README

October 11, 2019

1 Ebsom

1.1 Overview

This repository implements a novel empirical Bayesian model for low-frequency variant calling with high-coverage pooled-sample/high-ploidy/somatic DNA sequencing data. This approach works well on simulated data but is **not currently working on the real data for which it was designed. Use with caution.**

The basic approach is to jointly model the DNA sequencing error process and the global allele-frequency distribution, obtaining a maximum-likelihood estimate of each. The inferred global distribution of allele frequencies is employed as the prior distribution in calculating the posterior distribution of allele frequencies at individual candidate loci, making this an empirical Bayesian approach.

The sequencing error model is a fully-connected artificial neural network, where for each base-call the inputs are a number of covariates associated with that base-call. The output of the error model is a log-probability of the read being called as A, C, G, or T.

1.2 Dependencies

This module requires the following Python modules:

- `tensorflow 1.14+`
- `h5py`
- `pysam`

`Cython` is also required.

1.3 Model details

Suppose we have L BAM alignment files, each representing a different library, sequenced in a single multiplexed sequencing lane. Let there be P positions in each alignment file (possibly spanning multiple chromosomes). Let the number of reads aligning to position j for library i be C_{ij}. For each base-call $\{B_{ijk}, 1 \leq k \leq C_{ij}\}$, we have compiled a vector $\vec{X}_{ijk}$ of associated covariates, which may include things like the base quality, mapping quality, read number, position along the read, and potential for contamination at the site (allowing for index-swapping amongst multiplexed libraries). Let ψ be the function that maps these covariates, together with the unobserved true base, to probabilities of A, C, G, and T, such that the probability of B_{ijk} given $\vec{X}_{ijk}$, error-model parameters θ_e, and true base Z is $\psi(B_{ijk} \mid \vec{X}_{ijk}, Z; \theta_e)$.

Let the global allele frequency spectrum (i.e., the global distribution of allele frequencies, including frequencies 0 and 1) be $f(s, \theta_a)$, where θ_a are the parameters of the allele frequency distribution. Assume that at each position in each alignment file, there is a single major and minor allele. Then the global likelihood is

$$L(\theta_e, \theta_a | \mathbf{B}, \mathbf{X}) = \Pr(\mathbf{B} \mid \mathbf{X}; \theta_e, \theta_a) = \prod_{i=1}^{L} \prod_{j=1}^{P} \int_0^1 ds f(s, \theta_a) \cdot \\ \prod_{k=1}^{C_{ij}} \left[s\psi(B_{ijk} \mid \vec{X}_{ijk}, a_{ijk}; \theta_e) + (1-s)\psi(B_{ijk} \mid \vec{X}_{ijk}, A_{ijk}; \theta_e) \right],$$

where a_{ijk} and A_{ijk} respectively represent the minor and major alleles at a given site.

This likelihood is maximized to obtain maximum-likelihood estimates $\hat{\theta}_e$ and $\hat{\theta}_a$. The distribution of allele frequencies implied by $\hat{\theta}_a$ is adopted as the empirical prior distribution, which can be used to calculate the posterior distribution of allele frequencies at a given site:

$$\Pr(s | \boldsymbol{B_{ij}}, \boldsymbol{X_{ij}}) = \frac{f(s, \hat{\theta}_a) \prod_{k=1}^{C_{ij}} \left[s\psi(B_{ijk} \mid \vec{X}_{ijk}, a_{ijk}; \hat{\theta}_e) + (1-s)\psi(B_{ijk} \mid \vec{X}_{ijk}, A_{ijk}; \hat{\theta}_e) \right]}{\int_0^1 dt f(t, \hat{\theta}_a) \prod_{k=1}^{C_{ij}} \left[t\psi(B_{ijk} \mid \vec{X}_{ijk}, a_{ijk}; \hat{\theta}_e) + (1-t)\psi(B_{ijk} \mid \vec{X}_{ijk}, A_{ijk}; \hat{\theta}_e) \right]}.$$

1.4 Implementation notes

We assume that the global distribution of allele freuqencies $f(s, \theta_a)$ is a beta distribution with additional point-mass at 0, thus having three parameters: a, b, and z, where $a > 0$ and $b > 0$ are the typical beta distribution parameters, and $0 \leq z \leq 1$ is the probability that the minor allele frequency is 0. Since we assume we know the major and minor alleles at each site, we fold the distribution about $1/2$.

In practice, it is difficult to calculate the likelihood (or log-likelihood) exactly, owing to the integral inside of the products. To simplify calculation, we discretize the beta distribution to replace the integral with a sum. Discrete frequencies are more highly concentrated near 0 than 0.5, following a near-exponential spacing. Probabilities for each discrete frequency are calculated as the integral of the beta-distribution PDF between neighboring frequency midpoints.

The sequencing error function ψ takes the form of a three-layer fully connect artificial neural network with log-softmax output, so that θ_e is comprised of the weight and bias terms of each layer.

We optimize the likelihood using the stochastic optimization algorithm ADAM, a variant of stochastic gradient descent with momentum. This converges much more quickly than quasi-Newton methods like L-BFGS-B, which must evaluate the entirety of the data before updating parameters.

The minor allele at each position is assumed to be the second-most common base-call at the position. If there is a tie, it is chosen randomly from the three non-major bases during each evaluation in optimization.

All calculations are performed in log-space.