DIPP 3 Cities - Analysis of Technical Replicates

Lexi Ardissone

June 29, 2016

Introduction

In this document samples that were replicated across extractions, PCR-amplicon generation, sequencing run and lane were evaluated. The goal of this Rmd is 2-fold:

1. Assess microbial variance due to technical effects and hence determine if there is considerable batch-effect.
2. Estimate the limit of quantification for measuring relative abundance of bacterial taxa through 16S rRNA sequencing using Illumina HiSeq 2000 platform.

Description of Technical Replicates

Load data

```
#load phyloseq object containing all D3C samples
load('~/Desktop/D3C/D3C_2.0/1_inputs/input_files/dipp_6-20-16.Rdata')
dipp
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 50649 taxa and 2095 samples ]
## sample_data() Sample Data:       [ 2095 samples by 103 sample variables ]
## tax_table()   Taxonomy Table:    [ 50649 taxa by 8 taxonomic ranks ]
```

```
#Select TechReps samples
dipp.reps = subset_samples(dipp, TechReps == 'TRUE')
#remove reps due to reseuencing (at least 1 rep is 'defective...that's why it was resequenced)
dipp.reps = subset_samples(dipp.reps, TechRep_type != 'reseq')
#remove reps with <20,000 reads (being consistent with previous sample filtering parameters)
dipp.reps2 = prune_samples(sample_sums(dipp.reps) >= 20000, dipp.reps)
#extract metadata
reps.meta = data.frame(sample_data(dipp.reps2))
```

A total of 53 unique samples from 6 subjects were replicated giving a total of 301 sequenced replicates.

The number of replicates for each replicate type, sequencing run, and lane and the number of subjects these replicates correspond to is as follows:

```
##      site TechRep_type HS_run illumina_lane Reps nsamples nsubjects
## 1 Tampere         lane      3           ALL   12        3         3
## 2   Turku   extraction      2             1   30       30         3
## 3   Turku         lane      1             1   49       49         3
## 4   Turku         lane      1             2   50       50         3
```

```
## 5   Turku         lane      1             3   50       50         3
## 6   Turku         lane      1             4   50       50         3
## 7   Turku          pcr      2             1   30       30         3
## 8   Turku          run      2             1   30       30         3
```

Below is a list of the relevant `TechRep_type` comparisons:

- when `TechRep_type == lane`:
  - In order to assess lane to lane variability, compare `TechRep_type==lane` (4) for each `sample_id`.
  - all reps sequenced of HiSeq run 1 across 4 different lanes
  - all reps were from the same DNA extract and PCR reaction (same barcode)
  - e.g. `illumina_id` for 4 lane reps: `HS1_L1_B009`, `HS1_L2_B009`, `HS1_L3_B009`, `HS1_L4_B009`
  - Expectation: some lane variability, but overall very similar
- when `TechRep_type == run`:
  - In order to assess HiSeq run to HiSeq run variability, compare `TechRep_type==run` (1) to `TechRep_type==lane` (4) for each `sample_id`
  - all `run` reps were sequenced on HiSeq run 2, lane 1 (compared to HiSeq run 1 for `lane` reps)
  - all `run` reps were from the same DNA extract and PCR reaction (same barcode) as the `lane` reps
  - e.g. `illumina_id`: 1 run rep `HS2_L1_B009` vs. 4 lane reps `HS1_L1_B009`, `HS1_L2_B009`, `HS1_L3_B009`, `HS1_L4_B009`
  - Expectation: more variability between run and lane reps than within lane reps
- when `TechRep_type == pcr`:
  - In order to assess variability due to different PCR reactions, compare `TechRep_type==pcr` (1) to `TechRep_type==run` for each `sample_id`
  - all `pcr` reps were sequenced on HiSeq run 2, lane 1 (same as all `run` reps)
  - all `pcr` reps were from different PCR reactions as `run` reps but from the same DNA extract (eliminates any lane or extraction effects)
  - e.g. `illumina_id`: 1 pcr rep `HS2_L1_B006` vs. 1 run rep `HS2_L1_B009`
  - Expectation: some variability between PCR reps within a given sample from the sample DNA extract (but not as much as to be expected with the extraction rep)
- when `TechRep_type == extraction`:
  - In order to assess variability due to different DNA extractions, compare `TechRep_type==extraction` (1) to `TechRep_type == run/pcr` (2) for each `sample_id`
  - all `extraction` reps were sequenced on HiSeq run 2, lane 1 (same as all `run` and `pcr` reps)
  - all `extraction` reps were from different extractions and PCR reactions as `run` and `pcr` reps (wanted to eliminate any lane effect, but can't have the same barcode on reps in the same lane, so the comparison is really extraction+pcr variability)
  - e.g. `illumina_id`: 1 extraction rep `HS2_L1_B064` vs. 1 run rep `HS2_L1_B009` and 1 pcr rep `HS2_L1_B006`
  - Expectation: more variability between the extraction rep than the run and pcr reps

---

Assessing Variation Attributed to Technical Effects

In this section we aim to quantify the amount of microbial variation is attributed to each technical process within each sample with technical replicates.

What are batch effects?

> *"Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to biological or scientific variables in a study"* Leek et al., 2010

This can include different sequencing runs, lanes, personnel, extractions, reagent batches, and a myriad of other technical variants involved in sample prep and sequencing.

Luckily though, high-throughput technologies return enough data to detect and even remove batch effects.

Normalization is a data analysis technique that adjust global properties of measurements for individual samples so that they can be more appropriately compared BUT normalization does not remove batch effects!

Ordination analysis was performed using commonly used dissimilarity/distance measures (Bray & Unifrac - weighted & unweighted) with NMDS and PCoA, respectively.

Bray Dissimilarity quanitfies the *compositional* dissimilarity between 2 samples (considers presence/absence & abundance).

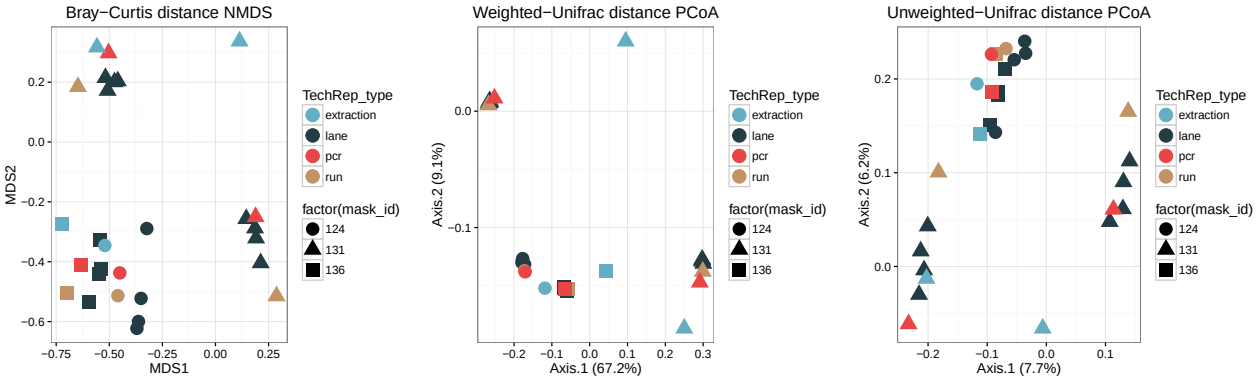
Jaccard Dissimilarity quantifies the dissimilarity between 2 samples (considers presence/absence).

Weighted-Unifrac a *quantitative* (accounts for abundance) assessment of microbial composition which accounts for phylogenetic relatedness of community members.

Unweighted-Unifrac a *qualitative* (presence/absence) assessment of microbial composition which accounts for phylogentic relatedness of community members.

Visualizing all data points on an ordination plot is a bit difficult, so a subset has been plotted here. This includes all technical replicates for 4 different specimens. This subset includes 3 samples, each from a different subject collected ~3-4 months of age, and an additional sample from one of these three subjects but collected ~15 months of age. Lane technical replicates for each sample appear very close and have consdierable overlap. Whereas, sequencing run and PCR technical replicates show a bit more diversion, while still extraction replicates show the greatest amount of diversion. Subject and age of sampling differences are apparent, which can be seen be looking at the clustering of technical replicates which correspond to each subject.

```
##   sample_id  site mask_id Reps age
## 1        24 Turku     131    7 131
## 2        25 Turku     131    7 447
## 3       188 Turku     136    7 133
## 4       393 Turku     124    7 171
```



Microbial Variation Attributed to Technical Processes

Non-parametric, permutational MANOVA (*adonis* function) was applied in order to determine the amount of microbial variation attributed to each technical process included in this experimental design. Samples

were subsetted from a phylsoeq object based on relevant comparisons. Then OTUs and metadata tables were extracted, which served as input for the `adonis()` function (included in the *vegan* package). All tests were stratified by `sample_id` which would restrict assessment of variation only within replicates from the same sample. Only code for `adonis()` using the *Bray* distance metric is shown, though analyses were also performed using *Jaccard*, *Unweighted Unifrac*, and *Weighted Unifrac* distances. These functions are not being executed in the Rmd, and the primary results are reported in a summary table below.

For this section reads were transformed to proportions and ~51K OTUs were compressed to the Species level, reducing the number of taxa to consider ~3,000.

```
#transform counts to proportions
dipp.reps2.prop = transform_sample_counts(dipp.reps2, function(x) x/sum(x))
#Compress OTUs to Species names (reduces from ~50K OTUs to 3,113)
dipp.reps2.prop.s = tax_glom(dipp.reps2.prop, taxrank = 'Species')
#melt phyloseq object
dipp.reps.m = psmelt(dipp.reps2.prop.s)
```

adonis lane

Lane replicates for each sample were performed on the same sequencing run, amplicons were generated in the same PCR reaction, and DNA was from the same extraction, thus eliminating extraction, PCR, or sequencing run effects.

```
#extract otus & meta for each subset --> used for adonis
adon.reads.lane = data.frame(otu_table(dipp.reps2.prop.s.lane), check.names=F)
adon.meta.lane = data.frame(sample_data(dipp.reps2.prop.s.lane))

#adonis on lane, only lane samples
adonis(adon.reads.lane~illumina_lane, data=adon.meta.lane, method='bray', strata=adon.meta.lane$sample_
```

adonis sequencing run

Sequencing run replicates for each sample were derived from the some DNA extract and PCR reaction, thus elimiating any effect of these processes on these results. However, because one cannot achieve a sequencing run replicate without also getting a lane replicate, lane effects cannot be eliminated from the analyses. Therefore, lane was included as a covariate in addition to sequencing run.

```
#extract otus & meta for each subset --> used for adonis
adon.reads.run = data.frame(otu_table(dipp.reps2.prop.s.runVlane), check.names=F)
adon.meta.run = data.frame(sample_data(dipp.reps2.prop.s.runVlane))

#adonis on run + lane, only lane + run samples
##all run reps came from the same PCR reaction and extraction per sample
##run reps were sequenced on different runs, and therefore also different lanes.
##there's no way of avoiding this, so we'd expect R2 for lane to increase slightly
adonis(adon.reads.run~illumina_lane+HS_run, data=adon.meta.run, method='bray', strata=adon.meta.run$samp
```

adonis PCR run

PCR run replicates were derived from the same DNA extract and were sequenced on the same lane of the same sequencing run, thus eliminating any effects of these technical aspects in this comparison.

```
#extract otus & meta for each subset --> used for adonis
adon.reads.pcrA = data.frame(otu_table(dipp.reps2.prop.s.pcrVrun), check.names=F)
adon.meta.pcrA = data.frame(sample_data(dipp.reps2.prop.s.pcrVrun))
#adon.reads.pcrB = data.frame(otu_table(dipp.reps2.prop.s.pcrVlane), check.names=F)
#adon.meta.pcrB = data.frame(sample_data(dipp.reps2.prop.s.pcrVlane))
##B is for looking at cummulative effects

#adonis on run + lane + PCR, only lane + run + PCR samples
##all PCR reps came from the same extraction per sample
##PCR reps were sequenced on different runs from lane reps, and therefore also different lanes
adonis(adon.reads.pcrA~PCR_run, data=adon.meta.pcrA, method='bray', strata=adon.meta.pcrA$sample_id)
```

adonis extraction

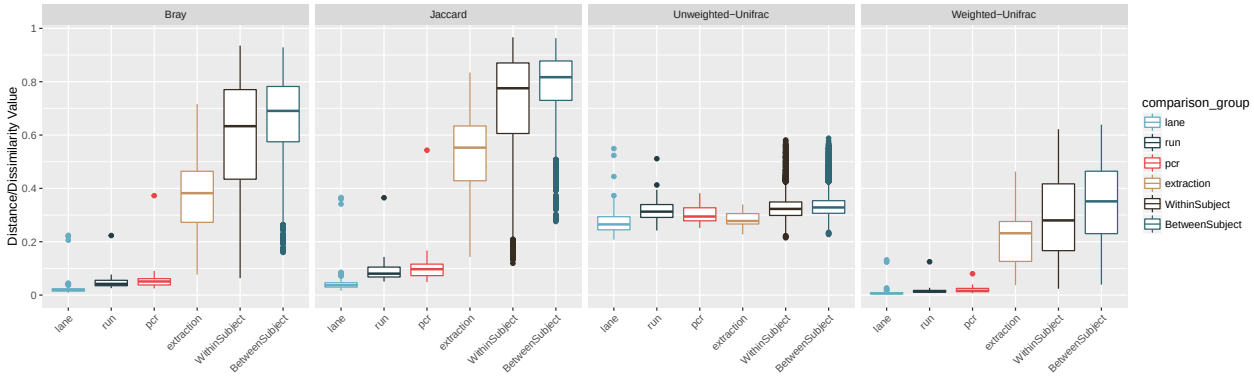
All extraction replicates were derived from the same sample specimen, but DNA was extracted at different times. However, all DNA extracts were processed in the same PCR batch and were sequenced in the same lane of the same sequencing run, thus eliminating these effects from analyses.

```
#extract otus & meta for each subset --> used for adonis
adon.reads.extractionA = data.frame(otu_table(dipp.reps2.prop.s.extractVpcr), check.names=F)
adon.meta.extractionA = data.frame(sample_data(dipp.reps2.prop.s.extractVpcr))
#adon.reads.extractionB = data.frame(otu_table(dipp.reps2.prop.s.extractVlane), check.names=F)
#adon.meta.extractionB = data.frame(sample_data(dipp.reps2.prop.s.extractVlane))
##B is for looking at cummulative effects

#adonis on run + lane + PCR + extraction, all samples
adonis(adon.reads.extractionA~extraction_cycle, data=adon.meta.extractionA, method='bray', strata=adon.m
```

adonis *R2*	Bray	Jaccard	Unweighted Unifrac	Weighted Unifrac
Lane	0.104%	0.084%	1.99%	0.077%
p-value	0.001	0.001	0.001	0.001
Run	0.220%	0.169%	3.79%	0.107%
p-value	0.001	0.001	0.001	0.001
PCR	1.34%	1.28%	4.04%	1.04%
p-value	0.001	0.001	0.001	0.001
Extraction	7.24%	5.86%	2.55%	12.0%
p-value	0.001	0.001	0.001	0.001

Distance Statistics



The distance/dissimiliarity metrics calculated for were plotted for each type of technical replicate considered. It is obvious that the distance/dissimilarity between lane, run and PCR technical replicates is quite small compared to that between different samples from the same subject ('WithinSubject' or intra-subject variation) and that between samples from different subjects ('BetweenSubject' or inter-subject variation). Likewise, the distance/dissimlarity between extraction replicates was much smaller than WithinSubject and BetweenSubjects samples, although the magnitude of this difference is less than that observed with lane, run and PCR technical replicates. These observations were statistically significant across all dissimilarity/distance metrics used.

Although, these effects were not as striking when Unweighted-Unifrac distances were used; recall, this is a phylogenetic presence/absence measurement, and this result could indicate a few things: 1) presence/absence differences within and between subjects is expected to be minimized - these samples are of the same specimen type (stool) and from subjects of a similar age and demographic; 2) presence/absence are highly influenced by samples that are included on the same lane. The latter would indicate the presence of batch effects, though it should be somewhat reassuring that biological variation still mostly exceeds (at least for lane and extraction replicates) technical variation. In the following section, we will see that this is mostly contributed by low-abundant taxa. So these presence/absence technical effects can be removed from analysis.

The result that dissimilarity/distance between biological samples exceeds that of technical replicat es indicates that while technical processes do have an effect on the microbial variation detected, the effect of biological covariates exceeds that effect and should be detectable. A summary of the results of applying pairwise-posthoc Kruskal-Nemenyi test to detect statistical significane between technical and biological replicates is summarized in the table below; p-values are presented in the order Bray:Jaccard:Unweighted-Unifrac:Weighted-Unifrac.

p-value	**lane**	**run**
run	1.00:1.00:**<0.001**:1.00	-
PCR	1.00:1.00:0.101:1.00	1.00:1.00:0.739:1.00
extraction	0.30:0.30:0.989:**<0.001**	0.42:0.42:**0.013:<0.001**
WithinSubject	**<0.001:<0.001:<0.001:<0.001**	**<0.001:<0.001**:0.148:**<0.001**
BetweenSubject	**<0.001:<0.001:<0.001:<0.001**	**<0.001:<0.001:<0.001:<0.001**

p-value	**PCR**	**extraction**
run	-	-
PCR	-	-
extraction	0.70:0.70:0.723:0.13	-
WithinSubject	**<0.001:<0.001**:0.056:**<0.001**	**<0.001:<0.001:<0.001**:0.123
BetweenSubject	**<0.001:<0.001:0.003:<0.001**	**<0.001:<0.001:<0.001:0.001**

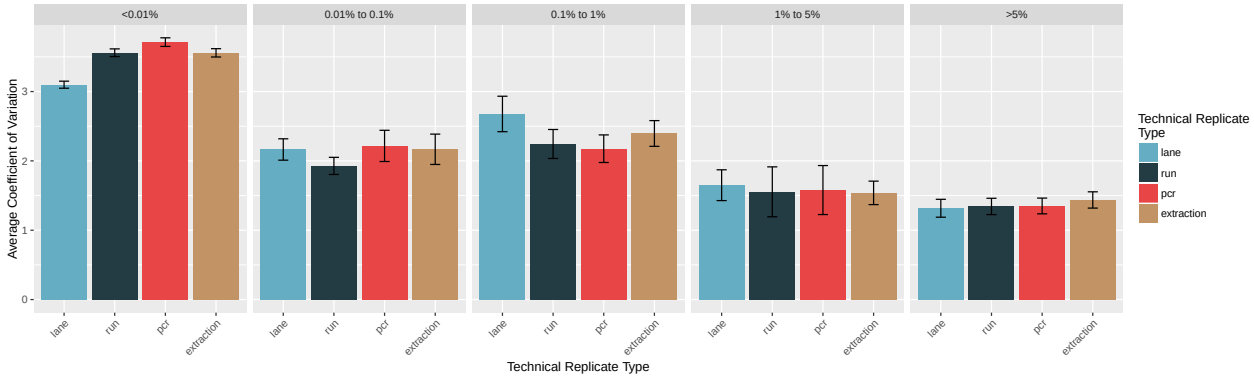
p-value	**WithinSubject**
BetweenSubject	**<0.001:<0.001:<0.001:<0.001**

Estimating the Limit of Quantification

Evaluating taxa variance across technical replicates

> Technical processes have a larger effect on low abundant taxa.

Because variance is a function of the mean, coefficient of variance (`stdev/mean`) is calculated as it normalizes variance to the mean. Coefficient of variance for each taxa across all technical replicates was calculated. Coefficient of variance is greatest among taxa whose average relative abundance is <0.01% and decreases as average relative abundance increases. That is taxa with an average relative abundance >5% have the lowest coefficient of variance and are thus less sensitive to errors introduced in technical processes of data generation.



Average coefficient of variance between technical replicates for each taxa; error bars represent standard errors.

Previous approach - using variance only as a threshold for selecting taxa

Calculate mean, median, standard deviation, and variance for each replicate for each species. The commands are hidden in this document for aesthetic reasons but can be found in the .Rmd file. Also, these calculations take several minutes to run, so the output has been provided as `dipp.rep.stats.csv` for convenience.

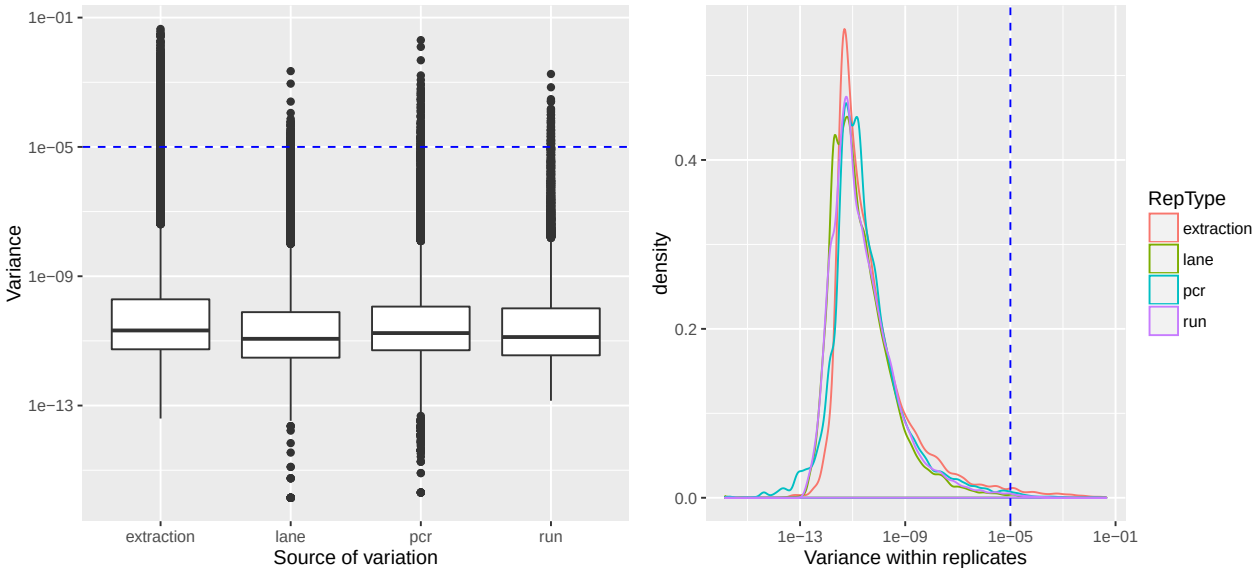
> Overall, within replicates, variance is small. This is reassuring that while technical and processing variation is present, it is not very large. Additionally, variance may serve as a useful value in determining OTU filtering parameters.

> Variance within replicates rarely exceeds 1e-5 for any given species. Therefore, a variance threshold of 1e-5 would be a justified cutoff for OTU selection. That is, if a given OTU's variance across all samples does not exceed 1e-5, then the variance observed is expected solely due to the fact that samples were processed in different batches. OTUs with a variance that exceeds 1e-5 would have additional sources of variation (i.e. biological sources) beyond that expected from technical variation alone.

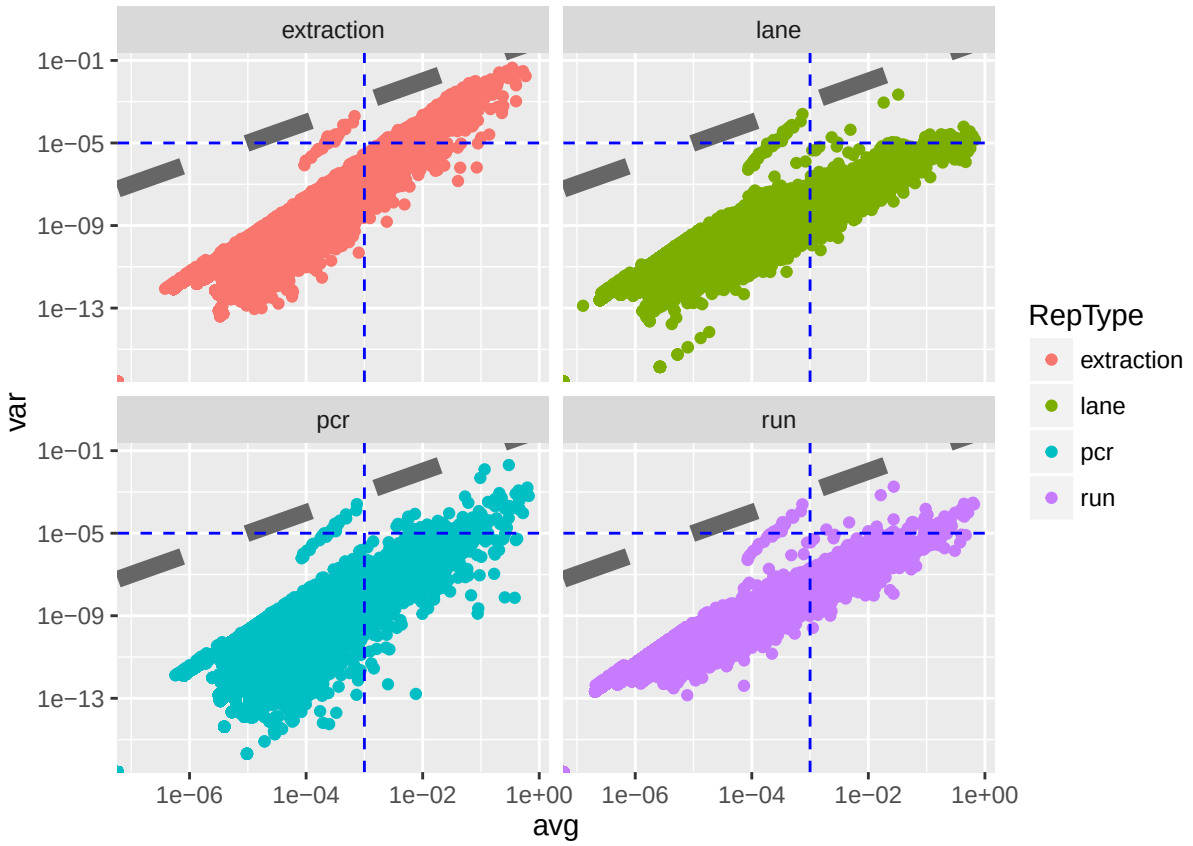
> **DISADVANTAGES:**

This approach does no consider the fact that variance is a function of the mean, and that this approach is not sensitive to the fact that lower abundant taxa will thus have a lower variance. Taxa with a lower variance could therefore be removed from further analysis with the assumption that their variance does not exceed that expected to technical processes. This is an invalid assumption, which can be remedied by using coefficient of variance as described in the section above.

This plots below represent the variance for each taxa across the different technical replicate types used.



This plot shows the linear relationship between the mean and the average for each taxa across all technical replicates. Selecting taxa based on differences in variance alone will thus be biased towards more abundant taxa.



Limit of Quantification

Based on the results above, we expect the variance observed between biological samples for each taxa to exceed the variance calculated for each taxa between technical replicates. Therefore, in this section we calculate the ratio between technical and bioloigcal variance for each taxa. When the ratio of the coefficient of variance for biological samples to technical replicates is ~1, biological variance cannot be discerned from technical variance; thus these taxa should not be considered when intepretting results that compare biological samples. However, taxa with a ratio >1 have variance which exceeds that expected to be contributed by technical processes and can be investigated for biological differences.

By setting the CV ratio threshold to 2.25, taxa can be filtered to only include OTUs from genera that have a coefficient of variance between biological samples that exceeds that observed between technical replicates. Filtering in this manner results in a dataset that includeds 172 genera which correpsonds to 519 species, or 11,583 OTUs. This data set is used in further statistical analyses which attempt to detect differences between biological covariates of interest.

This approach has several advantages:

1. *more inclusive of low abundant taxa* Rather than excluding taxa below some threshold (i.e. >0.1%), this approach would allow for low abundant taxa to be retained so long as the variance of that taxa across biological samples is not similar to the variance of that taxa across technical replicates. For example, several species of Firmicutes are thought to be beneficial but typically comprise a low relative abundance (usually ~0.1%). This approach is less likely to exclude taxa with a similar scenario.
2. *reduces the number of features to consider* this approach reduces the number of OTUs to consider (49,445 to 11,583) and corresponds to a number of species which is more similar to the number of species thought to colonize the gut (~500). Below is a table comparing the number of features at various classification levels before and after applying this filtering approach:

Level	Raw counts	Filtered by CV ratio
Phylum	87	12
Family	492	91
Genus	1,376	172
Species	2,100	519
OTU	49,445	11,583