Methods

Nate Olson

2017-11-03

0.1 Two-Sample Titration Design

Samples from a vaccine trial were selected for use in the study (Harro et al. 2011). Five trial participants were selected based on the following criteria no *Escherichia coli* detected in stool samples using qPCR and 16S metagenomic sequencing before exposure (pre-exposure) to Enterotoxigenic *Escherichia coli* (ETEC)) and timepoints with the highest concentration of *E. coli* after exposure (post-exposure) (Pop et al. 2016, Fig. 1A). For the two-sample titration post-exposure samples were titrated into pre-exposure samples with log_2 changes in pre to post sample proportions (Fig. 1B). Unmixed samples were diluted to 12.5 $ng/\mu L$ in tris-EDTA buffer prior to making two-sample titrations. Initial DNA concentration was measured using NanoDrop ND-1000 (Thermo Fisher Scientific Inc. Waltham, MA USA).

By using a two-sample titration mixture design the expected relative abundance of a feature can be determined using the following equation (1). Where θ_i, is the proportion of post-exposure DNA in titration i, C_{ij} is the relative abundance of feature j in titration i, and C_{post_j} and C_{pre_j} are the relative abundance of feature j in the unmixed pre- and post-exposure samples.

$$C_{ij} = \theta_i C_{post_j} + (1 - \theta_i) C_{pre_j} \tag{1}$$

0.2 Titration Validation

qPCR was used to validate the volumetric mixing of the unmixed samples and check of differences in the proportion of prokaryotic DNA across the titrations. To ensure that the two-sample titrations were volumetrically mixed according to the mixture design independent ERCC plasmids were spiked into the unmixed pre- and post-exposure samples (**TODO** Table ERCC) (Baker et al. 2005) (NIST SRM SRM 2374). The ERCC plasmids were resuspendended in 100 $ng/\mu L$ tris-EDTA buffer and 2 $ng/\mu L$ was spiked into the approporiate unmixed sample. Plasmid abundance was quantified using TaqMan gene expression assays (FAM-MGB) (Catalog # 4448892, ThermoFisher) specific to each ERCC plasmids using the TaqMan Universal MasterMix II (Catalog # 4440040, ThermoFisher Waltham, MA USA). To check for differences in the proportion of bacterial DNA in the pre- and post-exposure samples, bacterial DNA concentration in the titrations was quantified using the Femto Bacterial DNA quantification kit (Zymo Research, Irvine CA). All samples were run in triplicate along with a standard curve. An in-house standard curve consisting of log_{10} dilutions of *E. coli* DNA was used as the standard curve. All qPCR assays were performed using the QuantStudio Real-Time qPCR (ThermoFisher). The amplification data and Ct values were exported from the QuantStudio™ Design and Analysis Software v1.4.1 as tsv files for statistical analysis. Statistical analysis was performed using the R programming language.

0.3 Sequencing

The 45 samples (seven titrations and two unmixed samples for the five biological replicates) were processed using a standard 16S rRNA amplicon sequencing workflow based on the Illumina 16S library protocol (16S Metagenomic Sequencing Library Preparation, posted date 11/27/2013, dowloaded from https://support.illumina.com). The protocol consisted of an initial 16S rRNA PCR followed by a separate sample indexing PCR prior to normalization and pooling.

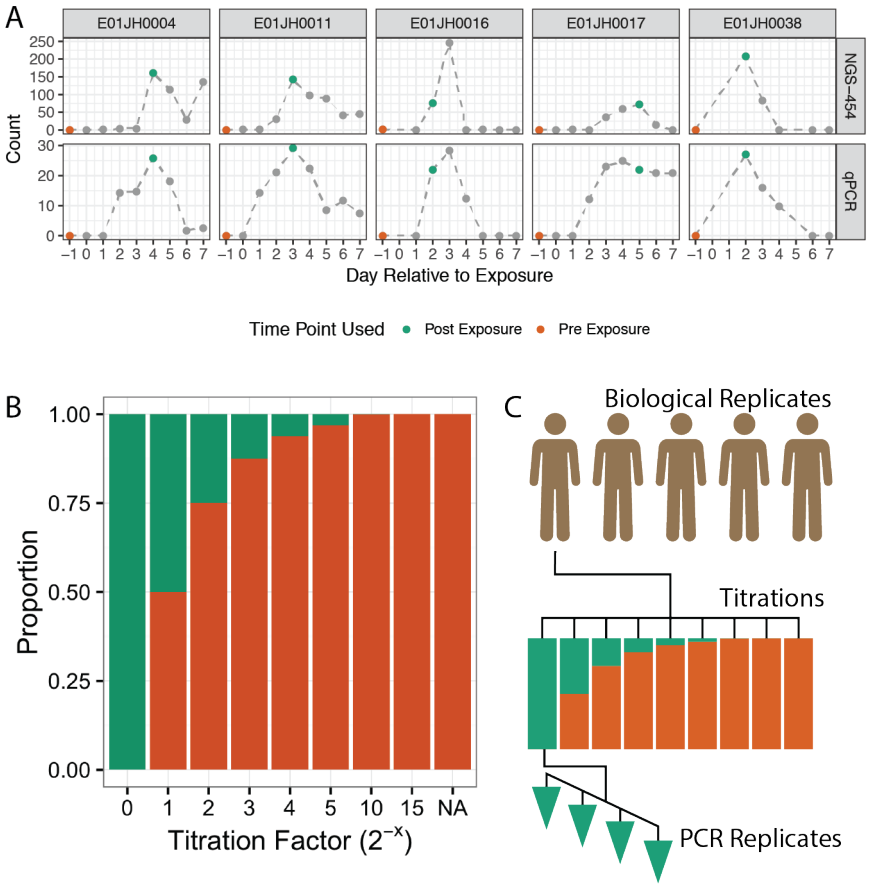

Figure 1: Sample selection and experimental design for two-sample titration 16S rRNA metagenomic sequencing assessment dataset. A) Pre- and post-exposure samples from five participants in a vaccine trial (Harro et al. 2011) were selected based on *Escherichia coli* abundance measured using qPCR and 454 16S rRNA metagenomics sequencing (454-NGS), data from (Pop et al. 2016). Pre- and post-exposure samples are indicated with orange and green data points. Grey indicates other samples from the vaccine trial time series. B) The pre-exposure samples were titrated into post-exposure samples following a log_2 dilution series. The NA titration factor represents the unmixed pre-exposure sample. C) Pre- and post-exposure samples from the five vaccine trial participants were used to generate independent two-sample titration series. The result was a total of 45 samples, 7 titrations + 2 unmixed samples times 5 biological replicates. Four replicate PCRs were performed for each of the 45 samples resulting in 190 PCRs.

A total of 192 PCRs were run including four PCR replicates per sample and 12 no template controls. The 16S PCR targeted the V3-V5 region, Bakt_341F and Bakt_806R (Klindworth et al. 2012). The V3-V5 target region is 464 bp, with forward and reverse reads overlaping by 136 bp (Yang, Wang, and Qian 2016) (http://probebase.csb.univie.ac.at). The primer sequences include additional overhang adapter sequences to facilitate library preparation (forward primer 5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGC-CTACGGGNGGCWGCAG - 3' and reverse primer 5'- GTCTCGTGGGCTCGGAGATGTGTATAAGA-GACAGGACTACHVGGGTATCTAATCC - 3'). The 16S targeted PCR was performed according to the Illumina protocol using the KAPA HiFi HotStart ReadyMix reagents (KAPA Biosystems, Inc. Wilmington, MA). The resulting PCR product was verified using agarose gel electrophoresis. Quality control DNA concentration measurements were made after the initial 16S rRNA PCR, indexing PCR, and normalization. DNA concentration was measured using SpextraMax Accuclear Nano dsDNA Assay Bulk Kit (Part# R8357#, Lot 215737, Molecular Devices LLC. Sunnyvale CA, USA) and flourescent measurements were made with a Molecular Devices SpectraMax M2 spectraflourometer (Molecular Devices LLC. Sunnyvale CA, USA).

The 16S rRNA PCR product was used to generate sequencing libraries. The initial PCR products were purified using AMPure XP beads (Beckman Coulter Genomics, Danvers, MA) following the manufactures protocol. After purification the 192 samples were indexed using the Illumina Nextera XT index kits A and D (Illumina Inc., San Diego CA). After purification the 192 samples were indexed using the Illumina Nextera XT index kits A and D (Illumina Inc., San Diego CA). Prior to pooling the purified sample concentration was normalized using SequalPrep Normalization Plate Kit(Catalog n. A10510-01, Invitrogen Corp., Carlsbad, CA), according to the manufactuers protocol. The pooled library concentration was measured using the Qubit dsDNA HS Assay Kit (Part# Q32851, Lot# 1735902, ThermoFisher, Waltham, MA USA). Due to the low concentration of the pooled amplicon library the modified protocol for low concentration libraries was used. The library was run on a Illumina MiSeq and base calls were made using Illumina Real Time Analysis Software version 1.18.54.

0.4 Sequence Processing

Sequence data was processed using four bioinformatic pipelines, Mothur (Schloss et al. 2009), QIIME (Caporaso et al. 2010), DADA2 (Callahan et al. 2016), and unclustered sequences as a control. Code used to run the bioinformatic pipelines is available at https://github.com/nate-d-olson/mgtst_pipelines. The Mothur (version 1.37, http://www.mothur.org/) pipeline used was based on the MiSeq SOP (Schloss et al. 2009; Kozich et al. 2013). As a different 16S rRNA region was sequenced than the region the SOP was developed for the procedure was modified to account for smaller overlap between the forward and reverse reads relative to the amplicons used in the protocol. The Mothur pipeline included an initial pre-processing step where forward and reverse reads were merge using the Needleman-Wunsch algorithm. Low quality reads were identified based on presence of ambiguous bases, reads that failed alignment to the SILVA reference database (V119, https://www.arb-silva.de/) (Quast et al. 2012), and chimeras were filtered from the dataset. Chimera filtering was performed using UChime without a reference database (Edgar et al. 2011). OTU clustering was performed using the OptiClust algorithm with a clustering threshold of 0.97 (Westcott and Schloss 2017). The RDP classifier implemented in mothur was used for taxonomic classification against the mothur provided version of the RDP v9 training set (Q. Wang et al. 2007). The QIIME open-reference clustering pipeline for paired-end Illumina data was performed according to the online tutorial (http://nbviewer.jupyter.org/github/biocore/qiime/blob/1.9.1/examples/ipynb/illumina_overview_tutorial.ipynb) using QIIME version 1.9.1 (Caporaso et al. 2010). Briefly the QIIME pipeline uses fastq-join to merge paired-end reads (Aronesty 2011), the Usearch algorithm (Edgar 2010) and Greengenes database version 13.8 with a 97% similarity threshold (DeSantis et al. 2006) was used for open-reference clustering. DADA2 a R native pipeline was also used to process the sequencing data (Callahan et al. 2016). The pipeline includes a sequence inference step and taxonomic classification using the DADA2 implementation of the RDP naive bayesian classifier (Q. Wang et al. 2007) and the SILVA database V123 provided by the DADA2 developers (Quast et al. 2012, https://benjjneb.github.io/dada2/training.html).
The unclustered pipeline was based on the mothur *de-novo* clustering pipeline, where the paired-end reads were merged and filtered using the `make.contigs` command and then dereplicated. Reads were aligned to the reference Silva alignment (V119, https://www.arb-silva.de/) and reads failing alignment were excluded from

the dataset. To limit the size of the dataset the most abundant 40,000 OTUs (comparable to the mothur dataset), across all samples, were used as the unclustered dataset. Taxonomic classification of the unclustered sequences was performed using the same RDP classifier implmented in mothur used for the *de-novo* pipeline.

0.5 Data Analysis

- To generate summaries of QA metrics for the 384 datasets in the study (192 samples with forward and reverse reads) used the bioconductor `Rqc` package (REF) to calculate the quality metrics used in the following analysis.
- negative binomial model was used to calculate the average relative abundance across PCR replicates.
- log changes between titrations and pre- and post-exposure samples were calculated using EdgeR (M. D. Robinson, McCarthy, and Smyth 2010; McCarthy et al. 2012).

0.5.1 Theta Inference

To account for differences in the proportion of bacterial DNA in the pre- and post-exposure samples. A linear model was used to infer θ in equation (2), where $\mathbf{C}$ is a vector of counts for a set of features, $\mathbf{C}_{obs_j}$ observed counts for titration j, with $\mathbf{C}_{pre_j}$ cand $\mathbf{C}_{post_j}$ representing the vector of counts for the same features for the unmixed pre- and post-exposure samples. To summarize counts across PCR replicates and account for differences in sequencing depth, negative binomial relative abundance estimates were used to infer θ. 16S rRNA sequencing count data is know to have a non-normal mean-variance relationship resulting in poor model fit for standard linear regression. Generalized linear models provide an alternative to standard least-squares regression however, the above model is additive and therefore unable to directly infer θ_j in log-space. To address this issue we fit the model using a standard least-squares regression then obtained non-parametric 95 % confidence intervals for the θ estimates by bootstraping with 1000 replicates. To limit the impact of uninformative and low abundance features a subset of features were used to infer θ. Features used were individual specific. To use a feature was observed in at least 14 of the 28 total titration PCR replicates (4 pcr replicates per titration, 7 titrations), greater than 1 log_2 fold-change between the pre- and post-exposure samples, and present in all four or none of the pre- and post-exposure PCR replicates.

$$C_{obs_j} = \theta_j(C_{post_j} - C_{pre_j}) + C_{pre_j} \tag{2}$$

0.6 Quantitative Assessment

To quantitatively assess the count table values the expected relative abundance and log fold-change values were compared to the relative abundance estimates calculated using a negative binomial model and the EdgeR log fold-change estimates. Equation (1) and the inferred θ values were used to calculate the expected feature relative abundance. The error rate bias and variance for the relative abundance estimates were compared across pipelines and biological replicates. Error rate was defined as $(exp - obs)/exp$. Mixed effects models were used to compare feature-level error rate bais and variance across pipelines accounting for individual effect. Feature-level bias and variance were evaluated using the median error rate and robust COV, $IRQ/median$, respectively. Large feature-level error rate bias and variance outliers were observed, these outliers were excluded from the mixed effects model to minimize biases in the model due to poor fit and were characterized independently.

To assess differential abundance log fold-change estimates, log fold-change between all titrations were compared to the expected log fold-change values for the pre-specific and pre-dominant features. When assuming the feature is only present in pre-exposure samples the expected log fold-change is independent of the observed counts for the unmixed samples. Expected log fold-change between titrations i and j is calculated using (3), where θ is the proportion of post-exposure bacterial DNA in a titration. Pre-dominant and pre-specific features were defined as features observed in all four pre-exposure PCR replicates and a log fold-change

between pre- and post-exposure samples greater than 5.
Pre-specific features were not observed in any of the post-exposure PCR replicates and pre-dominant features were observed in one or more of the post-exposure PCR replicates. Only individuals with consistent inferred and estimated θ values were included in the log fold-change analysis, E01JH0004, E01JH0011, and E01JH0016.

$$logFC_{ij} = log_2\left(\frac{1-\theta_i}{1-\theta_j}\right) \tag{3}$$

0.7 Qualitative Assessment

For the qualitative measurement assessment we evaluated features only observed in either the unmixed samples, unmixed-specific features, or the titrations, titration-specific features. Features are unmixed- or titration-specific due to differences in sampling depth (number of sequences) between the unmixed samples and titrations or an artifact of the feature inference process.

We tested if sampling alone could explain feature specificity. For unmixed-specific features we used a binomial test and for titration-specific features we used Monte-Carlo simulation and a Bayesian hypothesis test. For both tests p-values were adjusted for multiple comparisons using the Benjamini & Hochberg method (Benjamini and Hochberg 1995). To determine if sampling alone can explain unmixed-specific features the binomial test was used to test the following hypothesis;

H_0 - Given no observed counts and the total abundance for a titration the true proportion of a feature is **equal to** the expected proportion.

H_1 - Given no observed counts and the total abundance for a titration the true proportion of a feature is **less than** the expected proportion.

To test if titration-specific features could be explained by sampling alone we used Monte-Carlo simulation and a Bayesian hypothesis test. For the simulation we assumed a binomial distribution given the observed total abundance and a uniform distribution of proportions, 0 to the minimum expected proportion. The minimum expected proportion, $\pi_{min_{exp}}$, is calculated using the mixture equation (1) and the minimum observed feature proportion for unmixed pre-exposure, $\pi_{min_{pre}}$, and post-exposure $\pi_{min_{post}}$ samples for each individual and pipeline. For features not present in unmixed samples the assumption is that the feature proportion is less than $\pi_{min_{exp}}$.

We formulated our null and alternative hypothesis for the Bayesian test as follows,

H_0 - Given the total abundance for a sample and minimum expected proportion the true proportion of a feature is **less than** the minimum expected observed proportion.
H_1 - Given the total abundance for a sample and minimum expected proportion the true proportion of a feature is **greater than or equal to** the minimum expected proportion.

The following equations @ref(eq:probPi, eq:probC) were used to calculate the p-value for the Bayesian hypothesis test assuming equal priors, i.e. $P(\pi < \pi_{min_{exp}}) = P(\pi \geq \pi_{min_{exp}})$.

$$p = P(\pi < \pi_{min_{exp}} | C \geq C_{obs}) = \frac{P(C \geq C_{obs} | \pi < \pi_{min_{exp}}) P(\pi < \pi_{min_{exp}})}{P(C \geq C_{obs})} \tag{4}$$

$$P(C \geq C_{obs}) = P(C \geq C_{obs} | \pi < \pi_{min_{exp}}) P(\pi < \pi_{min_{exp}}) + P(C \geq C_{obs} | \pi \geq \pi_{min_{exp}}) P(\pi \geq \pi_{min_{exp}}) \tag{5}$$

Aronesty, Erik. 2011. "Ea-Utils: Command-Line Tools for Processing Biological Sequencing Data." *Expression Analysis, Durham, NC*.

Baker, Shawn C, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, et al. 2005. "The External Rna Controls Consortium: A Progress Report." *Nature Methods* 2 (10). Nature Publishing Group: 731–34.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 289–300.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods*. Nature Publishing Group.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (5). Nature Publishing Group: 335–36.

DeSantis, Todd Z, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. 2006. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with Arb." *Applied and Environmental Microbiology* 72 (7). Am Soc Microbiol: 5069–72.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster Than Blast." *Bioinformatics* 26 (19). Oxford University Press: 2460–1.

Edgar, Robert C, Brian J Haas, Jose C Clemente, Christopher Quince, and Rob Knight. 2011. "UCHIME Improves Sensitivity and Speed of Chimera Detection." *Bioinformatics* 27 (16). Oxford Univ Press: 2194–2200.

Harro, Clayton, Subhra Chakraborty, Andrea Feller, Barbara DeNearing, Alicia Cage, Malathi Ram, Anna Lundgren, et al. 2011. "Refinement of a Human Challenge Model for Evaluation of Enterotoxigenic Escherichia Coli Vaccines." *Clinical and Vaccine Immunology* 18 (10). Am Soc Microbiol: 1719–27.

Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. 2012. "Evaluation of General 16S Ribosomal Rna Gene Pcr Primers for Classical and Next-Generation Sequencing-Based Diversity Studies." *Nucleic Acids Research*. Oxford Univ Press, gks808.

Kozich, James J, Sarah L Westcott, Nielson T Baxter, Sarah K Highlander, and Patrick D Schloss. 2013. "Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the Miseq Illumina Sequencing Platform." *Applied and Environmental Microbiology* 79 (17). Am Soc Microbiol: 5112–20.

McCarthy, Davis J., Chen, Yunshun, Smyth, and Gordon K. 2012. "Differential Expression Analysis of Multifactor Rna-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): –9.

Pop, Mihai, Joseph N Paulson, Subhra Chakraborty, Irina Astrovskaya, Brianna R Lindsay, Shan Li, Héctor Corrada Bravo, et al. 2016. "Individual-Specific Changes in the Human Gut Microbiota After Challenge with Enterotoxigenic Escherichia Coli and Subsequent Ciprofloxacin Treatment." *BMC Genomics* 17 (1). BioMed Central: 1.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2012. "The Silva Ribosomal Rna Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (D1). Oxford University Press: D590–D596.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26: –1.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-

Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23). Am Soc Microbiol: 7537–41.

Wang, Qiong, George M Garrity, James M Tiedje, and James R Cole. 2007. "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy." *Applied and Environmental Microbiology* 73 (16). Am Soc Microbiol: 5261–7.

Westcott, Sarah L, and Patrick D Schloss. 2017. "OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units." *mSphere* 2 (2).

Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. "Sensitivity and Correlation of Hypervariable Regions in 16S rRNA Genes in Phylogenetic Analysis." *BMC Bioinformatics* 17 (1). BioMed Central: 1.