Library size distribution for grouped microbiome data

Christina Jin ‘21

Mentored by: Johnny Hong

May 8th, 2020

BACKGROUND

- **Library size:** number of mapped reads of microbes for an observation.
- **Scientific interest:**
  - Whether microbial compositions of different groups are the same.
  - Might result in inaccurate inference due to differences in library size.
- **Permutation testing:**
  - Small sample size; complicated distribution of the microbial community
  - Assumes observations are exchangeable under H_0
  - E.g.: PERMANOVA: a non-parametric method that tests whether groups are significantly different based on a categorical factor.
- **Project Goal:** Test whether library size distribution depends on group membership.

METHOD: Data Collection

- **Qiita**: an open-source microbiome data platform
- Recent 28 datasets with publications and categorical groupings on microbiome researches were drawn from QIITA website.
  - OTU (Operational Taxonomic Unit) table:
    - Removed samples with total OTU counts ≤ 1000
  - Sample information table:
    - # samples * # categories

	11815.PWW.3417.SkinEtOH	11815.JCK.10101.SkinEtOH	11815.JCK.10104.SkinEtOH
4479946	0	0	0
145205	2	0	0
4436710	0	39	31
244331	0	0	0

(OTU Table)

sample_name	sample_type
11815.BN.593.FecalFTA	feces
11815.BN.593.SkinEtOH	skin
11815.BN.593.TongueEtOH	saliva
11815.BN.594.FecalFTA	feces
11815.BN.594.SkinEtOH	skin

(Sample Info Table)

METHOD: Library Size Normalization

1. Total Sum Scaling

- Sum of OTU units per observation.
- Sensitive to outliers
- May incorrectly bias taxa that are sampled preferentially as sequencing yield increases.

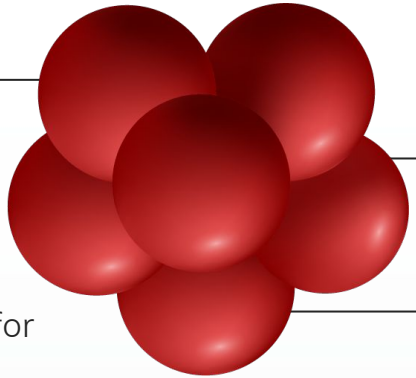
2. Cumulative Sum Scaling (Paulson et al, 2013)

- Sum of OTU units per observation, up to a certain percentile.
- The more abundant bacteria might have been preferentially sequenced.
- Highly dependent on the threshold of the percentile.

3. Geometric Mean of Pairwise Ratios (Chen et al, 2018)

- Considers the relative size across different samples
- More robust to outliers

METHOD: Assessment



Kruskal-Wallis Test

- Rank-based non-parametric method for testing whether samples originate from the same distribution.
- P value cutoff of 0.05

Effect Size (η^2)

- η^2 = (H-#groups-1)/(#obs-#groups),
- Indicates the percentage of variance in library sizes explained by groups.

Plots

- Qualitative(visual) comparison of the library size distribution plots across different groups

RESULTS & DISCUSSION

- Meta-analysis (309 entries)

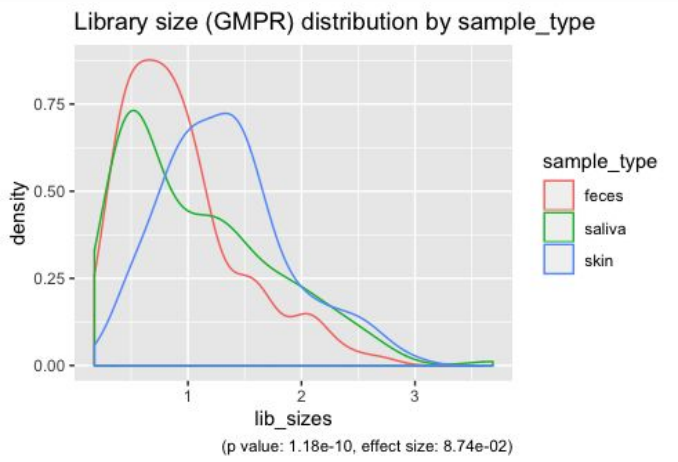
study_id	summary	method	group	p_val	effsize	category	interest
11815	bat	TSS	host_sex	0.7868881	-0.0021357	demographics	FALSE
11815	bat	GMPR	host_sex	0.5967447	-0.0016591	demographics	FALSE
11815	bat	CSS	host_sex	0.7761824	-0.0021179	demographics	FALSE

- In general, it is common that library size distribution is different across different groups.
- Body site is a main group of interest that tends to exhibit large differences:

method2	num_group	pct_signif_p	pct_large_effsize
CSS	63	47.62%	15.87%
GMPR2	**63**	**55.56%**	**20.63%**
TSS	63	39.68%	7.94%

category	num_group	pct_signif_p	pct_large_effsize
body_site	**5**	**100%**	**60%**
chemicals	7	42.86%	28.57%
demographics	13	38.46%	7.69%
disease	7	28.57%	0%
food	5	100%	20%
geo_loc	10	60%	10%
habit	1	0%	0%
temp	2	0%	0%
time	7	71.43%	42.86%

RESULTS & DISCUSSION (cont.)



- "Microbial richness associated with bat skin is significantly greater than gut or oral microbial communities." (Lutz, 2018)
- Library size could have been the reason too.

- **Limitations:**
  - Small sample size may result in large p-values.
  - It's hard to eliminate partial effect/ensure independence, since there are too many groups recorded in each study.

Thanks for listening!