Supplemental Methods

HMM parameter estimation by the expectation-maximization algorithm

After reads are mapped into the genome reference, we extract at random a training data set with N pairs of reads and their mapped reference segments, denoted as $D=\left\{\left(\mathbf{r}^d,\mathbf{t}^d\right)|\, d=1,2,\cdots,N\right\}$. Because the ground-truth alignments are not available, we design an iterative strategy to train the hidden markov model. Firstly, the parameter estimation procedure assigns a probability to every possible alignment of a read $\mathbf{r}$ and a reference $\mathbf{t}$ conditional on the training data D. Secondly, the model parameters are updated according to the probabilistically weighted alignments using an expectation-maximization algorithm described in the following. To specify the parameter estimation procedure, two kind of missing variables are defined. The first one $\gamma_{i,j,k,l,\pi}$ is the probability that a length k homopolymer ending at the read position i is aligned to a length l homopolymer ending at the target position j at the hidden state π. The second one $\xi_{i,j,k,l,\pi',\pi}$ is the probability that a length k homopolymer ending at the read position i is aligned to a length l homopolymer ending at the target position j at the hidden state π transitioned from the previous hidden state π'.

The iterative strategy first calculates $\left\{\gamma^d_{i,j,k,l,\pi}\right\}$ and $\left\{\xi^d_{i,j,k,l,\pi',\pi}\right\}$, $1\le d\le N$, $1\le i\le n_d$, $1\le j\le m_d$, $n_d=\left|\mathbf{r}^d\right|$ and $m_d=\left|\mathbf{t}^d\right|$ using the forward-backward algorithm (26). The forward variable $F\left(i,j,k,l,\pi\right)$ is the probability summarizing over all possible alignments ending at the hidden state π between the prefixes $\mathbf{r}_{[1:i]}$ and $\mathbf{t}_{[1:j]}$ in which a k bp homopolymer at the read position i is aligned to a l bp homopolymer in the reference position j. The forward variable is computed through the forward algorithm,

$$F\left(i,j,k,l,\pi\right)=\sum_{\pi'}F\left(i-k,j-l,\pi'\right)p\left(\pi\,|\,\pi'\right)p\left(\beta_i\,|\,\alpha_j,\pi\right)p\left(k\,|\,l,\beta_i,\pi\right)$$

$$F\left(i,j,\pi\right)=\sum_{k,l}F\left(i,j,k,l,\pi\right)$$

The backward variable $B\left(i,j,k,l,\pi\right)$ is the probability of all possible alignments of the suffixes $\mathbf{r}_{[i+1:n]}$ and $\mathbf{t}_{[j+1:m]}$ starting at the hidden state π, where a k bp homopolymer at the

read position i is aligned to a l bp homopolymer at the reference position j. The backward variable is computed through the backward algorithm,

$$B(i,j,k,l,\pi) = \sum_{\pi',di,dj} p(\pi'|\pi)\, p\left(\beta_{i+di} \mid \alpha_{j+dj}, \pi'\right) p(di \mid dj, \beta_{i+di}, \pi') B(i+di, j+dj, di, dj, \pi')$$

where $0 \le di \le k_{\max}$, $0 \le dj \le l_{\max}$, $k_{\max}$ is the maximum length of homopolymers starting at the read position $i+1$ and $l_{\max}$ is the maximum length of homopolymers starting at the reference position $j+1$.

With the forward variable $F^d(i,j,k,l,\pi)$ and the backward variable $B^d(i,j,k,l,\pi)$ for the d th data, the missing variables $\gamma^d_{i,j,k,l,\pi}$ and $\xi^d_{i,j,k,l,\pi',\pi}$ are computed by the following equations,

$$\gamma^d_{i,j,k,l,\pi} = \frac{F^d(i,j,k,l,\pi) B^d(i,j,k,l,\pi)}{p\left(\mathbf{r}^d, \mathbf{t}^d\right)}$$

$$\xi^d_{i,j,k,l,\pi',\pi} = \frac{\sum_{k',l'} F^d(i-k, j-l, k', l', \pi')\, p(\pi|\pi')\, p\left(\beta_i \mid \alpha_j, \pi\right) p(k \mid l, \beta_i, \pi) B^d(i,j,k,l,\pi)}{p\left(\mathbf{r}^d, \mathbf{t}^d\right)}$$

where k' is the length of homopolymers ending at the read position $i-k$ and l' is the length of homopolymer ending at the reference position $j-l$. The probability that $\mathbf{r}^d$ generated from $\mathbf{t}^d$ is calculated as $p\left(\mathbf{r}^d, \mathbf{t}^d\right) = \sum_\pi F(n,m,\pi)$.

Secondly, the iterative parameter estimation counts the occurrences of the hidden state transition events and homopolymer emission events occurring within the training data D. Three counting tables, T for the occurrences of the state transitions, L for the occurrences of the length calling, and B for the occurrences of the base calling, are defined to update the model parameters using the following equations,

$$T_{\pi',\pi} = \sum_{i,j,k,l,d} \xi^d_{i,j,k,l,\pi',\pi}$$

$$L_{k,l} = \sum_{i,j,k',l',d,\pi} \gamma^d_{i,j,k',l',\pi} \cdot \delta(k',k) \cdot \delta(l',l)$$

$$B_{\beta,\alpha} = \sum_{i,j,k,l,\pi} \gamma^d_{i,j,k,l,\pi} \cdot \delta(r_i,\beta) \cdot \delta\left(t_j,\alpha\right)$$

where r_i is the i-th nucleotide on $\mathbf{r}^d$, t_j is the j th nucleotide on $\mathbf{t}^d$, $\delta(x_1,x_2)$ is an Kronecker delta function, and $\delta(x_1,x_2)=1$ if $x_1 = x_2$ and 0 otherwise. The hidden state transition probabilities and the base call rate matrix are updated by the maximum likelihood estimations from the counting tables of T and B.

The update of the length call rate matrix is described in the following. In the proposed probabilistic model, the flow intensity is modeled as a latent variable. Assuming the raw flow data is not available, we update the parameters of the Laplace distribution, log-normal distribution and additive power model using two nested EM algorithms, an inner EM algorithm embedded in the outer EM algorithm. A missing variable $p(f \mid k,l)$ is defined as the probability that the flow intensity f is the immediate value linking the called length k and the input length l. At the E-step we sample the flow intensity space by the size $\Delta = 0.01$ and calculate the missing variable by the following equation,

$$p(f \mid k,l) = \frac{p(k \mid f)\, p(f \mid l)}{\sum_{f'} p(k \mid f')\, p(f' \mid l)}.$$

At the M-step the homopolymer length dependent scale parameters are calculated by

$$b_l = \frac{\sum_{f,k} p(f \mid k,l) \cdot L_{k,l} \cdot |f - l|}{\sum_{f,k} p(f \mid k,l) \cdot L_{k,l}}.$$

To avoid the sparsity, we take the data range from $l = 0$ to l^* of which $L_{l^*,l^*} \geq 1000$ to estimate the parameters. The parameters of the power-law model $b_l = c_0 + c_1 \times l^{c_2}$ are fitted by a quasi-Newton method. The variance of the log-normal model is updated by the following equation,

$$\sigma_0^2 = \frac{\sum_{f,k} p(f \mid k,0) \cdot L_{k,0} \cdot \ln^2 f}{\sum_{f,k} p(f \mid k,0) \cdot L_{k,0}}$$

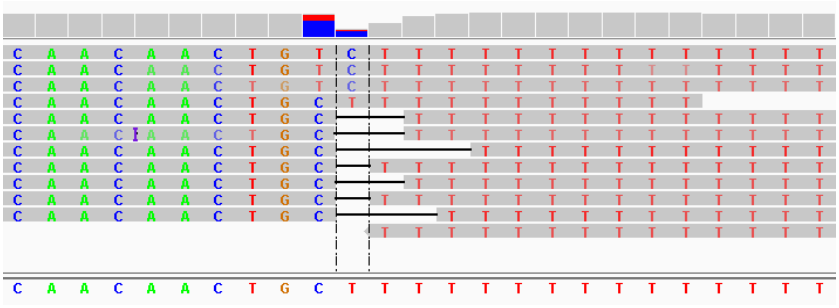
The iteration terminates when it meets the maximum number of iterations or the increment of the likelihood is less than the threshold (0.001).

Supplemental Experiments

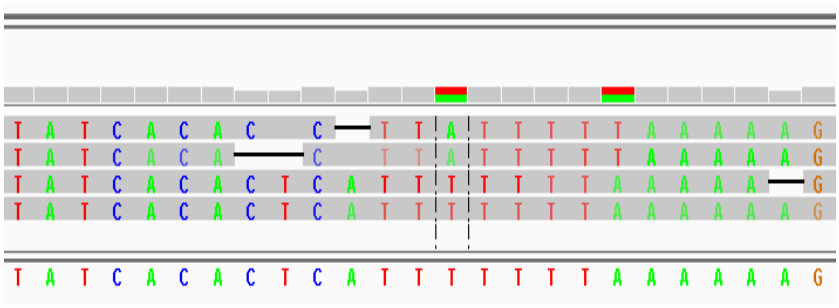
Simulation on the 454 data. We downloaded the reference genome of *E. coli* substrain MG1655 (NC_000913.2) and its whole genome resequencing data (SRR001355) generated by the 454 sequencer. The data consist of 256,503 reads with an average length of 244bp, and an average coverage of about 9.6-fold. Our simulation strategy followed three steps. First, we generated an artificial reference genome by randomly mutating bases on the *E. coli* MG1655 genome, creating a set of ground-truth SNPs. Second, we ran SNP-calling programs to predict SNPs using both the 454 resequencing data and the artificial reference genome. Finally, we assessed the accuracy of the predictions.

In this simulation, we sampled 10,000 random positions in the genome and mutated each base into one of the other three bases with an equal probability. For the resequencing data, we filtered out about 7.4% of poor quality reads and mapped the remaining reads onto the artificial reference genome using Megablast. We then applied PyroHMMsnp, Samtools, and VarScan, to call SNPs. Next, we calculated the sensitivity and specificity and plotted them in Supplemental Figure 10. Overall, all three methods performed well because long homopolymers are rare in the *E. coli* genome, as illustrated in Supplemental Figure 11, and flow intensities of short homopolymers of the 454 data have high fidelity. Using the logarithmic-scale y-coordinates to zoom in the increasing progression in Supplemental Figure 10, we found that the curve for PyroHMMsnp climbed quickly towards the top-left corner of the plot, achieving high sensitivity at low false discovery rates. For example, at a low false discovery rate of 0.5%, the sensitivity for PyroHMMsnp is about 99.8%, better than that of Samtools (98.2%) and VarScan (97.8%). We did not show the results of Atlas-SNP2 because it had the lowest accuracy. Overall, we conclude that the realignment-based approaches outperform the others in the detection of SNPs.

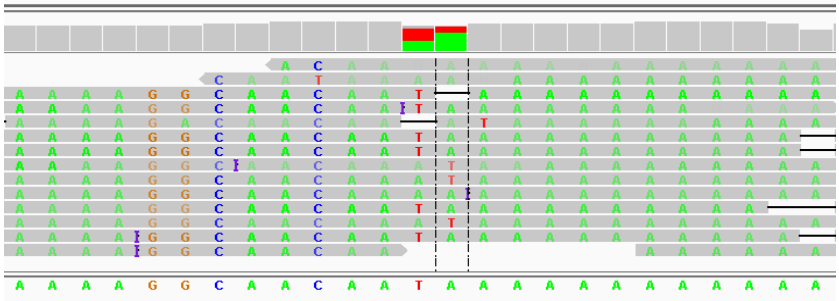
A.



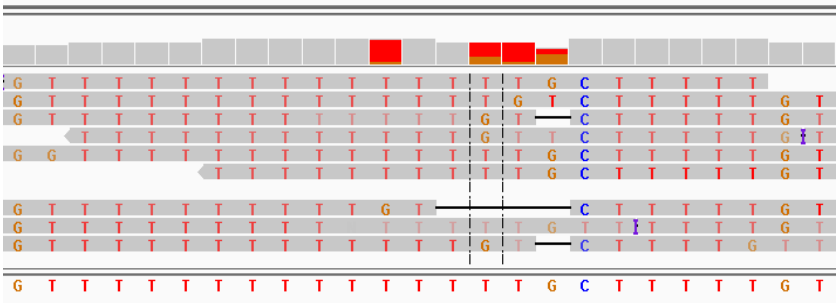
B.



C.

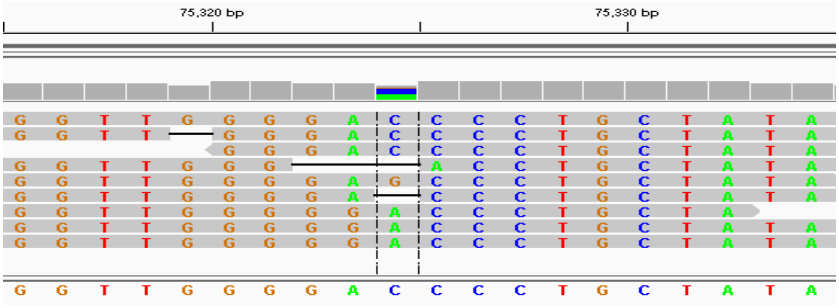


D.

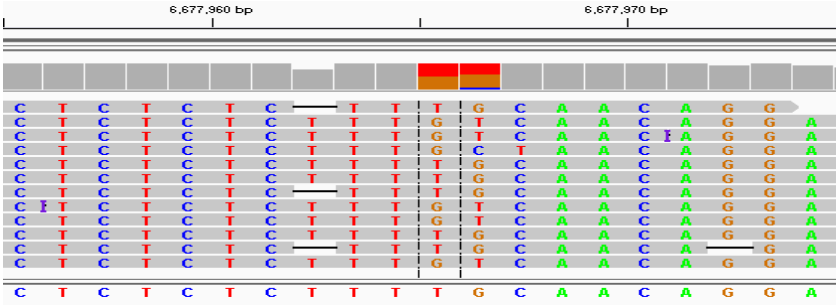


Supplemental Figure S1. The mis-alignments of the raw mapping results caused by homopolymer undercalls or overcalls. The reads were from the 454 sequencing data and visualized by the IGV viewer.

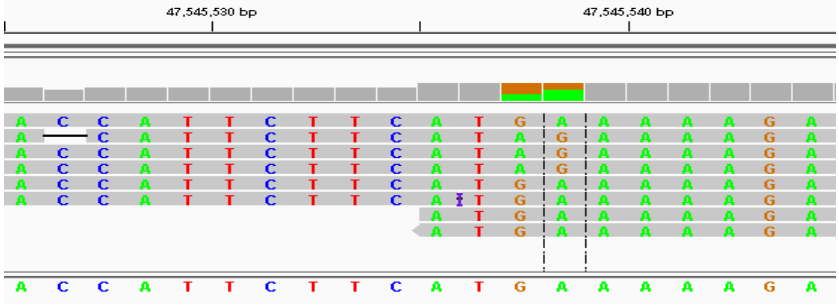
A.



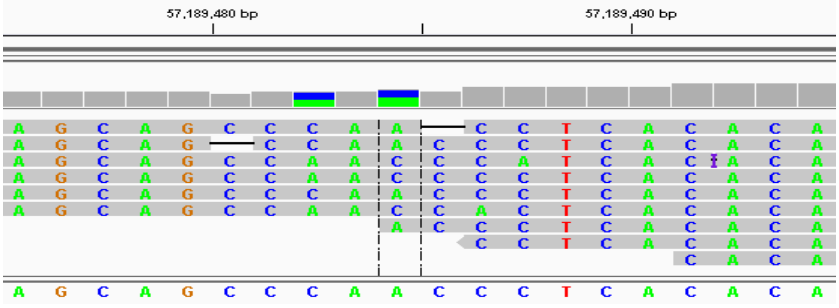
B.



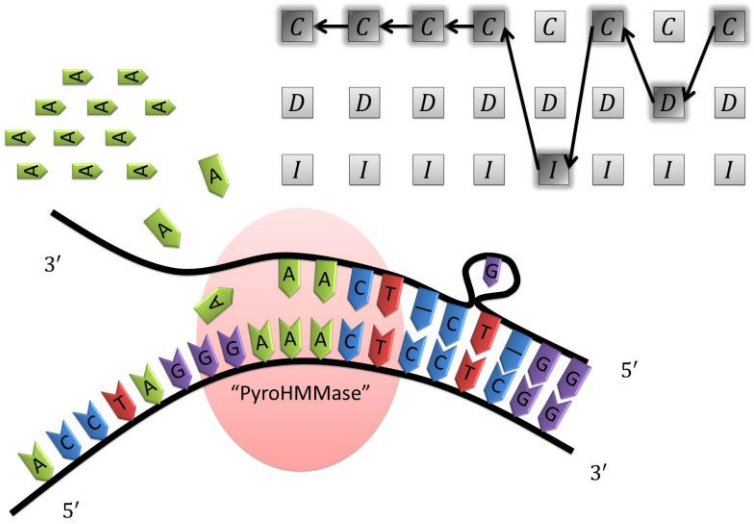
C.



D.

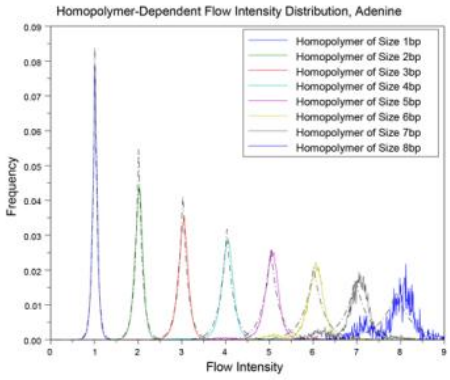


Supplemental Figure S2. The mis-alignments of the raw mapping results caused by homopolymer undercalls, overcalls, or CAFIEs. These examples were from the BWA-SW mapping results of the Ion Torrent sequencing data.

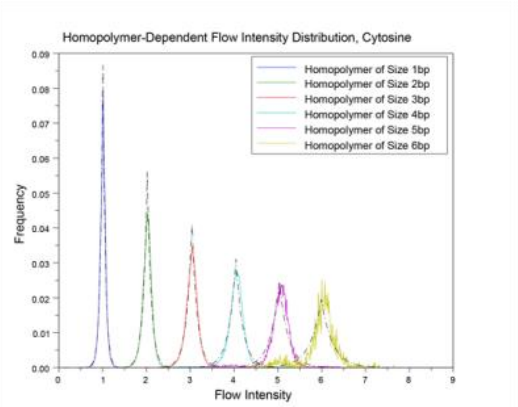


Supplemental Figure S3. The virtual pyrosequencing machine. The virtual machine, named PyroHMMase, takes input of a DNA template sequence and produce homopolymer reads according to the characteristics of pyro-sequencing. The top-right light gray nodes represent the hidden states of the hidden markov model. The path of dark gray nodes consists of the decoded states at which the virtual machine has been.

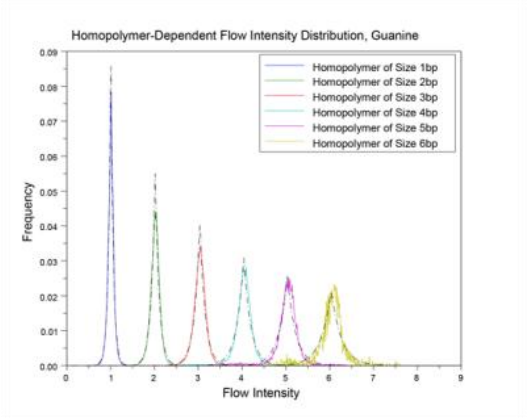
A



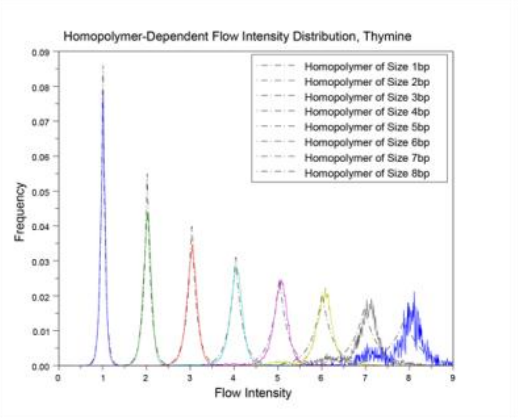
B



C

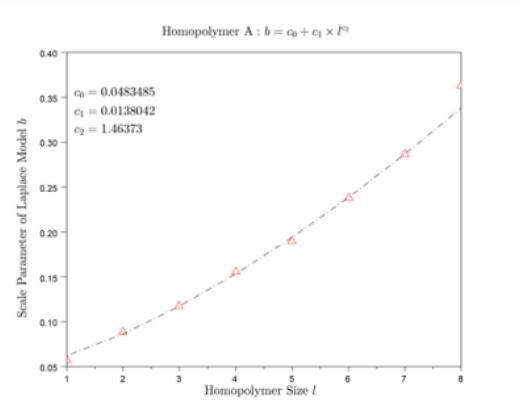


D

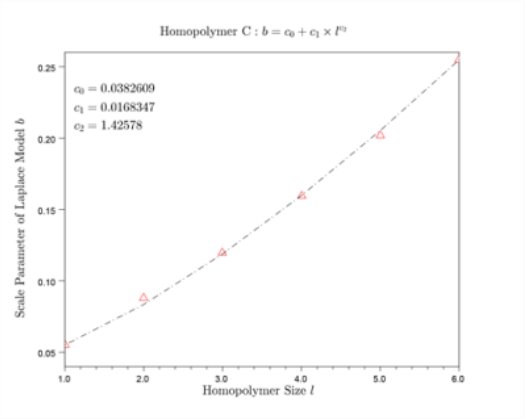


Supplemental Figure S4. The empirical distributions and model-fitted curves of homopolymer flow intensities for the 454 sequencing data. The plots for poly(dA), poly(dC), poly(dG), and poly(dT) are shown in figures A, B, C, and D, respectively.

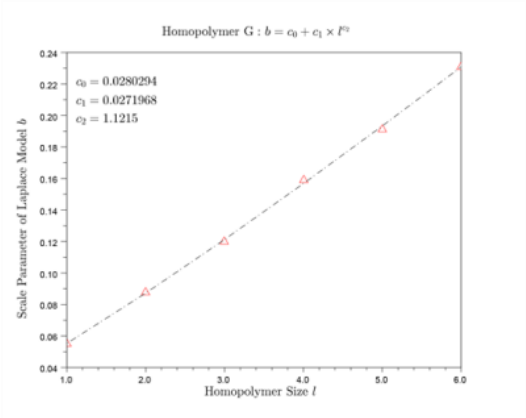
A



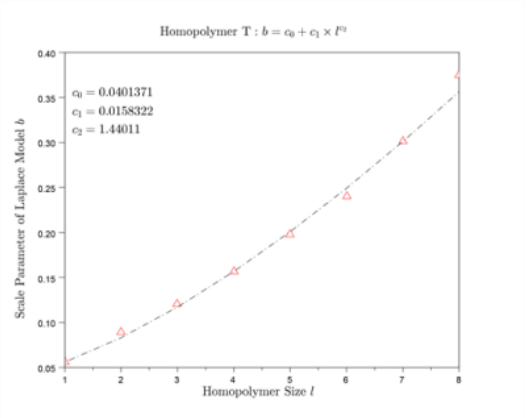
B



C

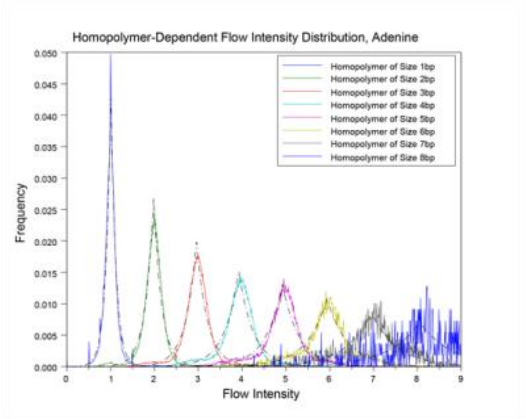


D

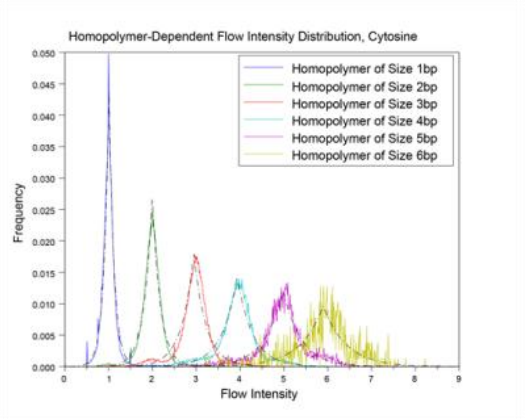


Supplemental Figure S5. The scale parameters of the estimated Laplace distributions and the model fitted curves of the homopolymer flow intensities for the 454 resequencing data. The plots for poly(dA), poly(dC), poly(dG), and poly(dT) are shown in figures A, B, C, and D, respectively.

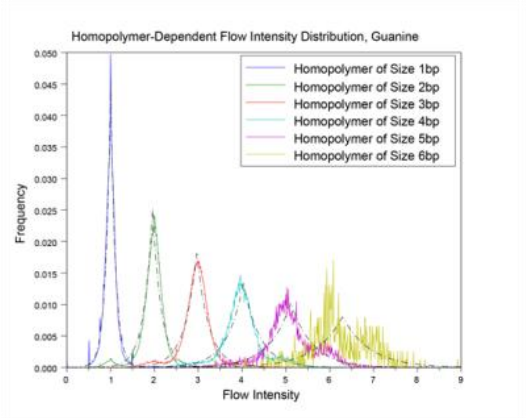
A



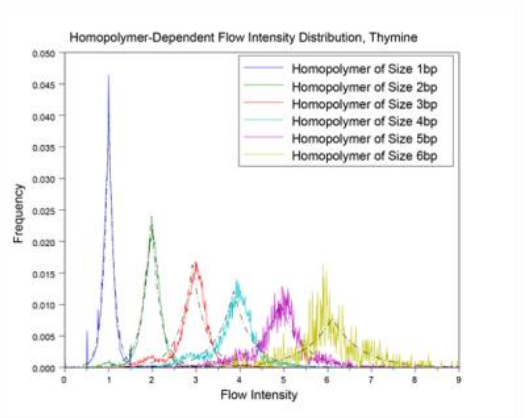
B



C

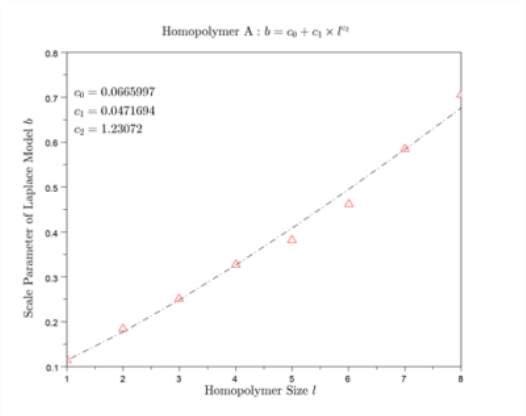


D

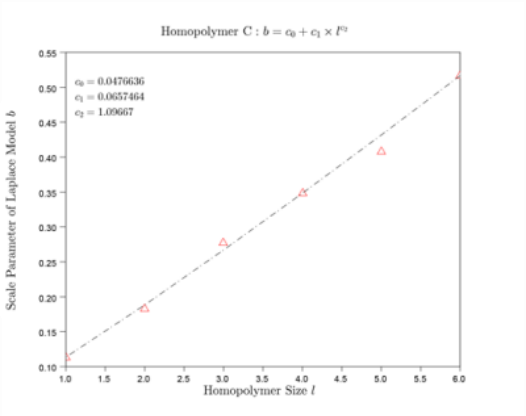


Supplemental Figure S6. The empirical distributions and model-fitted curves of the homopolymer flow intensities for the Ion Torrent sequencing data. The plots for poly(dA), poly(dC), poly(dG), and poly(dT) are shown in figures A, B, C, and D, respectively.

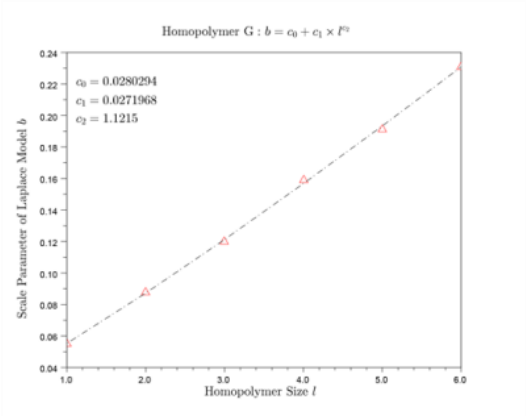
A



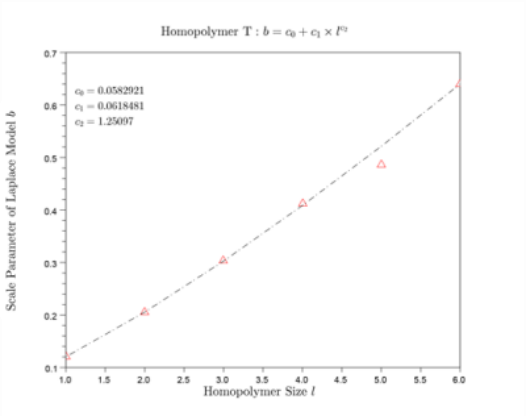
B



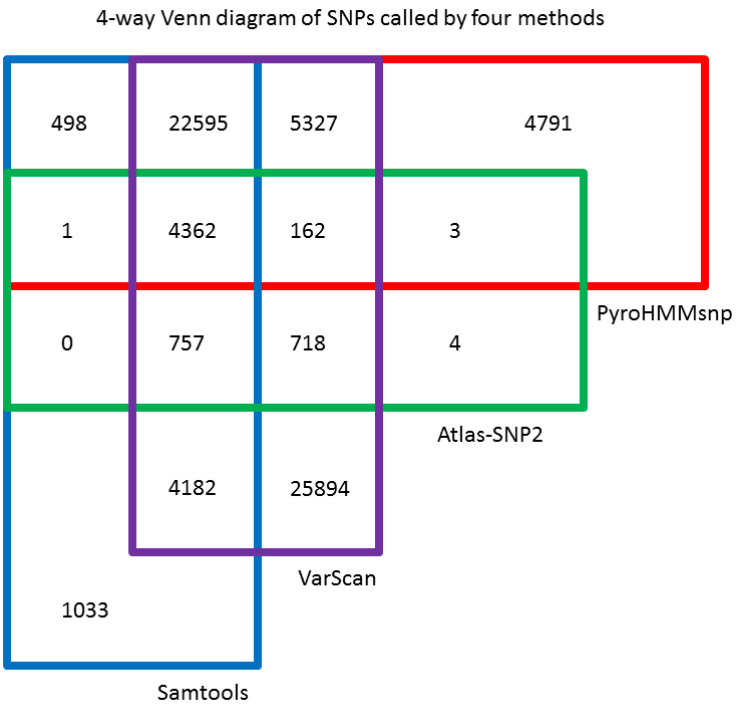
C

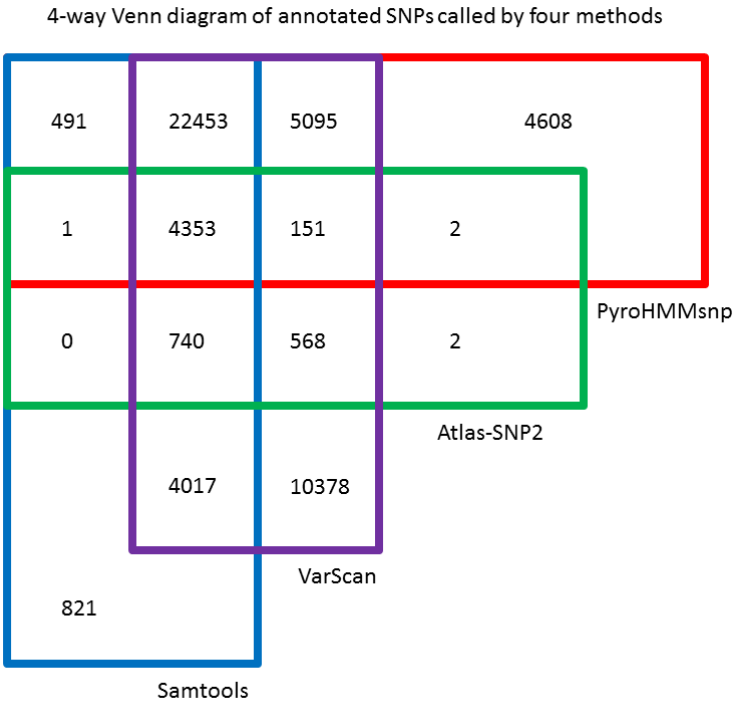


D



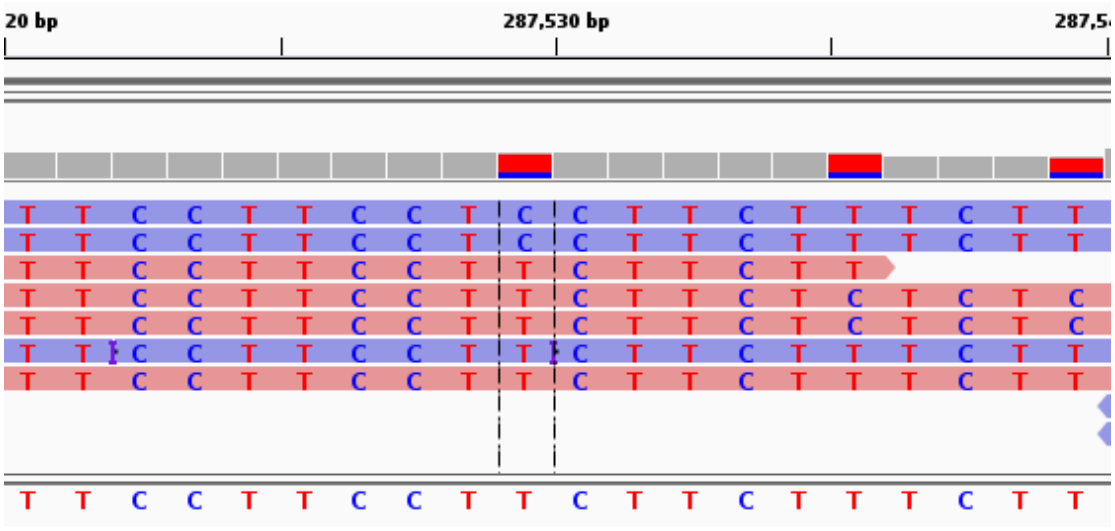
Supplemental Figure S7. The scale parameters of the estimated Laplace distributions and the model fitted curves of homopolymer flow intensities for the Ion Torrent resequencing data. The plots for poly(dA), poly(dC), poly(dG), and poly(dT) are shown in figures A, B, C, and D, respectively.



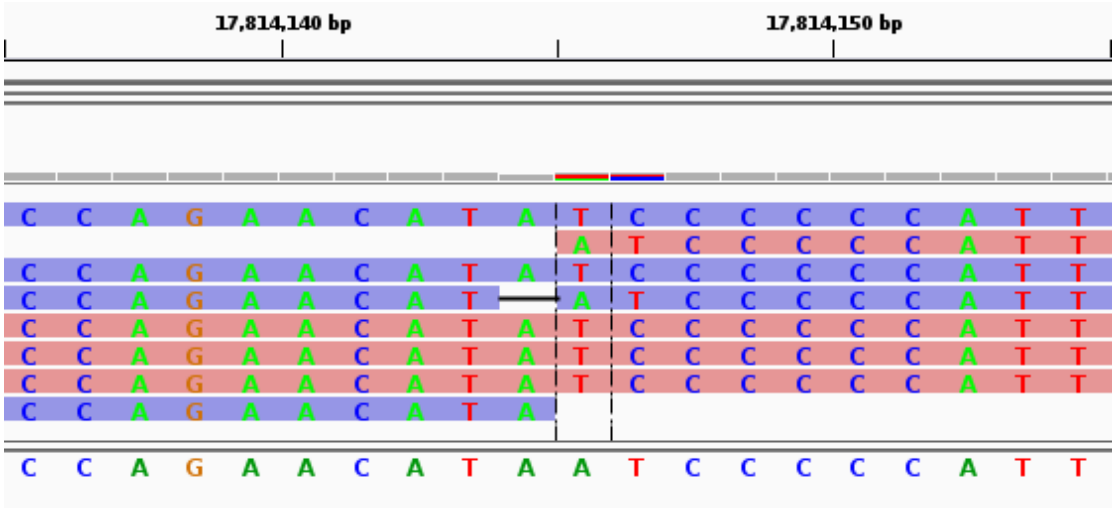


Supplementary Figure S8. The 4-way Venn diagrams of the SNPs called by the four programs (top) and also found in dbSNPs (bottom).

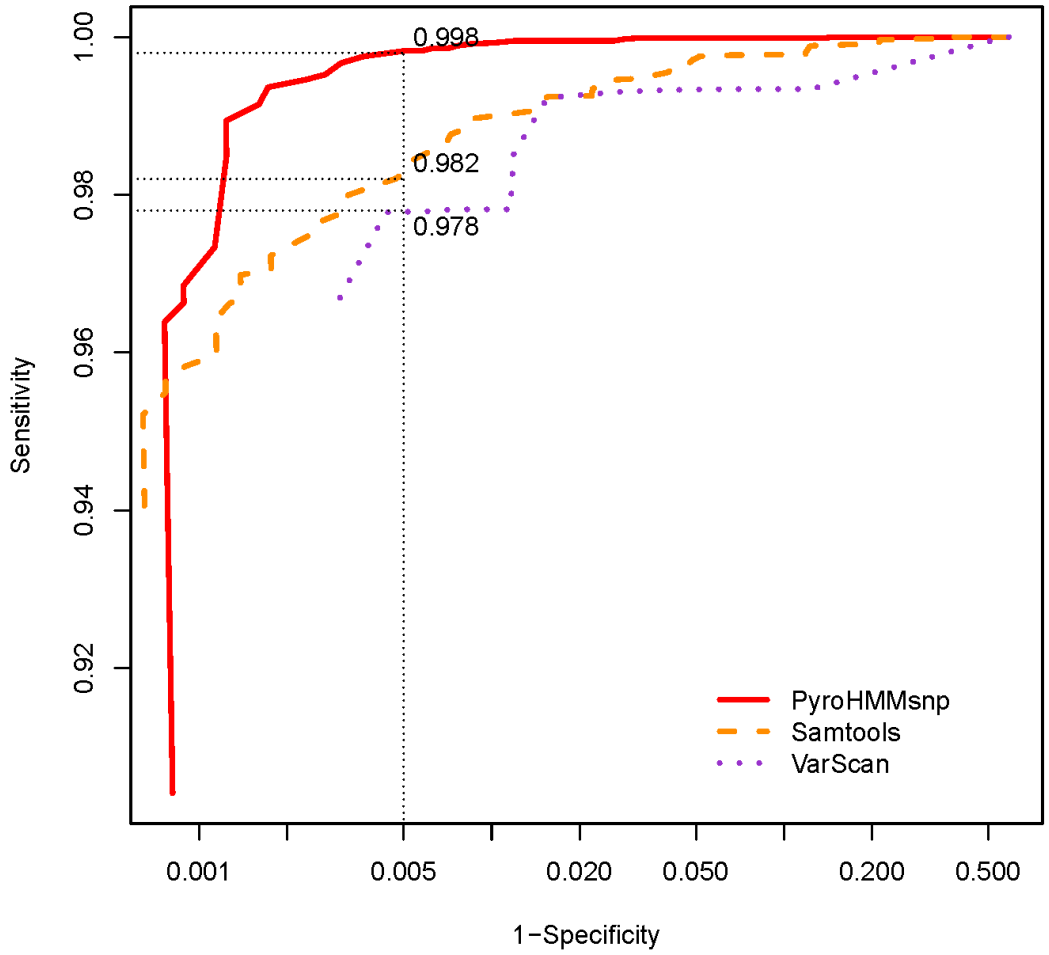
A.



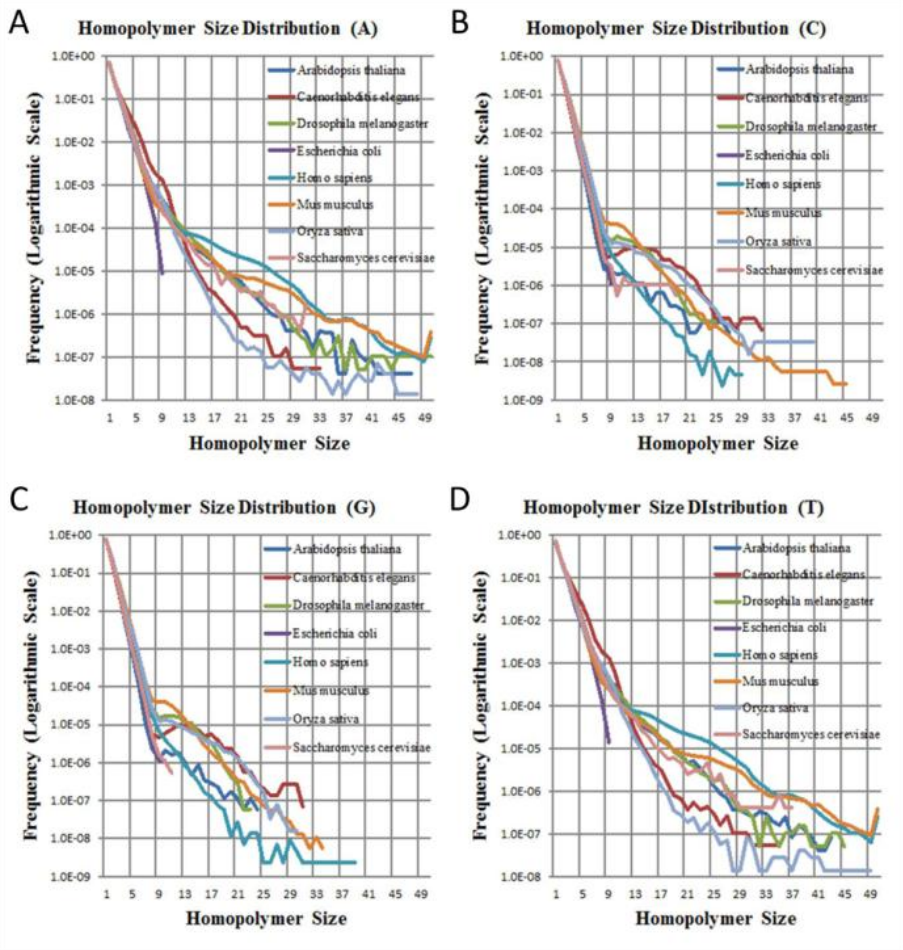
B.



Supplementary Figure S9. Examples that show annotated SNPs called by Samtools but not by PyroHMMsnp. (A) The dbSNP rs2316543. (B) The dbSNP rs544597 and rs658674.



Supplementary Figure S10. The sensitivity-versus-specificity (log-scaled) plot of SNP-calling accuracies for haploid genomes using the 454 resequencing data.



Supplemental Figure S11. The empirical distributions of the homopolymer length across the reference genomes of eight model organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli* sub-strain MG1655, *Homo sapiens*, *Mus musculus*, *Oryza sativa*, and *Saccharomyces cerevisiae.*