A cross-sectional comparison of gut metagenomes between dairy workers and community controls

Pauline Trinh[1,3], Sarah Teichman[2], Marilyn C. Roberts[3], Peter M. Rabinowitz[3] and Amy D. Willis[1,2*]

Abstract

Background As a nexus of routine antibiotic use and zoonotic pathogen presence, the livestock farming environment is a potential hotspot for the emergence of zoonotic diseases and antibiotic resistant bacteria. Livestock can further facilitate disease transmission by serving as intermediary hosts for pathogens before a spillover event. In light of this, we aimed to characterize the microbiomes and resistomes of dairy workers, whose exposure to the livestock farming environment places them at risk for facilitating community transmission of antibiotic resistant genes and emerging zoonotic diseases.

Results Using shotgun sequencing, we investigated differences in the taxonomy, diversity and gene presence of 10 dairy farm workers and 6 community controls' gut metagenomes, contextualizing these samples with additional publicly available gut metagenomes. We found no significant differences in the prevalence of resistance genes, virulence factors, or taxonomic composition between the two groups. The lack of statistical significance may be attributed, in part, to the limited sample size of our study or the potential similarities in exposures between the dairy workers and community controls. We did, however, observe patterns warranting further investigation including greater abundance of tetracycline resistance genes and prevalence of cephamycin resistance genes as well as lower average gene diversity (even after accounting for differential sequencing depth) in dairy workers' metagenomes. We also found evidence of commensal organism association with tetracycline resistance genes in both groups (including *Faecalibacterium prausnitzii*, *Ligilactobacillus animalis*, and *Simiaoa sunii*).

Conclusions This study highlights the utility of shotgun metagenomics in examining the microbiomes and resistomes of livestock workers, focusing on a cohort of dairy workers in the United States. While our study revealed no statistically significant differences between groups in taxonomy, diversity and gene presence, we observed patterns in antibiotic resistance gene abundance and prevalence that align with findings from previous studies of livestock workers in China and Europe. Our results lay the groundwork for future research involving larger cohorts of dairy and non-dairy workers to better understand the impact of occupational exposure to livestock farming on the microbiomes and resistomes of workers.

Keywords Shotgun metagenomics, Livestock farming, Antibiotic resistance, Microbiome, Virulence factors

*Correspondence:
Amy D. Willis
adwillis@uw.edu
Full list of author information is available at the end of the article

Trinh *et al. BMC Genomics* (2024) 25:708

Page 2 of 18

Background

Next-generation sequencing has facilitated the study of entire microbial communities of culturable and unculturable microorganisms, revealing the profound impact that the human gut microbiome has on immune homeostasis [1–3], disease development [4–7], and even resistance to pathogen invasion [8–11]. The human gut microbiota is influenced by both host genetics [12, 13] and environmental factors, including diet [14, 15], geography [16], and medications [17, 18]. Recent research suggests that environmental factors outweigh host genetics in shaping the gut microbiome [19, 20]. Consequently, environments that are rich in antibiotic resistant organisms, antibiotic residues, antibiotic resistance genes (ARGs), and/or zoonotic pathogens, such as livestock farms, may pose significant risks to public health, as these environments may serve as hotspots for antibiotic resistance and zoonotic disease emergence and propagation [21, 22]. Studies examining changes in the human microbiome and resistome in response to occupational exposure to livestock on farms may shed light on the potential risks of these environments for transmission and spread of zoonotic diseases and antibiotic resistance.

Modern farming practices and agricultural intensification have been linked to the emergence and amplification of zoonotic diseases and antimicrobial resistance (AMR), with livestock potentially serving as intermediate hosts for pathogens [23, 24]. Transmission of both zoonotic pathogens and antibiotic resistance genes can occur through direct or indirect contact at the human-animal interface, placing livestock workers and those in contact with these workers at risk of transmission and infection [25, 26]. Several shotgun metagenomic studies have looked at the effect of occupational exposure to animal agriculture on ARG carriage, finding higher prevalence of ARGs as well as evidence of transmission of ARGs from animal farming environments to workers [27–29]. While these studies highlight some potential impacts of exposure to ARG-rich animal farming environments, they either focused primarily on understanding the presence of ARGs in total community DNA without contextualizing ARGs to particular species of bacteria, or they used cultured isolates of a single bacterial species (e.g., *Escherichia coli*) to understand species-level antibiotic resistance transmission [27–29]. Furthermore, these studies did not examine virulence factor genes, which encode for functions that can cause disease and assist an organism with persisting within a host [30]. While virulence factors have historically been associated with pathogens [30] they have also been identified on commensal or non-pathogenic genomes [31, 32], and their transmission can occur between pathogens and commensals by mobile genetic elements transmission [33, 34].

To better understand the effect of the livestock farming environment on the human gut microbiome of workers — including virulence factors, taxonomic associations of ARGs, and the role of commensal organisms in ARG transmission — we compared dairy worker and community control gut microbiomes using shotgun metagenomic sequencing. We studied differences with respect to diversity, taxonomic composition, and the carriage of virulence factor and antibiotic resistant genes. We additionally evaluated potential taxonomic affiliations of genes conferring resistance to beta-lactams (cephamycin and cephalosporins) and tetracyclines through reconstruction of their genomic context, and assessed differences in taxonomic context based on group association.

Materials and methods

Study participant selection

We performed metagenomic sequencing on a subset of stool samples from participants in the Healthy Dairy Worker study. The Healthy Dairy Worker study is a prospective cohort study that focuses on the effects of dairy farm exposure on the fecal and nasal microbiome, and immune and respiratory function of dairy farm workers. The study began recruitment of subjects on a rolling basis in May 2017 and involves collection of fecal and nasal samples, as well as health history data on participants at baseline enrollment, 3, 6, 12, and 24 months. Dairy workers were recruited from 3 conventional large ($> 5,000$ animals) farms in the Yakima Valley of Eastern Washington State and community controls were recruited from surrounding communities. Recruitment of both community controls and dairy workers was done through snowball sampling where research participants assisted in identifying other potential participants. Eligibility to be a participant as a dairy worker required subjects to have been working on a dairy farm for at least 6 months. Eligibility as a community control required participants to have no prior work experience on a dairy farm in the previous 5 years, to have not lived on a dairy farm, and to have no current household member who worked on a dairy farm in the previous 5 years. Participants were consented by bilingual study staff and received an incentive payment for enrollment and subsequent sampling. Participants were asked to participate in self-reported surveys collecting information on health and work history. Sample collection and study activities were approved by the University of Washington Institutional Review Board under STUDY00000042. Study protocols have been previously described [35].

To conduct the current cross-sectional metagenomics study, we selected shotgun sequencing data of 16 fecal samples (the maximum possible with budget constraints) taken from the Healthy Dairy Worker study cohort.

Trinh *et al. BMC Genomics* (2024) 25:708

Page 3 of 18

These samples came from 10 dairy workers and 6 community controls, all sampled at baseline enrollment. We selected the 10 dairy worker samples through simple random sampling of study subjects that met our exclusion criteria (no antibiotic use within 3 months of baseline enrollment). All dairy worker samples were selected from workers on a single farm, and all identified as white Hispanic or Latino males (both the numbers of females working on the participating dairy farms and recruitment of females into the study was low). Selection of the 6 community control samples was done using simple random sampling among community participants who had no antibiotic use within 3 months of sample collection and baseline enrollment, and who covariate-matched our dairy workers on sex and ethnicity. The unbalanced sampling of each group was designed to over-sample dairy workers, as community control samples could be supplemented with additional healthy subjects' metagenomics data from publicly available data (i.e., The Human Microbiome Project).

Study enrollment and baseline sample collection began in 2018 for these 16 participants. The collection of study samples occurred at least one year after the Food and Drug Administration completed implementation of the Guidance for Industry (GFI) no. 213 which restricted the use of antibiotics in animal agriculture for growth promotion purposes and transitioned medically important antibiotics used in drinking water and feed from over-the-counter status to Veterinary Feed Directive (VFD) or prescription status [36, 37].

Sampling, shotgun metagenomic library preparation and sequencing
Stool samples were self-collected by participants using a stool specimen collection kit. Participants were instructed to store stool samples in their refrigerators and to return their stool samples within 24 hours of collection to study staff. Samples were stored at −20°C by field staff at a partner study site for 1-6 months before before being packaged with dry ice and transported to the University of Washington for extraction and storage at −20°C. DNA extraction was performed using the MoBio DNeasy PowerLyzer PowerSoil Kit (Qiagen) following manufacturer's protocols, and quantification of the resulting DNA was conducted using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher/Invitrogen). Extracted DNA samples were packaged on dry ice and transported to the Fred Hutchinson Cancer Research Center for sequencing.

Sequencing libraries were prepared from 250pg gDNA with a quarter reaction workflow using the Nextera XT Library Prep Kit (Illumina, San Diego, CA) and 12 cycles of indexing PCR. Indexed libraries were pooled by volume and post-amplification cleanup was performed with 0.8X Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN). The library pool size distribution was validated using the Agilent High Sensitivity D5000 ScreenTape run on an Agilent 4200 TapeStation (Agilent Technologies, Inc., Santa Clara, CA). Additional library QC and cluster optimization was performed using Life Technologies- Invitrogen Qubit® 2.0 Fluorometer (Life Technologies-Invitrogen, Carlsbad, CA, USA). The resulting libraries were sequenced on the Illumina HiSeq 2500 to generate paired-end 150nt sequences for each fragment. Image analysis and base calling were performed with Illumina Real Time Analysis software v1.18.66.3, followed by demultiplexing of dual-indexed reads, removal of adapters and primers, and generation of FASTQ files with bcl2fastq Conversion Software v1.8.4 [38].

Profiling taxonomic composition
We performed profiling of the microbial composition of the metagenomic short reads of our dairy workers and community control samples with primers, adapters, and host sequences removed using `MetaPhlAn3` v3.0.14 [39]. `MetaPhlAn3` estimates relative abundances by mapping reads to a reference database of clade-specific marker genes from `ChocoPhlAn` v30 (published in January 2019) [39, 40]. `MetaPhlAn3` performs this read mapping against marker genes using `bowtie2` v2.3.5.1 [41, 42]. Default parameters were used when running `MetaPhlAn3` with an additional flag `-t rel_ab_w_read_stats` for outputting relative abundances with estimated number of reads mapping to each clade.

Metagenomic assembly and processing of contigs
We conducted *de novo* assembly and processing of contigs using `anvi'o` v6.2 [43]. `anvi'o` integrates a suite of bioinformatics tools for the processing, analyzing, and visualization of metagenomics, pangenomics, and phylogenomics studies. We used the `anvi'o` Snakemake [44] metagenomics workflow obtained from "anvi-run-workflow" [45] with "–workflow metagenomics" to conduct our metagenomic assembly and processing of contigs. `Illumina-utils` [46] was used to apply the guidelines of Minoche et al. [47] with a default parameter $p = 0.75$ for quality filtering of reads based on Q-scores to trim and identify low quality reads. Details on the exact trimming and quality score filtering guidelines have been described in the literature [46, 47]. Further processing of the individual assemblies using `anvi'o` v6.2 included removal of human host reads using a GRCh38 reference, performing individual assembly of contigs for each sample metagenome [43] using `MEGAHIT` v1.2.9 [48], identifying open read frames (ORFs) using `Prodigal` v2.6.3 [49], predicting gene-level taxonomy using

Trinh *et al. BMC Genomics* (2024) 25:708

Page 4 of 18

Centrifuge [50], functional annotation of genes using NCBI's Clusters of Orthologous Groups (COGs) [51] and `Pfams` [52], searching for sequences using `DIAMOND` v0.9.14 [53], identifying single copy core genes (SCGs) using `HMMER` v3.3 [54] and built-in `anvi'o` Hidden Markov Model (HMM) profiles for bacteria and archaea, recruiting reads using `bowtie2` v2.3.5.1 [41], and generating BAM files with `samtools` v1.10 [55]. Prediction of the approximate number of genomes in a metagenomic assembly using SCGs was done using the anvi'o script "anvi-display-contigs-stats". Workflows using Snakemake with full parameter details can be found at the URL https://github.com/statdivlab/hdw_mgx_supplementary/.

Metagenome annotation of virulence factors and antibiotic resistance genes

We used `ABRicate` v1.0.1 [56] to perform a mass screening of our *de novo* assembled gene calls identified from our assembled contigs for antibiotic resistance genes and virulence factor genes. `ABRicate` uses the Basic Local Alignment Search Tool (BLAST) [57] to annotate genes from a user-specified reference database. We used the Virulence Factor Database v6.0 [58] and the Comprehensive Antibiotic Resistance Database (CARD) v4.0 [59] as reference databases in our search. Genes were considered present in a given metagenome if they met conservative minimum thresholds of 90% identity and 100% coverage.

To compare gene abundances across samples we perform a normalization of gene abundances by creating a measure of relative gene abundance. Relative gene abundances were calculated within a metagenome by taking the mean coverage of a target ARG or VF gene divided by the sum of all mean coverages of all protein coding genes identified in a given metagenome. Here mean coverage indicates the average depth of coverage across a gene calculated by adding up the coverage of each nucleotide in a gene, and dividing by the length of the gene. ARG relative abundances were further aggregated by their antibiotic classes by summing the relative abundances of genes within each antibiotic class for each metagenome. We focused our analyses to antibiotic classes that were identified by the World Health Organization (WHO) as Critically Important Antibiotics (CIA) [60].

Reconstruction of genomic context of ARGs

We used our results from `ABRicate` to extract ARG target sequences from each metagenomic assembly. These sequences were extracted using `samtools` [55] and were used as "query" sequences in our genomic context reconstruction analyses. ARG query sequences were used to produce query neighborhoods that reassociated unassembled or unbinned reads that are graph-adjacent to the query sequence. To prepare our raw metagenomic short reads for genomic context reconstruction, we removed adapters and quality trimmed the reads using `fastp` [61] before removing human host reads using `bbduk` [62] and the masked human k-mer data [63]. Using our quality trimmed and filtered short reads and our query sequences of interest, we constructed the genomic context of each query sequence using `MetaCherchant` [64]. `MetaCherchant` uses a de Bruijn graph assembly approach to build genomic context of query sequences. We used the "environment-finder" tool in parallel and set k-mer length to 31, minimum coverage to 5, and max radius to 1000. Taxonomic annotation of sequences corresponding to graph nodes was done using `kraken2` v2.1.2 [65]. Taxonomic affiliation of genes was based on `kraken2` annotations of surrounding graph nodes for a particular query sequence.

In order to understand whether the ARGs identified using `ABRicate` have historically been found on plasmids or microbial chromosomes, we cross-referenced the ARGs identified in our study with ARGs identified by the Resistance Gene Identifier (RGI) v5 [66]. The `RGI` integrates with the CARD database to predict AMR genes and their mutations in complete chromosome sequences, predicted genomic islands, complete plasmid sequences, and whole genome shotgun assemblies taken from National Center for Biotechnology Information (NCBI) databases. This is accomplished through prediction of ORFs using Prodigal [49], alignment to CARD reference sequences using either `BLAST` [57] or `DIAMOND` [53], and the use of either protein homolog or protein variant models. The results from `RGI`'s exhaustive search are maintained and updated for each antibiotic resistance gene catalog on the CARD database.

Comparison with the Human Microbiome Project

To contextualize our study cohort, we also considered data from the Human Microbiome Project (HMP). Information on the study's protocol, sampling, and sequencing procedures have been previously described [67–69]. Briefly, DNA was extracted using the Mo Bio PowerSoil DNA Isolation Kit and nucleic acid samples were quantified and checked for purity of the DNA with only samples with a minimum of 50-100 ng of DNA used. Libraries were prepared following a standard protocol from Illumina with modifications outlined in detail by the HMP study [69]. Processing of raw reads into contigs was conducted by the HMP study in the following steps: (1) sequencing of raw reads was performed using the Illumina GAIIx platform with 101bp paired-end reads, (2) all samples were screened for human contamination using NCBI's BMTagger tool, with $\sim$ 49% of reads targeted for

Trinh *et al. BMC Genomics* (2024) 25:708

Page 5 of 18

removal as human, (3) samples underwent quality control assessments, including identification of outliers by mean contig and ORF density, human hits, rRNA hits and size, and (4) samples passing QC were assembled into contigs using IDBA-UD v1.1.0.

To identify samples, we utilized the `curated-MetagenomicData` (cMD) package [70], which provides curated and uniformly processed microbiome data. Raw sequences are downloaded by the `cMD` team and processed to produce taxonomic profiles using Metaphlan3 v3.0.0. We filtered for samples from the original HMP "Healthy Human Subjects" (HHS) study, and subset to first visits. In order to match our all-male cohort, we further subset to only male samples, resulting in 47 HMP samples for comparison. For taxonomic analyses we utilized abundance data from `cMD`, which used `MetaPhlAn3` v3.0.0. Identification of antibiotic resistance genes and virulence factor genes was conducted using assembled contigs from the HMP portal, which were imported into `anvi'o` v6.2, where `Prodigal` v2.6.3 was used to identify ORFs. Annotation of ORFs for virulence factors and antibiotic resistance genes from the VFDB v6.0 and CARD v4.0 databases was conducted using `ABRicate` v1.0.1. We note that while information on occupation is not publicly available for the 47 HMP participants, the distribution of job types across the entire HMP cohort did not include any categories related to farm or animal work, and no more than 15% indicated an "other" category of occupation [71, 72]. Additionally, recruitment of individuals from the HMP focused on individuals in the general populations of two U.S. cities, St. Louis, Missouri and Houston, Texas [67]. We conclude that it is unlikely that any of the 47 HMP participants used in our study have occupational exposure to livestock farming. Information on race and ethnicity were not publicly available at the individual level for HMP participants, and thus, we were unable to adjust for race and ethnicity in our regression analyses.

Statistical analyses

Differences in demographic and sequencing characteristics between groups were evaluated using two-sample t-tests allowing for heteroskedasticity (continuous variables) and χ^2 tests for independence (categorical variables). To test for differential abundance (at the species and phyla level) between dairy workers and community controls, we used `radEmu` v1.2.0 [73] with species coverages as the response and an indicator for dairy work (the predictor of interest) and age (a potential confounder) as predictors. We conducted robust score tests and applied a false discovery rate (FDR) correction using the `qvalue` v2.26.0 [74] package. Secondary analysis of species differential abundance that included both data from this study as well as HMP data were performed similarly, but with the inclusion of an indicator for the participant being from the HMP study, thus adjusting for cohort and batch effects. `radEmu` accounts for differential sequencing depth, and is robust to the differential detectability of bacteria taxa and unobserved species in samples. By estimating fold-changes, `radEmu` addresses limitations of analyzing bacterial proportions from high-throughput sequencing data [75].

For our α-diversity analysis, we estimated the Shannon Diversity Index (SDI) using the `DivNet` v0.4.0 [76] model applied to `MetaPhlAn3` relative abundances using an identity design matrix. We compared the SDI of dairy worker and community control metagenomes with `betta` [77], using an indicator for dairy work and age as predictors. Our β-diversity analysis estimated Bray-Curtis dissimilarities using `MetaPhlAn3` relative abundances, and tested for differences in β-diversity using the `testBetaDiversity` function as implemented in `DivNet` with an indicator for dairy work as a predictor. This test was performed using a bootstrapped pseudo-F test with 10,000 bootstrap iterations. `DivNet` accounts for uncertainty in estimating SDI and Bray-Curtis dissimilarity arising from sample-to-sample variation and taxon co-occurrence. We estimated gene richness using `breakaway` v4.8.2 [78] and compared gene richness between dairy worker and community control metagenomes using function `betta` as integrated in `geneshot` v0.6.2 [79, 80], using an indicator for dairy work as a predictor. In addition to our gene richness analysis, we investigated species richness using single-copy core genes and `breakaway` and `betta`, testing for differences in species richness between dairy worker and community control metagenomes. Estimating gene and genome richness with `breakaway` accounts for unequal sequencing effort across samples by estimating unobserved species, and comparing estimated richness using `betta` accounts for uncertainty and heteroskedasticity in total richness estimates.

Relative abundances of ARGs were calculated using the mean coverage for a given target gene divided by the sum of all mean coverages of all the protein coding genes identified in a given metagenome. We applied a centered log-ratio (CLR) transformation on our relative abundance data (pseudocount 1×10^{-15}), filtering out genes in the database that had zero abundance across all samples. Differential abundance testing of antibiotic resistance genes in dairy workers compared to community controls was performed using linear regressions with CLR-transformed relative abundances as the response and an indicator for dairy work (the predictor of interest) and age (as a potential confounder) as predictors. We use robust standard errors implemented in `rigr` v1.0.4 [81].

Trinh *et al. BMC Genomics* (2024) 25:708

Page 6 of 18

This approach targets a similar estimand to `radEmu`, but can be parallelized for large-scale analysis [75]. Testing for differential presence of virulence factor genes and ARGs between dairy workers and community controls was performed using `happi` v0.8.7 [82], adjusting for depth as a quality variable. To conduct gene enrichment testing, we used `happi`'s likelihood ratio testing (LRT) procedure with default parameters, setting the number of permutations used to 1000 and contamination probability $\varepsilon = 0.05$. Comparisons of gene presence (response) between dairy workers and community controls using `happi` included an indicator for dairy work and age as predictors. For differential gene enrichment testing between our study cohort and the HMP participants we used indicators for dairy work and study cohort (which adjusts for batch/cohort variability) as well as age and BMI in the models. We performed FDR corrections for multiple comparisons for all tests using the `qvalue` v2.26.0 `R` package for q-value estimation. All statistical analyses were conducted using `R` v4.1.2.

Results

Study description

At baseline enrollment, the dairy worker cohort averaged 10 years (SD 5.2) of experience in the dairy industry. Compared to community controls, dairy workers were notably younger, with a mean age of 38.40 years compared to 49.50 years for controls (t-test $p = 0.06$, see Table 1). Both groups had similar proportions of current smokers (67% for controls vs. 70% for dairy workers, $\chi^2 = 1$). Furthermore, there were no statistically significant differences between dairy workers and community controls in terms of body mass index (BMI) or consumption habits of alcohol, dairy products, vegetables, eggs, beef, chicken, lamb, and fish (Table 1) at the 5% level. All community controls reported occupations as field workers in non-animal agriculture at the time of sample collection and study enrollment.

In our study cohort, the total number of paired reads in metagenomes ranged from 18–33 million per sample, with community control samples yielding significantly more paired reads compared to dairy worker samples (27.1×10^6 vs. 22.8×10^6, t-test $p = 0.04$, Table 2). An average of 88.3% (SD 1.1%) of paired reads corresponding to a mean of 24×10^6 paired reads (SD 4.3×10^6) (Table 2, Supplementary Table S1) passed quality filtering and trimming. Host reads made up 0.05%-1.55% of samples. Removal of host reads resulted in an average of 24.2×10^6 (SD 2.9×10^6) paired reads in community control samples and 20.1×10^6 (SD 4.0×10^6) paired reads in dairy worker samples used for our analyses.

To provide context for our study cohort, we compared our study participants with data from 47 healthy males participating in the Human Microbiome Project (HMP). Information available for these individuals was limited to age, sex, and BMI (Table 1). On average, HMP participants were notably younger than our study cohort (26.5 years vs. 42.6 years, t-test $p < 0.001$, see Table 1) and had lower BMI ($\chi^2 < 0.001$). Sequencing depths for the HMP males ranged from 21–239 million paired reads per sample. The average number of paired reads for the HMP cohort after quality filtering and host sequence removal was significantly higher than that of our study cohort (111×10^6 vs. 21.6×10^6, t-test $p < 0.001$, Table 2).

Taxonomic profiling of dairy worker and community control metagenomes

The 16 metagenomic samples were composed of 9 distinct phyla: Firmicutes, Bacteroidetes, Actinobacteria, Verrucomicrobia, Proteobacteria, Euryarchaeota, Spirochaetes, unclassified Eukaryota and Synergistetes. Of these phyla, Firmicutes, Bacteroidetes, and Actinobacteria were the 3 most abundant phyla found across all samples (Fig. 1A). The large representation of Firmicutes, Bacteroidetes, and Actinobacteria reflected similar community compositions observed in healthy subjects from the Human Microbiome Project [68]. We also note that while the majority of the phyla identified are from the domain Bacteria, we observed organisms from the domains Archaea (Euryarchaeota) and Eukaryota as well. We detected Euryarchaeota organisms in 5 dairy worker and 6 community control samples and unclassified Eukaryota organisms in low abundances in 2 dairy worker samples from our study. We conducted differential abundance testing comparing phylum-level differences between dairy workers and community controls, controlling for age, and found no significant differences at the 5% FDR level in phylum abundances (Supplementary Table S2).

At the species-level, we identified 272 different species across the 16 metagenomes. The most prevalent bacteria species observed were *Prevotella copri*, *Faecalibacterium prausnitzii*, *Eubacterium rectale*, *Ruminococcus bromii*, and *Bacteroides vulgatus* (Fig. 1, right). These five species have been previously shown to be highly abundant organisms found in healthy human gut microflora [83–87]. Differential abundance testing revealed no statistically significant differences in the abundances of these five organisms between occupational groups (adjusting for age) at an FDR of 5% (Supplementary Table S3). Abundance patterns for the five species with the most extreme test statistics (Supplementary Table S3) showed higher abundances of *Ruminoccocus lactaris* ($\hat{\beta} = 3.20$, $q = 0.19$) in dairy workers and higher abundances of *Methanobrevibacter smithii* ($\hat{\beta} = -1.12$, $q = 0.19$), *Methanosphaera stadtmanae* ($\hat{\beta} = -11.20$, $q = 0.19$),

Trinh *et al. BMC Genomics* (2024) 25:708

Page 7 of 18

Table 1 Demographic and behavioral characteristics of community controls (CC) and dairy workers (DW) from the Healthy Dairy Worker (HDW) cohort and participants from the Human Microbiome Project (HMP). *p*-values correspond to tests of equality of mean (continuous variables) or distribution (categorical variables) across dairy workers and community controls. Limited metadata (at the level of individuals) was made publicly available from the HMP participants, and has been denoted as n/a (data not available) where appropriate

	Community (N=6)	Dairy (N=10)	HMP (N=47)	*p*-value (DW vs. CC)
Age (years)				
Mean (SD)	49.5 (10.7)	38.4 (8.0)	26.5 (4.7)**	0.06
Body Mass Index				
18.5-24.9	0 (0%)	2 (20.0%)	33 70.2%)[a]	0.41
25-29.9	4 (66.7%)	4 (40.0%)	10 (21.3%)[a]	
30.0+	2 (33.3%)	4 (40.0%)	4 (8.5%)[a]	
Current Smoker				
No	4 (66.7%)	7 (70.0%)	n/a	1
Yes	2 (33.3%)	3 (30.0%)		
Alcohol Consumption				
Never	1 (16.7%)	2 (20.0%)	n/a	0.15
Rarely <1/week	4 (66.7%)	2 (20.0%)		
1-7 times/week	1 (16.7%)	6 (60.0%)		
Dairy Consumption				
Rarely <1/week	1 (16.7%)	2 (20.0%)	n/a	1
1-7 times/week	5 (83.3%)	8 (80.0%)		
Vegetable Consumption				
Rarely <1/week	0 (0%)	1 (10.0%)	n/a	1
1-7 times/week	6 (100%)	9 (90.0%)		
Eggs Consumption				
Rarely <1/week	1 (16.7%)	2 (20.0%)	n/a	1
1-7 times/week	5 (83.3%)	8 (80.0%)		
Beef Consumption				
Rarely <1/week	3 (50.0%)	1 (10.0%)	n/a	0.23
1-7 times/week	3 (50.0%)	9 (90.0%)		
Chicken Consumption				
Rarely <1/week	1 (16.7%)	0 (0%)	n/a	0.79
1-7 times/week	5 (83.3%)	10 (100%)		
Lamb Consumption				
Never	5 (83.3%)	7 (70.0%)	n/a	1
Rarely <1/week	1 (16.7%)	3 (30.0%)		
Fish Consumption				
Never	0 (0%)	1 (10.0%)	n/a	0.59
Rarely <1/week	3 (50.0%)	3 (30.0%)		
1-7 times/week	3 (50.0%)	6 (60.0%)		
Occupation				
field worker	6 (100%)	0 (0%)	n/a	-
dairy worker	0 (0%)	10 (100%)		

**two-sample t-test *p*-values < 0.001 for HMP compared to HDW

[a] $\chi^2 < 0.001$ for HMP compared to HDW

Trinh *et al. BMC Genomics* (2024) 25:708

Page 8 of 18

Table 2 Summary statistics related to sequencing data for the dairy worker (DW) and community controls (CC) from the Healthy Dairy Worker (HDW) cohort and participants from the Human Microbiome Project (HMP). $M = \times 10^6$, $k = \times 10^3$

	Community (N=6)	Dairy (N=10)	HMP (N=47)	*p*-value (DW vs. CC)
No. of pairs sequenced				
Mean (SD)	27M (3.1M)	22.8M (4.3M)	n/a	0.04
Total pairs post-filtering				
Mean (SD)	24.2M (2.9M)	20.1M (4.0M)	n/a	0.03
Total pairs post-host removal				
Mean (SD)	24.2M (2.9M)	20.1M (4.0M)	111M (26.4M)**	0.03
No. of contigs				
Mean (SD)	58.6k (8k)	43.4k (16k)	105k (49k)**	0.02
No. genes (Prodigal)				
Mean (SD)	240k (26k)	181k (59k)	221k (95k)‡	0.02

**two sample t-test $p < 0.001$ for HMP compared to HDW

‡two sample t-test $p = 0.37$ for HMP compared to HDW

Clostridium sp. CAG:167 ($\hat{\beta} = -4.21$, $q = 0.19$), and *Prevotella* sp. CAG:873 ($\hat{\beta} = -8.10$, $q = 0.19$) in community controls. We additionally conducted differential abundance testing comparing dairy workers and non-dairy workers with the inclusion of metagenomic data from 47 HMP male subjects with no evidence of disease. A comparison with and without the inclusion of HMP cohort showed similar effect sizes (Pearson $\hat{\rho} = 0.71$), but some variation in *p*-values (Pearson $\hat{\rho} = 0.55$) (Supplementary Figure S1).

We further investigated differences in the community structures of dairy worker and community control metagenomes by examining differences in $\alpha-$ and $\beta-$ diversities. A comparison of the species-level $\alpha-$ diversity using Shannon diversity showed no significant difference in the $\alpha-$diversity of dairy worker metagenomes compared to community control metagenomes ($\hat{\alpha}_{DW} - \hat{\alpha}_{CC} = -0.19$, betta $p = 0.24$). Similarly, a comparison of differences in the community composition ($\beta-$diversity) of dairy worker and community control metagenomes using the Bray-Curtis dissimilarity metric showed no evidence of differences in the true group centroids at the 5% level (DivNet $p = 0.40$) (Supplementary Figure S2). We additionally analyzed differences in gene-level richness (the number of unique genes) between dairy worker and community control metagenomes using geneshot [79, 80] and breakaway [78] to estimate the gene-level richness of each sample, finding significantly lower gene-level richness in dairy worker metagenomes compared to community control metagenomes ($\hat{C}_{DW} - \hat{C}_{CC} \approx -2.0 \times 10^5$, breakaway $p = 0.003$, [77]). To contextualize this finding, we also estimated the species richness in each metagenome using single-copy core genes [43], finding that on average there were 55 fewer species in dairy worker metagenomes compared to community control metagenomes, but that this difference was not statistically significant at the 5% level (breakaway $p = 0.31$).

Identification of virulence factor genes

Through mass screening of contigs across our 16 metagenomes using the Virulence Factor Database (VFDB) [58], we identified 37 different virulence factor genes across 4 samples (3 community controls and 1 dairy worker; Supplementary Table S4). We found that samples with the highest number of identified Virulence Factor Database (VFDB) genes were also those with higher sequencing depth (Fig. 2, right). On average, community control samples had higher number of virulence factor genes identified than dairy workers (mean = 9.2, sd 10.1 vs. mean = 0.3, sd 0.9, p = 0.08). Using happi [82], which accounts for unequal sequencing effort, we tested for differential enrichment of virulence factor genes between dairy worker and community control metagenomes, adjusting for age. No virulence factor genes were significantly enriched between dairy worker and community control metagenomes at the 5% FDR level (Supplementary Table S5). We note that 3 community control metagenomes had higher numbers of identified virulence factor genes compared to samples of similar sequencing depth.

For comparison, we also considered the number of virulence factor genes identified in the HMP HHS cohort.

Trinh *et al. BMC Genomics* (2024) 25:708

Page 9 of 18
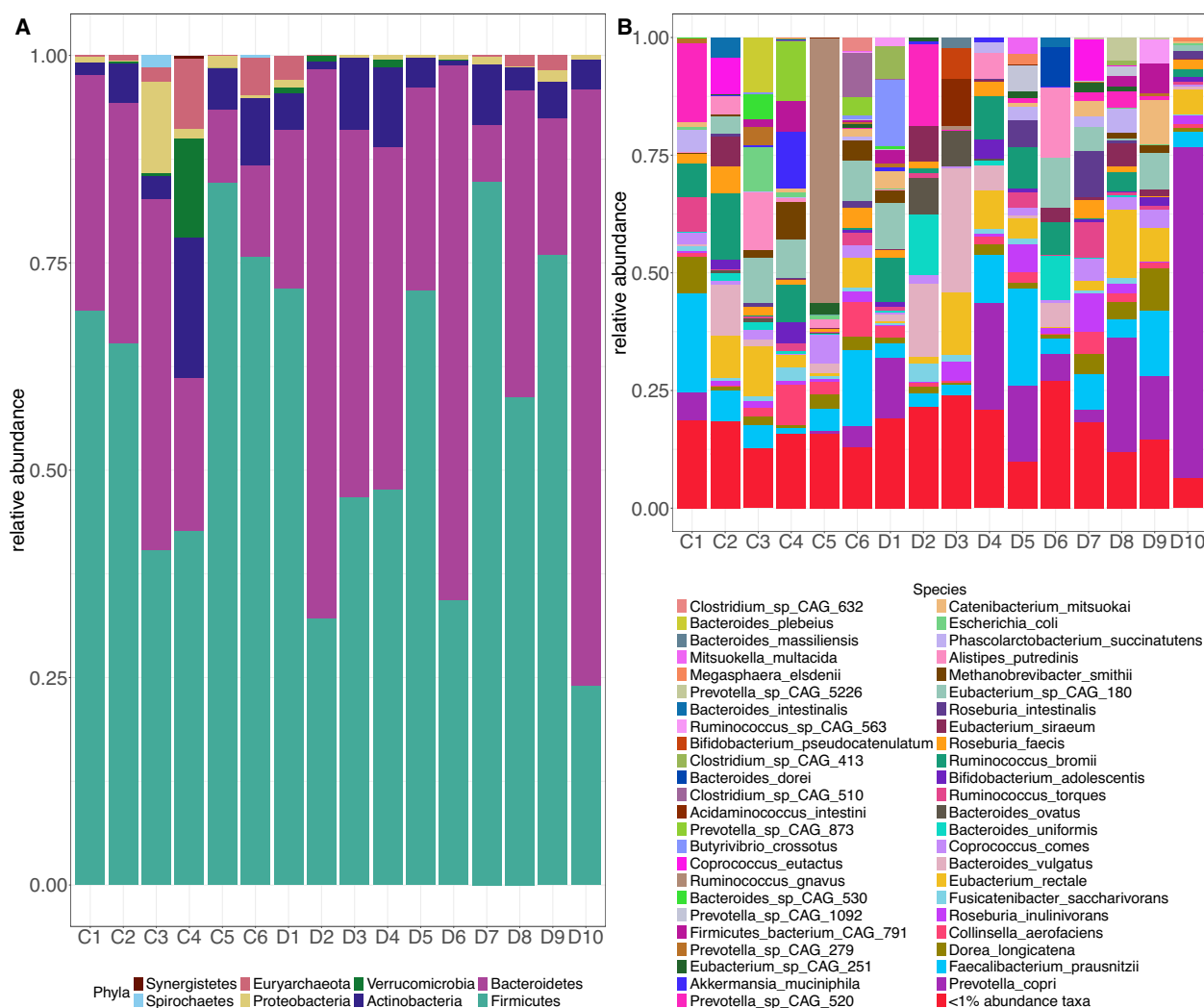
Fig. 1 Stacked barplots of relative abundances show the most abundant phyla (**A**) and species (**B**) within each metagenome. At the phylum-level (**A**), Firmicutes, Bacteroidetes, and Actinobacteria are the most abundant phyla across all samples. At the species-level (**B**), the 5 most abundant and prevalent species across community control and dairy worker metagenomes were *F. prausnitzii*, *E. rectale*, *P. copri*, and *Eubacterium* sp. CAG-180. Species with relative abundances less than 1% were grouped together. There was insufficient evidence to suggest major differences in the taxonomic composition of dairy worker metagenomes compared to community controls

When comparing the male subjects from the HMP cohort to our all-male dairy worker cohort, we found that the HMP males had a range of 4-36 virulence factor genes, which was higher than the range of 0-19 virulence factor genes found in our study metagenomes (Fig. 2A). Testing for differential enrichment of virulence factor genes between dairy workers and non dairy workers, adjusting for HMP cohort membership, age and BMI and accounting for sequencing depth did not show evidence of significantly enriched virulence factors genes between dairy workers and non dairy workers at the 5% FDR level (Supplementary Table S6).

Identification and taxonomic associations of antimicrobial resistance genes (ARGs)

Screening of the 16 metagenomes using the Comprehensive Antibiotic Resistance Database (CARD) identified 85 distinct ARGs across the 16 metagenomes, conferring resistance to at least 17 different antibiotic classes (Supplementary Figure S3; Supplementary Table S7). On average, a higher number of ARGs were identified in community control metagenomes compared to dairy worker metagenomes (mean = 26.5, sd 20.5 vs. mean = 8.5, sd 3.7, p = 0.09) (Fig. 2B). However, differences in the number of ARGs identified may

Trinh *et al. BMC Genomics* (2024) 25:708

Page 10 of 18
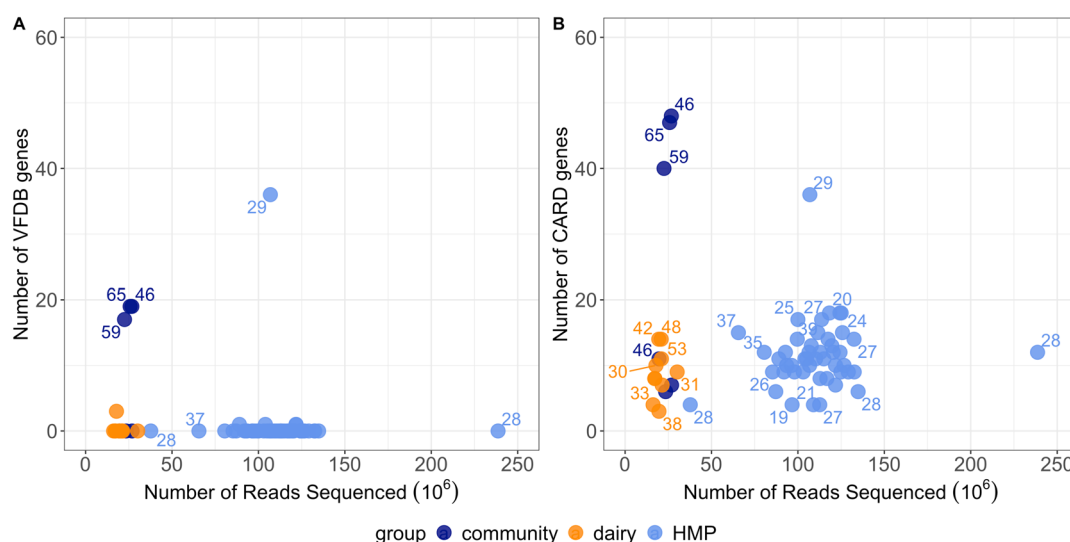
Fig. 2 For each metagenome, we compare the sequencing depth with the number of identified (**A**) VFDB genes and (**B**) CARD genes. Ages (years) of each subject have been labeled. Samples with deeper sequencing had higher numbers of identified genes from the CARD and VFDB databases and higher numbers of estimated genomes. Within the community control group, 3 samples had the highest number of identified CARD genes out of all samples studied, whereas the remaining 3 community control samples within the community control group appeared to be indistinguishable from dairy workers in the number of identified CARD genes. The number of CARD and VFDB genes identified in our study cohort appeared to be similar in range to the number of CARD and VFDB genes identified in the HMP HHS cohort despite higher sequencing depths on average per sample in the HMP study cohort

be due, in part, to differences in sequencing depth, as metagenome samples with the highest number of ARGs identified also had higher numbers of sequenced reads (Fig. 2B). We therefore used happi to test for differences in the enrichment of ARGs between dairy worker and community control metagenomes, while accounting for differences in sequencing depth and adjusting for age. No ARGs were differentially enriched at 5% FDR, but the following ARGs had the largest magnitude test statistics: *sat4* (happi LRT $\chi^2 = 0.001, q = 0.10$) a plasmid-mediated streptothricin acetyltransferase and streptothricin resistant determinant, *tet*(W) (happi LRT $\chi^2 = 0.05, q = 1$) a tetracycline resistance gene associated with both conjugative and non-conjugative DNA, and *emrB* (happi LRT $\chi^2 = 0.14, q = 1$) a translocase gene in the emrB-TolC efflux protein in *Escherichia coli* (Supplementary Table S8). The 3 community metagenomes that we found to have higher numbers of virulence factor genes identified in their metagenomes also had higher numbers of ARGs identified compared to other metagenomes of similar sequencing depth. To contextualize our study cohort, we compared the number of ARGs found in our study metagenomes with the number of ARGs identified in the HMP HHS. Overall, the range of ARGs identified in our study cohort (3-48 ARGs) was similar to the range of ARGs identified in the HMP cohort (4-36 ARGs) (Fig. 2B). Testing for differential enrichment of ARGs between dairy workers and non dairy workers while adjusting for HMP study

membership, age and BMI and accounting for sequencing depth did not show evidence of significantly enriched ARGs between dairy workers compared to non dairy workers at the 5% FDR level (Supplementary Table S9).

We further focused our analyses to ARGs conferring resistance to antibiotic classes considered critically important to human medicine by the World Health Organization (WHO) [60]. Across our study metagenomes, we identified 37 different ARGs conferring resistance to 8 antibiotic classes described in the WHO's list of Critically Important Antimicrobials (CIA): aminoglycosides, fluoroquinolones, macrolides, tetracyclines, cephalosporins, cephamycins, glycopeptides, and sulfonamides (Fig. 3). The most frequently occurring types of antibiotic resistance genes found across the 16 metagenomes were genes that typically confer resistance to tetracyclines ($n = 15$), aminoglycosides ($n = 14$), cephamycins ($n = 13$), and macrolides ($n = 12$) (Fig. 3). Genes that commonly confer tetracycline resistance appeared to dominate the resistomes of both dairy workers and community controls with 11 distinct tetracycline resistance genes identified across 15 of our study metagenomes. We compared relative abundances of genes aggregated by antibiotic class between both groups and found that, after adjusting for age, dairy workers' metagenomes had higher centered logratio transformed relative abundances of tetracycline ($\hat{\beta}_{DW} = 2.7$, $q = 1$), cephamycin ($\hat{\beta}_{DW} = 5.4$, $q = 1$),
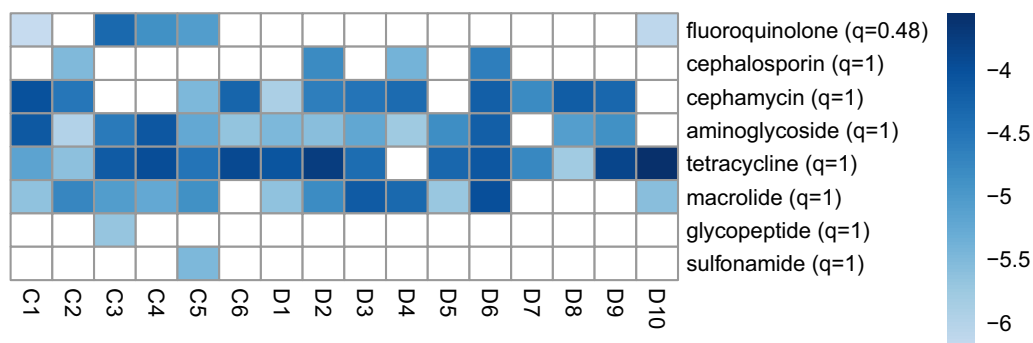
Fig. 3 We identified ARGs from 8 antibiotic classes (rows) listed as critically important to human medicine by the WHO. $\log_{10}$ transformed relative abundances of antibiotic resistance genes grouped by these antibiotic classes are colored from lower (light blue) to higher (darker blue) relative abundances in each metagenome. Antibiotic resistance classes (rows) have been ordered by ascending q-values. White squares denote undetected antibiotic resistance genes. Visual inspection displays patterns of increased abundance of tetracycline resistance genes and macrolide resistance genes in dairy worker metagenomes. Additionally, cephamycin resistance genes had a higher occurrence in dairy workers as these genes were identified in 90% of dairy worker samples compared to 67% of community control samples

macrolide ($\hat{\beta}_{DW} = 3.9$, $q = 1$), sulfonamide ($\hat{\beta}_{DW} = 1.6$, $q = 1$), and cephalosporin ($\hat{\beta}_{DW} = 4.1$, $q = 1$) resistance genes than community controls' metagenomes (Fig. 3; Supplementary Table S10); however, these differences were not significant at the 5% FDR. Similarly, the lower CLR-transformed relative abundance of fluoroquinolone ($\hat{\beta}_{DW} = -19.2$, $q = 0.5$), aminoglycoside ($\hat{\beta}_{DW} = -3.11$, $q = 1$), and glycopeptide ($\hat{\beta}_{DW} = -5.1$, $q = 1$) resistance genes in dairy workers' metagenomes compared to community controls' metagenomes was also not significant at an FDR of 5%.

To understand whether there were differences in taxonomic affiliation of ARGs between groups, we assessed the taxonomic context of tetracycline and beta-lactam resistance genes. We identified 6 different genes (*cblA-1*, *cfxA2*, *cfxA3*, *cfxA4*, *cfxA5*, *cfxA6*) that encode for beta-lactamases and confer resistance to beta-lactam antibiotics. Additional details on the presence of each beta-lactam resistance gene in each of our study metagenomes are found in Supplementary Table S7. These 6 beta-lactam genes have typically been identified on the chromosomes of *Bacteroides* spp. [66]. Taxonomic annotation of the genomic context of these genes in dairy worker and community control metagenomes confirmed their association with organisms from the phylum Bacteroidetes such as *Prevotella copri*, *Bacteroides fragilis*, and *Bacteroides uniformis*. Additionally, we observed no differences in taxonomic affiliation of these beta-lactam genes between dairy workers and community controls (Supplementary Table S11).

We identified 9 *tet* genes (efflux genes: *tet*(B), *tet*(G), *tet*(40); and ribosomal genes: *tet*(M), *tet*(O), *tet*(Q), *tet*(W), *tet*(W/N/W), and *tet*(32)) that encode for efflux pumps or ribosomal protection proteins conferring

resistance to tetracycline antibiotics. These genes have normally been associated with plasmids [66], which are small, extra-chromosomal DNA molecules that facilitate genetic sharing between and within species [88], but can also be found in chromosomes. Taxonomic annotation of the assembly graphs for these tetracycline resistance genes demonstrated affiliation of these genes with a variety of both commensal (e.g., *Lawsonia intracellularis*, *Ligilactobacillus animalis*, *Trueperella pyogenes*, *Schaalia turicensis*, and *Faecalibacterium prausnitzii*) and pathogenic (e.g., *Campylobacter* spp., *Clostridium* spp.) bacteria. Full annotations of these ARGs to affiliated bacterial organisms can be found in Supplementary Table S11. Finally, while these tetracycline resistance genes were affiliated with a wide range of commensal and pathogenic bacteria, we found no differences in the taxonomic context of tetracycline resistance genes identified in community controls compared to dairy workers.

Discussion

Using shotgun metagenomics sequencing, we investigated differences in taxonomy, diversity, and the presence of genes (especially ARGs and virulence factors) between dairy workers and community controls' gut microbiomes. The use of shotgun metagenomics data allowed us to circumvent some of the limitations of amplicon sequencing, and enabled us to investigate abundance and presence of a variety of genes as well as their taxonomic context. While the results of our investigation revealed no statistically significant differences at the 5% FDR level in the taxonomic composition, antibiotic resistance and virulence factor gene carriage, and relative abundances of ARGs, we observed several patterns for further investigation including greater abundance of tetracycline resistance

Trinh *et al. BMC Genomics* (2024) 25:708

Page 12 of 18

genes and higher occurrence of cephamycin resistance genes in dairy workers' metagenomes; evidence of commensal organism association with tetracycline resistance genes; and lower gene richness and genome diversity in dairy workers' metagenomes.

Previous metagenomic studies of livestock workers in China and Europe have found increased abundance and carriage of antibiotic resistance genes in individuals occupationally exposed to animal farming environments, raising concerns that these environments could be hotspots for antibiotic resistance and zoonotic disease emergence [27–29, 89, 90]. Cross sectional studies of pig farmers and slaughterhouse workers in the Netherlands ($n_{\text{workers}} = 70$, $n_{\text{controls}} = 46$) [28] and China ($n_{\text{workers}} = 4, n_{\text{controls}} = 5$) [90] found that the resistomes of these animal workers were dominated by tetracyclines, aminoglycosides, beta-lactam and macrolide resistance genes. Another cross-sectional study of live poultry market workers in China found higher abundance of ARGs, lower Shannon diversity, and greater enrichment of beta-lactam and lincosamide resistance genes in these workers compared to controls ($n_{\text{workers}} = 18$, $n_{\text{controls}} = 18$) [89], and a longitudinal study of veterinary students with exposure to swine farms observed similar patterns of increased total abundance of ARGs and increased abundances of beta-lactam, aminoglycoside, and tetracycline resistance genes within 3 months $n = 14$ [27].

Contrary to these previous studies of livestock workers, we found no significant difference in the abundance of ARGs between dairy workers and community controls, though we did observe patterns of greater abundance of tetracycline resistant genes in dairy workers' metagenomes that was directionally consistent with findings in these previous farm studies [27–29, 89]. In addition, we found more frequent occurrences of cephamycin (beta-lactam) resistant genes identified in the dairy worker population compared to community controls. These patterns are interesting to highlight since tetracyclines are commonly administered on dairy farms for treating gastrointestinal and respiratory diseases in dairy cows [91] and beta-lactam antibiotics such as ceftiofur are frequently used to treat metritis, a common post-partum uterine inflammatory disease [92]. It is also worth noting that the patterns observed in our study reflect the potential impacts of occupational exposure to livestock farming without the use of antibiotics for growth promotion, as the samples used in this metagenomics study were collected at least one year after the full implementation of the FDA's GFI no. 213 policy banning the use of antibiotics for growth promotion purposes.

Our study also highlighted the potential for commensal organisms to serve as ARG reservoirs for pathogenic bacteria. By reconstructing the genomic context of each antibiotic resistance gene then taxonomically annotating this context, we were able to confirm the association of chromosome-mediated ARGs (e.g., *cblA*-1, *cfx*A2, *cfx*A3, *cfx*A4, *cfx*A5, *cfx*A6) with previously recognized carriers of these genes (e.g., *Bacteroides* spp.) [66]. With the same approach applied to primarily plasmid-mediated ARGs (e.g., *tet*(B), *tet*(G), *tet*(W/N/W), *tet*(32), *tet*(M), *tet*(O), *tet*(Q), and *tet*(W)), we found that resistance genes were associated with both commensal and pathogenic organisms. These observations suggest the potential for sharing of ARGs between commensal organisms and pathogens through conjugation. Furthermore, our results corroborate findings from a recent study that compared ARGs identified in 1,354 culture commensal strains and 45,403 pathogen strains from the human gut and found evidence of 64,188 shared ARGs that mapped to 5,931 mobile genetic elements [93]. Some of the mobile genetic elements identified [93] had also been previously identified in data from ruminant guts, soil, and other human body sites [93]. While commensal organisms may serve as ARG reservoirs for pathogenic bacteria, they may also assist in preventing pathogenic invasion through indirect (enhancement of host immune defenses) and direct (competition of nutrients and niche) mechanisms [8–11]. Further research is needed to better understand the complex dynamic that commensal organisms balance in promoting both pathogen resistance and antibiotic resistance emergence.

Our results also demonstrated evidence of lower average gene richness (and some evidence of lower genome diversity) in dairy workers, even after accounting for differential sequencing depth. Lower gene richness has been associated with increased intestinal inflammation and metabolic disorders [79, 94, 95]. A common occupational hazard facing dairy workers is inhalation of dusts and aerosols containing endotoxins or other proinflammatory substances that can result in airway inflammation and decreased pulmonary function [96–98]. Several studies have proposed a gut-lung axis linking pulmonary inflammation to intestinal inflammation based on epidemiological and clinical observations of the co-occurrence of these diseases [99–101]. There is therefore the possibility that the lower gene richness observed in dairy workers points towards increased intestinal inflammation linked to possible increased airway inflammation from exposure to aerosols and endotoxins. Further investigation to explore the possibility of increased intestinal and airway inflammation of this cohort is warranted.

Our study had several limitations. The most significant limitation was its small sample size, and therefore relatively low power to reject false null hypotheses. Corroborating our findings, especially those regarding patterns of greater tetracycline and cephamycin resistance gene in

Trinh *et al. BMC Genomics* (2024) 25:708

Page 13 of 18

our dairy cohort, with a larger sample size is desirable. Another major limitation of our study was the comparability of the community controls with the dairy workers. The community controls in our metagenomic study occupationally identified as field workers in non-animal agriculture industries, and agricultural and dairy workers both experience occupational exposure to animal manure (e.g., as fertilizer) and antibiotics (e.g., streptomycin and oxytetracycline are commonly sprayed to control fire blight disease in Eastern Washington [102, 103]). Similar occupational exposures in the dairy workers and controls may reduce the effect sizes of group differences compared to comparisons of dairy workers and non-agricultural workers (e.g., office workers).

That our cohort is comprised of exclusively Hispanic or Latino males limits the generalizability of our findings. The American Community Survey, conducted by the US Department of Agriculture in 2021 [104], indicates that livestock workers across the US are 22% female and only 48% non-Hispanic Whites. The results observed from our study therefore may not be broadly generalizable to livestock worker cohorts.

We additionally note the higher average age of community controls in our cohort compared to dairy workers, with 3 community controls having both higher ages and the highest number of identified ARGs and virulence factors. Antibiotic resistance genes have been previously shown to have an age-related cumulative effect, with older age groups harboring higher abundances of ARGs [105]. To mitigate this, our gene enrichment and gene abundance analyses adjusted for age as a potential confounder (note also that all our analyses adjust for unequal sequencing depth across samples). Our analyses didn't reveal any statistically significant differences in ARG presence between groups of the same age who differ in their dairy worker status. A secondary analysis that supplemented our cohort with the Human Microbiome Project cohort (to contextualize the dairy workers with an alternative control group) also did not suggest that there are differences in ARG presence between groups of the same age, BMI and study cohort who differ in their dairy worker status. One limitation of this secondary analysis is that the HMP data was not processed alongside our study's samples, and technical artifacts inducing falsely significant results are a substantial concern when pooling data across studies in general. A comparison of the extraction, sequencing and data processing protocols between the HMP data and our study cohort show notable differences in sequencing platforms (HMP used an Illumina GAIIx platform with 101bp paired-end reads compared to our study's use of an Illumina HiSeq 2500 with 150bp paired-end reads), quality control and filtering methods (HMP removed duplicate and low-quality reads while also filtering out outliers by mean contig and ORF density, human hits, rRNA hits and size whereas our study employed a Q-score based algorithm to identify and filter out low quality reads), host removal tools used (HMP used BMTagger while our study used bowtie2), and *de novo* assemblers used (HMP used IDBA-UD compared to our study's use of MEGAHIT). However, as our ARG enrichment results were null (both within our study and when supplemented with HMP data), technical artifacts do not appear to be driving significantly different associations. Additionally, adjusting for study using an indicator for the participant being from the HMP cohort in our analyses addresses the potential differential detection of microbial taxa and genes across the two studies that might arise due to differences in extraction, library preparation protocols, and bioinformatics processing.

Another limitation of this secondary analysis using the HMP data is the potential misclassification of occupation in the HMP participants. In the absence of publicly available data on occupation, we relied on previous studies and limited published information on the distribution of job types across the HMP cohort. The HMP's 2012 phase one study participants of HHS had an over-representation of students with job categories reported in research, healthcare, education, sales, clerical/administrative, management, and an "other" category [72]. The "other" category made up no more than 15% of the responses collected. Additionally, recruitment of individuals from the HMP was conducted in two urban areas [67]. As many dairy workers live isolated in rural areas, having poor access to public transportation and many lacking personal vehicles [106], it is unlikely that any of the 47 HMP participants used in our study had occupational exposure to livestock farming. We further note that there is only limited publicly available demographic data on the HMP participants, and therefore we could only adjust for age and BMI as potential confounders. Thus, while there is a possibility of unmeasured confounding in our study, we can rule out the risk of confounding bias as the cause of significant associations between ARG presence and dairy work.

Finally, while cross-sectional studies can be advantageous for conducting cost-effective comparisons of populations, they only capture differences between groups at a single time point. Therefore, our study cannot provide information about long-term changes to the microbiome that are induced by occupational exposure to livestock farming. We also note that our choice to use shotgun metagenomics to studying antibiotic resistance limits our conclusions to genotypic potential for resistance, and not phenotypic resistance. Complementary future work includes pairing shotgun sequencing with phenotypic resistance profiles (e.g., using culture-based approaches).

Trinh *et al. BMC Genomics (2024) 25:708*

Page 14 of 18

Conclusions

In this study of occupational exposure to dairy farming, we observed no significant differences in antibiotic resistant gene or virulence factor presence in dairy workers compared to controls, but several patterns warranting further investigation, including greater abundance of tetracycline resistance genes and higher occurrence of cephamycin resistance genes in dairy workers' metagenomes; evidence of commensal organism association with tetracycline resistance genes; and lower average gene richness in dairy workers' metagenomes. This work demonstrates the depth and scope of utilizing shotgun metagenomics to investigate microbiomes and resistomes, and provides a foundation for further investigations into the impact of exposure to zoonotic pathogens, antibiotic resistant organisms, and ARGs on the microbiomes and resistomes of livestock workers.

Abbreviations

ARG(s)	Antibiotic Resistance Gene(s)
WHO	World Health Organization
HDW	Healthy Dairy Worker Study
VF	Virulence Factors
VFDB	Virulence Factor Database
CARD	Comprehensive Antibiotic Resistance Database
CIA	Critically Important Antimicrobials
LRT	Likelihood Ratio Test
CLR	Centered Log-Ratio
SCG(s)	Single Copy Gene(s)
DNA	Deoxyribonucleic acid
sd	standard deviation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10562-1.

> Supplementary Material 1.
>
> Supplementary Material 2.

Availability of data and materials

The data supporting the conclusions of this article along with code for reproducing our results are made available at https://github.com/statdivlab/hdw_mgx_supplementary. The sequencing data has been made available on the Sequence Read Archive under PRJNA971196.

Declarations

Author details

[1]Department of Biostatistics, University of Washington, Seattle, USA. [2]Department of Statistics, University of Washington, Seattle, USA. [3]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, USA.

References

1. Artis D. Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. Nat Rev Immunol. 2008;8(6):411–20. https://doi.org/10.1038/nri2316
2. Ivanov II, Littman D.R. Modulation of immune homeostasis by commensal bacteria. Curr Opin Microbiol. 2011;14(1):106–14 . https://doi.org/10.1016/j.mib.2010.12.003
3. Arrieta M-C, Finlay B. The commensal microbiota drives immune homeostasis. Front Immunol. 2012;3. https://doi.org/10.3389/fimmu.2012.00033
4. Geng J, Ni Q, Sun W, Li L, Feng X. The links between gut microbiota and obesity and obesity related diseases. Biomed Pharmacother. 2022;147:112678. https://doi.org/10.1016/j.biopha.2022.112678
5. Yang J-Y, Lee Y-S, Kim Y, Lee S-H, Ryu S, Fukuda S, Hase K, Yang C-S, Lim HS, Kim M-S, Kim H-M, Ahn S-H, Kwon B-E, Ko H-J, Kweon M-N. Gut commensal bacteroides acidifaciens prevents obesity and improves insulin sensitivity in mice. Mucosal Immunol. 2017;10(1):104–16. https://doi.org/10.1038/mi.2016.42
6. Nakanishi Y, Sato T, Ohteki T. Commensal gram-positive bacteria initiates colitis by inducing monocyte/macrophage mobilization. Mucosal Immunol. 2015;8(1):152–60. https://doi.org/10.1038/mi.2014.53
7. Cunningham AL, Stephens JW, Harris D.A. Gut microbiota influence in type 2 diabetes mellitus (t2dm). Gut Pathog. 2021;13(1):50. https://doi.org/10.1186/s13099-021-00446-0
8. Abt MC, Pamer E.G. Commensal bacteria mediated defenses against pathogens. Curr Opin Immunol. 2014;29:16–22. https://doi.org/10.1016/j.coi.2014.03.003
9. Khan R, Petersen FC, Shekhar S. Commensal bacteria: An emerging player in defense against respiratory pathogens. Front Immunol. 2019;10. https://doi.org/10.3389/fimmu.2019.01203
10. Buffie CG, Pamer E.G. Microbiota-mediated colonization resistance against intestinal pathogens. Nat Rev Immunol. 2013;13(11):790–801. https://doi.org/10.1038/nri3535
11. Kamada N, Seo S-U, Chen GY, Núñez G. Role of the gut microbiota in immunity and inflammatory disease. Nat Rev Immunol. 2013;13(5):321–35. https://doi.org/10.1038/nri3430
12. Xu F, Fu Y, Sun T-y, Jiang Z, Miao Z, Shuai M, Gou W, Ling C-w, Yang J, Wang J, Chen Y-m, Zheng J-S. The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features

Trinh *et al. BMC Genomics* (2024) 25:708

Page 15 of 18

for complex human diseases. Microbiome. 2020;8(1):145. https://doi.org/10.1186/s40168-020-00923-9

13. Qin Y, Havulinna AS, Liu Y, Jousilahti P, Ritchie SC, Tokolyi A, Sanders JG, Valsta L, Brożyńska M, Zhu Q, Tripathi A, Vázquez-Baeza Y, Loomba R, Cheng S, Jain M, Niiranen T, Lahti L, Knight R, Salomaa V, Inouye M, Méric G. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. Nat Genet. 2022;54(2):134–42. https://doi.org/10.1038/s41588-021-00991-z

14. Liu H, Han M, Li SC, Tan G, Sun S, Hu Z, Yang P, Wang R, Liu Y, Chen F, Peng J, Peng H, Song H, Xia Y, Chu L, Zhou Q, Guan F, Wu J, Bu D, Ning K. Resilience of human gut microbial communities for the long stay with multiple dietary shifts. Gut. 2019;68(12):2254–5. https://doi.org/10.1136/gutjnl-2018-317298

15. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh P.J. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505(7484):559–63. https://doi.org/10.1038/nature12820

16. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon J.I. Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222–7. https://doi.org/10.1038/nature11053

17. Blaser M.J. Antibiotic use and its consequences for the normal microbiome. Sci (NY). 2016;352(6285):544–5. https://doi.org/10.1126/science.aad9358

18. Maier L, Pruteanu M, Kuhn M, Zeller G, Telzerow A, Anderson EE, Brochado AR, Fernandez KC, Dose H, Mori H, Patil KR, Bork P, Typas A. Extensive impact of non-antibiotic drugs on human gut bacteria. Nature. 2018;555(7698):623–8. https://doi.org/10.1038/nature25979

19. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, Costea PI, Godneva A, Kalka IN, Bar N, Shilo S, Lador D, Vila AV, Zmora N, Pevsner-Fischer M, Israeli D, Kosower N, Malka G, Wolf BC, Avnit-Sagi T, Lotan-Pompan M, Weinberger A, Halpern Z, Carmi S, Fu J, Wijmenga C, Zhernakova A, Elinav E, Segal E. Environment dominates over host genetics in shaping human gut microbiota. Nature. 2018;555(7695):210–5. https://doi.org/10.1038/nature25973

20. Gacesa R, Kurilshikov A, Vich Vila A, Sinha T, Klaassen MAY, Bolte LA, Andreu-Sánchez S, Chen L, Collij V, Hu S, Dekens JAM, Lenters VC, Björk JR, Swarte JC, Swertz MA, Jansen BH, Gelderloos-Arends J, Jankipersadsing S, Hofker M, Vermeulen RCH, Sanna S, Harmsen HJM, Wijmenga C, Fu J, Zhernakova A, Weersma R.K. Environmental factors shaping the gut microbiome in a dutch population. Nature. 2022;604(7907):732–9. https://doi.org/10.1038/s41586-022-04567-7

21. Gao F-Z, He L-Y, Chen X, Chen J-L, Yi X, He L-X, Huang X-Y, Chen Z-Y, Bai H, Zhang M, Liu Y-S, Ying G-G. Swine farm groundwater is a hidden hotspot for antibiotic-resistant pathogenic acinetobacter. ISME Commun. 2023;3(1):34. https://doi.org/10.1038/s43705-023-00240-w

22. Xin H, Gao M, Wang X, Qiu T, Guo Y, Zhang L. Animal farms are hot spots for airborne antimicrobial resistance. Sci Total Environ. 2022;851:158050. https://doi.org/10.1016/j.scitotenv.2022.158050

23. Jones BA, Grace D, Kock R, Alonso S, Rushton J, Said MY, McKeever D, Mutua F, Young J, McDermott J, Pfeiffer D.U. Zoonosis emergence linked to agricultural intensification and environmental change. Proc Natl Acad Sci U S A. 2013;110(21):8399–404. https://doi.org/10.1073/pnas.1208059110

24. Manyi-Loh C, Mamphweli S, Meyer E, Okoh A. Antibiotic use in agriculture and its consequential resistance in environmental sources: Potential public health implications. Molecules. 2018;23(4). https://doi.org/10.3390/molecules23040795

25. Grout L, Baker MG, French N, Hales S. A Review of Potential Public Health Impacts Associated With the Global Dairy Sector. GeoHealth. 2020;4(2):1–46. https://doi.org/10.1029/2019gh000213

26. Marshall BM, Levy S.B. Food animals and antimicrobials: Impacts on human health. Clin Microbiol Rev. 2011;24(4):718–33. https://doi.org/10.1128/CMR.00002-11

27. Sun J, Liao X-P, D'Souza AW, Boolchandani M, Li S-H, Cheng K, Luis Martínez J, Li L, Feng Y-J, Fang L-X, Huang T, Xia J, Yu Y, Zhou Y-F, Sun Y-X, Deng X-B, Zeng Z-L, Jiang H-X, Fang B-H, Tang Y-Z, Lian X-L, Zhang R-M, Fang Z-W, Yan Q-L, Dantas G, Liu Y-H. Environmental remodeling of

human gut microbiota and antibiotic resistome in livestock farms. Nat Commun. 2020;11(1):1427. https://doi.org/10.1038/s41467-020-15222-y

28. Van Gompel L, Luiken REC, Hansen RB, Munk P, Bouwknegt M, Heres L, Greve GD, Scherpenisse P, Jongerius-Gortemaker BGM, Tersteeg-Zijderveld MHG, García-Cobos S, Dohmen W, Dorado-García A, Wagenaar JA, Urlings BAP, Aarestrup FM, Mevius DJ, Heederik DJJ, Schmitt H, Bossers A, Smit L.A.M. Description and determinants of the faecal resistome and microbiome of farmers and slaughterhouse workers: A metagenome-wide cross-sectional study. Environ Int. 2020;143:105939. https://doi.org/10.1016/j.envint.2020.105939

29. Duarte ASR, Röder T, Van Gompel L, Petersen TN, Hansen RB, Hansen IM, Bossers A, Aarestrup FM, Wagenaar JA, Hald T. Metagenomics-based approach to source-attribution of antimicrobial resistance determinants – identification of reservoir resistome signatures. Front Microbiol. 2021;11. https://doi.org/10.3389/fmicb.2020.601407

30. Niu C, Yu D, Wang Y, Ren H, Jin Y, Zhou W, Li B, Cheng Y, Yue J, Gao Z, Liang L. Common and pathogen-specific virulence factors are different in function and structure. Virulence. 2013;4(6):473–82. https://doi.org/10.4161/viru.25730. PMID: 23863604

31. Holden M, Crossman L, Cerdeño-Tárraga A, Parkhill J. Pathogenomics of non-pathogens. Nat Rev Microbiol. 2004;2(2):91. https://doi.org/10.1038/nrmicro825

32. Pallen MJ, Wren B.W. Bacterial pathogenomics. Nature. 2007;449(7164):835–42. https://doi.org/10.1038/nature06248

33. Kim H, Kim M, Kim S, Lee YM, Shin S.C. Characterization of antimicrobial resistance genes and virulence factor genes in an arctic permafrost region revealed by metagenomics. Environ Pollut. 2022;294:118634. https://doi.org/10.1016/j.envpol.2021.118634

34. Pan Y, Zeng J, Li L, Yang J, Tang Z, Xiong W, Li Y, Chen S, Zeng Z, Gilbert J.A. Coexistence of antibiotic resistance genes and virulence factors deciphered by large-scale complete genome analysis. mSystems. 2020;5(3):00821–19. https://doi.org/10.1128/mSystems.00821-19

35. Carmona J, deMarcken M, Trinh P, Frisbie L, Ramirez V, Palmandez P, Vedal S, Sack C, Rabinowitz P. A cross sectional study of respiratory and allergy status in dairy workers. J Agromedicine. 2023;0(0):1–8. https://doi.org/10.1080/1059924X.2023.2171522. PMID: 36704933

36. Food and Drug Administration, et al. New animal drugs and new animal drug combination products administered in or on medicated feed or drinking water of food-producing animals: recommendations for drug sponsors for voluntarily aligning product use conditions with GFI# 209. Guidance for Industry# 213. 2014. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cvm-gfi-213-new-animal-drugs-and-new-animal-drug-combination-products-administered-or-medicated-feed. Accessed 1 Apr 2024.

37. Food and Drug Administration. FACT SHEET: veterinary feed directive final rule and next steps. 2015. https://www.fda.gov/animal-veterinary/development-approval-process/fact-sheet-veterinary-feed-directive-final-rule-and-next-steps. Accessed 1 Apr 2024.

38. Illumina: bcl2fastq Conversion Software. http://support.illumina.com/downloads/bcl2fastq_conversion_software_184.html. Accessed 26 June 2022

39. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. eLife. 2021;10:1–42. https://doi.org/10.7554/eLife.65088

40. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9(8):811–4. https://doi.org/10.1038/nmeth.2066

41. Langmead B, Salzberg S.L. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923

42. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 2018;35(3):421–32. https://doi.org/10.1093/bioinformatics/bty648

43. Eren A.M, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont T.O. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:1319. https://doi.org/10.7717/peerj.1319

Trinh *et al. BMC Genomics* (2024) 25:708

Page 16 of 18

44. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2. https://doi.org/10.1093/bioinformatics/bts480

45. Shaiber A, Willis AD, Delmont TO, Roux S, Chen LX, Schmid AC, Yousef M, Watson AR, Lolans K, Esen ÖC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, Eren A.M. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol. 2020:21–292. https://doi.org/10.1101/2020.04.29.069278

46. Eren AM, Vineis JH, Morrison HG, Sogin M.L. A filtering method to generate high quality short reads using illumina paired-end technology. PLoS ONE. 2013;8(6):1–6. https://doi.org/10.1371/journal.pone.0066643

47. Minoche A, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. Genome Biol. 2011;12(11):112. https://doi.org/10.1186/gb-2011-12-11-r112

48. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674–6. https://doi.org/10.1093/bioinformatics/btv033

49. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser L.J. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119. https://doi.org/10.1186/1471-2105-11-119

50. Kim D, Song L, Breitwieser FP, Salzberg S.L. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26(12):1721–9. https://doi.org/10.1101/gr.210641.116

51. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale D.A. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4(1):41. https://doi.org/10.1186/1471-2105-4-41

52. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Res. 2018;47(D1):427–32. https://doi.org/10.1093/nar/gky995

53. Buchfink B, Xie C, Huson D.H. Fast and sensitive protein alignment using diamond. Nat Methods. 2015;12(1):59–60. https://doi.org/10.1038/nmeth.3176

54. Eddy S.R. Accelerated profile hmm searches. PLoS Comput Biol. 2011;7(10):1–16. https://doi.org/10.1371/journal.pcbi.1002195

55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup .G.P.D.P. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352. https://academic.oup.com/bioinformatics/article-pdf/25/16/2078/531810/btp352.pdf

56. Seemann T. ABRicate: mass screening of contigs for antibiotic resistance genes. 2016. https://github.com/tseemann/abricate. Accessed 1 Apr 2024.

57. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic Local Alignment Search Tool. J Mol Biol. 1990;215(3):403–40. https://doi.org/10.1016/S0022-2836(05)80360-2

58. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. Vfdb: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33(1):325–8. https://doi.org/10.1093/nar/gki008

59. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJV, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright G.D. The Comprehensive Antibiotic Resistance Database. Antimicrob Agents Chemother. 2013;57(7):3348–57. https://doi.org/10.1128/aac.00419-13

60. Scott HM, Acuff G, Bergeron G, Bourassa MW, Gill J, Graham DW, Kahn LH, Morley PS, Salois MJ, Simjee S, Singer RS, Smith TC, Storrs C, Wittum T.E. Critically important antibiotics: criteria and approaches for measuring and reducing their use in food animal agriculture. Ann N Y Acad Sci. 2019;1441(1):8–16. https://doi.org/10.1111/nyas.14058

61. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):884–90. https://doi.org/10.1093/bioinformatics/bty560

62. Bushnell B. BBTools software package. 2014. https://sourceforge.net/projects/bbmap/. Accessed 1 Apr 2024.

63. Bushnell B. Masked version of hG19. https://zenodo.org/record/1208052#.Yq-O-i-B1Yh. Accessed 26 June 2022.

64. Olekhnovich EI, Vasilyev AT, Ulyantsev VI, Kostryukova ES, Tyakht A.V. MetaCherchant: Analyzing genomic context of antibiotic resistance genes in gut microbiota. Bioinformatics. 2018;34(3):434–44. https://doi.org/10.1093/bioinformatics/btx681

65. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol. 2019;20(1):257. https://doi.org/10.1186/s13059-019-1891-0

66. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran H-K, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar GV, McArthur A.G. Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020;48(D1):517–25. https://doi.org/10.1093/nar/gkz935

67. Aagaard K, Petrosino J, Keitel W, Watson M, Katancik J, Garcia N, Patel S, Cutting M, Madden T, Hamilton H, Harris E, Gevers D, Simone G, McInnes P, Versalovic J. The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters. FASEB J Off Publ Fed Am Soc Exp Biol. 2012;27. https://doi.org/10.1096/fj.12-220806

68. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PSG, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Michael Dunne W, Scott Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney LJ, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Kinder Haake S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, King NB, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo C-C, Lozupone CA, Dwayne Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavromatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Sankar P, Fah Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR, Wilczek-Boney K, Wu Y, Wylie KM, Wylie T, Yan: Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14. https://doi.org/10.1038/nature11234

69. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D, Petrosino JF, Abubucker S, Badger JH, Chinwalla AT, Earl AM, FitzGerald MG, Fulton RS, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL,

Trinh *et al. BMC Genomics* (2024) 25:708

Page 17 of 18

Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi VR, Brooks P, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PS, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Michael Dunne W, Scott Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney L, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, Kinder-Haake S, King NB, Knight R, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo C-C, Lozupone CA, Dwayne Lunsford R, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavrommatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Sankar P, Fah Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR, Wilczek-Boney K. A framework for human microbiome research. Nature. 2012;486(7402):215–21. https://doi.org/10.1038/nature11209

70. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, Huttenhower C, Morgan M, Segata N, Waldron L. Accessible, curated metagenomic data through experimenthub. Nat Methods. 2017;14(11):1023–4. https://doi.org/10.1038/nmeth.4468

71. W. DJ, Yoshiki V-B, Daniel M, Zhenjiang X, Elaine W, Rob K. Turning participatory microbiome research into usable data: Lessons from the american gut project. J Microbiol Biol Educ. 2016;17(1):46–50. https://doi.org/10.1128/jmbe.v17i1.1034

72. NIH Human Microbiome Project - Core Microbiome Sampling Protocol A (HMP-A). https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/variable.cgi?study_id=phs000228.v4.p1 &phv=158614. Accessed 20 May 2024

73. Clausen DS, Willis AD. Estimating fold changes from partially observed outcomes with applications in microbial metagenomics. 2024. arXiv:2402.05231

74. Storey JD, Bass AJ, Dabney A, Robinson D. Qvalue: Q-value Estimation for False Discovery Rate Control. R package version 2.26.0. 2021. http://github.com/jdstorey/qvalue. Accessed 1 Apr 2024.

75. McLaren MR, Nearing JT, Willis AD, Lloyd KG, Callahan B.J. Implications of taxonomic bias for microbial differential-abundance analysis. bioRxiv. 2022. https://doi.org/10.1101/2022.08.19.504330

76. Willis AD, Martin B.D. Estimating diversity in networked ecological communities. Biostatistics. 2020;23(1):207–22. https://doi.org/10.1093/biostatistics/kxaa015. https://academic.oup.com/biostatistics/article-pdf/23/1/207/42208957/kxaa015.pdf

77. Willis A, Bunge J, Whitman T. Improved detection of changes in species richness in high diversity microbial communities. J R Stat Soc Ser C Appl Stat. 2017;66(5):963–77. https://www.jstor.org/stable/44682601

78. Willis A, Bunge J. Estimating diversity via frequency ratios: Estimating diversity via ratios. Biometrics. 2015;71. https://doi.org/10.1111/biom.12332

79. Minot SS, Willis A.D. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. Microbiome. 2019;7(1):110. https://doi.org/10.1186/s40168-019-0722-6

80. Minot SS, Barry KC, Kasman C, Golob JL, Willis A.D. Geneshot: Gene-Level Metagenomics Identifies Genome Islands Associated With

Immunotherapy Response. Genome Biol. 2021;22(1):1–10. https://doi.org/10.1186/s13059-021-02355-6

81. Emerson SS, Williamson BD, Wolock CJ, Okonek T, Chen YT, Willis AD, Spieker AJ, Hee Wai T.Y. rigr: Regression, Inference, and General Data Analysis Tools in R. GitHub; 2022

82. Trinh P, Clausen DS, Willis A.D. happi: a hierarchical approach to pangenomics inference. Genome Biol. 2023;24(1):214. https://doi.org/10.1186/s13059-023-03040-6

83. Machado D, Barbosa JC, Domingos M, Almeida D, Andrade JC, Freitas AC, Gomes A.M. Revealing antimicrobial resistance profile of the novel probiotic candidate faecalibacterium prausnitzii dsm 17677. Int J Food Microbiol. 2022;363:109501. https://doi.org/10.1016/j.ijfoodmicro.2021.109501

84. Kumari M, Singh P, Nataraj BH, Kokkiligadda A, Naithani H, Azmal Ali S, Behare PV, Nagpal R. Fostering next-generation probiotics in human gut by targeted dietary modulation: An emerging perspective. Food Res Int. 2021;150:110716. https://doi.org/10.1016/j.foodres.2021.110716

85. Baxter NT, Schmidt AW, Venkataraman A, Kim KS, Waldron C, Schmidt T.M. Dynamics of human gut microbiota and short-chain fatty acids in response to dietary interventions with three fermentable fibers. bioRxiv. 2018;10(1):1–13. https://doi.org/10.1101/487900

86. Wexler H.M. Bacteroides: the good, the bad, and the nitty-gritty. Clin Microbiol Rev. 2007;20(4):593–621. https://doi.org/10.1128/CMR.00008-07

87. Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, Armanini F, Manghi P, Bonham K, Zolfo M, De Filippis F, Magnabosco C, Bonneau R, Lusingu J, Amuasi J, Reinhard K, Rattei T, Boulund F, Engstrand L, Zink A, Collado MC, Littman DR, Eibach D, Ercolini D, Rota-Stabelli O, Huttenhower C, Maixner F, Segata N. The prevotella copri complex comprises four distinct clades underrepresented in westernized populations. Cell Host Microbe. 2019;26(5):666–6797. https://doi.org/10.1016/j.chom.2019.08.018

88. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nat Rev Microbiol. 2021;19(6):347–59. https://doi.org/10.1038/s41579-020-00497-1

89. Wang Y, Lyu N, Liu F, Liu WJ, Bi Y, Zhang Z, Ma S, Cao J, Song X, Wang A, Zhang G, Hu Y, Zhu B, Gao G.F. More diversified antibiotic resistance genes in chickens and workers of the live poultry markets. Environ Int. 2021;153:106534. https://doi.org/10.1016/j.envint.2021.106534

90. Tong C, Xiao D, Xie L, Yang J, Zhao R, Hao J, Huo Z, Zeng Z, Xiong W. Swine manure facilitates the spread of antibiotic resistome including tigecycline-resistant tet(x) variants to farm workers and receiving environment. Sci Total Environ. 2022;808:152157. https://doi.org/10.1016/j.scitotenv.2021.152157

91. Jeamsripong S, Li X, Aly S, Su Z, Pereira R, Atwill E. Antibiotic resistance genes and associated phenotypes in Escherichia coli and Enterococcus from cattle at different production stages on a dairy farm in Central California. Antibiotics. 2021;10:1042. https://doi.org/10.3390/antibiotics10091042

92. Taylor EA, Jordan ER, Garcia JA, Hagevoort GR, Norman KN, Lawhon SD, Piñeiro JM, Scott H.M. Effects of two-dose ceftiofur treatment for metritis on the temporal dynamics of antimicrobial resistance among fecal escherichia coli in holstein-friesian dairy cows. PLoS ONE. 2019;14(7):0220068. https://doi.org/10.1371/journal.pone.0220068

93. Forster SC, Liu J, Kumar N, Gulliver EL, Gould JA, Escobar-Zepeda A, Mkandawire T, Pike LJ, Shao Y, Stares MD, Browne HP, Neville BA, Lawley T.D. Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome. Nat Commun. 2022;13(1):1445. https://doi.org/10.1038/s41467-022-29096-9

94. Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, Almeida M, Quinquis B, Levenez F, Galleron N, Gougis S, Rizkalla S, Batto J-M, Renault P, Doré J, Zucker J-D, Clément K, Ehrlich SD, Blottière H, Leclerc M, Juste C, de Wouters T, Lepage P, Fouqueray C, Basdevant A, Henegar C, Godard C, Fondacci M, Rohia A, Hajduch F, Weissenbach J, Pelletier E, Le Paslier D, Gauchi J-P, Gibrat J-F, Loux V, Carré W, Maguin E, van de Guchte M, Jamet A, Boumezbeur F, Layec S, consortium AM, consortium members A.M. Dietary intervention impact on gut microbial gene richness. Nature. 2013;500(7464):585–8. https://doi.org/10.1038/nature12480

Trinh *et al. BMC Genomics* (2024) 25:708

Page 18 of 18

95. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clément K, Doré J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker J-D, Raes J, Hansen T, Guedon E, Delorme C, Layec S, Khaci G, van de Guchte M, Vandemeulebrouck G, Jamet A, Dervyn R, Sanchez N, Maguin E, Haimet F, Winogradski Y, Cultrone A, Leclerc M, Juste C, Blottière H, Pelletier E, LePaslier D, Artiguenave F, Bruls T, Weissenbach J, Turner K, Parkhill J, Antolin M, Manichanh C, Casellas F, Boruel N, Varela E, Torrejon A, Guarner F, Denariaz G, Derrien M, van Hylckama Vlieg JET, Veiga P, Oozeer R, Knol J, Rescigno M, Brechot C, M'Rini C, Mérieux A, Yamada T, Bork P, Wang J, Ehrlich SD, Pedersen O, consortium M. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500(7464):541–6. https://doi.org/10.1038/nature12506

96. Nonnenmann MW, Gimeno Ruiz de Porras D, Levin J, Douphrate D, Boggaram V, Schaffer J, Gallagher M, Hornick M, Reynolds S. Pulmonary function and airway inflammation among dairy parlor workers after exposure to inhalable aerosols. Am J Ind Med. 2017;60(3):255–63. https://doi.org/10.1002/ajim.22680

97. Davidson ME, Schaeffer J, Clark ML, Magzamen S, Brooks EJ, Keefe TJ, Bradford M, Roman-Muniz N, Mehaffy J, Dooley G, Poole JA, Mitloehner FM, Reed S, Schenker MB, Reynolds S.J. Personal exposure of dairy workers to dust, endotoxin, muramic acid, ergosterol, and ammonia on large-scale dairies in the high plains western united states. J Occup Environ Hyg. 2018;15(3):182–93. https://doi.org/10.1080/15459624.2017.1403610

98. Stoleski S, Minov J, Karadzinska-Bislimovska J, Mijakoski D, Atanasovska A, Bislimovska D. Asthma and chronic obstructive pulmonary disease associated with occupational exposure in dairy farmers - importance of job exposure matrices. Open Access Maced J Med Sci. 2019;7(14):2350–9. https://doi.org/10.3889/oamjms.2019.630

99. Wang H, Liu J-S, Peng S-H, Deng X-Y, Zhu D-M, Javidiparsijani S, Wang G-R, Li D-Q, Li L-X, Wang Y-C, Luo J-M. Gut-lung crosstalk in pulmonary involvement with inflammatory bowel diseases. World J Gastroenterol. 2013;19(40):6794–804. https://doi.org/10.3748/wjg.v19.i40.6794

100. Keely S, Talley NJ, Hansbro P.M. Pulmonary-intestinal cross-talk in mucosal inflammatory disease. Mucosal Immunol. 2012;5(1):7–18. https://doi.org/10.1038/mi.2011.55

101. Raftery AL, Tsantikos E, Harris NL, Hibbs M.L. Links between inflammatory bowel disease and chronic obstructive pulmonary disease. Front Immunol. 2020;11:2144. https://doi.org/10.3389/fimmu.2020.02144

102. Duffy B, Holliger E, Walsh F. Streptomycin use in apple orchards did not increase abundance of mobile resistance genes. FEMS Microbiol Lett. 2014;350(2):180–9. https://doi.org/10.1111/1574-6968.12313

103. Vidaver AK. Uses of antimicrobials in plant agriculture. Clin Infect Dis. 2002;34(Supplement 3):107–10. https://doi.org/10.1086/340247

104. Farm Labor. https://www.ers.usda.gov/topics/farm-economy/farm-labor/#demographic. Accessed 20 May 2024

105. Wu L, Xie X, Li Y, Liang T, Zhong H, Ma J, Yang L, Yang J, Li L, Xi Y, Li H, Zhang J, Chen X, Ding Y, Wu Q. Metagenomics-based analysis of the age-related cumulative effect of antibiotic resistance genes in gut microbiota. Antibiotics. 2021;10(8). https://doi.org/10.3390/antibiotics10081006

106. Panikkar B, Barrett M-K. Precarious essential work, immigrant dairy farmworkers, and occupational health experiences in vermont. Int J Environ Res Public Health. 2021;18(7). https://doi.org/10.3390/ijerph18073675

Publisher's Note