frontiers | Frontiers in Genetics

MINI REVIEW
published: 20 July 2022
doi: 10.3389/fgene.2022.865765

Metagenomic Predictions: A Review 10 years on

Elizabeth M Ross and Ben J Hayes*

Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane, QLD, Australia

Metagenomic predictions use variation in the metagenome (microbiome profile) to predict the unknown phenotype of the associated host. Metagenomic predictions were first developed 10 years ago, where they were used to predict which cattle would produce high or low levels of enteric methane. Since then, the approach has been applied to several traits and species including residual feed intake in cattle, and carcass traits, body mass index and disease state in pigs. Additionally, the method has been extended to include predictions based on other multi-dimensional data such as the metabolome, as well to combine genomic and metagenomic information. While there is still substantial optimisation required, the use of metagenomic predictions is expanding as DNA sequencing costs continue to fall and shows great promise particularly for traits heavily influenced by the microbiome such as feed efficiency and methane emissions.

Keywords: metagenomics, microbiome, prediction, methane, feed efficiency

OPEN ACCESS

Edited by:
Luciana Regitano, Brazilian Agricultural Research Corporation (EMBRAPA), Brazil

Reviewed by:
Juliana Petrini, University of São Paulo, Brazil
Piush Khanal, Michigan State University, United States

****Correspondence:***
Elizabeth M Ross
e.ross@uq.edu.au

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *30 January 2022*
Accepted: *01 June 2022*
Published: *20 July 2022*

Citation:
Ross EM and Hayes BJ (2022) Metagenomic Predictions: A Review 10 years on. Front. Genet. 13:865765. doi: 10.3389/fgene.2022.865765

INTRODUCTION

The host associated microbiome is known to influence many traits (**Figure 1A**), including health traits (Cho and Blaser, 2012; Rothschild et al., 2022), enteric methane production (Ross et al., 2013b; Wallace et al., 2015; Hess et al., 2020; Hess et al., 2021), feed efficiency (Wang et al., 2015; Wen et al., 2021), carcass traits (Maltecca et al., 2019) and even neurological traits (Kho and Lal, 2018). The metagenome is the cumulative genomes of the cells which make up the microbiome. Metagenomics is the study of that genome population. Metagenomic predictions use the variation in metagenomes to predict the phenotype of a host (Ross et al., 2013b). While the exact mechanisms through which microbiomes effect the host phenotype are not always known, some direct effects, such as methanogens producing methane (Morgavi et al., 2010), and some indirect effects, such as the modulation of the host immune system (Levy et al., 2017), have been identified in various species. While metagenomic predictions rely on these underlying causative effects, knowledge of them is not required for accurate metagenomic predictions, as the relationships calculated are purely mathematical, and currently do not consider biological relationships.

This brief review examines work done to explore the predictive potential of the host associated microbiome, with a focus on ruminant livestock traits.

PHENOTYPIC TRAITS

The metagenomic prediction method (**Figure 1B**) was originally inspired by genomic prediction, which use relationships between samples derived from DNA marker genotypes to predict unknown phenotypes using best linear unbiased prediction (BLUP). Metagenomic predictions were first reported in 2012 (Ross et al., 2012b) where they were used to predict sample type and inflammatory

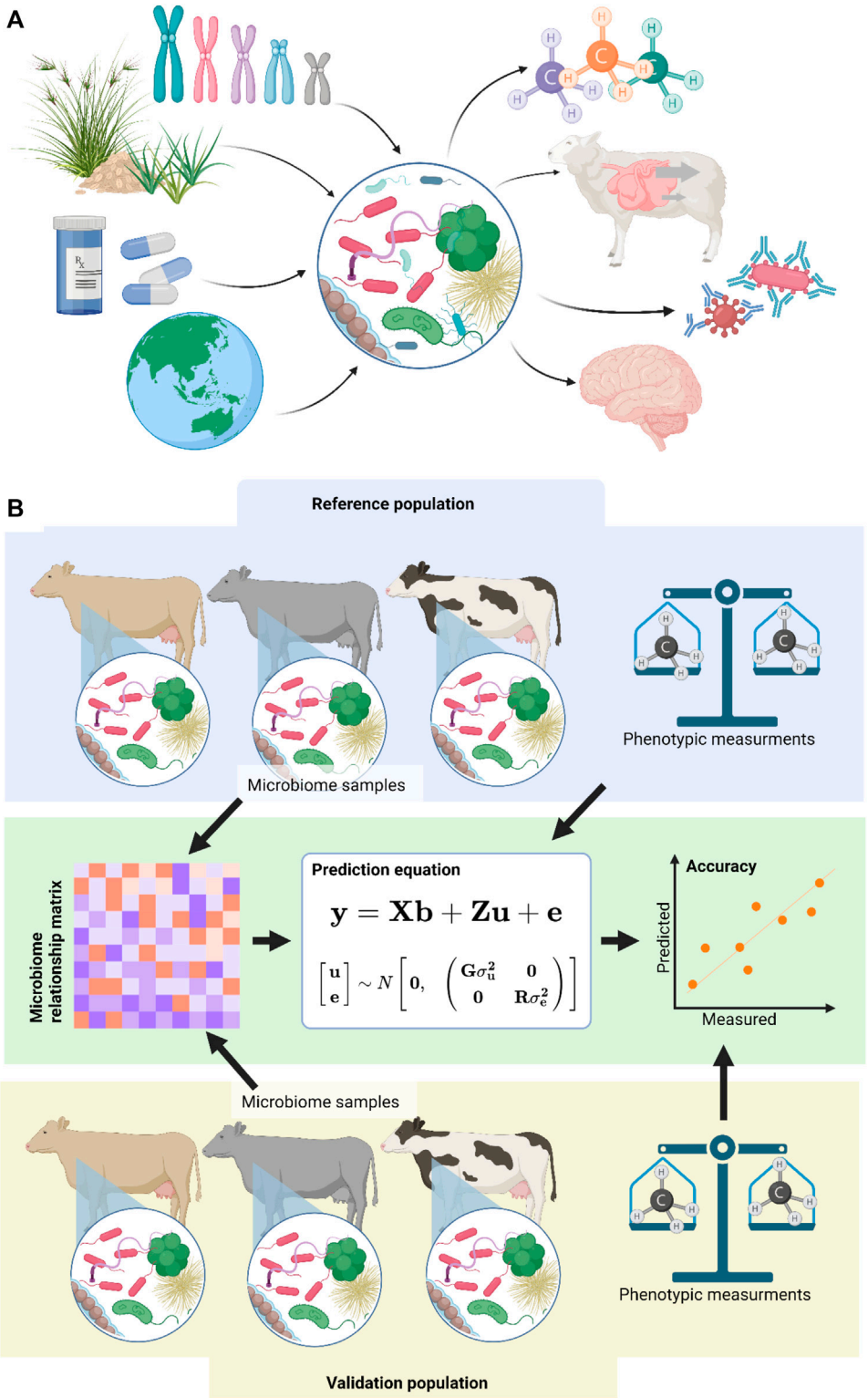

FIGURE 1 | (A) The host associated microbiome is influenced by a range of factors including host genetics, diet, drugs and medication, and physical location. In turn the host associated microbiome is thought to influence several phenotypes including enteric methane production, feed conversion efficiency, immune function, and even neurological traits. **(B)** Metagenomic predictions use a reference population of microbiome samples and measured phenotypes to predict unknown phenotypes in a difference validation population with only microbiome samples. The accuracy of the prediction can be evaluated by comparing measured phenotypes in the validation population to these predicted by the model. Image generated in BioRender.

	Taxa 1	Taxa 2	Taxa 3	Taxa 4	Taxa 5	Taxa 6	Taxa 7	Taxa 8	Taxa 9	Taxa 10	Taxa 11	Taxa 12
Animal 1	690	42	33	73	216	670	46	64	104	58	490	98
Animal 2	2475	786	1555	1124	2967	3086	555	1541	443	93	1174	2638
Animal 3	993	489	455	176	2506	1720	566	723	168	242	633	2583
Animal 4	343	455	70	467	267	196	5	307	219	10	88	278
Animal 5	762	37	1014	732	2145	1449	704	794	253	14	1618	1472
Animal 6	377	82	372	233	707	182	26	49	5	24	331	331
Animal 7	3705	855	1326	1403	3426	2620	460	1741	764	355	2224	127
Animal 8	559	39	5	1506	1961	2219	944	355	304	231	2121	2117
Animal 9	220	53	352	375	646	578	268	173	78	62	395	34
Animal 10	1762	1316	329	843	261	1466	341	1759	364	43	478	1042
Animal 11	327	27	102	252	350	128	77	398	5	34	523	121
Animal 12	2600	1623	369	1726	699	51	313	997	1042	398	1162	743
Animal 13	1835	817	1828	1294	2528	1813	71	2216	831	290	2169	914
Animal 14	406	81	70	88	618	485	195	126	22	47	271	90
Animal 15	2057	860	619	1321	2228	1147	477	1198	407	252	1443	1546
Animal 16	206	563	636	239	406	1304	375	884	132	79	77	1288
Animal 17	71	1552	1385	944	278	1536	266	2222	536	313	1509	809

FIGURE 2 | An example of the count matrix that can be used to capture the variation in the microbial population. Note that the total number of reads for each animal may vary, as in this example, and should be standardised. While this example only includes 12 taxa, many thousands of taxa can be included in the count matrix.

bowel disease status. Soon after, they were used to predict methane production levels from cattle (Ross et al., 2013b; **Table 1**), which has since been replicated in sheep (Ross et al., 2020; Hess et al., 2021).

Metagenomic predictions were subsequently used to predict residual feed intake (Wang et al., 2015). In chickens metagenomic variation of the caecum was found to be associated with residual feed intake, but not other gut locations (Wen et al., 2021). Carcass traits in pigs have been predicted with moderate to high accuracy from gut microbiomes by Maltecca et al. (2019). Recently, research in sheep used metagenomic predictions to predict methane yield in sheep in Australia (Ross et al., 2020) and in New Zealand (Hess et al., 2021). Additional studies have found further associations between microbiome variation and methane in cattle (Difford et al., 2018; Zhang et al., 2020; Andrade et al., 2022).

Methane and residual feed intake are expected to have a direct link to the gut metagenome composition, and hence a relationship between metagenome variation and phenotypic variation for these traits might be expected. On the other hand, a recent study in sheep found that the metagenome did not explain any of the phenotypic variance of dairy traits in sheep (Martinez Boggio et al., 2022), despite some rumen bacteria being associated with milk characteristics (Martinez Boggio et al., 2021). Conversely, Gebreyesus et al. (2020) found that the rumen metagenome was predictive of ketone bodies in milk, an indicator of ketosis, which has previously been associated with rumen microbiome changes (Zhu et al., 2018). This may suggest that either the link between the phenotype of interest and the microbiome needs to be particularly strong for metagenomic predictions to work, or that more sophisticated models are required.

Microbiome associated traits have also been predicted in humans. Body mass index was predicted within and across populations from gut samples (Ross et al., 2013b; Rothschild et al., 2022), as was Crohn's disease (Asgari et al., 2018) and ulcerative colitis (Ross et al., 2013b). Four other health related traits including menopausal status and smoking status were linked to skin microbiome variation in Carrieri et al. (2021), while Rothschild et al. (2022) predicted a number of traits including smoking status and type II diabetes. Overall, it is apparent that microbiome variation is associated with a large range of host phenotype traits, although the exact causal relationship is not always clear-cut.

THE COUNTS MATRIX

All metagenomic predictions begin with a "counts matrix" which attempts to approximate the proportion of different taxa in each animal's microbiome (**Figure 2**). This is challenging given many of the species in the microbiome are unknown. Originally, metagenomic predictions used short read shotgun sequences aligned to a reference genome of microbial species (or sequence assemblies) to approximate the proportion of different taxa (Ross et al., 2013b). This approach should be more accurate, for cattle at least, now that a good proportion of the rumen microbes have been fully sequenced (e.g., Seshadri et al., 2018).

Methods which use 16S sequencing and reduced representation sequencing have also been successfully used in metagenome predictions (Hess et al., 2020; Ross et al., 2020), although if a species is not represented in the 16S database, or not captured by the selected primer, it will not appear in the counts matrix. Another approach, used by Maltecca et al. (2019), aligned sequence reads to operational taxonomic units (OTUs). Other methods that provide the ability to classify reads such that they represent some sort of taxonomical groupings could equally be implemented for the generation of the relationship matrix, such as amplicon sequence variants (Callahan et al., 2017), or any other method that achieves the same end point of a count matrix which is able to capture relative changes of microbial abundances.

Once the count matrix that represents different microbial species abundance is formed, a co-variance matrix was calculated from the count matrix which was used to predict phenotypes in a non-overlapping group of individuals. The

TABLE 1 | Summary of studies which have used rumen metagenomic profiles to predict phenotypes in ruminants.

Study	Species (N#)	Phenotype	Method	Within or between Countries	Accuracy	
					Microbiome Only	Microbiome and Genome
Ross et al. (2013a)	Dairy cattle (62)	Enteric methane	Co-variance matrix and BLUP	Within	$<0^{NS}$–0.79	-
			Random Forests	Within	0.33	-
Wang et al. (2015)	Dairy cattle (28)	Residual feed intake	Co-variance matrix and BLUP	Within	0.08^{NS} - 0.49	0.38–0.57
Delgado et al. (2019)	Dairy cattle (61)	Feed efficiency	Linear effects	Between	0.19	-
		Dry matter intake	Linear effects	Between	0.39	-
Ross et al. (2020)	Sheep (99)	Enteric methane	Co-variance matrix and BLUP with microbiome	Within	$<0^{NS}$–0.14	0^{NS}–0.25
			Co-variance matrix and BLUP with metabolome	Within	0.13–0.25	0.16–0.27
Hess et al. (2020)	Sheep (340)	Enteric methane	Principle component analysis	Within	0.17–0.51*	-
Hess et al. (2021)	Sheep (1702)	Enteric methane	Correlation matrix and BLUP	Between	$<0^{NS}$ - 0.13^{NS}	$<0^{NS}$ - 0.13^{NS}
			Correlation matrix and BLUP	Within	0.40–0.57	0.53–0.60

#Number of animals used in the entire study, including both reference and validation populations.

**Not cross validated.*

^{NS}Not significantly different to 0.

methods above were the starting point for metagenomic predictions, but future work will likely find that there are more optimal approaches. For example, as more and more rumen microbes have their genomes completely sequenced, alignment at a species level becomes possible.

METAGENOMIC RELATIONSHIP MATRIX CALCULATION METHODS

The original metagenomic prediction method used a relationship co-variance matrix among animals that was calculated as **XX**'/m, where **X** is the (standardised) count matrix described above and m is the number of contigs or taxa used to make the count matrix. Subsequently (Hess et al., 2020; Hess et al., 2021) used a correlation matrix to overcome convergence issues. Earlier work (Ross et al., 2012a) used the Canberra method (Lance and Williams, 1966) to generate a distance matrix. Other distance methods should be explored in the larger emerging datasets to optimise the representation of microbiome similarities, and improve convergence of the models.

The relationship matrix can be used in a best linear unbiased prediction approach (BLUP) to predict trait performance for individual animals. If the relationship matrix is appropriately standardised, The BLUP methods assume rare species contribute equally to the relationship matrix as highly abundant species.

The BLUP method begins with fitting a linear mixed model to the data:

$$\boldsymbol{y} = \boldsymbol{W\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

Where **y** is vector of phenotypic records, **W** is a design matrix allocating records to fixed effects such as sex, age, cohort, **β** are the values for these fixed effects, **Z** is a design matrix allocating records to individuals, and **e** is a vector of random errors. The **u** are random effects for individuals assumed distributed $N(0, \boldsymbol{MRM}\sigma_m^2)$, where **MRM** is the metagenome relationship matrix, and σ_m^2 is the variance in the trait associated with the metagenome. Note that with model, and an appropriate experimental design, the effects of sex, age, cohort and so on can be disentangled from metagenome effects. This equation can be solved for metagenome predictions ($\hat{\boldsymbol{u}}$) for each individual using BLUP, and also for the effect of each individual species/OTU/contig in the metagenome as

$$\hat{\boldsymbol{g}} = \boldsymbol{X'MRM}^{-1}\hat{\boldsymbol{u}}/m$$

While the effects of differential weighting based on abundance have not been explored, one hypothesis could be that similarities within the metagenomic relationship matrix should be weighted by species abundance. This would also reduce the effect of random variation in the less abundant species affecting the observed relationships, especially with the low sequencing depth that is required for larger scale use of metagenomic predictions of phenotypes.

Something that is not easily captured by the co-variance relationship model is non-linear effects. Some machine learning methods can capture non-linear effects, however the number of samples required is large since the effect distribution needs to be estimated from the data. In genomic predictions of complex genomes non-linear effects can be captured by manipulation of the relationship matrix to represent interactions and non-linear patterns, for example a an organism with low abundance might have an effect with small changes in abundance, whereas an organism with high abundance might only have an effect with large changes in abundance. Including non-linear effects have resulted in an increased accuracy for some traits in genomic prediction

(Yadav et al., 2021). A similar approach could be applied to microbiome samples to overcome this limitation, however the proportion of non-linear effects, and whether they in fact matter at all, is currently unknown.

A limitation of the relationship matrix-based prediction approach is that all metagenome species are assumed to have a small, but non-zero effect on the trait. . An advantage of relationship-based models is that they are feasible with smaller datasets, as thousands of effects do not need to be estimated from the data. As sequence generation costs continue to fall, the limitation of sample sizes may soon be resolved and non-relationship matrix-based prediction methods that are able to place more emphasis on associated species may prove more appropriate for metagenomic predictions.

PREDICTION METHODS

The limitations for relationship matrix-based predictions bring us to the next logical step–the use of prediction models that allow different weightings on different features, including the possibility of zero effect. In genomic predictions, given sufficient reference set size, Bayesian prediction models such as BayesR outperform relationship-based models. The first step toward using such methods in metagenome predictions this has already been completed by Zhang et al. (2020) who used Bayesian methods to examine methane variance explained by the metagenome in dairy cattle. In genomic prediction, methods such as BayesRC have been used to include biological priors to increase the accuracy of prediction further from SNP data (Macleod et al., 2016). These methods could be directly applied to large metagenomic datasets to allow species that are known to be associated with the trait of interest to be more highly weighted in the model. For example, taxa in studies that have been correlated with the target trait, such as the ~500 taxa that Delgado et al. (2019) used for prediction, could be treated as a separate class in a BayesRC type approach.

Another option would be to increase the weighting (or treat as a separate class in BayesRC) species which contain the genes that are used in the relevant biological pathway. For example, for the prediction of enteric methane production, species which contain the methanogenesis pathway, or alternate hydrogen sinks (for examples see Morgavi et al., 2010) such as propionate formation, could be given larger priors in the same manner that SNP with biological priors can be more heavily weighted in some Bayesian prediction models. These taxa could be identified by mining the genome ontology terms in the fully sequenced rumen bacterial genome assemblies or the metagenome associated genomes.

A number of studies have proposed the use of machine learning for phenotypic prediction from metagenomes, mostly in humans and in pigs (e.g., Maltecca et al., 2019). These methods were recently reviewed by Marcos-Zambrano et al. (2021). At least in pigs, machine learning approaches such as random forest and gradient boosting, gave similar accuracies of prediction as BLUP. Some other recent examples of the use of machine learning to predict phenotypes from the microbiome include Asgari et al. (2018); Lo and Marculescu (2019); Fukui et al. (2020); and Carrieri et al. (2021). Carrieri et al. (2021) used the skin microbiome to predict a range of phenotypes in humans by applying explainable artificial intelligence. Reflux disorders were predicted by Lo and Marculescu (2019), and inflammatory bowel disease was a mutual target of a number of studies (Asgari et al., 2018; Lo and Marculescu, 2019; Fukui et al., 2020). These approaches illustrate the power of machine learning for phenotypic prediction when the dataset is large, however a direct comparison of these methods with BLUP based predictions has not been well examined. Ross et al. (2013b) compared BLUP and random forests in the same dataset, with BLUP outperforming random forests in both animal and human associated microbiomes. Rothschild et al. (2022) compared ridge regression-based predictions to gradient boosted decision trees, and found that the two methods were mostly comparable, but gradient boosted decision trees outperformed the regression on binary traits. Given the computational expense of machine learning, a significant benefit in terms of prediction accuracy would be required to justify their use over more basic methods, which is likely dependant on sample size.

COMBINING METAGENOMIC AND OTHER PREDICTION SYSTEMS

A handful of studies have examined the effect of combining genomic and metagenomic predictions. The first study to do so was Wang et al. (2015) who combined SNP data and metagenomic data in a small study to predict residual feed intake with higher accuracy than either method alone. While not used for metagenomic predictions, Difford et al. (2018) found that the variance explained by both the genomic and metagenomic data for dairy cows was greater than when either one was examined alone, as did Zhang et al. (2020) on the same dataset using a Bayesian method. Saborío-Montero et al. (2021) concluded that not only are both the genome and metagenome important for explaining the phenotypic variation, but that the interaction between genome and metagenome is also important. Recently, Ross et al. (2020) combined metagenomic and genomic predictions for studying enteric methane production in sheep. The study had a limited number of biological replicates ($N = 99$) but illustrated a proof of principle that the accuracy of the phenotypic prediction of enteric methane production was increased when metagenomic, or metabolomic, predictions from the rumen were included. Subsequently this finding was validated in a much larger cohort of animals ($N = 1702$) by Hess et al. (2021). Recently in pigs (Aliakbari et al., 2022) the accuracy of prediction for a number of traits including residual feed intake and back fat depth were shown to increase when both genetic and microbiome information was used in the prediction model.

The metagenome itself has some heritable components (Wallace et al., 2019; Abbas et al., 2020; Grieneisen et al., 2021; Cardinale and Kadarmideen, 2022), with heritabilities of individual genera up to 0.59 (Martínez-Álvaro et al., 2022). Therefore, there is expected to be overlap when selecting either based on metagenomic and genomic prediction values.

The extent of host-metagenome interaction is important to consider when metagenomic predictions are used in selection. If there is no host genome -metagenome interaction, such selection may shift the current population mean (e g., through culling), but will not result in genetic improvement Conversely, if host-metagenome interaction is extensive, selection on metagenome predictions will result in genetic gain. Host genome -metagenome interaction could also explain why the accuracy of genomic and metagenomic predictions is not fully additive, that is, they are not detecting independent factors.

The additional accuracy observed when combining genomic and metagenomic information of the phenotype prediction accuracy is probably partially due to the metagenomic predictions capturing part of the environmental variation. The environmental variation is, by definition, not captured by genomic predictions. It is important to understand the interactions between different genomic and metagenomic predictions of a trait. Each trait is likely to have a unique profile of genomic, metagenomics and uncaptured environmental variation, that needs to be understood through experimentation. Key to this understanding is that unlike in genomic predictions where the aim is to predict the heritable component of the phenotype, metagenomic predictions usually aim to predict the phenotype itself.

PREDICTION PITFALLS–CAUSE AND EFFECT

A careful interpretation of the values generated by metagenomic predictions is needed. Critically, as opposed to genomic predictions, the effect direction of metagenomic predictions is not necessarily known. That is: is the microbiome affecting the phenotype, or is the phenotype affecting the microbiome? In the case of enteric methane production, it is most likely that the microbiome is affecting the phenotype, as there is not a documented mechanism for methane concentration to affect microbiome composition, but the mechanism for microbiomes to affect methane production is well understood. Not all traits are so clear-cut.

We can take lessons in the directionality of the metagenome's effects not only from livestock research, but also from human and medical research. A recent study has illustrated the pitfalls of assuming that microbiome differences are causing host phenotype variation in humans. Yap et al. (2021) examined the link between autism spectrum disorder, diet, and the microbiome. They concluded that although variation in the microbiome is associated with autism spectrum disorder, it is not the cause. Rather, dietary preferences and limitations that are caused by autism spectrum disorder affect the microbiome composition. Therefore, it is the phenotype affecting the microbiome, not the other way around. Understanding this is critical for the correct use of metagenomic predictions, where the misuse of the information, such as attempting to alter the microbiome to reduce symptoms associated with autism spectrum disorder would be detrimental to the patients without any benefit.

POPULATION DIFFERENCES

Another challenge facing the use of metagenomic predictions is the effect of environment and location. Hess et al. (2021) showed that sheep from New Zealand could not be used to predict methane in sheep from Australia, but that within country predictions were successful. Conversely, Delgado et al. (2019) used a Spanish dairy herd to predict the feed efficiency of an Australian dairy herd with success. A cause behind the phenomena that geographically separate populations may show poorer prediction accuracy than expected given their relationships in the relationship matrix is that there may be strain level differences in the species that make up the microbiome. Different geographical regions may have strains of bacteria that carry a different subset of genes compared to those found in other locations. This could result in a breakdown between the association with the trait in the reference population, and the prediction ability in the validation population because the sequence that is being quantified is not connected to the same causal gene in both populations.

There is also the possibility that not all causal agents exist in all populations. This is equivalent to having fixed alleles in genomic predictions, where there is no variation in the genome/metagenome at that position in the discovery population, and so it is not used in the prediction even if it is present in the validation population. Where low across-population prediction accuracies are observed it may be that it is only possible to overcome this hurdle by the inclusion of phenotyped individuals from the same location as the target population.

MEASURING ACCURACY AND MICROBIABILITY

Prediction accuracies for metagenomic predictions of continuous traits are generally reported as the Pearson's correlation (r) between the predicted and the observed phenotype of the validation set for continuous traits. This is opposed to genomic predictions where r is scaled by the heritability of the trait by dividing by the square root of the narrow sense heritability. Analogous to the heritability is the microbiability. The microbiability is the proportion of the variance in the phenotype that can be attributed to the metagenomic relationship matrix. The microbiability however has substantial limitations including that it does not capture non-additive relationships, as pointed out by Rothschild et al. (2022).

The microbiability varies considerably across traits, for reasons described above. For example Aliakbari et al. (2022) gave estimates of 0.11 residual feed intake, 0.20 feed conversion ratio, and 0.02 backfat in pigs, while He et al. (2022) reported 0.42 for back fat. Hess et al. (2020) used both 16S sequencing, and reduced representation using restriction enzyme sequencing to calculate the microbiability for methane emission level. They revealed two things: that there is a substantial difference in the microbiability of the same dataset based on whether the data was derived from 16S or reduced representation sequencing; and also that the

restriction enzyme chosen for reduced representation sequencing had a large impact on the microbiability. Saborío-Montero et al. (2021) found that the method used to calculate the microbiome relationships also affects the microbiability. This would suggest that the microbiability is strongly reflective of the method used and thus any comparison of microbiability between studies should be done with extreme caution. It also suggests that this measure could be a useful tool to compare methods.

PARAMETERS AFFECTING ACCURACY OF METAGENOME PREDICTIONS

The microbiability is one parameter affecting the accuracy of metagenomic predictions–the higher this is (the greater proportion of the phenotypic variance captured by microbiome variation), the higher the accuracy of prediction.

Sample size is another key parameter. Prediction methods are limited by the number of samples that are available for use. The first metagenomic predictions had very limited biological replicates available and far too few to estimate effects for each species individually. Sequencing costs have plummeted as technology has advanced, and new methods of metagenomic profiling have become available (e.g., Hess et al., 2020). Thus, the limitation on sample numbers has moved from the sequencing cost to the cost of phenotyping, especially for traits that are expensive or difficult to measure such as enteric methane production. One exception is the study of Rothschild et al. (2022) where more than 30,000 samples were used in metagenomic predictions. That study demonstrated that accuracy plateaued with approximately 4,000 samples for most traits. For example, for BMI, the r value (the square root of the reported r^2) was approximately 0.32 for 2000 samples, 0.36 at 4,000 samples, and 0.38 for 8,000 samples. Thus, although increases in accuracy continue to be observed, there are diminishing returns as the sample same increases.

Building on theory that was developed to deterministically predict the accuracy of genomic selection with BLUP models (Daetwyler et al., 2008; Hayes et al., 2009), we would expect the accuracy of metagenome predictions to be, for BLUP predictions at least:

$$r = \sqrt{\frac{N_P m^2}{N_P m^2 + P}}$$

Where N_p is the number of samples with phenotypes, m^2 is the microbiability, and P is the number of independent entities in the microbiome population. Approximations of P could be the number of OTUs, or the number of principal components required to capture >99% of the variance in the metagenome relationship matrix described above (e.g., Kittelmann et al., 2014).

In deriving the accuracy of metagenomic predictions, it is important to note that the microbiability changes with time. For example, Maltecca et al. (2019) found that the accuracy of metagenomic predictions for backfat (in pigs at 22 weeks of age) from samples at weaning were lower than from samples taken at the same time as phenotyping. For example, accuracy of prediction of back fat increased from $r = 0.42$ when microbiome samples from weaning to $r = 0.48$ when microbiome samples were taken at week 22.

UTILITY OF METAGENOMIC PREDICTIONS

The most basic use of metagenomics is the direct inference of the phenotype. Such direct inference could be used for direct selection, or diagnosis for intervention. For example, metagenomic predictions could be used to identify and remove high methane emitting cattle from a herd to lower a producer's overall carbon footprint. Metagenomic predictions could also be used to select breeding animals with favourable traits such as high feed conversion efficiency (provided there was considerable host genome–metagenome interaction), or to diagnose conditions which may cause a shift in rumen ecology, such as sub-acute acidosis.

The microbiome could also be used for genomic selection in the future by generating proxy traits. Proxy traits are traits which approximate the true trait of interest. For example, metagenomic predictions could be used to generate predicted methane emission levels for large numbers of cattle. Those cattle could then be genotyped and genomic estimated breeding values for metagenomic-methane proxy traits could be calculated. Selection pressure could then be applied through breeding from the most desirable animals. Given that there have been several studies which have identified that there is a heritable aspect to the rumen metagenome, at least some of the changes to a low methane rumen should be able to be inherited, resulting on the ability to select for low methane emitting animals. This could be a useful approach for any trait where the microbiome is easier to measure than the trait itself, of which methane is a key example.

The development of metagenomic predictions to rapidly build large databases of proxy phenotypes to develop genomic breeding values would be enabled if the target microbiome is optimised for ease of sampling. In the case of enteric methane production, proxy databases could be generated using rumen metagenome samples, which are quicker and cheaper to obtain then direct phenotyping or could be obtained from saliva-based microbiomes. Tapio et al. (2016) investigated this using qPCR (quantitative polymerase chain reaction) and concluded that buccal swabs could be used as a predictor of the rumen microbiome population. The hypothesis behind this assertion is that as the animals ruminate, they deposit the rumen bacterial population in the mouth. This could provide a more user-friendly method of microbiome collection than currently available from the rumen itself.

CONCLUSION

Metagenomic predictions can be used to predict the phenotype of traits that are associated with microbiome variation. Their use is still in its infancy with many areas left to explore and optimise. With large sample numbers now able to be sequenced, metagenomic predictions offer an opportunity for use as proxy traits that can take

to place of challenging phenotypes that are expensive and/or difficult to measure on large numbers of individuals, such as enteric methane from ruminants. Future work should focus on dramatically increasing the size of the populations being studied. Testing new machine learning based prediction methods will become possible as the size of datasets increases. The anticipated outcome of larger populations with optimised predictions methods will be more accurate predictions that can be implemented by industry as proxy phenotypes for selection and culling.

AUTHOR CONTRIBUTIONS

ER and BH conceptualised and wrote the article.

FUNDING

This project was funded by Meat and Livestock Australia (P.PSH. 2010) and the Queensland Department of Agriculture and Fisheries.

ACKNOWLEDGMENTS

The authors are thankful to the Queensland Alliance for Agriculture and Food Innovation Centre for Animal Science staff and students for support and discussions. The authors thank Dr. Bailey Engle for grammatical editing of the manuscript, and ongoing encouragement and support.

REFERENCES

Abbas, W., Howard, J. T., Paz, H. A., Hales, K. E., Wells, J. E., Kuehn, L. A., et al. (2020). Influence of Host Genetics in Shaping the Rumen Bacterial Community in Beef Cattle. *Sci. Rep.* 10, 15101. doi:10.1038/s41598-020-72011-9

Aliakbari, A., Zemb, O., Cauquil, L., Barilly, C., Billon, Y., and Gilbert, H. (2022). Microbiability and Microbiome-wide Association Analyses of Feed Efficiency and Performance Traits in Pigs. *Genet. Sel. Evol.* 54, 29. doi:10.1186/s12711-022-00717-7

Andrade, B., Bressani Donatoni, F., Cuadrat, R., Cardoso, T. F., Malheiros, J., Oliveira, P., et al. (2022). Stool and Ruminal Microbiome Components Associated with Methane Emission and Feed Efficiency in Nelore Beef Cattle. *Front. Genet.* doi:10.3389/fgene.2022.812828

Asgari, E., Garakani, K., Mchardy, A. C., and Mofrad, M. R. K. (2018). MicroPheno: Predicting Environments and Host Phenotypes from 16S rRNA Gene Sequencing Using a K-Mer Based Representation of Shallow Sub-samples. *Bioinformatics* 34, i32–i42. doi:10.1093/bioinformatics/bty296

Callahan, B. J., Mcmurdie, P. J., and Holmes, S. P. (2017). Exact Sequence Variants Should Replace Operational Taxonomic Units in Marker-Gene Data Analysis. *Isme J.* 11, 2639–2643. doi:10.1038/ismej.2017.119

Cardinale, S., and Kadarmideen, H. N. (2022). Host Genome-Metagenome Analyses Using Combinatorial Network Methods Reveal Key Metagenomic and Host Genetic Features for Methane Emission and Feed Efficiency in Cattle. *Front. Genet.* 13, 795717. doi:10.3389/fgene.2022.795717

Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Gardiner, L.-J., Murphy, B., Mayes, A. E., et al. (2021). Explainable AI Reveals Changes in Skin Microbiome Composition Linked to Phenotypic Differences. *Sci. Rep.* 11. doi:10.1038/s41598-021-83922-6

Cho, I., and Blaser, M. J. (2012). The Human Microbiome: at the Interface of Health and Disease. *Nat. Rev. Genet.* 13, 260–270. doi:10.1038/nrg3182

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-wide Approach. *PLOS One* 3, e3395. doi:10.1371/journal.pone.0003395

Delgado, B., Bach, A., Guasch, I., González, C., Elcoso, G., Pryce, J. E., et al. (2019). Whole Rumen Metagenome Sequencing Allows Classifying and Predicting Feed Efficiency and Intake Levels in Cattle. *Sci. Rep.* 9, 11. doi:10.1038/s41598-018-36673-w

Difford, G. F., Plichta, D. R., Løvendahl, P., Lassen, J., Noel, S. J., Højberg, O., et al. (2018). Host Genetics and the Rumen Microbiome Jointly Associate with Methane Emissions in Dairy Cows. *PLoS Genet.* 14, e1007580. doi:10.1371/journal.pgen.1007580

Fukui, H., Nishida, A., Matsuda, S., Kira, F., Watanabe, S., Kuriyama, M., et al. (2020). Usefulness of Machine Learning-Based Gut Microbiome Analysis for Identifying Patients with Irritable Bowels Syndrome. *Jcm* 9, 2403. doi:10.3390/jcm9082403

Gebreyesus, G., Difford, G. F., Buitenhuis, B., Lassen, J., Noel, S. J., Højberg, O., et al. (2020). Predictive Ability of Host Genetics and Rumen Microbiome for Subclinical Ketosis. *J. Dairy Sci.* 103, 4557–4569. doi:10.3168/jds.2019-17824

Grieneisen, L., Dasari, M., Gould, T. J., Björk, J. R., Grenier, J.-C., Yotova, V., et al. (2021). Gut Microbiome Heritability Is Nearly Universal but Environmentally Contingent. *Science* 373, 181–186. doi:10.1126/science.aba5483

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009). Increased Accuracy of Artificial Selection by Using the Realized Relationship Matrix. *Genet. Res.* 91, 47–60. doi:10.1017/s0016672308009981

He, Y., Tiezzi, F., Jiang, J., Howard, J. T., Huang, Y., Gray, K., et al. (2022). Use of Host Feeding Behavior and Gut Microbiome Data in Estimating Variance Components and Predicting Growth and Body Composition Traits in Swine. *Genes* 13, 767. doi:10.3390/genes13050767

Hess, M. K., Donaldson, A., Henry, H. M., Robinson, D. L., Hess, A. S., Mcewan, J. C., et al. (2021). "Across-country Prediction of Methane Emissions Using Rumen Microbial Profiles," in Proceedings of the Association for the Advancement of Animal Breeding and Genetics, 2nd-4th November, 2021, 163–166.

Hess, M. K., Rowe, S. J., Van Stijn, T. C., Henry, H. M., Hickey, S. M., Brauning, R., et al. (2020). A Restriction Enzyme Reduced Representation Sequencing Approach for Low-Cost, High-Throughput Metagenome Profiling. *PLOS One* 15, e0219882. doi:10.1371/journal.pone.0219882

Kho, Z. Y., and Lal, S. K. (2018). The Human Gut Microbiome - A Potential Controller of Wellness and Disease. *Front. Microbiol.* 9, 1835. doi:10.3389/fmicb.2018.01835

Kittelmann, S., Pinares-Patiño, C. S., Seedorf, H., Kirk, M. R., Ganesh, S., Mcewan, J. C., et al. (2014). Two Different Bacterial Community Types Are Linked with the Low-Methane Emission Trait in Sheep. *PLOS One* 9, e103171. doi:10.1371/journal.pone.0103171

Lance, G. N., and Williams, W. T. (1966). Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses"). *Comput. J.* 9, 60–64. doi:10.1093/comjnl/9.1.60

Levy, M., Blacher, E., and Elinav, E. (2017). Microbiome, Metabolites and Host Immunity. *Curr. Opin. Microbiol.* 35, 8–15. doi:10.1016/j.mib.2016.10.003

Lo, C., and Marculescu, R. (2019). MetaNN: Accurate Classification of Host Phenotypes from Metagenomic Data Using Neural Networks. *BMC Bioinforma.* 20. doi:10.1186/s12859-019-2833-2

Macleod, I. M., Bowman, P. J., Vander Jagt, C. J., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting Biological Priors and Sequence Variants Enhances QTL Discovery and Genomic Prediction of Complex Traits. *BMC Genomics* 17, 144. doi:10.1186/s12864-016-2443-6

Maltecca, C., Lu, D., Schillebeeckx, C., Mcnulty, N. P., Schwab, C., Shull, C., et al. (2019). Predicting Growth and Carcass Traits in Swine Using Microbiome Data and Machine Learning Algorithms. *Sci. Rep.* 9, 6574. doi:10.1038/s41598-019-43031-x

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., et al. (2021). Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. *Front. Microbiol.* 12, 634511. doi:10.3389/fmicb.2021.634511

Martinez Boggio, G., Christensen, O. F., Legarra, A., Allain, C., Meynadier, A., and Marie-Etancelin, C. (2022). "Rumen Bacteria Do Not Provide Improved

Genetic Evaluation of Dairy Traits in Sheep," in Proceedings of the 12th World Congress on Genetics Applied to Livestock Production (Rotterdam, Netherlands: WCGALP.

Martinez Boggio, G., Meynadier, A., Daunis-I-Estadella, P., and Marie-Etancelin, C. (2021). Compositional Analysis of Ruminal Bacteria from Ewes Selected for Somatic Cell Score and Milk Persistency. *PLOS One* 16, e0254874. doi:10.1371/journal.pone.0254874

Martínez-Álvaro, M., Auffret, M. D., Duthie, C. A., Dewhurst, R. J., Cleveland, M. A., Watson, M., et al. (2022). Bovine Host Genome Acts on Rumen Microbiome Function Linked to Methane Emissions. *Commun. Biol.* 5, 350. doi:10.1038/s42003-022-03293-0

Morgavi, D. P., Forano, E., Martin, C., and Newbold, C. J. (2010). Microbial Ecosystem and Methanogenesis in Ruminants. *Animal* 4, 1024–1036. doi:10.1017/s1751731110000546

Ross, E. M., Hayes, B. J., Tucker, D., Bond, J., Denman, S. E., and Oddy, V. H. (2020). Genomic Predictions for Enteric Methane Production Are Improved by Metabolome and Microbiome Data in Sheep (*Ovis aries*). *J. Anim. Sci.* 98, skaa262. doi:10.1093/jas/skaa262

Ross, E. M., Moate, P. J., Bath, C. R., Davidson, S. E., Sawbridge, T. I., Guthridge, K. M., et al. (2012a). High Throughput Whole Rumen Metagenome Profiling Using Untargeted Massively Parallel Sequencing. *BMC Genet.* 13, 53. doi:10.1186/1471-2156-13-53

Ross, E. M., Moate, P. J., and Hayes, B. J. (2012b). *Toward Using Rumen Metagenomic Profiles to Predict Methane Emissions from Dairy Cows*. Melbourne, Australia: Tallygaroopna: Australasian Dairy Science Symposium.

Ross, E. M., Moate, P. J., Marett, L. C., Cocks, B. G., and Hayes, B. J. (2013b). Metagenomic Predictions: from Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle. *PLoS One* 8, e73056. doi:10.1371/journal.pone.0073056

Ross, E. M., Moate, P. J., Marett, L., Cocks, B. G., and Hayes, B. J. (2013a). Investigating the Effect of Two Methane-Mitigating Diets on the Rumen Microbiome Using Massively Parallel Sequencing. *J. Dairy Sci.* 96, 6030–6046. doi:10.3168/jds.2013-6766

Rothschild, D., Leviatan, S., Hanemann, A., Cohen, Y., Weissbrod, O., and Segal, E. (2022). An Atlas of Robust Microbiome Associations with Phenotypic Traits Based on Large-Scale Cohorts from Two Continents. *PLOS One* 17, e0265756. doi:10.1371/journal.pone.0265756

Saborío-Montero, A., Gutiérrez-Rivas, M., López-García, A., García-Rodríguez, A., Atxaerandio, R., Goiri, I., et al. (2021). Holobiont Effect Accounts for More Methane Emission Variance Than the Additive and Microbiome Effects on Dairy Cattle. *Livest. Sci.* 250, 104538.

Seshadri, R., Leahy, S. C., Leahy, S. C., Attwood, G. T., Teh, K. H., Lambie, S. C., et al. (2018). Cultivation and Sequencing of Rumen Microbiome Members from the Hungate1000 Collection. *Nat. Biotechnol.* 36, 359–367. doi:10.1038/nbt.4110

Tapio, I., Shingfield, K. J., Mckain, N., Bonin, A., Fischer, D., Bayat, A. R., et al. (2016). Oral Samples as Non-invasive Proxies for Assessing the Composition of the Rumen Microbial Community. *PLOS One* 11, e0151220. doi:10.1371/journal.pone.0151220

Wallace, R. J., Sasson, G., Garnsworthy, P. C., Tapio, I., Gregson, E., Bani, P., et al. (2019). A Heritable Subset of the Core Rumen Microbiome Dictates Dairy Cow Productivity and Emissions. *Sci. Adv.* 5, eaav8391. doi:10.1126/sciadv.aav8391

Wallace, R. J., Rooke, J. A., Mckain, N., Duthie, C.-A., Hyslop, J. J., Ross, D. W., et al. (2015). The Rumen Microbial Metagenome Associated with High Methane Production in Cattle. *BMC Genomics* 16, 839. doi:10.1186/s12864-015-2032-0

Wang, M., Pryce, J. E., Savin, K., and Hayes, B. J. (2015). "Prediction of Residual Feed Intake from Genome and Metagenome Profiles in First Lactation Holstein-Friesian Dairy Cattle," in Proceedings of the Association for the Advancement of Animal Breeding and genetics, Lorne, Victoria, Australia, 28-30 September 2015, 89–92.

Wen, C., Yan, W., Mai, C., Duan, Z., Zheng, J., Sun, C., et al. (2021). Joint Contributions of the Gut Microbiota and Host Genetics to Feed Efficiency in Chickens. *Microbiome* 9, 126. doi:10.1186/s40168-021-01040-x

Yadav, S., Wei, X., Joyce, P., Atkin, F., Deomano, E., Sun, Y., et al. (2021). Improved Genomic Prediction of Clonal Performance in Sugarcane by Exploiting Non-additive Genetic Effects. *Theor. Appl. Genet.* 134, 2235–2252. doi:10.1007/s00122-021-03822-1

Yap, C. X., Henders, A. K., Alvares, G. A., Wood, D. L. A., Krause, L., Tyson, G. W., et al. (2021). Autism-related Dietary Preferences Mediate Autism-Gut Microbiome Associations. *Cell* 184, 5916–5931. e5917. doi:10.1016/j.cell.2021.10.015

Zhang, Q., Difford, G., Sahana, G., Løvendahl, P., Lassen, J., Lund, M. S., et al. (2020). Bayesian Modeling Reveals Host Genetics Associated with Rumen Microbiota Jointly Influence Methane Emission in Dairy Cows. *Isme J.* 14, 2019–2033. doi:10.1038/s41396-020-0663-x

Zhu, Z., Kristensen, L., Difford, G. F., Poulsen, M., Noel, S. J., Abu Al-Soud, W., et al. (2018). Changes in Rumen Bacterial and Archaeal Communities over the Transition Period in Primiparous Holstein Dairy Cows. *J. Dairy Sci.* 101, 9847–9862. doi:10.3168/jds.2017-14366

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.