

1 Umap and Bimap: quantifying genome and methylome mappability

2 Mehran Karimzadeh^{1,4}, Carl Ernst², Anshul Kundaje³, and Michael M. Hoffman^{1,4,5}

3 ¹Princess Margaret Cancer Centre, Toronto, ON, Canada

4 ²Department of Human Genetics, McGill University, Montreal, QC, Canada

5 ³Department of Computer Science, Stanford University, Stanford, CA, USA

6 ⁴Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

7 ⁵Department of Computer Science, University of Toronto, Toronto, ON, Canada

8 December 19, 2016

9 **Abstract**

10 **Motivation:**

11 Short-read sequencing enables assessment of genetic and biochemical traits of individual genomic
12 regions, such as the location of genetic variation, protein binding, and chemical modifications. Ev-
13 ery region in a genome assembly has a property called *mappability* which measures the extent to
14 which it can be uniquely mapped by sequence reads. In regions of lower mappability, estimates of
15 genomic and epigenomic characteristics from sequencing assays are less reliable. At best, sequenc-
16 ing assays will produce misleadingly low numbers of reads in these regions. At worst, these regions
17 have increased susceptibility to spurious mapping from reads from other regions of the genome
18 with sequencing errors or unexpected genetic variation. Bisulfite sequencing approaches used to
19 identify DNA methylation exacerbate these problems by introducing large numbers of reads that
20 map to multiple regions. While many tools consider mappability during the read mapping process,
21 subsequent analysis often loses this information. Both to correct assumptions of uniformity in down-
22 stream analysis, and to identify regions where the analysis is less reliable, it is necessary to know
23 the mappability of both ordinary and bisulfite-converted genomes.

24 **Results:**

25 We introduce the Umap software for efficiently identifying uniquely mappable regions of any genome.
26 Its Bimap extension identifies mappability of the bisulfite-converted genome. With a read length of
27 24 bp, 15.5% of the unmodified genome and 30% of the bisulfite-converted genome is not uniquely
28 mappable. This complicates interpretation of functional genomics experiments using short-read se-
29 quencing, especially in regulatory regions. For example, 42% of human CpG islands overlap with
30 regions that are not uniquely mappable. Similarly, in some ENCODE ChIP-seq datasets, up to
31 30% of peaks overlap with regions that are not uniquely mappable. We also explored differentially
32 methylated regions from a case-control study and identified regions that were not uniquely map-
33 pable. In the widely used 450K methylation array, 962 probes are not uniquely mappable. Genome
34 mappability is higher with longer sequencing reads, but most publicly available ChIP-seq and re-
35 duced representation bisulfite sequencing datasets have shorter reads. Therefore, uneven and low
36 mappability remains a concern in a majority of existing data.

37 **Availability:**

38 A Umap and Bimap track hub for human genome assemblies GRCh37/hg19 and GRCh38/hg38,
39 and mouse assemblies GRCm37/mm9 and GRCm38/mm10 is available at
40 <http://bimap.hoffmanlab.org> for use with the UCSC and Ensembl genome browsers. We have de-
41 posited in [Zenodo](https://zenodo.org/record/60940) the current version of our software (<http://doi.org/10.5281/zenodo.60940>)
42 and the mappability data used in this project (<http://doi.org/10.5281/zenodo.60943>). In addi-
43 tion, the software (<https://bitbucket.org/hoffmanlab/umap>) is freely available under the GNU
44 General Public License, version 3 (GPLv3).

45 **Contact:**

46 michael.hoffman@utoronto.ca

47 **1 Introduction**

48 High-throughput sequencing enables low-cost collection of high numbers of sequencing reads but
49 these reads are often short. Short-read sequencing limits the fraction of the genome that we can

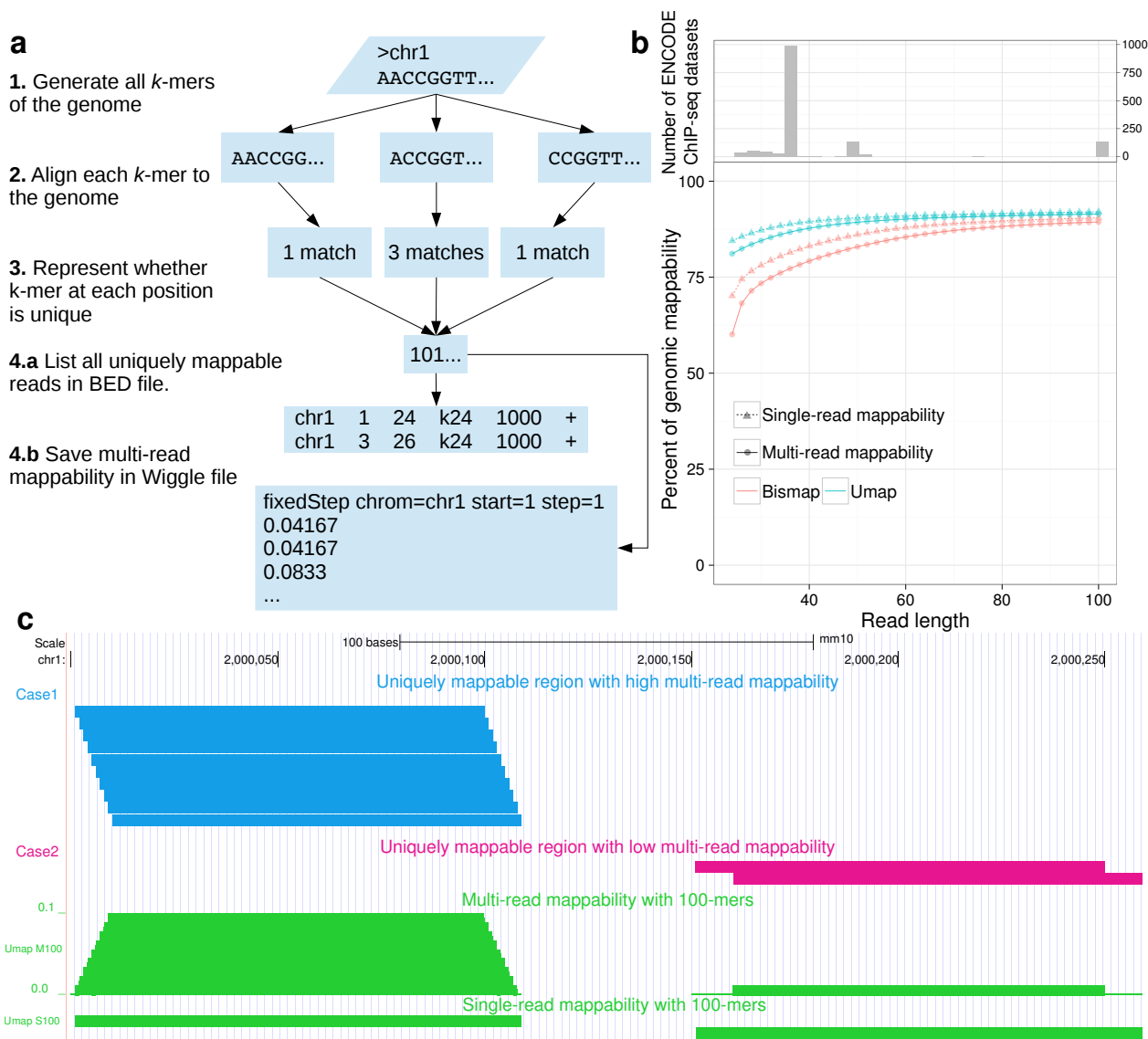


Figure 1: **Mappability of the genome by Umap.** (a) The Umap workflow identifies all unique k -mers of a genome given a read length of k . (b) Mappability of the human genome and methylome for read lengths between 24 and 100. (c) All of the uniquely mappable reads in two regions with high and low multi-read mappability is shown. In *Case 1* (blue), all possible reads covering the region are uniquely mappable. In *Case 2* (magenta), only two reads out of 10 are uniquely mappable.

50 unambiguously sequence by aligning the reads to the reference genome (Figure 1b). Still, we can
 51 identify much of the regulatory regions of the genome such as transcription factor binding sites,
 52 histone modifications and other important regulatory regions. However, reads that are ambiguously
 53 mapped produce a false positive signal that misleads analysis. Some regions of the genome with
 54 low complexity including repeat elements are not uniquely mappable at a given read length. Other
 55 regions overlap few uniquely mappable reads, and consequently the mappability is low. To map the

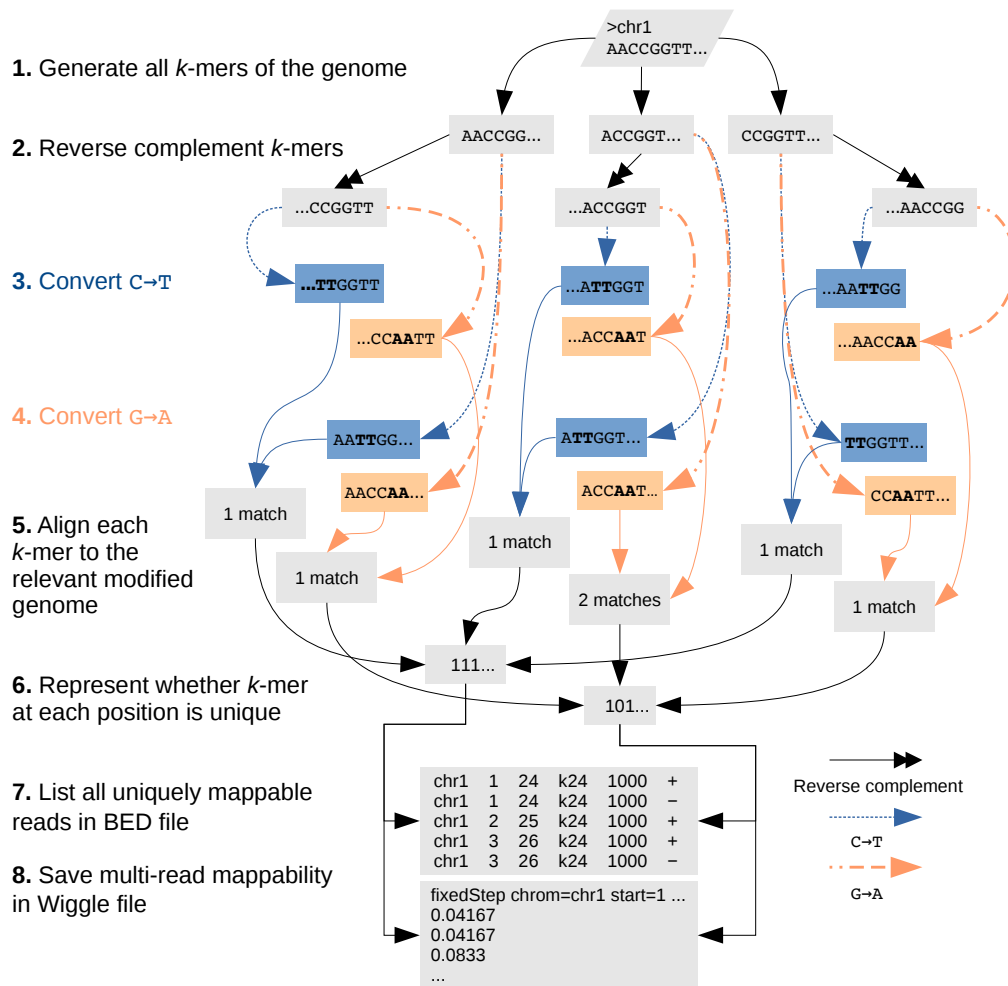


Figure 2: **Mappability of the methylome by Bismap.** Bismap identifies uniquely mappable k -mers of a bisulfite-converted genome. It simulates the same changes that may occur in bisulfite treatment on the + strand (C→T) and - strand (G→A). To account for sequence of the - strand, we generate an extra set of reverse-complemented chromosomes and then simulate bisulfite conversion on these chromosomes. We don't simulate reverse complementation after bisulfite conversion, because the experimental protocol does not involve post-conversion DNA amplification. We then align k -mers by disabling complement search and combine the resulting data to quantify the mappability of a bisulfite-converted genome.

56 regions with low mappability, a high sequencing depth is required to assure that sequencing reads
 57 completely overlap with few uniquely mappable reads in that region. If sequencing depth is low and
 58 genomic variation or sequencing error is high, the signal from a low mappability region is biased by
 59 reads falsely mapped to that region.

60 Most short-read alignment algorithms determine if any read maps to one or more regions in
 61 the genome. However, one must consider this in context of the surrounding regions, even if a read
 62 uniquely maps. A single nucleotide change might change a read from uniquely mappable to not. A

63 uniquely mappable read that aligns to a region with low mappability, has a high chance of mapping
64 incorrectly due to genetic variation or sequencing error.

65 In bisulfite sequencing, this problem increases. Bisulfite treatment reduces unmethylated cyto-
66 sine to uracil (sequenced as T) while 5-methylcytosine remains intact (sequenced as C). Bisulfite
67 treatment significantly increases the number of repeated short sequences in the genome. Many
68 regions uniquely mappable in an unmodified genome no longer uniquely map after bisulfite conver-
69 sion. Incorrect mapping of bisulfite sequencing reads creates a false methylation signal that can bias
70 downstream analysis and interpretation. When confounding factors such as read length, sequencing
71 depth or mutation rate differ among cases, this bias becomes even more evident.

72 In an unmodified human genome, 15.5% of the 24-mers do not map uniquely (Figure 1b). This
73 quantity increases to 30% for a bisulfite-converted genome (Figure 1b). In certain cases, the differ-
74 ence between a uniquely mappable and a non-uniquely mappable read can be only one nucleotide.
75 Sequencer base-calling errors and genetic variation often affect alignment, but we cannot comprehen-
76 sively account for them. These biases further exacerbate alignment when the read length is shorter,
77 emphasizing the importance of considering genomic mappability in any analysis involving short-read
78 sequencing. While previous tools such as the GEM mappability software¹ identify mappability of
79 the genome, no existing software solves the methylome mappability problem. To solve this problem,
80 we developed the Umap software, with a bisulfite mappability extension called Bimap.

81 **2 Methods**

82 **2.1 Single and multi-read mappability**

83 Umap efficiently identifies the uniquely mappable reads of any genome for a range of sequencing read
84 lengths. The Bimap extension of Umap produces uniquely mappable reads of a bisulfite-converted
85 genome. Both Umap and Bimap produce an integer vector for each chromosome that efficiently
86 defines the mappability for any region and can be converted to a browser extensible data (BED)
87 file. One way to assess mappability of a genomic region is by the **single-read mappability** — the
88 fraction of that region which overlaps with at least one uniquely mappable k -mer.

89 Analysis of sequencing data involves inferences about a base's genetic or regulatory state from
90 observations of all reads overlapping that base. Therefore, we must consider the mappability of all

91 reads overlapping a position or region, when estimating how many mapped reads we might expect.
92 Single-read mappability assumes that uniquely mappable reads are uniformly distributed in the
93 genome, while in reality we observe frequent localized enrichment of uniquely mappable reads.

94 A region can have 100% single-read mappability, but a below-average number of uniquely map-
95 pable reads that can overlap that region (Figure 1c). For example, a 1 kbp region with 100% single-
96 read mappability can be mappable due to a minimum of 10 unique non-overlapping 100-mers or a
97 maximum of 1100 unique highly overlapping 100-mers. Therefore, we define the **multi-read map-**
98 **pability** — the probability that a randomly selected read of length k in a given region is uniquely
99 mappable. For the genomic region $G_{i:j}$ starting at i and ending at j , there are $j - i + k + 1$ different
100 k -mers that overlap with $G_{i:j}$. The multi-read mappability of $G_{i:j}$ is the fraction of those k -mers
101 that are uniquely mappable (Figure 1c).

102 2.1.1 Mappability of the unmodified genome

103 Umap uses three steps to identify the mappability of a genome for a given read length k (Figure 1a).
104 First, it generates all possible k -mers of the genome. Second, it maps these unique k -mers to the
105 genome with Bowtie² version 1.1.0. Third, Umap marks the start position of each k -mer that aligns
106 to only one region in the genome. Umap repeats these steps for a range of different k -mers and
107 stores the data of each chromosome in a binary vector X with the same length as the chromosome's
108 sequence. For read length k , $X_i = 1$ means that the sequence starting at X_i and ending at X_{i+k}
109 is uniquely mappable on the + strand. Since we align to both strands of the genome, the reverse
110 complement of this same sequence starting at X_{i+k} in the - strand is also uniquely mappable.
111 $X_i = 0$ means that the sequence starting at X_i and ending at X_{i+k} can be mapped to at least two
112 different regions in the genome.

113 Eventually, Umap merges data of several read lengths to make a compact integer vector for each
114 chromosome (Figure 1a, step 3). In this vector, non-zero values at position X_i indicate the smallest
115 k -mer that position X_i to X_{i+K} is uniquely mappable with, where K is the largest k -mer in the
116 range. For example $X_i = 24$ means that the region X_i to X_{i+24} is uniquely mappable. This also
117 means that any read longer than 24 nucleotides that starts at X_i is also uniquely mappable.

118 Umap translates these integer vectors into six-column BED files for the whole genome (Figure 1a,
119 step 4). Additionally, Umap can calculate single-read mappability and multi-read mappability for

120 specified regions in any input BED file.

121 Although Bowtie can align with mismatches, here we do not use this capability. By defining
122 mappability with exact matches only, we provide baseline identification of regions that are not
123 uniquely mappable no matter how high the sequencing coverage. Nonetheless, the Umap software
124 allows users to change alignment options, including mismatch parameters.

125 **2.2 Mappability of the bisulfite-converted genome**

126 To identify the single-read mappability of a bisulfite-converted genome, we create two altered genome
127 sequences (Figure 2). In the first sequence, we convert all cytosines to thymine (C→T). In the
128 other sequence we convert all guanines to adenine (G→A). Our approach follows those of Bismark³
129 and BWA-meth⁴. We convert the genome sequence this way because bisulfite treatment converts
130 un-methylated cytosine to uracil which is read as thymine. Similarly the guanine that is base-
131 pairing with the un-methylated cytosine in the – strand converts to adenine. These two conversions,
132 however, never occur at the same time on the same read. We identify the uniquely mappable regions
133 of these two genomes separately, and then combine the data to represent the single-read mappability
134 of the + and – strands in the bisulfite-converted genome. For an unmodified genome, however, the
135 mappability of the + and – strand is identical by definition.

136 Bismap requires special handling of reverse complementation of C→T or G→A converted genomes.
137 Conversion of C→T on the sequence 5′-AATTCGG-3′ produces 5′-AATTTGG-3′. In the Bowtie
138 index, the reverse complement of the latter would be 5′-CCAAAATT-3′. For the purpose of identify-
139 ing the mappability of the bisulfite-converted genome, however, we expect the reverse complement
140 to be derived from the original converted sequence, yielding 5′-CCGGAATT-3′, and then after C→T
141 conversion, 5′-TTGGAATT-3′. Both + and – strands undergo bisulfite treatment simultaneously,
142 and there is no DNA replication to create new reverse complements after bisulfite treatment. To han-
143 dle this issue, Bismap creates its own reverse complemented chromosomes and suppresses Bowtie’s
144 usual reverse complement mapping.

145 Umap and Bismap each take ~ 200 core-hours on a 2.6 GHz Intel(R) Xeon CPU E5-2650 v2
146 processor to run for some read length. This is a massively parallelizable task, so on a computing
147 cluster with 400 cores, the task takes only 30 min of wall-clock time.

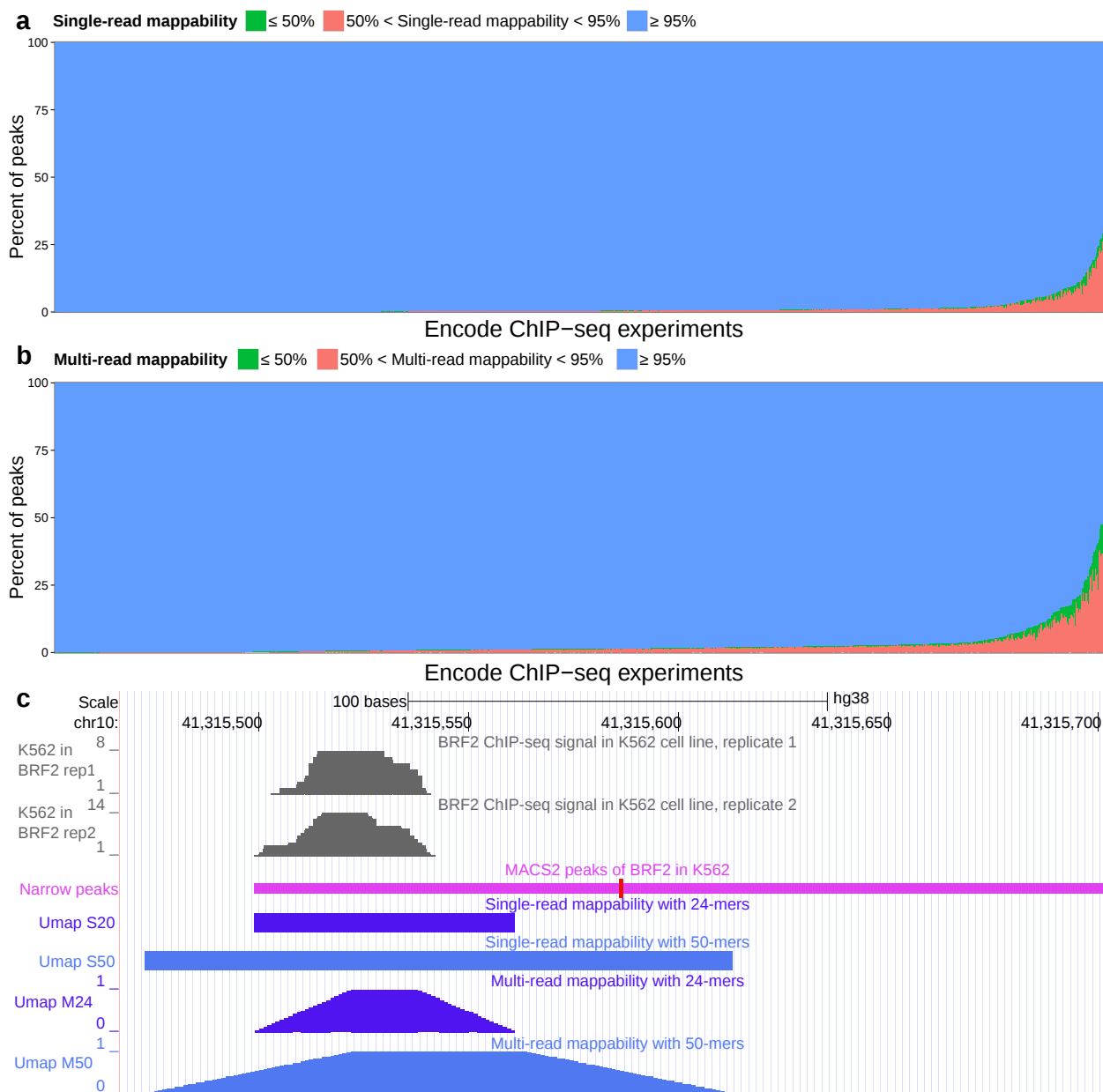


Figure 3: **Mappability of ChIP-seq peaks in 1193 ENCODE datasets.** (a) Single-read mappability and (b) multi-read mappability for narrow peaks identified in ENCODE ChIP-seq datasets. (c) A BRF2 narrow peak identified by MACS (purple) that is not uniquely mappable. Signal tracks (gray) show two different replicates of this ChIP-seq experiment in K562 chronic myeloid leukemia cells (ENCODE accessions ENCFF000YHB and ENCFF000YHD, read length 26 bp). Umap tracks show single-read and multi-read mappability for two different read lengths of 24 bp and 50 bp.

148 2.3 ENCODE ChIP-seq experiments

149 We downloaded ENCODE⁵ chromatin immunoprecipitation-sequencing (ChIP-seq) FASTQ files
 150 from the ENCODE Data Coordination Center⁶ and aligned them to GRCh38 using Bowtie⁷ 2. We
 151 switched to Bowtie 2 for this analysis because it supports gapped alignment, which we didn't need

152 for mappability calculations.

153 We used Samtools⁸ to remove duplicated sequences and those with a mapping quality of < 10 .
154 This assures that the probability of correct mapping to the genome for any read is > 0.9 . Pooling
155 replicates from the same experiment, we used MACS⁹ version 2 with `--nomodel` and `--qvalue`
156 `0.001` options to identify ChIP-seq peaks. Finally, Umap measured single-read mappability and
157 multi-read mappability within the peaks.

158 2.4 CpG islands

159 We downloaded CpG islands¹⁰ for GRCh38 from the UCSC Genome Browser¹¹ (http://epigraph.mpi-inf.mpg.de/download/CpG_islands_revisited). These CpG islands come from a hidden
160 Markov model (HMM) fitted to genomic G+C content. We then annotated CpG features around
161 the CpG islands following published definitions^{10,12} (Table 1). Then we used Umap and Bismap to
162 measure mappability across these annotations.
163

Annotation	Definition
CpG island	HMM fitted to G+C content
CpG shore	2 kbp area surrounding CpG islands
CpG shelf	2 kbp area surrounding CpG shores
CpG resort	Collection of islands, shores and shelves

Table 1: CpG annotations.

164 2.5 Whole-genome bisulfite sequencing analysis

165 First, we obtained datasets of whole-genome bisulfite sequencing of murine mammary tissues¹³
166 from the Sequence Read Archive (accession numbers SRR1946823, SRR1946824, SRR1946819, and
167 SRR1946820). Second, we trimmed Illumina TruSeq adapters from FASTQ files with Trim Galore¹⁴.
168 Third, for each experiment, we break down sequencing reads to produce two different FASTQ files
169 with read lengths of 50 bp and 100 bp. For example, if the read length of an experiment is 182 bp
170 and we want to generate a FASTQ file with read length of 50 bp, each sequencing read would
171 produce three different 50-bp sequencing reads (we would not use the remaining 32 bp). We aligned
172 these modified FASTQ files with BWA-meth⁴ to the GRCm38 genome. We extracted CpG-context
173 methylation using PileOmeth¹⁵. We use BSmooth¹⁶ (version 0.4.2) for identifying differentially

174 methylated regions. Finally, we used Bismap to measure mappability of differentially methylated
175 regions with at least four CpG dinucleotides.

176 **2.6 Other methylation assays**

177 DiseaseMeth¹⁷, a human methylation database, provides access to 17,024 methylation datasets from
178 88 different human diseases. These data are a collection of experiments using various platforms, in-
179 cluding 2728 assays using the Illumina Infinium HumanMethylation27 (27K) BeadChip, and 9795
180 assays using the Illumina Infinium HumanMethylation450 (450K) BeadChip. DiseaseMeth anno-
181 tates the genomic position of CpG dinucleotides covered by a 122 bp probe in the GRCh37 (hg19)
182 genome. To identify the mappability of these probes, we extended the genomic position of the CpG
183 dinucleotide to 61 bp on each direction (each probe is a 122 bp fragment centered on the annotated
184 CpG island¹⁸). We then measured single-read and multi-read mappability of these probes with
185 Bismap and a read length of 122 bp. In addition, we examined whether the exact 122-mer probe
186 sequence mapped uniquely.

187 DiseaseMeth also contains 71 experimental datasets using reduced representation bisulfite se-
188 quencing (RRBS)¹⁹. For CpG dinucleotides captured in RRBS experiments and annotated by Dis-
189 easeMeth, we examined the multi-read mappability for read lengths of 24 bp, 36 bp, 50 bp, and
190 100 bp.

191 **2.7 Umap and Bismap track hub**

192 We used read lengths of 24 bp, 36 bp, 50 bp, and 100 bp to generate mappability tracks for unmod-
193 ified and bisulfite-converted genomes of human (GRCh37 and GRCh38) and mouse (GRCm37 and
194 GRCm38). We store uniquely mappable regions of these genomes in bigBed format as a track hub
195 that can be loaded to UCSC or Ensembl genome browsers. The track hub contains one supertrack for
196 Umap and one supertrack for Bismap. The track hub is available at <http://bismap.hoffmanlab.org>.

197 **3 Results**

198 **3.1 Mappability of ENCODE ChIP-seq peaks**

199 ChIP-seq identifies proteins present in chromatin at particular loci and often involves short-read
200 sequencing. The ENCODE Project⁵ has performed around 1200 ChIP-seq assays on approximately
201 200 chromatin binding factors in more than 60 different human cell types. To show how mappability
202 affects downstream analysis of experiments such as ChIP-seq, we quantified the mappability of nar-
203 row peaks identified in ENCODE ChIP-seq experiments. Among 1193 experiments, most peaks map
204 uniquely. For some experiments, however, a high number of peaks overlap with non-uniquely map-
205 pable regions. Most of these experiments correspond to ChIP-seq of histone modifications with read
206 lengths from 24 bp to 36 bp. BRF2 (ENCODE accessions ENCFF000YHB and ENCFF000YHD) is
207 one of the few transcription factors with a high number of peaks that do not map uniquely (**Fig-**
208 **ure 3c**). This experiment used a read length of only 26 base pairs. The peak identified by MACS2
209 (**Figure 3c**) extends into a region that is not uniquely mappable. Although the ChIP-seq signal
210 is completely within a uniquely mappable region, MAC2 identifies a much broader peak than is
211 warranted.

212 **3.2 Mappability of CpG islands**

213 CpG islands substantially overlap transcription start sites and differentially methylated regions¹⁰.
214 Because CpG islands have a high number of CpGs, they are highly affected by bisulfite conversion.
215 Thus we investigated CpG islands and the neighboring CpG shores and CpG shelves.

216 Even with a relatively long read length of 100 bp, 3163/167,694 CpG annotations have zero
217 uniquely mappable bases, as calculated by Bismap. For shorter read lengths, even more of the
218 bisulfite-converted genome lacks unique mapping. For a read length of 100 bp, 16,396 CpG annota-
219 tions are not uniquely mappable with Bismap. This represents 9.8% of all CpG annotations. The
220 average single-read mappability of CpG annotations that are not uniquely mappable is 52%.

221 CpG islands and regions around them are often not uniquely mappable, to a lesser extent, in
222 an unmodified genome. For example, the average single-read mappability of CpG annotations that
223 are not uniquely mappable in the unmodified genome is 55% with a read length of 100 bp. This is
224 substantially lower than the average single-read mappability of the genome (92%). Also, there are

225 216 CpG islands that have some overlap with uniquely mappable regions of the unmodified genome,
226 but are not uniquely mappable in the bisulfite-converted genome.

227 The difference in genomic mappability and CpG island annotation mappability is even more
228 extensive for shorter read lengths. For example, for a read length of 24 bp, more than 80% of CpG
229 island annotations are not uniquely mappable, but the percent of the genome that is not uniquely
230 mappable is only 30% (Figure 4).

231 **3.3 Mappability of differentially methylated regions**

232 Many studies measure differences in methylation associated with a disease phenotype. These studies
233 test whether each CpG's methylation status correlates with the phenotype. Collective difference of
234 CpG dinucleotides in a given region, however, may provide higher statistical power in assessing the
235 association of methylation profile with disease states²⁰. Cluster of CpG dinucleotides are also a
236 more predictive feature of disease states than differences in individual CpGs²⁰. BSmooth¹⁶ is one
237 of the tools that identifies differentially methylated regions by estimating a smoothed methylation
238 profile.

239 We compared differences in CpG methylation of basal and luminal alveolar murine mammary
240 tissues¹³ using BSmooth¹⁶. Out of a total of 965,181 CpG dinucleotides identified with a read length
241 of 50 bp (see Methods), 4,091 of them are not uniquely mappable. For a read length of 100 bp, out of
242 a total of 1,136,993 CpG dinucleotides, 1,980 are not uniquely mappable. For the same experimental
243 setup, BSmooth identified 3082 differentially methylated regions for a read length of 50 bp and 3990
244 regions for a read length of 100 bp. For a read length of 100 bp, only 2 differentially methylated
245 regions were not uniquely mappable, while for a read length of 50 bp, 21 differentially methylated
246 regions were not uniquely mappable. This is a proof of principle that differential methylation analysis
247 can identify false signals that are not even uniquely mappable.

248 DiseaseMeth¹⁷ catalogs publicly available methylome datasets, including 12,073 using array
249 technologies. The cost-efficiency of these approaches has driven wide adoption. Many of these
250 datasets, however, include probes with low mappability in the bisulfite-converted genome. The
251 widely used Illumina Infinium methylation arrays use 122 bp probes centered on certain CpG
252 dinucleotides¹⁸. Out of the 27,574 probes in the Illumina Infinium HumanMethylation27 (27K)
253 BeadChip, 181 are not uniquely mappable. Additionally, 83 uniquely mappable probes have low

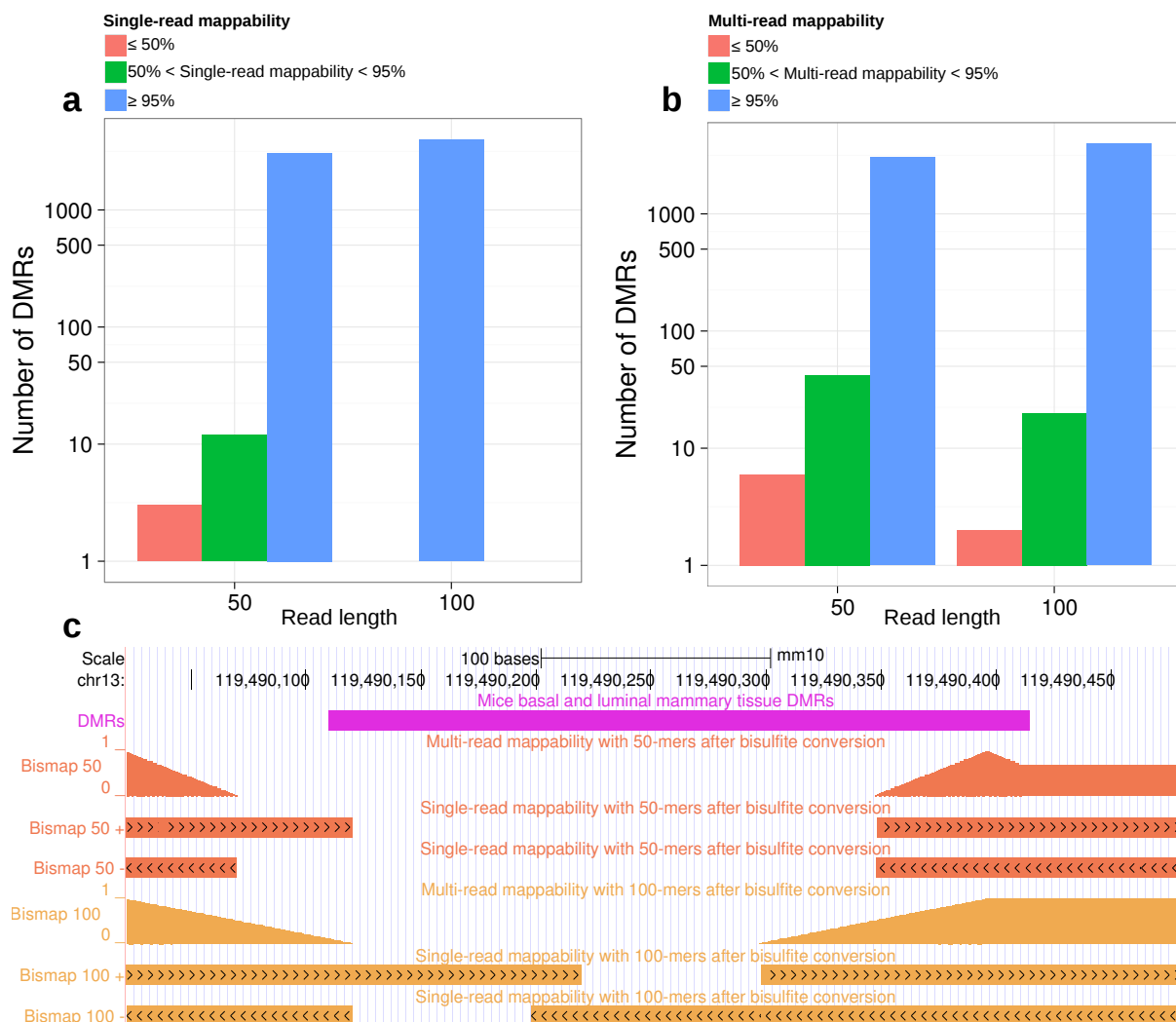


Figure 5: **Mappability of differentially methylated regions of mice mammary basal and luminal alveolar tissues.** (a) Single-read and (b) multi-read mappability of differentially methylated regions. (c) Example of a differentially methylated region identified with 50-nucleotide sequencing reads that is not uniquely mappable.

254 multi-read mappability, meaning that single nucleotide polymorphisms or mutations can result in
 255 probe multi-mapping (Figure 6a). Similarly, 962 probes in the Illumina Infinium HumanMethylation450 (450K) BeadChip are not uniquely mappable, and another 1,097 probes with single-read
 256 mappability, have low multi-read mappability (Figure 6b).

258 In addition, many publicly available RRBS datasets exist. In RRBS, only DNA fragments between
 259 40 bp and 220 bp are selected. The majority of selected fragments, however, are approximately
 260 50 bp²¹. Even with a read length of 100 bp, 329,799 (0.96%) of CpG dinucleotides annotated in RRBS
 261 experiments found in DiseaseMeth did not map uniquely (Figure 6c).

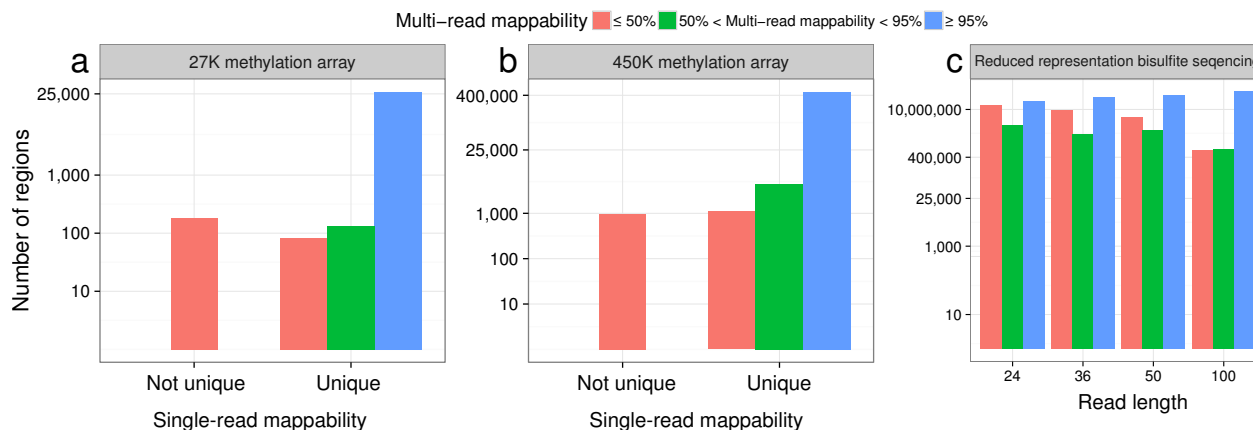


Figure 6: **Mappability of targeted methylation assays.** Multi-read mappability of probes in (a) the Illumina Infinium HumanMethylation27 (27K) BeadChip and (b) the Illumina Infinium HumanMethylation450 (450K) BeadChip. (c) Multi-read mappability of CpG dinucleotides found in DiseaseMeth RRBS datasets.

262 4 Discussion

263 4.1 The importance of considering mappability in analysis

264 In several examples we showed how mappability must be considered in analysis of sequencing data.
265 One needs to examine, however, the extent of genomic variation which affects mappability calcula-
266 tions. Genetic variants specific to each sample make it impossible to know the exact mappability.
267 We introduced a measure called multi-read mappability for addressing this issue. Genomic regions
268 with higher multi-read mappability are less prone to be biased by genetic variants and sequencing
269 errors.

270 In ENCODE ChIP-seq experiments using short read lengths, we found many examples where
271 signal was within a uniquely mappable region but peaks identified by peak caller had substantial
272 overlap with non-uniquely mappable regions. This shows how important it is to use the mappability
273 information in analysis pipeline of various experiments especially when the read length is small. In
274 fact, we initially developed Umap as part of the ENCODE uniform analysis pipeline⁵ to avoid such
275 problems.

276 We simulate bisulfite conversion assuming complete conversion of all cytosines, just as alignment
277 algorithms such as Bismark³ or BWA-meth⁴ do. In practice, chemical resistance or sample-specific
278 genetic variation may retard complete bisulfite conversion. This makes it impossible to estimate
279 the exact mappability for a bisulfite converted sample. When performing bisulfite sequencing on

280 different mouse strains, using the same reference genome for each introduces massive bias in bisulfite
281 sequencing data analysis²². Ideally, one would align data from each strain to a reference genome
282 specific to that strain. When one lacks a strain-specific reference genome, Bismap at least allows us
283 to quantify how and where genetic variation affects reliability of bisulfite sequencing results.

284 While paired-end sequencing with lengths greater than 100 bp has become more common, most
285 publicly available datasets such as ENCODE have used shorter reads. Out of 3,483 ENCODE ChIP-
286 seq experiments, 3,033 use single-ended sequencing, and 2,228 have read lengths of 36 bp or shorter.
287 Out of the 142 ENCODE RRBS datasets, 140 (98.6%) have a read length of 36 bp or shorter. In
288 addition, commonly used array technologies such as the 450K array uses 122 bp probes and multi-
289 read mappability of some of the probes is low. This allows multi-mapping due to genetic variation
290 and decreases data quality in these regions. Although non-uniquely mappable methylation array
291 probes constitute only a small fraction of all probes (0.66% in the 27K array and 0.2% in the 450K
292 array), one must still use caution when interpreting methylation signal—or the lack thereof—in
293 these regions.

294 In our analysis of whole genome bisulfite sequencing data of mouse mammary tissue, among the
295 CpG dinucleotides that had a minimum coverage of 3 reads in all of the 5 different whole genome
296 bisulfite sequencing datasets, ~0.15% were not uniquely mappable with 50 bp and 100 bp reads.
297 Such CpG dinucleotides must be excluded from analysis. RRBS usually involves filtering fragments
298 to only include those that are 40 bp–220 bp, and most RRBS reads are 50 bp or less²¹. This causes
299 a major issue for mapping of these reads.

300 4.2 Other methods for mappability

301 Bias Elimination Algorithm for Deep Sequencing (BEADS²³) also defines a mappability measure
302 that is obtained by identifying uniquely mappable 35-mers of the genome. Based on the assumption
303 that each read identifies a longer 200-mer, BEADS extends uniquely mappable 35-mers to 200 bp,
304 and calculates the fraction of reads that span a given genomic position. BEADS uses a cutoff of 25%
305 mappability to filter signals that might bias a study. Extending the 35-mer mappability to 200 bp,
306 however, defines the exact mappability for neither 35-mers nor 200-mers.

307 PeakSeq²⁴, uses an algorithm similar to Umap and identifies the single-read mappability in 1 kbp
308 windows of the genome. PeakSeq filters out ChIP-seq signals with low mappability in each window

309 by comparing it to a simulated background of reads with Poisson distribution.

310 Model-based one and two Sample Analysis and inference for ChIP-Seq Data (MOSAiCS)²⁵
311 uses a mappability measure similar to multi-read mappability for preprocessing of data. While
312 Umap’s multi-read mappability calculates the percent of uniquely mappable k -mers that span each
313 nucleotide, MOSAiCS calculates the percent of *extended uniquely mappable k -mers* for calculating
314 its mappability score. In comparison to other mappability measures, Umap’s multi-read mappability
315 has the advantages of specificity to an exact read length and efficient calculation for any read length.

316 Acknowledgements

317 We would like to thank Scott M. Lundberg for providing us with GRCh38-aligned BAM files of
318 the ENCODE ChIP-seq datasets. We also thank Carl Virtanen and Zhibin Lu at the University
319 Health Network High Performance Computing Centre and Bioinformatics Core for technical as-
320 sistance. This work was supported by the Canadian Cancer Society (703827 to M.M.H.), Ontario
321 Institute for Cancer Research (OICR), the Natural Sciences and Engineering Research Council
322 of Canada (RGPIN-2015-03948 to M.M.H. and RGPIN-435512-2013 to C.E.), the University of
323 Toronto McLaughlin Centre (MC-2015-16 to M.M.H.), and the Princess Margaret Cancer Founda-
324 tion.

325 References

- 326 [1] T. Derrien, J. Estelle, S. Marco Sola, D. G. Knowles, et al. Fast computation and applications of genome mappability. *PLoS*
327 *ONE*, 7(1):e30377, Jan 2012.
- 328 [2] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the
329 human genome. *Genome Biol.*, 10(3):R25, 2009.
- 330 [3] F. Krueger and S. R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*,
331 27(11):1571–1572, Jun 2011.
- 332 [4] B. S. Pedersen, K. Eyring, S. De, I. V. Yang, and D. A. Schwartz. Fast and accurate alignment of long bisulfite-seq reads. *ArXiv*
333 *e-prints*, January 2014.
- 334 [5] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74,
335 Sep 2012.
- 336 [6] The encode data coordination center. <https://www.encodeproject.org/>. Accessed: 2016-06-05.
- 337 [7] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, Apr 2012.

- 338 [8] H. Li, B. Handsaker, A. Wysoker, T. Fennell, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25
339 (16):2078–2079, Aug 2009.
- 340 [9] Y. Zhang, T. Liu, C. A. Meyer, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.
- 341 [10] H. Wu, B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg. Redefining CpG islands using hidden Markov models. *Bio-*
342 *statistics*, 11(3):499–514, Jul 2010.
- 343 [11] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, et al. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006,
344 Jun 2002.
- 345 [12] M. Bibikova, B. Barnes, C. Tsan, V. Ho, et al. High density DNA methylation array with single CpG site resolution. *Genomics*,
346 98(4):288–295, Oct 2011.
- 347 [13] C. O. Dos Santos, E. Dolzhenko, E. Hodges, A. D. Smith, and G. J. Hannon. An epigenetic memory of pregnancy in the mouse
348 mammary gland. *Cell Rep*, 11(7):1102–1109, May 2015.
- 349 [14] Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. Accessed: 2016-06-05.
- 350 [15] PileOmeth. <https://github.com/dpryan79/PileOMeth>. Accessed: 2016-06-05.
- 351 [16] K. D. Hansen, B. Langmead, and R. A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially
352 methylated regions. *Genome Biol.*, 13(10):R83, 2012.
- 353 [17] L. Jie, L. Hongbo, S. Jianzhong, W. Xueting, et al. DiseaseMeth: a human disease methylation database. *Nucleic acids research*,
354 40:D1030–5, 2011.
- 355 [18] Illumina. CpG Loci Identification. [http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/
356 technote_cpg_loci_identification.pdf](http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_cpg_loci_identification.pdf).
- 357 [19] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, et al. Reduced representation bisulfite sequencing for comparative high-
358 resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877, 2005.
- 359 [20] M. D. Robinson, A. Kahraman, C. W. Law, H. Lindsay, et al. Statistical methods for detecting differentially methylated loci and
360 regions. *Front Genet*, 5:324, 2014.
- 361 [21] Z. Sun, J. Cunningham, S. Slager, and J. Kocher. Base resolution methylome profiling: considerations in platform selection, data
362 preprocessing and analysis. *Future Medicine*, 7(5):813–828, 2015.
- 363 [22] P. Wulfridge, B. Langmead, A. P. Feinberg, and K. Hansen. Choice of reference genome can introduce massive bias in bisulfite
364 sequencing data. *bioRxiv*, 2016.
- 365 [23] M. S. Cheung, T. A. Down, I. Latorre, and J. Ahringer. Systematic bias in high-throughput sequencing data and its correction
366 by BEADS. *Nucleic Acids Res.*, 39(15):e103, Aug 2011.
- 367 [24] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, et al. PeakSeq enables systematic scoring of ChIP-seq experiments
368 relative to controls. *Nat. Biotechnol.*, 27(1):66–75, Jan 2009.
- 369 [25] P. F. Kuan, D. Chung, G. Pan, J. A. Thomson, et al. A Statistical Framework for the Analysis of ChIP-Seq Data. *J Am Stat*
370 *Assoc*, 106(495):891–903, 2011.