* **PACKAGE CONTENTS**

pscan\_chip.cpp *PscanChIP source file*

fasta\_to\_raw.cpp *Fasta\_to\_raw script source file*

jaspar\_2014.wil *JASPAR 2014 matrices*

*jaspar\_2016.wil JASPAR 2016 matrices*

transfac.wil *Free TRANSFAC matrices*

BG *This folder contains pre-computed background values for Jaspar and Transfac matrices, used to assess motif enrichment on different cell lines (see main paper on how background values are computed). The available backgrounds are for the same cell lines included on the PscanChIP web site (the list is at the end of this document)*

* **INSTALL**

To compile PscanChIP you need the source file contained in the pscan\_chip.tar.gz archive, a C++ compiler (like GNU g++) and the GNU Scientific Library (gsl - <http://www.gnu.org/software/gsl/>) installed in your system.

1 - Extract the archive pscan\_chip.tar.gz with the command

> tar -xvf pscan\_chip.tar.gz

and access the pscan\_chip folder with the command

> cd pscan\_chip

2 - Compile the pscan\_chip.cpp source file with

> g++ pscan\_chip.cpp -o pscan\_chip -O3 -lgsl -lgslcblas

If all is fine the compiler shall issue no error messages and you should find a pscan\_chip executable file in the folder you are in.

3 - If you want to work with genomes or genome assemblies different from hg19, hg18, mm10 and mm9 you need to compile the fasta\_to\_raw.cpp script as well, in order to convert the genome FASTA file(s) to the “raw” format used by Pscan\_ChIP. To compile the script just type:

> g++ fasta\_to\_raw.cpp -o fasta\_to\_raw -O3

Again, you should find the new executable file fasta\_to\_raw in your folder

* **GENOME FILES**

Files needed to work on the most recent human and mouse genome assemblies (hg19, hg18, mm10 and mm9) can be downloaded from http://www.beaconlab.it/pscan\_chip\_dev/download The respective archive(s) (.tar.gz) have to be expanded in the folder containing the PscanChIP executable file.

To add more genomes or assemblies you need to prepare a folder containing the genome in the particular “raw” format used by PscanChIP. First of all create a new folder, within the PscanChIP main one, and call it in accordance to the genome release, e.g. dm3. Put the genome FASTA file(s) in this new folder. Run the “fasta\_to\_raw” script on the fasta file(s) using the following syntax:

> ../fasta\_to\_raw file1 [file2] ... [fileN]

where file1, file2 … fileN are the FASTA files containing the sequences from your genome of interest.

After fasta\_to\_raw completes, a number of files named XXX.raw should appear in the folder, where XXX represents the name of each chomosome as defined in the headers of the FASTA file(s). When the “.raw” files have been produced, the original FASTA files can be removed from the folder.

ATTENTION: PscanChIP accepts as input a list of genomic coordinates in bed format, i.e. chromosome/start/end. The name of chromosomes in the .raw files have to match those that will be used to define coordinates in the input files. For example, chromosome X can be referred to as “chrx” or “chrX”. One possible solution is to edit the input files to match the chromosome annotation used. Or, when dealing e.g. to “chrx” while the corresponding sequence file name is “chrX.raw”, it may be useful to create symbolic links with alternative names pointing to the correct file.

In our example, with a “chrX.raw” file and a possible “chrx” nomenclature ambiguity, one could do something like

> cp -s chrX.raw chrx.raw

to create a chrx.raw symbolic link from “chrx.raw” to the “chrX.raw” file.

* **USING PSCAN-CHIP**

Mandatory options:

*-r [regionfile]*

the BED file with the regions to be analyzed TFs enrichment (e.g. peaks from a ChIP-Seq experiment). PscanChIP will compute the central position of each region and consider the genomic region surrounding it in its computations. The default length of the region surrounding the center is 150bp but it can be modified using the *-s* option. For optimal results we suggest to use summits’ coordinates when available instead of peak coordinates.

*-g [folder]*

the genome folder to which the BED file refers to. The directory must contain the genome files in RAW format (one file per chromosome).

*-M [matrixfile]*

the file containing the motif matrices to be used by Pscan\_ChIP. See the Matrix File section for further info.

Other options:

*-s [size]*

the genomic regions size, default is 150bp. Leaving the default value assures optimal results in most cases. Beware that changing the region length makes the available background file(s) inconsistent, since they were computed for regions of 150 bp. Thus, to change the genomic region size to be analyzed you will also need to produce new background file(s) for the new region size. All in all, for a normal ChIP-Seq experiment it’s better to leave this parameter untouched.

*-m [matrixname]*

use this option to select a matrix from *matrixfile* and make Pscan\_ChIP run in *single matrix* mode.

*-bg [bgfile]*

Background file, needed to compute global *pvalues*.

*-ss*

Single strand mode.

Output files will be written in a *regionfile*.res file, with *regionfile* being the name of the BED file passed with the *-r* parameter. When running in *single matrix mode* the output file will have a “.ris” extension instead.

- **MATRIX FILE**

The file containing the motif profiles to be used by PscanChIP must be formatted as in the example:

>ID1 NAME1

A\_1 A\_2 ..... A\_n

C\_1 C\_2 ..... C\_n

G\_1 G\_2 ..... G\_n

T\_1 T\_2 ..... T\_n

>ID2 NAME2

A\_1 A\_2 ..... A\_n

C\_1 C\_2 ..... C\_n

G\_1 G\_2 ..... G\_n

T\_1 T\_2 ..... T\_n

..and so on, where A\_i, etc. are the frequencies of the four nucleotides in the columns of the matrix. These values can be either integers or floating point values, they will be automatically rescaled to frequencies summing to one in each column.

The NAME field may be omitted. You can refer to the files \*.wil and transfac.wil in the PscanChIP folder as examples.

- **EXAMPLES**

1) Running PscanChIP with a precomputed background file using Jaspar matrices and human genome (hg19):

> pscan\_chip -r input.bed -g hg19 -M jaspar\_2014.wil -bg BG/K562.jaspar.bg

2) Running PscanChIP in *single matrix mode* to obtain the position of the best matches for a given matrix within the input regions (one match per region).

> pscan\_chip -r input.bed -g hg19 -M jaspar\_2014.wil -m MA0493.1

3) Preparing a new background file for a custom set of matrices or for a new set of accessible genomic locations.

> pscan\_chip -r background.bed -g hg19 -M mymatrices.wil

The *background.bed.res* file obtained can be used as a background file for successive PscanChIP runs using the same *mymatrices.wil* matrices file.

- **BACKGROUNDS**

This is the list of pre-computed backgrounds for Jaspar and Transfac binding profiles collections found in the BG folder of PscanChIP. If the cells/tissue on which your ChIP-Seq experiment was performed is not on the list, you can either choose what seems the closest one (e.g. HepG2 for liver cells), or select the "mixed" background, built using a random selection of regions from different cells or, if your ChIP-Seq regions are restricted to or mostly come from gene promoters, you can select "Promoters" as a background. Alternatively you can compute new backgrounds following the instructions at point 3 of the EXAMPLES section. A summary description of cell/tissue types is available [here](http://genome.ucsc.edu/ENCODE/cellTypes.html).

|  |  |  |
| --- | --- | --- |
| Cell Line | Jaspar BG | Transfac BG |
| \*MIXED\* | mixed.jaspar.bg | mixed.transfac.bg |
| \*PROMOTERS\* | promoters.jaspar.bg | promoters.transfac.bg |
| AG10803 | Ag10803.jaspar.bg | Ag10803.transfac.bg |
| AoAF | Aoaf.jaspar.bg | Aoaf.transfac.bg |
| CD20 | Cd20.jaspar.bg | Cd20.transfac.bg |
| GM06990 | Gm06990.jaspar.bg | Gm06990.transfac.bg |
| GM12865 | Gm12865.jaspar.bg | Gm12865.transfac.bg |
| H7-hESC | H7es.jaspar.bg | H7es.transfac.bg |
| HAEpiC | Hae.jaspar.bg | Hae.transfac.bg |
| HA-h | Hah.jaspar.bg | Hah.transfac.bg |
| HA-sp | Hasp.jaspar.bg | Hasp.transfac.bg |
| HCF | Hcf.jaspar.bg | Hcf.transfac.bg |
| HCM | Hcm.jaspar.bg | Hcm.transfac.bg |
| HCPEpiC | Hcpe.jaspar.bg | Hcpe.transfac.bg |
| HEEpiC | Hee.jaspar.bg | Hee.transfac.bg |
| HepG2 | Hepg2.jaspar.bg | Hepg2.transfac.bg |
| HFF | Hff.jaspar.bg | Hff.transfac.bg |
| HIPEpiC | Hipe.jaspar.bg | Hipe.transfac.bg |
| HMF | Hmf.jaspar.bg | Hmf.transfac.bg |
| HMVEC-LLy | Hmvecb.jaspar.bg | Hmvecb.transfac.bg |
| HMVEC-dBl-Ad | Hmvecdblad.jaspar.bg | Hmvecdblad.transfac.bg |
| HMVEC-dBl-Neo | Hmvecd.jaspar.bg | Hmvecd.transfac.bg |
| HMVEC-dLy-Neo | Hmvecf.jaspar.bg | Hmvecf.transfac.bg |
| HPAF | Hpaf.jaspar.bg | Hpaf.transfac.bg |
| HPdLF | Hpdlf.jaspar.bg | Hpdlf.transfac.bg |
| HPF | Hpf.jaspar.bg | Hpf.transfac.bg |
| HRCEpiC | Hrce.jaspar.bg | Hrce.transfac.bg |
| HSMM | Hsmm.jaspar.bg | Hsmm.transfac.bg |
| HUVEC | Huvec.jaspar.bg | Huvec.transfac.bg |
| HVMF | Hvmf.jaspar.bg | Hvmf.transfac.bg |
| K562 | K562.jaspar.bg | K562.transfac.bg |
| NB4 | Nb4.jaspar.bg | Nb4.transfac.bg |
| NH-A | Nha.jaspar.bg | Nha.transfac.bg |
| NHDF-Ad | Nhdfad.jaspar.bg | Nhdfad.transfac.bg |
| NHDF-neo | Nhdfneo.jaspar.bg | Nhdfneo.transfac.bg |
| NHLF | Nhlf.jaspar.bg | Nhlf.transfac.bg |
| SAEC | Saec.jaspar.bg | Saec.transfac.bg |
| SKMC | Skmc.jaspar.bg | Skmc.transfac.bg |
| SK-N-SH\_RA | Sknshra.jaspar.bg | Sknshra.transfac.bg |
| Th1 | Th1.jaspar.bg | Th1.transfac.bg |

- **REFERENCE**

If you find PscanChIP useful for your research please cite us:

Zambelli F, Pesole G, Pavesi G.

PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments.

Nucleic Acids Res. 2013 Jul;41(Web Server issue):W535-43. doi: 10.1093/nar/gkt448.

Epub 2013 Jun 7. PubMed PMID: 23748563; PubMed Central PMCID: PMC3692095.