User Manual for Accurity

Yao Su, Zhihui Luo, Yu S. Huang

Jan 30th, 2017

1 Introduction

Accurity is a software that infers tumor purity and ploidy from a pair of tumor-normal genome sequencing data.

For citation: Z. Luo*, X. Fan*, Y.Su, YS. Huang (2017). Accurity: Accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants. Bioinformatics(submitted).

Website: http://www.yfish.org/software

2 License

2.1 FOR-PROFIT PURPOSES

If you plan to use Accurity in any for-profit application, you are required to purchase a license. To do so, please contact yuhuang@simm.ac.cn to discuss your application.

2.2 ACADEMIC NON-COMMERCIAL RESEARCH PURPOSES

Under this license (included in the software release), you can use the program for free as long as you are using Accurity strictly for non-profit purposes.

3 Install

3.1 Prerequisites

Install R, Python, samtools, and R packages **ggplot2**, **grid**, **scales**, **glmnet**, **MASS**, **RColorBrewer** and **colorspace**.

Running Accurity requires a project-specific **configure** file, details below. Change the path of R and samtools in file **configure** according to your OS environment.

3.2 Unpack Accurity.tar.gz

Uncompress the tarball by

```
tar -xvzf Accurity.tar.gz
```

Accurity is a package that contains a few binary executables, and R/Python scripts. All binary executables were compiled for a Linux platform (ubuntu 14 and 16 tested). It also contains a sample **configure** file. Denote the software full path as *Accurity_path* in file **configure**(described below).

3.2.1 A folder of reference data files

From the website above, there is another download, containing a directory *refData/* of size 13G. subfolder *refData/hs37d5/* contains GC-counting binary files about the reference genome, used for the GC-correction step. Subfolder *refData/1000g/* contains common (allele frequency >5%) SNPs from the 1000 Genome project. All these reference data is based on reference genome version hs37d5. We included the reference genome (the original hs37d5 and hs37d5 with only chromosome 1 to 22) in the tarball. Users can contact us if you need GC-counting binary files of a different reference genome.

Folder *refData/* can be be put anywhere you like. Denote the full path as *path_to_reference_data*.

4 Project organization and file configure

4.1 Project organization

For an example, you have a project directory **project_A** with a pair of matched tumor and normal samples.

```
project_A/
    sample_1_cancer.bam
    sample_1_cancer.bam.bai
    sample_1_normal.bam
    sample_1_normal.bam.bai
```

***.bam.bai** files (bam index) are not required. Accurity will use samtools to generate them if not present.

4.2 Setup the configure file

Copy the sample **configure** file from the Accurity package into your project folder and modify it accordingly. An example looks like this:

```
project_name    CRC
reference_human_genome  hs37d5
```

```
readlength_(bp) 101
window_size_(bp)        500
path_to_reference_data  /home/user/Accurity/refData
path_to_reference_genome_data   /simm/huangyulab/AccurityTestData/simulation/hs37d5_name
full_path_to_samtools   /simm/program/bin/samtools
full_path_to_freebayes  /simm/program/bin/freebayes
Accurity_path   /simm/program/cancer_purity/Accurity
```

Meaning of the fields in the **configure** file:

reference_genome_name the version of the reference genome to which the bam files are aligned to. It is assumed that all samples under the same directory are aligned to the same reference genome.

read_length the length in base pair of the read

window_size the window size in base pair for segmentation. The segmentation program (BIC-seq) first calculates the number of reads for each window and then perform segmentation over the genome. A small window size often leads to a large number of small segments. The recommended window size is 500bp.

reference_index_folder_path the directory where the reference genome index data (pre-generated by accurity) is stored. See section 3.2.

reference_genome_fasta_path path to the reference genome fasta file

samtools_path path to the samtools program.

freebayes_path path to the freebayes program.

accurity_path path to the Accurity software. See section **??**.

4.3 Run Accurity

Accurity is composed of several C++ binaries. To make it easy to run, we have written a Python wrapper, **main.py**, that wraps all binary executables in a pipeline.

Get help:
./main.py -h

Run Accurity from scratch, given two bam files and an output folder **sample_1_infer**:

./main.py -c path_to_configure_file -t sample_1_cancer.bam -n sample_1_normal.bam -o sample_1_infer

Resume Accurity from a specific step:

./main.py -c path_to_configure_file -t sample_1_cancer.bam -n sample_1_normal.bam -o sample_1_infer $-s$ S

Step S has five choices, 0 to 4, with 0 equivalent to running from the very beginning. The series of steps are: 0 Generate directory and bam index files; 1 Get read locations and heterozygous germline SNVs; 2 Correction for GC bias; 3. Segment the genome; 4 Infer purity and ploidy. Given a specific S, all steps larger than or equal to S will be carried out.

Change the default lambda value for BIC-seq:

./main.py -c path_to_configure_file -t sample_1_cancer.bam -n sample_1_normal.bam -o sample_1_infer ***−l lambda***

lambda is an integer parameter for BIC-seq, controlling the sensitivity of segmentation. Default is 4.

Override previous results:

./main.py -c path_to_configure_file -t sample_1_cancer.bam -n sample_1_normal.bam -o sample_1_infer ***−−clean 1***

4.4 Accurity output

After running Accurity, the directory structure will look like:

project_A/
sample_1_infer/
sample_1_cancer/
sample_1_cancer.bam
sample_1_cancer.bam.bai
sample_1_normal/
sample_1_normal.bam
sample_1_normal.bam.bai

The italic folders are output folders created for each bam file. All major results are stored in the output directory *sample_1_infer*. File *sample_1_infer/infer.out.tsv* contains the output. An example looks like:

```
purity  ploidy  depth
0.10256 2.5     4.9707
logL    period  best_no_of_copy_nos_bf_1st_peak first_peak_int
2.3657e+07      50      2       975
```

There are two cases where purity and ploidy can not be inferred. 1) The cancer genome contains too few somatic copy number alterations; 2) The noise level is too high, or the noise level is moderate but the sample purity is very low (<0.05).