Genome-wide mapping of gene-microbe interactions in the murine lung microbiota based on quantitative microbial profiling

Cecilia J. Chung  
Max Planck Institute for Evolutionary Biology

Britt M. Hermes  
Max Planck Institute for Evolutionary Biology

Yask Gupta  
Columbia University Irving Medical Center

Saleh Ibrahim  
Khalifa University of Science and Technology

Meriem Belheouane  
Research Center Borstel - Leibniz Lung Center

John F. Baines  (✉ baines@evolbio.mpg.de)  
Kiel University

Research Article

Keywords: mouse, lung microbiota, host genetics, QTL mapping, quantitative microbial profiling, Lactobacillus, Interleukin 10

Posted Date: November 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2302161/v1

License: ☭ This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background: Mammalian lungs comprise a complex microbial ecosystem that interacts with host physiology. Previous research demonstrates that the environment significantly contributes to bacterial community structure in the upper and lower respiratory tract. However, the influence of host genetics on the makeup of lung microbiota remains ambiguous, largely due to technical difficulties related to sampling, as well as challenges inherent to investigating low biomass communities. Thus, innovative approaches are warranted to clarify host-microbe interactions in the mammalian lung.

Results: Here, we aimed to characterize host genomic regions associated with lung bacterial traits in an advanced intercross mouse line (AIL). By performing quantitative microbial profiling (QMP) using the highly precise method of droplet digital PCR (ddPCR), we refined 16S rRNA gene amplicon-based traits to identify and map candidate lung-resident taxa using a QTL mapping approach. In addition, the two abundant core taxa *Lactobacillus* and *Pelomonas* were chosen for independent microbial phenotyping using genus-specific primers. In total, this revealed seven significant loci involving eight bacterial traits. The narrow confidence intervals afforded by the AIL population allowed us to identify several promising candidate genes related to immune and inflammatory responses, cell apoptosis, DNA repair, and lung functioning and disease susceptibility. Interestingly, one genomic region associated with *Lactobacillus* abundance contains the well-known anti-inflammatory cytokine *Il10*, which we confirmed through the analysis of *Il10* knockout mice.

Conclusions: Our study provides the first evidence for a role of host genetic variation contributing to variation in the lung microbiota. This was in large part made possible through the careful curation of 16S rRNA gene amplicon data and the incorporation of a QMP-based methods. This approach to evaluating the low biomass lung environment opens new avenues for advancing lung microbiome research using animal models.

Background

Healthy lungs house a diverse and complex microbial ecosystem that contributes to critical aspects of host biology. Previous surveys of lung microbiota reveal microbial alterations in the context of disease, including cystic fibrosis [1], asthma [2], chronic obstructive pulmonary disease [3, 4], lung cancer [5, 6], as well as COVID-19 [7–9]. Thus, revealing the fundamental forces that govern the assembly and stability of bacterial communities in the lung is of critical importance for understanding its role in health and disease. Environmental factors, including smoking status [3, 10], infant feeding mode [11], early life stress [12], household [13, 14], and antibiotic use [15, 16] are identified as significant contributors to upper and lower respiratory tract bacterial community structure. However, there is a paucity of research exploring the influence of host genetics on the makeup of lung microbiota.

The role of host genetics on bacterial community structure at other body sites has been demonstrated using twin designs [17], comparison of mouse inbred strains [18, 19], genome-wide association studies [20–24], and quantitative trait locus (QTL) analyses [25, 26]. These studies are largely limited to the gut, although researchers are beginning to successfully apply these methods to low-biomass communities. For example, we previously used a novel QTL mapping approach in a murine skin microbiota study whereby we extended bacterial trait mapping to both 16S rRNA gene copy (DNA) as well as 16S rRNA gene transcripts (RNA) [27]. Microbial profiling based on RNA as template preferentially reflects living/active cells, and accordingly increased the number of significant associations detected between the host and resident skin microbes [27].

Microbiome research on the lung environment still lags, largely due to technical sampling challenges unique to this site [28–32]. Bronchoalveolar lavage (BAL), the established best practice for sampling lung microbiota, is invasive, costly, requires sedation, and poses unnecessary risks to healthy subjects [31]. These barriers impede sampling large numbers of individuals, which is required to detect biological signals with methods such as GWAS [31]. Non-invasive sampling
strategies, such as sputum or tracheal aspirate collection, are more accessible, but complicated by poorly defined methods for sputa processing, ambiguous origins of collected microbiota, and a substantial risk of contamination from the oropharynx [31, 33]. Moreover, the low biomass of bacterial samples collected from the lower airways poses further challenges for handling contamination. The so-called “kitome” of nucleic acid extraction kits and reagents, as well as laboratory environments, are well-documented sources of contamination that can radically affect data interpretation, as contaminants tend to be preferentially amplified and sequenced over true microbial signal in low biomass samples [34–40]. Thus, novel approaches are needed to advance our understanding of dynamic host-microbe interactions in the mammalian lung.

In this study, we aimed to improve the experimental profiling of resident lung microbes for QTL mapping of the lung microbiota, using a mouse advanced intercross line (AIL) that was previously successful for genetic mapping in both the gut [41] and low biomass environment of the skin [27, 42]. For this, we employed a strategy to first screen for taxa that are likely to be true lung residents using 16S rRNA gene amplicon profiles at the transcript (RNA) level, followed by measurements of overall bacterial load and selected individual taxa using the highly precise method of droplet digital PCR (ddPCR). QTL linkage mapping of lung microbiota using ddPCR-based estimates revealed significant associations with host loci, whose confidence intervals contain genes related to immune and inflammatory responses, cell apoptosis, and DNA repair. Further, a significant association between *Lactobacillus* abundance and a region of the mouse genome containing *Il10*, a well-known anti-inflammatory cytokine, was confirmed through the analysis of *Il10* knockout mice. These data suggest that incorporating quantitative profiling from ddPCR bacterial load measurements for use in linkage mapping may improve study reliability, and thus open new avenues for advancing lung microbiome research.

**Results**

**Ail Mouse Population And Overall Study Design**

We analyzed 242 lung tissue samples derived from the 15th generation of a previously established AIL population, as described by Belheouane et al [27]. In brief, the AIL consisted of MRL/MpJ, NZM2410/J, BXD2/TyJ, and CAST/EiJ mice (Jackson Lab, Maine, USA). To create a heterogenous intercross line, mice were intercrossed in equal strain and sex distributions [27, 42].

To map genomic regions associated with bacterial traits in the murine lung, we carried out a nested strategy to identify and map candidate resident taxa while minimizing the influence of potential contamination (see Methods). First, we screened for putative bacterial lung residents by analyzing 16S rRNA gene amplicon profiles at the transcript (RNA) level, thereby preferentially identifying live/active taxa. These data were further curated through the application of the “decontam” R package [34], which incorporates information from negative controls and absolute quantification of bacterial load to identify possible contaminants in metagenomic sequencing data. For this purpose, we applied ddPCR to obtain precise total bacterial load measurements. Next, a core measurable microbiota (CMM) was defined based on the processed sequences for further analysis. In a second step, we measured the bacterial loads of individual candidate taxa using ddPCR. Finally, linkage mapping was performed on the following panel of traits: (1) CMM based on conventional relative abundance estimates, derived from RNA template (herein: CMM-RA), (2) CMM based on relative abundance corrected by quantitative microbiome profiling (QMP; i.e., ddPCR estimates of bacterial load), derived from RNA template (herein: CMM-QMP), (3) ddPCR estimates of candidate taxa derived from DNA template (herein: ddPCR-DNA), and (4) ddPCR estimates of candidate taxa derived from RNA template (herein: ddPCR-RNA).

**16s Rrna Gene Sequencing And Ddpcr To Define Bacterial Traits**
To identify resident candidate taxa from the lung, we first performed 16S rRNA gene amplicon sequencing on both DNA and RNA reverse transcribed into complementary DNA (cDNA) as template (see Methods). After sequence processing, we determined the DNA-based data to be of insufficient quality/quantity, similar to other reports of lung samples [43]. However, rather than adopting a two-step PCR protocol as in Barfod et al. [43], which may be more prone to amplifying contaminants in low biomass samples [44], we instead narrowed the 16S sequencing data analysis to the transcript (RNA) level. This reflects metabolically active cells [45, 46], and may better reveal true resident lung bacteria interacting with the host, as demonstrated in our QTL mapping study exploring gene-microbe interactions in the skin of this same AIL [27]. To assess for potential contamination, we employed the “frequency” method from the “decontam” R package (v.1.4.0) using a threshold of 0.1 (see Methods). This analysis resulted in the removal of potential contaminant amplicon sequence variants (ASVs), including the common contaminants *Halomonas* and *Shewanella*. In total, we analyzed 8,414,939 sequences, and after normalizing sequencing coverage to 4,000 sequences per sample, a total of 20,772 ASVs remained in the data set. We first analyzed murine lung bacterial community composition at the phylum and genus levels based on 16S rRNA gene profiles (Fig. 1). The most abundant phyla include *Firmicutes*, *Proteobacteria*, *Actinobacteria*, and *Bacteroidetes* (Fig. 1A). The two most abundant genera classified to the genus-level include *Lactobacillus*, with a mean relative abundance of 13.20% and *Pelomonas*, with a mean relative abundance of 7.16% (Fig. 1B). These two taxa were thus additionally selected as bacterial candidate traits for targeted linkage mapping using ddPCR (see below).

Next, we defined a CMM [25, 26] (see Methods) consisting of 58 taxa ranging from the genus- to phylum-level, as well as 13 ASVs. The number of CMM traits represents a small fraction of the total number of lung microbiota (e.g., 1.74% of genera and 0.06% of ASVs), but their abundances represent 55.16% and 22.38% of the respective taxa at these levels.

Because 16S rRNA gene amplicon profiling is prone to biases and errors related to relative abundance compositional analysis [47], we performed QMP using ddPCR-based bacterial load estimates at the RNA level to account for these concerns. 16S rRNA sequence-derived relative abundances of the CMM traits were then subsequently transformed into corrected, absolute quantitative abundances [47] (see Methods), i.e., CMM-QMP traits. Thereafter, the quantitative profiles of the 58 taxa and 13 ASVs were included as an independent set of bacterial traits.

Finally, as indicated above, we selected two candidate bacterial traits, *Lactobacillus* and *Pelomonas*, for further independent analysis, as they were the two most abundant classifiable genera and are frequently identified as lung residents [43, 48, 49]. Droplet digital PCR is ideal for quantifying low biomass samples, as the process of fractionating a sample into thousands of individual droplets, in which independent PCR reactions occur, allows for the amplification of even very low levels of target strains [4, 50–52]. This was demonstrated by e.g., Gobert et al., who effectively measured low levels of *Lactobacillus* in fecal samples using ddPCR [50]. We thus generated load estimates for *Lactobacillus* and *Pelomonas* using genus-specific primers adapted for ddPCR (Table 3). This was performed at both the DNA and RNA level, as (i) ddPCR is sensitive enough to allow for absolute quantification and (ii) taxon-specific primers are expected to be less prone to contaminating taxa than universal PCR primers. Importantly, these specific estimates are significantly correlated to those based on 16S rRNA gene sequencing for both *Lactobacillus* (ddPCR-RNA vs. CMM-RA: Spearman's $r = 0.5848$, $p < 2.2 \times 10^{-16}$; ddPCR-RNA vs. CMM-QMP: $r = 0.7131$, $p < 2.2 \times 10^{-16}$) and *Pelomonas* (ddPCR-RNA vs. CMM-RA: $r = 0.2519$, $p = 7.411 \times 10^{-05}$; ddPCR-RNA vs. CMM-QMP: $r = 0.2916$, $p = 3.796 \times 10^{-06}$), whereby the correlation is stronger for the CMM-QMP traits.

### Summary Statistics And Evaluation Of Sources Of Variation For Lung Bacterial Traits
Prior to genetic mapping, it is important to determine whether (i) a given phenotype displays sufficient variation between individuals, and (ii) any residual variation remains after accounting for potential covariates such as cage, sex, or age. Summary statistics for each of the four categories of bacterial traits are provided in Additional file 1: Tables S1-3. This reveals considerable interindividual variation among traits. In particular, traits with large mean abundances, such as Lactobacillus and Pelomonas, display a large range, from 0-88.15% and 0-82.85% across the dataset for the CMM-RA measurements (Additional file 1: Tables S1-3).

Next, we evaluated the influence of host and environmental factors on bacterial trait variation in each of the four categories by constructing a mixed effects model that includes sex and age as fixed explanatory variables, cage as a random effect, and the bacterial trait values as responses (see Methods; Additional file 1: Table S4-6). These variables are associated with variation in trait values to varying degrees. For example, cage explains 0% of the total variance for Lactobacillus, but explains 28.83% of the total variance for Pelomonas among CMM-RA traits. However, importantly, the remaining residual variation in all four categories of traits accounts for the greatest proportion of total variance after accounting for sex, age, and cage effects. This suggests that other variables differing between individual mouse hosts, such as genotype, may contribute to variation in bacterial trait abundances.

**QTL Mapping Of Lung Microbiota Traits**

To test for associations between resident lung bacteria and the host genome, we performed QTL mapping for the CMM-RA and CMM-QMP traits, as well as for the ddPCR-DNA and -RNA for Lactobacillus and Pelomonas. In total, this yielded seven significant \((p \leq 0.05)\) and six suggestive \((p \leq 0.1)\) host loci involving eight bacterial traits (CMM-RA: one significant, five suggestive; CMM-QMP: four significant, one suggestive; ddPCR-DNA/RNA: two significant), with narrow confidence intervals ranging from 0.08 to 3.59 Mb, with an average of 1.80 Mb (Table 1, Additional file 2). The number of protein-coding genes within these confidence intervals ranged from one to 36. For example, two significant associations were detected among the CMM-QMP traits at the genus-level, including Pelomonas and Streptococcus at chromosomes 11 and 16, for which the confidence intervals contained one and eight genes, respectively. For ddPCR-DNA and -RNA, Lactobacillus and Pelomonas were each significantly \((p < 0.05)\) associated with a single genomic locus (Fig. 2, Table 1).
### Analysis Of Candidate Genomic Regions

The narrow confidence intervals afforded by the AIL population allowed us to identify several promising candidate genes. In Table 2, we report a list of candidate genes from significant QTLs and their functions, summarized from experimental evidence. Most genes within the candidate regions are related to immune response, inflammatory response, cell apoptosis, and/or DNA repair. A number of genes are notable due to their role in lung functioning and disease susceptibility. For the QTL on chromosome 1 associated with variation in *Lactobacillus* ddPCR-DNA, the *Mk2* (mitogen-activated protein kinase-activated protein kinase 2) and the *Il10* (*Interleukin 10*) genes are the two closest genes in proximity to peak SNP UNC1677482. In humans, MK2 is a downstream product of the p38MAPK pathway acting as a pro-inflammatory kinase, and induces various signals such as cytokines in response to lipopolysaccharide (LPS)- and virus-induced infections [53–56]. MK2 is involved in transcriptional and post-transcriptional regulation of cytokine...
expression and was previously shown to affect the stability of \textit{Il10} transcript [54]. IL-10, a well-known anti-inflammatory cytokine, is e.g., also involved in mitigating disease severity in \textit{Mycobacterium tuberculosis} infection [57]. For the \textit{Pelomonas} (ddPCR-RNA) QTL on chromosome 4, the peak SNP lies within an intergenic region. However, several genes are found within the confidence interval, including \textit{Pou3f2} and \textit{Mms22l}. POU family transcription factors were previously shown to be highly expressed in small-cell lung carcinoma (SCLC) cell lines and to contribute to neuroendocrine differentiation in non-small cell lung carcinoma (NSCLC) cell lines[58]. \textit{MMS22L} was shown to accelerate cancer cell growth in lung cancer cell lines [59]. In contrast to the other candidate genes, \textit{Klhl32} is poorly characterized. However, recent work by de Vries et al. [60] identified \textit{KLHL32} as a protein-coding gene that strongly associates with DNA methylation levels of a specific CpG-site (a cytosine base located adjacent to a guanine base) in patients with chronic obstructive pulmonary disease (COPD).
Table 2  
List of candidate genes within confidence intervals and their functions.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Gene</th>
<th>Functions (&amp; references)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA</td>
<td>Enterobacteriaceae / Enterobacteriales</td>
<td><em>Chrm</em>, <em>Chrng</em>  Nicotine dependence, smoking behaviors, lung cancer, COPD [61]; lung hypoplasia [62]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Ecel1</em>  Restrictive lung disease (affected by respiratory muscles) [63]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Dis3l2</em>  Lung function [64]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Kcnj13</em>  Development &amp; physiology of the respiratory system [65]; tracheal tubulogenesis [66]</td>
</tr>
<tr>
<td>QMP</td>
<td>Deltaprotebacteria</td>
<td><em>Per2</em>  NSCLC [67, 68]; lung tumorigenesis [69]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Twist2</em>  Lung cancer [70]; lung adenocarcinoma [71]; Pneumonia [72]</td>
</tr>
<tr>
<td></td>
<td><em>Pelomonas</em></td>
<td><em>Ptchd3</em>  Asthma [73]</td>
</tr>
<tr>
<td></td>
<td><em>Streptococcus</em></td>
<td><em>Grik1</em>  Lung metastasis (of colorectal cancer in vivo) [74]; lung cancer [75]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Bach1</em>  Lung cancer [76–78]; cystic fibrosis [79]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Map3k7cl</em>  NSCLC [80]; pulmonary cell development [81]</td>
</tr>
<tr>
<td></td>
<td><em>ASV8_Propionibacterium</em></td>
<td><em>Nalcn</em>  Respiratory rhythm [82]; NSCLC [83]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Fgf14</em>  Lung Adenocarcinomas [84, 85]; lung functioning &amp; phenotype [86]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Ubac2</em>  COPD [87]; cystic fibrosis [88]; asthma [89]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Zic2</em>  Lung adenocarcinoma [90]; NSCLC [91]; SCLC [92, 93]</td>
</tr>
<tr>
<td>ddPCR-DNA</td>
<td><em>Lactobacillus</em></td>
<td><em>Mk2</em>  Lung cancer [94]; inflammatory pulmonary diseases [95]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Il10</em>  Tuberculosis [96]; asthma [97]; NSCLC [98, 99]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Il19</em>  Antimicrobial defense in airway epithelial cells [100]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Plgr</em>  COPD-like phenotype airway inflammation [101]</td>
</tr>
<tr>
<td>ddPCR-RNA</td>
<td><em>Pelomonas</em></td>
<td><em>Pou3f2</em>  NSCLC [58]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Mms22l</em>  Lung carcinogenesis [59]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Klh32</em>  COPD [60]</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Fut9</em>  Bronchopulmonary dysplasia [102]</td>
</tr>
</tbody>
</table>

Key: COPD: chronic obstructive pulmonary disorder, NSCLC: non-small cell lung cancer, SCLC: small cell lung cancer

**Evaluation of Lactobacillus in a Il10 knockout model**

Given the promising association detected between *Lactobacillus* load (DNA) and a locus containing the well-known anti-inflammatory cytokine IL-10, we aimed to confirm this potential gene-microbe association in an *Il10* knockout model using the same phenotyping method as for our mapping population. We accordingly performed ddPCR to quantify the
*Lactobacillus* load at the DNA level using genus-specific primers in *II10*+/+, *II10*+/− and *II10*−/− mice. Interestingly, we observe significant differences in *Lactobacillus* according to genotype, with *II10*+/+ mice displaying higher abundances than *II10*+/− and *II10*−/− mice (Fig. 4).

**Discussion**

In this study, we applied a rigorous combination of experimental- and data curation procedures that enabled us to carry out the first systematic assessment of host genetic effects on the mammalian lung microbiota. From the panel of traits defined for mapping, we identified seven significant and six suggestive host gene-microbial associations. Importantly, we found that incorporating quantitative bacterial load estimates in defining microbial phenotypes to be more effective in identifying gene-microbe associations than 16S rRNA gene amplicon profiles alone. While we identified only two significant associations among CMM-RA traits, five further significant associations were revealed after incorporating bacterial load information (i.e., CMM-QMP and ddPCR). Previous studies of the lung microbiota have identified two of the bacterial taxa involved, *Lactobacillus* [32, 49, 104] and *Pelomonas* [43], and these may represent key inhabitants of this host habitat. Moreover, we confirmed an association between *Lactobacillus* and a genomic region containing the *II10* gene in an independent *II10* knockout mouse model.

*Lactobacillus* are known to inhabit the mammalian lung and are generally regarded as probiotic bacteria [48]. Additionally, previous research showed *Lactobacillus* to modulate host immune responses and to reduce injury during lung infection [105-107]. Our targeted mapping approach revealed *Lactobacillus* load to be associated with a region containing the important candidate genes *II10* and *Mk2*. The functional relationship between *Lactobacillus* and these genes remains unclear, although a number of studies suggest a clinically relevant link. A previous study administered the *Lactobacillus casei* strain Shirota (LcS) to 2-week-old mice, which were then challenged with ovalbumin to induce allergic symptoms in the lungs [108]. This revealed modified immune response, including increased IL-10 levels, in mice administered LcS compared to controls. Although it seems that the effect is dependent on species or strain [109, 110], research continues to demonstrate a beneficial role of *Lactobacillus* in immunity. For example, nasal administration of *L. casei* aided lung recovery from *S. pneumoniae* infection [107], and oral administration of *L. rhamnosus* caused changes in cytokine levels that aided recovery from lung injury and inflammation induced by a synthetic analog of viral double stranded RNA, poly(I:C) [106]. Given this evidence, it is possible that *Lactobacillus* and/or its metabolites modulate the phosphorylation of p38MAPK, perhaps through bacterial byproducts or components of the bacterial cell wall [111, 112], which affect downstream signals that support lung function and recovery. Alternatively, *Mk2* gene expression and subsequent IL-10 production may affect the growth or modify the abundance of *Lactobacillus* in the lung.

Likewise, *Pelomonas* was previously identified as a member of the murine lung bacterial community [43]. In humans, recent studies identified *Pelomonas* in the oropharynx of patients infected with SARS-CoV-2 virus [113] and also in breast cancer patients [114], with associations with multiple cytokines and immune genes. Yet, the role of *Pelomonas* in the mammalian lung remains largely unknown. A potential mechanism might be via γ-aminobutyric acid (GABA) found in the pulmonary neuroendocrine cells (PNEC) distributed along the alveolar airway epithelium [115]. GABA acts as a mediator in mucus production and airway smooth muscle toning and contraction [116–118]. Interestingly, *Pelomonas* was found to contribute to contraction frequency in hydra [119]. Although the potential interplay between *Pelomonas* and GABA in the lungs awaits experimental investigation, we speculate that this may contribute to crosstalk between the host and bacteria. Moreover, given its association with *Pou3f2* and *Mms22l*, *Pelomonas* might play a role in susceptibility to lung cancer and other related lung diseases. Additionally, POU transcription factors are specifically expressed in small cell lung cancer (SCLC), contributing to accelerating cell growth, and *POU3F2* was revealed to maintain the proneuronal/neuroendocrine phenotype of SCLC [120]. *MMS22L* was found to be over-expressed in clinical
and esophageal cancers, playing a role in growth and survival of cancer cells [59]. This gene might impact the efficacy of DNA-damaging agents, as the knockout of the gene enhances cancer cell apoptosis [59]. Based on these observations, both Pou3f2 and Mms22l might serve as cancer therapy targets, which could be aided by the mechanistic understanding of a possible interaction with Pelomonas.

Other QTL intervals also include interesting genes associated with lung or respiratory tract development, functioning, and/or diseases that are potentially supported by the host-microbe interactions. Enterobacteriaceae and Enterobacteriales were found to be associated with QTL on chromosome 1 with a narrow confidence interval of 1.79 Mb. Within this interval, Chmd and Chmg were previously shown to contribute to nicotine dependence [61], and Kcnj13 was shown to take part in smooth muscle morphogenesis and trachea development in the mouse respiratory tract [66]. Yin et al. [66] found that mice deficient in Kcnj13 developed shorter trachea due to loss of function of the potassium channel KCNJ13, which is critical during trachea tubulogenesis. Another candidate region on chromosome 16 is associated with variation in Deltaproteobacteria abundance. Within this interval, the core circadian clock gene Per2 was found near the peak SNP. Interestingly, members of the Per subfamily act as tumor suppressor genes in mice, and the downregulation and loss of PER2 is associated with tumor proliferation and metastasis, including NSCLC [67–69]. In contrast, increased PER2 expression inhibits cell growth and NSCLC growth [67, 68].

The success of our mapping study relied on our approach to mitigate contamination and the employment of quantitative microbial profiling. Studies comparing ddPCR and qPCR find that while qPCR has a lower detection limit, it yields a worse signal-to-noise ratio when tested with negative controls [52], and yields inconsistent results when tested against intentionally contaminated samples diluted to replicate low-biomass conditions [121]. Further, Taylor et al. [122] note that ddPCR is also particularly advantageous for studies requiring long sample processing times, e.g. where a study cohort might be constructed slowly over time, as the method is robust to batch effects.

In our study, we find that incorporating ddPCR-based quantitative profiling improved the detection of gene-microbe interactions. In particular, it is known that QTL mapping significance thresholds are influenced by the underlying phenotype distribution [123]. Our significance thresholds were assigned based on random and repetitive shuffling of microbial phenotype values across genotypes. The probability of a given genotype/sample to be randomly assigned its original phenotype value is higher when interindividual variation is low, ultimately leading to higher significance thresholds. Notably, our QTL analysis using CMM-QMP phenotypes were assigned lower QTL significance thresholds compared to the QTL analysis using CMM-RA phenotypes (see e.g., Pelomonas CMM-RA vs. CMM-QMP QTL on chromosome 11, Fig. 3), as variation within the CMM-QMP is greater than that within the CMM-RA (Additional file 1: Tables S1-2). Thus, it is possible that the increase in interindividual variability among CMM-QMP traits, made possible through ddPCR measurements, better reflects the underlying biological distribution of the trait, which also contributes to lower significance thresholds.

The effectiveness of our mapping approach is congruent with previous employing absolute abundance profiling to characterize resident microbial communities [124–129]. Vandeputte et al. [47] found microbial load to be a key feature of Crohn's increase disease, with disease being characterized by a reduced bacterial load rather than differences in abundance among disease-associated taxa per se. Similarly, Sibila et al. [130] identified an association between bacterial load and airway inflammation in patients with bronchiectasis. A subsequent clinical trial conducted using the same patient cohort revealed that those with a high bacterial load showed significant improvement after receiving antibiotic therapy in comparison to those with a low bacterial load [130].

Conclusions
In summary, by combining the unique resource of a high-resolution mouse genetic mapping population together with experimental and computational advances in studying low biomass microbial communities, we demonstrate a novel role for host genetic variation in shaping lung microbiota composition. We find several promising associations for the commonly identified lung taxa *Lactobacillus* and *Pelomonas*, suggesting functional relevance for these taxa that may be exploited for future preventative/therapeutic purposes. These approaches outlined here may find useful application in future experimental models of host-microbe interactions in the lung.

**Methods**

**Animal husbandry**

The analysis of G\(^{15}\) intercross mice was approved by the "Ministerium für Energiewende, Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein" in Kiel, Germany (reference number: V 312–72241. 122–5 (12-2/09)). The G\(^{15}\) AIL mouse population was generated by intercrossing four strains, MRL/MpJ, NZM2410/J, BXD2/TyJ, and CAST/EiJ, with equal sex and strain distributions for 15 generations as previously described [27, 42]. The analysis of *Il10* KO and wildtype C57BL/6 mice was performed according to approved animal protocols and institutional guidelines of the Max Planck Institute for Evolutionary Biology in Plön. Mice were maintained and handled in accordance with FELASA guidelines and German animal welfare law (Tierschutzgesetz § 11, permit from Veterinäramt Kreis Plön: 1401–144/PLÖ–004697).

*Il10* KO and wildtype C57BL/6 mice (Jackson Laboratories, Maine, USA) were mated at age of 8-10 weeks to produce F\(^{1}\) mice, using both directions of the cross to reduce potential "grandmother," or legacy effects [131, 132]. Heterozygous F\(^{1}\) mice were mated within each cross at 10 weeks and included 11 pairs from each cross. F\(^{2}\) mice were weaned at 3 weeks; males and females were housed in separate cages according to family. Mice were maintained in individual ventilated cages (IVCs), type II long (Tecniplast, Greenline) in a specific pathogen-free facility (MPI für Evolutionsbiologie, Plön, Germany) with a 12-h light/dark cycle. Decalcified water and food (1324, fortified, from Altromin) were provided *ad libidum*. An average of ten mice (range: 9-15) were selected from each sex of each genotype from each cross for tissue extraction at 17 weeks of age.

**Sampling and nucleic acid extraction**

The entire lower respiratory tract was dissected and preserved in RNALater (Thermo Fisher Scientific) at 4°C overnight. Samples were stored at -20°C after removing RNALater. Approximately the bottom half of the left lobe was obtained for nucleic acid extraction, using tools sterilized with 70% ethanol, RNase-Away (Thermo Fisher Scientific), and sterilizing beads (Fine Science Tools) heated to at least 120°C. Lung tissues were first homogenized using Lysing Matrix E (MP Biomedicals) and nucleic acid extraction was conducted using the AllPrep 96 DNA/RNA kit (QIAGEN) with on-column DNase I treatment (QIAGEN), according to manufacturer's protocol, with the exception of TCEP instead of β-mercaptoethanol for the lysis step. A total of 40µL of RNA was eluted by adding 20µL of RNase-free water twice; 30µL of EB buffer was added twice for DNA elution. Concentrations were measured using NanoDrop 1000 (Thermo Fisher Scientific), with RNA samples diluted to equal concentrations (200ng/µL). Reverse transcription was performed according to the manufacturer's instructions using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) using 10µL of template RNA.

**16S rRNA gene amplicon sequencing**

The V1-V2 hypervariable region of the 16S rRNA gene was amplified using primers 27F and 338R, barcoded with unique eight-based MIDs (multiplex identifiers), using a dual indexing approach on Illumina MiSeq platform for both cDNA and DNA template [27]. Three negative extraction controls (NEC) from each extraction plate and one microbial community
DNA standard (20ng/µL) (ZymoBIOMICS) were included. Host reads were removed from DNA extracts using “Kneaddata” (v.0.6.1), which removes host reads based on the provided reference database, *Mus musculus*. However, the sequence qualities and read counts were not sufficient after this procedure, and we thus did not further proceed with these sequences generated from the extracted DNA (Additional file 3: Table S8). Sequences were processed using the “DADA2” package (v.1.14.1) [133] with the Ribosomal Database Project (RDP) training set 16 [134] for taxonomic classification, resulting in abundance tables of amplicon sequence variants (ASVs). Results were merged with metadata including sex, age, weight, and cage information, using “phyloseq” (v.1.30.0) [135] for further analysis in R (v.3.6.3).

**ddPCR**

A 20µL ddPCR master mix was prepared with QX200™ ddPCR™ EvaGreen® SuperMix (BioRad) following the manufacturer’s instructions (BioRad), with a final primer concentration of 120nM and with 10ng of nucleic acid template. PCR was performed on Bio-Rad C1000 Touch Thermal Cycler with the following conditions: 95°C for 5 min, 40 cycles at 95°C for 15 sec and 60°C for 1 min, 4°C for 5 min, 90°C for 5 min, and incubation at 10°C. Final products were transferred to QX200™ Droplet Reader and quantified as gene copies (per 20µL) using Bio-Rad QuantaSoft (v.1.7.4.0917). *Lactobacillus* and *Pelomonas* loads were quantified using genus specific primers on 242 DNA and cDNA samples each (Table 3).

**Table 3**

<table>
<thead>
<tr>
<th>Primer</th>
<th>5’ – 3’</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total bacteria</td>
<td>63F</td>
<td>GCAGGCCTAACACATGCAAGTC</td>
</tr>
<tr>
<td></td>
<td>355R</td>
<td>CTGCTGCTCCCCGTAGGAGT</td>
</tr>
<tr>
<td><em>Lactobacillus</em></td>
<td>F-Lacto</td>
<td>GAGGCAGCAGTAGGGAATCTTC</td>
</tr>
<tr>
<td></td>
<td>R-Lacto</td>
<td>GGCCAGTTACTACCTCTATCCTTCTTC</td>
</tr>
<tr>
<td><em>Pelomonas</em></td>
<td>357F</td>
<td>CGGGTTGTAAACCGCTTTGT</td>
</tr>
<tr>
<td></td>
<td>550R</td>
<td>CGGGGATTTTCACCTCTGTCT</td>
</tr>
</tbody>
</table>

**Contamination assessment and defining core measurable microbiota**

To assess for potential contamination, total bacterial load estimates were measured by ddPCR and used for the “frequency” method in the R package “decontam” (v.1.6.0) with a threshold of 0.1, which examines the frequency of ASVs in relation to input sample concentration to identify potential contaminants [34, 38]. ASVs identified as likely contaminants, including those belonging to *Halomonas* and *Shewanella*, were removed from the dataset prior to further analysis. After the decontamination process, samples were rarefied to even sampling depth of 4,000 reads per sample. A core measurable microbiota (CMM) was defined by bacterial taxa and ASVs present in at least 25% of the samples, with a minimum relative abundance threshold of at least 1%. The final CMM among 242 samples included 13 ASVs and 58 taxa from the genus- to phylum-level (CMM-RA).

**Quantitative microbiome profiling**

We used ddPCR-based total bacterial load (RNA level) estimates for quantitative microbiome profiling (QMP), whereby 16S rRNA gene relative abundances were corrected using bacterial load measurements and transformed to “absolute” or
quantitative abundances. For this, we used R function *rarefy_even_sampling_depth* (https://github.com/raeslab/QMP) [47], which rarefies samples to even sampling depth, defined as the ratio between sequencing depth and bacterial load (here, based on ddPCR). Among the QMP dataset (CMM-QMP), taxa and ASVs included in the CMM were further selected for mapping.

### QTL mapping

Prior to mapping, summary statistics were performed on all traits including CMM-RA, CMM-QMP, and ddPCR-DNA/RNA in R studio (v.1.2.1335) with R (v.3.6.3) (Additional file 1: Tables S1-3). In order to perform a log\(_{10}\) transformation of relative abundance values, a value of 0.5 was added to the absolute abundances of all CMMs prior to converting the absolute abundances into relative abundances. Then, linear mixed effects analysis was performed on these traits using “lme4” (v.1.1-10) [137]. Variance was estimated using “r.squaredGLMM” in “MuMIn” (v.1.43.6) [138] and “VarCorr” in “lme4” for fixed and random effects, respectively.

Linkage mapping was performed using “DOQTL” (1.6.0) [139] and “QTLRel” (0.2.14) [140] in R (v3.2.2), whereby we fit an additive model that regresses the log\(_{10}\)-transformed traits on the four founder haplotype contributions. A kinship matrix was defined using “kinship.probs” in “DOQTL” to adjust for the kinship between [27]. A 3D-array of founder haplotype contributions (sample # x 4 founders x marker #) [27] and kinship matrix, along with sex, age, and cage information as calculated above, were included for linkage mapping.

Permutations were run in R (v.3.2.2) for each trait by shuffling the genotypic data to define significance thresholds at both the 90\(^{th}\) and 95\(^{th}\) percentiles of LOD scores [27]. Permutations were run 10,000 times, a ten-fold increase from minimum recommendations by Gatti et al. [139]. QTL confidence intervals were defined at 1.5 LOD score drops on either side of the QTL peak. After QTL mapping, genes located within confidence intervals were examined and then plotted using “get.mgi.features” and “gene.plot” in “DOQTL” to identify potential candidate genes.

### Abbreviations

AIL: Advanced intercross line; ASV: Amplicon sequence variant; BAL: Bronchoalveolar lavage; CMM: Core measurable microbiota; ddPCR: Droplet digital PCR; LOD: Logarithm of odds; PCR: Polymerase chain reaction; QMP: Quantitative microbiome profiling; qPCR: quantitative PCR; QTL: Quantitative trait loci; RA: Relative abundance.

### Declarations

**Ethics approval and consent to participate**

The analysis of G\(_{15}\) intercross mice was approved by the “Ministerium für Energiewende, Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein” in Kiel, Germany (reference number: V 312– 72241. 122–5 (12-2/09)). The analysis of Il10 KO and wildtype C57BL/6 mice was performed according to approved animal protocols and institutional guidelines of the Max Planck Institute for Evolutionary Biology in Plön. Mice were maintained and handled in accordance with FELASA guidelines and German animal welfare law (Tierschutzgesetz § 11, permit from Veterinäramt Kreis Plön: 1401–144/PLÖ–004697).

**Consent for publication**

Not applicable

**Availability of data and material**
Sequencing data generated and analyzed is available in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA856849 (https://dataview.ncbi.nlm.nih.gov/object/PRJNA856849?reviewer=3ko45p5hm7jq0b3bkn2vstedtc).

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

This work was supported by the German Science Foundation (DFG) grants EXC2167 “Precision Medicine in Chronic Inflammation (PMI)”, CRC 1182, and the Leibniz Science Campus Evolutionary Medicine of the Lung (EvoLUNG).

**Authors’ contributions**

JFB and SI developed and designed the study and acquired funding. YG designed the pipeline for QTL mapping. CJC conducted the study experiments and was responsible for methodologies and investigations, data analyses and visualization, and software implementation. CJC, BMH, MB, and JFB contributed to formal data interpretation and study validation. Supervision was provided by JFB and MB. Original draft preparation was by CJC, BMH, and JFB with input from MB. All authors contributed to final review and edits.

**Acknowledgements**

We thank Sven Künzel for assistance in lab work and mouse breeding; Anastasia Vock and Anika Jonas for their help in the animal facility; Marie Vallier for helpful discussions; and Shauni Doms for assistance in bioinformatic analyses.

**Authors details**

1 Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany. 2 Section of Evolutionary Medicine, Institute for Experimental Medicine, Kiel University, Arnold-Heller-Str. 3, 24105 Kiel, Germany. 3 Division of Nephrology, Department of Medicine, Columbia University Irving Medical Center, New York, NY, 10032, USA. 4 College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, UAE. 5 Research Center Borstel, Evolution of the Resistome, Leibniz Lung Center, Parkallee 1-40, 23845 Borstel, Germany.

**References**


25. Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, et al. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of


72. Ding N, Liu D, Duan X, Zhang J, Ma S, Chen Y. Twist2 Reduced NLRP3-Induced Inflammation of Infantile Pneumonia via Regulation of Mitochondrial Permeability Transition by FOXO1. International Archives of Allergy and Immunology. 2022:1-16.


Lung bacterial community composition of AIL mice: **A** phylum level, **B** genus level. The two most abundant genera that are classified to the genus-level include *Lactobacillus* and *Pelomonas* (shown in **bold** in legend)—these taxa were...
subsequently selected as bacterial candidate traits for targeted linkage mapping.

Figure 2

Manhattan plots from linkage mapping of bacterial traits. From inner to outer circles: *Enterobacteriales* CMM-RA, *Enterobacteriaceae* CMM-RA, *Deltaproteobacteria* CMM-QMP, *Pelomonas* CMM-QMP, *Streptococcus* CMM-QMP, *ASV8_Propionibacterium* CMM-QMP, *Pelomonas* ddPCR-RNA, and *Lactobacillus* ddPCR-DNA. CMM-RA traits are shown in yellow, CMM-QMP in blue, and ddPCR-DNA/RNA in green. Peak SNPs are highlighted with red dots. LOD score on the y-axis indicates the \(-\log p\) value of the association between a locus and a phenotypic trait.
Overview and comparison of the Manhattan plots of *Lactobacillus* and *Pelomonas* traits from CMM-RA, CMM-QMP, and ddPCR-DNA and ddPCR-RNA, with each column dedicated to the chromosome where the QTL was detected; *Lactobacillus* ddPCR-DNA at chromosome 1, *Pelomonas* ddPCR-RNA at chromosome 4, and *Pelomonas* CMM-QMP at chromosome 11, from left to right. Solid black lines indicate significant thresholds and dashed lines indicate suggestive thresholds of each trait. Red vertical lines indicate confidence intervals of the QTL.

Figure 3
Figure 4

*Lactobacillus* load according to *Il10* genotype. Bacterial loads determined by ddPCR (DNA) were log$_{10}$-transformed and compared using Kruskal-Wallis tests with Wilcoxon signed-rank post-hoc tests. *P* values were corrected for multiple testing according to Benjamini and Hochberg [103]. Error bars depict standard error. *p* < 0.01 **, *p* < 0.05 *, not significant “ns”.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx
- Additionalfile2.xlsx
- Additionalfile3.xlsx
- GraphicalabstractG15QTL.png