Dindel: Accurate Indel Calls from Short Read Data Supplementary Information

Cornelis A. Albers[1,2,†] Gerton Lunter[3] Daniel G. MacArthur[1]
Gil McVean[4] Willem H. Ouwehand[2,1] Richard Durbin[1]

[1] Wellcome Trust Sanger Institute, Hinxton, UK
[2] Department of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge, UK
[3] Wellcome Trust Centre for Human Genetics, Oxford, UK
[4] Department of Statistics, University of Oxford, Oxford, UK
[†] Genome Campus, Hinxton, Cambridge, CB10 1HH, UK

1 Generation of candidate haplotypes

Figure 2 outlines the procedure used to the generate candidate haplotypes. We use a heuristic approach to define the haplotype blocks. We use the reference haplotype sequence to define the first haplotype block, and then proceed by adding the reads sequentially in the order in which they were mapped to the reference sequence. Given a current set of haplotype blocks, the procedure for adding a read R is as follows: given the mapped position of the read R, the first block b_1 overlapping with the read R and the part of the read R_{b_1} overlapping with block b_1 is determined. If the substring R_{b_1} is not present as one of the subhaplotypes in the block, it is added as a subhaplotype to the haplotype block with frequency 1. If it is present, the frequency of the subhaplotype is increased by one. This procedure is repeated with the remaining part of the read, $R_{\neg b_1}$, until the whole of the read has been added to the haplotype block model. We let the start of a read always create a new haplotype block, or split an existing haplotype block. The frequencies the two haplotype blocks resulting from a split are obtained by marginalisation. Thus, for high coverage data the resulting haplotype block model is generally highly similar to the pileup as generated by SAMtools. However, for low coverage data the resulting halpotype block model will contain fewer blocks with longer subhaplotypes.

2 Probabilistic realignment model

In this section we describe the realignment model we use to compute the likelihood of observing a read given a haplotype sequence. We first introduce our notation. Capital symbols refer to random variables, lower case ones refer to observed values of the corresponding random variable. R_i^b refers to base b in read i, and r_i^b refers to the corresponding observed value: $R_i^b \in \{A, C, G, T\}$. H_p^b refers to base b in candidate haplotype p, and l_p is the number of bases in haplotype p. H_p^b is a hidden variable that is not directly observed. The set of candidate haplotypes is denoted by $\mathcal{H} = \{\mathbf{H}_p\}$. $X_i^b \in \{L, 1, \ldots, l_p, R\}$ is the position of base b in read i in the haplotype H_p^b the read is to be aligned to. L is a special state indicating that base b is aligned to the left of the candidate haplotype, ie left of the haplotype window. R is a special state indicating that base b is aligned to the right of the candidate haplotype, ie right of the haplotype window. $I_i^b = \{0, 1\}$ is an indicator variable that specifies whether base b in the read is part of an inserted sequence ($I_i^b = 1$) with respect to the haplotype. q_i^b is the *a priori* probability that base b in read i was correctly called; we infer it from the base quality reported in the alignment file for that read. q_i is the probability that read i was correctly mapped. q_i and q_i^b are assumed to be given by the read mapper that aligned the reads to the reference sequence.

We use the Bayesian network framework to define the realignment model. Figure S2 illustrates it with an example haplotype and read sequence. The model has two components. The first component models how every base in the read is aligned with

respect to the haplotype. Here there are two possibilities: either a read base is aligned to a base on the haplotype, the read base is part of an insertion with respect to the haplotype, or the read base is aligned to the left or the right of the haplotype. The variables X_i^b and I_i^b together fully specify the alignment of base b in read i. The second component models for a given alignment of the read base what the probability is of observing the observed base $R_i^b = r_i^b$.

We will now specify the conditional probability tables for the alignment component of the model. Based on the initial alignment of the read to the reference sequence (as produced by the read mapper), we choose a base b_0 in the read that serves as an anchor point. b_0 is chosen as the base that is in the middle of the segment of the read that overlaps with the candidate haplotype according to the initial alignment of the read mapper (and which may be refined by the realignment procedure), subject to the constraint that it is at least 10 bp away from the boundaries of the candidate haplotype. This constraint can always be satisfied as we only realign reads that have an overlap of at least 20 bases with the haplotype according to the initial alignment of the read mapper. For unmapped reads (for which the mate is mapped) b_0 is chosen to be the middle base in the read. We use the mapping quality of read i for this: we assume that the probability that base b_0 is aligned within the window defined by the coordinate of the leftmost base in the haplotype and the rightmost base in the haplotype is given by the mapping quality, which encodes the probability that the read was mapped correctly. Choosing the prior this way is formally incorrect since the base qualities may have been incorporated in the mapping quality, and we will use the base qualities again in the observation model component. However, in practice we find that this heuristic gives good results as it does account for haplotype sequences that are not unique in the genome. For single-ended reads the prior distribution is given by

$$P(1 \leq X_i^{b_0} \leq l_p | q_i) = q_i.$$

For paired-end reads of which the mate is mapped to the same chromosome, we use the aligned position of the mate and the library insert size distribution to parametrize the insert size distribution:

$$P(1 \leq X_i^{b_0} \leq l_p | q_i^{\text{mate}}, c_{\text{mate}}) = \frac{q_i^{\text{mate}}}{Z} P(IS(X_i^{b_0}, c_{\text{mate}}) | \text{insert size distribution}),$$

where $IS(X_i^{b_0}, c_{\text{mate}})$ is the insert size given the position $X_i^{b_0}$ of the read and the position of its mapped mate c_{mate}, and q_i^{mate} is the mapping quality of the mate. The normalization constant Z is given by

$$Z = \sum_{X_i^{b_0}=1}^{l_p} P(IS(X_i^{b_0}, c_{\text{mate}}) | \text{insert size distribution}).$$

The insert size distribution of the library is estimated empirically from all reads mapped

in pairs. Here we assume that the mate is mapped to the correct position c_{mate} and does not require substantial realignment.

Indels in the reads with respect to the haplotype, i.e. sequencing errors, are modeled by specifying the transition probabilities between X_i^b and I_i^b for neighbouring bases in the read. Consider two adjacent bases in the read b and b'. We will consider the cases $b, b' > b_0$ and $b, b' < b_0$ separately. We will first treat the first case. The table that models the insertions and deletions is given by $P(X_i^{b'}|X_i^b, I_i^{b'}, I_i^b)$. The next position $X_i^{b'}$ depends on both $I_i^{b'}$ and I_i^b in order to model insertions with respect to the haplotype. We have

$$P(X_i^{b'}|I^{b'}, X_i^b, I^b, b' > b > b_0) = \begin{cases} X^b, I^{b'} = 0, I^b = 0 & \Rightarrow \quad P(d)\delta(X^{b'}, X^b + d), d \in \{0, 1, \ldots, \Delta\} \\ X^b, I^{b'} = 1, I^b = 0 & \Rightarrow \quad \delta(X^{b'}, X^b) \\ X^b, I^{b'} = 1, I^b = 1 & \Rightarrow \quad \delta(X^{b'}, X^b) \\ X^b, I^{b'} = 0, I^b = 1 & \Rightarrow \quad \delta(X^{b'}, X^b + 1) \end{cases} \tag{1}$$

Here δ is the Kronecker delta function. The top transition represents the cases where there is no insertion, but where there could be a deletion ($d > 1$). $\Delta - 1$ is the maximum length of read-deletion (ie, a sequencing error). $d = 0$ corresponds to a base-extension error during sequencing. We assume the following distribution for sequencing deletion errors:

$$P(d) \propto \exp(|d - 1|),$$

and choose the proportionality constant such that $P(d = 1)$ corresponds to the probability of not having a deletion due to a sequencing error. The second line represents a transition from a base that is generated by a base on the haplotype to an inserted sequence, indicated by $I^{b'} = 1$. As long as $I_i^b = I_i^{b'} = 1$, we have $X^{b'} = X^b$ (third transition), so that the model 'remembers' the position in the haplotype of the last base that was aligned to the haplotype. The last line represents the transition from a read base that is part of an inserted sequence to a next base that is again generated by a base on the haplotype. The case $b' < b < b_0$ is analogous except that $X_i^{b'} = X^b - \delta$, where $\delta \in \{0, 1, \ldots, \Delta\}$ and corresponding changes to the other transitions in order to maintain the consistency in terms of the remembered position for insertions. An advantage of choosing the base b_0 to be the root of the Bayesian network is that the transition probability for cases where the next base b' is not aligned to the haplotype but b is aligned to the haplotype becomes easier to quantify. In Fig. S2 we see for example that read base $b = 24$ is aligned to haplotype base 19. We thus have $P(X_i^{b=25} = R|X_i^{b=24} = l_p, \cdots) = 1.0$. Similarly we have $P(X_i^{b=3} = L|X_i^{b=4} = 1, \cdots) = 1.0$.

We assume a first-order Markov property for insertions in the read with respect to the reference to maintain computational tractability. Thus, we have

$$P(\mathbf{I}_i|\theta) = P(I_i^{b_0}|\theta_0(X_i^{b_0})) \prod_{b=b_0+1}^{l_p} P(I_i^b|I_i^{b-1}, \theta(X_i^{b-1})) \prod_{b=0}^{b_0-1} P(I_i^b|I_i^{b+1}, \theta(X_i^{b+1}))$$

The two parameters are the transition probabilities $\theta_0(X_i^b) \equiv P(I_i^{b'} = 1|I_i^b = 0, hplen(X_i^b|\mathbf{H}_p))$ and $\theta_1 \equiv P(I_i^{b'} = 1|I_i^b = 1)$; here $b' = b + 1$ for $b > b_0$ and $b' = b - 1$ for $b < b_0$. θ_0, which is analogous to the 'gap open' penalty in pairwise alignment HMMs, is chosen according to table S1 and is dependent on the homopolymer run length of the haplotype sequence, through the function $hplen(X_i^b|\mathbf{H}_p)$. The effect is that in a long homopolymer tract an indel in the read with respect to the haplotype will be more likely than in complex sequence. θ_1 is analogous to the 'gap-extension' penalty, which we have chosen as $\theta_1 = \exp(-1)$.

The probability of observing a base $R_i^b = r_i^b$ given its base quality and the alignment of that base with respect to the haplotype. Here we mean by alignment specifically that X_i^b and I_i^b are given for every base in the read. We parametrize the probability of observing R_i^b as follows:

$$P(R_i^b = r_i^b|X_i^b, \mathbf{h}_p, I_i^b, q_i^b) = \begin{cases} I_i^b = 0, & X_i^b \notin \{L, R\} & : & q_i^b \delta(r_i^b, h_p^{X_i^b}) + (1 - q_i^b) Unif(r_i^b) \\ & X_i^b \in \{L, R\} & : & q_i^b \\ I_i^b = 1 & & : & q_i^b \end{cases} \tag{2}$$

Thus, if the base-quality q_i^b is low the read base is primarily drawn from a uniform distribution over the different nucleotides. If a read-base is part of an inserted sequence with respect to the haplotype (i.e., a sequencing error) or of it does not map to the haplotype ($X_i^b \in \{\text{left}, \text{right}\}$) we assume it was observed without error.

We use the Viterbi algorithm to infer the maximum-likelihood alignment $\{X_i^b, I_i^b\}$:

$$P_{\max}(\mathbf{R}_i|\mathbf{H}_p) \equiv \max_{X_i^b, I_i^b} P(\mathbf{R}_i = \mathbf{r}_i, \mathbf{X}_i, \mathbf{I_i}|\mathbf{H}_p, \theta). \tag{3}$$

Although it would be more appropriate to do a summation rather than a maximization over the hidden variables, we do the latter for the following reasons. First, maximization is more computationally efficient than summation, and second, having a particular alignment will allow us to count how many reads are covering a particular sequence variant in each haplotype. In homopolymers we explicitly correct for the fact that we do maximimzation rather than summation: we compute the probability of seeing an indel in the full homopolymer run according to the error model and then associate this increased probability only with the first nucleotide in the homopolymer run.

3 Bayesian EM algorithm

For the analysis of pooled samples the number of different haplotypes segregating in the underlying set of samples is unknown, and therefore we applied a Bayesian EM algorithm that does not overfit the number of segregating haplotypes. The Bayesian EM algorithm is the variational EM algorithm as described by Bishop (2007). As described in the main text, we partition the set of candidate into (potentially overlapping) subsets of candidate

haplotypes $\mathcal{H}_k$, with $k = 1, \ldots, K$ ranging over the K different subsets. Each subset corresponds to a hypothesis that a specific subset of the candidate sequence variants segregate in the pool of samples. The subset $\mathcal{H}_k$ consists of all candidate haplotypes that can be constructed from combinations of these sequence variants (including the reference allele for each variant). We define n_k as the number of candidate haplotypes in subset k, so that $\mathcal{H}_k = \{\mathbf{H}_k^{l=1}, \ldots, \mathbf{H}_k^{l=n_k}\}$. Here each $\mathbf{H}_k^l$ refers to candidate haplotype sequence l in subset k. The reference candidate haplotype (the reference sequence without any sequence variants) is included in every subset k as the first haplotype, which implies $\mathbf{H}_{k=1}^1 = \mathbf{H}_{k=2}^1 = \cdots = \mathbf{H}_{k=K}^1$.

We apply the Bayesian EM algorithm independently to each subset k to infer the haplotype frequencies $\mathbf{f}_k = \{f_k^1, \ldots, f_k^{n_k}\}$. Since it is applied to each subset independently, we will now consider one such subset, and drop the subscript k. We denote the vector of haplotype frequencies for that subset by the vector π. The idea is that each candidate haplotype represents a cluster, and that each read can be thought of as having been generated by one of the clusters. The clustering model assumes a prior Dirichlet distribution for the haplotype frequencies

$$\mathrm{Dir}(\pi|\alpha_0) \propto \prod_{l=1}^{n_k} \pi_l^{\alpha_0 - 1},$$

with α_0 controlling the sparsity of the clustering. The default value used in Dindel is $\alpha_0 = 0.001$. We define the responsibilities r_{il}, with i indexing reads and l haplotypes, as

$$r_{il} \equiv \frac{\rho_{il}}{\sum_{j=1}^{n_k} \rho_{ij}}.$$

Here ρ_{il} is defined through the relation

$$\log \rho_{il} = \mathbb{E}[\log \pi_l] + \log P_{\max}(\mathbf{R}_i|\mathbf{H}^l), \tag{4}$$

which follows from the variational approximation of the clustering model. The approximate posterior $q^*(\pi) = \mathrm{Dir}(\pi|\alpha)$ is also Dirichlet, with components

$$\alpha_k = \alpha_0 + \sum_{i=1}^{R} r_{il}. \tag{5}$$

Following the formulation of (Bishop, 2007), an iteration of the algorithm consists of the following operations:

E-step The E-step consists of computing the responsibilities ρ_{il} using Eq. 4, and subsequently the Dirichlet parameters α_k using Eq. 5, given the current estimates $\mathbb{E}[\log \pi_l]$.

M-step The M-step consists of updating the haplotype frequencies π_l:

$$\mathbb{E}[\log \pi_l] \leftarrow \psi(\alpha_l) - \psi(\hat{\alpha}),$$

where $\psi(\cdot)$ is the digamma function and $\hat{\alpha} = \sum_l \alpha_l$.

After convergence of the algorithm, we approximate the probability of the read data by

$$Z \equiv P(\mathbf{R}_i|\pi) \approx \sum_{il} \rho_{il}. \tag{6}$$

Running the Bayesian EM algorithm for each subset $\mathcal{H}_k$ provides the probability of the data Z_k through Eq. 6 and the haplotype frequencies $\mathbf{f}_k$ for each subset k. Since a candidate haplotype may be included in multiple subsets $\mathcal{H}_k$, the posterior haplotype frequency is obtained by computing the weighted average over the subsets with weighing factor Z_k. Similarly, the posterior probability of an indel segregating in the sample can be estimated by summing the Z_k of all subsets $\mathcal{H}_k$ where the variant was defined to be segregating, and normalizing by $\sum_k Z_k$.

4 Indel errors in Illumina reads

Basal error rate: Illumina reads from the 1000 Genomes Pilot 1 (low coverage across 179 individuals) were mapped using Stampy (Lunter and Goodson, 2010), a read mapper designed to be sensitive in the presence of indel mutations. Only reads with a mapping quality exceeding 10 were kept. The resulting data set had modal coverage of 530. We excluded all regions with coverage exceeding twice the mode (1060). Indels identified by the read mapper were collected, and filtered to include only indels that were called at least 8 bp from either end of the read. This data set underlies all analyses related to indel error rates. To estimate the basal indel rate, we first removed all indels that were supported by reads from more than one library, under the assumption that these were likely to represent true variation. This resulted in 48,283,494 calls in 2.836Gb, or 0.017 indels per site. Dividing by the (modal) coverage gives the estimate indel error rate per nucleotide per read, 3.2×10^{-5}.

Expected rate of singletons: Under the assumption of a neutrally evolving population under the standard coalescent without population structure, the number of segregating sites is proportional to the product of half the heterozygosity, θ, and the total expected coalescent tree, length, $2(1 + 1/2 + \ldots + 1/n - 1)$ where n is the sample size. From the neutral form of the site frequency spectrum, it is seen that the fraction of segregating sites that occur at frequency 1 in the sample is $1/(1 + 1/2 + \ldots + 1/n - 1)$. It follows that the fraction of singletons (per site) in a population is constant and equal to the heterozygosity. For indels, $\theta_{\text{indel}} = \theta_{\text{snp}}/8$ approximately (Lunter, 2007), giving the estimate 1.25×10^{-4}.

Allele frequency spectrum in homopolymer context: We called indels that were observed (at identical sites, and with identical lengths and sequence) in at least 3 individuals. This effectively filters out false calls assuming the background indel error rate of 0.017 per site (estimated average false discovery rate $0.0173/1.25\times10^{-4} = 0.04$). When more than one indel is observed at a single locus, which is a frequent occurrence in long homopolymer runs, we selected the indel supported by the largest number of reads. As estimate of the allele frequency we used the number of individuals in which a particular indel (with given position, length and sequence) was observed. While this procedure leads to under-estimates of frequencies in the presence of homozygous genotypes, our interest is in the low end of the frequency spectrum, and so does not affect the current analysis.

Allele frequency spectra not explained by recurrent mutations: While indel errors would explain the observed allele frequency spectra, it remains a priori possible that recurrent mutations on a polymorphic background would change the allele frequency spectrum in a similar way. Indeed, a high mutation rate would be expected to lower the relative proportion of singletons in a population, which is consistent with observations. Fig. S3 shows the allele frequency spectrum in a neutrally evolving Wright-Fisher population ($N_e = 1000$), with mutation rate $\theta = 2N_e u = 1$ and 10. Only for the higher mutation rate $\theta = 3$ does the allele frequency spectrum start to resemble the observed distribution. However, $\theta = 3$ is 5 orders of magnitude above the basal mutation rate ($\theta = 1.25 \cdot 10{-}4$), and close to that of microsatellites (Weber and Wong, 1993), contrary to expectation for the comparatively short homopolymers; moreover, virtually all sites with $\theta = 1$ would be expected to be polymorphic, which is not what we observe (1000 Genomes Consortium, personal communication).

An Illumina indel read error model: Table S1 lists error rate estimates using the three methods described in the main text. We used this data to arrive at a mixed interpolated and extrapolated indel error model. Because only approximate bounds are at our disposal, the resulting model involves a regrettable but unavoidable degree of arbitrariness.

For complex sequence, singleton indels are clearly in majority errors, and their frequency does not increase up to homopolymers of length 4. The upper bound estimate also does not vary in this regime, and varies between 0.038 and 0.048. It appears conservative to use the singleton-based rates in this regime.

The upper bound estimate for homopolymers of length 10 is expected to be an overestimate, but nevertheless, it is only twice higher than the estimate based on the allele frequency spectrum, which is expected to be an under-estimate. We therefore use the upper bound estimate in our model.

To close the gap between length 4 and length 10 homopolymers, we chose error rates to exceed the singleton-based rates, first by a modest amount, while for longer homopolymers

in such a way to progressively approach the upper bound estimate. We have no data to support these choices, but they seem to represent conservative estimates, assuming that indel error rates vary smoothly as a function of homopolymer run length, which all our data appears to support.

Beyond 12 base-pair homopolymers, the supposedly upper bound rate estimate start to decrease, for reasons indicated in the main text. In fact, up to 10 bp the curve traced by the upper bound indel rate estimate is concave, while beyond 10 bp it is convex. The slope is steepest at 0.36, going from 9 to 10 bp homopolymers. Since extreme value statistics are sensitive to noise and upwardly biased, we used the interval 8-10 instead, and chose 0.30 as the slope for the error model beyond 8 nt homopolymers. The figure between brackets is the final error rate model, converted to per-base-pair and per-read values.

References

Bishop, C. M. 2007. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, corrected at sixth printed edition.

Lunter, G. 2007. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, 23(13).

Lunter, G. and Goodson, M. 2010. Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. submitted.

Weber, J. and Wong, C. 1993. Mutation of short tandem repeats. *Hum Mol Genet*, 2:1123–8.

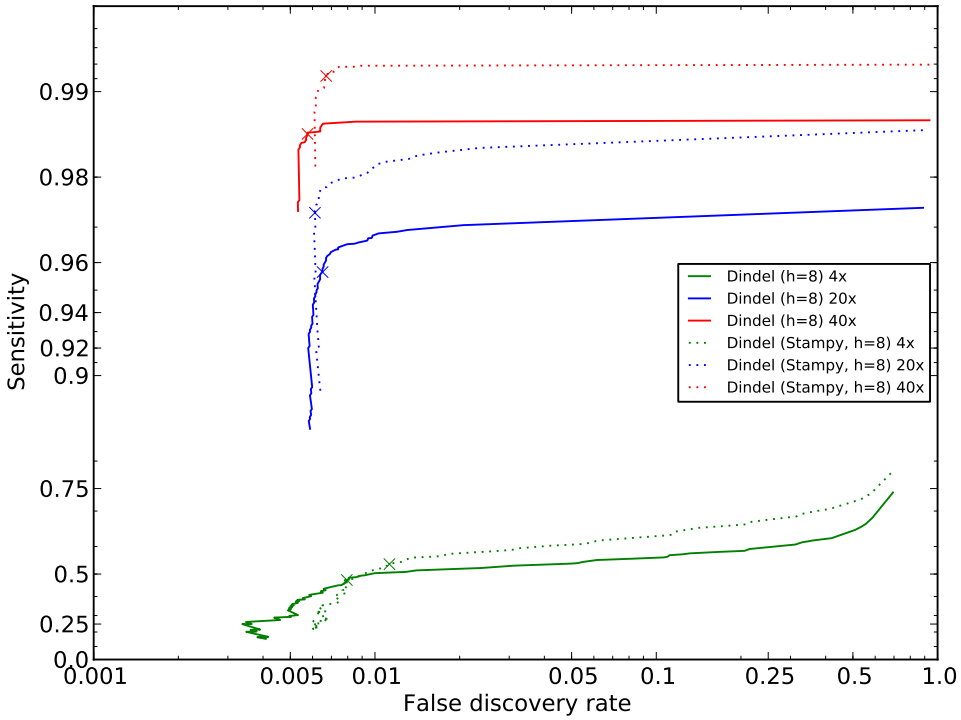

Figure S1. Comparison of the Stampy read mapper with BWA 0.5.7 on the simulated data set of Fig. 4B. Dindel applied to the alignments and candidate indels produced by the Stampy read mapper results in higher sensitivity while maintaining low false discovery rates compared to Dindel applied to BWA.

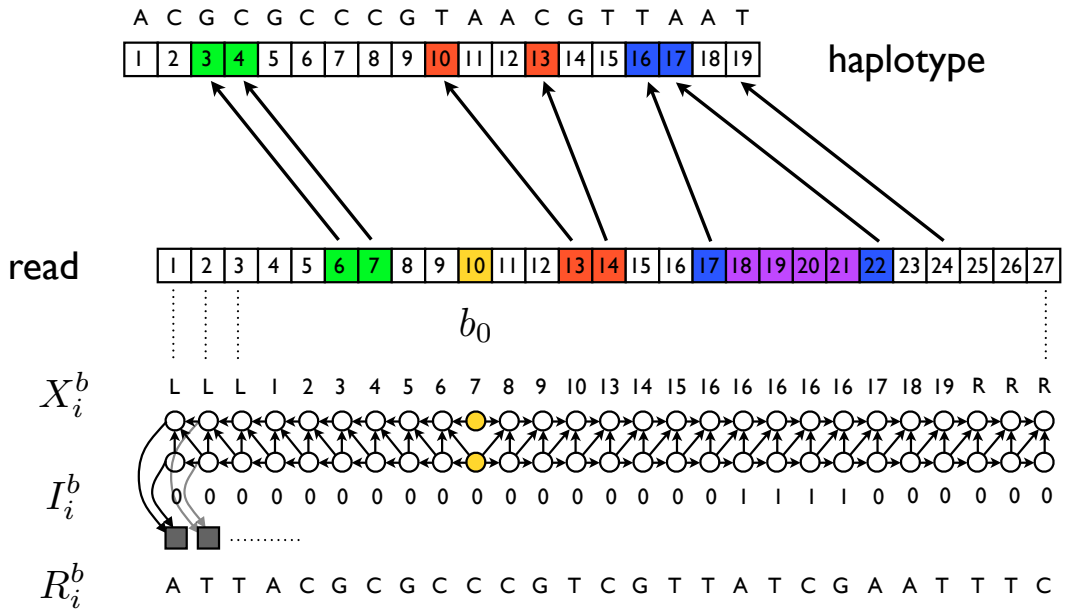

Figure S2. Probabilistic alignment model. Top horizontal bar represents a haplotype. The numbers inside the boxes represent the position of each base in the haplotype; above the horizontal bar an example haplotype sequence is shown. Below that is a horizontal bar that represents a read, with in each box a number that represents the position of each base in the read. The read sequence is shown in the bottom of the figure, as an instantiation of the observed variables R_i^b in the model which represent the observed read bases. In between is shown the Bayesian network, where the circles represent random hidden variables, the gray squares represent the observed variables, and the arrows represent conditional dependencies between the variables. The hidden variables are X_i^b, the position of each read base in the haplotype, and I_i^b, which indicates whether base b in read i is part of an insertion with respect to the haplotype. b_0 (marked in yellow) is the base in the read for which a prior probability of $X_i^{b_0} \neq \{L, R\}$, ie the probability that the read is aligned to the window specified by the haplotype, and $I_i^{b_0} = 0$, the probability that this base is part of an insertion, will be specified. For the hidden variables here the most likely value given the haplotype sequence and the read sequence, as inferred using the Viterbi algorithm is shown respectively above and below the corresponding variables in the Bayesian network. The colors in the read and the haplotype illustrate the different alignments possible. The green bases are an example of the case where consecutive bases in the read (here 6 and 7) are aligned to consecutive bases in the haplotype (3 and 4) without gap. The red bases are an example of the case where consecutive bases in the read (13 and 14) are aligned to non-consecutive bases 10 and 13 in the haplotype, signifying a deletion with respect to the haplotype. Blue shows the case where the read contains an inserted sequence with respect to the haplotype. For the inserted bases $I_i^{b=18,19,20,21} = 1$, and for those bases X_i^b is not incremented until the first base (read base 22) that is not part of the insertion and aligned to the haplotype.

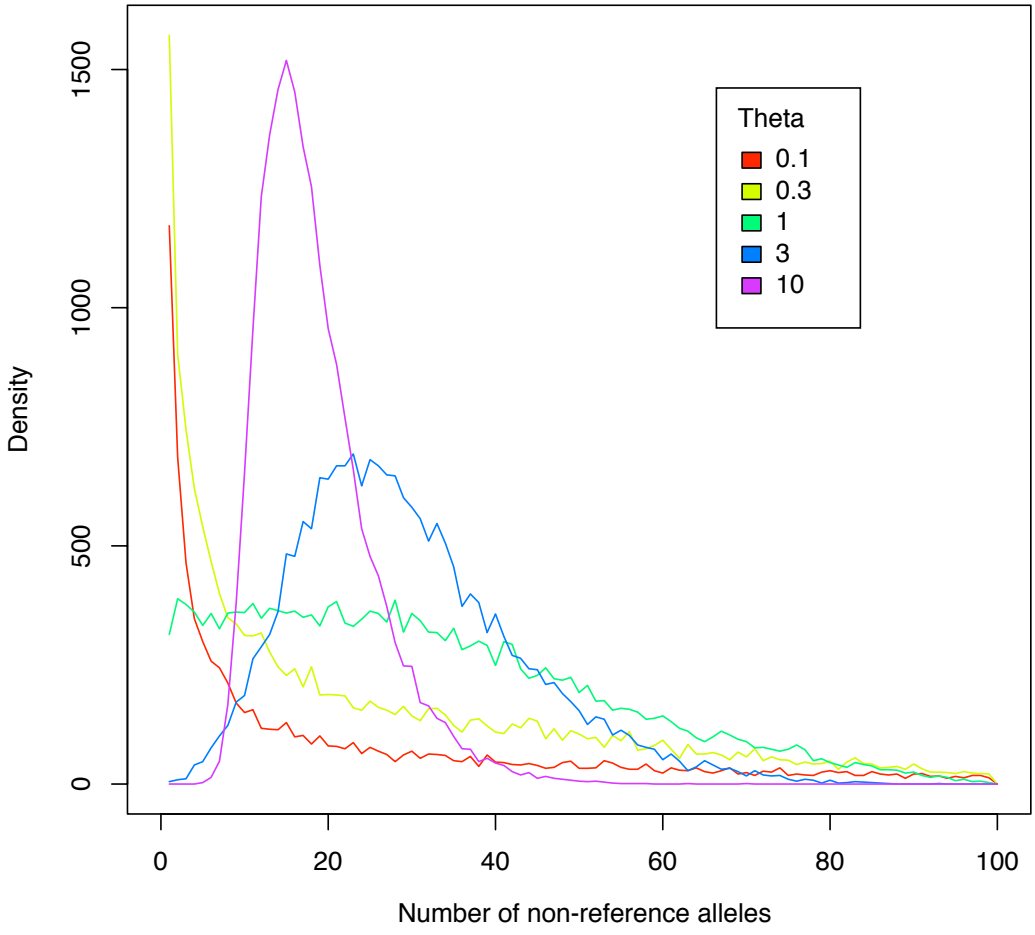

Figure S3. Allele frequency spectra under a Wright-Fisher model ($N_e = 1000$) and recurrent mutations, for different mutation rates θ. Although low-frequency alleles become less dominant in a regime of high mutation rates, the outline of the $1/f$ distribution remains visible, and does not appear to converge to the binomial-shaped distribution observed in the indel calls from Illumina data, even at extremely high mutation rates.

Context	Singletons	Allele frequency spectrum	All indel calls	Error model
Complex	**0.017**	-	0.048 (>)	0.017 ($3.2 \cdot 10^{-5}$)
2	**0.017**	-	0.042 (>)	0.017 ($3.2 \cdot 10^{-5}$)
3	**0.016**	-	0.038 (>)	0.017 ($3.2 \cdot 10^{-5}$)
4	**0.018**	-	0.043 (>)	0.02 ($3.8 \cdot 10^{-5}$)
5	0.023	-	0.057 (>)	0.03 ($5.7 \cdot 10^{-5}$)
6	0.031	-	0.11 (>)	0.08 ($2.0 \cdot 10^{-4}$)
7	0.041	-	**0.20** (>)	0.17 ($3.8 \cdot 10^{-4}$)
8	0.068	-	**0.42** (>)	0.40 ($7.5 \cdot 10^{-4}$)
9	0.093	-	**0.70** (>)	0.70 ($1.3 \cdot 10^{-3}$)
10	0.099 (<)	0.5 (<)	**1.06** (>)	1.00 ($1.9 \cdot 10^{-3}$)
11	0.101 (<)	0.55 (<)	**1.34** (>)	1.30 ($2.4 \cdot 10^{-3}$)
12	0.110 (<)	0.58 (<)	1.46 (>)	1.60 ($3.0 \cdot 10^{-3}$)
13	0.106 (<)	0.69 (<)	1.36	1.90 ($3.6 \cdot 10^{-3}$)
14	0.110 (<)	0.64 (<)	1.29	2.20 ($4.2 \cdot 10^{-3}$)
15	-	0.73 (<)	1.28	2.50 ($4.7 \cdot 10^{-3}$)
16	-	0.69 (<)	1.19	2.80 ($5.3 \cdot 10^{-3}$)
17	-		1.17	3.10 ($5.8 \cdot 10^{-3}$)
18	-		1.0	3.40 ($6.4 \cdot 10^{-3}$)

Table S1. Estimates of indel error rates per site, in Illumina short-read sequences across 180 individuals with a modal total coverage of 530×, and stratified by sequence context (complex sequence, or occurring in homopolymer runs of lengths 2 to 18). Three methods were used to estimate error rates (column 2,3 and 4). The last column describes the error model used in the realignment HMM, with the figure in parentheses representing the per-base and per-read error rate. Error rate estimates are marked (<) or (>) if the estimates are expected to be under- or over-estimates respectively. Estimates in bold were used to guide the error model.