ICeCAP

ICeCAP is written in ANSI-C and provides flexible, multi-scale analysis of *BOTH* Hi-C and Capture-Hi-C data. Dynamical, pointer based memory allocation and mapping of sub-genomic regions allow interaction maps to be resolved computationally at single restriction fragment, "meta-fragment" or uniform resolution. The output files can be converted to *hic* format with the $Juicer_tools$ (not included). The significant interactions, generated by a local background Poisson model along the lines of the HiC pipeline HiCCUPS, do not require GPU acceleration but progress in this respect is ongoing on a newer version of ICeCAP. ICeCAP has been developed at the Institute of Cancer Research and the Wellcome Centre for Human Genetics by Gabriele Migliorini.

Requirements:

****

ICeCap runs in a Linux environment running bash shell with GNU environment default functions (gawk, sort, sed, etc...) available at run time. Compilation is done using the GNU Compiler (GCC). This may be changed by editing the make file (see below). Read alignment is carried out using bowtie2, version 2.2.X. This should be available in the PATH environtment variable. To check the available version of bowtie type "bowtie2 –version" at the command line. Bowtie reference files should be generated using the same reference genome as the one used during alignment (see the bowtie manual for details). Bed formatted mapability files for the species under study should be placed in the subfolder "reference_map". See the README file in that folder. The mapability files should be generated by the same version of the aligner below. A script to help building mapability files is included. Perl version v5.10.1 or later is expected, earlier versions may also work. R version 3.X.X or later is expected. Free disk space that is at least double the size of the input fastq files is recommended, more if intermediate files are kept (see options). A considerable amount of RAM may also be needed depending on the number of fragments or bins being analysed. Analysis of data from a typical experiment involving the human genome digested with HindIII may require around 40 gigabytes of RAM. A minimum 1kb resolution is the default setting, though higher resolution is in principle possible, when considering .e.g. four base cutter enzyme or other digestion protocols.

1 Compilation:

****

Unzip/untar the file ICeCap.tar.gz:

> tar -zxvf ICeCap.tar.gz

Run the make file found in the ICeCap directory to compile the ICeCap binary file. The ICeCap directory can either be in your home folder or can be installed with administration privileges in the root directory. Make sure you add the ICeCap folder to your PATH environment in this case. In the ICeCap folder, at the command line prompt, type:

> make

That is all. Test shell scripts for submitting ICeCAP on a suitable queuing environment are provided.

2 Running ICeCap:

****

ICeCap takes unaligned sequencing data in the FASTQ format. The data is processed in two stages:

*Firstly, sequencing reads are processed to take into account ligation junctions, aligned using bowtie2 and paired up after alignment. Valid ditags, along the lines of other NGS pipelines, e.g. HICCUP, are identified and stored in SAM-like format files. Allele specific calibration of reads is available in the option list (see below).

*Secondly, a matrix of interacting intervals is populated. The resolution can be choosen to be fragment based or uniform-bin based. Iterative correction and interaction distance normalisation is carried out along the lines of the Sinkhorn-Knopp algorithm. A list of significant interactions based on the HICCUPs algorithm, namely a local Poisson backgroup approach, is produced. In the output/stat folders, a file for every captured fragment appears, listing the genome-wide location of the di-tag ends, the score, defined as -log(Pval) and the FDR value. A combined list of significant interactions is also generated.

2.1 **STAGE 1: read alignment, pairing, deduplication, and ditag validation.

To process sequencing data from, e.g., a non-capture Hi-C experiment, using the HindIII restriction enzyme, run in your system, as a *single* line command, e.g.:

- /path/to/binary/ICeCap -E AAGCTT -C 1
-L /path/to/your/fastqfiles one strand/
-R /path/to/your/fastqfiles one strand/
-P /path/to/your/bowtie/reference/directory/
-l sampleid

-T 100
-f /path/to/your/ICeCap/directory/
-D /path/to/your/outputsam/folder/

To process capture Hi-C data a BED formated file listing the captured intervals needs to be provided using the -B option.

• ICeCap -B path/to/baits.bed -E AAGCTT -C 1
-L /path/to/your/inputfastq/file(s)
-R /path/to/your/inputfastq/file(s)
-P /path/to/your/bowtie/reference/directory/
-l sampleid
-T 100
-f /path/to/your/ICeCap/directory/
-D /path/to/your/outputsam/folder/

ICeCap will carry out a virtual digestion the reference genome generating a list of fragments. Multiple files can be concatenated passing multiple fastq files on both the left and right strand flag, e.g. $-Lfile1\_1.fastq, file2\_1.fastq -Rfile1\_2.fastq, file2\_2.fastq$. Make sure you submit the command above with a suitable queuing environment, e.g. MOAB, LSF etc.

STAGE 1 will produced SAM files (without headers) listing valid ditags to be used for subsequent analysis. A header can be found in the "data" directory. The following output files will be produced in the output directory, e.g. /sampleid/data (indicated with the -D option).

File name — File contents

---

sampleid/data/SIZES — Ditag sizes.
sampleid/data/GSIZE — Valid ditag sizes.
sampleid/data/BSIZE — Invalid ditag sizes.
sampleid/data/sampleid.png — Plot of bona-fide and filtered di-tags size distributions
sampleid/data/Invalid_ditag_chart.pdf — Pie chart with the relative ratio of invalid di-tags filtered out. sampleid/data/sampleid.COUNTS.txt — Statistics about the numbers of aligned reads and valid ditags.
sampleid/data/sampleid.pairs.chr*.sam.bfide — SAM-like file listing valid ditags.
sampleid/data/sampleid.pairs.chr*.sam.bfide.ontarget — SAM-like file listing valid ditags corresponding to enriched ditags in
the case of capture Hi-C or all ditags in the case of non-capture Hi-C.

2.2 **STAGE 2: iterative correction, distance bias correction, interaction calling.

To run this stage, use the -N option to allocate di-tags to memory and post-processing analysis. The -S option will run statistical analysis on the two pools of data obtained: enriched to non-enriched (E-N) as well as the enriched to enriched (E-E) interactions. The argument to the -S option indicats the type of distribution that will be used as a local background model in the three-filtered, HiCCUPs like, method.

Example of interaction analysis in capture Hi-C data carried out on chromosomes 1, 2 and 3.
ICeCap -N -S poi -l sampleid
-s 1 -e 3 -G 0 -Z 1 -Q path/to/mappability/files
-B path/to/baits.bed -P path/to/bowtiereference/files

Mandatory fields are the -N flag, the reference fasta file directory (-P), the sample id (-l), that should correspond to those used in the STAGE 1 above, and the (-f) path to the ICeCap folder. If no baits BED file is specified (with the -B option), the pipeline will run in non-capture mode and allocate all fragment pairs to memory. Note that the mappability files should be in the ICeCap folder, in a folder called $reference\_map$ or should be specified by thwe $-Q$ flag. Setting -G 0 indicates that a restriction based (non-uniform) binning will be used. Alternatively, if the -G option is set to 1, a uniform binning is chosen, with a size of r*Z bp. (see the -r and -Z flags below). The -Z parameter will corse grain the analysis. So, for example, -G 0 -Z 1 corresponds to single fragment binning. Similarly, -G 0 -Z 4 will correspond to a "meta-fragment" based binning, where four consecutive fragments are combined. The -Z option can also be used with uniform bins, e.g. -G 1 -Z 5 will correspond to a uniform binning grid of mesh size 5*r=5kb. (default value of the -r option is 1000 bp). Interactions passing the FDR threshold, together with the associated q-values can be found in the "sampleid/stats/fdr" directory for further analysis and for plotting purposes. A Wash-U formatted file with all significant interactions will also be generated in the output folder, as specified by the -O option. Per fragment iWash files are grouped in chromosome start/end subfolder(s) for individual locus analysis.

3 ICeCap Options:

****
The following options can be passed to the ICeCap program.
****
*GENERAL OPTIONAL ARGUMENTS**
-h, –help Print this manual and exit.
-V, –version Print the version number and exit.
-B, –baits path to file path to a BED file [without header lines] listing fragments that were enriched in the capture Hi-C experimental protocol. If this option is not supplied, analysis is carried out on all fragments as in non-capture Hi-C.

****
*MANDATORY ARGUMENTS FOR STAGE 1**
-E, –site-sequence nucleotide sequence string Specifies the enzyme recognition site sequence. Currently, only enzymes that cut at palindromic sequences are supported. The default value is AAGCTT, which corresponds to the HindIII recognition seuqence.
-C, –cut-position integer Number of bases from the start of the recognition site sequence to the cut point.
(e.g for HindIII use -C 1)

-AAGCTT- -A AGCTT- -TTGCAA- -TTGCA A-

(e.g. for MboI use -C 0)
-GATC- - GATC- -CTAG- -CTAG -

-L, –input-fastq prefix string Path to the first set of fastq files . Sequencing read data contained in these files will be aligned. Alignment will be carried out on upaired reads. Reads will subsequently be paired according to matching read names.

-R, –input-fastq prefix string Path to the first set of fastq files . Sequencing read data contained in these files will be aligned. Alignment will be carried out on upaired reads. Reads will subsequently be paired according to matching read names.

-P, –ref-bowtie prefix string Path to the directory containing the Bowtie indices and reference fasta files. See the botwie2 manual for details on how to create these.

-H –prefixtobowtie Default is Human reference Genome,hg19. Change to other reference build and/or species. Max chromosomes 25. Please edit the main.c otherwise.

-T, –read-length integer Length of the reads (in bases) found in the input fastq files (please provide this number accurately).

-Q, –refmapability prefix string Path to the directory containing the files with mappability values [BED format]. The standard prefix is the one provided in the sample mappability folder provided with ICeCap.

-f, –dir-icecap path to directory Path to the ICeCap (tools) directory containing the "referenceC" sub-directory. The "referenceC" subdirectory contains scripts and templates used by ICeCap at various stages. The default is the untar ICeCap directory.

-l Sample Name/ID of the Biological replicate.

****

*OPTIONAL ARGUMENTS FOR STAGE 1**

-D, –output-sam folder: The output directory where the files listing ditags in SAM format, as well as intermediate processing files (in the subdirectory "BOWTIE"), will be written. This may include a prefix corresponding to the beginning of the file names. The default is the current working directory.

-b, –keep-intermediate Keep intermediate processing files Note, these may take up around ten times the amount of disk space of the input fastq files. Default is yes.

-A –allele-specific-HiC. If yes, the CIGAR string is read and used to calibrate positions of the reverse reads before the matefixing step. This flag is experimental. Please use with care. Only Insertions/Deletions are considered for recalibration in the present version.

-J Number of cores used by Bowtie2.

****

*MANDATORY ARGUMENTS FOR STAGE 2**

-N, this flag will point to STAGE 2. It does not require an argument. -P, –ref-bowtie prefix string Path to the directory containing the Bowtie indices and reference fasta files. See the botwie2 manual for details on how to create these.

-l Sample Name/ID, if specified in STAGE 1.

-f, –dir-icecap path to directory Path to the ICeCap (tools) directory containing the "reference_C" sub-directory. The "reference_C" subdirectory contains scripts and templates used by ICeCap at various stages. The default is the untar ICeCap directory.

-I Please provide a directory path here. IT should match the -D path, as given on stage 1, and corresponds to the folder where the Sample Name/ID folder and data subfolder containing the bona fide sam files generated in stage 1 are located.

****

*OPTIONAL ARGUMENTS FOR STAGE 2**

-o Output prefix of choice to mark a specific instance of ICeCAP. -G, –uniform-bin-size integer If this option is set, the interaction matrix will be constructed using bins of a uniform size in r bases (bp) as specified by the integer passed with the -r option. Note that the RAM usage increases as a square of the number of bins so setting a low value here may consume a lot of RAM. If this option is not set, the interaction matrix will correspond to fragments produced by the digestion of the genome with the chosen enzyme.

-Z, –combine-intervals integer If this option is set, fragments or bins will be combined into larger intervals for the purpose of calculating interactions. The number of fragments or bins to be combined is indicated by the argument passed to this option.

-S, –distribution string Carry out statistical analysis on read count matrices. This is mutually exclusive with the –alignment-script option. The argument passed to this option specifies the distribution that will be used to calculate interaction p-values. The following distribution can be specified: "poiss"

-I, –input-sam folder: You can provide here the folder where you have generated the output files of STAGE 1. This flag, when used, should match the path specified in STAGE 1 by the -D flag.

-O, –output-bed prefix string The output directory where the IBED file listing statistically significant interactions. The default is the ICeCap/SampleId directory.

-v, –trans-interactions Include trans-chromosomal interactions in the interaction matrix.

-s, –lowest-chromosome-number integer Analysis will be implemented for chromsomes numbering between the values set by –lowest-chromosome-number and –highest-chromosome-number, inclusively. The default value is 1.

-e, –highest-chromosome-number integer Analysis will be implemented for chromsomes numbering between the values set by –lowest-chromosome-number and –highest-chromosome-number, inclusively. The default value is 24

-r –minimum resolution integer. Number of base pairs. Default is 1000bp=1kb.

-w, –weight-threshold float Interactions for which the iteratively corrected weight falls below this threshold will be discareded to facilitate convergence of the iterative correction procedure. The default value is 0.1.

-m, –map-threshold float Fragments/bins with an average mappability value below this threshold will be discarded from further analysis. The default value is 0.20.

-t, –tolerance float Tolerance used to determine Iterative Correction convergence. The default value is 10-5.

-M, –max-iterations integer Maximum number of iterations allowed for the Iterative Correction. Default value is 100.

-W, –max-ditag-length integer Maximum valid ditag length in bases (bp). Ditags longer than this will be discarded. Default value is 800 bp.

-U, –q-threshold The minium q-value threshold used to filter the final list of interactions. The default value is 0.05.

-p, –p-threshold The minium p-value threshold used to filter the final list of interactions. At present no p-value threshold is implemented.

-c, –trimming Not all fragments enter the statistical analysis. A threshold on coverage is done and all fragments with, e.g. -c 0.05, more than 5 percent of populated cis other end fragments will be considered for statistical analysis.Coverage is measured as a ratio of reads in populated vs total number of cells per chromosome.

****

4 Graphical Interface

Please make sure to have the X11 libraries running on your desktop environment. At compile time, an executable named 'ICeCap_plot' is generated. Please use the command line help option.

Updates

Version 1.4 is the simplest version of the code. We currently improving on the float to double precision handling, that improves iteratively convergence performance of the code. Version 1.0-1.4 also includes a primitive graphical interface, used to plot heatmaps, topologically associating domains, significant interactions and Ensembl Genes. It is under current development and is based on the ANSI-C X11 graphical interface.