We leverage the assumed error rate of letters in the sequence to judge whether a particular alignment S is "correct" by measuring the probability that the observed disagreements across sites are due to read errors alone.

For now we focus on the observed agreement/disagreement in just one haplotype h. At a particular site i in h, there is a majority vote, and a minority vote. We call each minority vote a "disagreeing letter". Let the coverage at site i be C_i^h (which can change dynamically as reads move in and out of h). We assume read errors are uniformly distributed across all letters of all reads with constant probability p per letter. Then in a correct alignment, the distribution for the number of disagreements at site i in haplotype h is Poisson with mean $\lambda_i^h = pC_i^h$. For a given alignment S, let k_i^h be the number of minority votes at site i, haplotype h. Then, under the assumption that S is a correct alignment, the probability that the k_i^h minority votes at site i are due to read errors alone is the tail of the Poisson distribution,

$$T_P(\lambda_i^h, k_i^h) = e^{-\lambda_i^h} \sum_{j=k_i^h}^{\infty} \frac{(\lambda_i^h)^j}{j!}. \tag{1}$$

(This can also be expressed as 1 minus the cumulative Poisson distribution at k_i^h, and computed by summing towards infinity until the terms are too small to change the floating-point sum.) In essence, our goal is to make this probability as *large* as possible across all sites—to maximize the probability that all minority votes are due to read errors alone. To create an objective function that accomplishes this goal, we take the negative log probability of Equation (1) across

all sites i, and sum:

$$f_s^h(S) = -\sum_i \log T_P(\lambda_i, k_i), \tag{2}$$

where the subscript s in f_s refers to the fact that this is a *site*-based objective. Our goal is then to find S that minimizes $f_s(S)$.

There may be a subtle bias in the above because Equation (2) is meant to measure the log probability of an entire alignment S, and by summing the log probability at each site (equivalent to multiplying the probabilities at each site), it treats each site as having an independent T_P. However, sites are not independent because we cannot swap individual letters between haplotypes, we can only swap entire reads. When we do so, the T_P value at all the sites for that read will change in a correlated manner. To account for this, we can "transpose" the argument above: rather than looking at the errors (ie., minority votes) across one site, we can instead look at all the disagreements that an individual read r has with the majority votes across all its sites. To that end, given a read r that has l_r sites, the expected number of read errors in read r is $E_r = pl_r$. Thus, in a correct alignment S, the number of sites that read r disagrees with the majority is Poisson with mean E_r. By a similar argument leading to Equation (1), if S is a correct alignment then observing k_r disagreements with the majority vote across its sites is the tail of the Poisson distribution at k_r evaluated as $T_P(E_r, k_r)$. Then, a *read*-based objective f_r (that—hopefully!—accounts for the correlation between sites when a read r is swapped between haplotypes) can be described by

$$f_r(S) = -\sum_r \log T_P(E_r, k_r). \tag{3}$$

It will not be clear which objective gives better results until we experiment with both.

There is still one more potential bias: all of the above applies to just one haplotype. The total objective will be the sum of $f(S)$ across all haplotypes, and such a sum assumes each haplotype is independent. This assumption is not true, because when a read is moved from one haplotype to another, clearly the changes across r's sites are correlated between the two haplotypes involved in the move. For now we ignore this dependency since it's not clear how to remove it.

Our method has several advantages over other methods. For one, there are *absolutely no restrictions* on any of the parameters: there can be an arbitrary number reads, and an arbitrary number of sites; coverage at all sites can be arbitrarily high or low; read length can be arbitrarily large or small; the number of haplotypes is unrestricted; and the error rate p can be arbitrarily high, although of course our ability to recover the correct alignment will degrade as p increases. We need make no arbitrary judgement calls, assumptions, or parameter choices to make the algorithm "work". No empirical experiments are required to set parameters that, once so restricted, may adversely affect the quality of the result. Finally, since simulated annealing is a *random* search algorithm, we can judge the quality of the result by running it many times and seeing how the alignment changes: if the alignment stays largely constant across multiple independent runs of the algorithm, we can be confident that the objective function has a well-defined optimum alignment and that we are finding that alignment. (Whether the objective is the best way to find the "correct" alignment is a different matter that can require experimen-

tation.) Furthermore, low-confidence regions of the alignment (ie., reads or regions whose final resting place change between full runs of the aligner) can be highlighted as requiring higher coverage.