Supplementary Information:

Ultrasensitive Detection of Rare Mutations using Next-generation Targeted Resequencing

Authors:

Patrick Flaherty[1,5†], Georges Natsoulis[1†], Omkar Muralidharan[3], Mark Winters[4], Jason Buenrostro[1], John Bell[1], Sheldon Brown[6], Mark Holodniy[4,7], Nancy Zhang[3], Hanlee P. Ji[1,2]

† These authors contributed equally to this work.

Supplementary Material and Methods

Description of Binomial model

Supplementary Tables

Supplementary Figures

Binomial Model

The simple binomial probability forms the basis of our hierarchical model and is used to compute sample specific quantities. The number of mutant reads r_i out of n_i total reads in run $i = 1,\ldots,N$ is

$$\Pr\left(r_1,\ldots,r_m|\theta,n_1,\ldots n_m\right)=\prod_{i=1}^{m}\Pr\left(r_i|\theta,n_i\right)=\prod_{i=1}^{m}\mathrm{B}\left(n_i,r_i\right)\theta^{r_i}\left(1-\theta\right)^{n_i-r_i},$$

where $B(x,y)=\binom{x}{y}$.

We estimate the true fraction of the sample containing a mutation, θ, by the maximum likelihood method. The log-likelihood is

$$\log\Pr\left(r_1,\ldots,r_N|\theta,n_1,\ldots,n_N\right)=\sum_{i=1}^{N}\log\mathrm{B}\left(n_i,r_i\right)+\sum_{i=1}^{N}r_i\log\theta+\sum_{i=1}^{N}\left(n_i-r_i\right)\log\left(1-\theta\right).$$

Taking the derivative with respect to the parameter and setting equal to zero gives

$$\frac{\partial\log\Pr\left(r_1,\ldots,r_N|\theta,n_1,\ldots,n_N\right)}{\partial\theta}=\frac{1}{\theta}\sum_{i=1}^{N}r_i-\frac{1}{1-\theta}\sum_{i=1}^{N}\left(n_i-r_i\right)=0.$$

Solving for θ gives $\hat{\theta}_{Bino\mathrm{MLE}}=\frac{\sum_{i=1}^{N}r_i}{\sum_{i=1}^{N}n_i}$, the maximum likelihood estimate.

Under the assumption that the read depth for all runs is equal, $n=n_i\ \forall i$, the MLE reduces to

$\hat{\theta}_{Bino\mathrm{MLE}}=\frac{\sum_{i=1}^{N}r_i}{Nn}=\frac{\bar{k}}{n}$, where $\bar{k}$ is the average across runs.

Hierarchical Beta-Binomial Model

Since the number of runs per sample, N, may be quite small we use a hierarchical model to incorporate information from adjacent positions to estimate the null model. The hierarchical model (Supplementary Figure SS10) is

$$\boldsymbol{r_{ij}}|\boldsymbol{\theta_{ij}},\boldsymbol{n_{ij}}\sim\mathrm{Binomial}\left(\boldsymbol{\theta_{ij}},\boldsymbol{n_{ij}}\right)$$

$$\theta_{ij}|\mu_j,M\sim\mathrm{Beta}\left(\mu_j,M\right)$$

The corresponding probability distribution functions are

$$\Pr(r_{ij}|\theta_{ij}, n_{ij}) = \binom{n_{ij}}{r_{ij}} {\theta_{ij}}^{r_{ij}} (1-\theta_{ij})^{n_{ij}-r_{ij}}$$

$$\Pr(\theta_{ij}|\mu_j, M) = B(\mu_j M, (1-\mu_j)M) {\theta_{ij}}^{\mu_j M-1} (1-\theta_{ij})^{(1-\mu_j)M-1}$$

We have used the mean, sample-size parameterization of the Beta distribution rather than the standard scale parameterization in order to use a common M for all positions. The 1-1 conversion is $\alpha_j = \mu_j M$ and $\beta_j = (1-\mu_j)M$.

Maximum Likelihood Estimate for the Beta-Binomial Model

The complete data likelihood of the Beta-Binomial model is

$$\Pr(r, \theta|\mu, M) = \prod_{i=1}^{N}\prod_{j=1}^{J} \Pr(r_{ij}, \theta_{ij}|\mu_j, M, n_{ij}) = \prod_{i=1}^{N}\prod_{j=1}^{J} \Pr(r_{ij}|\theta_{ij}, n_{ij}) \Pr(\theta_{ij}|\mu_j, M)$$

Therefore, the complete data log-likelihood is

$$\ell_c(r, \theta|\mu, M, n) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log \Pr(r_{ij}|\theta_{ij}, n_{ij}) + \log \Pr(\theta_{ij}|\mu_j, M)$$

$$\begin{aligned}\ell_c(r, \theta|\mu, M, n) = \sum_{i=1}^{N}\sum_{j=1}^{J} & \log \Gamma(n_{ij}+1) - \log \Gamma(r_{ij}) - \log \Gamma(n_{ij}-r_{ij}) + r_{ij} \log \theta_{ij} \\ & + (n_{ij}-r_{ij}) \log(1-\theta_{ij}) + \log \Gamma(M) \\ & - \log \Gamma(\mu_j M) - \log \Gamma\left((1-\mu_j)M\right) + (\mu_j M - 1) \log \theta_{ij} + \Big((1-\mu_j)M \\ & - 1\Big) \log(1-\theta_{ij})\end{aligned}$$

Our data set provides observations on r_{ij} for all positions $j = 1, \dots, J$ and all replicates $i = 1, \dots, N$, but θ_{ij} is unobserved. The log-likelihood is then

$$\ell(r|\mu, M) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log \int \Pr(r_{ij}|\theta_{ij}) \Pr(\theta_{ij}|\mu_j, M) d\theta_{ij}$$

The logarithm cannot move through to the probability distributions and we are forced to compute $N \times J$ integrals in order to compute the log-likelihood.

We use the EM algorithm to compute a local MLE for the parameters. The algorithm alternates between computing the expected value of the unobserved variable θ and maximizing the likelihood with respect to the parameters $\phi = \{\mu, M\}$.

E-Step: Setting the derivative of the complete log-likelihood with respect to θ to zero gives

$$\frac{\partial \ell_c}{\partial \theta_{ij}} = \frac{r_{ij}}{\theta_{ij}} - \frac{n_{ij} - r_{ij}}{1 - \theta_{ij}} + \frac{(\mu_j - 1)}{\theta_{ij}} - \frac{\left((1 - \mu_j)M - 1\right)}{1 - \theta_{ij}} = 0$$

The update for θ is then

$$\hat{\theta}_{ij} = \frac{r_{ij} - \mu_j M - 1}{M + n_{ij} - 2}$$

M-step: Setting the derivative of the complete log-likelihood with respect to $\phi = \{\mu, M\}$ to zero is done separately for each parameter. The update for M does not have an analytical solution, but an optimization algorithm using the Hessian matrix is possible.

$$\frac{\partial \ell_c}{\partial M} = \sum_{i=1}^{N}\sum_{j=1}^{J} \psi(M) - \mu_j \psi(\mu_j M) - (1 - \mu_j)\psi\left((1 - \mu_j)M\right) + \mu_j \log \theta_{ij} + (1 - \mu_j)\log(1 - \theta_{ij})$$

Simplifying yields

$$\frac{\partial \ell_c}{\partial M} = NJ\psi(M) - N\sum_{j=1}^{J}\left[\mu_j \psi(\mu_j M) + (1 - \mu_j)\psi\left((1 - \mu_j)M\right)\right] + \sum_{i=1}^{N}\sum_{j=1}^{J}\left[\mu_j \log\left(\frac{\theta_{ij}}{1 - \theta_{ij}}\right) + \log(1 - \theta_{ij})\right]$$

The second derivative is

$$\frac{\partial^2 \ell_c}{\partial M^2} = \sum_{i=1}^{N}\sum_{j=1}^{J} \psi_1(M) - {\mu_j}^2 \psi_1(\mu_j M) - (1 - \mu_j)^2 \psi_1\left((1 - \mu_j)M\right)$$

The derivative of the complete log-likelihood with respect to μ_j is

$$\frac{\partial \ell_c}{\partial \mu_j} = \sum_{i=1}^{N} M\left(\psi\left((1 - \mu_j)M\right) - \psi(\mu_j M)\right) + M \log\left(\frac{\theta_{ij}}{1 - \theta_{ij}}\right)$$

The second derivative is

$$\frac{\partial^2 \ell_c}{\partial {\mu_j}^2} = -NM^2\left(\psi_1\left((1-\mu_j)M\right) + \psi_1(\mu_j M)\right)$$

EM Algorithm

1. Initialize $\phi^0 = \{\mu^0, M^0\}$ to $\hat{\mu}_j^0 = \frac{\sum_{i=1}^{N} r_{ij}}{\sum_{i=1}^{N} n_{ij}}$, $\widehat{M}^0 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J} n_{ij}}{NJ}$
2. Begin EM iterations, increment loop counter (i)
    a. E-step: $\hat{\theta}_{ij}{}^{i} = \frac{r_{ij} - \widehat{\mu_j}^{i-1}\widehat{M}^{i-1} - 1}{\widehat{M}^{i-1} + n_{ij} - 2}$
    b. M-step:
        i. Update M: $\widehat{M}^i \leftarrow \text{argmax}_M\, \ell_c\left(r, \hat{\theta}^i | \mu, M, n\right)$ s.t. $M \in [0, \infty]$
        ii. Update each μ_j: $\widehat{u_j}^i \leftarrow \text{argmax}_{\mu_j}\, \ell_c\left(r, \hat{\theta}^i | \mu, M, n\right)$ s.t. $\mu_j \in [0,1]$
3. Repeat while change in log-likelihood is large: $\delta\ell_c = \frac{{\ell_c}^i - {\ell_c}^{i-1}}{\left|{\ell_c}^{i-1}\right|} > 1 \times 10^{-4}$

The maximization step is carried out by the interior point algorithm.

Normal approximation Hypothesis Test

Each position was tested whether the reference error rate and the observed error rate were significantly different by a Normal z-test.

Given the parameters estimated from reference read data ($\hat{\phi}_0 = \{\hat{\mu}_0, \widehat{M}_0\}$), the average read depths for the reference data (n_0), and r, n observed data for the sample.

Compute the null distribution standard deviation for the Beta-Binomial model (iterated method of moments)

$$\widehat{var_0}\left(\frac{r}{n}\right) = \hat{\sigma}_0^2 = \frac{\hat{\mu}_0(1-\hat{\mu}_0)}{n_0}\left(1 + \frac{n_0 - 1}{\widehat{M}_0 - 1}\right)$$

Compute the z-statistic for the observed sample data at that position, j,

$z_j = \frac{\frac{r_j}{n_j} - \hat{\mu}_{j0}}{\hat{\sigma}_{j0}}$.

Compute the associated p-value from the normal distribution and the test statistic.

The power curve estimates were derived by the same method using the iterated method of moments estimate for the alternative hypothesis as well.

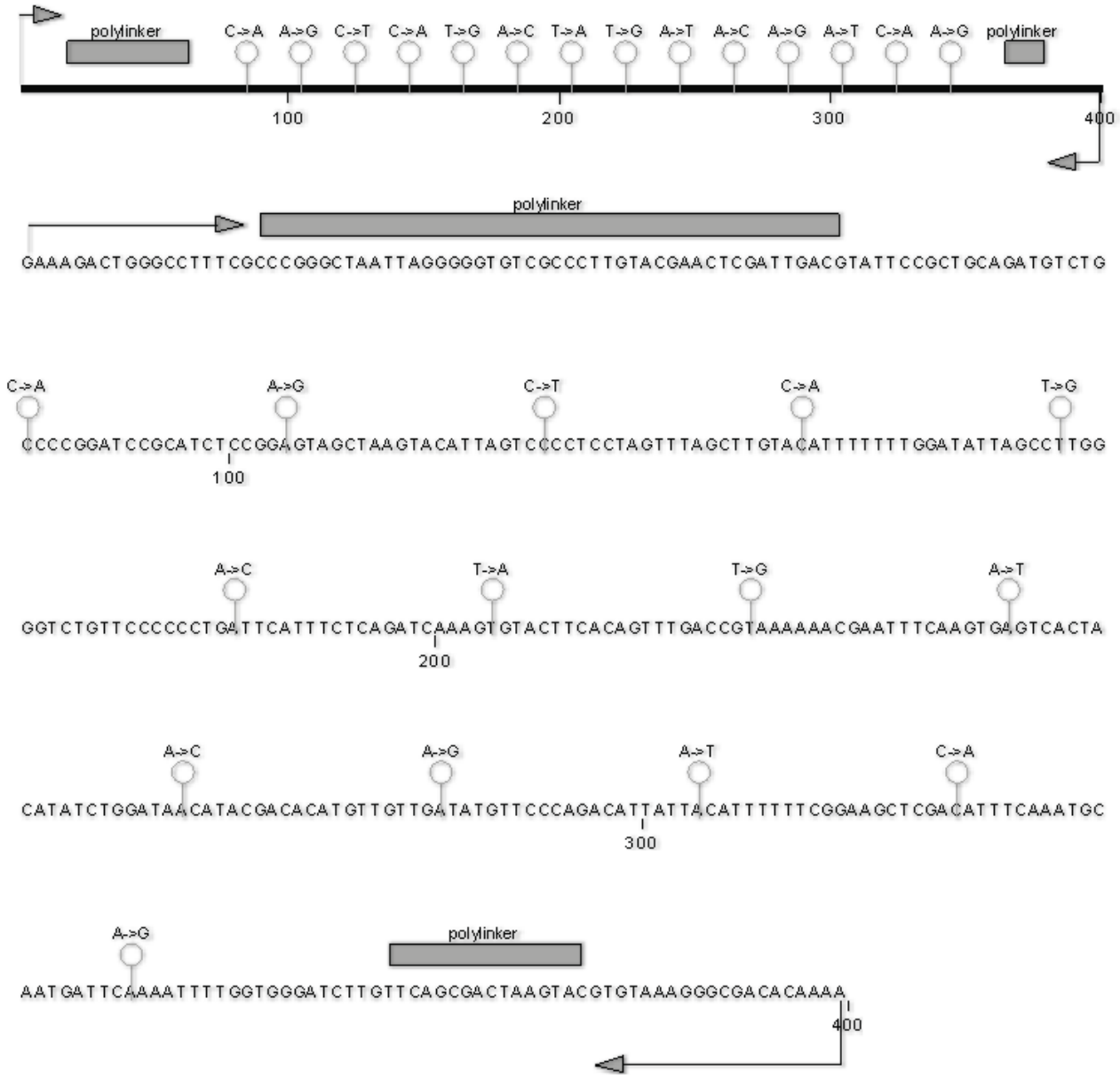
Supplementary Table S1. Indexing accuracy assessment.

Amplicon number	Barcode index sequence for Read 1 and 2	Number of Mapped Mate Pair Reads		
		Correct barcode on both Read 1 and Read 2	Correct amplicon for both Reads 1 and 2 barcode	Incorrect amplicon for both Reads 1 and 2 barcode
1	'AAT'	1,664,047	1,621,535	72
2	'ACT'	1,112,980	1,081,317	419
3	'AGT'	991,500	197,819	286
4	'ATT'	984,946	872,663	543
5	'CAT'	785,230	751,807	467
6	'CCT'	967,119	901,199	663
7	'CGT'	1,018,853	746,441	666
8	'CTT'	969,806	720,371	509
9	'GAT'	1,981,114	1,868,341	1,081
10	'GCT'	1,214,134	827,694	1,641
11	'GGT'	1,289,662	1,120,478	898
12	'GTT'	1,260,348	1,224,791	392
13	'TAT'	1,369,396	1,203,679	646
14	'TCT'	1,061,735	984,029	864
15	'TGT'	1,651,456	1,384,276	1,012
16	'TTT'	1,155,397	804,634	1,096
Total Mapped Mate Pair Reads		19,477,723	16,311,074	11,255

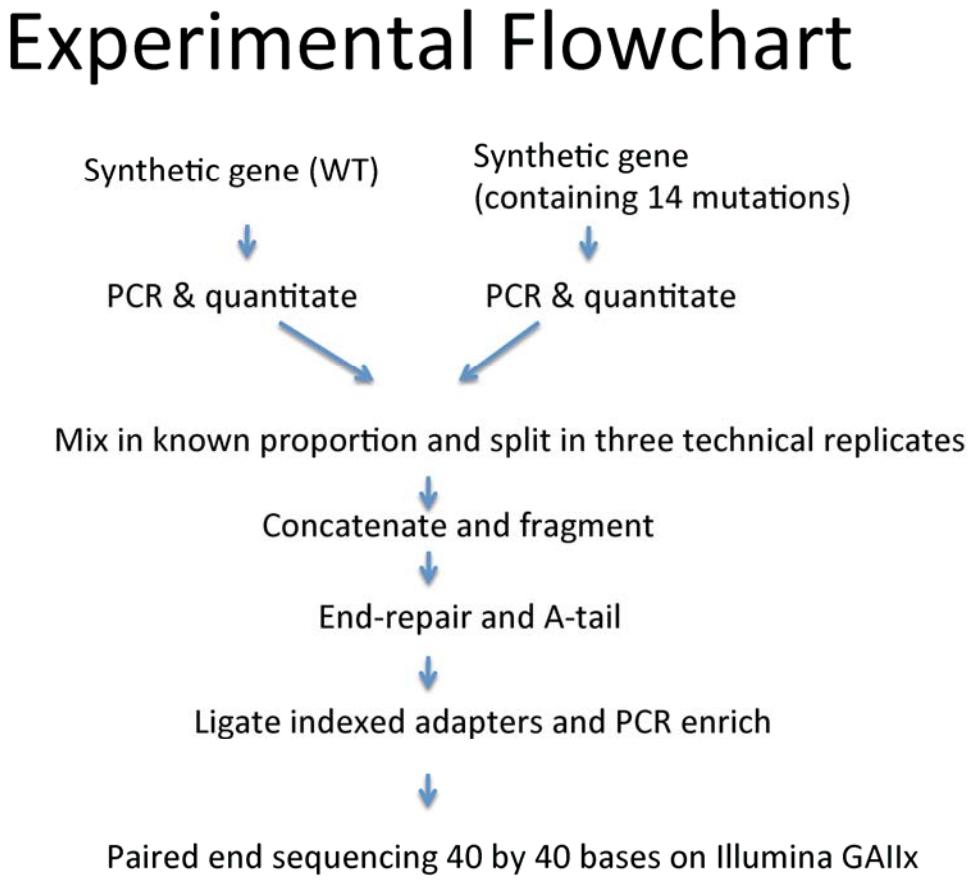
Supplementary Table S2. H1N1 variants.

		B23	B23 5X Dilution	B23 25X Dilution	BN1 (rep1)	BN1 (rep2)	BN3	BN4	BN5	BN6	BN7	BN8	BN9	ref1	ref2	ref3	Mutation average
Mutations with 0.1% sample fraction or greater	Lane 1	74	4	5	23	23	24	34	60	50	33	41	40	0	1	0	40
	Lane 2	87	6	7	25	21	26	42	62	51	31	51	54	0	0	0	45
	Consensus for both lanes	60	4	4	21	17	18	32	36	37	24	40	35	0	0	0	32
Non-synonymous mutations	Lane 1	50	3	2	13	14	13	17	43	32	21	23	27	0	1	0	25
	Lane 2	65	6	3	15	13	16	24	42	35	18	31	35	0	0	0	29
	Consensus for both lanes	42	2	2	13	11	10	16	25	23	15	23	24	0	0	0	20

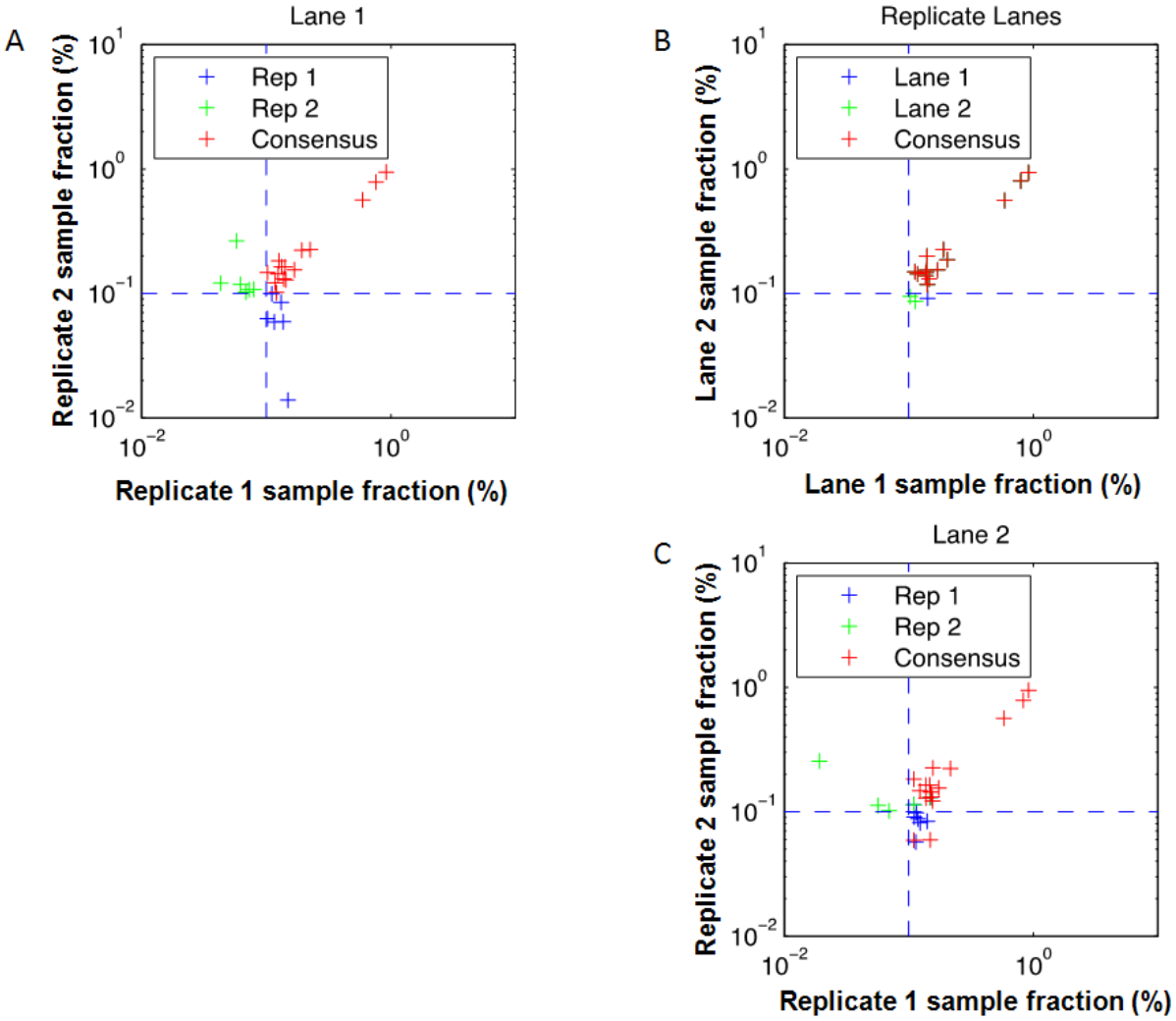
Supplementary Figure S1. Synthetic gene sequence. The synthetic gene reference is shown and the companion sequence with 14 known mutant positions is shown marked on the sequence. A polylinker sequence facilitates cloning into a DNA vector.



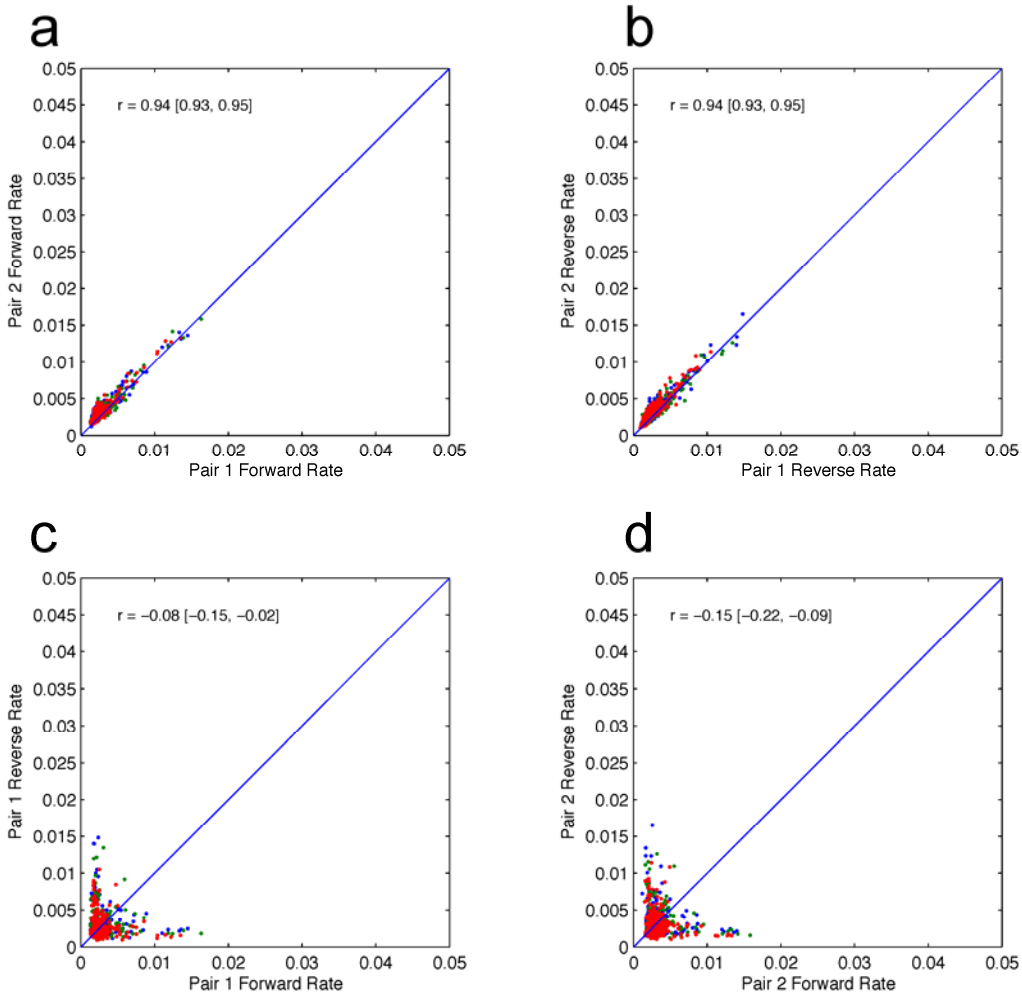
Supplementary Figure S2. Sequencing experiment design. The overall design of the synthetic sequence mutation analysis is outlined. We used this experiment to create known sample fraction admixtures of the mutant to the reference synthetic sequence.



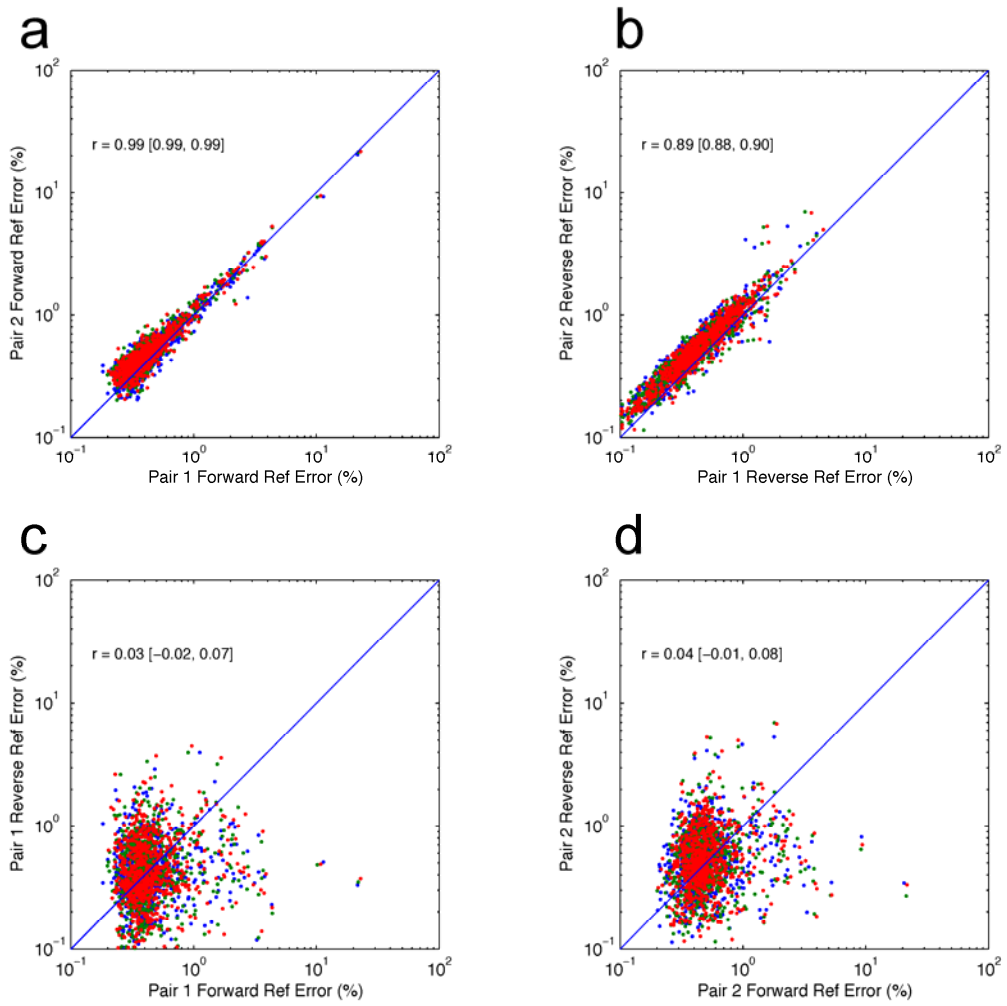
Supplementary Figure S3. Scatter plots showing the intra-lane and inter-lane reproducibility. One sample, BN1, was run twice on lane 1 and twice on lane 2 allowing us to compare called variant positions within lanes and between lanes. Each replicate of the reference is coded in a different color: red, blue, green. (A, C) Within lanes, the calls are reproducible for the sample fraction greater than 0.1%. However, the lower values of the variant fractions show some divergence between replicates. (B) A similar observation is present when comparing positions that are called in both replicates within individual lanes. For variant fractions greater than 0.1%, the calls are reproducible, but below that detection limit, the calls are less reproducible.



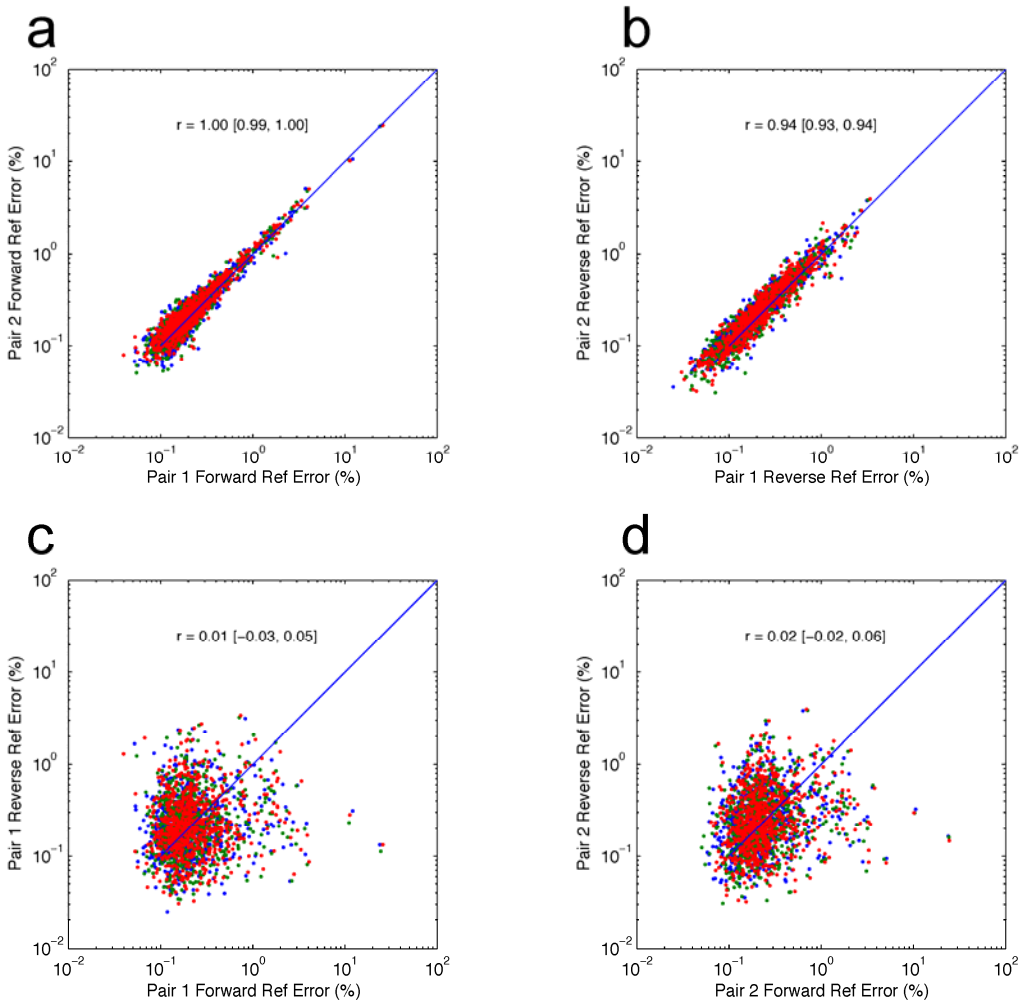
Supplementary Figure S4. Synthetic sequence data comparison of error rates between strands and matched pairs. The correlation between paired ends is high for both sequence reads in the forward direction and reverse direction (a and b). There is no correlation between forward and reverse reads for the first in the pair sequence (c) as well as forward and reverse reads in the second pair (d). Each replicate of the reference is coded in a different color: red, blue, green.



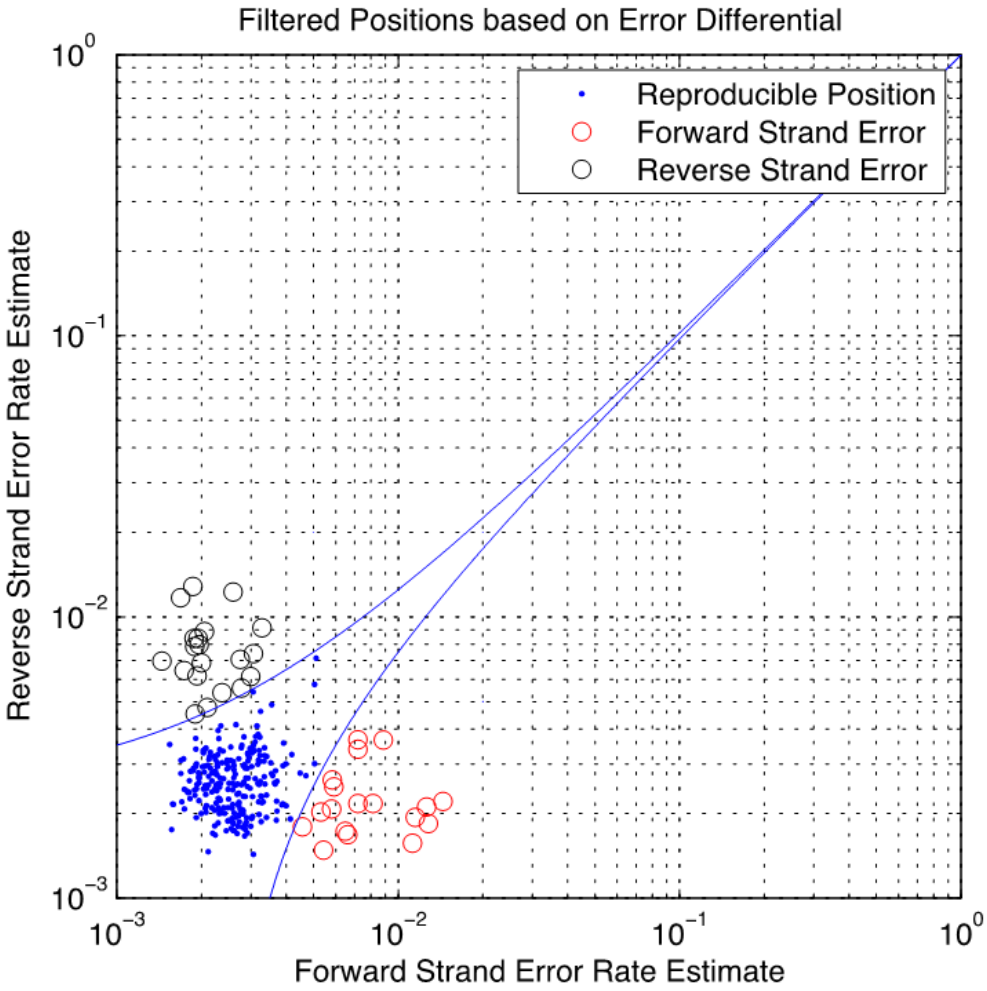
Supplementary Figure S5. Clinical sample error rates between strands and matched pairs lane 1. The correlation between paired ends is high for both sequence reads in the forward direction and reverse direction (a and b), similar to the synthetic DNA data. There is no correlation between forward and reverse reads for the first in the pair sequence (c) as well as forward and reverse reads in the second pair (d). Each replicate of the reference is coded in a different color: red, blue, green.



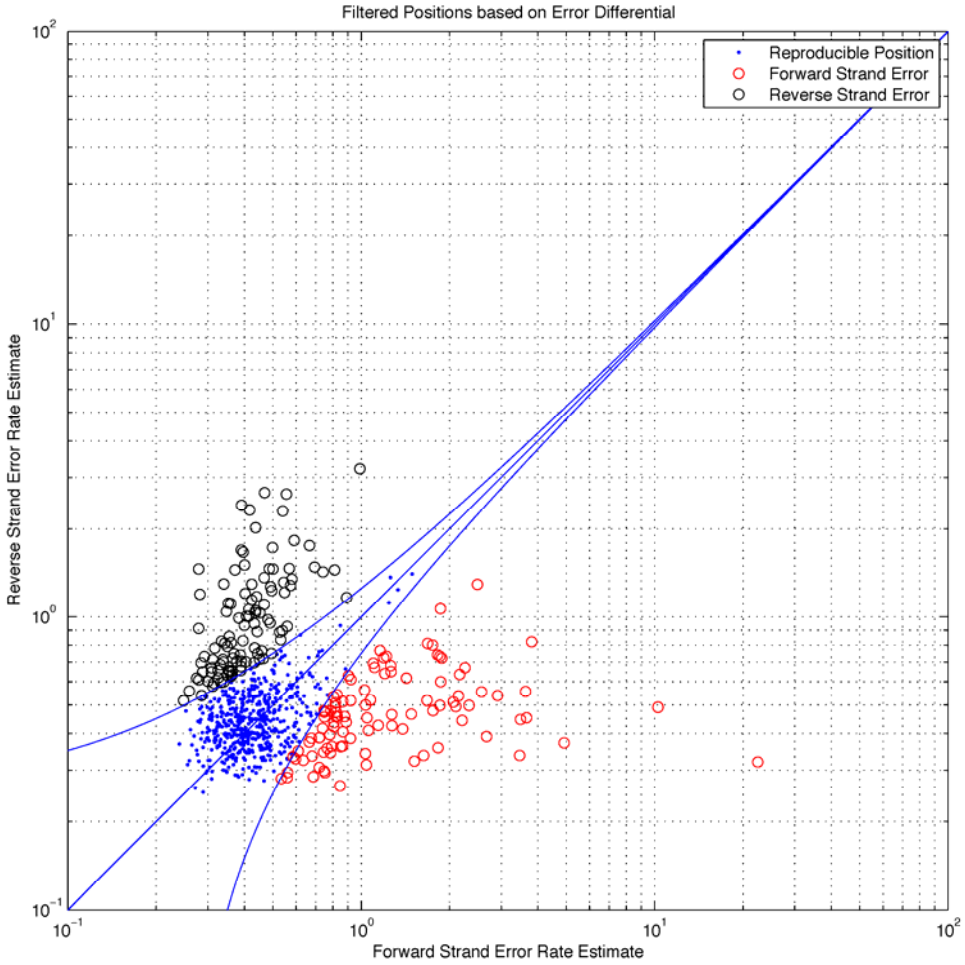
Supplementary Figure S6. Clinical sample error rate differential between forward direction reads and reverse direction reads for lane 2. Positions called error-prone in the forward direction (red) or reverse direction (black) are filtered. The error-prone direction reads are removed from the data set and only the lower error rate direction reads are retained for that position. The correlation between paired ends is high for both sequence reads in the forward direction and reverse direction (A and B), similar to the synthetic DNA data. Each replicate of the reference is coded in a different color: red, blue, green. There is no correlation between forward and reverse reads for the first in the pair sequence (C) as well as forward and reverse reads in the second pair (D). Each replicate of the reference is coded in a different color: red, blue, green.



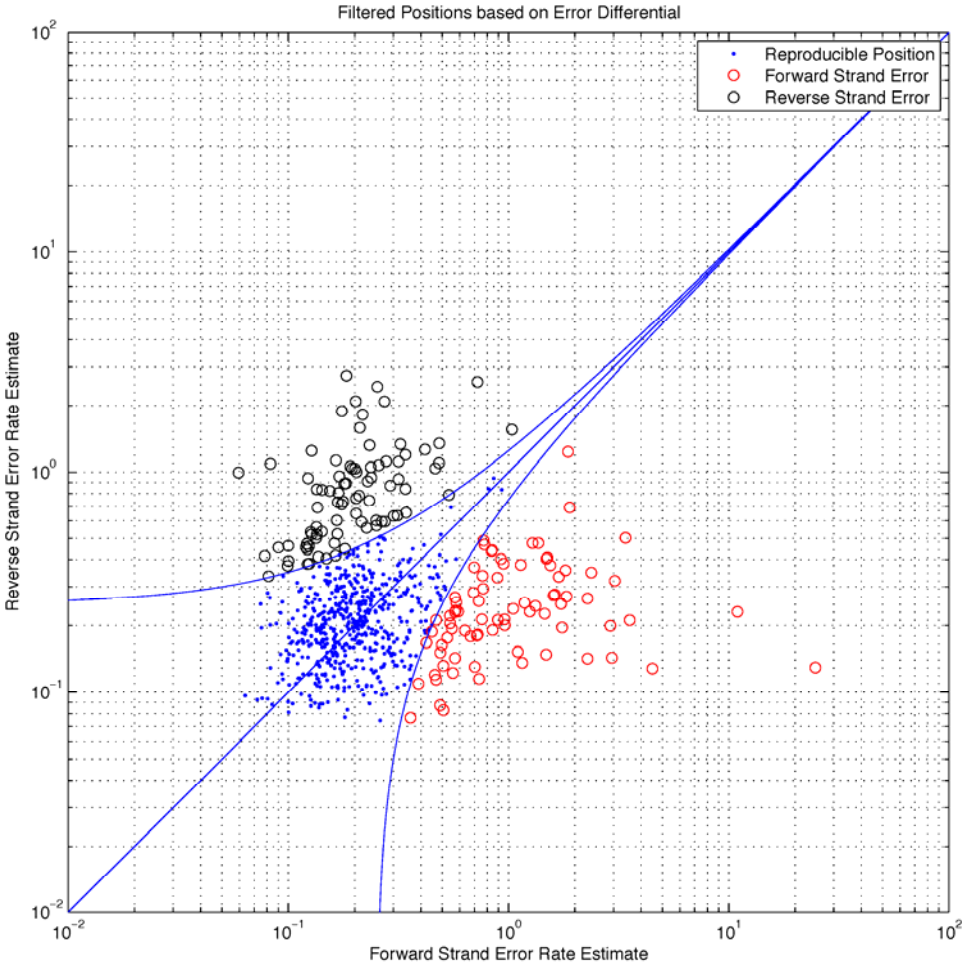
Supplementary Figure S7. Synthetic reference positions are filtered based on error rate differential between forward direction reads and reverse direction reads. Positions called error-prone in the forward direction (red) or reverse direction (black) are filtered. The error-prone direction reads are removed from the data set and only the lower error rate direction reads are retained for that position.



Supplementary Figure S8. Clinical sample reference lane 1 data positions are filtered based on error rate differential between forward direction reads and reverse direction reads. Positions called error-prone in the forward direction (red) or reverse direction (black) are filtered. The error-prone direction reads are removed from the data set and only the lower error rate direction reads are retained for that position.



Supplementary Figure S9. Clinical sample reference lane 2 data positions are filtered based on error rate differential between forward direction reads and reverse direction reads. Using Lane 2 data, positions called error-prone in the forward direction (red) or reverse direction (black) are filtered. The error-prone direction reads are removed from the data set and only the lower error rate direction reads are retained for that position.



Supplementary Figure S10. Graphical model representation of the Beta-Binomial model. In the graphical model, nodes represent random variables and edges indicate functional relationships. An edge from one node to another indicates that the distribution of the variable the arrow is pointing to is conditionally dependent on the node the arrow is pointing from. A shaded node indicates the random variable is observed. Unobserved nodes are not shaded and nodes with a filled circle inside are parameters. Parameters can be considered random variables with diffuse (improper) priors. The square plate surrounding nodes indicates the random variables contained within are replicated the number of times indicated within the plate and are exchangeable. This graphical model is hierarchical; theta~Beta (mu,M) and k~Binomial (n,theta).