bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Sources of PCR-induced distortions in high-throughput sequencing datasets

Justus M Kebschull[1,2], Anthony M Zador[2,*]
1 Watson School of Biological Sciences
2 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
∗ E-mail: zador@cshl.edu

Abstract

PCR allows the exponential and sequence specific amplification of DNA, even from minute starting quantities. Today, PCR is at the core of the most successful DNA sequencing technologies and is a fundamental step in preparing DNA samples for high throughput sequencing. Despite its importance, we have little comprehensive understanding of the biases and errors that PCR introduces into pools of DNA molecules. Understanding PCRs imperfections and their impact on the amplification of different sequences in a complex mixture is particularly important for a proper understanding of high-throughput sequencing data. We examined the effects of bias, stochasticity, template switches and polymerase errors introduced during PCR on sequence representation in next-generation sequencing libraries. Using Illumina sequencing results of a pool of diverse PCR amplicons with a defined structure, we searched for signatures of each process. We further developed quantitative models for each process and compared predictions of these models to our experimental data. We find that PCR stochasticity is the major force skewing sequence representation after amplification of a pool of unique DNA amplicons. PCR errors become very common in later cycles of PCR but have little impact on the overall sequence distribution as they are confined to small copy numbers. PCR template switches are rare and confined to low copy numbers. Our results will have particular relevance to single cell sequencing, in which sequences are represented by only one or a few molecules.

Author summary

High throughput sequencing technologies are used both qualitatively to determine the genomic sequence of an organism and quantitatively to measure the amount of specific DNA sequences present in complex mixtures. To prepare a sample for high throughout sequencing, the input DNA needs to be amplified by PCR. Amplification can introduce skews, biases and errors into the DNA pool leading to misrepresentation of the amounts of sequences in the sequencing results. Here we investigated four potential sources of such misrepresentation and find that, when molecule numbers are low early in PCR, the random amplification of some sequences and not others has a large impact on sequencing results.

Introduction

DNA sequencing technologies have rapidly improved during the last two decades. Due to decreased cost and increased speed, DNA sequencing is now a standard technique in molecular biology both for sequence determination and quantification.

Before a DNA sample can be sequenced, a sequencing library must be prepared from the sample. Although the steps in library preparation vary, the protocol invariably involves PCR amplification. Understanding and accurately quantifying sequencing results thus depends critically on understanding the effects of PCR.

Surprisingly, no study has comprehensively investigated sources of sequence misrepresentation in sequencing datasets, nor has there been a coherent theoretical and experimental investigation of more than one source of sequence misrepresentation after PCR. A recent paper experimentally investigates the effects

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

of extreme base composition on Illumina sequencing and pinpoints PCR during library preparation as a critical source of the observed sequence bias [1]. Similarly detailed studies of other unintended properties of PCR are necessary.

Here, using pools of carefully designed amplicons and Illumina sequencing, we theoretically and experimentally investigate several processes known to cause sequence misrepresentation after PCR amplification. First, we studied the effects of PCR bias focussing on variable PCR amplification efficiencies as a function of the GC content of individual sequences. Second, we investigated stochasticity with which each DNA molecule is amplified at each cycle of PCR. Third, we focussed on template switching during PCR as a process producing novel sequences during amplification. Lastly, we studied PCR errors and their impact on sequencing results. For each process, we formulated a mathematical framework, looked for signatures of the process in sequencing data, and compared our theoretical predictions with the experimental data.

We find that PCR stochasticity is the major force skewing sequence representation after amplification of a pool of unique DNA amplicons. PCR errors become very common in later cycles of PCR but have little impact on the overall distribution as they are confined to small copy numbers. Template switches are rare.

Materials and Methods

DNA oligos and PCR. We ordered three ultramers from Integrated DNA Technologies: BC1-BC1, BC2-BC2 and the adapter (Table 1). BC1-BC1 and BC2-BC2 contain the Illumina P5-SBS3T sequence followed by a 20nt barcode, the AttL sequence and another 20nt barcode. The adapter is $5'$ phosphorylated and contains a 15nt barcode followed by the reverse complement of the Illumina P7-SBS8 sequence. We pooled the two types of barcode pairs in equal proportions and ligated them to the $3'$ adapter using CircLigase ssDNA ligase (epicentre; previously sold as Thermophage ssDNA ligase) as previously described [2]. The ligation reaction was cleaned up with Agencourt RNAClean XP beads (Beckman Coulter) according to the manufacturers instructions. The ligated products were subjected to 25 cycles of PCR using $47\mu l$ Accuprime Pfx SuperMix (Invitrogen), $1\mu l$ of $10\mu M$ forward and reverse primers each (Table 1) and $1\mu l$ input. Cycling was performed in a BioRad MyCycler Thermal Cycler using standard Accuprime protocol with $58°C$ annealing temperature and 30 seconds extension time. The PCR product was gel extracted and sequenced on a single lane of a HiSeq 2000 machine at PE101.

Data processing. Illumina sequencing resulted in 60 million reads passing filter. We merged the paired end reads into their consensus sequence with the Pear tool [3] using standard settings and requiring a consensus sequence of 101 nt. We trimmed and preprocessed the remaining 15 million consensus reads using Matlab requiring a perfect match to the constant AttL region and detected 152798 unique sequences (*preprocessing.m*). Code for the analysis of GC content, position of candidate biased sequences and simulation of biased PCR can be found in *bias.m*.

Calculation of the PDF of copy numbers after PCR. It is computationally prohibitive to calculate the exact PDF of copy numbers after 25 cycles of PCR. Instead we approximated the PDF by starting with the exact PDF after 15 cycles. Let us use the vector $\mathbf{S}$ to denote the distribution of copies of a given barcode-pair after j cycles. Each element $\mathbf{S}(i)$ of $\mathbf{S}$ is the probability of having $i-1$ copies. Thus, $\mathbf{S}(1)$ is the probability of having 0 copies, $\mathbf{S}(2)$ of having 1 copy, etc. If eg $\mathbf{S}(3) = 1$, it means the probability distribution is a delta function at exactly two copies; if $\mathbf{S}(2) = 0.5$ and $\mathbf{S}(3) = 0.5$, it means there is a 50-50 chance of having two or three copies. (Note that the largest nonzero element of $\mathbf{S}$ must in general be $\leq 2^j$, so the length of $\mathbf{S}$ must be $\leq 2^{j+1}$.)

Our approach is to find an updating matrix $\mathbf{M}$ such that the distribution of copies $\mathbf{S}'$ on the next cycle is given by $\mathbf{S}' = \mathbf{MS}$.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

This formulation exploits the Markovian nature of the process, namely it does not matter how we ended up with k copies on a given cycle; all that matters is we have k copies. To determine the elements of **M**, we first consider an easier problem. Suppose there are k copies on a given cycle; what is the expected distribution of copies on the next cycle? The number of new copies is given by a binomial distribution with parameters k and P; the distribution of total number of copies is k + the number of new copies. Thus, for cases of the distribution **S** where $\mathbf{S}(i)$ is a delta function (S(i)=1), we have $B(i,P) = \mathbf{MS}(i)$ where B(i,P) is a binomial shifted by i. This means that the ith column of **M** is B(i,P).

The calculation of the PDF after 15 cycles can be found in *exactpdf.m*.

To approximate the PDF after more than 15 cycles, we used the exact PDF after 15 cyles and applied a Gaussian with a mean of $n * 1.9$ and a standard deviation of $\sqrt{n * 0.9 * 0.1}$ to every copy number value n. We added the resulting Gaussians weighted by the probability of the copy number to which they were applied for every cycle. Code for this approximate PDF can be found in *approxpdf.m*.

Position of template switched reads. We identified half of all template switched reads by comparing dinucleotide anchor sequences between the two barcodes in each barcode pair. We only considered reads where every dinucleotide anchor was either RR or YY, and then checked for barcodes of type 1 joined to barcodes of type 2 and vice versa. Code can be found in *tempswitch.m*.

Position of single nucleotide errors. We approximated the overall rate of single nucleotide errors in two ways: (1) by determining the minimum hamming distance of each sequence in the shoulder and tail to the plateau sequence and (2) by identifying single nucleotide errors in the dinucleotide anchor sequences, where we quantified the occurrence of RY or YR anchors, that by definition are not part of the input barcode pool. Code can be found in *PCRerrors.m*. Code for Supplemental figure 1 can be found in *mismatchvhamming.m*.

Simulation of PCR stochasticity and errors. The simulation was performed as described in the main text. Details and code can be found in *GWanderrors.m*.

Results

Experimental system. To study the effects of PCR on sequence representation in Illumina libraries, we designed an experiment in which we reduced sequencing library preparation to a single PCR on known, but diverse input sequences. We synthesized DNA oligonucleotides containing two regions of random sequence, i.e. two *barcodes*, joined by a constant region. These barcode pairs are flanked by two constant sequences that are required for Illumina sequencing and also act as PCR primer binding sites during library preparation (Fig 1 A).

We subjected about 3000 of these oligonucleotides to 25 cycles of PCR and sequenced the resulting library. The combinatorial space of all possible barcode sequences is large enough (40 random barcode nucleotides; $4^{40} \approx 10^{24}$ possible combinations) that we expect every input molecule to have a unique sequence. This implies that every barcode pair is present at equal abundance in the input DNA pool and should therefore be read out at approximately equal read counts in the sequencing results.

We plotted the experimental sequencing read counts for every barcode pair (Fig 1 B) sorted from the most abundant to least abundant barcode, that is by sequence rank where rank 1 is the most abundant sequence. The most abundant barcode pairs are present at similar read counts, forming a plateau in the plot. This plateau is followed by a shoulder of barcode pairs with intermediate abundance and a long tail of low abundance barcode pairs.

Given an equal abundance of all sequences before amplification, we would have naively expected a flat sequence trace only. The presence of a shoulder and tail in the experimental data suggests that PCR

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

amplification or Illumina sequencing has introduced sequence misrepresentation into the dataset. As sequencing is essentially linear and thus less likely to introduce large shifts in the sequence distribution, we suspected inaccurate PCR amplification was the source of the observed misrepresentation. Conceptually, there are a variety of potential artifacts that can be introduced by PCR, each of which will have a different impact on sequence representation after amplification (Fig 2). We will address each possibility in turn.

Perfect PCR. For comparison, we first considered the sequence distribution we should expect given perfect PCR amplification. Perfect PCR faithfully amplifies every molecule in the input DNA pool, simply doubling molecules at every cycle. Relative abundances of different sequences will thus be preserved during amplification.

Mathematically, perfect PCR can be summarized as

$$n(j) = N_0 2^j \tag{1}$$

where $n(j)$ is the number of molecules of a particular amplicon after j cycles and N_0 is the initial copy number of this amplicon.

If every amplicon is unique before amplification, then $N_0 = 1$ and every sequence will be present at $n(j) = 2^j$ copies after j cycles of PCR. Plotting copies against sequence rank results in a plateau of height 2^j reads. If sequencing is deep enough to overcome Poisson sampling effects, we expect a similar plateau when plotting read counts against sequence rank for the Illumina sequencing results but with a plateau height that reflects sampling during sequencing (Fig 2 C).

The experimental data deviate substantially from this expectation. A trivial explanation for this disparity is that not all sequences were present at equal proportions in the input DNA pool. We experimentally controlled for this possibility by ligating a large excess of a third barcode to the $3'$ end of the DNA oligonucleotides before amplification. Every individual oligonucleotide is therefore uniquely labeled with a random sequence, making it possible to count how many copies of each sequence were present before amplification [4]. We found that all sequences from the plateau were present at a single copy before PCR (compare *preprocessing.m*). We therefore conclude that the naive model of perfect PCR does not describe the experimentally observed data.

PCR bias. PCR amplification efficiencies are not identical for every sequence. Sequence composition and secondary structure can introduce amplification biases. A high fraction of G or C can reduce amplification efficiency [1, 5, 6], causing uneven amplification of different sequences in PCR. PCR bias could therefore give rise to underrepresented sequences after amplification. These would appear on a sequence trace as a shoulder or tail (Fig 2 D).

Assuming different amplification efficiencies for different sequences, we can express the expected copy number of sequence x after j cycles of PCR as

$$E(n_x(j)) = N_{0,x} c_x^j \tag{2}$$

where $c_x = [1, 2]$ is the PCR efficiency of sequence x. PCR bias, that is a PCR efficiency c_x smaller than the average efficiency of all sequences $\langle c_x \rangle$, has been reported in sequences with a GC content higher than 65 %, although this value depends on cycling conditions and the specific polymerase used [1].

For large n_x we can approximate PCR as continuous in $n_x(j)$. Under this assumption, we can define

$$n_x(j) = N_{0,x} c_x^j \tag{3}$$

As GC bias causes uneven amplification efficiencies, we speculated that molecules that were underrepresented after amplification (i.e. the sequences in the shoulder) would be rich in GC (Fig 3 A). If GC bias were causing the observed differences in read counts, regions of high read counts should have lower GC contents than regions of low read counts. However, the experimental distribution of GC contents is not

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

significantly different from a simple binomial sampling model in both plateau (high read counts), shoulder (intermediate read counts) and even tail sequences (low read counts) (Fig 3 B).

Sequences with a GC content as low as 65% may be subject to greatly reduced amplification efficiency [1]. *A priori* we expect about 2% of all input sequences to show such a high GC content. To investigate whether GC bias is acting on these extreme sequences, we focus on the position of these sequences in the experimental sequence trace. Assuming the worst case scenario, in which sequences with GC content less than 65% are all perfectly amplified and all sequences with a GC content greater than 65% are all equally poorly amplified, we would expect to find all poorly amplified sequences in the shoulder and tail. However, we find that the location of sequences with a GC content greater than 65% in the 10000 most abundant barcode pairs is not different to a random shuffling of these positions in the experimental data (Fig 3 C). This is in stark contrast to a simulation of the described worst case scenario, where all GC biased sequences are found in the tail region (Fig 3 C and D).

In conclusion, we find no indication of GC bias in our experimental results. This shows that GC bias is not an important force in skewing sequence representation, although we note that in our experiments we implicitly minimized the contribution of GC bias by designing barcodes with an equiprobable distribution of bases. The observed shoulder and tail are therefore unlikely to be formed by GC bias.

Stochastic amplification of low copy number amplicons. A second source of uneven amplification in PCR is stochasticity. If PCR were perfect, every single molecule would be replicated every cycle. However, PCR is imperfect, so each molecule undergoes replication with a probability of less than 1. For example, if $P_{amplification} = 0.9$, per cycle then out of every ten molecules amplified, PCR fails to replicate one. This is not particularly concerning when PCR is used on DNA mixtures where every sequence is present in high copy numbers. In this case, the expected $1 + 0.9 * 1 + 0.1 * 0 = 1.9$ fold increase of molecule number of per cycle is sufficient to describe the behaviour of PCR.

However, when sequences are present at very low copy numbers, stochastic amplification may have a significant impact on sequence representation. Consider an example. First, consider a *lucky* amplicon that undergoes replication on the first cycle, so that on cycle 2 there are 2 copies. Further suppose that both copies are *lucky* and again undergo replication, so on cycle 3 there are 4 copies. Compare this to an *unlucky* amplicon, which fails to get copied both cycles (for $P_{amplification} = 0.9$, this happens $(1-p)^2 = 0.1^2 = 0.01$ or 1% of the time). If their luck evens out and both amplicons get amplified equally during subsequent cycles, the lucky barcode will appear at a copy number of about 4 times more than the unlucky one. This suggests that the distribution of copy numbers for $P_{amplification} = 0.9$ will range over more than a factor of 4. Stochasticity in PCR could therefore explain the shoulder observed in the sequencing trace (Fig 4 A).

We can express this consequence of PCR stochasticity using the recursive expression

$$n(j+1) = n(j) + B(n(j), P_{amplification}) \tag{4}$$

where $B(n(j), P_{amplification})$ is a binomially distributed random variable with $n(j)$ trials and $P_{amplification}$ is the probability of success [7]. This expression is equivalent to modeling PCR as a Galton-Watson process, a stochastic branching process. In this formulation, every node represents one copy of a certain sequence and can give rise to one or two new branches, where each branch corresponds to success or failure of amplification on a given cycle.

Assuming this branching model and a realistic $P_{amplification} = 0.9$, we generated the exact probability distribution function (PDF) of the copy number n of a single sequence, starting with a single molecule at cycle 0. After 15 cycles of PCR, the PDF has a clear global maximum (Fig 4 B): As we would expect, most molecules are amplified most of the time. Interestingly, two further local maxima are discernible at copy numbers of 0.5 and 0.25 of the global maximum, corresponding to sequences that missed out on amplification during either one or two of the first two cycles of PCR.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

During the first few cycles of PCR molecule numbers are low and stochasticity has a large effect. We therefore expected to find the origin of the observed local maxima in the early cycles of PCR. After one cycle of PCR, the PDF is trivial. After two cycles, the PDF still shows only one maximum corresponding to molecules amplified in both cycles. After three cycles, the PDF shows two peaks at $n = 4$ and $n = 8$ (Fig 4 C). To reach copy number 8, the molecules have successfully been amplified at every cycle as $2^3 = 8$. To reach $n = 4$, molecules must have failed to replicate during one cycle. The biggest contribution to the probability of $n = 4$ comes from paths where molecules missed out on the first amplification cycle. This fact is immediately obvious when one considers all of the trees giving rise to four branches after three cycles and keeping in mind that a failure to amplify is less likely than success (Fig 4 D). After 4 cycles, the same structure as observed after 15 cycles becomes apparent (Fig 4 E). The PDF shows a global maximum at copy number 16 and two local maxima at 0.5 (copy number 8) and 0.25 (copy number 4) of that copy number. When we explicitly calculated the probabilities for these copy numbers, the dominating terms are sequences that failed to amplify on either the first or first two cycles of PCR. A third local maximum is apparent at $n = 12$, which is smoothed out in later cycles. This reasoning confirms that the local maxima in the PDF after 15 cycles correspond primarily to molecules that did not amplify during the first or first two cycles of the PCR reaction.

To test the hypothesis that stochasticity early in the PCR reaction generates the observed shoulder in the sequence trace, we approximated the PDF of copy numbers after 25 cycles of PCR with a constant PCR efficiency of $P_{amplification} = 0.9$. We sampled from this PDF to create a profile of read counts vs sequence rank. With a simulated PCR input of 2900 different sequences, we were able to reproduce a shoulder very similar to the one previously observed in experimental data (average correlation coefficient of $R^2 = 0.9689$; Fig 4 C). However, the smooth transition of shoulder to tail present in experimental data is missing.

In conclusion, PCR stochasticity has a large impact on sequence representation after PCR amplification and could give rise to most of the experimentally observed shoulder but not the tail.

Template switching. We next investigated processes producing *new* sequences during library preparation. If a new species is generated during PCR amplification, it will often be amplified in subsequent PCR cycles like one of the original input sequences. It will, however, lag behind these original sequences by at least one cycle and will thus be observed less frequently after amplification than most input sequences. Generation of new sequences during PCR could therefore account for the shoulder and tail of the sequence trace.

PCR template switching produces hybrid sequences of two already present sequences. DNA polymerase can jump from one template to another in a region of complementarity without aborting the nascent DNA strand during PCR [8]. This nascent strand therefore has a new hybrid sequence, where one piece is complementary to the old template and the other piece is complementary to the new template (Fig 5 A).

To gain a quantitative understanding of template switching, we set up a mathematical model of the process with the background of otherwise perfect PCR. We assume a bimolecular reaction. At any given time, every molecule is amplified by a polymerase. When two molecules collide, template switching occurs with probability s_0. When $s_0 \ll \frac{1}{N_j}$, where $N_j = N_0 * 2^j$ is the number of template molecules at cycle j, the per molecule probability of template switching on cycle j is

$$s_j = s_0 * N_j \tag{5}$$

From this the total number of template switches in cycle j is

$$S_j = s_j * N_j = s_0 * N_j^2 \tag{6}$$

As template switched molecules get amplified in every cycle after their generation, the total number of

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

template switched molecules after m cycles is

$$Q_m = \sum_{j=1}^{m} S_j 2^{m-j} = s_0 * N_0^2 * 2^m * \sum_{j=1}^{m} 2^j \approx s_0 * N_0^2 * 2^{2m+1} \tag{7}$$

The model predicts that the number of template switches per cycle grows with the square of the total number of molecules in solution N_j. As N_j increases exponentially with j, the probability of template switching increases exponentially with $2 * j$. We therefore expect template switches to become increasingly common in late cycles of PCR. As such, they will not accumulate to levels comparable to the original input sequences and should be detectable mostly in the tail of the sequence distribution.

To test this prediction experimentally, we searched for signatures of template switching in the sequencing results. We designed two different classes of barcodes (Fig 5 B). Barcodes of type 1 (BC1) are different from barcodes of type 2 (BC2) at six positions where the sequence is restrained to either a purine (R=A,G) or a pyrimidine (Y=C,T) base. Based on these anchors, BC1 and BC2 can be faithfully distinguished from each other. We started the PCR reaction with a pool of barcode pairs that were either BC1-BC1 pairs or BC2-BC2 pairs. In the absence of template switching, we would expect to observe only the initial barcode pair types in the sequencing dataset. However, when we detect a BC1-BC2 or BC2-BC1 pair, this barcode pair must have arisen from a template switch across the constant region between the two barcodes.

We find that template switched reads are present only at low read counts in the tail of the experimental sequence distribution (Fig 5 C). This is in agreement with the our prediction that template switched sequences are created late in PCR.

As essentially all observed template switches occur at read counts less than 3, we were unable to estimate s_0 from the data. However, our data suggest it is small. We estimate that each sequence must be present at least 1000 times to be detected by sequencing and thus must arise at cycle 15 or earlier (plateau is about 10^4 reads high, 25 cycles of PCR, so every unique input sequence is present about 10^7 times). For $j = 15$ and $N_0 = 3000$ equation 6 reads $S_j = s_0 * 9.7 * 10^{15}$, suggesting that s_0 is on the order of 10^{-15} so that we will detect some template switches at one or two copies. Assuming $s_0 = 10^{-15}$, a simulation of perfect PCR with template switching cannot account for the experimentally observed shoulder (Fig 5 D).

These results indicate that PCR template switching is a rare event in dilute solutions, and only becomes common late in PCR. By then, the newly generated sequences have lost out on many amplification cycles and are present at much lower copy numbers than original sequences. They are therefore detected at low copy numbers in sequencing. Template switched reads do not account for the observed shoulder, and only a small fraction of the sequences in the tail.

PCR errors. A second source of new sequences during PCR are amplification errors. During synthesis of a new DNA strand DNA polymerase makes errors including single nucleotide substitutions and, at a lower rate, small insertions or deletions [9]. Polymerase error rates strongly depend on experimental conditions, and estimates of error rates vary with the method used to determine polymerase errors. Wild type Taq polymerase is the best studied polymerase used in PCR and is generally used as a relative standard for polymerase fidelity. Estimates of Taq fidelity vary, but are on the order $P(\text{Error per nucleotide}) = 10^{-4}$ [10, 11]. AccuPrime Pfx polymerase [12], as used in our experiments, is estimated by the manufacturer to have a fidelity 26x higher than Taq polymerase. We therefore expect AccuPrime Pfx to introduce PCR errors at a probability of roughly

$$P(\text{Error per nucleotide}) = 4 * 10^{-6} \tag{8}$$

The probability of one or more errors in the 2x20 barcode nucleotides is therefore

$$P(\text{At least one error per barcode pair}) = 1 - \left(1 - 4 * 10^{-6}\right)^{40} \approx 1.6 * 10^{-4} \tag{9}$$

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Every roughly 6250 molecules, a new molecule with at least one error will be produced. These new sequences are subsequently amplified during the remaining cycles of PCR just like any other DNA molecule. However, as they, by definition, lag by at least one cycle of PCR, they are less abundant than original sequences. Moreover, the probability of producing new sequences due to PCR errors is linearly dependent on the number of amplified molecules and thus increases exponentially during PCR. Taken together, these considerations suggest that PCR errors are responsible for part of the shoulder but get increasingly more abundant with low copy number. We therefore predict that the lower part of the shoulder and parts of the tail are formed by PCR errors (Fig 6 A).

To test this prediction experimentally we used a Hamming distance metric to quantify sequences that arose from PCR errors. The Hamming distance between two sequences is defined as the number of substitutions necessary to go from one to the other. For example, a sequence with a single PCR error will have a Hamming distance of one to its original parent sequence. Using this metric, we cannot directly differentiate between PCR errors and Illumina sequencing errors. However, if a sequence appears more than three times in our dataset, it is unlikely that it arose due to an Illumina sequencing error and therefore is either real or result of a template switch or a polymerase error: At a lower bound quality score of $Q_{phred} = 30$, that is a base calling error rate of 10^{-3} per nucleotide, the probability of introducing the *same* sequencing error three times into different copies of the same 40nt barcode pair is $40 * 0.001^3 = 4 * 10^{-8}$. At a coverage on the order of 10^4 for *real* sequences — that is, reads from the plateau of the sequence trace — this implies that the probability of introducing the same sequencing error three times into a real barcode pair is $4 * 10^{-8} * 10^4 = 4 * 10^{-4}$. With only roughly 3000 real barcode pairs, sequencing errors in sequences present three or more times are negligible. At fewer than three counts per sequence, we cannot exclude the possibility that some of the observed mismatches arise from sequencing errors, especially in the singlet region.

To further reduce the contribution of sequencing errors to our dataset, we sequenced using paired end reads that each span the entire length of the barcode pair, so that every base was sequenced twice. We then determined the consensus sequence of paired end reads using the PEAR tool [3], analysing only those reads for which a consensus over the whole molecule could be found. Assuming independence of paired end reads, this procedure eliminates the majority of sequencing errors. Taking both read quality and paired end matching arguments into consideration, we consider the Hamming distance metric to be a good proxy for PCR errors.

We defined the plateau sequences from rank 1 to 2900 as original parent sequences, and find that we can account for about two thirds of all other sequences by a single nucleotide substitution in these original sequences. In contrast, the minimum Hamming distances between all the parent sequences are significantly different from each other, such that they could not be related to each other by PCR errors (Fig 6 B).

To quantify PCR errors in a more unbiased fashion without defining a set of parent sequences, we quantified PCR errors by scoring deviations from the expected sequence features of BC1 and BC2 sequences. Each of the defined anchors in BC1 or BC2 is a pair of two purines or two pyrimidines. Any mixed anchor sequence, e.g. AT or CG, therefore must be the result of a substitution. Using these mismatches as measure of PCR errors, we find that PCR errors are depleted in the plateau region of the experimental sequence trace, relative to randomly shuffled control positions. Outside the plateau region, the PCR error rate varies but is close to or higher than the randomly shuffled control (Fig 6 C). Interestingly, we found that the frequency of PCR errors as calculated from these mismatches is lower than expected from the Hamming distance metric even after correction for errors to which the mismatch metric is blind (e.g. purine to purine or pyrimidine to pyrimidine switches; Supplementary Fig 1 A). We found that error frequency is not uniform across all barcode positions as judged using the Hamming distance metric. This explains why we underestimated overall PCR error rate using the mismatches (Supplementary Fig 1 B).

Based on the relatively steady rate of PCR errors outside the plateau, we hypothesized that most sequences found at the bottom of the plateau and in the tail of the sequence trace arose from PCR

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

errors. To test this hypothesis, we simulated erroneous PCR with an overall amplification efficiency of $P_{amplification} = 0.9$ in a deterministic model. Assuming that every individual error is rare, we simulated 25 cycles of PCR on an input of 2900 sequences using an estimated error rate of $P(\text{Error per nucleotide}) = 1.5 * 10^{-5}$. The resulting sequence rank plot shows a plateau followed by a steep drop off which then softens into a long tail (Fig 6 D). This simulation does not account for the experimentally observed smooth top of the shoulder, but does agree closely with the experimentally observed trace at the bottom of the shoulder.

Taken together these data confirm our theoretical predictions. PCR errors are relatively common but in absolute numbers happen predominantly late in PCR. They are thus confined to the tail of the sequence distribution where they make up a large fraction of sequences.

Stochasticity and PCR errors explain much of the observed dataset. From the above analysis, we expected stochasticity of amplification and PCR errors to explain most of the observed sequence distribution. To test this hypothesis we simulated PCR by a Galton Watson process, and added PCR errors at a rate of $P(\text{Error per nucleotide}) = 1.5 * 10^{-5}$ (Fig 7). Again we assumed that each individual error is rare. Simulation and experimental data show a very good fit, confirming that the observed distribution can be explained by PCR stochasticity and PCR errors alone.

Discussion

We have systematically investigated potential sources of sequence misrepresentation in next-generation sequencing libraries, focussing on bias, stochasticity, template switching and errors introduced by PCR. We find that PCR stochasticity in the first two or three cycles of PCR greatly affects sequence representation of low copy number sequences after amplification. Polymerase errors and to a lesser extent template switches generate new sequences which occur predominantly at low read counts, producing the observed tail in the read distribution.

Each of the processes examined here has been previously described. GC bias [1, 6, 13] and polymerase errors are often voiced as a concern about PCR. PCR stochasticity has been noted in theoretical models of PCR [7, 14, 15], as well as in practical applications of PCR in forensics [16, 17]. Template switches in PCR reactions have been studied experimentally [8].

Sequence misrepresentation and biases are also commonly observed in PCR amplified sequencing samples. However, it is largely unclear what effect each of the known PCR imperfections has on sequence misrepresentation. In the absence of such knowledge, a number of strategies have been developed to improve on the skew in sequence representation in sequencing data. Techniques like multiple displacement amplification [18], antisense RNA amplification [19] or multiple annealing and looping-based amplification cycles [20] aim to minimize the use of PCR. With these approaches, amplification is linear or quasilinear which means that errors, biases and failures to amplify accumulate less rapidly than in exponential PCR amplification. In contrast, single molecule barcoding techniques [4] still rely on PCR, but compensate for misrepresentations *post hoc*. Individual molecules are uniquely labeled before amplification, so that the number of input molecules can be precisely quantified after sequencing. These approaches are doubtlessly effective at reducing sequence misrepresentation, but do not provide insights into what processes exactly they are avoiding or compensating for.

Our findings provide a comprehensive background in which to understand PCR induced misrepresentations in sequencing data. Studying four processes in the same system allowed us to assess their relative importance and revealed that stochasticity and polymerase errors were the dominating effects at high and low read counts, respectively. GC bias is well known to introduce misrepresentations into sequencing data [1], but its effects are minimized in our experimental system by the randomness of barcode design. Additionally, we expect that the use of Accuprime polymerase and of a thermocycler with relatively slow

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ramp rates (2.5 deg/sec) further diminished the detrimental effects of GC bias. Template switches occur at low read counts only.

We designed the experimental system used in this study to disentangle different sources of PCR induced misrepresentation in sequencing datasets. Our results are especially relevant to the single cell sequencing community. Single cell sequencing approaches operate in low copy number regimes. Stochasticity in PCR amplification might therefore explain some of the uneven coverage observed in copy number analyses of single cells [21] or variation of transcript abundance in single cell RNAseq [22, 23]. Indeed, variation of transcript abundance has been observed to increase with decreasing copy number [22]. However, our system differs from many real world applications of high throughput sequencing, so further work will be needed to assess how our findings generalize.

Acknowledgments

The authors would like to acknowledge Barry Burbach, Ian Peikon and Diana Gizatullina for technical support and Alex Koulakov and Paul Masset for insightful discussions. Additionally, we would like to thank Hassana Oyibo who provided the inspiration for this work.

References

1. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, et al. (2010) Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. Genome biology 12.

2. Li TW, Weeks KM (2006) Structure-independent and quantitative ligation of single-stranded DNA. Analytical biochemistry 349: 242-246.

3. Zhang J, Kobert K, Flouri T, Stamatakis A (2013) PEAR: a fast and accurate illumina Paired-End reAd mergeR. Bioinformatics (Oxford, England) .

4. Shiroguchi K, Jia TZ, Sims PA, Xie X (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proceedings of the National Academy of Sciences of the United States of America 109: 1347-1352.

5. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic acids research 36.

6. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome biology 14.

7. Stolovitzky G, Cecchi G (1996) Efficiency of DNA replication in the polymerase chain reaction. Proceedings of the National Academy of Sciences of the United States of America 93.

8. Odelberg S, Weiss R, Hata A, White R (1995) Template-switching during DNA synthesis by thermus aquaticus DNA polymerase i. Nucleic acids research 23: 2049-2057.

9. Eckert K, Kunkel T (1991) DNA polymerase fidelity and the polymerase chain reaction. PCR methods and applications 1: 17-24.

10. Tindall K, Kunkel T (1988) Fidelity of DNA synthesis by the thermus aquaticus DNA polymerase. Biochemistry 27: 6008-6013.

11. Keohavong P, Thilly W (1989) Fidelity of DNA polymerases in DNA amplification. Proceedings of the National Academy of Sciences of the United States of America 86: 9253-9257.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

12. LifeTechnologies I (2002). High fidelity and high specificity accuprime pfx dna polymerase gives you both. URL `http://tools.lifetechnologies.com/content/sfs/brochures/711-021834%20AccuPrime%20Brochu.pdf`.

13. Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. BioTechniques 52: 87-94.

14. Jagers P, Klebaner F (2003) Random variation and concentration effects in PCR. Journal of Theoretical Biology 224.

15. Hassibi A, Kakavand H, Lee T (2004) A stochastic model and simulation algorithm for polymerase chain reaction (PCR) systems. Proc of IEEE Workshop on Genomics Signal Processing and Statistics .

16. Gill P (2001) Application of low copy number DNA profiling. Croatian medical journal 42: 229-232.

17. Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, et al. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic acids research 24: 3189-3194.

18. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences 99: 5261-5266.

19. Phillips J, Eberwine J (1996) Antisense RNA amplification: A linear amplification method for analyzing the mRNA population from single living cells. Methods in Enzymology 10: 283-288.

20. Zong C, Lu S, Chapman AR, Xie XS (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science (New York, NY) 338: 1622-1626.

21. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90-94.

22. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nature biotechnology 30: 777-782.

23. Brennecke P, Anders S, Kim JK, Koodziejczyk AA, Zhang X, et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. Nature methods 10: 1093-1095.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

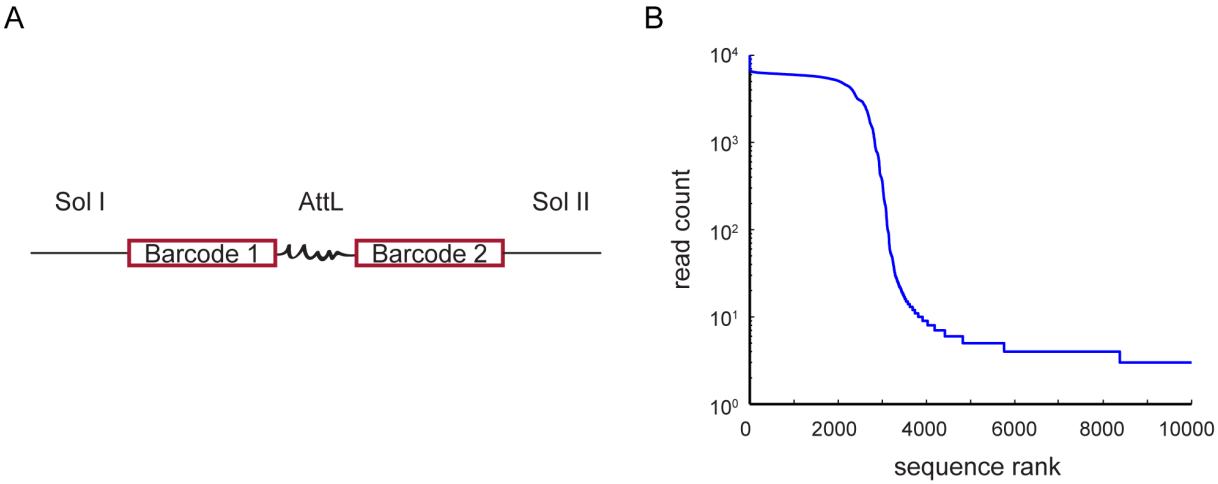

Figure 1. PCR can grossly affect sequence representation in Illumina library generation. (A) Structure of the amplicons used in this study. Each barcode is 20 random basepairs long. Sol I and Sol II are Illumina primer binding sites. (B) Result of oligo sequencing experiment, where sequences sorted by their abundance are plotted against their read counts. A wide shoulder is apparent.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

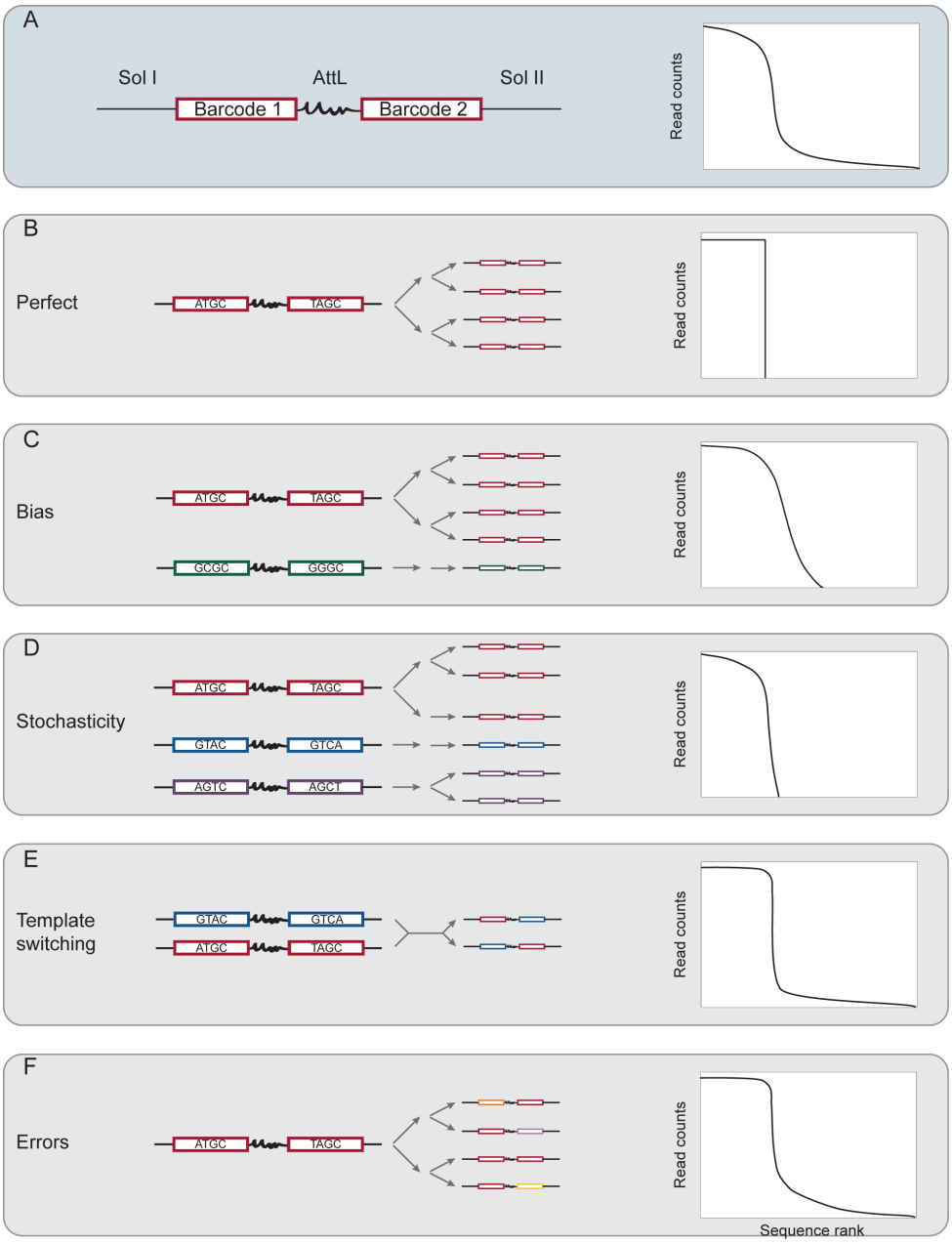

Figure 2. Errors and biases in PCR and their theoretical impact on sequence representation. (A) Structure of the amplicons used in this study and a schematic of the experimental results. (C-F) Schematic representation of perfect and different modes of skewed forms of PCR as well as their expected impact on sequencing data.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

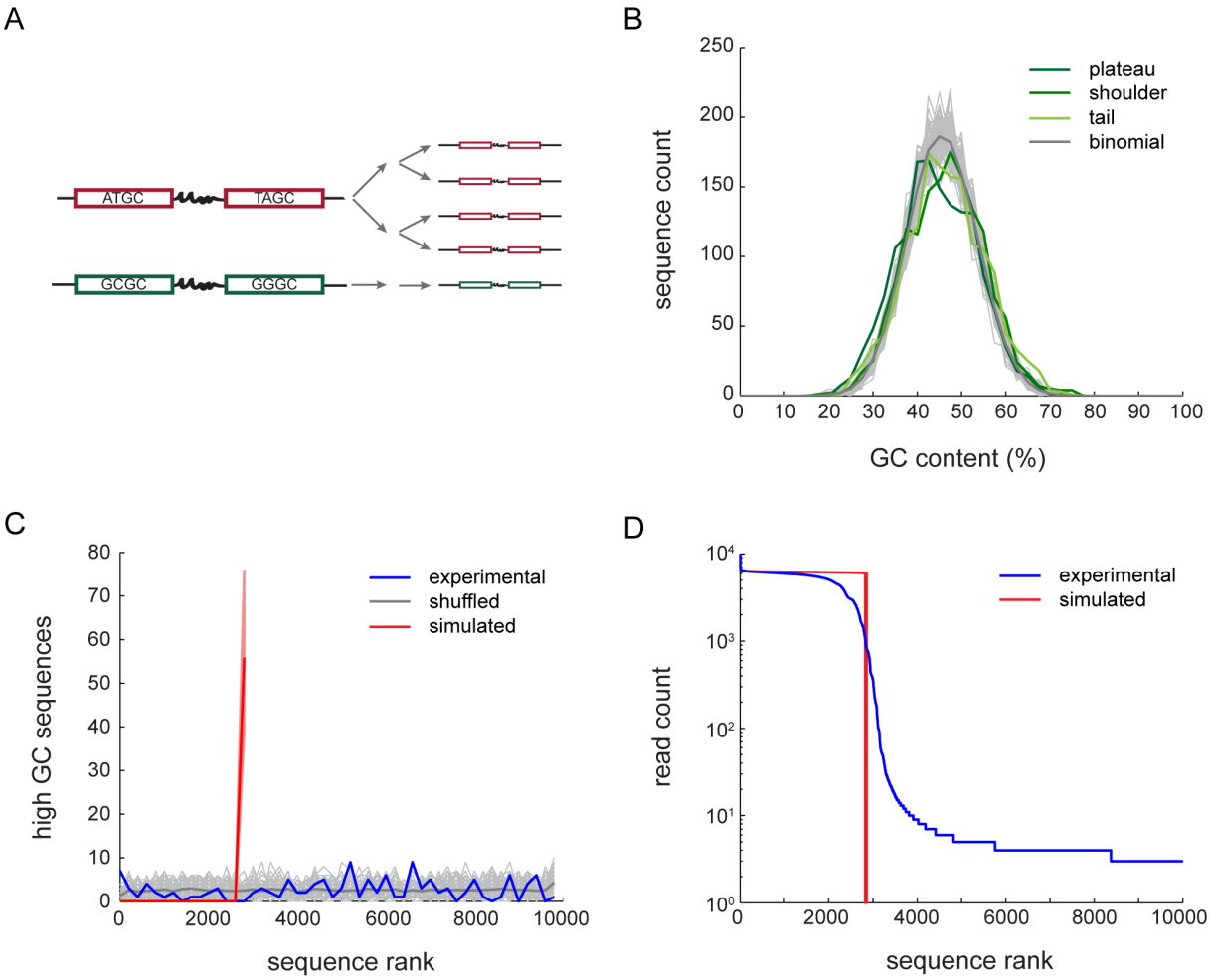

Figure 3. GC Bias. (A) Schematic of two cycles of GC biased PCR. (B) GC content distribution for 1500 sequences in plateau, shoulder and tail of the sequence trace. Compared to 100 simulations of a binomial sampling model with $P(GC) = 0.45$ (light grey) and the average (dark grey). (C) Position of sequences with a GC content $> 65\%$ in sequence rank space (blue) compared to a randomly shuffled control (100 individual iterations grey, average dark grey). The expected distribution of sequences with GC content $> 65\%$ based on a worst case simulation of perfect PCR with GC bias as reported in [1] that was applied to 2900 input sequences is plotted in red. All values are given in 200 sequence wide bins. (D) The same simulation (red) is plotted with the observed sequence profile (blue) and differs substantially.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

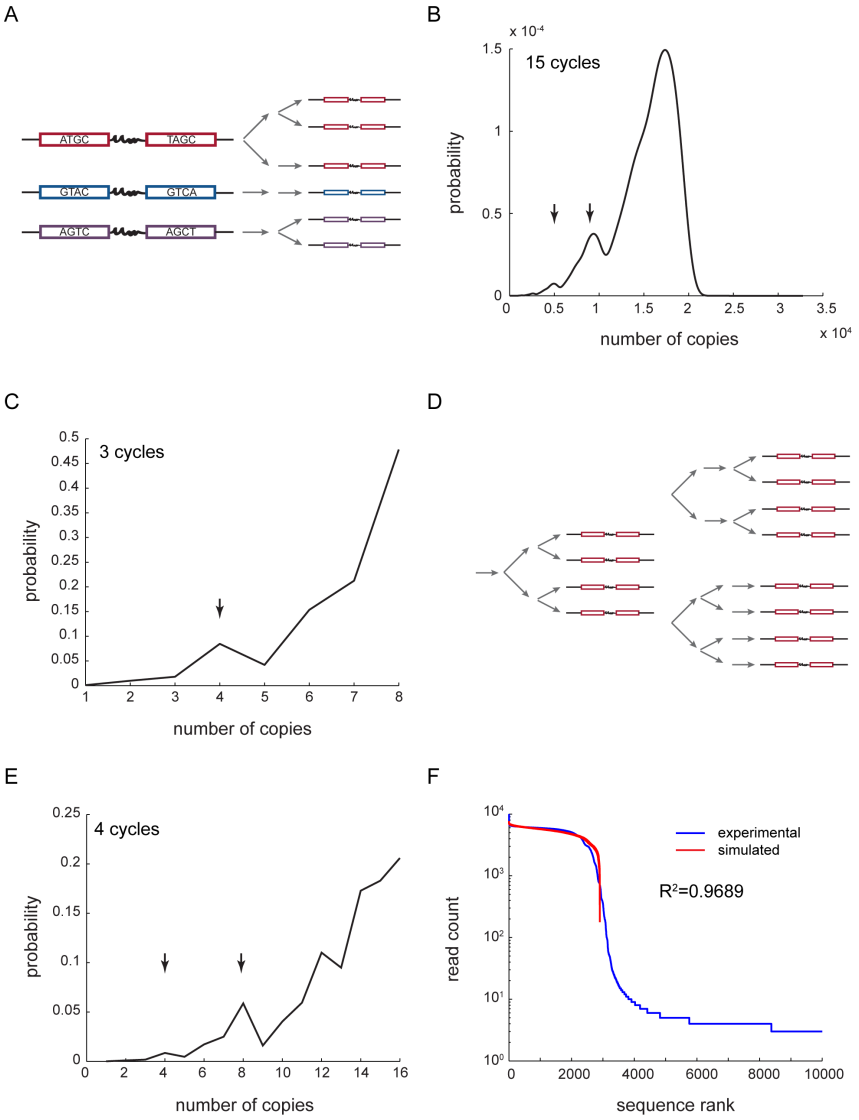

Figure 4. PCR stochasticity. (A) Schematic of two cycles of stochastic PCR amplification. (B) The exact probability distribution of sequence copy number after 15 cycles of PCR with $P(amplification) = 0.9$. Arrows indicate local maxima mentioned in text. (C) Probability distribution of sequence copy number after 3 cycles. (D) Schematic of how to obtain a copy number of 4 in three cycles of PCR. (E) Probability distribution of sequence copy number after 4 cycles. (F) A sample of 2900 sequences of the approximate probability distribution after 25 cycles of PCR (blue) correlates closely with the 2900 most abundant sequence reads of the experimental data (red)

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

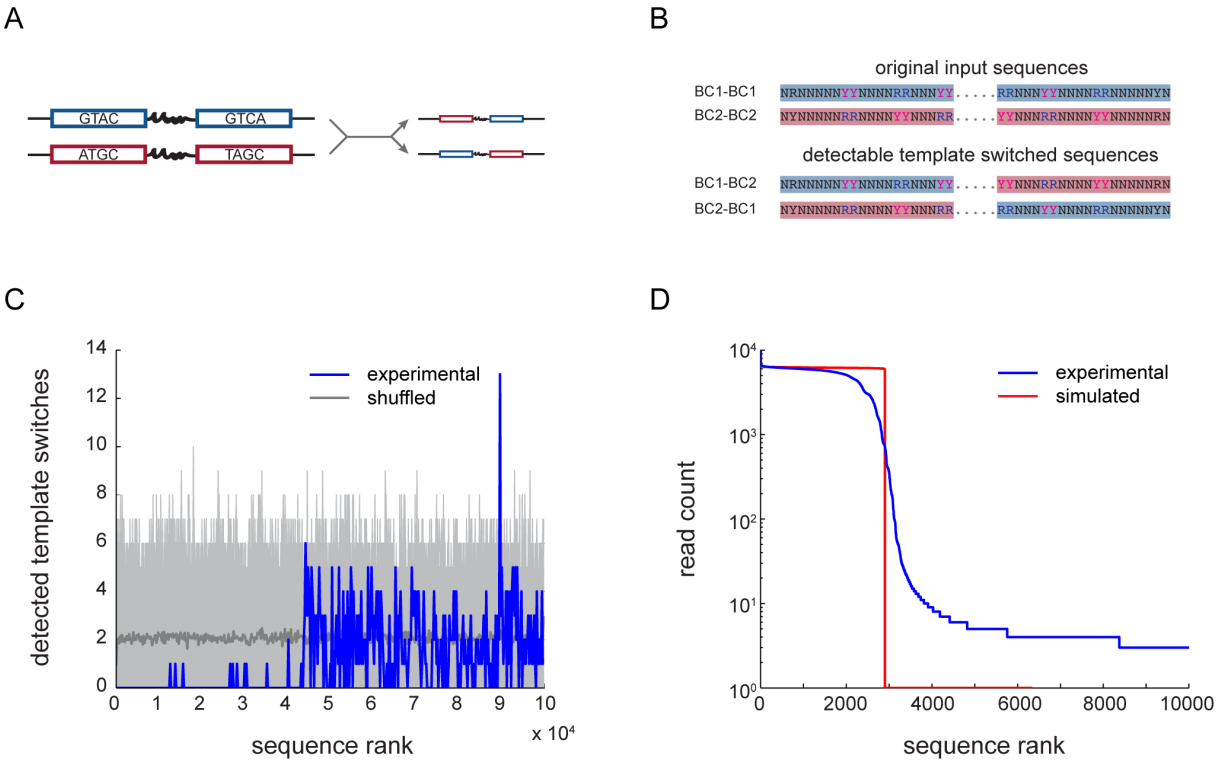

Figure 5. Template switching. (A) Schematic of one cycle of PCR with template switching. (B) Purine and Pyrimidine anchors used to detect template switches between BC1-BC1 and BC2-BC2 barcodes. (C) Position of detected template switched sequences (blue) in sequence rank space compared to randomly shuffled positions (100 individual iterations grey, average dark grey). All values are given in 200 sequence wide bins. (D) Observed sequence profile and simulated sequence profile for perfect PCR with template switches.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

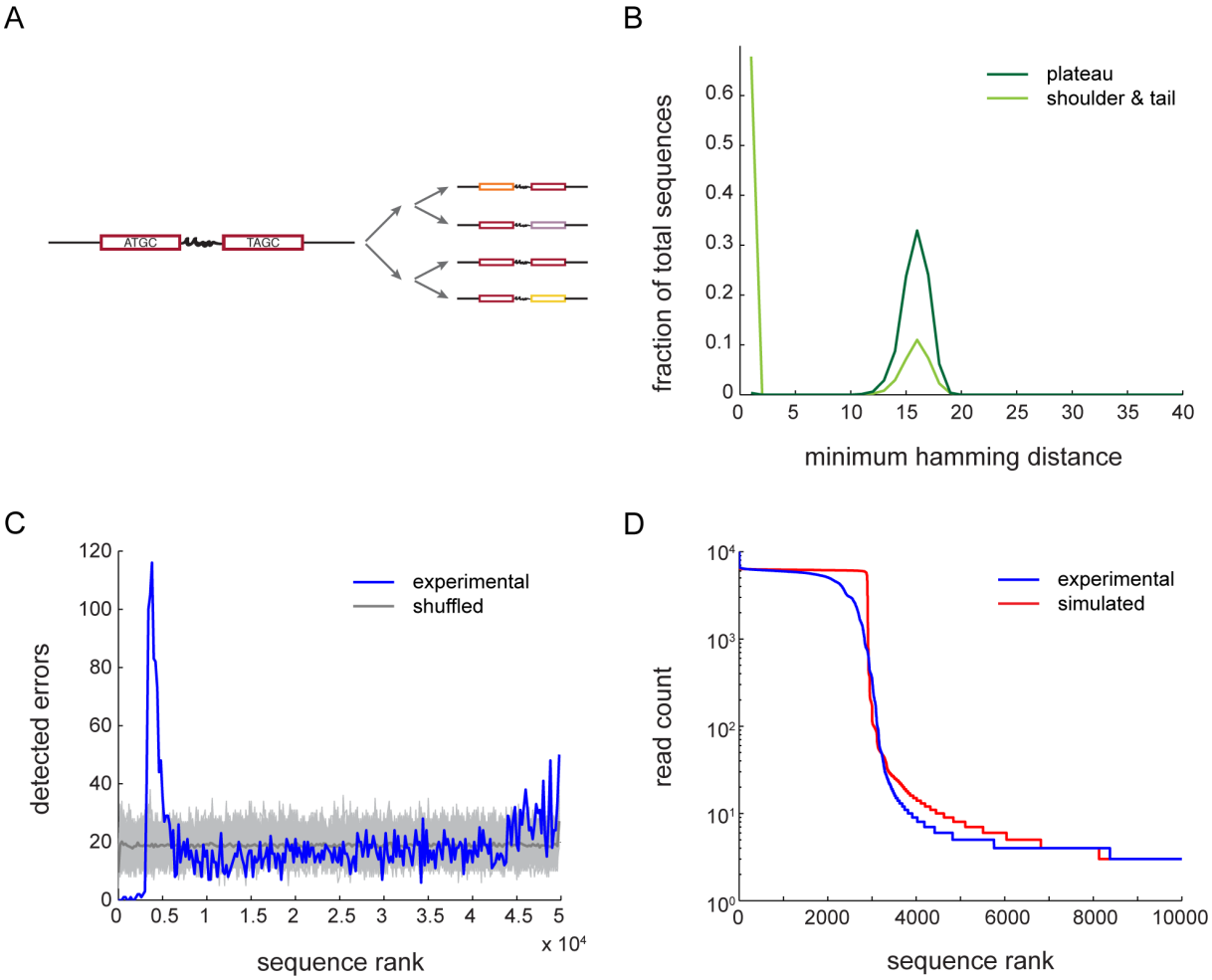

Figure 6. PCR errors. (A) Schematic of two cycles of PCR with polymerase errors. (B) Histogram of the minimum Hamming distance from sequences in the plateau to other plateau sequences (dark green) and sequences from shoulder and tail (rank 2900 to 10000) to plateau sequences (light green). (C) Position of errors detected using mismatches to anchor sequences in the barcodes (blue). These are compared to randomly shuffled positions (individual iterations grey, average dark grey). All values are given in 200 sequence wide bins. (D) Observed sequence profile and simulated sequence profile for perfect PCR with polymerase errors.

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

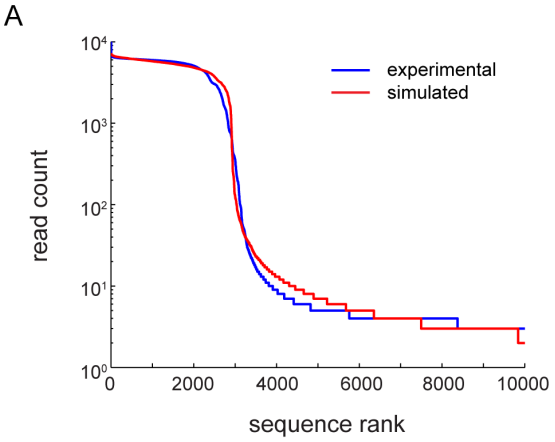

Figure 7. PCR errors and stochasticity appear to explain a large fraction of observed data. (A) PCR is simulated as a Galton Watson process with polymerase errors added on at $1.5 * 10^{-5}$ substitutions per nucleotide. Simulated and observed sequence profile match closely.

Tables

Table 1. DNA oligonucleotides used

Name	Sequence
BC1-BC1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNR NNN NNY YNN NNR RNN NYY ACG CCC CCA ACT GAG AGA ACT CAA GGG CAC GCC CTG GCA CCC GCA CRR NNN YYN NNN RRN NNN NYN
BC2-BC2	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNY NNN NNR RNN NNY YNN NRR ACG CCC CCA ACT GAG AGA ACT CAA GGG CAC GCC CTG GCA CCC GCA CYY NNN RRN NNN YYN NNN NRN
adapter	NNN NNN NNN NNN NNN AGA TCG GAA GAG CGG TTC AGC AGG AAT GCC GAG ACC GAT CTC GTA TGC CGT CTT CTG CTT G
F primer	AAT GAT ACG GCG ACC ACC GAG ATC T
R primer	CAA GCA GAA GAC GGC ATA CGA GAT C

bioRxiv preprint doi: https://doi.org/10.1101/008375; this version posted August 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplemental figures

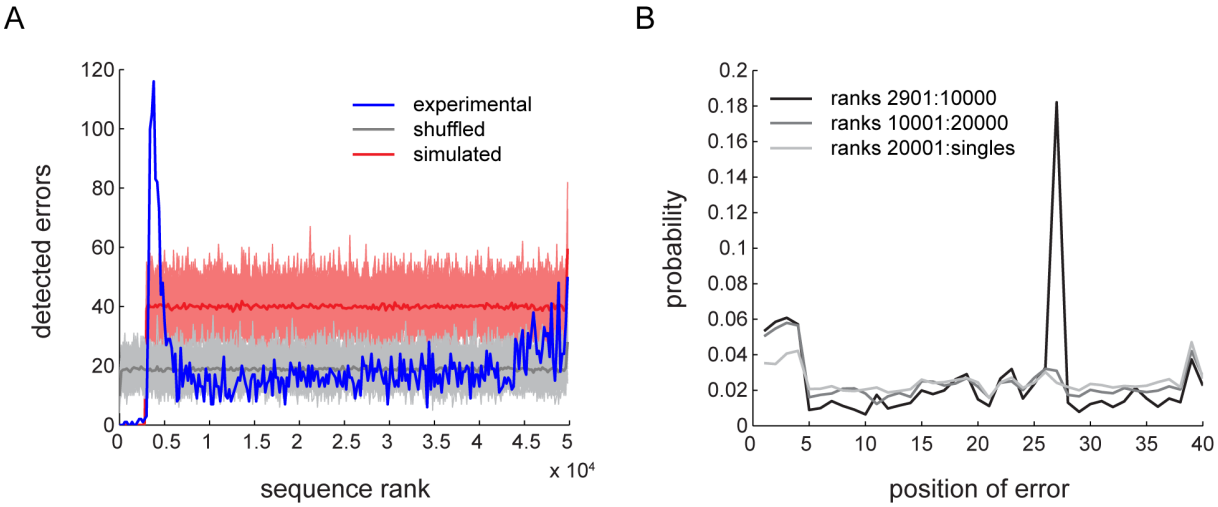

Figure S1. PCR errors are unevenly distributed in the barcodes. (A) Detection of PCR errors by using mismatches to the anchor sequences as a proxy (blue) underestimates the rate of errors as measured by Hamming distance of 1. If both Hamming distance and mismatches were preforming equally well, we would expect to detect about 40 errors per 200 sequences using the mismatch metric (red). Instead we detect about 20 (blue). (B) The position of PCR errors as detected by Hamming distance is non uniform across the barcodes and across windows in the sequence profile, explaining the mismatch observed in (A).