Vireo Documentation

Release 0.2.3

Yuanhua Huang

Mar 22, 2020

Contents

CHAPTER 1

About Vireo

This document gives an introduction and usage manual of Vireo (Variational inference for reconstructing ensemble origins), a Bayesian method to demultiplex pooled scRNA-seq data without genotype reference.

Many single-cell genomics studies can be enhanced by multiplexing donors, where cells from multiple genetically distinct individuals ("donors") are assayed together in a mixed population. Such experimental designs can be powerful for reducing cost of studies assaying cells from many donors and for enabling robustness to inevitable batch effects.

When cells from multiple donors are mixed together, the donors that each cell origins are unknown at the time of sequencing. However, natural genetic variation (specifically, single nucleotide polymorphisms, SNP) act as natural barcodes capturing the donor identity of each cell. Appropriate computational methods, such as Vireo here, can infer the donor identity for each cell and thus "demultiplex" a population of cells from multiple donors in preparation of the dataset for further downstream analysis.

Vireo can use variant information extracted from single-cell RNA-seq data, which can be most platforms, including droplet, 10X genomics, smart-seq, to probabilistically assign single-cell transcriptomes to specific donor individuals, and to detect doublets that are from two different donors. Crucially, Vireo does not require any genotype reference for input donors to demultiplex, but keeps a flexibility to use the genotype information if it is available in any sub set or all donors.

Briefly, Vireo is a Bayesian hierarchical model, where the donor identity of each cell and the genotype of each donor is unknown variables. In addition, binomial base model is applied to model the minor allele reads distribution, which also accounts for the sequencing errors. A mean field variational inference is employed here approximate the joint posterior distribution of cell identity, donor genotype, and the minor allele rates.

The input data are two matrices, A and D for ALT and depth (i.e. ALT + REF), respectively, and the size in a typical experiment is 10,000 common SNPs across 10,000 cells. Computation on this size won't be fast, however, this data set is highly sparse (~99% missing values). Our impeletation of Vireo takes the benefits of sparse matrix, hence largely saves memory space and computing time.

CHAPTER 2

Quick Resources

Latest version on GitHub https://github.com/huangyh09/vireo

Scripts for simulation https://github.com/huangyh09/vireo/tree/master/simulate

All releases https://pypi.org/project/vireoSNP/#history

CHAPTER 3

Issue reports

If you find any error or suspicious bug, we will appreciate your report. Please write them in the github issues: https://github.com/huangyh09/vireo/issues

CHAPTER 4

References

Yuanhua Huang, Davis J. McCarthy, and Oliver Stegle. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. **Genome Biology** 20, 273 (2019)