

HPeak: an HMM-based ChIP-Seq Analysis Program

Steve Qin and Jianjun Yu

05/26/08

Introduction

This program is for the purpose of defining genome-wide ChIP-enriched peaks using short sequence reads generated by ChIP-Seq assay. The underlying algorithm is based on a two-state HMM.

Set up

tar xvf creates a directory called HPeak-1.0, which contains all perl scripts and C++ source code plus two subdirectory, /data/ and /example. The first contains all working information files. The second contains sample data. One can either include the /HPeak-1.0/ path to the appropriate configuration file of your operating system, or add path in each command. E.g.,
`perl ~/program/HPeak/scripts/HPeak.pl.`

To get DNA sequence data using the `-seq` option, one needs to download the human genome sequence files from either UCSC genome site (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/chromFa.zip>) or HPeak site, see Download page). The sequence files should be extracted to data/chromFa/ folder.

To get detail genomic annotation information using `-ann` option, one needs to download additional information files from HPeak site (see Download page). The package includes one refgene file (refFlat.out) and a set of phastCons score files, one per chromosome. One should move the refFlat.out to script/data/ folder and the phastCons score files to script/data/phastCons/ folder. In order to quickly lookup the conservative scores for individual base positions, conservation scores for alignments of 16 vertebrate genomes with human are downloaded from UCSC genome site (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons17way/>) and converted to ASCII files.

In the examples shown in this document, we assume that the path has already been correctly set up so it is omitted from the commands.

For your convenience, we included pre-compiled executables for the C++ programs needed by HPeak. If for some reasons you need to recompile it. Use:

```
g++ -o chiphmm chiphmm.cpp
g++ -o hmminus chiphmmminus.cpp
```

Syntax:

```
perl HPeak.pl < -format FORMAT -t TFILE -n NAME > [Options]
```

Options:

-h	Show this help message
-format	Format of tag file, "ELAND", "BED" or "Custom". REQUIRED When choosing "Custom", the following information needs to be specified in the order of: 1. The column to store genome file where match was found 2. The column to store base position of match 3. The column to store direction of match The column numbers start at 1 and separate with comma. Example: the ELAND format can be presented as -format custom[7,8,9]
-t TFILE	Treatment file name. REQUIRED
-n, -name	Experiment name used to generate output file. REQUIRED
-c CFILE	Control file name.
-fmin	Minimal DNA fragment size. DEFAULT: 100
-fmax	Maximal DNA fragment size. DEFAULT: 300
-w, -window	Window size (bp). DEFAULT: 25
-s, -sig	P-value threshold for peak detection. DEFAULT: 1e-3
-wig	Whether to generate WIG file for UCSC genome browser
-seq	Whether to extract peak sequences
-ann	Whether to extract nearest gene information for peaks

Example:

```
perl HPeak.pl -format ELAND -t example.txt -n output -fmin 100 -fmax 300 -w 25 -s 1e-3 -wig -seq -ann
```

Parameters:

-f

input file format:

HPeak currently allows three types of input formats: ELAND, BED or Custom. ELAND is for files of s_N_eland_result.txt format produced by Illumina Genome Analyzer Pipeline software suite (see the figure below for illustration).

>PATHBIO-SOLEXA_20F0CAAXX:1:39:231:694	TTTTTTGAGGAATATATGTATATATNTGTTGGGT	U0	1	1	0	hs_ref_chr5.fa	68396743	F	DD
>PATHBIO-SOLEXA_20F0CAAXX:1:39:856:492	TCTCCCAATTGACCTTTGGGATATGNGNATAAAATT	U0	1	0	0	hs_ref_chr7.fa	128899421	F	DD
>PATHBIO-SOLEXA_20F0CAAXX:1:39:869:502	TGTTCTTGCCTATGTGTCAAAGTTTANANAATAGTA	U0	1	1	1	hs_ref_chr4.fa	84399937	F	DD
>PATHBIO-SOLEXA_20F0CAAXX:1:39:833:492	TAACACCACATTTATAAAGAAAATGNATAGGAGGT	U0	1	0	1	hs_ref_chr12.fa	98914367	F	DD
>PATHBIO-SOLEXA_20F0CAAXX:1:39:949:397	TCTACTCTATTCTGTATTTAATTGNTNTGTTGAGT	U1	0	1	0	hs_ref_chr12.fa	118576351	R	DD 2T
>PATHBIO-SOLEXA_20F0CAAXX:1:39:860:394	TATGACAGTGACAGTGACGTGTGATNANGGCCTTTT	U1	0	1	1	hs_ref_chr11.fa	104497572	R	DD 2T
>PATHBIO-SOLEXA_20F0CAAXX:1:39:932:440	GTGAGAAAAGCACATGTGGATTAAANANAATGTTT	U0	1	0	1	hs_ref_chr7.fa	77857072	F	DD
>PATHBIO-SOLEXA_20F0CAAXX:1:39:862:457	GGTGGGGGGAGGAGGGGGATATACCNANCATTAGGT	U0	1	0	0	hs_ref_chr2.fa	166775065	F	DD
>PATHBIO-SOLEXA_20F0CAAXX:1:39:848:496	TATTTTGGGCTTACATTCCCTGCATNTNAGTTCATG	U0	1	0	0	hs_ref_chr12.fa	127130339	R	DD
>PATHBIO-SOLEXA_20F0CAAXX:1:39:782:258	GAATTGACCCATAGTCATTGCAGAANGNAACAGCTG	U0	1	0	1	hs_ref_chr14.fa	101330447	F	DD

Columns 7-9 contain mapping information we used:

7. Genome file in which match was found.
8. Position of match (bases in file are numbered starting at 1).
9. Direction of match (F=forward strand, R=reverse).

BED format is used in UCSC genome browser. In addition to the three required columns of chromosome, start, end positions, HPeak requires a fourth column which contains the strand information (+: positive strand, -: negative strand). **They have to occupy the first four columns and need to be in the right order.** The sequences /rows do not need to be sorted. More

columns are allowed, but will be ignored by HPeak. Detailed information about the BED format can be found at <http://genome.ucsc.edu/goldenPath/help/customTrack.html#BED>. An example of BED format input file is shown below.

chr1	51131	51155	+
chr1	52543	52567	-
chr1	60254	60278	-
chr1	77358	77382	-
chr1	78157	78181	+

HPeak allows great flexibility in terms of the input format. Most non-standard formats can be used with the “CUSTOM” format option. The square bracket right after custom contains the column numbers in which chromosome, start and strand information are stored. For example, ELAND format is equivalent to “CUSTOM[7,8,9]”. Note that no space is allowed. It is required that the string “chr” and “.fa” (part of genome sequence files in which a match is achieved for the read) be present in the chromosome column. Other rows will be ignored. The strand can be denoted as either “F/R” or “+/-”.

HPeak support the new Illumina Genome Analyzer Pipeline 0.3.0. In this version, the final eland result for each lane is summarized in files called s_N_export.txt. One can use “CUSTOM[9,10,11]” to read this type of files.

-t

-c

input filenames

In addition to data files, one needs “input files” to run HPeak. These files contain location and names of sequencing data files. Each line in these files represents one data file. If there are multiple data files specified in the input file, they will be merged so sequences from multiple sources can be combined. A sample file is shown below.

<pre>/data/GERALD/s_1_eland_result.txt /data/GERALD/s_2_eland_result.txt /data/GERALD/s_3_eland_result.txt</pre>
--

-fmin

Minimum fragment width

The lower bound of the length of size-selected DNA fragments. The default value is 100 bp. This number has to be a multiple of the window size described below.

-fmax

Maximum fragment width

The upper bound of the length of size-selected DNA fragments. The default value is 300 bp. This number has to be a multiple of the window size described below.

-w, -window

window size

This program partitions the genome into small segment so number of read coverage is counted in each segment. This strategy allows comparison across samples. Larger window size reduces

computation time and file size but lowers resolution in defining enriched regions. The default value is 25bp.

–s. –sig

significance level

This is the p-value threshold to determine whether a peak is significantly ChIP-enriched. Multiple comparisons are adjusted using the Bonferroni method, i.e., p/N , N is total number of regions. The default significance level is 0.001 (same as in Robertson et al. Nature Methods 2007).

–wig

whether to generate WIG format coverage profile file for the enriched regions.

–seq

whether to generate FASTA format sequence files for the enriched regions.

–ann

whether to generate detailed annotation information for the enriched regions.

Detailed annotation including GC%, conservation rate, genomic features (exon, intron, intergenic,...) and up- and down- stream gene name. Details about this file can be found in the Optional output files section.

Output files:

.allregions.txt:

This is the main output file. It is in BED format indicating chromosome, start and end location, and the length (in bp) of all enriched regions. The last column contains the maximum coverage among all bins in this region. Note that the first column only contains the chromosome number, and X and Y are replaced by 23 and 24 for easier numerical manipulations.

.sum

This file contains three main parts. The first part lists all the parameter values entered. The second part contains total number of reads and uniquely mapped reads from treated and control samples. The third part summarizes the number of enriched regions, total length of DNA covered by these regions, and range of read coverage in these regions.

.log

This file contains all the raw output information. It is for debugging purpose.

Optional output files:

.seq

This is a FASTA format file containing the sequence of all enriched regions. Such a file is useful for motif scan which is often a follow up study. Note that we used the unmasked Build 36.1 finished human genome assembly (hg18, Mar. 2006).

.annotation.txt

This file contains detailed annotation of the enriched regions. A typical file is shown below:

Chromosome	Coverage	Height	GC%	Masked bases%	Conservation score	Location	GeneName	GB Acc	Strand	Distance	GeneName	GB Acc	Strand	Distance
chr1:1314451-1314700	250	8	GC%: 52.8	Masked%: 68.8	0+/-0	Intron	CCNL2	NM_030937	-					
chr1:2328626-2329000	375	8	GC%: 60.5	Masked%: 26.7	0+/-0	Intron	PEX10	NM_153818	-					
chr1:8847376-8847575	200	8	GC%: 46.5	Masked%: 72.0	0+/-0	Intron	ENO1	NM_001428	-					
chr1:8855476-8855800	325	8	GC%: 44.0	Masked%: 0.0	0+/-0	Intron	ENO1	NM_001428	-					
chr1:10998826-10999050	225	9	GC%: 48.4	Masked%: 100.0	0+/-0	Intron	TARDBP	NM_007375	+					
chr1:11003926-11004200	275	10	GC%: 53.1	Masked%: 98.9	0+/-0	Intron	TARDBP	NM_007375	+					
chr1:11051076-11051350	275	8	GC%: 47.6	Masked%: 0.0	0.2+/-0.38	Exon	EXOSC10	NM_002685	-					
chr1:11051901-11052100	200	8	GC%: 55.5	Masked%: 0.0	0+/-0.01	Intron	EXOSC10	NM_002685	-					

The columns are: peak genomic position, peak length, peak max height, GC content, repeated sequence percentage, mean and standard deviation of conservative scores for the enriched region, relationship with nearest genes including whether the peak is located within the gene or between genes, gene name, GB accession number, strand, distance to gene transcription start site. If the enriched region is located between genes, nearest gene on each end will be provided.

.wig

This file contains WIG format coverage profile file for the enriched regions. One can directly upload the generated file onto UCSC genome browser for visualization.

Sample data

In the /example/ subdirectory, there are two sample files, sample.chip.txt and sample.mock.txt. these two files are subsets of the corresponding datasets from the Johnson et al. NRSF ChIP-Seq study (Johnson et al. 2007). The data are downloaded from Illumina website: http://www.illumina.com/downloads/Illumina_ChIP-Seq_Demo_Data_Johnson_Science_2007.zip

The format of these data is slightly different from the standard ELAND format. So one needs to use the "CUSTOM" format option. Two input files: chip.inp and mock.inp are provided. A sample command can be found in the file readme.

Acknowledgement

We thank Dr. Chris Maher, Shanker Kalyana-Sundaram, Terrence Barrette and members of the Chinaniyan Lab for valuable suggestions and comments on earlier versions of this program.

Reference

Qin ZS, Yu J, Maher CA, Kalyana-Sundaram, S, Yu J, Chinnaiyan, AM (2008) HPeak: An HMM-based algorithm for defining read-enriched peaks from massively parallel sequencing data. <http://www.sph.umich.edu/csg/qin/HPeak>

Contact

Comments, suggestions, questions are welcomed, and should be directed to Steve Qin.
Email: qin@umich.edu. Phone: 734-763-5965.