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. 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

1 Foreword **2**

2 Installations **2**

2.1 Auto-install R packages 3

3 Generating output from seqOutATACBias and ATACorrect **4**

3.1 Downloading reference genome and read data. 4

3.2 Run seqOutATACBias to generate output 4

3.3 Run ATACorrect to generate output 9

4 Output analysis **15**

4.1 Read depth 15

4.2 Read number 17

4.3 Interval set coordinate comparison 19

4.4 ATACorrect scaling and read depth is only applied to supplied peak regions 23

4.5 ATACorrect simulated data is within 175 base pairs of the `peaks` file regions 23

*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")
}
eulertest = require(eulerr)
```

```
## Loading required package: eulerr
```

```
if(eulertest==FALSE){
  install.packages('eulerr')
}
```

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, FIMO results for ESR1 on chromosome 21 and combined ESR1/REST FIMO results for 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), ESR1 motifs for chromosome 21 (ESR1_rm_chr21_fimo.txt), and combined ESR1/REST motifs for chromosome 21 (ESR1_REST_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
```

```
## 2023-02-09 17:24:42 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/C1_gDNA_rep1_chr21.bam [20970
## 2023-02-09 17:24:52 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/hg38_chr21.fa [47488490/47488
## 2023-02-09 17:24:53 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/ESR1_rm_chr21_fimo.txt [92464
## 2023-02-09 17:24:54 URL:https://data.cyverse.org/dav-anon/iplant/home/jacobwolpe/ESR1_REST_chr21_fimo.txt [188
```

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
```

```
## Cleanup (-c or --cleanup)= TRUE
## Read Length (-r or --readlength)= 72
## Input (-i or --input)= C1_gDNA_rep1_chr21.bam
## Genome (-g or --genome)= hg38_chr21.fa
## Processors (-p or --processors)= 3
## Command = masks
## Outfile = C1_gDNA_rep1_chr21
## Starting mask generation...
## Starting Mask Generation
## ################### Creating mappability file using tallymer ###################
## # dna=yes
## # indexname="hg38_chr21.sft"
## # prefixlength=automatic
## # storespecialcodes=false
## # inputfile[0]=hg38_chr21.fa
## # indexname=hg38_chr21.sft
## # outtistab=true,outsuftab=true,outlcptab=true,outbwttab=false,outbcktab=false,outdestab=true,outsdstab=true,o
## # parts=4
## # maxinsertionsort=3
## # maxbltriesort=1000
## # maxcountingsort=4000
```

```
## # lcpdist=false
## # sizeof (GtUword)=64
## # wildcardranges of length 1=4
## # wildcardranges of length 10=4
## # wildcardranges of length 20=1
## # wildcardranges of length 100=13
## # wildcardranges of length 10000=1
## # wildcardranges of length 50000=25
## # wildcardranges of length 100000=2
## # wildcardranges of length 150000=1
## # wildcardranges of length 5010000=1
## # init character encoding (uint32, 11678192 bytes, 2.00 bits/symbol)
## # totallength=46709983
## # numofsequences=1
## # specialcharacters=6621364
## # specialranges=52
## # realspecialranges=52
## # wildcards=6621364
## # wildcardranges=52
## # realwildcardranges=52
## # occurrences(a)=11820664
## # occurrences(c)=8185244
## # occurrences(g)=8226381
## # occurrences(t)=11856330
## # automatically determined prefixlength=10
## # maxinsertionsort=3
## # maxbltriesort=1000
## # maxcountingsort=4000
## # storespecialcodes=false
## # cmpcharbychar=false
## # totallength=46709983
## # sizeof (leftborder)=4194308 bytes
## # sizeof (countspecialcodes)=1048576 bytes
## # sizeof (distpfxidx)=349520 bytes
## # sizeof (bcktab)=5592404 bytes
## # largest bucket size=35844
## # widthofpart[0]=10022175
## # widthofpart[1]=10022207
## # widthofpart[2]=10022099
## # widthofpart[3]=10022138
## # create suffix_sort_space: suftab uses 64bit values: maxvalue=46709983,numofentries=10022207
## # compute part 0: 10022175 suffixes,228270 buckets from 0..228269
## # used workspace for sorting: 0.11 MB
## # countinsertionsort=20391
## # countbltriesort=200224
## # countcountingsort=403
## # countshortreadsort=0
## # countradixsort=0
## # counttqsort=36
## # compute part 1: 10022207 suffixes,296263 buckets from 228270..524532
## # countinsertionsort=33162
## # countbltriesort=241342
## # countcountingsort=353
## # countshortreadsort=0
## # countradixsort=0
## # counttqsort=22
## # compute part 2: 10022099 suffixes,312559 buckets from 524533..837091
## # countinsertionsort=32166
## # countbltriesort=263259
## # countcountingsort=337
## # countshortreadsort=0
## # countradixsort=0
```

```
## # counttqsort=21
## # compute part 3: 10022138 suffixes,211484 buckets from 837092..1048575
## # countinsertionsort=19282
## # countbltriesort=186005
## # countcountingsort=435
## # countshortreadsort=0
## # countradixsort=0
## # counttqsort=35
## # construct mer buckets for prefixlength 7
## # numofcodes = 16384
## # indexfilename = hg38_chr21.tal_72
## # alphasize = 4
## # mersize = 72
## # numofmers = 1109447
## # merbytes = 18
## ################################################################################
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 4
## # plus-offset: 2
## # minus-offset: 2
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced hg38_chr21_72.4.2.2.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_not_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_unscaled.bigWig
## Clean up C1_gDNA_rep1_chr21_unscaled
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNNXXXXXXXXXXXXXXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNNXXXXXXXXXXXXXXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNNXXXXXXXXXXXXXXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNXXXXXXXXXXXXXXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNXXXXXXXXXXXXXXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNXXXXXXXXXXXXXXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNNXXXXXXXXXXXXXXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNNXXXXXXXXXXXXXXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNNCNNXXXXXXXXXXXXXXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
```

```
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXNNNCNNNNXXXXXXXXXXXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXNNNCNNNNXXXXXXXXXXXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXNNNCNNNNXXXXXXXXXXXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXNNCNNNNNXXXXXXXXXXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXNNCNNNNNXXXXXXXXXXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXNNCNNNNNXXXXXXXXXXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCNNNNNNNXXXXXXXXXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCNNNNNNNXXXXXXXXXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCNNNNNNNXXXXXXXXXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXNNNNNNNNXXXXXXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXNNNNNNNNXXXXXXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXNNNNNNNNXXXXXXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXNNNNNNNXXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXNNNNNNNXXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXNNNNNNNXXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
```

```
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXXNNNNNNNXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXXNNNNNNNXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXXNNNNNNNXXXX.bigWig
## # tallymer produced/found hg38_chr21.tal_72.gtTxt.gz
## # kmer-size: 46
## # plus-offset: 23
## # minus-offset: 23
## # chrom: "chr21"
## # - 46709984 bases
## # seqtable produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXNNNNNNNXXXXX.tbl
## # tabulate C1_gDNA_rep1_chr21.bam
## # scale C1_gDNA_rep1_chr21.bam
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXNNNNNNNXXXXX_scaled.bed
## # scale produced C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXNNNNNNNXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNXXXXXXXXXXXXXXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNXXXXXXXXXXXXXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNNXXXXXXXXXXXXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNNNXXXXXXXXXXXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNNNNXXXXXXXXXXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCNNNNNNNXXXXXXXXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXNNNNNNNXXXXXXXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXNNNNNNNXXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXNNNNNNNXXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXNNNNNNNXXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXNNNNNNNXXXXX
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNXXXXXXXXXXXXXXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNXXXXXXXXXXXXXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNNXXXXXXXXXXXXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNNNXXXXXXXXXXXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNNNNXXXXXXXXXXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCNNNNNNNXXXXXXXXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXNNNNNNNXXXXXXXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXNNNNNNNXXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXNNNNNNNXXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXNNNNNNNXXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXNNNNNNNXXXXX.bigWig
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNXXXXXXXXXXXXXXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXNNNNNNCNNXXXXXXXXXXXXXXXXXXXXX.bedGraph
```

```
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNNCNNNXXXXXXXXXXXXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXNNNCNNNNXXXXXXXXXXXXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXNNCNNNNNXXXXXXXXXXXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCNNNNNNNXXXXXXXXXXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXNNNNNNNNXXXXXXXXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXNNNNNNNXXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXXNNNNNNNXXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXNNNNNNNXXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXNNNNNNNXXXXXXX.bedGraph
## Clean up C1_gDNA_rep1_chr21_XXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXXXNNNNNNNXXXXXX.bedGraph
## Cleanup (-c or --cleanup)= TRUE
## Read Length (-r or --readlength)= 60
## Input (-i or --input)= C1_gDNA_rep1_chr21_union.bedGraph
## Genome (-g or --genome)= hg38_chr21.fa
## Processors (-p or --processors)=
## Command = scale
## Outfile = C1_gDNA_rep1_chr21_union
## Starting rule ensemble scaling...
## R version 4.2.1 (2022-06-23)
## [1] "Reading unscaled bed file..."
## [1] "Reading bed file C1_gDNA_rep1_chr21_union.bedGraph"
## [1] "Applying rule ensemble model"
## [1] "Writing rule ensemble scaled bed file..."
## [1] "Writing rule ensemble scaled bigWig file..."
```

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_Bias
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 ATACorrect_
bigWigToBedGraph ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_corrected.bw ATACorrect_ESR1_output/C1_gDNA_rep1_chr21
bigWigToBedGraph ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_uncorrected.bw ATACorrect_ESR1_output/C1_gDNA_rep1_chr

TOBIAS ATACorrect --bam C1_gDNA_rep1_chr21.bam --genome hg38_chr21.fa --peaks ESR1_REST_chr21.bed --outdir ATACor
bigWigToBedGraph ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_corrected.bw ATACorrect_ESR1_REST_output/C1_gDNA_
bigWigToBedGraph ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_uncorrected.bw ATACorrect_ESR1_REST_output/C1_gDN
```

```
## # TOBIAS 0.15.1 ATACorrect (run started 2023-02-09 17:35:24.534663)
## # Working directory: /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3
## # Command line call: TOBIAS ATACorrect --bam C1_gDNA_rep1_chr21.bam --genome hg38_chr21.fa --peaks ESR1_chr21.
##
## # ----- Input parameters -----
## # bam:   C1_gDNA_rep1_chr21.bam
```

```
## # genome:    hg38_chr21.fa
## # peaks: ESR1_chr21.bed
## # regions_in:    None
## # regions_out:   None
## # blacklist: None
## # extend:    100
## # split_strands: False
## # norm_off:  True
## # track_off: []
## # drop_chroms:   [’chrM’, ’chrMT’, ’M’, ’MT’, ’Mito’]
## # k_flank:   12
## # read_shift:    [4, -5]
## # bg_shift:  100
## # window:    100
## # score_mat: DWM
## # bias_pkl:  None
## # prefix:    C1_gDNA_rep1_chr21
## # outdir:    /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_o
## # cores: 3
## # split: 100
## # verbosity: 3
##
##
## # ----- Output files -----
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_output/C1_gD
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_output/C1_gD
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_output/C1_gD
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_output/C1_gD
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_output/C1_gD
##
##
## 2023-02-09 17:35:24 (41790) [INFO]   ----- Processing input data -----
## 2023-02-09 17:35:24 (41790) [INFO]   Reading info from .bam file
## [E::idx_find_and_load] Could not retrieve index file for ’C1_gDNA_rep1_chr21.bam’
## 2023-02-09 17:35:24 (41790) [WARNING]    No index found for bamfile - creating one via pysam.
## 2023-02-09 17:35:24 (41790) [INFO]   Reading info from .fasta file
## 2023-02-09 17:35:24 (41790) [WARNING]    No additional chromosomes were removed. Consider using ’--drop-chroms
## 2023-02-09 17:35:24 (41790) [INFO]   Processing input/output regions
## 2023-02-09 17:35:25 (41790) [STATS]  genome: 1 regions | 46709983 bp | 100.00% coverage
## 2023-02-09 17:35:25 (41790) [STATS]  input_regions: 8791 regions | 46701394 bp | 99.98% coverage
## 2023-02-09 17:35:25 (41790) [STATS]  output_regions: 7144 regions | 1640423 bp | 3.51% coverage
## 2023-02-09 17:35:25 (41790) [STATS]  peak_regions: 8589 regions | 8589 bp | 0.02% coverage
## 2023-02-09 17:35:25 (41790) [STATS]  nonpeak_regions: 8791 regions | 46701394 bp | 99.98% coverage
## 2023-02-09 17:35:25 (41790) [STATS]  blacklist_regions: 0 regions | 0 bp | 0.00% coverage
##
## 2023-02-09 17:35:25 (41790) [INFO]   ----- Estimating normalization factors -----
## 2023-02-09 17:35:25 (41790) [INFO]   Normalization was switched off
## 2023-02-09 17:35:25 (41790) [STATS]  CORRECTION_FACTOR:  1.00000
##
## 2023-02-09 17:35:25 (41790) [INFO]   Started estimation of sequence bias...
## 2023-02-09 17:35:25 (41790) [INFO]   Progress: 0%
## 2023-02-09 17:35:25 (41790) [INFO]   Progress: 9%
## 2023-02-09 17:35:26 (41790) [INFO]   Progress: 15%
## 2023-02-09 17:35:26 (41790) [INFO]   Progress: 29%
## 2023-02-09 17:35:27 (41790) [INFO]   Progress: 47%
## 2023-02-09 17:35:27 (41790) [INFO]   Progress: 63%
## 2023-02-09 17:35:28 (41790) [INFO]   Progress: 80%
## 2023-02-09 17:35:28 (41790) [INFO]   Progress: 95%
## 2023-02-09 17:35:29 (41790) [INFO]   Progress: 100%
## 2023-02-09 17:35:29 (41790) [INFO]   Finalizing bias motif for scoring
##
## 2023-02-09 17:35:29 (41790) [INFO]   ----- Correcting reads from .bam within output regions -----
```

```
## 2023-02-09 17:35:29 (41790) [INFO]   Correction progress: 0%
## 2023-02-09 17:35:31 (41790) [INFO]   Correction progress: 1.0%
## 2023-02-09 17:35:31 (41790) [INFO]   Correction progress: 2.0%
## 2023-02-09 17:35:32 (41790) [INFO]   Correction progress: 3.0%
## 2023-02-09 17:35:33 (41790) [INFO]   Correction progress: 4.0%
## 2023-02-09 17:35:34 (41790) [INFO]   Correction progress: 5.0%
## 2023-02-09 17:35:34 (41790) [INFO]   Correction progress: 6.0%
## 2023-02-09 17:35:35 (41790) [INFO]   Correction progress: 7.0%
## 2023-02-09 17:35:36 (41790) [INFO]   Correction progress: 8.0%
## 2023-02-09 17:35:37 (41790) [INFO]   Correction progress: 10.0%
## 2023-02-09 17:35:38 (41790) [INFO]   Correction progress: 11.0%
## 2023-02-09 17:35:39 (41790) [INFO]   Correction progress: 12.0%
## 2023-02-09 17:35:39 (41790) [INFO]   Correction progress: 13.0%
## 2023-02-09 17:35:40 (41790) [INFO]   Correction progress: 14.0%
## 2023-02-09 17:35:41 (41790) [INFO]   Correction progress: 15.0%
## 2023-02-09 17:35:41 (41790) [INFO]   Correction progress: 16.0%
## 2023-02-09 17:35:42 (41790) [INFO]   Correction progress: 18.0%
## 2023-02-09 17:35:43 (41790) [INFO]   Correction progress: 20.0%
## 2023-02-09 17:35:44 (41790) [INFO]   Correction progress: 22.0%
## 2023-02-09 17:35:45 (41790) [INFO]   Correction progress: 23.0%
## 2023-02-09 17:35:45 (41790) [INFO]   Correction progress: 24.0%
## 2023-02-09 17:35:46 (41790) [INFO]   Correction progress: 25.0%
## 2023-02-09 17:35:46 (41790) [INFO]   Correction progress: 26.0%
## 2023-02-09 17:35:47 (41790) [INFO]   Correction progress: 27.0%
## 2023-02-09 17:35:47 (41790) [INFO]   Correction progress: 28.0%
## 2023-02-09 17:35:48 (41790) [INFO]   Correction progress: 29.0%
## 2023-02-09 17:35:48 (41790) [INFO]   Correction progress: 30.0%
## 2023-02-09 17:35:49 (41790) [INFO]   Correction progress: 31.0%
## 2023-02-09 17:35:50 (41790) [INFO]   Correction progress: 32.0%
## 2023-02-09 17:35:50 (41790) [INFO]   Correction progress: 33.0%
## 2023-02-09 17:35:51 (41790) [INFO]   Correction progress: 34.0%
## 2023-02-09 17:35:51 (41790) [INFO]   Correction progress: 35.0%
## 2023-02-09 17:35:52 (41790) [INFO]   Correction progress: 36.0%
## 2023-02-09 17:35:53 (41790) [INFO]   Correction progress: 37.0%
## 2023-02-09 17:35:53 (41790) [INFO]   Correction progress: 38.0%
## 2023-02-09 17:35:54 (41790) [INFO]   Correction progress: 39.0%
## 2023-02-09 17:35:54 (41790) [INFO]   Correction progress: 40.0%
## 2023-02-09 17:35:55 (41790) [INFO]   Correction progress: 41.0%
## 2023-02-09 17:35:56 (41790) [INFO]   Correction progress: 42.0%
## 2023-02-09 17:35:56 (41790) [INFO]   Correction progress: 43.0%
## 2023-02-09 17:35:57 (41790) [INFO]   Correction progress: 44.0%
## 2023-02-09 17:35:58 (41790) [INFO]   Correction progress: 45.0%
## 2023-02-09 17:35:59 (41790) [INFO]   Correction progress: 46.0%
## 2023-02-09 17:35:59 (41790) [INFO]   Correction progress: 47.0%
## 2023-02-09 17:36:00 (41790) [INFO]   Correction progress: 48.0%
## 2023-02-09 17:36:01 (41790) [INFO]   Correction progress: 49.0%
## 2023-02-09 17:36:02 (41790) [INFO]   Correction progress: 50.0%
## 2023-02-09 17:36:02 (41790) [INFO]   Correction progress: 51.0%
## 2023-02-09 17:36:03 (41790) [INFO]   Correction progress: 52.0%
## 2023-02-09 17:36:04 (41790) [INFO]   Correction progress: 53.0%
## 2023-02-09 17:36:04 (41790) [INFO]   Correction progress: 54.0%
## 2023-02-09 17:36:05 (41790) [INFO]   Correction progress: 55.0%
## 2023-02-09 17:36:06 (41790) [INFO]   Correction progress: 56.0%
## 2023-02-09 17:36:07 (41790) [INFO]   Correction progress: 57.0%
## 2023-02-09 17:36:08 (41790) [INFO]   Correction progress: 58.0%
## 2023-02-09 17:36:08 (41790) [INFO]   Correction progress: 59.0%
## 2023-02-09 17:36:09 (41790) [INFO]   Correction progress: 60.0%
## 2023-02-09 17:36:10 (41790) [INFO]   Correction progress: 61.0%
## 2023-02-09 17:36:10 (41790) [INFO]   Correction progress: 62.0%
## 2023-02-09 17:36:11 (41790) [INFO]   Correction progress: 63.0%
## 2023-02-09 17:36:12 (41790) [INFO]   Correction progress: 64.0%
## 2023-02-09 17:36:12 (41790) [INFO]   Correction progress: 65.0%
```

```
## 2023-02-09 17:36:13 (41790) [INFO]   Correction progress: 66.0%
## 2023-02-09 17:36:14 (41790) [INFO]   Correction progress: 67.0%
## 2023-02-09 17:36:15 (41790) [INFO]   Correction progress: 68.0%
## 2023-02-09 17:36:15 (41790) [INFO]   Correction progress: 69.0%
## 2023-02-09 17:36:16 (41790) [INFO]   Correction progress: 70.0%
## 2023-02-09 17:36:17 (41790) [INFO]   Correction progress: 71.0%
## 2023-02-09 17:36:17 (41790) [INFO]   Correction progress: 72.0%
## 2023-02-09 17:36:18 (41790) [INFO]   Correction progress: 73.0%
## 2023-02-09 17:36:19 (41790) [INFO]   Correction progress: 74.0%
## 2023-02-09 17:36:20 (41790) [INFO]   Correction progress: 75.0%
## 2023-02-09 17:36:21 (41790) [INFO]   Correction progress: 76.0%
## 2023-02-09 17:36:22 (41790) [INFO]   Correction progress: 77.0%
## 2023-02-09 17:36:23 (41790) [INFO]   Correction progress: 78.0%
## 2023-02-09 17:36:24 (41790) [INFO]   Correction progress: 79.0%
## 2023-02-09 17:36:24 (41790) [INFO]   Correction progress: 80.0%
## 2023-02-09 17:36:25 (41790) [INFO]   Correction progress: 81.0%
## 2023-02-09 17:36:26 (41790) [INFO]   Correction progress: 82.0%
## 2023-02-09 17:36:27 (41790) [INFO]   Correction progress: 83.0%
## 2023-02-09 17:36:28 (41790) [INFO]   Correction progress: 84.0%
## 2023-02-09 17:36:29 (41790) [INFO]   Correction progress: 85.0%
## 2023-02-09 17:36:30 (41790) [INFO]   Correction progress: 86.0%
## 2023-02-09 17:36:31 (41790) [INFO]   Correction progress: 87.0%
## 2023-02-09 17:36:32 (41790) [INFO]   Correction progress: 88.0%
## 2023-02-09 17:36:34 (41790) [INFO]   Correction progress: 89.0%
## 2023-02-09 17:36:34 (41790) [INFO]   Correction progress: 90.0%
## 2023-02-09 17:36:36 (41790) [INFO]   Correction progress: 91.0%
## 2023-02-09 17:36:36 (41790) [INFO]   Correction progress: 92.0%
## 2023-02-09 17:36:38 (41790) [INFO]   Correction progress: 93.0%
## 2023-02-09 17:36:38 (41790) [INFO]   Correction progress: 94.0%
## 2023-02-09 17:36:40 (41790) [INFO]   Correction progress: 95.0%
## 2023-02-09 17:36:40 (41790) [INFO]   Correction progress: 96.0%
## 2023-02-09 17:36:42 (41790) [INFO]   Correction progress: 97.0%
## 2023-02-09 17:36:42 (41790) [INFO]   Correction progress: 98.0%
## 2023-02-09 17:36:42 (41790) [INFO]   Correction progress: 99.0%
## 2023-02-09 17:36:43 (41800) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
## 2023-02-09 17:36:43 (41800) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
## 2023-02-09 17:36:43 (41790) [INFO]   Correction progress: 100.0%
## 2023-02-09 17:36:43 (41790) [INFO]   Correction progress: done!
## 2023-02-09 17:36:44 (41800) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
## 2023-02-09 17:36:44 (41800) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
##
## 2023-02-09 17:36:45 (41790) [INFO]   Verifying bias correction
## 2023-02-09 17:36:45 (41790) [STATS]  BIAS    pre-bias variance forward:  0.0049439
## 2023-02-09 17:36:45 (41790) [STATS]  BIAS    post-bias variance forward: 0.0000628
## 2023-02-09 17:36:45 (41790) [STATS]  BIAS    pre-bias variance reverse:  0.0049210
## 2023-02-09 17:36:45 (41790) [STATS]  BIAS    post-bias variance reverse: 0.0000621
##
## 2023-02-09 17:36:45 (41790) [INFO]   Finished ATACorrect run (total time elapsed: 0:01:20.924270)
## # TOBIAS 0.15.1 ATACorrect (run started 2023-02-09 17:36:46.995684)
## # Working directory: /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3
## # Command line call: TOBIAS ATACorrect --bam C1_gDNA_rep1_chr21.bam --genome hg38_chr21.fa --peaks ESR1_REST_c
##
## # ----- Input parameters -----
## # bam:   C1_gDNA_rep1_chr21.bam
## # genome:    hg38_chr21.fa
## # peaks: ESR1_REST_chr21.bed
## # regions_in:    None
## # regions_out:   None
## # blacklist: None
## # extend:    100
## # split_strands: False
## # norm_off:  True
```

```
## # track_off: []
## # drop_chroms:   [’chrM’, ’chrMT’, ’M’, ’MT’, ’Mito’]
## # k_flank:   12
## # read_shift:    [4, -5]
## # bg_shift:  100
## # window:    100
## # score_mat: DWM
## # bias_pkl:  None
## # prefix:    C1_gDNA_rep1_chr21
## # outdir:    /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_R
## # cores: 3
## # split: 100
## # verbosity: 3
##
##
## # ----- Output files -----
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_REST_output/
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_REST_output/
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_REST_output/
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_REST_output/
## # /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vignette/output_mk3/ATACorrect_ESR1_REST_output/
##
##
## 2023-02-09 17:36:46 (41805) [INFO]   ----- Processing input data -----
## 2023-02-09 17:36:46 (41805) [INFO]   Reading info from .bam file
## 2023-02-09 17:36:46 (41805) [INFO]   Reading info from .fasta file
## 2023-02-09 17:36:46 (41805) [WARNING]    No additional chromosomes were removed. Consider using ’--drop-chroms
## 2023-02-09 17:36:46 (41805) [INFO]   Processing input/output regions
## 2023-02-09 17:36:48 (41805) [STATS]  genome: 1 regions | 46709983 bp | 100.00% coverage
## 2023-02-09 17:36:48 (41805) [STATS]  input_regions: 20445 regions | 46689713 bp | 99.96% coverage
## 2023-02-09 17:36:48 (41805) [STATS]  output_regions: 12710 regions | 3513628 bp | 7.52% coverage
## 2023-02-09 17:36:48 (41805) [STATS]  peak_regions: 20270 regions | 20270 bp | 0.04% coverage
## 2023-02-09 17:36:48 (41805) [STATS]  nonpeak_regions: 20445 regions | 46689713 bp | 99.96% coverage
## 2023-02-09 17:36:48 (41805) [STATS]  blacklist_regions: 0 regions | 0 bp | 0.00% coverage
##
## 2023-02-09 17:36:48 (41805) [INFO]   ----- Estimating normalization factors -----
## 2023-02-09 17:36:48 (41805) [INFO]   Normalization was switched off
## 2023-02-09 17:36:48 (41805) [STATS]  CORRECTION_FACTOR:  1.00000
##
## 2023-02-09 17:36:48 (41805) [INFO]   Started estimation of sequence bias...
## 2023-02-09 17:36:48 (41805) [INFO]   Progress: 0%
## 2023-02-09 17:36:48 (41805) [INFO]   Progress: 7%
## 2023-02-09 17:36:49 (41805) [INFO]   Progress: 11%
## 2023-02-09 17:36:49 (41805) [INFO]   Progress: 18%
## 2023-02-09 17:36:50 (41805) [INFO]   Progress: 28%
## 2023-02-09 17:36:50 (41805) [INFO]   Progress: 40%
## 2023-02-09 17:36:51 (41805) [INFO]   Progress: 49%
## 2023-02-09 17:36:51 (41805) [INFO]   Progress: 61%
## 2023-02-09 17:36:52 (41805) [INFO]   Progress: 72%
## 2023-02-09 17:36:52 (41805) [INFO]   Progress: 81%
## 2023-02-09 17:36:53 (41805) [INFO]   Progress: 90%
## 2023-02-09 17:36:53 (41805) [INFO]   Progress: 99%
## 2023-02-09 17:36:54 (41805) [INFO]   Progress: 100%
## 2023-02-09 17:36:54 (41805) [INFO]   Finalizing bias motif for scoring
##
## 2023-02-09 17:36:54 (41805) [INFO]   ----- Correcting reads from .bam within output regions -----
## 2023-02-09 17:36:54 (41805) [INFO]   Correction progress: 0%
## 2023-02-09 17:36:57 (41805) [INFO]   Correction progress: 1.0%
## 2023-02-09 17:36:59 (41805) [INFO]   Correction progress: 2.0%
## 2023-02-09 17:37:01 (41805) [INFO]   Correction progress: 3.0%
## 2023-02-09 17:37:02 (41805) [INFO]   Correction progress: 4.0%
## 2023-02-09 17:37:04 (41805) [INFO]   Correction progress: 5.0%
```

```
## 2023-02-09 17:37:05 (41805) [INFO]   Correction progress: 6.0%
## 2023-02-09 17:37:07 (41805) [INFO]   Correction progress: 7.0%
## 2023-02-09 17:37:09 (41805) [INFO]   Correction progress: 8.0%
## 2023-02-09 17:37:11 (41805) [INFO]   Correction progress: 9.0%
## 2023-02-09 17:37:12 (41805) [INFO]   Correction progress: 10.0%
## 2023-02-09 17:37:14 (41805) [INFO]   Correction progress: 11.0%
## 2023-02-09 17:37:14 (41805) [INFO]   Correction progress: 12.0%
## 2023-02-09 17:37:16 (41805) [INFO]   Correction progress: 13.0%
## 2023-02-09 17:37:17 (41805) [INFO]   Correction progress: 14.0%
## 2023-02-09 17:37:18 (41805) [INFO]   Correction progress: 15.0%
## 2023-02-09 17:37:19 (41805) [INFO]   Correction progress: 16.0%
## 2023-02-09 17:37:20 (41805) [INFO]   Correction progress: 17.0%
## 2023-02-09 17:37:21 (41805) [INFO]   Correction progress: 18.0%
## 2023-02-09 17:37:23 (41805) [INFO]   Correction progress: 19.0%
## 2023-02-09 17:37:23 (41805) [INFO]   Correction progress: 20.0%
## 2023-02-09 17:37:25 (41805) [INFO]   Correction progress: 21.0%
## 2023-02-09 17:37:25 (41805) [INFO]   Correction progress: 22.0%
## 2023-02-09 17:37:26 (41805) [INFO]   Correction progress: 23.0%
## 2023-02-09 17:37:27 (41805) [INFO]   Correction progress: 24.0%
## 2023-02-09 17:37:28 (41805) [INFO]   Correction progress: 25.0%
## 2023-02-09 17:37:29 (41805) [INFO]   Correction progress: 26.0%
## 2023-02-09 17:37:30 (41805) [INFO]   Correction progress: 27.0%
## 2023-02-09 17:37:31 (41805) [INFO]   Correction progress: 28.0%
## 2023-02-09 17:37:32 (41805) [INFO]   Correction progress: 29.0%
## 2023-02-09 17:37:33 (41805) [INFO]   Correction progress: 30.0%
## 2023-02-09 17:37:34 (41805) [INFO]   Correction progress: 31.0%
## 2023-02-09 17:37:35 (41805) [INFO]   Correction progress: 32.0%
## 2023-02-09 17:37:37 (41805) [INFO]   Correction progress: 33.0%
## 2023-02-09 17:37:37 (41805) [INFO]   Correction progress: 34.0%
## 2023-02-09 17:37:39 (41805) [INFO]   Correction progress: 35.0%
## 2023-02-09 17:37:39 (41805) [INFO]   Correction progress: 36.0%
## 2023-02-09 17:37:41 (41805) [INFO]   Correction progress: 37.0%
## 2023-02-09 17:37:42 (41805) [INFO]   Correction progress: 38.0%
## 2023-02-09 17:37:43 (41805) [INFO]   Correction progress: 39.0%
## 2023-02-09 17:37:44 (41805) [INFO]   Correction progress: 40.0%
## 2023-02-09 17:37:46 (41805) [INFO]   Correction progress: 41.0%
## 2023-02-09 17:37:47 (41805) [INFO]   Correction progress: 42.0%
## 2023-02-09 17:37:49 (41805) [INFO]   Correction progress: 43.0%
## 2023-02-09 17:37:50 (41805) [INFO]   Correction progress: 44.0%
## 2023-02-09 17:37:52 (41805) [INFO]   Correction progress: 45.0%
## 2023-02-09 17:37:52 (41805) [INFO]   Correction progress: 46.0%
## 2023-02-09 17:37:55 (41805) [INFO]   Correction progress: 47.0%
## 2023-02-09 17:37:55 (41805) [INFO]   Correction progress: 48.0%
## 2023-02-09 17:37:58 (41805) [INFO]   Correction progress: 49.0%
## 2023-02-09 17:37:58 (41805) [INFO]   Correction progress: 50.0%
## 2023-02-09 17:38:01 (41805) [INFO]   Correction progress: 51.0%
## 2023-02-09 17:38:01 (41805) [INFO]   Correction progress: 52.0%
## 2023-02-09 17:38:04 (41805) [INFO]   Correction progress: 53.0%
## 2023-02-09 17:38:04 (41805) [INFO]   Correction progress: 54.0%
## 2023-02-09 17:38:07 (41805) [INFO]   Correction progress: 55.0%
## 2023-02-09 17:38:07 (41805) [INFO]   Correction progress: 56.0%
## 2023-02-09 17:38:10 (41805) [INFO]   Correction progress: 58.0%
## 2023-02-09 17:38:12 (41805) [INFO]   Correction progress: 59.0%
## 2023-02-09 17:38:13 (41805) [INFO]   Correction progress: 60.0%
## 2023-02-09 17:38:15 (41805) [INFO]   Correction progress: 61.0%
## 2023-02-09 17:38:16 (41805) [INFO]   Correction progress: 62.0%
## 2023-02-09 17:38:18 (41805) [INFO]   Correction progress: 63.0%
## 2023-02-09 17:38:19 (41805) [INFO]   Correction progress: 64.0%
## 2023-02-09 17:38:21 (41805) [INFO]   Correction progress: 65.0%
## 2023-02-09 17:38:22 (41805) [INFO]   Correction progress: 66.0%
## 2023-02-09 17:38:24 (41805) [INFO]   Correction progress: 67.0%
## 2023-02-09 17:38:25 (41805) [INFO]   Correction progress: 68.0%
```

```
## 2023-02-09 17:38:26 (41805) [INFO]   Correction progress: 69.0%
## 2023-02-09 17:38:28 (41805) [INFO]   Correction progress: 70.0%
## 2023-02-09 17:38:29 (41805) [INFO]   Correction progress: 71.0%
## 2023-02-09 17:38:30 (41805) [INFO]   Correction progress: 72.0%
## 2023-02-09 17:38:32 (41805) [INFO]   Correction progress: 73.0%
## 2023-02-09 17:38:34 (41805) [INFO]   Correction progress: 74.0%
## 2023-02-09 17:38:36 (41805) [INFO]   Correction progress: 75.0%
## 2023-02-09 17:38:38 (41805) [INFO]   Correction progress: 76.0%
## 2023-02-09 17:38:39 (41805) [INFO]   Correction progress: 77.0%
## 2023-02-09 17:38:42 (41805) [INFO]   Correction progress: 78.0%
## 2023-02-09 17:38:43 (41805) [INFO]   Correction progress: 79.0%
## 2023-02-09 17:38:46 (41805) [INFO]   Correction progress: 80.0%
## 2023-02-09 17:38:47 (41805) [INFO]   Correction progress: 81.0%
## 2023-02-09 17:38:50 (41805) [INFO]   Correction progress: 82.0%
## 2023-02-09 17:38:52 (41805) [INFO]   Correction progress: 83.0%
## 2023-02-09 17:38:54 (41805) [INFO]   Correction progress: 84.0%
## 2023-02-09 17:38:56 (41805) [INFO]   Correction progress: 85.0%
## 2023-02-09 17:38:58 (41805) [INFO]   Correction progress: 86.0%
## 2023-02-09 17:39:01 (41805) [INFO]   Correction progress: 87.0%
## 2023-02-09 17:39:03 (41805) [INFO]   Correction progress: 88.0%
## 2023-02-09 17:39:06 (41805) [INFO]   Correction progress: 89.0%
## 2023-02-09 17:39:08 (41805) [INFO]   Correction progress: 90.0%
## 2023-02-09 17:39:10 (41805) [INFO]   Correction progress: 91.0%
## 2023-02-09 17:39:12 (41805) [INFO]   Correction progress: 92.0%
## 2023-02-09 17:39:15 (41805) [INFO]   Correction progress: 93.0%
## 2023-02-09 17:39:18 (41805) [INFO]   Correction progress: 94.0%
## 2023-02-09 17:39:21 (41805) [INFO]   Correction progress: 95.0%
## 2023-02-09 17:39:23 (41805) [INFO]   Correction progress: 96.0%
## 2023-02-09 17:39:26 (41805) [INFO]   Correction progress: 97.0%
## 2023-02-09 17:39:27 (41805) [INFO]   Correction progress: 98.0%
## 2023-02-09 17:39:28 (41805) [INFO]   Correction progress: 99.0%
## 2023-02-09 17:39:29 (41815) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
## 2023-02-09 17:39:29 (41815) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
## 2023-02-09 17:39:29 (41805) [INFO]   Correction progress: 100.0%
## 2023-02-09 17:39:29 (41805) [INFO]   Correction progress: done!
## 2023-02-09 17:39:31 (41815) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
## 2023-02-09 17:39:31 (41815) [INFO]   Closing /Users/guertinlab/Documents/tn5_Bias/230203_ATACorrect_sOAB_vigne
##
## 2023-02-09 17:39:32 (41805) [INFO]   Verifying bias correction
## 2023-02-09 17:39:32 (41805) [STATS]  BIAS    pre-bias variance forward:  0.0050726
## 2023-02-09 17:39:32 (41805) [STATS]  BIAS    post-bias variance forward: 0.0000494
## 2023-02-09 17:39:32 (41805) [STATS]  BIAS    pre-bias variance reverse:  0.0050998
## 2023-02-09 17:39:32 (41805) [STATS]  BIAS    post-bias variance reverse: 0.0000491
##
## 2023-02-09 17:39:32 (41805) [INFO]   Finished ATACorrect run (total time elapsed: 0:02:45.567174)
```

4 Output analysis

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

4.1 Read depth

First, we compare the read depth output of both methods using a bar chart.

```
sOAB_read_depth = fread('C1_gDNA_rep1_chr21_RE_scaled.bedGraph')
sOAB_read_depth = sum(sOAB_read_depth$V4)

sOAB_unscaled_read_depth = fread('C1_gDNA_rep1_chr21_not_scaled.bed')
sOAB_unscaled_read_depth = sum(sOAB_unscaled_read_depth$V5)
```

```
ATACorrect_ESR1_read_depth =
  fread('ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_corrected.bedgraph')
ATACorrect_ESR1_read_depth =
  sum(ATACorrect_ESR1_read_depth$V4)

ATACorrect_uncorrected_ESR1_read_depth =
  fread('ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_uncorrected.bedgraph')
ATACorrect_uncorrected_ESR1_read_depth =
  sum(ATACorrect_uncorrected_ESR1_read_depth$V4)

ATACorrect_ESR1_REST_read_depth =
  fread('ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_corrected.bedgraph')
ATACorrect_ESR1_REST_read_depth =
  sum(ATACorrect_ESR1_REST_read_depth$V4)

ATACorrect_uncorrected_ESR1_REST_read_depth =
  fread('ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_uncorrected.bedgraph')
ATACorrect_uncorrected_ESR1_REST_read_depth =
  sum(ATACorrect_uncorrected_ESR1_REST_read_depth$V4)

barchart_comparison = c(sOAB_read_depth, sOAB_unscaled_read_depth,
                        ATACorrect_ESR1_read_depth, ATACorrect_uncorrected_ESR1_read_depth,
                        ATACorrect_ESR1_REST_read_depth, ATACorrect_uncorrected_ESR1_REST_read_depth)
names(barchart_comparison) = c('seqOutATACBias', 'seqOutBias unscaled',
                               'ATACorrect ESR1', 'ATACorrect ESR1 uncorrected',
                               'ATACorrect ESR1/REST', 'ATACorrect ESR1/REST uncorrected')

pdf(file = 'sOAB_ATACorrect_RD_comparison.pdf')
par(mar=c(16, 5, 3, 1))
barplot_comparison= barplot(barchart_comparison,
                            ylim = c(min(barchart_comparison),max(barchart_comparison)+40000), las = 2)
text(x = barplot_comparison, y = barchart_comparison + 20000,
     labels = as.integer(unname(barchart_comparison)))
title(ylab = 'Read depth', line = 4)
dev.off()
```

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

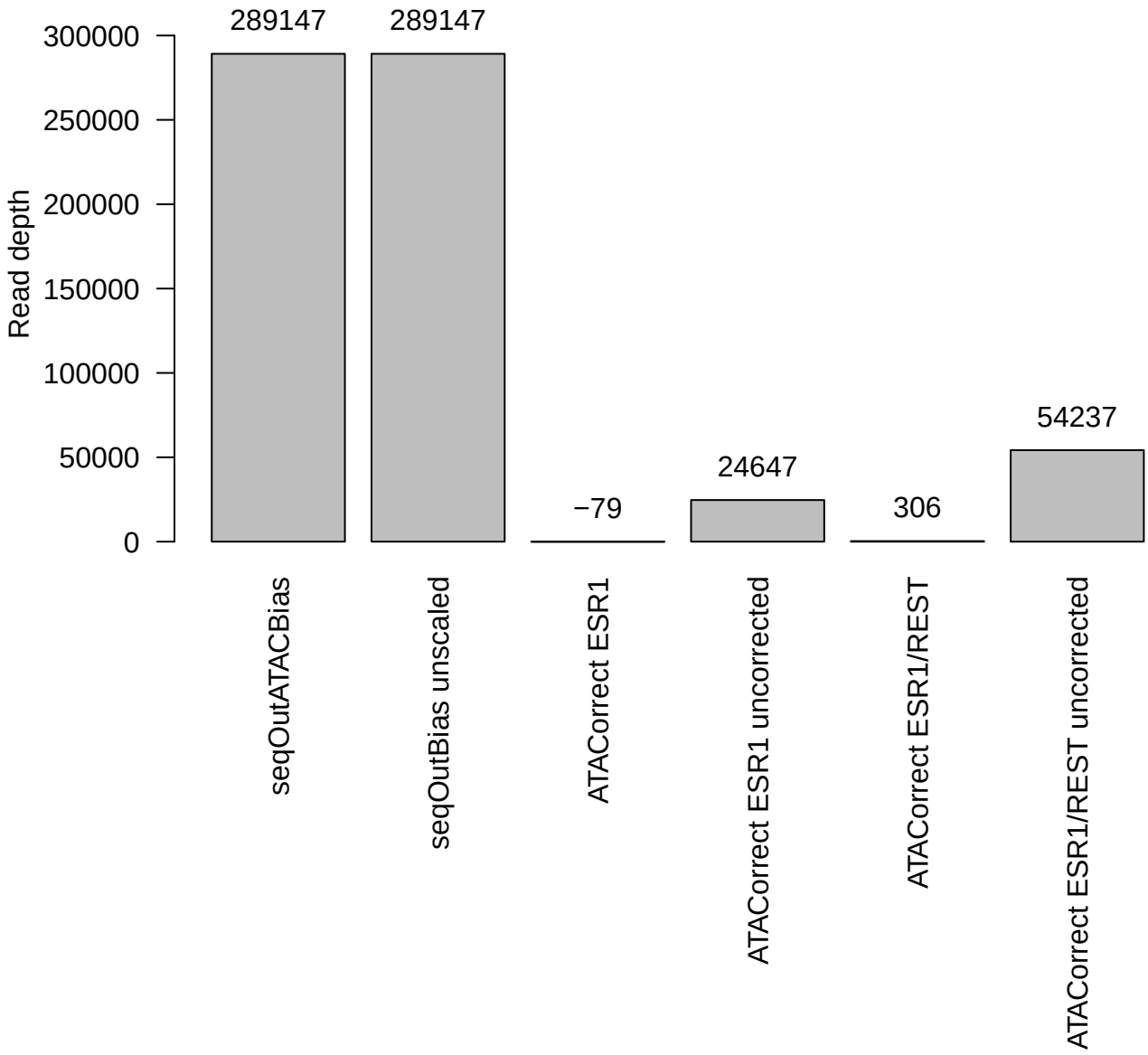

Figure 1: Bias correction method read depth comparison after analysis

As seen in Figure 1, read depth between seqOutATACBias and ATACorrect is very different. Where seqOutATACBias maintains the same read depth between scaled and unscaled output, ATACorrect read depth changes based on the number of peak regions in the input bed file. In many tested scenarios, we find that ATACorrect corrected output has a negative read depth, which is a preclusion to effective bias correction comparison with seqOutATACBias. An important feature of seqOutATACBias is that output reads have the same depth (signal) as the input, meaning that no signal is lost by correction. This is a central aspect of the seqOutATACBias correction method: reads are scaled based on Tn5 preference for or against the sequence environment of the read. Therefore, theoretically perfect bias correction could be determined by the number of reads divided by mappable positions in the genome. To better understand the output from these two methods, we next look at number of reads in both outputs.

4.2 Read number

To calculate the number of reads, or Tn5 insertion locations, in each data set we simply count the number of rows for each output file.

```
sOAB_read_number =
  fread('C1_gDNA_rep1_chr21_RE_scaled.bedGraph')
sOAB_read_number =
  nrow(sOAB_read_number)

sOAB_unscaled_read_number =
  fread('C1_gDNA_rep1_chr21_not_scaled.bed')
sOAB_unscaled_read_number =
  nrow(sOAB_unscaled_read_number)

ATACorrect_ESR1_read_number =
  fread('ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_corrected.bedgraph')
ATACorrect_ESR1_read_number =
  nrow(ATACorrect_ESR1_read_number)

ATACorrect_ESR1_uncorrected_read_number =
  fread('ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_uncorrected.bedgraph')
ATACorrect_ESR1_uncorrected_read_number =
  nrow(ATACorrect_ESR1_uncorrected_read_number)

ATACorrect_ESR1_REST_read_number =
  fread('ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_corrected.bedgraph')
ATACorrect_ESR1_REST_read_number =
  nrow(ATACorrect_ESR1_REST_read_number)

ATACorrect_ESR1_REST_uncorrected_read_number =
  fread('ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_uncorrected.bedgraph')
ATACorrect_ESR1_REST_uncorrected_read_number =
  nrow(ATACorrect_ESR1_REST_uncorrected_read_number)

barchart_comparison = c(sOAB_read_number, sOAB_unscaled_read_number,
                        ATACorrect_ESR1_read_number, ATACorrect_ESR1_uncorrected_read_number,
                        ATACorrect_ESR1_REST_read_number, ATACorrect_ESR1_REST_uncorrected_read_number)
names(barchart_comparison) = c('seqOutATACBias', 'seqOutBias unscaled',
                               'ATACorrect ESR1', 'ATACorrect ESR1 uncorrected',
                               'ATACorrect ESR1/REST', 'ATACorrect ESR1/REST uncorrected')

pdf(file = 'sOAB_ATACorrect_ReadNumber_comparison.pdf')
par(mar=c(16, 5, 3, 1))
barplot_comparison= barplot(barchart_comparison, ylim = c(0,max(barchart_comparison)+55000), las = 2)
text(x = barplot_comparison, y = barchart_comparison + 35000, labels = as.integer(unname(barchart_comparison)))
title(ylab = 'Number of reads', line = 4)
dev.off()
```

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

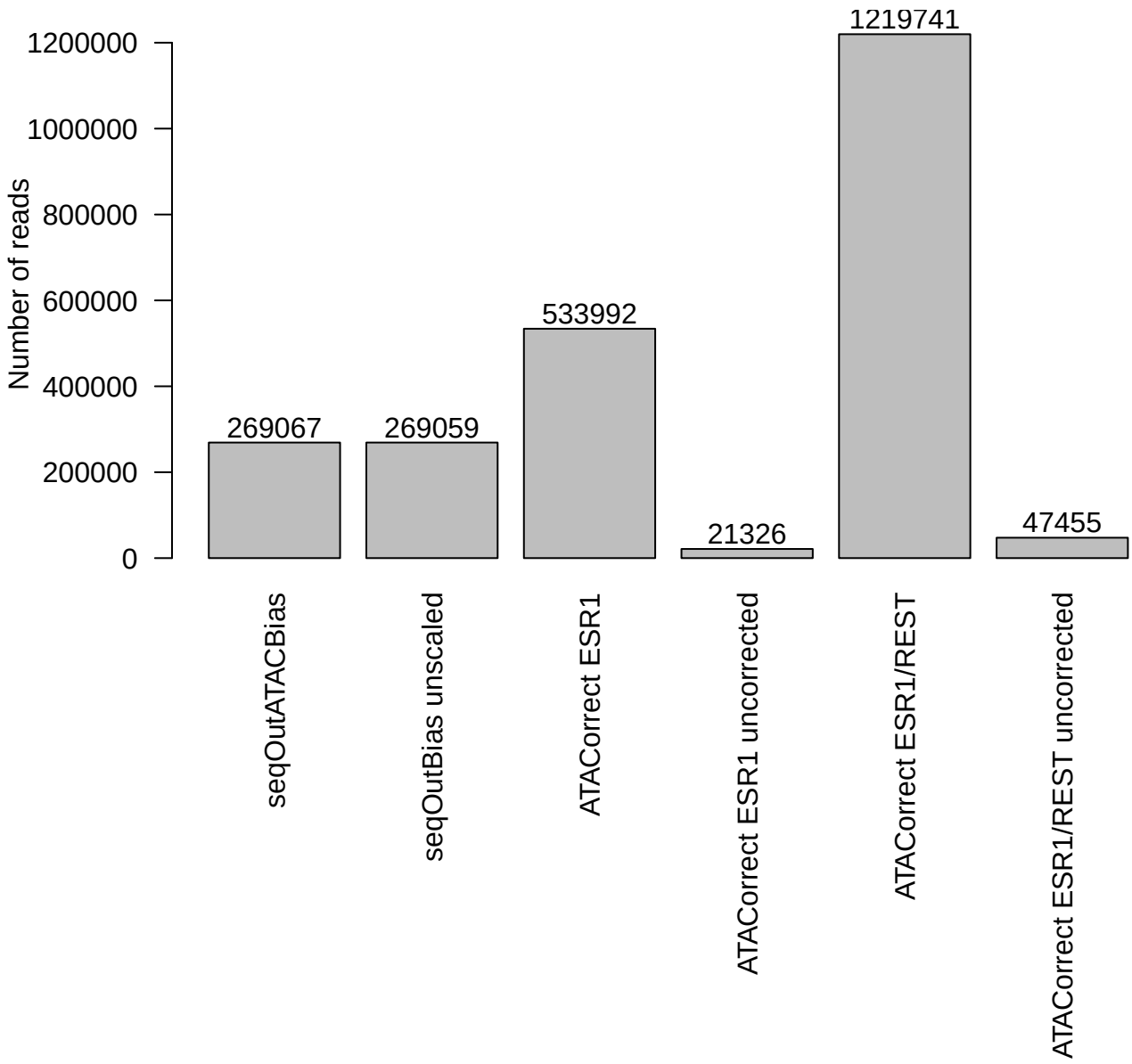

Figure 2: Comparison between number of reads in seqOutATACBias and ATACorrect output

Figure 2 clearly shows that seqOutATACBias scaled output has many fewer reads than corrected ATACorrect output, and many more reads than uncorrected ATACorrect output. Based on the publication accompanying TOBIAS (https://doi.org/10.1038/s41467-020-18035-1), we postulate that these additional reads in the ATACorrect corrected output are simulated data based on the modeled Tn5 bias within the peaks input region set. Additionally, we see in both Figure 1 and Figure 2 that ATACorrect read depth and read number are dependent on the number of peak motifs input. To understand how these sets of reads are different, we next explore their genomic coordinates.

4.3 Interval set coordinate comparison

Because we know that the read count and depth are very different between the two methods, we aim to compare the coordinates of the reads from each method. We compare these populations using an Euler diagram, to show which reads are the same in both methods and which are different.

```
sOAB_unscaled_reads = fread('C1_gDNA_rep1_chr21_not_scaled.bed')

ATACorrect_ESR1_uncorrected_reads = fread('ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_uncorrected.bedgraph')
```

```
ATACorrect_ESR1_REST_uncorrected_reads = fread('ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_uncorrected.bedgra

x = list(sOAB_unscaled_reads$V2,
         ATACorrect_ESR1_uncorrected_reads$V2,
         ATACorrect_ESR1_REST_uncorrected_reads$V2)
names(x) = c('seqOutATACBias', 'ATACorrect ESR1', 'ATACorrect ESR1/REST')

pdf(file = 'seqOutATACBias_ATACorrect_Uncorrected_overlap.pdf', width = 12,
    height = 12)
plot(euler(x, shape = "ellipse"), quantities = list(cex = 0.75, lineheight=10), legend = TRUE,
     main = 'Unscaled read coordinate comparison')
dev.off()
```

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

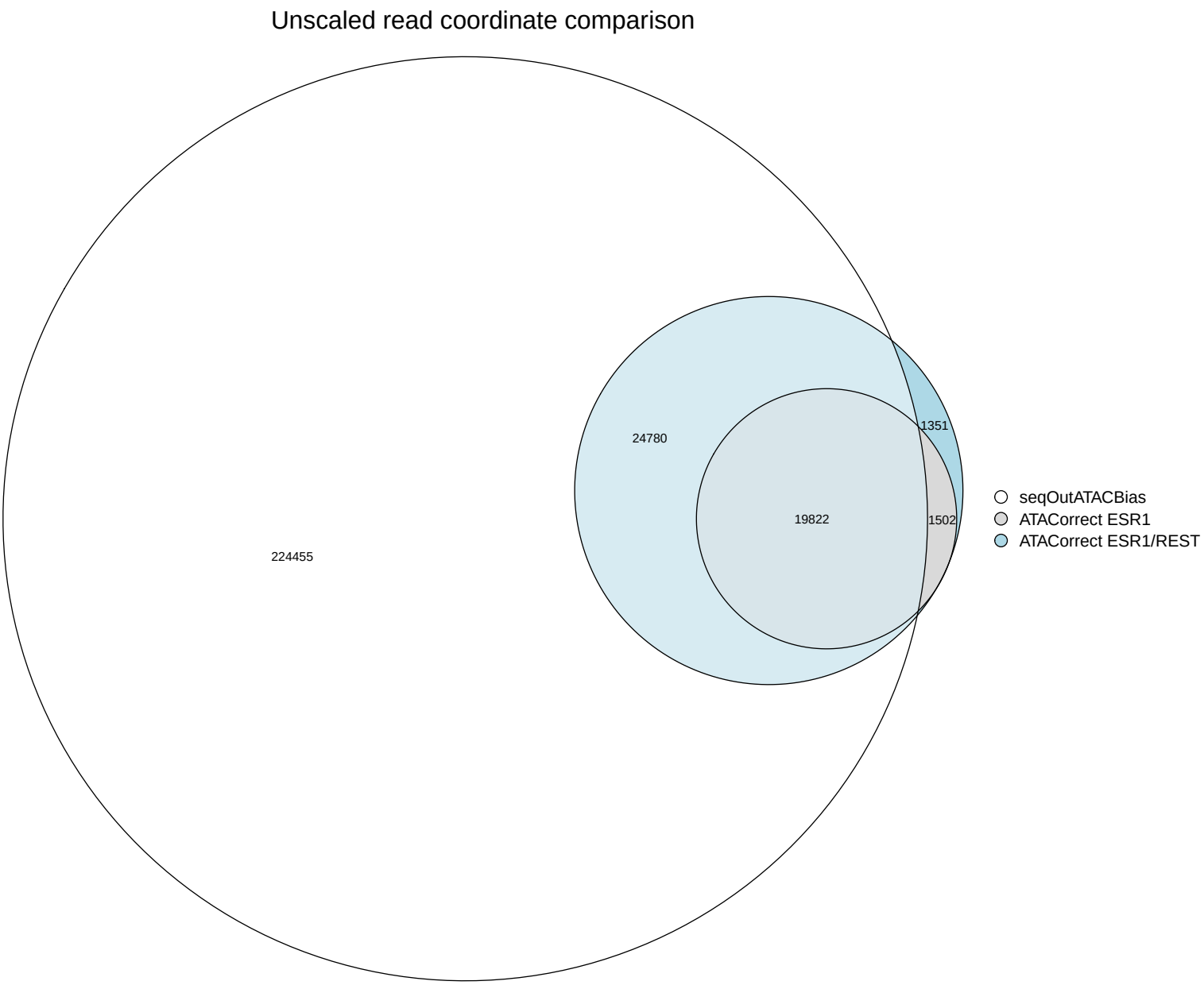

Figure 3: Read identity overlap between seqOutATACBias unscaled and ATACorrect uncorrected output

In Figure 3 we observe that most uncorrected ATACorrect reads are found in seqOutATACBias unscaled reads. The small number of ATACorrect uncorrected reads which are not present in seqOutATACBias unscaled output are likely aligned to regions determined to be unmappable by seqOutBias. This could be determined by using ATACorrect's `blacklist` input, but was not included for brevity. Additionally, the population of ATACorrect uncorrected reads is much smaller than the population found in seqOutATACBias, and the size of this population is based on the number of input `peak` motifs. Next, we visualize how the seqOutATACBias population fits into the larger corrected ATACorrect population.

```
sOAB_reads = fread('C1_gDNA_rep1_chr21_RE_scaled.bedGraph')

ATACorrect_ESR1_corrected_reads = fread('ATACorrect_ESR1_output/C1_gDNA_rep1_chr21_corrected.bedgraph')
```

```
ATACorrect_ESR1_REST_corrected_reads = fread('ATACorrect_ESR1_REST_output/C1_gDNA_rep1_chr21_corrected.bedgraph')

x = list(sOAB_unscaled_reads$V2,
         ATACorrect_ESR1_corrected_reads$V2,
         ATACorrect_ESR1_REST_corrected_reads$V2)
names(x) = c('seqOutATACBias', 'ATACorrect ESR1', 'ATACorrect ESR1/REST')

pdf(file = 'seqOutATACBias_ATACorrect_Corrected_overlap.pdf', width = 12,
    height = 10)
plot(euler(x, shape = "ellipse"), quantities = list(cex = 0.75, lineheight=10), legend = TRUE,
     main = 'Scaled read coordinate comparison')
dev.off()
```

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

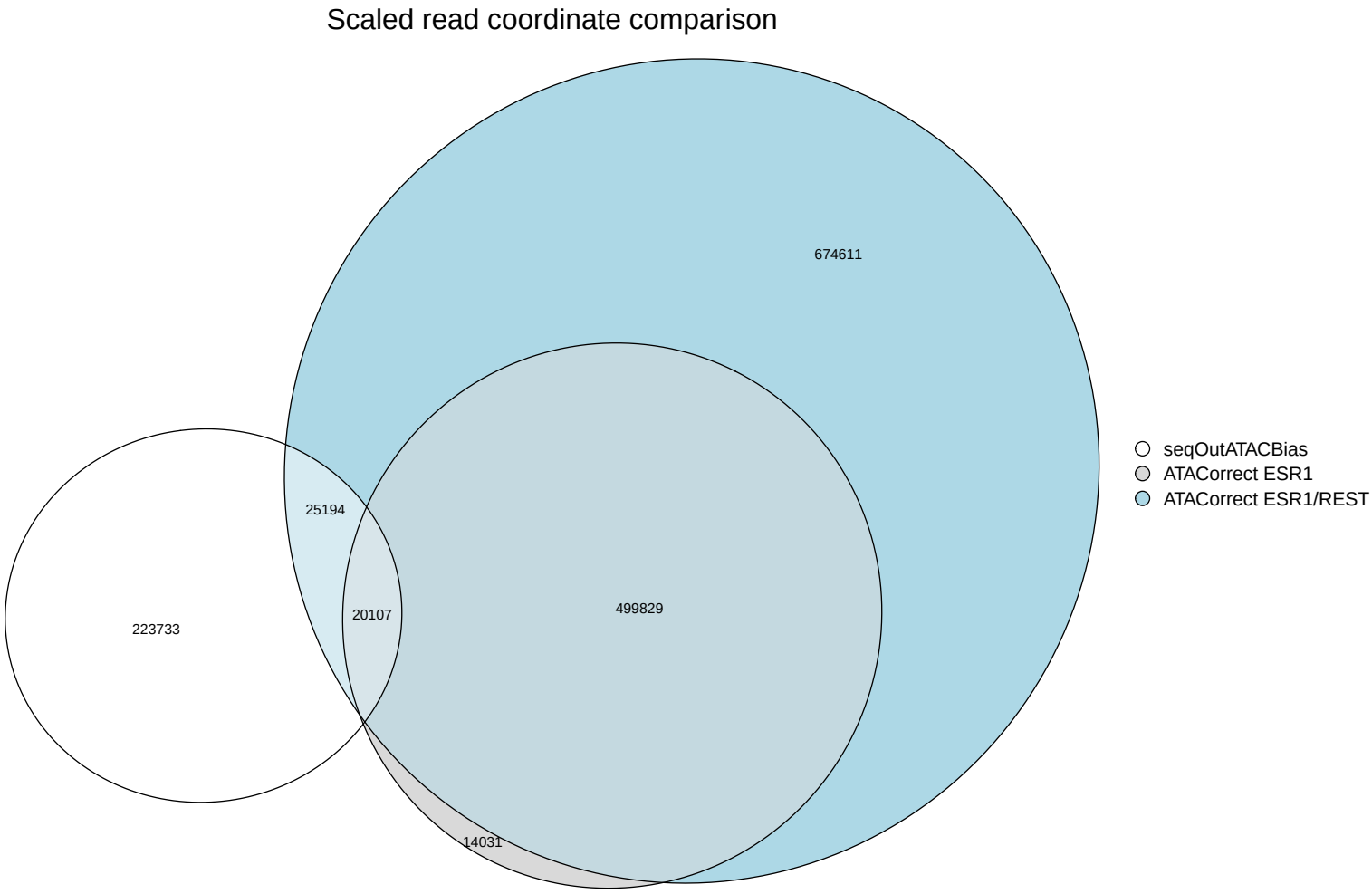

Figure 4: Genomic coordinate overlap between seqOutATACBias and ATACorrect scaled output

Figure 4 shows that bias corrected seqOutATACBias and ATACorrect output share a small fraction of the same reads, given identical input. This means that a comparison of bias correction between the two methods would rely on functionally examining the difference of bias between two separate populations of reads, with as little as 4% overlap in some tested conditions.

4.4 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%20Comparison

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

```
## 2023-02-09 17:39:35 URL:https://raw.githubusercontent.com/guertinlab/Tn5bias/master/Manuscript_Vignette/figs/s
```

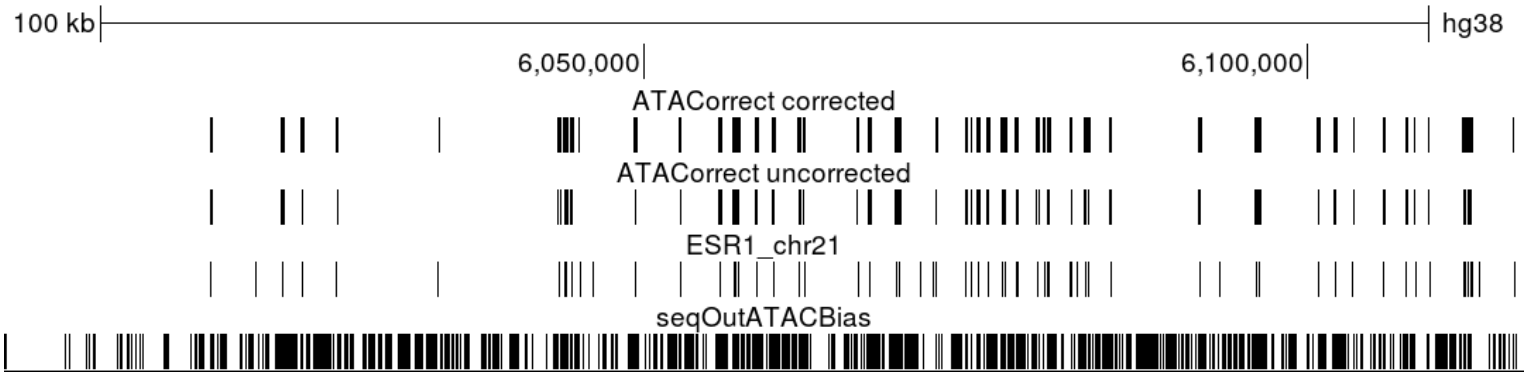

Figure 5: Browser snapshot of seqOutATACBias output, uncorrected ATACorrect output, corrected ATACorrect output, and the 'peaks' file (ESR1_chr21) used as input for the two ATACorrect output files

The browser snap shot seen in Figure 5 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. In our final analysis, we show that simulated ATACorrect reads are confined to the regions nearby the input `peaks` file.

4.5 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. Recall that 93% of uncorrected reads are also found in seqOutATACBias output as well.

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

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 > ATACorrec
ls -l ATACorrect_175bp_simulated_peaks.bed
```

```
## -rw-r--r--  1 guertinlab  staff  0 Feb  9 17:39 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.

In summary, this analysis shows several factors which inhibit direct comparison between ATACorrect and seqOutATACBias. The first is that seqOutATACBias output is scaled to the read depth of the input, while ATACorrect frequently includes negative read depth or very low read depth. Second, 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. Third, 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, without changing read depth, adding new reads, or removing other reads, and which gives this output in a single file for implementation in subsequent, nonrestricted analysis.