Chapter 4

Result

4.1 Documentation and procedure of making *napham*

4.1.1 Raw Materials used

Napham is a traditional cuisine of the Bodo tribe. Small and locally available fishes are used for the preparation of *napham*. A total of eighteen fish belonging to 13 genera and 10 families were documented in the present study. A combination of four or five types of fish is used according to the availability of fish. But most commonly used fish were Puntius spp., *Amblypharyngodon mola*, *Lepidocephalichthys guntea*, *Parambassis* spp., *Trichogaster* spp., *Channa* spp., and *Danio rerio*.

The fishes documented and identified were:

1. **Family: Cyprinidae**

Fish name: *Amblypharyngodon mola* (Hamilton, 1822), Local name-na maowa; *Esomus danrica* (Hamilton, 1822), local name-borali; *Danio rerio* (Hamilton, 1822), local name-na miji; *Puntius chola* (Hamilton, 1822), *Puntius terio* (Hamilton, 1822), *Puntius sophore* (Hamilton, 1822), local name-na phitikri.

2. **Family: Cobitidae**

Fish name: *Lepidocephalichthys guntea* (Hamilton, 1822), local name-na balabathia

3. **Family: Bagridae**

Fish name: *Mystus bleekeri* (Day, 1877), *Mystus carcio* (Hamilton, 1822), local name- na thengwna

4. **Family: Mastacembelidae**

Fish name: *Macrognathus pancalus* (Hamilton, 1822), local name- na thuri;

5. **Family: Ambassidae**

Fish name: *Parambassis ranga* (Hamilton, 1822) and *Parambassis lala* (Hamilton, 1822), local name-na chandanga

6. **Family:Nandidae**

Fish name: *Nandus nandus* (Hamilton, 1822), local name-na thotha

7. Family: Badidae

Fish name: *Badis badis* (Hamilton, 1822), local name-na duthumwi

8. Family: Gobiidae

Fish name: *Glossogobius giuris* (Hamilton, 1822), local name-na mutura

9. Family: Osphronemidae

Fish name: *Trichogaster fasciata* (Bloch & Schneider, 1801*) Trichogaster lalius* (Hamilton, 1822), local name –na bingsi

10. Family: Channidae

Fish name: *Channa gachua* (Hamilton, 1822), local name- na gwri.

Plant materials used in *napham*:

The other raw materials added in preparation of *napham* are tender shoots of *Colocassia esculanta* (L.) Schott. (Local name- thaso), Ash gourd, {Loc name: khumbra; *Scietific name: Benncasa hispida* (Thunb.) Cogn.)}, leaf of mwitha (Scietific name: *Hibiscus sabdariffa* L.) and papaya (*Carica papaya*) are used as additives. But the majority of the informants reported the use of *Colocasia esculanta*.

4.1.2 Procedure (Photoplate I)

The fishes were gutted, washed, drained, dried in the sun, and smoked under low flame. The raw material was then pounded with mortar and pestle (called *uwal* and *gaihen* in the local Bodo language). Young stems of *Colocassia esculanta* were used as an additive during pounding to prepare a mixture paste. The container made of the hollow bamboo stem that is open from one side and closed from the other by an inter-node is used to insert the raw mixture paste. The thin layer of traditional alkali or *kharwi* covers the mixture produced after pounding (*kharwi* is a local alkali prepared by burning dried banana stem, mustard plant/black lentil plant remains. pH: 7 to 9). The content was then tightly pressed and covered with banana leaves. After inserting the content, the opening of the bamboo container was tightly sealed with a paste prepared from ash and mud mixture to ensure the anaerobic condition. The preparation was kept for fermentation for three months. The recorded shelf life of *napham* was from six months to one year, and sometimes this period may extend further. In the present survey, people preferred

using glass bottles instead of bamboo stem containers. The tradition of allowing the fermentation in the hollow bamboo stem is less seen among the new generation of *napham* makers. All the vendors in the study area used bottles except the vendor in Jornagra Pt. I.

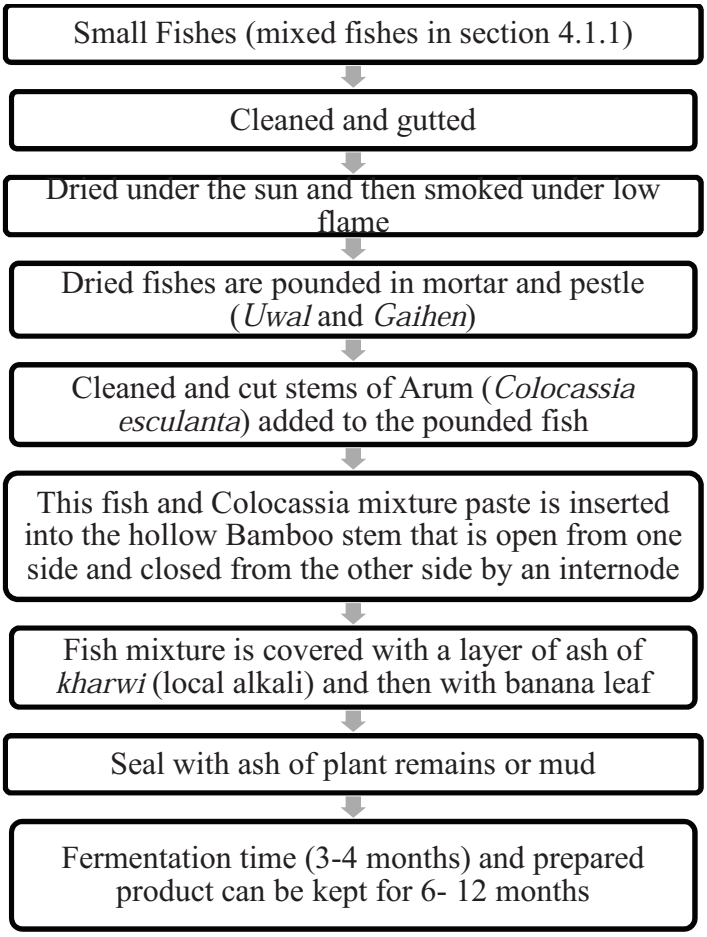


Figure 4.1 Procedure of traditional method of preparation of *napham*

4.1.3 Culinary preparation from *napham*

Napham is used as a flavoring ingredient or condiment in many traditional recipes of Bodo cuisine. It is cooked by tempering it with onion, garlic, and sometimes various vegetables. *Napham* soup is prepared by boiling it in water and mixing it with garlic, ginger, salt, turmeric, coriander, and onion leaves. It is also used as an additive in dishes like *sobai* (Black lentil) and *narji* (Dried jute leaves). In rural areas, it forms a good source of protein in the diet. Some cuisines prepared from *napham* are: *napham-bathwn*, *napham thaso-bisong*, *napham thaso-bithorai, napham bidwi*, *napham thaso-athing, napham narji, and napham sobai* (Photoplate II).

4.1.4 Cultural values and livelihood prospects of *napham*

The demand for *napham* is high in the markets of Kokrajhar, Dotma, Kochugaon, and Gossaigaon. *Napham* is sold in the range of Rs.1000 to Rs.1500 per container depending on the size of the container. Amongst the Bodo tribe, there is a strong tradition of the fishery. The Bodo people love fishing and ensure sustainability by using every bit of the harvest. The leftover small fishes are generally used to prepare *napham* so that no amount of fish remains wasted. It is also preserved for the period when the availability of fish is decreased.

4.2 Biochemical analysis

4.2.1 Proximate composition of *napham*

Table 4.1(a) proximate composition of *napham*

Proximate composition	Fermented fish *napham* Average ± SD
Moisture content (%)	38.8 ± 3.6
Ash content (%)	15.5 ± 2.9
Crude protein (%)	30.3 ± 4.3
Crude Fat (%)	24.4 ± 0.6
pH	7.03 ± 0.3

Table 4.1(b) Moisture, ash, crude protein contents and pH during different stages of fermentation

Sample name	Moisture content Average ± SD	Ash (%) Average ± SD	Crude protein (%) Average ± SD	pH
Raw sample	25.5 ± 0.02	16.7 ± 0.02	51.9 ± 2.2	6.5 ± 0.01
First month	40.3 ± 0.9	18.6 ± 0.3	31.5 ± 3.4	7.16 ± 0.01
Second month	41.2 ±1.1	16.3 ± 0.2	30.6 ± 3.5	7.15 ± 0.4
Sixth month	36.6 ±1.02	16.2 ± 0.2	29.7 ± 6.3	7.12 ± 0.02
Twelve month	36.4±1.3	11.8±0.21	29.47±5.8	7.2 ± 0.005
Kruskal-Wallis test	H = 8.95, df=3, P=0.03	H=9.46,df=3, P=0.02	H=0.128,df=0.13 P=0.98	H=8.2,df=3 P=0.04
Significance	Significant	Significant	No significant difference	Significant

For statistical analysis, the Kruskal Wallis test was carried out with PAST software to see the significant differences in proximate composition in four samples belonging to four different months of fermentation. A significant difference was found in moisture content, ash content, and pH, and no significant difference was found in protein content.

4.2.2 Analysis of Minerals and trace elements

The macro minerals present were Calcium, Sodium and Magnesium. The micro minerals present were Iron, Zinc, Copper, Potassium, Manganese, Molybdenum, and Chromium. Ultra- trace minerals present were Nickel and Cobalt. The detail of constituent and composition of minerals analyzed in *napham* is shown in Table 4.2.

Table 4.2 Mineral and trace element profile of *napham*

Sl. No.	Elements	Concentration in µg/g
1.	Iron (Fe)	9.1 ± 0.3
2.	Cobalt (Co)	0.21 ± 0.01
3.	Copper (Cu)	4.7 ± 0.02
4.	Potassium (K)	10.7 ± 0.7
5.	Manganese (Mn)	2.8 ± 0.001
6.	Chromium (Cr)	0.19 ± 0.001
7.	Nickel (Ni)	0.13 ± 0.1
8.	Calcium(Ca)	68.9 ± 0.2
9.	Zinc (Zn)	3.2 ± 0.004
10.	Molybdenum (Mo)	0.08 ± 0.006
11.	Sodium (Na)	67.3 ± 0.2
12.	Cadmium (Cd)	0.07 ± 0.02
13.	Lead (Pb)	0.16 ± 0.1
14.	Arsenic (As)	nd
15.	Magnesium (Mg)	32.2 ± 0.9

*The values are given as average and ± indicates standard deviation, nd =Not Detected

4.2.3 Analysis of Amino acid

Amino acids are the significant components of fish. The dominant amino acids detected in F1 were aspartic acid, asparagine, and serine. In F2, glutamine, aspartic acid, and alanine were detected in high concentration. Aspartic acid was present in

dominant amino acid in all the samples (chromatogram of HPLC in appendix annexure II). Tryptophan, threonine, serine, phenylalanine, valine, lysine, leucine, isoleucine, methionine, and histidine were the nine essential amino acids detected in *napham*. The non-essential amino acids detected were proline, aspartic acid, carnosine, tyrosine, glutamic acid, alanine, glycine, asparagine, and arginine. The detailed composition of all amino acids present in *napham* is presented in Table 4.3.

Table 4.3 Composition of amino acid in raw sample, F1 and F2

Sl. No.	Composition of Amino acid in	Raw sample (pg/mg) Average + SD	Fermented product (*napham*) F1 (pg/mg) Average + SD	Fermented product (*napham*) F2 (pg/mg) Average + SD
1.	Aspartic acid	16.3 ± 2.4^{a}	253.3 ± 3.1^{b}	297.3 ± 9.3^{c}
2.	Threonine *	35.9 ± 5.2^{a}	152.7 ± 7.1^{b}	96.2 ± 26.7^{c}
3.	Serine *	25.3 ± 12.1^{a}	155.7 ± 8.3^{b}	85.5 ± 17.2^{c}
4.	Asparagine	93.7 ± 16.4^{a}	170.7 ± 5.2^{b}	118.4 ± 11.4^{c}
5.	Glutamic acid	115.5 ± 12.2^{a}	104.01 ± 4.4^{b}	123.6 ± 7.1^{c}
6.	Glutamine	66.9 ± 10.4^{a}	40.4 ± 1.3^{b}	40.6 ± 6.5^{b}
7.	Glycine	10.4 ± 0.8^{a}	40.2 ± 2.3^{b}	6.5 ± 3.6^{c}
8.	Alanine	7.02 ± 0.7^{a}	41.5 ± 0.2^{b}	4.2 ± 3.1^{a}
9.	Valine *	78.2 ± 10.9^{a}	80.6 ± 5.5^{b}	70.7 ± 2.3^{c}
10.	Cysteine	15.4 ± 1.2^{a}	15.8 ± 1.5^{a}	24.7 ± 2.1^{b}
11.	Methionine *	17.5 ± 4.1^{a}	133.9 ± 2.2^{b}	29.3 ± 7.4^{c}
12.	Iso leucine *	21.02 ± 5^{a}	122.3 ± 1.8^{b}	58.1 ± 3.6^{c}
13.	Leucine *	65.7 ± 10^{a}	106.9 ± 3.6^{b}	33.8 ± 23^{c}
14.	Tyrosine	35.7 ± 2^{a}	11.4 ± 1.7^{b}	2.6 ± 1.6^{c}
15.	Phenylalanine	11.8 ± 1^{a}	11.2 ± 0.2^{a}	8.8 ± 1.6^{a}
16.	Tryptophan *	8.1 ± 0.6^{a}	8.1 ± 0.7^{a}	5.9 ± 1.6^{b}
17.	Histidine *	1.7 ± 0.2^{a}	8.6 ± 0.4^{b}	2.2 ± 1.2^{a}
18.	Lysine *	8.4 ± 0.5^{a}	9.46 ± 0.1^{a}	16.5 ± 1.2^{b}
19.	Arginine	5.7 ± 0.5^{a}	5.66 ± 0.5^{a}	3.1 ± 2.1^{b}
	Total	640.1	1471.2	1027.7

*Denotes essential amino acids ± SD (n=3). The different letters in each raw denote significant difference ($p \leq 0.05$) otherwise no difference

4.2.4 Fatty acid composition

During the study, twenty-five fatty acids were detected in F1, twenty-six fatty acids in F2, and 32 in raw samples.

The major fatty acids present in F1 were Palmitic acid, Alpha-Linolenic acid (PUFA), Alpha-Linoleic acid (PUFA), Linoleic-acid (PUFA), and 4, 7, 10, 13, 16, 19-Docosahexaenoic acid (DHA).

The major fatty acids present in F2 were Palmitic acid, Linoleic-acid, 11-Hexadecenoic acid and 4. 7, 10, 13, 16, 19 Docosahexaenoic acid (DHA).

The intermediate compounds like12-Methyltetradecanoatel-Phenylalanine, Methyl Ester, Methyl Arachidonate and L-Leucine, N-Capryloyl-, Methyl Ester were present in a higher percentage in all the samples. The other intermediate compounds detected in F1 & F2 are represented in table 4.5 (chromatogram of GC-MS, Retention time in Appendix, Annexure II). The detailed composition of all fatty acids present in *napham* is given in Table 4.4.

Table 4.4 Composition of fatty acids in raw sample, F1 and F2

Sl. No.	Composition of Fatty acids (%)	Raw sample Average + SD (%)	Fermented product (*napham*) F1 Average + SD (%)	Fermented product (napham) F2 Average ± SD (%)
1.	Arachidonic acid	0.6 ± 0.08[a]	0.8 ±0.6[a]	0.14 ± 0.01[a]
2.	4,7,10,13,16,19-Docosahexaenoic acid(DHA)	3.7 ± 0.14[a]	8.03 ± 3.8[b]	5.07 ± 5.5[b]
3.	Lauric Acid	0.5 ± 2.01[a]	0.6 ± 2.2[a]	nd
4.	5,8,11,14,17-Eicosapentaenoic acid	0.3 ± 0.6[a]	0.4 ± 0.6[a]	2.2 ± 9.7[b]
5.	5,8,11,14-Eicosatetraenoic acid	1.4 ± 0.04[a]	1.9 ± 0.02[a]	1.9 ± 0.8[a]
6.	Eicosanoic acid	0.2 ± 0.14[a]	0.4 ± 0.5[a]	1.1 ± 0.03[b]
7.	11-Eicosenoic acid	nd	5.5 ± 1.8[a]	0.83 ± 0.06[b]
8.	Margaric acid	2.4 ±1.03[a]	3.3 ± 0.5[b]	3.1 ± 0.70[b]
9.	Palmitic acid	14.4 ± 0.1[a]	16.01±1.9[b]	9.07 ± 0.6[c]
10.	Z-7-Hexadecenoic acid	1.4 ± 0.01[a]	2.1 ± 0.1[a]	nd
11.	11-Hexadecenoic acid	5.1 ± 0.09[a]	2.9 ± 0.03[b]	6.2 ± 2.5[a]
12.	6-Hexadecenoic acid	0.3 ± 0.2[a]	0.3 ± 0.2[a]	nd
13.	Nonadecanoic acid	1.12 ± 0.5[a]	0.4 ± 0.2[a]	1.3 ± 0.5[a]
14.	Pentadecanoic acid	1.04 ± 1.6[a]	0.3 ± 0.8[a]	1.4 ± 0.1[a]
15.	Myristic acid	0.7 ± 1.4[a]	2.1 ± 1.03[b]	2.4 ± 0.03[b]
16.	Tridecanoic acid	0.8 ± 0.01[a]	0.2 ± 0.04[a]	0.2 ± 0.8[a]
17.	Linoleic-acid	1.4 ± 3.5[a]	10.4 ± 1.3[b]	10.3 ± 0.4[b]
18.	9-Octadecenoic acid	6.9 ± 2.4[a]	1.1 ± 0.8[b]	1.1 ± 0.01[b]
19.	Alpha-Linoleic acid	6.3 ± 0.1[a]	9.5 ± 2.02[b]	1.03 ± 0.6[c]
20.	Oleic acid	0.4 ± 0.5[a]	0.8 ± 0.1[a]	nd

Continued from table 4.4

Sl. No.	Composition of Fatty acids (%)	Raw sample Average + SD (%)	Fermented product (*napham*) F1 Average + SD (%)	Fermented product (napham) F2 Average ± SD (%)
21.	Stearic Acid	11.3 ± 0.5^{a}	2.02 ± 0.7^{b}	1.8 ± 0.72^{b}
22.	Oxalic acid	nd	1.9 ± 0.5	nd
23.	10-Undecenoic acid	nd	0.38 ± 0.5	nd
24.	Gamma-Linolenic acid	1.7 ± 1.8^{a}	2.5 ± 0.7^{b}	2.5 ± 0.5^{b}

*Denotes essential amino acids ± SD (n=3). The different letters in each raw denote significant difference ($p \leq 0.05$) otherwise no difference.

Table 4.5 Intermediate compounds found in *napham*

Sl. No.	Intermediate compounds (%)	Intermediate compounds present in *napham* Average ± SD (%)
1.	(Diisopropylamino) Phosphanylboron Chloride	1.6 ± 0.1
2.	12-Methyltetradecanoatel-Phenylalanine, Methyl Ester	13.1 ± 0.3
3.	4,4'-Ethylenebis(2,6-Di-Tert-Butylphenol)	0.19 ± 0.4
4.	5-Cholesten-3beta-Yl Isobutyl Carbonate	1.5 ±1.2
5.	7,8-Carbonyldioxy-2-(Trifluoromethyl)Bicyclo[2.2.2]Octa2,5-Diene	0.6 ± 3.6
6.	9-Octadecenamide, (Z)- (CAS) Oleoamide	0.4 ± 0.1
7.	Androstan-7-One, (5.Alpha.)-	0.4 ± 0.1
8.	Anisole, 2-(Benzyloxy)-5-(2-Nitrovinyl)-	0.7 ± 0.7
9.	Cholest-5-En-3-Ol (3.Beta.)-, Acetate (CAS) Cholesteryl Acetate	1.7 ±1.6
10.	Cholesta-4, 6-Dien-3-Ol, (3.Beta.)- (CAS) 4,6-Cholestadien3.Beta.-Ol	0.3 ± 2.4
11.	Dianhydromannitol	0.5 ± 0.1
12.	DL-Phenylalanine, N-Formylundecenoate	0.6 ± 6.1

13.	Dodecane, 1,1'-Oxybis- (CAS) Didodecane Ether	1.2 ± 0.1
14.	L-Leucine, N-Capryloyl-, Methyl Ester	1.9 ± 0.9
15.	L-Valine, N-(3-Cyclopentylpropionyl)-, Methyl Ester	0.8 ± 2.5
16.	Methyl (Z)-5,11,14,17-Eicosatetraenoate	1.3 ± 0.1
17.	Methyl 10d-Hydroxyoctadecanoate	1.05 ± 4.3
18.	Methyl Arachidonate	10.9 ± 5.8
19.	Methyl Eicosa-5,8,11,14,17-Pentaenoate	3.5 ± 1.3
20.	Octadecanamide	0.3 ± 0.07
21.	Palmitic Acid Vinyl Ester	0.5 ± 3.6
22.	Phthalic Acid, Methyl 2-Phenylethyl Ester	0.5 ± 1.5
23.	Propyl Carbonate Cholesterol	1.04 ± 2.01
24.	Thiosulfuric Acid (H2S2O3), S-(2-Aminoethyl) Ester	0.4 ± 0.03
25.	Tyramine, N-formyl	1.3 ± 0.04
26.	Retalin	21.1 ± 7.5

4.3 Microbial Analysis: Enumeration of microbial community

Napham is a fermented food product that harbors a complex microbiome with a diverse microbial community. In the culture-dependent method total seven selective and enriched culture media were used to enumerate the microbial load in raw material and in four fermented samples: 1. Raw Sample, 2. One-month-old sample (S1), 3. Two months old sample (S2), 4. Sixth months old sample (S3) and 5. Twelve months old sample (S4). The result of enumeration of microbial load at different stages of fermentation is given in Table 4.6 (a).

Table 4.6 Enumeration of microbial communities in four different stages of fermentation

Media	Raw sample R1	S1	S2	S3	S4
PCA	3.6 ± 0.6	5.9 ± 0.09	7.2 ± 0.19	6.5 ± 0.03	4.6 ± 0.3
PCA with 10% NaCl	3.71 ± 0.08	6.94 ± 0.02	6.97 ± 0.03	6.8 ± 0.3	6.7± 0.6
Nutrient Agar with 1% Casien hydrolysate	3.8 ± 0.05	6.8 ± 0.02	6.9 ± 0.04	5.7± 0.08	5.8 ± 0.1
M-17 Agar	3.7 ± 0.009	6.2 ± 0.01	6.3 ± 0.04	0.3 ± 0.5	2.3 ± 0.01
PDA	1.9 ± 0.4	2.6 ± 0.4	2.1 ± 0.3	2.1 ± 0.3	1± 0.3
MRS Agar supplemented with $CaCO_3$	nd	5.9 ± 0.2	6.54±0.4	5.6 ± 0.5	nd
EMB Agar	2.7 ± 0.02	4.6 ± 0.7	4.9 ± 0.18	nd	nd
Kruskall wallis test H=19.71 df=43.67 P=0.00184 Significant difference in microbial count amongst different samples of different months					

The values denotes log cfu/g of sample, nd =Not Detected.

4.4 Whole metagenome analysis

Three samples S1 (1 month old) and S2 (2 months old) and one 12 month old sample were outsourced to study the microbial diversity of the fermented fish *napham*. The concentration and quality of extracted DNA was evaluated in 2% agarose gel (Fig: 4.2) and the total amount of extracted DNA is given in table 4.7. Out of three samples, visible bands were seen only for samples S1 and S2 but no visible band was seen for one year old sample given in lane-1.Therefore, only two samples S1 and S2 were continued for library preparation, sequencing and bioinformatics pipeline.

Figure 4.2 Gel profile of the samples S1 (lane-3), S2 (lane-2) and one year (lane-1) sample in 2 % Agarose

Table 4.7 Raw data of DNA isolation of samples S1 and S2

Sample Name	Qubit concentration ng/µl	Nanodrop concentration ng/µl	260/280	Total Conc.	Result
S1	8.96	77.02	1.87	313.6	DNA band visible
S2	8.40	41.65	1.76	294	DNA band visible

4.4.1 Raw read of the sample

The sequencing of the sample on IlluminaHiSeq provided the read length of 250x2 nucleotide base pairs. The raw reads obtained from Illumina sequencing platform after Demultiplexing was subjected to FastQC program (latest version.0.11.8) to check the quality of the reads with default parameters. The base quality (PhredScore; Q), base composition, GC content, ambigious bases (other than A, T, G, C) and adapter dimers were thoroughly checked prior to the Bioinformatics analysis. The total sequence read obtained after sequencing and QC was 37,947,065. The raw read summery of samples is given below in table 4.8.

Table 4.8 Raw reads summary

Sample	Mean read quality (Phred score)	Number of reads	% GC	% Q<10	% Q 10–20	% Q 20–30	% Q >30	Number of bases (MB)	Mean reads length
S1	39.345	19167516	37.2	0.09	1.635	2.64	95.64	2875.13	150
S2	39.39	18779549	35.5	0.07	1.57	2.56	95.8	2816.93	150

4.4.2 Base quality score distribution

Base quality of each cycle for all samples is shown in figures 4. 3 (a & b). The x-axis represents sequencing cycle and y-axis represents percentage of total reads. The quality of left and right end of the paired-end read sequences of the sample is shown in these figures. It can be seen that more than 80% of the total reads have phred score greater than 30 (>Q30; error-probability >=0.001). The phred score distribution of the sample is provided in table 4.9.

Table 4.9 Raw read summary with Phred quality score distribution

Sample Name	Q0-Q10	Q10-Q20	Q20-Q30	>= Q30
S1	0.09	1.63	2.64	95.64
S2	0.07	1.57	2.56	95.80

4.4.3 Base composition distribution

The composition of nucleotides in the sequence read for each sample is shown in the figure 4.6(a&b). The x-axis represents sequencing cycle and y-axis represents nucleotide percentage. The base composition of left and right end of the paired-end read sequences are calculated. Since the target sequence is that of region sequence composition bias is observed in the sample. Overall base compositions of these samples are provided in Table 4.10.

Table 4.10 Base composition distribution of the samples

Sample Name	A	C	G	T
S1	31.54	18.57	18.67	31.23
S2	32.37	17.73	17.80	32.10

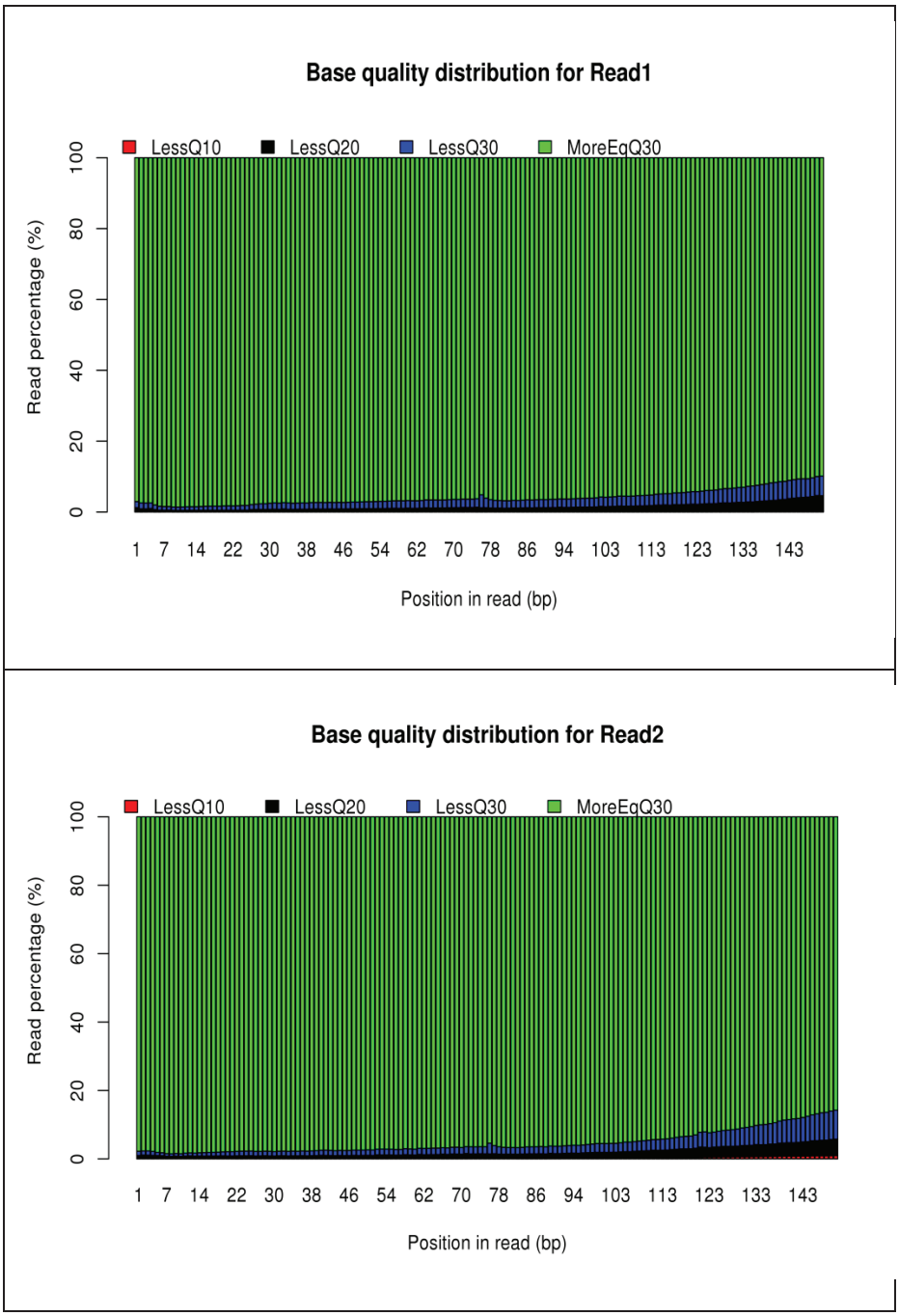


Figure 4.3 (a) Base quality distribution of sample S1 (replicates R1 & R2)

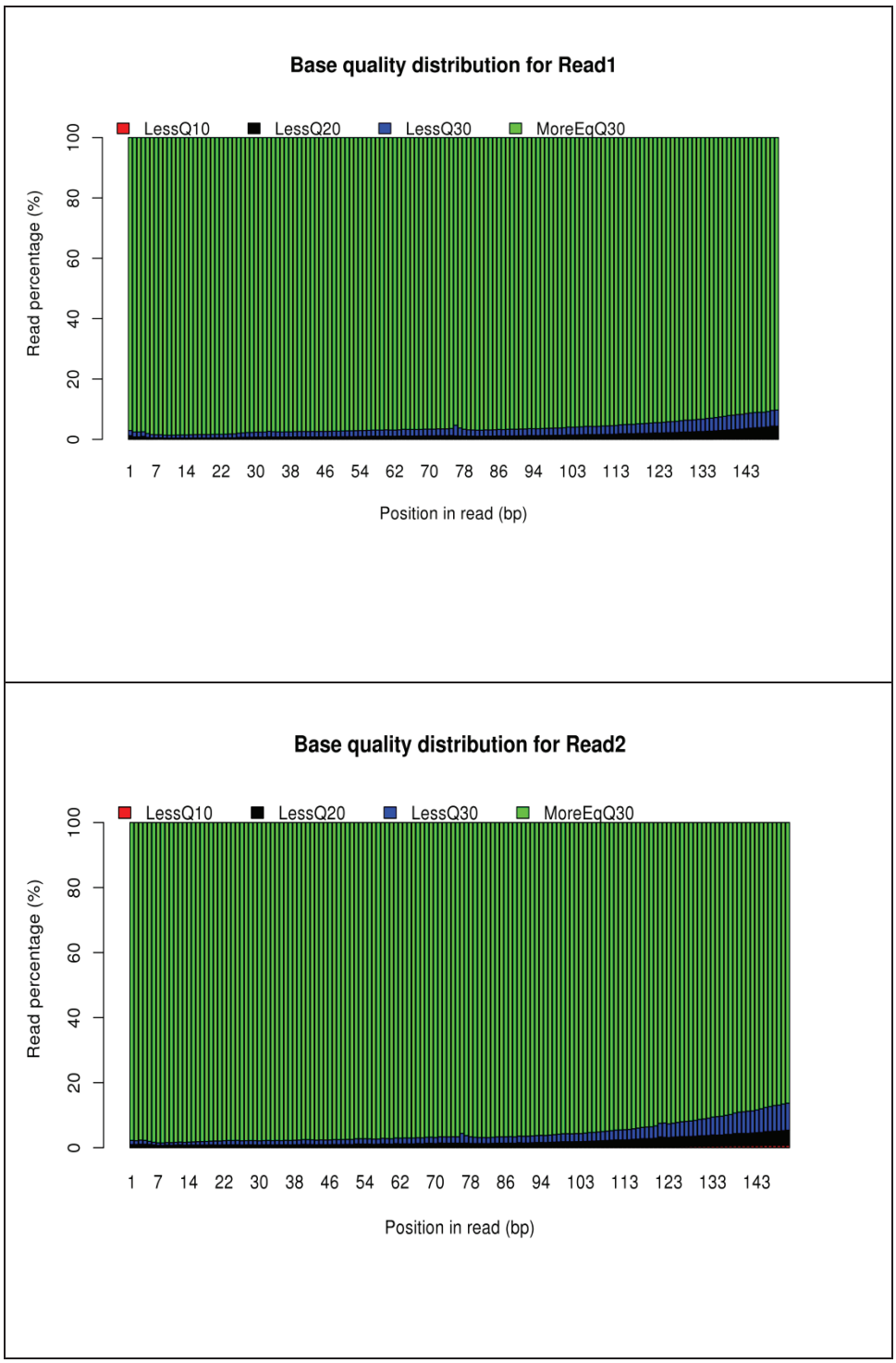


Figure 4.3(b) Base quality distribution of sample S2 (replicates R1&R2)

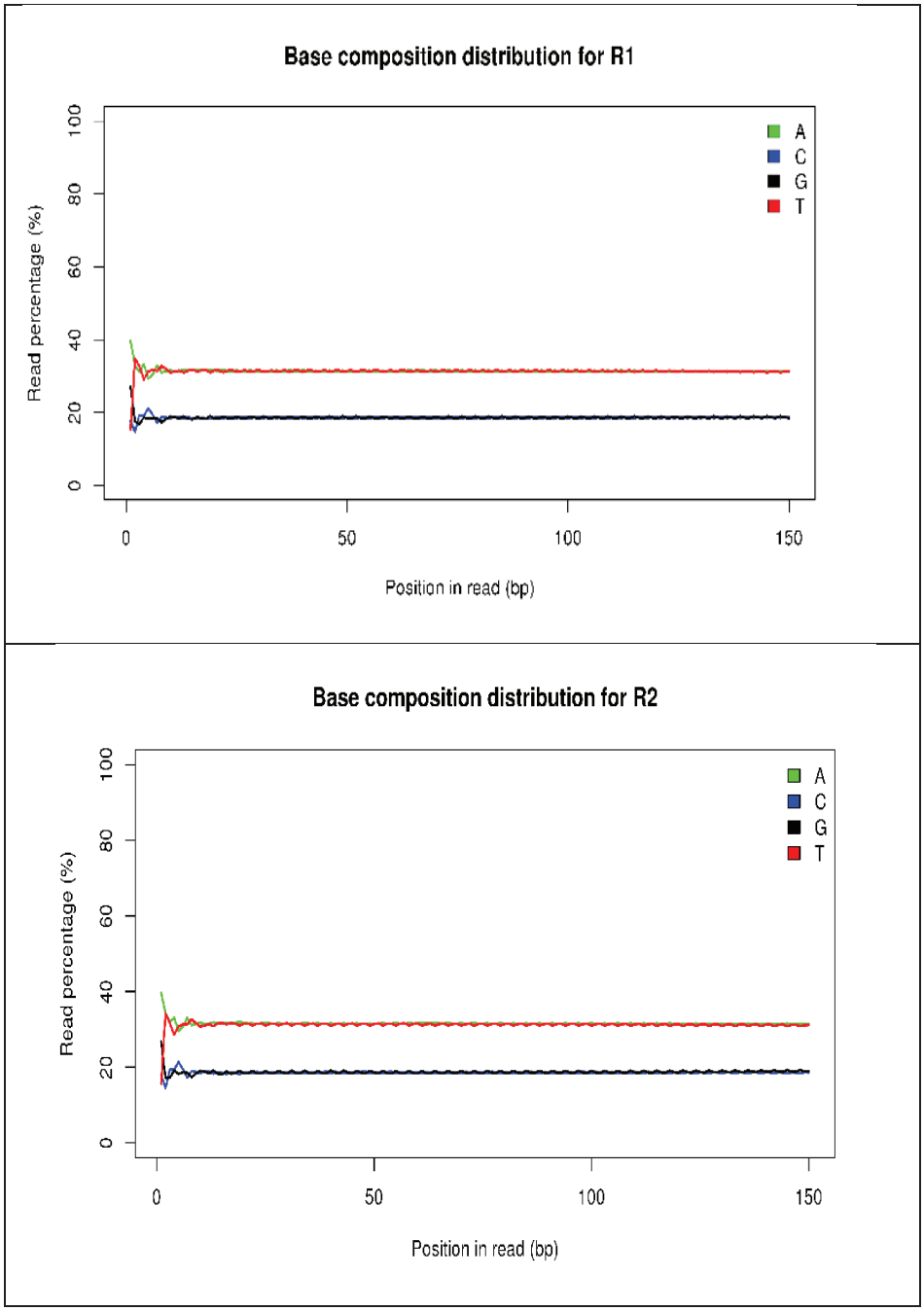


Figure 4.4(a) Base composition distribution of sample S1 (replicates R1 &R2)

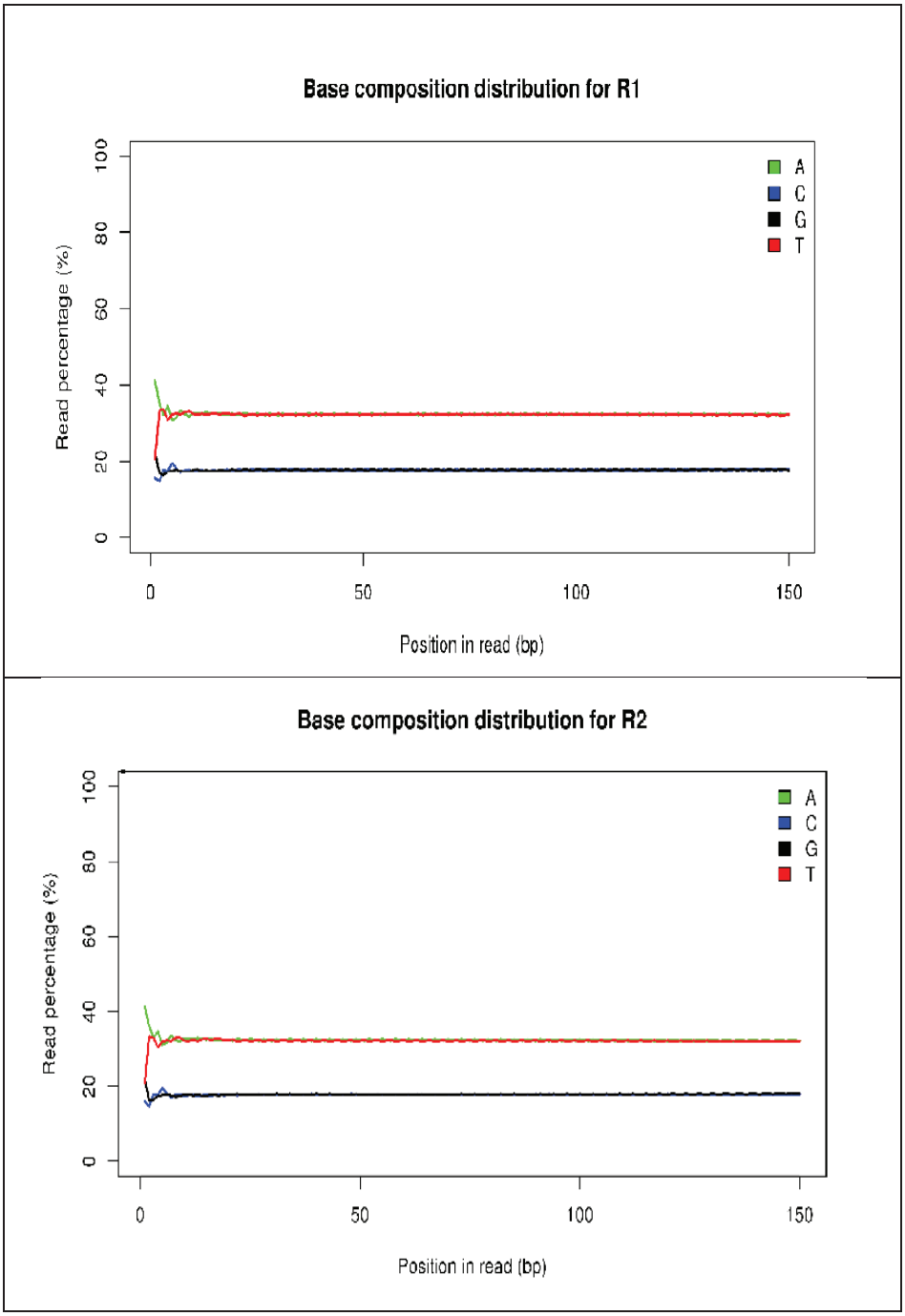


Figure 4.4 (b) Base composition distribution of sample S2 (replicates R1 &R2)

4.4.4 GC distribution

The average GC content distribution of the sequenced read of the samples is shown in the figure 4.5(a&b). The x-axis represents average GC content in the sequence and y-axis represents percentage of sequences. It was observed that the read had GC content in the range 30-60%.

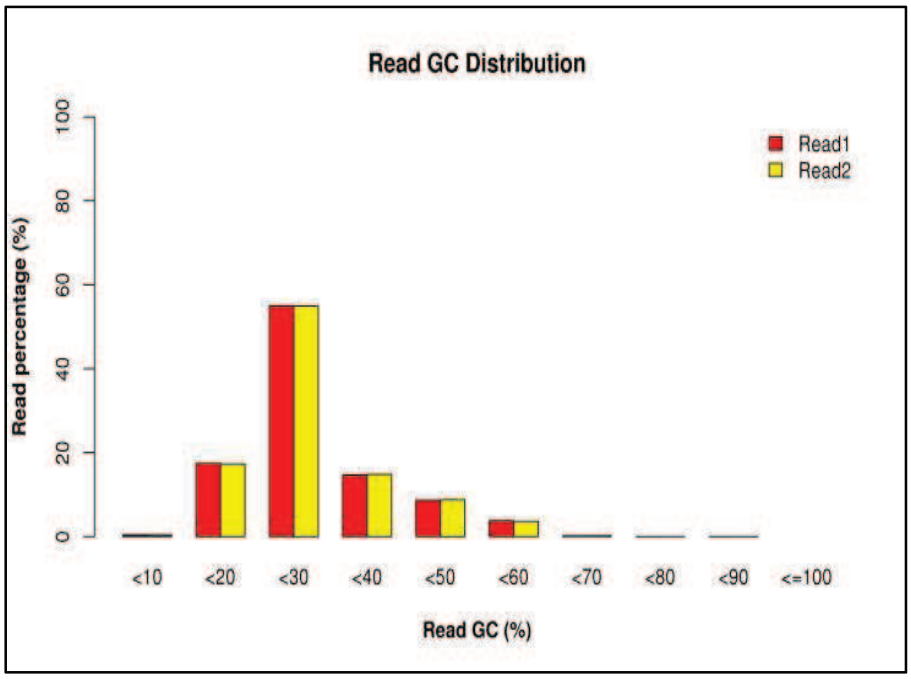


Figure 4.5(a) GC distribution of sample S1

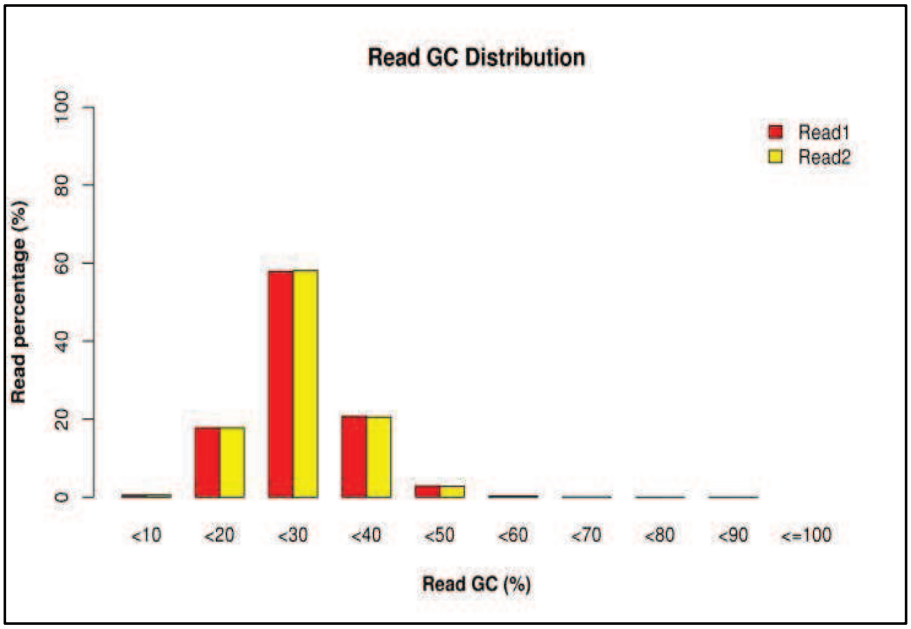


Figure 4.5(b) GC distribution of sample

4.4.5 *De novo* Metagenome Assembly

De novo metagenome assembly was carried out for the sample by assembling contigs from the reads using MetaSPAdes program. Further contigs were linked by the assembly algorithm to create scaffolds. Bad or mis-assemblies were removed from the result. Assembly was performed with default Kmer sizes 21, 33, 55 using de-bruijn graph method. In-house PERL and Python code were used to parse the fastq files for

the downstream analysis. Table 4.11 shows the summary of assembled contigs length in the samples. The total contigs for S1 is 40317, contigs at length<150is 2715, contigs at length150-500 is 25377 and contigs length >500 is 12225. For sample S2 total contigs is 103397, contigs at length, 150 is 3891, contigs at 150-500 length is 86759 and contigs length>500 is 12747.

Table 4.11 Trimmed and Consensus Read Summary

Sample Name	Total Contigs	Contigs Length <150	Contigs Length 150-500	Contigs Length >500
S1	40317	2715	25377	12225
S2	103397	3891	86759	12747

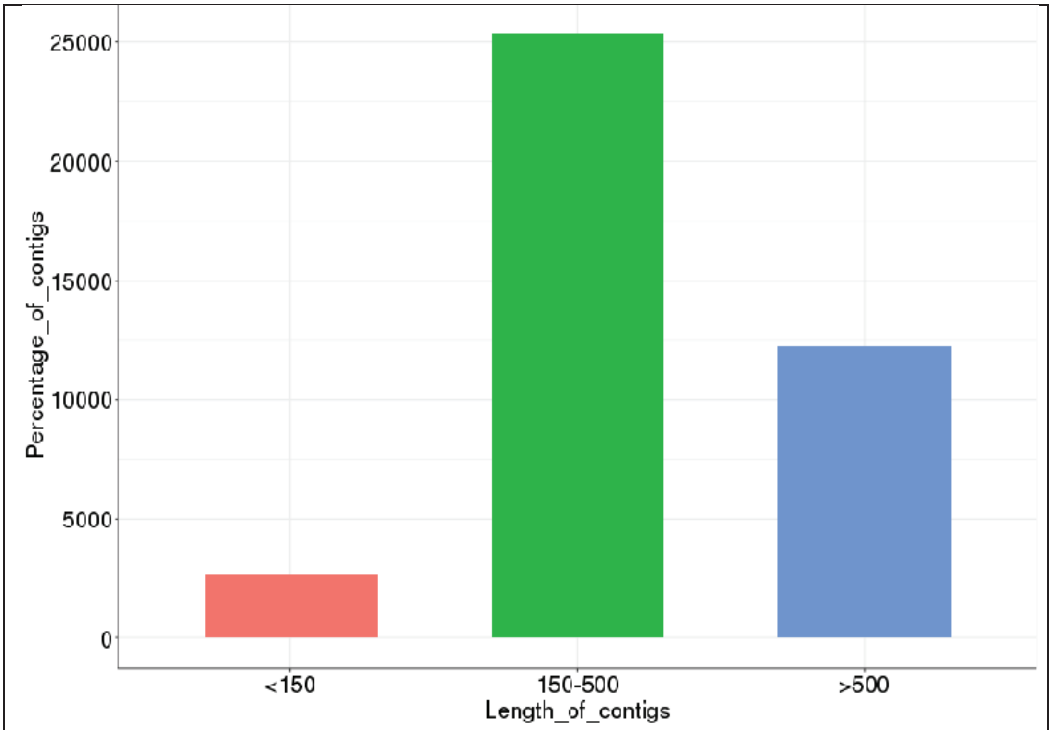


Figure 4.6(a) Contig Length distribution of sample S1

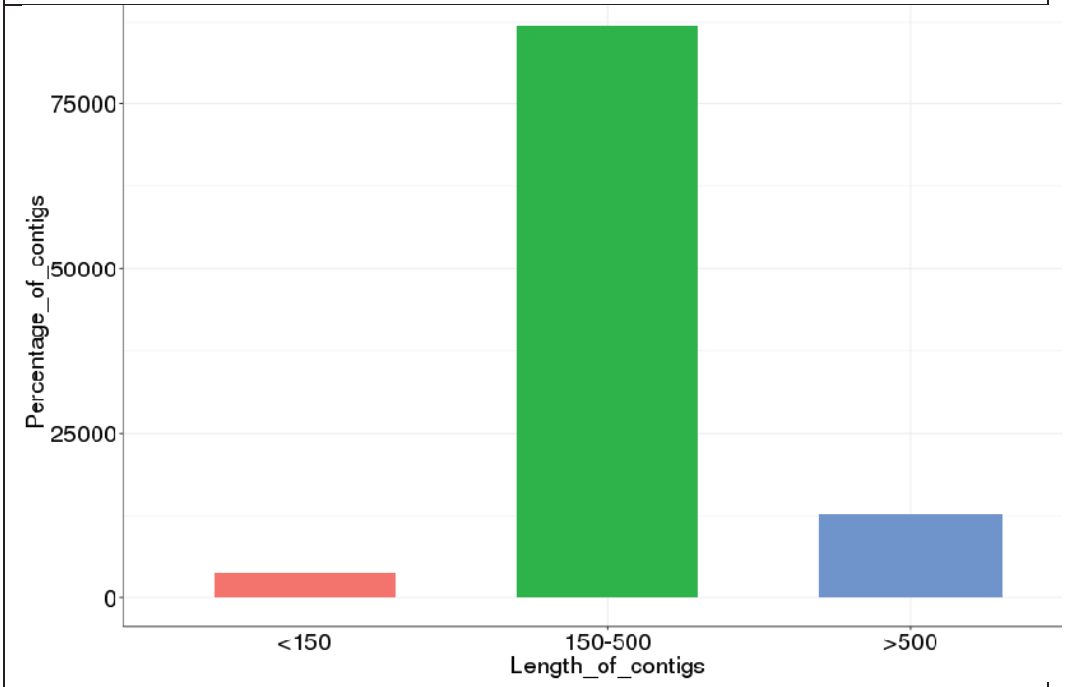


Figure 4.6(b) Contig Length distribution of sample S2

4.4.6 Analysis of dominant population

Rarefaction curve was generated by comparing the species abundance between the samples based on number of leaves in the taxonomy and number of sequence occurred. The curve is made for all taxa include Bacteria, Archaea, Eukaryote, Viruses, unclassified and other sequences. The curve length difference (shown in X-axis of plot) occurs between samples due to variation in the number of sequence between samples. Y-axis shows the number of leaves in taxon tree between samples.

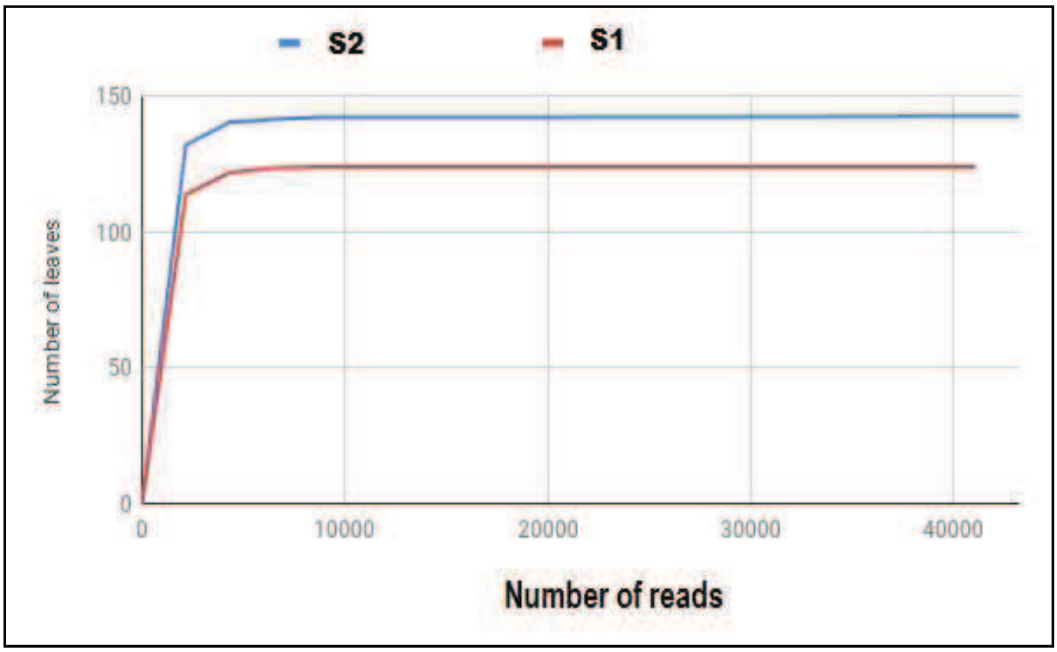


Figure 4.7 Rarefaction curves created in MEGAN

4.4.7 Taxonomic Analysis

The contigs obtained from the assembly were used as in put to Meta Gene Annotator (MGA) for the prediction of open reading frames (ORFs). The predicted ORF of S1 is 2940704 and ORF of S2 is 532641.The ORF obtained from the samples was queried to DIAMOND BLASTX program with optimum e-value of 1e-5. Taxonomic profiling for the entire metagenomic sample was performed using NCBI taxonomy data sets. The taxonomy tree was generated based on neighbor-joining method using MEGAN software. Taxonomic relative abundance in the sample based on contig abundance and phylum to species top 20 taxonomy abundance was determined.

S1 summarized seven Phyla and S2 consisted of eleven Phyla that corresponded to Prokaryotes and Eukaryotes. According to the annotation, it could be defined that

Firmicutes, Actinobacteria, Proteobacteria were dominant in S1, and Firmicutes Ascomycota Actinobacteria<phylum> & Proteobacteria were dominant in S2. Fermicutes were most abundant in both S1 and S2. It showed 81% relative abundance in S1, and 84% relative abundance in S2. The high abundance of Firmicutes, Actinobacteria, Proteobacteria, and Ascomycota indicates that these phyla played a key role in the fermentation of *napham*. Out of three Phyla, Firmicutes, Actinobacteria, Proteobacteria are Bacterial groups, and Ascomycota is a fungal group. In S1 the other phyla present were Chordata (0.45%), Planctomycetes (0.06%) and Cnidaria (0.05%). In S2 Chordata (0.52%), Arthropoda (0.33%), Planctomycetes (0.23%), Cnidaria (0.2%), Cyanobacteria (0.69%), Bacteroidetes (0.064%) and Chloroflexi (0.061%) were present.

Table 4.12 (a) Taxonomy read and relative abundance of Prokaryotes and Eukaryotes at Phylum level in sample S1

Sl. No.	Taxonomy (Phylum)	Sequence read S1	Relative abundance (%)
1.	Firmicutes	41770	81.59
2.	Actinobacteria <phylum>	7907	15.44
3.	Proteobacteria	1177	2.29
4.	Ascomycota	233	0.45
5.	Chordata	45	0.08
6.	Planctomycetes	35	0.06
7.	Cnidaria	27	0.05

Table 4.12(b) Taxonomy read and relative abundance of Prokaryotes and Eukaryotes at Phylum level in sample S2

Sl. No.	Taxonomy (Phylum)	Sequence read in S2	Relative abundance (%)
1.	Firmicutes	65775	84.4
2.	Ascomycota	5913	7.6
3.	Actinobacteria <phylum>	3736	4.8
4.	Proteobacteria	1372	1.7
5.	Chordata	408	0.5
6.	Arthropoda	263	0.3
7.	Planctomycetes	181	0.23
8.	Cnidaria	158	0.2
9.	Cyanobacteria	54	0.1
10.	Bacteroidetes	50	0.1
11.	Chloroflexi	48	0.1

At class level, the Class Bacilli was most abundant in both S1 and S2. In S1 it formed 81.45% of all the Classes followed by Actinobacteria (15.44%), Gammaproteobacteria (1.91%), Eurotiomycetes (0.55%), Alphaproteobacteria (0.28%), Clostridia (0.26%), Betaproteobacteria (0.07%), Planctomycetia (0.026%) and Actinopteri (0.05%). The total of Nine Classes was seen in S1. In S2 Bacilli forms 84.52% amongst all Classes followed by Eurotiomycetes (6.43%), Actinobacteria (4.72%), Saccharomycetes (1.15%), Arachnida (0.97%), Alphaproteobacteria (0.81%), Clostridia (0.511%), Gammaproteobacteria (0.51%), Actinopteri (0.49%), Betaproteobacteria (0.27%), Planctomycetia (0.21%), Deltaproteobacteria (0.17%), Thermoleophilia (0.08%). A total of eleven Classes were detected in S2.

In S1, total 15 Orders were detected, and the most abundant was Bacillales which formed 62 % of the total number of sequence reads. The relative abundance of other orders were Lactobacillales (18.9 %), Micrococcales (14.6%), Enterobacterales (1.22%), Corynebacteriales (0.68%), Eurotiales (0.44%), Clostridiales (0.3%), Streptomycetales (0.21 %) and Rhizobiales (0.18 %).

In S2, total 28 Orders were obtained and the most abundant order was Lactobacillales (48.6%) followed by Bacillales (34.45%), Eurotiales (6.4%), Micrococcales (3.03%), Caudovirales (1.19%), Saccharomycetales (1.14%), Corynebacteriales (0.93 %), Rhizobiales (0.58%), Clostridiales (0.48%) and some unclassified order (0.44 %).

At Family level, total 24 families were observed in S1 and 46 families in S2. Out of 24 families in S1 the top 5 most abundant families were Staphylococcaceae, Bacillaceae, Lactobacillaceae, Enterococcaceae and Aspergillaceae. Staphylococaceae forms 26.4% of the total reads. In S2 the top five families according to their abundance were Staphylococcaceae, Enterococcaceae, Lactobacillaceae, Bacillaceae, and Aspergillaceae.

At the genus level, total of 34 genera were obtained from S1 and 74 genera from S2. In S1 the relative abundance of genera is given in table 4.13 (a). The most abundant genera was *Staphylococcus* (36%) followed by *Oceanobacillus* (15%), *Virgibacillus* (12.5%), *Brevibacterium* (9.44%), *Pediococcus* (5.8%), *Enterococcus* (4.8 %), *Yaniella* (3.8%), *Bacillus* (3%), *Carnobacterium* (2.5%), *Lactobacillus* (2.1 %). The 34 genera obtained in fermented fish sample S1 with abundance above 0.01% are given below in table 4.13 (a).

In S2 also the most abundant genus was Staphylococcus (30%) followed by other top 9 genera including *Enterococcus* (18.6 %), *Lactobacillus* (11.6%), *Oceanobacillus* (6.5%), *Aspergillus* (6.4%), *Pediococcus* (3.9%), *Lactococcus* (3.9 %), *Tetragenococcus* (2.8 %), *Weissella* (2.4 %), *Vagococcus* (2.4%). The other genus whose abundance percentage was above 0.01% in S2 is given in table 4.13 (b). *Aspergillus, Hyphopichia and Candida<Debaryomycetaceae>*were three fungal Genus detected in two samples.

Table 4.13(a) Taxonomy read and relative abundance of bacterial and fungal Communities at Genus level in sample S1

Sl. No.	Taxonomy (Genus)	Sequence read in S1	Relative abundance (%)
1.	*Staphylococcus*	14898	32.9
2.	*Oceanobacillus*	6215	13.8
3.	*Virgibacillus*	5667	12.5
4.	*Brevibacterium*	4267	9.4
5.	*Pediococcus*	2636	5.8
6.	*Enterococcus*	2166	4.8
7.	*Yaniella*	1736	3.8
8.	*Bacillus*	1356	3
9.	*Carnobacterium*	1118	2.5
10.	*Lactobacillus*	935	2.06
11.	*Pisciglobus*	758	1.6
12.	*Enteractinococcus*	551	1.2
13.	*Lactococcus*	509	1.1
14.	*Macrococcus*	267	0.6
15.	*unknown*	236	0.3
16.	*Weissella*	235	0.5
17.	*Aspergillus*	221	0.5
18.	*Paenibacillus*	198	0.4
19.	*Streptococcus*	126	0.3
20.	*Corynebacterium*	125	0.3
21.	*Salinicoccus*	123	0.3
22.	*Arthrobacter*	120	0.3
23.	*Streptomyces*	106	0.2
24.	*Vagococcus*	83	0.2
25.	*Mycobacterium*	77	0.2
26.	*Nocardia*	67	0.1
27.	*Listeria*	51	0.1
28.	*Clostridium*	41	0.1
29.	*Microbacterium*	39	0.1
30.	*Kocuria*	37	0.1
31.	*Tetragenococcus*	34	0.1
32.	*Leuconostoc*	33	0.1

Table 4.13(b) Taxonomy read and relative abundance of bacterial and fungal Communities at Genus level in sample S2

Sl. No.	Taxonomy(Genus)	Sequence read in S2	Relative Abundance (%)
1.	*Staphylococcus*	19201	25.3
2.	*Enterococcus*	14155	18.6
3.	*Lactobacillus*	8843	11.6
4.	*Oceanobacillus*	4923	6.5
5.	*Aspergillus*	4827	6.4
6.	*Pediococcus*	3015	3.9
7.	*Lactococcus*	3000	3.9
8.	*Tetragenococcus*	2158	2.8
9.	*Weissella*	1827	2.4
10.	*Vagococcus*	1814	2.4
11.	*Brevibacterium*	1531	2.01
12.	*Carnobacterium*	778	1.02
13.	*Macrococcus*	759	0.99
14.	*Hyphopichia*	690	0.9
15.	*Pisciglobus*	673	0.9
16.	*Bacillus*	578	0.8
17.	*Ornithinibacillus*	447	0.6
18.	*Mycobacterium*	366	0.5
19.	*Lentibacillus*	310	0.4
20.	*Yaniella*	246	0.32
21.	*Sporosarcina*	240	0.31
22.	*unknown*	240	0.31
23.	*Corynebacterium*	237	0.31
24.	*Paucisalibacillus*	220	0.3
25.	*Virgibacillus*	192	0.25
26.	*Streptococcus*	172	0.22
27.	*Jeotgalicoccus*	170	0.22
28.	*Kocuria*	151	0.19
29.	*Clostridium*	141	0.18
30.	*Lysinibacillus*	138	0.18
31.	*Paenibacillus*	130	0.17
32.	*Massilia*	116	0.15
33.	*Streptomyces*	110	0.14
34.	*Gracilibacillus*	110	0.14
35.	*Enteractinococcus*	108	0.14
36.	*Salinicoccus*	107	0.14
37.	*Candida<Debaryomycetaceae>*	105	0.13
38.	*Listeria*	100	0.13
39.	*Enterobacter*	92	0.12

At species level, total 47 species were summarized in S1. The most abundant species with highest reads in S1 was S*taphylococcus xylosus* which formed 14.2% of the total population in S1. This was followed by *Oceanobacillus sojae* (12.5%), *Staphylococcus lentus* (7.9%), *Staphylococcus nepalensis* (7.8%), *Pediococcus pentosaceus* (7.7%), *Staphylococcus saprophyticus* (7.2 %), *Yaniella halotolerans* (5.3%), Enterococcus *faecalis* (4.7%), *Brevibacterium linens* (4.4%) and *Carnobacterium* sp.ZWU0011 (3.1%).

Total 153 species were summarized in S2. The abundant species detected in S2 were *Oceanobacillus oncorhynchi* (6.4%) followed by *Staphylococcus xylosus* (6.2 %), *Virgibacillus alimentarius* (5.7%), *Staphylococcus lentus* (5.2%), *Lactobacillus brevis* (4.7%), *Lactobacillus plantarum* (4.6%), *Lactococcus lactus* (4.3%), *Pediococcus pentosaceus* (4.2%), *Staphylococcus saprophyticus* (4.1%) and *Staphylococcus nepalensis* (3.9 %). The microbial species detected in S1 and S2 is given in table 4.14 (a) and 4.14 (b).

Table 4.14 (a) Microbial species detected in S1

Genus	Species
Brevibacterium	*Brevibacterium epidermidis, B. iodinum, B. linens*, *B. sandarakinum*, *B. siliguriense* and *Brevibacterium sp.VCM10*
Enterococcus	*Enterococcus faecalis* and *Enterococcus faecium*
Lactococcus	*Lactococcus garvieae* and *Lactococcus lactis*
Lactobacillus	*Lactobacillus brevis* and *Lactobacillus plantarum*
Oceanobacillus	*Oceanobacillus jeddahense*, *O. oncorhynchi*,*O. sojae* and *O. timonensis*
Pedococcus	*Pediococcus pentosaceus* and *P. acidilactici*
Staphylococcus	*Staphylococcus aureus, S. carnosus, S. cohnii, S. epidermidis, S. equorum, S. gallinarum, S. haemolyticus, S. hominis, S. lentus, S. nepalensis, S. saprophyticus,Staphylococcus sp.ZWU0021*, *S. succinus* and *S. sciuri*

1 species each	*Vagococcus teuberi,*
	Weissella paramesenteroides,
	Yaniella halotolerans,
	*Carnobacterium*sp.ZWU0011,
	Enteractinococcus helveticum,
	Listeria monocytogenes,
	Macrococcus caseolyticus,
	Aspergillus taichungensis
2 Phage viruses	*Staphylococcus phagepSco-10*, *Staphylococcus virus Sextaec* and *uncultured Caudoviralesphage*

Table 4.14(b) Microbial species detected in S2

Genus	Species
Aspergillus	*Aspergillus taichungensis, A.ruber* and *A. terreus*
Bacillus	*Bacillus* sp.VT-16-64, *B.thuringiensis, B. wiedmannii, B. cereus*
Brevibacterium	*Brevibacterium iodinum*, *B. antiquum*, *B.aurantiacum, B.casei*, *B. epidermidis*, *B.linens*, *B.sandarakinum*, *B. siliguriense*, *Brevibacterium* sp. 239c and *Bravibacterium* sp.VCM10
Candida	*Candida orthopsilosis & Candida parapsilosis*
Enterococcus	*Enterococcus avium, E. casseliflavus, E. devriesei, E. faecalis, E. faecium,E. gallinarum, E. hermanniensis, E.hirae, E.malodoratus, E. phoeniculicola, E. pseudoavium, E. thailandicus, E. gilvus, Enterococcus* sp.3H8_DIV0648, *Enterococcus* sp.6C8_DIV0013, *Enterococcus* sp.kppr-6
Jeotgalicoccus	*J. halophilus*, *J. psychrophilus, J.saudimassiliensis*, *J. saudimassiliensis*
Lactobacillus	*Lactobacillus.brevis, L. pentosus, L. plantarum* and *Lactobacillus* sp.SYF10-1a,
Lactobacillus	*Lactococcus garvieae* and *Lactococcus lactis*

Lentibacillus	*Lentibacillus. amyloliquefaciens, L. halodurans, L. jeotgali, L. persicus* and *L. sediminis*
Pedococcus	*Pediococcus pentosaceus* and *P. acidilactici*
Staphylococcus	*S. xylosus,S.aureus*, *S.carnosus, S. cohnii, S.edaphicus, S.epidermidis, S.equorum, gallinarum,S.haemolyticus, S.hominis,S.lentus,S.nepalensis,S.vitulinus,S.saprophyticus, S.sciuri, S.succinus, Staphylococcus* sp.LCT-H4, *Staphylococcus* sp.ZWU0021 and *Staphylococcus* sp.NAM3COL9,
Mycobacterium	*Mycobacterium abscessus* and *Mycobacteriumsp.M26*
Oceanobacillus	*O.damuensis*,*O.iheyensis*,*O.jeddahense*,*O.limi*,*O.manasiensis* ,*O.massiliensis*, *O.oncorhynchi*, *O.picturae*,*O.rekensis*, *O.senegalensis*,*O. sojae*, *O. timonensis*, *Oceanobacillus* sp.*Castelsardo*
Ornithinibacillus	*O. californiensis*, *O. contaminans*, *O.halophilus*, *O. scapharcae*
Tetragenococcus	*T.shalophilus, T.muriaticus* and *T.solitarius*
Vagococcus	*V. fluvialis*, *V.lutrae, V.penaei* and *V.teuberi*
Virgibacillus	*V.alimentarius*, *V. chiguensis*, *V.dakarensis*, *V. halodenitrificans*, *V. ndiopensis*,*V. necropolis*, *V. pantothenticus*, *V. phasianinus*, *V. proomii*, *V. siamensis*, *V. subterraneus*, V. *salinus*, *Virgibacillus* sp.IO3-P2-C2, *Virgibacillus* sp.IO3-P3-H5, *Virgibacillus* sp.SK37
Weissella	*W. hellenica*, *W. jogaejeotgali* and *W. paramesenteroides*
Phage virus	*Staphylococcus phagepSco-10,Staphylococcus phagevB_Sau_Clo6, Staphylococcus virus Sextaec, Staphylococcus virusP108* and *Staphylococcus phagevB_Sau_S24*

4.4.8 Analysis of Functional Diversity

To explore the metabolic potential of the microbiome of fermented fish S1 and S2 during fermentation, gene or protein functions of all the ORF from DIAMOND BLASTX output was parsed using in-house PERL script. The ORF obtained from the samples were queried to DIAMOND BLASTX program with optimum e-value of 1e-5. Further, functional annotation of all the Contigs are carried out by SEED Classification. MEGAN software was used to assign the function of each contigs. The protein functions of each contig having highest alignment score from DIAMOND BLASTX results were considered for functional assignment. The SEED classification of the samples S1and S2 were illustrated in Table 4.15.The level of reads categorized by SEED in two samples was compared. In S1 total 4083 and in S2 4530 sequence reads were categorized into seed functional category. As the *napham* fermentation progressed from 1st month to second month, matching levels of the metagenomic sequence reads to SEED functional categories increased due to the increase in bacterial abundance. The data shows that carbohydrate metabolism and fermentation were key categories for *napham* fermentation. Carbohydrate metabolism yielded an average of 19.4 % of all matches. The results showed that the *napham* microbiome had high metabolic versatility with respect to amino acid and protein metabolism. The protein metabolism yielded an average of 6.5% of all matches. Amino acids yielded an average of 12.1%. The top 10 functional diversity in S1 and S2 is given in figure 4.8.

1333 proteins were identified in S1. 5-FCL-like protein, 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100), Aldehyde dehydrogenase (EC 1.2.1.3), Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1) were the most active enzymes and involved in metabolic functions in S1.

1565 proteins have been identified in S2. 5-FCL-like protein, PTS system, cellobiose-specific IIC component (EC 2.7.1.69), Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1), 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) were the most active enzymes involed in metabolic pathways in S2. The Accession no. obtained from NCBI for WGS SRA Bioproject PRJNA689966 are SAMN17224118 YN3 (Tax ID: 496924) and SAMN17224117 YN2 (Tax ID: 496924).

Table 4.15 SEED based functional annotation for samples S1 and S2

Sl. No.	SEED Subsystem	Protein coding Sequences in S1 (%)	Protein coding Sequences in S2 (%)
1.	Carbohydrates	18.91	19.4
2.	Cofactors, Vitamins, Prosthetic Groups, Pigments	15.82	16.8
3.	Amino Acids and Derivatives	12.86	11.6
4.	Protein Metabolism	6.245	6
5.	Unclassified	4.55	4.5
6.	RNA Metabolism	4.26	3.9
7.	Miscellaneous	3.40	3.3
8.	Fatty Acids, Lipids, and Isoprenoids	3.18	3.1
9.	Nucleosides and Nucleotides	2.93	3.1
10.	Cell Wall and Capsule	2.91	2.9
11.	Respiration	2.89	2.7
12.	Stress Response	2.86	2.7
13.	DNA Metabolism	2.49	2.6
14.	Virulence	2.13	2.2
15.	Metabolite damage and its repair or mitigation	2.00	1.9
16.	Membrane Transport	1.66	1.8
17.	Regulation and Cell signaling	1.56	1.6
18.	Cell Division and Cell Cycle	1.07	1.1
19.	Metabolism of Aromatic Compounds	0.93	0.9
20.	Phosphorus Metabolism	0.93	0.9
21.	Virulence, Disease and Defense	0.85	0.9
22.	Sulfur Metabolism	0.73	0.8
23.	Predictions based on plant-prokaryote comparative analysis	0.61	0.6

Table 4.15 Continued…..

Sl. No.	SEED Subsystem	Protein coding Sequences in S1 (%)	Protein coding Sequences in S2 (%)
24.	Potassium metabolism	0.51	0.61
25.	Iron acquisition and metabolism	0.51	0.55
26.	Nitrogen Metabolism	0.49	0.52
27.	Thiamin	0.46	0.5
28.	Phages, Prophages, Transposable elements, Plasmids	0.41	0.5
29.	Motility and Chemotaxis	0.36	0.48
30.	Mitochondrial electron transport system in plants	0.24	0.22
31.	Dormancy and Sporulation	0.19	0.19
32.	Phages, Prophages, Transposable elements	0.12	0.19
33.	Nucleotide sugars	0.12	0.15
34.	Plant cell walls and outer surfaces	0.12	0.11
35.	Transcriptional regulation	0.09	0.08
36.	Central metabolism	0.09	0.08
37.	Secondary Metabolism	0.09	0.08
38.	Plant Glucosinolates	0.09	0.06
39.	Arabinose Sensor and transport module	0.07	0.02
40.	Autotrophy	0.02	0.02
41.	Photosynthesis	0.02	0
42.	Plastidial (cyanobacterial) electron transport system	0.02	0
43.	Not assigned	0.02	0

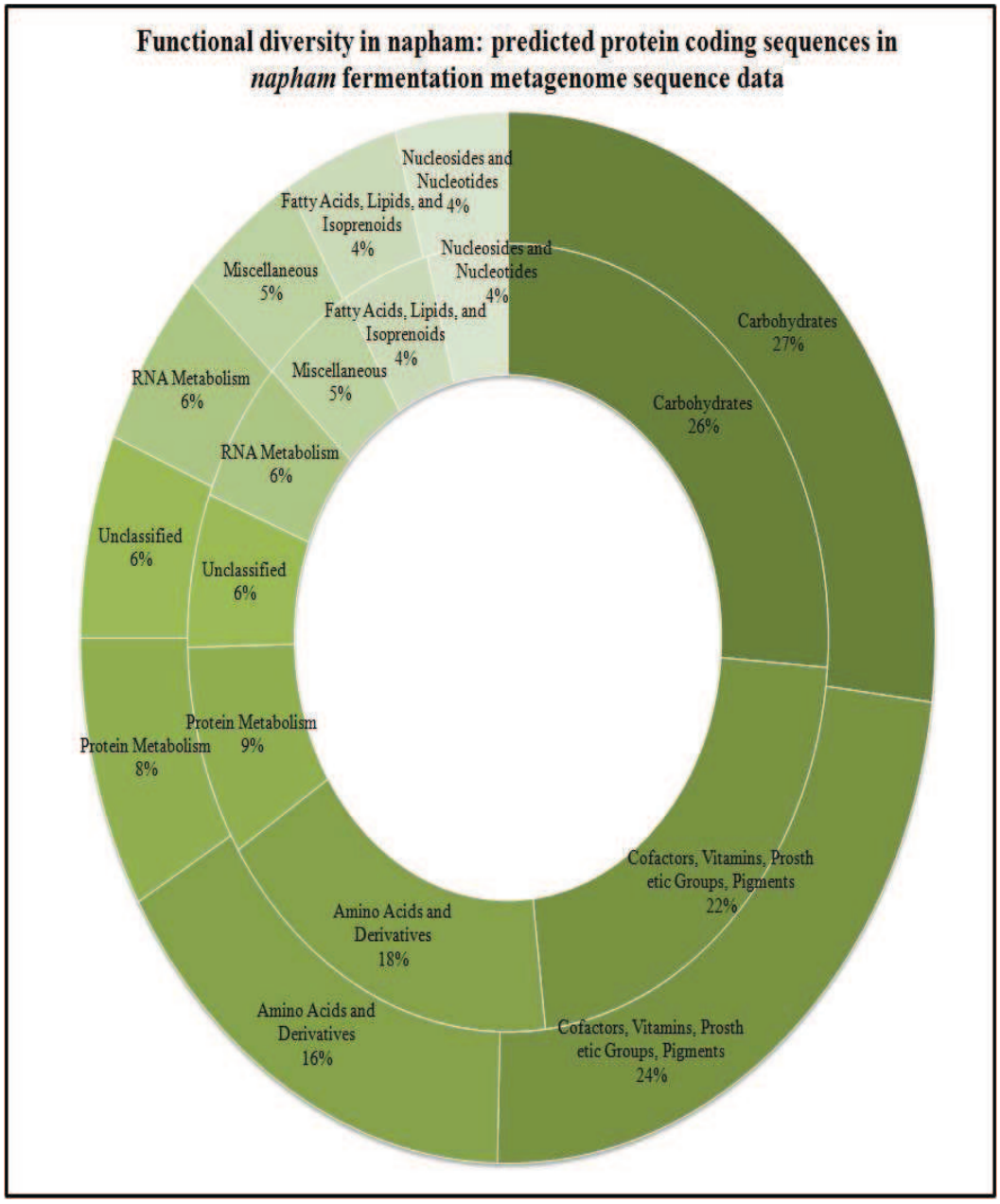


Figure 4.8 SEED based annotation showing top10 functional diversity and predicted protein coding sequences in *napham* fermentation metagenome of two samples S1(inner circle) and S2(outer circle)

4.5 16*S* metagenome analysis

16S rDNA amplicon sequencing is widely used for microbial community comparison among samples from various natural or endozoic environments such as soil, water, host intestine etc. The microbial diversity of two *napham* samples, which were more than three months old were analysed by 16 S rRNA metagenome. The samples were denoted as S3 and S4.

The concentration and quality of extracted DNA evaluated in 2 % agarose gel and the total amount of DNA extracted is given in Table 4.16 a. Out of two samples visible bands were seen in figure given in appendix (annexure IV). Two samples S3 and S4 were continued for library preparation, sequencing and further bioinformatics pipeline. A dataset consisting of 296878 filtered high-quality and classifiable 16S rRNA gene sequences, and an average of 148439 sequences were obtained for each individual sample. All sequences were clustered with representative sequences, and a 97 % sequence identity cut-off was used.

4.5.1 Sequencing and data processing

Amplicon was sequenced using Illumina paired-end chemistry and Illumina platform to generate 250bp paired-end raw reads (Raw PE), and then assembled and pretreated to obtain Clean Tags. The chimeric sequences in Clean Tags were detected and removed to obtain the Effective Tags finally. The data output has been shown in table 4.16 and data pre-processing and QC stat is given in Table 4.16(a & b).

Table 4.16(a) Output of DNA from samples

Sample Details					
Sl. No.	Sample ID	Conc. µg/µl	Sample Volume (µl)	Yield (µg)	Purity ($A_{260/280}$)
1.	S3	50.2	50	2.51	1.7
2.	S4	81.8	50	4.09	1.75

Table 4.16 (b) Data pre-processing and QC stat

Sample	Sample S3	Sample S4
Raw PE (#)	2,05,515	1,89,272
Raw Tags (#)	1,89,098	1,76,011
Clean Tags (#)	1,55,840	1,44,321
Effective Tags (#)	1,53,527	1,43,351
Base (nt)	657,61,204	614,14,151
AvgLen (nt)	428	428
Q20	97.18 %	97.23 %
Q30	94.41 %	94.36 %
GC	52 %	52.55 %
Effective	74.7 %	75.74 %

Here Raw PE means Pair End reads; Raw Tags means tags merged from PE reads; Clean Tags means tags after filtering; Effective Tags means tags after filtering chimera; Base means base number of Effective tags; AvgLen means average length of Effective Tags; Q20 and Q30 mean the percentage of base quantity that greater than 20 and 30; GC (%) means GC content in Effective Tags; Effective (%) means the percentage of Effective tags in Raw PE.

4.5.2 OTU analysis and species annotation

In order to analyze the species diversity in each sample, all Effective Tags were grouped by 97% DNA sequence similarity into OTUs (Operational Taxonomic Units) and these OTUs are then annotated. The evolutionary tree of genus is shown in figure 4.11.

4.5.3 OTU clustering, species annotation and statistical analysis of annotation

Based on 97% DNA sequence identity between the reads, 171 OTU for bacterial population was determined. During the construction of OTUs, basic information from

different samples had been collected, such as Effective Tags data, low-frequency Tags data and annotation data of Tags. The statistical dataset is showed as in Figure 4.11.

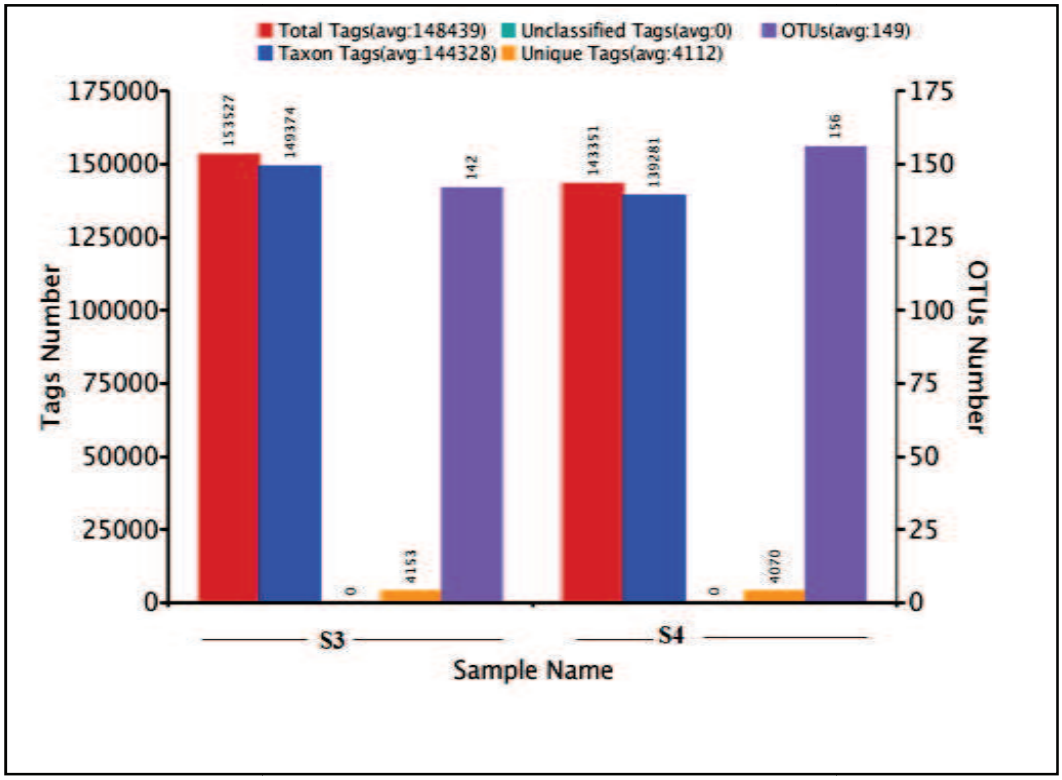
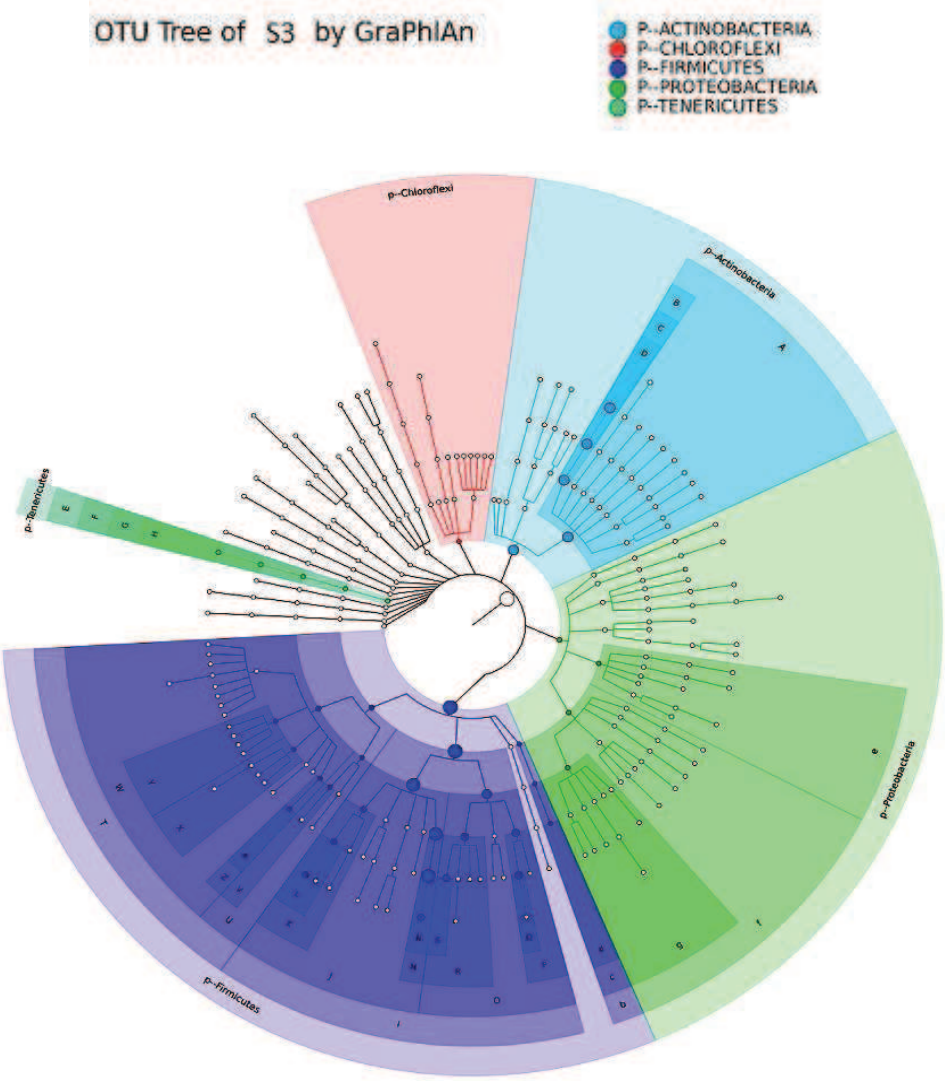


Figure 4.9 Statistical analysis of the tags and OTUs number of samples S3 and S4

The Y1-axis titled "Tags Number" means the number of tags; "Total tags"(Red bars) means the number of effective tags; "Taxon Tags" (Blue bars) means the number of annotated tags; "Unclassified Tags" (Green bars) means the number of unannotated tags; "Unique Tags" (Orange bars) means the number of tags with a frequency of 1 and only occurs in one sample. The Y2-axis titled "OTUs Numbers" means the number of OTUs which displayed as "OTUs" (Purple bars) in the above picture to identify the numbers of OTUs in different samples.

4.5.4 GraPhlAn display

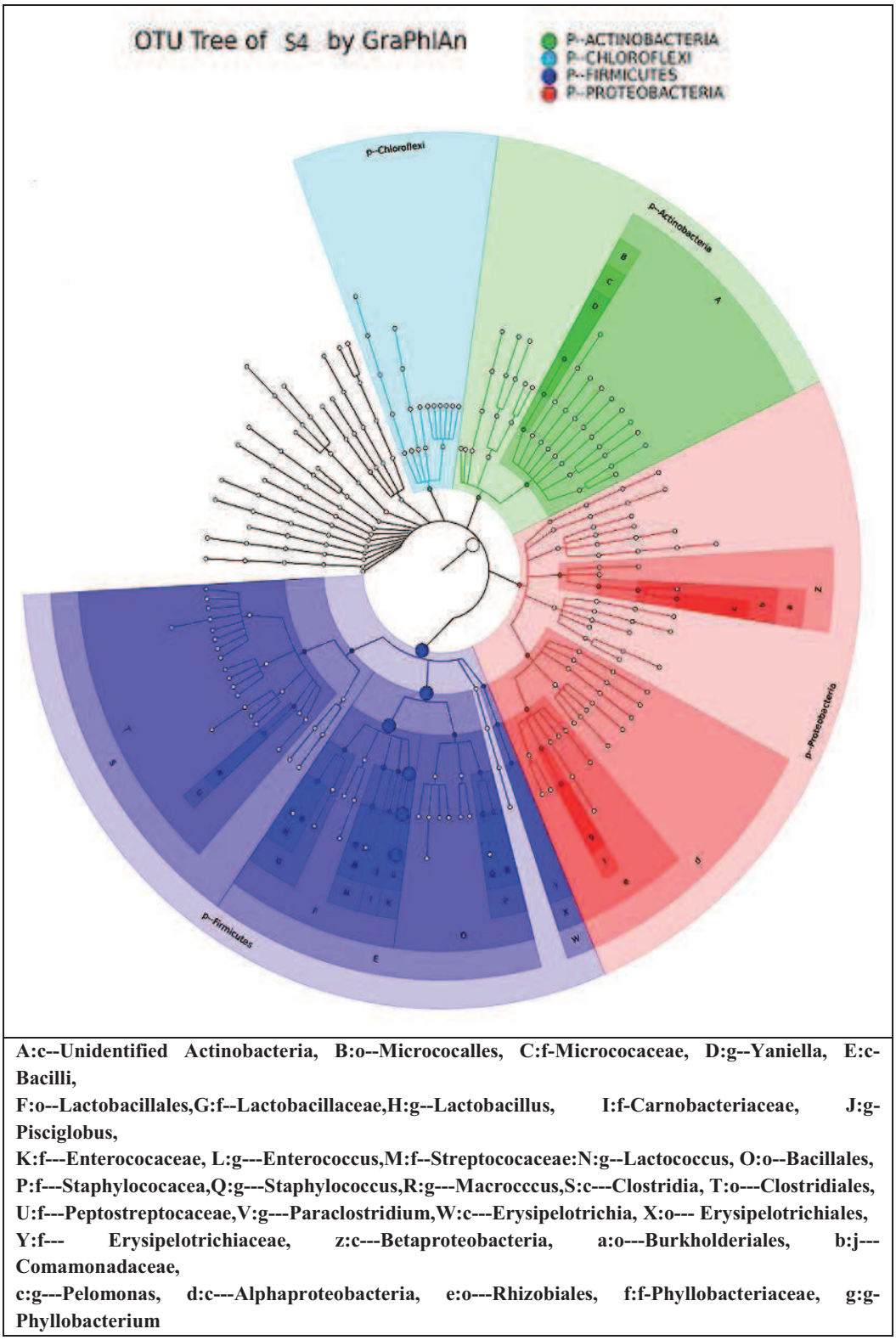
Tree graph of species annotation for each sample were constructed by GraPhlAn . The OTU trees of S3 and S4 are shown in Figure 4.10(a) and Figure 4.10(b).



A:c--unidentified Actinobacteria, B:o—Micrococcales, C:f—Micrococcaceae, D:g—Yaniella, E:c—Mollicutes, F:o—Haloplasmatales, G:f—Haloplasmataceae, H:g—Haloplasma, I:c—Bacilli, J:o—Lactobacillales, K:f—Lactobacillaceae, L:g—Lactobacillus, M:f—Enterococcaceae, N:g—Enterococcus, O:o—Bacillales, P:f—Staphylococcaceae, Q:g—Staphylococcus, R:f—Bacillaceae, S:g—Lentibacillus, T:c—Clostridia, U:o—Thermoanaerobacterales, V:f-Thermoanaerobacteraceae, W:o—Clostridiales, X:f--Clostridiaceae 1, Y:f—Lachnospiraceae, Z:f—Heliobacteriaceae, a:g—Hydrogenispora, b:c—Erysipelotrichia, c:o—Erysipelotrichales, d:f—Erysipelotrichaceae, e:c—Deltaproteobacteria, f:c—Alphaproteobacteria, g:o--Rhizobiales

Figure 4.10(a) S3 OTU annotation tree construct by GraPhlAn

Note: Different taxonomic ranks range inside out. The size of circles stands for abundance of species. Different colors stand for different phylum. Solid circles stand for the top 40 species in high abundanc



A:c--Unidentified Actinobacteria, B:o--Micrococalles, C:f-Micrococaceae, D:g--Yaniella, E:c-Bacilli,
F:o--Lactobacillales,G:f--Lactobacillaceae,H:g--Lactobacillus, I:f-Carnobacteriaceae, J:g-Pisciglobus,
K:f---Enterococaceae, L:g---Enterococcus,M:f--Streptococaceae:N:g--Lactococcus, O:o--Bacillales,
P:f---Staphylococacea,Q:g---Staphylococcus,R:g---Macrocccus,S:c---Clostridia, T:o---Clostridiales,
U:f---Peptostreptocaceae,V:g---Paraclostridium,W:c---Erysipelotrichia, X:o--- Erysipelotrichiales,
Y:f--- Erysipelotrichiaceae, z:c---Betaproteobacteria, a:o---Burkholderiales, b:j---Comamonadaceae,
c:g---Pelomonas, d:c---Alphaproteobacteria, e:o---Rhizobiales, f:f-Phyllobacteriaceae, g:g-Phyllobacterium

Figure 4.10(b) S4 OTU annotation tree construct by GraPhlAn

Note: Different taxonomic ranks range inside out. The size of circles stands for abundance of species. Different colors stand for different phylum. Solid circles stand for the top 40 species in high abundance.

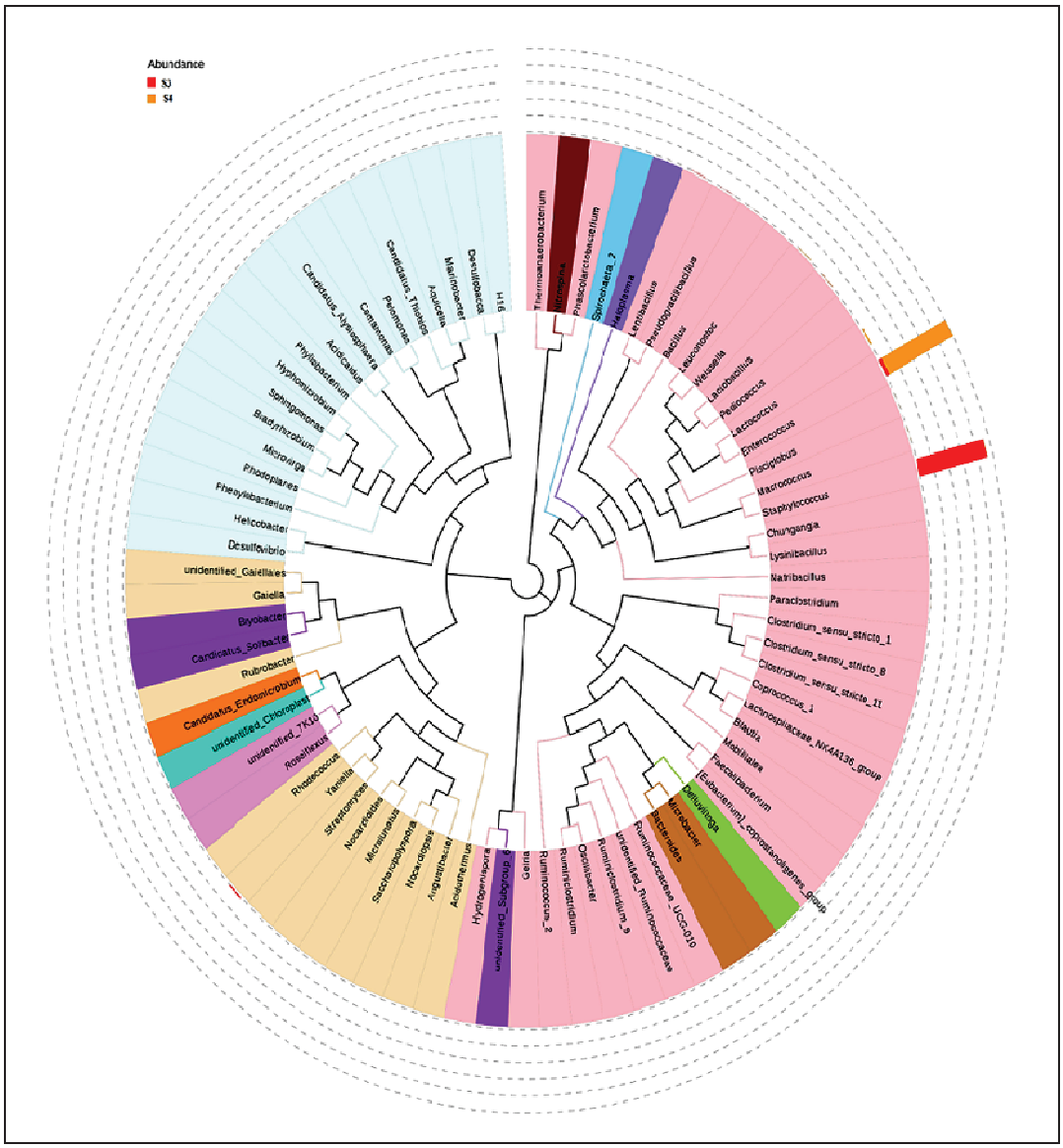


Figure 4.11 The evolutionary tree in genus

Note: Different colours of the branches represent different phyla. Relative abundance of each genus in each sample was displayed outside the circle and different colours represent different groups.

4.5.5 Phylogenetic Tree

Specific species (showing the top 10 genus in high relative abundance by default) were selected to make the taxonomy tree by independent R&D software. Phylogenetic trees in single samples S3 and S4 are shown in figure 4.12 (a & b).

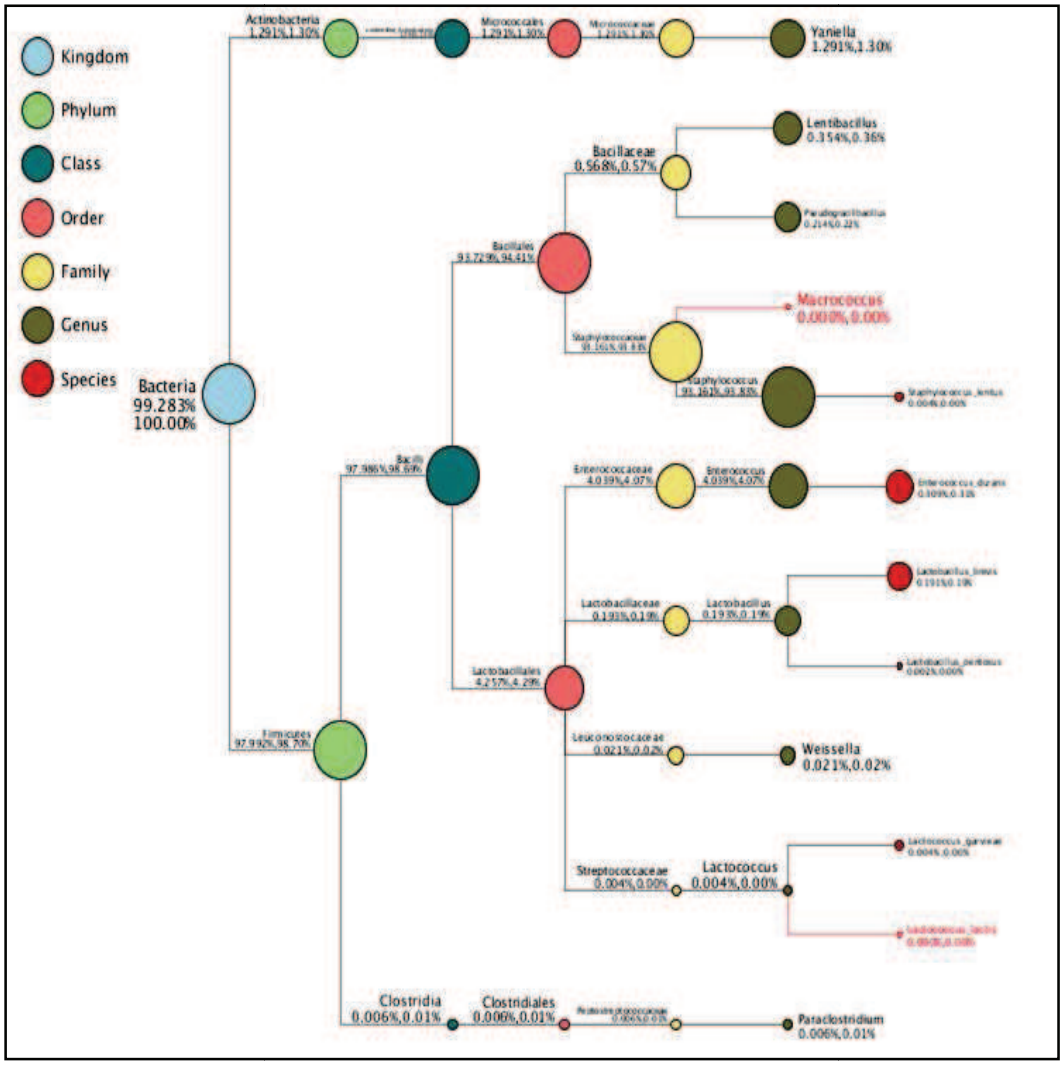


Figure 4.12(a) Phylogenetic tree –S3

Note: Different colours represent different taxonomic ranks. The sizes of circles stand for the relative abundance of species. The first number below the taxonomic name represents the percentage in the whole taxon, while the second number represents the percentage in the selected taxon.

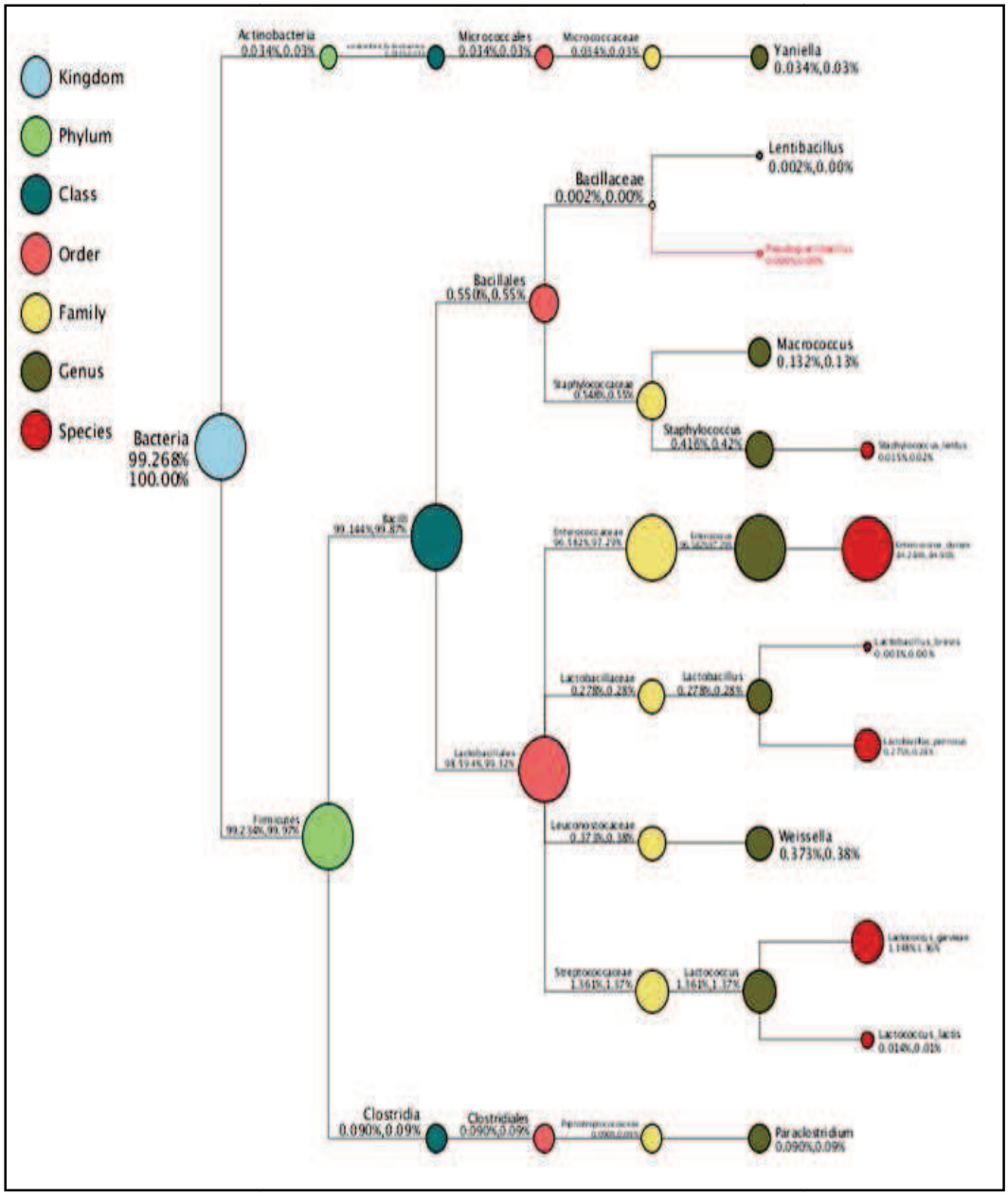


Figure 4.12(b) Phylogenetic tree – S4

Note: Sectors with different colors represent different groups. The size of the sector represents the relative abundance. The first number below the taxonomic name represents the percentage in the whole taxon, while the second number represents the percentage in the selected taxon.

4.5.6 Species distribution and Species relative abundance layout

The top 10 species in the different taxonomic ranks were selected to form the distribution histogram of relative abundance. At phylum level, 15 taxa were summarized from the sample S3 and 16 taxa for S4 given in Table 4.17. According to the annotation, it can be defined that Firmicutes, Actinobacteria and Proteobacteria were most abundant in both S3 and S4. Fermicutes alone forms 98.36 % of the total population in S3 and 99.45 % in S4.

Table 4.17 Taxonomy and relative abundance of Bacteria at Phylum level in samples S3 and S4

Taxonomy(Phylum)	Relative abundance in S3 (%)	Relative abundance level in S4 (%)
Firmicutes	98.36	99.5
Actinobacteria	1.40	0.1
Proteobacteria	0.14	0.3
Chloroflexi	0.02	0.06
Acidobacteria	0.03	0.04
Tenericutes	0.02	0.003
Bacteroidetes	0.001	0.01
Thermotogae	0.006	0.005
Cloacimonetes	nd	0.005
Nitrospirae	0.002	0.003
Elusimicrobia	0.001	0.003
Spirochaetes	nd	0.002
Cyanobacteria	0.001	0.002
Nitrospinae	nd	0.002
Fusobacteria	nd	0.002

Class Bacilli was the most abundant at Class level in both S3 and S4. In S3 it formed 98.22% of the total population followed by unidentified Actinobacteria (1.34%). In S4, Class Bacilli formed 99.26% of the total population followed by Clostridia (0.15%).

At order level, the Bacillales, Lactobacillales were the most abundant orders followed by Micrococcales. In S3, Bacillales formed 93.9 % of the total population, and in S4 Lactobacillales formed 98% of the total population.

At family level, Staphylococcaceae was most abundant in S3 and formed 93% of the total population followed by Enterococcaceae (4.03%) and Micrococaceae (1.3 %). In S4 Enterococcaceae (96.5 %) was the most abundant followed by Streptococcaceae (1.36 %) and Staphylococcaceae (0.54 %).

The taxonomical annotation could define total 80 genera in two samples. 64 genera were assigned in S3 and 80 in S4. Some unidentified genera were also detected in both the samples. The top 100 genera were selected and the evolutionary tree was drawn using the aligned sequences. Top 10 genera in S3 and S4 are given below in Table 4.18.

Table 4.18 Relative abundance at genus level in samples S3 and S4

Taxonomy (Genus)	Relative abundance in S3 (%)	Relative abundance in S4 (%)
Enterococcus	4.03	96.6
Staphylococcus	93.2	0.4
Lactococcus	0.004	1.4
Yaniella	1.3	0.03
Weissella	0.02	0.4
Lentibacillus	0.4	0.002
Lactobacillus	0.2	0.3
Pseudogracilibacillus	0.2	0
Macrococcus	nd	0.13
Paraclostridium	0.005	0.1
Others	0.7	0.7

At species level, only the top 15 abundant species have been analysed which are as follows: *Enterococcus durans, Lactobacillus brevis*, *Rhodococcus erythropolis, Clostridium leptum*, *Lactococcus garvieae*, *Staphylococcus lentus, Bradyrhizobium elkanii*, *Clostridium beijerinckii*, *Bacterium Ellin651*, *Lactobacillus pentosus*, *Pediococcus pentosaceus*, *Lactococcus lactis, Marinobacter alkaliphilus*, *Bacteroides graminisolvens* and *Leuconostoc citreum. Lactococcus lactis*, *Marinobacter alkaliphilus*, *Bacteroides graminisolvens*, *Leuconostoc citreum* were detected only in

S4. The analysis result of species annotation is displayed in KRONA Graphs (Figure4.13 a & b). The top 10 species of bacterial communities in the different taxonomic ranks were selected to form the distribution histogram of relative abundance for S3and S4 depicted in Figure 4.17 (a, b, d and e)

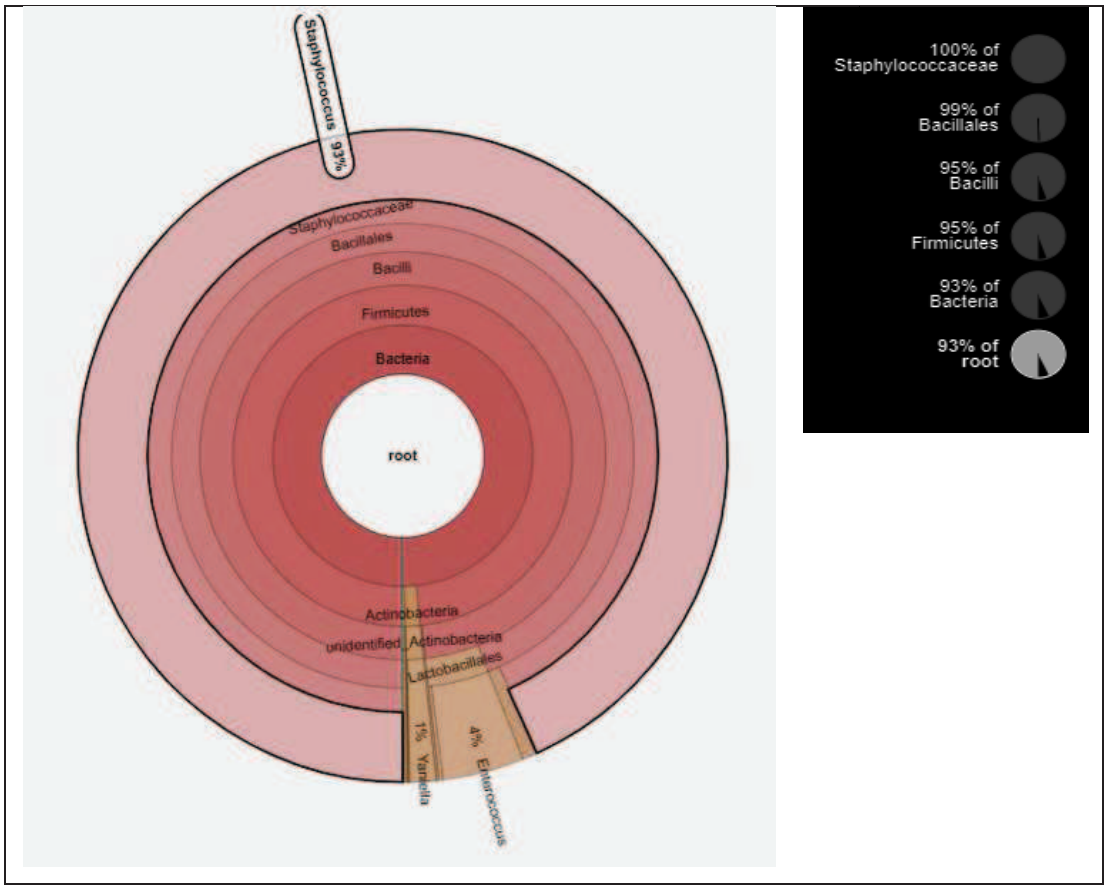


Figure 4.13(a) KRONA Graph of S3 showing dominant species *Staphylococcus* sp.

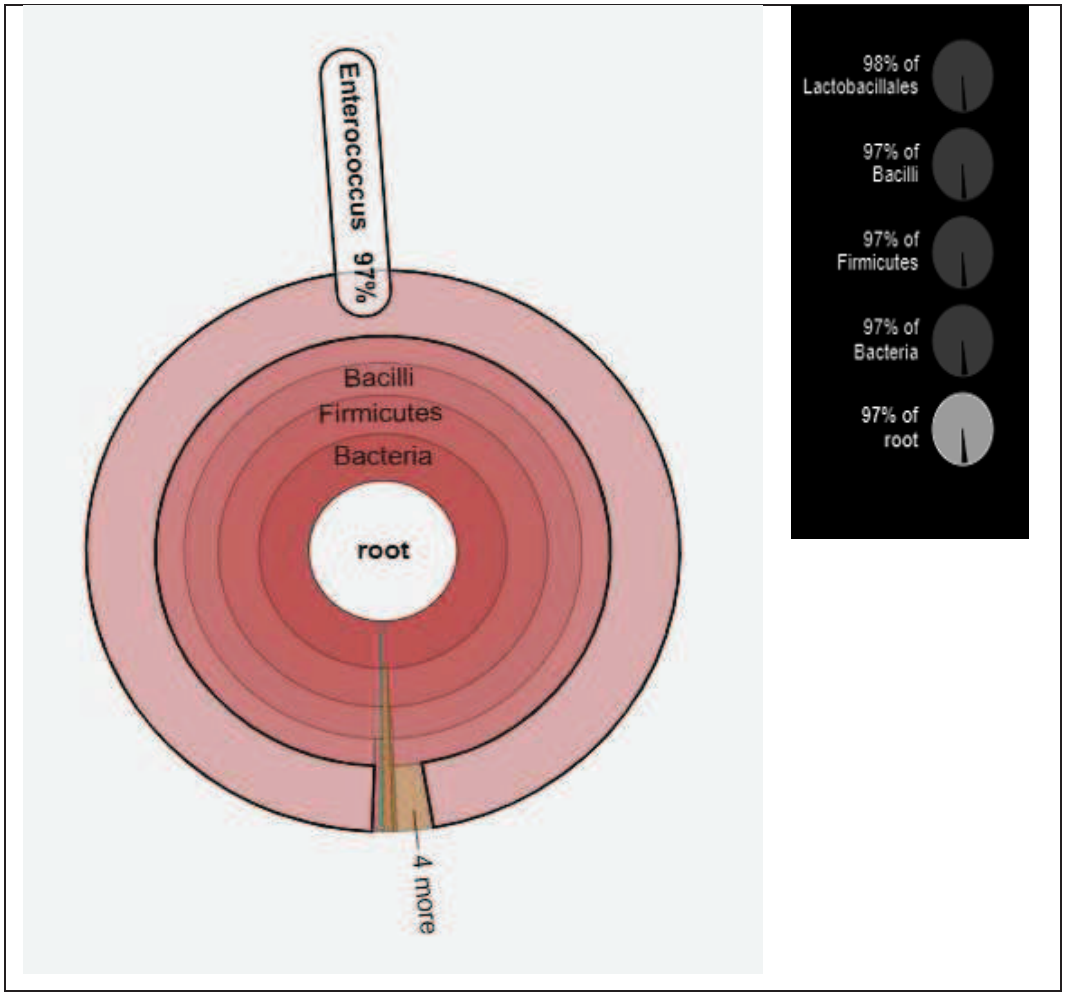
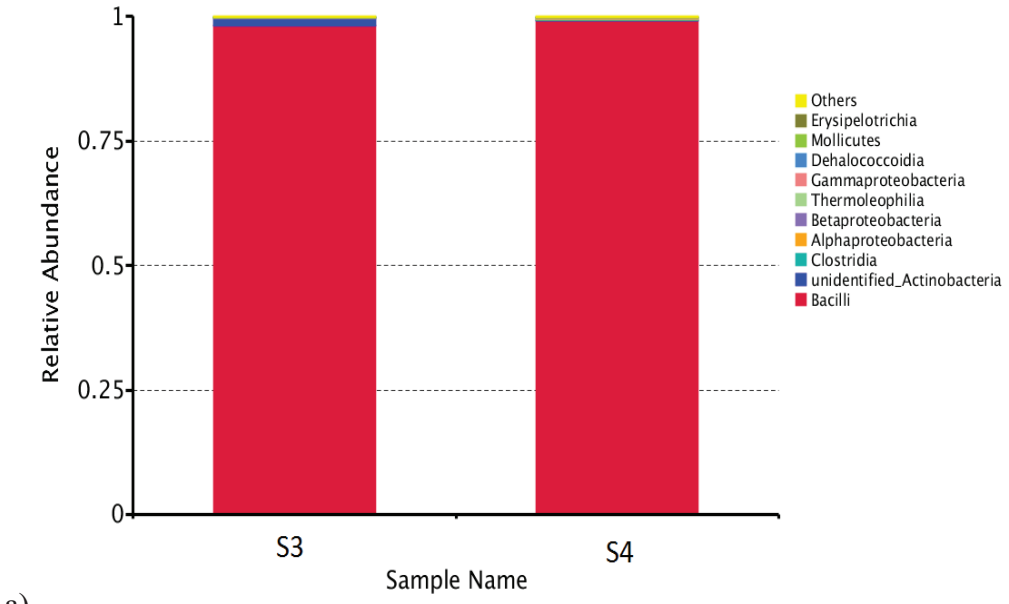
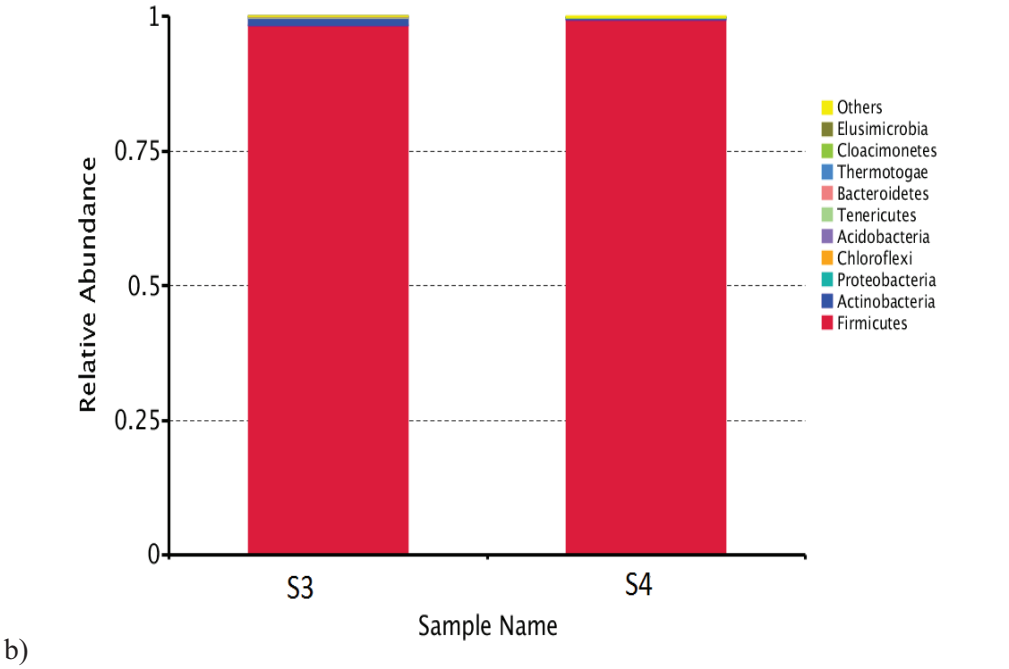


Figure: 4.13(b) KRONA Graph of S4 KRONA Graph of S4 showing dominant species *Enterococcus* sp.

KRONA visually displays the analysis result of species annotation. KRONA tool Circles from inside to outside stand for different taxonomic ranks, and the area of sector means respective proportion of different OTU annotation results.



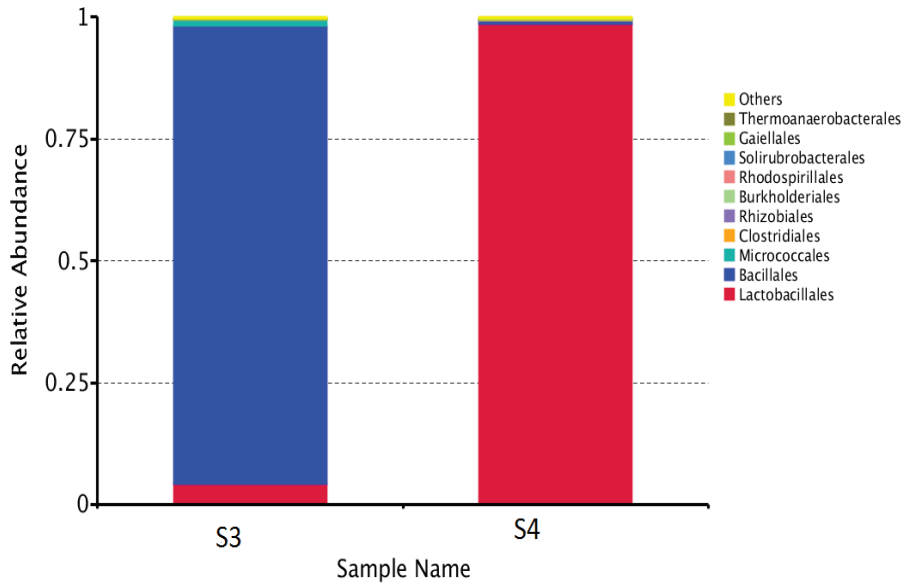
a)



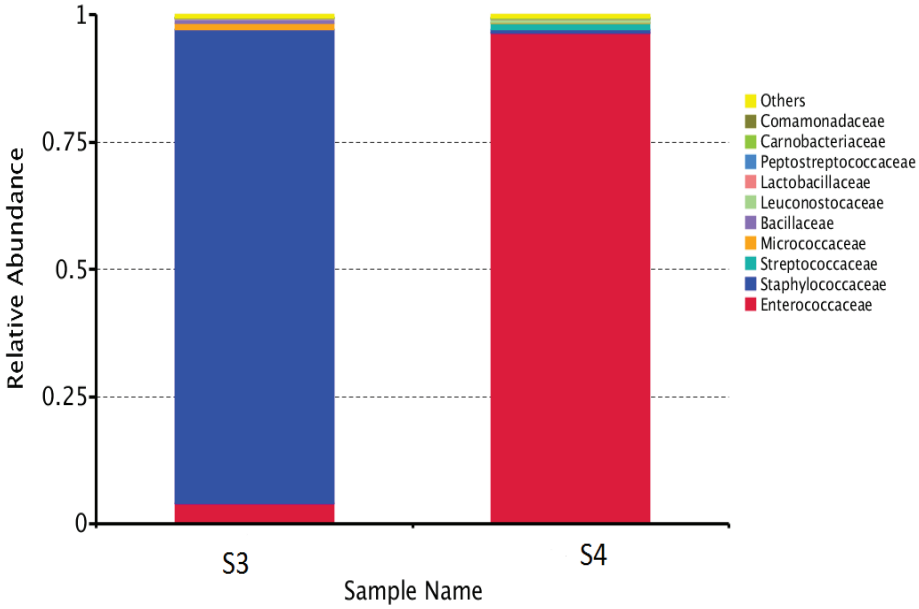
b)

Figure 4.14(a & b) Relative abundance of bacterial communities in two different stages of fermentation (S3and S4) of *napham* at (a) Phylum level (b) Class level

Note: Plotted by the "Relative Abundance" on the Y-axis and "Samples Name" on the X-axis



c)



d)

Figure 4.14(c & d) Relative abundance of bacterial communities in two different stages of fermentation (S3and S4) of *napham* at (c) Order level (d) Family level

4.5.7 Alpha Diversity Indices

In general, OTUs generated at 97% sequence identity were considered to be homologous in species. Statistical indices of alpha diversity when the clustering threshold is 97% were summarized as in Table 4.18 (Number of reads chosen for normalization : cutoff=139281). Alpha diversity was applied in analyzing complexity of species diversity for a sample through 6 indices, including Observed-species, Chao1, Shannon, Simpson, ACE, and Good-coverage (Table 4.19). All these indices in the samples were calculated with QIIME (Version 1.7.0) and displayed with R software (Version 2.15.3). The plot for observed species, Shannon, Simpson, Chao1 and whole tree is given in Figure 4.15 (a, b, c & d)

Table 4.19 Alpha Diversity Indices

Sl. No.	Sample Name	S3	S4
1	Observed Species	129	156
2	Shannon	0.583	0.881
3	Simpson	0.139	0.274
4	Chao1	134	168.214
5	ACE	136.705	163.51
6	Good's_Coverage	1	1
7	PD_Whole_Tree	12.649	16.263

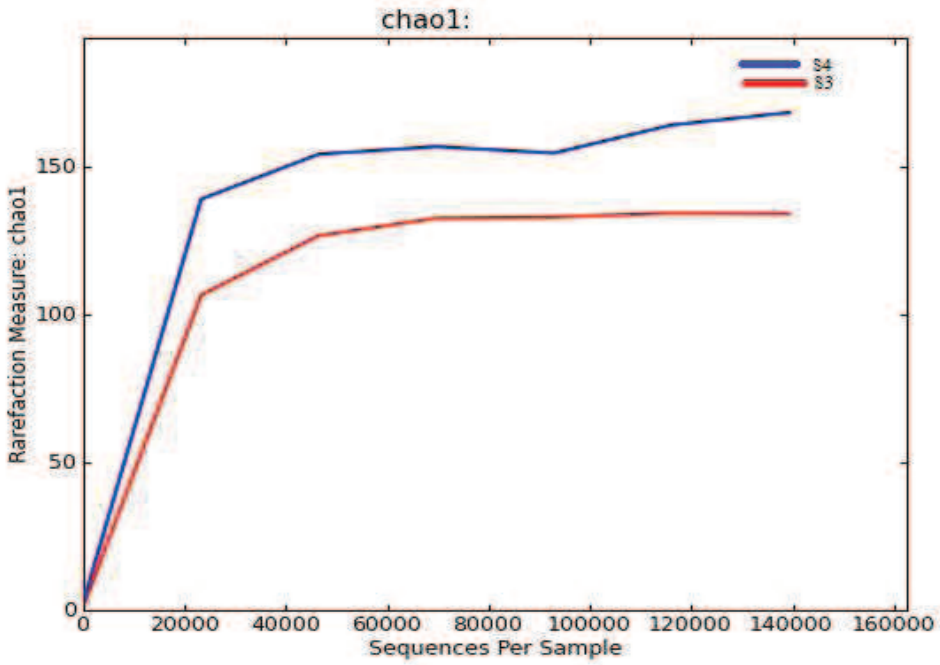


Figure 4.15(a) Alpha diversity: Chao1 index in S3 and S4

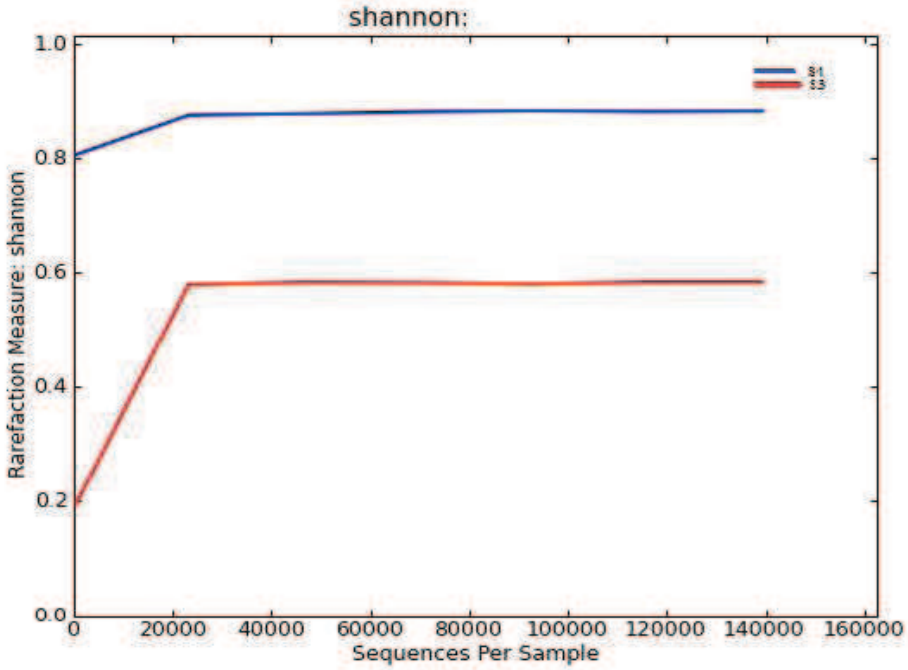


Figure 4.15(b) Alpha diversity: Shannon index in S3 and S4

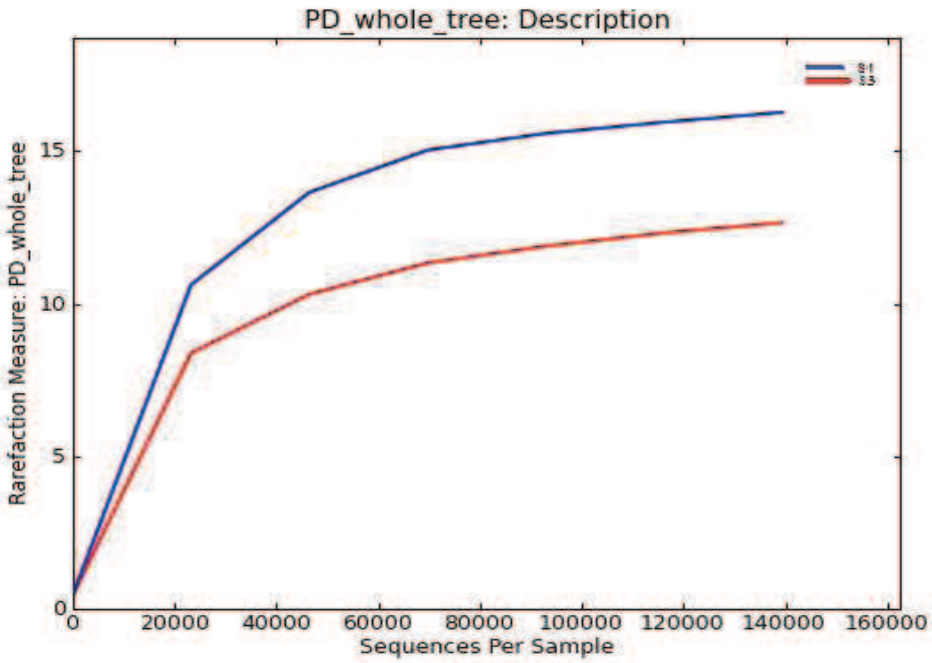


Figure 4.15(c)Alpha diversity: PD_whol_ tree in S3 and S4

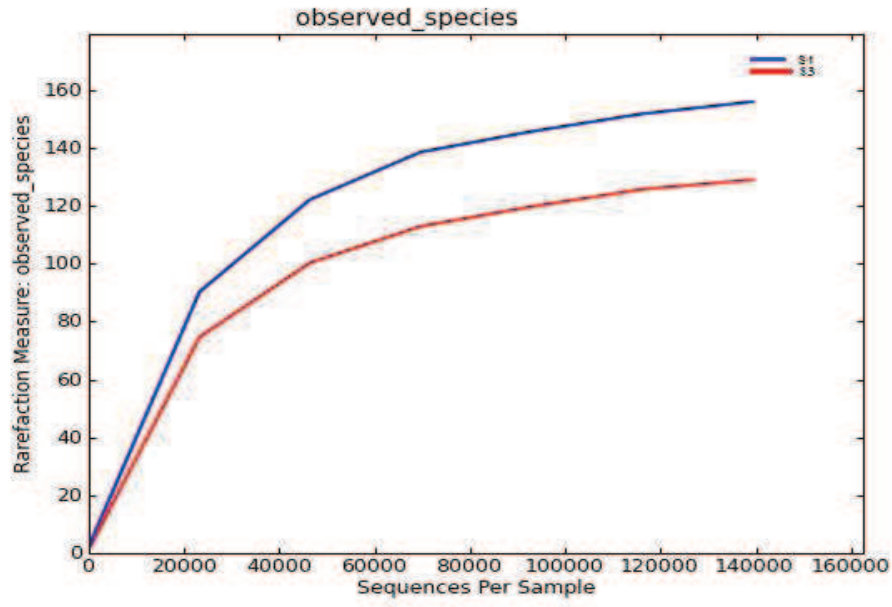


Figure 4.15(d) Observed species in S3 and S4

*The figures in 4.29 depicts the Alpha diversity analysis at clustering threshold of 97%

4.5.8 Species Diversity Curves

Rarefaction Curves (Figure 4.19) and Rank abundance curves (Figure 4.20) were used for indicating the biodiversity of the samples. Rarefaction Curve was created by selecting certain amount of sequencing data randomly from the samples, then counting the number of the species they represent. If the curve is steep means a lots of the species remain to be discovered. If the curve becomes flatter means a credible number of samples have been detected and only scarce number of species remains to be sampled. Rank abundance curve displayed relative species abundance of the samples S3 and S4. It showed the species richness and evenness. The diagrammatic representation of species diversity and rank abundance of samples are given in figure 4.16 and figure 4.17. The 16S rRNA metadata is submitted in NCBI as SRA bioproject with submission no. SUB9303737 is SAMN20088403:16SrRNA metagenome of *napham* (TaxID: 870726)

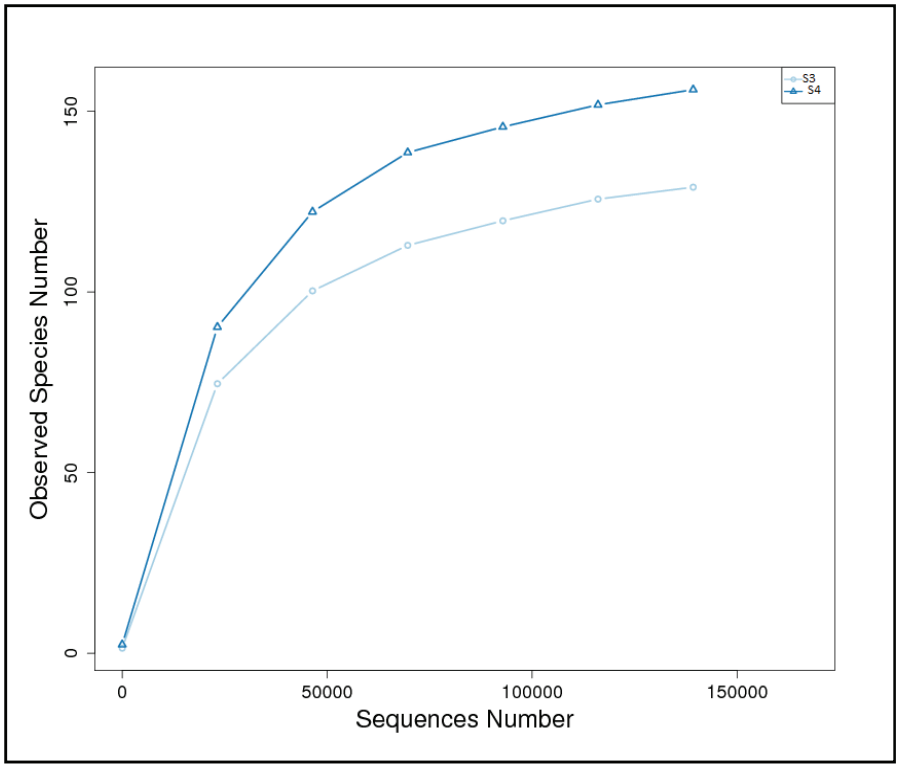


Figure 4.16 Rarefaction curve observed at 97 % similarity

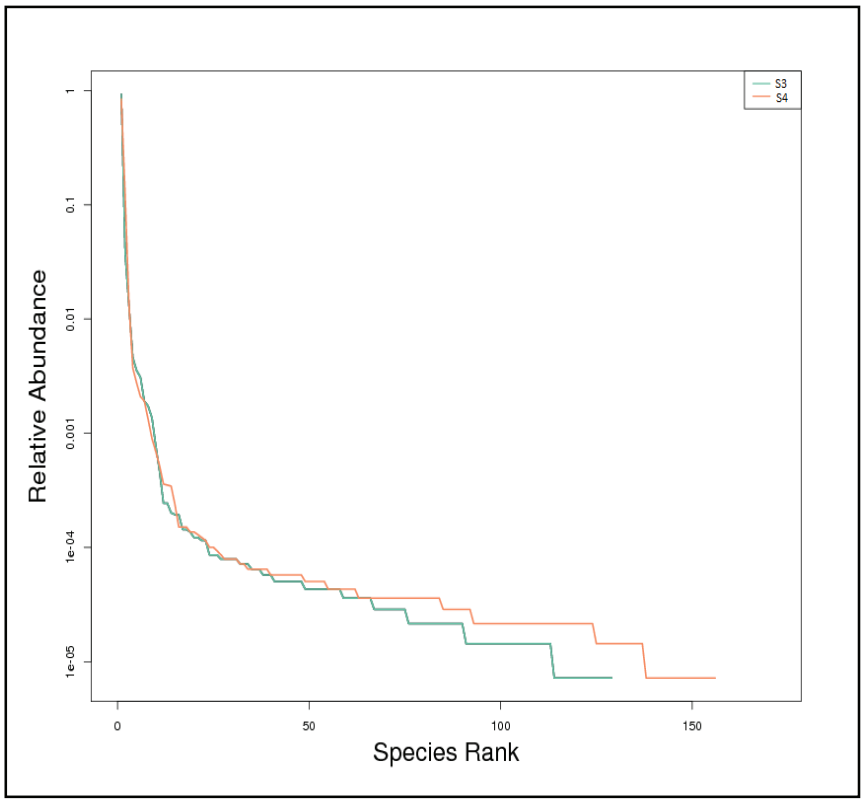


Figure 4.17 Rank Abundance curve observed at 97 % similarity