TCGA miRNA scanning script: Documentation

scan_tcga_files_for_beta.awk

Introduction:

A script for locating and taking means of Beta_value fields in TCGA methylation study files. The usual mode of operation is for chromosome ID and position values to be read from an input file and the beta values falling within a defined window around the range are averaged. These notes describe the various options which can be set for runs.

Implementation:

The script is written in awk (1), a text processing language which is an integral part of all Unix-type operating systems. These include Linux dialects and MacOs X. The scripts have been developed and run under MacOS X 10.8 to 10.10 (Yosemite), but should function identically in other environments.

Input files:

All input files should have normal Unix line terminators ('\n'). Files originating from MicroSoft Excel or other sources may need pre-treatment to correct this.

(a) Chromosome ID and position file - this is the main run time parameter and should contain lines like:

```
#Chr Start     End
10   131763530 131763587
11   65414680  65414818
17   4433208   4433295
18   8707361   8707438
18   8707439   8707493
19   1030236   1030346
19   14673048  14673096
20   56273715  56273790
5    322864    322936
[...]
```

Where the values are tab separated. The lines may contain other values but by default, these will be ignored since the script only uses the first 3 columns. An exception is if the genename_field parameter (below) is set to indicate which input column contains a gene name which will then be written to the output listing. This permits outputs from programs like the DMAP (2) suite (diffmeth and identgeneloc) to be used without modification. The header line is expected, but can be omitted (see skip_lines parameter).

(b) A file giving TCGA barcodes for different conditions (primary or metastatic), e.g.

```
TCGA-BF-A1PU-01A-11D-A19D-05
TCGA-BF-A1PV-01A-11D-A19D-05
TCGA-BF-A1PX-01A-12D-A19D-05
TCGA-BF-A1PZ-01A-11D-A19D-05
TCGA-BF-A1Q0-01A-21D-A19D-05
[...]
```

This file is associated with the parameter named 'barcode_file' and has no default.

Command Line Parameters:

Required to set various values needed for script operation, some have defaults, some are necessary. See 'Example Runs' below for details on how these are used.

barcode_file: the list of barcodes associated with a condition (primary vs. metastatic) - switches script operation between those conditions. There is no default value, the script will fail with an error message if this parameter is not given.

genename_field: to indicate which input column contains a gene name to be written to the output before the beta value means. The value is established by counting columns in the input to find which value is required (1-based). The default (genename_field=0) is to omit this. In fact, any arbitrary field from the input could be written out this way, but the header will still state 'Gene'.

margin: the window either side of the region defined in the input file in which beta values will be averaged. This is to allow a wider region around differentially methylated fragments to be included in calculations. Default = 500.

tcga_tail: the file extension used for generating TCGA methylation data file names. These files are simply named by the barcode + the extension so that the complete name can be generated by the tcga_dir value (below) the barcode (from the barcode_file) and this extension. Default = ".txt".

skip_lines: number of lines to skip at top of gene list input file. This is to skip over header lines if present. Should be set to 0 if there is no header. Defaults to 1;

tcga_dir: the location of TCGA methylation data files relative to the current working directory. Can be a relative or absolute path, but should end with a '/'. Defaults to `"DNA_methylation/JHU_USC__HumanMethylation450/Level_3/".`

f**ullheader**: the default behaviour is to identify columns with the unique patient ID field from the barcode (e.g. for `TCGA-EB-A41A-01A-11D-A24V-05` the column header would be `A41A`). Setting fullheader=1 will cause the complete barcode to be used as a column header. Defaults to fullheader=0.

unseen: for diagnostic purposes. Setting this to '1' should generate a list of any files which are not found, printed after other output. Defaults to 0 = no list.

Output:

A header line of tab-separated patient identifiers from each barcode in the barcode_file list. Optionally the complete barcode can be used (fullheader=1). Then, for each chromosomnal region in the input, a tab separated list of the average of all Beta_value fields from the section of that chomosome which corresponds to (start-margin) to (end+margin).

Tab separated values are suitable for loading into spreadsheets (MicroSoft Excel) or into other downstream processing tools.

Example Runs:

In Unix environments the script outputs are written to the file 'stdout' which defaults to the terminal. Unix convention allows this to be diverted into a file with the '>' or '>>' operators, or to be piped through other commands with the '|' operator. It is useful to test the command by piping through the more utility (" | more") before doing the actual run. These strategies are not included in the examples below.

Unix convention allows long commands to be spread over multiple lines by using '\' to indicate continuation. The examples here use this because of the command length. If you type the command as a single long line, then these should be omitted.

(a) Look through a file of identgeneloc (DMAP (2)) output `3lists_hyper_geneloc.txt` for primary beta values. The gene names are in the 40th field of identgeneloc for this run, so gene names are being written to the output:

```
awk -f scan_tcga_files_for_beta.awk barcode_file=TCGA_SKCM_primary.txt \
genename_field=40 3lists_hyper_genloc.txt
```

(b) use only the top 5 lines of 3lists_hyper_genloc.txt in this case don't skip any header lines. Don't include gene names in output and do the run for metastatic barcodes:

```
head -5 3lists_hyper_geneloc.txt | awk -f scan_tcga_files_for_beta.awk \
 barcode_file=TCGA_SKCM_metastatic skip_lines=0
```

(c) same as (a) but TCGA methylation data files are in directory /home/TCGA_meth_data/:

```
awk -f scan_tcga_files_for_beta.awk barcode_file=TCGA_SKCM_primary.txt \
 genename_field=40 tcga_dir="/home/TDGA_meth_data" \
 3lists_hyper_genloc.txt
```

(d) same as (a), but don't allow a margin around the input file regions:

```
awk -f scan_tcga_files_for_beta.awk barcode_file=TCGA_SKCM_primary.txt \
 genename_field=40 margin=0 3lists_hyper_genloc.txt
```

(e) same as (a), but allow a larger margin (1000 bp before and after) around input file regions:

```
awk -f scan_tcga_files_for_beta.awk barcode_file=TCGA_SKCM_primary.txt \
 genename_field=40 margin=1000 3lists_hyper_genloc.txt
```

(f) same as (e), but print full barcodes as column headers:

```
awk -f scan_tcga_files_for_beta.awk barcode_file=TCGA_SKCM_primary.txt \
genename_field=40 margin=1000 fullheader=1 3lists_hyper_genloc.txt
```

Execution Times:

Awk is an interpreted language so its performance will never reach that of compiled code. The ground work of searching through various barcode files and TCGA expression data files is done by invoking the Unix grep and awk (again) utilities and are thus executed efficiently. On the development machine (Mac

Pro, 2.8GHz Quad-Core Intel Xeon system with 32Gb RAM) completing a search for 10 hypermethylated genes/regions through 366 metastatic barcodes took about 1 hour. In that instance, the files were on a remote server and the operation would probably be faster if they had been on a local disk. While this performance is not brilliantly fast, it is still easier, more reliable and faster than manually working through the same data.

Installation:

The script scan_tcga_files_for_beta.awk should be put in an accessible place. This could be in among the data files or in the Unix local directory /usr/local/bin/. In the example runs above, it is in with the data files. If put elsewhere then the path should be included in the script name. E.g:

`awk -f /usr/local/bin/scan_tcga_files_for_beta.awk` etc.

Runtime Problems:

Hard to know what will go wrong in the hands of others.

(a) No output produced: check that the tcga_dir path are correct. If the files can't be found then the script will fail silently.

(b) Output for beta value means is '-': no methylation values were found for that barcode and region (chromosome, start, end).

(c) No output: check that the various files have correct Unix line terminators. Performing:

```
od -a <mydatafile> | more
```

will give an extensive byte by byte display of every character in the file <mydatafile> and line endings should appear as 'nl'. If they appear as 'cr nl' or 'cr' then they will not process correctly and will need to be changed. See the tr program for efficient ways to do this.

References:

1. A. V. Aho, B. W. Kernighan, P. J. Weinberger, The AWK Programming Language, Addison-Wesley, 1988. ISBN 0-201-07981-X

2. Stockwell, P.A., Chatterjee, A., Rodger, E.J. and Morison, I.M. "DMAP: Differential Methylation Analysis Package for RRBS and WGBS data" Bioinformatics (2014) DOI: 10.1093/bioinformatics/btu126.

Contact:

Peter A. Stockwell
peter.stockwell@otago.ac.nz
Dept of Biochemistry, University of Otago,
Dunedin,
New Zealand.
29-Oct-2015