ONLINE DATA SUPPLEMENT

Community analysis and co-occurrence patterns in airway microbial communities during health and disease

Gisli G Einarsson [1,3*], Jiangchao Zhao[4,7], John J LiPuma[4,5], Damian G Downey[1,6],

Michael M Tunney [1,2*] and J Stuart Elborn [3*]

[1] *Halo Research Group, Queen's University Belfast, Belfast, United Kingdom; g.einarsson@qub.ac.uk (GGE); m.tunney@qub.ac.uk (MMT)*
[2] *School of Pharmacy, Queen's University Belfast, Belfast, United Kingdom*
[3] *Centre for Experimental Medicine, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom; s.elborn@qub.ac.uk (JSE)*
[4] *Department of Pediatrics and Communicable Diseases, University of Michigan Medical School, Ann Arbor, MI, 48109, United States of America; jlipuma@med.umich.edu (JJL)*
[5] *Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, 48109, United States of America*
[6] *Northern Ireland Regional Adult Cystic Fibrosis Centre, Belfast City Hospital, Belfast Health & Social Care Trust (BHSCT), Belfast, United Kingdom; Damian.Downey@belfasttrust.hscni.net (DGD)*
[7]Department of Animal Science, University of Arkansas, Fayetteville, AR, 72701, *United States of America; jzhao77@uark.edu (JZ)*

Correspondence: Gisli Einarsson, PhD, Centre for Experimental Medicine, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom.
E-mail: g.einarsson@qub.ac.uk
Tel: +44(0)28 90 975876 Fax: +44 (0) 28 90 972671

* joint senior authors

Detailed Methods:

Methods:

Cohorts and Sample Data Processing

Sample information for the HMP dataset was gathered from the information available at http://hmpdacc.org, followed by retrieval of the relevant sequence data from the NCBI's sequence read archive (*SRA*) (http://www.ncbi.nlm.nih.gov/sra) using the "SRAdb" library in the Bioconductor (http://bioconductor.org/biocLite.R) package as implemented in the R environment (http://www.r-project.org). Following retrieval of the HMP datasets, all the sequence data was concatenated prior to initiation of data processing and analysis.

For each sample, low quality amplicons of less than 200 nucleotides, potential chimeric sequences and singleton (i.e., those represented by a single sequence) sequences were removed, followed by screening and removal of all amplicons of non-bacterial origin from the dataset inferring various datasets obtained from the NCBI Reference Sequence Database repository (http://www.ncbi.nlm.nih.gov/refseq/). Following quality filtering, samples were normalized to 2000 sequence reads to allow for comparison between samples that contained unequal number of amplicon reads. Taxonomic tables (OTU tables; OTU = Operational taxonomic units are defined according to clustering on the basis of DNA 97% sequence identity) for the community as a whole were prepared using the UCLUST algorithm (1) during analysis in the QIIME pipeline (v 1.8.0) (2). For OTU and taxonomic assignments of sequences, the open-reference OTU picking method as implemented in QIIME was applied, followed by sequences being assigned to their corresponding taxonomic ranks using the Ribosomal Database Project (RDP) Naïve Bayesian Classifier v.2.2 (3), at the default confidence threshold of 0.8, trained on the SILVA reference database (release 111) (4).

454-FLX Titanium sequence data for all ten sites of the upper airways used in this study were acquired from the HMP 16S rRNA sequence depository (http://hmpdacc.org/HMQCP/). 454-FLX Titanium sequence data for all corresponding BE samples used in this study were deposited to the EBI Sequence Read Archive (http://www.ebi.ac.uk/ena/) as a part of a previous study conducted by our group (13). 454-FLX Titanium sequence data for CF samples (sputum and mouthwash) were deposited to MG-RAST sequence depository (http://metagenomics.anl.gov/) as a part of a previous study conducted by our group (12, 15). Furthermore, to increase the number of stable patients in our CF cohort, raw data from the University of Michigan were provided by Professor John LiPuma. To avoid duplication of previously deposited sequence data, we have provided one FASTA (.fna) file containing all our demultiplexed sequences used in this study, which is available from the MG-RAST database (http://metagenomics.anl.gov/linkin.cgi?metagenome=4633193.3) with MG-RAST ID: mgm4633193.3. A mapping file associated with the current study is provided as Supplementary File S1.

Comparison between community structures

To assess the distribution of taxa within the airways as a whole, we ranked taxa according to their mean relative abundance from the normalised dataset into three main categories. The "Generalist" or "core" taxa were determined according to individually assigned OTUs at the sequence level and here represents the fraction of samples that an OTU was observed in to be considered to belong to the "core" microbiota, with the cut-off set as those taxa occurring in at least >70% of all samples from the main OTU table as implemented in QIIME. The second group consisted of "Niche Specialists" and was defined as those taxa showing strong niche association (with the majority of the sequence counts originating in a single cohort). The third group consisted of "Chronic Airways Disease Specialists" and was defined as those taxa representing

the majority of the relative abundance and occurrence associated with chronic disease (CF or BE). Community metrics were assessed on the basis of taxa richness, i.e. the number of normalised OTUs per sample, and by the nonparametric Shannon-Wiener (SW) diversity index.

Co-occurrence analysis

To assess if non-random co-occurrence patterns existed within our sample cohort, we initially performed calculations based on the *C*-score (checkerboard units) calculations (5) under a null model with preserved site frequencies. The *C*-score measures the average number of checkerboard units, or co-occurrences, between all possible OTU pairs and enabled determination of whether taxa/genera tended to aggregate together more than would be expected due to chance alone.

For further simplification of this complex dataset and analysis of potential co-occurrence between the more prevalent groups, we retained data for taxa that accounted for >0.5% of sequences in the 1057 samples of the combined upper-airways for inter-sample comparisons with CF, CFMW and BE. For niche associated co-occurrence analysis, we retained taxa that accounted for >0.1% of all sequences in each individual cohort. By applying a filtering step prior to the generation of any potential co-occurrence networks, we reduced both the overall complexity of the data, as well as the effect of false-positive correlations arising from spurious associations caused by poorly represented OTUs (containing many zeros in large community/abundance based matrices) within the dataset.

Network inference was generated by calculating all possible Spearman´s rank correlation coefficients (ρ) between taxon pairs. To further limit potential false-positive, or spurious associations between corresponding taxon pairs, we performed multiple testing correction, which produced adjusted p-values according to the Benjamini-Hochberg-Yekutieli false discovery rate (FDR) correction. We considered a valid co-occurrence between two different

taxa if the Spearman's correlation coefficient (ρ) was both >0.5 and statistically significant (adjusted p<0.005). In the reconstructed co-occurrence networks, all nodes represent taxa that show at least 97% identity, with the edges (i.e., connections) corresponding to a significant correlation between nodes (i.e., taxa; based on ρ and significance according to the adjusted p-value).

Community variance - Principal Component Analysis

To compare the effect each taxon had on the observed variance within its environment, we conducted principal component analysis (PCA) on the normalized OTU abundance measures.

Technical variance - using unweighted and weighted UniFrac distance metrics and Bray-Curtis dissimilarities

To assess if differences in primer combination for amplicon generation, e.g. a subset of samples were processed using V1V2 region of the 16S rRNA marker gene compared to V1V3 for the majority of the other samples, affected the results we looked at factors associated with the microbiota community structure, as assessed using PERMANOVA with the adonis function (9999 permutations) of the unweighted and weighted UniFrac distances metrics and Bray-Curtis dissimilarities vegan package in R (6).

Statistical analysis

All statistical analyses and graphical representations were carried out in Graphpad Prism (ver. 5.00) and the R environment (http://www.r-project.org) using vegan (version 2.4-1) (6), igraph (7), ggplot2 (version 2.1.0) (8), phyloseq (version 1.18.0) (9), ampvis (version 1.26.0) (10), stringr (version 0.6.2) (11), reshape2 (version 1.2.2) (12), grid (version 3.0.1) (13) and vegan (2.4-3) packages. Post analysis and visualisation of co-occurrence networks was performed

within the Gephi package (release 0.8.2) (14). A Spearman rank correlation coefficient (ρ) was calculated to measure the strength of association between different taxa, as implemented in the Hmisc (version 3.12-2) (15) in the R environment (http://www.r-project.org).

Community richness and diversity was compared between three or more cohorts by the Kruskal-Wallis test, followed by post-hoc testing using the non-parametric Mann-Whitney test with Bonferroni adjustment to evaluate differences between two sample cohorts. $P<0.05$ values were deemed statistically significant. Species assignment (closest neighbor - "best hit") to short amplicon sequences from a number of the most prevalent taxa in both upper and lower airways was attempted through a local implementation of BLASTn algorithm against RefSeq microbial reference sequence database release 61 (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/microbial/).

Comparison between community structures

To assess the distribution of taxa within the airways as a whole, we ranked taxa according to their mean relative abundance from the normalised dataset into three main categories. The "Generalist" or "core" taxa were determined according to individually assigned OTUs at the sequence level and here represents the fraction of samples that an OTU was observed in to be considered to belong to the "core" microbiota, with the cut-off set as those taxa occurring in at least 60% of samples from either disease associated cohorts (CF, BE and CFMW) and health associated cohorts (upper airways) and in >70% of all samples (combined) from the main OTU table as implemented in QIIME. The second group consisted of "Niche Specialists" and was defined as those taxa showing strong niche association (with the majority of the sequence counts originating in a single cohort). The third group consisted of "Chronic Airways Disease Specialists" and was defined as those taxa representing the majority of the relative abundance and occurrence associated with chronic disease (CF or BE). Community metrics were assessed

on the basis of taxa richness, i.e. the number of normalised OTUs per sample, and by the nonparametric Shannon-Wiener (SW) diversity index.

Community Structures affected by technical variation

Comparison into if use of different overlapping primer pairing (V1V2 vs. V1V3) for a subset of sample belonging to the CF cohort the p-value of <0.05 indicates that the grouping of samples by cohort and primer pairing is statistically significant. The R^2 value indicates that approximately 3-6% of the variation in distances is explained by the effect of primer pairing (i.e. V1V2 vs. V1V3). The results confirm that there is limited effect caused by the different primer pairs, with the main affect, accounting for 23-51% of the variability depending on the distance metric used, being due to the niche the corresponding samples originated from (Supplementary Table 3). We further explored if the how this affected samples originating from the CF cohort and were generated using both V1V2 and V1V3 variable regions. The results showed that the effect within this cohort the variance explained accounted for 3-6% dependent on the metric used.

Community differences within sites - species assignment and distribution

Despite inherent limitations associated with the common methodologies used to assign taxonomic ranking to the species level, we performed species assignment to provide a snapshot of potential differences that may occur between health and disease associated communities. We examined a number of the main taxa associated with "core" communities (i.e., *Streptococcus* spp., *Prevotella* spp., and *Veillonella* spp.), as well as bacterial taxa important in communities of chronic airways diseases (i.e., *Pseudomonas* spp., and *Haemophilus* spp.). We extracted all sequences belonging to the genera of interest and assigned each sequence to its proposed species followed by adjustment for their intra-genus relative abundances.

Relative abundance of the top 20 taxa differed significantly within the thirteen cohorts (Supplementary Figure S2). Similarly, when we assigned different sequence types at the species

level, there were differences in distribution of species-level taxa between upper and lower airways cohorts for a number of the most relevant taxa (Supplementary Table S5 [a-e]). For example, members of the genus *Streptococcus* spp. were assigned to thirteen different "species" on the basis of >1% intra-genus abundance with *Streptococcus pseudopnuemoniae* the most abundant sequence type in both the healthy upper airways and BE sputum. In contrast, in CF sputum and CFMW, *Streptococcus oralis* contributed most to intra-genus sequence abundance. *Prevotella melaninogenica* accounted for the majority of the relative abundance in all four groups. For *Veillonella* spp. the largest number of sequences was assigned to *V. atypica* in CF and BE, accounting for 53.5% and 52.4%, respectively, of the relative abundance for the genus. For the upper airways and CFMW, *V. parvula* accounted for the largest proportion of sequences within the genera, 69.12% and 54.85%, respectively. As expected, *Pseudomonas aeruginosa* accounted for the majority of the assigned sequences belonging to the genus *Pseudomonas* within CF (99.95%) and BE (99.96%) sputum. Similarly, in CF and BE sputum, *Haemophilus influenzae* accounted for 94.02% and 97.80%, of the intra-genus relative abundance, respectively. However, *H. parainfluenzae* was the dominant taxa within the upper airways and CFMW communities accounting for 73.37% and 80.11% of the intra-genus abundance, respectively (Supplementary Table S5 a-e).

References

1. Edgar RC. Search and clustering orders of magnitude faster than blast. *Bioinformatics* 2010;26:2460-2461.
2. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI. Qiime allows analysis of high-throughput community sequencing data. *Nature methods* 2010;7:335-336.
3. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 2007;73:5261-5267.
4. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. Silva: A comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic Acids Research* 2007;35:7188-7196.
5. Stone L, Roberts A. The checkerboard score and species distributions. *Oecologia* 1990;85:74-79.
6. Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M. The vegan package. *Community ecology package Disponível em: http://www R-project org Acesso em* 2007;10:2008.
7. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal Complex Systems* 2006;1695:1695.
8. Wickham H. Ggplot2: Elegant graphics for data analysis. 2009. Springer, New York; 2009.
9. McMurdie PJ, Holmes S. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one* 2013;8:e61217.
10. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. Back to basics – the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS ONE* 2015;10:e0132783.
11. Wickham H. Stringr: Make it easier to work with strings., 2010. *URL http://CRAN R-project org/package= stringr R package version 04* 2010.
12. Wickham H. Reshaping data with the reshape package. *J Stat Softw* 2007;21:1-20.
13. Murrell P. The grid graphics package. *R News* 2002;2:14-19.
14. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. ICWSM; 2009.
15. Harrell Jr F. Hmisc: Harrell miscellaneous. R package version 3.12-2. 2013.