## Appendix C2

### Supplement to Section 3.5.4.1 of What is the Burden of Antimicrobial Resistance Genes in Selected Ready-to-Eat Foods?

#### ARG predictions from RGI

We processed all 256 samples' quality-controlled, host-filtered unassembled short read sequence data with RGI BWT. We also processed (with RGI MAIN) all the samples' assembled metagenomes which had been derived from those.

For both the RGI BWT (short-read) and RGI MAIN (contigs) results, we also performed in-house filtering of the results.

As stated in Section 3.5.2, due to the nature of a significant proportion of the samples' assembled metagenomic data, we could not use the predictions from RGI MAIN as the production results.

##### RGI results in a selection of samples designated as high-quality assemblies

We could not use the RGI MAIN results (from assembled contigs) as the production data, due to issues with too many of the samples' assemblies (see main report Section 3.5.2). However, to help guide future work, it is useful to gain insight into the RGI MAIN results which arise from notionally "good" assemblies. This may indicate the worth of undertaking deeper sequencing for a large sample set.

Therefore we investigated the extent to which the RGI BWT results agree with the RGI MAIN results, in a small number of samples selected for their apparently high-quality assembly metrics.

These seven samples (three dairy, three produce, one meat) have among the highest percentage of reads back-mapped to the contigs (indeed the three dairy samples have the highest of all 256 samples). All seven also have a high percentage of reads back-mapped to the predicted ORF sequences, indicating a low or negligible level of host DNA-contamination (Table 1).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample ID | Food type | Food category | number of reads (post host-filtering) | contigs  back-mapped reads % | rank of 256 (contigs-mapping) | ORFs  back-mapped reads % | rank of 256 (ORFs-mapping) |
| 343587 | dairy | semi-skimmed milk | 15,803,729 | 99.0 | 1 | 87.0 | 1 |
| 343701 | dairy | probiotic yoghurt drink | 31,310,738 | 98.8 | 3 | 84.3 | 2 |
| 343624 | dairy | whole milk | 14,454,308 | 99.0 | 2 | 83.3 | 3 |
| 2672534 | produce | olives in brine | 23,626,929 | 89.0 | 35 | 80.1 | 4 |
| 2672718 | produce | red peppers | 23,182,503 | 91.0 | 25 | 78.5 | 8 |
| 343520 | produce | soya milk unsweetened | 487,646 | 90.8 | 27 | 75.7 | 18 |
| 2672664 | meat | ham not smoked | 7,125,112 | 89.3 | 34 | 73.0 | 27 |

Table 1. Seven samples selected for assessment of assembly and short-read derived ARG predictions.

The RGI BWT predictions from unassembled reads and RGI MAIN from assemblies for these seven samples are summarised in Table 2. All of the data for RGI BWT refers to the post-filtered results (see Methods). In contrast to the detailed, multi-stage filtering procedure we applied to the RGI BWT results, we consider that the filtering we applied to the RGI MAIN results is quite crude with a percentage-identity cutoff (90%) that may be too harsh for some genes. Therefore, we also show the pre-filtered gene-name counts as well as those post-filter.

For simplicity we have summarised the gene predictions in terms of unique ARG names. This is not the same as unique ARGs, because there is a 1:many relationship between an ARG name and the gene(s) of that name (in this context, strictly the genes refer to unique reference sequences). Therefore, even when gene names are in agreement, the predictions do not necessarily refer to the same gene (but often do). Predictions in good agreement gernerally would of course be expected to have good agreement at the ARG name level.

Clearly, the intersection between the two methods is remarkably small. For three of the samples, the number of RGI MAIN positive gene (names) is far smaller than the RGI BWT positives, and this is clearly not due to our own post-filtering. For the four samples where the numbers of unique gene names are similar, the intersection is very small, so these sets are of roughly the same size but mostly contain different names. One of the samples is completely ARG-negative by the assembly analysis (but has 2 ARG names pre-filter), but positive for 74 different ARG names by the short-read analysis.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample ID | number of read pairs in sample (post QC and host-filtering) | no. of +ve read pairs,  RGI BWT (filtered) | +ve rate RGI BWT (%) | no. of +ve gene names, RGI BWT (either read) | no. of +ve gene names,  RGI MAIN (unfiltered) | no. of +ve gene names,  RGI MAIN (filtered) | no. of +ve gene names by both RGI BWT and RGI MAIN (filtered) | no. of +ve reads, RGI BWT for genes also +ve for RGI MAIN (filtered) | read pairs per +ve gene name (RGI BWT filtered) | read pairs per +ve gene name (RGI BWT + MAIN filtered) |
| 343587 | 15,803,729 | 32,660 | 0.207 | 55 | 7 | 4 | 2 | 5,151 | 593.8 | 2,575.5 |
| 343701 | 31,310,738 | 81 | 0.000 | 4 | 2 | 2 | 0 | 0 | 20.25 | - |
| 343624 | 14,454,308 | 28,277 | 0.196 | 74 | 2 | 0 | 0 | 0 | 382.1 | - |
| 2672534 | 23,626,929 | 512 | 0.002 | 70 | 69 | 68 | 3 | 34 | 7.3 | 11.3 |
| 2672718 | 23,182,503 | 10,659 | 0.046 | 127 | 121 | 112 | 23 | 1,419 | 83.9 | 61.7 |
| 343520 | 487,646 | 84 | 0.017 | 17 | 6 | 5 | 0 | 0 | 4.9 | - |
| 2672664 | 7,125,112 | 2,083 | 0.029 | 47 | 42 | 38 | 6 | 621 | 44.3 | 103.5 |

Table 2. Comparison of (filtered) RGI BWT, unfiltered RGI MAIN and filtered RGI MAIN results.

It might be supposed that true positives would be more frequent among those genes (names) predicted by both methods than by only one method. For RGI BWT, if it is also supposed that read counts tend to be lower for false positives than for true positives, then it follows that higher mean read frequencies (per gene name) would occur for the consensus-predicted gene names than for those predicted only by RGI BWT. The final two columns of Table 2 compare these frequencies, for the four samples where the number of consensus gene names is non-zero. The mean frequencies of the consensus-predicted genes are higher in three of the four cases; this comparison would need to be made for many more samples for any conclusions to be drawn.

A more direct approach to evaluating the RGI BWT positives (especially those which are not positive by RGI MAIN) is to independently check the reads (individual or pairs) which have been classified as positive (after our in-house filtering stage). That is, to not assume correctness of the RGI MAIN results in the context of this data; we have earlier described concerns about the suitability of these assembled metagenomes in general for production use for identifying ARGs. Even though the seven test samples were selected on the basis of a high proportion of back-mapped reads, the proportion which does not map to predicted ORFs, though a minority, is not insignificant at 13-27% (Table 1).

Clearly, independently checking all of the RGI BWT ARG-positive reads would be a very large analysis beyond the scope of this project, but here we highlight some examples from the first two samples in Table 2.

**Sample #343701**

This sample has the following four ARG names predicted as positive by RGI BWT (after our in-house filtering), with numbers of match read pairs (or single reads) shown:

* AAC(6')-Ie-APH(2'')-Ia (1 x R1 only)
* adeF (1 x R2 only)
* APH(3')-Ia (1 x pair)
* patB (71 x R1 only, 2 x R2 only, 5 x pair)

Whereas, the two predicted positive by RGI MAIN (irrespective of whether post-filtering is done) are:

* APH(2'')-Ie
* poxtA

i.e. there is strictly no agreement, though the first ARG in each list might be assumed to be very similar (in ARO, which confirms that the two share an aminoglycoside phosphotransferase function, these are sibling terms), and we do not discuss AAC(6')-Ie-APH(2'')-Ia here for that reason. We also do not discuss the APH(2'')-Ie and poxtA ARGs, as these might be assumed to be absent from the former list due to the limitations of using short reads instead of longer assembled sequences.

We used the **adeF** (R2 read sequence), **APH(3')-Ia** (both R1 and R2) and one of the R1, R2 pairs of **patB** as queries in BlastX searches versus the NCBI nr protein database (this is often a more direct way of obtaining an indication of gene/protein, since searching versus the nucleotide databases will often match a segment of a genome sequence without the desired annotations). (For reference, for these first 3 examples we include the query sequences used.)

*adeF*:

The 151 bp query made 100%-identical matches at the protein sequence level (a 48 amino-acid segment) with a number of sequences with annotations which were either "multidrug efflux RND transporter permease subunit" or "efflux RND transporter permease subunit", in several species of *Comomonas* (soil and other environmental organisms). According to ARO adeF (ARO:3000777) "AdeF is the membrane fusion protein of the multidrug efflux complex AdeFGH". This seems to confirm this identification - *which notably arose from a single R2 read* - at least at the functional level, even if the precise component of the system is not of identical name.

>putative adeF

ATAGTCTGGCCATGTTCGCCATGGTGCTGGTGATCGGTATCGTGGTGGACGATGCCATCGTGGTGGTGGAAAACGTCGAGCGCATCATGGCCGAGGAAGGCCTCCCCCCCAAGGAAGCCACCATCAAGGCCATGGGCCAGATCCAGGGCGC

*APH(3')-Ia*:

The R1 and R2 reads total 302 bp, and a single sequence consisting of them spliced (separated by 10 x "N", with the R2 as it appears in the SAM output - i.e. the reverse complement of the actual read) was used as the query. There were many hits of 100% identity over a 50 amino-acid segment as well as a second segment where 25 / 26 amino acids were identical, to proteins annotated as "APH(3') family aminoglycoside O-phosphotransferase"; the first is annotated as "aph(3')-Ia (plasmid)" in *Klebsiella pneumoniae*. This too is therefore deemed a true positive, *arising from a single read pair*.

>putative APH(3')-Ia

TTACATAAACAGTAATACAAGGGGTGTTATGAGCCATATTCAACGGGAAACGTCTTGCTCGAGGCCGCGA

TTAAATTCCAACATGGATGCTGATTTATATGGGTATAAATGGGCTCGCGATAATGTCGGGCAATCAGGTG

CGACAATCTATNNNNNNNNNNCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCAC

GTTAAGGGATTTTGGTCATGAACAATAAAACTGTCTGCTTACATAAACAGTAATACAAGGGGTGTTATGA

GCCATATTCAACGGGAAACGTCTTGCTCGAGG

*patB*:

The R1 and R2 reads total 296 bp, and a single sequence consisting of them spliced (separated by 10 x "N") was used as the query. It was subsequently apparent than in this particular pair, R1 and R2 (reverse complement) were identical, so in effect this represented 148 bp of unique sequence. Nonetheless this made 100%-identical matches in a 47 amino-acid segment to a number of proteins, annotated as "multidrug ABC transporter ATPase/permease" in various *Lactobacillus* species, also "Lipid A export ATP-binding/permease proteinMsbA" in *Lactiplantibacillus*, "ATP-binding cassette domain-containing protein" (i.e. ABC-containing protein) in *Lacticaseibacillus rhamnosus*, and others. Thus this too seems credible, since according to ARO (ARO:3000025), "PatB is an ABC transporter of *Streptococcus pneumoniae* that interacts with PatA to confer fluoroquinolone resistance." *Streptococcus* is, like *Lactobacillus,* a member of order Lactobacillales.

>putative patB

GGCTAACATCCACGATTTCATCGTGAGCTTGCCAGAAGGCTACGATACCTTTGTTTCAGATGAACAAAGTGTTTTCTCGGCAGGACAAAAGCAGTTGATGTCGATTGCGCGAACTATTTTGACCAATCCGCGGTTGTTAATTC

**Sample #343587**

We provide an overview of this sample, which has many more ARGs predicted (55 by RGI BWT, post-filtering), from > 32,000 reads (post-filtering). The results are summarised in Appendix MMM. That shows, for each ARG, the numbers of R1 read matches, the numbers of R2, and the number of pairs where both reads matched well enough to pass the filtering, and agreed on the name of the ARG. Of the 55 ARGs, APH(3'')-Ib and APH(6)-Id were also present in the post-filtered RGI MAIN results (see also Table 2). A third ARG, adeF, had also been present in the RGI MAIN results but had been removed by our in-house filtering.

We have selected some gene names at random, which are positive by RGI BWT but absent from the RGI MAIN predictions.

*adeF* (positive by both RGI BWT and RGI MAIN):

For the sake of comparison, we first examined a single read (an R1 whose R2 was negative) and a different read pair, which had been identified by both RGI BWT *and* RGI MAIN (adeF; also tested in the previous sample), again using BlastX versus nr. The single R1 query made various high-identity matches in a number of *Pseudomonas* species annotated as "efflux RND transporter permease subunit", including a 98% identical (49 / 50 amino acids) match with a *P. alkylphenolica* protein "multidrug efflux RND transporter permease subunit". Similar results were obtained with the paired-sequence query, whose hits included "efflux RND transporter permease subunit" from *P. fragi*, consisting of two segments both of 100% identity and respectively 50 and 42 amino acids long.

The remaining reads (or pairs) investigated had been identified by RGI BWT as ARGs but which were *negative* in the RGI MAIN results (whether post-filtered or not).

*mexF*:

This was positive by RGI BWT (post-filtering) by merit of only *a single R1 read* (151 bp), whose R2 partner was negative. According to ARO, "MexF is the multidrug inner membrane transporter of the MexEF-OprN complex. mexF corresponds to 2 loci in *Pseudomonas* aeruginosa PAO1 (gene name: mexF/mexB) and 4 loci in Pseudomonas aeruginosa LESB58 (gene name: mexD/mexB)" (ARO:3000804). BlastX hits included several 98%-identical matches (49 / 50 amino acids), annotated as "efflux RND transporter permease subunit" in various *Pseudomonas* species (some unnamed), including *P. alkylphenolica.* This therefore also appears to be a credible positive.

*arnA*:

ArnA is a bifunctional enzyme, and one of its functions is synthesis of UDP-4-keto-pentose (Gatzeva-Topalova et al., 2004). ARO: "arnA modifies lipid A with 4-amino-4-deoxy-L-arabinose (Ara4N) which allows gram-negative bacteria to resist the antimicrobial activity of cationic antimicrobial peptides and antibiotics such as polymyxin. arnA is found in E. coli and P. aeruginosa" (ARO:3002985). 711 read pairs were positive by RGI BWT (almost 1,000 R1). As a query, we used a positive read pair of aggregate length 302 bp. This made a perfect match (two segments of 50 amino acids each, both of 100% identity) with "bifunctional UDP-4-keto-pentose/UDP-xylose synthase" of *Pseudomonas fragi*, and other similar hits (e.g. a less similar match in *Acinetobacter baumannii*) . Indeed, at least one is annotated as "arnA" ("bifunctional UDP-4-amino-4-deoxy-L-arabinose formyltransferase/UDP-glucuronic acid oxidase ArnA" in *Yersinia pseudotuberculosis* has two segments 90% and 92% identical). This therefore appears to be a sound positive.

*Erm*(36):

Erm(36) was positive by RGI BWT by merit of *a single read pair*, with both R1 and R2 matching. In ARO, Erm(36) (synonym "ErmD") is annotated as "ErmD confers MLSb phenotype" (*sic*) (ARO:300605). The associated literature concerns plasmid-borne macrolide resistance in *Micrococcus* (Liebl et al., 2002) and describes a putative 281-residue protein with similarity to 23S rRNA adenine N(6)-methyltransferases. The nr matches we found via BlastX are annotated as "23S ribosomal RNA methyltransferase Erm", "MULTISPECIES: 23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(36) [Micrococcaceae]", etc and the best match is with a *Microccocus* protein where there are two 100%-identical segments of 39 and 50 amino acids. This therefore appears to be a sensible result.

*mexI*:

According to ARO, "MexI is the inner membrane transporter of the efflux complex MexGHI-OpmD" (ARO:3000808). In our RGI BWT results, mexI was positive due to only *a single read pair* (both R1 and R2 matched; both are relatively short). A BlastX of the aggregate query sequence (totalling 265 bp) yields several matches that are 100% identical over two segments of 43 amino acids, and are annotated as "MexW/MexI family multidrug efflux RND transporter permease subunit" or similar. However, this does not necessarily indicate a clear positive, as membership of this family does not necessarily mean that this DNA is part of a mexI gene itself; indeed, some of the hits have multiple annotations such as "MULTISPECIES: MexW/MexI family multidrug efflux RND transporter permease subunit [unclassified Pseudomonas]" coupled with "Efflux pump membrane transporter BepE". However, within the constraints of the short read query, it seems like a reasonable putative positive, likely to be accurate in terms of predicted function.

Thus far, the ARG assignments in these examples bear examination. In contrast, the following examples appear to be much less convincing.

*AAC(3)-IIe*:

135 read-pairs were positive for this ARG (slightly more R1 and R2 if treated individually). We selected *one* of the longest aggregate pairs (2 x 90 bp R1 and R2). This had a number of high-identity matches in nr, including several of 100% identity (over two segments of 37 + 31 amino acids). However, the annotation was consistently "Tn3 family transposase" or similar. According to ARO, "AAC(3)-IIe is a plasmid-encoded aminoglycoside acetyltransferase in E. coli" and has the synonym aacC2e (ARO:3004621). Tn3 transposases are in fact associated with a *different ARG type* (beta-lactamases), being one of the three components of the Tn3 transposon. This is distinctly different from AAC(3)-IIe.

We checked the match of the read pair with the CARD reference sequence ("Prevalence\_Sequence\_ID:129123|ID:3321|Name:AAC(3)-IIe|ARO:3004621") which had given rise to the identification. In brief, we concluded that this reference sequence includes a segment at its end which is part of a Tn3 transposase gene, adjacent to the AAC(3)-IIe coding sequence, and the match of the short reads was to this segment. This therefore is concluded to be a false positive (and potentially indicates a false negative elsewhere, if a corresponding beta-lactamase was not positive in this sample).

N.B.: we have only checked *one read-pair*, and this does not necessarily mean that all of the positives are also false, but that is possible (they could all be fragments of the Tn3 transposase).

*AAC(3)-IIb*:

15 read-pairs (22 individual R1, 18 individual R2) were positive for this ARG. We used an aggregate total 240 bp query (R1 + R2), and concluded that this pair too is a false positive. It clearly matches various chaperonin sequences, and again we confirmed that a segment of the CARD reference sequence ("Prevalence\_Sequence\_ID:58882|ID:925|Name:AAC(3)-IIb|ARO:3002534") contains a segment towards the end which represents this, but the bulk of the reference contains the AAC(3)-IIb annotated sequence. Again, this does not necessarily mean that all the positives for this ARG are false. (ARO: "AAC(3)-IIb is an aminoglycoside acetyltransferase in *E. coli*, *A. faecalis* and *S. marcescens*"; ARO:3002534).

*acrF*:

28 read-pairs (61 individual R1, 34 individual R2) were positive for this ARG. We used an aggregate total 286 bp query (R1 + R2), and concluded that this pair too is a false positive. According to ARO, "AcrF is a inner membrane transporter, similar to AcrB" (ARO:3000502), and is a component of AcrEF-TolC, conferring resistance to ciprofloxacin. However, the BlastX matches are consistently with *amino acid* ABC transporter permeases. Although these may share an inner-membrane transporter context, to the best of our knowledge these should be assumed to be different. We did also check the matched CARD reference ("Prevalence\_Sequence\_ID:139312|ID:1437|Name:acrF|ARO:3000502").

*eptA*:

eptA was positive by merit of a *single read pair*, of which both reads are 151 bp long. However, the aggregate query of 302 bp makes a perfect match (two 100%-identical segments of 50 and 50 amino acids) with "Biodegradative arginine decarboxylase" from *E. coli* and other similarly-annotated hits. However, EptA (synonym PmrC) "mediates the modification of Lipid A by the addition of 4-amino-4-deoxy-L-arabinose (L-Ara4N) and phosphoethanolamine, resulting in a less negative cell membrane and decreased binding of polymyxin B" (ARO:3003576). This is an unrelated function and we conclude that this is a false positive.

*Klebsiella pneumoniae KpnH*:

This was positive due to 14 read pairs (38 R1 and 17 R2 if considered separately). We used a 301-bp aggregated query sequence, whose hits are very consistently annotated as type I DNA topoisomerase (mostly in *Pseudomonas* species), including matches of 100% identity over segments of 49 and 50 amino acids. ARO annotates Klebsiella pneumoniae KpnH as "KpnH consists of ~511 residues, resembles EmrB of E. coli, and is probably a translocase in the KpnGH-TolC efflux protein in *K. pneumoniae*" (ARO:3004597). This is a very different function, and we conclude that this is a false positive.

A final caveat to these findings is that the annotations in the nr database cannot be assumed to always be correct. Annotation errors are not rare, and errors can propagate via annotation by sequence similarity, leading to groups of sequences incorrectly (but consistently) misannotated. Therefore were there are discrepancies, the RGI results could still be correct - notwithstanding that an apparent explanation for the discrepancies (segments of adjacent genes) has been identified in some cases.

**Summary of the comparison between RGI BWT and RGI MAIN results**

* Generally, far fewer ARGs are predicted by RGI MAIN, even where the samples' metagenomes appear to have a high assembly quality
* Even when numbers are similar for a given sample, there may be little overlap between RGI MAIN and RGI BWT
* By examination of an *ad hoc* selection of a few RGI BWT-positive ARGs in two samples, using a general reference database, there appear to be clear examples of:
  + the RGI BWT identification being well supported, even it rests on identification of a single read pair, or even a single read
  + conversely, of individual reads/pairs being apparently false positives
    - these may sometimes be the sole instance of the ARG in the sample - in which case the ARG itself appears to be a false positive in the sample;
    - or may be one of a much larger number of reads/pairs
    - the identified reads always appear to be well-matched to the CARD reference sequence
    - but some of the references appear to have segments of other genes at their 5' or 3' end; though relatively short, these can lead to a good match in the short-read context
  + some others which are less clear either way

It is important to bear the context in mind, which is the anecdotal nature of the "manual" checks; however, the above does represent all of those we have checked in detail. Of twelve individual read pairs (or single reads) examined across the two samples, 7 appear to be well founded (or perhaps sound as "putatives" in one or two cases) while 5 appear to be clear false positives. Since, for each of these two outcomes, there are some cases of "lone-determinant" reads (pairs), there is not necessarily much to be gained by using a high read (pair) count as a "weight of evidence" to designate more likely true positives.

Nonetheless, the fact remains that for an ARG to be viewed as genuinely positive in a sample requires *only one* of the reads/pairs to be considered a genuine match. The context is that in the sample #343587, 9 of the 54 ARGs are represented by a single read (or read pair), and a further 6 between 2 and 9 reads (pairs). 25 ARGs are represented by at least 50 reads (pairs). In #343624, 18 of the 74 ARGs are represented by a single read (or pair), and 29 by > 50. In #2672718 (which is another of the 7 examples with > 10,000 positive reads), 34 of the 127 ARGs are represented by 1 read (or pair), and 31 by > 50.

We would also like to emphasise our confidence in the matches between the metagenomic reads and the CARD reference sequences themselves, especially in the context of the detailed post-filtering we have applied. While it is unlikely that there will be no cases of poor matches between reads and references in the entire dataset of 256 samples, we believe that a huge majority of the matches are at a high level of sequence similarity, and thus sound. (Of course, inferences, such as functional inferences, from a particular level of similarity are not always guaranteed because there are no universal "rules".) The issues we have identified above relate to those particular reference sequences themselves, not in the degree of match between them and the reads.

A systematic benchmarking of RGI / CARD is of course well beyond the scope of this study (we also note the license terms: "*Licensee may not publish any work related to the evaluation of the CARD and the RGI without McMaster's approval*"). Most sequence-analysis tasks which aim to predict function from DNA/protein are subject to significant rates of false positives and false negatives, and the above places this in context.

#### ARG predictions from DeepARG

DeepARG is a tool published in 2018, which used a novel approaching implementing a deep-learning methodology to predict ARGs. As stated in the project proposal, our default position was to treat RGI / CARD predictions as the standard and DeepARG predictions (DeepARG-ss for short-read input data, and DeepARG-ls for assembled sequences) as more speculative. We ran it with a probability cutoff of P ≥ 0.8, and subsequently applied a filter of P ≥ 0.9 to these results.

Using the 7 example samples as described for RGI, it is apparent that both DeepARG tools predict a great many more positive ARGs compared to RGI, and the total number of ARG-positive reads is also much greater than by RGI (at least an order of magnitude in general). Although it is difficult to anticipate an expected volume of positive reads (and ARGs), we do view these numbers as very high.

We also noted what appears to be an approximate constant behaviour regarding the proportion of the numbers of reads which are matched per ARG for ARGs which are positive by DeepARG-ss (i.e. whether positive or negative by DeepARG-ls), compared to reads for ARGs positive by DeepARG-ls (as well DeepARG-ss).

That is, the mean number of reads per ARG for DeepARG-ss positives (general) is always (in these 7 samples) approximately 45% of the mean for ARGs which are also positive by DeepARG-ls (the last two columns in Table 3). This does not necessarily indicate a problem, but does mean that those predicted by both methods cannot be viewed as more reliable than those predicted only by DeepARG-ss. Otherwise, there would be an expectation that for some samples (where true positives are a much higher proportion of reads), this ratio of the means for the intersection, compared to the overall DeepARG-ss positive ARGs, would be expected to be significantly higher than in other samples. That would follow from false positives being over-represented in the general set (compared to the insersection subset), because this set includes the ss-only ARGs.

To put it another way, from the mean positive reads per ARG from DeepARG-ss, the mean positive reads for the DeepARG-ss-only, and ss/ls intersection, ARGs can be predicted. In fact, remarkably from the number of DeepARG-ss positive ARG names, the number of DeepARG-ls positive ARG names (irrespective of the DeepARG-ss results) can also be predicted quite well in these samples, since it is approximately one third of the former.

In one sense this indicates that the DeepARG methodology is very consistent, but it does seem surprising given the different natures of the two types of input data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample ID | number of reads | no. of +ve reads,DeepARG-ss | +ve rate DeepARG-ss (%) | no. of +ve gene names, DeepARG-ss | no. of +ve gene names, DeepARG-ls | no. of +ve gene names by both DeepARG-ss and DeepARG-ls | mean reads per DeepARG-ss +ve gene | mean reads per DeepARG-ss/-ls both +ve gene |
| 343587 | 15,803,729 | 138,909 | 0.88 | 214 | 68 | 59 | 649.1 | 1,476.1 |
| 343701 | 31,310,738 | 138,152 | 0.44 | 94 | 29 | 24 | 1,469.7 | 2960.7 |
| 343624 | 14,454,308 | 110,882 | 0.77 | 216 | 71 | 62 | 513.3 | 1,186.9 |
| 2672534 | 23,626,929 | 77,459 | 0.33 | 327 | 119 | 99 | 236.9 | 515.4 |
| 2672718 | 23,182,503 | 75,145.5 | 0.32 | 425 | 151 | 135 | 176.8 | 439.7 |
| 343520 | 487,646 | 980 | 0.2 | 132 | 37 | 29 | 7.4 | 18.0 |
| 2672664 | 7,125,112 | 43,010.5 | 0.6 | 277 | 110 | 92 | 155.3 | 310.9 |

Table 3. Comparison of DeepARG-ss and DeepARG-ls results.

##### Comparison of the DeepARG-ss results with RGI BWT

While noting that many more DeepARG-ss reads are positive compared to RGI BWT, we compared the read-by-read assignments of each, where both made an ARG-positive prediction; notwithstanding that as noted previously, there are some nomenclature differences between the two reference databases, so some differences would always be expected.

Again taking the sample **#343587** as an example, we found a large number (4,375) of individual read (R1, R2) assignments which agreed (26 different ARG names were involved). However, the number of assignments not in agreement was much larger: 27,453. On the other hand, this number arose from 55 unique pairings of different ARG names, and in fact one RGI/CARD name (adeF) accounted for most of these, and for 23,971 of the differing name assignments in all. For example, there are 9,232 instances of RGI/CARD 'adeF' paired with DeepARG-ss 'MEXF'; with adeF paired with 7,571 MEXB; 2,231 MEXD; 2,046 ACRB; 1,371 ADEB, etc. There is a very similar situation in sample **#343624**, with 8,090 reads in agreement (27 ARGs) and 26,702 differing (62 pairings), with adeF accounting for 24,950 of the disagreements (10,597 adeF-MEXB pairings alone). (These are both milk samples.)

We have already noted several instances of adeF assignments by RGI BWT which we confirmed as appearing to be sound. According to ARO, "AdeF is the membrane fusion protein of the multidrug efflux complex AdeFGH" (ARO:3000777). "MexB" also exists in ARO (3000378)/ CARD: "MexB is the inner membrane multidrug exporter of the efflux complex MexAB-OprM". Also, other Mex components (e.g. mexA, mexI, mexW) are also present in the RGI BWT predictions of these samples. It may be that there is indeed a naming difference between the two databases; or in the context of short reads, precise identifications may be difficult or not always possible, even if the functional essence of the ARG is identified - where the two methods are in agreement.

This does seem to particularly affect these adeF, MEX and related ARGs. For example, in sample **#2664796**, where these are not predicted by either method, there are 5,752 reads in agreement (albeit only arising from 2 ARGs) and 348 differing (5 unique pairings, including lmrD versus LMRC, and lmrD versus "multidrug\_abc\_transporter") (lmrD is an efflux pump and lmrC is an ABC-F subfamily protein). Similarly in sample **#2664672**, 5,825 reads agree (1 ARG: lmrD) and 353 differ (9 pairings). (The ARG names predicted within these two samples are highly similar; these are two cheddar cheese samples.)

Gatzeva-Topalova, P. Z., May, A. P. & Sousa, M. C. 2004. Crystal structure of Escherichia coli ArnA (PmrI) decarboxylase domain. A key enzyme for lipid A modification with 4-amino-4-deoxy-L-arabinose and polymyxin resistance. *Biochemistry,* 43**,** 13370-9.

Liebl, W., Kloos, W. E. & Ludwig, W. 2002. Plasmid-borne macrolide resistance in Micrococcus luteus. *Microbiology (Reading),* 148**,** 2479-2487.