Fall 12-15-2017

Next generation sequencing technologies for real-time genotyping and targeted sequencing for precision medicine

Priyanka Rawat

PRIYANKA RAWAT

Candidate

BIOMEDICAL ENGINEERING

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

DR. JEREMY EDWARDS , Chairperson

DR. ELIZABETH DIRK

DR. HEATHER CANAVAN

DR. PAYMAN ZARKESH-HA

Next-generation sequencing technologies for real-time genotyping and targeted sequencing tailored for precision medicine.

By

Priyanka Rawat

Bachelors of Technology, Biotechnology, Mahrshi DayanandUniversity, 2008
M.S., Biomedical Engineering, University of New Mexico, 2015

Dissertation

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biomedical Engineering

The University of New Mexico
Albuquerque, New Mexico

Dec, 2017

DEDICATION

This dissertation is dedicated to my sister Parul, who is no more with me but who has always instilled inspiration in me to bring out my best even in the rockiest waters. I know you are watching me from over there, I ensure you that I will try to live life with as zeal as you wanted to. To my parents, who have always been there as a rock-solid support. To my father, who always motivated me to achieve my dreams and supported me throughout with all his capacity. To my mother, who made so many sacrifices just to keep her children succeed in their endeavors. To my sister Prerna, who always told me that she will be there anywhere when I needed her and did so. To my brother Rahul, although younger than me, but you have always given me support no matter what and stands as my pillar.

Finally, to my husband Pranav, who started this journey with me, lived each day with me. Thank-you for motivating me when I was low, appreciating my efforts and taking care of us as a family.

To my son Sarvang, you are the most precious of all, you have brought the strength in me to face hardest of times, to achieve the best and to hope for the best.

ACKNOWLEDGMENTS

I would like to acknowledge my committee for their constant guidance and motivation to finish my projects. I would like to thank my advisor and mentor Dr. Jeremy Edwards, for giving me this opportunity to work with you, for your support and guidance without which it would have been a difficult path. To Dr. Payman Zarkesh-Ha, for having faith in me on a subject in which I had no expertise, our discussions made it easy for me to learn semi-conductor basics and electronics. To Dr. Elizabeth Dirk, your motivational discussions and never to give-up attitude made me believe in myself , without you, I would have lost my path easily, I am truly indebted to you. To Dr. Heather Canavan, your professional guidance, be it in my project, dissertation structure, for preparation of my exams or directions for professional career has been very valuable to shape up my doctoral journey. I would also like to acknowledge Edwards Lab members for being supportive in experiments, lab discussions, feedback and encouragement. To Linda Marie Bugge, for always being there to solve any problem and extending her timely support.

Next-generation sequencing technologies for real-time genotyping and targeted sequencing tailored for precision medicine.

By

Priyanka Rawat

Bachelors of Technology, Biotechnology, Mahrshi DayanandUniversity, 2008

M.S., Biomedical Engineering, University of New Mexico, 2015

Ph.D., Optical Science & Engineering, University of New Mexico, 2017

ABSTRACT

Astounding success of Human genome project and accelerating success of sequencing technologies have enabled $1000 genome goals possible. But, this is still far-fetched from the reach of many resource refrained populations with high genetic variations causing lethal genetic diseases. Based on present technology principles, I have developed prototypes for affordable, scalable and customizable point-of-care genotyping and targeted sequencing. Ion-sensitive field effect transistors with novel read-out and signal amplification techniques are used for laying foundation of possible ISFET based allele-arrays. Sequencing-by-synthesis based full-fledge sequencer is made with novel immobilization, flow-cell and data acquisition methods discussed in Chapter - 2 and 3. In chapter-4, I have developed mis-assembly validation protocol for de-novo sequenced Kibdelosporangium MJNF-24 genome. During this validation, I found for the first time that assembly of this prokaryotic genome sequenced with long-read technology is not completely accurate. I characterize the mis-assemblies and annotate the errors by using open-source tools and scripts.

Chapter 1

Introduction

By

Priyanka Rawat

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

"It is far more important to know what person the disease has than what disease the person has"-Hippocrates

Prognosis based on predictive and preventive, personalized medicine model and abandoning the traditional model of treating pathologies assures reduced burden on the current health care system ((Gonzalez-garay, 2015). Personalized medicine model is based on diagnosis of precision biological markers which impact disease pathologies, drug metabolism and treatment outcomes. These pharmacogenetic markers are genetic variations caused mostly due to single nucleotide polymorphisms (SNP) which differ between individuals(Alwi, 2005).One way that ensures successful implementation of precision biology is early diagnosis and treatment based on pharmacogenetic markers. These markers have been studied to alter drug metabolism from person to person for complex diseases (Ventola, 2013). Single nucleotide polymorphisms may determine or cause complex diseases, alter the efficacy and safety of drugs and are associated with treatment outcomes(Sim, Kacevska, & Ingelman-Sundberg, 2012). Thus, tailoring drug therapy according to individual genotype could aid in reduc-

ing drug related fatalities by implementation of accurate drugs and dosage(Shenfield, 2004). All this could be made possible by rapidly developing next generation sequencing technologies which have the potential of accelerating the diagnosis of diseases and pharmacogenetic markers by genome analysis. In the last decade development has accelerated by leaps and bound in genomic technologies. The cost of sequencing has decreased 14000-fold between1999 to 2009 (Scott, 2013). The success of human genome project launched in 1987(Steward et al., 2017) gave insights to variations in genomic position causing phenotypic differences in humans. Population based sequencing projects like 1000 genomes help create comprehensive databases for population, geographic and environmental specific variations in human genome(Gonzaga, 2012). Large population based projects which explore the medical value of whole genome sequencing such as UK 100,000 genome projects(100K Genomes. Sequencing 100000 Genomes. 2014, 2014). Although, this data provides insights to what genes or coding regions can cause diseases for those who had already had the diseases but not for those who can develop diseases in future. Genomic sequencing can help in predicting genetic diseases in patients making it possible for early treatments. Veritas genetics (Goodwin, Mcpherson, & Mccombie, 2016) is the first health-care company which has made it possible to sequence whole human genome for $1000. But, the downsides include high annotation times (12-16 weeks), possibility of finding novel variants leading to confused result interpretations and also delayed prognosis of crucial etiologies.

Next generation Sequencing Technologies: We are at the apex of next generation sequencing technologies. After the Nobel laureate, Frederick Sanger developed the di-deoxy chain termination method conjugated with electrophoretic size separation of DNA for sequencing(Sanger & Nicklen, 1977), rapid developments in NGS sequencing dramatically reduced the price of sequencing per base to $0.0024 by mid-1990s(Goodwin, McPherson, & McCombie, 2016). Successful and a decade longer

human genome project motivated development of many NGS technologies which led to reduction in sequencing costs dramatically from $2400 per/Mb (ABI 3130xL) to 0.031per/Mb (Ion-torrent Proton II)(Niedringhaus, Milanova, Kerby, Snyder, & Barron, 2012),(Quail et al., 2012). $1000 genome project was started as an initiative to propel technology development towards economic direction and establishing a benchmark for routine, affordable and personal genome sequencing. This led to rapid and dramatic decrease in cost per genome from $95,263,072 (Sept.01) to $1,245 (Oct.15)(Goodwin, McPherson, et al., 2016).

A generalized sequencing technology process can be divided into preparation, detection and analysis. Preparation involves molecular methods of DNA manipulation and NGS library preparation (emulsion PCR in most technologies(Pang, Macdonald, Yuen, Hayes, & Scherer, 2014)), detection or read-out based on optical imaging or electro-chemical readout followed by bio-informatics analysis(Head et al., 2014). Most of the high throughput commercially available technologies share technological commonalities and critical differences. Many technologies are under developments which have the same goals but with different attributes. Each system has an advantage over another in either cost reduction, time or efficiency and accuracy (Table1) .

Whole genome sequencing

Whole genome sequencing involves shearing or fragmenting the genome, enzymatically modifying and amplifying the fragments, immobilization on solid support, sequencing and data acquisition.Table 1 summarizes the costs, total run time and expected read lengths for sequencing technologies. I will discuss 4 major technologies here which dominate the market now. Illumina sequencing is based on sequencing by synthesis based on incorporation of reversible terminator dNTPs. After DNA fragmentation, several enzymatic modifications, clusters are generated on a flow cell from single stranded DNA and sequencing is done by addition of 3-end blocked nucleotides with fluorophores to the immobilized templates. Fluorescent image are registered and

Company/Platform	Technology	Price	Cost $/run	Cost $/Mb	Run Time	Output per run	Read Length	Accuracy %	Refrences
ABI 3130xL		95k	$4/800bp	2,400	20min to 3hrs	1.984k	400-900b	100.00%	Allseq.com
Roche 454	Optical	500k	$7,000	10	24hrs.	0.7G	700b	99.99%	Allseq.com
ABi SOLiD 5500xL		665k	~$10k	0.105	6 days	180Gb	2x60bp		Allseq.com
Illumina MiSeq	Optical	125k	$1.4k	0.093	65hrs.	15Gb	2x300		refs(Quail et al., 2012)
Illumina Hiseq Hiseq X		1M*	$12k	0.007	3days	1.8Tb	2x150		Allseq.com
Ion torrent PGM 318	pH-	80k	250Reagent	0.375	4-7hrs.	2Gb	200b		refs(Quail et al., 2012)
Ion torrent PGM 316	"	50k	250Reagent	0.549	3-5hrs.	1Gb	400b		Allseq.com
Ion-Torrent Proton I	"	149k	300Reagent	0.999	2-4hrs	10Gb	200b		Allseq.com
Ion-Torrent Proton II	"	149k	300Reagent	0.031	2-4hrs.	32Gb	100b		Allseq.com
PacBio RSII	Optical	695k	$400	1.45	240min.	500Mb-1Gb.	10-15kb	>99.99%	refs(Quail et al., 2012)
PacBio Sequel	"	700k	$850	0.085	240min.	5Gb-10Gb	10-15kb	>99.99%	refs(Quail et al., 2012)
Oxford Nanopore		125K	$1,000	0.01495	2 days	100GB	150kb	~90% ID	refs(Goodwin, Mcpherson, et al., 2016)

Figure 1.1: Table1

used for base-calling(Liu et al., 2012). Ion- torrent covers 25% market and is based on electronic detection of incorporated natural nucleotides. Fragmented and enzymatically modified strands of DNA are amplified on acrylamide beads and extension is done.

pH is measured with ISFET based pH meter. Illumina and Ion-torrent both are short-read technologies which enable WGS, exome-sequencing, transcriptome and targeted amplicon sequencing. Illuminas paired-end technology has become a validation measure for de-novo assemblies by other platforms(Vezzi, Narzisi, & Mishra, 2012). Pacbio and Oxford Nanopore are long-read technologies with read length from 10kb-15kb and 150kb respectively. Pacbio sequences a circular single molecule with hairpin adaptors ligated at both ends of double-stranded template (Figure 1.1 c). These templates are immobilized to zero-mode waveguide pores and sequencing is done by incorporation of one nucleotide at a time. Pacbio, like Illumina uses sequencing-by-

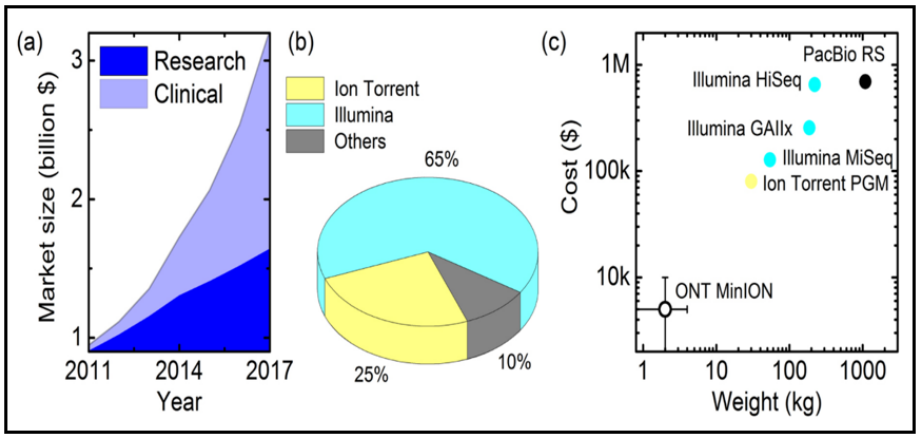


Figure 1.2: a) Market-size analysis of sequencing services in research and Clinical sector. b) Market-segmentation showing dominance of Illumina and Ion-torrent . c)Cost and size comparison of major sequencing technologies (Figure adapted from :The emergence of nanopores in next-generation sequencing L. J. Steinbock and A. Radenovic 2015 Nanotechnology 26 074003 doi:10.1088/0957-4484/26/7/074003.).

synthesis technology but sequences single molecule at a time rather than clusters of templates(Goodwin, Mcpherson, et al., 2016). In the world of desktop size sequencing machines, Oxford Nanopore MinIon is a pocket-size sequencer based on nanopore sequencing. Chnages in ionics flux across the nanopore placed in electrical field are detected when DNA is passed through the pore. Minimal library preparation is required in Nanopore sequencing and no fluorophore tags are used. There are theoretically no limits to read-length in this technology but high error rates and detection speed are major technological hurdles(Dewey et al., 2013). Although, Illumina and Ion-torrent cover major sequencing market, higher cost of the instruments and reagents is still a deciding factor for organizations to pursue sequencing projects. Oxford Nanopore is cheaper and smaller in size comparison to other 3 technologies but still have high error rates and limitations for being used in clinical settings.

Targeted sequencing: Targeted sequencing is investigating specific sequence in the genome coding genes, exons or desired genetic sequences by sequencing. It reduces the overall cost of sequencing and facilitates finding low-levels of variation which might be otherwise not detected. Several key technologies have developed

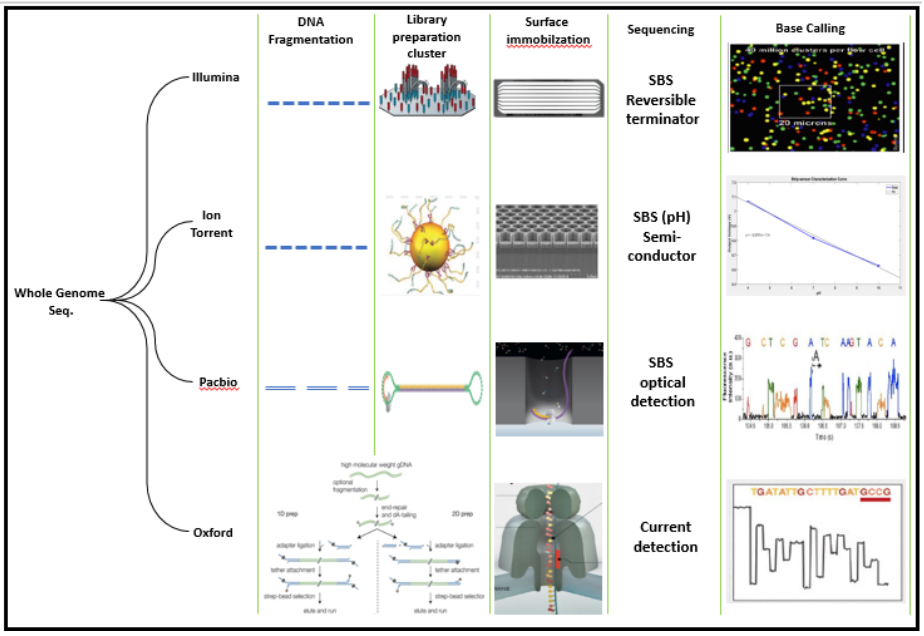


Figure 1.3: Commercial whole-genome sequencing technologies and principal of sequencing for each technology.

platforms specific for targeted sequencing (Ion-torrent and MiSeq(Illumina) for both clinical and research purposes. The technologies are based on capturing the genomic sequence of interest by hybridizing to the probe in solution(Dapprich et al., 2016), on solid-support(Albert et al., 2007) or multiplexed pcr based amplification technologies.(Porreca et al., 2007).

Ion-torrent(Zhang et al., 2015) and Illumina have both introduced targeted panels for various genetic diseases along with options for customizable panels by the users. Nimblegen Seq Cap and Agilent sure select (Samorodnitsky et al., 2015)are other companies that provide target capture kits for sequencing.

Although, targeted sequencing has been limited to read-length or amplicon size, there have been methods to capture longer sequences (utpo $\tilde{2}$0kb) of specific genes of interest with surrounding genetic sequences to not miss any important association information(Dapprich et al., 2016).

Technologies under development: There are a few emerging technologies that are exploring new horizons to push the limitations for cost-effective reliable systems.

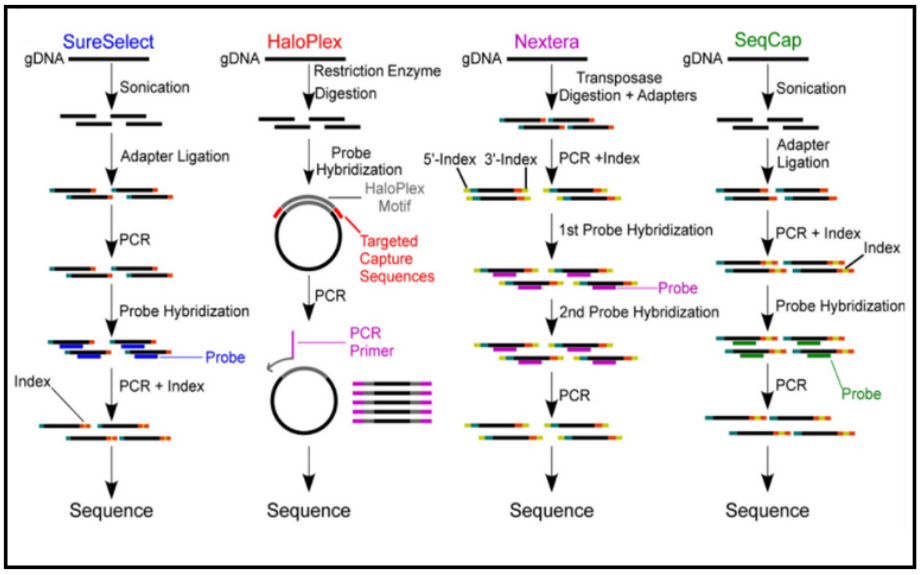


Figure 1.4: Targeted sequencing methods. The library preparation protocols by different technologies are shown. Sure Select and SeqCap use similar technology. Nextera uses transposons and Haloples uses enzymes respectively. Figure adapted from .

Most of these technologies are based on nanopore sequencing (Electronic Biosciences, Oxford Nanopore, Genia, Stratos Genomics and Two pore guys) or optical or transmission electron microscopy (Light Speed Genomics, Electron Optics and ZS Genetics). Only, two are ISFET based pH sensing systems in which one (Genapsys) focuses on manufacturing cheaper sequencing machine and other on infectious disease diagnostics systems (DNA electronics) but still rely on complex pcr protocols.

Whole genome sequencing versus Targeted sequencing for precision medicine

Whole genome sequencing involves sequencing of complete genomes, which has its advantages and disadvantages. Whole genome sequencing can detect all kinds of mutations and novel mutations which are hard for WES or targeted sequencing to investigate (Pang et al., 2014) and since the foundation of precision medicine lies in detecting and studying the genome-wide associations of coding and non-coding regions related to genetic variations, WGS has an advantage over targeted sequencing methods to study non-coding regions too. It has been shown that some of the genetic

diseases are an interplay of rare causative variants of large effect along-with common variants(Dewey et al., 2013). Thus, WGS can unravel rare mutations and its effect on common diseases which is not possible with targeted sequencing of only exonic regions. On the other hand, major limitations of WGS in application in clinical utility include costs, time to reporting and storage of huge data. Routine sequencing of whole genomes for clinical purposes is limited majorly due to these reasons.

Targeted sequencing provides insights to molecular profiling of certain diseases (Shao et al., 2016), and is crucial in determining a persons response to targeted therapies. Current clinical practices involve various probe-based, fluorescence-based and amplification based gene detection technologies which are expensive, limited to one or two genes and time-consuming(Shao et al., 2016). Overcoming the major limitations of whole-genome sequencing for higher cost and massive data, targeted sequencing also enables higher sequencing depth across targeted regions and increased accuracy conquering intricacies of complex regions(Dapprich et al., 2016).

- Sequencing and Genotyping: Whole genome sequencing is determining the exact sequence of DNA while genotyping is to determine the variants an individual possesses by use of specific biological markers. Sequencing ensures the order of nucleotides by incorporation of complementary nucleotides on a single stranded DNA. There have been many platforms developed for sequencing (Table 1) which are costly, time-consuming and require downstream data analysis. Genotyping has been mainly dependent on microarrays and PCRs with optical detection13. Biomarkers are immobilized on the microarrays and DNA to be queried is hybridized. If there is a matching, the event is registered through light by optical imaging. Platforms such as Affymetrix, Agilent technologies, and Illumina Microarrays are bulky desktop systems and costs tens of thousands of dollars. Each run again costs hundreds of dollars per test and requires interpretation of data by some skilled personnel before information can be used

by a physician for disease diagnosis. These microarray setups and sequencing systems require dedicated facilities for functioning which limit the portability. Current NGS technology was not developed with a purpose of diagnosis, on-site analysis and easier clinical interpretations. Although, there are a few emerging technologies for point-of-care analysis like Biotage PyroMark, Cepheid GeneXpert, CombiMatrix Corp., DirectIf Diagnostics Solutions and Nanogen Inc.s Nanochip NC40015. Most of these technologies follow the footprints of current massively parallel sequencing technologies for microarray applications (Biotage for 454 pyrosequencing), are based on optical imaging (Cepheid GeneXpert), quantitative fluorescent-PCR detection and electrochemical detection system which makes them unsuitable for point-of-care diagnostics (POCD) (Chen et al., 2014) purpose. Some technologies suffer from scalability, costly instruments and lack of accuracy(Craw & Balachandran, 2012). Above all, these systems are designed for infectious disease diagnosis at the point of care rather than pharmacogenetic markers(Garner et al., 2010). POCD devices should be scalable, sensitive to the target, mass manufacturable and ideally be disposable(Garner et al., 2010). Rapid and accurate diagnosis with consumer oriented features makes it suitable for on-site use in clinical settings.

Routine targeted sequencing: Targeted routine sequencing for genetic diseases for most prevalent diseases in developing countries with a centralized model to serve populations routinely for particular genes may soon be the norm in the genomic diagnostics. Economic burden on developing nations due to costs related to healthcare is increasing. Almost 700,000 babies are born with a genetic disorder or congenital disorder each year worldwide. Among these 300,000 suffer from hemoglobinopathies only(Christianson, Howson, & Modell, 2006). In the last decade, deaths due to non-communicable diseases have increased by 8 million world-wide in last decade(Weatherall, Dc, & Weatherall, 2012). Genetic dis-orders such as sickle-

cell anemia with 180,000 babies born per year in sub-saharan Africa and similar high numbers of off-springs with thalassemia disorders in Asian countries(Weatherall et al., 2012) require routine genetic tests on clusters of populations. Developing countries have poor diagnosis, treatment and management facilities for such health disparities. Routine targeted sequencing for such monogenic diseases for parents is necessary to rule out birth of such offsprings. Genetic diseases not only have higher economic, but also emotional burden on the family along with the suffering of the child. Due to lack of proper facilities, physical and financial resources, diagnostics based on genetic testings are rarely available. Present sequencing technologies are expensive, require high-end instrumentations, power resources and trained personnel to sequence. In 2014, Illumina announced to have achieved $1000 genome goal, but that did not include the machine costs of 1M, and other complexities of the procedure involved. Also, most successful sequencing technologies are not customized for routine clinical screening. Veritas genetics have shown that $1000 whole genome sequencing is a reality, but this remains challenging for many developing countries. Also, timings for bio-informatics analysis for whole genome is 12-16 weeks.

In this dissertation, we have made an attempt to explore methods by which we can develop affordable, customizable and scalable sequencing apparatus. For routine diagnostics, it is important that the total cost is affordable for even small laboratories or clinics. The instrument should be end-user customizable allowing flexibility to tailor different diagnostic tests on one machine with same reagents. To be able to adjust scalability, is not what we can find in todays sequencing machines. We have made attempts to make same sequencing machine to be able to run for single experiments or for multiple experiments simultaneously.

Our first attempt has been to make develop point-of-care SNP genotyping devices. We have developed prototype for single nucleotide extension with ISFET based sensors for point-of-care diagnostics. Label free electro-chemical detection of DNA by

ion-sensitive field effect transistors (ISFET) fabricated with complementary metal oxide sensor technology (CMOS) (Devadhasan, Yoo, & Kim, 2015) has the potential to mark the beginning of an era of semiconductor genetics. ISFETs are the best candidates for genetic POCD devices owing to the compatibility to DNA reaction chemistries, specificity and parallel scalability properties. Ion-torrent and emerging technologies (Genapsys) have leveraged this property to revolutionize the sequencing market, but the POCD genotyping market still remains unexplored. For our first prototype we focused on developing allele specific targeted SNP genotyping ISFET POCD device with properties discussed below:

- Novel read-out technique- Novel read-out technique for signal amplification and suitable configuration of ISFET for higher sensitive genotyping device is developed.

- Targeted SNP genotyping- It has been shown that with whole genome sequencing, rare mutations could be discovered which might influence the course of treatment or divert the mainstream pathology interpretations21.Therefore, allele specific ISFET arrays(Figure 1.3) with immobilized probes can reduce the genotyping time and increase the accuracy.

- Low-cost - Use of natural dNTPs and enzymes along with ISFET could bring down the cost to $50-100 per test.

- Low-power- Highly sensitive ISFET and signal amplifying readout ensures higher sensitivity at low-power (10μA/pH at VGS 1.5V and VDS .5V).

Our next attempt is to build a sequencing machine based on sequencing by synthesis chemistry with 3blocked reversible terminators as used in Illumina technology. Based on three attributes discussed in Chapter2 we have made a cheaper, easy to assemble, scalable and customizable sequencing apparatus. We have used off-the shelf

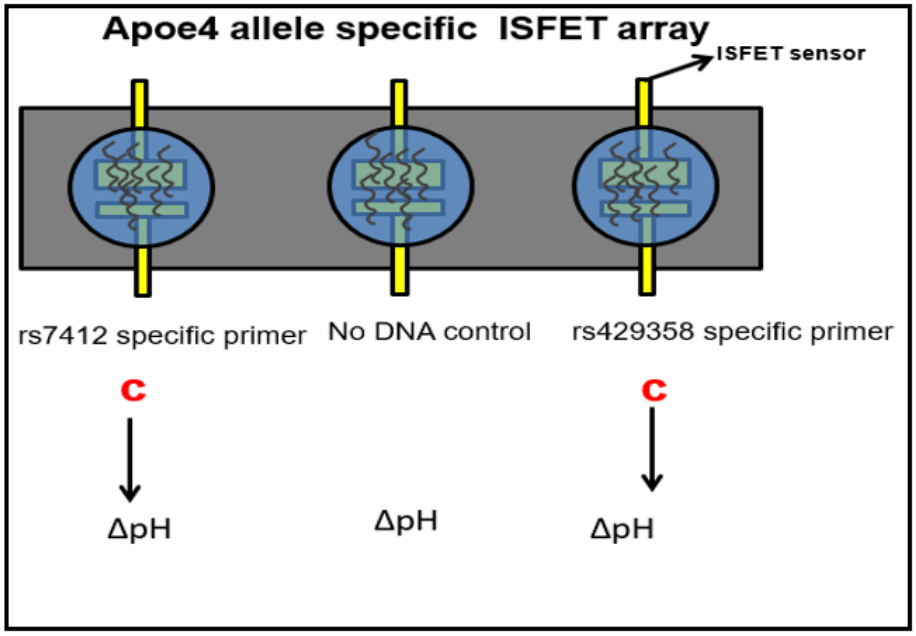


Figure 1.5: Proposed allele-specific ISFET-array for real-time genotyping. Apoe4 allele is characterized by presence of two SNPs. Primers are hybridized to the surface of ISFET specific to the SNPs. On hybridization to the template there is a Ph change detected. Based on the output form each sensor, genotyping is done.

electronics and fluidics. Instead of using expensive lasers we have used light emitting diodes. Higher expenditure associated with flow-cell and reagents is replaced by making cheaper in-house flow-cell of silicon and reagents. We have developed our own chemistry for immobilizing sequencing molecules on the glass-surface and loading in the flow-cell. This system can be assembled easily and with some more efforts can be made in a sophisticated sequencing machine. It can be cutomised for targeted sequencing for shorter amplicons or longer amplicons with changes in library preparation. Multi-sample run can be done by simple barcoding the samples and thus useful for multiple diagnostics tests. We demonstrate the competency of our system by sequencing a panel of 327 inherited genes with 10,500 amplicons with optimal results.

Thus, in this dissertation we present a sequencing prototype based on sequencing by synthesis and optical detection of extension events. The efforts have been made to make an end-user customizable, affordable and scalable sequencer. Further, reagents

are formulated and optimized in the lab, novel surface immobilization techniques have been developed which can be altered for specific applications (whole genome v's targeted sequencing). In-house flow-cell is developed to allow scalability per sequencing run. In efforts to bring down costs further and design a system which can be assembled in the absence of high-end electronics and fluidic components, we have used commonly used components. Chapter 2 discusses the components and assembly of the 3 prototypes. Chapter 3 discusses the library preparation, biochemical surface preparation, data-acquisition and bio-informatics data analysis.

Illumina has introduced HiSeq X Ten systems which lower down the cost of whole genome sequencing to $1000- $1500 but it does not include costs of instrument, consumables and expensive reagents. Also, Illumina and Ion-torrent acquire most of the sequencing market and thus there is monopoly of the technologies which results to higher costs. We have proved that with few optimizations, same technologies can be used to make cheaper and scalable apparatus.

Information contained in sequencing data is not useful without Bio-infomatics analysis. Leading sequencing technologies have their own specified methods of analysis and thus assembly of genomes. As discussed, long-read technologies are known to generate golden standard genomes but the accuracy of such genome has not been validated. Thus, in Chapter 4 I have validated de-novo sequenced and assembled draft and complete genome assembly of Kibdelosporangium MJ-NF24 actinobacteria. Draft genome is a hybrid assembly of multiple sequencing technology whereas complete genome is assembled with long-read Pacbio data. I have shown that the accuracy and correctness of Pacbio data is just little higher than the draft assembly and has its own share of mis-assemblies.

References

T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., Gibbs, R. a. (2007). Direct selection of human genomic loci by microarray hybridization. Nature Methods, 4(11), 903905. https://doi.org/10.1038/nmeth1111 Alwi, Z. Bin. (2005). The Use of SNPs in Pharmacogenomics Studies. The Malaysian Journal of Medical Sciences: MJMS, 12(2), 412. Chen, X., Sullivan, P. F., Duca, F. A., Lam, T. K. T., Beigh, M., Craw, P., Toumazou, C. (2014). Review Article. Digest of Technical Papers - IEEE International Solid-State Circuits Conference, 12(2), 14. https://doi.org/10.1109/ISSCC.2010.5433834 Christianson, A., Howson, C., & Modell, B. (2006). March of Dimes. Global report on birth defect. The hidden toll of dying and disabled children. New York, 1016. Craw, P., & Balachandran, W. (2012). Isothermal nucleic acid amplification technologies for point-of-care diagnostics: a critical review. Lab on a Chip, 12(14), 24692486. https://doi.org/10.1039/c2lc40100b Dapprich, J., Ferriola, D., Mackiewicz, K., Clark, P. M., Rappaport, E., DArcy, M., Durbin, R. (2016). The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. BMC Genomics, 17(1), 486. https://doi.org/10.1186/s12864-016-2836-6 Devadhasan, J. P., Yoo, I. S., & Kim, S. (2015). Overview of CMOS image sensor use in molecular diagnostics. Current Applied Physics, 15(3), 402411. https://doi.org/10.1016/j.cap.2015.01.009 Dewey, F. E., Pan, S.,

Wheeler, M. T., Stephen, R., Ashley, E. A., & Dphil, M. (2013). DNA sequencing: Clinical applications of new DNA sequencing technologies, 125(7), 931944. https://doi.org/10.1161/CIRCULATIONAHA.110.972828.DNA Garner, D. M., Bai, H., Georgiou, P., Constandinou, T. G., Reed, S., Shepherd, L. M., Toumazou, C. (2010). A multichannel DNA SoC for rapid point-of-care gene detection. Digest of Technical Papers - IEEE International Solid-State Circuits Conference, 53, 492493. https://doi.org/10.1109/ISSCC.2010.5433834 Gonzaga, C. (2012). Human genome sequencing in health and disease. Annu Rev Med., 63, 3561. https://doi.org/10.1146/annurev-med-051010-162644.Human Goodwin, S., Mcpherson, J. D., & Mccombie, W. R. (2016). Coming of age: ten years of next- generation sequencing technologies. Nature Publishing Group, 17(6), 333351. https://doi.org/10.1038/nrg.2016.49 Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), 333351. https://doi.org/10.1038/nrg.2016.49 Head, S. R., Kiyomi Komori, H., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. BioTechniques, 56(2), 6177. https://doi.org/10.2144/000114133 Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Law, M. (2012). Comparison of next-generation sequencing systems. Journal of Biomedicine and Biotechnology, 2012. https://doi.org/10.1155/2012/251364 Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., & Barron, A. E. (2012). NIH Public Access, 83(12), 43274341. https://doi.org/10.1021/ac2010857.Landscape Pang, A. W. C., Macdonald, J. R., Yuen, R. K. C., Hayes, V. M., & Scherer, S. W. (2014). Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. G3 (Bethesda, Md.), 4(1), 635. https://doi.org/10.1534/g3.113.008797 Porreca, G. J., Zhang, K., Li, J.

B., Xie, B., Austin, D., Vassallo, S. L., Shendure, J. (2007). Multiplex amplification of large sets of human exons. Nature Methods, 4(11), 931936. https://doi.org/10.1038/nmeth1110 Quail, M. M., Smith, M. E., Coupland, P., Otto, T. D. T., Harris, S. R. S., Connor, T. R., Salzberg, S. (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics, 13(1), 1. https://doi.org/10.1186/1471-2164-13-341 Samorodnitsky, E., Datta, J., Jewell, B. M., Hagopian, R., Miya, J., Wing, M. R., Roychowdhury, S. (2015). Comparison of custom capture for targeted next-generation DNA sequencing. Journal of Molecular Diagnostics, 17(1), 6475. https://doi.org/10.1016/j.jmoldx.2014.09.009 Sanger, F., & Nicklen, S. (1977). DNA sequencing with chain-terminating. Pnas, 74(12), 54635467. https://doi.org/http://dx.doi.org/10.1073%2Fpnas.74.12.5463 Scott, G. (2013). The neurotechnology revolution has arrived. Futurist, 47(5), 67. https://doi.org/10.1038/464674a Shao, D., Lin, Y., Liu, J., Wan, L., Liu, Z., Cheng, S., He, J. (2016). A targeted next-generation sequencing method for identifying clinically relevant mutation profiles in lung adenocarcinoma. Scientific Reports, 6(November 2015), 22338. https://doi.org/10.1038/srep22338 Shenfield, G. M. (2004). Genetic polymorphisms, drug metabolism and drug concentrations. The Clinical Biochemist. Reviews / Australian Association of Clinical Biochemists, 25(4), 2036. Sim, S., Kacevska, M., & Ingelman-Sundberg, M. (2012). Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. The Pharmacogenomics Journal, 13(10), 111. https://doi.org/10.1038/tpj.2012.45 Steward, C. A., Parker, A. P. J., Minassian, B. A., Sisodiya, S. M., Frankish, A., & Harrow, J. (2017). Genome annotation for clinical genomic diagnostics: strengths and weaknesses. Genome Medicine, 9(1), 49. https://doi.org/10.1186/s13073-017-0441-1 Ventola, C. L. (2013). Role of pharmacogenomic biomarkers in predicting and improv-

ing drug response: part 1: the clinical significance of pharmacogenetic variants. P & T: A Peer-Reviewed Journal for Formulary Management, 38(9), 54560. Vezzi, F., Narzisi, G., & Mishra, B. (2012). Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. PLoS ONE, 7(12), 111. https://doi.org/10.1371/journal.pone.0052210 Weatherall, D. J., Dc, W., & Weatherall, J. (2012). The inherited diseases of hemoglobin are an emerging global health burden The inherited diseases of hemoglobin are an emerging global health burden, 115(22), 43314336. https://doi.org/10.1182/blood-2010-01-251348 Zhang, B., Ryan Penton, C., Xue, C., Wang, Q., Zheng, T., & Tiedje, J. M. (2015). Evaluation of the ion torrent personal genome machine for gene-targeted studies using amplicons of the nitrogenase gene nifH. Applied and Environmental Microbiology, 81(13), 45364545. https://doi.org/10.1128/AEM.00111-15

Online reference list

- 100K Genomes. Sequencing 100000 Genomes. 2014:https://www.genomicsengland.co.uk/

Chapter 2

Design and construction of DNA Sequencing Prototypes

By

Priyanka Rawat

Table of Contents

List of Figures

List of Tables

Chapter 2

Design and construction of DNA sequencing prototypes

2.1 Introduction

In this chapter, I will discuss the design and construction of the DNA sequencing prototypes. Based on the sequencing technology there were three versions built (prototype-1, prototype-2 and prototype-3): First two prototypes were based on ISFET sensor based DNA sequencing and the last prototype was based on optical detection of the sequencing events (prototype-3). I started with something very simple (a commercial pH sensor and meter to sequence 50bp Oligos) and then added one part (fluidics, electronics and software) at a time to achieve the final design (prototype-2). Figure 2.1 presents the various parts of the prototype-1 and 2 with all components. Prototype-1 and 2 had 4 major parts: Fluidics, electronics, sensors (ISFET), software and bio-chemical interface to the ISFET chip (DNA Immobilizing surface/set-up). Fluidics included the bottles, pipings, fittings, control valves and pumps. A single syringe pump was also used in prototype-2 with pressured argon to drive the fluid. Electronics included ISFET sensor (used to read the pH) with op-

amp (operational amplifier) electronics, data acquisition card and control electronics (Arduino microcontroller and control relay array). For control, automation and data acquisition an Arduino IDE (sketch) along with LabVIEW and Visual Basic 6 was used. For data analysis Microsoft Excel and MATLAB was used. Chemical aspect of the surface preparation, DNA immobilization and data analysis are discussed in Chapter-3. The prototype-3 fluidics and control electronics was developed based on prototype-1 and 2.

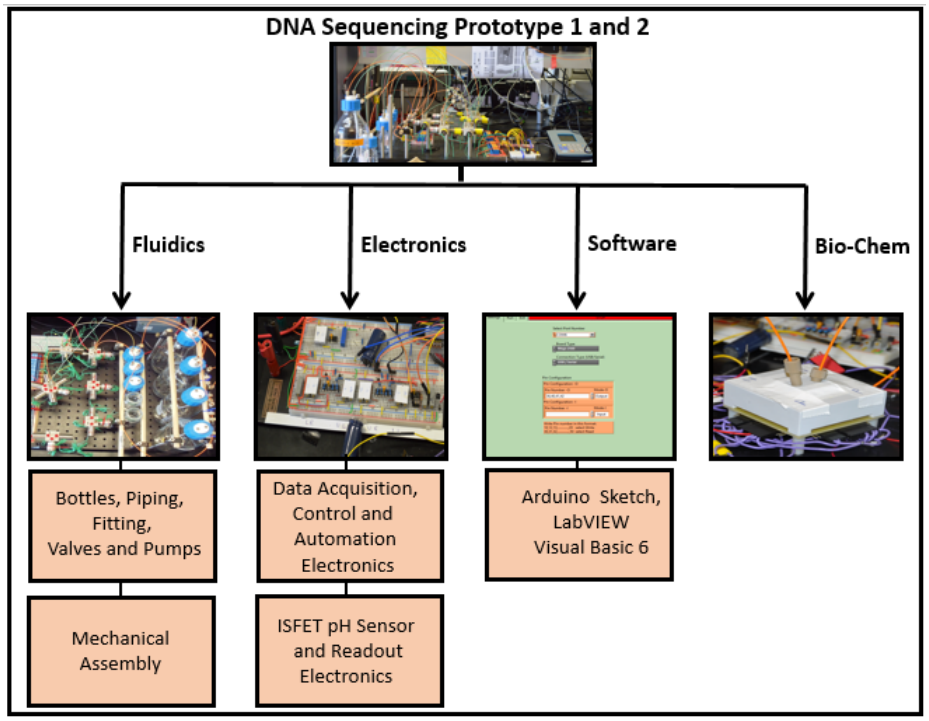


Figure 2.1: Prototype-1 and 2 Parts and Components.

The prototypes were developed in steps. In the first step the sequencing (sequence 50bp Oligos) was verified using a commercially available pH meter and sensor; this way not only the bio-chemistry was tested but also was obtained a valuable optimization on several experimental parameters such as response time (required to obtain a stable pH), volume (amount of liquid needed), amount of DNA required vs sensitivity (of the ISFET sensor), any dependence on environment (temperature and effect of air) and ions diffusion rate (ions are diffused at their own rate in liquid which affects

the reading of pH over time). The experiments I performed are discussed in the next session (2.2). The electronics, fluidics and software (for the prototypes) was developed based on the information obtained.

2.2 Isothermal single nucleotide extension with ISFET Sensor

The incorporation of single nucleotide event can result in change of pH due to release of hydrogen ion after hydrolysis of dNTP molecule incorporation. This change can be modeled by equation 2.1

$$H_a dNTP + DNA = DNA_{+1} + H_b PP_i + H_c c = ab \tag{2.1}$$

Where **a** is the number of dNTP, **b** is the number of pyrophosphates formed and **c** is the number of hydrogen ions released. This reaction should be in unique equilibrium as the hydrogen ions released should be equal to nucleotides incorporated into the DNA strand and those consumed by pyrophosphate. This is the basic principle for DNA sequencing using ISFET sensor. But to achieve that, DNA must be coated (DNA Immobilization provides controlled number of DNA molecules and thus, controlled pH variations which can be measured efficiently.) on some platform where the reaction can take place.

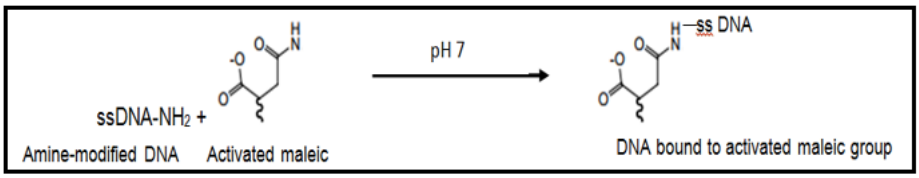


Figure 2.2: Maleic anhydride activated group is used to covalently bind amine-modified DNA oligomers at neutral pH.

I also did a brief study on DNA immobilization chemistry to identify the best

platform for future genotyping experiments and compatibility with the sensor. Several DNA immobilization chemistries like covalent bond formation with maleic anhydride (Figure 2.2) and polymerization with acrylamide (Figure 2.3) were used to identify the best platform.

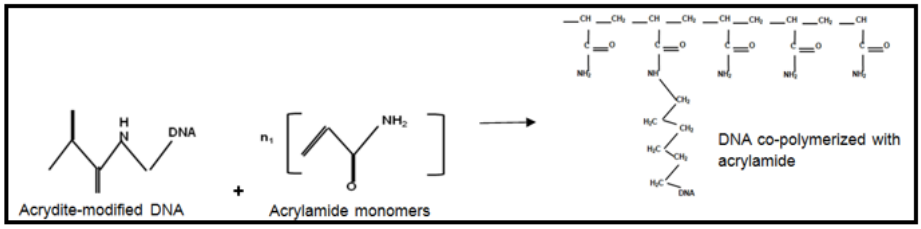


Figure 2.3: Co-polymerization of acrydite-modified DNA with acrylamide monomers.

For preliminary DNA sequencing with ISFET: I used three different experimental setups (the experiments are discussed in detail in chapter 3 section 3.2.2) to validate and optimize multiple base incorporations and single nucleotide incorporations on multiple copies of single stranded oligomers. Oligomers were purchased, immobilized and isothermal extension was done. The three experimental setups were:

1. 50bp amine-modified oligomer was covalently immobilized on maleic anhydride activated 96 well-plates. 20bp primer was hybridized to the oligomer and multiple base extensions and single nucleotide extensions with controls were studied (Figure 2.4).

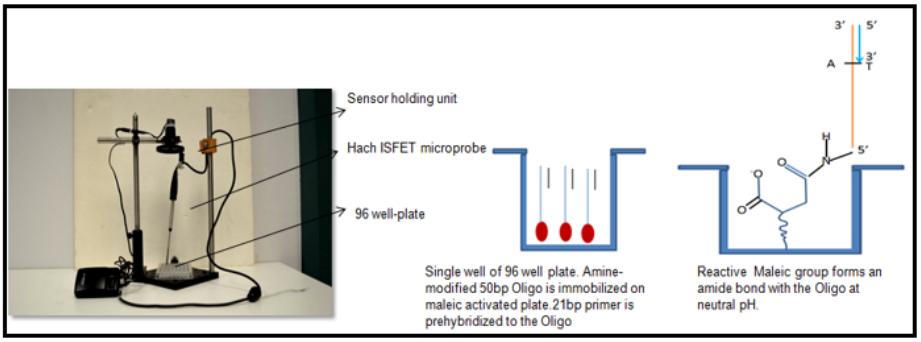


Figure 2.4: HACH ISFET sensor is immobilized in a sensor holding unit made specifically for this set-up. Pierce binding maleic anhydride coated 96 well plates are used for DNA Immobilization.

I used HACH ISFET sensor and meter to read the pH. I developed a unique holder for these experiments. I used the holder in two modes: First was Sturdy Mode in which the holder kept the sensor upright in sturdy position and second was Stir Mode in which the holder moved the sensor in smooth circular manner. The smooth circular manner was useful to study the ion diffusion (check if the sensor motion was any useful to stabilize the pH). I found out that there wasnt a significant difference between the two modes. Figure 2.5 presents Setup-1; the sensor, meter and holder with the specs. I named the holder pH rocker because it supposed to change the pH readings when used in stir mode but it did not, which was a good thing.

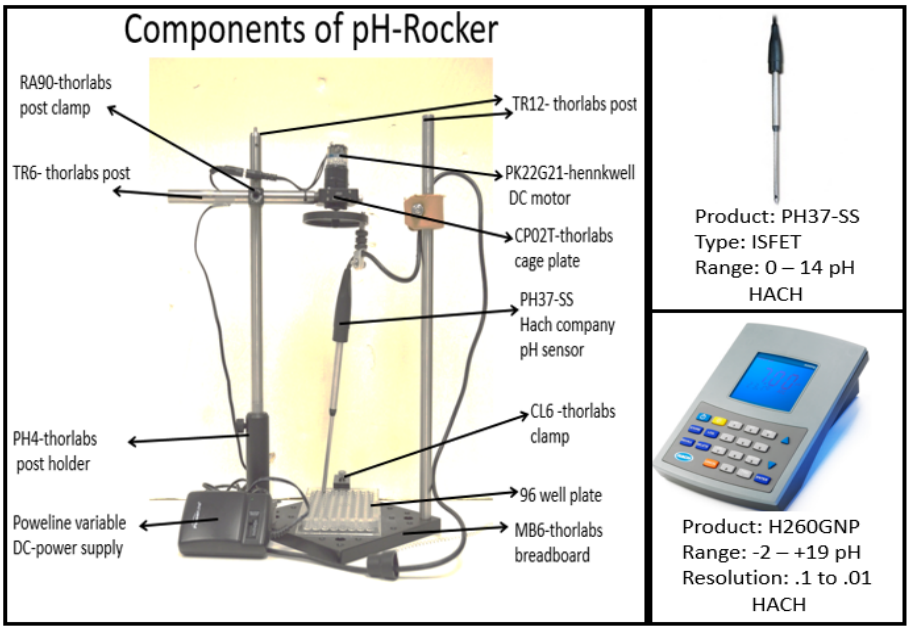


Figure 2.5: Setup-1; pH sensor, pH meter and pH rocker (holder).

2. In second experiment the Unmodified 50bp Oligo was used in 50μl 1X reaction buffer pre-hybridized with 20bp primers. Extension with Bst. polymerase at 37^{o}C was done. Sentron ISFET microprobe was used in this case which was capable of sensing pH in small volumes (20μl) making it ideal for single nucleotide resolution experiments. A low-pressure argon head was used to displace

atmospheric air and hence carbon-dioxide which could alter the pH readings. To mimic genotyping experiments, different experiments such as insertion and non-insertion events were done. See section 3.2. 2. Figure 2.6 presents the setup-2; Micro FET sensor, meter and setup with the specifications.

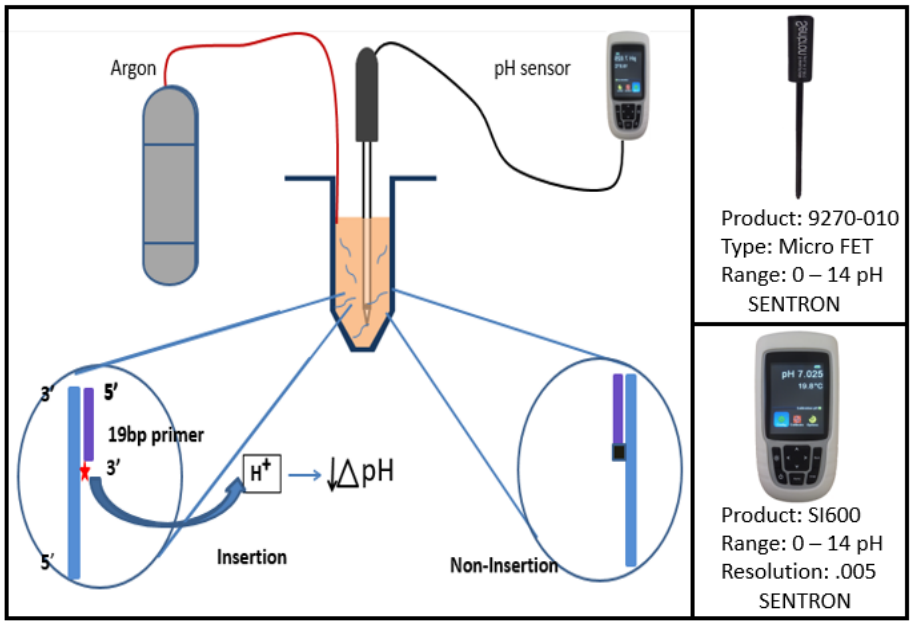


Figure 2.6: Setup-2; Micro FET pH sensor and pH meter. 50μl tube is used to do genotyping reaction in with DNA in solution. Experiment is designed to detect pH changes during single nucleotide extension when complementary and uncomplimentary dNTP is added.

3. After, successful single nucleotide resolution detection in solution, DNA immobilization was achieved by co-polymerizing acrydite modified DNA pre-hybridized with 20bp primer with 10% acrylamide 4mm gel beads, this was the third and final experiment. Setup-2 was used for pH measurements. To determine if beads were feasible for these experiments, pH changes on single nucleotide extension with and without DNA (beads) were analyzed. Mean difference of 0.067pH was observed for two experiments for .8μM of DNA and no DNA control. Multiple base extension reaction experiments were done with beads and it gave a normalized pH difference of .5 for .8μM DNA. Results were confirmed with 15% PAGE-gel analysis. Also, beads were stained with SYBR gold to confirm the

presence of DNA with Biorad UV imager (Figure 2.7). Further details on this experiment are discussed in chapter 3.

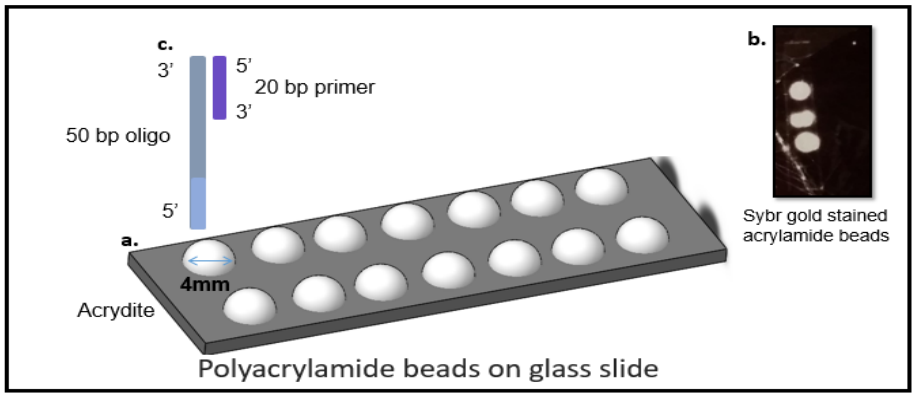


Figure 2.7: a) 4mm acrylamide beads with DNA co-polymerized, b) SYBR gold stained beads to confirm co-polymerization, c) Descriptive picture showing acrydite-modified oligo pre-hybridized with primers.

These experiments were performed for optimizing the experimental variables for SNP detection with ISFET sensor. Average pH changes for multiple and single nucleotide extension reaction were characterized in controlled environment.

Based on the information obtained, prototype-1 and 2 were designed and built with a customized ISFET sensor and fluidics. The two prototypes had similar fluidics and control electronics but different ISFET sensors (and readout circuits) and software. In next few sections I will discuss the design and build of the prototype-1 and 2 and later prototype-3.

2.3 Design and construction of DNA Sequencing Prototype-1

The prototype-1 had three parts (as shown in Figure 2.1): Fluidics, Electronics (ISFET sensor, readout electronics, control electronics and data acquisition hardware) and Software. Prototype-1 was a precursor of final Prototype-2. Prototype-1 incorporated a commercial pH sensor with a fully automated fluidics system to obtain a

reliable DNA sequencing but before going into that I will first discuss the ISFET sensor and its background with signal amplification electronics briefly.

2.3.1 ISFET pH Sensor and novel pH-to-current readout circuit

The ion-sensitive field effect transistor (ISFET), is a micro sensing element, combining electro-chemistry with microelectronic technology, was first proposed and fabricated by [Bergveld 1970]. Because of its compatibility with complementary metal-oxide-semiconductor (CMOS) manufacturing technologies, portability and the label-free detection of bio-molecule binding the ISFET technology has been investigated intensively [Barbaro 2006, Lee 2009, Uslu 2004]. The gate insulator of the ISFET senses the specific ion concentration (Figure 2.8), and generates an interface potential at the gate, which causes a change in drain source current depending upon the bias of ISFET. The sensitivity of an ISFET is proportional to $\Delta I_{DS}/\Delta V_T$, where ΔI_{DS} is the change in the drain source current and ΔV_T is the change of interface potential at gate, due to the capture or adsorption of analytes. For an ISFET at a given bias, it is always desirable to have a large sensitivity, which means higher ΔI_{DS} for a given ΔV_T. This is mainly due to the limitations on signal-to-noise ratio (SNR) in ISFET sensors. Among many focus areas to increase the sensitivity of ISFET, the selection of biasing regime of ISFET [Shoorideh 2012, Chapman 2011], and the readout technique [Hammond 2004, Yang 2007]are the key ones. Different readout techniques that are proposed in literature [Chan 2007, Wang 2012] for ISFET sensors include differential voltage readout through ISFET and reference FET (REFET), threshold voltage readout, temperature compensated threshold voltage readout, voltage based time to digital conversion readout, differential current mode readout. The goal of our ISFET sensor setup was to maximize the sensitivity of the device that would be used for DNA sequencing, where the SNR of high density arrays of small ISFETs were

required for accurate readout. The similar readout circuit was used for single ISFET based sensor which I used in prototype-1.

ISFET Sensor Structure and Biasing

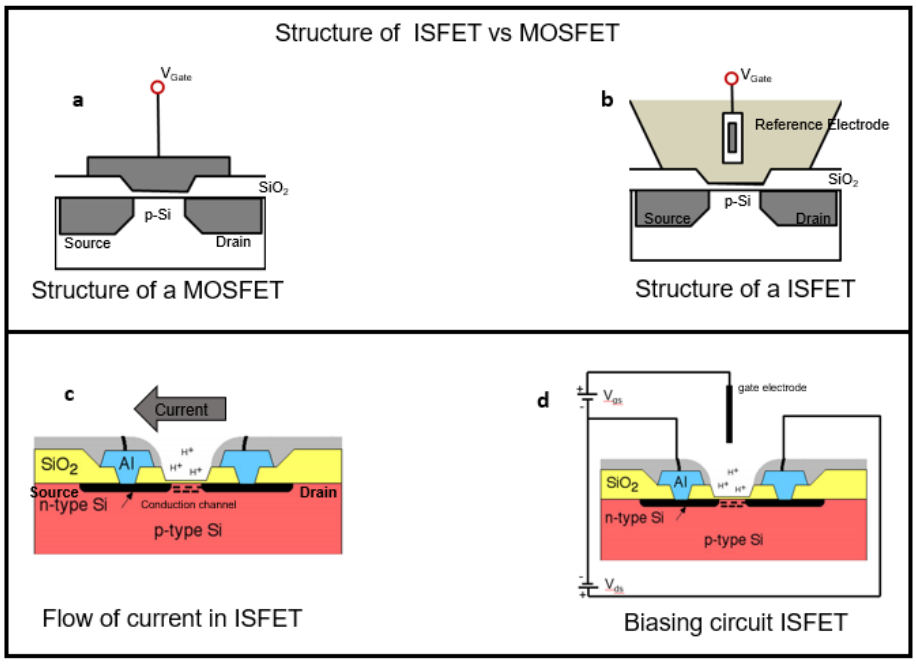


Figure 2.8: a) Structure of MOSFET (Comparison of a MOSFET and an ISFET: A reference electrode in analyte works as gate in ISFET. When an ionic layer of protons is formed due to changes in pH at the ion sensitive layer, current flow from drain to source increases), b) Structure of ISFET, c) Current flow in ISFET and d) ISFET biasing circuit.

Figure 2.8 presents the structure of ISFET and MOSFET. ISFET are very like MOSFET. It has been shown in previous work to allow pH sensing in low power (nano-watts) condition when fabricated in unmodified CMOS making them advantageous for portable applications. An ISFET has same physical primitives as MOSFET (Figure 2.8) and can be integrated into CMOS circuits. This system has been well-established as an alternative for solid-state pH measuring systems as it is sensitive to specific ionic-concentration which can be measured as change in voltage due to change in current at the drain-source terminals.

The current through ISFET can be modeled with that of MOSFET as stated by

[Bergveld 1970]. The threshold voltage of ISFET is given by[Bergveld 2003]:

$$V_T = E_{ref} - \psi_0 + \chi_{solution} - \frac{\phi_{si}}{q} - \frac{Q_{ox} + Q_{ss} + Q_B}{C_{ox}} + 2\phi_f \tag{2.2}$$

where, E_{ref} is the constant potential of reference electrode, φ_0 is the interface potential at solution/oxide interface, $\chi_{solution}$ is the surface dipole potential of the solvent, ϕ_{Si} is he work function of the silicon, Q_{ox}, Q_{ss} and Q_B are the oxide charge, oxide interface charge and bulk charge respectively, $2\phi_f$ is the potential for inversion of channel and C_{ox} is the oxide capacitance /unit area.

Equation (2.2) can be simplified[Shepherd 2005] to be:

$$V_T = E_{th(MOS)} - V_{chem} \tag{2.3}$$

Where

$$V_{chem} = \gamma + 2.303.\alpha..U_T.pH \tag{2.4}$$

Here, γ is a group of pH independent chemical potential, α is a dimensionless sensitivity parameter and $UT = \frac{kT}{q}$ is the thermal potential. In an ideal case, where the highest sensitive interface materials (Ta_2O_5) is used on the gate (the ISFET sensor in Prototype 1 is Ta_2O_5), α is approximately equal to 1 and UT at room temperature is about 26mV. Based on (2.3), for every unit of pH change, we can achieve about 60mV of change (theoretically) in threshold voltage when Ta_2O_5 is used as gate interface material. Therefore, as the pH of gate liquid increase, threshold voltage of ISFET increases too. This increment of threshold voltage causes reduced drain source current IDS of ISFET for a given bias condition.

Bias dependent ISFET sensitivity:

The current through drain-source of ISFET can be modeled with that of MOSFET . This drain-source current strongly depends on the V_{GS} and V_{DS} bias voltage sources

of ISFET. The sensitivity of an ISFET can be defined as the change of ISFET current due to the change of gate liquid pH i.e. ISFET ability to detect pH change. This pH sensitivity of ISFET is related with the change of ISFET current through the following relation:

$$S_{ISFET} = \frac{\Delta I_{DS}}{\Delta pH} = \frac{\Delta I_{DS}}{\Delta V_{GS}} \times \frac{\Delta V_T}{\Delta pH} \tag{2.5}$$

Where

$$\frac{\Delta I_{DS}}{\Delta V_{GS}} = K'_n \frac{W}{L} V_{DS} \tag{2.6}$$

where S_{ISFET} is the sensitivity of ISFET device. For Ta_2O_5 as interface material in ISFET gate, a threshold voltage change of around 60mV for every unit of pH change can be achieved (theoretically). A sensitivity of around 35μA/pH for an ideal ISFET with W/L=2, V_T=0.3V and K'_n=160μA/V^2 can be achieved. Here I_{DS} is the drain source current, W is the ISFET channel width, L is the channel length and K'_n is the current driving capacity of an ISFET.

The sensitivity of ISFET depends on many factors as shown in the equation 2.5. To explore this relationship, we used Spice Simulation; the result is shown in Figure 2.9. The SPICE simulation results for a long-channel N-type ISFET using TSMC's

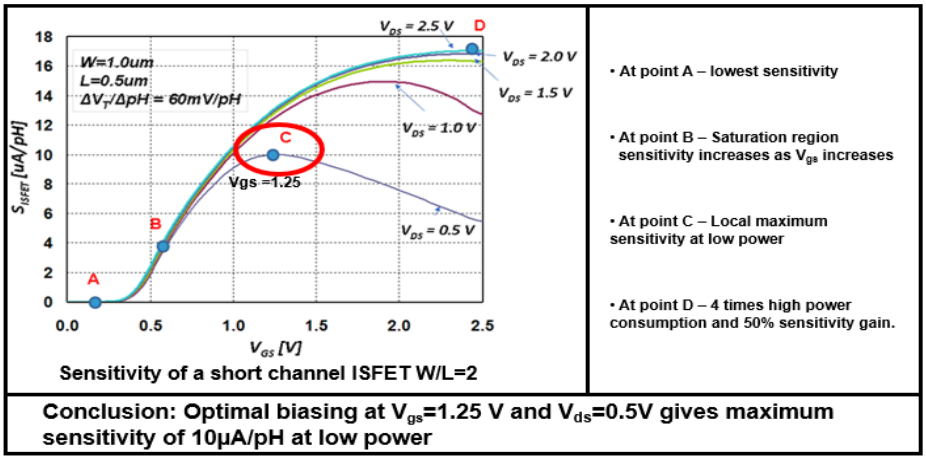


Figure 2.9: Sensitivity (S_{ISFET}) of a short channel ISFET with W/L = 2.

0.25μm technology model is shown in Figure 2.9, when L = 10μm and W/L = 2.

Due to mobility degradation effect on ISFET, the sensitivity doesn't stay constant for a fix V_{DS}. The short channel effects become more prominent when the channel length was reduced to L = 0.5μm as shown in Figure 2.9. Because of many issues in the short channel device, including, channel length modulation, velocity saturation and drain induced barrier lowering (DIBL), the sensitivity of ISFET degrades when the device was shrunk. For example, the maximum sensitivity of ISFET reduces from 37μA/ΔpH to about 17μA/ΔpH when the channel length reduces from 10μm to 0.5μm. The pros and cons of bias points A, B, C and D in Figure 2.9, are discussed as:

Bias point **A**: At this bias point ISFET operates in cutoff/ sub-threshold region of operation. The sensitivity from ISFET in this point is quite low. But the power consumption at this point is also very low. Therefore, this point is suitable for ultra-low power applications where higher sensitivity is not an issue.

Bias point **B**: Here ISFET operates in saturation region of operation and the non-idealities due to short channel effects did not take into effect. The sensitivity of ISFET around this region is a linear function of V_{GS} and hence sensitivity increases around bias point **B** as V_{GS} increases. However, the sensitivity of about 4μA//ΔpH may not be sufficient for many applications.

Bias point **C**: This point shows a local maximum of ISFET sensitivity. Beyond this bias point, the sensitivity degrades due to mobility degradation of carrier from short channel effects. We can bias ISFET to this sensitivity by applying V_{DS} = 0.5V and V_{GS} = 1.25V. The pH-to-current sensitivity at point **C** is near 10μA/ΔpH. This bias point gives us both high sensitivity and low power operation. This concept was used to design the ISFET sensor used in Prototype-2. For Prototype-1 a commercial chip was used, which was calibrated using the same concept (calibration is discussed in Chapter 3).

Bias point **D**: As we keep increasing of the voltage of V_{GS} and voltage of V_{DS}, IS-

FET enters the region where short channel effects i.e. Mobility degradation, DIBL etc. becomes so significant that the sensitivity of ISFET becomes saturated or degrades. For the given TSMCs 0.25μm technology node- this bias point gives us maximum sensitivity from ISFET sensor at a price of high power consumption i.e. four times higher power than point **C** for a sensitivity gain of 50% only.

Therefore, for low power operation with high sensitivity, we should operate ISFET sensor at point **C** with optimum voltage of V_{GS} and V_{DS}. But for application that requires higher sensitivity we must operate ISFET at bias point **D** where both V_{DS} and V_{GS} are close to maximum for a given technology node.

Novel pH-to-current readout circuit

The readout circuit is also an important part of the ISFET sensor. The circuit has two jobs: provide an appropriate bias (per the findings of previous section (**C**)) and readout the current, convert it into a voltage and amplify it.

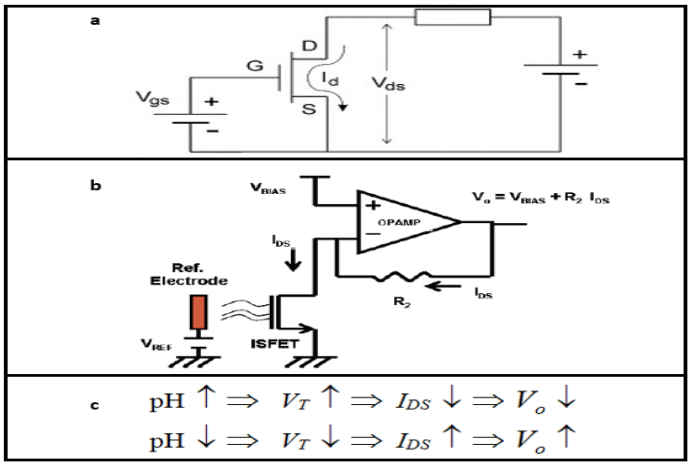


Figure 2.10: a) ISFET biasing circuit, b) ISFET biasing circuit with an op-amp and c) Process flow chart; change in pH conversion to change in current and change in current conversion to change in voltage.

To readout the current from pH-to-current ISFET, I use the circuit shown in Figure 2.10. Here, the gate of ISFET is biased by voltage source of V_{REF} and the drain is biased with the voltage source of V_{BIAS}, given from the analysis given in the

previous section. When a bio-chemical (pH change due to DNA sequencing) is applied at the floating gate of ISFET, the threshold voltage of ISFET gets changed. This threshold voltage changes results in a change of drain source current. The readout circuit converts this drain-source current change into an equivalent voltage change at output. The output voltage of the readout circuit is:

$$V_0 = V_{BIAS} + R_2 I_{DS} \tag{2.7}$$

Based on equation 2.5, we can amplify the readout voltage at output, V_0 by simply changing R_2 to a larger value. For a given bias condition of ISFET, we can describe this pH-to-current conversion based sensing with the flow chart presented in Figure 2.10 c.

Toumazou et.al[Wang 2012] introduced the concept of DNA sequencing by detecting change in pH on incorporation of nucleotides by DNA polymerase in complementary strand and release of H+ ion . This technology was further commercialized by Ion-torrent. Although, Ion-torrent has a very successful sequencing platform, ISFETs can further be leveraged for real-time genotyping applications owing to scalability, size and composition dependent sensitivity which can be tailored for specific applications. I utilized the same concept in developing the Prototype-1.

Prototype-1 ISFET Sensor and readout circuit

I used a commercially available pH sensor (strip-sensor) manufactured by Micropto. The specifications are given in Figure 2.11. Figure 2.11 is little crowded; it is presenting lot of information: Top picture presents the dimensions of the sensor fixture in millimeters, top table presents the pin output number and type, center picture on very right presents gate orientation, bottom table presents the specification of the sensor and bottom box presents the specifications for the op-amp circuit.

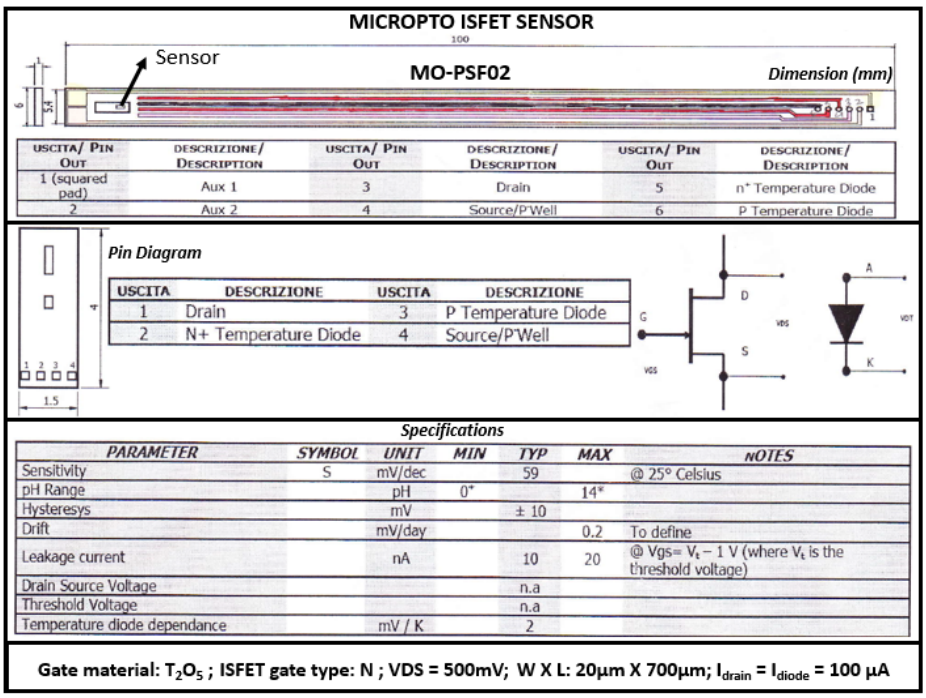


Figure 2.11: ISFET strip-sensor.

The op-amp based readout circuit is presented in Figure 2.12. I used op-amp 276 to biase, read, convert and amplify the signal (voltage). Any change in pH is converted to a change in output voltage. The voltage is digitized using National Instrument USB6001 DAQ. I wrote a LabVIEW app to read and plot the voltage in real time. I used TekPower 4 channel RS-1345 power supply to power the circuit. All the circuit component and voltage budget is presented in the Figure 2.12.

First I characterized the sensor for lab conditions and then the sensor was incorporated with fluidics system to perform the experiments. I performed the characterization using 3 pH solutions: 4, 7 and 10. The pH solutions (for calibration) were prepared using the HACH pH sensor and cross checked with Sentron. The calibration was done on several occasions (different days and time). The sensor was also tested for drifts over a long period of continuous sensing. After analyzing the data, a calibration slope of 74mV/pH was achieved with a signal magnification of 13X.

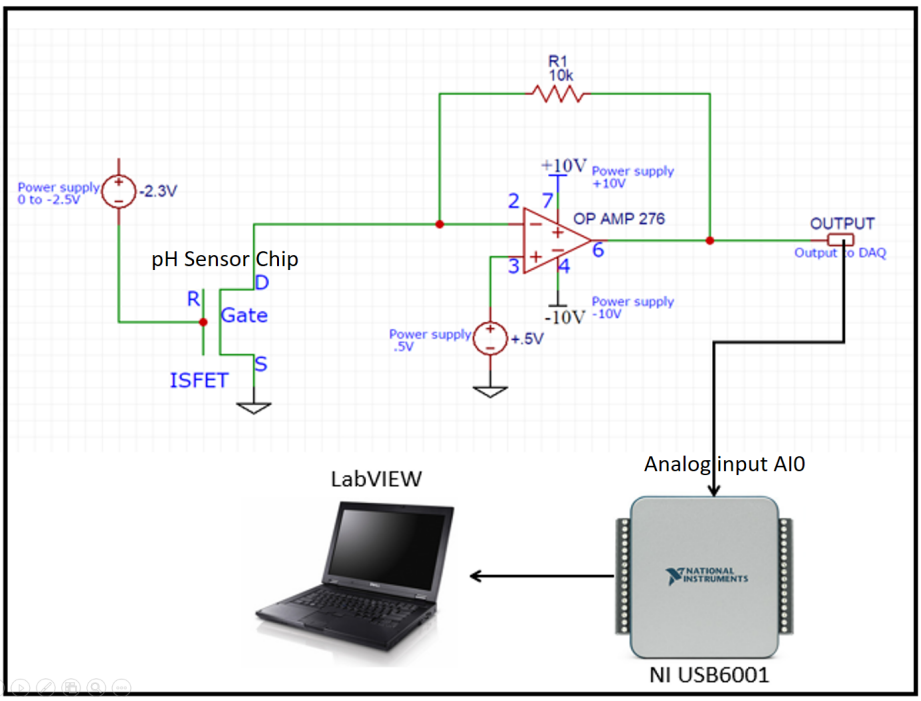


Figure 2.12: Op-amp based readout circuit. The output signal is digitized and collected using National Instrument DAQ and LabVIEW program.

2.3.2 Prototype-1 fluidics, control electronics and data acquisition

In this section I will discuss the fluidics, the electronics (which controlled the fluidics) and the data acquisition hardware.

Fludics

A major part of the fluidics system was the ISFET sensor holder which was used to hold the ISFET under a reaction chamber where fluids were inlet and outlet.

ISFET sensor holder:

Figure 2.13 presents the ISFET sensor and its holder. I designed a custom airtight holder for the sensor (it was not easy to incorporate the sensor in an airtight chamber where the reaction was taking pace). The holder had 3 parts: Reaction chamber, base and cover. The sensor was fixed in the horizontal groove in the base of the holder as shown in Figure 2.13. Above the ISFET chip around the sensor a square gasket was

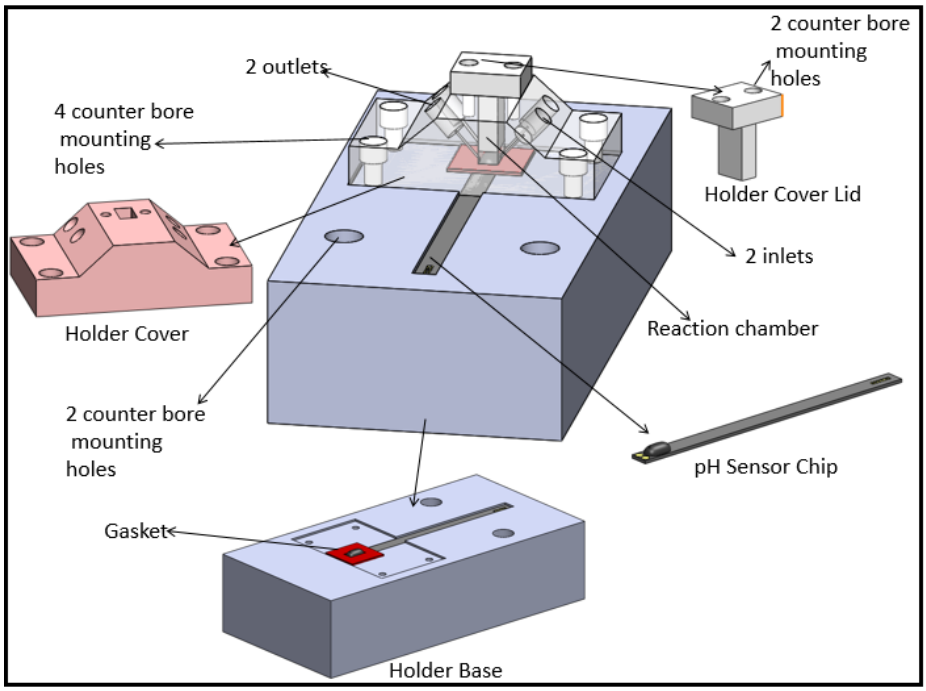


Figure 2.13: Different parts of the ISFET strip-sensor holder.

fixed to seal the reaction chamber from leaks. The gasket and chip was covered by a cover. The middle part of the cover was a rectangular volume-space which fitted right above the sensor, this volume was the reaction chamber as shown in the Figure 2.13. The cover was screwed tightly above the chip and gasket was sandwiched between the cover and chip making a tight seal. The chamber was sealed with a lid; the lid was designed such that the volume of chamber could be controlled and scaled if needed. A volume of 500μl was required for the experiment and testing. The block had 2 airtight inlets and 2 outlets. For the experiments, I used 1 inlet for the fluid, 1 inlet for the reference electrode, 1 outlet for the fluid and 1 for the Argon. The holder based was mounted on an aluminum breadboard with 1/4 inch screws, the cover was mounted on the base with 8/32 screws and lid was mounted on the cover with 4/40 screws.

Figure 2.14 presents the lab setup. The holder was made using PEEK (Polyetheretherketone) because it is chemically passive. The gasket, fitting and pipes were also chemically passive. The chamber was kept airtight filled with low pressured argon.

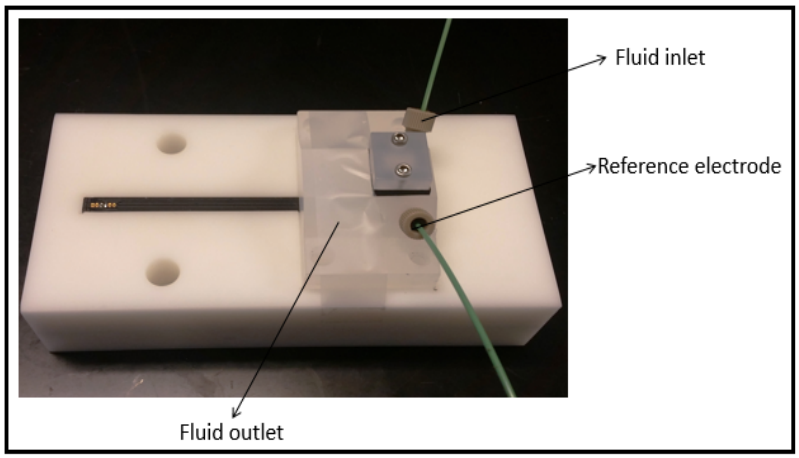


Figure 2.14: ISFET sensor holder lab setup.

A copper wire was used as a reference electrode.

Fluidics:

The fluidics was designed to target three requirements: The fluids must be delivered in controlled (well calibrated) and automated manner, the system must be scalable (additional resources/reagents could be added if required) and all the reagents (nucleotides, wash, cleave and argon) must be included as required by the experiments. Figure 2.15 presents the fluid flow diagram of a single sequence cycle.

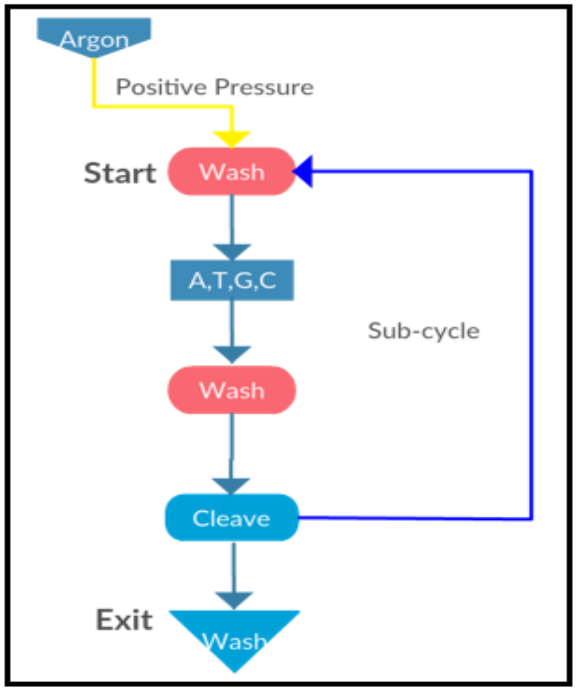


Figure 2.15: Fluid flow diagram for a single sequence cycle. In beginning a cycle starts with Wash (Argon is used to push the fluid) then a sub cycle is performed for each nucleotide.

Argon was used to create a positive pressure in the fluidics to push the fluid. Six different fluids (dNTPs A, C, G, T in buffer, buffer and cleave) and Argon were used in a sequence cycle and the number

of cycles depended on the number of the base pairs to be sequenced. One nucleotide was added per cycle. Each experiment had many number of cycles and each cycle had 4 sub-cycles as presented in the Figure 2.15. In the beginning of an experiment (first cycle) the wash was injected and then 4 sub cycles were completed as presented in the Figure 2.15 (Nucleotide-Wash-Cleave); Each sub-cycle began with a different nucleotide. The cycles were repeated depending on the number of base pairs to be sequenced. Each cycle and sub-cycle (in terms of functionality) was just an ON-OFF sequence of selection valves defined by the order presented in Figure 2.15. Figure 2.16

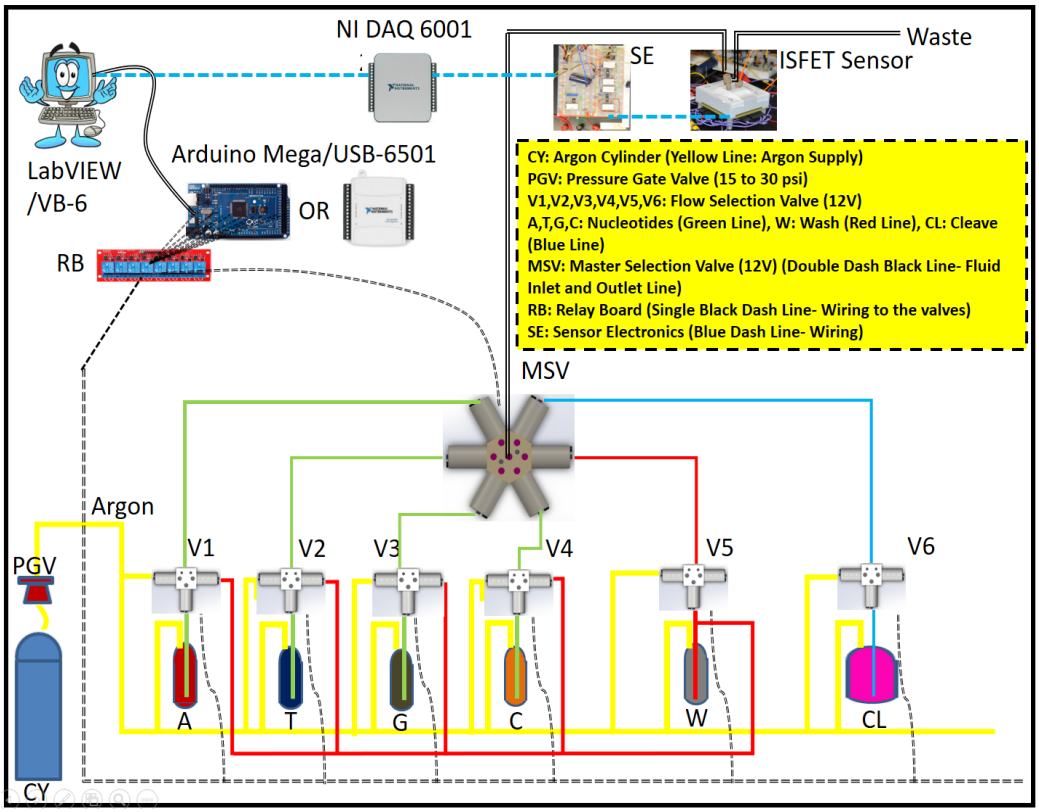


Figure 2.16: Prototype-1 and 2 schematics (fluidics). All the fluid channels are presented.

presents the schematics for the fluidics systems and Table 2.1 presents the different components of the Prototype-1 and 2. The fluidics system was a combination of two types of flow selection valves: 3way flow selection valves (3 valves in 1 valve; a valve-set) and a 6way flow section valve (6 valves in 1 valve). In Prototype-1, I used

pressured Argon (to push the fluid) and in Prototype-2, I used both a syringe pump and pressured Argon (both to push the fluid and syringe pump also to quantize the volume)).

Table 2.1: DNA Sequencing Prototype-1 and 2 component list.

Components	Quantity	Vendor	Comments
Mechanical Assembly			
Bread Board; MB1824	1	Thorlabs	Fluidics system base
Post; TR4	10	Thorlabs	Valve mounts
Post; TR8	4	Thorlabs	Bottle holder
Fluidics Assembly			
Tee w/ F-300 fittings; P-727	17	Idex	Material PEEK
Ferrule with SS ring; P-250	26	Idex	Material PEEK
Nut 1/4-28; P-255	26	Idex	Material PEEK
Tee pressure gauge; U-433	1	Idex	
Red tubing; 51085K41	1	Mcmaster Carr	1/16" Material PEEK
Yellow tubing; 51085K42	1	Mcmaster Carr	1/16" Material PEEK
Blue tubing; 51085K44	1	Mcmaster Carr	1/16" Material PEEK
Green tubing; 51085K48	1	Mcmaster Carr	1/16" Material PEEK
Brass pipe fitting; 50785K281	1	Mcmaster Carr	1/4" to 3/8"
100ml bottles; FB-800-100	4	Fisher Scientific	Material glass
1000ml bottles; FB-800-1000	2	Fisher Scientific	Material glass
Two port bottle caps; GL45	6	Fisher Scientific	Material glass
Flow selection Valve; 080T312-62	6	Bio Chem	1/16" Material PTFE
Flow selection valve; 080T612-62	1	Bio Chem	1/16" Material PTFE
Cavro XLP 6000 Syringe pump	1	Tecan	Only used in Prototype-2

There are many techniques for fluid injection, out of which one of the techniques is to pump pressurized air (Argon in our case) into a lab bottle that is sealed tightly

and connected to the reaction chamber. Pushing air into a lab-bottle will force the pressure in the bottle to increase and the pressure will force liquid from the lab-bottle to flow into the system. The lab-bottle approach is one used in Prototype-1 and 2 (Figure 2.16). This approach has some drawbacks though. The user does not know the amount of liquid that has been dispensed or how high the flow-rate is. The user would have to detect this using external sensors in feedback loops. If the system is a real-time system, then it must operate with the same conditions every time to perform the same operation. If the temperature changes and the channels are expanding the flow-rate is increased, and hence the user will lose track of the position of the liquid.

Since our system was to be used in the lab only (control environment) it was designed and calibrated for pressure vs time for a particular-volume approach. The temperature in the lab was very stable. I required a volume of 500μl for the reactions so I measured the flow-rate at 15 psi. The size (tubing size 1/16") and length of the tubing was fixed so it took 5 seconds to dispense 500μl of fluid. This was the calibration parameter: 100μl/second. Yellow line in Figure 2.16 is the Argon pressure line; it is connected to all the 3way flow selection valves. I used 3way flow selection valves (default close position) to inject the fluids into the system. The valve outlet could select between 3 inlets: A (or C, G, T), Wash and Argon. Valves V1 to V6 are 3way flow selection valves. V5 and V6 use only two inlets. The Argon pressure was controlled by the pressure gate valve at the Argon cylinder.

A 6way flow selection valve was used to select the which reagent (A, C, G, T, Wash or Cleave) would be injected in to the reaction chamber. The combination of 3way and 6way valve selection defines the fluid path. Both the valves were controlled independently and simultaneously. Figure 2.17 presents the selection of valves for 2 sub-cycles. For first sub-cycle valve-set V1s second value opens simultaneously with MSVs first valve for 5 seconds for A then these 2 valves close. The system incubates for 1 minute for the reaction to happen and to record the signal. Then the valve-

set V6s second valve opens simultaneously with MSVs sixth valve for 30 seconds for cleave then these two valves close after that valve V1s third valve and MSVs firth valve opens for 5 seconds for wash and then close and in the end V6s first valve and MSVs sixth valve opens for 5 seconds for Argon (to empty the system) and close. The process starts again for the next sub-cycle. The fluidics system I designed was

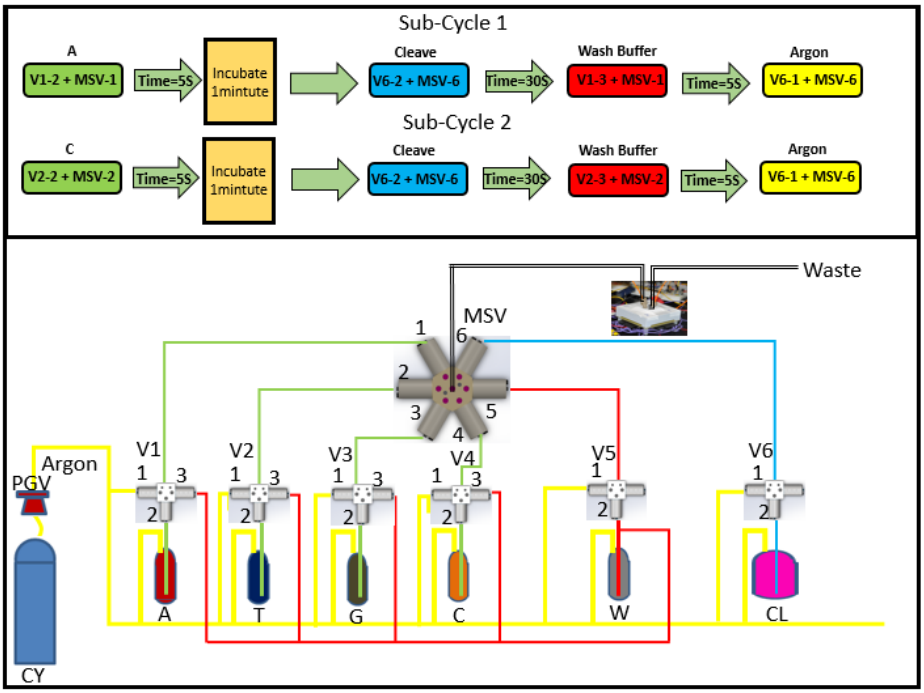


Figure 2.17: Top figure presents valve selection for two sub-cycles and bottom figure presents the valve numbering.

very straightforward and scalable. I used Thorlabs breadboard and posts to build the assembly. Breadboard was very useful because it provided full freedom in organizing the valves, tubing and bottles (see Table 2.1). I used the half-inch posts to make the holder for bottles and to mount the valves. Figure 2.18 presents Prototype-1. Breadboard kept everything sturdy, organized and clean. I also mounted the control electronics on the breadboard. The ISFET sensor holder was mounted on a different breadboard. The valves were controlled by the relay system which was controlled by the TTL-IO and the TTL-IO was controlled by the LabVIEW program. In the next section, I will discuss the control electronics.

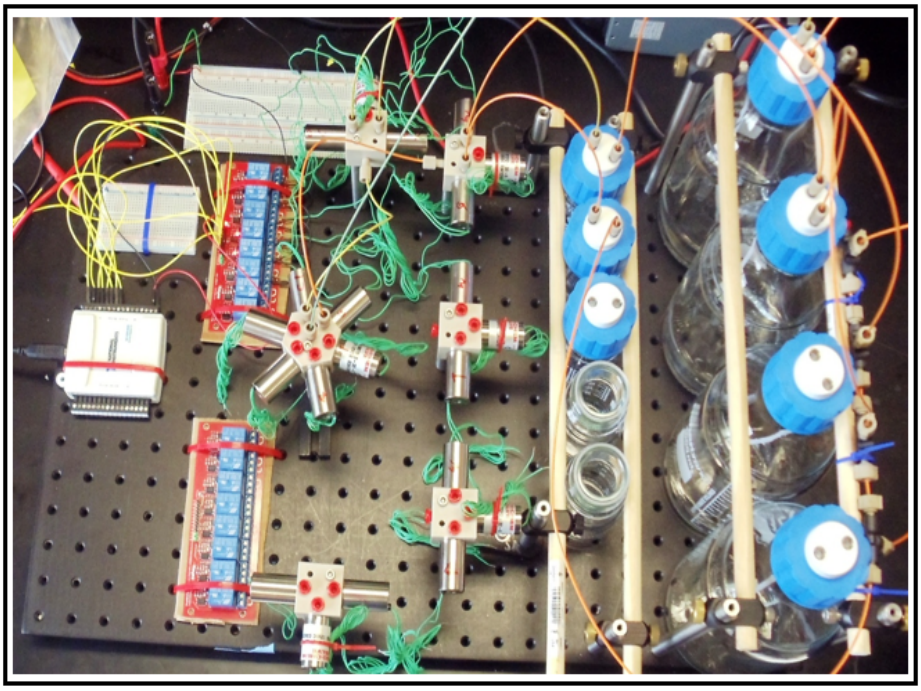


Figure 2.18: Prototype-1 fluidics: Mechanical assembly and Control electronics.

Control Electronics

In this section, I will discuss the control electronics. I used commercial off the shelf components to build the control electronics. Fluidics system required 24 independent valves plus the system had to be scalable so new valves could be included when needed. Table 2.2 presents the components.

Table 2.2: Control Electronics for Prototype-1 and 2 component list.

Components	Quantity	Vendor	Comments
DAQ USB 6501	1	National Instruments	Prototype-1 only
DAQ USB 6001	1	National Instruments	Prototype-1 only
Arduino Mega 2560	1	Sparkfun	Prototype-2 only
Analog Discovery (DAQ)	1	Digilent	Prototype-2 only
Keyes 8 Channel Relay Board	3	Sparkfun	12V
4 channel; RS-1345	1	TekPower	

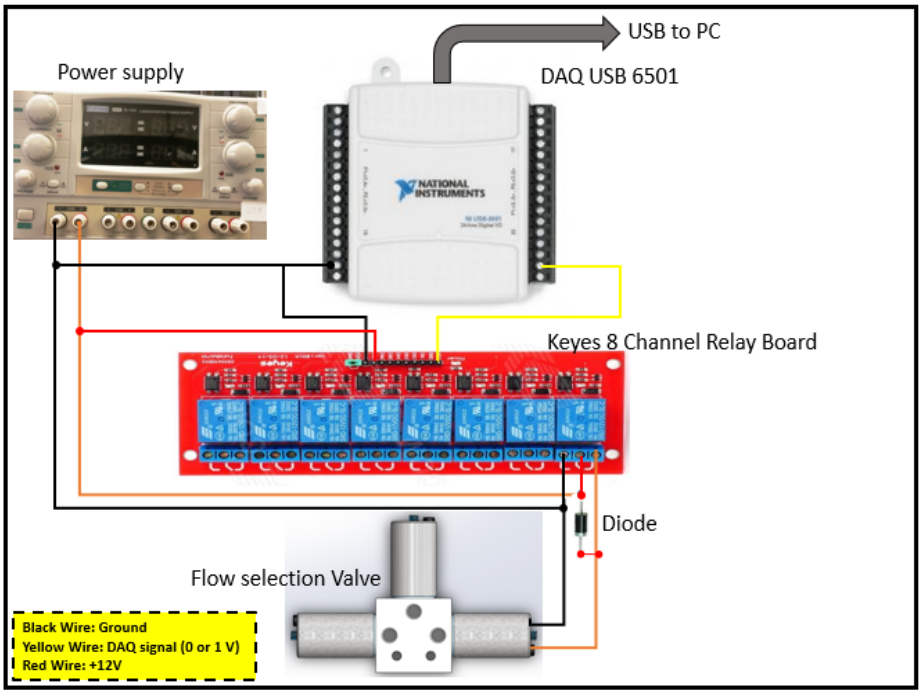


Figure 2.19: Wiring schematics for one valve for Prototype-1.

Figure 2.19 presents the wiring schematics for one valve. Similar wiring is applied to all 22 valves. I used 3 eight channel relay boards controlled by a single DAQ (TTL-IO) and power supply. The fluidics system was completely automated; controlled by a LabVIEW program. LabVIEW controlled National Instrument DAQ (USB 6501) and DAQ controlled the 8-channel relay board and the relay board controlled the valves using TTL-IO. In Prototype-1 I used 22 valves each operating at 12V and .15Amp. DAQ could not control the valves directly for two reasons: The valves were analog and DAQ was digital (TTL) and DAQ could only produce 0 or 5V at 8.5mAmp digital signal (ON-OFF). So, a level shifter (voltage) was needed.

DAQ USB 6501:

I used DAQ USB 6501 as an interface between the computer and valves. LabVIEW is a great graphical language. It is easy to write a scalable and complicated code in LabVIEW. National Instruments provide a vast range of hardware, which is easy to install and use. Considering all the good things LabVIEW and National Instruments (besides cost), I decided to use DAQ USB 6501. The DAQ provides 24 digital channels

(TTL-IO) which operate 0 or 5V at 8.5mAmp. The DAQ produced a signal which was either 0V (OFF) or 5V (ON). This signal was fed into the relay board. The DAQ was connected to the PC with a USB 2 cable.

Keyes 8 Channel Relay Board:

The valves required 1.8W (12V at .15Amp) of power to operate. DAQ couldnt handle such power so a level shifter was needed which could convert the DAQ TTL signal (5V) to an operating voltage (12V) for valve. And that is done by the relays, which operate at 12V. The board operates at 12V and takes the DAQ produced signal (either 0 or 1V) to shift the relay (mechanical relay) from OFF position to ON position (relay works as an ON/OFF switch) and vice versa. 22 valves were controlled by 3 relay boards (each relay board had 8 channels).

Diode:

The valves were a simple solenoid configuration with a stationary coil and moving magnet. When the current is applied to the coil it would generate a magnetic field which would make the magnet move against a spring to open the valve. When the current is removed, the magnet would move back (pushed by the spring) and close the valve. The moving magnet would generate a current (inductive kick) in the opposite direction of the primary current. If this current is allowed to flow into the circuit it could damage the relay board and DAQ (most probably DAQ; it is a very delicate device). A solution to this problem is to insert a diode between the two poles of the relay switch (in parallel with relay switch). The polarity of the diode should be such that it does not pass the current from power supply line to valve connector line but let the current pass in opposite direction (from valve connector line to power supply line). So, when the relay is ON the current does not flow through the diode, it flows through the switch and when the relay is OFF the current flows through the diode not through the switch (because it is Off; not connected). In this way, the solenoid is on a short while after the power is shut OFF.

Power Supply:

The components were powered by Tekpower RS-1345 power supply. The system had total 22 valves each consuming 1.8W of power. So, the total required power was 40W (ignoring the relay boards and considering all the valves were operating simultaneously) maximum. But all the valves were not operated simultaneously, only 2 to 3 valves were required at any time. RS-1345 power supply was more than enough to support the whole system.

Data Acquisition hardware and Software

In this section, I will discuss the data acquisition hardware and software. Software had two parts: Control and Automation software (used to control the fluidics through the control electronics) and Data acquisition software (used to acquire the data using data acquisition hardware).

Data Acquisition Hardware:

Dara acquisition hardware and schematics is shown in Figure 2.12. The output of the op-amp circuit was connected to the input (AI0) of National Instruments USB 6001 DAQ. The DAQ had 14-Bit resolution at 20kS/s: 14-Bit provided a resolution of .6mV, which was more than enough for our experiments. Our experiments were very slow so the speed of 20kS/s was way better than we needed.

Software: I wrote separate software for control (& automation) and data acquisition: Figure 2.20 presets the GUI for the software. Figure 2.20 a) present the GUI for fluidics control software. The software is in LabVIEW 2011. The software controls the DAQ (USB 6501) and DAQ controls the relay board and the relay boards control the valves and valves control the fluidics. All the valves can be configured independently for different ON/OFF timings.

Figure 2.21 b) presents the GUI for data acquisition software. The GUI reads the voltage from DAQ (USB 6001) and converts it into a pH (using the calibration

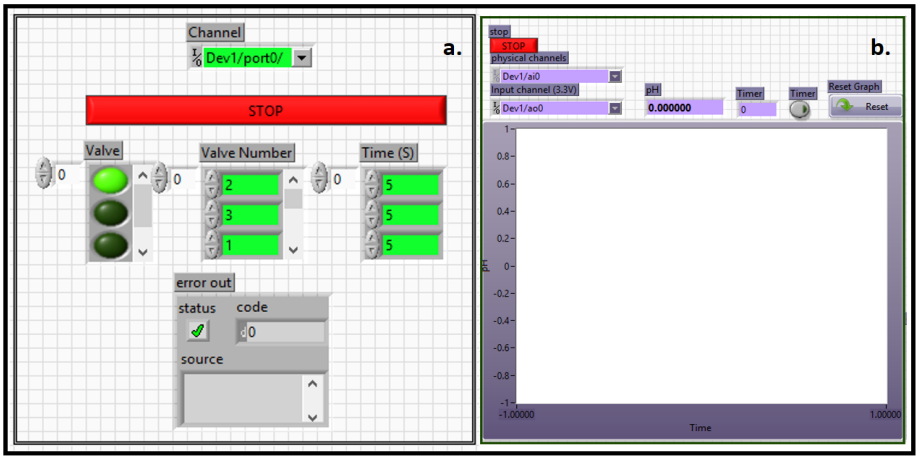


Figure 2.20: a) GUI for fluidics control software, b) GUI for data acquisition software.

curve) and plots it in the graph. The data could be acquired for any length of time but usually I recorded it for a minute or so.

2.4 Design and construction of DNA Sequencing Prototype-2

In this section, I will discuss the fluidics, the electronics (which controlled the fluidics) and the data acquisition hardware and software.

2.4.1 ISFET pH Sensor and novel pH-to-current readout circuit

The theory of ISFET sensor and novel readout circuit has already been discussed in section 2.3.1. In this section, I will discuss a new ISFET sensor [Mohammad 2015] we developed for the sequencing application.

Prototype-2 ISFET Sensor and readout circuit

A test chip was designed and fabricated using TSMC's 0.25μm technology. The test chip (4-core) consists of four sensing cores, shown in Figure 2.21. Each of the

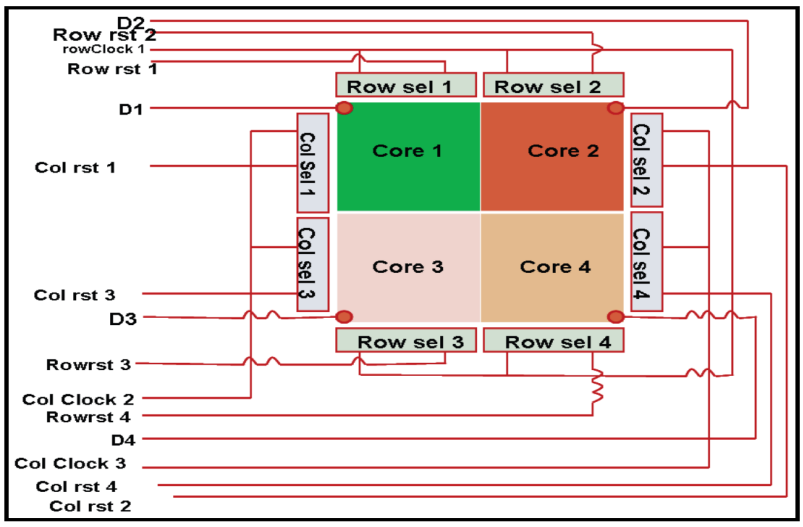


Figure 2.21: Architecture and wiring of 4-core ISFET sensor chip.

cores in test chip has 90×95-unit cells, which are accessed through column-select and row-select signals during readout process. The access transistors are designed to be large to ensure that it doesnt affect the measurement results. The cores contain four different unit cell specifications, which includes P-ISFET and N-ISFET for both small (with W/L=2.5) and as well as large (with W/L=24) ISFETs. The column-select and row-select registers for each core has its own reset signal to synchronize the readout scheme. This register works similar like a ring counter but with no feedback. The

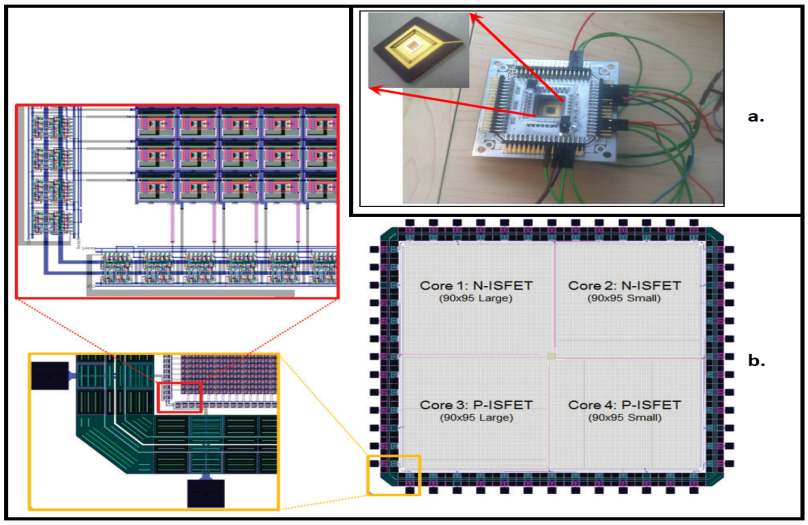


Figure 2.22: a) Presents wire-bounded ISFET sensor chip and b) Presents the layout of ISFET sensor chip including row and column select.

chip pad ring has electrostatic discharge (ESD) devices to protect the transistors in the core against electrostatic charge injection during chip handling. The test chip has a total of 52pins including 4 VDD, 4 GND, and several I/O signals. The layout of the chip is shown in Figure 2.22.

Novel pH-to-current readout circuit

The prospect of the current readout circuit has already been discussed in section 2.3.1.2 In this section I will briefly discuss the signal readout from the chip and its routing and amplification. The timing diagram of signals used to readout the chip

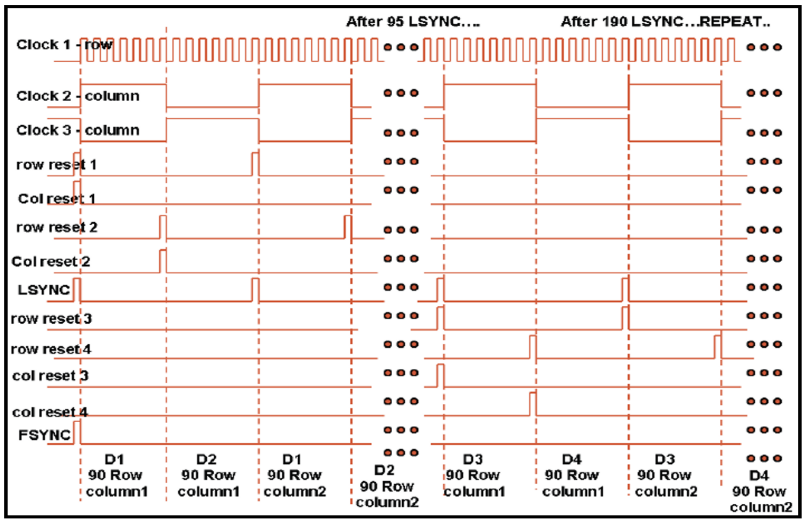


Figure 2.23: Signals timing diagram for chip readout.

is shown in Figure 2.23. There are eleven input signals of which 3 are clock signals and 8 are reset signals. The 4 data signals from 4 cores are the output signals. The readout process has been discussed in step by step, below:

1. Chip is biased with proper bias voltages VGS and VDS to ensure maximum sensitivity from the ISFET sensor, as described in section 2.3.1.1.

2. Row-select and column-select registers are activated with respective clock and reset signals to readout a unit cell in a sensing core. Column-select and row-

select registers generate column-select and row-select signals for each of the unit cell during readout process.

3. Finally, at the end of each row readout we activate an LSYNC signal and at the end of complete chip readout we activate an FSYNC signal. The LSYNC and FSYNC signals are used to synchronize the chip readout with image grabber.

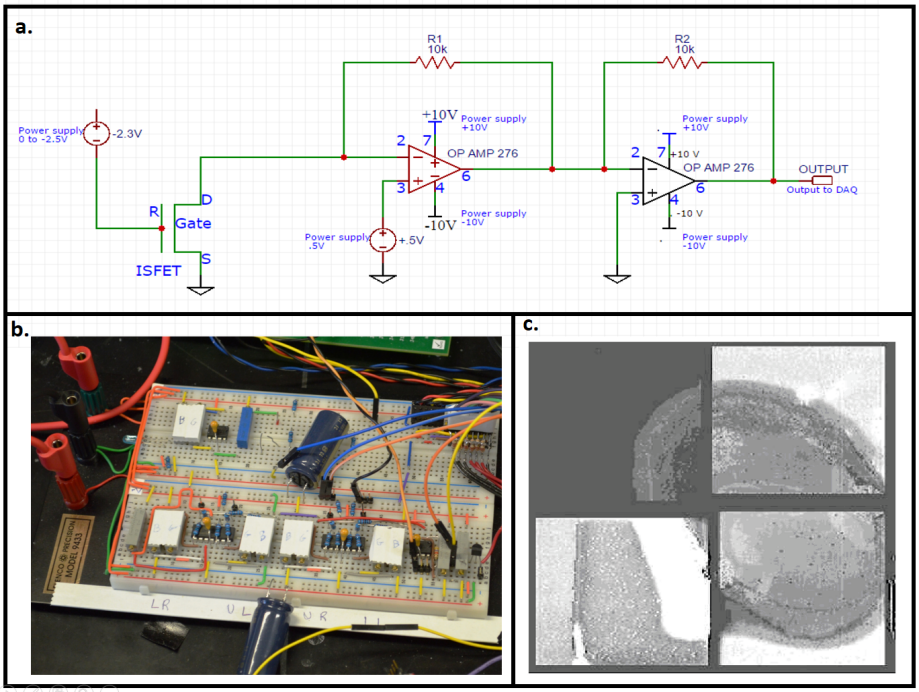


Figure 2.24: Figure **a** represents op-amp based readout circuit, Figure **b** represents the ISFET core readout circuit and op-amp amplifier and Figure **c** image generated as a result of pH sensing by ISFET chip.

With the application of signals in Figure 2.23, we then readout the data from the four cores and pass the time multiplexed video data through a single channel to an image grabber card. The image grabber displays (Figure 2.24 c) the ISFET sensor output with the help of clock, LSYNC and FSYNC signals. As we are reading 20 frames/sec from the chip, the image at display shows the real-time response of the pH interaction with the ISFET sensing gate surface.

Figure 2.24 represents the biasing, current readout and amplification circuit. Each core of the ISFET senor chip was handled by a separate circuit (as shown in the figure

2.24 b). The signal for each core was amplified separately and recoded by a Digilent Analog Discovery DAQ using a visual basic program

2.4.2 Prototype-2 fluidics, control electronics and data acquisition

Fluidics

The fluidics of Prototype-2 was like Prototype-1, the only addition was a syringe pump and a different custom ISFET sensor chip holder as shown in the Figure 2.25 b. The

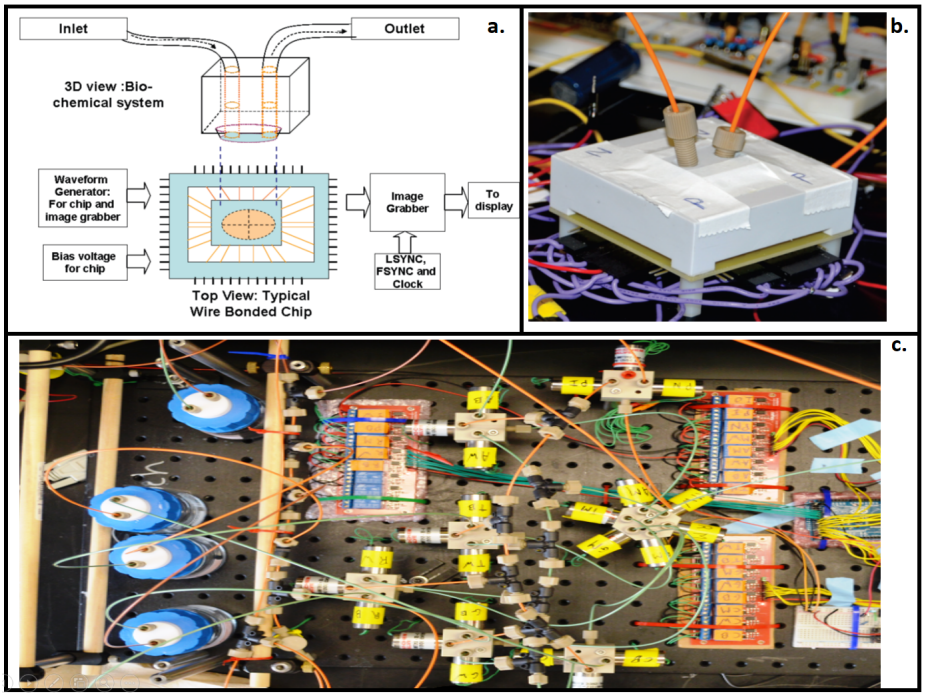


Figure 2.25: Figure **a** presents the setup showing mounted fluids system inlet and outlet and signal input and output from the chip, Figure **b** presents the sensor holder and Figure **c** presents the fluidics system.

ISFET sensor chip is shown in Figure 2.22 a. A custom cover-block was designed to incorporate the fluidics into the chip. The cover-block had an inlet and outlet which had its openings right above the chip. The reaction chamber was made by an elliptical airtight seal sandwiched between the chip base and cover-block (Figure 2.25 b). The fluidics schematics is presented in Figure 2.26. It is very like the Prototype-1. The

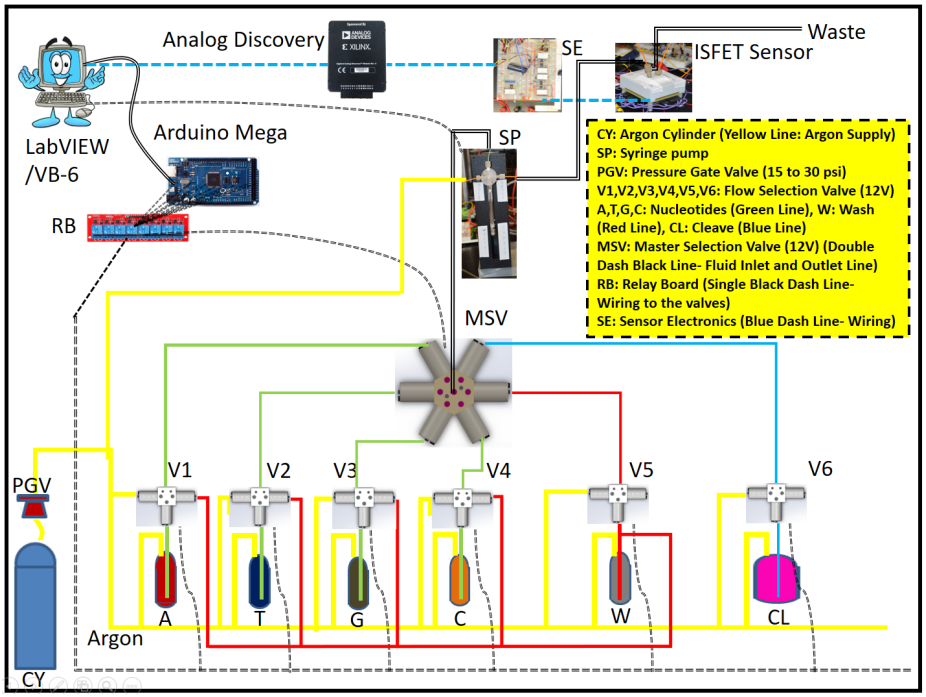


Figure 2.26: Prototype-2 schematics (fluidics). All the fluid channels are presented.

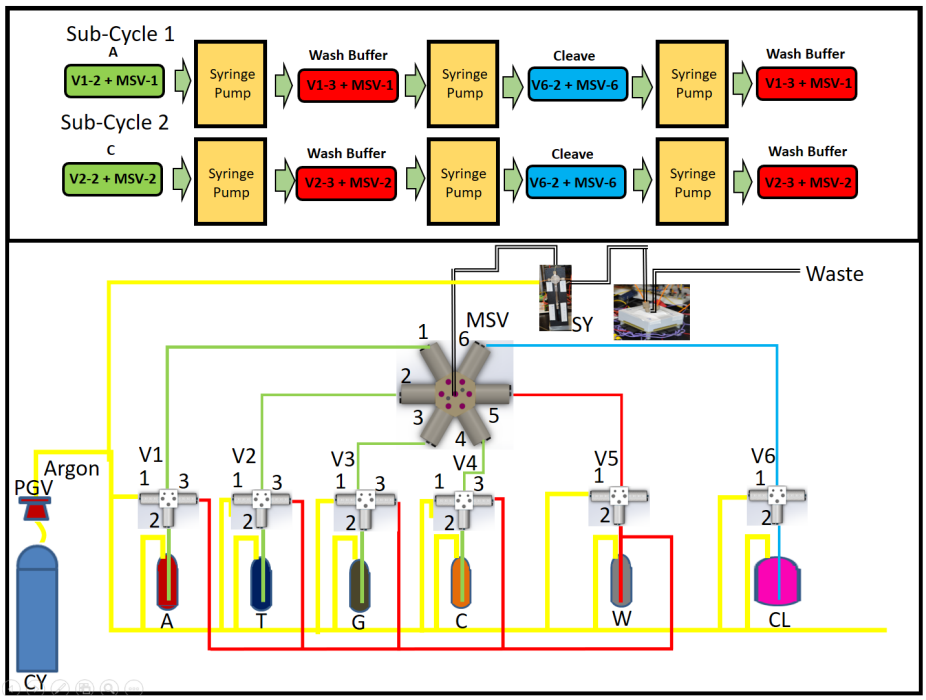


Figure 2.27: Top figure presents valve selection for two sub-cycles and bottom figure presents the valve numbering.

only addition was a syringe pump to control and automate the flow with accurately measuring the volume of 500μl. The computer directly controlled the syringe pump.

In Prototype-2, also a positive pressured Argon (to push the fluid) was used.

Figure 2.27 presents the selection of valves for 2 sub-cycles. For first sub-cycle valve-set V1s second value opens simultaneously with MSVs first valve then syringe pump operates and pushes the fluid for A. Reagents are incubated and signals are recorded simultaneously. Then the valve-set V1s third valve opens simultaneously with MSVs first valve then syringe pump operates and pushes the fluid for wash after that valve-set V6s second valve and MSVs sixth valve opens then syringe pump operates and pushes the fluid for cleave and then close and in the end V1s third valve and MSVs first valve opens then syringe pump valve opens and pushes the wash and close. The process starts again for the next sub-cycle.

Control Electronics

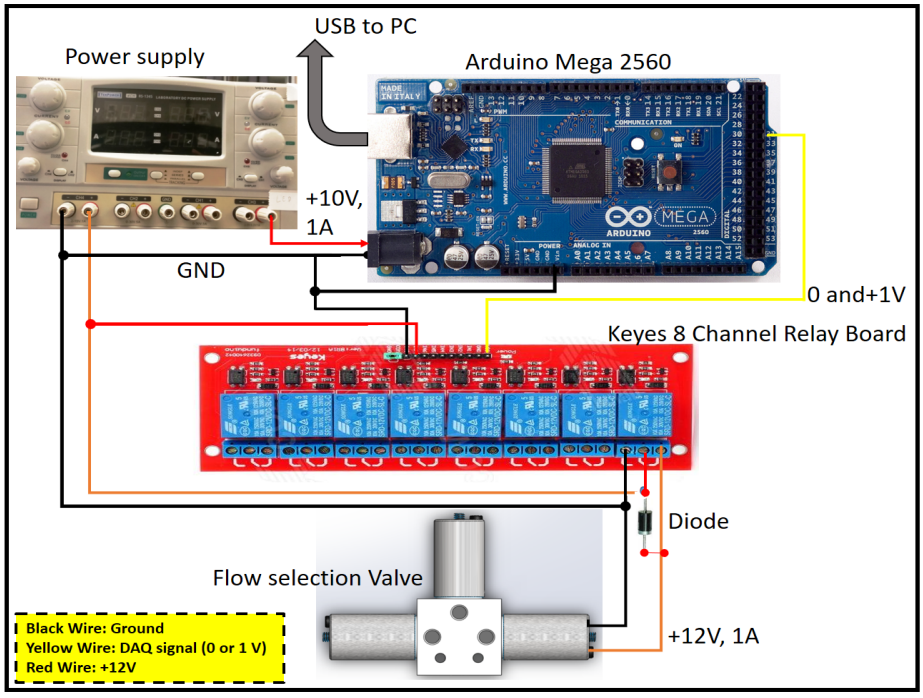


Figure 2.28: Wiring schematics for one valve in the Prototype-2.

In this section, I will discuss the control electronics. I used commercial off the shelf components to build the control electronics. Fluidics system needed 24 independent valves plus the system had to be scalable so new valves could be included when

needed. Table 2.2 presents the component list for the Prototype-2 and Figure 2.28 presents the wiring schematics for Prototype-2.

Figure 2.28 presents the wiring schematics for one valve. Similar wiring is applied to all 22 valves. I used 3 eight channel relay boards controlled by an Arduino Mega 2560 microcontroller and power supply. The fluidics system was completely automated; controlled by Visual Basic 6. Visual Basic controlled Arduino Mega 2560 and Arduino Mega 2560 controlled the 8-channel relay board and the relay board controlled the valves. In Prototype-2 I used 22 valves each operating at 12V .15Amp. Microcontroller could not control the valves directly for two reasons: The valves were analog and microcontroller was digital and it could only produce 0 or 5V at 20mAmp (TTL-IO). So, a level shifter (voltage) was needed.

Arduino Mega 2560:

The Arduino Mega 2560 provides a 5V power output, 56 configurable digital I/O pins and can be programmed via an USB connection to the computer. It can be used as independent controller or can be used as an interface to power the hardware controlled by Visual Basic, LabVIEW or MATLAB. The board can be powered by a 7-12V DC adapter or battery; I used the same power supply to power all the electronics and valves so the ground stays the same hence no ground loop current.

The Arduino board runs at a clock frequency of 16MHz. However, the clock can only be checked by the Arduino sketch every four micro seconds or with a clock frequency of 250kHz. This means that any program running on an Arduino cannot be controlled to a higher precision than 4 micro seconds. These features suited best to our requirements for the system. All other components of the control electronics were like the Prototype-1 which has already been discussed in section 2.3.2.2.

Data Acquisition hardware and Software

In this section, I will discuss the data acquisition hardware and software. Software had two parts: Control and Automation software (used to control the fluidics through the control electronics) and Data acquisition software (used to acquire the data using data acquisition hardware). Unlike the Prototype-1 I used a single software to control the fluidics and data acquisition in Prototype-2.

Data Acquisition Hardware:

Data acquisition hardware and schematics is shown in Figure 2.29. Data acquisi-

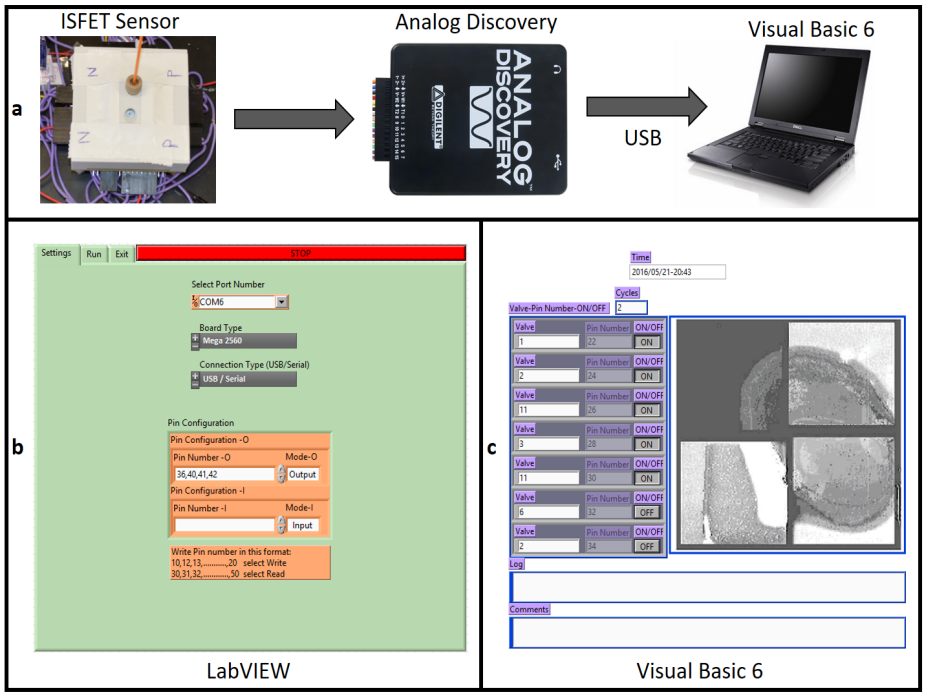


Figure 2.29: a) Data acquisition hardware, b) LabVIEW fluidics test GUI and c) Visual Basic 6 data acquisition software GUI.

tion hardware is presented in Figure 2.29 a. I used Analog Discovery 100MS/s USB Oscilloscope & Logic Analyzer with 14-Bit resolution. 14-Bit provided a resolution of .06mV, which was more than enough for our experiments. A high data acquisition speed was required because we were acquiring 20frames/sec, DAQ could acquire data at 100MS/s which was adequate for our experiments.

Software:

First, I wrote a test software in LabVIEW to check Arduino compatibility with valves, electronic requirements and functioning (testing the pin settings with valve settings), this is presented in the Figure 2.29 b. Once the fluidics was working I moved the control and automation software from LabVIEW to Visual Basic 6. Figure 2.29 c presets the GUI for the software. The software controls the Arduino Mega 2560 and microcontroller controls the relay board and the relay boards control the valves and valves control the fluidics.

Figure 2.29 c presents the GUI for data acquisition software. For readout and synchronization of the chip, the signals shown in Figure 2.23 were used. From the readout of the chip, four data signal - D1, D2, D3 and D4 were generated as output from four cores. An adder was used to add the four-time multiplexed data signals and to pass it through a single channel to the image grabber. The image grabber then digitized the data through the LSYNC, FSYNC and row clock signal, for the external display. The actual test setup with wiring of the chip is shown in Figure 2.25 b. The responses that we had from four cores were different as these four cores had different design dimensions and hence different response to the same pH change from the application of Rolonies in the ISFET sensing gate surface. A typical image generated by our ISFET arrays test chip is presented in Figure 2.29 c.

2.5 Design and construction of DNA Sequencing Prototype-3

In this section, I will discuss the optics, the fluidics, the electronics (which controlled the optics and fluidics), and the data acquisition hardware and software of Prototype-3. The Prototype-3 was completely a different approach from Prototype-1 and 2, where Prototype-1 and 2 were based on the pH sensing based DNA sequencing, Prototype-3 was based on light detection based DNA sequencing by SBS (SBS:

Sequencing by Synthesis).

DNA sequencing has revolutionized the world of precision medicine[Gonzalez2015]. Targeted sequencing of just protein coding regions enables sequencing of disease causing genes, and also unravel mutations [Gonzalez2015]. Among most successful technologies currently , Illuminas sequencing-by-synthesis approach has almost 98% of the market world-wide [Guzvic 2013]. Sequencing by synthesis as first used by Solexa in 2006 [Liu 2012].Solexa was later acquired by Illumina and has been a market leader in sequencing world-wide. SBS is widely accepted due to its accuracy by natural competition through use of 4 unique reversibly bound nucleotides [Chen 2013]. SBS has allowed for both short-read and long-read sequencing for paired-end. Recent developments with 2-channel SBS accelerate sequencing and data processing times by use of two colour fluorophores and reducing scanning time [Goodwin 2016] but costs of sequencing still remain high. Although, Illumina has made possible the $1000 genome sequencing and use in clinical setting [Veritas Genetics], sequencing in higher number of samples simultaneously and high cost of instrumentation still limits its wide-spread use in developing world. This dissertation proposes a prototype for in-house built sequencer which is comparatively cheaper than the current commercial sequencers, which can be customized as per user requirements and effectively uses 2-channel sequencing SBS chemistry for targeted re-sequencing of Inherited disease panels.

Genetic diseases are major economical burdens in developing countries like India [Balgir 2000], Pakistan, Africa, Middle-east and many more. Population based genetic pre-screening of population to ensure birth of healthy off-springs needs affordable and available DNA-Sequencing resources. This prototype can be assembled and used effectively for whole-genome sequencing or targeted sequencing, making it scalable as per requirements of the project. As Illumina uses expensive surface-chemistry to immobilize the clusters for sequencing, this sequencer works with cheap glass surfaces

modified chemically for library immobilization and used for sequencing. Lasers used for optical detection in Illumina system are replaced by light emitting diodes (LEDs) which are cheaper than lasers. The electronics and fluidic system comprises of cheap and scalable components. This system can be customized as per user requirements and needs.

Basic Principle:

2μm diameter acrylamide beads are immobilized with DNA-Sequences to be sequenced, by clonal amplification by emulsion-PCR (Chapter-3). Biotinylated sequencing primer is then hybridized to the beads, which in turn aids in immobilizing the acrylamide beads to the streptavidin coated surface. The glass slide with coated surface is assembled into the flow-cell (or pre-assembled before coating in 4-channel). The flow-cell is washed with wash buffer (1E). Extend reagent (nucleotides, reversible-terminator fluorescently labelled dNTPs, enzymes and buffer) is loaded into the flow cell at 45^{o}C-50^{o}C temperature. Reagents are incubated for 90sec in the flow-cell. After incubation, it is washed and images (of fluorescently labeled dNTPs) are taken manually by exposing for 300msec to LEDs. After taking pictures and looking at image quality, any adjustments are made for optimizing the exposure times, frame positions, reagent volumes and incubation timings. Another optimization step is done after cleave is incubated in the flow-cell to cleave the reversibly incorporated nucleotides, after cleave fluorescence should be zero. The basic idea of light detection based DNA sequencing by two channel SBS is presented in Figure 2.30. Two channel corresponds to the two wavelengths (red and green) used to excite the two fluorophore dyes (Cy3 and Cy5) at 550nm and 650nm. The whole process goes as following: The 2μm diameter acrylamide beads (immobilized with DNA-Sequences to be sequenced) are exposed with white light from halogen lamp and a bright field image is captured. This image works as a reference image. The positions of bead clusters are identified and recorded as X-Y coordinates.

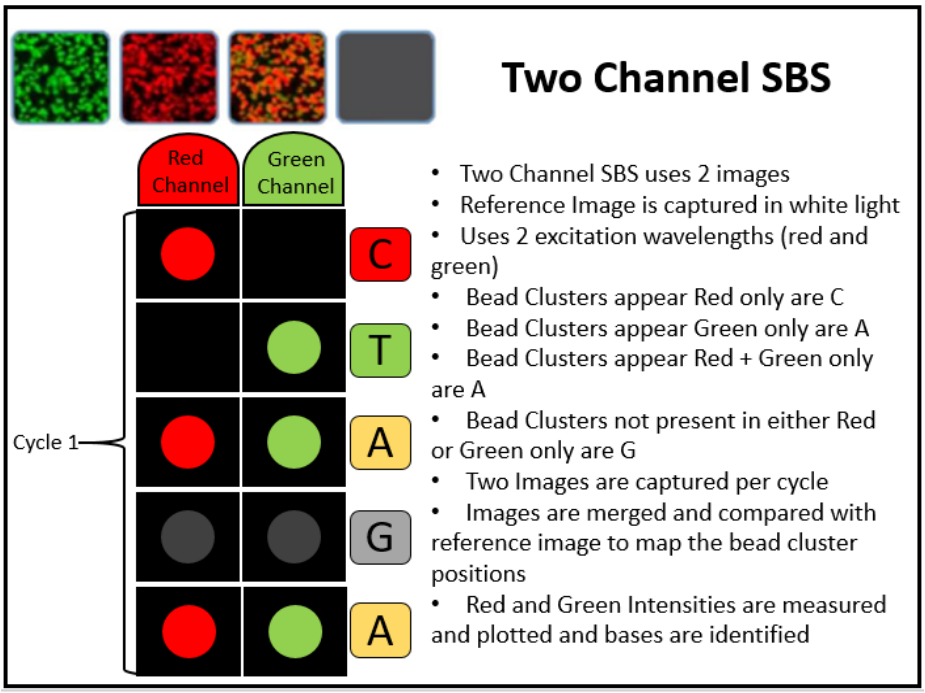


Figure 2.30: Light detection based DNA sequencing by Two Channel SBS.

White light is turned off and beads are exposed with red wavelength for 300ms to excite Cy3 then a red image is captured, same process is repeated for green light. In the end two images are saved, one for red and one for green. A single image is produced by merging red and green images for processing. Later, during image processing the red and green images are compared to the reference image and the bead intensities are recorded and plotted for the bead positions. The DNA bases are identified based on the intensities and positions. The positions where only red is present is identified as only C, the positions where only green is present is identified as only T, the positions where red and green is present is identified as A and the positions where neither red nor green is present is identified as G. This process runs for all the base pairs and this is how all the base pairs are sequenced. With this scheme, many different DNA sequences can be sequenced in parallel. In a single cycle, all 4 nucleotides are sent and incorporated and sequenced unlike the Prototype-1 and 2 where only one nucleotide was sent at a time. This is way faster and simpler than the pH based sequencing. And this is way the fluidics, electronics and data acquisition and control was much

simpler in Prototype-3.

2.5.1 Optics

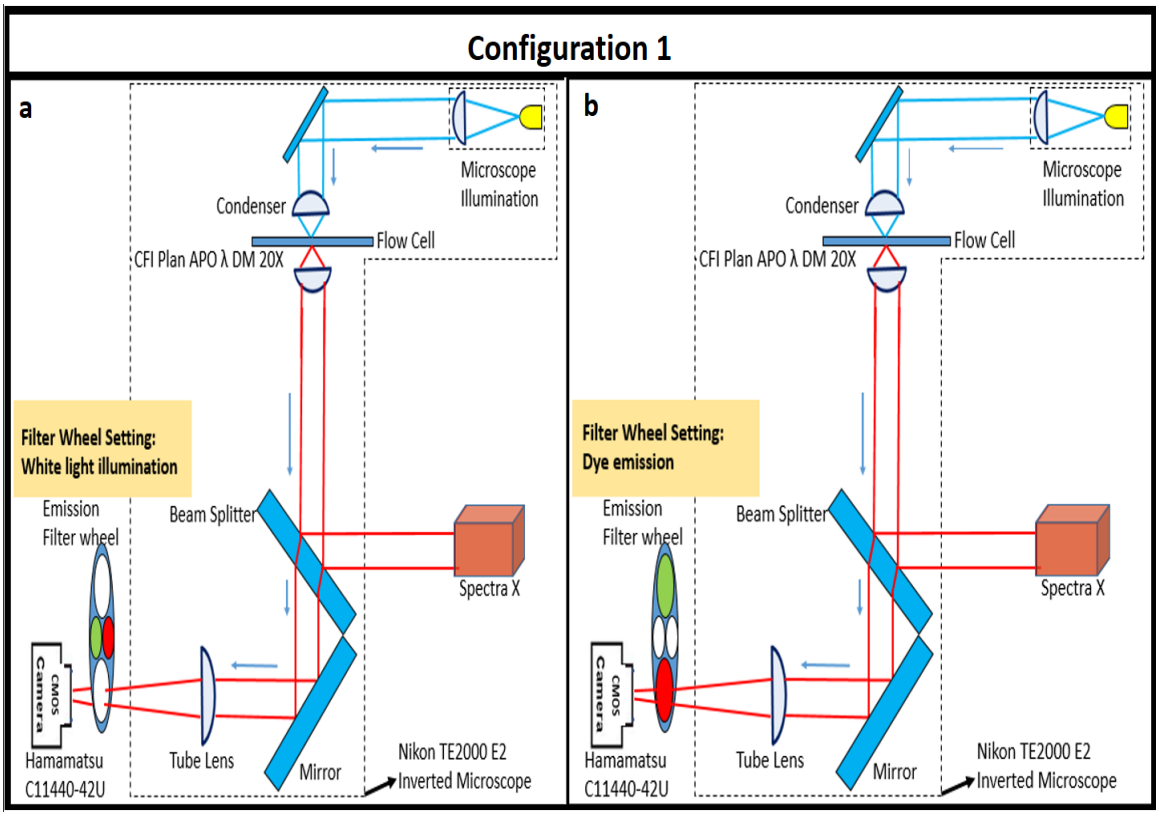


Figure 2.31: Optical path of Prototype-3. Configuration-1: a) Presents optical path during white light illumination, b) Presents optical path during dye emission.

We used two different configurations (1 and 2) for Prototype-3. The optical path was similar in both the configurations. First configuration was designed to use single channel flow cell and second configuration was designed to use 4. The optics and optomechanics of Prototype-3 was very simple, a schematic of the optical path in configuration-1 is presented in Figure 2.31. The configuration-1 was built using Nikon inverted microscope TE2000-E2 and configuration-2 was built using Nikon inverted microscope Eclipse Ti-E. A full list of optical components (configuration-1 and 2) is presented in Table 2.3.

Table 2.3: Optical and Optomechanical components of Prototype-3.

Components	Quantity	Vendor	Comments
Optical			
Inverted Microscope; TE2000-E2	1	Nikon	Configuration-1 only
Inverted Microscope; Eclipse Ti-E	1	Nikon	Configuration-2 only
CFI Plan Apo DM 20X; MRD30200	1	Nikon	.75 NA Objective only
Condenser Lwd Lens; MEL36200	1	Nikon	Diaphot
X-Cite 120 LED Boost	1	Excelitas	Ultra bright LEDs
Spectra X	1	Lumencor	Light Engine
FF01-531/40-25	1	Semrock	Green
ET560/20x	1	Chroma	Green
ET660/40	1	Chroma	Red
ET640/30x	1	Chroma	Red
C11440-42U	1	Hamamatsu	CMOS Camera
Andor ixon +	1	Andor	CCD Camera
Optomechanical			
T-FLMC Cassette Holder; MEV51100	1	Nikon	Motorized Filter Changer
Lambda 10-3	1	Sutter	Filter Changer Driver
Linear Motor Stage; HLD117	1	Prior	X-Y Stage
ProScan II	1	Prior	X-Y Stage Driver
Piezo Stage; Z piezo	1	PI	Z-Focus
Piezo Amplifier; EE665-CR	1	PI	

Nikon inverted microscopes were chosen because they provided a great flexibility and quality. Inverted microscopes were ideal for our system because it was easier with inverted configuration to build a customized flow cell setup right over and between the objective and condenser lens. Nikon microscopes provide full flexibility to build customized optomechanical instrumentation around and within them. Nikon provides an inlet in the back of the microscopes for any addition light sources. Also, microscopes have sufficient space under the objective to mount a filter wheel and Z-piezo stage

without blocking any normal functionality. The overall structure of the microscopes is very sturdy which is ideal for our system.

Configuration-1 White Light Optical Path:

The optical path was similar in configuration-1 and 2, only few components were different and arranged differently. In configuration-1 Nikon inverted microscope TE2000-E2 was used with Spectra X Light Engine for dye excitation and Hamamatsu C11440-42U CMOS camera was used for imaging. The emission filters were installed directly before the camera in an automated filter changer cassette. The optical path of prototype-3 configuration-1 is presented in Figure 2.31 a. Two light sources were used to illuminate the flow cell: Spectra X Light Engine and Halogen lamp (White light). First time when the flow cell was mounted, Halogen lamp was used to adjust its X-Y and Z position using X-Y stage (Prior) and manual Z (microscope focus knob), looking through the 10X Nikon Eyepiece. Then a good spot was found (where the bead clusters looked most promising for experiment) by scanning the flow cell using the X-Y stage, this position was recorded. Once the position was recorded Z auto focus was initiated (a PI z-piezo stage was used) and its Z position of best focus was also recorded. The halogen lamp was in the top part of the microscope. The light from halogen lamp was collected by collector-lens (Nikon) and focused into the image plane by Nikon condenser lens. Both collector lens and condenser lens were the generic parts came with the microscope. The light through the image plane was collected by an infinite corrected 20X .75 NA Nikon objective with 1mm working distance. The objective and Hamamatsu CMOS camera provides a FOV of 665m X 665m. The collected light went through the beam splitter and reflected of the mirror (inside the microscope) and passed through the filter changer and focused by the tube lens on the CMOS plane of the camera and a bright field image of the sample plane was saved. This Image would serve as reference image in image processing. The bright field image contained all the information regarding the beads positions

and frame position in X-Y and Z.

Configuration-1 Dye Excitation and Emission Optical Path:

We used Cy3 and Cy5 dye (at wavelength of 550nm and 650nm) for green and red. Both the wavelengths were produced by the broad band Spectra X Light Engine. We did not use any excitation filters for this arrangement. The optical path is given in Figure 2.31 b. The optical path of excitation and emission was similar. After bright field image the halogen lamp was turned off and Spectra X was turned on for 300ms. The light from Spectra X outputted through a liquid light fiber cable. The fiber cable was terminated in front of a collimator lens (not shown). The collimated beam then directed through the back of the microscope to the beam splitter (inside the microscope) and reflected of towards the back aperture of the objective. The objective focused the light in the image plane and excited the dye Cy3 and 5. The emission light from the dye and the reflected light was again collected by the objective and directed through the beam splitter and reflected of the mirror towards the tube lens. After tube lens light went through the emission filters one by one (for green and red) and imaged at the camera plane. Two Images were saved one of green and one for red. The filters were changed automatically by the automated filter changer controlled by the Nikon Image Software.

Configuration-2 White Light Optical Path:

The optical path of configutation-2 is very like configuration-1 and presented in the Figure 2.32. In this configuration, we replaced the TE2000-E2 with Eclipse-Ti-E and replaced the light source Spectra X with X-Cite 120 LED Boost and placed the filter changer under the objective instead in front the camera. We used two sets of excitation and emission filters for red and green. We replaced the Hamamatsu CMOS camera with Andor ixon + CCD camera. We could use four cell flow chamber in this configuration. The optical path of Prototype-3 configuration-2 is presented in Figure 2.31 a. Two light sources were used to illuminate the flow cell: X-Cite 120 LED

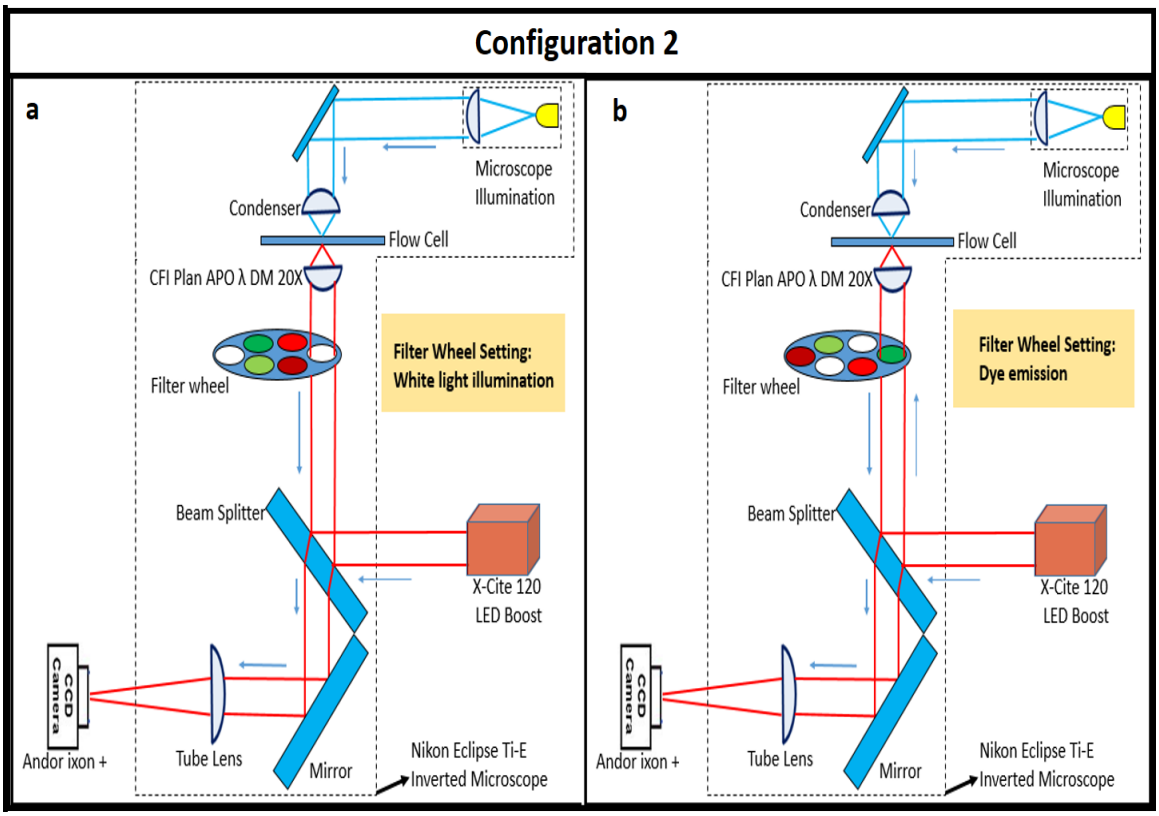


Figure 2.32: Optical path of Prototype-3. Configuration-2: a) Presents optical path during white light illumination, b) Presents optical path during dye emission.

Boost and Halogen lamp (White light). The process of mounting the flow cell and optimization of settings is same as discussed previously: The halogen lamp was in the top part of the microscope. The light from halogen lamp was collected by collector-lens (Nikon) and focused into the image plane by Nikon condenser lens. Both collector lens and condenser lens were the generic parts came with the microscope. The light through the image plane was collected by an infinite corrected 20X .75 NA Nikon objective with 1mm working distance. The objective and Andor ixon + CCD camera provides a FOV of 650μm×650μm. The collected light went through the filter wheel (where no filter was installed) then beam splitter and reflected of the mirror (inside the microscope) and focused by the tube lens on the CCD plane of the camera and a bright field image of the sample plane was saved.

Configuration-2 Dye Excitation and Emission Optical Path:

We used same dyes Cy3 and Cy5 for this configuration. Both the wavelengths were produced by the excitation filters installed in the filter wheel under the objective. The optical path is given in Figure 2.32 b. The optical path of excitation and emission was similar but different filters were used for incoming excitation and outgoing emission lights. After bright field image the halogen lamp was turned off and LED Boost was turned on. The light form LED Boost was collimated and coupled through the back of the microscope to the beam splitter (inside the microscope) and reflected of towards the filter wheel. The filter wheel was automatically controlled using Sutter driver. First excitation filter for dye Cy3 was inserted in the optical path and dye was exposed for 300ms. The objective focused the light (from the filter) in the image plane and excited the Cy3. The emission light from the Cye3 and the reflected light was again collected by the objective and directed through the filter wheel, now the emission filter was inserted in the optical path. The emission filter cut off all the wavelengths and passed only the emission wave length ($\sim$ 550nm). The light then passed through the beam splitter and reflected of the mirror towards the tube lens. Tube lens focused the light on the CMOS plane and the image was recorded. Same process was repeated for Cy5. And in the end two images were saved for red and green. This process was repeated for all the cycles and a flow chart of the cycle is presented in Figure 2.35.

Optomechanical Setup:

Figure 2.33 presents the optomechanical setup for configuration-1: Figure **1** presents the single channel flow cell with the holder. We designed and fabricated the holders for the flow cell. The flow cell was installed directly above the objective and under the condenser on X-Y stage. The objective was mounted directly below the flow cell on a z-piezo stage (for auto focus). In the beginning the optical axis of the objective was aligned thought the center of the chamber. Once the experiment was started the chamber was moved to a desired location.

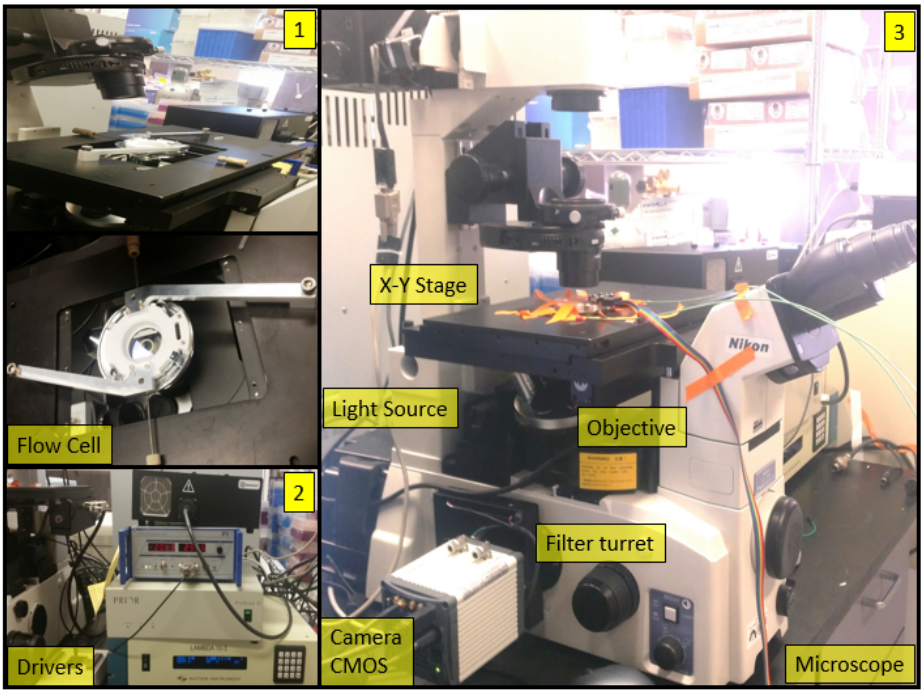


Figure 2.33: Optomechanical setup configuration-1 for Prototype-3.

Figure **2** presents all the drivers and controllers. On top, we have Spectra X then it is PI piezo driver and then Prior X-Y stage driver and in the bottom, we have Sutter filter changer driver.

Figure **3** presents Nikon inverted microscope TE2000-E2 with Prior X-Y linear motor stage. The linear stage could travel up to 120mm×72mm with .05μm resolution and .15μm repeatability. The objective was installed on a turret under the flow cell. The light source was provided through the back port of the microscope; the light was reflected from the mirror block (inside the microscope) towards the back aperture of the objective. Hamamatsu camera and filter changer was installed on the side port of the microscope. All the fluids were installed on the breadboard on the side of the microscope.

Figure 2.34 presents the optomechanical setup for configuration-2. The optical path and basic setup of both the configuration was same. Configuration-1 was designed to use a single channel flow cell and configuration-2 was designed to use 4 channel flow cell. Few components of configuration-1 were replaced with new com-

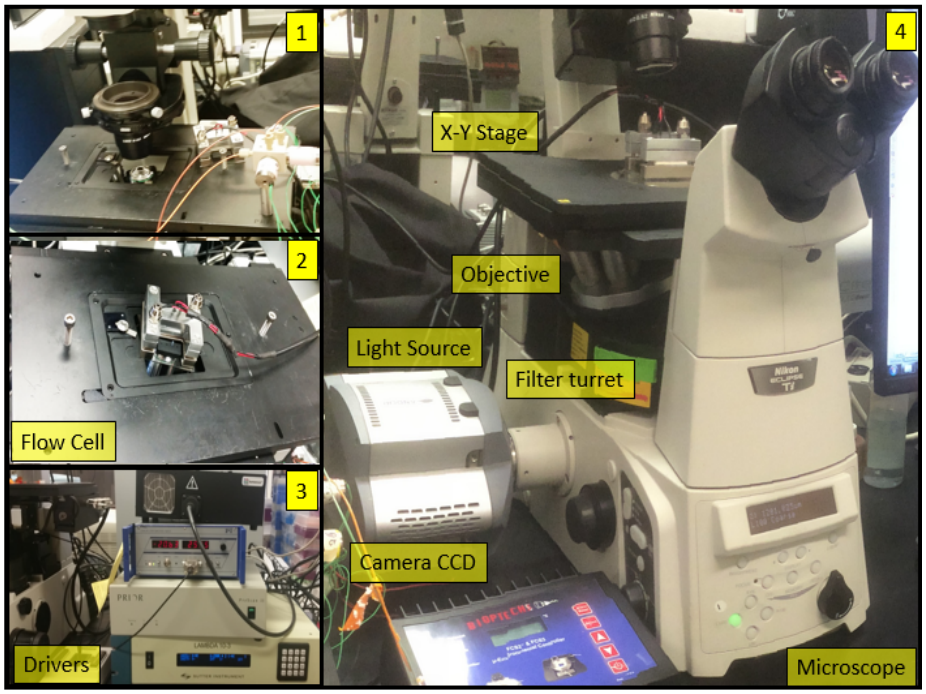


Figure 2.34: Optomechanical setup configuration-2 for Prototype-3.

ponents to increase the effectiveness and adaptability, other components were kept same. We replaced microscope TE2000 with Eclipse Ti (both from Nikon), camera from Hamamatsu to Andor and light source from Spectra to LED boost. We moved the filter changer from in front of the camera to under the objective. The filter changer was mounted between the camera and microscope in configuration-1 and below the objective in configuration-2. The objective was mounted directly below the flow cell on a z-piezo stage (for auto focus). And the light source was mounted on the back of the microscope.

Figure **1** and **2** present the 4-channel flow cell, the flow cell was mounted directly above the objective. Once it was mounted X-Y stage was used to find the desired location for the experiment and once that was found the coordinates (of the location) were noted down and the temperature controller was installed on the flow cell. Figure **3** presents all the divers as already been described.

Figure **4** presents Nikon inverted microscope Eclipse Ti with Prior X-Y linear motor stage. The objective was installed on a piezo stage on a turret under the flow

cell. The light source was provided through the back port of the microscope; the light was reflected from the mirror block (inside the microscope) towards the back aperture of the objective. The filter changed was installed under the turret right on the optical axis of the microscope. Andor camera was installed on the side port of the microscope. All the fluids were installed on the breadboard on the side of the microscope like configuration-1.

2.5.2 Prototype-3 fluidics, control electronics and data acquisition

Fluidics

Figure 2.35 presents the complete flow chart for Prototype-3. It is presented for two cycles; in first cycle, all the optimization and adjustments are done regarding mounting the flow cell, finding a good spot to start the experiment, recording the position coordinates of X-Y and Z, take bright filed image and then take a first set of red and green images. In second cycle and all the coming cycles the procedure is same: We started from flowing the reagents then incubate for 90sec (time for the dNTPs to incorporate) then flow the wash to wash off all the free reagent. Activate the autofocus, find the good image plane and take the red and green images. And in the end of the cycle flow the cleave to cut of the incorporated part. Before the cleave is flown the process from auto focus to image acquisition can be repeated from many number of FOVs if required. The flow cells are very big with dimensions of 10sec of millimeters but a single FOV only covers about 650μm×650μm so if we want to cover the complete flow cell we can do it by programmatically moving the X-Y stage in steps of FOV to cover the whole flow cell. Once the flow cell is the X-Y stage moves back to the starting position (which was recorded previously) and the cleave flows to start the next cycle. This whole process is repeated for the number equal to

the number of base pairs needed to be sequenced. Figure 2.36 presents the fluid flow diagram for prototype-3.

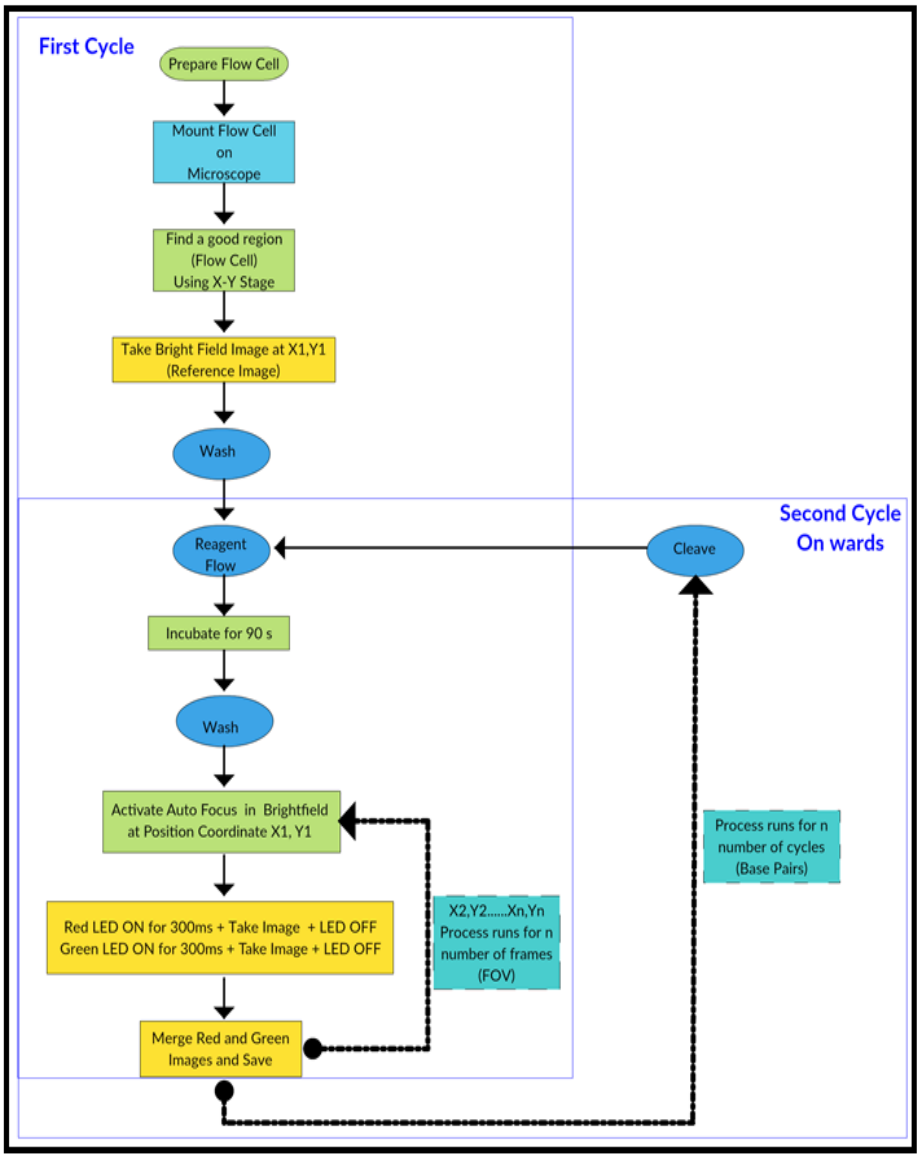


Figure 2.35: Flow Chart for Prototype-3.

Fluid-flow sequence:

After the system is prepared for run (prime and prime reagents), flow-cell is assembled into the fluidics and washed. Reagents are loaded into the respective valves and flown to the flow-cell in the chronological sequence shown in Figure 2.36.

One sequencing cycle consist of the extension reaction in the reaction chamber in

the flow cell and then cleaving the reagents, preparing for another cycle. Figure 2.36 shows the chronological sequence of flow of reagents for one complete cycle. Air is used to separate the reagents from mixing with each other. Reagents are drawn into the system before injecting to the flow-cell. Volume calibration is done for each flow-cell for reagent volume (volume). To prevent saturation of the reagents during the reaction chemical reagents (extend and cleave) are pushed by the following air and wash channel. The sequence of flow of reagents consists of two processes. First the

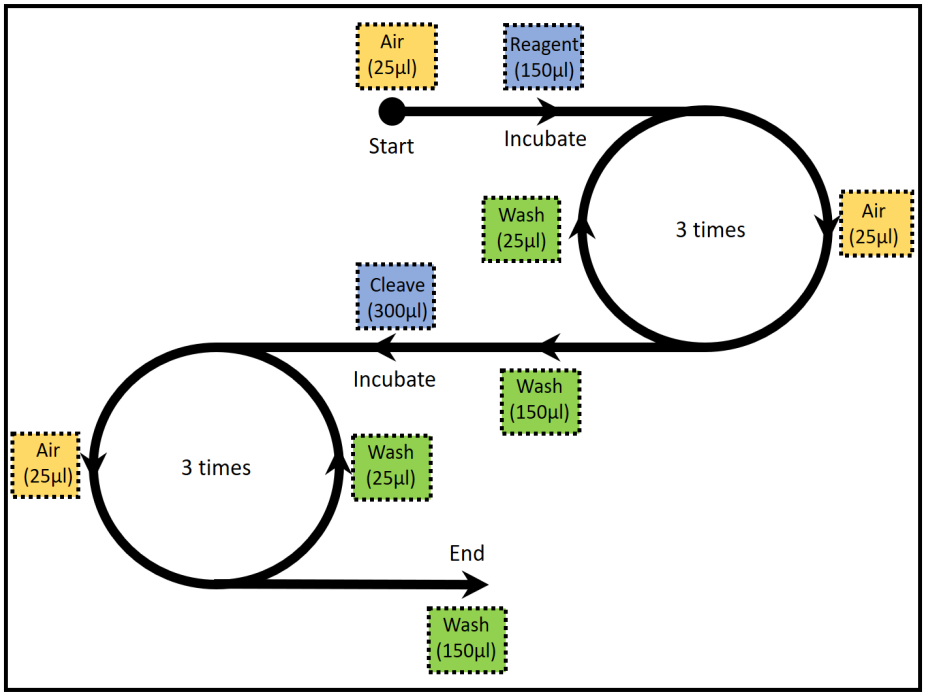


Figure 2.36: Chronological fluid flow diagram for a one sequencing cycle.

reagents are drawn into the system in a sequence (Figure 2.36) and then pushed into the flow-chamber. 150μl of reagent is preceded by 25μl of air-bubble and succeeded by an air-wash channel. Extend is incubated in the chamber for 90sec, then 10μl of the reagent is pushed to prevent saturation and introduction of new reagent. This is repeated 3 times. 120μl of the reagent is incubated in the channel and fresh 10μl reagent is pushed after 90sec of incubation period. After incubation, the channel is washed. Picture is taken and cleave is pushed into the reaction chamber and reaction is processed similarly. The reagents, wash and cleave were kept at 4^{o}C and flow cell

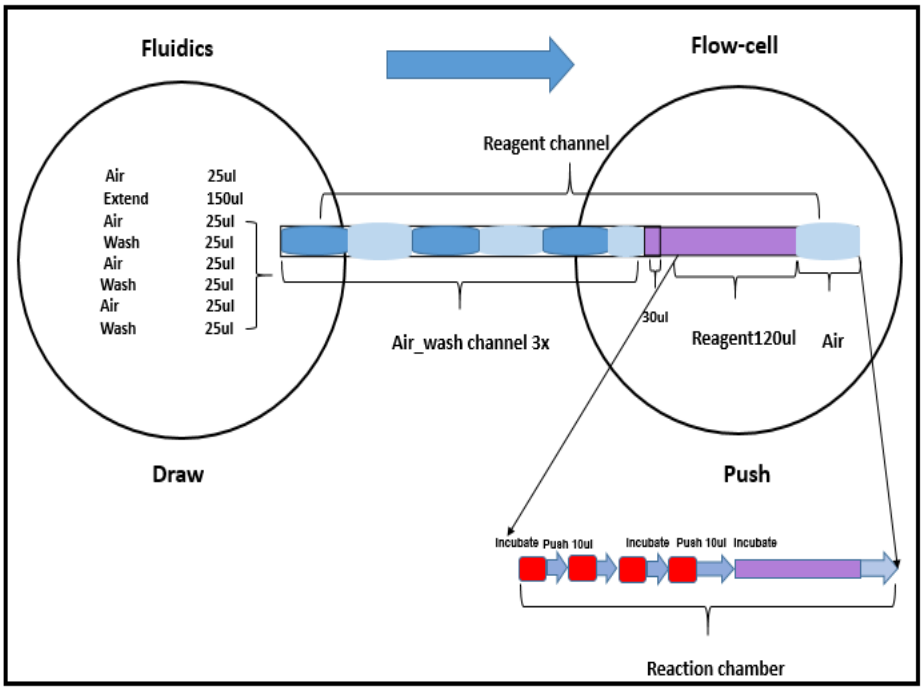


Figure 2.37: Fluid flow from fluidics to reaction chamber.

was kept at 45^{o}C to 50^{o}C. A syringe pump was used to push the liquids.

Fluidics

The fluidics of Prototype-3 is different from Prototype-1 and 2 but uses similar components. Prototype-3 fluidics is much simpler and smaller. A component list of Prototype-3 is presented in Table 2.4. Most of the components are similar as Prototype-1 and 2. We used two different flow cells for the s prototype: Configuration-1 used a single channel flow cell and configuration-2 used a 4 channel. We used a syringe pump to move the liquid and air in the system. The idea of using a syringe pump was the following: Syringe pump was both accurate (volume) and power full and it was easy to control it with a MATLAB code. We used the syringe pump after the flow cell in the end of the fluidics; syringe pump was connected to the flow-cell which was in turn connected to the fluidics system. In this way, the fluid was pushed in to the flow-cell facilitated by the pump. Figure 2.37 presents the fluidics schematics.

Table 2.4: DNA Sequencing Prototype-3 Component List.

Components	Quantity	Vendor	Comments
Mechanical Assembly			
Bread Board; MB1824	1	Thorlabs	Fluidics system base
Post; TR4	10	Thorlabs	Valve mounts
Post; TR8	4	Thorlabs	Bottle holder
Fluidics Assembly			
Tee w/ F-300 fittings; P-727	1	Idex	Material PEEK
Ferrule with SS ring; P-250	20	Idex	Material PEEK
Nut 1/4-28; P-255	20	Idex	Material PEEK
Red tubing; 51085K41	1	Mcmaster Carr	1/16" Material PEEK
Yellow tubing; 51085K42	1	Mcmaster Carr	1/16" Material PEEK
Orange tubing; 51085K44	1	Mcmaster Carr	1/16" Material PEEK
Green tubing; 51085K48	1	Mcmaster Carr	1/16" Material PEEK
Brass pipe fitting; 50785K281	1	Mcmaster Carr	1/4" to 3/8"
1000ml bottles; FB-800-1000	3	Fisher Scientific	Material glass
Two port bottle caps; GL45	3	Fisher Scientific	Material glass
Flow selection Valve; 080T312-62	3	Bio Chem	1/16" Material PTFE
Cavro XLP 6000 Syringe pump	1	Tecan	
Configuration-1			
Flow Cell; FCS2	1	BioTech	Single Channel
Configuration-2			
Home-made	1	Home Made	4 Channel

Figure 2.38 presents the prototype-3 schematics for configuration-1 (single channel flow cell). In comparison to prototype-1 and 2 we used only 3 liquids and air in prototype-3. We used wash buffer, cleave and extend (reagent). All the liquids were kept in 1000ml bottles. Cleave and extend were kept at 4^{o}C. All the bottles were connected to 3-way flow selection valves with 3 inlets and 1 outlet. We used 3 sets (A, B and C) of selection valves. The valve set C valve inlet V4 was connected to

Extend, V5 was kept open for air and V6 was connected to the Wash through a Tee. The outlet of valve set C was connected to the inlet valve V9 of valve set A. From valve set A the extend was connected to the flow cell. The valve set B valve inlets

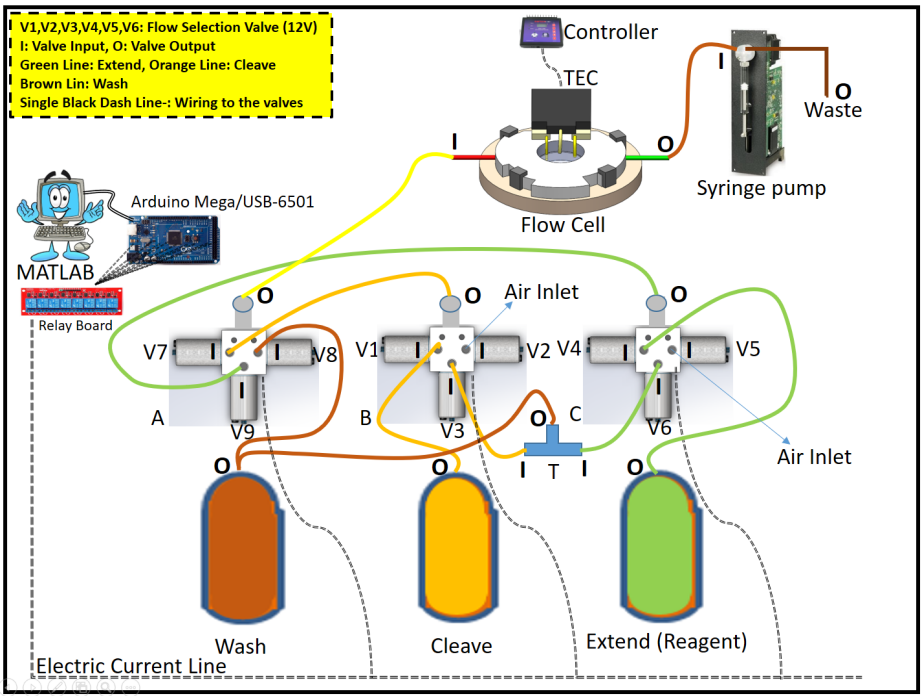


Figure 2.38: DNA Sequencing Prototype-3 schematics. All the fluid channels are presented.

V1 was connected to Cleave, V2 was kept open for air and V3 was connected to the Wash through a Tee. The outlet of valve set B was connected to the inlet valve V7 of valve set A. From valve set A the Cleave was connected to the flow cell. The valve set A valve inlet V7 was connected to the outlet of valve set B for Cleave and air, V9 was connected to the outlet of valve set C for Extend and air and V8 was connected to the Wash. The outlet of valve set A was connected to the inlet of flow cell and the outlet of flow cell was connected to the inlet of the syringe pump and the outlet of the syringe pump was a waste.

Figure 2.39 presents the configuration-1 with a single channel flow cell. The flow cell was kept on between 40^{o}C to 50^{o}C. The configuration-2 was like configuration-1 only the single channel flow cell was replaced with 4 channel flow cell.

Single channel flow cell Figure 2.39 presents the explode view of the single

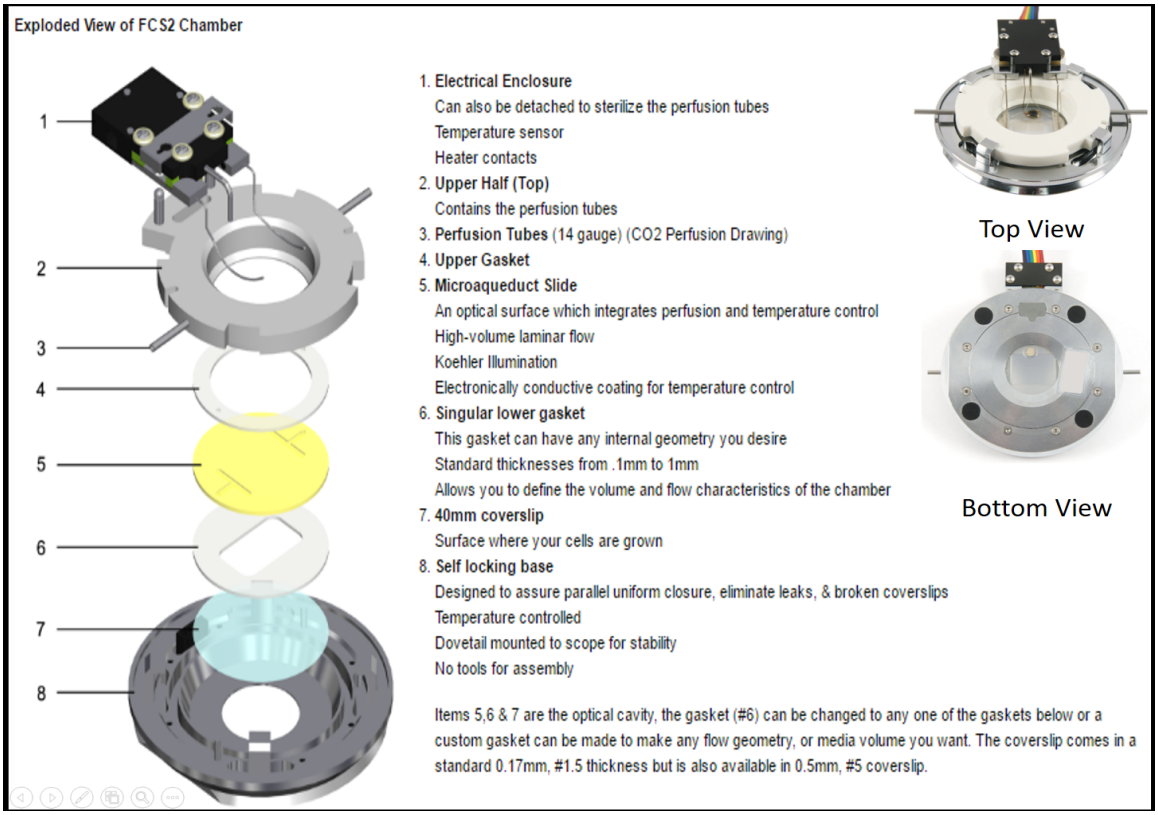


Figure 2.39: FCS2 single channel flow cell.

chamber FCS2 flow cell assembly from BiopTech with all its components. The flow cell comes with a temperature controller. The flow cell sits right above the microscope objective on a 2D X-Y stage as presented in the Figure 2.33. The flow cell was clamped to the stage. The flow chamber was constructed by the singular gasket of 14mm×22mm×.1mm. For more information please go to reference 19 (Figure 2.38). A 40mm coverslip was chemically treated to bind amine-functionalize and coated with streptavidin. Biotinylated acrylamide beads with clonally amplified DNA molecules were immobilized on the surface and the cover slip was loaded into the flow-cell. Same procedure was used in the 4-channel flow-cell except the cover slip was chemically functionalized after assembling into the flow-cell. For details please see the Chapter 3.

Four channel flow cell Single channel flow cell Figure 2.40 presents the 4-

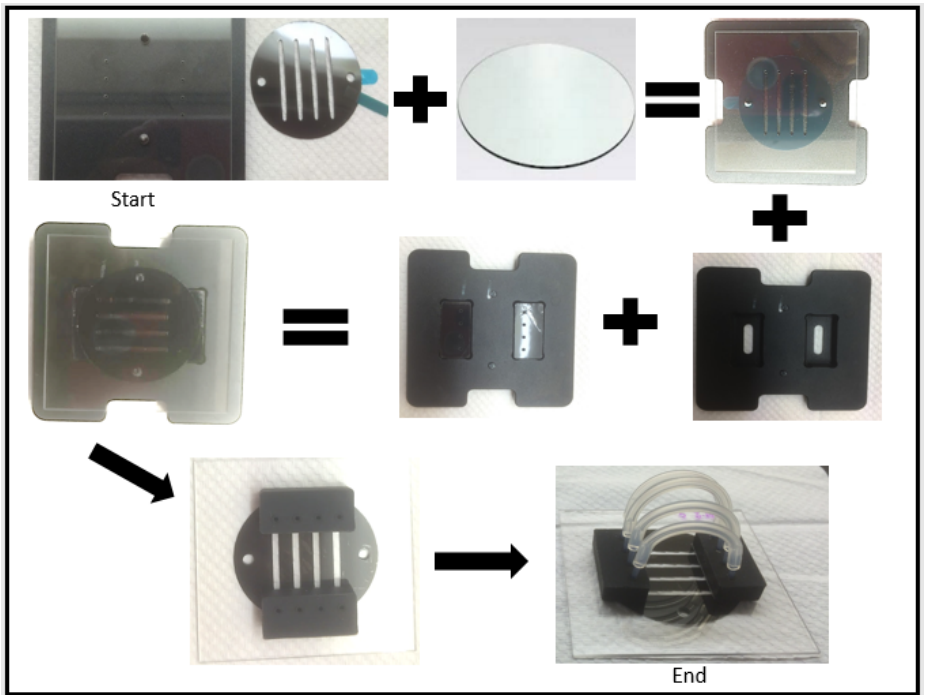


Figure 2.40: 4-Channel flow cell assembly procedure.

channel flow cell assembly procedure. The flow cell was an in-house made apparatus designed to achieve the following:

1. Scalability (4-channels could facilitate 4 simultaneous sequencing reactions).

2. Cost-effectiveness (total-cost for the flow cell is lower than the Bioptech) and easy and low maintenance. LEDs are used instead of lasers which are cheaper than lasers and have longer life.

3. Customizability.

This flow cell had 4 major components: **Component-1:** Square glass with circular holes for connecting two ends of the flow-channel **Component-2:** Pressure sensitive adhesive circular flow-channel to complete the enclosure with 1. **Component-3:** Silicone mounts for each flow-channel **Component-4:** 40mm circular coverslip Figure 2.41 presents the assembly procedure of the 4-channel flow cell which is as the following: All the accessories for assembling the flow-cell were ethanol washed and plasma

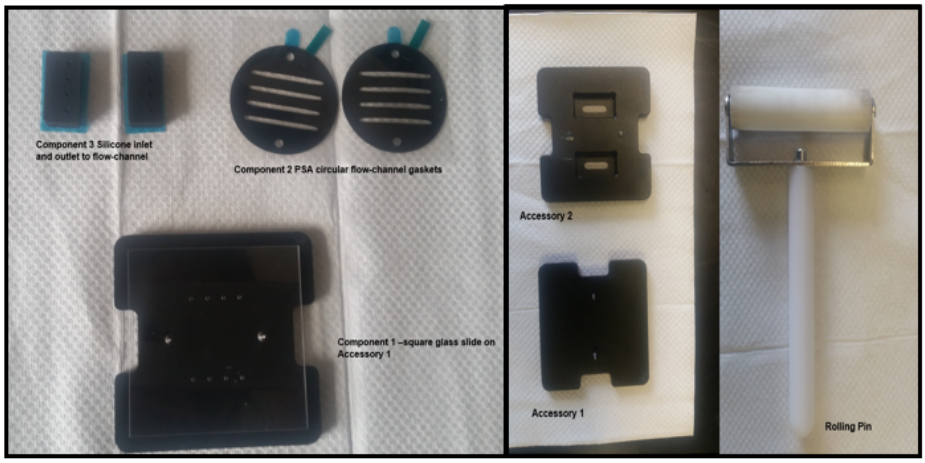


Figure 2.41: 4-Channel flow cell parts.

treated. (Fig 2.41). Component 1 was assembled on accessory 1. Accessory one had screws to keep component 1 in place for accurate assembly of the flow-cell. PSA circular 4-channel flow-cell was attached to the Component 1, PSA had adhesives on both sides. The other side was for circular cover slip. After carefully assembling these three components, the component 4 was assembled in accessory two. The assembled flow-cell was kept upside down on it and pressed against it. Then, the complete assembly was taken out and connected to fluidics with the PEEK piping. Each channel had its own set of inlet and outlet. The fluid was pushed into the flow-cell by the Cavro syringe pump which was controlled by the MATLAB. Fluid calibration was done for each of the flow-channel to ensure the accurate reagent volumes and ensure that complete extension and cleaving occur. Fluid volumes were dictated by various parameters selected for the sequencing runs. These are described as following:

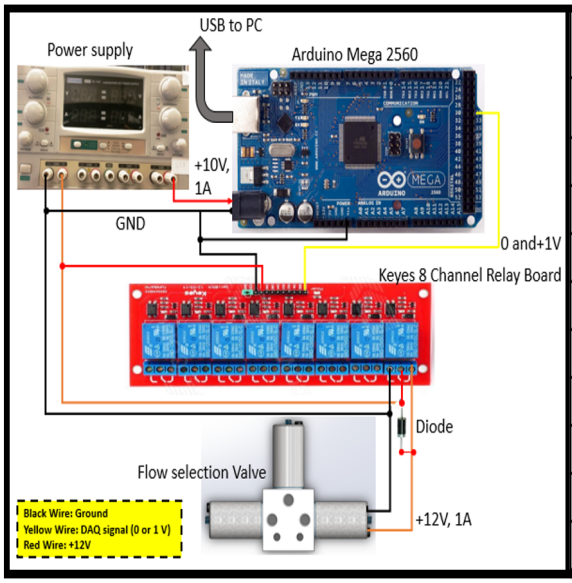
- **Push volume:** Volume to maintain the wash and air channel.
- **Slow volume:** Volume to push the reagent during incubation period to avoid saturation
- **Total volume:** Volume of the reagent for sequencing

For calibration, volumes are tested with colored fluids to maximize reagents used in

the flow-channel during the sequencing run. Appropriate volumes are calculated by multiple runs and thus used for the experiments.

Control Electronics

The electronics of Prototype-3 was very like Prototype-1 and 2; similar components were used. I used Arduino Mega 2560 microcontroller with 8 channel relay board. We even used the same power supply. Figure 2.42 presents the wiring schematics and table 2.5 presents the components list. Figure 2.42 presents the wiring schematics for



Wiring Schematics Prototype-3

Arduino Pins	Relay Board	Valve	Reagent
30	1	1	Cleave
31	2&5	2&5	Air
32	3	3	Wash
33	4	4	Extend
34	8	8	Wash
35	6	6	Wash
36	9	9	Valve Set C
37	7	7	Valve Set B
45	PWM	Temp (5V)	TEC

Figure 2.42: Wiring schematics for one valve on the left and Pin number list on the right.

one valve. Similar wiring is applied to all 9 valves (in set A, B and C). I used 1 eight channel relay boards controlled by an Arduino Mega 2560microcontroller and power supply. The fluidics system was completely automated; controlled by MATLAB code script. MATLAB controlled Arduino Mega 2560 and Arduino Mega 2560 controlled the 8-channel relay board and the relay board controlled the valves. I used 8 digital pins and 1 PWM pin; valves were controlled by the digital pins and TEC controller (temperature controller) was controlled by the PWM. I used 9600 Baud rate and 8-bit protocol. Images were taken by the Nikon Imaging software.

Table 2.5: Control electronics component list for Prototype-3.

Components	Quantity	Vendor	Comments
Arduino Mega 2560	1	Sparkfun	
Keyes 8 Channel Relay Board	3	Sparkfun	12V
Power Supply 12V	2	Sparkfun	
Configuration-1			
FCS2/3 Chamber Controller	1	BioTech	Single Channel
Configuration-2			
Home-made	1	Home Made	4 Channel

Data Acquisition hardware and Software

Data was taken in the form of images by Hamamatsu camera in configuration-1 and Andor camera in configuration-2. Both the cameras along with their optical paths have been already discussed in the optics section.

Software

Two programs were used for data acquisition and control: MATLAB was used to control the fluidics and TEC and NIS-Elements Advance Research (from Nikon) was used for image acquisition S0888754315300410.(n.d.), Z piezo focus stage control and X-Y stage control. During experiments both the programs were called by an independent macro script.

Macro

A macro - an executable sequence of commands - can make the work very effective. NIS-Elements provides a C-like programming language utilizing its internal set of commands. The sequence of commands can be created either by recording the performed actions, by writing the commands inside the macro editor, or by modifying the command history (the history is recorded automatically during the work). The created macro can be saved to an external (*.mac) file for later use. The image

processing was done my Image-J.

2.6 Discussion

In this chapter, sequencing prototypes were discussed with the aim of genotyping at point-of-care diagnostics. Simple, scalable and customizable prototypes were built for sequencing solutions for low-power, low-sample and cost-effectiveness. Prototypes were developed with novel signal amplification read-out electronics to ensure cheaper, scalable, sensitive and customizable sequencing device which is hard for current available technologies. ISFET based prototypes along-with novel pH to current read-out circuit enables genotyping in low power and low DNA conditions which is a limitation in current commercial sequencers.

Fluorescence based sequencing is currently dominating the market (Illumina) for various NGS applications. With flexible chemistry, DNA immobilizing surface modifications and hardware, customizable and relatively cheaper sequencer prototype was developed.

2.7 Summary

ISFET based commercial sensors were used to optimize the experimental variables for sequencing experiments with ISFET chips and for genotyping experiments. Various chemical modifications were explored for immobilizing the DNA on the surface. ISFET sensors were characterized for the sensitivity and specificity and novel-readout circuit was developed for amplifying the original out-put for better signal and detection in low sample limits. Meanwhile, automated fluidics was designed for controlled delivery of sequencing reagents. Fluorescence based technique was used for detection in another prototype developed for targeted sequencing of exome. Chemical modifications were optimized to immobilize DNA-loaded acrylamide beads on the glass

surface. The glass coverslip was assembled in the in-house made and commercial flow-cells and sequencing was done. The fluidic system developed previously was used with much simpler version. Nikon-TE and Nikon-eclipse were used for imaging the fluorescence after the sequencing reaction.

References

[Bergveld 1970] Bergveld, P. (1970). Development of an Ion-Sensitive Solid-State Device forNeuro-physiological Measurements. *IEEE Transactions on Bio-Medical Engineering.* **Click on the number to return to that page.** 21, 23

[Barbaro 2006] Barbaro, M., Bonfiglio, A., et .al. (2006). A CMOS, Fully Integrated Sensor for Electronic Detection of DNA Hybridization. *IEEE Electron Device Letters.*, Vol. 27, No.7. **Click on the number to return to that page.** 21

[Lee 2009] Lee, C., Kim, S. K., Kim, M., et. al. (2009). Ion-Sensitive Field-Effect Transistor for Biological Sensing. *Sensors.* **Click on the number to return to that page.** 21

[Uslu 2004] Uslu, F., Ingebrandt, S., Mayera, D.,Bcker-Meffert, S., et. al. (2009). Label free fully electronic nucleic acid detection system based on a field-effect transistor device. *Biosensors and Bioelectronics.*19, 2204. **Click on the number to return to that page.** 21

[Shoorideh 2012] Shoorideh, K., & Chui, C. O. (2009). Optimization of the Sensitivity of FETBased Biosensors via Biasing and Surface Charge Engineering. *IEEE Transactions on Electron Devices.* Vol. 19, 2204. **Click on the number to return to that page.** 21

[Chapman 2011] Chapman, R. A., Fernandes, P. G., et. al. (2011). Comparison of Methods to Bias Fully Depleted SOI-Based MOSFET Nano-ribbon pH Sensors.

IEEE Transactions on Electron Devices. Vol. 58, No. 6. **Click on the number to return to that page.** 21

[Hammond 2004] Hammond, P. A., Ali, D., & Cumming, D. R. S. (2004). Design of a Single-Chip pH Sensor Using a Conventional 0.6-μm CMOS Process. *IEEE Sensors Journal.* Vol. 4, No. 6, pp 706-712. **Click on the number to return to that page.** 21

[Yang 2007] Yang, C., & Liao, Y. (2011). An ISFET Interface Circuitry for Biomedical Applications. *IEEE Conference on Electron Devices and Solid-State Circuits.* pp 1083 - 1086. **Click on the number to return to that page.** 21

[Chan 2007] Chan, P. K., & Chen, D. Y. (2007). A CMOS ISFET Interface Circuit With Dynamic Current Temperature Compensation Technique. *IEEE Transactions on Circuits and SystemsI.* Vol. 54, No. 1, pp 119-129. **Click on the number to return to that page.** 21

[Wang 2012] Wang, K., Liu, Y., &Toumazou, C., Georgiou,P. (2012). A TDC Based ISFET Readout for Large-Scale Chemical Sensing Systems. *IEEE Biomedical Circuits and Systems Conference.* pp 176-179. **Click on the number to return to that page.** 21, 27

[Bergveld 2003] Bergveld, P. (2003). Thirty years of ISFETOLOGY, what happened in the past 30 years and what may happen in the next 30 years. *Sensors and Actuators B.* **Click on the number to return to that page.** 23

[Shepherd 2005] Shepherd, L. M., Toumazou, C. (2005). A Biochemical Trans-linear Principle With Weak Inversion ISFETs. *IEEE Transactions on Circuits and Systems-I.* Vol. 52, No. 12. **Click on the number to return to that page.** 23

[Mohammad 2015] Mohammad, M. U., Zarkesh-Ha, P., Edwards, J., S.,Coelho, E., Rawat, P. (2015). A Highly Sensitive ISFET Using pH-to-Current Conversion for

Real-Time DNA Sequencing. *IEEE Explore.* **Click on the number to return to that page.** 40

[Gonzalez2015] Gonzalez, M. L. (2015). The road from next generation sequencing to personalize medicine. *Medicine.* **Click on the number to return to that page.** 11(5), pp 523-544. 50

[Guzvic 2013] Guzvic, M. (2013). The History of DNA Sequencing. *Journal of Medical Biochemistry.* 32(4), pp 301-312. **Click on the number to return to that page.** 50

[Liu 2012] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Law, M. (2012). TComparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology.* **Click on the number to return to that page.** 50

[Chen 2013] Chen, F., Dong, M., Ge, M., Zhu, L., Ren, L., Liu, G., & Mu, R. (2013). The History and Advances of Reversible Terminators Used in New Generations of Sequencing Technology. *Genomics, Proteomics and Bioinformatics.* 11(1), pp 34-40. **Click on the number to return to that page.** 50

[Goodwin 2016] Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics.* 79(11), pp 333-351. **Click on the number to return to that page.** 50

[Veritas Genetics] https://www.nikoninstruments.com/Products/Software/NIS-Elements-Advanced-Research **Click on the number to return to that page.** 50

[Balgir 2000] Balgir, R., S. (2000). The burden of haemoglobinopathies in India and the challenges ahead. *Current Science.* 79(11), pp 1536-1547. **Click on the number to return to that page.** 50

Online reference list

- BIOPTECH FCS2 System: http://www.bioptechs.com/product/fcs2-system/
- Nikon Imaging Software: https://www.nikoninstruments.com/Products/Software/NIS-Elements-Advanced-Research

Chapter 3

Calibration, Materials, Methods and Results

By

Priyanka Rawat

Table of Contents

List of Figures

List of Tables

Chapter 3

Calibration, Materials, Methods and Results

3.1 Introduction

In this chapter I will discuss calibration, materials and methods along with the results and analysis. The discussion is organized based on chapter 2. In calibration, the hardware (sensor etc.) and experimental parameters are discussed, in material and methods, the bio-chem of the experiments is discussed. This chapter is in continuation with chapter 2. Experimental details, sequencing process and prototype functioning for each of the prototypes will be discussed in this chapter. For ISFET based prototypes, calibration data of the sensors, sensitivity and specificity for single and multiple nucleotide extension reactions and other experimental conditions are explored and optimized. For fluorescence based prototypes, optimization is done for library amount, flow-cell volumes, reagent volumes per reaction and then image acquisition parameters. For both prototypes, various surface chemistry options are explored for robust and controlled immobilization of the sequencing molecules.

3.2 Isothermal single nucleotide extension with ISFET Sensors: optimization and characterization of the variables

3.2.1 Optimization and characterization of the variables

We use two COTS (PH37-SS from Hach and MicroFET 9270-010 from Sentron; commercial off the shelf) ISFET sensors (which came as complete assembly, ready to use, with digital pH meter), one commercial strip-sensor (MO-PSF2 from Micropto; with limited calibration data and no readout circuit and pH meter) and one in-house built sensor (4-core ISFET sensor chip). All the sensors with their corresponding setups (electronics and software) have already been discussed in Chapter 2. COTS sensors were characterized by the vendors so any additional characterization was not necessary. Strip-sensor (MO-PSF2) came with very limited characterization data (see Figure 2.11) so it was required to have a more detailed characterization (characterization curve and sensitivity). 4-core ISFET sensor chip was in-house built so it required a full characterization. The readout circuits have already been discussed in detail in Chapter 2.

Strip-sensor (MO-PSF2) characterization:

Strip-sensor MO-PSF2 was bought from Micropto. Inc. It was a n-channel ISFET sensor with T_2O_5 gate material. The gate dimensions were $20\mu m \times 700\mu m$. I used a drain-source voltage of 500mV at 100μAmp of current in the start. I performed the characterization using 3 pH solutions: 4, 7 and 10. The pH solutions (for calibration) were prepared using the HACH pH sensor and cross checked with Sentron. The calibration was done on several occasions (different days and time). The sensor was also tested for drifts over a long period of continuous sensing. After analyzing the data, a calibration slope of 74 ± 1.5mV/pH was stabilized. The readout circuit for

the calibration and data acquisition has already been discussed in Chapter 2. The characteristic curve is presented in Figure 3.1. The characterization data suggests the

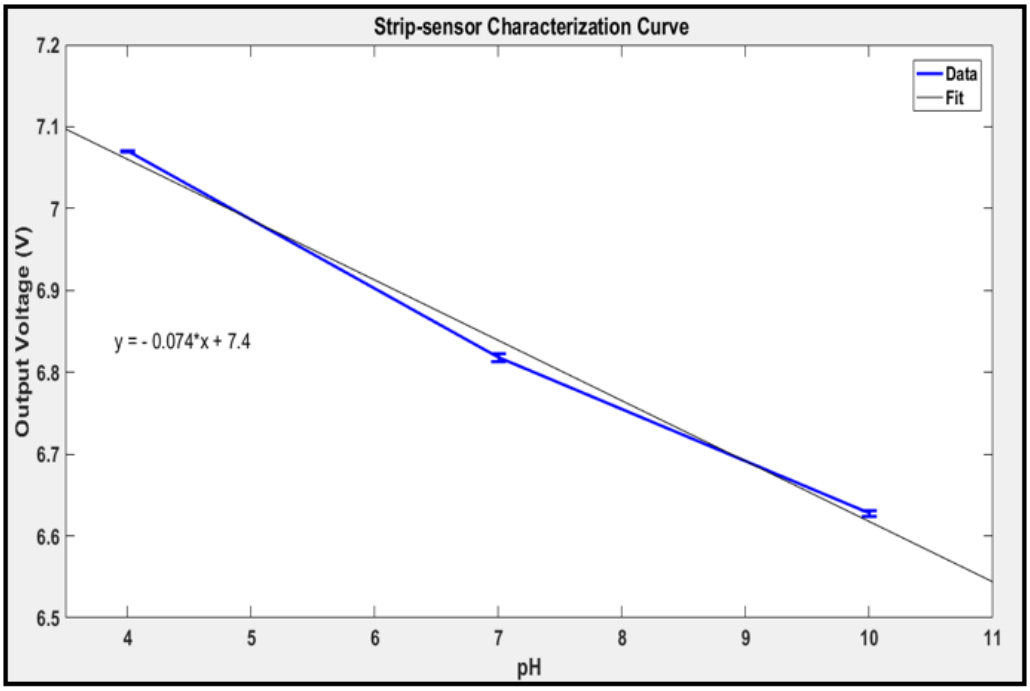


Figure 3.1: Strip-sensor characterization curve. A slope of 74mV/pH is presented.

sensor resolution of .04pH. The sequencing experiments usually go for hours so it is very important for the sensor to be consistent. I did a brief study on the stability of the sensor vs time (over different days and time). The study suggests that the sensor was consistent with an average error of .015pH over a period of 3 hours. The vendor data sheet gives a value of .2pH/day.

4-core ISFET sensor chip characterization:

A test chip was designed and fabricated using TSMC's 0.25μm technology. The test chip (4-core) consists of four sensing cores, shown in Figure 2.21 (Chapter 2) and 3.4. Each of the cores in test chip has 90×95 unit cells. The cores contained four different unit cell specifications, which included P-ISFET and N-ISFET for both small (with W/L=2.5) and as well as large (with W/L=24) ISFETs. Core 1 and 2 were P-ISFET (core 1 small and core 2 large) and core 3 and 4 were N-ISFET (core 3 small and core 4 large). T_2O_5 was used as the gate interface material. Two different

channel types (N and P) were made with two different sizes to explore the optimum sensitivity based on biasing; all this has already been discussed in Chapter 2 along with circuit, electronics and software.

The characterization depends on many factors such as bias voltage (V_{ds}), reference voltage (V_{gs}), ISFET channel type (N or P) and channel size (W/L). So, we based the characterization on the results of SPIC simulation. First, we found out the optimum bias voltage for any reference voltage for 3pH values for four cores. Here we were looking for the maximum mean pixel value for each core for 3 different pH values averaged over 5 different references voltages.

Table 3.1: Mean pixel value for 4 cores. Each data point is an average of 75 data points.

Bias Voltage V_{ds} vs Core mean pixel value at 3pH values												
Bias Voltage (V)	**pH**											
	7.36				**7.74**				**8**			
	Core mean pixel value (V)				**Core mean pixel value (V)**				**Core mean pixel value (V)**			
	1	2	3	4	1	2	3	4	1	2	3	4
3	246	318	529	571	245	317	528	567	255	324	542	567
3.5	253	156	596	618	255	157	596	610	256	157	594	602

Table 3.1 presents the data for 2 different bias voltages; 3 and 3.5V. Column 1 to 4 presents the pixel mean value for 4 cores. Each value is an average of 5 different reference voltages (1, 1.5, 2, 2.5 and 3V) and each voltage value has 15 repetitive data point at each pH (7.36, 7.74 and 8.00) so each mean pixel value is an average of 75 (15×5) data points per pH. The calibration data suggests that the core 3 and 4 has the highest mean pixel value (signal) and the value is optimum at 3.5V bias. So, we chose the bias voltage to be 3.5V for our next calibration step.

Next step was to find the best reference voltage (V_{gs}). Best reference voltage was found for all four cores using the same data. The result is presented in Table 3.2.

Table 3.2: Mean pixel value for core 3 and 4 at 3.5V bias.

Core 3 & 4 mean pixel value vs pH at 3.5V biase										
pH	Core 3 mean pixel value (mV)					Core 4 mean pixel value (mV)				
	Reference Voltage (V)					Reference Voltage (V)				
	1	1.5	2	2.5	3	1	1.5	2	2.5	3
7.36	716	714	547	549	458	494	610	631	635	721
7.74	713	712	548	546	465	487	598	627	626	712
8.00	712	710	547	542	463	483	589	621	614	705

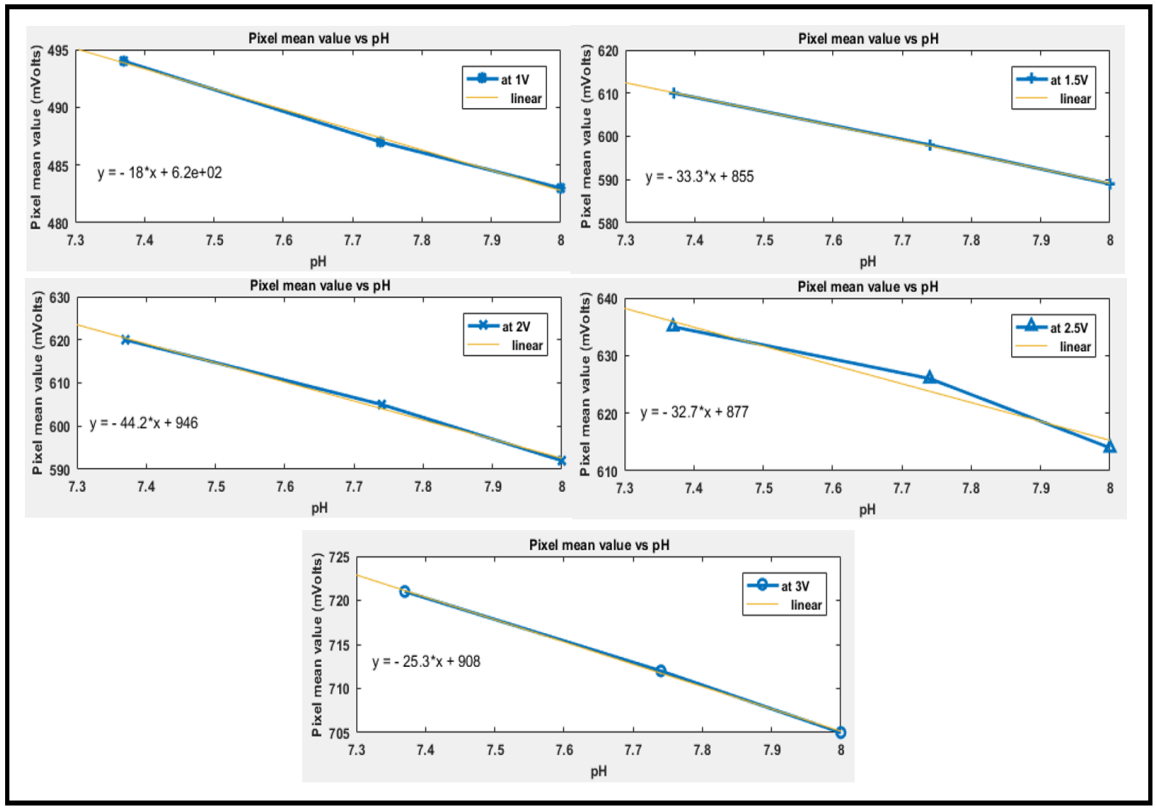


Figure 3.2: Core 4 mean pixel value vs pH at 1, 1.5, 2, 2.5 and 3V reference voltages at 3.5V bias.

Table3.2 presents the mean pixel value for core 3 and 4 at 3.5V bias. Left most column presents the pH (7.36, 7.74 and 8.00) at which the values were acquired. The right two columns present the mean pixel value for core 3 and 4 at reference voltage of 1, 1.5, 2, 2.5 and 3V. Each value is an average of 15 repetitive data point at each

pH (7.36, 7.74 and 8.00). The calibration data suggests that the core 3 and 4 (N-ISFET small and N-ISFET large) has very similar response but core 4 over all better response over all the pH values.

Figure 3.2 presents 5 plots between core 4 mean pixel value vs pH for 5 different reference voltages. The curves are almost linear with negative slope presenting sensitivity (slope; mV/pH) at each reference voltage. The calibration suggests (Figure

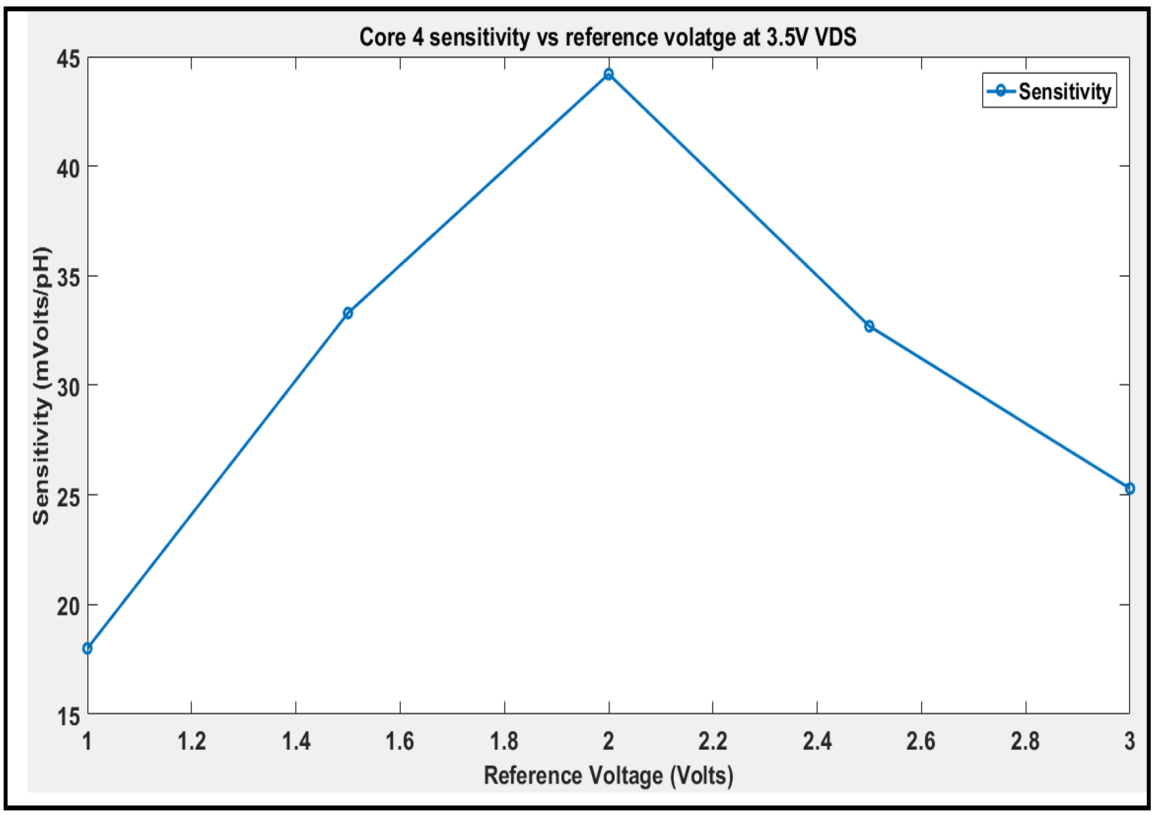


Figure 3.3: Sensitivity(mV/pH) vs reference voltage (Vgs). The sensitivity is maximum at 2V reference.

3.3) that core 4 N-ISFET large is the most sensitive among all with mean sensitivity of 44mV/pH, the optimum bias V_{ds} is 3.5V and the optimum reference voltage V_{gs} is 2V.

Figure 3.4 presents the image of the chip for all five reference voltages at 7.36 pH and 3.5 bias voltage. The image is very like the CMOS sensor in our cell phone cameras. In cell phones the sensor is sensitive to the brightness of the light in our case

it is sensitive to the pH values. The image was acquired in real time at the refresh rate of 30 frames per second. Image represent the four cores and each core has 90×95 unit cells, which mean 90×95 pH sensors which are simultaneously sensing the pH of the reaction. Top first image represents the position and type of all 4 cores. Core 3 and 4 has the maximum brightness and core 4 has the highest contrast for all 5 reference voltages.

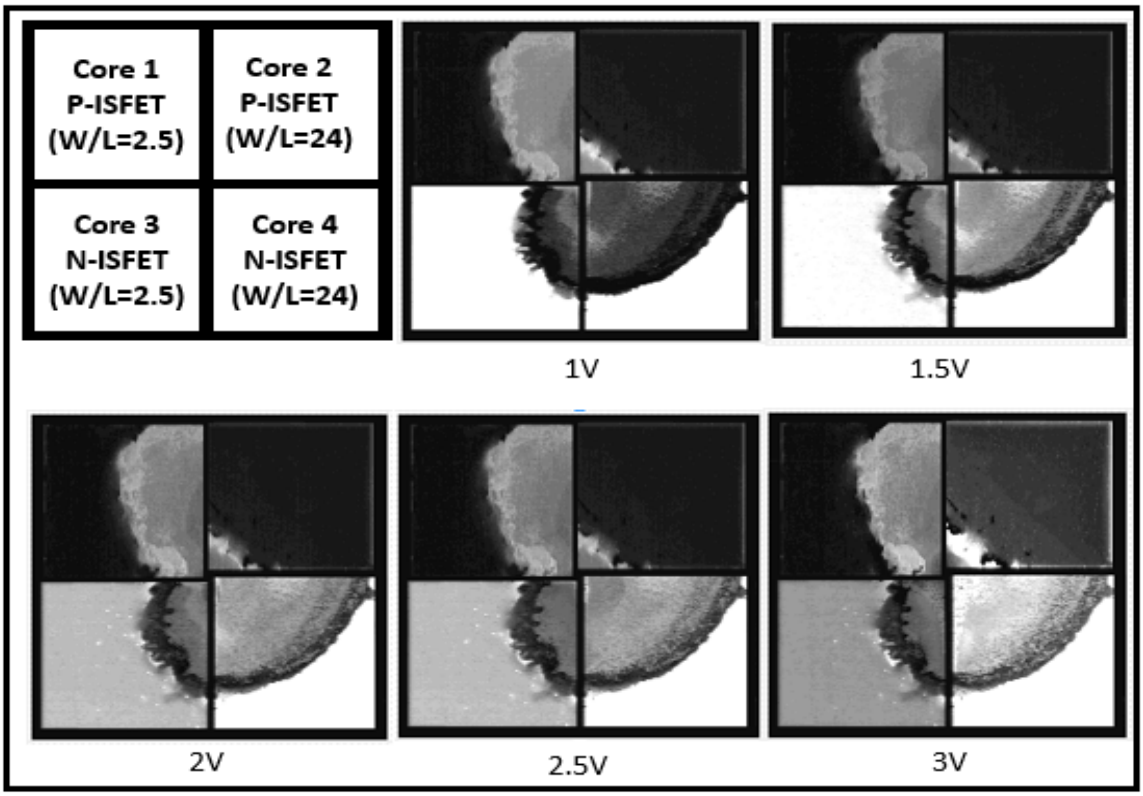


Figure 3.4: Chip Image (all 4 cores) for all 5 reference voltages at 7.36pH and 3.5V bias voltage.

Other experiment variables:

In nucleotide sequencing, complexity increases with increasing optimization. For developing the ISFET based genotyping prototype, preliminary tests were done with commercial sensors and Oligos to detect single nucleotide changes and optimize the experimental variables. 50bp Oligos were purchased from IDT along with 20bp primers.

Immobilization chemistries:

Amine or acrydite modified Oligos were purchased to optimize the immobilization of Oligo strands in controlled manner and to correlate the number of extensions occurred. Immobilization was done on chemical surfaces and in acrylamide gels. Extensions were also optimized in solution. Best suitable chemistries were explored for genotyping experiments and single and multiple nucleotide extensions.

Reaction volumes:

Working reaction volumes were optimized as per sensor detection limits. HACH sensor needed larger volumes for detection approx. 500μl and sentron could detect changes in pH in 20μl making it more convenient for tube-PCR.

Effect of CO_2:

As the purpose was to detect sequencing as a function of pH changes, any kind of buffering due to chemical reagents or environmental factors such as carbon-dioxide can alterd the pH. For this Argon was used to displace any atmospheric air.

I did a brief study on DNA immobilization chemistry to identify the best platform for future genotyping experiments and compatibility with the sensor. Several DNA immobilization chemistries like covalent bond formation with maleic anhydride (Chapter 2 Figure 2.2) and polymerization with acrylamide (Chapter 2 Figure 2.3) were used to identify the best platform. For preliminary DNA sequencing with ISFET: I used three different experimental setups for this study. Multiple base incorporations and single nucleotide incorporations on multiple copies of single stranded oligomers were optimized and confirmed. Oligomers were purchased and immobilized and extension was done in solution. The three experiments are discussed next.

3.2.2 Validation and optimization of multiple and single base incorporations

Multiple and single nucleotide extensions with 96 well-plate

Along with surface immobilization and experimental variable optimization, physical set-ups were designed as discussed in Chapter 2 Figure 2.2 and 2.3. Isothermal extension is done at 37^oC with Bst polymerase. In set-up (shown in Chapter 2 Figure 2.4), 50bp 1μM amine-modified DNA is immobilized on maleic anhydride coated 96 well-plates (Thermofisher, 15108). .4μl of Bst pol(NEB), 10mM ammonium sulphate, 10mM KCl, 2mM $MgSO_4$ are added along with 21bp of 2μM of primer. pH was measured after adding .5mM dNTP for multiple base extensions and single nucleotide extensions with controls were studied (Chapter 2 Figure 2.4). For pH sensing

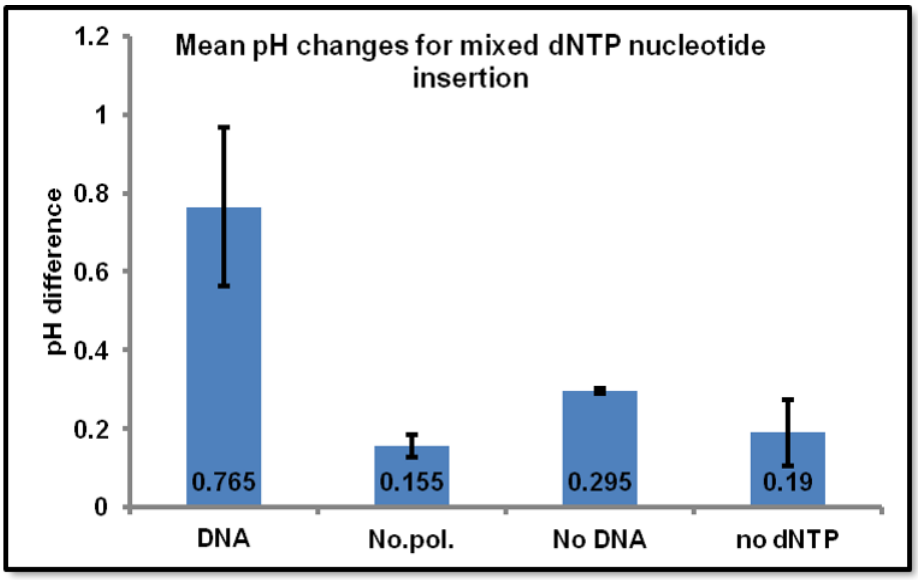


Figure 3.5: Data for Set-up 1. Avg. pH changes for multiple dNTP extension experiments for n=3.DNA is 1μM per control.

HACH ISFET microprobe was used. The probe had a resolution of 0.01pH change detection. Although, this sensor could confirm the extension events Figure 3.5), it lacked consistency, needed higher volumes and thus, more reagents for detection of pH. Also, DNA immobilization was not consistent and regular in 96 well-plate.

Insertion and Non-Insertion with tube-pH

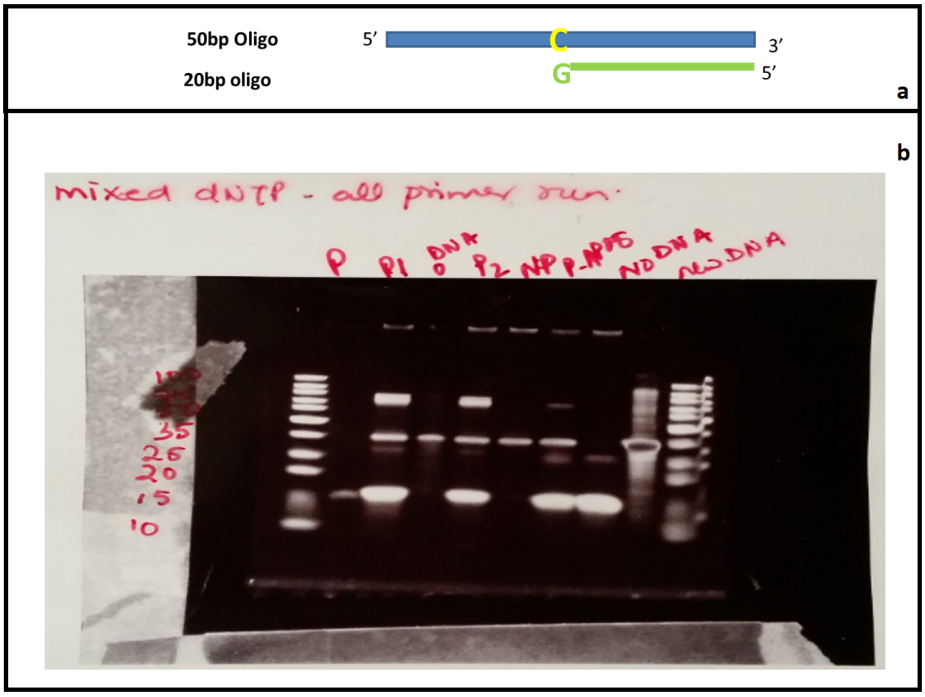


Figure 3.6: 50bp oligo is hybridized to 20bp oligo and multiple base extension is done. b. Gel -Image of sequencing product of 50bp oligo with controls.

Sentron microprobe was used in this experiment which could detect pH changes in volumes as low as 20μl, which could facilitate detection in tube-PCR. Insertion and Non-Insertion events were tested in this set-up. In second experiment the Unmodified 50bp Oligo was used in 50μl 1X reaction buffer pre-hybridized with 20bp primers. Extension with Bst. polymerase at 37^oC was done. Sentron ISFET microprobe was used in this case which was capable of sensing pH in small volumes (20μl) making it ideal for single nucleotide resolution experiments. A low-pressure argon head was used to displace atmospheric air and hence carbon-dioxide which could alter the pH readings. To mimic genotyping experiments, different experiments such as insertion and non-insertion events were done. The results were confirmed by15% PAGE- gel electrophoresis (Figure 3.6). Figure 3.7 present the experiment setup on left and result on the right. An average pH difference of .16 and .06 was reported with insertion vs non-insertion.

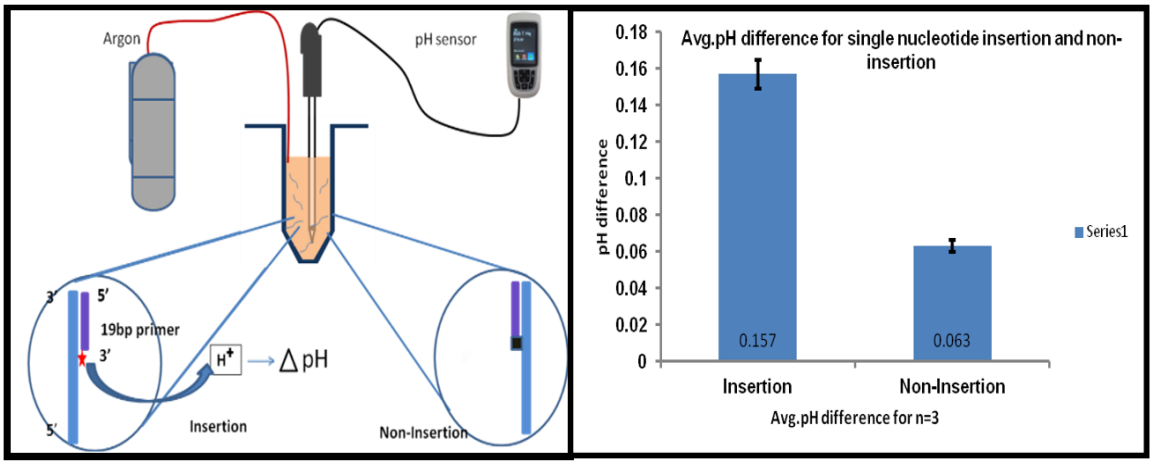


Figure 3.7: Setup 2 with Sentron microprobe. 50l tube is used to do genotyping reaction with DNA in solution. Experiment is designed to detect pH changes during single nucleotide extension when complementary and uncomplimentary dNTP is added. The data is for n=3.

Capture based genotyping

After, successful single nucleotide resolution detection in solution, DNA immobilization was achieved by co-polymerizing acrydite modified DNA pre-hybridized with 20bp primer with 10% acrylamide 4mm gel beads. Set up 2 was used for pH measurements. To determine if beads are feasible for these experiments, pH changes on single nucleotide extension with and without DNA (beads) were analyzed. Mean difference of 0.067 pH was observed for two experiments for .8μM of DNA and no DNA control. Multiple base extension reaction experiments were done with beads and it gave a normalized pH difference of .5 for .8μM DNA. Results were confirmed with 15% PAGE-gel analysis. Also, beads were stained with SYBR gold to confirm the presence of DNA with Biorad UV imager (Figure 3.8).

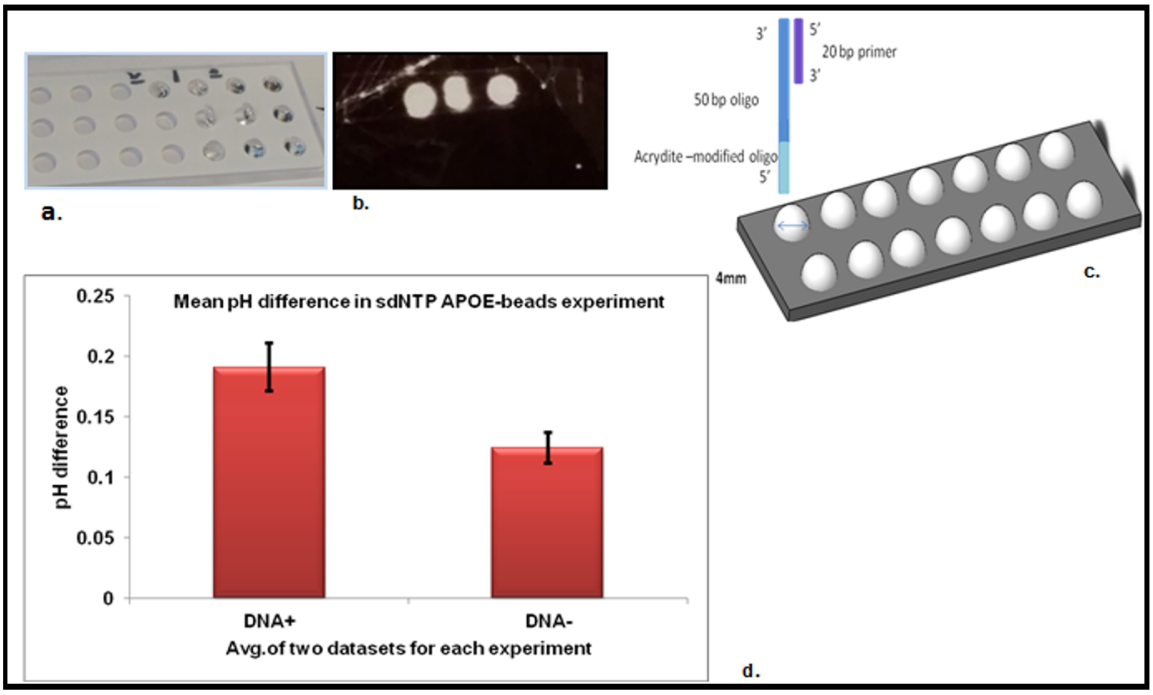


Figure 3.8: a) 4mm acrylamide beads with DNA co-polymerized. b) SYBR gold stained beads to confirm co-polymerization. c) Descriptive picture showing acrydite-modified oligo pre-hybridized with primers in the bead array. d) Avg. pH difference for n=2.

3.3 Materials and Methods for Prototype-3

Prototype-3 is designed to achieve large-scale sequencing goals for targeted re-sequencing. For sequencing, DNA molecules need to be immobilized on a surface. Streptavidin coated glass-cover slides were used to immobilize biotinylated acrylamide beads. These cover slides were then assembled in the flow-cells to carry out the sequencing reactions. For this prototype, we used two flow-cells. Flow-cell 1 was Bioptechs which was a single-channel flow-cell. Flow-cell 2 was an in-house made flow-cell which was four-channel. Although, we used only one channel for these experiments, this flow-cell could be used for multiple sequencing reactions simultaneously. Details about flow-cell-2 could be found in Chapter-2 section 2.5.2.1. These flow-cells were used as chambers for holding glass-slide which was used for immobilization of sequencing molecules (DNA molecules) and facilitate controlled reagent delivery and

temperature for chemical reactions as part of sequencing by synthesis.

Experimental Design

Every sequencing apparatus has following components:

1. Surface to immobilize the sequencing molecules.
2. Library of desired sequencing molecules
3. Fluidics/System for reagent delivery
4. Data acquisition - Imaging/electronic signals

Flow-chart below explains the experiment design for this system briefly.

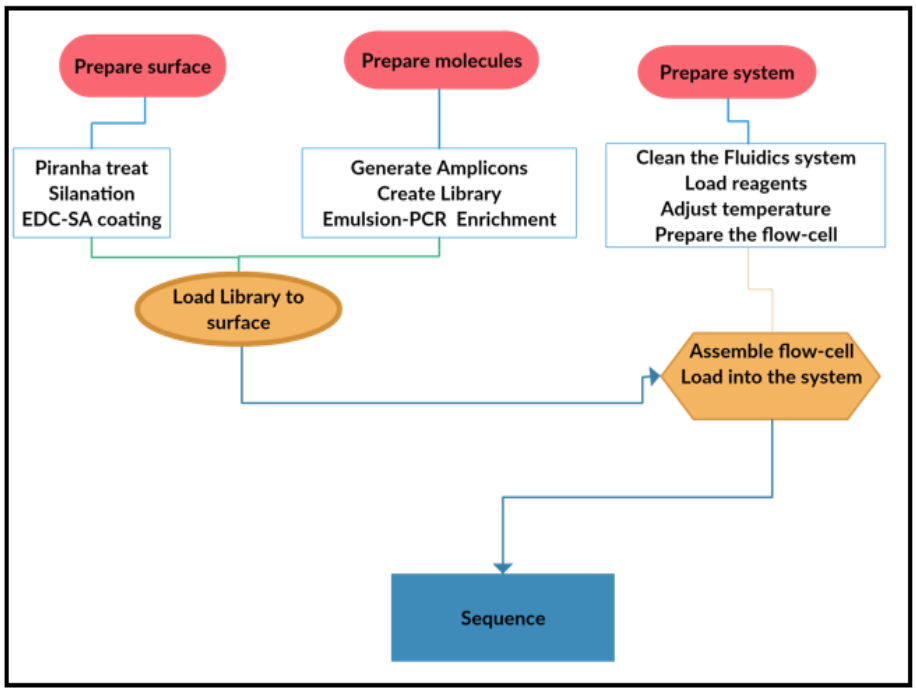


Figure 3.9: Flow-chart describing experimental design of Prototype-3. Glass-surface is treated and prepared for loading streptavidin molecules. Library is loaded after PCR. Fluidics system is prepared and loaded with reagents and flow-cell is assembled and system is prepared for sequencing.

3.3.1 Library Preparation and flow-cell preparation

1. **Multiplex-PCR and amplicon generation** Since this sequencing system is for massive and parallel sequencing purposes, multiple genes could be sequenced. 328 genes associated with 700 unique diseases are chosen to be sequenced simultaneously. Genes are associated with neuromuscular, cardiovascular, developmental and metabolic diseases. The library of exons was prepared in the steps described below:

   Pre-amplification target enrichment:

   Multiplex PCR was performed for generating amplicons from the human DNA. For Targeted sequencing, inherited disease panel of Ion-torrent was used (Ion Ampliseq inherited disease panel, 4477686). This panel contains 3 primer pools with each pool having ∼3500 primer pairs which target exons of 328 genes associated with 700 unique inherited diseases. Avg amplicon length was 197bp (125-225bp). Control DNA was purchased from Promega (G3041, Human genomic DNA). 10ng of DNA was used for each primer pool and PCR was carried out as per instructions. Briefly, 10ng of human DNA per primer pool was mixed with amplification mix. Amplification was carried out by activating the enzyme at 99^oC for 2min, and denaturing at 99^oC for 15sec, anneal and extension at 60^oC for 8min. Amplicons were mixed from all three inputs and partial digestion was done by adding 3μl of FuPa reagent in 30μl sample.

   Partial Digestion and phosphorylation:

   Partial digestion was carried out for 1 hour as per the following temperature profile 55^oC for 20min, 55^oC for 30min and 60^oC for 20min by adding 3μl of Fupa reagent. The digested amplicons was phosphorylated by T4PNK enzyme(For, Pcr, & Technologies, 2012).

   Ligation of sequencing adapters:

Sequencing adapters were ligated to the amplicons by adding 3μl DNA Ligase, 6μl of Switch solution and 3μl of the adapters with 33μl of the amplified library. Ligation was carried out at 22^oC for 30min, 68^oC for 5min and 72^oC for 5min and then held at 10^oC.

Library Purification:

Adapter ligated Amplicons were purified by adding 67.5μl (1.5x) AMPure XP Reagent (Beckman Coulter, Cat. No. A63880) to each library. After incubating for 5min, the beads were pelleted by incubating in the magnetic rack for 2min.Supernatant was discarded and beads were washed with freshly prepared 70% ethanol. Repeat washing.

Amplification of the library:

50μl of the Platinum PCR SuperMix HiFi was added to the beads and mixed thoroughly. 2μl of the Primer Mix was added, amplification mix was vortex thoroughly and kept on the magnetic rack for 2 min. $\sim$50μl of the amplification mix was collected and added to new tube. Amplification was carried out as hold for 2min at 98^oC, 5X of 98^oC for 15sec,64^oC for 1 min and then hold at 10^oCC.

Dual-bead based size selection of the library:

25μl of Agencourt Ampure beads (.5X) was added to 50μl of the amplified library. Mix the library thoroughly and incubate for 5 minutes. Beads were pelleted after keeping on the magnetic rack. Supernatant was collected. Larger amplicons were bound to the beads and smaller were eluted. Another purification was done to purify undesired smaller fragments, and retain only desired size. 60μl (1.2X) of beads were added to the purified library from the first purification, incubated for 5min and pelleted at magnetic rack after 2min. The supernatant was removed and beads were kept. The beads were washed with

70% ethanol, dried and library was eluted with 50μl of low TE.

Library Quantification:

10μl of Library was quantified with Qubit.

2. **Clonal amplification with emulsion-PCR and library enrichment**

Size-selected and purified amplicons were clonally amplified by emulsion-PCR. Emulsion-PCR was a single molecule PCR (Nakano M, 2003). Water-in-oil emulsion allowed for droplet formation and each droplet served as a partitioned chamber for one template. This reduces primer-product formation, product-product hybridization, and prevents chimeric product formation during complex gene library formation ((Keke Shao, 2011). For our application, we had 225bp amplicons for 328 unique genes from the library made in section 1. 1μm hydrogel beads with primers complimentary to adapters ligated to library during library formation were added to em-PCR.

Emulsion-PCR was carried out as per manufacturers instructions. Briefly, 6μl of .5ng/μl of amplified library is added to 94μl of nuclease free water. 100μl of 1μm hydrogel beads were added to the library along with enzyme and primer mix. Emulsion was carried out after adding 2400μl of amplification mix and 200μl of Reaction Oil to make water-in-oil emulsion. After PCR, 150μl of breaking solution (butanol) was added. This breaks the emulsion and beads with amplified molecules were then used to enrich the clonally amplified beads.

Enrichment of the clonally amplified beads:

Clonally amplified beads were enriched to separate beads with no templates, and beads with templates. Streptavidin coated magnetic beads were used to separate beads bearing amplified templates. The library was then used for sequencing.

Sequencing primer hybridization:

Biotinylated sequencing primer was added to the enriched library. 10μl of 100μM primer in TE is added to 2μl of beads in 8μl PBS. Primer is annealed at 60^oC for 15min. and immobilized on the glass surface as mentioned in next section.

3. **Surface Immobilization of the library**

Biotinylated acrylamide beads with amplified library in section 2 were immobilized on a glass surface. The surface is prepared chemically to activate the surface for controlled immobilization of sequencing molecules.

Chemical activation of the glass surface:

Circular glass coverslips were treated with 3:1(v/v) piranha solution for 30min to activate the surface with hydroxyl functionalities for silane modification. Amine groups were attached to the surface by treating with 1% - 2% 3-aminopropyl-triethoxysilane (Sigma Aldrich, 440140) in water for 30min. After washing with milliQ water it was air-dried and cured at 100^oC for 30minu. 95μl of 1M EDC in PBS was used with 5μl of 1mg/ml Streptavidin (Sigma-Aldrich, 85878-1MG) and incubated at room temperature for 1 hour. The glass surface was ready for immobilization of biotinylated library prepared in section 2 (Figure 3.10). 20μl of sequencing primer hybridized library was immobilized on surface and incubated for 1 hour at room temperature. The glass slide was washed with 100μl of 1E wash buffer and dried. The glass-slide was ready for sequencing.

Flow-cell preparation:

Two flow-cells were used in Prototype-3. First configuration was based on the commercial single channel Bioptech flow-cell assembly and the second configuration was an in-house made 4-channel flow cell assembly. The flow cells were assembled as per the sequence in Figure 2.39 and Figure 2.40 Chapter 2. Before assembly, the parts were cleaned and washed with 100% ethanol and water and

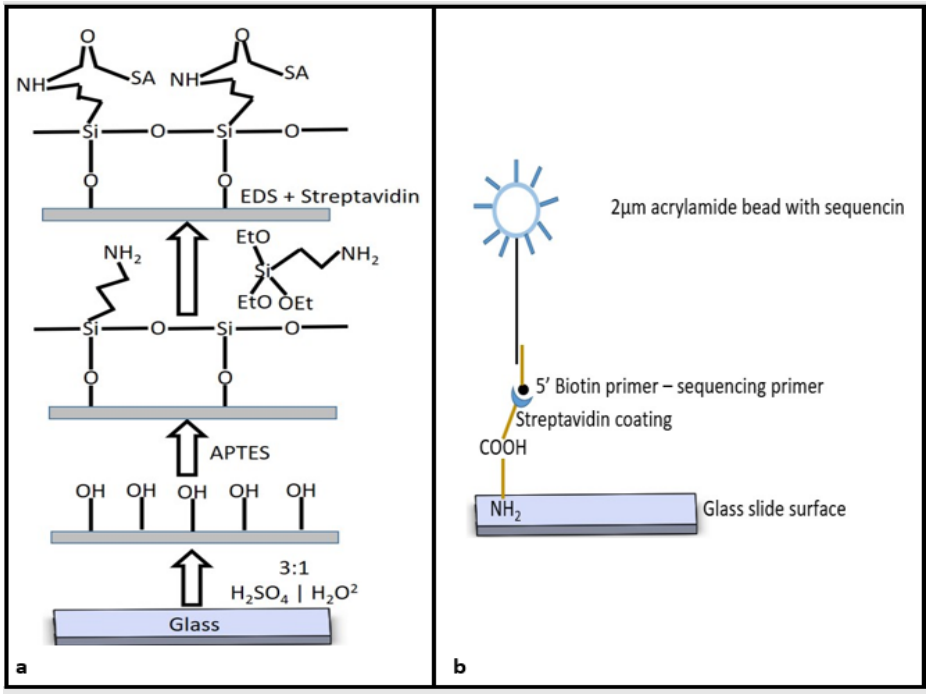


Figure 3.10: a) Glass-activation of the coverslip by amine-functionalization and streptavidin molecule immobilization. b) Biotinylated sequencing primer-attached acrylamide beads loaded onto the activated surface.

dried with air. Plasma treatment was also done to avoid any contamination. Circular cover slip with immobilized library was immobilized in the flow-cell.

Fluid-channel formation in the two configurations/biochemical reaction chamber:

Bioptech flow-cell served as a single channel flow cell. This flow-cell was used because it ensured uniform temperature control over the reaction chamber channel. The upper half of the flow-cell was connected to the perfusion tubes which facilitate the flow of reagents in the fluid flow channel (Figure 3.11). A fluid pathway was made by using a micro-aqueduct slide with a sealing gasket on top of cover slip with immobilized DNA molecules. Fluid access to this flow channel was made through two 14-gauge needle stock tubes protruding from the sides of the chamber top. In second configuration, the fluid channel was formed with pressure sensitive adhesive circular flow channel over the coverslip and silicone

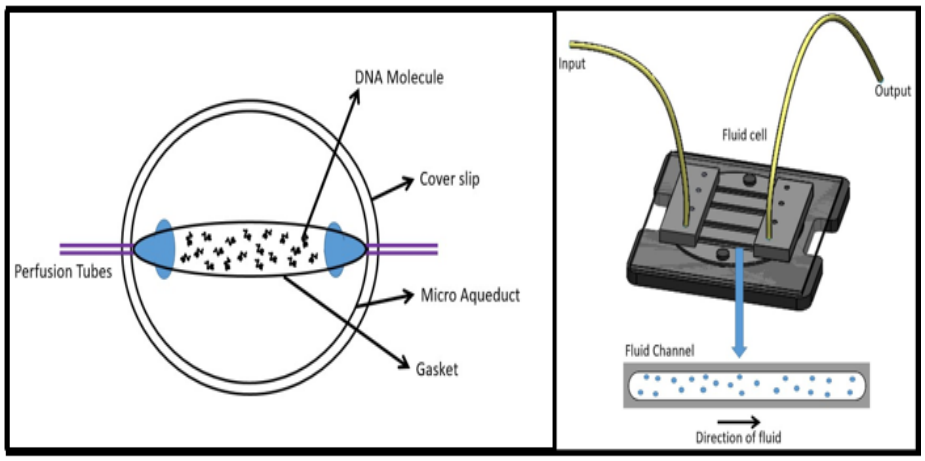


Figure 3.11: Fluid channel formed in 4-channel in-house made flow-cell.

inlet and outlet channels.

Temperature control:

In both the configurations, temperature was controlled for each cleave and extension cycle. For Bioptech flow-cell, commercially available electrical enclosure (060319-2, FCS2 STARTER SET) was used to heat the electrically conductive transparent thin film of Inidium-Tin Oxide (ITO) on Microaqueduct slide (Bioptech, 130119-5) through two electrical contacts (busbars). A 6W temperature controller was used to modulate temperature profile during the sequencing cycles. This passed regulated current flow through the ITO Coating and caused the surface of the microaqueduct slide to heat. The heat was transferred through the reagent to the DNA molecules on the coverslip thereby providing first surface thermal control. The self-locking base of the chamber was also temperature regulated to provide peripheral heat as well. The controller received 5V power from Arduino. In second configuration, a flat aluminum heating block was used which was designed to cover a single biochemical chamber It is controlled by an in-house made temperature controller and connected to a 12V power supply. The controllers were modulated with a Matlab program through modulated current from Arduino in Configurtion-1 (Figure 3.12 a) and modulated current

directly through USB to controller in configuration-2 (Figure 3.12 b). respectively. Both the controllers maintained a fixed temperature of $45 - 50^oC$ as specified in the program.

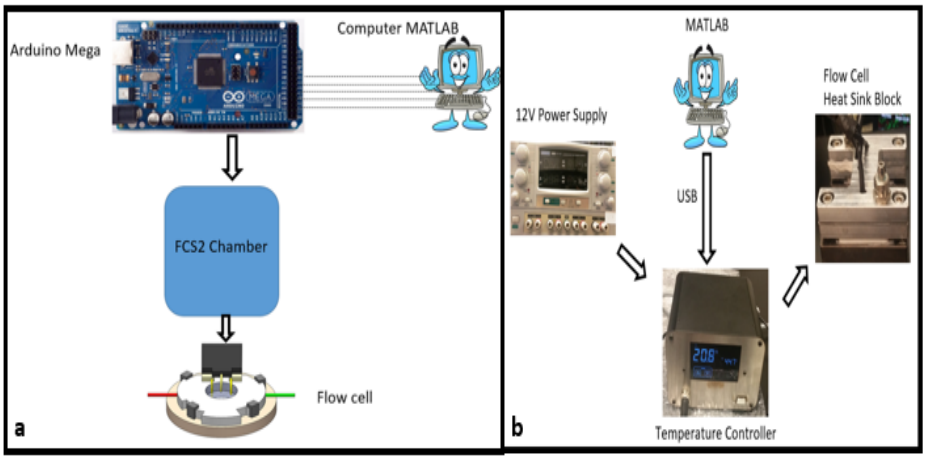


Figure 3.12: a) Temperature controller configuration-1, b) Temperature controller configuration 2.

3.3.2 Fluorescence Imaging

Two high-speed and automated fluorescence imaging configurations were assembled. Two inverted microscopes Nikon inverted microscope TE2000-E2 and Nikon inverted microscope Eclipse Ti-E with motorized stages (Prior, H117 Proscan stages) were used. SpectraX Light engine and X-Cite 120 LED Boost (Excelitas) were used for light source and Hamamatsu CMOS Camera (C11440-42U) and Andor ixon+1 (Andor CCD Camera) were used for Image acquisition. Hamamatsu had 2048×2048-pixel number with 665μm×665μm FOV and 1024×1024 in Andor ixon. Images were acquired with 20X CFI Plan Apo DM 20X; MRD30200 objective in 100-300ms exposure with 5-50 EM gain on the CMOS camera. Images were taken from Bioptech Imaging aperture of 22mm. The configuration-1 and 2 have been discussed in great detail in Chapter 2 section 2.5.1

3.3.3 Flow-cell loading into the system and volume optimization

Reagents:

For sequencing with Prototype-3 set-up, Buffer 1E was used as wash buffer, 0.15M TCEP was used as Cleaving reagent. Therminator DNA polymerase (M02261S, NEB) was used for sequencing for incorporating reversible-terminator ddNTPs.

Optimization of integrated fluidic system:

An integrated and automated fluidic system was built to deliver reagents in controlled and sequential manner to the flow cell. The reagents were delivered in precise volumes through a Cavro XLP 6000 Syringe pump (Tecan) which in turn was connected to 12V power supply and controlled by Matlab through a USB connection. Optimization of volumes of reagents for each of the biochemical reaction chamber was done before actual sequencing reactions for maximum results.

3.4 Sequencing by synthesis

Sample preparation:

Chemically activated slides prepared in section 3.3.1 and assembled in to the flow-cells as mentioned in section 3.3.1. Immobilization of the library was done after the cover slip is washed with IX 1E wash buffer. 10μl of 10μM sequencing primer was annealed with 2μl of acrylamide beads in 8μl of 1M PBS at 60^oC for 15min. After annealing, the beads are directly added to the streptavidin coated glass slide for 1 hour at room temperature for capturing. After capturing the beads, the slide was washed with 100μl of 1X Wash buffer to remove additional unbound beads. Fluidics inlet and outlet were connected to the assembled flow-cell and flow-cell was assembled to the microscope stage.

System preparation:

Prior to loading the assembled flow-cell in the system, system was washed by connecting inlet and outlet connections. 5.5ml 0.15M TCEP as cleaving reagent in 15ml tube and 5.5ml extension solution with fluorescent reversible terminator dNTPs and DNA polymerase enzyme was loaded into the system. System was primed and a FOV was selected and bright field image was captured as discussed in section.

3.4.1 Principal of Sequencing

Fluorescently labelled reversible-terminators are the essence of sequencing-by-synthesis. Single-stranded DNA molecules were immobilized on the streptavidin coated glass surface with biotinylated sequencing primer. Fluorescently-labelled dNTPs were added to 3 end of the primer which allowed only one incorporation at a time preventing mis-incorporations. The 3end was blocked by a hydroxyl group which did not allow further incorporations. Therminator DNA polymerase was used for extension reaction. After the extension, hds occurred, camera captured fluorescence signals to detect the fluorescence after each incorporation. 3 hydroxyl group was cleaved by 0.15M TCEP. Washing was done after cleaving to ensure that all cleaved fluorophores were washed and not carried to further reactions.

3.4.2 Data Acquisition

33 or 34 sequencing cycles were done to sequence 34-36 base-pair reads. For each base, a fluorophore activity was captured. 2 frames were captured each cycle with different FOV. A bright-field image was captured for each FOV (Figure 3.13) and was used as a reference-image for Base-calling as discussed in next section. X and Y co-ordinates were specified and focus (Z) was adjusted for proper data acquisition. Manual sequencing cycles were run and colored images were obtained to ensure optimization of each experimental variable. Auto-run was set-up for continuous 35-36 cycles. Finally, when colored images were obtained image processing was done to

prepare Image files for data/base calling.

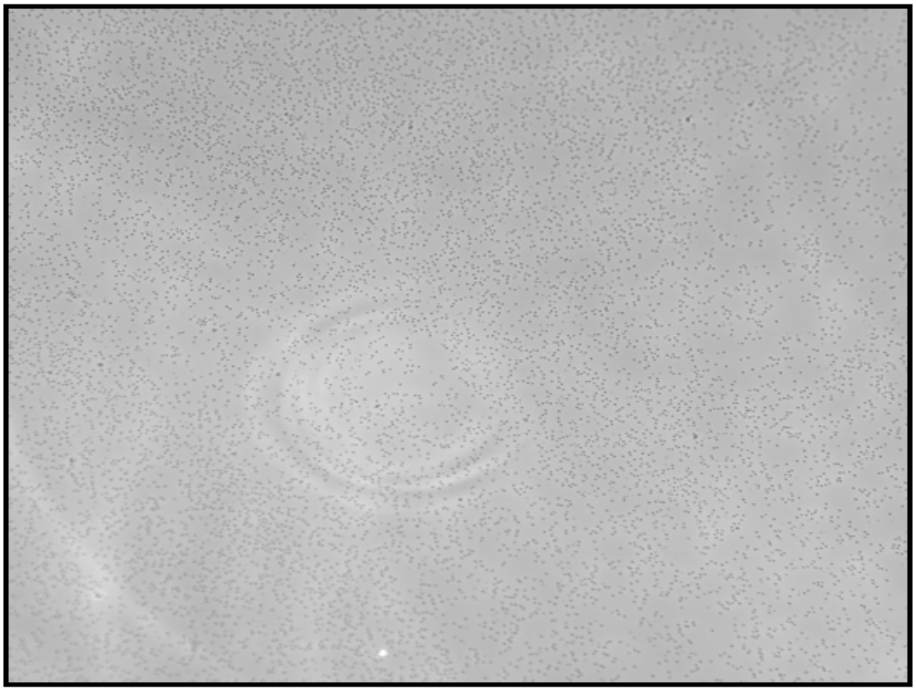


Figure 3.13: Brightfield image of field of view chosen for first frame for a sequencing run.

3.4.3 Image analysis

The colored Images obtained after 36 cycles were processed to do the base-calling. The colored Images and 2 format files were converted to tiff files which had the data in form of dot matrix with pixel intensities. Bright-field image taken as the base-image was aligned to one of the Tiff image and processed by using ImageJ to produce a base image that was sharper and clearer for to be used by the program for aligning the colored images and calculating the intensities and base-calling.

ImageJ processing

A single Image was processed as a reference to be used for aligning images (Figure 3.14) of other cycles for the same slide/flow-cell/filed-of-view. Usually, a brightfield image was used as the reference image to minimize any shifts of the X-Y coordinates

during sequencing run. Tiff files images were sharpened and processed for making individual beads accessible for the program for aligning pixel coordinates to the images. Usually, for sequencing, multiple cycles were run, each cycle with multiple

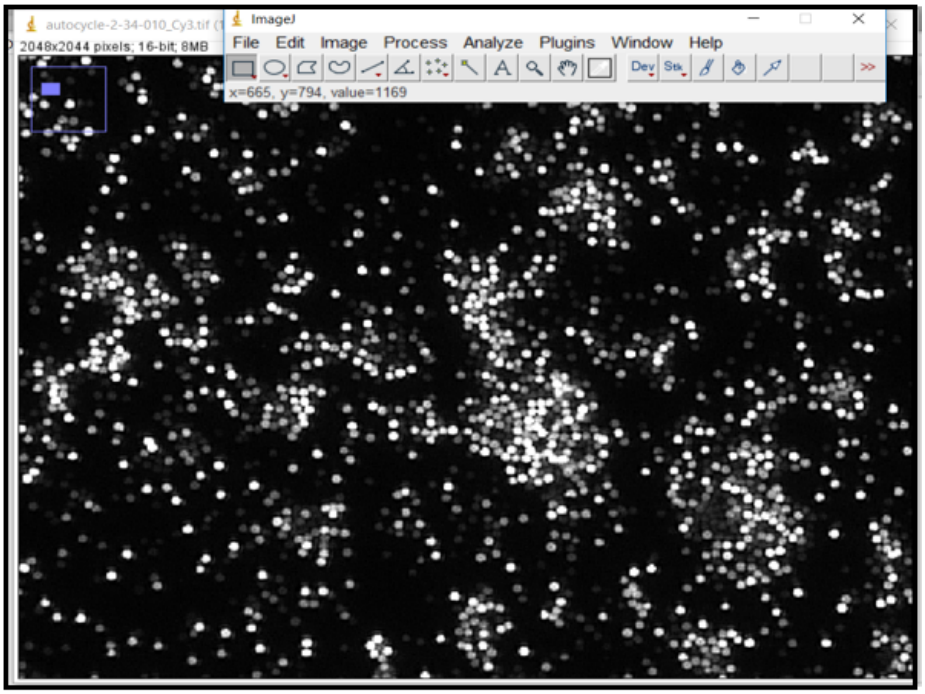


Figure 3.14: Cy3 and Cy5 Tiff images merged and processed by using ImageJ. Pixel intensities and image resolution was optimized.

frames. For our data, we had used two frames per cycle. This way each cycle would have 4 images, two for Cy3 and two for Cy5. BF Image for each frame was used as a reference for any off-sets of X-Y coordinates per image and provides as a consistent data for position of beads against which the addition of a base could be measured.

Image threshold

Image thresholding was done by pixel-by-pixel thresholding for a bead. Each individual bead comprises many number of individual pixels. The pixel intensity was defined by integer intensity from some min to some max value. Center pixel of a bead is calculated and used as intensity for this bead. These center pixel intensities are used for base-calling.

Base calling

The process starts with one reference white field image, one red image (Cy3) and one green image (Cy5). The red and green images were merged and saved as Tiff file. The saved image and the reference image was aligned and the total number of beads (objects) were counted and saved. This image was also used to extract the position of the beads. Next for each position the individual red and green images were aligned (one by one) and the pixel intensities (above threshold) were recorded and converted into the bases (red for C, green for T, yellow (red + green) for A and grey (no red or green) for G. The bases were also recorded by the position of individual beads. This is base calling for one cycle. The same process was repeated for all the consecutive cycles and bases were recorded for each position in a string (text) file.

One bead one base one signal

Colored images were scanned to calculate the intensities per bead per cycle for 2 images. (Figure 3.15) These scanned images are then aligned to reference image. Bead location and corresponding intensity are determined for each image and recorded. Based on the pixel intensity profile of the bead at a position bases are called. This is done for 35 cycles. Each base was given position in the 35bp read per the cycle sequence and thus 35-37bp reads sequences were determined.

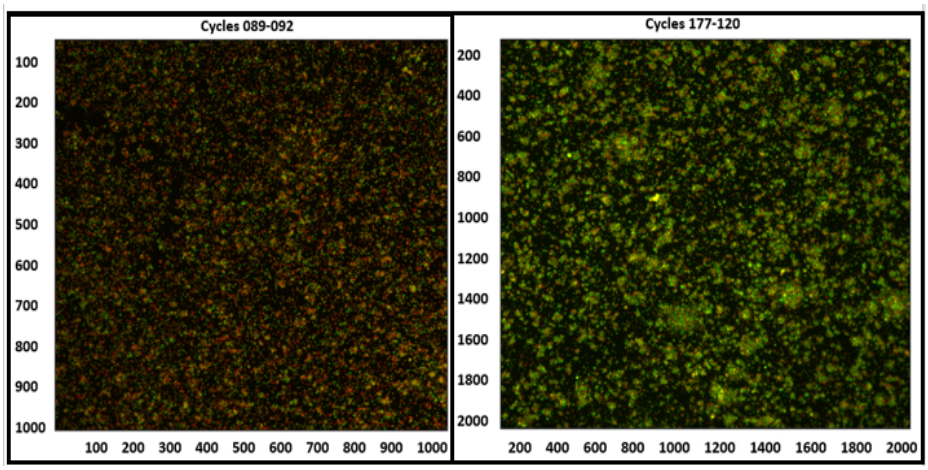


Figure 3.15: Raw Images from system configuration-1 and system configuration-2 for base-calling.

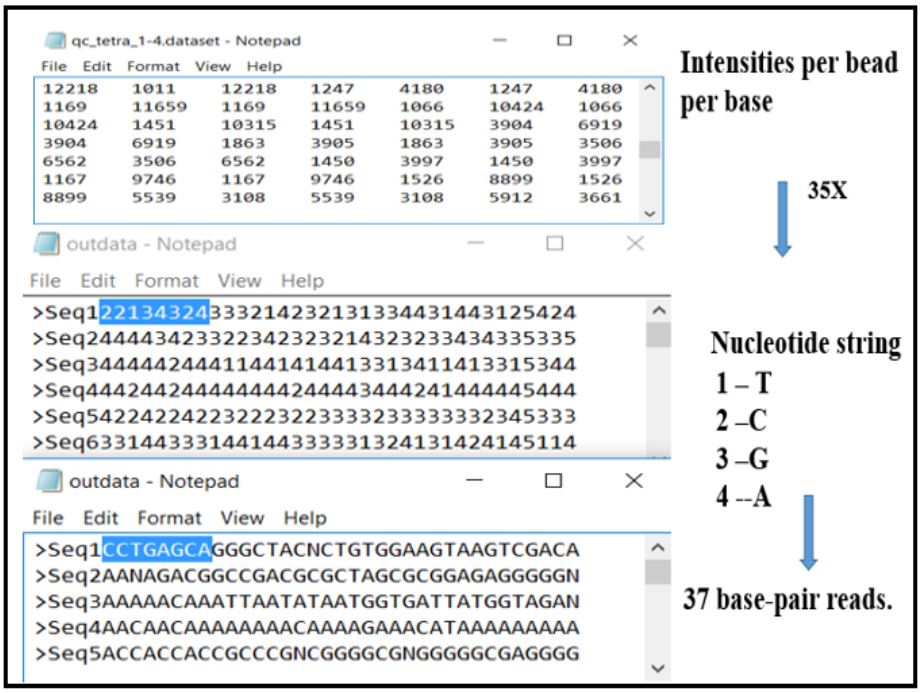


Figure 3.16: Base-calling: 2 images per sequencing cycle and 2 tiff images are aligned to reference image. Intensities are calculated for each bead per cycle based on center pixel intensity. A fasta file is created of the numerical string as an output per sequencing run.

The Quality of the base is determined by determining the mean intensity of the 2 images per cycle. Base-call intensity was plotted against this mean intensity and distance from the mean determines the quality of the base.

3.4.4 Data Analysis

Reads Quality

For evaluating the overall base quality of the sequenced data, we aligned the reads to NCBI human reference build 37 with bowtie-2 algorithm version 2.3.1. The reads mapped to the genome with 40% overall alignment and ~95% accuracy. 48,551 reads with mean read length of 35.63bp were mapped with 16% of mapped reads being non-specific matches. This implies that 16% of reads align at other positions with same mapping quality. 99% of the mapped reads were non-perfect matches containing one or more mismatches. Bowtie-2 had higher mismatch fraction of 0.4 at the 30-35bp

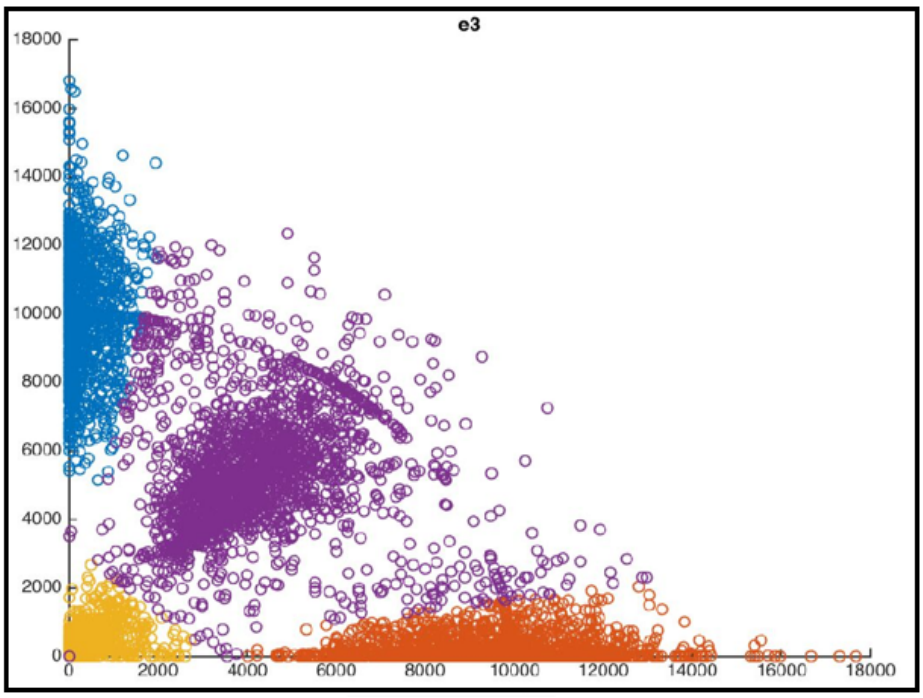


Figure 3.17: Intensity of bases in 2 images plotted against the mean intensity. The farther the base from the mean, lower the quality of the bases.

position in the reads mostly (Figure 3.19). This suggests that sequencing quality decreases with the increasing length of the reads. We can see in figure 3.18 that 0.94% of total reads in our sequencing data differ from the reference. On mapping mismatches of reads to the we can see that A and T mismatches are called mostly. Few of these errors could also be true single nucleotide polymorphisms and there are few mismatches called.

Targeted amplicons coverage statistics

Aligned reads were filtered with BED file to measure the coverage across the amplicons included in the panel. As there were 10,500 primers pairs for 326 genes. Total mapped bases numbered to 65,589 among which 34,483 were in target region with total specificity of 52.57% for 22 autosomes (Figure 3.23), X and Y chromosomes. For comparison basis, we also used CLC version 10.0 to map the reads to human reference and then filter for targeted regions. Out of total 104,469 mapped bases, total 46,608

Nucleotide differences in reads relative to reference

Nucleotide in reference	% read bases that differ
A	0.75
C	0.90
G	1.10
T	0.65
-	68.20
Total	0.94

Nucleotide Mapping

	Read: A	Read: C	Read: G	Read: T	Read: -	Total
Reference: A	33,592	85	29	100	39	33,845
Reference: C	34	13,770	26	22	43	13,895
Reference: G	21	34	15,082	78	34	15,249
Reference: T	80	27	84	34,425	34	34,650
Reference: -	39	36	38	35	69	217
Total	33,766	13,952	15,259	34,660	219	97,856

Figure 3.18: Error in reads when aligned to reference human genome hg-19. 0.94% of bases in reads differ from reference. Nucleotide mapping suggests that A and T are called in high numbers than G and C.

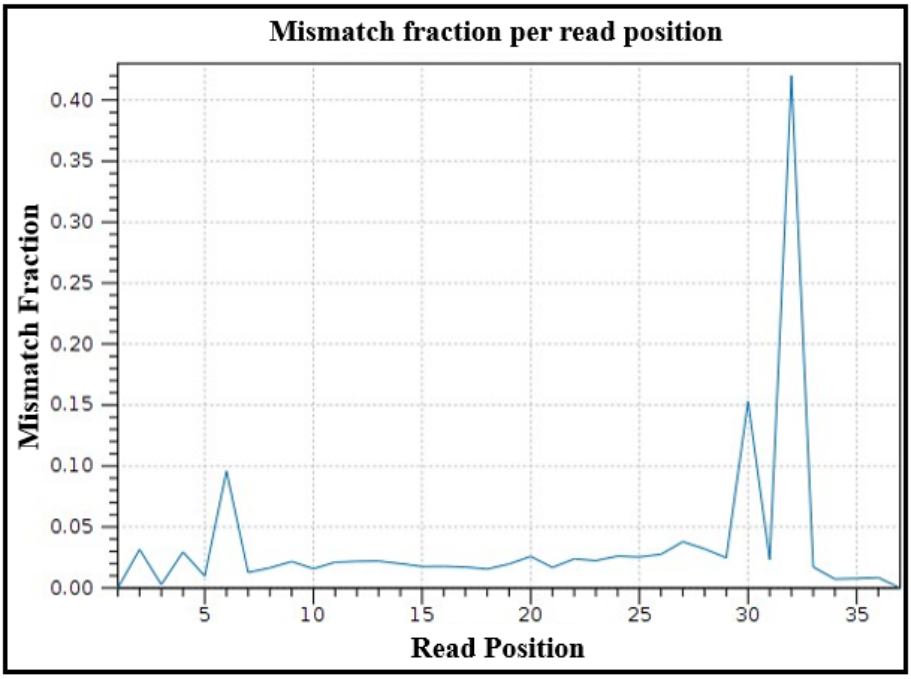


Figure 3.19: Mismatch fraction in bowtie-2 mapping.

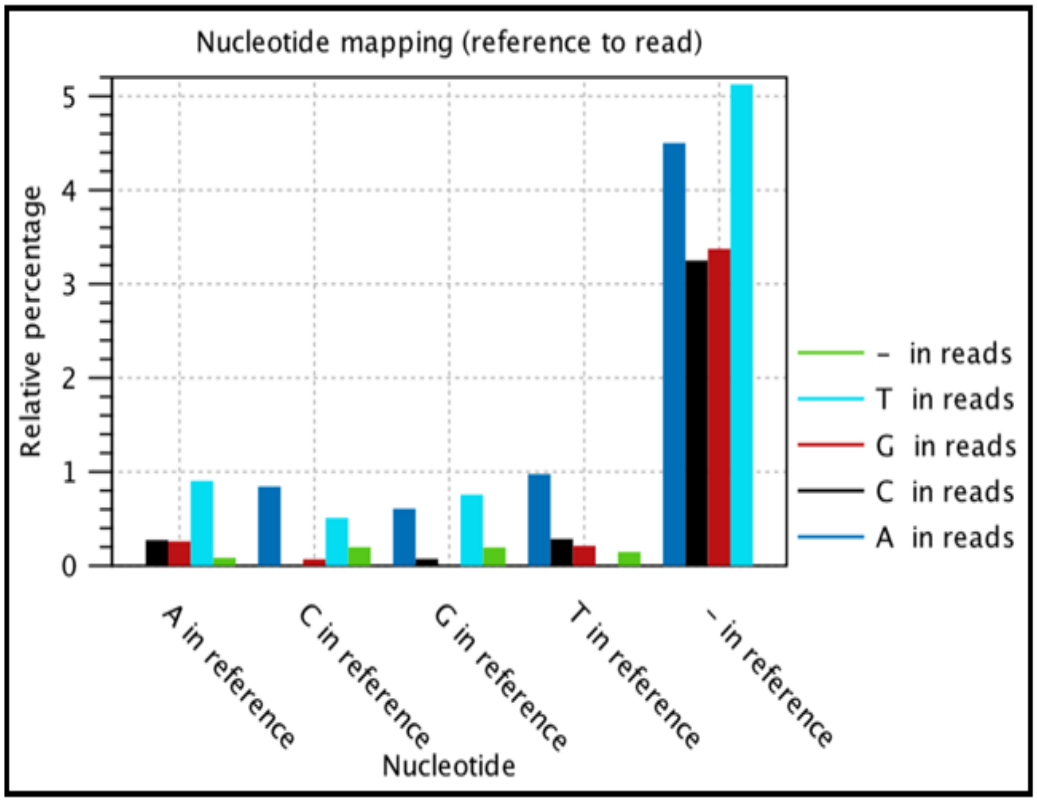


Figure 3.20: Relative errors in mapping from reference to reads.

bases map in target regions with 44.61% of the bases in target region. Decrease in specificity can be due to higher number of non-perfect matches as compared to Bowtie-2. Although, total mapped reads are more than bowtie-2 in CLC but most of them are non-perfect matches. The coverage of amplicons per chromosome ranged from 0 to 80X (Figure 3.21 and 3.22) giving high confidence in some of the SNPs called and demonstrated accuracy of the data. Chromsome X has highest coverage across the amplicons in both mappings. Chromosome 6 has coverage upto 80X in CLC and 52X in Bowtie-2. 16,21, and Y chromsomes have lowest coverage as the number of amplicons in the panel in these chromosomes is lower than other chromsomes with 263, 159 and no amplicons for Y chromosomes respectively. 5.9% bases of reference are covered at 1X coverage. Accuracy of the sequencing data was also determined by plotting GC content of the reference to mapped reads coverage(Chen, Liu, Yu, Chiang, & Hwang, 2013). Figure 3.25 shows that as the GC content increases, the number

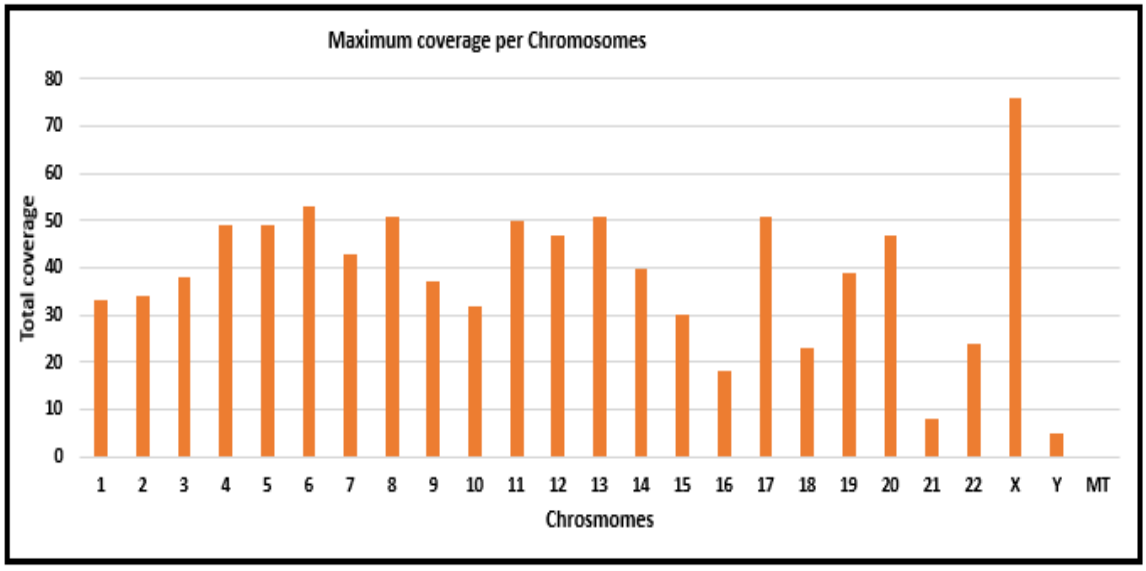


Figure 3.21: Maximum coverage per chromosome in Bowtie-2.

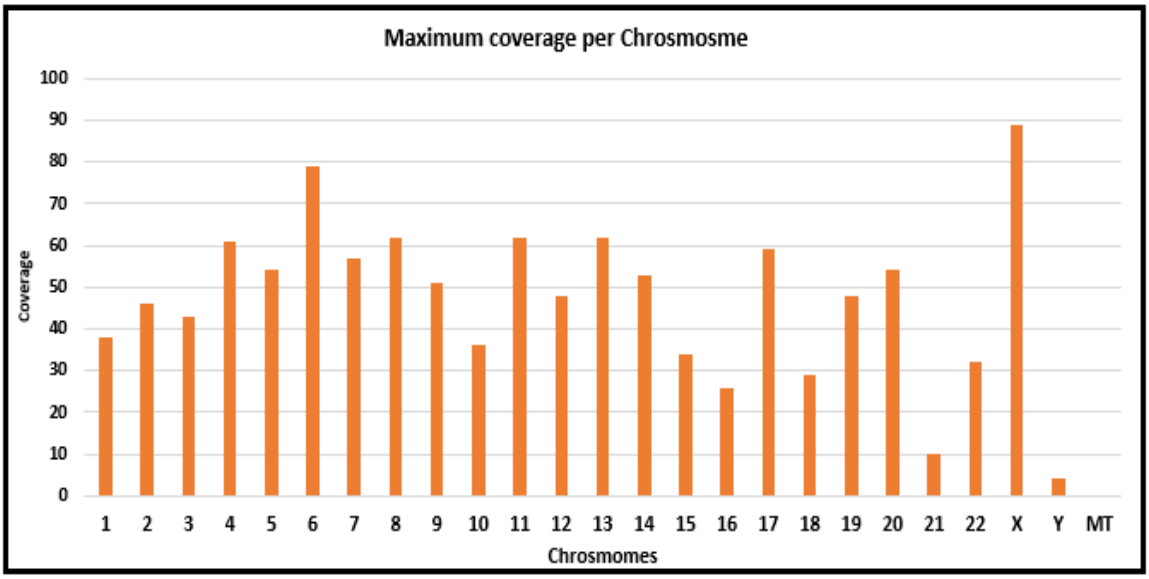


Figure 3.22: Maximum coverage per Chromsome CLC.

of reads also increases thus, there is no observed GC-bias in the data confirming accuracy.

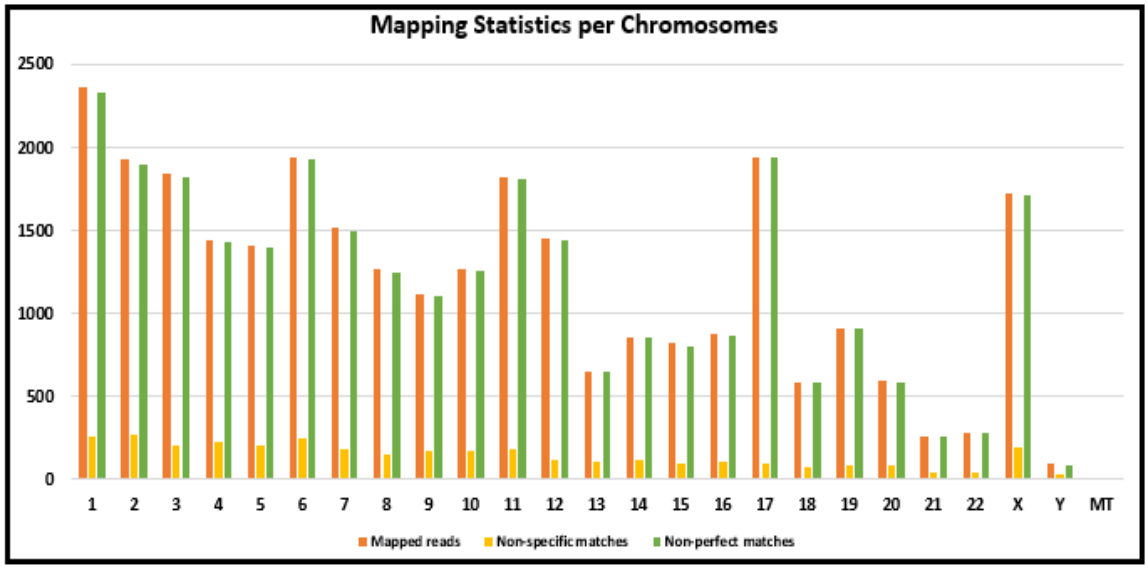


Figure 3.23: Total mapped reads per Chromosome. Bowtie-2 is used for mapping reads to human genome hg-19.

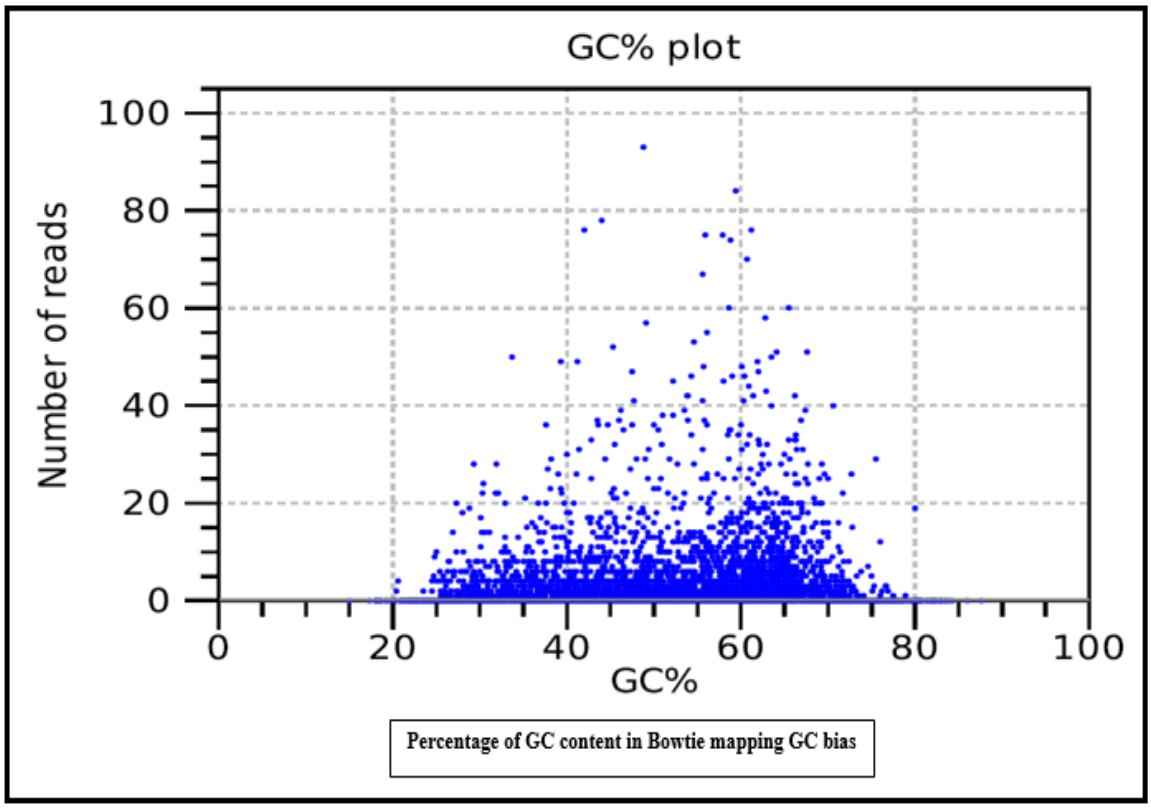


Figure 3.24: GC-bias plots of GC content of bases in reference and number of reads aligned to it.

Variant calling

35bp-37bp reads were obtained from 6 datasets as a result of sequencing. The dataset was combined to one fasta file. A fastq file was generated with a perl script. The reads were aligned to the human exome hg19 data by CLC and Bowtie2. Variant calling was done by CLC basic variant calling and bwa mpile-up. Bowtie-2 was used in default mode and then parameters were tweaked to allow for more mismatches. Mapping in default mode resulted in 48% reads aligning more than one time and allowing more mismatches allowed 80% reads mapping more than one time. Annotation was done with CLC version 10.0.

Mismatch annotation:

To analyse the accuracy of the data, we analyzed the data for mismatches and annotate with potential SNPs as should be present in exome data. Since, our data was low-coverage (∼5X-6X) and single sample, to gain confidence in our SNP calling we used two methods to confirm our data. SNPs were validated with dbSNP database for hg-19 human sequence. Further, to achieve agreement between two methods, positive calling rate(Hwang, Kim, Lee, & Marcotte, 2015) was used as a comparison method for accuracy. CLC-basic variant calling module and samtools mpileup were used for SNP calling. Basically, 60,628 reads of 35-37 base-pair length were mapped to human reference genome NCBI build 37. The mapping was done by CLC version 10.0 (www.clcbio.com) and Bowtie-2 (bowtie-bio) algorithm with very sensitive local option which gave 48% over-all alignment rate with 29000 reads mapping once or more than once. To increase the alignment rate, maximal mismatches were allowed (N1 option, bowtie-2) resulting in 80% reads mapped with 48551 reads mapping once or more than once. CLC mapping for basic variant calling resulted in 71.56% or 43388 reads mapped total. Samtools mpileup version was used to sort, align and create index of mapped reads for variant calling. Mapping quality adjustments were disabled and max depth is kept 1 (-C 0, -d 1). SNPs were called with bcftools called

and filtered with vcfutils. Annotation was done with hg-19 SNPs with CLC. The vcf files obtained were then filtered for insertions and deletions. This data was then further analyzed for concordance and differences in two methods.

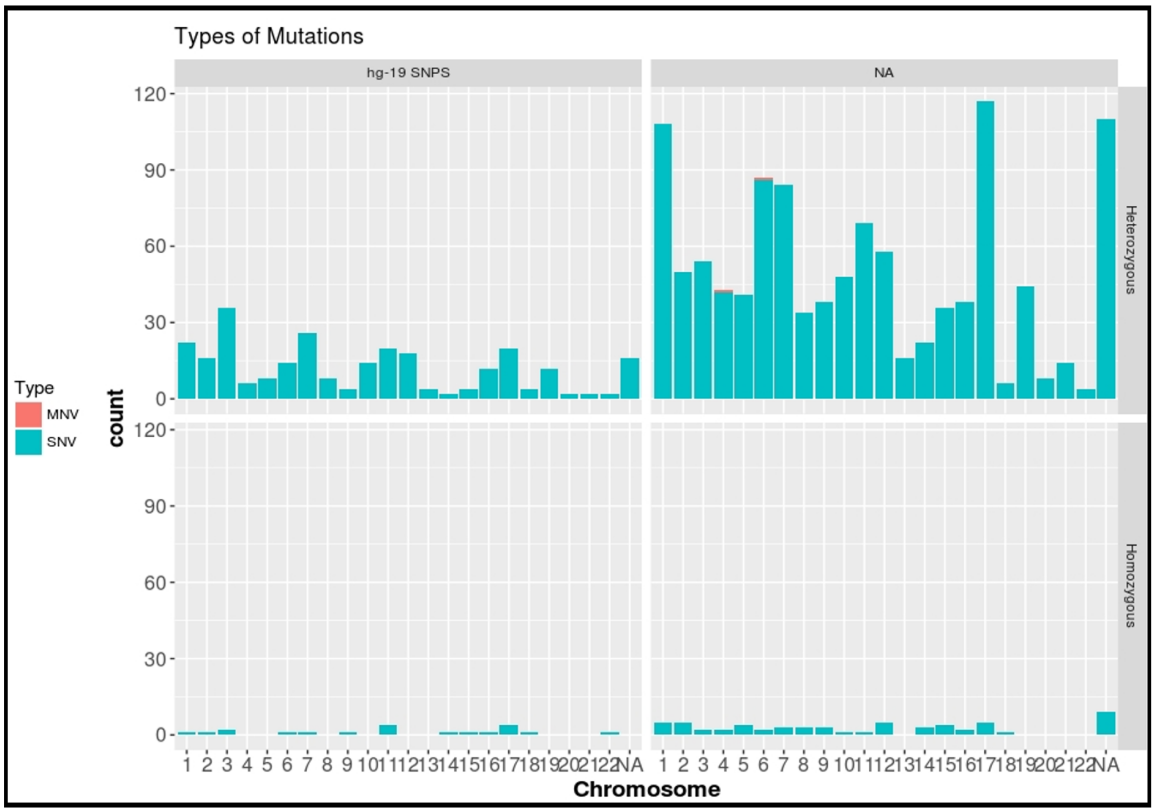


Figure 3.25: SNPs called by Bowtie-2-samtools-mpileup. The SNPs in dbSNP database are labelled in hg-19 SNPS quadrant for both heterozygous and homozygous SNPs. NA here denotes X-chromosme. The data is filtered with BED file and also insertions and deletions are removed.

Figure 3.26 shows that CLC calls more SNPs than samtools mpileup and this is because stricter variant filtering done by bcftools. Although, there are more positive calls (SNPS annotated in dbSNP database) made by CLC, it also calls more false SNPs. On allowing more mismatches in bowtie-2 mappings the number of dbSNP called increases, and gives same number of SNPs as CLC when variant calling is done by CLC (Appendix Figure A1.2). CLC does call some replacement SNPs which are not called by Bowtie-2. These SNPs are multiple nucleotide changes from the reference and overlap with hg-19 SNPs database (Table A 1.2). Most of the mutations when

categorized are nonsynonymous single nucleotide variations which cause the amino-acid changes and thus phenotypic differences (Hwang et al., 2015) in individuals. Since, this is a human exome inherited disease panel, such mutations are expected from this data. The exome panel used in this study has 10,300 primers for 326 genes.

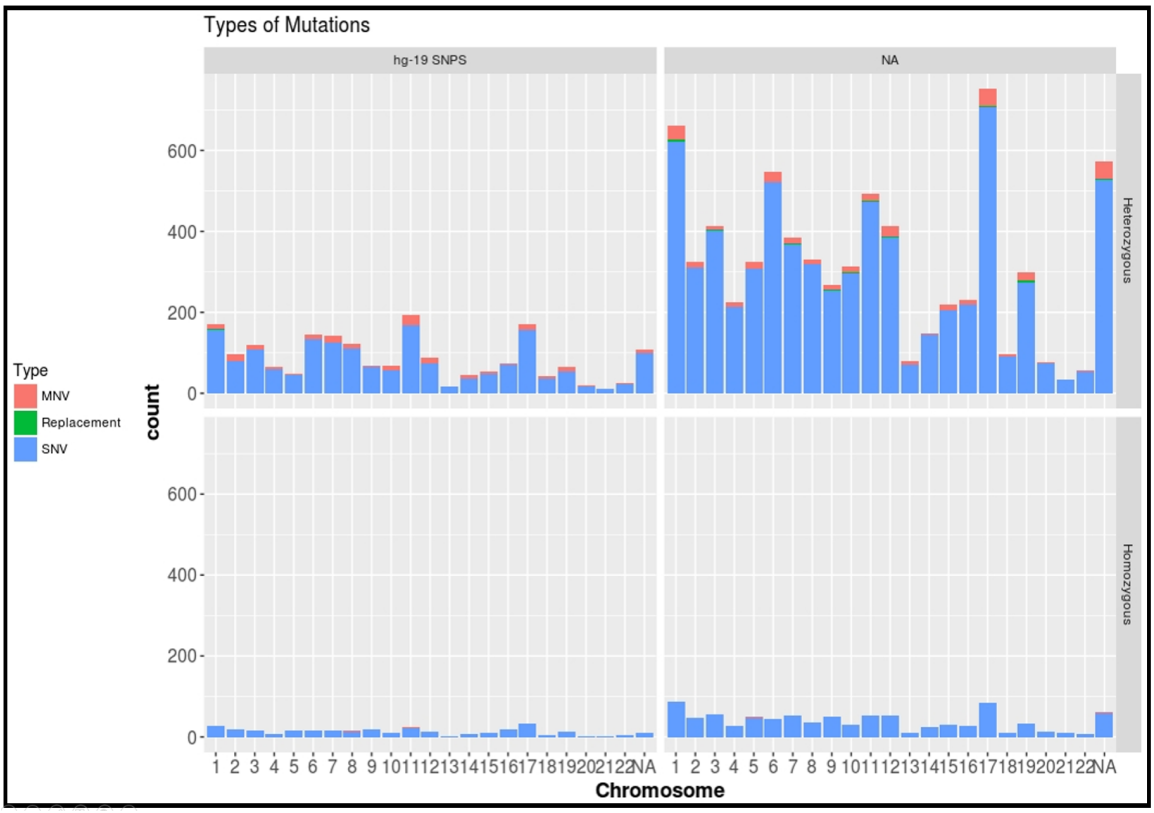


Figure 3.26: Total SNPs called by CLC-Basic variant calling. The SNPs in dbSNP database are labelled in hg-19 SNPS quadrant for both heterozygous and homozygous SNPs. NA here denotes X-chromosme.

With our sequencing data, we are able to cover 293genes specified in BED file with additional 2444 genes (Figure 3.28 b). When the data is filtered with BED file we cover 271genes (Figure 3.28 a.), cumulatively for CLC and Bowtie data. Bowtie-2-mpileup vcf file annotated with CLC workbench yields the highest number of genes. It is also observed that after filtering with BED file, most of the homozygous true calls are lost in Bowtie and CLC both (Supplemetary A1.2). We could call variants for 212 genes total in concordance with CLC and Bowtie-2 after filtering the vcf file

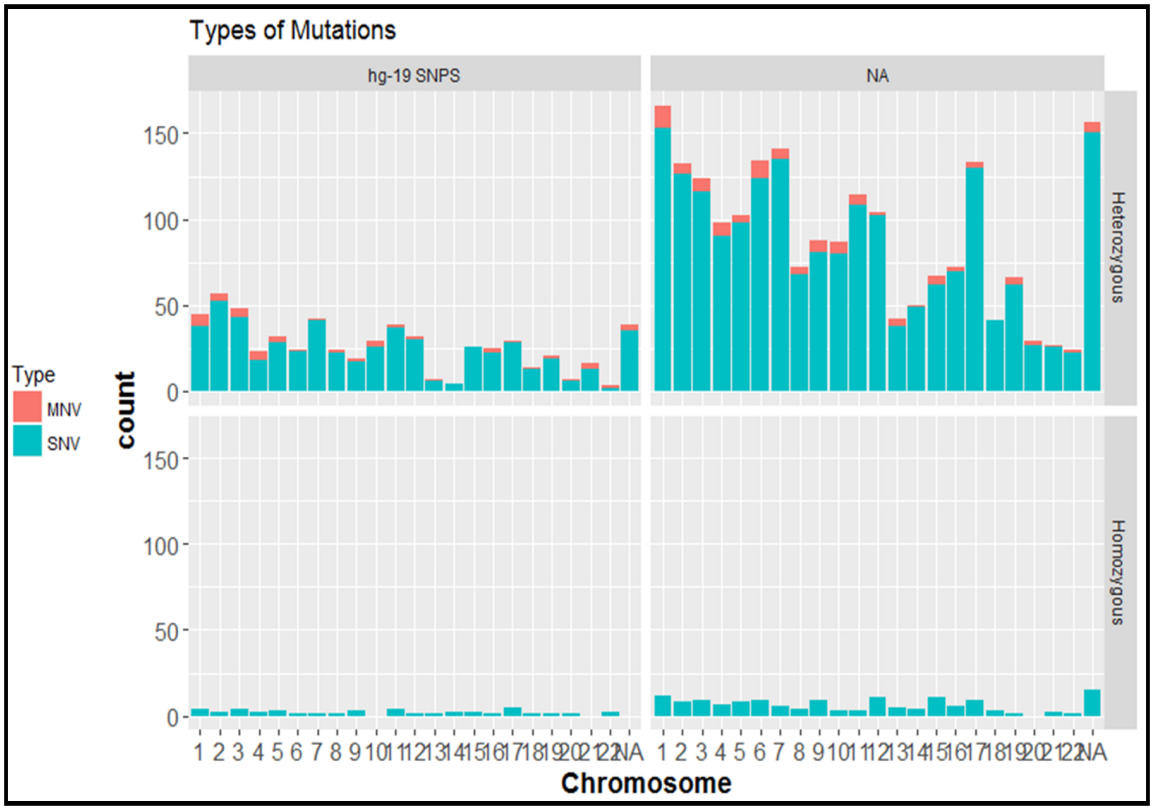


Figure 3.27: Total SNPs called by Bowtie-2 when more mismatches are allowed. The SNPs in dbSNP database are labelled in hg-19 SNPS quadrant for both heterozygous and homozygous SNPs. NA here denotes X-chromosme.

with BED file and removing insertions and deletions., number of total genes covered when data not filtered with BED file is 293 genes in concordance with both CLC and bowtie-2-samtools-mpileup (Figure 3.28 b). 212 genes (Figure 3.28a) called are in concordance with both methods of calling mismatches. CLC has 26 unique genes and 5 genes in common with Bowtie only, whereas Bowtie has no unique genes but 3 in common with BED file only, when filtered with BED file. When data is not filtered with BED file, 45 genes are called by CLC and Bowtie both which are not in BED file. 76 genes in BED file are called by Bowtie alone in common to BED file, thus, demonstrating that 291 genes are covered by Bowtie only when not filtered. Bowtie-2 calls total 2280 unique genes. When same data is used for comparing true SNPs from hg-19 database, 876 dbSNPs are in concordance with data both BED file

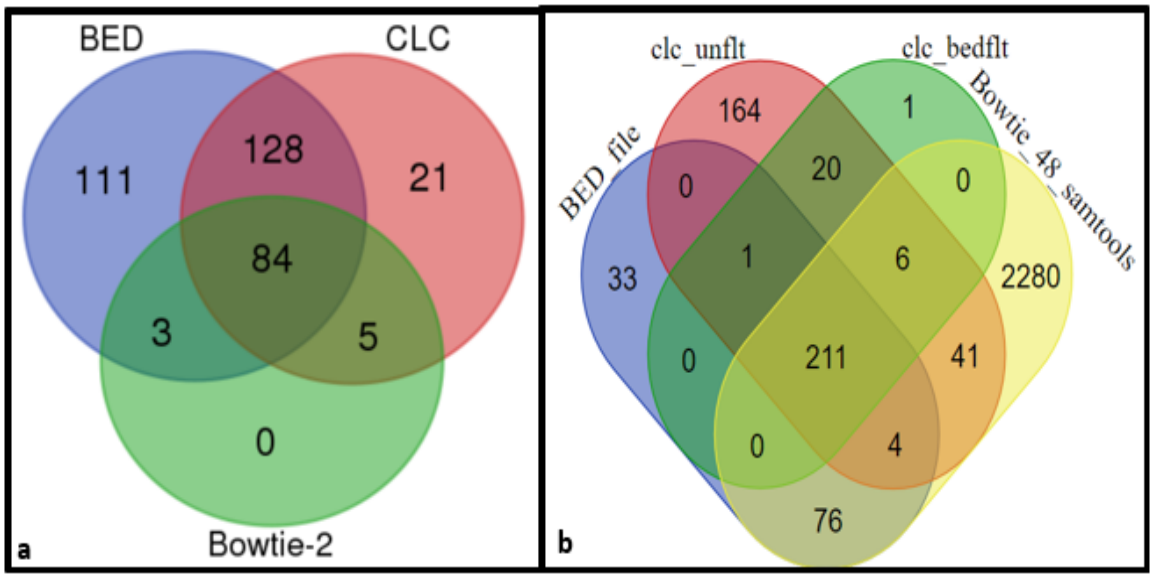


Figure 3.28: Total number of genes covered by Bowtie-2 and CLC after filtering for insertions and deletions.

filtered and non-filtered. 43 dbSNPs are unique to Bowtie and CLC has 431 unique dbSNPs (Figure 3.29).

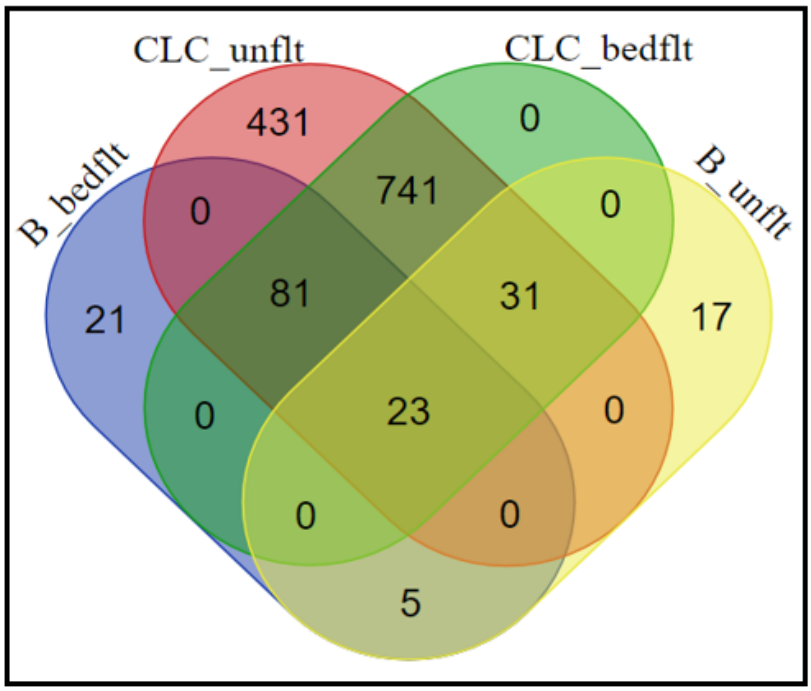


Figure 3.29: Total dbSNP Ids called by CLC and Bowtie. B_unflt-bowtie data not filtered with BED file. CLC_unflt-CLC data not filtered with BED file.

Transition vs. transversion ratios

Transition and transversion ratios are considered as a bench-mark measure for SNP calling (Liu et al., 2012). Ti/Tv ratio gives an estimate of false positives called or estimate of accuracy of SNP calling. The whole-genome transition to transversion ratio for humans is around 2.0-2.2 (Baes et al., 2014) and little higher in whole exomes for about 2.8(Ebersberger, Metzler, Schwarz, Paabo, & Pa, 2002). Low-coverage sequencing gives less confidence in variant calling and more false positive rates but on other hand, with stricter variant calling modules for high coverage sequencing data could lead to loosing true positive calls and sometimes novel SNP calls. We calculated Ti/tv ratios for targeted regions within the sequenced data as specified by the BED file. The samtools-mpileup SNP calls had a variation of 0.3-0.5 ti/tv ratio and CLC had constant ratio of 0.5. Such low Ti/Tv ratio indicates high number of false positives as evident in Figure 3.31. Also, we are calling mismatches and then annotating them instead of calling high quality SNPs. The low-ratio could also be accounted because of targeted regions including intronic, intergenic, UTR, non-coding exonic and intronic, upstream and splicing regions along with exonic regions. One way to increase or improve ti/tv ratio would be to separately calculate for dbSNPs and non-dbSNPs called by two methods. Mismatch calling allows more insertions and deletion called as in CLC and bowtie-N1. Bowtie-2 sensitive option allows less number of insertions and deletions (Figure 3.30 c). CLC calls more insertions whereas Bowtie-2 mismatch calling option calls more deletions. Bowtie-2 sensitive option allows less mismatches, thus only 1bp insertions and deletions in equal numbers are called. Bowtie-2-mismatch option calls highest number of substitutions, and has higher variation in ts/tv ratio (Figure 3.31 c). CLC calls less number of substitutions than Bowtie N1 and has constant ts/tv ratio (Figure 3.31b). All three have highest number of A>T insertions and T>A deletions. As more mismatches are allowed there are higher number of substitutions.

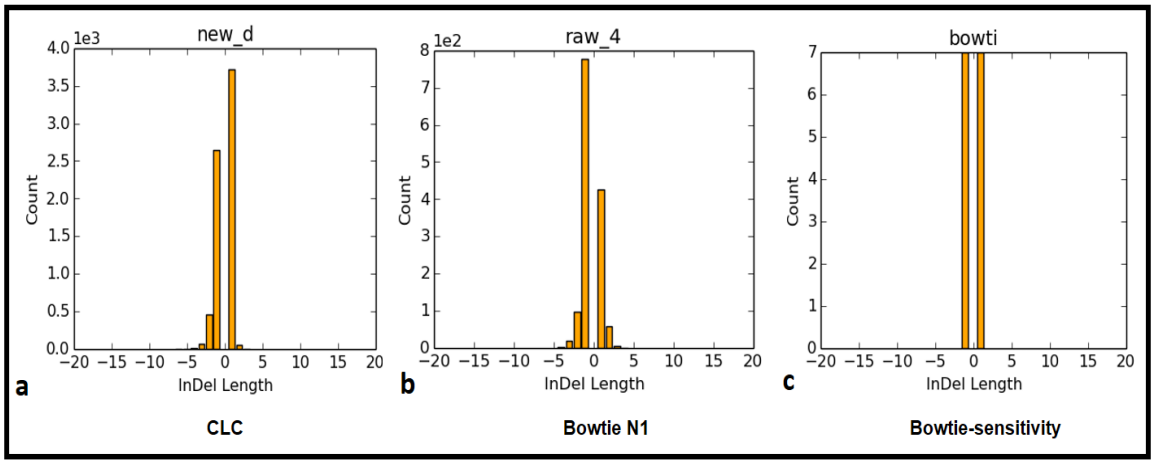


Figure 3.30: Number of Insertions and Deletions called by CLC, Bowtie-1 mismatch calling and Bowtie-2 sensitive calling.

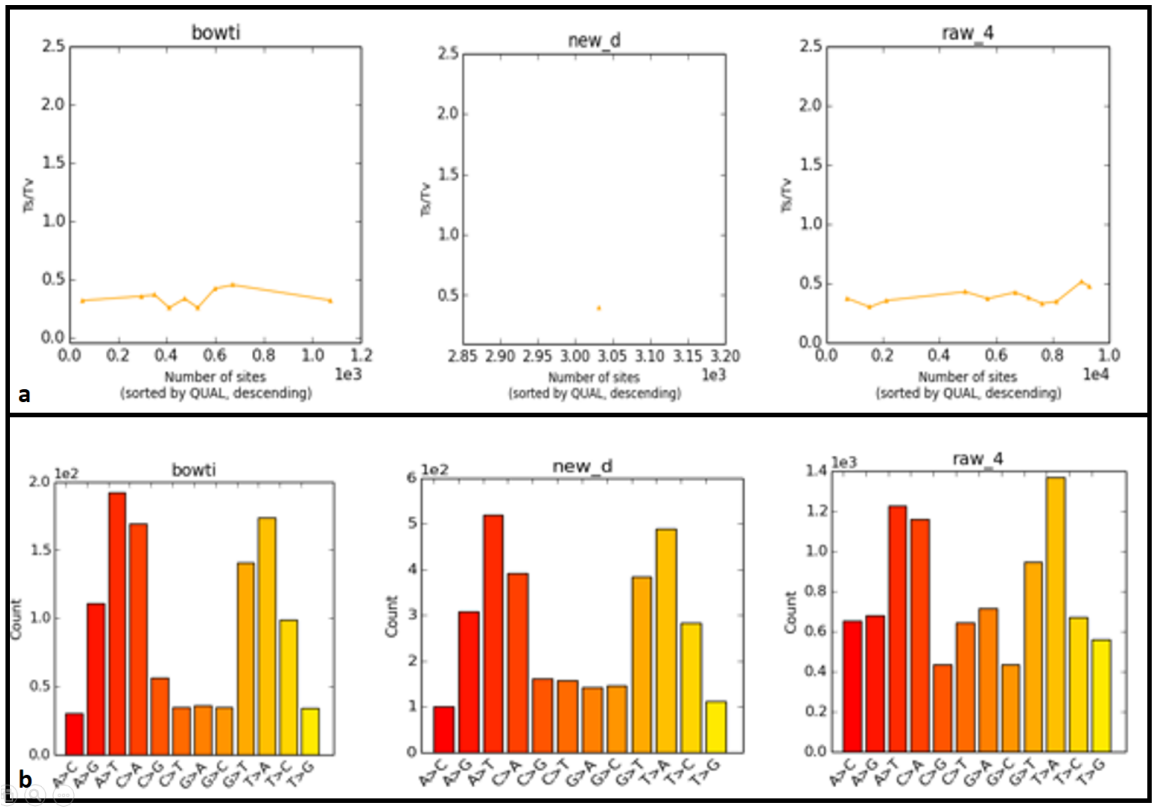


Figure 3.31: .) Ts/tv ratio of Bowtie sensitive, CLC and Bowtie-N1, b) Substitutions in Bowtie sensitive, CLC and Bowtie-N1.

3.5 Discussion

We have developed a sequencing prototype for cheaper, scalable and customizable sequencing for whole genome or targeted re-sequencing approaches. This prototype is an attempt to bring down the sequencing cost due to expensive instrumentation, lack of customization and scalability. Our prototype is assembled with affordable electronics and fluidics components which can be easily assembled and automated with any interface based programming. We have developed and optimized new technique for immobilization of sequencing molecules on regular microscope glass coverslips thus, eliminating the need for expensive flow-cells and chemical reagents. Libraries can be made as per user preference and chemistry can be customized for targeted, whole genome, exome or RNA sequencing. Firstly, we used commercially available Bioptech flow-cell for creating a chemical chamber and temperature control for sequencing which served as single channel flow-cell. Then, for scaling purposes we developed in-house flow-cell with 4-channels which can be run simultaneously to increase the throughput of sequencing experiments. Light-emitting diodes are used instead of expensive lasers. We eliminate major challenges of optical detection based sequencing-by-synthesis sequencing technologies as follows:

1. Cross-talk between fluorophore emission spectra We use 2- fluorophores to detect 4-bases. Each fluorophore is excited at a time and image is taken. The two images are then merged with NIS element software. Thus, we do not require four lasers to excite the 4 dyes simultaneously and introduce overlapping emission spectra. Also, incomplete excitation due to lag time introduced when lasers are warming-up is eliminated by using LEDs.

2. Eliminate homopolymer issues By using reversible terminator dNTPs we ensure that only one nucleotide is added at a time, so that there are no multiple incorporation and wrong signals are not detected.

3. Phasing due to inefficient cleaving or T-fluorophore accumulation- Our design of single channel and 4-channel flow-cells streamline the flow of reagents. We optimize the volumes and introduce the air-bubbles to separate the reagents from mixing and to ensure complete cleaving and washing after extension. Manual cycle optimization before each sequencing run ensured that cleaving was perfect and there was no carry over to next cycles.

   Wash buffer 1E and Cleaving reagent were made in lab thus, reducing the cost of reagents. We developed our base-calling pipeline to convert intensities to base-calls and quality scores. To demonstrate the efficiency of sequencing, we sequenced Ion-torrent inherited disease panel which covers 326 genes. Accuracy of our data was analyzed by mapping to hg-19 human reference 0.94% bases differing from the reference. We could call total 1350 dbSNP Ids for both CLC and Bowtie-2 samtools mpileup data covering 238 genes. Low ts/tv rates can be due to more false positives called. Although, our data is low-coverage we have demonstrated that our prototype works and with some optimization can be used for clinical settings and routine sequencing in clinical and research settings.

3.6 Future work

The prototype presented can be optimized further for increased data coverage by allowing more frames to be captured per cycle. Bio-informatic analysis can be optimized for calling mismatches or real variants. High coverage and multiple sample targeted sequencing can be done by using 4-channel flow cells.

References

Ariella Zivelin, Nurit Rosenberg, Hava Peretz, Yonit Amit, N. K. and U. S. (1997). Improved Method for Genotyping Apolipoprotein E Polymorphisms by a PCR-Based Assay Simultaneously Utilizing Two Distinct Restriction Enzymes. Clinical Chemistry, 43(9), 16571659. Baes, C. F., Dolezal, M. A., Koltes, J. E., Bapst, B., Fritz-Waters, E., Jansen, S., Gredler, B. (2014). Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. BMC Genomics, 15(1), 948. https://doi.org/10.1186/1471-2164-15-948 Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLoS ONE, 8(4). https://doi.org/10.1371/journal.pone.0062856 Demers, L. M. (1999). Laurence M. Demers* Betty Smith. Clinical Chemistry, (9), 15791580. Ding, X., Boney-montoya, J., Owen, B. M., Bookout, A. L., Coate, C., Mangelsdorf, D. J., & Kliewer, S. A. (2013). NIH Public Access, 16(3), 387393. https://doi.org/10.1016/j.cmet.2012.08.002. Ebersberger, I., Metzler, D., Schwarz, C., Pbo, S., & Pa, S. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. American Journal of Human Genetics, 70(6), 14901497. https://doi.org/10.1086/340787 For, C., Pcr, M., & Technologies, L. (2012). 5IIIIIIIIIIIIIIIII: I, 1(61). Hixson, J. E., & Vernier, D. T. (1990). Restriction isotyping of human apolipoprotein E by gene amplification and cleavage with HhaI. Journal of Lipid Research, 31(3), 545548. Hwang,

S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific Reports, 5(December), 17875. https://doi.org/10.1038/srep17875 Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., & Shyr, Y. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics, 13 Suppl 8(Suppl 8), S8. https://doi.org/10.1186/1471-2164-13-S8-S8 Masterman, T., Zhang, Z., Hellgren, D., Salter, H., Anvret, M., Lilius, L., Hillert, J. (2002). APOE genotypes and disease severity in multiple sclerosis. Mult Scler, 8(2), 98103. https://doi.org/10.1191/1352458502ms787oa Mckeith, I. (2004). Early diagnosis of Alzheimers disease: update on combining genetic and brain-imaging measures, 333341. Pearson, J. V, Huentelman, M. J., Halperin, R. F., Tembe, W. D., Melquist, S., Homer, N., Craig, D. W. (2007). Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. American Journal of Human Genetics, 80(1), 12639. https://doi.org/10.1086/510686 Rohn, T. T. (2014). Is apolipoprotein E4 an important risk factor for vascular dementia? International Journal of Clinical and Experimental Pathology, 7(7), 35043511. https://doi.org/10.13188/2376-922X.1000004 Tchernitchko, D., Legendre, M., Delahaye, A., Cazeneuve, C., Niel, F., Goossens, M., Girodon, E. (2003). Clinical Evaluation of a Reverse Hybridization Assay for the Molecular Detection of Twelve MEFV Gene Mutations. Clinical Chemistry, 49(11), 19421945. https://doi.org/10.1373/clinchem.2003.021212 Zhong, L., Xie, Y., Cao, T., Wang, Z., Wang, T., Li, X., Chen, X. (2016). A rapid and cost-effective method for genotyping apolipoprotein E gene polymorphism. Molecular Neurodegeneration, 18. https://doi.org/10.1186/s13024-016-0069-4

Chapter 4

A Misassembly validation protocol for *denovo* hybrid assembly for microbial genomes

By

Priyanka Rawat

Table of Contents

List of Figures

Chapter 4

A Misassembly validation protocol in microbial *denovo* hybrid assembly

4.1 Abstract

Prokaryotic *de novo* genome assemblies specifically those which do not have finished reference genomes, are difficult to assemble owing to repetitive content, transposons, tandem-repeats and no estimation of the correct assembly size (Wetzel, Kingsford, & Pop, 2011). For such assemblies, conclusive result of most of the assemblers is a highly compressed assembly with collapsed repeats and incorrect copy numbers (Phillippy, Schatz, & Pop, 2008). Many de-novo genome assemblies do not proceed beyond draft level genomes. Thus, we have information which is not complete and accurate. Long-read technologies may provide resolution to repeats but vigorous approaches used in assembling might introduce mis-assemblies and traditional may introduce fragmentation (Kamath, Shomorony, Xia, Courtade, & Tse, 2016). Due to lack of standard procedures and methods to validate and evaluate assemblies, there might be multiple answers to one question with different sequencing technologies and assemblers. In this paper, we analyze two assemblies for a single genome of *de novo* sequenced *Kibdelosporangium* Actinobacteria (Ogasawara et al., 2015) with different technologies and approaches. While seeking accuracy and

completeness of the genome, work-flow for mis-assembly detection is developed and two assemblies are analyzed. It is shown that different technologies and assemblers give different results for the same genome. Thus, need for developing standards for validation of *de novo* sequenced genome is demonstrated.

4.2 Introduction

The genus *Kibdelosporangium* is one of the rarest actinobacterial genera which are source of medicinally useful bio actives (Manuscript & Nanostructures, 2008). This genera does not have extensively sequenced and complete representative genomes (Ogasawara et al., 2015). This makes it difficult to estimate the complete genome size, assembly contiguity and accuracy. Being the fifth largest Actinobacterial genome to be sequenced, makes it more challenging to sequence and assemble. The genome is extensively sequenced with short-read (Illumina), Ion-torrent and long-read (Pacbio RSII) technologies (Ogasawara et al., 2015) for the first draft genome. Briefly, Illumina paired-end data (2X 12 M reads,250 bp), Ion Torrent (3.6 M,200 bp), and PacBio data (~ 38X coverage) were assembled using MIRA and polished using in-house made custom script. This resulted in high-quality 11.75 Mbp (21 scaffold N-50,1.6 Mbp) draft genome assembly. This assembly lacked information about scaffold orientation, gaps and overlaps between contigs, and linkage information. With the development in sequencing chemistry and greater read-length output in sequencing technologies, another sequencing attempt was made with Pacbio RSII (Menlo Park, CA) sequencing platform (20 Kb insert size library were sequenced with 2 SMRT (single-molecule real-time) cells). It resulted in 275718 reads (N50, 8476 bp) which were assembled with Hierarchical Genome Assembly Process 2 (HGAP2) protocol from

SMRT Analysis version 2.0 package (Chin et al., 2013). The result was a complete linear (12,113,479bp ~117X coverage) single contig assembly gaining ~363bp of sequence from initial draft assembly. The goal of this chapter is to study as to why despite such extensive sequencing and higher depth of coverage these scaffolds could not be completed as a complete genome. These results leave few important questions unanswered-

1. Why, despite of extensive sequencing in first attempt, 21 scaffolds could not be completely assembled as a finished genome?
2. If, the Pacbio assembly, 12 Mbp, single contig, is completely accurate or just another estimation?
3. If this bacterial genome is circular or linear as depicted by pacbio sequencing?
4. If Pacbio assembly is right size and accurately assembled?

As, well said by Ian Korf at University of California, Davis – "When a species' genome is newly assembled, no one knows what's real, what's missing, and what's experimental artifact"(Baker, 2012). In this paper, we use two assemblies of *Kibdelosporangium* MJNF24 genome and investigate our curiosities based on questions asked above. Open-source software are used for detection of mis-assemblies. 12 Mbp Pacbio assembly is used as a reference to seek information about scaffold orientation, gaps and overlap analysis. Characterization of the two assemblies is done for repeats, transposons and basic characteristics. Validation is done by two approaches-reference-based mapping and coverage based feature analysis, this is done to avoid any biases and to simultaneously assess the two approaches. First 21 scaffolds are aligned to 12 Mbp assembly and analyzed

for gaps, overlaps and orientation of the contigs. Secondly, mate-pair validation is done by using Illumina short reads. Finally, the information from both approaches is combined and used.

4.3 Materials and Methods

4.3.1 Sequencing and Assembly

DNA was isolated from samples and sequencing was done at National center for genome resources (Santa Fe, New Mexico). 20 kilo base-pair library was prepared and sequenced with 2 SMRT cells. 275,718 reads were generated with 5127 mean length. The reads were error-corrected and *de-novo* assembled by HGAP2 protocol.

4.3.2 Assembly Characterization- Read-Quality statistics

All paired-end reads to be used for analysis were quality-trimmed by using BBmap 36.67. Reads were trimmed to remove adapters and quality score of 30 by using quality trimming option. Total of 32419784 paired-end reads were input, 29465910 reads (90.89%) were quality trimmed and kept. 2953874 reads (9.11%) were removed due to lower quality.

4.3.3 Assembly Characterization- Assembly Statistics

The precise behavior and estimation of mis-assemblies was hard to predict without prior knowledge of assembly statistics. Basic assembly statistics like GC-content, cumulative length, N50 and advanced statistics like NA50 (N50 of aligned bases to the reference) provided estimation for length and assembly composition on assembly and aligned parts of

the assembly. For reference based analysis, assembly metrics on aligned genome helped quantifying the best and worst alignments to narrow down the mis-assembly evaluation. Alignment to the reference assembly showed that 21 scaffolds were compressed and aggressively concatenated, which when broken into contigs resulted in increased alignment length mapping to the reference (combined1_21 broken). The regions of aggressive compression could be narrowed down for mis-assembly detection.

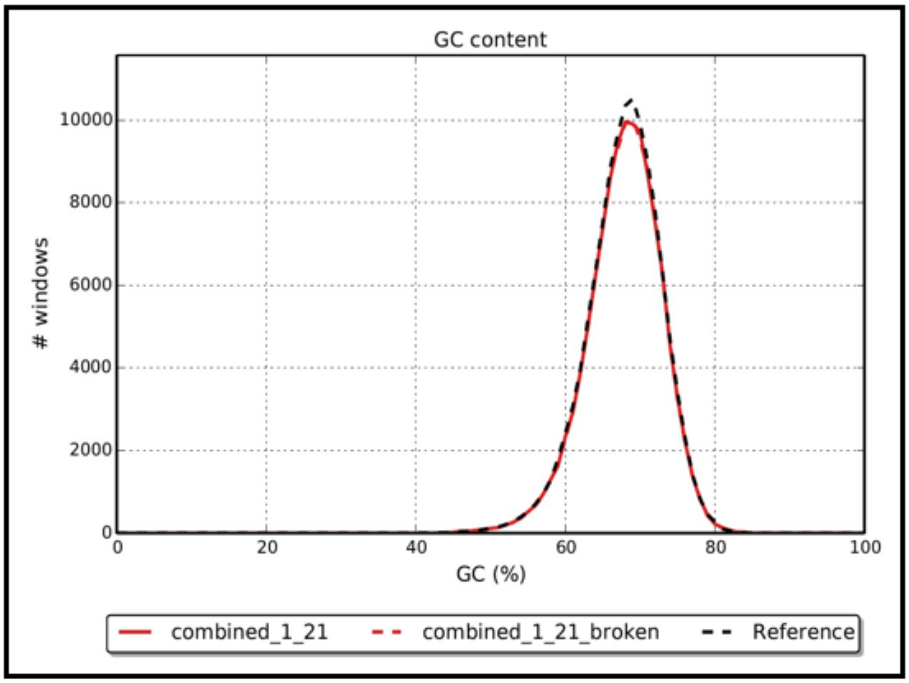


Figure 4.1: GC-content of Pacbio assembly as reference. Combined_1_21 is total 21 scaffolds. Combined 1_21_broken is

N50 was falsely used as assembly quality metrics, as it could be increased due to forced scaffolding and resulting in mis-assemblies. NA50 metrics introduced by QUAST (Gurevich, Saveliev, Vyahhi, & Tesler, 2013) could be used to correctly calculate N50 on aligned parts of reference rather than the whole genome. Figure 4.3 shows that value of

NGAx decreases as compared to NGX when computed on aligned reference only. Here, NGx is showing the percentage of total assembly covered per contig length but NGAx refers to the total percentage of reference assembly alignment covered by contig length. The decrease in total percentage shows that there are mis-assemblies present in the 21 scaffolds and there are forced alignments.

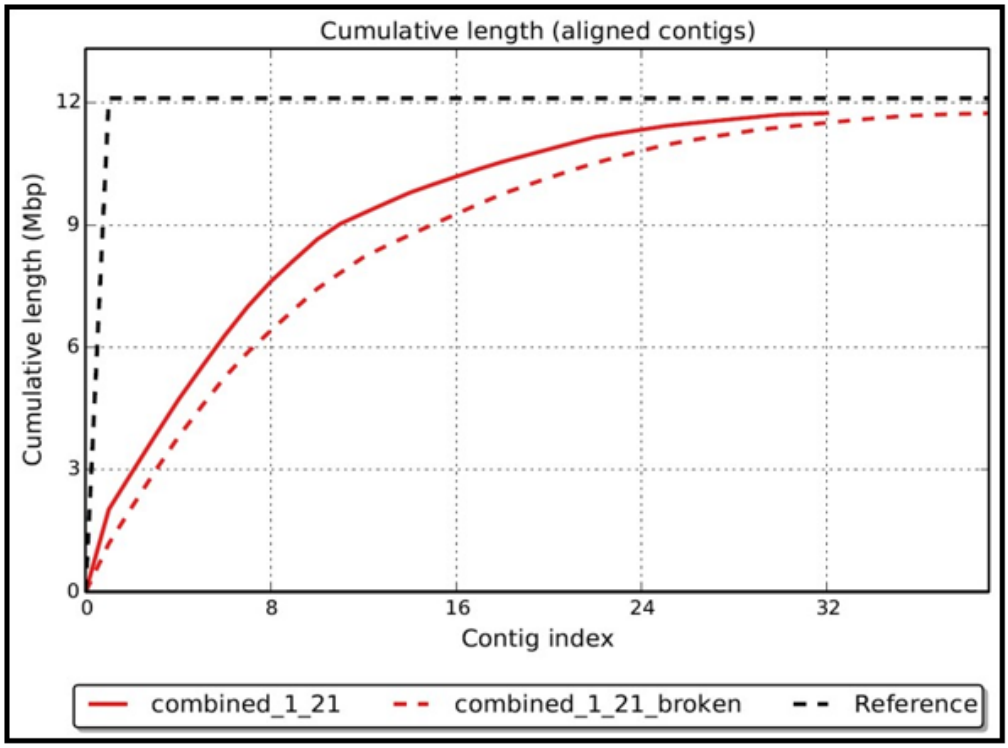


Figure 4.2: Length of aligned contigs when mapped against the reference.

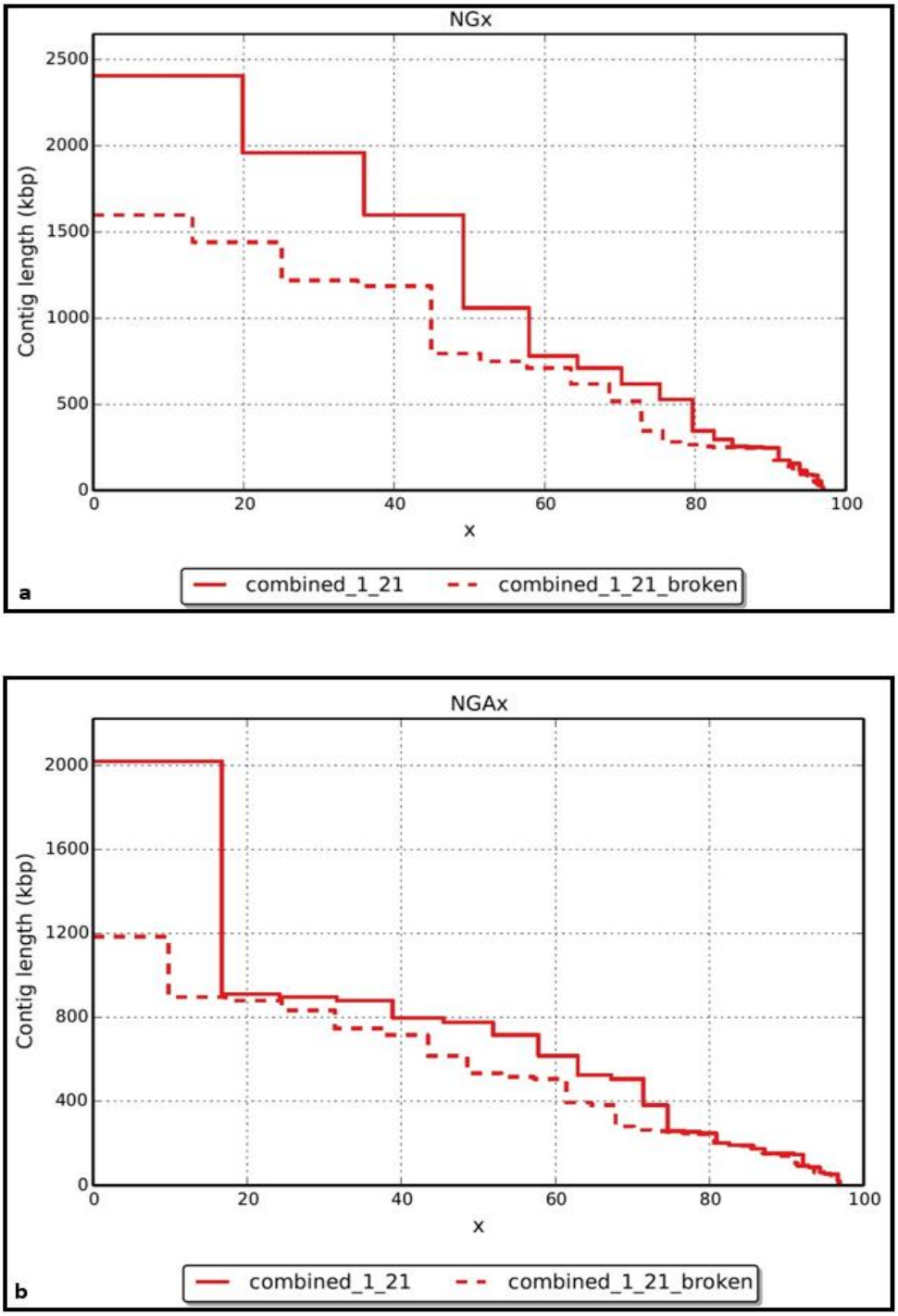


Figure 4.3: a) N50 calculated on total percentage of reference covered by alignment per contig length. b) NA50 calculated on total percentage of reference aligned per contig length.

4.3.4 Assembly Characterization - Mapping Statistics

Q30 trimmed paired-end reads were aligned by BWA version 0.7.10 to the 21 scaffolds and Pacbio assembly for analyzing coverage based features and per-base scores in an effort to analyze erroneous regions of the assemblies (Chawla, Kumar, & Shankar, 2016). Basic coverage statistics were calculated using FRCbam (Vezzi, Narzisi, & Mishra, 2012b). A sorted and indexed bam was provided to FRCbam for 21 contigs and pacbio assembly. As shown in figure 4.4, basic coverage statistics of mapped paired-ends in both assemblies were almost similar. Few paired-end reads mapped at different contigs in 21 scaffolds with both ends at different contigs, suggesting mis-assembled contigs in first assembly. In graph 2, total number of reads aligned to the two mappings are shown, as Pacbio assembly has higher number of paired-end reads mapped in wrong distance, which suggests some re-arrangement or relocation errors in the assembly which will be discussed further in section 4.3.6.2.

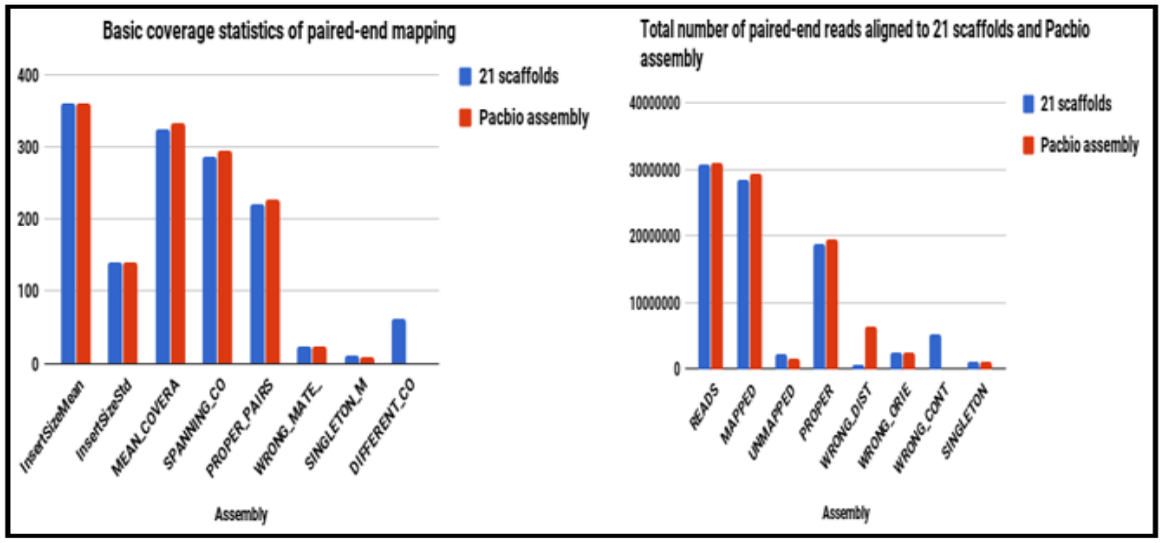
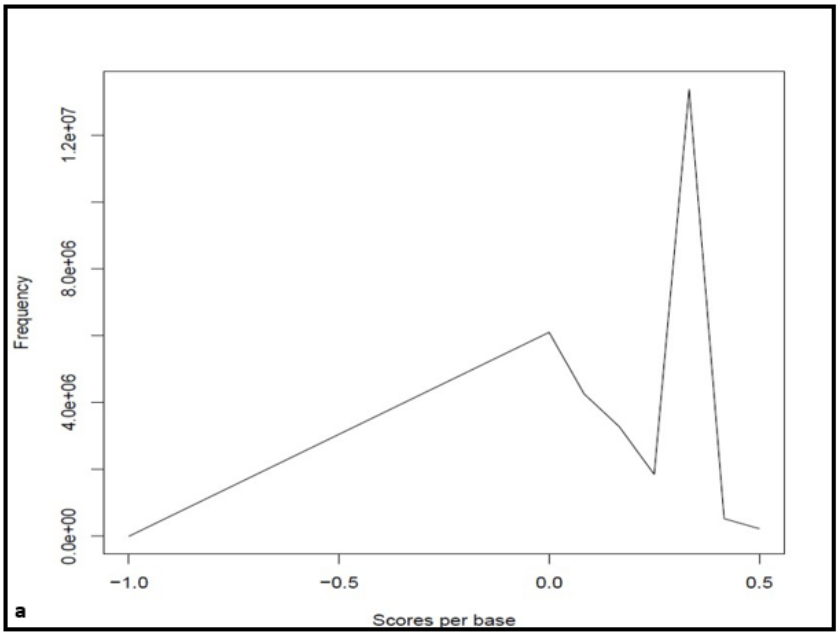


Figure 4.4: Basic coverage statistics of mapped paired-end reads to the two assemblies.

Per-base statistics were calculated by reapr (Hunt et al., 2013) to analyze the accuracy at per-base level by only considering perfectly and uniquely mapped reads with correct orientation and correct distance. Reapr version was used first to align the Q30 trimmed paired-end reads with SMALT (Ponsting & Ning, 2010). Reapr broke 21 scaffolds to 28 contigs and analysis was done on the 28 contigs. Based on per base- read-coverage, fragment-depth and size on inner and outer fragments and FCD-error (fragment coverage depth errors), scores per-base were calculated for a quantitative comparison of the two assemblies. Lower the scores, lower the quality of the base.



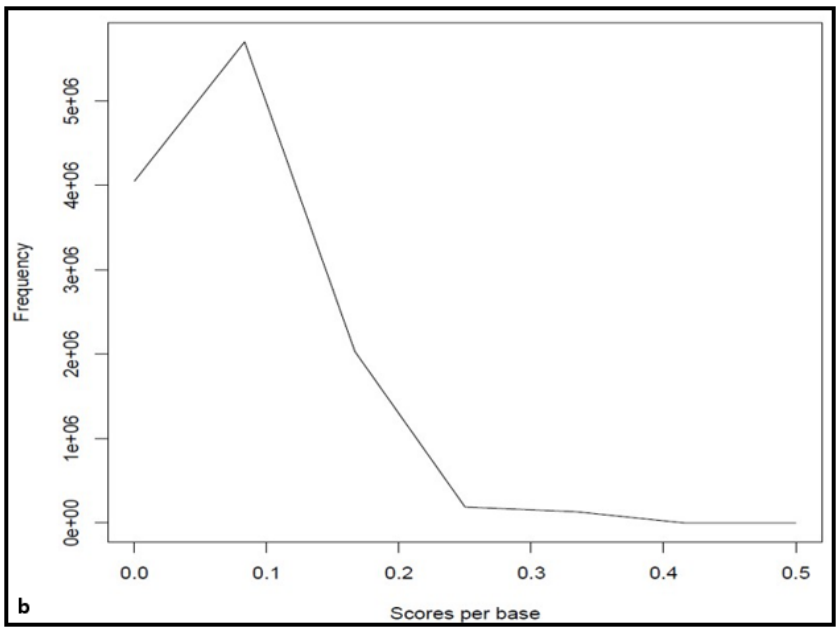


Figure 4.5: Scores per base calculated by REAPR. a) Score per base for 28 contigs from Kibdelosporangium. b) Scores per base for pacbio assembly.

4.3.5 Assembly Characterization - Repeat-Analysis

Most of the mis-assemblies in *de novo* genomes are due to inability of assemblers to correctly position the repeats and copy numbers (Lin & Liao, 2015). Longer repetitive transposons, tandem-repeats and insertion sequences can lead to in-correct orientation, placement or wrong copy number estimation in the assembly (Lupski & Weinstock, 1992). This could lead to an overly compressed alignment, and mis-assemblies, when multiple copies of a repeat are collapsed to one location. Repeat analysis is done by searching for repetitive patterns (Figure 4.6) by Tandem repeat database (Koren, Treangen, Hill, Pop, & Phillippy, 2014) in both assemblies.

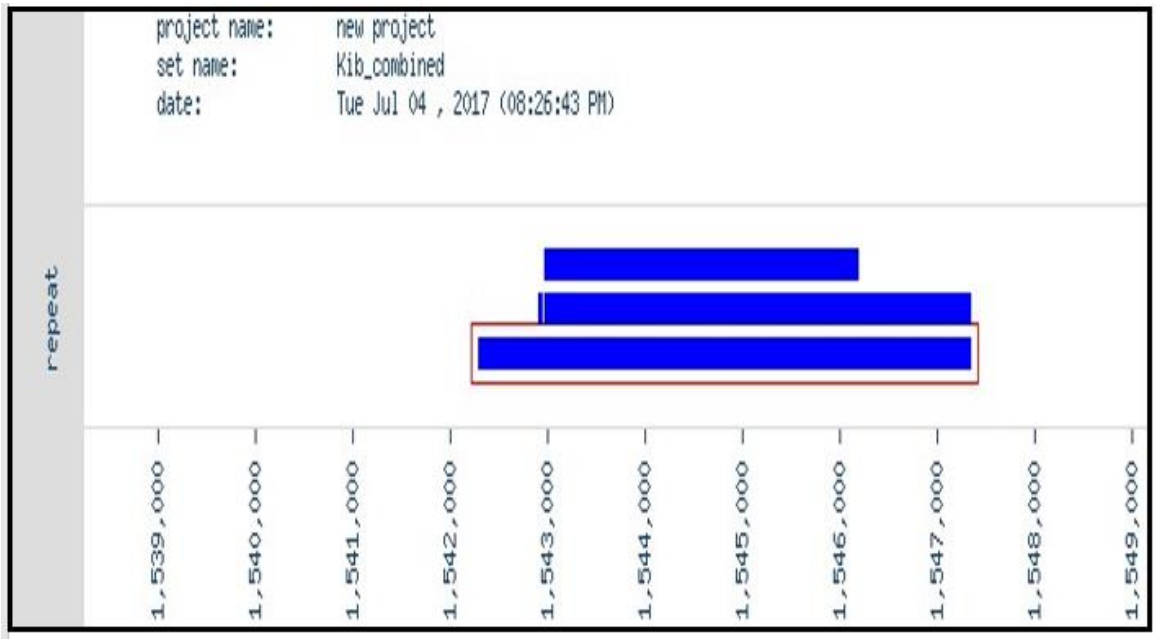


Figure 4.6: 1936bp repeat in Contig 18 aligned with another 700 and 400 bp repeats.

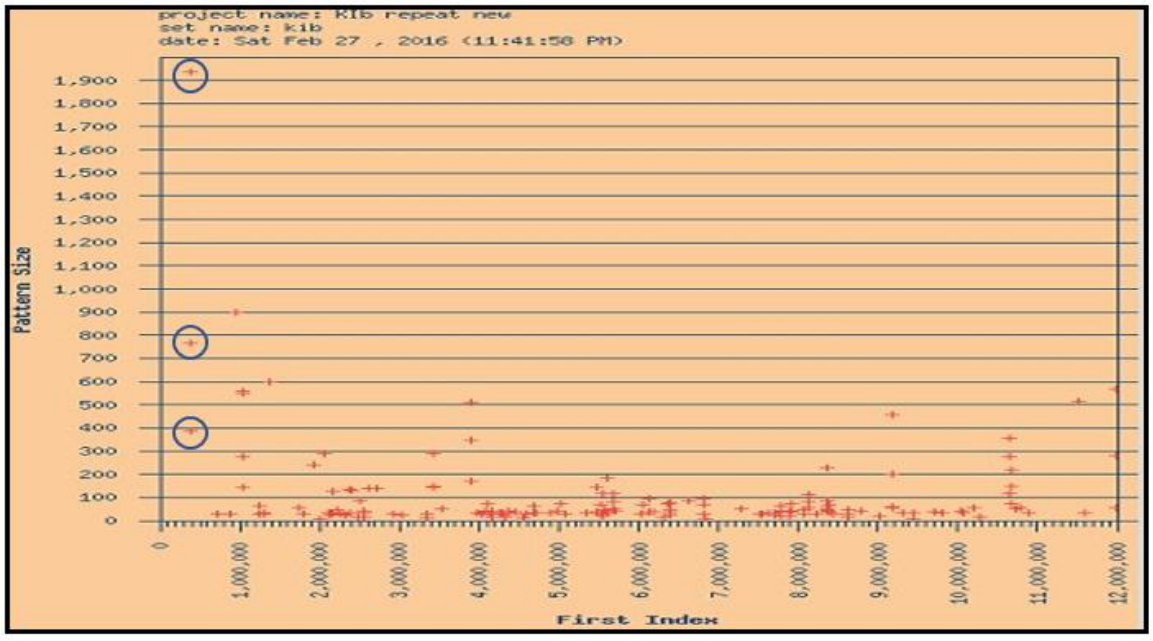


Figure 4.7: Copy-number of various repeats in 12 Mbp Pacbio assembly. 1936bp repeat in contig 18 Figure 4.6 is present in the starting of the assembly along with two other 700bp and 400bp repeat copies aligning to it.

Pacbio has 2.6 copies of 1937 mer totaling 5782bp, 4.2 copies of 765 mer totaling 2408bp and 11.3 copies of 385mer 1948bp. These small repeats might be the flanking sequences spanning terminal inverted repeats which were difficult to assemble in first draft and thus, contig 18 was mis-assembled (Figure 4.6).

Transposons and Insertion sequences analysis

Repetitive transposons like elements posed problems in assembling the scaffolds in first assembly. Analyzing such elements can be helpful in analyzing the orientation of the repeats combined with paired-end data. Insertion sequences flank transposon elements, analysis for such sequences is done by IS-saga (Varani, Siguier, Gourbeyre, Charneau, & Chandler, 2011). MITE Digger(Lin & Liao, 2015) is used to search for miniature inverted repeats transposable elements which pose problems while positioning of the repeats in the assembly. These are present in high numbers in the genome and can cause mis-placement of the real repeats. This program found 6 True MITE sequences in the pacbio assembly of total 1024bp sequence with total 14 copies in the genome (Figure A1.7). Further analysis with MITE sequences will be done in section 4.5

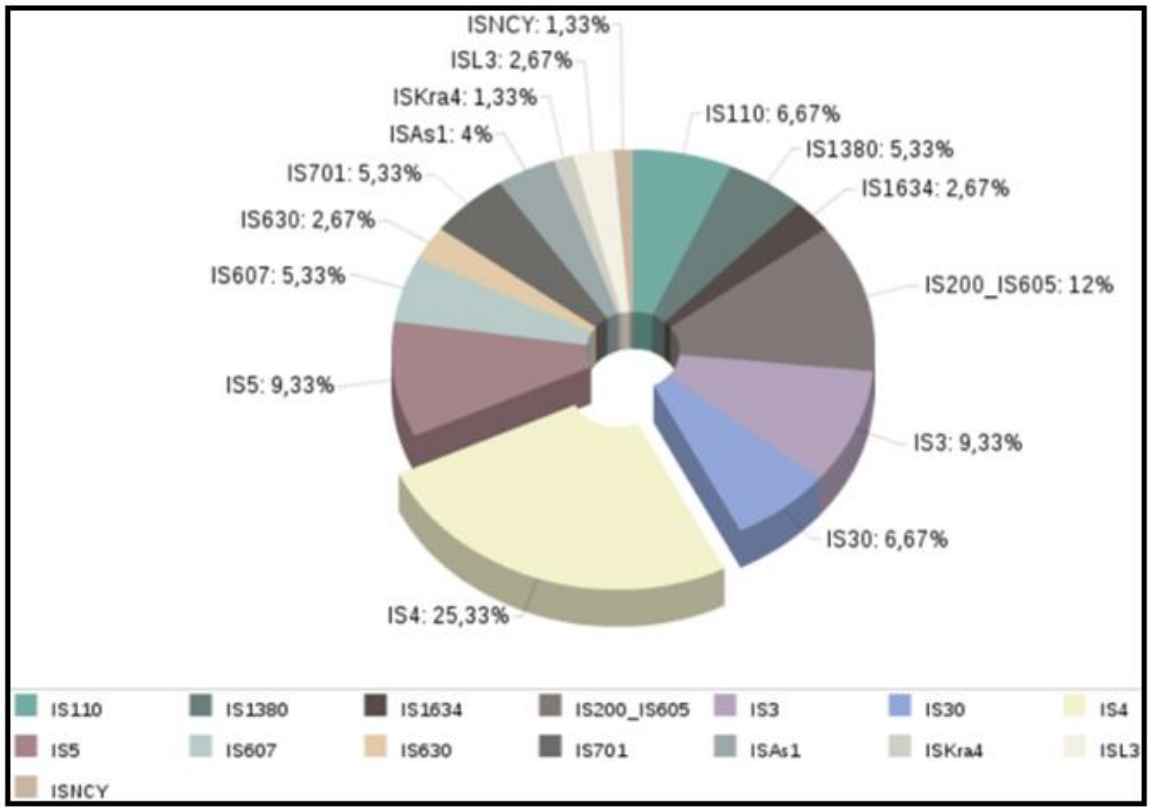


Figure 4.8: Insertion sequences in Kibdelosporangium Pacbio assembly.

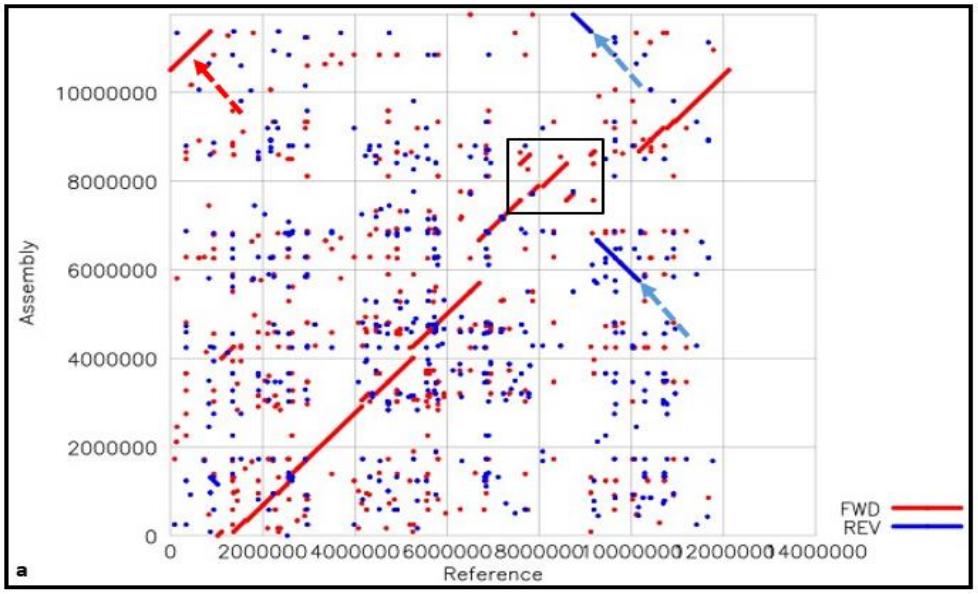
4.3.6 Assembly Evaluation

In this paper, the goal is to first figure out major genomic re-arrangements, relocations, gaps and overlaps in 21 scaffolds based on reference (Pacbio assembly) analysis. QUAST and misFinder are used for this analysis. Mapping statistics results show that Pacbio assembly and contigs have similar mapping patterns with each having its own set of errors. Since, we are not sure of the accuracy of Pacbio assembly either, non-reference based analysis is also done by FRCbam and REAPR on both 21 scaffolds and Pacbio assembly. Later these genomes based re-arrangements are queried for causes which led to these errors in the assembly. Paired-end data is used to compute compression-expansion based statistics

(FRCbam) and per-base statistics (REAPR) to zero on the real error regions. These regions are further probed for presence of any transposons, repetitive or hard to sequence areas to finalize what made the assemblers to fail at these regions and to produce errors.

4.3.6.1 Reference-based: Alignment based

21 scaffolds were mapped to Pacbio assembly using CLC version 8.0. (Figure A1.2). Quast is an assembly quality assessment tool which just does not rely on N50 metrics but introduces new metrics such as NA50, and more metrics based on the mapping with the reference (Varani et al., 2011). 21 contigs were aligned to the pacbio assembly and analyzed for mis-assemblies. NA50 is calculated by QUAST for 21 scaffolds and broken 21 scaffolds on basis of alignment. Fig 4.3 shows that N50 metrics is higher than it should be due to aggressive concatenation of the contigs. Mummer alignment dot plots further confirm this. Alignment plot shows that the there are some inversions and collapsed regions. 21 scaffolds are broken to 28 contigs based on the NA50 metrics calculated after alignment as discussed in section 4.3.3. Unaligned portions are removed, new blocks are created and N50 is calculated. 5 contigs align partially with 1000bp to 2300bp of region not aligned to the assembly.



Kib 21: Dot-plot between Pacbio assembly and 21 scaffolds.

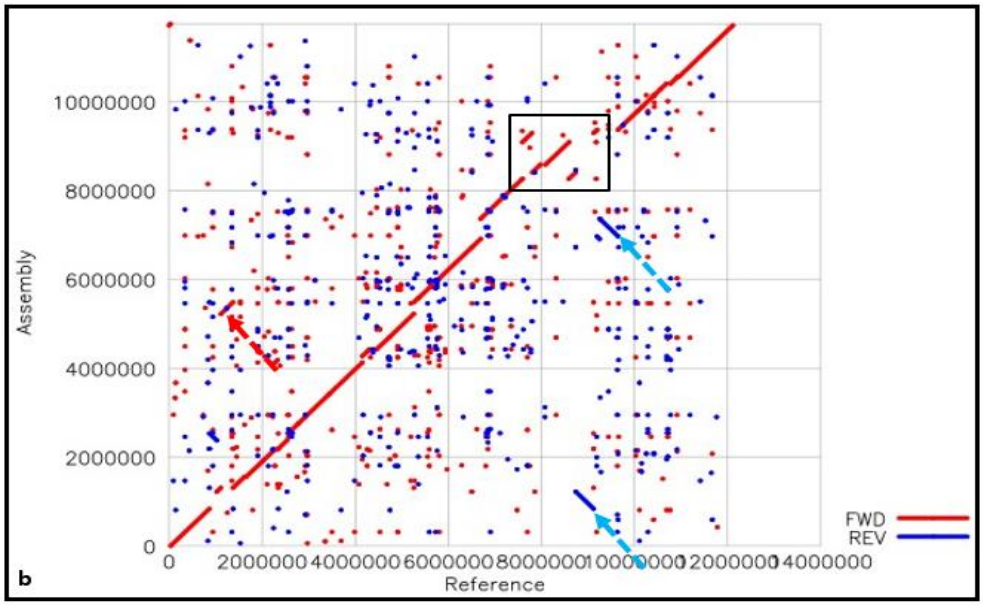
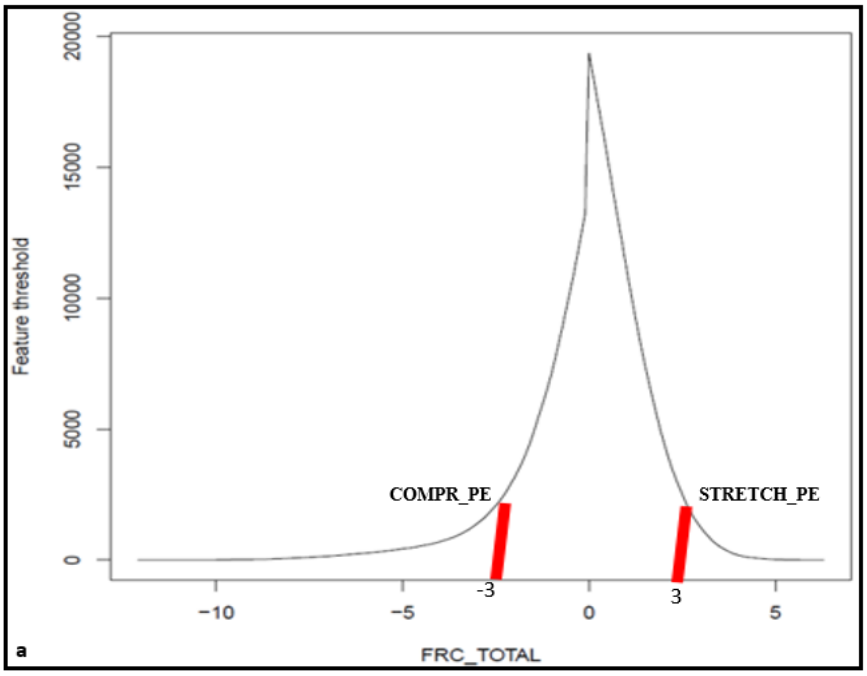


Figure 4.9: shows the MUMmer alignment dot plots (Varani et al., 2011) of 21 scaffolds to the reference before calculating misassembly (a) breakpoints and after breaking the assembly (b) into 28 contigs.

Due to forced concatenation of scaffolds, there are inversions (blue arrows), and translocations (red-arrows). Also, there are deletions in the assembly. The blue arrows in the second plot show that there are inversions still there after scaffolds are broken into the contigs, but smaller in size and one additional. This might explain incorrect positioning of repeats while scaffolding. Black box shows that there is overlap at a position on both sides of the contig suggesting tandem repeats. Such patterns are usually seen in circular contigs. Thus, we suspect that this chromosome is circular not linear as Pacbio data suggests. This will be discussed more in detail in section 4.5.

4.3.6.2 Without Reference: Coverage based

After characterizing the assemblies for genome-level re-arrangements, reference free approaches were done to rule-out any errors due to bias of reference. For this approach, paired-end reads were aligned against both assemblies. It has been shown that for high coverage prokaryotic genomes, mate-pair analysis effectively validates the assembly contiguity (Wetzel et al., 2011). Fragment-size based features such as Compression-Expansion statistics (CE-Statistics) of aligned mate-pairs(Vezzi et al., 2012b) was calculated (Figure 4.10). Mapping based features such as mapped, un-mapped, properly-paired, correct orientation, stretched-pairs (on all aligned reads) as shown in Figure 4.4 provide estimation of error prone area in the assembly.



Kib 21 scaffolds

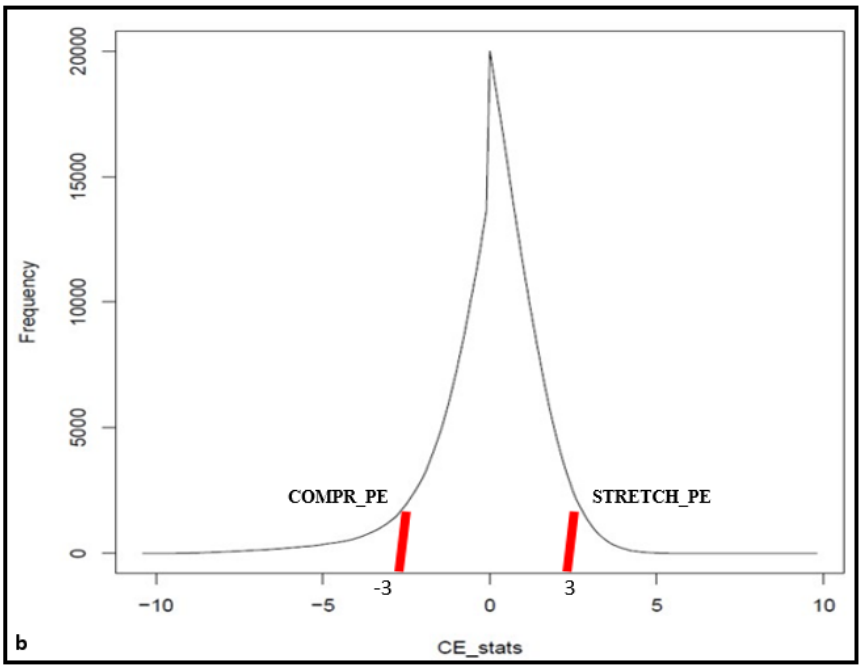


Fig 4.10: a) CE-statistics of Kib 21 scaffolds and (b) Pacbio assembly. The comparison shows that 21 scaffolds assembly has more paired-end reads which are compressed and

less stretched pairs. Pacbio assembly has equal numbers of compressed and stretched pairs thus suggesting more genomic relocations errors.

FCD-error cut-off is calculated per base based on perfectly and uniquely mapped reads (Hunt et al., 2013) and two assemblies are compared for total contiguity or accuracy. Read-depth, orientation, type of paired-mapping and fragment depth are metrics used to score the bases with 0 being the worst and 1 being the error-free demonstrating local accuracy and regions of bases that are highly accurate.

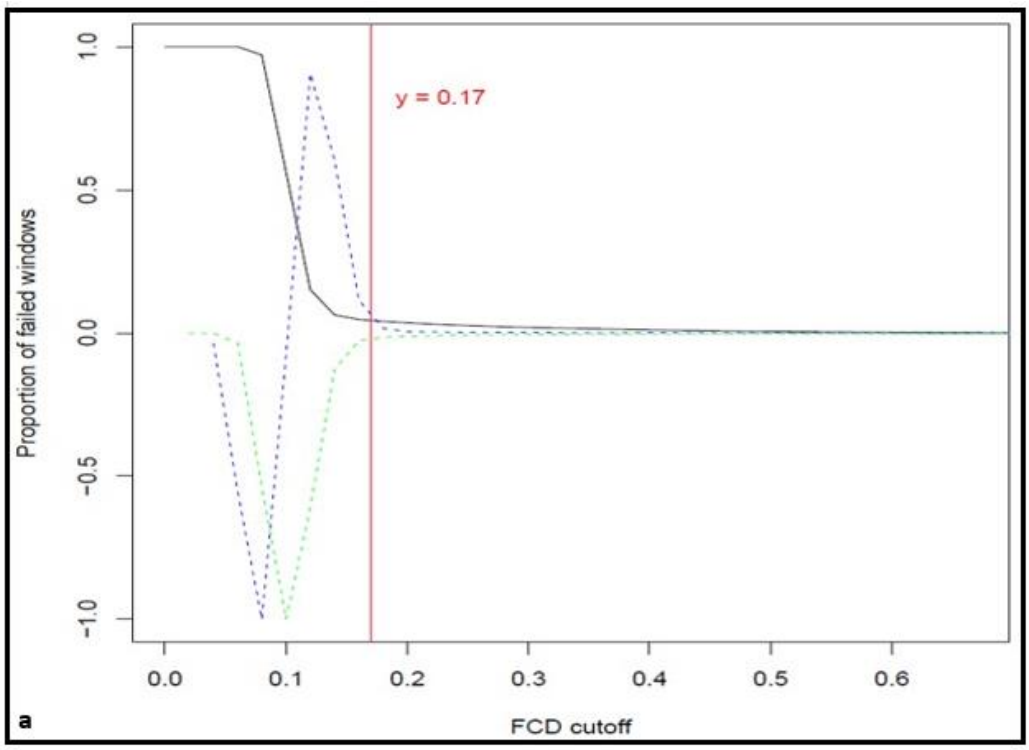


Figure 4.11: a) FCD-error cut-off for Pacbio assembly. The black line is the theoretical cut-off calculated. Blue and Green line are first and second derivatives of the black line

normalised to fall between -1 and 1(Hunt et al., 2013). Red line shows the FCD-cut off for Pacbio assembly.

As it is obvious from the Figure 4.11 a and b that Kib has 0.17 cut-off and Pacbio has 0.19. This suggests high error-rates in the assemblies and lower accuracy, with Pacbio assembly being slightly better. More detailed discussion about types of errors and mis-assemblies is discussed in results in section 4.4.

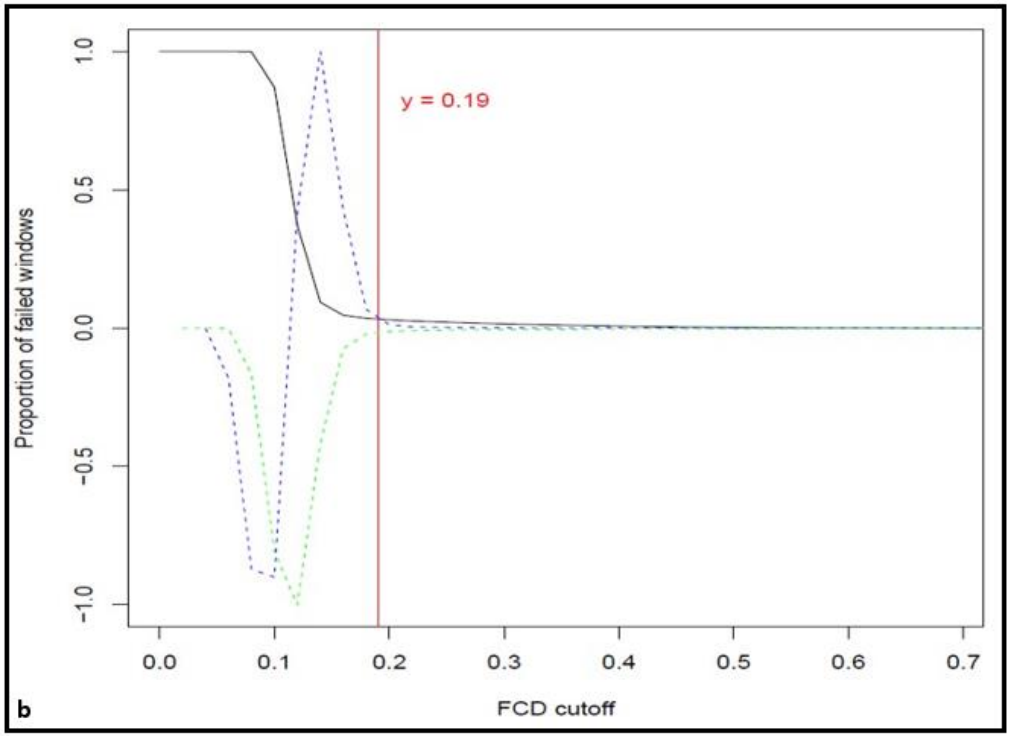


Figure 4.11: b) FCD-error cut-off for 21 scaffolds. The black line is the theoretical cut-off calculated. Blue and Green line are first and second derivatives of the black line normalised to fall between -1 and 1(Hunt et al., 2013). Red line shows the FCD-cut off for Kib.

4.3.6.3 Cumulative approach: Align and validate errors with paired-reads

MisFinder (Varani et al., 2011) is used for this purpose to simultaneously detect errors based on alignment with reference and validate with paired-end read data. This approach is used to distinguish between possibly false errors by first two approaches. MisFinder also tries to distinguish between structural variation differences and true errors. MisFinder finds 23 errors out of which 9 errors are in gap regions and outputs new assembly in 33 contigs.

4.4 Results

Based on reference-mapping and coverage based analysis results will be discussed in the following manner. First, the gaps, overlaps and genomic-rearrangements are discussed in respect to alignment. Then, state of paired-ends mapping is analyzed at these errors to interpret the reasons for the erroneous regions. Lastly, if these errors are caused due to insertion sequences, repeats or transposons, this is discussed. This manner is adopted to lay-out the results of analysis in the simplest and contiguous way.

4.4.1 Gaps, Overlaps and Genomic Re-arrangements

On aligning the 21 scaffolds to pacbio assembly, there are gaps due to missing regions in the scaffolds but presence in pacbio. When the nucleotide sequences in the gaps and overlaps are blast against the assembly, multiple copies of repeats interspersed in the genome in as small as 27bp are seen. Such multi-reads flank large sequences as seen in

Figure 4.12, repeat induced gaps and overlaps are caused due to the inability of the assemblers to correctly place these repeat copies. This genome has short-interspersed repeats spanning throughout the assembly which either cluster at the boundaries of two sequences and assemblers leaves the unique sequence creating gaps, or multiple copies of multi-reads at the ends of large nucleotide sequences causing them to overlap.

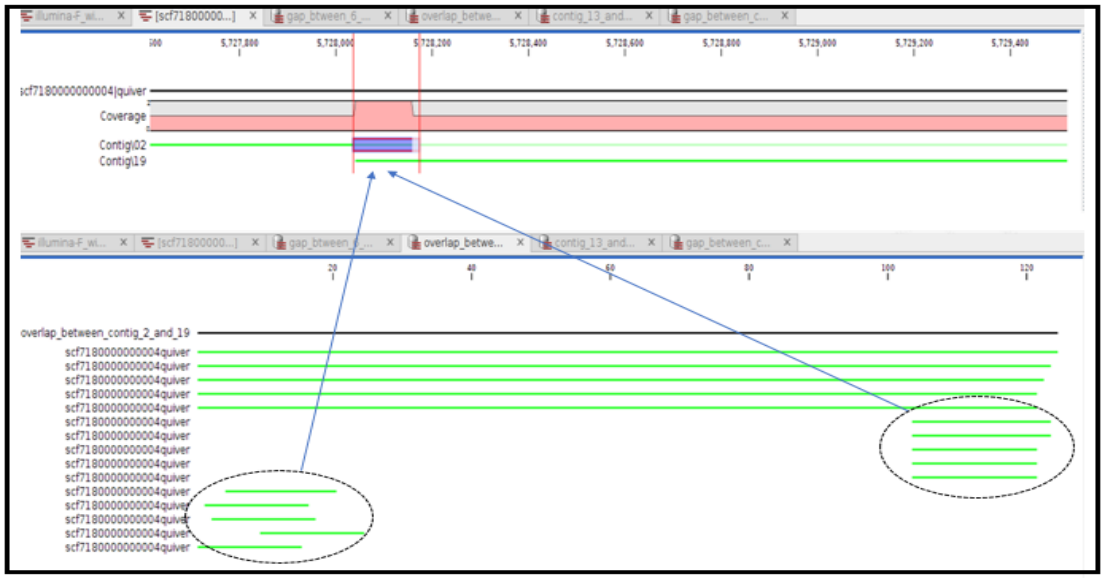


Figure 4.12: Repeat induced overlap in Contig 2 and Contig 19. On blast with reference genome there are multiple copies of the repeat and associated multi-reads in Pacbio.

MisFinder reports 9 mis-assemblies or misjoins in the gap regions. MisFinder reports 4 errors in Contig1 (Figure 4.13). One of them is a gap insertion at 264029bp-264207bp coordinates in Contig1.There is an overlap of 864bp in contig1 at this position with coordinates 263166bp-264016p. By analyzing with mate-pairs, it can be seen in Figure 4.14, that most of the mates are truncated and clipped near the gap region. The overlap suggests that there are two copies of this region oriented in wrong direction, thus,

collapsing of the paired-end reads. When this 864bp sequence is blast against the ISFinder database, it matches with IS4 sequences. ISQuest annotation of Kibdelosporangium shows that there is orf06931 near region 4724677bp and 4725933bp which contains region 4724529bp and 4725388bp (area on pacbio contig overlapping the error region). Thus, it can be concluded that this mis-assembly has been due to repetitive nature of terminal inverted repeats associated with these insertion sequences of IS4 family. Pacbio contig (Figure A1.4) also shows same mate-pair pattern at the region which aligns to Contig1 gap region suggesting that even pacbio assembly could not correctly place these inverted repeats and have collapsed repeat assembly as predicted by REAPR at region 4724385bp to 4724538bp. This suggests the aggressive nature of the long-read technology assemblers and concatenation of reads in collapsed repeats.

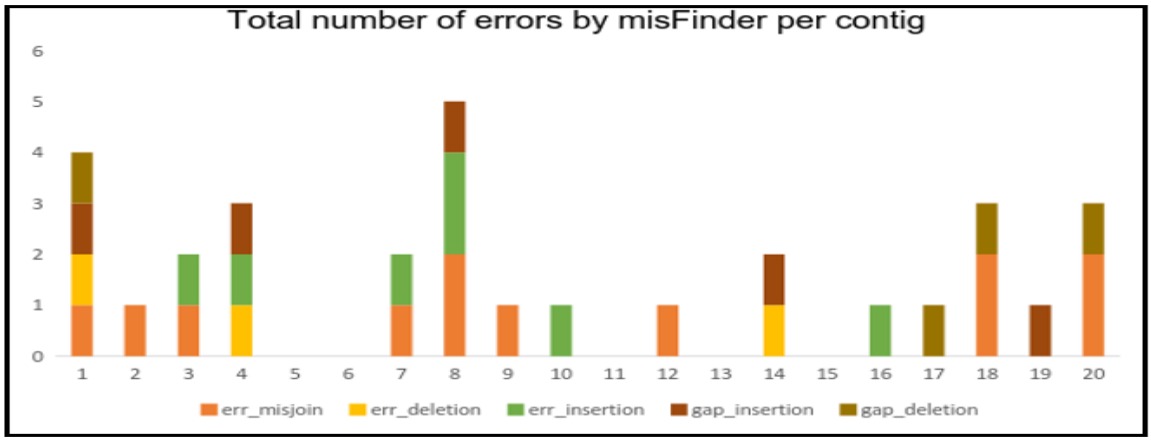


Figure 4.13: Error regions results by misFinder. As can be seen Contig1 has 4 errors out of 5 miskinds

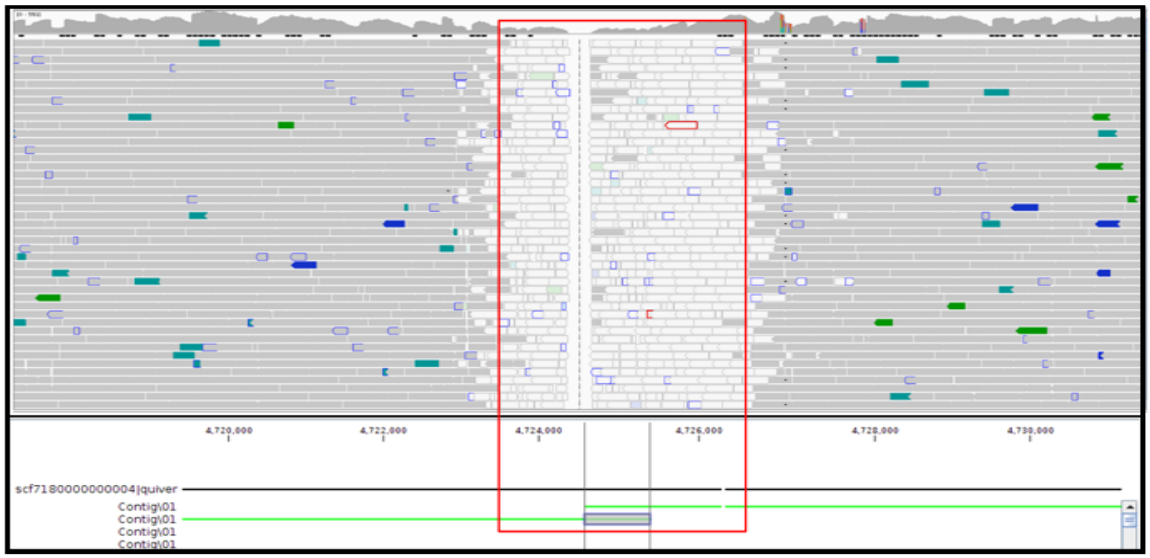


Figure 4.14: Insertion gap error in Contig 1: As there is gap due to truncation of paired-end reads. The pairs collapse on each side of the gap resulting into high-coverage near the gap region. It can also be seen that there are clipped pairs and disagreements.

QUAST reports this mis-assemblies as a scaffold-gap size mis-assembly interpreting wrong estimation of the copy numbers of the repeats at the right side of the contig. It also reports relocation mis-assembly in which Contig1 sequence from 798366bp to 1058046bp aligns 113176bp to 1372856bp of the pacbio contig.

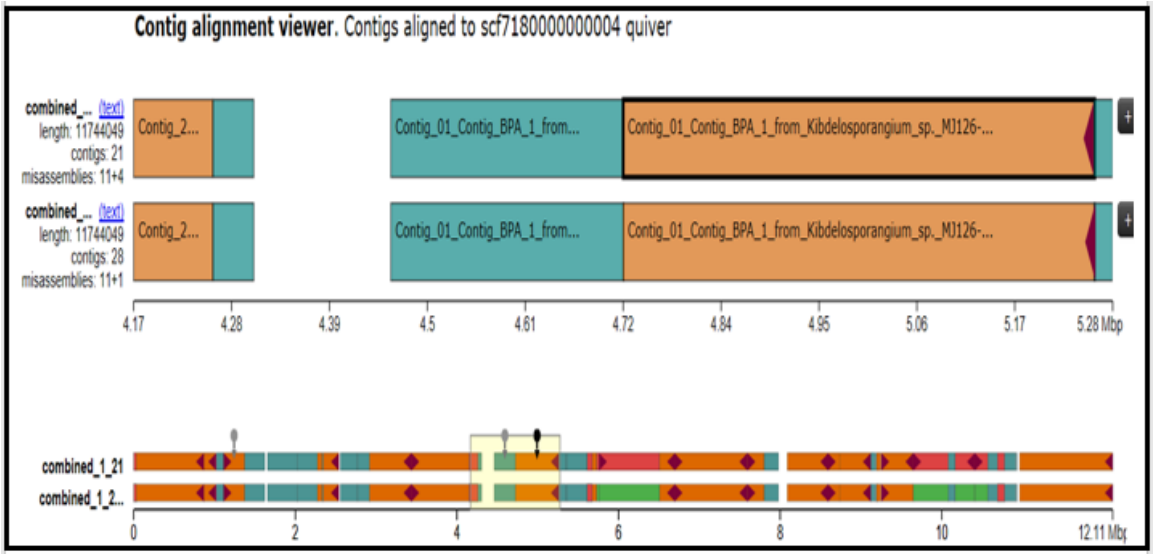


Figure 4.15 Icarus view of QUAST output. Mis-assembled Contig1 at the right side with mis-assembly shown in red triangles. This shows the scaffold-gap size mis-assembly 21 scaffolds are broken into 28 contigs and represented in order of alignment to the reference.

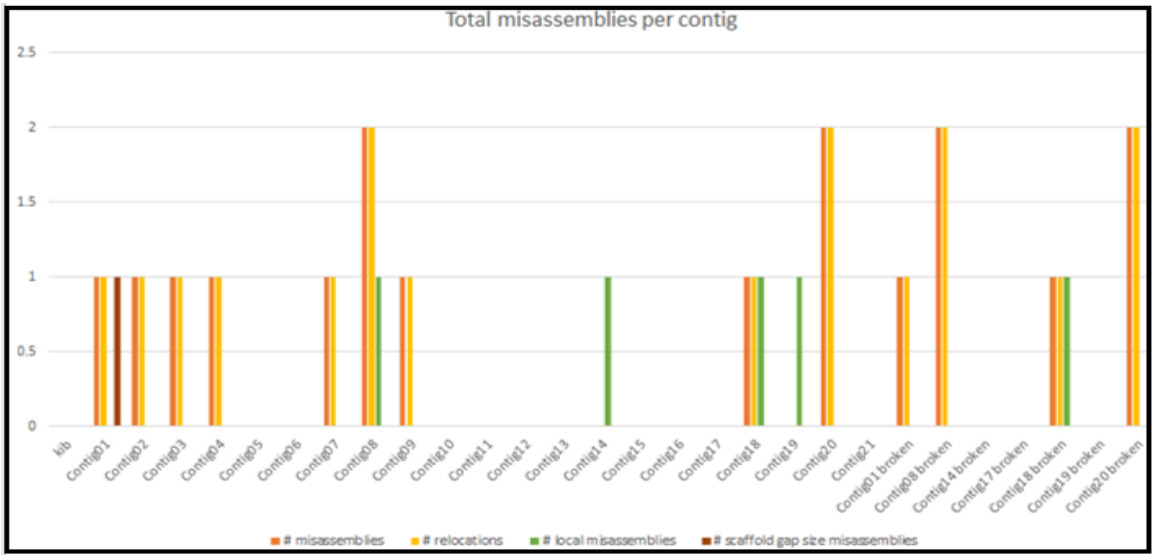


Figure 4.16: Total mis-assemblies as a result of QUAST analysis with 21 scaffolds aligned to reference Pacbio assembly.

Relocation mis-assemblies are the most prominent in this genome which are due to the overlapping of left and right flanking regions or are apart by more than 1kbp (Figure 4.16). Contigs 8,14,18 and Contig-18-broken are mis-assembled either due to two or more alignments covering the breakpoint, or gap between the left and right flanking regions is less than 1kbp or left and right flanking regions are on the same chromosomes. misFinder categorizes these errors to insertions, deletions and misjoin errors. Most of the misjoin errors are reported due to incorrect joining of segments as shown by mate-pairs aligning to positive and negative strands on the same or different contigs.

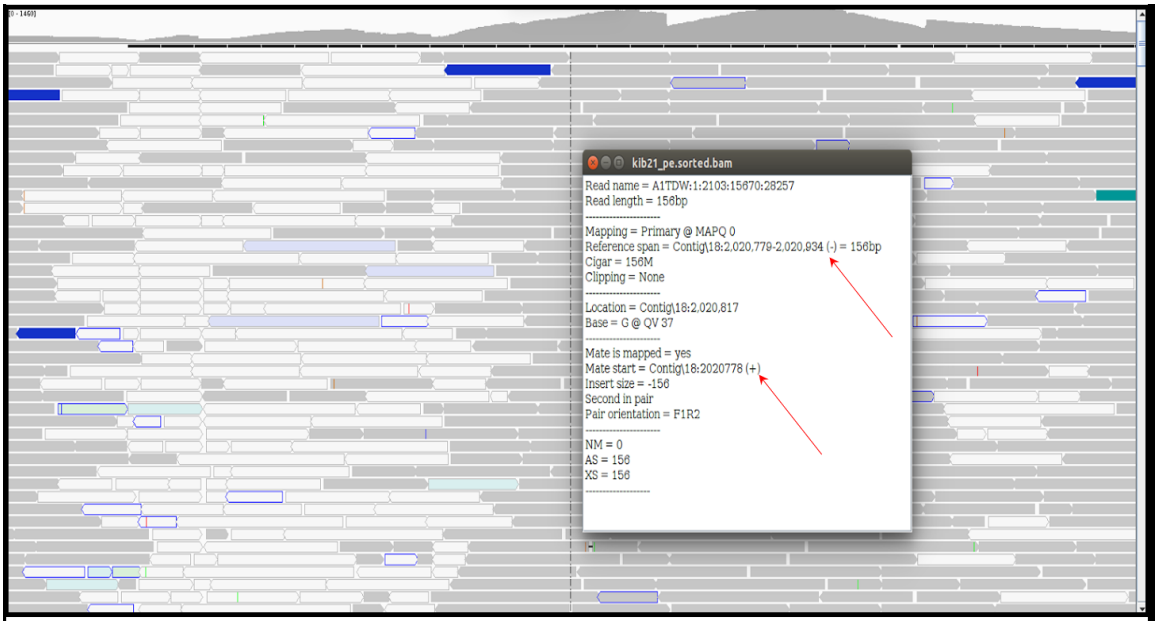


Figure 4.17: Misjoin error by misFinder in Contig18. Paired-end mates aligning to positive and negative strand. Change this figure from presentation and add arrowsat negative and positive signs in the box

Several sequence coverage gaps are found in the assembly. These might be due to the large segmental duplications with higher similarity, or terminally inverted repeats which are positioned in wrong directions (Varani et al., 2011). OLC based graphs could aggressively assemble different reads with same repeat boundaries in either wrong-orientation, collapsed or fragmented.

4.4.2 Coverage based features

An accurate assembly comprises of higher number of accurate bases. Mate-pair analyses is done regarding read-depth, read-coverage, orientation and fragment coverage provides an estimate of how correctly base is positioned and thus, how correctly the genome is assembled. As discussed in section 4.3.6.2, FRCbam and REAPR calculated mis-assemblies based on features mentioned above. Reapr uses paired-end reads to analyze the basic assembly statistics like read-coverage, read-orientation and coverage depth to score each base and to pinpoint the mis-assemblies. Reapr version 1.0.18 is used for the computing errors. FRCbam computes the features based on the observed distances between paired-end reads and local read coverage. Quality of assembly is estimated by calculating compression/expansion statistics of the paired-end reads. Q30 Illumina reads are mapped to 21 contigs and Pacbio contig. Coverage based features along-with CE-statistics are used to categorize assembly corresponding to erroneous regions, assign them a feature and plot this behavior for whole assembly. FRC (feature-response curve) is calculated for both assemblies which provides an estimation of accuracy and contiguity of the assemblies.

Features (errors) are computed per 1kbp sliding window based on compression and expansion of the paired-ends when mapped to the assembly. CE plot statistics for reference and kib21 are plotted Types of features determine the contiguity of the assembly. Figure 4.10 shows the graphs for CE statistics distribution only on paired-end reads aligned for reference and Query assembly. As we can see in Figure 4.18 that contigs have higher number of compressed paired-ends. Contigs1, 4, 8 and 18 have highest number of errors. Compressed paired-ends are a result of forced assembly where a region is forcefully joined and is a possible repeat-collapse. Missing repeat copies could cause higher read coverage locally. As shown in Appendix Figure A1.5.

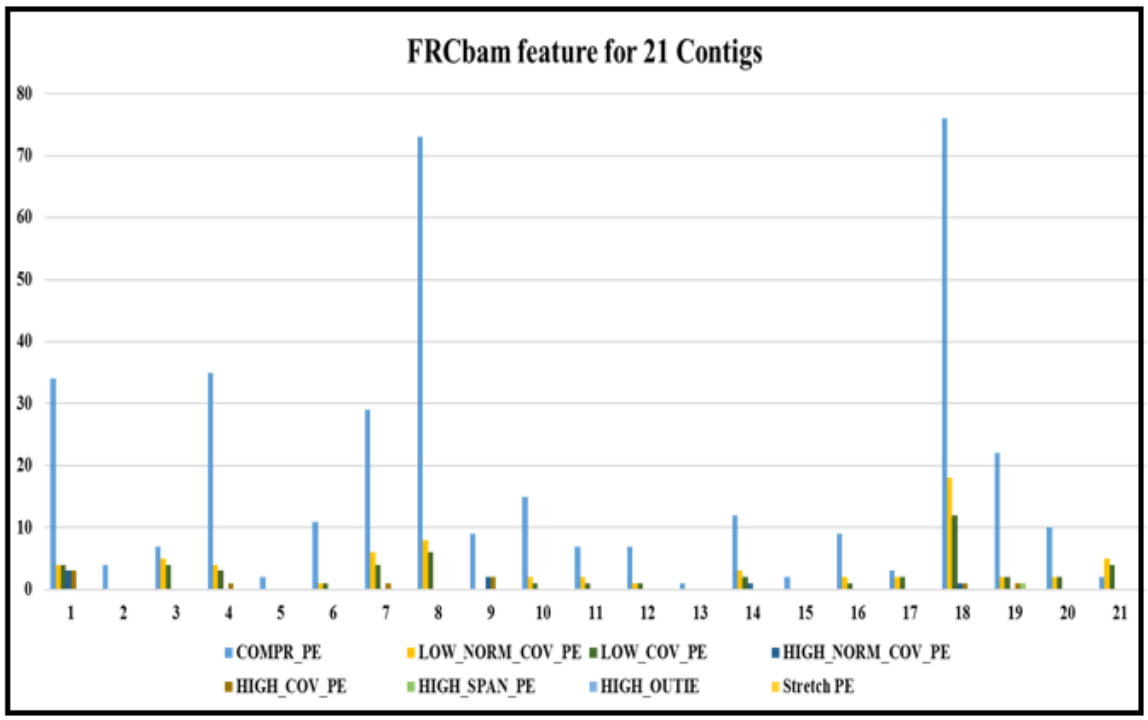


Figure 4.18: Total coverage based features in 21 scaffolds.

Most of the contigs have compressed paired ends, contig18 having the highest numbers. This feature indicates sequences assembled in wrong copy-numbers due to high number of

repeats. Figure 4.18 shows compression paired-ends in Contig1. Other most common mis-assembly is LOW_NORM_COVG_PE and LOW_COV_PE which are due to the few paired-end aligning at these regions, thus indicating insertions or inversions.

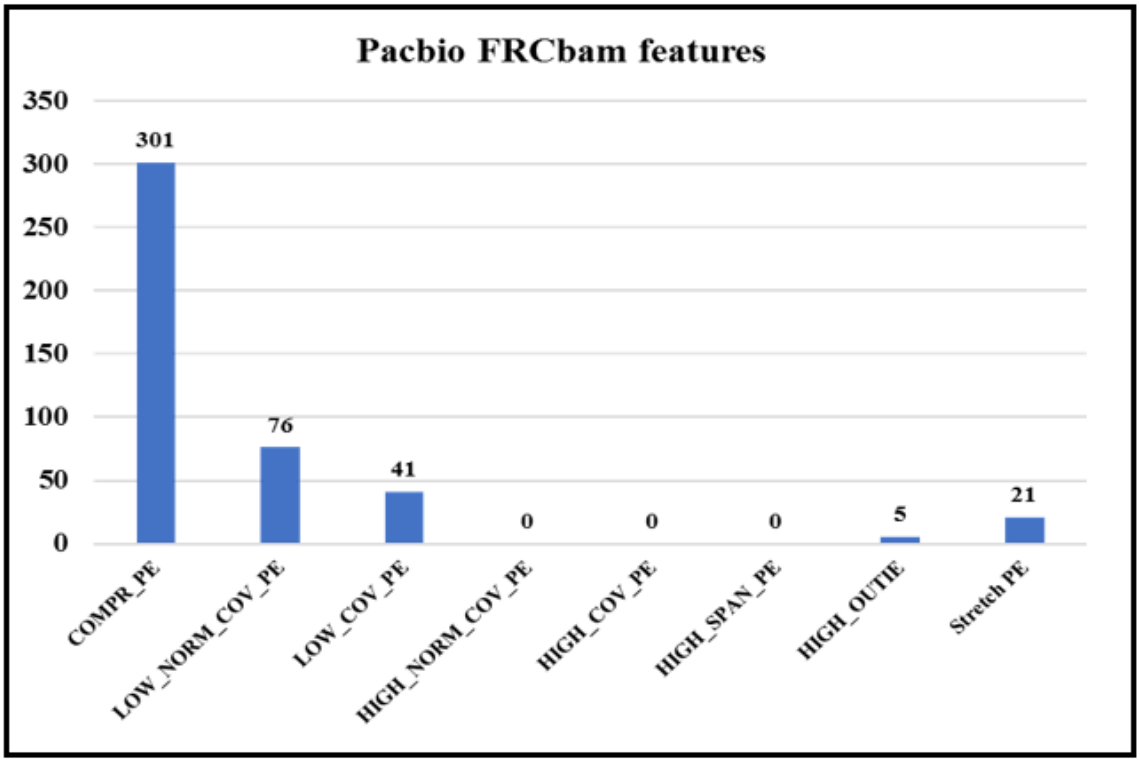


Figure 4.19: Total number of features in Pacbio assembly.

Although, Pacbio is used in this study as reference assembly, it does have its own share of mis-assemblies as shown in Figure 4.19. High number of compressed-paired end reads show that this assembly has highly mis-assembled repetitive regions which is unexpected in long-read technologies. On manually investigating this, it was found that there are highly repetitive terminally inverted repeats and there are either repeat induced gaps or overlaps. Reapr supports FRCbam findings although, reapr has less number of features (Figure 4.21) as each base is scored on multiple metrics. HIGH_OUTIE feature in Pacbio assembly

suggests that there is inversion in the pacbio assembly. To confirm this finding we made a dot-plot of pacbio against itself and it confirmed multiple inversion events (Figure 4.20) as output by FRCbam (Appendix1 Table A1.1) at position 5757401bp and 5758599bp.

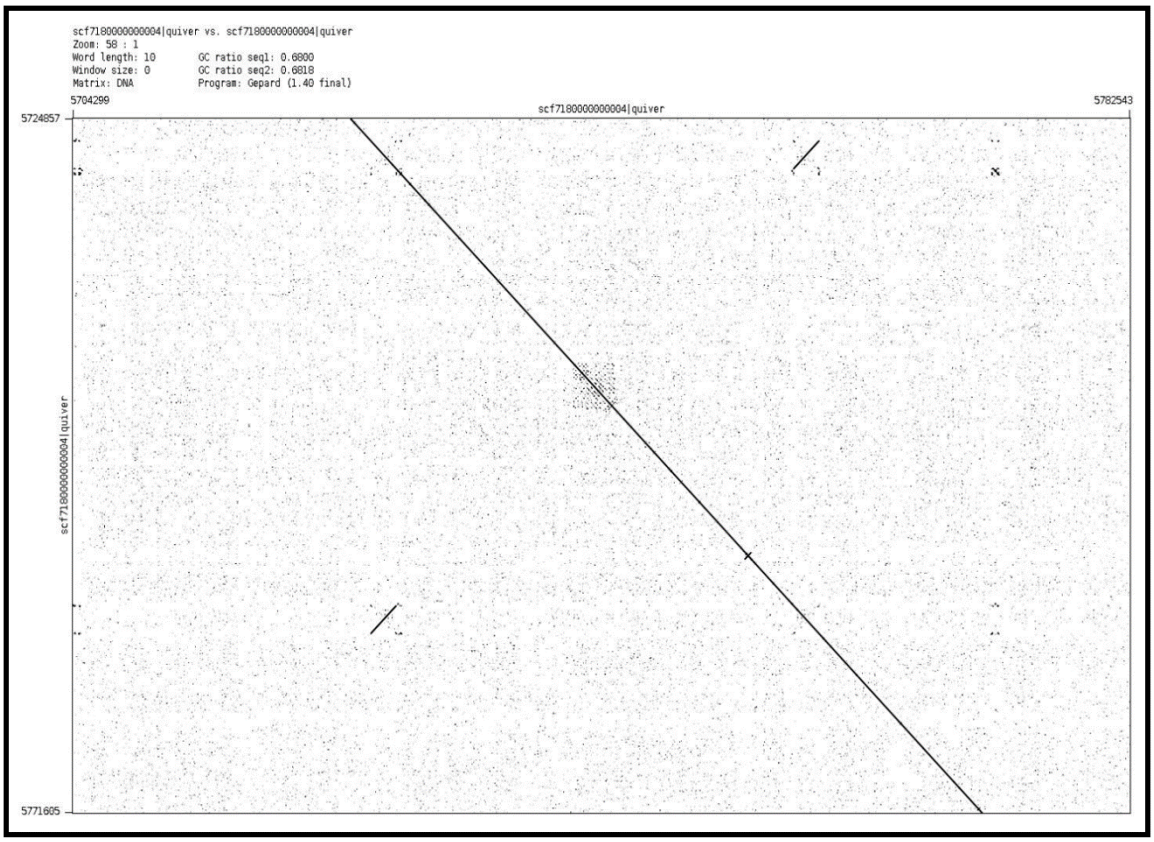


Figure 4.20: 8 Dot-plot of Inversion events in Pacbio assembly of 1198bp as predicted by FRCbam at positions 5757401bp and 5758599bp.

High number of compressed pairs show collapsed repeats as confirmed by REAPR in pacbio assembly (Figure 4.21). Stretched pairs result to fragmented assembly suggesting genomic relocation of a sequence as do LINK mis-assemblies reported by REAPR. As

discussed in Section 4.4.1, QUAST reports multiple relocation mis-asemblies in the contigs suggesting wrong placement of the sequence.

High number of FCD errors in 28 contigs suggest scaffolding errors. HIGH_OUTIE suggest that there is inversion or translocation present in the pacbio assembly, Low_covg_PE suggests error in connecting the contigs, present in highly fragmented assemblies Large number of wrongly oriented and wrong-distance pairs show inversions and insertion errors, which are present in pacbio containing misjoins, High span PE shows systematic error in correctly merging the contigs and wrong copy number estimation of the repeats. Also, there are misjoins present in both assemblies as the mates are mapped in nearby locus or at different locus (Appendix A1.7).

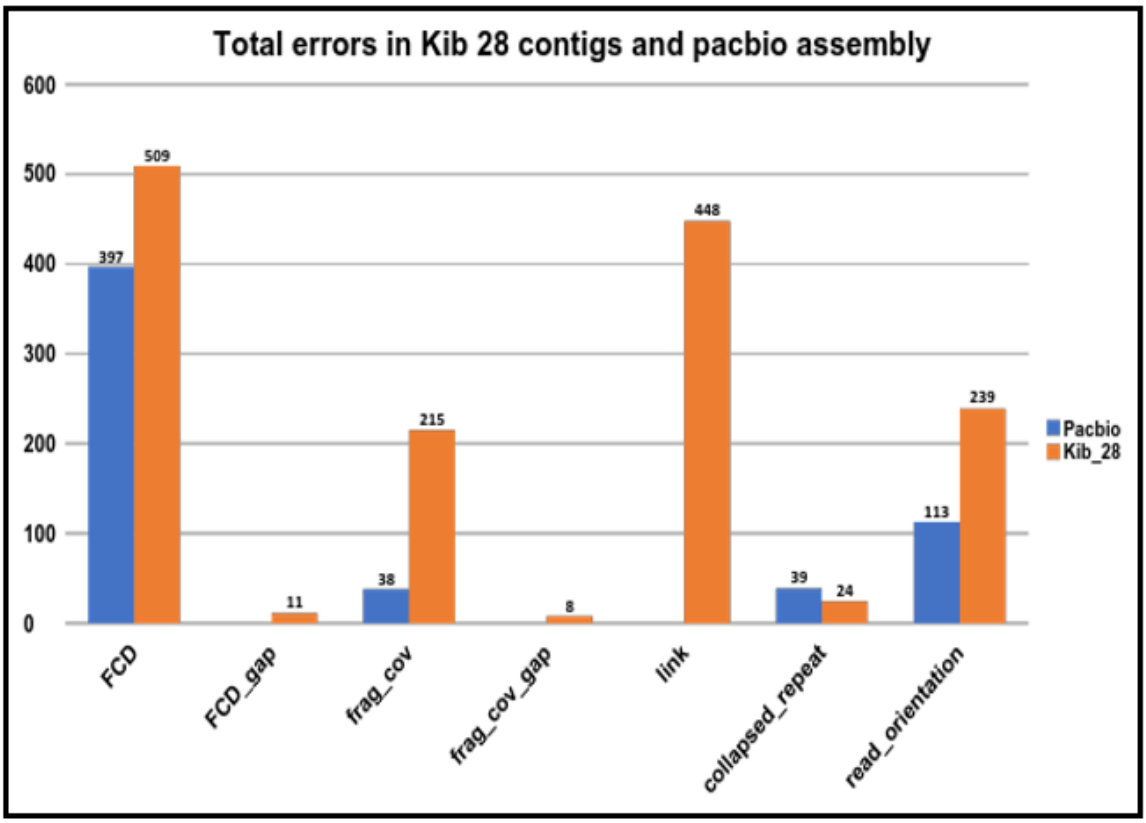


Figure 4.21: Total number of features in Kib 28 contigs and Pacbio assembly.

4.5 Circularity of the genome

With the advances in long-read technologies there are more and more complete bacterial genomes being sequenced (Rhoads & Au, 2015). Also, there are many linear bacterial genomes being sequenced (He et al., 2016), but if this information is correct and complete remains a question, as we study in this chapter. The first attempt to sequence and assemble *Kibdelosporangium* sequence resulted in high-quality draft genome with 21 scaffolds with no orientation information. Pacbio assembly suggests that the genome is linear based on BLAST results of the contig by itself (Appendix Figure A1.8). But, results looked different

when 28 contigs are aligned against the pacbio assembly. Here we present several arguments in favor of circularity of the genome.

1. Missing 225bp sequence in Pacbio assembly

Contig 18 aligns in forward and reverse directions at the end of the Pacbio assembly (Appendix Figure A1.2). QUAST results suggest that there is a -225bp overlap between Alignment 1 (10959991bp and 1211374bp (pacbio)) and (1153489bp and 1153496bp, (Contig18)) and Alignment 2 (1bp and 32338bp (Pacbio) and 1153722bp and 1186059bp (Contig18)) as seen in Figure 4.22.

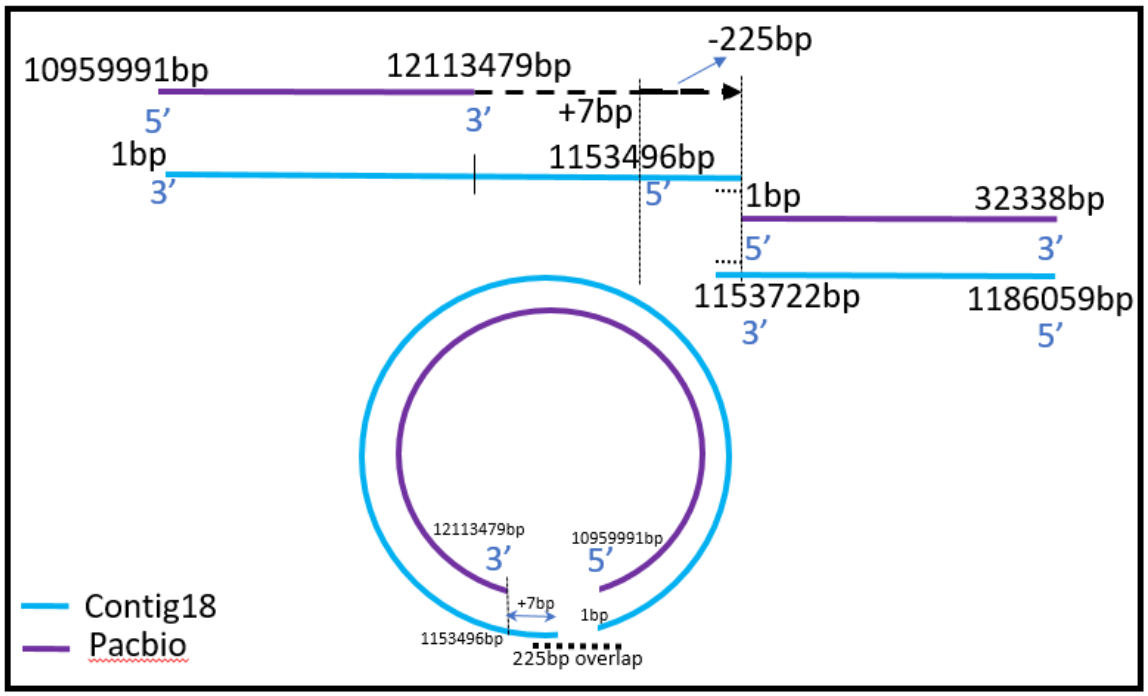


Figure 4.22: -225 bp alignment in Contig18 which is absent in Pacio assembly.

2. Presence of OriC site

Circular genomes have OriC sites near dnaA sequence. We could find dnaA sequence at middle of the assembly at positions 6086637bp - 6088187bp. The OriC sequence is almost always near to the dnaA sequence.

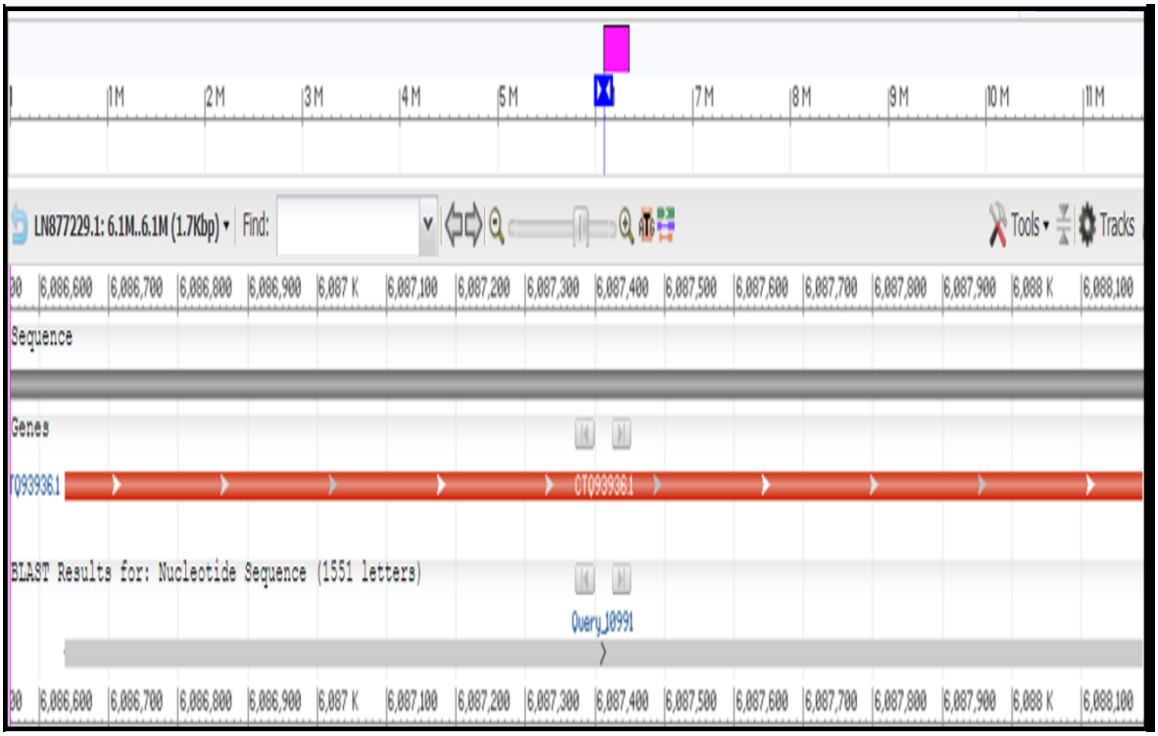


Figure 4.23: dnaA sequence in 12 Mbp pacbio assembly.

On plotting GC-skew and AT-skew plots (Figure 4.24) for pacbio assembly we could find regions for OriC and terC both. Usually, for circular chromosomes there is genomic polarity because of regions of high AT and high GC biasing on lagging and leading strands (Lobry & Louarn, 2003). Nucleotide composition mapping across the genome can be used to locate oriC and terC sites as a shift in GC skews. It has been documented that GC-skews have not been found in linear bacterial chromosomes, symbionts or archeal genomes

(Lobry & Louarn, 2003). Also, on doing further analysis we could find that OriC site is present near 5973773bp to 5974124bp.

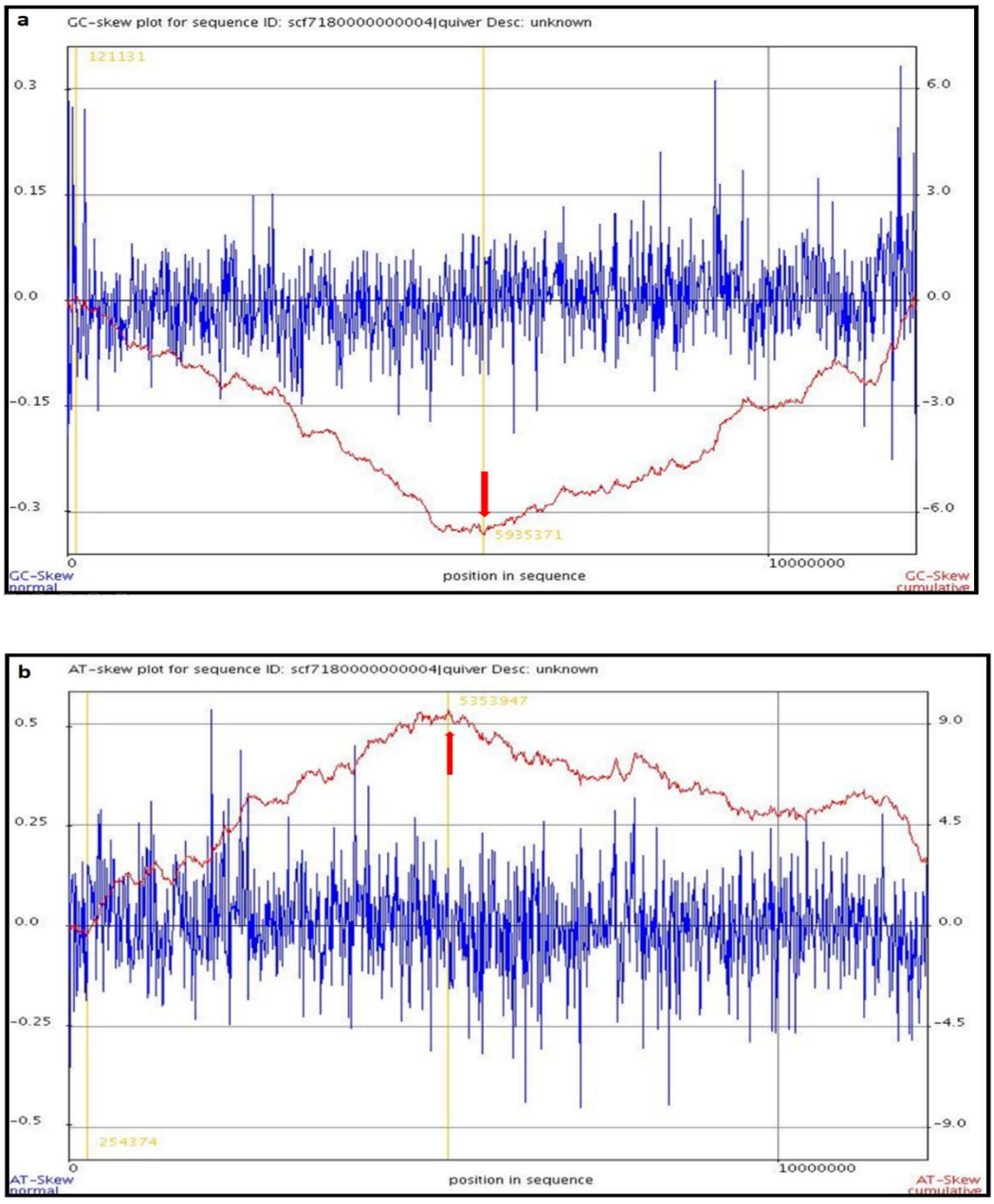
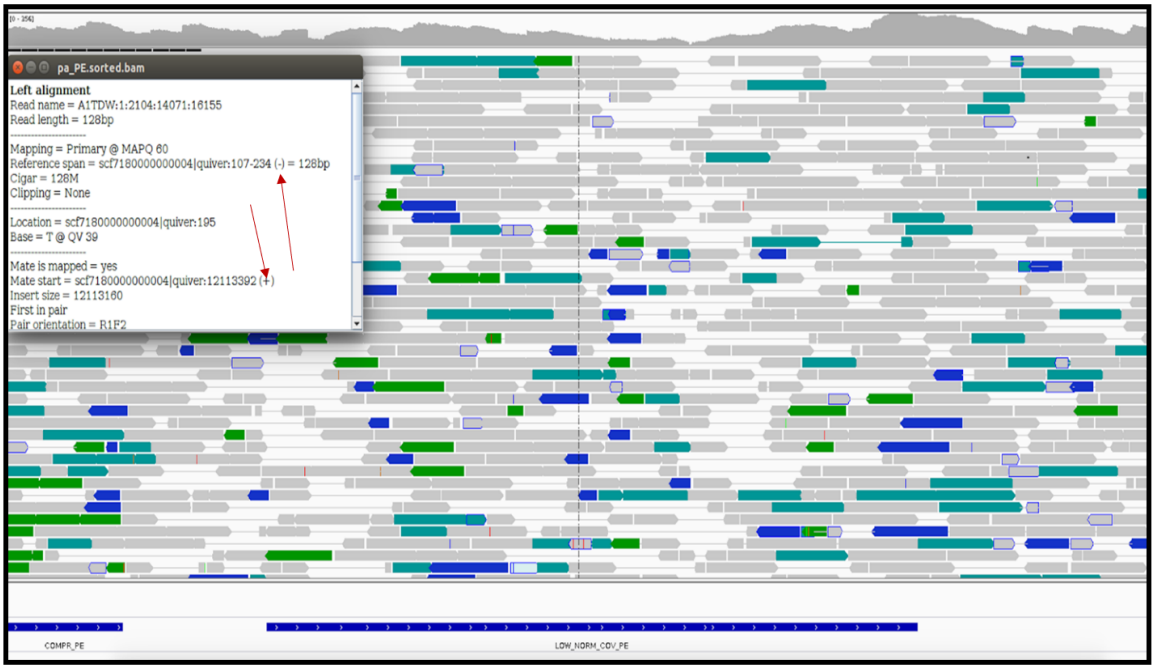


Figure 4.24: Genskew plots of Pacbio assembly, a) shows GC-skew and possible OriC site, b). AT-skew plot confirming a terC sequence opposite to OriC sequence in the graph. Red arrows show the position of OriC site and terC site respectively

3. Mate-pair validation

Another supporting evidence for circularity is found during mate-pair validation when paired-end reads are aligning on negative and positive strands at the end and beginning of the assembly. Figure 4.25 shows one such example. This suggests that there are regions which must be overlapping in the end and at the beginning of the assembly for possible circularism.



Gigure 4.25: Paired-end reads aligned to Pacbio assembly. The reads are colored according to mapping orientation on the forward and reverse strands. Red arrows are both mates on reverse strand, turquoise colored reads are both mates on positive strand and green colored are on positive and negative strand. Insert box shows details of one of the pairs whose first

mate aligns at the end at position 12113392bp on forward strand and other mate at position in beginning at from 107bp to 234bp on negative strand.

4. Miniature inverted terminal repeat analysis

MITE analysis of Pacbio assembly resulted in 6 sequences of 1024 bp total and 171bp each. This 1024bp sequence appears 14 times in pacbio assembly and 6 times in contig 18. BLAST with contig 18 showed that this sequence from pacbio is missing a significant repeat copy (Figure 4.26)

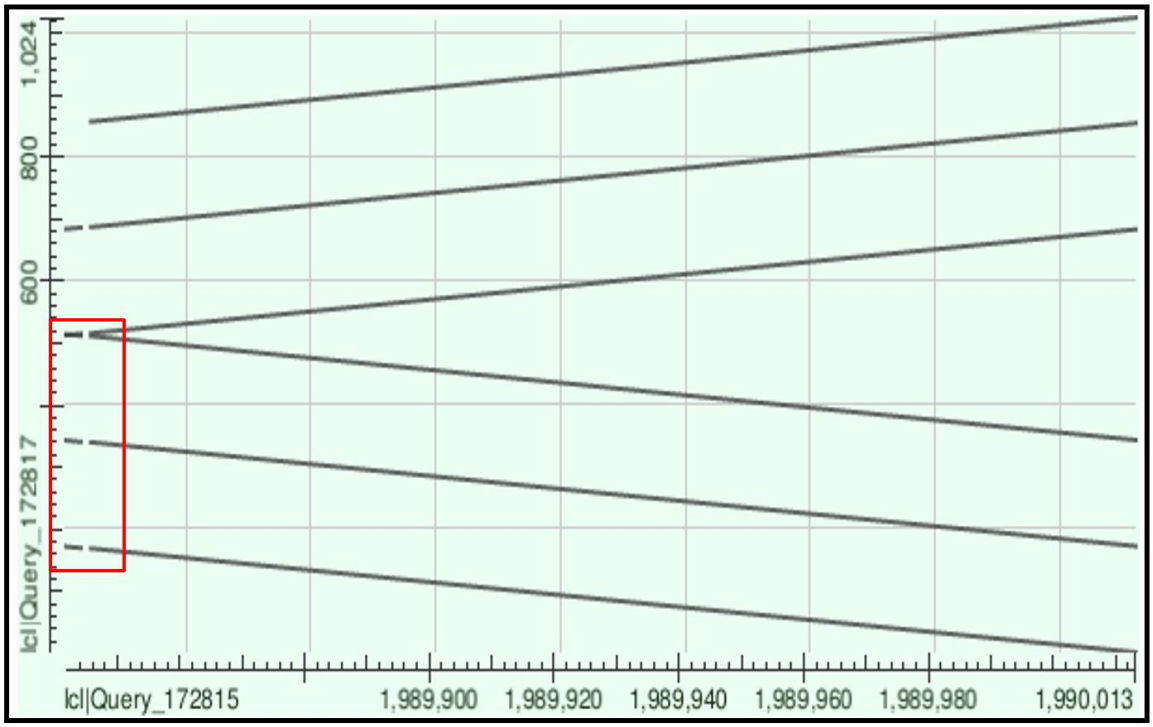


Figure 4.26: 1024 base-pair sequences from pacbio aligned against Contig18. Third and fourth sequence merge to one sequence but are missing a joining sequence. On adding contig18 to pacbio polished assembly and doing blast against it, we obtained dot plot

confirmation as of a circular contig as shown in Figure 4.27. The two diagonals at the corners suggest circularism but still there is sequence missing.

It might be possible that the terminal inverted repeats are part of tandem repeat copies found in section 4.3.5, as QUAST analysis and BLAST results suggest. The 225bp overlap might be a part of these repeats and is missing from the Pacbio assembly.

To gain full picture of Pacbio and 21 scaffolds, we do a dot-plot analysis (Figure 4.27). Right side are all contigs and left side are pacbio assembly sequences with 4,6,7,8 and 17 contigs being inverted. Contigs 2, 9, 11 and 20 have both inversion and re-arrangement. . Contig18,19,20 and 21 show that there might be circularization in this genome.

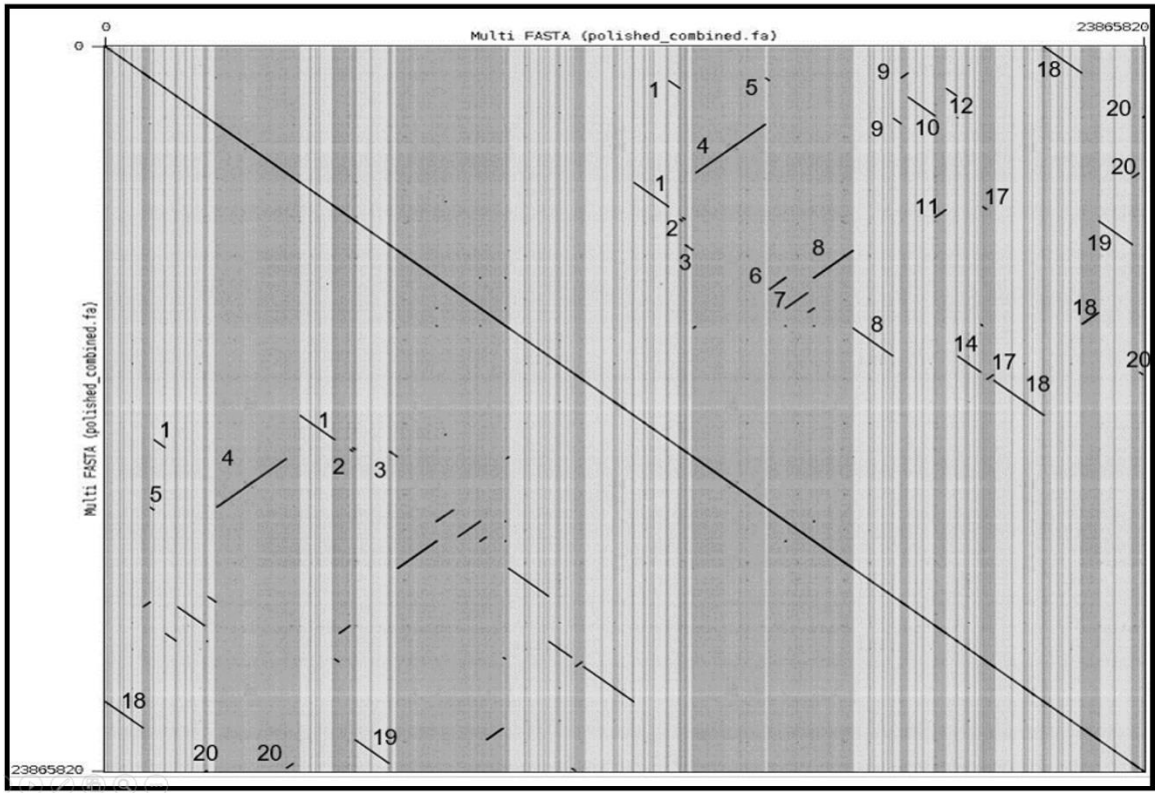


Figure 4.27: Multi-fasta Dot-plot representation of pacbio assembly and 21 scaffolds when blast against itself.

4.6 Discussion

Assembly contiguity is a crucial factor for distinguishing complete and finished genomes but precision of the de novo sequenced assemblies cannot be compromised. In this study, we have a single contig Pacbio assembly which suggests that Kibdelosporangium MJ-NF124 genome is 12 Mbp single continuous contig and is linear. Our results prove that this is not completely accurate. Pacbio assembly is missing crucial repeat sequences for complete circularization of the genome which is partially contained in contig18 from first assembly. Miniature inverted terminal repeats are mis-assembled and lack complete sequence which is contradicting as pacbio is shown to produce golden standard complete genomes (Rhoads & Au, 2015). But, our analysis shows that pacbio is not accurate. The assembly quality is slightly better than the draft genome, with high number of collapsed repeats, and major genomic re-arrangements. Draft genome assembly is fragmented and has locus re-arrangements as evident by paired-end mapping to different contigs. Currently, vastly used metrics such as N50 for assessing the contiguity of the assemblies is shown (Vezzi, Narzisi, & Mishra, 2012a) as more of an artefactual number than the real measures to define quality of the assembly. Aggressive concatenation of the contigs 21 scaffolds increases the N50 resulting in mis-assemblies as evident by QUAST results (Figure 4.28). More than 50% of the genome is mis-assembled, fragmented and has low coverage regions. Looking deeply, we could find multi-reads copies in the genome which caused repeat-induced overlaps and gaps in the scaffolds and thus, the complete genome could not be

assembled. Some of the errors observed in draft genome duplicated in pacbio assembly at the same regions. These repeats are either the insertion sequences related to type II transposons or inverted repeats in high number copies. Feature based analysis provides relationship between the features and causes of mis-assemblies. Pacbio, although has longer read connectivity but has major genomic relocations, whereas the contigs have local mis-assemblies.

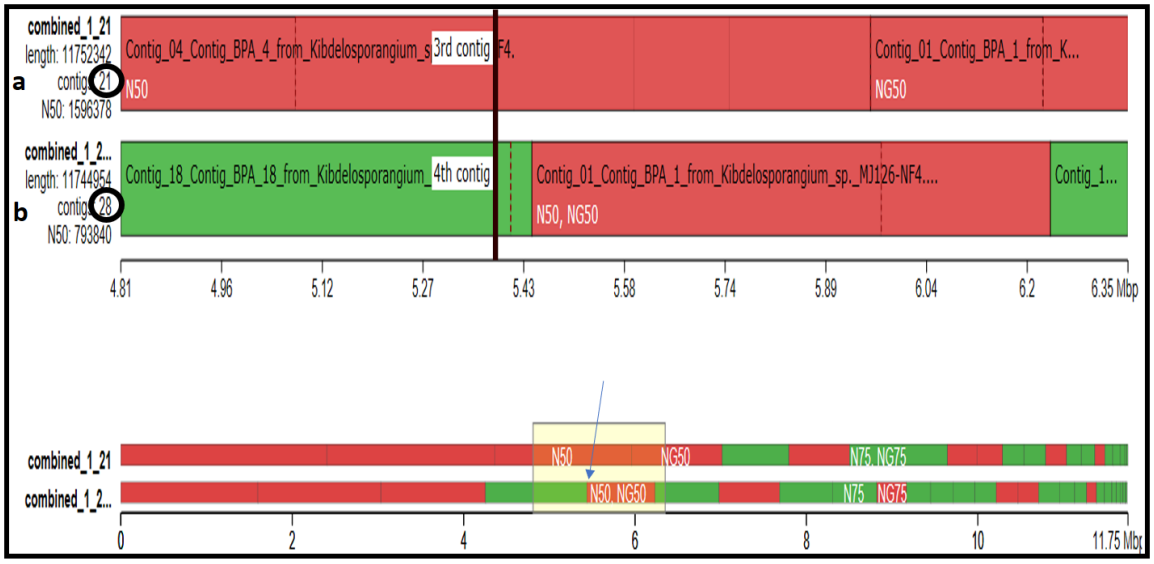


Figure 4.28: a) 21 scaffolds aligned to pacbio and b) 28 contigs aligned to pacbio after mis-assembly detection. Complete alignment schematic showing perfectly aligned(green) and mis-assembled (red) regions. Bottom portion shows that how N50 is increased and mis-assemblies are introduced in 21 scaffolds.

21 scaffold assembly is mis-assembled mostly due to repetitive content and transposons.

Most of the paired-end reads were badly oriented because of the repeats. Positioning of correct orientation of repeats is lacking. Contigs17 have mate-pairs aligned to another contigs20,21 and multiple contigs respectively, suggesting re-location mis-assemblies. Pacbio assembly has its share of mis-assemblies owing to collapsed repeats which are evident by significant numbers of broken read-pairs. FRCbam results and manual inspection suggest genome locus misplaced in the assembly and incorrect copy numbers. Draft genome could not be assembled as a complete genome due to high repetitive content and longer terminal repetitive elements in this genome

Pacbio assembly is just another estimation and not completely accurate. There is sequence missing and high number of genomic locus re-arrangement along with collapsed repeats, as evident by 25% of paired-end broken reads. Also, Pacbio reports this assembly as a linear contig, but we could find OriC sites, missing -225bp sequence aligning Contig18 forwards and reverse strand on the same alignment as to pacbio (also confirmed by unmapped broken-pairs), and multiple paired-ends that align to end and starting of the Pacbio assembly suggesting circularity of the genome.

Thus, we conclude that this genome is circular, that Pacbio assembly is not accurate and that first draft assembly could not be completed in first place due to presence of flanking repeats at the transposons and leading to gaps, overlaps or mis-assemblies.

Future work

Broken illumina pairs if suggest that repetitive elements are still not correctly assembled. Unmapped pairs suggest that Pacbio is missing some information. Algorithms that can deal with correct copies of repeats based on k-mer length compared with illumina and Pacbio

should be used. Hybrid assembly with paired-end information and long-read information should be used to resolve and accurately count small repeat structures.

References

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, *9*(4), 333–7. Retrieved from http://dx.doi.org/10.1038/nmeth.1935

Chawla, V., Kumar, R., & Shankar, R. (2016). Identifying wrong assemblies in de novo short read primary sequence assembly contigs. *Journal of Biosciences*, *41*(3), 455–474. https://doi.org/10.1007/s12038-016-9630-0

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., … Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569. https://doi.org/10.1038/nmeth.2474

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

He, J., Sundararajan, A., Devitt, N. P., Schilkey, F. D., Ramaraj, T., & Iii, E. M. (2016). Complete Genome Sequence of Streptomyces venezuelae ATCC 15439 , Producer of the Methymycin / Pikromycin Family of Macrolide Antibiotics , Using PacBio Technology. *Genome Announcements*, *4*(3), 3–4. https://doi.org/10.1128/genomeA.00337-16.Copyright

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., & Otto, T. D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, *14*(5), R47. https://doi.org/10.1186/gb-2013-14-5-r47

Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A., & Tse, D. N. (2016). HINGE: Long-Read Assembly Achieves Optimal Repeat Resolution. *bioRxiv*, 62117. https://doi.org/10.1101/062117

Koren, S., Treangen, T. J., Hill, C. M., Pop, M., & Phillippy, A. M. (2014). Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics*, *15*(1), 126. https://doi.org/10.1186/1471-2105-15-126

Lin, H. H., & Liao, Y. C. (2015). Evaluation and validation of assembling corrected pacbio long reads for microbial genome completion via hybrid approaches. *PLoS ONE*, *10*(12), 1–13. https://doi.org/10.1371/journal.pone.0144305

Lobry, J. R., & Louarn, J. M. (2003). Polarisation of prokaryotic chromosomes. *Current Opinion in Microbiology*, *6*(2), 101–108. https://doi.org/10.1016/S1369-5274(03)00024-9

Lupski, J. R., & Weinstock, G. M. (1992). Short, interspersed repetitive DNA sequences in prokaryotic genomes. *Journal of Bacteriology*, *174*(14), 4525–4529.

Manuscript, A., & Nanostructures, S. P. C. (2008). NIH Public Access. *Nano*, *6*(9), 2166–2171. https://doi.org/10.1021/nl061786n.Core-Shell

Ogasawara, Y., Torrez-Martinez, N., Aragon, A. D., Yackley, B. J., Weber, J. A., Sundararajan, A., … Melançon, C. E. (2015). High-Quality Draft Genome Sequence of Actinobacterium Kibdelosporangium sp. MJ126-NF4, Producer of Type II Polyketide Azicemicins, Using Illumina and PacBio Technologies. *Genome Announcements*, *3*(2), e00114-15. https://doi.org/10.1128/genomeA.00114-15

Phillippy, A. M., Schatz, M. C., & Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology*, *9*(3), R55. https://doi.org/10.1186/gb-2008-9-3-r55

Ponsting, H., & Ning, Z. (2010). SMALT - A New Mapper for DNA Sequencing Reads. *F1000Posters*, 1. https://doi.org/10.7490/F1000RESEARCH.327.1

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V., & Chandler, M. (2011). ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology*, *12*(3), R30. https://doi.org/10.1186/gb-2011-12-3-r30

Vezzi, F., Narzisi, G., & Mishra, B. (2012a). Feature-by-feature - evaluating De Novo sequence assembly. *PLoS ONE*, *7*(2). https://doi.org/10.1371/journal.pone.0031002

Vezzi, F., Narzisi, G., & Mishra, B. (2012b). Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. *PLoS ONE*, *7*(12), 1–11. https://doi.org/10.1371/journal.pone.0052210

Wetzel, J., Kingsford, C., & Pop, M. (2011). Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics*, *12*(1), 95. https://doi.org/10.1186/1471-2105-12-95

Appendix 1

Quast results:

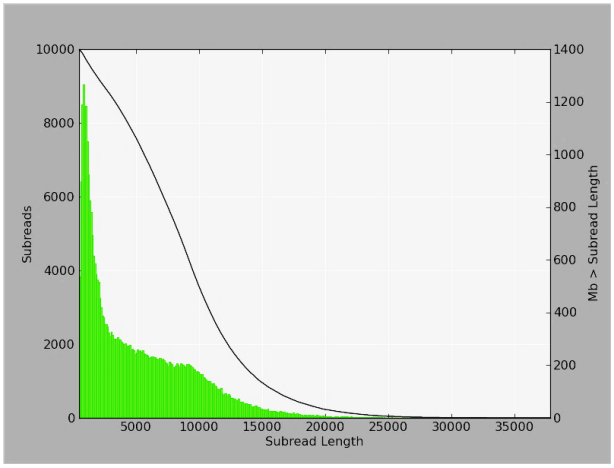
1.file:///F:/Priyanka/kib/QUAST_full_output/kib_quast/icarus_viewers/alignment_viewer.html

2. De novo assemblies are prone to mis-assembled regions due to sequencing and assembly errors, repetitive content, transposons and insertion sequences. Aggressive assemblers produce longer contigs and scaffolds but are more likely to join regions in the wrong order and orientation. Though researchers are working out ways to identify telltale signs of misassemblies and correct them, errors are hard to detect.

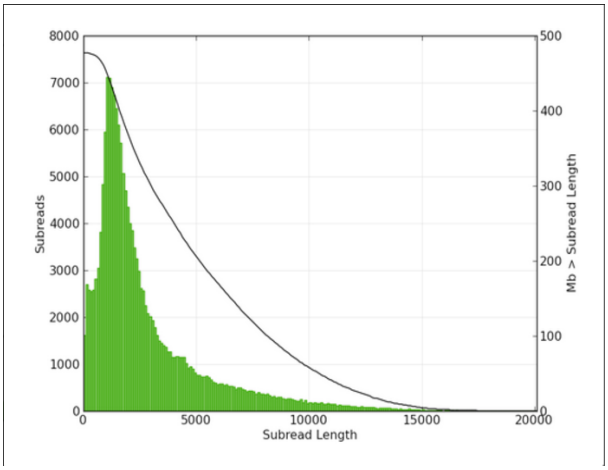
Comparison of old Pacbio and newer Pacbio assembly sequencing data

scf718000	.	HIGH_OUTIE_PE	15001	16199	.	+	.	Name=HIGH_OUTIE_PE
scf718000	.	HIGH_OUTIE_PE	2789401	2790399	.	+	.	Name=HIGH_OUTIE_PE
scf718000	.	HIGH_OUTIE_PE	5757401	5758599	.	+	.	Name=HIGH_OUTIE_PE
scf718000	.	HIGH_OUTIE_PE	11727201	11728399	.	+	.	Name=HIGH_OUTIE_PE
scf718000	.	HIGH_OUTIE_PE	12095601	12096999	.	+	.	Name=HIGH_OUTIE_PE

Table A1.1: High outie features shown by FRCbam suggesting inversions which are confirmed by dot-plot (Figure 4.4.8) at highlighted position in table.



Number of SMRT Cells: 1
Library Protocol: 2kb - 10kb
Number of Reads: 173,690
Total Number of Base Pairs: 486,086,375



Total Number of Reads: 275,718
Total Number of Bases: 1,413,776,574
Mean Subread length: 5,127
Read Length N50: 8,476
Sequence Coverage: Approximately 117X (Based on a genome size estimate of 12 Mbps)

Read length distribution of Pacbio sequencing reads.

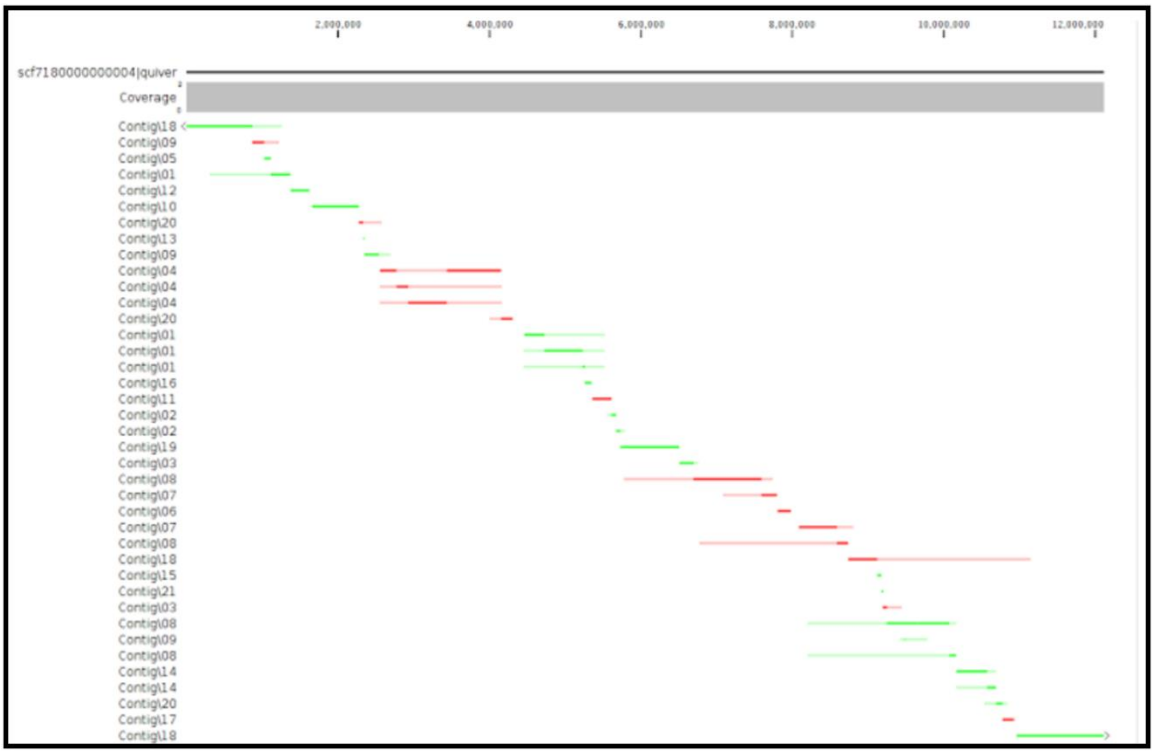


Figure A1.2

ISA2s8	IS[illegible]3		Azospirillum sp.	30.2	5.0
ISGeob5	IS5	IS427	Geodermatophilus obscurus	30.2	5.0
ISAs12	ISAs1		Aeromonas salmonicida	30.2	5.0
ISMno33	IS1182		Methylobacterium nodulans	30.2	5.0
ISAau2	IS21		Arthrobacter aurescens	30.2	5.0
ISCur1	IS1380		Corynebacterium urealyticum	30.2	5.0
ISMex35	IS5	IS5	Methylobacterium extorquens	30.2	5.0
ISStma3	IS481		Stenotrophomonas maltophilia	30.2	5.0
ISThsp20	IS5	IS427	Thiomonas sp.	30.2	5.0
ISAzo40	IS1380		Azoarcus sp.	30.2	5.0
ISLxc3	IS30		Leifsonia xyli	30.2	5.0
ISLxc4	IS30		Leifsonia xyli	30.2	5.0
ISCro6	IS4	IS4	Citrobacter rodentium	30.2	5.0
ISNisp3	IS3	IS407	Nitrobacter sp.	30.2	5.0
ISSpo2	IS1380		Silicibacter pomeroyi	30.2	5.0
ISDra7	IS4	ISPepr1	Deinococcus radiodurans	30.2	5.0
ISFsp4	IS200/IS605	IS605	Frankia sp.	30.2	5.0

Figure A1.3: ISfinder blast results for 264027bp Contig 1 864bp sequences.

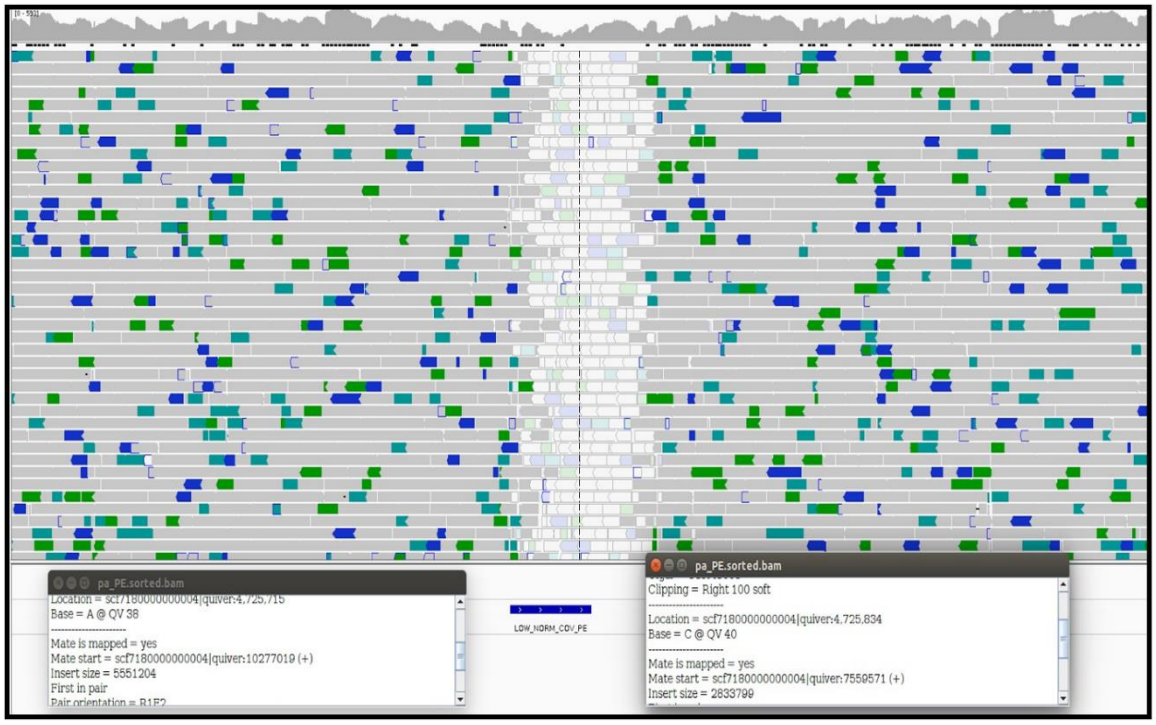


Figure A1.4: Pacbio region 472409bp to 4724589bp corresponding to insertion gap error at Contig1.

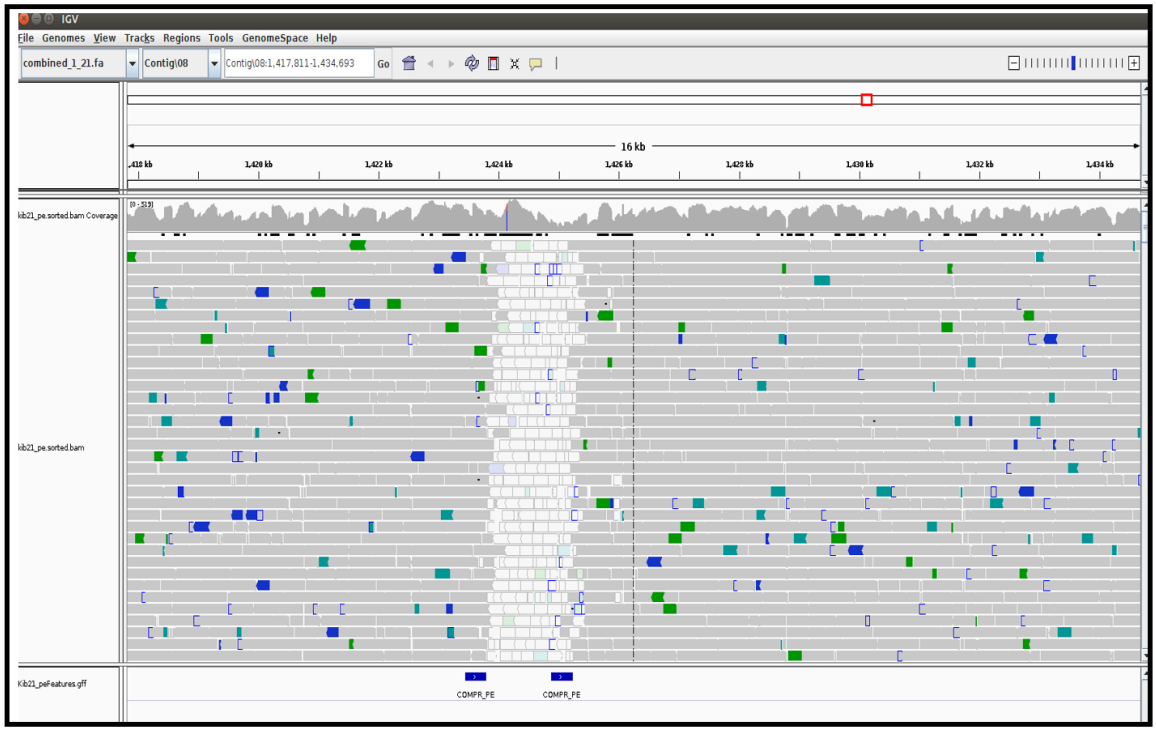


Figure A1.5: Compressed-paired ends at both ends resulting in high coverage area on either side.

Figure A1.6: Mate-pairs aligning to the beginning and end of the assembly

Figure A1.7: Mates aligning to different locus in Pacbio assembly suggesting genomic rearrangements.

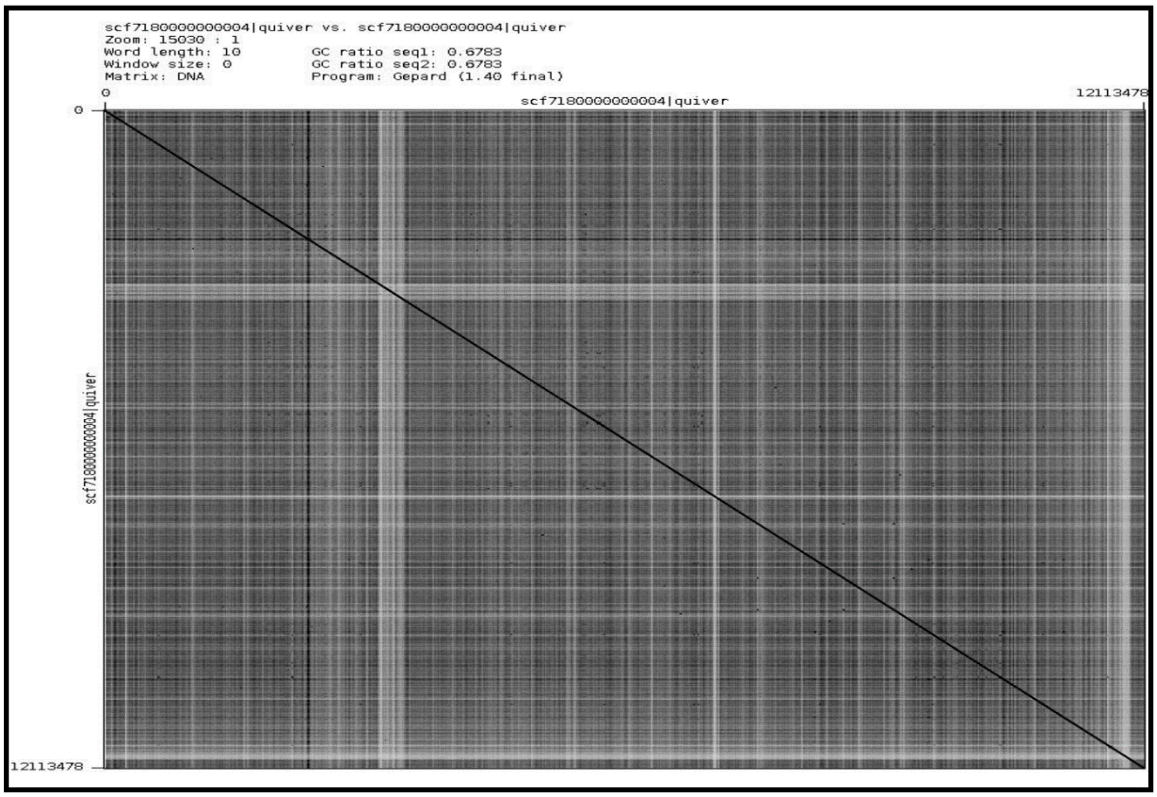


Figure A1.8

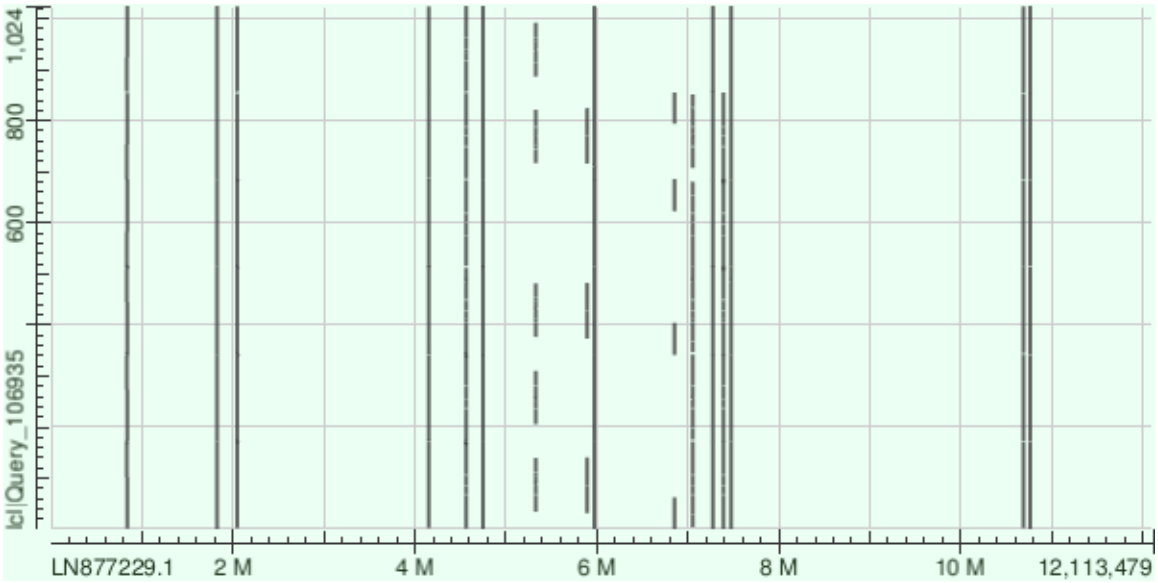
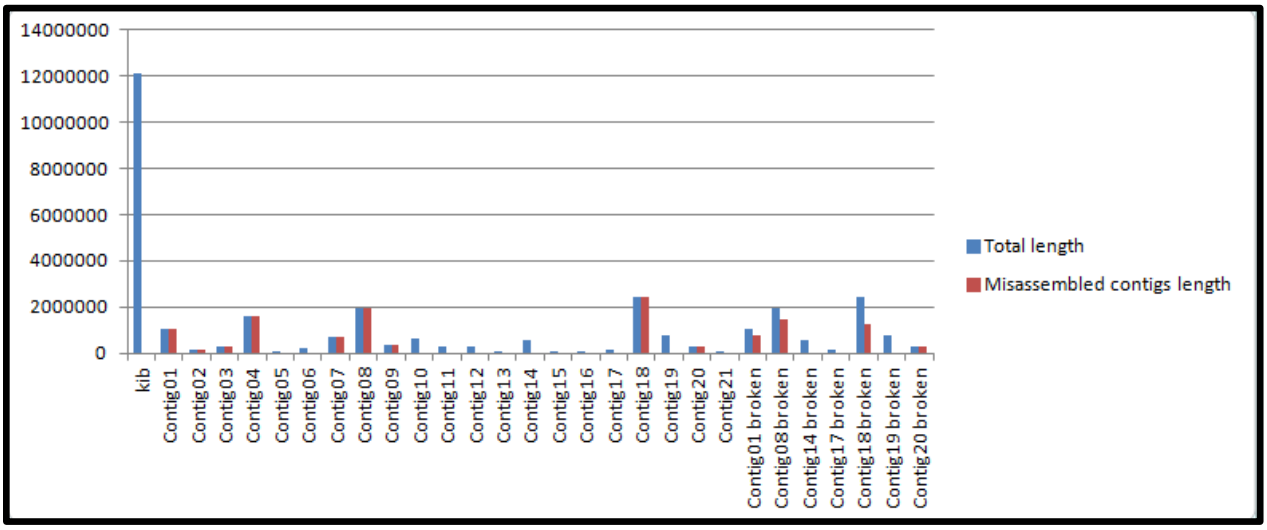


Figure A1.9: 1024 base-pair sequence MITEs in Pacbio assembly



Most of the broken pairs align to a repetitive region. Unmapped paired-end reads are selected and mapped to the assembly, it shows many reads aligned to one high coverage fragment which when blasted against the original assembly gives highly repetitive matches throughout the single contig.

Supplementary Table –

ACRONYMS USED –

Combined 1_21 -21 scaffolds without any break-points

Combined 1_21_broken – 21 scaffolds broken at positions where mis-assemblies are detected by QUAST.