CS31810, Computational Bioinformatics Assignment

Runar Reve (rur7)[1]

[1]**Aberystwyth University**

Corresponding author:
Author[1]

Email address: rur7@aber.ac.uk

ABSTRACT

Understanding the microbiome of the cow rumen can help us understand how it can breakdown and extract energy and nutrients from the plant material the cow consumes. By analysing 4 genomes that has been collected and assembled by a previous study on cow rumen Stewart et al. (2018), then compare it with different measures with other genomes, we can get a better insight what is in a rumen and how it is able to metabolise plant material. Here I suggest using a combination of GC-content, N50, and k-mers to extract metadata from genomes. Then by using these measures we can compare each genome visually or with correlation method such as Pearson correlation coefficient. All this is done in relation to an assignment for my module CS31810, Computational Bioinformatics.
The pipeline with all my algorithms for this analysis can be accessed in supplementary zip folder or not analysed version at GitHub: `https://github.com/RunarReve/AberCS31810Work`.

INTRODUCTION

With a better understanding of how animals, especially ruminants, can digest plant materials that often is hard to break down and into energy and nutrients. With this information we can improve our productivity and increase the amount of foodstuff we are able to produce with fewer resources. It is not only useful for the food industry, it can also be used to produce biofuel and in other biotechnology fields. Previous research have looked into the importance of the microbiome in the gut and its importance to its larger host with gut-brain interaction Martin et al. (2018). By understanding the microbiomes of other organisms, we may infer this knowledge to understand the human gut.

In this paper I will run various simple studies on 4 sequenced genomes that has been collected by (Stewart et al., 2018). In additional to these 4 rumen genomes I have selected 3 other genomes to compare against: Hay bacillus (AKA Bacillus Subtilis) (Borriss et al., 2018), Ecoli K12(Blattner et al., 1997), and SARS-Cov-2 (Wu et al., 2020). These are selected as they are common bacteria, expected to find in these types of data. The SARS-Cov-2 virus responsible for the current pandemic is used as a control as it is not expected to be significantly correlated with any of the data, most of the data is appear to be bacteria and not viruses due to their length of the genomes.

To automate and replicate my studies I have connected the programs I have created and used into a pipeline(mainPipe.sh). My pipeline with all my scripts can be accessed in the supplementary zipped (containing the complete study) file or cloned from Github(Pre-run study): `https://github.com/RunarReve/AberCS31810Work`. By using this pipeline it makes it easy to add additional genomes, by adding a new fasta file to the '/origin/' directory with the '*.fa' suffix.

MATERIALS AND METHODS

For my analysis I used in total 7 genomes, 4 incomplete genomes from cow rumen(RUG001, RUG233, RUG272, and RUG644) (Rumen Uncultured Genomes), and 3 complete genomes (Hay bacillus, Ecoli K12, and SARS-Cov-2) used to compare the 4 incomplete genomes against. I will mainly focus my analysis on the rumen genomes, as these are the focus of this study and thereby not pay close attention to the additional complete genomes and their relation between themselves.

GC-content

GC-content is a simple method to quickly insight on the structure of a sequence by retrieving the frequency of guanine (G) and cytosine (C) in the genome. This can be used as a simple comparison between genomes if they might be related. In my study, I retrieve GC-content for each scaffold and for the whole genome by summing over each scaffold. I will be focusing my attention on the whole genome, as this makes it easier to compare with other genomes. But the information on each scaffold is made available with the pipeline and might be an interest for future research.

N50

Checking the N50 is a good method of examining the assembly quality of each sequence, similar to finding the median length for each set of scaffolds. To get a better understanding on the assembly quality in the genome, I have added various versions of N50 (such as N25 and N90) to observe the changes in the scaffolds. For my additional sequences this will not be useful, as they are complete genomes will always return their one complete scaffold.

K-mers

To get a more complete insights into each genome to better compare against other genomes, I used k-mers. To get sufficient information I ran K-mers for all k between 1 to 4. Larger k will slow down the pipeline notably and with some experimentation does not provide any significant new information compared to the time it takes. But with these first k-mers we can filter out uninteresting sequences and focus on a select few (e.g. two sequences that seems similar) and run larger k-mers for more precise comparison. I got the k-mers for each scaffolds, but for ease of comparing I am focusing on the summed output over all scaffolds and normalizing the values to the frequency they are observed. Further investigation could be to look deeper into each scaffold and/or filtering out scaffolds that sticks out. Comparing the k-mers between genomes can be done visually with the raw data or plots of the data, but this gets unreliable and tedious with larger k.

Pearson correlation

With Pearson correlation coefficient I can compare different genomes with same K-mers studies and get a numeric score on the correlation between them. This allows us with a more reproducible and scalable comparison between sets of data, compared to a visual comparisons. I decided to use Pearson in stead of the similar Spearman's rank correlation coefficient, as my output data already has been normalized to the frequency of each K-mer.

Lower K might output high correlation based on there being less variability in the data to compare. That is why it might be advisable to ignore k=1 for a more reliable correlation.

RESULTS

Getting to know the data

First thing in any study is to analyse basic statistics of the data, and that is what I first did to get to know mine, see table 1. By counting the number of scaffolds we understand how scattered each genome are. The length of each RUG genome are relatively similar to each other, but not to any of the complete genomes. This does not mean we can conclude that these are not comparable, as the RUG genomes are not complete and there are unknown parts that is needed to use the length for a reliably comparison. The shortest scaffold are all similar while the longest scaffold reflects the number of scaffolds: the fewer scaffolds the more complete the scaffold are. By using these few factors we can conclude that RUG644 are the most complete of the rumen genomes while RUG272 are the least complete.

With the GC-content we can identify and comparing individual genomes against each other. As statistically the scattering of SNPs and the various scaffolds covering different parts of the genome will normally not be significantly different (of course the more complete genome are more likely to be the correct GC-content). With this we see that the only RUG644 and Ecoli K12 are similar, while the rest have largely different GC-content. This might be an indicator that RUG644 are a, or related to, Ecoli bacteria and could be further investigated.

Genome	N Scaffolds	Total length	Shortest	Longest	GC-Content
RUG001	144	2,125,227	2010	100,711	68.89%
RUG233	160	2,452,148	2197	82,421	53.04%
RUG272	268	2,260,494	2003	46,014	47.08%
RUG644	54	2,118,824	2106	184,293	50.43%
Ecoli K12	1	4,641,653	NA	NA	50.79%
Hay bacillus	1	4,215,607	NA	NA	43.51%
SARS-Cov-2	1	29,904	NA	NA	37.97%

Table 1. General information about each sequences (NA: No difference from total length)

Assembly Quality

To continue analysing and checking the Assembly quality of the data I used N50 with other check points to see how the length of scaffolds changes. Se table 2 for the data. As we discussed in the previous section we can better see the quality of each genome. RUG644 has the highest quality, as the scaffold length does not decrease as quickly for each point compared to the the others. While RUG272 loses significant length of scaffolds for each point.

Genome	N10	N25	N50	N75	N90
RUG001	66kB	44kb	24kb	13kb	7kb
RUG233	56kB	34kb	24kb	14kb	7kb
RUG272	33kB	23kb	13kb	6kb	3kb
RUG644	122kB	112kb	60kb	40kb	24kb

Table 2. N50 and other Nx to check the quality of each sequence. It is rounded to the nearest significant value.

Comparing Genomes

To compare the genomes I extracted different k-mers (1 to 4). This makes it possible to compare genomes of various completeness by comparing the frequency of each set of SNPs. Bar plots of 2-mers can be seen in figure 1, bar plots for all k-mers can be found in the supplementary materials for each study. Then I comparing a set of k-mers of the same k, either visually or with with a correlation algorithm such as Pearson correlation. The top correlating genome for each RUG genome calculated with Pearson can be seen in table 3, full list can be found in supplementary document ”pearsonCorr.tsv”.

The higher the k the more reliable this method will be, as can be observed in in row 2 and 4 with the p-value. As discussed previously RUG644 and Ecoli K12 have a high correlation and might be a further indicator that these genomes are related. By observing further into the relevant supplementary material and running with larger k we can see these genomes continues having correlation close to 0.6 with 5-mers.

Genome1_underGenome2;K	Pearson correlation	P-Value
RUG644_EcoliK12;K2	0.83	8.06e-05
RUG233_EcoliK12;K2	0.80	1.95e-4
RUG272_HayBacillus;K4	0.77	3.31e-51
RUG001_RUG233;K2	0.67	4.75e-3

Table 3. Top correlations for each RUG sequence, ignoring K1. Complete list in supplementary: pearsonCorr.tsv

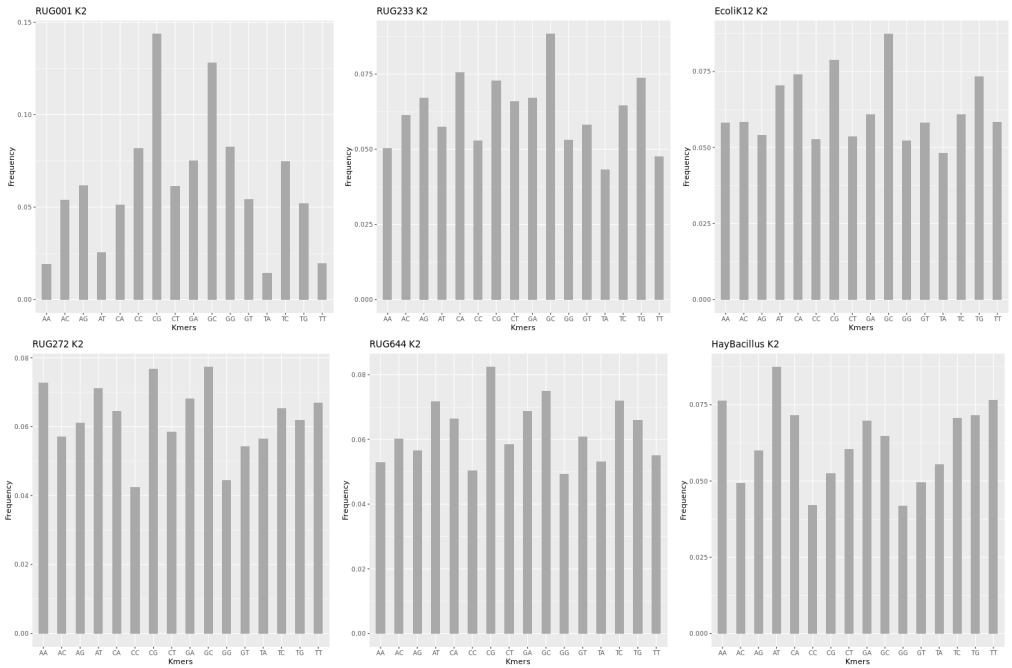

Figure 1. 2-mers (2-mers as it gives comprehensive figures) for 6 of our investigated genomes

DISCUSSION AND CONCLUSIONS

There is a lot of benefits of understanding the content of cow rumen, as cattle provides a source of food for large portion of all humans and better understanding could increase the overall food security for the world population. From the data I have analysed we get a small insight into this microbiome. The multiple methods used points to that Ecoli prevalent in the rumen from my small set of data, especially highly correlated with RUG644. This is not any new finding, as it is widely known that Ecoli is commonly found in the guts of vertebrates Tenaillon et al. (2010). Hay bacillus can be found in soil and therefor not surprising it has some correlation, most significantly with RUG272. As expected the added virus, SARS-Cov-2, did not get any significant correlations against the rumen genomes with k-mers (k¿=2), 0.23 with RUG272 (k=4) as seen in the relevant supplementary file.

This study does not fully determine that the correlating bacteria are the most related organism for each rumen genome, as I only compared against a small section of genomes. Further research is needed and could be to add more bacteria to the study, inspect what strain of Ecoli most likely to be prevalent, or extract ORFs for further inspections between genomes and/or other databases.

REFERENCES

Blattner, F. R., Plunkett, G., et al. (1997). The complete genome sequence of escherichia coli k-12. *science*, 277(5331):1453–1462.

Borriss, R., Danchin, A., et al. (2018). Bacillus subtilis, the model gram-positive bacterium: 20 years of annotation refinement. *Microbial biotechnology*, 11(1):3–17.

Martin, C. R., Osadchiy, V., et al. (2018). The brain-gut-microbiome axis. *Cellular and molecular gastroenterology and hepatology*, 6(2):133–148.

Stewart, R. D., Auffret, M. D., et al. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature communications*, 9(1):1–11.

Tenaillon, O., Skurnik, D., et al. (2010). The population genetics of commensal escherichia coli. *Nature Reviews Microbiology*, 8(3):207–217.

Wu, F., Zhao, S., et al. (2020). A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269.