1 Materials and Methods

1.1 A finite mixture and latent Dirichlet model

We describe the sampling and sequencing procedure of metagenomics as a generative model: First, in a metagenomic dataset, microbial organisms can be represented by their genome set $\{g_1, g_2, ..., g_m\} \triangleq G$. These genomes are subject to a particular discrete distribution $(\theta_1, \theta_2, ..., \theta_m) \triangleq \vec{\theta}$, where $\sum_{i=1}^{m} \theta_i = 1$ and each mixture probability θ_j of a genome g_j is proportional to its abundance a_j and base length l_j, e.g. $\theta_j \propto a_j l_j$ according to Xia et.al 2012. In fact, the metagenome M can be noted as $M = \sum_{i=1}^{m} \theta_i g_i$. Second, randomly choose a genome g_j with the multinomial probability $\vec{\theta}$ from metagenome M. Third, for the chosen genome g_j, randomly generate a read r_k. The generation of reads from the genome g_j can be reasonably approximated by the genome g_j's component distribution $(\phi_{1,g_j}, \phi_{2,g_j}, ..., \phi_{K_j,g_j}) \triangleq \vec{\phi_{g_j}}$, where ϕ_{i,g_j} means the probability of generating a component i in genome g_j, e.g. $p\{component\ i|G = g_j\}$, and K_j represents the number of components that can be generated by genome g_j. And we assume the total number of components in metagenome M is K. We denote all the component distributions of target genomes as distribution set $\{\vec{\phi_{g_1}}, \vec{\phi_{g_2}}, ..., \vec{\phi_{g_m}}\} \triangleq \Phi$.

The above procedure will repeat for N times to generate N metagenomic reads. In a generative model, the total number of metagenomic reads can be approximated by Poisson distribution, e.g. $N \sim Poisson(\lambda)$, where λ notes the inverse of the average number of reads of metagenomic datasets. As we can see, both sampling procedures of the second and third steps are subject to multinomial distributions, e.g. $g_j \sim Mult(\vec{\theta})$ and $r_k \sim Mult(\vec{\phi_{g_j}})$ respectively. For calculation's sake, we follow the suggestion of Pritchard et.al 2000 in using the Dirichlet distribution as the prior distribution of genome mixture probability in the metagenome because of the Dirichlet-Multinomial conjugacy. However, we do not assume prior distributions for the component distributions of genomes since we can approximate their full conditional probabilities via sequence alignment results. The procedure can be shown as the pseudocode in Procedure 1:

In most real studies, our knowledge of the genome set G is limited, so the last component g_m of G is often reserved as a collective of unknown genomes. Hence, we can induce the formula of genome relative abundance(GRA) with known genomes $\{g_1, g_2, ..., g_{m-1}\}$ and their genome sizes $\{l_1, l_2, ..., l_{m-1}\}$ respectively, where the relative abundance of the known genome g_j is

$$a_j = \frac{\#\ of\ genome\ g_j}{\#\ of\ all\ known\ genomes}$$

Noting that $\sum_{j=1}^{m-1} a_j = 1$ and $a_s = t\frac{\theta_s}{l_s}$, where t is the propotional coefficient, we have $\sum_{j=1}^{m-1} \frac{\theta_s}{l_s} = t$, thus the GRA formular under the metagenomic generative model is:

$$a_j = \frac{\theta_j}{l_j \sum_{s=1}^{m-1} \frac{\theta_s}{l_s}}$$

Procedure 1 A finite mixture and latent Dirichlet model for metagenomics

```
Require: the hyperparameter $\vec{\alpha}$, the Poisson parameter $\lambda$, the genomes $G$
    present in the community
Ensure: the read dataset $R$
  sample the genome mixture probability $\vec{\theta} \sim Dir(\vec{\alpha})$ for a metagenome
  sample the total number of reads $N \sim Poisson(\lambda)$ for a metagenome
  generate the component distributions $\Phi$ according to each genome in $G$
  repeat
    1. sample a genome $g_j \sim Mult(\vec{\theta})$
    2. sample a read $r_k \sim Mult(\vec{\phi_{g_j}})$
  until the total number of metagenomic reads N is reached
  return the read dataset $\{r_1, r_2, ..., r_N\} \triangleq R$
```

1.2 Mixture Parameter Inference and Gibbs Sampling

In this section, we assume hidden variables $(z_1, z_2, ...z_N) \triangleq \vec{z}$, where each entry z_i is the index of the genome which the read r_i is from, and then formulate an approximate inference algorithm of Gibbs Sampling to emulate the probability distribution of $\vec{z}$ given the observations of metagenomic reads, e.g. $p\{\vec{z}|R; \vec{\alpha}, \Phi\}$. We can estimate the values of parameters $\vec{\theta}$ using the samples of $p\{\vec{z}|R; \vec{\alpha}, \Phi\}$ after the burn-in period of Gibbs sampling, and then calculate the genome relative abundance via the above formula.

Gibbs sampling generates an instance of each dimension z_i of $\vec{z}$ in turn, subject to their full conditional $p\{z|z_{\vec{-}i}, R; \vec{\alpha}, \Phi\}$ respectively, where $z_{\vec{-}i}$ notes all other dimensions of $\vec{z}$ except z_i. It can be shown (Gelman et al., 1995) that the sequence of samples $\{\vec{z}_1, \vec{z}_2, ...\}$ constitutes a Markov chain whose stationary distribution is $p\{\vec{z}|R; \vec{\alpha}, \Phi\}$. In real application of metagenomic data, we just use only one sample like $\vec{z}^*$ after the burn-in period to obtain the parameter inference, as the reads number is enormous.

To derive the full conditional $p\{z|z_{\vec{-}i}, R; \vec{\alpha}, \Phi\}$, we firstly calculate the probability $p\{\vec{z}|\vec{\alpha}\}$. Starting with the probability of $\vec{z}$ conditioned on the genome mixture probability $\vec{\theta}$, and noting that the hidden genome index are generated according to multinomial trials, we have

$$p\{\vec{z}|\theta\} = \prod_{i=1}^{M} \theta_i^{n_i}$$

where n_i refers to the number of reads whose corresponding genome index is g_i. Noting that we assume $p\{\vec{\theta}|\vec{\alpha}\}$ subjects to Dirichlet distribution, so we have

$$\int p\{\vec{\theta}|\vec{\alpha}\} d\vec{\theta} = \int \frac{1}{\triangle(\vec{\alpha})} \prod_{i=1}^{M} \theta_i^{\alpha_i - 1} d\vec{\theta} = 1$$

where $\triangle(\vec{\alpha}) = \frac{\prod_{k=1}^{dim\,\alpha} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{dim\,\alpha} \alpha_k)}$. Then by conditional probability formula and integrating out $\vec{\theta}$, we obtain

$$p\{\vec{z}|\vec{\alpha}\} = \int p\{\vec{z}|\vec{\theta}\}p\{\vec{\theta}|\vec{\alpha}\}d\vec{\theta} = \int \frac{1}{\triangle(\vec{\alpha})}\prod_{i=1}^{M}\theta_i^{n_i+\alpha_i-1}d\vec{\theta} = \frac{\triangle(\vec{n}+\vec{\alpha})}{\triangle(\vec{\alpha})}$$

Secondly, we obtain the probability of reads conditioned on the hidden variables with the knowledge of component distribution set Φ as

$$p\{\vec{r}|\vec{z};\Phi\} = \prod_{i=1}^{N} p\{r_i|z_i,\Phi\} = \prod_{i=1}^{N}\phi_{i,g_i} = \prod_{k=1}^{K}\prod_{m=1}^{M}(\phi_{k,g_m})^{n_{k,m}}$$

where we assume the generation of each read is independent with other reads. We can estimate ϕ_{i,g_i} by the ratio of high quality hits of r_i to all the high quality read hits on the target genome g_i, that is

$$\phi_{i,g_i} \approx \frac{\#\ of\ r_i\ that\ hit\ genome\ g_i\ with\ high\ quality}{\#\ of\ reads\ that\ hit\ genome\ g_i\ with\ high\ quality}$$

Thus, we have

$$\begin{aligned}
p\{z_i = t|\vec{z_{-i}}, R; \vec{\alpha}, \Phi\} &= \frac{p\{\vec{z},\vec{r}\}}{p\{\vec{z_{-i}},\vec{r}\}} \\
&= \frac{p\{\vec{r}|\vec{z}\}p\{\vec{z}\}}{p\{\vec{r_{-i}}|\vec{z_{-i}}\}p\{\vec{z_{-i}}\}p\{r_i\}}(independence\ in\ R) \\
&\propto \frac{p\{\vec{r}|\vec{z}\}}{p\{\vec{r_{-i}}|\vec{z_{-i}}\}}\frac{p\{\vec{z}\}}{p\{\vec{z_{-i}}\}} \quad (p\{r_i\}\ is\ constant\ to\ \vec{z}) \\
&= \frac{\prod_{k=1}^{K}\prod_{m=1}^{M}(\phi_{k,g_m})^{n_{k,m}}}{\prod_{k=1}^{K}\prod_{m=1}^{M}(\phi_{k,g_m})^{n_{k,m,-i}}}\frac{\triangle(\vec{n}+\vec{\alpha})}{\triangle(\vec{n_{-i}}+\vec{\alpha})} \\
&= \phi_{i,g_t}\cdot\frac{(n_{t,-i}+\alpha_t)}{\sum_{j=1}^{M}(n_{j,-i}+\alpha_j)} \\
&\propto \phi_{i,g_t}\cdot(n_{t,-i}+\alpha_t)
\end{aligned} \tag{1}$$

as we know that $\Gamma(a+1) = a\Gamma(a), n_k = n_{k,-i}+1$. Finally, we can infer the genome mixture parameters $\vec{\theta}$ using the frequency vector $\vec{n}$ of the sample generated by Gibbs sampling. For Multinomial-Dirichlet conjugacy, we have

$$p\{\vec{\theta}|\vec{z},\vec{\alpha}\} = Dir(\theta|\vec{n}+\vec{\alpha})$$

and we can estimate $\vec{\theta}$ by the average of its distribution:

$$\theta_k = \frac{n_k+\alpha_k}{\sum_{i=1}^{M}(n_i+\alpha_i)}$$

The pseudocode of Gibbs sampling is as following:

Algorithm 2 Gibbs sampling for metagenomic model

```
Require: reference genomes $M$, metagenomic reads $R$, hyperparameter $\vec{\alpha}$
Global data: count statistics $\{n_m\}$, component distributions $\Phi$, memory for
  full conditionals $p\{z_i|z\vec{_{-i}}, R; \vec{\alpha}, \Phi\}$
Ensure: mixture probability $\vec{\theta}$
  //initialization:
  obtain component distributions $\Phi$ according to alignment results
  zero all count statistics $\{n_m\}$
  for $i = 1$ to $N$ do
    sample a genome index $z_i = m \sim Mult(M)$
    increment sampled genome count $n_m = n_m + 1$
  end for
  //Gibbs sampling
  while not finished do
    for $i = 1$ to $N$ do
      decrement target genome count $n_m = n_m - 1$
      sample a genome index $z_i = \tilde{m} \sim p\{z_i|z\vec{_{-i}}, R; \vec{\alpha}, \Phi\}$
      increment sampled genome count $n_{\tilde{m}} = n_{\tilde{m}} + 1$
    end for
    if converged and $L$ samples generated then
      return mixture probability $\vec{\theta}$ according to the equation
    end if
  end while
```