TransgeneR: A tool for transgenic integration and recombination sites discovery

Guofeng Meng

2018-03-20

Contents

0.1 Introduction

TransgeneR is designed to find the transgenic integration information in the animal genome using the whole genome sequencing data or PCR-based sequencing data. In many case, the transgenic sequences can have multiple integration sites and even recombination between transgenic sequences. Therefore, transgeneR is supposed to answer following question: * where are the transgenic sequences integrated on the genome? * Is there transgenic recombination? * How many transgenic sequences are integrated on the genome?

To use transgeneR:

```
library(devtools)
install_github("menggf/transgeneR")
```

Please note that "bowtie2" should be installed in the users' computers. And the bowtie2 geneome reference has been built the studied animals before using transgeneR.

0.2 usage

TransgeneR is an one-stop analysis pipeline.

The output of transgeneR are stored in a directory set by "output.dir".

To do the whole analysis, it has following steps:

- Build the bowtie2 reference for transgenic sequence. The output is store in a directory "insert_ref/";
- Map the homologous regions of transgenic sequences in genome. Output is a file "homo.txt";
- Reads local alignment to both genome and transgenic sequences: in this step, users need to pre-install bowtie2 and build the genome reference of studied animals. This step will generate two file: aln_genome.sam and aln_insert.sam.
- Assign the reads to genome, transgenic sequence or both and collect the clipping parts of read for second-round alignments. The output will be store in a dictory "temp_files/";
- Second-round alignments. The output are stored as "temp_files/fragment_genome.sam" and "temp_files/fragment_insert.sam";
- Connect the break sites in either transgenic sequence, genome or between transgenic sequence and genome;
- Make the plot for the split reads in the integration sites. Figures are created in "sites/*.pdf";

- If whole genome sequencing data are used, it can estimated both the full and imcomplete integration; and calculate their copy numbers. One figure is drawn as “plot_fragment.pdf”

In same case that users wish to re-run part of the analysis, user can just delete the output in mentioned steps and this will make it to skip the steps with outputs.

0.3 Output

The output are files located in “output.dir”. They have a structure of: output.dir:/

- aln_genome.sam (or aln_genome.sam.gz)
- aln_insert.sam (or aln_insert.sam.gz)
- assign.txt : read assignment in genome and transgenic sequence
- homo.txt : homologous annotation of transgeneic sequence
- mapping_summary.txt : reads alignment information
- report.txt : the predicted results
- warning.txt : the warning information
- plot_fragment.pdf the copy information of transgenic sequence
- insert_ref/ : the bowtie2 reference for transgenic sequence
- temp_files/: some temparory files
- sites/: the transgenic integration or recombination information
- – site1.pdf: the plot for first site
- – site2.pdf: the plot for second site