

epic2 and sicer give the same results

By default, epic2 and SICER should give slightly different results. There are three reasons for this:

1. epic2 chooses an effective genome fraction (EGF) that is slightly different than the EGF used by SICER.

2. SICER contains a logical bug: SICER includes reads falling outside of the genome boundaries in their count of total number of reads and this affects computed p-values (see below). epic2 has a flag that allows users to run the original SICER algorithm with the bug included (\texttt{--original-algorithm}). This flag is mainly included for debugging purposes, as this bug is undesired for regular ChIP-seq analyses.

3. SICER and epic2 identify out of bounds reads in different ways.

Differences in EGF

The effective genome length is the total length of mappable regions in the genome (Xu, et al., 2014) and depends on both the characteristics of the genome, such as the number of unknown nucleotides (Ns) and the amount of repeat sequences, and the characteristics of the sequencing library, such as the read length. The EGF is the effective genome length divided by the actual genome size.

epic2 contains four precomputed values of the EGF for ~80 genomes. These four values correspond to four different read lengths (36, 50, 75, 100); the values were computed by running jellyfish (Marcais and Kingsford, 2011) with the following options:

\begin{verbatim}
  jellyfish count -t 25 -m {read_length} -s {genome_length} -L 1 -U 1 \
  \qquad --out-counter-len 1 --counter-len 1 {fasta}
\end{verbatim}

Here, \{read_length\} is the read length, \{fasta\} is the FASTA file containing the genome sequence, and \{genome_length\} is the number of nucleotides in the genome FASTA file. The resulting values differ from those used in SICER, which uses values computed based on the approach described by Koehler, et al. (2011). To illustrate, SICER uses the default value of 0.74 for the hg18 version of the human genome, whereas epic2 uses values 0.8247546892509732, 0.8665593467306165, 0.8960340102187695, and 0.9054491980427741 for hg18 and read lengths 36, 50, 75, and 100, respectively.

The EGF is used in the Poisson background model of random reads and to modify the p-values for regions with an input count of zero (see below). Consequently, the choice of EGF affects which regions become candidate islands and the regions’ p-values.
Logical bug
SICER uses two factors computed from the global library characteristics to modify each region’s p-value. The scaling factor, which is the total number of ChIP reads (#ChIP) divided by the total number of input reads (#input), is multiplied with all p-values; the zero scaler, which is #input divided by EGF, is multiplied with the p-value for each region with an input count of zero. Because of SICER’s bug, #ChIP and #input can differ slightly between epic2 and SICER, resulting in differences in p-values and the ordering of regions between the two programs.

To illustrate these differences, we implemented a feature allowing epic2 to run with SICER’s original logical bug. This feature is accessed by providing the flag --original--algorithm when running epic2. For the tests described in the following sections, we ran epic2 with this flag.

Computing out of bounds reads
SICER discards all reads where the end falls outside of the chromosome boundaries. epic2 discards all reads where the corresponding DNA fragment’s estimated mid point falls outside of math.ceil(chromosome_size / bin_size)*bin_size. The consequence is that epic2 can have more reads than SICER has within the last bin of the chromosome, resulting in epic2 potentially identifying longer regions towards the end of the chromosomes than SICER does. This difference might also affect the ordering of the results, as the length of the region is used to give regions without input a pseudo count.

The tests
To illustrate that SICER and epic2 do indeed give the same results, we ran the two programs on three publicly available datasets. To avoid differences due to different EGF values, we used the same EGF for both SICER and epic2.

The three datasets used were:
1. the canonical test data (ChIP and input) included in the SICER software package;
2. H3K27me3 data (ChIP and input) from human aorta, provided by the Roadmap Epigenomics project; and
3. H3K27me3 and H3K4me3 data (ChIP and input) from the human keratinocyte cell line HaCaT.

These tests and their results should be reproducible by running the Snakemake workflow available here: https://github.com/endrebak/epic2_supplementaries/tree/master/workflows/show_same_results. Other user-provided files can be tested by updating the file sample_sheet.txt within this workflow. Note that all
files used in these tests are in BED format, as this is the only format SICER accepts. All three tests use ChIP and input data.

Ensuring that the regions are the same

To determine whether the regions detected by SICER and epic2 are exactly the same, we used the subtract operation from the bedtools program (Quinlan and Hall, 2010). In the following code, epic2_results.bed and SICER_results.bed are the output from epic2 and SICER, respectively.

```
bedtools subtract -a epic2_results.bed -b SICER_results.bed > locs_only_in_epic2.bed
bedtools subtract -b epic2_results.bed -a SICER_results.bed > locs_only_in_SICER.bed
```

If there is any difference in the output from epic2 and SICER with respect to location, one or both of the output files (locs_only_in_epic2.bed or locs_only_in_SICER.bed) will be non-empty.

Ensuring the ordering is the same

To test whether the regions identified by SICER and epic2 have the same ranks, as determined by the regions’ p-values, we used the unix tool diff to compare the FDR-sorted list of significant regions from SICER and epic2 to ensure these were exactly the same.

Parameters used in the tests

- EGF: 0.85
- Genome: hg38
- Remove all duplicate reads
- Bin size: 200
- Max gaps: 3 (or 600 nucleotides)
- FDR-cutoff: 0.05
- –original-algorithm

Test 1: SICER example data

ChIP: https://raw.githubusercontent.com/biocore-ntnu/epic2/master/examples/test.bed
Input: https://raw.githubusercontent.com/biocore-ntnu/epic2/master/examples/control.bed

Here, epic2 finds 166 islands and so does SICER. They have the exact same sort order.
Test 2: Roadmap Epigenomics aorta data


Here epic2 finds one region more than SICER does. The region is located at the end of chromosome 15 (length chr15: 101991189). Remember that SICER has a stricter out-of-bounds cutoff than epic2 so this is to be expected.

The other differences are the following (output from `diff`):

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>4539</td>
<td>a</td>
</tr>
<tr>
<td>4540</td>
<td></td>
</tr>
<tr>
<td>&gt;</td>
<td>chr6</td>
</tr>
<tr>
<td>7666</td>
<td>d</td>
</tr>
<tr>
<td>7666</td>
<td></td>
</tr>
<tr>
<td>&lt;</td>
<td>chr6</td>
</tr>
<tr>
<td>7832</td>
<td>d</td>
</tr>
<tr>
<td>7832</td>
<td></td>
</tr>
<tr>
<td>&lt;</td>
<td>chr8</td>
</tr>
<tr>
<td>8080</td>
<td>a</td>
</tr>
<tr>
<td>8080</td>
<td></td>
</tr>
<tr>
<td>&gt;</td>
<td>chr8</td>
</tr>
<tr>
<td>43489</td>
<td>d</td>
</tr>
<tr>
<td>43489</td>
<td></td>
</tr>
</tbody>
</table>

For these two regions, epic2 and SICER disagree slightly on the length of the regions. Both regions are close to the chromosome boundaries (chr6: 170805979, chr8: 145138636).

Test 3: HaCaT dataset

In this test we used two ChIP-files and one Input file. We are mixing H3K27me3 and H3K4me3, which is something you should not do in practice, but this does not matter when ensuring that the programs produce the same results.

Collection and citation info: https://zenodo.org/record/2548491/

H3K27me3.bed.gz: https://zenodo.org/record/2548491/files/Satrom-H3K27me3.bed.gz
H3K4me3.bed.gz: https://zenodo.org/record/2548491/files/Satrom-H3K4me3.bed.gz
Input.bed.gz: https://zenodo.org/record/2548491/files/Satrom-Input.bed.gz

The islands are the same in number (342411) and exactly equal down to the last decimal:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>head −2</td>
<td></td>
</tr>
<tr>
<td>sicer_results/satrom/satrom_chip--W200--G600--islands--summary--FDR1.0</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>28400</td>
</tr>
<tr>
<td>6.43921723185</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>38000</td>
</tr>
<tr>
<td>1.21630995399</td>
<td></td>
</tr>
<tr>
<td>head −3</td>
<td></td>
</tr>
<tr>
<td>epic2_results/satrom/fdr_list.csv</td>
<td></td>
</tr>
<tr>
<td>#Chromosome Start End PValue Score Strand ChIPCount InputCount FDR log2FoldChange</td>
<td></td>
</tr>
</tbody>
</table>

4
There are no differences in ordering.

References


