Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Institute of Fundamental Sciences

A comparison of next-generation sequencing protocols for microbial profiling

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Genetics

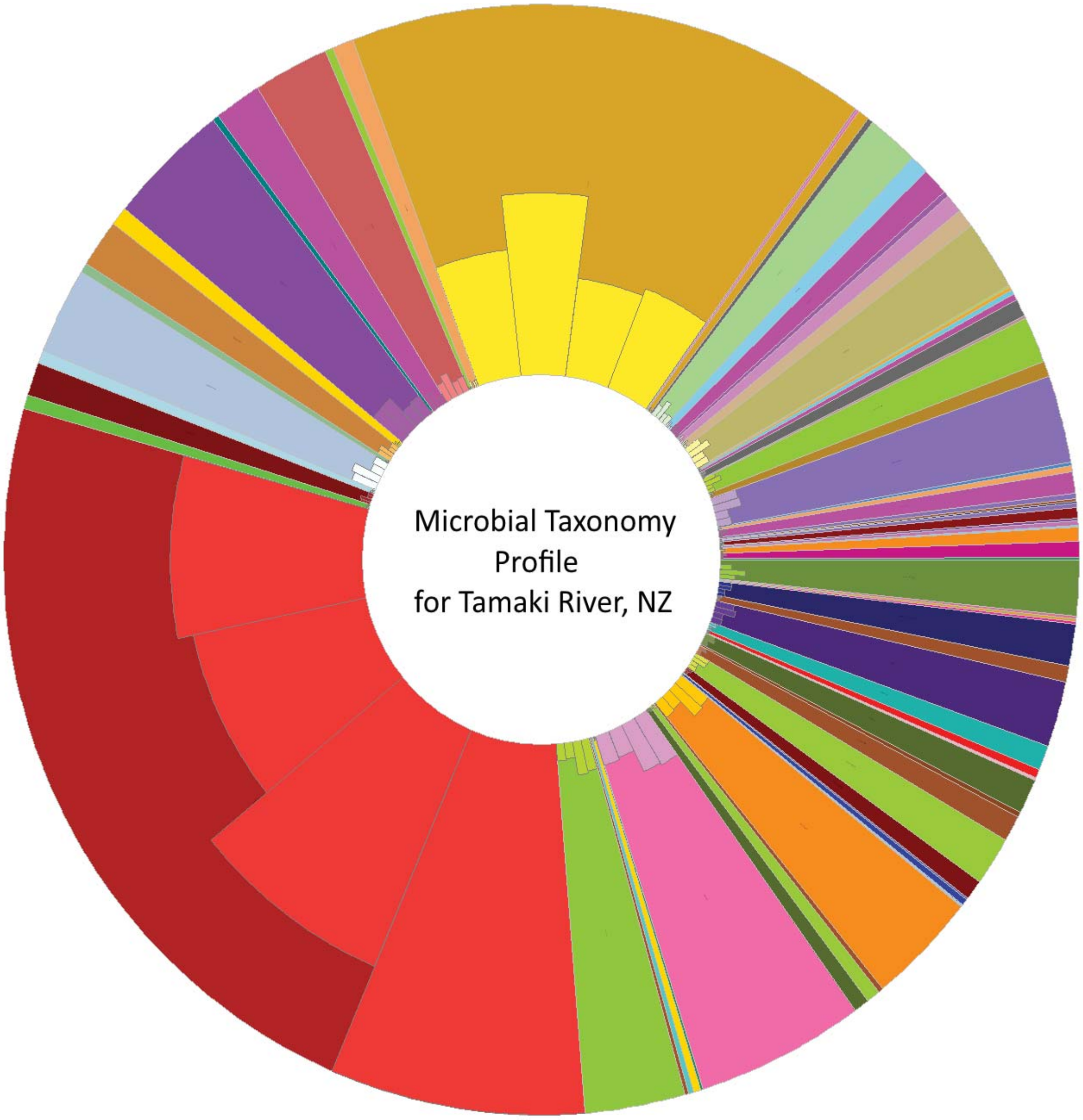
By

Yang Fong (Richard)

2016

Massey University, Palmerston North

New Zealand



One of the responsibilities faced by the Environmental Genome Project is to provide the science base upon which society can make better informed risk management decisions.

-Samuel Wilson-

Abstract

The introduction of massive parallel sequencing has revolutionized analyses of microbial communities. Illumina and other Whole Genome Shotgun Sequencing (WGS) sequencing protocols have promised improved opportunities for investigation of microbial communities. In the present work, we compared and contrasted the findings from different NGS library preparation protocols (Illumina Nextera, Nextera-XT, NEXTFlex PCR-free and Ion-Xpress-400bp) and two sequencing platforms (MiSeq and Ion-Torrent). Short reads were analysed using the rapid database matching software PAUDA and visualization software MEGAN5, which provides a conservative approach for taxonomic identification and functional analyses. In analyses of a Tamaki River water sample, biological inferences were made and compared across platforms and protocols. For even a relatively small number of reads generated on the MiSeq sequencing platform important pathogens were identified in the water sample. Far greater phylogenetic resolution was obtained with WGS sequencing protocols than has been reported in similar studies that have used 16S rDNA Illumina sequencing protocols. TruSeq and Nextera-XT sequencing protocols produced similar results. The latter protocol offered cheaper, and faster results from less DNA starting material. Proteobacteria (alpha, beta and gamma), Actinobacteria and Bacteroidetes were identified as major microbial elements in the Tamaki River sample. Our findings support the emerging view that short read sequence data and enzymatic library prep protocols provide a cost effective tool for evaluating, cataloguing and monitoring microbial species and communities. This is an approach that complements, and provides additional insight to microbial culture “water testing” protocols routinely used for analysing aquatic environments.

Acknowledgement

There are many people I would like to express my gratitude and cordial thanks in helping me out in preparing my Master's Thesis. This dissertation would not have been possible without your support and strong collaboration between different academia backgrounds.

To my lovely wife, I know I could not have done this without your constant encouragement and great patience at all times. I know that during this challenging period, you have been understanding and have given me much help, showered me with your love and support and there are no words to express my appreciation for having you by my side. To my family members; my parents, in-laws and brother, I would like to send my appreciation and would like to say thank you for being patient and for your unequivocal moral support during my master degree course. To my mentor Trish McLenachan, without you my thesis would be incomplete! Thank you for your valuable guidance, advice and patience with my writing. I send you immeasurable and deepest gratitude for your contribution in making this study worthwhile and possible.

To my principal supervisor Professor Peter-James Lockhart, co-supervisor Professor Nigel French and my bioinformatics co-supervisor Dr Patrick Biggs, I would like to thank you for being very supportive, understanding, patient and for providing precious academic and technical advice. To my principal supervisor Peter, I would like to thank you for your guidance throughout my course especially in having faith in me to finish my dissertation on time. Despite project challenges you have continued to encourage me with your knowledge and invaluable thoughts both on an academic and personal level. Special thanks to Dr Patrick Biggs for your continual support and enlightenment regarding bioinformatics analyses.

For financial support, I would like to acknowledge and thank the Institute of Veterinary, Animal and Biomedical Sciences (IVABS), Institute of Fundamental Sciences (IFS), Ministry of Health (MOH), Protozoal Research Unit (PRU, Hopkirk Research Institute). Lastly to all my fellow friends and colleagues, you guys kept me going and most importantly were understanding and patient with my workload and study. You guys are awesome!

Table of Contents

Abstract .... **I**

Acknowledgement .... **II**

List of Acronyms .... **VI**

List of Figures .... **XI**

List of Tables .... **XVII**

1 Introduction **19**

1.1 Background **19**

1.2 Common communicable diseases in New Zealand **21**

1.3 Overview of Metagenomics **23**

1.3.1 What is metagenomics? **23**

1.3.2 Types of microbial sequencing methods **24**

1.3.3 Water Metagenomics **30**

1.4 Overview of DNA Sequencing Technologies **32**

1.4.1 Illumina High-throughput Sequencing System **35**

1.4.1.1 Illumina Sequencing Chemistry **35**

1.4.1.2 Sequencing by synthesis (SBS) Illumina Sequencer **37**

1.4.1.3 Life-Technologies Ion Semiconductor Sequencing System **44**

1.5 Metagenomic analyses **49**

1.6 A role for metagenomics in studying freshwater environment. **51**

1.7 Project Outline **52**

2 Materials and Methods **53**

2.1 Sampling Sites **53**

2.2 Sample Collection **55**

2.3 DNA Extraction **55**

2.4 Colorimetric, microscopy and PCR tests **56**

2.5 Pre-NGS library validation and quantification **57**

2.6 Construction of NGS Metagenomics libraries **58**

2.6.1 Preparation of libraries for Illumina sequencing **58**

2.6.1.1 Nextera and Nextera-XT DNA Sample Preparation Method **58**

2.6.1.2 NEXTFlex PCR-free (Illumina Compatible) **60**

2.6.1.3 Ion-Torrent Library Preparation **63**

2.6.2 Illumina Sequencing **64**

2.6.2.1 MiSeq Sequencing System 64

2.6.2.2 Ion-Torrent Sequencing 65

2.7 Metagenomic data analysis 67

2.7.1 Pre-processing the metagenomics raw reads 67

2.7.2 Primary Data Analyses 69

2.7.3 Comparative Outputs and Functional Analyses 70

3 Results **72**

3.1 Microbiological tests conducted for water quality 72

3.2 DNA extraction 74

3.3 Optimisation of water filtration and DNA extraction protocols 75

3.4 Metagenomic library preparations 78

3.4.1 Nextera and Nextera-XT DNA Library Construction 78

3.4.2 NEXTFlex PCR-free DNA Library Construction 80

3.4.3 Ion-Torrent PGM Library Preparation 82

3.5 Next Generation Sequencing 84

3.5.1 Illumina Sequencing 84

3.5.1.1 MiSeq Sequencing System 84

3.5.2 Ion-Torrent PGM Sequencing 87

3.5.3 Summary for different NGS platforms and sample preparation protocols. 88

3.6 Additional QC checks 89

3.6.1 FastQC analysis 89

3.6.2 Quality Assessment using SolexaQA 99

3.6.3 Summary of results from both QC software 106

3.7 Secondary data analysis 110

3.7.1 Taxonomy classification of metagenomics reads 110

3.7.2 Functional analysis of metagenomic data using SEED and KEGG 122

3.7.2.1 SEED hierarchy with MEGAN5 122

3.7.2.2 KEGG pathway with MEGAN5 125

4 Discussion **151**

4.1 Sampling and filtration strategy 151

4.2 Optimization of NGS library preparation workflow. 152

4.2.1 Overcoming poor DNA yields from low biomass samples. 152

4.2.2 Issues with the next-generation sequencing library preparation protocols 155

4.3 Performance comparison of Illumina MiSeq and Ion-Torrent sequencers **159**

4.4 Comparison of running costs based on different workflows **163**

4.5 Computational challenges in our metagenomics analyses. **165**

5 Conclusion **168**

6 Future work **169**

References .... **XIX**

Appendix .... **XXVIII**

List of Acronyms

%	Percent
°C	Degrees Celsius
µl	Microlitre(s)
µM	Micromolar
100 PE	2 x 100 base pair paired-end read
150 PE	2 x 150 base pair paired-end read
250 PE	2 x 250 base pair paired-end read
300 PE	2 x 300 base pair paired-end read
A	Adenine
A260	Nanodrop absorbance at 260 nanometres
A280	Nanodrop absorbance at 280 nanometres
AFLP	Amplified Fragment Length Polymorphism
ATL	A-Tailing Mix
ATM	Amplicon Tagment Mix
ATP	Adenosine Triphosphate
BAM	Binary Alignment Matrix
BGI	Beijing Genomics Limited
BIPES	Illumina Multiplexed Paired-end Sequencing Adapter
BLAST	Basic Local Alignment Search Tool
bp	Base pair(s)
C	Cytosine
CCD	Charge-coupled Device
cDNA	Complementary Deoxyribonucleic Acid
contig	Continuous Sequence
CTA	A-Tailing Control

CTE	End-Repair Control
CTL	Ligation Control
ddNTP	Dideoxy Nucleotide Triphosphate
dH_2O	Distilled Water
DNA	Deoxyribonucleic Acid
dNTP	Deoxy Nucleotide Triphosphate
ds	Double Stranded
EB	Elution Buffer
eDNA	Environmental Deoxyribonucleic Acid
EDTA	Ethylenediamine Tetra-Acetic Acid
emPCR	Emulsion Polymerase Chain Reaction
ERP	End-Repair mix
EtBr	Ethidium Bromide
E-value	A parameter that describes the number of expected matches when searching a sequence database of a particular size and composition
FC	Flowcell
fq	Fastq File Format
g	Gram(s)
G	Guanine
Gb	Gigabytes
gDNA	Genomic DNA
HiFi	High fidelity enyzme
HMW	High Molecular Weight
HT1	Hybridization Buffer
Inc.	Incorporated
Indel	Small Insertion or deletion
ISFET	Ion Sensitive Field Effect Transistor

KEGG	Kyoto Encyclopedia of Genes and Genomes
LCA	Lowest Common Ancestor
LIG	Ligation Mix
Log^{10}	Logarithm to the base 10
M	Molar
Mb	Megabytes
MDA	Multiple Displacement Amplification
MEGAN	Metagenome Analyzer
MGS	Massey Genome Service
min	Minute(s)
ml	Millilitre(s)
mm	Millimetre(s)
mM	Millimolar
MPSS	Massive Parallel Signature Sequencing
mRNA	Messenger Ribonucleic Acid
mtDNA	Mitochondrial Deoxyribonucleic Acid
ng	Nanogram(s)
NGS	Next-generation Sequencing
No	Number
NPM	Nextera PCR Master Mix
NPS	Non-point Source
nt	Nucleotide
NT	Neutralize Tagment Buffer
NZGL	New Zealand Genomics Limited
OTU	Operational Taxonomic Unit
PAUDA	Protein Alignment Using a DNA Aligner

PCoA	Principal Coordinate Analysis
PCR	Polymerase chain reaction
pDNA	Pseudo DNA
PE	Paired-end
PGM	Personal Genome Machine
PhiX	Bacteriophage PhiX174
PMM	PCR Master Mix
pmol	Picomole(s)
PPC	PCR Primer Cocktail
PP_i	Pyrophosphate
Q_{10}	Phred Quality Score 1 error in 10
Q_{20}	Phred Quality Score 1 error in 100
Q_{30}	Phred Quality Score 1 error in 1000
QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
Q-score	Phred Quality Score
RNA	Ribonucleic Acid
rpm	Revolutions per Minute
rRNA	Ribosomal Ribonucleic Acid
RSB	Resuspension Buffer
RTA	Real-Time Analysis
s	Second(s)
SAM	Sequence Alignment Map
SBS	Sequencing by Synthesis
SCIMM	Sequence Clustering with Interpolated Markov Models
SEED	Database infrastructure for comparative genomics in MEGAN5 software

SMRT	Single Molecule Real Time
spp.	Species
SPRI	Solid Phase Reversible Immobilization
ss	Single Stranded
STL	Stop Ligation Buffer
T	Thymine
TAE	Tris-Acetate EDTA buffer
TAP	Taxonomic Assignment Pipeline
Taq	*Thermus aquaticus*
TB	Tuberculosis
TD	Tagmentation Buffer
TE	Tris EDTA Buffer
V	Volts
WGS	Whole Genome Shotgun Sequencing

List of Figures

Figure 1 – Upper figure shows the sequencing-by-synthesis (SBS) workflow from sample preparation to sequencing meanwhile the bottom imaging show the base calling detection via 4- and 2-channel imaging detection technology using red and green laser filters. The latter has better accuracy and faster processing time (Figure provided by Illumina Inc). **39**

Figure 2 - In 2-channel imaging there are only two images captured (red and green filters) in determining the four nucleotides bases; the red colour represents the C base, the green colour represents the T base, meanwhile yellow (combination of green and red) represents the A base and lastly for the G base, it is blue-gray in colour with no specific filter colour coding (Figure provided by Illumina Inc)..**40**

Figure 3 – Paired-end sequencing (left) showing Read 1 and Read 2 primers starting the elongation or extension of the DNA template after hybridization. The schematic on the right indicates how paired-end sequencing data can be used for elucidating the genome arrangement when aligned against a reference sequence. Paired-end sequencing can produce more accurate information due to the high number of overlapping regions of sequences and is particularly useful for difficult-to-sequence genome regions. **40**

Figure 4 – A schematic overview of the HiSeq 2000 instrument showing the reagents compartment, optical modules with dual surface imaging technology and flow cell compartments that can hold two independent flow cells for a single sequencing run. All these improvements are now controlled by an integrated touch screen monitor with a simple intuitive interface. **42**

Figure 5 – Different sequencing chemistries available for various MiSeq sequencing projects. The projected output number of reads passing filter and quality scores are based on the Illumina internal sequencing PhiX control library with a cluster density of between 850 – 980 k/mm^2 using 2 x 250 bp version 2 chemistry, and between 1200 – 1400 k/mm^2 using 2 x 300 bp version 3 chemistry. **43**

Figure 6 – Ion-Torrent PGM sequencer, A) touch screen control, B) Ion-chip loading deck clamping mechanism, C) special material grounding plate, D) power button, E) Reagent bottles, F) Wash bottles..**47**

Figure 7 – Technology behind semiconductor sequencing, A) CMOS sensor build on a wafer shape polycarbonate die, B) underlying electronics and sensors board, C) upper surface of the Ion-chip showing location for addition of sequencing reagents, D) A schematic diagram showing the technology behind semiconductor sequencing with DNA template releasing H$^+$ ions which change the pH of the well - this signal is transformed into potential voltage and sensed by the under lying sensor and electronics, E) electron micrograph showing connection between miniscule well and ISFET sensor, F) schematic diagram for the sensor detection workflow in two-dimensional array. **48**

Figure 8a – Satellite image from Google Map showing the location of our water collection site on the Tamaki River near Dannevirke, Manawatu..**54**

Figure 8b – Higher resolution satellite image from Google Map showing the GPS coordinates for the water collection site (40°09'43.2"S 176°03'50.5"E) and driveway entrance (40°09'38.1"S 176°03'50.5"E). **54**

Figure 9 - One litre “grab” water samples were filtered through 0.22 and 0.44 µm filters. The microbes were then washed from the filters and their DNA extracted into a 20 µl volume of buffer prior to NGS library construction. **57**

Figure 10 – (a) Nextera sample preparation uses a ‘transposase’ enzyme to fragment and tag DNA in a single step. (b) Primer adapters for read 1 and 2, along with individually bar-coded index i7 and i5, are added for PCR amplification before sequencing. **59**

Figure 11 - A) Agilent bioanalyzer priming station, B) priming station base plate aligning to a correct position C) syringe lock clip was set to lowest position D) the DNA 1000 chip showing the position of the wells........ **61**

Figure 12 – With the NEXTFlex protocol, 1-3 µg of starting material (gDNA) is sheared to smaller fragments. The end-repair process and size selection process are merged into a single step via the SPRI beads system that binds to the DNA accordingly to the concentration of magnetic beads. After adenylation, a new enzymatic mix is employed to enhance the adapter ligation step prior to cluster generation, without the need for a PCR enrichment step. **62**

Figure 13 – Ion-Torrent sequencing can be achieved within hours due to the speed of semiconductor ion sequencing (A). Genomic DNA is fragmented and size selected using a SPRI bead system, before the adaptor ligation step (B). Next the adapter-ligated DNA is bound to Ion Sphere particles and amplified (C). These products are then loaded onto an Ion Chip and sequenced on the Ion-Torrent machine (D)........ **63**

Figure 14 - Illumina MiSeq instrument is the only ‘all in one’ sequencer capable of producing clusters and sequencing under ‘one roof’. Sample preparation and automated real time data analysis required less than a day for sequencing 2 x 150bp paired-end reads. **65**

Figure 15 - Ion-Torrent PGM utilises semi-conductor sequencing chemistry where loaded DNA samples are supercharged with ionic electrical charges prior to sequencing. Additions of DNA bases then release a charged ion one at a time which causes a spike in the pH gradient characteristic of a particular base. The pH changes are detected and the relevant base is called by the instrument. **66**

Figure 16 - Raw reads produced from different platforms and methodologies were pre-processed: the metagenomic data was quality checked, filtered, trimmed and binned before taxonomic classification and annotation. **68**

Figure 17 - PAUDA analysis, protein reference sequences from the NCBI nr database are pre-processed with index code (pDNA) for computational analysis (PAUDA-build) before alignment with DNA reads (PAUDA-run) prior to generating outputs as BLASTX alignments. **70**

Figure 18 - Functional analyses of meta-data workflow. Blasted NCBI PAUDA data were loaded into MEGAN5 for both SEED and KEGG analyses to investigate their biological roles and also to group them together to identify different clusters of genes and functional metabolic pathways. **71**

Figure 19 – Gel electrophoresis of DNA extracted from filters with different pore sizes of 1.0 µm, 0.8 µm, 0.44 µm, 0.22 µm and 0.1 µm. From left, L = 1Kb+ ladder, PC = Positive control (*E.coli*), NC = Negative control, L1 = 1.0 µm filter, L2= 0.8 µm filter, L3 = 0.45 µm filter, L4 = 0.22 µm filter and L5 = 0.1 µm filter. The red box indicates the filters we chose for our protocol. **73**

Figure 20 - Gel electrophoresis of hmwt DNA from the duck pond water (Massey University, Palmerston North) and a positive control of *E.coli* at 5 µg. From left, L = High molecular weight DNA mass ladder, NC = Negative Control, PC = Positive control (*E.coli*), L1 = 0.45 µm filter (duck pond water), L2 = 0.45 µm filter (*E.coli* 5 µg + Milli-Q water), L3 = 0.22 µm filter (duck pond water), L4 = 0.22 µm filter (*E.coli* 5 µg + Milli-Q water). **74**

Figure 21 - DNA extracted from multiple filters (3 x 0.45 and 1 x 0.22 µm) together with positive and negative controls on a 1% (w/v) agarose gel. The band intensity for each of the filters can be compared to the previous gel (single filtration, Figure 20). The amount of DNA recovered was similar across all 4 filters. DNA from filters 1 and 2 appears to be running at a higher molecular weight compared to the DNA from filters 3 and 4 and the positive control. This result could be due to salt, or other contaminants in the final elution. All gel wells were loaded with 2 µl of purified DNA product. **76**

Figure 22 - Fragmented genomic DNA from a multiple filtration protocol. The sheared genomic DNA was within the recommended DNA peak size range of 400 to 800 bp which indicates that the fragmentation process had been successful and is suitable for MiSeq paired-end sequencing. FU: arbitrary fluorescent unit. **77**

Figure 23 A) Bioanalyzer profile for Nextera libraries indicating fragment size range of 200 to 800 bp following PCR enrichment and B) gel view showing most of the fragments were between 300 to 500 bp. FU: arbitrary fluorescent unit. **79**

Figure 24 A) Bioanalyzer profile showing that the Nextera-XT library fragments were larger than those obtained with the Nextera procedure and B) gel visualisation indicating most amplified products between 600 to 800 bp. FU: arbitrary fluorescent unit. **79**

Figure 25 – A total of 5 µl of genomic DNA from Tamaki River was loaded into lanes 1 and 2. We observed the presence of hmw DNA (yellow square) in both lanes. Both bands are strong with minimal degradation. **81**

Figure 26 – Bioanalyzer DNA 1000 profile of the PCR-free protocol showing size distribution of the library fragments. These ranged between 350 and 700bp. After nebulization the concentration of the total amount of DNA dropped from 1.2 µg to 0.93 µg. FU: arbitrary fluorescent unit........**81**

Figure 27 – A) Size distribution for 1µg of gDNA after fragmentation for 20 minutes at 25°C and 10 minutes at 70°C. After fragmentation, the majority of the DNA fragments were < 800 bp. B) Size distribution for enriched NGS library after size selection and emulsion-PCR amplification. The majority of the final library fragments were between 200 – 400 bp in size. FU: arbitrary fluorescent unit........**83**

Figure 28 – FastQC analysis on sequence quality and Kmer content of read 1 and 2 data obtained with the Nextera protocol. Per base sequence quality scores for read 2 (C) were lower than for read 1 after position 120bp. For both read 1 (B) and read 2 (D) Kmer content was high at the beginning of the raw reads and also high after position 40bp **91**

Figure 29 – Quality report for data produced using the Nextera-XT protocol. A) Quality per sequence for read 1 showed high quality with less than a 0.1% error rate. C) Read 2 showed lower quality scores compared to read 1 but still passed the quality metrics score (less than 0.1% error rate). The high kmer content likely due to Nextera-XT transposase enzyme was evident in both read 1(B) and read 2 (D). **92**

Figure 30 – A) Sequence quality for read 1 was high with almost all sequences exhibiting a phred quality score above 28 and a 0.01% error rate. C) Sequence quality for read 2 was good before 200 bases and reduced to a lower phred score of 20 after 230 bases. B) The presence of 5-mer repetitive sequences likely due to primer adapter sequences or possibly dimer contamination. These kmers were also present in read 2 (D). **94**

Figure 31 – Preliminary FastQC report for Ion-Torrent data indicating quality across length of all sequences. A) Sequences were generally of high quality up to position 200 bases before dropping to lower quality after 250 bases. B) The distribution of GC content over all sequences and peak at ~ 62% with at least 250,000 reads C) The distribution of quality scores for all DNA sequences indicating region of sequences with lowest error rate and we have at least 600,000 reads with a phred score above 30, D) This visualisation shows the presence of repetitive sequences among the reads located at 250 - 400 bp. **96**

Figure 32 – A) Quality distribution graph for read 1 sequences shows that the majority of the raw metagenomic sequences sequenced using different Illumina library protocols were of good quality (Phred score of 30 and above) with 99.9% accurate base-calling. The Ion-Torrent data was of lower quality. B) Quality distribution graph for Illumina read 2 sequences. The graph again shows that the majority of the raw metagenomic sequences were of good quality with approximately 1% error rate of incorrect base-calling. **97**

Figure 33 – SolexaQA cumulative plots and histograms showed that the majority proportion of our reads were of high quality. A) Almost 80% of our reads from read 1 have more than 100 bases and approximately 40% are at 150 bases, B) The higher quality reads are reflected on the histogram showing majority proportion of reads at 150 bases. Meanwhile we observe a drop in sequence quality for read 2 compared to read 1 where approximately 60% of reads are less than 100 bases and only 10% reads are at 150 bases (C) and this was further reflected on the histogram for read 2 (D)...**101**

Figure 34 – Similar to the Nextera protocol, both cumulative plots and histograms for Nextera-XT protocol showed that the majority of our reads were of high quality. A) Approximately 75% of Nextera-XT reads (read 1) have more than 100 bases and less than 40% of the reads have 150 bases, (B) The 150 base reads can be seen on the histogram with 1% error rate. In comparison, the read 2 quality drop earlier compared to read 1 where approximately 50% of reads now are less than 100 bases and only less than 10% reads are at 150 base reads long (C) and again this was shown on histogram plot for read 2 (D). **102**

Figure 35 – Histogram plots which show that the quality of read 1 data was better than read 2 with most of the longest fragments (251 bp) being generated from read 1 with <0.01 error rate. A slight drop in sequence quality can be observed between reads with 60 and 80 bp long fragments. This is indicated by a small spike in both read 1 and read 2 histograms.....**103**

Figure 36 – Histograms with 1% error rate showing the length of the contiguous read data generated from the Ion-Torrent PGM after trimming. For trimming, the data was run through LengthSort set to 75bp with p-value of 0.01 and this gave 2,229,013 (41.06%) good quality single reads (phred score >20) and 3,199,123 (58.9%) discarded reads (Table 11). The figure shows only the relative proportion of sequencing reads after trimming with maximum fragments length at 52 bp...**105**

Figure 37 – Combined cumulative SolexaQA plots for metagenomic data obtained using different library preparation protocols. A) Summary for read 1 showed a comparison of sequences from two sequencing platforms and different library protocols utilised in this project. We observed a mixture of high and poor quality data across the library protocols and across the platforms (B) Cumulative plot for Read 2 data indicating that the majority of sequences (~50%) with Q_{20} scores were less than 100 bases in length. There is no data for Read 2 for Ion-Torrent because it was only a 400bp single read run since paired-end read chemistry is still not available...**107**

Figure 38 – Twenty most represented bacterial genera in the Tamaki River sample (number of reads indicated via log scale algorithm). The top three bacterial genera were *Pseudomonas*, *Yersinia* and *Serratia*. The presence of *E. coli* in our Tamaki River sample (via colorimetric result, see result section 3.1) was also consistent with our NGS data as the genus *Escherichia* was within the top 10 bacterial genera...**112**

Figure 39 – Forty most represented species (99.5% hit/0.01 cut-off point) common to three different libraries (Nextera, Nextera-XT and NEXTFlex PCR-free). We observed two of the most abundant *Pseudomonas fluorescens and Yersinia enterocolitica* across all preparation methods. *E.coli* was observed to be the 11th most abundant species which is consistent with our calorimetric result (see result section 3.1). The number of sequencing reads for NEXTFlex PCR-free, Nextera and Nextera-XT protocols were similar. Pie Chart showing the relative proportion of taxa identified in the Illumina libraries..**113**

Figure 40 – Relative proportions of identified taxa generated from the Illumina MiSeq instrument (Nextera, Nextera-XT and NEXTFlex PCR-free protocols) **115**

Figure 41 – Forty most abundant genera identified in the Ion-Torrent library. The profile is similar to that observed with the Illumina libraries in that *Pseudomonas fluorescens* (981,825 reads) and *Yersinia enterocolitica* (334,391 reads) species were most abundant, followed closely by *Pseudomonas putida* (145,957 reads), *Pseudomonas aeruginosa* (103,367 reads) and *Pseudomonas sp, CMAA1215* (93,998 reads). The low abundance of the bacterium *Esherichia coli* (43,901 reads) was also evident...**116**

Figure 42 - Pie chart indicating relative proportion of the most abundant genera in the Ion-Torrent library. According to the analysis, *Pseudomonas fluorescens* (981,825 reads) made up more than 30% of the total bacterial population found in Tamaki River Water sample. .**118**

Figure 43 – Histogram of entire microbial profile (instead of just top 20 species) found in the Tamaki River obtained from all metagenomic datasets generated from different library preparation protocols and sequencing platforms. The bacterial composition for the 40 most abundant species was similar for different NGS protocols. The PCR-free protocol which used a longer read length (250 PE) sequencing chemistry had better sensitivity in detecting more species compared to other protocols. Five species: *Herbaspirillum seropedicae*, *Mesorhizobium* opportunistum, *Brevundimonas subvibrioides*, *Yersinia aldovae* and *Pseudomonas coronafaciens* were only present in the Ion-Torrent data. In addition, the bacterial '*phiX174'* was absent in the Ion-Torrent data which require further explanation. .**120**

Figure 44 – A word cloud for all the metagenomic datasets originated from the MEGAN5taxonomy profiles (A: Nextera, B: Nextera-XT, C: PCR-free and D: Ion-Torrent) showing the most abundant bacterial species in the Tamaki River sample..........................**121**

Figure 45 – Summary of SEED subsystems analysis showing different library sequences assigned to different categories of biological niche. These categories were carbohydrate synthesis, amino acid and derivatives synthesis, protein, DNA and RNA metabolism along with virulence factors and associated disease. There was no normalisation of the sequences for this data analyses. **124**

Figure 46 – KEGG analysis for Tamaki River showing main classifications of functional content for metagenome datasets. This analysis involved assigning sequence reads to KEGG orthology categories - metabolism, cellular processes, genetic information, environmental processing, risk of human disease and organismal system. There was no normalisation of the sequences for this data analyses. **133**

Figure 47 – KEGG analysis for the "metabolism" pathway indicating "carbohydrate and energy metabolism" categories and subcategories. The relative number of reads assigned to different functional nodes is shown. The figures in red indicate the number of sequencing reads assigned to the respective functional content network by the KEGG orthology system**.** Here we have chosen six important categories for the functional analysis: Gluconeogenesis, the Citrate cycle, Fructose and Mannose metabolism, amino and nucleotide sugars metabolism, oxidative phosphorylation and finally the Nitrogen metabolism. **135**

Figure 48 – KEGG analysis on "Human Disease" functional hierarchy indicating potential associations of infectious disease from our metagenomics datasets. Functional nodes highlighted in red show six potential infectious diseases (*Vibrio cholera, Helicobacter pylori, Salmonella, Bordetella pertussis, Legionella and Mycobacterium tuberculosis*) of interest in our project together with the number of sequencing reads assigned to it. Most of the assigned sequencing reads are very low due to lower coverage and average quality sequences. **137**

Figure 49 – Tricarboxylic Acic Cycle (TCA) or also known as Krebs cycle is an important metabolic pathway for generation of energy in many bacterial species. The figure shows genes mapped to the TCA pathway from our metagenomics reads. **142**

Figure 50 – The nitrogen metabolism cycle is used by many bacteria for processing organic and inorganic nitrogen compounds for ammonification, mineralisation, nitrification and denitrification processes. Reads mapping to key pathway steps have been indicated. **144**

Figure 51 – Pathogenesis-associated colonization of *V.cholerae* cycle. This cycle shows the dual life cycle of *V.cholerae* in the aquatic environment and in the host during virulent phase when colonizing the human small intestine. **146**

Figure 52 – *V.cholerae* infection pathway (in human) indicating the steps required for virulence i.e. secretion of Cholera toxin (CTX). The highlighted area is where our metagenomics reads have been mapped. **148**

Figure 53 – Microbial attributes co-occurrence chart plotted from KEGG analysis classification based on reads from MiSeq Nextera, MiSeq Nextera-XT, MiSeq NEXTFlex PCR Free and Ion-Torrent PGM sequencing protocols. The chart indicates common functional gene group relationships in the Tamaki river microorganisms. **150**

Figure 54 – The workflow above shows the main steps followed in the current project. Different library preparation protocols were used for NGS sequencing. All were able to detect a wide range of microbial species. **152**

List of Tables

Table 1 – Screening for *Cryptosporidium*, *Giardia* and *E.coli* in Tamaki River grab samples collected in November 2011. Only one sample (number 3) tested positive for coliform bacteria with the colorimetric test. Samples 4 and 5 tested positive for *Cryptosporidium* and *Giardia* with the PCR test....**72**

Table 2 - Quality and quantity measurement for a single filter and multiple (0.45 and 0.22 µm) filters. We used both Qubit and Nanodrop instruments for this assessment. Most of the sample purities were within an acceptable range for the construction of a NGS library (1.8 to 2.0). The concentration of DNA in the 0.22 µm final pooled samples from multiple filters was significantly higher compared to that obtained with single 0.45 µm filter....**78**

Table 3 – Qubit quantification readings obtained from Qubit fluorometer for protein, RNA and DNA assays. Both Nextera and Nextera-XT libraries showed an acceptable level of protein and RNA (less than 1 ng/µl) with total DNA concentration of 51 ng/µl. **80**

Table 4 – Quantification of protein, RNA and dsDNA levels made with a Qubit fluorometer. The Tamaki River sample had less than 1% RNA and protein contamination. The average library fragment size was at 581bp. **82**

Table 5 – The Sequencing Analysis Viewer (SAV) summary report indicated that we had a total data output of 2.19 Gb with less than 0.6% error rate and 99.4% base-calling accuracy. We obtained an optimal cluster density of 961 k/mm^2 with an average passing filter Q_{30} score of 82.9% for the Tamaki river water sample. **85**

Table 6 – A total of 2.61 Gb was generated for this run with error rates less than 0.5% and 99.5% accuracy for nucleotide base calling. We obtained a high cluster density of 1121 k/mm^2 for which 85% data was categorised as 'good quality'....**86**

Table 7 – Run summary indicating that there was a total of 2.34 Gb of data generated from the 2x250 bp paired-end sequencing run. The table indicates a 0.6% total error rate with the final library loading molarity of 2nM . We observed a total cluster density of 1003k/mm^2 with 93.4% passing the quality filter. **86**

Table 8 – Summary statistics indicating the amount of raw data output, number of raw reads along with the percentage of wells with ISP beads. The table also shows that most reads had a length of 147.7 bp and phred quality mean score of 34.7. 90.9%. of the reads had an AQ20 read length score. These scores are similar to Phred-like scores. Here, AQ20 quality refers to a phred-like score of 20 or better, where there is one error rate per 100 bp....**87**

Table 9 – Summary of NGS raw data output from different instruments and library preparations. All NGS libraries were normalised to 2nM concentration before being loaded for sequencing....**88**

Table 10 – Pre-processing (FastQC) quality assessment for sequences obtained using Nextera, Nextera-XT, NEXTFlex PCR free and Ion Xpress-400bp) sequencing protocols....**98**

Table 11 – Summary of SolexaQA reports software for Nextera, Nextera-XT, NEXTFlex PCR free and lastly Ion Xpress-400bp data....**108**

Table 12 – SEED subsystems classification on four metagenomic dataset: MiSeq Nextera (dataset 1), MiSeq Nextera-XT (dataset 2), MiSeq NEXTFlex PCR Free (dataset 3) and Ion-Torrent PGM (dataset 4). The number of assigned reads were filtered under standard correlation of 0.01 error rate to each functional biological nodes. Please note highlighted area shows large proportion of the NGS reads were binned to ‘unknown’ or ‘unassigned’ due to ambiguous nucleotide base-calling i.e. homopolymeric regions with many repetitive sequences. **123**

Table 13 – Tabulated data showing the number of matching reads to KEGG hierarchy pathway system from four datasets (Dataset 1 – MiSeq Nextera, Data 2 – MiSeq Nextera-XT, Dataset 3 – MiSeq NEXTFlex PCR Free and Dataset 4 – Ion-Torrent PGM). Please note the percentage of sequence reads matched to each pathway was calculated by dividing the individual reads from each pathway by the overall total number of reads from each dataset. **126**

Table 14 - Summary statistics for the number of metagenomic sequences used for the KEGG analysis in this project. The percentage of matching KEGG was calculated from the number of reads assigned to all KEGG hierarchy divided by the total number of reads used in the KEGG analysis..**127**

Table 15 – Enzymes found in our metagenomic datasets from the TCA pathway, along with number of sequencing reads assigned to KO and EC numbers **131**

Table 16 – Metabolic enzymes responsible for nitrogen metabolism found in our metagenomic sample with KEGG orthology and enzyme nomenclature (EC) numbers and their essentiality. **132**

Table 17 – Genes associated with *V. cholerae* pathogenesis and its functionality in our metagenomics datasets...**139**

Table 18 – Genes and enzymes from our metagenomics dataset linked to the *V. cholerae* infection pathway..**140**

Table 19 – Summary of specifications of NGS platforms compared in our metagenomics project. **164**