Comparison of seqOutATACBias and ATACorrect bias correction output

Jacob B. Wolpe* Michael J. Guertin†

Abstract

This vignette outlines an analysis which compares output from ATACorrect and seqOutATACBias. We find that ATACorrect does not scale individual reads, but instead removes all reads outside of input peak regions, optimizing data for later footprinting and downstream analysis of a narrowly defined region set. ATACorrect also creates simulated read data within input peak regions to further offset Tn5 bias within these defined genomic locations. Finally, ATACorrect bias correction itself is based on the input peak regions, and changes when these regions change. Because ATACorrect requires these input peak regions, it is a tool highly optimized for footprinting analysis, but does not broadly correct observed Tn5 sequence bias in a data set. In contrast, seqOutATACBias operates by scaling all individual reads of a data set to correct Tn5 local and regional bias. This bias correction is useful for any application of the input data set going forward, including identification of open genomic regions for further investigation.

Contents

*Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America
†Department of Genetics and Genome Sciences, University of Connecticut, Farmington, Connecticut, United States of America

1 Foreword

This vignette examines the differences between seqOutATACBias and ATACorrect (a TOBIAS tool) output from a test set of read data on chromosome 21. It requires that seqOutATACBias and its dependencies are in path. Because seqOutATACBias and TOBIAS are written in different python versions, a conda virtual environment (python 3.7, named 'TOBIAS_venv') is used to install and run TOBIAS, which must also be in path using this method. If you wish to reproduce this analysis, you **must** have a conda virtual environment named 'TOBIAS_venv' with TOBIAS installed. This vignette is split into 3 sections: a check to make sure software dependencies are installed, running seqOutATACBias and ATACorrect on identical input reads and reference genome, and finally analysis of their output files. Results from this analysis reveal that while both tools operate on similar input, their output is drastically different, as is their intended use.

2 Installations

In order to run this vignette, you must have the following installed and added to PATH:
seqOutBias (https://github.com/guertinlab/seqOutBias/archive/refs/heads/master.zip)
Rust >= 1.32.0
genometools >= 1.6.1
pyfaidx >= 0.7.1
GNU parallel >= 20220722
GNU wget >= 1.21.3
bedtools >= 2.30.0
bigWigToBedGraph >= 438
bedGraphToBigWig >= 2.9
wigToBigWig >= 2.8 Pandoc >= 2.19.2
conda 22.11.1
TOBIAS 0.15.1
R >= 4.2.1
R Packages:
- R data.table package >= 1.14.2
- bigWig R package

Check to see if you have the required dependencies in PATH. The following will print a message if a dependency cannot be called:

```
if ! command -v wget &> /dev/null
then
    echo "wget could not be found"
elif ! command -v faidx &> /dev/null
then
    echo "faidx could not be found"
elif ! command -v parallel &> /dev/null
then
    echo "GNU parallel could not be found"
elif ! command -v bigWigToBedGraph &> /dev/null
then
    echo "bigWigToBedGraph could not be found"
elif ! command -v bedGraphToBigWig &> /dev/null
then
    echo "bedGraphToBigWig could not be found"
elif ! command -v gt &> /dev/null
then
    echo "Genome tools could not be found"
elif ! command -v rustc &> /dev/null
then
    echo "Rust could not be found"
elif ! command -v seqOutBias &> /dev/null
```

```
  then
      echo "seqOutBias could not be found"
  elif ! command -v wigToBigWig &> /dev/null
  then
      echo "wigToBigWig could not be found"
  elif ! command -v seqOutATACBias &> /dev/null
  then
      echo "seqOutATACBias could not be found"
  elif ! command -v conda &> /dev/null
  then
      echo "conda could not be found"
  else
      source activate TOBIAS_venv
  fi
  if ! command -v TOBIAS &> /dev/null
  then
      echo "TOBIAS could not be found"
  else
      echo "Checked dependencies installed"
  fi
```

```
## Checked dependencies installed
```

If you find that any of these dependencies are not in PATH, you may install them from the following:

seqOutBias: https://github.com/guertinlab/seqOutBias/archive/refs/heads/master.zip seqOutATACBias: https://github.com/guertinlab/Tn5bias/tree/master/seqOutATACBias_setup
Rust: https://www.rust-lang.org/
genometools: http://genometools.org/
R: https://rstudio-education.github.io/hopr/starting.html
pyfaidx: https://pypi.org/project/pyfaidx/
GNU parallel: https://www.gnu.org/software/parallel/
bedtools: https://bedtools.readthedocs.io/en/latest/
bigWigToBedGraph: http://hgdownload.soe.ucsc.edu/admin/exe/
bedGraphToBigWig: http://hgdownload.soe.ucsc.edu/admin/exe/
bigWig R package: https://github.com/guertinlab/bigWig
wigToBigWig: https://anaconda.org/bioconda/ucsc-wigtobigwig
GNU wget: https://www.gnu.org/software/wget/
conda: https://docs.conda.io/projects/conda/en/latest/user-guide/install/macos.html
TOBIAS: https://github.com/loosolab/TOBIAS

2.1 Auto-install R packages

Install the `data.table`, `bigWig`, `devtools`, and `eulerr` R packages, if necessary:

```
tabletest = require(data.table)
```

```
## Loading required package: data.table
```

```
if(tabletest==FALSE){
  install.packages('data.table')
}
bigWigtest = require(bigWig)
```

```
## Loading required package: bigWig
```

```
if(bigWigtest==FALSE){
  install.packages('devtools')
  devtools::install_github("andrelmartins/bigWig", subdir="bigWig")
}
```

3 Generating output from seqOutATACBias and ATACorrect

This section prepares the data to compare the output from seqOutATACBias and ATACorrect. The first section downloads the chromosome 21 reference genome (hg38), aligned unscaled chromosome 21 read files in BAM format, and FIMO results for ESR1 on chromosome 21 from cyverse. Next, we use each model to generate the output that will be analyzed and compared.

3.1 Downloading reference genome and read data.

Download the reference genome for chromosome 21 (hg38_chr21.fa), aligned deproteinized ATAC-seq read file from cyverse (C1_gDNA_rep1_chr21.bam), and ESR1 motifs for chromosome 21 (ESR1_rm_chr21_fimo.txt).

```
#To test this vignette with a subset (chr 21) genome and reads:
wget -nv https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/C1_gDNA_rep1_chr21.bam
wget -nv https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/hg38_chr21.fa
wget -nv https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/ESR1_rm_chr21_fimo.txt
wget -nv https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/ESR1_REST_chr21_fimo.txt
```

```
## 2023-03-07 16:17:53 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/C1_gDNA_rep1_chr21.bam [2
## 2023-03-07 16:19:07 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/hg38_chr21.fa [47488490/4
## 2023-03-07 16:19:10 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/ESR1_rm_chr21_fimo.txt [9
## 2023-03-07 16:19:17 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/ESR1_REST_chr21_fimo.txt
```

3.2 Run seqOutATACBias to generate output

Here we run seqOutATACBias to scale ATAC-seq reads using our rule ensemble model.

```
seqOutATACBias masks -i=C1_gDNA_rep1_chr21.bam -g=hg38_chr21.fa -p=3 -r=72

seqOutATACBias scale -i=C1_gDNA_rep1_chr21_union.bedGraph -g=hg38_chr21.fa
```

3.3 Run ATACorrect to generate output

We now use ATACorrect to generate output for later comparison. Because ATACorrect requires the genomic coordinates of interest as an input field ('peaks' file), we first convert the downloaded coordinates from FIMO format to bed.

```
options(scipen = 100)
source('https://raw.githubusercontent.com/guertinlab/Tn5bias/master/Manuscript_Vignette/Vignette_Scripts/Tn5_
library(data.table)

bed_peaks = FIMO.to.BED('ESR1_rm_chr21_fimo.txt')
write.table(bed_peaks[,1:4],file= 'ESR1_chr21.bed',
              sep = '\t', quote = FALSE, row.names = FALSE, col.names = FALSE)

bed_peaks = FIMO.to.BED('ESR1_REST_chr21_fimo.txt')
write.table(bed_peaks[,1:4],file= 'ESR1_REST_chr21.bed',
              sep = '\t', quote = FALSE, row.names = FALSE, col.names = FALSE)
```

Next, we run ATACorrect using the newly generated bed files and previously downloaded data. The first run of ATACorrect uses all identified ESR1 motifs on chromosome 21. The second run includes REST motifs in the peak file, in addition to the ESR1 motifs. Once ATACorrect has finished running, we convert the bigwig output into bedgraph format for comparison.

```
source activate TOBIAS_venv
TOBIAS ATACorrect --bam C1_gDNA_rep1_chr21.bam --genome hg38_chr21.fa --peaks ESR1_chr21.bed --outdir ATACorr
bigWigToBedGraph ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_corrected.bw ATACorrect_ESR1_output/C1_gDNA_rep1_c
bigWigToBedGraph ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_uncorrected.bw ATACorrect_ESR1_output/C1_gDNA_rep1

TOBIAS ATACorrect --bam C1_gDNA_rep1_chr21.bam --genome hg38_chr21.fa --peaks ESR1_REST_chr21.bed --outdir AT
bigWigToBedGraph ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_corrected.bw ATACorrect_ESR1_REST_output/C1_g
bigWigToBedGraph ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_uncorrected.bw ATACorrect_ESR1_REST_output/C1
```

4 Output analysis

This section compares the output and potential uses of both seqOutATACBias and ATACorrect.

4.1 ATACorrect scaling and read depth is only applied to supplied peak regions

To illustrate the differences between seqOutATACBias scaling and ATACorrect footprinting correction, we first download an illustrative browser snapshot from the following UCSC browser session:
https://genome.ucsc.edu/s/JacobWolpe/sOAB%20ATACorrect%20Raw%20Comparison

```
wget -nv https://github.com/guertinlab/Tn5bias/raw/master/Manuscript_Vignette/figs/sOAB_ATACorrect_comparison
```

```
## 2023-03-07 16:33:14 URL:https://raw.githubusercontent.com/guertinlab/Tn5bias/master/Manuscript_Vignette/fi
```

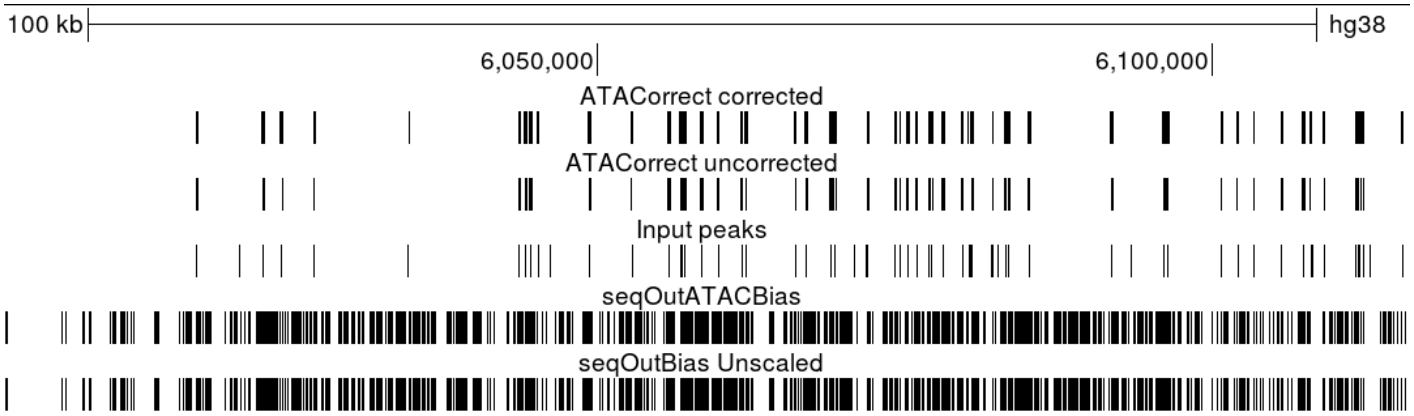

Figure 1: Browser snapshot of unscaled seqOutBias output, seqOutATACBias output, uncorrected ATACorrect output, corrected ATACorrect output, and the 'peaks' file (Input peaks) used as input for the two ATACorrect output files

The browser snapshot seen in Figure 1 exemplifies the incompatibility between seqOutATACBias and ATACorrect outputs. In this image, we see that ATACorrect corrected and uncorrected output is only present at genomic locations contained in the input `peaks` file (ESR1_chr21). This is in contrast to seqOutATACBias scaled output, which is present at all read locations in the original input `BAM` file. seqOutATACBias was designed as a general method used to scale individual reads of a given data set to reduce the Tn5 sequence bias of each read. ATACorrect refines input data to the specific regions of interest, removes all other reads and adds theoretically unbiased reads not in the original data set to optimize footprinting algorithms. This is why the evaluation of both methods is so drastically different: ATACorrect was measured for its ability to correctly uncover footprints and improve footprint depth; seqOutATACBias was measured for its ability to return average signal to theoretical random cleavage. Next, we show that simulated ATACorrect reads are confined to the regions nearby the input `peaks` file.

4.2 ATACorrect simulated data is within 175 base pairs of the peaks file regions

To determine what percent of simulated data produced by ATACorrect is within `peak` regions, we first expand the peak regions by 175 base pairs in both directions.

```
peaks = fread('ESR1_chr21.bed')
peaks$V2 = peaks$V2 - 175
peaks$V3 = peaks$V3 + 175
write.table(peaks[,1:4],file= 'ESR1_chr21_175bp.bed',
              sep = '\t', quote = FALSE, row.names = FALSE, col.names = FALSE)
```

Next, we use `bedtools intersect` to remove all uncorrected reads from the corrected reads file.

```
bedtools intersect -v -a ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_corrected.bedgraph -b ATACorrect_ESR1_outp
```

Now, we determine which of these simulated reads are within the bounds of the `peaks` file we generated.

```
bedtools intersect -v -a C1_gDNA_rep1_chr21_corrected_NO_UNCOR_READS.bedgraph -b ESR1_chr21_175bp.bed > ATACo
ls -l ATACorrect_175bp_simulated_peaks.bed
```

```
## -rw-r--r--  1 guertinlab  staff  0 Mar  7 16:33 ATACorrect_175bp_simulated_peaks.bed
```

As this file is empty, no ATACorrect simulated reads are farther than 175 base pairs from a region in the original `peaks` file. This result shows that ATACorrect simulated output is entirely based on the input peak regions, to the exclusion of all other data within the input `BAM` file.

We can visualize this by zooming in on the previous browser snap shot. First, we download the image.

```
wget -nv https://github.com/guertinlab/Tn5bias/raw/master/Manuscript_Vignette/figs/sOAB_ATACorrect_SimReads.p
```

```
## 2023-03-07 16:33:16 URL:https://raw.githubusercontent.com/guertinlab/Tn5bias/master/Manuscript_Vignette/fi
```

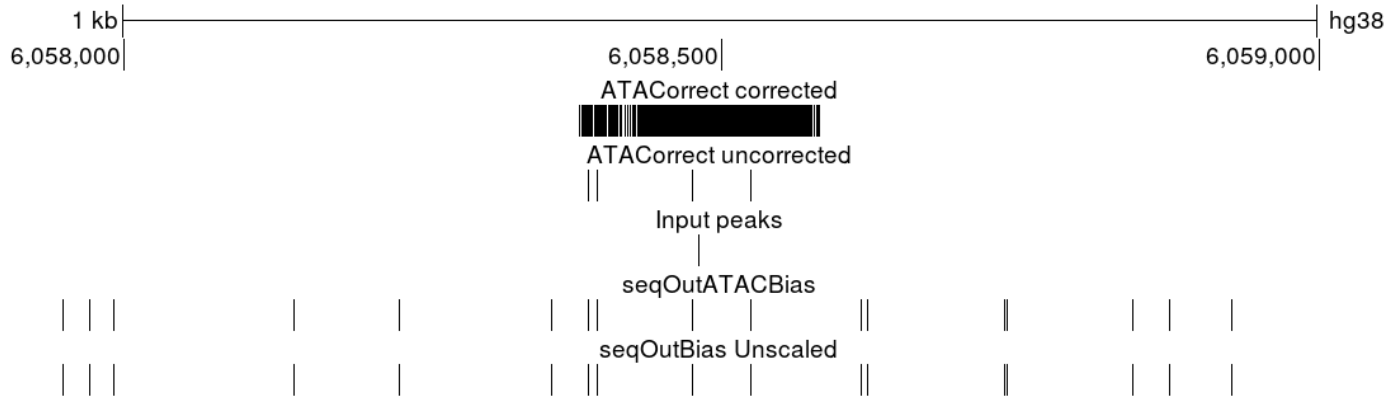

Figure 2: Zoomed in browser snapshot of unscaled seqOutBias output, seqOutATACBias output, uncorrected ATACorrect output, corrected ATACorrect output, and the 'peaks' file (Input peaks) used as input for the two ATACorrect output files

This shows that ATACorrect corrected output within 175bp of the coordinates within the peaks file is heavily dependent on simulated data, which is not found in either the ATACorrect unscaled output, seqOutATACBias, or seqOutBias unscaled output.

4.3 ATACorrect bias correction changes based upon which peaks are input

To show that ATACorrect scaling is dependent on the input peaks file, we first remove all simulated data from the ATACorrect output. Next, we only examine reads in the ESR1 peak file, as regions in the REST peak file will not be present in both.

```
bedtools intersect -a ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_corrected.bedgraph -b ATACorrect_ESR1_RE

bedtools intersect -a ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_corrected.bedgraph -b ATACorrect_ESR1_output/

bedtools intersect -a ESR1_REST_C1_gDNA_rep1_chr21_corrected_NO_SIM_READS.bedgraph -b ESR1_C1_gDNA_rep1_chr21

bedtools intersect -a ESR1_C1_gDNA_rep1_chr21_corrected_NO_SIM_READS.bedgraph -b ESR1_REST_C1_gDNA_rep1_chr21
```

Finally, we plot these scaled reads.

```
ESR1_reads = fread('ESR1_C1_gDNA_rep1_chr21_corrected_NO_SIM_REST_READS.bedgraph')
ESR1_REST_reads = fread('ESR1_REST_C1_gDNA_rep1_chr21_corrected_NO_SIM_REST_READS.bedgraph')

pdf(file = 'sOAB_ATACorrect_scaled_read_comparison.pdf')
plot(ESR1_reads$V4, ESR1_REST_reads$V4, xlab = 'ESR1 Scaled Reads', ylab = 'ESR1 REST Scaled Reads', pch = 16
dev.off()
```

```
## pdf
##   2
```

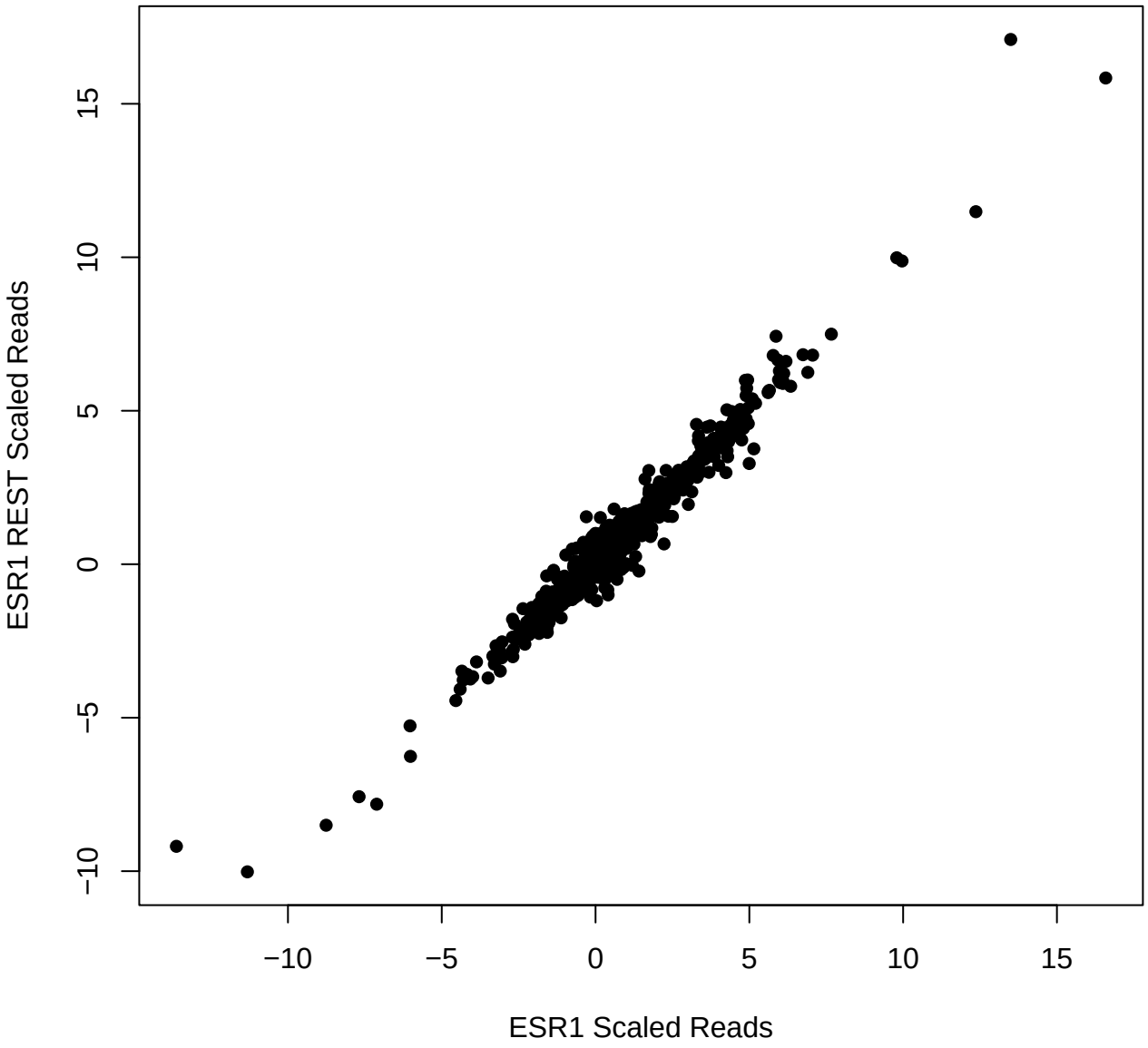

Figure 3: Dot plot of ATACorrect scaled read signal in the ESR1 peak regions from both ESR1 input and combined ESR1 and REST regions.

The above figure shows that ATACorrect scaling is dependent on, and changes with the input peak regions. Thus, comparison between seqOutATACBias and ATACorrect is altered based on which regions are used as input into ATACorrect. In summary, this analysis shows several factors which inhibit direct comparison between ATACorrect and seqOutATACBias. First, ATACorrect requires a `peaks bed` file input so that output is confined to these regions and all other output is lost, whereas seqOutATACBias scales all individual reads in a data set and gives their scaled values as output. Second, seqOutATACBias only operates on input reads, while ATACorrect creates simulated reads, expanding the size of the input data set within the regions of interest. As such, seqOutATACBias and seqOutBias are the only two software packages known to scale all individual reads of an input data set, adding new reads, or removing other reads, and which gives this output in a single file for implementation in subsequent, nonrestricted analysis.