EXPRESSION OF TN5 TRANSPOSASE FOR NEXT GENERATION SEQUENCING PROTOCOL OF HIV AND SELECTED ONCOVIRUSES

A dissertation submitted in fulfilment of the requirements
For
Master of science degree in Microbiology

Submitted by

Mathobo Phindulo

(14004968)

To

The Department of Microbiology
School of Mathematical and Natural Sciences
University of Venda

Supervisor: Prof. Pascal Obong Bessong
Co-Supervisor: Dr. Lufuno Grace Mavhandu-Ramarumo

March 2021

DECLARATION

I, Mathobo Phindulo, hereby declare that this report for the award of Master of Science degree in Microbiology at the University of Venda is my own work. It has not been submitted before for any degree examination at this or any other University. It is my own work at execution and all the reference materials contained therein have been duly acknowledged.

Mathobo Phindulo (Student)  Date: 11 March 2021

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation to the following people and organization:

- My supervisor: Professor Pascal Obong Bessong, I am thankful for his endless support, encouragement, and opportunity to learn from his vast knowledge in the field of molecular microbiology.

- My co-supervisor: Dr Lufuno Grace Mavhandu–Ramarumo, for the scientific guidance and advice she offered when needed.

- Dr Nontokozo Daphyne Matume, for her advice, alternative ideas and support that allowed me to understand more in what I was doing.

- Members of the AIDS Virus Research Laboratory, University of Venda for their support throughout the study.

- My family for being there for me when I needed them the most.

- Special thanks to God almighty.

- The National Research Foundation for supporting the study and providing me with a studentship (UID 113465). The University of Venda is also acknowledged for partial financial support (SMNS/17/MBY/25).

DEDICATION

To my father Mathobo Mbulaeni Samuel and mother Mathobo Lindelani Rosina.

Table of Contents

LIST OF ABBREVATIONS

HIV	Human immunodeficiency virus
HPV	Human papillomavirus
KSHV	Kaposi's Sarcoma associated herpes virus
NGS	Next Generation Sequencing
mRNA	Messenger ribonucleic acid
PEI	Polyethlenimine
IPTG	Isopropyl β-D-1-thiogalactopyranoside
MEDS	Mosaic End double-stranded
HMW	High molecular weight
RSB	Resuspension buffer
PCR	Polymerase chain reaction
ATM	Amplicon tagment mix
cDNA	Complementary deoxyribonucleic acid
EB	Elution buffer
ES	End Sequences
OE	Outside Ends
IE	Inside Ends
CBD	Chitin Binding Domain
PEG	Polyethylene Glycol
DMF	Dimethylformamide

LIST OF FIGURES

LIST OF TABLES

SCIENTIFIC COMMUNICATION FROM THIS DISSERTATION

- Phindulo Mathobo, Nontokozo Daphney Matume, Pascal Obong Bessong. Application of Next Generation Sequencing in HIV Drug Resistance Studies in Africa, 2005 – 2019: A Systematic Review. **Submitted to Scientific African (Manuscript Number: SCIAF-D-21-00325)**.

CHAPTER 1

GENERAL INTRODUCTION AND STUDY RATIONALE

1.1 GENERAL INTRODUCTION

Next generation sequencing (NGS) is a sequencing technology that provides high throughput by rapid acquisition of thousands to millions of short nucleotides sequences (Laethem et al., 2015). Recent introduction of NGS technologies capable of producing millions of DNA sequence reads in a single run is rapidly changing the landscape of genetics, providing the ability to answer questions more rapidly than before (Mardis E. 2008). Several NGS platforms are available, and they use different sequencing technologies. For example, Illumina sequencers use sequencing by synthesis of fluorescent, reversible terminators, and Pacific Bioscience use fluorescent nucleotides in their single molecule real-time technology (Deurenberg et al., 2017).

A great advantage of NGS is that a single protocol can be used to sequence any nucleic acid. No need for target specific primers as it is the case with Sanger sequencing. Next generation sequencing can process millions of sequences reads in parallel and only little DNA input is required. In a single run, the whole genome of a microorganism can be sequenced at random. Next generation sequencing has the disadvantage of producing shorter read length (30-300 bp) which can impact the utility of data for various application. Libraries may contain errors that decrease the data quality, and thus can disrupt the data interpretation. Limitations of use of these technology in developing countries include limited molecular laboratory infrastructure, lack of skilled personnel, high capital investment and running costs.

All currently available sequencing platforms requires some level of DNA pre-processing into a library suitable for sequencing. In Illumina platforms, fragmentation and tagmentation of the genome is performed before sequencing. In this process, libraries are prepared, in which fragments of nucleic acid are fused to adaptors and barcodes to distinguish the DNA of the sequence isolates after sequencing (Deurenberg et al., 2017).

Fragmentation of nucleic acid can be performed in several ways, which can be mechanical or enzymatic. An example of a mechanical approach is the Adaptive Focused Acoustics technology followed by ligation of adaptors from Covaris (United

Sates). The transposons in Nextera XT Library Preparation Kit from Illumina is an example of an enzymatic approach. The advantage of this method is that fragmentation and fusion of the adaptor to the DNA or RNA fragments are performed in one step. More importantly, relatively low concentration of input DNA is needed to produce a library (Mardis E. 2008).

Sanger sequencing for many years has been the routine approach to detect HIV drug resistance viruses in the management of patients. A limitation of Sanger sequencing is its inability to detect viral minority drug resistant populations (Beck et al., 2014). More sensitive technology such as NGS can detect clinically relevant minority resistant populations as low as 1% of the viral population (Bensode et al., 2013).

TN5 transposase is used in the Nextera XT Library Preparation Kit, a common kit used in DNA library preparation for the Illumina platform. The cost to sequence on the Illumina platform can therefore be reduce by laboratories producing their own TN5 transposase.

1.2. STUDY RATIONALE

The most rapid and scalable NGS library preparation strategy available to date is based on a hyperactive version of the Tn5 transposase (Adey et al., 2010). Next generation sequencing is limited by its high cost of reagents to execute large scale projects. The cost of sequencing platforms has decreased but limiting factors such as costs of reagents for library preparations remains. Tn5 transposase is an enzyme that is used for library preparation, it mediates the fragmentation of dsDNA and ligates synthetic oligonucleotides at both ends. In-house cloning, expression, and purification of hyperactive Tn5 transposase can be viable way of lowering the sequencing cost in DNA library preparation. The costs of library preparation currently limit its use in small or medium sized research projects. Low costs of reagents are needed to generate sequence libraries and to harness the full potential of current NGS technology and to make it affordable for laboratory in resource limited setting. This study described (1) the trend in the uptake of NGS in HIV-1 drug resistance studies in Africa and to identify countries, institutions using the technology, and sequencing platforms being applied

(2) The expression of Tn5 transposase, and test in applicability in library preparation for NGS on the Illumina platform.

1.3. OBJECTIVES OF THE STUDY

1. To determining the trend in the uptake of NGS in HIV-1 drug resistance studies in Africa and to identify countries and institutions using the technology.

2. To over-express Tn5 transposase for NGS library preparation and sequencing of HIV and selected oncoviruses.

1.4. REFERENCES

Bansode V, McCormack GP, Crampin AC, Ngwira B, Shrestha RK, French N, Glynn JR, Travers SA (2013). Characterizing the emergence and persistence of drug resistant mutations in HIV-1 subtype C infections using 454 ultra deep pyrosequencing. *BMC Infectious Diseases*, doi: 10.1186/1471-2334-13-52.

Beck IA, Deng W, Payant R, Hall R, Bumgarner RE, Mullins JI, Frenkel LM (2014). Validation of an oligonucleotide ligation assay for quantification of human immunodeficiency virus type 1 drug-resistant mutants by use of massively parallel sequencing. *Journal of Clinical Microbiology*, doi: 10.1128/JCM.00306-14.

Mardis ER, (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, *24*(3): 133-141.

Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, García-Cobos S, Kooistra-Smid AM, Raangs EC, Rosema S, Veloo AC, Zhou K, (2017). Application of next generation sequencing in clinical microbiology and infection prevention. *Journal of Biotechnology*, *243*: 16-24.

CHAPTER 2

Application of Next Generation Sequencing in HIV Drug Resistance Studies in Africa, 2005-2019: A Systematic Review

ABSTRACT

Objective: The development of HIV drug resistance is a significant challenge in maintaining suppressed viral load in the management of patients. Next generation sequencing (NGS) is a more sensitive approach to determine the burden of HIV drug resistance. We aimed to describe the uptake of NGS in HIV drug resistance studies in Africa.

Methodology: Electronic search was done for studies published between 2005 and 2019, from three databases: Pubmed, Web of Science and Google scholar. The search approach was carried out according to PRISMA guidelines. Studies included in the analysis were those that reported the use of NGS on HIV drug resistance using samples from Africa or in which the studies were done in Africa. Only articles published in English were included in the analysis.

Results: Four thousand one hundred and eighty articles were identified according to the search criteria. Out of these, 238 studies met the inclusion criteria and were included in the analysis. Thirty studies (12.6%) used NGS, 194 (81.5%) used Sanger sequencing, and 14 (5.9%) used both techniques. Evidence of in-country application of NGS was observed in six studies (13.6%), all from South Africa. In other studies, NGS was either done outside of Africa using samples obtained from Africa or the country in which NGS was done was not indicated. From 2005 to 2012, only one study was reported on HIV drug resistance using NGS; however, 44 studies were published by 2019. Out of the 54 African countries, investigators from 13 countries (24.1%) published data on HIV drug resistance using NGS, as at end of 2019.

Conclusion: There is an uptake in the application of NGS in HIV drug resistance studies in Africa, albeit in a slower pace, with investigators from about a quarter of African countries applying the technology for this purpose.

Key Words: HIV drug resistance; Next generation sequencing; Viral suppression; Systematic review; Africa.

2.1. INTRODUCTION

Next generation sequencing (NGS), referred to as massively parallel or deep sequencing, provides a high-throughput approach by rapidly sequencing thousands to millions of short nucleotides sequences. It has had an extremely high impact on genomic research since its first introduction in the market in 2005 (Morozava and Marra, 2008). The different types of NGS platforms present different chemistries, and generate enormous amount of data per run, thus enabling large genomes to be sequenced cost effectively, and by allowing high resolution detection of rare variants (Brumme and Poon, 2017).

Throughout Africa, antiretroviral therapy (ART) is becoming more accessible, but the development of drug resistant viruses in the pre-treated or treated population poses a challenge to the maintenance of viral suppression sustainably. Resistant viruses that make up as little as 1% of the viral population within an individual have been shown to be clinically important as they can expand rapidly under selective pressure exerted by exposure to ART (Bansode et al., 2013). The presence of resistant viruses has been demonstrated to have a negative impact on the response to ART and shortens the time to the virologic failure (Gonzalez et al., 2013). For many years, Sanger sequencing has been the routine approach used to detect HIV-1 drug resistant viruses in the management of patients (Simen et al., 2014). A shortcoming of Sanger sequencing is its inability to detect minority resistant viral populations.

Drug resistance surveillance is important to inform procedures and policies in HIV treatment, and more important in low resource settings where first-line ART options are limited and second- or third-line regimens are not always available or more expensive (Inzaule et al., 2016). More sensitive detection methods are essential for a better estimate of the prevalence of drug-resistant variants and their impact on treatment and treatment options (Hauser et al., 2015; Lapointe et al., 2015). Therefore, it is important to understand the burden and spread of resistant viruses using sensitive methods to guide treatment management and monitoring (Sendagire et al., 2009).

High-throughput technologies can clearly contribute significantly to more extensive surveillance of drug resistance globally (Laethem et al., 2015). In resource-rich

settings, drug resistance testing has enabled personalized strategies for the treatment of HIV-1 infected patients to improve virologic response in patients failing ART (Inzaule et al., 2016). This is also an expectation in low resource settings. This review was aimed at determining the uptake in the application of NGS in HIV drug resistance studies in Africa.

2.2. METHODOLOGY

2.2.1. Search Strategy

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach was used, save for meta-analysis. An electronic search was carried out to identify HIV drug resistance studies in Africa. Pubmed, Web of Science and Google Scholar databases were used to search for published articles from 2005, when NGS was introduced, to 2019. Articles were searched on all three databases using the following search strategy. For Pubmed: (((((((Human Immunodeficiency Virus) OR (HIV)) AND (drug resistance testing)) AND (sequencing)) OR (next generation sequencing)) OR (ultra-deep sequencing)) OR (deep sequencing)) AND Country name. Google Scholar: “Human Immunodeficiency Virus” OR HIV AND “drug resistance testing” AND sequencing OR “next generation sequencing” OR “ultra-deep sequencing” OR “deep sequencing” AND "Country name". Web of Science: (((((((Human Immunodeficiency Virus) OR (HIV)) AND (drug resistance testing)) AND (sequencing)) OR (next generation sequencing)) OR (ultra-deep sequencing)) OR (deep sequencing)) AND Country name. All search options were accompanied with an African country (name). This was repeatedly done for all 54 African countries.

2.2.2. Inclusion and exclusion criteria

Published full-text studies, abstracts, case report and thesis on HIV drug resistance from African countries were selected for examination of their relevance. Articles included in the analysis met the following criteria: studies that have used sequencing for HIV drug resistance genotyping, studies in which samples were collected from Africa and sequencing was carried out within or outside of the African continent; studies in which sequencing was carried out in Africa but samples were not from Africa, studies published between 2005 and 2019; and in English.

2.2.3. Data Extraction

We extracted data on authors, research questions and factors influencing the problem, rationale, background, country of study participants, sample size, sample type, sequencing method, sequencing institution, results, and conclusions. Figure 1 shows the PRISMA flow diagram used in sourcing, identifying and selection of studies for the current analysis. Significant statistical differences were assessed at 95% confidence interval.

2.2.4 PRISMA flow diagram

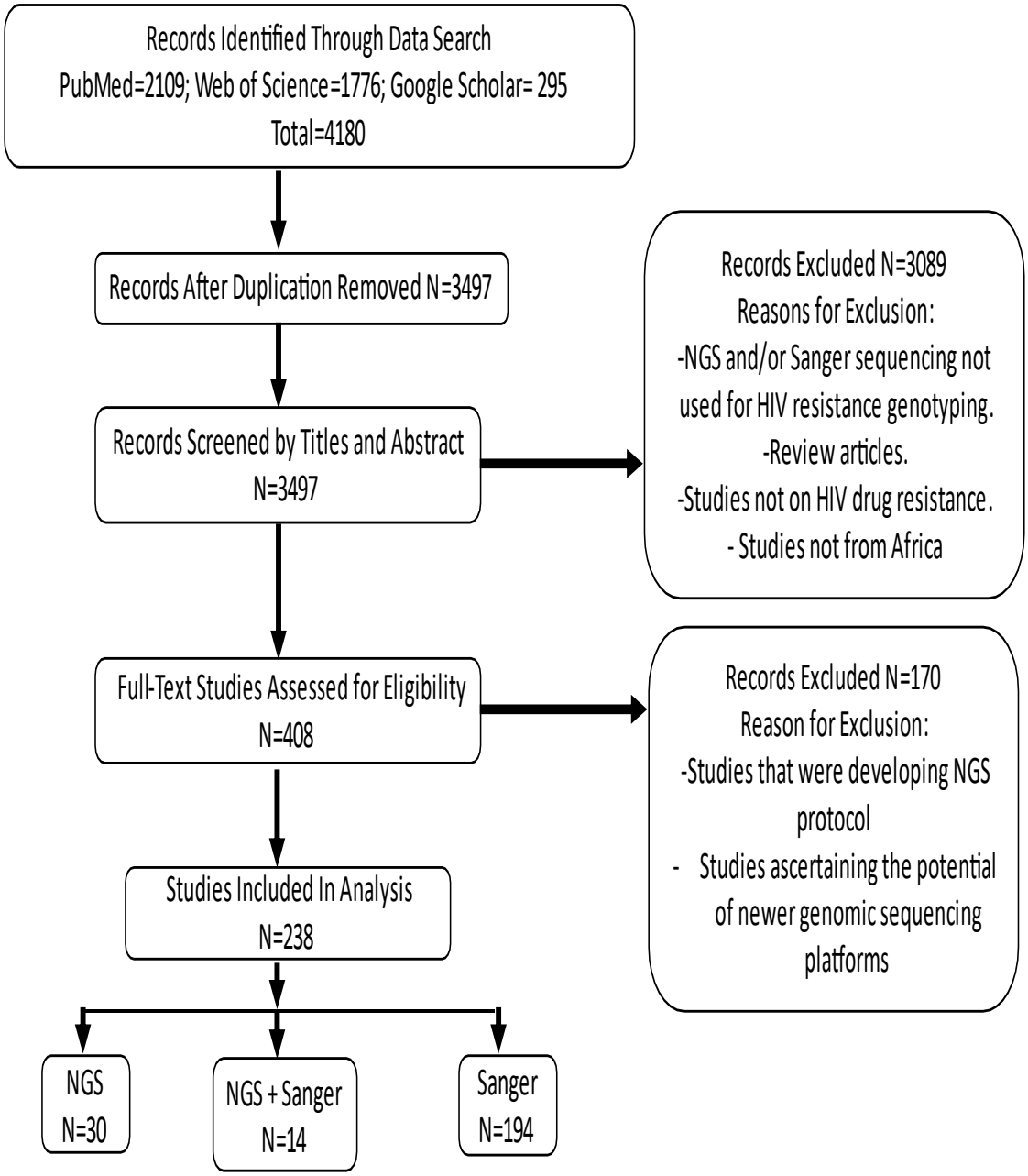

Figure 2.1: PRISMA flowchart on the selection of studies included for analysis

2.3. RESULTS

2.3.1 Studies eligible for analysis

A total of 4180 studies were using the search criteria. After eliminating duplicates, the number of studies was reduced to 3497. Studies included for analysis were research articles, abstracts, theses, case reports and research letters. Out of 3497 studies, 238 met the inclusion criteria and were further analysed. Studies that met the inclusion criteria were published between 2011 – 2019. In about 81.5% (194/238) of studies, Sanger sequencing was used, 12.6% (30/238) used NGS and 5.9% (14/238) used both NGS and Sanger sequencing.; The of 3255 studies that were excluded from analysis was because they were not on HIV drug resistance, review articles, or NGS or Sanger sequencing was not used as a sequencing method. Other studies were excluded because they were not carried out in Africa or did not use samples from Africa or did not involve African scientists.

2.3.2 Use of NGS in HIV drug resistance in Africa (2005 – 2019)

In 6 studies (13.6%) NGS was done in Africa; for seven studies (15.9%) NGS was done outside of Africa; for 31 studies (70.5%), the location where NGS was done was not specified. From 2005 to 2012 only one study was reported on HIV drug resistance using NGS in Africa, but by 2019 there were 44 studies reported. It took seven years from when NGS technology became available for the first study to be reported. The difference in the number of studies between 2005 and 2012, within which the first study using NGS was reported, and the number of studies from then up till 2019 was significant (p= 0.0014). Generally, there is an uptake in use of NGS in HIV drug resistance studies in Africa (Figure 2), but this is taking place a slow pace.

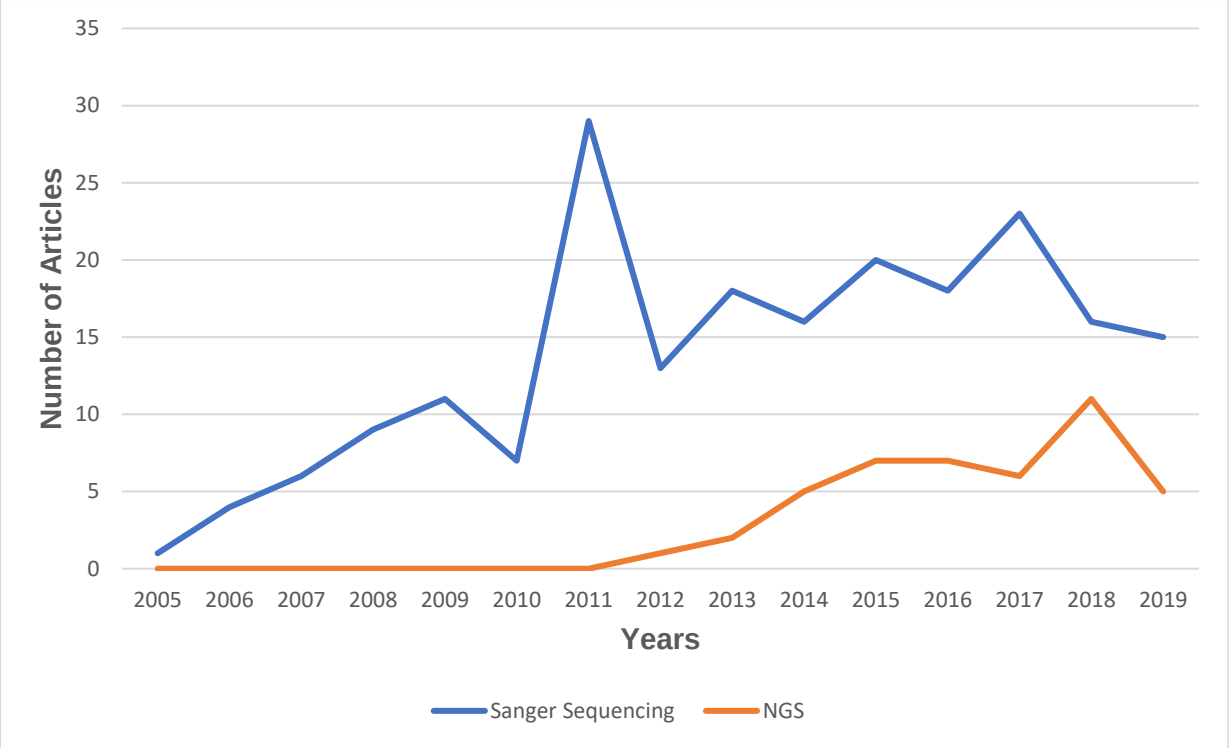

Figure 2.2: Graph showing the trend in uptake of articles published in Africa using sequencing for HIV drug resistance genotyping between 2005 – 2019 per year. Number of articles reported in a year are increasing over the years. Statistics on Sanger sequencing was included for comparative purposes.

The relative aggregations (r^2) of the trendlines for NGS shows a near linear increase in uptake (Figure 2.3).

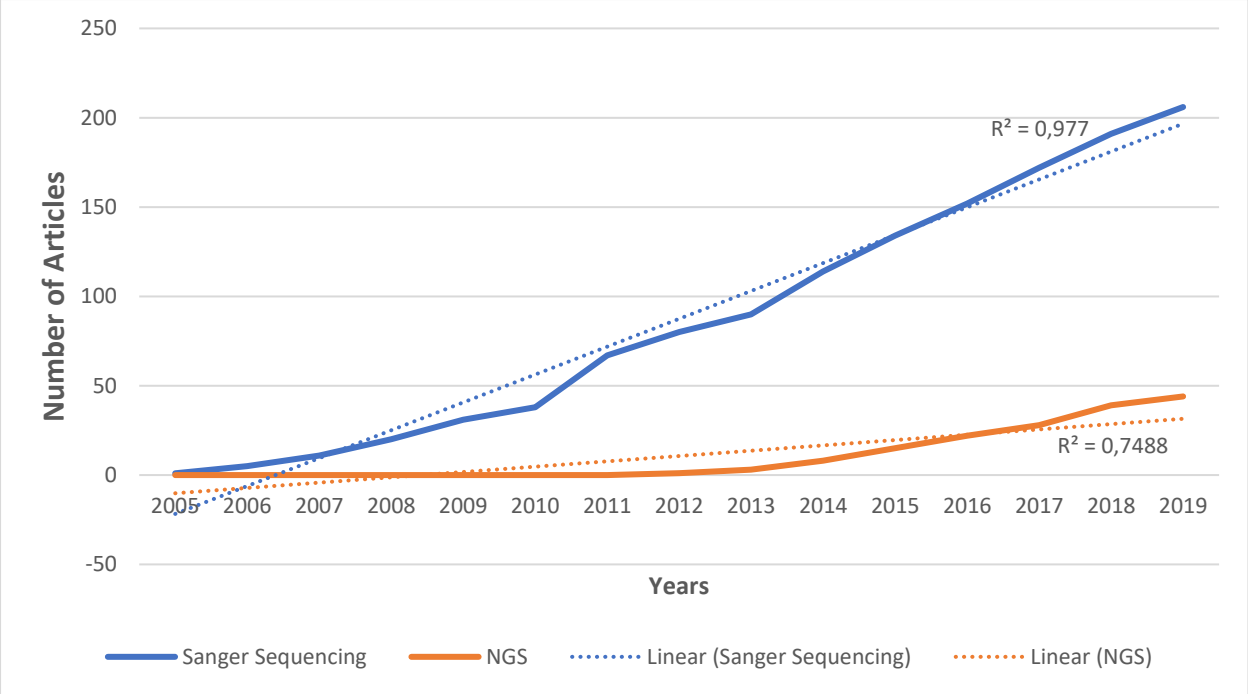

Figure 2.3: Cumulative line graph showing the trend in uptake of studies published in Africa using NGS for HIV drug resistance genotyping between 2005 – 2019. An increasing trend is observed, although at a slower rate compared to Sanger sequencing.

2.3.4 Distribution in the use of NGS in HIV drug resistance in Africa (2005-2019)

Studies were published from 13 out of the 54 African countries on HIV drug resistance using NGS (Table 1). For the 13 countries from which NGS was used, South Africa (29.5%) have the highest number of studies published, followed Uganda (13.6%). One study each (2.3%) was published from Botswana, Ghana, Mozambique, Nigeria, and Tanzania (Table 1).

Table 2.1: Proportion of studies on HIV drug resistance applying according to African countries from 2005-2019.

Countries	**Number of HIV drug resistance studies using NGS (n=44)**
Individual Countries	
Botswana	1 (2.3%)
Cameroon	4 (9.1%)
Ethiopia	2 (4.5%)
Ghana	1 (2.3%)
Kenya	5 (11.4%)
Malawi	2 (4.5%)
Mozambique	1 (2.3%)
Nigeria	1 (2.3%)
South Africa	13 (29.5%)
Tanzania	1 (2.3%)
Uganda	6 (13.6%)
Zambia	1 (2.3%)
Multiple Countries	
Kenya and Uganda	2 (4.5%)
Mozambique and Kenya	1 (2.3%)
PASER-M (Kenya, Nigeria, South Africa, Uganda, Zambia)	2 (4.5%)
AIDS Clinical Trials Group A5208 (Botswana, Kenya, Malawi, South Africa, Uganda, Zambia, Zimbabwe	1 (2.3%)

The proportion of studies published based on Africa regions, in decreasing order was as follows: Southern Africa (21/48; 43.8%), East Africa (19/48; 39.6%), Central Africa (4/48; 8.3%), and West Africa (4/48; 8.3%). No published data was available for North Africa for the period considered (Figure 4). Figure 5 shows the geographical distribution of studies of HIV drug resistance in Africa with NGS application. Out of 44 studies that used NGS, 13.6% (6/44) indicated that NGS was carried out within Africa,

29.5% (13/44) indicated that NGS was performed out of African continent, and 56.8% (25/44) did not indicate where NGS was carried out. In six studies, the institution in which NGS was applied was indicated. These includedStellenbosch University, University of the Western Cape; University of KwaZulu-Natal, Africa Health Research Institute, Africa Centre, and Inqaba Biotechnological Industries, all in South Africa.

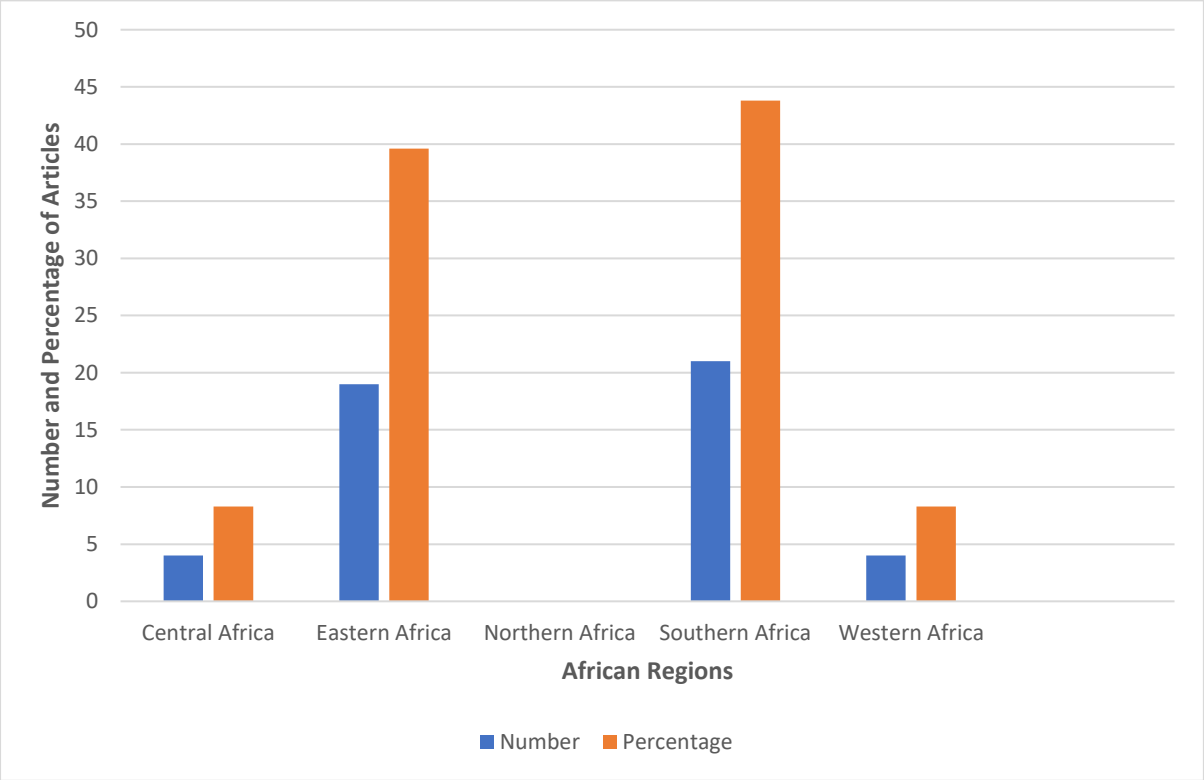

Figure 2.4: Bar graph showing numbers and percentage of articles published on HIV drug resistance using NGS according to different African regions. Southern African region has the most articles published followed by East Africa. No data for Northern Africa was identified for the period under review.

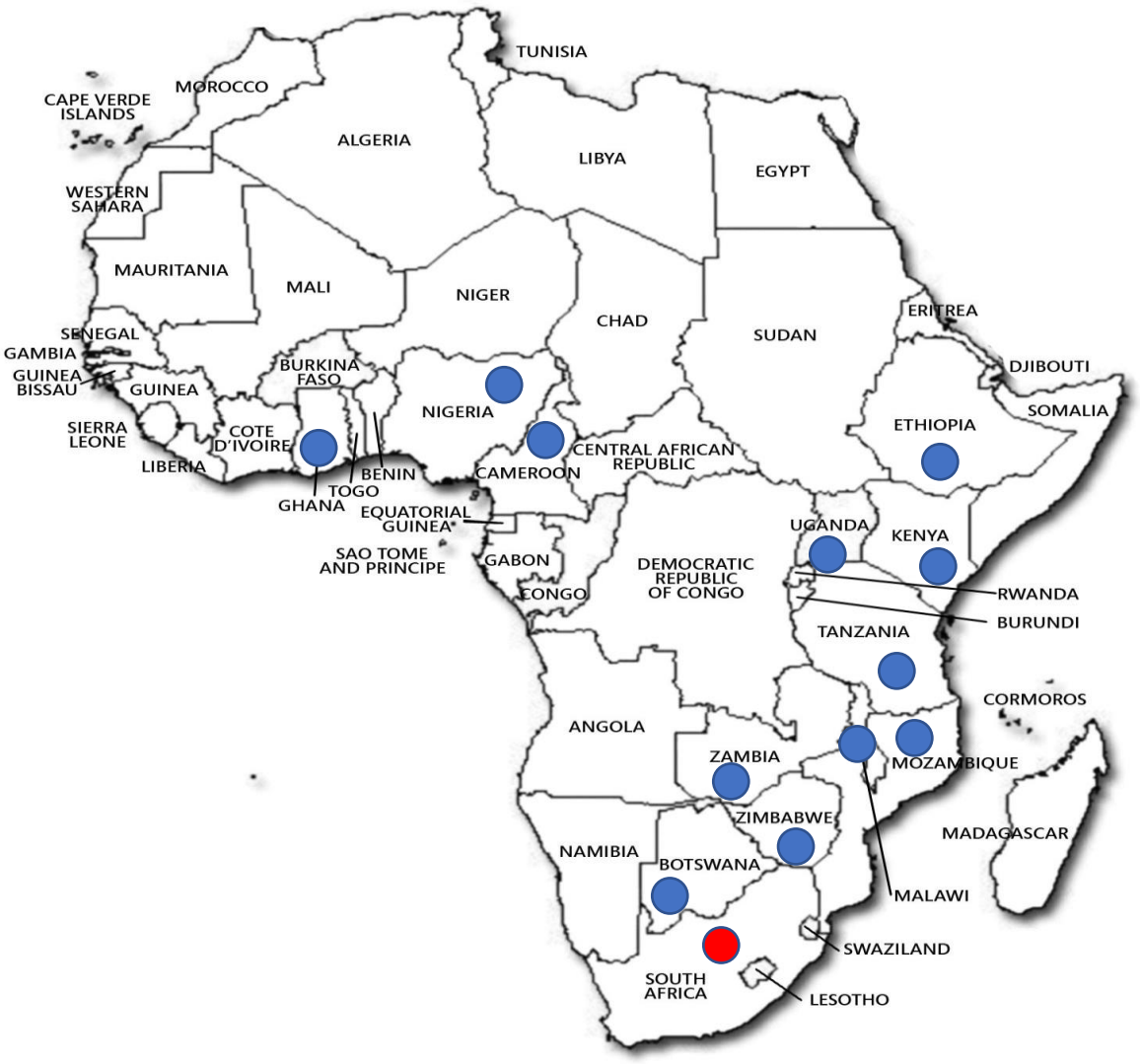

Figure 2.5: Map of Africa showing countries that published studies on HIV drug resistance using NGS between 2005 and 2019. The red circle indicates that NGS was done in-country as evident from the literature, while blue circles indicate that NGS was either done outside of the country or not evident from the literature. No data for Northern Africa was identified for the period under review.

2.3.5 Type of NGS platform used in HIV drug resistance and institutions in Africa (2005-2019)

Forty-four studies used NGS in HIV drug resistance sequencing. Of these, more than one NGS platforms were used in two studies, while one study used three different platforms. Sequencing platforms were used at different frequencies: 454 ultra-deep (39%), Illumina (50%), Ion Proton (4%), and unspecified platforms (7.0%).

2.5. DISCUSSION AND CONCLUSION

The Africa continent remains disproportionately affected by HIV, and access to treatment is expanding across the continent (UNAIDS, 2020). The development of drug resistance poses a challenge in achieving sustained viral suppression in the treatment HIV infection (Bessong et al., 2021). The application of NGS has been recommended by the WHO Working Group on HIV drug resistance because of its capability in identifying minority resistance variants of clinical importance (Mbunkah et al., 2020). However, this is not always possible in many African settings because of limited resources to acquire instrumentation and the maintenance thereof, and the lack of expertise in data analysis and interpretation. This review was aimed at determining the trend in uptake of NGS in HIV drug resistance studies in Africa and to identify countries and institutions applying the technology for this purpose.

A systematic review of the literature showed that the use of NGS in HIV drug resistance studies in Africa has increased very slowly over the years. Our analysis showed that the first study to report on HIV drug resistance from data generated through NGS was published sevens year after the NGS technology became available. Thus far, majority of studies was observed from Southern and Eastern Africa, with South Africa being the leading country. The relatively high proportion of NGS application in South Africa could be attributed to its higher economic status as an upper middle-income country compared to the others, and also perhaps because of the need to seek solutions to its higher burden of HIV infection, and as a country with one the largest HIV treatment programmes in the world. Collaboration was identified as a strategy used by many African investigators to access NGS. For example, the participation of investigators from Botswana, Malawi, Mozambique, and Zambia in multi-country studies was noted.

Several barriers, such as high capital investment, limited molecular laboratory infrastructure, cost of reagents, and lack of skilled personnel limit the use of NGS in Africa (Inzaule et al., 2016; Chimukangara et al., 2017). We propose a few approaches through which these limitations can be reduced: (1) Access to centralized laboratories can resolve the lack equipment and of skilled personnel and reduce cost through high batch sequencing. (2) The cost of the test reagents can also be significantly reduced,

for example, by in-house production, evaluation, and standardization of a key enzyme, Tn5-transposase, used in DNA library preparation (3) Establishing research collaborations and partnerships with biotechnology companies involved in NGS activities.

In conclusion, there is a slow uptake in the application of NGS in HIV drug resistance studies in Africa, despite evidence for its superior beneficial attributes for a better understanding of the burden of resistant viruses in pre-treated and treated populations.

2.5. REFERENCES

Bansode V, McCormack GP, Crampin AC, Ngwira B, Shrestha RK, French N, Glynn JR, Traver SA (2013). Characterizing the emergence and persistence of drug resistant mutations in HIV-1 subtype C infections using 454 ultra deep pyrosequencing. BMC Infectious Diseases, 13. doi: 10.1186/1471-2334-13-52.

Bessong P O, Matume N D, Tebit D M (2021). Potential challenges to sustained viral load suppression in the HIV treatment programme in South Africa: a narrative overview. AIDS Research and Therapy, 18(1), doi.org/10.1186/s12981-020-00324-w.

Brumme CJ, Poton AF (2017). Promises and pitfalls of Illumina sequencing for HIV resistance genotyping. Virus Research, 239: 97-105.

Chimukangara B, Varyani B, Shamu T, Mutsvangwa J, Manasa J, White E, Chimbetete C, Luethy R, Katzenstein D (2017). HIV drug resistance testing among patients failing second line antiretroviral therapy. Comparison of in-house and commercial sequencing. Journal of Virological Methods, 243:151-157.

Gonzalez S, Tully CD, Gondwe C, Wood C (2013). Low-abundance resistant mutations in HIV-1 subtype C antiretroviral therapy-naive individuals as revealed by pyrosequencing. Current HIV Research, 11(1):43-49.

Hauser A, Kuecherer C, Kunz A, Dabrowski PW, Radonić A, Nitsche A, Theuring S, Bannert N, Sewangi J, Mbezi P Dugange F (2015). Comparison of 454 ultra-deep sequencing and allele-specific real-time PCR with regard to the detection of emerging drug-resistant minor HIV-1 variants after antiretroviral prophylaxis for vertical transmission. PLOS One, doi: 10.1371/journal.pone.0140809.

Inzaule SC, Ondoa P, Peter T, Mugyenyi PN, Stevens WS, de Wit TFR, Hamers RL (2016). Affordable HIV drug-resistance testing for monitoring of antiretroviral therapy in sub-Saharan Africa. Lancet Infectious Diseases, 16(11), doi: 10.1016/S1473-3099(16)30118-9.

Lapointe HR, Dong W, Lee GQ, Bangsberg DR, Martin JN, Mocello AR, Boum Y, Karakas A, Kirkby D, Poon AFY, Harrigan PR (2015). HIV drug resistance testing by high-multiplex “wide” sequencing on the MiSeq instrument. Antimicrobial Agents and Chemotherapy, 59(11): 6824-6833.

Mbunkah HA, Bertagnolio S, Hamers RL, Hunt G, Inzaule S, De Wit TFR, Parades R, Parkin NT, Jordan MR, Metzner KJ, WHO HIVResNet Working Group (2020). Low-Abundance Drug-Resistant HIV-1 Variants in Antiretroviral Drug-Naive Individuals: A Systematic Review of Detection Methods, Prevalence, and Clinical Impact. Journal of Infectious Diseases, 221(10):1584-1597.

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

Sendagire H, Easterbrook PJ, Nankya I, Arts E, Thomas D, Reynolds SJ (2009). The challenge of HIV-1 antiretroviral resistance in Africa in the era of HAART. AIDS Reviews, 11(2):59-70.

Van Laethem K, Theys K, Vandamme AM (2015). HIV-1 genotypic drug resistance testing: digging deep, reaching wide? Current Opinion in Virology, 14: 16-23.

Simen BB, Braverman MS, Abbate I, Aerssens J, Bidet Y, Bouchez O, Gabriel C, Izopet J, Kessler HH, Stelzl E, Di Giallonardo F (2014). An international multicenter study on HIV-1 drug resistance testing by 454 ultra-deep pyrosequencing. Journal of Virological Methods, 204: 31-37.

UNAIDS, 2020. Global HIV and AIDS statistics – 2020 fact sheet. https://www.unaids.org/en/resources/fact-sheet. Accessed on 18 February 2021.

CHAPTER 3

OVER-EXPRESSION OF TN5 AND ITS APPLICATION IN LIBRARY PREPARATION FOR NEXT GENERATION SEQUENCING

ABSTRACT

Introduction: Next generation sequencing (NGS) provides a high-throughput approach by rapidly sequencing thousands to millions of short nucleotides sequence. Transposase 5 (Tn5) enzyme is the key reagent in library preparation for Illumina platforms, in which it mediates the fragmentation of ds-DNA and ligates synthetic oligonucleotides at both ends. The most rapid and scalable NGS library preparation strategy available to date

is based on a hyperactive version of the Tn5 transposase. NGS is a new sequencing technology, but it is limited by its high cost of reagents to execute large scale projects. This study was aimed at over-expressing Tn5 transposase for application in NGS library preparation and sequencing of HIV-1 and selected oncoviruses.

Method: Tn5 transposase was expressed in T7 express *lysY/I*^{q} Competent *E*. *coli* cells and purified using chitin column. Transposase 5 activity was tested by fragmentation of targeted DNA before it was applied in DNA library preparation. DNA library preparation was done for HIV-1, HPV and KSHV DNA using both expressed Tn5 protocol and Nextera XT library preparation kit in parallel. The libraries were then sequenced in an Illumina MiniSeq platform.

Results: Transposase 5 was successfully expressed and purified. Final concentration of purified Tn5 was 2.60 mg/ml. Expressed Tn5 was active and libraries of viral DNA were prepared. A MiniSeq run generated about 4.46 Gb data, with 44-50 million reads passing filter, 170-220 kmm^{2} cluster passing filter giving a QC30 of > 90.90%. Sequences generated from both protocols had good quality score as determined from FastQC software.

Conclusion: A procedure for over expression of Tn5 used for DNA library preparation for NGS has been demonstrated. In addition, the enzyme was shown to fragment viral DNA, although this activity was not seen to be optimal to generate a high number of NGS sequence reads

Key Words: Tn5 transposase, Next Generation Sequencing, Library Preparation

3.1. BACKGROUND

Protein expression is the way in which protein are synthesized, modified, and regulated in living organisms. Synthesis and regulation of proteins in the cell depends on the functional need for the protein. DNA contains the blueprint of protein. These blueprints are undecoded by a regulated process called transcription in which they are decoded into messenger RNA (mRNA). Messenger RNA are then translated into proteins through the process called translation. Both prokaryotic and eukaryotic protein expression occur in these two processes (Overton, 2014).

There are different types of systems that are used to express proteins, these include bacterial protein expression, insect protein expression, mammalian protein expression and cell free protein expression systems (Imataka and Mikami., 2009). It is important to choose the right type of protein expression system for the specific application for successful recombinant protein expression. The type of protein, the requirements for functional activity and the desired yield are some of the factors that are considered when selecting the system (Mikami et al., 2008). Traditional strategies for recombinant protein expression involve transfecting cells with DNA vector that contains the template and then culturing the cells so that they transcribe and translate the desired protein (Mikami et al., 2008). Proteins are expressed for different reasons. When researchers are investigating protein, they usually produce a functional protein of interest. Some of the reasons why researchers express protein involves investigating different aspects of the protein such as function, structure, localization, protein interaction or modification. In this study, Tn5 transposase was expressed and used for DNA library preparation for NGS.

3.2. LITERATURE REVIEW

3.2.1. STRUCTURE OF TN5 TRANSPOSASE

Tn5 transposase is a member of the 154 family of transposases that contain an YREK motif around the E of the DDE residues (Reznikoff, 2003). Tn5 transposase is 476 residues long (Steiniger et al., 2004). It is a prokaryotic composite transposon that consist of two insertional sequences, IS50R and IS50L, flanking a region of DNA containing three antibiotic resistance genes. Tn5 transposase consist of three distinct domains, N-terminal, catalytic, and C-terminal. The N-terminal domain is composed of only α helices and turns that are used for DNA binding. The N-terminal domain consists of the first 70 amino acids. The active site of Tnp is found within the 300 amino acids of the catalytic domain. This domain contains a ribonuclease H like motif, and α/β/α fold with a mixed β sheet of five strands in which strand 2 is antiparallel to the other four strands (Murzin et al., 1995). The C-terminal is primarily responsible for protein-protein interaction and is comprised entirely of α helices and turns (Steiniger et al., 2004)

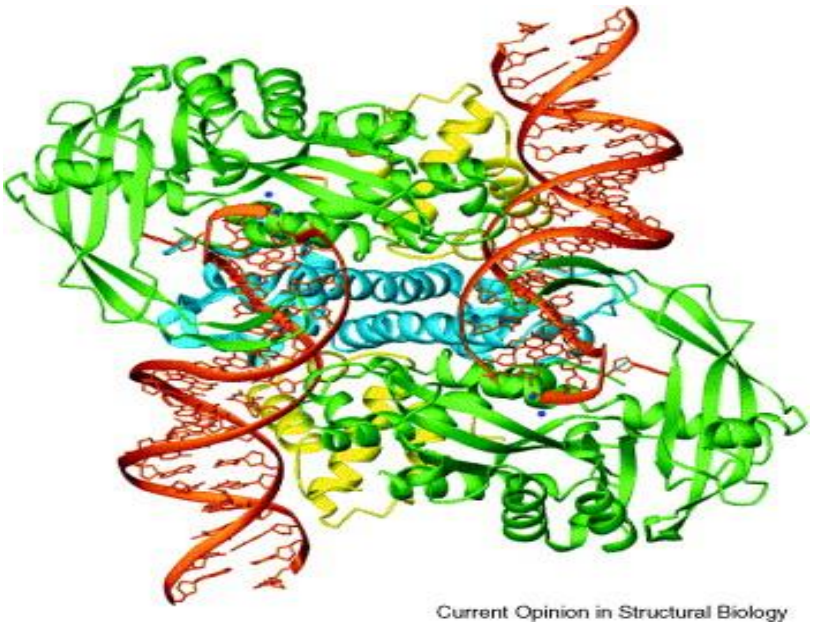

Figure 3.2.1: Tn5 transposase structure (Steiniger et al., 2004).

3.2.2. EXPRESSION SYSTEMS

Natural sources of proteins rarely meet the requirements for quality, ease of isolation and cost of protein purification (Sorensen and Mortense, 2005). Instead, living cells and their cellular machinery are usually harnessed as factories to build and construct proteins based on supplied genetic template. Recombinant cell factories are used constantly for production of protein preparation bound for downstream purification and processing (Sorensen and Mortense, 2005). The main purpose of expression of recombinant protein is often to obtain a high degree of accumulation of soluble product in the expression system (Sorensen and Mortense, 2005). Before the recombinant protein is expressed, additional information is required, which includes, lysis of the cell, purification, and analysis of recombinant protein (Hunt, 2005). Challenges encountered during recombinant protein expression include cellular stress response and accumulation of target protein into soluble aggregates known as inclusion bodies (Sorensen and Mortense, 2005). Inclusion bodies generally contains proteins that are misfolded and are biologically inactive.

There are different strategies of recombinant protein expression depending on the expression system. Traditional strategies involve transfecting a cell with a DNA vector containing the template of interest and then culturing the cells to transcribe and translate the desired protein. After translation of desired protein, the cells are lysed to extract the expressed protein for subsequent purification and processing.

The type of a protein that is being expressed determines the expression system and the requirements for functional activity, and the desired yield play a role. Different expression systems have their own advantages and disadvantages in relation to the cost, ease of use and their post translational modification profile. There is a wide selection of expression systems available as described below.

3.2.2.1. Bacterial protein expression

Bacterial expression is the most common expression system used to produce recombinant proteins (Hunt, 2005). Bacterial expression systems are commonly employed because they are easy to culture, their ability to grow rapidly and produce high yields of recombinant protein. *Escherichia coli* is the popular bacterial protein expression system. Since bacterial expression system are prokaryotes, eukaryotic

protein that are expressed using this system are often non-functional because they are not equipped to accomplish the required post translational modification (Andersen and Krummen, 2002). Another limiting factor is the inability of *E. coli* to secrete the protein products (Chen, 2012).

3.2.2.2. Mammalian protein expression

Mammalian protein expression systems are used to produce eukaryotic protein that have the most native structure and activity due to its physiological relevant environment (Imataka and Mikami, 2009)**.** This type of expression can do proper protein folding and post translational modification to the expressed protein (Khan, 2013). Mammalian protein expression systems are coupled with more demanding culture conditions (Imataka and Mikami, 2009). Expression of recombinant protein in mammalian cells requires a suitable cell line and appropriate vectors that must act as transport vehicle to carry the gene of interest into the required cell line of the expression system.

3.2.2.3. Insect protein expression

Insect protein expression is like mammalian systems in terms of modification and can be used for high level protein expression. There are several systems that can be used to produce recombinant baculovirus, which can then be utilized to express the protein of interest in insect cell. The most popular vehicle to produce large quantities of recombinant protein for functional and structural studies is baculovirus-mediated insect cell expression (Hunt, 2005). Insect cells have been used in different protein expression applications, particularly related to high throughput expression of sequence for functional screening (Andersen and Krummen, 2002).

3.2.2.4. Cell free protein expression

Cell free protein expression is the system that express protein in vitro using translation compatible extracts of whole cell. Cell lysate containing the whole cell extract are needed for transcription and translation, and even post translational modification. Proteins are produced much faster than those produced in vivo since it does not require time to culture the cells. A disadvantage is that it is more expensive to produce recombinant protein using this type of expression system.

3.2.3. PROTEIN PURIFICATION

Protein purification is the process where by one or few proteins are separated from a complex mixture, usually cells, tissues, or whole organisms. Protein purification process usually separates proteins in protein sizes, physico-chemical properties, binding affinity, and biological activity. Protein production and purification of recombinant protein are intimately linked. The choice of protein expression system affects the amplification and isolation of the protein, but also the technique in which the product produced can be subsequently purified (Ana et al., 2014). Chromatography is one of the techniques that is used to purify protein, it is a physical method that separates components between two phases, one stationary phase and one mobile phase. Chromatography is a well-established technique in the purification of proteins, and it is considered inexpensive and purifies proteins in high yields recovery at high purities. Due to high recovery yields and high purity, affinity chromatography is one of the most efficient strategies in purification of recombinant proteins.

3.2.4. DNA LIBRARY PREPARATION

DNA library preparation is the process whereby a DNA of an organism is fragmented into short lengths. Tn5 transposase mediates the fragmentation of dsDNA and ligates synthetic oligonucleotides at both ends of fragmented DNA (Adey et al., 2010). Tn5 transposase has improved turnaround time and reduce DNA input and increase throughput unlike most common library preparation methods. Common library preparation methods follow the basic procedure with minor variations. Basic procedure is long, and it includes several steps: First, sonication is used to fragment DNA or nebulization or shearing to desired sizes. Then, the fragmented DNA is repaired and end-polished including blunt-end and A-tailing. The final step includes ligation of specific adaptors to DNA libraries (Head et al., 2015; Dijk et al., 2014). Significant amount of sample is lost when using this method and high amount of DNA is required, over 200 ng.

Newly developed methods take advantage of an in-vitro transposition reaction. Using Tn5 transposase and a free transposon end that contains a transposase recognition site mosaic end (ME) and the sequencing adaptors to form a transposome complex (Caruccio, 2011). Targeted DNA is fragmented when target dsDNA is incubated with transposome complex. Sequenceable DNA libraries results when the transferred

strand of transposon end including the sequence adaptor is covalently attached to the 5' end of the target fragment. Multiplexed sequencing on a single instrument run can be done by incorporating barcodes on the libraries and thus significantly saving cost. Figure 1.2 shows the DNA library preparation for NGS.

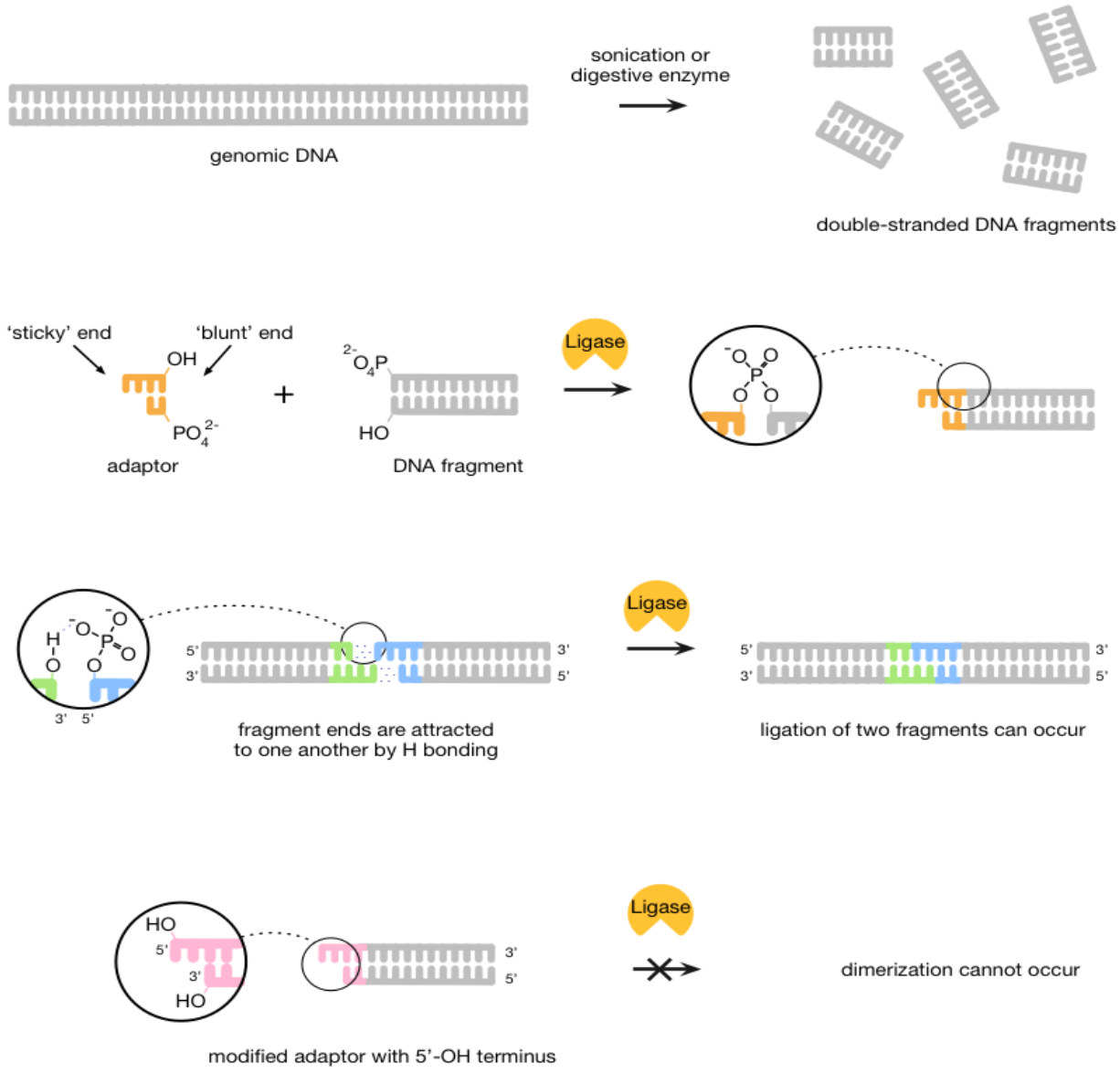

Figure 3.2.2: DNA library preparation for NGS (Ahmadian and Svahn., 2011)

3.2.5. NEXT GENERATION SEQUENCING

Next generation sequencing methods have revolutionized biology by providing detailed insights into genome organization, the regulation of gene expression and genetic variation (Bianca et al., 2018). With its ultra-high throughput, scalability, and

speed, NGS enables researchers to perform a wide variety of applications and study biological systems at a deeper gene level than ever possible. Next generation sequencing can sequence larger genomes within a single day compared to Sanger sequencing which requires over a decade to deliver a final draft of the sequences

All NGS platforms perform sequencing of millions of small fragments of DNA in parallel. Next generation sequencing captures a broad spectrum of mutations than Sanger sequencing (Behjati and Tarpey, 2013).

3.3. SPECIFIC OBJECTIVES

The specific objectives of this study were:

1. To express Tn5 transposase in T7 express *lysY/I^q* competent *E. coli*
2. To evaluate the expressed Tn5 transposase through fragmentation of HIV, HPV and KSHV DNA
3. To evaluate the sequences generated with the expressed Tn5 transposase and commercial Illumina transposase for reproducibility

3.4. MATERIALS AND METHODS

3.4.1. Ethical Consideration

The investigation reported here-in are a component of a larger study that was approved by the Health and Clinical Trials Research Ethics Committee of the University of Venda (SMNS/18/MBY/06)

3.4.2. Plasmid Verification

pTXB1-Tn5 plasmid, an IPTG inducible expression of transposon Tn5 fused to Mxe intein and chitin binding domain, was obtained from Addgene (USA) in stab culture format. A stab containing the plasmid was cultured on agar plate containing ampicillin (100 µg/mL) overnight at 37°C. Single colonies were inoculated into 4 mL Luria Bertina (L.B) broth medium and incubated overnight at 37°C on a shaking platform. Minipreps were prepared from culture is required soon thereafter or, the bacterial cells were aliquoted and stored in 50% glycerol at -80°C for subsequent experiments.

A miniprep was done in order isolate plasmid DNA. An Aliquot of 1 mL overnight culture was used for Miniprep using QIAprep Spin Miniprep Kit, following the manufactures instructions (QIAprep Spin Miniprep Kit). Following a miniprep, the presence of the plasmid was verified by way of two methods: restriction fragment length polymorphism and sequencing.

3.4.2.1 Restriction Digestion of Plasmid DNA

Two restriction enzymes, ApaI and BamHI, were used to digest the pTXB1-Tn5 plasmid. Restriction site of ApaI is at 1414 bp and BamHI is at 5099 bp within the pTXB1-Tn5 plasmid genome. In the first reaction, pTXB1-Tn5 plasmid was digested by each enzyme independently. In the second reaction, a double digestion was carried out with pTXB1-Tn5 plasmid exposed to both enzymes simultaneously in the same tube. All reactions were carried out in a 20 µL reaction. The reaction mix comprised of: 1µL of 1 µg pTXB1-Tn5 plasmid DNA, 2 µL 10X buffer (FastDigest, Thermo Scientific), 1µL each restriction enzyme (FastDigest, Thermo Scientific), and the reaction volume made up to 20 µL with sterile distilled water. The reaction mix was incubated at 37°C for 1 hour. Digestion was stopped by heat inactivation at 65°C for 15 minutes. One percent agarose gel electrophoresis was done to visualize the digestion outcomes.

3.4.2.2. Sequence Analysis of the Plasmid

Transposase 5 gene within the plasmid was sequenced by Sanger sequencing. Sequencing was done to confirm if the plasmid contained Tn5 gene. Two primers were used, Tn5ME-A: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG and Tn5ME-B: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG; to generate a sequence length of 850 bp.

3.4.3. Expression of T7 express *lysY/I^q* Competent *E. coli* cells

Twenty millilitres of T7 express *lysY/I^q* Competent *E. coli* cells were cultured overnight at 37°C. A colony was inoculated into 2 mL L.B broth medium without antibiotics, incubated overnight at 37°C, shacking at 222 rpm. Overnight 2 mL culture was inoculated into 50 mL of L.B broth with shaking at 222rpm. Cells were harvested at D_{600} of 0.324 (Log Phase). The culture was chilled on ice for 15 minutes. Culture was centrifuged for 10 minutes at 4000 rpm. Supernatant was discarded and the pellet containing bacterial cells was resuspended in 20 mL ice cold 0.1M $CaCl_2$. Cells were kept on ice for 30 minutes. Cells were centrifuged for 10 minutes at 4000 rpm. Supernatant was discarded and cell pellets were resuspended in 6 ml of 0.1 M $CaCl_2$ solution with 15% glycerol and stored at -80˚C.

3.4.4. Transformation of *E. coli* Competent Cells by pTXB1-Tn5

pTXB1-Tn5 plasmid was transformed in T7 express *lysY/I^q* competent *E. coli*. Five microliters of plasmid DNA was added into a sterile 1.5 mL tube on ice and 50 µl of T7 express *lysY/I^q* Competent *E. coli* cells were added to the tube. The contents were mixed by flicking the tube and placed on ice for 30 minutes. For cells to take in the pTXB1-Tn5 clone, the mixture was heat-shocked by placing the tubes immediately from ice to a heating block at 42°C for 10 seconds. After heat-shocking, the tube was placed on ice for 5 minutes. Nine hundred and fifty microliters of room temperature Super Optimal broth with Catabolite repression (SOC) medium were added into the tubes containing cells transformed with pTXB1-Tn5 clone. The reaction mixture was then incubated at 37°C for 1 hour while shacking at 250 rpm. After incubation, 20 µL of each transformed culture was plated into L.B agar plates containing ampicillin (100 µg/mL) and incubated overnight at 37°C.

3.4.5. Colony PCR

Heat-boiling method was used to extract plasmid DNA from transformed cells. A bacterial colony was picked and swirl it into 25 µL of dH_2O in a microcentrifuge tube. The mixture was boiled at 100°C for 5 minutes. The mixture was spin down for 2 minutes at high speed in centrifuge. Twenty microliters of the supernatant were transferred into a new microcentrifuge tube. Two microlitres of the supernatant was used as template in a 25 µL PCR.

One round PCR was done to amplify Tn5 gene from transformed cells with pTXB1 plasmid. Amplified region was a partial Tn5 gene of 850 bp, using primers in Table 2.1. PCR was carried out in 25 µL reaction. The PCR reagents used were: 10X standard Tag (Mg-free) reaction buffer, 25 mM $MgCl_2$, 10 mM dNTPs, and 10 µM of each primer. The final concentration of PCR reagent in PCR reaction in 25 µl was 10x Standard Tag (Mg-free) reaction buffer, 1.5 mM MgCL2, 0.2 µM of each primer, 200 µM of dNTPs, 1.25 units/50 µL Taq DNA polymerase. PCR cycling conditions was as follow: initial denaturation for 30 seconds at 95°C, 30 cycles of 95°C for 30 seconds, 63°C for 30 seconds, 72°C for 1 minutes, final elongation for 7 minutes at 72°C

Table 3.1: Primer sets for amplifying partial Tn5 gene

PRIMER NAME	SEQUENCE (5' → 3')
Tn5ME-A	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Tn5ME-B	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

3.4.6. Tn5 Expression in T7 Express lysY/Iq Competent E. *coli*

A pre-culture for expression was prepared by inoculating a single colony from transformation culture plates into 5 mL L.B broth containing ampicillin (100 µg/mL) and grown overnight at 37°C, shacking at 250 rpm. Five millilitres of overnight pre-culture were added into 1L of L.B broth with ampicillin and grown at 37°C and shacking at 200 rpm. The culture was grown until the OD_{600} reached 0.6 to 0.9 (about 3 hours). After the culture reached the desired OD_{600} value, the culture was chilled to 10°C. Transposase 5 expression was induced by adding Isopropyl-β-D-1-thiogalactopyranoside (IPTG) to the final concentration of 0.25 mM into 1L culture. After induction, the culture was additionally grown at 30°C for 6 hours shacking at 200

rpm. Before induction, 1 mL was removed from the culture and during expression, 1 mL of the culture was removed at 1-hour interval and placed into microcentrifuge tubes. After 6 hours of expression, cells were harvested by centrifugation at 5000 rpm, 4°C for 10 minutes. Supernatant was discarded and the pellet was stored at -80°C overnight. One millilitres of culture samples collected were analysed on sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE). Expression was confirmed by analysing the 1 mL culture samples aliquoted one hour before and six hours during expression on SDS-PAGE.

3.4.7. Cell Lysis and Protein Purification

Cell pellets were thawed on ice with 20 mL HEGX (20 mM HEPES-KOH at pH 7.2, 0.8 M NaCl, 1 mM EDTA, 10% glycerol, 0.2% Triton X-100) buffer. One protease inhibitors tablet was added on cell pellets. Cell pellets were lysed by sonication with pulse: on for 15 seconds, off for 30 seconds and AMP of 30%. Lysate were centrifuged at 5300 rpm for 30 minutes at 4°C. Pellets were discarded. One point six millilitres of neutralized polyethyleneimine (PEI) were added dropwise on a magnetic stirrer on the supernatant. To pellet the DNA, supernatant was centrifuged at 5300 rpm for 20 minutes at 4°C. Supernatant was kept on ice and pellets discarded.

Chitin column was prepared by adding 10 mL of chitin resins into a 50 mL purification column. Once the chitin resins settled on the column, the column was washed by 50 mL of HEGX buffer. To bind Tn5 with chitin beads, supernatant from the cell lysis was loaded to the column and allowed to bind for 30 minutes at 4°C with agitation. The column was washed three times with 25 mL of HEGX buffer. After washing, 24 mL HEGX buffer with 100 mM DTT was added to the column. The column was left closed for 36 hours at 4°C while shacking to effect cleavage of Tn5 from chitin beads. Transposase 5 was eluted from the column.

3.4.8. Bradford Assay

Eluted aliquots were tested for protein concentration using a Bradford assay (Bio-Rad protein assay). One microliter of each fraction eluted was added to 50 µL of assay solution. The fraction with the strongest blue colour were pooled together and concentrated using MacroSep Advance Centrifugal Device 10K MWCO (Pall Life Science, USA). Concentrated Tn5 was quantified using Qubit instrument (Invitrogen).

3.4.9. Tn5 Assembly with Mosaic End Double Stranded (MEDS)

Transposase 5 was assembled with preannealed MEDS oligonucleotides in solution. Tn5 assembly was done by mixing 0.5 vol of 50 µM MEDS (A or B) in Tris and EDTA (TE) buffer, 1 vol of 80% glycerol, 0.5 vol of 50 µM Tn5 = A280 = 2.60 mg/ml. The solution was mixed and incubated at 37°C for 60 minutes, shacking. After incubation, Tn5+A-MEDS and Tn5+B-MEDS solutions were mixed to get 12.5 µM A/B-MEDS Tn5. The solution was stored at -20°C.

3.4.10. Dilution of Tn5

Once Tn5 have been assembled to MEDS, different concentration of Tn5 were prepared. Different concentrations were prepared to determine the desired concentration of Tn5 that will produce results of good quality. Concentration obtained after assembly was 2.065 mg/ml. concentrations that were used are 0.65, 0.37, 0.26, 0.13, 0.076, and 0.026 mg/ml.

3.4.11. Tn5 Activity

Assay of Tn5 activity reaction was assembled by mixing 2 µL buffer (10 mM Tris-HCl pH 7.5, 10 mM $MGCl_2$, and 25% dimethylformamide), 12 µL H_2O, 1 µL of DNA (10 ng/µL) and 1 µL A/B - MEDS Tn5. The reaction was incubated at 55°C for 7 minutes. Tn5 reaction was stopped by adding 5 µL of 0.2% SDS and incubated again at 55°C for 7 minutes. Eleven microliters of Tn5 fragmentation reaction were used directly in the subsequent enrichment PCR amplification of libraries for the Illumina Sequencer. Enrichment PCR was done using 2X KAPA 2G Robust HotStart Ready Mix (KAPA Biosystems). The reaction contained 11 µL of the fragments, 15 µL of 2X KAPA 2G Robust HotStart Ready Mix and 5 µL of each Index (Index 1 (i7) and Index 2 (i5)). Cycling conditions of the enrichment PCR are in table 1. E-gel electrophoresis (Thermo Fisher) was used to analyze the samples.

3.4.12. Fragmentation and Tagmentation Using In-House Tn5 Transposase

Purified HIV, HPV and KSHV DNA (10 ng/µL) using AMPure XP beads was used for fragmentation and tagmentation reaction, carried out using in-house Tn5 transposase. In fragmentation reaction, 2 µL buffer (10 mM Tris-HCl pH 7.5, 10 mM $MGCl_2$, and

25% dimethylformamide), 12 µL H_2O, 1 µL of DNA (10 ng/µL) and 1 µL A/B - MEDS Tn5 was mixed. The reaction was incubated at 55°C for 7 minutes. Tn5 reaction was stopped by adding 5 µL of 0.2% SDS and incubated again at 55°C for 7 minutes. Eleven microliters of Tn5 fragmentation reaction were used directly in the subsequent enrichment PCR amplification of libraries for the Illumina Sequencer. Enrichment PCR was done using 2X KAPA 2G Robust HotStart Ready Mix. The reaction contained 11 µL of the fragments, 15 µL of 2X KAPA 2G Robust HotStart Ready Mix and 5 µL of each Index (Index 1 (i7) and Index 2 (i5)). Cycling condition used were the same as those for Nextera XT DNA Library preparation kit. Then libraries were pooled for sequencing on the MiniSeq instrument.

3.4.13. Different ratio of AMPure XB Beads

Since low concentrations of libraries were being obtained. Different ratios of beads were used to prevent losing libraries during washing step. Ratio that was used to purify the libraries were 1:0.5, 1:1, 1:2. AMPure XB beads were using following the manufactures protocol

3.4.14. Normalize Libraries

Amplified libraries were normalized to 1 nM before pooling the libraries. To normalize libraries, a formula below was used to calculate DNA concentration in nM.

$$\text{Concentration in nM} = \frac{concentration\ in\ ng/\mu L}{660\ g/mol \times average\ library\ size} \times 10^6$$

After normalizing to 1 nM, 5 µL of each normalized library were pooled into one 2 mL micro centrifuge tube. Libraries were denatured and diluted to 1.8 pM and spiked with 20% of Phix (Illumina). Libraries were subjected to sequencing using MiniSeq instrument (Illumina).

3.4.15. Tagmentation Reaction and Final PCR Amplification Using Nextera XT DNA Library Preparation Kits

Fragmentation and tagmentation were carried out with Nextera XT DNA library preparation kit (Illumina) to be able to compare the performance of the commercial kit with the in-house transposase 5. When using Nextera XT DNA library preparation kit, 5 µL of 0.2 ng/µL of purified DNA, 10 µL of tagment DNA buffer (TD) and 5 µL of amplicon tagment mix (ATM) was added to a final volume of 20 µL to fragment the input DNA. The solution was incubated for 5 minutes at 55°C. to stop the reaction, 5 µL of neutralize tagment buffer (NT) was added to inactivate DNA fragmentation. Fragmented libraries were amplified by adding 5 µL of each Index 1 adapters, 5 µL of each Index 2 adapters, and 15 µL of Nextera PCR Master mix (NPM) into fragmented libraries. The PCR program was as follows: 3 min at 72°C, 30 sec at 95°C, and then twelve cycles of 10 sec at 95°C, 30 sec at 55°C, 30 sec at 72°C, and 3 min at 72°C.

3.4.16. Sequence Analysis

As part of data processing steps and end paring, sequences were de-multiplexed automatically, on a MiniSeq. Fastq files were generated for each sample representing the two paired end reads. FastQC software was used to validate sequence quality. Generated fastqc files were imported into Geneious software version 2019.2.3 for downstream analysis. Sequence reads were cleaned up by trimming regions with poor quality. Sequence reads for each sample were mapped to their reference sequence to determine the coverage of generated sequence reads.

3.5. RESULTS AND DISCUSSION

3.5.1 Plasmid Verification

3.5.1.1 Plasmid Digestion

pTXB1-Tn5 plasmid was digested by two restriction enzymes, BamH1and Apa1. pTXB1-Tn5 plasmid is circular and 9 Kb in size. Single and double plasmid digestions were done. It was observed that digestion with single restriction enzyme linearize the plasmid when viewed on agarose gel and the plasmid band size was 9 kb. In the double digestion, a single band was observed which was approximately 4.5 Kb. This is because when these two restriction enzymes digest, two bands of the same size were obtained. The first restriction enzyme linearizes the plasmid and the second restriction enzyme digest the plasmid at approximately 4.5 kb.

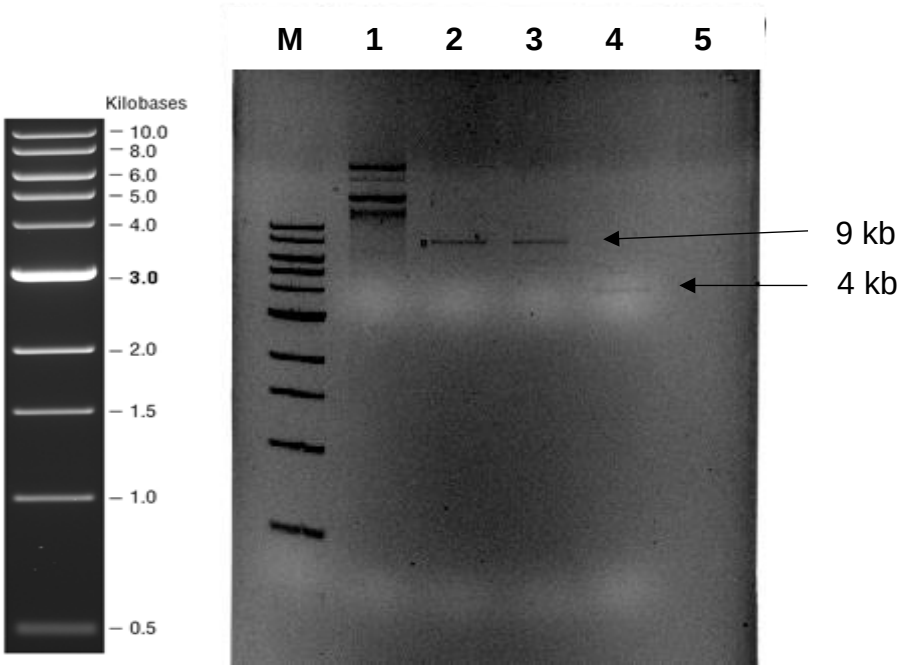

Figure 3.5.1: pTXB1 plasmid digestion using three restriction enzymes. M is the Quick Load 1 kb DNA ladder, 1- is the undigested circular plasmid, 2 - digested plasmid by Apa1, 3 - digested plasmid by BamH1, 4 - double digestion by Apa1 and BamH1. A circular plasmid was linearized by single digestion and cut into half of approximately 4.5 Kb each by double digestion.

3.5.1.2 Tn5 Sanger Sequencing

Tn5 gene within pTXB1 plasmid was sequenced by Sanger sequencing. Primers designed to sequence Tn5 gene covered partial portion of 850 bp. Forward and reference sequence were aligned on pTXB1 Tn5 plasmid reference sequence, and both aligned within Tn5 gene where it was located on the pTXB1 Tn5 plasmid reference sequence. Forward and reverse sequences aligned to the location of Tn5 gene on the pTXB1 reference sequence are shown in Figure 3.5.2.

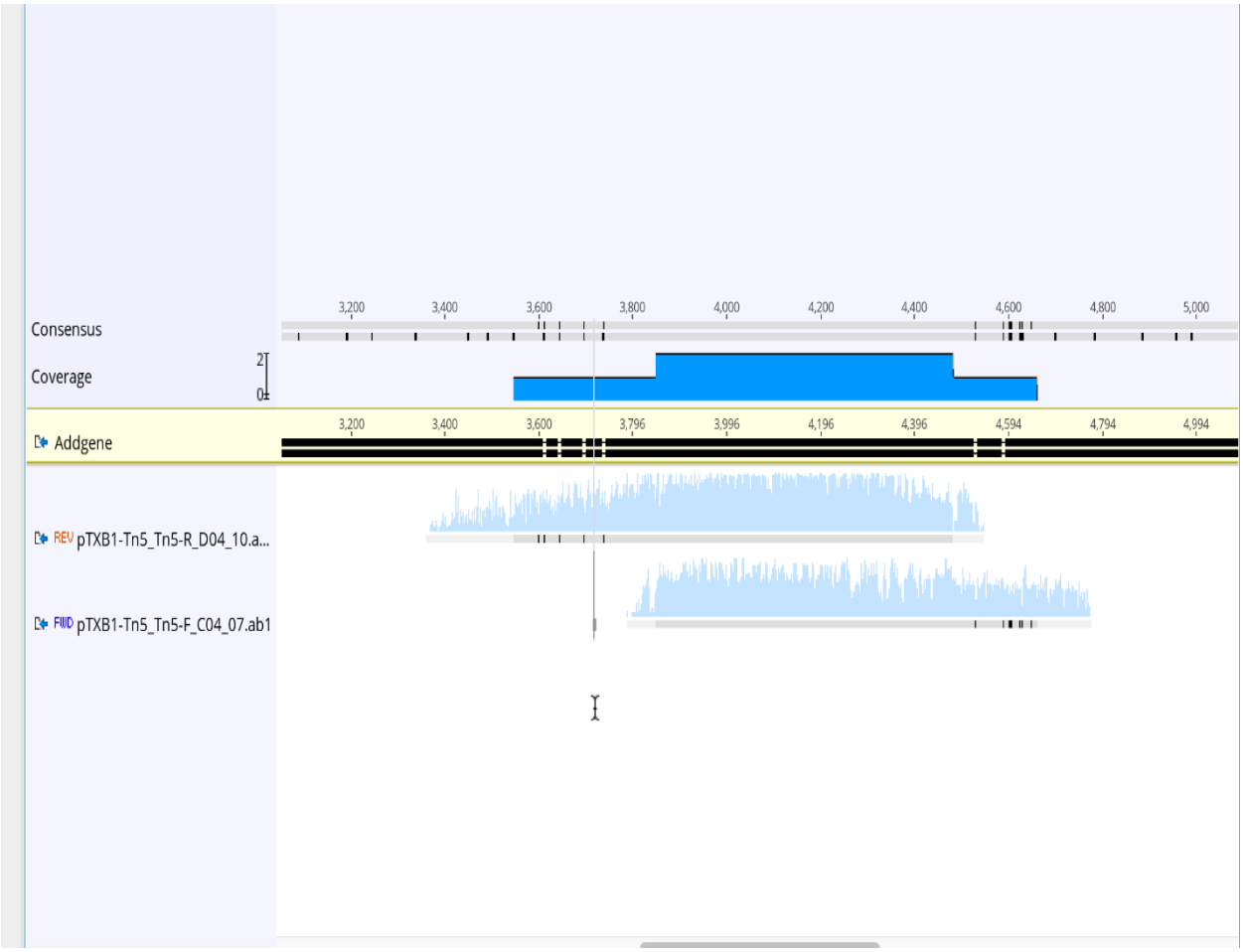

Figure 3.5.2: Partial Tn5 gene of 850 bp sequences mapped to pTXB1-Tn5 plasmid full genome. Forward and reverse sequences aligned to the location of Tn5 gene on the plasmid

3.5.2 Transformation of T7 express *lysY/I*^{q} Competent *E. coli* with pTXB1-clone

In-house T7 express *lysY/I*^{q} competent *E. coli* cells were successfully produced from T7 express *lysY/I*^{q} Competent *E. coli* cells stock from BioLabs Inc. After it was confirmed that the pTXB1-Tn5 plasmid contains Tn5 gene, In-house T7 express *lysY/I*^{q} competent *E. coli* cells were subsequently transformed by pTXB1-Tn5 plasmid. Figure 3.5.3 shows the expected cream white colonies appearing on (100 µg/ml) ampicillin, LB agar plates after an overnight incubation at 37°C

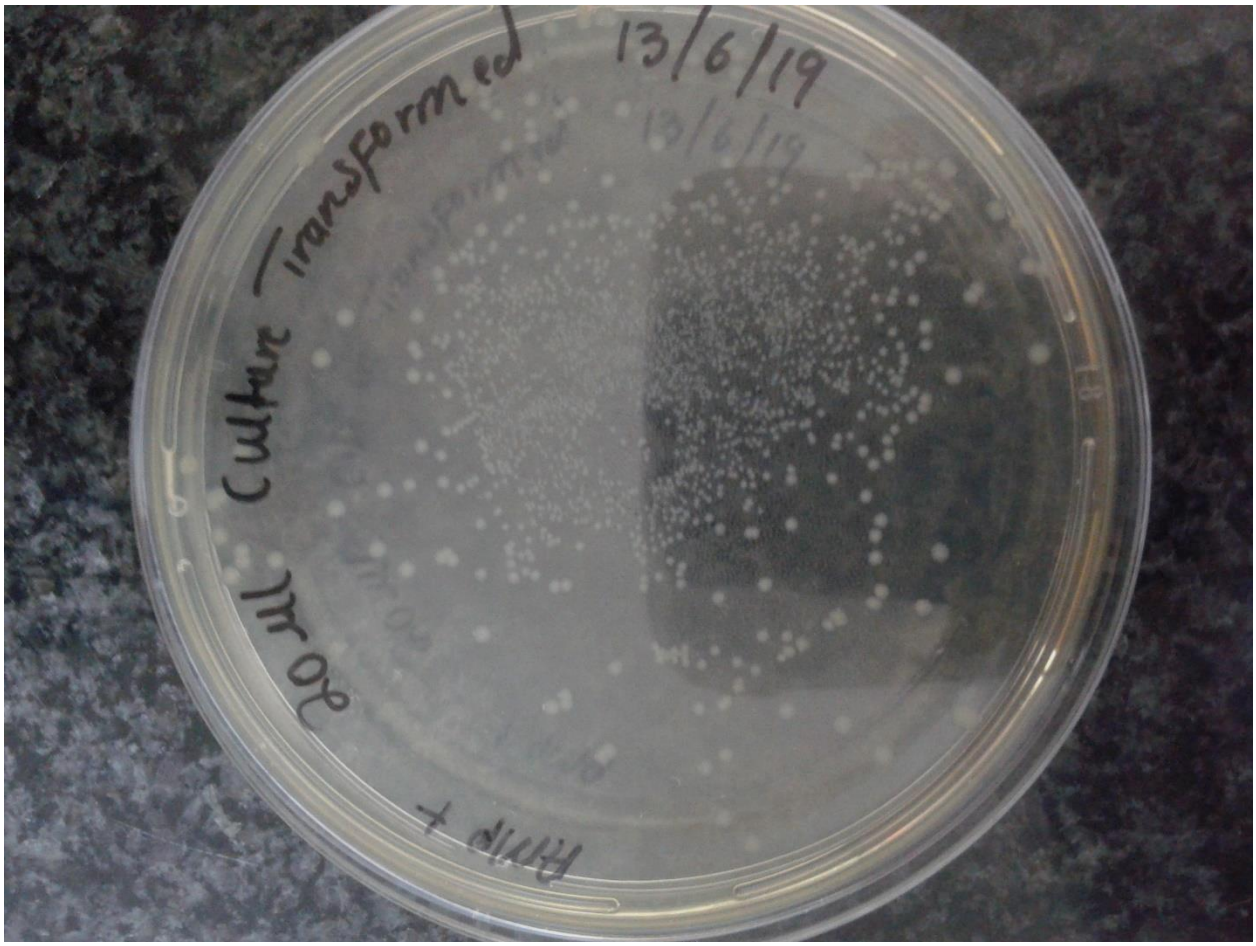

Figure 3.5.3: Ampicillin (100 µg/ml) LB agar plates with T7 express *lysY/I*^{q} Competent *E. coli* – pTXB1 clone colonies, grown overnight at 37°C.

T7 express *lysY/I*^{q} Competent *E. coli* cells were able to grow on L.B agar plate containing (100 µg/ml), that is because pTXB1-Tn5 plasmid have a resistance gene of ampicillin. Ampicillin inhibited the growth of T7 express *lysY/I*^{q} Competent *E. coli* cells that were not successfully transformed by pTXB1-Tn5 plasmid, since they do not have a resistance gene against ampicillin.

3.5.3 Colony PCR

pTXB1-Tn5 Plasmid DNA was isolated from transformed T7 express *lysY/I*^{q} Competent *E. coli* cells and Tn5 partial gene was successfully amplified using one round polymerase chain reaction (PCR). Figure 3.5.4 shows amplified partial Tn5 gene on 1% agarose gel.

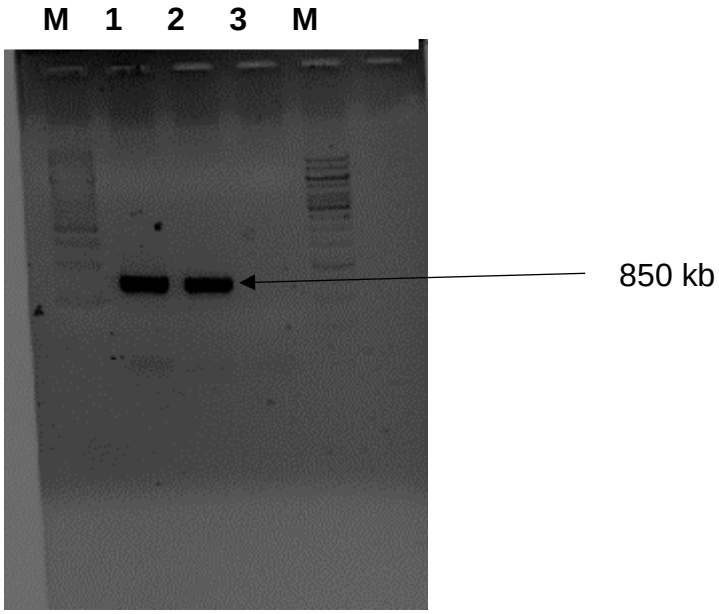

Figure 3.5.4: 10% gel electrophoresis, one round PCR product of the partial Tn5 gene. M is the molecular ladder called Quick Load 1 kb ladder, 2 and 3 is the amplified partial Tn5 gene, and 4 is negative control. Partial Tn5 gene of 850 bp was successfully amplified.

3.5.4 Expression and Purification of Tn5

Expression of Tn5 on T7 express *lysY/I*^{q} Competent *E. coli* was induced by 0.25 mM IPTG. T7 express *lysY/I*^{q} Competent *E. coli* cell harboring the clone pTXB1 and contained all necessary transcriptional factors produced several proteins including Tn5. Figure 3.5 shows expression profile of T7 express *lysY/I*^{q} Competent *E. coli* cells.

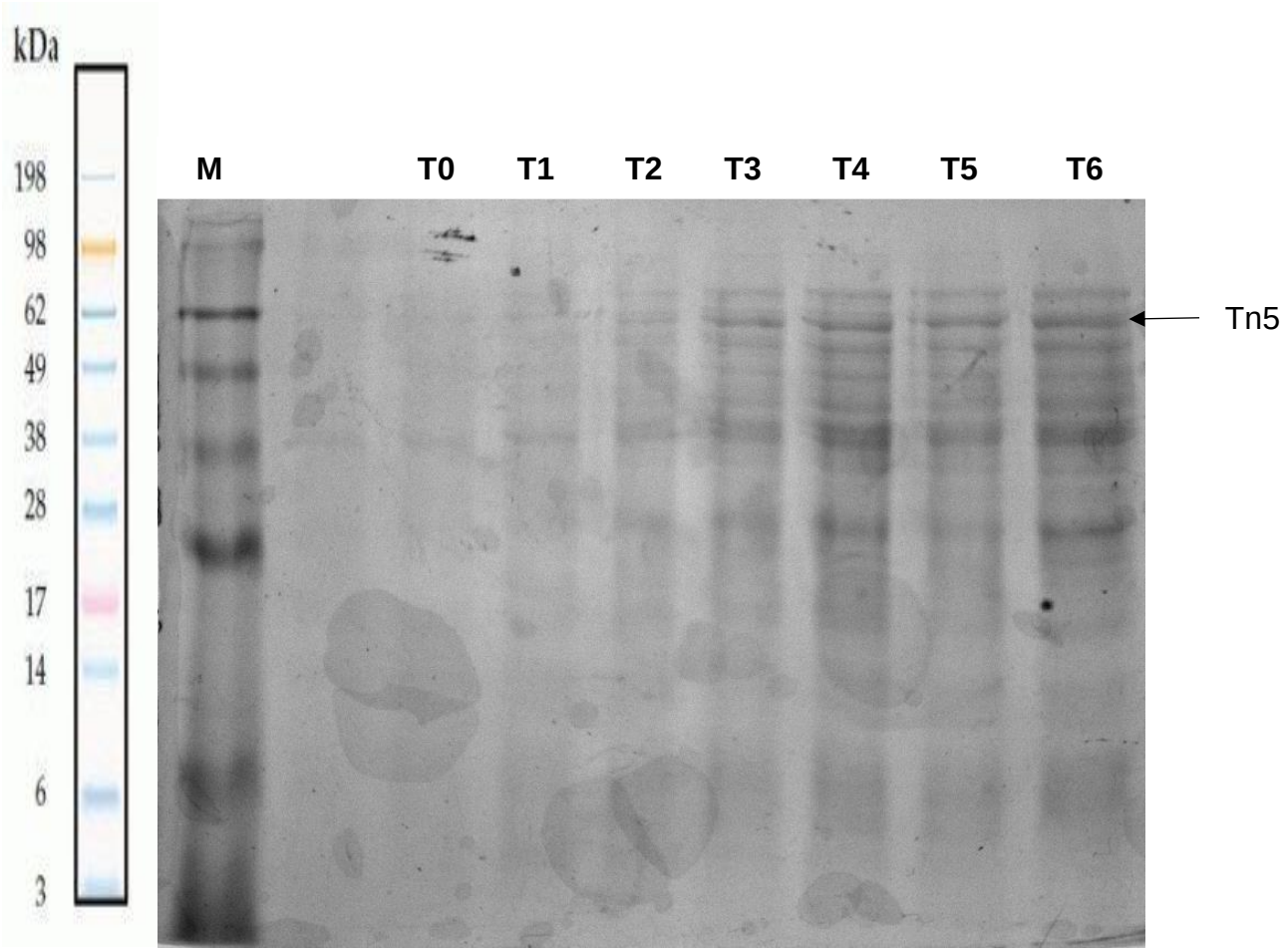

Figure 3.5.5: 10% SDS-PAGE gel showing the expression of Tn5 in T7 express *lysY/I*^{q} Competent *E. coli* at time interval 0-6 hours. Lane M indicates the SeeBlue Plus2 Prestained Standard protein marker protein marker. Lane T0-T6 represent protein expression time intervals. The band representing the expressed Tn5 is seen at ≈53 Kda and it is indicated by a red arrow.

After confirming expression, T7 express *lysY/I*^{q} Competent *E. coli* cells were lysed, and Tn5 enzyme was purified using the chitin column system. It was observed that low concentration of Tn5 was eluted from chitin column and high concentration of Tn5 was still bound to the chitin beads.

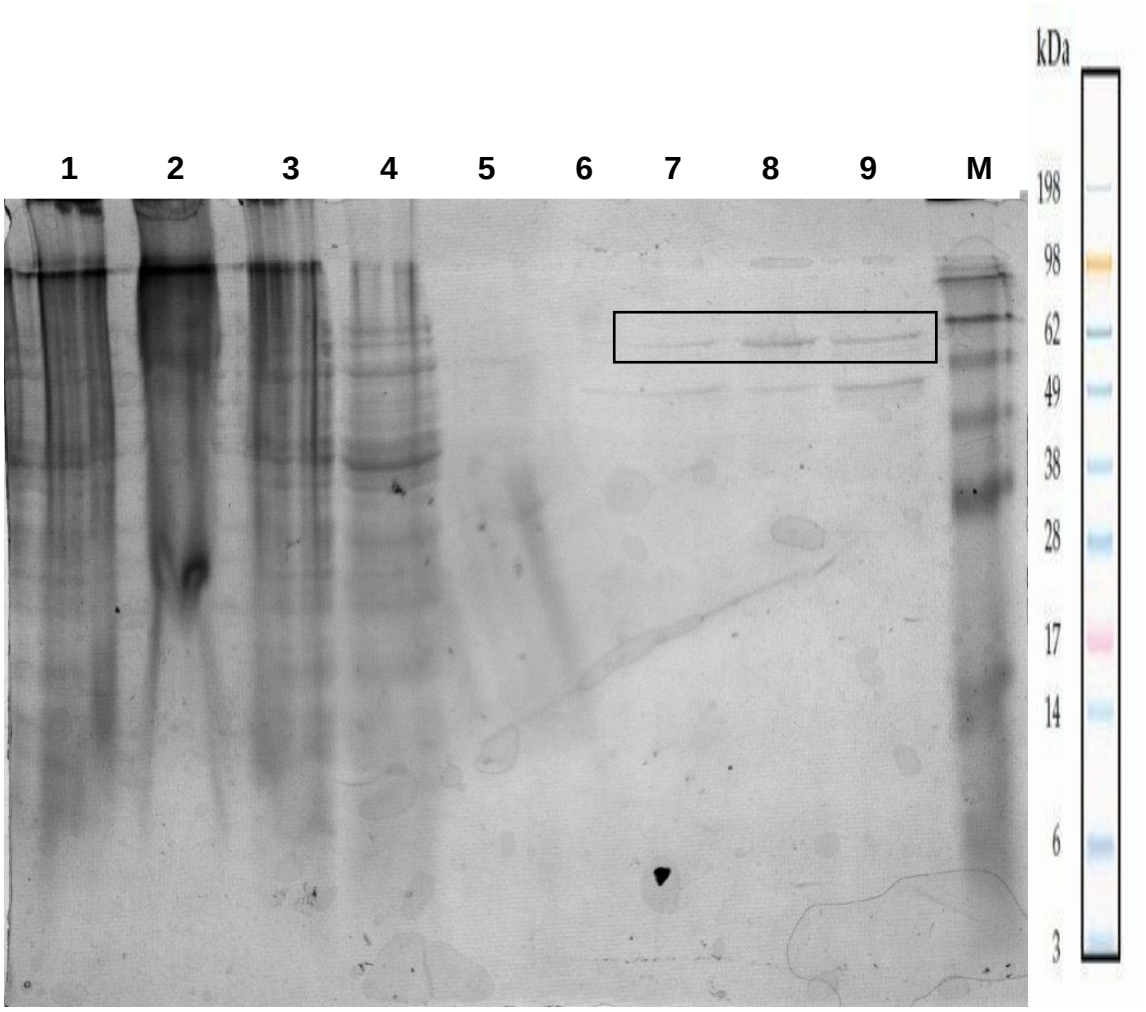

Figure 3.5.6: 10% SDS-PAGE run of Tn5 purification. Supernatant after precipitation (1), pellet after precipitation (2), and flow through (3), protein wash (4), Lane 5 have no sample loaded, DTT eluted fraction (6), protein elution (7), chitin resin (8), concentrated protein (9), M indicating the SeeBlue Plus2 Prestained Standard protein marker. Low concentration of Tn5 was eluted from chitin column and high concentration of Tn5 was still bound to the chitin beads.

Although Tn5 transposase was successfully expressed and purified, low concentration of Tn5 was obtained from purification. An unknown band of protein was observed from purified Tn5 which indicate that the Tn5 was not pure. A previous study indicated that bacterial transposon Tn5 encodes two proteins, the transposase and a related protein, the transposition inhibitor, whose relative abundance determines, in part, the frequency of Tn5 transposition (Reznikoff W.S., 1993). It was also noted that high concentration of Tn5 was still bound to the chitin beads. In a study done by Hennig et al., (2018), it was also difficult for them to purify Tn5 from chitin column and very low yields of the enzyme were obtained.

3.5.5 Bradford Assay

Bradford assay was done for eluted Tn5 fractions to determine which fractions have high concentration of Tn5. Blue colour on the tube indicates the presence of Tn5 within the fraction. The stronger the blue color the higher the concentration of Tn5 on the fraction. Purified Tn5 fractions with high Tn5 concentrations were pulled together and concentrated using MacroSep Advance Centrifugal Device 10K MWCO concentrated Tn5 was quantified using Qubit 3.0 Fluorometer and the concentration was 2.60 mg/mL.

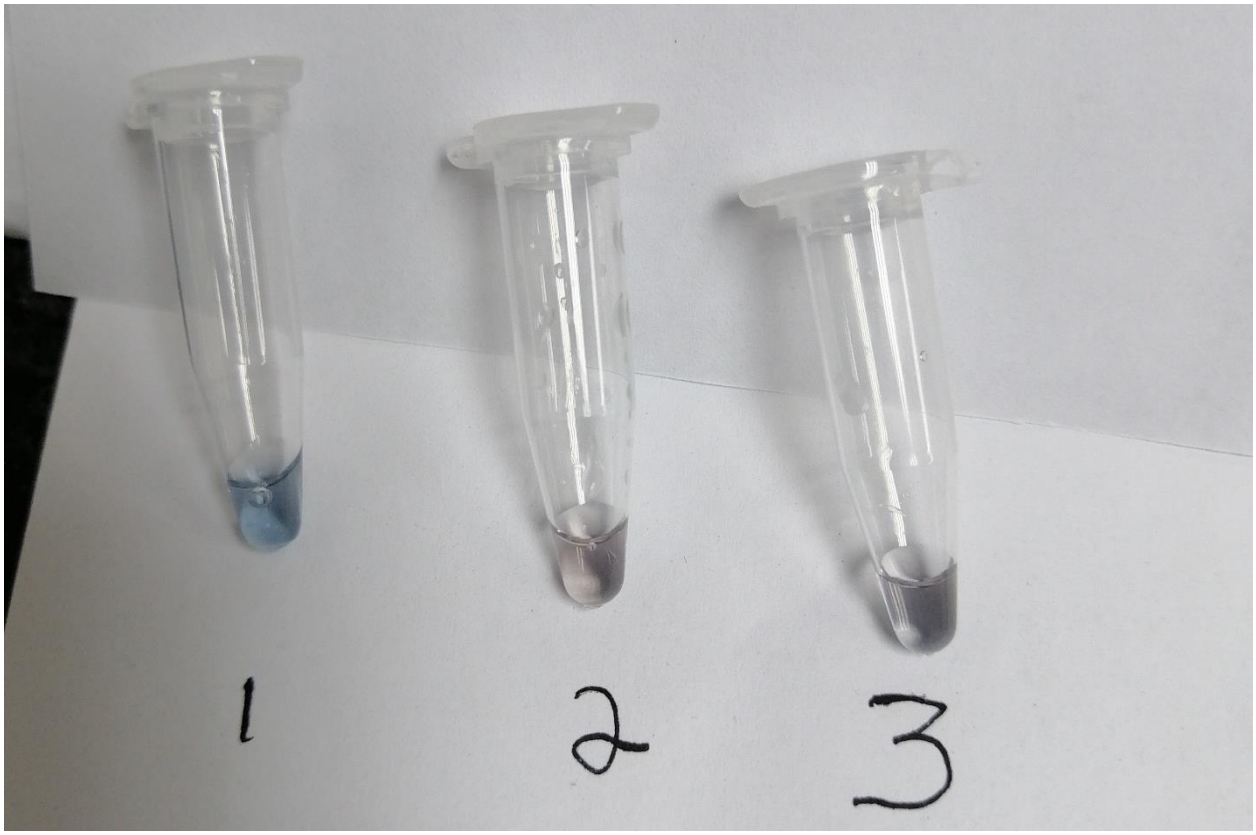

Figure 3.5.7: Representative results of protein concentration testing using Bradford assay. The blue colour indicates the presence of protein and the stronger the blue color indicates high protein concentration from the elution. Tube 1 have high concentration of protein.

3.5.6 Tn5 Activity Assays

Transposase 5 assembly with pre-annealed MEDS oligonucleotides was done prior to fragmentation and tagmentation step. Firstly, activity of the expressed Tn5 was tested by performing fragmentation reactions of different concentrations of Tn5 (0.50, 0.37, 0.26, 0.13, 0.076, 0.026 mg/mL) to determine Tn5 concentration that will yield better fragmentation while keeping DNA concentration constant at 10 ng/µL. Different DNA concentration were fragmented while keeping Tn5 constant (0.26 mg/mL) to determine DNA concentration that will result in better fragmentation. Unassembled Tn5 stock fragmented input DNA completely, on the other hand, assembled Tn5 did not fragment the DNA completely as the concentration of Tn5 decreases, see Figure 3.5.8.

M 1 2 3 4 5 6 7 8 9 10

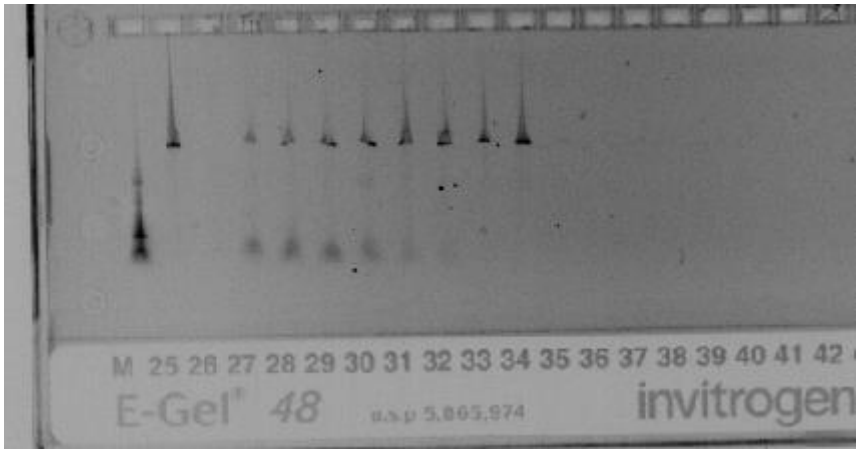

Figure 3.5.8: 10% E-Gel showing fragmentation of 1.6 Kb HIV-1 DNA using In-house Tn5 enzyme. M indicating Quick-Load LMW ladder, Unfragmented DNA (1), unassembled stock Tn5 2.60 mg/ml (2), Tn5 0.50 mg/ml (3), Tn5 0.37 mg/ml (4), Tn5 0.26 mg/ml (5), Tn5 0.13 mg/ml (6), Tn5 0.076 mg/ml (7), Tn5 0.026 mg/ml (8), control with no Tn5 input (9), Unfragmented DNA (10). As the concentration of Tn5 decreases from lane 3 to lane 8, the concentration of unfragmented DNA increases.

Tagmentation results of the fragmented DNA are shown in Figure 3.5.9. It was observed that tagmentation reaction of Tn5 (0.26 mg/ml) had a strong smear compared to other Tn5 concentrations. Libraries obtained had fragmentation distribution size of 100 to 250 bp.

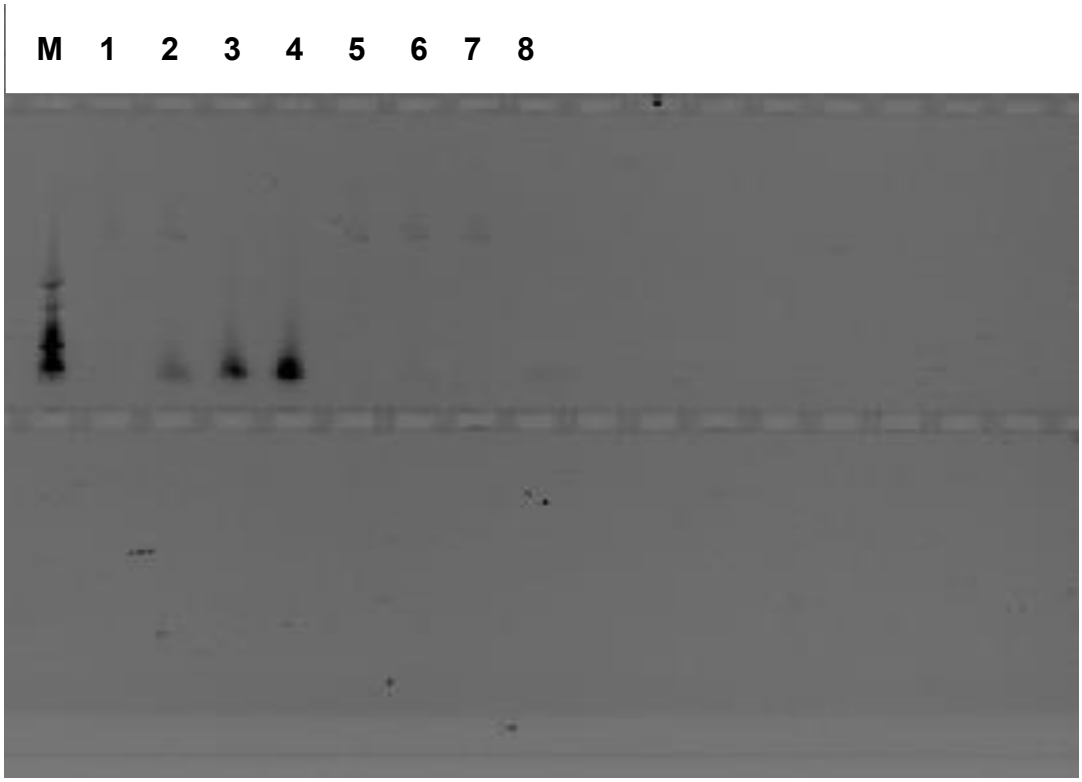

FIGURE 3.5.9: 10% E-Gel showing the tagmentation of HIV-1 DNA using In-house Tn5 enzyme. M indicating Quick Load LMW ladder, Unassembled stock Tn5 2.60 mg/ml (1), Tn5 0.50 mg/ml (2), Tn5 0.37 mg/ml (3), Tn5 0.26 mg/ml (4), Tn5 0.13 mg/ml (5), Tn5 0.076 mg/ml (6), Tn5 0.026 mg/ml (7), control with no Tn5 input (8). Lane 4 with 0.26 mg/ml concentration of Tn5 had strong smear band compared to other libraries prepared with different Tn5 concentration.

Purified libraries were quantified using Qubit 3.0 Fluorometer. It was observed that purified libraries had low concentrations. Although different concentration of Tn5 was used while keeping DNA concentration constant (Table 3.2), and different concentration of DNA used while keeping Tn5 concentration constant (Table 3.3), DNA libraries concentrations were not increasing. Different ratio of AMPure XP beads for DNA libraries cleanup was used to rule out the theory that libraries are being lost during cleanup stage (Table 3.4). This step was done to identify desired concentration of both DNA and Tn5, only HIV DNA was used at this stage.

Table 3.2: Concentration of libraries prepared using different Tn5 concentration.

Samples	Tn5 (mg/mL)	Purified Libraries (ng/µL)
HIV A3 (10 ng/ µL)	0.26 mg/mL	0.091
HIV A4 (10 ng/ µL)	0.20 mg/mL	0.052
HIV A5 (10 ng/ µL)	0.15 mg/mL	0.021

Table 3.3: Concentration of libraries prepared using different DNA concentrations input.

Samples	Tn5 (mg/mL)	Libraries (ng/µL)	Purified Libraries (ng/µL)
HIV A3 (1 ng/µL)	0.26	3.91	0.065
HIV A4 (5 ng/µL)	0.26	4.04	0.110
HIV A5 (10 ng/µL)	0.26	3.41	0.078

Table 3.4: Concentration of libraries purified using different ratio of AMPure XP Beads.

Samples (10 ng/µL)	Tn5 (mg/mL)	AMPure XP beads	Libraries (ng/µL)	Purified Libraries (ng/µL)
HIV A3	0.26	1:0.05	4.84	0.051
HIVA4	0.26	1:1	5.76	0.086
HIV A5	0.26	1:2	4.31	0.194

3.5.7 Fragmentation and Tagmentation using In-house Tn5 and Nextera DNA Library Preparation Kit

After Tn5 activity assay, 0.26 mg/ml of expressed Tn5 was identified as desired concentration for fragmentation reactions. HIV, HPV and KSHV DNA (10 ng/µL) were fragmented and tagmented using both in-house Tn5 protocol and Nextera XT DNA library preparation kit to compare fragmentation profiles. It was observed on 1% E-gel electrophoresis that in-house Tn5 had short smears that were between 50 to 250 bp where as Nextera XT DNA library preparation kit had long smears that were distributed between 200 to 700 bp (Figure 3.5.10).

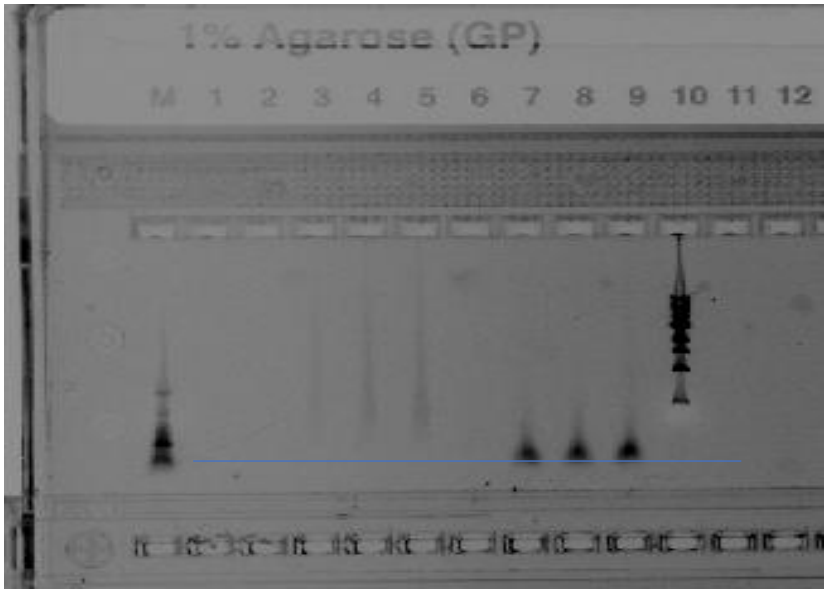

Figure 3.5.10: A representative of libraries prepared using in-house Tn5 protocol and Illumina Nextera DNA library preparation kit. M indicating Quick-Load LMW ladder, libraries prepared by Nextera Kit (3 – 5), libraries prepared by in-house Tn5 protocol (7 – 9). Long and well distributed smears were observed on libraries prepared using Nextera kit while short smear of distribution size between 50 – 250 bp was observed on libraries prepared using in-house Tn5 protocol.

Table 3.5: Concentrations of libraries prepared using in-house Tn5 protocol and Nextera XT DNA library preparation kit.

Samples	**In-house Tn5 (ng/µL)**	**Nextera XT DNA library prep kit (ng/µL)**
HIV D07 (10 ng/µL)	0.114	1.99
HIV E05 (10 ng/µL)	0.092	3.61
HIV F02 (10 ng/µL)	0.119	3.06
HPV D1 (10 ng/µL)	7.62	4.21
HPV D3 (10 ng/µL)	8.65	3.15
HPV D5 (10 ng/µL)	3.65	5.30
KSHV BM-115 (10 ng/µL)	0.455	4.76
KSHV BM-189 (10 ng/µL)	0.213	2.68
KSHV BM-408 (10 ng/µL)	0.090	4.52

Concentrations of amplified and purified libraries are shown in Table 3.5. Libraries prepared using in-house Tn5 protocol had lower concentration compared to libraries prepared using Nextera XT DNA library preparation kit. Low concentration of libraries prepared using in-house Tn5 protocol indicate that there is low number of fragments which were successfully amplified during tagmentation step.

In-house expressed Tn5 have shown potential of fragmenting viral DNA during library preparations as compared to Nextera XT DNA library prep kit. It was observed that libraries prepared using In-house Tn5 had fragment size that was not well distributed compared to libraries obtained using Nextera XT DNA library prep kit. Quantification of libraries indicated that libraries prepared by In-house Tn5 had low concentration compared to the required concentration before pulling the libraries for sequencing. A similar study was done in which high quality sequencing libraries were obtained; PEG was used as crowding agent reagent (Picelli *et al*., 2014). In the presence of a crowding agent, when high amounts of macromolecules such as proteins and polymer are added to a solution, they sequester water, lowering molecular diffusion rates and ultimately increasing the efficiency of biological reactions (Zimmerman and Pheiffer 1983)

3.5.8 Sequencing Analysis

An Illumina MiniSeq run for 2X150 bp paired end reads using V3 reagent kit is expected to generate about 6.6 to 7.5 Gb data, with 14 – 16 million paired end reads passing filter, 170 – 220 kmm^2 cluster passing filter giving a QC 30 of >70%. The run was successful and about 4.46 Gb of data was obtained with the error rate of 0.72%, and QC30 of 90.90%

Libraries prepared using In-house Tn5 protocol and Nextera XT DNA library preparation kit were sequenced using MiniSeq instrument. Quality check using FastQC software v0.11.8 showed that generated sequences using In-house Tn5 protocols had lower number of reads compared to sequences generated using Nextera XT DNA library preparation kit. But no sequences were flagged for poor quality for both in-house Tn5 and Nextera XT.

Table 3.6: Characteristics of sequences generated using the in-house Tn5 and Nextera XT protocols.

Sequence name	Total sequence	Sequence length	Sequence flagged for poor quality	%GC
HIV D07 (In-house Tn5)	40	43 – 151	0	40
HIV E05 (In-house Tn5)	74	39 – 151	0	40
HIV F02 (In-house Tn5)	54	41 – 151	0	37
HIV D07 (Nextera XT)	138 321	35-151	0	38
HIV E05 (Nextera XT)	120 998	35-150	0	40
HIV F02 (Nextera XT)	120 288	35-151	0	40
KSHV-BM 115 (In-house Tn5)	26	45 – 151	0	35
KSHV-BM 189 (In-house Tn5)	4	151	0	36
KSHV-BM 408 (In-house Tn5)	4	39 – 151	0	44
KSHV-BM 115 (Nextera XT)	123 458	35-151	0	38
KSHV-BM 189 (Nextera XT)	133 201	35-151	0	38
KSHV-BM 408 (Nextera XT)	119 789	35-151	0	40
HPV S93 (In-house Tn5)	2	150	0	45
HPV S94 (In-house Tn5)	2	109	0	33
HPV S95 (In-house Tn5)	0	0	0	0
HPV S93 (Nextera XT)	141 256	35-151	0	40
HPV S94 (Nextera XT)	117 632	35-151	0	40
HPV S95 (Nextera XT)	127 852	35-151	0	39

The y-axis shows the quality scores of the sequences on the graphs, such as in figure 3.5.11. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). Generate sequences using both In-house Tn5 protocol and Nextera XT DNA library preparation kit have higher score in base calling and the graph is within green regions which indicates good quality calls. The quality score and coverage figures of generated sequences are shown below:

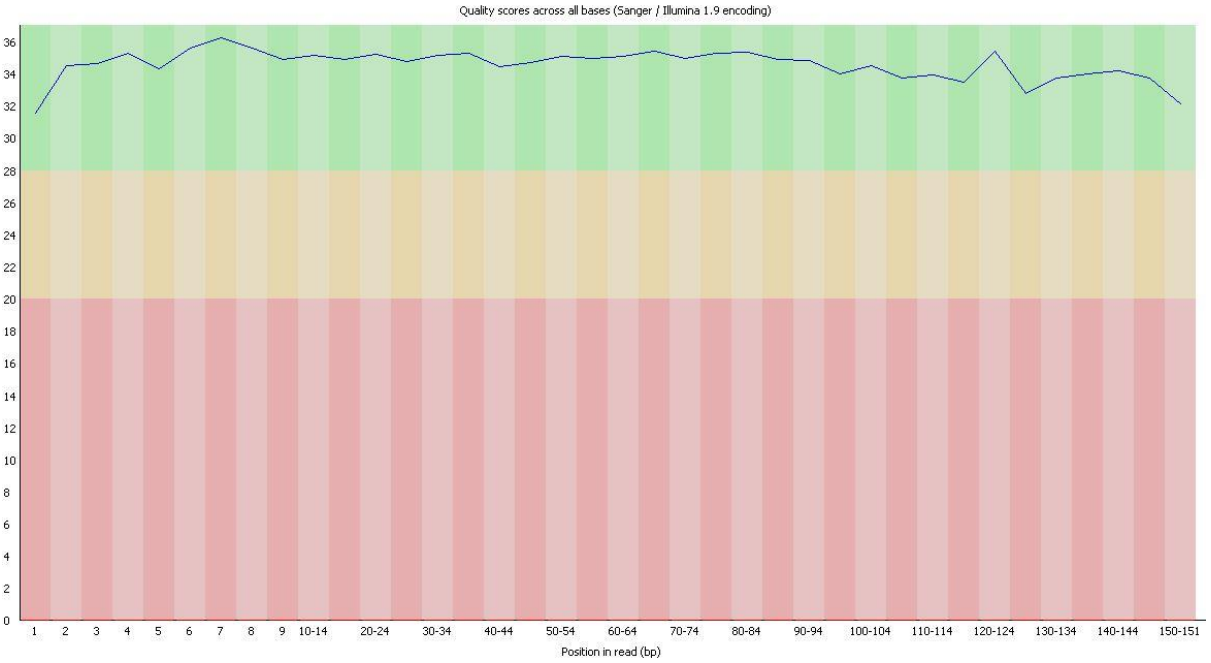

Figure 3.5.11: Showing quality scores from FastQC across all bases at each position for sequences sequenced with in-house Tn5 protocol, sample HIV E05. All bases are of good quality.

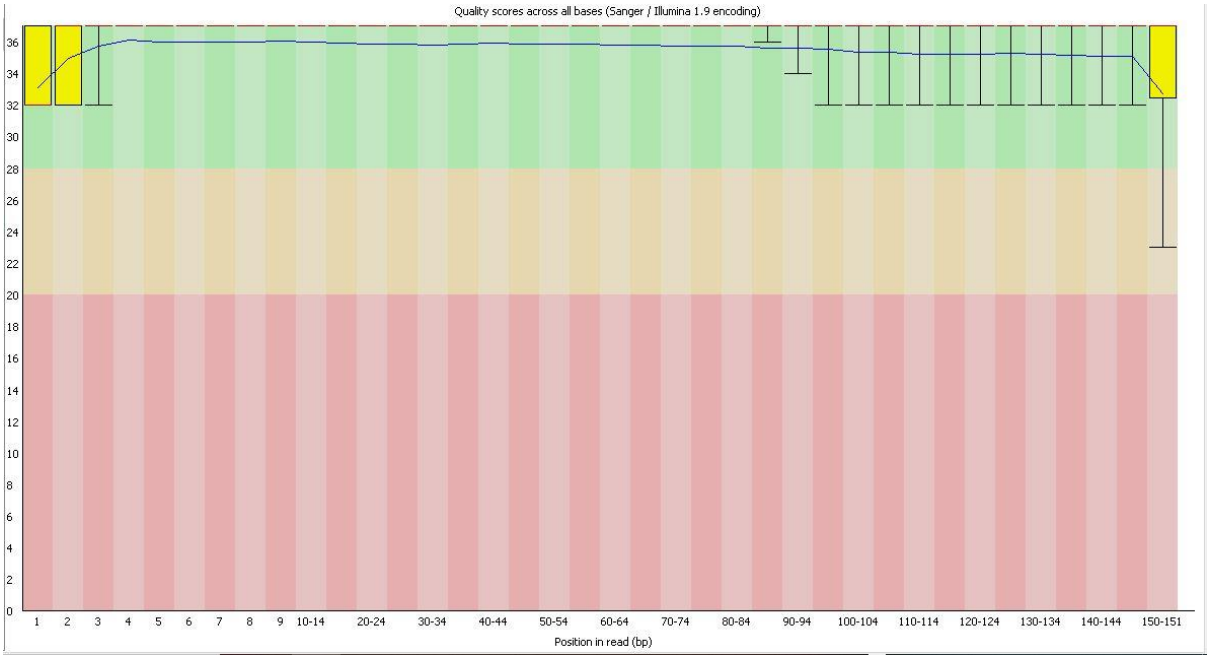

Figure 3.5.12: Showing quality scores from FastQC across all bases at each position for sequences sequenced using Nextera XT DNA library preparation kit, sample HIV E05. All bases are of good quality except the last two bases.

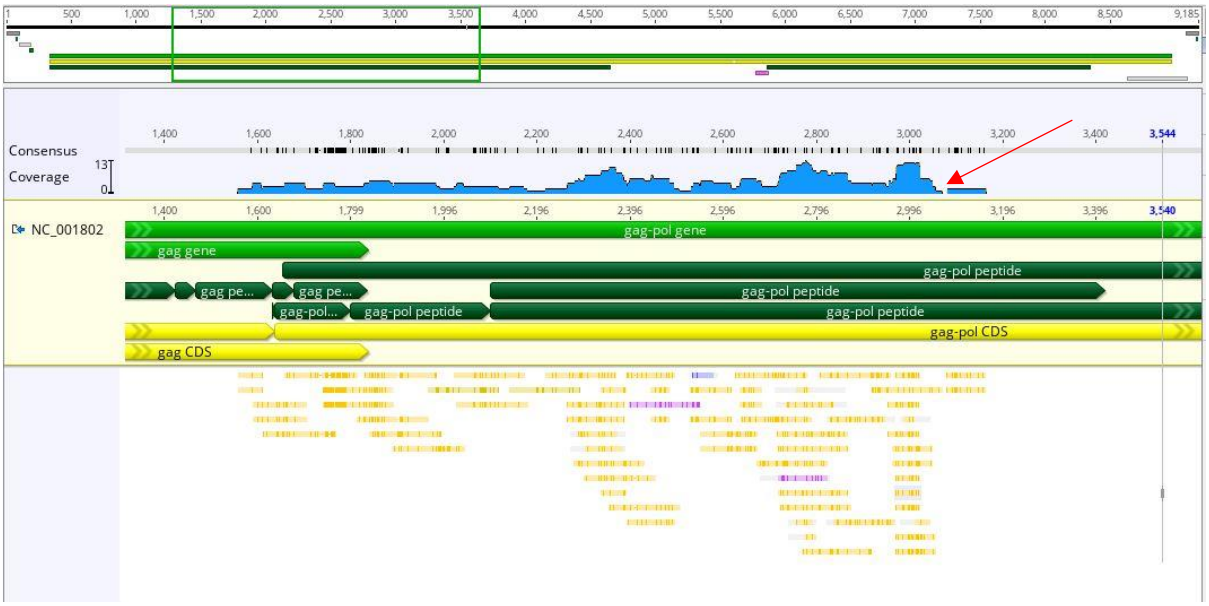

Figure 3.5.13: Coverage of generated sequences of HIV-1 gag/pol region (≈1.6 Kb) using In-house Tn5 protocol in blue colour, sample HIV E05. There is a gap at the end of reads coverage indicated by red arrow.

Generated paired reads (in orange colour) were trimmed at 1% error rate and mapped to HIV-1 reference sequence. Paired reads mapped to gag/pol region, but a gap of 10 bases was observed at the end of reads coverage.

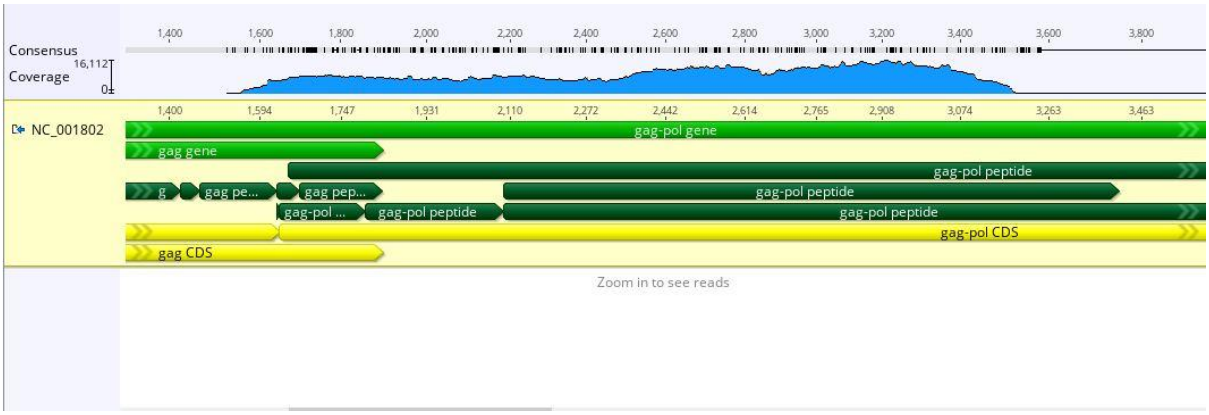

Figure 3.5.14: Coverage of generated sequences of HIV gag/pol region (≈1.6 Kb) using Nextera XT DNA library preparation kit in blue colour, sample HIV E05. Generated sequences had a full coverage for the entire gag/pol region which was sequenced.

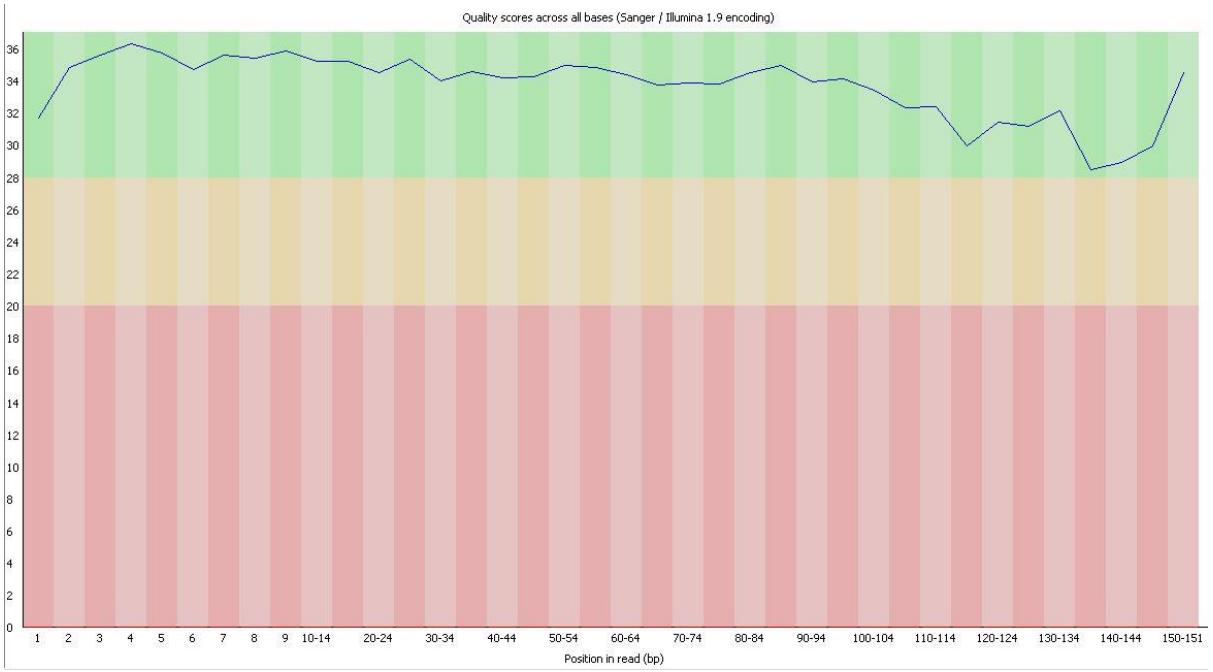

Figure 3.5.15: Showing quality scores from FastQC across all bases at each position for sequences sequenced using In-house Tn5 protocol, sample HIV D07. All bases are of good quality.

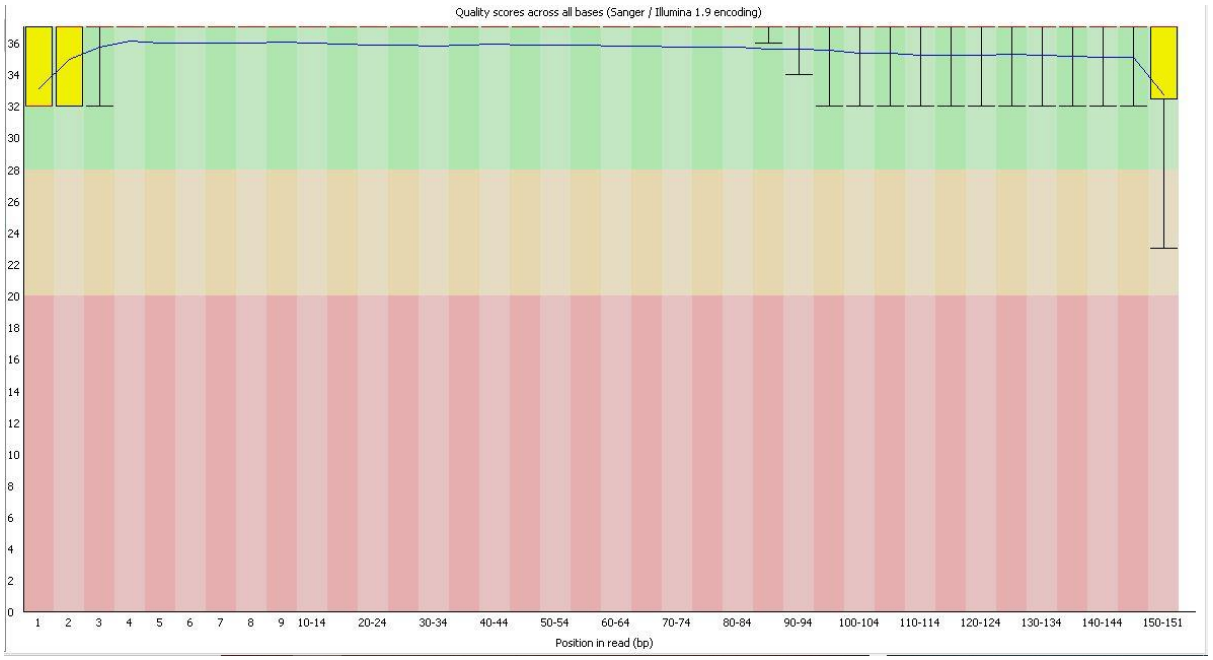

Figure 3.5.16: Showing quality scores from FastQC across all bases at each position for sequences sequenced using Nextera XT DNA library preparation kit, sample HIV D07. All bases are of good quality.

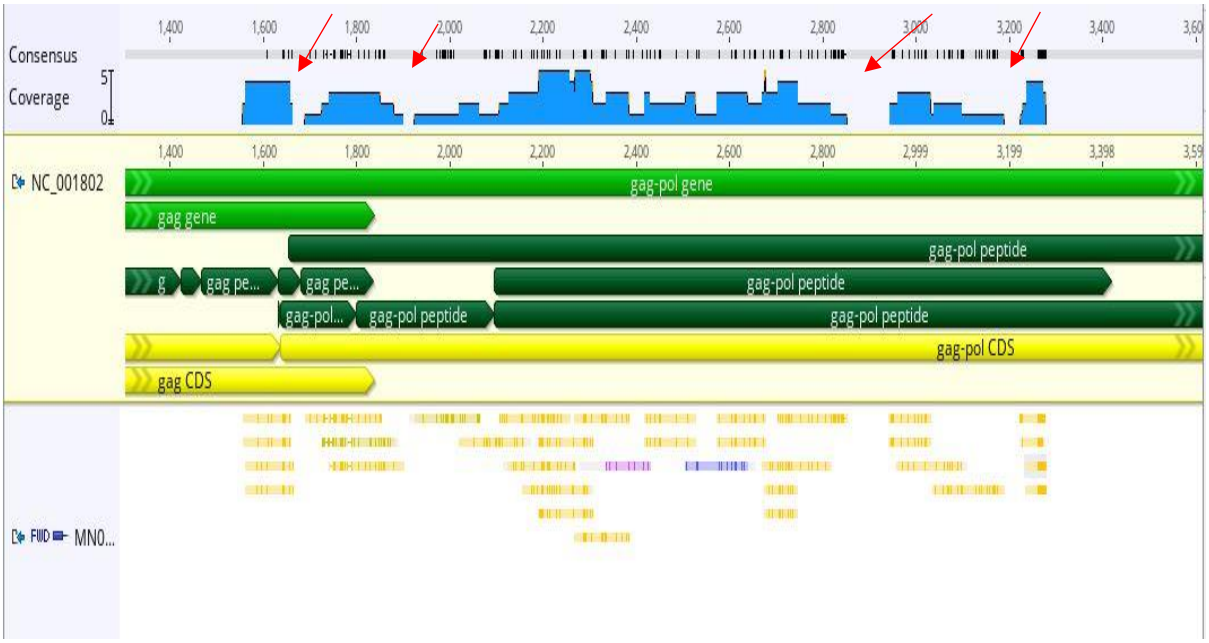

Figure 3.5.17: Coverage of generated sequences of HIV gag/pol region (≈1.6 Kb) using in-house Tn5 protocol in blue colour, sample HIV D07. There are gaps within the sequenced region indicated by red arrows.

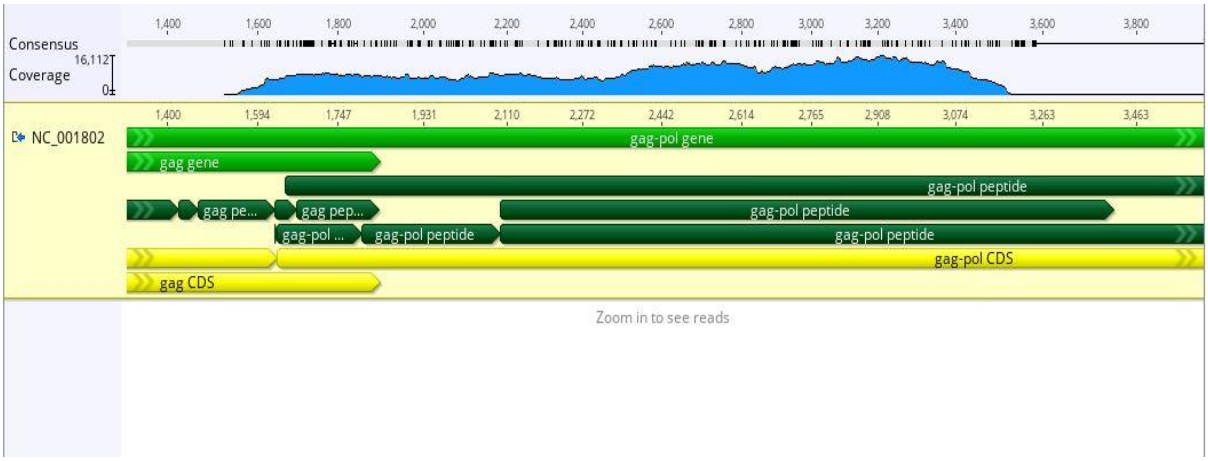

Figure 3.5.18: Coverage of generated sequences of HIV gag/pol region (≈1.6 Kb) using Nextera XT DNA library preparation kit in blue colour, sample HIV D07. Generated sequences had a full coverage for the entire gag/pol region which was sequenced.

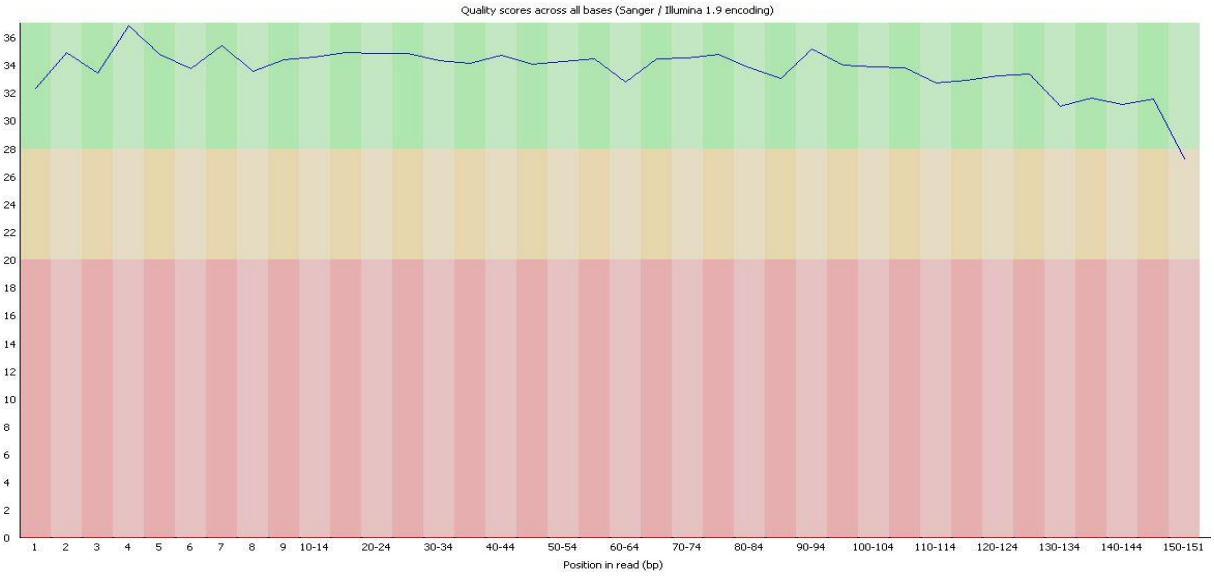

Figure 3.5.19: Showing quality scores from FastQC across all bases at each position for sequences sequenced using in-house Tn5, sample HIV F02. All bases are of good quality except bases at position 150 and 151.

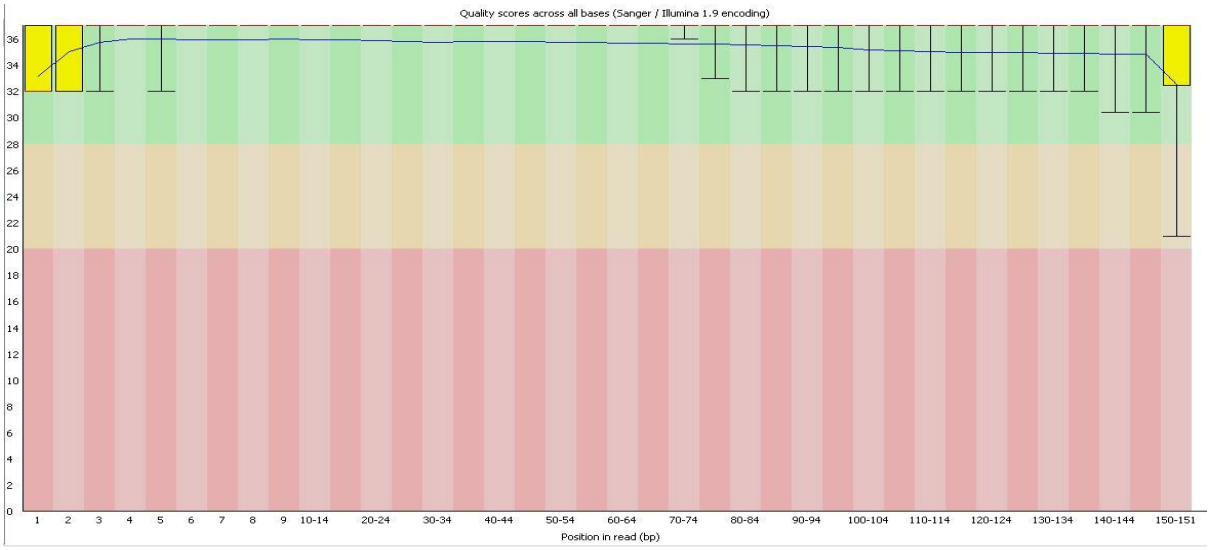

Figure 3.5.20: Showing quality scores across all bases at each position for sequences sequenced using Nextera XT DNA library preparation kit, sample HIV F02. All bases are of good quality.

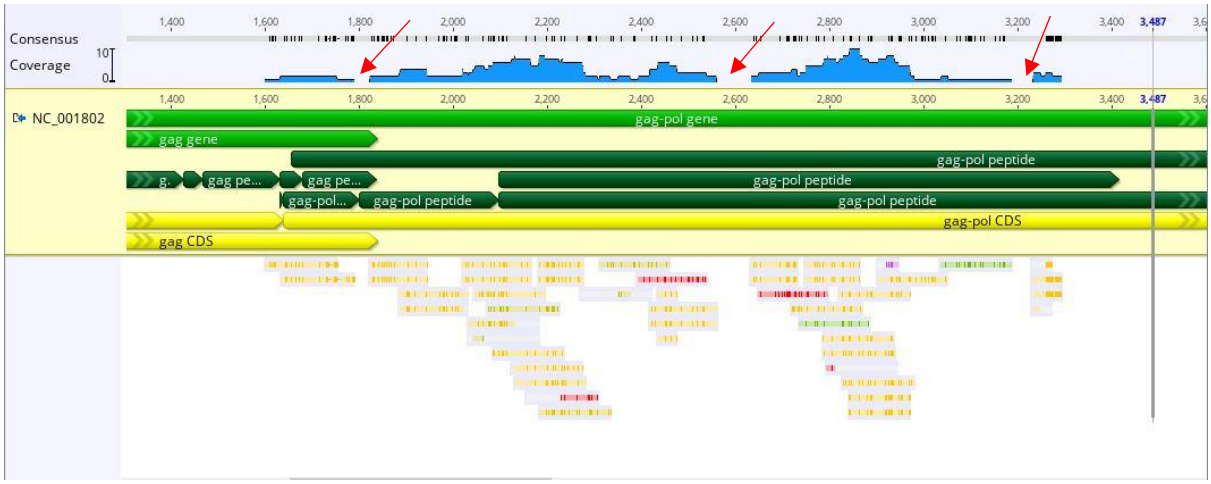

Figure 3.5.21: Coverage of generated sequences of HIV gag/pol region (≈1.6 Kb) using In-house Tn5 protocol in blue colour, sample HIV F02. There are gaps within the sequenced region indicated by red arrows.

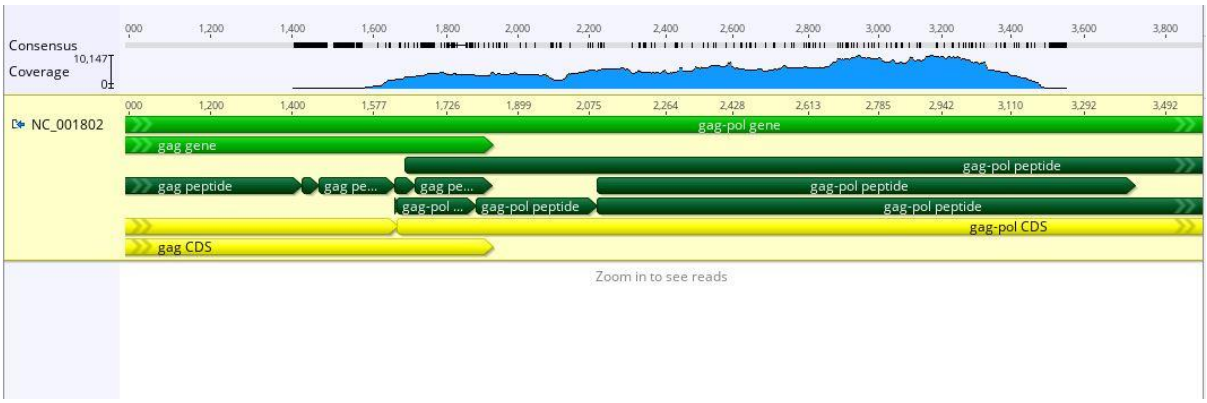

Figure 3.5.22: Coverage of generated sequences of HIV gag/pol region (≈1.6 Kb) using Nextera XT DNA library preparation kit in blue colour, sample HIV F02. Generated sequences had a full coverage for the entire gag/pol region which was sequenced.

Fewer sequenced reads were obtained on HPV and KSHV DNA libraries, and one HPV sample had no sequence reads. When mapped to their respective reference sequences, no reads were assembled to the reference sequences. One sample (KSHV-BM 115) that had 26 reads, 15 of 26 reads assembled to reference sequence. Most of the sequence reads that were obtained had poor quality on their bases as determined by FASTTQC software.

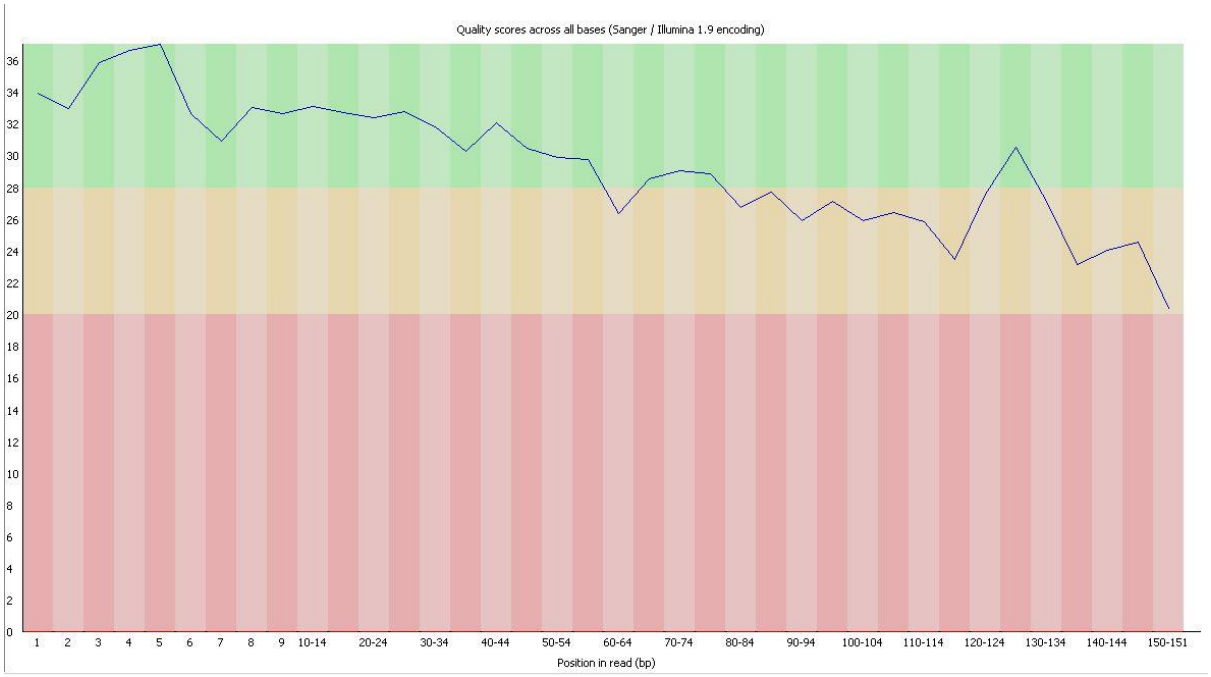

Figure 3.5.23: Showing quality scores from FastQC across all bases at each position for sequences sequenced using in-house Tn5, sample KSHV BM 115. All bases are of good quality.

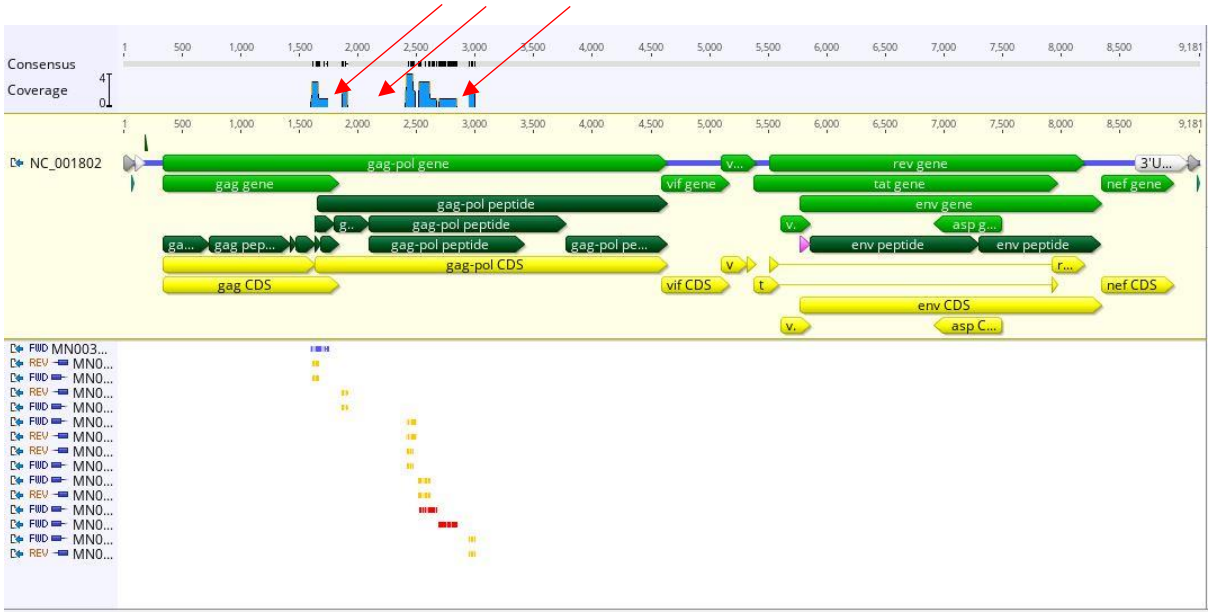

Figure 3.5.24: Coverage of generated sequences of KSHV region (≈320 bp) using In-house Tn5 protocol in blue colour, sample KSHV BM 115. There are gaps within the sequenced region indicated by red arrows.

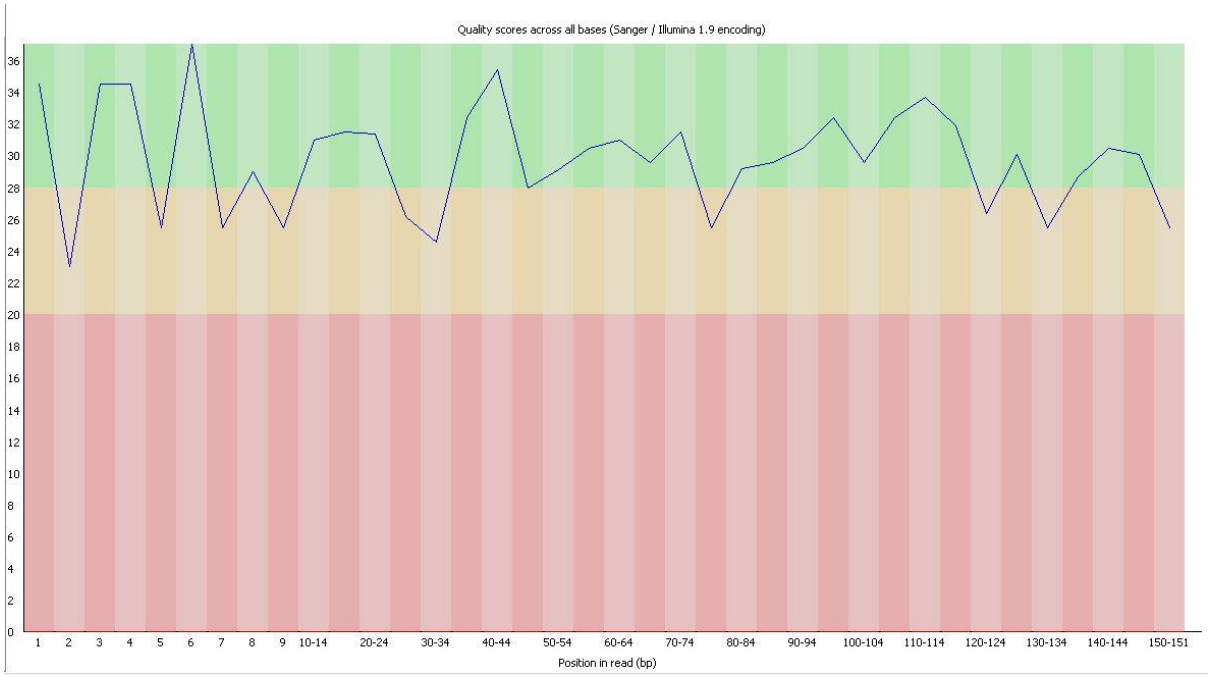

Figure 3.5.25: Showing quality scores from FastQC across all bases at each position for sequences sequenced using in-house Tn5, sample KSHV BM 189. All bases are of good quality.

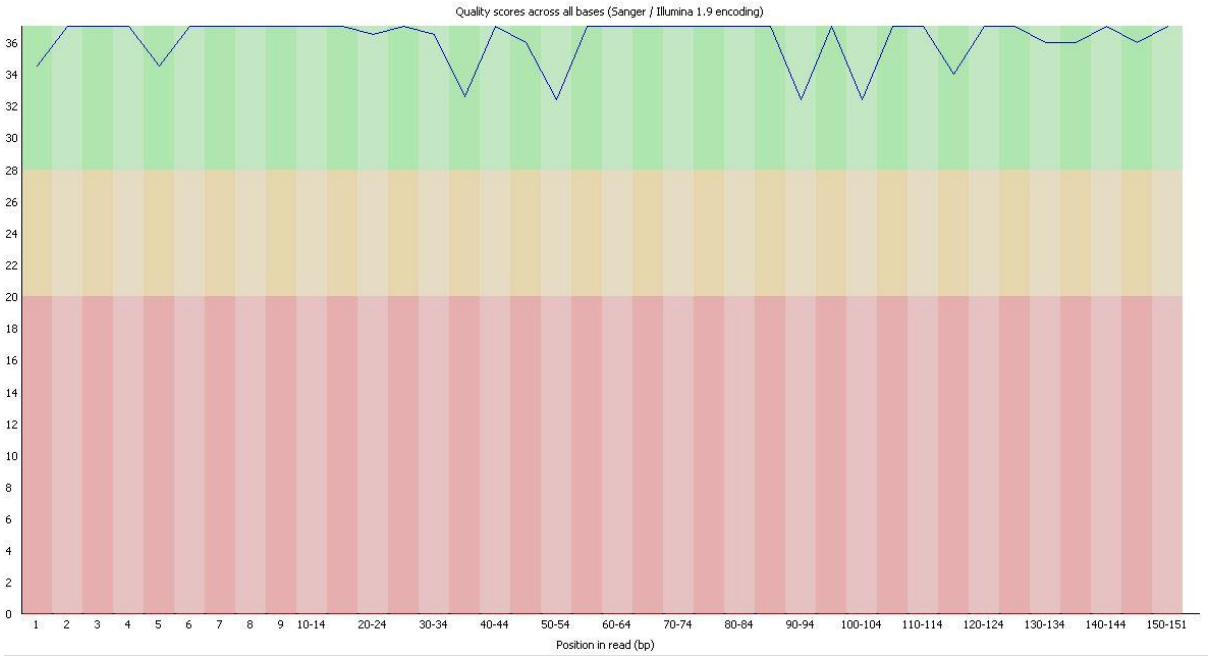

Figure 3.5.26: KSHV 408 Showing quality scores from FastQC across all bases at each position for sequences sequenced using in-house Tn5, sample KSHV BM 408. All bases are of good quality.

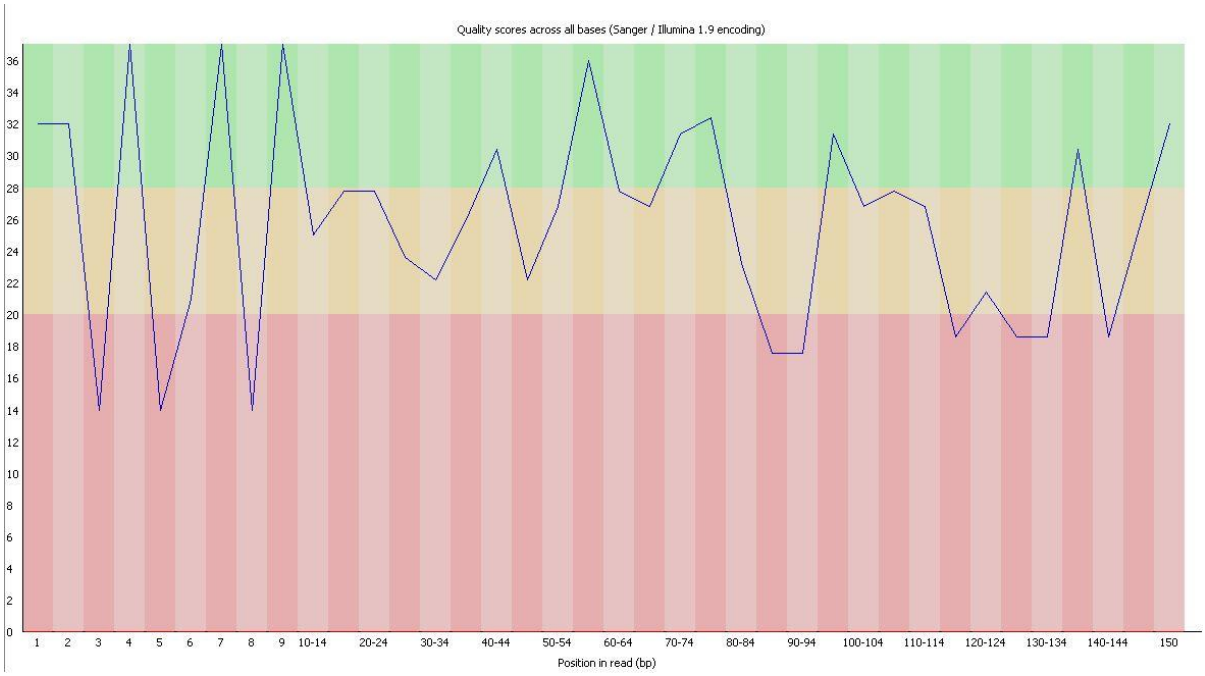

Figure 3.5.27: Showing quality scores from FastQC across all bases at each position for sequences sequenced using in-house Tn5, HPV S93. Few bases are of bad quality.

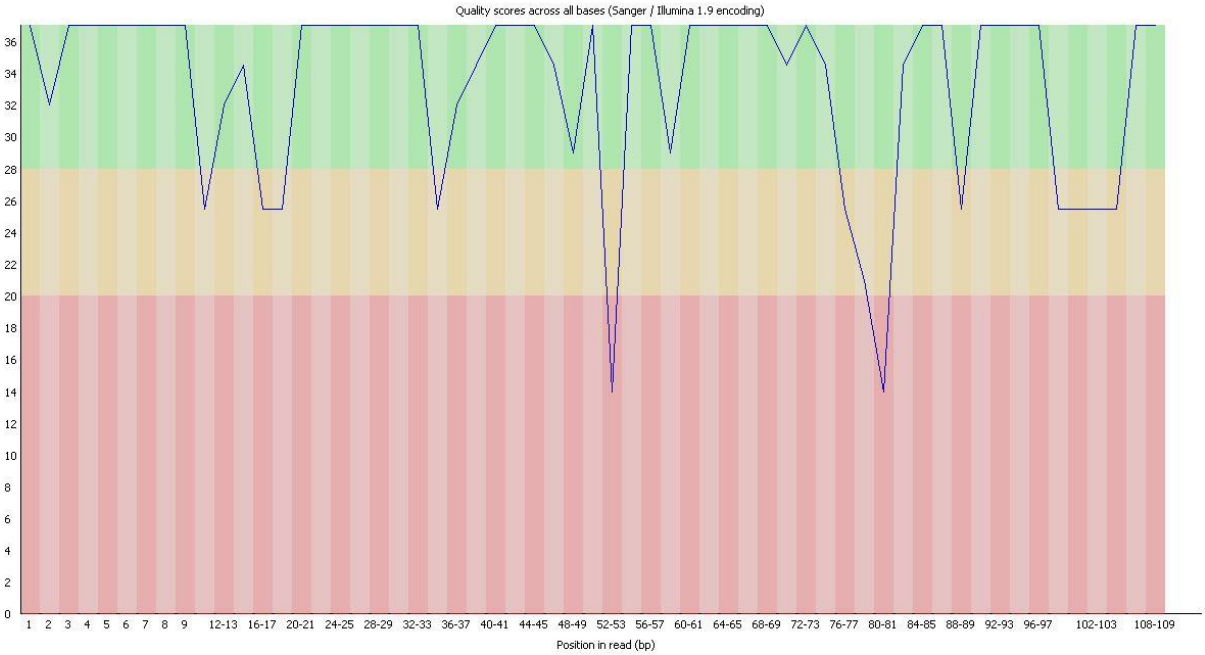

Figure 3.5.28: Showing quality scores from FastQC across all bases at each position for sequences sequenced using in-house Tn5, sample HPV S94. All bases are of good quality except two bases at position 52-53 and 80-81.

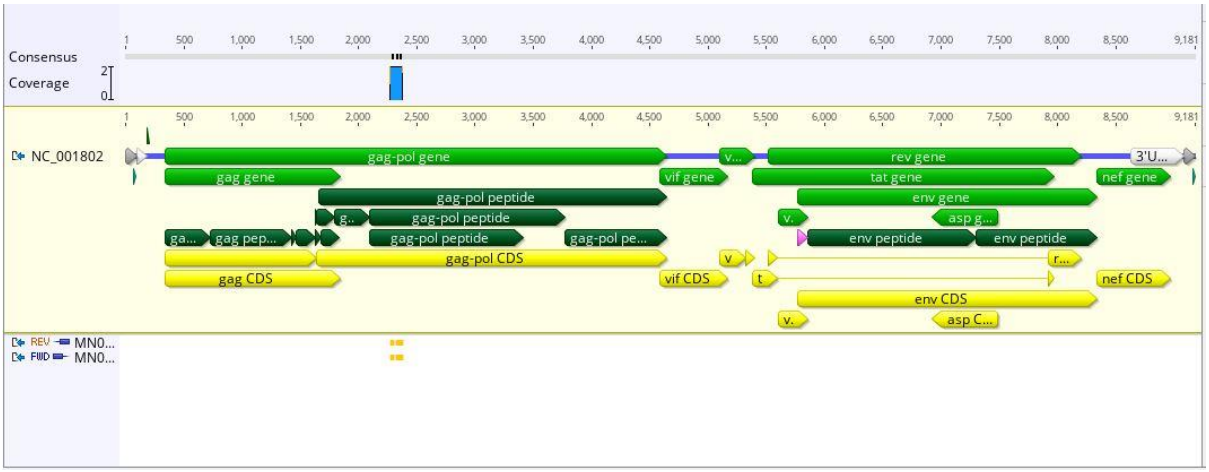

Figure 3.5.29: Coverage of generated sequences of HPV region (≈430 bp) using In-house Tn5 protocol in blue colour, sample HPV S94. Coverage of 151 bp was obtained because only two reads (forward and reverse) were mapped to the reference sequence.

Transposase 5 was successfully expressed from T7 express *lysY/I*^{q} competent E. *coli* cells and purified. Viral DNA libraries of HIV, HPV and KSHV that were prepared using In-house expressed Tn5 were fragmented, but this activity was not optimal to generate high number of reads. Overall, the sequencing run was successful and good data were generated with the QC30 of > 90.90%. Generated sequences prepared from both protocols had good quality score as determined by FastQC software, no sequence was flagged for poor quality. Low number of sequences reads resulted from poor fragmentation activity of in-house expressed Tn5 led to gaps when sequencing reads mapped to a reference sequence, thus because generated reads could not cover the whole targeted DNA region.

A limitation of this study was that a bioanalyzer, for quality control analysis of prepared libraries, would have been a better option, than E-gel, to visualize fragmentation, giving a better idea of the fragment sizes of libraries prepared using the in-house Tn5 before sequencing.

3.5.9 Future directions

Since low number of reads were obtained, protocol optimization is required to obtain high number of reads. After quantification of purified libraries, low concentrations were obtained, which indicates that few fragments might have been obtained from fragmentation. Tn5 activity might have also played a role in obtaining low fragments, because during Tn5 purification, two bands were observed, that is Tn5 and an unknow protein. An unknown protein might be affecting Tn5 activity. Therefore, Optimization of Tn5 purification will be required to rule out Tn5 activity in this challenge.

In conclusion, a procedure for over expression of Tn5 used for DNA library preparation for NGS has been demonstrated. In addition, the enzyme was shown to fragment viral DNA, although this activity was not seen to be optimal to generate a high number of NGS sequence reads. Further work will endeavor to enhance the production throughput of the produced Tn5 and optimized its fragmentation activity to yield better libraries for quality sequences in terms of the number of reads. The capability to produce Tn5 in-house will allow researchers in resource constrained laboratories to harness the full potential of NGS technology cheaper.

3.7 REFRENCE

Adey A, Morrison HG, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, *11*(12): R119.

Ahmadian A, Svahn H (2011). Massively Parallel. Sequencing Platforms using Lab on a Chip Technologies. *Lab Chip*, 11: 2653 − 2655.

Andersen DC, Krummen L (2002). Recombinant protein expression for therapeutic applications. *Current Opinion in Biotechnology*, *13*(2): 117-123.

Behjati S, Tarpey PS (2013). What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, *98*(6): 236-238.

Caruccio N, 2011. Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. In *High-Throughput Next Generation Sequencing*: 241-255.

Chen R, 2012. Bacterial expression systems for recombinant protein production: E. coli and beyond. *Biotechnology Advances*, *30*(5): 1102-1107.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, *56*(2): 61.

Hennig BP, Velten L, Racke I, Tu CS, Thoms M, Rybin V, Besir H, Remans K, Steinmetz LM (2018). Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3: Genes, Genomes, Genetics*, *8*(1): 79-89.

Hunt I, (2005). From gene to protein: a review of new and enabling technologies for multi-parallel protein expression. *Protein Expression and Purification*, *40*(1): 1-22.

Imataka H, Mikami S (2009). Advantages of human cell-derived cell-free protein synthesis systems. *Seikagaku* 81(4): 303–307.

Khan, K.H., 2013. Gene expression in mammalian cells and its applications. *Advanced Pharmaceutical Bulletin*, *3*(2), p.257.

Mikami S et al. (2006). An efficient mammalian cell-free translation system supplemented with translation factors. *Protein Expr Purif*, 46(2): 348–357.

Mikami S et al. (2008). A human cell-derived *in vitro* coupled transcription/translation system optimized for production of recombinant proteins. *Protein Expr Purif* 62(2):190–198.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*(4): 536-540.

Overton TW, 2014. Recombinant protein production in bacterial hosts. *Drug discovery today*, 19(5): 590-601.

Picelli S, Björklund ÅK, Reinius B, Sagasser S, Winberg G, Sandberg R (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, *24*(12): 2033-2040.

Reznikoff WS, 1993. The TN5 transposon. *Annual review of microbiology*, *47*(1): 945-964.

Reznikoff WS, (2003). Tn5 as a model for understanding DNA transposition. *Molecular Microbiology*, *47*(5): 1199-1206.

Sørensen HP, Mortensen KK, (2005). Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. *Microbial Cell Factories*. doi: 10.1186/1475-2859-4-1.

Steiniger-White M, Rayment I, Reznikoff WS, (2004). Structure/function insights into Tn5 transposition. *Current Opinion in Structural Biology*, *14*(1): 50-57.

ThermoFisher. Overview of Protein Expression. Retrieved from: http://www.thermofisher.com/za/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-protein-expression-systems.html

Van Dijk EL, Jaszczyszyn Y, Thermes C (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research*, *322*(1): 12-20.

Zimmerman SB, Pheiffer BH (1983). Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or Escherichia coli. *Proceedings of the National Academy of Sciences*, *80*(19): 5852-5856.