Data

Our aim was to be able to infer genome abundance using mHMM depth coverage, beginning with the following four genes: Frr, RplB, RplK, and RpsT. We had depth coverage data for 25 different genomes. For each genome, Frr was associated with 104 mHMMs, RplB had 186, RplK had 65, and RpsT had 22. Each mHMM corresponds to a subsequence of the gene, and hence, a discrete depth coverage when run against the Shakya dataset. When adjusted for overlaps and hence data correlation, however, Frr had an effective sample size of around 19. RplB was around 34, RplK was around 14, and RpsT was around 6.

Variance within genes

We first took the Frr depth coverages for different genomes and sought to determine what canonical distribution it would best fit. Our candidates were the normal, gamma, negative binomial, and poisson distributions. To make our determination, we fit our data to each of the families using R's fitdistrplus library. Then, we made a qqplot of the data points against their quantiles from each of the fitted distributions. In the aggregate, we concluded that the negative binomial distribution produced qq-plots with the highest correlation coefficients, meaning that our data most closely approximates a negative binomial. This was consistent with the fact that the depth coverage is discrete, non-negative, and tends to have a higher variance than its mean.

Fitted with a negative binomial distribution, the depth coverage data for the Frr gene had, on average, across genomes, a qq-plot correlation coefficient of 0.982. For RplB it was 0.984, for RplK it was 0.977, and for RpsT it was 0.955; all reasonably high. The relatively smaller correlation coefficient in the case of RpsT can be attributed to its much smaller effective sample size of around 6.

Variance between genes

An interesting observation we made is that the variance for each gene was generally lower than the overall variance where all the data were pooled together. This led us to investigate further the group effect of each gene.

Location-wise, the four genes had a different spread along each individual genome. We suspected that gene location (we simply knew that they varied, not where they are) would influence the depth coverage data. We wanted to see for each genome, whether the depth coverage statistically varied across genes. We performed a one-way ANOVA analysis for each of the 25 genomes. For 23 of the genomes, the resulting p-value was below 0.05 and allowed us to reject, with this confidence level, the null hypothesis that the means were the same. This was suggestive that gene location does have an effect, in accordance with the group variances between lower than an overall variance.

Generalized Linear Mixed Effects Model

Since gene location is not constant, we decided to model the gene effect as a random effect, with a presupposed normality. The random effect accounts for the detectable variance across genes. A model for depth coverage must include, of course, a fixed effect representing the true genome abundance. The variances within genes was already determined above to follow a negative binomial distribution. So, we fit a generalized linear effects model, using R's glmer.nb function, with a random effect, a fixed effect, and negative binomial errors. The model is shown below:

$$y_{ij} \sim NegBin(\mu_i, \theta)$$
$$log(\mu_i) = \eta + \zeta_i, \quad \zeta_i \sim N(0, \sigma^2)$$

where y_{ij} is the jth coverage depth for the ith gene, coming from a negative binomial distribution with mean of μ_i and dispersion parameter (not equal to the variance) θ. A log-link function was used to relate the mean for each gene with the linear predictor. The linear predictor is an equation having a fixed effect η and a random effect ζ_i for each gene. The fixed effect η, also called the intercept of the equation, can be interpreted as an overall mean when converted from log-space. The variance σ^2 of the random effects, is the portion (additive?) of the total variance attributed to between-group variation. The remainder can be considered within-group variation.

Note that the dispersion parameter θ is global for each model. Negative binomial is not part of the exponential family for which GLMMs are intended. It becomes amenable to GLMM when the dispersion parameter is held fixed. So, the glmer.nb function, which is relatively experimental, first estimates the dispersion parameter and then fits the GLMM. The variance for each gene differs, since for a negative binomial it's a function of the dispersion parameter (global) and the mean (which is different by gene).

Since η is what we were originally looking for as the estimated overall mean (which estimates the true population size), we are interested also in the standard error of this estimator. Denote SE(η) as the standard error of η. Since η is in log space, then exp(η) is the estimated overall mean in response scale, and exp(η) - exp(η +SE(η)) would be the change in the response-scale mean at +1 SD in log space.

After fitting this model, we compared its AIC score with that of a non-generalized linear mixed effects model where the errors are normally distributed, and that of a regular linear model which holds the genes to be fixed effects. We found that the AIC score for the generalized linear mixed effects model was lowest among the three candidate models for all but 6 of the genomes, and within those 6, the AIC spread was marginal. So, we concluded that the GLMM is suitable for our data.

The results for this analysis are included in 5 tables stored in /cbcb/lab/mpop/mHMMs/bacterial_core_genes/vetted_perfect_models/updated_parent_models/yancy_Shakya_analyses/glmm_analysis
Results1.txt: includes the sample mean, sample standard deviation, and ANOVA p-value for each genome
Results2.txt: includes the 4 r values (correlation coefficient) for qqplots of each of the 4 genes' empirical quantiles versus the fitted negative binomial distribution's theoretical quantiles
Results3.txt: includes the estimated overall mean of the GLMM model when converted from log-space, also a converted form of the standard error, corresponding to the formula exp(η) - exp(η +SE(η)) (be careful when interpreting this… looking at the raw log-space values might be safer)
Results4.txt: includes the mean of each gene (incorporating random effects) when converted from log-space; this can be contrasted with the overall mean, also converted, in results3. Also includes the

GLMM-estimated standard deviation of the random effect, converted via the formula exp($\eta + \sigma$) - exp(η); be careful when interpreting this.
Results5.txt: includes the AIC scores for the GLMM model, the linear model, and the linear mixed effects model, plus a column saying whether the AIC score for the GLMM model is in fact the lowest of the 3

Autocorrelation and scale analysis