Supplementary Data for "A new strategy to reduce allelic bias in RNA-Seq read-mapping"

Ravi Vijaya Satya*, Nela Zavaljevski, and Jaques Reifman*
DoD Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, US Army Medical Research and Materiel Command, Fort Detrick, MD 21702, USA

*Corresponding authors: rvijaya@bhsai.org, jaques.reifman@us.army.mil

1. Construction of the enhanced reference

Our approach to construct the enhanced reference assumes a fixed read length *r*. If read lengths are variable, the shorter reads will map to multiple segments within the enhanced reference, resulting in lower mapping quality.

1.1 Objectives in constructing the enhanced reference

The following are the main design principles used in the construction of the enhanced reference:

Objective 1: The enhanced reference should contain all possible haplotypes within any length-*r* window of the genome. This guarantees that all possible reads from any individual have an exact match within the enhanced reference. It ensures that all possible error-free reads from that region of the genome are equally likely to be mapped irrespective of whether they carry the reference alleles, non-reference alleles, or any combination of both. If an *r*-window contains a single polymorphic locus, this results in adding just one enhanced segment (Fig. 1A in the main text). Conversely, if there are *k* polymorphic loci within the *r*-window, this will result in the addition of 2^k-1 enhanced segments.

Objective 2: Within any window of length ≥*r*, none of the added segments in the enhanced reference should be identical within themselves or with the original reference genome. This is necessary for two reasons: 1) to ensure that the length of the added segments is kept to a minimum, and 2) to eliminate avoidable scenarios in which a read matches two different segments of the enhanced reference equally well, thereby reducing the perceived mapping quality of the read.

1.2 Approach

In any window of length *r* with *k* polymorphic loci, the total number of haplotypes possible is 2^k, assuming that there are only two possible alleles at each locus. Therefore, to ensure that all possible haplotypes are represented in the enhanced reference, we need to add 2^k-1 segments to the original genome to build a complete enhanced reference. For simplicity in illustration, we can represent the two alleles at each locus with '0' and '1', where '0' indicates the reference allele and '1' indicates the non-reference allele. Fig. S1 shows a scenario with 3 polymorphic loci within an *r*-window. The 8 possible haplotypes within this window are {000,001,010,011,100,101,110,111}. By definition, the reference genome carries the haplotype 000. Fig. S1B shows the remaining 7 haplotypes missing from the reference. Including segments of length *r*-1 on either side of each of these haplotypes ensures that any *r*-window overlapping any of the haplotypes is part of an enhanced segment. Fig. S1B shows these haplotypes extended by *r*-1 on either side. While these segments satisfy Objective 1, they do not satisfy Objective 2: specifically, Fig. S1B shows two instances in which regions of length greater than *r* are identical between segments. For instance, the reference genome and the segment containing the haplotype 100 are identical to the right of S_1.

To satisfy Objective 2, we need to exclude the redundant regions while making sure that Objective 1 is still satisfied. There are many possible ways of achieving this: Fig. S1C shows only one of the possible solutions. In the example above, the segment containing haplotype 100 extends only *r*-1 bases to the right of S_1, thereby ensuring that it is not identical to the reference genome in any *r*-window.

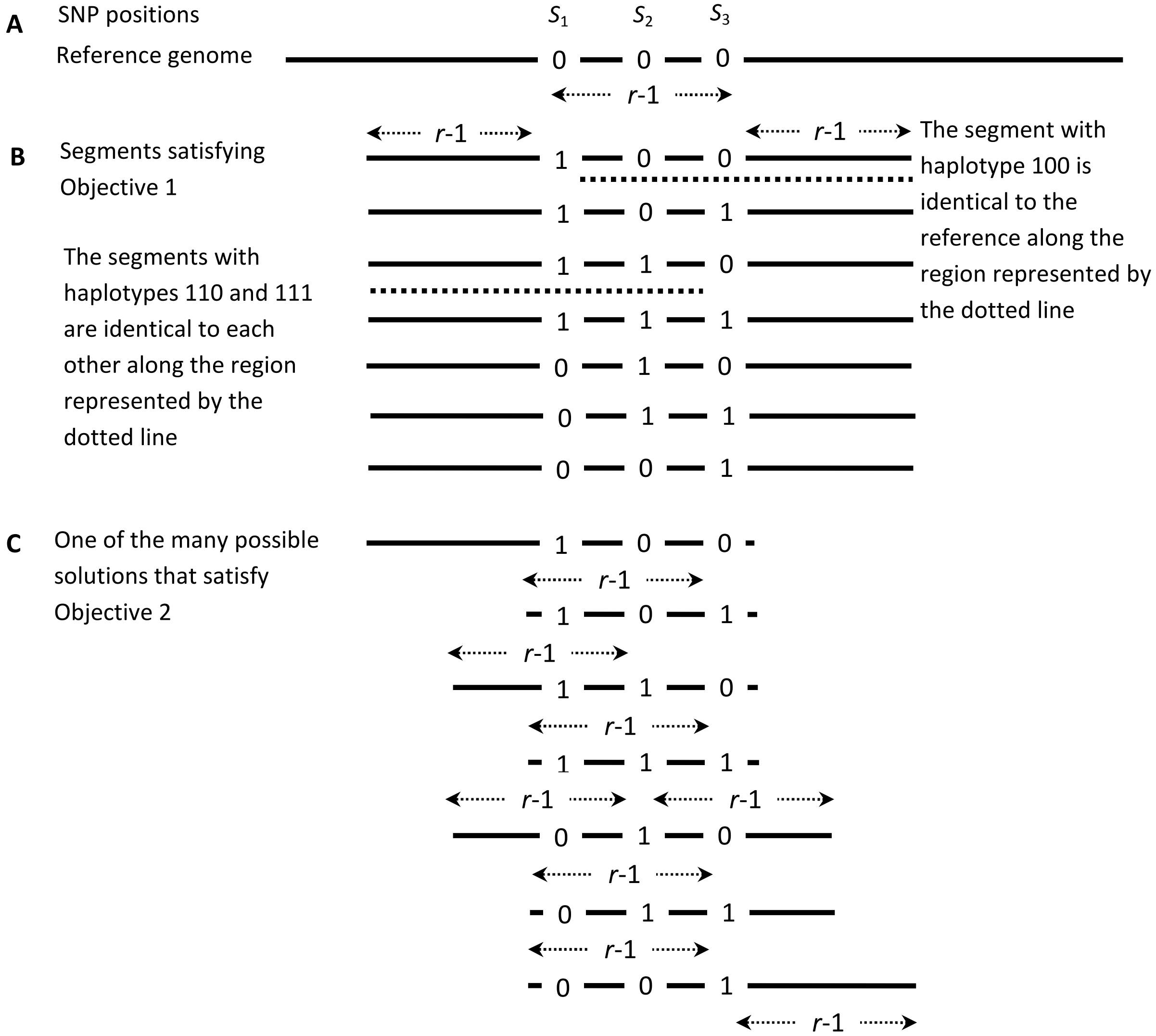

A
SNP positions
Reference genome
S1
S2
S3
0
0
0
r-1
B
Segments satisfying Objective 1
r-1
r-1
The segment with haplotype 100 is identical to the reference along the region represented by the dotted line
The segments with haplotypes 110 and 111 are identical to each other along the region represented by the dotted line
1 0 0
1 0 1
1 1 0
1 1 1
0 1 0
0 1 1
0 0 1
C
One of the many possible solutions that satisfy Objective 2
1 0 0
r-1
1 0 1
r-1
1 1 0
r-1
1 1 1
r-1
r-1
0 1 0
r-1
0 1 1
r-1
0 0 1
r-1

Fig. S1. An instance with three SNPs within a single *r*-window. The symbol '0' indicates the reference allele and '1' indicates the non-reference allele. (**A**) The reference genome contains the haplotype '000' by definition. (**B**) The seven possible remaining haplotypes. (**C**) One of the many possible solutions for adding enhanced segments that satisfy both objectives.

1.3 Algorithm

The solution presented in Fig. S1C can be formalized and generalized for k polymorphic loci using a greedy algorithm. This algorithm scans the reference genome from left to right. At each polymorphic locus, the algorithm adds all the necessary segments that carry the non-reference allele at that locus. For instance, the solution shown in Fig. S1C is built by adding segments with haplotypes {100, 101, 110, and 111} at S_1, {010 and 011} at S_2, and {001} at S_3. Fig. S2 gives a high-level description of this algorithm.

```
Procedure build_enhanced_reference
Inputs: Reference genome $G$, sorted list of polymorphic loci $S$, maximum read length $r$
Output: Enhanced reference $G_e$
$G_e \leftarrow G$
for $i$ from 1 to $|S|$
        $Pos \leftarrow S(i)$
        $K \leftarrow$ number of SNPs in the window $G[Pos, Pos+r-1]$
        $V \leftarrow$ list of all possible $2^{k-1}$ binary vectors of length $k$ such that for each $v \in V$, $v[0]=1$
        for each $v \in V$
                $start \leftarrow S(i + \max\{j | v[j]=1\}) - r + 1$
                $end \leftarrow Pos + r - 1$
                $E \leftarrow G[start, end]$
                for each $j | v[j]=1$
                        $E[S(i+j) - start] \leftarrow \mathrm{NonRef}(S(i+j))$
                end for
                $G_e \leftarrow G_e \cup E$
        end for
end for
```

Fig. S2. Algorithm for constructing the enhanced reference.

Lemma 1 below proves that the procedure *build_enhanced_reference* satisfies Objective 1, and Lemma 2 proves that it satisfies Objective 2.

Lemma 1: Every possible haplotype within any r-window is part of the original reference genome or an enhanced segment added by the algorithm.

Proof: Consider any window $W[i, i+r-1]$ of length r. By definition, the reference genome consists of the haplotype in which each SNP is ‘0’. Let H be a haplotype within W with at least one SNP S such that $H[S]=1$. We will show that H will be a part of an enhanced segment added by the algorithm. Let S_l be the left most SNP in W such that $H[S_l]$ =1, and S_m be the right most SNP in W such that $H[S_m]$ =1. By definition, $H[S_p]=0$ for every SNP S_p such that $i \le S_p < S_l$ and every SNP S_p such that $S_m < S_p \le (i+r-1)$. When the algorithm reaches $S_{l,}$, it adds an enhanced segment for each haplotype in the window $[S_l , S_l + r-1]$. Therefore, it includes an enhanced segment E such that $E\ [S_{l,}]=1$, $E\ [S_{m,}]=1$, $E[S_k]=H[S_k]\ \forall\{S_k | \ S_l \le S_k \le S_m\}$, and $E[S_{p,}]=0$ for all other SNPs within E. Hence, E is identical to H in every SNP position. Now, it is enough to show that E covers the entire window W. According to the algorithm, the left end of E will be

(S_m-r+1). Since $S_m \leq (i+r-1)$ by definition, $i \geq (S_m-r+1)$, and hence E covers the left end of W. The right end of E is (S_l+r-1). Since $S_l \geq i$, $(S_l+r-1) \geq (i+r-1)$, and E covers the right of W. Hence the haplotype H will be covered by the enhanced segment E. ◊

Lemma 2: In any r-window, an enhanced segment added by the algorithm is neither identical to the reference nor identical to any other enhanced segment.

Proof: It can be trivially shown that each enhanced segment added by the algorithm is different from the original reference in any r-window. We will show by contradiction that no two enhanced segments can be identical in an r-window. Let as assume that there are two enhanced segments E_1 and E_2 that are identical in an r-window $W[i, i+r-1]$. Let S_l be the left most SNP in W such that $E_1[S_l]=E_2[S_l]=1$, and S_m be the right most SNP in W such that $E_1[S_m]=E_2[S_m]=1$. Let us also assume that E_1 was added by the algorithm at a SNP S_1 and E_2 at a SNP S_2. Since $E_1[S_p]=0$ for any SNP $S_p<S_1$, we know that $S_l \geq S_1$. Similarly, we also know that $S_l \geq S_2$. Also, $S_m-S_1 \leq (r-1)$ and $S_m-S_2 \leq (r-1)$, since $E_1[S_m]$ or $E_2[S_m]$ cannot be '1' otherwise.

> *Case 1*: $S_1=S_2$. If S_1 and S_2 are the same SNP, then $S_1=S_2= S_l$ by definition. This is not possible, since the algorithm only adds one enhanced segment for each haplotype at any SNP, and hence E_1 and E_2 will be different for at least one SNP within W.
>
> *Case 2*: $S_1<S_2$. The right end of E_1 will be (S_1+r-1), and the right end of E_2 will be (S_2+r-1). However, $i>S_1$, since $E_1[S_1]=1$ and $E_2[S_1]=0$, and E_1 and E_2 will not be identical within W otherwise. But, if $i>S_1$, E_1 and E_2 can overlap in at most r-1bases, as E_1 ends at (S_1+r-1). Hence, E_1 and E_2 cannot be identical in an r-window.
>
> *Case 3*: $S_1<S_2$. Similar to *Case 2*. E_1 and E_2 cannot be identical in an r-window.

Hence, our initial assumption is wrong. There can be no two enhanced segments that are identical in an r-window. ◊

1.4 Limitations and assumptions

The algorithm makes the following assumptions:

- *Each SNP position is bi-allelic*: The great majority (>99.99%) of the common SNPs in the human genome have only two known alleles. Therefore, this assumption does not limit the effectiveness of the algorithm, in practice.
- *All the reads have the same fixed read length r*: The algorithm can still be used even if the read lengths are variable; in that case r should be set to the maximum possible read length. In this case, Objective 2 will not be satisfied. However, this will have a very minimal impact on the ability to map any read, since all possible haplotypes are present in the enhanced reference. The only negative impact is that reads shorter than r can then have an exact match with multiple segments in the enhanced reference. This results in a lower mapping quality for these reads.
- *The number of SNPs within any r-window is bound by a constant*: The number of possible haplotypes within a window increases exponentially with the number of polymorphic loci within the window. Therefore, this approach is only practical when the number of polymorphic loci within an r-window is much smaller than r, ideally not more than 5 or 6. This, in fact, is the case in the human genome. In the very few instances where there are a large number of possible SNPs within an r-window, we can

take the first few SNPs and ignore the remaining ones while constructing the enhanced segments at any position.

2. Results on simulated reads

2.1 Mapping statistics for simulated reads with mapping qualities ignored

The following figures and tables provide detailed results for the simulated reads mapped using MAQ and BWA. These figures are only intended to compare the effect of the three alternative reference construction strategies in accurately identifying allele-specific expression. These results are not intended to compare the performance of the MAQ and BWA programs. In obtaining these results, BWA was run with default settings, while some of the default settings for the MAQ program were modified as described in the main text.

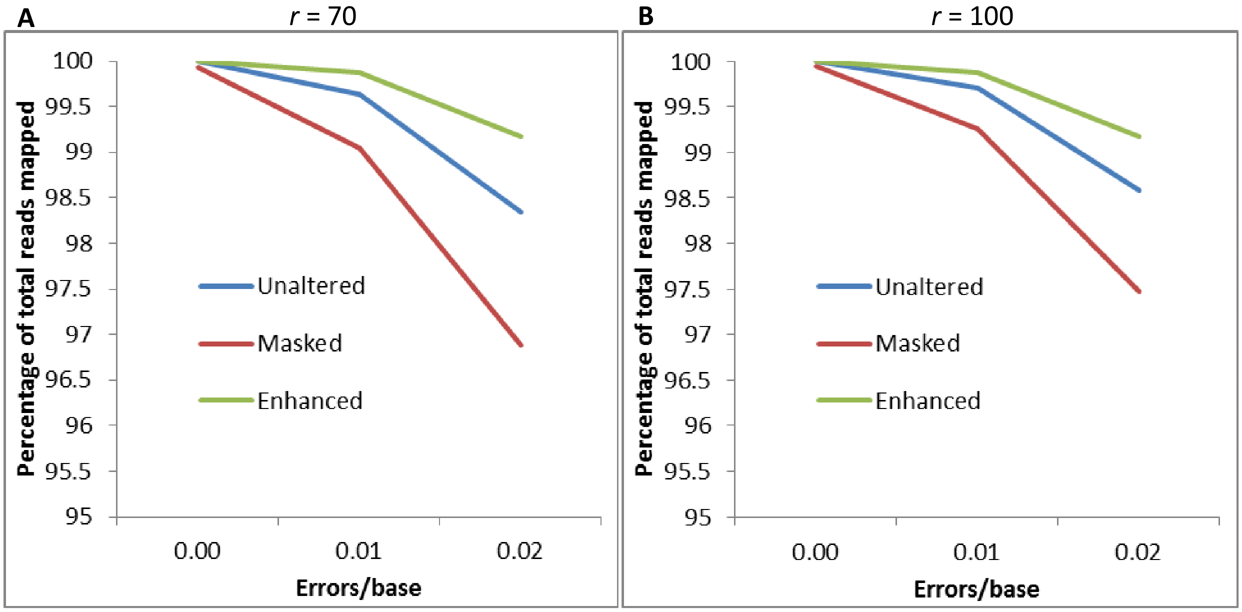


Fig. S3. Percentage of simulated reads that could be mapped using MAQ to each of the three references for: (**A**) read length 70 and (**B**) read length 100. The enhanced reference approach consistently outperforms the other two approaches.

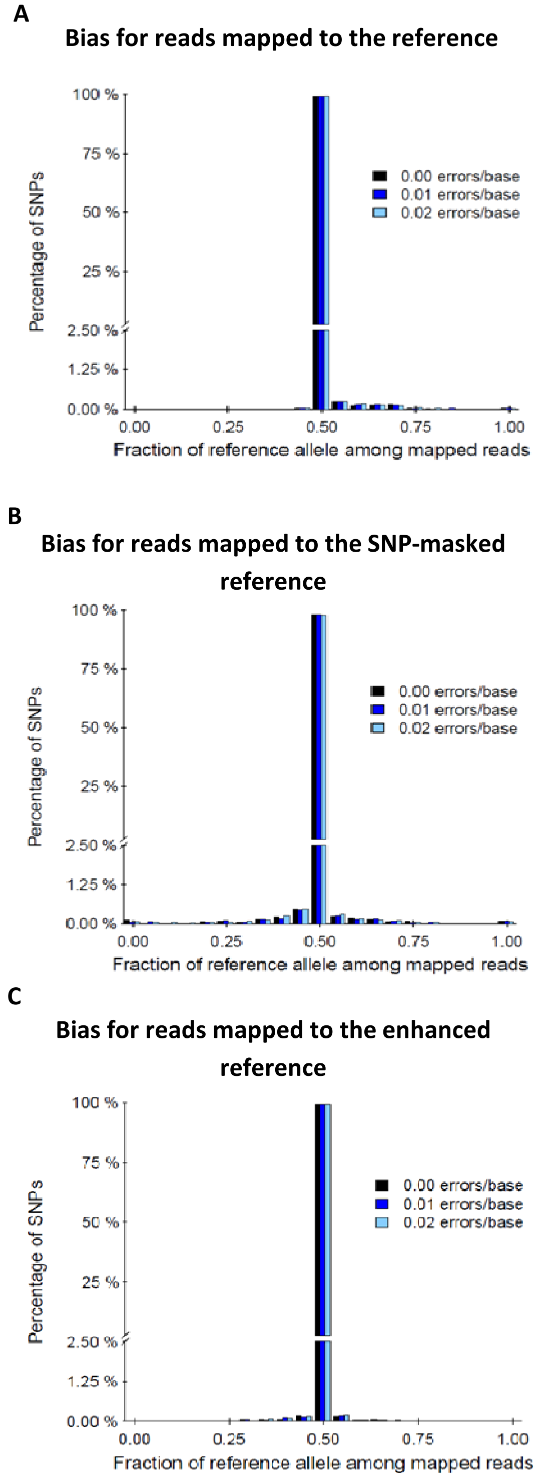


Fig. S4. Read-mapping biases for simulated 70-bp reads mapped using MAQ. (**A**) Mapping against the unaltered reference shows that there is significant bias towards the reference allele (*i.e.*, to the right). (**B**) Mapping against the masked reference results in many loci that are biased in both directions. (**C**) Mapping against the enhanced reference reduces the number of biased loci.

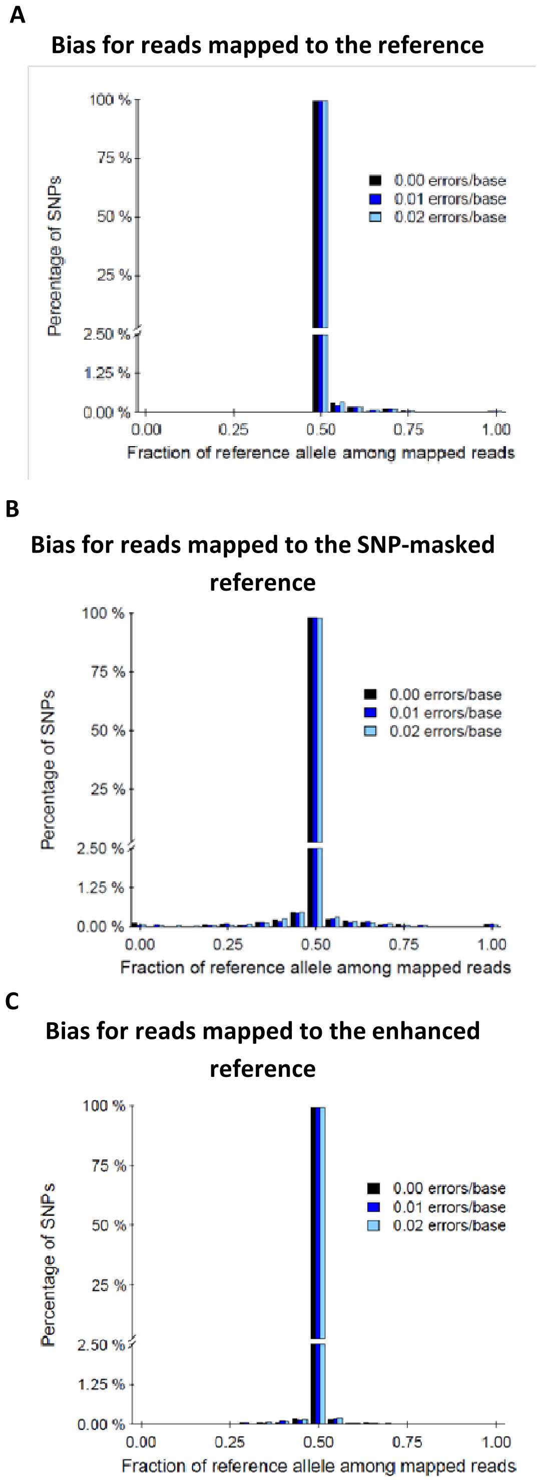


Fig. S5. Read-mapping biases for simulated 100-bp reads mapped using MAQ. (**A**) Mapping against the unaltered reference shows that there is significant bias towards the reference allele (*i.e.*, to the right). (**B**) Mapping against the masked reference results in many loci that are biased in both directions. (**C**) Mapping against the enhanced reference reduces the number of biased loci.

A

```
Query  1         CTGATTCTGGCCACCACCATCCCCATGCCTGCCGG  35
                 |||||||||||||||||||||||||||||||||||
Sbjct  16250569  CTGATTCTGGCCACCACCATCCCCATGCCTGCCGG  16250603
```

B

```
Query  1         CTGATTCTGGCCACCACCATCCCCATGCCTGCCGG  35
                 |||||||||||||||||||||||||||||||||||
Sbjct  16228822  CTGATTCTGGCCACCACCATCCCCATGCCTGCCGG  16228856
```

Fig. S6. An instance where a simulated error-free read has an exact match at two different places in the genome. The SNP rs6650119 (A/G) is at position 16250587 in chr1. (**A**) Alignment of the read to the intended location, with the SNP position shown in bold. (**B**) Alignment of the read to an alternate location. The read with the alternate allele ‘G’ does not have an exact match anywhere in the reference genome. A larger 70 bp read around the read with reference allele uniquely matches the intended location.

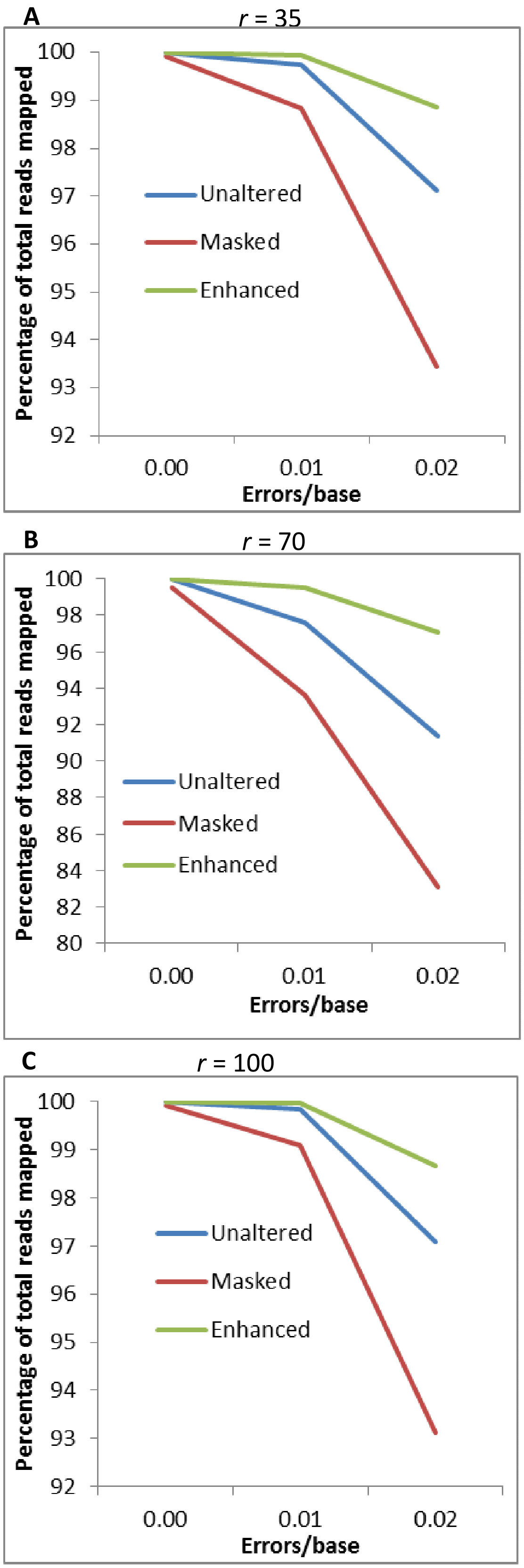


Fig. S7. Percentages of simulated reads that could be mapped using BWA for different read lengths.

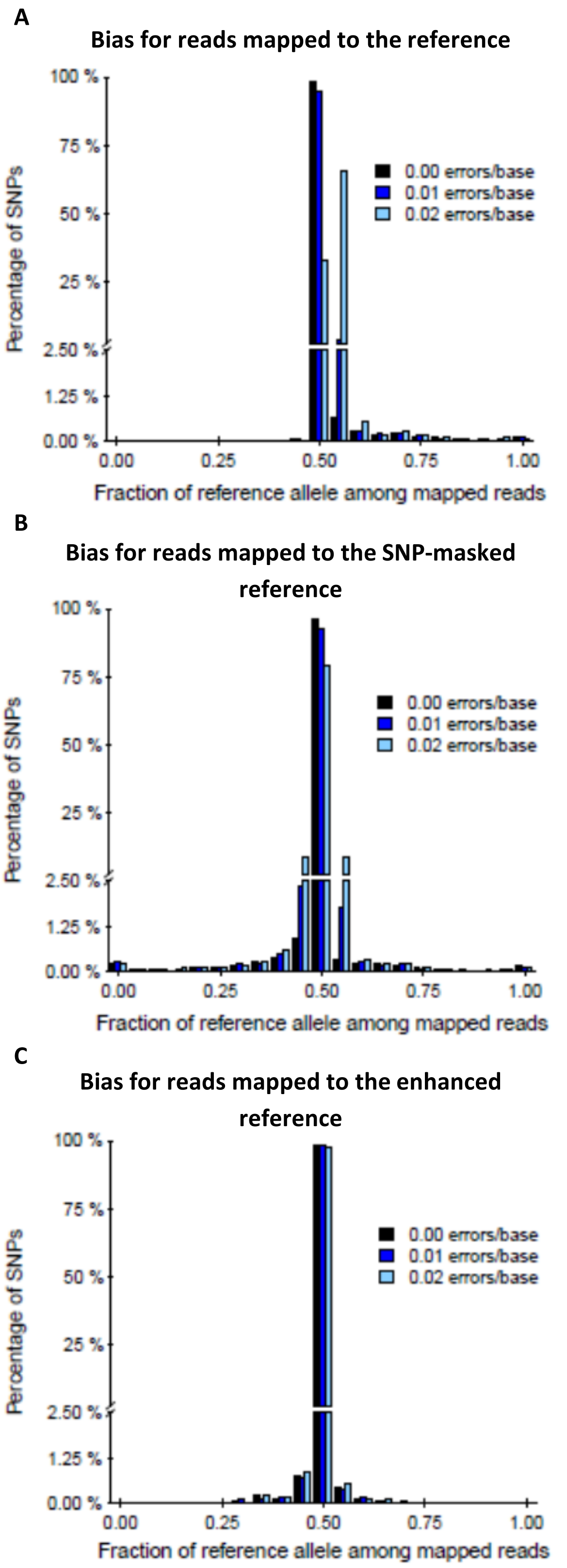


Fig. S8. Read-mapping biases for simulated 35-bp reads mapped using BWA.

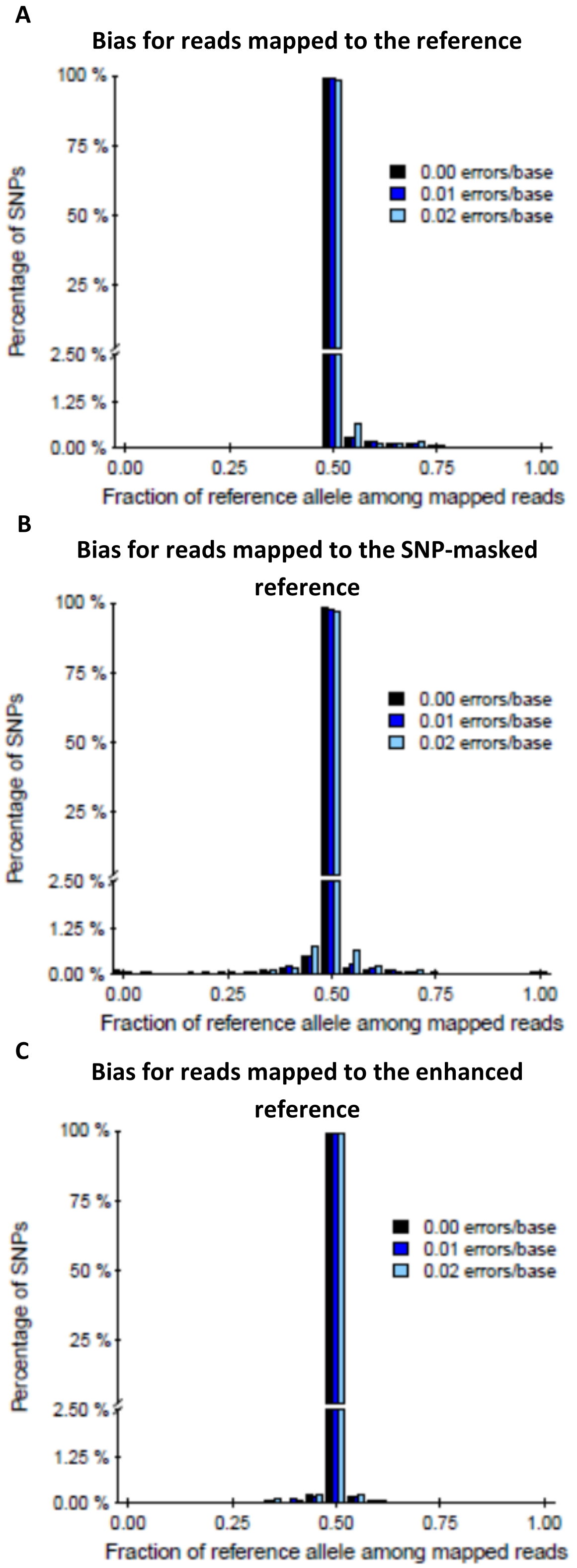


Fig. S9. Read-mapping biases for simulated 70-bp reads mapped using BWA.

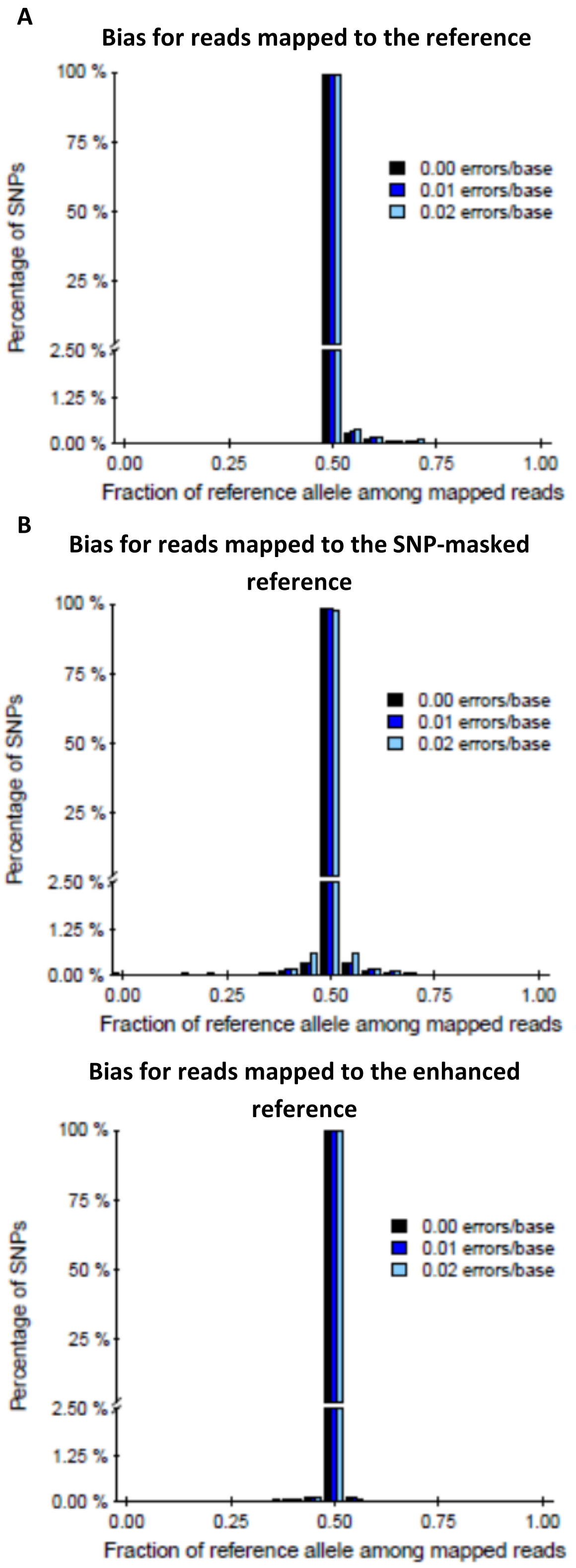


Fig. S10. Read-mapping biases for simulated 100-bp reads mapped using BWA.

Table S1. The numbers of biased loci (*i.e.*, loci showing a difference in the proportions of the mapped reads) when equal numbers of simulated reads from each allele were mapped using the three methods. The difference in the frequencies of the two alleles in the mapped reads is given by Δp. Δp ≥ 10% implies that ≥ 55% of the mapped reads carry one allele while ≤ 45% carry the other. Similarly, Δp ≥ 5% implies that ≥ 52.5% of the mapped reads carry one allele while ≤ 47.5% carry the other. The enhanced reference approach consistently outperforms the other two methods, keeping the numbers of biased loci low even at the higher error rates. The numbers from BWA mappings are very similar to the ones from MAQ mappings, except for slightly biased loci in case of longer reads with high error rates. These differences between the MAQ mappings and BWA mappings are likely due to the different settings used to run these programs. However, it is interesting to notice that these differences are minimal for the enhanced reference method, which shows that the enhanced reference method is able to handle these reads better than the other two methods.

MAQ	**Δp ≥ 10%**			**Δp ≥ 5%**			**Δp ≥ 2%**		
Errors/base	**0.00**	**0.01**	**0.02**	**0.00**	**0.01**	**0.02**	**0.00**	**0.01**	**0.02**
***r*=35**									
Unaltered	141	156	986	194	498	6312	268	5025	8950
Masked	319	364	467	396	669	2033	462	3392	5478
Enhanced	116	114	115	178	188	197	248	322	2162
***r*=70**									
Unaltered	72	71	74	89	91	91	124	128	779
Masked	159	161	161	204	197	206	250	296	730
Enhanced	40	41	39	63	64	65	110	106	116
***r*=100**									
Unaltered	49	49	52	67	64	72	86	85	268
Masked	103	106	105	144	145	142	177	193	337
Enhanced	24	26	26	41	43	40	63	73	77
BWA									
***r*=35**									
Unaltered	143	156	972	189	494	6304	255	5029	8948
Masked	320	368	472	396	669	2020	469	3394	5486
Enhanced	118	121	118	186	182	211	288	373	2215
***r*=70**									
Unaltered	71	78	78	87	91	126	117	128	3820
Masked	163	159	165	195	206	275	244	357	1759
Enhanced	41	41	41	67	64	70	115	116	179
***r*=100**									
Unaltered	49	48	57	64	70	80	82	86	3371
Masked	101	107	103	145	151	197	181	236	1370
Enhanced	23	26	25	40	42	39	71	71	88

2.2 Mapping statistics for simulated reads that are mapped with mapping quality >0

The enhanced reference approach is designed to ensure that any read uniquely matches either the reference sequence or one of the enhanced segments corresponding to the window in the reference that the read came from (provided there are no other regions in the reference that are highly similar to this window). However, by design, each enhanced segment is 1-mismatch away from either the reference or other enhanced segments. This guarantees that any read mapping to an enhanced segment or reference window corresponding to an enhanced segment with *d* mismatches will also have at least one other hit with *d*+1 mismatches. This inevitably reduces the mapping quality score of a read that maps to an enhanced segment or the corresponding window in the reference. In theory, the mapping quality, although reduced, should still be >0, as the read should still have a single best match. However, we find that the mapping quality computations vary significantly from program to program, and some programs might assign a mapping quality of zero even when the read has a single best match. Specifically, the mapping quality computed by MAQ is zero in many instances when the read has one perfect hit and one 1-mismatch hit, as in the following instance:

```
----------------------------------------------------------------------------------------------------------------------
rs3855952_chr1_77689_REF_-_A_0   16
     gi|89161185|ref|NC_000001.9|NC_000001  77590 0    100M  *    0
     0
     CCTGATGGCAGAGAAGCAAACACCAGTCGGAGAGCTGGGGTCCTCCCAGCCCTCTTGGCCCTGTG
GCCAATTTTTTCTTCAATAGCCTCATAAAATCACA
     BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB   MF:i:0    NM:i:0    UQ:i:0
     H0:i:1    H1:i:1
----------------------------------------------------------------
```

In the above alignment in SAM format, the underlined zero indicates the mapping quality. The entry “H0:i:1” indicates that the read has one perfect hit and “H1:i:1” indicates that the read has one 1-mismatch hit.

However, the same read is assigned a mapping quality score of 23 when mapped using BWA:

```
----------------------------------------------------------------
rs3855952_chr1_77689_REF_-_A_0   16
     gi|89161185|ref|NC_000001.9|NC_000001  77590 23   100M  =    77590
     0
     CCTGATGGCAGAGAAGCAAACACCAGTCGGAGAGCTGGGGTCCTCCCAGCCCTCTTGGCCCTGTG
GCCAATTTTTTCTTCAATAGCCTCATAAAATCACA
     BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB   XT:A:U    NM:i:0    X0:i:1
     X1:i:1    XM:i:0    XO:i:0    XG:i:0    MD:Z:100
----------------------------------------------------------------
```

Figures S11 through S16 show the read-mapping biases when reads with mapping quality of zero were eliminated. Table S2 shows the number of biased loci at different levels of bias. All three approaches resulted in more biased loci than when the mapping qualities were ignored (Table S1). The enhanced reference approach still outperformed the other two approaches, producing the least number of biased loci in most cases. Figures S11, S13, and S15 show that the mapping qualities computed by MAQ negatively affect the performance of the enhanced reference approach. Figures S12, S14, and S16 show that mapping qualities computed by BWA do not have such a negative effect.

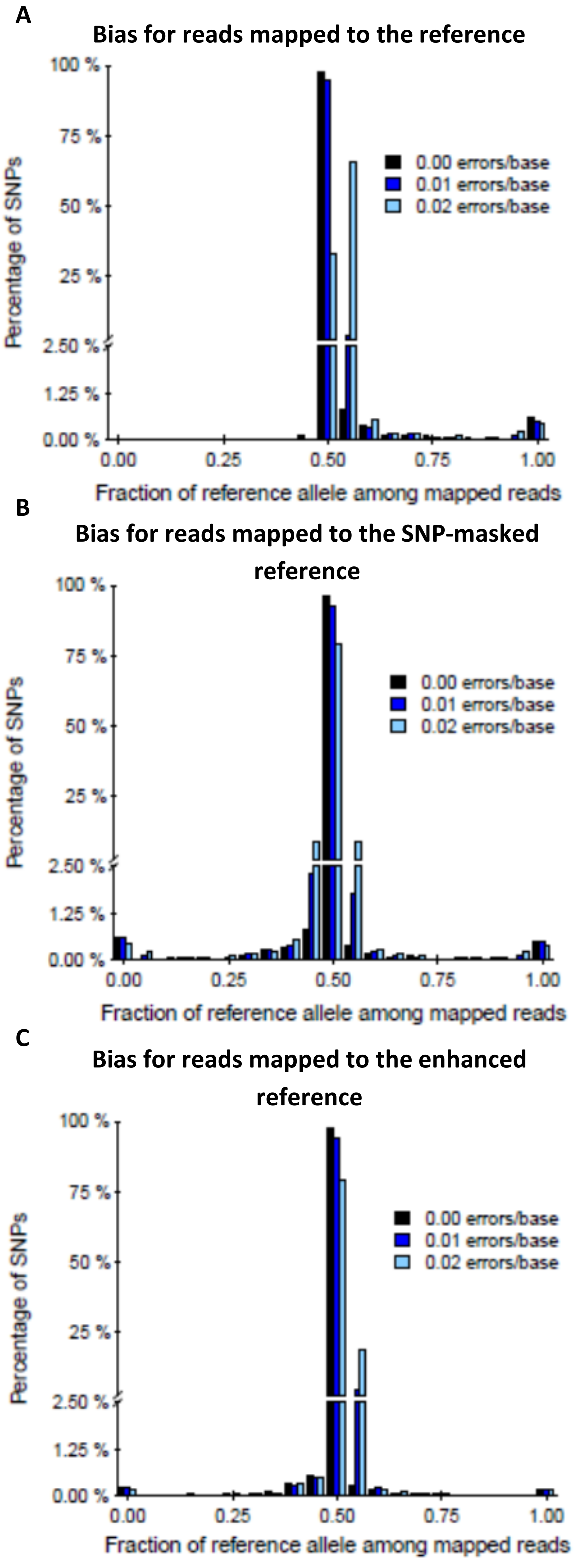


Fig. S11. Read-mapping biases for simulated 35-bp reads mapped using MAQ with mapping quality > 0.

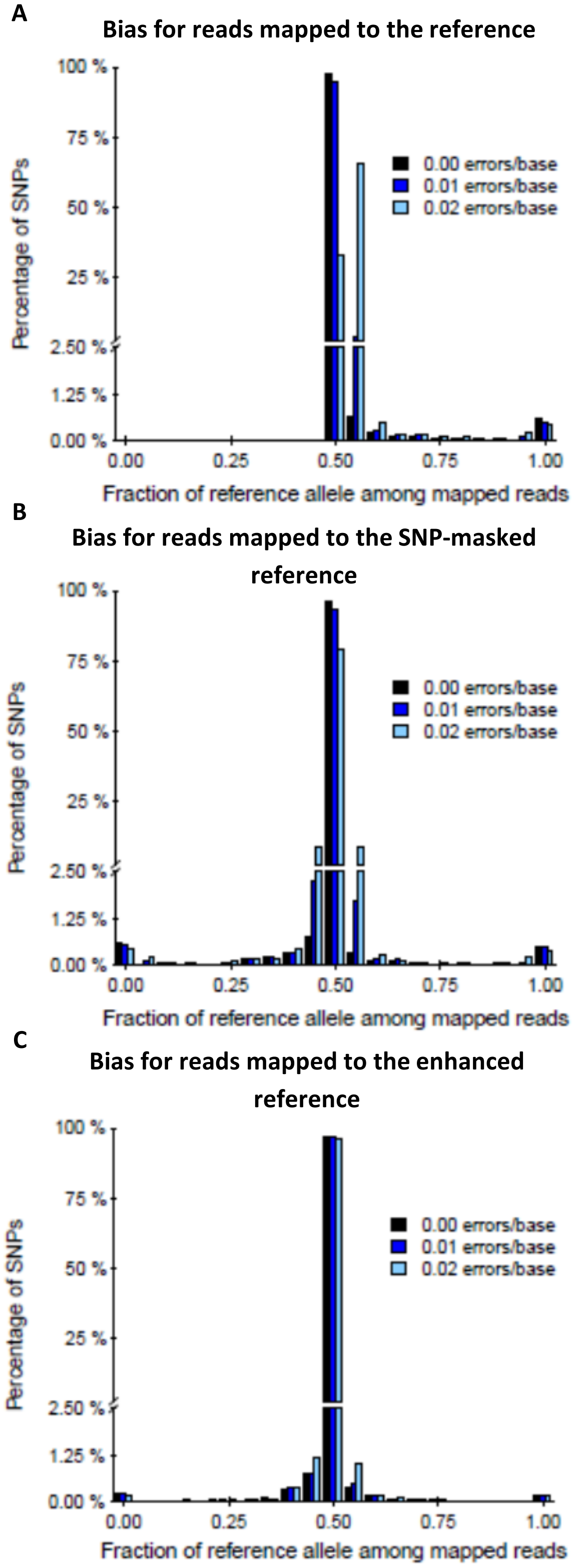


Fig. S12. Read-mapping biases for simulated 35-bp reads mapped using BWA with mapping quality > 0.

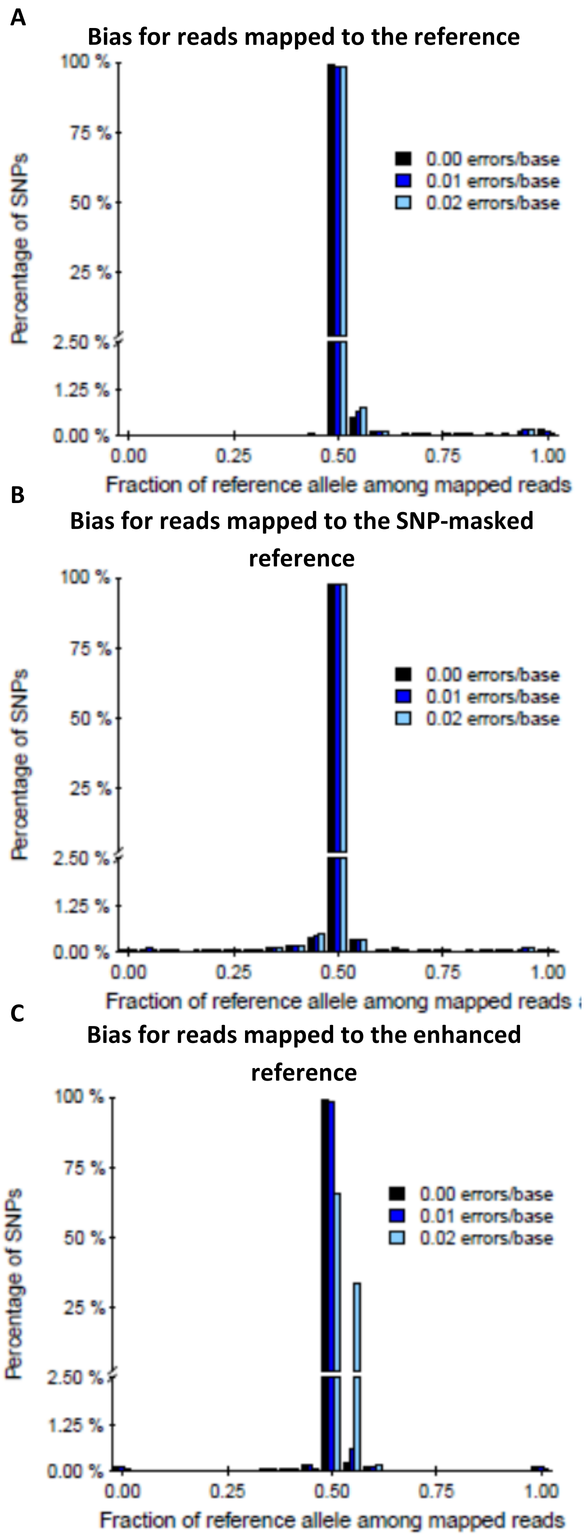


Fig. S13. Read-mapping biases for simulated 70-bp reads mapped using MAQ with mapping quality > 0.

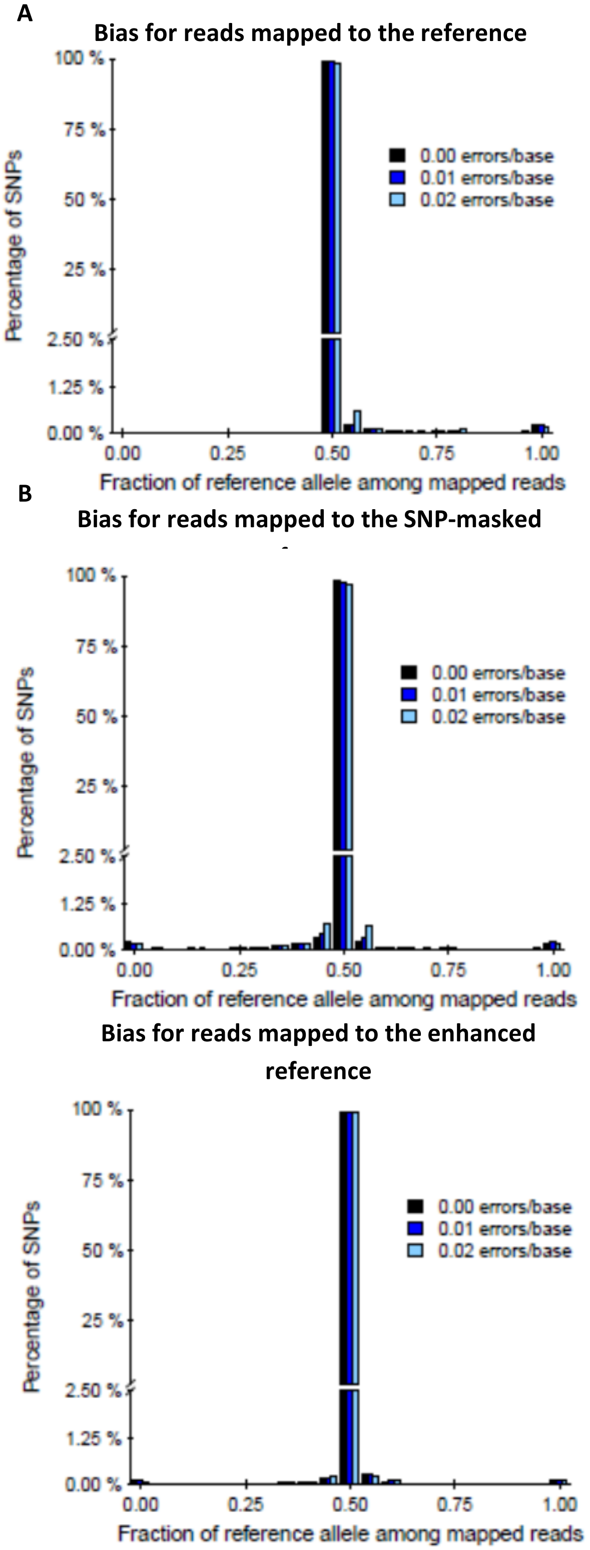
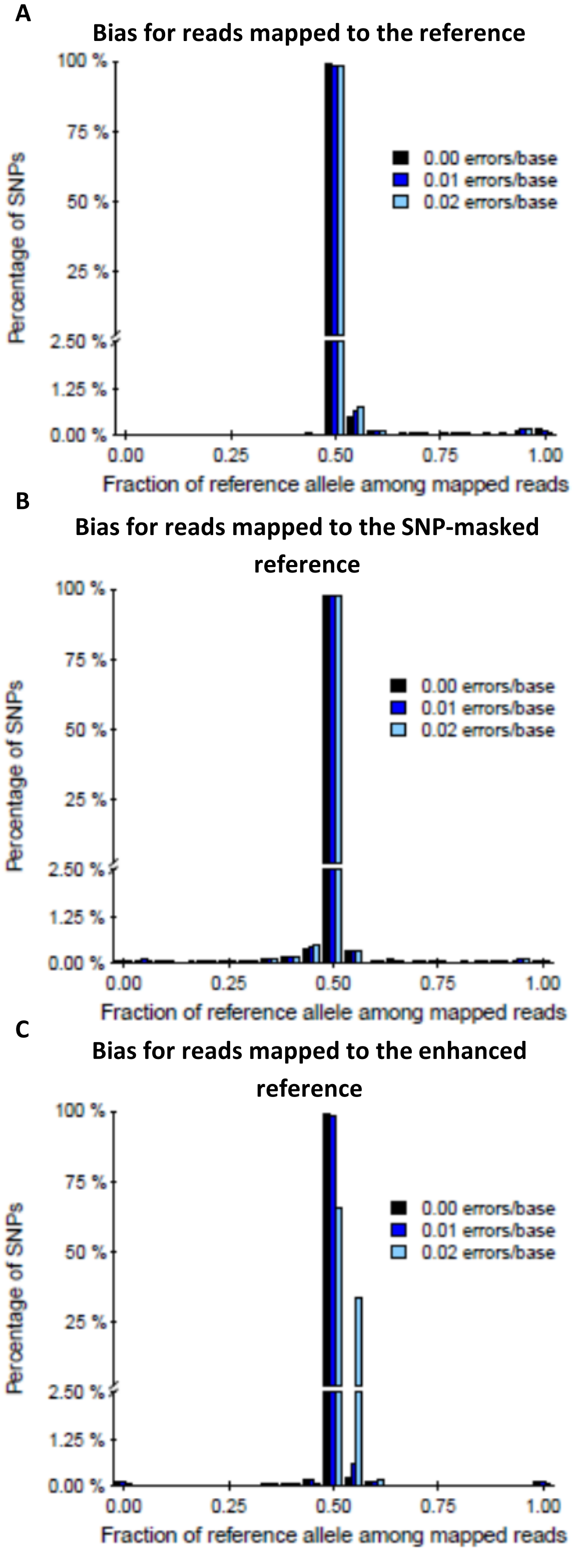


Fig. S14. Read-mapping biases for simulated 70-bp reads mapped using BWA with mapping quality > 0.



A

Bias for reads mapped to the reference

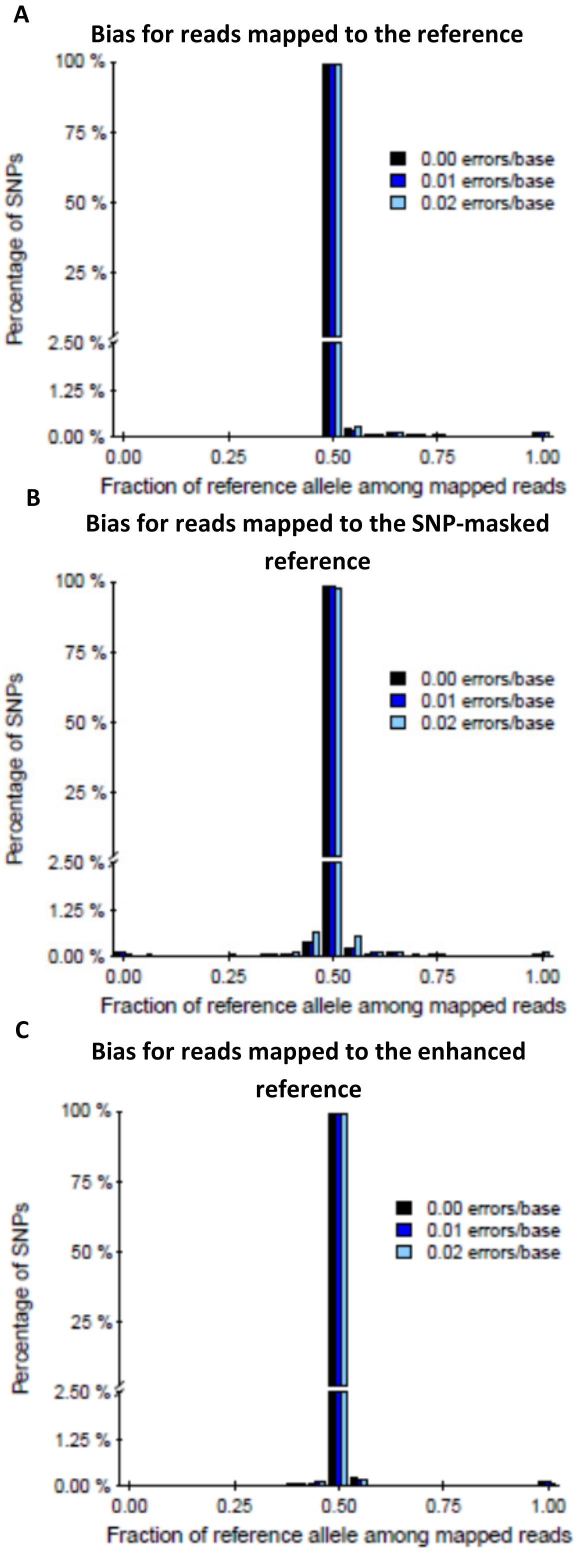
B

Bias for reads mapped to the SNP-masked reference

C

Bias for reads mapped to the enhanced reference

Fig. S15. Read-mapping biases for simulated 100-bp reads mapped using MAQ with mapping quality > 0.



A

Bias for reads mapped to the reference

B

Bias for reads mapped to the SNP-masked reference

C

Bias for reads mapped to the enhanced reference

Fig. S16. Read-mapping biases for simulated 100-bp reads mapped using BWA with mapping quality > 0.

Table S2. The numbers of biased loci (*i.e.*, loci showing a difference in the proportions of the mapped reads) when equal numbers of simulated reads from each allele were mapped using the three methods, and the reads with a mapping quality of zero were eliminated. The difference in the frequencies of the two alleles in the mapped reads is given by Δp. $\Delta p \geq 10\%$ implies that $\geq 55\%$ of the mapped reads carry one allele while $\leq 45\%$ carry the other. Similarly, $\Delta p \geq 5\%$ implies that $\geq 52.5\%$ of the mapped reads carry one allele while $\leq 47.5\%$ carry the other. The enhanced reference approach still outperformed the other two methods in most cases for the MAQ mappings and in almost all cases for the BWA mappings.

MAQ	**$\Delta p \geq 10\%$**			**$\Delta p \geq 5\%$**			**$\Delta p \geq 2\%$**		
Errors/base	**0.00**	**0.01**	**0.02**	**0.00**	**0.01**	**0.02**	**0.00**	**0.01**	**0.02**
***r*=35**									
Unaltered	233	234	1060	287	570	6332	364	5047	8972
Masked	404	419	520	476	733	2080	548	3450	5522
Enhanced	180	187	222	229	580	1935	280	4953	6494
***r*=70**									
Unaltered	101	105	106	147	160	167	235	243	898
Masked	207	193	187	253	241	240	338	380	835
Enhanced	75	74	101	104	141	3403	214	4065	8320
***r*=100**									
Unaltered	75	77	81	99	103	110	179	184	390
Masked	155	155	148	198	192	197	279	294	450
Enhanced	49	52	76	152	211	2182	1047	2378	7117
BWA									
***r*=35**									
Unaltered	206	219	1042	242	530	6336	263	5042	8969
Masked	390	411	515	458	715	2060	521	3442	5514
Enhanced	184	192	199	261	279	364	462	1013	2992
***r*=70**									
Unaltered	100	100	106	114	115	149	131	135	3719
Masked	195	186	181	229	235	291	269	383	1766
Enhanced	78	79	80	110	109	113	235	272	691
***r*=100**									
Unaltered	68	70	72	78	78	89	89	90	3376
Masked	135	131	129	170	167	218	191	241	1385
Enhanced	52	53	53	71	73	70	138	187	440

2.3 Frequency of high-SNP-density windows

A single *r*-window with *k* SNPs results in 2^k-1 enhanced segments generated for that window. Therefore, windows with a large number of SNPs generate too many enhanced segments. As this situation is somewhat undesirable, it is helpful to see how often this situation occurs in actual SNP datasets. Supplementary Tables S3 and S4 show that this situation occurs very rarely.

Table S3. The number of 100-bp windows with ≥*k* SNPs for different values of *k* in Chr1. Non-coding SNPs are also shown for comparison, even though it is not necessary to build enhanced segments for these non-coding SNPs as these regions are never expressed, unless they are part of some non-coding RNA. The number of windows with ≥*k* SNPs rapidly decreased with increasing *k*. There are only two 100-bp windows with ≥6 exonic/UTR SNPs. Even when non-coding SNPs were included, there were only 34 such windows in HapMap release 22, and 38 such windows in HapMap release 28.

***K* = no. SNPs within a 100-bp window**	**No. of windows with ≥*k* SNPs**		
	Exonic+UTR Yoruba SNPs in HapMap release 22	**All Yoruba SNPs in HapMap release 22**	**All SNPs in HapMap release 28**
1	9362	294968	326026
2	1651	55577	64194
3	238	7249	8908
4	36	937	1210
5	6	159	202
6	2	34	38
7	0	5	6
8	0	0	0

Table S4. The number of 100-bp windows with ≥*k* SNPs for different values of *k* in all 24 chromosomes. Non-coding SNPs are also shown for comparison. As in case of Chr1, the number of windows with ≥*k* SNPs rapidly decreased with increasing *k*. There are only 39 windows with ≥6 exonic/UTR SNPs. Even when non-coding SNPs were included, there were only 476 such windows in all chromosomes.

K = no. SNPs within a 100-bp window	***No. of windows with ≥k SNPs***	
	Exonic and UTR Yoruba HapMap release 22 SNPs	***All HapMap Yoruba release 22 SNPs***
1	94548	3788495
2	16772	730412
3	2463	97536
4	424	12230
5	103	1991
6	39	476
7	17	135
8	9	46
9	4	20
10	0	10

2.4 Mapping bias in high-SNP-density windows

We selected the 39 windows in Supplementary Table S4 with ≥6 exonic or UTR SNPs to evaluate mapping bias. These 39 windows contained a total of 134 SNPs (some of these windows overlap with each other). Evaluating mapping bias in these windows is complicated because there are multiple SNPs in each window. Therefore, it is not possible to determine which haplotypes should be used in generating the simulated reads. To allow for unbiased testing, we included all haplotypes in generation of the simulated reads, and generated one read from each 100-bp window from each strand of each possible haplotype. The exact number of haplotypes that any SNP is involved in depends on the number of nearby SNPs and exact distances between the nearby SNPs. Therefore, this procedure resulted in different numbers of simulated reads that overlap different SNPs. This scheme generated a total of 273008 reads, with a minimum of 2288 reads overlapping each SNP. At each SNP locus, equal numbers of reads contained the reference and the non-reference alleles. We mapped the reads using the MAQ program, and ignored mapping qualities. Figures S17 and S18 show the percentage of mapped reads and the mapping bias for the different reference construction methods.

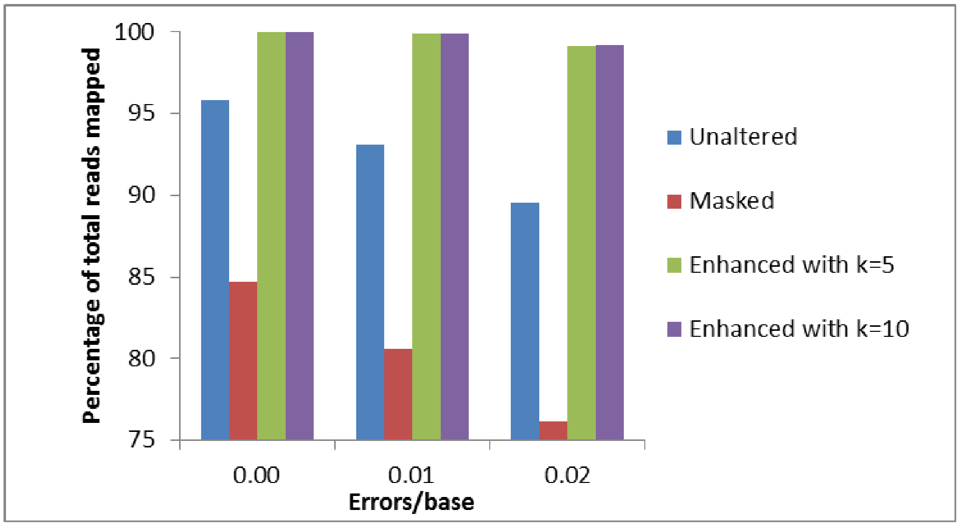


Fig. S17. Percentages of simulated reads that could be mapped to 100-bp windows with ≥6 exonic or UTR SNPs. A significant percentage of the reads cannot be mapped to the unaltered reference and masked reference approaches, and these mappings worsened as the error rate increased. However, the enhanced reference approach was able to map >99% of the reads, even at higher error rates. The limit on the maximum number of SNPs in a 100-bp window (*k*) has almost no impact on the performance of the enhanced reference approach. Setting *k* =5 affected only 0.05% of the reads as compared with setting *k*=10.

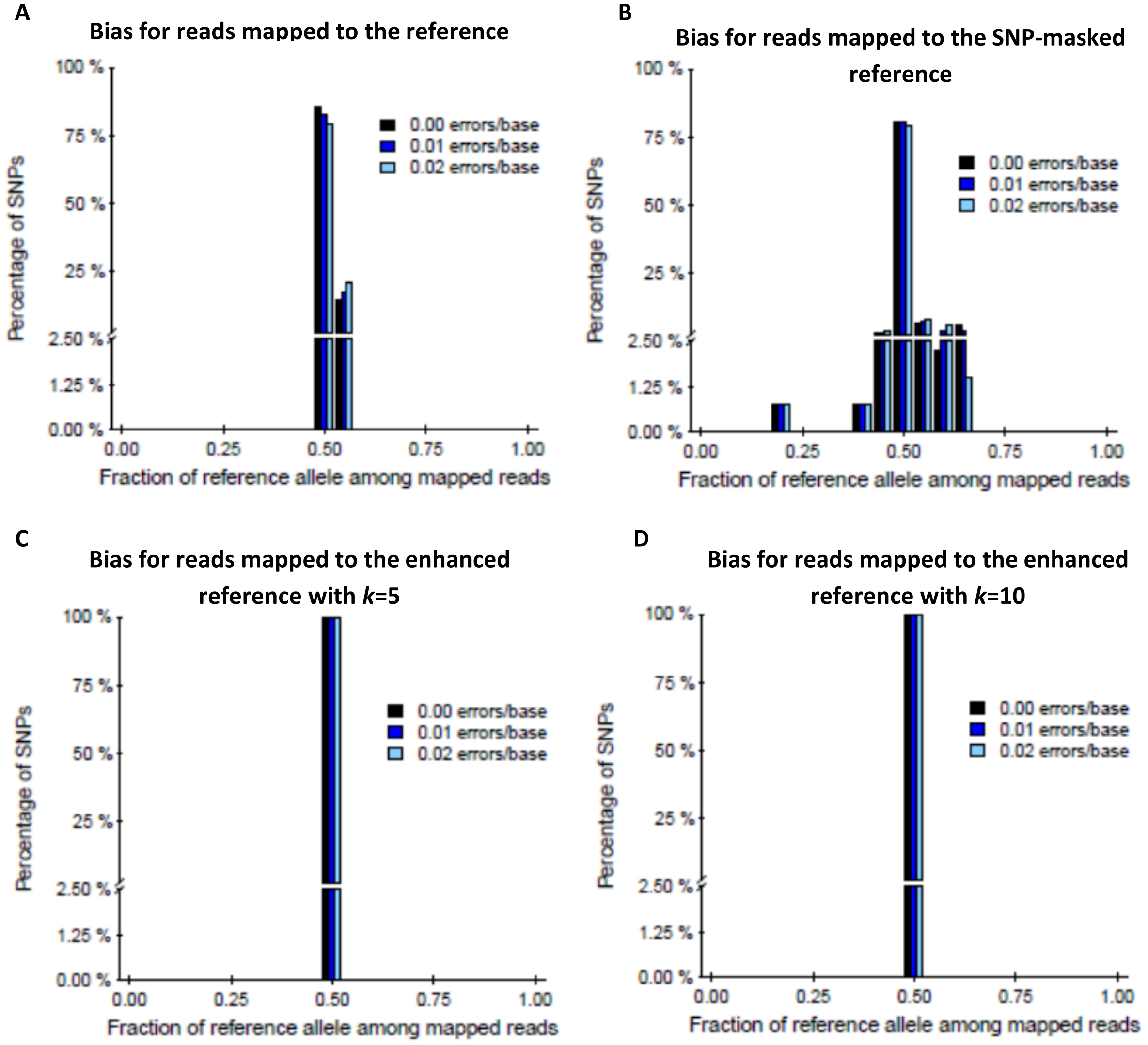


Fig. S18. Read-mapping biases for simulated reads at loci with ≥6 exonic and UTR SNPs within 100 bp. A significant proportion of the loci were biased in both the unaltered reference and masked reference approaches. When the enhanced reference was constructed using a limit of 5 SNPs within 100 bp, there was no significant bias at any error rate. Similarly, when the enhanced reference was constructed with a limit of 10 SNPs within 100 bp, there was no bias even at high error rates.

3. Results on RNA-Seq data

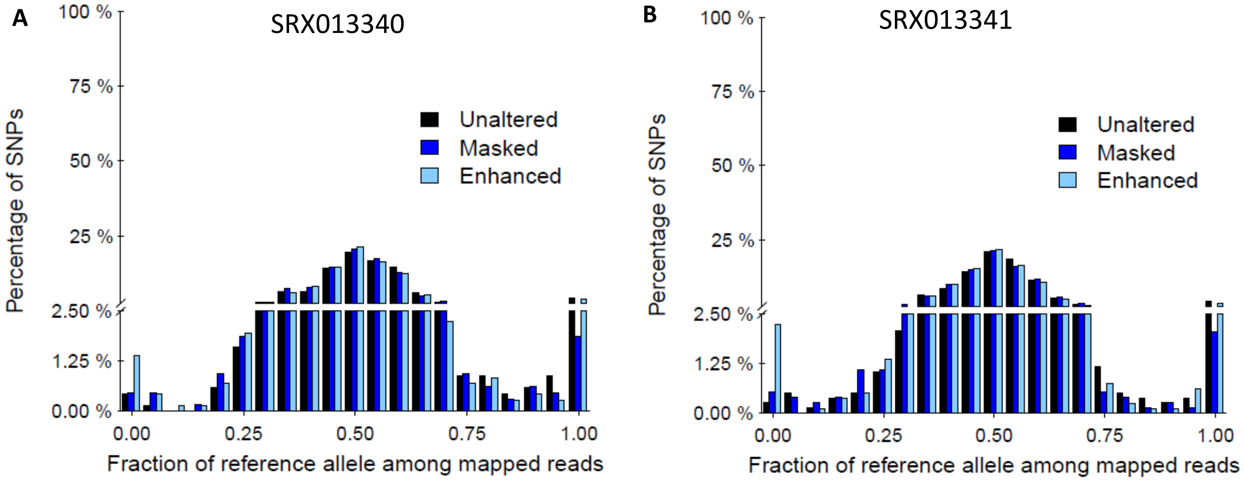


Fig. S19. Histograms of ASE for the two individuals in the actual RNA-Seq data. The histograms cover exonic and UTR loci with ≥20 mapped reads. As would be expected in any actual data set, some loci do exhibit ASE. However, only a small fraction of these loci (listed in Tables S5 and S6) show statistically significant ASE. Most of the loci with significant ASE come from the tails in the histograms. The unaltered reference and masked reference approaches show much larger number of loci specific to the reference allele than to the non-reference allele, indicating a bias toward the reference allele. These differences are much smaller for the enhanced reference approach.

Table S5. List of loci with significant allele-specific expression in individual GM19238 at a false discovery rate (FDR) of 1%. Reads were mapped to the enhanced reference. The column labeled 'other reads' mainly consists of sequencing errors or mapping errors.

SNP ID	Chromosome	Position (hg 18)	Ref Reads	Non-Ref Reads	Other Reads	Gene
rs1158	chr10	22865294	23	5	0	PIP4K2A
rs2574943	chr10	51695173	0	23	0	LOC728532
rs7914886	chr10	97345055	45	1	1	LOC643981
rs2848622	chr11	57100600	305	1	0	LOC390183
rs7309149	chr12	12895367	73	0	0	RPS6P1
rs3759294	chr12	31835958	61	0	0	LOC440093
rs3888051	chr12	62503257	87	0	0	LOC341315
rs916974	chr12	111931395	42	18	2	OAS2
rs11619791	chr13	24568984	4	140	0	PABPC3
rs11160859	chr14	105181102	109	4	2	IGHG2
rs2731202	chr14	106106026	44	0	0	IGHV5-51
rs12441946	chr15	70978916	1	35	0	LOC729686
rs4784800	chr16	55957003	236	86	2	CCL22
rs7214234	chr17	1708330	25	0	0	LOC642502
rs4791596	chr17	14549410	988	3	13	LOC388339
rs525911	chr17	38065445	2	19	0	TUBG2
rs1139405	chr17	77092614	1027	687	6	ACTG1
rs12610462	chr19	22343302	0	259	0	LOC342994
rs7417535	chr1	16006574	1	33	1	LOC440567
rs6677535	chr1	40371584	360	0	0	LOC728602
rs17459	chr1	74944339	28	4	1	CRYZ
rs3819946	chr1	74948474	34	9	0	CRYZ

rs273259	chr1	78866406	10	31	1	IFI44L
rs2794041	chr1	143823777	39	0	0	SEC22B
rs6682136	chr1	148861639	59	5	1	ENSA
rs2223477	chr1	169575456	0	30	0	TOP1P1
rs7513402	chr1	234714134	3	242	1	EDARADD
rs1807676	chr22	25613521	0	101	0	LOC100130624
rs7578292	chr2	136673203	34	3	0	LOC389053
rs17044594	chr3	18556112	58	0	0	LOC131185
rs6792846	chr3	134790181	23	5	0	CDV3
rs16858473	chr3	185350316	99	0	3	LOC440991
rs10022054	chr4	43595905	0	87	1	LOC402175
rs17008180	chr4	106625532	200	1	1	EEF1AL7
rs7662486	chr4	113928795	53	0	0	LOC441034
rs17008716	chr4	165338000	52	0	3	LOC646954
rs6875717	chr5	92629704	0	20	0	LOC391811
rs2009646	chr5	108148857	1	365	1	LOC643534
rs1061837	chr5	150541530	22	6	0	CCDC69
rs11953084	chr5	177415300	202	1	12	LOC653314
rs11953029	chr5	177415790	329	3	6	LOC653314
rs2734945	chr6	29963924	169	2	0	HLA-H
rs9461576	chr6	30337105	31	3	3	FLJ45422
rs1058026	chr6	31429664	410	321	3	HLA-B
rs8084	chr6	32519013	396	562	2	HLA-DRA
rs7192	chr6	32519624	217	329	0	HLA-DRA
rs701831	chr6	32657379	106	366	1	HLA-DRB1
rs9273655	chr6	32737187	181	2	2	HLA-DQB1
rs7739387	chr6	34730399	3	17	0	LOC100129061
rs2814966	chr6	34820210	130	0	0	LOC646785
rs13319	chr6	107124347	7	26	1	AIM1
rs2430028	chr7	122108911	24	0	0	LOC645979
rs11986385	chr8	30094837	470	0	1	LOC648729
rs10106469	chr8	30330002	62	3	0	LOC100131210
rs20583	chr9	33016572	26	8	0	DNAJA1
rs9410092	chr9	140190645	0	21	0	LOC643224
rs7205	chrX	1468324	122	186	2	SLC25A6
rs16981209	chrX	19464216	28	4	0	SH3KBP1
rs909713	chrX	47547916	0	20	0	WASF4
rs6624600	chrX	71296606	69	1	0	FLJ44635
rs3088376	chrX	118647434	97	23	1	SEPT6

Table S6. List of loci with significant allele-specific expression in individual GM19239 at an FDR of 1%. Reads were mapped to the enhanced reference. The column labeled ‘other reads’ mainly consists of sequencing errors or mapping errors.

SNP ID	**Chromosome**	**Position**	**Ref Reads**	**Non-Ref**	**Other Reads**	**Gene**
		(hg 18)		**Reads**		
rs2228064	chr10	45198056	4	19	1	ALOX5
rs2574943	chr10	51695173	0	35	0	LOC728532
rs1049455	chr10	103910465	19	4	0	NOLC1
rs2089910	chr11	1830980	223	0	1	LSP1
rs2071461	chr11	11330536	0	32	0	CSNK2A1P
rs2848622	chr11	57100600	206	1	0	LOC390183
rs6591717	chr11	62091484	315	236	3	EEF1G
rs7927338	chr11	93561259	20	0	0	LOC729494
rs11831812	chr12	4077714	0	32	0	LOC399988
rs4765775	chr12	4284387	46	19	0	CCND2
rs3759294	chr12	31835958	38	1	0	LOC440093
rs14105	chr13	27137970	17	48	0	POLR1D

rs2731202	chr14	106106026	610	17	4	IGHV5-51
rs2589529	chr15	33317322	1	60	0	LOC723972
rs12441946	chr15	70978916	0	26	1	LOC729686
rs7214234	chr17	1708330	36	0	0	LOC642502
rs4791596	chr17	14549410	706	1	5	LOC388339
rs4792757	chr17	16461158	22	0	0	LOC644422
rs199455	chr17	42154400	1	168	0	LOC644315
rs16952692	chr18	46693267	27	0	0	ME2
rs10250	chr19	4052062	14	37	1	MAP2K2
rs17627	chr19	44615792	469	630	5	RPS16
rs3170545	chr19	55128527	9	40	0	ATF5
rs8647	chr19	55128792	10	81	1	ATF5
rs6677535	chr1	40371584	393	0	0	LOC728602
rs13306758	chr1	43165406	31	0	0	SLC2A1
rs631045	chr1	78330102	0	35	1	LOC729768
rs630245	chr1	78330252	29	0	0	LOC729768
rs4000303	chr1	96686353	0	43	3	LOC440595
rs3871984	chr1	143823858	49	0	0	SEC22B
rs4950386	chr1	145142175	0	26	0	LOC644131
rs7546434	chr1	145142963	0	26	0	LOC644131
rs7695	chr1	154413950	11	44	0	SEMA4A
rs2488896	chr1	164513043	0	47	0	LOC284685
rs2223477	chr1	169575456	1	52	0	TOP1P1
rs6570	chr21	45130540	51	22	0	ITGB2
rs1807676	chr22	25613521	0	101	0	LOC100130624
rs8177832	chr22	37807512	4	24	0	APOBEC3G
rs7599670	chr2	41900536	0	21	0	LDHAL3
rs1997	chr2	42431310	24	2	0	COX7A2L
rs17014852	chr2	127537677	48	0	0	BIN1
rs7578292	chr2	136673203	35	1	0	LOC389053
rs10031608	chr4	12948262	48	0	0	HSP90AB2P
rs7662486	chr4	113928795	64	0	0	LOC441034
rs7662013	chr4	174791881	0	171	0	LOC100128266
rs7404	chr5	133335760	90	45	2	VDAC1
rs11953084	chr5	177415300	219	2	20	LOC653314
rs11953029	chr5	177415790	198	4	2	LOC653314
rs1736924	chr6	29800990	0	124	0	HLA-F
rs2734945	chr6	29963924	76	2	0	HLA-H
rs2428512	chr6	29964309	200	2	1	HLA-H
rs1051336	chr6	32520570	448	293	2	HLA-DRA
rs9276436	chr6	32822061	127	1	0	HLA-DQA2
rs762815	chr6	32837620	0	50	1	HLA-DQB2
rs2071888	chr6	33380833	56	26	0	TAPBP
rs2814966	chr6	34820210	103	0	0	LOC646785
rs10947623	chr6	36750814	40	0	0	LOC389386
rs13296	chr6	44326098	231	162	0	HSP90AB1
rs1059307	chr6	86444607	15	92	0	SNORD50B
rs10266655	chr7	24705083	33	12	0	DFNA5
rs3087615	chr7	102739359	26	7	1	PMPCB
rs4726719	chr7	143975161	575	0	2	LOC100132804
rs11986385	chr8	30094837	219	0	0	LOC648729
rs2719323	chr8	34300101	76	3	0	CYCSP3
rs9410092	chr9	140190645	1	45	0	LOC643224
rs7205	chrX	1468324	94	186	2	SLC25A6

Table S7. Comparison of the three mapping methods at each locus that was significant at 1% FDR in at least one out of the three methods for individual GM19238. In most loci, the enhanced reference approach was able to map the largest number of reference and non-reference reads.

			Unaltered			Masked			Enhanced						
SNP ID	**Chromosome**	**Position (hg 18)**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Gene**	**Unaltered Significant**	**Masked Significant**	**Enhanced Significant**
rs7662486	chr4	113928795	42	0	0	8	0	1	53	0	0	LOC441034	Yes	No	Yes
rs10106469	chr8	30330002	51	1	0	14	0	2	62	3	0	LOC100131210	Yes	No	Yes
rs16981209	chrX	19464216	28	4	0	28	4	0	28	4	0	SH3KBP1	Yes	Yes	Yes
rs16978523	chr18	41929062	46	52	1	14	51	1	52	68	1	TRNAK-CUU	No	Yes	No
rs10022054	chr4	43595905	0	46	1	0	53	1	0	87	1	LOC402175	Yes	Yes	Yes
rs2073687	chr11	8663809	208	12	0	15	11	0	200	165	0	SNORA45	Yes	No	No
rs1807676	chr22	25613521	0	0	0	0	0	0	0	101	0	LOC100130624	No	No	Yes
rs4784800	chr16	55957003	236	84	2	235	84	2	236	86	2	CCL22	Yes	Yes	Yes
rs14408	chr11	298314	36	0	0	10	1	0	50	45	0	IFITM2	Yes	No	No
rs11619791	chr13	24568984	5	3	0	2	1	0	4	140	0	PABPC3	No	No	Yes
rs1053492	chr15	41849094	28	35	2	1	22	1	26	47	1	PDIA3	No	Yes	No
rs11160859	chr14	105181102	92	4	2	47	4	2	109	4	2	IGHG2	Yes	Yes	Yes
rs12610462	chr19	22343302	0	13	2	0	8	3	0	259	0	LOC342994	No	No	Yes
rs8084	chr6	32519013	396	556	2	393	556	2	396	562	2	HLA-DRA	Yes	Yes	Yes
rs3088376	chrX	118647434	97	22	1	97	22	1	97	23	1	SEPT6	Yes	Yes	Yes
rs1139405	chr17	77092614	1003	667	6	776	669	5	1027	687	6	ACTG1	Yes	No	Yes
rs17008180	chr4	106625532	218	0	3	2	0	3	200	1	1	EEF1AL7	Yes	No	Yes
rs1792624	chr11	93561458	21	0	0	0	0	0	14	0	0	LOC729494	Yes	No	No
rs9461576	chr6	30337105	21	2	1	17	1	1	31	3	3	FLJ45422	Yes	No	Yes
rs11953029	chr5	177415790	324	1	6	1	1	9	329	3	6	LOC653314	Yes	No	Yes
rs6677535	chr1	40371584	357	0	0	233	0	0	360	0	0	LOC728602	Yes	Yes	Yes
rs13319	chr6	107124347	7	26	1	7	26	1	7	26	1	AIM1	No	Yes	Yes
rs2848622	chr11	57100600	302	1	2	10	1	2	305	1	0	LOC390183	Yes	No	Yes
rs7739387	chr6	34730399	3	16	0	3	16	0	3	17	0	LOC100129061	No	No	Yes
rs7914886	chr10	97345055	54	0	1	1	0	2	45	1	1	LOC643981	Yes	No	Yes
rs1061837	chr5	150541530	22	6	0	21	6	0	22	6	0	CCDC69	Yes	No	Yes
rs9276976	chr6	33081772	17	4	0	17	4	0	17	4	0	HLA-DOA	Yes	No	No
rs7359861	chr19	40834027	94	37	2	76	40	2	94	98	2	COX6B1	Yes	Yes	No
rs701831	chr6	32657379	101	20	0	0	12	1	106	366	1	HLA-DRB1	Yes	No	Yes
rs7513402	chr1	234714134	3	78	1	3	78	1	3	242	1	EDARADD	Yes	Yes	Yes
rs1042448	chr6	33162320	96	4	0	83	4	0	96	106	0	RPL32P1	Yes	Yes	No
rs3819946	chr1	74948474	34	9	0	31	9	0	34	9	0	CRYZ	Yes	Yes	Yes
rs6624600	chrX	71296606	75	0	0	76	1	0	69	1	0	FLJ44635	Yes	Yes	Yes
rs7192	chr6	32519624	217	325	0	214	325	0	217	329	0	HLA-DRA	Yes	Yes	Yes
rs2574943	chr10	51695173	0	20	0	0	21	0	0	23	0	LOC728532	Yes	Yes	Yes

rs1042665	chr5	137930238	33	30	0	13	38	0	34	50	0	SNORD63	No	Yes	No
rs2794041	chr1	143823777	39	0	0	11	0	0	39	0	0	SEC22B	Yes	No	Yes
rs2223477	chr1	169575456	0	11	0	0	10	0	0	30	0	TOP1P1	No	No	Yes
rs12441946	chr15	70978916	1	0	0	0	1	0	1	35	0	LOC729686	No	No	Yes
rs11986385	chr8	30094837	477	0	0	26	0	1	470	0	1	LOC648729	Yes	Yes	Yes
rs1803621	chr12	6517370	2694	1690	18	2269	1695	21	2705	2541	18	LOC100133042	Yes	Yes	No
rs9509472	chr13	20433586	29	3	0	1	1	2	37	21	0	LOC440125	Yes	No	No
rs7768	chr10	120917783	16	37	0	8	37	0	16	38	0	PRDX3	No	Yes	No
rs6792846	chr3	134790181	23	1	0	14	1	0	23	5	0	CDV3	Yes	No	Yes
rs525911	chr17	38065445	2	1	0	1	1	0	2	19	0	TUBG2	No	No	Yes
rs7309149	chr12	12895367	53	1	1	31	1	1	73	0	0	RPS6P1	Yes	Yes	Yes
rs1051470	chr12	117067615	44	20	0	40	19	0	43	32	0	PEBP1	Yes	No	No
rs4791596	chr17	14549410	991	1	10	6	1	9	988	3	13	LOC388339	Yes	No	Yes
rs1736924	chr6	29800990	89	30	0	80	28	1	90	89	0	HLA-F	Yes	Yes	No
rs7214234	chr17	1708330	23	0	0	11	0	0	25	0	0	LOC642502	Yes	No	Yes
rs6875717	chr5	92629704	0	2	0	0	2	0	0	20	0	LOC391811	No	No	Yes
rs273259	chr1	78866406	10	30	1	10	30	1	10	31	1	IFI44L	No	No	Yes
rs2731202	chr14	106106026	44	0	0	38	0	0	44	0	0	IGHV5-51	Yes	Yes	Yes
rs9273655	chr6	32737187	182	2	3	164	2	3	181	2	2	HLA-DQB1	Yes	Yes	Yes
rs1049230	chr19	6702281	22	5	2	20	5	2	22	10	2	TRIP10	Yes	No	No
rs6682136	chr1	148861639	59	5	1	58	5	1	59	5	1	ENSA	Yes	Yes	Yes
rs909713	chrX	47547916	0	12	0	0	13	0	0	20	0	WASF4	No	No	Yes
rs916974	chr12	111931395	42	17	2	38	15	2	42	18	2	OAS2	Yes	No	Yes
rs17626	chr19	44618361	413	293	16	274	292	16	414	402	15	RPS16	Yes	No	No
rs20583	chr9	33016572	26	2	0	20	2	1	26	8	0	DNAJA1	Yes	Yes	Yes
rs3759294	chr12	31835958	67	0	0	0	0	0	61	0	0	LOC440093	Yes	No	Yes
rs1158	chr10	22865294	23	2	0	23	2	0	23	5	0	PIP4K2A	Yes	Yes	Yes
rs7578292	chr2	136673203	30	1	0	0	0	0	34	3	0	LOC389053	Yes	No	Yes
rs17044594	chr3	18556112	60	0	0	49	0	0	58	0	0	LOC131185	Yes	Yes	Yes
rs9410092	chr9	140190645	0	0	0	0	0	0	0	21	0	LOC643224	No	No	Yes
rs2734945	chr6	29963924	169	1	0	27	1	0	169	2	0	HLA-H	Yes	Yes	Yes
rs7404	chr5	133335760	64	34	0	42	33	0	65	43	0	VDAC1	Yes	No	No
rs17008716	chr4	165338000	47	0	3	8	0	3	52	0	3	LOC646954	Yes	No	Yes
rs3888051	chr12	62503257	81	0	0	5	0	1	87	0	0	LOC341315	Yes	No	Yes
rs6568	chr2	127170700	32	10	0	30	10	0	32	12	0	LOC100130248	Yes	Yes	No
rs17459	chr1	74944339	28	4	1	28	4	1	28	4	1	CRYZ	Yes	Yes	Yes
rs1058026	chr6	31429664	399	302	2	227	292	3	410	321	3	HLA-B	Yes	No	Yes
rs2009646	chr5	108148857	1	173	1	0	174	1	1	365	1	LOC643534	Yes	Yes	Yes
rs7205	chrX	1468324	111	81	2	4	98	2	122	186	2	SLC25A6	No	Yes	Yes
rs2814966	chr6	34820210	121	0	0	6	0	0	130	0	0	LOC646785	Yes	No	Yes
rs2230659	chr1	45851471	36	9	0	34	11	0	36	17	0	NASP	Yes	Yes	No
rs1136853	chr11	310805	10	29	0	6	29	0	10	30	0	IFITM3	No	Yes	No

rs11953084	chr5	177415300	187	0	9	116	0	10	202	1	12	LOC653314	Yes	Yes	Yes
rs2430028	chr7	122108911	14	0	0	2	0	0	24	0	0	LOC645979	No	No	Yes
rs16858473	chr3	185350316	104	0	2	0	0	3	99	0	3	LOC440991	Yes	No	Yes
rs7417535	chr1	16006574	0	14	1	1	10	1	1	33	1	LOC440567	No	No	Yes

Table S8. Comparison of the three mapping methods at each locus that was significant at 1% FDR in at least one out of the three methods for individual GM19239. In most loci, the enhanced reference approach was able to map the largest number of reference and non-reference reads.

			Unaltered			**Masked**			**Enhanced**						
SNP ID	**Chromosome**	**Position (hg 18)**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Ref Reads**	**Non-Ref Reads**	**Other Reads**	**Gene**	**Unaltered Significant**	**Masked Significant**	**Enhanced Significant**
rs10250	chr19	4052062	17	16	1	3	21	0	14	37	1	MAP2K2	No	Yes	Yes
rs7662486	chr4	113928795	50	0	2	3	0	0	64	0	0	LOC441034	Yes	No	Yes
rs1049455	chr10	103910465	19	4	0	19	4	0	19	4	0	NOLC1	Yes	Yes	Yes
rs10031608	chr4	12948262	55	0	0	25	0	0	48	0	0	HSP90AB2P	Yes	Yes	Yes
rs6591717	chr11	62091484	322	67	3	230	101	3	315	236	3	EEF1G	Yes	Yes	Yes
rs8177832	chr22	37807512	4	24	0	4	24	0	4	24	0	APOBEC3G	Yes	Yes	Yes
rs1807676	chr22	25613521	0	1	0	0	1	0	0	101	0	LOC100130624	No	No	Yes
rs4792757	chr17	16461158	14	0	0	1	0	0	22	0	0	LOC644422	No	No	Yes
rs3170545	chr19	55128527	9	39	0	9	39	0	9	40	0	ATF5	Yes	Yes	Yes
rs4765775	chr12	4284387	46	19	0	46	19	0	46	19	0	CCND2	Yes	Yes	Yes
rs14408	chr11	298314	48	1	0	15	1	0	64	52	1	IFITM2	Yes	No	No
rs17014852	chr2	127537677	48	0	0	44	0	0	48	0	0	BIN1	Yes	Yes	Yes
rs13296	chr6	44326098	230	69	0	165	76	0	231	162	0	HSP90AB1	Yes	Yes	Yes
rs2228064	chr10	45198056	4	19	1	4	19	1	4	19	1	ALOX5	No	Yes	Yes
rs13306758	chr1	43165406	31	0	0	31	0	0	31	0	0	SLC2A1	Yes	Yes	Yes
rs7662013	chr4	174791881	0	18	0	0	14	0	0	171	0	LOC100128266	No	No	Yes
rs4950386	chr1	145142175	0	3	0	0	3	0	0	26	0	LOC644131	No	No	Yes
rs1051336	chr6	32520570	445	207	1	410	207	1	448	293	2	HLA-DRA	Yes	Yes	Yes
rs631045	chr1	78330102	0	12	0	0	12	0	0	35	1	LOC729768	No	No	Yes
rs4845	chr1	148547174	34	42	1	6	26	1	36	46	1	MRPS21	No	Yes	No
rs3087615	chr7	102739359	26	7	1	26	7	1	26	7	1	PMPCB	Yes	Yes	Yes
rs11953029	chr5	177415790	184	2	1	1	2	3	198	4	2	LOC653314	Yes	No	Yes
rs6677535	chr1	40371584	390	0	0	254	0	0	393	0	0	LOC728602	Yes	Yes	Yes
rs2848622	chr11	57100600	187	0	0	7	1	0	206	1	0	LOC390183	Yes	No	Yes
rs7927338	chr11	93561259	17	0	0	2	0	0	20	0	0	LOC729494	No	No	Yes
rs4726719	chr7	143975161	562	0	3	1	0	5	575	0	2	LOC100132804	Yes	No	Yes
rs11831812	chr12	4077714	0	1	0	0	1	0	0	32	0	LOC399988	No	No	Yes
rs1042448	chr6	33162320	57	2	0	50	2	0	57	61	0	RPL32P1	Yes	Yes	No
rs9276436	chr6	32822061	117	1	1	17	1	1	127	1	0	HLA-DQA2	Yes	No	Yes

rs2574943	chr10	51695173	0	24	0	0	26	0	0	35	0	LOC728532	Yes	Yes	Yes
rs1997	chr2	42431310	24	2	0	19	2	0	24	2	0	COX7A2L	Yes	Yes	Yes
rs1042665	chr5	137930238	50	49	0	19	46	0	50	62	0	SNORD63	No	Yes	No
rs12441946	chr15	70978916	0	0	0	0	0	1	0	26	1	LOC729686	No	No	Yes
rs17101478	chr14	23683279	128	118	0	34	120	0	137	142	0	PSME2	No	Yes	No
rs2223477	chr1	169575456	1	24	0	1	24	0	1	52	0	TOP1P1	Yes	Yes	Yes
rs11986385	chr8	30094837	227	0	0	11	0	0	219	0	0	LOC648729	Yes	No	Yes
rs6570	chr21	45130540	51	22	0	51	22	0	51	22	0	ITGB2	Yes	Yes	Yes
rs9509472	chr13	20433586	45	2	0	1	1	1	36	15	0	LOC440125	Yes	No	No
rs630245	chr1	78330252	29	0	0	18	0	0	29	0	0	LOC729768	Yes	No	Yes
rs4000303	chr1	96686353	1	23	0	0	26	1	0	43	3	LOC440595	Yes	Yes	Yes
rs17627	chr19	44615792	467	530	4	459	540	4	469	630	5	RPS16	No	No	Yes
rs10947623	chr6	36750814	42	0	0	15	0	0	40	0	0	LOC389386	Yes	No	Yes
rs8647	chr19	55128792	10	79	1	10	79	1	10	81	1	ATF5	Yes	Yes	Yes
rs4791596	chr17	14549410	729	1	4	12	1	5	706	1	5	LOC388339	Yes	No	Yes
rs1736924	chr6	29800990	1	35	0	0	37	1	0	124	0	HLA-F	Yes	Yes	Yes
rs2071888	chr6	33380833	56	25	0	55	25	0	56	26	0	TAPBP	Yes	Yes	Yes
rs2428512	chr6	29964309	180	1	0	11	1	3	200	2	1	HLA-H	Yes	No	Yes
rs2089910	chr11	1830980	223	0	1	219	0	1	223	0	1	LSP1	Yes	Yes	Yes
rs7214234	chr17	1708330	35	0	0	16	0	0	36	0	0	LOC642502	Yes	No	Yes
rs10266655	chr7	24705083	33	12	0	33	12	0	33	12	0	DFNA5	Yes	No	Yes
rs4518636	chr8	100973453	23	0	0	10	0	0	22	20	0	COX6C	Yes	No	No
rs2488896	chr1	164513043	0	32	0	0	32	0	0	47	0	LOC284685	Yes	Yes	Yes
rs2731202	chr14	106106026	607	9	4	471	9	4	610	17	4	IGHV5-51	Yes	Yes	Yes
rs2071461	chr11	11330536	0	14	0	0	14	0	0	32	0	CSNK2A1P	No	No	Yes
rs7695	chr1	154413950	11	44	0	11	44	0	11	44	0	SEMA4A	Yes	Yes	Yes
rs3759294	chr12	31835958	49	1	0	0	1	0	38	1	0	LOC440093	Yes	No	Yes
rs1158	chr10	22865294	23	3	0	23	3	0	23	8	0	PIP4K2A	Yes	Yes	No
rs7578292	chr2	136673203	38	0	0	0	1	0	35	1	0	LOC389053	Yes	No	Yes
rs7546434	chr1	145142963	0	1	0	0	3	0	0	26	0	LOC644131	No	No	Yes
rs14105	chr13	27137970	17	48	0	16	46	0	17	48	0	POLR1D	Yes	Yes	Yes
rs9410092	chr9	140190645	1	1	0	1	0	0	1	45	0	LOC643224	No	No	Yes
rs2734945	chr6	29963924	77	2	0	11	2	0	76	2	0	HLA-H	Yes	No	Yes
rs7404	chr5	133335760	90	18	2	48	19	2	90	45	2	VDAC1	Yes	Yes	Yes
rs6568	chr2	127170700	23	6	0	21	6	0	23	7	0	LOC100130248	Yes	No	No
rs2589529	chr15	33317322	1	34	0	1	30	0	1	60	0	LOC723972	Yes	Yes	Yes
rs7205	chrX	1468324	79	102	3	3	90	2	94	186	2	SLC25A6	No	Yes	Yes
rs16952692	chr18	46693267	27	0	0	27	0	0	27	0	0	ME2	Yes	Yes	Yes
rs762815	chr6	32837620	0	12	1	0	12	1	0	50	1	HLA-DQB2	No	No	Yes
rs2814966	chr6	34820210	103	1	0	4	0	0	103	0	0	LOC646785	Yes	No	Yes
rs3871984	chr1	143823858	49	0	0	43	0	0	49	0	0	SEC22B	Yes	Yes	Yes
rs2230659	chr1	45851471	29	8	0	23	13	0	29	23	0	NASP	Yes	No	No

rs1059307	chr6	86444607	15	90	0	15	90	0	15	92	0	SNORD50B	Yes	Yes	Yes
rs199455	chr17	42154400	0	6	0	0	5	1	1	168	0	LOC644315	No	No	Yes
rs11953084	chr5	177415300	219	2	13	125	2	16	219	2	20	LOC653314	Yes	Yes	Yes
rs2719323	chr8	34300101	61	1	0	0	1	0	76	3	0	CYCSP3	Yes	No	Yes
rs7599670	chr2	41900536	0	0	0	0	0	0	0	21	0	LDHAL3	No	No	Yes