Assessment 3

s4536711@binf-training.biosci.uq.edu.au
Mkdir assessment3
cd assessment3
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR630/SRR630565/SRR630565_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR630/SRR630565/SRR630565_2.fastq.gz
fastqc SRR630565_1.fastq.gz SRR630565_2.fastq.gz

Sequencing platform from which sequence reads were generated

> zcat SRR630565_1.fastq.gz | head

The Illumina 1.8+ platform outputs quality scores on the Phred+33 scale, with quality values ranging between 0 and 41. Based on the quality scores '?' = Phred+33, score of 30, the sequencing platform for the SRR630565 reads is Illumina 1.8+ (paired end sequencing).

Length of each of the read

> zcat SRR630565_1.fastq.gz | head -n 2 | tail -n 1 | wc -c

- 102 – 1 = 101 (exclude end of line character)

> zcat SRR630565_2.fastq.gz | head -n 2 | tail -n 1 | wc -c

- 102 – 1 = 101

Number of reads

> zcat SRR630565_1.fastq.gz | wc -l

- 6083000 total number of lines in the first file

> zcat SRR630565_2.fastq.gz | wc -l

- 6083000 total number of lines in the second file

> zcat SRR630565_1.fastq.gz | head

- Each read has 4 lines so 6083000 / 4 = 1520750

> zcat SRR630565_1.fastq.gz | grep '@SRR' | wc

- 1520750 3041500 41879542

G+C content

> zcat SRR630565_1.fastq.gz | sed -n '2~4p' | awk '{gc+=gsub(/[GCgc]/,"")} {total+=length} END {print (gc/total)*100}'

- 56.104

> zcat SRR630565_2.fastq.gz | sed -n '2~4p' | awk '{gc+=gsub(/[GCgc]/,"")} {total+=length} END {print (gc/total)*100}'

- 56.1741

The GC content of both files is similar because they are sequencing opposite ends of the same DNA fragments. However, it is normal for there to be small differences in GC content due to natural variation and sequencing biases.

Assemble the genome de novo

> mv ~/week8/bbmap .
> velveth Assembly69 69 -shortPaired -fastq.gz -separate SRR630565_1.fastq.gz SRR630565_2.fastq.gz
> velvetg Assembly69 -ins_length 353 -exp_cov auto
> bbmap/stats.sh in=Assembly69/contigs.fa > Assembly69/contigs.fa.stats
> cd Assembly93
> more contigs.fa.stats

```
s4536711@binf-training[Assembly69] more contigs.fa.stats
A       C       G       T       N       IUPAC   Other   GC      GC_stdev
0.3243  0.1767  0.1814  0.3177  0.0002  0.0000  0.0000  0.3581  0.0500

Main genome scaffold total:             168
Main genome contig total:               186
Main genome scaffold sequence total:    2.782 MB
Main genome contig sequence total:      2.782 MB        0.020% gap
Main genome scaffold N/L50:             15/58.671 KB
Main genome contig N/L50:               16/55.733 KB
Main genome scaffold N/L90:             46/19.687 KB
Main genome contig N/L90:               49/18.765 KB
Max scaffold length:                    183.425 KB
Max contig length:                      183.425 KB
Number of scaffolds > 50 KB:            20
% main genome in scaffolds > 50 KB:     61.03%


Minimum         Number          Number          Total           Total           Scaffold
Scaffold        of              of              Scaffold        Contig          Contig
Length          Scaffolds       Contigs         Length          Length          Coverage
--------        --------------  --------------  --------------  --------------  --------
    All                    168             186       2,782,188       2,781,640    99.98%
    100                    168             186       2,782,188       2,781,640    99.98%
    250                    132             150       2,775,440       2,774,892    99.98%
    500                    108             126       2,766,462       2,765,914    99.98%
   1 KB                     90             106       2,752,341       2,751,822    99.98%
 2.5 KB                     67              76       2,714,271       2,714,030    99.99%
   5 KB                     62              68       2,699,028       2,698,893   100.00%
  10 KB                     55              59       2,653,553       2,653,438   100.00%
  25 KB                     41              45       2,419,504       2,419,389   100.00%
  50 KB                     20              24       1,697,936       1,697,821    99.99%
 100 KB                      6               6         770,464         770,464   100.00%
```

(a) Brief description of your approach (name(s) of program(s), key parameters used)
69 is the k-mer size used to construct the initial assembly graph. **-shortPaired**: Indicates that paired-end reads are being used, enhancing the assembly's ability to connect fragments. **-fastq.gz -separate**: Indicates that the input files are in compressed FASTQ format, with the paired-end reads in separate files (SRR630565_1.fastq.gz and SRR630565_2.fastq.gz). Setting insert length to 353 allowed velvetg to use paired-end information more effectively. -exp_cov auto, calculated the expected coverage based on the input data, improving its ability to identify true genomic regions versus low-coverage noise. This prevented the loss of genomic segments and helped achieve a more accurate assembly size. BBMap uses the stats.sh script to generate statistics about the assembled contigs in the contigs.fa file. The output is redirected to a file (Assembly69/contigs.fa.stats), which provides detailed information on metrics such as N50, maximum contig length, and GC content, helping to evaluate the quality of the assembly.

(b) N50 scaffold length

- 58.671 KB

(c) Maximum scaffold length

- 183.425 KB

(d) Total scaffold length

- 2,782,188 KB

(e) Total number of scaffolds

- 2.782 MB

(f) Percentage of reads that are assembled into contigs
> cat Log

- 2178034/3041500 reads = 71.6%

(g) Mean coverage for all contigs
> R
> A <- read.table ("stats.txt", header=T)
> summary(A$short1_cov)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9.589  22.466  35.000  57.216  61.929 280.786
```

- Mean = 57.216%

Ab initio gene prediction
> genemark -opn -m Staph_aureus_JKD6008.mat Assembly69/contigs.fa

(a) Brief description of your approach (name(s) of program(s), key parameters used)
The approach involves using GeneMark with an ab initio gene prediction model designed for Staphylococcus aureus, to predict genes in the assembled genome contig.fa. The key parameters include using the prokaryotic gene prediction mode (-opn) and specifying a related species' model to improve prediction accuracy, leveraging the genomic similarities between the two species.

(b) Total number of predicted genes
> grep -c '>' contigs.fa.orf
- 5142

> genemark -on -m Staph_aureus_JKD6008.mat Assembly69/contigs.fa
> mv contigs.fa.orf contigs.cds.fasta
> grep -c '>' contigs.cds.fasta
- 2571

> genemark -op -m Staph_aureus_JKD6008.mat Assembly69/contigs.fa
> mv contigs.fa.orf contigs.prot.fasta
> grep -c '>' contigs.prot.fasta
- 2571

(c) Average gene length
```
total_length=$(grep -v '^>' contigs.fa.orf | awk '{sum += length($0)} END {print sum}')
num_genes=$(grep -c '^>' contigs.fa.orf)
average_length=$(echo "$total_length / $num_genes" | bc)
echo "Average gene length: $average_length bp"
```
Average gene length: 613 bp
```
total_length=$(grep -v '^>' contigs.cds.fasta | awk '{sum += length($0)} END {print sum}')
num_genes=$(grep -c '^>' contigs.cds.fasta)
average_length=$(echo "$total_length / $num_genes" | bc)
echo "Average gene length: $average_length bp"
```
Average gene length: 920 bp
```
total_length=$(grep -v '^>' contigs.prot.fasta | awk '{sum += length($0)} END {print sum}')
num_genes=$(grep -c '^>' contigs.prot.fasta)
average_length=$(echo "$total_length / $num_genes" | bc)
echo "Average gene length: $average_length bp"
```
Average gene length: 306 bp

(d) Length and function of the longest gene
> cd assessment3
> mv ../assessment3/Assembly69/contigs.fa .
> mv ../assessment3/Assembly69/contigs.cds.fasta .
> mv ../assessment3/Assembly69/contigs.prot.fasta .
> nano contigs.cds.fasta (delete the first 4 lines)
> nano contigs.prot.fasta (delete the first 4 lines)
> awk '/^>/ {if (seqlen > max) {max = seqlen; longest_header = header; longest_seq = seq} header = $0; seqlen = 0; seq = ""} !/^>/ {seqlen += length($0); seq = seq $0} END {if (seqlen > max) {longest_header = header; longest_seq = seq}; print longest_header; print longest_seq}' contigs.prot.fasta

- **Length = 1650586−1634756= 15830**

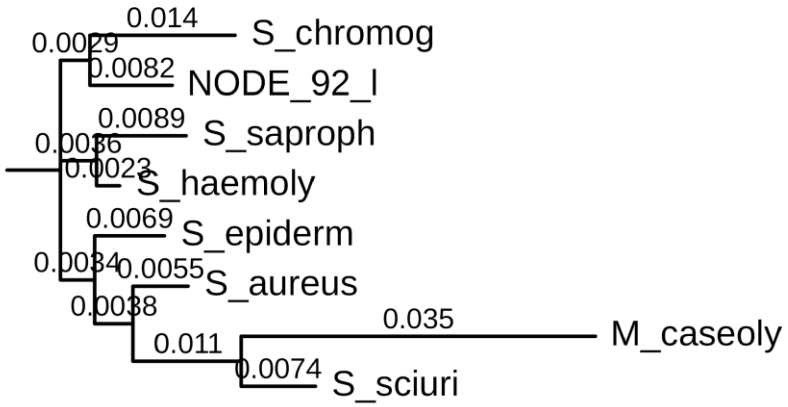
Use orf_1522 (NODE_1_length_46816_cov_16.656635, 1634756 - 1650586) translated
→ NCBI → Blastp → G5 domain-containing protein [Staphylococcus simulans]
This record represents a single, non-redundant, protein sequence which may be annotated on many different RefSeq genomes from the same, or different, species. The longest gene encodes a G5 domain-containing protein, which is involved in cell wall maintenance, adhesion, or carbohydrate binding. This functional annotation is based on the protein's similarity to sequences found in *Staphylococcus simulans*. G5 domain-containing proteins are commonly associated with interactions with host cells or structural roles within the bacterial cell wall.

Phylogenomic analysis (maximum two pages; 8)
> makeblastdb -dbtype nucl -in contigs.fa
> blastn -query gene_query.fa -db contigs.fa -outfmt 7 -evalue 1e-10
16S_rRNA NODE_92_length_817_cov_181.892288 98.983 885 710 1594 885
> samtools faidx -i contigs.fa NODE_92_length_817_cov_181.892288:1-885 > My_16S.fasta
> cp /opt/BINF7001/2024/Prac9_2024/Bact_16S_rRNA.fa .
> cat My_16S.fasta >> Bact_16S_rRNA.fa
> grep '>' Bact_16S_rRNA.fa
> cp contigs.prot.fasta proteins
> orthofinder -og -f proteins
> cd proteins/OrthoFinder/Results_Oct01/Comparative_Genomics_Statistics
> cat Statistics_Overall.tsv

(a) Brief description of your approach (name(s) of program(s), key parameters used)
The phylogenomic analysis involved identifying the **16S ribosomal RNA** sequence from the assembled genome, followed by comparative genomic analysis using **OrthoFinder.** -dbtype nucl specifies that the input (contigs.fa) is a nucleotide database. -outfmt 7 formats the BLAST output in a tabular form. -evalue 1e-10 filters for highly significant matches, reducing false positives. The **16S rRNA** was identified in a contig (NODE_92_length_817_cov_181.892288) with a high sequence identity (98.983%) over an alignment length of 885 bases. faidx extracts a specific region from the genome sequence. -i generates the reverse complement of the sequence if needed (in this case, extracting bases 1 to 885). Then, incorporate the identified 16S rRNA sequence into a known set of 16S sequences (Bact_16S_rRNA.fa), preparing it for further phylogenetic analysis. The entire set of predicted protein sequences was used for orthologous group identification with **OrthoFinder**. -og specifies that OrthoFinder should identify orthologous groups in the provided protein sequences.

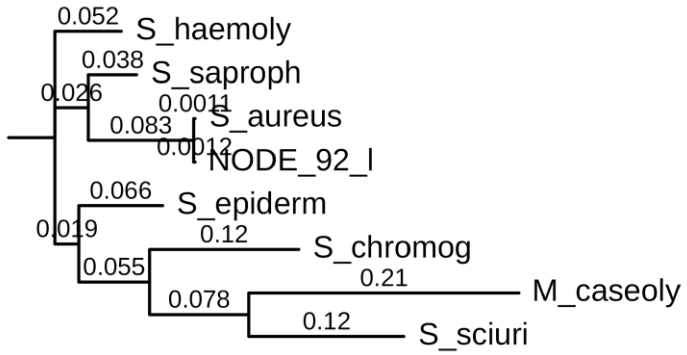
(b) Total number of homologous protein groups, and number of single-copy groups (putative orthologous groups)
> cat Statistics_Overall.tsv
2930, 1026

(c) 16S rRNA gene tree (8 taxa, rooted using outgroup)
> muscle -in Bact_16S_rRNA.fa -out Bact_16S.rRNA.aln
> readseq -a -f17 Bact_16S.rRNA.aln > Bact_16S.rRNA.nex
> cat mb_nuc_param.txt >> Bact_16S.rRNA.nex
> nano Bact_16S.rRNA.nex (to edit the MrBayes)
> mb Bact_16S.rRNA.nex
> cat Bact_16S.rRNA.nex.con.tre



(d) protein tree of pykA (8 taxa, rooted using outgroup)

> grep 'pyruvate kinase' *faa
> grep WP_041636071.1 OrthoFinder/Results_Oct01/Orthogroup_Sequences/*.fa
> grep '>' OrthoFinder/Results_Oct01/Orthogroup_Sequences/OG0001167.fa
> cp OrthoFinder/Results_Oct01/Orthogroup_Sequences/OG0001167.fa ../pykA.prot.fasta
> muscle -in pykA.prot.fasta -out pykA.prot.aln
> readseq -a -f17 pykA.prot.aln > pykA.prot.nex
> cat mb_nuc_param.txt >> pykA.prot.nex
> nano pykA.prot.nex (to edit the MrBayes)
> mb pykA.prot.nex
> **cat pykA.prot.nex.con.tre**
> sed -i 's/WP_041636/M_caseoly/g; s/WP_048539/S_sciuri/g; s/WP_037574/S_chromog/g; s/WP_048793/S_epiderm/g; s/YP_500311/S_aureus/g; s/orf_1716/NODE_92_l/g; s/WP_001830/S_saproph/g; s/WP_011275/S_haemoly/g' pykA.prot.nex.con.tre



(e) one key difference/similarity between the two trees

In both trees, M_caseoly and S_sciuri appear to form the outgroup. An outgroup helps determine the direction of evolutionary changes and identify the root of the tree. In the 16S rRNA tree, M_caseoly and S_sciuri are positioned as the longest branch length from the other taxa, suggesting that they are less closely related to the other species in the study. Similarly, in the pykA protein tree, M_caseoly and S_sciuri also serve as the outgroup. This indicates a consistent phylogenetic relationship where M_caseoly and S_sciuri are more closely related to each other than to the other taxa in both trees. A key difference between the two trees is the placement of some taxa within the tree. For example, in the 16S rRNA tree: S_chromog and NODE_92_l are grouped together. However, in the pykA protein tree: S_aureus is grouped with the client's gene instead of being closer to S_epiderm. In the 16S rRNA gene tree, the branch lengths are short overall, indicating a closer evolutionary relationship or slower mutation rates among these taxa. For example, the branches leading to S_sciuri and M_caseoly are short, suggesting a recent common ancestor. In the pykA protein tree, branch lengths leading to M_caseoly and S_sciuri are longer, implying more significant evolutionary changes or divergence in the pykA gene compared to the 16S rRNA gene. The branch lengths also show more variation in the pykA tree, indicating differing evolutionary pressures on the pykA protein across these species.

(f) one plausible explanation as to why such a difference occurs

The 16S rRNA gene is a part of the ribosomal RNA complex and plays a role in protein synthesis. It is highly conserved across different species due to its essential function, and any mutations in this region can be disastrous. This high level of conservation leads to slower mutation rates, which is why the 16S rRNA tree shows shorter branch lengths and more consistent evolutionary relationships. pykA encodes for pyruvate kinase, an enzyme involved in glycolysis. This gene is subject to different selective pressures across different bacterial species due to variations in metabolic requirements and environmental factors. Thus, the pykA gene may evolve at a faster rate than the 16S rRNA gene, leading to longer branch lengths and differences in species relationships in the phylogenetic tree. In bacteria, genes like pykA can be transferred horizontally between species, which can obscure the true evolutionary history of the organisms. If horizontal gene transfer has occurred, the pykA tree may show groupings that do not reflect the organism's evolutionary lineage, leading to the observed differences in the tree topology compared to the more conserved 16S rRNA gene. In contrast, the 16S rRNA gene is rarely involved in horizontal gene transfer, making it a more reliable marker for determining phylogenetic relationships. This could explain why taxa are grouped differently in the pykA tree compared to the 16S rRNA tree.

Limitations
The limitation of short reads restricts the assembly's ability to span large repetitive regions, often resulting in a fragmented genome assembly. Despite achieving an average contig coverage of approximately 57.216%, the assembly still suffer from uneven coverage across different genomic regions. Areas with low coverage can lead to incomplete assembly representation. Furthermore, the selection of a fixed k-mer size (69) influences the quality of the assembly. While this size might work well for the current dataset, it is not be optimal for all genomic regions. Additionally, setting the insertion length to 353 bp, based on sequencing data, leverages paired-end information but could introduce misassemblies if there are inaccuracies in the insert size estimation. The use of velvetg with the -exp_cov auto parameter relies on the program's estimation of expected coverage from the input data. While this automated method filters out low-coverage noise, it struggles to differentiate between authentic genomic regions and sequencing errors, leading to loss of valid regions in the final assembly.

For gene prediction, an ab initio approach with the GeneMark model tailored for *Staphylococcus aureus*. Relying on a model specific to *S. aureus* may result in missing species-specific genes in the client's protein or inappropriately annotating non-functional regions as genes. The 16S rRNA sequence identification was based solely on sequence similarity, which is effective for identifying regions homologous to known 16S rRNA sequences but fails to account for strain-specific mutations or novel elements within the 16S rRNA gene. Thus, this method will not capture strain-specific evolutionary nuances. Moreover, the BLAST search focused on identifying the single best-hit region within the assembled genome, overlooking the sequence diversity that exists in multi-copy rRNA genes found in bacterial genomes. The phylogenetic tree constructed using the 16S rRNA gene provides an initial understanding of evolutionary relationships but represents only a fragment of the organism's genome. Similarly, analyzing the phylogenetic relationships using a single protein-coding gene, such as *pykA*, does not accurately reflect the full evolutionary history, especially if gene duplication events have occurred.

Recommendations
It is highly recommended to incorporate long-read sequencing technologies, such as PacBio or Oxford Nanopore, in future genome assembly efforts. Alternatively, a hybrid assembly approach that combines both short- and long-read data should be employed. Long-read technologies offer the advantage of spanning repetitive regions and complex genomic structures that short reads cannot resolve, thereby significantly reducing fragmentation in the genome assembly and improving its overall accuracy. While the current parameters, including the k-mer size and insert length, yielded a strong assembly, exploring a broader range of k-mer sizes and insert lengths could further enhance assembly quality and optimize coverage uniformity.

For gene prediction, using a model trained on *Staphylococcus aureus* falls short of capturing the unique genetic characteristics of ***Staphylococcus simulans***. To push the boundaries of accuracy, a custom gene prediction model specifically trained for the client's protein is necessary. This investment in a species-specific model will dramatically enhance prediction accuracy, facilitating the identification of unique species-specific genes while minimizing the risk of incorrectly annotating non-functional regions. This recommendation will ensure future gene predictions are exemplary.

Additionally, for a more holistic understanding of evolutionary relationships, a multilocus or whole-genome phylogenetic analysis should be conducted instead of relying on a single gene, like 16S rRNA or *pykA*. Each gene in an organism's genome is subject to different evolutionary pressures, selection constraints, and horizontal gene transfer events. By analyzing multiple loci or the entire genome, a more accurate picture of the evolutionary history can be developed, providing deeper insights into the genetic diversity and relationships among the taxa under study. This recommendation will not just trace a singular lineage; it has the potential to map the entire evolutionary landscape, enabling future researchers to decipher complex genomic narratives and answer questions that single-gene analysis simply cannot. By incorporating these cutting-edge recommendations, it is ensured that future research will set new standards in genome assembly and evolutionary analysis.

Introduction
Hello everyone, today I'll be discussing de novo genome assembly, species identification, and phylogenetic analysis for a bacterial genome using paired-end Illumina reads. The goal is to assemble the genome, identify the species, and analyze its evolutionary relationships using both the 16S ribosomal RNA gene and a pykA housekeeping gene for comparison. These findings will help place the bacterium in context with related species.

Methodology and Techniques Used
I downloaded the paired-end sequencing data from the client and performed a quality check using FastQC to ensure the data was suitable for assembly. For de novo genome assembly, I used Velvet, setting a k-mer size of 69, an insert length of 353, and automatic coverage calculation to maximize assembly accuracy. For species identification, I used BLAST to search for homologous sequences against a reference database, particularly focusing on identifying the 16S ribosomal RNA sequence to pinpoint the species. After extracting the 16S rRNA sequence, I compared it with known sequences to build a phylogenetic tree using MUSCLE and MrBayes. Finally, I selected a housekeeping gene, used OrthoFinder to extract its sequence, aligned it with known sequences, and inferred a second phylogenetic tree to compare with the 16S rRNA tree.

Key Finding 1 - Genome Assembly
I successfully assembled the genome de novo, achieving an N50 scaffold length of around 58 KB, with a maximum scaffold length of 183 KB. The GC content was 35.81%, typical for many bacterial genomes. Over 61% of the genome was contained in scaffolds larger than 50 KB, indicating a high degree of genome completeness. This assembly provided a solid foundation for downstream analysis, particularly for gene prediction and phylogenetic work.

Key Finding 2 - Gene Prediction
For gene prediction, I focused on protein sequences to narrow this down to 2,571 key genes. The longest gene NODE92 is analyzed using BLASTp and found to encode a G5 domain-containing protein that is associated with bacterial cell wall maintenance, adhesion, and carbohydrate binding for the ***staphylococcus simulans*** organism.

Key Finding 3 - Phylogenomic Analysis:
The phylogenetic tree constructed identified *M_caseoly* and *S_sciuri* as the outgroup, consistent across both the 16S ribosomal RNA and pykA protein trees. However, I observed differences in the placement of other taxa, specifically with *S_aureus* grouping differently in the pykA tree compared to the 16S tree. Also, the evolutionaly distances are longer in the pykA tree. This finding suggests evolutionary pressures affect the protein-coding genes more than the conserved 16S rRNA gene.

Conclusions
In conclusion, I successfully assembled the client's genome using short-read sequencing data, achieving a high-quality assembly with 61% of the genome in scaffolds larger than 50 KB. Through gene prediction, I identified key functional genes, including the longest gene encoding a ***staphylococcus simulan*** protein. Finally, my phylogenomic analysis revealed the evolutionary relationships between the protein I studied and related proteins from other species.

Limitations and Recommendations
In this study, the limitations include using short-read sequencing technology. Although an average contig coverage of 57%, this uneven coverage across different genomic regions poses a challenge, as areas with low coverage can lead to incomplete genome assembly. Additionally, the fixed velvet parameters are not ideal for all genomic regions because of the risk of misassembly if the estimation is inaccurate. The automated expected coverage in velvetg can struggle to distinguish true genomic regions from sequencing errors. For gene prediction, the GeneMark model was designed for Staphylococcus aureus. This approach can overlook species-specific genes in the client's strain. Similarly, identifying the 16S ribosomal RNA sequence was based purely on sequence similarity, which is effective for known homologs but may fail to capture strain-specific mutations. Additionally, using single-gene phylogenetic analysis, with 16S and pykA, limits the scope of evolutionary inference, as it doesn't fully capture the complexities of gene duplication or horizontal gene transfer events.To improve future research, I highly recommend including long-read sequencing technologies like Oxford Nanopore to overcome the limitations of short reads. For gene prediction, training a custom model would enhance prediction accuracy by identifying unique species-specific genes and avoiding incorrect annotations. Finally, for phylogenomic analysis, a whole-genome approach would provide a comprehensive understanding of evolutionary relationships. By analyzing multiple genes or the entire genome, we can map the full evolutionary landscape, offering deeper insights into the genetic history of the taxa under study.