University of Leicester

Just how good is third generation sequencing for complex genomes?

Esma Karaarslan

Report of work submitted for the M.Sc. in Molecular Genetics in the Department of Genetics, University of Leicester

July 2013

Word count: 8,407

DECLARATION

All sentences or passages quoted in this project dissertation from other people's work have been specifically acknowledged by clear cross referencing to author, work and page(s). I understand that failure to do this amount to plagiarism and will be considered grounds for failure in this module and the degree examination as a whole.

Name: Esma KARAARSLAN

Signed:

Date: 31. 07. 2013

Contents

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Edward Hollox, to introduce me new trends in genetics. Furthermore, this dissertation would not have been possible without the help, support, and patience of my supervisor. I would also like to thank PhD students, Shamik Polley, Razan Abujaber and Ezgi Kucukkilic who took part in my study for generously sharing their time and ideas. Additionally, my sincere appreciation also extends to all my family. They were always supporting and encouraging me with their best wishes, also understanding me during the long years of my education. My research would not have completed without their helps. Lastly, I owe sincere and earnest thankfulness to my invaluable network of supportive and generous friends. Thanks for everyone who are involved both directly and indirectly in the completion of my thesis.

ABSTRACT

Just how good is third generation sequencing for complex genomes?

Esma Karaarslan

The sequencing technologies have been pioneers in field of genomics. Sanger sequencing as a first generation sequencing (FGS) technique has been the first significant method for Human Genome Project. Numbers of publishes had done through this method but it had to be improved in terms of read length, time consuming and high cost. To reduce cost and increase throughput, second generation sequencing (SGS) techniques have appeared on market. However, still read length has been the main challenge for complex genome sequencing. In this case, third generation sequencing techniques (TGS) have provided large-scale sequencing and real time single molecule detection to exceed previous limitations. Our study covers of testing Pacific Bioscience Single Molecule Real Time Sequencing Technique (SMRT) as a TGS method to analyse both the single base and the molecular level of complex genomes, also assemble a contiguous BAC sequence across the macaque beta-defensin region, and perform a preliminary analysis of sequence variation. Analyses of restriction enzyme maps and BAC end sequencing have referred misassembling and also data analysis of gene content has confirmed sequence error in TGS. As a result, this study has demonstrated that SMRT technology has some gaps for sequence and assembling accuracy even though it provides long read length and direct assembly. Therefore, in future, enzyme kinetics can be improved to increase the accuracy of single molecule sequencing in real time. Furthermore, if higher level dimensional data is accurately analysed and assimilated, TGS will contribute towards a better understanding of large-scale DNA sequencing.

ABBREVIATIONS

µg	Microgram
µM	Micromolar
µl	Microlitre
ng	Nanogram
ml	Mililitre
g	Gram
DNA	Deoxyribonucleic acid
ddNTP	Dideoxynucleotides
ddATP	DideoxyAdinine-tri-phosphate
ddGTP	DideoxyGuanine-tri-phosphate
ddTTP	DideoxyThymine-tri-phosphate
ddCTP	DideoxyCytosine-tri-phosphate
SGS	Second generation sequencing
PCR	Polymerase chain reaction
SBS	Sequencing-by-synthesis
SNP	Single nucleotide polymorphisms
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
TGS	Third generation sequencing
PGM	Personal genomic machine
PacBio	Pacific Biosciences
SMRT	Single Molecule Real Time
ZMW	Zero mode waveguide
BAC	Bacterial artificial chromosome
CBCS	Clone-by-clone shotgun
WGS	Whole-genome shotgun
HIV	*Human immunodeficiency virus* and
AIDS	Acquired immunodeficiency syndrome
CNVs	Copy number variations
OR	Olfactory repeat

List of Figures

List of Tables

Chapter 1: INTRODUCTION

1.1 Sequencing technologies

Oswald Theodore Avery determined deoxyribonucleic acid (DNA) as genetic material in 1944. James D. Watson and Francis Crick introduced DNA structure as double helical strand composed of four nucleotide bases in 1953 (Liu *et al.*, 2012). The sequencing technologies were developed in field of genomics (Kircher & Kelso, 2010). The first and second generation sequencing techniques, have made it possible to analyse whole genome sequences, different isoforms of genes, chromatin conformation, nucleic acid structure, point mutation, copy number variation, transcriptome and methylome detections that gives information to understand the association between DNA and protein (Schadt *et al.*, 2010), (Meyerson *et al.*, 2010) (Kircher & Kelso, 2010). However, complex region of mammalian genome comprising repetitive elements, inversions and duplications are still obstacles in front of complex genome sequencing due to , this information is helpful for the researchers and health care practitioners to find out key information for many biological questions - hence, the sequencing technologies should be low cost, fast and with accurate results (Liu *et al.*, 2012). There was tremendous development in the sequencing technologies from the past thirty years that makes possible to characterize vast amount of genomic data (Liu *et al.*, 2012), However, the third generation sequencing (TGS) offers some advantages over Sanger and second generation sequencing limitations and attempt to overcome these obstacles with (i) the real time detection of biological process at single molecule resolution, which does not require PCR-induced bias; (ii) large-scale sequenced read lengths enable to de novo assembly; (iii) considerably low cost and quick time for preparation process; (iv) direct RNA sequencing instead of cDNA sequencing (Ozsolak, F., 2012; McCarthy, A., 2010). The study describes about previous technologies and their limitations compare to the third generation technologies.

1.1.1 First Generation Sequencing

In previous days, the first generation techniques was carried out by capillary electrophoresis based on semi automated implementation (Shendure & Ji, 2008). In 1975, the first method for sequencing DNA was introduced by Frederick Sanger. In 1977, two techniques Maxam and Gilbert: chemical cleavage and Sanger: dideoxy sequencing were published for DNA sequencing (Pareek *et al.*, 2011), had major advantages to determine exonic mutations, copy number variations and other DNA alterations (Meyerson *et al.*, 2010). The automated Sanger sequencing was introduced by Caltech (Pareek *et al.*, 2011). The Sanger's method is flexible and easily automated for multiple platforms including high throughput sequencing (Williams, Vincent et al., 2009). The automated DNA sequencing method was successful and used to sequence the first human genome (Pareek *et al.*, 2011). Sanger method is used by addition of dideoxynucleotides (ddNTP) which acts as inhibitor on DNA polymerase and the nucleotides are incorporated into growing DNA strand (Sanger *et al.*, 1977). The nucleotide fragments with four colours radiolabelled nucleotides end with particular ddNTP (ddATP, ddGTP, ddTTP, or ddCTP) at 3' end (Schadt, *et al.*, 2010). The nucleotide fragments are run via capillary electrophoresis; the fragments are separated on acrylamide gel results with pattern of bands by distribution of nucleotides (Sanger *et al.*, 1977). The labelled fluorescent nucleotides are identified by Laser and printed out from an automated sequencer. The main limitations of this approach are related to its high cost and time consuming (Schadt, *et al.*, 2010).

Maxam and Gilbert: chemical cleavage technology is another method which was published along with the Sanger (Pareek *et al.*, 2011), used to break the DNA sequence with chemical agent at each base followed with terminally labelled nucleotide. After partial cleavage, every single base is extended from the labelled end (Maxam & Gilbert, 1977). The four different reactions are used to cleave the four nucleotides guanines, adenines, cytosines and thymine and these nucleotides are resolved according to their size by polyacrylamide gel electrophoresis (Maxam & Gilbert, 1977). The gel was analysed by autoradiography to read the sequence of DNA molecules which are labelled with radioactive elements by the pattern of bands. At least 100 bases are sequenced by cleavage technique. The sequencing gels are resolved is the main limit of this technique (Maxam & Gilbert, 1977).

1.1.2 Second Generation Sequencing

The second generation sequencing (SGS) techniques overcome the first generation techniques (Kircher & Kelso, 2010). The SGS techniques were used to sequence the whole human genome of read length 35 - 400 bp with more speed which reduces cost and time compare to Sanger technique. The SGS techniques platforms include 454 sequencing, Solexa technology, SOLID, and IonTorrent (Schatz *et al.*, 2010).

Roche 454 sequencer was introduced by Roche in 2005 (Balzer *et al.*, 2010) was first commercially available SGS technique based on pyrosequencing technology (Shendure & Ji, 2008) and done by three steps – template preparation, constructing libraries and sequencing (Rothberg & Leamon, 2008) was shown in Figure 1. pyrosequencing technology considered as good method for single nucleotide polymorphisms (SNP) and was not good enough for standard sequencing needs cause of short read lengths (Rothberg & Leamon, 2008).

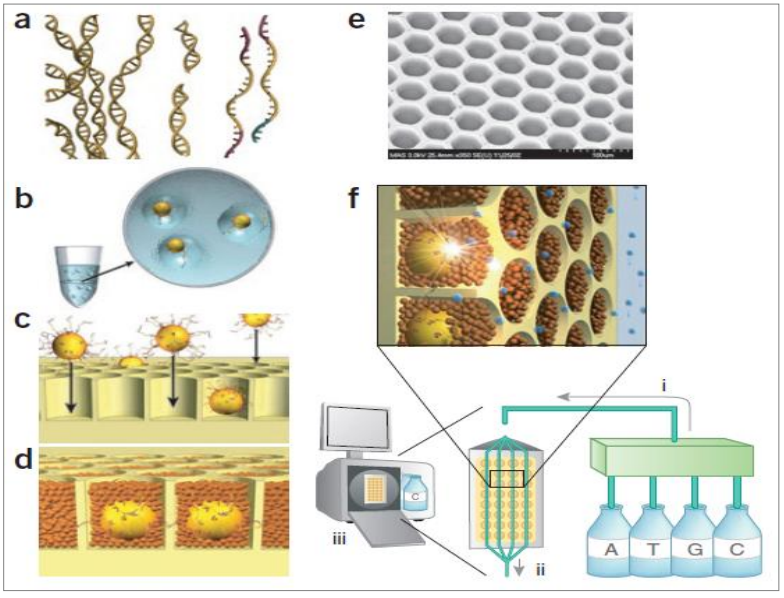


Figure 1.1: The mechanism of Roche 454. a) Fragmentation of Genomic DNA, ligate to adapters and separate DNA into single standard; b) DNA fragments are captured on 28µm beads and amplify emulsion PCR; c) Water is broken in oil emulsion, the beads are centrifused and deposited into picolitre reactors; d) Add DNA polymerase as well as sulfurylase and luciferase enzymes and loaded into beads; e) Fiber optic slide with wells before deposition of beads; f) The

454 sequencing machine containing (a) Reagents are flowed across the wells; (ii) A fiber-optic slide with samples deposition in wells; (iii) CCD camera for imaging sequencing within the wells, computer to interface the information and to control instrument (Rothberg & Leamon, 2008).

This method based on sequencing by synthesis (SBS) approach that utilizes bioluminescence which is captured by charged coupled device (CCD) camera that incorporate nucleotide base with in growing chain. The SBS approach starts with the construction of short and adapter-flanked fragments libraries and was captured on 28μm beads. The single stranded DNA sequence amplified by emulsion based PCR and generates million copies of DNA fragments on each bead. Then the pyrosequencing method was performed and the addition of nucleotides followed by DNA polymerase as well as sulfurylase and luciferase enzymes results in releasing pyrophosphate molecules. The enzymes catalyzes the incorporated nucleotides in the DNA strand and drive flash light which is proportional to the sum of incorporated nucleotides and followed by well wash with apyrase to remove unincorporated nucleotides (Shendure & Ji, 2008)(Dale *et al.,* 2012, Rothberg & Leamon, 2008). 454 is an pyrosequencing approach which depend on light detection that does not require electrophoresis to identify the nucleotide base in the DNA strand. Hence, unlike electrophoresis - the physical length which required to get accurate resolution of distinct fragments that limit miniaturization – it can be possible to reduce pyrosequencing to any reaction volume which is capable to generate levels of light which are detectable. The light which was released also allowed the sequencing possibly to be done in parallel (Rothberg & Leamon, 2008). The decreasing base quality is the limitation caused due to drop of sequencing runs that was effected by the polymerases and luciferases (Kircher & Kelso, 2010). The major limitation of pyrosequencing performs poorly on homopolymer sequences and short-read lengths (Chen F *et al.,* 2006).

Illumina genome analyzer or solexa platform utilize SBS approach and capable of sequencing DNA in low cost and high throughput compare to 454 technology (Strausberg et al., 2008). The solexa technique begin with preparation of library constructing adaptor-flanked fragments similar to 454 technology but varies with 454 in template preparation and amplification process (Shendure & Ji, 2008). To amplify fragment DNA, Illumina relies on solid substrate through bridge PCR as shown in figure 2. The amplification of fragment DNA was done on solid phase

PCR by flexible linker. At each cycle of sequencing, the primer extension reaction involves the addition of four modified nucleotides with fluorescent reversible terminators which forms at 3'-OH position (Shendure & Ji, 2008). The information of the nucleotide sequence was recorded at when flow cell interrogate with the laser beam and produce fluorescent signal at each base. Cleavage of the fluorescent label and termination of the moiety from the incorporated nucleotide occurs and followed for next nucleotide.

Finally multiple cycles results to identify each single base and built the sequence (Shendure & Ji, 2008) (Rothberg & Leamon, 2008). The major limitations of this technology were short read lengths and substitution errors caused by modified substrate and polymerase during the incorporation of nucleotide based on SBS reactions inefficient (Strausberg et al., 2008).

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencer was introduced by J.S. and colleagues in 2005 (Shendure & Ji, 2008) and it is based on polonator technology (Shendure *et al.*, 2005). Similar to 454, the preparation of this process starts with constructing short adaptor flanked fragments and emulsion based PCR. The clonal amplification was done by emulsion based PCR with the fragment DNA on 1μm paramagnetic beads. SOLiD SBS approach is done by DNA ligase other than polymerase and at each sequencing round involved addition of fluorescently labelled octamers (Shendure & Ji, 2008). The di-nucleotide is identified at the base 5 position within the octamer that ligated to the primer followed with fluorescent detection. The imaging was produced in four channels after ligation. The nucleotides are cleaved at the labelled position at the linkage between 5 and 6 bases of the octamer and remain free end for the next cycle of ligation. Performing many cycles will lead to discontiguous sequence with gaps. The primer sequence is denatured and repeated with several bases so that the discontiguous sequence is interrogated after next cycle of ligation (Shendure & Ji, 2008). Significantly, the limitation of this technique was obviously read length, emulsion based PCR can be difficult to handle and technically challenging, 1-μm bead arrays with high-resolution limit one pixel per sequencing (Shendure & Ji, 2008).

IonTorrent sequencing platform is based on pyrosequencing approach and implemented in the personal genomic machine (PGM). This technology is fast, low expensive, affordable and

produce long reads with several hundred bases (Golan & Medvedev, 2013). This technology is exactly the same principle as 454, except the addition of each nucleotide is detected by the release of an H+ ion rather than a flash of light. The signal that generate on each well was measured that indicates the number of incorporation. The disadvantages of this technique was produce errors while base calling which will be difficult for resequencing project that make confusion of SNPs (Golan & Medvedev, 2013)

1.1.3 Third generation sequencing

TGS techniques are based on Single Molecule Sequencing (SMS), assembling large read lengths within minutes and low cost, thus can provide an integration of genomics, transcriptomics, metabolomics, epigenomics (Schadt, 2010; Ozsolak, 2012) (McCarthy, 2010). Pacific Biosciences Single Molecule Real Time (SMRT) sequencing and Helicos True Single Molecule Sequencing have been developed by third generation technologies. Although, the SGS techniques were truly revolutionized in the genomics field, DNA sequencing and also in knowing the information about genome-related diseases (Schadt, 2010; McCarthy, 2010), this technology have the drawback of short read lengths range between 30 to 450 bases and also have inherent error rate (McCarthy, 2010). The TGS techniques overcome these issues, with potential long read lengths in short time with lower cost. The single molecule sequencing does not require PCR amplification bias, was an advantage to avoid errors while constructing libraries or amplification process (Kircher & Kelso, 2010).

Heliscope sequencer does not perform clonal amplification. This platform utilizes fluorescence detection system that interrogates single nucleotide bases by SBS approach (Shendure & Ji, 2008). The libraries are constructed by fragmentation of DNA, the poly-A tailed templates are captured by hybridization to poly-T oligomers that results to form an disordered sequence of primed single-molecule sequencing templates. Each round contain addition of polymerase and fluorescently labelled nucleotides results with template extension of primer duplexes, the Cy3 labeled nucleotides undergo chemical cleavage and detected by imaging full array to identify subset of array coordinates (single nucleotide bases) (Shendure & Ji, 2008).

The Pacific Biosciences (PacBio) is a new field with novel technology **single molecule real time (SMRT)** sequencing. The SMRT is the first third generation sequencing technology which was commercially available in late 2010, was done by the instrument zero mode waveguide (ZMW). The ZMW contain many chambers, the diameter of each well hole is 70 nm with 100 nm in depth. The single molecule DNA polymerase complexed with the template was affix at the bottom of the well (McCarthy, 2010) as shown in the Figure 1.2. The diffused nucleotides are labelled with fluorescent colours undergoes incorporation, bases with each of different fluorescent dye was identified by the detector through the signal emitted out of the ZMW (McCarthy, 2010) (Quail *et al.*, 2012). Addition of each base catalyze by the enzyme polymerase, cleaves the fluorescent tag. The main principal of single nucleotide sequencing technology is to avoid unwanted background noise which was created by the biological building materials, ZMW technology come out with the solution for this problem (McCarthy, 2010). The translations were occurred efficiently when the substrates are used at very high concentrations was the limit which faces during translation. The advantages of this technology is to provide long reads up to 10,000 bases in length, nearly 10,000 to 20,000 times faster than SGS technology and easy to assemble of unknown genome. This method is used to find out the rare mutations among people to the molecular level. One of the main advantages of this technology was possible to study biological processes in real time at particular concentrations (McCarthy, 2010).

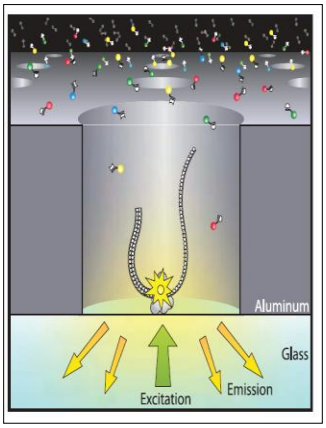


Figure 1.2: The single molecule synthesis in real time with zero mode waveguide (Eid *et al.* 2009)

	First generation	Second generation	Third generation
Fundamental technology	Size-separation of specifically end labeled DNA fragments, produced by SBS or degradation	Wash-and-scan	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Average across many copies of the DNA molecule being sequenced	Average across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800-1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base, Low cost per run	Low cost per base, High cost per run	Low-to-moderate cost per base, Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because	Complex because of

	of large data volumes and because short reads complicate assembly and alignment algorithm	large data volumes and because technologies yield new types of information and new signal processing challenges

Table 1.1: Comparison of sequencing technologies: FGS, SGS, and TGS (Schadt & Turner, 2010).

The table 1.1 describes the differences among the three sequencing techniques FGS, SGS and TGS. From the above table, it is clearly observed that there is a huge development of TGS technologies in improving read length and throughput characteristics compare to SGS technologies (Schadt & Turner, 2010).

1.2 Complex genomic regions

It is challenging to determine structural variations, complex and dynamic interactions between genes and environmental regions, structural diversity among species and structural differences among chromosomes (Redon et al, 2006; Johnson et al., 2001; Lupski, 2007). Structural variations include translocations, inversions and copy number variations (CNVs). The dynamic regions of different healthy individuals have significant amount of CNVs. The human beta-defensin regions are good example for structural variation which is informative for further studies on different structural variable regions of humans and other mammals. The beta-defensin regions are assembled on large repeat units of CNVs containing Olfactory repeat (OR) receptors and retroviral elements of REPP (for repeat proximal) and REPD (repeat distal) regions which are located at 8p23.1 on chromosome band (Hollox et al., 2008). Many beta-defensin genes are embedded on large segmental repeat regions which are varied in copy number of 260kb (Hollox et al., 2012) (see figure 1.3). Remarkable structural diversity is located at the subtelomeric regions. In this case, subtelomeric duplication at the beta-defensin repeat region sponsors large-scale rearrangement such as segmental duplication and polymorphic inversion (Mefford & Eichler, 2009).

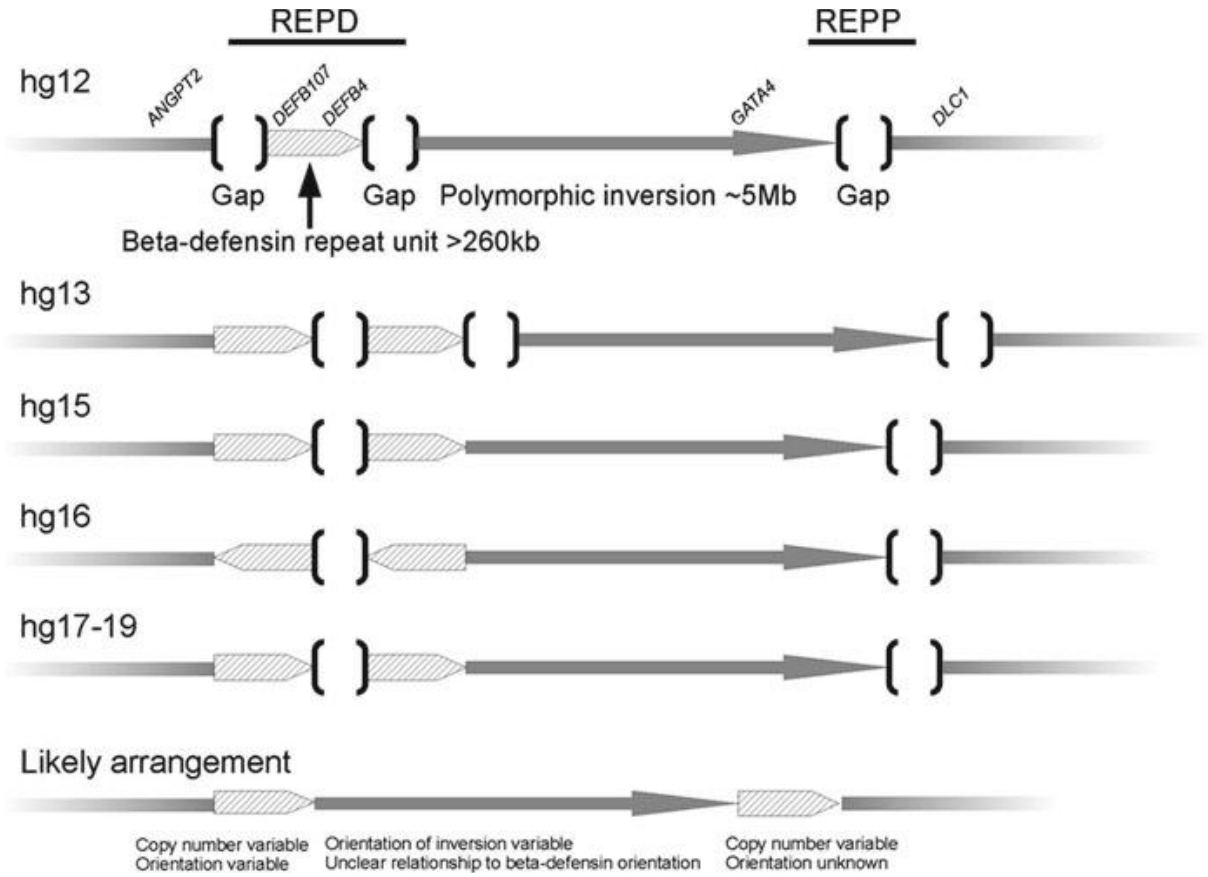


Figure 1.3: The beta-defensin region at 8p23.1 in different reference assemblies of human genome. Reference assemblies (hg12-19) taken from UCSC genome browser (http://genome.ucsc.edu).

In identifying and characterising the CNVs of rhesus macaque genome gives the information to understand the relationship among CNVs and disease susceptibility (Lee *et al.*, 2008). Macaque CNVs of the beta-defensin gene cluster, containing the *DEFB4* gene, are relevant to human CNVs that have been implicated in susceptibility to Crohn's disease, an inflammatory bowel disease of the gastrointestinal system (Fellermann *et al.*, 2006) (see figure 1.4). In this case, the rhesus macaque, as a proper model organism, contributes towards understanding the possible correlation between beta-defensin gene dosage and microbial influences (Lee *et al.*, 2008).

Macaque CNV locus	Macaque CNV obs.	Human CNV obs.	Genes of interest	Phenotypic effect of CNV
Chr. 3: 180.2–180.3	1	28	*PRSS1*	Hereditary pancreatitis susceptibility
Chr. 4: 29.5–29.6	6	21	*LOC347981*	Correlated with gene expression level
Chr. 4: 31.0–31.1	1	20	*LOC282956*	Correlated with gene expression level
Chr. 4: 32.1–32.6	7	231	*HLA-DRB5, HLA DQA1, HLA-DQA2*	Correlated with gene expression level
Chr. 8: 8.0–8.7	6	124	*DEFB4*	Correlated with gene expression level; Crohn's disease susceptibility; psoriasis susceptibility
Chr. 10: 86.5–86.5	1	25	*CGI-96*	Correlated with gene expression level
Chr. 16: 55.9–55.9	2	207	*HDAC5, MGC3130*	Correlated with gene expression level
Chr. 19: 47.2–47.4	9	3	*CYP2A6*	Correlated with gene expression level; correlated with protein level; lung cancer susceptibility

Table 1.2: **The functional effects of CNVs in both macaque and human** (Lee *et al*., 2008)

Rhesus macaque genome

The rhesus macaque (*Macaca mulatta*) is a species from Indian origin. The genetical and physiological features of rhesus macaque are similar to humans and used in biomedical research mainly in the field of physiology, behavioural biology, cardiological disorders and infectious disease (Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.*, 2007). Hence, the human genome assembly used to compare with rhesus macaque assembly to correct the errors (Zhang *et al.*, 2012). The nucleotide sequence alignment between human and rhesus genome shows 90.76% identity on average by excluding insertion and deletions. The orthologs of human and rhesus have 97.5% similarity on both the levels of nucleotide and amino acid sequence (Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.*, 2007), Hence, the rhesus macaque are very essential as model organism for *human immunodeficiency virus* (*HIV*) and acquired immunodeficiency syndrome (AIDS) of their susceptibility (Lee *et al.*, 2008).

The assembly of Rhesus macaque genome by whole genome shotgun (WGS) approach exits many gaps (Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.*, 2007), (Lee *et al.*, 2012). WGS sequencing approach begin with fragmentation of DNA into small fragments of different sizes, these fragments are cloned and generate reads from both ends of the sequence.

These reads are collected from both the ends and assembled them into larger contigs based on overlapping DNA fragments by computational approach (Brown TA *et al.,* 2002). This approach poses problem in assembling random reads due to repetitive elements (Brown TA *et al.,* 2002). Most of the gaps occur in genome assembly due to repetitive elements or insufficient reads (Batzoglou *et al.,* 2002). The isolation of individual BAC clones and sequence complete genome by finishing that improves the quality of data (Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.,* 2007), (Lee *et al.,* 2012).

1.3 Large Scale genome sequencing with bacterial artificial chromosome (BAC) insert clones

The large scale genome sequencing was introduced by Human Genome Project in the year 1990 (Gardiner, 2002), that require bacterial based large insert clones for sequencing (BAC as tools for genome sequencing). The clone-by-clone shotgun (CBCS) and whole-genome shotgun (WGS) are the two techniques used for large scale genome sequencing. Large insert DNA clones provide unit by unit sequencing, which accurately enables the assembly of complex genome contigs, which have long-range repetitive regions such as beta- defensin gene clusters (Adams *et al*., 2000). Moreover, to maintain DNA for a long period, large insert DNA clones are necessary, because entire genome sequences can be covered by inserting large DNA clones instead of a large number of small insert clones (Zhang & Wu, 2001).

Bacterial artificial chromosome (BAC) is an plasmid F factor based on *Escherichia coli,* used to construct DNA libraries of complex genome which is capable to maintain DNA fragments (>300 kilo base pairs) of human genome (Shizuya *et al.,* 1992). The F plasmids have potential to reduce recombination among DNA fragments due to low copy number. The bacterial DNA with F factors is capable to maintain and clone large DNA fragments (Shizuya *et al.,* 1992). The BAC with F factor is circular DNA which prevents shearing during purification (Zhang & Wu, 2001), (Shizuya *et al.,* 1992). Due to these advantages, in recent years, BAC have become the main element for large-scale genome sequencing, physical mapping and gene cloning. BAC based physical mapping has an essential role to play in large-scale genome sequencing to select minimal overlapping clones and building clone path for sequencing genome and sequence

assembly (Zhang & Wu, 2001).

The aims and objectives

The aim of this project is to examine the sequence results generated by new technologies for both the single base and the molecular level of complex genomes, also to assemble a contiguous BAC sequence across the macaque beta-defensin region, and perform a preliminary analysis of sequence variation. The study includes following techniques, isolation of BAC DNA; predicting restriction enzyme maps using bioinformatics tools such as "NEBcutter" and "restrictionmapper"; analysing restriction fragment digestion pattern using both normal agarose gel (for small piece of DNA fragments) and pulsed-field gel electrophoresis (for larger DNA fragments); Sanger and single molecule real time sequencing qualities will be compared in particular regions; analysis of gene structure and content by using bioinformatics databases.

Chapter 2: MATERIALS and METHODS

2.1. BAC clones

Three large BAC clones, BAC 201P10 (96.7 kb); BAC 246K23 (169 kb); BAC 148I5 (191 kb), were provided from CHORI-250 BAC library and screened by PhD student Barbara Ottolini using PCR probes from DEFB2L gene.

2.2. Growing and isolation of BAC DNA

BAC DNA samples were isolated with two different purification methods: CsCl purification and Macherey-Nagel, Nucleobond BAC 100 plasmid purification methods. Samples purified by CsCl isopycnic centrifugation were obtained from PhD student Barbara Ottolini.

BAC colonies from glycerol stock were incubated into 10 ml Luria Bertani (LB) medium supplemented with appropriate antibiotic (chloramphenicol at 12.5 μg/ml) at 37 °C with 200-300 rpm shaking overnight. The concentrations of 10X diluted cultures were measured with spectrometer. After OD reached 0.2, cultures were used to DNA isolation with "Macherey-Nagel Low Copy Plasmid Purification Maxi Kit" (see protocol 1 at appendix).

2.3. Analysis of restriction digestion pattern

"Nebcutter V2.0" and "restrictionmapper" bioinformatics tools were used to form predicted restriction enzyme maps for each BAC DNA sequence. According to these maps, restriction enzymes, available from departmental stocks, which cut BAC DNAs up to 25 fragments were chosen to set up digestion reactions.

2.3.1. Restriction enzyme digestion reaction

Reactions were set up to confirm predicted restriction digestion maps of BAC DNAs. 500-1000 ng BAC DNA, 10X restriction enzyme buffer (NEB), 100X BSA (if required), 5 units restriction enzyme and distilled water were used to set up total 50 μl digestion reaction. All components

were pipetted into a PCR tube and then incubated overnight at 37 °C.

2.3.2. Agarose gel electrophoresis

Agarose gel electrophoresis was used to visualise fragmentation of BAC DNAs after digestion reactions. Approximately 1-20 kb fragments was visualised by agarose gel electrophoresis. A 0.8% agarose gel was prepared with 1X TBE buffer. The digestion samples in different dilution (20 µl sample, 15 µl sample +5 µl NEB buffer, 10 µl sample +10 µl NEB buffer, 5 µl sample +15 µl NEB buffer) were prepared and 4 µl 6X loading buffer was added. The samples and 24 kb max ladder were loaded into the gel and run at 150 V for 6 hours in the electrophoresis tank containing 1X TBE buffer. Then, ethidium bromide staining was carried out.

2.3.3. Pulsed-field gel electrophoresis (PFGE)

Agarose gel electrophoresis cannot resolve large DNA fragments which are more than 20 kb because of using single electrical field. In this case, pulsed-field gel electrophoresis takes an advantage providing conventional direction of electric-field for more than 20 kb fragments (Shaffer&Lupski 2000). A 1% gel was prepared with 100 ml 0.5X TBE buffer and 1 g agarose (special for pulsed-field gel). 2000 ml 0.5X TBE was poured into the CHEF-DR III chamber. The cooling module was set up at 14 °C. After samples and pulsed-field gel (IPG) ladder were loaded, the run parameters were entered as initial switch time: 1 and final switch time: 25, then the gel was run 16 hours at 120 V. Then, ethidium bromide staining was carried out.

2.3.4. Ethidium bromide staining

Ethidium ions are positively charged, therefore if EtBr is in gel the Et^+ ions migrate in the opposite direction to DNA – can result in weak staining in long gel runs. Because of this drawback, post run ethidium bromide staining was carried out. After electrophoresis, the gel was stained in 500 µl, 1X TBE buffer with 50 µl EtBr for 30 minutes on a shaking incubator and then cleaned with water for 10 minutes. Finally, the gel was visualised under UV light.

2.3.5. Calculation of fragment sizes

Each fragment distance of marker was measured to draw graph and the reference formula was obtained to calculate restriction enzyme fragments (see example of a graph at appendix).

2.4. Analysis of gene structure and content

2.4.1 BAC end sequencing

End sequencing was carried out to identify sequences at the two ends of Rhesus genomic DNA inserted into BAC clones. The general map of BAC vector (pTARBAC2.1) was taken from BACPAC Resources Centre (BPRC) to determine primers used for end sequencing.

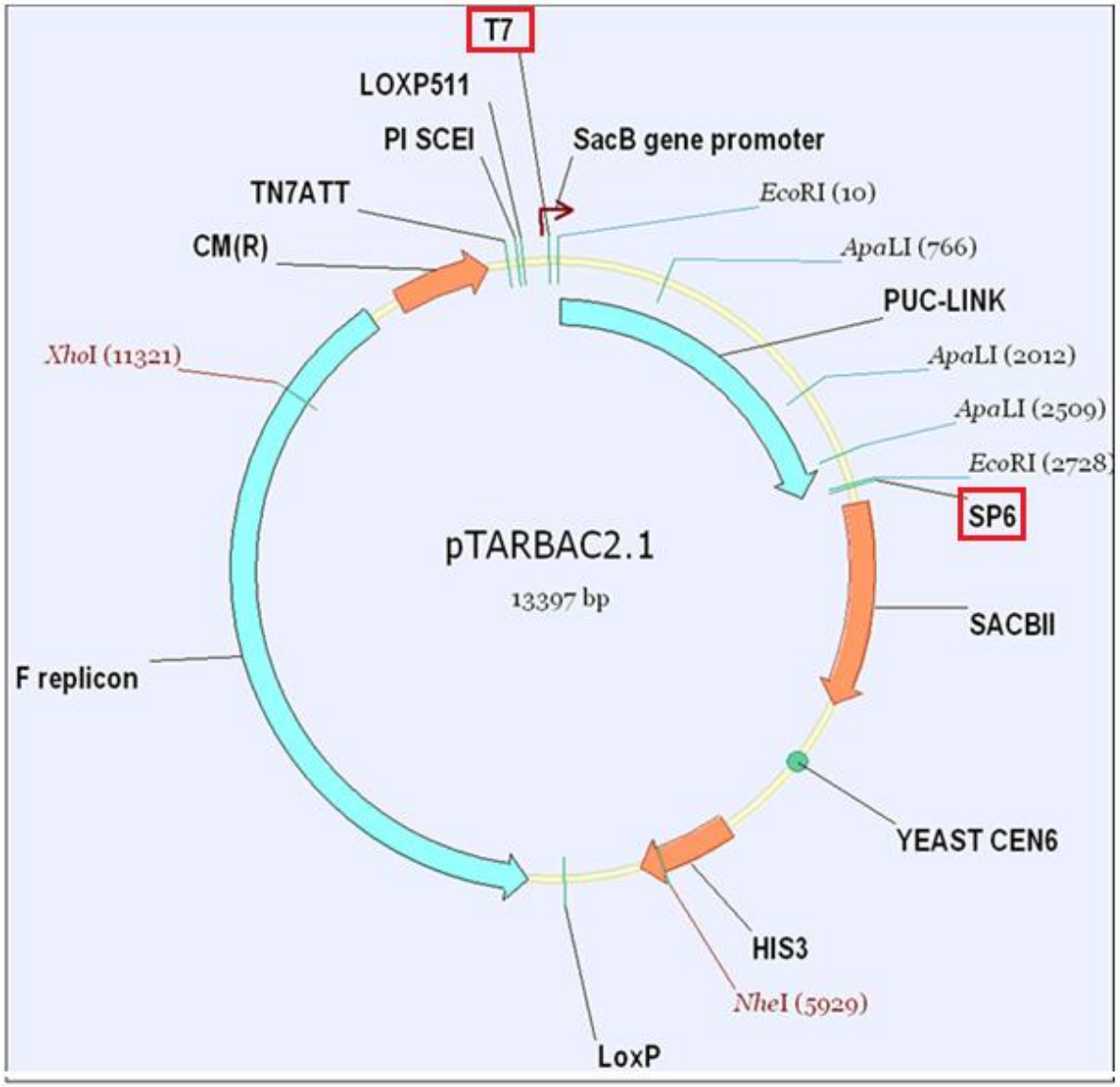


Figure 2.1: The map of pTARBAC2.1 vector adapted from bacpac.chori.org

SP6 (5'-3')	T7 (5'-3')	Tm (°C)	Cycles
ATTTAGGTGACACTATG	TAATACGACTCACTATAGG	50	25

Table 2.1: Primers for BAC end-sequencing

According to the table 2.1, SP6 and T7 primers were used to carry out end sequencing. The protocol of "high throughput direct end sequencing of BAC clones" (Kelley J. M. et al., 1999) was followed. Per reaction was containing 12 µl of BigDye Termterminator mix (+4 mM extra $MgCl_2$) , 1 µl of 20 pmoles primer, 1 µg BAC DNA and distilled water to made 30 µl of total volume. The PCR reaction was set up for initial denaturation of 96 °C for 2 minutes, denaturation of 95 °C for 10 seconds, followed by an annealing temperature of 50°C for 5 seconds, an extention of 60°C for 4 minutes, final extention of 72°C for 10 minutes and holding at 4 °C. This cycling protocol was used on Applied Biosystem Veriti 96 well Thermal Cycler. After BigDye reaction, Edge Biosystems Performa DTR (Dye Terminator Removal) gel filtration columns were used to clean up reaction from excess dyes. (see appendix for protocol2). The Protein and Nucleic Acid Chemistry Laboratory (PNACL) at University of Leicester analysed the samples using ABI 3730 DNA sequencer.

2.4.2. Analysis of beta-defensin gene cluster

2.4.2.1. DNA samples of rhesus macaque

Sixteen rhesus genomic DNA samples were obtained from University of California, Saint Davis.

The Sanger sequencing results of beta-defensin genes were blast with TGS result of each BAC DNA to determine base changes, deletions or insertions.

2.4.2.2 PCR

Rhesus macaque and BAC 246K23 samples were amplified for a individual SNP at the *DEFB104* gene and a deletion at DEFB107 gene. Each PCR reaction comprised of 0.5 µl of 10 ng DNA template, 0.6 µl of 10X buffer, 0.3 µl of 2.5 mM dNTPs, 0.3 µl of 10mM forward and reverse primers, 0.06 µl of 5 u/µl Taq DNA polymerase and distilled water to complete 6 µl of total volume. The PCR reaction was carried out 30-35 cycles for initial denaturation of 95 °C for 30 seconds, denaturation of 95 °C for 30 seconds, followed by an annealing temperature of 62-67°C for 30 seconds, an extention of 70°C for 30 seconds and final extention of 72°C for 5 minutes. Applied Biosystems, Veriti 96 well Thermal Cycler was used to run PCRs. The table 2.1 and 2.2 demonstrate each primer and PCR conditions.

Forward Primer	Reverse Primer	Tm (°C)	Cycles
TTTGAATTGGACAGAATA TGTGG	ACTGAATCGTACAAAACCC TGA	62	35

Table 2.2: Primers for SNP at *DEFB104* gene

Forward Primer	Reverse Primer	Tm (°C)	Cycles
AGGGTATCTCCTTGTAGCATTGG	GAATTTGGCCTGGGCAAT	67	35

Table 2.3: Primers for deletion at DEFB107 gene

2.4.2.3 Restriction enzyme digestion

PCR products were digested with *Hpy188III* to detect SNP at *DEFB104*. 18 µl restriction reaction mix containing 1.8 µl of 10X NEB buffer, 0.2 µl of BSA (if required), 1-2 unit enzyme, 6 µl PCR product and distilled water, was set up and run in MJ Research thermo cycler at 37°C overnight.

2.4.2.4 Agarose gel electrophoresis

To visualise SNP at *DEFB104* agarose gel electrophoresis was used after PCR and digestion reactions. A 3% agarose gel was prepared with 1X TBE buffer containing 5 µl/100ml ethidium bromide. 18 µl samples with 3.6 µl 6X dye were loaded into the gel and run at 120 V for 1.5 hours in the electrophoresis tank containing 1X TBE buffer. Then, the gel was visualised under UV light.

2.4.2.5 DNA sequencing

DNA sequencing reactions were carried out to identify some deletions at DEFB107 gene. After PCR reactions and analysis of agarose gel electrophoresis, DNA products were excised from gel and purified using QIAQuick Gel Extraction Kit (see protocol 3 at appendix). Purified DNAs were used to set up sequencing reaction. Per reaction was containing 2 µl of Big Dye Terminator buffer (+4 mM extra $MgCl_2$) , 1 µl of Big Dye Terminator Ready reaction mix, 1 µl of 3.2 µM primer , 2-4 µl of 3-10 ng DNA and distilled water to made 10 µl of total volume .

The sequencing reactions were carried out for initial denaturation of 96 °C for 1 minute, denaturation of 96 °C for 10 seconds, followed by an annealing temperature of 50°C for 5 seconds, an extention of 60°C for 4 minutes and holding at 4 °C. This cycling protocol was used on Applied Biosystem Veriti 96 well Thermal Cycler. After BigDye reaction, Edge Biosystems Performa DTR (Dye Terminator Removal) gel filtration columns were used to clean up reaction from excess dyes. (see protocol 2 at appendix). The Protein and Nucleic Acid Chemistry Laboratory (PNACL) at University of Leicester analysed the samples using ABI 3730 DNA sequencer.

Chapter 3: RESULTS

3.1. Isolation of BAC DNA

BAC DNA samples were isolated to analyse restriction digestion maps and gene contents. It was able to provide an option to test TGS data from PasBio assembly. They were obtained by two different purification methods, CsCl isopycnic centrifugation and Macherey-Nagel, Nucleobond BAC 100 plasmid purification kit.

BAC DNAs	Concentration of CsCl isopynic centrifugation (ng/µl)	Concentration of plasmid purification kit (ng/µl)
BAC 201P10	240	547
BAC 246K23	286	495
BAC 148I5	145	139

Table 3.1: Concentrations of BAC DNAs after different isolation methods.

Table 3.1 showed the results of concentrations of BAC DNAs by isopycnic centrifugation and plasmid purification kit after isolation. According to these results, the concentrations of CsCL BAC DNAs were lower compared to plasmid purification kit.

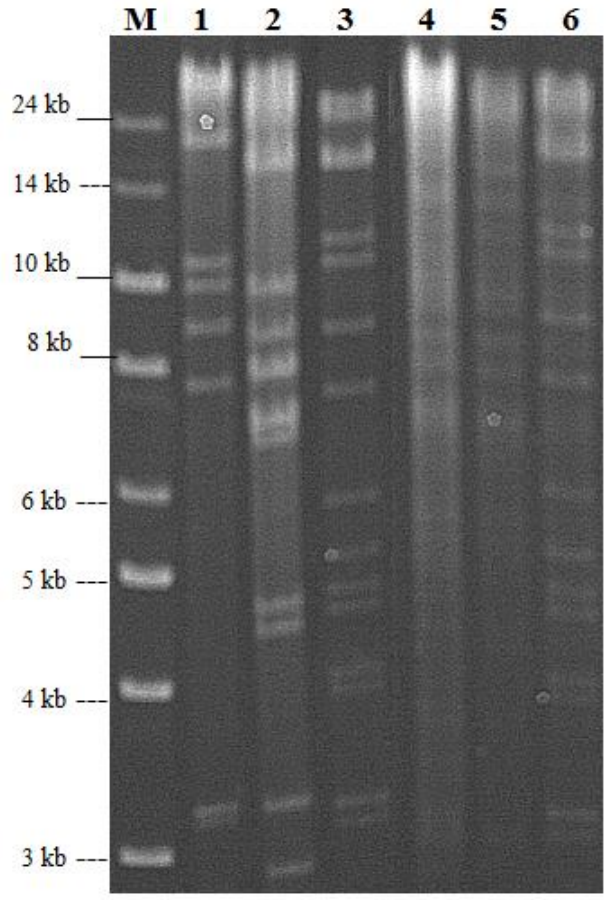


Figure 3.1 Agarose gel electrophoresis image of BAC DNAs digested with KpnI restriction enzyme. The lane M represents 24 kb max DNA ladder; (1) BAC 201P10 DNA; (2) BAC 246K23 DNA; (3) BAC 148I5 DNA purified with CsCl isopycnic centrifugation; (4) BAC 201P10 DNA; (5) BAC 246K23 DNA; (6) BAC 148I5 DNA isolated with plasmid purification kit.

Figure 3.1 is the image with the result of agarose gel electrophoresis of BAC DNAs. The left side of the image represent the molecular weight of DNA in kilobase maximum up to 24 kb. The upper side of the image represents BAC DNA samples, these include BAC 201P10 (1), BAC 246K23 (2) and BAC 148I5 (3) DNA were purified by CsCl isopycnic centrifugation, whereas, BAC 201P10 (4), BAC 246K23 (5) and BAC 148I5 (6) isolated with plasmid purification kit. In the above image of agarose gel electrophoresis, the bands are clearly observed in 1, 2 and 3 lanes which correspond to CsCl isopycnic centrifugation results. Although, there are few bands which observed in lane 6, the other two lanes does not appear any bands. According to these results, the fragments of CsCl BAC DNAs were clearly observed compare to plasmid purification BAC DNAs. As a result, the high concentration of BAC DNAs isolated by plasmid purification kit

indicates the presence of chromosomal DNA which is not required. Therefore, CsCl BAC DNAs were used to analyse restriction digestion pattern, which is requires high quality of DNA, while plasmid purification BAC DNAs were used for PCR reactions which do not require pure DNA.

3.2. Analysis of restriction digestion pattern

Each BAC DNA was digested with different enzyme to analyse restriction maps. The descriptive statistical analysis were done on excel sheet to calculate fragments which are obtained from each gel and compared with expected fragments depending on TGS assemblies. The maps of expected fragments were formed by using NEBcutter V2.0 online tool. Small fragments, approximately less than 1.5 kb, could not be observed on the gel because they were run away after long running time. Hence, they were not considered for calculations.

3.2.1. BAC 201P10 DNA

BAC 201P10 DNA, which is 96.710 kb, digested with *KpnI* and *NotI* restriction enzymes.

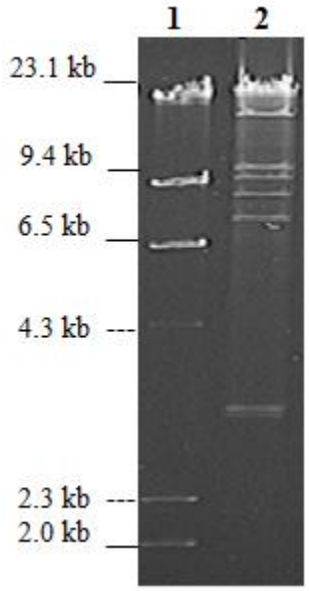


Figure 3.2a: Agarose gel electrophoresis image for BAC 201P10 DNA sample. (1) λ *HindIII* marker; (2) BAC 201P10 DNA digested with *KpnI* restriction enzyme.

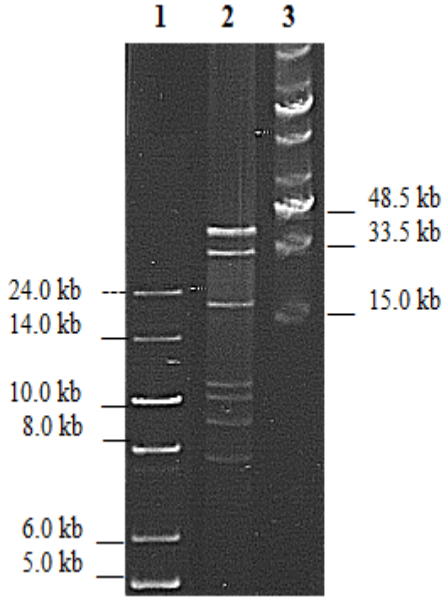


Figure 3.2b: Pulsed-field gel electrophoresis image for BAC 201P10 DNA sample. (1) 24 kb max DNA marker; (2) BAC 201P10 DNA digested with *KpnI* restriction enzyme; (3) mid-range I PFG ladder.

The Figure 3.2a is the image of agarose gel electrophoresis with (1) λ *HindIII* marker and (2) BAC 201P10 DNA. The outcome of this gel was clearly shown that the first two bands are not resolved properly. The fragments were calculated by the reference formula and more than 20 kb fragments are not clearly observed in the gel. The pulsed field gel electrophoresis (PFGE) were preferred to run the BAC 201P10 DNA sample with *KpnI* restriction enzyme and result with expectation fragments was shown in the figure 3.2b. The gel was run with two markers and one BAC sample and the lanes in the gel was represented as (1) 24 kb max DNA marker, (2) BAC 201P10 DNA sample with *KpnI* restriction enzyme and (3) mid range I PFG marker. If pulsed-field gel result was compared with agarose gel result, it can be observed that fragments which are more than 20 kb were better resolved on pulsed-field gel. The largest two fragments, 44 kb and 32 kb respectively, can be obviously seen on the pulsed-field gel, whereas these two fragments

cannot be determined on the agarose gel. After digestion with *KpnI*, it would be formed 7 fragments but we obtained 9 fragments. This showed that there should be 2 extra fragments.

S.No	pfg size (kb)	expected size (kb)
1	46.7	44.715
2	**32.0**	**extra band**
3	18.6	18.583
4	**12.4**	**extra band**
5	11.5	11.062
6	9.9	8.903
6	8.6	6.898
7	3.3	3.142
8	3.1	3.086
9		0.321

Table 3.2: The size comparison among pulsed-field gel fragments and expected fragments of BAC 201P10 digested with *KpnI*

Table 3.2 showed the sizes in kb comparing among pulsed-field gel fragments and expected fragments. According to this table, in the pulsed-field gel most of the fragmentation was come with expected result except two extra fragments which sizes were 32 kb and 12 kb respectively.

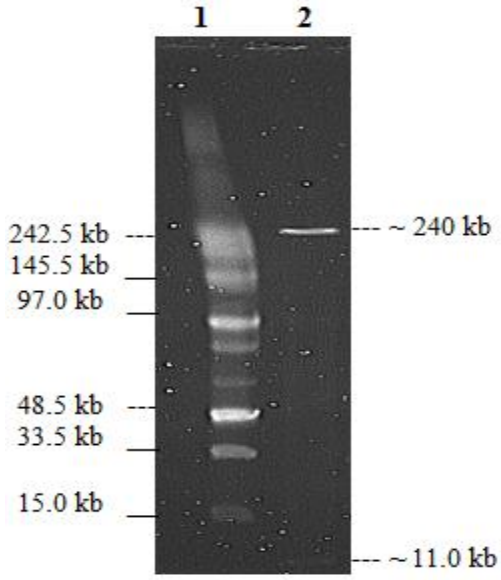


Figure 3.3: Pulsed-field gel electrophoresis image for BAC 201P10 DNA digested with *NotI* restriction enzyme. (1) Mid-range I PFG DNA marker**;** (2) BAC 201P10 DNA digested with *NotI* restriction enzyme**.**

S.No	pfg size (kb)	expected size (kb)
1	240.0	96.710
2	11.0	-

Table 3.3: The size comparison among pulsed-field gel fragments and expected fragments

The figure 3.3 is an pulsed-field gel electrophoresis image for BAC 201P10 DNA digested with *NotI* restriction enzyme. The two lanes of the gel are (1) Mid-range I PFG DNA marker and (2) BAC 201P10 DNA digested with *NotI* restriction enzyme. The left and right side of the image was mentioned the DNA in kb. The image was clearly observed and identified two bands, one with ~240 kb and the other with ~11.0kb. However, the expected fragment was one with 96.76 kb (seen in table 3.3).

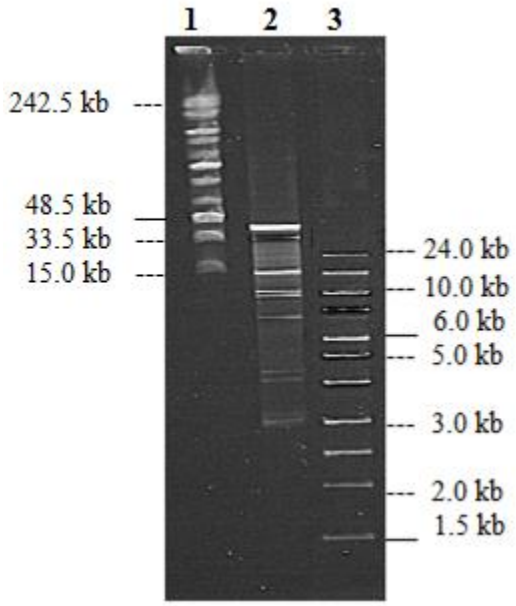
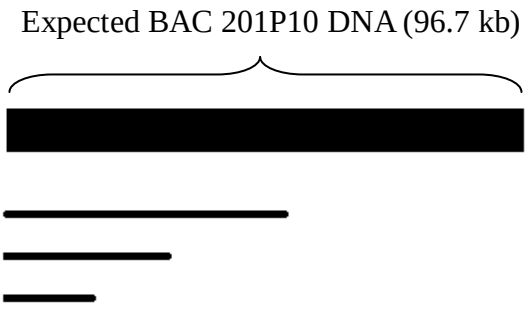


Figure 3.4: Pulsed-field gel electrophoresis image for BAC 201P10 DNA double digested with *KpnI* and *NotI* restriction enzymes. (1) Mid-range I PFG DNA marker; (2) BAC 201P10 DNA digested with *KpnI* and *NotI* restriction enzymes; (3) 24kb max DNA ladder.

S.No	pfg size (kb)	expected size (kb)
1	45.7	44.715
2	**31.7**	**extra band**
3	16.0	18.583
4	**11.9**	**extra band**
5	10.8	11.062
6	8.3	6.898
6	4.7	4.590
7	4.3	4.313
8	3.1	3.142
9	2.9	3.086
10		0.321

Table 3.4: The size comparison between pulsed-field gel fragments and expected fragments for BAC 201P10 after *KpnI* and *NotI* double digestion

The figure 3.4 is an image of pulsed-field gel electrophoresis for BAC 201P10 DNA double digested with Kpn and NotI restriction enzymes. The gel have three lanes, first lane with mid-range I PFG DNA marker, second lane is BAC 201P10 DNA digested with *KpnI* and *NotI* restriction enzymes and third lane with 24kb max DNA ladder. 8.903 kb fragment seen on the table 3.2 was digested with *NotI* and 4.590 kb and 4.313 kb fragments seen on the table 3.3 were obtained. By analysing images of *KpnI-NotI* double digestion and *NotI* single digestion, it can be seen that after *KpnI-NotI* double digestion, *NotI* cut BAC 201P10 DNA on only one region as expected. However, after single digestion with NotI, two fragments were formed and one of them was bigger than double BAC 201P10 DNA size. These two different results provided a crucial advantage to understand the structure of BAC 201P10 DNA.



(a)

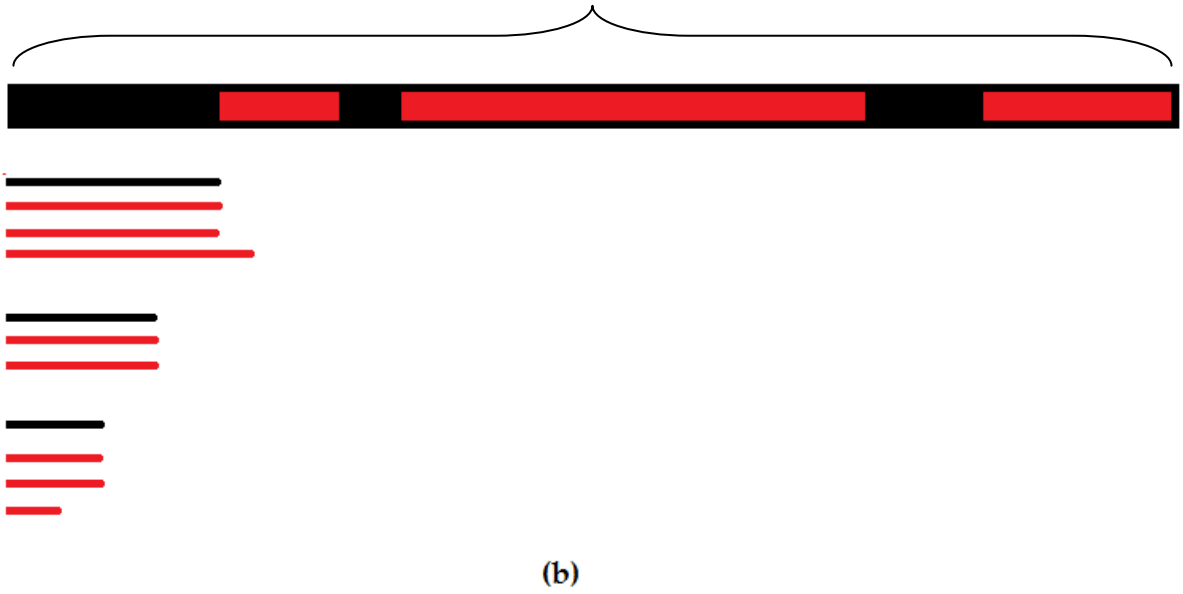


Figure 3.5: Two alternative diagrams for BAC 201P10 DNA structure. (a) Expected structure and fragments; (b) Actual structure and fragments. Red boxes and fragments refer repetitive regions.

The above two figures of 3.5a and b were drawn manually based on the prediction of all digestions which include *KpnI*, *NotI and KpnI-NotI* digestions depending on TGS assembly. In all these digestions, the expected whole BAC 201P10 DNA size was 96.7kb. However, the actual size of BAC 201P10 DNA was nearly 250 kb. In the figure 3.5b, the red lines indicate repetitive regions which were obtained by *NotI* digestion reaction, According to these results, it conclude, TGS assembly was not exactly matched with the real BAC 201P10 DNA that contain some insertions and repetitive regions.

3.2.2 BAC 246K23 DNA

BAC 246K23 DNA, which is 169 kb, digested with *SalI, PmeI, PvuI, EcoRV* and *PshaI* restriction enzymes.

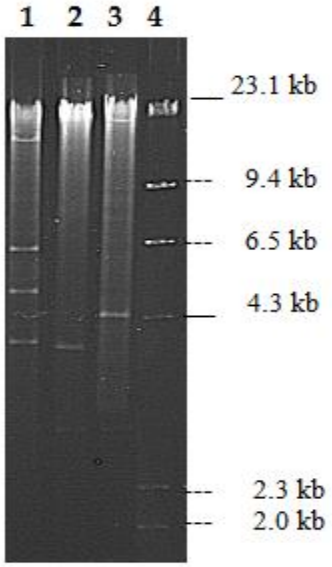


Figure 3.6: Agarose gel electrophoresis image for BAC 246K23 DNA digested with *SalI, PmeI and PvuI* restriction enzymes. (1) BAC 246K23 DNA digested with *SalI* restriction enzyme; (2) BAC 246K23 DNA digested with *PmeI* restriction enzyme (3) BAC 246K23 DNA digested with *PvuI* restriction enzyme (4) λ *HindIII* marker.

Figure 3.6 is an normal agarose gel electrophoresis image for BAC 246K23 DNA digested with *SalI, PmeI* and *PvuI* restriction enzymes corresponds to first, second and third lanes, the fourth lane represented as λ *HindIII* marker. The outcome of the gel was able to predict small fragments that can be seen clearly in the first three lanes, whereas, the large fragments were unable to predict due to unresolved fragments. Hence, there was a drawback to derive expected fragments, thus, the pulsed-field gel electrophoresis technique was carried out further to predict large fragments.

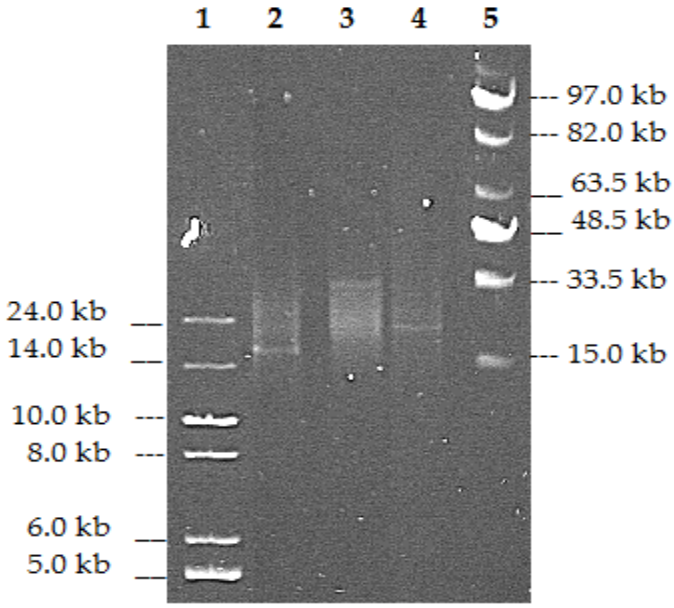


Figure 3.7: Pulsed-field gel electrophoresis image for BAC 246K23 DNA digested with *SalI, PmeI and PvuI* restriction enzymes. (1) 24 kb max DNA ladder; (2) BAC 246K23 DNA digested with *SalI* restriction enzyme; (3) BAC 246K23 DNA digested with *PmeI* restriction enzyme; (4) BAC 246K23 DNA digested with *PvuI* restriction enzyme; (5) Mid-range I PFG ladder

Figure 3.7 is an pulsed-field gel electrophoresis image for BAC 246K23 DNA digested with *SalI, PmeI* and *PvuI* restriction enzymes that corresponding to second, third and fourth lanes. The first and fifth lanes are represented as 24 kb max DNA ladder and mid range I PFG ladder respectively. Although the pulsed-field gel electrophoresis was carried out to predict large fragments but the outcome of the gel was not clear to calculate fragment sizes. Thus, only normal gel results were used to form table 3.5.

predicted size (kb)	expected size (kb)
	87.680
	51.204
	14.800
7.6	6.384
6.1	5.031
4.5	3.906

(a) *SalI*

predicted size (kb)	expected size (kb)
	58.009
	31.976
	29.783
	26.182
	19.229
4.1	3.826

(b) *PmeI*

predicted size (kb)	expected size (kb)
	63.892
	51.793
	48.868
2.3	4.506

(c) *PvuI*

Table 3.5 : The units sizes of predicted and expected fragments, and represented the variations among three tables. (a) *SalI*; (b) *PmeI*; (c) *PvuI*

The sub-table (a), (b) and (c) from the above table 3.5 corresponds to BAC 246K23 DNA digested with *SalI, PmeI* and *PvuI* restriction enzymes respectively, shows sizes of predicted and expected fragments and compared with all three tables. According to these three tables, it was clearly observed that the predicted fragments are small in all three digestions and does not contain large fragments which was expected.

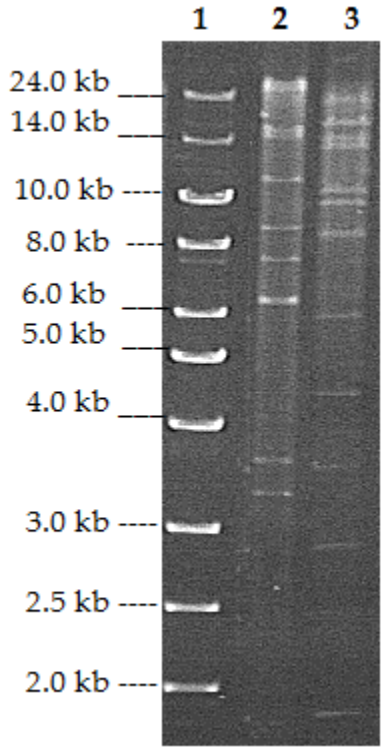


Figure 3.8: Agarose gel electrophoresis image for BAC 246K23 DNA digested with *EcoRV and PshAI* restriction enzymes. (1) 24 kb max DNA ladder; (2) BAC 246K23 DNA digested with *EcoRV* restriction enzyme; (3) BAC 246K23 DNA digested with *PshAI* restriction enzyme.

The figure 3.8 is an agarose gel electrophoresis image for BAC 246K23 DNA digested with *EcoRV* and *PshAI* restriction enzymes, which are corresponding to second and third lanes, the first lane was represented as 24 kb max DNA ladder.

S.No	predicted size (kb)	expected size(kb)
1	20.9	32.624
2	18.8	27.416
3	15.3	21.990
4	14.6	19.744
5	12.1	10.654
6	9.6	10.044
7	8.2	9.639
8	7.0	9.070
9	3.7	8.925
10	3.2	6.271
11		6.232
12		4.293
13		3.575
14		3.252
15		0.705
16		0.342
17		0.229

(a) *EcoRV*

S.No	predicted size (kb)	expected size (kb)
1	18.8	45.808
2	16.9	23.437
3	15.3	21.777
4	14.6	14.739
5	**missing**	**13.392**
6	11.5	10.610
7	10.5	10.078
8	9.2	8.617
9	6.4	5.939
10	4.8	4.517
11	3.7	3.663
12	2.7	2.880
13	1.7	1.901
14		0.819
15		0.325
16		0.116
17		0.086
18		0.086
19		0.043
20		0.043
21		0.043
22		0.043
23		0.043

(b) *PshAI*

Table 3.6a and b: BAC 246K23 DNA digested with *EcoRV and PshAI* restriction enzymes and their calculation in unit sizes of predicted and expected fragments. (a) *EcoRV*; (b) *PshAI*

In the table 3.6a, the agarose gel of BAC 246K23 DNA digested with *EcoRV* restriction enzyme was able to unpredicted 10 fragments out of 17 expected fragments, whereas, in table 3.6b, the BAC DNA digested with *PshAI* restriction enzyme predicted 13 fragments out of 23 expected fragments. Hence, all expected fragments were not obtained with *EcoRV and PshAI* digestions and predicted fragment sizes were not large as expected sizes. Overall analysis of all digestion reactions, it was determined that the real structure of BAC 246K23 DNA was not matched the expected structure which depends on TGS assembly.

Ends	Coordinates	Length (bp)
PshAI-PshAI	135037-135079	43
“	135080-135122	43
“	135123-135165	43
“	135166-135208	43
“	135209-135251	43
“	135252-135337	86
“	135338-135423	86

Table 3.7: Recurrent fragments in BAC 246K23 DNA digested with *PshAI*

In the table 3.7, the recurrent fragments of 43bp and 86bp digested with *PshAI* might refer repetitive regions in TGS of BAC 246K23 genome assembly. However, it could not be provided an alternative structure due to the presence of inconsistent fragments after each digestion.

3.2.3. BAC 148I5 DNA

BAC 148I5 DNA, which is 191.2 kb, digested with *SalI, PmeI* and *PvuI* restriction enzymes.

The below figure 3.9 is an agarose gel electrophoresis image for BAC 148I5 DNA digested with restriction enzymes. The gel was run with two restriction enzymes *SalI*, and *PmeI*, which are corresponding to second and third lane. The outcome of the result was small fragments compare to expected fragments.

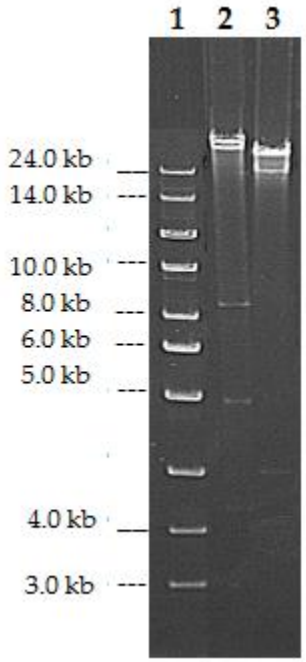


Figure 3.9: Agarose gel electrophoresis image for BAC 148I5 DNA digested with *SalI, and PmeI* restriction enzymes separately. (1) 24 kb max DNA ladder; (2) BAC 145I5 DNA digested with *SalI* restriction enzyme; (3) BAC 145I5 DNA digested with *PmeI* restriction enzyme.

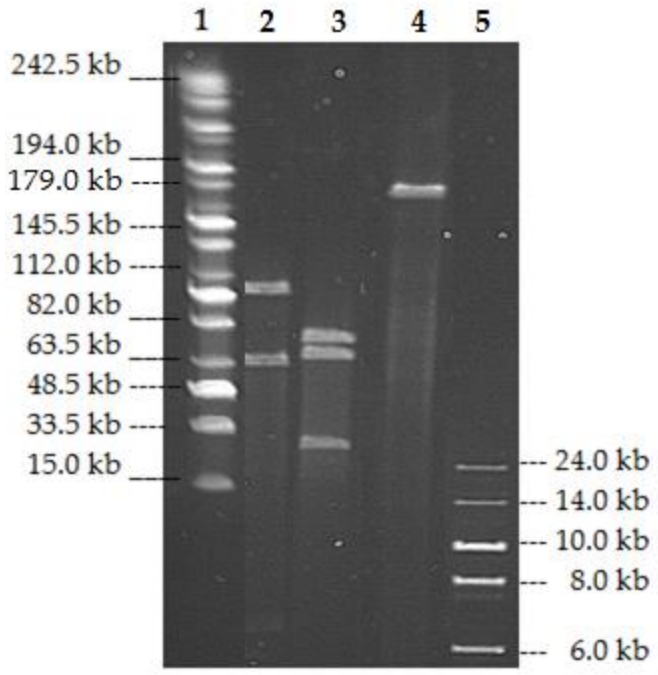


Figure 3.10: Agarose gel electrophoresis image for BAC 148I5 DNA digested with *SalI, and PmeI* restriction enzymes separately. (1) 24 kb max DNA ladder; (2) BAC 145I5 DNA digested with *SalI* restriction enzyme; (3) BAC 145I5 DNA digested with *PmeI* restriction enzyme.

The figure 3.10 is image of pulsed-field gel electrophoresis for BAC 148I5 DNA digested with *SalI*, *PmeI* and *PvuI* restriction enzymes. The out come of the gel with prediction of large fragments that was matched to the expected fragments. The results of normal gel was used to calculate small fragments sizes (<20kb), whereas, the pulsed-field electrophoresis gel results with prediction of large fragment sizes (>20kb).

ng size (kb)	pfg size (kb)	expected size (kb)
23.6	105.6	106.486
22.3	67.0	66.159
7.0		6.392
missing		**5.594**
4.2		3.921
2.5		2.657

(a)

S.No	Ends	Coordinates	Length (bp)
1	*SalI-SalI*	1991-68149	66.159
2	*SalI-SalI*	68150-72070	3.921
3	*SalI-SalI*	72071-78462	6.392
4	*SalI-SalI*	78463-81119	2.657
5	*SalI-SalI*	81120-187605	106.486
6	*SalI-SalI*	187606-1990	5.594

(b)

ng size (kb)	pfg size (kb)	expected size (kb)
22.3	79.09	79.555
19.9	67.09	71.870
17.9	27.89	31.046
missing		**5.591**
2.9		3.147

(c)

S.No	Ends	Coordinates	Length (bp)
1	*PmeI-PmeI*	2646-33691	31046
2	*PmeI-PmeI*	33692-36838	3147
3	*PmeI-PmeI*	36839-108708	71870
4	*PmeI-PmeI*	108709-188263	79555
5	*PmeI-PmeI*	188264-2645	5591

(d)

pfg size (kb)	expected size (kb)
178.65	186686
	4523

(e)

	Ends	Coordinates	Length (bp)
1	PvuI-PvuI	69348-73870	4523
2	PvuI-PvuI	73871-69347	186686

(f)

Table 3.8: The size comparison among agarose gel fragments, pulsed-field gel fragments and expected fragments for BAC 148I5 DNA. (a) Fragmentation result of *SalI* digestion; (b) Coodinates of *SalI* fragments on BAC 148I5 DNA sequence; (c) Fragmentation result of *PmeI* digestion; (d) Coodinates of *PmeI* fragments on BAC 148I5 DNA sequence; (e) Fragmentation result of *PvuI* digestion; (b) Coodinates of *PvuI* fragments on BAC 148I5 DNA sequence.

The above table 3.8 showed the two different sub-tables for each restriction enzyme of *SalI, PmeI and PvuI*. The table (a) and (c) represented the predicted fragment sizes of normal agarose gel and pulsed-field electrophoresis gel, comparing with expectation fragments, whereas, the table (e) represented the predicted fragment sizes of pulsed-field electrophoresis gel, comparing with expectation fragments. The table (b) showed coordinates of fragments in BAC TGS assembly corresponding to table (a). Same way the table (d) shows coordinates of fragments in BAC TGS assembly corresponding to table (c). After careful observation, it was identified missing fragment of length 5.591 kb in between the coodinates of *PmeI* restriction enzyme digestion was shown in table (d) which was marked in yellow colour. The results from each restriction enzyme was identified the missing fragment in *SalI* (5.594 kb) and *PmeI* (5.591 kb). In *PvuI* restriction enzyme, 178 kb fragment was obtained instead of 186 kb (expected fragment) – it shows nearly 8 kb was missing in *PvuI* restriction enzyme. Overall analysis, almost the TGS of BAC 148I5 assembly fragments were matched with the actual results except nearly 6 kb fragment at the end of the sequence.

3.3 Analysis of gene structure and content

3.3.1 Bac End Sequencing

3.3.1.1 Comparison of Bac end sequencing and rhesus reference genome

BAC DNA & Primer	Identity	Chr	Start	End
BAC 201P10 - SP6	99.2%	8	8066198	8066874
BAC 201P10 - T7	96.5%	8	8150444	8151111
BAC 246K23 - SP6	98.8%	8	8140658	8141449
BAC 246K23 - T7	99.6%	8	7957551	7958383
BAC 148I5 - SP6	100.0%	5	131821372	131821863
BAC 148I5 - T7	100.0%	5	131997541	131997983

Table 3.9: The percentages and coordinates of sequence alignment between rhesus macaque assembly (reference genome) and BAC end-sequencing (FGS) using UCSC genome browser.

The above table 3.9 showed percentage identity, corresponding to their coordinates sequence and chromosome number between rhesus macaque assembly (reference genome) and BAC end-sequencing (FGS) using USCS genome browser. The BAC 148I5 genome with the primers SP6 and T7 showed high similarity (100%) with reference genome, BAC 148I5 with primer SP6 is located on chromosome (Chr) 5 at the region 131821372-131821863, whereas, the BAC 148I5 with T7 was located at the region chr5: 131997541-31997983. The BAC 246K23 with T7 primer showed 99.6% similarity with Rhesus macaque genome which was located on Chr 8 between the region 7957551-7958383, whereas, the BAC 201P10 with SP6 primer showed 99.2% similarity with reference genome at the region 8066198 – 8066874 on same chromosome. The BAC

201P10 DNA with the primer T7 and the BAC 246K23 with SP6 primer showed more than 96% similarity with the rhesus macaque genome on same chromosome. The BAC 148I5 on chromosome 5 indicated misassembling, since, the beta-defensin gene cluster was located on chromosome 8 in the macaque genome.

3.3.1.2 The comparison of bac end sequencing and TGS

BAC DNA & Primer	Identity	Gaps	Score	E-Value
BAC 201P10 – SP6	N/A	N/A	N/A	N/A
BAC 201P10 – T7	669/674 (99%)	5/674 (0.75%)	656	0.0
BAC 246K23 – SP6	720/720 (100%)	0/720 (0%)	720	0.0
BAC 246K23 – T7	833/834 (99.8%)	1/833 (0.12%)	830	0.0
BAC 148I5 – SP6	394/395 (99.7%)	1/395 (0.25%)	391	0.0
BAC 148I5 – T7	443/443 (100%)	0/443 (0%)	443	0.0

Table 3.10: The alignment results between bac end sequence (FGS) and TGS using blast.

The table 3.10 demonstrated the results of BLAST alignment between FGS and TGS data. The BAC 201P10 sequence (FGS) with SP6 primer was not matched any region in TGS assembly. Therefore, it was considered that this sequence region can support extra fragments obtained by *KpnI* restriction digestion. The BAC 246K23 genome with SP6 primer and BAC 148I5 with T7 primer shows 100% similarity, whereas, the BAC 201P10 with T7 primer, BAC 246K23 with T7 primer and BAC 148I5 with SP6 primer has more than 99% similarity. The expectation value for all the BAC DNAs are same (0). The BAC 246K23 with T7 primer shows highest score 830 with one gap (0.12%) at 812 base, whereas, the BAC 148I5 with SP6 primer shows lowest score with one gap penality (0.25%) at 391 base. The gap penalities in BAC 201P10 sequence with T7 primer show highest 0.75% compare to other sequences.

BAC DNA & PRIMER	Deletion in TGS	Insertion in TGS	Base change in TGS	Total
BAC 201P10-SP6	N/A	N/A	N/A	N/A
BAC 201P10-T7	-	5/674 (0.74%)	-	5/674 (0.74%)
BAC 246K23-SP6	-	-	-	0
BAC 246K23-T7	-	1/834 (0.12%)	-	1/834 (0.12%)
BAC 148I5-SP6	1/395 (0.25%)	-	-	1/395 (0.25%)
BAC 148I5-T7	-	-	-	0

Table 3.11: Deletion, insertion and base changes in TGS depending on bac end sequencing

This table 3.11 showed deletions, insertions and base changes in BAC DNAs of TGS data compared with FGS(end sequencing). Two BAC sequences with T7 primer showed insertions in the sequence and only 1 deletion with 0.25% in BAC 148I5 genome with SP6 primer. The sequence of BAC 201P10 with T7 primer have 5 insertions out of 674 (0.74%), whereas, the BAC 246K23 sequence with T7 primer have 1 insertion out of 834 (0.12%). There is no changes in base in all the BAC DNAs and the right side lane of the table shows total percentage of deletion, insertions and base changes.

3.3.2 Sequence similarity of BAC genome of TGS assemblies

3.3.2.1 Sequence similarity of BAC 201P10 genome with BAC 246K23 genome

S.No	Identities	Location of matching regions (1) BAC 201P10 and (2) BAC 246K23	
1	28850/29564 (98%)	18271- 47630	**(1)**
		139626 -169002	**(2)**
2	7917/8104 (98%)	208 - 8271	**(1)**
		136918 -128946	**(2)**
3	7265/7820 (93%)	49078 - 56746	**(1)**
		93380 - 85710	**(2)**
4	1867/2105 (89%)	42298 - 44284	**(1)**
		103193 - 101150	**(2)**
5	1398/1499 (93%)	47620 - 49088	**(1)**
		97137 - 95567	**(2)**
6	1336/1471 (91%)	56880 - 58305	**(1)**
		85593 - 84181	**(2)**

Table 3.12: Sequence similarity of BAC 201P10 and BAC 246K23 genome

The table 3.12 showed the sequence similarity of BAC 201P10 with BAC 246K23 genome. The both genomes were analysed by BLAST algorithm. The third lane in the table shows the location of matching regions in both genomes of BAC 201P10 and BAC 246K23. The second lane shows the identity of both genome sequence corresponding to the location of matching regions. In this table, overall identity sequence bases of BAC genomes (BAC 201P10 and BAC 246K23) of TGS assemblies showed high similarity in the region which was approximately 50kb.

3.3.2.2 Sequence similarity of BAC 246K23 genome with BAC 148I5 genome

S.No	Identities	Location of matching regions (1) BAC 246K23 and (2) BAC 148I5	
1	10687/10712 (99%)	128944 - 139600	**(1)**
		78820 - 68141	**(2)**
2	1122/1261 (89%)	4421 - 5665	**(1)**
		142962 - 141724	**(2)**
3	870/910 (96%)	4421 - 5311	**(1)**
		86311 - 85405	**(2)**
4	875/926 (94%)	94690 - 95605	**(1)**
		85405 - 86309	**(2)**
5	1130/1316 (86%)	94320 - 95623	**(1)**
		141724 - 142971	**(2)**
6	943/1148 (82%)	113571 - 114707	**(1)**
		142963 - 141874	**(2)**

Table 3.13: Sequence similarity of BAC 246K23 and BAC 148I5 genome

The table 3.13 showed the sequence similarity of BAC 246K23 with BAC 148I5 genome by BLAST analysis. The second and third lane in the table showed sequence identity and location of matching regions between BAC 246K23 and BAC 148I5 genome sequence. The region at 128944-139600 of BAC 246K23 sequence showed high similarity 99% with the sequence at 78820-68141 region of BAC 148I5 genome. The region at 4421-5311 of BAC 246K23 genome shows 96% similarity with BAC 148I5 genome at 86311-85405 regions and the regions at 94690-95605 of BAC 246K23 genome shows 94% sequence similarity with the BAC 148I5 genome at 85405-86309 region. The regions at second, fifth and sixth row in the table shows more than 80% sequence similarity. Overall identity sequence bases of BAC genome (BAC 246K23 and BAC 148I5) of TGS assemblies showed similarity in the region which was nearly 16kb .

3.3.2.3 Sequence similarity of BAC 201P10 genome with BAC 148I5 genome

S.No	Identities	Location of matching regions (1) BAC 201P10 and (2) BAC 148I5
1	7917/8112 (98%)	208-8271 **(1)** 70838-78818 **(2)**
2	827/921 (90%)	64179-65076 **(1)** 69022-69883 **(2)**

Table 3.14: Sequence similarity between BAC 201P10 and BAC148I5 genome

The table 3.14 described the sequence similarity between BAC 201P10 and BAC148I5 genome by BLAST analysis. In both genome, regions at two location shows sequence similarity. The region at 208-8271 of BAC 201P10 genome showed 98% sequence similarity with BAC 14815 genome at 70838-78818 region and the other matching region 64179-65076 of BAC 201P10 showed 90% similarity with BAC 14815 genome at 69022-69883 region. Overall identity sequence bases of BAC genome (BAC 201P10 and BAC 148I5) of TGS assemblies showed similarity in the region which was nearly 9 kb .

3.3.3 Analysis of rhesus beta-defensin gene cluster

Sanger sequence result of each beta-defensin gene and third generation sequence of each BAC DNA were aligned by using BLAST to identify any deletion, insertion or base change at TGS genome assemblies.

Beta-defensin gene	BAC DNA	Identities	Location at BAC	Strand
DEFB2L	**BAC 201P10**	765/834		+/-
		92%	57845-57031	
“	**BAC 246K23**	770/845		+/+
		91%	84618-85453	
DEFB103	**BAC 246K23**	766/773		+/+
		99%	73683-74452	
DEFB104	**BAC 246K23**	2556/2565		+/+
		99%	31082-33646	
DEFB105	**BAC 246K23**	1838/1840		+/+
		99%	14052-15891	
DEFB106	**BAC 246K23**	1422/1436		+/+
		99%	20340-21774	
DEFB107	**BAC 246K23**	2249/2342		+/+
		96%	951-3285	
DEFB109	**BAC 201P10**	256/308		+/-
		83%	90783-90488	
“	**BAC 246K23**	258/300		+/+
		86%	13586-13882	
“	**BAC 148I5**	260/304		+/+
		86%	159001-159302	
SPAG11A	**BAC 246K23**	4870/4896		+/+
		99%	47851-52739	
SPAG11B	**BAC 246K23**	4664/4681		+/+
		99%	39470-44150	

Table 3.15: The length and coordinates of beta-defensin genes at BAC DNAs depending on BLAST

In the table 3.15, *DEFB103, DEFB104, DEFB105, DEFB106,DEF107, SPAG11A-B* located on only BAC 246K23. While *DEFB2L* was at both BAC 201P10 and 246K23, all three BAC contained *DEFB109* gene. The alignment results demonstrated a SNP at *DEFB104* gene and insertions at *DEFB107* which are located on BAC 246K23.

3.3.3.1 SNP detection at DEFB104

```
Query  1      AGTGAGAAGTGAATTTGAATTGGACAGAATATGTGGTTATGGGACTGCCCGCTGCCGGAA
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  33436  AGTGAGAAGTGAATTTGAATTGGACAGAATATGTGGTTATGGGACTGCCCGCTGCCGGAA
33495

Query  61     CAAATGTCGAAGCCAAGAATACAAAATTGGAAGATGTCCCAACTCCTATGCATGCTGTTT
              |||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  33496  CAAATGTCGAAGTCAAGAATACAAAATTGGAAGATGTCCCAACTCCTATGCATGCTGTTT
33555

Query  121    GAGAAAATGGGATGAGAGCTTACTGAATCGTACAAAACCCTGA  163
              |||||||||||||||||||||||||||||||||||||||||||
Sbjct  33556  GAGAAAATGGGATGAGAGCTTACTGAATCGTACAAAACCCTGA  33598
```

Table 3.16: Blast result between DEFB104 and BAC 246K23. Query is Sanger sequence of DEFB104 ; sbjct is TGS of BAC 246K23

To confirm the SNP (AGC→AGT/ serin →serin) in different rhesus samples PCR and digestion reaction were carried out.

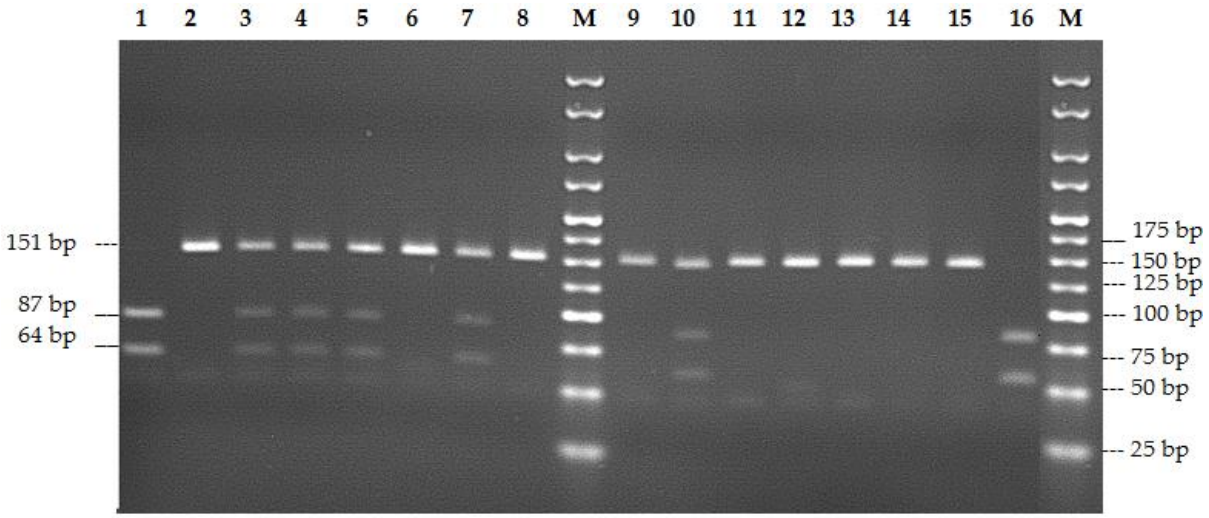


Figure 3.11: The image of agarose gel for the SNP at beta-defensin DEFB104 cluster.
Lanes 1-16 rhesus macaque DNA samples from UCSD; (M) Hyper ladder V

The above table 3.11 is the image of agarose gel electrophoresis for the SNP at beta-defensin DEFB104 gene. The lanes from 1 to 16 in the gel represent rhesus macaque DNA samples and lane M represents hyper ladder V. *Hpy188III* restriction enzyme cut the fragment near thymine residues, therefore two bands at 87bp and 64 bp were appeared in the gel. The band which appear

at 151bp represented the uncut fragment which contains cytosine residue instead of thymine.

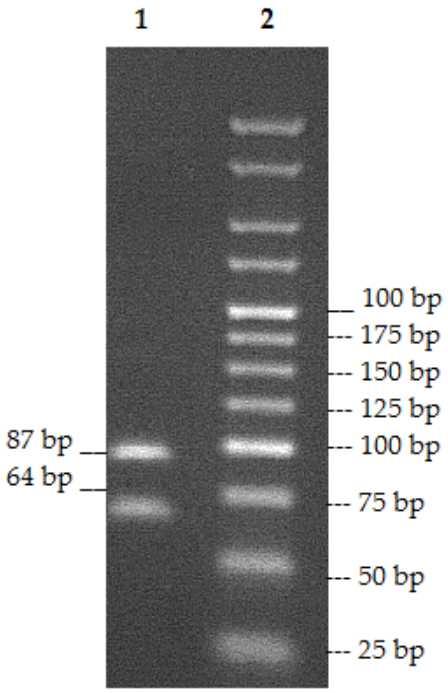


Figure 3.12: The image of agarose gel for the SNP at DEFB104 gene in BAC 246K23.
(1)BAC 246K23 DNA sample; (2) Hyper ladder V.

The above table 3.12 is the image of agarose gel electrophoresis for the SNP at beta-defensin DEFB104 cluster. The lane 1 represents BAC 246K23 DNA sample and lane 2 represents hyper ladder V. The restriction enzyme cut the fragment near thymine residues, therefore two bands at 87bp and 64 bp were appeared in the gel.

Samples	Homozygous C	Homozygous T	Heterozygous	Allels
Rh1		√		**TT**
Rh2	√			**CC**
Rh3			√	**CT**
Rh4			√	**CT**
Rh5			√	**CT**
Rh6	√			**CC**
Rh7			√	**CT**

Rh8	√				**CC**
Rh9	√				**CC**
Rh10				√	**CT**
Rh11	√				**CC**
Rh12	√				**CC**
Rh13	√				**CC**
Rh14	√				**CC**
Rh15	√				**CC**
Rh16		√			**TT**
BAC 246K23		√			**TT**

Table 3.17: Genotyping results of SNP at DEFB104 gene

The above table 3.17 describes genotyping result of a SNP in DEFB104 gene. In this gene, three homozygous T (TT alleles) are located at Rh1, Rh16 and BAC genome and 5 heterozygous alleles were found in Rh3, Rh4, Rh5, Rh7 and Rh10 genome. Compare to homozygous T and heterozygous alleles, the homozygous C alleles are more in number.

3.3.3.2 Detection of insertions at DEFB107 on BAC 246K23

```
Query  1       ACAGCCTAGGAAGGGTATCTCCTTGTAGCATTGGAAGCTGGACTGACATGGTTTCAGATA
               |||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  100859  ACAGCCTAGG-AGGGTATCTCCTTGTAGCATTGGAAGCTGGACTGACATGGTTTCAGATA
100801

Query  61      ATCCAAACTTTGCAGATCAAAGAGAGACGGTG-CAGAGAGATTTGTCCCACTGATATCGC
               |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  100800  ATCCAAACTTTGCAGATCAAAGAGAGACGGTGCCAGAGAGATTTGTCCCACTGATATCGC
100741

Query  120     AGCCAGAGAATCTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTT
               |||||||||||||||||||||||||||||| |||||||||||||||||| ||||||||||
Sbjct  100740  AGCCAGAGAATCTTCACCTCTTTCATTCTT-CAGCTGGTGCTTAGTTTTAAATCTTTCTT
100682

Query  180     TCTCTTTTTGCAGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTAC
               |||||||||||||||||||| ||||||||    |||||| ||||||||| |||||||||
Sbjct  100681  TCTCTTTTTGCAGCAATATGGATGTTAATTAATTCAGCTCATACAGCCCC-AATCTTTAC
100623

Query  234     TTCAAAGGTAAGACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGC
               |||||||||||||||| |||| |||||||||| |||||||||||||||||||| ||| 
Sbjct  100622  TTCAAAGGTAAGACATTTCAGCCTTCACAGTGAC-TTCCATTCTCTTACAAATTAGGTGC
100564

Query  291     TCTATGAATTGCTGTCCTGGCTAAGTAAAGAAAGCAATTTGTCTTTAGTTAT-CTATTAA
               |||||||||||||||||||||||||||||||||| ||||||||| ||  | |||||||
Sbjct  100563  TCTATGAATTGCTGTCCTGGCTAAGTAAAGAAAGCA-TTTGTCTTT-GTAGTGCTATTAA
100506

Query  350     TTAGTAACTATAAGA-TAAAATTACATTGCCCAGGCCAAATTCAGAC-TTCCTC-ATTAC
               ||||||||||||||| |||||| |||||||||||||||||||||||| |||||| ||| 
Sbjct  100505  TTAGTAACTATAAGAATAAAAT-ACATTGCCCAGGCCAAATTCAGACCTTCCTCTATTTA
100447

Query  407     CATAAC-TCC  415
               |||||| ||| 
Sbjct  100446  CATAACCTCC  100437
```

Table 3.18: Blast result between DEFB107 and BAC 246K23. Query is Sanger sequence of DEFB107 ; sbjct is TGS of BAC 246K23

Alignment between DEFB104 and BAC 246K23 showed insertion of AATT at DEFB104 located on TGS of BAC 246K23. To confirm this insertion at rhesus samples, sequencing results were analysed by using EBI CLUSTAL 2.1 multiple sequence alignment

```
tgs             GCAGATCAAAGAGAGACGGTGCCAGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 120
RH2             GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCNCNNCCAGAGAAT 65
RH4             GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 63
RH8             GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 63
RH10            GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 62
BAC             GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 62
RH15            GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 63
RH5             GCAGATCANNNAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 62
RH7             GCAGATCANNNAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 64
RH6             GCAGATCANANAGAGANNGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 63
RH13            GCAGATCANANAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 63
RH14            GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 62
RH16            GCAGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 62
RH9             GNNNANCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 67
RH11            GCAGATCANAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 63
RH12            GNNGATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 63
RH1             GCAGANCANAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 64
RH3             GNNNATCAAAGAGAGACGGTGC-AGAGAGATTTGTCCCACTGATATCGCAGCCAGAGAAT 64
                *   * **   *****  **** ********************** *  *********

tgs             CTTCACCTCTTTCATTCTT-CAGCTGGTGCTTAGTTTTAAATCTTTCTTTCTCTTTTTGC 179
RH2             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 125
RH4             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 123
RH8             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 123
RH10            CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 122
BAC             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 122
RH15            CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 123
RH5             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 122
RH7             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 124
RH6             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 123
RH13            CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 123
RH14            CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 122
RH16            CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 122
RH9             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 127
RH11            CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 123
RH12            CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 123
RH1             CTTCNCCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 124
RH3             CTTCACCTCTTTCATTCTTGCAGCTGGTGCTTAGTTTTTAATCTTTCTTTCTCTTTTTGC 124
                **** ************** ***************** ********************

tgs             AGCAATATGGATGTTAATTAATTCAGCTCATACAGCCCC-AATCTTTACTTCAAAGGTAA 238
RH2             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 179
RH4             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 177
RH8             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 177
RH10            AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 176
BAC             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 176
RH15            AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 177
RH5             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 176
RH7             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 178
RH6             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 177
RH13            AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 177
RH14            AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 176
RH16            AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 176
RH9             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 181
RH11            AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 177
RH12            AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 177
RH1             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGNNA 178
RH3             AGCAATATGG-TGTTAATT----CAGCTC-TACAGCCCCCAATCTTTACTTCAAAGGTAA 178
                ********** ********    ****** ********* ************ ** *  *

tgs             GACATTTCAGCCTTCACAGTGAC-TTCCATTCTCTTACAAATTAGGTGCTCTATGAATTG 297
```

```
RH2             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 236
RH4             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 234
RH8             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 234
RH10            GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 233
BAC             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 233
RH15            GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 234
RH5             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 233
RH7             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 235
RH6             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 234
RH13            GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATAAATTG 234
RH14            GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 233
RH16            GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 233
RH9             GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 238
RH11            GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACAAATTAG-TGCTCTATGAATTG 234
RH12            GACATT-CAGC-TTCACAGTGACCTTCCATTCTCTTACANATTAG-TGCTCTATGAATTG 234
RH1             NACATT-CAGC-TTC-CAGTGACCTTCNNTTCTCTTACAAATTAN-TGCTCTATGAATTG 234
RH3             GACATT-CAGC-TTCACAGTGACCTTACATTCTCTTACAAATTAN-TGCTNTATGANTTG 235
                 ***** **** *** ******* **   ********** ****  **** *** * ***

tgs             CTGTCCTGGCT
RH2             CTGTCCTGGCT
RH4             CTGTCCTGGCT
RH8             CTGTCCTGGCT
RH10            CTGTCCTGGCT
BAC             CTGTCCTGGCT
RH15            CTGTCCTGGCT
RH5             CTGTCCTGGCT
RH7             CTGTCCTGGCT
RH6             CTGTCCTGGCT
RH13            CTGTCCTGGCT
RH14            CTGTCCTGGCT
RH16            CTGTCCTGGCT
RH9             CTGTCCTGGCT
RH11            CTGTCCTGGCT
RH12            CTGTCCTGGCT
RH1             CTGTCCTGGCT
RH3             CTGTCCTGGCT
                ***********
```

Table 3.19: Multiple sequence alignment. RH1-16 : Sanger sequence of rehesus samples for DEFB107 gene; BAC: Sanger sequence of BAC 246K23 for DEFB107 gene; TGS: third generation sequence of BAC 246K23 for DEFB107 gene.

From table 3.19, multiple sequence alignment result showed that Sanger sequences of all rhesus samples and BAC 246K23 did not include insertion of AATT. The presence of this insertion part at only TGS data might indicate sequence error at the TGS assembly.

Chapter 4: DISCUSSIONS

Sequencing techniques are used to analyse whole genome, different isoforms of genes, chromatin conformation, nucleic acid structure, point mutation, copy number variation, transcriptome and methylome detections to understand the interactions among nucleotides and proteins (Schadt *et al.*, 2010), (Meyerson *et al.*, 2010) (Kircher & Kelso, 2010). The FGS of Sanger technique was the first significant method used for sequencing genome for Human Genome Project. Although number of research work published by this method, but had to be improved in terms of read length, time consuming and high cost (Schadt *et al.*, 2010). The SGS techniques were introduced into the market by reducing time and cost, however, still read length has been the main challenge for complex genome sequencing (Scatz, 2010). In this case, TGS has provided large-scale sequencing and real time detection to exceed previous limitations (McCarthy, 2010 Our study covers of testing PacBio Single Molecule Real Time Sequencing Technique (SMRT) as a TGS method to analyse both the single base and the molecular level of complex genomes, also assemble a contiguous BAC sequence across the macaque beta-defensin region, and perform a preliminary analysis of sequence variation.

The BAC DNAs were isolated and analyzed. Restriction fragment digestion patterns had matched only one of BAC assembly with the TGS assemblies. The digestion analysis was done with three BAC DNAs, BAC 201P10, BAC 246K23 and BAC 148I5. Among these three analysis, the restriction digestion of BAC 148I5 obtained expected fragments compare to TGS assemblies, whereas, the BAC 201P10 and BAC 246K23 was not matched with TGS assemlies. The restriction analysis of BAC 201P10 exists with additional repetitive regions; whereas, the BAC 246K23 DNA exists with missing regions.

The preliminary analysis of sequence alignment was done between the reference genome rhesus macaque assembly and BAC end-sequencing of FGS by USCS genome browser and shows high similarity. The sequence alignments between FGS data and TGS data by BLAST tool, that shows high similarity in all BAC DNAs, but the BAC 201P10 with SP6 primer sequence data was not available. The sequence analysis shows high identity more than 99% with high score, less gap penalities and low e-value (0). The comparative analysis of FGS and TGS data shows

indel regions in BAC end sequencing. The analysis of single nucleotide polymorphism in beta-defensin DEFB104 gene was carried out and confirmed an actual SNP without any sequence error. Eventually, insertions at of DEFB107 gene demonstrated sequence error in TGS. As a result, our analysis supported that errors and misassembling in third generation sequence. All these drawbacks confirm TGS limitations.

Sequencing long read length was the main aim for most sequencing technologies to resolve complex, dynamic genomes. SMRT technology of TGS developed by Pasific Biosystems presented zero mode waveguide approach to long sequence reads and high speed assembly. Our study demonstrated that, the data of TGS by SMRT technology shows some gaps in the sequence and assembling accuracy even though it provides long read length and direct assembly. Therefore, in future, enzyme kinetics can be improved to increase the accuracy of single molecule sequencing in real time (Eid *et al*. 2009). Furthermore, if higher level dimensional data is accurately analysed and assimilated, TGS will contribute towards a better understanding of large-scale DNA sequencing (Schadt, *et al.*, 2010). The limitations of SMRT technology is high raw data error rate of SMRT represents deletions, insertions, misassembling and misalignment. Too short incorporation and intervals between reads and contamination of unlabelled molecules causes erroneous deletions during single molecule sequencing. In addition, disassociation of the phosphate group from cognate nucleotides is a major source of insertion errors. Despite the accuracy level being adequate for a consensus on resequencing, it would be a challenge for the alignment of repetitive regions and de novo assembly. (Eid *et al*. 2009). An additional challenge of this technology is that of analysing the sequenced data. SMRT sequencing data does not function in the same way as the sequencing data described previously; thus, it is necessary to engage in advanced probabilistic modelling. (Schadt, *et al.*, 2010).

Conclusion

In this preliminary study, we have able to isolate BAC DNAs by and perform restriction enzyme maps for each BAC DNA sequence by using the online bioinformatics tool "Nebcutter V2.0" and "restrictionmapper". According to these maps, restriction enzymes were selected to carried out restriction digestion pattern analysis with normal agarose gel and PFGE. We have also performed BLAST analysis to identify sequence similarity among FGS and TGS data. And have also performed analysis of single nucleotide polymorphisms with beta-defensin gene DEFB104 cluster. We are highly interested to examine the quality of the DNA sequence produced by this new approach to analyze beta-defensin gene cluster as a complex region in the rhesus macaque genome. The final outcome of this work covers errors and misassembling in TGS data generated by PacBio assembly. In future, analysis of TGS data can be improved and compared with SGS results to provide better resolution for complex genomic regions.

References:

1. Adams, M.D. *et al.,* 2000. The genome sequence of Drosophila melanogaster. *Science.* **287,** 2185-2195.

2. Balzer, S., Malde, K., Lanzen, A., Sharma, A., Jonassen, I., 2010. Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. *Bioinformatics (Oxford, England).* **26,** i420-5.

3. Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 5, Mapping Genomes. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21116/

4. Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 6, Sequencing Genomes. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21117/

5. Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., Lander, E.S., 2002. ARACHNE: a whole-genome shotgun assembler. *Genome Research.* **12,** 177-189.

6. Chen, F., Alessi, J., Kirton, E., Singan, V. and Richardson, P. (2006) Comparison of 454 sequencing platform with traditional Sanger sequencing: a case study with de novo sequencing of Prochlorococcus marinus NATL2A genome. Poster LBNL 59003. Plant & Animal Genome XIV Conference, January 14–18, 2006 (San Diego, CA).

7. Dale, J. W., Schantz, M. V., Plant, N. 2012: *From Genes to Genomes, Concepts and Applications of DNA technology*. Chichester, West Sussex: John Wiley & Sons Ltd.

8. Eid, J. *et al.,* 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science.* **323,** 133-138.

9. Fellermann, K. *et al.,* 2006. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *American Journal of Human Genetics.* **79,** 439-448.

10. Gardiner, R.M., 2002. The Human Genome Project: the next decade. *Archives of Disease in Childhood.* **86,** 389-391.

11. Golan, D. & Medvedev, P., 2013. Using state machines to model the Ion Torrent sequencing process and to improve read error rates. *Bioinformatics (Oxford, England).* **29,** i344-i351.

12. Hollox, E.J. *et al.*, 2008. Definsins and the dynamic genome: what we can learn from structural variation at human chromosome ban 8p23.1. Cold Spring Harbor Laboratiry Press.

13. Hollox, E. J., 2012. The challenges of studying complex and dynamic regions of the human genome. *Methods Mol Biol.* **838,** 187-207

14. Johnson, M.E. *et al.*, 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature.* **413,** 514.

15. Kelley, J. M.; Field, C. E.; Craven, M. B.; Bocskai, D.; Kim, U. J.; Rounsley, S. D.; Adams, M. D., 1999. High throughput direct end sequencing of BAC clones. *Nucleic Acids Research,* **27** pp. 1539 - 1546

16. Kircher, M. & Kelso, J., 2010. High-throughput DNA sequencing--concepts and limitations. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology.* **32,** 524-536.

17. Lupski, J. R., 2007. Genomic rearrangements and sporadic disease. *Nature Genetics.* **39 Suppl 1,** S43-S47.

18. Lee, A.S., Gutierrez-Arcelus, M., Perry, G.H., Vallender, E.J., Johnson, W.E., Miller, G.M., Korbel, J.O., Lee, C., 2008. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Human Molecular Genetics.* **17,** 1127-1136.

19. Lee, K.T., Byun, M.J., Kang, K.S., Hwang, H., Park, E.W., Kim, J.M., Kim, T.H., Lee, S.H., 2012. Single nucleotide polymorphism association study for backfat and intramuscular fat content in the region between SW2098 and SW1881 on pig chromosome 6. *Journal of Animal Science.* **90,** 1081-1087.

20. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology.* **2012,** 251364.

21. Maxam, A.M. & Gilbert, W., 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America.* **74,** 560-564.

22. Mefford, H.C. & Eichler, E.E., 2009. Duplication hotspots, rare genomic disorders, and common disease. *Current Opinion in Genetics & Development.* **19,** 196-204.

23. McCarthy, A., 2010. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chemistry & Biology.* **17,** 675-676.

24. Meyerson, M., Gabriel, S., Getz, G., 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews.Genetics.* **11,** 685-696.

25. Osoegawa, K., Woon, P.Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J.J., de Jong, P.J., 1998. An improved approach for construction of bacterial artificial chromosome libraries. *Genomics.* **52,** 1-1.

26. Ozsolak, F., 2012. Third-generation sequencing techniques and applications to drug discovery. *Expert Opinion on Drug Discovery.* **7,** 231.

27. Pareek, C.S., Smoczynski, R., Tretyn, A., 2011. Sequencing technologies and genome sequencing. *Journal of Applied Genetics.* **52,** 413-435.

28. Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* **13,** 341-2164-13-341.

29. Redon, R., *et al.*, 2006. Global variation in copy number in the human genome. *Nature.* **444,** 444-454.

30. Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., Batzer, M.A., Bustamante, C.D., Eichler, E.E., Hahn, M.W., Hardison, R.C., Makova, K.D., Miller, W., Milosavljevic, A., Palermo, R.E., Siepel, A., Sikela, J.M., Attaway, T., Bell, S., Bernard, K.E., Buhay, C.J., Chandrabose, M.N., Dao, M., Davis, C., Delehaunty, K.D., Ding, Y., Dinh, H.H., Dugan-Rocha, S., Fulton, L.A., Gabisi, R.A., Garner, T.T., Godfrey, J., Hawes, A.C., Hernandez, J., Hines, S., Holder, M., Hume, J., Jhangiani, S.N., Joshi, V., Khan, Z.M., Kirkness, E.F., Cree, A., Fowler, R.G., Lee, S., Lewis, L.R., Li, Z., Liu, Y.S., Moore, S.M., Muzny, D., Nazareth, L.V., Ngo, D.N., Okwuonu, G.O., Pai, G., Parker, D., Paul, H.A., Pfannkoch, C., Pohl, C.S., Rogers, Y.H., Ruiz, S.J., Sabo, A., Santibanez, J., Schneider, B.W., Smith, S.M., Sodergren, E., Svatek, A.F., Utterback, T.R., Vattathil, S., Warren, W., White, C.S., Chinwalla, A.T., Feng, Y., Halpern, A.L., Hillier, L.W., Huang, X., Minx, P., Nelson, J.O., Pepin, K.H., Qin, X., Sutton, G.G., Venter, E., Walenz, B.P., Wallis, J.W., Worley, K.C., Yang, S.P., Jones, S.M., Marra, M.A., Rocchi, M., Schein, J.E., Baertsch, R., Clarke, L., Csuros, M., Glasscock, J., Harris, R.A., Havlak, P., Jackson, A.R., Jiang, H., Liu, Y., Messina, D.N., Shen, Y., Song, H.X., Wylie, T., Zhang, L., Birney, E., Han, K., Konkel, M.K., Lee, J., Smit, A.F., Ullmer, B., Wang, H., Xing, J., Burhans, R., Cheng, Z., Karro, J.E., Ma, J., Raney, B., She, X., Cox, M.J., Demuth, J.P., Dumas, L.J., Han, S.G., Hopkins, J., Karimpour-Fard, A., Kim, Y.H., Pollack, J.R., Vinar, T., Addo-Quaye, C., Degenhardt, J., Denby, A., Hubisz, M.J., Indap, A., Kosiol, C., Lahn, B.T., Lawson, H.A., Marklein, A., Nielsen, R., Vallender, E.J., Clark, A.G., Ferguson, B., Hernandez, R.D., Hirani, K., Kehrer-Sawatzki, H., Kolb, J., Patil, S., Pu, L.L., Ren, Y., Smith, D.G., Wheeler, D.A., Schenck, I., Ball, E.V., Chen, R., Cooper, D.N., Giardine, B., Hsu, F.,

Kent, W.J., Lesk, A., Nelson, D.L., O'brien, W.E., Prufer, K., Stenson, P.D., Wallace, J.C., Ke, H., Liu, X.M., Wang, P., Xiang, A.P., Yang, F., Barber, G.P., Haussler, D., Karolchik, D., Kern, A.D., Kuhn, R.M., Smith, K.E., Zwieg, A.S., 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science (New York, N.Y.).* **316,** 222-234.

31. Rothberg, J.M. & Leamon, J.H., 2008. The development and impact of 454 sequencing. *Nature Biotechnology.* **26,** 1117-1124.

32. Shaffer, L. G. and Lupski, J. R., 2000. Molecular mechanisms for constitutional chromosomal rearrangements in humans. Annual Review of Genetics, 34 pp.297-329

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

34. Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., Simon, M., 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America.* **89,** 8794-8797.

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

36. Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology.* **26,** 1135-1145.

37. Strausberg, R.L., Levy, S., Rogers, Y.H., 2008. *Emerging DNA sequencing technologies for human genomic medicine*. Drug Discovery Today. 13, 569-577

38. Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. *Human Molecular Genetics.* **19,** R227-40.

39. Schatz, M.C., Delcher, A.L., Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. *Genome Research.* **20,** 1165-1173.

40. *Williams, Vincent. Cell and Tissue Based Molecular Pathology a Volume in the Series Foundations in Diagnostic Pathology [Book Review] [online]. Australian Journal of Medical Science, Vol. 30, No. 4, Nov 2009: 178-179. Availability: <http://search.informit.com.au/documentSummary;dn=653110943375323;res=IELHEA > ISSN: 1038-1643. [cited 26 Jul 13].*

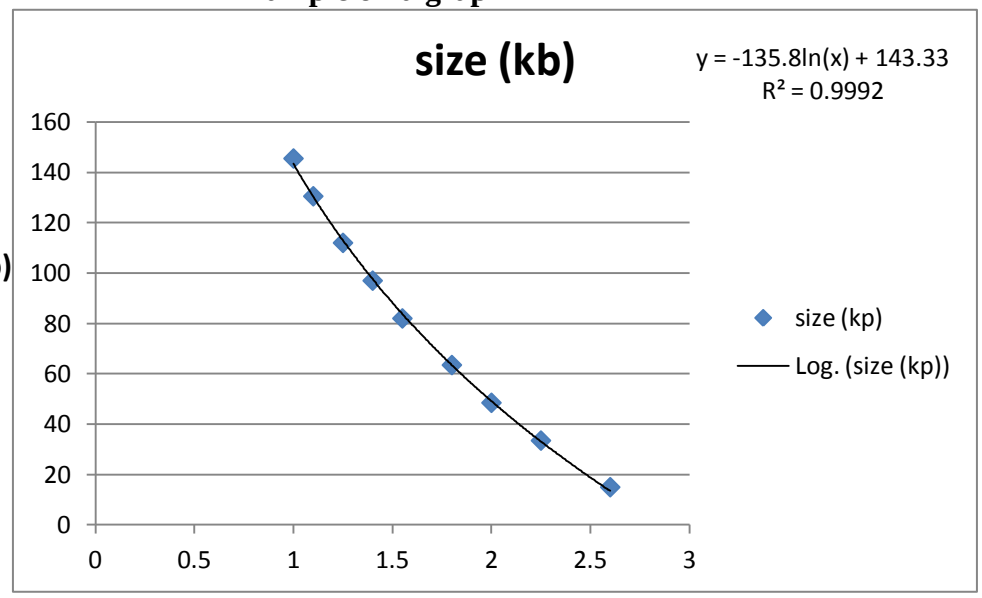
41. Zhang, H. & Wu, C., 2001. BAC as tools for genome sequencing. *Plant Physiology and Biochemistry.* **39,** 195-209.

42. Zhang, X., Goodsell, J., Norgren, R.B.,Jr, 2012. Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics.* **13,** 206-2164-13-206.

Appendices

Example of a graph

PFG ladder distance (cm)	size (kb)
1	145.5
1.1	130.5
1.25	112
1.4	97
1.55	82
1.8	63.5
2	48.5
2.25	33.5
2.6	15



This graph shows how size of fragments were calculated. The size of PFG ladder fragments were used as reference to calculate restriction enzyme fragments. In first step, distance of ladder fragments were calculated and then a graph was formed by using excel. The below formula was obtained from this graph:

$$y = -135.8\ln(x) + 143.33$$

After that, each fragments of restriction enzymes were calculated using this formula.

Protocol 1

MACHEREY-NAGEL Low-copy plasmid purification (Maxi / BAC, Mega)

	Maxi (AX 500 / BAC 100)	Mega (AX 2000)

1. Cultivate and harvest bacterial cells

Harvest bacteria from an LB culture by centrifugation at **4,500–6,000 x *g*** for **15 min** at **4 °C**.

2. Cell lysis

Carefully resuspend the pellet of bacterial cells in **Buffer S1 + RNase A**. Please see section 6.3 regarding difficult-to-lyse strains.

Maxi	Mega
24 mL	**90 mL**

Add **Buffer S2** to the suspension. Mix gently by inverting the tube 6–8 times. Incubate the mixture at room temperature (18–25 °C) for 2–3 min (max. 5 min). Do not vortex, as this will release contaminating chromosomal DNA from cellular debris into the suspension.

Maxi	Mega
24 mL	**90 mL**

Add pre-cooled **Buffer S3 (4 °C)** to the suspension. Immediately mix the lysate gently by inverting the flask 6–8 times until a homogeneous suspension containing an off-white flocculate is formed. Incubate the suspension on ice for 5 min.

Maxi	Mega
24 mL	**90 mL**

3. Equilibration of the column

Equilibrate a NucleoBond® AX 500 (Maxi), BAC 100 (Maxi), or AX 2000 (Mega) Column with **Buffer N2**. Allow the column to empty by gravity flow. Discard flowthrough.

Maxi	Mega
6 mL	**20 mL**

4. Clarification of the lysate

Clear the bacterial lysate by following EITHER **option 1** or **option 2**, described below. This step is extremely important; excess precipitate left in suspension may clog the NucleoBond® Column in later steps.

Note: For purification of BAC DNA it is recommended to follow option 1.

Note: Complete removal of precipitated protein and cell debris is essential to avoid clogging of the NucleoBond® Column.

Place a **NucleoBond® Folded Filter** in a funnel of appropriate size.

Wet the filter with a few drops of Buffer N2 and load the bacterial lysate onto the wet filter. Either collect the flow-through in a separate vessel and proceed with step 5 **OR** position funnel and filter directly on top of the NucleoBond® Column to clear and load the lysate in one time-saving step (skip step 5).

5. Binding

Load the cleared lysate from step 4 onto the NucleoBond® Column. Allow the

column to empty by gravity flow.
Optional: You may want to save all or part of the flow-through for analysis.

6. Washing
Wash the column with **Buffer N3.** Repeat as indicated. Discard flow-through.

Maxi	Mega
2 x 18 mL	**2 x 50 mL**

7. Elution
Elute the plasmid DNA with **Buffer N5**. Preheating Buffer N5 to 50 °C prior to elution may improve yields for high-molecular weight constructs such as BACs. We recommend precipitating the eluate as soon as possible (step 8). Nevertheless, the eluate can be stored in closed vials on ice for several hours. In this case the eluate should be preheated to room temperature before the plasmid DNA is precipitated.

Maxi	Mega
15 mL	**25 mL**

Optional: Determine plasmid yield by UV spectrophotometry in order to adjust the desired concentration of DNA (step 10).

8. Precipitation
Add **room-temperature isopropanol** to precipitate the eluted plasmid DNA. Mix carefully and centrifuge at ≥ **15,000 x *g*** for **30 min** at **4 °C**. Carefully discard the supernatant.

Maxi	Mega
11 mL	**18 mL**

9. Wash and dry DNA pellet
Add **room-temperature 70 % ethanol** to the pellet. Vortex briefly and centrifuge at ≥ **15,000 x *g*** for **10 min** at **room temperature (18‑25 °C)**.

Maxi	Mega
5 mL	**7 mL**

Carefully remove ethanol from the tube with a pipette tip. Allow the pellet to dry at **room temperature (18‑25 °C)** no less than the indicated time.
Drying for longer periods of time will not harm the quality of plasmid DNA but overdrying may render the DNA less soluble.

10‑20 min	**30‑60 min**

10. Reconstitute DNA
Dissolve pellet in an appropriate volume of buffer TE or sterile deionized H2O. Depending on the type of centrifugation tube, dissolve under constant spinning in a sufficient amount of buffer for 10–60 min (3D-shaker).
Determine plasmid yield by UV spectrophotometry. Confirm plasmid integrity by agarose gel electrophoresis.

Protocol 2

1. Centrifuge the Performa Gel Filtration Cartridge for 3 minutes at 850 *x g*.

• The time and speed of centrifugation are important.

• The drier the packing (longer centrifugation times and/or higher *g* forces), the longer it takes to recover product and the lower the overall recovery.

• Conversely, shorter spin times and lower speeds result in elution volumes higher than the input sample volume.

2. Transfer the cartridge to the provided 1.5-ml microcentrifuge tube and add the sample to the packed column. Be sure the fluid runs into the gel.

• If using a microcentrifuge or other centrifuge which uses a fixed angle rotor, place the sample in the center of the slanted gel bed surface to obtain optimal performance.

3. Close the cap and centrifuge for 3 minutes at 850 *x g*. Retain eluate.

• Up to 4 µl may be lost during sample processing.

• If the volume loss is greater than 4 µl, this is an indication of an overly dry gel. To optimize recovery of sample, repeat centrifugation.

Protocol 3

QIAquick Gel Extraction Kit Protocol

1. Excise the DNA fragment from the agarose gel with a clean, sharp scalpel.

Minimize the size of the gel slice by removing extra agarose.

2. Weigh the gel slice in a colorless tube. Add 3 volumes of Buffer QG to 1 volume of gel (100 mg ~ 100 µl).

For example, add 300 µl of Buffer QG to each 100 mg of gel. For >2% agarose gels, add 6 volumes of Buffer QG. The maximum amount of gel slice per QIAquick column is 400 mg; for gel slices >400 mg use more than one QIAquick column.

3. Incubate at 50°C for 10 min (or until the gel slice has completely dissolved). To help dissolve gel, mix by vortexing the tube every 2–3 min during the incubation.

IMPORTANT: Solubilize agarose completely. For >2% gels, increase incubation time.

4. After the gel slice has dissolved completely, check that the color of the mixture is yellow (similar to Buffer QG without dissolved agarose).

If the color of the mixture is orange or violet, add 10 µl of 3 M sodium acetate, pH 5.0, and mix. The color of the mixture will turn to yellow.

The adsorption of DNA to the QIAquick membrane is efficient only at pH ≤7.5.

Buffer QG contains a pH indicator which is yellow at pH ≤7.5 and orange or violet at higher pH, allowing easy determination of the optimal pH for DNA binding.

5. Add 1 gel volume of isopropanol to the sample and mix.

For example, if the agarose gel slice is 100 mg, add 100 µl isopropanol. This step increases the yield of DNA fragments <500 bp and >4 kb. For DNA fragment between 500 bp and 4 kb, addition of isopropanol has no effect on yield.

Do not centrifuge the sample at this stage.

6. Place a QIAquick spin column in a provided 2 ml collection tube.

7. To bind DNA, apply the sample to the QIAquick column, and centrifuge for 1 min.

The maximum volume of the column reservoir is 800 µl. For sample volumes of more than 800 µl, simply load and spin again.

8. Discard flow-through and place QIAquick column back in the same collection tube.

Collection tubes are re-used to reduce plastic waste.

9. (Optional): Add 0.5 ml of Buffer QG to QIAquick column and centrifuge for 1 min.

This step will remove all traces of agarose. It is only required when the DNA will subsequently be used for direct sequencing, in vitro transcription or microinjection.

10. To wash, add 0.75 ml of Buffer PE to QIAquick column and centrifuge for 1 min.

Note: If the DNA will be used for salt sensitive applications, such as blunt-end ligation and direct sequencing, let the column stand 2–5 min after addition of Buffer PE, before centrifuging.

11. Discard the flow-through and centrifuge the QIAquick column for an additional 1 min at ≥10,000 x *g* (~13,000 rpm).

IMPORTANT: Residual ethanol from Buffer PE will not be completely removed unless the flow-through is discarded before this additional centrifugation.

12. Place QIAquick column into a clean 1.5 ml microcentrifuge tube.

13. To elute DNA, add 50 µl of Buffer EB (10 mM Tris·Cl, pH 8.5) or H2O to the center of the QIAquick membrane and centrifuge the column for 1 min at maximum speed. Alternatively, for increased DNA concentration, add 30 µl elution buffer to the center of the QIAquick membrane, let the column stand for 1 min, and then centrifuge for 1 min.