# Genomic Biome Analysis, Microbiome Focused

Bioinformatics Boot Camp

Ben Kellman, Jia Liu, Jeff Yuan, & Yuan Zhao

9/23/13

# Background

## microbiome

Web definitions

A microbiome is the totality of microbes, their genetic elements (genomes), and environmental interactions in a defined environment. A defined environment could, for example, be the gut of a human being or a soil sample. ...
http://en.wikipedia.org/wiki/Microbiome

- Biologically significant aspect of metabolic function
- Shown to influence gene expression of host
- High potential for variability

# Research Question

- Identical twins (monozygotic) share identical DNA whereas fraternal (dizygotic) do not
- Epigenetic, non-sequence modifications are known to vary from individual to individual, even twins
- So…

- Is there any correlation between the type of twin pair and microbiome composition?

Transcriptome Analysis of Human Supragingival Dental Plaque Microbiome from 19 Twin Pairs with and less without Dental Caries. Ong, et al. 2011

# Dataset

- Metagenomic data from oral microbiome samples (Illumina Genome Analyzer II)
- Twin pairs of monozygotic or dizygotic twins
- GSM780826/7 (monozygotic; accession SRX091838/9) aka 3000/3001
- GSM780830/1 (dizygotic; accession SRX091842/3) aka 3004/3005
- Short 0.01% random sampling for most analyses

http://www.ebi.ac.uk/ena/data/view/SRX091836-SRX091873

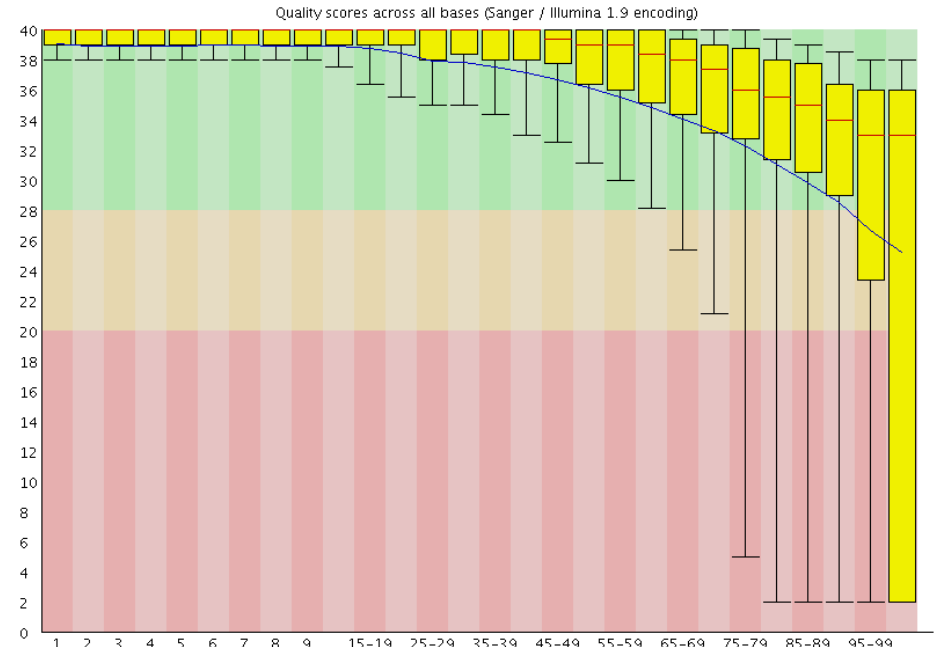
# Methods of Analysis

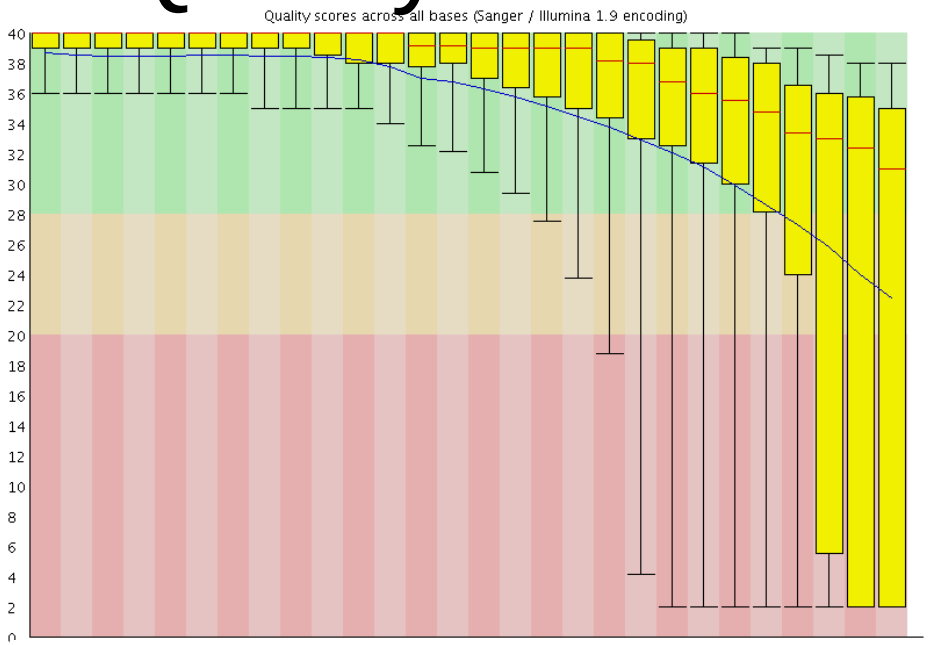
- FastQC
  - Determine composition of data
  - As well as quality of data

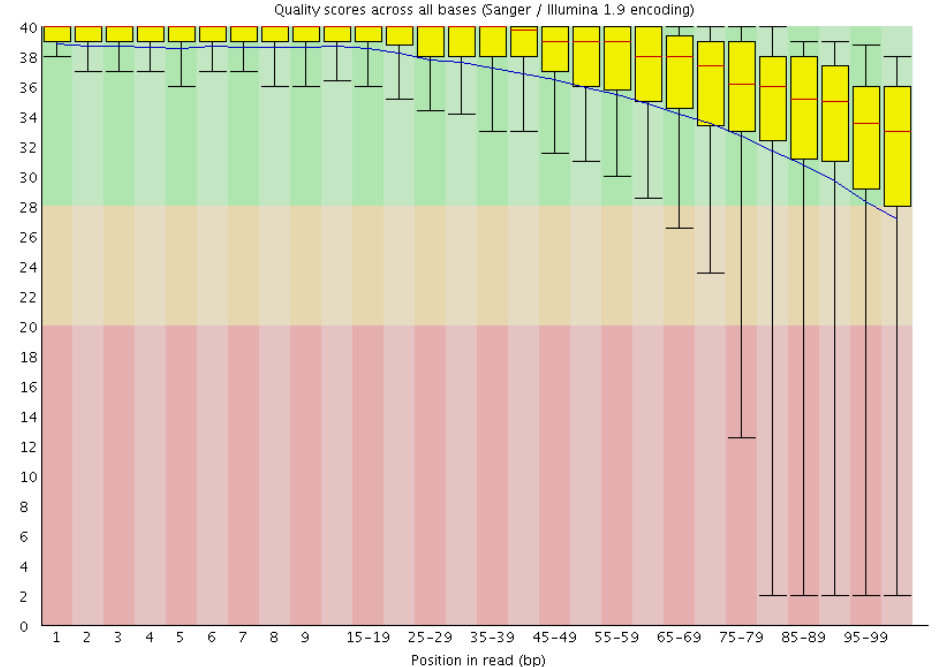
- BLAST / MEGAN
  - Alignment tool to determine microbiome diversity and distribution

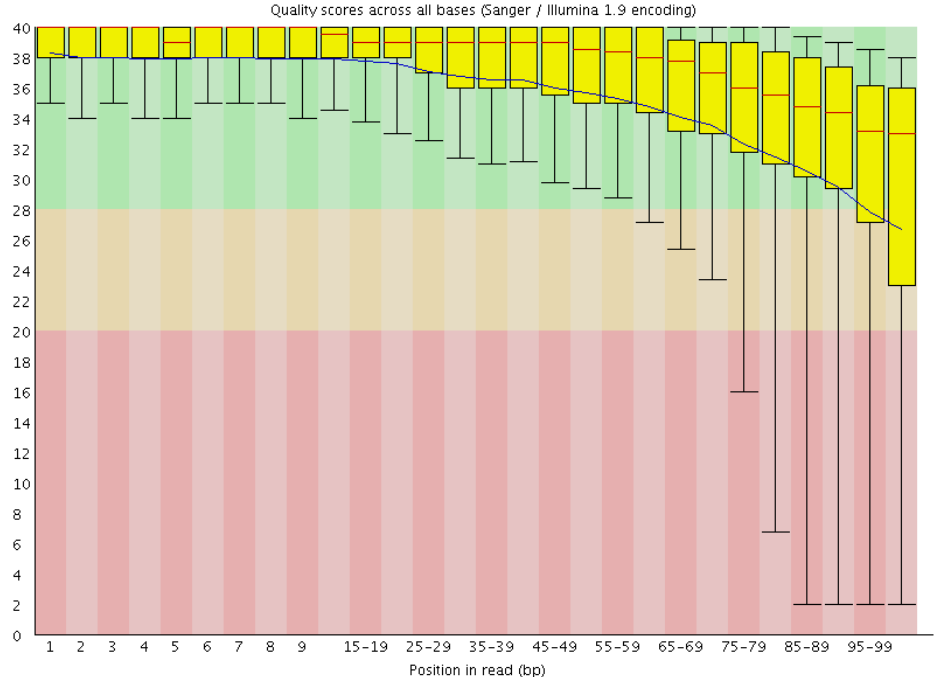
- HOMER
  - Motif finding algorithm for direct comparison of enriched motifs

# Per Base Quality

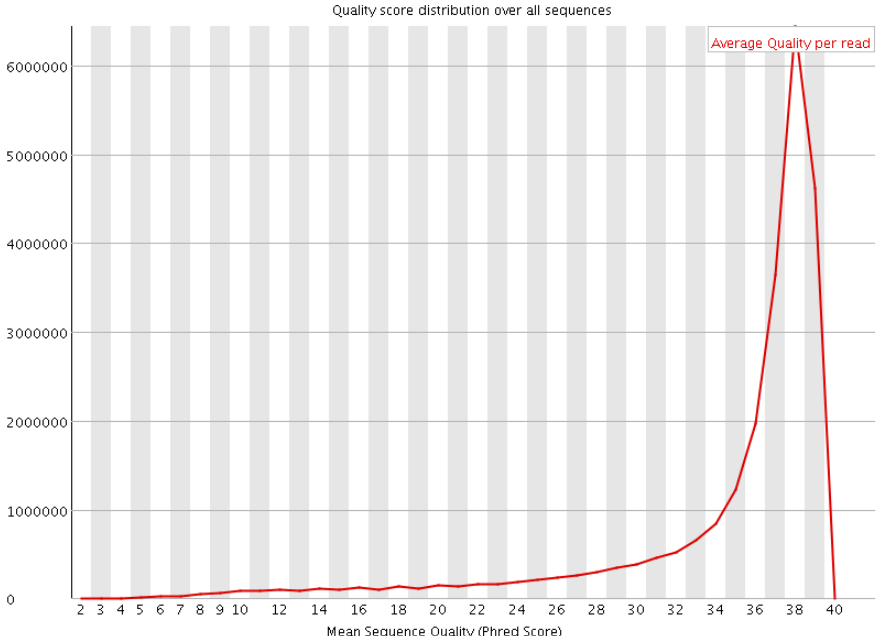


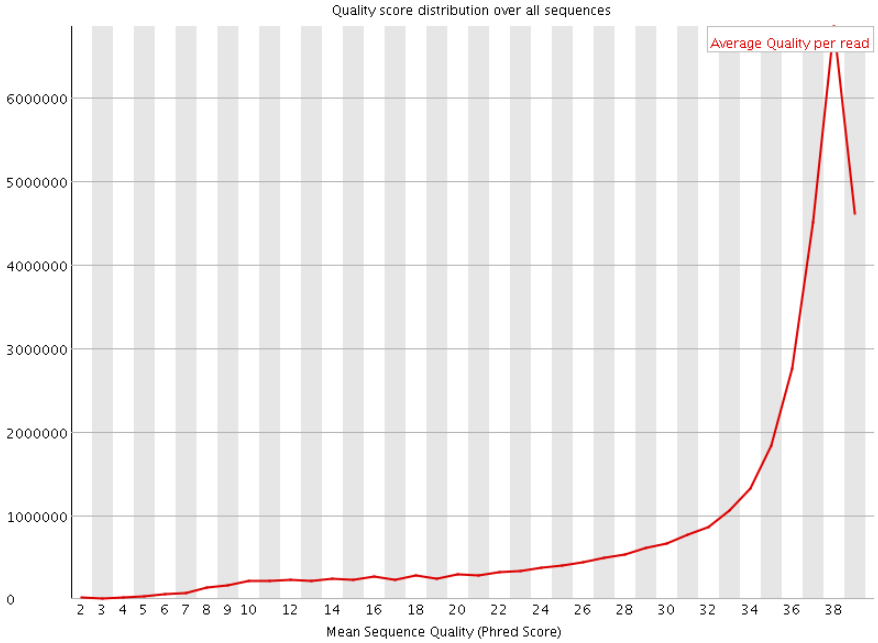



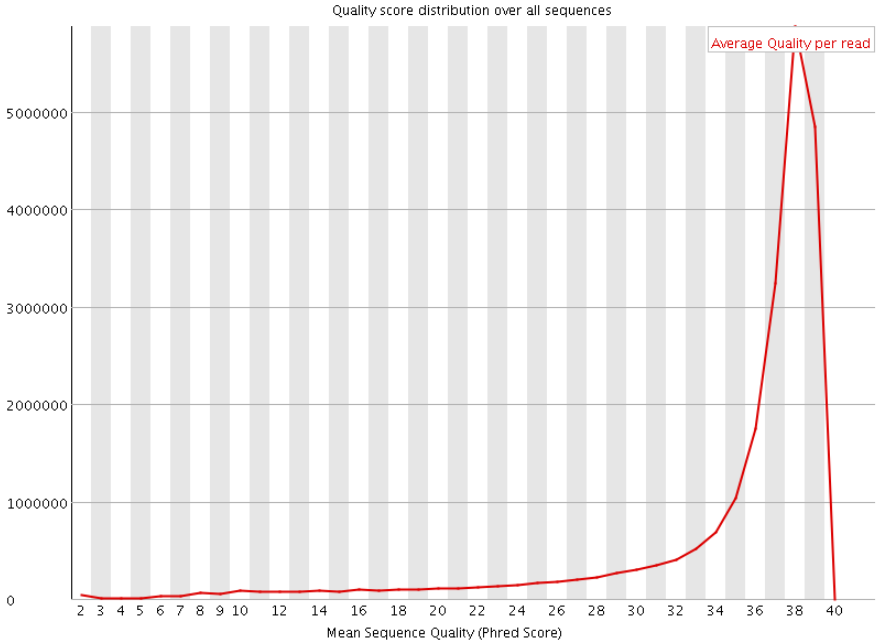



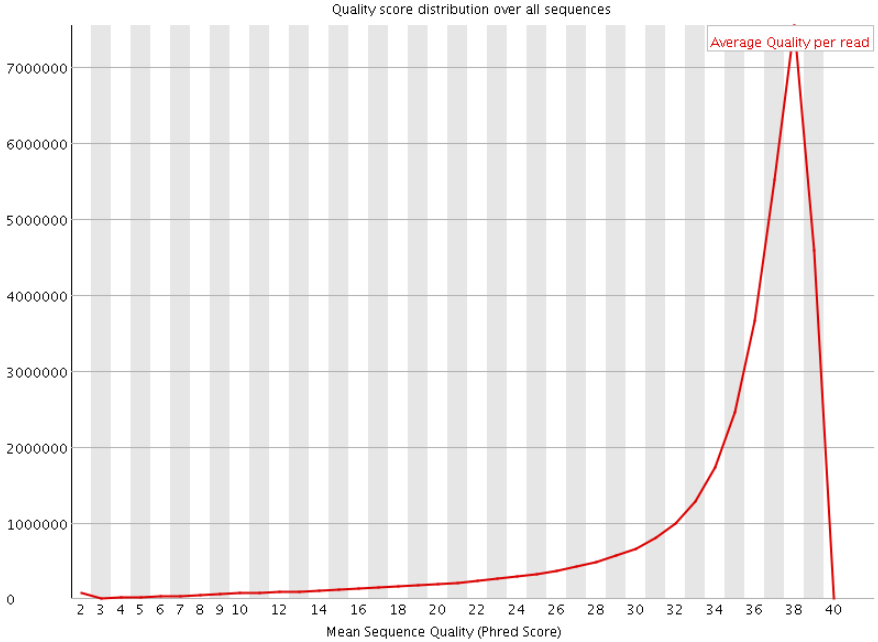



# Per Sequence Quality

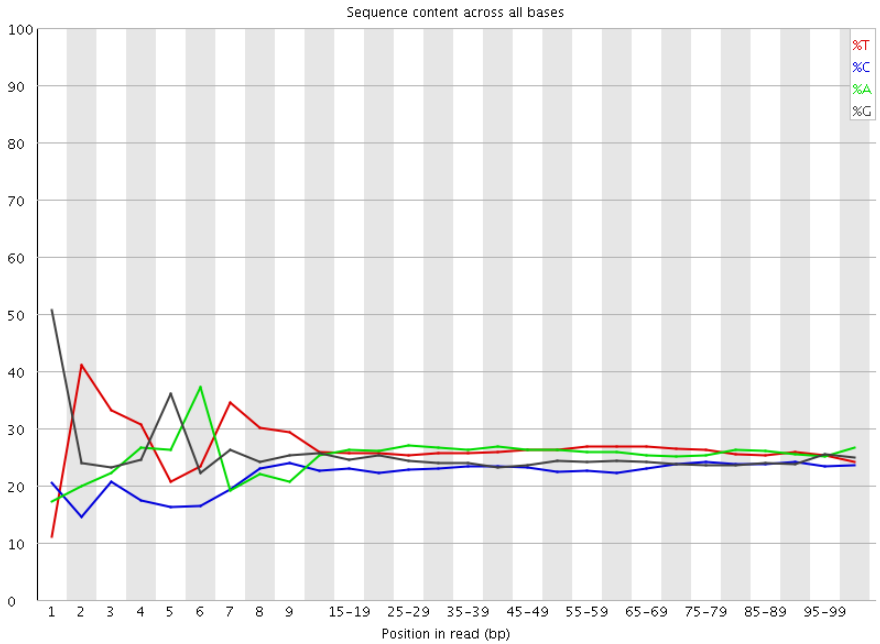


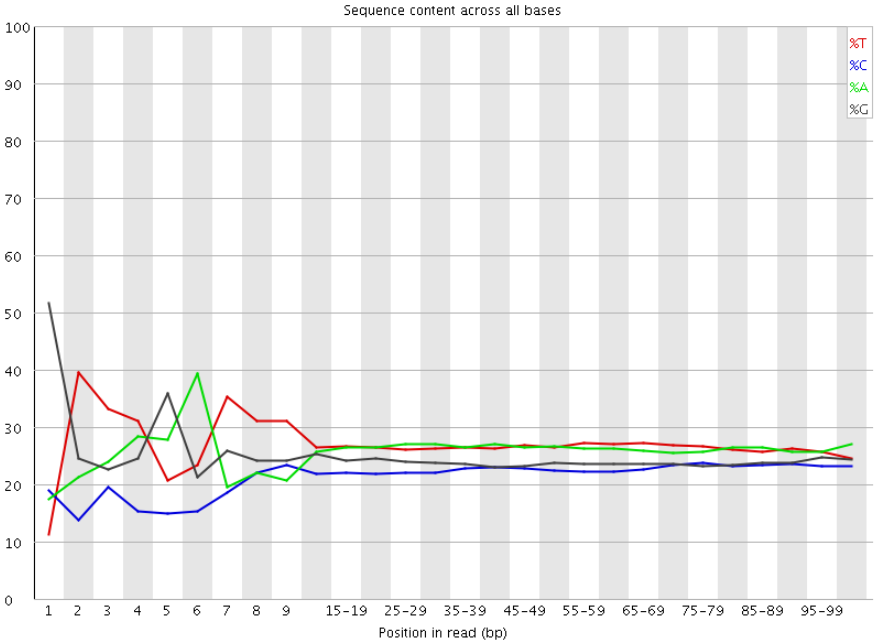



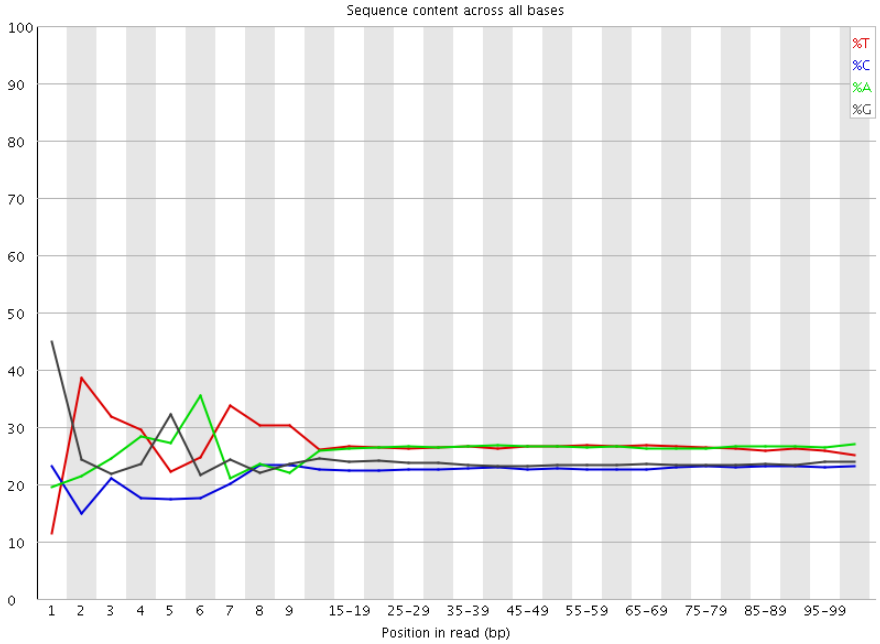



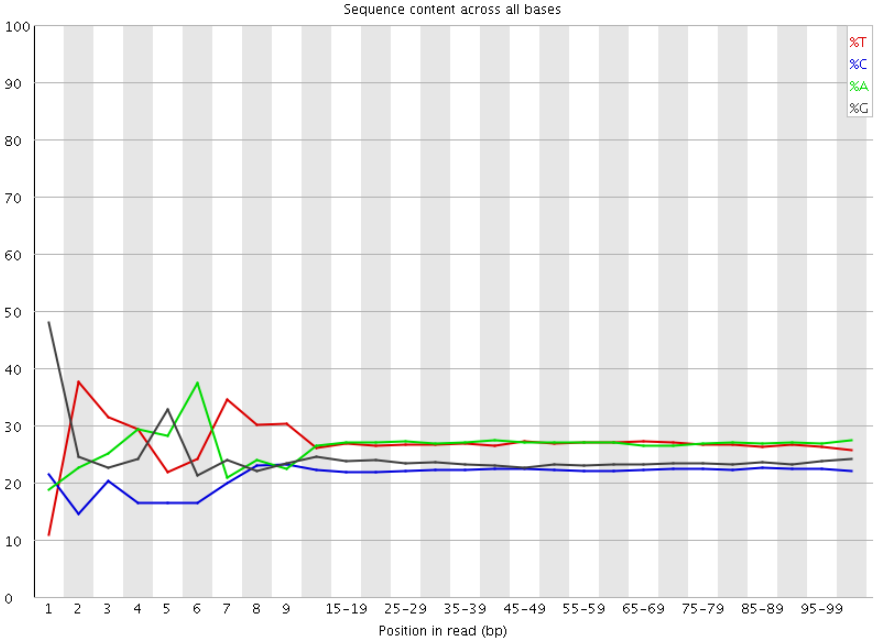



# Per Base Sequence Content

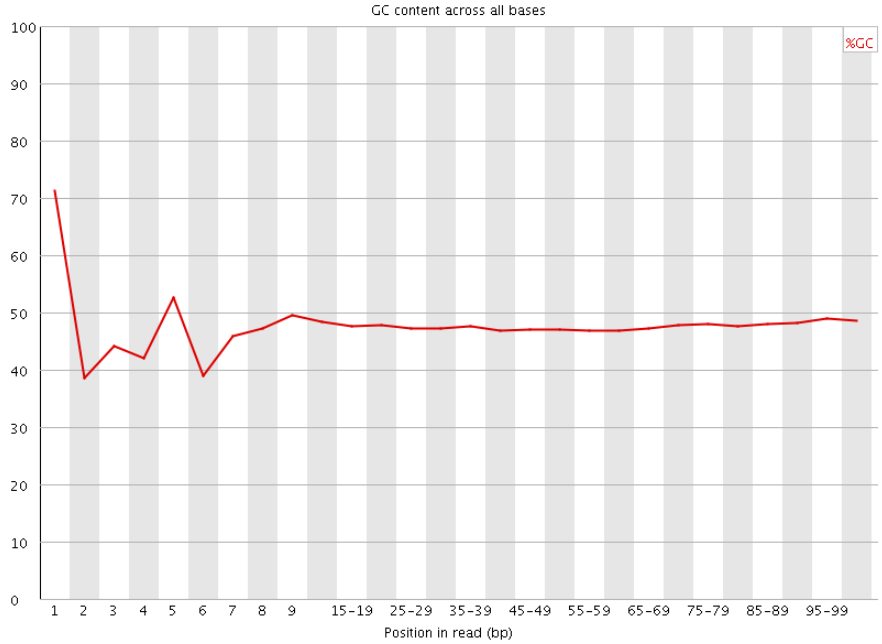

Sequence content across all bases

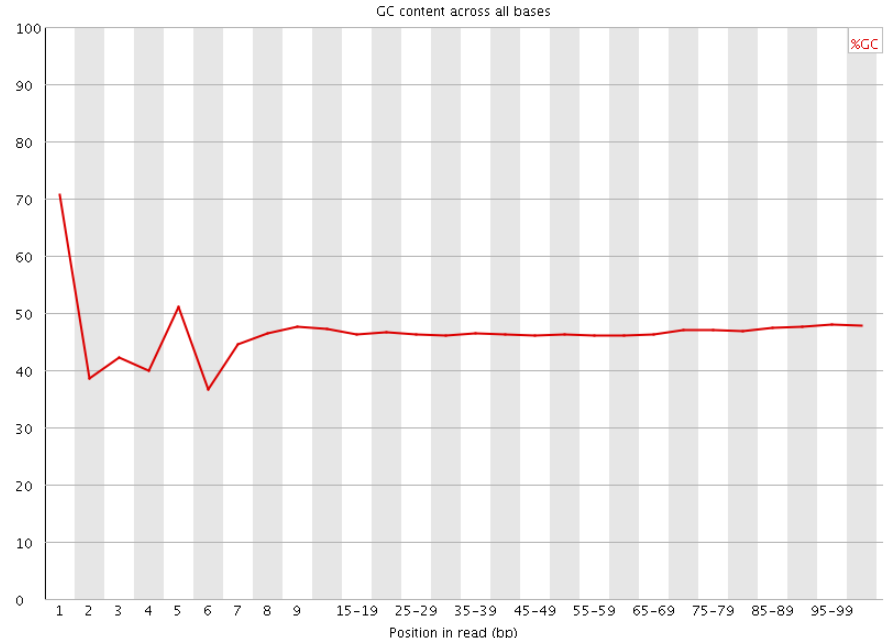

Sequence content across all bases

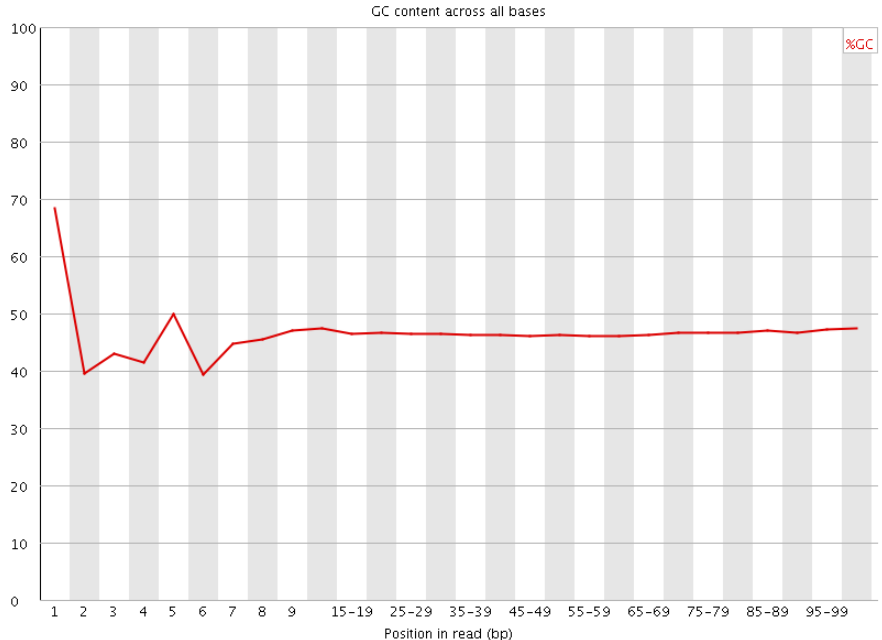

Sequence content across all bases

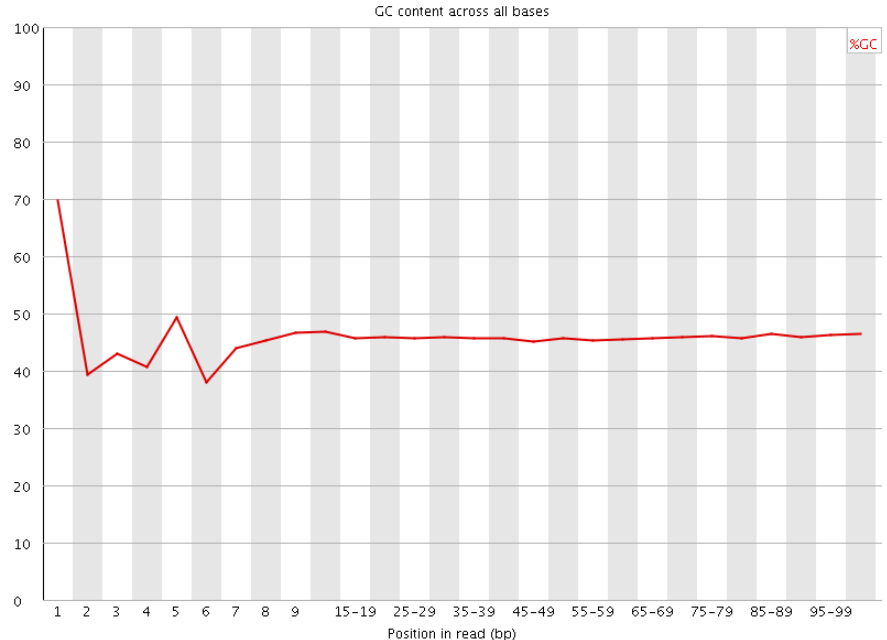

Sequence content across all bases

# Per Base GC Content

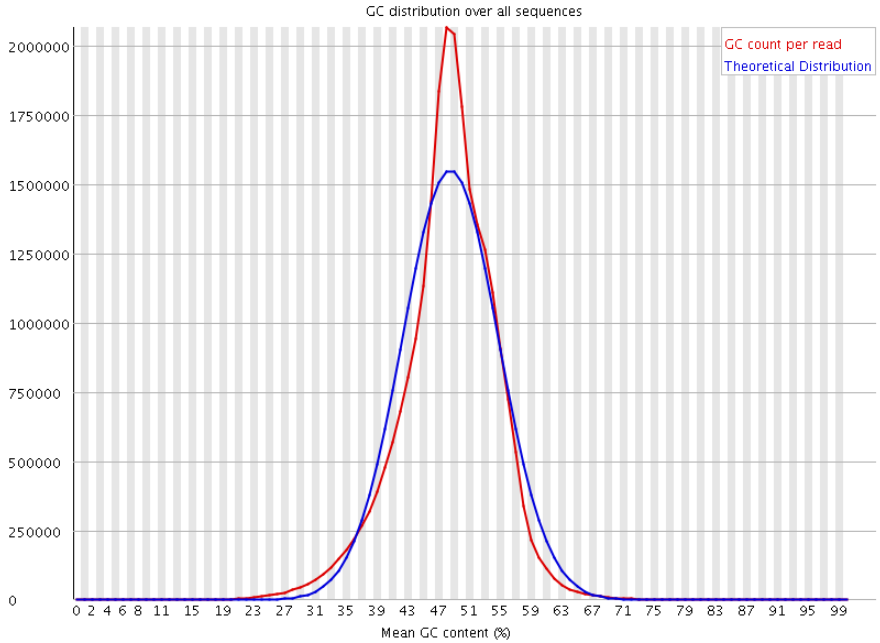


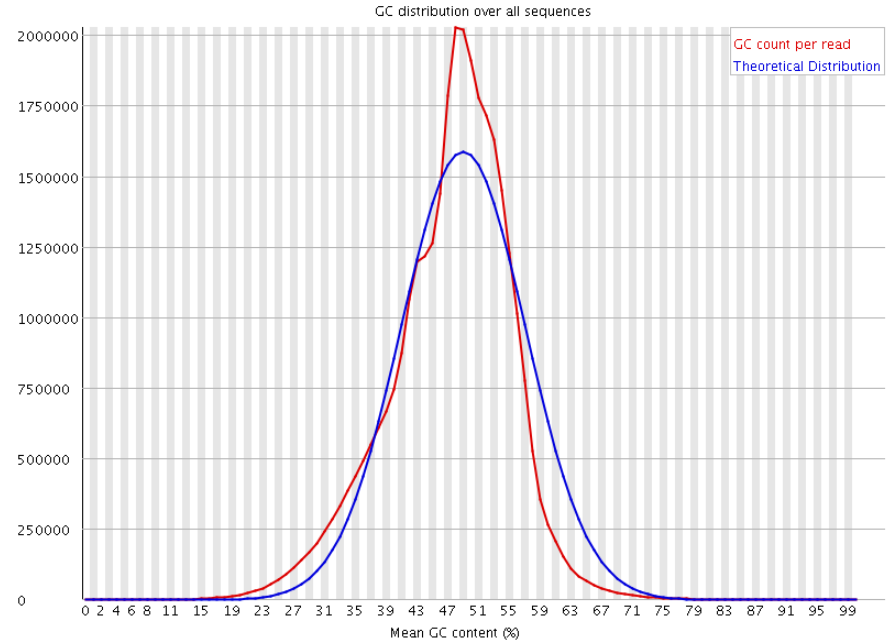


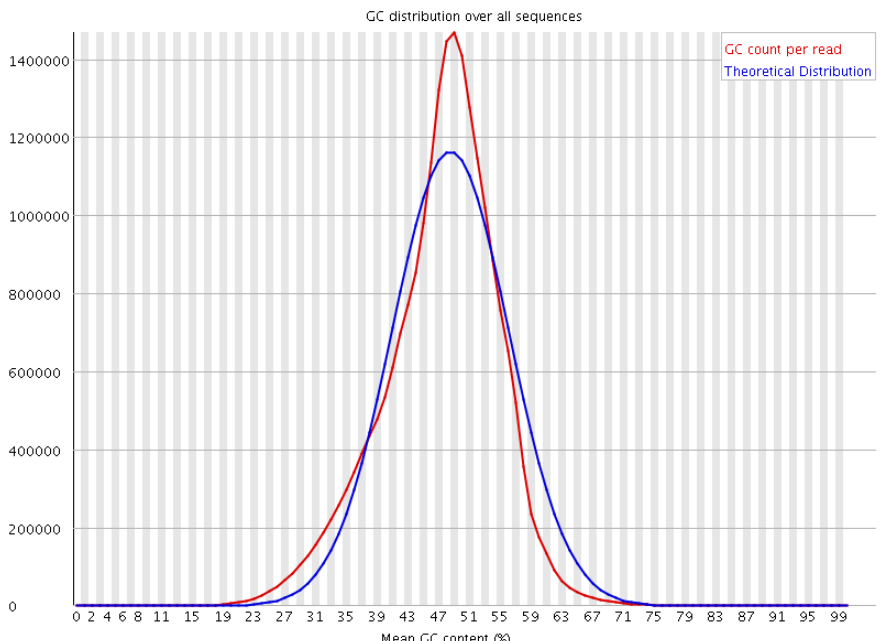


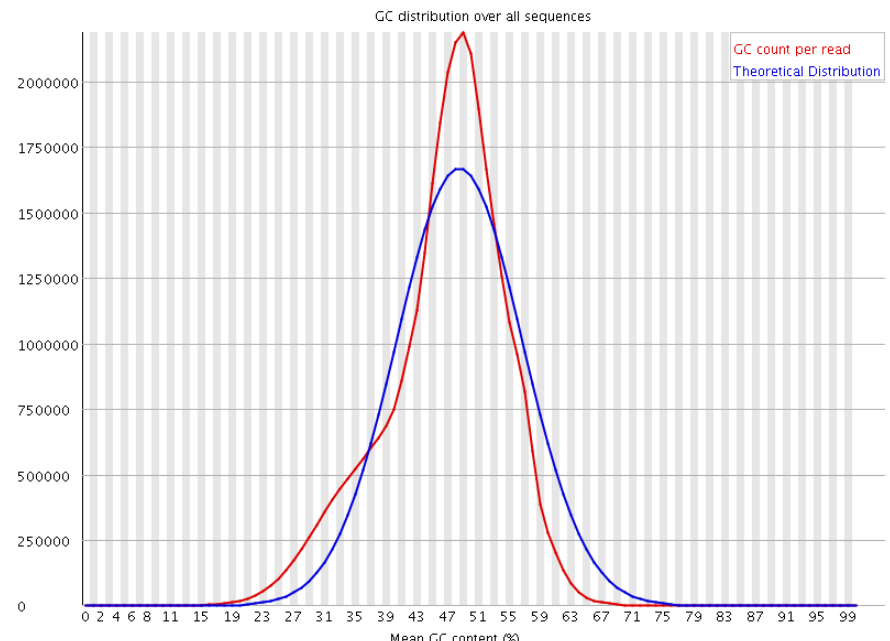


# Per Sequence GC Content

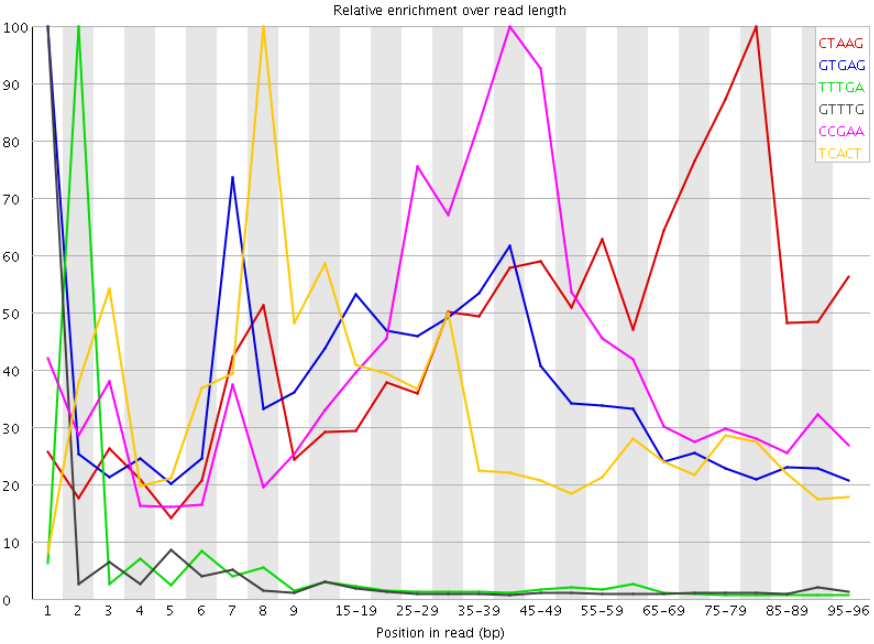


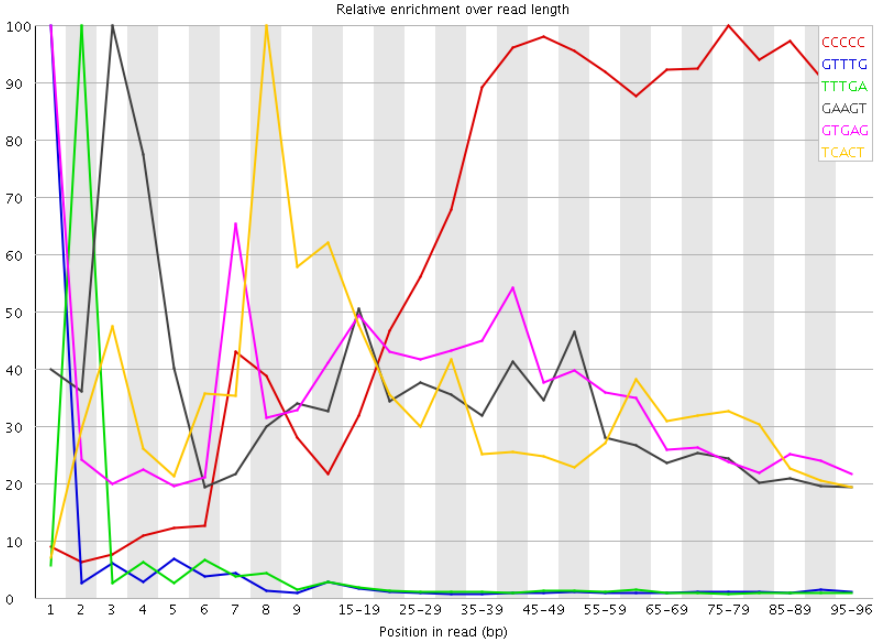



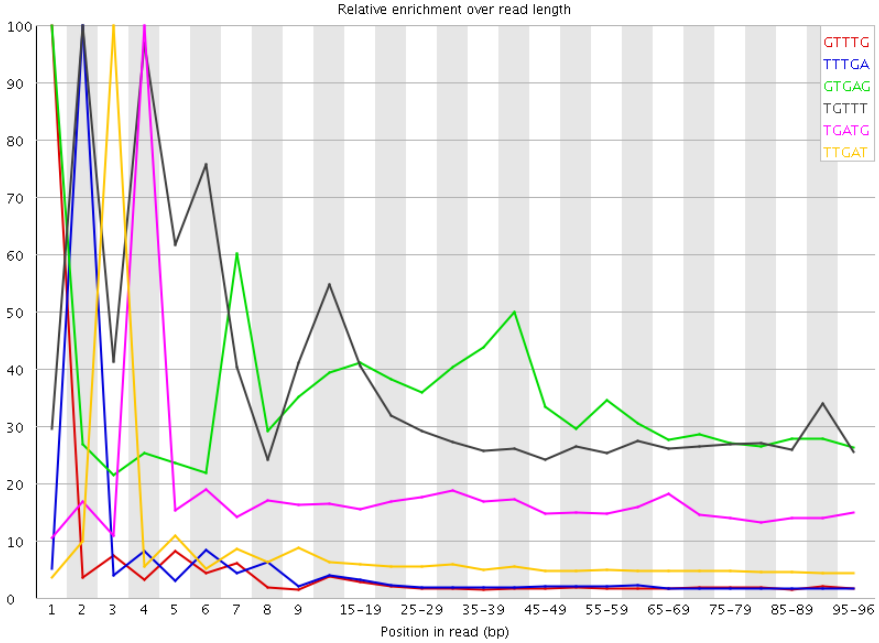



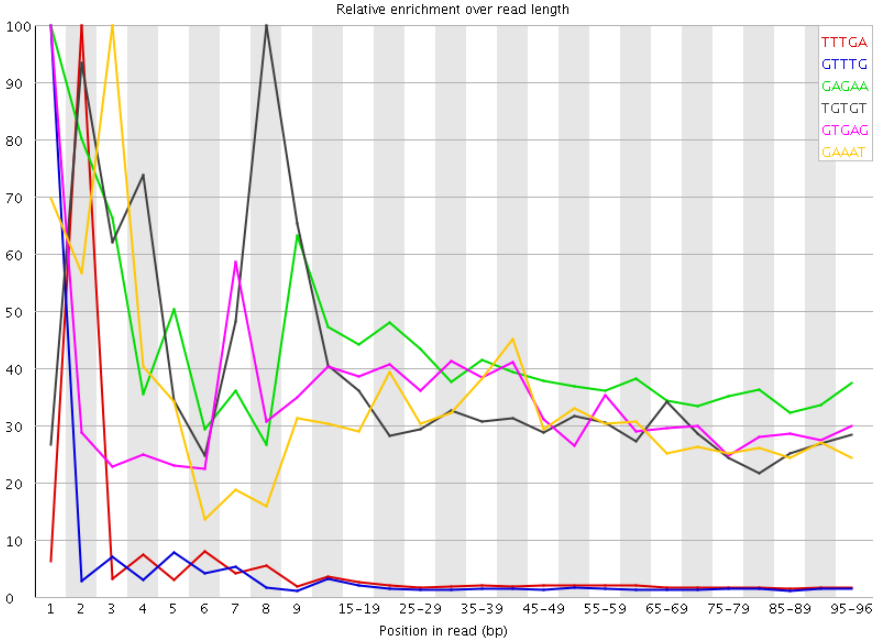



# K-mer Profiles

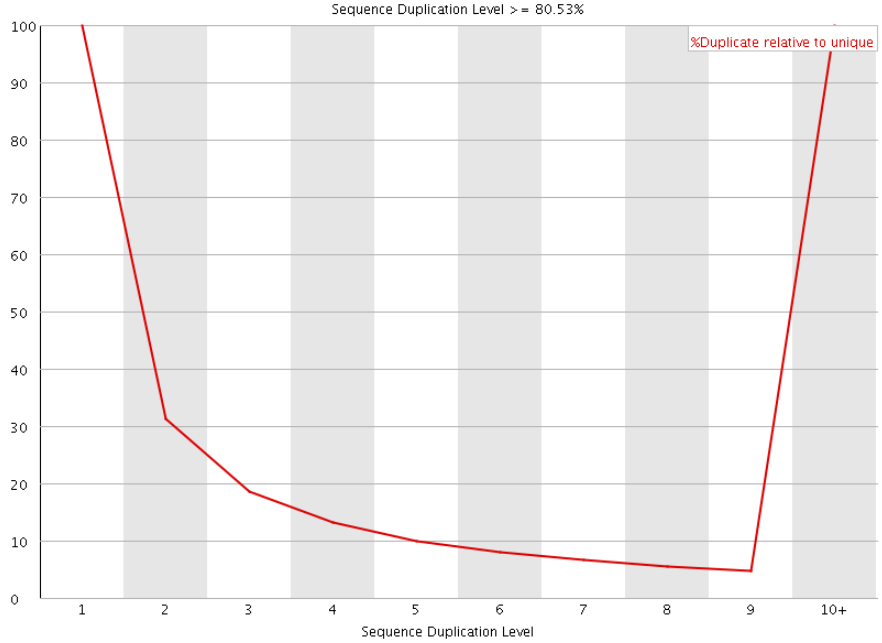


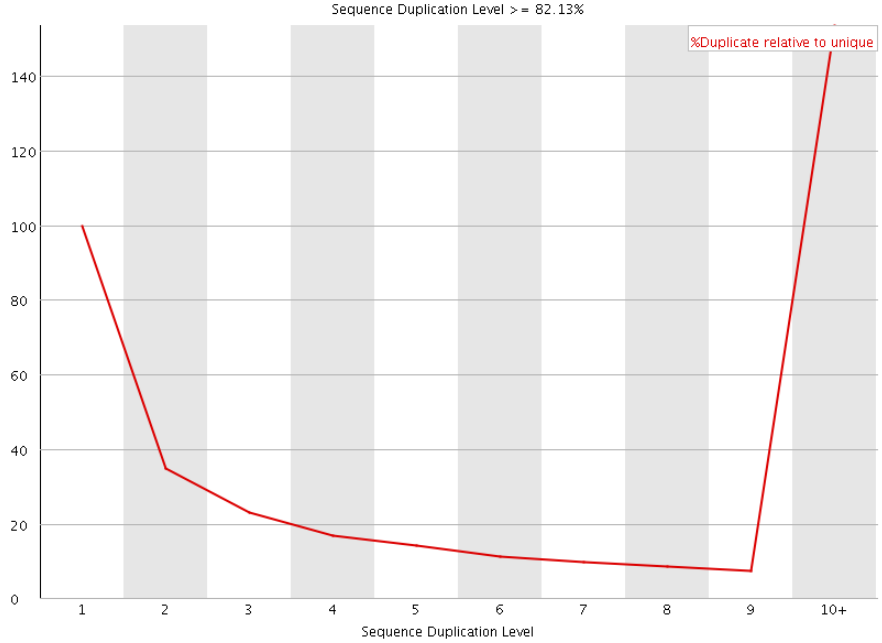



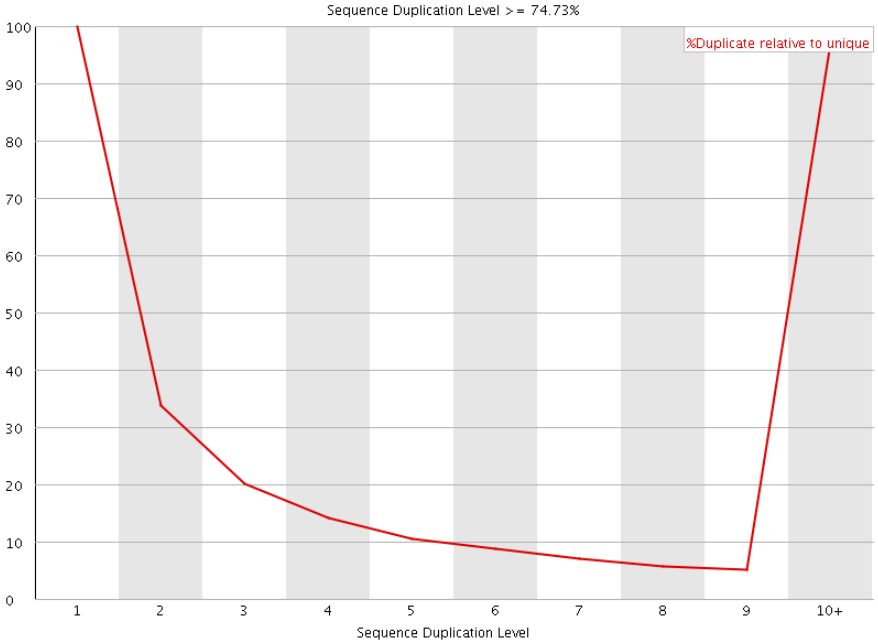



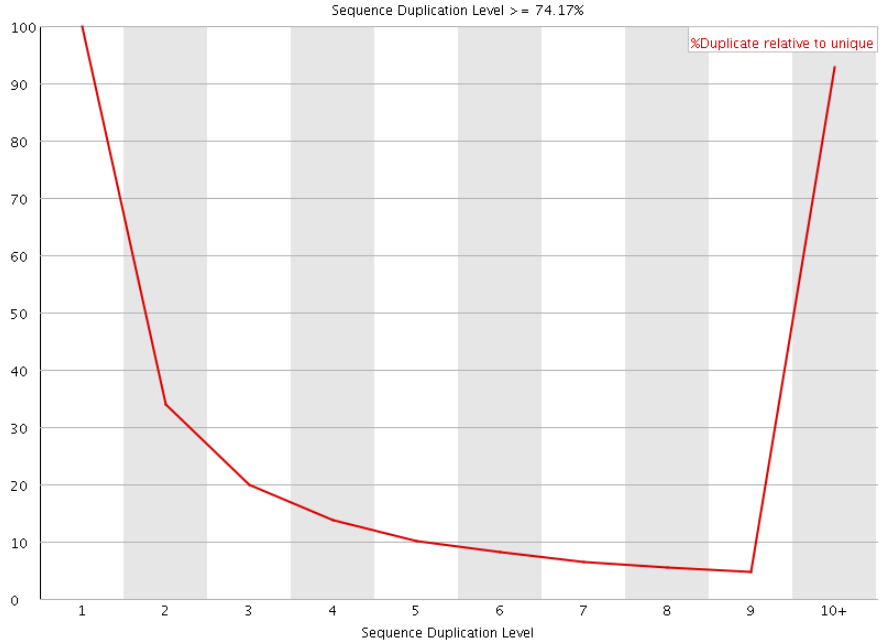



# Duplication Levels



# Bowtie2 output

```
jeyuan@gordon-ln1:/oasis/projects/nsf/csd399/jeyuan/twintestdata
  GNU nano 2.0.9            File: align.err

Time loading reference: 00:00:08
Time loading forward index: 00:00:19
Time loading mirror index: 00:00:10
Multiseed full-index search: 00:02:34
24324914 reads; of these:
  24324914 (100.00%) were unpaired; of these:
    23969933 (98.54%) aligned 0 times
    129054 (0.53%) aligned exactly 1 time
    225927 (0.93%) aligned >1 times
1.46% overall alignment rate
Time searching: 00:03:11
Overall time: 00:03:11

^G Get Help  ^O WriteOut  ^R Read File^Y Prev Page^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is ^V Next Page^U UnCut Tex^T To Spell
```

# Microbiome Data Analysis

**16S rDNA:** The genes coding for 16S rRNA are referred to as 16S rDNA and are used in reconstructing phylogenies.

**Align sequences to OTUs**

↓

**Build phylogenetic tree**

↓

**Use UniFrac for diversity analysis**

UPARSE
Fast, accurate OTU clustering

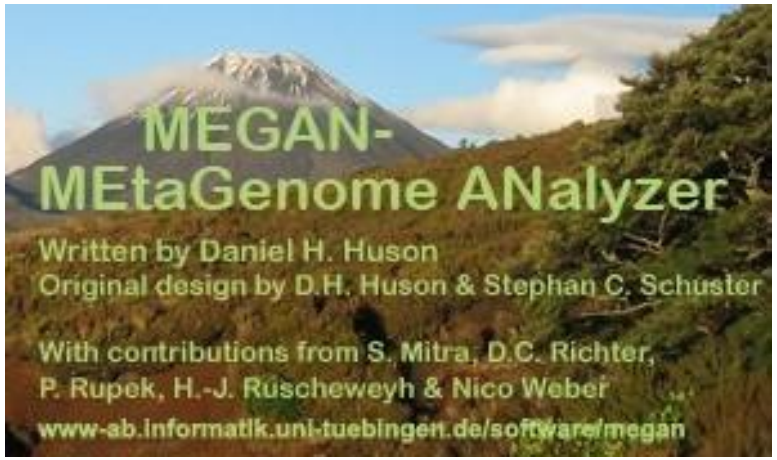
http://drive5.com/uparse/

**FastTree**

http://www.microbesonline.org/fasttree/

UniFrac

http://bmf.colorado.edu/unifrac/




- Taxonomic analysis, functional analysis and comparative analysis
- Requires a BLAST search beforehand

http://ab.inf.uni-tuebingen.de/software/megan/

**Our problems:**

➢ *No 16S rDNA sequence:* The dataset was generated to study gene expression differences between samples. rRNA sequence was removed.

➢ *File size limit:* 20 M

NIH HUMAN MICROBIOME PROJECT

Login

REFERENCE GENOMES | MICROBIOME ANALYSIS | IMPACTS ON HEALTH | TOOLS & TECHNOLOGY | ETHICAL IMPLICATIONS | OUTREACH | HMPDACC DATA BROWSER

home > blast

Feedback

**Instructions**

Select the Genome database you would like to Blast against. You may select "All Reference Genomes", a particular body site, or an individual organism.

Select the type of Blast you would like to perform.

Enter your Nucleotide or Protein Sequence into the text box or upload a file in plain text format containing the sequence or sequences you would like to analyze.

Give the search Job a descriptive title.

To begin, please either select a Reference Genome BLAST Database from the menu below, or select an individual organism:

Oral

Select an individual organism

- (●) BLASTN - nucleotide sequence against nucleotide sequence of predicted genes in this genome
- ( ) BLASTP - protein sequence against amino acid sequence of predicted genes in the genome
- ( ) TBLASTN - protein sequence against the entire genome sequence
- ( ) BLASTX - translated nucleotide query against amino acid sequence of predicted genes in the genome
- ( ) TBLASTX - translated nucleotide query against the entire genome sequence

Paste nucleotide or protein sequence below:

Upload File:

http://www.hmpdacc.org/resources/blast.php

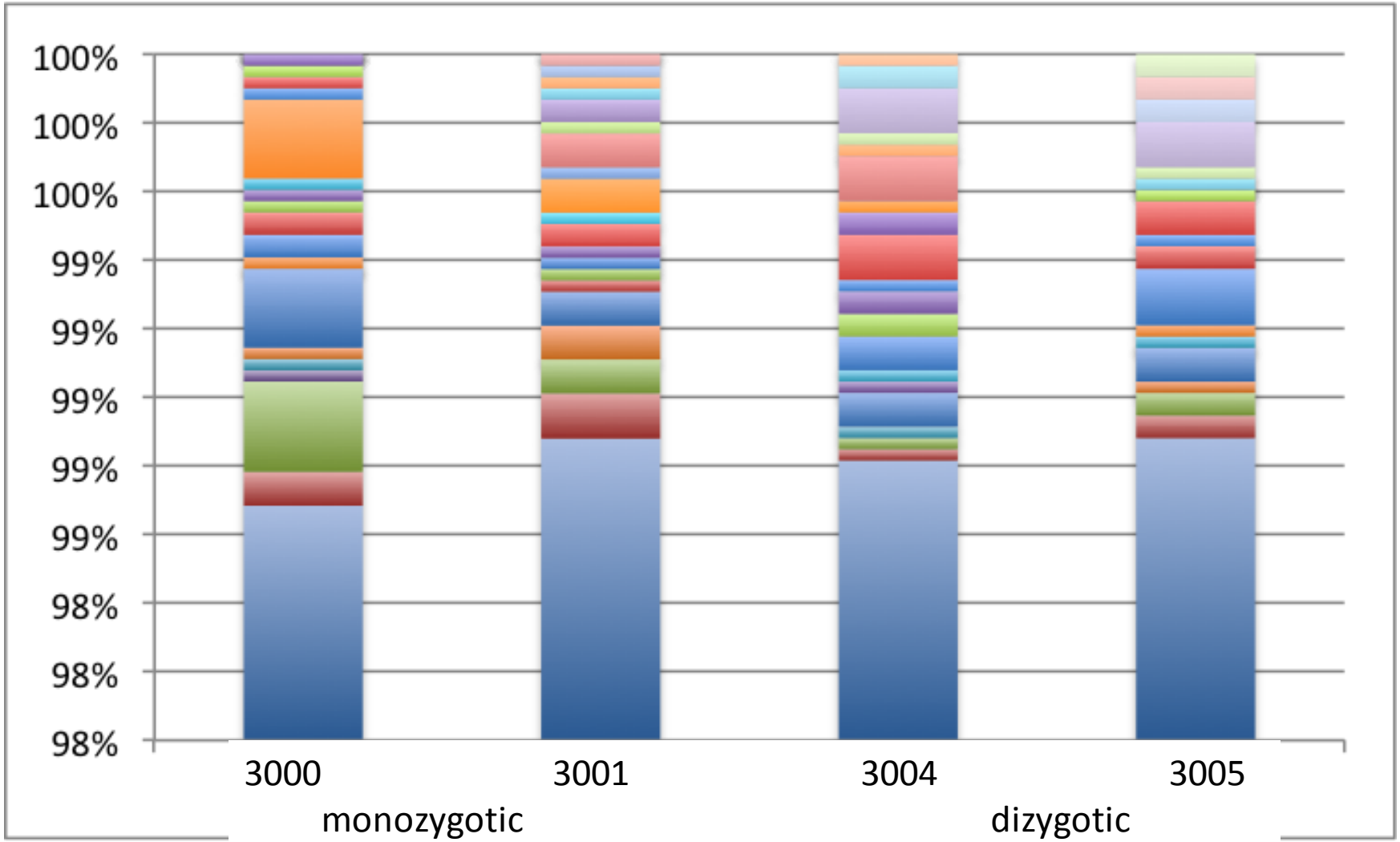
# Results of BLAST search against NCBI-NR database

| Sample No. | Monozygotic(MZ) or Dizygotic(DZ) | Caries Positive/Negative |
|---|---|---|
| 0998 | MZ | N |
| 0999 | MZ | N |
| 1000 | MZ | P |
| 1001 | MZ | P |
| 1002 | DZ | N |
| 1003 | DZ | N |

# BLAST Alignment Results

| Sample ID | Organism Matches | Different Types |
|---|---|---|
| **3000** | 40 | 11 |
| **3001** | 41 | 14 |
| **3004** | 40 | 14 |
| **3005** | 34 | 12 |

# Microbiome Relative Diversity



# HOMER Results

- Used the 0.01% random sampling from BLAST analysis
- Performed *de novo* motif finding against a random background
- Also performed motif finding against hg19 reference background

- Did not find any motifs meaningfully enriched in one twin pair type vs another

# Things We Learned

NGS is a multifacite analysis, pipelines don't exist yet.

fastq-dump

- Paired-ended != Single-ended reads
- sra-tools

sickle

- .bash_profile

SOAPdenovo-31mer != SOAPdenovo 31mer != SOAPdenovo

# Things We Learned

EMBOSS::getorf

.bash_profile

pfam_scan.pl

Bioperl

velvet

Bioperl

Tautological Analysis Teaching Allocation Substitution

Subversive History Instanciation of Transcription

# Summary

- *Question*: Is there a statistically meaningful difference in the microbiomes of monozygotic twins compared to dizygotic twins?

- SUMMARY OF FINDINGS

# Conclusions

- Ultimately, there is insufficient evidence to reject our null hypothesis, therefore, we still believe that our initial hypothesis could be true…

- However….

# Considerations & Future Work

- Assemble metagenome for deeper analysis
- More time to analyze entirety of data rather than a small random sampling and extrapolating conclusions
- Pick a less outrageously ambitious Boot Camp project in the first place