# Appendix 6: Assessing Assemblies – Back-mapping Reads

As part of its ARG-detection/prediction procedure, the RGI MAIN software predicts open reading frames (ORFs) in the input sequences (in this context, the assembled contigs).

Back-mapping is usually performed using the *assembled contigs* as the references. In the context of our analyses, it was also useful to back-map the reads using only the *predicted ORF sequences* as references.

The proportion of reads which can be back-mapped to the sequences which arose from that collection of reads, indicates the quality of the assembly. Here we describe some detailed analysis of the back-mapping results, concerning both the mapping to contigs and mapping to ORFs.

The purpose is to aid evaluation of:

* intrinsic quality of the assemblies, i.e.
  + completeness
  + correctness
* the amount of contamination by the host (food-organism) DNA

Note that in principle the intrinsic quality could be perfect, but the assembled data might consist mostly of host DNA, thus most/all of the assembled contigs would be from that organism, not bacteria.

This is important because the **success of the host-read filtering is unknown**. If a low percentage of reads has been removed from the sample by the host-filtering procedure, that can be because the host-filtering was completely successful: there were few host-DNA reads in the first place. Alternatively, there may be a large proportion of host reads in the sample, but the filtering failed to detect them. This is more likely to occur when the reference genome sequence used is from a different species, or perhaps a different variety.

Therefore, we present data regarding the various relationships observed, for each sample, between:

* total number of reads prior to host DNA filtering
* total number of reads after host DNA filtering
* proportion of reads back-mapped to the assembled contigs
* proportion of reads back-mapped to the ORF sequences specifically, within those contigs

Two characteristics of an assembled metagenome that can lead to a poor back-mapping performance, i.e. a low percentage of reads mapped to the assembled sequences (whether for contigs or ORFs) are:

* incompleteness, causing fragmentation
  + i.e. the reads may have been assembled correctly as far as is possible, but a high proportion could not be incorporated into the final assemblies
  + some reads will be absent from any contigs (did not overlap with others) or could be assembled only into very short segments which are discarded by the assembler
* incorrectness
  + i.e. the reads have been assembled wrongly, which can lead to contig sequences to which not many reads (relatively speaking) align well

Note that here, 'completeness' ('read-completeness') refers to how completely the set of read pairs has been incorporated into the contig sequences (rather than how completely a series of genomes in a metagenome have been sampled).

Both of the above problems can occur even with a 'pure' sample of metagenomic sequence data (e.g. only bacterial DNA). Additionally, the issue of host DNA-contamination can have an effect on both. This is firstly in the sense of there being relatively little bacterial DNA remaining, if all host DNA is successfully detected and removed (leading to short fragments/incompleteness). Secondly, if a significant amount of host sequences remain, this may have an influence on incorrectness. However, even with high levels of host DNA in the sequence data, an assembly with high read-completeness and low incorrectness is possible (but many of the contigs will be non-bacterial).

**Figure 1** and **Figure 2** show the results of back-mapping to the contigs, and to the ORFs, for all samples.

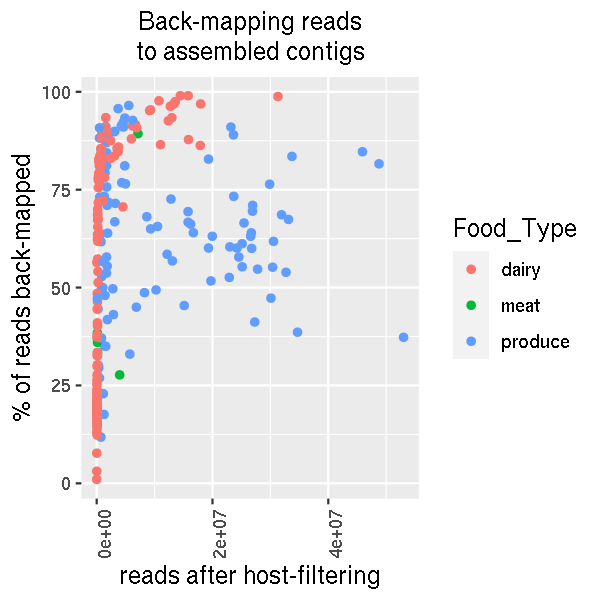


Figure 1 Number of post-QC, host-filtered reads mapping back to assembled contigs.

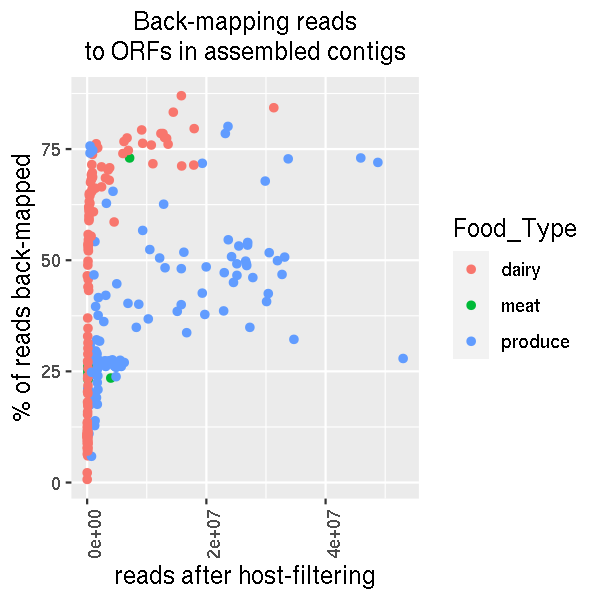


Figure 2 Number of post-QC, host-filtered reads mapping back to ORFs on assembled contigs.

Although it is clear that there is some kind of positive correlation between the total number of reads used for the assemblies and the % back-mapped, that is much clearer on the same two plots with the read count on a log scale (Figure 3 and Figure 4):

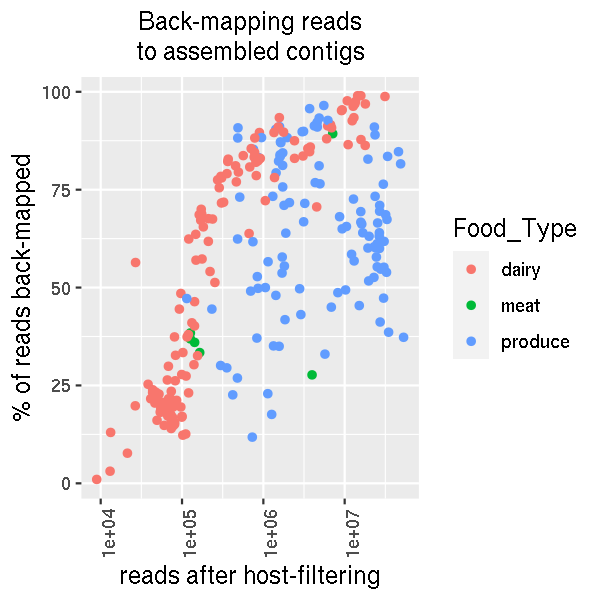


Figure 3 Number of post-QC, host-filtered reads mapping back to assembled contigs, plotted on a log scale

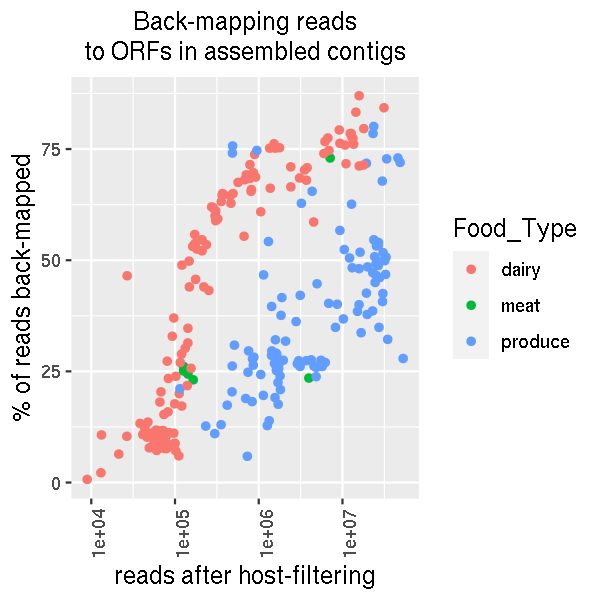


Figure 4 Number of post-QC, host-filtered reads mapping back to ORFs on assembled contigs.

This indicates that for many dairy samples, it is easier to obtain notionally better assemblies compared to many of the produce samples.

Also pertinent is the % of reads back-mapped to the ORFs as a function of the % back-mapped to the complete contigs, as follows.

Bacterial genomes have a high gene density, in contrast to eukaryote genomes (i.e. of all the host organisms of these samples) where large intergenic regions are generally present. Further, many eukaryote genes include introns, i.e. non-coding segments. For these purposes, the eukaryote gene coding segments (exons) can be considered as "ORFs".

Therefore, it could be hypothesised that in "pure" bacterial assemblies, where host-organism DNA has not been included in the sequencing data, a high proportion of the nucleotides in the contigs will be in ORFs (essentially the same proportion in any metagenome).Therefore the ORF back-mapping % in any such assembly would attain some theoretical maximum as assembly quality increases to notional "perfection". That theoretical maximum is biologically defined by the proportion of bacterial genomes' nucleotides which occur in ORFs.

In an assembled metagenome where a significant proportion of the contigs are of eukaryote origin, a significantly smaller percentage of the nucleotides are in coding sequences. Also, a high proportion of exons will probably not be predicted as ORFs, although the expectation is that many 5' exons might be, albeit this may be partially subject to the details of the gene-finding algorithm - which is designed to predict **prokaryote** ORFs (for our data, the Prodigal software performed this, via RGI MAIN).

Therefore, a collection (i.e. one sample) of "pure" bacterial contigs will have a very high, maximal back-mapping % **if** it suffers neither the incompleteness nor incorrectness problems. The contig-mapping would be ~ 100%, and the ORF-mapping would be some high value *k*% (biologically defined).

A collection of "pure" bacterial contigs which does suffer from incompleteness and/or incorrectness will be expected to have a much lower contig backmapping %; but the ORF-mapping % might be expected to also be ~ *k*% of the *contig-mapping* %.

The to-ORFs % back-mapping is shown as a function of the to-contigs % back-mapping below (Figure 5):

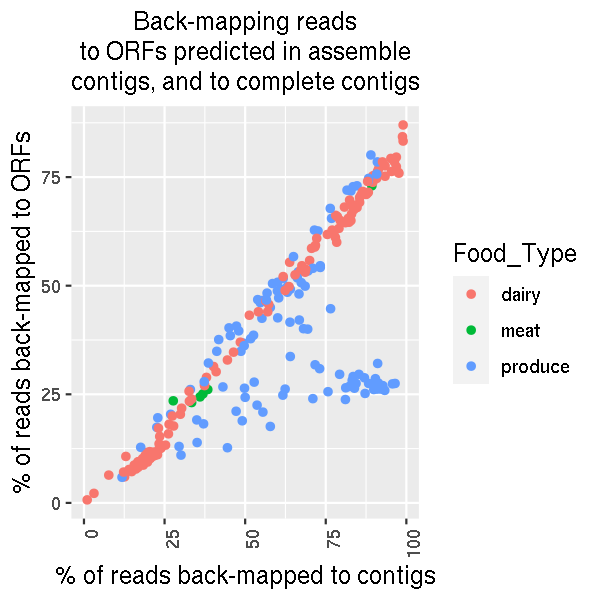
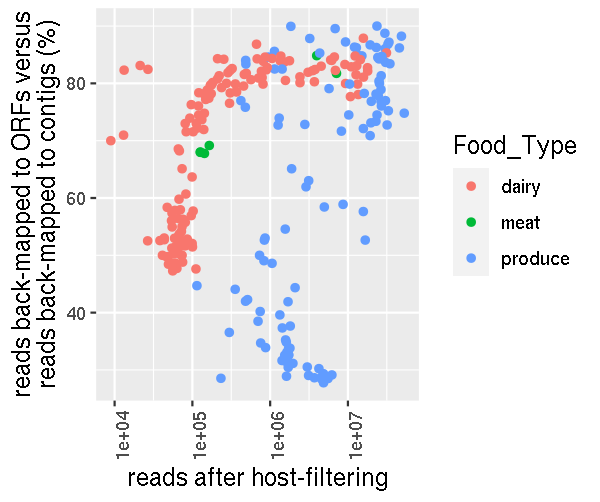


Figure 5 . Percentage of reads back-mapping to ORFs, as a function of percentage of reads back-mapping to contigs

The mostly linear relationship (but see below) is at least consistent with the above hypotheses. Therefore, **those samples where the to-ORFs backmapping is a much lower proportion than *k*% of the to-contigs backmapping can be hypothesised to be the samples where the host-filtering was least successful**, i.e. the assembly has high eukaryote-contamination.

It also indicates that *k%* may be in the region of 80-90% (i.e. gradient is 0.8 to 0.9). These considerations are a simplification, since they ignore non-host eukaryote DNA - e.g. microfungi that may occur on produce - but nonetheless serve as a useful indicator.

At a first glance, it appears that it is nearly all "produce" types which are thus affected, i.e. are considerably below the linear trend. However it is more complicated, because there are in fact a large number of dairy samples overlaid in the region *x* = 12.5 to 25, *y* = 7 to 12.5, i.e. the ORFs : contigs mapping ratio is about 50% there. This is much more visible in the plot below with the read count on a log scale Figure 6:

  
Figure 6 Back-mapping reads to ORFs predicted in assembled contigs, and to complete contigs

This does clarify that the poor-performing (in terms of the ORFs:contigs ratio) produce samples are in fact much worse than the poor-performing dairy, because those produce samples have relatively high read counts but the worst ORFs:contigs backmapping ratio. It also suggests that for dairy overall, there is a strong association between the number of reads and the ORFs:contigs backmapping ratio. If the hypothesis about the cause of this ratio being low is correct, then that suggests that lower read counts are associated with higher levels of eukaryote read contamination in dairy (but not in produce, where there is no such association) - the reasons for this are not obvious, and there are 5 dairy outliers which do not follow this, and have a high ORFs : contigs ratio despite extremely low read counts.

Note that in the above plot, the samples placed high on the *y-*axis do **not** necessarily have a high back-mapping %. They could have a very low back-mapping %, but the point is that the to-ORFs % is high relative to the to-contigs %.

To see which do have a high back-mapping %, the following plot may be more useful (  
Figure **7**):

  
Figure 7 Back-mapping reads to ORFs predicted in assembled contigs, and to complete contigs

In summary, the set of samples which can be hypothesised to be of high quality in terms of back-mapping to the contigs are those to the right of some arbitrary point on the *x*-axis, e.g. > 80%. Those samples fall neatly into two groups, those with high (> 75% of contig-mapping rate; these include dairy and produce) and low (< 40% of contig-mapping rate; produce only) to-ORF mapping ratios, which can be assumed to have relatively low and high host-read contamination, respectively.

The data for each sample can be found in the Appendix 6; those used in the above plots are 'reads\_post\_host\_filter' (the read count), 'backmap\_pc\_contigs' and 'backmap\_pc\_ORFs' (there is no column which explicitly contains the ORFs:Contigs mapping %).