A manual for hot_scan

Contents

What it is hot_scan? 1
Requirements 1
Availability 1
Configuring hot_scan 2
Running hot_scan 2
How To Cite 4
License and Copyright 4

What it is hot_scan?

hot_scan is a free software to detect genomic regions unusually rich (hotspot) in a given pattern via *scan statistics*.

Requirements

hot_scan is designed to run on linux-based operating system. It is implemented in Perl[1] and R[2] languages. There are only few non-standard Perl module required to run hot_scan:

1. Math::GSL::SF
2. AnyEvent::ForkManager
3. File::Path

These modules can be installed via The Comprehensive Perl Archive Network (CPAN) at http://search.cpan.org. Make sure these modules are properly installed if there are no error messages.

Availability

The source code is available at https://github.com/aholanda/psa

[1] http://www.perl.org
[2] www.r-project.org

Configuring hot_scan

After download the compressed folder **hot_scan.tar.gz** to an appropriate directory, extract it using following command: **tar -zxvf hot_scan.tar.gz**. Where,

- BY.R: a R script;
- events/ : a directory to put the events file in BED format as decribed in following section;
- test/ : a directory with some test data files;
- cs/ : a directory with mappable chromosome size file for *Mus musculus* (mm9 and mm10) and *Homo sapiens*(hg19);
- hot_scan: the main program.

Running hot_scan

Before running hot_scan, you need to first provide a file in BED[3] format within **events/** directory. Only the first three fields are required, where:

1. **chrom**: the name of the chromosome (e.g. chr1, chr2, chrM);
2. **chromStart**: the start position of the event in the chromosome;
3. **chromEnd**: the end position of the event in the chromosome.

Parameters

The main script is hot_scan and the essential parameters needed to run it are as follows:

-m: window of width *m* to scan on chromosome;

-c: file with the name and mappable chromosome size;

-e: events file name;

-o: output directory name;

-p: max parallel forking count (default: 8);

-s: significant level (default: 0.05);

-a: adjust *p-values* using Benjamini & Yekutieli (2001) (default: no).

[3] http://genome.ucsc.edu/FAQ/FAQformat.html

Notes:
(1) The chromosome size file should be tab delimited and structured as follows:
<chromName><TAB><chromSize>

For example, Human (hg19):
chr1 225280621
chr2 238204518
...

(2) The chromosome size files for *Mus musculus* (mm9 and mm10) and *Homo sapiens* (hg19) are in **cs/** folder.

(3) The events file should be a file in BED format as defined before.

Output files

The output files are put in the directory name specified by user ($-o$ option). In the output folder, there will be two folders named as **scan_out_c/** and **tracks_c/**. The log files are in the **scan_out_c/**. A BED file having the name of the corresponding events file is created in **tracks_c/** folder with all hotspot information. Thus, for each hotspot, there will be one row having the chromosome, start, end and *p-value* with the corresponding hotspot id. The BED file in **tracks_c/** folder can be upload into the Genome Browser on the Add Custom Tracks page[4].

Example of running `hot_scan` on test data

Change the working directory to that which contains the `hot_scan` package:
$ cd hot_scan

Copy events files from test directory to events directory:
$ cp test/cMyc_AIDrv.bed events/

Execute `hot_scan` using the following command:
$ perl hot_scan.pl -m 500 -c cs/mm9.txt -e cMyc_AIDrv.bed -p 15 -s 0.05 -o cMyc_AIDrv

The command line option for `hot_scan` above takes six parameters: i) a window width size (specified by $-m$); ii) chromosome size file ($-c$); name of the file[5] with events ($-e$); iv) number of fork to launch ($-p$); v) significant level ($-s$) and vi) name of the directory ($-o$) in which are reported all logs and hotspots informations in BED format file as described in former section.

[4] http://genome.ucsc.edu/cgi-bin/hgCustom
[5] the file must be inside of the **events/** directory

How To Cite

SIlva IT, Rosales RA, Holanda AJ, Jankovic M and Nussenzweig MC. Identification of chromosomal translocation hotspots via *scan statistics*.

If you find a bug, please send an email to isilva@rockefeller.edu or rrosales@usp.br.

License and Copyright

`hot_scan` is free software and it is licensed under the GNU General Public License[6]

[6] `http://www.gnu.org/copyleft/gpl.html`