LLNL-TR-680776

Technical Report on Modeling for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data

K. McLoughlin

January 11, 2016

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-TR-680776

Technical Report on Modeling for Quasispecies Abundance Inference with Confidence Intervals from Metagenomic Sequence Data

DHS Bioforensics Program

IAA No.: HSHQPM-13-X-00219

Principal Investigator and Correspondent

Kevin McLoughlin

Lawrence Livermore National Laboratory (LLNL), Livermore, CA

925-423-5486, mcloughlin2@llnl.gov

Submission date: March 4, 2014

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

1 Introduction

The overall aim of this project is to develop a software package, called MetaQuant, that can determine the constituents of a complex microbial sample and estimate their relative abundances by analysis of metagenomic sequencing data. The goal for Task 1 is to create a generative model describing the stochastic process underlying the creation of sequence read pairs in the data set. The stages in this generative process include the selection of a source genome sequence for each read pair, with probability dependent on its abundance in the sample. The other stages describe the evolution of the source genome from its nearest common ancestor with a reference genome, breakage of the source DNA into short fragments, and the errors in sequencing the ends of the fragments to produce read pairs.

We use a Bayesian framework to model the generation of read pairs from source genomes. This means that, rather than simply fitting maximum likelihood estimates for the abundances and other parameters, we aim to fit a probability distribution for each parameter. This will allow us to estimate confidence intervals for each abundance level.

2 The eXpress model for RNA-Seq data

MetaQuant's generative model is inspired in part by the model underlying the RNA-Seq analysis tool eXpress [1], which is diagrammed in Figure 1. In this graphical model diagram, the circles represent the random variables (either hidden or observed) and the parameters of their probability distributions; the arrows represent conditional dependence relationships. The box surrounding the four variables on the left indicates that the variables are instantiated N times, once for each observed read pair. The random variables are the fragment length L, the target genome T, the fragment start position P, and the observed fragment (read pair) sequence F. The circle for F is shaded to indicate that it is observed; unshaded circles correspond to hidden variables and parameters.

Under the eXpress model, fragments are generated as follows:

- A fragment length l is chosen from the range [1, $L_{\max}$] according to a categorical distribution, with probability λ_l. $L_{\max}$ is the maximum expected fragment length, which depends on the procedure used for shearing cDNA molecules, along with their sequence composition.
- A target (transcript) t is sampled from the set of all possible targets $\mathcal{T}$ with probability τ_t^l. τ_t^l is determined by the relative abundance of the target ρ_t and the length of the target sequence $l(t)$, according to a relation to be described below.
- The fragment starting position p within the target sequence is chosen from the range [1, $l(t) - l + 1$], with probability π_p^{tl}. π_p^{tl} depends on the sequence context surrounding position

p in target t; it describes the tendency for DNA strands to break preferentially at the centers of particular k-mer sequences.

- The ends of the fragment are sequenced, with substitutions, insertions and deletions produced randomly by errors in the sequencing process. The probability of producing the observed read pair sequences f is ϕ_f^{ptl}; it depends primarily on the actual fragment sequence, which is determined by the target sequence and the fragment position and length.

In the eXpress generative model, the fragment length is selected first. This is somewhat counterintuitive; however, it simplifies the model because it reflects the fact that the target sampling probabilities τ_t^l depend on the target and fragment lengths. Given equal abundances, longer target sequences are more likely to be sampled, and targets shorter than the fragment length will not be sampled at all.

The sampling probabilities are also affected by sequence bias, which causes some fragments to be selected more than others. The sequence bias is described by a set of normalized weights w_p^{tl}, which in turn are parameterized by a set of third order Markov chain probabilities specified for each position in a 21-base window surrounding the fragment start position. Details of how the weights w_p^{tl} are calculated can be found in [2]. Given the weights, the conditional probability of starting a fragment at position p is

$$\pi_p^{tl} = \mathbb{P}[P = p | T = t, L = l] = \frac{w_p^{tl}}{\sum_{q=1}^{l(t)-l+1} w_q^{tl}}$$

while the target sampling probability is

$$\tau_t^l = \mathbb{P}[T = t | L = l] = \frac{\rho_t \sum_{p=1}^{l(t)-l+1} w_p^{tl}}{\sum_{u \in \mathcal{T}} \rho_u \sum_{q=1}^{l(u)-l+1} w_q^{ul}}$$

The probability of producing the observed read pair sequence, ϕ_f^{ptl}, is parameterized by sets of probabilities for insertions or deletions of a given length, and emission probabilities for observing a base at a particular position given the true bases at the same position and the one preceding it.

Under the above model, the complete data likelihood given the observed set of fragment read pairs $\mathcal{F} = (f_1, f_2, \ldots, f_N)$ is given by

$$\mathcal{L}(\lambda, \tau, \pi, \phi | \mathcal{F}) = \prod_{n=1}^{N} \lambda_{l_n} \tau_{t_n}^{l_n} \pi_{p_n}^{t_n l_n} \phi_{f_n}^{p_n t_n l_n}$$

The eXpress software fits maximum likelihood (ML) estimates for the parameters of the above model, including the relative abundances, using an online expectation-maximization algorithm [3]. The input data for model fitting are alignments of read pairs to target sequences, produced by a tool such as Bowtie or BWA. In the context of model fitting, “online” means that each sequence read is examined exactly once; there is no need to make multiple passes over the data set, as in the conventional batch EM algorithm. As a result, online algorithms such as eXpress are much faster and require much less memory than batch algorithms, especially when applied to the large datasets produced with high-throughput sequencing. Note that the model parameters describing sequencing errors, fragmentation bias and fragment length distribution are fit independently for every sequence dataset; there is no need to store error models for particular sequencers.

3 Limitations of the eXpress model for metagenomic data

While the eXpress model captures many aspects of the stochastic process underlying all types of high-throughput sequencing data, it was primarily designed for gene expression analysis. Therefore, it incorporates some assumptions that limit its applicability to metagenomic data:

1. The model assumes that the true target sequences are known; or at least, that they do not deviate greatly from a known set of reference transcriptome sequences. In a metagenomic sample, the constituent genomes will almost never be identical to any reference genome sequences. At best, some organisms present in the sample may share recent common ancestors with reference strains of the same species. Therefore, a model for metagenomic data must allow for evolutionary divergence between the true genomes and the reference genomes.

2. Most mRNA transcripts are relatively short linear sequences, on the order of a few kilobases in length. Typical fragment lengths are on the order of 300 bp, a substantial fraction of the target length. Therefore, the fragment length has a significant effect on the target sampling probability. By contrast, the microbial targets in a metagenomic sample are mostly circular genomes and plasmids, with lengths ranging from hundreds to thousands of kilobases, so the target sampling probability is nearly independent of the fragment length.

3. A substantial fraction (on the order of 10-20%) of the possible transcripts may be expressed in a given tissue sample. Therefore, there is no reason to favor sparse solutions. For metagenomic data, by contrast, we expect to see nonzero abundances for only a tiny fraction of the targets in a reference genome database.

To investigate how eXpress performs on metagenomic data, we used it to analyze a mock community dataset prepared as a benchmarking standard for the members of the Human Microbiome Program consortium (HMP) [4, 5]. The dataset was prepared by sequencing a mixture of genomic DNA from 21 bacterial species and one fungus, in staggered quantities selected to yield specific numbers of 16S rRNA copies ranging from 1,000 to 1,000,000. The mixture was sequenced on an Illumina GA II, yielding 7,932,819 75-bp single-end reads; the dataset is available at the NCBI SRA, under accession SRR172903.

We used BWA version 0.7.5a-r405 to align the reads to a reference genome database described in [6], obtained from Washington University, St. Louis (WUSTL). We converted the SAM file output by BWA into binary (BAM) format, and ran eXpress on the BAM file with default parameter settings. eXpress assigned nonzero fragment counts to target sequences from 511 distinct microbial strains. We grouped the strains by genus for the 19 genera represented in the mock community dataset, and drew bar plots of the relative abundance estimates, expressed as fragments per kilobase per million reads mapped (FPKM). Figures 2 and 3 show the results for two genera, *Bacillus* and *Lactobacillus*. The red bar in each plot indicates the strain that was actually contained in the mock community mixture.

For *Bacillus*, we note that the largest abundance was assigned to the *B. cereus* strain that was actually present. However, eXpress also estimated substantial abundances for several other *B. cereus* strains and even strains of other *Bacillus* species. This is happening because a large number of reads map equally well to the genomes of these different strains; in the absence of other evidence, eXpress apportions the fragment counts among the multiple genomes. For *Lactobacillus*, the correct strain is not even included in the top 5 abundance estimates.

In the mock community sample, all the bacterial strains used had genome sequences that were included in the reference genome database. Therefore, assumption (1) above is in fact valid for this dataset, though it will not generally be true for real environmental metagenomic samples. The main problem with applying the eXpress model to this dataset is assumption (3). For metagenomic data, we seek to explain the observed set of read sequences using a minimal set of constituent genomes. In order to do this, we need a modeling approach that encourages sparse solutions, without imposing an *a priori* cap on the number of constituents.

Because our ultimate goal is to estimate abundances and other parameters together with confidence intervals, MetaQuant uses a Bayesian modeling approach. In a Bayesian model, we assume a prior distribution for each of the parameters; the goal of model fitting is to infer a posterior distribution for each parameter, given the observed data. Maximum *a posteriori* (MAP) estimates for the parameters and confidence intervals then follow directly from the posterior distributions.

4 Target sampling with the Chinese restaurant process

An approach that has become popular in text mining, genome assembly, and other fields is to model the observed data as being generated by a stochastic procedure known as the Chinese restaurant process (CRP) [7, 8]. It is based on the metaphor of a Chinese restaurant containing an infinite number of tables of infinite capacity, in which a series of N customers enters one at a time. The first customer sits at a table chosen at random. Each succeeding customer sits at an occupied table with a probability proportional to the number of diners already sitting there; or chooses an unoccupied table with probability proportional to a parameter α. After $n - 1$ customers have taken seats, if n_i is the number of diners sitting at table i, the n^{th} customer sits at some occupied table j with probability $\frac{n_j}{n-1+\alpha}$, or at an unoccupied table with probability $\frac{\alpha}{n-1+\alpha}$.

The CRP naturally groups observations into clusters, if we associate customers with observations and tables with clusters. The parameter α controls the sparsity of the clusters; it can be set large to produce a large set of small clusters, or set small to produce a small number of large clusters. The cluster sizes fall off according to a power-law distribution, whose rate parameter depends on α. Since species populations in complex microbial samples also tend to follow power laws, the CRP is an atractive approach for modeling relative target abundances.

To turn the CRP into a tractable model for the relative abundances, we use an equivalent formulation called a stick-breaking construction [9]. Imagine that we have a stick of length 1. We sample a value V_1 between 0 and 1 from a $\mathrm{Beta}(1, \alpha)$ distribution and break the stick at position V_1. Let $g_1 = V_1$ be the length of this portion; then the remainder of the stick has length $1 - V_1$. Now we sample a second value V_2 from the beta distribution and break off a fraction V_2 of the remainder of the stick; this yields pieces with length $g_2 = V_2(1 - V_1)$ and $(1 - V_1)(1 - V_2)$. We continue this process recursively, yielding a series of fragments of length $g_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$. It can be proven that this construction produces an infinite series of g_i values that sums to 1, and that these are equivalent to the sampling probabilities in the CRP.

The CRP can be used to model two different stages in our generative model, as diagrammed in Figure 4. In the first stage, a "seed" genome S is sampled from a reference database $\mathcal{S}$, with probability derived from a stick-breaking construction of hidden variables $U_i, i = 1, \ldots, |\mathcal{S}|$, with sparsity parameter α. In the second stage, a target variant T is sampled from a cluster of genomes derived clonally from the seed genome, according to probabilities generated by another stick-breaking construction of variables V_i, with sparsity parameter β. The number of targets generated from each seed is unbounded, and the total number of seeds is bounded only by the size of the database $|\mathcal{S}|$; but the *likely* values of each number become smaller as the associated CRP sparsity parameters decrease.

Under this "nested CRP" model, the relative abundances of target sequences ρ_{st} are generated by:

$$\begin{aligned}
U_i &\sim \mathrm{Beta}(1, \alpha) && i = 1, 2, \ldots, |\mathcal{S}| \\
V_i &\sim \mathrm{Beta}(1, \beta) && i = 1, 2, \ldots \\
\rho_{st} &= U_{i(s)} \prod_{j=1}^{i(s)-1} (1 - U_j)\ V_{i(t)} \prod_{j=1}^{i(t)-1} (1 - V_j)
\end{aligned}$$

where $i(s)$ is the index of seed sequence s in the set of all seed sequences $\mathcal{S}$, and $i(t)$ is the index of target sequence t in the set of all variants of s.

Each target variant is associated with a phylogenetic distance (branch length) parameter ν_t, which is drawn from a Gamma(ζ_1, ζ_2) distribution. The branch length is used to compute substitution probabilities between the aligned bases in the reference sequence and the variant sequence, according to a continuous time-reversible Markov chain model of molecular evolution [10]. Although it is unlikely that the variant genomes are descended directly from a reference seed genome, time-reversible models generate the same probabilities for this scenario as for the more likely scenario in which the variants and reference seed all derive from a common ancestor. The Markov chain rate parameters κ_{b_1,b_2} and equilibrium frequencies θ_b, for $b,\ b_1,\ b_2\ \in\ \{A, C, G, T\}$ are assumed to be shared by the variants of each seed genome, though not necessarily between seed genomes. We will discuss further details of the model for variant sequence evolution in section 6 below.

5 Bayesian modeling of position bias

As in the eXpress model, we want to capture the dependence of the fragment starting position within the target sequence on the sequence context surrounding the position. This requires us to go into more detail about how eXpress models position bias, and how this can be translated into a Bayesian framework. In eXpress, the probability of starting or ending a fragment at a particular position is assumed to depend on the sequence in a 21-base window centered on that position; this assumption comes from the empirical observation that certain combinations of bases are more likely to occur at fragment ends than others. Figure 5, taken from [2], illustrates the observed sequence biases in a set of paired-end reads from a yeast RNA-Seq experiment. Panel A shows sequence logos representing the nucleotide distributions in the windows surrounding the $5'$ and $3'$ ends of the fragments. Panel B shows normalized nucleotide distributions, after factoring out differences in abundance between transcripts, and panel C shows the position-independent background nucleotide frequencies for the yeast transcriptome. The bias weights w_p^{tl} are shown in panel D; these are the ratios of the abundance-corrected nucleotide frequencies to the background frequencies at each position p. The actual bias model in eXpress is somewhat more complicated than this, in that weights are calculated based on a third-order Markov chain model for each position in the 21-base window; the overall weight is a product of tetramer probabilities at each position, divided by the expected tetramer frequencies.

In MetaQuant, we can follow a similar model, except that we want to infer distributions for the bias parameters rather than ML estimates. We assume that π_p^{tl}, the probability that a fragment starts at position p given the target t and fragment length l can be factorized as follows:

$$\pi_p^{tl} = \prod_{x=-10}^{10} \psi^{5'}_{x,t[p+x-3:p+x]} \psi^{3'}_{x,t[p+x+l-4:p+x+l-1]}$$

where $t[i:j]$ is the subsequence of the target between positions i and j inclusive. The parameters $\psi^{5'}_{x,b_1b_2b_3b_4}$ and $\psi^{5'}_{x,b_1b_2b_3b_4}$ are grouped into sets of categorical probabilities for each position and preceding 3-mer, such that

$$\sum_{b_4 \in \{A,C,G,T\}} \psi^{5'}_{x,b_1b_2b_3b_4} = 1$$
$$\sum_{b_4 \in \{A,C,G,T\}} \psi^{3'}_{x,b_1b_2b_3b_4} = 1$$

To put these into a conjugate Bayesian framework, we specify that each group of parameters $\psi^{5'}_{x,b_1b_2b_3b_4}$ or $\psi^{3'}_{x,b_1b_2b_3b_4}$ for $b_4 \in \{A, C, G, T\}$ is drawn from a Dirichlet prior distribution with hyperparameter 4-vector $\gamma^{5'}_{x,b_1b_2b_3}$ or $\gamma^{3'}_{x,b_1b_2b_3}$ respectively. Using a conjugate prior greatly simplifies the calculations when we fit the model to data using an approach such as variational Bayes.

As in the eXpress model, sequence position bias affects the target sampling probability τ^l_t. Longer sequences are more likely to be sampled, as are sequences with higher fragment selection probabilities. We account for both those effects by making $\tau^l_t \propto \rho_{st} \sum_p \pi^{tl}_p$. Normalizing this expression gives us:

$$\tau^l_t = \frac{\rho_{st} \sum_{p=1}^{l(t)-l+1} \pi^{tl}_p}{\sum_{t' \in \mathcal{T}} \rho_{s't'} \sum_{p'=1}^{l(t')-l+1} \pi^{t'l}_{p'}}$$

6 Sequence evolution and sequencing errors

As we mentioned earlier, eXpress models variations between reference target sequences and observed read sequences as being entirely due to sequencer errors. We need to account for sequencing errors in the MetaQuant model as well; however, the modeling is complicated by the fact that some of the variation is due to evolutionary divergence of the "true" target sequence from the reference sequence. The true target sequence is not observed directly; it is inferred by grouping the observed read pairs into clusters that share common sets of variations. The variations between reads within a cluster are explained by sequencing errors.

To describe the evolution of the target variant sequences from the seed reference sequences, we use the Tamura-Nei (TN93) model [10, 11]. The TN93 model imposes more restrictions on mutation rates than the generalized time-reversible (GTR) Markov chain model; but unlike the GTR model, allows us to express the nucleotide substitution probabilities as closed-form functions of the branch length ν_t. It is characterized by three rate parameters, which are different for each seed sequence s but assumed to be the same for all target variants of s: the general substitution rate κ^s_S and parameters κ^s_R and κ^s_Y for purines and pyrimidines respectively. Together with the equilibrium frequencies θ^s_A, θ^s_C, θ^s_G and θ^s_T, also dependent on the seed sequence s, with $\theta^s_A + \theta^s_C + \theta^s_G + \theta^s_T = 1$, these define the instantaneous rate matrix R^s:

$$\begin{array}{ccccc}
 & A & C & G & T \\
A & - & \theta^s_A \kappa^s_S & \theta^s_A(\kappa^s_S + \kappa^s_R/\theta^s_R) & \theta^s_A \kappa^s_S \\
C & \theta^s_C \kappa^s_S & - & \theta^s_C \kappa^s_S & \theta^s_C(\kappa^s_S + \kappa^s_Y/\theta^s_Y) \\
G & \theta^s_G(\kappa^s_S + \kappa^s_R/\theta^s_R) & \theta^s_G \kappa^s_S & - & \theta^s_G \kappa^s_S \\
T & \theta^s_T \kappa^s_S & \theta^s_T(\kappa^s_S + \kappa^s_Y/\theta^s_Y) & \theta^s_T \kappa^s_S & -
\end{array}$$

In the above, $\theta^s_R = \theta^s_A + \theta^s_G$, $\theta^s_Y = \theta^s_C + \theta^s_T$, and the diagonal entries are understood to be set so that each row sums to zero.

The substitution probability matrix $\omega^s(\nu_t)$ is derived as a matrix exponential: $\omega^s(\nu_t) = e^{R^s \nu_t}$. The (i, j) entry $\omega^s_{ij}(\nu_t)$ is the conditional probability of finding base j at some position in a target sequence at branch length ν_t from the seed, given that the seed sequence has base i at that position. It can be expressed concisely as:

$$\omega^s_{ij}(\nu_t) = \delta_{ij} e^{-(\kappa^s_i + \kappa^s_S)\nu_t} + \frac{\theta^s_j \epsilon_{ij}}{\sum_k \theta^s_k \epsilon_{jk}} e^{-\kappa^s_S \nu_t}(1 - e^{-\kappa^s_i \nu_t}) + \theta^s_j(1 - e^{\kappa^s_S \nu_t})$$

where κ^s_i is κ^s_R or κ^s_Y according to whether base i is a purine or pyrimidine; δ_{ij} is the Kronecker delta (1 if and only if $i = j$); and ϵ_{ij} is the “Watson-Kronecker” delta (1 if and only if bases i and j are both purines or both pyrimidines).

To model the production of the observed sequence fragments f from the target sequence t, subject to sequencing errors, we follow a Bayesian version of the approach used in eXpress. The probabilities of each type of error - substitutions, insertions and deletions - are defined by sets of parameters. Given a pair of aligned target and fragment sequences t and f, the substitution parameters $\chi^S_{f[k],t[k-1],t[k]}$ indicate the probability of observing fragment base $f[k]$ at alignment position k, conditional on the *pair* of bases $t[k-1 : k]$ in the target sequence. This formulation is based on the observation that, for many NGS platforms, miscall error rates depend on the true base as well as the immediately preceding ($5^{'}$) base. There are 16 groups of 4 χ^S_{hij} parameters, each with the constraint $\sum_h \chi^S_{hij} = 1$. Each group of χ^S_{hij} parameters is distributed according to a 4-dimensional Dirichlet prior with hyperparameter ξ^S_{ij}. Insertion parameters χ^I_m are categorical probabilities of insertions of length m for $m \in \{0, 1, \ldots, I_{\max}\}$, with the constraint that $\sum_{m=0}^{I_{\max}} \chi^I_m = 1$. They are associated with an $I_{\max}$+1-dimensional Dirichlet prior with hyperparameter ξ^I. Similarly, deletion parameters χ^D_m are categorical probabilities of deletions of length m for $m \in \{0, 1, \ldots, D_{\max}\}$, with the constraint that $\sum_{m=0}^{D_{\max}} \chi^D_m = 1$; and are associated with a $D_{\max} + 1$-dimensional Dirichlet prior with hyperparameter ξ^D.

We are now left with the problem of combining genome evolution and sequencing errors into our generative model. Since the target sequences are not observed directly, we treat them as latent variables in the model. For now, we will assume that indels in the observed fragment sequences arise exclusively from sequencing errors rather than from divergence between the seed and target sequences; this assumption will be revisited later once we have applied MetaQuant to real datasets with quasispecies variation. This assumption conveniently allows us to use a common coordinate system for seed-fragment and target-fragment alignments, an example of which is diagrammed in Figure 6. Each successive aligned element is assigned an index, as shown in the figure; blocks of inserted or deleted bases are indexed as atomic elements rather than by their individual base positions.

If we treat the elements of the seed - target - fragment alignment as being independent of one another, we can express the conditional probability of observing the fragment sequence as:

$$\phi_f^{pstl} \equiv \prod_{k \in \mathcal{P}(s,f)} \omega^s_{t[k-1],s[k-1]} \omega^s_{t[k],s[k]} \chi^S_{f[k],t[k-1],t[k]} \prod_{k \in \mathcal{I}(s,f)} \chi^I_{l(k)} \prod_{k \in \mathcal{D}(s,f)} \chi^D_{l(k)}$$

where $\mathcal{P}(s,f)$ is the set of paired bases in the alignment; $\mathcal{I}(s,f)$ is the set of inserted blocks; $\mathcal{D}(s,f)$ is the set of deleted blocks; and $l(k)$ is the length of indel block k.

7 The complete data likelihood

We now have all the ingredients to write down the complete data likelihood, which is the first step required for model fitting using variational Bayes, EM or other techniques. In terms of conditional probabilities, it is:

$$\begin{aligned}
\mathcal{L}(\mathbf{l}, \mathbf{p}, \mathbf{s}, \mathbf{t}, \mathbf{u}, \mathbf{v}, \nu | \mathcal{F}, \lambda, \alpha, \beta, \zeta, \psi, \chi, \kappa, \theta) = \prod_{n=1}^{N} & \cdot \mathbb{P}[L_n = l_n | \lambda] \\
& \cdot \mathbb{P}[S_n = s_n, T_n = t_n | L_n = l_n, \mathbf{V} = \mathbf{v}, \mathbf{U} = \mathbf{u}] \\
& \cdot \mathbb{P}[\mathbf{V} = \mathbf{v} | \beta] \mathbb{P}[\mathbf{U} = \mathbf{u} | \alpha] \\
& \cdot \mathbb{P}[\nu = \nu_{t_n} | \zeta] \\
& \cdot \mathbb{P}[P_n = p_n | S_n = s_n, T_n = t_n, L_n = l_n, \psi] \\
& \cdot \mathbb{P}[F_n = f_n | P_n = p_n, S_n = s_n, T_n = t_n, L_n = l_n, \nu = \nu_{t_n}, \theta, \kappa, \chi]
\end{aligned}$$

The conditional probabilities are as follows:

$$\mathbb{P}[L_n = l_n | \lambda] = \lambda_{l_n}$$

$$\begin{aligned}
\mathbb{P}[P_n = p_n | T_n = t_n, L_n = l_n, \psi] &\equiv \pi^{t_n l_n}_{p_n} \\
&= \prod_{x=-10}^{10} \psi^{5'}_{x, t_n[p_n+x-3:p_n+x]} \psi^{3'}_{x, t_n[p_n+x+l_n-4:p_n+x+l_n-1]}
\end{aligned}$$

$$\begin{aligned}\mathbb{P}[S_n = s_n, T_n = t_n | L_n = l_n, \mathbf{V} = \mathbf{v}, \mathbf{U} = \mathbf{u}] \equiv \tau_{t_n}^{l_n} \\ = \frac{\rho_{s_n t_n} \sum_{p=1}^{l(t_n)-l_n+1} \pi_{p_n}^{t_n l_n}}{\sum_{s',t'} \rho_{s't'} \sum_{p'=1}^{l(t')-l_n+1} \pi_{p'}^{t' l_n}}\end{aligned}$$

$$\begin{aligned}
&\rho_{s_n t_n} = U_{i(s)} \prod_{j=1}^{i(s)-1} (1 - U_j)\ V_{i(t)} \prod_{j=1}^{i(t)-1} (1 - V_j) \\
&\mathbb{P}[V_i = v_i | \beta] = \beta (1 - V_i)^{\beta - 1} \qquad\qquad i = 1, 2, \ldots \\
&\mathbb{P}[U_i = u_i | \alpha] = \alpha (1 - U_i)^{\alpha - 1} \qquad\qquad i = 1, 2, \ldots, |\mathcal{S}| \\
&\mathbb{P}[\nu = \nu_{t_n} | \zeta] = \frac{1}{\Gamma(\zeta_1)} \zeta_2^{\zeta_1} \nu_{t_n}^{\zeta_1 - 1} e^{-\zeta_2 \nu_{t_n}}
\end{aligned}$$

$$\begin{aligned}\mathbb{P}[F_n = f_n | P_n = p_n, S_n = s_n, T_n = t_n, L_n = l_n, \nu = \nu_{t_n}] \equiv \phi_{f_n}^{p_n s_n t_n l_n} \\ = \prod_{k \in \mathcal{P}(s,f)} \omega_{t_n[k-1], s_n[k-1]}^{s_n} \omega_{t_n[k], s_n[k]}^{s_n} \chi_{f_n[k], t_n[k-1], t_n[k]}^{S} \prod_{k \in \mathcal{I}(s,f)} \chi_{l(k)}^{I} \prod_{k \in \mathcal{D}(s,f)} \chi_{l(k)}^{D}\end{aligned}$$

where the substitution matrix elements are given by

$$\omega_{ij}^{s_n}(\nu_{t_n}) = \delta_{ij} e^{-(\kappa_i^{s_n} + \kappa_S^{s_n}) \nu_{t_n}} + \frac{\theta_j^{s_n} \epsilon_{ij}}{\sum_k \theta_k^s \epsilon_{jk}} e^{-\kappa_S^{s_n} \nu_{t_n}} (1 - e^{-\kappa_i^{s_n} \nu_{t_n}}) + \theta_j^{s_n} (1 - e^{\kappa_S^{s_n} \nu_{t_n}})$$

8 Next steps

Although the complete data likelihood appears complex, it is a product of (mostly) simple factors, each based on an exponential family distribution. Therefore, it should be straightforward to express the complete data log likelihood (CDLL) as a dot product of sufficient statistics and natural parameters, plus a log normalizer, and to fit the parameters using an online variational Bayes algorithm. The only problematic factor is the target sampling probability $\tau_{t_n}^{l_n}$, which uses a sum over possible fragment positions in the target to account for positional bias. We can take a number of approaches to deal with this; for example, we could alternate fitting the positional bias parameters with fitting the abundances and target sampling probabilities, in each case leaving the other set of parameters fixed. This is essentially the eXpress approach. Therefore, the next stage of this project will be to rewrite a simplified version of the CDLL in exponential family form, with conjugate priors for the associated parameters, and apply the well-developed machinery of variational Bayes methods to design an algorithm for fitting posterior distributions for those parameters.

References

[1] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, November 2012.

[2] Adam Roberts. *Ambiguous fragment assignment for high-throughput sequencing experiments*. PhD thesis, UC Berkeley, UC Berkeley, August 2013.

[3] Olivier Cappé and Eric Moulines. On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, June 2009.

[4] Human Microbiome Project Data Generation Working. PLOS ONE: Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PLoS One*, 2012.

[5] B J Haas, D Gevers, A M Earl, M Feldgarden, D V Ward, G Giannoukos, D Ciulla, D Tabbaa, S K Highlander, E Sodergren, B Methe, T Z DeSantis, The Human Microbiome Consortium, J F Petrosino, R Knight, and B W Birren. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21(3):494–504, March 2011.

[6] J Martin, S Sykes, S Young, K Kota, R Sanka, and N Sheth. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS One*, 2012.

[7] D Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983*, 1985.

[8] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, June 1995.

[9] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *arXiv.org*, October 2007.

[10] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, January 2004.

[11] K Tamura and M Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. 1993.

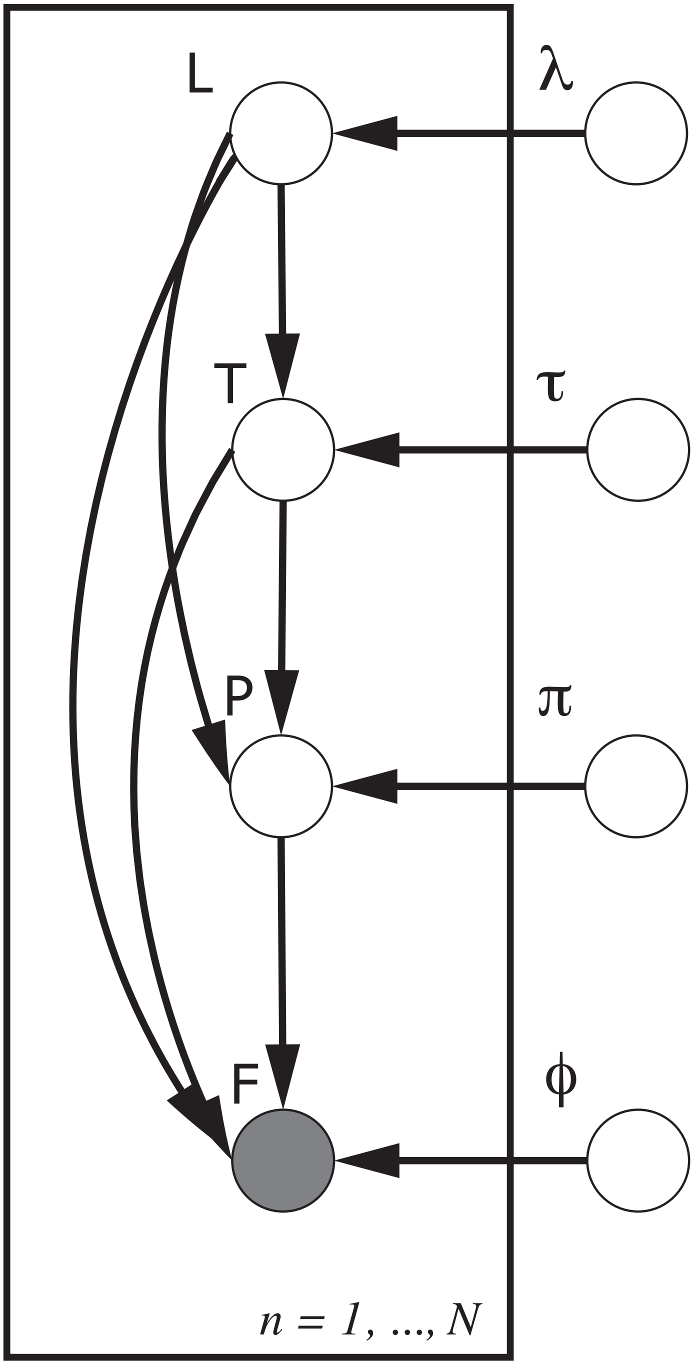


Figure 1: Graphical representation of eXpress model for RNA-Seq data

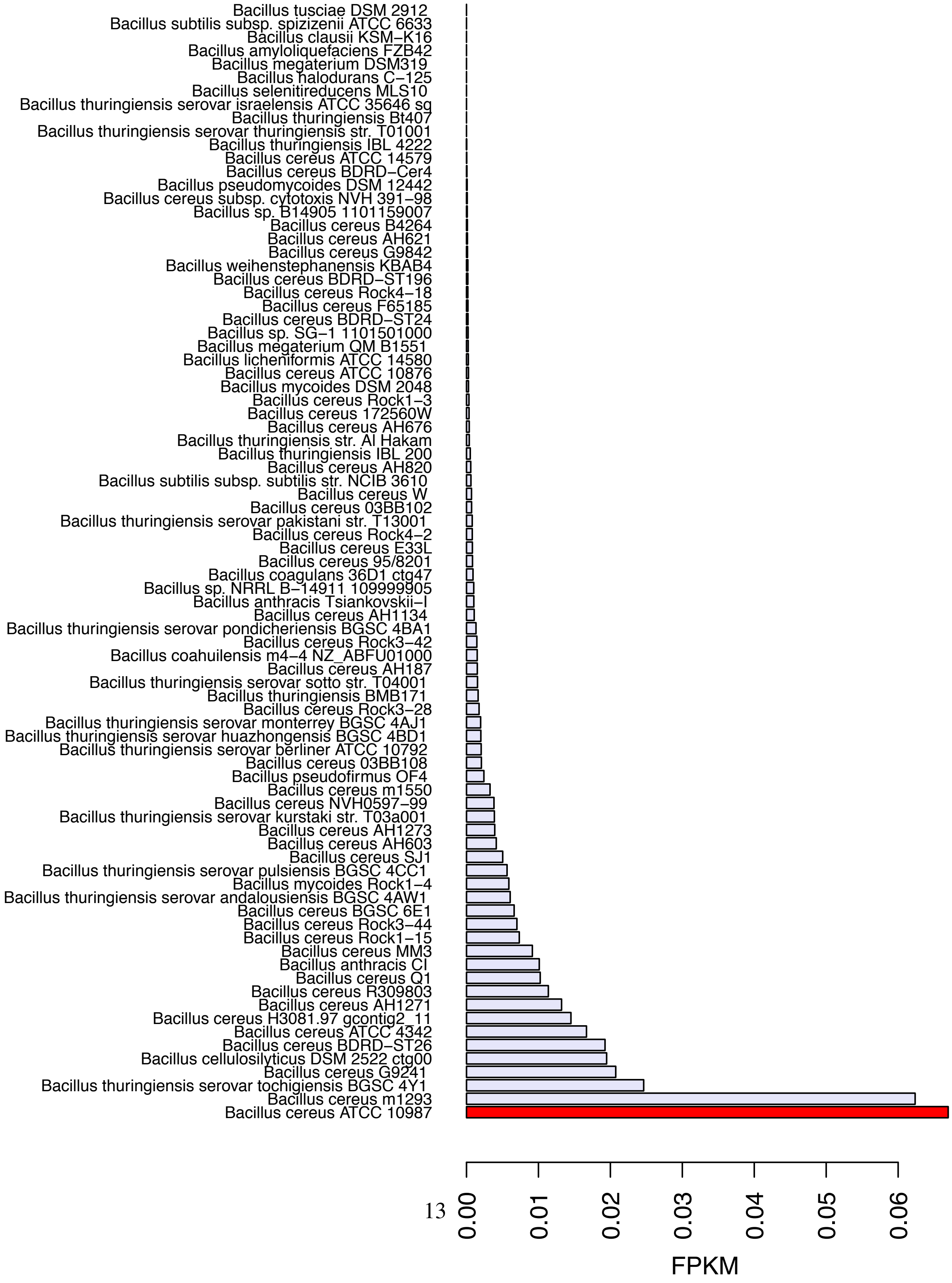


Figure 2: Abundance estimates for *Bacillus* strains computed by eXpress for mock community dataset

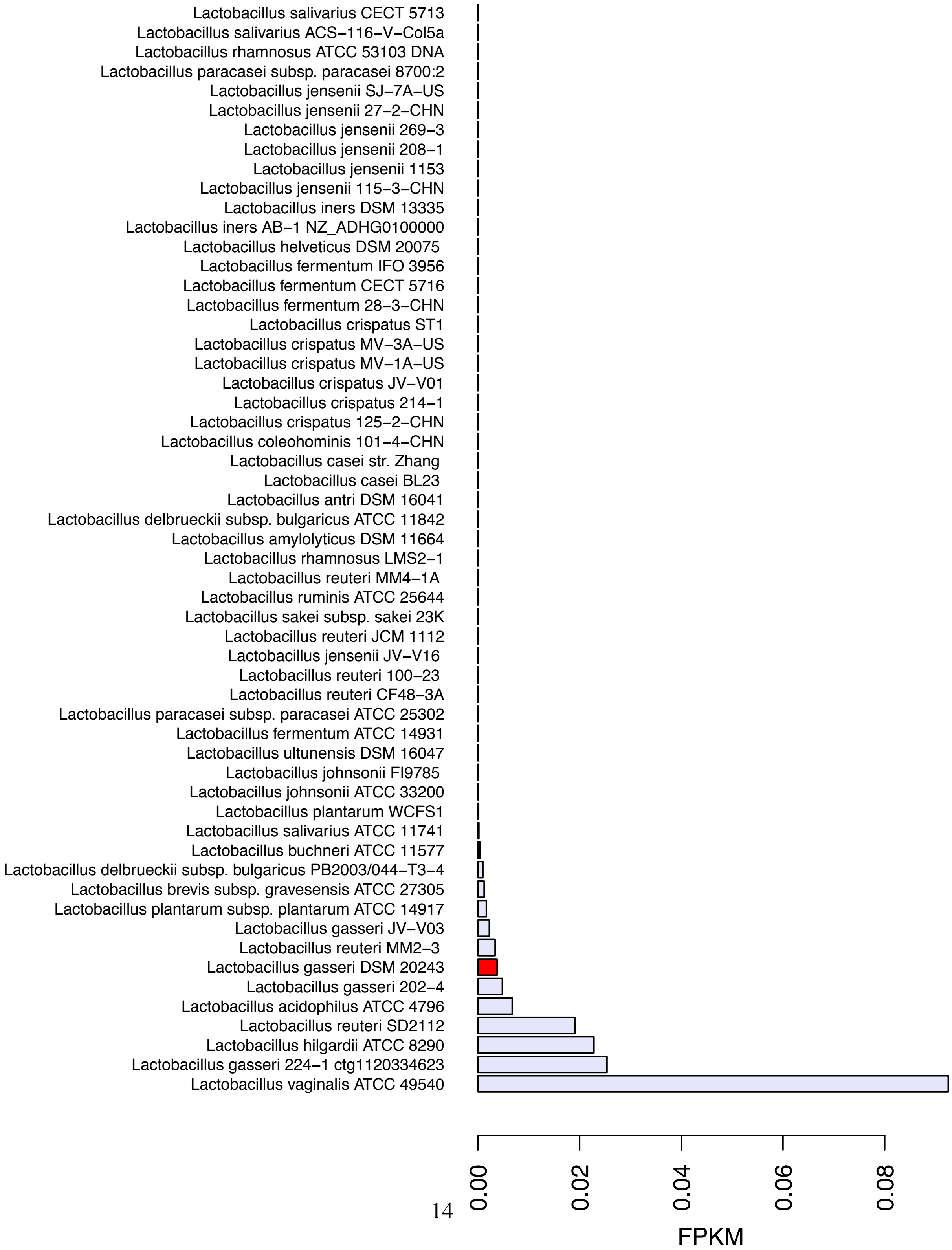


Figure 3: Abundance estimates for *Lactobacillus* strains computed by eXpress for mock community dataset

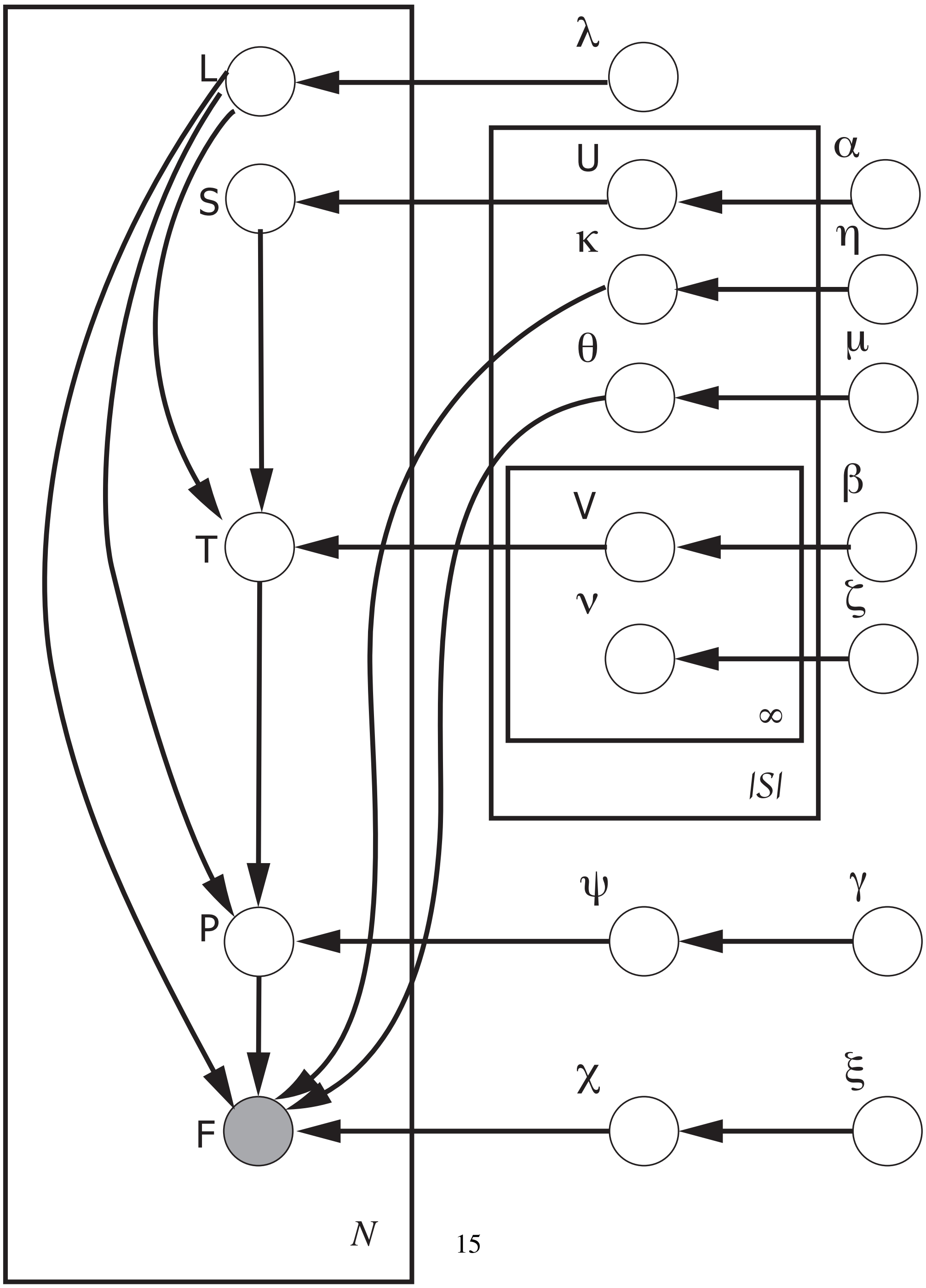


Figure 4: Graphical representation of MetaQuant generative model for metagenomic data

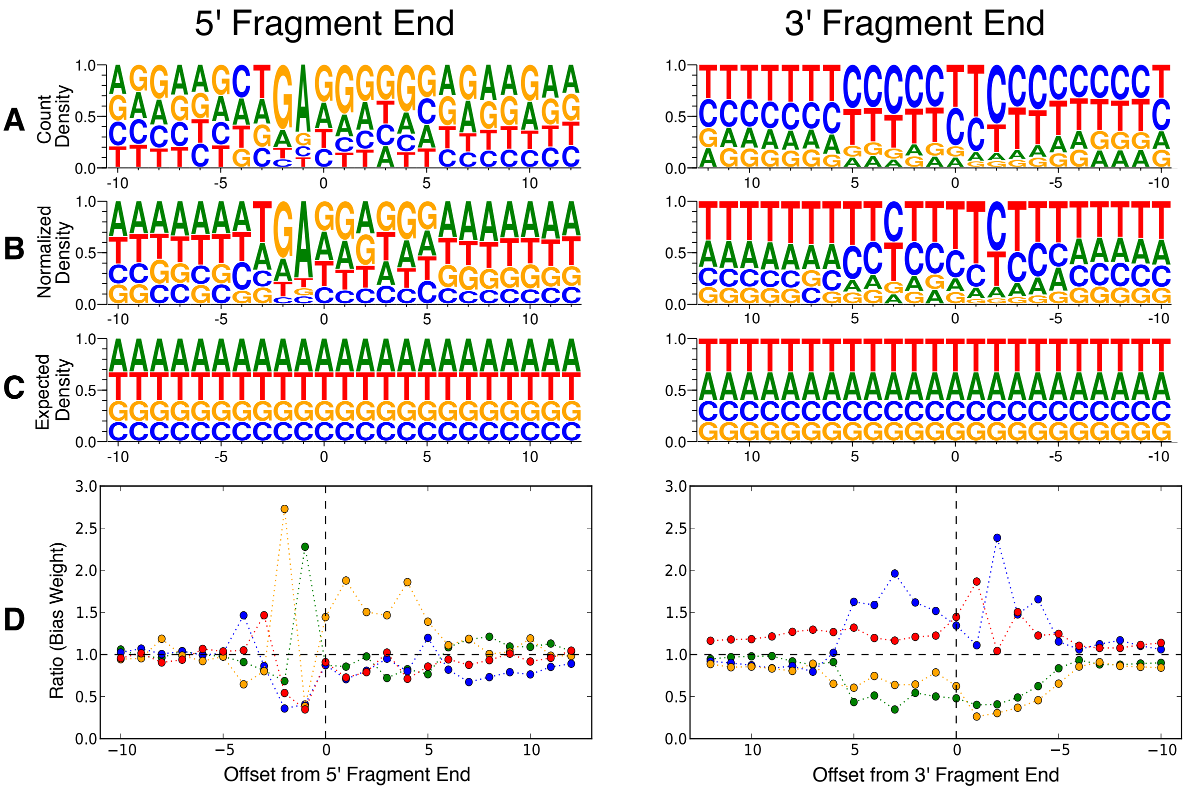
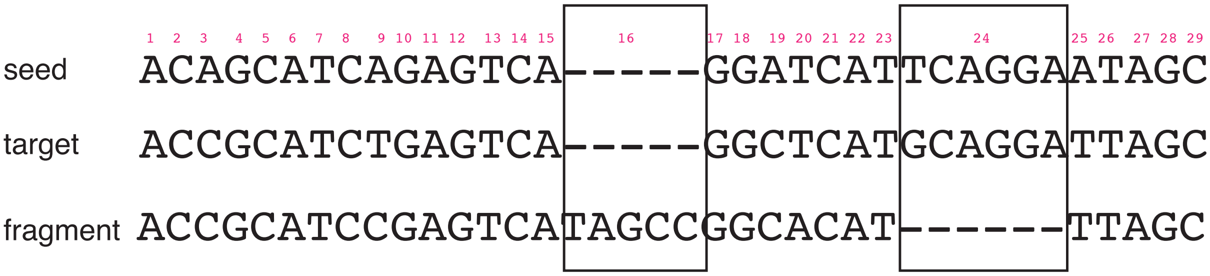


Figure 5: Nucleotide distribution surrounding fragment ends in a yeast transcriptome experiment, and calculation of position bias; from [2].



```
seed      ACAGCATCAGAGTCA-----GGATCATTCAGGAATAGC
target    ACCGCATCTGAGTCA-----GGCTCATGCAGGATTAGC
fragment  ACCGCATCCGAGTCATAGCCGGCACAT------TTAGC
```

Figure 6: Alignment coordinate system for seed, target, and fragment (read) sequences. Each aligned element (paired bases, deletions or insertions) is assigned an alignment index.