Supplementary Information
`EzMap`: A Simple Pipeline for Reproducible Analysis of the Human Virome

Patrick Czeczko, Steven C. Greenway, and A.P. Jason de Koning

March 13, 2017

1 Supplementary Methods

1.1 Pipeline overview

`EzMap` accepts input sequences in standard FASTQ format. It first removes low quality reads using PRINSEQ [Schmieder and Edwards, 2011]. The remaining reads are aligned to the host genome using an aligner of the user's choice (e.g., Bowtie2 [Langmead and Salzberg, 2012] or HISAT [Kim et al., 2015]) in order to subtract out irrelevant sequences from the host. SAMTools [Li et al., 2009] is then used to aggregate reads not mapped to the host genome. These reads are queried against a database of known viral sequences using BLAST [Altschul et al., 1990]. The maximum likelihood method of Xia et al. [2011] is then used to estimate the relative abundance of each viral genome while accounting for a number of factors. The method of Xia et al. [2011] was used for consistency with several of the early papers in the field of cfDNA-viromics (e.g., De Vlaminck et al. [2013]).

1.2 Maximum Likelihood Estimation of Genome Relative Abundance

`EzMap` estimates genome relative abundance (GRA) using the EM algorithm of Xia et al. [2011], which we independently implemented in Python. The method is fully described in the original publication. In brief, it assumes that reads, r_i, represent random draws from a mixture distribution, M, which is characterized by a set of genomes, g_i, and a set of mixing coefficients, π_i. The lengths of the genomes, l_i, are assumed to be known, and the mixing coefficients are estimated by maximum likelihood. Given an estimate of the mixing coefficients, GRA is then calculated as:

$$a_i = \frac{\hat{\pi}_i}{l_i \sum_{j=1}^{m} \frac{\hat{\pi}_j}{l_j}} \tag{1}$$

for m genomes.

To apply the method, the conditional probability of a read coming from a given genome is needed. This is calculated by obtaining an estimate of the

number of copies of read i in genome j, s_{ij}, and dividing this by the genome length l_j,

$$p(r_i|z_{ij} = 1) \approx \frac{\hat{s_{ij}}}{l_j} \tag{2}$$

where z_{ij} is a random variable denoting whether read i came from genome j. Following Xia et al. [2011], we estimate s_{ij} using the number of high-quality matches (e.g., by BLAST, Bowtie2, or HISAT) for read i in genome j, filtered using an adjustable E-value cutoff (as in De Vlaminck et al. [2013]).

1.3 Expectation step

In the E-step, the conditional expectation of the set of random variables, z_{ij}, is calculated for all i and j. Following Xia et al. [2011],

$$z_{ij}^{(t)} = \frac{p(r_i|z_{ij} = 1)\pi_j^{(t)}}{\sum_k p(r_i|z_{ik} = 1)\pi_k^{(t)}} \tag{3}$$

1.4 Maximization step

In the M-step, the estimates of the mixing coefficients are updated using

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t)}}{n} \tag{4}$$

where n is the total number of reads.

The E and M steps are then cycled until convergence.

2 Supplementary Results

`EzMap` can be run in both a cluster computing environment using Slurm, or on a workstation computer. Below we provide some sample run-times for the workstation version of `EzMap`.

Table 1: Complete run-times for `EzMap` on a standard workstation computer (Intel Xeon E5-2687W with 16 cores). Reads were all 50bp Illumina reads.

Num. reads	Using HISAT2 (min.)	Using Bowtie2 (Min.)
10,000,000	5.51	5.05
15,000,000	9.55	10.92
20,000,000	9.97	9.99
25,000,000	11.26	14.93

References

Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–359, 2012.

Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 12(4):357–60, Apr 2015. doi: 10.1038/nmeth.3317.

Heng Li et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

S F Altschul et al. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 1990. doi: 10.1016/S0022-2836(05)80360-2.

Li C Xia et al. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PloS one*, 6(12):e27992, 2011.

Iwijn De Vlaminck et al. Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell*, 155(5):1178–87, Nov 2013. doi: 10.1016/j.cell.2013.10.034.