Web-based Supplementary Materials for "BM-Map: Bayesian Mapping of Multireads for Next-Generation Sequencing Data"

Yuan Ji [1,*], *Yanxun Xu*[2], *Qiong Zhang*[3], *Kam-Wah Tsui*[3],
Yuan Yuan[4], *Clift Norris*[1], *Shoudan Liang*[4], *Han Liang* [4,*]

1. Department of Biostatistics, M.D. Anderson Cancer Ctr., Houston, Texas, U.S.A.

2. Department of Statistics, Rice University, Houston, Texas, U.S.A.

3. Department of Statistics, University of Wisconsin – Madison, Wisconsin, U.S.A.

4. Dept. of Bioinformatics and Computational Biology, M. D. Anderson Cancer Ctr., Houston, Texas, U.S.A.

A: Main idea of the BM-Map

As stated in the main text, the key idea is to borrow information stored in the unique reads to estimate the model parameters. We utilize three sources of information when mapping the multireads, the sequencing mismatch profiles, the likelihood of hidden nucleotide variations, and the expression levels of competing genomic locations. See Figure 1.

For a multiread that can be aligned to multiple genomic locations, the set of unique reads are defined as those with the starting or ending base falling in-between the range of the genomic location. See Figure 2.

B: Sequencing error rates α_{kt} and $\alpha_{l,kt}$

There are two options for estimating the values of α's in the Bayesian model. First, we can compute the observed position-specific error rates using all the unique reads. The 35 sample proportions are summarized in left panel of Figure 3. This option was used for the simulation studies in the main text.

Alternatively, individual quality scores are assigned to the bases of the mapped reads. We can compute a sample mismatch proportion for each of the quality scores. In the yeast RNA-Seq data, there are a total of 41 different quality scores. The right panel plots the sample mismatch proportion for each of the 41 quality scores using all the unique reads in the yeast genome data.

In our RNA-Seq data analysis, we used the values in the right panel as $\alpha_{l,kt} \equiv \alpha_{l,Q_k t}$, where $Q_k \in \{1, \ldots, 41\}$ is the quality score of the base (of unique read l) matched with the k-th position of location t.

C: Additional simulation setup

Table 1 presents the 21 simulations described in the simulation setup of the original paper.

D: Additional simulation results – ROC

Figure 4 presents 21 ROC curves along with the areas under the curve (AUC) for all the three methods in the simulation. The BM-Map is superior than the other two methods under comparison.

E: RPKM as gene expression

In the paper, we used read counts to quantify gene expression. An alternative approach is to compute *the number of reads that map per kilobase of exon model per million mapped reads (RPKM).* A larger RPKM indicates larger gene expression. A standard algorithm for computing the RPKM is given below. We compare the RPKMs based on three read mapping approaches: the BM-Map method, the proportional method, and the industry standard. Most practitioners simply discard all the multireads in computing the RPKM. We call this the naive method. In the BM-Map and proportional methods, we use the probabilities of mapping a multiread to each genomic location as the count number in the first step of computing the RPKM.

Algorithm for computing the RPKM:

1. Count the number of reads (multireads and unique reads) mapped to the genes in millions. Call that number m.

2. Count the number of bases in all the exons of the gene in kilo-bases. Call that number l.

3. Count the total number of reads that have been mapped to the entire genome in millions. Call that number t.

4. RPKM $= m/l/t$.

   We get one RPKM value per gene. In the BM-Map and proportional methods, we use the probabilities of mapping a multiread to each genomic location as the count number in the first step.

There are 5,862 known yeast genes, which result in 5,862 RPKMs for each of the three methods. Figure 5 compares the pairwise RPKMs among the three methods for both the yeast data and human data.

Table 1: The proposed 21 simulation cases with different configurations of sequence difference (*Diff*), hidden nucleotide variation (*Mut*), and imbalanced expression (*Exp*). There are a total of seven configurations of *Diff*, *Mut*, and *Exp*, for each of three sample sizes (see Section 3.1 of the main text).

Case Index	Sequence difference (*Diff*)	Hidden nucleotide variation variation (*Mut*)	Imbalanced expression (*Exp*)
1-3	Yes	No	No
4-6	No	Yes	No
7-9	No	No	Yes
10-12	Yes	No	Yes
13-15	Yes	Yes	No
16-18	No	Yes	Yes
19-21	Yes	Yes	Yes

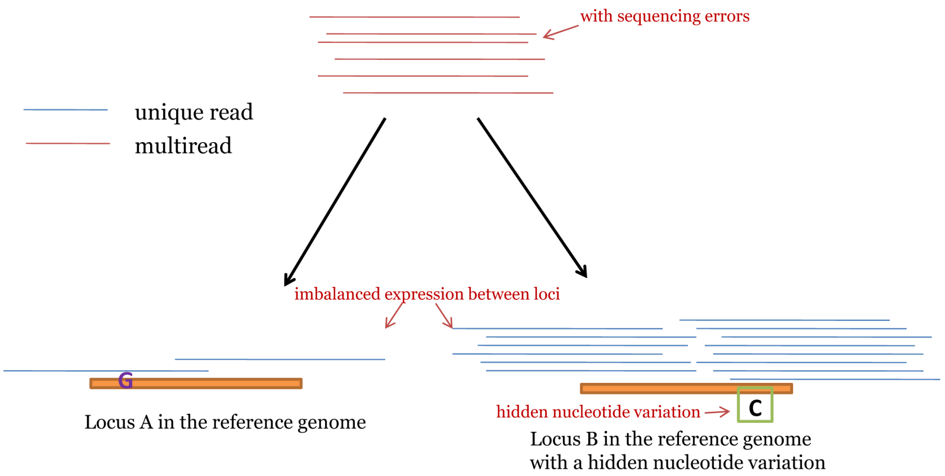

with sequencing errors
unique read
multiread
imbalanced expression between loci
G
Locus A in the reference genome
hidden nucleotide variation
C
Locus B in the reference genome
with a hidden nucleotide variation

Figure 1: (Colored) An illustration of the three sources of information used in the BM-Map method. (1) The short read may be sequenced with errors; (2) One genomic location may have a larger expression than the other (Locus B has more unique reads than locus A); (3) The competing genomic locations may have different sequences due to hidden nucleotide variation (boxed C in the locus B).

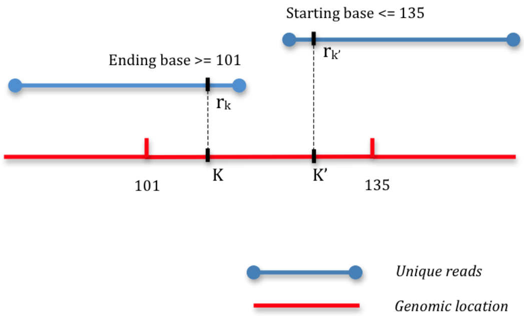


Figure 2: (Colored) For a given location spanning from position 101 to 135, any unique reads with starting base or ending base $\in [101, 135]$ will be included in our analysis.

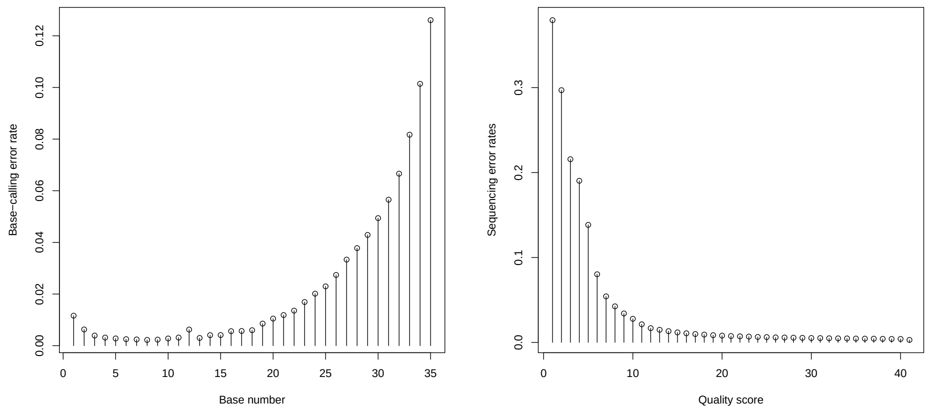
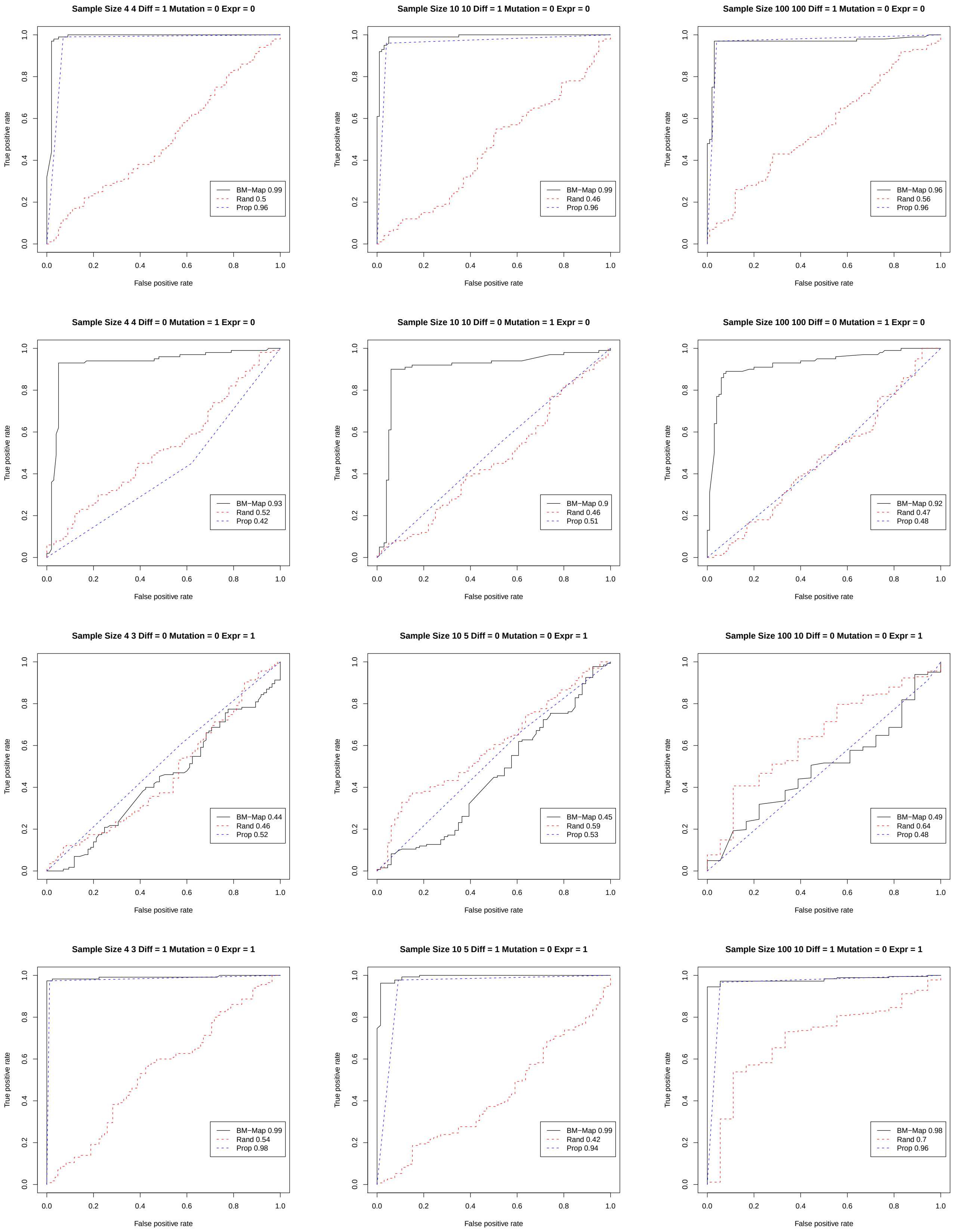
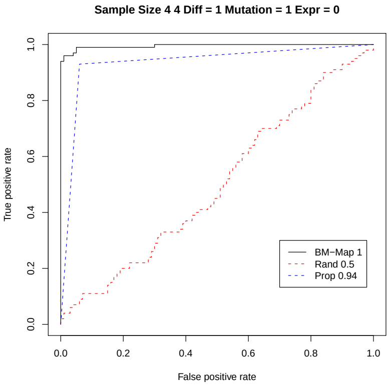


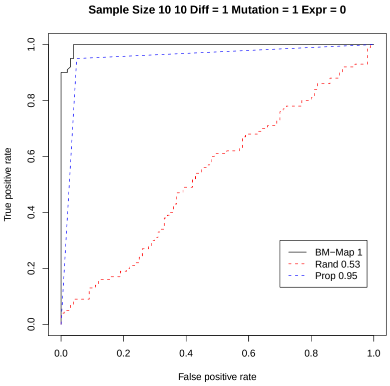
Figure 3: Observed sequencing error rates as a function of the base number (left panel) or quality score (right panel) using the unique reads from the yeast data set.

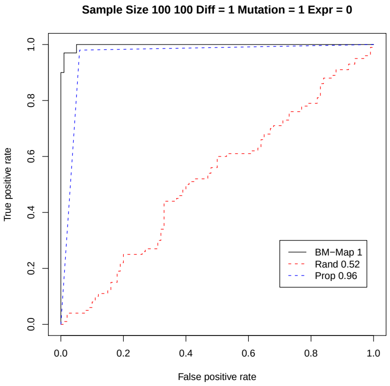
Figure 4: (Colored) The ROC curves and resulting area under the curve (AUC) for the three methods under comparison in the simulation. The three rows of ROC plots correspond to the scenarios 13-15, 16-18, and 19-21 in Table 1 of the manuscript. The sample sizes (number of unique reads) in each scenario is on top of each ROC plot.

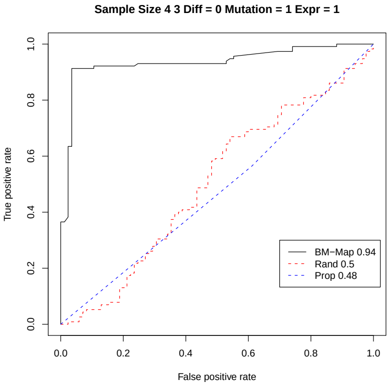


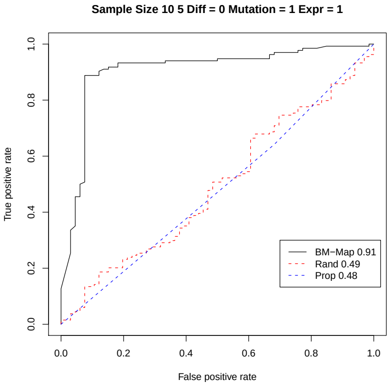
ccc

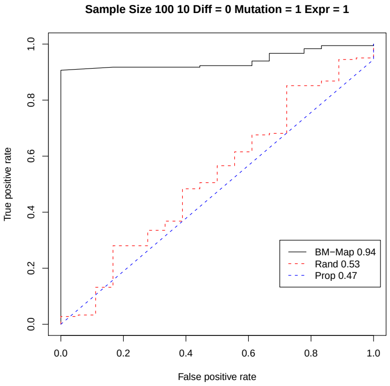

Sample Size 4 4 Diff = 1 Mutation = 1 Expr = 0

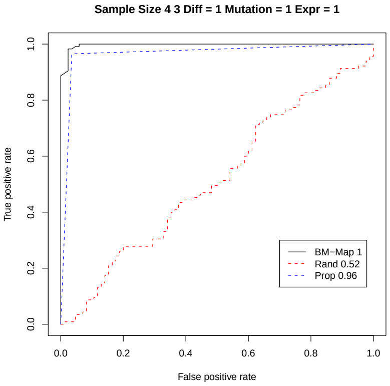

Sample Size 10 10 Diff = 1 Mutation = 1 Expr = 0

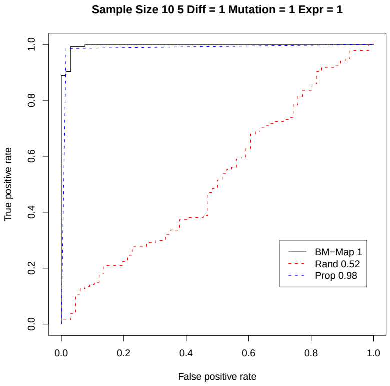

Sample Size 100 100 Diff = 1 Mutation = 1 Expr = 0

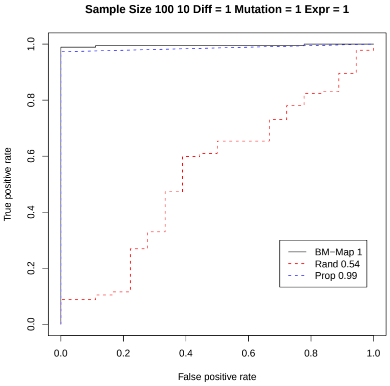

Sample Size 4 3 Diff = 0 Mutation = 1 Expr = 1


Sample Size 10 5 Diff = 0 Mutation = 1 Expr = 1


Sample Size 100 10 Diff = 0 Mutation = 1 Expr = 1


Sample Size 4 3 Diff = 1 Mutation = 1 Expr = 1


Sample Size 10 5 Diff = 1 Mutation = 1 Expr = 1


Sample Size 100 10 Diff = 1 Mutation = 1 Expr = 1

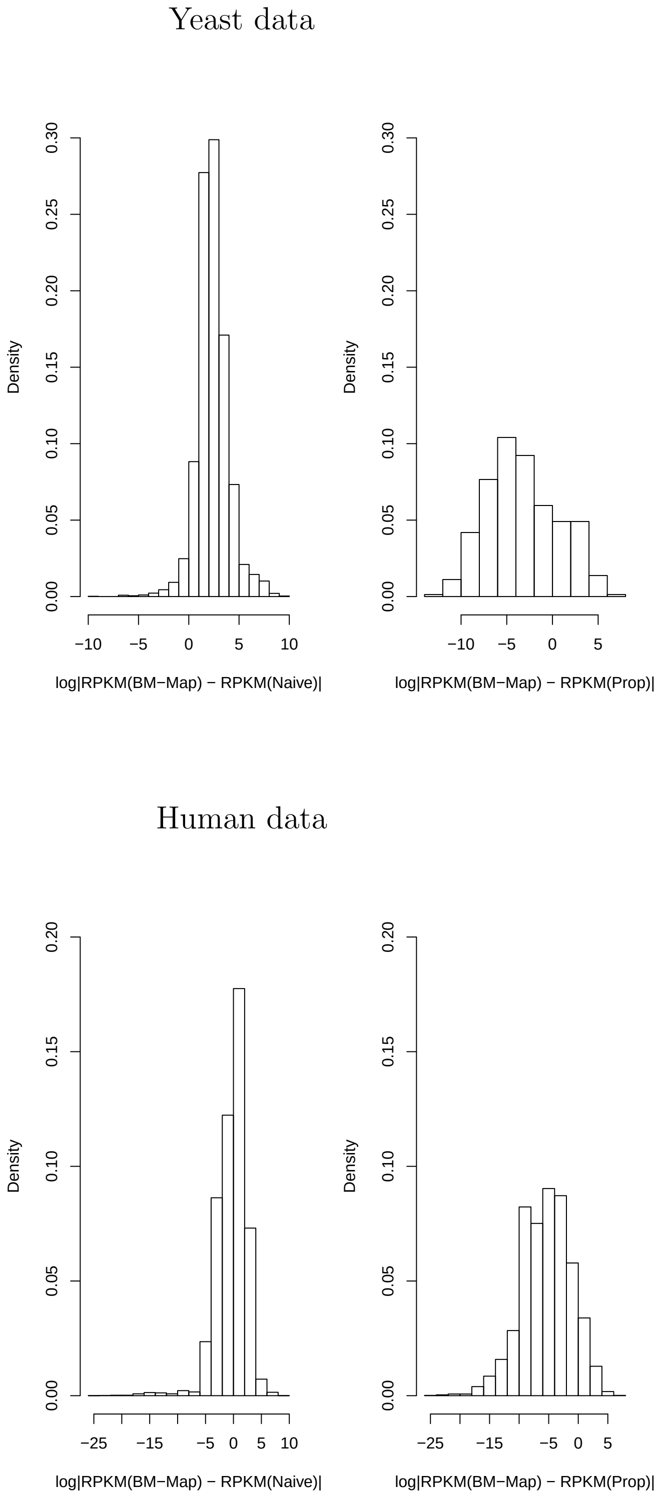


Figure 5: Results comparing the RPKMs from three methods. Shown are the log absolute differences in the RPKMs between the BM-Map and the Naive method (left panel) and between the BM-Map and the proportional method (right panel) for the yeast data (top) and the human data (bottom).