User Manual for BADGE 1.0.0 demo (Matlab)

For questions regarding using of BADGE 1.0.0, please contact gujh@vt.edu

Introduction

Bayesian Analysis of Dispersed Gene Expression (BADGE) is a computational method for RNA-Seq abundance quantification and differentially expressed gene (DEG) identification. BADGE explicitly model within-sample and between-sample variability in RNA-Seq read counts to improve numerical performance. Within-sample over-dispersion is typically caused by technical factors in sequencing technology, such as transcript length bias, GC content bias, and etc. Between-sample over-dispersion, on the other hand, is usually related to biological variations that are carried among individual samples. By using a Gibbs sampler to estimate posterior distributions of parameters in hierarchical Bayesian model, BADGE achieves improved performance in both abundance quantification and DEG identification compared to existing methods.

Citation

Jinghua Gu, Xiao Wang, Leena Halakivi-Clarke, Robert Clarke and Jianhua Xuan, "BADGE: A novel Bayesian model for accurate abundance quantification and differential analysis of RNA-Seq data", *RECOMB-seq: Fourth Annual Recomb Satellite Workshop on Massively Parallel Sequencing*, Pittsburgh, PA, USA, Mar. 31 – Apr. 1 2014.

Requirement

The Matlab package of BADGE method was developed under Win7 OS. Matlab version >=7.11.0 is recommended. The package also self-includes several external Matlab packages and functions for string operation and numerical calculations, which are:

'string' package: includes functions such as strsplit.m;

alogam.m: computes the logarithm of the Gamma function;

invpsi: computes the inverse of the Digamma (or Psi) function

Function files

Table 1. Function files in BADGE package 1.0.0

File name	Description	Source
Badge.m	Main function that implements BADGE algorithm	internal[1]
mhsample1d.m	Metropolis-Hastings sampling function that is re-developed based on Matlab function mhsample in order to sample a vector of independent parameters.	internal
calMeanU.m	Calculate the mean of parameter U, which is used by Eq. (3) in the RECOMB-seq paper.	internal
vectorize_params.m	Vectorize parameters for Gibbs sampling to boost speed	internal
Plot_ROC.m	Plot receiver-operating-characteristic curve	internal
alogam.m	Computes the logarithm of the Gamma function	external[2]
Invpsi.m	Computes the inverse of the Digamma (or Psi) function	external

1: internal functions are developed by the authors of BADGE method;
2: external functions are downloaded from public sites.

Usage

Simply execute Badge.m, which will automatically run a demo program to: 1. Simulate RNA-Seq counts based on human annotation file; 2. Estimate within-sample over-dispersion model; 3. Estimate between-sample over-dispersion model; Show trace plots of model parameters and test performance using receiver-operating-characteristic (ROC) curve and its area-under-the-curve (AUC).

1. Simulate over-dispersed RNA-Seq counts

Table 2. Parameter setting for the data simulation

Parameter	Value	Description
J1	10	Number of samples in group 1
J2	10	Number of samples in group 2
alpha	2	Gamma shape parameter for beta
alpha0	0.75	Gamma shape parameter for lambda
v	0.2	Gamma rate parameter for lambda
sigma	1.5	Within-sample over-dispersion parameter, sigma^2=1/tau

We use the parameter setting in Table 2 to simulate RNA-Seq data with both within-sample and between-sample variations. The .bed file (hg.19.test.bed) is extracted from the human annotation

file (hg19) that is downloaded from UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables), where more than 200 genes are selected for simulation study.

2. Within-sample over-dispersion model

Within-sample over-dispersion model estimate (also referred to as Poisson-Lognormal regression model) undergoes two stages: adaptive stage and stable stage. Adaptive stage is particularly useful for sampling parameters using Metropolis-Hastings sampling when no conjugate priors can be used (i.e., for U). We initially set the standard deviation of random walk proposal function to 1, which are adjusted every 100 iterations based on acceptance rate. We adjust the proposal standard deviation for 5 times (5 stages) and fix it in the stable sampling. Please see the supplementary materials for RECOMB-seq paper for more details.

3. Between-sample over-dispersion model

Similar strategy is used for sampling between-sample over-dispersion model (i.e., α and α_0 in Poisson-Gamma-Gamma model), where we learn the optimal proposal standard deviation in adaptive stage (10 stages), which are later fixed for stable sampling process. For the Gibbs sampler, we use thin=10, which means to record one sample in every 10 iterations so that to reduce auto-correlation of Gibbs samples. Finally, the first 300 samples are removed for burn-in and the remaining samples are used to calculate the posterior mean of model parameters.

4. Trace plots and ROC curves

Fig. 1 gives the trace plots of sampled model parameters (of the second model, i.e., Poisson-Gamma-Gamma model), which have quite good convergence to their ground truth values. The area-under-the-curve (AUC) of the ROC curve (Fig. 2) is 0.913.

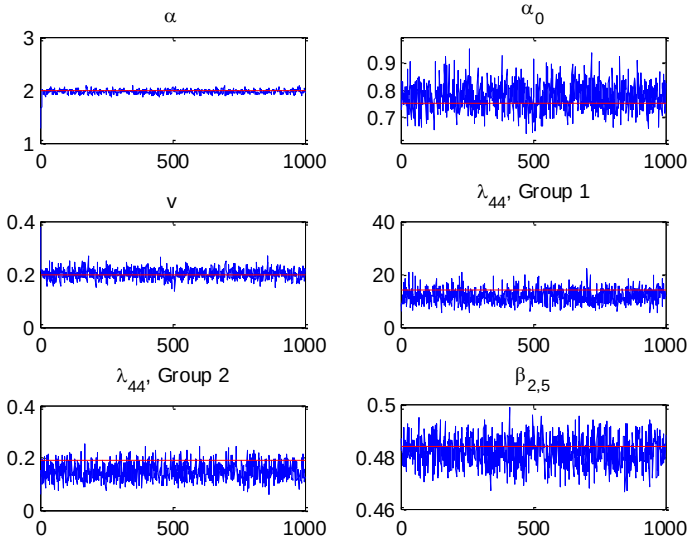

Fig. 1 Trace plots of model parameters show good convergence.

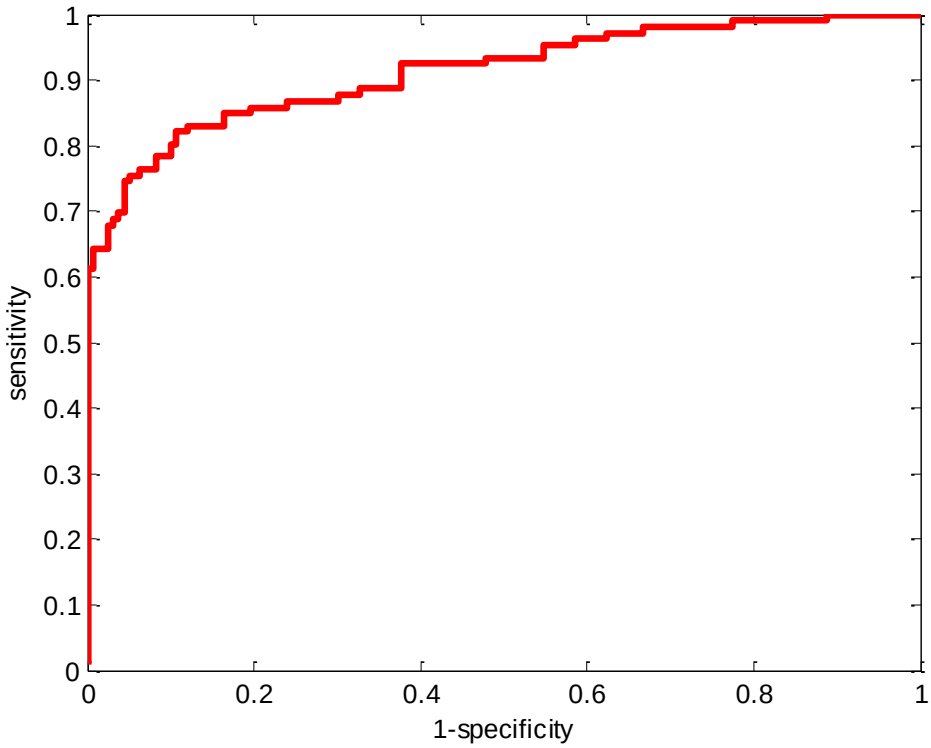

Fig. 2 ROC curve of demo code for differential gene identification.