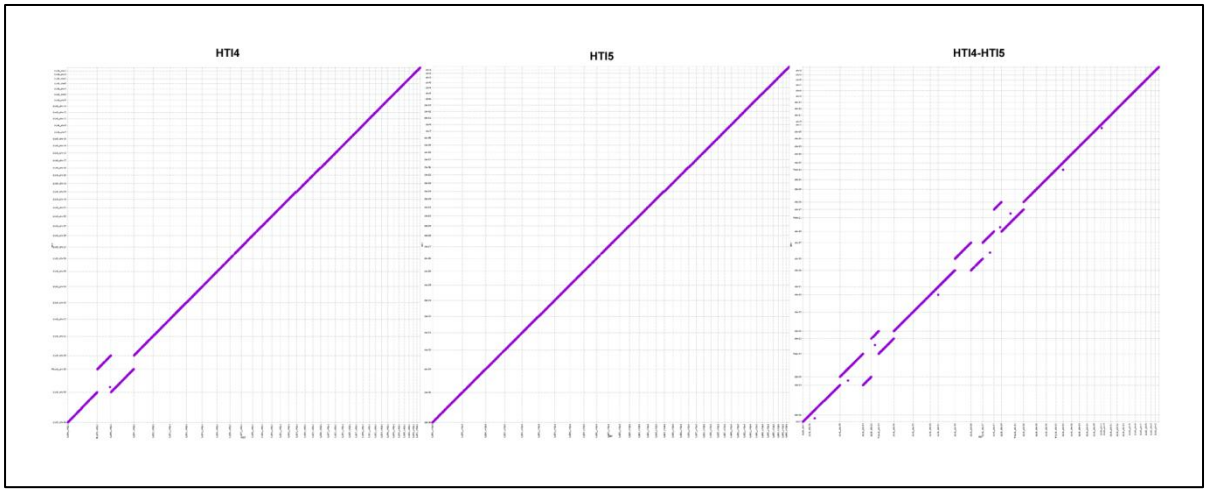
The genomes of these two passages had about 11.3% gaps as compared to 3.67% gaps in reference LdBPK282A1 from Genbank. Chromosome wise comparison of early and late passage genome with reference genome didn't show much difference (Figure 4.1).

Figure 4.1: Chromosome wise comparisons of early, late passages of *L. donovani* to a reference genome of *L. donovani* LdBPK282A1



When the whole genome was compared with reference genome revealed the small prominent changes in the genome of the late passage Figure 4.2.

Figure 4.2: Early and Late passage whole genome comparisons with reference genome of *L.donovani* LdBPK282A1



RATT was used to transfer annotation from LdBPK282A1 strain (Downing et al., 2011) which annotated 7563 and 7552 predicted protein coding genes from genomes of early and late passages. Whereas there were 7967 protein coding genes in NCBI reference genome shown in Table 4.2.

Table 4.2: Statistics of Gene prediction in early and late passage genome of *Leishmania donovani*

Genome name	**Gene**	**Protein coding genes**	**ncRNA**	**Pseudo genes**	**snRNA**	**SnoRNA**	**tRNA**
HTI-4 Early passage (2nd)	7656	7563	14	73	2	1	67
HTI-5 Late passage (25th)	7643	7552	15	82	2	1	66
LdBPK282A1 (GCA_000227135.2)	8079	7967	37	54	0	0	64

We assume that less number of genes from early and late passage genomes in comparison with reference genome is due to presence of gaps in assembly. A comparison of pseudogenes was done on early, late and reference genomes which predicted 73, 82 and 54 pseudogenes respectively. Our comparison on number of pseudogenes with other *L. donovani* suggested that number of pseudogenes were similar to the Srilankan *L. donovani* strains by (Zhang et al., 2014a). Large clusters of calpain-like proteases and amastins are found to occur in large numbers in reference and our assembled genomes. There were large clusters of amastins which were found in more numbers shared among the genomes. The details are shown in Table 4.3.

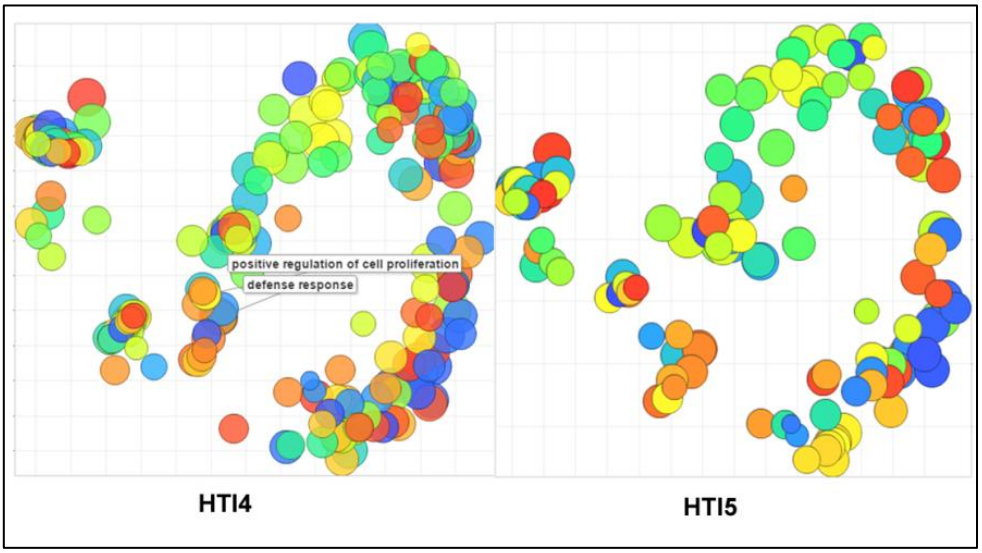
Table 4.3: Clusters of genes in early, late passages and Genbank strain LdBPK282A1 of *L. donovani*

Gene Product	Family Size			Distribution on Chromosomes		
	HTI4	HTI5	LdBPK 282A1	HTI4 (early passage)	HTI5 (late passage)	LDBPK282A1 (reference genome)
Kinesins	51	50	49	Scattered	Scattered*	Scattered*
Protein Kinases	259	258	255	Scattered	Scattered*	Scattered*
MAP Kinases	17	17	19	Scattered*	Scattered*	Scattered
Amastins	14	15	26	10,24,28,30,34,36	10,24,28,30,34,36	8,24,28,29,30,34
PSA2 (GP46) metalloproteases	28	29	29	Scattered*	Scattered	Scattered
Serine peptidases	17	18	13	Scattered	Scattered	Scattered
Protein phosphatises	120	118	86	Scattered	Scattered	Scattered
Tuzins	4	4	6	8, 29*,34	8, 29*,34	8,29,34
Amino acid permeases	15	15	18	Scattered*	Scattered*	Scattered
HSP	11	12	10	Scattered*	Scattered*	Scattered
Calpain-like cysteine peptidase	29	31	26	4,17,18,20(7), 21,25,27(5),31(6), 32,33,36	4,17,18,20(8),21, 25,27(4),31(7),32, 33,34,36	4,14,18,20(8),21, 25,27,30,31(5), 32,33,34
Phosphoglycan β 1,3 galactosyltransferases	3	4	10	2*,14,31*,36*#	2*,14,31*,36*#	2,14,31,36
Dynein heavy and light chain	44	44	44	Scattered	Scattered	Scattered
Helicases	84	84	72	Scattered	Scattered	Scattered
Pteridine transporters	1	1	2	6,10*#	6,10*#	6,10
Microtubule-associated proteins	72	72	72	Scattered	Scattered	Scattered
ABC transporters	39	42	39	Scattered*	Scattered	Scattered*
Vesicle transporters	4	4	4	11,23,31,32	11.23,31,32	11,23,31,32
DNAJ protein/chaperone	61	61	29**	Scattered	Scattered	Scattered**
Long-chain fatty acid CoA ligases	9	9	9	1,3,13,19*,28,36*#	1,3,13,19*,28,36*#	1,3**,13,19*,28, 36*#
Cyclophilins	1	15	13**	1,6,16,18,22,23,24, 25,30,31,33,35,36	1,6,16,18,22,23,24, 25,30,31,33,35,36	1,6,16,22,23,25, 30,31,33,35,36
Histone acetyl transferase/ histone deacetylase	8	8	7**	8,14,16,21,24,26,28	8,14,16,21,24,26,28	8,14,16,21,24,28
Nucleoside hydrolase	4	4	3**	14,18,26,29	14,18,26,29	14,18,29

***: one or more genes not present due to assembly gaps.**
#: Gene absent in the scaffold due to assembly gaps.
****: Gene prediction error**

The Gene ontology analysis identified genes responsible for defense system (GO: 0006952) and positive regulation of cell proliferation (GO: 0008284) were missing in late passage genome (Figure 4.3).

Figure 4.3: Comparison of early and late passage genome of *L. donovani* showing absence of defense system and positive regulation of cell proliferation genes in late passage



4.3.2. Chromosome copy number variation identified

In *leishmania* it's already known that copy number variants play an important role in gene regulation (Dumetz et al., 2017, Leprohon et al., 2009). The copy number variant analysis on the basis of read depth mapped binning ratio approach detected various deletion and insertion events on the genomes. It was reported that copy number variants help to maintain virulence in different environmental conditions (Dumetz et al., 2017). While analyzing the CNV events, approximately 365 CNV events were found in early passage that ranged from 0.7 kb to 271 kb in size. More number of events were overlapping with protein coding regions. It was found that a total of 230 deletion events and 135 duplication events were identified in the early passage. Whereas 234 deletions events and 146 duplication events were identified in the late passage genome. We found more CNV events in chromosomes 5, 6, 8, 15 and 31 in the genome of early passage as per earlier reports (Laffitte et al., 2016;Iantorno et al., 2017). High number of CNV events plays a major part in the gene expression regulation of in vitro cultured promastigotes (Dumetz et al., 2017).

ABC transporters, Amino acid transporters, amastins, GP63, calpain like cysteine are the major protein was having changes in both the passages. More changes were observed in the chromosomes of 5, 6, 8, 15 and 31 in early passage genomes as reported by (Iantorno et al., 2017, Laffitte et al., 2016).

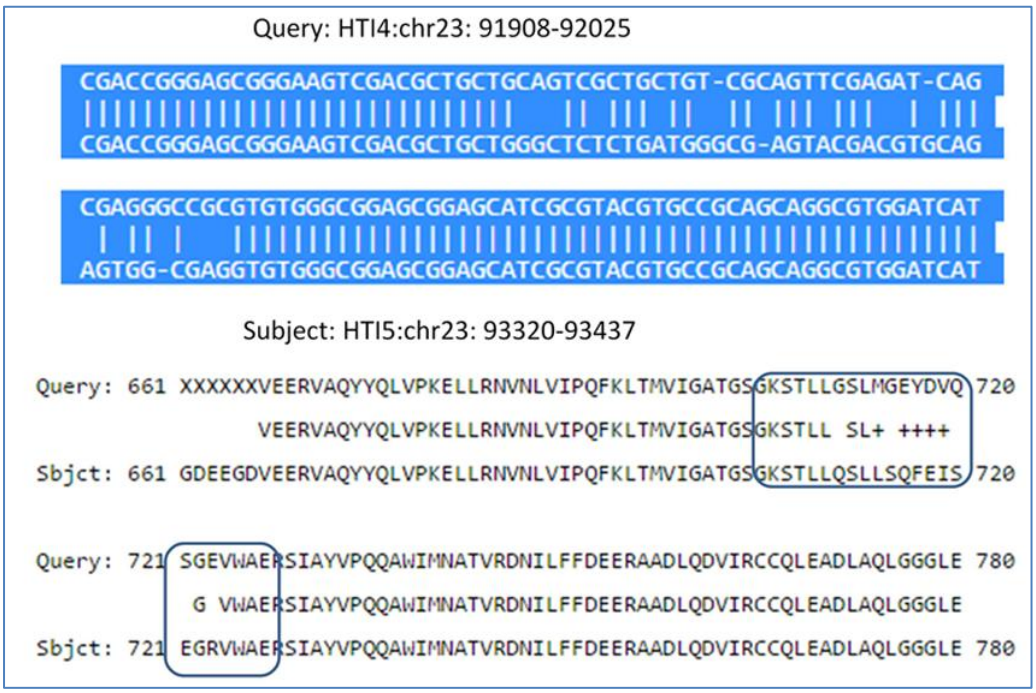
Interestingly more CNV events were observed in chromosome 3, 4, 13, 16 and 20 of the late passage genome. Chromosome 20 displayed more duplication events. Major insertion event from chromosome 20 was found only in late passage detected with the change in Phosphoglycerate kinase

B, cytosolic fragment, while this was absent in early passage genome. Noticeable protein coding genes falling in the variant regions on chromosome 13 of late passage genome were some acetyl transferases including histone acetyltrasferase, N-acetyl transferase subunit ARD1, RAS-related protein RAB5, mitogen activated protein kinase 2, etc. which may have a role in cell cycle progression and morphogenesis of Leishmania (Wiese, 1998;Yadav et al., 2016).

4.3.3. Single nucleotide polymorphism in transporter genes

The evolution of pathogenicity in microbes has been attributed to ordered changes in the functionality of the genes as a result of physiological constraints encountered by the organism in their immediate environment. A KEGG pathway analysis didn't show major differences in the number of genes involved, although detailed mutational analysis revealed interesting changes linked directly or indirectly to loss of virulence in the later passage. While studying the polymorphisms from early and late passage genomes, we found 4390 and 4356 heterozygous loci. ABC transporters play an important role in drug resistance and pathogen virulence (Glavinas et al., 2004). 42 copies of ABC transporter coding genes are found in early and late passage genomes. ABC transporter from the late passage at genome [chr23:91219-96174] has undergone modification in the nucleotide level, which makes ABC transporter inactive. There were also duplication events involving 12000 bp in the chromosomal location: [Chr23:86801-98800] detected in the late passage. From our early passage genome, ABC transporter from Chr23:91908-92025 inside a larger gene locus was found to be duplicated at another locus of early passage genome: chr23:89807-101617. Figure 4.4 shows the changes in ABC transporter. Accordingly we also found CNV events in chromosome23 at location 86801-98700 of early passage and that supports our finding as well.

Figure 4.4: Nucleotide and protein sequence comparison in *L. donovani* ABC transporter gene in Chr23 of early and late passages. Polymorphisms in nucleotide and protein sequences are showed in upper and lower panel.

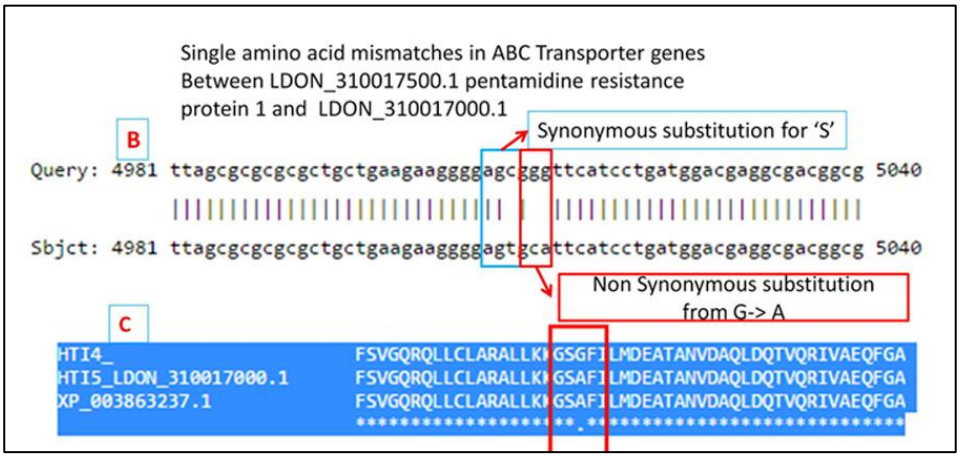


Another ABC transporter gene [early passage (Chr31; gene id LDON_310017500.); late passage genome(Chr31; with gene id of LDON_310017000.1)] had substitution at 5012th position (GGG -> GCA) a non-synonymous substitution which leads to a change from G->A amino acid at 1671th position in late passage. Polymorphism in ABC transporter is represented in Figure 4.5.

Figure 4.5: Chromosome 31 showing substitution in ABC transporter gene in early passage genome of *L. donovani*

We found that same gene had two substitutions at the 5010th position (from AGC -> AGT) leading to synonymous substitution. Another polymorphism containing ABC transporter gene is shown in Figure 4.6.

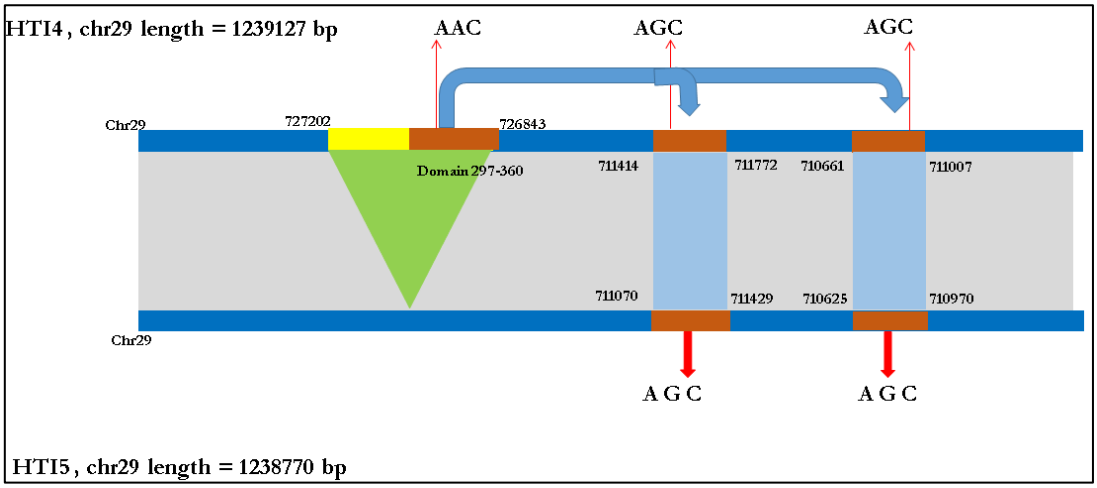
Figure 4.6: Comparison of ABC transporter genes in early, late passages along NCBI reference genome. B contains the nucleotide alignment of early and late passage genome showing substitutions in the nucleotide level whereas in C the changes in the amino acid among the early, late and reference *L. donovani*



Another interesting observation includes, reference genome of *L.donovani (*LdBPK282A1) containing A in that particular loci. The same A variant was observed in late passage genome. Whereas in early passage it was containing G as substitution Figure 4.6 (c) represents the changes in gene. We found copy number variant in chromosome 31 and the same was also reported by *Imamura* et al. as well (Imamura et al., 2016a). This gene codes for pentamidine resistance transporter protein. Major facilitator Protein (MFS class) is responsible for solute transport *via* membrane.

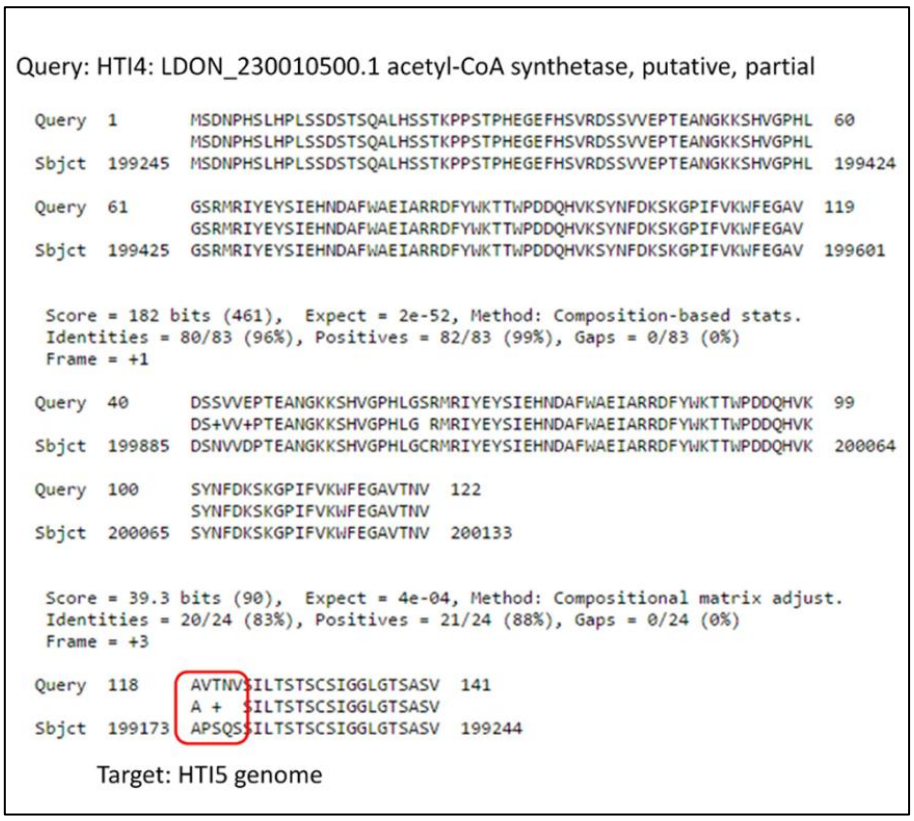
A small segment of the gene (297-360 bp) coding for MFS transporter is copied at two places both in late and early passage genome with one point non-synonymous mutation (AAC->AGC) resulting in N ->S . The MFS domain duplication is represented in Figure 4.7. Domain duplication in the early passage promastigotes (HTI4) indicate enhanced intracellular survival strategy.

Figure 4.7: Domain duplication of MFS transporter identified in chromosome 29 of early and late passages of *L. donovani*



Functional acetyl-CoA synthatase gene in early passage genome underwent modification in chromosome 23: 199275 -199598 in late passage and chromosome 23: 198391-19851 of early passage leading to loss of function (Figure 4.8).

Figure 4.8: Mutation in the acetyl-CoA synthatase gene from chromosome 23 in comparison with early and late passage genome



We found other important gene Calpain like proteases play an important role in infection process. A single insertion was identified in 172519th position in late passage chromosome 27 (Chr27:172458-188744) leads to frameshift mutation causing this gene to become a pseudogene. The corresponding functional gene in early passage is present at Chr27:172465-188750 (Figure 4.9). In case of Genbank L. donovani genome, the protein sequence is more identical to the HTI5 protein sequence. Recent reports have pointed towards the regulatory role played by expressed pseudogenes in cancer cells and parasites (Wen et al., 2011). Interestingly, parasite CALPs may serve other functions in the intracellular form which determine disease outcome and host responses (Branquinha et al., 2013) and are thus potent drug targets. This may open up new avenues in understanding Leishmania biology. The implications of these modifications need further investigation.

Figure 4.9: Detection of Frameshift mutation in the calpain like protease of chromosome 27 in late passage genome of *L. donovani*

4.4. Summary

In this study we attempted to identify *L. donovani*-specific genetic factors governing adaptability to host environment and pathogenicity using genomics approach. Using early and late passage genomes we detected subtle changes in the genome. This is accompanied by loss of function of drug resistance genes; defense related genes; appearance of pseudogenes. Copy number variants are observed in the genome. These findings help us to develop drug targets for disease management.

Publication from this work:

1. Genome plasticity in cultured *Leishmania donovani*: Comparison of early and late passages.

Roma Sinha[1#a†,] Mathu Malar [C2, 3†], Raghwan Kumar[1,#b†], Subhadeep Das[2, 3], Sonali Das[1], Md. Shadab[1#c], Rukhsana Chowdhury[1], Sucheta Tripathy[2, 3*,] Nahid Ali[1*] [Accepted in Frontiers in Microbiology 2018, First Author].

References

Abyzov A, Urban AE, Snyder M, Gerstein M, 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974-84.

Altschul SF, Gertz EM, Agarwala R, Schaffer AA, Yu YK, 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res* **37**, 815-24.

Andrews S, 2010. FastQC: a quality control tool for high throughput sequence data.

Au KF, Underwood JG, Lee L, Wong WH, 2012. Improving PacBio long read accuracy by short read alignment. *PLoS One* **7**, e46679.

Augusto Corrêa Dos Santos R, Goldman GH, Riaño-Pachón DM, 2017. ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics* **33**, 2575-6.

Bailey TL, Boden M, Buske FA, *et al.*, 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202-W8.

Bankevich A, Nurk S, Antipov D, *et al.*, 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* **19**, 455-77.

Bendtsen JD, Nielsen H, Von Heijne G, Brunak S, 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology* **340**, 783-95.

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM, 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology* **33**, 623.

Bioinformatics B, 2011. FastQC: a quality control tool for high throughput sequence data. *Cambridge, UK: Babraham Institute*.

Bleidorn C, 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and biodiversity* **14**, 1-8.

Blin K, Medema MH, Kazempour D, *et al.*, 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Research* **41**, W204-W12.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W, 2010. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9.

Boetzer M, Pirovano W, 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics* **15**, 211.

Bolger AM, Lohse M, Usadel B, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20.

Braslavsky I, Hebert B, Kartalov E, Quake SR, 2003. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences* **100**, 3960-4.

Brouwer H, Coutinho PM, Henrissat B, De Vries RP, 2014. Carbohydrate-related enzymes of important Phytophthora plant pathogens. *Fungal Genetics and Biology* **72**, 192-200.

Browning SR, Browning BL, 2011. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* **12**, 703.

Buermans H, Den Dunnen J, 2014. Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **1842**, 1932-41.

Butler J, Maccallum I, Kleber M, *et al.*, 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* **18**, 810-20.

Calderón CE, Ramos C, De Vicente A, Cazorla FM, 2015. Comparative genomic analysis of Pseudomonas chlororaphis PCL1606 reveals new insight into antifungal compounds involved in biocontrol. *Molecular Plant-Microbe Interactions* **28**, 249-60.

Chaisson MJ, Tesler G, 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC bioinformatics* **13**, 238.

Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. *Proceedings of the German conference on bioinformatics*, *1999*: Hanover, Germany, 45-56.

Chin C-S, Peluso P, Sedlazeck FJ, *et al.*, 2016a. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods* **13**, 1050.

Chin CS, Peluso P, Sedlazeck FJ, *et al.*, 2016b. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050-4.

Cingolani P, Platts A, Wang LL, *et al.*, 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92.

Cock PJ, Grüning BA, Paszkiewicz K, Pritchard L, 2013. Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. *PeerJ* **1**, e167.

Dong S, Raffaele S, Kamoun S, 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Current opinion in genetics & development* **35**, 57-65.

Downing T, Imamura H, Decuypere S, *et al.*, 2011. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res* **21**, 2143-56.

Dumetz F, Imamura H, Sanders M, *et al.*, 2017. Modulation of aneuploidy in Leishmania donovani during adaptation to different in vitro and in vivo environments and its impact on gene expression. *mBio* **8**, e00599-17.

Edge P, Bafna V, Bansal V, 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**, 801-12.

Ekblom R, Wolf JB, 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary applications* **7**, 1026-42.

Emanuelsson O, Brunak S, Von Heijne G, Nielsen H, 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols* **2**, 953-71.

Erwin DC, Ribeiro OK, 1996. *Phytophthora diseases worldwide*. American Phytopathological Society (APS Press).

Ewels P, Magnusson M, Lundin S, Käller M, 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-8.

Fiers W, Haegeman G, Iserentant D, Min Jou W, 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*.

Garg S, Martin M, Marschall T, 2016. Read-based phasing of related individuals. *Bioinformatics* **32**, i234-i42.

Giardine B, Riemer C, Hardison RC, *et al.*, 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-5.

Glavinas H, Krajcsi P, Cserepes J, Sarkadi B, 2004. The role of ABC transporters in drug resistance, metabolism and toxicity. *Current drug delivery* **1**, 27-42.

Gruenwald NJ, Goss EM, Press CM, 2008. Phytophthora ramorum: a pathogen with a remarkably wide host range causing sudden oak death on oaks and ramorum blight on woody ornamentals. *Molecular Plant Pathology* **9**, 729-40.

Gurevich A, Saveliev V, Vyahhi N, Tesler G, 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-5.

Guzvic M, 2013. The History of DNA Sequencing. *Journal of Medical Biochemistry* **32**, 301-12.

Haas B, 2007. TransposonPSI: an application of PSI-blast to mine (Retro-) transposon ORF homologies. In.

Haas BJ, Kamoun S, Zody MC, *et al.*, 2009a. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature* **461**, 393-8.

Haas BJ, Kamoun S, Zody MC, *et al.*, 2009b. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature* **461**, 393-8.

Hackl T, Hedrich R, Schultz J, Förster F, 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-11.

Heather JM, Chain B, 2016. The sequence of sequencers: the history of sequencing DNA. *Genomics* **107**, 1-8.

Holley RW, Apgar J, Everett GA, *et al.*, 1965. Structure of a ribonucleic acid. *science*, 1462-5.

Huang X, Madan A, 1999. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-77.

Hunt M, Newbold C, Berriman M, Otto TD, 2014. A comprehensive evaluation of assembly scaffolding tools. *Genome biology* **15**, R42.

Iantorno SA, Durrant C, Khan A, *et al.*, 2017. Gene expression in Leishmania is regulated predominantly by gene dosage. *mBio* **8**, e01393-17.

Imamura H, Downing T, Van Den Broeck F, *et al.*, 2016a. Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *Elife* **5**.

Imamura H, Downing T, Van Den Broeck F, *et al.*, 2016b. Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent. *Elife* **5**.

Ivors KL, Hayden KJ, Bonants PJM, Rizzo DM, Garbelotto M, 2004. AFLP and phylogenetic analyses of North American and European populations of *Phytophthora ramorum*. *Mycological Research* **108**, 378-92.

Jiang RH, Tyler BM, 2012. Mechanisms and evolution of virulence in oomycetes. *Annual review of phytopathology* **50**, 295-318.

Jones P, Binns D, Chang H-Y, *et al.*, 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-40.

Käll L, Krogh A, Sonnhammer EL, 2004. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology* **338**, 1027-36.

Keller O, Odronitz F, Stanke M, Kollmar M, Waack S, 2008. Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC bioinformatics* **9**, 278.

Kleigrewe K, Gerwick L, Sherman DH, Gerwick WH, 2016. Unique marine derived cyanobacterial biosynthetic genes for chemical diversity. *Natural product reports* **33**, 348-64.

Klioutchnikov G, Kriventseva EV, Zdobnov EM, 2017. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol*.

Knoll AH, 2008. Cyanobacteria and earth history. *The Cyanobacteria: Molecular Biology, Genomics, and Evolution* **484**.

Koren S, Schatz MC, Walenz BP, *et al.*, 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology* **30**, 693-700.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM, 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-36.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL, 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567-80.

Kurtz S, Phillippy A, Delcher AL, *et al.*, 2004a. Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Kurtz S, Phillippy A, Delcher AL, *et al.*, 2004b. Versatile and open software for comparing large genomes. *Genome biology* **5**, R12.

La Roche J, Van Der Staay G, Partensky F, *et al.*, 1996. Independent evolution of the prochlorophyte and green plant chlorophyll a/b light-harvesting proteins. *Proceedings of the National Academy of Sciences* **93**, 15244-8.

Laffitte M-CN, Leprohon P, Papadopoulou B, Ouellette M, 2016. Plasticity of the Leishmania genome leading to gene copy number variations and drug resistance. *F1000Research* **5**.

Lassmann T, Hayashizaki Y, Daub CO, 2009. TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* **25**, 2839-40.

Leao T, Castelão G, Korobeynikov A, *et al.*, 2017. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea. *Proceedings of the National Academy of Sciences* **114**, 3198-203.

Leprohon P, Légaré D, Raymond F, *et al.*, 2009. Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant Leishmania infantum. *Nucleic acids research* **37**, 1387-99.

Li H, 2011a. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-93.

Li H, 2011b. wgsim-Read simulator for next generation sequencing. *Github Repository*.

Li H, 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li H, Durbin R, 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-60.

Li H, Handsaker B, Wysoker A, *et al.*, 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-9.

Li H, Ruan J, Durbin R, 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-8.

Li L, Stoeckert CJ, Roos DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-89.

Liu B, Shi Y, Yuan J, *et al.*, 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint arXiv:1308.2012*.

Lohse M, Bolger AM, Nagel A, *et al.*, 2012. R obi NA: A user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic acids research* **40**, W622-W7.

Maxam AM, Gilbert W, 1977. A new method for sequencing DNA. *Proceedings of the national academy of sciences* **74**, 560-4.

Mccall LI, Zhang WW, Matlashewski G, 2013. Determinants for the development of visceral leishmaniasis disease. *PLoS Pathog* **9**, e1003053.

Miclotte G, Heydari M, Demeester P, Audenaert P, Fostier J. Jabba: Hybrid error correction for long sequencing reads using maximal exact matches. *Proceedings of the International Workshop on Algorithms in Bioinformatics, 2015*: Springer, 175-88.

Miller JR, Koren S, Sutton G, 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-27.

Mitchell TJ, 2006. Streptococcus pneumoniae: infection, inflammation and disease. In. *Hot Topics in Infection and Immunity in Children III.* Springer, 111-24.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M, 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**, W182-5.

Moss NA, Bertin MJ, Kleigrewe K, Leão TF, Gerwick L, Gerwick WH, 2016. Integrating mass spectrometry and genomics for cyanobacterial metabolite discovery. *Journal of industrial microbiology & biotechnology* **43**, 313-24.

Mulkidjanian AY, Koonin EV, Makarova KS, *et al.*, 2006. The cyanobacterial genome core and the origin of photosynthesis. *Proceedings of the National Academy of Sciences* **103**, 13126-31.

Myers G, Brown D, Morgenstern B, 2014. Algorithms in bioinformatics.

Nakagawa I, Kurokawa K, Yamashita A, *et al.*, 2003. Genome sequence of an M3 strain of Streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome Res* **13**, 1042-55.

Newman DJ, Cragg GM, 2016. Natural products as sources of new drugs from 1981 to 2014. *Journal of natural products* **79**, 629-61.

Nurk S, Bankevich A, Antipov D, *et al.*, 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of computational biology* **20**, 714-37.

Pancrace C, Jokela J, Sassoon N, *et al.*, 2017. Rearranged biosynthetic gene cluster and synthesis of hassallidin E in Planktothrix serta PCC 8927. *ACS chemical biology* **12**, 1796-804.

Pareek CS, Smoczynski R, Tretyn A, 2011. Sequencing technologies and genome sequencing. *Journal of applied genetics* **52**, 413-35.

Parra G, Bradnam K, Korf I, 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-7.

Parra G, Bradnam K, Ning Z, Keane T, Korf I, 2008. Assessing the gene space in draft genomes. *Nucleic Acids Research* **37**, 289-97.

Paszkiewicz K, Studholme DJ, 2010. De novo assembly of short sequence reads. *Briefings in bioinformatics* **11**, 457-72.

Peltola H, Söderlund H, Ukkonen E, 1984. SEQAID: A DNA sequence assembling program based on a mathematical model.

Peng Y, Leung HC, Yiu S-M, Chin FY, 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420-8.

Phillippy AM, 2017. New advances in sequence assembly. In.: Cold Spring Harbor Lab.

Pierleoni A, Martelli PL, Casadio R, 2008. PredGPI: a GPI-anchor predictor. *BMC bioinformatics* **9**, 392.

Pryszcz LP, Gabaldón T, 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic acids research* **44**, e113-e.

Quevillon E, Silventoinen V, Pillai S, *et al.*, 2005. InterProScan: protein domains identifier. *Nucleic acids research* **33**, W116-W20.

Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE, 2002. Whole-genome analysis of photosynthetic prokaryotes. *Science* **298**, 1616-20.

Salmela L, Rivals E, 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**, 3506-14.

Salmela L, Walve R, Rivals E, Ukkonen E, 2016. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics* **33**, 799-806.

Sanger F, Brownlee G, Barrell B, 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of molecular biology* **13**, 373IN1-98IN4.

Sanger F, Coulson AR, 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**, 441IN19447-446IN20448.

Sanger F, Nicklen S, Coulson AR, 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences* **74**, 5463-7.

Saunders DG, Win J, Kamoun S, Raffaele S, 2014. Two-dimensional data binning for the analysis of genome architecture in filamentous plant pathogens and other eukaryotes. *Plant-pathogen interactions: methods and protocols*, 29-51.

Schornack S, Van Damme M, Bozkurt TO, *et al.*, 2010. Ancient class of translocated oomycete effectors targets the host nucleus. *Proc Natl Acad Sci U S A* **107**, 17421-6.

Shendure J, Porreca GJ, Reppas NB, *et al.*, 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *science* **309**, 1728-32.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM, 2015a. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM, 2015b. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-2.

Simpson JT, Durbin R, 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22**, 549-56.

Smit A, Hubley R, Green P, 2016. RepeatMasker Open-4.0. 2015. *Google Scholar*.

Sommer DD, Delcher AL, Salzberg SL, Pop M, 2007. Minimus: a fast, lightweight genome assembler. *BMC bioinformatics* **8**, 64.

Sperschneider J, Gardiner DM, Dodds PN, *et al.*, 2016. EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist* **210**, 743-61.

Stam R, Jupe J, Howden AJ, *et al.*, 2013. Identification and characterisation CRN effectors in Phytophthora capsici shows modularity and functional diversity. *PLoS One* **8**, e59517.

Stanke M, Diekhans M, Baertsch R, Haussler D, 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-44.

Stanke M, Waack S, 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215-ii25.

Steinbiss S, Silva-Franco F, Brunk B, *et al.*, 2016. Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res* **44**, W29-34.

Tatusova T, Dicuccio M, Badretdin A, *et al.*, 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic acids research* **44**, 6614-24.

Tyler BM, Tripathy S, Zhang X, *et al.*, 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261-6.

Van Der Auwera GA, Carneiro MO, Hartl C, *et al.*, 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 11.0. 1-.0. 33.

Walker BJ, Abeel T, Shea T, *et al.*, 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963.

Weber T, Blin K, Duddela S, *et al.*, 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic acids research* **43**, W237-W43.

Welker M, Von Döhren H, 2006. Cyanobacterial peptides—nature's own combinatorial biosynthesis. *FEMS microbiology reviews* **30**, 530-63.

Whisson SC, Boevink PC, Moleleki L, *et al.*, 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* **450**, 115.

Win J, Morgan W, Bos J, *et al.*, 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *The Plant Cell* **19**, 2349-69.

Winnenburg R, Baldwin TK, Urban M, Rawlings C, Köhler J, Hammond-Kosack KE, 2006. PHI-base: a new database for pathogen host interactions. *Nucleic acids research* **34**, D459-D64.

Xue W, Li J-T, Zhu Y-P, *et al.*, 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC genomics* **14**, 604.

Yang Z, 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-91.

Yin L, An Y, Qu J, *et al.*, 2017a. Genome sequence of Plasmopara viticola and insight into the pathogenic mechanism. *Scientific Reports* **7**.

Yin L, An Y, Qu J, *et al.*, 2017b. Genome sequence of Plasmopara viticola and insight into the pathogenic mechanism. *Sci Rep* **7**, 46553.

Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y, 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research* **40**, W445-W51.

Zehr JP, Bench SR, Carter BJ, *et al.*, 2008. Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *science* **322**, 1110-2.

Zhang WW, Ramasamy G, Mccall L-I, *et al.*, 2014a. Genetic analysis of Leishmania donovani tropism using a naturally attenuated cutaneous strain. *PLoS pathogens* **10**, e1004244.

Zhang WW, Ramasamy G, Mccall LI, *et al.*, 2014b. Genetic analysis of Leishmania donovani tropism using a naturally attenuated cutaneous strain. *PLoS Pathog* **10**, e1004244.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT, 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* **16**, 1099-108.

Chapter 5 – Conclusion

Advancement in sequencing technology opened up new avenues for sequencing a variety of organisms ranging from microbes to humans. Sequencing technologies evolved from short read to long read technology and from low throughput to high throughput. The cost of sequencing has also fallen dramatically making it affordable for every small lab. Simultaneously, the tools available for analyzing the genomes are also evolving at a rapid pace. Genome assembly is the key step, the accuracy of which is largely responsible for making the right analysis and conclusions.

The emergence of third generation sequencing technologies e.g.; long reads generation has made a paradigm shift in genome science. By using long read sequencing, problems with genome assemblies such as large gaps, misassemblies can be curtailed which was not possible with short reads.

The number of assemblers available now is huge and choosing the right assembler for the right data type leads to success in a genome project. Handling complex repeats, heterozygous regions in eukaryotes make the assembly process challenging. The higher the organism, the more complex is the genome and so also is the assembly process. Genome assembly algorithms are mainly classified into two types; The De Bruijn Graph method and Overlap Layout Consensus (OLC) method. Short read assemblers such as Allpaths, SOAPdenovo2, Abyss, velvet are based on De Bruijn graphs. CAP3, ARACHNE, NEWBLER, CELERA, PCAP are the assemblers which are based on Overlap Layout Consensus.

In our sequenced genomes, we optimized assemblies using Allpaths, SPAdes, Celera, and FALCON. By using Allpaths assembler we assembled Cyanobacterial genomes which were collected from various parts of eastern India. When we assess the genome assemblies of cyanobacteria's, the ones generated with Allpaths assembler had good assembly metrics such as higher N50, larger scaffolds, and more core ortholog genes.

Falcon long read assembler which is based on Overlap Layout Consensus, assembles and prepares the contig bubble graph and stores the information of haplotypes. Illumina reads can be used along with the long reads to phase the genome and achieve haplotype phased assembly. Polishing with illumina reads helps to reduce indel errors from Pacbio assembled genome.

5.1. Genome Assembly of photosynthetic prokaryotes

We sequenced *T. bouteillei* and *L. confervoides* using illumina sequencing technology. The assembled genomes were more complete in comparison with other related de novo genomes. Assembling genomes for the first time with Allpaths never resulted in the best assembly. However, a second round involving sheared larger jump libraries from the first assembly; bioinformatically resulted in the optimal assembly. Recently, we found that spades hybrid assembler produces the best result for mixed data types. During genome mining, various secondary metabolite clusters, NRPs were identified which codes for antifungal and antibacterial activity. These can be exploited for further commercialization process.

5.2. Genome Assembly of Eukaryotic Plant pathogen

We have sequenced genomes of two eukaryotic tree pathogens of *Phytophthora ramorum* species e.g.; Pr102, a pathogen on Oak and ND886, a pathogen on Camellia plant. In Pr102 isolate where an older chemistry, P5-C3 was used, needed a lot of effort in error correction. A 3-way error correction protocol resulted in more percentage of good quality reads. The assembly obtained from the error corrected reads contained number of effectors, transposable elements and no gaps when compared with the older assembly version V1 (Chapter 3a).

For isolate ND886, a newer sequencing chemistry, P6-C4 was used. This increased the read accuracy dramatically and the error correction was the not the most critical step in this assembly. FALCON assembler proved to be the best long read assembler producing an assembly of 60.5 Mb. additionally, genotypic phasing and haplotype block detection was done using a series of sophisticated bioinformatics tools. One of the trickiest part in haplotype phasing and variant calling is to differentiate between the true read errors and the variants. One needs to be very careful while dealing with this and presence of additional high coverage Illumina reads are a must for this step. Our overall assembly resulted in 345 haplotype blocks, 222,892 phased variants, some new decameric repeats and several paralogs of one effector gene that was probably lying in the gap region of the earlier assembly. So, it is imperative for higher eukaryotic genome assembly to have high quality short reads as well as long reads to be able to phase haplotypes and obtain complex repetitive sequences (Chapter 3 b).

When only short reads are available, Spades prove to be a better assembler. Combining a first round of assembly and providing that to SPADES as long read proves to improve the assembly statistics as evidenced by our analysis on *Phytophthora plurivora* genomes (Chapter 3 C).Genome sequence of *P. plurivora* is consistent with the genome architecture of other sequenced *Phytophthora* species, and there is evidence of polyploidy.

5.3. Genome Assembly of Eukaryotic Human Pathogen

For assembly of genomes, that already has a high quality reference; it is much easier to get a chromosome level assembly. The work presented here highlights the importance of newly emerging pathoadaptive factors like transporters and calpain-like proteins in *Leishmania* virulence. Our work can add to the pool of information on adaptive genomics in *L. donovani* and careful mining of these genes can be used for virulence surveillance at least throughout the Indian subcontinent. The genome sequences of early and late passage promastigotes can act as references for future genomic studies on *Leishmania*, particularly related to virulence and drug resistance. Annotated leads from this study and further annotation and functional characterization of hypothetical proteins can provide novel drug/vaccine targets for disease management.

5.4. Scope of Future Work-:

During the course of this dissertation, we were successful in developing genome assembly pipelines for various ranges of organisms from prokaryotic to complex eukaryotic heterozygous pathogens. There are scopes to take this research ahead in the near future like:

- Complete genome assemblies can give more insights into biology and genetics of the organisms. Several important genes can be taken from the Cyanobacterial genomes that we have assembled and can be exploited for commercial purposes.
- Diploid genome assembly of *P. ramorum* ND886 will serve as a good reference for population genomics studies. Divergence of haplotypes and genetic information can be studied from the phased genome.
- Newly sequenced *P. plurivora* genome can be used to check the bio-control activity for controlling pathogenicity.
- The reference based genome assembly was performed for generating chromosome level assembly. Our reference based STLab assembler that can be an excellent resource for generating complete chromosome level assemblies for genomes having high quality references. Many cues can be drawn from our work on genome assembly of early and late phase assemblies of *Leishmania donovani*. This information can be used to design potential drug targets.