ACGTNacgtn<PLUS | MINUS> fields

ACGTNacgtn<PLUS | MINUS> fields consist of all base counts at a certain pileup position, comma separated in the order A, C, G, T, N (sequencing forward strand), a, c, g, t, n (sequencing reverse strand). The suffixes PLUS, and MINUS define the PCR template strand from which read pairs originate. Here it is noteworthy that read pairs with orientation F1R2 originate from a plus stranded PCR template, while those with orientation F2R1 are derived from minus stranded PCR templates. The sequencing strand is encoded by F (forward strand, i.e. sequencing direction was 5` to 3` with respect to the reference sequence), and R (reverse strand, i.e. sequencing direction was 5` to 3` with respect to the reverse complement of the reference sequence). Thus, the ACGTNacgtnPLUS fields are calculated using read pairs with orientation F1R2, while ACGTNacgtnMINUS fields are calculated using read pairs with orientation F2R1. On the other hand the entries for ACGTN are calculated using F<1 | 2> reads, and acgtn entries are calculated using R<1 | 2> reads. For determining counts of the ACGTNacgtn<PLUS | MINUS> fields we only consider reads that exceed a mapping quality of 1, and bases exceeding a base quality of 1.

PCR- and Sequencing strand count matrices

For calculating the count matrices we consider six possible mutations (C to A, C to G, C to T, T to A, T to C T to G) in 16 possible triplet contexts, where the base up- and downstream of the mutational position can be either of A, C, G, or T. For each possible mutation and triplet context we pool it's reverse complement, since on double stranded DNA level a mutation in a certain triplet context is equal to it's reverse complement on the opposite strand. This leaves us with 96 possible sources of PCR-, and sequencing strand bias. Using the information stored in the ACGTNacgtn<PLUS | MINUS> fields we are able to calculate for each of the 96 fields the number of alternative bases observed on either the forward or reverse PCR- or sequencing template strand and write this information into the respective error matrix. Afterwards, we define a mutation in a certain base triplet context as being biased following the decision tree in Figure 1:

Let n be the number of bases showing evidence for the SNV, n_f be the number of bases showing evidence for the SNV on forward stranded reads, f_r be the number of bases showing evidence for the SNV on reverse stranded reads, n_m be the overall number of SNVs found in the triplet context, $f_f := n_f/n$, and $f_r := n_r/n$. Furthermore we define a set of parameters to control the strictness of our bias filter: Let α be the significance threshold of a binomial test (default 0.01), b_{min} be the minimal frequency of the more abundant strand on which a SNV is being observed (default 0.53), n_{min} be the minimal number of reads observed for a SNV in a certain triplet context (default 20), $n_{m.min}$ be the minimal number of SNVs found in a certain triplet context (default 4), and b_{max} be the minimal frequency of the more abundant strand on which a SNV is being observed to flag the field as strongly biased (default 0.63).

Biased mutation marking

After defining SNVs in certain base triplet context as being weakly or strongly biased, individual SNVs within the biased contexts are investigated individually and marked if necessary: Let f_{max} be the maximal fraction of reads observed on the opposite strand, when checking for a read bias towards a certain strand (default 0.1), $n_{max.weak}$ be the maximal number of reads observed on the opposite strand when checking for a read bias towards a certain strand on a field which was defined as weakly biased (default 0), and $n_{max.strong}$ be the maximal number of reads observed on the opposite strand when checking for a read bias towards a certain strand on a field which was defined as strongly biased (default 1). The decision tree in Figure 2 shows how an individual SNV is being marked as biased.

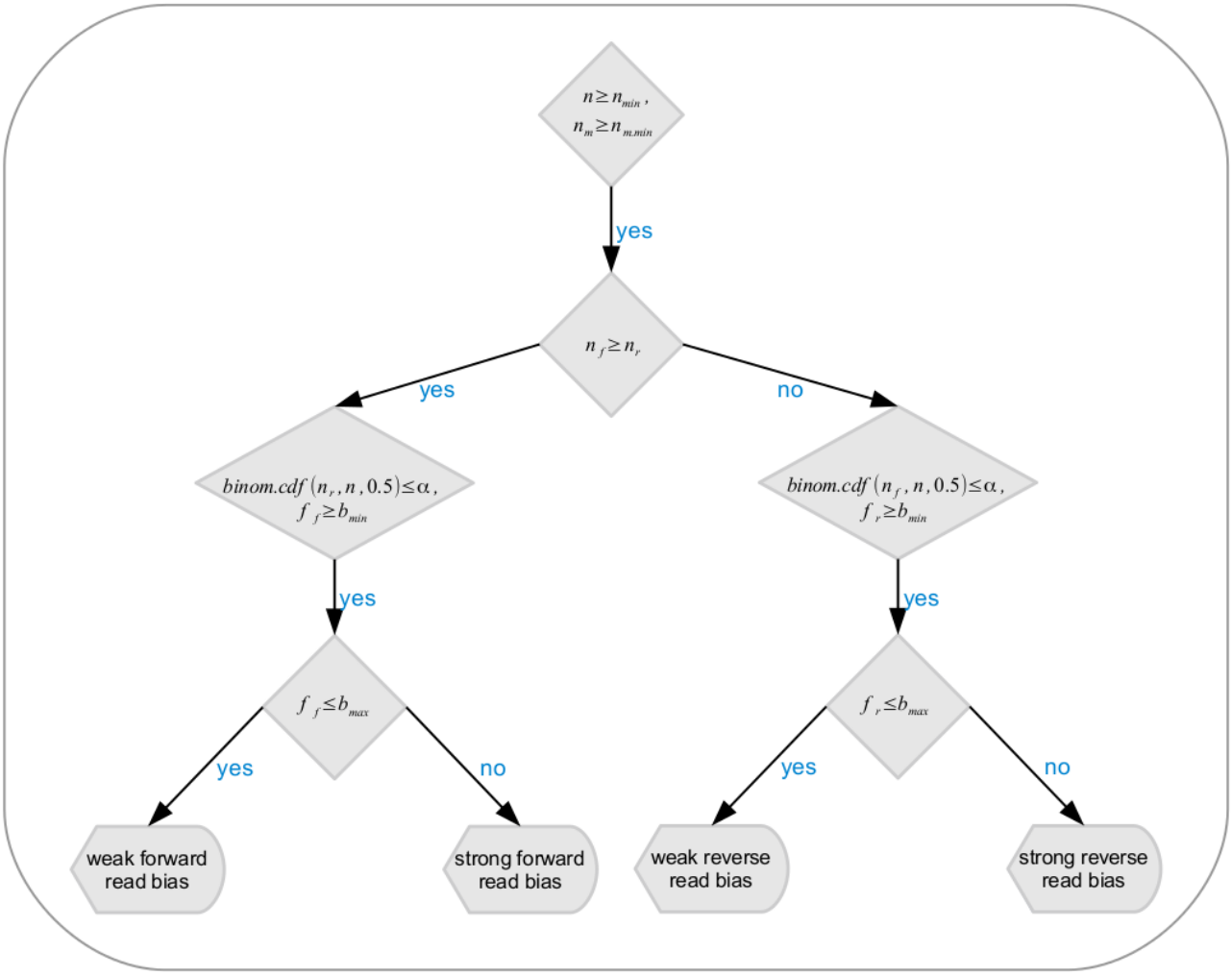

Figure 1: Decision tree outlining the scheme for read bias discovery.

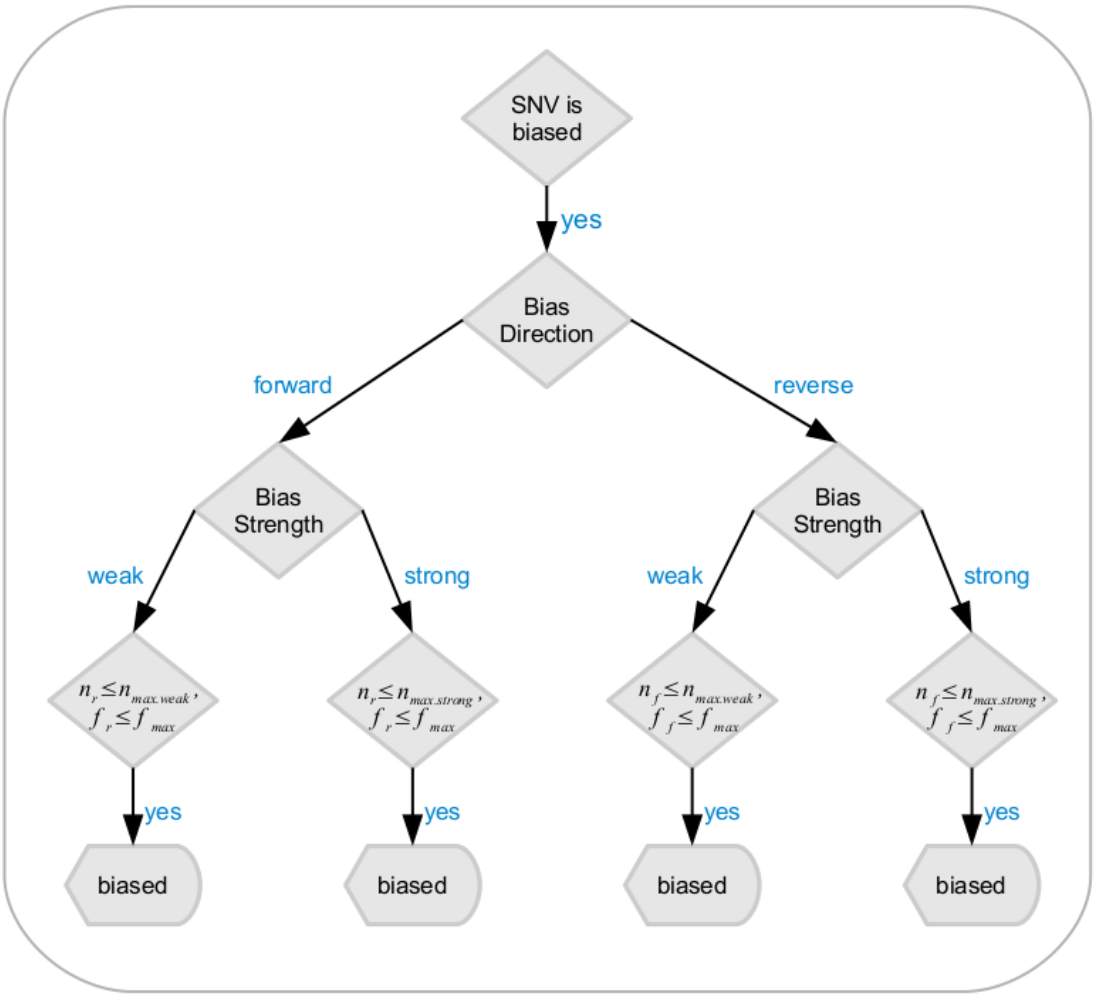

Figure 2: Decision tree outlining the process of marking individual SNVs as being biased