EXAM Bioinformatics 2018

1. You are a system administrator. While the **who** command is useful to know who is logged in, it does not tell you *what* users are doing. You’ve recently discovered that the **w** command gives more information than **who. (2P)**

You execute the **w** command and the following gets printed to your screen:

13:35:25 up 62 days, 21:22, 5 users,  load average: 1.22, 1.11, 1.11

USER     TTY LOGIN@   IDLE JCPU WHAT

oskar    pts/0 4days    6:47m 6:47mRscript

juliette pts/6    3days 1:56m 1.88s  sshd: juliette [priv]

admin    pts/12 08:36    3:28m 0.13s sshd: admin [priv]

hadrien  pts/15 12:52    1:09 3:16 w

erik     pts/16 1:33     0.07s 0.07s bowtie2

Please write a small script that:

* Saves the output of **w** to a file, raw\_users.txt
* Processes raw\_users.txt to print back to your screen online to columns USER and WHAT
* *Hint:* ***tr -s ‘ ‘****will squeeze repeated blank space together*

Example good output:

USER     WHAT

Oskar    Rscript

Juliette sshd: juliette [priv]

Admin    sshd: admin [priv]

Hadrien  w

Erik     boowtie2

**Answer**: w > raw\_users.txt | sed 1d |

1. Your grant to sequence the alpaca genome got funded. You decide to create a project directory on your computer and to use version control in it. The directory will be called **agp** (3P)
2. Describe the steps and commands that you would use, from creating a directory to publishing it on github

**Answer**:

Assuming that one already has a github account:

1. Create agp directory in local computer: mkdir agp
2. Change directory to agp: cd agp
3. Initialize Git: git init
4. Check to make sure that git was initialized: ls -a
5. Create different directories according to need, for example, data, installations, results, etc directories
6. Once into a subdirectory also check to make sure that git was initialized: ls -a
7. You can now run operations such as creating files, downloading data, etc
8. After every activity in a directory/file, add the changes to git: git add name\_of\_file
9. Commit the changes to git: git commit - m "description of changes"
10. Check status: git status to see if all changes have been added and committed to Git
11. You have new file called **alpaca.sh**. Describe how you would make that file visible on the github repository created in a)

**Answer**:

1. Login to Git and create a repository called agp
2. Connect agp repositories on server and local computer. Make GitHub a remote repository for the local one: copy URL to the repository then run: git remote add origin https://github.com/username/agp.git
3. Check that the command has worked: git remote -v
4. Push changes in the local repository to online repository: git push origin master

c. While working on your **alpaca.sh** script, you realize you’ve made and committed a mistake in **alpaca.sh** at your last commit. How would you revert your changes?

**Answer**:

1. Use git log command to get a list of commits and their IDs
2. Run command: git revert [wrong\_commit\_ID]
3. Type in the new commit message.
4. Save and close
5. You have run the blast command on 8 samples. Now you have 8 blast tabulated output files (sample\_1.tab, sample\_2.tab, ...) from blast. (2P)

Now, you want to find, for each file, all the results corresponding to a virus hit, and redirect that in files called “sample\_1\_viruses.txt”,  “sample\_2\_viruses.txt”, ...

How would you proceed? (2P)

1. Describe the process of going from sample to metagenome to sequence classification using a read based classifiers. Include sequencing technology choice, initial informatics and analysis. Be sure to describe in great detail, including references, the different parts and why you chose to do as suggested. (2P)

**Answer**:

A metagenome refers to total genomic information within a sample. Genomic information in a metagenome may include genomes of different living organisms and viruses, mobile genetic elements, plasmids, free DNA and RNA. Genetic information of a sample (metagenome) is characterized by using sequencing technologies, such as high throughput and shotgun sequencing followed by assigning taxonomic information to sequence reads that are obtained.

The workflow may be as follows:

A. The first part (after sampling):

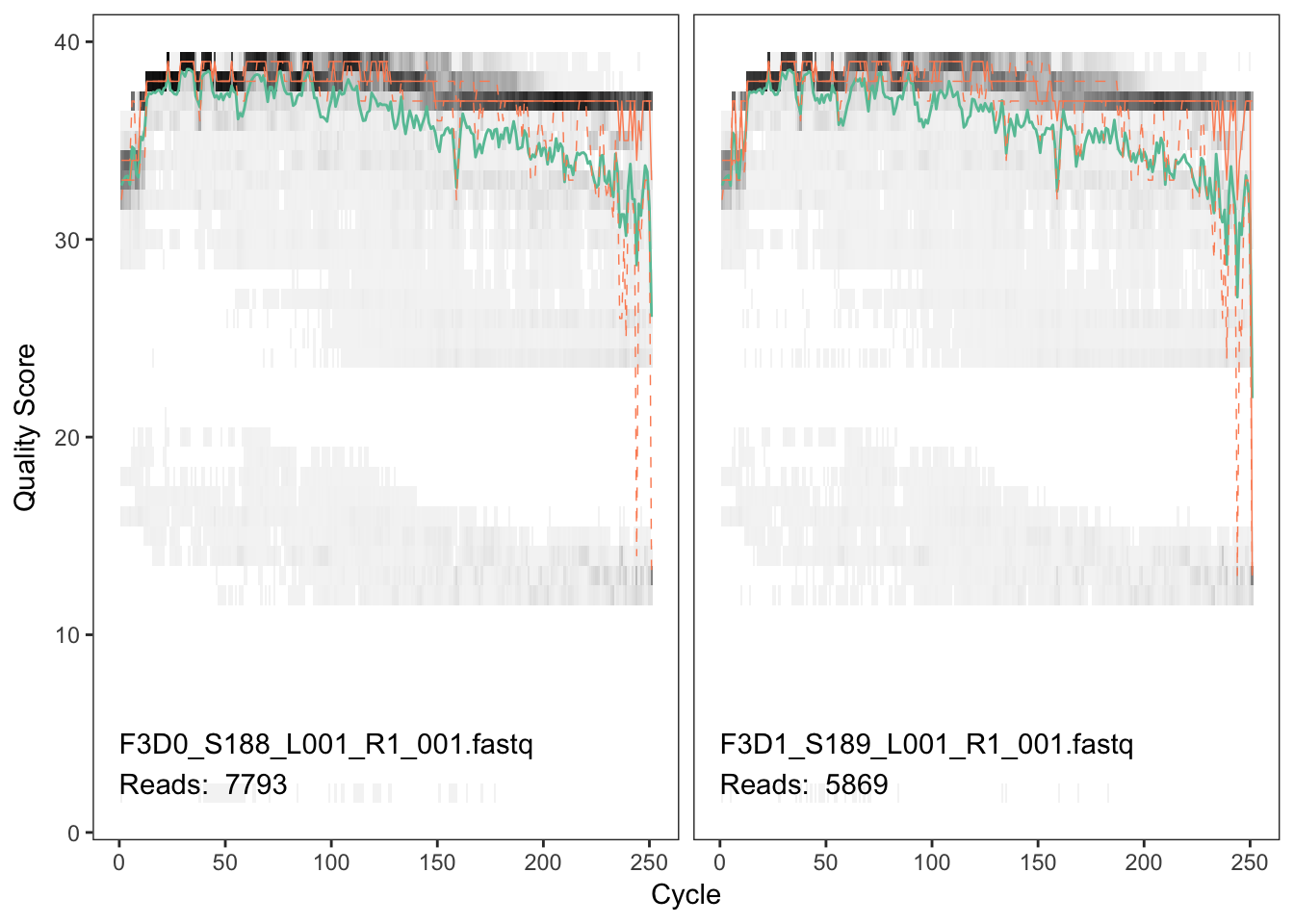
1. DNA extraction - total DNA extraction
2. DNA sequencing. Choice of sequencing platform depends on research question(s), resources available and other technicalities but the mostly nowadays Illumina and Ion torrent platforms are used for metagenomics studies.

B. Bioinformatics part:

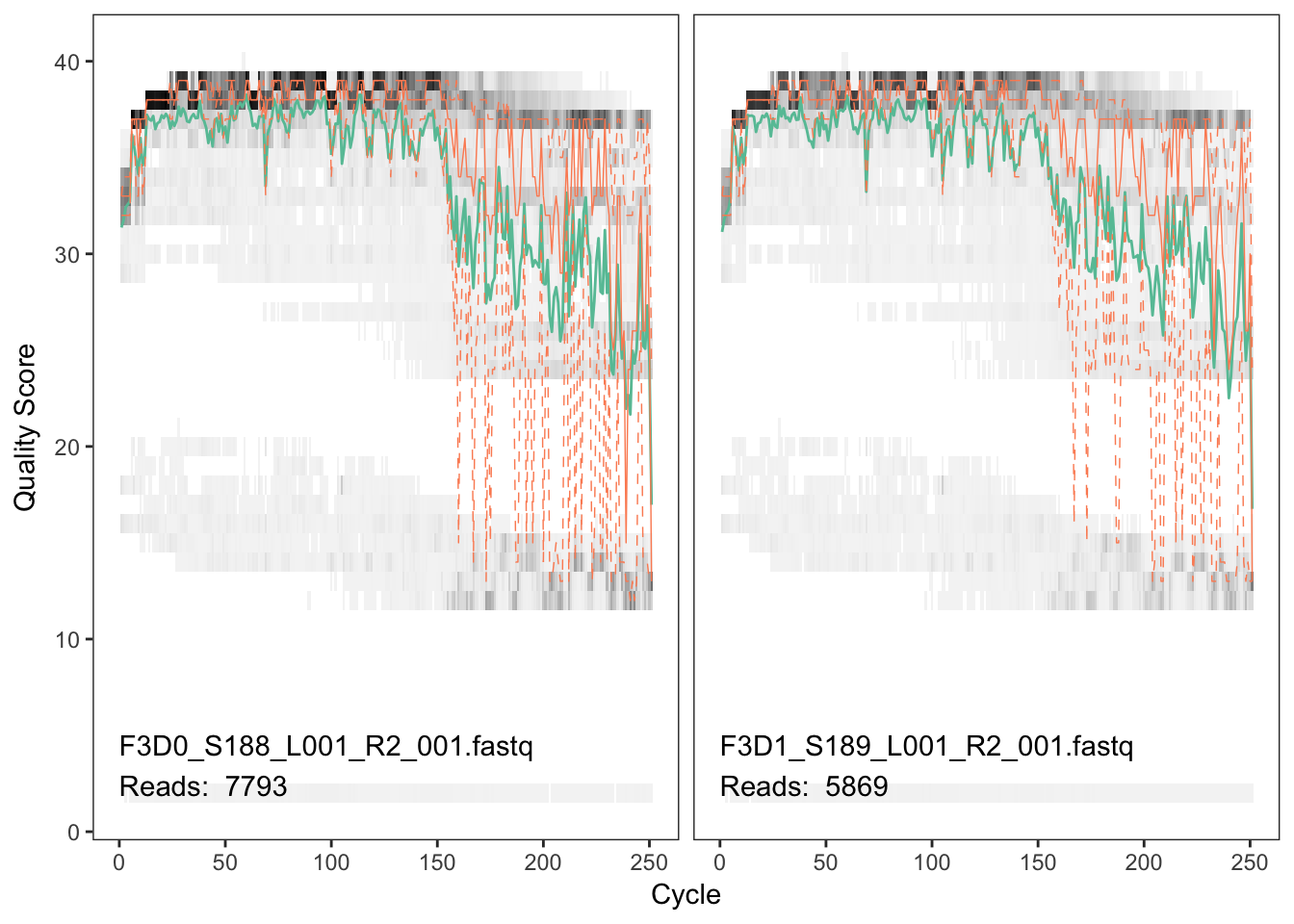
1. Sequence filtering (Quality control) is the first step after obtaining raw reads. Quality control is done to remove alien sequences, low complexity sequences and biologically irrelevant sequences depending on the research question(s). Look at the average read length and quality, remove poor quality regions and trim adapters if present. This step is of utmost importance because if the sequences are of low/poor quality, they may classify wrongly or produce chimeric reads. Even small errors can lead to reclassification. Rechecking and sanity checks are also very important. Scythe and Sickle programs may be used for quality control of the reads
2. Assembly of reads by using Megahit and binning to map the reads to the assemlby to get coverage information by using bowtie2 and metabat
3. Quality check of bins and visualization to see how many complete bins were obtained
4. Annotation (gene prediction, comparative genomics)
5. Statistical analysis

References:

1. Tully BJ, Graham DE and Heidelberg JF (2018) Data Descriptor: The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Science daya. 5: 170103
2. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholts P and Tyson GW (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology. 2:1533–1542
3. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H and Lam TW (2016). MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler driven by Advanced Methodologies and Community Practices. Methods.
4. After initial QC of 16S rRNA data youfind R1:



And R2



Suggest settings to use for the filtering step in DADA2 with motivation for the different settings. Is there another way to handle that data that would also work well with DADA2? (1P)

**Answer**:

R1 reads have a better quality while a large portion of R2 reads have a bad quality. R1 reads can be trimmed from position 240, R2 reads from position 160. Standard filtering parameters: maxN=0, truncQ=2, rm.phix=TRUE and maxEE=2 may be used. maxN=0 because DADA2 requires no Ns and the maxEE parameter sets the maximum number of “expected errors” allowed in a read.

Quality control may also be done with other tools like Scythe and Sickle then visualized with FASTQC or MULTIQC programs.

1. Analyze the following protein sequence. What is the function? Do you have an idea of the structure? Give the evidences that you find (patterns, profiles, structure predictions…) (3P)

>prot\_A

VHLTDAEKAAVSCLWGKVNSDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNAKVKAHGKKVITAFNDGLNHLDSLKGTFASLSELHCDKLHVDPENFRLLGNMIVIVLGHHLGKDFTPAAQAAFQKVVAGVATALAHKYH

**Answer**:

Prosite Proscan returned 1 hit corresponding to prot\_A. The protein has a globin family domain, which is 146 amino acids long with two metal (Iron) binding sites. Globins family proteins are heme-containing proteins involved in binding and/or transporting oxygen widely distributed in many organisms. These include Hemoglobins, Myoglobins and Leghemoglobins.

Upon querying Interpro with prot\_A sequence, the protein was observed to belong to a homologous superfamily of globins, had a globin domain and was shown to be involved with oxygen transport and iron ion binding.

BLAST against UniProtKB/SwissProt database at NCBI returned a top hit with accession number P02088.2 (E-value 3x106-105). It was a 147 amino acids long protein from a mouse named Hemoglobin subunit beta-1, Uniprot accession number P02088.

Blast search in UniProt database at Uniprot returned best hit protein Hemoglobin subunit beta-1 from the Mouse gene Hbb-b1 (PP02088). Its function - involved in oxygen transport from the lung to the various peripheral tissues. Secondary structure of the protein was shown to contain 9 helices and one turn. The tertiary structure was also viewed at PDB and was observed to contain alpha helices and turns.

Consensus secondary structure prediction at NPS showed presence of alpha helices, Extended strands and random coils.

1. First create a data frame / dataset as df1:

df1 <- data.frame(Chr=paste0('chr',1:9), Gene=paste0('gene',19:11))

Second createanother data frame/ dataset as df2:

df2 <- data.frame(Chr=paste0('chr',2:10),Position=paste0('pos',22:30))

Q1. Write a command to merge df1 and df2 into df3 by common column name “Chr” (1P)

**Answer**: data\_frame\_name <- merge(df1, df2, by="Chr")

Q2. If column names are different in 2 datasets? Write another command to merge df1 and df2 into df4, If the column name in df2 is "Chr\_ID" instead of "Chr". (2P)

**Answer**: data\_frame\_name <- merge(df1, df2, by.x="Chr", by.y="Chr\_ID")

1. The following GWAS study was submitted to a Journal. This GWAS aims at identifying significant associations with a new trait called the non-coagulation (NC) of milk in the Swedish Red cows. In this study, a total of 382 Swedish Red cows have been studied. The GWAS was done with the bovine High-Density (HD; N=777,000 SNP) and also with imputed sequences. Overall, one can conclude that a very high-density of marker was used. Since the authors focused on Chromosome 18 to explain their results, they called their study a region-wise association study (RWAS). There are no other differences between GWAS and RWAS. The provided Manhattan plot (Figure 1A) shows the results for the Bovine HD overlaid with the imputed sequences. The most significant association with NC milk is named TagSNP1. Although the study looks good, no information was provided about the genotypes.

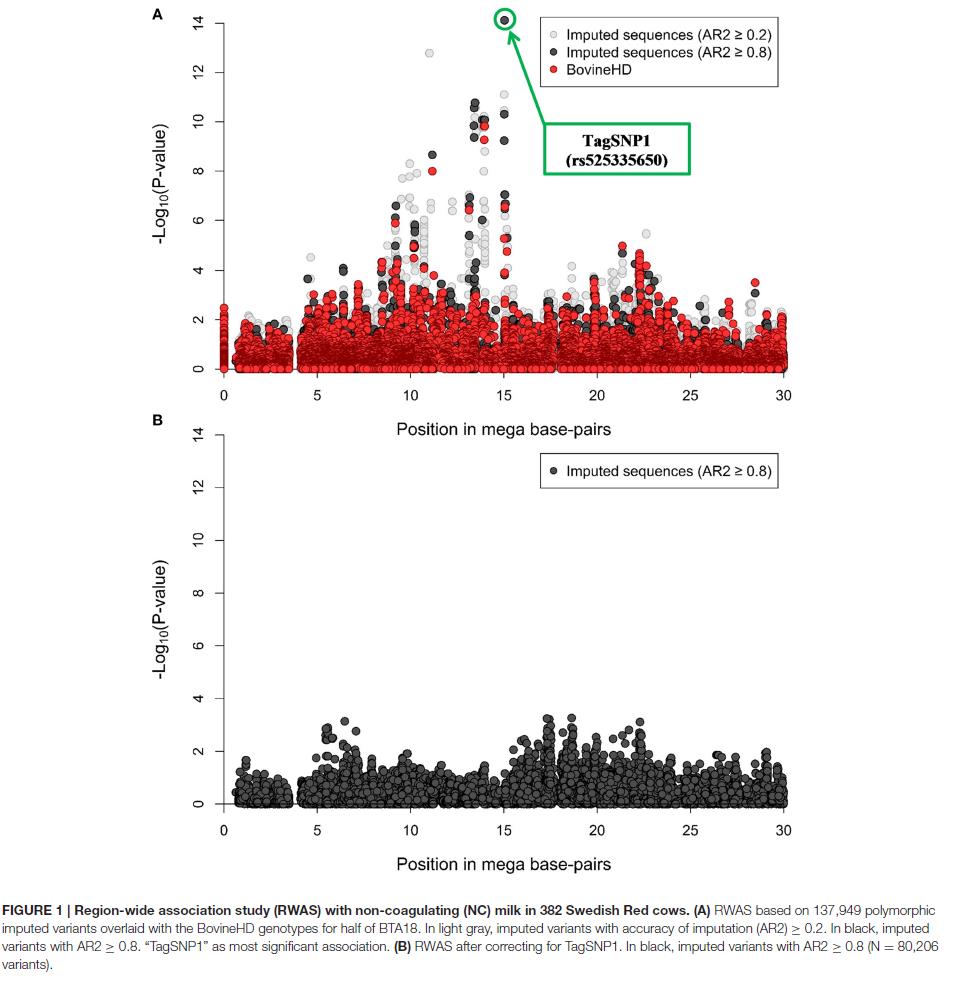
QuestionA: How can the reliability of SNPs be evaluated? Which steps are required? What is the consequence of not describing these steps? (2P)

**Answer**:

Reliability of the SNPs can be evaluated by:

1. Doing a quality control (QC) for call rate, minor allele frequency, HWE, inheritance
2. Using double controls on the platform
3. Call rate might be problem due to calling in the platform algorithms
4. Doing a validation study using data from another population
5. Make sure the p-values were corrected for false positives (e.g. permutation)

If the steps are not described it may lead to false positives.



Question B: In figure 1B, the authors state that the GWAS was corrected for TagSNP1, thus no genetic variation was left in this region. By looking at figure1A and then at figure 1B, what information was crucial in this study? Explain briefly what this information implements. (2P)

**Answer**: Population structure such as age of cows, batch effect and phenotype information is important for the study since it helps interpretation of results obtained

1. In the analysis of differentially expressed genes, the final outcome is, for each gene, an effect size and a p-value. Which one of the following statements is not correct regarding differential expression analysis of two samples, X and Y, using RNA-seq? (2P)

**Answer**: Statement A is not correct because RPKMs should be used to compare samples in the same library.

A-The p-value is the probability that the observed difference (or larger) in counts or RPKMs would occur by random chance for that gene

B- Multiple testing correction should be performed

C- For a gene with two isoforms A and B, it is possible that isoform A is more highly expressed than isoform B in sample X, but that isoform B is more highly expressed than isoform A in sample Y

D- The effect size is defined by the read length used in the RNA-seq

1. On a RNAseq analysis, what are the differences between RPKM, TPM and TMM normalization? (2P)

**Answer**:

In RNAseq analysis, normalization aims at correcting for read length differences, sequencing depth, batch effect and technical bias arinsing from sequencing instruments/technologies before comparing the data.

RPKM (Reads Per Kilobase Million) corrects for size and length of the data when comparing expression between different genes within the same sample. RPKM results into different total normalized counts per sample.

TPM (Transcript Per Million) normalization for sequencing depth and length but produces same total TPM-normalized counts per sample.

TMM (Trimmed mean of M values) corrects for size and RNA composition between samples but do not take gene length into account. TMM normalization is suitable for differential expression between sample groups for the same gene.

1. A salmon breeder is interested in genomic selection after hearing some seminars from SLU. He has a big breeding population where he has been selecting for growth. He is interested to start selecting for disease resistance.

From his earlier work and studying the literature he knows that growth has a heritability of about 0,40 but is controlled by many genes (at least 1000 QTL). Disease resistance has a much lower heritability (about 0,05) but this may be controlled by far fewer genes (max 20).

He wants to estimate what he can achieve with genomic selection in terms of accuracy.

He reads in the literature that the effective population size (Ne) of salmon is 2500. The salmon genome is about 28 Morgan

On the basis of the information above what accuracy would he achieve if he started genomic selection for respectively growth and disease resistance if he had a training population of 1000? How about 2000? (give a short explanation for your answers)

**Answer**:

Genomic selection accuracies for the two traits and training populations are as follows in Table 1 below:

Table 1: Genomic selection accuracy for growth and resistance to disease

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Specie | Genome size(cM) | Ne | H2 | QTL | Np | Me | r(accuracy) |
| Salmon | 28 | 2500 | 0.4 | 1000 | 1000 | 11162.00909 | 0.534522 |
| Salmon | 28 | 2500 | 0.4 | 1000 | 2000 | 11162.00909 | 0.666667 |
| Salmon | 28 | 2500 | 0.05 | 20 | 1000 | 11162.00909 | 0.845154 |
| Salmon | 28 | 2500 | 0.05 | 20 | 2000 | 11162.00909 | 0.912871 |

From the table above, for growth, he can achieve a higher genomic selection accuracy of about 0.7 with a training population of 2000 than 0.5 with a training population of 1000. For disease resistance, a higher genomic selection accuracy (0.9) is obtained with a training population of 2000 than a training population on 1000.

When he ponders a bit more he realises that he can only select in his own breeding program. That breeding program has had a stable effective population size (Ne) of 150. How does this affect his estimates under a): how much do they change?

Remember: Ne is a theoretical measure which does not relate to the actual number of animals. E.g. we have millions of Holstein cows but Ne > 100. This is because over many decades only a small proportion of animals contribute to the next generation. (3P)

**Answer**:

Table 2: Genomic selection accuracy for growth and disease resistance in salmon with two different Ne

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Specie** | **Genome size(cM)** | **Ne** | **H2** | **QTL** | **Np** | **Me** | **r(accuracy)** |
| Salmon | 28 | 2500 | 0.4 | 1000 | 1000 | 11162.00909 | 0.534522 |
| Salmon | 28 | 2500 | 0.4 | 1000 | 2000 | 11162.00909 | 0.666667 |
| Salmon | 28 | 2500 | 0.05 | 20 | 1000 | 11162.00909 | 0.845154 |
| Salmon | 28 | 2500 | 0.05 | 20 | 2000 | 11162.00909 | 0.912871 |
|  |  |  |  |  |  |  |  |
| Salmon | 28 | 150 | 0.4 | 1000 | 1000 | 863.3861819 | 0.562681 |
| Salmon | 28 | 150 | 0.4 | 1000 | 2000 | 863.3861819 | 0.693503 |
| Salmon | 28 | 150 | 0.05 | 20 | 1000 | 863.3861819 | 0.845154 |
| Salmon | 28 | 150 | 0.05 | 20 | 2000 | 863.3861819 | 0.912871 |

An effective population size of 150 does not do much to improve genomic selection accuracy or either traits.

Explain why genomic selection for disease is a better idea than genomic selection for growth?

**Answer**: Genomic selection for diseases is a better idea than selection for growth because:

1. The trait growth is controlled by many genes, it is hard to get a definitive marker for it and the linkage with the marker will become weaker after a few generations compared to disease resistance that is controlled by fewer genes
2. Genomic selection for growth has a lower accuracy (0.7) than that for disease resistance, even when using a training population of 2000. While for disease resistance, with a training population of 1000 a reasonable genomic selection accuracy is obtained (0.8).
3. You are analysing some *E.coli* Illumina data, that consist of two files: ecoli\_R1.fastq

and ecoli\_R2.fastq

a) Why are there two files? What do R1 and R2 represent? (1P)

**Answer**: There are two files because Illumina does paired end sequencing producing two files for each run. R1 and R2 represents forward and reverse reads, respectively.

b) You ran fastqc to check the quality of your sequencing data and obtain the following output files, in the qc directory: ecoli\_R1\_fastqc.html ecoli\_R1\_fastqc.zip ecoli\_R2\_fastqc.html ecoli\_R2\_fastqc.zip.

With which command would you extract the content of the .zip files? (1P)

**Answer**: tar -xvf \*fastqc.zip

c) In these .zip files are a text file named summary.txt , which look like this:

PASS Basic Statistics ecoli\_R1.fastq PASS Per base sequence quality ecoli\_R1.fastq PASS Per sequence quality scores ecoli\_R1.fastq FAIL Per base sequence content ecoli\_R1.fastq ...

How would you identify which statistics failed? (1P)

**Answer**: Failed statistics are preceded by the word FAIL

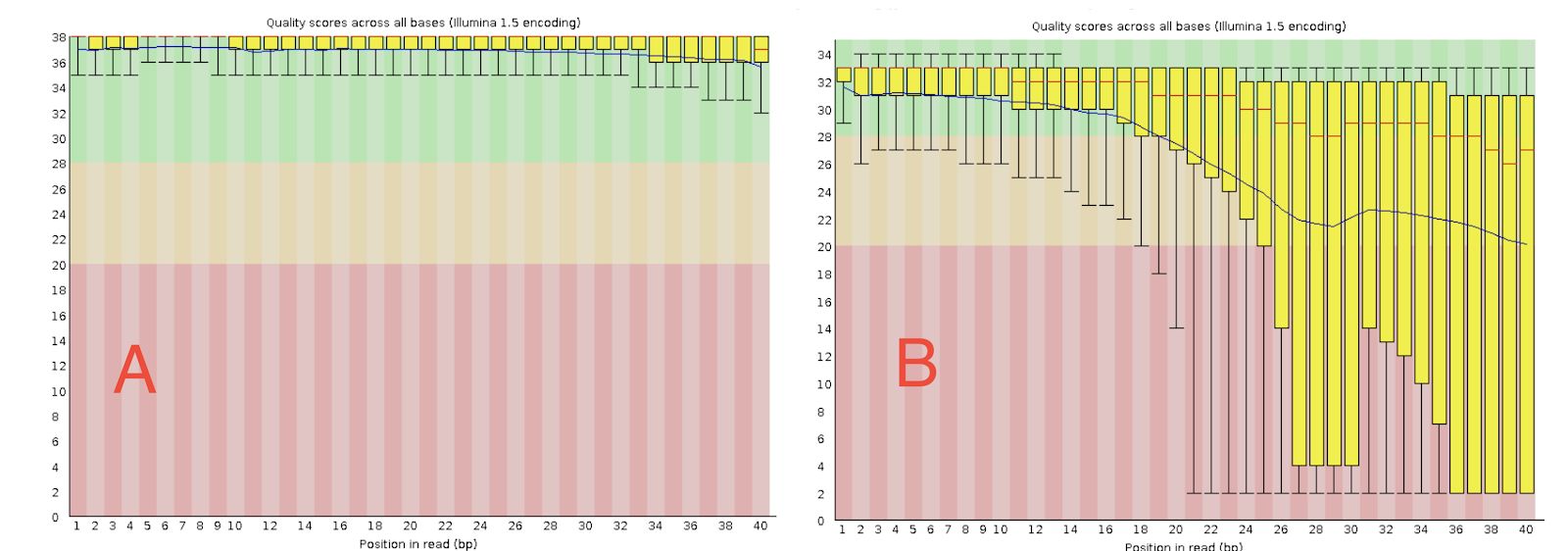
d) You know that you use TruSeq universal adapters and would like to remove them. What is the sequence of the TruSeqadapter?(1P)

**Answer:**

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'

e) In the image below, which fastqc file has the better quality?(1P)

**Answer**: File A has better quality than file B because its quality scores are high, ranging 36-38 while file B quality scores are as low as 2. A good quality FASTQ file file has a quality score of at least 20



1. Concerning Eukaryotic Annotation

A- Cite two different structural annotation approaches (1P)

B- Cite two methods to describe the function of an unknown protein sequence (1P)

**Answer**:

A

1. *Ab initio* prediction approaches - uses statistical models to predict structural annotation of a genome/sequence
2. Comparative homology approaches - uses other available information from related species/sequences for annotation

B

1. Sequence based methods that may rely on similarity and/or evolutionary relationships. Similarity based methods infer protein function by comparing its signatures with similar proteins with known functions, for example, PSI BLAST and HMM while evolutionary relationship based methods infer function of unknown protein based on phylogeny, synteny and clustering with other known proteins
2. Structure based methods compares unknown protein to known structures. Comparison may be done for whole/global structures, local structure or active sites.