Encyclopedia of Metagenomics
DOI 10.1007/978-1-4614-6418-1_723-4

Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads

Fengzhu Sun* and Li Charlie Xia
Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Dana and David Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, USA

Synonyms

Genome Relative Abundance estimation using Mixture Model theory (GRAMMy)

Introduction

Accurate estimation of microbial community composition based on metagenomic sequencing data is fundamental for subsequent metagenomic analysis. However, it is also a challenging computational problem because of the mixed nature of metagenomes and the fact that only a small fraction of them get sequenced.

With the advents of next-generation sequencing (NGS) technologies, there has been significant increase in sequencing capacity yet reduction in single read length. This paradigm shift in sequencing technologies has impacted downstream analyses. Specifically, the identification of the origin of a read becomes more difficult for several reasons. First, a large number of short reads cannot be uniquely mapped to a specific location of one genome. Instead, they map to multiple locations of one or multiple genomes. These ambiguities are directly associated with the read length reduction in NGS technologies. Second, communities usually consist of many microbes with similar genomes, different only in some parts, making it indeed impossible to determine the origin of a particular short read based solely on its sequence.

Despite these difficulties, NGS read sets have brought in richer abundance information of microbial communities than traditional datasets because of the significant increase in the number of reads. Along with the increase of read set size, efforts to assemble more reference genomes are ongoing. In addition, new experimental techniques, such as single-cell sequencing approaches, are being developed to sequence reference genomes directly from environmental samples. In face of the challenges from short reads and the opportunities from fast-expanding reference genome databases, GRAMMy is a statistical framework developed to accurately and efficiently estimate the relative abundance of microbial organisms within the community (Xia et al. 2011).

Description

The GRAMMy Framework

The GRAMMy framework is based on a mixture model for the short metagenomic sequencing and an expectation-maximization (EM) algorithm, as outlined in the model schema and the analysis flowchart in Figs. 1 and 2. GRAMMy accepts a set of shotgun reads as well as external references

*Email: fsun@dornsife.usc.edu

DOI 10.1007/978-1-4614-6418-1_723-4

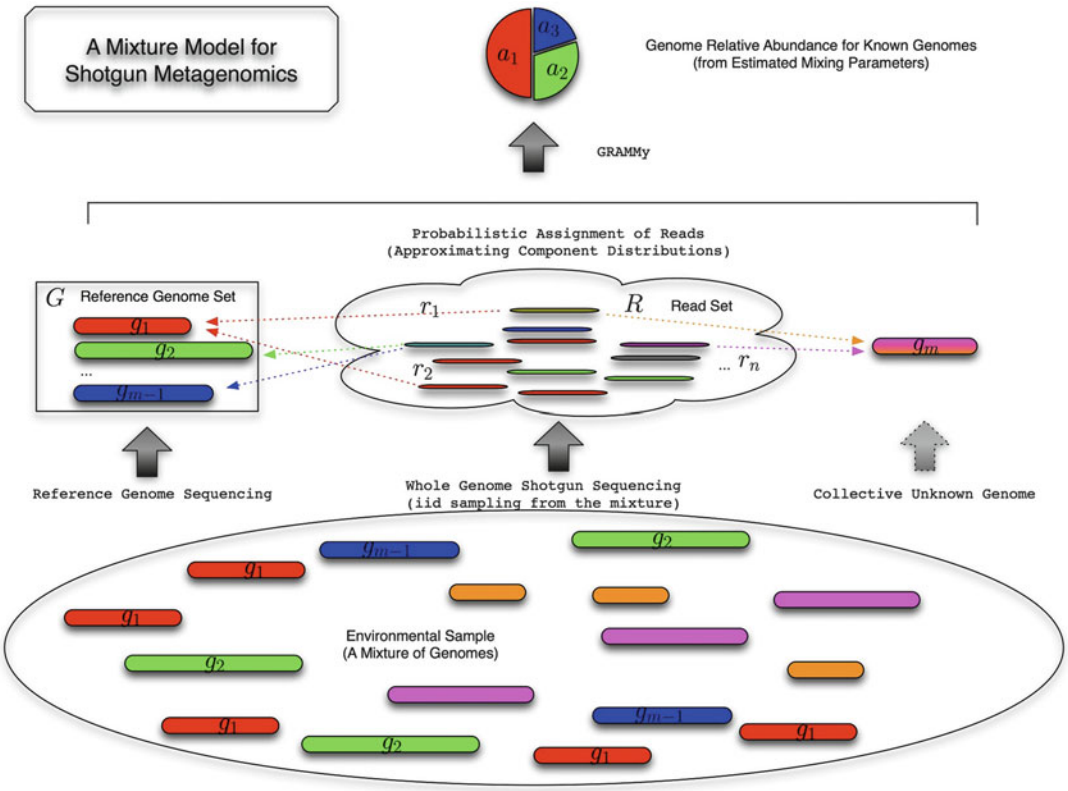

Fig. 1 The GRAMMy model. A schematic diagram of the finite mixture model underlies the GRAMMy framework for shotgun metagenomics. In the figure, "iid" stands for "independent identically distributed"

(e.g., genomes, scaffolds, or contigs) as inputs and subsequently performs the maximum-likelihood estimation (MLE) of the genome relative abundance (GRA) levels.

In the typical GRAMMy workflow, which is shown in Fig. 2, the end user starts with the metagenomic read set and reference genome set and then chooses between mapping-based ("map") and k-mer composition-based ("k-mer") assignment options (He and Xia 2007). In either option, after the assignment procedure, an intermediate matrix describing the probability that each read is assigned to one of the reference genomes is produced. This matrix, along with the read set and reference genome set, is fed forward to the EM algorithm module for estimation of the GRA levels. After the calculation, GRAMMy outputs the GRA estimates as a numerical vector, as well as the log-likelihood and standard errors for the estimates. If the taxonomy information for the input reference genomes is available, strain (genome) level GRA estimates can be combined to calculate high taxonomic level abundance, such as species- and genus-level estimates.

Accurate GRAMMy Estimates with EM Algorithm

Formally, GRA is defined as the normalized abundance for m unique genomes, where the relative abundance for the jth known genome is

$$a_j = \frac{\#\text{j-th genome}}{\#\text{known genomes}}$$

DOI 10.1007/978-1-4614-6418-1_723-4

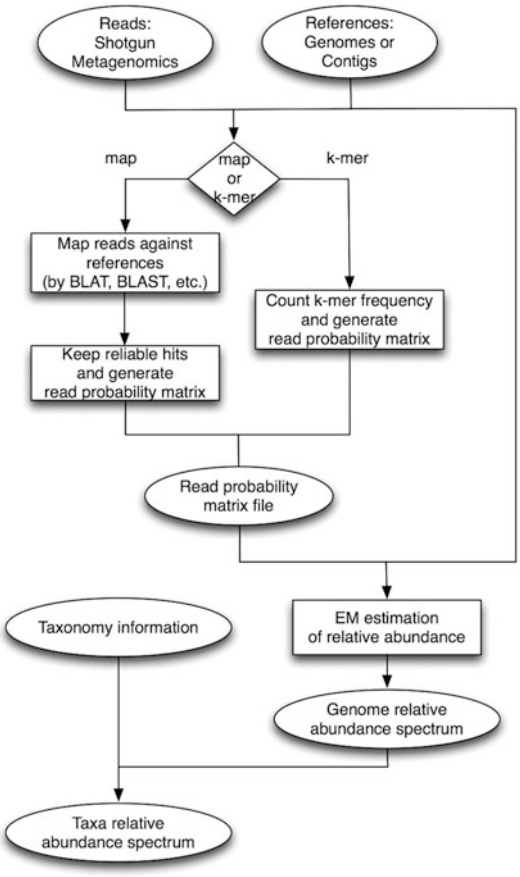

Fig. 2 The GRAMMy flowchart. A typical flowchart of GRAMMy analysis pipeline employs "map" and "k-mer" assignment

Note that g_m is a collective surrogate for unknown genomes and cannot be estimated in the model. Knowing length l_j, a_j is one-to-one related to the corresponding mixing parameter π_j by

$$a_j = \frac{\pi_j}{l_j \sum_{k=1}^{m-1} \frac{\pi_k}{l_k}}$$

Mixing component distributions are needed to solve for mixing parameter π, which are $p(r_i|z_{ij} = 1; \mathbf{g})$'s – i.e., the probabilities of generating a read r_i from g_j. They are approximated empirically. The first approach is to use the number of high-quality hits s_{ij} from BLAST, BLAT, or other mapping tools and approximate by $\frac{s_{ij}}{l_j}$; the second approach is to use k-mer composition as detailed in the original study (Xia et al. 2011). The EM algorithm to calculate π iterates between E-step

$$z_{ij}^{(t)} = \frac{p\left(r_i|z_{ij} = 1; g\right)\pi_j^{(t)}}{\sum_{k=1}^{m} p(r_i|z_{ik} = 1; g)\pi_k^{(t)}}$$

DOI 10.1007/978-1-4614-6418-1_723-4

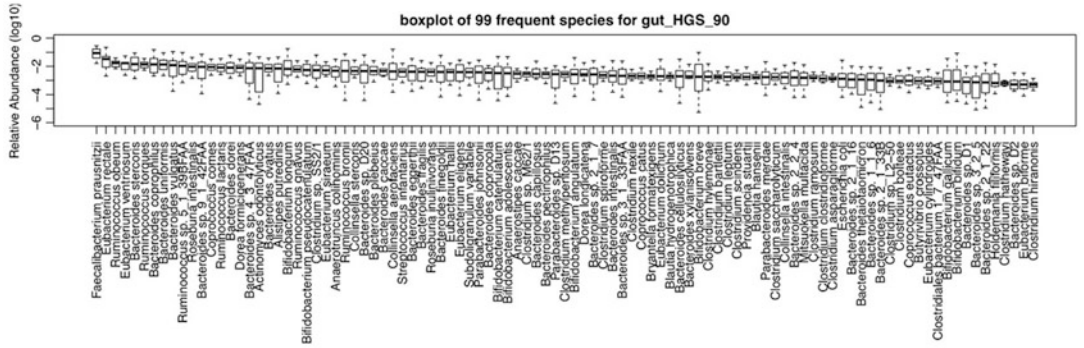

Fig. 3 Frequent species of human gut microbiome. The 99 species occurring in at least 50 % of the 33 human gut samples with a minimum relative abundance of 0.05 % were selected

and M-step

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n}$$

until convergence, where n is the total number of reads and z_{ij}'s are entries in the missing read origin matrix $\boldsymbol{Z}$. The estimated mixing parameters π are then converted back to GRA estimates **a**.

GRAMMy Estimates for Human Gut Metagenomes

The human gastrointestinal tract harbors the largest group of human symbiotic microbes. Figure 3 shows the 99 most frequent species of human gut based on the GRAMMy analysis of the 33 metagenomic samples collected from three human gut metagenome projects (Gill et al. 2006; Kurokawa et al. 2007; Turnbaugh et al. 2009). The medians of estimated average genome lengths for these metagenomes range from 2.8 to 3.7 Mbp. Among the top ten most frequent species, there are eight from the *Firmicutes* phylum including members of *Faecalibacterium*, *Eubacterium*, and *Ruminococcus* genera, and two from the *Bacteroides* genus of the *Bacteroidetes* phylum. *Firmicutes* and *Bacteroidetes* dominate the human gastrointestinal tract. Species' relative abundance displays a long-tail distribution, suggesting that many are detected across samples, though most of them are not highly abundant. The abundance levels of some species are highly variable (with larger box size), while most others are relatively constant.

Conclusions

GRAMMy is a rigorous probabilistic framework for accurately and efficiently estimating genome relative abundance (GRA) based on shotgun metagenomic reads. Users have a wide choice of mapping and alignment tools to assign reads to references. The method is particularly suitable for NGS short read datasets due to its better handling of read assignment ambiguities. GRAMMy tools are packaged as a C++ extension to Python, which can be downloaded freely from GRAMMy's homepage: http://meta.usc.edu/softs/grammy.

DOI 10.1007/978-1-4614-6418-1_723-4

Cross-References

- ▶ Approaches in Metagenome Research-Progress and Challenges
- ▶ Computational Approaches for Metagenomic Datasets
- ▶ Computational Tools for Taxonomic Assignment
- ▶ Extended Local Similarity Analysis (eLSA) of Biological Data
- ▶ Marine Bacterial, Archaeal and Protistan Association Networks
- ▶ Metagenomic Research: Methods and Ecological Applications
- ▶ Metagenomics, Metadata and MetaAnalysis
- ▶ Molecular Ecological Network of Microbial Communities

References

Gill SR, Pop M, Deboy RT, et al. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778):1355–9.

He PA, Xia L. Oligonucleotide profiling for discriminating bacteria in bacterial communities. Comb Chem High Throughput Screen. 2007;10(4):247–55.

Kurokawa K, Itoh T, Kuwahara T, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res. 2007;14(4):169–81.

Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4.

Xia LC, Cram JA, Chen T, et al. Accurate genome relative abundance estimation based on shotgun metagenomic reads. PLoS One. 2011;6(12):e27992.