High-resolution QTL mapping with Diversity Outbred mice identifies genetic variants that impact gut microbiome composition

Florencia Schlamp, David Y Zhang, Elissa Cosgrove, Petr Simecek, Matthew Edwards, Julia K Goodrich, Ruth E Ley, Gary A Churchill, Andrew G Clark.

RL: Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen 72026, Germany.

1. ABSTRACT

The composition of the gut microbiome is impacted by a complex array of factors, from nutrient composition and availability, to physical factors like temperature, pH and flow rate, as well as interactions among the members of the microbial community. Many of these factors are affected by the host, raising the question of how host genetic variation impacts microbiome composition. Human studies confirm a role for host genetics, but opinions vary on just how important is host genetic variation in determining gut microbiome composition. The mouse model, by allowing far better control of genetics, nutrition, and other environmental factors, has provided an excellent opportunity to extend this work, and the Diversity Outbred (DO) mice in particular present a chance for mapping host genetic variants that influence specific attributes of microbiome composition. Here we apply 16s sequencing to fecal samples of 247 DO mice and perform QTL mapping on microbial abundances. In addition to finding a number of novel associations, the concordance with heritabilities and associations with both human and mouse studies was striking, including the phylum Tenericutes, family Ruminococcaceae, as well as Staphylococcus and Turicibacter.

2. INTRODUCTION

The gastrointestinal tract of all vertebrates, including humans, harbors a complex ecological community of highly diverse microbes referred to as the gut microbiota. The microbiota colonizes the gut for the first time during the birth of the host and its composition is influenced by many factors during the host's life such as disease, diet, and antibiotics [Dash *et al.* 2019, Battaglioli *et al.* 2018, Dudek-Wicher *et al.* 2018, Francino 2015]. Variation in the human gut microbiome composition has also already been associated with host immune responses [Round and Mazmanian 2009, Veiga *et al.* 2010, Garrett *et al.* 2010], metabolic phenotypes [Turnbaugh *et al.* 2009, Ridaura *et al.* 2013], and diseases such as obesity [Ley *et al.* 2005], heart disease [Fava 2006], and diabetes [Wen *et al.* 2008]. Given the roles of the gut microbiome in complex human diseases, it is important to characterize the factors that impact microbiome composition.

While it is clear that the gut microbiome composition is strongly impacted by environmental exposures [Rothschild *et al*. 2018], the role of host genetics has only recently been implicated [Goodrich *et al.* 2014a, Blekhman *et al.* 2015, Goodrich *et al.* 2016]. Studies have identified

multiple genetic variants significantly associated with specific bacterial taxa abundances [Goodrich *et al.* 2017, Bonder *et al.* 2016, Igartua *et al.* 2017, Davenport *et al.* 2015, Turpin *et al.* 2016, Wang *et al.* 2016, Rothschild *et al*. 2018], despite the observation that generally the primary determinants of microbiome composition are non-genetic [Rothschild *et al*. 2018]. Human genetic studies have significant limitations for accurate assessment of genetic effects on the microbiome, including accessibility to large and diverse sample populations as well as a general lack of control over confounding variables. One major limitation is that there is minimal control of diet and other environmental factors, and so only the strongest genetic effects can be detected.

The mouse model, with the ability to control diet, along with well-defined genetic differences among inbred strains, provides a better opportunity to dissect genetic and environmental factors impacting microbiome composition. Quantitative trait locus (QTL) mapping efforts show that gut microbiota composition is a polygenic trait, with clearly mappable genetic factors influencing the gut microbiome composition [Benson *et al.* 2010, McKnite *et al.* 2012, Snijders *et al.* 2016]. Standard QTL mapping approaches have low mapping resolution, however, and advanced intercross lines provide one excellent means of improving mapping resolution. Belheouane *et al.* [2017] performed genetic and 16S rRNA gene analysis of skin microbiomes of a collection of 15-generation advanced intercross lines, and demonstrated that the improved mapping resolution also improved the specificity and significance of genetic associations. It is clear that the mouse model will provide further opportunities to dissect the means by which the host genome can modulate microbiome composition. A logical next step is a mapping experiment to identify portions of the genome that influence functional pathways that modulate the microbiome.

Here we extend the analysis of the link between the host genome and microbiome using the Diversity Outbred mouse model. The Diversity Outbred (DO) population is a heterogeneous mouse stock derived from the same eight progenitor lines (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ) used to establish the Collaborative Cross (CC) [Collaborative Cross Consortium 2012]. Mice from the CC lines at early stages of inbreeding were used to establish the DO population, which is maintained by randomized outbreeding among 175 mating pairs. The result is each individual DO mouse represents a unique combination of segregating alleles, whose genome is a unique mosaic of the original eight progenitor lines. The advantages of this outbreeding include normal levels of heterozygosity — similar to the human genetic condition — and substantially increased genetic mapping resolution [Churchill *et al.* 2012]. The CC/DO mice founder progenitor lines have already proven to be successful in identifying genetic associations with intestinal microbiome composition [O'Connor *et al.* 2014].

In this study, motivated by the high level of environmental control of the laboratory mouse and the improved mapping resolution of the Diversity Outbred mouse system, we identified genetic underpinnings of the gut microbiota of 247 Diversity Outbred mice. We uncover strong evidence of host genetic factors influencing the composition of many specific attributes of the gut microbiome. These included not only associations between specific host genetic variants and abundances of particular bacterial taxa, but also associations with functional molecular pathways.

3. RESULTS

3.1. Variation of gut microbiota

High-throughput sequencing of fecal samples from 247 three month old male mice from the Diversity Outbred Mouse Panel generated 15,149,384 16S rRNA gene sequences that passed the quality filtering criteria after demultiplexing (see **Materials and Methods**). On average, 61,334 sequences were obtained per sample (ranging from 17,658 to 135,803 sequences). Sequences were sorted into 57,014 operational taxonomic units (OTUs) at 97% identity against the Greengenes 8_13 database using open-reference OTU picking. Next, OTUs were summarized at five levels of taxonomy (phylum, class, order, family, genus). In order to focus on the most abundant microbes, only the taxa present in at least 50% of samples (i.e. present in 124 samples or more) were used for all following analysis, leaving a total of 80 taxa to test at the five levels of taxonomy (7 phyla, 9 classes, 12 orders, 21 families, and 31 genera). The most predominant taxa at the phylum level were Firmicutes (average relative abundance = 48.64%) and Bacteroidetes (46.41%), which is consistent with previous findings [Benson *et al.* 2010, McKnite *et al.* 2012, Org *et al.* 2015]. The relative abundances of these taxa were highly variable, with Firmicutes ranging from 11% to 94%, and Bacteroidetes ranging from 1% to 88% (**Figure 1**).

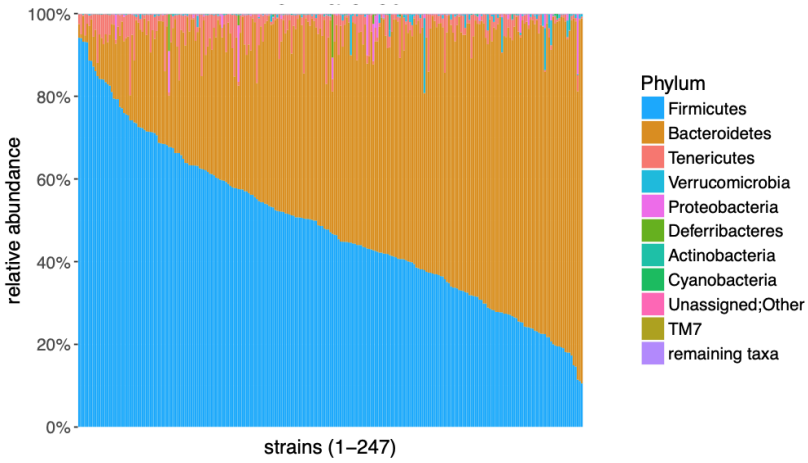

Figure 1. Relative abundances of top ten most abundant phyla across the 247 mouse strains. Relative abundances shown, mouse strains sorted by phylum Firmicutes, the most abundant phylum.

The top 8 most abundant genera were present in at least 99% of the samples. The two most abundant genera were an unidentified genus within Bacteroidales family S24-7 (average relative abundance = 43.89%, ranging from 1% to 88%) and another unidentified genus within Clostridiales (32.35%, ranging from 4% to 78%), consistent with previous findings [Shin *et al.* 2016]. Stacked bar plots and box plots depicting relative abundance frequencies for all five taxonomic levels are available in **Figure S1**.

When dealing with uneven sequence counts across samples, microbiome studies commonly normalize the data by rarefying sequence counts, which consists of randomly selecting from each sample an equal number of sequences without replacement [Weiss *et al.*

2017]. It has been argued, however, that rarefaction is not an ideal approach due to valuable data being discarded [McMurdie and Holmes 2014]. Therefore, we decided to present our analysis of the non-rarefied data using sequence counts per sample as a covariate, noting also that the rarefied data provided highly concordant results (see **Supplemental Material**).

3.2. Heritability estimation

Each of the 247 individual mice used in this study was genetically unique. The unit of inference for phenotypes was the relative abundance of each taxon in each individual, while the units of genetic inference were the SNP genotypes at each of 57,973 sites for each mouse using the mouse array. We estimated narrow-sense "SNP" heritability (h^2) using a linear mixed model in R-package *lme4qtl* [Ziyatdinov *et al.* 2018]. A linear mixed model was used to predict whether the effects of the autosomal genotype on the phenotype is proportional to the genetic similarity between the mice, after adjustment for known factors. Thus, calculations were based on the kinship matrix (genetic similarity), expression of a phenotype (taxon abundance) across all samples, and additional covariates (such as sequencing lane, read counts, and cage effect). Significance was assessed by an exact (restricted) likelihood ratio test using R-package *RLRsim* [Scheipl *et al.* 2008]. More details can be found in **Materials and Methods**. In total, 27 of the 80 tested taxa were heritable (nominal p-value < 0.05), with 3 additional taxa having statistically suggestive heritabilities of 20% or more (nominal p-value < 0.1) (**Table S1A**). Proportion variance estimates for kinship and cage for all taxa are presented in **Figure 2**.

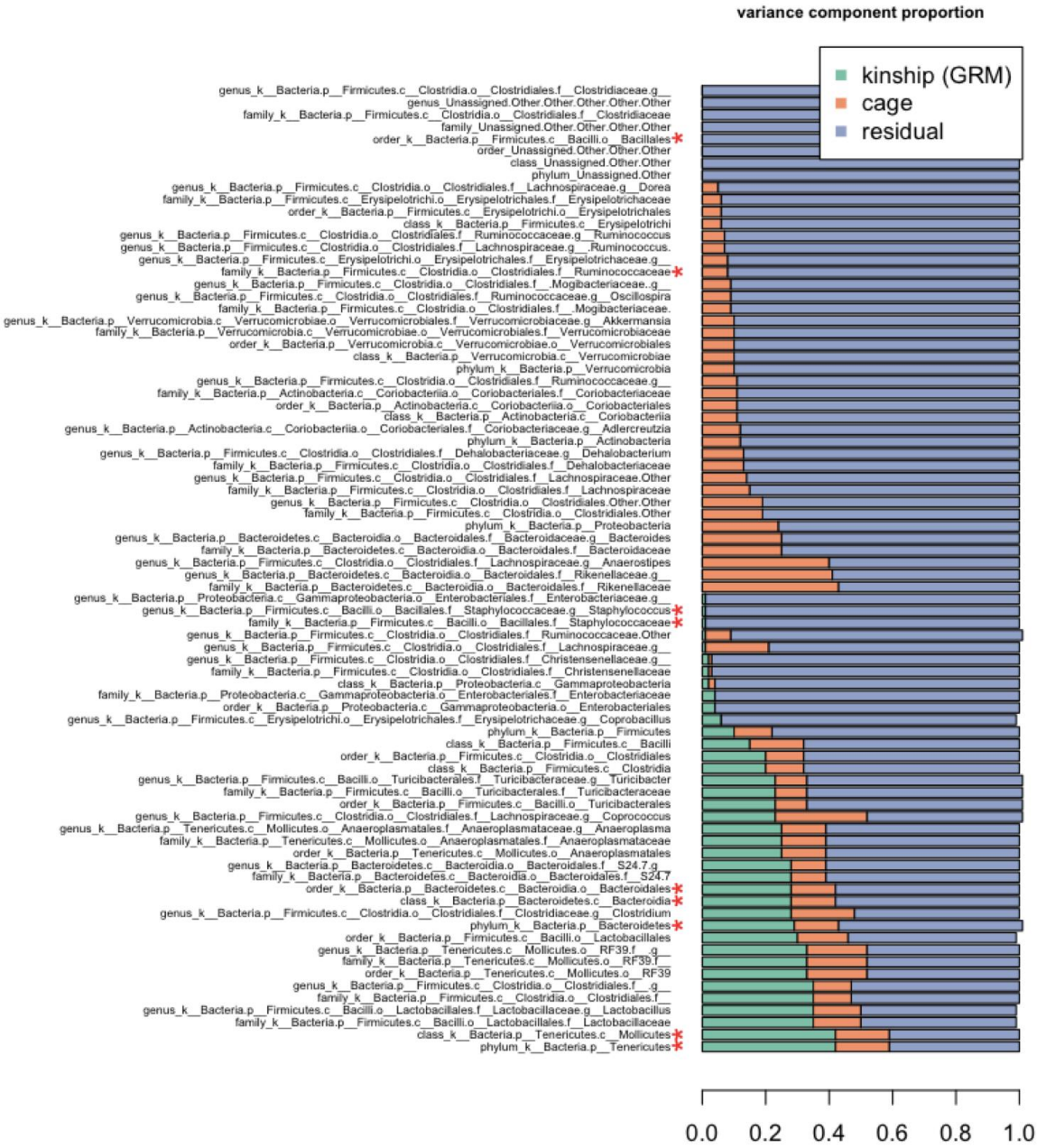

Figure 2. Proportion variance estimates for kinship and cage for all taxa. Proportion of variance for each taxon that can be explained by additive effects (heritability) using a kinship or Genomic Relationship Matrix (GRM) (green), cage effects (orange), and unexplained residual effects (blue). Taxa marked with a red asterisk have statistically suggestive QTL (adj. *p-value* < 0.1).

The most heritable taxon was the class Mollicutes with a heritability estimate of 39% (*p*-value of 0.002) (**Table 1**). Within Mollicutes, an unidentified genus in order RF39 was also found to be highly heritable, with a heritability of 34% (*p*-value 0.010) and the genus *Anaeroplasma* has a heritability of 28% (*p*-value 0.013). Within class Clostridia, an unidentified genus in order Clostridiales showed a heritability of 38% (*p*-value 0.0106). Furthermore, the genus *Lactobacillus* within class Bacilli and the entire Firmicutes phylum were also heritable, at 36% (*p*-value 0.008) and 23% (*p*-value 0.049) respectively. The genus *Turicibacter* within class

Bacilli had high heritability estimates as well at 35% (*p*-value 0.0043) and 28% (*p*-value 0.029) respectively. Given the large proportion of the microbiome is composed of either Firmicutes or Bacteroidales, their proportions are strongly negatively correlated. This means that the high heritability of Firmicutes abundance implies also a high heritability of the order Bacteroidales (31%, *p*-value 0.013), as well as an abundant unidentified genus in family S24-7 (heritability of 32%, *p*-value 0.014).

Table 1. Heritability of taxa at five taxonomic levels. Only showing ranked results with heritability above 20%. Results with *p*-value < 0.05 (statistically significant) are bolded. When results were identical across taxa in the same phylogenetic branch, only the lowest (most specific) taxon was kept. The designations p_, c_, o_, f_, and g_ are for phylum, class, order, family, and genus, respectively. Complete table of heritability results, including rarefied data, can be found in **Tables S1A-B**.

		Heritability %	*p*-value
Taxa	**p_Tenericutes;c_Mollicutes**	**39%**	**0.002**
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Unclassified;g_Unclassified	**38%**	**0.011**
	p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_*Lactobacillus*	**36%**	**0.008**
	p_Firmicutes;c_Bacilli;o_Turicibacteriales;f_Turicibacteraceae;g_*Turicibacter*	**35%**	**0.043**
	p_Tenericutes;c_Mollicutes;o_RF39;f_Unclassified;g_Unclassified	**34%**	**0.010**
	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_S24.7;g_Unclassified	**32%**	**0.014**
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Clostridiaceae;g_*Clostridium*	**31%**	**0.022**
	p_Tenericutes;c_Mollicutes;o_Anaeroplasmatales;f_Anaeroplasmataceae;g_*Anaeroplasma*	**28%**	**0.013**
	p_Firmicutes;c_Clostridia;o_Clostridiales	**28%**	**0.029**
	p_Firmicutes;c_Bacilli	**28%**	**0.029**
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_*Coprococcus*	**25%**	**0.019**
	p_Firmicutes	**23%**	**0.049**
	p_Firmicutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_*Coprobacillus*	23%	0.063
	p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae	20%	0.071

3.3. QTL Mapping

QTL mapping of the bacterial taxa at the five taxonomic levels revealed significant findings that suggest statistically significant associations between host genotype and abundances of certain taxa. QTL regions on autosomes were found using the R-package *lme4qtl* [Ziyatdinov *et al.* 2018]. Significance was assessed first by comparison of models with and without genotype via a likelihood ratio test, followed by a genome-wide permutation test. The reported *p*-values were corrected for multiple testing across SNPs (but not across taxa). In total, genetic associations with 3 taxa were found to be statistically significant (adj. *p*-value < 0.05), and genetic associations with 3 additional taxa were statistically suggestive (adj. *p*-value < 0.1) (**Table 2**, **Table S2A**).

We found statistically significant QTL associated with the abundance of family Ruminococcaceae, order Bacillales, and genus *Staphylococcus* (**Table 2**). We also found statistically suggestive QTL associated with phylum Bacteroidetes, order Bacteroidales, and class Mollicutes. Multiple QTL for various taxa overlapped with the QTL regions for their parent taxa, such as QTL hit for genus *Staphylococcus* (which is below the taxonomic branch for order Bacillales) overlapping the QTL hit for order Bacillales (**Table 2**). These overlaps are a common occurrence in both the significant and non-significant QTL (**Table S2A**).

Table 2. QTL regions for taxa at five taxonomic levels. Only showing ranked results with adj. *p*-value < 0.1 (statistically suggestive). Results with adj. *p*-value < 0.05 (statistically significant) are bolded. When results were overlapping across taxa in the same phylogenetic branch (such as p_Bacteroidetes and o_Bacteroidales), permutations were calculated only for the lowest (most specific) taxon. The designations p_, c_, o_, f_, and g_ are for phylum, class, order, family, and genus, respectively. Complete table of QTL results, including rarefied data, can be found in **Tables S2A-B**.

Taxa	chr[a]	maxlod[b]	pos[c]	from[d]	to[e]	*p*-value	adj.*p*-value	perm.*p*-value
p_Bacteroidetes	5	7.11	118.27	118.01	118.42	2.97E-05	0.089	NA
p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales	5	7.20	117.73	117.43	117.76	2.46E-05	0.085	0.105
	5	6.84	117.79	117.79	117.80	5.06E-05	0.085	0.192
	5	7.49	118.58	118.01	118.81	1.38E-05	0.085	0.061
p_Firmicutes;c_Bacilli;o_Bacillales	**19**	**8.37**	**27.02**	**26.55**	**27.42**	**2.38E-06**	**0.042**	**NA**
p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_*Staphylococcus*	**19**	**8.30**	**27.04**	**26.51**	**27.46**	**2.73E-06**	**0.023**	**0.008**
	19	**7.97**	**27.82**	**27.61**	**28.20**	**5.36E-06**	**0.023**	**0.022**
	19	6.68	32.10	31.83	32.28	6.82E-05	0.075	0.243
	19	6.56	32.43	32.43	32.46	8.73E-05	0.078	0.292
p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae	**5**	**7.12**	**31.93**	**31.90**	**32.16**	**2.91E-05**	**0.046**	**0.169**
	5	**7.27**	**32.52**	**32.27**	**33.36**	**2.14E-05**	**0.046**	**0.111**
	2	**6.87**	**170.57**	**170.51**	**170.64**	**4.71E-05**	**0.048**	**0.09**
p_Tenericutes;c_Mollicutes	1	7.00	121.32	120.23	125.20	3.66E-05	0.089	0.155

[a] Chromosome in which lies the QTL
[b] Maximum LOD score within the QTL
[c] Position in the chromosome where the maximum LOD score is found
[d] Chromosomal position where the QTL begins
[e] Chromosomal position where the QTL ends

Looking at specific QTL, we identified the genes *Insig2* and *Ksr2* on the highest point in the region for the class Mollicutes (chr1:121,315,223, LOD = 7.002) and the order Bacteroidales (chr5:117,733,508, LOD = 7.203) respectively. INSIG2 plays a central role in the pathway by which the circadian clock regulates liver lipid metabolism [Zhang *et al.* 2017] and *Ksr2* has been implicated in being associated with BMI and severe early-onset obesity through large scale GWAS studies [Milaneschi *et al.* 2019].

3.4. OTU level analysis

Next, we decided to increase the specificity of the taxonomic classifications to operational taxonomic units (OTUs) by compiling all OTUs identified within taxa that had statistically suggestive QTL (**Table 2**). We filtered out OTUs that were present in less than 50% of the

samples, resulting in 362 OTUs. QTL mapping performed on these selected OTUs resulted in 59 OTUs with at least one statistically suggestive association. Additionally, 99 OTUs were found to be heritable (h^2 > 20%, *p*-value < 0.05), of which 28 OTUs also had statistically suggestive QTL (**Tables S3 and S4**). Proportion variance estimates for kinship and cage for all tested OTUs are presented in **Figure S3**.

QTL associations to OTUs varied compared to overlapping QTL regions associated to taxa at higher taxonomic levels, some were sharper and stronger, others were less specific and wider (**Table 3**). These results are interesting because a sharper QTL peak associated with an OTU may suggest that the overlapping QTL region associated with the broader taxonomic group is being driven by that specific OTU. On the other hand, if the overlapping QTL region associated with the broader taxonomic group is smaller and more specific than the region seen on an individual OTU, this might suggest a cumulative effect of multiple sub-taxonomies driving a stronger signal at the broader taxonomic level. For example, QTL for OTU 338796 and New.CleanUp.ReferenceOTU 170146 within family Ruminococcaceae were both statistically significant and overlapped with the QTL region for Ruminococcaceae, but the QTL for the OTUs were both wider.

Table 3. QTL regions for OTUs. Only showing OTUs with adj. *p*-value < 0.1 (statistically suggestive) and with a QTL region overlapping QTL from higher-level taxonomies. Results with adj. *p*-value < 0.05 (statistically significant) are bolded. Complete table of QTL results for OTUs can be found in **Tables S4**.

		chr	maxlod	pos	from	to	*p*-value	adj.*p*-value	perm.*p*-value
Taxa	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_421792	5	5.87	118.69	118.63	118.82	3.26E-04	0.076	0.642
	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_460953	5	5.79	118.69	118.58	118.79	3.84E-04	0.078	NA
	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_190835	5	5.67	118.67	118.58	118.74	4.77E-04	0.076	NA
	p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;OTU_209408	5	7.02	118.67	118.50	118.82	3.53E-05	0.064	NA
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_NR.OTU100	5	8.22	31.93	30.25	32.06	3.23E-06	0.064	0.023
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_338796	**2**	**8.15**	**170.57**	**169.64**	**170.96**	**3.74E-06**	**0.023**	**0.023**
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_NR.OTU95	5	7.65	31.83	31.34	32.00	1.01E-05	0.076	0.088
		5	7.36	32,27	32.11	32.40	1.80E-05	0.076	0.048
		5	7.36	32.27	32.11	32.39	1.80E-05	0.076	0.081
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_336810	2	6.32	170.54	170.48	170.56	1.39E-04	0.100	NA
	p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;OTU_NCR.OTU170146	**5**	**7.63**	**36.01**	**32.27**	**36.22**	**1.05E-05**	**0.030**	**NA**

3.5. Comparison to other studies

Results from other published studies on heritabilities of the various bacterial taxa in the gut microbiome of mice, pigs, and humans were compiled and compared with our results (**Figure 3**). A full comparison of heritabilities among all analyzed taxa in our study and other studies can be found in **Table S5.**

Family S24-7 within order Bacteroidales had a high heritability in our study (h^2 = 0.32) and it has been reported as heritable and significant in both mice (h^2 = 0.60 in Org *et al.* 2015) and

humans (h^2 = 0.33 in Turpin *et al.* 2016) (**Figure 3**). Genus *Lactobacillus* was found to have a high and significant heritability (h^2 = 0.36) and it was also found to be highly heritable in one mouse study (h^2 = 0.74 in O'Connor *et al.* 2014) and both highly heritable and significant in a pig study (h^2 = 0.34 in Camarinha-Silva *et al.* 2017) and multiple human studies (h^2 = 0.36 in Davenport *et al.* 2015, 0.26 in Turpin *et al.* 2016, and 0.15 in Lim *et al.* 2017). Genus *Turicibacter*, also within class Bacilli, was found to have a high and significant heritability as well (h^2 = 0.57) and was found to be highly heritable in one mouse study (h^2 = 0.54 in Org *et al.* 2015) and display a strong QTL association in another (Kemis *et al.* 2019). *Turicibacter* was also found to be significantly heritable in human studies (h^2 = 0.26 in Turpin *et al.* 2016 and 0.36 in Goodrich *et al.* 2016). Under the class Clostridia and still within the phylum Firmicutes, family Christensenellaceae in our rarefied dataset had a high, significant heritability (h^2 = 0.31) that did not appear in our non-rarefied dataset. In human studies, Christensenellaceae has been found to be highly heritable and statistically significant (h^2 = 0.64 in Turpin et al. 2016, 0.42 in Goodrich et al. 2016, and 0.31 in Lim et al. 2017). Additionally, we found genus *Clostridium* to have a high and significant heritability (h^2 = 0.31) that was also seen in other human studies as well (h^2 = 0.24 in Goodrich *et al.* 2016 and 0.46 in Davenport *et al.* 2015). The genus *Coprococcus* was also significant and highly heritable from our study (h^2 = 0.25) as well as in various human studies (h^2 = 0.46 in Davenport *et al.* 2015 and 0.16 in Lim *et al.* 2017).

The clade within the phylum Tenericutes including genus *Anaeroplasma* gave high and statistically significant heritabilities in both our rarefied and non-rarefied datasets (**Figure 3**). Few other studies found similar results either because they did not include these taxa in their study or their results gave weaker heritabilities with non-significant *p*-values. Nonetheless, one mouse study did find a high heritability for genus *Anaeroplasma* (h^2 = 0.48 in O'Connor *et al.* 2014). In human studies, significant heritabilities were found for phylum Tenericutes (h^2 = 0.34 in Goodrich *et al.* 2016 and 0.23 in Lim *et al.* 2017), class Mollicutes (h^2 = 0.32 in Goodrich *et al.* 2016 and 0.23 in Lim *et al.* 2017), and order *RF39* (h^2 = 0.31 in Goodrich *et al.* 2016).

Both our non-rarefied and rarefied datasets gave insignificant heritability estimates of 0.02 for the genus *Akkermansia* and all the way up its taxonomic branch to phylum Verrucomicrobia, yet estimates from Org *et al.* 2015 were as high as h^2 = 0.92, and heritability of *Akkermansia* from O'Connor *et al.* 2014 was h^2 = 0.62 in mice. Studies conducted using human microbiome samples show a diversity of heritability estimates for this taxonomic branch: moderately high and significant (h^2 = 0.30 for both in Turpin *et al.* 2016), low and significant (h^2 = 0.15 for Verrucomicrobia and h^2 = 0.14 for *Akkermansia* in Goodrich *et al.* 2016), and close to zero and non-significant (up to h^2 = 0.01 for *Akkermansia* in Davenport *et al.* 2015, and 0.05 for Verrucomicrobia and 0.06 for *Akkermansia* in Lim *et al.* 2017).

	Heritability														
	Mouse								Human						Pigs
	Our		Org					O'Connor	Goodrich	Davenport			Turpin	Lim	Camarinha-
	nonR	R	All	M	F	Avg	One		'16	W	S	C			Silva
p_Bacteroidetes	**0.31**	**0.32**	0.53	0.82	0.73	0.00	0.02		0.08				**0.33**	**0.25**	0.20
└c_Bacteroidia	**0.31**	**0.32**	0.53	0.82	0.73	0.00	0.02		0.08				**0.33**		0.20
└o_Bacteroidales	**0.31**	**0.32**	0.53	0.82	0.73	0.00	0.02		0.08		0.00	0.00	**0.33**	**0.25**	0.00
└f_S24-7	**0.32**	**0.32**	0.60	0.86	0.82	0.00	0.00						**0.33**		
p_Firmicutes	**0.23**	0.22	0.56	0.71	0.77	0.15	0.16		0.00	0.00	0.00	0.00	0.18	0.10	
└c_Bacilli	**0.28**	**0.24**	0.68	0.74	0.76	0.01	0.00		0.03				0.19	**0.35**	
└o_Lactobacillales	**0.32**	**0.35**	0.77	0.79	0.55	0.00	0.00		0.00				0.10	**0.33**	
└f_Lactobacillaceae	**0.36**	**0.37**							0.04		0.13		**0.26**	**0.17**	0.00
└g_*Lactobacillus*	**0.36**	**0.37**						0.74	0.04	**0.36**	0.00	0.19	**0.26**	**0.15**	**0.34**
└o_Turicibacterales	**0.35**		0.54	0.75	0.82	0.12	0.12		**0.39**				**0.26**		
└f_Turicibacteraceae	**0.35**		0.54	0.75	0.82	0.12	0.12		**0.39**				**0.26**		
└g_*Turicibacter*	**0.35**		0.54	0.75	0.82	0.12	0.12	0.29	**0.39**	0.00	0.19	0.13	**0.26**		0.00
└c_Clostridia	**0.28**	**0.27**	0.58	0.80	0.77	0.00	0.03		0.03				0.22	0.07	0.00
└o_Clostridiales	**0.28**	**0.27**	0.58	0.80	0.77	0.00	0.03		0.03	0.00	0.00	0.00	**0.33**	0.07	0.00
└f_Christensenellaceae	0.18	**0.31**							**0.42**				**0.64**	**0.31**	
└f_Clostridiaceae	0.00	0.00	0.61	0.83	0.80	0.09	0.05		**0.30**	**0.35**		0.00	**0.35**	**0.34**	
└g_*Clostridium*	**0.31**								**0.24**	0.10	**0.46**	0.04	0.20	0.14	0.10
└f_Lachnospiraceae	0.07	0.08	0.52	0.60	0.69	**0.36**	0.07		0.16	0.13	0.00	**0.29**	0.17	**0.15**	0.09
└g_*Coprococcus*	**0.25**	0.29	0.28	0.61	0.55	0.00	0.02	0.19	0.09	**0.46**	0.06	**0.26**	0.04	**0.16**	
p_Tenericutes	**0.39**	**0.39**							**0.34**				0.06	**0.23**	
└c_Mollicutes	**0.39**	**0.39**							**0.32**				0.18	**0.23**	
└o_Anaeroplasmatales	**0.28**	**0.29**													
└f_Anaeroplasmataceae	**0.28**	**0.29**													
└g_*Anaeroplasma*	**0.28**	**0.29**						0.48							
└o_RF39	**0.34**	**0.32**							**0.31**				0.18		
p_Verrucomicrobia	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.15**				**0.30**	0.05	
└c_Verrucomicrobiae	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.14**				**0.30**		
└o_Verrucomicrobiales	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.14**				**0.30**		
└f_Verrucomicrobiaceae	0.02	0.11	0.54	0.85	0.92	0.13	0.33		**0.14**				**0.30**	0.06	
└g_*Akkermansia*	0.02	0.11	0.54	0.85	0.92	0.13	0.33	0.62	**0.14**	0.00	0.01	0.00	**0.30**	0.06	

Figure 3. Comparison of taxon heritabilities across mouse, human, and pig studies. The green shading over heritability estimates ranges from lowest heritability estimate (white) to highest heritability estimate (green) in a given study. Statistically significant results are shown in bold font. For our Diversity Outbred study, we report both non-rarefied (nonR) and rarefied (R) results. For Org *et al.* [2015] we report results using all mice (All), just males (M), just females (F), an average per strain (Avg), and a single mouse per strain (One). Org *et al.* [2015] and O'Connor *et al.* [2014] did not report significances. For Goodrich *et al.* [2016] the estimates are calculated by the ACE model, bold values indicate estimates with a 95% confidence interval not overlapping 0. For Davenport *et al.* [2015] the estimates are the proportion of variance explained (PVE) estimates ("chip heritability"), we report winter (W), summer (S), and combined seasons (C) datasets, and bold values indicate estimates with a standard error not overlapping 0. For Turpin *et al.* [2016] and Lim *et al.* [2017] estimates are polygenic heritability (H2r). For Camarinha-Silva *et al.* [2017] estimates are narrow-sense heritability (h^2). Grey indicates that the taxon was not observed or excluded in a given study. Figure adapted from Goodrich *et al.* [2016]. Selected comparisons shown, full comparison found in **Table S5**.

In addition to comparing our heritability estimates with other studies, we also contrasted our QTL mapping results of the gut microbiome with those from previous findings (**Figure 4**). A full comparison of QTL among all analyzed taxa in our study and other studies can be found in **Table S5**.

We identified statistically significant QTL associations for the order Bacillales as well as for the family Staphylococcaceae and the genus *Staphylococcus* within Bacillales in chromosome 19; another mouse study also found statistically significant QTL associations for all of the same taxa but on chromosome 17 [McKnite *et al.* 2012]. A human microbiome study found statistically

significant QTL regions for the class Bacilli, which comprise the above mentioned order and families [Blekhman *et al.* 2015].

Family Ruminococcaceae has been previously found to have significant QTL associations both in mice (chromosome 12, [Benson *et al.* 2010]) and humans ([Blekhman *et al.* 2015]). In our study, Ruminococcaceae was identified to be associated with chromosomes 2 and 5. We also identified a QTL hit for the phylum Bacteroidetes in chromosome 5 while another mouse study identified a significant hit in chromosome 14 [Wang *et al.* 2015]. Within Bacteroidetes, even though we did not find any significant QTL results for the genus *Bacteroides*, many other mouse studies did (chr 4 in McKnite *et al.* 2012, chr 9,16,18 in Leamy *et al.* 2014, chr 1 in Wang *et al.* 2015, and chr 11 in Bubier *et al.* 2018) as well as a human study [Blekhman *et al.* 2015].

Phylum Tenericutes had a significant hit in chromosome 1 in both our non-rarefied and rarefied datasets, and family Lachnospiraceae had a statistically suggestive QTL in chromosome 10 in our rarefied dataset but not in our non-rarefied dataset. Both of these taxa had significant QTL in a human study [Blekhman *et al.* 2015].

	QTL/GWAS signals							
	Mouse							Human
	Our nonR	Our R	Benson	McKnite	Leamy	Wang '15 H	Bubier	Blekhman
p_Bacteroidetes	5	5	no	no		**14**		no
└c_Bacteroidia	5	5						no
└o_Bacteroidales	5	5	no					no
└f_Bacteroidaceae			no	no				no
└g_*Bacteroides*			no	**4**	**9,16,18**	**1**	**11**	**9**
p_Firmicutes			no	no				no
└c_Bacilli			no					**2,10,14**
└o_Bacillales	**19**		no	**17**				
└f_Staphylococcaceae	**19**			**17**				
└g_*Staphylococcus*	**19**			**17**				
└c_Clostridia			no					no
└o_Clostridiales			no			3		no
├f_Lachnospiraceae		10	no	no				**1**
└f_Ruminococcaceae	**2,5**	**2,5**	**12**	no				**10**
p_Tenericutes	1	1						**6**
└c_Mollicutes	1	1						

Figure 4. Comparison of taxa with QTL associations across mouse and human studies. Associations with each taxon are marked in blue if statistically suggestive and bolded if statistically significant, or dark grey if not significant. The chromosome number were the QTL were found are denoted in each box. Light gray indicates that the taxon was not observed or excluded in a given study. For our Diversity Outbred study, we report both non-rarefied (nonR) and rarefied (R) results. Figure adapted from Goodrich *et al.* [2016]. Selected comparisons shown, full comparison found in **Table S5**.

3.6. Gene level analysis

Examining the QTL mapping results from previous studies, it was apparent that although different studies might all have found significant QTL regions for a particular bacteria taxon, they identified different genomic positions as showing associations. In order to identify common pathways shared by different QTL regions, we ran a cumulative geneset pathway analysis on the genes within our identified regions and the genes within the regions indicated in other

studies. In total, there were 60 significant QTLs with an additional 256 suggestive hits across the six taxonomic levels (phylum, class, order, family, genus, and OTU) (**Table S2 and S4**).

Of the analyzed gene subsets, the collection of genes within QTL among the taxa and OTUs that fall under the family Ruminococcaceae returned the most significant results. The Ingenuity Pathway Analysis (IPA) software was employed to analyze and categorize our geneset (IPA®, QIAGEN Redwood City, CA). Overall, 372 genes from 58 statistically significant and suggestive Ruminococcaceae QTL (**Table S6**) were submitted to IPA. A core analysis to find associated pathways and diseases generated multiple gene networks that revealed genes strongly associated with ovarian, breast, and colon cancer pathways (**Figure 5A-B)**.

Five genes, *Vegfa*, *Kat2b*, *Smad4*, *Fgfr2*, and *Yes1,* from our gene set were found to be highly linked to ovarian cancer (*p*-value = 5.43 x 10^{-6}) (**Figure 5A**). Although all of these genes have been identified to be related to ovarian cancer pathways, we recognized *Smad4* as a well-known and prevalent tumor suppressor gene. SMAD4 mediates the TGF-beta signaling pathway and occurs frequently in pancreatic and colorectal cancers with malignant progression and appears occasionally in other human cancers [Miyaki *et al.* 2003]. Additionally, five genes from our input gene set were found to be significantly associated with breast cancer pathways (*p*-value = 1.17 x 10^{-5}) (**Figure 5B**). Though the results are sparse and under studied, Ruminococcaceae abundance and breast cancer have been linked in previous studies. One study shows that Ruminococcaceae abundance was significantly higher in postmenopausal breast cancer patients when compared to normal healthy patients [Yang *et al.* 2017].

Genes from our significant Ruminococcaceae QTL were also found to be associated with colon carcinoma (*p*-value = 6.67 x 10^{-5}) and colorectal carcinoma (*p*-value = 1.75 x 10^{-4}) (**Figure 5A**) and associations between these bacteria and colorectal cancer (CRC) have been studied before. Ruminococcaceae was found to be significantly less abundant in cancerous colorectal tissue compared to healthy intestinal lumen [Chen *et al.* 2012]. Furthermore, another study showed findings that suggest Ruminococcaceae provides beneficial effects against risk of colorectal cancer [Ericsson *et al.* 2015].

In addition to having genes associated with specific cancers, various genes from our Ruminococcaceae gene set were found to have direct interactions with the well known and prevalent cancer gene *Tp53*. **Figure 5C** from IPA depicts a gene network containing 21 genes from the Ingenuity Knowledge Base and 14 genes from our input gene set, 5 of which (*Sorbs1*, *Stau1*, *Cox15*, *Ran*, and *Glb1*) have direct interactions with the widely known tumor suppressor gene *Tp53*. TP53 has been shown to be a critical player in tumor development and how tumor cells avoid apoptosis, and mutations in *Tp53* have been identified in numerous types of cancers [Petitjean *et al.* 2007, Greenblatt *et al.* 1994, Levine *et al.* 1991, Volgenstein *et al.* 2010].

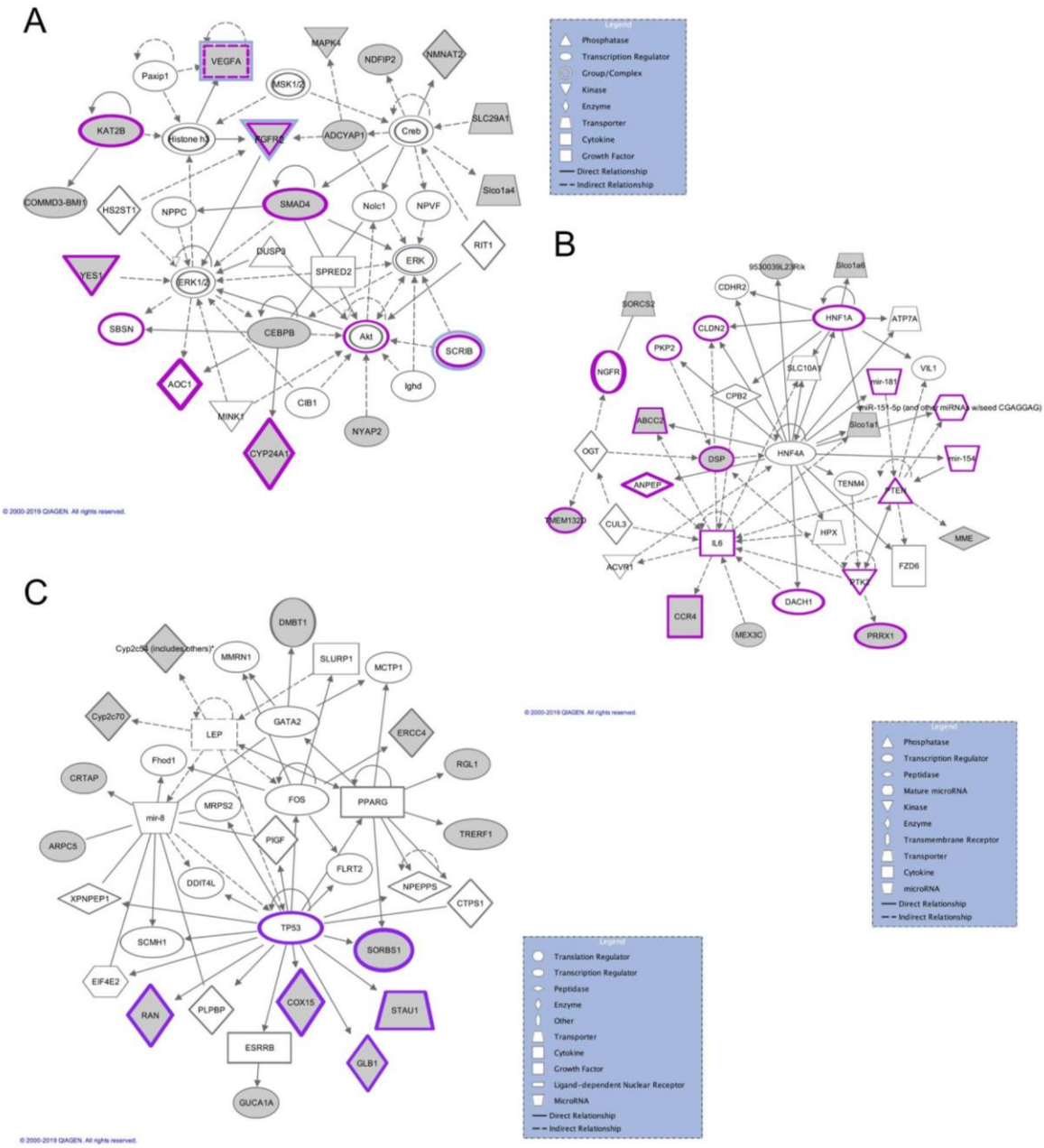

Figure 5. Ingenuity Pathway Analysis (IPA) interaction network generated from genes within Ruminococcaceae QTL. Genes circled in color are all part of specific associated pathways as specified below. Genes colored in gray belong to our dataset whereas un-colored genes are other closely associated genes added by IPA. Refer to **Tables S7A-C** for a list of these associated genes from our dataset. (**A**) The network shows genes found within Ruminococcaceae QTL strongly associate with pathways related to ovarian cancer (circled in pink) and colon carcinoma and colorectal carcinoma (circled in light blue). (**B**) The network shows genes found within Ruminococcaceae QTL strongly associate with pathways related to breast cancer (circled in pink). (**C**) The network shows genes found within Ruminococcaceae QTL strongly associated with hallmark cancer gene *Tp53* (circled in purple).

We compiled the results from all of our significant QTL under the order Bacillales and used the genes from within the QTL regions to run gene set functional pathway analyses and found these bacteria to be highly associated with pathways involved in lipid and sphingolipid metabolism. IPA identified our input genes *Vldlr* and *Sgms1* to be related to multiple lipid

metabolism pathways (p-value = 5.23 x 10^{-8} to 5.14 x 10^{-4}) (**Figure S4**, **Table S8A**) and DAVID functional annotation tool [Huang et al. 2008] identified our genes *Vldlr*, *Sgms1*, and *Asah2* to be related to lipid metabolism (p-value = 0.0048) as well as genes *Sgms1* and *Asah2* to be related to sphingolipid metabolism (p-value = 0.0063) (**Table S8B**). Associations between gut microbiota and host lipid metabolism have been investigated previously, and proof of causality between specific microbial associations with lipid metabolism and sphingolipid production has been demonstrated [Ghazalpour *et al.* 2016, Heaver *et al.* 2018, Johnson *et al.* 2019, Brown *et al.* 2019].

4. DISCUSSION

There exists a complex and multifaceted relationship between the gut microbiome and its host's genome, where recent studies are beginning to show the true magnitude of these connections. Our results seek to further understand this relationship by identifying functional and disease pathways that may be associated with specific bacterial abundances in the mouse gut microbiome.

SNPs with the highest LOD in the QTL regions for Mollicutes and Bacteroidales were found to lie within genes *Insig2* and *Ksr2* respectively. *Insig2* encodes a transmembrane protein that releases SREBP proteins to the endoplasmic reticulum where they exert control over lipid metabolism [Paschos *et al.* 2017]. Relationships between gut microbiome and lipid metabolism have already been established [Velagapudi *et al.* 2010, Li *et al.* 2008], and our reported association between Mollicutes and *Insig2* further suggest some kind of interaction between Mollicutes abundance and lipid metabolism. Gene *Ksr2* is known to be associated with BMI and early-onset obesity, as *Ksr2* variants impair cellular fatty acid oxidation and glucose oxidation, often leading to hyperphagia, low heart rate, reduced basal metabolic rate, and severe insulin resistance [Pearce *et al.* 2013]. This provides potential pathways by which *Ksr2* may lead to severe cases of obesity. Additionally, Bacteroidetes relative abundance has been shown to be 50% lower in genetically obese *ob*/*ob* mice compared to lean mice while Firmicutes relative abundance was higher by a corresponding amount [Ley *et al.* 2005]. The association we find between Bacteroidales and *Ksr2* may suggests a potential relationship between Bacteroidales abundance and risk for obesity.

Using the total set of genes from within all 58 statistically significant and suggestive QTL regions for taxa within the family Ruminococcaceae, we identified multiple networks, each of 35 functionally interrelated genes, enriched in disease pathways for ovarian, breast, and colon cancer. Evidence of functional associations between ovarian cancer and Ruminococcaceae is lacking, but various studies have confirmed findings showing increased Ruminococcaceae abundance in breast cancer patients compared to normal healthy individuals [Fernández *et al.* 2018, Zhu *et al.* 2018]. While these studies did not uncover a directionality to this association, the significant differences in microbiome composition could be used as independent biomarkers of breast cancer [Zhu *et al.* 2018]. In addition to specific links between the family Ruminococcaceae and breast cancer, associations between the gut microbiome and breast cancer have been flagged [Fernández *et al.* 2018]. This includes associations between perturbations in the gut microbiome and circulating estrogen levels and metabolites, produced

by several bacteria including *Ruminococcaceae* and also known as the estrobolome, which can affect the risk for breast cancer [Plottel and Blaser 2011, Fuhrman *et al.* 2014]. Indeed, the gut microbiome can influence estrogen metabolism through enterohepatic circulation [Adlercreutz et al. 1984, Flores *et al.* 2012], and thus could be implicated in breast cancer by interacting with estrogen metabolism [Minelli *et al.* 1990, Goedert *et al.* 2015]. Outside the gut microbiome, a study looked at the relationship between the breast tissue microbiome and breast cancer and also found significantly different microbiome composition and functions between women with benign and malignant breast disease [Hieken *et al.* 2016]. In aggregate, these studies support a role for microbes in the risk of breast carcinogenesis and our study extends this relationship by identifying specific genes involved in breast cancer pathways that may mediate this connection.

Our functional gene networks also revealed genes involved in colon cancer pathways. Ruminococcaceae abundance has been shown to be negatively correlated with risk for colorectal cancer (CRC) [Chen *et al.* 2012, Ericsson *et al.* 2015]. Looking beyond the specificity of Ruminococcaceae, various other studies have shown strong evidence for a link between the gut microbiome and risk for CRC. Microbiota in the colon form biofilms that line the mucosal surface, and a study has shown evidence suggesting that this biofilm structure may impact cellular proliferation and cancer growth by affecting the metabolome and down-regulating or up-regulating the production and release of metabolites favorable for tumor cells [Johnson *et al.* 2015]. General decreased microbial community diversity has been shown to be significantly correlated with risk for CRC in a study that compared CRC case subjects to control healthy subjects [Ahn *et al.* 2013]. Additionally, a study identified the enrichment and depletion of several bacterial populations associated with CRC and used this information in addition to known clinical risk factors for CRC to build a predictive model for evaluating risk for CRC. Used as a screening tool, this new predictive model that included microbial abundances improved accuracy by more than 50 folds [Zackular *et al.* 2014]. This not only confirms the existence of strong associations between the gut microbiome and CRC, but also raises the possibility that these data may be used as a potential diagnostic tool for clinical purposes.

In addition to revealing potential disease pathways associated with Ruminococcaceae, our geneset pathway analysis also unveiled connections between multiple genes to the well-characterized cancer gene *Tp53.* Genes *Sorbs1* and *Stau1* found in our QTL analysis have been shown to be down-regulated in cells that have undergone p53-mediated immortalization and transformation as a direct or indirect result of Ras signaling activity [Boiko *et al.* 2006]. Furthermore, another study showed through gene ontology analysis that p53 regulates various mitochondrial bioenergetic pathways including the up-regulation of our gene *Cox15* involved in ATP synthesis [Mak *et al.* 2017]. The same study also found that p53 regulates various genes involved in cardiac tissue function including the down-regulation of our gene *Ran* involved in major signal transduction pathways [Mak *et al.* 2017]. P53 was further found to decrease the activity of mouse SA beta-Galactosidase protein (encoded by our gene *Glb1*) in mouse mesothelial cells as well as in mouse embryonic fibroblast cells [Pietruska *et al.* 2007, Wang *et al.* 2007]. With multiple genes from within our significant Ruminococcaceae QTL exhibiting interactions with the popular tumor suppressor gene *Tp53*, it is highly suggestive that Ruminococcaceae abundance may be in some way linked to cancer development and tumor cell proliferation.

Similar geneset pathway analysis was conducted for the QTL under the order Bacillales and significant associations were found between various genes and lipid metabolism. Although specific interactions between Bacillales and lipid metabolism have not been thoroughly studied before, previous studies have elucidated a relationship between the gut microbiome and the metabolome. One study discovered increased energy metabolites in conventionally raised mice compared to germ free mice and further found microbiome composition to influence levels of various lipid classes, most significantly on triglyceride and phosphatidylcholine molecular species [Velagapudi *et al.* 2010]. Furthermore, systems biology analysis comparing human baby microbiota to normal microbiota in mice found that metabolism of dietary lipids was specifically influenced by the microbiome [Li *et al.* 2008]. In mouse, a study confirmed the microbiome to exert a strong impact on the metabolism of bile acids with increased bile acid levels in various gut compartments in germ free mice, suggesting that gut microbiome composition may affect host lipid metabolism through bile acid metabolism [Claus *et al.* 2008].

Comparing our results with other studies, we found little overlap in the specific bacterial taxa studied as well as the calculated heritabilities and QTL results. This is most likely due to the limited number of existing studies discussing heritabilities and QTL mappings of bacteria within the gut microbiome. Additionally, the absence of a standardized methodology for performing these studies leads to use of different procedures and analytical methods, making it increasingly difficult to compare results across studies [Goodrich *et al.* 2017]. Ultimately, the current state of the field for profiling different characteristics of the gut microbiome is still rapidly evolving and as it matures and more studies are undertaken, it will become easier to compare and validate results.

Although our results support the claim that host genetics can impact the gut microbiome composition in ways that are relevant to the health of the host, our study has some limitations. The biggest limitation to the power of the study is its relatively small sample size (n = 247 DO mice). Conducting QTL mapping with small sample sizes may lead to the 'Beavis effect' which is a failure to detect QTL of small effect sizes as well as an overestimation of effect size of the QTL that are discovered [Miles *et al.* 2008]. Our study also shares all the weaknesses common to the Diversity Outbred design: since the genome of each mouse is a unique mosaic of the 8 strains from the CC population, the genotype of each DO mouse is irreproducible. This limits the amount and manner of phenotyping that can be done, and it makes replicating results within the DO population difficult. However, this limitation could be partially circumvented by using the CC lines as a form of validation, since they can provide reproducible genotypes [Svenson *et al.* 2012]. Another limitation is the current lack of experimental validations of associations between disease pathways (such as those for ovarian, breast, and colon cancer) and specific taxa within gut microbiome composition, making it difficult to confirm any associations we find between genes and bacterial abundances.

Our results provide insight into the complex interplay between host genetics and the gut microbiome, and isolate potential associations between microbial taxa and QTL that may be involved in pathological disease phenotypes. Additional studies are required to verify associations between specific genes and taxon abundance in the gut microbiome, such as performing gene knockouts and observing the effects on microbiome composition. While most of the variation in the gut microbiome composition is not due to genetics but rather

environmental factors [Rothschild *et al*. 2018], attributes of the gut microbiome that are clearly heritable may provide important insights about host-microbiome interactions and mechanisms that impact microbiome composition. The direct genotype-phenotype association approach in this study could be applied to illuminate novel associations between genetic variants and their effects on microbial abundances involved in the microbiome through the mechanism of a complex disease of interest. Understanding the interactions between a host's genome and its microbiome composition may also aid in our understanding of complex diseases and their mechanisms and potentially aid in developing medical treatments.

5. MATERIALS AND METHODS

5.1. Animal population and sample collection

Male mice from the Diversity Outbred Mouse Panel were obtained from The Jackson Laboratory (Bar Harbor, ME, USA) at 6 weeks of age. Mice were group housed (5 animals per cage) for 2 weeks of post-travel acclimation, and then single housed at identical conditions. All mice were reared on chow diet. Fecal pellets from 249 mice were collected at 3 months old (two samples were later discarded, leaving a final analyzed dataset of 247 mice). Pellets were stored in Eppendorf tubes placed on dry ice and moved to a -80°C freezer until processing.

5.2 Microbial DNA extraction, 16S rRNA gene PCR, and sequencing

Microbial community DNA was extracted from one single frozen pellet per sample using the MO BIO PowerSoil-htp DNA Isolation Kit (MO BIO Laboratories, Inc., cat # 12955-4), but instead of vortexing, samples were placed in a BioSpec 1001 Mini-Beadbeater-96 for 2 minutes. We used 10-50 ng of sample DNA in duplicate 50 µl PCR reactions with 5 PRIME HotMasterMix and 0.1 µM forward and reverse primers. We amplified the V4 region of 16S using the universal primers 515F and barcoded 806R and the PCR program previously described [Caporaso *et al.* 2011], but with 25 cycles. We purified amplicons using the Mag-Bind® E-Z Pure Kit (Omega Bio-tek, cat # M1380) and quantified with Invitrogen Quant-iT™ PicoGreen® dsDNA Reagent, and 100 ng of amplicons from each sample were pooled and paired end sequenced (2x250bp) on an Illumina MiSeq instrument at Cornell Biotechnology Resource Center Genomics Facility.

5.3. 16S data processing

We performed demultiplexing of the 16S rRNA gene sequences and OTU picking using open source software package Quantitative Insights Into Microbial Ecology (QIIME) version 1.9.0 with default methods [Caporaso *et al.* 2010]. The total number of sequencing reads was 15,149,384, with an average of 61,334 sequences per sample and ranging from 17,658 to 135,803. Open-reference OTU picking at 97% identity was performed against the Greengenes 8_13 database. 12% of sequences failed to map in the first step of closed-reference OTU picking. The taxonomic assignment of the reference sequence was used as the taxonomy for each OTU. 'NR' within taxa names represents New Reference OTUs defined as those with sequences that failed to match the reference and are clustered *de novo*. Random subsamples were used to create a new reference OTU collection and 'NCR' represents New Clean-up Reference OTUs that failed to match the new reference OTU collection [Rideout et al. 2014].

For the non-rarefied data, read count was used as an additional covariate during QTL mapping to reduce the effect of sequencing depth. A rarefied dataset was also used for heritability estimates and QTL mapping, as explained in **Supplemental Material**. Two extreme outliers were omitted from further analysis, yielding a total of 247 samples. To differentiate the non-rarefied taxa from the rarefied taxa, we use 'NonR' to represent the non-rarefied dataset and 'R' to represent the rarefied dataset.

For heritability estimates and QTL mapping, a filter was applied across all 247 samples that removed any taxon that was not present in more than 50% of the samples. Relative

abundance of reads (number of reads clustered to each taxa divided by the total number of reads in a given sample) was used as the tested phenotype.

Stacked bar plots of the most abundant taxa within each taxonomic level were plotted with R-package *ggplot2*. A box-plot was first generated for each taxonomic level depicting the abundances of the taxa within that taxonomic level across the 247 samples (**Figure S1**). The top ten taxa with the highest average abundances are selected to be plotted in the stacked bar plot, ordered by the most abundant taxon. A heatmap that correlates similarities between taxa from the non-rarefied and rarefied datasets based on the Pearson correlation coefficient was plotted using the R-package *corrplot* (**Figure S2**).

5.4. SNP genotyping

SNP genotyping was done at the Jackson Laboratories on each of the 247 mice using The Mega Mouse Universal Genotyping Array (MegaMUGA). A total of 57,973 SNPs passed QC metrics and were used in the heritability and mapping analysis reported here.

5.5 Heritability calculations

Heritabilities of the various bacterial taxa were quantified and calculated on autosomes using a linear mixed model as implemented in R-package *lme4qtl* via the relmatLmer() function [Ziyatdinov *et al.* 2018] (https://github.com/variani/lme4qtl). This linear mixed model enables us to decompose variability into genetic and environmental components. The variance of the genetic component is expected to be $\sigma^2_g K$, where K is a kinship matrix normalized as proposed in Kang et al. 2010. The kinship matrix in specified via the "relmat" argument in relmatLmer(). To account for the potentially confounding effects of shared cages during acclimation (as noted above in Section 5.1), we also included cage as a random effect in our model. Thus, the model included estimates of variance of the genetic component (σ^2_g) and the cage component (σ^2_{cage}), and the residual variance due to unspecified environmental factors (σ^2_{rs}).

The narrow sense heritability was then estimated as:

$$h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_{cage} + \sigma^2_{rs}}$$

Sequencing lane was included as a covariate in both non-rarefied and rarefied datasets. For our non-rarefied dataset, narrow sense heritabilities were calculated using the number of read counts as an additional covariate. Significance of heritability estimates was assessed by conducting a restricted likelihood ratio test using the exactRLRT() function in the R-package *RLRsim* [Scheipl *et al.* 2008], as applied in Supplementary Note 3 in Ziyatdinov *et al.* 2018.

5.6. QTL Mapping

For QTL mapping, the relative abundances were rank Z-transformed using R-package *DOQTL* [Gatti *et al.* 2014] and then mapped using a linear mixed model in R-package

lme4qtl::relmatLmer() [Ziyatdinov *et al.* 2018] on autosomes with kinship included as random effect to account for genetic relatedness among animals. For the bacterial taxa from the five taxonomic levels, we generated QTL mappings for all taxa that passed the 50% zero cut-off (i.e. those present in at least 50% of the mice), with the taxa designated as the phenotype. Sequencing lane (fixed effect) and cage (random effect) were included in both non-rarefied and rarefied datasets. We included read count as an additional covariate (fixed effect) for our non-rarefied dataset. Significant and suggestive associations were identified in a two-step procedure. First, we applied likelihood ratio tests comparing models with and without genotype. *P*-values derived from these tests were adjusted for multiple testing across SNPs (within a given taxon) using R function p.adjust() with method "BH" [Benjamini and Hochberg 1995]. In the second step, we conducted permutation tests (1000 permutations) for taxa that had associations with adjusted *p*-value < 0.1 in the first step.

For every bacterial taxon from the five taxonomic levels with a statistically significant QTL association, we mapped the OTUs belonging to that taxon. We applied a 50% zero cut-off filter to only retain common OTUs. With the OTUs obtained, we generated QTL mappings and assessed significance just as we had done for the five taxonomic levels.

5.7. Gene Set Pathway Analysis

We used the open-source online DAVID annotation tool [Huang et al. 2008] and the Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, CA) software to conduct gene set pathway analysis. We used DAVID v6.8 and their functional annotation tool to reduce large gene sets into smaller groups of functionally related genes. A list of gene names was uploaded onto the website with the identifier parameter set to 'official_gene_symbol' and the species *Mus musculus* selected. DAVID then outputs a list of categories, such as functional, gene ontology, tissue expression, and others, which contained subsets of the inputted gene set. Within each category, DAVID also lists more specific categories and by displaying the genes for each sub-category, we were able to view which of the genes from our gene list were found to be associated with various different classifications. From the association results, a *p*-value filter allowed us to view only the results above a certain EASE *p*-value threshold, a modified Fisher-Exact *p*-value score. We chose the groupings with shown higher significance and reinforced the results outputted by DAVID with KEGG pathway database [Kanehisa *et al.* 2017] by simply confirming the presence of each gene in their organized category, as by DAVID, in KEGG's online database.

Using IPA, a new "core analysis project" was created and then our list of genes was uploaded as a dataset with parameters chosen to fit the formatting of our gene list. Before running the analysis, we set the reference set to be Ingenuity® Knowledge Base and then ran our analysis. IPA uses multiple categories to classify the inputted gene set and we focused on their disease and functions category. Others include expression, regulatory effects, and other canonical pathways. Additionally, IPA generates networks of genes proven to be either directly or indirectly related to each other. We chose the most significant network outputted and identified the intersection of that network with the network relating the genes in our QTL with the respective disease.

5.8. Data Availability

All scripts will be on GitHub. 16S data are on the Short Read Archive.

6. Acknowledgments

The authors want to thank Noah Clark, Jessica L. Sutter, Qiaojuan Shi, Emily Davenport, and Afrah Shafquat for all the help and advice provided. FS was supported by a Presidential Life Science Fellowship (PLSF) from Cornell University. This work was supported by NIH grant R01 GM 070683.

7. Author Contributions

F.S., A.G.C, G.A.C, and R.E.L. conceived the study. A.P. provided the samples. F.S. extracted and generated the 16S rRNA gene sequencing data. F.S., E.C., P.S., J.K.G., R.E.L., G.A.C, and A.G.C. conceived the computational and statistical analyses. F.S., D.Y.Z., E.C., M.E., and P.S. performed the computational and statistical analyses. F.S., D.Y.Z., E.C, and A.G.C. wrote the manuscript.

8. Supplemental Material

Figure S1 - **Taxa relative abundance frequencies** - Stacked bar plots and box plots depicting relative abundance frequencies of the top ten most abundant taxa for each of five taxonomic levels. Relative abundance frequencies are plotted for taxa levels from both the non-rarefied and the rarefied datasets.

Figure S2 - **Correlation plot between non-rarefied and rarefied taxa.** Heatmap depicting the Pearson correlations between the relative common taxa abundances in non-rarefied (NonR) and rarefied (R) data, revealing that the same taxa from both non-rarefied and rarefied datasets always group closer together than with other taxa, followed by taxa belonging to the same clade.

Figure S3 - Proportion variance estimates for kinship and cage for all OTUs. Proportion of variance for each OTU that can be explained by additive effects (heritability) using a kinship or Genomic Relationship Matrix (GRM) (green), cage effects (orange), and unexplained residual effects (blue). Taxa marked with a red asterisk have statistically suggestive QTL (adj. *p-value* < 0.1).

Figure S4 - **IPA network for Bacillales QTL** - Genes circled in purple are all part of the lipid metabolism pathway. Genes colored in gray belong to our dataset whereas un-colored genes are other closely associated genes added by IPA. Refer to **Tables S6A** for a list of these associated genes from our dataset.

Table S1 - **Heritability results at 5 taxonomic levels** - Complete heritability measurements (h^2) as well as their respective *p*-values for all tested taxonomies at the 5 taxonomic levels from the non-rarefied (**A**) and rarefied (**B**) datasets.

Table S2 - **QTL results at 5 taxonomic levels** - QTL regions and their respective *p*-values at the 5 taxonomic levels from the non-rarefied (**A**) and rarefied (**B**) datasets.

Table S3 - **Heritability results at OTU level in non-rarefied dataset** - Complete heritability measurements (h^2) as well as their respective *p*-values for all tested taxonomies at the OTU level from the non-rarefied dataset

Table S4 - **QTL results at OTU level in non-rarefied dataset** - QTL regions and their respective *p*-values at the OTU level from the non-rarefied dataset

Table S5 - **Comparison of heritabilities and QTL with other studies** - Comparison of taxa heritabilities and QTL across mouse, human, and pig studies.

Table S6 - **Ruminococcaceae genes used for gene-set analysis** - List of 372 genes from all significant QTL for taxa and OTUs under family Ruminococcaceae. Some genes were found in multiple QTL and all sources for those QTL are listed.

Table S7 - **Genes included in networks from Figure 5** - List of all genes that are part of the IPA networks from **Figure 5A** (**A**), **Figure 5B** (**B**), and **Figure 5C** (**C**). All sources for each gene are listed as well.

Table S8 - **Genes from gene set analysis using QTL under Bacillales** - Using the set of genes from all significant QTL from taxa under Bacillales, (**A**) lists the genes from IPA network in **Figure S4** and (**B**) shows results from DAVID functional annotation tool.

Link to Supplemental Figures:
https://docs.google.com/document/d/1mSQ0rNyCYPdyTOfBuI3Q4dgRDqaVg88M5KmGLpnExBw/edit
Link to Supplemental Tables:
https://docs.google.com/spreadsheets/d/1sgH4jg3rQDC2USu1u_LCHNzC652Nh9QQ_WT59V5hbbQ/edit
Link to Supplemental Material:
https://docs.google.com/document/d/1oQVyICh1MaH282ipYWRtWDQuY6C0yn1P6AQlxfJXzvc/edit

8. Citations

Adlercreutz, H., M. O. Pulkkinen, E. K. Hämäläinen and J. T. Korpela (1984). "Studies on the role of intestinal bacteria in metabolism of synthetic and natural steroid hormones." Journal of Steroid Biochemistry 20(1): 217-229.

Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J.J., Hayes, R.C., Yang, L. (2013). "Human gut microbiome and risk for colorectal cancer." Journal of the National Cancer Institute 105(24), 1907–1911.

Battaglioli, E. J., Hale, V. L., Chen, J., Jeraldo, P., Ruiz-Mojica, C., Schmidt, B. A., … Kashyap, P. C. (2018). "*Clostridioides difficile* uses amino acids associated with gut microbial dysbiosis in a subset of patients with diarrhea." Science Translational Medicine 10(464), eaam7019.

Belheouane, M., Y. Gupta, S. Künzel, S. Ibrahim and J. F. Baines (2017). "Improved detection of gene-microbe interactions in the mouse skin microbiota using high-resolution QTL mapping of 16S rRNA transcripts." Microbiome 5(1): 59.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B (Methodological) 57(1): 289-300.

Benson, A. K., S. A. Kelly, R. Legge, F. Ma, S. J. Low, J. Kim, M. Zhang, P. L. Oh, D. Nehrenberg, K. Hua, S. D. Kachman, E. N. Moriyama, J. Walter, D. A. Peterson and D. Pomp (2010). "Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors." Proceedings of the National Academy of Sciences USA 107(44): 18933-18938.

Blekhman, R., J. K. Goodrich, K. Huang, Q. Sun, R. Bukowski, J. T. Bell, T. D. Spector, A. Keinan, R. E. Ley, D. Gevers and A. G. Clark (2015). "Host genetic variation impacts microbiome composition across human body sites." Genome Biology 16(1): 191.

Brown, E. M., Ke, X., Hitchcock, D., Jeanfavre, S., Avila-Pacheco, J., Nakata, T., … Xavier, R. J. (2019). "Bacteroides-Derived Sphingolipids Are Critical for Maintaining Intestinal Homeostasis and Symbiosis." Cell Host & Microbe 25(5), 668-680.e7.

Boiko, A. D., S. Porteous, O. V. Razorenova, V. I. Krivokrysenko, B. R. Williams and A. V. Gudkov (2006). "A systematic search for downstream mediators of tumor suppressor function of p53 reveals a major role of BTG2 in suppression of Ras-induced transformation." Genes & Development 20(2): 236-252.

Bonder, M. J., A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, A. V. Vila, P. Deelen, T. Vatanen, M. Schirmer, S. P. Smeekens, D. V. Zhernakova, S. A. Jankipersadsing, M. Jaeger, M. Oosting, M. C. Cenit, A. A. M. Masclee, M. A. Swertz, Y. Li, V. Kumar, L. Joosten, H. Harmsen, R. K. Weersma, L. Franke, M. H. Hofker, R. J. Xavier, D. Jonkers, M. G. Netea, C. Wijmenga, J. Fu and A. Zhernakova (2016). "The effect of host genetics on the gut microbiome." Nature Genetics 48(11): 1407.

Bubier, J., V. Philip, C. Quince, J. Campbell, Y. Zhou, T. Vishnivetskaya, S. Duvvuru, R. Hageman Blair, J. Ndukum, K. D. Donohue, C. Phillips, C. Foster, D. Mellert, G. Weinstock, C. T. Culiat, E. J. Baker, M. A. Langston, B. Hara, A. V. Palumbo, M. Podar and E. J. Chesler (2018). "Systems genetic discovery of host-microbiome interactions reveals mechanisms of microbial involvement in disease." bioRxiv.

Camarinha-Silva, A., M. Maushammer, R. Wellmann, M. Vital, S. Preuss and J. Bennewitz (2017). "Host Genome Influence on Gut Microbial Composition and Microbial Prediction of Complex Traits in Pigs." Genetics 206(3): 1637-1644.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld and R. Knight (2010). "QIIME allows analysis of high-throughput community sequencing data." Nature Methods 7(5): 335-336.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer and R. Knight (2011). "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample." Proceedings of the National Academy of Sciences USA 108 Suppl 1(Suppl 1): 4516-4522.

Chen, W., F. Liu, Z. Ling, X. Tong and C. Xiang (2012). "Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer." PLOS ONE 7(6): e39743.

Churchill, G. A., D. M. Gatti, S. C. Munger and K. L. Svenson (2012). "The Diversity Outbred mouse population." Mammalian genome : official journal of the International Mammalian Genome Society 23(9-10): 713-718.

Claus, S. P., T. M. Tsang, Y. Wang, O. Cloarec, E. Skordi, F.-P. Martin, S. Rezzi, A. Ross, S. Kochhar, E. Holmes and J. K. Nicholson (2008). "Systemic multicompartmental effects of the gut microbiome on mouse metabolic phenotypes." Molecular Systems Biology 4: 219-219.

Consortium, C. C. (2012). "The genome architecture of the Collaborative Cross mouse genetic reference population." Genetics 190(2): 389-401.

Dash, N. R., Khoder, G., Nada, A. M., and Bataineh, M. T. A. (2019). "Exploring the impact of *Helicobacter pylori* on gut microbiome composition." PLOS ONE 14(6): e0218274.

Davenport, E. R., D. A. Cusanovich, K. Michelini, L. B. Barreiro, C. Ober and Y. Gilad (2015). "Genome-Wide Association Studies of the Human Gut Microbiota." PLOS ONE 10(11): e0140301.

Dudek-Wicher, R. K., Junka, A., & Bartoszewicz, M. (2018). "The influence of antibiotics and dietary components on gut microbiota." Przegląd Gastroenterologiczny 13(2), 85–92.

Ericsson, A. C., S. Akter, M. M. Hanson, S. B. Busi, T. W. Parker, R. J. Schehr, M. A. Hankins, C. E. Ahner, J. W. Davis, C. L. Franklin, J. M. Amos-Landgraf and E. C. Bryda (2015). "Differential susceptibility to colorectal cancer due to naturally occurring gut microbiota." Oncotarget 6(32): 33689-33704.

Fava, F., J. Lovegrove, R. Gitau, K. Jackson and K. Tuohy (2006). "The gut microbiota and lipid metabolism: implications for human health and coronary heart disease." Current Medicinal Chemistry 13(25): 3005-3021.

Fernández, M. F., I. Reina-Pérez, J. M. Astorga, A. Rodríguez-Carrillo, J. Plaza-Díaz and L. Fontana (2018). "Breast Cancer and Its Relationship with the Microbiota." International journal of environmental research and public health 15(8): 1747.

Flores, R., J. Shi, B. Fuhrman, X. Xu, T. D. Veenstra, M. H. Gail, P. Gajer, J. Ravel and J. J. Goedert (2012). "Fecal microbial determinants of fecal and systemic estrogens and estrogen metabolites: a cross-sectional study." Journal of Translational Medicine 10(1): 253.

Francino, M. P. (2015). "Antibiotics and the Human Gut Microbiome: Dysbioses and Accumulation of Resistances." Frontiers in Microbiology 6, 1543.

Fuhrman, B. J., H. S. Feigelson, R. Flores, M. H. Gail, X. Xu, J. Ravel and J. J. Goedert (2014). "Associations of the fecal microbiome with urinary estrogens and estrogen metabolites in postmenopausal women." Journal of Clinical Endocrinology and Metabolism 99(12): 4632-4640.

Garrett, W. S., J. I. Gordon and L. H. Glimcher (2010). "Homeostasis and inflammation in the intestine." Cell 140(6): 859-870.

Gatti, D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, P. Simecek, N. Goodwin, R. Cheng, D. Pomp, A. Palmer, E. J. Chesler, K. W. Broman and G. A. Churchill (2014). "Quantitative trait locus mapping methods for diversity outbred mice." G3 (Bethesda, Md.) 4(9): 1623-1633.

Ghazalpour, A., I. Cespedes, B. J. Bennett and H. Allayee (2016). "Expanding role of gut microbiota in lipid metabolism." Current Opinion in Lipidology 27(2): 141-147.

Goedert, J. J., G. Jones, X. Hua, X. Xu, G. Yu, R. Flores, R. T. Falk, M. H. Gail, J. Shi, J. Ravel and H. S. Feigelson (2015). "Investigation of the Association Between the Fecal Microbiota and Breast Cancer in Postmenopausal Women: a Population-Based Case-Control Pilot Study." JNCI: Journal of the National Cancer Institute 107(8): djv147-djv147.

Goodrich, J. K., E. R. Davenport, M. Beaumont, M. A. Jackson, R. Knight, C. Ober, T. D. Spector, J. T. Bell, A. G. Clark and R. E. Ley (2016). "Genetic Determinants of the Gut Microbiome in UK Twins." Cell Host & Microbe 19(5): 731-743.

Goodrich, J. K., E. R. Davenport, A. G. Clark and R. E. Ley (2017). "The Relationship Between the Human Genome and Microbiome Comes into View." Annual Review of Genetics 51: 413-433.

Goodrich, J. K., S. C. Di Rienzi, A. C. Poole, O. Koren, W. A. Walters, J. G. Caporaso, R. Knight and R. E. Ley (2014). "Conducting a microbiome study." Cell 158(2): 250-262.

Goodrich, Julia K., Jillian L. Waters, Angela C. Poole, Jessica L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, Jordana T. Bell, Timothy D. Spector, Andrew G. Clark and Ruth E. Ley (2014). "Human Genetics Shape the Gut Microbiome." Cell 159(4): 789-799.

Greenblatt, M. S., W. P. Bennett, M. Hollstein and C. C. Harris (1994). "Mutations in the p53 Tumor Suppressor Gene: Clues to Cancer Etiology and Molecular Pathogenesis." Cancer Research 54(18): 4855.

Heaver, S. L., Johnson, E. L., & Ley, R. E. (2018). "Sphingolipids in host-microbial interactions." Current Opinion in Microbiology 43, 92–99.

Hieken, T. J., J. Chen, T. L. Hoskin, M. Walther-Antonio, S. Johnson, S. Ramaker, J. Xiao, D. C. Radisky, K. L. Knutson, K. R. Kalari, J. Z. Yao, L. M. Baddour, N. Chia and A. C. Degnim (2016). "The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease." Scientific Reports 6: 30751-30751.

Huang, D. W., B. T. Sherman and R. A. Lempicki (2008). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature Protocols 4: 44-57.

Igartua, C., E. R. Davenport, Y. Gilad, D. L. Nicolae, J. Pinto and C. Ober (2017). "Host genetic variation in mucosal immunity pathways influences the upper airway microbiome." Microbiome 5(1): 16.

Johnson, Caroline H., Christine M. Dejea, D. Edler, Linh T. Hoang, Antonio F. Santidrian, Brunhilde H. Felding, J. Ivanisevic, K. Cho, Elizabeth C. Wick, Elizabeth M. Hechenbleikner, W. Uritboonthai, L. Goetz, Robert A. Casero, Jr., Drew M. Pardoll, James R. White, Gary J. Patti, Cynthia L. Sears and G. Siuzdak (2015). "Metabolism Links Bacterial Biofilms and Colon Carcinogenesis." Cell Metabolism 21(6): 891-897.

Johnson, E. L., Heaver, S. L., Waters, J. L., Kim, B. I., Bretin, A., Goodman, A. L., … Ley, R. E. (2019). "Sphingolipid production by gut Bacteroidetes regulates glucose homeostasis." BioRxiv, 632877.

Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato and K. Morishima (2017). "KEGG: new perspectives on genomes, pathways, diseases and drugs." Nucleic Acids Research 45(D1): D353-D361.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti and E. Eskin (2010). "Variance component model to account for sample structure in genome-wide association studies." Nature Genetics 42: 348.

Kemis JH, Linke V, Barrett KL, Boehm JF, Traeger LL, Keller MP, Rabaglia ME, Scheiler KL, Stapleton DS, Gatti DM, Churchill GA, Amador-Noguez D, Russell JD, Yandel BS, Broman KW, Coon JJ, Attie AD, and Rey FE. 2019 Genetic determinants of gut microbiota composition and bile acid profiles in mice. bioRxiv https://www.biorxiv.org/content/10.1101/571075v1

Leamy, L. J., S. A. Kelly, J. Nietfeldt, R. M. Legge, F. Ma, K. Hua, R. Sinha, D. A. Peterson, J. Walter, A. K. Benson and D. Pomp (2014). "Host genetics and diet, but not immunoglobulin A expression, converge to shape compositional features of the gut microbiome in an advanced intercross population of mice." Genome Biology 15(12): 552-552.

Levine, A. J., J. Momand and C. A. Finlay (1991). "The p53 tumour suppressor gene." Nature 351(6326): 453-456.

Ley, R. E., F. Bäckhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight and J. I. Gordon (2005). "Obesity alters gut microbial ecology." Proceedings of the National Academy of Sciences USA 102(31): 11070-11075.

Li, M., B. Wang, M. Zhang, M. Rantalainen, S. Wang, H. Zhou, Y. Zhang, J. Shen, X. Pang, M. Zhang, H. Wei, Y. Chen, H. Lu, J. Zuo, M. Su, Y. Qiu, W. Jia, C. Xiao, L. M. Smith, S. Yang, E. Holmes, H. Tang, G. Zhao, J. K. Nicholson, L. Li and L. Zhao (2008). "Symbiotic gut microbes modulate human metabolic phenotypes." Proceedings of the National Academy of Sciences 105(6): 2117.

Lim, M. Y., H. J. You, H. S. Yoon, B. Kwon, J. Y. Lee, S. Lee, Y.-M. Song, K. Lee, J. Sung and G. Ko (2017). "The effect of heritability and host genetics on the gut microbiota and metabolic syndrome." Gut 66(6): 1031.

Mak, T. W., L. Hauck, D. Grothe and F. Billia (2017). "p53 regulates the cardiac transcriptome." Proceedings of the National Academy of Sciences USA 114(9): 2331-2336.

McKnite, A. M., M. E. Perez-Munoz, L. Lu, E. G. Williams, S. Brewer, P. A. Andreux, J. W. M. Bastiaansen, X. Wang, S. D. Kachman, J. Auwerx, R. W. Williams, A. K. Benson, D. A. Peterson and D. C. Ciobanu (2012). "Murine Gut Microbiota Is Defined by Host Genetics and Modulates Variation of Metabolic Traits." PLOS ONE 7(6): e39191.

McMurdie, P. J. and S. Holmes (2014). "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible." PLOS Computational Biology 10(4): e1003531.

Milaneschi, Y., W. K. Simmons, E. F. C. van Rossum and B. W. J. H. Penninx (2019). "Depression and obesity: evidence of shared biological mechanisms." Molecular Psychiatry 24(1): 18-33.

Miles, C. M. and M. Wayne (2008). "Quantitative trait locus (QTL) analysis." Nature Education 1(1): 208.

Minelli, E. B., A. M. Beghini, S. Vesentini, L. Marchiori, G. Nardo, R. Cerutti and E. Mortani (1990). "Intestinal Microflora as an Alternative Metabolic Source of Estrogens in Women with Uterine Leiomyoma and Breast Cancer." Annals of the New York Academy of Sciences 595(1): 473-479.

Miyaki, M. and T. Kuroki (2003). "Role of Smad4 (DPC4) inactivation in human cancer." Biochemical and Biophysical Research Comm. 306(4): 799-804.

O'Connor, A., P. M. Quizon, J. E. Albright, F. T. Lin and B. J. Bennett (2014). "Responsiveness of cardiometabolic-related microbiota to diet is influenced by host genetics." Mammalian genome : official journal of the International Mammalian Genome Society 25(11-12): 583-599.

Org, E., B. W. Parks, J. W. J. Joo, B. Emert, W. Schwartzman, E. Y. Kang, M. Mehrabian, C. Pan, R. Knight, R. Gunsalus, T. A. Drake, E. Eskin and A. J. Lusis (2015). "Genetic and environmental control of host-gut microbiota interactions." Genome Research 25(10): 1558-1569.

Paschos, G. K. and G. A. FitzGerald (2017). "Circadian Clocks and Metabolism: Implications for Microbiome and Aging." Trends in Genetics 33(10): 760-769.

Pearce, Laura R., N. Atanassova, Matthew C. Banton, B. Bottomley, Agatha A. van der Klaauw, J.-P. Revelli, A. Hendricks, Julia M. Keogh, E. Henning, D. Doree, S. Jeter-Jones, S. Garg, Elena G. Bochukova, R. Bounds, S. Ashford, E. Gayton, Peter C. Hindmarsh, Julian P. H. Shield, E. Crowne, D. Barford, Nick J. Wareham, S. O'Rahilly, Michael P. Murphy, David R. Powell, I. Barroso and I. S. Farooqi (2013). "KSR2 Mutations Are Associated with Obesity, Insulin Resistance, and Impaired Cellular Fuel Oxidation." Cell 155(4): 765-777.

Petitjean, A., M. I. W. Achatz, A. L. Borresen-Dale, P. Hainaut and M. Olivier (2007). "TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes." Oncogene 26(15): 2157-2165.

Pietruska, J. R. and A. B. Kane (2007). "SV40 Oncoproteins Enhance Asbestos-Induced DNA Double-Strand Breaks and Abrogate Senescence in Murine Mesothelial Cells." Cancer Research 67(8): 3637.

Plottel, Claudia S. and Martin J. Blaser (2011). "Microbiome and Malignancy." Cell Host & Microbe 10(4): 324-335.

Ridaura, V. K., J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, M. J. Muehlbauer, O. Ilkayeva, C. F. Semenkovich, K. Funai, D. K. Hayashi, B. J. Lyle, M. C. Martini, L. K. Ursell, J. C. Clemente, W. Van Treuren, W. A. Walters, R. Knight, C. B. Newgard, A. C. Heath and J. I. Gordon (2013). "Gut microbiota from twins discordant for obesity modulate metabolism in mice." Science 341(6150): 1241214-1241214.

Rideout, J. R., Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez, A. Robbins-Pianka, J. C. Clemente, J. A. Gilbert, S. M. Huse, H.-W. Zhou, R. Knight and J. G. Caporaso (2014). "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences." PeerJ 2: e545-e545.

Rothschild, D., O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem, D. Zeevi, P. I. Costea, A. Godneva, I. N. Kalka, N. Bar, S. Shilo, D. Lador, A. V. Vila, N. Zmora, M. Pevsner-Fischer, D. Israeli, N. Kosower, G. Malka, B. C. Wolf, T. Avnit-Sagi, M. Lotan-Pompan, A. Weinberger, Z. Halpern, S. Carmi, J. Fu, C. Wijmenga, A. Zhernakova, E. Elinav and E. Segal (2018). "Environment dominates over host genetics in shaping human gut microbiota." Nature 555: 210.

Round, J. L. and S. K. Mazmanian (2009). "The gut microbiota shapes intestinal immune responses during health and disease." Nature Reviews Immunology 9(5): 313-323.

Scheipl, F., S. Greven and H. Küchenhoff (2008). "Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models." Computational Statistics & Data Analysis 52(7): 3283-3299.

Shin, J., S. Lee, M.-J. Go, S. Y. Lee, S. C. Kim, C.-H. Lee and B.-K. Cho (2016). "Analysis of the mouse gut microbiome using full-length 16S rRNA amplicon sequencing." Scientific Reports 6: 29681.

Snijders, A. M., S. A. Langley, Y.-M. Kim, C. J. Brislawn, C. Noecker, E. M. Zink, S. J. Fansler, C. P. Casey, D. R. Miller, Y. Huang, G. H. Karpen, S. E. Celniker, J. B. Brown, E. Borenstein, J. K. Jansson, T. O. Metz and J.-H. Mao (2016). "Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome." Nature Microbiology 2: 16221.

Svenson, K. L., D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan and G. A. Churchill (2012). "High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population." Genetics 190(2): 437-447.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight and J. I. Gordon (2009). "A core gut microbiome in obese and lean twins." Nature 457(7228): 480-484.

Turpin, W., O. Espin-Garcia, W. Xu, M. S. Silverberg, D. Kevans, M. I. Smith, D. S. Guttman, A. Griffiths, R. Panaccione, A. Otley, L. Xu, K. Shestopaloff, G. Moreno-Hagelsieb, G. E. M. P. R. Consortium, M. Abreu, P. Beck, C. Bernstein, L. Dieleman, B. Feagan, K. Jacobson, G. Kaplan, D. O. Krause, K. Madsen, J. Marshall, P. Moayyedi, M. Ropeleski, E. Seidman, S. Snapper, A. Stadnyk, H. Steinhart, M. Surette, D. Turner, T. Walters, B. Vallance, G. Aumais, A. Bitton, M. Cino, J. Critch, L. Denson, C. Deslandres, W. El-Matary, H. Herfarth, P. Higgins, H. Huynh, J. Hyams, D. Mack, J. McGrath, A. D. Paterson and K. Croitoru (2016). "Association of host genome with intestinal microbial composition in a large healthy cohort." Nature Genetics 48: 1413-1417.

Veiga, P., C. A. Gallini, C. Beal, M. Michaud, M. L. Delaney, A. DuBois, A. Khlebnikov, J. E. T. van Hylckama Vlieg, S. Punit, J. N. Glickman, A. Onderdonk, L. H. Glimcher and W. S. Garrett (2010). "Bifidobacterium animalis subsp. lactis fermented milk product reduces inflammation by altering a niche for colitogenic microbes." Proceedings of the National Academy of Sciences of the United States of America 107(42): 18132-18137.

Velagapudi, V. R., R. Hezaveh, C. S. Reigstad, P. Gopalacharyulu, L. Yetukuri, S. Islam, J. Felin, R. Perkins, J. Borén, M. Orešič and F. Bäckhed (2010). "The gut microbiota modulates host energy and lipid metabolism in mice." Journal of Lipid Research 51(5): 1101-1112.

Vogelstein, B., Sur, S. & Prives, C. (2010) "p53: The Most Frequently Altered Gene in Human Cancers." Nature Education 3(9):6.

Wang, J., S. Kalyan, N. Steck, L. M. Turner, B. Harr, S. Künzel, M. Vallier, R. Häsler, A. Franke, H.-H. Oberg, S. M. Ibrahim, G. A. Grassl, D. Kabelitz and J. F. Baines (2015). "Analysis of intestinal microbiota in hybrid house mice reveals evolutionary divergence in a vertebrate hologenome." Nature Communications 6: 6440.

Wang, J., L. B. Thingholm, J. Skiecevičienė, P. Rausch, M. Kummen, J. R. Hov, F. Degenhardt, F.-A. Heinsen, M. C. Rühlemann, S. Szymczak, K. Holm, T. Esko, J. Sun, M. Pricop-Jeckstadt, S. Al-Dury, P. Bohov, J. Bethune, F. Sommer, D. Ellinghaus, R. K. Berge, M. Hübenthal, M. Koch, K. Schwarz, G. Rimbach, P. Hübbe, W.-H. Pan, R. Sheibani-Tezerji, R. Häsler, P. Rosenstiel, M. D'Amato, K. Cloppenborg-Schmidt, S. Künzel, M. Laudes, H.-U. Marschall, W. Lieb, U. Nöthlings, T. H. Karlsen, J. F. Baines and A. Franke (2016). "Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota." Nature genetics 48(11): 1396-1406.

Wang, L., L. Yang, M. Debidda, D. Witte and Y. Zheng (2007). "Cdc42 GTPase-activating protein deficiency promotes genomic instability and premature aging-like phenotypes." Proceedings of the National Academy of Sciences of the United States of America 104(4): 1248-1253.

Weiss, S., Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, E. R. Hyde and R. Knight (2017). "Normalization and microbial differential abundance strategies depend upon data characteristics." Microbiome 5(1): 27.

Wen, L., R. E. Ley, P. Y. Volchkov, P. B. Stranges, L. Avanesyan, A. C. Stonebraker, C. Hu, F. S. Wong, G. L. Szot, J. A. Bluestone, J. I. Gordon and A. V. Chervonsky (2008). "Innate immunity and intestinal microbiota in the development of Type 1 diabetes." Nature 455(7216): 1109-1113.

Yang, J., Q. Tan, Q. Fu, Y. Zhou, Y. Hu, S. Tang, Y. Zhou, J. Zhang, J. Qiu and Q. Lv (2017). "Gastrointestinal microbiome and breast cancer: correlations, mechanisms and potential clinical implications." Breast Cancer 24(2): 220-228.

Zackular, J. P., M. A. M. Rogers, M. T. Ruffin and P. D. Schloss (2014). "The Human Gut Microbiome as a Screening Tool for Colorectal Cancer." Cancer Prevention Research 7(11): 1112.

Zhang, Y., R. Papazyan, M. Damle, B. Fang, J. Jager, D. Feng, L. C. Peed, D. Guan, Z. Sun and M. A. Lazar (2017). "The hepatic circadian clock fine-tunes the lipogenic response to feeding through RORα/γ." Genes & Development 31(12): 1202-1211.

Zhu, J., M. Liao, Z. Yao, W. Liang, Q. Li, J. Liu, H. Yang, Y. Ji, W. Wei, A. Tan, S. Liang, Y. Chen, H. Lin, X. Zhu, S. Huang, J. Tian, R. Tang, Q. Wang and Z. Mo (2018). "Breast cancer in postmenopausal women is associated with an altered gut metagenome." Microbiome 6(1): 136.

Ziyatdinov, A., M. Vázquez-Santiago, H. Brunel, A. Martinez-Perez, H. Aschard and J. M. Soria (2018). "lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals." BMC bioinformatics 19(1): 68-68.