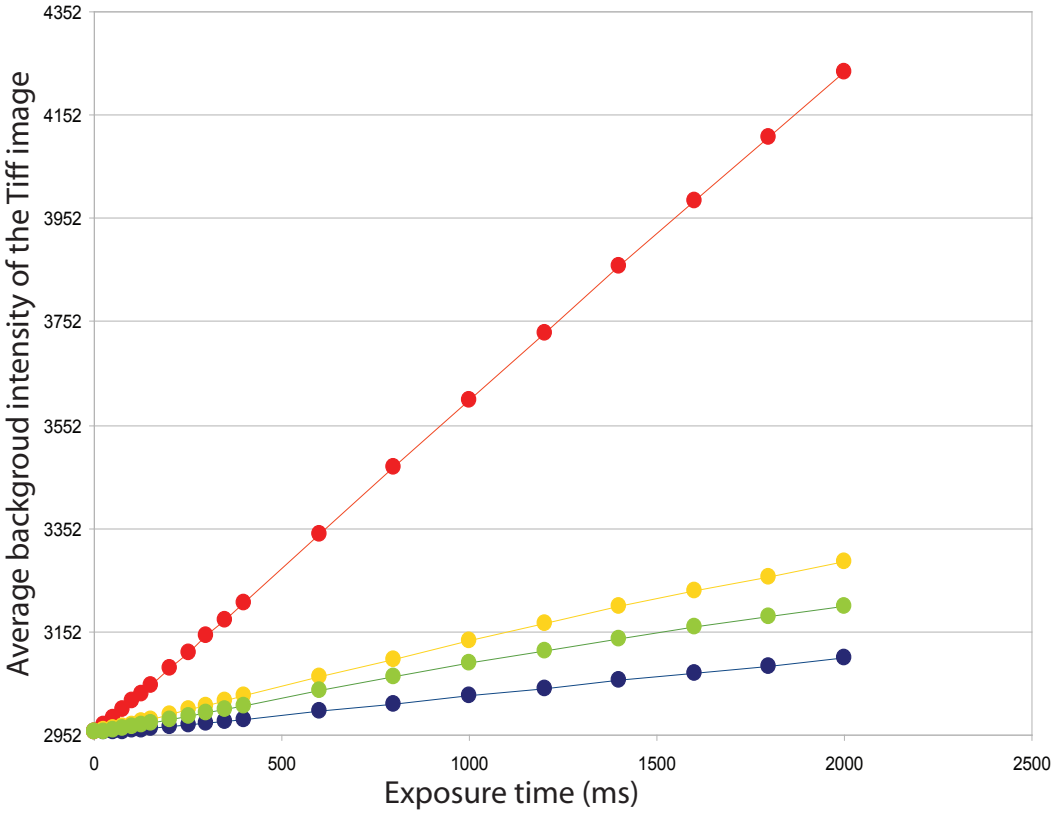
nature | methods

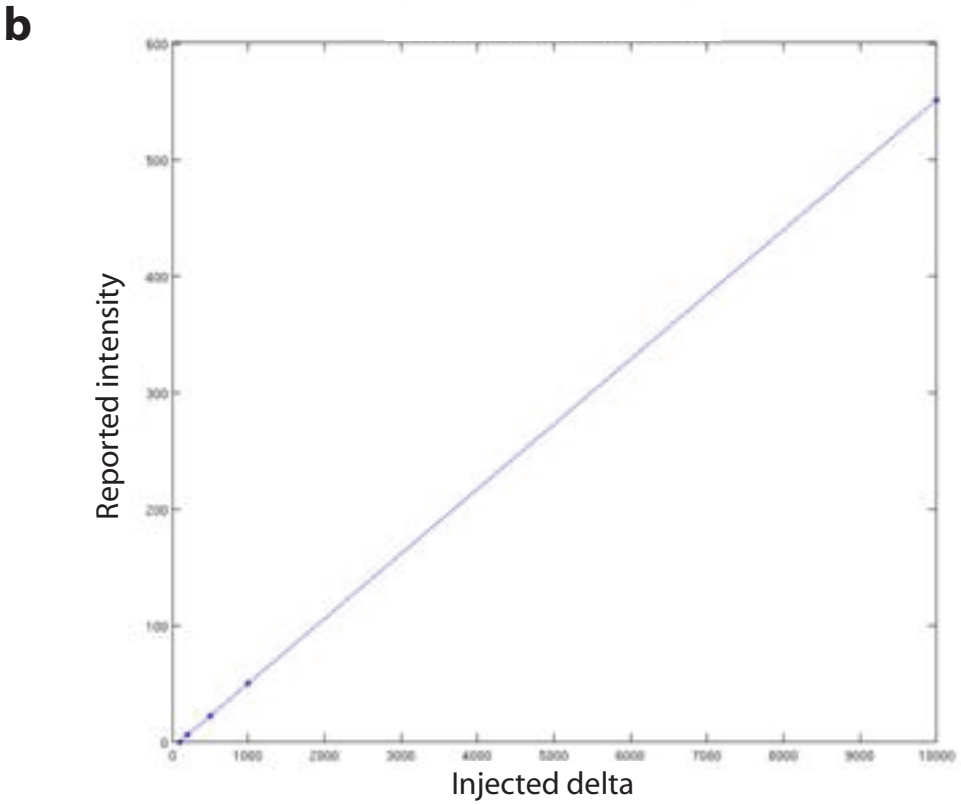
Alta-Cyclic: a self-optimizing base caller for next-generation sequencing

Yaniv Erlich, Partha P Mitra, Melissa delaBastide, W Richard McCombie & Gregory J Hannon

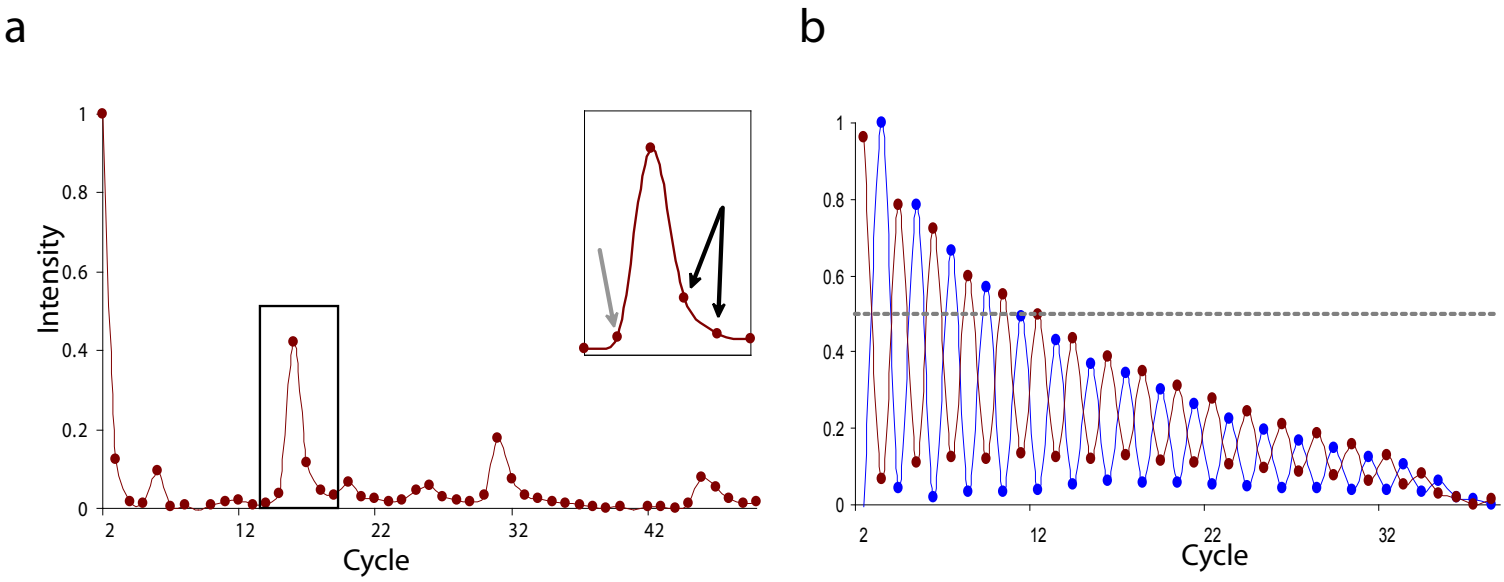
Supplementary figures and text:

Supplementary Figure 1

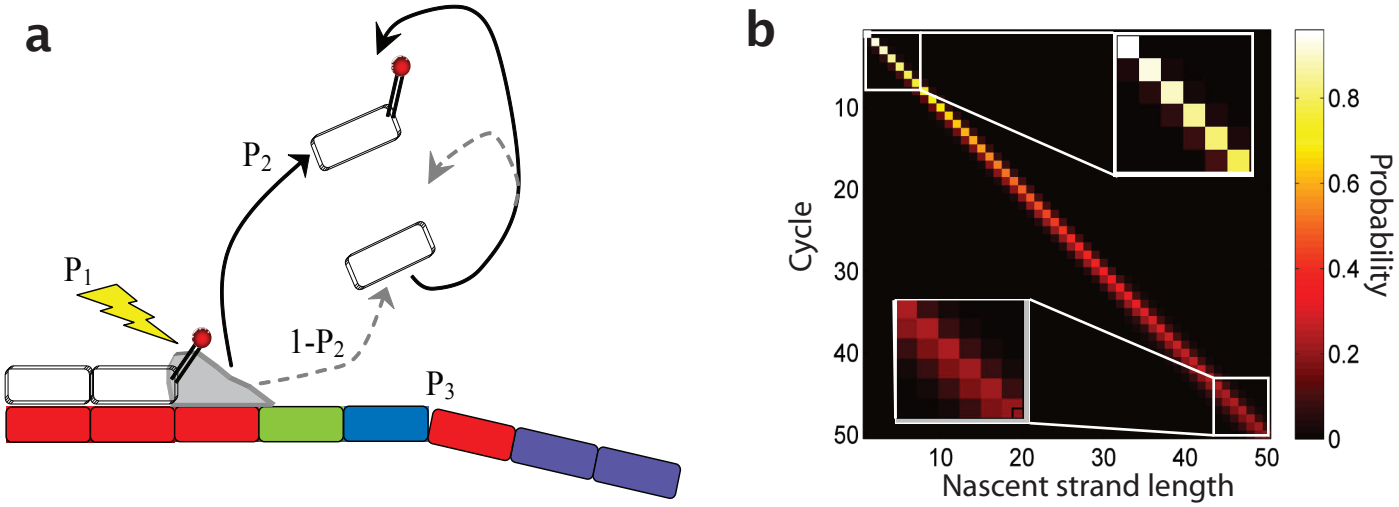




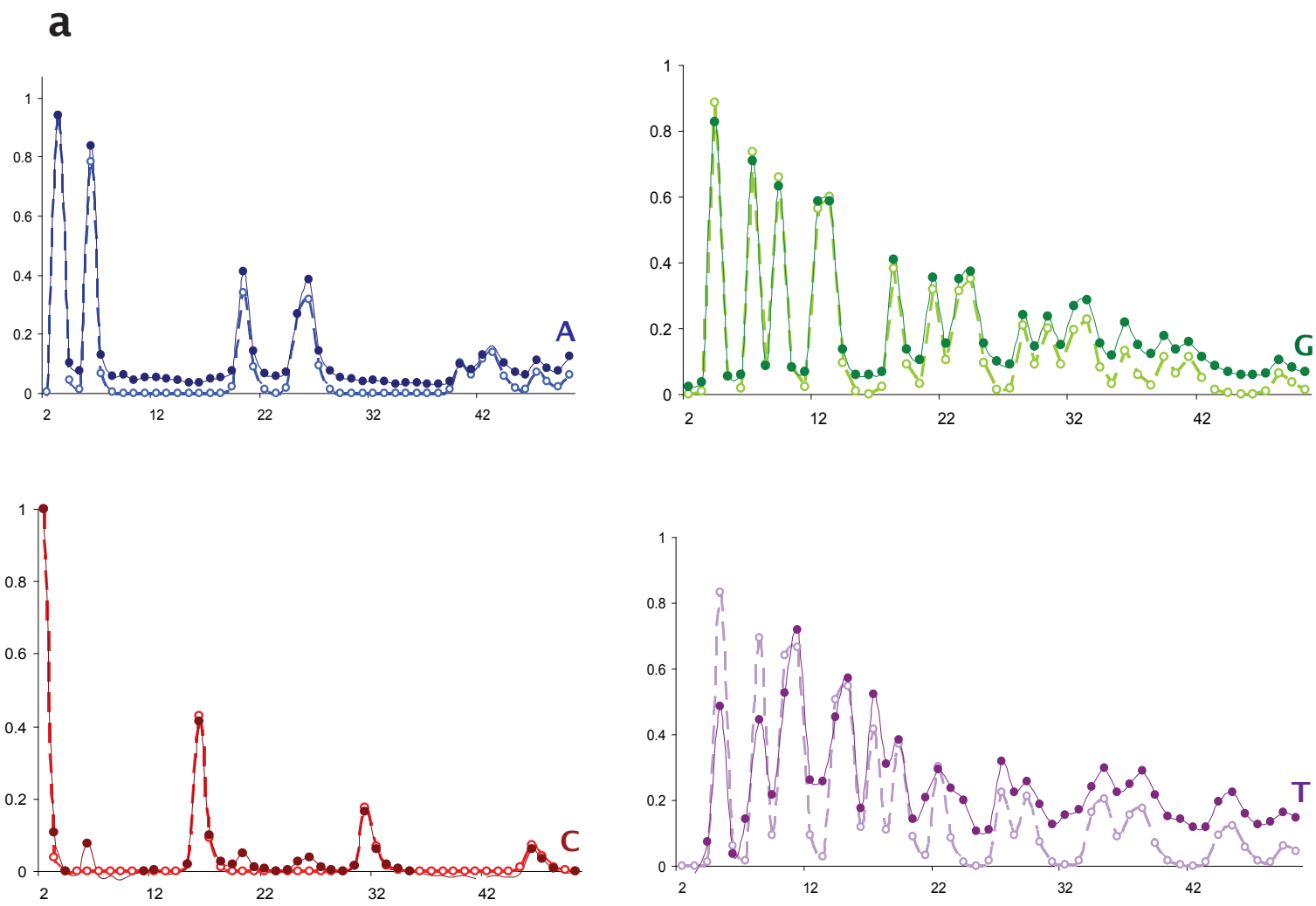
Supplementary Figure 1. Illumina image processing is linear. (a) The optic chain is linear. A series of images with different exposure times were taken from an intact flow cell, and the background intensity was measured for all the channels (A – blue, C – red, G – yellow, T – green). The linear relationship between the input and the output indicates a linear optic chain. (b) Firecrest applies a linear transformation to the image. Artificial TIFF images were created with constant background and a few brighter pixels (spatial delta functions) with different brightness levels. The images were processed by Firecrest and the reported intensities were compared to the input pixels. Again, the linear relationships between the input and output indicate linear processing.



Supplementary Figure 2. Output of the impulse response test. (a) Shown are the averaged intensities of the cytosine channel from DNA clusters with the delta function sequence: GCAGTAGTGTTGGTTCTG-TAGTGGAATGTGCGGTTGTTGAGAATTCAGTA (the first cycle is not shown), after crosstalk correction and normalization. Phasing appears as an anticipation signal that precedes the position of the C in the sequence (gray arrow) and persists in subsequent cycles (black arrows). The diffusive properties of the phasing are shown by relative increase in the residual signal in adjacent cycles to the actual C position. (b) Signal decay (fading) is reflected in intensity reads from microsatellite sequences. Shown are the average intensities of the cytosine (red) and adenine (blue) channels from DNA clusters with the microsatellite sequence ACAC… (the first cycle is not shown) after crosstalk correction and normalization. In the absence of fading the signals should converge to half of their initial intensities (gray line). Nevertheless, the signal exponentially decays toward zero, which indicates material loss or another mechanism that gradually disrupts the signal.

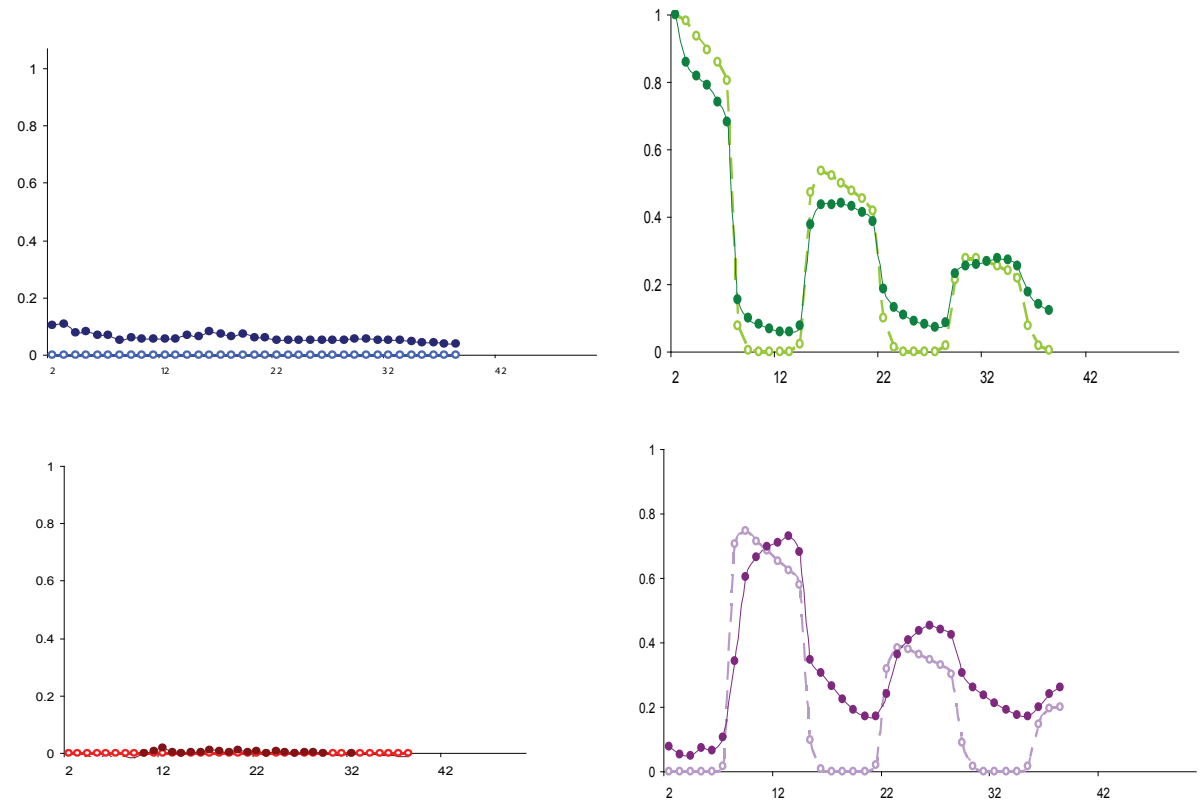


Supplementary Figure 3. Random walk model. (a) A schematic illustration of the random walk model is shown. In the initial state the last nucleotide of the nascent DNA strand (white rectangles) is blocked and has a fluorophore label (red ball). The block is removed with probability P1, allowing a nucleotide to be incorporated in the next cycle. We assume two types of nucleotides in the mix: blocked and contaminating non-blocked species. A blocked nucleotide is incorporated with probability P2, and the non-blocked with probability (1- P2). If a non-blocked nucleotide is incorporated the process continues (gray arrows), until a blocked nucleotide is incorporated (black arrows). In addition, the template is lost with P3 probability, due to strand breakage or another processes. (b) Heat map representation of the R matrix calculated by our random walk model with P1=0.98, P2=0.99 and P3=0.01 for 50 cycles. The rows correspond to sequencing cycles and columns to possible nascent strand lengths. The color indicates the probability. In an idealized situation, the diagonal would be white and all other cells would be black. Notably, in the first cycles (upper left corner) there is almost no variation in length of the nascent strands, whereas in later cycles (bottom right corner) the variation increases and degrades the correct signal.

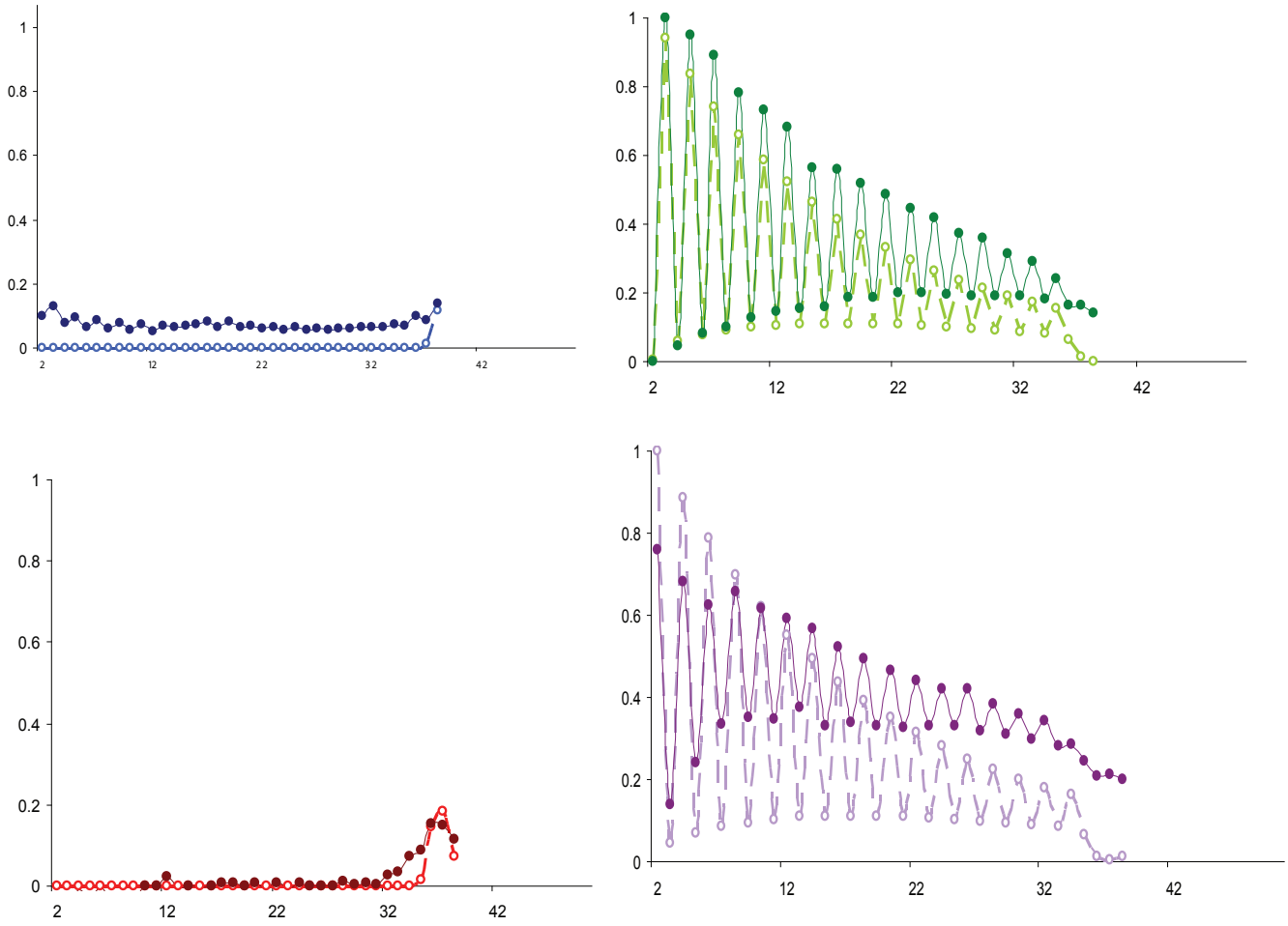


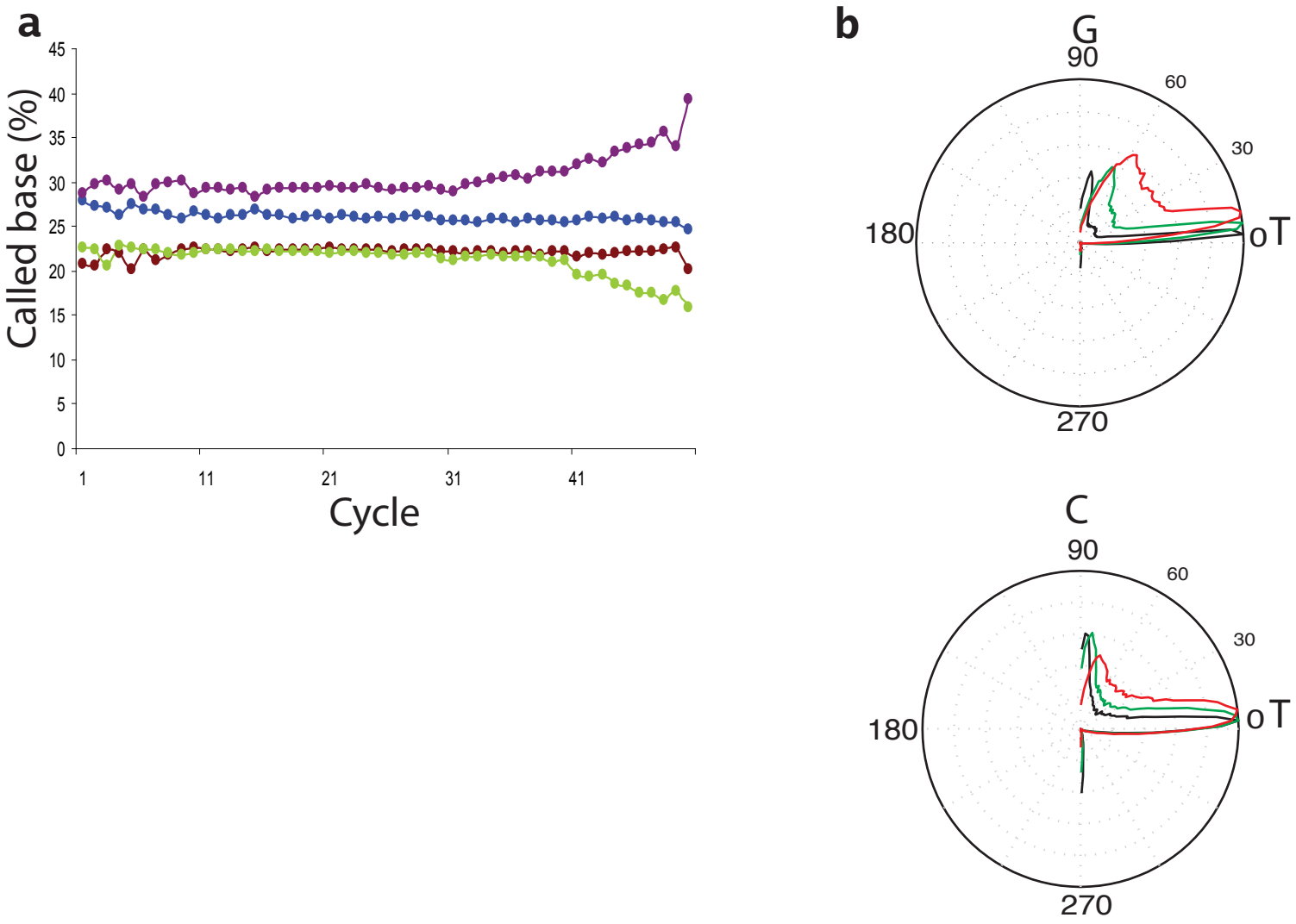
Supplementary Figure 4. The model prediction and the anomaly in the T channel The model prediction (broken lines) is plotted versus the averaged intensity reads (solid lines). The reference data was corrected for fluorophore crosstalk and normalized (A – blue, C –red, G – green, T – purple). (a) Outputs are given for the delta function sequence: GCAGTAGTGTTGGTTCTGTAGTGGAATGTGC GGTTGTTGAGAATTCAGTA (the first cycle is not shown). There is an overall agreement between the model and reference data. Note the T channel anomaly. (b) Shown are the analyses of the theta function sequence: G[7]T[7]G[7]T[7]…(the first cycle is not shown). The intensity of the T channel is very strong in the last cycles where G nucleotides are called. This presumably stems from extensive crosstalk from the G to the T channel. (c) Shown are the analyses of the microsatellite sequence: GTGT…(the first cycle is not shown). Again, the T channel intensities are very strong compared to the prediction. The increase in the intensities of the A and C channels around cycle 38 cycle are because of small deletions in the microsatellite sequences that causes some DNA cluster to report the adaptor sequence in that cycles.

b

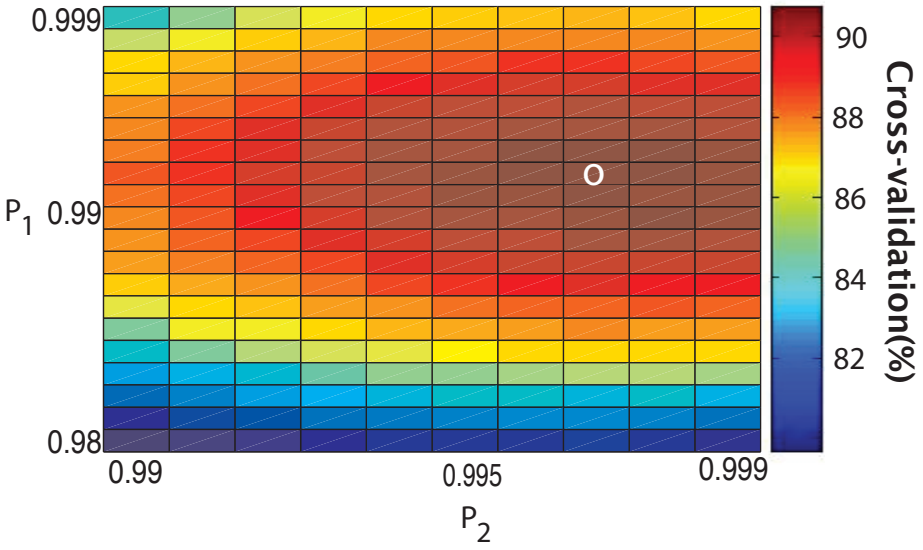


c

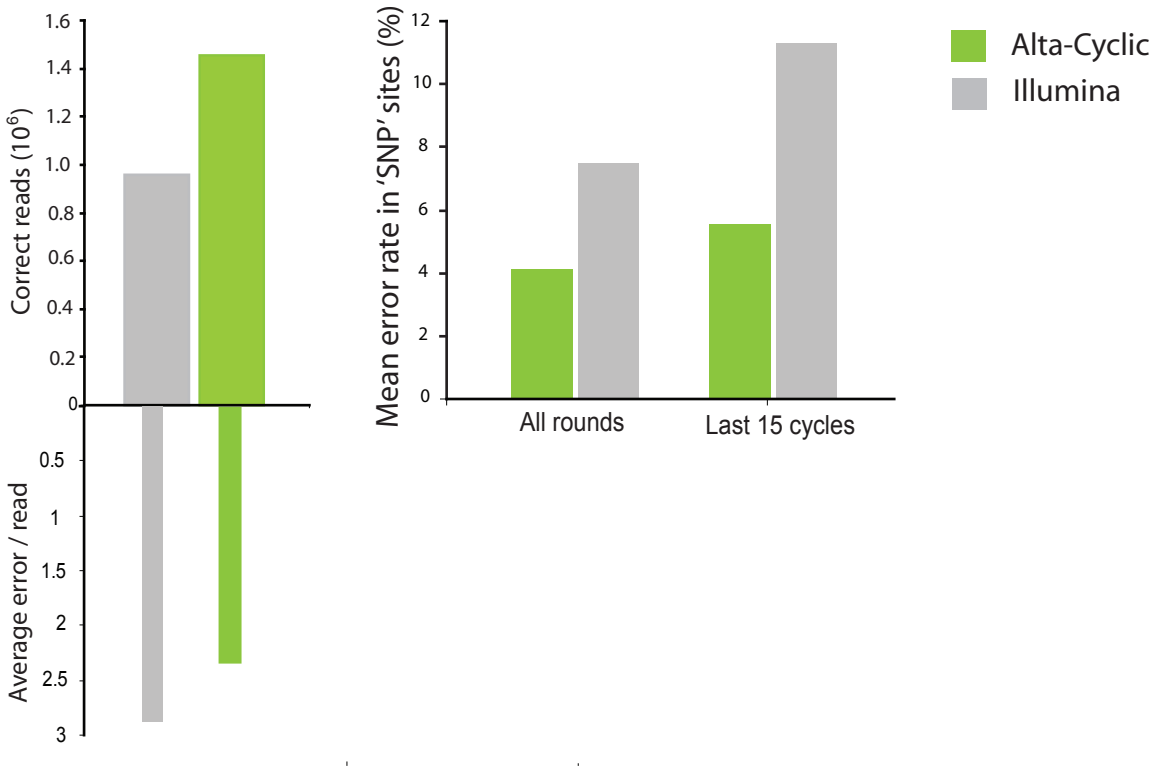




Supplementary Figure 5. Crosstalk matrix changes explain the anomaly. (a) The percentage of called bases in the phi-X library is plotted as a function of cycle number using the Illumina base caller (A – blue, C – red, G – green, T – purple). The T and the G calls have strong opposite trends. A more moderate opposite trend is observed between the C and A calls. (b) Polar histograms present the ratio between channel intensities correlated with the base preference. The upper polar histogram exhibits the ratios between the G and the T channel. A strong G signal with a weak T one will appear in the bins that are close to 90 degrees and the opposite will appear close to zero degrees. In the first cycle (black) the two lobes are orthogonal which indicates correct crosstalk correction. In later cycles (green and red) the G lobe starts to migrate toward the T lobe, which indicates a change in the crosstalk matrix. In contrast, the polar histogram of the ratio between the C and T channels does not exhibit any major crosstalk change. The size of the lobes is increased because of phasing. In both histograms the main peak for each cycle was normalized to value of 1 for clarity.



Supplementary Figure 6. Grid search for optimization of the random walk de-convolution.. The Y-axis corresponds to P1 and the X-axis to P2. The color of each cell indicates the cross-validation rate that was achieved. The white circle shows the values that were chosen



Supplementary Figure 7. Comparison between Alta-Cyclic and Illumina base caller on GAI platform. The improvement by Alta-Cylcic for the phi-X library on 50 cycles on GAI. The left graph shows the improvement in the number of fully correct reads and the reduction of the average error rate. The right graph shows a comparison of error rates at the artificial SNP sites between the Illumina base-caller and Alta-Cyclic.

	1.. ..50
Δ on randm context	CAGTCGGCCGTCGGTATCCTGGTGGTGGCTAGGCTGTCTCTTTCCACGGC GCAGTAGTGTTGGTTCTGTAGTGGAATGTGCGGTTGTTGAGAATTCAGTA CGCCTTACAATTCAAAGTCCATATAACTTTGAATAACCTTACATCGATAT CTAGCCGCGACAACATAGCAGGCACGAGAGTCGACGGACAGCGGATGCGA
Δ on homeo-polymer con-text	ACAAAAAAAAAAAACAAAAAAAAAAAAAAACAAAAAAAAAAAAAACAAAAA AGAAAAAAAAAAAAGAAAAAAAAAAAAAAGAAAAAAAAAAAAAAGAAAAA TCTTTTTTTTTTTTCTTTTTTTTTTTTTTTCTTTTTTTTTTTTTTTCTTTTT TGTTTTTTTTTTTTGTTTTTTTTTTTTTTTGTTTTTTTTTTTTTTTGTTTTT
MS	ACACACACACACACACACACACACACACACACACACATTG GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGCCA
Θ	AAAAAAACCCCCCCCAAAAAAACCCCCCCCAAAAAAACCC GGGGGGGTTTTTTTGGGGGGGTTTTTTTGGGGGGGTTT

Supplementary Table 1. The sequences of the controlled input set. Twelve DNA fragment were synthesized and used as sequencing templates for determination and validation of the noise model. The first four sequences are delta function sequences on a random context. The second group contains delta function sequences on a homopolymer context. The positions of the delta function are underlined. The third set comprises dinucleotide microsatellite sequences (MS) and the last contains the theta function sequences.

Supplementary data

Creating a mathematical model for Solexa non-stationary noise factors

The Illumina GA platform has almost no published literature regarding the exact data processing or the noise factors that are applied or introduced at each stage. Therefore, we started by constructing a comprehensive model describing the signal and non-stationary noise factors. Alta-Cyclic relies on this model, and the training stage is designated to find the best parameters that fit the model to the noise. In the following sections we describe the steps we took in order to developed the model.

Demonstrating linearity of the intensity values

First, we showed that the intensity values that are generated by the platform are outcome of a linear transformation to the number of received photons. We checked whether the machine optic channel is linear – specifically that the reported intensities in the digital tiff images are linearly correlated with the number of received photons. We instructed the sequencer to take images with different exposure times of an intact flow cell to which none of the sequencing chemistry had been applied (see **Supplementary Methods** for recipe). The following exposure times (ms) were used: 1, 25, 50, 75, 100, 125, 150, 200, 250, 300, 350, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000. The captured images were taken from 30 tiles of lanes 4,5 and 7. The median intensity level of every image was determined using the libTiff library and the averaged value from all 30 tiles was calculated. Linear regression of the data found that r^2 for all channels was > 0.995 (**Supplementary Fig. 1a**). Thus, we concluded that the optical channel is linear.

The Illumina analysis pipeline generates intensity files as part of its image analyzer routine, using a program called Firecrest. The intensity files contain the location of each DNA cluster and a measurement of the four signal intensities for each cycle. Critically, the reported intensities are not directly derived from the image, but Firecrest applies an image-processing step. We checked whether Firecrest processing applies a linear transformation to the data. We exploited impulse response analysis to study the transfer function of Firecrest. A series of artificial images with spatial “delta functions” of different heights was generated. We send these images to the analyze_image software that is part of Firecrest, and we used the command line option ‘dump filtered image’ and ‘object list’ option to get back the reported ‘intensity’. We found a linear correlation between the input spike and the reported ‘intensity’ (**Supplementary Fig. 1b**). Thus, we can demonstrate that the intensity values are linear correlated with the number of

received photons. This simplifies our mathematical formalization, and allows us to use the intensity flies as input to our base caller.

Impulse response analysis as a tool for revealing noise factors

Impulse response analysis is a common tool in signal processing to characterize distortions in communication channels. The analysis is performed by injection of a sharp pulse, called delta function, as the input to the system under investigation. The output, measured as a function of time, reflects the system transfer function. The change between the input and output corresponds to the distortion. In sequencing, a delta function corresponds to sequences in which one type of nucleotide dwells in a context of different type of nucleotides. The output is the signal intensities of the corresponding fluorophore as a function of cycle number. For example, the sequences: …AAAACAAAA... or …AGTGCGGTG... are both instances of delta functions with respect to the C channel.

Our impulse analysis was largely designated to characterize the properties of phasing. A delta sequence after phasing will have two imperfections. The first is anticipation signal – an increase in the signal intensity in the fluorophore channel corresponds to the delta function during cycles before actual cycle, which corresponds to that nucleotide position. The second is memory signal – persistence of residual signal in the same channel after calling the nucleotide that corresponds to the delta function. Note, that due to the non-stationary nature of this noise factor, we assume the imperfections will escalate in later cycles. To monitor phasing noise, we created 8 delta sequences. We also created two additional types of sequences: dinucleotide microsatellites, and theta function sequences (**Supplementary Table 1**). Dinucleotide microsatellites, which can also be thought of as a train of delta functions, transmit signal in only two channels that correspond to their alternating constituent bases. Because of the diffusive properties of phasing, the two channels should converge during the run toward half of their initial intensities. The theta function sequences consist of two interleaved short homopolymers – repeats of one nucleotide followed by a repeats of another nucleotide. In total, we synthesized 12 DNA fragments and cloned these as minigenes that contained Illumina sequencing primers and anchors for bridge amplification **(**see **Supplementary Methods**). These were verified by conventional sequencing and excised from their host plasmids prior to loading onto the flow cell. In this way, we avoided any PCR amplification step – prior to bridge amplification itself – that could introduce errors. These templates were sequenced for 50 cycles on a GAI.

For analyzing non-stationary noise factors, we used the sequence files and the intensity files that are generated by the standard Illumina pipeline. We annotated each cluster. Then, we averaged the signal intensities for DNA clusters with the same annotation. This procedure generated twelve 50-by-4

matrices, in which each row corresponds to a cycle and each column to a fluorophore channel. We found that there is extensive crosstalk between the fluorophore channels, reflecting overlapping emission spectra. Specifically, the fluorophore attached to adenine strongly leaked into the cytosine channel, and signal leaked from the guanine channel to the thymidine channel. Using the delta function sequences, the anticipatory and memory signals, which correspond to phasing, were detected, and these escalated with cycle number (**Supplementary Fig. 2a**)**.** Contrary to naïve predictions, the microsatellite sequences did not yield signals that converged to half of their original intensities. Instead, the signals decayed towards zero or faded (**Supplementary Fig. 2b**). In fact, after logarithmic transformation the signal intensity showed linear behavior, which implies an exponential decay. Since we could not attribute this effect to phasing, we introduced another noise factor – fading. We found that fading is correlated with the scanning buffer that is used, and switching to the new Illumina scanning buffer decreased the decay rate from ~3% to 1% on the GAI machines. Hence, fading is mainly an outcome of material loss, but we cannot exclude additional mechanisms that might also contribute to fading, such as nascent strand melting or existence of a very small fraction of nucleotides that are irreversibly blocked.

Random walk model for phasing and fading

We developed a random walk model to describe phasing and fading **(Supplementary Fig. 3a).** At the biochemical level, phasing and fading are directly correlated with the nascent strand length. Phasing causes some variation in nascent strand lengths in the DNA cluster, which accumulate throughout the run, and fading exponentially decreases the number of strands in the cluster as a function of cycle. Our random walk model describes the overall effect by three parameters, p_1, p_2 and p_3 with simple probabilistic interpretations. p_1 donates the probability of block removal that permits further polymerization. If the blocked is removed, two mutually exclusive events are considered: (a) incorporation of blocked nucleotide with probability p_2 or (b) incorporation of a contaminating, non-blocked nucleotide with probability of 1- p_2. Therefore, in a particular cycle the nascent strand stays at the same length without any synthesis with probability 1-p_1, grows by a single nucleotide with probability p_1x p_2, grows by two nucleotides with probability p_1x(1- p_2)x p_2, and so on. The third parameter, p_3, donates the probability of stand loss, which leads to signal decay. Note that p_1, p_2 and p_3 are time, cluster-size and sequence invariant, which is a simplification of the real situation. It should be possible to include these effects in a more complete model, but at the cost of increased parametric complexity. The propagator of the random walk model is given by:

$$R(t,n) = e^{-p_3 t} \int_{-\pi}^{\pi} \left[1 - p_1 + \frac{p_1 p_2 e^{i\omega}}{1-(1-p_2)e^{i\omega}} \right]^t e^{-i\omega n} \frac{d\omega}{2\pi} \qquad (1)$$

Thus, R, is a matrix that gives the probability of a nascent strand to be n-nucleotides long after t cycles (**Supplementary Fig. 3b)**. R can be decomposed to the effect of phasing and fading:

$$D \times P = R \tag{2}$$

D, the "fading" matrix, is T-by-T diagonal matrix capturing the exponential decay of the signal, where T is the total number of synthesis cycles. P is T-by-N matrix that donates the phasing. P(t,n) corresponds to the probability of finding a nascent strand with a length n after t cycles. N is the length of the longest nascent strand, and it cannot be longer than the template length. In the ideal case, D and P would be identity matrices.

The relationship between a DNA sequence a j-th DNA and the received intensities from that cluster is given by:

$$\eta_j \cdot D \times P \times S_j \times G^* = I_j \tag{3}$$

In this model, η_j is a scalar that represents the size of the j-th DNA cluster. D and P are the fading and phasing matrices described above. S is N-by-4 matrix that contains a binary representation of the DNA sequence of the j-th cluster. Each row of S has one element that equals 1 and three elements that equal 0. For instance, the short sequence, 'ACCGT,' would be represented as:

$$S_j = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

G is 4-by-4 matrix that represents the crosstalk between the channels, and star indicates transposition. In order to maintain the format of this crosstalk matrix as reported by Illumina, each row of G corresponds to the response of the four channels for a fluorophore. I_j donates the intensity signal from the j-th cluster and is T-by-4 matrix.

In the model described above, one can consider Illumina sequencing as a process that forward transforms the sequence data - nucleotide identity as a function of position - to the fluorophore channel response as a function of cycle number. Therefore, in order to estimate S, the input sequence, an inverse transform must be applied to the intensity matrix. The accuracy of this inverse transform relies heavily on a correct estimation of phasing and the crosstalk matrices, and is described by:

$$\eta_j \cdot (\hat{P} \times \hat{D})^+ \times D \times P \times S_j \times G^T \times \hat{G}^{T^{-1}} = (\hat{P} \times \hat{D})^+ \times I_j \times \hat{G}^{T^{-1}} \quad (4)$$

In this equation, + represents the pseudo inverse. If we assign Y as the right hand term of (4), upon correct estimation of the phasing and the crosstalk matrices, (4) can be written as:

$$\eta_j \cdot \Sigma \times S_j = Y \quad (5)$$

Where $\sum$ is a band diagonal matrix. The algorithm to find the sequence is dependent upon the nature of $\sum$. In a case in which the bandwidth of $\sum$ is greater than one, a dynamic programming can be applied to find the most likely sequence call[1]. In a case in which $\sum$ is very close to a diagonal matrix, one can find the best sequence call by detecting the strongest signal of Y in each row.

Testing the parametric model with a constant cross talk matrix

We tested the Illumina signal model in (3) using the averaged intensity files generated from our 12 DNA fragments. We obtained an estimation for the crosstalk matrix, G, using the reported values from Illumina. In order to find our experimental parameters we employed a least mean squares fitting procedure, and found that p1 was ~0.99, p2 was ~0.995, and p3, the decay was 3%-5% in each round. We calculated the $\sum$ matrix using these values found that the bandwidth of the matrix was very close to one. Therefore, after de-convolution of phasing, calling a base relies only on the four intensities values that were detected in that specific cycle. Thus, one can normalize the intensities in each cycle, and one does not have to estimate the decay rate. A code that simulates the R matrix (phasing and decay) and $\sum$ upon different p1, p2 and p3, and presents them graphically is available upon request from the authors.

In general, we found good agreement between our model and the reference data **(Supplementary Fig. 4a)**. However, when we closely examined the predictions for the T channels, we found an extensive anomaly. For almost all the sequences in our input set the T channel was stronger than predicted. These deviations were especially strong in the GT microsatellite and the GT theta function sequences **(Supplementary Fig. 4b,c)**. We probed the phenomenon more carefully by analyzing sequencing data generated from a phi-X genomic library, which often serves as a control for Illumina sequencing runs. We called phi-X bases according to the P matrix that was derived from fitting our model and the cross talk matrix that is given be Illumina. We found that the overall composition of the called bases changed dramatically as cycle number increased. Specifically, the percentage of T calls increased during late cycles, whereas the percentage of G calls dropped. We observe the same trend with C and A calls, albeit to a lesser extent, with the percentage of C calls increasing and the percentage of A calls decreasing at late cycles. The same trends were

observed if we used the standard Illumina base caller rather than our own algorithm **(Supplementary Fig. 5a)**.

Incorporating cycle-dependent cross talk

The observed deviation from the model was reminiscent of the relationship between the crosstalk in the emissions from G and T and from A and C. We analyzed whether the crosstalk between these fluorescence channels changes over time. The intensity files from the phi-X lane were de-convoluted from the random walk, and the ratios between all the possible 6 pairs of channels were measured. The ratios were normalized using a polar coordinate transformation, binned, and enumerated **(Supplementary Fig. 5b)**. For instance, signals with a strong C component and a weak T component were presented in the 90 degree bin of the CT histogram, and the opposite instances were presented in the 0 degree bin. The 45-degree line represents the criterion that separates signals from each of the two nucleotides. In the case of accurate crosstalk correction, the two lobes of the histogram should be orthogonal with respect to each other, and close to the axes. Indeed, we found that in early cycles all lobes were orthogonal. In later cycles, the lobe width increased, as expected, due to length spread but the lobes remained orthogonal in most signal pairs. However, the G lobe of the GT histogram underwent a strong shift as cycle number increased. A more subtle shift was noted in the A lobe of the AC histogram. These shifts reflect actual changes in the crosstalk matrix over the course of a run. We therefore revised our model to include a dependency of the crosstalk matrix on cycle number:

$$(\eta_j \cdot D \times P \times S_j) \times G^*(t) = I_j \qquad (6)$$

And the inverse model is given by:

$$\left(\eta_j \cdot (\hat{P} \times \hat{D})^+ \times D \times P \times S_j \right) \times G^*(t) \times \hat{G}(t)^{*^{-1}} = (\hat{P} \times \hat{D})^+ \times I_j \times \hat{G}(t)^{*^{-1}} \qquad (7)$$

Due to extensive variations from run to run and between different machines, the exact description of G as function of cycle is yet to be determined. Thus, we used an SVM in order to address cross-talk changes.

Additional results

Benchmarking Alta-Cyclic on GAI machines

We performed simulated a SNP calling experiment on phi-X on the GAI machine using a 50 cycle run. Surprisingly, this yielded similar values for phasing as the GAII run (p1=0.9925, p2=0.9975) (**Supplementary Fig. 6)**. This may permit us to reduce the area of the grid search and speed training in the future. Overall, Alta-Cyclic increased the number of correct 50nt reads by a factor of 1.52, reporting ~1,450,000 correct reads, whereas the Illumina base caller reported ~950,000 correct reads. The rate of miscalling SNPs with the Illumina

pipeline was around 5.5% across all the cycles; whereas, the Alta-Cyclic miscall rate was 4.1%. The difference in the error rates was elevated for the last 15 cycles. The Illumina base caller missed 11% of the SNPs, but Alta-Cyclic only missed 7% (**Supplementary Fig. 7)**. Note that these results obtained with the old scanning buffer.

Supplementary methods

Alta-Cyclic software

The Alta-Cyclic skeleton was written mainly in Perl. It uses BLAT[2] for building the training set, Perl Data Language (PDL)[3] for matrix manipulation, and libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) for machine learning. The SVD decomposition is based on the Golub-Reinsch algorithm that was implemented using the GNU Scientific Library (GSL). The Alta-Cyclic skeleton code and related code are open source. Alta-Cyclic can be downloaded from http://hannonlab.cshl.edu/.

Computer cluster architecture

We tested Alta-Cyclic on CSHL's High Performance Computing Cluster (HPCC). The system is an IBM E1350 blade-based cluster running Linux. It comprised 500 compute nodes, but Alta-Cyclic used only ~81 nodes for training, and ~100 nodes for base-calling. Each compute node contained 2 dual-core Opteron 64 processors running at 2.0GHz, a local 73GB SFF SCSI disk, and memory from 2 to 8GB. We found that the last SVM training step should be executed on a node with at least 4GB of memory. Sun Grid Engine (SGE) was used to submit jobs to the compute nodes. The training process took typically 10 hours. Note that Alta-Cyclic requires SGE in order to submit jobs.

Illumina sequencing

The 78-cycle GAII run took approximately one week, and additional reagents were added as needed. We modified the machine "recipe" to run for 80 cycles (see **Supplementary Methods** for recipe). For data storage, we installed a Western-Digital 2Tb external hard-drive using FireWire. However, after 78 cycles the machine stopped the run due to insufficient storage space.

Benchmarking and sample preparation

The phi-X, HepG2, and *Tetrahymena* libraries were sequenced on the same flowcell. The HepG2 sample preparation started by isolation of HepG2 nuclei. RNA was extracted with trizol and subsequently small RNA was isolated using the Ambion mirVana kit. 5µg of small RNA was TAP treated according to manufacturers protocol, but reducing the enzyme to 1µl. C-tailing was conducted using the ambion polyA tailing kit, substituting the ATP with CTP. Solexa 5' linker was ligated using Ambion T4 RNA ligase according to manufacturers description. Reverse transcription was conducted with Superscript III according to manufacturers description using the reverse complement of the IDT linker1 sequence extended by 10 guanine nucleotides. cDNA was amplified, the correct size range (100-300 bp) was gel extracted and submitted for sequencing. We trimmed the sequences by finding a block of 'CCCCCC' with up to 2 substitutions. We used Nexalign (Lassmann, T. et al. ,Unpublished) to find perfect matches to the human genome, and we eliminated matches from

sequences with blocks of 78xT, which is a recurrent sequencing artifact on Illumina machines. We collapsed matches from the same sequence to multiple locations, and enumerated the results. This procedure was the same for Illumina base caller and Alta-Cyclic.

For the Terahymena sequences, we trimmed the first cycle that represented an additional T nucleotide at the 5' end, and we extracted the following 60 nucleotides. We aligned the 60nt long sequence to the Tetrahymena MAC genome to find perfect matches.

The phi-X genome (NC_001422.1) was randomly mutated in ~50 locations, which corresponds to 1% SNPs. After training and base calling, Blat was used to find sequences with less than 4 substitutions to the mutated genome. We filtered reads that are aligned to SNP locations, and extracted the nucleotide identity of these locations, while we also saved the cycle number in which the SNP was called. We compared the SNP identity to the intact genome and calculated the number of errors. For the distribution of number of errors, we used Blat for alignment with minScore=25, minMatch=1, tileSize=6, minIdentity=25. We filtered the best matches for each read, but we did not include reads that align with gaps.

Recipe for 80 cycles

Note: the machine will stop after 78 with 2Tb storage device

```
<?xml version="1.0" ?>
<RecipeFile>

        <!-- 80Cycle_GA2_v1.xml -->
        <!-- This recipe is for use with the GAII system -->

        <!-- Single Read Recipe -->
        <!-- Exposure Time: Total = 1375 ms (A=500, C=350, G=350, T=175) -->
        <!-- No. Tiles Per Column: 50 -->
        <!-- No. Cycles: 80 -->
        <!-- ImageCyclePump config file should be set to "TRUE" -->
        <!-- Edited by Yaniv Erlich erlich@cshl.edu 4th Apr 2008          -->


  <TileSelection>

    <Incorporation>
        <Lane Index="1"><RowRange Min="1" Max="50"/></Lane>
        <Lane Index="2"><RowRange Min="1" Max="50"/></Lane>
        <Lane Index="3"><RowRange Min="1" Max="50"/></Lane>
        <Lane Index="4"><RowRange Min="1" Max="50"/></Lane>
        <Lane Index="5"><RowRange Min="1" Max="50"/></Lane>
        <Lane Index="6"><RowRange Min="1" Max="50"/></Lane>
        <Lane Index="7"><RowRange Min="1" Max="50"/></Lane>
        <Lane Index="8"><RowRange Min="1" Max="50"/></Lane>
    </Incorporation>

  </TileSelection>

  <ChemistryDefinitions>

    <Chemistry Name="Warning">
        <UserWait Message="Please Ensure that You have Previously Run the FirstBase
Recipe.    Press OK to Continue, or CANCEL to Stop." />
```

```
    </Chemistry>

    <Chemistry Name="CompleteCycle">
        <PumpToFlowcell     Solution="7"     AspirationRate="250"     DispenseRate="2500"
Volume="125" />
        <Temp Temperature="55" Duration="120000" />
        <PumpToFlowcell  Solution="6" AspirationRate="250" DispenseRate="2500" Volume="75"
/>
        <Wait Duration="240000" />
        <PumpToFlowcell  Solution="6" AspirationRate="250" DispenseRate="2500" Volume="25"
/>
        <Wait Duration="240000" />
        <PumpToFlowcell  Solution="6" AspirationRate="250" DispenseRate="2500" Volume="25"
/>
        <Wait Duration="240000" />
        <Temp Temperature="22" Duration="120000" />
        <PumpToFlowcell  Solution="5" AspirationRate="250" DispenseRate="2500" Volume="75"
/>
        <PumpToFlowcell  Solution="4" AspirationRate="250" DispenseRate="2500" Volume="75"
/>
        <PumpToFlowcell     Solution="5"     AspirationRate="250"     DispenseRate="2500"
Volume="125" />
        <Temp Temperature="55" Duration="120000" />
        <PumpToFlowcell  Solution="1" AspirationRate="250" DispenseRate="2500" Volume="75"
/>
        <Wait Duration="240000" />
        <PumpToFlowcell  Solution="1" AspirationRate="250" DispenseRate="2500" Volume="25"
/>
        <Wait Duration="240000" />
        <PumpToFlowcell  Solution="1" AspirationRate="250" DispenseRate="2500" Volume="25"
/>
        <Wait Duration="240000" />
        <Temp Temperature="22" Duration="120000" />
        <PumpToFlowcell  Solution="5" AspirationRate="250" DispenseRate="2500" Volume="75"
/>
        <PumpToFlowcell  Solution="4" AspirationRate="250" DispenseRate="2500" Volume="75"
/>
        <PumpToFlowcell  Solution="3" AspirationRate="250" DispenseRate="2500" Volume="60"
/>
        <TempOff />
    </Chemistry>

    <Chemistry Name="End">
        <PumpToFlowcell     Solution="2"     AspirationRate="250"     DispenseRate="2500"
Volume="500" />
    </Chemistry>

  </ChemistryDefinitions>

  <Protocol>

    <ChemistryRef Name="Warning" />

                <!--      Cycle 1       -->
    <Incorporation ExposureA="500" ExposureC="350" ExposureG="350" ExposureT="175" />

                <!--      Cycle 2       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 3       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 4       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 5       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />
```

```
                <!--      Cycle 6       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 7       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 8       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 9       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 10       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 11       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 12       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 13       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 14       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 15       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 16       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 17       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 18       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 19       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 20       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 21       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 22       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 23       -->
```

```
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 24       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 25       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 26       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 27       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 28       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 29       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 30       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 31       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 32       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 33       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 34       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 35       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 36      -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 37       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 38       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 39       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 40       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />
```

```
               <!--      Cycle 41       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 42       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 43       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 44       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 45       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 46       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 47       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 48       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 49       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 50       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 51       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 52       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 53       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 54       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 55       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 56       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 57       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

               <!--      Cycle 58       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />
```

```
                <!--      Cycle 59       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 60       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 61       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 62       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 63       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 64       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 65       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 66       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 67       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 68       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 69       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 70       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 71      -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 72       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 73      -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 74      -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 75       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 76       -->
```

```
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 77      -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 78      -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 79      -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />

                <!--      Cycle 80       -->
    <Incorporation    ChemistryName="CompleteCycle"    ExposureA="500"    ExposureC="350"
ExposureG="350" ExposureT="175" />



    <ChemistryRef Name="End" />

  </Protocol>

</RecipeFile>
```

--- end of recipe ---

Generation of sequences for impulse response analysis

We used miniGene technology (IDT) to synthesis the sequences for testing the impulse response. We synthesized a controlled input set of 12 DNA fragments.. The Illumina p5, p7 and SBS3 priming site were synthesized as part of the fragments and two NlaIII sites flanked each DNA fragment. Each of the four fragment types were cloned into pZero (Invitrogen). We checked the accuracy of clones using conventional ABI sequencing. For Illumina sequencing, the plasmids were cut with 2ul of NlaIII (NEB), and the 3' prime overhangs were removed with 1ul of T4 DNA polymerase (NEB). 100-200bp fragments were selected using 1% agarose gel (Roche) and extracted. The concentration of the DNA from the plasmids was measured and equal amounts were mixed and diluted to 2ng/ul in TE.

Analysis of the intensity files for impulse response

Each sequence read of the Illumina sequence files was matched to the twelve possible sequences and the annotation was determined according to the best match. We filtered out sequences with ambiguous base called, represented as a "." in Illumina output. Then, the intensity files were scanned and only DNA clusters in which the intensity values of the first round were stronger than the average passed the filter. We averaged all the intensities according to their annotation. Then, the intensity values for the first cycle were removed, and the strongest point was normalized to 1 and the median of the minimum intensity values from all the cycles was normalized to 0. We used a grid computer to scan different values of p1,p2 and p3 and to extract the sum of the squared error from

different types of sequence. We excluded the delta sequences on homopolymer background since we noticed that they were subject to small deletions and insertions, presumably during bridge amplification. These variations could affect the calculation of the random walk parameters.

Recipe for taking images with different exposure time:

```
<?xml version="1.0" ?>
<RecipeFile>

       <!-- 36Cycle_v1.xml -->

       <!-- Service protocol made by Yaniv Erlich 12/19/07 -->
       <!-- Exposure time: variable -->
       <!-- No. tiles per column: 15 -->
       <!-- No. Cycles: No Chemistrey -->


  <TileSelection>

    <Incorporation>
       <Lane Index="4"><RowRange Min="1" Max="15"/></Lane>
       <Lane Index="5"><RowRange Min="1" Max="15"/></Lane>
       <Lane Index="7"><RowRange Min="1" Max="15"/></Lane>
    </Incorporation>

  </TileSelection>

  <ChemistryDefinitions>

  </ChemistryDefinitions>

  <Protocol>
              <!--      Going from smaller to bigger to avoid bleaching       -->
    <Incorporation ExposureA="1" ExposureC="1" ExposureG="1" ExposureT="1" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="25" ExposureC="25" ExposureG="25" ExposureT="25" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="50" ExposureC="50" ExposureG="50" ExposureT="50" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="75" ExposureC="75" ExposureG="75" ExposureT="75" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="100" ExposureC="100" ExposureG="100" ExposureT="100" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="125" ExposureC="125" ExposureG="125" ExposureT="125" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="150" ExposureC="150" ExposureG="150" ExposureT="150" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="200" ExposureC="200" ExposureG="200" ExposureT="200" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="250" ExposureC="250" ExposureG="250" ExposureT="250" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="300" ExposureC="300" ExposureG="300" ExposureT="300" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="350" ExposureC="350" ExposureG="350" ExposureT="350" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="400" ExposureC="400" ExposureG="400" ExposureT="400" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="600" ExposureC="600" ExposureG="600" ExposureT="600" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="800" ExposureC="800" ExposureG="800" ExposureT="800" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="1000" ExposureC="1000" ExposureG="1000" ExposureT="1000" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="1200" ExposureC="1200" ExposureG="1200" ExposureT="1200" />
       <Wait Duration="120000" />
    <Incorporation ExposureA="1400" ExposureC="1400" ExposureG="1400" ExposureT="1400" />
```

```
        <Wait Duration="120000" />
    <Incorporation ExposureA="1600" ExposureC="1600" ExposureG="1600" ExposureT="1600" />
        <Wait Duration="120000" />
    <Incorporation ExposureA="1800" ExposureC="1800" ExposureG="1800" ExposureT="1800" />
        <Wait Duration="120000" />
    <Incorporation ExposureA="2000" ExposureC="2000" ExposureG="2000" ExposureT="2000" />


  </Protocol>

</RecipeFile>
```

--- end of recipe ---

Supplementary references

1. Kailath, T. & Poor, H.V. Detection of stochastic processes. IEEE Transactions on Information Theory, 44, 2230-2231 (1998).
2. Kent, W. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).
3. Glazebrook, K. et al. Perl Data Language. The Perl Journal 5, 5 (1997).