Tutorial for ***demuxlet***

1 Introduction

1.1 Overview

This tutorial provides streamlined instructions for using the tool ***demuxlet***. For a more detailed description of all of the options available to use with ***demuxlet***, please refer to the README.

demuxlet is a software tool to deconvolute sample identity and identify multiplets when multiple samples are pooled by barcoded single cell sequencing. ***demuxlet*** requires the following input files:

(1) a SAM/BAM/CRAM file produced by the standard 10x sequencing platform, or any other barcoded single cell RNA-seq (with proper --tag-UMI and --tag-group) options
(2) a VCF/BCF file containing the genotype (GT), posterior probability (GP), or genotype likelihood (GL) to assign each barcode to a specific sample (or a pair of samples) in the VCF file.

1.2 Additional resources

The README for ***demuxlet*** is available here:

https://github.com/statgen/demuxlet

If you have questions about using demuxlet or suggestions for future releases, please contact chun.ye@ucsf.edu.

2 Getting Started

2.1 Installing ***demuxlet***:

2.1.1 Install htslib

```
$ git clone https://github.com/samtools/htslib.gitcd

$ cd path/to/htslib

$ autoheader

$ autoconf

$ ./configure # optional, --prefix=/path/file

$ make

$ make install # optional, DESTDIR=/path/file
```

2.1.2 Install demuxlet (demuxlet and htslib should be installed in same directory)

```
$ git clone https://github.com/statgen/demuxlet.git

$ cd /path/to/demuxlet

$ mkdir m4

$ autoreconf -vfi
$ ./configure # optional, --prefix=/path/file

$ make

$ make install # optional, DESTDIR=/path/file
```

2.2 Running ***demuxlet***:

```
$ cd demuxlet

$ ./demuxlet --sam $bam --vcf $vcf  --field $(GT or GP or PL) --out $filename
```

Add bam file name for $bam and vcf file name for $vcf. Use <(zcat $vcf) if vcf file is compressed

The options for --field are individual genotypes (GT), posterior probability (GP), or genotype likelihood (PL). If using GT option for --field, you must include --geno-error, which is the genotype error rate

2.3 ***demuxlet*** output

The ***demuxlet*** software produces three output files.

1. [prefix].best

   The .best file contains the assignments of the best sample identity (singlet, SNG-<sample name>; doublet, DBL-<sample IDs>; ambiguous, AMB-< >) in the BEST column for each cell barcode identified in the BARCODE column along with details of the statistics used to determine the best identity.

1. [prefix].single

   The [prefix].single file contains the statistics for matching each cell with each possible sample

For complete descriptions of the columns in each output file, please see the ***demuxlet*** README.

3 Analyzing the sample dataset

You can download a VCF files for Jurkat and 293T cell lines that can be used to run demuxlet on the publicly available BAM file for a 50:50 mixture of 293T and Jurkat cells from the Chromium^{TM} Cell Ranger^{TM} pipeline.

3.1 Download datasets

BAM file from 10x data for jurkat:293T sample

```
$  wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/jurkat:293t_50:50/jurkat:293t_50:50_possorted_genome_bam.bam
```

Index file for BAM file from 10x data for jurkat:293T sample

```
$ wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/jurkat:293t_50:50/jurkat:293t_50:50_possorted_genome_bam_index.bam.bai
```

VCF file with genotype data for 293T and Jurkat cells. See section 4 for original sources of VCF files.

```
$ wget   https://github.com/emccarthy23/demuxlet/blob/master/tutorial/jurkat_293T_exons_only.vcf.gz
```

3.2 Run demuxlet

```
$ cd /path/file/demuxlet
```

```
$ ./demuxlet --sam /path/file/jurkat:293t_50:50_possorted_genome_bam.bam --vcf <(zcat /path/file/293T_jurkat_merged_sorted_updated_header.vcf.gz)  --field GP --out /path/file/10x_293T_jurkat_demultiplex
```

3.3 Secondary analysis in R

In this analysis, we will use R to produce a t-SNE (t-Distributed Stochastic Neighbor Embedding) plot of the cells from the 293T:Jurkat 10x experiment with the cells colored by the assignments from the demuxlet pipeline. The analysis requires the Cell Ranger R package.

You can find instructions for downloading the Cell Ranger package here: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/rkit

3.3.1 Download and clean the Cell Ranger output

Download the data set from the 10x website and unzip the files. You need to use "--force-local" while unzipping since the file names contain a ":".

```
$ wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/jurkat:293t_50:50/jurkat:293t_50:50_analysis.tar.gz
$ wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/jurkat:293t_50:50/jurkat:293t_50:50_filtered_gene_bc_matrices.tar.gz
$ wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/jurkat:293t_50:50/jurkat:293t_50:50_raw_gene_bc_matrices.tar.gz
$ wget http://cf.10xgenomics.com/samples/cell-exp/1.1.0/jurkat:293t_50:50/jurkat:293t_50:50_metrics_summary.csv
$ tar -x -z --force-local -f jurkat:293t_50:50_analysis.tar.gz
$ tar -x -z --force-local -f jurkat:293t_50:50_filtered_gene_bc_matrices.tar.gz
$ tar -x -z --force-local -f jurkat:293t_50:50_raw_gene_bc_matrices.tar.gz
```

Clean up the file directory tree so that the files are correctly organized for the latest Cell Ranger pipeline.

```
$ mv filtered_matrices_mex outs/filtered_gene_bc_matrices
$ mv matrices_mex outs/raw_gene_bc_matrices
$ mv analysis_csv outs/analysis
$ mv jurkat:293t_50:50_metrics_summary.csv outs/metrics_summary.csv
```

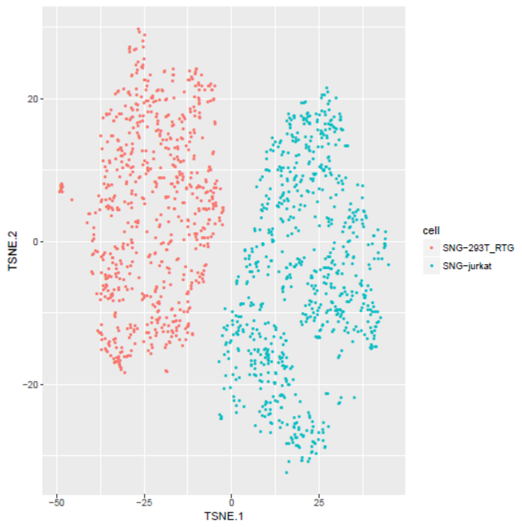
3.3.2 TSNE plot

This cleaned 10x data can be fed into the Cell Ranger pipeline. Cell Ranger provides a tutorial for creating TSNE plots and adding cell markers here:

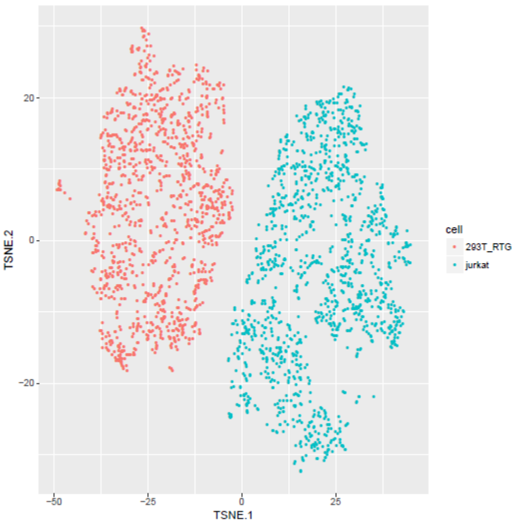
http://cf.10xgenomics.com/supp/cell-exp/cellrangerrkit-PBMC-vignette-knitr-2.0.0.pdf

In order to use the cell assignment from demuxlet to label the cells, you can use the statistics information to adjust the cutoff for singlets to provide the identification for the optimum number of cells. To illustrate this point, below we show two TSNE plots with the cells labelled by the identity from demuxlet using either the identification as "SNG-[cell-type]" in the BEST column which labels 1455/3388 cells versus using the identification of singlets by comparing the likelihood of the cell being a singlet to the likelihood of the cell being a doublet (i.e. SNG.LLK1 - LLK12 > 1) which labelled 2491/3388 cells.

Cell identification by BEST column

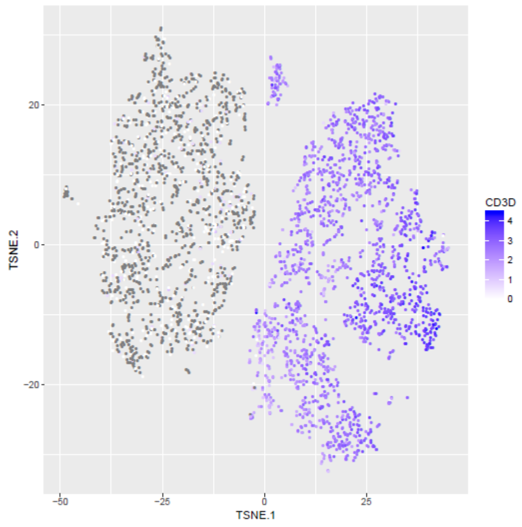


Cell identification by comparing likelihoods

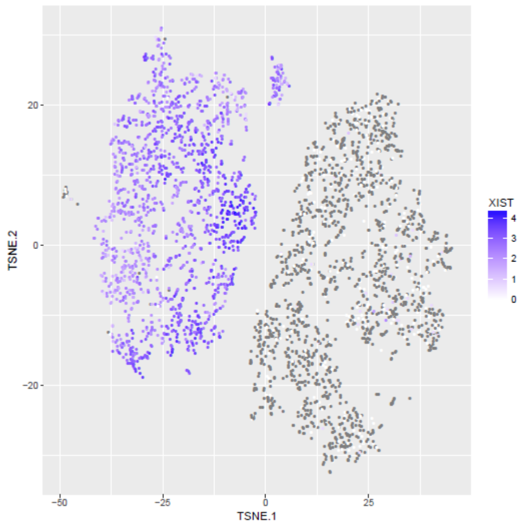


The identification of Jurkat versus 293T by demuxlet for this dataset can be verified by comparing the identification of the Jurkat and 293T clusters by expression of the CD3D and XIST markers, respectively.

Cell identification by expression of CD3D



Cell identification by expression of XIST



4 Source of 293T and Jurkat VCF file.

1. 293T VCF

   Source website:

   http://hek293genome.org/v2/data.php

Source file:
http://bioinformatics.psb.ugent.be/downloads/genomeview/hek293/SNP/293T_RTG.vcf.gz

2. Jurkat VCF

   Source website:

   https://zenodo.org/record/400615#.WYIh7IQrLIV

   Source file:

   https://zenodo.org/record/400615/files/jurkat_final_variant_calls.tar.gz

3. 293T:jurkat VCF file generation

   We used the CrossMap tool to liftover the 293T vcf file from hg18 to hg19. The tetraploid genotype for the Jurkat vcf was collapsed to a diploid genotype before being merged with the 293T vcf file and the resulting file was filtered to contain only the exon positions.