Higher grade questions
Genome Analysis 1MB462
Agapi Eleni Simaiaki

Reads quality control

1. What is the structure of a FASTQ file? How is the quality of the data stored in the FASTQ files and how are paired reads identified?

A FASTQ file contains four lines per sequence:

i. Begins with @ and a sequence identifier. A short description might be present as well.
ii. Raw sequence.
iii. Begins with + and might contain a seq identifier and an additional description.
iv. Quality values represented by an ASCII character - one per each nucleotide of the sequence Increasing order of quality control ASCII characters: !"#$%&'()*+,-./0123456789:;<=>? @ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Paired reads are identified with utilizing the number of the paired-end sequence file. The paired-end files typically have the same name and sequence identifier but differ on the additional number identifying them. That number is typically 1 or 2. In our case for example the paired-end reads were named E745-1.L500_SZAXPI015146-56_1_clean.fq.gz and E745-1.L500_SZAXPI015146-56_2_clean.fq.gz.

2. What is read preprocessing?

Typically, raw sequencing data may contain errors or low-quality reads. Hence, it is critical to filter out low-quality reads, sequencing adapters, and correct any sequencing errors to ensure that the assembly is of a reliable quality

3. What parameters are of interest when looking into read quality?

After running fastq for all the reads an .html file for each is created containing important information and statistics. Some of the parameters of interest found in the html files are:

- Basic statistics: Contain an overview with the total number of reads, their length and the sequences flagged as poor quality.
- Per base sequence quality: Shows the error likelihood at each position averaged over all measurements. Three colored bands are observed which characterize the observed score based on the quality score:
  - reliable (28-40, green)
  - less reliable (20-28, yellow)
  - and error prone (1-20, red)
- Per sequence GC content: Shows the GC content across each sequence and compares it to the theoretical normal distribution created by the input reads
- Per base N content: Shows if unknown bases are present in the sequencing reads
- Sequence Length Distribution: Length distribution of the reads
- Sequence Duplication Levels: Shows duplication levels, it is a good estimate if enrichment bias exists

- Overrepresented sequences: Shows if some sequences are overrepresented
- Adapter content: Shows if adapters are present in the sequences. If that the case they need to be removed/trimmed.

4. How is the quality of your data?

The provided data already have been preprocessed as a result the basic statistics described above seem quite good.

5. How important is read preprocessing for downstream analyses and why?

Preprocessing is a very important step as it ensure that the quality of the reads is good for downstream analysis. If that not the case biased results might occur leading to false assemblies, mappings or any other downstream analysis is performed.

6. What can generate the "fails" in FastQC that you observe in your data? Can these cause any problems during subsequent analyses?

As the reads were provided to us already trimmed, no "fails" were observed.

Reads preprocessing

7. How many reads have been discarded after trimming?

None, as the trimming took place on already trimmed data and minor changes were applied.

8. How can this affect your future analyses and results?

Removing a high amount of reads can lead to lower depth dequencing and hence lower significance or even bias in the output.

9. How is the quality of your data after trimming?

As the reads were already trimmed the additional trimming did not really add positively to their quality, on the contrary it affected a bit negatively the Sequence Length Distribution leading to a variety of 51-91 long reads for the E745-1.L500_1_paired.fq.gz compared to 90 before trimming and 35-90 for E745-1.L500_2_paired.fq.gz compared to 90 that was before.

10. What quality threshold did you choose for the leading/trailing/sliding window parameters, and why?

- Sets a minimum quality score from the beginning of the sequence required to keep a base at 20 to ensure good quality of reads-green band quality score (LEADING:20)
- Specifies the minimum quality score from the end of the sequence required to keep a base at 20 to ensure good quality of reads-green band quality score (TRAILING:20)

- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (Default: SLIDINGWINDOW:4:15). The smaller the window the more accurate the outcome especially in short reads and the more computational expensive the process is .

Genome and Metagenome Assembly
The questions here are answered based on the basic analysis and Flye assembler.

11. What does it mean to assemble genomic reads and what should the output be?

Assembling genomic reads means aligning and merging short sequences that have been extracted by the genomes of a number of individuals to reconstruct the original sequence. The output is ideally one continuous sequences with all the genomic information of the species or fewer longer species that contain the total of the original genome information.

12. What information can you get from the plots and reports given by the assembler (if you get any)?

Flye gives three main output report files, two coning information on the graphs used for the creation of the assembly assembly_graph.gfa and assembly_graph.gv and one text file containing information on the total number of contigs, their length, coverage, if they are circular and repetitive, multiplicity, alternative group and their graph path.

13. What intermediate steps generate informative output about the assembly?

Flye produces many intermediate binary files and additional information for each sub-step of the assembly process. More specifically:

- 00-assembly/ folder: Information on the process of the assembly containing draft_assembly.fasta and draft_assembly.fasta.fai files
- 10-consensus/ folder: Information on the consensus assembly produced from all provided reads
- 20-repeat/ folder: Information on the various repeat runs for the construction of the final graph
- 30-contigger/ folder: Contains information on the contigs created, their stats and the final graphs.
- 40-polishing/ folder: Contains corrected and final contigs after additional processing and filtering along with some final stats

14. How many contigs do you expect? How many do you obtain?

I expected around 5-6 given the length of the genome reference genome ~2,9 Mb and considering that plasmids are also present. With Flye I obtained 13.

15. What are “contigs”, “unitigs” and “scaffolds”?

- Contig: a set of DNA sequences that overlap (sequencing reads) and lead to a contiguous representation of a genomic region.
- Unitig: an assembly of fragments for which there are no competing choices in terms of internal overlaps. Simply they contain unique sequences.
- Scaffolds: is a series of contigs aligned in a relation to each other to bridge the gaps of contigs and provide a better base for the final assembly construction.

16. What are the k-mers? What are the problems and benefits of choosing a small or a large k-mer?

K-mers are short contiguous sequences of length k that are extracted from longer sequences of DNA (sequencing reads in our case).

The length of k-mers can affect the resulting assembly. Short k-mers can lead to increased sensitivity of the to repetitive regions but require higher computational power and might lead to more sequencing errors. On the contrary long k-mers are not that computational expensive and can lead to more accurate assemblies but might miss on the repetitive parts. The selection of the k-mer length is a trade-off computatntional expense, assembly accuracy and senesitivity of the repetitive regions.

17. Some assemblers can include a read-correction step before doing the assembly. What is this step doing?

Such step aims to detect errors in the reads such as sequences artifacts or PCR biases, correct for them and aims to improve the overall quality and prepare the reads for a more efficient assembly construction.

18. Do you expect the same result between different assemblers, for the same data?

All in all, the majority of the final assembly should be close enough between different assemblers as the same input is utilized. However, differences regarding minor deviations in the length of the contigs or even their number (especially in when referring to short ones) of them are expected to occur.

19. What does it mean that coverage of a genome is 98% at depth of 40X?

- Coverage: How much of the total sequence has been covered of reads. In our case 98% of the genome has been covered by reads.
- Depth: How many times a specific nucleotide has been covered by a read (has been sequenced). In our case the 98% of the genome has been sequenced at least 40x.

20. Would you expect to see other letters than A, T, G, C in the DNA sequence of an assembly? If so, what would they mean?

- N: This letter is typically seen when there is no information on the exact nucleotide in the position. In the assemblies is typically seen in the scaffolds as the exact length and sequences connecting the contigs are not known and are arbitrarily selected.
- Degenerate bases: Additional information that characterize the base in a specific position in a more general why, most likely because the sequencing was not accurate enough there. For example, R represents purines: A or G and Y represents pyrimidines T or C.

Assembly evaluation

21. What are the measures that we can use to summarize the quality of an assembly? Name and explain at least three.

Some of the main interest metrics regarding assembly evaluation are:

- the number of large contigs and their length,
- N50 (length of a contig, so that when all contigs sorted by that specific contigs all together cover at least 50% of the assembly),
- the number of various types of misassemblies,
- the genome fraction (percentage of assembled genome compared to a reference)

To obtain those metrics the QUAST software was utilzed. QUAST provides a wide range of metrics regarding bith the contigs produced and their comparison with the reference genome if provided. In our case the reference genome was provided for bestter estimation of the assembly.

22. How does your assembly compare with the reference assembly? What could have caused the differences?

The Flye (corrected with Pilon) assembly seems to be nicely compared with the reference genome. The genome fraction assembled reaches 85.182% with a total of 190 missassemblies. Few structural difference such as relocations (176), translocation (12) and inversion (2) are spotted as well as 458 indels. A total of 9739 bases were found to be mismatched. Such deviations are expected when considering that not all the bacteria, even though they come from the same strain of Enterococcus faecium of one patient are the same. They may differ in the their genomic and plasmid sequences due to genomic plasmicity due to various factors such as horizontal gene transfer, mutations, and mobile genetic elements or even plasmid within the same environment and host.

23. Why do you think your assembly is better/worse than the public one?

The aim of the present project is to delve into de novo genome assembly practices. The present analysis was designed for educational purposes therefore the outcome cannot easily surpass the outcome of the reference genome which has been produced by scientist specialized on the field. However, the produced assembly here seems to be very close to the reference one, which has been validated extensively.

Annotation

29. What is the difference between structural and functional annotation?

Structural annotation is the process during which the precise locations of various genomic regions are found and characterized. Such regions can be open reading frames, coding regions, exons, introns, repeats, regulatory regions, promoters, start and stop codons etc.

Functional annotation is the next step after structural annotation during which specific functions are assigned to the genomic regions found by the structural annotation. By assigning specific biological processes to the structural regions we can also assess the quality of annotation.

30. What types of features are detected by the software? Would you trust the prediction of some features over others and why?

For the present project the software Prokka for annotation of bacterial genomes was used. In the present project Prokka gave information on features such as genes, coding sequences, rRNAs, tRNAs, and tmRNAs (mRNAs). In general Prokka has a high trust level regarding the identification of well-

studied and characterized regions. However Prokka also predicts hypothetical features for novel or less characterized regions. In such cases further investigation through additional experimental or computational methods is in need.

31. How many features of each kind are detected in your contigs? Do you detect the same number of features as the authors? How do they differ?

Prokka output for the Flye Pilon assembly:

organism: Enterococcus faecium E745
contigs: 11
bases: 3190176
CDS: 3109
gene: 3198
rRNA: 18
tRNA: 70
tmRNA: 1

Prokka output with Spades assembly (finaly selected for downstream analysis):

organism: Enterococcus faecium E745
contigs: 34
bases: 3148669
CDS: 3068
gene: 3143
rRNA: 15
tRNA: 59
tmRNA: 1

Paper / NCBI RefSeq assembly: GCF_001750885.1 (Celera Assembler v. 8.1; SPAdes v. 3.0):
organism: Enterococcus faecium E745
contigs: 7
bases: 3168410
CDS: 3093
gene: 3182
rRNA: 18
tRNA: 70
tmRNA: 1

As we can see both assemblies, are quite similar regarding the features they contain. Flye Pilon assembly appears to be a bit closer to the number of features found in the assembly constructed in the paper but the differences all in all are not that extreme. Spade assembly appears to have a slightly fewer features in all categories except for the tmRNA that seems to be 1 for all assemblies. It seems that in general SPades assembly in stringer with the overall information obtained by Prokka when compared to the Flye Pilon and the paper results.

32. Why is it more difficult to do the functional annotation in eukaryotic genomes?

The eukaryotic genome is generally much larger than the prokaryotic and contains large amounts of non-coding DNA, including introns, regulatory elements, and repetitive sequences, which can complicate the identification of functional elements. Additionally, the eukaryotic genome is characterized by complex gene structure allowing for alternative splicing and the production of various protein isoforms, abundance in repetitve sequences including the transposable elements and complex regulatory elements such as promoters, enhancers, silencers, and insulators that can be found far away from the genes making prediction of the functional tricky to be predicted solely on the sequence data. Considering all the above the functional annotation of such genomes require complex algorithm that add to the overall challenges in a computational way and require further experimental validation and integration of additional data such as transcriptomics, proteomics and epigenomics to reach a significantly good annotation level.

33. How many genes are annotated as 'hypothetical protein'? Why is that so? How would you tackle that problem?

Out of 3143 genes found in the assembly, 1328 were characterized as hypothetical proteins. Those regions have no known function or homologs in existing protein databases but they meet the criteria to be considered as potential functional regions (such as an open reading frame and codon usage). Typically those regions lack experimental evidence or significant similarity to proteins with known functions. To make sure if those regions are indeed functional further investigation is in need. Some approaches could be comparative genomics analysis across related species or stains, integration of transcriptomics and proteomics data to identify possible expression of those regions, synteny analysis if those regions are closely located to well-known genes or conserved areas they might indicate some functionality and experimental validation to study expression, locaion and function of the hypothetical proteins.

34. How can you evaluate the quality of the obtained functional annotation?

As discussed above, tools like Prokka are mainly based on well characterized functional regions to predict similar on the studied assembly. To evaluate the produced evaluation approaches like experimental validation incorporating PCR and Sanger sequencing for example for selected genes or functianal analysis are useful. Furthermore, gene structure validation with RNA-seq data under various growth phases, and comparison with already existed annotation for the same species and identifying consistences can also help assess the quality of the annotation.

Mapping

39. What are SAM files and what information can we find in them? How are BAM files different from SAM files and why do we always want to have BAM instead of SAM files?

SAM files: Sequence Alignment/Map TAB-seperated files that contain text information (readable by a text viewer software) on the sequence alignment data to a reference genome. They contain information sunch as metadata about the sequencing experiment, reference genome, and alignment parameters and seperate information for each alignment such as the quary name, the alignment properties, the reference name and position, the alignmnet in a CIGAR (Concise Idiosyncratic Gapped Alignment Report string: a sequence of numbers and letters indicating continuities or discontinuities in the alignment) string form, the sequence read and quality score.

BAM files: compressed, binary format, of SAM files making them more efficient in terms of storage space and faster to read and manipulate. Typically they can include index files that allowe for faster and more efficient access to specific regions. They are also preferred due to the advantage they hold regarding the required storage size which is smaller that the SAM files.

40. What percentage of your reads map back to your contigs? Why do you think that is?

In all reads, the percentage of reads mapped back to the assembly's contigs was around 99%. That can be due to the fact that prokaryotic genome does not contain many repetitive or complex regions (at least compared to eukaryotic) to affect the alignment and high coverage length (quality) of the RNA-seq data. The increased percentage of mapped reads is also an indication that the assembly is of good quality so that RNA-seq data align to it.

41. What can cause mRNA reads not to map properly to genes in the chromosome? Do you expect this to differ between prokaryotic and eukaryotic projects?

Some of the reasons why in general RNA reads might not map properly to genes in chromosomes can be errors in sequencing, poor-quality reads and incomplete or low quality assembly. In our case we have tried to minimize as much as possible all the mentioned errors with evaluation in each steo as has been described above. Some other reasons why mapping of mRNAs might not be successful could be:

- the presence of chimeric reads that are typically formed by the artificial joining of two different sequences
- post-transcriptional modifications which can add confusion in mapping onto the genome
- repetitive regions and gene duplications that can cause ambiguity in read mapping

All those reason can affect both eukaryotes and prokaryotes; however, as eukaryotes have a more complex genome with higher number of repetitive regions, more complex regulatory regions, and RNA editing mechanisms and lower gene density the mapping can be a challenge. Additionally, eukaryotic genome contains introns and is subjected to alternative splicing that creates multiple mRNA isoforms for a single gene. This might lead splice variants read to not map properly on the genome.

42. Why can there be reads that don't align to genes? How does that relate to the type of sequencing data you are mapping?

Some of the reasons why there can be reads that do not align to gens ca be:

- Sequencing artifacts such as chimeric reads, contamination of previous samples or adapters (Illumina) do not align to the assembly (RNA-seq & WGS).
- Reads from repetitive region, transposable elements can lead to poor alignment (WGS mainly)
- Short reads with genomic variation such as indels and SNPs might not align properly (WGS mainly, non-coding or expressed regions typically)
- Non-coding regions not well-annotated they might align to the assembly but not correspond to any annotated region. That includes novel transcripts as well. (RNA-seq & WGS)

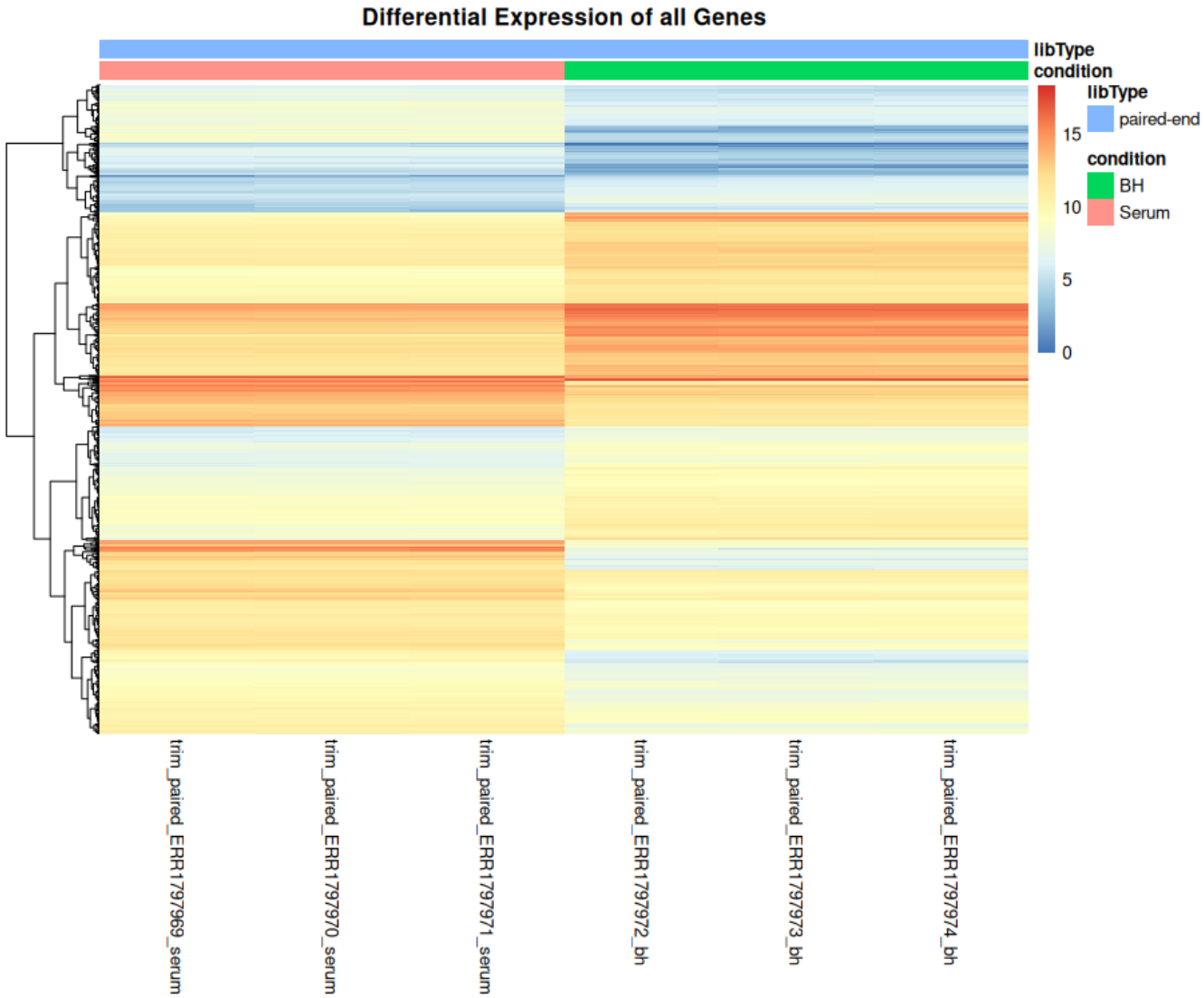
Long sequencing technologies such as PacBio and Nanopor can help over come the difficulties regarding mapping to low complexity regions such as repetitive and transposable elements as they allow for more information to be considered in the alignment procedure. That can however lead to the

exclusion of extensive information that might hinder the proper mapping of the reads. Short reads on the contrary might be more helpful regarding transcriptomics data allowing for an escape from possible gene editing variants due to their short length. Nevertheless, the short read technologies might be more prone to sequencing artifact due to the polymerisation step that they require. Addditionally, the short sequence might in some cases might not be significant enough to distinguish and map to the correct region allowing for missalignments. Finally, in all scenarios, the quality of the annotation and the assembly are of pivotal importance to ensure that all the aligned reads correspond to annotated genomic regions.

43. What do you interpret from your read coverage differences across the genome?

The difference in the read coverage across the genome signifies the the extent to which a genomic region is expressed as the reads come from transcriptomic data. RNA-seq reads coverage differencies can give insight to the expression levels of between different genes and help identify regulatory mechanisms when also focusing on non-coding regions.

44. Do you see big differences between replicates?



The heatmap above shows the normalized count reads between two different conditions, serum and BH, for all the genes found in the *Enterococcus faecium* E745. As we can see across the same condition, the replicates seem to have similar expression in the majority of the genes with some minor exceptions in lowly expressed genes (top blue part). The heatmap above is a qualitative approach to assess also the quality of the replicates. If deviations were present further investigation or even experimental replication and integration of more repeats would be in need.

(Optional for all papers, questions 45-48)

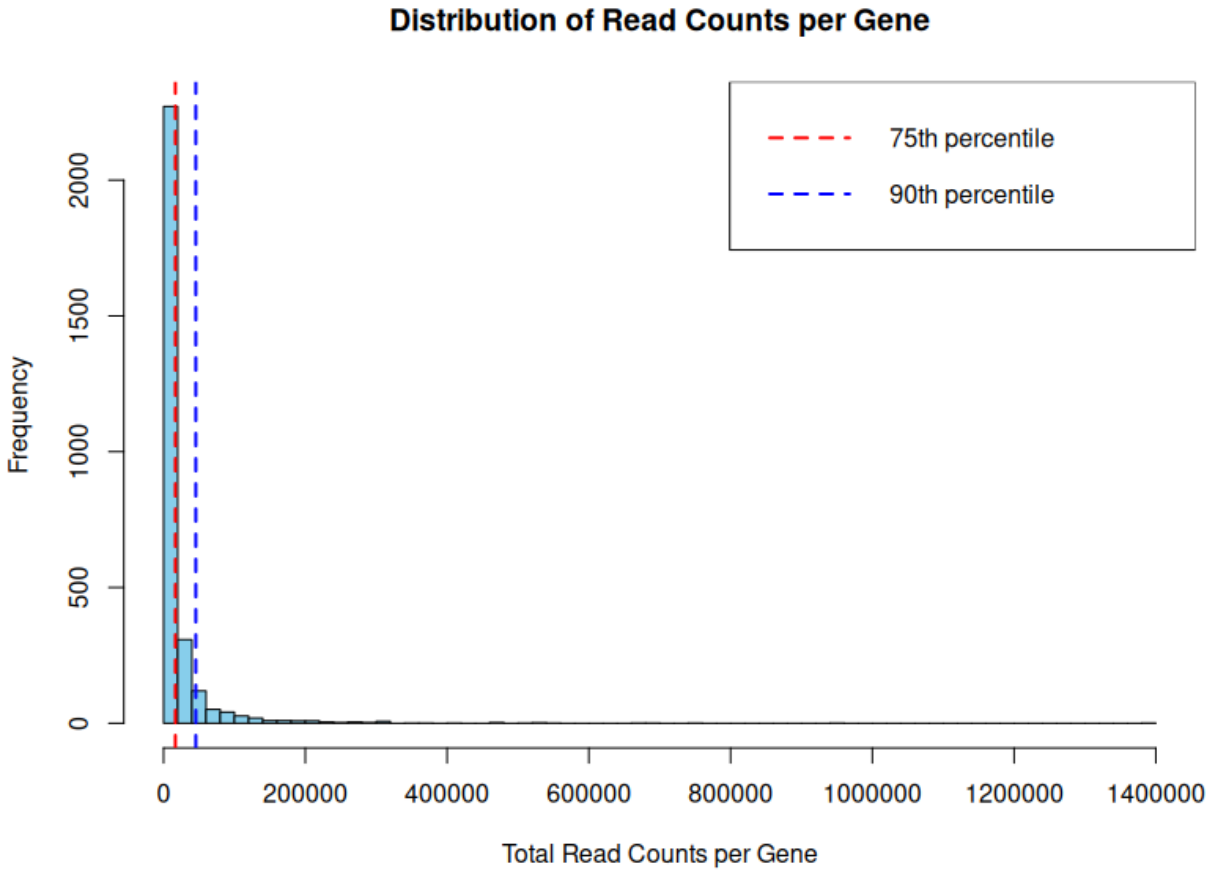
Post-mapping analyses

Variant calling analysis was not performed here.

45. What are VCF files and what information can we find in them? How are BCF files different from VCF files?
46. Why is filtering important for variant calling? What parameters can we use to filter out low quality variants and what are they?
47. How many SNPs and INDELs do you get?
a. What is the quality of those variants?
b. How are the variants distributed along the genome?
48. What is the difference between the variant quality, the mapping quality and the fastq quality?

Read counting

50. What is the distribution of the counts per gene (plot a histogram of read counts per gene)? Are most genes expressed? How many counts would indicate that a gene is expressed?



By fitting the counts in a negative binomial distribution that has been found to describe the best the over-dispersion in the RNA-seq data, we have:

- Median counts per gene: 4540.5
- 75th percentile threshold: 16628.75
- Number of genes above 75th percentile: 728
- 90th percentile threshold: 45509.6
- Number of genes above 90th percentile: 292

The percentile 75% and 90% are common in balancing sensitivity and specificity. Those percentiles help indicate genes that are expressed at a higher level relative to the total of the data without distinguishing between conditions. All reads per gene were summed and plotted in the form of a histogram where the percentiles mentioned were calculated. This approach is based on the fact that the highly expressed genes regardless the condition will surpass the lower expressed genes as a sum. Additionally, no normalization took place after first checking that are no high deviations between the replicates per condition (heatmap on deseq data of all genes without normalization or transformation – not shown here). This approach is a qualitative way to roughly estimate the variance of the reads between all genes.

Since RNA-seq provides information on which genes are expressed in the cells, compared to the genes found in the genome with annotation, the reads with counts>0 indicate that the genes are expressed. By considering the fact that some minor biological or technical error is present we can arbitrarily consider that genes with counts>10 are indeed expressed. Therefore out of 3067 coding regions found in the genome of the bacteria 2912 are found to be expressed (counts>10).

Expression analyses

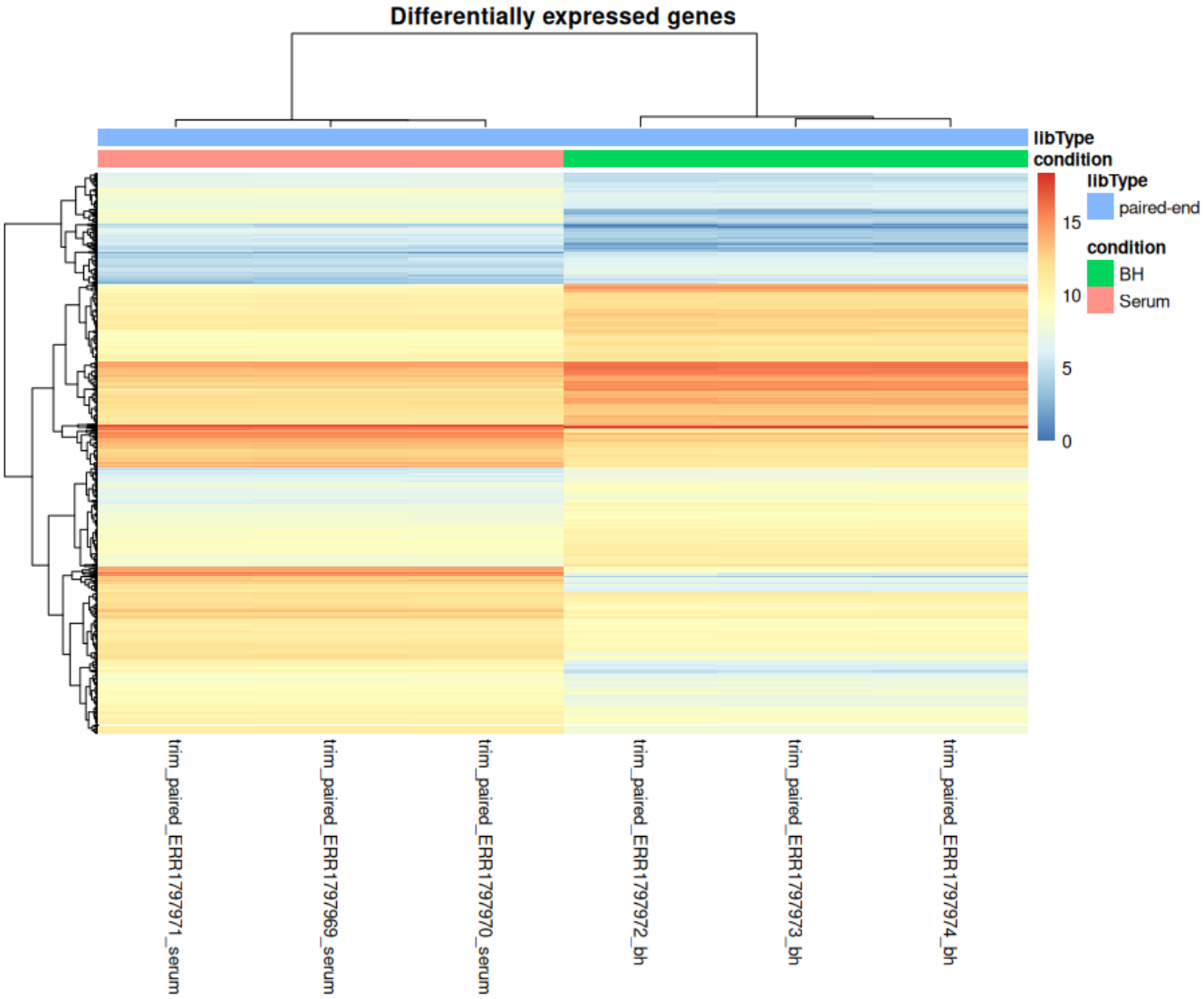
51. If your expression results differ from those in the published article, why could it be?

In the present analysis, the 20 top upregulated and 20 top downregulated genes were compared to the total of 860 identified as differentially expressed in the paper. Regarding the downregulated genes, signifying a decrease to their expression level in serum compared with the control condition (BH), 11 out of the top 20 (3 marked as hypothetical proteins) matched the ones found in the paper. Similarly with the upregulated ones (5 marked as hypothetical proteins). Deviations are expected since not the same data were utilized, in the paper all reads were employed while here only the Illumina and PacBio reads were used in the construction of the assembly, and different tools for the assembly Celera assembler (vs Flye corrected with Pilon and Spades here) and the Rockhopper (vs HTSeq and DESeq2 here) for the differential expression were utilized.

52. How do the different samples and replicates cluster together?

As we can see in the heatmap below, the different samples/genes (rows) are clustered based on the degree of change in their normalized count reads across the different conditions. More specifically, three big clusters are observed which contain genes with small (top part – shades of blue), high (middle part - shades of red) and medium (bottom part - light blue and yellow shades) expression difference. In all clusters an additional level of division can be observed where in each one, the top part indicates upregulation of genes in serum versus the control, and the bottom part indeicates

downregulation. Regarding the different replicates in the different conditions, a distinct clustering is apparent with all the replicates for the serum condition and control BH condition clustering into two groups.



53. What effect and implications does the p-value selection have on the expression results?

The p-value signifies the cut-off for statistical significance. Generally:

- Lower threshold allows for decrease in the number of false positive and increase in the number of false negative results.
- Higher threshold allows for increase in the number of false positive and decrease in the number of false negative results. That might lead to the identification of more genes as differentially expressed.

A trade-off between sensitivity and specificity is necessary to select a p-value that will maximize the detection of more true positives and avoid false positives. Finally as differential expression is simultaneously testing thousands of genes multiple testing correction is in need to control the false discovery rate.

54. What is the q-value and how does it differ from the p-value? Which one should you use to determine if the result is statistically significant?
Q-value is the adjusted p-value for multiple hypothesis testing which controls for the positive false discovery rate compared to p-value that represents the probability of obtaining a specific result or more extreme under the null hypothesis. Q-value basically represents the proportion of false discoveries expected among the tests considered significant. Q-value is preferred when multiple hypotheses take place in parallel, such as in our case with the differential expression.

55. How did you sort your differential expression results? Why?

Arbitrarily, I selected for the top 20 up and 20 downregulated genes. That took place by sorting the normalized count reads in increasing and decreasing manner based on their log2fold change values for all the replicates per condition. Log2fold cgange in the logarithm of the ratio of the expression level of a gene in one condition to its expression level in another condition. The base 2 allows for equally scaled differences. Hence the ordering based in the log2fold change allows tio identify the genes with the highest ratio of difference and therefore the most differentially expressed ones.

56. Do you need a normalization step? What would you normalize against? Does DESeq do it?

Normalization is an important step in differential expression analysis as it ensures that the expression levels are comparable across all samples and conditions. It allows to overpass differences in the depth of the sequencing between different samples, the amount of mapped reads to longer genes and any differences that might exist between the different samples and thei RNA-composition.

A simple normalization technique is to account for the total number of reads per sample. Other approaches can take into consideration the sequencing depth and gene length and other variations of those.

DESeq2 applies the median of rations methods which basically divides the counts with a sample-specific factor defined by the median ratio of gene counts relative to geometric mean per gene. The input of DESeq2 should be raw count reads and it internally applies filters for normalization of the counts.

57. What would you do to increase the statistical power of your expression analysis?

To increase the statistical power of our expression analysis some approaches would be:

- increase of the sample size could help identify true difference in gene expression by providing better estimates of variability and gene expression
- use good quality RNA-seq data, minimize sequencing artifacts and focus on biological variation
- utilize data with high sequencing depth for better estimates of the expression levels
- use suitable for RNA-seq data normalization and statistical methods like DESeq or edgeR
- optimize the experimental design / ensure that the samples of the studied conditions are properly selected regarding their collection conditions so that noise is minimized