Supplementary Note

Robust methods for differential abundance analysis of marker gene survey data

Joseph Nathaniel Paulson[1,2], O. Colin Stine[3], Héctor Corrada Bravo[1,2,4*], & Mihai Pop[1,2,4*]

September 4, 2013

Contents

[1]Applied Mathematics and Scientific Computing, University of Maryland, College Park, MD
[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD
[3]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD
[4]Computer Science Department, University of Maryland, College Park, MD

1 Introduction

In this supplementary note we include the discussion of our simulation study, a detailed comparison of differential abundance methods on oral microbiome data from the Human Metagenomics Project and include a discussion on rarefaction and ambiguous read assignment to OTUs.

2 Zero-inflated Gaussian mixture-model

2.1 The model

Our zero-inflated Gaussian (ZIG) mixture model is motivated by the observed relationship between depth of coverage and the number of OTUs detected (Supplementary Fig. 3). In this section, we provide full details for our method (Supplementary Fig. 5).

Count data is modeled from two populations, each with n_A and n_B samples and with m features (OTUs). The raw count for sample j and feature i is denoted by c_{ij}. The class indicator function is defined as $k(j) = I\{j \in \text{group}A\}$.

The zero-inflated model is defined for the continuity-corrected $\log_2$ of the raw count data.

$$y_{ij} = \log_2(c_{ij} + 1)$$

as a mixture of a point mass at zero $I_{\{0\}}(y)$ and a count distribution $f_{count}(y; \mu, \sigma^2) \sim N(\mu, \sigma^2)$. Given mixture parameters π_j, we have that the density of the zero-inflated Gaussian distribution for feature i, in sample j with s_j total counts is

$$f_{zig}(y_{ij}; s_j, \beta, \mu_i, \sigma_i^2) = \pi_j(s_j) \cdot I_{\{0\}}(y_{ij}) + (1 - \pi_j(s_j)) \cdot f_{count}(y_{ij}; \mu_i, \sigma_i^2)$$

We specify the mean model as:

$$E(y_{ij}|k(j)) = \pi_j \cdot 0 + (1 - \pi_j) \cdot \left(b_{i0} + \eta_i \log_2(\frac{\hat{s_j^l} + 1}{N}) + b_{i1}k(j) \right).$$

In this case, parameter b_{i1} is an estimate of fold-change in mean normalized counts between the two populations. The term including the logged normalization factor $\log_2\left(\frac{\hat{s_j^l}}{N}\right)$ captures OTU-specific normalization factors through parameter η_j. This can capture feature specific biases, for instance in PCR amplification efficiency[31, 32, 33]. The model can also be specified *without* OTU-specific normalization, in which case the term including the normalization factor is treated as an offset in the linear model. This is equivalent to defining a model on logged *normalized* count data without including the normalization offset term in the linear model.

For large marker gene survey studies in clinical and epidemiological settings, it is essential to include possible sources of confounding error when testing the association between the abundance of taxonomic features and a clinical phenotype of interest (disease, for instance). Our linear model methodology can easily incorporate these confounding covariates in a straightforward manner. Other zero-inflated models have been developed mixing the Poisson and Binomial distributions. These models have had applications to ecological count data[34].

Based on the observation that the number of zero-valued features of a sample depends on its total number of counts, we model the mixture parameters $\pi_j(s_j)$ as a binomial process:

$$\log \frac{\pi_j}{1-\pi_j} = \beta_0 + \beta_1 \cdot \log(s_j)$$

The linear model of the binomial process above can also include covariates that capture variability in the sampling process as appropriate. Note however, that the detection model *does not* depend on class indicator function $k(j)$.

We highlight an example of the effect of the ZIG model on differential abundance in one OTU annotated as *Granulicatella para-adiacens* found in the Human Microbiome Project dataset (Supplementary Fig. 6).

2.2 Expectation-Maximization Algorithm

Denote the full set of estimates as $\theta_{ij} = \{\beta_0, \beta_1, b_{i0}, \eta_i, b_{i1}\}$. Maximum-likelihood estimates are approximated using the EM algorithm where we treat mixture membership $\Delta_{ij} = 1$ if y_{ij} is generated from the zero point mass as latent indicator variables. The log-likelihood in this extended model is then

$$l(\theta_{ij}; y_{ij}, s_j) = (1-\Delta_{ij}) \log f_{count}(y; \mu_i, \sigma_i^2) + \Delta_{ij} \log \pi_j(s_j) + (1-\Delta_{ij}) \log\{1-\pi_j(s_j)\}.$$

E-Step: Estimates responsibilities $z_{ij} = Pr(\Delta_{ij} = 1)$ given current estimates $\hat{\theta}_{ij}$ as

$$\hat{z}_{ij} = \frac{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij})}{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij}) + (1-\hat{\pi}_j) f_{count}(y_{ij}; \hat{\theta}_{ij})}$$

Notice $\hat{z}_{ij} = 0 \; \forall \; y_{ij} > 0$.

M-Step: Estimates parameters θ_{ij} given current estimates $\hat{z}_{ij}$:

To compute b, we use weighted least squares, with weights $1 - \hat{z}_{ij}$. Note that only samples with $y_{ij} = 0$ potentially have weights < 1. Estimates of standard error are also obtained using $1 - \hat{z}_{ij}$ as weights. The mixture parameter is estimated as $\hat{\pi}_j = \sum_{i=1}^{G} \hat{z}_{ij}/G$, from which we estimate β, using least squares on the logit model as:

$$\log \frac{\hat{\pi}_j}{1-\hat{\pi}_j} = \beta_0 + \beta_1 \log(s_j).$$

From the estimated fold-change (b_{1i}) and its standard error, we construct a moderated t-statistic by Empirical Bayes [25] and use a parametric t-distribution to obtain p-values for the test $b_{1i} = 0$. Notice that this only incorporates the count component of the zero-inflated mixture model. We interpret this test as the expected difference in abundance between groups conditioned on feature detection.

The moderated t-statistic is defined as $t_i = \frac{b_{1i}}{(\tilde{s}_i^2 / \sum_j (1-z_{ij}))^{1/2}}$, where $\tilde{s}_i^2$ is obtained by pooling all features' variances as described in [25], $\tilde{s}_i^2 = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}$ where s_i^2 and d_i are respectively the observed feature variance and degrees of freedom and d_0 and s_0^2 are estimated using the method of moments incorporating all feature variances and degrees of freedom. We found that by using a log-Normal distribution, the moderated t-test was appropriate. As in the previous Metastats version, we use the q-value method [35] to correct for multiple testing.

We chose to use a log-normal distribution in the count component of the mixture instead of a generalized linear model, *e.g.* negative binomial [18,19], for both computational and statistical reasons. On the computational side, we would need to estimate a weighted generalized linear model using an iterative method at each maximization step. That is too computationally intense and numerically unstable. On the statistical side, we find that the log-normal distribution is appropriate since the type of marker gene survey study we are targeting tend to have moderate to large sample sizes (Supplementary Fig. 2). This is consistent with recent observations in the literature [36].

3 Comparison of differential abundance detection methods

We compared Metastats[13], Lefse[14], DESeq[18] and edgeR[19] along with metagenomeSeq using oral microbiota data from the Human Microbiome Project[24] to identify differentially abundant OTUs in tongue and subgingival plaque samples. Metastats and edgeR declared the largest number of OTUs to be significant (533 and 524, respectively), while metagenomeSeq (360) and, especially, DESeq (20) and Lefse (8) declared fewer significant OTUs (Supplementary Table 2). We illustrate the range in depth of coverage that metagenomeSeq takes into account (Supplementary Fig. 8).

Overall, metagenomeSeq and DESeq showed high agreement in fold-change estimates (Supplementary Fig. 9A). Specifically, metagenomeSeq and DESeq fold-change estimates were very similar on features exhibiting low sparsity, but DESeq fails to declare as significant the majority of dense features declared significant by metagenomeSeq (only 20 of 244). We found that the mean-dispersion trend estimated by DESeq for this dataset was uncharacteristic of those estimated from RNAseq data (Supplementary Fig. 9B) and DESeq consistently overestimated dispersion for dense features resulting in a large number of failed discoveries (*e.g.* the feature shown in Supplementary Fig. 9C is not declared as differentially abundant by DESeq). Features with high sparsity drive the poor dispersion estimate in DESeq and are also features where fold-change estimates between metagenomeSeq and DESeq disagreed (*e.g.* Supplementary Fig. 9D). By controlling for low sequencing depth, metagenomeSeq is able to detect these population differences appropriately.

The edgeR method consistently estimated larger fold-changes in comparison with both metagenomeSeq (Supplementary Fig. 10A) and DESeq (Supplementary Fig. 10B). The edgeR method includes total sample counts as a term in a generalized linear model, while our model includes the CSS normalization term in the log-normal linear model of the count distribution. Artifacts arising from normalization using total counts lead to many false differential abundance predictions made by edgeR (*e.g.* Supplementary Fig. 10C), which are avoided by using our proposed normalization method. The dispersion trend estimate in edgeR is also uncharacteristic for this dataset (Supplementary Fig. 10D) and the deviation is again driven by feature sparsity.

Metastats was more consistent with edgeR (Supplementary Fig. 11A) than metagenomeSeq (Supplementary Fig. 11B) or DESeq (Supplementary Fig. 11C) due to its use of total normalization. Even though we are testing for overall differential abundance between oral microbiota, regardless of sequencing site, Metastats consistently called features as significant where differences are specific to sequencing site (*e.g.* samples sequenced in JCVI as shown in Supplementary Fig. 12D). Large survey studies may obtain samples from a variety of locations, over heterogeneous populations. Analysis methods used in these studies require the ability to interpret differential abundance taking into account the heterogeneity of these populations. Both RNAseq methods and metagenomeSeq use linear models that can include possible confounding sources of variability, in this case sequencing site or gender, to aid interpretation in differential abundance testing. metagenomeSeq and both RNAseq methods are able to detect site-specific differential abundance between microbiota using an interaction model. Metastats was not designed to carry out this kind of test. Similarily, Lefse uses an ad-hoc heuristic approach to account for subpopulations in large marker studies that is overly conservative and prone to low sensitivity.

4 Ambiguous read assignment to OTUs

Ambiguous read assignment is an important consideration in testing differential abundance in count data, and in particular RNAseq[37, 38, 39, 40]. There are two sources for ambiguous read assignment in RNAseq data that may apply to marker gene survey data. In isoform ambiguity, where a sequence read could be generated from sequencing one of multiple possible isoforms for a gene. In this case gene-level abundance measurements are convolutions of isoform-level abundances which may bias differential abundance inferences in the presence of differential abundance at the isoform level. In marker gene survey data, reads are clustered based on sequence similarity and the resulting clusters define features over which differential abundance testing is performed. The type of convolution occurring in RNAseq data would occur when an OTU defined by clustering contains two distinct functional OTUs. We have chosen a sequence similarity threshold (99%) that was previously shown to be more stringent than similarity at species-level[20] thus reducing the possibility of convolution to occur. We therefore believe that this type of ambiguity does not arise frequently in this setting. On the other hand, less stringent sequence similarity thresholds, which would increase the frequency of convolution, still exhibit high sparsity as previously reported[41, 42]. We believe sparsity does indeed drive the improvement in results we see as methods using non-zero inflated negative binomial models, including Cuffdiff2, are not suitable.

However, there is a second source of ambiguity in RNAseq data that can occur in marker gene survey data. In RNASeq analysis there is ambiguity for some reads with multiple potential 'optimal' mappings along the genome, so called 'multi-mapped' reads. This analogously occurs in marker gene survey data in assigning reads to OTU clusters and subsequently the count observed for given OTUs. In this case, reads may be assigned to more than one cluster if it is within the given similarity threshold for more than one cluster representative sequence. The default option in our pipeline, based on DNAClust, does not guarantee that a sequence can be uniquely placed. Reads are assigned to a particular cluster by choosing the best alignment and largest OTU representative center for a given set of clusters and can have more than one possible placement. However, there an option in DNAClust, 'non-overlapping', that results in less ambiguously assigned reads by restricting reads that are within the radius of two or more clusters to not get assigned. This is similar to a commonly used approach in RNAseq analysis of discarding multi-mapped reads.

To test the effect of a potentially ambiguous read mapping to a cluster we re-ran DNAClust with the 'non-overlapping' option on the full HMP dataset to compare rarefaction and sparsity results observed earlier. The rarefaction effect (association between depth of coverage and the number of detected features) and sparsity is essentially unchanged (Supplementary Fig. 14). The least sparse sample after filtering OTUs (less than 5 positive samples or reads present) and samples ($<$1,000 total counts or $>$ 35,000) in the 'non-overlapping' run is 97.48% non-positive while we observed 97.46% with default DNAclust options.

We subset the data to the same subgingival and tongue samples and trimmed OTUs (less than 5 positive samples) as previously performed for the subgingival and tongue analysis. We observed 23,275 OTUs, a total of 410 fewer OTUs. We reran DESeq and the zero-inflated Gaussian mixture model on this less ambiguous dataset and compared fold-change estimates between DESeq and the zero-Inflated Gaussian mixture (Supplementary Fig. 10A). We observed the same phenomena that ZIG is adjusting fold-changes on sparse OTUs as previously described. Also, we observed that the

over dispersion estimates are similar to our previous run (Supplementary Fig. 15).

5 Discussion

Rarefaction is a common phenomenon in molecular surveys of bacterial communities[43], where the number of taxonomic features detected in a sample depends on the amount of sequencing performed. This large variation in the number of taxonomic features detected in each sample, contributes to the inherent sparsity of metagenomic data where most features are only found in a few samples, as previously reported[20,21,22].

Recent publications have relied on machine learning techniques, such as random forests, to identify microbiota signatures correlated with phenotypic observations[41, 42]. Our work targets a complementary task the feature-by-feature assessment of differential abundance based on an appropriately defined linear model that accounts for specific microbiota features and confounding factors. The methods developed here, in particular the ZIG mixture model, can be incorporated into machine-learning based predictive models that seek to identify multiple features for specific phenotypes.

Other types of analyses adversely affected by high sparsity and sampling bias include clustering and co-occurrence network discovery. While the normalization approach presented here can help control biases in analyses based on simple correlation measures, methods developed to specifically discover significant correlations between sparse features in marker gene survey data are better suited for the task[20,21].

References

[31] Gonzalez, J. M., Portillo, M. C., Belda-Ferre, P. & Mira, A. Amplification by pcr artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS ONE* **7**, e29973 (2012).

[32] Wu, J.-Y. *et al.* Effects of polymerase, template dilution and cycle number on pcr based 16 s rrna diversity analysis using the deep sequencing method. *BMC Microbiology* **10**, 255 (2010).

[33] Von Wintzingerode, F., Gbel, U. B. & Stackebrandt, E. Determination of microbial diversity in environmental samples: pitfalls of pcr-based rrna analysis. *FEMS Microbiology Reviews* **21**, 213–229 (1997).

[34] Hall, D. B. Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030–1039 (2000). URL `http://www.ncbi.nlm.nih.gov/pubmed/11129458`.

[35] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

[36] Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics* **14**, 91 (2013).

[37] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).

[38] Wang, X., Wu, Z. & Zhang, X. Isoform abundance inference provides a more accurate estimation of gene expression levels in rna-seq. *Journal of Bioinformatics and Computational Biology* **8 Suppl 1**, 177–192 (2010).

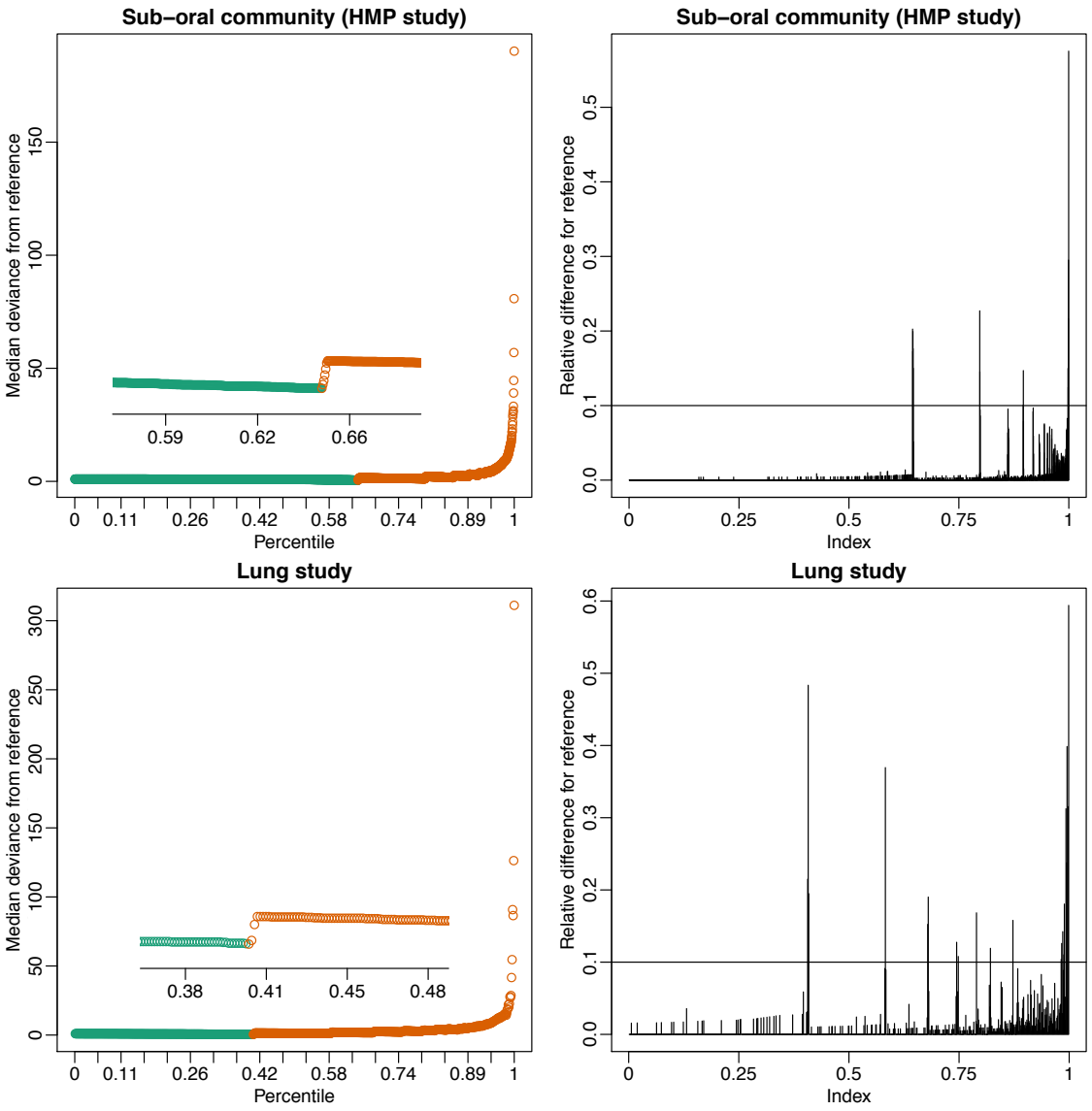
[39] Salzman, J., Jiang, H. & Wong, W. H. Statistical modeling of rna-seq data. *Statistical Science* **26**, 62–83 (2011).

[40] Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature Biotechnology* 4653 (2012).

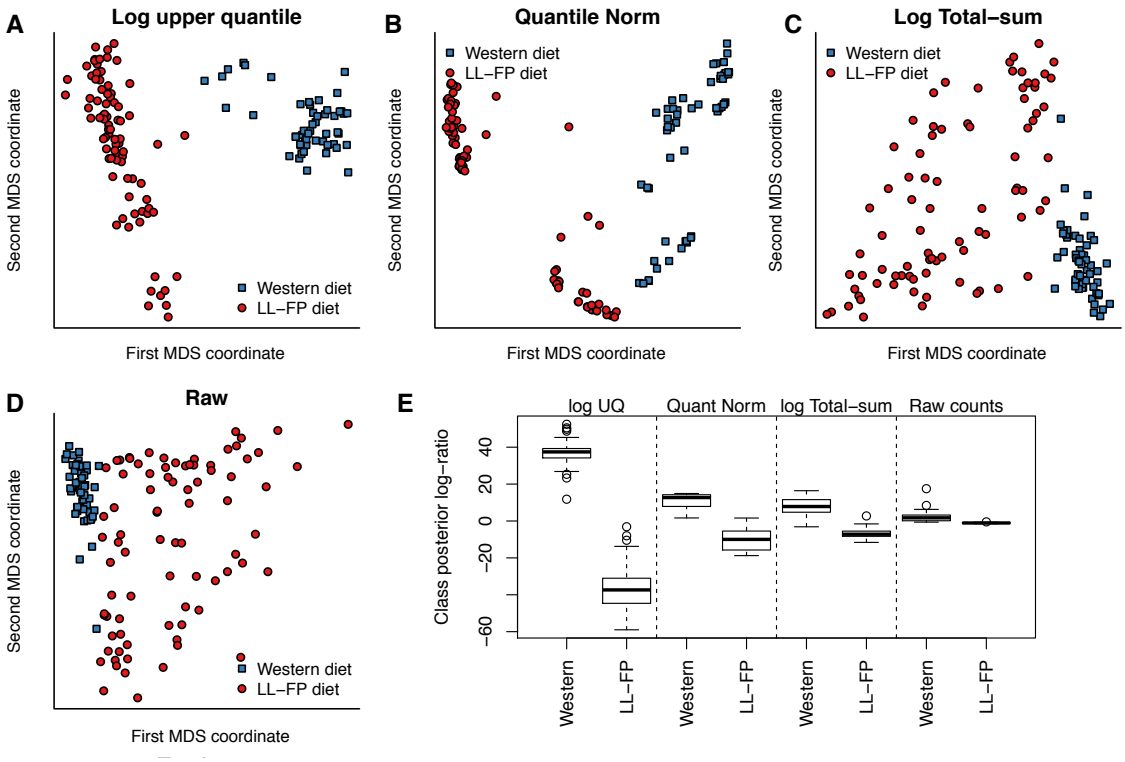
[41] Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C. & Knight, R. Human-associated microbial signatures: examining their predictive value. *Cell host & microbe* **10**, 292–296 (2011).

[42] Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

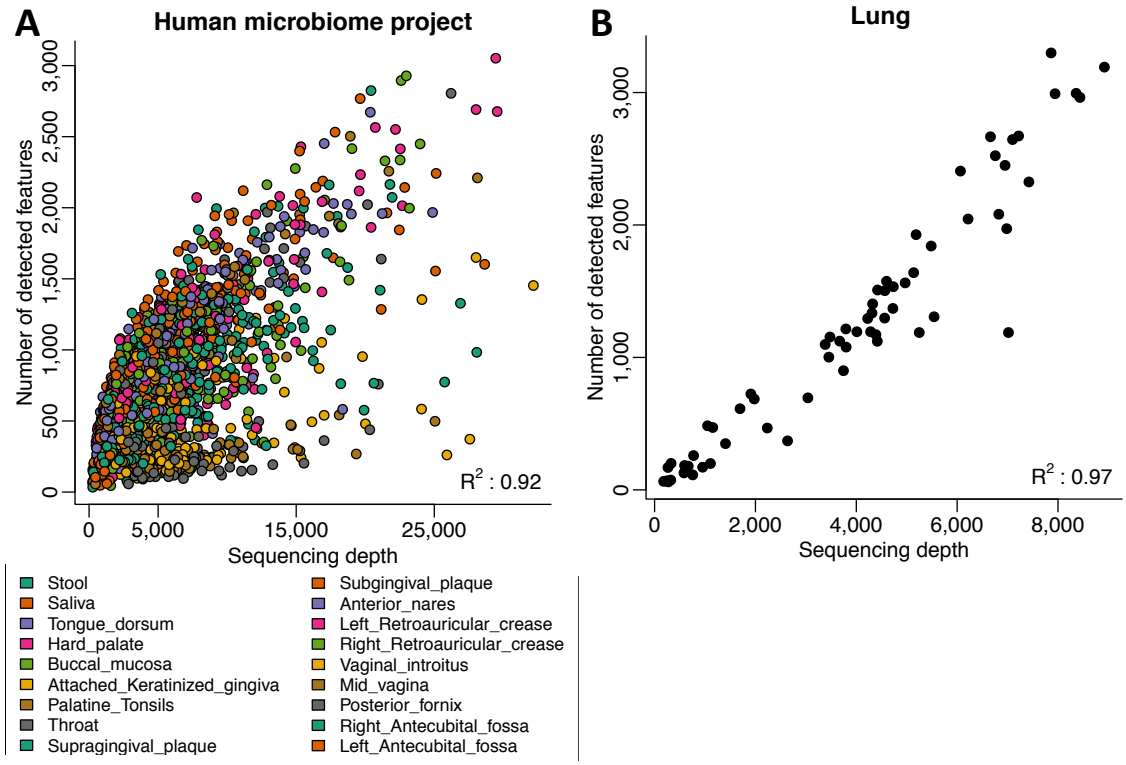
[43] Hughes, J. B. & Hellmann, J. J. The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods in Enzymology* **397**, 292–308 (2005). URL `http://www.ncbi.nlm.nih.gov/pubmed/16260298`.



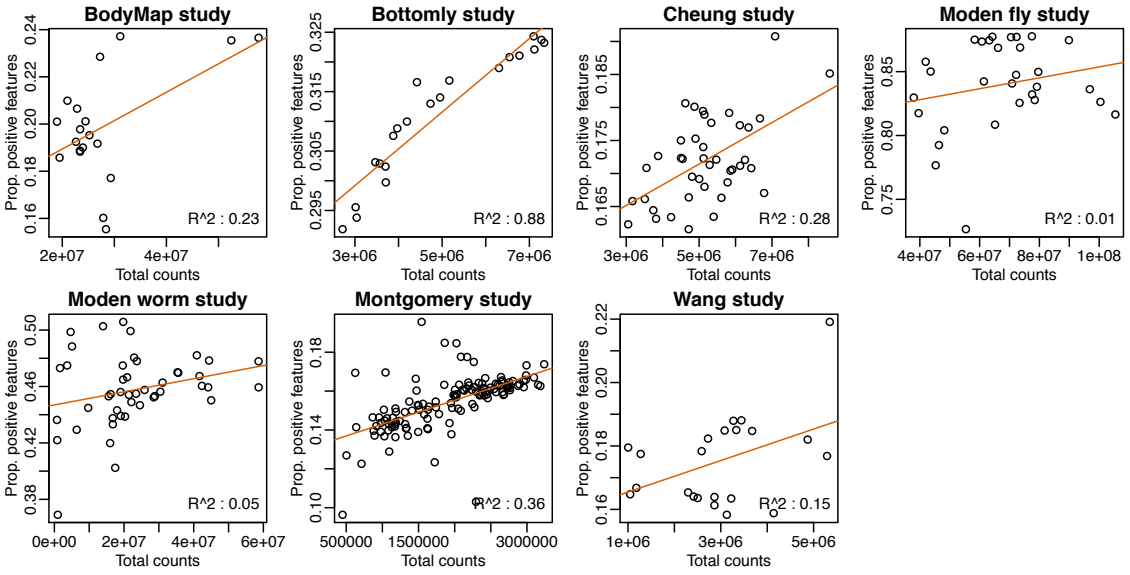
Supplementary Figure 1. Data-driven adaptive method for selecting normalization scale quantile. Left, we plot d_l (see Online Methods) for our oral sub-community and lung microbiome datasets. In the oral sub-community (HMP) dataset we observe sample count distributions differ greatly from the reference at the 65th percentile where relative difference of the median deviation is greater than 10%. For the lung microbiome dataset, this occurs at the 41st percentile. Right, relative difference of the median deviation of sorted sample counts from reference for the oral sub-community and lung microbiome datasets. We observe that the raw sample counts follow similar distributions up to the 65th and 41st percentile respectively. The reference is calculated as the row means of raw sorted counts.



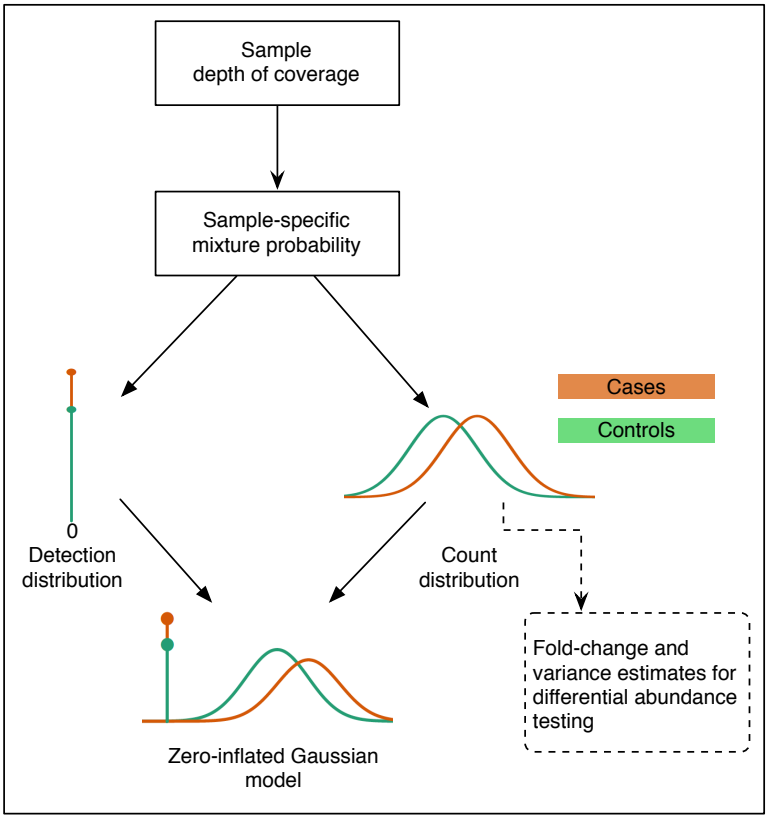
Supplementary Figure 2. Effect of normalization on clustering analysis. We plot the first two principal coordinates in a multi-dimensional scaling analysis components for count data normalized by **(A)** logged upper quantile, **(B)** quantile normalization, **(C)** logged total-sum scaling, and **(D)** the raw counts. Orange points represent samples on the LF/PP diet and green the Western diet. **(E)** Class posterior probability log-ratio for Western diet obtained from linear discriminant analysis (LDA). Each box in the plot corresponds to the distribution of leave-one-out posterior probability of assignment to the "Western" cluster for samples of each type across normalization methods. Clustering analysis is improved significantly by CSS normalization and Logged upper quantile scaling.



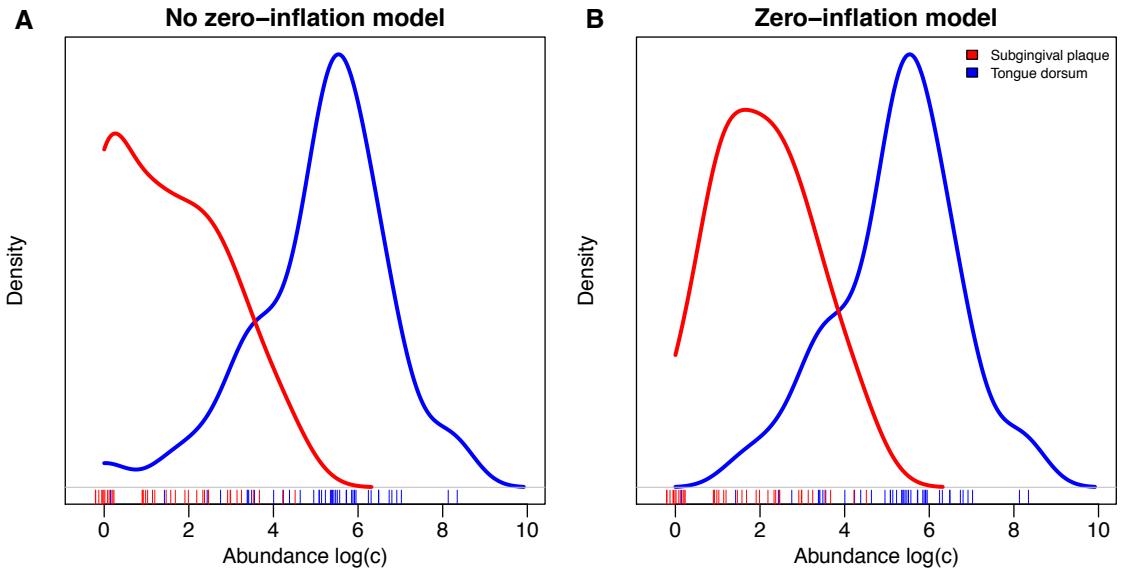
Supplementary Figure 3. The number of OTUs detected in a sample depends on sequencing depth and phenotypic characteristics. We plot the number of detected OTUs in a sample as a function of sequencing depth for the Human Microbiome Project **(A)** and the lung microbiota study **(B)**. There is a strong dependency between sequencing depth and number of detected OTUs. Neither dataset shows that the number of OTUs stabilizes for samples with high depth, indicating that in both cases sequencing depth may not be sufficient for comprehensive profiling of the microbial community. We found that a large proportion of variability in the number of OTUs detected is explained by sequencing depth. We found that including clinical covariates, *e.g.* body site sampled, we obtained greatly improved fits with higher adjusted R2. The same dependency is observed in all studies analyzed for this paper.



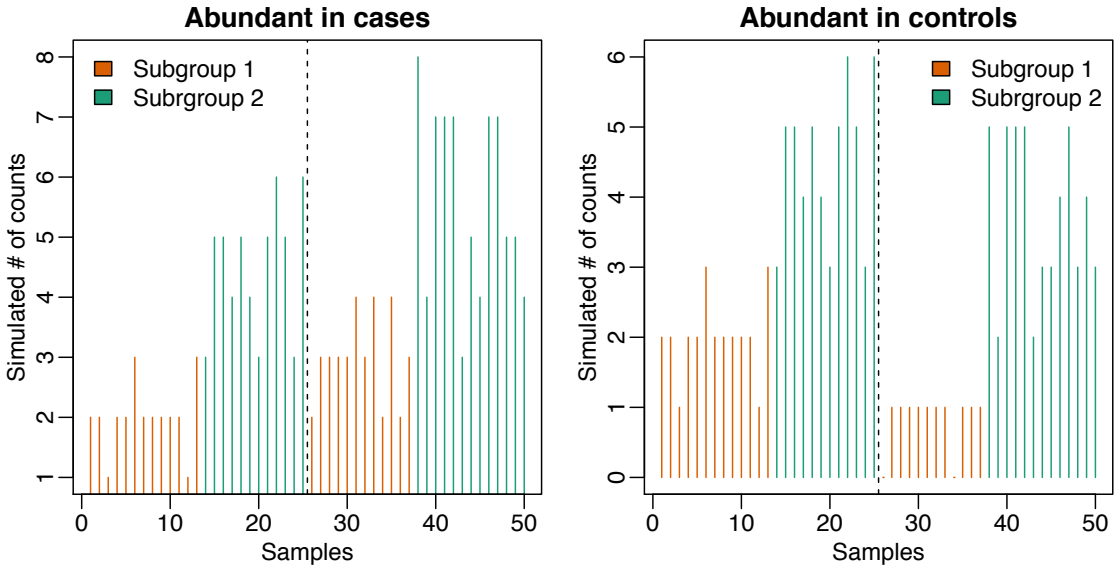
Supplementary Figure 4. Effect of sequencing depth on the number of genes detected in RNAseq. We plot the number of genes detected in a RNAseq sample as a function of sequencing depth. In the lower corner of each plot is the adjusted R^2 value representing how much of the variation in each sample's detected genes is described by the depth of coverage. The proportion of genes detected in any particular sample is much higher in RNAseq datasets (15-85%) compared to marker gene survey data samples (1-3%). Depths of coverage are also much larger in RNAseq. Datasets obtained from Recount[23] at: http://bowtie-bio.sourceforge.net/recount/



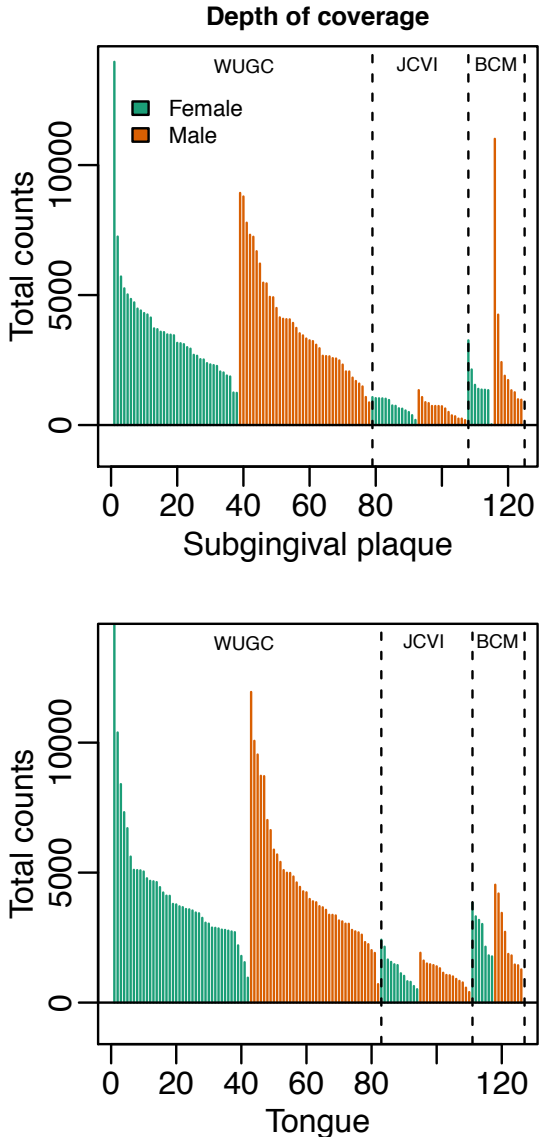
Supplementary Figure 5. The zero-inflated Gaussian mixture model. A graphical representation of the zero-inflated Gaussian mixture model. Counts are modeled with a mixture distribution with components corresponding to normally-distributed log-abundances in each group, *e.g.* cases (orange) and controls (green), and a spike-mass at zero corresponding to a detection distribution that depends on each sample's sequencing depth. We use fold-change and variance estimates derived from the count component of the mixture to test for differential abundance between groups.



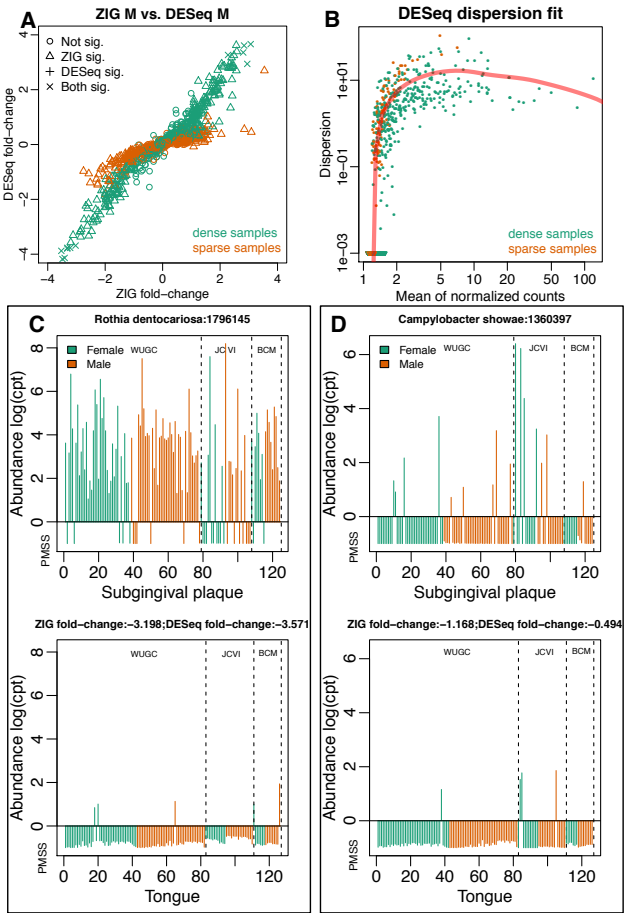
Supplementary Figure 6. Illustration of the effect of zero-inflated Gaussian mixture model on differential abundance. (A) Kernel density estimates of log-counts for an *Granulicatella para-adiacens* OTU in samples from the Human Microbiome Project. We see that subgingival plaque samples have a large number of zero counts that result in large differential abundance estimate when a model without zero-inflation is used. **(B)** Weighted kernel density estimates for the same samples. Weights are obtained from the posterior probabilities for zero counts due to under-sampling of the microbial community. After accounting for zero-inflation, the differential abundance estimate is moderated for this feature. This plot also suggests that the log-normal distributional assumption used in this paper is supported.



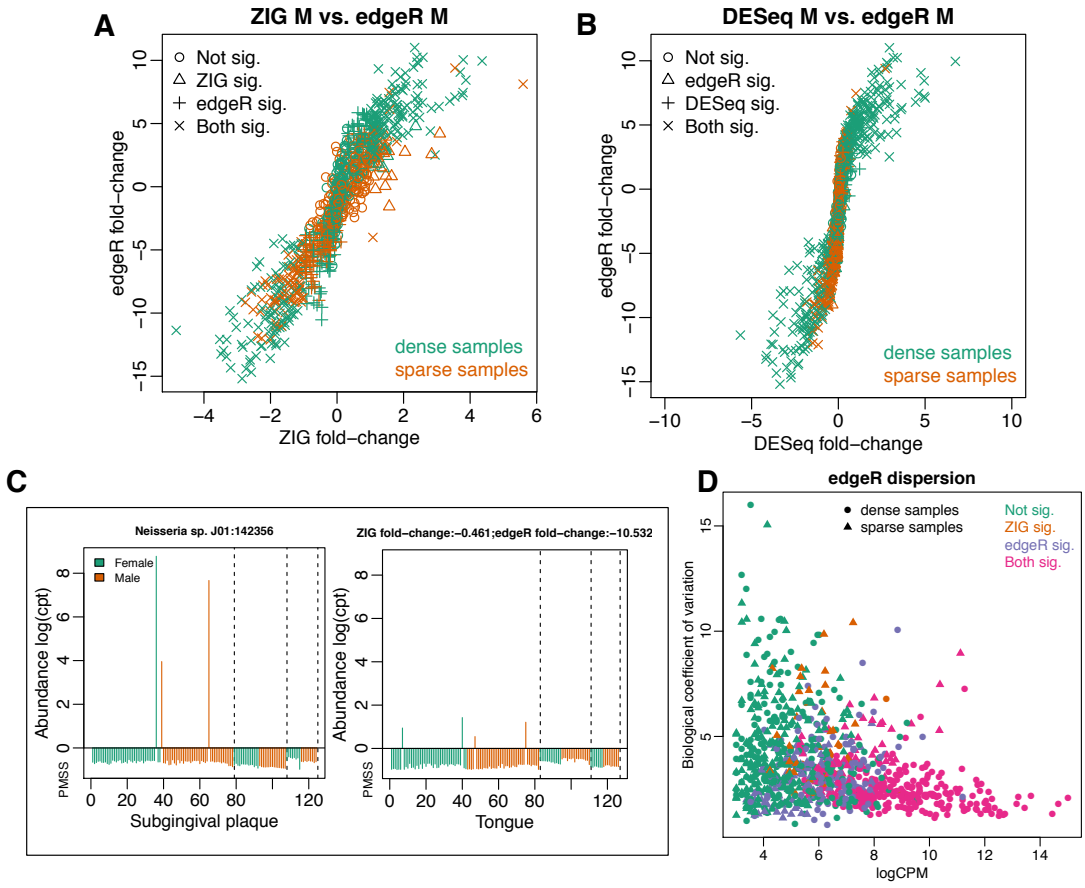
Supplementary Figure 7. Abundance of two simulated features from the subgroup simulation. Plot of simulated counts for two OTUs. Samples are ordered by case control and subgroup population. The dotted line separates the two populations, controls and cases and the colors for each line represents the subgroup. Notice the trend from controls to cases is consistent across subgroups, but Lefse is confounded by the cross population subgroups.



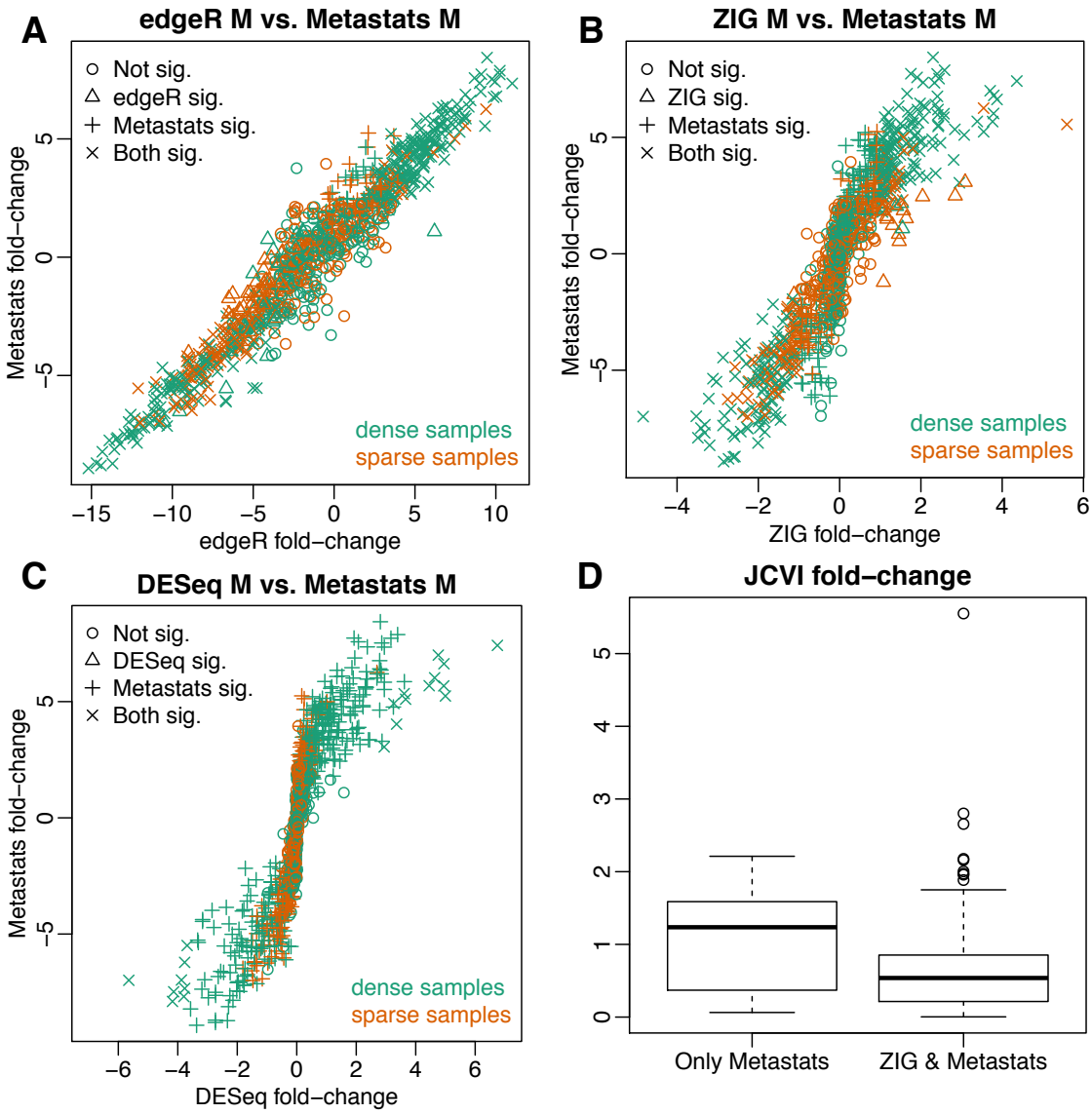
Supplementary Figure 8. Sample sequencing depth for Oral Sub-community Samples. Samples are ordered by sequencing center (WUGC, JCVI, BCM), sex (female, male), and raw depth of coverage. The top graphs represent subgingival plaque samples, the bottom represent tongue samples.



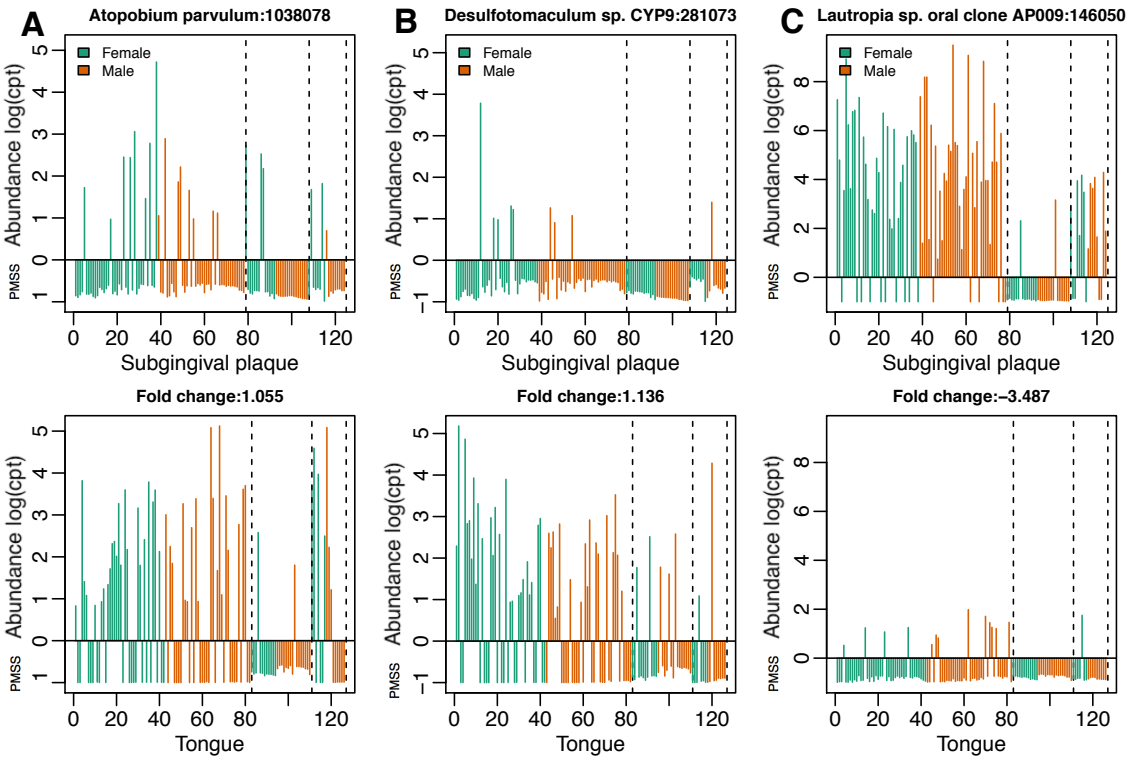
Supplementary Figure 9. Comparison of metagenomeSeq differential abundance detection to detection by DESeq. (A) Fold-change estimates for metagenomeSeq and DESeq. We see that metagenomeSeq and DESeq agree in these estimates for dense features (green), while for sparse features (orange) metagenomeSeq adjusts fold-changes weighting zeros according to sample depth of coverage. **(B)** Estimate of dispersion as a function of mean feature counts by DESeq. Sparse features, indicated in orange, display high dispersion relative to dense features (green). This makes DESeq overestimate dispersion in general resulting in a large number of missed discoveries. **(C)** A feature where dispersion overestimates by DESeq lead to a false negative call. Each panel plots the CSS normalized counts with samples ordered by sequencing center (WUGC, JCVI, BCM), sex (female, male), and depth of coverage. The top panel shows subgingival plaque samples, the bottom panel shows tongue samples. The bottom strip in each panel indicates the posterior probability estimates for zeros due to community sub-sampling (PMSS). Depth of coverage for this dataset is given in Supplementary Fig. 8. **(D)** A sparse feature where weighting of zero counts by metagenomeSeq results in a differential abundance call missed by DESeq. Panels follow convention in sub-figure **(C)**.



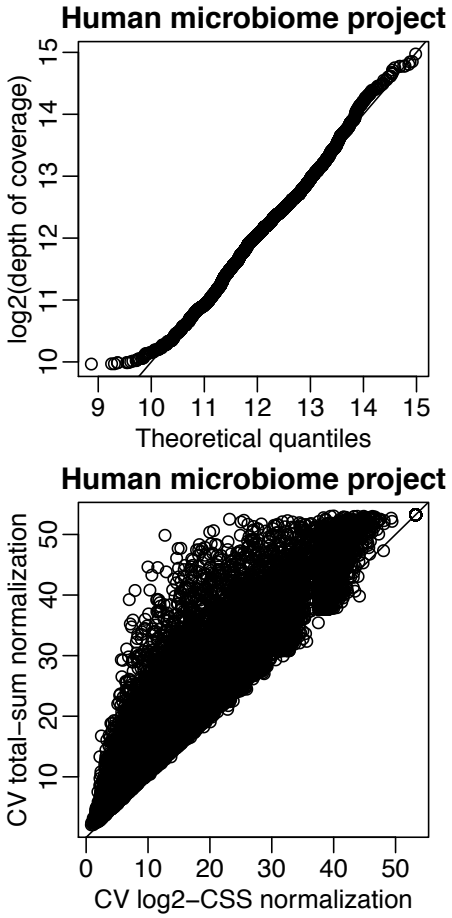
Supplementary Figure 10. Comparison of edgeR differential abundance detection to detection in metagenomeSeq and DESeq. **(A)** Fold-change estimates for metagenomeSeq and edgeR. We see that fold-changes are not consistent across features with edgeR significantly overestimating many fold changes. **(B)** Fold-change estimates for DESeq and edgeR. Again, edgeR overestimates many fold-changes. **(C)** Feature not considered significant according to metagenomeSeq, but edgeR normalization using total counts results in a significant call. Panels follow convention in Supplementary Fig. 9C. Counts plotted are CSS normalized counts. **(D)** Dispersion as estimated by edgeR. Dispersion does not follow typical RNAseq experiments. Features declared significant by metagenomeSeq, but not edgeR (orange) have high variability estimates in edgeR and tend to be sparse features (triangles). metagenomeSeq and edgeR agree on abundant dense features (magenta circles). Features declared significant by edgeR but not metagenomeSeq (purple) have moderate abundance driven by few high-count features resulting from normalization artifacts.



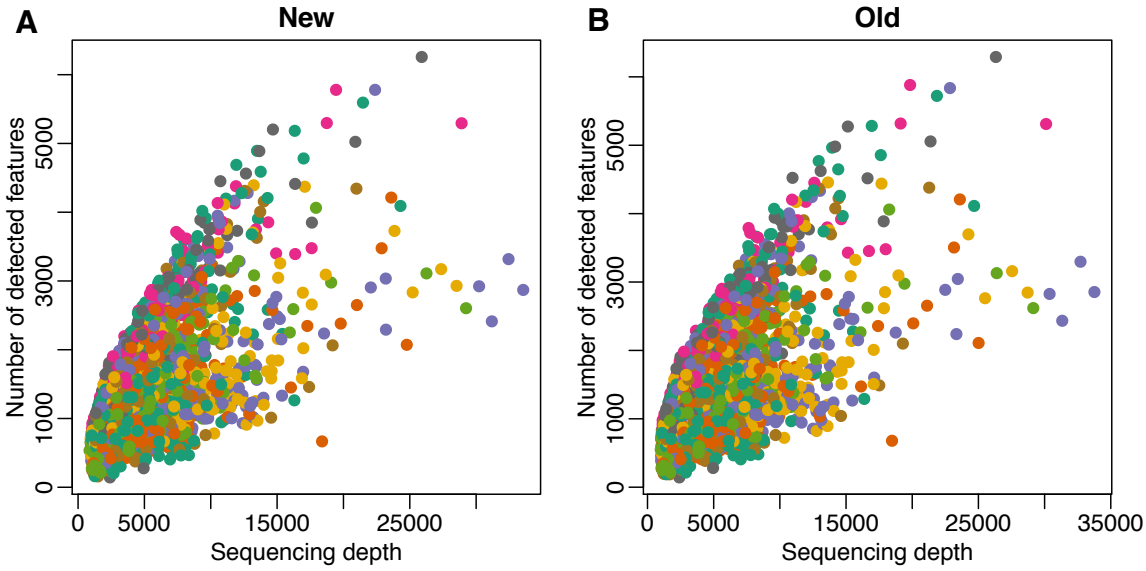
Supplementary Figure 11. Comparison of Metastats differential abundance detection to detection in metagenomeSeq, DESeq and edgeR. (A) Fold-change estimates for Metastats and edgeR showing high consistency between the two methods. **(B)** Fold-change estimates for metagenomeSeq and Metastats where Metastats consistently over-estimates fold-changes especially for sparse features (orange). **(C)** Fold-change estimates for DESeq and Metastats, where again Metastats overestimates fold-changes for sparse features (orange). **(D)** Sequencing site fold-change estimates from metagenomeSeq for features detected as differentially abundant by the original Metastats, but not by metagenomeSeq. Many differential abundance discoveries in Metastats are solely due to site specific effects.



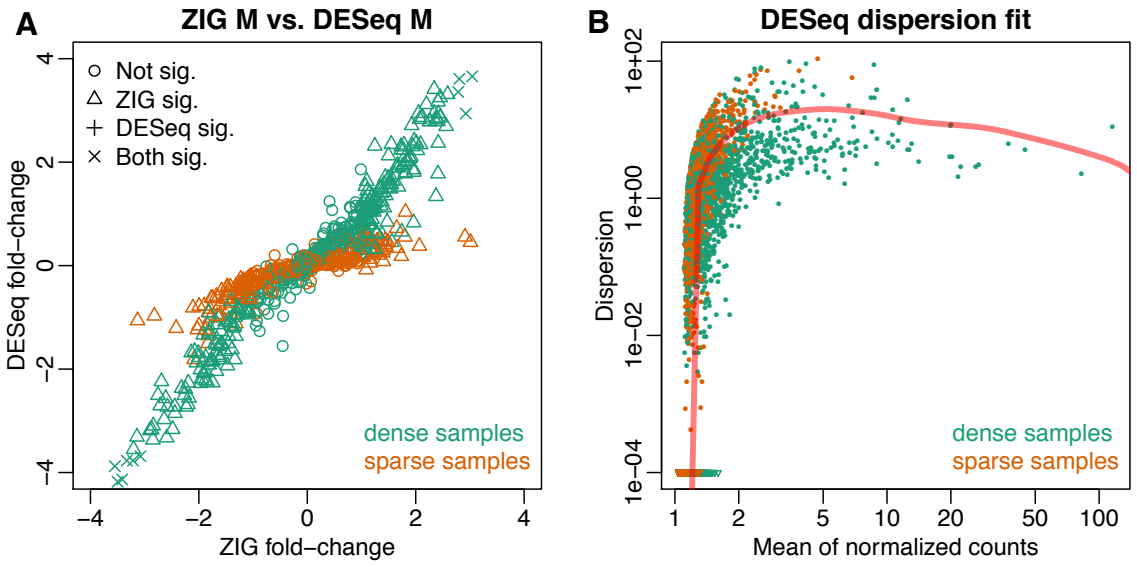
Supplementary Figure 12. Novel species detected as differentially abundant in tongue and subgingival microbiomes. Plot of CSS normalized counts for three OTUs. Samples are ordered by sequencing center (WUGC, JCVI, BCM), sex (female, male), and depth of coverage. Underneath each panel are probabilities of missing sequences due to subsampling (PMSS). The top graphs represent subgingival plaque samples, the bottom represent tongue samples. Depth of coverage of the samples is given in Supplementary Fig. 8 for comparison.



Supplementary Figure 13. Depth of coverage follows a log-Normal distribution and cumulative sum scaling normalization controls dispersion. Top, quantile-quantile plot of sample sequencing depth and a log-Normal distribution. We see that raw sample depth of coverage closely follows a log-Normal distribution. Bottom, coefficient of variation for total-sum normalized counts versus CSS normalized counts. We observe that dispersion is always greater in total-sum normalized counts.



Supplementary Figure 14. Effect of unambiguously placing reads in OTU centers on rarefaction. Plot of the same samples for the 99% 'perfect clusters' **(A)** and 99% 'exact clusters' **(B)** run through DNAclust after trimming OTUs to be positive in at least five samples and removing outlying samples. Notice that rarefaction and sparsity is not affected by the DNAclust option.



Supplementary Figure 15. Effect of unambiguously placing reads in OTU centers on differential abundance. Plots are as Supplementary Fig. 9A-B but were plotted using OTUs created running the 'perfect clustering' option in DNAclust. **(A)** Fold-change estimates for metagenomeSeq and DESeq. We see that metagenomeSeq and DESeq agree in these estimates for dense features (green), while for sparse features (orange) metagenomeSeq adjusts fold-changes weighting zeros according to sample depth of coverage. **(B)** Estimate of dispersion as a function of mean feature counts by DESeq. Sparse features, indicated in orange, display high dispersion relative to dense features (green). This makes DESeq overestimate dispersion in general resulting in a large number of missed discoveries.

	Zero-Inflated Gaussian Model (ZIG)				
TRUTH	Called DA	Not Called DA	Total	**Sensitivity**	**Specificity**
Differentially Abundant (DA)	2,376	124	2,500	**0.950**	**0.962**
Not DA	1,802	45,698	47,500		
Total	4,178	45,822			

	Lefse				
	Called DA	Not Called DA	Total	**Sensitivity**	**Specificity**
DA	24	2,476	2,500	**0.010**	**1.000**
Not DA	0	47,500	47,500		
Total	24	49,976			

Supplementary Table 1: Comparison of metagenomeSeq with Lefse - subpopulation simulation. We simulated data from two populations where each population consisted of two subpopulations representing a case-control study where cases and controls were collected from multiple sites. Using linear modeling, ZIG was more sensitive than Lefse, while retaining specificity in settings where groups tested include confounding subpopulations.

	ZIG	DESeq **A**	DESeq **B**	DESeq **C**	edgeR **A**	edgeR **B**	edgeR **C**	Metastats **A**	Metastats **B**	Metastats **C**	Lefse **A**	Lefse **B**	Lefse **C**
More abundant in plaque	186	12	0	0	145	0	57	168	0	125	6	0	0
More abundant in tongue	174	0	8	0	1	172	149	0	169	71	0	2	0
Not differentially abundant	607	172	165	607	40	2	401	18	5	411	180	172	607
Total number of differentially abundant features	360	20			524			533			8		

Supplementary Table 2: Comparison of DESeq, edgeR, original metastats, and Lefse with ZIG. DESeq finds only 20 and interesting features with our criteria, FDR<0.05 and fold-change of at least 1. edgeR and the original Metastats call many more, 524 and 533 features respectively, as differentially abundant. Many of these are biased by a small number of non-zero, large valued counts. Lefse only declared 8 features significant due to heuristic false positive controls.

A: Intersection of features that ZIG finds are differentially abundant in the subgingival plaque
B: Intersection of features that ZIG finds are differentially abundant in the tongue
C: Intersection of those that ZIG does not declare as differentially abundant