**Multi-Kingdom Spike-seq (MK-SpikeSeq) protocol v1.0**

**Chitong Rao 2019/12**

The backbone of MK-SpikeSeq is the addition of a defined number of bacterial, fungal and archaeal cells to the sample of interest. These cells serve as internal controls for sample processing and subsequent normalization. Here we outline the protocol for applying MK-SpikeSeq to mammalian stool samples, however, this method can easily be adjusted for other samples types by changing the spike-in taxa and the quantity added. The criteria for spike-in usage are:

1) The spike-in taxa should absent in the sample

2) Once added, the spike-in should account for 0.1%~10% of the sample quantity (assuming the sequencing depth is >10K per sample; higher depth of sequencing could yield broader range of detection).

To test these qualifications, when studying a new type of samples, multiple representative test samples (ideally covering both upper and lower bounds of absolute abundance) should be used for DNA extraction. The targeted ribosomal DNA (e.g., 16S) of these test samples could be quantified using quantitative PCR (qPCR) with the following qPCR primers and literature reviewed (or empirically Illumina sequenced) to confirm the absence of selected spike-in taxa. To identify an appropriate quantity of spike-in cells, 10-fold serial dilutions of defined number of spike-in cells should be DNA extracted and qPCR quantified in the same way so that a select dilution of spike-in would suffice the 0.1%~10% abundance requirement for most test samples (see the stool example below). Some qPCR primers we use are:

16S\_V3-F AGACACGGYCCARACTCCTA

16S\_V3-R GCTGCTGGCACGKAGTTAG

Arch344F ACGGGGYGCAGCAGGCGCGA

Arch517R GWATTACCGCGGCKGCTG

ITS1mod TCCGTWGGTGAACCWGCGG

ITS2R GCTGCGTTCTTCATCGATGC

For human and mouse stool samples we recommend using *Salinibacter ruber* DSM 13855, *Haloarcula hispanica* ATCC 33960 and *Trichoderma reesei* ATCC 13631 as spike-in taxa for bacteria, archaea and fungi, respectively. We systematically checked the absence/rarity of these taxa against several major public depositories of host-associated microbiome data (see table below). These three strains are also publicly available and have completed genomes; *S. ruber* and *H. hispanica* have 1 and 3 copies of 16S gene per genome, respectively (*T. reesei* fungal rDNA copy number is undetermined, usually 10~100).

|  |  |  |  |
| --- | --- | --- | --- |
| **searched database/study** | **searched taxon** | **hit samples (total samples)** | **host-associated samples** |
| Qiita (host & environmental 16S database) | **g\_\_Salinibacter** | **234 (178,259)** | **11** |
| Qiita (host & environmental 16S database) | **g\_\_Haloarcula** | **121 (178,259)** | **25** |
| Qiita (host & environmental 16S database) | g\_\_Salmonella | 38,257 (178,259) | 23,028 |
| Qiita (host & environmental 16S database) | g\_\_Bacteroides | 117,181 (178,259) | 73,579 |
| Human Microbiome Project (HMP) 16S | **g\_\_Salinibacter** | **0 (2,910)** | **/** |
| Human Microbiome Project (HMP) 16S | **g\_\_Haloarcula** | **0 (2,910)** | **/** |
| Human Microbiome Project (HMP) 16S | g\_\_Salmonella | 0 (2,910) | / |
| Targeted Host-associated Fungi (THF) ITS v1.6.1 | **g\_\_Trichoderma;s\_\_reesei** | **0** | **0** |
| Human Microbiome Project (HMP) metagenomics | **Trichoderma reesei** | **1/215** | **2/472 samples ; 5/27,091,491,028 reads** |
| various mammalian mycobiome studies | **g\_\_Trichoderma;s\_\_reesei** | **0** | **0** |

Spike-in cells are grown according to supplier’s instructions. *S. ruber* and *H. hispanica* are recovered in ATCC agar medium 1270 and then inoculated in liquid broth, both grown at 37C for about two weeks until dense. *T. reesei* is grown on Potato Dextrose Agar medium and fungal spores are collected using swabs. These spike-in cells are harvested in DNA-free PBS, passed through 70 um cell strainers and quantified based on optical density unit (ODU, 1 ODU = 1 OD \* 1 mL) at 600nm. As slight differences of rDNA quantity per ODU spike-in cells will occur between batches (especially for fungal cells), it is suggested to make large enough homogenous stocks of spike-in cells for one study.

Stool samples are first weighted and aliquoted into bead-beating DNA extraction tubes. We aim for 50~100 mg each for human stool samples and 5~20 mg each for mouse stools. Our pilot qPCR experiments suggest to spike in 10^-2 ODU *S. ruber*, 10^-5 ODU *H. hispanica* and 10^-4 ODU *T. reesei* to each 100 mg human stool or 10 mg mouse stool (we observe that raw mouse stools usually have higher microbial density than human ones). Spike-in volumes are calculated based on the measured optical densities of spike-in cells (e.g., 10 uL of *S. ruber* 1 OD stocks equals 10^-2 ODU); different spike-in cells could be pre-mixed for convenience so that small volumes (10~100 uL) of total spike-in cells are added to samples before DNA extraction. Spike-in quantities are also adjusted to the sample weights (e.g., add half volume of spike-in cell mixture to a 50 mg stool sample compared to a 100 mg sample), so that the weight information would not be required in the final normalization step. Alternatively, a uniform spike-in quantity could be added regardless of sample abundances (i.e., same volume of spike-in cells for each sample) and the information of sample weights will be used for normalization in the end.

The above spike-in quantities were chosen based on: 1) 10^-2 ODU *S. ruber* usually takes ~1% of raw reads from sequencing, and 2) 10^-5 ODU *H. hispanica* equals roughly 10^4 cells and 10^-4 ODU *T. reesei* equals roughly 10^3 cells, which would be low enough to accommodate to the usually low archaeal and fungal signal in stools, but still high enough for robust pipetting of consistent number of cells. Using such spike-in quantities, the spike-in rDNA abundances are pre-determined using qPCR against DNA standards. To make these DNA standards, a plasmid is constructed using Gibson cloning to contain tandem single copy of each rDNA, and standard curves are generated in rDNA-specific qPCR using 10-fold serial dilutions of the plasmid DNA as template. Our measurement of the 10^-2/10^-5/10^-4 ODU multi-kingdom spike-in mix is: bac16S\_Srub : arch16S\_Hhis : ITS1\_Tree ~ 10,000,000 : 50,000 : 20,000. Note that the cross-kingdom ratios are experimentally quantified but the exact numbers are arbitrary (as practically DNA extraction efficiency is not 100%, the exact rDNA copy number per sample is less meaningful/reliable than the relative changes between samples in one study). These spike-in rDNA quantities will be used as scaling factors in the later cross-kingdom normalization step.

Following spike-in addition, DNA extraction and amplicon library preparation are standard. We use ZymoBIOMICS DNA extraction kit for DNA extraction, and elute DNA in 50 uL water. For rDNA amplicon sequencing, we do two-step PCR using Nextera-based sequencing adapters. Below are the primers and reaction conditions we use in the two-step PCR:

* 1st PCR primers:

Uni16S\_V3F TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-CCTACGGGNGGCWGCAG

Uni16S\_V4F TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG- GTGYCAGCMGCCGCGGTAA

Uni16S-V4R GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-GGACTACNVGGGTWTCTAAT

Arch516F TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-TGYCAGCCGCCGCGGTAAHACCVGC

Arch915R: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-GTGCTCCCCCGCCAATTCCT

ITS1F TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-CTYGGTCATTTAGAGGAAGTAA

ITS2R GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-GCTGCGTTCTTCATCGATGC

* 1st PCR recipe (total 10ul)

SuperFi 2x mastermix 5ul

Forward primer 1uM 2ul

Reverse primer 1uM 2ul

gDNA template 1ul

* 1st PCR program (change cycle number\* and annealing temperature\*\* as needed)

98C 30s

30~40 cycles\* of

98C 10s

50C\*\* 10s

72C 1min

72C 5min

4C hold

\*[[Fecal samples: Uni16S\_V4/V3V4 – 30cycle, Arch16S\_V4V5 – 40cycle, ITS1 – 40cycle]]

\*\*[[50C for Uni16S\_V4/V3V4 and ITS1 primers, 60C for Arch16S\_V4V5]]

Purify 1st PCR products using Ampure XP beads. Alternatively (and more conveniently), perform 50-fold dilution of 1st PCR products by adding 2ul PCR reaction into 98ul water in a 96-well plate and mix.

* 2nd PCR primers:

Nextera\_i5 AATGATACGGCGACCACCGAGATCTACAC-[i5]-TCGTCGGCAGCGTC

Nextera\_i7 CAAGCAGAAGACGGCATACGAGAT-[i7]-GTCTCGTGGGCTCGG

* 2nd PCR recipe (total 10ul)

SuperFi 2x mastermix 5ul

N7XX primer 1uM 1ul

S5XX primer 1uM 1ul

Diluted 1st PCR product 3ul

* 2nd PCR program

98C 30s

10 cycles of

98C 10s

55C 10s

72C 30s

72C 5min

4C hold

Purify and quantify each 2nd PCR products before pooling. Alternatively (and more conveniently), combine equal volume of each 2nd PCR product (usually leads to <4-fold variation in raw reads count). Purify using either Ampure XP beads or Zymo DNA Clean & Concentrator kit.

The prepared amplicon libraries are qPCR quantified using NEBNext Library Quant kit for Illumina, before further pooling and loading on Illumina Miseq (or other platforms).

Illumina raw reads are demultiplexed ([skewer](https://doi.org/10.1186/1471-2105-15-182)), paired-end joined ([PEAR](https://cme.h-its.org/exelixis/web/software/pear/), only for overlapping reads generated by Miseq long-read sequencing), adapter trimmed ([cutadapt](https://cutadapt.readthedocs.io/en/stable/)), quality filtered ([vsearch](https://doi.org/10.7717/peerj.2584)), dereplicated and denoised ([usearch](https://doi.org/10.1093/bioinformatics/btq461)) and finally derived into an OTU table ([FAST](https://github.com/ZeweiSong/FAST)). 16S and ITS OTUs are mapped against public databases like SILVA and UNITE, respectively ([QIIME 2](https://docs.qiime2.org/)). Representative scripts for Nextseq and Miseq data are also provided.

Spurious bacterial OTUs based on taxonomic annotations are removed from the archaeal table. Amplicon samples with <1K read counts or <10 spike-in counts are excluded from further analyses. Counts of spike-in OTUs (see below, based on taxonomic annotation) are used in normalization for each other taxon:

k\_\_Bacteria;p\_\_Bacteroidetes;c\_\_Cytophagia;o\_\_Order II;f\_\_Rhodothermaceae;g\_\_Salinibacter

k\_\_Archaea;p\_\_Euryarchaeota;c\_\_Halobacteria;o\_\_Halobacteriales;f\_\_Halobacteriaceae;g\_\_Haloarcula

k\_\_Fungi;p\_\_Ascomycota;c\_\_Sordariomycetes;o\_\_Hypocreales;f\_\_Hypocreaceae;g\_\_Trichoderma

Normalized OTU abundance = OTU read count / Spike-in read count \* kingdom-specific scaling factor