NAME

PeaKDEck v1.1 – a Perl/Tk kernel density estimator based peak caller for DNaseI-seq data.

SYNOPSIS

Numerical sorting: sorts sam format file by read position.
Sam filter: groups sam reads by chromosome, filters data for mapq & uq scores, and optionally removes PCR read duplication.
Random read selection: random selection of n reads from sam file.
Density analysis: generates continuous density track from mapped, ordered sam file.
Peak calling: identifies read density peaks in mapped, ordered sam file.
Size ordering of peaks: orders bed files by peak size score.

ARGUMENTS

Numerical sorting

Sorts sam format reads by base start position, irrespective of chromosome. Equivalent to the Linux/Unix/osx command: [sort -n --key=4,5 filename.sam > sortedFilename.sam]. For fast results, memory corresponding to ~2.5 times file size should be available.

Sam filtering

Sorts sam files by chromosome, in the order that chromosomes appear in the chromosome size file. The chromosome size file is mandatory. The chromosome size file is a plain text, tab-separated file in the format:

chr1	249250621
chr2	243199373
chr3	198022430
....	………………..
chrN	size(bp)

Mandatory settings

chromosome size file [/path/to/chromosomeSizeFile.txt]
Specifies the path to the text file containing tab-separated list of chromosome names and sizes.

Optional settings

MapQ Limit [integerValue]
Specifies a mapq cutoff score for filtering. Reads with a mapq score less than the supplied value will be removed from the resulting filtered file. By default, -q is set to zero, so no filtering for mapq scores will occur.

UQ limit [integerValue]
Specifies a UQ base mismatch score for filtering. Reads with mismatch scores greater than this value will be removed from the filtered dataset. By default, -u is set to 10000, so that no filtering by uq score will occur.

Sam header (opt) [samHeaderFile.sam]
Specifies a file containing a sam header, which if set, will be included at the beginning of the newly filtered file.

PCR delete [ON|OFF]
Allows PCR duplicate reads to be removed from sam file. Reads are considered PCR duplicates if adjacent reads have identical chromosome, start position, mapq score, and sequence. By default -PCR is set to OFF, so no filtering of PCR duplicates will occur. To detect PCR duplicates, chromosomes must be in numerical order (see Numerical sorting above).

Random read selection

Randomly selects a target number of reads from a specified sam file. Selected reads are printed to STDOUT by default.

Mandatory settings

Number [integer]

Specifies the target number of reads to be randomly selected from the given sam file. The number of reads must by a positive whole number.

Density analyzer

Creates a smoothed, unitless read density track in wig format, representing the distribution of reads in the given sam file. Sam files must be grouped by chromosome, and ordered by read start position (see Numerical sorting and Sam filtering above). The order of chromosomes in the density track is determined by the order in which they appear in the mandatory chromosome size file (see Sam filtering for chromosome size file format). By default, the results are printed to STDOUT.

Mandatory settings

chromosome size file [/path/to/chromosomeSizeFile.txt]

Specifies the path to the text file containing tab-separated list of chromosome names and sizes.

Optional settings

1/2 bin size [positiveInteger]

Specifies the one-tailed size of the smoothing bin. By default, -t is set to 150, giving a bin size of 300 bp. This value determines both the size of sampling bin, and the width of the Gaussian probability density function used to calculate read densities, and must by a positive whole number.

Step size [positiveInteger]

Specifies the size of steps by which the probability density function and sampling bin move along the genome. By default, -STEP is set to 100. Smaller step sizes proportionately increase the number of calculations carried out, and therefore the time taken for the analysis. -STEP must be a positive whole number.

Sigma [positiveInteger]

Specifies the standard deviation of the probability density function. This value determines how broadly the read density scores are spread over each sampling bin, and therefore determines the degree of smoothing that occurs. By default -d is set to 50, and must be a positive whole number.

Min thresh [positiveInteger]

Specifies a low threshold, below which read density scores won't be included in probability density function calculations. By default, -t is set to the number of reads expected to occur in the set bin size if the number of reads in the dataset were randomly distributed. All reads present in the data set will be included in the analysis if -t is set to 0. -t must be a non-negative whole number.

Max thresh [positiveInteger]

Specifies a high threshold, above which read density scores won't be included in probability density function calculations. By default, -m is set to 100000000, ensuring that no reads will be excluded from analysis in default settings.

Offset [integer]

Specifies a track offset. All positions in the resulting wig file will be offset by this value. For DNaseI-seq data, the read start sites are considered DNaseI cutting sites, and so by default, -o is set to 0. If the centre of the DNA fragment is considered the point of interest (for example, in ChIP-seq), setting -o to half the average fragment size may give a more precise depiction of signal localisation.

Peak calling

Identifies peaks in the provided sam file, and provides output in bed format to STDOUT. Sam files must be grouped by chromosomes, and ordered by read position (see Numerical sorting and Sam filtering above). The order of chromosomes in the peak file is determined by the order in which they appear in the mandatory chromosome size file (see Sam filtering for chromosome size file format).

Mandatory settings

chromosome size file [/path/to/chromosomeSizeFile.txt]

Specifies the path to the text file containing tab-separated list of chromosome names and sizes.

Optional settings

Bin size [positiveInteger]

Specifies the size of the central sampling bin. By default, -bin is set to 300, which represents the expected average feature size. -bin must by set to a positive whole number

Back size [positiveInteger]

Specifies the size of the background sampling bin. By default, -back is set to 3000, ten times the size of the central sampling bin. -back must be set to a positive whole number and must be larger than the size of the central bin.

Step size [positiveInteger]

Specifies the size of steps by which the sampling bin moves along the genome. By default, -STEP is set to 100. Smaller step sizes proportionately increase the number of calculations carried out, and therefore the time taken for the analysis. -STEP must be a positive whole number.

Flat thresh [positiveInteger]

Specifies a flat threshold for peak calling in reads per bin. When -FLAT is set, the threshold calculated by PeaKDEck for peak calling is overridden, and the value given by -FLAT is used in its place. FLAT must by a positive number.

Blueprint file (opt) [/path/to/blueprintFile.bed]

This option provides the path to a bed file which contains a list of contiguous genomic loci indicating the sites of known open chromatin sites, tagged with the number of cell types with open chromatin at that site. The format is as follows:

```
chr1    1       10099   C#1     0
chr1    10100   10330   #1      37
chr1    10331   10344   C#2     0
chr1    10345   10590   #2      4
chr1    10591   16099   C#3     0
```

where the columns respectively indicate the chromosome name, site start position, site end position, element name, and number of cell types with open chromatin at that site. When this file is provided, PeaKDEck calculates signal-to-noise ratio, and calculates the background probability distribution from sites selected from loci with no known open chromatin.

npBack [positiveInteger]

Sets the number of sites to randomly select to calculate the background probability distribution. By default this is set to 50000 sites. -npBack must be a positive whole number.

p-value [probabilityValue]

Specifies the positive limit of the probability distribution for selecting the corrected read density for peak threshold. By default, -sig is set to 0.001. -sig must be a positive number between 0 and 1.

PVAL score [ON|OFF]

Peaks are scored with the maximum corrected read density recorded during that peak by default. Setting -PVAL to ON converts this corrected read density to a probability value from the background probability distribution used to calculate the threshold. This value represents the probability that a corrected read density of that magnitude belongs to the background probability distribution.

Top peak selection

Sorts peak bed files in descending order by corrected read density score. By default, the sorted peaks are printed in bed format to STDOUT. The target file must by in bed format.

Mandatory settings

Optional settings

Number [positiveInteger]

This specifies the number of peaks to include in the resulting bed file, from the highest scoring peak downwards. By default, -n is set to ALL, and all the peaks are printed to the output file. -n must be either 'ALL' or a positive whole number.

DESCRIPTION

PeaKDEck is a utility written in Perl, mainly intended for use in the identification of peaks in mapped DNaseI-seq data. It also includes a set of utilities for processing and manipulation of this data from the mapping stage forwards. It works on data in sam format.

PeaKDEck selects a threshold read density for peak calling by constructing a probability distribution of background read density scores using kernel density estimation. It selects a threshold by selecting a read density that is 'significantly' outside this background probability distribution. All measurements of read density are corrected for local background variation in signal intensity.

FAQs

What are the system requirements?

The command line and GUI PeaKDEck applications have been tested on OSX (Mountain Lion), Ubuntu 12.04 LTS, Windows XP and Windows 7. The system requirements are largely dependent on the size of data files being used. We recommend at least 4GB memory for basic use with small data files. For the numerical sorting of sam files, ~(file size * 2.5) free memory is required for efficient sorting. For the command line applications, we recommend Perl v5.12 or later. On Windows, PeaKDEck was tested with Strawberry Perl.

How do I install PeaKDEck?

PeaKDEck GUI: On Windows, PeaKDEck should run without the need to install Perl, or other additional software. On Linux and OSX platforms, the X Window System (X11 or Quartz) must be installed.

PeaKDEck command line: to use the PeaKDEck command line application, Perl must be installed on your computer. We recommend Perl v5.12 or later. On Linux and OSX platforms, no other software is required to run the command line application. On the Windows platform, a pseudorandom number generating module (Math::Random::MT) is required, and is available through CPAN for Strawberry Perl users, and PPM (Math-Random-MT) for ActiveState users.

Which short read file formats does PeaKDEck work with?

At present, PeaKDEck only works with files in the SAM format (see samtools.sourceforge.net/SAMv1.pdf for details).

Which application should PeaKDEck be opened with?

On the OSX platform, after launching the PeaKDEck GUI application (having installed XQuartz), you may be prompted to choose an application with which to open PeaKDEck. PeaKDEck should be opened with the Terminal application (located at /Applications/Utilities/Terminal.app).

Why is the GUI freezing?

As yet, the PeaKDEck GUI is not a multi-threaded program. As such during data processing, the GUI may appear frozen or unresponsive. Particularly on the Windows platform. For now, this is expected behaviour. The GUI will refresh when new status updates are available, and will return to full responsiveness when data processing has finished.

EXAMPLES

see SYNOPSIS above.

CAVEATS

AUTHOR

Michael McCarthy
michaeltmccarthy@gmail.com

ACKNOWLEDGEMENTS