# Identification of CT editing sites

For each genomic position with reference base C, we used the likelihood ratio test to assess how well the alignment data was explained by an assumption that the reads originated from a template mixture containing C and T alleles at certain proportion. We defined the likelihood function

|  |  |  |
| --- | --- | --- |
|  |  |  |

where denotes the pileup of all bases aligned at this genomic position, denotes the position of base on the read, and is the CT edit fraction. The probability of observing base given edit fraction is given by

|  |  |  |
| --- | --- | --- |
|  |  |  |

where is the probability of observing base on read position given the template the read originated from had base at the genomic position which maps to. For the latter probability, the following model was assumed

|  |  |  |
| --- | --- | --- |
|  |  |  |

where the overall mismatch rate and the confusion matrix were estimated from the data as functions of the position on the read as described below. When the likelihood is maximized over and compared to the likelihood of data under the condition=0 (no edit), then the quantity

|  |  |  |
| --- | --- | --- |
|  |  |  |

has the distribution with one degree of freedom. Similarly to one can also define likelihoods of data under the assumption that the reads originated from a mixture of C and one of the other (non-T) bases: and, with and denoting the A and G fraction in the template mixture, respectively. The CT edit was called at a genomic position with reference base C if the major allele was T or if the major allele was C and the following two conditions were satisfied:

1. , where is the second (besides ) minor allele observed
2. p-value of derived assuming distribution was below the threshold of .

The p-value threshold was chosen as to ensure that the probability of any CT edit being called accidentally stays below 1% (applying Bonferroni correction for repetitive testing with 8320 C sites gives 0.01/8320 = ). All calculations were performed using the mpileup function of the samtools package [34] coupled with custom perl scripts. Likelihood maximizations were performed using a numerical procedure.

# Estimating mismatch rates

The mismatch rates used in Eq. (3) have been obtained for each sample individually from alignment of this sample’s reads to the respective template. Pre-processing of reads and alignment were performed as described in Section “Read analysis and identification of editing sites”. For each position on the reference, pileup of aligned bases was obtained using samtools [34]. Each base in the pileup was compared to the reference base and counts of matches and mismatches were collected over genomic positions. These counts were binned as functions of position on the read. For each fixed position on read, the overall mismatch rate was calculated as

|  |  |  |
| --- | --- | --- |
|  |  |  |

and the confusion matrix was given by

|  |  |  |
| --- | --- | --- |
|  |  |  |

where denotes the count of base observed at read position given the reference base was . Note that , so that also , as expected.

While collecting mismatch counts for error estimates one needs to exclude genomic positions with known polymorphisms, such as highly edited C sites or obvious SNPs. A site was excluded if the number of reads mapping to it was at least 100 and one of the following conditions was satisfied:

1. Major allele was different from reference allele
2. Reference and major alleles were C, the first minor allele was T, and depth of the T allele was more than 10% of the combined depth of C and T alleles

Variation at all other genomic sites was considered erroneous and contributed to mismatch rate estimates. Once these estimates were used to call CT edits as described above, some of these sites turned out to be statistically significant edits. Mismatch rates obtained in this way reflect errors arising from the experimental procedure to the extent allowed by the adopted strategy of read pre-processing and alignment.