High-throughput rumen microbial profiling using genotyping-by-sequencing

M.K. Hess¹, S.J. Rowe¹, T.C. Van Stijn¹, R. Brauning¹, A.S. Hess¹, M.R. Kirk², G.T. Attwood², P.H. Janssen² & J.C. McEwan¹

¹AgResearch Limited, Invermay Agricultural Centre, Puddle Alley, Mosgiel, New Zealand melanie.hess@agresearch.co.nz (Corresponding Author)
²AgResearch Limited, Grasslands Research Centre, Tennent Dr, Palmerston North, New Zealand

Summary

The composition of the rumen microbial community has been associated with host traits such as feed efficiency and methane production. Gaining an understanding of the rumen microbial profile of a large number of animals may help to select more efficient and environmentally-friendly animals. A current barrier to obtaining rumen microbial profiles on a large number of animals is the cost and time necessary to obtain these profiles using traditional sequencing methods. We explored the potential of using genotyping-by-sequencing (GBS) to produce cost-effective and high-throughput rumen microbial profiles by developing a pipeline to BLAST reads against the Hungate1000 dataset of rumen microbial genomes and assign their taxonomies. This pipeline was then tested on GBS data from 236 high- or low-methane sheep rumen samples using either ApeKI or PstI restriction enzymes. On average, 2.4 and 2.8 million reads per sample were obtained for ApeKI and PstI, respectively, with average lengths of 71 and 81 base pairs. The average hit rate at the genus level for ApeKI was 5.3±1.7% and for PstI was 6.8±1.8% which is in line with similar WGS metagenomic studies. The accuracy of taxonomic assignment at the genus level was estimated at >96% and increased for higher taxonomic levels. Samples from high-methane animals had lower hit rates at the genus level than samples from low-methane animals (ApeKI: 1.2%; PstI 1.0%; p<0.0001) due to the species represented in the Hungate1000 dataset. These results suggest that GBS is a promising technique for low-cost, high-throughput rumen microbial profiling.

Keywords: genotyping-by-sequencing, rumen, microbiome, metagenomics, efficiency

Introduction

Ruminants have a symbiotic relationship with the microbes in their rumen, which aid in important biological functions, primarily providing energy in the form of volatile fatty acids from the breakdown of plant cellulose, along with microbial protein as a source of amino acids. Although similar bacterial and archaeal species have been found in the rumens of diverse species (Henderson et al., 2015), the prevalence of different microbial species within the rumen varies between individuals and is heritable (Roehe et al., 2016; Rowe et al., 2015; Ross et al., 2013). Differences in microbial composition have been linked to traits such as feed efficiency and methane production in multiple ruminant species (Kittelmann et al.,
2014; Shi et al., 2014 Danielsson et al., 2017; Jami et al., 2014). Obtaining a rumen microbial profile (RMP) may allow us to predict these traits with higher accuracy, thereby making greater gains towards more efficient and environmentally-friendly ruminants.

Two common approaches to obtain RMPs are through targeted and whole genome shotgun (WGS) sequencing. Targeted sequencing first amplifies specified taxonomically-informative genes from a metagenome sample, such as the gene encoding the small subunit ribosomal RNA (ssu rRNA) of microbes. Such targeted approaches typically distinguish taxonomic groups well and perform best with long sequence reads due to having some regions that are highly conserved and other (informative) regions that are more variable (Franzén et al., 2015). There are databases of ssu rRNA sequences with taxonomic assignments that extensively cover both culturable and unculturable microbes (DeSantis et al., 2006). In contrast, WGS doesn’t target any specific gene but can capture any part of the microbial, host or feed genome. Assignment of sequence reads to a taxonomic group typically involves aligning reads to a database of sequences with known taxonomies; however, because whole genome assemblies are typically required for this step, the databases aren’t as extensive as those available for ssu rRNA sequencing and primarily contain cultured microbes (Kelly et al., 2014).

Although both ssu rRNA and WGS approaches successfully generate RMPs, their cost prohibits their implementation in a commercial selection programme. Genotyping-by-sequencing (GBS) is a next-generation sequencing technique that reduces genome complexity by digestion of genomic DNA by restriction enzymes, followed by the sequencing of fragments within a given size range (Elshire et al., 2011). GBS holds promise as a technique for rapid, high-throughput and cost-effective sequencing of microbial samples as it can cost less than half the price of targeted sequencing and only a fraction of the cost of WGS. This paper describes the development of a rumen microbial profiling pipeline using the Hungate1000 genomes dataset (Kelly et al., 2014) and a WGS dataset of rumen samples from 8 high- and 8 low-methane sheep (Shi et al., 2014). This pipeline was then tested on GBS data from 236 rumen samples from high- and low-methane sheep to further evaluate the potential of GBS as a low-cost technique to obtain RMPs on a large number of animals.

Material and methods

A pipeline for analysis of rumen microbial sequence reads was developed based on two available datasets: the Hungate1000 dataset (Kelly et al., 2014) and the WGS dataset used by Shi et al. (2014). The Hungate1000 dataset consists of nucleotide and protein assemblies of 409 rumen microbes. The WGS dataset included WGS for 8 high- and 8 low-methane sheep. WGS reads were trimmed using trim_galore (Krueger 2015) for paired reads with a minimum length of 40 base pairs. GBS uses single-end reads so only the forward strand of each pair was retained. Each sample had ~220 million reads after this quality control. One million reads were randomly sampled from each of the high- and low-methane samples and the first or last 92 nucleotides were retained from each read and used as query sequences in BLAST, as described below.

In silico Restriction Enzyme Digestions

In silico restriction enzyme digestions were performed on the nucleotide assemblies of the Hungate1000 dataset for two single digestions: ApelI and PstI, using custom code in R. Fragments that were between 65 and 195 base pairs in length represented parts of the genome that may be captured using GBS. These fragments were then trimmed to a maximum of 92
base pairs starting arbitrarily at the 3’ or 5’ end of the sequence to represent sequences that may be obtained by GBS.

**Bioinformatic Pipeline Development**

To optimize the parameters for aligning query sequences to the Hungate1000 database, analyses were run using NCBI’s command line BLAST v 2.2.28+ (Camacho et al., 2009) with default parameters except for e-value = 0.02. The evaluated parameters were 1) a BLAST algorithm that queries based on nucleotide sequence (BLASTN) vs. amino acid sequence (BLASTX) and 2) the BLAST database (full vs. filtered). Query sequences were either the digested Hungate1000 or the WGS query sequences. The database contained the Hungate1000 nucleotide assemblies (or predicted proteins for BLASTX). The filtered database had those assemblies for which taxonomy was not known to the genus level removed.

Query sequences were assigned to a taxonomic node based on an algorithm similar to MEGAN (Huson et al., 2007) and implemented in R. All hits with a bit score (which measures the quality of the alignment) above 50 were retained. For a given query sequence, the maximum bit score of hits was found and hits with a bit score within 10% of that bit score were considered the top hits. The query sequence was then assigned to the lowest taxonomic node that encompassed all top hits. The taxonomic levels of interest were kingdom (or domain), phylum, class, order, family, genus and species. Some query sequences were not assigned to any taxonomic node because they either had no hits that were above the bit score threshold, or their top hits contained hits to both bacterial and archaeal sequences.

The performance of different BLAST parameters was evaluated through accuracy, runtime and hit rate. Accuracy (calculated as the proportion of correct hits at a given taxonomic level) was evaluated by BLASTing the (in silico) digested fragments from the Hungate1000 genomes against the Hungate1000 database, ignoring the perfect hit between the query sequence and the genome it originated from. Runtime was the System + User time taken to BLAST 1 million sequences from a WGS sample, and hit rate was the proportion of query sequences that were assigned at a given taxonomic level. If a sequence was assigned at a lower taxonomic level, it was considered in the evaluation of higher taxonomic levels e.g. if a query was assigned at the taxonomic level of species, the genus of that species was considered when evaluating accuracy at the genus level.

**GBS on DNA from Rumen Samples**

A collection of 236 rumen samples, consisting of 2 samples from each of 118 low- or high-methane sheep (Kittelmann et al., 2014) were available for GBS. The freeze dried rumen samples were extracted using a combined bead-beating, phenol and column purification protocol as described in Kittelmann et al. (2014) to provide high quality nucleic acids for GBS. The DNA separately digested by *Ape*KI and *Pst*I restriction enzymes (Elshir et al., 2011), and barcodes identifying each sequence to each sample were attached. Samples were grouped into two libraries based on the restriction enzyme used. Sequences between 193 and 318 bp (equivalent to 65 to 195 base pair inserts) were selected using a Pippin Prep (SAGE Science, Massachusetts, USA) and each library was run across 2 lanes on the same flow cell on an Illumina HiSeq2500 machine. After sequencing, the sequence reads were demultiplexed using GBSX (Herten et al., 2015), and trimmed using trim_galore (Krueger 2015) for single reads with a minimum length of 40 base pairs. The reads were then run through the pipeline developed in the previous section. The hit rate was evaluated at each taxonomic level for the two libraries separately, and also by whether the sample was from a high- or low-methane individual. The difference in hit rate at the genus level between
samples from high- and low-methane sheep was tested using a 2-tailed t-test with unequal variances and $\alpha = 0.05$.

## Results and Discussion

### In silico Restriction Enzyme Digestions

The average proportion of each bacterial or archaeal genome that is expected to be captured by GBS is shown in Table 1. In silico digestion by *ApeIKI* captured a larger proportion of the Hungate1000 genomes than *PstI* and all Hungate1000 genomes were captured to some degree. There was one Hungate1000 genome that was not captured at all by *PstI*; however, other genomes assigned to the same species were captured, indicating this may have been due to an incomplete genome rather than digestion by *PstI* not capturing any of the genome. These results indicate that, provided that the portion of the genome that is captured can distinguish reads from different taxa, *PstI* may be the better restriction enzyme to use for the purpose of developing low-cost GBS sampling of the rumen microbiome. Using *PstI* may allow us to reduce read depth per sample compared to *ApeKI* due to less redundancy in available sequence from a given microbe.

<table>
<thead>
<tr>
<th>Restriction enzyme</th>
<th>Proportion of genome captured (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td><em>ApeKI</em></td>
<td>8.6</td>
</tr>
<tr>
<td><em>PstI</em></td>
<td>0.3</td>
</tr>
</tbody>
</table>

### Bioinformatic Pipeline Development

Using BLASTX resulted in a higher hit rate than using BLASTN (Table 2), but it took over 11 days to process 1 million reads compared to ~15 minutes for BLASTN. The accuracy of taxonomic assignment was similar between the two BLAST methods, and was lowest for reads assigned at the species level but increased >96% for reads assigned to genus level or higher (Table 3). Although accuracy was similar for the two methods, hits with BLASTN tended to be at lower taxonomic levels than BLASTX (i.e. with BLASTN ~80% of hits were assigned to the species or genus level, while for BLASTX these were only ~40% of hits). The remainder of these analyses will focus on BLASTN analyses because the computation time for BLASTX was deemed unreasonable for a high-throughput GBS pipeline.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sample(^1)</th>
<th>Hit rate by taxonomic level (%)(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kingdom</td>
<td>Phylum</td>
</tr>
<tr>
<td><em>ApeKI</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>PstI</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Proportion of microbial genomes captured by in silico restriction enzyme digestion.

Table 2: Hit rate for BLASTX and BLASTN by taxonomic level.
<table>
<thead>
<tr>
<th>Method</th>
<th>Restriction enzyme</th>
<th>Accuracy by taxonomic level (%)</th>
<th>Kingdom</th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genus</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTN</td>
<td>ApeKI</td>
<td>100.0</td>
<td>99.8</td>
<td>99.7</td>
<td>99.5</td>
<td>98.2</td>
<td>96.1</td>
<td>83.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>100.0</td>
<td>99.9</td>
<td>99.8</td>
<td>99.6</td>
<td>98.5</td>
<td>97.8</td>
<td>86.3</td>
<td></td>
</tr>
<tr>
<td>BLASTX</td>
<td>ApeKI</td>
<td>100.0</td>
<td>99.5</td>
<td>98.9</td>
<td>98.9</td>
<td>96.9</td>
<td>96.0</td>
<td>83.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PstI</td>
<td>100.0</td>
<td>99.0</td>
<td>98.7</td>
<td>98.7</td>
<td>96.8</td>
<td>96.8</td>
<td>86.3</td>
<td></td>
</tr>
</tbody>
</table>

Hit rates (Table 4) and accuracy were similar when using the full Hungate1000 database or the database filtered to remove sequences not assigned a taxonomy at the genus level. Runtimes for the database, including the BLASTN and summarizing steps, took 27 minutes, ~4 minutes faster than when the full database was used. Therefore, for the analysis pipeline used on the GBS samples, the filtered database was used.

**GBS on DNA from Rumen Samples**

Read lengths ranged from 40 to 92 base pairs with an average length of 71 base pairs for ApeKI and 81 base pairs for PstI. There were an average of 2.4 million reads per sample for ApeKI and 2.8 million for PstI. The hit rates for the GBS samples (Table 5) were consistent with the hit rates obtained from the WGS dataset and those found in other WGS studies (Watson et al., 2017). Samples from low-methane sheep had a higher hit rate than the high-methane samples for both restriction enzymes (ApeKI: \( p = 1.4 \times 10^{-8} \); PstI: \( p = 4.0 \times 10^{-5} \)). Overall, the consistency between the WGS and GBS results suggest that GBS is a suitable technique to obtain rumen microbial profiles.

**Table 3: Accuracies for BLASTX and BLASTN by taxonomic level.**

**Table 4: Average hit rate by taxonomic level for two BLAST databases.**

**Table 5: Average hit rate for GBS samples by taxonomic level.**

1 Sample = Sample from high- or low-methane sheep
2 Hit rate at a given level includes hits below that level, e.g. genus hit rate includes matches at species level
<table>
<thead>
<tr>
<th>ApeKI</th>
<th>High</th>
<th>4.9</th>
<th>4.9</th>
<th>4.8</th>
<th>4.8</th>
<th>4.7</th>
<th>4.7</th>
<th>1.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>6.2</td>
<td>6.1</td>
<td>6.1</td>
<td>6.1</td>
<td>5.9</td>
<td>5.9</td>
<td>2.2</td>
</tr>
<tr>
<td>PstI</td>
<td>High</td>
<td>6.5</td>
<td>6.4</td>
<td>6.4</td>
<td>6.4</td>
<td>6.3</td>
<td>6.3</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
<td>7.4</td>
<td>7.3</td>
<td>7.3</td>
<td>2.3</td>
</tr>
</tbody>
</table>

1. RE = Restriction enzyme
2. Sample = Sample from high- or low-methane sheep
3. Hit rate at a given level includes hits below that level, e.g. genus hit rate includes matches at species level

Overview

GBS was able to capture a defined proportion of the genome of each species represented in the Hungate1000 genomes dataset when using either ApeKI or PstI restriction enzymes and we were able to find a significant difference between the proportion of reads assigned to the genus level between samples from high- and low-methane sheep, indicating that GBS is an appropriate tool for obtaining RMPs. Samples differed in hit rate at the genus level based on host methane production levels because of the taxa represented in the Hungate1000 dataset. Kittelmann et al. (2014) found that sheep with low methane production had a higher prevalence of *Sharpea* and *Kandleria* species in their rumen than sheep with high methane production. Genomes from these genera are represented in the Hungate1000 database and this partially explains the difference in hit rate between the two methane groups. Further evaluation of the potential of GBS for rumen microbial profiling could examine the impact of including additional genomes in the BLAST reference database, although this would be associated with increased computation time.

The next steps for this research are to establish how many reads are needed per sample before we have a reduced ability to distinguish between different rumen microbial profiles. The highest hit rates were obtained by the BLASTX analysis, but the runtime was deemed too long. The program DIAMOND can perform a similar analysis to BLASTX in a fraction of the time (Buchfink et al., 2015). However, a consequence of the short runtime is that there are not as many hits returned, possibly influencing accuracy. The sensitivity of DIAMOND will be evaluated in subsequent analyses. If the hit rate can be improved at the genus level using a program such as DIAMOND, the number of reads required to achieve a given sensitivity can be reduced, which will allow more samples per lane and reduced GBS cost per sample.

Acknowledgements

Financial support was provided by the AgResearch Curiosity fund, the MBIE funded Genomics for Production and Security Programme, The microbes to predict methane project funded by Global Partnerships in Livestock Emissions Research (SOW-AGR-GPLER-SP5), the Pastoral Greenhouse Gas Research Consortium (PGgRc.co.nz) and New Zealand Agricultural Greenhouse Gas Research Centre. (NZAGRC.org.nz).

List of References


Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., GRC Collaborators & Janssen, P.H., 2015. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Scientific Reports 5:14567


structure and function. ISAG Conference, Dublin, Ireland.